

2023

The Effects of Disinformation Upon National Attitudes Towards the EU and its Institutions

Alex Murphy

Follow this and additional works at: <https://arrow.tudublin.ie/scschcomdis>



Part of the [Computer Engineering Commons](#)



This work is licensed under a [Creative Commons Attribution-Share Alike 4.0 International License](#).

The Effects of Disinformation Upon National Attitudes Towards the EU and its Institutions

Alex Murphy - D19125527

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computing (TU059)

March 2023

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (TU059), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: *Alex Murphy*

Date: *28th March 2023*

ABSTRACT

This work explores the effects of misinformation and disinformation upon national attitudes towards the EU. Several nations, in particular the Russian Federation, have been working for decades to spread narratives that debase the political processes of healthy democracies around the world. There is strong evidence to show that extensive efforts have been made to disrupt the inner workings and overall membership of the EU, to support disruptive policies in the United States such that political deadlock is maintained indefinitely. These efforts are largely based on the spreading of misinformation and disinformation across social networks that have done very little to attempt to protect the discourse that they provide.

A wide range of academic work has explored this area. Efforts have, in the main, focused on utilising machine learning approaches to identify bot accounts sharing propaganda, or bot networks creating social network cascades around sensitive referenda such as Brexit, or vibrant issues such as trans rights. Other work has explored the effects on user behaviour following their exposure to falsehoods. Heretofore no work has sought to explore wider trends on a national level. This work will therefore focus on using a statistical approach to exploring the strength of the correlation between propaganda efforts and changes to national attitudes over time.

This study collects the relevant data from multiple sources and explores the quantity of propaganda efforts over a number of years by utilising the 2 datasets sourced from the EUvsDisinformation group. These data provide a detailed explanation of when and where the listed propaganda was first encountered, as well as what nations and issues it touches upon. This data is used to quantify how much propaganda was targeted in the study period, at a particular nation or centred on a particular topic. For the national perspective, Eurobarometer survey data is used over TBD years. This national overview allows us to examine to what extent the propaganda can be said to be having an effect.

The results will determine the strength of the correlation between the efforts expended by hostile actors and the resultant change in attitudes in the examined nations. Further exploration is performed using Machine Learning approaches to determine if any other insights can be gained and to strengthen the experiment conclusions.

Key Words: Misinformation, Disinformation, Propaganda, social media, Political Disruption, Brexit, January 6th coup attempt, 2016 election, Social Media Bots, Botnets, Social Media cascade, The EU, Anti-EU sentiment, k-NN, Machine Learning.

ACKNOWLEDGEMENTS

Damian Gordon for his help, support and guidance, this work could not have been completed without his insight and knowledge, his effusive confidence in my abilities as well as gentle guidance throughout the work has, in no small manner, led to its successful completion. Whatever success is contained within this document is largely due to his tireless support at all hours of the day and night.

Several teachers and staff have been instrumental during my time in TUD, Andrea Curley for her constant patience whenever I had questions and for also giving me the space to decide to switch major at the last moment. Doctor Emma Murphy for her help and support throughout the process. Doctor Deirdre Lawless for giving me the confidence to continue with college, her friendship and gentle prodding had a powerful impact on my ability to continue.

It is only through the efforts of others that I have been able to climb so far. Brighter shores beckon, perhaps greater achievements are yet to come but certainly, it will only be with the help of my family, friends, and teachers. I thank them all, they have made me who I am.

DEDICATION

This work is dedicated to my family. My wonderful wife Xiaocui Zhou, who has been by my side throughout the four years of blood sweat and tears that have been liberally expended in the pursuit of this masters. I could not have done it without her and would not be able to enjoy the future that it is sure to create for us if she were not my partner.

I also dedicate the work to my young daughter Freya Murphy; little does she know the chaos she brought into the house when she arrived with us in the summer of 2022! I will constantly remind her, in years to come, how often she kept me awake for hours at night while I was desperately trying to complete some work. Her smiles and hugs have been a great tonic throughout the process, those precious moments sustained me on the darkest days of the masters.

Finally, to my parents, Garry and Jennifer Murphy who offered their services in babysitting Freya from January 2023 to give me the space to try and get as much work done as possible. In no small part, they have provided the structure that was needed week after week to be able to attempt something on this scale.

To each of them I say this: I could not have done this without you. You were my rock and I only hope your faith will be rewarded as I finally set out to achieve some the potential that I was lacking in my earlier years.

Alex Murphy 02/03/23

Table of Contents

DECLARATION	2
ABSTRACT	3
ACKNOWLEDGEMENTS	5
DEDICATION	6
TABLE OF FIGURES	9
1. INTRODUCTION	11
1.1. PROJECT BACKGROUND.....	11
1.2. PROJECT DESCRIPTION	13
1.3. PROJECT AIMS AND OBJECTIVES	16
1.4. THESIS ROADMAP.....	16
2. LITERATURE REVIEW	18
2.1. INTRODUCTION.....	18
2.2. MISINFORMATION OVERVIEW	18
2.3. BREXIT AND THE TARGETING OF MAJOR EVENTS	20
2.4. USER BEHAVIOURS.....	22
2.5. TWITTER, FACEBOOK, AND MACHINE LEARNING APPROACHES	23
2.6. KNOWLEDGE GAPS.....	25
2.7. KEY REFLECTIONS	26
3. PROJECT DESIGN	27
3.1. INTRODUCTION	27
3.2 ADDRESSING THE RESEARCH QUESTION	28
3.3. DATA SELECTION	28
3.4 THE DATASETS	30
3.4.1. <i>EUvsDisinformation Dataset</i>	31
3.4.2. <i>Kaggle EUvsDisinformation data</i>	36
3.4.3. <i>Eurobarometer Data</i>	41
3.4.4. <i>Exploratory analysis</i>	54
3.5. THE KAGGLE DATASET	68
3.6. MACHINE LEARNING APPROACHES	68
3.7. KEY REFLECTIONS	69

4. PROJECT DEVELOPMENT	71
4.1. INTRODUCTION.....	71
4.2. THE EUvsDISINFORMATION DATASETS.....	71
4.3. THE EUROBAROMTER DATASET.....	73
4.4. EUROBAROMTER DATA FINAL STEPS.....	74
4.5. EUROBAROMTER DATA PREPARATION.....	77
4.6. PLOTTING THE EUROBAROMTER DATA.....	81
4.7. EXPLORATORY STATISTICS.....	89
4.8. MACHINE LEARNING.....	108
4.9. KEY REFLECTIONS.....	111
5. RESULTS AND EVALUATION	112
5.1. INTRODUCTION.....	112
5.2. THE DATASETS.....	112
5.3. CHOOSING A STATISTICAL APPROACH.....	114
5.4. THE MANOVA TEST AND RESULTS.....	120
5.5. MACHINE LEARNING RESULTS.....	127
5.6. KEY REFLECTIONS.....	135
6. CONCLUSIONS AND FUTURE WORK	136
6.1. INTRODUCTION.....	136
6.2. PROBLEM DEFINITION.....	136
6.3. LIMITATIONS.....	137
6.4 DESIGN CHOICES.....	138
6.5 FUTURE WORK.....	139
6.6. THE NECESSITY OF THIS WORK.....	140
BIBLIOGRAPHY	142
APPENDIX A	151

TABLE OF FIGURES

FIGURE 3-1- STAGES OF THE EXPERIMENT	27
FIGURE 3-2 - THE EUvsDISINFORMATION MAIN PAGE	29
FIGURE 3-3: THE EUvsDISINFORMATION WEBSITE DATABASE AS DISPLAYED THROUGH FIREFOX ON 10/02/23.....	33
FIGURE 3-4: THE TQDM FOR LOOP USED FOR THE WEBSCRAPING.	33
FIGURE 3-5: THE TQDM PROGRESS BAR, AT THE END OF THE PROCESS IN THIS CASE.	33
FIGURE 4-1- GLIMPSE OF THE DATA TO BE USED FOR A K-NN APPROACH.	108
FIGURE 4-2- THE STRUCTURE OF THE DATA USED IN BOTH THE EXPERIMENT AND THE K- NN APPROACH.	109
FIGURE 4-3- SHARE OF THE COUNTRY FACTOR DATA IN THE MACHINE LEARNING DATAFRAME.	109
FIGURE 4-4- PRE-PROCESSING COMMAND. IN THIS CASE APPLIED ON COLUMNS WITH INDEX POSITION 2:.....	109
FIGURE 4-5- APPLICATION OF PRE-PROCESSING TO THE TEST DATA. IN THIS CASE THE IGNORED COLUMN IS THE TARGET COLUMN.	110
FIGURE 4-6- PRE-PROCESSING APPLIED TO TEST DATA.	110
FIGURE 4-7- SUMMARY OF K-NN TEST DATA AFTER PRE-PROCESSING HAD BEEN APPLIED.	110
FIGURE 4-8- TRAIN AND TEST DATA SPLIT.	111
FIGURE 5-1- THE DATES_SCRAPE DATASET.....	114
FIGURE 5-2- CURRENT STRUCTURE OF DATES_KAGGLE	115
FIGURE 5-3- QUESTION D71, IE DATA PIVOTED WIDE	115
FIGURE 5-4- ADD COLUMNS TO DATA BEFORE TESTING.....	116
FIGURE 5-5- THE MEMBER STATE'S COUNTRY CODE HAS BEEN ADDED. NA VALUES WERE UNUSED MISSING PERCENTAGE VALUES AND CAN BE IGNORED.	116
FIGURE 5-6- THE ORIGINAL RBIND OF INDIVIDUAL RESPONSES AND THE EVENTUAL DESIGN OF ALL MEMBER STATE'S RESPONSES IN A SINGLE DATAFRAME	117
FIGURE 5-7- THE TEST DATAFRAME DISPLAYING THREE OF THE FOUR MEMBER STATES IN THE COUNTRY COLUMN.	117
FIGURE 5-8- MANOVA SUMMARY FOR QUESTION QA8	120
FIGURE 5-9- SUMMARY OF THE ANOVA FOR QUESTION QA8.....	121

FIGURE 5-10- MANOVA SUMMARY FOR QUESTION D71	122
FIGURE 5-11- SUMMARY OF ANOVA FOR QUESTION D71	122
FIGURE 5-12- MANOVA SUMMARY FOR QUESTION D73.....	123
FIGURE 5-13- SUMMARY ANOVA FOR 4 OF THE RESPONSES TO QUESTION D71	124
FIGURE 5-14- SUMMARY OF ANOVA FOR 4 OF THE 9 RESPONSES TO QUESTION D78. 126	
FIGURE 5-15- SUMMARY OF ANOVA FOR THE LAST OF THE 9 RESPONSES TO QUESTION D78.....	127
FIGURE 5-16- EXAMPLE TRAINING DATA AFTER PRE-PROCESSING AND A TRAINING AND TEST SPLIT.....	128
FIGURE 5-17 - EXAMPLE TEST DATA AFTER PRE-PROCESSING AND A TRAINING AND TEST SPLIT.	128

1. INTRODUCTION

1.1. Project Background

Propaganda, misinformation, disinformation, and conspiracy theories have found fertile ground in social media. Indeed, one of the unintended consequences of the spread of internet literacy, once thought to be the bastion of a freer, more informed society, has been the speed with which rumour and outright lies can spread across social media networks (Thompson, 2018). That recent examples such as hydroxychloroquine & Ivermectin so easily entered our cultural consciousness, and with such rapidity; is a testament to the ease with which rumour can spread in our online social spaces (Wang *et al.*, 2021). Suddenly family and friends could be exposed to insidious narratives that proved immune to all reason, that cascaded through social media networks, that evolved and developed to become a serious threat to an effective global response to COVID-19 (Karhu *et al.*, 2022).

The ease with which these narratives travelled through social media and their persistence when confronted with science and scepticism will fuel academic study for decades to come. Of note however, is the source of many of these narratives is seen to have been nations eager to disrupt the healthy democracies of the world (Eady *et al.*, 2023). There is significant evidence that coordinated and large-scale efforts were undertaken by Russian organisations, to disrupt Western nations; researchers have been focused on issues such as how the information travelled through social networks, how can these messages be auto detected through Machine Learning, or what strategies best counteract the effects of the messages; there has been little to no work on the Marco scale effects of these propaganda efforts beyond those of the individual user themselves (Mocanu *et al.*, 2014). Issues such as Brexit and elections in the United States, France and Italy, to name a few; have been targeted in the lead up to the polling day, in some cases, years in advance (Bruno *et al.*, 2021). However, the national impact of these propaganda efforts is still largely unexplored. This work seeks to address this with a preliminary investigation of how national attitudes towards the EU have changed in those nations that have been targeted by propaganda to a larger degree.

Research Background, Context, and Scope

With the advent of social media in the early 2000s, we created a new method to share vast amounts of data, much of it private, at a rate never seen before. The implications of this new media space were poorly understood in the early days of this technology, however companies such as Facebook and Twitter have been able to leverage this media space into vast, somewhat opaque, social media empires (Stukal *et al.*, 2019). While these companies have been expanding and profiting publicly by utilising the data that has been freely given by users'; other disruptive entities have woken up to the potential of how these networks can be used for their purposes.

Recent events such as the COVID-19 pandemic, necessitated massive social media campaigns to educate the public (Venegas-Vera *et al.*, 2020, Tasnim *et al.*, 2020) while also sparking significant changes to users' social media behaviour (Puri *et al.*, 2021, Haggag *et al.*, 2021, Karhu *et al.*, 2022). Beyond public health, less benevolent tools have also been developed. Propaganda, misinformation, and disinformation have found fertile ground in social media and are capable of spreading at unprecedented rates (Howard *et al.*, 2019, DiResta *et al.*, 2018).

Disinformation and propaganda are prevalent on Twitter (Wang *et al.* 2021) and often utilise networks of bots (a computer program that performs automatic repetitive tasks... especially: one designed to perform a malicious action)¹ have been created specifically to spread propaganda across social media (Sanovich, 2017; Howard & Kollanyi, 2016) or to target specific hot button issues such as Brexit (ibid). Recent political upsets such as Brexit, the election of Tump and the attempted American coup, require a deeper understanding of how propaganda and disinformation spreads through social media and the effects it has upon national politics.

There is a wealth of academic work available regarding disinformation in social media, the obvious next question is, what has been the effect of disinformation campaigns that have been uncovered? Do higher or increasing levels of disinformation have a significant affect upon the targeted populations and if so, can that affect be measured?

Attempts to sanitise social media spaces of disinformation and propaganda are in their infancy but have largely relied on Machine Learning (Bruno *et al.*, 2021), human moderation, or some combination of the two. Paradoxically, the difficulty of

¹ <https://www.merriam-webster.com/dictionary/bot>

removing a falsely held concept or belief, whether through exposure to disinformation and propaganda, is not as easy as simply correcting the information, indeed the user may be *more* resistant to truth when told that what they are consuming is a falsehood (Garrett & Weeks, 2013). The challenges of countering propaganda and the rapidity with which social media cascades² can be sparked have overwhelmingly been beyond these efforts at mitigation.

Researchers are catching up to the scale of the problem but efforts thus far, have focused on small scale examples (Bastos & Mercea, 2019), or contrived situations (Bail *et al.*, 2020) that have failed to find any significant effects upon users that are already strictly partisan. Other work has focused on the effects on users' behaviour (Conover *et al.*, 2011; De keersmaecker & Roets, 2017).

This leads to the obvious question. What evidence is there for the large-scale effects of these disruptive efforts? Can we see changes in attitudes following concerted efforts made by propagandist organisations? In this work, we seek to address the lesser focus on the larger scale and longer-term effects of propaganda and to assess to what extent exposure can be seen to have a dosage response.

1.2. Project Description

This work explores the correlation between those EU member states that have been targeted by higher levels of propaganda as demonstrated in the publicly available EUvsDisinformation database; and those states' attitudes towards the EU over time as measured in the standard Eurobarometer survey over the study period.

It seeks to measure whether any discernible variability can be seen in those nations that have been targeted to a greater degree than other states which were less targeted. Can any statistically significant difference be seen in member state's attitudes towards the EU after exposure to higher levels of propaganda when compared to other EU members?

There is a growing body of research on quantities and types of disinformation that pervade social media (Llewellyn *et al.*, 2018, Vosoughi *et al.*, 2018), this work seeks

² A social media cascade refers to a situation where a piece of content (such as a post, tweet, or video) gains momentum and spreads rapidly across social media platforms through a chain reaction of shares, likes, and comments. As more and more people engage with the content, it reaches a wider audience and can quickly become viral.

to measure the effects, if any, on EU member states. This necessitates several assumptions. Firstly, in the case of EUvsDisinformation data, that the data available is some indication of the ground truth as regards quantities of propaganda and who or what was targeted in each instance; and secondly, that the Eurobarometer survey can be said to be a reasonably accurate representation of a member state's citizen's attitudes towards the EU while having sample sizes of no greater than circa 1000 citizens per member state per year.

Eurobarometer surveys are a long-established tool of research regarding the EU (Skirka *et al.*, 2020; Boldureanu *et al.*, 2020;) and can comfortably be assumed to be some measure of the ground truth, in so far as possible. The EUvsDisinformation data is curated by a politically motivated entity in response to the findings of East StratCom Taskforce³, its data, while comprehensive, is slightly more biased than Eurobarometer group. However, the data is of reasonable fidelity and is of sufficient quality to warrant its' inclusion in this work.

There are several limitations with survey data of this kind, including questions such as, how much faith can be placed in social media as a representation of users' attitudes? Hargittai (2015) argues that a users' choice of social media service is indicative of a broad range of their behaviours and how "internet savvy" they are. The researchers suggest that a user's social media of choice necessarily biases the research that can be based on such data.

Very little academic work on disinformation has focused on the larger scale of national trends. Researchers have focused on specific domains such as Twitter (Dutta *et al.*, 2021), issues such as Brexit, (Bruno *et al.*, 2021, Great Britain & Intelligence and Security Committee Report, 2020), or national elections (Ferrara, 2017, Bingle *et al.*, 2020). Larger scale and longer-term effects are rarely studied. This work seeks to remedy that.

The quality of the different datasets is not on equal footing. While the Eurobarometer dataset is of the highest quality and the basis for a wide range of academic work, details regarding the EUvsDisinformation data regarding collection and quality of the data are not easy to acquire. The organisation itself (*EUvsDisinfo*

³ https://www.eeas.europa.eu/eeas/questions-and-answers-about-east-stratcom-task-force_en

project, part of the European External Action Service: East StratCom Task Force)⁴ is a part of a European council's efforts to 'challenge Russian's ongoing disinformation campaigns'⁵.

Finally, while there is a wide selection of work related to the United States and other parts of the world in this domain, this work is delimited to EU data only. Additionally, the data is limited to the years 2015 to 2022 only as this captures key recent events including Brexit (2016), the French presidential election (2017) and the COVID-19 pandemic.

In summary, this project will utilise three key datasets:

1. The EUvsDisinformation database publicly available on the project webpage
2. An EUvsDisinformation dataset that was made available on Kaggle that contains data from a narrower date range but with much greater detail.
3. Several years of Eurobarometer survey data with questions selected as they pertain to national attitudes towards the EU.

The project firstly will quantify the amount of disinformation that has been targeted at individual EU nations over time. Did nations such as the UK pre-Brexit, and France and Italy before various elections during the period captured by the data, have an increase in the disinformation that was targeted towards them? Those nations that have significantly more disinformation targeted towards them, will then be explored in the more detailed Kaggle dataset to determine whether there are other patterns and strategies being employed.

Finally, once patterns have been discerned in the disinformation databases, those nations will have their survey data examined in the Eurobarometer data to explore whether there was a statistically significant change in their attitudes towards the EU over the period that correlates with the timing of the disinformation campaign. Following the completion of the statistical testing required to test the research question, an appropriate Machine Learning technique will also be applied to explore further evidence in support of the findings.

⁴ *ibid*

⁵ <https://www.consilium.europa.eu/en/press/press-releases/2015/03/19/conclusions-russia-ukraine-european-council-march-2015/>

1.3. Project Aims and Objectives

This project aims are as follows:

1. Identify and collect all relevant datasets and attempt to quantify the amount of disinformation that nations were targeted with.
2. To explore whether there are any discernible patterns in the more detailed disinformation data available in the Kaggle dataset. The years covered by this dataset are much more limited in scope but contain much richer data. Distinctive styles and patterns will be discovered such as what key topics were targeted for these propaganda efforts.
3. Those nations who have been the target of disinformation will have their Eurobarometer data explored to see whether there has been a statistically significant correlation between their changes in attitude, and the amount of disinformation they were targeted by.

1.4. Thesis Roadmap

The second chapter, the *Literature Review* chapter, examines current academic work in this area and establishes the study within the context of that research as well as identifying some gaps in the current body of work.

The next chapter, the *Project Design* Chapter, describes the gathering, exploration and preparation of the data and details why and how the data has been prepared in such a manner. The different approaches required due to the disparate data sources is examined and justified in detail with images used to clarify the approach used in challenging areas or to illuminate areas of interest to the research question.

The *Project Development* Chapter details the implementation of statistical exploration of the data and examines what can be said about the data prior to the later statistical tests. Detailed examination of the data is conducted and explained along with images of pertinent results to help illustrate the structure of the data overall.

The *Project Evaluation* Chapter presents the results, and they are examined in the context of the previous chapters to help quantify to what extent the results demonstrate

and support the hypothesis. Detailed examination of the results on a question-by-question basis is produced in this chapter.

The Final Chapter, is the *Conclusions and Future Work* Chapter that details the conclusions that have been reached based on the results obtained in the experiment and discusses the implications for future work conducted on the basis of the current database, including various other approaches that can be used to further strengthen the results, as well as other research questions raised by the results.

There are also additional appendices that contain any images and plots that were generated in the completion of this work but that were not used in the main body of the text.

2. LITERATURE REVIEW

2.1. *Introduction*

This chapter will explore recent work in the area of disinformation and will highlight some of the main approaches as well as seminal works in this area. A discussion of the findings and implications of the literature in this area will also be presented.

2.2. *Misinformation Overview*

Wang *et al.* (2021) explored the effects of exposure to misinformation on Twitter users and through an exploration of language distance analysis, found that such users did display a statistically significant change in the language use overall, with repeated exposures leading to a more pronounced change. Similarly, Dutta *et al.* (2021) explored the change in Twitter user's behaviour following exposure to the Internet Research Agency (IRA), a known Russian internet propaganda organisation (Prier, 2017, Eady *et al.*, 2023); and presented further supporting evidence that exposure to propaganda accounts and tweets, had a statistical effect upon user's behaviour and that direct exposure to the IRA tweets had a greater and longer lasting effect overall. Dutta *et al.* references details released by Twitter following the 2016 United States Presidential Election, that indicates that 3,841 bot accounts were deployed on twitter. (*Update on Twitter's Review of the 2016 US Election*, 2018) This led to over 1 million users engaging with messages posted by those accounts, which then led to a further 73 million engagements (Thompson, 2018) all based on posts that originated with bots. DiResta *et al.* also examined the tactics and known activities of the IRA. (DiResta *et al.*, 2018) They note that social media providers have been far from forthcoming in sharing the activities of organisations such as the IRA.

As a counter to this and other findings in this area, Bail *et al.* (2020), found that exposure to IRA tweets may not have been all that effective in changing user's attitudes. In a study of over one thousand Democratic and Republic voters and a Bayesian regression tree approach, they suggest that there is no evidence of a significant impact on user's attitudes and suggest that perhaps the users targeted by IRA operatives may have not been particularly suitable to this type of approach.

However, their work is in the minority and the vast majority of the work in this area has found significant affects and changes in user's behaviours following exposure to misinformation, disinformation and propaganda.

Mocanu *et al.* (2014), in a robust study of 2.3 million Facebook users who had consumed information outside the curated mainstream media and found that while the quality of the information that these users consumed was not on a par with curated media, the ideas spread over Facebook and had the same longevity. They present strong evidence that misinformation, disinformation and propaganda are effective tools that have equal staying power as more accurate information.

Bakshy *et al.* in their 2015 article examined over 10 million deidentified US Facebook account to examine how these users shared their views. They explored how it is a user may get exposed to an alternative viewpoint and suggest that homogenous groupings of like-minded users, such as family groups, and the Facebook news feed, are the two most common methods for sharing of 'cross-cutting' information. They note that Facebook differs somewhat from other social media in that the structure of groups tends to be focused on friends and family as opposed to political beliefs or shared hobbies as is often the case in Twitter. They suggested that Facebook users may have up to 20% of their connections as being from ideologically different users, presenting a broad range of opportunities to be exposed to differing viewpoints.

Conover *et al.* in a 2011 paper suggested that users purposefully inject alternative views into the social media landscape of users with alternate viewpoints. This was in contrast to the highly partisan political tweets produced in the run up to the 2010 US congressional elections. They differentiated between the retweet and mention tools of Twitter and found that while retweets are highly polarised, mentions were not. Mentions were seen to cross ideological lines, but the effects were not explored in this work.

'Fake news': incorrect but hard to correct by De keersmaecker & Roets (2017) conducted a study on the impact of cognitive ability of users who are exposed to information that corrects incorrectly held beliefs that they previous professed. They found that the user's cognitive ability overall correlated with their ability to adjust their views based on new information. The implications on how difficult it is to correct misinformation, disinformation and propaganda are clear. A further examination of the difficulties of expunging misinformation was explored in a 2015 paper by Exker *et al.*, who argue that the primacy of the information is the main deciding factor of user

credence. While later information may be presented that counteracts previous information, it is more difficult for a user to replace a primary interaction with a later interaction even when that interaction suggests that earlier interactions are falsely held beliefs.

2.3. Brexit and the Targeting of Major Events

Several researchers have explored Brexit as a political event that was ripe for the employment of misinformation, disinformation and propaganda strategies. A wide range of work has focused on this contentious political issue. A 2021 paper by Bruno *et al.* 'Brexit and bots: characterizing the behaviour of automated accounts on twitter during the UK election' (Bruno *et al.*, 2021), examines 10 million tweets from 1 million users from 20th November 2019 to the 23rd December 2019 during and just after the UK general election (12th December). They found that almost 10% of the total users tweeting in the week of the election were bots, this was an increase from the usual background level of 2% found in their data. Similarly, Howard & Kollanyi (2016) uncovered a truly massive number of tweets that were strongly partisan, were targeted towards one or other extreme of the Brexit debate and jumped massively in number during the same period as found by Bruno *et al.* (2021).

Professor Philip Howard, professor of internet studies at the Oxford Internet Institute⁶, has produced a significant body of work in this area and is the lead author of one of the seminal works that explored the disruptive use of bots in the lead up to the Brexit referendum (Howard & Kollanyi, 2016), finding that leave hashtags proliferated through social media networks on a much larger scale than remain hashtags, that there were differing levels of automation between the two sides and that only 1% of the accounts overall produced over 30% of the total tweets. The effects of disinformation and propaganda in the lead up to the 2016 US election were also investigated. The work examined whether polarising content was being purposefully targeted at swing states. They found that overall Twitter users were sharing more 'misinformation, polarizing and conspiratorial content than professionally produced new' (ibid). They concluded that such content was targeted at swing states more than uncontested states. They presented their expert opinion to the United States senate in 2019 exposing IRA

⁶ <https://www.oii.ox.ac.uk/people/faculty/>

efforts to exploit division and create disruption often by supporting and empowering both sides of a political issue such as Black Lives Matter (S. Rep. No. 10, 2019), going so far as to organise opposing rallies on both sides of the political spectrum in order to create disruption (ibid), they also found that peaks in disruptive activity often coincided with major events in the United States political calendar (ibid).

Bastos and Mercea in a 2019 paper found that a large number of bot accounts were activated in the weeks leading up to the Brexit referendum and that those accounts almost immediately ceased to function following the vote. They identified 13,493 (ibid) accounts tweeted in relation to the referendum that disappeared immediately afterward. They suggest that these bots are extremely effective at producing content that moves rapidly through social media networks and that clusters of bots can be effective in simulating consensus by sharing among other bot accounts and networks. A recent paper by Eady *et al.* (2023) corroborated these findings and presented evidence that exposure to misleading tweets increased with the proximity to the 2016 American Presidential election. Their work however did not find evidence supporting a change in users' voting intentions or attitudes across a range of topics, following exposure to IRA content. They also found that a small number of users was responsible for over 70% of the exposures uncovered.

Other nations and political events have also been the focus of misinformation, disinformation and propaganda campaigns. Ferrara (2017) examined disinformation in the lead up to the 2017 French Presidential election through exploring a locally known political issue dubbed '*MacronLeaks*'. Following the use of machine learning techniques and cognitive behavioural modelling based on a database of 17 million tweets; they uncovered that the users who engaged in the sharing of hashtags related to *MacronLeaks* were usually members of international alt-right Twitter communities as opposed to national French voters. They also presented evidence that a significant number of bots that had previously been utilised in the promotion of Donald Trump in the 2016 presidential election were reactivated and utilised in the run up to the French election.

Forelle *et al.* (2015) similarly present evidence of IRA attempts to affect public opinion in Venezuela and found that while IRA bot accounts overall constitute only a small percentage of the overall political tweeting, they still have a disproportionate part to play in radical opposition.

Historic global events such as COVID-19 have also been seen to be fertile ground for misinformation, disinformation, and propaganda. In a 2020 paper, Jamieson and Albarracín found that the media that a user consumes correlates with the accuracy of a user's understanding of COVID-19. They also found that conservative media tended to express conspiratorial and anti-science views. Karhu *et al.* (2022) in an online survey of 172 individuals, found an increase in social media overall that coincided with the COVID-19 pandemic and that an increased awareness of division around health policies implemented during the COVID-19 pandemic emerged.

2.4. User Behaviours

Endres and Panagopoulos, in a 2015 paper expressed a counter to the view that 'Cross-pressured partisans'⁷ are commonly viewed as persuadable'. They found that while a significant effort is made to contact and impact cross-party partisans in elections in the United States, it appears from their research that the effect is not to change the subject's partisan viewpoints but rather perhaps to disrupt their participation in the election. They noted an increase in abstentions and spoiled votes following exposure to cross-partisan views conducted in the study. This suggests that while misinformation, disinformation and propaganda may not actively change a subject's views in their entirety, it may be enough to disrupt their participation in the political process itself. Supporting evidence was presented by Iyengar & Westwood of Stanford and Princeton universities (2015) in their study on the effects of polarization and extreme views held by in-group members when considering out-group members.

In a paper focused on the effects of conspiratorial thinking, Imhoff *et al.* (2021) present evidence in support of the view that belief in conspiracy correlates with a subject's willingness to pursue their political goals through non-standard approaches, such as violence or terrorism.

O'Callaghan *et al.* (2013) explore the existence and strength of a 'filter bubble' in relation to right wing media as available on YouTube using a Non-Negative Matrix Factorization approach and explore both English and German examples. They argue

⁷ People with partisan political views that are exposed to views from the polar opposite of their political perspective.

and present evidence for the existence of a filter bubble for extreme right-wing views that is facilitated by the YouTube ‘related videos’ algorithm. Thus, a user who is exposed to an extreme right-wing video can find themselves rapidly drawn into a deeper pool of such videos.

Ördén argues in their 2022 paper that securing cyberspace from threats such as disinformation requires a new approach that they dub ‘Cyber Sovereignty’. They argue that the very nature of the international internet, beyond a nation’s borders, creates unique challenges in protecting users from exposure to disruptive content.

Prelog & Bakić-Tomić, in an astute and succinct 2020 study explore the difference in reactions to disinformation between younger and older generations. They found that younger generations were largely unconcerned with sharing and spreading fake news among their social networks and present a worrying picture of how little the truth of the media, shared by younger users, concerns them.

2.5. Twitter, Facebook, and Machine Learning approaches

Many of the Machine learning approaches to identify misinformation, disinformation and propaganda in the Twitter ecosystem stem from a DARPA competition that challenged participants to find the most effective tool to find such tweets in real time (Subrahmanian et al., 2016)

Twitter, by releasing the public API⁸ have become the mainstay of academic work in Machine Learning approaches to misinformation, disinformation and Propaganda such as classification, identification or sentiment analysis. The API constitutes at most 1% of the total tweets in the network at any one time, but it has become the bedrock for a wide range of research. During this study, Twitter was unexpectedly purchased and the ability to access the API freely has now come into question (Stokel-Walker, 2023) with a new paid system being introduced. This is sure to have drastic effects on the ability for researchers to produce good quality science on a similar scale as has been the case in the last decade.

Sanovich in their 2017 working paper, and later in a 2019 policy memo along with other researchers (Stukal *et al.*, 2019), examined the behaviour of bot accounts on Twitter through the application of supervised Machine Learning algorithms to 36

⁸ <https://developer.twitter.com/en/docs/twitter-api>

million tweets, posted by 1.88 million users, using a list of 86 keywords to identify the bot accounts. Whereas Kümpel *et al.* (2015), examined how news is shared on social media in their review of multiple papers addressing the matter.

Garret and Weeks (2013) explore the effects of some of the recent approaches that use Machine Learning in order to flag disreputable tweets in real time and find that the user's beliefs have a significant impact on their willingness to believe a tweet that has been marked as of questionable veracity. Similarly, Glenski *et al.* (2013) find that bots and humans react differently to flagged tweets. They find that trusted news sources had the highest proportion of human interaction on Twitter while less reputable sources had a higher proportion of bot accounts associated with retweets and reactions.

A different approach was taken by Farkas and Bastos (2018). They examined twitter accounts, previously identified as bots account and subsequently deleted by twitter, to determine the function of the bot and troll accounts. Their paper offers insights into the targeting of certain users for the spreading of disinformation such as pro-Brexit German speakers, and the use of a wide variety of hashtags dependant on the disinformation that is being spread. Contrary to their expectations, they found that much of the material focused on spreading disinformation on the back of local news and issues as opposed to the expected behaviour of focusing on larger international issues. Llewellyn conducted similar work attempting to identify bot and propaganda accounts on active Twitter accounts (Llewellyn *et al.*, 2018). While Squire (2021) examined the monetization of right-wing views and how the spreading of these views has economic impetus.

In a recent paper, a linguistic approach was taken, using the bounded confidence (BC) model Douven & Hegselmann (2021) attempted to apply a complex model that seeks to identify non-benevolent actors in a social network to help identify bots and trolls that are sharing misinformation and disinformation. Their work utilises a modified BC model that accounts for truthfulness in the simulated actors. Their work demonstrates a complex relationship between not only the bad faith actors and the targeted users, but also 'free rider' users. These users do not necessarily have an agenda but are susceptible to the opinions of those around them and unwilling to challenge or update their beliefs which, they argue, is an essential requirement for the successful spread of misinformation & disinformation.

Several papers have examined how misinformation & disinformation spread through either rumour and the unfolding and spreading of debate within social networks (Vosoughi et al., 2018; Santagiustina & Warglien, 2021; Garrett, 2011). Hahn in their 2020 paper, examined how the interconnectedness of the social network itself can cause clusters of users to be less discerning of information they are presented with. Santagiustina & Warglien (2021) examined how arguments around Brexit evolved over time and sought to develop means to classify users based on the types of argumentations that they use. They note that Twitter has an antagonistic design that emphasises the polarisation of its users into opposing viewpoints.

An interesting paper by Matei *et al.* (2017) explores the limitations of big data approaches to ground truth and utilised Eurobarometer survey data to explore how accurate some approaches have been. They identify that factors effecting the statistical outcomes can easily be hidden from the data. They point to ‘age-related inequalities’ that cause an inherent bias in the data collection.

Pherson *et al.* (2021) present an overview of methods utilised in the intelligence community for the combating of disinformation and asses the strengths and weaknesses of several different approaches. They argue that a successful strategy to defend against such foreign activities must be couple with increase media literacy.

In an earlier paper demonstrating how long Machine Learning approaches have been utilised in efforts to combat misinformation, disinformation and propaganda, Ratkiewicz *et al.* (2011), present a method that claims a 96% success rate in the identification of misinformation in tweets.

Other approaches such as combined detection methods may lead greater efficacy as argued by Volkova & Jang (2018).

2.6. Knowledge Gaps

The majority of work reviewed relates to individual domains, individual issues, or individual states. Studies rarely encompass two of those domains and no research was found that covered all domains. There is an ongoing research endeavour exploring how disinformation spreads throughout social networks, or the identification of examples in real time. Other work has focused on the targeting of individual hot topics such as Brexit, COVID-19 or various elections. Little work has been undertaken exploring the effects of such disinformation efforts on anything greater than the scale

of individual users or single issues. There is undeniable value to utilising the widest range of methods as possible in order to identify the sources of misinformation, disinformation and propaganda but without a focus on the long-term effects of such activities, it is impossible to gauge the utility of such endeavours. This work seeks to address this by exploring the effects of what has been observed on a state scale. The exploration and quantification of the effects of these activities will foster a greater understanding of their dangers.

2.7. Key Reflections

The academic literature in this area has been seen to be focused on three main areas, firstly, Machine Learning approaches to classifying and identification of misinformation, disinformation and propaganda especially in relation to Twitter data. As was discussed above, this is mainly due to the open nature the Twitter API (accurate at the time of writing). Secondly, studies in this area are centred upon gauging the impact of exposure to misinformation, disinformation, propaganda or alternative viewpoints upon users. These studies have also tended to focus Twitter and, in the main, have small numbers of subjects. Finally, the third approach in this area has been to examine the impact of major political events upon the sharing of information across social media.

It was noted that there is a dirge of literature related to exploration of macro scale exploration of the patterns identified in the micro scale. This then was to be the focus of this study.

3. PROJECT DESIGN

3.1. Introduction

In this section the design of the experiments to be undertaken in this project will be outlined. The key research question will be stated using a null hypothesis and alternative hypotheses, and an exploration of the potential approaches to seek evidence for this hypothesis will be outlined. Additionally, the types of datasets necessary for this research will be outlined as well the as variables necessary in order to test the hypothesis will be discussed.

The design of the exploratory analysis to be undertaken in the study will be discussed in detail and the implications of the analysis in relation to the study. The potential range of machine learning approaches that can be used to explore the data will also be outlined. The design of the interpretation of the results as well as any limitations in methods will also be delineated.

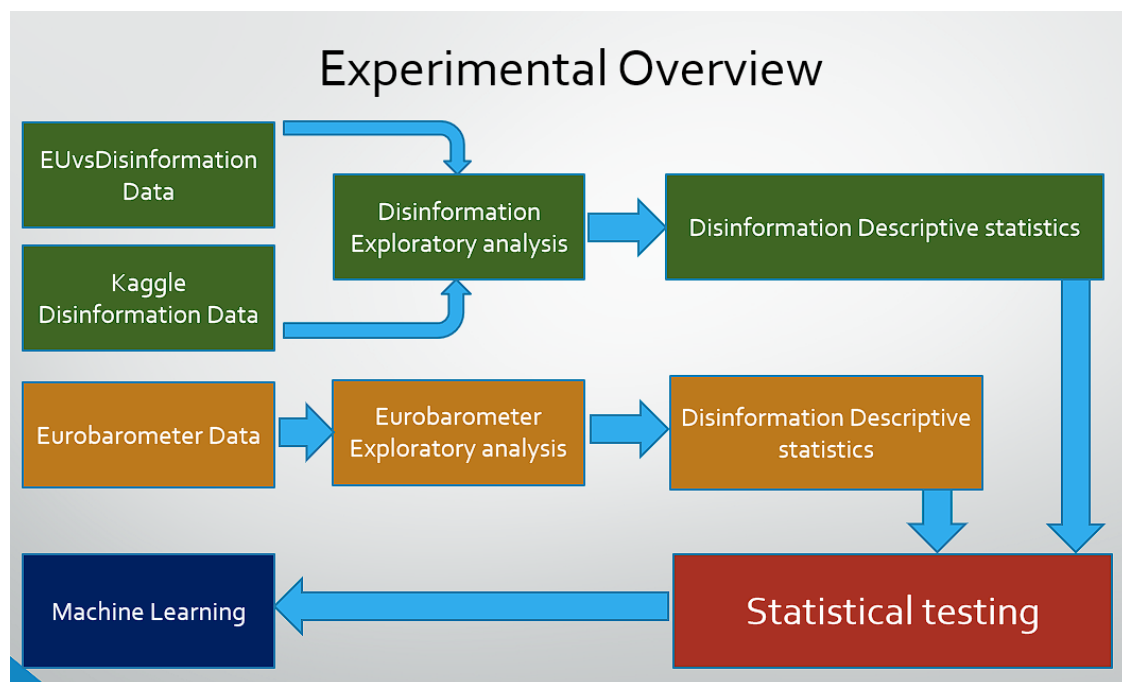


Figure 3-1- Stages of the experiment

3.2 Addressing the Research Question

The goal of this research is to explore whether EU member states that received a larger proportion of disinformation based on a range of datasets, including two EUvsDisinformation datasets, (one of which was scraped from the EUvsDisinformation website⁹ which will be called the Scraped EUvsDisinformation Dataset and the other dataset from the Kaggle website that will be called the Kaggle EUvsDisinformation Dataset) have a statistically significant change in their attitudes towards the EU and EU institutions, as measured in selected questions from an additional dataset, which is the Eurobarometer survey data that covers the years 2015 to 2022.

Propaganda efforts have been utilised in recent years around events such as the Charlie Hebdo attack (Farkas & Bastos, 2018), terrorist organisations such as Islamic State (Chatfield *et al.*, 2015), or the monetization of far-right propaganda (Squire, 2021). This work seeks to uncover evidence of the effects of these practices upon national opinion and to quantify the impact they have had on member states.

The key hypotheses are as follows:

- H_0 : There is no statistically significant correlation between nations that have been the target of larger amounts of disinformation, based on the data contained in the EUvsDisinformation datasets, and that EU member state's change in attitude towards the EU, as measured in the Eurobarometer data between the years of 2015 and 2021.
- H_A : There is a statistically significant correlation between nations that have been the target of larger amounts of disinformation, based on the data contained in the EUvsDisinformation datasets, and that EU member state's change in attitude towards the EU, as measured in the Eurobarometer data between 2015 and 2021.

3.3. Data Selection

The project began with the discovery of the EUvsDisinformation website in early 2022. After discovering and exploring the database available on the website,

⁹ https://euvsdisinfo.eu/disinformation-cases/?offset=100&per_page=100

the initial question was formulated: Can affects caused by disinformation be recognised in other data? After discovering the EUvsDisinformation database, a search of appropriate datasets available on Kaggle, and data.europa.eu, was undertaken. The majority of the available datasets were designed in relation to the information credibility in relation to hoaxes and rumours¹⁰, fake news detection¹¹ and classification of misinformation¹². Indeed, the majority of datasets in this area are focused on detection and classification rather than the macro scale affects that these events may cause.

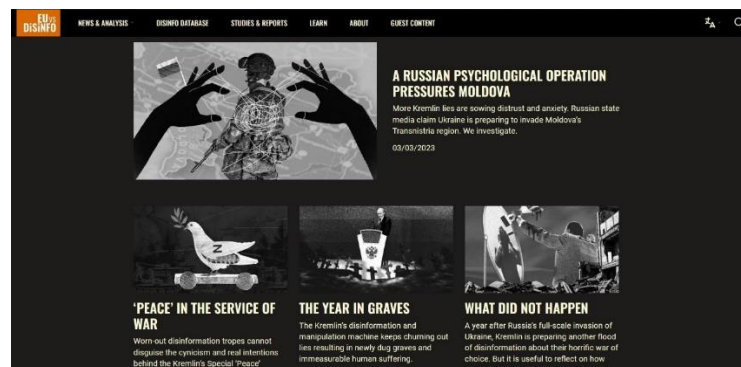


Figure 3-2 - The EUvsDisinformation main page

Extensive academic work has been conducted in these areas and a wide variety of approaches (Bastos & Mercea, 2019, O’Callaghan *et al.*, 2013) and methods are regularly being developed (Douven & Hegselmann, 2021, Hahn *et al.*, 2020), that seek to improve the accuracy of detection systems utilised by social media platforms to weed out disruptive content.

In conducting background research in this area, it was determined that there has been much less focus on the implications of the effects of disinformation upon users who have been exposed to it. Some limited studies have sought to explore the effects of disinformation and propaganda on user’s online habits (Mocanu *et al.*, 2014), or to explore whether or not a subject’s voting intentions changed after being exposed to views espoused by the opposite extreme of their political position (Iyengar & Westwood, 2015). However, there is little to no work on the effects of disinformation and propaganda on the National or International scale. This work

¹⁰ <https://data.europa.eu/data/datasets/5840066288ee38426dc65bb3?locale=en>

¹¹ https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k?select=EXTRA_RussianPropagandaSubset.csv

¹² <https://www.kaggle.com/code/sasakitetsuya/how-can-we-classify-fake-news>

seeks to explore whether there is any discernible change in national attitudes after exposure to disinformation and propaganda.

In order to answer this question, specific data would be required:

1. Firstly, data that captured the quantity of disinformation and propaganda that was targeted *per nation*.
2. Secondly, data that captured national attitudes over time
3. Thirdly, data that measured the change in national attitudes towards something that can reasonably be said to be the target of disinformation campaigns.

The majority of the datasets available on Kaggle and data.europa did not match these criteria. Most were text-based and focused on Machine Learning classification of different news articles. However, there was one dataset¹³ available that did contain data useful for the study that will be outlined in the next section.

3.4 The Datasets

The first step required to explore the research question as to whether or not there is a statistically significant correlation between nations that have been the target of larger amounts of disinformation and their attitudes to the EU, was to identify inclusion and exclusion criterion for the types of datasets that would be useful in this study. So key characteristics of this selection criterion included:

- The dataset pertains to either quantities of disinformation per nation or attitudes of an EU member state.
- The reporting must be between the years 2015 and 2022.
- The dataset must identify a country for each record.
- The data must be sufficiently rich to be useful for varied approaches to answer the research question.
- The data contained within the datasets must align with one another or be capable of being shaped into aligned data.

¹³ <https://www.kaggle.com/datasets/corrieaar/disinformation-articles>

- Data pertaining to national attitudes must have a wide range of source data to better understand the results regarding the research question.

Given these criteria a widespread search process was undertaken to find datasets that would be suitable, and several datasets were eliminated on the basis of this criteria, including a Twitter based database in French and English¹⁴ and a text based fake news dataset¹⁵. Both of which were within the scope of the research but did not satisfy the criteria as stated above in that they were aimed towards disinformation classification or detection utilising Machine Learning approaches. However, three key sources were identified as being suitable for this research:

- The current website EUvsDisinformation Data¹⁶
- The Kaggle EUvsDisinformation Data¹⁷
- The relevant Eurobarometer Data¹⁸

The diverse nature of the sources required an individualised approaches to each as outlined in the following sections.

3.4.1. EUvsDisinformation Dataset

The EUvsDisinformation website contains a wide variety of articles, studies and reports related to their ongoing efforts to detect and counter disinformation across European media. The bulk of their work has centred around eastern Europe more so since the invasion of Ukraine in February 2022. The website also provides a range of educational tools to better educate users on how to identify disinformation. Finally, they offer a disinformation database that lists the full collection of disinformation that they have collected since their creation in 2015. There are no details on the website regarding the methodology used in the collection of their data.

¹⁴ <https://data.europa.eu/data/datasets/5840066288ee38426dc65bb3?locale=en>

¹⁵ https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k?select=EXTRA_RussianPropagandaSubset.csv

¹⁶ As available at: <https://euvsdisinfo.eu/disinformation-cases/>

¹⁷ As available at: <https://www.kaggle.com/datasets/corrieaar/disinformation-articles>

¹⁸ As available at (for example):

https://data.europa.eu/data/datasets/s2532_95_3_95_eng?locale=en

Each listed instance of disinformation is hyperlinked to a summary of that instance, a proof of why the instance is a falsehood, with a short paragraph explaining where the narrative originated from, as well as other metadata. The first goal of this research in respect to this website was to scrape the data in its entirety and to use this for this project.

This necessitated using a webscraping libraries from Python using Requests¹⁹ and BeautifulSoup²⁰. While other approaches such as Selenium²¹ were available, using Requests and BeautifulSoup was the approach that allowed for a straightforward implementation given the way structure of the data on the website. A significant amount of time was spent on this endeavour since the EUvsDisinformation website security was prone to identifying the Web scraping process as a DDOS attack. This meant that the data had to be scraped over three separate sessions, with the data appended to a `.csv` file. This was necessary as Cloudflare locked out the scraping program IP address for 24 hours after about 1/3 of the data had been downloaded.

The EUvsDisinformation website data is organised in simple `<html>` tables containing only 10 rows and 4 columns (Date, Title, Outlets & Country). There were however, 14,957 rows with each table page only displaying 10 rows at a time. This meant that any attempt to scrape the data would require careful design so that it would not have to loop over the entire database at only 10 rows per table page. Through exploration, it was determined that the best approach would be to set the URL on the webpage to display 100 rows per page and scrape each of these larger pages. This significantly reduced the number of requests that were made to the website and decreased the risk of having Cloudflare block the IP address for 24 hours.

¹⁹ <https://pypi.org/project/requests/>

²⁰ <https://pypi.org/project/beautifulsoup4/>

²¹ <https://pypi.org/project/selenium/>

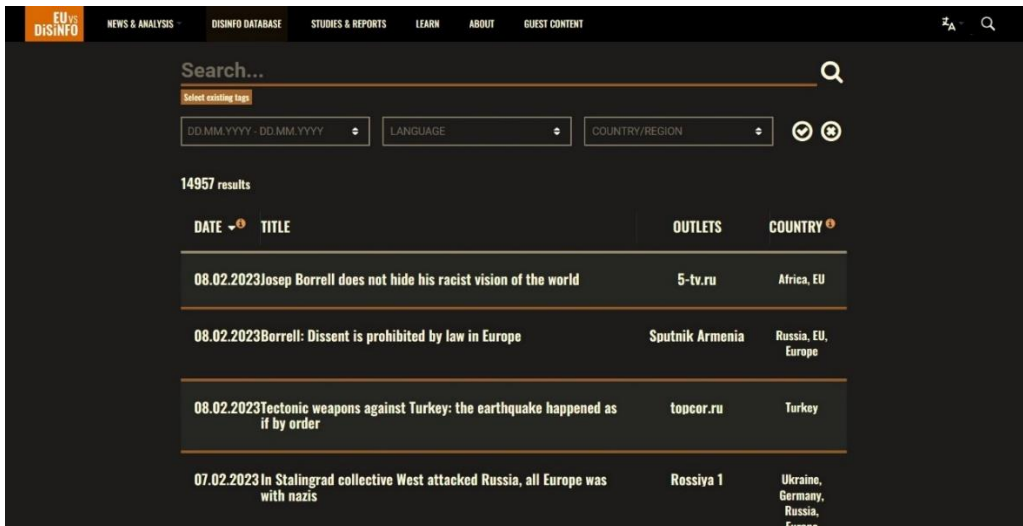


Figure 3-3: The EUvsDisinformation website database as displayed through Firefox on 10/02/23. (Note, an example pertaining to the Turkish earthquake has been added, a mere 4 days after the event)

As part of a scraping program a progress bar was developed as a further addition to the process to indicate how much of the process was completed and how much remained. Each successful webscrape could take up to 10 minutes to run and a progress bar meant that there was no danger of thinking that the process had crashed and losing what had already been downloaded. The Tqdm package²² was used for this purpose, by wrapping processes in a tqdm for loop, a simple progress bar would be displayed.

```
#Loop to iterate over table page, scrape the table and concat it to the dataframe
try:
    for i in tqdm (range(10000, 14000, 100)):
```

Figure 3-4: The tqdm for loop used for the webscraping.

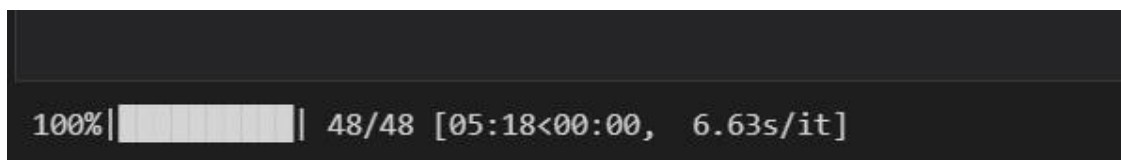


Figure 3-5: The tqdm progress bar, at the end of the process in this case.

²² <https://pypi.org/project/tqdm/>

The first step in this process required BeautifulSoup to:

1. Set the URL: In this case by setting the number of table rows per page and then later using the offset, this increased the data collected by an order of magnitude with less IP requests being sent to the website.
2. Create a Header: It was necessary to create a header to give the website security system the impression that the scraper was a browser.
3. Check the Data: It was necessary to confirm that the data was available to be downloaded.
4. Set the Page: It was necessary to set the part of the page that was targeted, a table in this case.

```
Beautiful soup steps
+ Code + Markdown

# url target
url = 'https://euvsdisinfo.eu/disinformation-cases/?offset=0&per_page=100'

# header to trick webpage into thinking I am a browser to prevent code 403: access denied
header = {
  'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; rv:91.0) Gecko/20100101 Firefox/91.0'
}

# using request to use the url and header together to access the webpage
r = requests.get(url, headers=header)

# Scraping the table with panda's read_html to check if everything is working
dfs = pd.read_html(r.text)
# Check the status code, 200 = everything working
print(r.status_code)
```

Figure 3.5: Steps in Using BeautifulSoup.

Early attempts ran into significant challenges, for example, while scraping a single page, or several pages at a time, the system worked as expected. The scale of the scraping required meant that the scraper was being IP blocked by security as a potential DDOS attacker. Two solutions were attempted, neither of which fully solved the issue. Firstly, a list of proxy IP addresses were used to pass into the function that would rotate after each preceding IP address was blocked the Cloudflare security. There are significant lists of public IP addresses available but several attempts using this method failed due to the unreliability of the publicly available addresses.

A better solution was to add a time delay with a random interval to each iterative page request to space out the requests and attempt to mask the scale of the scraping.

This was more successful and allowed the scraping of approximately one third of the data at a time before being IP blocked for 24 hours.

```
# Counts and elements needed
# the number of records to offset by as we iterate, this functions as the table number when added to the base_url
offset = 100
# delays to be used with sleep()- increase to range 4-10 to see if that prevents the security block
delay = randint(4,10)

# -----

# Block to set off the process
# the base_url that we are starting from:
#base_url = 'https://euvsdisinfo.eu/disinformation-cases/?offset=0&per_page=100'
#base_url2 = 'https://euvsdisinfo.eu/disinformation-cases/?offset=4900&per_page=100'
base_url3 = 'https://euvsdisinfo.eu/disinformation-cases/?offset=4900&per_page=100'
r = requests.get(base_url3, headers=header)
dfs = pd.read_html(r.text) #read the table with pandas
df = dfs[0] #set the dataframe as the first table found (in this case there is only 1)
# set the col heads in the dataframe as those on the webpage
col_headings = df[['Disinformation Detection Date.', 'Date', 'Title', 'Outlets', 'Countries and/or Regions discussed in the disinformation.', 'Country' ]]

# -----
```

Figure 3.6- Initial steps for web scraping

The `base_url` was the start point of each day's scraping. The offset was used to pick up from where the last scrape had been blocked. The data was then appended to a spreadsheet as it was scraped.

```
#Loop to iterate over table page, scrape the table and concat it to the dataframe
try:
    for i in tqdm (range(10000, 14800, 100)):
        offset_url = 'https://euvsdisinfo.eu/disinformation-cases/?offset=' + str(i) + '&per_page=100'
        sleep(delay) #sets the delay to avoid being blocked by webpage
        r = requests.get(offset_url, headers=header)
        dfs = pd.read_html(r.text)
        df_loop = dfs[0]
        df = pd.concat([df, df_loop]) #concat the table page to ongoing dataframe
        col_headings = df[['Disinformation Detection Date.', 'Date', 'Title', 'Outlets', 'Countries and/or Regions discussed
except ValueError:
    print('Blocked again! Last page completed: ', i)

print('Finished, last page complete: ', i)
```

100%|██████████| 48/48 [05:18<00:00, 6.63s/it]

Finished, last page complete: 14700

Python

Figure 3.7- Final block for scraping EUvsDisinformation data.

The final development of the code functioned well enough bar the limits of being blocked by Cloudflare each time after collecting around a third of the data. The image above demonstrates the final time the block was used, the range is set to capture the last table pages. An exception was included in the loop to capture the last page that was scraped before being blocked so that it could be picked up at that point on the following attempt. The final process was finished over 96 hours with one third of the data being scraped each time followed by a 24-hour block from the website. The data was then merged and later accessed in Rstudio.

	Disinformation Detection Date. Date	Title	Outlets	Countries and/or Regions discussed in the disinformation. Country
95	16.11.2015	Without military, financial, informational, po...	evrazia.org	US
96	16.11.2015	Assad is not the enemy of France, ISIL is, say...	RIA	France, Syria
97	16.11.2015	Assad is not the enemy of France, ISIL is, say...	sputniknews.com	France, Syria
98	15.11.2015	Media have forgotten about the 286 women who w...	Webtribune.rs	Ukraine
99	15.11.2015	A dozer was driving over soldiers' graves in U...	rusvesna.su	Ukraine

Figure 3.8- The scraped data as viewed in Python.

The scraped data was structured as follows:

Name	Description
Date	The date of the claim
Title	A text description of the claim. Usually about a sentence long, no more than 20 words
Outlets	The source of the claim
Country	A comma separated list of countries that were targeted in the disinformation instance.

Since the data had been scraped directly from the website, this data was compact and complete. There were no missing values across the dataset, and it required a minimum of wrangling. More detailed data is available by utilizing the website metadata, but it was decided that the main focus of this dataset would be the count of countries targeted by disinformation.

3.4.2. Kaggle EUvsDisinformation data

The second set of data was the EUvsDisinformation data as available on Kaggle²³. The original poster of this dataset responded after being contacted for further information and explained that the original data had been made available at a

²³ <https://www.kaggle.com/datasets/corrieaar/disinformation-articles>

hackathon several years ago. The data is extremely rich and has many possible applications given the wide variety of dimensions that each instance contains.

It contains significantly more detail than is currently available on the public facing website. The richer data contained in this dataset is also somewhat accessible through website meta data, it was decided however to use only the basic data that was successfully scraped, and the richer data available on Kaggle. This more complex dataset (`Kaggle_Disinfo`), used in conjunction with the more simplistic but wider ranged data (`Scrape_Disinfo`), provided both the breadth and depth necessary for this analysis.

The data was downloaded and simply loaded in R:

```
Kaggle_import |
Kaggle_Eu = read.csv2("C:/Users/alexm/OneDrive/Desktop/TU059/Semester
3/Data/EuVSDisinfo/Kaggle_EUvsDisinfo/data.csv", header = TRUE, sep = ",", na.strings =
c("", "['None']"), fill = FALSE)
# Header included
# na specified as blank cell and "['None']" since that is used in the original dataset
```

Figure 3.9- Loading the Kaggle data.

The data contained a significant number of blank cells. Within the 7.3K rows there was a staggering 59K missing values, as well as cells that were marked as “['None']”. These were converted to ‘NA’ during the import process. The data contained over 7K observations with 37 columns. This would have to be significantly pruned in order to be of use for the study purposes:

```
[1] "X" "claims_id" "claim_published" "first_appearance" "review_id"
[6] "is_part_of" "claim_reviewed" "review_published" "review_name" "html_text"
[11] "text" "issue_id" "keyword_id" "keyword_name" "country_id"
[16] "country_name" "appearances" "has_parts" "creative_work_id" "type"
[21] "url" "author" "claim" "web_archive_url" "abstract"
[26] "in_language" "start_time" "end_time" "organization_id" "location"
[31] "organization_name" "image_id" "image_type" "image_content_url" "language_id"
[36] "language_name" "language_code"
```

Figure 3.10- Column names of raw Kaggle data

Many of the columns merely contained a repetitive code word. So, for example, `language_id` and `language_name` were in essence, the same data, while

language_code was merely a factor representation of the same data. Other columns such as start_time and end_time were empty in over 90% of the dataset. Finally, some rows were dependant on external sources such as images or out of date URLs.

There were 37 columns in total, most of which proved to be of little use:

Name	Description
X	An index coerced when imported
Claims_id	An id number: '/claims/100'
Claim_published	Date of the disinformation instance
First appearance	Code number for initial source of the claim
Review_id	code number of the review
Is_part_of	Unclear data, a code number for issue but no explanation given. Example: '/issues/177'
Claim_reviewed	Text Paragraph reviewing the claim
Review_published	Date of when the review was published
Review_name	Sentence summary of the response to disinformation claim Example: 'the protests in Hong Kong are US-funded'
First appearance	Code number for initial source of the claim
Review_id	code number of the review
Is_part_of	Unclear data, a code number for issue but no explanation given. Example: '/issues/177'
Html_text	An extract html element contained a response to the claim in text.
Text	The same html_text with html tags removed
Issue_id	Repeat of 'is part of' column
Keyword_id	List of keyword codes. Example: ['/keywords/61', '/keywords/76']
Keyword_name	List of specific keywords. Example: ['Conspiracy', 'Terrorise', 'Donald Trump']

Country_id	List of country codes. Example: ['/countries/4', '/countries/10']
Country_name	List of country names: ['Ukraine', 'Russian', 'UK']
Appearances	Unclear data: '/news_articles/598', '/media_objects/1847'
Has_parts	Unclear data. Seems to be a list of elements of claims: ['/claims/75', '/claims/73', '/claims/33']
Creative_work_id	Unclear data. Repeats appearances column
type	Unclear data. A non-functional html address: 'http://schema.org/NewsArticle'
url	The non-functional url where the claim was originally found.
Author	A code for different author organisations but without a key
Claim	A code for different claims but without a key
Web_archive_url	A non-functional https address for the original claim
Abstract	Text harvested from the https source and stored as a string. Mix of languages, mostly Russian
In_language	List of languages stored as codes: '/language/3'. No key provided but later column lists the languages as a string
Start_time	Mostly blank cells, unclear data. Possibly the start time in seconds for the disinformation contained in the source.
End_time	Mostly blank cells, unclear data, seems to be the end time in seconds related to the column before.
Organization_id	Code for source organisation but without a key. Later columns contain the country as a string. '/organizations/262'
Location	Unclear data. Possibly the location of the source of the claim
Organization name	The source of the claim. Example: 'southfront.org'. Mostly blank cells

Image_id	Code for each image: '/image_object/23'
Image_type	Unclear data, mostly blank cells but clearly related to previous column. Addresses are non-functional: 'http://schema.org/ImageObject'
Language_id	Code for language contained in claim: '/languages/3'. NO key provided but next column lists language as string
Language_name	List of language or languages used in claim.
Language_code	3 letter code for language used in claim but lists only one language even if more are used.

In the end, of the 37 rows the dataset was cut down to 6 columns:

Name	Description
Claim published	Equivalent to the date of the event.
Keyword name	The key issues that the event targeted.
Country name	The country or countries that were targeted or mentioned.
Abstract	A description of the issue addressed in the event. This was later removed as it contained strings mostly in other languages.
Organisation name	Equivalent to organisation in the scraped data.
Language name	The language that the event was originally written in.

Subsequently a large number of 'N/A' values were discovered throughout the remaining data. Most rows had some missing element in one or more column. Arguably all rows with an N/A could be removed, but as further exploration and experiment planning developed, it was decided that only the bare minimum of rows were of value to the work. Removing all rows containing an N/A value left less than 200 rows. The abstract column alone contained 6.5k N/A values, therefore it was decided to leave the dataframe intact and to use only the values that were needed for the eventual experiment.

claim_published	keyword_name	country_name	abstract	organization_name	language_name
0	14	26	6544	0	4150

Figure 3.11- Kaggle dataframe missing values. Total number of rows was over 7000.

Finally, some small changes were made to the data on the dataframes, dates were correctly stored as date values, factors were applied where necessary. The large text data in `Kaggle$abstract` and `Scrape$Title`, were left as part of the data so that they could be utilised for extra plots such as word clouds or data mining. The final dataset was 6 columns wide (as seen in the above image) and contained 7.3K rows, listed by date with both `keyword_name` and `country_name` column containing comma separated lists. This would be modified later in preparation for the eventual experiment.

3.4.3. Eurobarometer Data

Having prepared the Kaggle and scraped data, the next task was to prepare the Eurobarometer survey data. This task proved reasonably complicated, the scale of the data that was initially collected proved unwieldy, each question that was added created a significant amount of work in terms of wrangling and preparing the data. Issues arose such as how to clean data effectively over thousands of pages contained in long lists that were not easily accessible in R. Other issues arose around the use of French text in the data that was incorrectly encoded in RStudio. Each of these challenges had to be handled in turn.

Standard Eurobarometer surveys are conducted twice a year. Each of the 28 Pre-Brexit EU member states, 27 member states following Brexit, have circa 1000 respondents polled on upwards of 170 questions that vary in each biannual survey. The structure of the data and the way it is packaged for public sharing led to a host of problems. Other than Standard Eurobarometer surveys, there are a number of Flash Eurobarometers that also occur each year that tend to explore current events. These flash surveys were not used as the data contained within them was not replicated over the study years.

B2 ¹ A1	B Country	
Q1NAT ¹ A1	Q1NAT What is your nationality? Please tell me the country(ies) that applies(y). (MULTIPLE ANSWERS POSSIBLE)	
QA1a.1 ¹ A1	QA1a.1 How would you judge the current situation in each of the following?	The situation in (OUR COUNTRY) in general
QA1a.2 ¹ A1	QA1a.2 How would you judge the current situation in each of the following?	The situation of the (NATIONALITY) economy
QA1a.3 ¹ A1	QA1a.3 How would you judge the current situation in each of the following?	The situation of the European economy
QA1a.4 ¹ A1	QA1a.4 How would you judge the current situation in each of the following?	Your personal job situation
QA1a.5 ¹ A1	QA1a.5 How would you judge the current situation in each of the following?	The financial situation of your household
QA1a.6 ¹ A1	QA1a.6 How would you judge the current situation in each of the following?	The employment situation in (OUR COUNTRY)
QA1a.7 ¹ A1	QA1a.7 How would you judge the current situation in each of the following?	The provision of public services in (OUR COUNTRY)
QA2a.1 ¹ A1	QA2a.1 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	Your life in general
QA2a.2 ¹ A1	QA2a.2 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	The situation in (OUR COUNTRY) in general
QA2a.3 ¹ A1	QA2a.3 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	The economic situation in (OUR COUNTRY)
QA2a.4 ¹ A1	QA2a.4 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	The financial situation of your household
QA2a.5 ¹ A1	QA2a.5 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	The employment situation in (OUR COUNTRY)
QA2a.6 ¹ A1	QA2a.6 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	Your personal job situation
QA2a.7 ¹ A1	QA2a.7 What are your expectations for the next twelve months: will the next twelve months be better, worse or the same	The economic situation in the EU
QA3a ¹ A1	QA3a What do you think are the two most important issues facing (OUR COUNTRY) at the moment? (MAX. 2 ANSWERS)	(IF 'SPLIT A')

Figure 3.12- Example Eurobarometer data as downloaded by the author. Questions useful to the work were highlighted in green.

Index		Eurobarometer 92.3																																
		VOLUME A Pondéré Weighted													Terrain/Fieldwork : 14 - 29/11/2019																			
QA1a.3 Comment jugez-vous la situation actuelle de chacun des domaines suivants ?		QA1a.3 How would you judge the current situation in each of the following?																																
La situation de l'économie européenne		The situation of the European economy																																
		UE28	UE28-UK	BE	BG	CZ	DK	D-W	DE	D-E	EE	IE	EL	ES	FR	HR	IT	CY	LV	LT	LU	HU	MT	NL	AT	PL	PT	RO	SI	SK	FI	SE	UK	
TOTAL		27382	26372	1012	1039	1013	1022	1035	1540	505	1001	1013	1008	1008	1014	1013	1023	505	1000	1008	510	1011	501	1006	1018	1008	1003	1058	1007	1007	1001	1023	1010	
Très bonne		1034	1060	20	141	73	59	24	37	13	30	70	73	19	4	76	25	23	29	67	17	114	20	56	121	77	33	125	54	30	14	23	21	
Very good		4%	4%	2%	13%	7%	6%	2%	2%	3%	7%	7%	7%	2%	1%	8%	2%	5%	3%	7%	3%	11%	4%	5%	12%	8%	3%	12%	5%	3%	1%	2%	2%	
Plutôt bonne		11826	11883	499	474	572	583	551	812	256	626	587	420	343	294	560	289	253	638	679	284	591	292	643	577	579	613	546	627	521	535	532	305	
Rather good		43%	45%	49%	46%	57%	57%	53%	53%	51%	63%	58%	42%	34%	29%	55%	23%	50%	64%	67%	50%	59%	58%	64%	57%	57%	61%	52%	62%	52%	54%	52%	30%	
Plutôt mauvaise		8854	8603	386	111	235	198	338	506	171	134	202	342	405	417	256	486	425	173	114	158	205	70	230	242	180	219	223	257	314	323	251	306	
Rather bad		32%	33%	38%	11%	23%	19%	33%	33%	34%	13%	20%	34%	40%	41%	25%	48%	25%	17%	11%	31%	20%	14%	23%	24%	18%	22%	21%	26%	31%	32%	25%	30%	
Très mauvaise		1907	1746	59	47	34	23	32	47	14	12	35	107	88	113	55	141	36	12	17	14	29	5	19	37	40	4	55	34	39	12	18	95	
Very bad		7%	6%	6%	5%	3%	2%	3%	3%	3%	1%	3%	11%	9%	11%	5%	14%	7%	1%	2%	3%	3%	1%	2%	3%	4%	1%	5%	3%	4%	1%	2%	10%	
NSP		3761	3080	48	266	99	158	90	138	51	198	118	66	153	187	66	83	64	148	130	38	71	113	58	41	132	133	109	35	103	117	199	283	
DK		14%	12%	5%	25%	10%	16%	9%	9%	10%	20%	12%	6%	15%	19%	7%	8%	13%	15%	13%	7%	7%	23%	6%	4%	13%	13%	10%	4%	10%	12%	19%	28%	
Total 'Bonne'		12859	12943	518	615	645	642	575	849	269	657	657	493	362	297	636	314	277	667	746	301	706	312	698	698	657	647	671	681	550	549	555	326	
Total 'Good'		47%	49%	51%	59%	64%	63%	55%	55%	53%	66%	65%	49%	36%	29%	63%	30%	53%	67%	74%	59%	70%	62%	69%	69%	65%	64%	67%	55%	55%	54%	32%		
Total 'Mauvaise'		10761	10349	445	159	269	222	370	553	185	146	237	449	493	530	311	626	165	185	131	171	234	75	250	279	220	223	278	291	353	335	269	401	
Total 'Bad'		39%	39%	44%	16%	26%	21%	36%	36%	37%	14%	23%	45%	49%	49%	52%	30%	62%	32%	18%	13%	34%	23%	15%	25%	27%	22%	23%	26%	29%	35%	33%	27%	40%

Figure 3.13- Example EB question

Each individual question recorded the number of each Member State's citizen's response to the question. Responses were recorded depending on the question. In the pictured example, a scale of *Bad* to *Very Good* was used. Other questions had different options. The data also recorded the EU 28 responses overall, and in the example above, the EU28-UK (EU 28 *minus* UK). The structure of the table itself as well as differences between tables, across the range of years, led to a host of problems in RStudio. The scale of the data, with up to 30 member or prospective member states, answering up to 170 questions meant that each years' survey was a large file. 7 years' worth of such files were needed and had to be loaded into RStudio. The first attempts however, either totally failed due to the size and structure of the data or crashed RStudio. A new approach was developed that avoided these issues.

The first step was to decide on which questions that were available over the 7-year period would be directly related to the hypothesis in this experiment. This

necessitated tabulating the questions that were used in each of the 11 surveys, determining the questions that were consistent, or at least partially consistent over the survey period and exploring any patterns in the data overall.

A	B	C	D	E	F	
EB 83- Spring 2015	EB 84- Autumn 2015	EB 85- spring 2016	EB 86- Autumn 2016	EB 87- Spring 2017	EB 88- Autumn 2017	FROM EB92
D71b.2'IA1	D71b.2'IA1	D71b.2'IA1	D71b.2'IA1	D71b.2'IA1	D71b.2'IA1	
D71b.3'IA1	D71b.3'IA1	D71b.3'IA1	D71b.3'IA1	D71b.3'IA1	D71b.3'IA1	
					C1'IA1	
C2'IA1	C2'IA1	C2'IA1	C2'IA1	C2'IA1	C2'IA1	C2 - Political interest index
QA1a.1'IA1	QA1a.1'IA1	QA1a.1'IA1	QA1a.1'IA1	QA1a.1'IA1	QA1a.1'IA1	QA1a.1 How would you judge the current situation in each of the following?
QA1a.2'IA1	QA1a.2'IA1	QA1a.2'IA1	QA1a.2'IA1	QA1a.2'IA1	QA1a.2'IA1	QA1a.2 How would you judge the current situation in each of the following?
QA1a.3'IA1	QA1a.3'IA1	QA1a.3'IA1	QA1a.3'IA1	QA1a.3'IA1	QA1a.3'IA1	QA1a.3 How would you judge the current situation in each of the following?
QA1a.4'IA1	QA1a.4'IA1	QA1a.4'IA1	QA1a.4'IA1	QA1a.4'IA1	QA1a.4'IA1	QA1a.4 How would you judge the current situation in each of the following?
QA1a.5'IA1	QA1a.5'IA1	QA1a.5'IA1	QA1a.5'IA1	QA1a.5'IA1	QA1a.5'IA1	QA1a.5 How would you judge the current situation in each of the following?
QA1a.6'IA1	QA1a.6'IA1	QA1a.6'IA1	QA1a.6'IA1	QA1a.6'IA1	QA1a.6'IA1	QA1a.6 How would you judge the current situation in each of the following?
QA1a.7'IA1	QA1a.7'IA1	QA1a.7'IA1	QA1a.7'IA1	QA1a.7'IA1	QA1a.7'IA1	QA1a.7 How would you judge the current situation in each of the following?
QA1b.1'IA1	QA1b.1'IA1	QA1b.1'IA1	QA1b.1'IA1	QA1b.1'IA1	QA1b.1'IA1	
QA1b.2'IA1	QA1b.2'IA1	QA1b.2'IA1	QA1b.2'IA1	QA1b.2'IA1	QA1b.2'IA1	
QA1b.3'IA1	QA1b.3'IA1	QA1b.3'IA1	QA1b.3'IA1	QA1b.3'IA1	QA1b.3'IA1	
QA1b.4'IA1	QA1b.4'IA1	QA1b.4'IA1	QA1b.4'IA1	QA1b.4'IA1	QA1b.4'IA1	
QA1b.5'IA1	QA1b.5'IA1	QA1b.5'IA1	QA1b.5'IA1	QA1b.5'IA1	QA1b.5'IA1	
QA1b.6'IA1	QA1b.6'IA1	QA1b.6'IA1	QA1b.6'IA1	QA1b.6'IA1	QA1b.6'IA1	

Figure 3.14- Example of the tabulation of a selection of questions from 5 of the EB surveys

Eurobarometer 92-2019 (EB92-2019) was chosen as the initial reference list of questions. It was chosen as it is the midpoint of the range, and this meant that there were a number of questions that were consistent in the years before and the years following EB 92-2019.

“QA12.1: And please tell me if you tend to trust or tend not to trust these European institutions: The European Parliament”.

44 questions were chosen from EB92-2019 as being of potential interest or likely to be of use for the work. Example questions such as QA12.1, were clearly of use in answering whether a Member State had had a change in their trust in the EU over the study period. Other questions were less clearly of use but may have included interesting insights into the data:

“QA7: What does the EU mean to you personally? (Multiple answers Possible)”

After compiling a list of questions, the task of checking each Eurobarometer survey and tabulating whether the question was included in that year’s survey began. It

was discovered immediately that not only were question codes *not* consistent across the years but that also codes could be reused for different questions in different years. Further, EB93-2020 had its own code system with no relation to the following or previous years. Following EB93-2020 an entirely new system was implemented, with a simplification of the codes. The range of possible codes for each question meant that very specific lists would have to be used when importing the required data into R and many hours of careful analysis and determining translation processes between the codes would be necessary.

QA8a.14 IA1	QA8a.10 IA1	QA8a.14 IA1	QA6a.10 IA1	QA6a.14	T52	QA6b_10	QA6a_11	QA6b_10	QA6a.14 I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell	The EU
-------------	-------------	-------------	-------------	---------	-----	---------	---------	---------	---	--------

Figure 3.15- Change in one question over time. The code changes each year, and the question is missing in 2022.

The final challenge encountered while tabulating the questions was in the variation of questions from year to year. Many of the original 44 applicable questions from EB93-19 were not present in other EB years or were only occasionally present. Very few questions overall were entirely consistent across the 10-year range.

With the list of questions in hand, they were loaded into RStudio. This presented its own challenge. The structure of each Eurobarometer contained an index page that listed the selected questions from that year with hyperlinked connections to switch to specific question sheets. The questions themselves were formatted in such a way that RStudio had difficulty in formatting the data in a readable manner, finally, the scale of each year’s survey results, with up to 170 questions meant that RStudio took several minutes to load in the data per year.

The first attempts to load the data into RStudio used the `Rio` and `openxlsx` packages. Both were capable of loading the data but took significant amounts of time in the first tests without having set any parameters. Initial attempts to load the data without arguments crashed RStudio. After analysis, the `openxlsx` package was settled on as it performed better.

The eventual method after much trial and error was to pass the string values of the pages from the specific EB spreadsheet that referred to the questions of use to the study, in total the data was loaded from EB83-2015 to EB97-2022.

```
## [r EB97 Import]

sheets_97 <- c("D71_1", "D71_2", "D71_3", "C2", "QA1_3", "QA2_2", "QA2_3", "QA2_7",
"QA5", "D73_1", "D73_2", "D73_3", "QA7",
"QA11_2", "SD18a", "SD18b", "D72_1", "D72_2", "QB1_1", "QB3_1", "QB3_4", "QB3_6", "QB3_7",
"QD1a_2",
"QD1a_3", "QD1a_4", "QD2_1", "QD3a", "QD3b", "QD3T", "QD5_1", "QD5_2", "QD7", "QD8", "D78"
)

openxlsx_97 <- lapply(sheets_97, read.xlsx, xlsxFile = "C:/Users/alexm/OneDrive/Desktop/TU059/Semester
3/Data/EB/EB97- Summer 2022/eb97_Volume_A.xlsx")
names(openxlsx_97) <- sheets_97 # Sets the names as the above list of pages

# EB data loaded one year at a time
# Loaded in reverse order (for no particular reason)
# Sheets are selected with strings and others are skipped.
# Questions have slightly different codes each year (because of course they do)
# This data will then be cleaned and appended to question DF.
...
```

Figure 3.18- Code block to import EB97-2022 using only the specific questions needed for the study.

The data was loaded as a ‘Large List’ by `openxlsx`. This resulted in data that could be accessed through the list index. The data was not yet in dataframe format and was still listed as separate years:

Data		
▶ openxlsx_83	Large list (33 elements,	1.4 M... 🔍
▶ openxlsx_84	Large list (39 elements,	1.6 M... 🔍
▶ openxlsx_85	Large list (34 elements,	1.4 M... 🔍
▶ openxlsx_86	Large list (41 elements,	1.7 M... 🔍
▶ openxlsx_87	Large list (39 elements,	1.7 M... 🔍
▶ openxlsx_88	Large list (39 elements,	1.6 M... 🔍
▶ openxlsx_89	Large list (42 elements,	1.8 M... 🔍
▶ openxlsx_90	Large list (42 elements,	1.8 M... 🔍
▶ openxlsx_91	Large list (41 elements,	1.7 M... 🔍
▶ openxlsx_92	Large list (44 elements,	1.8 M... 🔍
▶ openxlsx_93	Large list (39 elements,	1.7 M... 🔍
▶ openxlsx_94	Large list (31 elements,	1.5 M... 🔍
▶ openxlsx_95	Large list (35 elements,	1.7 M... 🔍
▶ openxlsx_96	Large list (35 elements,	1.6 M... 🔍
▶ openxlsx_97	Large list (35 elements,	1.7 M... 🔍

Figure 3.19- Eurobarometer questions loaded year by year as Large Lists

Each year's data was loaded as a list of questions from that year. EB92-2019, as the original reference year, contained the most questions (44), other years contained varying numbers of questions that aligned with EB92-2019 with EB94-2020 having the lowest number of questions with 31. Each large list was constituted as a list of the individual questions that had been loaded into Rstudio. Each question could be accessed by the Large List index or string value such that, question D71_1 for openxlsx_96 (EB96-2021), could be accessed with the code: `View(openxlsx_96[["D71_1"]])`. However, these list elements had yet to be cleaned in order to make the data accessible.

Name	Type	Value
openxlsx_97	list [35]	List of length 35
D71_1	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
D71_2	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
D71_3	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
C2	list [14 x 44] (S3: data.frame)	A data.frame with 14 rows and 44 columns
QA1_3	list [21 x 44] (S3: data.frame)	A data.frame with 21 rows and 44 columns
QA2_2	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
QA2_3	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
QA2_7	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
QA5	list [40 x 44] (S3: data.frame)	A data.frame with 40 rows and 44 columns
D73_1	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
D73_2	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
D73_3	list [15 x 44] (S3: data.frame)	A data.frame with 15 rows and 44 columns
QA7	list [40 x 44] (S3: data.frame)	A data.frame with 40 rows and 44 columns
QA11_2	list [21 x 32] (S3: data.frame)	A data.frame with 21 rows and 32 columns
SD18a	list [20 x 44] (S3: data.frame)	A data.frame with 20 rows and 44 columns
SD18b	list [20 x 43] (S3: data.frame)	A data.frame with 20 rows and 43 columns
D72_1	list [21 x 43] (S3: data.frame)	A data.frame with 21 rows and 43 columns
D72_2	list [21 x 44] (S3: data.frame)	A data.frame with 21 rows and 44 columns
QB1_1	list [21 x 32] (S3: data.frame)	A data.frame with 21 rows and 32 columns
QB3_1	list [15 x 32] (S3: data.frame)	A data.frame with 15 rows and 32 columns
QB3_4	list [15 x 32] (S3: data.frame)	A data.frame with 15 rows and 32 columns
QB3_6	list [15 x 32] (S3: data.frame)	A data.frame with 15 rows and 32 columns
QB3_7	list [15 x 32] (S3: data.frame)	A data.frame with 15 rows and 32 columns
QD1a_2	list [21 x 44] (S3: data.frame)	A data.frame with 21 rows and 44 columns
QD1a_3	list [21 x 44] (S3: data.frame)	A data.frame with 21 rows and 44 columns
QD1a_4	list [21 x 44] (S3: data.frame)	A data.frame with 21 rows and 44 columns
openxlsx_97[["C2"]]		

Figure 3.20- EB97-2022. Questions are contained in a Large List and can be accessed with their index or string value.

The questions in the list were loaded by openxlsx exactly as they were laid out by the original xlsx file. This meant that the empty cells that were part of the spreadsheet, were loaded as NA values in R. Attempting to remove the NA values by hand for each question for each year was not a viable solution so an algorithmic solution was developed.

X1	X2	X3	X4	X5	X6	X7	X8	Eurobarometer-.97.5	X10	
1	NA	VOL A weighted	NA	NA	NA	NA	TerrainFieldwork: 17/06 - 17/07/2022	NA	NA	
2	NA	D71.1. Quand vous retrouvez avec des amis ou des pro...	NA	NA	NA	NA	D71.1. When you get together with friends or relatives, woul...	NA	NA	
3	NA	De sujets de politique nationale	NA	NA	NA	NA	National political matters	NA	NA	
4	NA	Base: Ensemble	NA	NA	NA	NA	Base: All Respondents	NA	NA	
5	NA	NA	NA	NA	NA	NA	NA	NA	NA	
6	< Back to content	NA	UE27 EU27	BE	BG	CZ	DK	D-W	D-E	
7	Total	26468	1009	1038	1015	1037	1211	1507	296	
8	NA	Fréquentement	6800	205	333	301	324	426	534	108
9	NA	Frequently	0.26	0.2	0.32	0.3	0.31	0.35	0.35	0.37
10	NA	Occasionnellement	14447	619	559	594	590	690	854	164
11	NA	Occasionally	0.55000000000000004	0.62	0.54	0.57999999999999996	0.56999999999999995	0.56999999999999995	0.56999999999999995	0.5500
12	NA	Jamais	5171	183	140	117	124	91	115	23
13	NA	Never	0.19	0.18	0.13	0.12	0.12	0.08	0.08	0.08
14	NA	Ne sait pas	51	3	7	2	0	4	5	1
15	NA	Don't know	-	-	0.01	-	-	-	-	-

Figure 3.21- EB97-2020- Question D71.1. 44 columns long, all character values without useful headings and containing many NA values.

On exploring the questions that had been loaded, it was apparent that there was a lot of variation in the spreadsheet structure. D71.1 (above) had 44 columns and 15 rows. This covered the available replies to that specific survey question and recorded the raw numbers of responses per Member state as well as the equivalent percentage. Other pages had up to 30 rows depending on the available response to the specific question. However, the first 6 to 9 rows, containing meta data, usually shared the same layout across years.

A function was written to do the following:

1. Search for the string value (green box): 'UE27 EU27'
2. Remove all rows above the location where it was found. This removed all the Meta data contained on the question page and the first column containing NA values (red box).
3. Set the country cells as the new column headings (Yellow box) and then remove those string values.
4. Convert all values to numeric, bar the factor values (blue box):
5. Reset the index for the new Dataframe.

	X1	X2	X3	X4	X5	X6	X7	X8
1	NA	NA	VOL A weighted	NA	NA	NA	NA	Terrain/Fieldwor
2	NA	D71.1. Quand vous retrouvez avec des amis ou des pro...	NA	NA	NA	NA	NA	D71.1. When yo
3	NA	De sujets de politique nationale	NA	NA	NA	NA	NA	National politica
4		Base: Ensemble	NA	NA	NA	NA	NA	Base: All Respon
5		NA	NA	NA	NA	NA	NA	NA
6	<<Back to content	NA	UE27 EU27	BE	BG	CZ	DK	D-W
7		Total	26696	1103	1036	1020	996	1289
8	NA	Fréquemment	6925	197	302	414	316	492
9	NA	Frequently	0.26	0.18	0.28999999999999998	0.41	0.32	0.38
10	NA	Occasionnellement	14254	666	597	565	565	712
11	NA	Occasionally	0.53	0.6	0.57999999999999996	0.55000000000000004	0.56999999999999995	0.55000000000000004
12	NA	Jamais	5481	240	128	41	115	83
13	NA	Never	0.21	0.22	0.12	0.04	0.11	7.0000000000000004
14	NA	Ne sait pas	36	0	9	0	0	1
15	NA	Don't know	-	-	0.01	-	-	-

Figure 3.22- Reseting the DataFrame

A major stumbling block was encountered in this approach due to the inner workings of R. The function, as initially designed searched 3 possible locations for the necessary string, in EB97-2020 for example, 'UE27\nEU27' could be located in either the 5th, 6th or 7th row.

The first design of the function attempted to iteratively search from lower numbered rows to higher numbered rows. But this proved challenging as there is an issue in R when searching for string values, if an 'N/A' is found in the location, R will throw an error and end the process. This proved to be challenging, and was solved by changing the search approach, rather than searching from the lower numbered rows to higher numbered rows, the search criteria searched higher numbers and then descended

through the row numbers. This deftly avoided the issue in R rather than solving it, the function now worked as intended.

A second issue was caused by a change in the structure of the spreadsheet in 2020 specifically. The earlier function would not work with that year’s data. This meant that a second function had to be written for those sheets to load them successfully into R. Further issues with this dataset were uncovered later. In total, three functions each with three different `if/else` statements, totally nine different variations to the function, were needed to search different row locations, or different strings depending on the style used that year. Applying the function to each Eurobarometer year resulted in large lists containing all the question for that year cleaned of all meta data, with new row headings, and with all data converted to the correct format.

Year	Level	UE28 EU28	BE	BG	CZ	DK	D- W	DE	D- E	EE	IE	EL	ES	FR	HR	IT	CY	
1	2016 Aut	TOTAL	27705.00	1022.00	1012.00	1004.00	1006.00	1011.00	1531.00	520.00	1005.00	1006.00	1008.00	1011.00	1000.00	1062.00	1021.00	500.00
2	2016 Aut	Fréquemment	6218.00	175.00	200.00	165.00	324.00	304.00	462.00	159.00	210.00	201.00	378.00	190.00	208.00	195.00	179.00	106.00
3	2016 Aut	Frequently	0.23	0.17	0.20	0.16	0.32	0.30	0.30	0.31	0.21	0.20	0.37	0.19	0.21	0.18	0.18	0.21
4	2016 Aut	Occasionnellement	15047.00	619.00	601.00	668.00	546.00	593.00	894.00	299.00	622.00	482.00	521.00	469.00	476.00	639.00	532.00	226.00
5	2016 Aut	Occasionally	0.54	0.61	0.59	0.67	0.54	0.59	0.58	0.57	0.62	0.48	0.52	0.46	0.47	0.60	0.52	0.45
6	2016 Aut	Jamais	6390.00	226.00	200.00	168.00	133.00	108.00	165.00	59.00	171.00	319.00	108.00	352.00	316.00	227.00	310.00	163.00
7	2016 Aut	Never	0.23	0.22	0.20	0.17	0.13	0.10	0.11	0.11	0.17	0.32	0.11	0.35	0.32	0.22	0.30	0.33
8	2016 Aut	NSP	50.00	2.00	11.00	3.00	3.00	7.00	10.00	3.00	2.00	4.00	0.00	0.00	0.00	0.00	1.00	6.00
9	2016 Aut	DK	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01

Figure 3.23- Example of a cleaned and formatted Eurobaromter question from EB86- 2016

A year column would be needed to be able to record which Eurobarometer survey the data had come from. A simple code chunk was used for each chunk with a year and season input being requested when executed. This year and code input would then be attached to each question located in the large lists, secondly all NA values were converted to a ‘0’.

```

r [r 2016]
# I GOT IT!!!
#-----
Year_var <- readline(prompt = "Input year & Season: ")
EB_85_2016_spring <- lapply(openxlsx_85, EB_Convert_ordered) # 2015 spring
EB_85_2016_spring <- lapply(EB_85_2016_spring, function(x) {x[is.na(x)] <-0; return(x) })
#-----
Year_var <- readline(prompt = "Input year & Season: ")
EB_86_2016_Aut <- lapply(openxlsx_86, EB_Convert_ordered) # 2015 AUT
EB_86_2016_Aut <- lapply(EB_86_2016_Aut, function(x) {x[is.na(x)] <-0; return(x) })
#-----

```

Figure 3.24- The `Year_var` was recorded when the chunk was executed and then applied in the `EB_convert_ordered` function. This added the year column to the start of each question.

Finally, after having loaded each individual Eurobarometer’s questions into RStudio, and then having cleaned and added useful columns to the data. Questions useful to the research hypothesis would have to be amalgamated into a dataframe that contained the responses to the questions across all available years, listed by country.

```

r [r Q1_Nationality]

Q1_Nationality <- bind_rows(
  #EB_83_2015_spring$D71a.2,
  #EB_84_2015_Aut$D71a.2,
  EB_85_2016_spring$Q1,
  EB_86_2016_Aut$Q1NAT,
  EB_87_2017_spring$Q1NAT,
  EB_88_2017_Aut$Q1NAT,
  EB_89_2018_spring$Q1NAT,
  EB_90_2018_Aut$Q1NAT,
  EB_91_2019_spring$Q1NAT,
  EB_92_2019_Aut$Q1NAT,
  EB_93_2020_spring$T2,
  EB_94_2020_Aut$Q1NAT,
  EB_95_2021_spring$Q062,
  #EB_96_2021_Aut$D71_2,
  #EB_97_2022_spring$D71_2
) # This rbinds all

colnames(Q1_Nationality)
Q1_Nationality <- as.data.frame(Q1_Nationality[, c(1:3, 40, 41, 43, 4:39, 42, 44:49 )]) # Rearrange
cols based on index

```

Figure 3.25- Combining questions from each year. Column rows varied year on year, the final step above, rearranges the columns as needed.

	Year	Level	UE28 EU28	UE28-UK EU28-UK	UE27 EU27	CH	BE	BG	CZ	DK	D-W	DE	D-E	EE	IE	EL
1	2015 Spring	TOTAL	27758.00	NA	NA	NA	1014.00	1063.00	1021.00	1020.00	1033.00	1554.00	521.00	1001.00	1018.00	999.00
2	2015 Spring	Plutôt confiance	11102.00	NA	NA	NA	482.00	592.00	435.00	586.00	427.00	610.00	163.00	553.00	443.00	261.00
3	2015 Spring	Tend to trust	0.40	NA	NA	NA	0.48	0.56	0.43	0.57	0.41	0.39	0.31	0.55	0.44	0.26
4	2015 Spring	Plutôt pas confiance	12730.00	NA	NA	NA	455.00	290.00	464.00	324.00	457.00	740.00	318.00	187.00	398.00	726.00
5	2015 Spring	Tend not to trust	0.46	NA	NA	NA	0.45	0.27	0.45	0.32	0.44	0.48	0.61	0.19	0.39	0.73
6	2015 Spring	NSP	3926.00	NA	NA	NA	76.00	181.00	123.00	110.00	149.00	203.00	40.00	261.00	177.00	13.00
7	2015 Spring	DK	0.14	NA	NA	NA	0.07	0.17	0.12	0.11	0.15	0.13	0.08	0.26	0.17	0.01
8	2015 Aut	TOTAL	27681.00	NA	NA	NA	1031.00	1035.00	1013.00	1001.00	1031.00	1548.00	517.00	1004.00	1004.00	1002.00
9	2015 Aut	Plutôt confiance	8732.00	NA	NA	NA	405.00	455.00	270.00	472.00	300.00	439.00	131.00	405.00	331.00	178.00
10	2015 Aut	Tend to trust	0.32	NA	NA	NA	0.39	0.44	0.27	0.47	0.29	0.28	0.25	0.40	0.33	0.18
11	2015 Aut	Plutôt pas confiance	15368.00	NA	NA	NA	559.00	358.00	642.00	413.00	639.00	971.00	338.00	288.00	521.00	813.00
12	2015 Aut	Tend not to trust	0.55	NA	NA	NA	0.54	0.35	0.63	0.41	0.62	0.63	0.66	0.29	0.52	0.81
13	2015 Aut	NSP	3581.00	NA	NA	NA	67.00	222.00	100.00	116.00	91.00	138.00	47.00	312.00	152.00	11.00
14	2015 Aut	DK	0.13	NA	NA	NA	0.07	0.21	0.10	0.12	0.09	0.09	0.09	0.31	0.15	0.01

Figure 3.26- The resultant dataframe: Question QA8- 'how much do you trust the EU?'

In the example question above, we can see that UE28 EU28 features in both 2015 Spring and 2015 Autumn. Later years utilised one of the other options listed in column 4 (UE28-UK, EU28-UK) or column 5 (UE27 EU27). Each nation has its' responses recorded year by year and is organised by the Level column.

After beginning the process of loading all 44 available questions into R and wrangling them into useful data it became clear that the scale of the effort required to load all the available questions went far beyond the time available. A small selection of pertinent questions was selected to be used in this analysis that were essential in exploring the research question. There are undoubtedly further interesting insights in the other questions, this is addressed in more detail in the Conclusions and Future Work chapter. The following questions, available across the range of Eurobarometer surveys used in the project, were selected as the most appropriate to the research question:

Question code (varies by year)	Description
D71a2:	'When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...?': 'European Political Matters'
D71a3:	'At the present time, would you say that, in general, things are

	going in the right direction or in the wrong direction, in...?': 'The EU'
D78 :	In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?'
QA8 :	'I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it.': 'The EU'

3.4.4. Exploratory analysis

Having wrangled the Eurobarometer questions, the scraped_EUvsDisinformation data and the Kaggle_EUvsDisinformation data we were ready to move on to exploratory analysis. The Eurobaromter data would be subjected to a more statistical analysis (see chapter 4) given that entirety of the data was numerical. The three datasets required different approaches due to the differences in the data contained within each.

The Scraped EUvsDisinformation data:

Name	Description
Date	Formatted as date: YYYY-mm-dd
Title	<chr> A description of the disinformation instance
Outlets	<chr> A record of the source of the disinformation
Country	<factor> A record of the country targeted in the disinformation instance. 118 factor levels.

The Kaggle Data:

Name	Description
Claim_Published	Formatted as date: YYYY-mm-dd
Keyword_Name	<chr> A description of the keywords contained in the disinformation instance
Country_name	<factor> A record of the country targeted in the disinformation instance. 98 factor levels
Organisation_name	<chr> A record of the source of the disinformation
Language_name	<chr> A record of the language used in the disinformation instance. 37 possible languages

Eurobarometer D71a.2: ‘When you get together with friends or relatives, would you say you discuss frequently, occasionally, or never about...? European Politics’

Name	Description
Year	<factor> The year and season of the survey. 15 factor levels
Level	<factor> The response to the questions. 20 factor levels
Country results (DE, UK, IE, PL)	<numeric> The results per year and per level to the question as a numeric value

Eurobarometer D71a.2: ‘When you get together with friends or relatives, would you say you discuss frequently, occasionally, or never about...? European Politics’

Name	Description
Year	<factor> The year and season of the survey. 15 factor levels
Level	<factor> The response to the questions. 20 factor levels
Country results (DE, UK, IE, PL)	<numeric> The results per year and per level to the question as a numeric value

Eurobarometer D71a.3: ‘At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...?’ The EU.

Name	Description
Year	<factor> The year and season of the survey. 15 factor levels
Level	<factor> The response to the questions. 19 factor levels
Country results (DE, UK, IE, PL)	<numeric> The results per year and per level to the question as a numeric value

Eurobarometer QA8: ‘I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it.’

Name	Description
Year	<factor> The year and season of the survey. 15 factor levels
Level	<factor> The response to the questions. 7 factor levels
Country results (DE, UK, IE, PL)	<numeric> The results per year and per level to the question as a numeric value

The scraped data required some small amount of wrangling to clean up the column headings and convert the data column to the correct format, there were no missing values to be removed in the data. However, the data that was of most use was contained in lists in the data as currently constructed, this would have to be reconstituted in order to ensure the maximum utility of the data:

Country
Ukraine
EU, US
Switzerland
Russia, Ukraine
Ukraine
Russia, US, Europe
Ukraine, Russia, US, EU
Ukraine, Russia, France, Germany, UK
Russia, Ukraine, France
Ukraine, Russia, France
Ukraine
Ukraine, US
Russia, Ukraine
Ukraine

Figure 3.27- Country data contained in lists in the Scraped data.

In order to maximise the fidelity of the data, a simple command was used to separate the rows and replicate the data for each list item. This converted the original Scaped dataframe, containing 15K rows into an atomic instance dataset containing the same columns but 66,409 rows, the Scraped data was now prepared and ready for exploratory analysis.

Date	Title	Outlets	Country
2023-01-17	Persecution of the Ukrainian Orthodox Church aims to open...	t.me	Ukraine
2023-01-17	NATO is Washington's territorial expansion mechanism	sputniknews.lat	EU
2023-01-17	NATO is Washington's territorial expansion mechanism	sputniknews.lat	US
2023-01-17	The World Economic Forum aims to reduce the human pop...	Laivaslaikrastis.lt	Switzerland
2023-01-17	Apartment building in Dnipro was destroyed by Ukrainian ai...	oroszhirek.hu	Russia
2023-01-17	Apartment building in Dnipro was destroyed by Ukrainian ai...	oroszhirek.hu	Ukraine
2023-01-17	Ukrainian air defense missile destroyed a residential buildin...	tsargrad.tv	Ukraine
2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	Russia
2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	US
2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	Europe
2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	Ukraine
2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	Russia
2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	US
2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	EU
2023-01-16	Zelenskyy sends his soldiers to be slaughtered for Western ...	almatareed.org	Ukraine

Figure 3.28- Detail of the atomic data after separating the Country list elements. Original rows have been replicated for each list item.

The Kaggle data contained a significant amount of missing data:



```
claim_published      keyword_name      country_name      abstract      organization_name      language_name
0                   14                26                6544         0                       4150
```

Figure 3.29- Sum of missing values in the Kaggle data

The `abstract` column in the Kaggle data (containing a paragraph about each disinformation instance in the original language, mostly in Russian, some Arabic, Italian and other languages). `Language_name` was also missing values in over half the available rows. This column contained a list of the languages that the disinformation instance had been originally written in. However, many of the cells contained the language name but were missing the abstract. Or contained a clearly national organisation but were still missing the `language_name`. The `abstract` and `language_name` columns both proved useless for our purposes.

However, removing all rows that contained a missing value left only 200 rows. It was decided to leave the dataframe as it was constructed and to utilise only the useful data contained in the dataset. The data values contained in `claim_published`, the `keyword_name` and the `country_name` were all used in the project.

A further issue was how the data was listed in the `keyword_name` and `country_name` columns in the Kaggle data:

keyword_name	country_name
['Abandoned Ukraine', 'Angela Merkel', 'Emmanuel Macron',...]	['Russia', 'Ukraine', 'The West', 'EU', 'Germany', 'France']
['Conspiracy', 'Terrorism', 'Donald Trump']	['Iran', 'United States', 'Saudi Arabia']
['Genocide', 'WWII', 'Historical revisionism', 'WWI', 'Austria-...]	['Ukraine', 'Poland']
['Russophobia', 'Encircling Russia', 'Operation Barbarossa']	['United States', 'The West', 'Czech Republic', 'Eastern Europe']
['European Parliament', 'WWII', 'Adolf Hitler', 'World War 2']	['Russia', 'Germany', 'Eastern Europe']
['Conspiracy', 'Climate', 'George Soros', 'migration']	['Spain']
['Conspiracy', 'Protest']	['United States', 'Hong Kong']
['Ukraine', 'Anti-Semitism']	['Ukraine']
['NATO', 'Encircling Russia']	['Russia', 'The West']
['Conspiracy', 'Protest']	['The West', 'Hong Kong']
['Western values', 'Conspiracy', 'Secret elites / global elites']	['US', 'The West', 'EU']
['Western values', 'Conspiracy', 'Secret elites / global elites']	['US', 'The West', 'EU']

Figure 3.30- Keyword_name and country_name lists contained in single cells

This data was inaccessible in this format and would have to be separated into atomic values to render it functional for the experiment. This process was further complicated by the inclusion of the square brackets.

```

# Separate out rows in Kaggle Data- IN USE

test <- clean_kaggle
test <- as_tibble(test)

test2 <- separate_rows(test, 'country_name', convert = TRUE, sep = ", ")
test3 <- separate_rows(test2, 'keyword_name', convert = TRUE, sep = ", ")
#test4 <- separate_rows(test3, in_language, convert = TRUE, sep = ", ") #This col is removed as
a duplicate

SEP_Kaggle <- test3
rm(test, test2, test3)

```

Figure 3.31- Initial step to separate the Kaggle data into atomic rows

The above code chunk failed to deal with the square brackets on the ends of the lists. Thus, any listed country that was at the start or end of a list, still had the square bracket remaining at either the start or the end of the string following execution of the code: ['Ukraine', 'Ireland'] became "[Ukraine" and "Ireland"].

A second code block was written utilising regex to remove any further extraneous data:

```

# (x SEP_Kaggle- Replace unneeded strings- IN USE)

# this block replaces all [' ', '] from the cols in SEP_kaggle.
# Needed since Ukraine, ['Ukraine & Ukraine'] are being recorded as separate values
# May have to come back to this and apply it to clean_kaggle instead of the separate DF.
# Had to come back fix the replacement value with nothing

SEP_Kaggle <- SEP_Kaggle %>%
  mutate(keyword_name = str_replace_all(keyword_name, "[[:punct:]]", ""))

SEP_Kaggle <- SEP_Kaggle %>%
  mutate(country_name = str_replace_all(country_name, "[[:punct:]]", ""))

SEP_Kaggle <- SEP_Kaggle %>%
  mutate(language_name = str_replace_all(language_name, "[[:punct:]]", ""))

```

Figure 3.32- Code block to remove any and all punctuation from Kaggle dataset columns

Initially the data was made completely atomic in that any row that could have a separated value was separated, however it was later determined that quite a lot of the data would not be of use and so these steps were delimited to the data that we required. The abstract and organisation name columns were not separated. There meant that some organisations were captured in lists, but they were not used at a later point so this was deemed acceptable. Adding unnecessary separations tended to create datasets of hundreds of thousands of rows. The two datasets, after having the atomic values coerced were arranged as follows:

- SEP_Kaggle: 23K rows with 5 columns:

	claim_published	keyword_name	country_name	organization_name	language_name
1	2019-12-13	Abandoned Ukraine	Russia	sputnik.by // lifenews.ru	Russian
2	2019-12-13	Angela Merkel	Russia	sputnik.by // lifenews.ru	Russian
3	2019-12-13	Emmanuel Macron	Russia	sputnik.by // lifenews.ru	Russian
4	2019-12-13	Ukrainian statehood	Russia	sputnik.by // lifenews.ru	Russian
5	2019-12-13	Vladimir Putin	Russia	sputnik.by // lifenews.ru	Russian
6	2019-12-13	Minsk agreements	Russia	sputnik.by // lifenews.ru	Russian
7	2019-12-13	Abandoned Ukraine	Ukraine	sputnik.by // lifenews.ru	Russian
8	2019-12-13	Angela Merkel	Ukraine	sputnik.by // lifenews.ru	Russian
9	2019-12-13	Emmanuel Macron	Ukraine	sputnik.by // lifenews.ru	Russian
10	2019-12-13	Ukrainian statehood	Ukraine	sputnik.by // lifenews.ru	Russian
11	2019-12-13	Vladimir Putin	Ukraine	sputnik.by // lifenews.ru	Russian
12	2019-12-13	Minsk agreements	Ukraine	sputnik.by // lifenews.ru	Russian
13	2019-12-13	Abandoned Ukraine	The West	sputnik.by // lifenews.ru	Russian
14	2019-12-13	Angela Merkel	The West	sputnik.by // lifenews.ru	Russian
15	2019-12-13	Emmanuel Macron	The West	sputnik.by // lifenews.ru	Russian

Figure 3.33- Kaggle data, each coloured blocked has been separated into atomic data such that each lists item has been extracted and the row on which it was originally contained, has had the data replicated.

- SEP_Scrape: 66K rows with 4 columns:

	Date	Title	Outlets	Country
1	2023-01-17	Persecution of the Ukrainian Orthodox Church aims to open...	t.me	Ukraine
2	2023-01-17	NATO is Washington's territorial expansion mechanism	sputniknews.lat	EU
3	2023-01-17	NATO is Washington's territorial expansion mechanism	sputniknews.lat	US
4	2023-01-17	The World Economic Forum aims to reduce the human pop...	Laivaslaikrastis.lt	Switzerland
5	2023-01-17	Apartment building in Dnipro was destroyed by Ukrainian ai...	oroszhirek.hu	Russia
6	2023-01-17	Apartment building in Dnipro was destroyed by Ukrainian ai...	oroszhirek.hu	Ukraine
7	2023-01-17	Ukrainian air defense missile destroyed a residential buildin...	tsargrad.tv	Ukraine
8	2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	Russia
9	2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	US
10	2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	Europe
11	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	Ukraine
12	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	Russia
13	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	US
14	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	EU
15	2023-01-16	Zelenskyy sends his soldiers to be slaughtered for Western ...	almatareed.org	Ukraine

Figure 3.34- Scraped data, the country col has been separated into atomic data such that each lists item has been extracted and the row on which it was originally contained, has had the data replicated.

The Date Column

The date range between the Scaped and Kaggle data were not initially aligned with one another. The date ranges are listed below:

```
range(as.Date(SEP_Scrape$Date)) |
# 2015-11-15 -> 2023-01-17

range(as.Date(SEP_Kaggle$claim_published))
# 2015-11-08 -> 2020-01-02

# EB date range: 2015 -> 2022
...
```

Figure 3.35- Date ranges across all three datasets

The scraped data dates were scraped from the website in late January and as can be seen in the image, the database is consistently being added to. The Kaggle data dates however were limited to 2020. The Eurobaromter data was delimited to 2022 at the design stage. The differences in the date ranges meant that date would have to be

subset to align the dates so that the data could be successfully validated. All further exploration was carried out on the dates delimited by the Eurobarometer data (2015 to 2022).

Plotting of the unaligned dates contained in the Scraped and Kaggle datasets demonstrated some concerning aspects of the data. Considering that both datasets were ostensibly sourced from the same organisation albeit at perhaps three years removed from one another, the date ranges did not align particularly well:

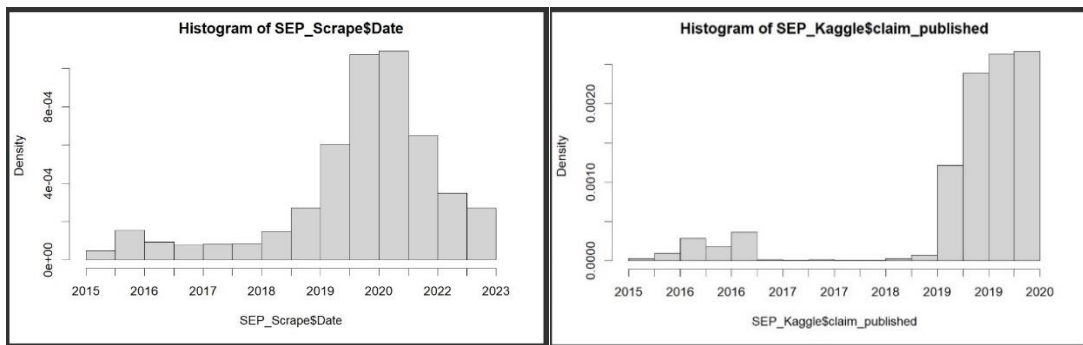


Figure 3.36- Plots of Unaligned Dates

The histograms displayed a problem misalignment for dates between 2017 and 2019. On the other hand, the scale was significantly different between the plots. The Scraped data also seemed to display a fairly normal, if skewed distribution, these aspects will be explored in chapter 4.

After aligned the dates between the two datasets, the data was somewhat more aligned and closer in scale. However, the lack of data between 2017 and 2019 for the Kaggle data raised serious questions as to the fidelity of the data overall.

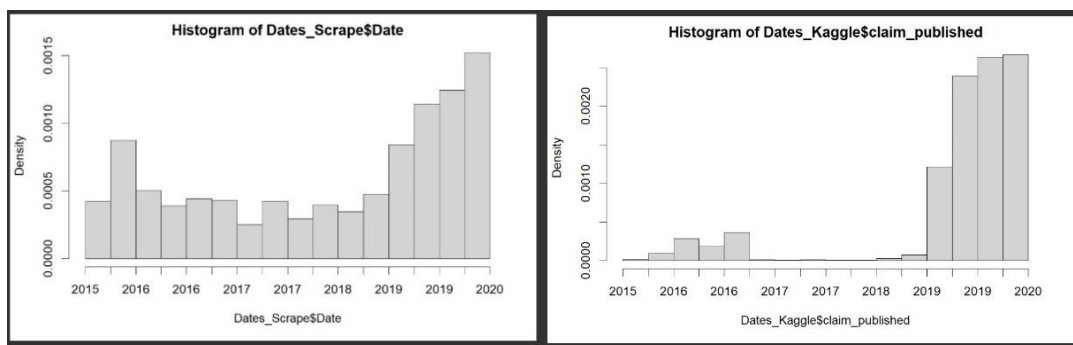


Figure 3.37- Plots of Aligned Dates

Using ggplot to better control the aesthetics of the plots confirmed that the date data was not particularly well aligned overall:

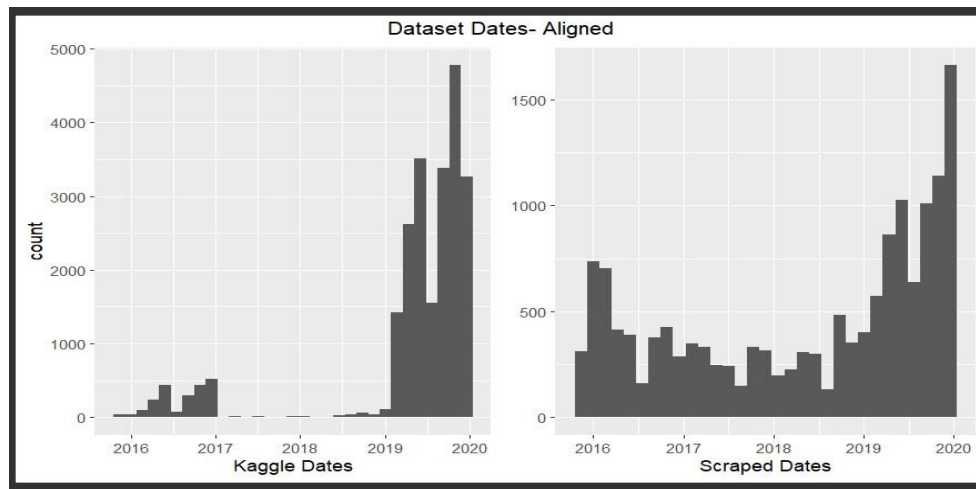


Figure 3.38- GGPlot of Unaligned Dates

It was decided that since the function of the two datasets was not similar, that the apparent lack of alignment, may be caused by the different goals of each dataset. The website data that had been scraped, seeks to be an overall, up-to-date overview of disinformation that EUvsDisinformation has detected, whereas the Kaggle date had been used for a Hackaton at some point in 2020 with a more Machine Learning focused goal. This potential issue as spotted in the data exploration would have to be explored in later analyses to see if combining the datasets together had any validity.

The Country Column

The most important data was most likely to be the country data contained in both datasets. Looking at the data contained in the scraped data we can see that Russia, Ukraine, and the United States are far and away the target of more disinformation than other nations. None of these nations were included in the Eurobarometer data and so these countries were excluded. The EU value was also of little use as while it targeted the EU as a single institution, the Eurobarometer data was recorded and utilised on a member state basis. Germany, the UK and Poland featured in the top 5 when other nations, that were not part of the study, were removed.

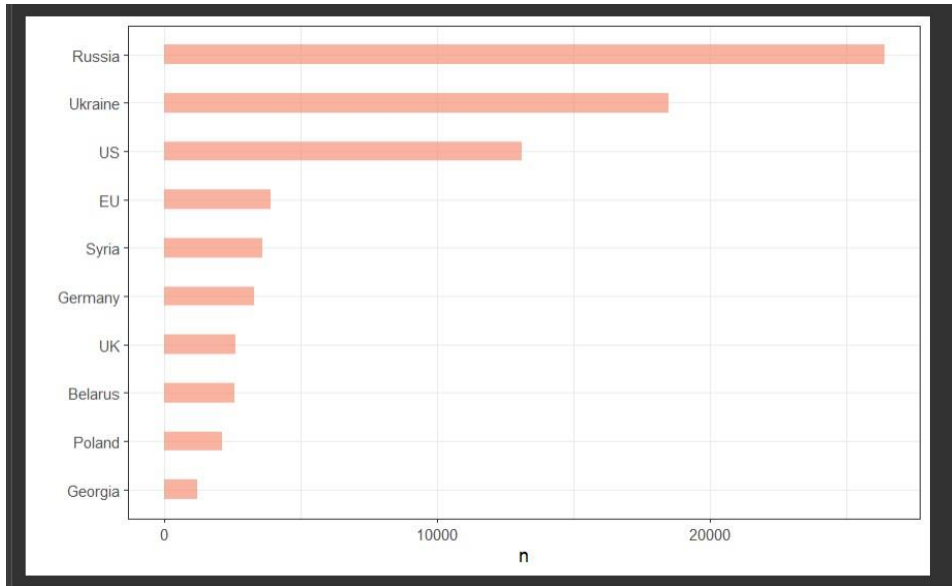


Figure 3.39- scraped Data- Sum of countries (Scraped Data)

Similar results were found in the Kaggle data. There were some variations but overall, the data was very similar:

- the United States was recorded as both the 'US', and as: 'United States'.
- We see the same top 4, Russia, Ukraine, US and EU
- We have the addition of The West and United States.
- Following that, Poland, UK and Germany feature in the top 5 again.

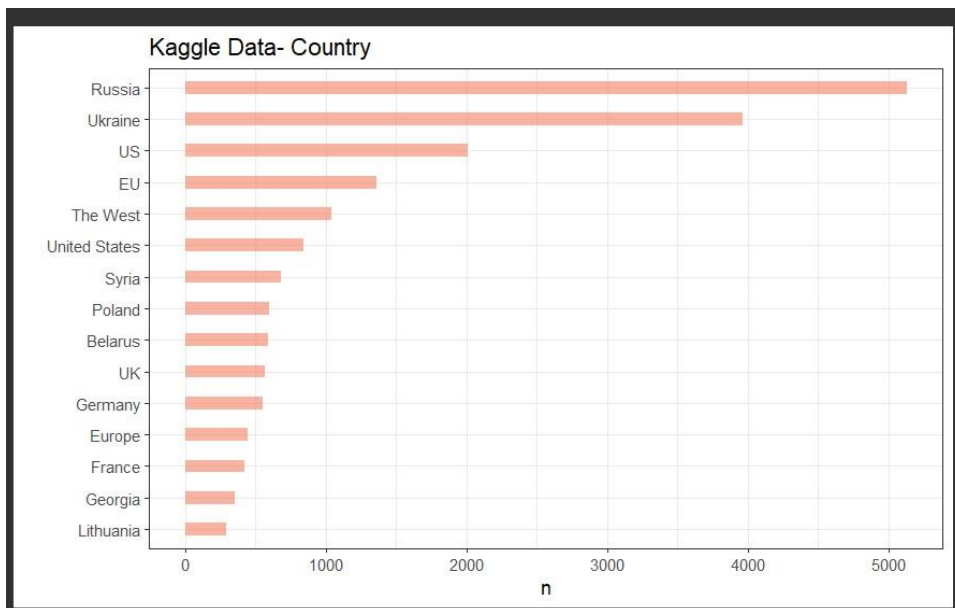


Figure 3.40- scraped Data- Sum of countries (Kaggle Data)

When plotted together, the similarities of the plots can easily be seen, the counts are also quite similar:

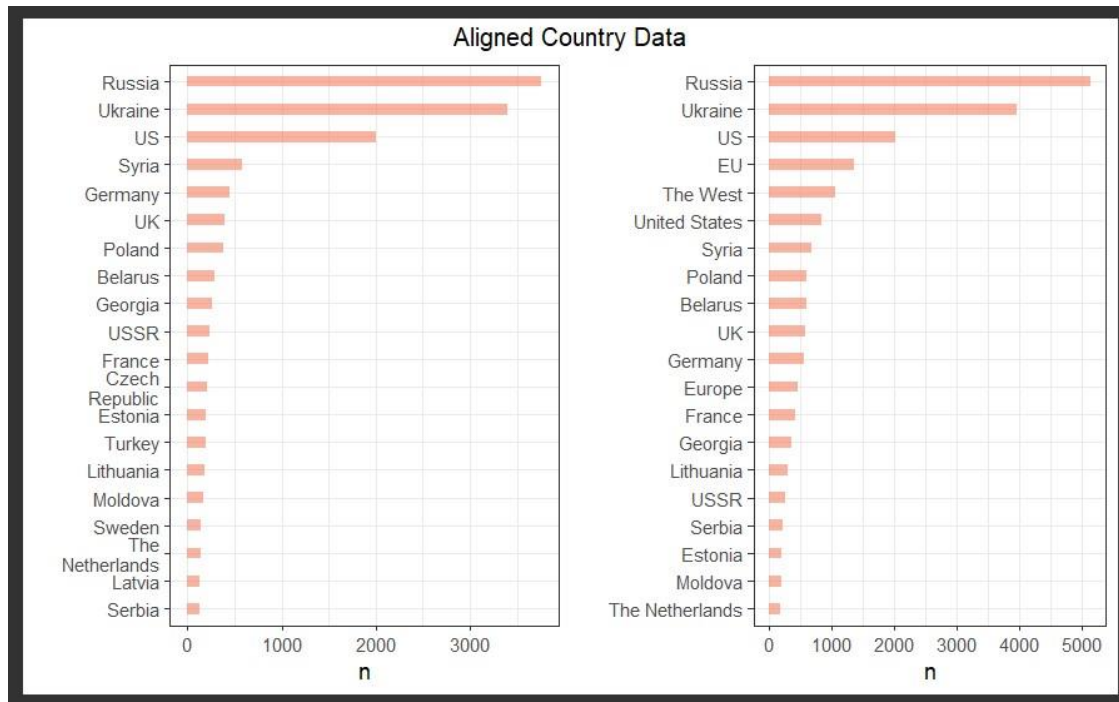


Figure 3.41- scraped Data- Sum of countries (Comparing Scraped and Kaggle Data)

Overall, the country data seems of high fidelity between the two datasets and several European member states have been targeted by disinformation during the study period.

The Outlet Column

The outlet data was initially expected to constitute a large part of the eventual experiment design, however on exploring the data, it was rejected as of little use overall. Both datasets contained a column that captured the original source of the disinformation however the structure of the columns was different and of little relation to the alternative dataset.

The scraped data attempted to capture both the title of the source, often a youtube or other media site video, or a newspaper article; as well as the source website. This resulted in extremely long strings that were too ungainly to be of real use:

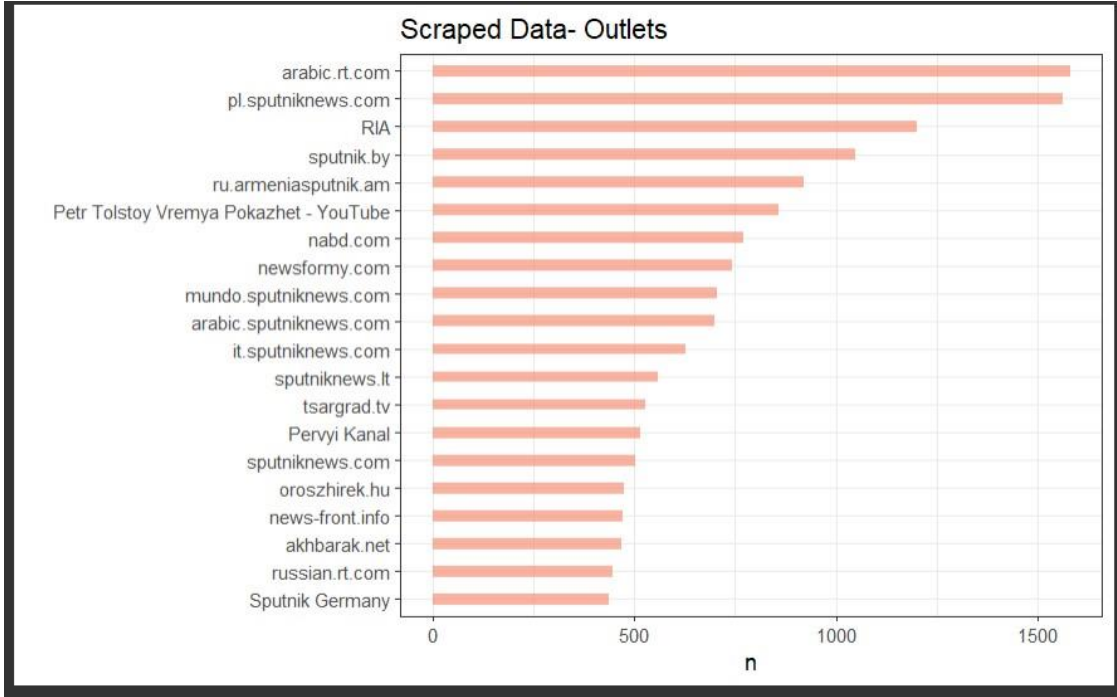


Figure 3.42- scraped Data- Sum of Outlets

‘Petr Tolstoy Vremya Pokazhet- Youtube’ is just one example of the troublesome structure to this data. The Kaggle data had a different but equally difficult style in how the data was recorded:

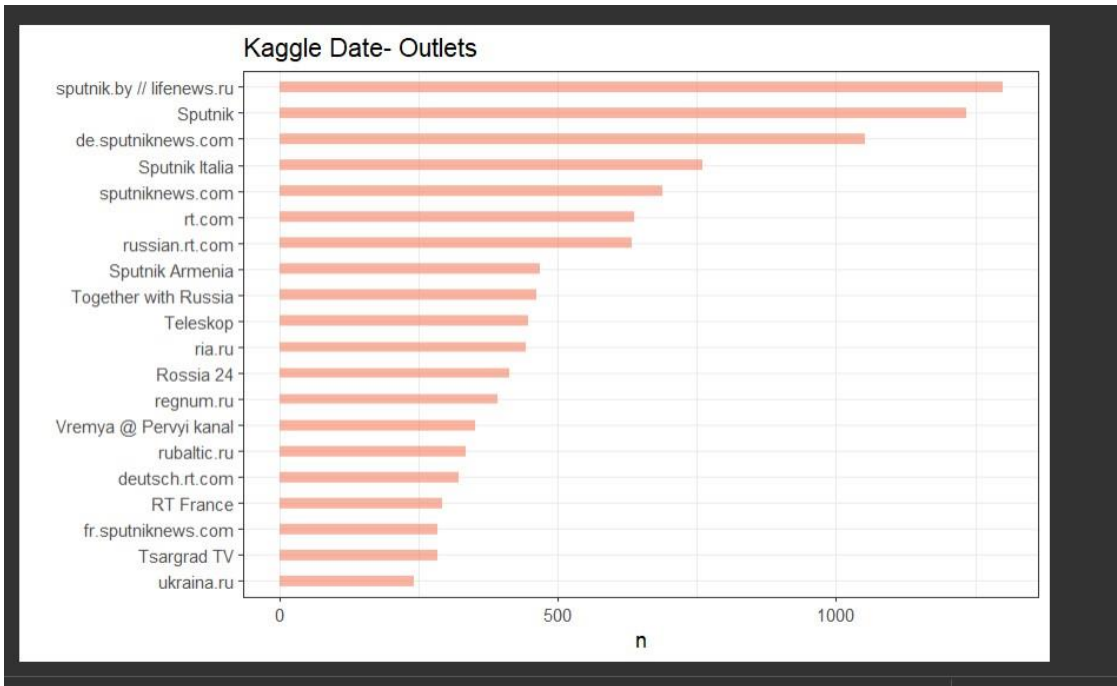


Figure 3.43- Kaggle Data outlets

As can be seen, there is little alignment between the two datasets at face value. However, looking at some of the recorded values we can see that they are alternative recordings of the same data as contained in the Scraped date. “Vremya Pervyi kanal” seems to be the same video as mentioned in the previous dataset. Other web sites such as sputnik news, or rt.com are also recorded in both but clearly utilising a different stylistic approach that would necessitate a complex system to extract the data with any sense of fidelity. Both columns were rejected for use in the experiment.

The Kaggle Data- Language Column

Only the Kaggle dataset contained language data. When this data is plotted, some interesting insights were gained:

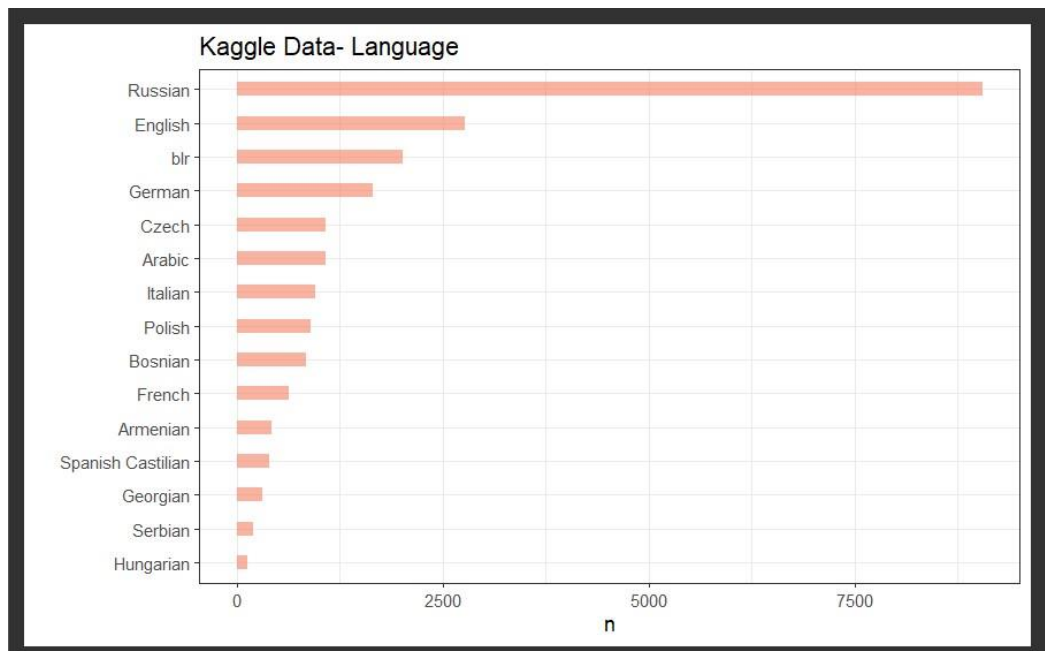


Figure 3.44- Kaggle sum of languages used.

Russian was by far the most used language. English features as the second most common language, this is presumed to be due to its international nature. Bulgarian is unexpectedly the third most common language, an anachronism that was not easily explained by the data. Similarly, Czech made a surprise placement in the top 5, almost equal to Arabic, although Arabic’s inclusion is better explained by the prevalence of Syria in the country data (see above). Within the top 10 we find German, English and Polish. While the other language are less easily explained, those three regions of Europe have featured yet again.

3.5. *The Kaggle Dataset*

The final column contained in the Kaggle data was the keyword data. This was unlikely to be used in the eventual study but does bear exploration as it helped with further understanding of the data overall:

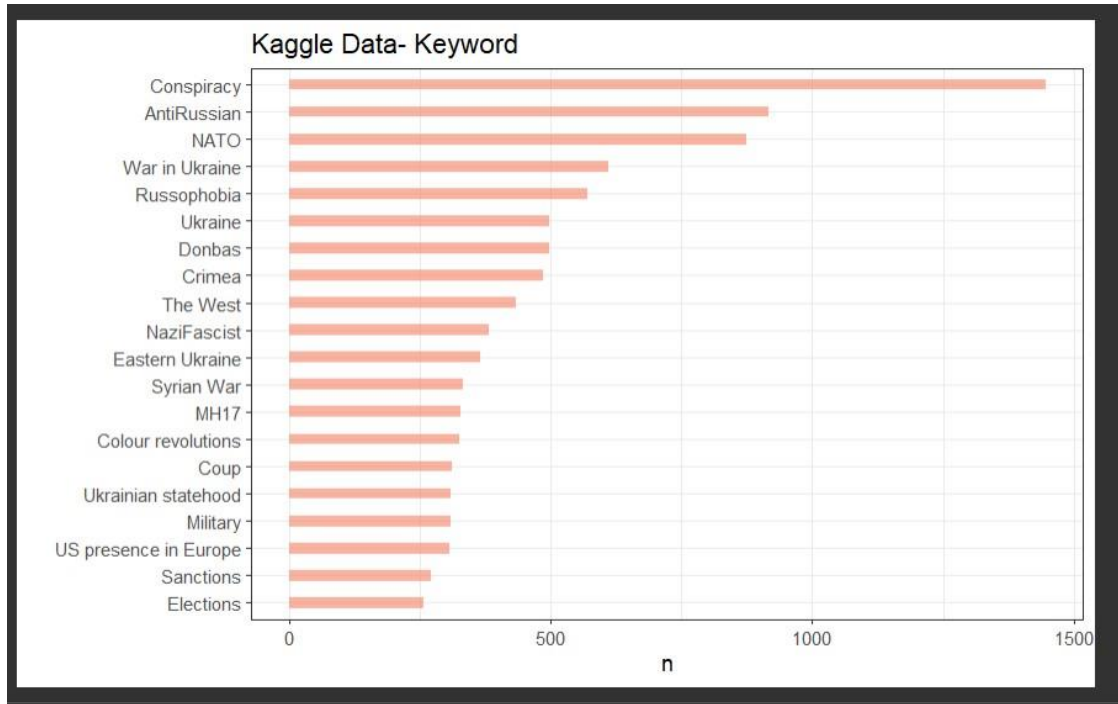


Figure 3.45- Kaggle Keyword Data

The range of keywords here is illuminating. A clear focus on creating a supposed sense of Anti-Russian bias can clearly be seen. While several extremely sensitive issues are also seen such as ‘MH17’, ‘Colour revolutions’, and ‘War in Ukraine’. Secondly, a very clear focus on Ukraine connected topics can be seen. Most interestingly, conspiracy features as the most common keyword and is by far the most used term overall. Ultimately this data was fascinating but of no direct use in this study. However, it does present several interesting possibilities for further research.

3.6. *Machine Learning Approaches*

Given that a wide range of data is available, and that Machine Learning is essentially the state of the art in terms of data analytics and data Science, it behoves us to at least explore its utility in relation to these datasets. A number of approaches are possible with the data that we have, later statistical exploration of the Eurobaromter

data would further clarify possible approaches that could be taken. Each of the approaches were tentative at this stage and the eventual application of (k-NN or LR) was only decided at a later stage following much more detailed analysis of the datasets as well as the results of the experiment.

- K-nearest neighbour, as a supervised, classification approach could be useful in identifying different types of disinformation that are targeted at different nations by utilising the wider data available in the Kaggle data to establish links between keywords, language and the target of the disinformation.
- Linear Regression, using the Eurobarometer data (see chapter 4) could be used to model the relationship between changes in attitude in one European Member State and others that may also have a change in their attitude to the EU during the time frame.
- Random Forest as an ensemble learning method could contain deeper insights that other methods in that it combines some of the strengths of some of the other approaches. By utilising random forests to balance the overfitting potential of decision trees alone, interesting insights could be gained in the interconnectedness of the Kaggle data given that it is much richer than the other datasets.
- Support Vector Machine, sadly beyond the scope of this study as the most useful data has been rejected for the purposed of the research question; but SVM could easily be used, especially in relation to the complex text data in the Kaggle and Scraped datasets to help classify the disinformation instances.
- An alternative approach utilising the same data would be Neural Networks. The weighted data approach structured into nodes that resemble biological neural networks, can be used to find unexpected and hidden links between the data points and the disinformation.

3.7. Key Reflections

Overall, the research question was based on correlation and so Machine Learning approaches were not expected to be the most appropriate approach. The bulk of the academic work in this area does however, focus on the use of ML in order to facilitate identification and classification of the Levithan and ever increasing amount of Tweets, Youtube videos and Facebook posts that are the field upon which disinformation is laid ad nauseum. Machine Learning was utilised at the end of the experiment to gain useful insight into the implications of the data, but this was not the original focus of the study.

The main finding in the explorations of the datasets in relation to the research question is that Germany, the UK and Poland feature in both datasets within the top 5 overall when other nations, outside the scope of this study are removed. Secondly, the date data is not very robust or at least, the collection criteria may have changed between the creation of the Kaggle dataset and the ongoing collection of the disinformation database. The date data in the Scraped data overall seems more robust in that there are no obvious date ranges that seem to be missing data.

The main focus of the study was refined, following the exploratory analysis, to be: ‘have there been any changes in the national attitudes of Germany, the UK and Poland in the study period given that they have been targeted by disinformation, as recorded in both the Kaggle and scraped disinformation datasets?’

4. PROJECT DEVELOPMENT

4.1. Introduction

In this chapter the structure of the Kaggle_EUvsDisinfo and Scraped_EUvsDisinfo data will be reviewed to better understand their functions regarding the research question of whether EU member states that received a larger proportion of disinformation based on these datasets have had a statistically significant change in their attitudes towards the EU as measured using the Eurobarometer dataset.

The final steps in wrangling and exploration of the Eurobarometer data will be explained in this chapter with extensive images to help better clarify the configuration of the data. Secondly, statistical exploration will be carried out and explained in detail regarding the implications and potential sources of some of the unexpected variations in the data. Finally, the results will be discussed as regards the potential impact on our statistical tests to be carried out in the next chapter.

4.2. The EUvsDisinformation Datasets

The Scraped_EUvsDisinfo data and the Kaggle_EUvsDisinfo data has been extensively wrangled by this stage in the project. The original data that was scraped from the EUvsDisinformation website²⁴ database has been separated into more atomic rows so that each country mentioned in each instance in the database can be measured individually rather than as part of a list of countries. Similarly, the Kaggle_EUvsDisinfo has been separated atomically by country, earlier approaches that further separated the data by keyword proved counter-productive as the dataset was expanded to hundreds of thousands of rows and became ungainly.

The key outcomes from the exploration of the EUvsDisinformation data thus far were seen to be that while several nations outside the scope of this study, dominated the counts of disinformation overall, that Germany, the UK and Poland all featured within the top nations that can be said to have been targeted based on the data available. This was further confirmed when looking at the main languages used after removing the languages outside the scope of the study (Russian, Bulgarian). Finally,

²⁴ https://euvsdisinfo.eu/disinformation-cases/?offset=100&per_page=100

while a simple exploration of key words and the outlets where the disinformation was first detected, contained in the Kaggle_EUvsDisinformation dataset, proved very illuminating, it was not useful in answering our research question. These issues will be addressed further in Chapter 6,.

	Date	Title	Outlets	Country
1	2023-01-17	Persecution of the Ukrainian Orthodox Church aims to open...	t.me	Ukraine
2	2023-01-17	NATO is Washington's territorial expansion mechanism	sputniknews.lat	EU
3	2023-01-17	NATO is Washington's territorial expansion mechanism	sputniknews.lat	US
4	2023-01-17	The World Economic Forum aims to reduce the human pop...	Laivaslaikrastis.lt	Switzerland
5	2023-01-17	Apartment building in Dnipro was destroyed by Ukrainian ai...	oroszhirek.hu	Russia
6	2023-01-17	Apartment building in Dnipro was destroyed by Ukrainian ai...	oroszhirek.hu	Ukraine
7	2023-01-17	Ukrainian air defense missile destroyed a residential buildin...	tsargrad.tv	Ukraine
8	2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	Russia
9	2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	US
10	2023-01-17	The EU-NATO agreement aims to weaken the European Uni...	oroszhirek.hu	Europe
11	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	Ukraine
12	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	Russia
13	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	US
14	2023-01-17	Kyiv passed from rejecting nuclear weapons to preparing a ...	sputniknews.lat	EU
15	2023-01-16	Zelenskyy sends his soldiers to be slaughtered for Western ...	almatareed.org	Ukraine

Figure 4.1- The current structure of the scraped EUvsDisinformation data.

	claim_published	keyword_name	country_name	organization_name	language_name
1	2019-12-13	Abandoned Ukraine	Russia	sputnik.by // lifenews.ru	Russian
2	2019-12-13	Angela Merkel	Russia	sputnik.by // lifenews.ru	Russian
3	2019-12-13	Emmanuel Macron	Russia	sputnik.by // lifenews.ru	Russian
4	2019-12-13	Ukrainian statehood	Russia	sputnik.by // lifenews.ru	Russian
5	2019-12-13	Vladimir Putin	Russia	sputnik.by // lifenews.ru	Russian
6	2019-12-13	Minsk agreements	Russia	sputnik.by // lifenews.ru	Russian
7	2019-12-13	Abandoned Ukraine	Ukraine	sputnik.by // lifenews.ru	Russian
8	2019-12-13	Angela Merkel	Ukraine	sputnik.by // lifenews.ru	Russian
9	2019-12-13	Emmanuel Macron	Ukraine	sputnik.by // lifenews.ru	Russian
10	2019-12-13	Ukrainian statehood	Ukraine	sputnik.by // lifenews.ru	Russian
11	2019-12-13	Vladimir Putin	Ukraine	sputnik.by // lifenews.ru	Russian
12	2019-12-13	Minsk agreements	Ukraine	sputnik.by // lifenews.ru	Russian
13	2019-12-13	Abandoned Ukraine	The West	sputnik.by // lifenews.ru	Russian
14	2019-12-13	Angela Merkel	The West	sputnik.by // lifenews.ru	Russian
15	2019-12-13	Emmanuel Macron	The West	sputnik.by // lifenews.ru	Russian

Figure 4.2- the current structure of the Kaggle_EUvsDisinfomration data

4.3. The Eurobarometer Dataset

The Eurobarometer data has, by this stage, been loaded into R on a per survey basis. Data from Eurobarometer 83 (EB83-2015) to Eurobarometer 97 (EB97-2022) have been sourced from the `data.europa` website and loaded into R using the `openxlsx` package. Extensive wrangling has already taken place in order to clean NA values and restructure the individual questions into a useful shape regarding the research question.

Finally, four questions from the original list of 44 questions based on Eurobarometer 93 (EB93-2020) were selected and those questions were extracted from the seven years' worth of available survey data and combined into single dataframes that captured all member state's replies to the questions as well as various averages depending on the structure of the EU at the time of the survey (EU28, EU28-UK, EU27 + UK). Also included though not used in this study were answers from some prospective members as well as ill-defined political regions such as Cyprus.

At this stage there are 4 questions cleaned, wrangled and amalgamated into single dataframe constituting each question over the 7 year period of the study:

1. D71a.2: 'When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...? : European Political Matters'
2. D73a.2: 'At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...? : The EU'
3. D78: 'In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?'
4. QA8: 'I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it.'

These questions were chosen on the basis of their closeness to one another in subject matter as well as their appropriateness and utility in answering the key research question in this project. There are approximately 40 further eligible questions available but these four were chosen as having the most impact on this study. In chapter 5, other questions, not used thus far, will be discussed. A single example is provided below:

Year	Level	UE28 EU28	UE28-UK EU28-UK	UE27 EU27	UE27 EU27	BE	BG	CZ	DK	D-W	DE	D-E	EE	IE	EL
2015 Spring	TOTAL	27758.00	NA	NA	NA	1014.00	1063.00	1021.00	1020.00	1033.00	1554.00	521.00	1001.00	1018.00	
2015 Spring	Très positive	1362.00	NA	NA	NA	42.00	194.00	53.00	75.00	59.00	81.00	16.00	31.00	122.00	
2015 Spring	Very positive	0.05	NA	NA	NA	0.04	0.18	0.05	0.07	0.06	0.05	0.03	0.03	0.12	
2015 Spring	Assez positive	9955.00	NA	NA	NA	396.00	392.00	330.00	325.00	429.00	624.00	180.00	463.00	458.00	
2015 Spring	Fairly positive	0.36	NA	NA	NA	0.39	0.37	0.32	0.32	0.41	0.40	0.34	0.46	0.45	
2015 Spring	Neutre	10527.00	NA	NA	NA	356.00	308.00	425.00	446.00	371.00	571.00	209.00	408.00	299.00	
2015 Spring	Neutral	0.38	NA	NA	NA	0.35	0.29	0.42	0.44	0.36	0.37	0.40	0.41	0.29	
2015 Spring	Assez négative	4374.00	NA	NA	NA	188.00	107.00	174.00	129.00	141.00	222.00	88.00	66.00	87.00	
2015 Spring	Fairly negative	0.15	NA	NA	NA	0.19	0.10	0.17	0.13	0.14	0.14	0.17	0.07	0.09	
2015 Spring	Très négative	1049.00	NA	NA	NA	25.00	41.00	35.00	35.00	18.00	37.00	25.00	16.00	32.00	
2015 Spring	Very negative	0.04	NA	NA	NA	0.02	0.04	0.03	0.03	0.02	0.03	0.05	0.01	0.03	
2015 Spring	NSP	491.00	NA	NA	NA	7.00	21.00	4.00	10.00	14.00	18.00	3.00	17.00	19.00	
2015 Spring	DK	0.02	NA	NA	NA	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.02	
2015 Spring	Total 'Positive'	11317.00	NA	NA	NA	438.00	586.00	383.00	400.00	488.00	705.00	196.00	494.00	580.00	
2015 Spring	Total 'Négative'	5424.00	NA	NA	NA	213.00	148.00	209.00	164.00	160.00	260.00	113.00	82.00	120.00	
2015 Aut	TOTAL	27681.00	NA	NA	NA	1031.00	1035.00	1013.00	1001.00	1031.00	1548.00	517.00	1004.00	1004.00	1

Figure 4.3- Sample question at this stage of the project. D78- 'In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?' NA values caused by changes in the EU average over time.

4.4. Eurobarometer Data Final Steps

The Eurobarometer data, currently arranged on a per question basis, included many columns that were not of relevance to the current project. In the above image we can see that there are a large number of NA values, as well as a wide variety of national columns, none of which were the top targets of disinformation according to the previously discussed exploration of the Scraped and Kaggle data (see chapter 3). The NA values were coerced values due to the changes in the EU code over the study period based on how Brexit effected the data. In 2015, which was pre-Brexit, all Eurobarometer data up to that point was a simple count of the current number of member states 'UE28 EU28' (UE for Union Européenne in French). However, following the Brexit referendum, this code changed over time as the political structure of the EU shifted alongside the complex and time consuming Brexit negotiations. Eurobarometer surveys began to use a new code: (UE28-UK EU28-UK) for a number of years, this then later changed to 'UE27 EU27' as Brexit was confirmed as a permanent change in the relationship between the EU and the UK. Whether this continues and for how long, remains to be seen. However, it did cause an issue in this dataset in that those columns were coerced as NA values when the data was combined.

Secondly, a large number of values were added at the end of the dataset that were related to nations outside the EU that are geographically linked, or regional designations of politically unclear regions:

CH - Switzerland

NO - Norway

BA - Bosnia and Herzegovina

IS - Iceland

XK - Kosovo

CY - Cypriot Community

Since these regions were only surveyed in some years but not others, the years in which they were not included created further NA values in the data. This data was not used in the current study.

A number of imputation approaches were considered to deal with these NA values, however since the records that contained them either didn't directly impact the key research questions in this study, or the salient information could be inferred from other records, they were deemed unnecessary for this study but left in place for possible future work. Finally, the nations that were not seen to be targets of disinformation according to the EUvsDisinformation data as seen in the scraped and Kaggle EUvsDisinformation datasets (see Chapter 3), need not be explored in this study. The data was left as is and subsets were used to explore the three nations that were the most targeted (Germany- DE, The United Kingdom-UK, and Poland- PL). Additionally, Ireland (Ireland- IE) was also included as a control value to measure the differences in changes in the selected questions compared to those countries that had been targeted more by disinformation.

CY(tcc)	CH	NO	BA	IS	XK	Cy(Tcc)
0.21	NA	NA	NA	NA	NA	NA
51.00	NA	NA	NA	NA	NA	NA
0.10	NA	NA	NA	NA	NA	NA
34.00	NA	NA	NA	NA	NA	NA
0.07	NA	NA	NA	NA	NA	NA
195.00	NA	NA	NA	NA	NA	NA
0.39	NA	NA	NA	NA	NA	NA
154.00	NA	NA	NA	NA	NA	NA
0.31	NA	NA	NA	NA	NA	NA
NA	1104.00	1112.00	1046.00	513.00	1067.00	508.00
NA	45.00	42.00	130.00	38.00	185.00	60.00
NA	0.04	0.04	0.12	0.08	0.17	0.12
NA	332.00	297.00	387.00	155.00	350.00	157.00
NA	0.30	0.27	0.37	0.30	0.33	0.31
NA	466.00	463.00	455.00	194.00	441.00	158.00
NA	0.42	0.42	0.43	0.38	0.41	0.31

Figure 4.4- Less common country codes included only in some Eurobaromter surveys causing NA values to be coerced when amalgamated.

4.5. Eurobarometer Data Preparation

The data was firstly extracted for each individual nation that had been the target of disinformation as detailed in Chapter 3. The individual nation columns were selected from the amalgamated question data and reconstituted in a new dataframe. This was done for each country resulting in four new dataframes per nation that contained only that nation's responses to specific question.

```
## D71a2 European Politics
### DE
(r D71a_Euro_politics_DE)

#-----
# Extracing only year, Level and Germany

D71a.2_Euro_DE <- D71a.2_European_politics[,c("Year", "Level", "DE")]

# confirm classes are set correctly
D71a.2_Euro_DE$Year <- as.factor(D71a.2_Euro_DE$Year)
D71a.2_Euro_DE$Level <- as.factor(D71a.2_Euro_DE$Level)

#-----
```

Figure 4.5- The data related to German (DE) is selected from the overall amalgamated question data (D71a2-European Politics) and sent to a new national question dataframe.

This new dataframe contained all answers across all EB (EB83-2015 to EB97-2022) to each question used in the study by a single nation alone. In the example below we see the German response to question D71a2: 'When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...? : The EU'.

	Year	Level	DE
1	2015 Spring	TOTAL	1554.00
2	2015 Spring	Fréquemment	380.00
3	2015 Spring	Frequently	0.25
4	2015 Spring	Occasionnellement	1001.00
5	2015 Spring	Occasionally	0.64
6	2015 Spring	Jamais	173.00
7	2015 Spring	Never	0.11
8	2015 Spring	NSP	0.00
9	2015 Spring	Don't know	0.00
10	2015 Aut	TOTAL	1548.00
11	2015 Aut	Fréquemment	523.00
12	2015 Aut	Frequently	0.34
13	2015 Aut	Occasionnellement	905.00
14	2015 Aut	Occasionally	0.58
15	2015 Aut	Jamais	116.00
16	2015 Aut	Never	0.08
17	2015 Aut	NSP	4.00
18	2015 Aut	Don't know	0.00
19	2016 Spring	TOTAL	1593.00

Figure 4.6- New national response dataframe containing only a single nation's response to the question.

Issues arose at this stage regarding the use of French in the original `xlsx` files. The Eurobarometer dataframes were arranged with both French and English text used in different columns. However, the column headings were not consistent across all years, small changes in how the headings were organised created complex problems that had to be resolved. For example, in most years, the question D71a.2 had had the following potential responses:

- Fréquemment (Frequently)
- Occasionnellement (Occasionally)
- Jamais (Never)
- Ne sait pas (Don't know)
- TOTAL

However, this had changed over the years, for example 'TOTAL' had switched to 'Total' for some years and then switched back in later years, therefore these

columns would have to be carefully combined in order to maintain the data across the years:

```
# Changing factor levels so they match fixes the problem
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "Don't know" = "DK"))
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "Fréquentment" = "Fréquentment\nFrequently"))
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "Jamais" = "Jamais\nNever"))
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "Occasionnellement" =
"Occasionnellement\nOccasionally"))
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "Don't know" = "Ne sait pas\nDon't know" ))
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "NSP" = "Ne sait pas" ))
D71a.2_Euro_DE <- mutate(data, Level = fct_recode(Level, "TOTAL" = "Total" ))
```

Figure 4.7- Mutating questions to maintain the data across the years by combining column titles using strings.

In the case of D71a2, this was not a particularly difficult problem but in the case of question D78: 'In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?', there were many more columns with many different changes that required a slow and careful approach in order to ensure the fidelity of the data:

```
#-----
# Changing factor levels so they match fixes the problem
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "TOTAL" = "Total" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Très positive" = "Très
positive\nVery positive" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Très négative" = "Très
négative\nVery negative" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Assez négative" = "Assez
négative\nFairly negative" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Assez positive" = "Assez
positive\nFairly positive" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "DK" = "Don't know" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "NSP" = "Ne sait pas" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "NSP" = "Ne sait pas\nDon't know"
))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Neutre" = "Neutre\nNeutral" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Total 'Négative'" = "Total
'Négative'\nTotal 'Negative'" ))
D78_Image_EU_DE <- mutate(D78_Image_EU_DE, Level = fct_recode(Level, "Total 'Positive'" = "Total
'Positive'\nTotal 'Positive'" ))
```

Figure 4.8- Question D78 mutating rows to combine the changing column names over time.

Finally, after having cleaned and wrangled the Eurobarometer data into individual member state's responses over the data range into a single, targeted question; the data was pivoted wide to create the columns necessary for further study:


```
D78_Image_EU_DE_wide <- pivot_wider(D78_Image_EU_DE, names_from = "Level", values_from = DE, values_fn = list)
```

Figure 4.9- Pivot the data into longer format to create new columns.

The final dataframes thus constituted a national response, to an individual question related to the research question, across the entire date range of the available Eurobarometer data. This data had had a date added to it and was arranged sequentially so that changes in national response could be measured and explored over time. The data for Ireland was also included as a potential control to compare the changes over time in national responses to the Eurobarometer questions. These steps were undertaken for each targeted country. At the end of this process, four questions were prepared with the four member state’s responses available for use in answering the research question. There were 16 datasets in all. An example of the construction of the final data is seen in the image below:

	Year	TOTAL	Les choses vont dans la bonne direction	Things are going in the right direction	Les choses vont dans la mauvaise direction	Things are going in the wrong direction	Ni l'un ni l'autre (SPONTANÉ)	Neither the one nor the other (SPONTANEOUS)	NSP	DK
1	2015 Spring	1554	440	0.28	590	0.38	399	0.26	125	0.08
2	2015 Aut	1548	280	0.18	871	0.56	319	0.21	78	0.05
3	2016 Spring	1592	202	0.13	953	0.60	353	0.22	83	0.05
4	2016 Aut	1531	285	0.19	815	0.53	352	0.23	80	0.05
5	2017 Spring	1605	487	0.30	896	0.56	151	0.10	70	0.04
6	2017 Aut	1565	486	0.31	800	0.51	170	0.11	108	0.07
7	2018 Spring	1509	484	0.32	785	0.52	126	0.08	114	0.08
8	2018 Aut	1461	379	0.26	807	0.55	161	0.11	114	0.08
9	2019 Spring	1487	74	0.05	1253	0.84	44	0.03	115	0.08
10	2019 Aut	1540	464	0.30	832	0.54	109	0.07	135	0.09
11	2020 Spring	1514	605	0.40	702	0.46	75	0.05	133	0.09
12	2020 Aut	1575	716	0.45	598	0.38	90	0.06	171	0.11
13	2021 Spring	1535	548	0.36	711	0.46	148	0.10	128	0.08
14	2021 Aut	1604	578	0.36	732	0.46	173	0.11	121	0.07
15	2022 Spring	1507	536	0.36	745	0.49	127	0.08	100	0.07

Figure 4.10- D73a.2- 'At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...?: The EU', arranged by individual country (DE in this case), sequentially.

4.6. Plotting the Eurobarometer Data

An initial exploration of the Eurobarometer data was undertaken to better understand the data before conducting statistical analysis needed to explore the research question. Several interesting insights were gained in conducting this exploration. Simple line plots were used first to gain an understanding of the overall configuration of the data.

Some of the key insights are detailed here, however, the entire range of the plots used has been added to the appendix. Only the most pertinent results are presented in the main body of this work.

Question D71a.2- 'When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...?: European Political Matters'

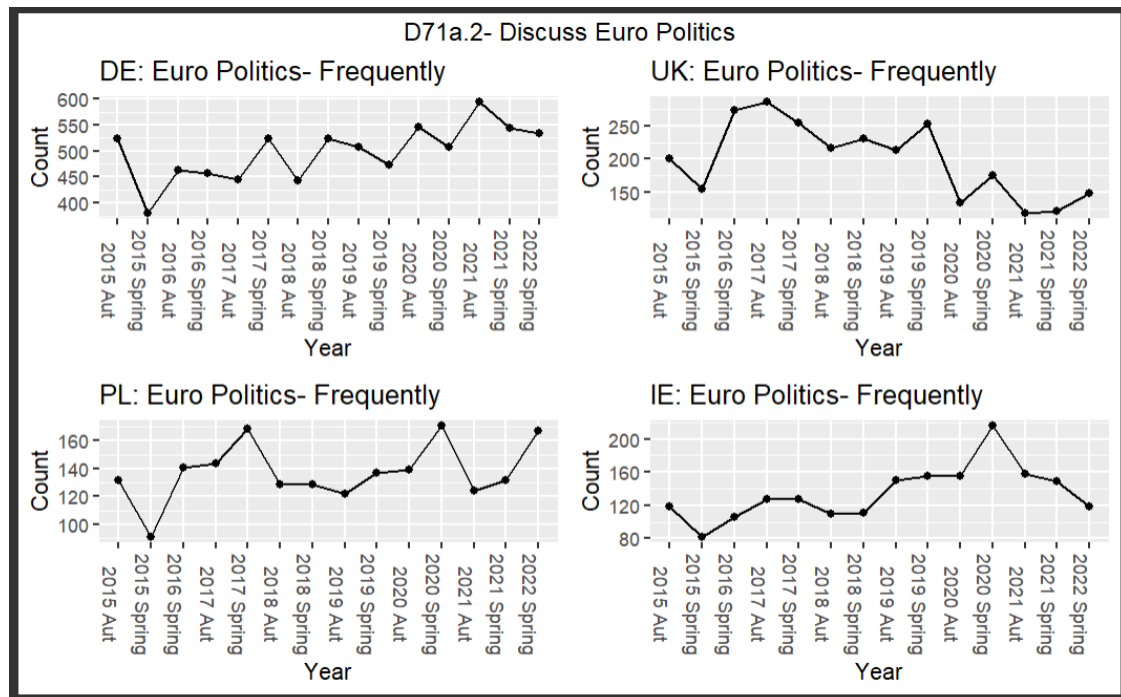


Figure 4.11- Changes over the study period to the question D71a.2

Changes across the four member states was markedly different. Of note was an increase in all four nations in Spring 2020 though of different strength. This is likely to have been due to the COVID-19 lockdown and the changes that was imposed to how citizens spent their time as well as potentially a new focus on global, or at least, extranational issues.

Interestingly, a slight increase across DE, PL and IE can be seen in the study period but this increase is not mirrored in the UK data. In fact, there was a marked *decrease* in the UK data in the same period. The Irish data noticeably has a peak in Spring 2020, mirrored in the PL data also. There is a sudden shift in UK response in 2015 – 2016, potentially due to the endless national discussions in the lead up to Brexit, as well as potentially the impact of directed disinformation campaigns.

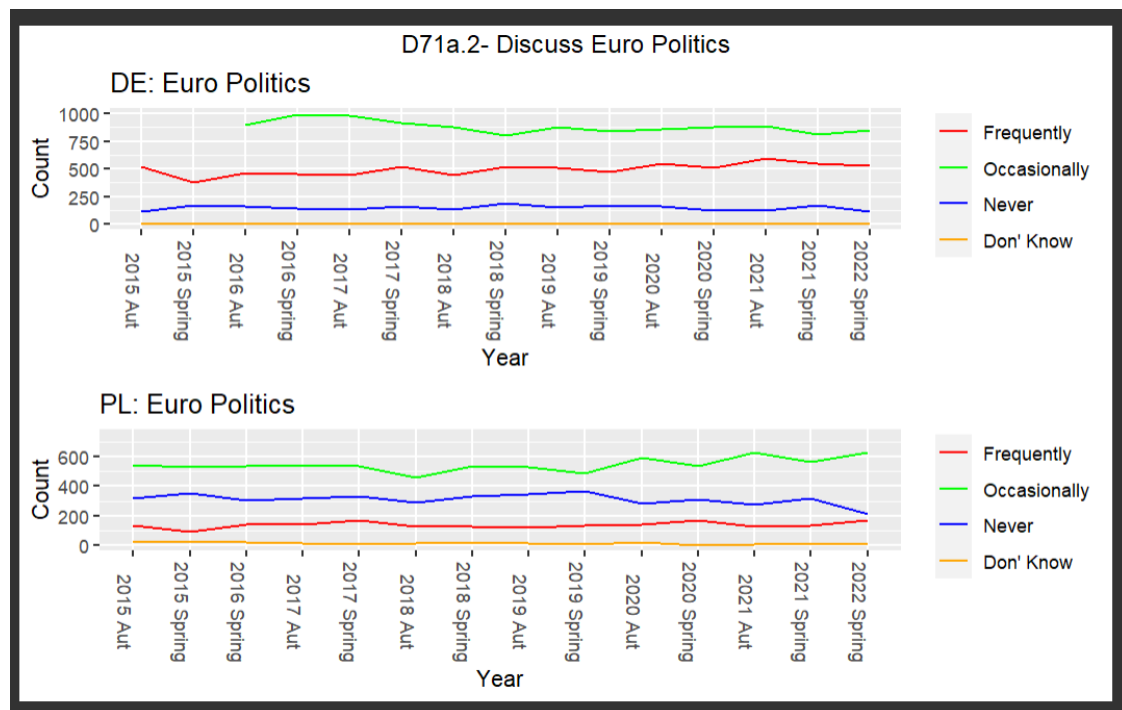


Figure 4.12- DE and PL- Responses to Discussion question

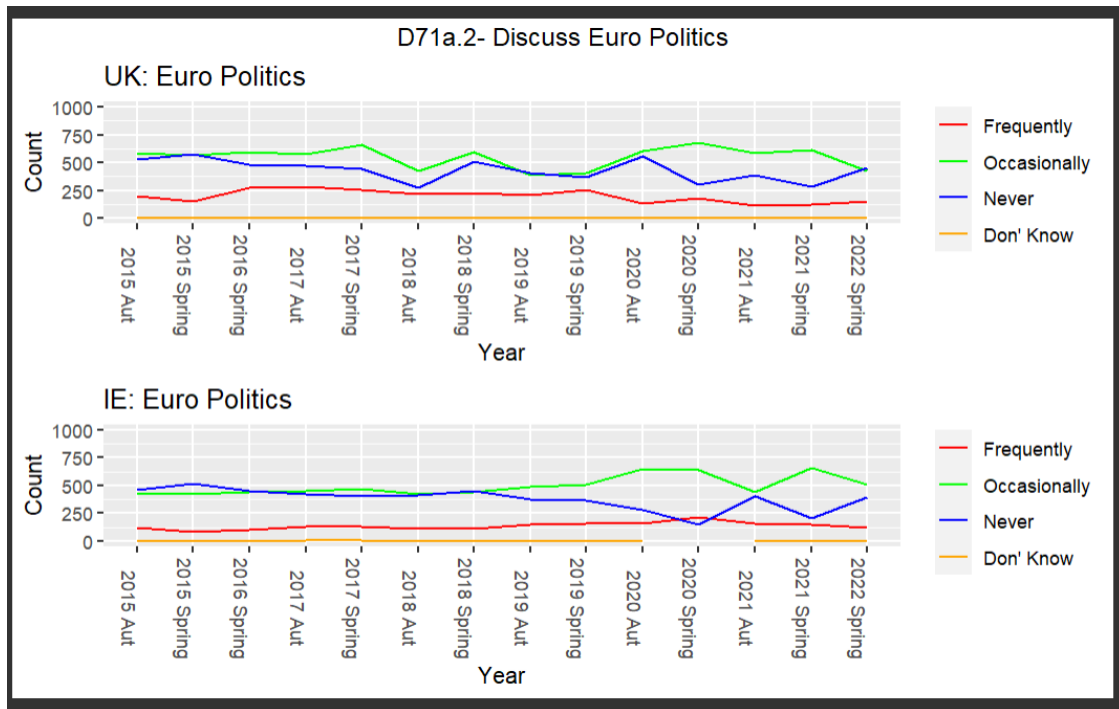


Figure 4.13- UK and IE - Responses to Discussion question

Looking at the full data, it appears that DE and PL remained fairly consistent in their responses over the period of study. DE has by far the most respondents expressing ‘frequently’ in answer to this question. IE and UK have had more changes in their respective response over the period, though those changes do not mirror one another.

Question D73a.2- 'At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...?: THE EU'

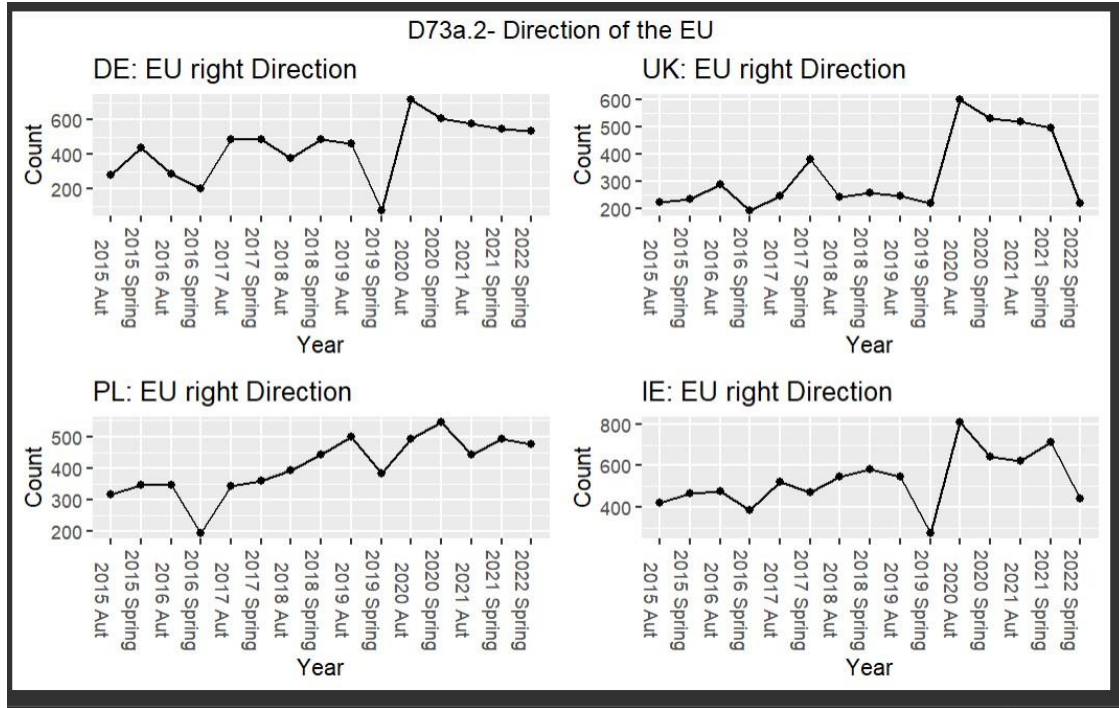


Figure 4.14- Changes over the study period to the question D73a.2

Question D73a.2 displays a fairly consistent change in the nations over the period with some distinct differences at points. Overall DE, PL and IE display similar changes throughout the study period, including a dip in 2016 during the year of the Brexit negotiations. However overall, there was a tendency for this answer to increase (as in PL most obviously) over time. There was however, a significant drop in 2019, assumed to be the early warnings of the pandemic, followed by a significant jump in satisfaction across all four nations in 2020.

The UK data is noticeably different to the other nations displayed though UK and IE display the most similarity as one another overall. The UK displays the lowest responses to the EU moving in the right direction throughout the study period. Interestingly there is a sharp drop in “Right Direction” in 2022 in both the UK and IE. This change is not mirrored in the other nations. Currently this is unexplained within the limits of this data. Looking at the entirety of the data, we are presented with a confusing image that seems to capture the disruption of the study years in great detail:

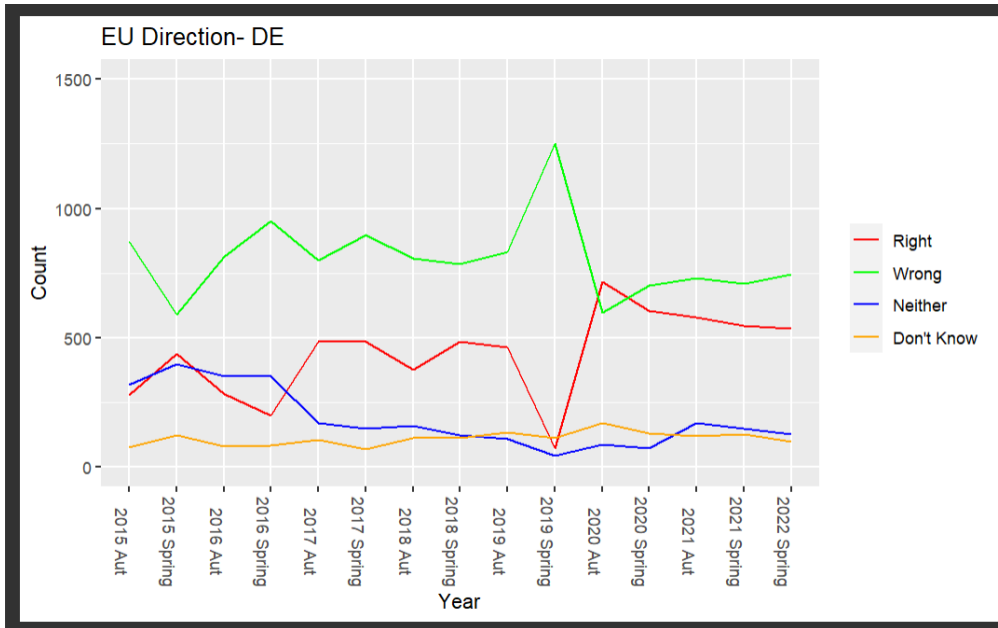


Figure 4.15- DE: Responses to the Right Direction answers

The DE data is the least changed over time, bar the Spring 2019 and Autumn 2020 data, this is presumed to capture the change in attitude as the pandemic was firmly on the horizon in 2019 and when the potential disaster was somewhat better managed in 2020.

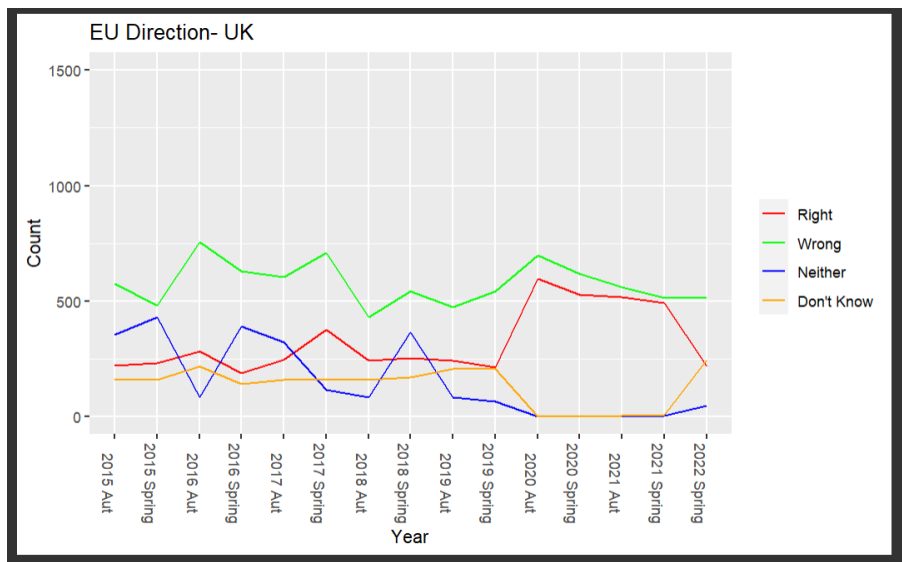


Figure 4.16- UK: Responses to the Right Direction answers

The UK is also seen to follow a sudden change in 2019, however this change is in the opposite direction to the DE data. Rather than thinking the EU was moving in the wrong direction, UK respondents felt that it was moving in the right direction at

that point. There was also an increase in ‘wrong’ direction at that point but nowhere near as pronounced as ‘right’ direction.

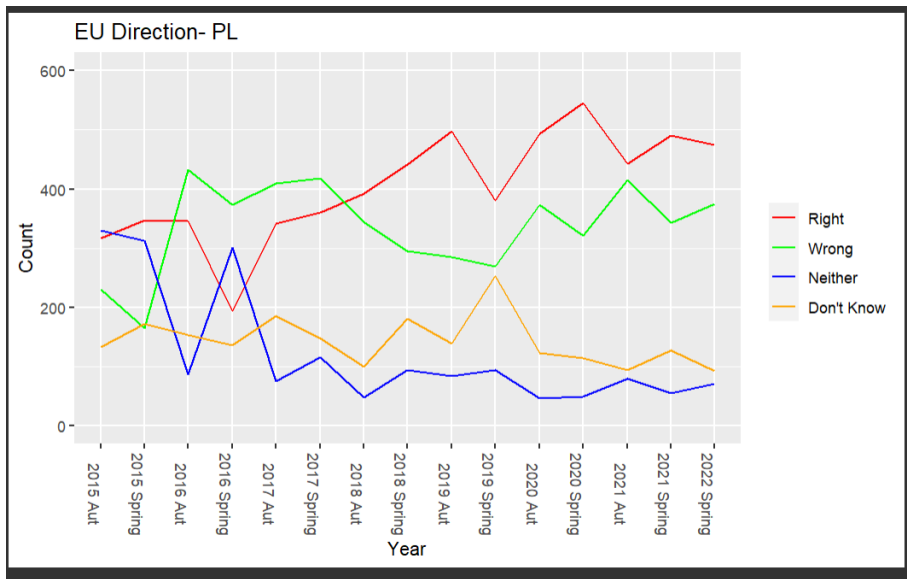


Figure 4.17- PL: Responses to the Right Direction answers

The Polish data, while following some of the pattern in the DE and UK data, remains much more confused. There seems to be a severe dip in ‘right direction’ in 2016, this is assumed to be a reaction to the Brexit referendum. However, that change is quickly replaced with strong and continuing support for ‘right direction’, a similar dip takes place in 2019 as in the other years. The PL data is unexpectedly volatile.

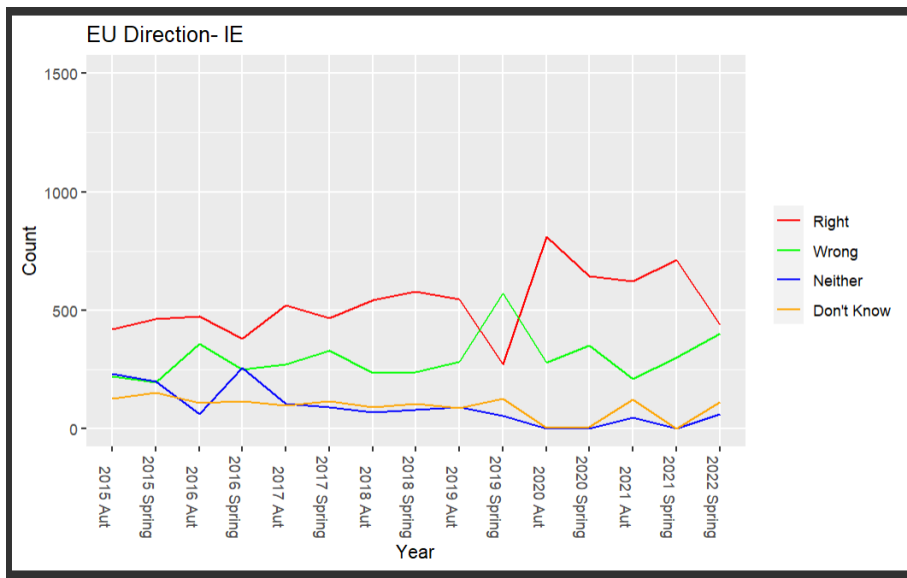


Figure 4.18- IE: Responses to the Right Direction answers

The Irish data follows the already established pattern with one key difference, the Irish respondents consistently rate the EU as moving in the ‘right direction’ bar the

disruptive period of 2019 as seen in the other data. There is the same drop in 2016 in line with Brexit but is much less pronounced than other nations.

Question QA8- 'I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it: The EU' .

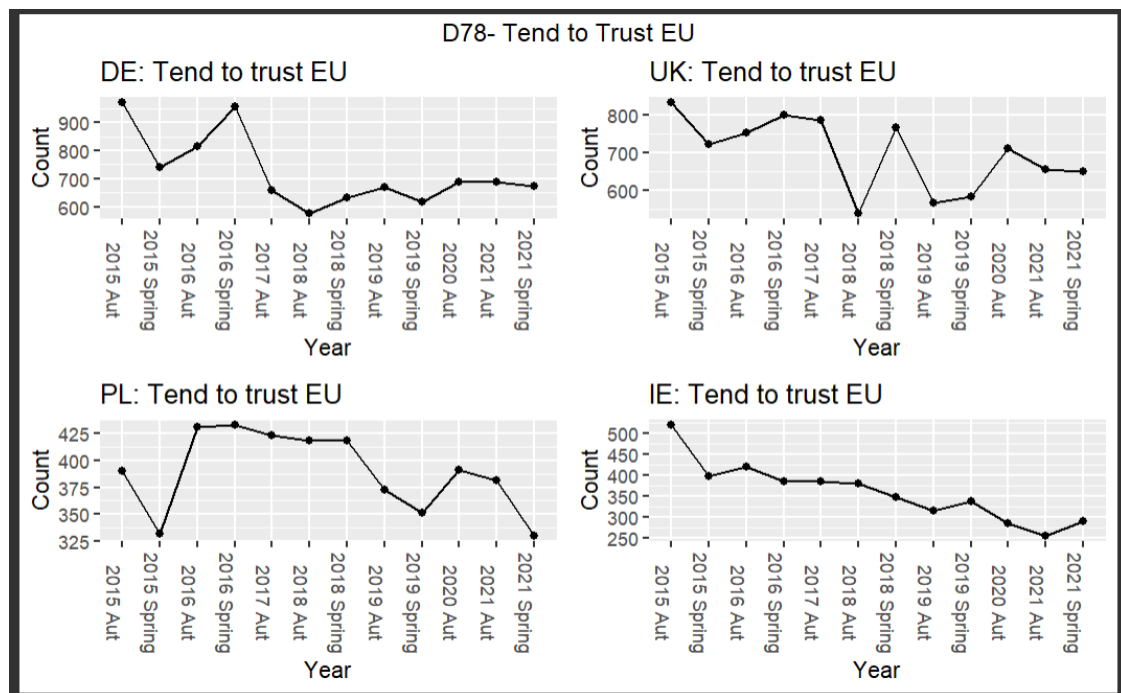


Figure 4.19- QA8- Trust the EU, 'tend to trust' across all 4 member states.

Question QA8 presents interesting results over the study period. Firstly, a general and slow decline in 'tend to trust' can be seen across all four member states. While there was a sudden and significant increase in DE and PL in 2016, coinciding with Brexit, that increase declines over time, significantly so in the case of DE. A similar decrease over time can be seen in Ireland but it is more gradual. The UK data is markedly different than the other nations with peaks and troughs during the study period. Noticeably there is a severe decrease in 2018 that is not seen in the other four member states. IE is the only nation seen to have a mild increase in the final year of data compared to DE and UK who have a mild decrease and PL which has a steep decline.

Individual plots of this question were not particularly illuminating not did they diverge sufficiently from the data demonstrated above, they have been included in the appendix.

Question D78: 'In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?'

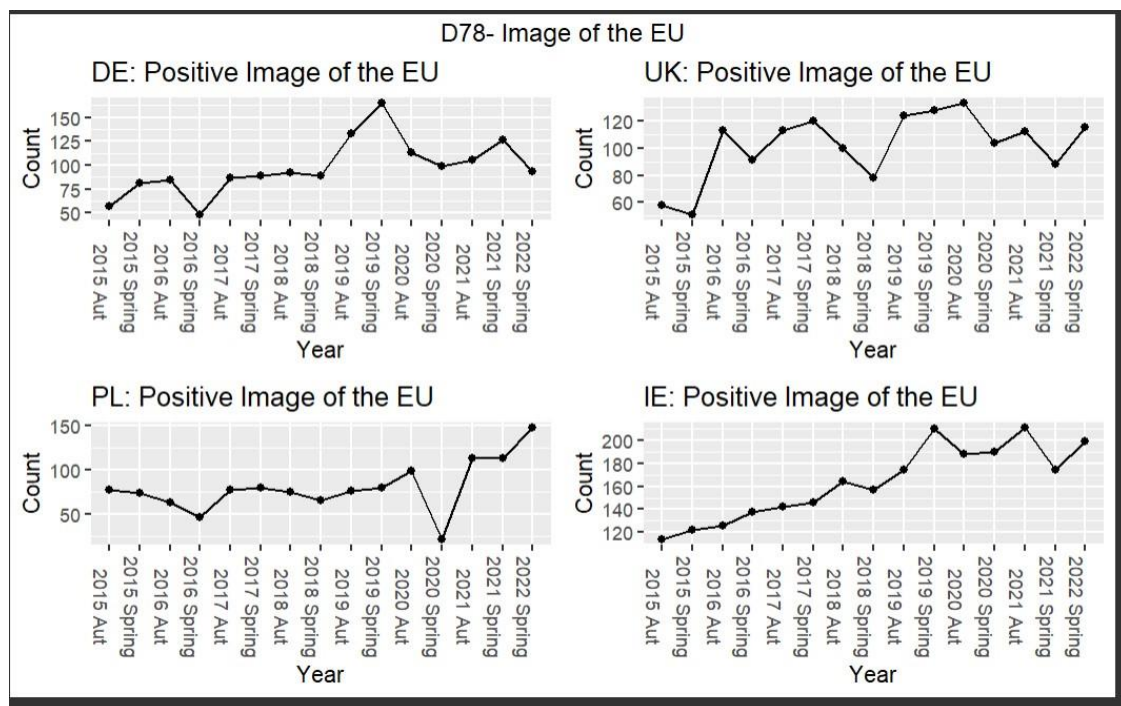


Figure 4.20- D78- Image of the EU- Positive

The responses to question D78, in general, follows a similar pattern across all four member states with some obvious departures. All four display a gradual increase overall in the positive image of the EU but with different peaks and troughs. The year 2016 was again a banner year for change in that DE, PL and UK all had noticeable decreases in their positive image of the EU, IE on the other hand had an increase. PL and DE diverge significantly in their responses in 2019-2020 with DE having a marked increase followed by a return to the slow increase seen before 2019. PL had a marked decrease followed by a return to the previous pattern.

The UK had much more volatile data overall with no clear increase or decrease seen across the full range of data. There was a noticeable increase in 2016 following Brexit and a similar slow pattern of increase in other years but this is counterbalanced

by a significant drop in 2018 and further drops in later years. IE, on the other hand, has had a fairly consistent increase overall with a noticeable peak in 2019. This is somewhat unexpected considering that all previous data points in other questions had displayed a turning against positive trust or positive image towards the EU in 2019. However, a general and consistent increase can be seen in the data overall.

With this exploration of the Eurobarometer data completed there are some noteworthy patterns:

- 1- 2015/2016 tended to be a year of significant changes in the data, aligning with the Brexit referendum in the UK.
- 2- 2019/2020 also tended to be a year of significant changes in the data, aligning with the start of Covid and the first summer of lockdown in 2020.
- 3- DE, PL and IE tended to follow one another's patterns fairly consistently if perhaps not in the same Eurobarometer year.
- 4- The UK tends to be an outlier in its responses with consistent differences in its answers overall.

4.7. Exploratory Statistics

Having completed the wrangling and exploration of the Eurobarometer dataset, exploratory statistics were next undertaken. Again, with four countries data, and four questions per country overall, a significant amount of plotting was necessary. Different approaches were used at different stages in this process, and ultimately, it was found that simple density plots capture the data in the clearest manner. Overlaid density plots were also constructed but these tended to confuse rather than amplify the understanding of the data, they are included in the appendix but will not be discussed in the main body of this work. A selection of the key plots that highlight interesting themes will be explained here with the remaining plots added to the appendix.

At this stage of the process, we have combined responses from member states that were the most targeted by disinformation based on the EUvsDisinformation data in both the Scraped and Kaggle EUvsDisinformation data. Those three nations (DE, UK, PL) are joined by Ireland (IE) as a baseline comparison since it does not appear to have been directly targeted based on the disinformation data. Statistical exploration is

undertaken in this section as a basis for the statistical testing that follows. The data was plotted and tested in its entirety but only the key plots are presented here, the remainder are contained in Appendix A. The questions are demonstrated in order of their statistical robustness.

The Most Robust Data

Question QA8 (below) was generally relative normal data and presented data that can be considered on a solid foundation for later statistical tests.

QA8: ‘I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it. The EU’

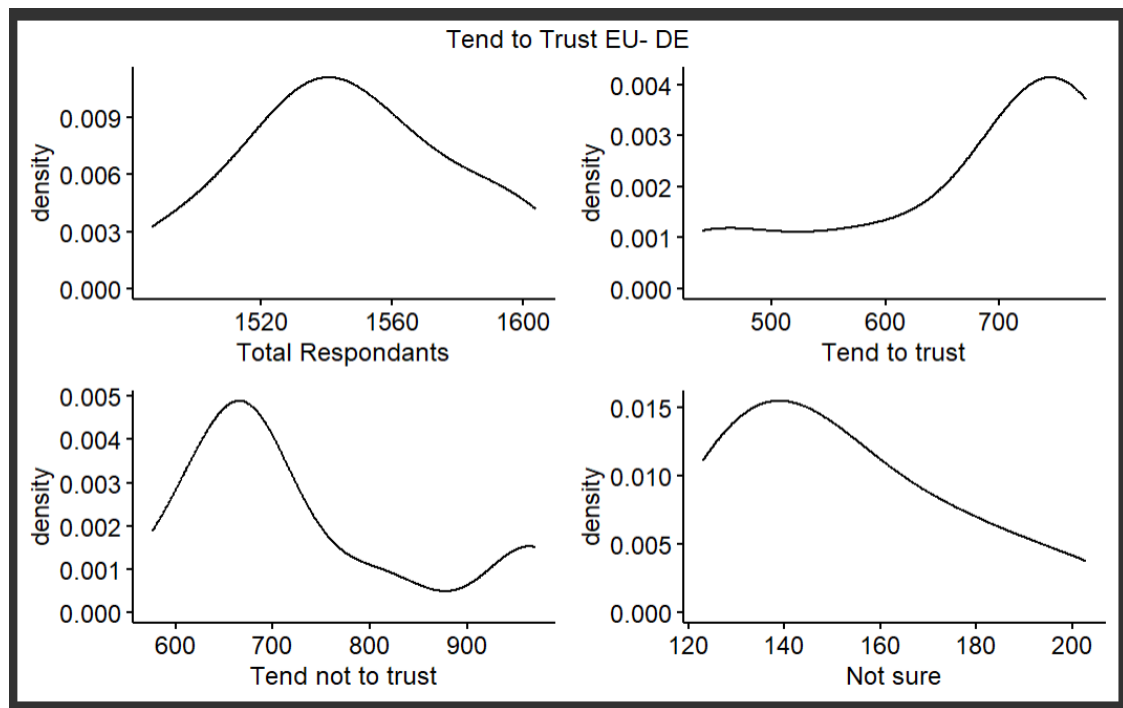


Figure 4.21- Density plots for DE responses to QA8- Tend to Trust EU

The data overall follows a relatively normal structure. In general, there are no obvious unexpected peaks in the charts, however, the “Tend to not trust’ has as second peak at a very low density of circa 0.002. in the responses for ‘Tend to trust’ and ‘Not sure’ both have their tails abruptly cut-off and will therefore be reported as skewed.

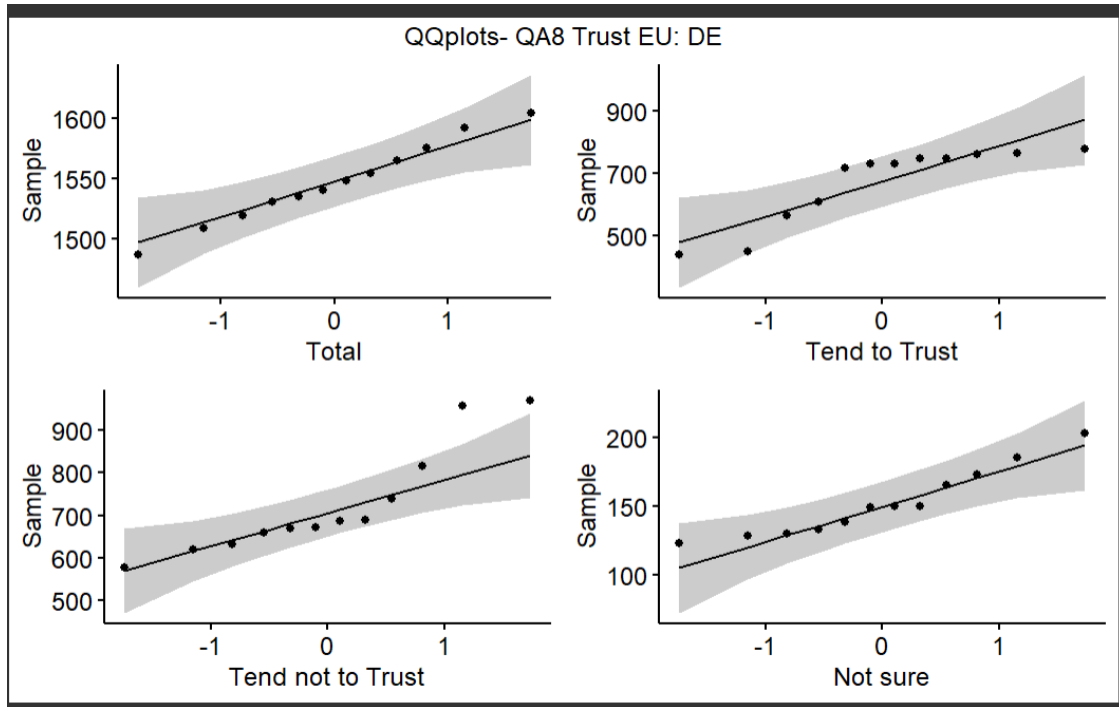


Figure 4.22- QQ plots for DE responses to QA8- Tend to Trust EU

The Q-Q plots strengthens the conclusions based on the density plots. There are some obvious outliers in ‘Tend not to Trust’ but, other than that, the data remains well within in expected limits.

```

"""(r Shapiro test)

shapiro.test(stats_QA8_TRUST_DE_wide$`Tend to Trust`)
# p-value 0.005- Non- Normal distribution

shapiro.test(stats_QA8_TRUST_DE_wide$`Tend not to Trust`)
# p-value 0.02- Non-Normal distribution

shapiro.test(stats_QA8_TRUST_DE_wide$TOTAL)
# p-value 0.99- Normal distribution

shapiro.test(stats_QA8_TRUST_DE_wide$`Not Sure`)
# p-value 0.30- Normal distribution
"""

```

Figure 4.23- Shapiro-Wilks normality test for QA8- DE

Looking at Shapiro-Wilks, the data has some issues. We can see that ‘Tend to Trust’ and ‘Tend not to Trust’ are both non-normal. This is an expected result for ‘Tend not to Trust’ given the earlier plots but is less expected for ‘Tend to Trust’. Any statistical test completed on this data will have to be balanced by these results.

```

# (r Skewness)
skewness(stats_QA8_TRUST_DE_wide$`Tend to Trust`)
# Left skewed
skewness(stats_QA8_TRUST_DE_wide$`Tend not to Trust`)
# Right skewed
skewness(stats_QA8_TRUST_DE_wide$TOTAL)
# Slightly right skewed
skewness(stats_QA8_TRUST_DE_wide$`Not Sure`)
# Slightly right skewed

[1] -1.025737
[1] 1.050022
[1] 0.05160774
[1] 0.7235501

```

Figure 4-24- Skewness test for QA8- DE

As expected, ‘Tend to Trust’ is left skewed and ‘Tend not to Trust’ is right skewed, as seen in the original density plots above. While the other responses are much less offset.

Overall, this data appears robust for Germany, if slightly less than normal overall. The data appears to be of sufficient normalcy and within expected parameters to be useful for later statistical testing.

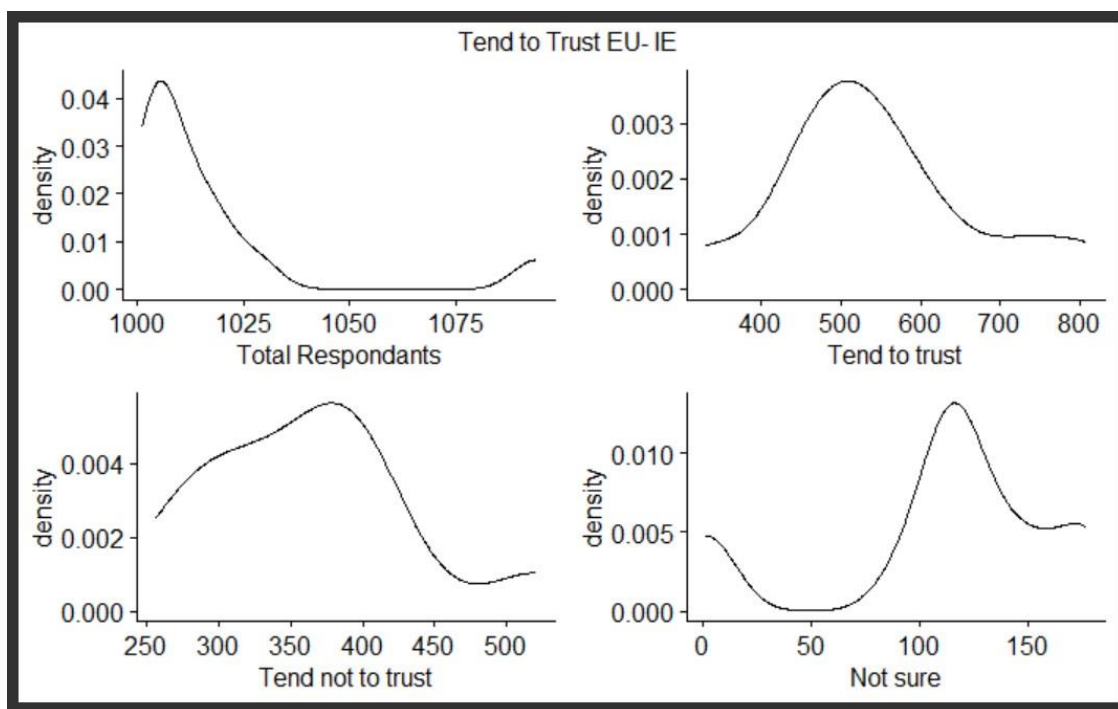


Figure 4.25- Density plots for IE responses to QA8- Tend to Trust EU

The Ireland (IE) responses to the same question were also relatively normal. Certainly ‘Tend not to Trust’ and ‘Tend to Trust’ appeared to display signs of normalcy. The only clear outlier based on the density plot was ‘Not sure’ that seemed to display an unusually high density at 0.010, as well as a bimodal structure.

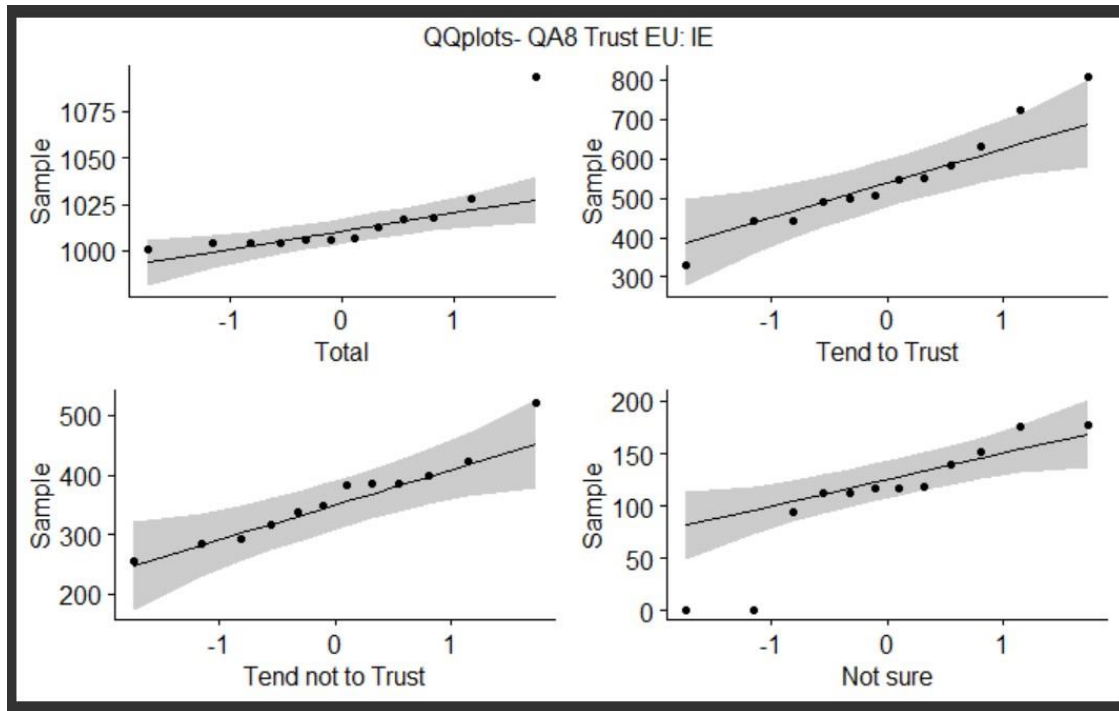


Figure 4.26- QQ plots for IE responses to QA8- Tend to Trust EU

Unexpectedly however, we can see in the Q-Q plot for total respondents an extreme outlier at over 1075. However, at this scale, that may not be as much of an outlier as it seems as it is perhaps only 50 more than the upper limit of the range of standard deviation. Further outliers, or near outliers can see seen in all 3 other plots. Overall, this data can be seen to be fairly robust but not as obviously sound as the DE data. Looking at the Shapiro Wilks test:

```

'''(r Shapiro test)

shapiro.test(stats_QA8_TRUST_IE_wide$`Tend to Trust`)
# p-value = 0.7777- Normal distribution
shapiro.test(stats_QA8_TRUST_IE_wide$`Tend not to Trust`)
# p-value = 0.6458- Normal distribution
shapiro.test(stats_QA8_TRUST_IE_wide$TOTAL)
# p-value = 8.48e-05- Not Normal distribution
shapiro.test(stats_QA8_TRUST_IE_wide$`Not Sure`)
# p-value = 0.02589- Not Normal distribution

'''

```

Figure 4.27- Shapiro Wilks test for IE responses to QA8- Tent to Trust EU

In this test we can see that ‘Tend to Trust’ and ‘Tend not to Trust’ are comfortably normal but that ‘Total respondents’ is extremely non-normal. This stat is included more to understand the data and is not used in later statistical test but is a surprising result, nonetheless. ‘Not sure’ is also seen to not be normal at p-value: 0.02, in this case only ‘Tend to Trust’ and ‘Tend not to Trust’ have the most statistical foundation for later tests.

Similar results were seen in both the UK and PL data. While there are some differences, overall, the data for question QA8 can be seen to be normally distributed and of sufficient robustness to be useful in answering the research question.

Data with Mixed Robustness

D71: ‘When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...? The EU’

The data across DE & UK for question D71 tended to be relatively normal, using UK as an example:

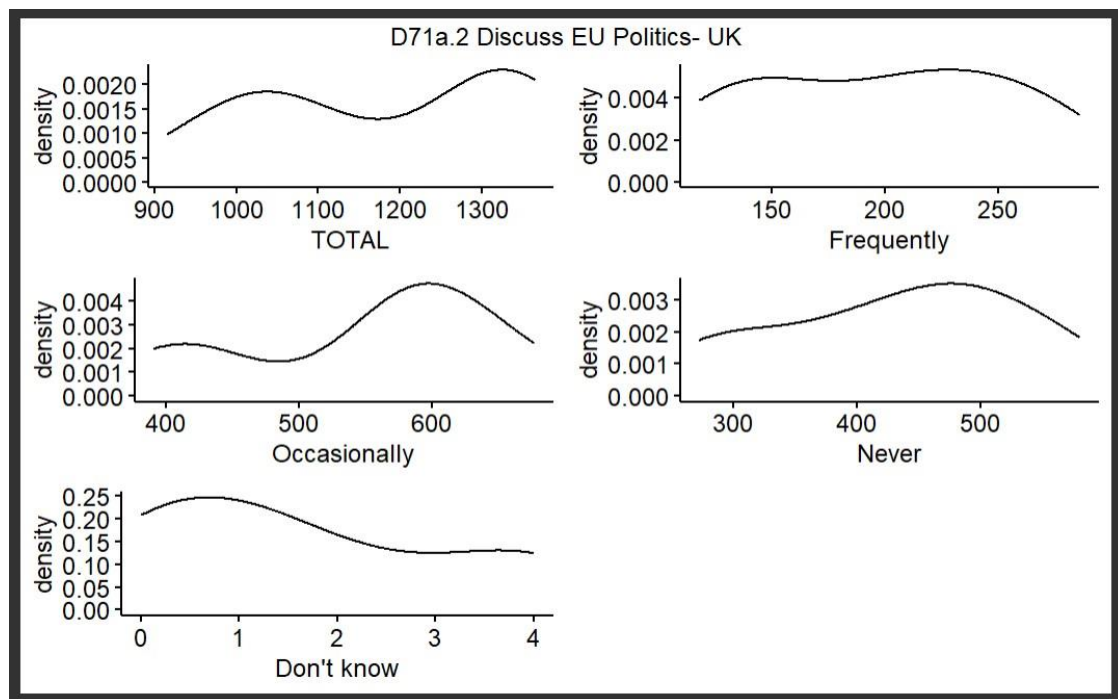


Figure 4.28- Density plots for UK to D71a.2 Discussion EU politics

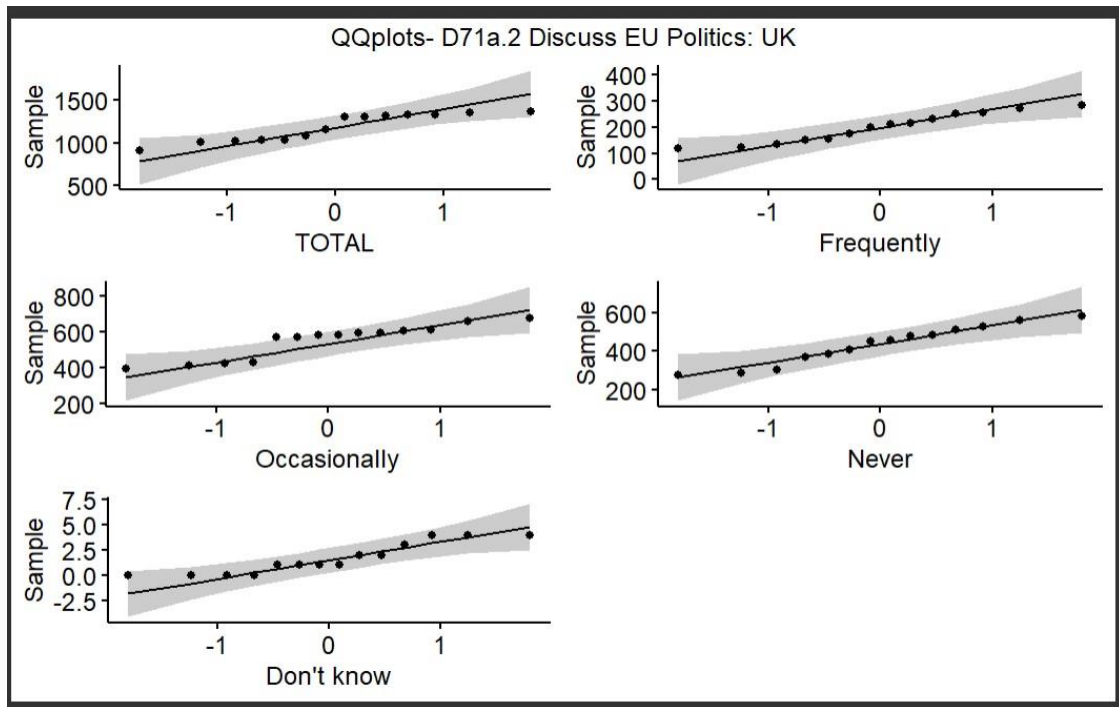


Figure 4.29 Q-Q plots for UK to D71a.2 Discussion EU politics

The data was well balanced without any major issues. ‘Total’ is bimodal and ‘Occasionally’ is trending in the same direction. Overall, the data can be seen to be smooth and without any evidence of extreme outliers. The other nations tended towards the same though there were some exceptions.

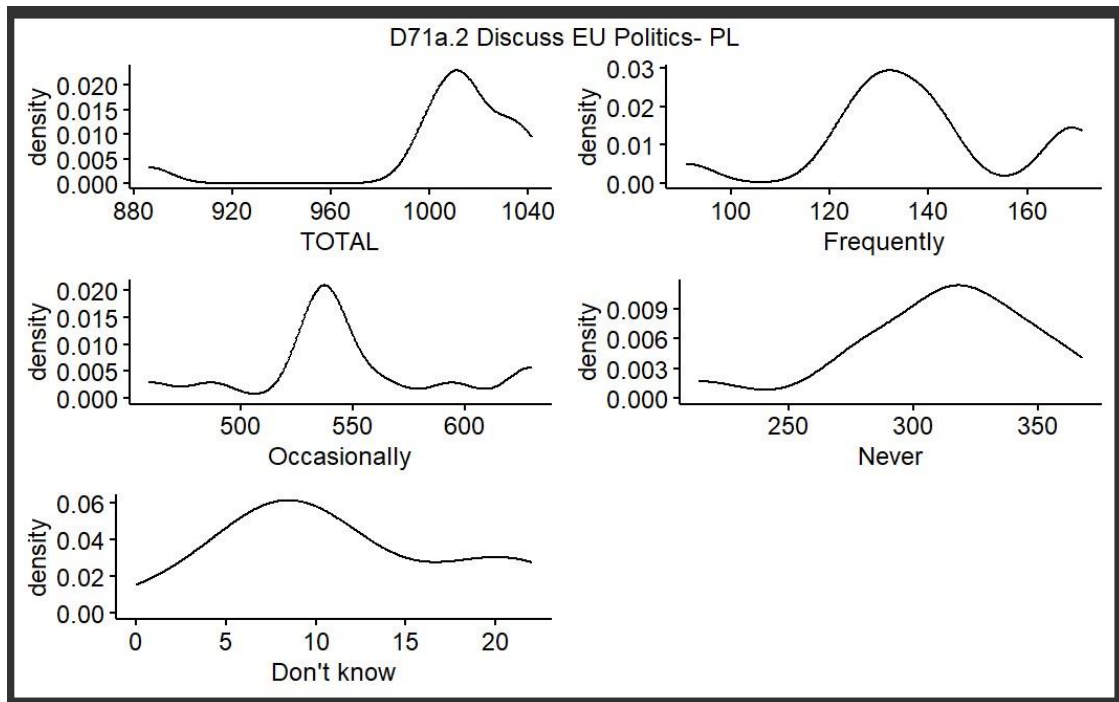


Figure 4.30- Q-Q plots for PL to D71a.2 Discussion EU politics

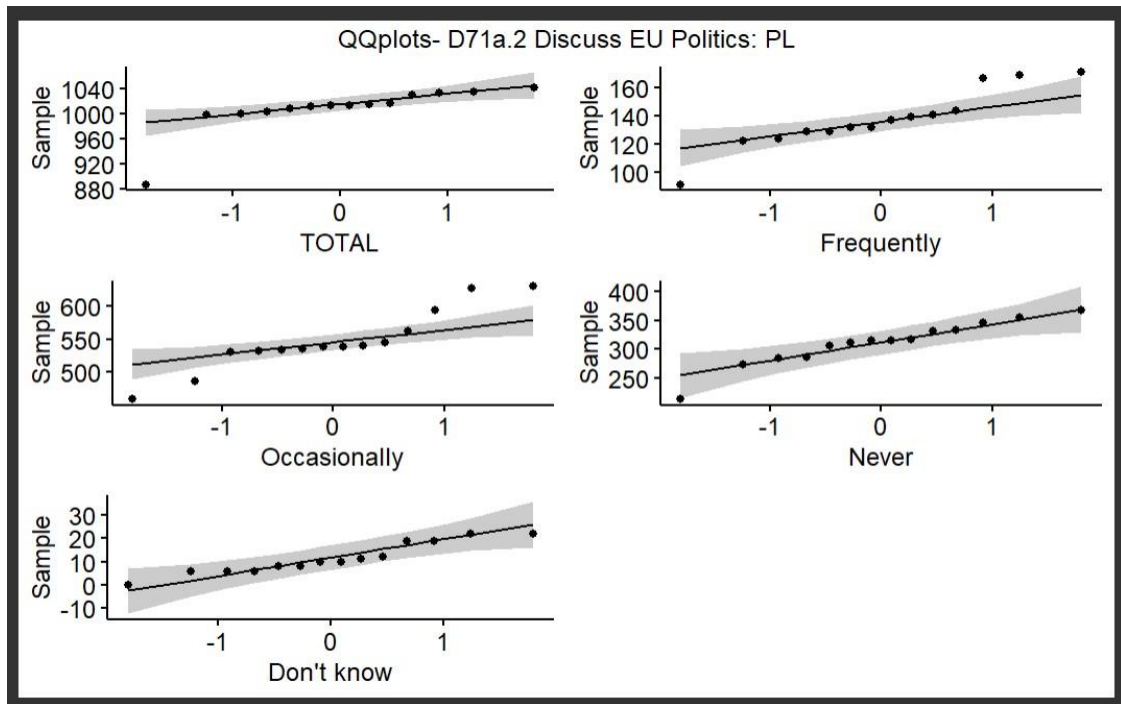


Figure 4.31- Q-Q plots for PL to D71a.2 Discussion EU politics

The Poland (PL) data had evidence of outliers in the density plots that was confirmed with the Q-Q plots. ‘Total’ can be seen to have one very low-level outlier, while ‘Occasionally’ and ‘Frequently’ also have outliers. The IE data similarly had extreme outliers.

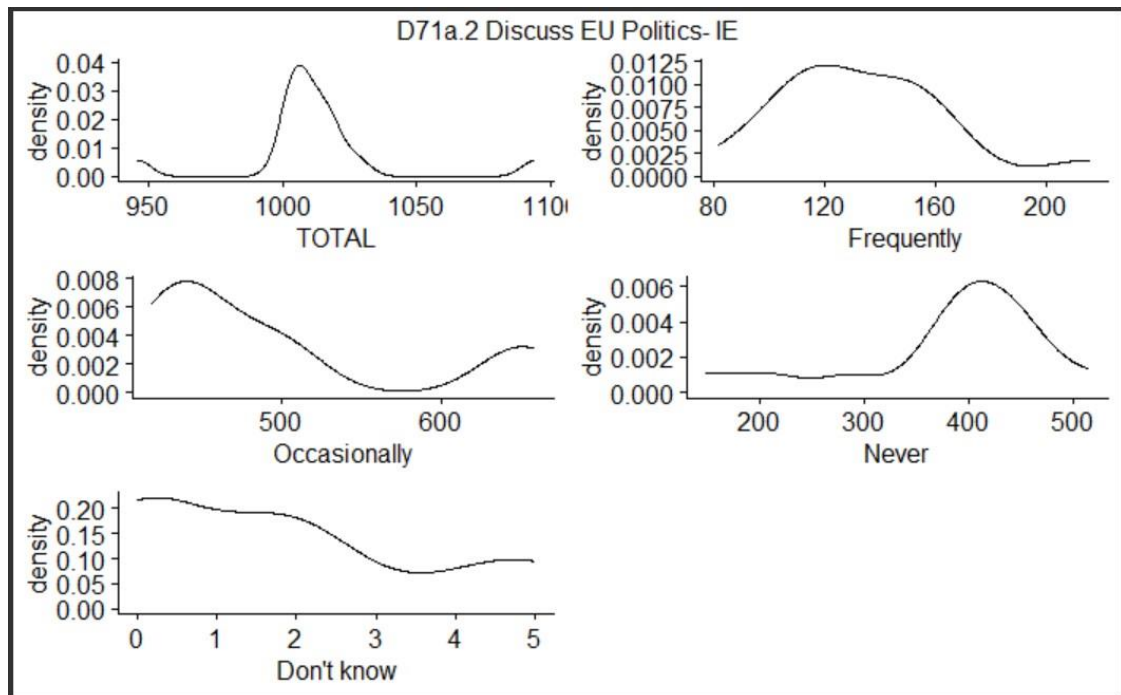


Figure 4.32- Density plots for IE to D71a.2 Discussion EU politics

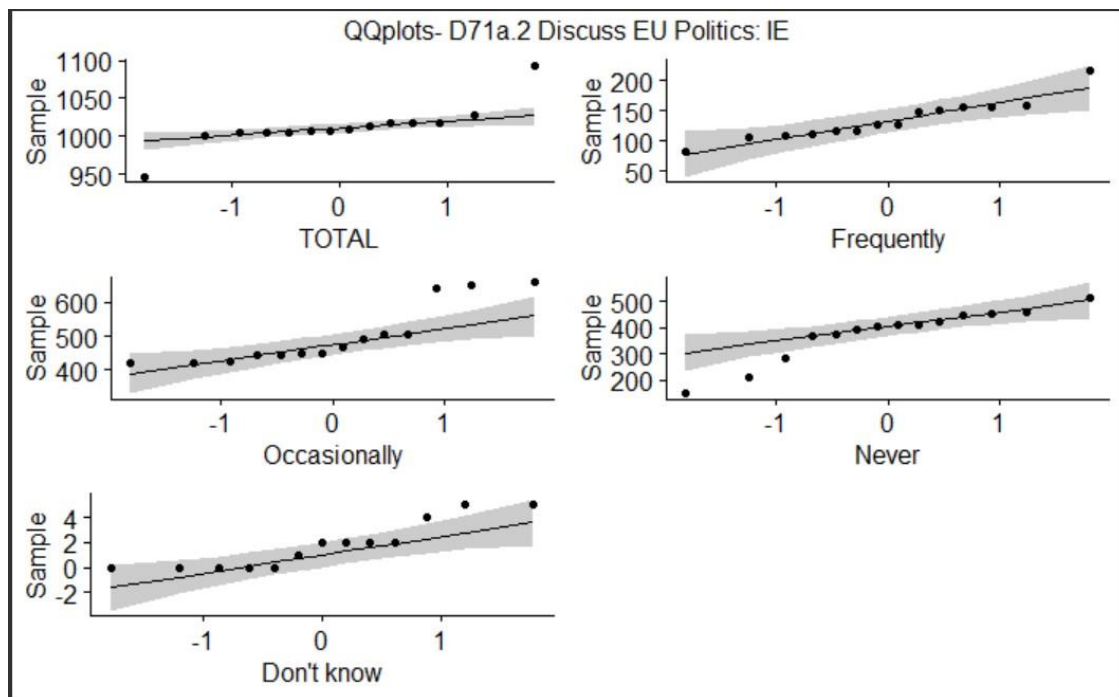


Figure 4.33- Q-Q plots for IE to D71a.2 Discussion EU politics

Outliers could be seen in ‘Total’, ‘Occasionally’ and ‘Never’. In the case of ‘Total’, the outliers straddled both sides of the Q-Q plot seriously calling into question the robustness of that data. This question was quite robust overall but with some outliers demonstrated in PL and IE data in particular. It was determined that the data was of sufficient utility to be included in the research question based on the statistical analysis but that any statistical testing based upon this question would not on as solid a foundation as QA8.

Unfortunately question D73 was in a similar position as D71. Two of the four member states had data that was quite normal and demonstrated few outliers that would impact the statistical testing. However, two of the member states again demonstrated data that was insufficiently robust to be considered on a very solid basis for later testing.

D73: At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...? The EU?

The UK data is used here to display the overall structure of both the UK and DE data, both were found to be similarly arranged and shaped.

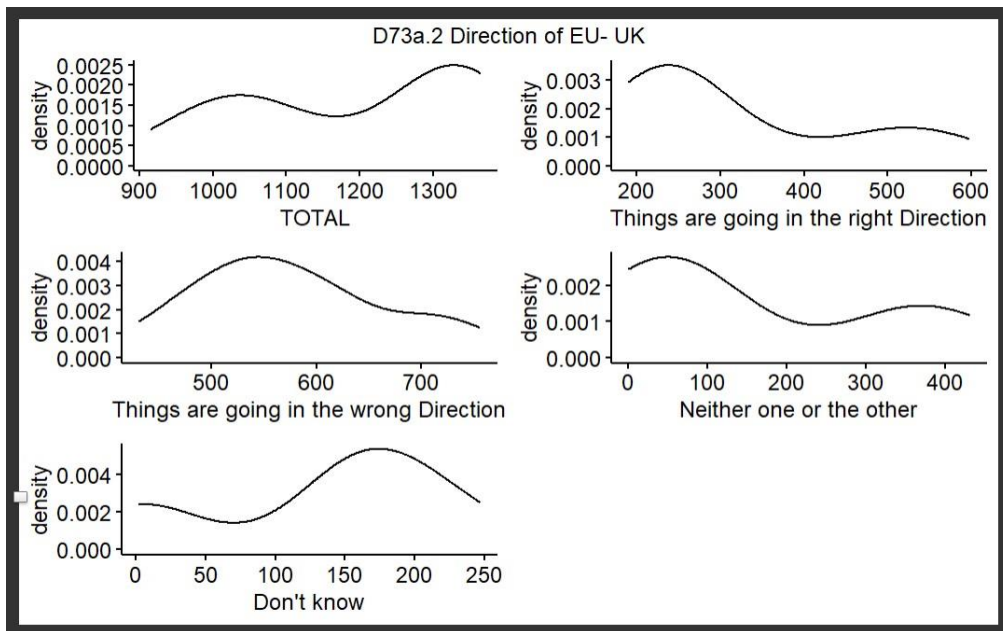


Figure 4.34- Density plots for UK to D73- Direction of the EU

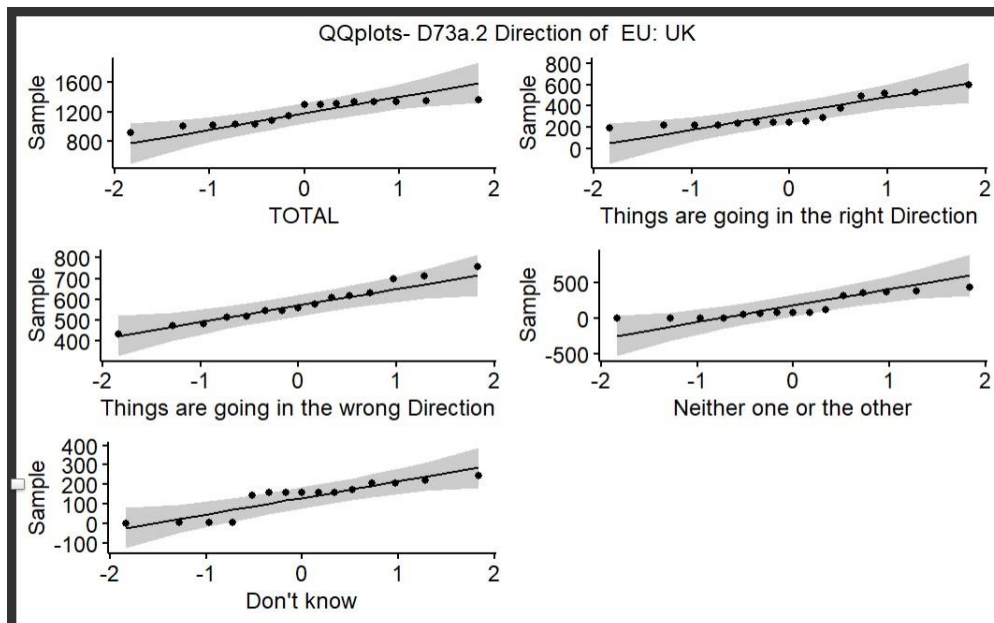


Figure 4.35- Q-Q plots for UK to D73- Direction of the EU

The data for both the UK and DE appear normal with few outliers. This is further confirmed when looking at the Q-Q plots. The data in both cases can be said to be of strong robustness. However, in looking at the IE and PL data, the level of confidence decreases:

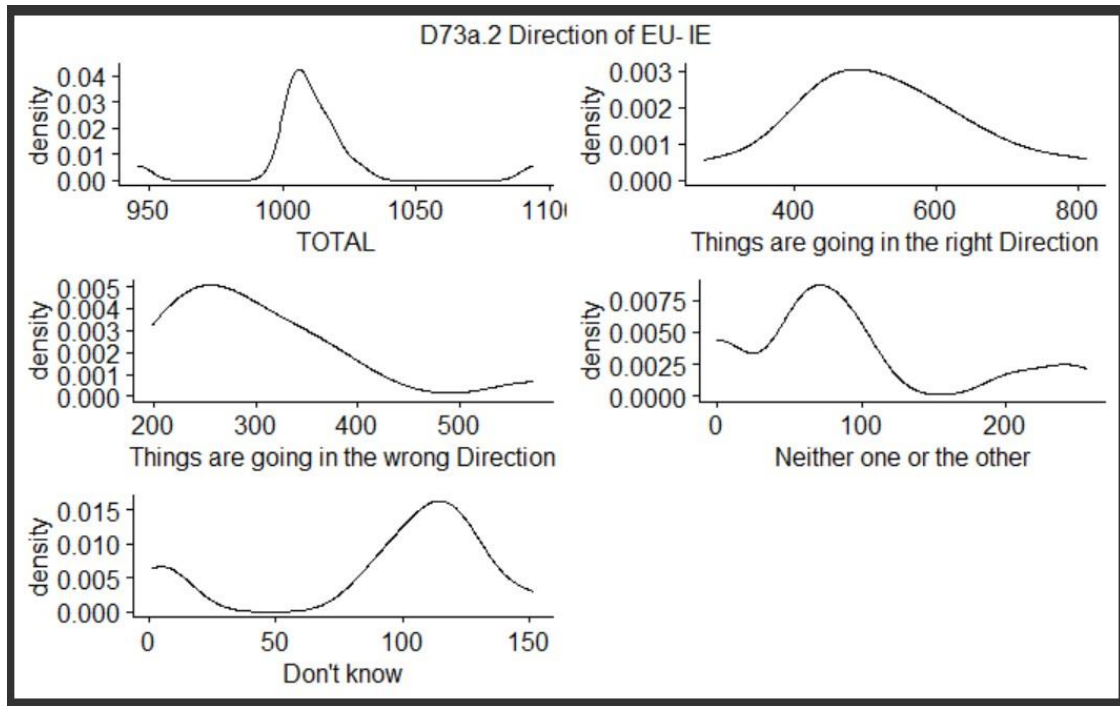


Figure 4.36- Density plots for IE to D73- Direction of the EU

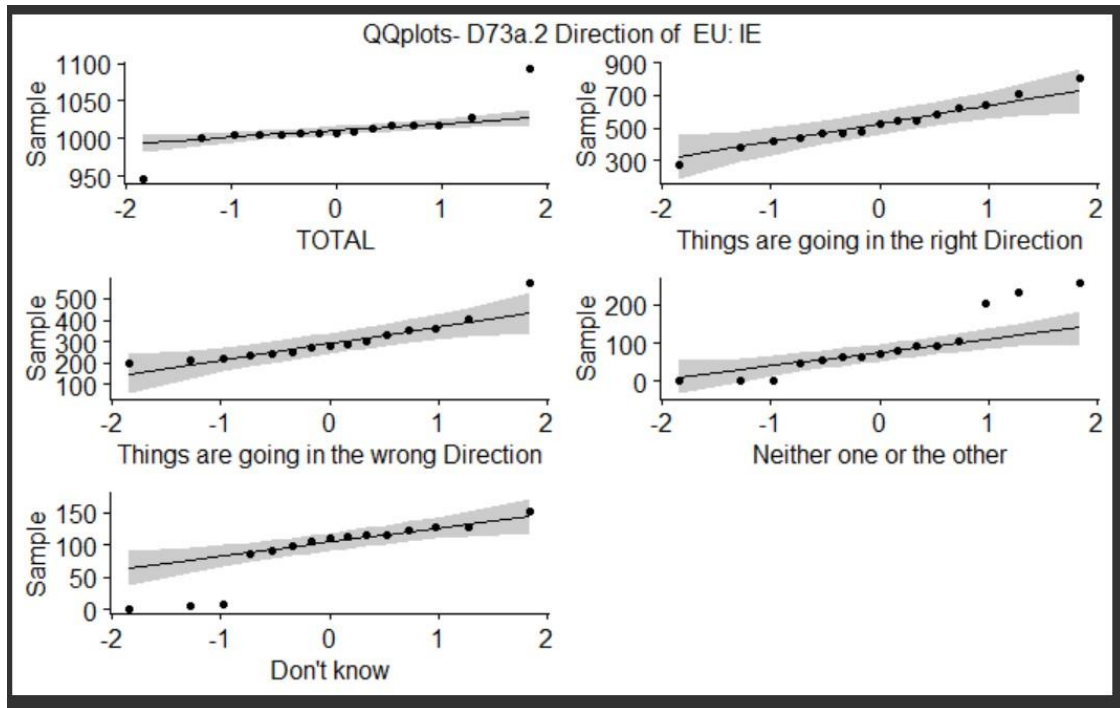


Figure 4.37- Q-Q plots for IE to D73- Direction of the EU

IE data serves here as an example of both IE and PL robustness. Once again, we can see unexpected outliers that influence the confidence of the data. In the Q-Q plot above, we can see that ‘Total’, ‘Don’t know’, ‘Things are going in the wrong direction’, and ‘Neither one or the other’; all have outliers either above or below the expected values. Given that the question is only five plots in total, four of them having extreme outliers is concerning in terms of their potential reliability to address the research question.

D73 was judged to have mixed robustness overall. The number of outliers across half the availability member state’s data could not be ignored.

Less Robust Data

D78: ‘In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?’

Question D78 proved more complex than QA8 with data of less robustness overall and increased complexity. The first challenge was that there are many more potential responses to the question thereby increasing the chance that some of the data would have sufficient issues to warrant its exclusion from the statistical tests used to answer the research question. While the statistical exploration demonstrated normal data overall for IE, PL and UK, the DE data was a much less solid foundation for later statistical tests. Using IE as an example of the more standardized data of PL and UK:

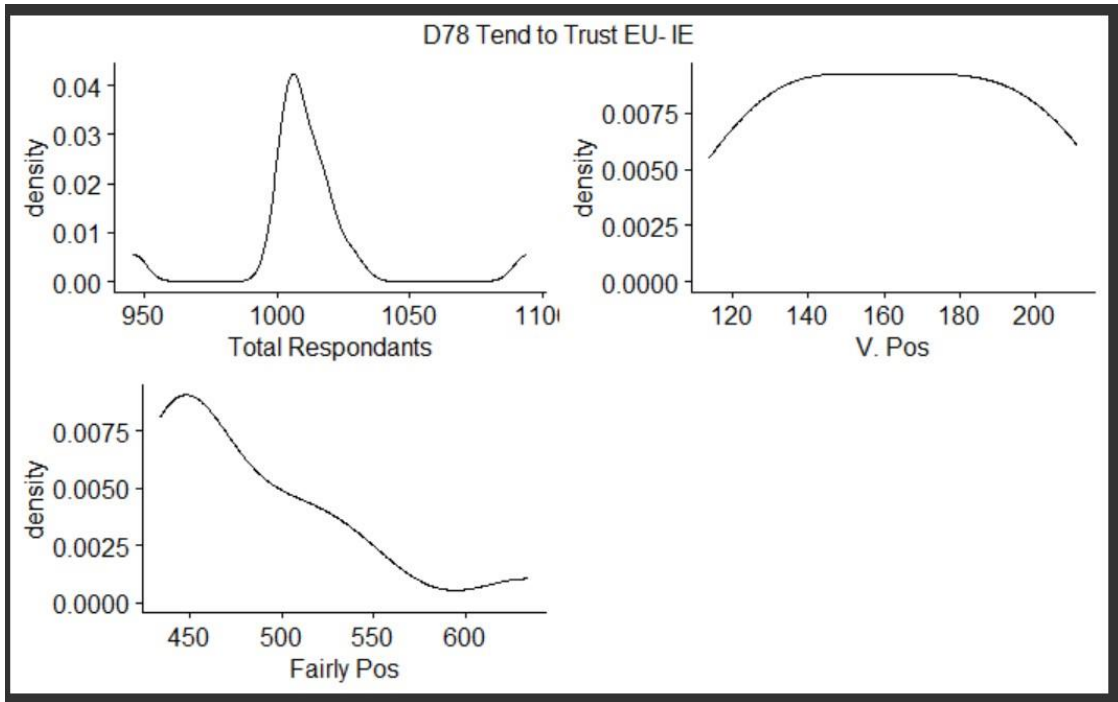


Figure 4.38- Density plots for IE first 3 responses to D78- EU Image

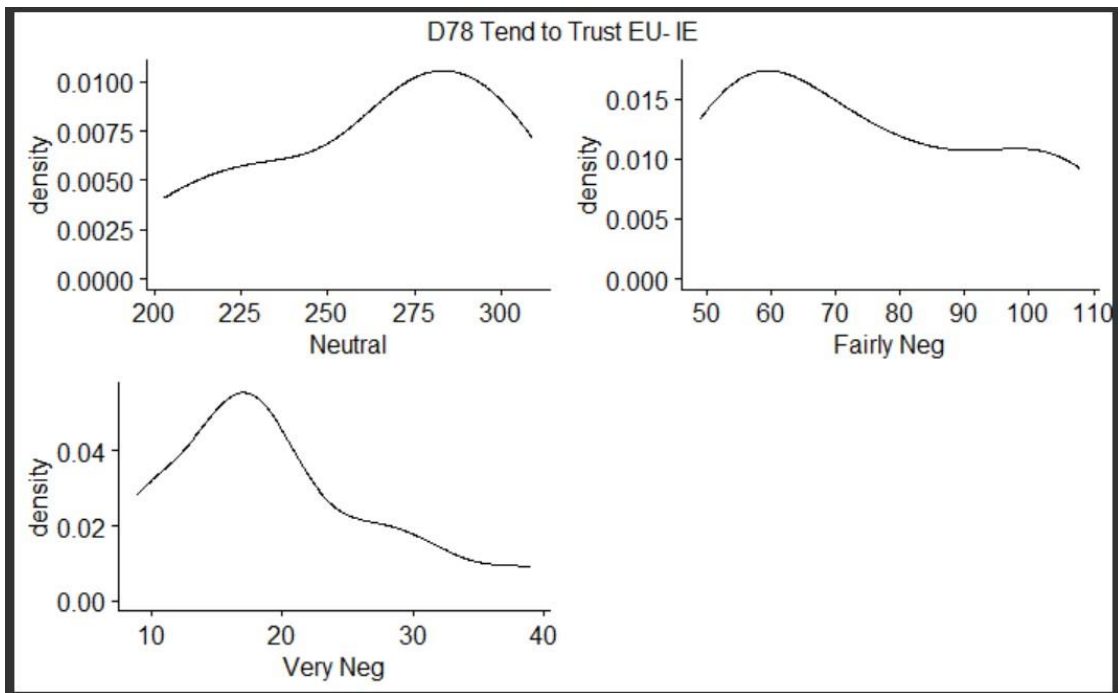


Figure 4.39- Density plots for IE second 3 responses to D78- EU Image

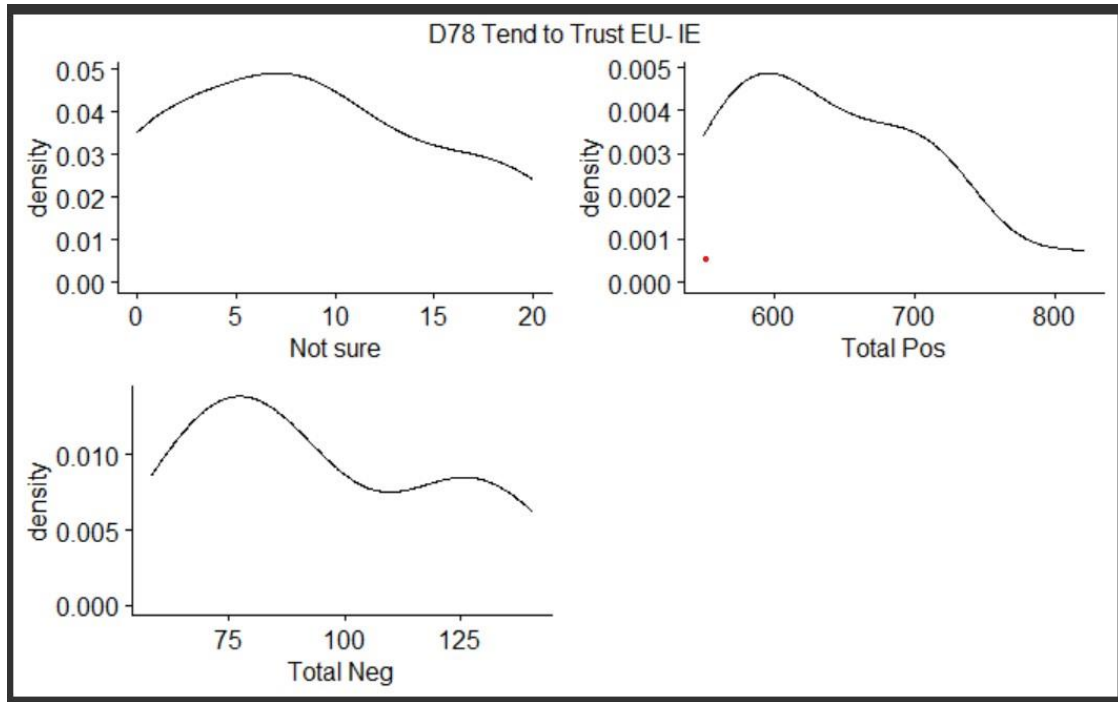


Figure 4.40- Density plots for IE last 3 responses to D78- EU Image

The data overall appears normal although with ‘Total Negative’ (the third graph in the above diagram) displaying a second peak at 125 respondents & ‘Total respondents’ have an unusually tall peak. The various other tests confirm the basic good overall shape of the data, ‘Total Respondents’ in particular has an extreme outlier that tallies with the pervious density plot.:

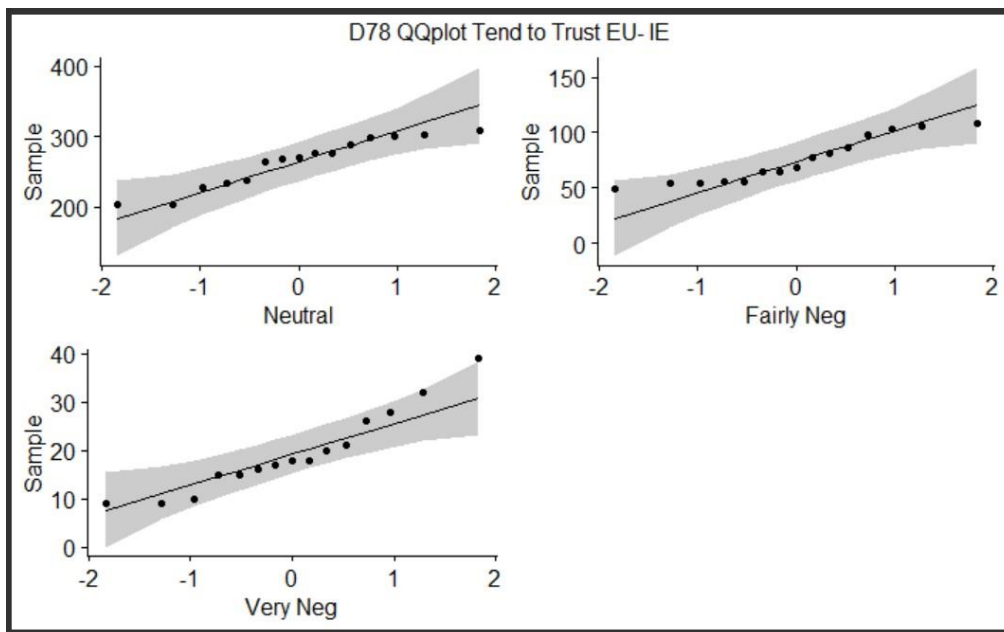


Figure 4.41- Q-Q plots for IE first 3 responses to D78- EU Image

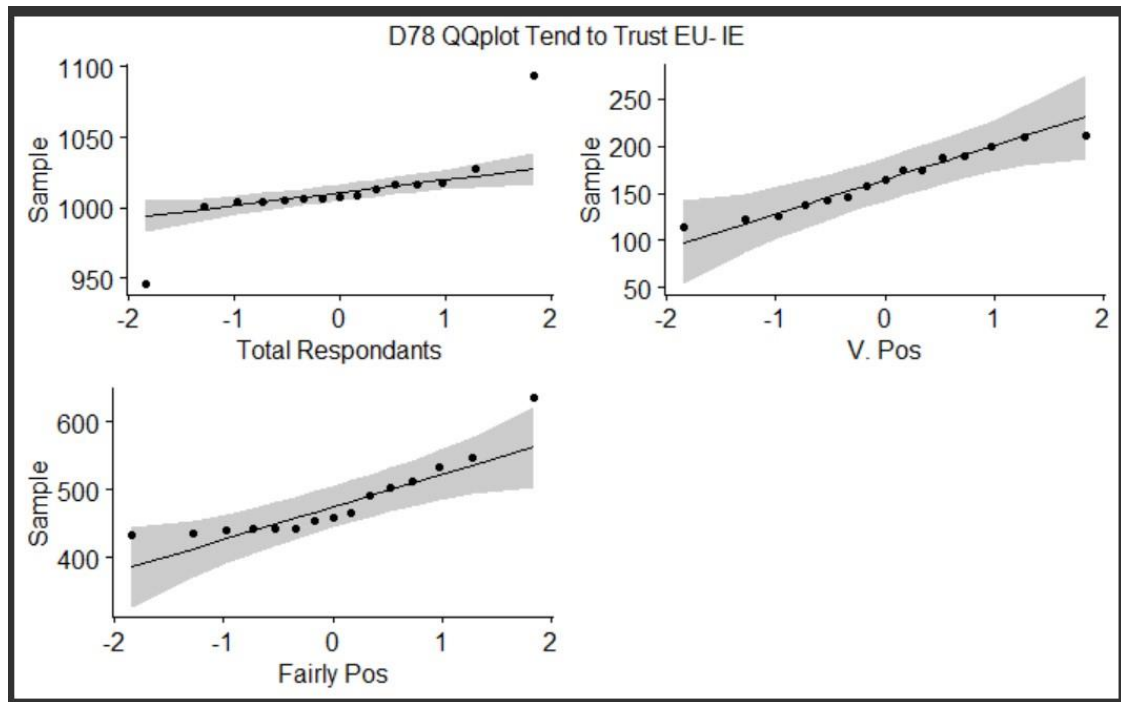


Figure 4.42- Q-Q plots for IE second 3 responses to D78- EU Image

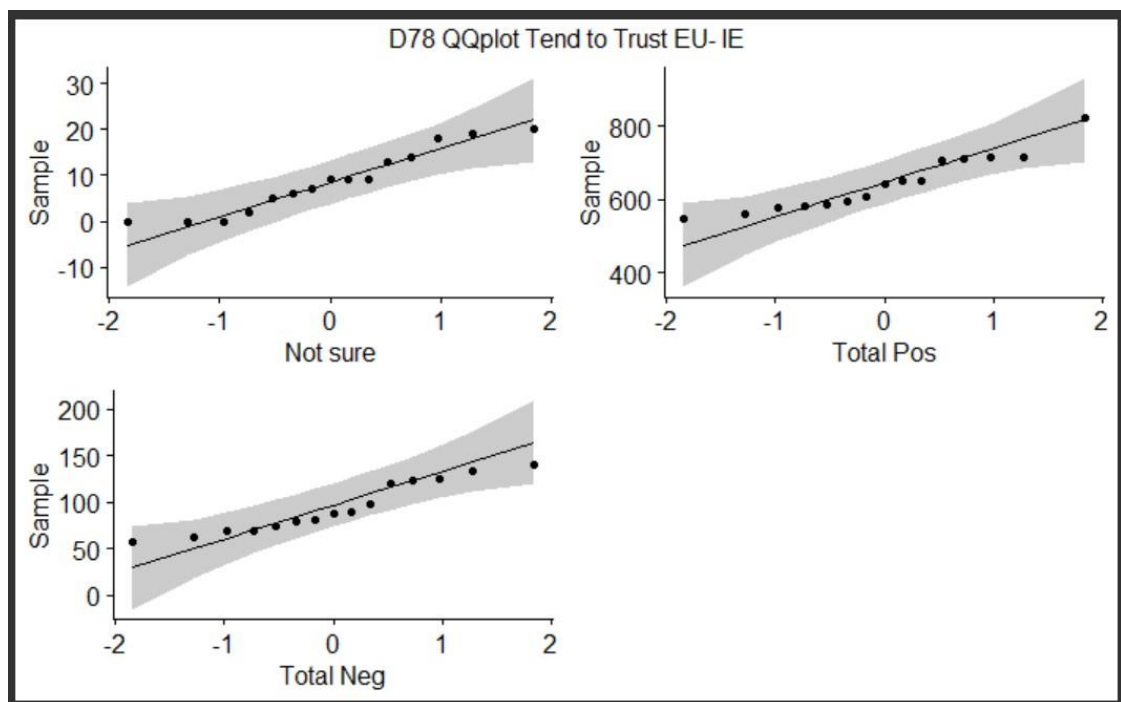


Figure 4.43- Q-Q plots for IE last 3 responses to D78- EU Image

Similar results are found with the Shapiro Wilkes, both 'TOTAL' and 'Fairly Positive' correspond to outliers seen in the Q-Q plots seen previously:


```

(r Shapiro test)
shapiro.test(stats_D78_TRUST_IE_wide$TOTAL)
# p-value = 0.0006623 Not Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Very Positive`)
# p-value = 0.5617 Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Fairly Positive`)
# p-value = 0.005817 Not Normal

shapiro.test(stats_D78_TRUST_IE_wide$Neutral)
# p-value = 0.1559 Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Fairly Negative`)
# p-value = 0.07824 Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Very Negative`)
# p-value = 0.2221 Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Not sure`)
# p-value = 0.2264 Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Total Positive`)
# p-value = 0.1634 Normal

shapiro.test(stats_D78_TRUST_IE_wide$`Total Negative`)
# p-value = 0.1533 Normal

```

Figure 4.44- Shapiro-Wilkes test of responses to D78- EU Image

```

(r Kurtosis)
kurtosis(stats_D78_TRUST_IE_wide$TOTAL)
#6.934243- leptokurtic- More outliers
kurtosis(stats_D78_TRUST_IE_wide$`Very Positive`)
#1.768125- Platykurtic- fewer and less extreme outliers
kurtosis(stats_D78_TRUST_IE_wide$`Fairly Positive`)
#4.511103- leptokurtic- More outliers
kurtosis(stats_D78_TRUST_IE_wide$Neutral)
#1.993345- Platykurtic- fewer and less extreme outliers
kurtosis(stats_D78_TRUST_IE_wide$`Fairly Negative`)
#1.658714- Platykurtic- fewer and less extreme outliers
kurtosis(stats_D78_TRUST_IE_wide$`Very Negative`)
#2.962176- Platykurtic- fewer and less extreme outliers
kurtosis(stats_D78_TRUST_IE_wide$`Not sure`)
#1.893067- Platykurtic- fewer and less extreme outliers
kurtosis(stats_D78_TRUST_IE_wide$`Total Positive`)
#2.866885- Platykurtic- fewer and less extreme outliers
kurtosis(stats_D78_TRUST_IE_wide$`Total Negative`)
#1.725742- Platykurtic- fewer and less extreme outliers

```

Figure 4.45- Kurtosis test of responses to D78- EU Image

A Kurtosis test also confirms the previous tests and finds that ‘Total’ and ‘Fairly Positive’ are both leptokurtic and are prone to more and more extreme outliers as has been found in the data. Overall, the IE, UK and PL data had similar structure and was found to be fairly robust overall. The DE data however was less well structured:

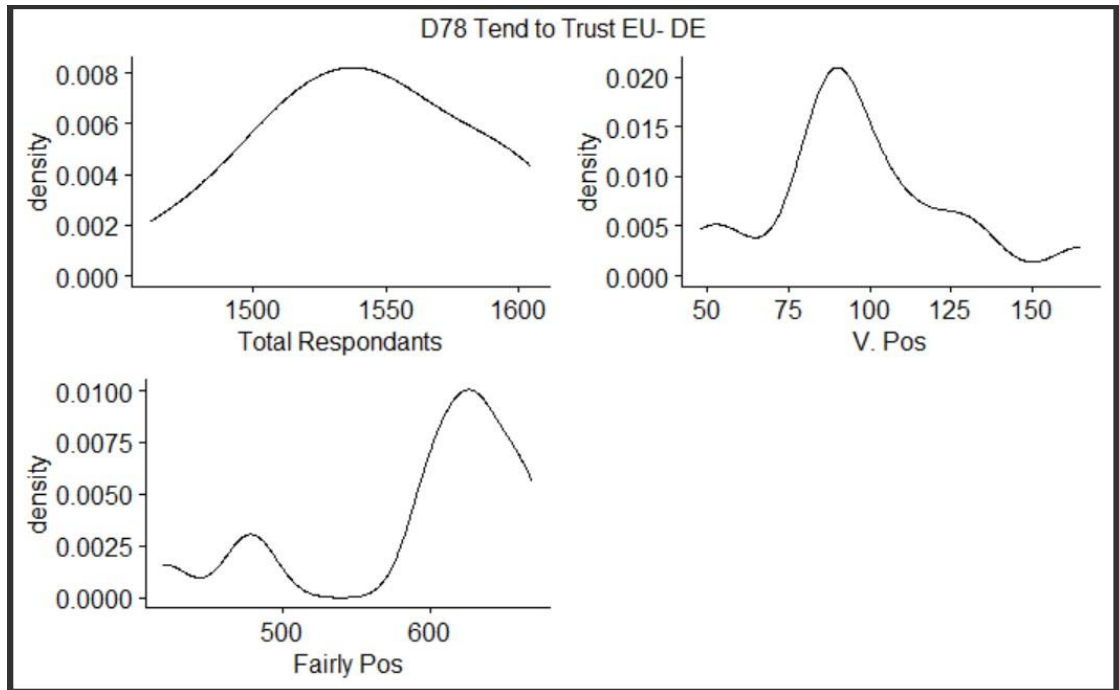


Figure 4.46- Density plots for DE first 3 responses to D78- EU Image

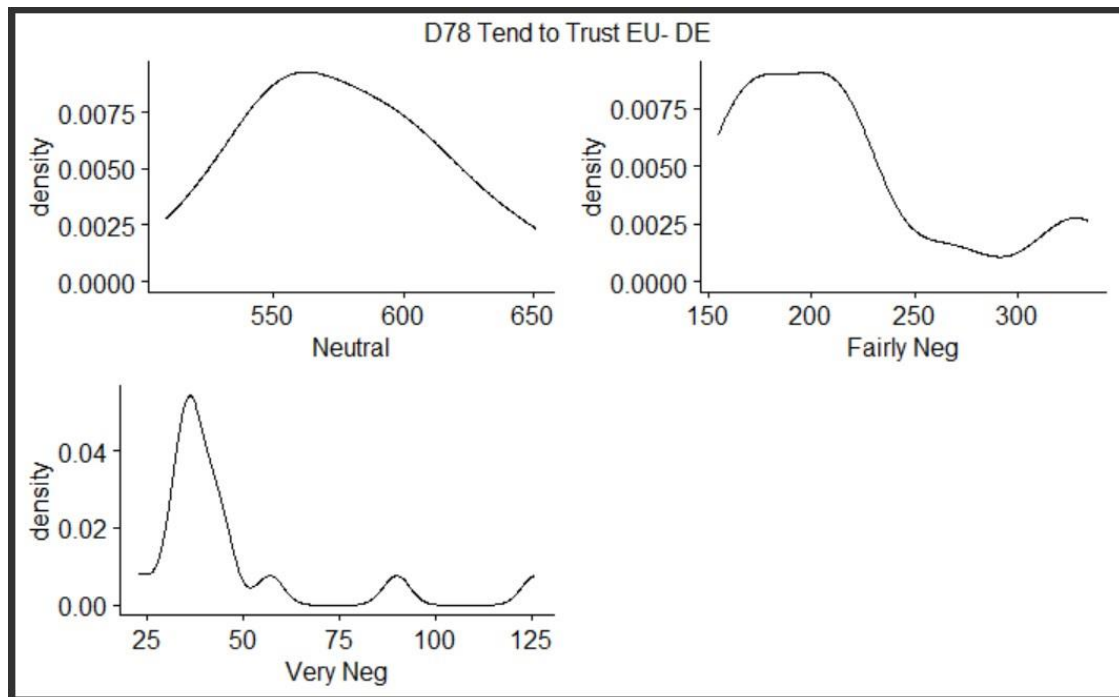


Figure 4.47- Density plots for DE second 3 responses to D78- EU Image

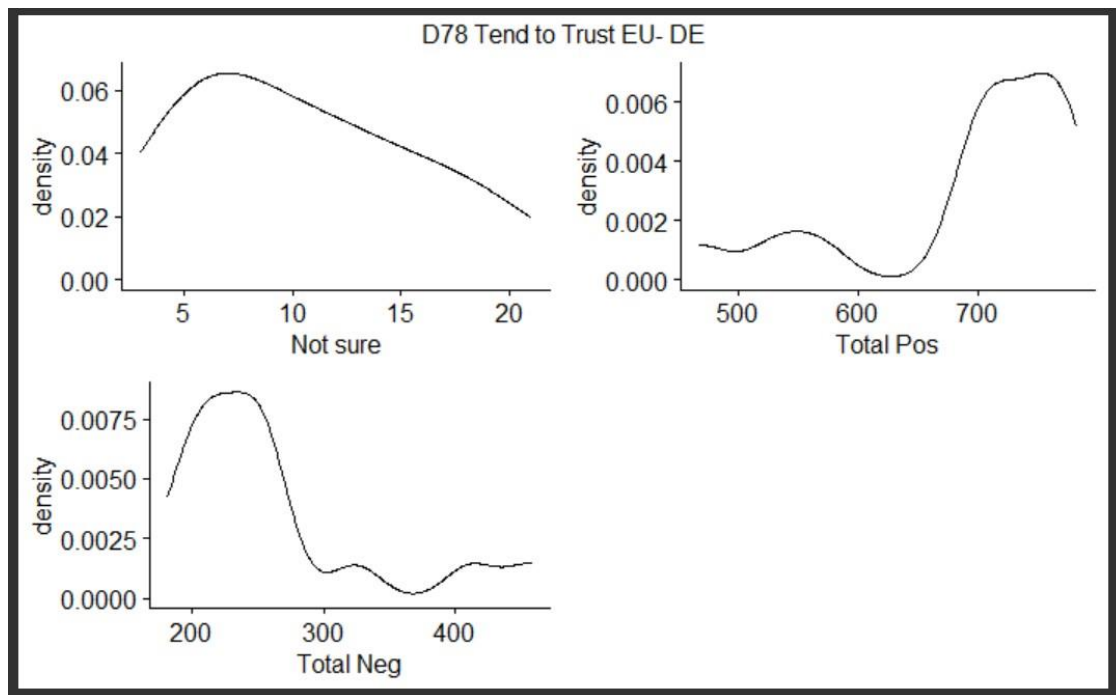


Figure 4.48- Density plots for DE final 3 responses to D78- EU Image

The DE data clearly demonstrated much less robustness on examination. This was a surprising outlier given that the data in all three other nations was of similar shape overall. The DE data was fairly normal in some of the responses ('Not sure', 'Neutral' & 'Total Respondents') but other responses had unexpected modality, 'Total Negative' and 'Very Negative' in particular had very unusual peaks in their plots indicating that there were many more outliers in that data.

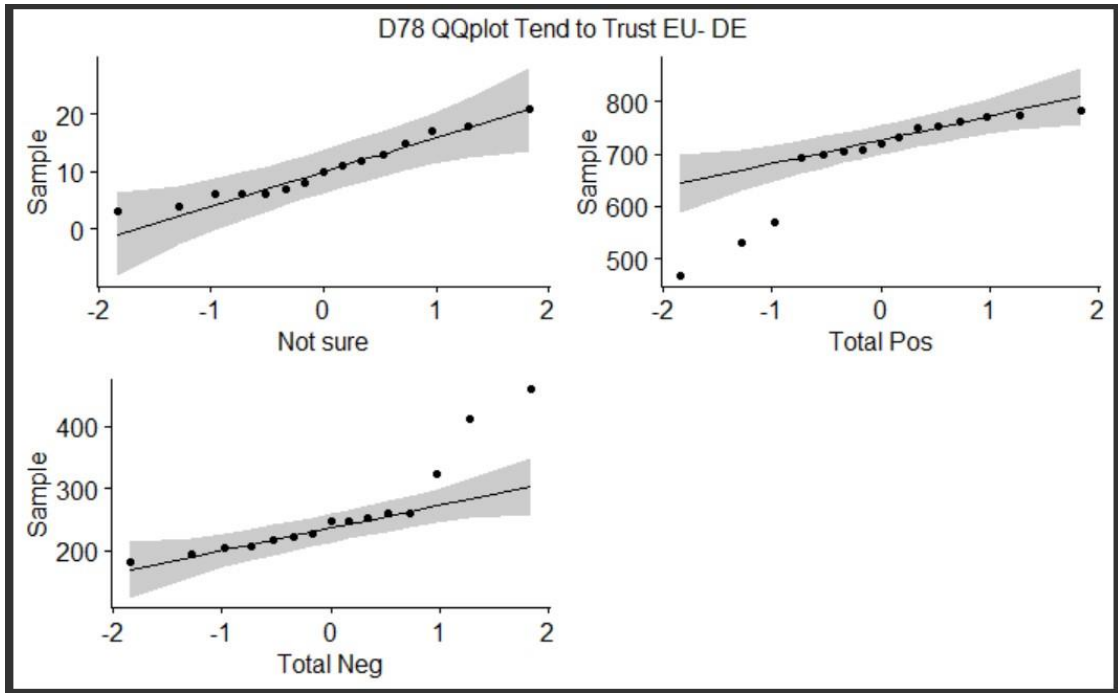


Figure 4.49- Q-Q plots for DE first 3 responses to D78- EU Image

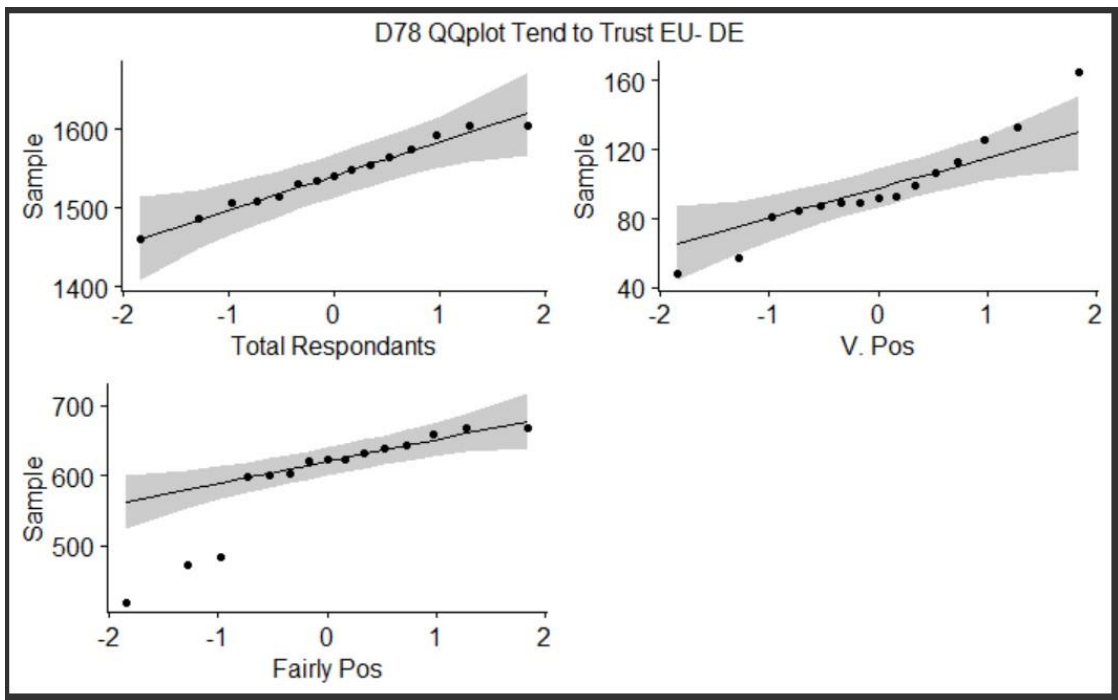


Figure 4.50- Q-Q plots for DE second 3 responses to D78- EU Image

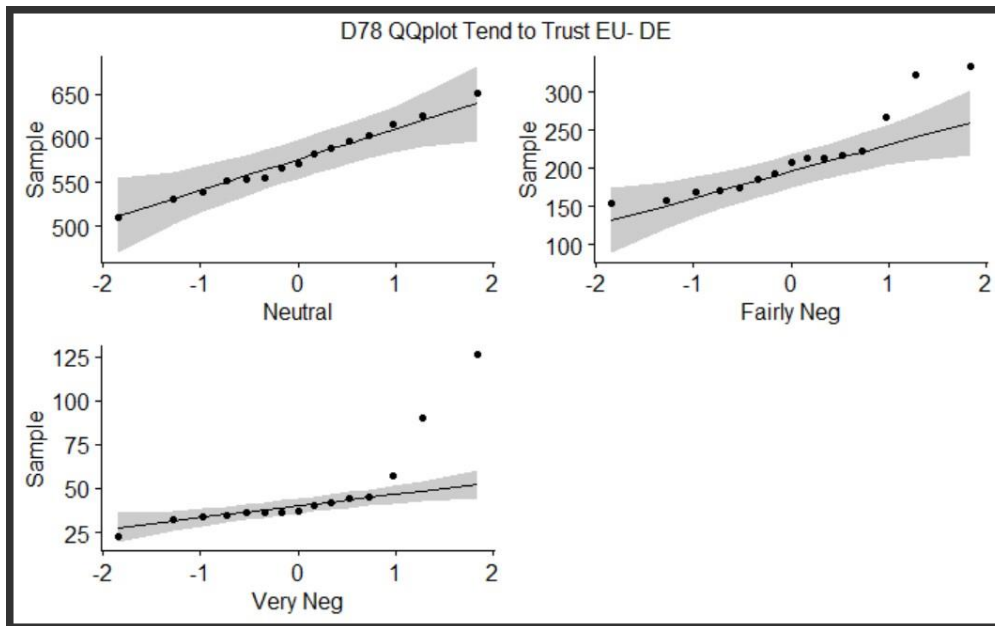


Figure 4.51- Q-Q plots for DE final 3 responses to D78- EU Image

Five Q-Q plots (‘Very Negative’, ‘Fairly Negative’, ‘Fairly Positive’, ‘Total Negative’, ‘Total Positive’) of the nine plots produced, displayed extreme outliers from the remainder of the data. Any statistical questions based upon this data would have to be considered tentative at best.

4.8. Machine Learning

The data was also sufficiently prepared to be utilized in a Machine Learning approach, it was deemed at this stage that k-NN would best suit the shape of the data given that the test dataframes are designed with a minimum of numeric data and country factor levels. The caret library²⁵ was used throughout.

```

data <- test_QA8_TRUST_EU #set the data

glimpse(data)

Rows: 48
Columns: 9
$ Year          <fct> 2015 Spring, 2015 Aut, 2016 Spring, 2016 Aut, 2017 Aut, 2018 Spring...
$ TOTAL         <dbl> 1554, 1548, 1592, 1531, 1565, 1509, 1519, 1487, 1540, 1575, 1535, 1...
$ `Tend to Trust` <dbl> 610, 439, 449, 566, 732, 747, 777, 718, 749, 760, 731, 765, 371, 30...
$ `Trust %`      <dbl> 0.39, 0.28, 0.28, 0.37, 0.47, 0.49, 0.51, 0.48, 0.49, 0.48, 0.47, 0...
$ `Tend not to Trust` <dbl> 740, 971, 958, 815, 660, 632, 576, 619, 668, 687, 672, 690, 724, 83...
$ `Not Trust %`  <dbl> 0.48, 0.63, 0.60, 0.53, 0.42, 0.42, 0.38, 0.42, 0.43, 0.44, 0.44, 0...
$ `Not Sure`    <dbl> 203, 138, 185, 150, 173, 130, 165, 150, 123, 128, 133, 149, 211, 17...
$ DK            <dbl> 0.13, 0.09, 0.12, 0.10, 0.11, 0.09, 0.11, 0.10, 0.08, 0.08, 0.09, 0...
$ Country       <chr> "Germany", "Germany", "Germany", "Germany", "Germany", "Germany", "...

```

Figure 4-1- glimpse of the data to be used for a k-NN approach.

²⁵ <https://cran.r-project.org/web/packages/caret/index.html>

The data to be used was the same data as was wrangled and prepared for the overall experiment. In the image above the Country column will constitute the target with ‘Germany’ used in the test (the column was subsequently changed to the correct <fct> format).

	Year	TOTAL	Tend to Trust	Trust %	Tend not to Trust	Not Trust %	Not Sure	DK	Country
1	2015 Spring	1554	610	0.39	740	0.48	203	0.13	Germany
2	2015 Aut	1548	439	0.28	971	0.63	138	0.09	Germany
3	2016 Spring	1592	449	0.28	958	0.60	185	0.12	Germany
4	2016 Aut	1531	566	0.37	815	0.53	150	0.10	Germany
5	2017 Aut	1565	732	0.47	660	0.42	173	0.11	Germany
6	2018 Spring	1509	747	0.49	632	0.42	130	0.09	Germany
7	2018 Aut	1519	777	0.51	576	0.38	165	0.11	Germany
8	2019 Spring	1487	718	0.48	619	0.42	150	0.10	Germany
9	2019 Aut	1540	749	0.49	668	0.43	123	0.08	Germany
10	2020 Aut	1575	760	0.48	687	0.44	128	0.08	Germany
11	2021 Spring	1535	731	0.47	672	0.44	133	0.09	Germany
12	2021 Aut	1604	765	0.48	690	0.43	149	0.09	Germany
13	2015 Spring	1306	371	0.29	724	0.55	211	0.16	UK
14	2015 Aut	1314	302	0.23	834	0.63	178	0.14	UK
15	2016 Spring	1352	404	0.30	800	0.59	148	0.11	UK
16	2016 Aut	1343	415	0.31	753	0.56	175	0.13	UK
17	2017 Aut	1334	380	0.29	788	0.59	166	0.12	UK
18	2018 Spring	1337	403	0.30	767	0.57	166	0.13	UK
19	2018 Aut	1015	317	0.31	539	0.53	159	0.16	UK

Figure 4-2- the structure of the data used in both the experiment and the k-NN approach.

The Machine Learning data was arranged equally in the data:

```
prop.table(table(data$Country)) # Perfectly balanced as all things should be
```

Germany	IE	PL	UK
0.25	0.25	0.25	0.25

Figure 4-3- Share of the country factor data in the Machine Learning dataframe.

A short command was written to pre-process the data and apply weighted averages across the available numeric data:

```
data_preprocess <- preProcess(data[, 2:9], method = c("scale", "center")) # preProcess the date using scaling
```

Figure 4-4- Pre-processing command. In this case applied on columns with index position 2:

The pre-processing step was applied to the test data and then checked to ensure that it had been applied as expected:

```
Created from 48 samples and 8 variables

Pre-processing:
- centered (7)
- ignored (1)
- scaled (7)
```

Figure 4-5- application of Pre-processing to the test data. In this case the ignored column is the target column.

The caret package ‘predict’ function was applied to the test date using the Pre-processing data that was previously defined. A glimpse at the data confirmed that the data has been correctly processed.

```
data_stand <- predict(data_preprocess, data[,2:9]) #apply caret Prediction to data
glimpse(data_stand)

Rows: 48
Columns: 8
$ TOTAL <dbl> 1.5433230, 1.5174843, 1.7069680, 1.4442747, 1.5906939, 1.3495329, ...
$ `Tend to Trust` <dbl> 0.62525519, -0.54994394, -0.48121885, 0.32286477, 1.46370136, 1.5...
$ `Trust %` <dbl> -0.439432203, -1.414307006, -1.414307006, -0.616682167, 0.2695676...
$ `Tend not to Trust` <dbl> 1.03038513, 2.23847547, 2.17048770, 1.42262225, 0.61199887, 0.465...
$ `Not Trust %` <dbl> 0.3213821, 1.7497468, 1.4640738, 0.7975036, -0.2499638, -0.249963...
$ `Not Sure` <dbl> 1.28395054, 0.08169193, 0.95101739, 0.30364737, 0.72906195, -0.06...
$ DK <dbl> 0.31711680, -0.48402037, 0.11683250, -0.28373608, -0.08345179, -0...
$ Country <fct> Germany, Germany, Germany, Germany, Germany, Germany, Germany, Ge...
```

Figure 4-6- Pre-processing applied to test data.

A second confirmation is that the mean of all values has been set to zero across the dataframe:

```

TOTAL          Tend to Trust      Trust %          Tend not to Trust  Not Trust %
Min.   :-0.8511  Min.   :-1.5739  Min.   :-1.8574  Min.   :-1.50085  Min.   :-1.8688
1st Qu.:-0.7929  1st Qu.:-0.7338  1st Qu.:-0.7718  1st Qu.:-0.83013  1st Qu.:-0.6309
Median :-0.7133  Median :-0.1754  Median : 0.2253  Median :-0.06788  Median :-0.2500
Mean   : 0.0000  Mean   : 0.0000  Mean   : 0.0000  Mean   : 0.00000  Mean   : 0.0000
3rd Qu.: 0.8188  3rd Qu.: 0.4964  3rd Qu.: 0.5354  3rd Qu.: 0.75713  3rd Qu.: 0.9880
Max.   : 1.7586  Max.   : 1.9860  Max.   : 2.6624  Max.   : 2.23848  Max.   : 1.8450
Not Sure          DK          Country
Min.   :-2.4523  Min.   :-2.2866  Germany:12
1st Qu.:-0.2790  1st Qu.:-0.3338  IE       :12
Median : 0.1002  Median : 0.1168  PL       :12
Mean   : 0.0000  Mean   : 0.0000  UK       :12
3rd Qu.: 0.5996  3rd Qu.: 0.5174
Max.   : 2.0608  Max.   : 2.5202
```

Figure 4-7- Summary of k-NN test data after pre-processing had been applied.

The final steps taken to prepare the data was to set a test and training split:

```
data_split <- createDataPartition(stand_data$Country, times = 1, p = 0.8, list = FALSE) #Split the
data for test and train
train_data <- stand_data[data_split,] # extract the train data
test_data <- stand_data[-data_split,] # remainder of the data is test data
```

Figure 4-8- Train and test data split.

With that, the data was now prepared for a k-NN test to hopefully confirm and strengthen the results in later stages of the study. These steps were completed for each question that would make up the overall experiment design. The k-NN approach would be subsequently conducted on each question in turn following the completion of the statistical testing in order to offer a secondary source of information regarding the research question.

4.9. Key Reflections

Following this chapter, the data is now cleaned, wrangled and correctly constructed for our experiment as utilised in the next chapter. The process of wrangling and preparing the Eurobarometer proved the most time-consuming stage in the project and necessitated a more careful and discriminating selection of questions to ensure that the most appropriate questions were used for the experiment. A detailed exploration of the Eurobarometer data as undertaken in this chapter confirmed that the majority of the data is well constituted and sufficiently robust for the proposed experiment. Overall, the data was seen to be very well suited for purposes of the experiment and has been sufficiently wrangled and prepared to ensure that the validity of the results overall can be considered as quite sufficient for the research question.

However, in relation to the Machine Learning data this was not as much the case. The small number of instances per question (circa 40 to 60) as well as only 4 factor levels, were found to be insufficient for very robust experimentation. The tests were carried out in turn, but their validity was found to not be on the same solid statistical footing as the data utilized in the statistical testing.

5. RESULTS AND EVALUATION

5.1. Introduction

In this chapter the overall results of this project will be presented and discussed. Those results will be evaluated in relation to the original research question and to what degree those results can be considered of statistical value based on the findings in the previous chapter (Chapter 4) will be explored.

The datasets will be re-examined, and the choices made in their cleaning and wrangling will be briefly touched upon. Finally, the suitability of the Eurobarometer survey questions will be discussed.

The choice of Statistical Approach will be discussed and its appropriateness in the case of this data will be delineated and justified. The limitations of the experimental design will be explored as well as the implications of those limitations upon the results. Finally, the implications of the results will be discussed.

5.2. The Datasets

As discussed in Chapter 3, three datasets are being used to conduct this experiment. Two of the datasets are sourced from the EUvsDisinformation group²⁶, one of which was scraped from the EUvsDisinformation website database²⁷. The second dataset, available on Kaggle²⁸, is based on an EUvsDisinformation Hackathon event (that dataset had been uploaded and made available by one of the attendees). The cleaning, wrangling and preparation of the data is explained in detail in Chapter 3.

The final dataset is the Eurobarometer survey²⁹ data between the years 2015 and 2022. This data has been cleaned and wrangled based on the findings after a detailed exploration of the EUvsDisinformation datasets. The three member states that have received the most disinformation during the study period have had their data extracted

²⁶ <https://euvsdisinfo.eu/>

²⁷ <https://euvsdisinfo.eu/disinformation-cases/>

²⁸ <https://www.kaggle.com/datasets/corrieaar/disinformation-articles>

²⁹ https://data.europa.eu/data/datasets/s2532_95_3_95_eng?locale=en

from the Eurobarometer data. The data for Ireland was also selected as a control group, Ireland was found to not have been a significant target of disinformation based on the EUvsDisinformation data, as explored in Chapter 3.

Of the circa 170 available questions in each Eurobarometer survey, 44 questions were initially chosen as being of value to the current study. Those questions were extracted from each Eurobarometer survey and loaded into RStudio. Of those 44 questions, four were chosen as being the most pertinent to the research question associated with this study. The responses for Germany (DE), The United Kingdom (UK), Poland (PL) and Ireland (IE) were extracted and cleaned and wrangled (as discussed in Chapter 4).

A detailed statistical analysis of those questions was carried out to gauge their suitability in answering the research question. Of the four questions, it was found that question QA8³⁰ was the most robust across all 4 member states, Question D71³¹ and question D73³² were also found to have robust data but with less solid foundations as questions QA8. Finally, question D78³³ was found to data of noticeably lesser robustness although this is balanced somewhat by the much greater selection of responses available to this survey question. In general, the data was deemed to be of sufficient robustness for the research question but with varying degrees of assuredness in the results.

³⁰ QA8: ‘I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it: The EU’.

³¹ D71: ‘When you get together with friends or relatives, would you say you discuss frequently, occasionally or never about...? European Politics’.

³² D73: At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...? The EU’.

³³ D78: ‘In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?’

5.3. Choosing a Statistical Approach

This study set out to test whether countries that had been the target of disinformation would have any change in their attitudes towards the EU as measured in selected Eurobarometer questions, and thus from a statistical perspective, it seeks to assess the correlation between data contained in multiple datasets. Therefore, the most suitable test in this instance is MANOVA (Multivariate analysis of Variance) as it provides both a regression analysis and analysis of variance, for multiple dependent variables by one or more factor variables or covariates.

The scraped EUvsDisinformation dataset after having the dates aligned with the date range available in both the Kaggle EUvsDisinformation Dataset and the Eurobarometer survey datasets, was arranged as follows:

Date	Title	Outlets	Country
2020-01-02	Crimean referendum was the will of the Crimeans, while Kos...	Sputnik Serbia - Latin	Kosovo
2020-01-02	Crimean referendum was the will of the Crimeans, while Kos...	Sputnik Serbia - Latin	Serbia
2020-01-02	The US is not a free country and is under occupation by mul...	Geopolitica.ru - Arabic	US
2020-01-02	The US is not a free country and is under occupation by mul...	Geopolitica.ru - Arabic	Russia
2020-01-02	Ukraine does not want to buy gas from Russia because its p...	russian.rt.com	Ukraine
2020-01-02	Ukraine does not want to buy gas from Russia because its p...	russian.rt.com	Russia
2020-01-02	The Treaty of Tartu has lost its force	arabic.rt.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	m.akhbarelyom.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	m.akhbarelyom.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	shorouknews.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	albawabhnews.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	akhbarak.net	USSR
2020-01-02	The Treaty of Tartu has lost its force	watan-m3k.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	capbonnews.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	shamra.sy	USSR
2020-01-02	The Treaty of Tartu has lost its force	nabd.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	elaosboa.news	USSR
2020-01-02	The Treaty of Tartu has lost its force	emeknes.com	USSR
2020-01-02	The Treaty of Tartu has lost its force	...	USSR

Figure 5-1- The Dates_Scrape Dataset

The Kaggle dataset, after having the dates aligned with the available data from the Scraped EUvsDisinformation dataset was arranged as follows:

claim_published	keyword_name	country_name	organization_name	language_name
2019-12-13	Abandoned Ukraine	Russia	sputnik.by // lifenews.ru	Russian
2019-12-13	Angela Merkel	Russia	sputnik.by // lifenews.ru	Russian
2019-12-13	Emmanuel Macron	Russia	sputnik.by // lifenews.ru	Russian
2019-12-13	Ukrainian statehood	Russia	sputnik.by // lifenews.ru	Russian
2019-12-13	Vladimir Putin	Russia	sputnik.by // lifenews.ru	Russian
2019-12-13	Minsk agreements	Russia	sputnik.by // lifenews.ru	Russian
2019-12-13	Abandoned Ukraine	Ukraine	sputnik.by // lifenews.ru	Russian
2019-12-13	Angela Merkel	Ukraine	sputnik.by // lifenews.ru	Russian
2019-12-13	Emmanuel Macron	Ukraine	sputnik.by // lifenews.ru	Russian
2019-12-13	Ukrainian statehood	Ukraine	sputnik.by // lifenews.ru	Russian
2019-12-13	Vladimir Putin	Ukraine	sputnik.by // lifenews.ru	Russian
2019-12-13	Minsk agreements	Ukraine	sputnik.by // lifenews.ru	Russian
2019-12-13	Abandoned Ukraine	The West	sputnik.by // lifenews.ru	Russian
2019-12-13	Angela Merkel	The West	sputnik.by // lifenews.ru	Russian
2019-12-13	Emmanuel Macron	The West	sputnik.by // lifenews.ru	Russian
2019-12-13	Ukrainian statehood	The West	sputnik.by // lifenews.ru	Russian
2019-12-13	Vladimir Putin	The West	sputnik.by // lifenews.ru	Russian
2019-12-13	Minsk agreements	The West	sputnik.by // lifenews.ru	Russian

Figure 5-2- Current structure of Datas_Kaggle

Both datasets present the country data as SUM of the disinformation instances. This data, across both datasets would be used in the MANOVA test.

The Eurobarometer data was arranged as thusly:

	Year	TOTAL	Fréquentment	Frequently	Occasionnellement	Occasionally	Jamais	Never	NSP	Don't know
1	2015 Spring	1018	82	0.08	419	0.41	516	0.51	2	0
2	2015 Aut	1004	118	0.12	423	0.42	457	0.46	5	0
3	2016 Spring	1004	106	0.11	447	0.44	450	0.45	2	0
4	2017 Spring	1009	127	0.13	468	0.46	409	0.41	5	0
5	2017 Aut	1001	127	0.13	448	0.45	422	0.42	4	0
6	2018 Spring	1007	111	0.11	443	0.44	454	0.45	0	0
7	2018 Aut	946	110	0.11	422	0.45	414	0.44	0	0
8	2019 Spring	1028	155	0.15	507	0.49	366	0.36	0	0
9	2019 Aut	1013	150	0.15	490	0.48	373	0.37	0	0
10	2020 Spring	1005	216	NA	641	NA	148	NA	NA	0
11	2020 Aut	1094	155	0.14	652	0.60	286	0.26	1	0
12	2021 Spring	1017	149	0.15	659	0.65	209	0.20	0	0
13	2021 Aut	1006	158	0.16	442	0.44	404	0.40	2	0
14	2022 Spring	1017	118	0.12	505	0.50	392	0.38	2	0

Figure 5-3- Question D71, IE data pivoted wide

The Eurobarometer data was arranged with each member state's responses to the selected questions arranged as separate datasets. This data would be used in the MANOVA test but would first have to be amalgamated such that multiple member state's data would be arranged on single dataframes. In order to differentiate each

member's data, a new column would have to be added that contained a member state's country code:

```

## Add cols
### Add cols DE
### Add cols UK
### Add cols PL
### Add cols IE

stats_D71a.2_IE_wide <- stats_D71a.2_IE_wide %>%
  mutate(Country = "IE")

stats_D73a.2_DIR_IE_wide <- stats_D73a.2_DIR_IE_wide %>%
  mutate(Country = "IE")

stats_D78_TRUST_IE_wide <- stats_D78_TRUST_IE_wide %>%
  mutate(Country = "IE")

stats_QA8_TRUST_IE_wide <- stats_QA8_TRUST_IE_wide %>%
  mutate(Country = "IE")

```

Figure 5-4- Add columns to data before testing.

The resulting dataframes were still arranged as individual member state's responses to the pertinent questions, the next step would be to amalgamate all member state's responses to the question together to allow the MANOVA test compare the member state's results to one another across a single dataframe:

	Year	TOTAL	Frequently	Freq%	Occasionally	Occ %	Never	Never %	Don't Know	DK	Country
1	2015 Spring	1018	82	0.08	419	0.41	516	0.51	2	0	IE
2	2015 Aut	1004	118	0.12	423	0.42	457	0.46	5	0	IE
3	2016 Spring	1004	106	0.11	447	0.44	450	0.45	2	0	IE
4	2017 Spring	1009	127	0.13	468	0.46	409	0.41	5	0	IE
5	2017 Aut	1001	127	0.13	448	0.45	422	0.42	4	0	IE
6	2018 Spring	1007	111	0.11	443	0.44	454	0.45	0	0	IE
7	2018 Aut	946	110	0.11	422	0.45	414	0.44	0	0	IE
8	2019 Spring	1028	155	0.15	507	0.49	366	0.36	0	0	IE
9	2019 Aut	1013	150	0.15	490	0.48	373	0.37	0	0	IE
10	2020 Spring	1005	216	NA	641	NA	148	NA	NA	0	IE
11	2020 Aut	1094	155	0.14	652	0.60	286	0.26	1	0	IE
12	2021 Spring	1017	149	0.15	659	0.65	209	0.20	0	0	IE
13	2021 Aut	1006	158	0.16	442	0.44	404	0.40	2	0	IE
14	2022 Spring	1017	118	0.12	505	0.50	392	0.38	2	0	IE

Figure 5-5- The member state's country code has been added. NA values were unused missing percentage values and can be ignored.

The initial experiment design at this stage was to compare member state's response on a one-to-one basis however it was quickly apparent that this approach

would result in needlessly complex and ungainly data. It was decided to also conduct the experiment using amalgamated questions containing all four-member state's response on a single dataframe as well as different combinations of member state's responses. The initial approach can be seen in the image below as well as the amalgamated approach:

```
(*)
# test_D71_DE_UK <- rbind(stats_D71a.2_DE_wide, stats_D71a.2_UK_wide)
# test_D71_DE_IE <- rbind(stats_D71a.2_DE_wide, stats_D71a.2_IE_wide)
# test_D71_DE_PL <- rbind(stats_D71a.2_DE_wide, stats_D71a.2_PL_wide)
# test_D71_UK_PL <- rbind(stats_D71a.2_UK_wide, stats_D71a.2_PL_wide)
# test_D71_UK_IE <- rbind(stats_D71a.2_UK_wide, stats_D71a.2_IE_wide)
# test_D71_PL_IE <- rbind(stats_D71a.2_PL_wide, stats_D71a.2_IE_wide)
test_D71_Euro_Pol <- rbind(stats_D71a.2_DE_wide, stats_D71a.2_UK_wide,
                           stats_D71a.2_IE_wide, stats_D71a.2_PL_wide)

# test_D73_DE_UK <- rbind(stats_D73a.2_DIR_DE_wide, stats_D73a.2_DIR_UK_wide)
# test_D73_DE_IE <- rbind(stats_D73a.2_DIR_DE_wide, stats_D73a.2_DIR_IE_wide)
# test_D73_DE_PL <- rbind(stats_D73a.2_DIR_DE_wide, stats_D73a.2_DIR_PL_wide)
# test_D73_UK_PL <- rbind(stats_D73a.2_DIR_UK_wide, stats_D73a.2_DIR_PL_wide)
# test_D73_UK_IE <- rbind(stats_D73a.2_DIR_UK_wide, stats_D73a.2_DIR_IE_wide)
# test_D73_PL_IE <- rbind(stats_D73a.2_DIR_PL_wide, stats_D73a.2_DIR_IE_wide)
test_D73_DIR_EU <- rbind(stats_D73a.2_DIR_DE_wide, stats_D73a.2_DIR_UK_wide,
                           stats_D73a.2_DIR_IE_wide, stats_D73a.2_DIR_PL_wide)
```

Figure 5-6- The original rbind of individual responses and the eventual design of all member state's responses in a single dataframe

The resultant dataframes combined all four member state's responses on a single dataframe that could then be used to complete a MANOVA test:

	Year	TOTAL	Frequently	Freq%	Occasionally	Occ %	Never	Neve %	Don't Know	DK	Country
15	2022 Spring	1507	534	0.35	854	0.57	115	0.08	5.00	0.00	Germany
16	2015 Spring	1306	155	0.12	569	0.44	580	0.44	1.00	0.00	UK
17	2015 Aut	1314	200	0.15	585	0.45	525	0.40	4.00	0.00	UK
18	2016 Spring	1352	273	0.20	594	0.44	482	0.36	3.00	0.00	UK
19	2017 Spring	1365	255	0.19	657	0.48	449	0.33	4.00	0.00	UK
20	2017 Aut	1334	286	0.21	572	0.43	476	0.36	0.00	0.00	UK
21	2018 Spring	1337	230	0.17	596	0.45	510	0.38	1.00	0.00	UK
22	2018 Aut	915	216	0.24	423	0.46	273	0.30	4.00	0.00	UK
23	2019 Spring	1032	253	0.24	409	0.40	368	0.36	2.00	0.00	UK
24	2019 Aut	1010	213	0.21	391	0.39	405	0.40	2.00	0.00	UK
25	2020 Spring	1153	176	0.15	677	0.59	301	0.26	0.00	0.00	UK
26	2020 Aut	1301	134	0.10	607	0.47	560	0.43	0.00	0.00	UK
27	2021 Spring	1020	121	0.12	614	0.60	284	0.28	1.00	0.00	UK
28	2021 Aut	1086	118	0.11	584	0.54	384	0.35	0.00	0.00	UK
29	2022 Spring	1034	149	0.15	427	0.41	457	0.44	1.00	0.00	UK
30	2015 Spring	1018	82	0.08	419	0.41	516	0.51	2.00	0.00	IE
31	2015 Aut	1004	118	0.12	423	0.42	457	0.46	5.00	0.00	IE
32	2016 Spring	1004	106	0.11	447	0.44	450	0.45	2.00	0.00	IE
33	2017 Spring	1009	127	0.13	468	0.46	409	0.41	5.00	0.00	IE
34	2017 Aut	1001	127	0.13	448	0.45	422	0.42	4.00	0.00	IE

Figure 5-7- The test dataframe displaying three of the four member states in the Country column.

At this stage, the data was correctly constructed and prepared for a MANOVA test. The test would be run on a question-by-question basis with all four-member state's data combined into a single dataframe. An ANOVA test would have been the appropriate choice if we were merely measuring the difference in variance for a single response variable, for example 'Frequently' in the example above. MANOVA on the other hand is the appropriate test when testing multiple different response variables. Each of our questions contains at least three dependant or response variables:

Question	Response Variables
D71	<ul style="list-style-type: none"> • Frequently • Occasionally • Never
D73	<ul style="list-style-type: none"> • Things are going in the right direction • Things are going in the wrong direction • Neither nor of the other • Don't know
D78	<ul style="list-style-type: none"> • Very positive • Fairly positive • Neutral • Fairly Negative • Very Negative • Not sure • Total Positive • Total Negative
QA8	<ul style="list-style-type: none"> • Tend to Trust • Tend not to Trust • Not Sure

Each question dataframe contains four independent variables or factor levels of equal size. The Factor Levels were as follows: DE = Germany, UK = United Kingdom, PL = Poland, IE = Ireland (as seen in figure 5-7).

The test data is now correctly constructed to be able to test whether the independent variable (Member state country code) has any change in the variance of responses when measured in the four questions selected from the Eurobarometer survey. The MANOVA test will use both an F-value and P-value to measure firstly whether any relationship exists between the independent and dependant variables and secondly, to measure the strength of that relationship. The F-value will present the strength of the relationship with a larger score indicating a larger variance between the group means, and the p-value will present the significance of the result with a score of < 0.05 considered to be a result that attains statistical significance. Thus, a higher F-value and lower P-value is the criteria by which the test will be considered to have been significant overall.

Since the chosen test is specifically MANOVA, there will also be a Pillai Trace score for each test. This test is best chosen in instances where data is found not to be the ideal shape for MANOVA tests. During the exploration in chapter 4 it was found that much of the data is of less-than-ideal structure overall, thus the Pillai Trace was chosen as the most appropriate test as opposed to Wilks Lambda or other possible statistical measurements. The test will provide a statistic range from 0 to 1 with values closer to 1 being considered as greater evidence that the independent variable has a statistical effect on the dependent variable.

The tests will be explored on a question-by-question basis beginning with the most robust question data and then in descending order of robustness as explored more fully in Chapter 4.

5.4. The MANOVA test and results

Question QA8: ‘I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it. The EU’

```
## (R)
#colnames(test_QA8_TRUST_EU)
# model_qa8 <- manova(cbind(TOTAL, `Tend to Trust`, `Tend not to Trust`, `Not Sure`)
test_QA8_TRUST_EU)
summary(model_qa8)
# summary.aov(model_qa8)
# plotmeans(test_QA8_TRUST_EU$`Tend to Trust` ~ test_QA8_TRUST_EU$Country)
# plotmeans(test_QA8_TRUST_EU$`Tend not to Trust` ~ test_QA8_TRUST_EU$Country)
# plotmeans(test_QA8_TRUST_EU$`Not Sure` ~ test_QA8_TRUST_EU$Country)

          Df Pillai approx F num Df den Df      Pr(>F)
Country    3  1.5678   11.768    12   129 9.476e-16 ***
Residuals 44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5-8- MANOVA summary for Question QA8

The above results suggest that there is a strong relationship between the independent and dependent variables in the QA8 data. The P-value sufficiently surpasses the level of significance at 9.46e-16, coupled with a strong Pillai measure of 1.56 and an F-value of 11.768, this question can comfortably be considered to strongly support the hypothesis that disinformation targeted at a nation has an appreciable and measurable effect on its attitudes towards the EU as measured in the Eurobarometer data.

Looking at the individual responses results we see other interesting aspects of the test:

```

Response Tend to Trust :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country   3 534725  178242   17.035 1.735e-07 ***
Residuals 44 460376   10463
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Tend not to Trust :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country   3 1363372  454457   56.323 4.131e-15 ***
Residuals 44  355023    8069
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Not Sure :
      Df Sum Sq Mean Sq F value Pr(>F)
Country   3  13295  4431.8   1.5715 0.2098
Residuals 44 124086  2820.1

```

Figure 5-9- Summary of the ANOVA for Question QA8

Both ‘Tend to Trust’ and ‘Tend not to Trust’ comfortably demonstrated significance in their P-value results added to that is highly rated Sum of Squares means. This strongly suggests a link between disinformation and a member state’s responses in those cases. However, on the other hand, ‘Not sure’ is seen not to surpass the level of significance while also have a low Sum of Square Means results. This is indicative of a result that does not support the hypotheses in that case.

This test presents strong support for the hypotheses and uncovers a strong link between those nations that have been the target of misinformation and their change in attitude towards the EU as measured in the Eurobaromter data.

Question D71: ‘When you get together with friends or relatives, would you say you discuss frequently, occasionally, or never about...? European Politics’

```

          Df Pillai approx F num Df den Df      Pr(>F)
Country   3 1.3552   14.557     9   159 < 2.2e-16 ***
Residuals 53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 5-10- MANOVA summary for Question D71

The above results suggest that there is a strong relationship between the independent and dependent variables in the D71 data, although not as strong as the previous QA8 test. The P-value sufficiently surpasses the level of significance at 2.2e-16, coupled with a strong Pillai measure of 1.3552 and an F-value of 14.557, this question can comfortably be considered support the hypothesis that disinformation targeted at a nation has an effect on its attitudes towards the EU as measured in the Eurobaromter data. Interestingly this result has a higher F-value and Pillai score but a lower P-value than the more robust data in QA8.

Unlike QA8, all three responses are seen to be robust in a summary ANOVA test:

```

Response Frequently :
      Df Sum Sq Mean Sq F value      Pr(>F)
Country   3 1318256  439419  227.32 < 2.2e-16 ***
Residuals 53  102452    1933
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Occasionally :
      Df Sum Sq Mean Sq F value      Pr(>F)
Country   3 1460253  486751  86.587 < 2.2e-16 ***
Residuals 53  297942    5622
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Never :
      Df Sum Sq Mean Sq F value      Pr(>F)
Country   3  663827  221276  40.711 8.697e-14 ***
Residuals 53 288071    5435
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 5-11- Summary of ANOVA for Question D71

All three responses surpass the level of significance, and all have strong Sum of Square Mean values. The individual responses strongly suggest a link between

disinformation and these responses to Eurobarometer questions however the overall result is somewhat less clearly indicated than question QA8.

In conclusion this test presents strong support for the hypotheses and uncovers a strong link between those nations that have been the target of disinformation and their change in attitude towards the EU as measured in the Eurobarometer data. The data is especially robust when viewed as individual questions in a summary ANOVA test.

Question D73: ‘At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...? The EU’

Question D73 was seen to have marginally less robust data than question D71, all the more so when measured against question QA8. When conducting the MANOVA test the following results were obtained:

```
          Df Pillai approx F num Df den Df      Pr(>F)
Country   3  1.2405   7.6141   15   162 1.117e-12 ***
Residuals 56
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5-12- MANOVA Summary for Question D73

This question also returns significant results but as expected, the results are on a less solid footing than the previous questions. The result demonstrates a significant P-value at 1.117e-12, though a lower score than previous questions and a high Pillai score, unexpected higher than previous questions however the F-value is quite low at 7.61. While this does not invalidate the result, it is of lesser significance than the previous questions. Looking at the questions individually we find some unusual results:

```

Response Things are going in the right Direction :
      Df Sum Sq Mean Sq F value Pr(>F)
Country 3 317121 105707  5.691 0.00179 **
Residuals 56 1040160 18574
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Things are going in the wrong Direction :
      Df Sum Sq Mean Sq F value Pr(>F)
Country 3 2486341 828780 67.914 < 2.2e-16 ***
Residuals 56 683391 12203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Neither one or the other :
      Df Sum Sq Mean Sq F value Pr(>F)
Country 3 77390 25797 1.848 0.149
Residuals 56 781721 13959

Response Don't know :
      Df Sum Sq Mean Sq F value Pr(>F)
Country 3 24073 8024.4 2.6548 0.05726 .
Residuals 56 169267 3022.6
---

```

Figure 5-13- summary ANOVA for 4 of the responses to question D71

Only 2 of the 4 responses in the above image are seen to be significant findings. ‘Things are going in the right direction’ and ‘things are going in the wrong direction’ are both seen to have significant results, the F-value for very low for right direction. ‘Neither one or the other’ is seen not to support the hypotheses while ‘don’t know’ is on the very edge of significance. These results based on the somewhat lesser robust data present a confused picture overall. While the overall MANOVA test seems strongly in support of the hypothesis, the individual questions are much less supportive.

In conclusion, this question cannot be seen to be in strong support of the hypotheses overall. While the results are varied and overall lend themselves to significance, the individual results suggest a far from stable picture.

Question D78: ‘In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?’

This data was noted as being the least robust overall when examined in Chapter 4. Testing this question was undertaken merely for completeness as the results were expected to be very confused and of little value overall in answering the research question. As discussed in Chapter 4, the much larger range of responses to this question seems to have impacted the quality of the data overall for the purposes of this study.

```

          Df Pillai approx F num Df den Df      Pr(>F)
Country   3  2.4431   23.885    27   147 < 2.2e-16 ***
Residuals 55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 5.14- MANOVA Summary for Question D78

The MANOVA test results are, as expected, not very clear. While the threshold for significance is comfortably passed with a P-value of $<2.2e-16$, and a higher F-value than all three other tests, the Pillai value invalidates the results. A value closer to 1 is considered to be a sign of variances of the dependant variables but with a Pillai result of 2.44, it must be concluded that this question does not support the hypotheses.

When looking at the summary ANOVA scores, it can be seen that each response is considered significant and that they each have high Sum of Square Means result. There is no clear reason to be able to reject this data overall apart from its less robust structure overall:

```

Response TOTAL :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 2827630  942543 124.75 < 2.2e-16 ***
Residuals 55 415541    7555
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Very Positive :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3  59093 19697.8  23.246 7.544e-10 ***
Residuals 55 46605   847.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Fairly Positive :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 597215 199072  66.734 < 2.2e-16 ***
Residuals 55 164068   2983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Neutral :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 753254 251085  131.2 < 2.2e-16 ***
Residuals 55 105260   1914

```

Figure 5-14- Summary of ANOVA for 4 of the 9 responses to Question D78

```

Response Fairly Negative :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 332391 110797  63.917 < 2.2e-16 ***
Residuals 55  95340   1733
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Very Negative :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 107901  35967  74.988 < 2.2e-16 ***
Residuals 55 26380    480
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Not sure :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 4223.2 1407.74  9.3481 4.32e-05 ***
Residuals 55 8282.6  150.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Total Positive :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country  3 678438 226146  47.973 2.275e-15 ***
Residuals 55 259270   4714
---

```

Figure 5-6- Summary of ANOVA for the next 4 of the 9 responses to Question D78

```

Response Total Negative :
      Df Sum Sq Mean Sq F value    Pr(>F)
Country   3 772275  257425  78.994 < 2.2e-16 ***
Residuals 55 179234    3259
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

1 observation deleted due to missingness

```

Figure 5-15- Summary of ANOVA for the last of the 9 responses to Question D78

The summary ANOVA results present a seemingly clear picture of strongly significant results with high Sum of Square Means values and some of the highest F-Values overall. It seems that the data is well-constructed, and no one result indicates that the results can be rejected.

In conclusion, it seems that the question D78, while on less robust statistical footing as explored in Chapter 4, does not support the hypothesis that nations targeted by disinformation can have that change measured in their responses to Eurobarometer surveys.

5.5. Machine Learning Results

As discussed in previous chapters, a k-NN approach would be used as a complimentary approach in answering the research question. The approach would be taken on a question-by-question basis. There were, however, some limitations to adapting the currently constituted data, designed in the main to function as data for a MANOVA test, to data used for Machine Learning. The main limitation would be the small number of available data points in the data. Each available dataframe was an amalgamation of each of four member state’s responses to one of four Eurobarometer questions that related to the research question. This results question dataframes of, at most, 60 rows divided equally by member states. While a wide range of data is currently still available in the Eurobarometer data, only the four member states chosen for this study have had their data comprehensively examined in order to confirm their validity in answering the research question. This limitation will be discussed in more detail in Chapter 6.

The data had already been pre-processed and split into training and test data in previous steps explored in Chapter 4. This was the structure of the data:

TOTAL	Tend to Trust	Trust %	Tend not to Trust	Not Trust %	Not Sure	DK	Country
1.5433230	0.62525519	-0.439432203	1.03038513	0.3213821	1.28395054	0.31711680	Germany
1.5174843	-0.54994394	-1.414307006	2.23847547	1.7497468	0.08169193	-0.48402037	Germany
1.7069680	-0.48121885	-1.414307006	2.17048770	1.4640738	0.95101739	0.11683250	Germany
1.5906939	1.46370136	0.269567654	0.61199887	-0.2499638	0.72906195	-0.08345179	Germany
1.3495329	1.56678900	0.446817618	0.46556367	-0.2499638	-0.06627836	-0.48402037	Germany
1.3925974	1.77296429	0.624067582	0.17269329	-0.6308611	0.58109166	-0.08345179	Germany
1.4830327	1.58053402	0.446817618	0.65383749	-0.1547395	-0.19575236	-0.68430467	Germany
1.6337584	1.65613162	0.358192636	0.75320423	-0.0595152	-0.10327093	-0.68430467	Germany
1.4615005	1.45682885	0.269567654	0.67475681	-0.0595152	-0.01078950	-0.48402037	Germany
1.7586454	1.69049417	0.358192636	0.76889372	-0.1547395	0.28515108	-0.48402037	Germany
0.4753241	-1.01727459	-1.325682024	0.94670788	0.9879523	1.43192083	0.91796968	UK
0.6734207	-0.79048178	-1.237057042	1.34417483	1.3688495	0.26665479	-0.08345179	UK
0.6346627	-0.71488417	-1.148432060	1.09837290	1.0831766	0.76605452	0.31711680	UK
0.5959046	-0.95542201	-1.325682024	1.28141689	1.3688495	0.59958795	0.11683250	UK
0.6088240	-0.79735429	-1.237057042	1.17159050	1.1784009	0.59958795	0.31711680	UK
-0.7778520	-1.38839011	-1.148432060	-0.02081036	0.7975036	0.47011394	0.91796968	UK
-0.7046424	-1.51896779	-1.325682024	0.20930209	1.0831766	0.32214365	0.71768538	UK
-0.7993843	-1.57394786	-1.325682024	0.12562483	1.0831766	0.35913623	0.71768538	UK
-0.7563198	-1.06538216	-0.705307149	0.56493041	1.8449711	-2.37831414	-2.28657901	UK
-0.4720943	-0.63241406	-0.439432203	0.59630938	1.5592981	-2.43380300	-2.28657901	UK

Figure 5-16- Example training data after pre-processing and a training and test split.

The above data is the training data for QA8 after pre-processing and the training and test split at 80/20. This left a dataframe of 40 rows and 8 columns. The test data was similarly constructed and resulted in a dataset of 8 rows and 8 columns.

TOTAL	Tend to Trust	Trust %	Tend not to Trust	Not Trust %	Not Sure	DK	Country
1.4442747	0.3228648	-0.61668217	1.4226223	0.7975036	0.3036474	-0.28373608	Germany
1.2547910	1.3674862	0.35819264	0.3975759	-0.2499638	0.3036474	-0.28373608	Germany
0.5097757	-1.4914778	-1.85743192	1.5219890	1.7497468	0.8215434	0.51740109	UK
0.4537919	0.4534425	0.09231769	0.8839499	0.9879523	-2.4153067	-2.28657901	UK
-0.8166100	-0.1925734	0.44681762	-0.6327003	-0.2499638	-0.7321447	-0.48402037	IE
-0.7864649	0.4465699	1.24444246	-1.1870621	-1.2974313	-0.3807152	-0.08345179	IE
-0.7347875	-0.2269360	0.26956765	-0.6536196	-0.3451881	-0.2327449	0.11683250	PL
-0.8079972	-0.1582109	0.44681762	-0.8889619	-0.7260854	0.1001882	0.51740109	PL

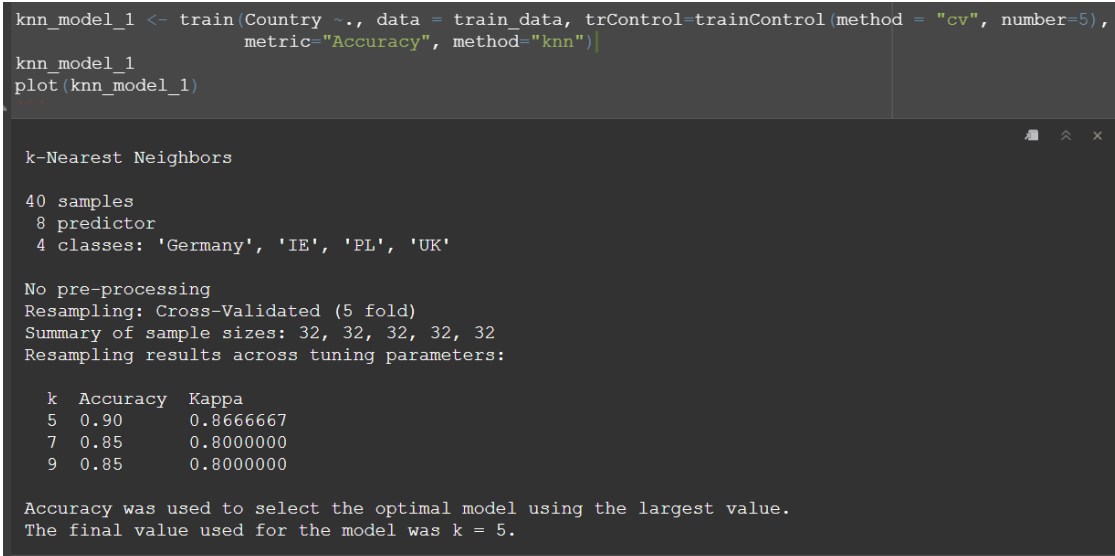
Figure 5-17 - Example test data after pre-processing and a training and test split.

Other dataframes were of similar length and construction apart from D78, which was noticeably wider due to the higher number of possible responses. The dataframes used were limited in scope due to the small size of the original dataframes and this was evident in the eventual results.

Question QA8: ‘I would like to ask you a question about how much trust you have in certain media and institutions. For each of the following media and institutions, please tell me if you tend to trust it or tend not to trust it. The EU’

After collecting, pre-processing and splitting the data as discussed in chapter 4, we were ready to conduct our test:

```
knn_model_1 <- train(Country ~., data = train_data, trControl=trainControl(method = "cv", number=5),
metric="Accuracy", method="knn")
knn_model_1
plot(knn_model_1)
```



```
k-Nearest Neighbors
40 samples
8 predictor
4 classes: 'Germany', 'IE', 'PL', 'UK'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 32, 32, 32, 32, 32
Resampling results across tuning parameters:

k Accuracy Kappa
5 0.90 0.8666667
7 0.85 0.8000000
9 0.85 0.8000000

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Figure 5.20- k-NN Results for Question QA8

The 40 samples and 8 predictors match our row and column numbers and the four factor levels as used in the MANOVA test are consistent. Pre-processing did occur previous to this step as explained in Chapter 4, there was no pre-processing included when the above command was run.

Cross-validation has been included with five attempts included to increase the robustness of the results overall. The summary of sample sizes confirms that 32 examples were used in each subsequent test. The test finds that k=5 has the highest accuracy at 0.86 compared to 0.8 for larger k values.

```

test_data <- test_data[complete.cases(test_data),] # To remove the NA values
test_data$Country <- as.factor(test_data$Country)
predict_knn_1 <- predict(knn_model_1, test_data)
confusionMatrix(predict_knn_1, test_data$Country, positive = "Germany")

```

Figure 5.21- Code block run to conduct the test

Due to some missing values found in the data when attempting to conduct this test with other questions, a step was added to forceable remove all missing values, a second step was included to confirm that ‘Country’ was correctly set as a factor. Finally, the test was applied to the test data and a Confusion Matrix was produced. The above steps are repeated for all other questions, following the above example, only results will be presented here.

```

Confusion Matrix and Statistics

      Reference
Prediction Germany IE PL UK
Germany      2  0  0  1
IE           0  2  0  0
PL           0  0  2  0
UK           0  0  0  1

Overall Statistics

      Accuracy : 0.875
      95% CI   : (0.4735, 0.9968)
      No Information Rate : 0.25
      P-Value [Acc > NIR] : 0.0003815

      Kappa : 0.8333

      Mcnemar's Test P-Value : NA

Statistics by Class:

      Class: Germany Class: IE Class: PL Class: UK
Sensitivity          1.0000    1.00    1.00    0.5000
Specificity          0.8333    1.00    1.00    1.0000
Pos Pred Value       0.6667    1.00    1.00    1.0000
Neg Pred Value       1.0000    1.00    1.00    0.8571
Prevalence           0.2500    0.25    0.25    0.2500
Detection Rate       0.2500    0.25    0.25    0.1250
Detection Prevalence 0.3750    0.25    0.25    0.1250
Balanced Accuracy    0.9167    1.00    1.00    0.7500

```

Figure 5.22 - k-NN prediction results for Question QA8

This data results in an overall accuracy of 0.87 with confidence interval values of 0.47 and 0.99. The P-value is comfortably significant and a high Kapp value suggest good predicative capability. However, the results per class are suspicious, the cleanliness of the results overall with a wide spread of 1.0 values suggests that this

data may not be of sound statistical merit. This is a function of the limitations of the data as currently constituted, and of the test design in this iteration.

Question D71: ‘When you get together with friends or relatives, would you say you discuss frequently, occasionally, or never about...? European Politics’

Question D71 has the same pre-processing and split decisions applied to it as QA8 but there are some differences in the k values:

```
47 samples
10 predictors
 4 classes: 'Germany', 'IE', 'PL', 'UK'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 38, 38, 36, 39, 37
Resampling results across tuning parameters:

 k Accuracy  Kappa
 5 0.8991919 0.8674427
 7 0.8628283 0.8171307
 9 0.8428283 0.7904275

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Figure 5.23 - k-NN results for Question Q71

Of note is the higher number of sample sizes, as well as the perhaps more realistic accuracy statistics that result from the test. There are more predicative columns in this data which may explain the more robust results. Similar to QA8, a k-value of 5 is determined to be the most accurate.

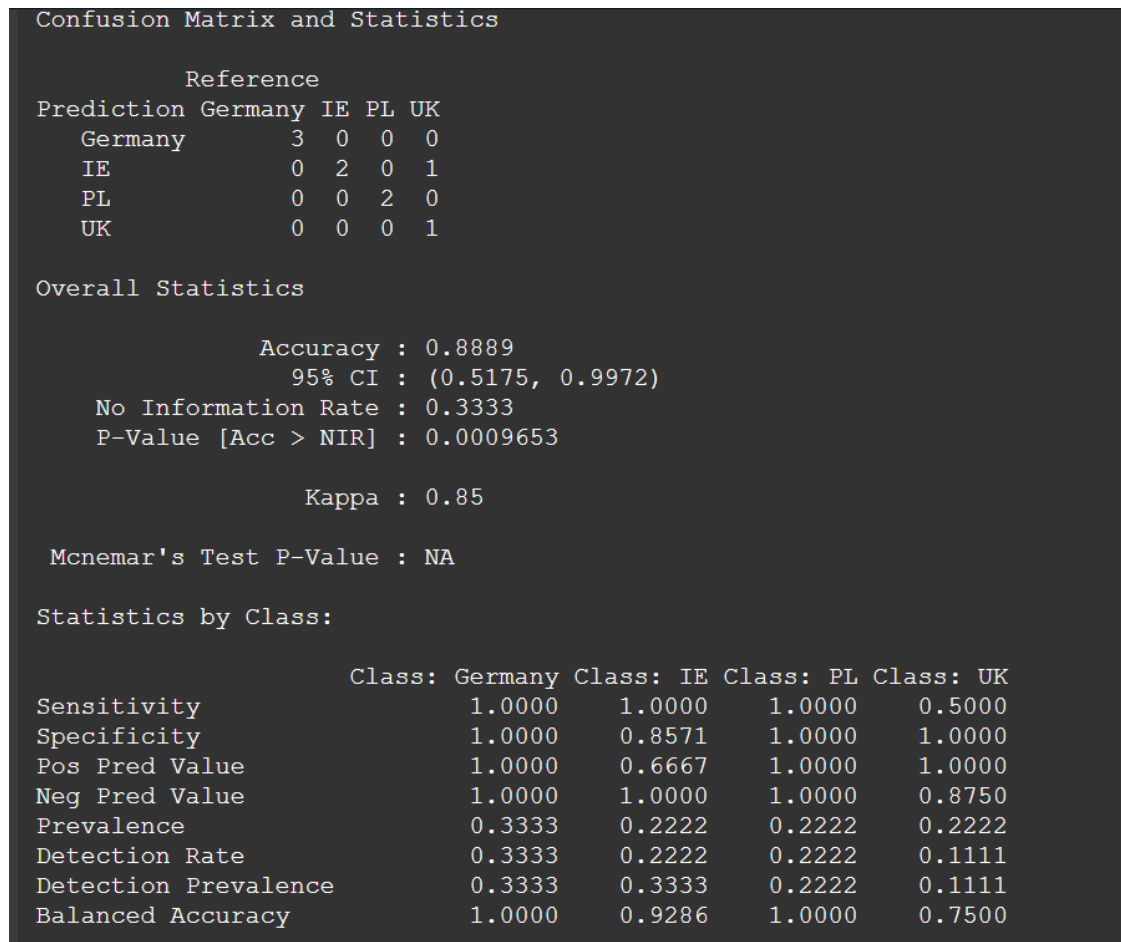


Figure 5.24 - k-NN prediction results for Question D71

The results are similar overall to QA8 with an accuracy of 0.88 and a significant p-value with a high Kappa score. However, once again the results for statistics by class call the results into question overall. This test does not seem robust and is similarly hampered by the current structure of the data and the Machine Learning design.

Question D73: ‘At the present time, would you say that, in general, things are going in the right direction or in the wrong direction, in...? The EU’

Question D73 has the same pre-processing and split decisions applied to it as previous questions but there are some differences in the k values:

```

48 samples
 9 predictor
 4 classes: 'Germany', 'IE', 'PL', 'UK'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 39, 38, 39, 39, 37
Resampling results across tuning parameters:

 k Accuracy  Kappa
 5  0.6854545 0.5760440
 7  0.7236364 0.6251086
 9  0.6454545 0.5250101

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 7.

```

Figure 5.25-k-NN prediction results for Question D73

The accuracy and Kappa overall declines with this test. Alternate sample sizes can also be seen in the Cross-validation steps. The eventual model is set as k = 7 with an accuracy of 0.72.

```

Reference
Prediction Germany IE PL UK
Germany      3  0  0  0
IE           0  1  0  0
PL           0  1  2  0
UK           0  0  1  3

Overall Statistics

Accuracy : 0.8182
95% CI : (0.4822, 0.9772)
No Information Rate : 0.2727
P-Value [Acc > NIR] : 0.0002617

Kappa : 0.7528

McNemar's Test P-Value : NA

Statistics by Class:

Class: Germany Class: IE Class: PL Class: UK
Sensitivity      1.0000  0.50000  0.6667  1.0000
Specificity      1.0000  1.00000  0.8750  0.8750
Pos Pred Value   1.0000  1.00000  0.6667  0.7500
Neg Pred Value   1.0000  0.90000  0.8750  1.0000
Prevalence       0.2727  0.18182  0.2727  0.2727
Detection Rate   0.2727  0.09091  0.1818  0.2727
Detection Prevalence 0.2727  0.09091  0.2727  0.3636
Balanced Accuracy 1.0000  0.75000  0.7708  0.9375

```

Figure 5.26- k-NN prediction results for Question D73

As with other tests, the overall results seem robust in that an accuracy of 0.81 with a significant p-value and a high Kappa are found. However, looking at the statistic by Class we again see that there are some unexpected results. The one for one sensitivity and specificity is in itself essentially an invalidation of the results but to have it sprinkled throughout the results quite so liberally future suggests that this data and the k-NN design are not suitable for this research question.

Question D78: ‘In general, does the EU conjure up for you a very positive, fairly positive, neutral, fairly negative or very negative image?’

Question D78 has, as discussed in Chapter 4, less robust data overall than the other questions that were utilised in attempting to answer the research question. As explained previously, the larger number of possible replies to the question is assumed to have caused problems with the statistics overall. The same was predicted to be the case when it k-NN test. Question D78 has the same pre-processing and split decisions applied to it as previous questions but there are some differences in the k values:

```
k-Nearest Neighbors
34 samples
16 predictors
 4 classes: 'Germany', 'IE', 'PL', 'UK'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 29, 28, 27, 26, 26
Resampling results across tuning parameters:

  k  Accuracy  Kappa
  5  0.975    0.9652174
  7  0.975    0.9652174
  9  0.975    0.9652174

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
```

Figure 5.27- k-NN prediction results for Question D78

This question, with the least robust data overall. A k-value of 9 was found to have a prediction value of 0.97, however given the robustness challenges of this question, the outcome must be discarded.

5.6. Key Reflections

The MANOVA testing provided robust support overall for the research questions and can comfortably be said to be in support of the hypothesis that EU member states that have been the target of disinformation can be seen to have their attitudes change as measured in the Eurobarometer survey data between the years 2015 and 2022. With a few exceptions in the case of a minimum of the response variables per question, the bulk of the data tested was found to be in support of the hypothesis.

Additionally, the Machine Learning approach was found to be restricted in its applicability due to the nature of the datasets available rather than the unsuitability of Machine Learning as an effective tool to answer or illustrate the research question. As it stands, the results cannot be said to have significant validity to this study in their current inception.

6. CONCLUSIONS AND FUTURE WORK

6.1. Introduction

The goal of this research was to explore whether or not it would be possible to find any statistically significant relationships between nations that have been the target of disinformation and their change in attitude towards the EU, therefore, in this chapter the implications of that research will be explored, as well as the conclusions that can be reached based on the current experiment design. Additionally, any limitations of the study design will be explored and considered, and finally, any further potential experiments based on the data already available will also be explained as well as any changes that would have been made in hindsight given the results following the end of the project.

6.2. Problem Definition

The formulation of this project began with the discovery of the EUvsDisinformation database in early 2022. A deeper dive into the academic work in the area of disinformation and propaganda followed. In that background research it was found that the majority of the work in this area is based on the freely available (at the time of writing) Twitter API³⁴. The largest part of the academic work was devoted more to the application of complex mathematical approaches to probabilistic models of identifying disinformation in tweets in real time. While the background reading was taking place, the January 6th coup attempt took place in the United States which led to the inception of the research question. Rather than focusing on the tweets or other social media that was being spread leading to events such as January 6th, could the effects of these disinformation and propaganda campaigns be seen in the attitudes of the public being targeted by these campaigns?

Overall, the research question has been robustly tested utilising the data in this study. The conclusions based on the exploration of the data, the statistical analysis of the data in preparation for testing, the testing itself and the Machine Learning

³⁴ <https://developer.twitter.com/en/docs/twitter-api>

approaches attempted; suggest that this data does support the research hypothesis that there is a relationship between the nations targeted with disinformation and those whose attitudes changed regarding the EU. It is worth noting that there a minority of the utilised questions were not in support of the hypothesis, particularly question D78. However, as discussed in Chapters 3 and 4, that data was not particularly robust, and this may have been potentially due to the higher number of available responses to the survey question.

The Machine learning approaches did not offer significant additional information and a searching processing for additional datasets, as well as much greater focus on wrangling the data and overall study design would be needed to be able to ensure the validity of the Machine Learning results with the current data. While the preliminary results of those tests that seemed to be designed well enough to have any validity do seem to support the hypothesis, the results were of low quality, and, therefore, cannot be considered of an equal footing as the rest of the work.

It seems reasonable at the conclusion of this work to reject the null hypothesis given the overlapping and consistent results seen in Chapter 5 when examining the four MANOVA and individual ANOVA results that were explored. There is solid evidence that there is a correlation between a member state's exposure to disinformation and that member state's change in attitudes towards the EU and its' institutions as measured in the Eurobarometer survey between the years of 2015 and 2022.

6.3. Limitations

Apart from the limitations of the Machine Learning approach as explained in Chapter 5, there are some limitations of the current study design. Firstly, the selection of the specific questions, while not random, was an influencing factor on the results of the study. Given that one of the questions was on a lower statistical footing when examined in detail, it is possible that a different selection of questions would have generated different results. While data for Ireland was included as a control group, this was not as robust as selecting an equal number of member states as had been targeted

according to the EUvsDisinformation datasets. Had there been a 50-50 split between targeted and non-targeted nations, the results could have been altered significantly.

Secondly, there were several significant and unprecedented global events during the study period that could not be controlled for. While disinformation was undoubtedly an aspect of the social media environment during that period, the effects of the Brexit referendum, of COVID-19 as well as, to a lesser extent, the Trump Presidency, and the end of the Merkel era in Germany, could not be controlled for in this study. Their effects upon national attitudes towards the EU remain unknowable in this data given the current design.

Finally, the scale of the data proved massive. Given more time, or a team of researchers, the complexities and intricacies of the data contained in seven years of Eurobarometer data might be explored in greater detail. As it was, only the three most targeted countries and only the four most significant questions could be used to attempt to illuminate the research question.

6.4 Design Choices

The inclusion of two datasets regarding disinformation was likely unnecessary. The EUvsDisinformation website provides much broader data than is initially visible on the public facing website. Had this data been scraped and wrangled, cleaned and explored, it would have perhaps provided data in much greater detail than was made available by trying to combine the two datasets together. Given that the Kaggle dataset was date limited, its inclusion proved to be costly in terms of time although it was arguably a somewhat independent confirmation of the data as found in the scraped data.

The Eurobarometer data proved extremely ungainly, and the initial research approach was concluded to simply be impossible given the time available. As explored in Chapter 4, the selection of the original questions as well as the cross tabulation of those question across all Eurobarometer years followed by the loading, wrangling and cleaning of the four questions that were used of the original 44 that had been selected; consumed the bulk of the time spent on this study by far. A more conservative selection of questions to focus only on those that directly had an influence on the research question would potentially have allowed a much greater examination of the selected questions in depth.

Questions such as QA2a.7 (*‘What are your expectations for the next twelve months: will the next twelve months be better, worse or the same, when it comes to...? The European Economy’*.) had previously been selected, loaded into R, cleaned and prepared for further study only to then be rejected as the size of the data proved too great for the time available. Ultimately, those questions remain viable for future work based on this data, but they proved a significant time sink when measured in terms of their application to the research question.

There were two rejected approaches to the Eurobarometer data that could have proven fruitful. Firstly, the inclusion of an EU average based on the ‘UE28 EU28’ values on each Eurobarometer question would have illuminated larger scale trends and would have increased the robustness of the study overall. Secondly, an amalgamation of several nations with similar levels of disinformation into groups of three or four member states could have provided further insights into the implications of the data overall. Ultimately it was decided to use a single nation (Ireland) as the control group based purely on its not having been directly targeted with disinformation according to the EUvsDisinformation data, though many different member states fall into that category and a different selection could have drastically changed the results. Conversely, the choice of a member state that is not as familiar to the author may have meant that significant events that may have affected the data may not have been known about. In 2017 for example there was an election in France, it would be interesting to see if this affected French attitudes towards the EU.

6.5 Future Work

The potential future work and possible applications of this data is extensive. The collection of Eurobarometer questions available is extremely broad and far reaching compared to the four questions that were utilised for this study. A wealth of data remains untouched and unexplored. Future work could easily include further questions in order to explore in greater detail the effects of disinformation upon member states’ attitudes.

Apart from the sheer number of available questions, a more comprehensive amalgamation of member states into clusters could be extremely interesting. There are any number of approaches that this data is currently capable of facilitating. For

example, do geographic regions have similar changes over time? Does disinformation targeting nations with similar linguistic structure such as Nordic languages or Latin languages tend to have similar changes in their attitudes? Do nations of similar GDP or population size have similar changes in their attitudes?

Additionally, the EUvsDisinformation data is currently ideally suited to examine whether there a noticeable and meaningful increase in disinformation targeted at an EU Member State before sensitive referenda, and a similar decline in disinformation efforts following the referenda as seen by Howard and Kollanyi (2016).

This is also a potentially useful Machine Learning approach. Can clusters of nations be effectively and accurately identified using this data and do those clusters attitude's change with one another over time? Additionally. comparing Supervised and Unsupervised Learning approaches may prove interesting in exploring the data.

6.6. The Necessity of this Work

Although most academic research has been focused on the identification and classification of tweets, very little work has tended to have focused on the Macro scale implications of disinformation and propaganda. This needs to change, while identification of disinformation is undoubtedly essential, a greater understanding of the effects of these efforts would significantly assist efforts to mitigate their effects. The identification of disinformation is not the overarching issue in terms of its detrimental effects, rather it is the insidious injection of doubt into the national and international discourse that is the most dangerous aspect. Those who stormed the Capitol building or who refused to wear a mask or feels themselves to be a sovereign citizen to whom the law does not apply, will not be swayed by having their tweets flagged as potentially harmful. But a deeper understanding of how disinformation and propaganda effects national discourse will go a long way to help fight against the effects of such campaigns. There was once a time where the most disingenuous ideas available only reached as far as the strength of a single person's voice, but those messages have been adapted, packaged and constitute a targeted attack upon the world of western democracies. Bad ideas do not wither and die in a social media eco-system designed to maximise user engagement in order to extend their exposure to advertising. Rather that

creates a witch's brew wherein those nations that seek to disrupt our way of life can easily disrupt the fragile health of our democracies.

It is hoped that this work in some way helps illustrate how effective these efforts can be. It behoves us all to be careful of what we consume and share. We are the vector for this intellectual disease.

Thank you for your time.

BIBLIOGRAPHY

- Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. <https://doi.org/10.1073/pnas.1804840115>
- Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*, *117*(1), 243–250. <https://doi.org/10.1073/pnas.1906420116>
- Bakshy, E., Messing, S., & Adamic, L. (2015). Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science (New York, N.Y.)*, *348*. <https://doi.org/10.1126/science.aaa1160>
- Bastos, M. T., & Mercea, D. (2019a). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38–54. <https://doi.org/10.1177/0894439317734157>
- Bastos, M. T., & Mercea, D. (2019b). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38–54. <https://doi.org/10.1177/0894439317734157>
- Bingle, Y., Burke, W., Harrack, M. A., Blankenship, L., Khoanam, N., Sokol, C., & Tabatabaei, S. S. (2020). Internet research agency's campaign to influence the U.S. 2016 elections: Assessing linguistic profiles via statistical analysis. *Journal of Computing Sciences in Colleges*, *36*(3), 31–42.

- Boldureanu, D., Roman, I., Sardaru, D., & Andruseac, G. G. (2020). Romanian Citizens' Attitudes and Opinions over the Course of the Covid-19 Pandemic. *2020 International Conference on E-Health and Bioengineering (EHB)*, 1–4. <https://doi.org/10.1109/EHB50910.2020.9280207>
- Bruno, M., Lambiotte, R., & Saracco, F. (2021). Brexit and bots: Characterizing the behaviour of automated accounts on Twitter during the UK election. *ArXiv:2107.14155 [Physics]*. <http://arxiv.org/abs/2107.14155>
- Chatfield, A. T., Reddick, C. G., & Brajawidagda, U. (2015). Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. *Proceedings of the 16th Annual International Conference on Digital Government Research*, 239–249. <https://doi.org/10.1145/2757401.2757408>
- Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.
- De keersmaecker, J., & Roets, A. (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, 65, 107–110. <https://doi.org/10.1016/j.intell.2017.10.005>
- DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R. C., Fox, R., Albright, J., & Johnson, B. (2018, December 17). *The tactics & tropes of the Internet Research Agency*. <https://www.semanticscholar.org/paper/The-tactics-%26-tropes-of-the-Internet-Research-DiResta-Shaffer/ccdf2c98b9a5372305d5075f6bea0a298058c287>
- Douven, I., & Hegselmann, R. (2021). Mis- and disinformation in a bounded confidence model. *Artificial Intelligence*, 291, 103415. <https://doi.org/10.1016/j.artint.2020.103415>

- Dutta, U., Hanscom, R., Zhang, J. S., Han, R., Lehman, T., Lv, Q., & Mishra, S. (2021a). Analyzing Twitter Users' Behavior Before and After Contact by the Russia's Internet Research Agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–24. <https://doi.org/10.1145/3449164>
- Dutta, U., Hanscom, R., Zhang, J. S., Han, R., Lehman, T., Lv, Q., & Mishra, S. (2021b). Analyzing Twitter Users' Behavior Before and After Contact by the Russia's Internet Research Agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–24. <https://doi.org/10.1145/3449164>
- Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1), Article 1. <https://doi.org/10.1038/s41467-022-35576-9>
- Ecker, U. K. H., Lewandowsky, S., Cheung, C. S. C., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of misinformation. *Journal of Memory and Language*, 85, 101–115. <https://doi.org/10.1016/j.jml.2015.09.002>
- Endres, K., & Panagopoulos, C. (2019). Cross-Pressure and Voting Behavior: Evidence from Randomized Experiments. *The Journal of Politics*, 81(3), 1090–1095. <https://doi.org/10.1086/703210>
- Farkas, J., & Bastos, M. (2018). IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News. *Proceedings of the 9th International Conference on Social Media and Society*, 281–285. <https://doi.org/10.1145/3217804.3217929>
- Ferrara, E. (2017). Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election. *First Monday*. <https://doi.org/10.5210/fm.v22i8.8005>

- Forelle, M., Howard, P. N., Monroy-Hernandez, A., & Savage, S. (2015). *Political Bots and the Manipulation of Public Opinion in Venezuela* (SSRN Scholarly Paper No. 2635800). <https://doi.org/10.2139/ssrn.2635800>
- Garrett, R. K. (2011). Troubling Consequences of Online Political Rumoring. *Human Communication Research*, 37(2), 255–274. <https://doi.org/10.1111/j.1468-2958.2010.01401.x>
- Garrett, R. K., & Weeks, B. E. (2013). The promise and peril of real-time corrections to political misperceptions. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1047–1058. <https://doi.org/10.1145/2441776.2441895>
- Glenski, M., Weninger, T., & Volkova, S. (2020). How humans versus bots react to deceptive and trusted news sources: A case study of active users. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 654–661.
- Great Britain & Intelligence and Security Committee. (2020). *Russia: Presented to Parliament pursuant to section 3 of the Justice and Security Act 2013*. <https://assets.documentcloud.org/documents/6999013/20200721-HC632-CCS001-CCS1019402408-001-ISC.pdf>
- Haggag, O., Haggag, S., Grundy, J., & Abdelrazek, M. (2021). COVID-19 vs social media apps: Does privacy really matter? *Proceedings of the 43rd International Conference on Software Engineering: Software Engineering in Society*, 48–57. <https://doi.org/10.1109/ICSE-SEIS52602.2021.00014>
- Hahn, U., Hansen, J. U., & Olsson, E. J. (2020). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, 197(4), 1511–1541. <https://doi.org/10.1007/s11229-018-01936-6>

- Hall Jamieson, K., & Albarracín, D. (2020). The Relation between Media Consumption and Misinformation at the Outset of the SARS-CoV-2 Pandemic in the US. *Harvard Kennedy School Misinformation Review*. <https://doi.org/10.37016/mr-2020-012>
- Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 63–76. <https://doi.org/10.1177/0002716215570866>
- Howard, P., Kollanyi, B., Bradshaw, S., & Neudert, L.-M. (2018). Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States? *ArXiv*. <https://www.semanticscholar.org/paper/Social-Media%2C-News-and-Political-Information-during-Howard-Kollanyi/eb70e45a5f4b74edc1e1fdfa052905184daf655c>
- Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (n.d.). *The IRA, Social Media and Political Polarization in the United States, 2012-2018*. 48.
- Howard, P. N., & Kollanyi, B. (2016). Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2798311>
- Howard, P. N., Kollanyi, B., Bradshaw, S., & Neudert, L.-M. (2018). *Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?* (arXiv:1802.03573). arXiv. <https://doi.org/10.48550/arXiv.1802.03573>
- Howard, P. N., & Kreiss, D. (2010). Political parties and voter privacy: Australia, Canada, the United Kingdom, and United States in comparative perspective. *First Monday*. <https://doi.org/10.5210/fm.v15i12.2975>
- Imhoff, R., Dieterle, L., & Lamberty, P. (2021). Resolving the Puzzle of Conspiracy Worldview and Political Activism: Belief in Secret Plots Decreases Normative but

- Increases Nonnormative Political Engagement. *Social Psychological and Personality Science*, 12(1), 71–79. <https://doi.org/10.1177/1948550619896491>
- Iyengar, S., & Westwood, S. J. (2015). Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science*, 59(3), 690–707. <https://doi.org/10.1111/ajps.12152>
- Karhu, M., Suoheimo, M., & Häkkinen, J. (2022). People's Perspectives on Social Media Use during COVID-19 Pandemic. *Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia*, 123–130. <https://doi.org/10.1145/3490632.3490666>
- Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks. *Social Media + Society*, 1(2), 2056305115610141. <https://doi.org/10.1177/2056305115610141>
- Linville, D. L. (n.d.). *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*. 21.
- Llewellyn, C., Cram, L., Favero, A., & Hill, R. L. (2018). Russian Troll Hunting in a Brexit Twitter Archive. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 361–362. <https://doi.org/10.1145/3197026.3203876>
- Matei, S., Rughinis, C., & Rughinis, R. (2017). Big Data, Old Users, Personal Worlds: A Survey of Challenges and Resistance to Big Data Analytics in the EU. *2017 21st International Conference on Control Systems and Computer Science (CSCS)*, 175–181. <https://doi.org/10.1109/CSCS.2017.31>
- Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., & Quattrociocchi, W. (2014). Collective attention in the age of (mis)information. *ArXiv:1403.3344 [Physics]*. <http://arxiv.org/abs/1403.3344>

- O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2013). The Extreme Right Filter Bubble. *ArXiv:1308.6149 [Physics]*.
<http://arxiv.org/abs/1308.6149>
- Ördén, H. (2022). Securitizing cyberspace: Protecting political judgment. *Journal of International Political Theory*, 18(3), 375–392.
<https://doi.org/10.1177/17550882211046426>
- Pherson, R. H., Mort Ranta, P., & Cannon, C. (2021). Strategies for Combating the Scourge of Digital Disinformation. *International Journal of Intelligence and CounterIntelligence*, 34(2), 316–341. <https://doi.org/10.1080/08850607.2020.1789425>
- Prelog, L., & Bakić-Tomić, L. (2020). The Perception of the Fake News Phenomenon on the Internet by Members of Generation Z. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 452–455.
<https://doi.org/10.23919/MIPRO48935.2020.9245169>
- Prier, J. (2017). Commanding the Trend: Social Media as Information Warfare. *Strategic Studies Quarterly*, 11(4), 50–85.
- Puri, M., Dau, Z., & Varde, A. S. (2021). *COVID and social media*.
<https://doi.org/10.1145/3494825.3494830>
- Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). *Detecting and Tracking Political Abuse in Social Media*.
- Sanovich, S. (2017a). *Computational Propaganda in Russia: The Origins of Digital Misinformation*.
- Sanovich, S. (2017b). *Computational Propaganda in Russia: The Origins of Digital Misinformation*. 26.

- Santagiustina, C., & Warglien, M. (2021). The Unfolding Structure of Arguments in Online Debates: The case of a No-Deal Brexit. *ArXiv:2103.16387 [Cs, Stat]*. <http://arxiv.org/abs/2103.16387>
- Skirka, A., Adamyk, B., Adamyk, O., & Valytska, M. (2020). Trust in the European Central Bank: Using Data Science and predictive Machine Learning Algorithms. *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, 356–361. <https://doi.org/10.1109/ACIT49673.2020.9208857>
- Squire, M. (2021). Monetizing Propaganda: How Far-right Extremists Earn Money by Video Streaming. *13th ACM Web Science Conference 2021*, 158–167. <https://doi.org/10.1145/3447535.3462490>
- Stokel-Walker, C. (2023, February 7). TechScape: Why Twitter ending free access to its APIs should be a ‘wake-up call’. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/07/techscape-elon-musk-twitter-api>
- Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2019). *The Use of Twitter Bots in Russian Political Communication*. 10.
- Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F., Stevens, A., Dekhtyar, A., Gao, S., Hogg, T., Kooti, F., Liu, Y., Varol, O., Shiralkar, P., Vydiswaran, V., ... Hwang, T. (2016). The DARPA Twitter Bot Challenge. *Computer*, 49(6), 38–46. <https://doi.org/10.1109/MC.2016.183>
- Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of Rumors and Misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health*, 53(3), 171–174. <https://doi.org/10.3961/jpmph.20.094>

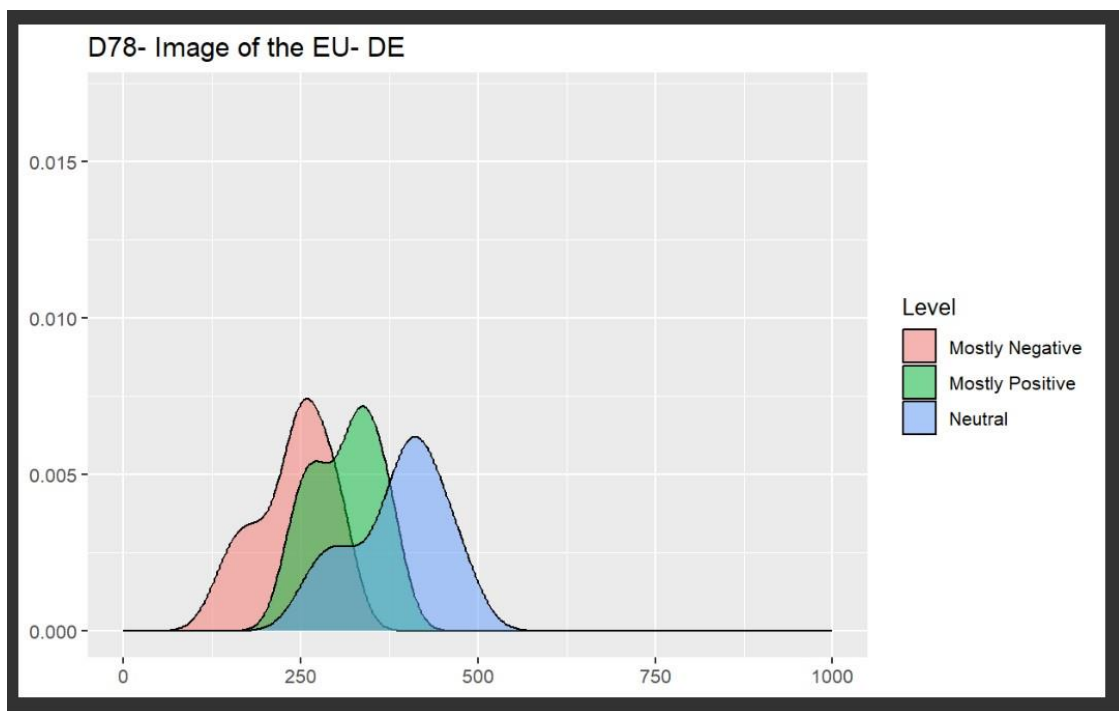
- Thompson, N. (n.d.). How Russian Trolls Used Meme Warfare to Divide America. *Wired*. Retrieved 24 April 2022, from <https://www.wired.com/story/russia-ira-propaganda-senate-report/>
- Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). *Online Human-Bot Interactions: Detection, Estimation, and Characterization* (arXiv:1703.03107). arXiv. <https://doi.org/10.48550/arXiv.1703.03107>
- Venegas-Vera, A. V., Colbert, G. B., & Lerma, E. V. (2020). Positive and negative impact of social media in the COVID-19 era. *Reviews in Cardiovascular Medicine*, 21(4), Article 4. <https://doi.org/10.31083/j.rcm.2020.04.195>
- Volkova, S., & Jang, J. Y. (2018). Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media. *Companion Proceedings of the The Web Conference 2018*, 575–583. <https://doi.org/10.1145/3184558.3188728>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wang, Y., Han, R., Lehman, T., Lv, Q., & Mishra, S. (2021). Analyzing behavioral changes of Twitter users after exposure to misinformation. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 591–598. <https://doi.org/10.1145/3487351.3492718>

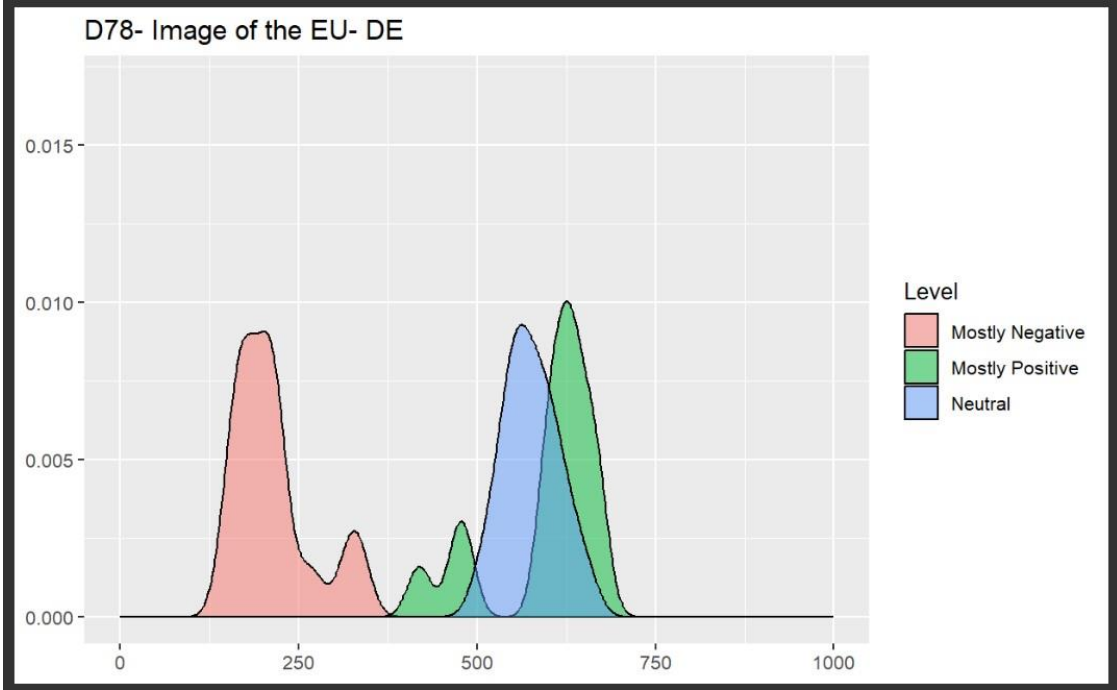
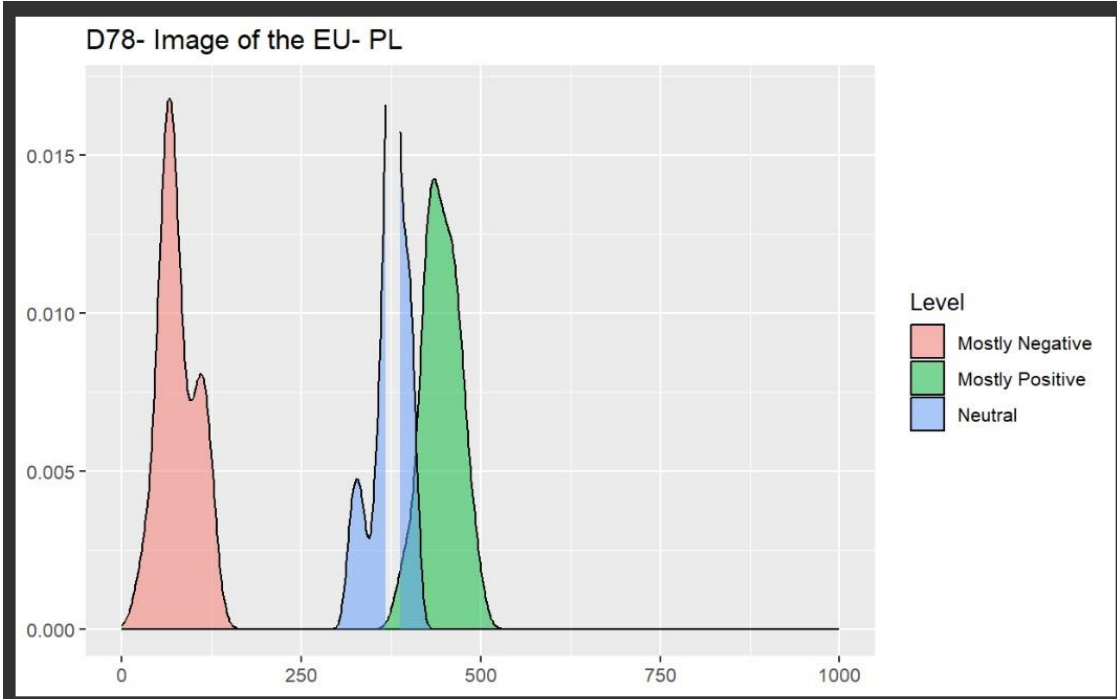
APPENDIX A

In this section, collections of images that were not included in the main body are presented. These images are included to ensure the completeness of the work overall, but they were felt to not be appropriate to the overall needs of the exploration and eventual testing of the data:

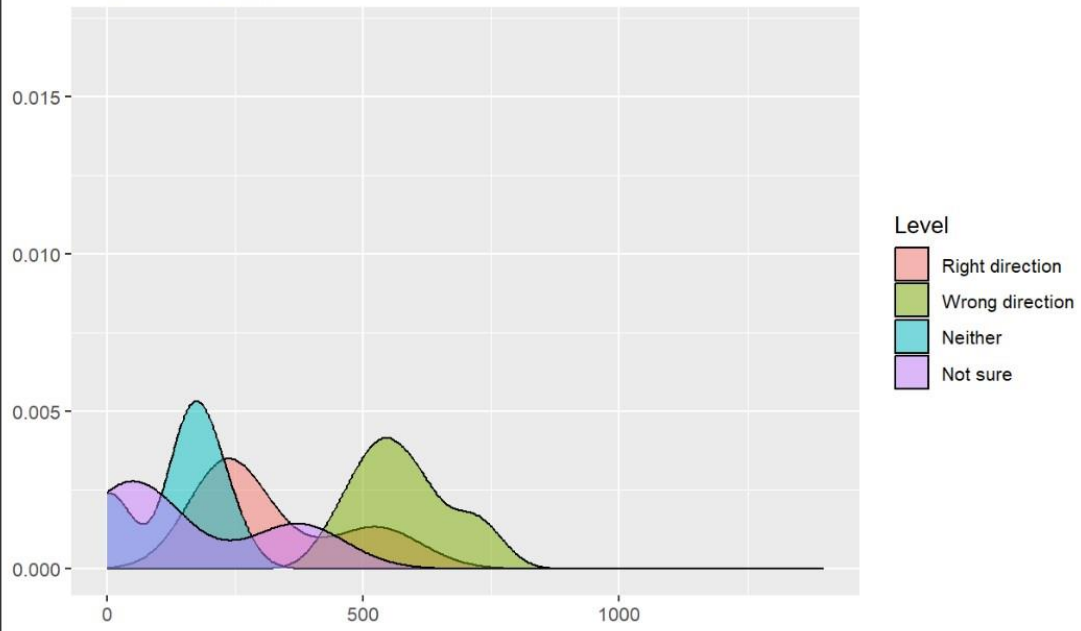
Overlaid Density plots

An overlaid density plotting schema was developed and written but the quality of the images was insufficient to warrant their inclusion.

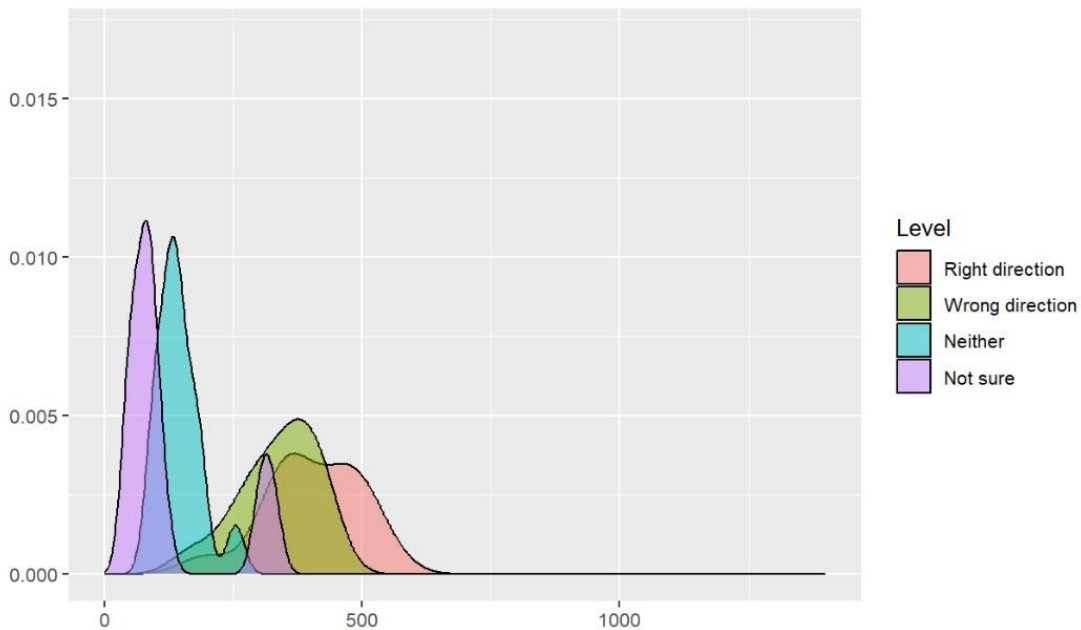


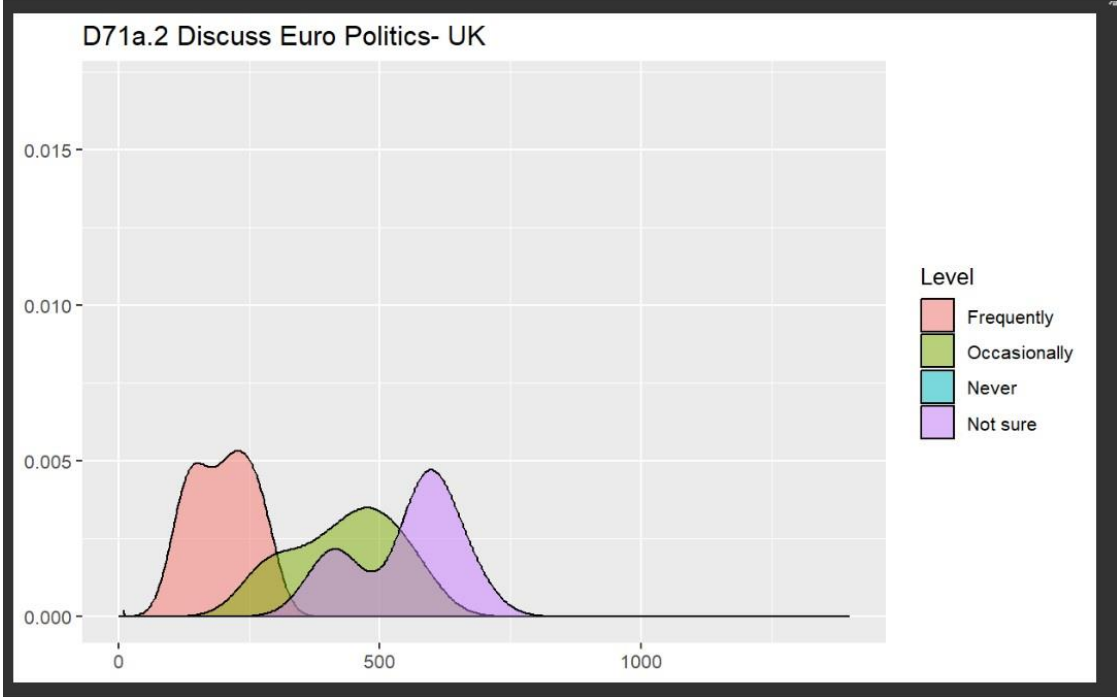
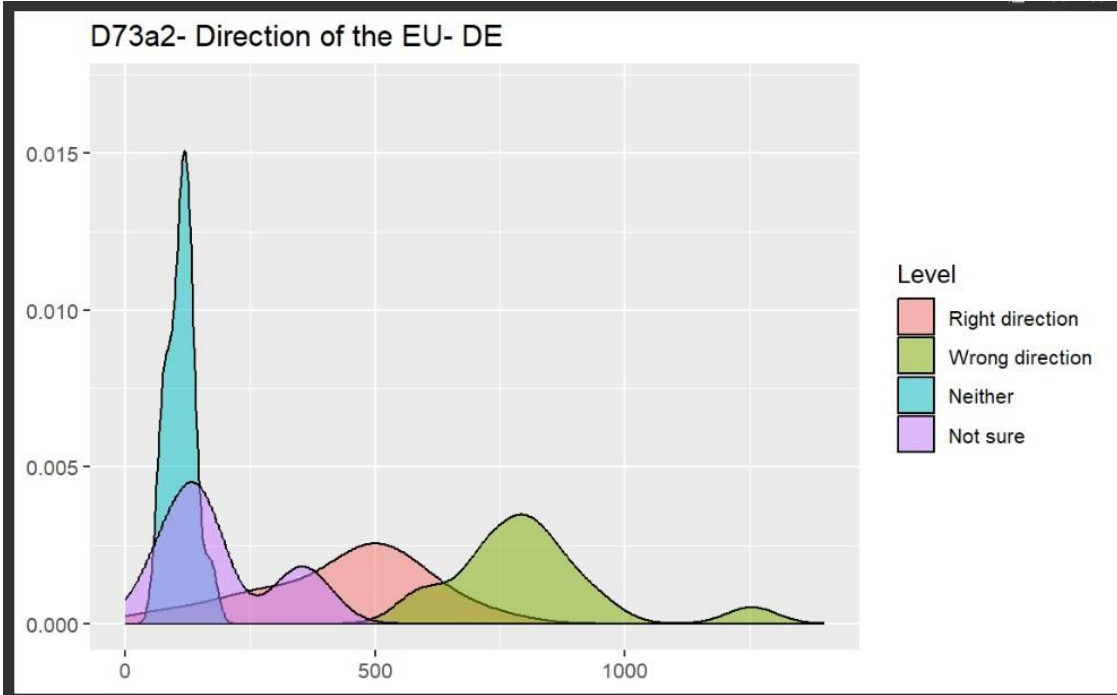


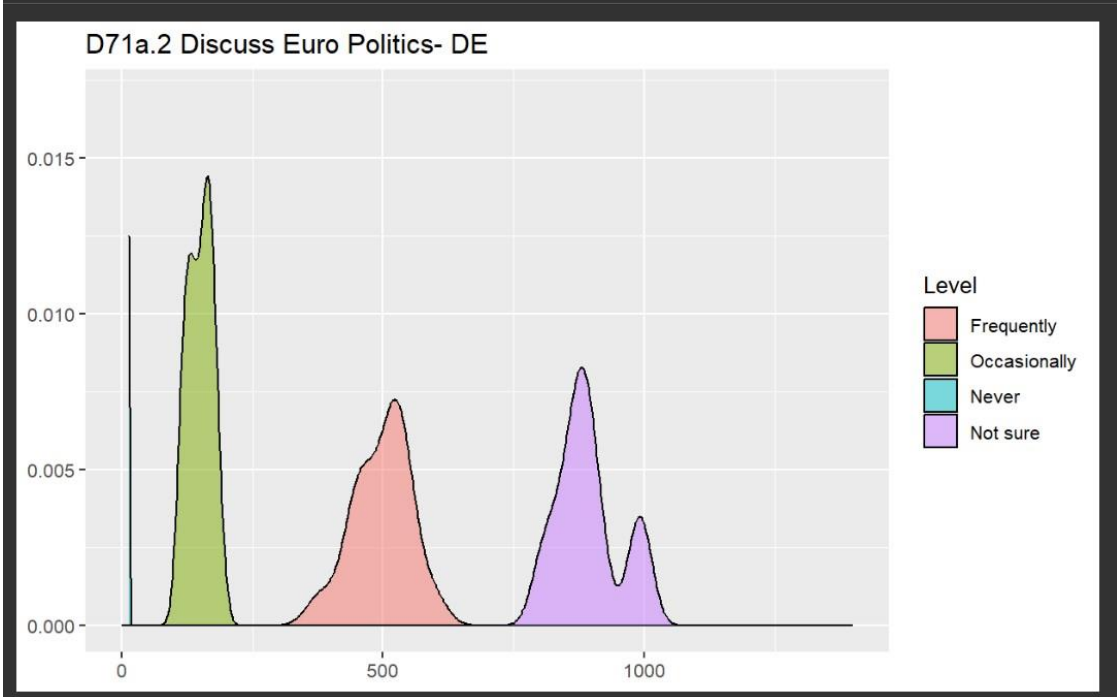
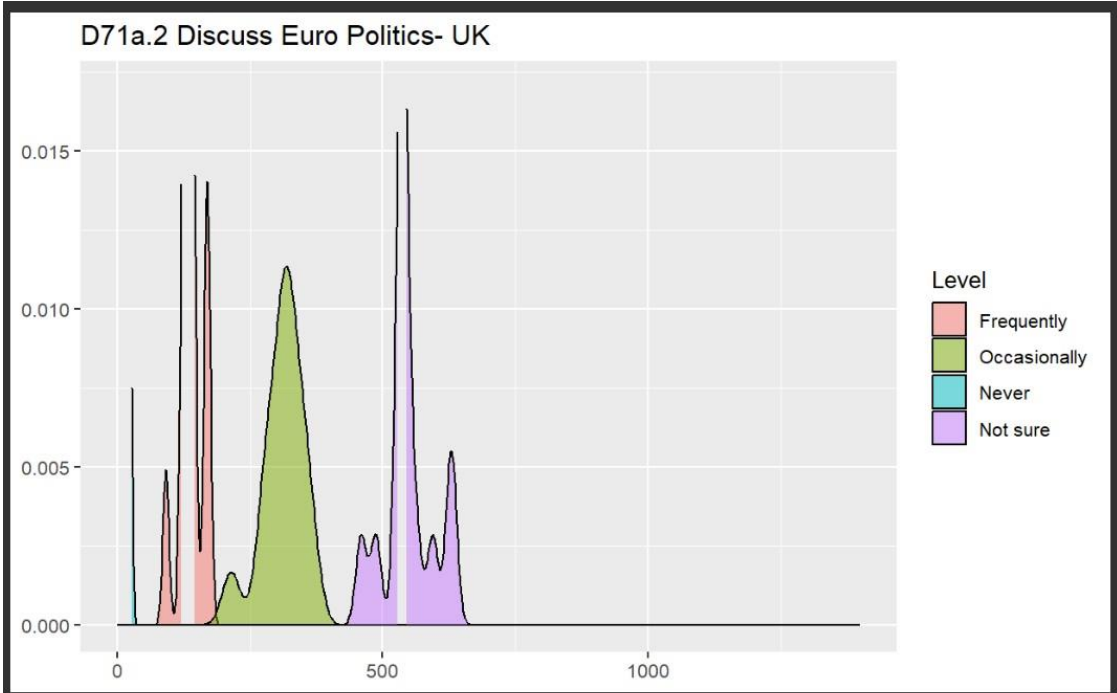
D73a2- Direction of the EU- UK



D73a2- Direction of the EU- POL

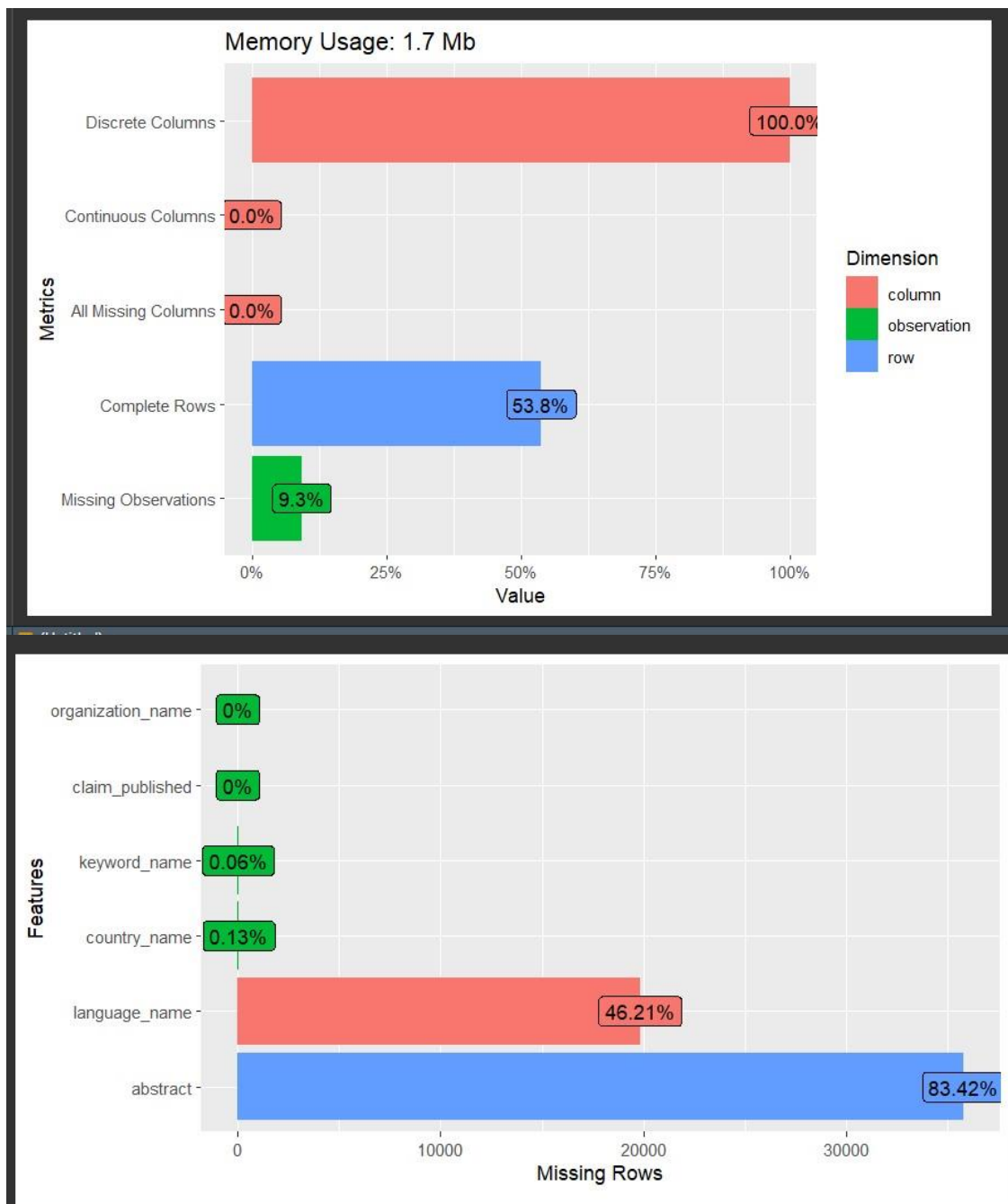




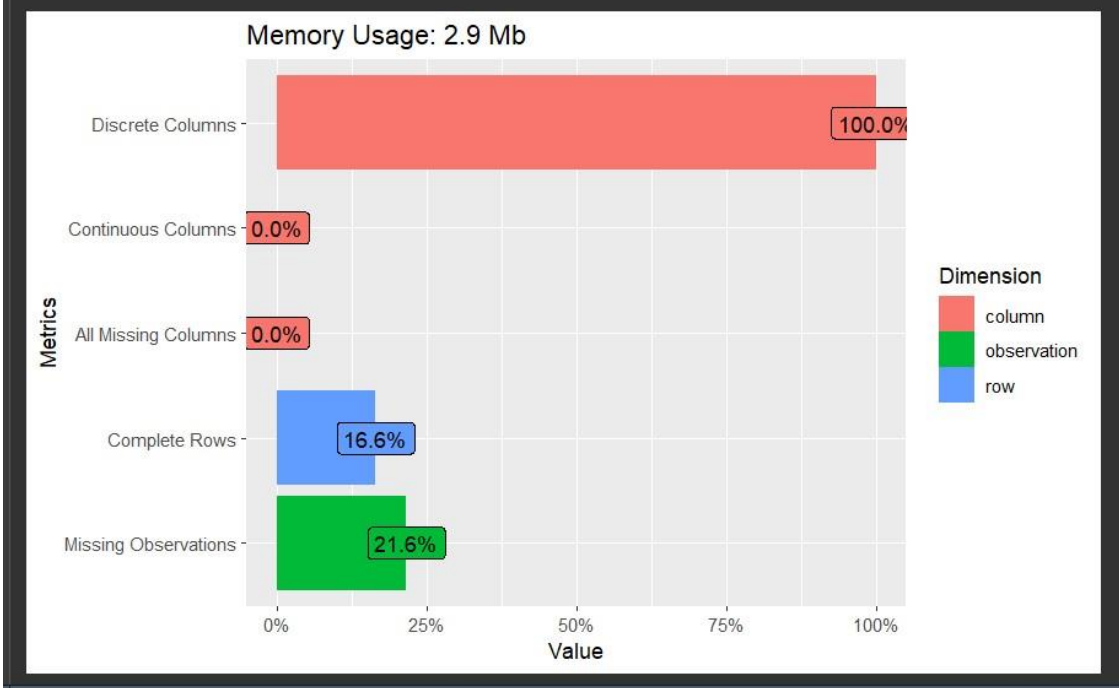


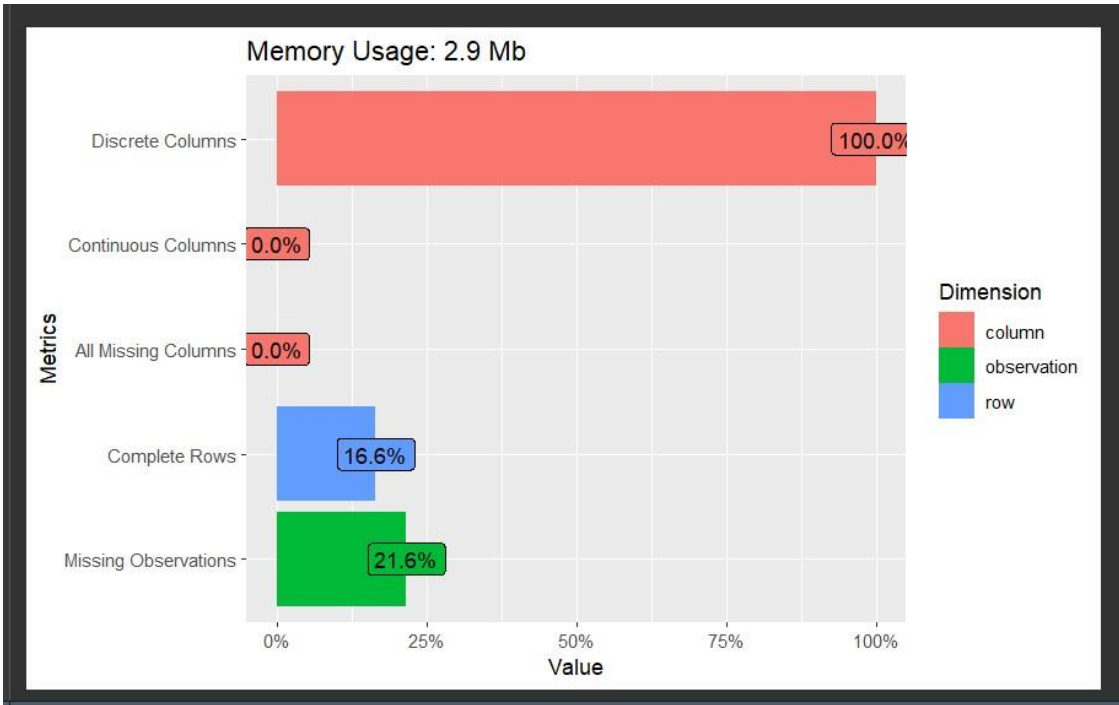
Data Explorer Library

The Data Explorer library³⁵ was found to be an extremely powerful tool to assist in exploring the data but the results were impracticable for the study. The inability to modify the plots easily meant that the information displayed tended to not cover the detail that was required:



³⁵<https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html>

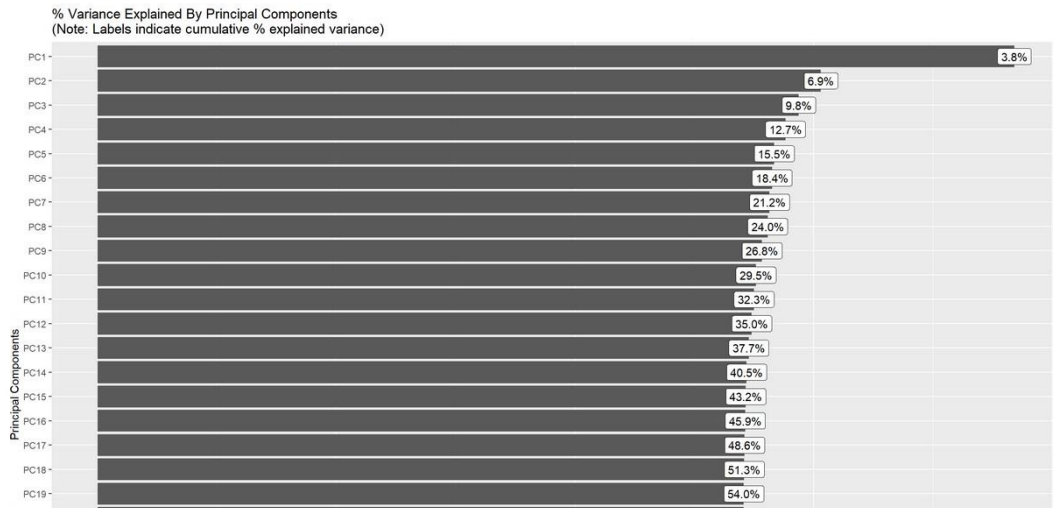




The library also produced automatic Machine Learning, but it was felt that it was not fitting to submit this as work completed by the author:

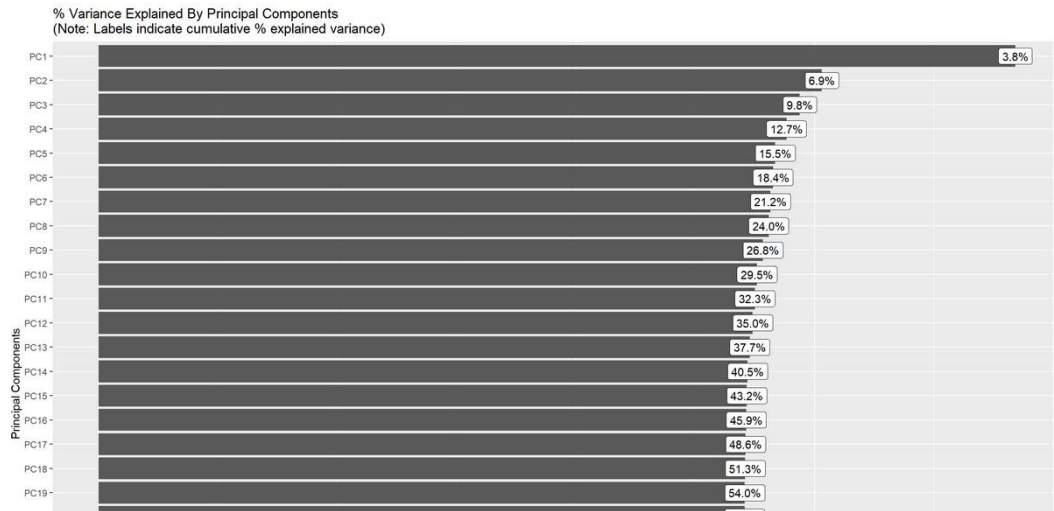
Principal Component Analysis

```
## 4 features with more than 50 categories ignored!
## claim_published: 549 categories
## keyword_name: 420 categories
## country_name: 98 categories
## organization_name: 495 categories
```



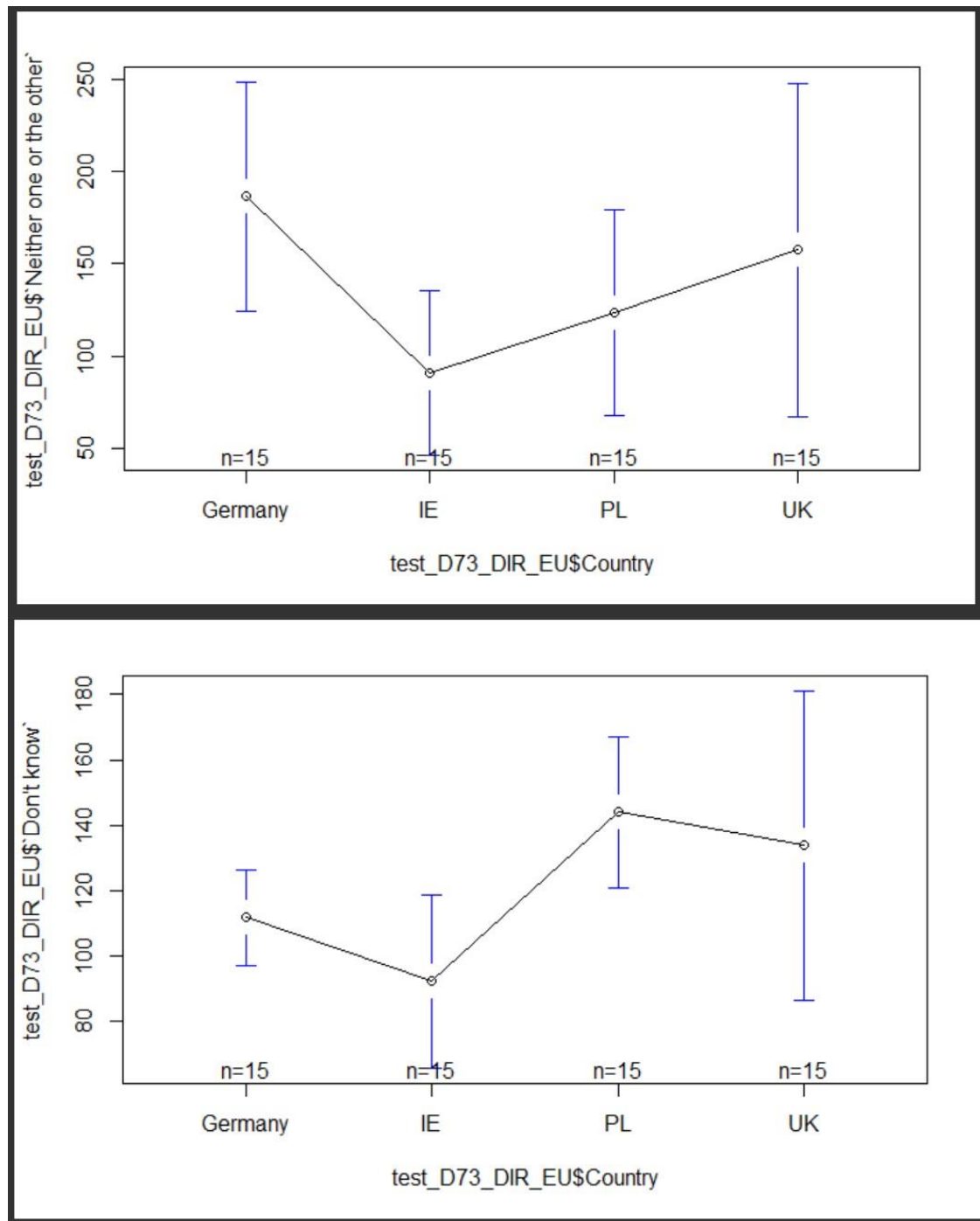
Principal Component Analysis

```
## 4 features with more than 50 categories ignored!  
## claim_published: 549 categories  
## keyword_name: 420 categories  
## country_name: 98 categories  
## organization_name: 495 categories
```



Plot Means

The gplots library³⁶ includes a feature that plots the means of the results of the statistical testing. These plots were felt not be conducive to the overall conclusions of the experiment. These plots themselves were lacking in detail and while they could be modified, it was felt that the main body had sufficient evidence of the results without adding these plots to each section.



³⁶ <https://cran.r-project.org/web/packages/gplots/index.html>

