Explaining Music Recommendations Through Narrative Visualizations

Onushkina, Yana G.

Award date:
2023

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science and Department of Innovation Sciences

Explaining Music Recommendations Through Narrative Visualizations

Master Thesis

Yana Onushkina

Supervisors:
Stef van den Elzen
Martijn Willemsen
Rianne Conijn

Eindhoven, August 2023

Abstract

Many people experience recommender systems as black-boxes, where it is not clear how the algorithm arrives at its output. A solution to this is adding an explanation, explaining the decision making process of the recommender. A number of studies have looked at creating explanations for recommender systems using different visualizations. In this paper we present a novel approach to explanations, using a relatively unexplored visualization style, narrative visualizations. Narrative visualizations help the user build a story line out of the presented data. In this research we investigate how effective, in terms of trust and understandability, this visualization style is at explaining why users got a certain set of recommended songs. To better test the research question, three conditions are used, one where users interact with a narrative based explanation, one where users interact with an explanation without narrative features and one where users do not get an explanation. The results show that having a visualization in the explanation increases trust and understandability compared to no explanation. Similarly, an explanation with narrative features is more effective at justifying the recommended songs than an explanation without narrative features. However, the findings also show that in terms of usability, having no explanation is considered to be the most usable compared to the other two conditions. This finding is logical as the no explanation tool has the least amount of information and no visualizations and hence also limited lag when interacting with the tool, hence it can be considered as the easiest to use.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recommendations are incorporated in many interactions on the internet such as streaming services, webshops, music providers and many others. Recommenders have become a part of our everyday lives and a significant amount of research has looked into different properties of recommenders and how these can be used to help users trust the recommendations. Commonly, this field of research investigates factors that make recommendations convincing. Such factors include choice availability, design, personal characteristics, interactivity and variability. For example, the number of items presented to the user and their quality has a significant effect on the satisfaction of the user, such as presenting too many items or lacking variability in recommendation quality can lead to choice-overload and lower choice satisfaction [4, 18].

However, the biggest challenge to acceptance of recommendations is that people are reluctant to trust a system that they do not understand the working of, the problem of lack of explanation and justification. Tackling this challenge is a field of research investigating how to show users that the recommendations are suitable, through providing an explanation of how the recommendations are made. The field of Explainable Artificial Intelligence (AI) focuses on helping people understand the algorithm behind the AI system, in this case a recommender system, through which it makes its decisions [34]. Currently, such models are used as a black box, an output is given, however, the why is lacking due to the complexity of understanding the model itself. The need for interpretation of models has been investigated in a number of studies starting from the 1970s, as summarised in a survey by Biran and Cotton [2]. The authors' findings show the trend of modern models becoming increasing more complicated and autonomous. These models receive more power by being able to make decisions without human supervision. The need for justification and explanation is the driving force for explainable AI in order to create transparency and accountability.

Explanations in recommender systems can help with acceptance of recommendations. Biran and Mekeown investigated how experts interpret the recommendations of a recommender system [3]. This work showed that even when disagreeing with the predictions, experts will still judge the recommender system as correct if the justification is convincing enough. Further literature on explanations shows that users are overwhelmingly more satisfied with recommendations when an explanation is provided [18]. Hence, explanations are needed to increase user satisfaction with the system, to improve the understandability of the system and to gain trust in the decisions made by the model.

Such explanations can consists of visualizations and text. In the current study a novel visualization style is used, narrative visualizations, which aids the user with understanding and has been shown to increase user trust in the provided information by creating a story line for a user to follow [16]. Storytelling is a very natural and key aspect of human communication, hence to engage a user and motivate them to study the visualization it has to have an intriguing and memorable story [32]. Narrative visualizations also allow for adaptability of the story line based on the user which optimizes the type and amount of information that the reader receives, leading to higher satisfaction and understandability of the information communicated in the visualization [16]. However, currently most of available visualizations lack a story telling feature [13]. Hence,

with this innovative visualization, the current study aims to investigate how effective this visualization is at providing an explanation to the users through looking at how it aids users' trust and understanding the recommendations provided by a recommender system, the primary research of the thesis. To answer this question, a study is conducted with a number of factors measuring features of the tool and users' engagement with it such as how useful the tool is, the perceived accuracy of the recommendations, how much a user interacted with the tool and more. These factors are used to answer the secondary research question, mainly how these factors affect trust and understandability of the users in the recommendations provided by the recommender system. Through answering this subquestion we are able to conclude how effective the tool in general is at improving the trust and understandability of users towards the recommender system. The recommender system is a music recommender system making recommendations of songs using data from Spotify.

The remainder of this paper is organized in the following way. In chapter 2 relevant literature is discussed and the research questions and hypotheses for the current study are derived. Chapter 4 the technical implementation of the proposed tool and its components are discussed along with a justification for the chosen methods. A walk through of this tool is given in chapter 3. After the development of the tool, the focus shifts to the study, the method of which is discussed in chapter 5. Following that are the results, chapter 6, and discussion/conclusion in chapter 7. The limitation of this study and further research as proposed in chapter 7.

Chapter 2

Literature Review

2.1 Related Work

A number of different explanation techniques have been developed in the existing literature for creating explanations for recommender systems. These can be used to serve different purposes, such as increasing transparent, trust or effectiveness of the recommender system. For these purposes different styles of recommendations can be used, such as an explanation focusing on proving why the algorithm works correctly would increase transparency. Tintarev & Masthoff (2015) discuss a number of such explanation styles, where which style is used depends on the goal of the explanation and the algorithm of the recommender system as the explanation has to be suited to the type of information the algorithm provides [37]. The most commonly used explanation style (7/23 papers), according to Tintarev & Masthoff, is a content-based explanation style. This style of explanation is used to explain a content-based algorithm where items are recommended based on their similarity to the items the user likes. Hence this style explains the recommendations by illustrating the similarity of the item and its features to those previously liked by the user. Other explanation styles include knowledge-based, case-based and demographic-based. Knowledge and utility based style is the second most popular type of explanation (6/23 papers), where the most common type of the knowledge-based explanation is case-based, where the explanation is given through providing previous examples, making this style similar to content-based. Content-based explanations are used for content-based recommenders, where the recommendation was made based on which other users have overlapping interests with the current user. Similarly the demographic-based style assumes that the input to the algorithm involved demographic data about the user, hence the explanation focuses on showing the similarity between the recommended items and those from the demographic of the user.

Tintarev & Masthoff (2015) have identified trust as one of the main goals for development of explanations in existing literature. These studies looked at trust as a measure of increase of users' confidence in the recommender system. A total of 9 out of 24 studies evaluated their explanation in terms of trust, as well as other factors. Trust in technology is a well-studied discipline and is known to have an affect on reliance on recommendations of a recommender systems [24]. This effect, however, is heavily dependent on the understanding of what the automation does under different circumstances, where users have shown that knowing when the automation is less reliable leads to higher trust in the system [28]. This is an interesting finding since most developers are not willing to disclose when the tool fails to provide accurate predictions in fears of the tool being viewed as less reliable and trustworthy [12].

Cramer et al. (2008) have looked into how explanations affect trust [8]. In this research Cramer et al. mainly investigated the effect of transperency on trust. For this they designed a CHIP (Cultural Heritage Information Personalisation) system, a system which based on an individual's art preference recommends art from the Rijksmuseum in the Netherlands. The artwork has been labelled by experts at the museum and fed into the algorithm, and an explanation tool

was created, which provides users with information on why this item was recommended to them. To test their system Cramer et al. used three conditions, one with no explanation, a textual explanation stating the similarities between the recommended art and the art the user liked, and lastly showing the users the confidence of the algorithm. The results showed that between the three conditions, no significant differences in trust were observed. However, it was found that trust had a relation to a number of measured variables such as satisfaction, understanding and willingness to accept the recommendations. The findings also showed that users trusted the system, however, that did not translate into them being willing to accept the recommendations, they would still rather choose themselves. As seen in the work of Cramer et al., trust is intertwined with user experience and satisfaction when interacting with an explanation for a recommender system. The research of Zhang & Curley (2018) further builds on this finding by designing a study where users interact with 4 recommender systems for digital cameras [39]. These recommender systems vary on the explanation they provide, ranging from no explanation to using narrative visualizations to illustrate the decisions made by the recommender system for the specific recommendation. After interacting with each recommender system, users rated it in terms of perceived personalization, trust and willingness to use the recommendation provided. The results show that the presence of an explanation increased trust, where the highest trust is attributed to the recommender system with the narrative visualization.

The design of the content of the explanation is the next important step after choosing the style. For this three main methods have been found in literature. Firstly, explanations can be provided through text, such as in the paper by Symeonidis et al. (2009) [36]. Symeonidis et al. present a movie recommender system with verbal explanations based on features of the movie, in the form of "Movie Z is recommended based on features x,y,z...". This work has found that even a simple explanation such as this can still be effective at justifying the recommendations to the users and increasing users' understanding. However, there are some challenges and limitations of creating textual explanations, such as lacking flexibility when the explanation relies on a template or being difficult to control when the explanation is dynamic. Due to these factors textual explanations are rarely used on their own. Some interesting work by Sevastjanova et al. (2018) [33] makes use of both textual and visual techniques, also known as double encoding, when creating an explanation. While looking at machine learning models, not recommender systems specifically, this paper is found to have an innovative approach worth further discussion. A benefit of this method, as stated by Sevastjanova et al., is that it serves a larger set of users, those preferring verbal or visual explanations. A variety of techniques for explanation generation and presentation are presented. The textual components can concern the exact same data as the visualization, they can summarize the visualization/data, they can provide insight into the visualization such as metadata or they can provide further detail on individual decisions depicted in the visualization. These features can be presented on demand, such a user hovering over them, they can be presented based on data driven factors, such as being the most interesting/influential points. The discussed design choices aid in creating an all round explanation, where the user is able to receive further information through interaction. In this interaction the user not only pays attention to patterns in the changing data but also the presented text, which make the explanation suitable for a larger audience.

The remaining literature on explanations focuses on using visual techniques, such as visualisations [14, 5, 20, 30, 27, 16, 23, 21, 31, 25]. The idea for the current study stemmed from a paper by Liang & Willemsen (2021) [25]. Liang & Willemsen use top songs retrieved from a user's Spotify to make recommendations and use interactive visualizations to investigate how these can help improve users' understanding of recommendations. By exploring how good a contour plot visualization, see figure 2.1, is at guiding the user through the recommendations compared to a bar chart as a baseline. The bar chart shows the valence and energy of the recommended songs along with an average of the top songs of the user. The contour plot shows a scatter plot of the recommended tracks and two contour plots, one representing the top songs of the user and the other the genre the user is exploring. With this visualization songs are visualized in a 2D space making it possible to see the relations between the recommended songs, the user's top songs and the songs in the genre. Having a 2D space coverts the relationships between songs, something most people can not visualize, into space and distances. Humans are good at understanding re-

lationships in terms of distance from one another, which is why this visualization is so effective illustrating relationships between songs. The idea behind the current approach is to use a similar visualization and interaction as part of an explanation. This explanation focuses on explaining the recommended items by making use of all features available, such as valence, energy, genre, used in the work of Liang & Willemsen, but also incorporating further features such as duration, popularity and instrumentalness.

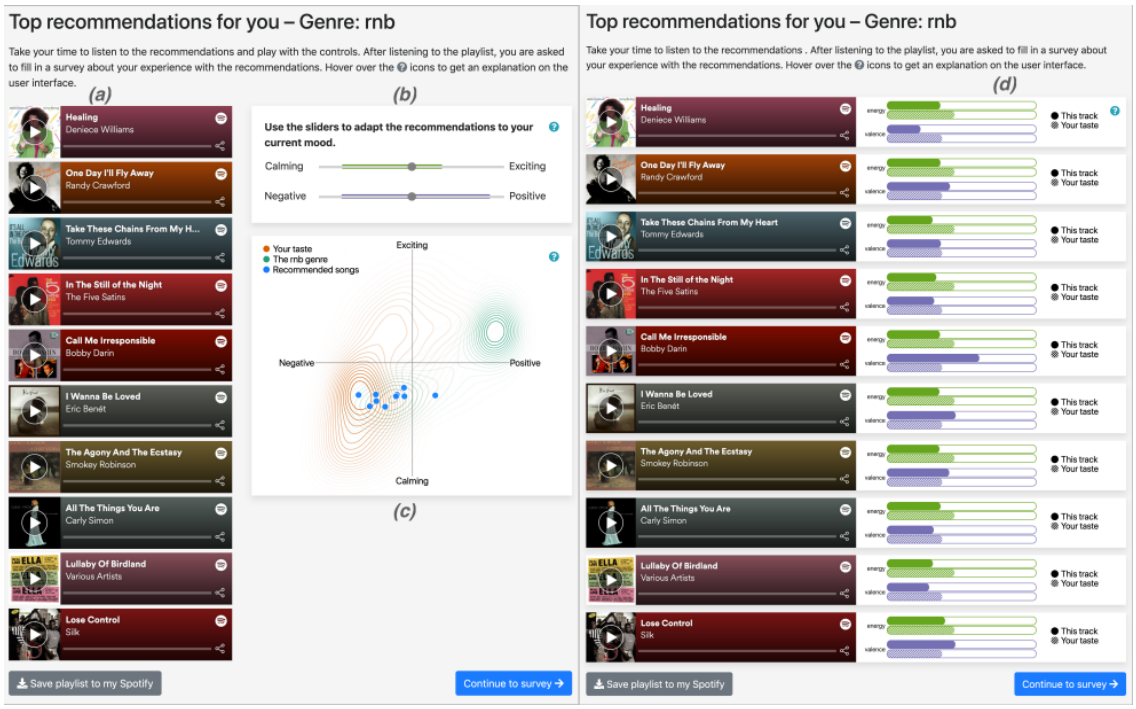

Figure 2.1: A contour plot and bar chart for recommended songs [25]

Building from this study, we further explore distance based visualizations, similar to the contour plot used by Liang & Willemsen, as distance is the one of the most popular methods of representing similarity in literature and while some research has looked into better techniques to visualizing similarity such as work by Fabrikant & Montello (2008), no single best approach has been found [11]. Therefore, we use distance to represent similarity as it has been shown to be an effective method of showing similarity relationships between points. One example of such a visualization is in a tool called PeerChooser. O'Donovan et al. (2008) [27] explain movie genres through a network visualization where the current user is the center of the network and other users are surrounding them. The distance from the user to others depicts their similarity in movie preferences. The user is able to tweak distances between themselves and their peers in order to personalize their recommendations further. Through this interaction the user is able to understand how proximity to others plays a role in their recommendations. The work of Liang & Willemsen (2021) also showed similar results for increased understandability, where users were able to retrieve more insights from the contour plot. The contour plot being perceived as more informative lead to increased understandability compared to the baseline bar plot. Kunkel et al. (2017) build on this spacial/distance based visualization with their tool [22]. The tool consists of a 3D map with pictures of movies scattered among it, see figure 2.2, an innovative version of a scatter plot. The idea behind this is that users are familiar with distances and heights, as these are encountered in everyday life. Hence, Kunkel et al. place items that users will enjoy on top of mountains, while less recommended items are placed on flatter lands, valleys and even in the ocean. The landscape also allows to visualize the genres of the recommended movies, by having similar genres placed closer together. Kunkel et al. found that users thought the landscape

is comprehensive and helpful in providing an overview of the item space. The transparency of the tool and accuracy of recommendations is also rated highly. The ability to interact with the recommendations, dragging them lower or higher increased satisfaction with the tool as users were able to personalize their recommendations further. This argument is convincing, however, personalization based on feedback is not incorporated into the tool for the current thesis. The presented literature makes a good case for space based visualizations, where users are able gain insights about similarity based on the proximity of items to each other. This explanation helps with building trust in the recommendation system.

Figure 2.2: Visualization use in [22] to explain movie recommendations

Zhang & Curley show that narrative visualizations are most effective in an explanation helping build trust in the recommendation system. Narrative visualizations are visualizations that use different techniques to create a narrative/storyline for the user. Narrative visualizations facilitate understanding by dividing the information presented to the user into logical blocks of information, which flow together. When taken separately these blocks tell a part of a story and when placed together they create a narrative. Narrative visualizations are not a new concept. This is because anywhere a visualization is used, a narrative feature can be added to it for additional support and explanation, which makes the visualization easier to interpret and understand [29]. The work of Belmonte (2014) is a great example of what a narrative visualization is [1]. In their paper Belmonte discusses a tool which finds relevant Twitter post data to supplement a text of a public speech. The visualization is primarily a script of the speech. As the user reads the paragraphs of the text, they are provided with a streamgraph of the trending hashtags on Twitter in that moment on top of the page. The visualization of the streamgraph allows the user to hover over it, such that activity in the streamgraph is coupled to the exact timestamp of the text, by highlighting the relevant parts of the text. This visualization can be seen in figure 2.3. There is also a bar chart, representing the popularity of the presented hashtags, and a map locating the where the Twitter activity is coming from. Each of the elements on the page builds a story, and together a storyline is created. Unfortunately, Belmonte did not conduct a user study or present any discussion on the effectiveness of their visualization, hence it is not possible to conclude how effective their tool is. However, a number of papers discussing design choices for narrative visualizations state the effectiveness of this style of visualizations in facilitating user understanding and trust [13]. One of the main goals of an explanation is to increase user trust, where this goal can be facilitated through the use of a visualization with narrative features as it is show to increase understandability and trust. Hence, in this paper, this style of visualization is used in order to increase understandability and trust through facilitating storyline building of the explanation.

2.1.1 Primary Research Question

As discussed, explanations aim to help the user understand why they got a certain set of recommendations. At the same time the presence of an explanation increases trust of users in the

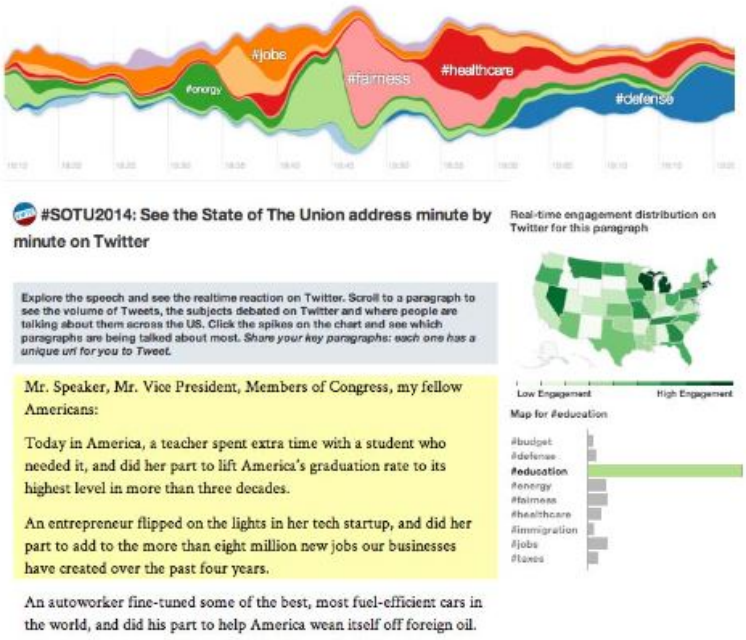

Figure 2.3: Narrative Visualization through an adapting script with supplementary visualizations from [1]

recommender system. In most literature, understandability and trust are not addressed together. Out of the found literature, one paper looked at both trust and understandability. The findings of Cramer et al. (2008) show that trust and understandability are significantly correlated ($r = 0.450$) [8]. Understandability is also correlated with the willingness to accept recommendations showing that understandability is related to system acceptance, which trust is also related to. Hence, in this study we look at the effectiveness of an explanation in terms of both understandability and trust. Furthermore, not a lot of literature looks into explanations using narrative features, hence, we are looking further into this with the intention to find if an explanation with narrative features is more effective than one without. From this we define the first research question.

RQ1: How effective, in terms of trust and understandability, are narrative visualizations in explaining recommendations made by a recommender system?

For the current work we need to create a baseline for effectiveness to draw valid conclusions regarding the narrative explanation. Hence, we create a total of 3 conditions; the control condition with no explanation, a non-narrative condition with an explanation without narrative features and a narrative condition as has been discussed throughout this chapter. The control condition helps us answer the question of "do we need an explanation?". The non-narrative allows to draw conclusions about the narrative features of the narrative explanation.

The main measure of effectiveness, used in this thesis, is trust and understandability. Ghidini et al. (2017) and Zhang & Curley (2018) show that narrative features lead to higher trust and understandability. Symeonidis et al. (2009) conclude that having an explanation, even a simple one, is better for user trust and understandability than no explanation. Based on these findings we formulate hypotheses H1 through H4.

- H1: non-Narrative will increase trust in the recommender, compared to Control
- H2: Narrative will increase trust in the recommender, compared to non-Narrative
- H3: non-Narrative will increase perceived understandability in the recommender, compared to Control

- H4: Narrative will increase perceived understandability in the recommender, compared to non-Narrative

We want to know how effective the narrative visualization is at improving trust and understandability. However, we are also interested in creating a tool which improves transparency, users' satisfaction, if the tool is persuasive and makes users confident in the decisions of the algorithm. Kouki et al. (2019) have derived a scale which measures reception of a tool, consisting of the 4 variables we desire [19]. Hence, we also look at how the tool with a narrative explanation is received compared to one with is no explanation or no narrative features.

- H5: non-Narrative will have a higher reception, compared to Control.
- H6: Narrative will have a higher reception, compared to non-Narrative.
- H7: Higher perceived understandability leads to higher trust.

2.2 Secondary Research Question

As the second part of this study, we aim to understand how different characteristics of the tool and the user contribute to the trust and perceived understandability built by the explanation.

2.2.1 Understandability

Transparency

Explainability is desired from recommender systems as users want to understand the recommendations they got. Previously it is believed that users want to understand why items they did not like were recommended to them, however, [35] found that users also want to know based on what parameters the items they did like were recommended to them. Giving users good recommendations is no longer enough, as they are looking for a justification and explanation of the decisions made by the system [35]. In the work of Sinha & Swearingen (2002), transparency is identified as an influential factor in explainability of a recommender system [35]. In this study 5 music recommendation systems were used each with varying degrees of transparency. While having a small sample, 12 people participated, the findings show a clear answer. Both mean liking and mean confidence were significantly higher for transparent systems [M = 3.51, 8.12] vs. non-transparent systems [M = 2.79, 6.89].

In the current work transparency is part of the reception metric developed by Kouki et al. Kouki et al. investigate the effectiveness of explanations for hybrid recommenders, recommenders using multiple styles of recommending together to produce recommendations. The authors produce 5 different explanation styles for this system, user-based, item-based, content-based, social-based and popularity-based. As part of the analysis, Kouki et al. are interested in the reception of their explanations, allowing them to understand which explanation was best received by the users. This metric evaluates what the user thought of the explanation in terms of how convincing it was.

- H8: Higher reception leads to higher perceived understandability.

Information Sufficiency

As shown in the work of Liang & Willemsen (2021), perceived understandability is influenced by perceived informativeness and control. In the current study no elements of control are built into the tool, hence this variable is not measured. However, informativeness is measured through information sufficiency.

- H9: Higher information sufficiency leads to higher perceived understandability.

Ease of use and cognitive load

It is also expected that if the tool is perceived to be easy to use, it is then easier to understand. Similarly, if the tool provides too much interaction and information, this is expected to affect understandability in a negative way, resulting in a high congnitive load. These effects are not found in existing literature, hence they are investigated in this study.

- H10: Higher ease of use leads to higher perceived understandability.
- H11: Lower cognitive load leads to higher perceived understandability.

Interaction with the tool

The explanation provided in the current work provides some ability for interaction, which reveals more information to the user. Hence, it is expected that the more users interact with the explanation, the better they understand why the items were recommended to them as they recieve more information.

- H12: Higher interaction with the tool leads to higher understandability.

Visualization familiarity

Lastly, understanding of the explanation is heavily dependant on understanding the visualizations provided. Hence, it is expected that users who have a history of interacting with visualizations will be able to understand the visualizations better, hence find the tool easier to use. This increase in ease to use then increases understandability (H10).

- H13: Higher visualization familiarity leads to higher ease of use.

2.2.2 Trust

There are several more factors influencing the trust one has in a recommender system. These are algorithm aversion, transparency and accuracy, both perceived and objective.

Algorithm Aversion

Algorithm aversion is a very common problem in the recommender systems literature. Algorithm aversion results in a decreased tolerance for error. A paper on algorithm aversion has shown that people are less tolerant to errors made by algorithms than they are to those made by humans [9]. Hence, algorithm aversion results in a lower trust for a recommender system. However, literature shows that trust in recommender systems with an explanation are less susceptible to the algorithm aversion bias [10]. In the current study it is expected that the aversion users have to algorithms will negatively impact how much they trust the recommender system.

- H14: Lower algorithm aversion leads to higher trust.

Transparency

Transparency improves trust, such as when a user receives an unexpected result, if the recommender system is transparent, by providing an explanation, the user is more willing to trust it [17]. Findings by Nilashi, et al. (2016) show that users find transparency equally important as recommendation quality [26]. Hence, transparency, measured through reception, is expected to have a postive effect on trust.

- H15: Higher reception leads to higher trust.

Accuracy

Accuracy is also an important factor in building trust in the system. With accuracy it is also looked at both perceived and stated accuracy of the system. Work by Yin et al. (2019) [38] finds that both types of accuracy influence a user's trust. However, this research also concluded that users rely more on perceived accuracy, as their interpretation of the objective accuracy can change depending on how accurate they perceive the system to be. Other research finds that objective accuracy does not influence trust [8]. The relationship between objective accuracy coming from the system and perceived accuracy requires further investigation for more conclusive results. However, from previous studies it is clear that perceived accuracy affects trust in a positive direction.

- H16: Higher objective accuracy leads to higher perceived accuracy.
- H17: Higher perceived accuracy leads to higher trust.

The hypotheses above are used to answer the second research question defined as:
RQ2: What is the effect of algorithm aversion, visualization familiarity, accuracy, perceived accuracy, reception of the tool, interaction with the tool, perceived ease of use, cognitive load and information sufficiency on trust and perceived understandability of the recommender system?

2.3 User-Centric Evalution

The research questions defined are evaluated using a user-centric evaluation technique as defined by Knijnenburg & Willemsen (2015) [18]. Most research tends to focus on the system itself, the algorithm behind it or the output of the system. Knijnenburg & Willemsen describe three main interaction components, the algorithm, the outputs of the system and inputs that the user is required to give. The authors state that a key factor of conducting a good study is to look at the interaction between all three components.

The framework contains five key concepts that should be measured during a user study with a recommender system, see figure 2.4. These are: Objective System Aspects (OSA), Subjective System Aspects (SSA), User Experience (EXP), Interaction (INT) and Personal and Situational Characteristics (PC and SC). OSA refers to elements that can be measured as part of the system. These are the inputs/outputs of the recommender system where several simple key features should be measured per study. SSA is the perception that the user has of the OSA, hence, SSA helps with understanding what the user perceived the system to be like. EXP refers to the evaluation of the recommender system of a feature of interest of the recommender system. INT measures how the user is interacting with the system, most often measured through clicks-stream data. PC and SC are items which might be influencing certain behaviors or opinions of the user. These characteristics are measured through a questionnaire either before or after the introduction of the system.

The user-centric model for the current research questions is shown in figure 2.5. For the OSA features we have the interaction if there is an explanation or not, if it is narrative or not and observed accuracy. Observed accuracy is expected to have an effect on the SSA metric of perceived accuracy (H16). The explanation presence and type are expected to have an effect on perceived understanding and trust, however, it is expected that this effect is moderated by all of the SSA metrics (H1-H6). Personal characteristic of visualization familiarity should effect the perceived ease of use of the system (H13). Algorithm aversion is expected to have a direct effect on trust (H14). Interactions with the tool are expected to help with understandability (H12). Lastly, the SSA metrics are expected to effect perceived understandability and trust as discussed in H8, H9, H11, H15 and H17.

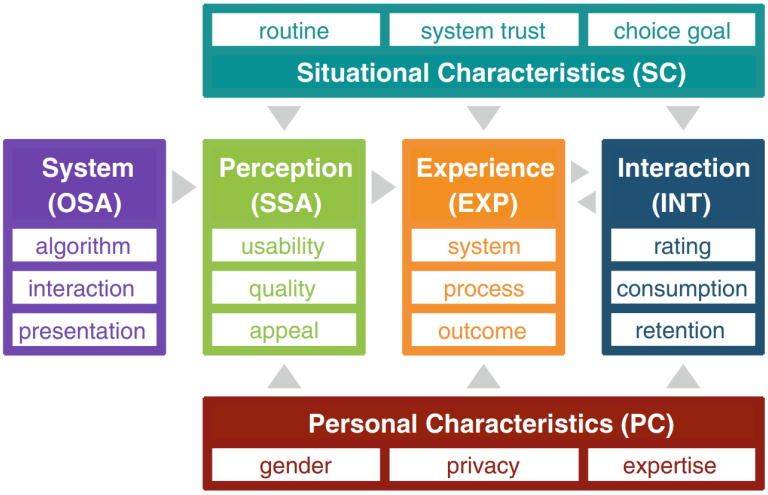

Figure 2.4: The User-Centric Evaluation Framework for Recommender Systems [18]

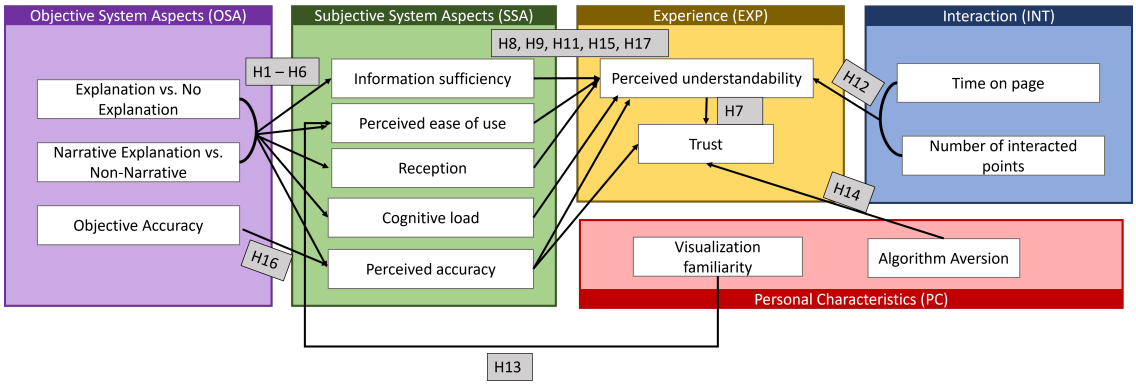

Figure 2.5: Model of the user-centric approach to evaluating the recommender and explanation of the current thesis

Chapter 3

System Design

The proposed system, *Recommendation Explanation*, is a tool which uses narrative features to build an explanation as to why a user is presented with a specific set of recommendations. It makes use of visualizations, one of those being a distance representing visualization such as a scatter plot, text and interaction possibilities for the user. The recommender system used to make recommendations makes use of features of the songs, based on which the recommendations are made. Hence, the explanation uses these features to further explain why this set of recommendations, more specifically the tool identifies top 5 most influential features and presents those. The tool itself is encapsulated in a system to function as an online study. This system guides the user through the steps of the study. Through this, users are able to participate in the study in one screen, instead of having to switch screens to complete the questionnaires and interact with the explanation. In this chapter, we provide an argumentation for the decisions made while developing the tool.

Firstly, we decided to create a tool from scratch instead of using an exciting one. This is because this research aims to combine multiple techniques from different studies. As discussed in chapter 2 the research idea stems from the paper of Liang & Willemsen (2021), hence it is logical to proceed with a similar set up of using a song recommender system and utilizing Spotify. The visualization used by Liang & Willemsen is also shown to be effective at helping users understand the relation between their songs and the recommended songs of a new genre. Hence, we proceed with this visualization as it was found to be intuitive for users, due to it representing the relationship it terms of distance, and the work of Fabrikant & Montello (2008) confirms that distance based plots are the most effective type of plot for representing similarity relationships to this date. However, in the current thesis, we are not interested in the genre exploration part of Liang & Willemsen's work, where this exploration is supported by the contour part of their plot. Hence, we will be using the underlying scatter plot only to show similarity between recommended songs and the songs of the user.

For the recommender system we did not intend to spend much time on developing a good recommender system as it is not an objective of this study. The best approach, and one explored first, is to use the recommender system provided by Spotify[1], however, at the time of creating this tool it is not possible to see the input to the Spotify recommender, the exact songs the algorithm used. Additionally, the implementation of the Spotify algorithm is not publicly available. This means that it is not possible to explain the recommendations in a model-agnostic or a model-specific manner as the input to the model would be needed for the former and knowledge of the implementation of the recommender for the latter. Due to this a different recommender has to be used. The recommender system of Liang & Willemsen is also not fitting for our work as it is specific to their implementation and extensions of it would be time consuming and challenging to implement. Hence, we chose to employ a simple recommender based on an existing simple implementation [2]. This approach is used as it implements a recommender system fitted exactly to the requirements of this study; the prediction is based on all available features for a song in Spotify,

[1] https://developer.spotify.com/documentation/web-api/reference/get-recommendations

[2] https://towardsdatascience.com/part-iii-building-a-song-recommendation-system-with-spotify-cf76b52705e7

the algorithm is content-based giving a recommendation based on items the user already likes and the implementation is done in Python. The preprosessing of the songs before recommendation and the implementation of the algorithm are further discussed in sections 3.4 and 3.5.

For the explanation of the recommendations we give two types of explanations. Firstly, the songs that a user likes, the songs fed to the algorithm as input for recommendations, are presented to the user. Through this the user can see how similar the recommendations are to the songs they already like, where this similarity between a set of songs explains why the songs are recommended. Secondly, top 5 features based on which the recommendations are made are shown to the user and what values the recommended and the top songs had for each feature. This second explanation is an addition to the explanations seen in previous literature. Take the work of Kunkel et al. (2017) with the map movie recommender explanation tool. The explanation in this tool uses other movies to explain why the movies are recommended to the user, hence showing which movies the recommendation is similar to. Not much information on properties of the movies is given aside from the genre. We think that other features can be of importance, such as users might like short movies, hence an influencing feature in the recommendation process would be the duration of the movie which is potentially interesting for a user to know and provides a further explanation into how a recommendation is derived. Hence, for the current thesis we use the features that are most influential during the recommendation process for a user especially as our recommended system recommends based on the provided features for the songs.

Deriving the top 5 features is done through classification by looking at what are the most common features among the recommendations. The goal is to find a set of input features that are important to predict the output, hence which features influence the category (recommended or not) an item is predicted to be in. To find these similarities a k-Neighbours, a Regression or a Random Forest Classifier are often used. Out of these, the Random Forest Classifier has been shown to be the most effective and accurate, hence it is used in our work [7, 6]. The derived features show the features that the recommended songs have in common. This is a good explanation for how the recommender system works since the recommender system gives the recommendations based on these features. Hence, if the recommended songs have feature X in common, there is a high chance that these songs were recommended because the top songs have this feature as well. This is not a perfect method as it is possible that the recommendations have a feature in common through chance. However, the derivation of influential features in recommendations is a complex challenge to which no perfect solution exists. The current method provides a simple, however, imperfect implementation which through personal experimentation produces logical results.

For displaying the top 5 features we use a parallel coordinates plot (PCP). This plot is used to show what values the recommended and the top songs have for the 5 most influential features. The plot is used for visualizing high dimensional data, making it easy to notice clusters and trends for the different dimensions. This is possible as a PCP makes it easy to see correlations in the data, lines where given value X on dimension A, on dimension B values are around Y, creating an overview of patterns in the data. This type of plot has been shown to be highly effective at showing clustering, through which users learn about similarity between the different items [15]. As this is the main objective of presenting the top 5 features, we show the similarity between the recommended and top songs through the top 5 features with a PCP.

Lastly, as discussed in chapter 2, this thesis looks at narrative visualizations specifically aiming to employ narrative techniques to explanations in order to test how it affects trust in recommender algorithms. From previous literature it is shown that with narrative visualizations it is important to build a storyline for all users. In order to build a storyline we spread information over several pages. This helps limit the amount of information the user consumes in one go, allowing them to learn simpler concepts and move on to more complicated ones, building the narrative. We do this by first showing users the scatter plot, since that is a visualization most users are likely to be familiar with and only later introducing the PCP. Through this users learn about the similarity of their songs based on the proximity to their liked songs and then build their understanding by adding the top 5 features. We also allow the users to obtain more information through interaction which further spreads the information out, whereby users can not see all information available on a page in one go and hence builds the narrative. This interaction allows users to learn on their

own pace and not get too overwhelmed with new information as they can choose when and how many points they hover over. Lastly, we make use of text in summary pages to help all users understand the information provided. The idea is providing users with the same information but in a different format which has been shown to be effective at reaching a wider range of audience in the work of Sevastjanova et al. (2018). This helps users with understanding the storyline we are trying to create.

In the remainder of this chapter the technical implementation of the system is discussed.

3.1 System Implementation

The system is built using a combination of Flask (version 2.2.2) and Dash (version 2.7.1), in Python 3.9.16. The architecture is set up using Flask in addition to the pages relevant to the study only, such as questionnaires and debriefing. The tool itself is made using Dash.

Flask comes with a number of out of the box functionalities which are relevant for this system. Firstly, Flask contains sessions, where a new session is created every time the application is opened. In these sessions it is possible to store information accessible at any stage of the application, such as the local identifier of the user. On top of that, Flask contains a number of compatible packages for creating forms, for displaying questionnaires, and interacting with databases. Flask can also host an application inside it, which is how the tool is incorporated into the system.

The tool is created using Dash, which runs on a Flask server. Dash is chosen as it is a recommended tool for creating advanced dashboards[3]. Since the tool that is discussed in this work is similar to a dashboard, this package is well suited. Dash comes with a lot of prebuilt functionalities. Since it is intended for creating dashboards, the tool comes with a large selection of plots and properties which can be changed in these plots. Dash also makes interactions easy to support. It is possible to easily execute functions when an event occurs. For example, when hovering over a point in the plot, to recolor the plot, recolor the supporting plot and highlight an element on the left side of the screen. Dash has some limitations such as not all plots support hovering to create an event. This caused the interaction the user can have with the system to be limited to the scatter plot.

3.1.1 Databases

A MySQL database is used, which is hosted on the EU server of PythonAnywhere and is accessed by the app using $flask\_sqlalchemy$. The database is used to store information about the user, the responses to the questionnaires and interactions with the plots. Additionally, interaction with the system is recorded. This concerns how much time users spent on each page of the system and the tool.

3.2 Spotify Authentication and Song Retrieval

3.2.1 Authentication flow

In order to be able to retrieve songs from the Spotify account of a user, they need to be authenticated. The protocol and how to set it up is documented by Spotify. First step is creating an application on the Spotify Developer Portal. The app contains a client id and a client secret, making it possible to make a connection to Spotify APIs through this app.

Once the app is set up the next step is creating the authentication workflow, documentation for which is provided by Spotify[4], see figure 3.1. The flow starts with the system requesting authentication from Spotify, if it is granted, when a token is given which is used to direct the user to Spotify. In Spotify the user is triggered to log-in and then are asked if they agree to their data being used for by the application. After submitting their reposnse, the user is redirected back to

[3] https://medium.com/spatial-data-science/the-best-tools-for-dashboarding-in-python-b22975cb4b83
[4] https://developer.spotify.com/documentation/web-api/tutorials/code-flow

the system. The steps of what happens where a user agrees or not are set up in the system. When a user does agrees, an authentication token is returned with which it is possible to query their data.

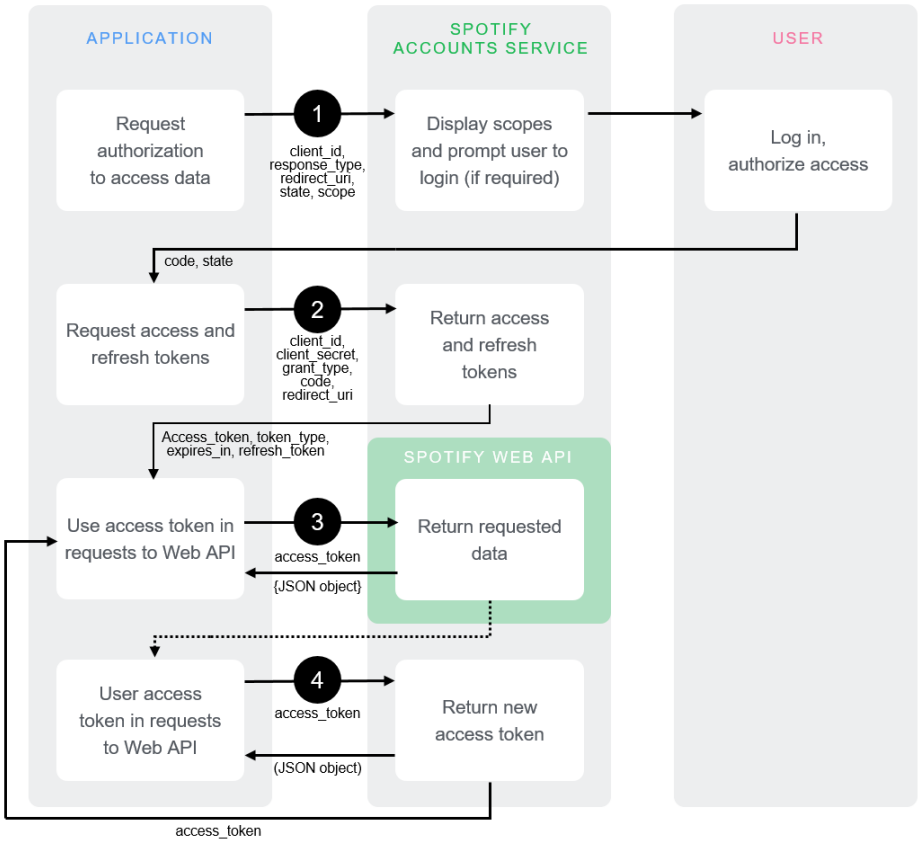

Figure 3.1: The Authentication Code Flow for Spotify `https://developer.spotify.com/documentation/web-api/tutorials/code-flow`

3.2.2 Data Extraction

The gathering of the data is done in two ways in the current system. To retrieve the top songs of the user and the information about them a package Spotipy is used[5]. This package makes extracting songs very straightforward, as it is possible to use a function for the information you want to retrieve, such as *current_user_top_tracks*(), pass the authentication information and the package handles the rest returning the output it receives from Spotify. However, when developing we noticed a problem with retrieving information about the artist when using the package. Due to this, the rest of the information gathering is done through API calls, using HTTP requests.

3.2.3 Extracted Information

An overview of the information that is available for retrieval can be viewed at[6]. For this system the following features are extracted using these APIs:

- Get Current User's Profile - Used to retrieve the Spotify account id, which is used to ensure that a user does not complete the study more than once with the same Spotify account.

[5]Documentation for Spotipy `https://spotipy.readthedocs.io/en/2.22.1/`
[6]Spotify Web API Documentation urlhttps://developer.spotify.com/documentation/web-api

- Get Tracks is used to obtain information about the top songs of the user. This includes: the popularity of the track, if the lyrics are explicit, information about the artist and the album the track is a part of.

- Get Audio Features retrieves features of the songs such as accousticness, danceability, duration and energy.

The extracted information is then put together into one large dataset of songs with all possible features. The recommendation can now be done using the dataset.

3.3 Database of songs

The database of songs is created specifically for this tool. It consists of over 23000 songs from a variety of genres such as avant-garde, blues, classical, country, electronic, folk, jazz, latin, new-age, pop-rock, rap, raggae and rnb. The same genres are used as in previous work of Liang et al. (2021), as these represent a varied set of genres likely to appeal to most users. From each genre 250-400 most popular artists of that genre are selected, and the top songs for each artist are retrieved. This is done using Spotify APIs. The database is tested and trimmed to an optimal size to ensure efficient loading speeds. The recommendations to the users are made from this database.

3.4 Song Processing and Analysis

The songs in the dataset are processed to create a format for recommending. This includes one-hot encoding all categorical features and performing sentiment analysis on the song name. The one-hot encoding is done using the *get_dummies* function of *pandas*, creating dummy variables for all categorical features of the song, such as if the song is explicit or not. The sentiment analysis is done using a package *TextBlob* which can analyze text based on subjectivity and polarity. This analysis turns the name of the song into a feature of interest as well, through the scores of subjectivity and polarity, where songs can be recommended based on the sentiment of the name. The song database of the tool is prepared in the same manner. After this, the songs are ready for the recommender algorithm.

3.5 The Recommender Algorithm

The recommender works in the following way. It gives a similarity score for each of the songs in the database based on each of the 60 top songs of the user. The similarity score is calculated as cosine similarity between the database and each top song using formula 3.1. After this, the top 10 songs in the database with the highest similarity to the top songs are selected. These are the top 10 recommended songs. During this process, the top songs that have the highest similarity to the recommended songs are recorded, to be displayed to the user in the scatter plot as discussed earlier in the chapter.

$$similarity = cosine\_similarity(\begin{bmatrix} song_1 & feature_1 & ... & feature_n \\ ... & ... & ... & ... \\ song_m & feature_m & ... & feature_{nm} \end{bmatrix}, \begin{bmatrix} topsong \\ feature_1 \\ ... \\ feature_n \end{bmatrix}) \quad (3.1)$$

After the recommendation is completed, the data is processed and prepared for the visualizations.

3.6 Preparation for plotting data

For the scatter plot the idea is to display where the songs are in relation to each other based on all of the features retrieved from Spotify. The dataset of recommended songs and top songs containes over 100 columns, hence, a dimensionality reduction technique is used to create a 2D plot. A number of dimensiontality reduction methods exit for non-linear data, where MDS and t-SNE are most popular. Multidimensional Scaling (MDS) focuses on preserving distances between different clusters, while t-SNE prioritizes within cluster distances. Both techniques were explored in this work and t-SNE is found to be the most fitting. This is because, the relationships between the songs in one cluster are more interesting for this use case than those between clusters. For this explanation it is enough to show that the songs are in different clusters, since we want to focus on placing similar songs closer together in a cluster, such that the user is able to see which songs the recommendations are most similar to and hence why the songs were recommended.

Chapter 4

System Walk Through

The *Recommendation Explanation* tool is designed with the primary goal of explaining to users why they receive a specific set of recommendations. This is done using narrative visualizations, which help build a story for the user. In order to understand the effectiveness of this technique, the tool is extended with 2 additional explanation styles: no visualization and a non-narrative visualization. To make this assessment, the tool is encapsulated in a system which guides the user through the steps of a study.

4.1 System Design

This section consists of 3 parts, where the first and second part of the study workflow are discussed and a section on the *Recommendation Explanation* tool. The later goes over the workflow of the tool for each of the three conditions.

4.1.1 Study Workflow Part 1

As shown in figure 4.1 the first part of the system consists of 5 pages. The user lands on the home page, where they are greeted. By selecting the 'Continue to informed consent' button, the user is redirected to page 2. On this page the informed consent form is displayed. The informed consent can be found in full in Appendix A. Upon reading the consent form, the user can select if they consent to the information in the form. This selection is mandatory, if a user does not select an option they are prompted to do so before continuing. In case the user does not consent, they are redirected to page 3_1 and complete the study. Otherwise, the user proceeds to page 3, where from they can proceed to login and authenticate themselves in Spotify.

The request to Spotify can be unsuccessful, such as if the user does not authenticate. In this case page 4_1 is presented, where the data of the user obtained so far is cleared and they can choose to try again. Alternatively, the user might have already interacted with the system before and is attempting to do so again. As this is not allowed, they are redirected to page 4_2. However, if the authentication is successful and this is a new user, then their top 60 tracks are obtained and stored and they are assigned to a condition. The user is redirected to page 4. On this page they fill in the Intake Questionnaire. On page 5 the user selects the 'Load data and go to tool' button, after which all the needed databases are prepared, recommendations and analysis are performed.

Lastly, the system also contains a page X. This page can be redirected to from any point in the system. If a user tries to navigate to a specific URL, without have completed the necessary steps before hand (such as if after the study they navigate back to the *Recommendation Explanation* tool), they will be directed to this page. The purpose of the page is to ensure that it is only possible to access the system step by step, instead of being able to navigate to a specific URL, ensuring a smooth experience for the user.

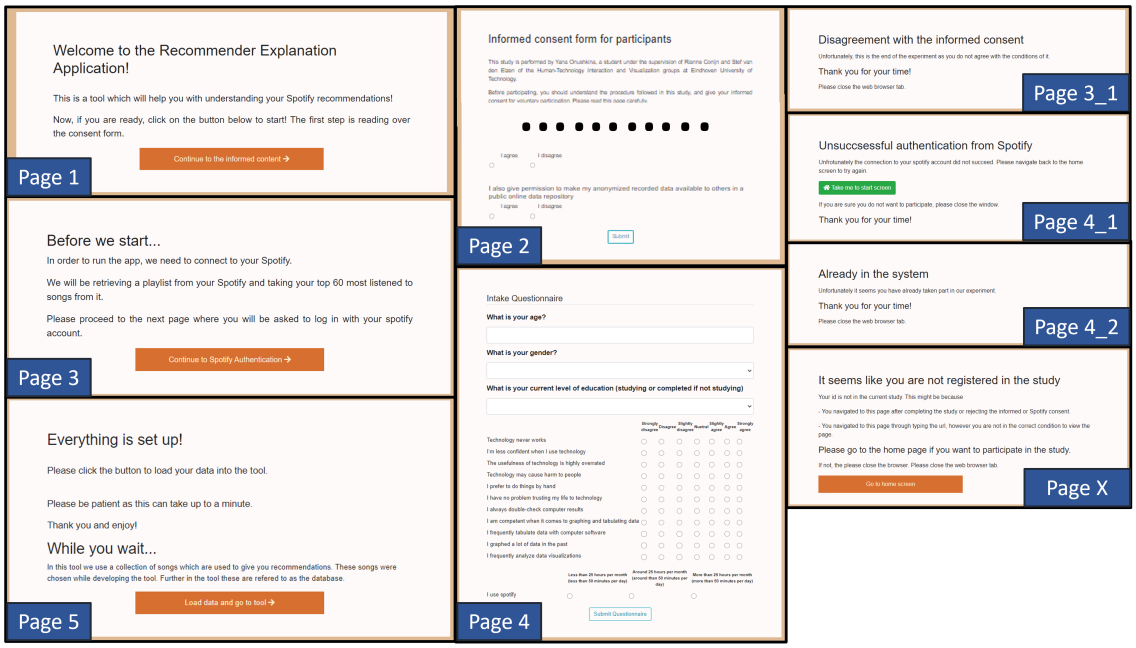

Figure 4.1: Overview of the first part of the system as discussed in section 4.1.1

4.1.2 *Recommendation Explanation* Tool

Condition 3: Narrative Visualization

An overview of the pages for the narrative condition are displayed in figure 4.4. On page 1 an introduction to the tool and the steps ahead is shown. To help users get a familiar overview of their songs, the explanation displays the songs in a similar format to Spotify, see figure 4.4 page 1. Here the user can see information about the song, such as the artist, the name, an image of the song cover and a video fragment they can play. This information is obtained from Spotify. This view also shows the confidence of the recommendation coming from the recommender system. This is shown to the user to help them build trust for the recommendations as they are able to see how fitting the system finds these recommendations to be. The recommended songs in this view are colored with a gradient going from top recommendation (darkest) to 10^{th} recommendation (lightest) to help the user with identifying the songs in the visualizations. The gradient also illustrates that there is an order to the recommendations, top 1 to top 10.

On page 2 the user interacts with the first visualization. The text on top of the page informs the user of what the visualizations is, what interactions are possible and what they are expected to do. In the middle of the page the visualization is shown. The scatter plot is used to visualize the recommended songs of the user along with their top songs, see figure 4.4 page 2 for an illustration of the discussed visualization. Through this, users are able to see which songs are close together, where distance is a meaningful property of this visualization. A set of recommended songs being close to some top songs tells the user that the songs were recommended to them because they are similar to what they already like. The colors of the dots were chosen to create a high contrast between the recommended and the top songs. The recommended songs are colored with a gradient to match the overview of the songs on the left side of the screen. On the left of the page the recommended songs are displayed. The user is able to listen to a sample of the song if available. Above the scatter plot is a legend and a mint green bar, where the color is chosen to complement the existing colors of the explanation.

The interaction on this page is done through hovering over the different points (songs) in the scatter plot. When hovering over one of the top 60 songs, this song will be highlighted in blue, other points repainted grey and information about this songs is displayed in the bar above the

plot, see figure 4.2. When hovering over a recommended song, the point is highlighted in blue in the plot and in the song view on the left side of the screen. Additionally, a different song in the plot is highlighted in yellow. This is the song which is the most similar to the recommended song that is hovered. Through this, the user is able to learn more about the points in plots, which songs are similar to which and what is the most similar song for each of their recommendations. The mint bar above the visualization shows additional information about the song, such as name, genre, popularity, duration and if it is a recommended song then which of the top songs is most similar to the hovered song.

Figure 4.2: Overview of the hovering functionality for one of the top 60 songs on the left and recommended song on the right on page 2, as discussed in Condition 3: Narrative Visualization 4.1.2

The user can then proceed onto page 3. Here the user sees a summary of the information they have received in the plot. This helps build a narrative by giving all of the information in a different format to the plot, which gives an opportunity to the user to understand all of the provided information. The view shows a table, where each row contains a recommended song, the genre of that song, its popularity, duration and the song that is most similar to it form the top 60 songs of the user. Above this table, the user sees a summary of the averages of these features for their top 60 songs. They see the most common genre of their songs, the average popularity and the average duration.

Page 4 moves the narrative further by introducing most influential features. The most influential features are those that explain the recommendations based on the top 60 songs. These features are listed and explained. In the middle of the screen, a similar plot is show as on page 2. The main difference here is the presence of the parallel coordinates plot below it. With this graph it is possible to see clustering and the distribution of the features of the songs. Through this, users are able to visualize why these features are influential, for example, most of the top 60 songs have a low valence, hence then it would make sense for the recommended songs to have low valence as well. Users are able to interact with this visualization by hovering over the points in the scatter plots, see figure 4.3. The lines in the PCP are colored to correspond to the points in the scatter plot. The last element on the page is a question. This question is used to test objective understandability. Based on the information provided so far are users able to correctly identify which feature is the most influential for this set of recommendation?

The tool has some limitations on interaction on this page. For example, the goal would be to allow for interaction with the PCP and the recommended song view, all linked together. If a line is hovered in the PCP, the song in the recommended songs view is highlighted and in the scatter plot and similarly for hovering over a song in the recommended songs view. However, due to limitations of the used system, it is not possible to implement this functionality. Unfortunately, this limitation was discovered late into the implementation, hence it was no longer feasible to change system.

A summary of the information from page 4 is presented on page 5. This is done to appeal to a wider set of users, since some might prefer to be able to read the information in a text form as well as see the visualization [33]. In the middle of the page, the averages and ranges of the values of the top 60 songs of the user for the most influential features are given. Below that, in a table view

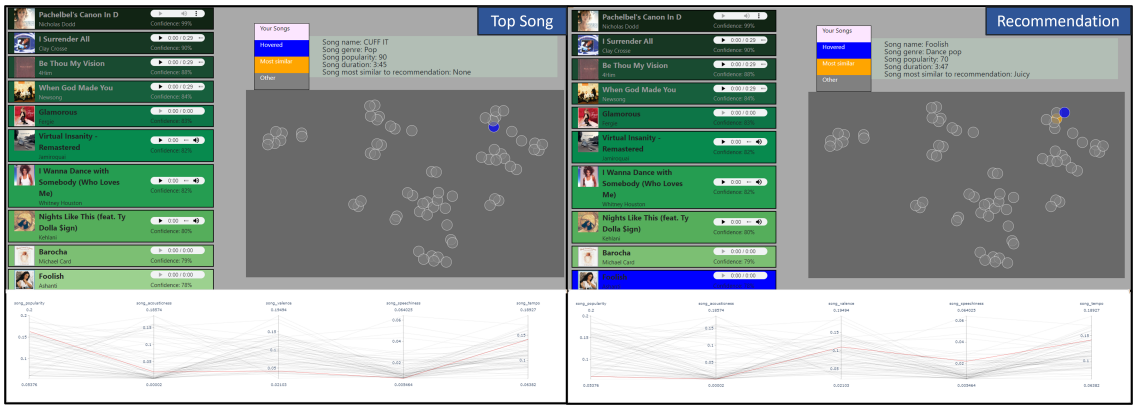

Figure 4.3: Overview of the hovering functionality for one of the top 60 songs on the left and recommended song on the right on page 4, as discussed in Condition 3: Narrative Visualization 4.1.2

the user can investigate the values for each of the 5 features for each of the recommended songs. The user can use the average values to put the values of the recommended songs into perspective. This summary also ensures that users do not miss any details, as going over all the details in the plot is much more time consuming.

Page 6 is identical to page 2, except for the data that is shown in the scatter plot. On this page the user can see their recommendations, top songs and all of the songs in the database of the tool in one visualization. This view allows the user to see in which clusters their songs lie, and to explore songs in a distance cluster. This further builds the narrative for the user as to why they got this specific set of recommendations. The songs of the user, both the top 60 and the recommended are likely to be close in this scatter plot, showing their similarity.

As the very last interaction with the tool, the users rate what they thought of the recommended songs. From here they proceed to the second part of the study.

Condition 2: Non-Narrative Visualization

As seen in figure 4.5, the first page gives an introduction to the *Recommendation Explanation* tool.

The key difference between the the narrative and non-narrative conditions is that in the narrative condition a narrative is built. This is done by spreading information over multiple pages and providing summaries about the insights from the shown plots. In the non-narrative condition, all the information is displayed on one page. This is the content of page 2. The content is the same as discussed in all of Section 4.1.2. The text on top gives an explanations of the two visualizations, a scatter plot and a parallel coordinates plot, the 5 most influential features are presented and defined, and the types of possible interactions are stated. In the middle of the page the scatter plot with the legend and the overview of the recommended songs are shown. Below that is the parallel coordinates plot and the understandability question. One difference to the narrative condition is the 'Show all' button above the scatter plot, see figure 4.5 page 2 a). In a). the same plot is shown as in the narrative condition page 2, figure 4.4. When clicking the button, the user sees a different plot, b). This plot shows the same view as in the narrative condition page 6. By selecting the 'Show less' button, the plot in view a). and the parallel coordinates plot are shown again.

Lastly, the user navigates to page 3 to rate the recommendations.

Condition 1: No Visualization/Control

An overview of the pages of the control condition of the *Recommendation Explanation* tool is presented in figure 4.6. On page 1 the user is introduced to the tool. When the user has interacted

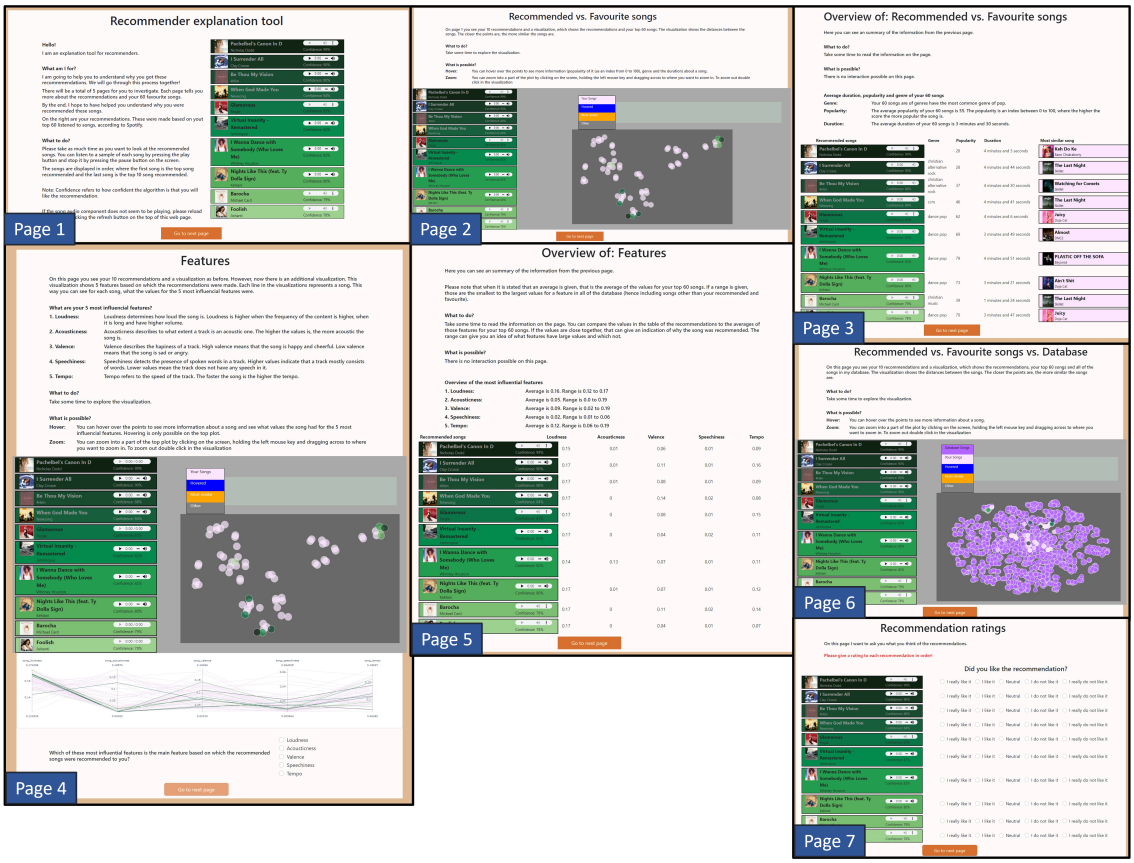

Figure 4.4: Overview of the pages of Condition 3: Narrative Visualization 4.1.2

with the page they proceed to page 2, to rate what they thought of the recommendations. Hence, this condition contains no explanation, only an overview of the recommendations.

After rating the recommendations, users proceed to the second part of the study.

4.1.3 Study Workflow Part 2

In the second part of the study, figure 4.7, the user starts with filling out the Post-Test Questionnaire. All questions on this page are mandatory, hence the user must provide an answer to each before continuing. On page 2 the user receives a debriefing. Based on the condition they were in, they will see the corresponding page: page 2_1 for the control condition, page 2_2 for the non-narrative condition and page 2_3 for the narrative condition. Lastly, the user arrives at page 3, where the study is concluded.

An element that is common to every page is interaction logging. On every page in the system, including the 4.1.2, the timestamp of the start and the end of the user's interaction with the page is recorded. For pages in the 4.1.2 containing a plot, the interactions with the points in the plot are recorded. The timestamp of the interaction and if the song is a recommended song, a top 60 song or a database song are recorded for further analysis.

Figure 4.5: Overview of the pages of Condition 2: Non-Narrative Visualization 4.1.2

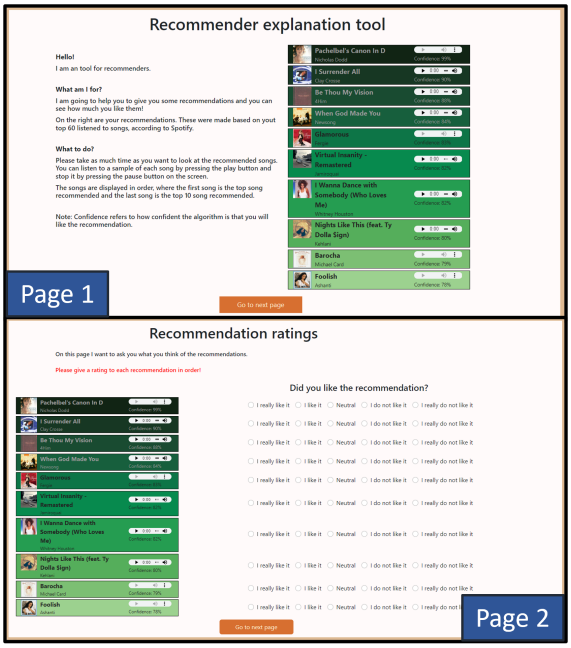

Figure 4.6: Overview of the pages of Condition 1: No Visualization/Control 4.1.2

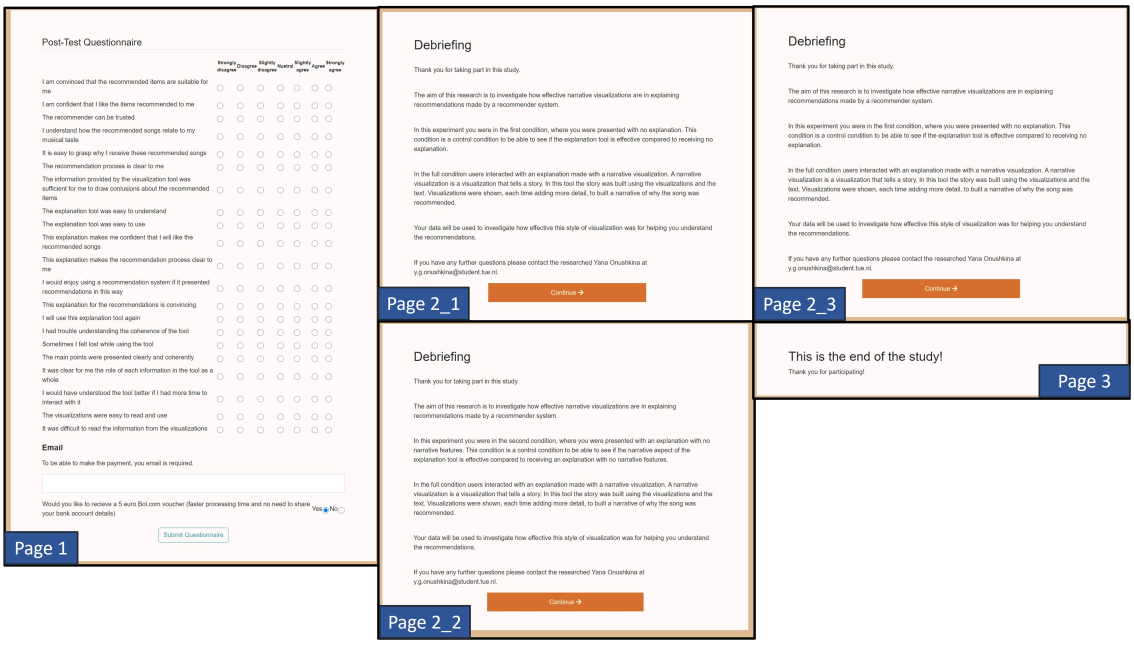

Figure 4.7: Overview of the second part of the system as discussed in section 4.1.3

Chapter 5

Method

5.1 Study design

In this section the method of the study is presented. The goal of the study is to evaluate the effectiveness of the tool in helping the user understand and trust the song recommendations they receive from a recommender system. For this the system described in Chapter 3 is used.

The study took place online with a between-subjects design. It is hosted on the server of PythonAnywhere in Europe. Users received a link to the application. The application guided users through the steps of the study as one would be in a live experiment. The steps the user is guided through were; introduced to the study, informed consent, Spotify authentication, intake questionnaire, interaction with the explanation, questionnaire per recommended song, post test questionnaire, debriefing and conclusion.

5.2 Users

A selection criteria is developed. Users have to fit the following requirements:

- Active premium Spotify account, the user must have listened to at least 60 songs.
- The user must not have any visual impairment.

The users were reimbursed with an amount of 5 euros for their participation. This amount is paid only in the case where users completed the whole study, from introduction to conclusion.

5.3 Study conditions

The conditions were created to test the effectiveness of the explanation. First factor of interest is if having an explanation leads to higher trust and understandability, hence control condition. The users in the first condition saw the recommendations made by the algorithm with no explanation. They have interacted limitedly with the tool, as described in Section 4.1.2. Secondly, the question if the created explanation, containing narrative features, is more effective at improving trust and understandability than an explanation without narrative features, non-narrative condition. This condition showed visualizations without any narrative features. Users saw all the same visualizations as in narrative condition, however, lacking the narrative aspect of the tool. The design of this condition of the tool is presented in Section 4.1.2. For the third condition the users interacted with the explanation as discussed in Section 4.1.2.

Construct	Items
Age	What is your age
Gender	What is your gender
Education	What is your highest level of education
Trust in technology	Technology never works I'm less confident when I use technology The usefulness of technology is highly overrated Technology may cause harm to people I prefer to do things by hand I have no problem trusting my life to technology I always double-check computer results
Visualization familiarity	I am competent when it comes to graphing and tabulating data I frequently tabulate data with computer software I graphed a lot of data in the past I frequently analyze data visualizations
Spotify usage	I use spotify

Table 5.1: Intake Questionnaire

5.4 Questionnaires

The study contained a total of three questionnaires, an intake questionnaire, questionnaire for confidence of each recommended item and a post-test questionnaire.

5.4.1 Intake questionnaire

Intake questionnaire contains all the data about the users before starting the study. These include some demographic questions about age, gender and education, as well as control items such as their existing trust in technology and familiarity with visualizations. Please note, in this work, we use the variable of existing trust in technology a user has before starting the study as a representation of their general algorithm aversion. For more details on the questions used in the Intake questionnaire please see Table 5.1.

- Demographic questions (age, gender, education).
- Trust in technology (General trust in technology scale; 7 items; Knijnenburg et al., 2012).
- Visualization familiarity (Visualization familiarity scale; 5 items; Kouki Santa Cruz et al., 2019).
- Spotify Usage (self-defined based on average monthly Spotify usage of 2022).

5.4.2 Questionnaire for each recommended item

To understand how accurate the users found the recommendations to be, this questionnaire is presented. For more details on the questions used to rate each item please see Table 5.2.

- Perceived accuracy (self-define; 1 item).

Construct	Items
Accuracy	I like the recommendation

Table 5.2: Questionnaire for each recommended item

5.4.3 Post-test questionnaire

The last questionnaire measured all of the variables of interest for the study. For more details on the questions used in the post-test questionnaire please see Table 5.3.

- Trust (Trust scale; 3 items; Nilashi et al., 2016).
- Perceived Understandability (Understandability scale; 3 items; Liang & Willemsen, 2021).
- Information sufficiency (Information sufficiency scale; 3 items; Nilashi et al., 2016).
- Perceived ease of use (Perceived ease of use scale; 2 items; Pu et al., 2011).
- Reception (Reception scale; 4 items; Kouki Santa Cruz et al., 2019).
- Use intentions (Use intentions scale; 1 item; Millecamp et al., 2019).
- Cognitive load (Cognitive load scale; 7 items; Macedo-Rouet et al., 2003).

5.4.4 Tool metrics

Aside from the questionnaires, the tool also recorded the interaction between the user and the tool. This interaction consisted of timestamps of landing on and leaving a page and hovering points in the visualizations.

Construct	Items
Trust	I am convinced that the recommended items are suitable for me I am confident that I like the items recommended to me The recommender can be trusted
Perceived Understandability	I understand how the recommended songs relate to my musical taste It is easy to grasp why I receive these recommended songs The recommendation process is clear to me
Information Sufficiency	The information provided by the visualization tool was sufficient for me to draw conclusions about the recommended items
Perceived ease of use	The explanation tool was easy to understand The explanation tool was easy to use
Reception	This explanation makes me confident that I will like the recommended songs This explanation makes the recommendation process clear to me I would enjoy using a recommendation system if it presented This explanation for the recommendations is convincing
Use intentions	I will use this explanation tool again
Cognitive load	I hand trouble understanding the coherence of the tool Sometimes I felt lost while using the tool The main points were presented clearly and coherently It was clear for me the role of each information in the tool as a whole I would have understood the tool better if I had more time to interact with it The visualizations were easy to read and use It was difficult to read the information from the visualizations

Table 5.3: Post test Questionnaire

5.5 Measurements

Further objective measurements were recorded by the tool. These measurements allowed for direct comparisons between perceived and observed features.

5.5.1 Question for understandability

The question "Which of these most influential features is the main feature based on which the recommended songs were recommended to you?" is added to test the objective understandability of the tool with the main goal of checking how well the users understood the data presented to them in the explanations. Presented with 5 options users have to choose which option which they believe has the most influence in the recommendation process. The five options were the 5 most influential features. Based on the visualizations presented to them, the users have to determine which of these features is the most influential. See figure 5.1 for an example of this question from the tool.

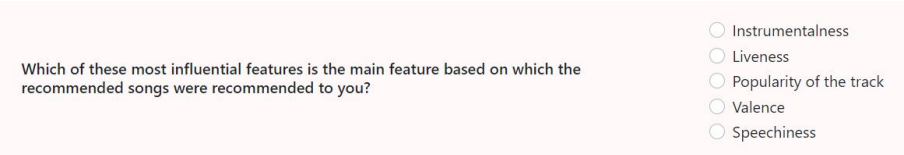

Figure 5.1: Example of question for understandably with options from the tool

5.5.2 Confidence of the algorithm

The recommender system outputs the confidence of each recommendation, see figure 5.2. This shows how fitting the recommender system finds this song is to the songs of the users. This measure showed the objective accuracy of the algorithm.

Figure 5.2: Example of how the confidence is displayed for a song in the tool

5.5.3 Time spent on a page

The time spent on a page and on each questionnaire is measured by the tool. Through this the participation and attention of the user is measured. In case a user spent a minute or less on each page, that indicates a lack of participation. This measurement additionally gives an insight into the complexity of the tool and the engagement of the user with the tool, as the more time they spend on a page the more they interact and engage with the tool. It is important to distinguish if the users found the tool to be complex or engaging, which we aim to do with the help of other measured variables, such as interaction with the plots and the perceived usability of the tool.

5.6 Interaction with plots

Lastly, the tool measured when a user hovered over a point in the visualization. Through this measure it is possible to further understand the level of interaction a user have with the tool. This measure provides metrics such as when this point is hovered and if the song is one that they were recommended, from their Spotify songs or from the database of songs.

5.7 Study procedure

The study started when users navigated to the home page of the tool. The users were greeted and directed to the informed consent form. Upon consenting to the study, the users have to consent to the use of their Spotify data by logging into their account and providing their account id. In case users did not consent, the study is terminated and the users thanked for their participation. Next, users filled in an intake questionnaire and proceeded to the tool. The tool users interacted with depended on the condition that they were assigned to. Upon completion of this interaction, users were asked about their opinion for each recommended song. Lastly, users completed a post test questionnaire after which they received a debriefing and the study is completed.

5.8 Data Analysis

The data analysis for this study is performed in Stata 17.

5.8.1 Outliers

A total of 151 people have participated in the study. After detecting outliers, a total of 132 participants remain. The 19 participants are removed due to the following outlier criteria.

- Participants must read the information provided in the pages of the tool.
- Participants must interact with visualizations if available.
- Participants must fill in the questionnaires honestly.
- Participants must participate once.

Two participants were removed from the study as they have participated twice. Their first attempt is kept in the study, as this is a between subjects design, hence on their second attempt they are already familiar with the tool.

One person is removed as they lack interaction through out. The amount of time they spent on each page of the tool is often far below average and they have no interaction with the plots. Another participant also seemed to speed through the tool at times and the questionnaires, producing patterned, such as responding with the same option to all questions, and illogical results. This participant would respond with a 5 out of 7 for a positively and negatively framed question. For example, for these questions:

- The visualizations were easy to read and use.
- It was difficult to read the information from the visualizations.

Further 5 participants spent far less time than average on answering the questionnaires and produced such patterned and illogical results. 9 participants were removed from the study as they either did not interact with the presented visualizations or did so limitedly (less than 10 points interacted per page). Lastly, one participant is removed from the study as they spent 78 hours on one of the pages. It is likely that they stepped away and came back after sometime, however, as time a participant spends interacting with the system is of interest to us, they were removed.

5.8.2 Factor Analysis

As we used a combination of defined scales in this study, not one existing scale, an exploratory factor analysis is performed to form and evaluate the variables/factors of interest. The most common type of factor analysis is a principle component analysis (PCA). In this study the iterated factor analysis (IFA) method is used. This method is similar to PCA with the addition that the analysis is iterated to obtain better estimates.

Code	Item	Factor 1	Factor 2	Uniqueness
algAv_1	Technology never works		0.5597	0.6846
algAv_2	I'm less confident when I use technology		0.6517	0.5723
algAv_3	The usefulness of technology is highly overrated		0.5239	0.7153
algAv_4	Technology may cause harm to people			0.9029
algAv_5	I prefer to do things by hand		0.4328	0.8083
algAv_6	I have no problem trusting my life to technology		-0.5768	0.6699
algAv_7	I always double-check computer results			0.9691
visFam_1	I am competent when it comes to graphing and tabulating data	0.7149		0.4758
visFam_2	I frequently tabulate data with computer software	0.8306		0.2995
visFam_3	I graphed a lot of data in the past	0.8877		0.2161
visFam_4	I frequently analyze data visualizations	0.8437		0.2918

Table 5.4: Factor Loading for the items of the Intake Questionnaire with loading score $\geq abs(0.3)$

For the analysis, the items from the Intake and the Post-Test questionnaires are looked at separately. This is because some of the items are similar in the two questionnaires, such as trust, however, in the Post-Test questionnaire the interest is in the trust in the recommender, where as in the Intake we are interested in the perspective of the participant before the study. The system is unaware of that, and is hence likely to group some of these items together.

Intake Questionnaire Variables

From the intake questionnaire we get the items for algorithm aversion and visualization familiarity. The IFA is performed on 11 items, only considering factors with an eigenvalue of at least 1. This indicates that the variance this factor explains is greater than the variance explained by a single observed item.

The results are shown in Table 5.4, where the distribution of the eigenvalues of these factors can be seen in figure 5.3. Two factors are identified, where one is related to trust and one to visualization familiarity, as expected. For visualization familiarity, codes $visFam\_1$ through 4, the items load to one factor, with high factor scores, showing a strong relevance to the factor, see figure 5.4 for a visual representation of this information. The low uniqueness scores indicate that $21-48\%$ of the explained variance of the item is unique. Therefore, the final variable visualization familiarity contains items $isFam\_1$ through 4.

For algorithm aversion, items $algAv\_1$ to 7, where 5 have a factor score of above 0.4 for factor 2. This means that items $algAv\_4, algAv\_5$ and $algAv\_7$ are not considered very relevant to that factor. Their uniqueness is also very high, hence they do not bring new explanation to the variance of the data. When looking at the questions, it can be seen that they are relatively similar. This can explain why only 3 out of the 7 items are loaded onto the algorithm aversion factor. The Cronbach's alpha of the items $algAv\_1-7$ is 0.612, which can be improved by removing items $algAv\_4$ and $algAv\_7$, resulting in a maximum alpha of 0.650. Hence both the factor analysis and the Cronbach's alpha result in the same set of items for the variable algorithm aversion, we use the factor analysis to create the variable.

Post-Test Questionnaire Variables

The results of the second factor analysis are presented in Table 5.5, the distribution of the eigenvalues of these factors can be seen in figure 5.3 and the loading plot in figure 5.4. Here a number of items have been removed due to low factor scores ($cognitiveLoad\_5$) and loading on multiple factors ($understand\_3, reception\_1, reception\_3, reception\_4, use\_intentions$). Hence, 14 items remain loading onto 2 factors. This is not what has been expected, as 7 different factors

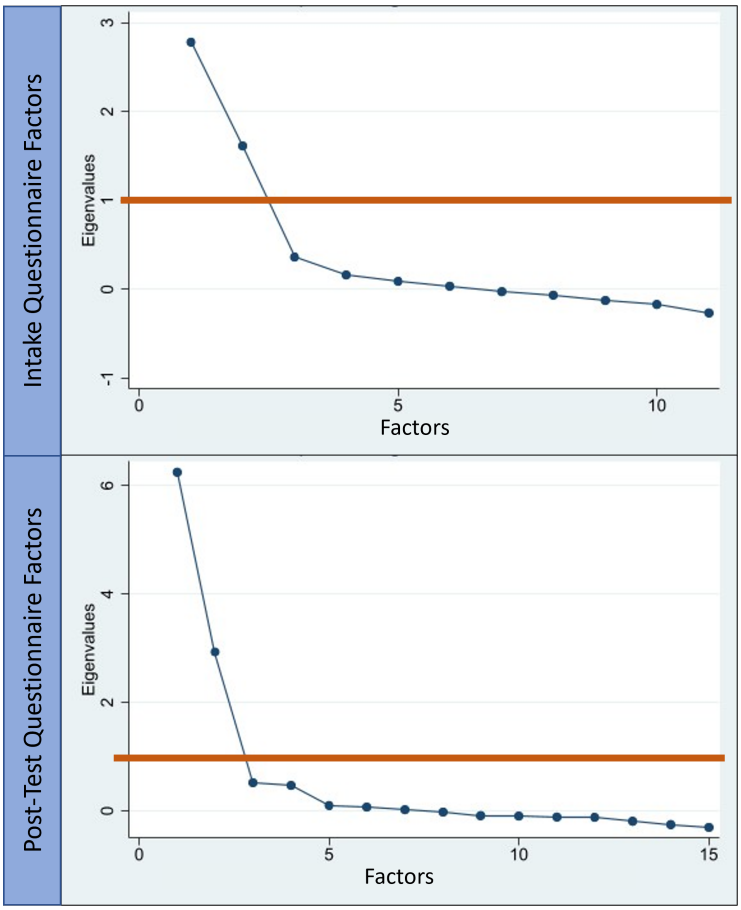

Figure 5.3: Overview of the eigenvalues of factors for factor analyses of Intake and Post-Test questionnaires

were intended out of these items. However, considering the questions asked in these items, the proposed factors are logical. There is a lot of overlap in the items, making it difficult to create 7 separate items. Additionally, due to the fact that the items have high factor scores, ranging from 0.63 to 0.92, it is clear that the items have a lot of common variance, hence separating them into multiple factors would lead to high correlations, which are highly undesirable. Hence, due to this the items will be combined into a smaller number of factors.

Inspecting the eigen values of the factors in this factor analysis we notice that there is a third factor with a value of 0.7697, just below the orange line in figure 5.3. Since we expected more factors, and this factor is close to 1, we consider the results with 3 factors. The outputs of this factor analysis, after removing items due to multiple loading and low factor scores, can be seen in Appendix B in Table B.1. Three separate factors are derived with high factor scores ($FS > 0.7$) and low uniquenesses ($U < 0.46$). In this factor analysis, we obtain a separate variable for trust and understandability, which is desired since the research question focuses on those two specifically and while we assume there is a relation between the two, they are not exactly the same. However, the correlation between these factors, specifically Factors 1 and 2 is relatively high, $r = 0.5497$. The correlation between two variables created as shown in Table 5.5 have a correlation of 0.3224. Since we want to avoid adding variables to a regression that explain each other, have high correlation, better than they explain the target variable, we create 2 variables based on the post test questionnaire.

The two created variables can be interpreted in the following way. Factor 1 contains items asking questions of how user friendly the tool is, hence a variable of *usability*. The second factor con-

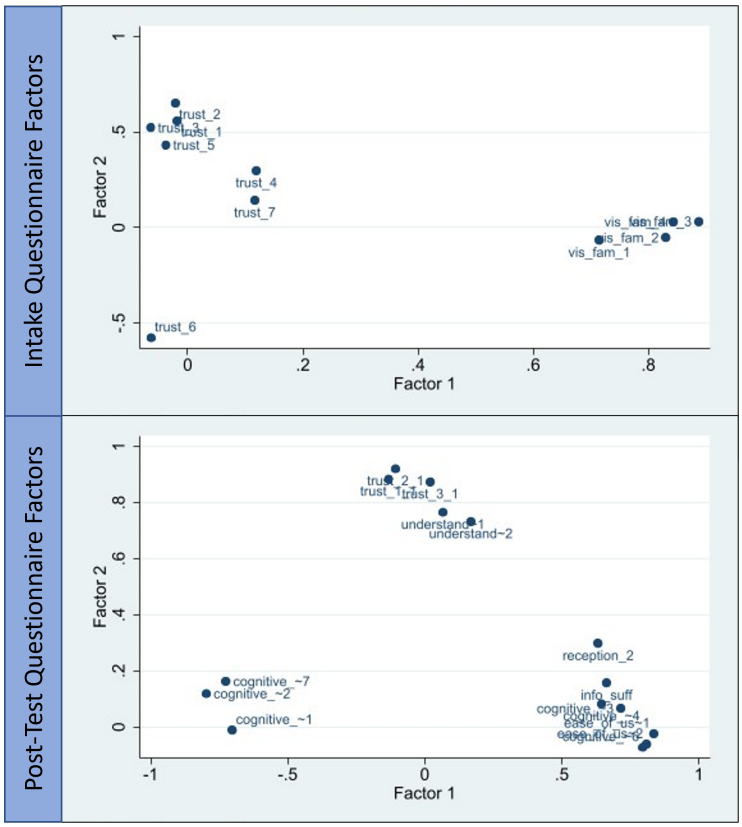

Figure 5.4: Loading plot of items for factor analyses of Intake and Post-Test questionnaires

tains items for trust and understandability, hence we create a joint variable of $trust\_understandability$.

5.8.3 Hypotheses Testing

Hypotheses testing is done through the use of t-tests for RQ1 and multiple-regression models for RQ1 and RQ2 with target, predictor and moderator variables. This analysis is supplemented with summary statistics about the data and visualizations.

Code	Item	Factor 1	Factor 2	Uniqueness
trust_1	I am convinced that the recommended items are suitable for me		0.8838	0.2722
trust_2	I am confident that I like the items recommended to me		0.9208	0.2004
trust_3	The recommender can be trusted		0.8743	0.2243
understand_1	I understand how the recommended songs relate to my musical taste		0.7657	0.3784
understand_2	It is easy to grasp why I receive these recommended songs		0.7326	0.3604
easeOfUse_1	The explanation tool was easy to understand	0.8349		0.3145
easeOfUse_2	The explanation tool was easy to use	0.8099		0.3707
reception_2	This explanation makes the recommendation process clear to me	0.6325		0.3971
cognitiveLoad_1	I hand trouble understanding the coherence of the tool	-0.7018		0.5026
cognitiveLoad_2	Sometimes I felt lost while using the tool	-0.7967		0.4087
cognitiveLoad_3	The main points were presented clearly and coherently	0.7155		0.4549
cognitiveLoad_4	It was clear for me the role of each information in the tool as a whole	0.6451		0.5446
cognitiveLoad_6	The visualizations were easy to read and use	0.7960		0.3960
cognitiveLoad_7	It was difficult to read the information from the visualizations	-0.7251		0.5194
infoSufficiency	The information provided by the visualization tool was sufficient for me to draw conclusions about the recommended items	0.6629		0.4729

Table 5.5: Factor Loading for the items of the Post-Test Questionnaire with loading score $\geq abs(0.3)$

Chapter 6

Results

In total 132 participants' data is used for the data analysis, with a mean age of 24.7, Table 6.1. The majority of the participants are female (82), 48 of the participants are male, one transgender male and one preferred not to disclose their gender. Of those, 48 participants are in the control condition, 51 are in the Non-Narrative condition and 37 are in the Narrative condition. The majority of the participants have a highest level of education (achieved or current) of bachelors (64) or masters (58). Five participants are in highschool and 5 are PhD candidates. All of these participants agreed to participate in the experiment, and 127 participants agreed to sharing their data externally. The majority (86) of the participants chose to get the compensation paid directly to their bank account, with the remaining 50 opting for a Bol.com card.

Variable	Mean	Std. dev.	Min	Max
Age	24.7	8.16	15	74

Table 6.1: Summary Statistics for age

6.1 Summary Statistics

The distribution of the responses of participants for their Spotify usage can be seen in Table 6.1. The majority of participants, 63 out of 132, use Spotify for more than 50 minutes a day. The distribution of responses is similar across the 3 conditions, where 13-25% of participant use Spotify less than 50 minutes a day, 27-33% use if for around 50 minutes a day and 42-52% use it for more than 50 minutes a day.

Spotify Usage	Frequency	Percent (%)
Less than 50 minutes per day	29	21.8
Around than 50 minutes per day	41	30.8
More than 50 minutes per day	62	47.4

Table 6.2: Summary Statistics for Spotify usage

6.1.1 Interaction

Interaction is broken down into number of points the participant hovered over in the available plots and the time they spent interacting. This time is split into interacting with the system, study related pages only, and with the *Recommendation Explanation* tool.

Duration on page	Mean (sec)	Std. dev. (sec)	Min (sec)	Max (sec)
Landing/Home	13.2	40.0	1	315
Informed Consent From	126.3	689.6	2	7448
Spotify Authentication	9.7	35.1	1	325
Intake Questionnaire	159.8	585.6	27	6741
Post-Test Questionnaire	134.3	57.1	50	381
Debriefing	22.5	39.9	1	390

Table 6.3: Summary of how long the participants stayed on the pages part of the system (study) only

The interactions with the study related pages are shown in Table 6.3. No anomalies are observed, hence as the interaction with these pages is not relevant for this study, we leave it out of the discussion. The interactions with the *Recommendation Explanation* tool are presented in Table 6.4 and in figure 6.1. On average, participants spent increasingly more time interacting with the tool with each condition, around 5 minutes for the control condition, 12 minutes for the non-narrative and 17 for the narrative condition. This is explained by the increased number of pages with each condition. Similarly participants spent more time on the welcome page with each condition, around 3 minutes for control, 5 minutes for non-narrative and 6 minutes for the narrative condition, as more information is presented with each condition.

Looking at the time spent interacting with the explanation pages, in the narrative condition participants spent more time in total (6 minutes) compared to the non-narrative condition (4 minutes). Looking at Table 6.5 and figure 6.2, it can be seen that the total number of points hovered in the non-narrative condition is almost half of those in the narrative condition. Hence, while being presented with the same information, the layout of the narrative explanation has encouraged users to interact more with the tool.

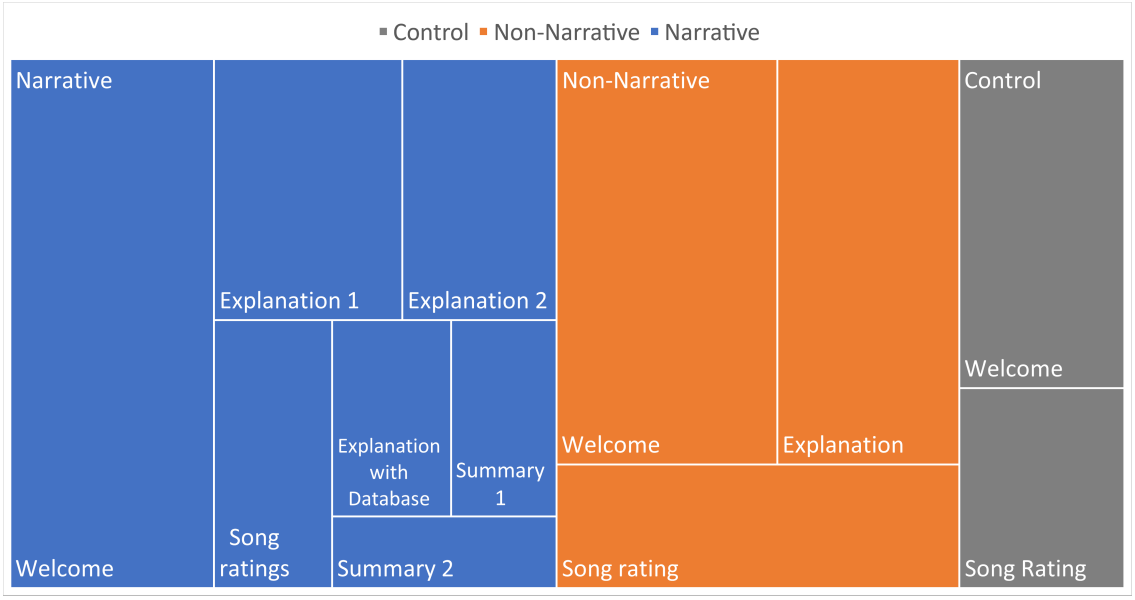

Figure 6.1: Overview of how long the participants stayed on the pages part of the *Recommendation Explanation* tool per condition

For the number of points interacted with in the non-narrative condition, we see that more points were interacted on the *a.* compared to the *b.* screen. It seems that on the *b.* screen users

Condition	Duration on page	Mean (sec)	Median	Std. dev. (sec)	Min (sec)	Max (sec)
control						
	Welcome	179.9	137	187.1	12	1147
	Song rating	109.2	66.5	116.4	17	640
	Total	289.6	222.5	254.6	57	1415
Non-Narrative						
	Welcome	296.5	140.5	609.7	35	3689
	Explanation	244.6	229	110.5	35	624
	Song rating	165.2	80	350.2	24	2244
	Total	731.0	538	768.5	112	4029
Narrative						
	Welcome	357.2	157	785.8	19	4754
	Explanation 1	162.8	118.5	163.0	27	957
	Summary 1	68.6	53	61.3	10	325
	Explanation 2	133.6	118	76.1	31	321
	Summary 2	53.4	34	81.7	3	485
	Explanation with Database	62		66.7	3	231
	Song ratings	104.9	85	72.8	19	277
	Total	1020.8	789	949.8	252	5128

Table 6.4: Summary of how long the participants stayed on the pages part of the *Recommendation Explanation* tool per condition

mostly hovered over the database songs. A similar pattern is seen in the narrative condition, where the most points hovered over on page 6 are the database songs. As the visualizations on page 2 *b.* in non-narrative and page 6 in the narrative contain the same information about the top songs and the recommendations as on page 2 *a.* and on page 2/4, it is logical for participants to have used this screen for exploring the database songs as those have not been presented before. There is a notable difference in the interaction of these 2 screens between the two conditions. In the non-narrative condition participants interacted with around 24 points on average, where as that number is around 3 times as large in the narrative condition.

For the narrative condition we can see that the participants spent similar amounts of time looking at the explanation 1 and 2 pages. Looking at the hovered points, the interaction in that time on the explanation 2 page is almost half of explanation 1. This indicates that participants have spent some of that time investigating the parallel-coordinates plot. On average users spent around a minute on the summary pages, which is the expected time as that gives enough time to read through all of the information on the page for an average user who would be assumed to have picked up most of the presented information from the plots themselves as the summary pages act as supplementary information for those who were not able to fully understand the information presented in the plots.

6.1.2 Check question

Table 6.6 shows the results participants got on the check question for their understanding of their top 5 most influential features. Both groups on average scored relatively low, $M_{Non-N} = 0.2$ and $M_N = 0.3$, hence most participants did not correctly identify the most influential feature in their recommendations. The mean is larger in the narrative condition indicating that the narrative

Condition	Page	Mean (#)	Median (#)	Std. dev. (#)	Min (#)	Max (#)
Non-Narrative						
	Page 2 a - all	90.1	90.5	55.2	1	209
	Page 2 a - recommended	20.8	18	17.4	0	76
	Page 2 a - top songs	69.3	68.5	42.6	1	172
	Page 2 b - all	28.1	15	43.1	0	187
	Page 2 b - recommended	0.9	0	2.0	0	10
	Page 2 b - top songs	3.0	0	6.9	0	39
	Page 2 b - database songs	24.3	13.5	38.6	0	171
	Total	236.5	220	147.5	12	736
Narrative						
	Page 2 - all	88.2	89	63.7	0	323
	Page 2 - recommended	21.7	17	16.4	0	72
	Page 2 - top songs	66.6	62	49.9	0	251
	Page 4 - all	47.8	31.5	49.7	0	179
	Page 4 - recommended	12.8	9.5	13.9	0	61
	Page 4 - top songs	34.9	21	39.5	0	161
	Page 6 - all	81.8	64	71.4	0	269
	Page 6 - recommended	1.3	0	2.2	0	9
	Page 6 - top songs	6.7	2.5	10.9	0	46
	Page 6 - database	73.8	59	63.2	0	217
	Total	435.6	411	293.0	38	1386

Table 6.5: Summary of how many points the participants interacted within the *Recommendation Explanation* tool per condition

explanation did improve their objective understanding, however, this effect is not statistically significant, $t(82) = -1.5, p = 0.1$.

Variable	Mean	Std. dev.	Min	Max
Check question - Non-Narrative	0.2	0.4	0	1
Check question - Narrative	0.3	0.5	0	1

Table 6.6: Summary Statistics for the Check Question

6.1.3 Accuracy

The average accuracy of the recommended songs, as given by the algorithm, is around 89%, where the lowest accuracy is 76.1%. Hence, the items recommended to the user are considered as highly relevant. However, based on the perceived accuracy, Table 6.7, participants did not agree with their recommendations. The average accuracy score given is 63% with a lowest of 36% on average over all of the recommended items. These results are similar among all three conditions. Hence, the narrative condition does not seem to have an effect on perceived accuracy.

A question of interest for this study, is how accuracy affects perceived accuracy (H16). The results of this test are presented in Table 6.8. Accuracy negatively affects perceived accuracy, however, this is not a significant effect, $r(102) = 0.005, p > 0.05$. This suggests that presenting participants with the confidence of the algorithm in a recommendation does not affect their confidence in the fittingness of the recommendation as the amount of variance of perceived accuracy

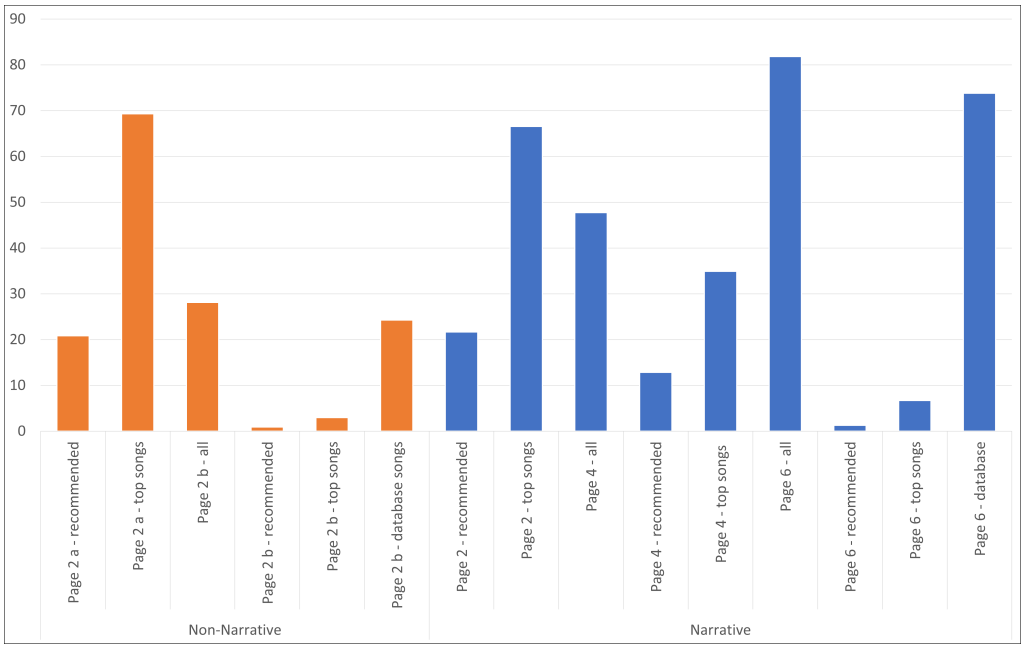

Figure 6.2: Overview of how many points the participants interacted within the *Recommendation Explanation* tool per condition

explained by accuracy is almost none.

6.1.4 Intake Questionnaire

Overview of the variables Algorithm Aversion and Visualization familiarity are presented in Table 6.9. These are the variables created using the factor analysis as discussed in Section 5.8.2. The distributions in figure 6.3 show that in general participants tend to have average visualization familiarity with a few individuals with really low or high scores. Similarly, most participants have average algorithm aversion with a few individuals with really low values, meaning low algorithm aversion.

6.1.5 Post-Test Questionnaire

Overview of the variables Trust/Understandability and Usability is presented in Table 6.10 and figure 6.4. Usability is distributed with a large number of people having an average score, some with above average and few with a high score (2.0). Similarly for below average where some people have low scores and only a few have a really low score of -2.4. For Trust/Understandability there are two peaks. One large peak is at scores around -1 and a smaller peak at around 1.5. This suggests that there is some factor that is effecting the scores of participants for trust/understandability. Looking at the variables per condition, there is a clear trend. For Trust/Understandability the mean improves from the control condition to the non-narrative to narrative. With usability an almost opposite effect is seen. The mean of usability for the control condition is around 0.6 higher than the mean for the non-narrative or narrative conditions. This suggests that participants found the tool to be very useful in the control condition and not that useful in the other two conditions.

Variable	Mean (%)	Std. dev. (%)	Min (%)	Max (%)
Accuracy	88.9	5.2	76.0	98.8
Accuracy_C	88.0	6.0	76.3	98.8
Accuracy_{Non-N}	89.3	4.5	76.8	98.0
Accuracy_N	89.2	4.9	76.1	97.3
Perceived Accuracy	63.2	11.0	36.0	86.0
Perceived Accuracy_C	62.7	11.8	36.0	86.0
Perceived Accuracy_{Non-N}	62.6	9.9	4.0	82.0
Perceived Accuracy_N	64.6	11.9	44.0	86.0

Table 6.7: Summary Statistics for Accuracy and Perceived Accuracy

Variable	Coefficient (β)	Std. err.	t	p	95% conf. interval
Accuracy	-0.2	0.2	-0.7	0.5	-0.6 - 0.3

Table 6.8: Regression results for H16: The effect of Accuracy on Perceived Accuracy

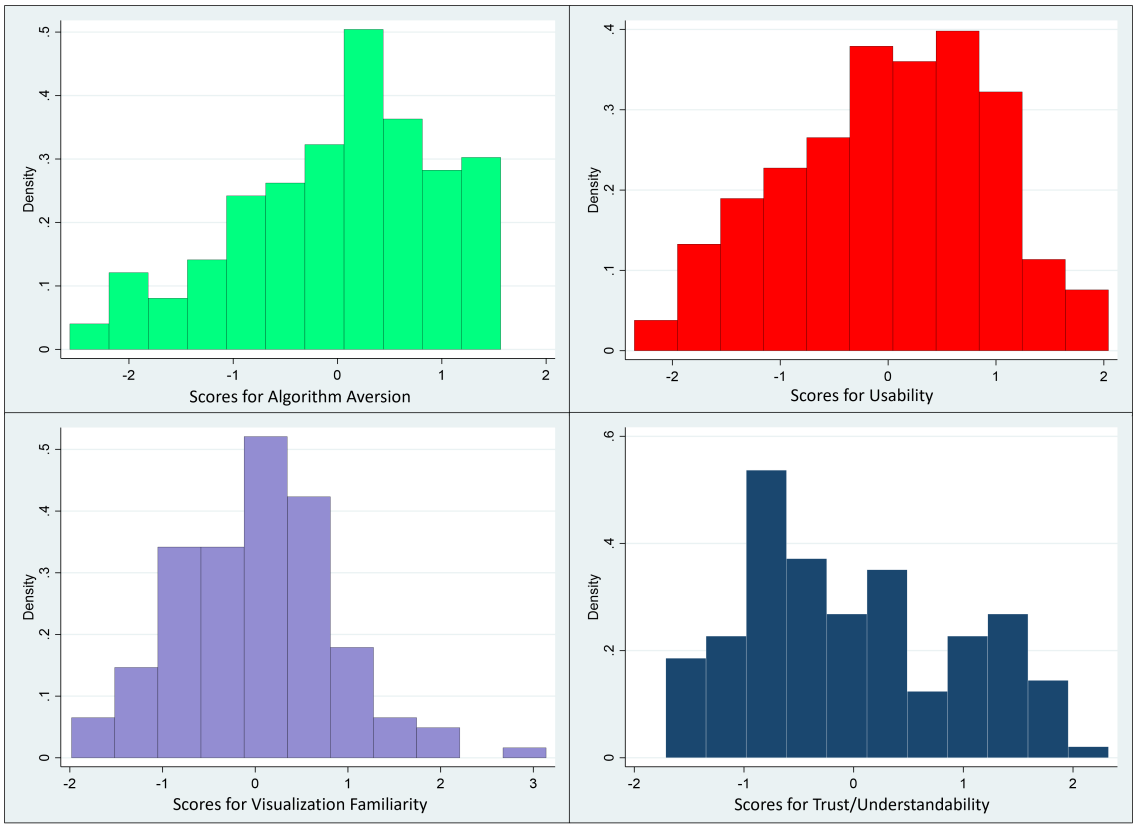

Figure 6.3: Distribution of variables from Intake and Post-Test Questionnaires

Variable	Mean	Std. dev.	Min	Max
Algorithm Aversion$_C$	0.1	0.9	-1.6	3.1
Algorithm Aversion$_{Non-N}$	-0.1	0.8	-1.7	1.8
Algorithm Aversion$_N$	0	0.8	-2.0	1.9
Visualization Familiarity$_C$	0	0.9	-2.6	1.6
Visualization Familiarity$_{Non-N}$	-0.1	0.9	-1.9	1.4
Visualization Familiarity$_N$	-0.1	1.0	-2.2	1.5

Table 6.9: Summary Statistics for the Intake Questionnaire variables

Variable	Mean	Std. dev.	Min	Max
Trust/Understandability$_C$	-0.1	1.0	-1.7	2.3
Trust/Understandability$_{N-N}$	0	0.9	-1.7	1.6
Trust/Understandability$_N$	0.1	1.0	-1.6	1.7
Usability$_C$	0.4	0.7	-1.7	1.8
Usability$_{N-N}$	-0.3	1.0	-2.4	1.7
Usability$_N$	-0.2	1.0	-2.1	2.0

Table 6.10: Summary Statistics for the Post-Test Questionnaire variables

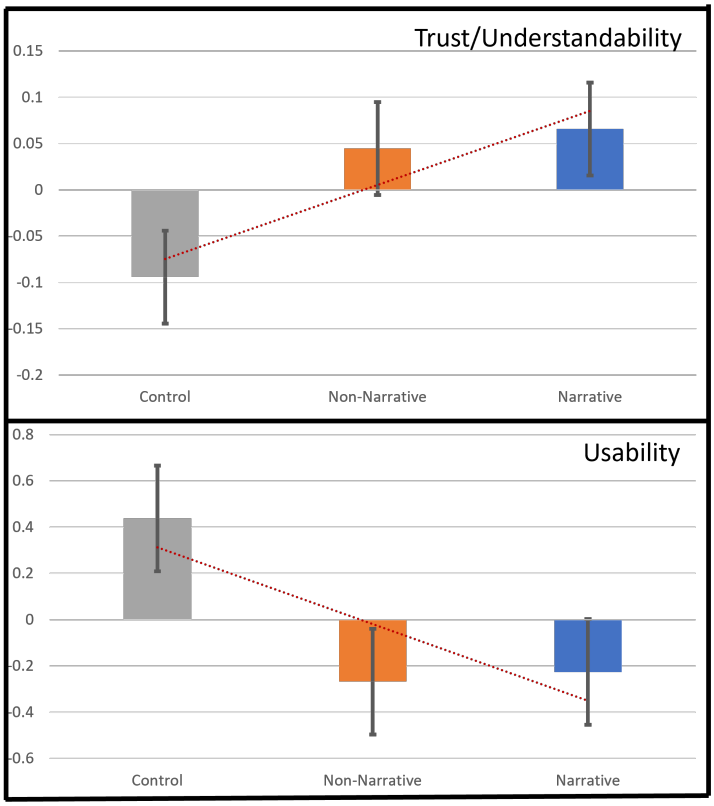

Figure 6.4: Overview Statistics for the Post-Test Questionnaire variables

6.2 Hypothesis Testing

The hypothesis testing is aimed at providing an answer to the 17 hypothesis derived for this research. However, due to the results of the factor analysis for the post-test questionnaire, a number of hypotheses have to be dropped as those effects have not been measured in the study/detected in the factor analysis. The remaining hypotheses are:

- H1: non-Narrative will increase trust/understandability in the recommender, compared to control
- H2: Narrative will increase trust/understandability in the recommender, compared to non-Narrative
- H5: non-Narrative will have a higher usability, compared to control.
- H6: Narrative will have a higher usability, compared to non-Narrative.
- H8: Higher usability leads to higher trust/understandability.
- H12: Higher interaction with the tool leads to higher perceived trust/understandability.
- H13: Higher visualization familiarity leads to higher usability.
- H14: Lower algorithm aversion leads to higher trust/understandability.
- H16: Higher objective accuracy leads to higher perceived accuracy.
- H17: Higher perceived accuracy leads to higher trust/understandability.

The hypothesis H1 - H4 were combined into 2 hypothesis, H1 and H2, as trust and understandability are one factor. For hypotheses H5, H6, H8, H15 instead of reception, we look at usability as reception overlaps with usability. For hypotheses H8, H12, H14, H15 and H17 we look at the target of trust/understandability instead of trust of understandability. H13 now looks at the target of usability instead of reception. Lastly, H16 is left unchanged.

Having reformulated some of the hypotheses, we proceed with the analyses. As discussed in section 5.8.3, the hypotheses testing is done using multiple regressions. For this, correlations between predictor variables are checked to ensure that those are not included together in the regression model. Table 6.11 shows correlations of all predictor variables. The highest correlation is between time and the control condition variables ($r(96) = -0.6$). The correlation between the interaction variable (int) and narrative condition is also over 0.5, $r(96) = 0.6$. This correlations tells us that around 50% of variance is common between these pairs of variables, hence needs to be monitored how these variables behave together in the regression.

	use	visFam	algAv	time	int	control	narrative	accuracy
visFam	-0.03							
algAv	0.02	-0.10						
time	-0.26	0.14	-0.14					
int	-0.08	0.07	-0.02	0.61				
control	0.36	0.02	0.06	-0.59	-0.62			
narrative	-0.08	0.05	-0.04	0.41	0.57	-0.42		
accuracy	0.01	-0.05	0.09	-0.04	-0.03	-0.10	0.12	
accuracyP	0.21	0.15	-0.20	0.23	0.06	-0.05	0.14	-0.10

Table 6.11: Correlation of predictor variables

Aside from the predictor variables, we are also interested in the interaction effects between certain variables across the conditions. The variables that are expected to vary depending on

the condition are time, interaction, usability and perceived accuracy. Hence, for these interaction variables are created and the correlation between those is presented in Table 6.12. The correlations between these and the predictor variables from Table 6.11 are not presented as no high correlations were found.

Table 6.12 shows that there are a number of high correlations between items. However, all of these items are from the same condition if they have high correlation, for example time_N and accuracy x N ($r(96) = 1.0$), time_C and accuracyP_C ($r(96) = 1.0$) and narrative and int_C ($r(96) = 79$). While the correlations are hence logical, we bare them in mind as we proceed with the analyses.

	use	time	int	accuracyP	control	narrative	use$_C$	use$_N$	int$_C$	time$_C$	time$_N$	accuracyP$_C$
time	-0.11											
int	-0.29	0.60										
accuracyP	0.21	0.06	0.23									
control	0.40	-0.62	-0.59	-0.05								
narrative	-0.15	0.58	0.43	0.14	-0.41							
use$_C$	0.56	-0.35	-0.48	0.02	0.56	-0.23						
use$_N$	0.47	0.14	-0.05	0.11	0.09	-0.22	0.05					
int$_C$	0	0.82	0.41	0.16	-0.32	0.79	-0.18	0.10				
time$_C$	0.37	-0.61	-0.50	-0.01	0.99	-0.41	0.51	0.09	-0.32			
time$_N$	-0.15	0.59	0.47	0.16	-0.41	0.94	-0.23	-0.20	0.8078	-0.4026		
accuracyP$_C$	0.40	-0.60	-0.52	0.10	0.97	-0.40	0.56	0.09	-0.32	0.98	-0.40	
accuracyP$_N$	-0.13	0.59	0.45	0.23	-0.40	0.98	-0.22	-0.16	0.80	-0.40	0.98	-0.39

Table 6.12: Correlation of interaction and predictor variables variables

6.2.1 Regression Analysis - Trust/Understandability

The results of the regression are shown in Table 6.13. The R^2 score of the model is 0.3. The model contains a lot of insignificant effects, hence in order to obtain a better fit, we remove some of the small effects. These variables that are used in the final regression are usability (use), visualization familiarity (visFam), perceived accuracy (accuracyP), interaction (int), control, narrative, time and interaction effect of time for the conditions (time_C and time_N). The results of the second regression are shown in Table 6.14.

Variable	β	Std. err.	t	p	95% conf.
use	0.4	0.1	2.8	0	0.1 - 0.7
visFam	-0.1	0.1	-1.3	0.2	-0.3 - 0.1
algAv	-0.1	0.1	-0.6	0.5	-0.3 - 0.1
time	0.5	0.2	2.0	0	0 - 1.0
int	-0.002	0.001	-1.8	0.1	-0.004 - 0.0002
accuracyP	5.9	1.2	5.0	0	3.5 - 8.2
control	0.3	1.9	0.1	0.9	-3.4 - 3.9
narrative	-1.9	2.4	-0.8	0.4	-6.6 - 2.9
use_C	0.3	0.3	1.2	0.2	-0.2 - 0.9
use_N	0.02	0.2	0.1	0.9	-0.4 - 0.5
time_C	-0.1	0.3	-0.5	0.7	-0.8 - 0.5
time_N	0.2	0.4	0.5	0.7	-0.6 - 0.9
int_N	0.002	0.001	1.8	0.08	-0.0002 - 0.005
accuracyP_C	0.9	1.5	0.6	0.5	-2.0 - 3.9
accuracyP_N	-0.3	1.7	-0.2	0.8	-3.7 - 3.0
cons	-4.4	1.8	-2.5	0	-7.9 - -0.9

Table 6.13: First Regression Model for the target variable Trust/Understandability

The second regression model has a R^2 of 0.7, which is a great improvement from the first model. There are 6 significant effects in this model. These are usability (use), visualization familiarity (visFam), perceived accuracy (accuracyP), interaction (int), control, interaction of time and the control condition. Out of these perceived accuracy has the largest effect ($\beta = 6.1, p < 0.05$). The control condition has the next largest effect at $\beta = 3.1, p < 0.05$. Usability and visualization familiarity have similar effect sizes, however, visualization familiarity has a negative coefficient of -0.2. Hence the more familiar a participant is with visualizations, that has a negative impact on their trust and understandability of the system compared to the non-narrative condition. Similarly, spending more time interacting with the tool in the control condition has a significant negative effect on the target variable relative to time spent in the non-narrative condition ($\beta = -0.5, p < 0.05$). Interaction has a really low effect size ($\beta = 0.001, p < 0.05$). The narrative condition results in a negative effect size ($\beta = -0.3, p > 0.05$) indicating that being in the control condition has a negative effect on trust and understandability. The log of the time interacted with the tool has an effect size of 0.3, not significant. Lastly, the interaction of time and the narrative condition has a positive effect on trust and understandability compared to the non-narrative condition. This effect is, however, not significant.

Based on the regression depicted in Table 6.14 we can see that there are 2 variables that play a role in explaining trust/understandability. First is the condition itself, where it can be seen that participants in the control condition have higher trust/understandability compared to the non-narrative condition, and those in the narrative have lower. The other variable is time, and more importantly interaction between time and condition. If users spend more time in the control condition this decreases their trust/understandability, while spending more time in the narrative condition increases the target variable, relative to the non-narrative condition. This relationship between the interaction of time and condition and its affect on the target variable per condition is shown in figure 6.5. In this plot we see that the more time a participant spends in the

Variable	β	Std. err.	t	p	95% conf.
use	0.2	0.1	3.1	0	0.1 - 0.4
visFam	-0.2	0.1	-2.9	0	-0.3 - -0.1
accuracyP	6.1	0.6	10.25	0	5.0 - 7.3
int	0.001	0	1.6	0.1	-0.0001 - 0.001
control	3.1	0.9	2.0	0.05	0 - 3.6
narrative	-0.3	1.8	0.3	0.8	-3.0 - 4.0
time	0.3	0.6	1.9	0.06	0 - 0.6
$time_C$	-0.5	0.2	-2.4	0.02	-1.0 - 0.1
$time_N$	0.03	0.2	0.11	0.11	-0.5 - 0.5
cons	-5.77	1.1	-5.3	0	-8- -3.2

Table 6.14: Final Regression Model for the target variable Trust/Understandability

narrative condition, the higher their trust and understandability. For the non-narrative and control conditions, the effect of time is similar, where in general the estimated trust/understandability in the control condition is higher than in the non-narrative and they increase in a similar pace with time. Hence, this shows that participants in the narrative condition must spend at least 6.3 log seconds (a bit less than 10 minutes) in order to trust and understand the tool as well as participants in the non-narrative condition, and around 11 minutes to trust and understand better than those in the control condition. A total of 21 (out of 37) participant in the narrative condition spent a bit less than 10 minutes using the tool and 19 of them spent more than 11 minutes. Hence, it would be expected that more than half of the participants in the narrative condition have higher trust and understandability.

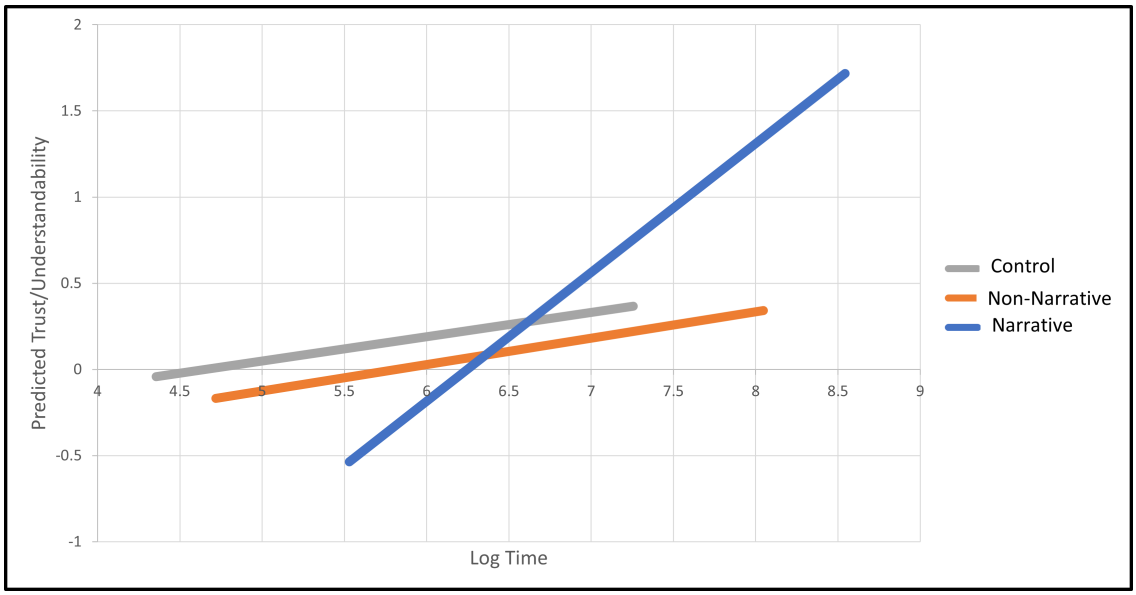

Figure 6.5: Estimated Means Plot of target variable Trust/Undestandability and predictor Log Time

With this, there is enough evidence to answer hypotheses H1, H2, H8, H12, H14 and H17. Firstly, looking at the control versus the non-narrative condition (H1) it was expected that the non-narrative condition, having a visualization versus not, leads to higher trust and understandability. This is also supported by the findings in figure 6.4 where the mean value of trust/understandability in the control condition is negative, while in the non-narrative condition the mean

is positive. However, the results of the regression, visualized in figure 6.5 show that participants in the control condition are expected to have higher trust/understandability of the tool than the non-narrative. Hence, this leads to the conclusion that having a visualization does not lead to better trust and understandability compared to having no visualization and even leads to decreased trust/understandability, hence H1 is rejected.

For the narrative compared to the non-narrative condition (H2) we see a large difference. The mean of the target variable, figure 6.4, in the narrative condition is larger than in the non-narrative condition. However, figure 6.5 shows that based on the regression, the trust/understandability in the narrative condition is on average much better than in the non-narrative condition, supporting H2, a narrative explantion is more effective in terms of trust/understandability compared to not having narrative features in the explanation and having no visualization.

Looking at hypotheses for the secondary research question, the results of the regression model in Table 6.14 support the hypotheses H8 and H17. Usability has a small positive and statistically significant effect ($\beta = 0.2$) on trust/understandability. Similarly perceived accuracy is also a significant predictor of the target variable, with the largest effect $\beta = 6.2$. Visualization familiarity has a negative effect $\beta = -0.2$, hence participant with more experience with plots trust and understand why they got the recommendations less than those who have less prior experience with visualizations. This is not the relation that was expected. Both H12 and 14 are rejected. Algorithm aversion (H14) is not a significant predictors of trust/understandability. Interaction with the tool (in terms of number of points hovered) while significant has a very small effect size, close to zero. Additionally, time itself is not a significant predictor, while the interaction of time and the control condition is and time and the narrative condition not. This does not provide conclusive evidence to support H12.

6.2.2 Regression Analysis - Usability

Predictors of usability and the effect of the conditions on usability is also a point of interest in the current research. Table 6.15 shows the results of a multiple regression with target variable usability. The model has an R^2 score of 0.3.

Variable	β	Std. err.	t	p	95% conf.
visFam	-0.1	0.1	-1.0	0.3	-0.3 - 0.1
algAv	0.1	0.1	0.6	0.5	-0.1 - 0.3
time	-0.1	0.3	-0.5	0.6	-0.7 - 0.4
int	0.002	0.001	1.6	0.1	-0.0004 - 0.004
accuracyP	4.9	1.5	3.3	0.001	2.0 - 7.9
control	4.6	2.1	2.2	0	0.4 - 8.8
narrative	1.5	2.6	0.6	0.6	-3.7 - 6.7
time_C	-0.3	0.4	-0.8	0.4	-1.0 - 0.4
time_N	-0.002	0.4	0	1.0	-0.8 - 0.8
int_N	-0.001	0.001	-0.5	0.6	-0.003 - 0.002
accuracyP_C	-3.2	2.0	-1.6	0.1	-7.1 - 0.7
accuracyP_N	-2.6	2.3	-1.1	0.3	-7.1 - 2.0
cons	-4.2	2.5	-1.7	0.1	-9.1 - 0.7

Table 6.15: First Regression Model for the target variable Usability

Similarly as before, the insignificant effects are taken out of the regression. For this case an intermediate model is used with variables interaction, perceived accuracy, control condition, narrative condition, time and interaction between time and conditions. In this model, both interaction effects of time and conditions are not significant ($p = 0.2$ and $p = 0.9$ respectively). Therefore, the interaction variables are taken out. The resulting model is shown in Table 6.16 where $R^2 = 0.3$. Hence, removing the insignificant variables has not changed the amount of variance of usability

that the model explains. Four out of the five predictors are significant, with narrative condition resulting in an insignificant negative effect ($\beta = -0.3, p = 0.3$). The largest effects are the control condition ($\beta = 0.8, p = 0.001$) and perceived accuracy ($\beta = 2.5, p = 0.002$) showing that in the control condition users score higher on usability compared to the non-narrative condition. Additionally higher perceived accuracy leads to higher usability. Interaction has a small positive and significant effect on usability, hence more interaction leads to higher usability scores. Lastly, the more time an individual spends using the tool, the less they perceive it to be usable.

Variable	β	Std. err.	t	p	95% conf.
int	0.001	0.0004	2.4	0.007	0.0001 - 0.002
accuracyP	2.5	0.8	3.3	0.002	1.1 - 4.3
control	0.8	0.2	3.6	0.001	0.3 - 1.3
narrative	-0.2	2.5	-1	0.3	-0.7 - 0.2
time	-0.3	0.1	-2.33	0.02	-0.6 - -0.05
cons	-0.1	0.8	-0.12	0.9	-1.9 - -1.6

Table 6.16: Final Regression Model for the target variable Usability

As we are interested in the effect on usability per condition, we present an estimated means plot with the target variable of predicted usability and condition on the x axis, figure 6.6. This plot clearly shows the desired relationship. Contrary to what is expected, the control condition results in the best usability scores, where all the scores are predicted above zero, hence above the average across all participants. The non-narrative condition gives usability of mostly below the average and the narrative condition results in very distributed scores. Based on the final regression for usability, narrative has a negative effect size compared to the the non-narrative condition, indicating that participants in the narrative condition find the tool less usable than in the non-narrative condition, however the effect is not significant. Hence, based on this H5 and H6 are rejected as the condition with no visualization is found to be most usable and a narrative visualization is not significantly more effective in terms of usability than either of the explanations.

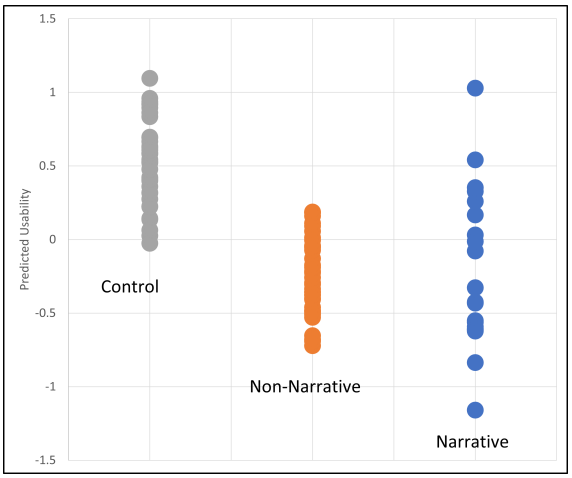

Figure 6.6: Estimated Means Plot of target variable Usability and predictor Condition

The remaining hypothesis, H13, is also rejected as visualization familiarity is not a significant predictor of usability. Unexpectedly, perceived accuracy of the recommendations is a good predictor of usability ($\beta = 2.5$) along with the control condition ($\beta = 0.8$) and spending time in the tool ($\beta = -0.3$). It is interesting to see that spending time in the tool is a significant predictor of usability but not of trust/understandability.

Chapter 7

Discussion

The results of the conducted study provide evidence for the primary and secondary objectives of the current thesis. The objectives of the thesis are: RQ1 - How effective, in terms of trust and perceived understandability, are narrative visualizations in explaining recommendations made by a recommender system? and RQ2 - What is the effect of algorithm aversion, visualization familiarity, accuracy, perceived accuracy, reception of the tool, interaction with the tool, perceived ease of use, cognitive load and information sufficiency on trust and perceived understandability of the recommender system? This latter objective, due to the changes in factors of the study, is rephrased to What is the effect of algorithm aversion, visualization familiarity, accuracy, perceived accuracy, usability, interaction with the tool on trust and perceived understandability of the recommender system? In this chapter we discuss the results for each objective further, answering the posed research question.

7.1 Effectiveness of the Three Explanations

The first research question looks at the effectiveness of the explanations. In the chapter on results (chapter 6), based on the presented evidence hypotheses H2 and H6 are supported and H1 and H5 are rejected. Based on hypotheses H1 and H2 we learn that having a visualization, and hence an explanation, compared to not does not have a conclusive effect on trust and understandability of the user towards the recommender system, while a visualization with narrative features increases trust and understandability compared to both a visualization without such features or no explanation at all. The former finding is surprising, as the work of Zhang Curley (2018) showed that the presence of an explanation leads to increased trust, hence our results seem to contradict these findings. However, a potential explanation for our finding comes from the design of the three conditions. Since the focus of the thesis is on narrative visualizations, to be able to investigate the effects of the narrative features all information in the explanation that was not part of the narrative-building techniques had to remain the same for the non-narrative condition. Hence, the explanation is optimized for the narrative condition, while for the non-narrative, due to a constraint of keeping as much as possible the same, the explanation contains all information in one page, which can be quite overwhelming for the participant. Hence, due to this, a participant can find the non-narrative explanation hard to understand and therefore not trust it.

The evidence found in support of H2 is in line with the literature discussed on narrative visualizations where earlier work found narrative visualizations to be most effective in increasing trust from the user [39]. However, interestingly the narrative explanation is found to be most effective, in terms of trust and understandability, only after the participants use the tool for at least 'enough' time, around 11 minutes. If the participant does not invest this time in interacting with the tool, they are expected to have lower trust and understandability of the recommendation process than in both of the other two conditions. This finding is interesting and also logical. The narrative explanation is built upon a storyline concept, where the participant is given more

information at every step and interaction. Hence, if a participant does not invest the time in reading the information provided by the tool and interacting with the given visualizations, they will not be able to follow the created story and hence are not expected to have high trust and understandability of the recommender.

While trust and understandability are the main objectives of this thesis, usability is also an important component of the effectiveness of an explanation. This is because if an explanation is not usable, then users will not come back to it. The results of usability for the three conditions are surprising. It was expected that since a narrative explanation increases trust and understandability the most, this leads to higher usability as perceived by the user. Our results show that the control condition, without any explanation, is seen as the most usable, the narrative explanation is the second most usable, and the non-narrative is the least usable. This discontinuity in the results for the narrative explanation suggests that the relationship between trust/understandability and usability is not as is expected in H8, discussed in section 7.2. While striking, these results also have an explanation. Firstly, the control condition is the simplest to interact with and use, since there are no plot and no explanation. Using the tool is straightforward and something they might already be familiar with from Spotify, get a list of recommendations and look over/listen to them. However, the non-narrative and the narrative conditions are much more complex, as it involves interacting with potentially a new concept to the participants, an explanation for their recommendations, along with being presented with graphs and information to read through. The tool itself is also not the most aesthetically pleasing and contained a lag of a few seconds depending on how busy the server was and how many participants were interacting with the tool at the same time. Hence all these factors together show why the two explanations are not seen as easy to use. On top of this for the narrative condition, there are a total of seven pages, which a participant could have found tedious to go through and would hence be challenging to use frequently. Lastly, for the non-narrative explanation, a similar argument as before can be applied as to why the explanation scores the worst for usability, as the components of this explanation are optimized for the narrative explanation.

In conclusion, the findings show that in terms of trust and understandability narrative visualizations are effective in explaining recommendations made by a recommender system, however, the usability of such an explanation is a point of improvement.

7.2 Factors Influencing the Effectiveness of Explanations

With the second objective we are interested to know which factors contributed to the effectiveness of the different explanations, looking at both trust/understandability and usability.

Firstly, as hypothesized in H8, higher usability leads to higher trust and understandability. This is surprising given the finding that the narrative condition has the highest trust/understandability but a low usability score. However, the effect size of usability on trust/understandability is the second smallest effect, where the dominant effect is perceived accuracy. Therefore, while this H8 is supported, the model shows that it is not the main the predictor of trust/understandability.

Perceived accuracy is a good predictor for the target variable (H17). In fact, for this target variable perceived accuracy is the best predictor. This relationship in existing literature was inconclusive as Yin et al. (2019) found perceived accuracy to affect trust, whereas Cramer et al. (2008) did not. The current thesis acts as evidence towards the results of Yin et al. (2019). While not part of the hypotheses for the study, perceived accuracy is also a good predictor of usability. Hence, it seems that if users think the recommendations are accurate, then they are likely to trust the system and find it more usable as they agree with the information shown to them.

The hypothesis (H12) is rejected since the effect size of the interaction (in number of points hovered) is so small it can be considered as zero and the relationship with time spent using the tool in inconclusive. One would expect that interacting with the tool leads to getting to know it better, extracting more information from it and hence understanding it better. An interesting finding was observed; time is not a significant predictor of trust/understandability, but it is of usability. This effect is negative, spending more time in the tool leads to lower usability. This

can be explained by the fact that participants spent most time in the non-narrative and narrative conditions, which have low usabilities.

Algorithm aversion was not shown to be a predictor of trust/understandability (H14). It was expected that if someone trusts algorithms in general, they should be more likely to trust the system. However, the found results coincide with the previous literature, where contradictory results are presented. Some work shows that higher algorithm aversion leads users to be less tolerable of mistakes made by a recommender [9], and hence trust it less, while other research shows that a recommender system with an explanation can help overcome algorithm bias [10]. The results of the current thesis suggest that this relationship is not clear. Additionally, the unclarity in this relationship could stem from the combined variable of trust and understanding, where literature discusses the effect of algorithm aversion on trust, not understandability.

Additionally for usability, it was expected that visualization familiarity would be a predictor of this target variable since the tool would be easier to understand and use if the user is proficient in visualizations. However, this hypothesis (H13) is not supported by the results of the study. However, time is found to be a good predictor of usability. The results show that the more users used the tool, the less usable they found it. This relationship can be explained by the fact that participants mostly spent more time using the tool in the non-narrative and narrative conditions, which have low usability scores.

Lastly, we expected a relationship between objective and perceived accuracy, however, no relationship was found (H16). Likely this is due to the fact that other factors have an effect on perceived usability, relationships which were not investigated in this thesis.

In conclusion, we find that usability and perceived accuracy have a positive effect on trust and understandability of the recommender system. The evidence for algorithm aversion suggests that there is no relation between algorithm aversion and trust/understandability. The effect of interacting with the tool are surprising and unexplainable within the current study.

7.3 Limitations and Future Work

There are a number of limitations with the current study. Firstly, as it is conducted online there is no control over the environment of the participants which could have an effect on their participation and responses. It is known that participants experienced some technical difficulties due to APIs not always functioning as expected. Hence, these outside factors could have affected the trust/understandability and usability scores for the tool. On top of this, as discussed above, the tool is not very aesthetically pleasing and had some significant lags at times which could have effects on trust and usability not measured by the current study.

A second limitation is that the non-narrative condition is designed to test what is the effect of not having narrative features, hence there was no focus on making it informative in an easy way. It was decided it would be identical to the narrative explanation but without the story-building text and distribution over multiple pages. This made the non-narrative condition provide a lot of information to the user at once. This has likely had effects on the effectiveness of the explanation. An idea for a future study is to create a non-narrative explanation optimized with suitable visualizations and rerun the study to see if the narrative is still better, keeping the narrative explanation the same.

A further improvement of the current study is in the questionnaires used. The study intended to generate 7 factors from the post-test questionnaire, however, only 2 were found during the factor analysis. An improvement of the items in the questionnaire is needed. With the improved items the study can be run again, to obtain the full scope of factors and investigate the originally hypothesized effects in detail.

Furthermore, the recommendation system used is a simple one, hence a future study can extend on this research by investigating how effective narrative visualizations are in explaining different recommender systems.

The study was also conducted in a technical university, hence it is expected that these findings do not translate very well to the general public. Additionally, for the number of conditions in

the study, a larger number of participants would be needed to draw stronger more generalizable conclusions. However, the research itself does show a potential for narrative visualizations in explaining recommender systems as even in this small-scale study we see quite positive results. Hence, for future work, we are interested to see more studies look into this type of visualization in order to investigate the effectiveness of it on a larger scale.

Bibliography

[1] Nicolas Garcia Belmonte. *Extracting and Visualizing Insights from Real-Time Conversations Around Public Presentations.* 2014. 6, 7

[2] Or Biran and Courtenay Cotton. Explanation and Justification in Machine Learning: A Survey. 1

[3] Or Biran N-Join and Kathleen Mckeown. Human-Centric Justification of Machine Learning Predictions. 2017. 1

[4] Dirk Bollen, Bart P Knijnenburg, and Martijn C. Willemsen. Understanding Choice Overload in Recommender Systems. 2010. 1

[5] Svetlin Bostandjiev, John O'donovan, and Tobias Höllerer. TasteWeights: A Visual Interactive Hybrid Recommender System. 2012. 4

[6] Diogo N. Cosenza, Lauri Korhonen, Matti Maltamo, Petteri Packalen, Jacob L. Strunk, Erik Næsset, Terje Gobakken, Paula Soares, and Margarida Tomé. Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. *Forestry*, 94(2):311–323, 2021. 13

[7] Raphael Couronné, Philipp Probst, and Anne Laure Boulesteix. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):1–14, 2018. 13

[8] Henriette Cramer, Vanessa Evers, Satyan Ramlal, · Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, Bob Wielinga, H Cramer, V Evers, · S Ramlal, · M Van Someren, S Ramlal, M Van Someren, L Rutledge, L Aroyo, N Stash, · L Aroyo, and B Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. 18:455–496, 2008. 3, 7, 10

[9] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. 2014. 9, 51

[10] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *International Journal of Human Computer Studies*, 58(6):697–718, 2003. 9, 51

[11] Sara Irina Fabrikant and Daniel R Montello. The effect of instructions on distance and similarity judgements in information spatializations. *International Journal of Geographical Information Science*, 22(4):463–478, 2008. 5

[12] Corey K Fallon, Anne K G Murphy, Laura Zimmerman, and Shane T Mueller. *The Calibration of Trust in an Automated System: A Sensemaking Process.* 2010. 3

[13] Eduardo Ghidini, Caroline Q Santos T Φ, Isabel Manssour, and Milene S Silveira. Analyzing Design Strategies for Narrative Visualization. 2017. 1, 6

[14] Chen He, Denis Parra, and Katrien Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 9 2016. 4

[15] Rassadarie Kanjanabose, Alfie Abdul-Rahman, and Min Chen. A Multi-task Comparative Study on Scatter Plots and Parallel Coordinates Plots. *Computer Graphics Forum*, 34(3):261–270, 6 2015. 13

[16] Amirsam Khataei and Diana Lau. Recommender Narrative Visualization. Technical report, 2013. 1, 4

[17] René F Kizilcec. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. 2016. 9

[18] Bart P Knijnenburg and Martijn C Willemsen. Evaluating Recommender Systems with User Experiments. In *Recommender Systems Handbook,*, pages 309–352. 2015. 1, 10, 11

[19] Pigi UC Kouki Santa Cruz, James Schaffer, Jay Pujara, Lise Getoor, and Pigi Kouki. Personalized Explanations for Hybrid Recommender Systems. page 12, 2019. 8

[20] Josua Krause, Adam Perer, and Kenney Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. *ACM*, pages 5686–5697, 2016. 4

[21] Jayachithra Kumar and Nava Tintarev. Using Visualizations to Encourage Blind-Spot Exploration. *IntRS Workshop*, 2018. 4

[22] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. A 3D Item Space Visualization for Presenting and Manipulating User Preferences in Collaborative Filtering. 2017. 5, 6

[23] Johannes Kunkel, Claudia Schwenger, and Jürgen Ziegler. NewsViz: Depicting and Controlling Preference Profiles Using Interactive Treemaps in News Recommender Systems. 10(20):2020. 4

[24] John D. Lee and Katrina A. See. Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. 3

[25] Yu Liang and Martijn Willemsen. Interactive Music Genre Exploration with Visualization and Mood Control. *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 175–185, 4 2021. 4, 5

[26] Mehrbakhsh Nilashi, Dietmar Jannach, Othman bin Ibrahim, Mohammad Dalvi Esfahani, and Hossein Ahmadi. Recommendation quality, transparency, and website quality for trust-building in recommendation agents. *Electronic Commerce Research and Applications*, 19:70–84, 9 2016. 9

[27] John O'donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. *PeerChooser: Visual Interactive Recommendation.* 2008. 4, 5

[28] Vlad L. Pop and Alex Shrewsbury. Individual Differences in the Calibration of Trust in Automation. *HUMAN FACTORS*, 57(4):545–556, 2014. 3

[29] Donghao Ren, Matthew Brehmer, Bongshin Lee, and Eun Kyoung Choe. *ChartAccent: Annotation for Data-Driven Storytelling.* 6

[30] Christian Richthammer and Günther Pernul. Explorative analysis of recommendations through interactive visualization. *Lecture Notes in Business Information Processing*, 278:46–57, 2017. 4

[31] Yuri Saito and Takayuki Itoh. MusiCube: A Visual Music Recommendation System featuring Interactive Evolutionary Computing. 2011. 4

[32] Edward Segel and Jeffrey Heer. Narrative Visualization: Telling Stories with Data. Technical report, 2010. 1

[33] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel Keim, and Mennatallah El-Assady. Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models. *Proc. of IEEE VIS Workshop on Visualization for AI Explainability (VISxAI)*, (July), 2018. 4, 20

[34] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum. Comput. Stud.*, 146, 2 2021. 1

[35] Rashmi Sinha and Kirsten Swearingen. The Role of Transparency in Recommender Systems. Technical report, 2002. 8

[36] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. MoviExplain: A Recommender System with Explanations. 2009. 4

[37] N. Tintarev and J. Masthoff. Explaining recommendations: Design and evaluation. *Recommender Systems Handbook, Second Edition*, pages 353–382, 1 2015. 3

[38] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the Effect of Accuracy on Trust in Machine Learn-ing Models. 2019. 10

[39] Jingjing Zhang and Shawn P Curley. Exploring Explanation Effects on Consumers' Trust in Online Recommender Agents. *International Journal of Human-Computer Interaction*, 34(5):421–432, 2018. 4, 49

Appendix A

Informed Consent Form

Information form for participants

This study is performed by Yana Onushkina, a student under the supervision of Martijn Willemsen, Rianne Conijn and Stef van den Elzen of the Human-Technology Interaction and Visualization groups at Eindhoven University of Technology.

Before participating, you should understand the procedure followed in this study, and give your informed consent for voluntary participation. Please read this page carefully.

About this study

This study has the goal to investigate how the results of a recommender system can be explained. You will interact with a tool that was built for this purpose. As part of the tool, we will ask you to log into your Spotify account and based on your top 60 songs will give you some song recommendations. Then we will help you with understanding these recommendations. As part of the study, you will be asked to fill in 2 questionnaires before and after using the tool.

This study will take 30 minutes to complete. The study makes use of your Spotify data, retrieving 60 songs and your account name. The songs are stored under an id number which will be randomly assigned to you, and hence cannot be connected to you outside of this study. The account name is encrypted before being stored. Through this it is not possible to know what your account name is. Additionally, your email address will be stored until you have successfully received the compensation. After this it will be removed.

Voluntary Participation

Your participation is completely voluntary. You can stop participation at any time, however, keep in mind that this will not be compensated. In order to obtain the monetary compensation, the study has to be fully completed. You can also withdraw your permission to use your data up to 4 days after completing this study. You will be paid 5 euro if you complete this study.

Confidentiality and use, storage, and sharing of data

This study has been approved by the Ethical Review Board of Eindhoven University of Technology. In this study demographic data (age, gender, level of education), personal data (email address) and experimental data (your responses to the questionnaires and logs of your interaction with the tool) will be stored. To protect your privacy, all data that can be used to personally identify you is stored on a GDPR complient server (PythonAnywhere) in Europe. Additionally, when retrieved this data will be stored on an encrypted server of the Human Technology Interaction group for at least 10 years. The anonymized dataset that, to the best of our knowledge and ability will not contain information that can identify you, will be made publicly available. This anonymized

dataset will also be placed in a public repository where other researchers can access it. From this data it will not be possible to derive your identity or any information about you.

Further information

If you want more information about this study, the study design, or the results, you can contact Yana Onushkina (contact email: y.g.onushkina@student.tue.nl). You can report irregularities related to scientific integrity to confidential advisors of the TU/e, whose contact information can be found on www.tue.nl.

Certificate of consent

By starting this study,
- I indicate that I have read and understood the study procedure, and I agree to voluntarily participate.
- I know that my participation is completely voluntary. I know that I can refuse to participate and that I can stop my participation at any time during the study, without giving any reasons. I know that I can withdraw permission to use my data up to 24 hours after the data have been recorded.
- I agree to voluntarily participate in this study carried out by the research group Human Technology Interaction of the Eindhoven University of Technology.
- I know that no information that can be used to personally identify me or my responses in this study will be shared with anyone outside of the research team.

- I agree
- I disagree

I also give permission to make my anonymized recorded data available to others in a public online data repository.

- I agree
- I disagree

Appendix B

Factor Loadings

Code	Item	Factor 1	Factor 2	Factor 3	Uniqueness
trust_1	I am convinced that the recommended items are suitable for me			0.9321	0.1921
trust_2	I am confident that I like the items recommended to me			0.9617	0.0999
trust_3	The recommender can be trusted			0.7941	0.2132
understand_3	The recommendation process is clear to me		0.7449		0.4464
easeOfUse_1	The explanation tool was easy to understand	0.7056			0.3448
easeOfUse_2	The explanation tool was easy to use	0.8129			0.3114
reception_2	This explanation makes the recommendation process clear to me		0.8752		0.1545
reception_4	This explanation for the recommendations is convincing		0.7483		0.2699
cognitiveLoad_2	Sometimes I felt lost while using the tool	-0.7193			0.4261
cognitiveLoad_6	The visualizations were easy to read and use	0.8475			0.3005
cognitiveLoad_7	It was difficult to read the information from the visualizations	-0.7538			0.4661

Table B.1: Factor Loading for the items of the Post-Test Questionnaire - Dropped Version