APPLICATION OF DATA ANALYTICS FOR PREDICTION OF SUICIDE RATES AT THE STATE AND NATIONAL LEVELS

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Industrial Engineering

by

Derek Benson

December 2018

COMMITTEE MEMBERSHIP

TITLE:	Application of Data Analytics for Prediction of Suicide Rates at the State and National Levels
AUTHOR:	Derek Ronald Benson
DATE SUBMITTED:	December 2018
COMMITTEE CHAIR:	Reza Pouraghabagher, Ph.D. Professor of Industrial Engineering
COMMITTEE MEMBER:	Roy Jafari, Ph.D. Professor of Industrial Engineering
COMMITTEE MEMBER:	Michael Whitt, Ph.D. Professor of Biomedical Engineering

ABSTRACT

Application of Data Analytics for Prediction of Suicide Rates at the State and National Levels

Derek Ronald Benson

The increasing suicide rate in the United States has amplified the need to assure that regions with high suicide risk receive adequate funding programs and related resources for prevention methods. The way in which organizations, dedicated to preventing suicides, distribute funding could be improved with the development of predictive models for suicide rate. In this study, a multiple linear regression model at a national level was developed to identify relevant factors associated with suicide. The national level model was developed in two phases; the first using response variable data and explanatory variable data from the same time period, and the second with the response variable data shifted one time period to create a more accurate model for prediction. The models had k-fold R-squared values of 0.676 and 0.675. The national model identified four variables to include in a predictive state level model: Foreclosure Rates, Violent Crime Rates, Gini ratio, and Consumption Volume. In the second part of this study, the use of Twitter data in a state level model was evaluated. Tweets terms relating to suicide were identified in fifteen states over a thirty-one-day period and used to calculate three variables: Tweet rate, Favorite rate, and Retweet rate. Each of these three variables for the terms “suicide” and “suicidal” underwent an Analysis of Variance test (ANOVA) to check for differences between states. Each ANOVA test resulted in a p-value less than 0.0001 providing strong evidence that there was a difference in Tweet rate, Favorite rate, and Retweet rate for the two search phrases analyzed among the states. Next, a Pearson Product-Moment correlation coefficient and Pearson Rho correlation coefficient were evaluated for each Twitter variable and the states’ historical suicide rates. All computed correlation coefficients were between -0.15 and 0.3 suggesting that there is, at best, a weak correlation between the Twitter variables and a state’s historical suicide rate. The results from the Twitter data analysis suggest that it is too early to accurately incorporate such data into a state level multiple linear regression model. The results of this study would help in further development of a state level model that allows organizations, dedicated to reducing suicides, allocate related resources more efficiently.

Keywords: Regression Modeling, Prediction, Suicide Rate, Twitter, Correlation

TABLE OF CONTENTS

Page

LIST OF TABLES…. vi
LIST OF FIGURES ix
CHAPTER
1. INTRODUCTION 1
1.1 Problem Description 1
1.2 Purpose of Study 2
2. LITERATURE REVIEW 3
2.1 Predictive Suicide Modeling Introduction 3
2.2 Traditional Suicide Factors 3
2.3 Use of Social Media Data in Suicide Modeling 5
2.4 Literature Review Conclusions 7
3. NATIONAL MULTIPLE LINEAR REGRESSION MODELING 9
3.1 Regression Modeling Methodology 9
3.2 Calculating Suicide Rate 10
3.3 Preparing Predictor Variables 11
3.4 Multiple Regression Model Creation 13
3.5 Regression Model Results 14
4. SOCIAL MEDIA ANALYSIS 20
4.1 Social Media Analysis Methodology 20
4.2 Social Media Results for English Phrase 1 (EP1) 23
4.2.1 Social Media Results for EP1 Tweet Rate 26
4.2.2 Social Media Results for EP1 Favorite Rate 31
4.2.3 Social Media Results for EP1 Retweet Rate 35
4.3 Social Media Results for English Phrase 2 (EP2) 39
4.3.1 Social Media Results for EP2 Tweet Rate 39
4.3.2 Social Media Results for EP2 Favorite Rate 47
4.3.3 Social Media Results for EP2 Retweet Rate 51
4.4 Comparison of Twitter Data to Historical State Suicide Rates 55
5. CONCLUSIONS 62
5.1 National Multiple Regression Model Conclusions 62
5.2 Social Media Conclusions 62
5.2.1 Differentiating Rates Between States 63
5.2.2 Relating Rates of Tweets to Historical Suicide Trends 63
5.3 Conclusion 64
5.4 Ethical Considerations 64
5.5 Study Limitations 65
5.6 Future Work 66
BIBLIOGRAPHY 67
APPENDICES
A. RSTUDIO CODE FOR NATIONAL MODEL 70
B. DATA TABLE EXCERPT FOR REGRESSION MODEL 76
C. RSTUDIO CODE FOR TWITTER DATA GATHERING 77
D. SAS EXAMPLE CODE FOR CALCULATING SUICIDE RATE 81
E. M CODE FOR PROCESSING TWITTER DATA 82
F. EXAMPLE EXCEL PIVOT TABLE FOR TWITTER DATA 83

LIST OF TABLES

Table Page

1. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model including all predictor variables and Quarter numbers 14
2. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters 15
3. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters and Property Crime Rates 15
4. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters, Property Crime Rates, and Graduation Percentage 16
5. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters, Property Crime Rates, Graduation Percentage, and Adjusted Personal Income 16
6. Significant variables found in the resulting model from the stepwise regression and their respective p-values for the first national multiple regression model 17
7. Significant variables found in the resulting model from the stepwise regression and their respective p-values for the second national multiple regression model 18
8. One-Way ANOVA test output for the transformed Tweet Rate for EP1 26
9. Means for the One-Way ANOVA for the transformed Tweet Rate for EP1 27
10. Connecting Letters Report for each state for the transformed Tweet Rate for EP1 27
11. One-Way ANOVA output after the exclusion of outliers for Tweet Rate for EP1 29
12. Means for the One-Way ANOVA for the transformed Tweet Rate for EP1 after the removal of selected outliers 30
13. Connecting Letters Report for each state for the transformed Tweet Rate for EP1 after the removal of selected outliers 30
14. One-Way ANOVA test output for the transformed Favorite Rate for EP1 31
15. Means for the One-Way ANOVA for the transformed Favorite Rate for EP1 31
16. Connecting Letters Report for each state for the transformed Favorite Rate for EP1 31
17. One-Way ANOVA output after the exclusion of outliers for Favorite Rate for EP1 33

18. Means for the One-Way ANOVA for the transformed Favorite Rate for EP1 after the removal of selected outliers 34
19. Connecting Letters Report for each state for the transformed Favorite Rate for EP1 after the removal of selected outliers 34
20. One-Way ANOVA test output for the transformed Retweet Rate for EP1 35
21. Means for the One-Way ANOVA for the transformed Retweet Rate for EP1 35
22. Connecting Letters Report for each state for the transformed Retweet Rate for EP1 36
23. One-Way ANOVA output after the exclusion of outliers for Retweet Rate for EP1 38
24. Means for the One-Way ANOVA for the transformed Retweet Rate for EP1 after the removal of selected outliers 38
25. Connecting Letters Report for each state for the transformed Retweet Rate for EP1 after the removal of selected outliers 38
26. One-Way ANOVA test output for the transformed Tweet Rate for EP2 42
27. Means for the One-Way ANOVA for the transformed Tweet Rate for EP2 43
28. Connecting Letters Report for each state for the transformed Tweet Rate for EP2 43
29. One-Way ANOVA output after the exclusion of outliers for Tweet Rate for EP2 45
30. Means for the One-Way ANOVA for the transformed Tweet Rate for EP2 after the removal of selected outliers 46
31. Connecting Letters Report for each state for the transformed Tweet Rate for EP2 after the removal of selected outliers 46
32. One-Way ANOVA test output for the transformed Favorite Rate for EP2 47
33. Means for the One-Way ANOVA for the transformed Favorite Rate for EP2 47
34. Connecting Letters Report for each state for the transformed Favorite Rate for EP2 47
35. One-Way ANOVA output after the exclusion of outliers for Favorite Rate for EP2 49
36. Means for the One-Way ANOVA for the transformed Favorite Rate for EP2 after the removal of selected outliers 50
37. Connecting Letters Report for each state for the transformed Favorite Rate for EP2 after the removal of selected outliers 50

38. One-Way ANOVA test output for the transformed Retweet Rate for EP2 51
39. Means for the One-Way ANOVA for the transformed Retweet Rate for EP2 51
40. Connecting Letters Report for each state for the transformed Retweet Rate for EP2 51
41. One-Way ANOVA output after the exclusion of outliers for Retweet Rate for EP2 53
42. Means for the One-Way ANOVA for the transformed Retweet Rate for EP2 after the removal of selected outliers 54
43. Connecting Letters Report for each state for the transformed Retweet Rate for EP2 after the removal of selected outliers 54
44. Summary Table of historical suicide trends for states in study 55
45. The Pearson Product-Moment Correlation Coefficient and Pearson Rho Correlation Coefficient for EP1's Tweet Rate, Favorite Rate, and Retweet Rate and a state's average age-adjusted suicide rate…. 58
46. The Pearson Product-Moment Correlation Coefficient and Pearson Rho Correlation Coefficient for EP2's Tweet Rate, Favorite Rate, and Retweet Rate and a state's average age-adjusted suicide rate…. 61
47. Excerpt of the Table created in RStudio used for the National Level Multiple Linear Regression Models 76
48. Example Excel Pivot Table for Alabama's Twitter Data for EP1 83

LIST OF FIGURES

Figure Page

1. Illustration of the data set transformation for the second multiple regression model created 13
2. Plot of Residuals versus Fitted, Residuals versus Leverage, Scale-Location, and Normal Q-Q for the first national multiple regression model 17
3. Plot of Residuals versus Fitted, Residuals versus Leverage, Scale-Location, and Normal Q-Q for the second national multiple regression model 19
4. Plot of the Tweet Rate data for EP1 for each state including a Box Plot (seen as the small red lines on graph)..... 23
5. Plot of the Favorite Rate data for EP1 for each state including a Box Plot (seen as the small red lines on graph)..... 24
6. Plot of the Retweet Rate data for EP1 for each state including a Box Plot (seen as the small red lines on graph)..... 24
7. Plot of the Tweet Rate data for EP1 after a natural log transformation including a Box Plot (see as the small red lines on graph) 25
8. Plot of the Favorite Rate data for EP1 after a natural log transformation including a Box Plot (seen as the small red lines on graph) 25
9. Plot of the Retweet Rate data for EP1 after a natural log transformation including a Box Plot (seen as the small red lines on graph) 26
10. One-Way ANOVA test plot for the transformed Tweet Rate data for EP1..... 28
11. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Tweet Rate data for EP1 28
12. One-Way ANOVA test plot for the transformed Tweet Rate data for EP1 after the removal of selected outliers 29
13. One-Way ANOVA test plot for the transformed Favorite Rate data for EP1..... 32
14. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Favorite Rate data for EP1 32

15. One-Way ANOVA test plot for the transformed Favorite Rate data for EP1 after the removal of selected outliers 33
16. One-Way ANOVA test plot for the transformed Retweet Rate data for EP1 36
17. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Retweet Rate data for EP1 37
18. One-Way ANOVA test plot for the transformed Retweet Rate data for EP1 after the removal of selected outliers 37
19. Plot of the Tweet Rate data for EP2 for each state including a Box Plot (seen as the small red lines on graph) 39
20. Plot of the Favorite Rate data for EP2 for each state including a Box Plot (seen as the small red lines on graph) 39
21. Plot of the Retweet Rate data for EP2 for each state including a Box Plot (seen as the small red lines on graph) 40
22. Plot of the Tweet Rate data for EP2 after a natural log transformation including a Box Plot (see as the small red lines on graph) 41
23. Plot of the Favorite Rate data for EP2 after a natural log transformation including a Box Plot (seen as the small red lines on graph) 41
24. Plot of the Retweet Rate data for EP2 after a natural log transformation including a Box Plot (seen as the small red lines on graph) 42
25. One-Way ANOVA test plot for the transformed Tweet Rate data for EP2 44
26. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Tweet Rate data for EP2 44
27. One-Way ANOVA test plot for the transformed Tweet Rate data for EP2 after the removal of selected outliers 45
28. One-Way ANOVA test plot for the transformed Favorite Rate data for EP2 48
29. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Favorite Rate data for EP2 48

30. One-Way ANOVA test plot for the transformed Favorite Rate data for EP2 after the removal of selected outliers 49
31. One-Way ANOVA test plot for the transformed Retweet Rate data for EP2 52
32. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Retweet Rate data for EP2 52
33. One-Way ANOVA test plot for the transformed Retweet Rate data for EP2 after the removal of selected outliers 53
34. Scatterplot Matrix of states' average suicide rate and EP1 Tweet Rate 56
35. Scatterplot Matrix of states' average suicide rate and EP1 Favorite Rate 57
36. Scatterplot Matrix of states' average suicide rate and EP1 Retweet Rate 57
37. Scatterplot Matrix of states' average suicide rate and EP2 Tweet Rate 59
38. Scatterplot Matrix of states' average suicide rate and EP2 Favorite Rate 60
39. Scatterplot Matrix of states' average suicide rate and EP2 Retweet Rate 60

Chapter 1

INTRODUCTION

1.1 Problem Description

In a recent study published by the Centers for Disease Control and Prevention, it was found that suicide rates in the United States have steadily increased from 1999 to 2014 (Hedegaard, Warner, & Curtin, 2016). This disturbing trend has garnered attention from the federal government, which in 2001 published a National Strategy for Suicide Prevention and an updated version in 2012. One area of focus in these strategies is improving the federal government's ability to collect and report data relating to suicides. This data can be found in the web-based Injury Statistics Query and Reporting System and in the Center for Disease Control and Prevention's Mortality Multiple Cause files.

One of the most recognizable government organizations dedicated to preventing suicide is the Substance Abuse and Mental Health Services Administration (SAMHSA). SAMHSA is federally funded and provides services such as the National Suicide Prevention Lifeline (Substance Abuse and Mental Health Services Administration, 2013). In addition, SAMHSA provides grants to fund research and programs dedicated to reducing suicides (Substance Abuse and Mental Health Services Administration, 2013). SAMHSA also pilots' programs including the Zero Suicide Model across the United States (Substance Abuse and Mental Health Services Administration, 2013). Beyond government organizations that provide funding and information with the goal of reducing suicides, there is a litany of private organizations that raise and distribute funds to curb suicides. Many of these organizations are a part of the National Action Alliance for Suicide Prevention, which is the primary group advancing the National Strategy for Suicide Prevention (Action Alliance, n.d.). One of the largest health organizations dedicated to preventing suicides is the American Foundation for Suicide Prevention (AFSP) (American Foundation for Suicide Prevention, 2017). AFSP funds various activities including research, school programs, and mental health programs in the hope of reducing the number of suicides (American Foundation for Suicide Prevention, 2017).

One problem that organizations from SAMHSA to the National Action Alliance for Suicide Prevention to AFSP is that they must decide what programs to fund. Accurate models that can

predict future suicide rates would help these organizations distribute funds in the most efficient manner.

1.2 Purpose of Study

Suicide risk predictive models have been an area of interest in academia for several decades. Many suicide risk predictive models developed in literature rely on predictor variables that can only be found in an individual's medical record or through an interview with the patient or their family (Phillips et al., 2002; McCarthy et al., 2015). The difficulty in obtaining the predictor variable data in these models limits their application. The federal government's collection of publicly accessible suicide data, that spans the last fifty years, offers the opportunity to develop a suicide risk predictive model based on easily accessible data, such as an area's unemployment rate, income rate, and education rate.

The purpose of this study will be to first identify significant factors relating to suicide rates in a national multiple regression model. By identifying national level significant factors, the study will have demonstrated the robustness of said variables helping justify their inclusion in a potential state level regression model for predicting suicide rates. The second purpose of this study is to evaluate the usage of social media data in a state level regression model for predicting suicide rates. The future development of a predictive state level regression model for suicide rate would help both government and private organizations distribute funds more efficiently. Examples of this could include SAMSHA choosing to pilot a new program in a state where suicide rates are likely to increase in the future or ASFP choosing to feature a story about an individual from a state to raise suicide awareness in the area.

Chapter 2

LITERATURE REVIEW

2.1 Predictive Suicide Modeling Introduction

The link between suicide rate and economic and environmental factors has been analyzed in academic research over the last century. Historically the data analyzed in this research is gathered through interviews with the deceased's family or through some form of national registry with more recent studies utilizing data from social media websites and other internet sources. The previous body of research has revealed several explanatory variables that appear to be correlated with suicide. This literature review will be broken up into two sections. The first section will cover articles that identify traditional economic and environmental factors that are associated with suicide. The second section will focus on more recent work in the field that incorporates social media or other internet related data. At the end of the two sections conclusions from the literature review will be drawn.

2.2 Traditional Suicide Factors

A good starting point for investigating factors associated with suicide is a two-part literature review published by Stack in 2000. This publication was a follow-up to a previous published literature review by Stack in the early 1980's and outlines recent developments in the sociology of suicide. The paper summarizes various studies findings and can be used to get a rough summary of factors associated with suicide. The factors can be broken up roughly into two categories: cultural and economic. Economic factors included unemployment, underemployment, family income, income inequality, cost of healthcare, and female participation in the labor force (S. Stack, 2000). Cultural factors included, but were not limited to, gender, alcohol consumption, religion, marital status, age, crime rates, holiday effects, depression, fertility rates, and urbanization (S. Stack, 2000).

One common theory applied in sociology when discussing suicide is general strain theory. General strain theory is a framework that can be used to classify factors associated with suicide and was developed to explain "analogous behavior" (Steven Stack & Wasserman, 2007). The

sources of strain according to the theory include loss, blocked goals, and exposure to negative stimuli (Steven Stack & Wasserman, 2007). Strains identified in papers include poor work relationships, loss of a loved one, and recent acute stress, but the intimate nature of these factors make interviews necessary to gather accurate information invalidating them for this paper's model (Steven Stack & Wasserman, 2007; Phillips et al., 2002). However, useful strains have been identified in literature such as the loss of a home or vehicle and issues with the justice system (Steven Stack & Wasserman, 2007). These strains have the potential to be interpreted through data on foreclosure and crime rates in the United States.

Additional studies have identified factors that increase risk of suicide. The factors identified in these studies include income level, employment status, educational achievement, and income inequality (Li, Page, Martin, & Taylor, 2011; Fountoulakis et al., 2015). A study conducted in Denmark identified several risk factors for suicide. The factors included unemployment, low income, and family medical history (Agerbo, Sc, Mortensen, & Sc, 2003). The relationship between suicide risk and mental health has been noted in many academic papers (Li et al., 2011; Phillips et al., 2002). Mental health issues are likely related to suicide rates, but due to the complex nature of this factor it will not be included in this paper's model. However, factors such as employment status, income level, and educational achievement can be incorporated into this paper's model.

The relationship between suicide and unemployment has garnered an extensive amount of attention in literature. A study by Classen and Dunn (2012) suggested that the length of time a person was unemployed was responsible for the relationship with suicide, not the loss of the job itself (Soares, 2009). While a study conducted by Andres that accounted for "country specific linear time trends" found suicide rates to not be related to unemployment or income levels but did find suicide rates to be related to alcohol consumption, fertility rates, and economic growth (Rodrı, 2005, p. 1). To further complicate the matter, Yong-Hwan Noh conducted a study in which the interaction between unemployment and income was analyzed (Noh, 2009). In this paper it was found that unemployment was positively associated with suicide rate in wealthier countries (Noh, 2009). The conclusion suggested by this outcome is that unemployment alone is not associated with suicide rates rather a loss in economic opportunity is associated with suicide rate (Noh, 2009). In recently

published papers by Fountoulakis, Coppola, and others a relationship between suicide and unemployment was observed (Fountoulakis et al., 2015; Coppola et al., 2016). The literature suggests that unemployment should be included in this paper's model; however, attention should be given to what other factors may interact with it.

2.3 Use of Social Media Data in Suicide Modeling

The popularity of using data derived from the internet in suicide models has increased as the platform has become more available. A study conducted by Biddle et al. found that just under half of the top website results when searching typical phrases researched by individual's contemplating suicide contained information about suicide methods (Biddle, Donovan, Hawton, Kapur, & Gunnell, 2008). In another study a combination of different search engines and key search words and phrases were used to find web pages relating to suicide (Recupero, Harms, & Noble, 2008). The study found that out of the 373 unique web pages found 41 contained pro-suicide information (Recupero, Harms, & Noble, 2008). These studies were used in a paper published by Luxton and others to justify the assertion that information on suicide and suicide methods is relatively easy to find on the Internet (Luxton, June, & Fairall, 2012).

Researchers have utilized data gathered from web-blogs, forums, and social media sites to develop predictive models for national suicide rates, risk of suicide for individuals, and even the chances a user may shift from posting about mental illness to practicing suicide ideation. An example of this can be seen in a study conducted by Choudhury and others. In the study, data was gathered from various "subreddits", which are subforums found on the popular internet forum Reddit. In the study the researchers were able to develop a logistic regression classifier using linguistic indicators to predict which users who post about mental illness would later post about suicide ideation (De Choudhury, Kiciman, Dredze, Coppersmith, & Kumar, 2016).

The value of utilizing data gathered from web-blogs in predictive models for suicide rates was demonstrated in two papers published on suicide rates in South Korea. Won and others carried out a univariate linear regression analysis to predict future suicide rates in South Korea (Won et al., 2013). The developed model utilized two social media variables, suicide-related and dysphoria-

related weblog entries, as well as variables containing information on the country's economy and climate (Won et al., 2013). It was found that both social media variables were significant in the model and the model could be used to accurately predict suicide occurrences (Won et al., 2013). In a follow up study a multivariate model was developed using similar variables to the original study with one of the key differences being an increase in the quantity of data used when developing the model (Lee et al., 2018). In this study suicide numbers were split into a seasonal and non-seasonal component and only the non-seasonal component was used in the model (Lee et al., 2018). The model produced in this paper had an accuracy of 82.9% with accuracy being defined "as the ratio of correct predictions to total predictions" (Lee et al., 2018, p. 347). These two studies both controlled for the celebrity effect on suicide and could be used to justify the use of social media data in a national level model.

One social media site that has garnered a particularly large amount of research regarding this topic is Twitter. Twitter, a popular social media web application, provides an Application Program Interface that allows access to all public "tweets" (messages posted by users). This data, being relatively easy to access, has been the subject of linguistic analysis and predictive model development. In a study conducted on data from the United States researchers found an association between the rate of "tweets" being classified as "at risk" and suicide rates in that geographical location (Jashinsky et al., 2014). The study converted a list of suicide risk factors, e.g. depressive feelings, into key phrases and words to look for in "tweets" (Jashinsky et al., 2014).

One issue that arose in this study was the need to remove tweets that contain the key words as well as words that negate the "at risk" tag, e.g. a tweet that contains the words "cut" and "myself" and "shaving", needs to be filtered out (Jashinsky et al., 2014). The researchers accomplished this by compiling a list of words that if found in a "tweet" would remove it from the database (Jashinsky et al., 2014). Another issue in the study was that many "tweets" gathered did not contain location information barring them from being included in the analysis (Jashinsky et al., 2014). Nonetheless, this study demonstrates that Twitter data may be a reliable way to assess suicide risk in certain geographical locations.

Another study that utilized Twitter data in suicide modeling was conducted using data gathered in Japan. In this study a logistic regression model was developed using traditional indicator variables such as education level and family income as well as data gathered from Twitter (Sueki, 2015). Researchers found and recorded “tweets” that contained the Japanese words that translate to the phrases “want to die” and “want to commit suicide” and incorporated the data into the logistic regression model (Sueki, 2015). These two social media variables were found to be significantly related to suicidal behavior and ideation (Sueki, 2015).

An important characteristic in studies that utilize data from web-applications such as Twitter is the ability to differentiate posts that are of concern and those that are not. One study that addressed this issue utilized machine learning to differentiate levels of concern relating to “tweets” (O’Dea et al., 2015). The study concluded that individuals do express suicidality on Twitter and both human coders and automated processes can determine the level of concern a “tweet” warrants (O’Dea et al., 2015). In another study researchers developed two logistic regression models that used linguistic predictor variables (O’Dea, Larsen, Batterham, Calear, & Christensen, 2017). One model was used to differentiate between “tweets” that are strongly concerning and those that are general (O’Dea et al., 2017). The other model developed was used to differentiate between “tweets” that are strongly concerning and those that are safe to ignore (O’Dea et al., 2017). These two studies establish that Twitter data can be reliably and electronically processed and used in predictive model development.

2.4 Literature Review Conclusions

One of the primary issues in the studies regarding traditional factors associated with suicide rates is multicollinearity among risk factors skewing results. For example, a citizen with a lower income level may be more likely to commit suicide, but they may also be more likely to commit a crime or may be less likely to reach certain educational achievements. In this case it is hard to determine how specific factors relate to the risk of suicide. The study conducted in this paper will attempt to address this issue by including a relatively large number of risk factors associated with suicide. These factors will include foreclosure rate, non-violent crime rate, unemployment rate,

income level, educational achievement, alcohol consumption rates, fertility rates, economic growth, and income inequality. The hope is that by including a larger number of explanatory variables in the multiple regression model some of the multicollinearity issues can be teased out and a better understanding of how certain explanatory variables relate to suicide can be achieved.

An important step in the development of suicide predictive models based on Social Media data will be determining the geographic size limits where significant changes in variables can still be observed. In this study Twitter data will be gathered and tested to see if a significant difference can be recognized between states, a necessary characteristic if it is to be included in a predictive state level multiple regression suicide model.

Chapter 3

NATION MULTPLE LINEAR REGRESSION MODELING

3.1 Regression Modeling Methodology

Linear regression modeling is the process of fitting a response variable to one explanatory variable. While multiple linear regression modeling is the process of fitting a response variable to multiple explanatory variables. It is said that the explanatory variable is significant when its use in a regression model can account for a significant portion of a response variable's variability. It is important to note that a linear regression model cannot establish a cause and effect relationship. The only way to establish a cause and effect relationship between a response and explanatory variable is through a controlled experiment. Multiple linear regression models are typically used for one of two purposes to either estimate a response variable based on known inputs or to predict a future response variable value based on current inputs.

The quality of a regression model can be interpreted from its R-squared value. The R-squared value is a measure of how much variability a respective model can account for regarding its response variable. An issue that arises in multiple linear regression modeling is that the addition of explanatory variables will always increase a model's R-squared value. The increase in R-squared however is not always "real" due to overfitting. Two metrics can be used to assess if a multiple linear regression model is overfitted, R-squared adjusted and Akaike information criterion (AIC), both measures include a penalty for each added explanatory variable.

Another issue that arises in multiple linear regression models is that explanatory variables can be correlated to one another. When explanatory variables are correlated to one another they overinflate their significance in a multiple linear regression model, this problem is known as multicollinearity. A measure of an explanatory variable's multicollinearity is its variable inflation factor (VIF). Ideally an explanatory variable's VIF will be 1; however, as issues with multicollinearity increase VIF increases. A general rule of thumb is that an explanatory variable with a VIF greater than 10 should be removed from a multiple linear regression model.

3.2 Calculating Suicide Rate

The first step in developing the national multiple linear regression model was calculating the crude suicide rates per quarter in the United States. The Mortality Multiple Cause files for the years 1991 to 2016 were downloaded and read into SAS 9.4. The data files were then processed. Files between the years of 1999 and 2016 contained a column in the data file corresponding to “Manner of Death” with the option of “2”, indicating suicide. Once the files containing the “Manner of Death” were read into SAS, the observations that were not coded as a “2” were filtered out. In addition, all columns not corresponding to Manner of Death and month of death were filtered out.

The Mortality Multiple Cause files between the years 1991 to 1998 do not contain a Manner of Death column. For these files all observations not containing 950, 951, 952, 953, 954, 955, 956, 957, 958, 959 in the first three columns of Underlying cause were filtered out. The list of numbers corresponds to the codes referencing suicide in the International Classification of Diseases 9th edition (ICD-9) (Public Health Surveillance and Environmental Health Branch Public Health Division, Alberta Health & Wellness, July 2006). It is important to note that these codes are normally led by an “E”, e.g. “E950”; however, the “E” is dropped in the Mortality Multiple Cause files. In addition, all columns not corresponding to the first three columns of the Underlying condition or the month of the death were removed.

Next, the data was sorted by month and the number of observations were counted for each month. Once the monthly counts were found, the months of January, February, and March were summed to represent Quarter 1 (Q1) while the months April, May, and June were summed to represent Quarter 2 (Q2). The months July, August, and September were summed to represent Quarter 3 (Q3) and the last three months (October, November, and December) were summed to represent Quarter 4 (Q4). After finding the total number of suicides per quarter the totals were divided by that year’s population and multiplied by 100,000 to find a crude suicide rate per 100,000 inhabitants. The United States population was taken from July 1st of the respective year and was found on the United States census website (Population Estimates Program Population Division, U.S. Census Bureau, Internet Release Date: April 11, 2000; Intercensal Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2000 to July 1, 2010;

and U.S. and World Population Clock). Once the quarterly suicided rates were calculated they were inserted into a separate excel file and loaded into RStudio.

3.3 Preparing Predictor Variables

Once the crude suicide rates were loaded into RStudio, predictor variable data was added into a data table along with its respective Quarter number. The first predictor variable added to the data table was quarterly foreclosure rates (Economic Research Federal Reserve Bank of St. Louis (Source: Board of Governors of the Federal Reserve System (US)). Specifically, this dataset tracked “Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks”.

Next, both violent crime and property crime rates were added to the data table. This data was found using the Uniform Crime Reporting Statistics tool provided by the Federal Bureau of Investigation. The tool could not be used to find statistics on the years 2015 and 2016, but the data is provided on year-specific pages on the Federal Bureau of Investigation’s website. The data from the Uniform Crime Reporting Statistics tool and the data found in the table “Crime in the United States, by Volume and Rate per 100,000 Inhabitants, 1997-2016” included the rate of Violent and Property Crime per 100,000 inhabitants. The rates per 100,000 were added to the data table. Since the data provided by these sources was on an annual basis the yearly rates were divided by four and assigned to the four quarters for each year. The assumption that crime rates were constant throughout the year was based on a paper published by the U.S. department of Justice (Lauritsen & White, 2014). The paper noted seasonal trends in property and violent crime rates but stated that the difference in seasonal high and low rates were less than 11% for household property crimes and less than 12% for violent crimes (Lauritsen & White, 2014).

The fourth predictor variable added was unemployment rate. The data was found on the Bureau of Labor Statistics. The data found was in monthly intervals. To translate this data into quarterly points the average unemployment rate was found for each quarter. The next variable added to the data table was total personal income per quarter in billions and was found on the Bureau of Economic Analysis website. The total personal income per quarter was adjusted to account for

inflation by multiplying the quarterly totals by the average Consumer Price Index for All Urban Consumers for the months covering the specific quarter. The Consumer Price Index for All Urban Consumers was found on the Economic Research Federal Reserve Bank of St. Louis's website.

The fifth predictor variable added to the data table was the percent change in Gross Domestic Product (GDP) by quarter. The GDP data was found on the Bureau of Economic Analysis website. The next predictor variable added was the Gini Ratio. This number was used to represent income inequality and was found on the United States Census Bureau website. Since the data on Gini Ratio found was recorded annually the ratio was repeated for each quarter within a year. The seventh predictor variable added was the percentage of people 25 years of age or over who had completed either High School or College. Again, this information was found on the United States Census Bureau website. Since the data was recorded annually the ratio was repeated for each quarter within the year.

Next, a measure of fertility was added to the data table. Annual birth data files were download from The National Bureau of Economic Research website and read into SAS 9.4. The number of observations per month were then counted. After counting the number of observations per month the counts were summed into quarters. Once the counts were summed into quarters the sums were divided by the national population of the corresponding year (same population used to calculate crude suicide rates). To make the resulting number easier to interpret they were multiplied by 10,000. The resulting number would be interpreted as the number of observations per quarter per 10,000 people.

The final predictor variable added to the data table was United States alcohol consumption. The annual historical data on the volume of alcohol consumed by the average United States citizen was divided by four and replicated four times then assigned to its corresponding quarter in the data table. The assumption that there is no seasonality in alcohol consumption is somewhat dubious. In a recent study conducted researchers found that participants were more likely to have had a drink within 30 days in the months of January and July when compared to other months (Cho, Johnson, & Fendrich, 2001).

3.4 Multiple Regression Model Creation

Next a multiple regression model was created using the data table created in section 3.3. The lm function in RStudio was used with the explanatory variables being Quarters, Foreclosure Rates, Violent Crime Rates, Property Crime Rates, Unemployment Rates, Adjusted Personal Income, Graduation Percentage, Fertility Measure, and Consumption volume. The method used was iteratively reweighted least squares for formulating the model. Then the model assumptions were tested, e.g. normally distributed residuals. In addition, the variable inflation factors (VIF) were calculated and variables were removed with VIFs greater than 10. The variable with the greatest VIF was removed and the model was run again until all variables in the model had VIFs that were less than 10.

The factors remaining included: Foreclosure Rates, Violent Crime Rates, Unemployment Rates, GDP Change, Gini Ratio, Fertility Measure, and Consumption Volume. Next two multiple regression models were created using the glm function. The first model created used the calculated suicide rate for a given quarter along with data from the same quarter for the other factors. The model produced using this data set would allow a researcher to estimate a given quarters suicide rate based on data gathered during the same quarter. While this model would help identify factors associated with suicide it would not help predict future suicide rates. The second model was based on a data set with suicide rate shifted one quarter forwards creating a model tailored to predicting future suicide rate. An illustration of the data set transformation can be seen in Figure 1.

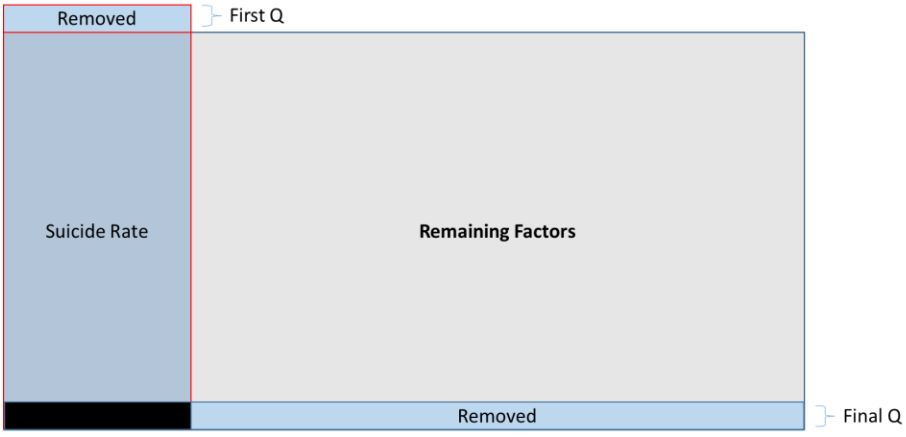

Figure 1. Illustration of the data set transformation for the second multiple regression model created.

Both national level multiple regression models were created and validated using the same steps. After creating the multiple regression models using the glm function the stepAIC function was used with the direction option set to “both”. The stepAIC function uses stepwise regression to minimize a model’s AIC. The stepAIC function identified what factors should be included in a final multiple regression model. The models were validated by finding their respective k-fold R-squared values. The train and traincontrol functions in the caret package were used to perform a 10-fold cross validation of the model with 100 repetitions.

3.5 Regression Model Results

The multiple regression model created using all the predictor variable data and Quarter number (e.g. Q1) has an adjusted R-square of 0.83. The residuals distribution was tested using an Anderson-Darling normality test. The residuals were found to be normally distributed with an Anderson-Darling p-value of 0.316. The calculated VIFs can be seen in Table 1.

Table 1. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model including all predictor variables and Quarter numbers.

Variable	VIF
Quarters	920.46**
Foreclosure Rates	19.61
Violent Crime Rates	95.84
Property Crime Rates	166.40
Unemployment Rates	15.07
Adjusted Personal Income	595.76
GDP Change	1.25
Gini Ratio	9.27
Graduation Percentage	184.01
Fertility Measure	1.75
Consumption Volume	10.24

** Indicates Variable to be removed

The multiple regression model excluding the Quarters variable has an adjusted R-square of 0.79. The residuals distribution was tested using an Anderson-Darling normality test. The

residuals were found to not be normally distributed with an Anderson-Darling p-value of 0.0012. The calculated VIFs can be seen in Table 2.

Table 2. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters.

Variable	VIF
Foreclosure Rates	12.95
Violent Crime Rates	58.11
Property Crime Rates	166.37**
Unemployment Rates	9.01
Adjusted Personal Income	107.39
GDP Change	1.24
Gini Ratio	8.95
Graduation Percentage	161.82
Fertility Measure	1.71
Consumption Volume	9.37

** Indicates Variable to be removed

The multiple linear regression model excluding the Quarters and Property Crime Rates variables has an adjusted R-square of 0.79. The residuals distribution was tested using an Anderson-Darling normality test. The residuals were found to not be normally distributed with an Anderson-Darling p-value of 0.0042. The calculated VIFs can be seen in Table 3.

Table 3. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters and Property Crime Rates.

Variable	VIF
Foreclosure Rates	12.91
Violent Crime Rates	34.33
Unemployment Rates	8.87
Adjusted Personal Income	107.07
GDP Change	1.23
Gini Ratio	8.12
Graduation Percentage	121.79**
Fertility Measure	1.69
Consumption Volume	9.08

** Indicates Variable to be removed

The multiple regression model excluding the Quarters, Property Crime Rates, and Graduation Percentage variables has an adjusted R-square of 0.78. The residuals distribution was tested using an Anderson-Darling normality test. The residuals were found to be normally distributed with an Anderson-Darling p-value of 0.14. The calculated VIFs can be seen in Table 4.

Table 4. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters, Property Crime Rates, and Graduation Percentage.

Variable	VIF
Foreclosure Rates	11.99
Violent Crime Rates	11.17
Unemployment Rates	8.21
Adjusted Personal Income	38.81**
GDP Change	1.17
Gini Ratio	8.00
Fertility Measure	1.68
Consumption Volume	5.87

** Indicates Variable to be removed

The multiple regression model excluding the Quarters, Property Crime Rates, Graduation, and Adjusted Personal Income variables has an adjusted R-square of 0.66. The residuals distribution was tested using an Anderson-Darling normality test. The residuals were found to be normally distributed with an Anderson-Darling p-value of 0.76. The calculated VIFs can be seen in Table 5.

Table 5. Calculated Variable Inflation Factors (VIFs) for the multiple linear regression model excluding Quarters, Property Crime Rates, Graduation Percentage, and Adjusted Personal Income.

Variable	VIF
Foreclosure Rates	8.42
Violent Crime Rates	4.29
Unemployment Rates	6.49
GDP Change	1.16
Gini Ratio	6.50
Fertility Measure	1.36
Consumption Volume	2.39

The first stepwise regression model ran with the variables Foreclosure Rates, Violent Crime Rates, Unemployment Rates, GDP Change, Gini Ratio, Fertility Measure, and Consumption Volume identified five variables. The identified variables were Foreclosure Rates, Violent Crime Rates, GDP Change, Gini Ratio, and Consumption Volume. The identified variables with their corresponding p-values can be seen in Table 6.

Table 6. Significant variables found in the resulting model from the stepwise regression and their respective p-values for the first national multiple regression model.

Variable	P-Value
Foreclosure Rates	2.22*10-8
Violent Crime Rates	8.02*10-9
GDP Change	0.0086
Gini Ratio	3.65*10-9
Consumption Volume	0.0014

The multiple regression model generated using these five variables has an adjusted R-squared value of 0.656. The residuals are normally distributed with an Anderson-Darling p-value of 0.774. Plots of the residuals versus fitted, a normal Q-Q plot of residuals, Scale-Location, and residuals versus leverage can be seen in Figure 2.

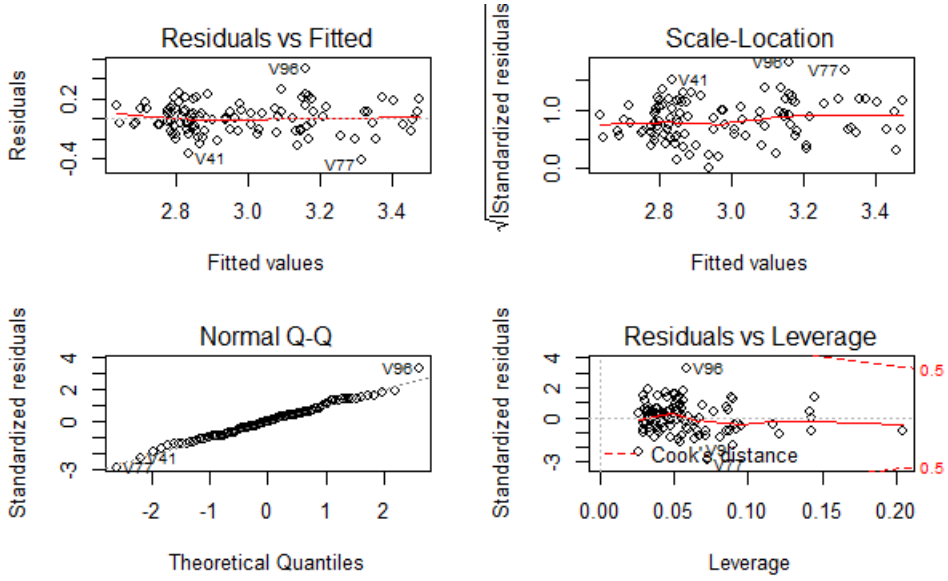

Figure 2. Plot of Residuals versus Fitted, Residuals versus Leverage, Scale-Location, and Normal Q-Q for the first national multiple regression model.

The plots in Figure 2 demonstrate that the assumptions for a multiple regression model are satisfied. The normal Q-Q plot seen in the bottom left closely follows a line demonstrating that the residuals are normally distributed. While the Residuals vs Fitted plot seen in the top left corner shows no signs of fanning suggesting that the residuals are randomly distributed. A 10-fold cross-validation was ran in RStudio using the *trainControl* and *train* functions with 100 repetitions. The resulting R-squared value found was 0.673.

The second stepwise regression model ran with the shifted data set also including the variables Foreclosure Rates, Violent Crime Rates, Unemployment Rates, GDP Change, Gini Ratio, Fertility Measure, and Consumption volume identified six variables. The identified variables were Foreclosure Rates, Violent Crime Rates, Unemployment Rates, Gini Ratio, Fertility Measure, and Consumption Volume. The identified variables with their corresponding p-values can be seen in Table 7.

Table 7. Significant variables found in the resulting model from the stepwise regression and their respective p-values for the second national multiple regression model.

Variable	P-Value
Foreclosure Rates	0.000143
Violent Crime Rates	8.01*10-8
Unemployment Rates	0.16413
Gini Ratio	0.000137
Fertility Measure	0.076925
Consumption	0.000786

The multiple regression model generated using these six variables has an adjusted R-squared value of 0.655. The residuals are normally distributed with an Anderson-Darling p-value of 0.428. Plots of the residuals versus fitted, a normal Q-Q plot of residuals, Scale-Location, and residuals versus leverage can be seen in Figure 3.

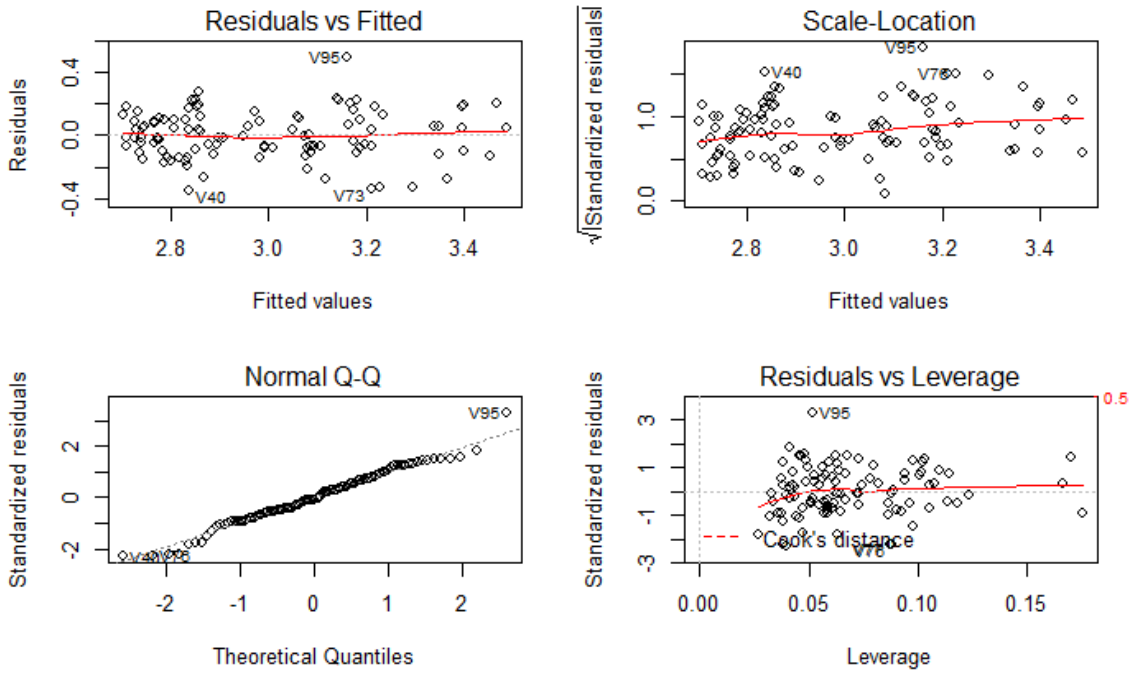

Figure 3. Plot of Residuals versus Fitted, Residuals versus Leverage, Scale-Location, and Normal Q-Q for the second national multiple regression model.

The plots in Figure 3 demonstrate that the assumptions for a multiple regression model are satisfied. The normal Q-Q plot seen in the bottom left closely follows a line demonstrating that the residuals are normally distributed. While the Residuals vs Fitted plot seen in the top left corner shows no signs of fanning suggesting that the residuals are randomly distributed. A 10-fold cross-validation was ran in RStudio using the *trainControl* and *train* functions with 100 repetitions. The resulting R-squared value found was 0.675.

Chapter 4

SOCIAL MEDIA ANALYSIS

4.1 Social Media Analysis Methodology

In the second portion of this paper, the validity of using social media data within a predictive suicide model at a state level is analyzed. There are two goals for the section of this paper. The first is to test if there is a significant difference in the rate of suicide related “tweets” from state to state. The second, assuming a significant difference in the rate of “tweets” from state to state is observed, is to see if a higher rate of “tweets” occurs in states with higher historical suicide rates.

Fifteen states were selected for the study and data was gathered from August 12, 2018, until September 11, 2018. The fifteen states selected were Alabama, Arizona, Florida, Georgia, Maine, Massachusetts, Minnesota, Montana, Nevada, New Hampshire, North Dakota, Oregon, South Carolina, South Dakota, and Vermont. States selected spanned most of the continental United States and contained sets of States near one another. By selecting nearing states, such as North and South Dakota, one can limit the effect of external factors that influence suicide such as sunlight hours and temperature. In addition, by selecting states from various regions of the United States a degree of robustness can be asserted in the results.

Five search phrases/words were decided upon and included “suicide”, “suicidal”, “Prozac”, “feel depressed”, and “feel hopeless”. The “tweets” were identified and gathered using RStudio and the rtweet package. The exact code used to collect “tweets” can be seen in Appendix C. The settings in RStudio were set to collect “tweets” in English, Spanish, French, traditional Chinese, and non-traditional Chinese. The code, ran once for each day spanning the testing period, identified and gathered “tweets” then placed them into a new Excel workbook. A workbook for each state and date combination was created. Within each workbook there are 25 worksheets, each one corresponding to a search phrase and language combination.

Once all the data for the study was gathered the files were split into 15 folders. Each folder contained the workbooks relating to a specific state. Next the data was appended in a new Excel Workbook using Power Query. A query that appended all of a given state’s data relating to a specific search phrase and language was created. The query also selected the columns of interest for the

study. The Twitter data gathered contained 88 columns ranging from information on a user's account to data on the number of times a "tweet" was "retweeted" or favorited. Four columns were selected for use in the analysis: the time the "tweet" was created, the text of the "tweet", the favorite count, and the "retweet" count.

After creating the query, it was saved to a worksheet in an Excel workbook. This process resulted in an Excel Workbook containing fifteen queries, one for each state that appended data pertaining to one search phrase and language combination. This process was carried out for the first two search phrases: "suicide" and "suicidal". The next step in the process was cleaning the data. The goal when cleaning the gathered data was to remove "tweets" that contained the respective search phrase but were intended to prevent suicide. After combing through several data tables, a list of words was selected. The list included the words "Prevention", "prevention", "hotline", "Intervention", and "intervention".

In Power Query the Text.Contains function was used to identify the "tweets" containing an exclusionary term. Once the terms were identified they were filtered out of the respective data set. An example of the Power Query M code can be seen in Appendix E. Next, the fifteen individual queries were appended into one data table with an additional column specifying an observation's respective state. This final data table was then loaded to an Excel Worksheet and a pivot table was created. The pivot table used the State column and date for rows and the count of observations collected, the sum of the count of favorites, and sum of the count of "retweets". An excerpt of the table can be seen in Appendix F.

After summarizing the data in a Pivot table, the population for each State in the study was found. The July 1, 2017 state population from the U.S. Census was used. The count of observations, sum of the count of favorites, and sum of the count of "retweets" for each day was divided by its respective state's population resulting in a variable representing the rate in which a citizen either "tweeted", favorited, or "retweeted" a "tweet" that satisfied the previous criteria. It is important to note that a "tweet" could be favorited or "retweeted" from a user outside of the state in which the "tweet" originated. After calculating these rates, a summary table in a separate Excel Workbook was created.

The summary table had a column specifying State, Date, State's Population, "TweetRate" (the number of "tweets" observed divided by the state's population multiplied by a 1,000), "FavRate" (the sum of the favorite count divided by the state's population multiplied by 1,000), "RetweetRate" (the sum of the favorite count divided by the state's population), and lastly a state's average age-adjusted suicide rate from the years 2016, 2015, and 2014. The average age-adjusted suicide rates for these years was found on the CDC's website (Suicide Mortality by State: 2016, Suicide Mortality by State: 2015, and Suicide Mortality by State: 2014). In instances were no observations were found for a specific state date combination a zero was input for all three measures.

Once the summary table was created, it was inserted into JMP Pro 13 and an Analysis of Variance (ANOVA) test was conducted for each of the calculated variables: "TweetRate", "FavRate", and "RetweetRate". The factor for each ANOVA test was the State name. The purpose of the ANOVA test was to check if there is a significant difference in variables between states. After conducting an initial analysis on the summary table for English phrase 1 it became apparent that the data contained multiple outliers (commonly referred to as points of influence). These outliers were often caused by an individual "tweet" being "retweeted" many times within an individual state. The most striking example of this occurred in Nevada of August 22nd where 28,533 instances of the same "tweet" was observed. The "tweet" "People getting creative w suicide now https://t.com/MAsQxWyZCW" was "retweeted" 49,615 times and accounted for the significant jump in all metrics on that respective day.

It was decided that the removal of such a "tweet" would be improper since it did not violate any of the conditions set in data collection and filtering. However, the existence of outliers in an ANOVA test can skew results thus a data transformation was selected. A 1/x and natural log transformation were tested to reduce the effect of outliers and ultimately the natural log transformation was chosen. The non-transformed distribution of each calculated variable was plotted in JMP. Then, the calculated variables were transformed and plotted again in JMP. Next, a single factor ANOVA test was conducted as well as a Tukey's test. The Tukey's test generated a Connecting Letters Report that uses the Tukey-Kramer HSD method to determine if variables are significantly different. Significance in the Tukey's test was set to a p-value of 0.05 and variables

that are not significantly different are assigned the same letter in the Connecting Letters Report. Finally, the remaining outliers, still present after the transformation, were removed to see if there was significant change in the ANOVA model or Tukey's test.

It was decided to only use the observations found in the English searches for the first two variables due to the low number of observations made in other languages. There were 137,299 English observations for the search word "suicide", but only 1,143 observations in French, 307 observations in Spanish, and 24 observations in each Chinese search. It was also decided that due to the complexity of text filtering for the other three search phrases/words ("Prozac", "feel depressed", and "feel hopeless") only the data gathered on the words "suicide" and "suicidal" would be analyzed in the results section.

4.2 Social Media Results for English Phrase 1 (EP1)

Figures 4, 5, and 6 are the non-transformed data for the three rates calculated for the first English search word "suicide".

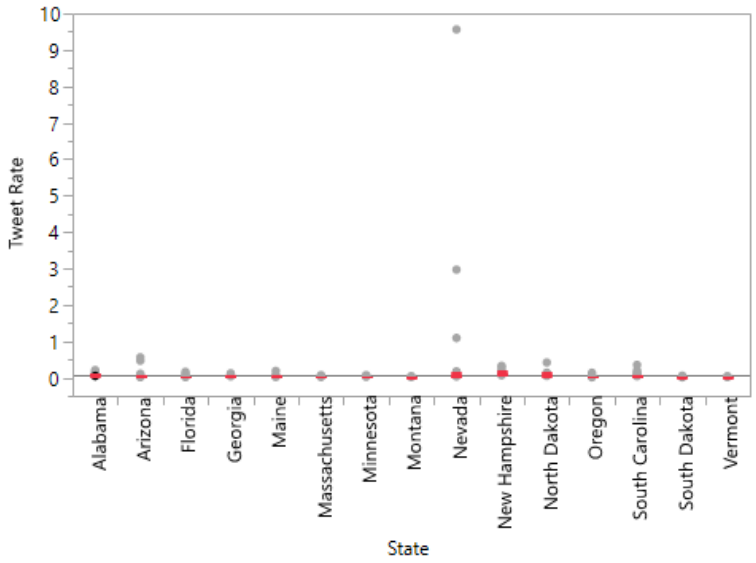

Figure 4. Plot of the Tweet Rate data for EP1 for each state including a Box Plot (seen as the small red lines on graph).

Figure 4 shows the presence of outliers in almost every state. The most egregious outliers being those for Nevada. The presence of outliers suggest that a data transformation may be appropriate.

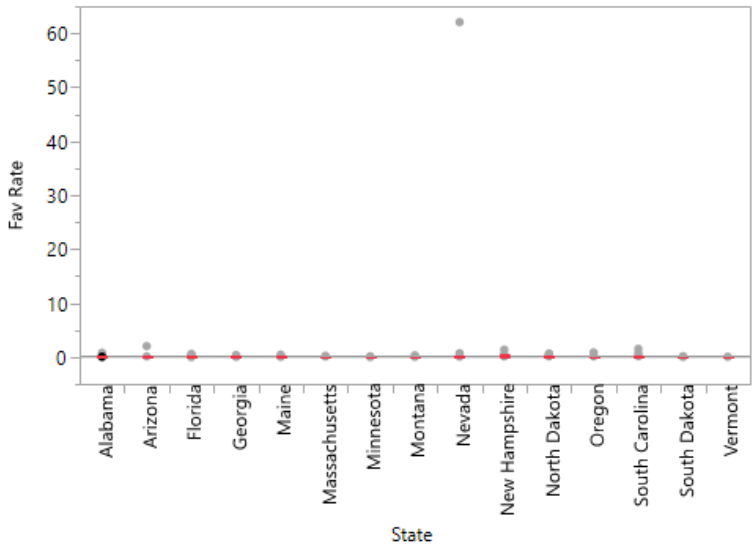

Figure 5. Plot of the Favorite Rate data for EP1 for each state including a Box Plot (seen as the small red lines on graph).

Figure 5 shows the presence of outliers in almost every state. The most egregious outliers being those for Nevada. The presence of outliers suggest that a data transformation may be appropriate.

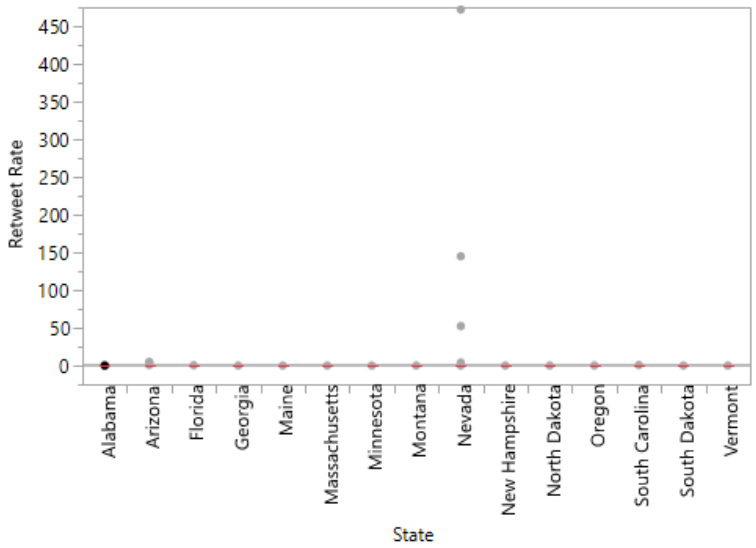

Figure 6. Plot of the Retweet Rate data for EP1 for each state including a Box Plot (seen as the small red lines on graph).

Figure 6 shows the presence of outliers in almost every state. The most egregious outliers being those for Nevada. The presence of outliers suggest that a data transformation may be appropriate.

Figures 7, 8, 9 are plots of each state's data after a natural log transformation. One Vermont data point of zero for Tweet Rate was excluded from the data set as the transformation could not be performed. Eight data points of zero were excluded after the data transformation for the Favorite Rate and twenty-one data points of zero were excluded from the Retweet Rate data set.

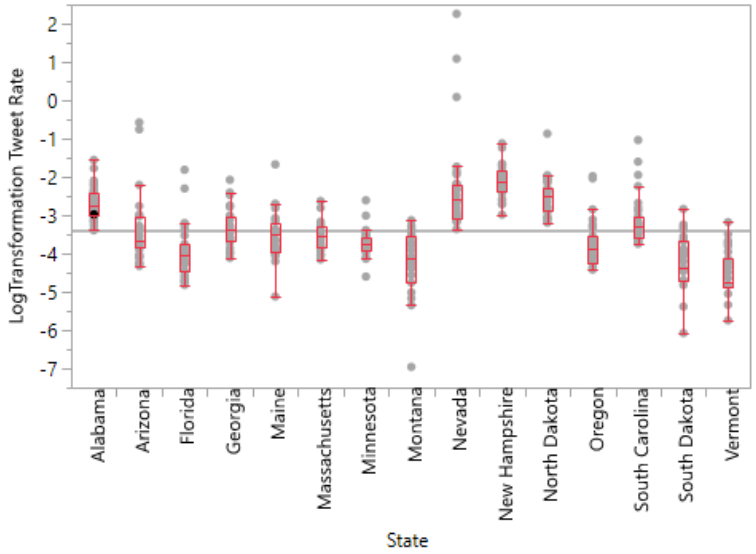

Figure 7. Plot of the Tweet Rate data for EP1 after a natural log transformation including a Box Plot (see as the small red lines on graph).

Figure 7 shows that after the data transformation outliers still remained; however, they are far closer to the data sets average than before the transformation.

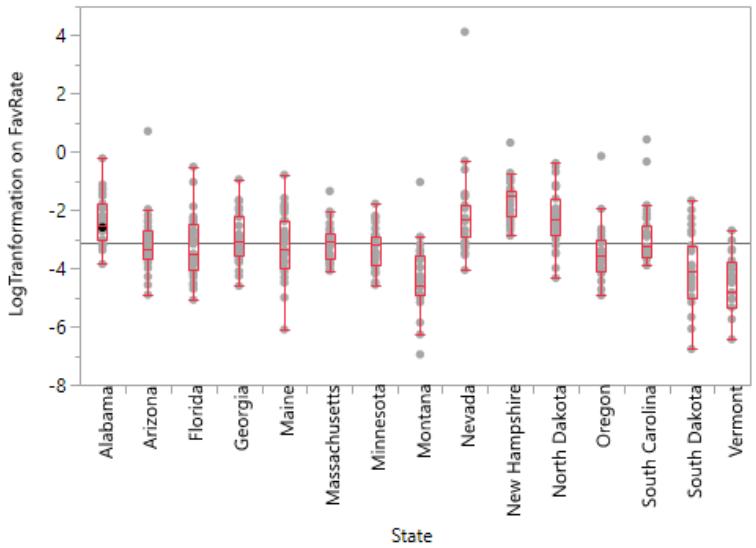

Figure 8. Plot of the Favorite Rate data for EP1 after a natural log transformation including a Box Plot (seen as the small red lines on graph).

Figure 8 shows that after the data transformation outliers still remained; however, they are far closer to the data sets average than before the transformation.

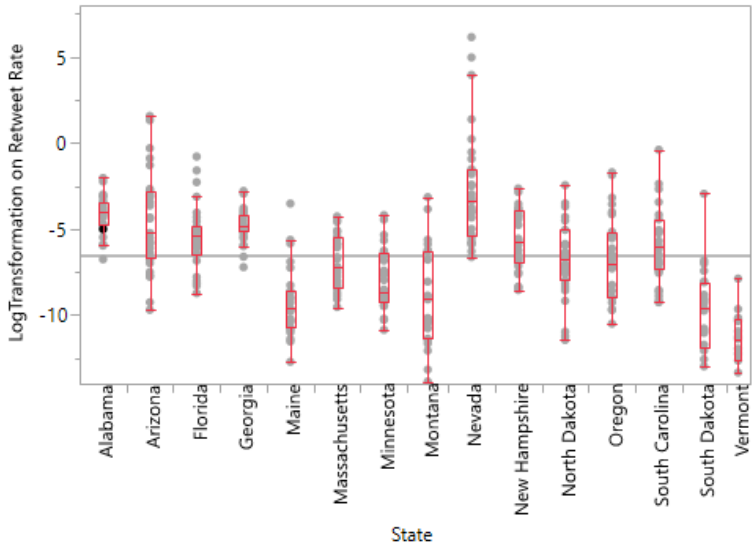

Figure 9. Plot of the Retweet Rate data for EP1 after a natural log transformation including a Box Plot (seen as the small red lines on graph).

Figure 9 shows that after the data transformation outliers still remained; however, they are far closer to the data sets average than before the transformation.

4.2.1 Social Media Results for EP1 Tweet Rate

The results of a One-Way ANOVA test on the transformed EP1 data for Tweet Rate can be seen in Tables 8 and 9. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 10.

Table 8. One-Way ANOVA test output for the transformed Tweet Rate for EP1.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	233.67798	16.6913	37.6917	<.0001*
Error	449	198.83409	0.4428		
C. Total	463	432.51207			

Table 9. Means for the One-Way ANOVA for the transformed Tweet Rate for EP1.

Level	Number	Mean	Std. Error	Lower 95%	Upper 95%
Alabama	31	-2.6898	0.11952	-2.925	-2.455
Arizona	31	-3.3660	0.11952	-3.601	-3.131
Florida	31	-3.9488	0.11952	-4.184	-3.714
Georgia	31	-3.3209	0.11952	-3.556	-3.086
Maine	31	-3.5299	0.11952	-3.765	-3.295
Massachusetts	31	-3.5540	0.11952	-3.789	-3.319
Minnesota	31	-3.7170	0.11952	-3.952	-3.482
Montana	31	-4.2119	0.11952	-4.447	-3.977
Nevada	31	-2.2738	0.11952	-2.509	-2.039
New Hampshire	31	-2.1145	0.11952	-2.349	-1.880
North Dakota	31	-2.5062	0.11952	-2.741	-2.271
Oregon	31	-3.7635	0.11952	-3.998	-3.529
South Carolina	31	-3.1368	0.11952	-3.372	-2.902
South Dakota	31	-4.2335	0.11952	-4.468	-3.999
Vermont	30	-4.5633	0.12150	-4.802	-4.324

Table 10. Connecting Letters Report for each state for the transformed Tweet Rate for EP1.

Level								**Mean**
New Hampshire	A							-2.114450
Nevada	A							-2.273809
North Dakota	A							-2.506184
Alabama	A	B						-2.689822
South Carolina		B	C					-3.136758
Georgia			C	D				-3.320923
Arizona			C	D				-3.366012
Maine			C	D	E			-3.529918
Massachusetts			C	D	E			-3.553950
Minnesota				D	E	F		-3.716977
Oregon				D	E	F		-3.763531
Florida					E	F		-3.948794
Montana						F	G	-4.211949
South Dakota						F	G	-4.233473
Vermont							G	-4.563268

Figure 10 is a plot depicting the One-Way ANOVA multiple comparison results for the transformed state data for EP1. The presence of outliers can skew results thus they were recognized, see Figure 11, and excluded from the data set. Figure 12 displays a plot for the One-Way ANOVA test run after the removal of most outliers.

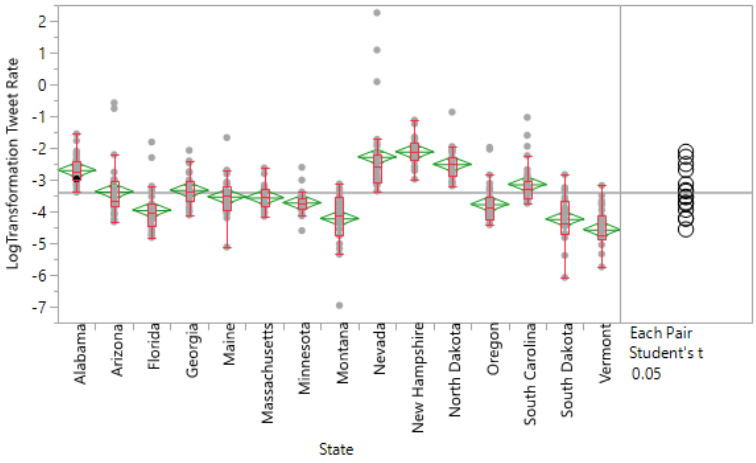

Figure 10. One-Way ANOVA test plot for the transformed Tweet Rate data for EP1.

Figure 10 suggests that there is a significant difference between states' Tweet Rates for EP1. In addition Figure 10, shows what data points are outliers.

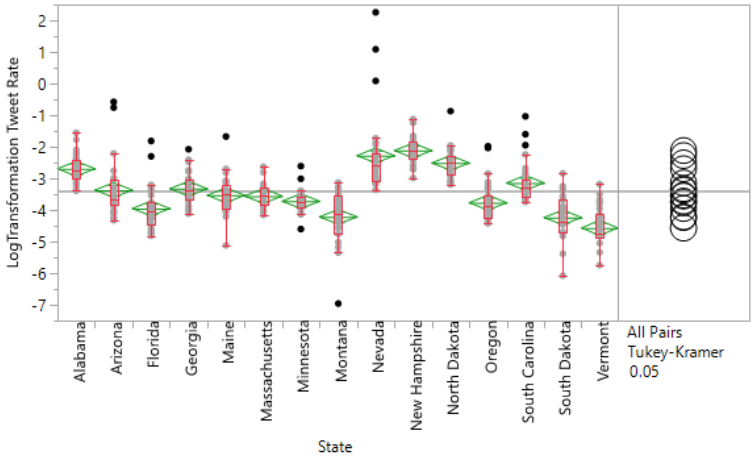

Figure 11. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Tweet Rate data for EP1.

Figure 11 shows which data points were selected to be removed before the ANOVA test was run again.

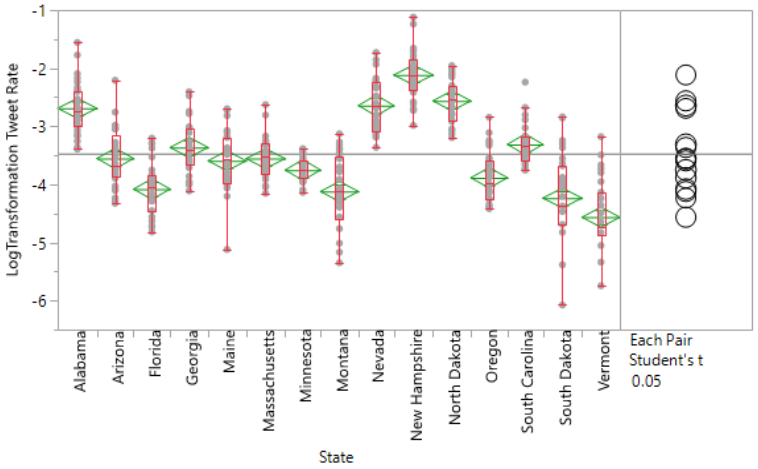

Figure 12. One-Way ANOVA test plot for the transformed Tweet Rate data for EP1 after the removal of selected outliers.

Figure 12 shows that most outliers have been removed from the data set and the ANOVA test now conducted has little chance of being skewed inappropriately.

The results of a One-Way ANOVA for Tweet rate with outliers removed can be seen in Tables 11 and 12. The outliers excluded were rows 47, 48, 80, 87, 110, 155, 204, 208, 216, 240, 258, 259, 260, 328, 344, 345, 389, 390, and 402. The test was significant at a p-value less than 0.0001. Table 13 depicts a Connecting Letters Report generated from a Tukey's test.

Table 11. One-Way ANOVA output after the exclusion of outliers for Tweet Rate for EP1.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	206.40901	14.7435	63.1609	<.0001*
Error	430	100.37394	0.2334		
C. Total	444	306.78295			

Table 12. Means for the One-Way ANOVA for the transformed Tweet Rate for EP1 after the removal of selected outliers.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-2.6898	0.08678	-2.860	-2.519
Arizona	29	-3.5522	0.08972	-3.729	-3.376
Florida	29	-4.0793	0.08972	-4.256	-3.903
Georgia	30	-3.3625	0.08821	-3.536	-3.189
Maine	30	-3.5919	0.08821	-3.765	-3.418
Massachusetts	31	-3.5540	0.08678	-3.725	-3.383
Minnesota	28	-3.7503	0.09131	-3.930	-3.571
Montana	30	-4.1204	0.08821	-4.294	-3.947
Nevada	28	-2.6401	0.09131	-2.820	-2.461
New Hampshire	31	-2.1145	0.08678	-2.285	-1.944
North Dakota	30	-2.5608	0.08821	-2.734	-2.387
Oregon	29	-3.8847	0.08972	-4.061	-3.708
South Carolina	28	-3.3093	0.09131	-3.489	-3.130
South Dakota	31	-4.2335	0.08678	-4.404	-4.063
Vermont	30	-4.5633	0.08821	-4.737	-4.390

Table 13. Connecting Letters Report for each state for the transformed Tweet Rate for EP1 after the removal of selected outliers.

Level									**Mean**
New Hampshire	A								-2.114450
North Dakota		B							-2.560778
Nevada		B							-2.640075
Alabama		B							-2.689822
South Carolina			C						-3.309257
Georgia			C	D					-3.362493
Arizona			C	D	E				-3.552221
Massachusetts			C	D	E				-3.553950
Maine			C	D	E				-3.591852
Minnesota				D	E	F			-3.750286
Oregon					E	F	G		-3.884673
Florida						F	G		-4.079320
Montana						F	G		-4.120447
South Dakota							G	H	-4.233473
Vermont								H	-4.563268

The results for the ANOVA tests before and after the removal of outliers still indicate that there is a significant difference between states' Tweet Rates for EP1. In addition, the Connecting Letters Reports indicate that not only is there a significant difference between states' Tweet Rates for EP1, but this difference can be used to break the states up into multiple groups.

4.2.2 Social Media Results for EP1 Favorite Rate

The results of a One-Way ANOVA test on the transformed EP1 data for Favorite Rate can be seen in Tables 14 and 15. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 16.

Table 14. One-Way ANOVA test output for the transformed Favorite Rate for EP1.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	285.11095	20.3651	20.4332	<.0001*
Error	442	440.52605	0.9967		
C. Total	456	725.63700			

Table 15. Means for the One-Way ANOVA for the transformed Favorite Rate for EP1.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-2.3788	0.17931	-2.731	-2.026
Arizona	31	-3.1495	0.17931	-3.502	-2.797
Florida	31	-3.2667	0.17931	-3.619	-2.914
Georgia	31	-2.9646	0.17931	-3.317	-2.612
Maine	31	-3.2341	0.17931	-3.586	-2.882
Massachusetts	31	-3.1128	0.17931	-3.465	-2.760
Minnesota	31	-3.3095	0.17931	-3.662	-2.957
Montana	30	-4.3945	0.18227	-4.753	-4.036
Nevada	31	-2.1453	0.17931	-2.498	-1.793
New Hampshire	31	-1.6364	0.17931	-1.989	-1.284
North Dakota	31	-2.2783	0.17931	-2.631	-1.926
Oregon	31	-3.4720	0.17931	-3.824	-3.120
South Carolina	31	-2.9374	0.17931	-3.290	-2.585
South Dakota	30	-4.1426	0.18227	-4.501	-3.784
Vermont	25	-4.6756	0.19967	-5.068	-4.283

Table 16. Connecting Letters Report for each state for the transformed Favorite Rate for EP1.

Level								Mean
New Hampshire	A							-1.636393
Nevada	A	B						-2.145250
North Dakota	A	B	C					-2.278271
Alabama	A	B	C	D				-2.378840
South Carolina		B	C	D	E			-2.937400
Georgia		B	C	D	E			-2.964559
Massachusetts			C	D	E			-3.112841
Arizona				D	E			-3.149452
Maine				D	E			-3.234055
Florida					E			-3.266688
Minnesota					E	F		-3.309527
Oregon					E	F		-3.472034
South Dakota						F	G	-4.142631
Montana							G	-4.394546
Vermont							G	-4.675619

Figure 13 is a plot depicting the One-Way ANOVA multiple comparison results for the transformed Favorite Rate data. The presence of outliers can skew results thus they were recognized, see Figure 14, and excluded from the data set. Figure 15 displays a plot for the One-Way ANOVA test run after the removal of most outliers.

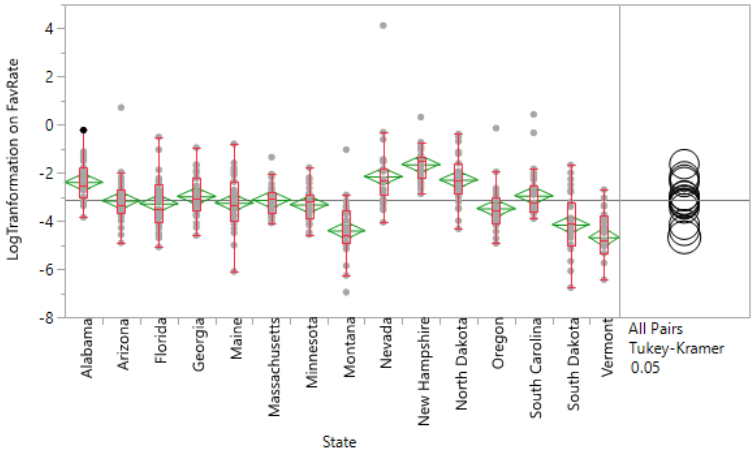

Figure 13. One-Way ANOVA test plot for the transformed Favorite Rate data for EP1.

Figure 13 shows the existence of outliers for most states in the data set. The existence of outliers can skew ANOVA test results.

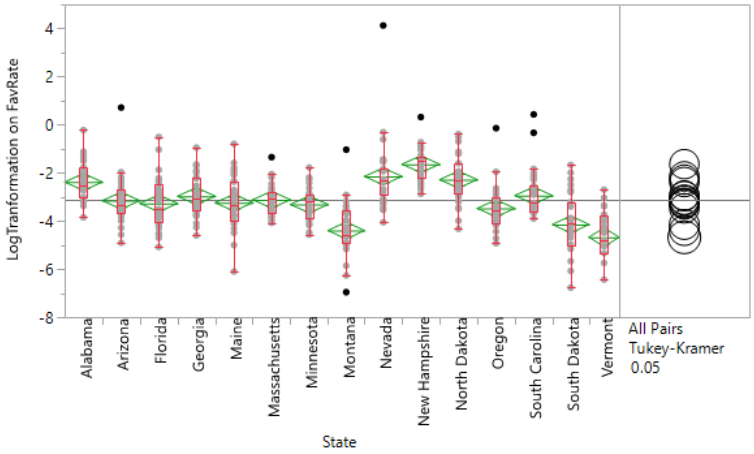

Figure 14. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Favorite Rate data for EP1.

Figure 14 shows the data points selected as outliers that are to be removed before the ANOVA test is rerun.

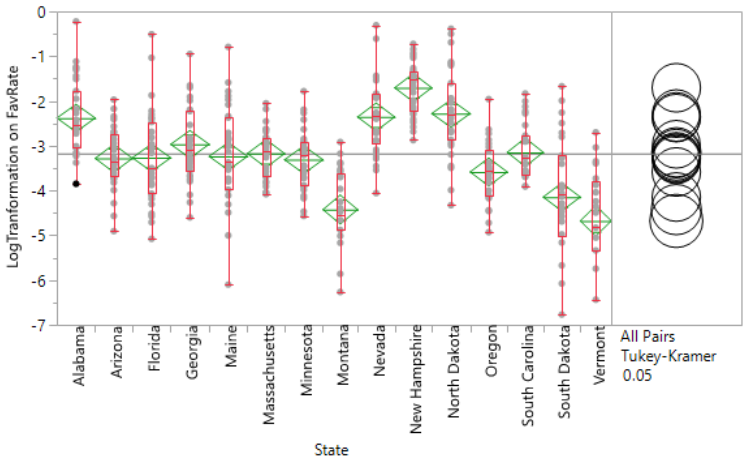

Figure 15. One-Way ANOVA test plot for the transformed Favorite Rate data for EP1 after the removal of selected outliers.

Figure 15 shows the data set after the removal of outliers. It can be seen that the data set now contains no noticeable outliers.

The results of a One-Way ANOVA for Favorite rate with outliers removed can be seen in Tables 17 and 18. The outliers excluded were rows 46, 184, 225, 245, 258, 308, 343, 389, and 402. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 19.

Table 17. One-Way ANOVA output after the exclusion of outliers for Favorite Rate for EP1.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	265.53271	18.9666	25.0017	<.0001*
Error	433	328.47912	0.7586		
C. Total	447	594.01183			

Table 18. Means for the One-Way ANOVA for the transformed Favorite Rate for EP1 after the removal of selected outliers.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-2.3788	0.15643	-2.686	-2.071
Arizona	30	-3.2783	0.15902	-3.591	-2.966
Florida	31	-3.2667	0.15643	-3.574	-2.959
Georgia	31	-2.9646	0.15643	-3.272	-2.657
Maine	31	-3.2341	0.15643	-3.542	-2.927
Massachusetts	30	-3.1718	0.15902	-3.484	-2.859
Minnesota	31	-3.3095	0.15643	-3.617	-3.002
Montana	28	-4.4231	0.16460	-4.747	-4.100
Nevada	30	-2.3544	0.15902	-2.667	-2.042
New Hampshire	30	-1.7016	0.15902	-2.014	-1.389
North Dakota	31	-2.2783	0.15643	-2.586	-1.971
Oregon	30	-3.5831	0.15902	-3.896	-3.271
South Carolina	29	-3.1435	0.16174	-3.461	-2.826
South Dakota	30	-4.1426	0.15902	-4.455	-3.830
Vermont	25	-4.6756	0.17420	-5.018	-4.333

Table 19. Connecting Letters Report for each state for the transformed Favorite Rate for EP1 after the removal of selected outliers.

Level							Mean
New Hampshire	A						-1.701638
North Dakota	A	B					-2.278271
Nevada	A	B					-2.354377
Alabama	A	B	C				-2.378840
Georgia		B	C	D			-2.964559
South Carolina			C	D			-3.143460
Massachusetts				D			-3.171782
Maine				D			-3.234055
Florida				D			-3.266688
Arizona				D			-3.278319
Minnesota				D			-3.309527
Oregon				D	E		-3.583097
South Dakota					E	F	-4.142631
Montana						F	-4.423094
Vermont						F	-4.675619

The results for the ANOVA tests before and after the removal of outliers still indicate that there is a significant difference between states' Favorite Rates for EP1. In addition, the Connecting Letters Reports indicate that not only is there a significant difference between states' Favorite Rates for EP1, but this difference can be used to break the states up into multiple groups.

4.2.3 Social Media Results for EP1 for Retweet Rate

The results of a One-Way ANOVA multiple comparison results on the transformed EP1 data for Retweet Rate can be seen in Tables 20 and 21. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 22.

Table 20. One-Way ANOVA test output for the transformed Retweet Rate for EP1.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	2074.9029	148.207	31.6694	<.0001*
Error	429	2007.6477	4.680		
C. Total	443	4082.5505			

Table 21. Means for the One-Way ANOVA for the transformed Retweet Rate for EP1.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-4.100	0.38854	-4.86	-3.34
Arizona	31	-4.552	0.38854	-5.32	-3.79
Florida	31	-5.468	0.38854	-6.23	-4.70
Georgia	31	-4.755	0.38854	-5.52	-3.99
Maine	31	-9.257	0.38854	-10.02	-8.49
Massachusetts	31	-7.011	0.38854	-7.78	-6.25
Minnesota	31	-7.912	0.38854	-8.68	-7.15
Montana	27	-8.848	0.41633	-9.67	-8.03
Nevada	31	-2.659	0.38854	-3.42	-1.90
New Hampshire	31	-5.596	0.38854	-6.36	-4.83
North Dakota	31	-6.739	0.38854	-7.50	-5.98
Oregon	31	-6.655	0.38854	-7.42	-5.89
South Carolina	31	-5.781	0.38854	-6.54	-5.02
South Dakota	29	-9.755	0.40171	-10.54	-8.97
Vermont	16	-11.428	0.54082	-12.49	-10.36

A connecting letters report generated by a Tukey's test can be seen in Table 22.

Table 22. Connecting Letters Report for each state for the transformed Retweet Rate for EP1.

Level								**Mean**
Nevada	A							-2.65875
Alabama	A	B						-4.10016
Arizona		B						-4.55204
Georgia		B						-4.75521
Florida		B	C					-5.46752
New Hampshire		B	C					-5.59610
South Carolina		B	C					-5.78120
Oregon			C	D				-6.65523
North Dakota			C	D				-6.73906
Massachusetts			C	D	E			-7.01146
Minnesota				D	E	F		-7.91195
Montana					E	F		-8.84785
Maine						F	G	-9.25684
South Dakota						F	G	-9.75490
Vermont							G	-11.42750

Figure 16 is a plot depicting the One-Way ANOVA multiple comparison results. The presence of outliers can skew results thus they were recognized, see Figure 17, and excluded from the data set. Figure 18 displays a plot for the One-Way ANOVA test run with the removal of most outliers.

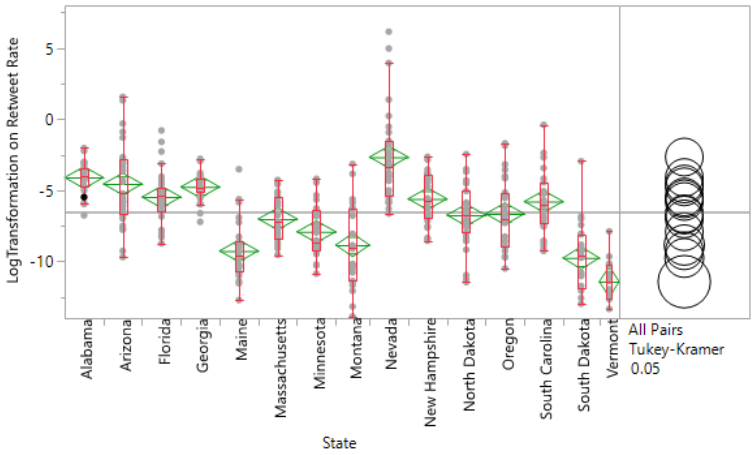

Figure 16. One-Way ANOVA test plot for the transformed Retweet Rate data for EP1.

Figure 16 shows the existence of outliers for a few states in the data set. The existence of outliers can skew ANOVA test results.

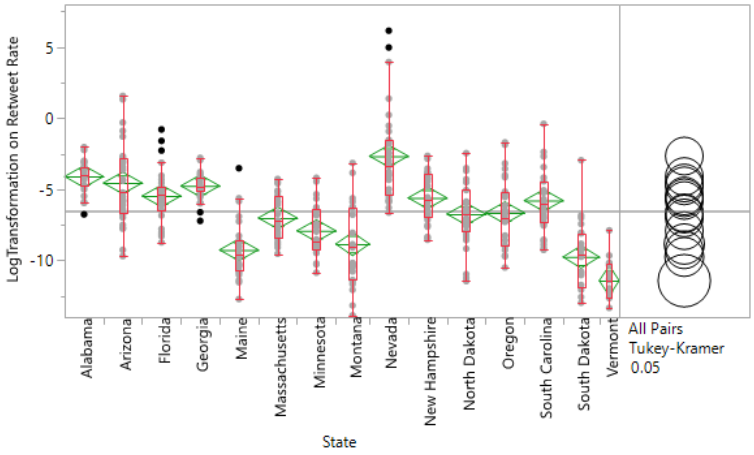

Figure 17. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Retweet Rate data for EP1.

Figure 17 shows the outliers selected to be removed from the data set before rerunning the ANOVA test.

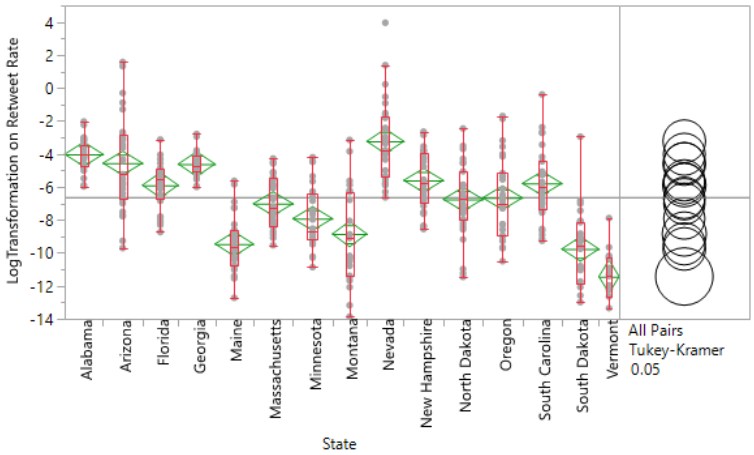

Figure 18. One-Way ANOVA test plot for the transformed Retweet Rate data for EP1 after the removal of selected outliers.

Figure 18 shows the data set after the removal of the selected outliers. It can be seen that one noticeable outlier still remains; however, it is unlikely that it will skew the ANOVA test's results.

The results of a One-Way ANOVA for Retweet rate with outliers removed can be seen in Tables 23 and 24. The outliers excluded were rows 8, 80, 87, 88, 101, 102, 155, 258, and 259. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 25.

Table 23. One-Way ANOVA output after the exclusion of outliers for Retweet Rate for EP1.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	1943.2928	138.807	33.1482	<.0001*
Error	420	1758.7316	4.187		
C. Total	434	3702.0244			

Table 24. Means for the One-Way ANOVA for the transformed Retweet Rate for EP1 after the removal of selected outliers.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	30	-4.011	0.37361	-4.75	-3.28
Arizona	31	-4.552	0.36753	-5.27	-3.83
Florida	28	-5.887	0.38672	-6.65	-5.13
Georgia	29	-4.606	0.37999	-5.35	-3.86
Maine	30	-9.448	0.37361	-10.18	-8.71
Massachusetts	31	-7.011	0.36753	-7.73	-6.29
Minnesota	31	-7.912	0.36753	-8.63	-7.19
Montana	27	-8.848	0.39382	-9.62	-8.07
Nevada	29	-3.226	0.37999	-3.97	-2.48
New Hampshire	31	-5.596	0.36753	-6.32	-4.87
North Dakota	31	-6.739	0.36753	-7.46	-6.02
Oregon	31	-6.655	0.36753	-7.38	-5.93
South Carolina	31	-5.781	0.36753	-6.50	-5.06
South Dakota	29	-9.755	0.37999	-10.50	-9.01
Vermont	16	-11.428	0.51158	-12.43	-10.42

Table 25. Connecting Letters Report for each state for the transformed Retweet Rate for EP1 after the removal of selected outliers.

Level										Mean
Nevada	A									-3.22605
Alabama	A	B								-4.01141
Arizona	A	B	C							-4.55204
Georgia	A	B	C							-4.60638
New Hampshire		B	C	D						-5.59610
South Carolina		B	C	D						-5.78120
Florida			C	D						-5.88712
Oregon				D	E					-6.65523
North Dakota				D	E					-6.73906
Massachusetts				D	E	F				-7.01146
Minnesota					E	F	G			-7.91195
Montana						F	G	H		-8.84785
Maine							G	H	I	-9.44817
South Dakota								H	I	-9.75490
Vermont									I	-11.42750

The results for the ANOVA tests before and after the removal of outliers still indicate that there is a significant difference between states' Retweet Rates for EP1. In addition, the Connecting Letters Reports indicate that not only is there a significant difference between states' Retweet Rates for EP1, but this difference can be used to break the states up into multiple groups.

4.3 Social Media Results for English Phrase 2 (EP2)

4.3.1 Social Media Results for EP2 Tweet Rate

Figures 19, 20, and 21 are the non-transformed data for the three rates calculated for the second English search word “suicidal”.

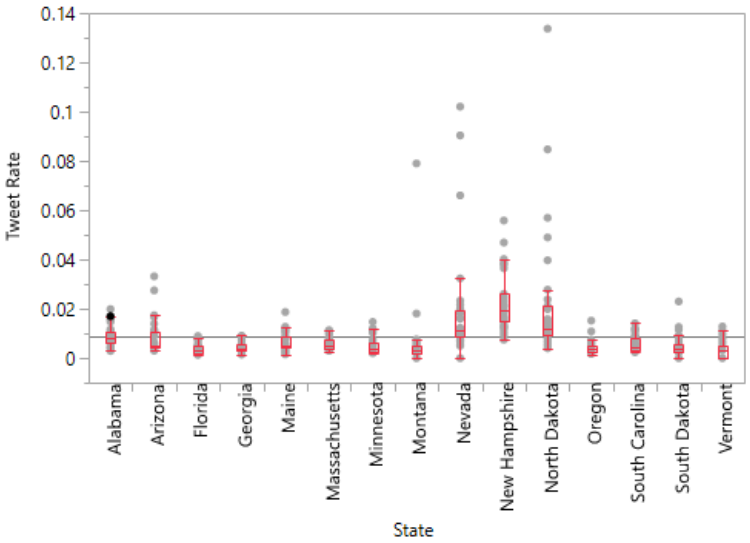

Figure 19. Plot of the Tweet Rate data for EP2 for each state including a Box Plot (seen as the small red lines on graph).

Figure 19 shows the existence of outliers in the data set for Tweet Rate for EP2. Some of the outliers are more than 10 times the data set’s average suggesting that a data transformation may be appropriate.

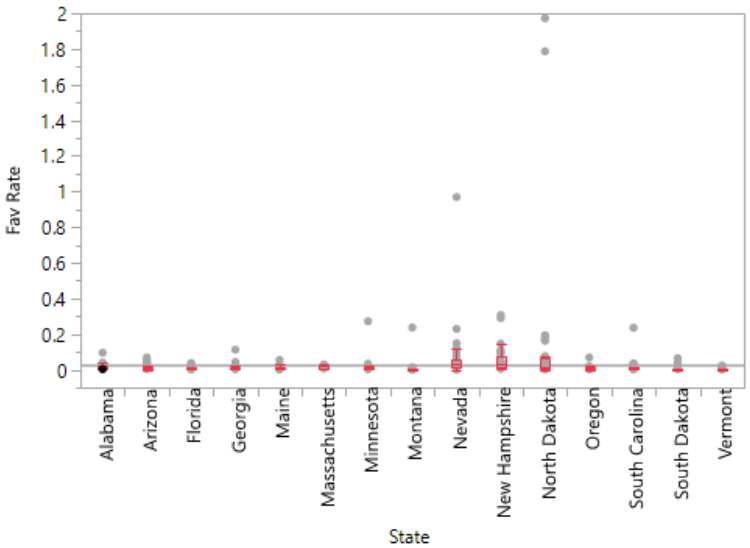

Figure 20. Plot of the Favorite Rate data for EP2 for each state including a Box Plot (seen as the small red lines on graph).

Figure 20 shows the existence of outliers in the data set for Tweet Rate for EP2. Some of the outliers are more than 10 times the data set's average suggesting that a data transformation may be appropriate.

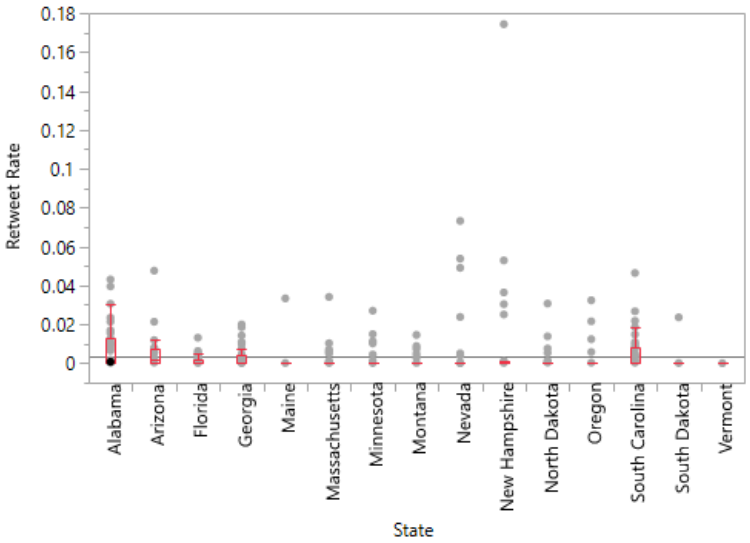

Figure 21. Plot of the Retweet Rate data for EP2 for each state including a Box Plot (seen as the small red lines on graph).

Figure 21 shows the existence of outliers in the data set for Tweet Rate for EP2. Some of the outliers are more than 10 times the data set's average suggesting that a data transformation may be appropriate.

Figures 22, 23, and 24 are plots of each state's data after a natural log transformation. Twenty-two data points were removed from the Tweet Rate data set as the transformation could not be performed on a zero. Fifty-three data points were excluded after the data transformation for the Favorite Rate and 86 data points of zero were excluded from the Retweet Rate data set.

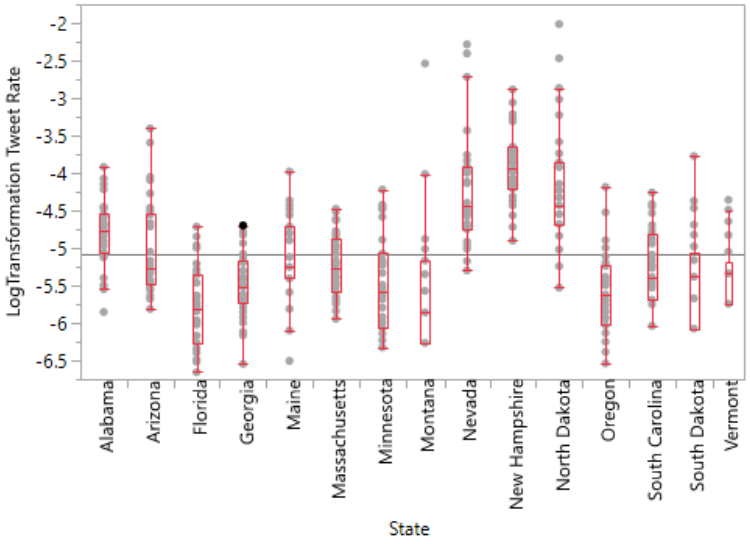

Figure 22. Plot of the Tweet Rate data for EP2 after a natural log transformation including a Box Plot (see as the small red lines on graph).

Figure 22 shows that the data set for Tweet Rate for EP2 still contains outliers, but these outliers are much closer to the data set's average.

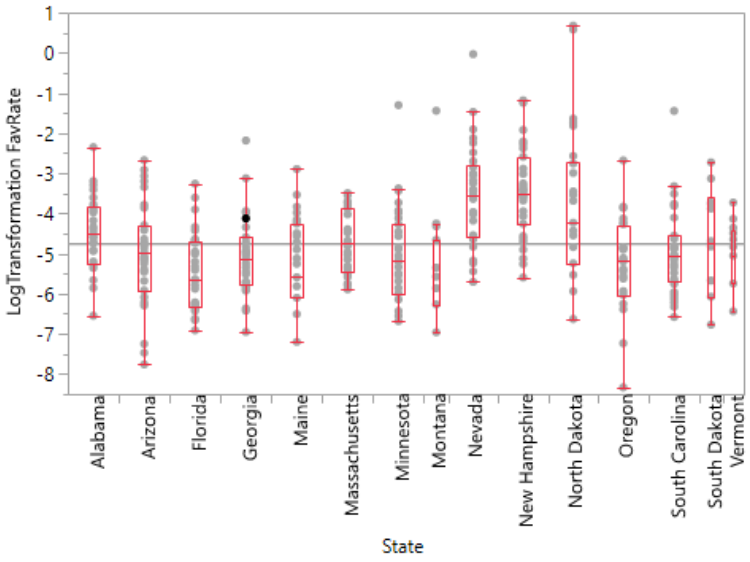

Figure 23. Plot of the Favorite Rate data for EP2 after a natural log transformation including a Box Plot (seen as the small red lines on graph).

Figure 23 shows that the data set for Favorite Rate for EP2 still contains outliers, but these outliers are much closer to the data set's average.

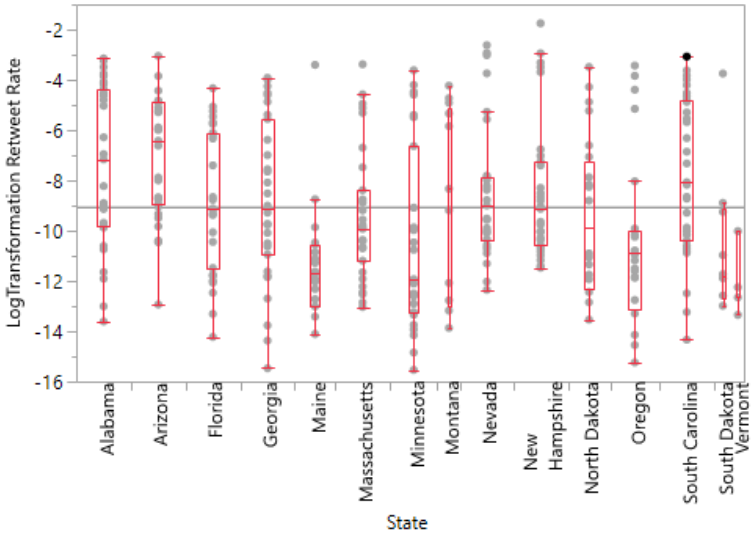

Figure 24. Plot of the Retweet Rate data for EP2 after a natural log transformation including a Box Plot (seen as the small red lines on graph).

Figure 24 shows that the data set for Retweet Rate for EP2 still contains outliers, but these outliers are much closer to the data set's average.

The results of a One-Way ANOVA test on the transformed EP2 data for Retweet Rate can be seen in Tables 26 and 27. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 28.

Table 26. One-Way ANOVA test output for the transformed Tweet Rate for EP2.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	135.99882	9.71420	28.0901	<.0001*
Error	428	148.01225	0.34582		
C. Total	442	284.01107			

Table 27. Means for the One-Way ANOVA for the transformed Tweet Rate for EP2.

Level	Number	Mean	Std. Error	Lower 95%	Upper 95%
Alabama	31	-4.7873	0.10562	-4.995	-4.580
Arizona	31	-5.0280	0.10562	-5.236	-4.820
Florida	31	-5.7904	0.10562	-5.998	-5.583
Georgia	31	-5.4715	0.10562	-5.679	-5.264
Maine	31	-5.1236	0.10562	-5.331	-4.916
Massachusetts	31	-5.2257	0.10562	-5.433	-5.018
Minnesota	31	-5.4733	0.10562	-5.681	-5.266
Montana	26	-5.5623	0.11533	-5.789	-5.336
Nevada	30	-4.2266	0.10737	-4.438	-4.016
New Hampshire	31	-3.9083	0.10562	-4.116	-3.701
North Dakota	31	-4.2141	0.10562	-4.422	-4.006
Oregon	31	-5.6016	0.10562	-5.809	-5.394
South Carolina	31	-5.2205	0.10562	-5.428	-5.013
South Dakota	25	-5.4151	0.11761	-5.646	-5.184
Vermont	21	-5.3720	0.12833	-5.624	-5.120

Table 28. Connecting Letters Report for each state for the transformed Tweet Rate for EP2.

Level						**Mean**
New Hampshire	A					-3.908277
North Dakota	A					-4.214061
Nevada	A					-4.226586
Alabama		B				-4.787265
Arizona		B	C			-5.027952
Maine		B	C	D		-5.123597
South Carolina		B	C	D		-5.220523
Massachusetts		B	C	D		-5.225744
Vermont			C	D	E	-5.371982
South Dakota			C	D	E	-5.415052
Georgia			C	D	E	-5.471493
Minnesota			C	D	E	-5.473336
Montana				D	E	-5.562293
Oregon				D	E	-5.601621
Florida					E	-5.790441

Figure 25 is a plot depicting the One-Way ANOVA multiple comparison results. The presence of outliers can skew results thus they were recognized, see Figure 26, and excluded from the data set. Figure 27 displays a plot for the One-Way ANOVA test run with the removal of most outliers.

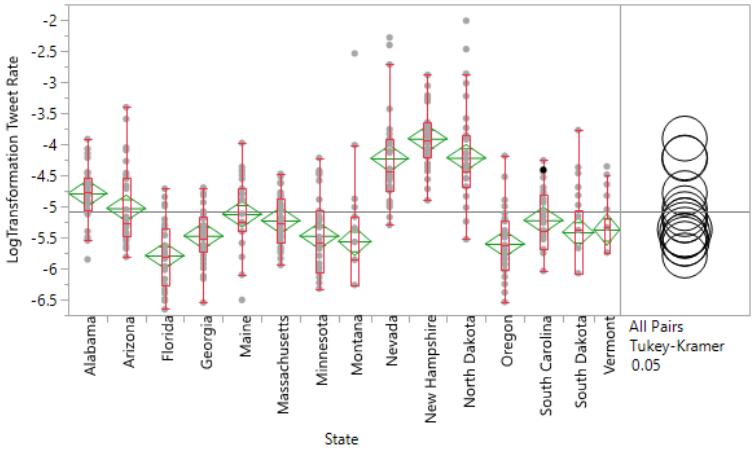

Figure 25. One-Way ANOVA test plot for the transformed Tweet Rate data for EP2.

Figure 25 shows the remaining the outliers in the Tweet Rate data set for EP2. The presence of outliers can skew ANOVA test results.

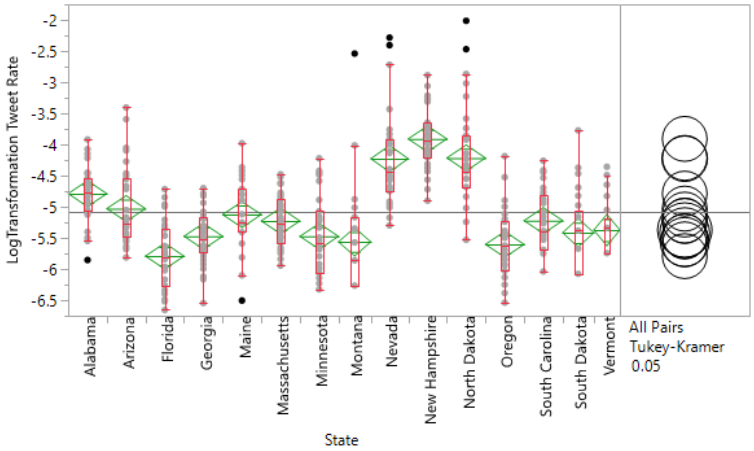

Figure 26. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Tweet Rate data for EP2.

Figure 26 shows the selection of outliers in the data set that are to be removed before rerunning an ANOVA test.

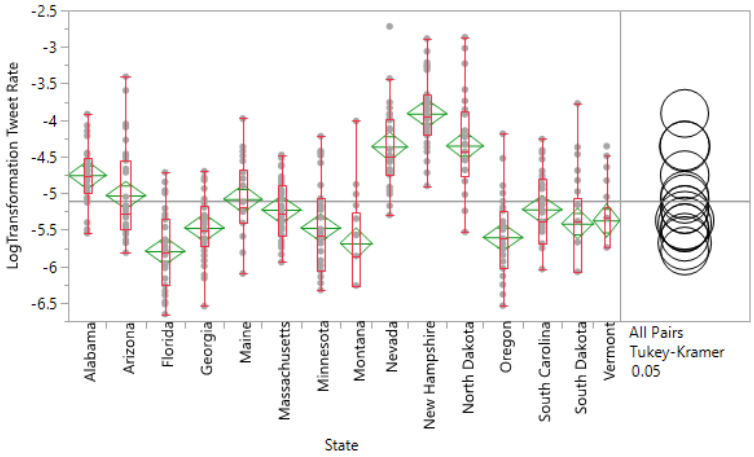

Figure 27. One-Way ANOVA test plot for the transformed Tweet Rate data for EP2 after the removal of selected outliers.

Figure 27 shows the data set for Tweet Rate for EP2 after the selection and removal of outliers in the previous step. The figure also shows that an outlier is still present; however, it is unlikely to skew the ANOVA test's results.

The results of a One-Way ANOVA for Tweet rate with outliers removed can be seen in Tables 29 and 30. The outliers excluded were rows 11, 140, 248, 274, 275, 328, and 333. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 31.

Table 29. One-Way ANOVA output after the exclusion of outliers for Tweet Rate for EP2.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	124.06588	8.86185	31.2691	<.0001*
Error	421	119.31381	0.28341		
C. Total	435	243.37968			

Table 30. Means for the One-Way ANOVA for the transformed Tweet Rate for EP2 after the removal of selected outliers.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	30	-4.7517	0.09719	-4.943	-4.561
Arizona	31	-5.0280	0.09561	-5.216	-4.840
Florida	31	-5.7904	0.09561	-5.978	-5.603
Georgia	31	-5.4715	0.09561	-5.659	-5.284
Maine	30	-5.0776	0.09719	-5.269	-4.887
Massachusetts	31	-5.2257	0.09561	-5.414	-5.038
Minnesota	31	-5.4733	0.09561	-5.661	-5.285
Montana	25	-5.6833	0.10647	-5.893	-5.474
Nevada	28	-4.3611	0.10061	-4.559	-4.163
New Hampshire	31	-3.9083	0.09561	-4.096	-3.720
North Dakota	29	-4.3502	0.09886	-4.545	-4.156
Oregon	31	-5.6016	0.09561	-5.790	-5.414
South Carolina	31	-5.2205	0.09561	-5.408	-5.033
South Dakota	25	-5.4151	0.10647	-5.624	-5.206
Vermont	21	-5.3720	0.11617	-5.600	-5.144

Table 31. Connecting Letters Report for each state for the transformed Tweet Rate for EP2 after the removal of selected outliers.

Level										**Mean**
New Hampshire	A									-3.908277
North Dakota		B								-4.350187
Nevada		B								-4.361137
Alabama			C							-4.751748
Arizona				D						-5.027952
Maine				D	E					-5.077576
South Carolina				D	E	F				-5.220523
Massachusetts				D	E	F				-5.225744
Vermont					E	F	G			-5.371982
South Dakota						F	G	H		-5.415052
Georgia						F	G	H		-5.471493
Minnesota						F	G	H		-5.473336
Oregon							G	H	I	-5.601621
Montana								H	I	-5.683258
Florida									I	-5.790441

The results for the ANOVA tests before and after the removal of outliers still indicate that there is a significant difference between states' Tweet Rates for EP2. In addition, the Connecting Letters Reports indicate that not only is there a significant difference between states' Tweet Rates for EP2, but this difference can used to break the states up into multiple groups.

4.3.2 Social Media Results for EP2 Favorite Rate

The results of a One-Way ANOVA test on the transformed EP2 data for Favorite Rate can be seen in Tables 32 and 33. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 34.

Table 32. One-Way ANOVA test output for the transformed Favorite Rate for EP2.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	170.52383	12.1803	8.7425	<.0001*
Error	397	553.11363	1.3932		
C. Total	411	723.63746			

Table 33. Means for the One-Way ANOVA for the transformed Favorite Rate for EP2.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-4.5326	0.21200	-4.949	-4.116
Arizona	31	-5.0088	0.21200	-5.426	-4.592
Florida	31	-5.4652	0.21200	-5.882	-5.048
Georgia	31	-5.0773	0.21200	-5.494	-4.661
Maine	31	-5.2695	0.21200	-5.686	-4.853
Massachusetts	31	-4.7270	0.21200	-5.144	-4.310
Minnesota	31	-5.0959	0.21200	-5.513	-4.679
Montana	15	-5.4828	0.30477	-6.082	-4.884
Nevada	30	-3.5772	0.21550	-4.001	-3.154
New Hampshire	31	-3.4655	0.21200	-3.882	-3.049
North Dakota	30	-3.8732	0.21550	-4.297	-3.450
Oregon	31	-5.2768	0.21200	-5.694	-4.860
South Carolina	31	-4.9627	0.21200	-5.379	-4.546
South Dakota	14	-4.7795	0.31546	-5.400	-4.159
Vermont	13	-5.1590	0.32737	-5.803	-4.515

Table 34. Connecting Letters Report for each state for the transformed Favorite Rate for EP2.

Level					Mean
New Hampshire	A				-3.465476
Nevada	A	B			-3.577216
North Dakota	A	B	C		-3.873179
Alabama		B	C	D	-4.532611
Massachusetts			C	D	-4.726993
South Dakota		B	C	D	-4.779536
South Carolina				D	-4.962703
Arizona				D	-5.008821
Georgia				D	-5.077327
Minnesota				D	-5.095860
Vermont			C	D	-5.158980
Maine				D	-5.269473
Oregon				D	-5.276776
Florida				D	-5.465199
Montana				D	-5.482780

Figure 28 is a plot depicting the One-Way ANOVA multiple comparison results. The presence of outliers can skew results thus they were recognized, see Figure 29, and excluded from the data set. Figure 30 displays a plot for the One-Way ANOVA test run with the removal of most outliers.

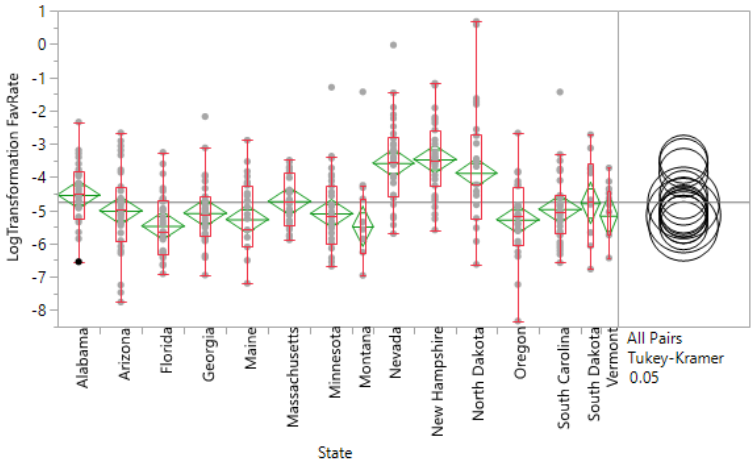

Figure 28. One-Way ANOVA test plot for the transformed Favorite Rate data for EP2.

Figure 28 shows the existence of outliers in the Favorite Rate data set for EP2. The existence of outliers can skew ANOVA test results.

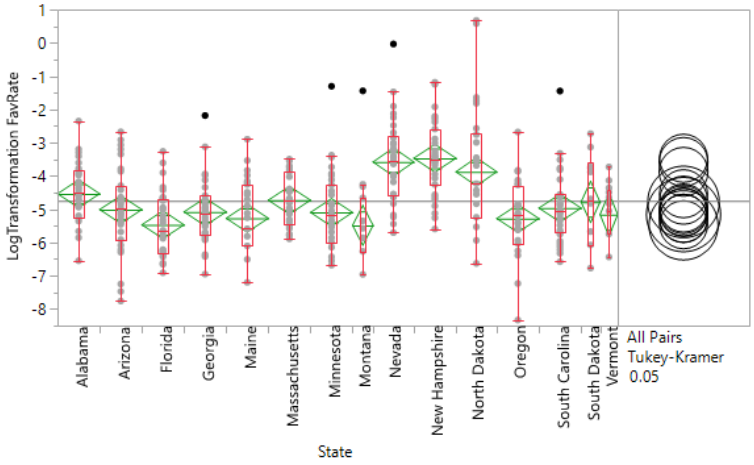

Figure 29. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Favorite Rate data for EP2.

Figure 29 shows the outliers that were selected for removal from the Favorite Rate data set for EP2.

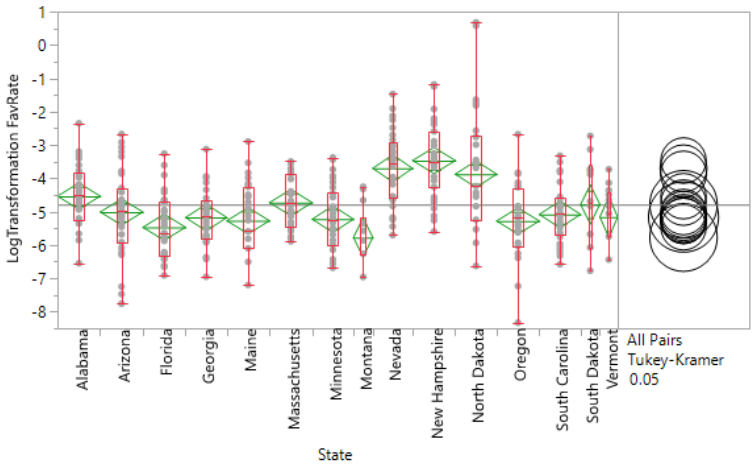

Figure 30. One-Way ANOVA test plot for the transformed Favorite Rate data for EP2 after the removal of selected outliers.

Figure 30 shows the data set after the removal of the outliers selected in the previous step. There are no longer any noticeable outliers in the data set helping assure the reliability of the ANOVA test's results.

The results of a One-Way ANOVA for Favorite rate with outliers removed can be seen in Tables 35 and 36. The outliers excluded were rows 110, 204, 248, 273, and 389. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 37.

Table 35. One-Way ANOVA output after the exclusion of outliers for Favorite Rate for EP2.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	174.16309	12.4402	10.0286	<.0001*
Error	392	486.26686	1.2405		
C. Total	406	660.42995			

Table 36. Means for the One-Way ANOVA for the transformed Favorite Rate for EP2 after the removal of selected outliers.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-4.5326	0.20004	-4.926	-4.139
Arizona	31	-5.0088	0.20004	-5.402	-4.616
Florida	31	-5.4652	0.20004	-5.858	-5.072
Georgia	30	-5.1738	0.20335	-5.574	-4.774
Maine	31	-5.2695	0.20004	-5.663	-4.876
Massachusetts	31	-4.7270	0.20004	-5.120	-4.334
Minnesota	30	-5.2223	0.20335	-5.622	-4.823
Montana	14	-5.7716	0.29767	-6.357	-5.186
Nevada	29	-3.6995	0.20682	-4.106	-3.293
New Hampshire	31	-3.4655	0.20004	-3.859	-3.072
North Dakota	30	-3.8732	0.20335	-4.273	-3.473
Oregon	31	-5.2768	0.20004	-5.670	-4.883
South Carolina	30	-5.0800	0.20335	-5.480	-4.680
South Dakota	14	-4.7795	0.29767	-5.365	-4.194
Vermont	13	-5.1590	0.30890	-5.766	-4.552

Table 37. Connecting Letters Report for each state for the transformed Favorite Rate for EP2 after the removal of selected outliers.

Level						Mean
New Hampshire	A					-3.465476
Nevada	A	B				-3.699493
North Dakota	A	B	C			-3.873179
Alabama		B	C	D		-4.532611
Massachusetts			C	D	E	-4.726993
South Dakota		B	C	D	E	-4.779536
Arizona				D	E	-5.008821
South Carolina				D	E	-5.079974
Vermont				D	E	-5.158980
Georgia				D	E	-5.173849
Minnesota				D	E	-5.222305
Maine				D	E	-5.269473
Oregon				D	E	-5.276776
Florida				D	E	-5.465199
Montana					E	-5.771581

The results for the ANOVA tests before and after the removal of outliers still indicate that there is a significant difference between states' Favorite Rates for EP2. In addition, the Connecting Letters Reports indicate that not only is there a significant difference between states' Favorite Rates for EP2, but this difference can used to break the states up into multiple groups.

4.3.3 Social Media Results for EP2 Retweet Rate

The results of a One-Way ANOVA test on the transformed EP2 data for Retweet Rate can be seen in Tables 38 and 39. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 40.

Table 38. One-Way ANOVA test output for the transformed Retweet Rate for EP2.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	642.4463	45.8890	5.2606	<.0001*
Error	364	3175.2058	8.7231		
C. Total	378	3817.6521			

Table 39. Means for the One-Way ANOVA for the transformed Retweet Rate for EP2.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-7.389	0.5305	-8.43	-6.35
Arizona	31	-7.033	0.5305	-8.08	-5.99
Florida	31	-8.755	0.5305	-9.80	-7.71
Georgia	31	-8.791	0.5305	-9.83	-7.75
Maine	23	-11.455	0.6158	-12.67	-10.24
Massachusetts	31	-9.462	0.5305	-10.51	-8.42
Minnesota	27	-10.376	0.5684	-11.49	-9.26
Montana	13	-8.917	0.8192	-10.53	-7.31
Nevada	30	-8.536	0.5392	-9.60	-7.48
New Hampshire	31	-8.492	0.5305	-9.54	-7.45
North Dakota	24	-9.465	0.6029	-10.65	-8.28
Oregon	28	-10.742	0.5582	-11.84	-9.64
South Carolina	31	-7.770	0.5305	-8.81	-6.73
South Dakota	10	-10.758	0.9340	-12.59	-8.92
Vermont	7	-11.937	1.1163	-14.13	-9.74

Table 40. Connecting Letters Report for each state for the transformed Retweet Rate for EP2.

Level						Mean
Arizona	A					-7.03339
Alabama	A	B				-7.38883
South Carolina	A	B	C			-7.76952
New Hampshire	A	B	C	D		-8.49199
Nevada	A	B	C	D		-8.53565
Florida	A	B	C	D	E	-8.75467
Georgia	A	B	C	D	E	-8.79087
Montana	A	B	C	D	E	-8.91730
Massachusetts	A	B	C	D	E	-9.46217
North Dakota	A	B	C	D	E	-9.46483
Minnesota			C	D	E	-10.37568
Oregon				D	E	-10.74164
South Dakota		B	C	D	E	-10.75803
Maine					E	-11.45492
Vermont			C	D	E	-11.93729

Figure 31 is a plot depicting the One-Way ANOVA multiple comparison results. The presence of outliers can skew results thus they were recognized, see Figure 32, and excluded from the data set. Figure 33 displays a plot for the One-Way ANOVA test run with the removal of most outliers.

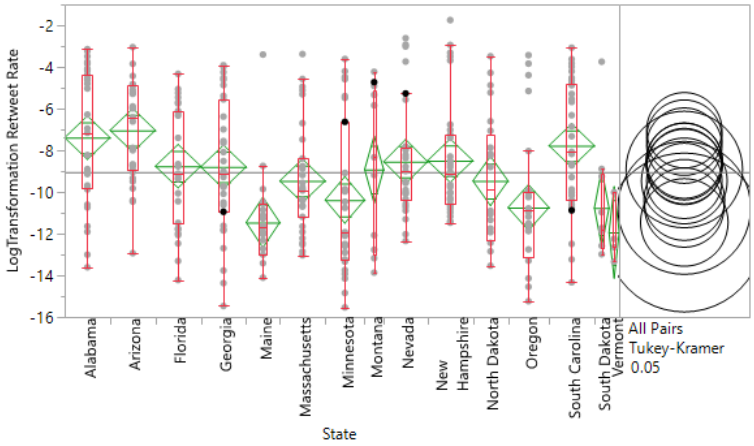

Figure 31. One-Way ANOVA test plot for the transformed Retweet Rate data for EP2.

Figure 31 shows the existence of outliers in the Retweet Rate data set for EP2. The existence of outliers can skew ANOVA test results.

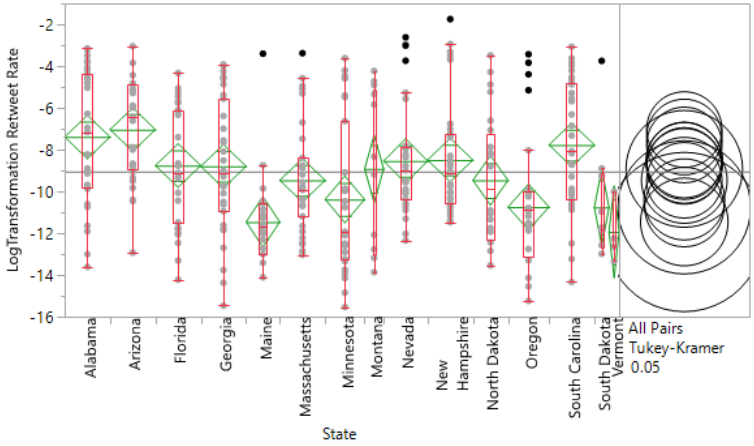

Figure 32. One-Way ANOVA test plot depicting the outliers selected for removal from the transformed Retweet Rate data for EP2.

Figure 32 shows the outliers that were selected to be removed from the Retweet Rate data set before the ANOVA test is rerun.

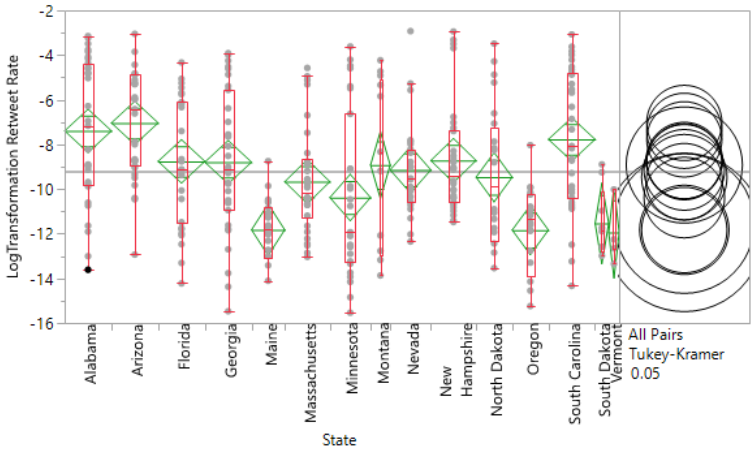

Figure 33. One-Way ANOVA test plot for the transformed Retweet Rate data for EP2 after the removal of selected outliers.

Figure 33 shows that even after the removal of outliers in the previous step some outliers still remain; however, the existence of these outliers is unlikely to skew the ANOVA test's results.

The results of a One-Way ANOVA for Retweet rate with outliers removed can be seen in Tables 41 and 42. The outliers excluded were rows 129, 164, 269, 270, 275, 288, 348, 354, 363, 371, and 423. The test was significant at a p-value less than 0.0001. In addition, a Connecting Letters Report can be seen in Table 43.

Table 41. One-Way ANOVA output after the exclusion of outliers for Retweet Rate for EP2.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
State	14	794.6007	56.7572	7.5101	<.0001*
Error	353	2667.7899	7.5575		
C. Total	367	3462.3907			

Table 42. Means for the One-Way ANOVA for the transformed Retweet Rate for EP2 after the removal of selected outliers.

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Alabama	31	-7.389	0.4938	-8.36	-6.42
Arizona	31	-7.033	0.4938	-8.00	-6.06
Florida	31	-8.755	0.4938	-9.73	-7.78
Georgia	31	-8.791	0.4938	-9.76	-7.82
Maine	22	-11.821	0.5861	-12.97	-10.67
Massachusetts	30	-9.665	0.5019	-10.65	-8.68
Minnesota	27	-10.376	0.5291	-11.42	-9.34
Montana	13	-8.917	0.7625	-10.42	-7.42
Nevada	27	-9.137	0.5291	-10.18	-8.10
New Hampshire	30	-8.717	0.5019	-9.70	-7.73
North Dakota	24	-9.465	0.5612	-10.57	-8.36
Oregon	24	-11.832	0.5612	-12.94	-10.73
South Carolina	31	-7.770	0.4938	-8.74	-6.80
South Dakota	9	-11.537	0.9164	-13.34	-9.74
Vermont	7	-11.937	1.0391	-13.98	-9.89

Table 43. Connecting Letters Report for each state for the transformed Retweet Rate for EP2 after the removal of selected outliers.

Level						Mean
Arizona	A					-7.03339
Alabama	A	B				-7.38883
South Carolina	A	B				-7.76952
New Hampshire	A	B	C			-8.71683
Florida	A	B	C			-8.75467
Georgia	A	B	C			-8.79087
Montana	A	B	C	D	E	-8.91730
Nevada	A	B	C	D		-9.13721
North Dakota	A	B	C	D	E	-9.46483
Massachusetts		B	C	D	E	-9.66499
Minnesota			C	D	E	-10.37568
South Dakota			C	D	E	-11.53725
Maine				D	E	-11.82104
Oregon					E	-11.83180
Vermont			C	D	E	-11.93729

The results for the ANOVA tests before and after the removal of outliers still indicate that there is a significant difference between states' Retweet Rates for EP2. In addition, the Connecting Letters Reports indicate that not only is there a significant difference between states' Retweet Rates for EP2, but this difference can used to break the states up into multiple groups.

The ANOVA tests conducted in this study all had p-values less than 0.0001 indicating that there was a significant difference between states in respect to all variables both before and after the removal of outliers. The Connecting Letters Reports demonstrates how many groups with significant differences the states could be broken into. Ideally, the Connecting Letters Report would break each state into its own specific group indicating that every state is significantly different from every other state; however, this does not occur. The two best performing variables were Retweet Rate for EP1 after the removal of outliers and Tweet Rate for EP2 after the removal of outliers. Both Retweet Rate for EP1 after the removal of outliers and Tweet Rate for EP2 after the removal of outliers were split into 9 groups in their respective Connecting Letters Report. While Favorite Rate for EP2 performed the worst after the removal of outliers only being split into 4 groups in its respective Connecting Letters Report.

4.4 Comparison of Twitter Data to Historical State Suicide Rates

Table 44 contains information about the states involved in this study. The table contains information regarding the state's population and age adjusted suicide rates.

Table 44. Summary Table of historical suicide trends for states in study.

State	Population	Age Adjusted Suicide Rate (2014)	Age Adjusted Suicide Rate (2015)	Age Adjusted Suicide Rate (2016)	Average Age Adjusted Suicide Rate (2014-2016)
Montana	1,050,493	23.9	25.3	25.9	25.0
Nevada	2,998,039	19.6	18.4	21.4	19.8
South Dakota	869,666	17.1	20.4	20.2	19.2
North Dakota	755,393	17.8	17.5	19	18.1
Oregon	4,142,776	18.6	17.8	17.8	18.1
Arizona	7,016,270	18	18.2	17.7	18.0
New Hampshire	1,342,795	17.8	16.5	17.2	17.2
Vermont	623,657	18.7	14.8	17.3	16.9
Maine	1,335,907	15.7	16	15.9	15.9
South Carolina	5,024,369	15.2	14.8	15.7	15.2
Alabama	4,874,747	14.5	14.9	15.7	15.0
Florida	20,984,400	13.9	14.4	14	14.1
Georgia	10,429,379	12.6	12.7	13.3	12.9
Minnesota	5,576,606	12.2	13.2	13.2	12.9
Massachusetts	6,859,819	8.2	8.9	8.8	8.6

The Pearson Product-Moment correlation coefficient was found for the combinations of EP1 Tweet Rate, EP1 Favorite Rate, and EP1 Retweet Rate with a state's average age-adjusted suicide rate using JMP. Scatterplot matrices for each combination can be seen in Figures 34, 35, and 36. In addition, the Spearman's correlation coefficient was found and noted in Table 45.

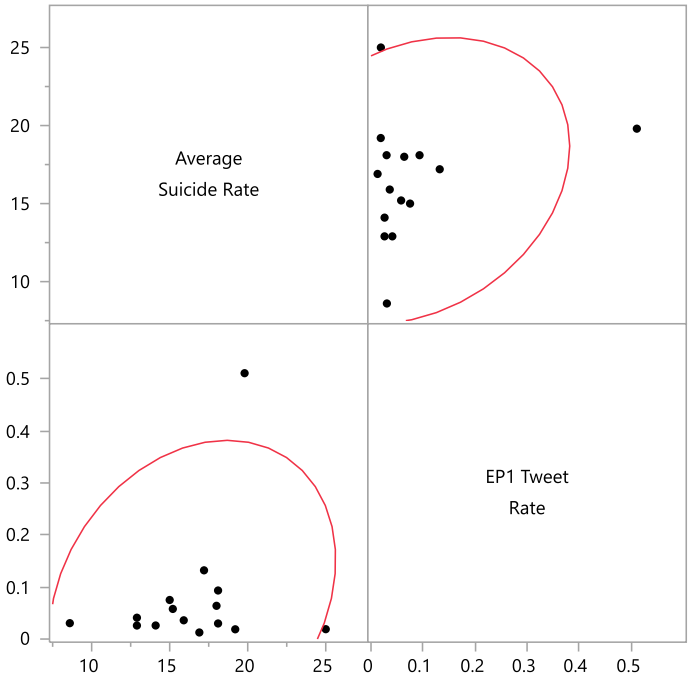

Figure 34. Scatterplot Matrix of states' average suicide rate and EP1 Tweet Rate.

Figure 34 shows that there is little correlation between a states' average suicide rate and states' Tweet Rate for EP1. Ideally, states with higher average suicide rates would also have had a higher Tweet Rate.

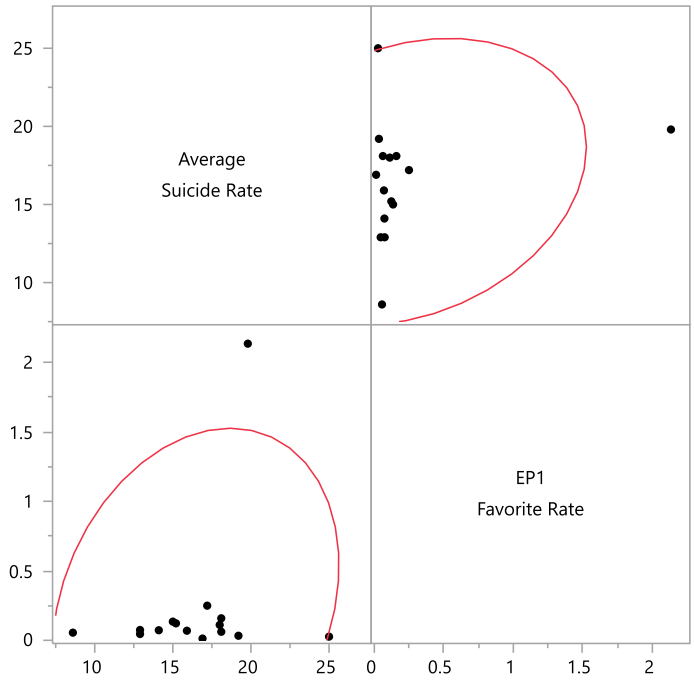

Figure 35. Scatterplot Matrix of states' average suicide rate and EP1 Favorite Rate.

Figure 35 shows that there is little correlation between a states' average suicide rate and states' Favorite Rate for EP1. Ideally, states with higher average suicide rates would also have had a higher Favorite Rate.

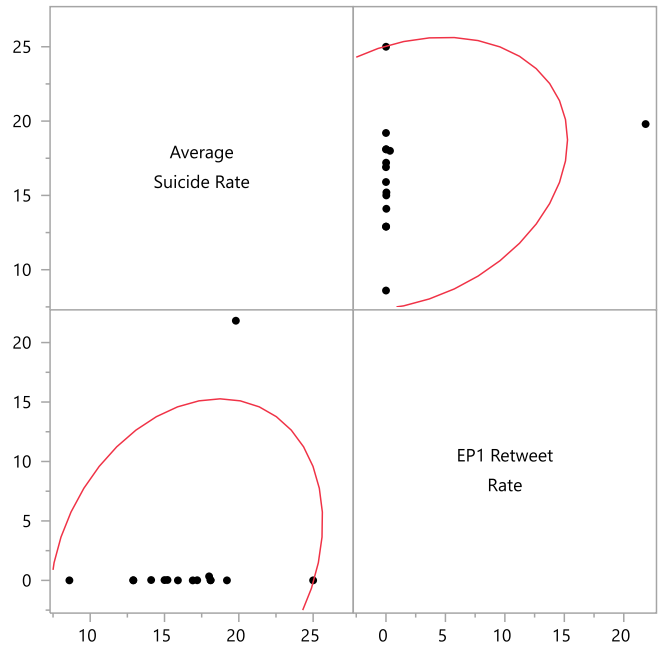

Figure 36. Scatterplot Matrix of states' average suicide rate and EP1 Retweet Rate.

Figure 36 shows that there is little correlation between a states' average suicide rate and states' Retweet Rate for EP1. Ideally, states with higher average suicide rates would also have had a higher Retweet Rate.

Table 45. The Pearson Product-Moment Correlation Coefficient and Pearson Rho Correlation Coefficient for EP1's Tweet Rate, Favorite Rate, and Retweet Rate and a state's average age-adjusted suicide rate.

Combination	Pearson Product-Moment	Pearson Rho
Tweet Rate-Suicide Rate	0.243	0.068
Favorite Rate-Suicide Rate	0.242	0.057
Retweet Rate-Suicide Rate	0.248	0.097

Two variables that are perfectly correlated will have a correlation coefficient of either 1 (perfect positive correlation) or -1 (perfect negative correlation) while two variables that are not correlated at all will have a correlation coefficient of zero. The correlation coefficients calculated show weak correlations for the Person Product-Moment method and very weak correlations for Pearson Rho method. The correlation coefficients found suggest that there is not a strong relationship between EP1 variables and a state's age-adjusted historical suicide rate.

The Pearson Product-Moment correlation coefficient was found for the combinations of EP2 Tweet Rate, EP2 Favorite Rate, and EP2 Retweet Rate with a state's average age-adjusted suicide rate using JMP. Scatterplot matrices for each combination can be seen in Figures 37, 38, and 39. In addition, the Spearman's correlation coefficient was found and noted in Table 46.

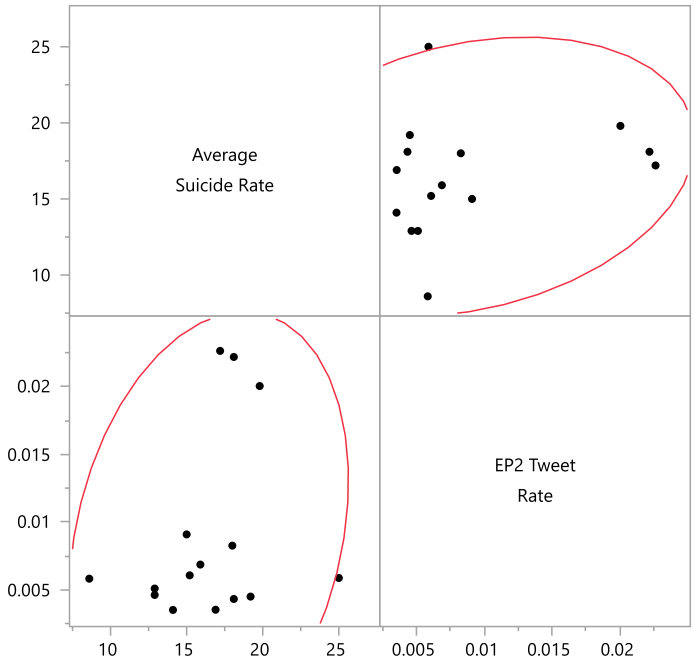

Figure 37. Scatterplot Matrix of states' average suicide rate and EP2 Tweet Rate.

Figure 37 shows that there is little correlation between a states' average suicide rate and states' Tweet Rate for EP2. Ideally, states with higher average suicide rates would also have had a higher Tweet Rate.

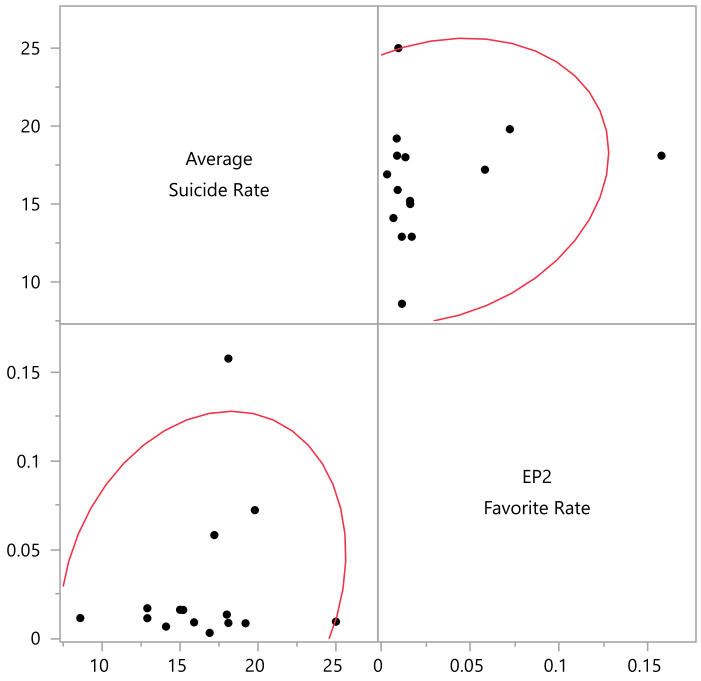

Figure 38. Scatterplot Matrix of states' average suicide rate and EP2 Favorite Rate.

Figure 38 shows that there is little correlation between a states' average suicide rate and states' Favorite Rate for EP2. Ideally, states with higher average suicide rates would also have had a higher Favorite Rate.

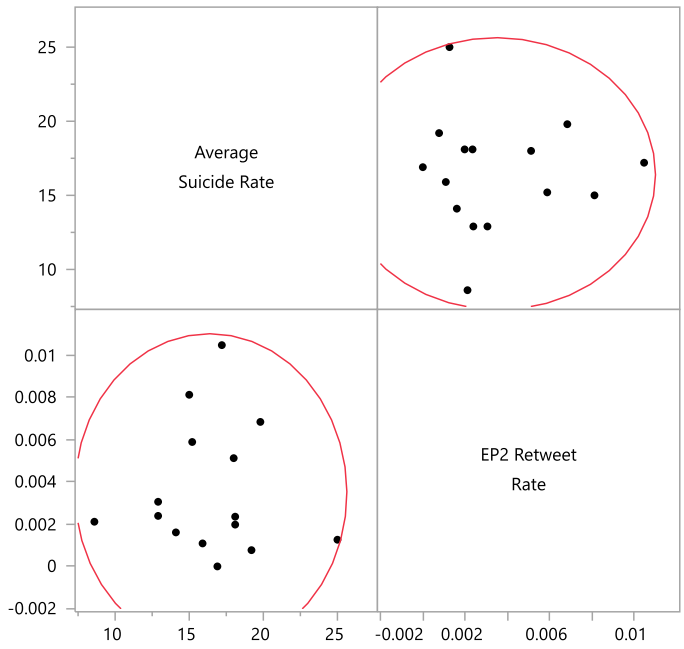

Figure 39. Scatterplot Matrix of states' average suicide rate and EP2 Retweet Rate.

Figure 39 shows that there is little correlation between a states' average suicide rate and states' Retweet Rate for EP2. Ideally, states with higher average suicide rates would also have had a higher Retweet Rate.

Table 46. The Pearson Product-Moment Correlation Coefficient and Pearson Rho Correlation Coefficient for EP2's Tweet Rate, Favorite Rate, and Retweet Rate and a state's average age-adjusted suicide rate.

Combination	Pearson Product-Moment	Pearson Rho
Tweet Rate-Suicide Rate	0.253	0.261
Favorite Rate-Suicide Rate	0.199	0.043
Retweet Rate-Suicide Rate	-0.008	-0.145

The correlation coefficients calculated show weak correlations for the Person Product-Moment method and very weak correlations for Pearson Rho method. The correlation coefficients found suggest that there is not a strong relationship between EP2 variables and a state's age-adjusted historical suicide rate.

Chapter 5

DISCUSSION

5.1 National Multiple Regression Model Conclusions

The first purpose of this study was to identify significant factors relating to suicide rates in a national multiple regression model. Two national multiple regression models were created one using the data from the same Quarter for suicide rates and predictor variables and one using data with Quarters for suicide rates offset by 1. The first national multiple regression model identified five factors: Foreclosure Rates, Violent Crime Rates, GDP Change, Gini ratio, and Consumption Volume. The second national multiple regression model identified six factors: Foreclosure Rates, Violent Crime Rates, Unemployment Rates, Gini ratio, Fertility Measure, and Consumption Volume. Both models identified Foreclosure Rates, Violent Crime Rates, Gini ratio, and Consumption Volume as factors relating to suicide rates. The first model identified GDP Change as a factor; however, its absence in the second model means that it may not be needed in a predictive state level multiple regression model. Fertility Measure and Unemployment Rates were both identified in the second model but were not identified in the first. Since the ultimate goal in the future would be developing a model for prediction variables identified in the second national level multiple regression model, but not in the first, should still be considered for use in the state level multiple regression model.

The variables identified by both models: Foreclosure Rates, Violent Crime Rates, Gini Ratio, and Consumption Volume should be used in a predictive state level regression model. While the variables Fertility Measure and Unemployment Rates should be considered for the state level model. Data on foreclosure rates, violent crime rates, fertility measures, and unemployment rates can be found at a state level; however, it may be difficult to locate reliable consumption volume data at the state level.

5.2 Social Media Conclusions

The goal of the Social Media Analysis conducted in this paper was two-fold. The first goal was to demonstrate that there is a significant difference in the rate of “tweets” relating to suicide

between states. If this could be demonstrated, it would further support the use of social media data in predictive suicide models. The second goal was to find evidence that suggests that historically higher rates of suicide are observed in states with higher rates of “tweets” relating to suicide. The analysis conducted in this paper suggest that there is a significant difference between states regarding “tweet” rates relating to suicide. However, there was little evidence to suggest that a higher rate of “tweets” relating to suicide had any correlation with a state’s historical suicide rate.

5.2.1 Differentiating Rates Between States

The results found in this study strongly suggest that there is a significant difference between the Tweet Rates, Favorite Rates, and Retweet Rates between states regarding “tweets” related to suicide. In all six ANOVA tests conducted, both on data with and without outliers, a p-value of <.0001 was found. In addition, the connecting letters reports, ranging from A to I in regards to the Retweet Rate for EP1 after the removal of outliers and Retweet Rate for EP2 after the removal of outliers, to A to D in regards to the Favorite Rate for EP2 before the removal of outliers, suggests that there was not only a measurable difference in states’ “tweeting” rates, but the difference was significant enough to break the states into multiple groups. The ability to separate states into groups based on Twitter data suggests that the use of this data in suicide prediction models may be appropriate.

5.2.2 Relating Rates of Tweets to Historical Suicide Trends

The results of this study do not suggest that the rate of “tweets”, rate of favorites, or rate of retweets of the two phrases examined are correlated to a state’s historical suicide rate. Both correlation tests performed, Pearson Product-Moment and Spearman Rho, reported low correlation coefficients. In general, the Pearson Product-Moment correlation coefficients were larger, being around 0.2 (excluding Retweet Rate for EP2), then the Spearman Rho coefficients, being between -0.15 and 0.1 (excluding Tweet Rate for EP2). The low correlation coefficients suggest that it may be inappropriate to include Twitter data in a state level multiple linear regression model.

5.3 Conclusion

The first purpose of this study was to identify significant factors relating to suicide rates in a national multiple linear regression model. The models created identified the factors Foreclosure Rates, Violent Crime Rates, GDP Change, Gini Ratio, Consumption Volume, Unemployment Rates, and Fertility Measure. After considerations to the intended use of these factors, the development of a model for prediction, it is recommended that the factors Foreclosure Rates, Violent Crime Rates, Gini ratio, and if possible, Consumption Volume be included in a state level model with consideration given to the factors Unemployment Rate and Fertility Measure.

The second purpose of this study was to evaluate the usage of social media data in a state level multiple linear regression model for predicting suicide rates. Although strong evidence that there is a significant difference in Twitter data between states was found, there was little evidence to suggest that these differences corresponded to differences in the states' suicide rates. Based on this outcome it would be recommended that further research be conducted before incorporating Twitter data into a predictive state level suicide model.

5.4 Ethical Considerations

The development of a predictive state level model for suicide rates that can assist organizations in distributing funding in the most "efficient" manner for suicide prevention does have several ethical concerns. The primary concern being that the most "efficient" use of suicide prevention funding may not be the most equitable. Organizations dedicated to funding programs in the hopes of reducing suicides, should be careful not to exclusively focus on reducing the greatest total number of suicides in the United States. In doing so, these organizations run the risk of neglecting rural areas, which despite having a relatively high suicide rate, have a relatively low number of suicides due to their smaller populations.

The way in which decision makers utilize the results of a predictive state level model also needs to be addressed. At best a predictive model can only give an approximate estimate of a future variable bounded by some confidence level. Since the results of a predictive model are not guaranteed to be correct decision makers should only use the results as one of many inputs when

making decisions. Researchers should also recognize the fact that there is no guarantee that the variables recognized in the national level model will behave the same in each state within the United States. The difference from state to state in the significance of the identified variables can lead to certain states receiving more funding than appropriate. To combat biasing based on the factors included in a state level model researchers should consistently test any model used to inform funding decisions to assure that the model works appropriately for each state.

5.5 Study Limitations

This study was subject to several limitations, both when creating the national model and collecting and analyzing the social media data. The primary limitation when creating the national model was the inconsistency in the time span certain variables covered. Variables ranged from being monthly, quarterly, and even annually. Converting monthly variables to quarterly variables was a straightforward process; however, converting annual variables to quarterly variables was, at best, questionable. As previously stated, crime rate, Gini ratio, and consumption rate were all found to be significant variables in the model and all three of these variables were originally annual metrics.

The two main limitations in the social media analysis portion of this study arose from limited data access and difficulty in text filtering. The difficulty in obtaining historical Twitter data lead to the collection of current Twitter data, which then had to be compared to historical suicide rates. It would have been ideal to compare the Twitter data gathered in this study to the actual suicide rate in the states over the same time period; however, this data is currently not available. Also, much of the Twitter data collected could not be analyzed in this study due to the complexity of text filtering. The three search phrases/words not analyzed in this study, “feel hopeless”, “feel depressed”, and “Prozac” all posed significant difficulty in removing false positives. Ultimately, it was decided that adequately filtering the results for these three phrases was beyond the scope of this study.

5.6 Future Work

Future work on this study would include conducting analysis on the three search phrases/words not analyzed and testing if the social media analysis results varied when compared to suicide rates in each state by age group. In addition, further work should be done in the text filtering portion of the social media analysis. A simplistic method of removing Tweets was used in this study to demonstrate that false positive Tweets could be identified and removed from the data set; however, more in depth text filtering methods should be explored. For example, future researchers could attempt to classify tweets by tone and exclude tweets based on this classification. Researchers could also investigate the use of machine learning algorithms to identify and remove false positive tweets. Also, more attention should be given to the identified points of influence recognized in the social media analysis. Future work could look for a relationship between the points of influence days and the number of suicides recorded on those days or look for similarities between Tweets that were retweeted many times. Future work could also include incorporating interaction terms in the national level model developed to gain further insight into the variables that influence suicide rates.

BIBLIOGRAPHY

Action Alliance. (n.d.). Retrieved from https://theactionalliance.org/about-us

Agerbo, E., Sc, M., Mortensen, P. B., & Sc, M. (2003). Suicide Risk in Relation to Socioeconomic , Demographic , Psychiatric , and Familial Factors : A National Register – Based Study of All Suicides in Denmark , 1981 – 1997, (April), 765–772.

Biddle, L., Donovan, J., Hawton, K., Kapur, N., & Gunnell, D. (2008). Suicide and the internet. *BMJ : British Medical Journal*, *336*(7648), 800–802. https://doi.org/10.1136/bmj.39525.442674.AD

Cho, Y. I., Johnson, T. P., & Fendrich, M. (2001). Monthly variations in self-reports of alcohol consumption. *J Stud Alcohol*, *62*(2), 268–272. https://doi.org/10.15288/jsa.2001.62.268

Classen, T, Dunn, R. (2012). The effect of job loss and unemployment duration on suicide risk in the United States: a new look using mass-layoffs and unemployment duration. *Health Economics*, *21(1)*, 338-350. https://doi.org/10.1002/hec

Coppola, I., Marangon, D., Gramaglia, C., Delicato, C., Di Marco, S., Gattoni, E., … Zeppegno, P. (2016). In a period of economical crisis who is at risk for attempted suicide? *European Psychiatry*, *33*, S598. https://doi.org/10.1016/j.eurpsy.2016.01.2231

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 2098–2110. https://doi.org/10.1145/2858036.2858207

Fountoulakis, K. N., Savopoulos, C., Apostolopoulou, M., Dampali, R., Zaggelidou, E., Karlafti, E., … Hatzitolios, A. I. (2015). Rate of suicide and suicide attempts and their relationship to unemployment in Thessaloniki Greece (2000-2012). *Journal of Affective Disorders*, *174*, 131–136. https://doi.org/10.1016/j.jad.2014.11.047

Hedegaard, H., Warner, M., & Curtin, S. C. (2016). Increase in suicide in the United States, 1999-2014. *NCHS Data Brief*, (241). https://doi.org/10.1073/pnas.1518393112

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, *35*(1), 51–59.

https://doi.org/10.1027/0227-5910/a000234

Lauritsen, J. L., & White, N. (2014). Seasonal Patterns in Criminal Victimization Trends, (June), 22.

Lee, K. S., Lee, H., Myung, W., Song, G.-Y., Lee, K., Kim, H., … Kim, D. K. (2018). Advanced Daily Prediction Model for National Suicide Numbers with Social Media Data. *Psychiatry Investigation*, *15*(4), 344–354. https://doi.org/10.30773/pi.2017.10.15

Li, Z., Page, A., Martin, G., & Taylor, R. (2011). Attributable risk of psychiatric and socio-economic factors for suicide from individual-level, population-based studies: A systematic review. *Social Science and Medicine*, *72*(4), 608–616. https://doi.org/10.1016/j.socscimed.2010.11.008

Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social media and suicide: A public health perspective. *American Journal of Public Health*, *102*(SUPPL. 2). https://doi.org/10.2105/AJPH.2011.300608

McCarthy, J. F., Bossarte, R. M., Katz, I. R., Thompson, C., Kemp, J., Hannemann, C. M., … Schoenbaum, M. (2015). Predictive modeling and concentration of the risk of suicide: Implications for preventive interventions in the us department of veterans affairs. *American Journal of Public Health*, *105*(9), 1935–1942. https://doi.org/10.2105/AJPH.2015.302737

Noh, Y. H. (2009). Does unemployment increase suicide rates? The OECD panel evidence. *Journal of Economic Psychology*, *30*(4), 575–582. https://doi.org/10.1016/j.joep.2009.04.003

O'Dea, B., Larsen, M. E., Batterham, P. J., Calear, A. L., & Christensen, H. (2017). A Linguistic analysis of suicide-related Twitter posts. *Crisis*, *38*(5), 319–329. https://doi.org/10.1027/0227-5910/a000443

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on twitter. *Internet Interventions*, *2*(2), 183–188. https://doi.org/10.1016/j.invent.2015.03.005

American Foundation for Suicide Prevention. (2017, September 07). Retrieved from https://afsp.org

Phillips, M. R., Yang, G., Zhang, Y., Wang, L., Ji, H., & Zhou, M. (2002). 1-s2.0-S0140673602116813-main, *360*, 1728–1736.

Recupero, P. R., Harms, S. E., & Noble, J. M. (2008). Googling suicide: Surfing for suicide information on the Internet. *The Journal of Clinical Psychiatry, 69*(6), 878-888. http://dx.doi.org/10.4088/JCP.v69n0601

Rodrı, A. (2005). Income inequality , unemployment , and suicide : a panel data analysis of 15 European countries, 439–451. https://doi.org/10.1080/0003684042000295304

Stack, S. (2000). Suicide: a 15-year review of the sociological literature. Part I: cultural and economic factors. *Suicide and Life-Threatening Behavior*, *30*(2), 145–162. https://doi.org/10.1111/j.1943-278X.2000.tb01073.x

Stack, S., & Wasserman, I. (2007). Economic Strain and Suicide Risk: A Qualitative Analysis. *Suicide and Life-Threatening Behavior*, *37*(1), 103–112. https://doi.org/10.1521/suli.2007.37.1.103

Substance Abuse and Mental Health Services Administration. (2013, May 13). Retrieved from https://www.samhsa.gov

Sueki, H. (2015). The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of Affective Disorders*, *170*, 155–160. https://doi.org/10.1016/j.jad.2014.08.047

Won, H. H., Myung, W., Song, G. Y., Lee, W. H., Kim, J. W., Carroll, B. J., & Kim, D. K. (2013). Predicting National Suicide Numbers with Social Media Data. *PLoS ONE*, *8*(4), 1–6. https://doi.org/10.1371/journal.pone.0061809

Appendix A

RSTUDIO CODE FOR NATIONAL MODEL

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```{r}
setwd("C:/Users/Derek/Desktop/Suicide_Thesis")
suicide_data <- read.csv(file = "Historical_suicide_rate_by_quarter.csv", header = T, sep = ",")
suicide_data
```
```{r}
stacking_sui_data <- cbind.data.frame(1:(27*4),stack(suicide_data[1:27]))
names(stacking_sui_data) <- c("Quarters", "Suicide_Rates", "Year")
stacking_sui_data
```
```{r}
#Need to begin adding variable data to the datatable
# Starting with Foreclosure Rates
foreclosure_data <- read.csv(file = "Foreclosure_rates.csv", header = T, sep = ",")
foreclosure_data[, 3] <- 5:112
names(foreclosure_data) <- c("Date", "Foreclosure_Rates", "Quarters")
foreclosure_data
```
```{r}
suicide_data_cut <- stacking_sui_data[5:108,]
suicide_data_cut
working_data_table <- merge.data.frame(suicide_data_cut, foreclosure_data, by = "Quarters")
working_data_table
```
```{r}
#need to add the next variable: property and violent crime rates
crime_rates <- read.csv(file = "Violent_and_nonviolent_crime_rates.csv", header = T, sep = ",")
crime_rates
names(crime_rates) <- c("Violent Crime Rate", "Property Crime Rate")
constant_qtr_rate_v <- crime_rates[, 1]/4
violent_crime_rates <- rep(constant_qtr_rate_v, each = 4)
constant_qtr_rate_p <- crime_rates[, 2]/4
property_crime_rates <- rep(constant_qtr_rate_p, each = 4)
crime_rates_data_table <- cbind.data.frame(violent_crime_rates, property_crime_rates)
crime_rates_data_table[, 3] <- 5:108
names(crime_rates_data_table) <- c("Violent_Crime_Rates", "Property_Crime_Rates", "Quarters")
crime_rates_data_table

```
```{r}
working_data_table_2 <- merge.data.frame(working_data_table, crime_rates_data_table, by = "Quarters")
working_data_table_2
```
```{r}
#next variable that needs to be added is unemployment rate
unemp_data <- read.csv(file = "unemployment_data.csv", header = T, sep = ",")
unempl_data <- unemp_data[24:50, 2:13]
unempl_data

Q1 <- c()
Q2 <- c()
Q3 <- c()
```
```

```
Q4 <- c()
total_Q <- c()
for (i in 1:27) {
  Q1[i] <- (unempl_data[i, 1] + unempl_data[i, 2] + unempl_data[i, 3])/3
  Q2[i] <- (unempl_data[i, 4] + unempl_data[i, 5] + unempl_data[i, 6])/3
  Q3[i] <- (unempl_data[i, 7] + unempl_data[i, 8] + unempl_data[i, 9])/3
  Q4[i] <- (unempl_data[i, 10] + unempl_data[i, 11] + unempl_data[i, 12])/3
}

index_1 <- seq(1, 105, 4)
index_2 <- seq(2, 106, 4)
index_3 <- seq(3, 107, 4)
index_4 <- seq(4, 108, 4)

total_Q[index_1] <- Q1
total_Q[index_2] <- Q2
total_Q[index_3] <- Q3
total_Q[index_4] <- Q4

unemp_data_table <- cbind.data.frame(total_Q[1:104], 5:108)
names(unemp_data_table) <- c("Unemployment_Rate", "Quarters")
unemp_data_table
```

```
```{r}
working_data_table_3 <- merge.data.frame(working_data_table_2, unemp_data_table, by = "Quarters")
working_data_table_3 <- working_data_table_3[, c(1, 2, 5, 6, 7, 8)]
working_data_table_3
```
```{r}
#add the next variable: income level adjusted by CPI
cpi_data <- read.csv(file = "CPI_Data.csv", header = T, sep = ",")
cpi_data <- cpi_data[937:1248, ]

index_cpi <- seq(1, 310, 3)
cpi_trim <- c()

for (j in 1:104) {
   cpi_trim[j] <- ((cpi_data[index_cpi[j], 2] + cpi_data[index_cpi[j] +1, 2] + cpi_data[index_cpi[j] +2, 2]))/3
 }

cpi_data_table <- cbind.data.frame(5:108, cpi_trim)
names(cpi_data_table) <- c("Quarters", "CPI")
cpi_data_table

personal_inc_data <- read.csv(file = "Personal_Income_Billions.csv", header = T, sep = ",")
adj_income_table <- cbind.data.frame(5:108, cpi_trim*personal_inc_data)
names(adj_income_table) <- c("Quarters", "Adjusted_Personal_Income")
adj_income_table
```
```{r}
working_data_table_4 <- cbind.data.frame(working_data_table_3, adj_income_table, by = "Quarters")
working_data_table_4 <- working_data_table_4[, c(1,2,3,4,5,6,8)]
working_data_table_4
```
```{r}
#next factor to add is change in GDP
gdp_data <- read.csv(file = "GDP_Change_Data.csv", header = F, sep = ",")
```
```

```
gdp_data <- gdp_data[2:105]
gdp_data_fixed <- t(gdp_data)
gdp_data_table <- cbind.data.frame(5:108, gdp_data_fixed)
names(gdp_data_table) <- c("Quarters", "GDP_Change")
gdp_data_table
```
```
```{r}
working_data_table_5 <- cbind.data.frame(working_data_table_4, gdp_data_table, by = "Quarters")
working_data_table_5
```
```{r}
#add next variable income inequality
gini_ratio <- read.csv(file = "Gini_Ratios.csv", header = T, sep = ",")
gini_ratio
gini_ratio_expanded <- rep(gini_ratio[, 2], each = 4)
gini_ratio_data_table <- cbind.data.frame(5:108, gini_ratio_expanded)
names(gini_ratio_data_table) <- c("Quarters", "Gini_Ratio")
gini_ratio_data_table
```
```{r}
working_data_table_6 <- cbind.data.frame(working_data_table_5, gini_ratio_data_table, by = "Quarters")
working_data_table_6
```
```{r}
#add next variable educational achievement
educ_ach <- read.csv(file = "Educational_Achievement_Data.csv", header = T, sep = ",")
educ_ach[, 1] <- rev(educ_ach[, 1])
educ_ach[, 2] <- rev(educ_ach[, 2])
educ_ach_expanded <- rep(educ_ach[, 2], each = 4)
educ_ach_data_table <- cbind.data.frame(5:108, educ_ach_expanded)
names(educ_ach_data_table) <- c("Quarters", "Graduation")
educ_ach_data_table
```
```{r}
working_data_table_7 <- cbind.data.frame(working_data_table_6, educ_ach_data_table, by = "Quarters")
```
```{r}
#add next variable fertility rate
fer_rate <- read.csv(file = "Fertility_measure.csv", header = F, sep = ",")
fer_rate <- rev(fer_rate$V1)
fer_rate_data_table <- cbind.data.frame(5:108, fer_rate)
names(fer_rate_data_table) <- c("Quarters", "Fertility_Measure")
```
```{r}
working_data_table_8 <- cbind.data.frame(working_data_table_7, fer_rate_data_table, by = "Quarters")
```
```{r}
#add final variable alcohol consumption rate
consum_data <- read.csv(file = "Consumption_data.csv", header = F, sep = ",")
consum_new <- rev(consum_data$V2)
consum_data_table <- cbind.data.frame(5:108, rep(consum_new/4, each = 4))
names(consum_data_table) <- c("Quarters", "Consumption")
consum_data_table
```
```{r}
finalized_data_table <- cbind.data.frame(working_data_table_8, consum_data_table, by = "Quarters")
finalized_data_table <- finalized_data_table[, -c(8, 11, 14, 17, 20)]
```
```

```
finalized_data_table
```
```{r}
#creating a multiple linear regression model including all variables
library(olsrr)
linearMod <- lm(formula = Suicide_Rates ~ Quarters + Foreclosure_Rates + Violent_Crime_Rates +
Property_Crime_Rates + Unemployment_Rate + Adjusted_Personal_Income + GDP_Change + Gini_Ratio + Graduation
+ Fertility_Measure + Consumption, data = finalized_data_table)
summary(linearMod)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod)
ols_test_normality(linearMod)
```
```{r}
library(car)
```
```{r}
vif(linearMod)
```
```{r}
library(olsrr)
linearMod <- lm(formula = Suicide_Rates ~  Foreclosure_Rates + Violent_Crime_Rates + Property_Crime_Rates +
Unemployment_Rate + Adjusted_Personal_Income + GDP_Change + Gini_Ratio + Graduation + Fertility_Measure +
Consumption, data = finalized_data_table)
summary(linearMod)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod)
ols_test_normality(linearMod)
```

```{r}
#remove Quarters from model
vif(linearMod)
```
```{r}
library(olsrr)
linearMod <- lm(formula = Suicide_Rates ~  Foreclosure_Rates + Violent_Crime_Rates +  Unemployment_Rate +
Adjusted_Personal_Income + GDP_Change + Gini_Ratio + Graduation + Fertility_Measure + Consumption, data =
finalized_data_table)
summary(linearMod)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod)
ols_test_normality(linearMod)
```
```{r}
#remove Property Crime Rate
vif(linearMod)
```
```{r}
library(olsrr)
linearMod <- lm(formula = Suicide_Rates ~  Foreclosure_Rates + Violent_Crime_Rates +  Unemployment_Rate +
Adjusted_Personal_Income + GDP_Change + Gini_Ratio + Fertility_Measure + Consumption, data =
finalized_data_table)
summary(linearMod)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod)
ols_test_normality(linearMod)
```

```{r}
#remove graduation
vif(linearMod)
```
```{r}
library(olsrr)
linearMod <- lm(formula = Suicide_Rates ~  Foreclosure_Rates + Violent_Crime_Rates +  Unemployment_Rate +
GDP_Change + Gini_Ratio + Fertility_Measure + Consumption, data = finalized_data_table)
summary(linearMod)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod)
ols_test_normality(linearMod)
```
```{r}
#remove Adjusted_Personal_Income
vif(linearMod)
```
```{r}
#running stepwise regression
library(tidyverse)
library(caret)
library(leaps
library(MASS)
full.model <- glm(formula = Suicide_Rates ~ Foreclosure_Rates + Violent_Crime_Rates + Unemployment_Rate +
GDP_Change + Gini_Ratio + Fertility_Measure + Consumption, data = finalized_data_table)

step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)

```
```{r}
linearMod2 <- lm(formula = Suicide_Rates ~ Foreclosure_Rates + Violent_Crime_Rates + GDP_Change + Gini_Ratio +
Consumption, data = finalized_data_table)
summary(linearMod2)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod2)
ols_test_normality(linearMod2)
```

```{r}
library(caret)
limited_data_table <- finalized_data_table[, c(2, 3, 4, 8, 10, 16)]
limited_data_table
```
```{r}
# K-fold cross-validation
library(caret)
library(rpart)
train_control <- trainControl(method = "repeatedcv", number = 10, repeats = 100)

model <- train(Suicide_Rates ~ Foreclosure_Rates + Violent_Crime_Rates + GDP_Change + Gini_Ratio + Consumption,
data = limited_data_table, trControl = train_control, method = "lm")

print(model)
```
```{r}
#creating "predictive model"


```

```
suicide_rates_shifted <- finalized_data_table[2:length(finalized_data_table[, 4]), 2]
factors_shifted <- finalized_data_table[1:103,]
predictive_table <- cbind.data.frame(suicide_rates_shifted, factors_shifted)
```
```{r}
#checking results should start at the second quarter in row one, but the rest of the table should remain the same
head(predictive_table)
head(finalized_data_table)
```
```{r}
predictive_table
```


```{r}
#running step-wise model
predictive.model <- glm(formula = suicide_rates_shifted ~ Foreclosure_Rates + Violent_Crime_Rates +
Unemployment_Rate + GDP_Change + Gini_Ratio + Fertility_Measure + Consumption, data = predictive_table)
step.model <- stepAIC(predictive.model, direction = "both",
                      trace = FALSE)
summary(step.model)
```
```{r}
linearMod2 <- lm(formula = suicide_rates_shifted ~ Foreclosure_Rates + Violent_Crime_Rates + Unemployment_Rate
+ Gini_Ratio + Fertility_Measure + Consumption, data = predictive_table)
summary(linearMod2)
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(linearMod2)
ols_test_normality(linearMod2)
```
```{r}
#trimmed predictive table
head(predictive_table)
head(predictive_table[, c(1, 4,5, 7, 11, 15, 17)])
```
```{r}
#k-fold analysis
trimmed_predictive_table <- predictive_table[, c(1, 4,5, 7, 11, 15, 17)]
train_control <- trainControl(method = "repeatedcv", number = 10, repeats = 100)

model <- train(suicide_rates_shifted ~ Foreclosure_Rates + Violent_Crime_Rates + Unemployment_Rate + Gini_Ratio
+ Fertility_Measure + Consumption, data = trimmed_predictive_table, trControl = train_control, method = "lm")

print(model)
```

Appendix B

DATA TABLE EXCERPT FOR REGRESSION MODEL

Table 47. Excerpt of the Table created in RStudio used for the National Level Multiple Linear Regression Models.

	Quarters	Suicide_Rates	Foreclosure_Rates	Violent_Crime_Rates	Property_Crime_Rates	Unemployment_Rate	Adjusted_Personal_Income	GDP_Change	by	Gini_Ratio	by.1	Graduation	by.2	Fertility_Measure	by.3	Consumption	by.4
V2	5	2.946226	3.18	189.550	1285.050	6.600000	671816.2	-1.9	Quarters	0.481	Quarters	78.4	Quarters	40.25241	Quarters	0.5750	Quarters
V3	6	3.170296	3.07	189.550	1285.050	6.833333	683830.8	3.1	Quarters	0.481	Quarters	78.4	Quarters	39.38247	Quarters	0.5750	Quarters
V4	7	3.176245	3.17	189.550	1285.050	6.866667	696139.0	1.9	Quarters	0.481	Quarters	78.4	Quarters	38.07631	Quarters	0.5750	Quarters
V5	8	2.946226	3.30	189.550	1285.050	7.100000	712335.9	1.8	Quarters	0.481	Quarters	78.4	Quarters	37.54501	Quarters	0.5750	Quarters
V6	9	3.083955	3.23	189.425	1225.925	7.366667	733574.4	4.8	Quarters	0.479	Quarters	79.4	Quarters	36.40354	Quarters	0.5750	Quarters
V7	10	3.034549	2.96	189.425	1225.925	7.600000	753396.2	4.5	Quarters	0.479	Quarters	79.4	Quarters	36.53649	Quarters	0.5750	Quarters
V8	11	3.000827	2.88	189.425	1225.925	7.633333	767848.6	3.9	Quarters	0.479	Quarters	79.4	Quarters	36.07921	Quarters	0.5750	Quarters
V9	12	2.862412	2.86	189.425	1225.925	7.366667	783756.3	4.1	Quarters	0.479	Quarters	79.4	Quarters	36.29751	Quarters	0.5750	Quarters
V10	13	2.971108	2.82	186.775	1185.000	7.133333	795292.6	0.8	Quarters	0.480	Quarters	80.2	Quarters	36.03488	Quarters	0.5575	Quarters
V11	14	3.066925	2.58	186.775	1185.000	7.066667	811341.3	2.4	Quarters	0.480	Quarters	80.2	Quarters	36.00317	Quarters	0.5575	Quarters
V12	15	3.121235	2.59	186.775	1185.000	6.800000	820378.2	2.0	Quarters	0.480	Quarters	80.2	Quarters	34.94447	Quarters	0.5575	Quarters
V13	16	2.929988	2.45	186.775	1185.000	6.633333	837429.5	5.4	Quarters	0.480	Quarters	80.2	Quarters	34.77105	Quarters	0.5575	Quarters
V14	17	2.919405	2.47	178.400	1165.050	6.566667	850375.9	4.0	Quarters	0.482	Quarters	80.9	Quarters	35.13413	Quarters	0.5450	Quarters
V15	18	3.134519	2.18	178.400	1165.050	6.200000	871642.0	5.6	Quarters	0.482	Quarters	80.9	Quarters	35.24810	Quarters	0.5450	Quarters
V16	19	3.040791	2.11	178.400	1165.050	6.000000	889117.1	2.4	Quarters	0.482	Quarters	80.9	Quarters	34.86945	Quarters	0.5450	Quarters
V17	20	2.885985	2.11	178.400	1165.050	5.633333	909531.3	4.6	Quarters	0.482	Quarters	80.9	Quarters	36.10374	Quarters	0.5450	Quarters
V18	21	2.981317	2.11	171.125	1147.625	5.466667	931571.5	1.4	Quarters	0.477	Quarters	81.7	Quarters	35.83988	Quarters	0.5375	Quarters
V19	22	3.075685	2.03	171.125	1147.625	5.666667	949758.4	1.4	Quarters	0.477	Quarters	81.7	Quarters	34.11413	Quarters	0.5375	Quarters
V20	23	3.047907	2.15	171.125	1147.625	5.666667	965658.7	3.5	Quarters	0.477	Quarters	81.7	Quarters	32.94360	Quarters	0.5375	Quarters
V21	24	2.820741	2.28	171.125	1147.625	5.566667	981780.5	2.9	Quarters	0.477	Quarters	81.7	Quarters	32.45550	Quarters	0.5375	Quarters
V22	25	2.833028	2.22	159.150	1112.750	5.533333	1008879.5	2.7	Quarters	0.477	Quarters	81.7	Quarters	31.42349	Quarters	0.5400	Quarters
V23	26	3.065281	2.17	159.150	1112.750	5.500000	1039287.4	7.2	Quarters	0.477	Quarters	81.7	Quarters	31.75294	Quarters	0.5400	Quarters
V24	27	3.014381	2.22	159.150	1112.750	5.266667	1056654.2	3.7	Quarters	0.477	Quarters	81.7	Quarters	31.52426	Quarters	0.5400	Quarters
V25	28	2.764031	2.41	159.150	1112.750	5.333333	1078624.2	4.3	Quarters	0.477	Quarters	81.7	Quarters	31.52878	Quarters	0.5400	Quarters
V26	29	2.895995	2.32	152.750	1079.075	5.233333	1105765.1	3.1	Quarters	0.470	Quarters	82.1	Quarters	31.02683	Quarters	0.5375	Quarters
V27	30	2.880684	2.19	152.750	1079.075	5.000000	1123082.1	6.2	Quarters	0.470	Quarters	82.1	Quarters	30.23318	Quarters	0.5375	Quarters
V28	31	2.891887	2.22	152.750	1079.075	4.866667	1145744.5	5.2	Quarters	0.470	Quarters	82.1	Quarters	43.10508	Quarters	0.5375	Quarters
V29	32	2.751475	2.34	152.750	1079.075	4.666667	1172021.9	3.1	Quarters	0.470	Quarters	82.1	Quarters	41.43902	Quarters	0.5375	Quarters

Appendix C

RSTUDIO CODE FOR TWITTER DATA GATHERING

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```{r}
#loading packages
options(java.parameters = "-Xmx8000m")
library(twitteR)
library(devtools)
library(rjson)
library(bit64)
library(httr)
library(ROAuth)
```
```{r}
#obtaining twitter access
twitter_setup <- options("httr_oauth_cache")
options(httr_oauth_cache = TRUE)
setup_twitter_oauth(consumer_key = "xxxxxxxxxxxxxxxxxxx", consumer_secret = "xxxxxxxxxxxxxxxxx", access_token
= "xxxxxxxxxxxxxxxxxxxxxxxxx", access_secret = "xxxxxxxxxxxxxxxxxxxxxxxxx")
```
```{r}
#loading packages
library(maps)
library(ggplot2)
library(ggmap)
library(mapproj)
library(dismo)
library(rtweet)
library(rJava)
library(xlsx)
```
```{r}
#getting access for rtweet
create_token(app = "my_app", consumer_key = "xxxxxxxxxxxxxxxxx", consumer_secret = "xxxxxxxxxxxxxxxxxxxx",
access_token = "xxxxxxxxxxxxxxxxxxx", access_secret = "xxxxxxxxxxxxxxxxxxx" )
```
```{r}
#installing github rtweet
devtools::install_github("mkearney/rtweet")
```
```{r}
#north midwest region
#montana <- lookup_coords(address = "montana")
#north_dakota <- lookup_coords(address = "north dakota")
#south_dakota <- lookup_coords(address = "south dakota")
#minnesota <- lookup_coords(address = "minnesota")
```
```{r}
#south west region
#california <- lookup_coords(address = "california")
#arizona <- lookup_coords(address = "arizona")
#nevada <- lookup_coords(address = "nevada")
#oregon <- lookup_coords(address = "oregon")
```
```{r}
```

```
#south east region
#florida <- lookup_coords(address = "florida")
#alabama <- lookup_coords(address = "alabama")
#georgia <- lookup_coords(address = "georgia")
#south_carolina <- lookup_coords(address = "south carolina")
```
```{r}
#north east region
#maine <- lookup_coords(address = "maine")
#new_hampshire <- lookup_coords(address= "new hampshire")
#vermont <- lookup_coords(address = "vermont")
#massachusetts <- lookup_coords(address = "massachusetts")
```
```{r}
#creating vector of state Geocodes
states <- c("montana", "north dakota", "south dakota", "minnesota", "arizona", "nevada", "oregon", "florida",
"alabama", "georgia", "south carolina", "maine", "new hampshire", "vermont", "massachusetts")
#list of abbreviated state names for file name creation
abbrev_state <- c("MT", "ND", "SD", "MN", "AZ", "NV", "OR", "FL", "AL", "GA", "SC", "ME", "NH", "VT", "MA")
#establishing search phrases
search_phrase1 <- "suicide"
search_phrase2 <- "suicidal"
search_phrase3 <- "Prozac"
search_phrase4 <- "feel depressed"
search_phrase5 <- "feel hopeless"
```
```{r}
#dates of interest
start_date <- "2018-09-11"
end_date <- "2018-09-12"
```
```{r}
#for loops to output for each state
 for (i in 1:length(states)){
   #Setting state geocoordinates
   state = lookup_coords(states[i])
   #setting state abbreviation for file naming purposes
   abbrev = abbrev_state[i]
   #searches for state Phrase 1
   search_phrase1_english <- search_tweets(search_phrase1, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "en")
   search_phrase1_spanish <- search_tweets(search_phrase1, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "es")
   search_phrase1_french <- search_tweets(search_phrase1, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "fr")
   search_phrase1_chinese_simplified <- search_tweets(search_phrase1, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-cn")
   search_phrase1_chinese_traditional <- search_tweets(search_phrase1, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-tw")
   #searches for state Phrase 2
   search_phrase2_english <- search_tweets(search_phrase2, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "en")
   search_phrase2_spanish <- search_tweets(search_phrase2, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "es")
   search_phrase2_french <- search_tweets(search_phrase2, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "fr")
   search_phrase2_chinese_simplified <- search_tweets(search_phrase2, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-cn")
```

```
  search_phrase2_chinese_traditional <- search_tweets(search_phrase2, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-tw")
  #searches for state Phrase 3
  search_phrase3_english <- search_tweets(search_phrase3, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "en")
  search_phrase3_spanish <- search_tweets(search_phrase3, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "es")
  search_phrase3_french <- search_tweets(search_phrase3, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "fr")
  search_phrase3_chinese_simplified <- search_tweets(search_phrase3, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-cn")
  search_phrase3_chinese_traditional <- search_tweets(search_phrase3, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-tw")
  #searches for state Phrase 4
  search_phrase4_english <- search_tweets(search_phrase4, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "en")
  search_phrase4_spanish <- search_tweets(search_phrase4, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "es")
  search_phrase4_french <- search_tweets(search_phrase4, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "fr")
  search_phrase4_chinese_simplified <- search_tweets(search_phrase4, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-cn")
  search_phrase4_chinese_traditional <- search_tweets(search_phrase4, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-tw")
  #searches for state Phrase 5
  search_phrase5_english <- search_tweets(search_phrase5, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "en")
  search_phrase5_spanish <- search_tweets(search_phrase5, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "es")
  search_phrase5_french <- search_tweets(search_phrase5, n = 100000, geocode = state, retryonratelimit = TRUE,
since = start_date, until = end_date, lang = "fr")
  search_phrase5_chinese_simplified <- search_tweets(search_phrase5, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-cn")
  search_phrase5_chinese_traditional <- search_tweets(search_phrase5, n = 100000, geocode = state,
retryonratelimit = TRUE, since = start_date, until = end_date, lang = "zh-tw")
  #Setting pathway
  pathway <- paste(c("C:/Users/Derek/Desktop/Twitterdata/", abbrev, as.character(start_date),
as.character(end_date), ".xlsx"), collapse="")
  #creating new excel workbook
  write.xlsx2(search_phrase1_english, file = pathway, sheetName = paste(c(abbrev, "EngPhrase1"), collapse=""),
append = FALSE)
  #writing search results into new sheets in the workbook for each phrase and language
  write.xlsx2(search_phrase1_spanish, file = pathway, sheetName = paste(c(abbrev, "SpnPhrase1"), collapse = ""),
append = TRUE)
  write.xlsx2(search_phrase1_french, file = pathway, sheetName = paste(c(abbrev, "FrPhrase1"), collapse = ""),
append = TRUE)
  write.xlsx2(search_phrase1_chinese_simplified, file = pathway, sheetName = paste(c(abbrev, "CHsPhrase1"),
collapse = ""), append = TRUE)
  write.xlsx2(search_phrase1_chinese_traditional, file = pathway, sheetName = paste(c(abbrev, "CHTPhrase1"),
collapse = ""), append = TRUE)
  write.xlsx2(search_phrase2_english, file = pathway, sheetName = paste(c(abbrev, "EngPhrase2"), collapse = ""),
append = TRUE)
  write.xlsx2(search_phrase2_spanish, file = pathway, sheetName = paste(c(abbrev, "SpnPhrase2"), collapse = ""),
append = TRUE)
  write.xlsx2(search_phrase2_french, file = pathway, sheetName = paste(c(abbrev, "FrPhrase2"), collapse = ""),
append = TRUE)
  write.xlsx2(search_phrase2_chinese_simplified, file = pathway, sheetName = paste(c(abbrev, "CHsPhrase2"),
collapse = ""), append = TRUE)
```

```
   write.xlsx2(search_phrase2_chinese_traditional, file = pathway, sheetName = paste(c(abbrev, "CHTPhrase2"),
collapse = ""), append = TRUE)
   write.xlsx2(search_phrase3_english, file = pathway, sheetName = paste(c(abbrev, "EngPhrase3"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase3_spanish, file = pathway, sheetName = paste(c(abbrev, "SpnPhrase3"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase3_french, file = pathway, sheetName = paste(c(abbrev, "FrPhrase3"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase3_chinese_simplified, file = pathway, sheetName = paste(c(abbrev, "CHsPhrase3"),
collapse = ""), append = TRUE)
   write.xlsx2(search_phrase3_chinese_traditional, file = pathway, sheetName = paste(c(abbrev, "CHTPhrase3"),
collapse = ""), append = TRUE)
   write.xlsx2(search_phrase4_english, file = pathway, sheetName = paste(c(abbrev, "EngPhrase4"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase4_spanish, file = pathway, sheetName = paste(c(abbrev, "SpnPhrase4"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase4_french, file = pathway, sheetName = paste(c(abbrev, "FrPhrase4"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase4_chinese_simplified, file = pathway, sheetName = paste(c(abbrev, "CHSPhrase4"),
collapse = ""), append = TRUE)
   write.xlsx2(search_phrase4_chinese_traditional, file = pathway, sheetName = paste(c(abbrev, "CHTPhrase4"),
collapse = ""), append = TRUE)
   write.xlsx2(search_phrase5_english, file = pathway, sheetName = paste(c(abbrev, "EngPhrase5"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase5_spanish, file = pathway, sheetName = paste(c(abbrev, "SpnPhrase5"), collapse = ""),
append = TRUE)
   write.xlsx2(search_phrase5_french, file = pathway, sheetName = paste(c(abbrev, "FrPhrase5"), collapse= ""),
append = TRUE)
   write.xlsx2(search_phrase5_chinese_simplified, file = pathway, sheetName = paste(c(abbrev, "CHSPhrase5"),
collapse = ""), append = TRUE)
   write.xlsx2(search_phrase5_chinese_traditional, file = pathway, sheetName = paste(c(abbrev, "CHTPhrase5"),
collapse = ""), append = TRUE)
   i = i + 1
   #setting system to sleep after each loop for five minutes to work around 15 minute query limits
   Sys.sleep(300)
  }
```

Appendix D

SAS EXAMPLE CODE FOR CALCULATING SUICIDE RATE

```
/* 2 in line 107 corresponds to suicide */
/* lines 65-66 coresond to month of death */
data sui_2016;
        infile "C:\Users\liblabs-user\Desktop\VS16MORT.DUSMCPUB";
        input Date 65-66 Cause_Death $ 107;
run;
proc print data = sui_2016 (obs = 6);
run;
proc sql;
        Create Table Clean_16 AS
        Select Date, Cause_Death
        From sui_2016
        where Cause_Death contains "2"
        order by Date;
        ;
Quit;
proc print data = Clean_16 (obs = 6);
run;
data Clean_16_Counts;
        set Clean_16;
        by Date;
        if first.Date then do;
        num_sui = 0;
        end;
        num_sui + 1;
        if last.Date then output;
keep Date num_sui;
run;
proc print data = Clean_16_Counts;
run;
```

Appendix E

M CODE FOR PROCESSING TWITTER DATA

```
let
    Source = Folder.Files("C:\Users\liblabs-user\Desktop\AlabamaData"),
    #"Removed Other Columns" = Table.SelectColumns(Source,{"Content"}),
    #"Added Custom" = Table.AddColumn(#"Removed Other Columns", "GetExcelData", each
Excel.Workbook([Content])),
    #"Removed Columns" = Table.RemoveColumns(#"Added Custom",{"Content"}),
    #"Expanded GetExcelData" = Table.ExpandTableColumn(#"Removed Columns",
"GetExcelData", {"Name", "Data", "Item", "Kind", "Hidden"}, {"GetExcelData.Name",
"GetExcelData.Data", "GetExcelData.Item", "GetExcelData.Kind", "GetExcelData.Hidden"}),
    #"Filtered Rows" = Table.SelectRows(#"Expanded GetExcelData", each ([GetExcelData.Item]
= "ALEngPhrase1")),
    #"Added Custom1" = Table.AddColumn(#"Filtered Rows", "PromoteHeaders", each
Table.PromoteHeaders([GetExcelData.Data])),
    #"Removed Other Columns1" = Table.SelectColumns(#"Added
Custom1",{"PromoteHeaders"}),
    #"Expanded PromoteHeaders" = Table.ExpandTableColumn(#"Removed Other Columns1",
"PromoteHeaders", {"created_at", "text", "is_retweet", "favorite_count", "retweet_count",
"retweet_favorite_count", "retweet_retweet_count", "retweet_followers_count"}, {"created_at",
"text", "is_retweet", "favorite_count", "retweet_count", "retweet_favorite_count",
"retweet_retweet_count", "retweet_followers_count"}),
    #"Changed Type" = Table.TransformColumnTypes(#"Expanded
PromoteHeaders",{{"created_at", type date}}),
    #"Removed Columns1" = Table.RemoveColumns(#"Changed Type",{"is_retweet",
"retweet_favorite_count", "retweet_retweet_count", "retweet_followers_count"}),
    #"Added Custom2" = Table.AddColumn(#"Removed Columns1", "Exclusionary (Prevention)",
each Text.Contains([text], "Prevention")),
    #"Added Custom3" = Table.AddColumn(#"Added Custom2", "Exclusionary (prevention).1",
each Text.Contains([text], "prevention")),
    #"Added Custom4" = Table.AddColumn(#"Added Custom3", "Exclusionary (Education)", each
Text.Contains([text], "Education")),
    #"Added Custom5" = Table.AddColumn(#"Added Custom4", "Exclusionary (education).1",
each Text.Contains([text], "education")),
    #"Added Custom6" = Table.AddColumn(#"Added Custom5", "Exclusionary (hotline)", each
Text.Contains([text], "hotline")),
    #"Added Custom7" = Table.AddColumn(#"Added Custom6", "Exclusionary (Intervention)",
each Text.Contains([text], "Intervention")),
    #"Added Custom8" = Table.AddColumn(#"Added Custom7", "Exclusionary (intervention).1",
each Text.Contains([text], "intervention")),
    #"Filtered Rows1" = Table.SelectRows(#"Added Custom8", each ([#"Exclusionary
(Prevention)"] = false) and ([#"Exclusionary (prevention).1"] = false) and ([#"Exclusionary
(Education)"] = false) and ([#"Exclusionary (education).1"] = false) and ([#"Exclusionary (hotline)"]
= false)),
    #"Filtered Rows2" = Table.SelectRows(#"Filtered Rows1", each ([#"Exclusionary
(Intervention)"] = false) and ([#"Exclusionary (intervention).1"] = false)),
    #"Removed Columns2" = Table.RemoveColumns(#"Filtered Rows2",{"Exclusionary
(Prevention)", "Exclusionary (prevention).1", "Exclusionary (Education)", "Exclusionary
(education).1", "Exclusionary (hotline)", "Exclusionary (Intervention)", "Exclusionary
(intervention).1"}),
    #"Added Custom9" = Table.AddColumn(#"Removed Columns2", "State", each "Alabama")
in
    #"Added Custom9"
```

Appendix F

EXAMPLE EXCEL PIVOT TABLE FOR TWITTER DATA

Table 48. Example Excel Pivot Table for Alabama's Twitter Data for EP1.

State	created_at	Count of text	Sum of favorite_count	Sum of retweet_count
Alabama	12-Aug	249	365	32680
	13-Aug	227	505	103647
	14-Aug	212	208	20409
	15-Aug	316	1599	39631
	16-Aug	314	208	64623
	17-Aug	271	241	81268
	18-Aug	164	233	89844
	19-Aug	169	194	5635
	20-Aug	243	236	13548
	21-Aug	188	335	74178
	22-Aug	271	1416	207498
	23-Aug	337	553	42425
	24-Aug	389	390	122001
	25-Aug	205	209	41369
	26-Aug	299	423	43464
	27-Aug	829	3912	638845
	28-Aug	1030	1078	521192
	29-Aug	558	351	146680
	30-Aug	442	1381	142024
	31-Aug	474	390	227673
	1-Sep	268	167	72650
	2-Sep	329	104	198517
	3-Sep	247	540	79103
	4-Sep	324	320	74761
	5-Sep	274	202	90144
	6-Sep	441	827	104351
	7-Sep	602	1158	244199
	8-Sep	368	364	148148
	9-Sep	329	433	100409
	10-Sep	451	1325	12508
	11-Sep	522	696	164060