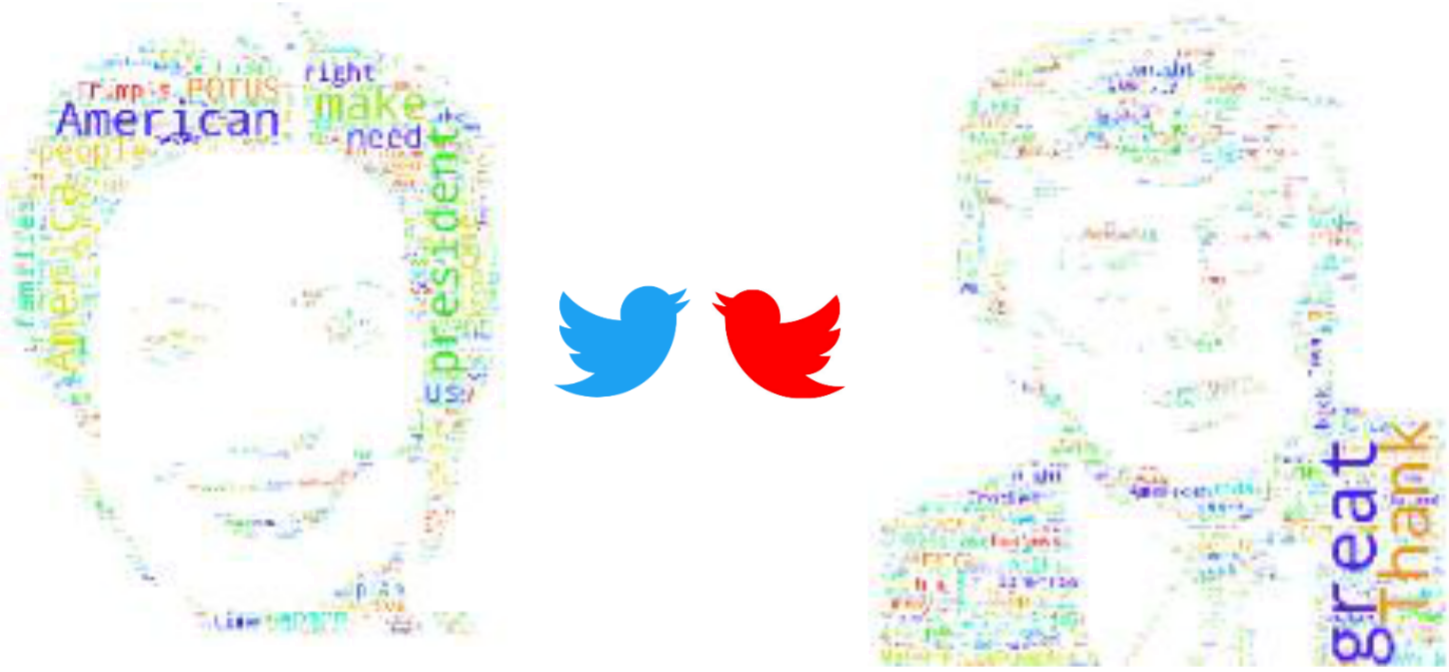
MS5103 Business Analytics Project

An investigation into whether Twitter can be an accurate Predictor of Elections versus Traditional Opinion Polls



Kevin Derrane: 12409118
Jamie O'Halloran: 12458152
Robert O'Sullivan: 12459088

Supervisor: Dr Michael Lang

NUI Galway
OÉ Gaillimh

Declaration of Originality

Project Details

Module Code: **MS5103** Assignment Title: **M.Sc. Business Analytics - Major Project**

Group Members: (please use BLOCK CAPITALS)

Student ID	Student Name	Contact Details (Email, Telephone)
12409118	Kevin Derrane	k.derrane1@nuigalway.ie 0857340313
12458152	Jamie O'Halloran	j.ohalloran12@nuigalway.ie 0894585652
12459088	Robert O'Sullivan	r.osullivan9@nuigalway.ie 0858139050

I/We hereby declare that this project is my/our own original work. I/We have read the University *Code of Practice for Dealing with Plagiarism*[*] and am/are aware that the possible penalties for plagiarism include expulsion from the University.

Signature **Date**

[*] http://www.nuigalway.ie/plagiarism

Executive Summary

The 2016 US Presidential election was very different to previous campaigns. Prior to this election, traditional campaigning methods were used to gather and predict votes. In 2016, Nominee Hillary Clinton was forecasted to be triumphant with all polls pointing towards a comfortable victory over opponent Donald Trump. In the end, the polls proved to be a poor election predictor with Trump being triumphant. The polls got it so wrong, but how? Opinion polls can lead to sampling bias and lead people to give false statements and potentially lie. Social media eliminates these false predictors allowing people to express their true opinion publicly while also giving them the chance to remain anonymous. Twitter was the most used social media platform by both candidates, with Donald Trump utilizing it the most. It is the ultimate goldmine of all things opinionated. This project focuses on comparing both nominees Twitter accounts and public tweets during the final Election Day vs traditional public opinion polls. The results of this project showed that there is moderately positive correlation between the popularity of each candidate on Twitter and the number of times they were mentioned is user's tweets on the final Election Day.

Table of Contents

1. Introduction

Opinion polls have been around for almost 200 years. The first example can be traced back to the 1824 US Presidential Election. A straw poll was conducted by the Harrisburg Pennsylvanian newspaper which showed nominee Andrew Jackson leading John Quincy Adams by 335 votes to 169. Andrew Jackson went on to win the most popular vote in that state and the whole country. Opinion polls gradually became more popular, leading them to become a worldwide phenomenon that is used to forecast election results today. Opinion polls are generally a good predictor for predicting the outcome of certain events They have been suitable predictors for the last 21 US elections where they have only been wrong three times in 1948, 1976 and most recently 2016.

The presence of Social Media in our everyday lives is continuing to grow. So, what is social media? It is primarily an internet based tool for sharing and discussing information. It is part of our daily lives with a need to Tweet about what we had for lunch, what we trained in the gym or who we think is going to win next week's Premier League match. Social media presents us with the opportunity to give our opinion on anything in the world that we like, express our views freely with millions of others online. Users are able to view your thoughts, comment on it or even exchange views and opinions with you. However, one of the most appealing aspects of social media is its ability to allow users to express their opinions anonymously if they wish to do so. It allows users to say what they wish online without the fear of any repercussions. Think about how many opinions you generate in your mind in one day. How many times have you wanted to say something in your daily life and you couldn’t? Any opinion a person expresses publically can be contradicted. Expressing yourself through the medium of social media enables you to be yourself. Nowadays, there is numerous amounts of platforms for our voice to be heard whether it be on Television, radio, Facebook, Twitter, Snapchat or the newspaper.

Twitter is by far the most popular and dominant social media platform for making your voice heard. Twitter allows you to express an array of emotions and opinions in 140 characters or less. With over 317 million active monthly users, 500 million tweets per day and over 100 million daily active users, it is no doubt that Twitter is and will continue to be the largest, dominate force in opinionated blogging to date. Accessibility and ease of use of Twitter couldn't be more straight forward. With the click of a button users can log in from anywhere in the world and Tweet about whatever they like. With over 80% of Twitter users accessing the platform from their mobile, they are able to follow
their favourite football team, presidential candidate along with their views as they walk along the street, as they have a coffee or as they're on the bus home from work. Access is basically anywhere and anytime. The options are limitless.

The 2012 US Election was the first US Election where there was a formidable social media presence. Social media began to take off in 2010 with millions of people around the world fascinated by the platforms Facebook and Twitter. Obama was not only the first African American president to be elected but he was the first nominee to identify and use social media as a major campaign strategy. It's easy to forget how rare social media was ten years ago. It is instinct to wake up in the morning now, roll over and check your iPhone for any notifications you may have. Ten years ago in 2007 when Obama announced his candidacy, Twitter had only just started and there wasn't even an iPhone yet. Obama managed to obtain 23 million followers in the five years that led up to his victorious 2012 US Election. Obama had 23 million followers while Romney had 1 million. Romney completely neglected the use of social media which may have been a contributing factor to his narrow loss. Obama is the first and perfect example that an effective social media campaign is based on the psychology and interaction of Tweets on a social media platform. He set the tone for future presidential elections as we will see in the 2016 election in which both candidates aggressively and actively used social media.

Below is a graph showing the increase in the number of monthly active users on Twitter from Q4 2010 to Q4 2016. The number of monthly active users on Twitter increased by 150 million from the beginning of the 2012 US Election to the 2016 US Election.

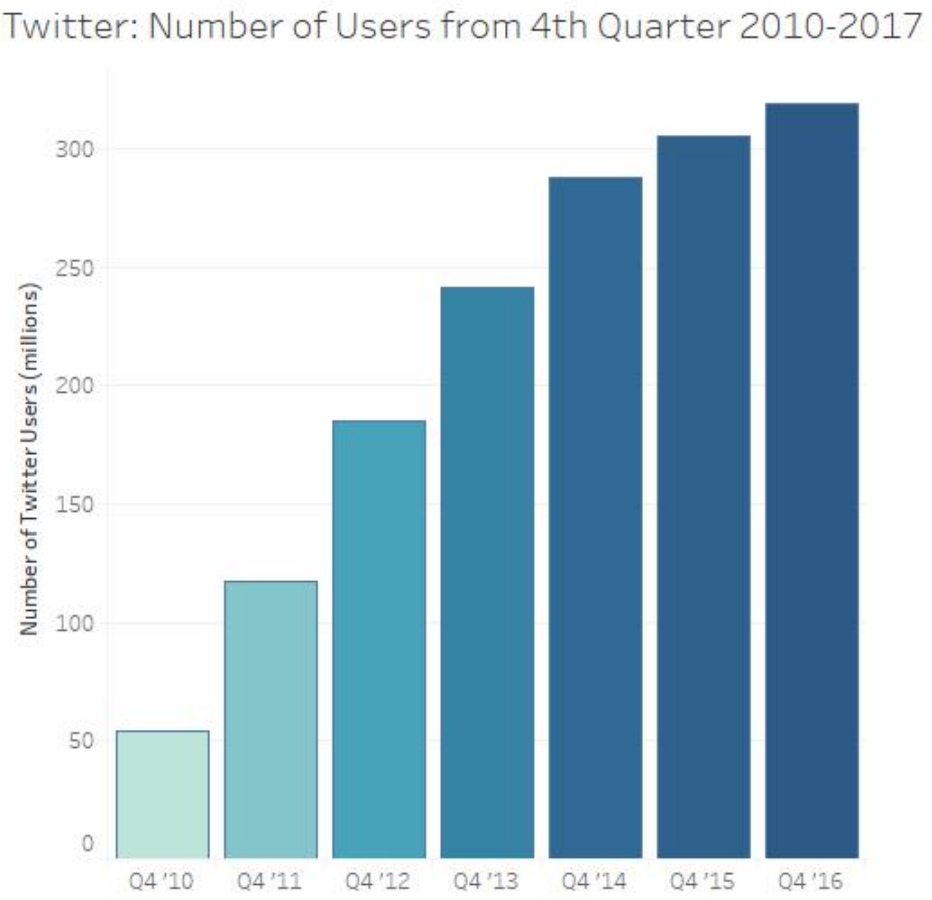


Figure 1.1.1

In 2016, we were presented with the social media battle between Donald Trump and Hilary Clinton with both candidates striving to be elected the next President of the United States of America. During the final day of the election, Trump had amassed over 13.5 million followers with Hillary close behind having just 10.5 million. A timeline showing their follower account can be seen below in diagram 1.0. It is a possibility that the difference in the 3 million followers ultimately had an impact on who the people decided would be the next President of the United States. Twitter allowed Trump and Hilary to express their opinions on a wide range of topics such as education, Terrorism and immigration. Users of Twitter view this information via their account and can interact with candidates' by retweeting, commenting or favouriting their Tweets. It allowed both candidates to trade blows with each other from wherever and whenever they wanted to. Before this, debates were the main source of communication and discussion between candidates. Twitter allows the people of America to follow their desired candidate in real-time, viewing their lives and views as they express themselves instantly. It also allows the candidates to engage in argumentative conversations while users view their remarks in real-time enabling them to retweet, comment or favourite.Trumps blunt commentary and headline grabbing remarks led him to put Twitter at the forefront of his campaign.

Below is a graph showing the number of Twitter followers each candidate had during the final month of the election. This graph shows the increasing use of Twitter by both candidates during the 2016 election.

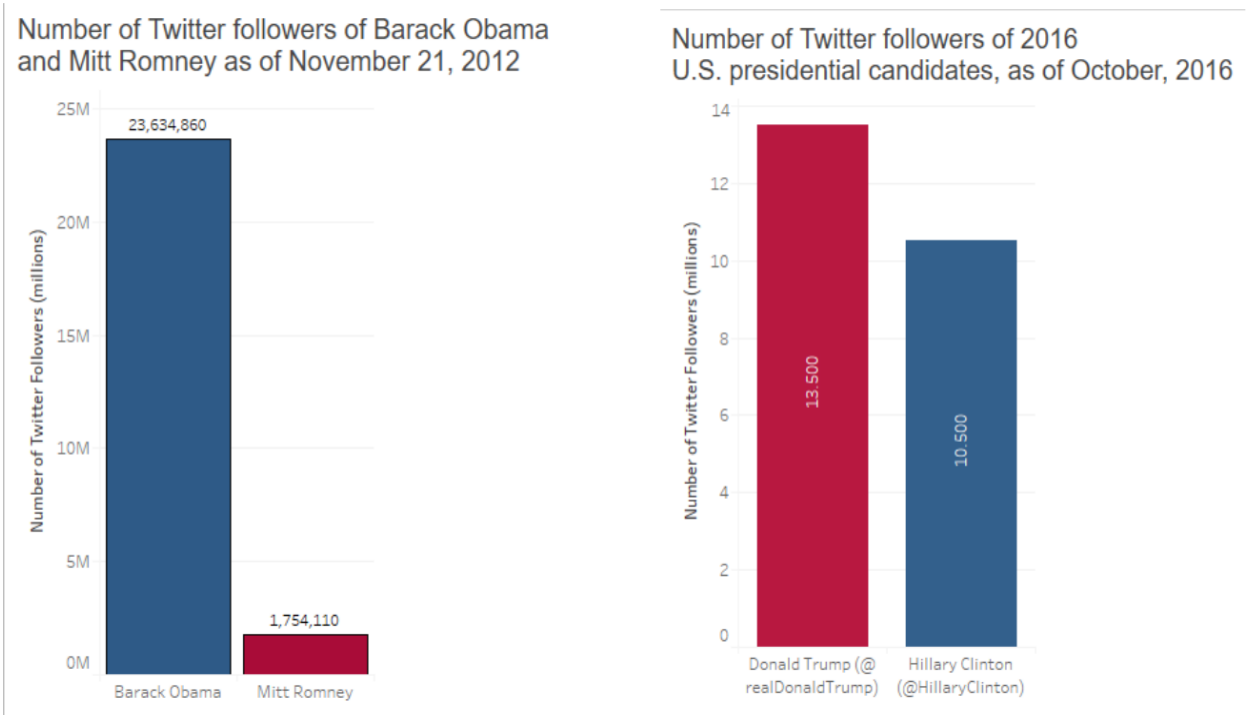


Figure 2.1.2

Our goals for this project can be seen in section 2 below. Here we define our research questions that we aim to answer and the goals we wish to achieve.

2. Problem outline & objectives

The objective of this project is to determine and analyze whether Twitter can be a better predictor for elections than traditional opinion polls. It focuses on comparing both nominees Twitter accounts and public tweets during the final Election Day vs traditional public opinion polls. We want to see if there is a correlation between a nominee's social media presence on Twitter and their result in an election. In order for this analysis to be carried out, three unique datasets had to be analysed using a number of tools and techniques.

Furthermore, an additional number of research questions to be answered were defined as follows:

1. Who was predicted to win the election in the traditional opinion polls?
2. Whose Twitter account is more popular?
3. On Election Day, who was predicted as the next US president according to their mentions on Twitter?
4. Does the sentiment of tweets towards the candidates affect their popularity on Twitter?

Additionally once we defined these research questions, our aims were as follows:

1. To create or find three datasets to help answer our above research questions.
2. To perform analysis on these datasets to see:
   a) Who was predicted to win according to the polls?
   b) Who was the more popular Twitter account?
   c) Who was the more popular candidate on Twitter on Election Day?
3. To perform sentiment analysis on a dataset.
4. To prepare a document stating the results of our analysis along with our conclusive results and findings.

To help us answer our research questions and achieve our aims, good project management skills were crucial and three datasets were to be worked on. The description for our project management techniques and the datasets can be seen in relevant sections, 3 and 4 below.

3. Project Management

Project management is one of the most overlooked features when starting work on a project. It is the organisational and strategic execution of everything that needs to be completed in a project in order to reach the ultimate goal in a specified time frame.

Every project is completely different to the next. Every project entails endless amounts of hours which ultimately leads you to the finished, final product. There isn't one perfect system for each and every project. Every project is unique and requires its own unique project management system in order to develop and materialise. There are a number of project management methods such as Scrum, Agile, Lean, Kanban, Six sigma and PRINCE2. These methods aid development and provide a wide range of benefits, each giving their own distinct contribution.

For our project we decided to go with Google docs and a combination of lean thinking and Kanban boards through Trello to manage our project.

3.1 Google Docs

Google docs was used as a sort of scrap book. We used Google Docs to add links, create headings, discuss sections of the project and post any relevant websites/information that we thought may be useful. Google docs was a key feature throughout our project. It was also a useful tool for communicating throughout the duration of our project as we could access it anywhere, anytime as long as we had a laptop and internet connection.

3.2 Lean thinking

Lean is a project management technique that ensures every part of your project is shipped with the same quality. Lean aided us in breaking up our project into minute pieces so that we can work on them individually. It also ensures a workflow for each task is laid out which is extremely important.

Lean's stages and flexibility has allowed us to manage and align our project so that each part of our project is completed, and completed well. Focusing and working on one thing at a time in a project can have its downfalls. Lean enabled us to mix and match what we worked on through various stages of our project. It's no strict deadline characteristics has allowed us to focus on various tasks in various phases of our workflow at the same time. It has allowed us to develop an overall view and build a system tailored to our team.

Consistency, quality and time management is the key to any successful project. Lean has enabled us to build a system tailored to our needs, allowing us to work on our

project week in week out focusing at the task at hand. Looking at your project from an oversight view perspective is essential in viewing tasks needed to be completed, tasks completed, and issues. This tool defined our projects structure allowing us to guide our project management for months on end.

3.3 Kanban boards

The worst thing you can do with a project is to just jump ahead and try get as much as possible completed in a little space of time. Kanban allowed us to align tasks to various stages. It benefited us greatly and allowed us team to adjust our workload accordingly, along with keeping time management in check. There were many stages during our project where tasks are half complete or a few lines of code needed to be finalised. Kanban allowed us to set these problems aside while allowing us to progress with our workflow, enabling efficient and effective team progress while keeping goals and objectives in check. Kanban incorporate the most important aspect of working on a project – keeping our workflow in check.

Kanban is always about putting your teams needs first. It's not about scheduled meetings or sticking to a pre-defined schedule. It's about focusing only at the task at hand. Focusing what is right in front of you is what matters most. It's no set times, no assigned roles and zen-like focus has led us to managing our project extremely efficiently. It has aided us greatly throughout our project

These are the four pillars of the Kanban philosophy:

1. Cards: translates to 'visual card'. Each task has a card that includes all relevant information about it, it ensures nothing is left out so that it is fully completed.
2. Cap on work in progress: This prevents our team from over-committing. Over-committing is a big issue and can lead teams to be side tracked and led astray. The cap on work in progress sets a limit on how many cards are in play at once.
3. Continuous flow: This enables movement down the list of backlogs in order of importance. The most important thing is that something is always being worked on and this incorporates that aspect.
4. Constant improvement: Analyse the flow to determine how efficiently you're working, and always strive to improve it.

A Kanban board is a work and workflow visualization tool that enables you to optimize the flow of your work. It allowed us to visually implement the lean development strategy. For implementing this we used Trello.

3.4 Trello

Trello is a project management tool that was perfectly designed for use throughout our project. It gave us a meaningful insight into how our team was working and allowed us to manage tasks accordingly to ensure nothing was forgotten about or ignored.
Dedicating a board to different sets of activities proved to be vital in managing our project as a whole. Within the board you can create all the lists you need, view them in real-time and tick them off as they are completed.

The Trello Kanban board allowed us to easily configure a specific series of lists that indicated what tasks needed to be worked on. This allowed us to name the lists accordingly taking into accord priority areas and time sensitives. It allowed us to see who was working on what, the work to do, left to do and work done. It also allowed us to see if any backlog was going to build up or if a member of the team was not pulling his weight. It gave us flexibility with our work. Without the Trello Kanban board our project would have been very unorganised. It is the perfect tool to ensure that your project is completed on time and with all the appropriate work done. A screenshot of our Trello board can be seen below in Figure 3.4.1 below. Link to our Trello board is here:

Figure 3.4.1

4. Description of Datasets

One of the main objectives of our project was to obtain the correct datasets to analyse to yield the most accurate results. At the start of our project, we intended to create our own dataset using Twitter's API. This would allow us to have full control over our datasets. However, early on when we began scraping tweets using the twitteR package with R programming and through python, we ran into difficulties with Twitters 15 day tweet history limitation. More about this can be seen below in the appendences section.

We were still determined to complete our project along the same lines so we turned to Kaggle which is a public online dataset hosting website. Here we discovered three datasets which would help us answer our research questions and achieve our aims. The three datasets are as follows: Collection of Presidential Election Polls from 2015-2016; Tweets from Hillary Clinton and Donald Trump and Final day election tweets. These datasets allowed us keep our original project aims and also expand on them. The three datasets allowed us to get the information we required in a number of different ways and also allowed us to gain a deeper and more accurate insight into the people's opinions on the result of the US election. These datasets and their field types are described in more detail below to help get a better understanding of the data we are working with.

4.1 Collection of Presidential Election Polls from 2015-2016

Our first dataset is a Collection of Presidential Election Polls from 2015-2016. This dataset is a collection of state and national polls conducted from November 2015 to November 2016 on the 2016 presidential election. It contains polls on all four candidates: Trump, Hillary, Johnson and McMullin. This dataset is in the CSV format. Data is presented on the raw and weighted poll results by state, date, pollster, and pollster ratings. These poll results were aggregated from HuffPost Pollster, RealClearPolitics, polling firms and news reports. It has a size of 2.95 MB. It original had 27 different fields types, but for our analysis we cleaned this data, more on this in section 5 below.

These fields include some of the following;

- **Pollster:** which is a person or company who conducts or analyses opinion polls;
- **Grade:** of the poll which is how good the poll is;
- **Samplesize:** which is the number of people polled;
- **Cycle:** indicates the year of the election;
- **Branch:** indicates what role the candidate is running for;
- **Type:** indicates the type of election, in this case polls-plus.
- **Matchup:** indicates what candidates competed against each other.
- **Forecastdate:** This is the date the poll was forecasted.
- **State:** indicates what state the poll was forecasted in;
- **Startdate:** indicates the start date of the poll;

- **Enddate:** indicates the end date of the poll;
- **Grade:** indicates the usefulness, creativity and accuracy of the poll;
- **Population:** indicates how many people voted in the poll;
- rawpoll_johnson, rawpoll_mcmullin, adjpoll_johnson, adjpoll_mcmullin, rawpoll_clinton, rawpoll_trump, adjpoll_trump and adjpoll_clinton; Further explained below.
- **URL:** This is the link to the specific poll.
- **poll_id:** This is the unique ID number assigned to the poll.
- **question_id:** This is the unique question assigned to the poll.
- **Createddate:** This outlines what date the poll was created.
- **Timestamp;** This outlines the exact time the poll was created.

The raw poll is the "direct" answers of the people. Interviews and opinion polls lead people to give biased answers while not saying what is truly on their mind. They give answers that will have the least impact on their lives. As a result of this, they apply weighting and adjustment procedures to the raw data to eliminate the bias problem, resulting in the adjacent (adj) poll data. Its value reflects calibration for historical statistical bias of the individual polls. The adj poll can be considered the “true” answers of the people. The adj poll can be likened to how a person would answer a question on social media. It takes away the bias problem.

4.2 Over 3000 tweets from Hillary Clinton and Donald Trump’s Twitter accounts.

Our second dataset is based on Tweets from Hillary Clinton and Donald Trump. This dataset is in the CSV format. It is 4.92mb in size. This dataset provides over 3000 recent tweets from their Twitter accounts. These tweets range from 5th of January 2016 to the 28th of September 2016. This dataset originally had 28 different field types.

These fields range from;

- id indicates the users unique identifier.
- handle indicates the Twitter username.
- text indicates the body of data that the user has typed.
- is_retweet indicates a boolean value if a tweet has been re-tweeted or not.
- original_author indicates the original author handle if a tweet has been retweeted.
- time indicates the timestamp that a tweet has been generated.
- in_reply_to_screen_name indicates the username that is being replied to.
- in_reply_to_status_id indicates the unique id of the tweet that is being replied to.
- in_reply_to_user_id indicates the unique identifier of the user that is being replied to.
- is_quote_status indicates a boolean value that represents if text contains quotation.
- lang indicates the language of the text.
- retweet_count indicates the number of re-tweets the tweet has achieved.

- favorite_count indicates the number of favorites the tweet has achieved.
- longitude indicates the geographic tag for longitude.
- latitude indicates the geographic tag for latitude.
- place_id indicates the location id of the author.
- place_full_name indicates the name of the location of the author.
- place_name indicates the name of the location of the author.
- place_type indicates the type of area of the location of the author.
- place_country_code indicates the country code of the author location.
- place_country indicates the country name of the author location.
- place_contained_within indicates metadata of location - null values.
- place_attributes indicates metadata of location - null values.
- place_bounding_box indicates a bounding box of coordinates i.e the location.
- source_url indicates the unique URL of the data.
- truncated indicates a boolean value if retweet has been edited.
- entities indicates metadata.
- extended_entities indicates metadata.

4.3 Tweets during the final day of the US election.

Our third dataset is also based on tweets. This time it is based on tweets during the final day of the US election. This dataset is also in CSV format. This dataset is more in line with the original aim of our project when we began trying to create a dataset. It contains just under 400k tweets, about 6% of the 6.5 million originally posted during the final Election Day. It is 206mb in size. This dataset has 35 different field types.

These fields include the following;

- **retweet_count:** which is the number of times this tweet has been retweeted;
- **favorite_count:** which is the number of times the tweet has been favorited;
- **text:** text of the tweet
- **created_at:** date and time of the tweet
- **geo:** a JSON object containing coordinates [latitude, longitude] and a `type'
- **lang:** Twitter's guess as to the language of the tweet
- **place:** a Place object from the Twitter API
- **coordinates:** a JSON object containing coordinates [longitude, latitude] and a `type'; note that coordinates are reversed from the geo field
- **user.favourites_count:** number of tweets the user has favorited
- **user.statuses_count:** number of statuses the user has posted
- **user.description:** the text of the user's profile description
- **user.location:** text of the user's profile location
- **user.id:** unique id for the user
- **user.created_at:** when the user created their account
- **user.verified:** bool; is user verified?
- **user.following:** bool; am I (Ed King) following this user?
- **user.url:** the URL that the user listed in their profile

5. Analysis of Datasets

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. We analysed all three datasets using R studio and its various techniques. These techniques are further discussed below in the Appendices section. Each dataset was analysed to help us answer our research questions and aims discussed above in section 2. Correctly analyzing the datasets was key in getting the results we wanted. Prior to analyzing the datasets they needed to be cleaned up.

5.1 Cleaning the Datasets.

Data cleaning, or data preparation is an essential part of statistical analysis. Prior to loading the data in R studio, we loaded the files in Excel and removed any unnecessary fields.

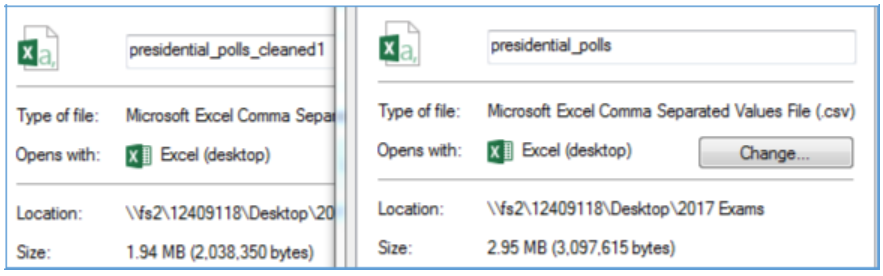
Our first dataset original had 27 different fields, but for our analysis we cleaned this data and removed candidates that that were not Trump or Hillary.

We removed the following fields;

- rawpoll_johnson
- rawpoll_mcmullin
- adjpoll_johnson
- adjpoll_mcmullin
- url
- multiversions

These fields were removed as they either contained NA values or were not used during our analysis. We used Microsoft Excel to do this. Cleaning the dataset allowed us to process the data faster and with more accuracy. It reduced the size and scope of the CSV file. It allowed us to remove any waste we did not need. It went from an original size of 2.95 MB to 1.94 MB, reducing its size by over 30%.

Showing difference in size after cleaning data using Microsoft Excel.



Our second dataset original had over 28 different field types. For our analysis, we removed any fields that were not used. We removed 12 fields in total including;

- latitude
- longitude
- place_id
- source_url.
- place_full_name
- place_name
- place_type
- place_country_code
- place_country
- place_contained_within
- place_attributes
- place_bounding_box

The original size decreased from 4.92mb to 4.60mb.These fields were removed as they either contained NA values or were not used during our analysis.

Our third and final dataset is the largest of them all. Cleaning this was very important to reduce the time for analysis and achieve the most accurate results. It originally had 35 different field variables. We managed to cut this down to 27 in Excel removing the following variables:

- user.listed_count
- user.time_zone
- user.profile_image_url
- user.profile_background_color
- user.geo_enabled
- geo
- place
- coordinates

This helped us get our file from an original size of over 206mb down to 147mb. These fields were removed as they either contained NA values or were not used during our analysis.

For both twitter datasets we ran the below command to clean up the text and remove any punctuation, control character and digits. They were not needed for the purpose of this analysis.

```
## function to clean text. convert to lower case, removing speacilcharacters/numbers/control characters
cleanText <- function(x){

  x <- tolower(x)
  x <- gsub("[[:punct:]]","",x) ## remove punctuation
  x <- gsub("[[:punct:]]","",x) ## remove control characters
  x <- gsub("\\d+","",x) ## remove digits

  return(x)
}
```

5.1.1 Transforming and loading the datasets.

To analyse each CSV file it first had to be loaded into Rstudio and then transformed. When we loaded it we set the header as true and the stringsAsFactors to false. The header is set to true as sometimes R studio assumes that the row names don't have a header and the columns do. Therefore the default is header=TRUE. This helps prevent any irregularities in the csv file. stringsAsFactors = FALSE tells R to keep character variables as they are rather than convert to factors. The command we used to load and transform the datasets can be seen below.

```
tweets <- read_csv('C:/Users/Kevin/Desktop/Project/clinton_trump_tweets.csv' ,header = T,stringsAsFactors = F)
```

5.1.2 Installing and loading the relevant packages

Prior to analyzing the datasets there was one final thing to do. We had to install and load our relevant packages. Image below showing how to do this.

```
# Install
install.packages("tm")
install.packages("SnowballC")
install.packages("wordcloud")
install.packages("RColorBrewer")
# Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

5.1.3 How to set directory for datasets in R:

```
> setwd("C:\\Users\\Kevin\\Desktop\\")
> poll <- read.csv("presidential_polls.csv", stringsAsFactors = FALSE,na.s
trings = "")
> setwd("C:\\Users\\Kevin\\Desktop\\")
> tweets <- read.csv("tweets.csv", stringsAsFactors = FALSE,na.strings = "")
```

5.2 Presidential Election Polls from 2015-2016 analysis

The first dataset we analysed was the Presidential Election Polls from 2015-2016 analysis. We decided to analyse this dataset first as it would give us the answer to our first research question which was who was predicted to win the election in the traditional opinion polls? The result here would shape the rest of our project.

To analyse this dataset we first loaded it up in R studio. After this we then used a number of techniques and packages from R to analyse the data. These techniques can be seen in appendix a.

5.2.1 Raw presidential poll results & adjacent presidential poll results.

The code below analyses the raw and adjacent polls for both Trump and Clinton. Ggplot offers us powerful graphics language for creating elegant and complex plots. Here ggplot takes on the variable data which was assigned to the csv file taken in previously. Aes or aesthetic mappings describe how variables in the data are mapped to visual properties of geoms. Here the aes is the month field. So the data is the poll and the aes is the month. Geom specifies the geometric objects that define the graph type. Smooth is the type of graph. y is the data which in this case is the rawpolls for each candidate. Both candidates are given their own colours and same fill types. The date is averaged out over 12 months with a month break between each result. The labs function below then declares the x axis as month, y axis as averaged poll results and the title.

```
# First a plot with the Averaged Poll Results Raw by Candidate
ggplot(data = poll, aes(month)) +
  geom_smooth(aes(y = rawpoll_clinton, colour = "Clinton",fill="Raw")) +
  geom_smooth(aes(y = rawpoll_trump, colour = "Trump",fill="Raw")) +
  scale_x_date(labels = date_format("%Y-%m"),
               date_breaks = "1 month")  +
  labs(x = "Month", y = "Averaged Poll Results (%)",
       title = "Presidential Election Poll Results From 12/2015 - 11/2016")
```

The code below is similar to above except for the second and third lines where the inputted data is taken as adjpoll_trump and adjpoll_clinton instead of raw polls for each.

```
# First a plot with the Averaged Poll Results  Adj by Candidate
ggplot(data = poll, aes(month)) +
  geom_smooth(aes(y = adjpoll_clinton, colour = "Clinton",fill="Adj")) +
  geom_smooth(aes(y = adjpoll_trump, colour = "Trump",fill="Adj")) +

  scale_x_date(labels = date_format("%Y-%m"),
               date_breaks = "1 month")  +
  labs(x = "Month", y = "Averaged Poll Results (%)",
       title = "Presidential Election Poll Results From 12/2015 - 11/2016")
```

The results for this code can be seen below in section 6.1.1

5.2.2 Raw & adjacent state predictions

The code below also analyses both the raw and adjacent polls for both Trump and Clinton. This time however it uses the maps package. First off the variable stateshapes is assigned to map.text which plots the map. The fortify package is then used to add information to data based on a fitted model which is the variable stateships declared below. Then polling data is added to the maps using means. The fields rawpoll and state are loaded in from the csv file. The function mapPlotp along with ggplot is used to plot the prediction map. Longitude is assigned to the x axis and latitude is assigned to the y axis. Hillary's results are to be color blue and Trumps red. The title is then assigned using ggtitle.

```
# plotting the map
stateshapes <- map.text("state", regions = ".", exact = FALSE, cex = 0.75,
                        add = FALSE, move = FALSE)

stateshapes <- fortify(stateshapes)
# Adding some polling data to the mapns
stateshapes$Hillary <- by(polls$rawpoll_clinton, tolower(polls$state), mean)[stateshapes$region]
stateshapes$Trump <- by(polls$rawpoll_trump,tolower(polls$state),mean)[stateshapes$region]
stateshapes$Win <- ifelse(stateshapes$Hillary > stateshapes$Trump,"Hillary","Trump")

# plot the prediction map
mapPlotp <- ggplot(data=stateshapes,aes(x=long,y=lat,fill= Win,group=group))
mapPlotp <- mapPlotp + scale_fill_manual(values = c("Hillary"="blue","Trump"="red"))
mapPlotp <- mapPlotp + geom_polygon(colour = "grey")
mapPlotp <- mapPlotp + coord_map(project="conic", lat0 = 30)
mapPlotp <- mapPlotp + ggtitle("2016 US Election Predictions by State Raw")
mapPlotp
```

The code below is similar to above except for the first and second lines where the inputted data is taken as adjpoll_trump and adjpoll_clinton instead of raw polls for each. This is all that is changed to get the adjacent result instead of the raw.

```
stateshapes <- fortify(stateshapes)
# Adding some polling data to the mapns
stateshapes$Hillary <- by(polls$adjpoll_clinton, tolower(polls$state), mean)[stateshapes$region]
stateshapes$Trump <- by(polls$adjpoll_trump,tolower(polls$state),mean)[stateshapes$region]
stateshapes$Win <- ifelse(stateshapes$Hillary > stateshapes$Trump,"Hillary","Trump")
```

The results for this code can be seen below in section 6.1.2 & 6.1.3.

5.3 Tweets from Hillary Clinton and Donald Trump analysis.

The second dataset we analysed was the tweets from Hillary Clintons and Donald Trump's twitter accounts. We decided to analyse this dataset to gain a better insight into how both candidates utilised Twitter to their advantage during the election.

To analyse this dataset we first loaded it up in R studio. After this we then used a number of techniques and packages from R to analyse the data. These techniques can be seen below.

5.3.1 Retweets vs Favourites.

This code below uses the popular R package ggplot. This code allowed us to create a graph comparing the number of favourite and retweets on their tweets. The title was assigned 'Clinton vs. Trump: Retweets and Favorites' using the ggtitle function. The x axis is assigned field retweet_count and the y axis is assigned field favourite_count. geom_point gives the graph its circle shape and is size. The graph was then saved to our documents using the ggsave function.

```
p <- ggplot(tweets, aes(x=retweet_count, y=favorite_count, colour=candidate)) +
  geom_point(alpha=0.5, size=3) +
  scale_color_discrete("") +
  xlim(0, 50000) +
  ylim(0, 100000) +
  xlab("Retweets") +
  ylab("Favorites") +
  ggtitle("Clinton vs. Trump: Retweets and Favorites") +
  theme_light(base_size=14) +
  theme(legend.position="top", legend.direction="horizontal")
```

The results for this code are seen below in section 6.2.1

5.3.2 Original tweet or retweet

The code below also uses the ggplot package. Here we compare whether or not a tweet is an original tweet or a retweet. The number of tweets is declared on the y axis and their twitter handle is declared on the x axis.

```
temp <- tweets %>% select(handle,is_retweet)
temp$is_retweet <- ifelse(temp$is_retweet=='False','Original Tweets','Retweets')
temp<- temp %>% group_by(is_retweet,handle) %>% summarise(n=n())
ggplot(temp, aes(x=handle, y=n, fill=is_retweet)) +
  geom_bar(position="dodge",stat='identity')
```

The results for this code are seen below in section 6.2.2

5.4 Tweets from final US Election day analysis.

Our last dataset to be analysed was the tweets from the final day of the us election. We analsyed this dataset with the aim in helping answer our research questions and aims. Our aim for this analysis was to calculate the number of mentions, retweets and favourites each candidate had on the final election day. The sentiment of these tweets would also be analysed.

5.4.1 Plots

The code below looks at the number of mentions each candidate had during the final Election Day. Frequency of the tweets is declared on the x axis and the hour of the day is declared on the y axis. Clinton is coloured blue and Trump red. This is declared in the scale_color_manual variable. The code for the other graphs retweets and favourites is very similar with either retweet or favourite substituting for mentions. Ggplot again is used to create the graph. Geom_line is used to make line plots and geom_point is used to make scatter plots. The lines and points on our graph below. The theme describes the size of the graph and its elements such as text.

```
## Mentions plot
p1 <-
data_for_plot %>%
    filter(metric=='Mentions') %>%
    ggplot(aes(hour,value,group=candidate,color=candidate))+
    geom_line(size=1.5,alpha=0.75)+
    geom_point(size=3,alpha=0.75)+
    scale_color_manual(values=c('blue','red'))+
    labs(x='Hour of day',y='Frequency',title='Number of Mentions',subtitle='Based on election day tweets') +
    theme_bw()+
    scale_x_continuous(breaks = seq(0,23,2))+
    theme(
        panel.grid.minor = element_blank(),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        legend.text=element_text(size=15),
        legend.position="top",
        legend.title = element_blank(),
        axis.text.x=element_text(size=15),
        axis.text.y=element_text(size=15),
        axis.title.x=element_text(size=15),
        axis.title.y=element_text(size=15),
        strip.text.x = element_text(size=15),
        strip.text.y = element_text(size=15),
        plot.title = element_text(size=15),
        plot.subtitle = element_text(size=15)
    )
png("C:/Users/Kevin/Desktop/Project/Mentions.png", width = 2220, height = 1020, units = 'px', res = 100)
print(p1)
dev.off()
```

The results of this code and the other plots can be seen below in section 6.3.1.

5.5 Sentiment analysis.

A sentiment analysis function was run in R to assign a score to a tweet depicting the overall sentiment of the tweets. “Three lexicons for sentiment analysis are combined here in a tidy data frame. The lexicons are the NRC Emotion Lexicon from Saif Mohammad and Peter Turney, the sentiment lexicon from Bing Liu and collaborators, and the lexicon of Finn Arup Nielsen. Words with non-ASCII characters were removed from the lexicons” **[U]**. These word-banks contain thousands of positive and negative words and emotions which were saved in the working directory and allowed the sentiment of tweets to be categorised into emotions, positive, negative and neutral, through the use of scores.

The nrc lexicon categorizes words in a binary fashion (“yes”/“no”) into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The bing lexicon categorizes words in a binary fashion into positive and negative categories. The AFINN lexicon assigns words with a score that runs between -4 and 4, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

All of this information is tabulated in the sentiments dataset, and tidytext provides a function get_sentiments() to get specific sentiment lexicons. Then the sentiment analysis is run. Code below.

```
## get sentiments of tweets using 3 methods bing/afinn/nrc
## run sentiment for clinton
clinton_sentiment_bing <- sapply(clinton_tweets,function (x) get_sentiment(x,method='bing'))
clinton_sentiment_afinn <- sapply(clinton_tweets,function (x) get_sentiment(x,method='afinn'))
clinton_sentiment_nrc <- sapply(clinton_tweets,function (x) get_sentiment(x,method='nrc'))

## run sentiment for trump
trump_sentiment_bing <- sapply(trump_tweets,function (x) get_sentiment(x,method='bing'))
trump_sentiment_afinn <- sapply(trump_tweets,function (x) get_sentiment(x,method='afinn'))
trump_sentiment_nrc<- sapply(trump_tweets,function (x) get_sentiment(x,method='nrc'))
```

```
## merge bing dataset
x <- data.frame(candidate='clinton',table(clinton_sentiment_bing))
names(x)[2] <- 'sentiment_value'
y <- data.frame(candidate='trump',table(trump_sentiment_bing))
names(y)[2] <- 'sentiment_value'
bing <- rbind(x,y) %>%
  group_by(candidate) %>% mutate(ratio=Freq/sum(Freq),sentiment_value=as.numeric(as.character(sentiment_value))) %>% data.frame()

## merge afinn dataset
x <- data.frame(candidate='clinton',table(clinton_sentiment_afinn))
names(x)[2] <- 'sentiment_value'
y <- data.frame(candidate='trump',table(trump_sentiment_afinn))
names(y)[2] <- 'sentiment_value'
afinn <- rbind(x,y) %>%
    mutate(sentiment_value=as.numeric(as.character(sentiment_value)),
           sentiment_value=ifelse(sentiment_value < -10,-10,ifelse(sentiment_value>10,10,sentiment_value))) %>%
    group_by(candidate,sentiment_value) %>%
    summarise(Freq=sum(Freq)) %>%
    mutate(ratio=Freq/sum(Freq)) %>%
    data.frame()

## merge rnc dataset
x <- data.frame(candidate='clinton',table(clinton_sentiment_nrc))
names(x)[2] <- 'sentiment_value'
y <- data.frame(candidate='trump',table(trump_sentiment_nrc))
names(y)[2] <- 'sentiment_value'
nrc <- rbind(x,y) %>%
    mutate(sentiment_value=as.numeric(as.character(sentiment_value)),
           sentiment_value=ifelse(sentiment_value < -4,-4,ifelse(sentiment_value>4,4,sentiment_value))) %>%
    group_by(candidate,sentiment_value) %>%
    summarise(Freq=sum(Freq)) %>%
    mutate(ratio=Freq/sum(Freq)) %>%
    data.frame()
```

More code screenshots in appendences section below showing the plotting of the graphs and also the emotions.

6. Findings and Results

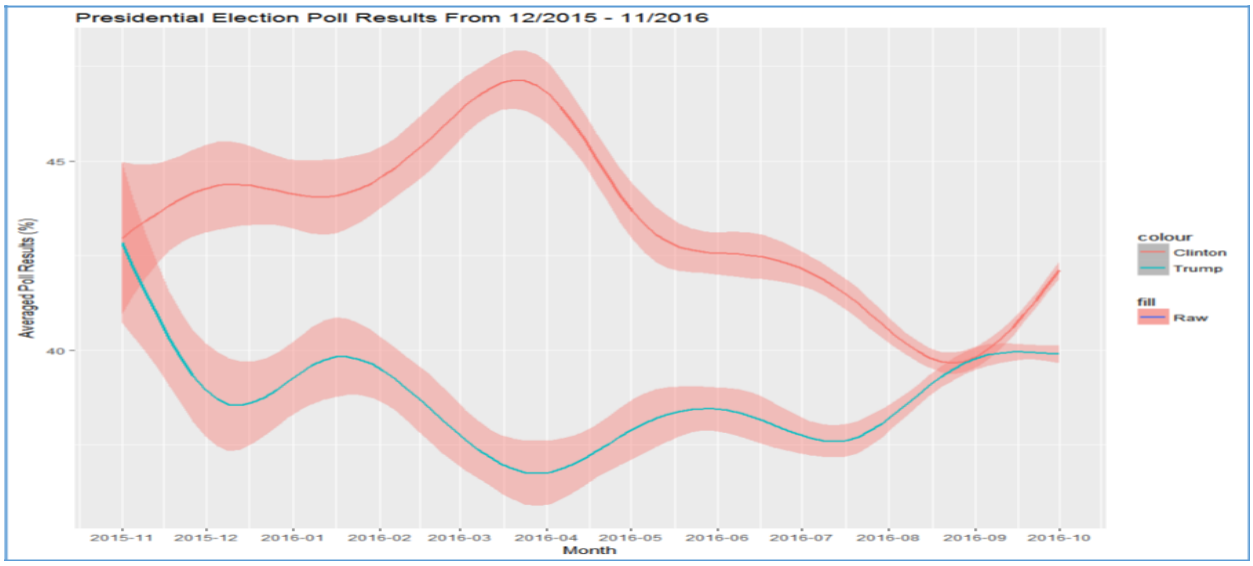
This section covers our findings and results which we obtained as a result of our analysis in section 5 above.

6.1 Presidential Election Polls from 2015-2016 findings and results

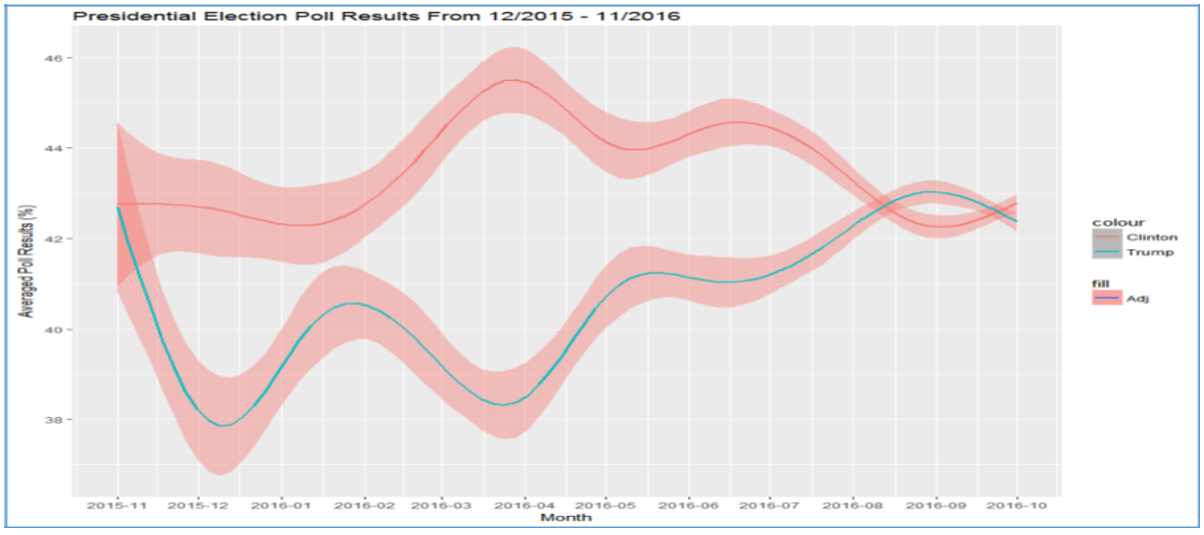
Below are the results of the code run above in section 5.1.1 and 5.1.2. Here we can see the huge difference between the raw and adjacent poll analysis.

6.1.1 Raw presidential poll results & adjacent presidential poll results

In the raw poll analysis below we can see Clinton maintaining a huge lead over Trump throughout most of the election. It narrows around the month of September but Clinton quickly rebounds. In the end Clinton is victorious.

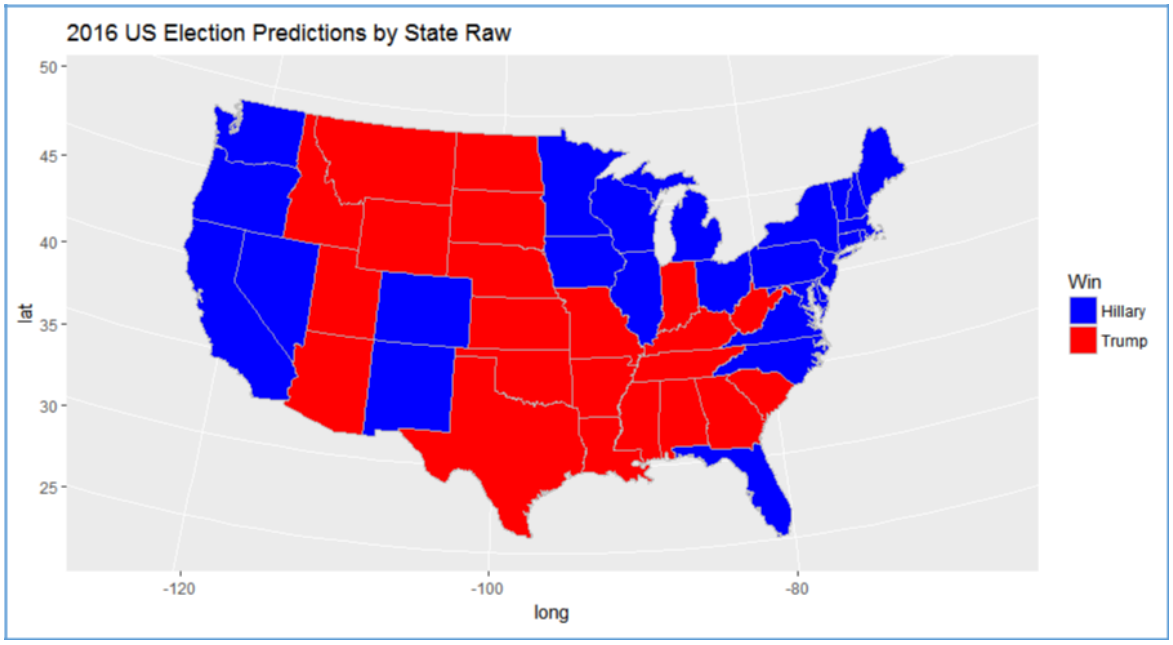


In the adjacent poll analysis we can see its a lot closer between both candidates. Trump narrows the lead towards the end of the election actually passes out Clinton for the first time. She narrows the gap and regains her lead narrowly coming into the final month.



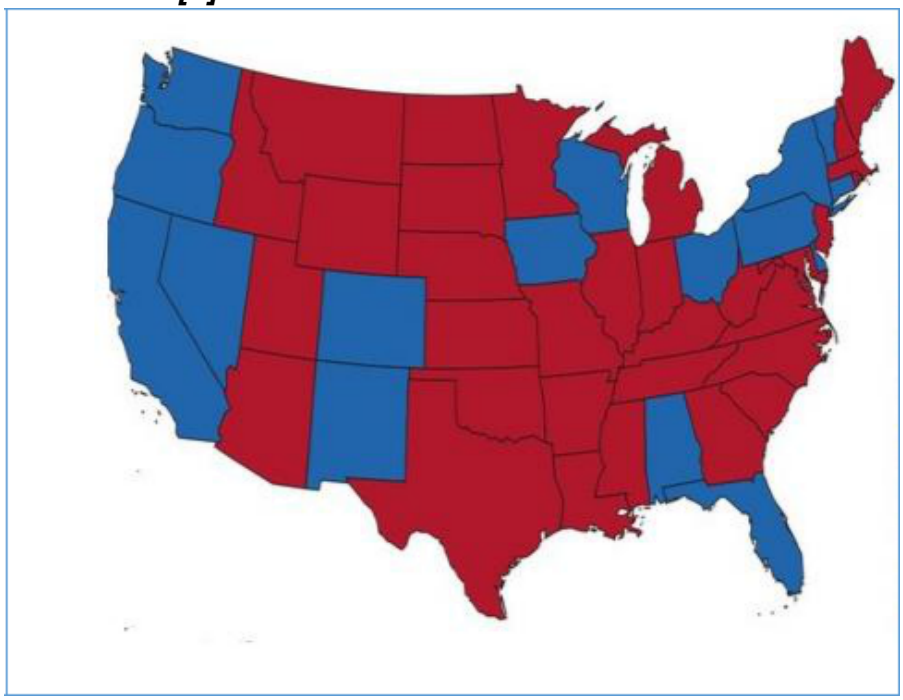
6.1.2 Raw & predicted 2016 election result according to an AI based computer

In the raw poll map analysis below we can see Hillary with a dominant win over Trump. She takes the key swing states and wins the electoral by a landslide. If we compare this with the predicted result from an AI based computer below we can see the results are very close. This was the generally predicted result for the election. It gives us a less accurate result.



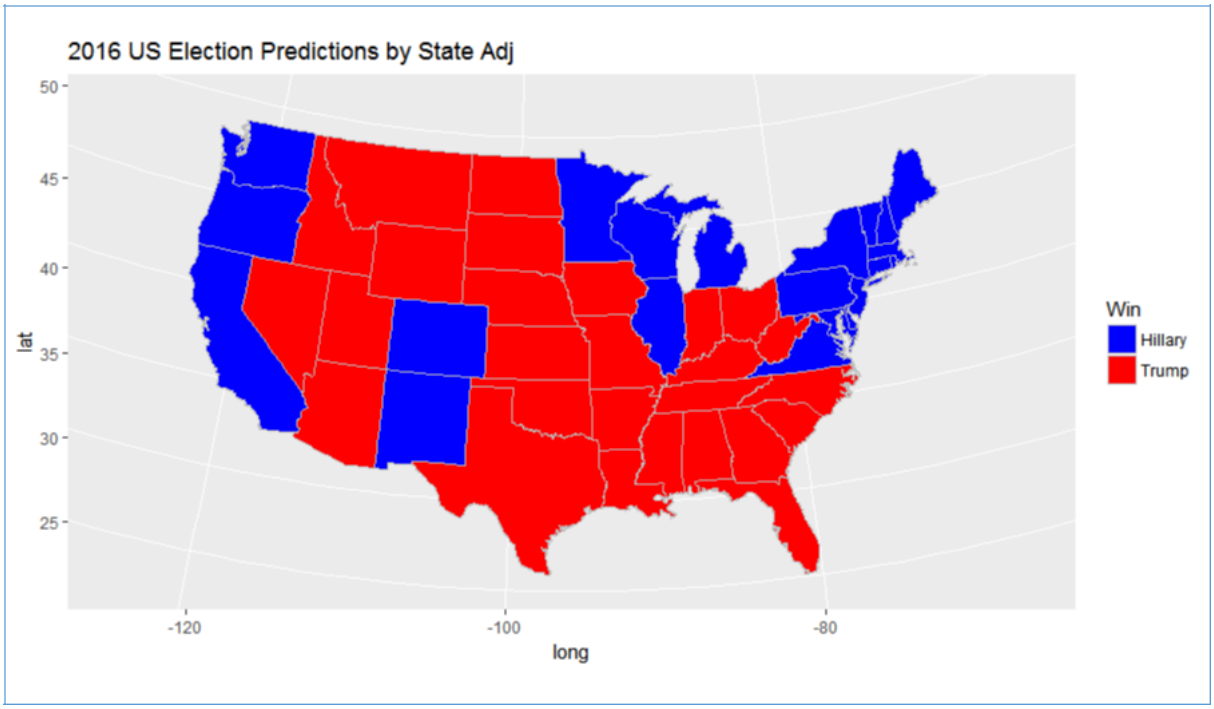
Predicted result of US Election

"A Predictive-Descriptive Artificial Intelligence-based expert computer system predicts a Democratic landslide victory in the 2016 presidential race, regardless of the nominated candidate." ***[T]***

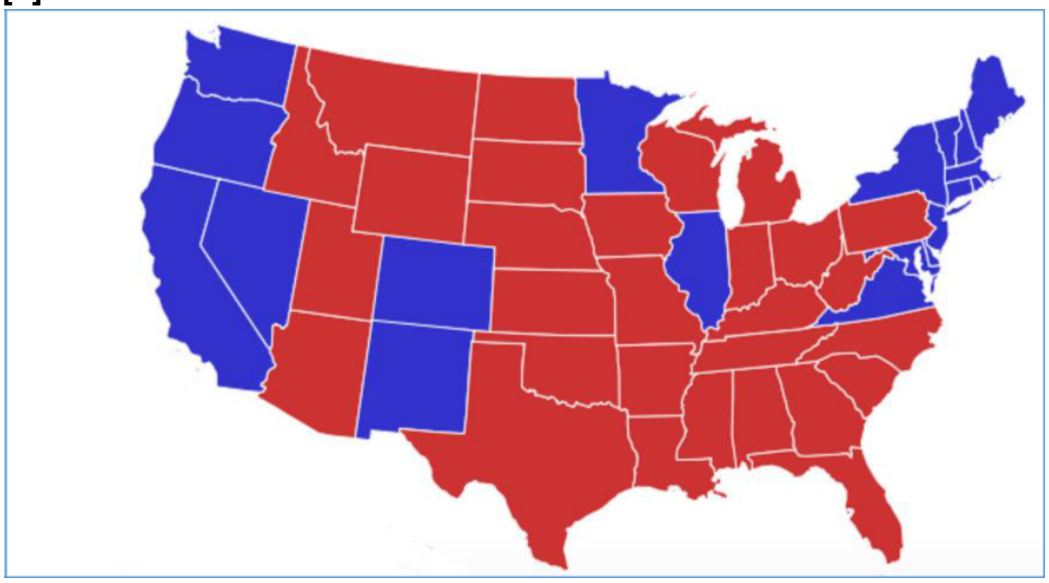


6.1.3 Adjacent & Actual 2016 election result.

In the adjacent map analysis below we can see a massive difference in the results in each state. The election has been turned upside down. Now, Donald Trump is the winner and clear leader. If we compare this to the map above there is a huge difference. If we compare this to the actual result of the 2016 US Election map shown below we can see that it is very close. The adjacent map gives us a more accurate result.



Actual result of US Election
[Z]

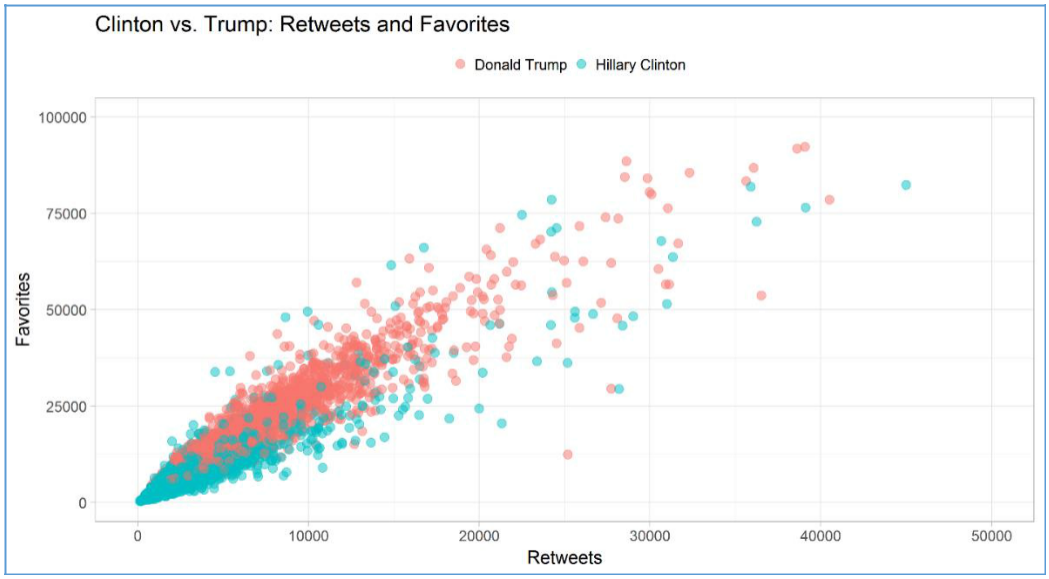


6.2 Tweets from Hillary Clinton and Donald Trump findings and results. .

Below are the results of the code run above in section 5.2.1 and 5.2.2. Here we get an insight into the popularity of each candidates Tweets on twitter.

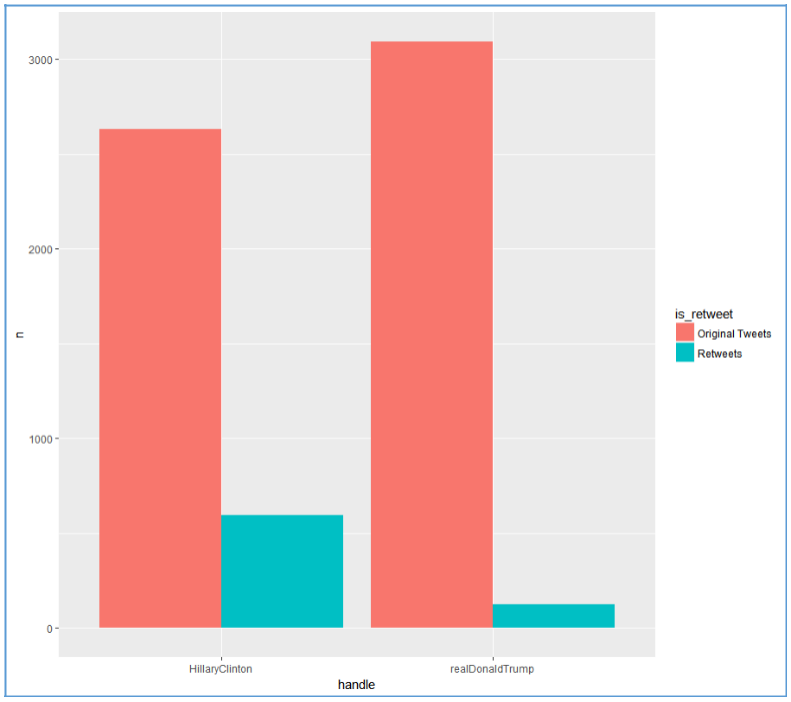
6.2.1 Clinton vs Trump retweets

In the graph below, we can see that on average Trump is a lot more popular in both aspects. This show us the wider reach of his Tweets. He is a lot more popular.



6.2.2 Original tweets or retweets.

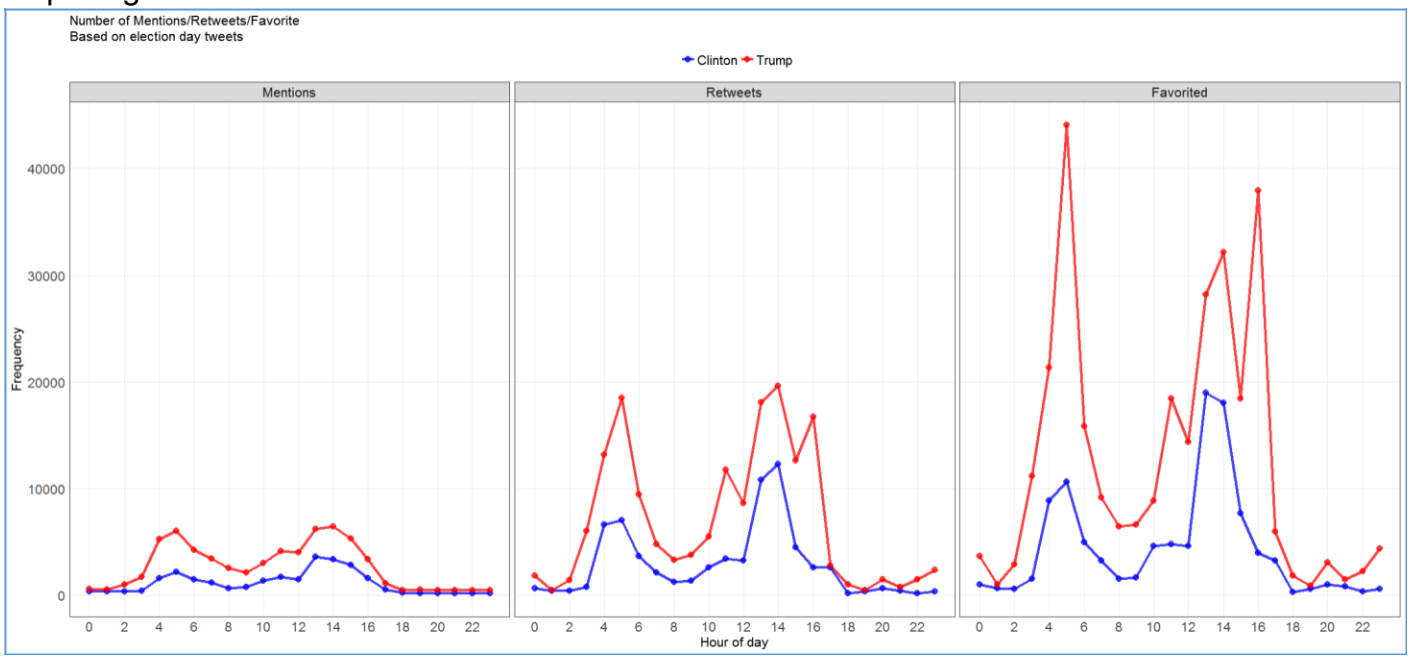
The graph below shows us how each candidate sues Twitter. Despite replying to Tweets more Trumps tweets have a wider range than Clintons. This show us his popularity advantage.



6.3 Tweets from final US Election day

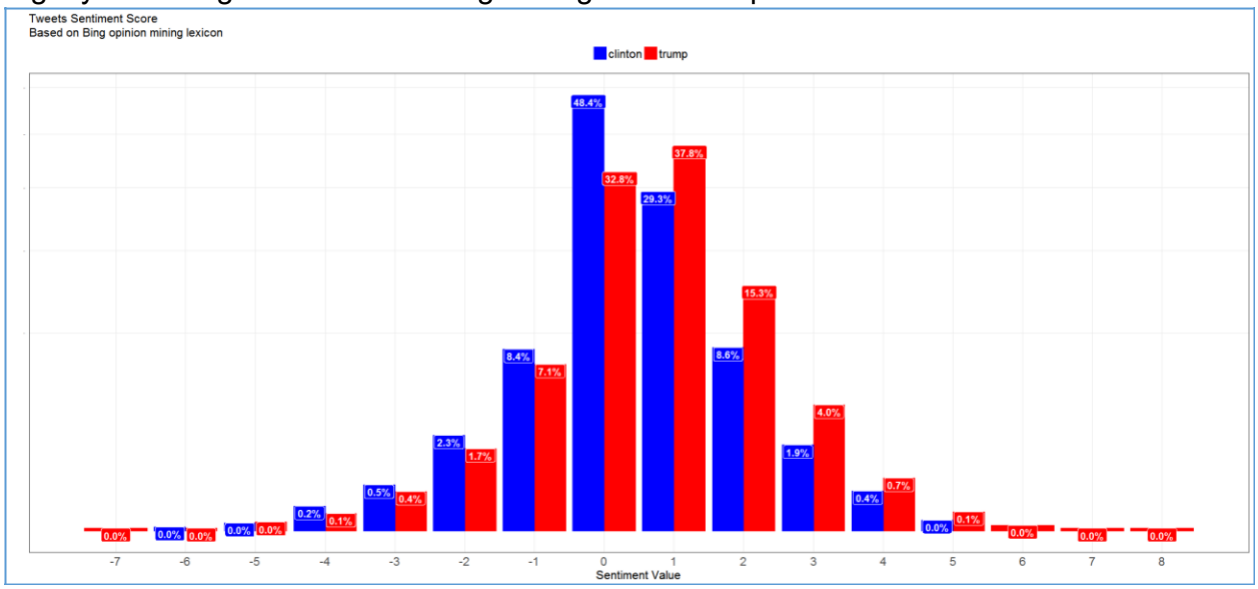
6.3.1 Mentions Retweets and Favourites

Our results below show the number of mentions, retweets and favourites for each candidate during the final Election Day. Here we can see just how more popular Trump is. It's in the magnitude of 10's of thousands of tweets. This graph was one of our most surprising results.

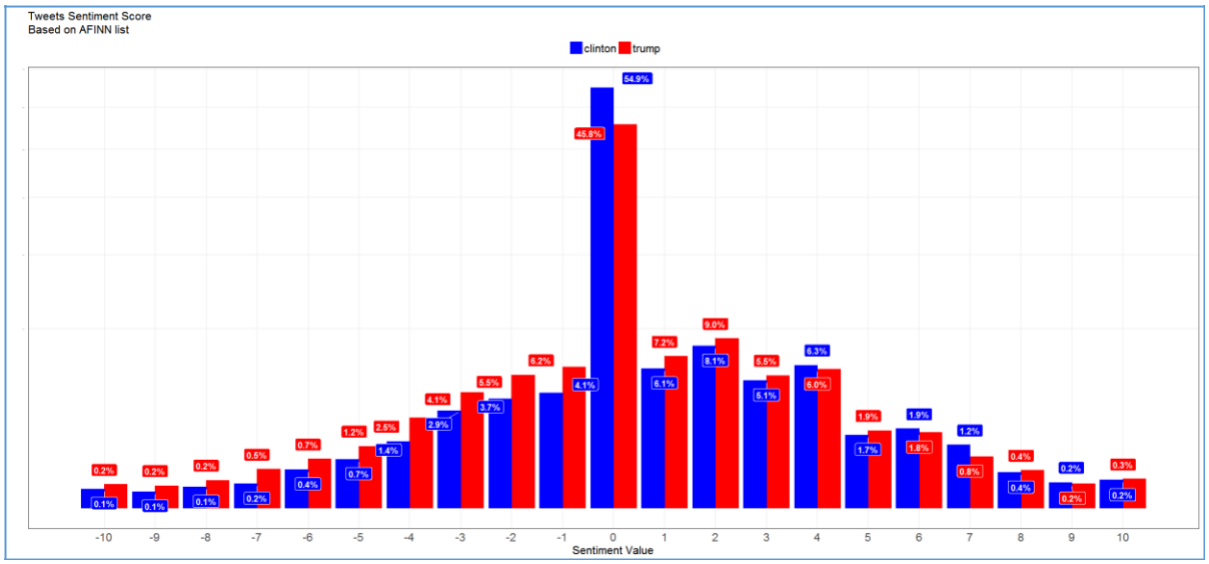


6.4 Sentiment Analysis.

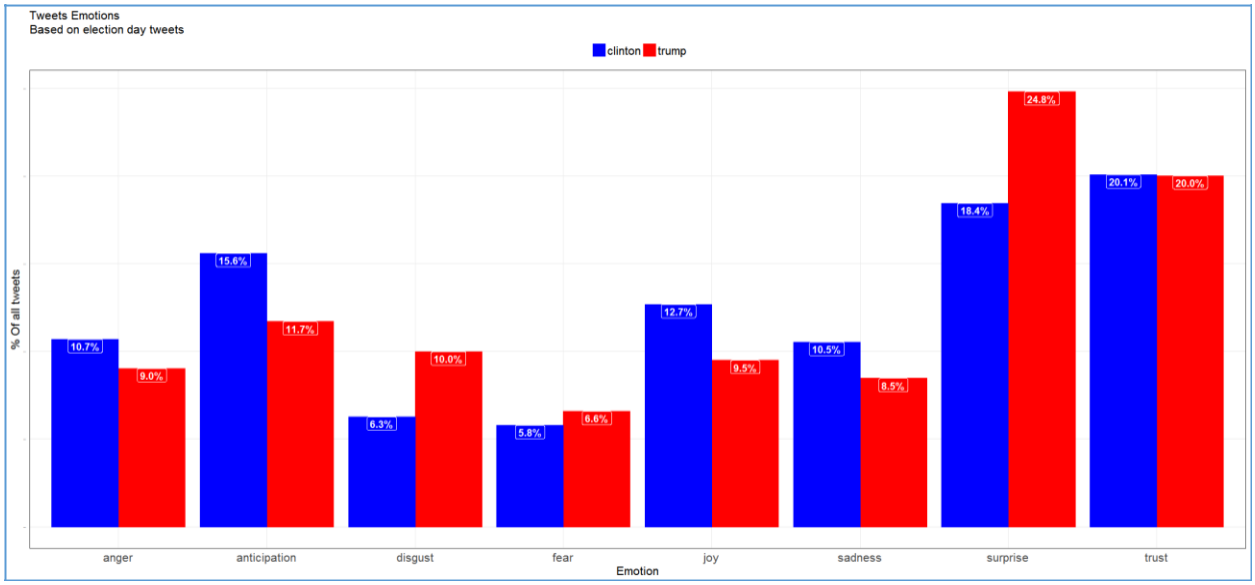
Our first result showing the sentiment analysis compared to the Bing lexicon. The results are very similar with Trump having the slight edge. Trump is slightly more positive and Clinton is slightly more negative. This shows good signs for Trump.



Our second result is again very similar. The sentiment value here is mainly neutral towards both candidates. Again Trump does have the higher sentiment values with Clinton again behind.



Our final sentiment analysis is to do with the emotions of the tweets. There are 8 different emotions ranging from anger to trust. Emotions towards Clinton are more on the negative side. Trump emotions are more of a surprise. This graph gives us a good insight into the emotions of people towards the candidates during the final day of the US election.



6.5 Results and key findings

Our key findings and results helped to answer our original research questions. Our research questions were as follows:

1. Who was predicted to win the election in the traditional opinion polls?

Our analysis of our first dataset showed us who, according to the polls, was going to win the elections. Our analysis gave us two different winners. The raw polls showed Clinton winning with the adjacent polls showing Trump winning. This was a very interesting finding. When people told the 'truth', Trump won. The raw data clearly shows Clinton prevailing as triumphant, this was the general predicament around the election. For this we needed to analyse our Twitter datasets and see if there is a correlation between the adjacent result and their popularity on Twitter.

2. Whose Twitter account is more popular?

Our analysis of our second data set showed us whose account was more popular on Twitter and whose tweets had more traffic. Our results showed us that although Trump did not reply to people as much as Clintons did, Trumps tweets had more retweets and favourites. This shows us that Trump utilized Twitter as much as he could.

3. On Election Day, who was predicted as the next US president according to their mentions on Twitter?

Our analysis of our final dataset aimed to answer two main questions stated above and below. Our results here showed us that Trump was mentioned more often, retweeted more and favourited more than Clinton. The difference was staggering. Trump was a lot more popular on the final day of the election than Clinton. It wasn't comparable.

4. Does the sentiment of tweets towards the candidates affect their popularity on Twitter?

For our final analysis we looked at the sentiment of the tweets. Here we discovered that generally both candidates had similar sentiment levels mostly in the neutral range. We found that Clintons' sentiment was slightly more negative.

7. Conclusion

In conclusion, we firmly believe that Twitter is the future of forecasting election results. Upon extensive analysis and research, we have come to this conclusion for a number of reasons:

- People generally express their true feelings online. In the graph in section 6.3.1 you can see that Trump is mentioned, retweeted and favourited substantially more than Clinton.
- The adjacent map on 6.1.3 also shows Trump winning. The adjacent results are the raw results of the poll with weighting applied. This is done to combat any statistical bias. People generally tend to lie so this weighting is applied.
- If we combine both of these results together, it can be concluded that Twitter is a more accurate tool for predicting the outcome of the election.

All of the opinion polls leading up to the election pointed at Hillary to succeed. As a result, it was decided that this project would be centered around analyzing whether Twitter analysis can be a more accurate predictor of elections than traditional opinion polls.

This boasts the question, is twitter a better predictor of elections than opinion polls? Is Twitter somewhat undermined as a future predictor of election results? It is hard to tell, but with the growing demand hinging towards social media there is no reason to believe that it will not have an immediate impact in the future. Our research above points towards this being true. We firmly believe that in the future Twitter will play a major part in the prediction of future elections.

The use of social media in recent years has continued to grow. It has the ability to impact a wide range of diverse areas, from Blogs, reviews and opinions, brand monitoring, communication and collaboration to social media sites. The days of surveys, questionnaires and opinion polls are well in the past. It is time to move forward, embrace the technological world we live in today. Taking full advantage of the benefits that social media has to offer is the key to opening up and broadening our minds in the era we live in today. As new tools are continuing to be developed in both the public and private sector every single day, we can expect to see the use of social media being implemented in these tools to further enhance their capabilities.

As Social Media develops, political organisations need to stay informed, up to date and realise the impact it can have on society allowing us to make more analytical decisions in real-time today. Political governments need to open their eyes around the world and take full advantage of the potential benefits that are right in front of them, which is – Twitter.

8. Further research & Recommendations

Upon completion of our analysis, we asked ourselves, what would we have done differently or if we were given more time, what extra research would we have undertaken?

- One of the first drawbacks during our project was being informed of the 15 day Twitter API limitation. This presented us with a wide range of problems. We weren't able to scrape the original Tweets that we intended to conduct analysis on. It led us to conduct research into finding appropriate datasets that were suitable for our project, in the end we found three excellent datasets. We would have done many things differently if we were to start from scratch and begin this project from the very start back in October. We would have been aware of the API limitation so we would have scraped Tweets in the days leading up to the Election, and most importantly scraped Tweets continuously throughout Election Day enabling us to create our own dataset.

- Implementing Python, SPSS, Minitab and IBM Watson Analytics would have been top of our to-do-list if given more time with this project. We feel they would have aided us greatly and allowed us to conduct more extensive, exploratory analysis. We would have learned the basic of these tools and techniques by watching various tutorial videos online.

- Conducting analysis on another dataset would also have been one of our main objectives if given more time with this project. We would have created maps for sentiment analysis, unfortunately we did not have the skills to create maps this time. It proved to be too intensive and time-consuming.

- For the Twitter dataset, we would have liked to use the fields latitude and longitude to calculate our users' location and analyse their Tweets. We would have liked to create a state map based on the sentiment analysis of their Tweets.

In the future implementing Twitter as a stepping stone for Election forecasting is a very realistic possibility for the near future. The popularity of Social Media is increasing at a huge rate with the number of followers on Twitter surging every single second.
Implementation of this concept around the world would turn out to be a good choice. It would eliminate voter fraud and biased votes.

The following is a description of how a bot on Twitter would work in regards to parsing and collecting Tweets on Election Day:

The bot would circulate Twitter on the day of the election. It would be programmed to analyse any Tweets relevant to the Election. In regards to voting, it would be programmed to pull and parse any keywords it finds suitable to be regarded as positive or negative to certain candidates. If positive, then it could count as a vote. The bot would be intelligent, predictive and cognitive. It would take time and effort to combine with artificial intelligence but we believe we are not far from this happening.

9. Appendences

This section will cover all of the work we did throughout our project.

9.1 Appendix A: Tools and Techniques used

Throughout this project we have used various tools and techniques. The have aided us in guiding our project in the right direction. The tools and techniques we have used were first presented to us across various modules throughout the year. Over the course of the last nine months, we have discovered many different ways that we can implement these tools and techniques throughout our project.

9.1.1 Microsoft Excel

Microsoft Excel was one of the first tools we used in our project. Excel is a spreadsheet program in the Microsoft Office system. The datasets that we obtained were very large at the beginning. They were in .csv format. Cutting them down was definitely required.
Excel allowed us to merge and tidy up each of the datasets before cleansing. Creating and formatting our datasets was our number one objective at the beginning and Excel allowed us to analyse data and make more informed decisions about how to move forward. We removed any unnecessary columns to ensure that our analysis was straight forward and to the point.

9.1.2 Tableau

Tableau is a visualisation tool that we made great use of. We downloaded to broaden our knowledge of various visualisation techniques that would aid us during the duration of our project. It was relatively easy to learn. Online tutorials as well as videos on YouTube allowed us to create interesting graphs with the software. We created graphs with Tableau such as comparing the number of Twitter followers of Barack Obama and Mitt Romney and comparing the number of Twitter followers of Donald Trump and Hilary Clinton. We learned about Tableau during Trevor Clohessy's Decision Theory Analysis module during first semester. 'A graph is worth a thousand words' – Clohessy, T

9.1.3 Python

Python was another program that we looked at. It was our first experience with any python code. We ran into many difficulties as we were not familiar with this program before. We decided to eliminate it from the project but it was a learning curve for us.

9.1.4 SPSS

SPSS is another tool that we thought would be beneficial to our project. We learned it during our Statistical Techniques for Business module during semester one. We ran into some difficulties as our dataset was incompatible for analyse by SPSS. They didn't go hand and hand together. We found that R Studio was a better route to go down. R provided us with a broader aspect of analysis and its wide range of techniques eventually led us to choosing it over SPSS.

9.1.5 R Studio

R studio was the main tool used throughout this project. We first encountered R during our Applied Customer Analytics back in semester one. This basic understanding of R allowed us to move forward and learn various ways we can implement it. We began to learn more advanced techniques that were required for this project.

A number of packages were required for this project. A brief description of some of the the Rpackages used are provided below:

- **TwitterR:** This package establishes a connection between R Studio and the Twitter API
- **StreamR:** This package provides access to Twitter so that tweets can be downloaded and parsed.
- **ROAuth:** This allows secure, authenticated access to Twitter using what is called a "handshake" which informs Twitter that the application is authentic by providing a unique PIN.
- **PlyR:** This package breaks large tasks down into smaller, more manageable tasks.
- **StringR:** This simplifies any string operations, and ensures all code is consistent.
- **ggplot2:** This package is used to generate graphical representations in R.
- **TM:** This package provides a framework for text mining applications in R.
- **Wordcloud:** This package enables unique word clouds to be generated.
- **XML:** This package was necessary for geo-coding when creating the map representing where tweets were sent from.
- **Maps:** This package allowed the data to be plotted on a world map.
- **SnowballC:** This package is used for text stemming.

9.1.6 Sentiment Analysis.

Sentiment analysis is essentially ''the process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.'' (Oxford, 2017) It has the ability to analyse peoples' sentiments, attitudes and emotions towards a certain event. This was particularly evident during the US Election.

The emotional tone behind a body of text is often overlooked. Fake news, untruthful political speeches and voter fraud became a consistent topic throughout the US Election. Throughout the US Election there have been many policy changes and campaign announcements with regard to 'building the wall', healthcare and employment. Political candidates Donal Trump and Hilary Clinton made great use of sentiment analysis by using it to view overall opinions about such changes in real-time. Twitter is the number one medium to monitor this change. Twitter has enabled these two candidates to alter their approach slightly in order to better relate to and develop a better, more honest friendship with their voters and constituents. It allowed them to let

their opinions and views be heard, let the public react and then fine tune their initial view based upon the reaction of their voters.

Sentiment analysis was chosen as one of the main focal points in our project due to its ability to accurately analysis large sets of unstructured data. The US Election was perhaps the hottest topic of 2016 with questions such as ‘Who are you voting for?’ and ‘Who do you think is going to win?’ being frequently asked questions. As a result, we came to a conclusion that this project would be centred around analysing the concept that is Twitters users’ opinions, or sentiment, a better predictor of elections than traditional opinion polls. How well did Twitter affect who would be crowned the next President of the United States of America.

9.1.7 Project Management

During the months October, November and December we would meet one or two times per week. Meeting every week was not possible so we took each week as it came. Assignment, exams and doing our SAP project often presented many obstacles. The use of Trello and Kanban allowed us to proceed in working on our project to the fullest possible extent. It enabled us to adjust our workload accordingly, week by week to adjust to and bypass everything that came in our way.

After our exams in December, we proceeded to meet two to three times per week from January onwards. Kevin had access to a lab down in the IT building so we used to meet down there. The lab had a wide range of computers and a large white board that we made great use of. The projector in the room enabled us to display our graphs on the white board. This allowed us to analyse graphs together as a whole, hearing other people’s opinions and overall helped our project take a step in the right direction.

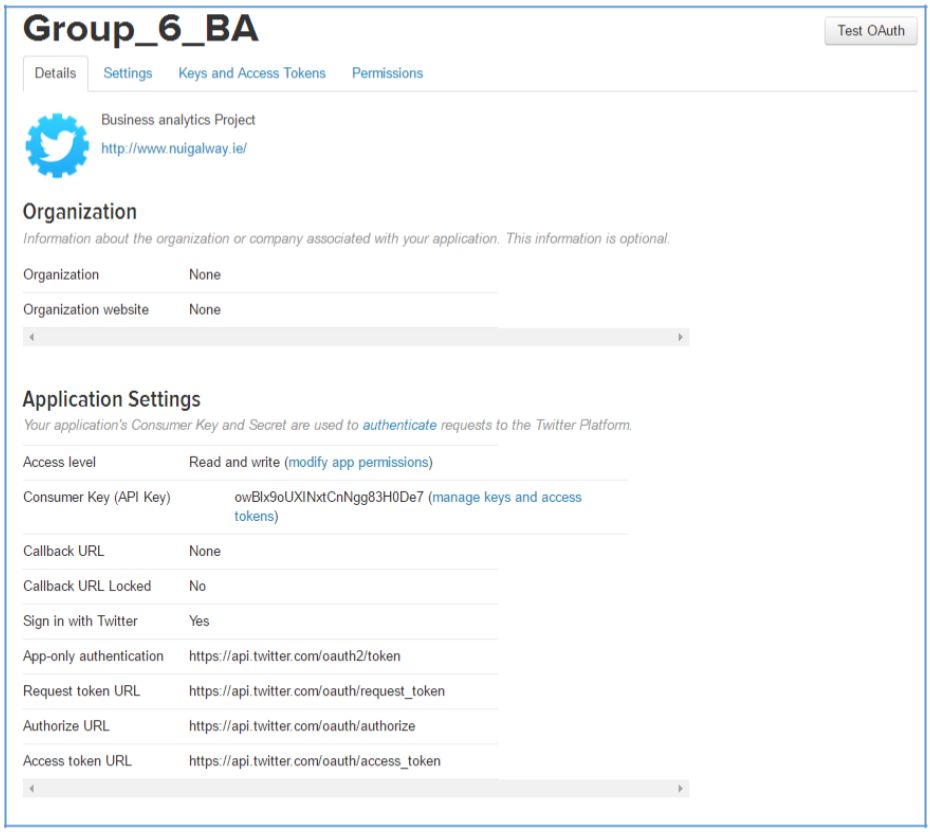
We used the white board to define the stages of our workflow. We were able to setup a way to move each task from one stage to the other, whether it was completed or not. Our first objective during our meetings would be to make two stacks of lists, with information stored on cards or pieces of paper. The first stack of lists would contain information about a task that needs to be done. The second stack would consist of tasks completed. As we would complete each card it would be then moved to the ‘complete’ stack. This allowed is to efficiently manage all tasks. No task was left unturned and everything was faced full on. This visualisation technique aided us greatly in defining what tasks needed to be completed and what tasks were left to do.

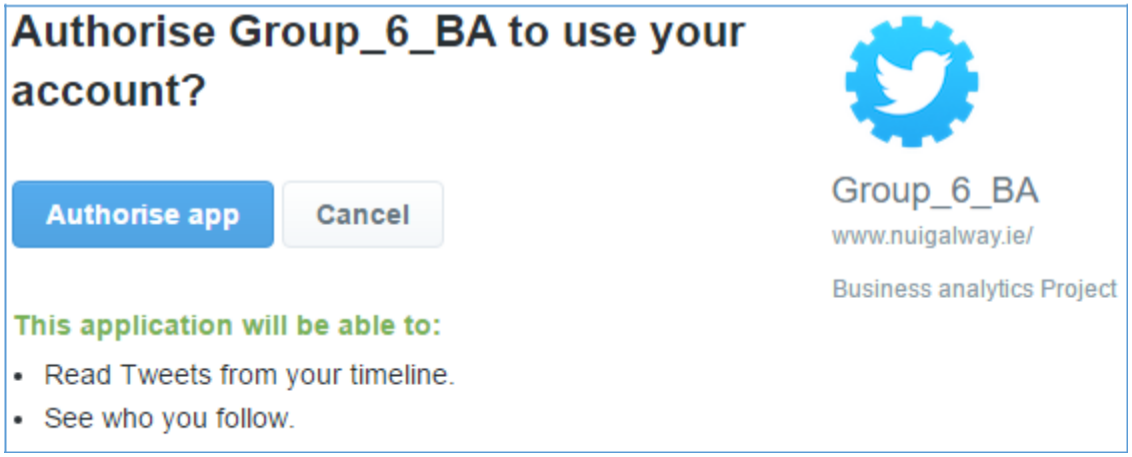
Appendix B: Code & Graphs

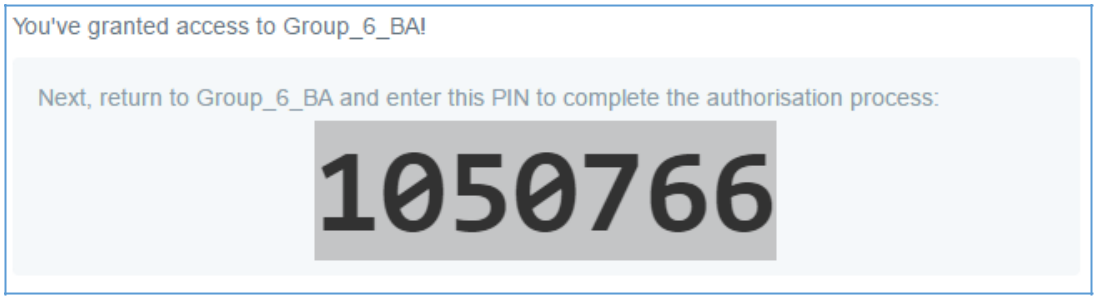
Below are some of the code and graphs that did not make into into our final document or were deemed irrelevant.

twitteR API for R studio implementation

```
1  install.packages("twitteR")
2  install.packages("ROAuth")
3  library("twitteR")
4  library("ROAuth")
5
```



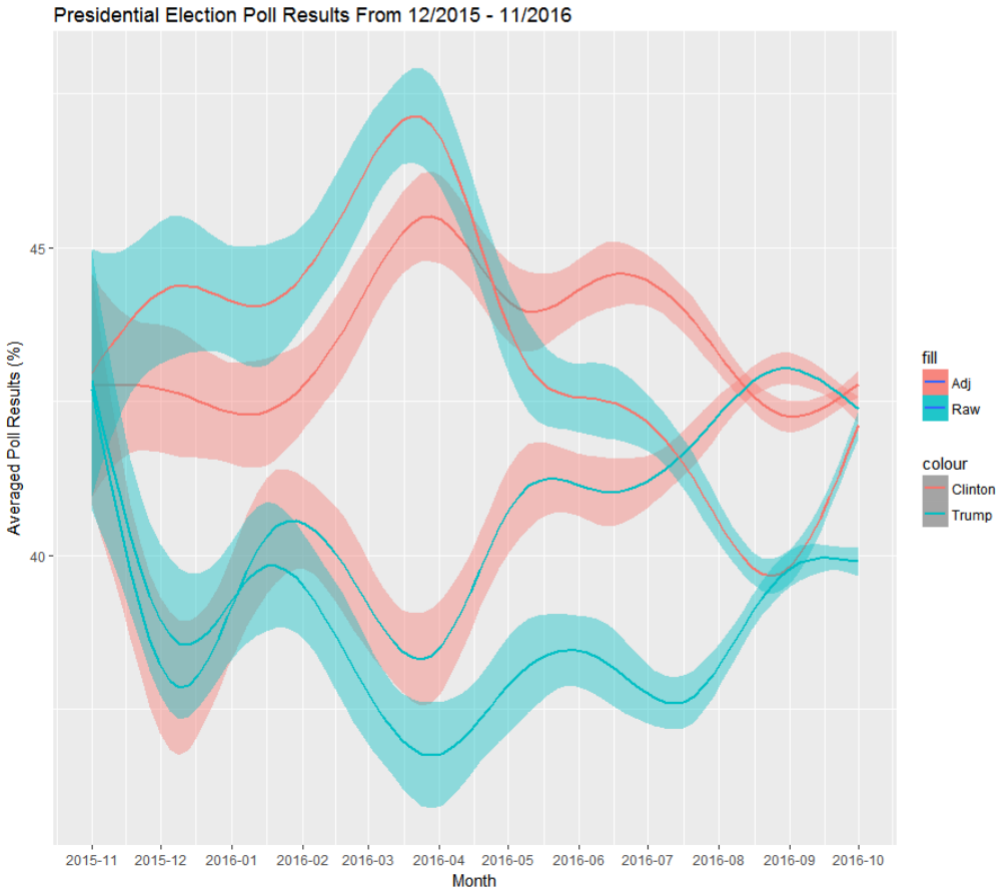





```
> cred <- OAuthFactory$new(consumerKey='owBlx9oUXINxtCnNgg83HODe7',
+                          consumerSecret='hOpgco7KEUnAzYbKCQ4bqaVpZgBsBOCNLgkhUcChFPBRF6Lvch',
+                          requestURL='https://api.twitter.com/oauth/request_token',
+                          accessURL='https://api.twitter.com/oauth/access_token',
+                          authURL='https://api.twitter.com/oauth/authorize')
> cred$handshake(cainfo="cacert.pem")
To enable the connection, please direct your web browser to:
https://api.twitter.com/oauth/authorize?oauth_token=JxYbOgAAAAAAOCl-AAABWOpVcQo
When complete, record the PIN given to you and provide it here: 1050766
> save(cred, file="twitter authentication.Rdata")
```

Dataset one additional code and graphs

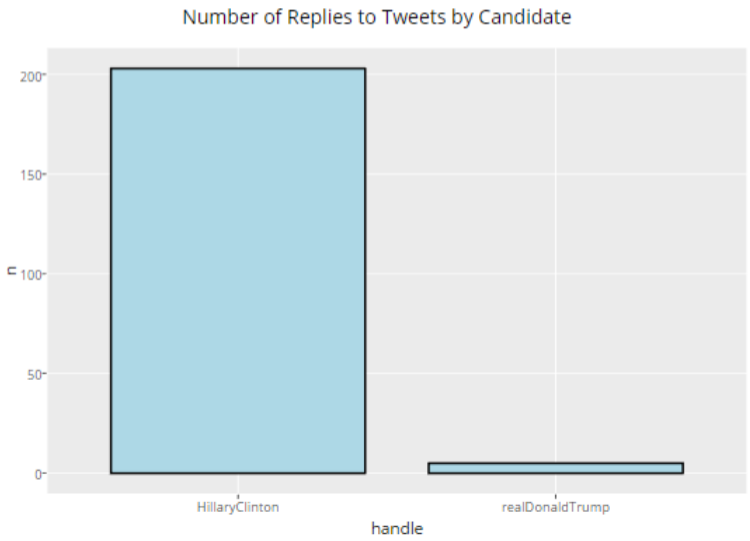
```
# First a plot with the Averaged Poll Results Raw and Adj by Candidate
ggplot(data = poll, aes(month)) +
  geom_smooth(aes(y = adjpoll_clinton, colour = "Clinton",fill="Adj")) +
  geom_smooth(aes(y = adjpoll_trump, colour = "Trump",fill="Adj")) +
  geom_smooth(aes(y = rawpoll_clinton, colour = "Clinton",fill="Raw")) +
  geom_smooth(aes(y = rawpoll_trump, colour = "Trump",fill="Raw")) +
  scale_x_date(labels = date_format("%Y-%m"),
               date_breaks = "1 month")  +
  labs(x = "Month", y = "Averaged Poll Results (%)",
       title = "Presidential Election Poll Results From 12/2015 - 11/2016")
```



Dataset two additional code and graphs

The code below also utilizes ggplot. The number of replies to tweets by each candidate is compared. The number of replies is along the y axis and their twitter handle is along the x axis. The temp variable is assigned the handle,in_reply_to_screen_name fields from the csv file. The temp variable is then over written by the temp_1 variable where it is then called by ggplot and written to the graph. geom_bar decides the type of graph and in this case a bar chart.

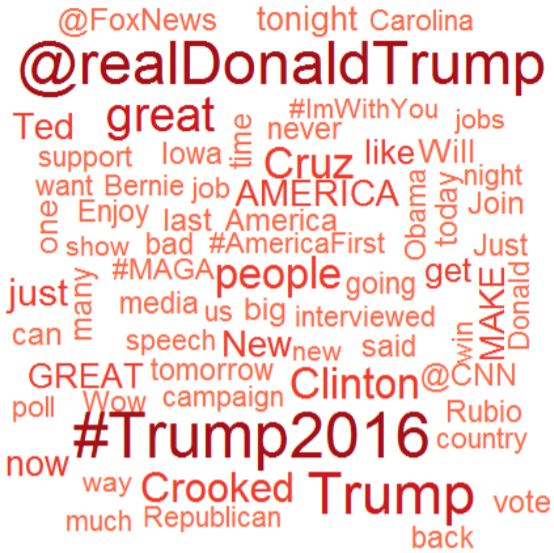
```
temp <- tweets %>% select(handle,in_reply_to_screen_name)
temp$in_reply_to_screen_name <- ifelse(temp$in_reply_to_screen_name=='',NA,temp$in_reply_to_screen_name)
temp <- na.omit(temp)
temp_1<-temp %>% group_by(handle) %>% summarise(n=n())
ggplotly(ggplot(temp_1, aes(x=handle, y=n)) +
           geom_bar(stat="identity", fill="lightblue", colour="black")+ggtitle("Number of Replies to
                                                                               Tweets by Candidate"))
```



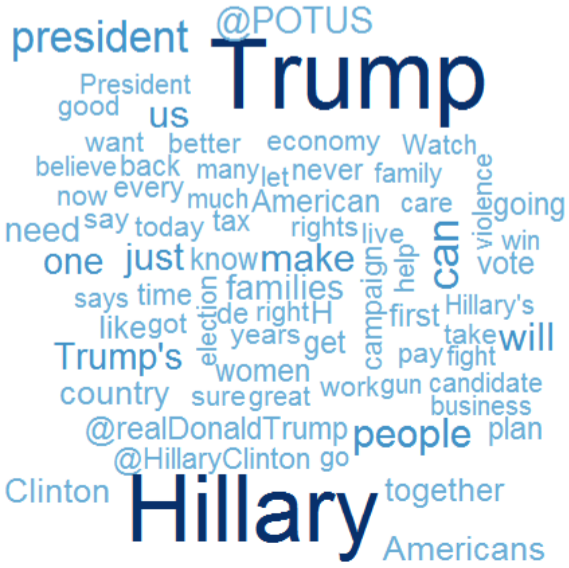
Wordcloud code:

```
get_words_with_freq <- function(tweet.text) {
  # separate the words (&amp; and &gt; appear quite a lot, especially in Trump's tweets)
  elements <- str_split(tweet.text, "(&amp;|&gt;|\\s)")[[1]]
  # remove urls
  elements <- grep("^http", elements, value = TRUE, invert = TRUE)
  # remove all punctuation except ', # and @ (because we're on twitter)
  tweet.words <- elements %>% gsub(pattern = "[^[:alnum:][:space:]'#@]",
                                   replacement = "")
  # strip trailing white spaces if any
  tweet.words <- str_trim(tweet.words)
  # remove empty strings
  tweet.words <- tweet.words[-which(tweet.words == "")]
  # remove stopwords (list of stopwords was obtained from http://www.ranks.nl/stopwords)
  stopwords.regex <- "^(a|about|above|after|again|against|all|am|an|and|any|are|aren't|as|at|be|
  because|been|before|being|below|between|both|but|by|can't|cannot|could|couldn't|did|didn't|do|
  does|doesn't|doing|don't|down|during|each|few|for|from|further|had|hadn't|has|hasn't|have|haven't|
  having|he|he'd|he'll|he's|her|here|here's|hers|herself|him|himself|his|how|how's|i|i'd|i'll|i'm|i've|
  if|in|into|is|isn't|it|it's|its|itself|let's|me|more|most|mustn't|my|myself|no|nor|not|of|off|on|once|
  only|or|other|ought|our|ours|ourselves|out|over|own|same|shan't|she|she'd|she'll|she's|should|shouldn't|
  so|some|such|than|that|that's|the|their|theirs|them|themselves|then|there|there's|these|they|they'd|they'll|
  they're|they've|this|those|through|to|too|under|until|up|very|was|wasn't|we|we'd|we'll|we're|we've|were|weren't|
  what|what's|when|when's|where|where's|which|while|who|who's|whom|why|why's|with|won't|would|wouldn't|you|you'd|
  you'll|you're|you've|your|yours|yourself|yourselves)$"
  tweet.words <- grep(stopwords.regex, tweet.words, value = TRUE,
                      ignore.case = TRUE, invert = TRUE)
  # get unique words with frequencies
  uniq.tweet.words <- tbl_df(table(tweet.words)) %>% arrange(desc(n))
  return(uniq.tweet.words)
}
```

```r
# Trump word cloud
# Choose red palette with use of RColorBrewer
pal <- brewer.pal(n = 9, name = "Reds")
pal <- pal[-(1:4)]
# word cloud
wordcloud(trump.words$tweet.words, trump.words$n,
          scale = c(4, 1), max.words = 75, colors = pal)
```




```r
#Clinton word cloud
# Choose bluepalette with use of RColorBrewer
pal <- brewer.pal(n = 9, name = "Blues")
pal <- pal[-(1:4)]
# word cloud
wordcloud(clinton.words$tweet.words, clinton.words$n,
          scale = c(4.5, 1), max.words = 75, colors = pal)
```



Dataset three additional code and graphs

```
install.packages(c('ggplot2','lubridate','dplyr','reshape2','tidyr','ggrepel'
library(ggplot2)
library(lubridate)
library(dplyr)
library(reshape2)
library(tidyr)
library(ggrepel)
library(scales)
library(snow)
library(parallel)
library(syuzhet) ## get_sentiment()
```

5.3.1 Gathering data for plotting

```
data$created_at <- ymd_hms(data$created_at)
data$hour <- hour(data$created_at)
data$is_clinton <- ifelse(grepl('clinton',tolower(data$text))==TRUE,1,0)
data$is_trump <- ifelse(grepl('trump',tolower(data$text))==TRUE,1,0)
```

```
## get data for plots
data_for_plot <-
    data %>%
    group_by(hour) %>%
    summarise(Clinton_Mentions=sum(is_clinton)
              ,Clinton_Retweets=sum(retweet_count[is_clinton==1])
              ,Clinton_Favorited=sum(favorite_count[is_clinton==1])
              ,Trump_Mentions=sum(is_trump)
              ,Trump_Retweets=sum(retweet_count[is_trump==1])
              ,Trump_Favorited=sum(favorite_count[is_trump==1])
    ) %>%
    melt(id='hour') %>%
    separate(variable,c('candidate','metric'),'_') %>%
    mutate(metric=factor(metric,levels=c('Mentions','Retweets','Favorited')))
```

Plotting all the metrics in on plot code. Use in section 6 above.

```
## plot all metircs on one plot
p <-
data_for_plot %>%
ggplot(aes(hour,value,group=candidate,color=candidate))+
geom_line(size=1.5,alpha=0.75)+
geom_point(size=3,alpha=0.75)+
facet_grid(~metric,scales='free')+
scale_color_manual(values=c('blue','red'))+
labs(x='Hour of day',y='Frequency',title='Number of Mentions/Retweets/Favorite',subtitle='Based on election day tweets') +
theme_bw()+
scale_x_continuous(breaks = seq(0,23,2))+
theme(
    panel.grid.minor = element_blank(),
    axis.ticks.x = element_blank(),
    axis.ticks.y = element_blank(),
    legend.text=element_text(size=15),
    legend.position="top",
    legend.title = element_blank(),
    axis.text.x=element_text(size=15),
    axis.text.y=element_text(size=15),
    axis.title.x=element_text(size=15),
    axis.title.y=element_text(size=15),
    strip.text.x = element_text(size=15),
    strip.text.y = element_text(size=15),
    plot.title = element_text(size=15),
    plot.subtitle = element_text(size=15)
)

png("C:/Users/Kevin/Desktop/Project/All.png", width = 2220, height = 1020, units = 'px', res = 100)
print(p)
dev.off()
```

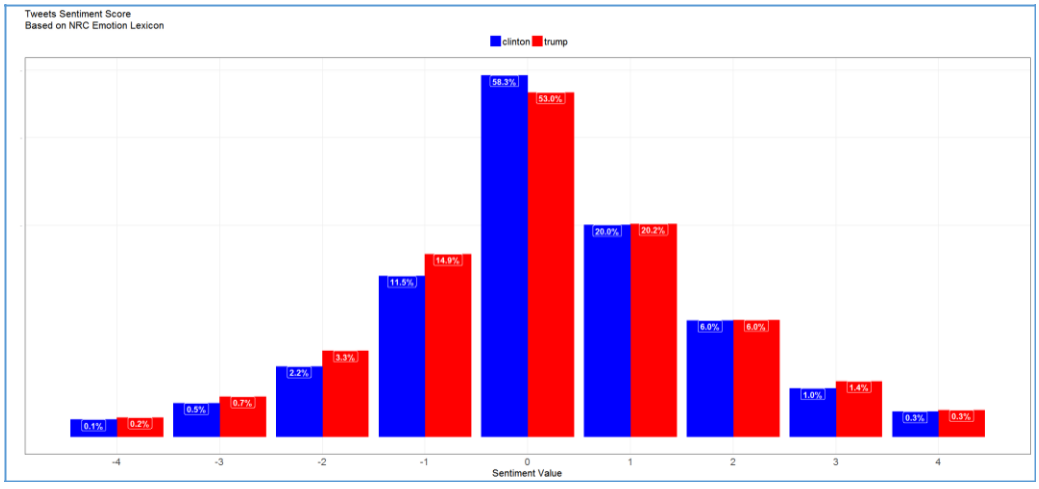
Prior to sentiment analysis. The text is cleaned. This calls the function cleanText mentioned above in section 5.

```
## get clinton and trump tweets
clinton_tweets <- data %>% filter(is_clinton==1) %>% select(text)
trump_tweets <- data %>% filter(is_trump==1) %>% select(text)

## clean text
clinton_tweets <- sapply(clinton_tweets,cleanText)
trump_tweets <- sapply(trump_tweets,cleanText)
```

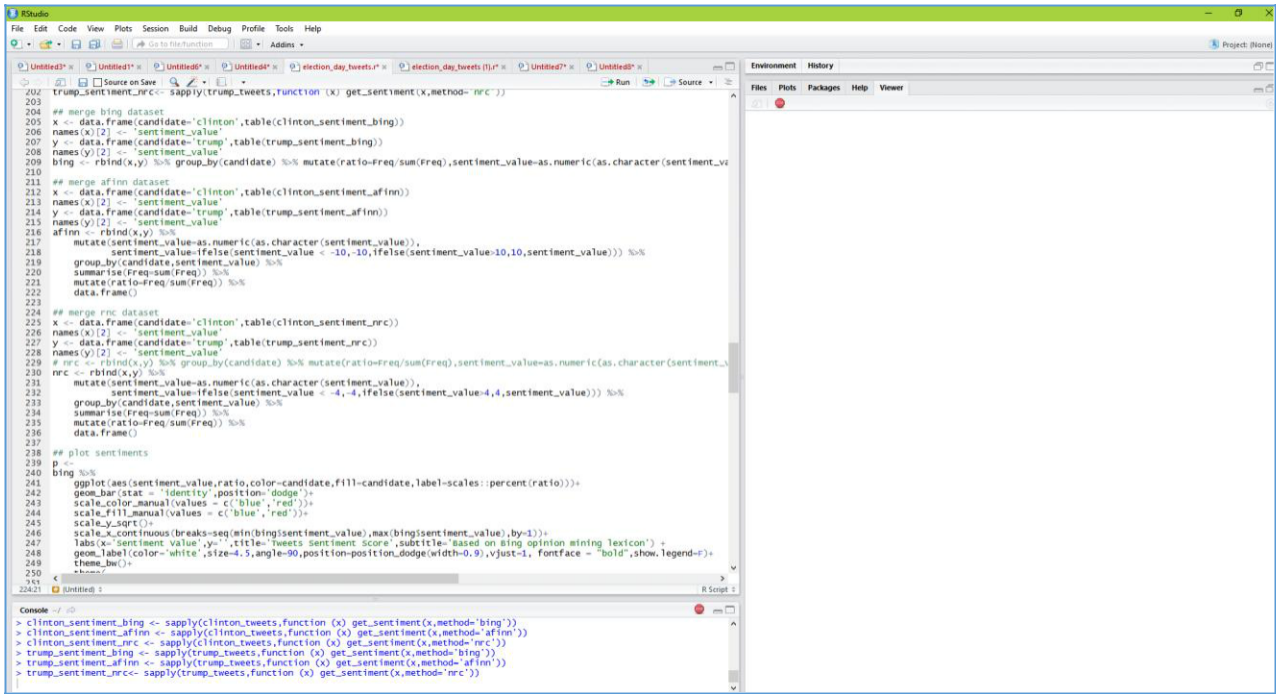
Sentiment Analysis code, merging the datasets.

```
## merge bing dataset
x <- data.frame(candidate='clinton',table(clinton_sentiment_bing))
names(x)[2] <- 'sentiment_value'
y <- data.frame(candidate='trump',table(trump_sentiment_bing))
names(y)[2] <- 'sentiment_value'
bing <- rbind(x,y) %>%
  group_by(candidate) %>% mutate(ratio=Freq/sum(Freq),sentiment_value=as.numeric(as.character(sentiment_value))) 
%>% data.frame()

## merge afinn dataset
x <- data.frame(candidate='clinton',table(clinton_sentiment_afinn))
names(x)[2] <- 'sentiment_value'
y <- data.frame(candidate='trump',table(trump_sentiment_afinn))
names(y)[2] <- 'sentiment_value'
afinn <- rbind(x,y) %>%
    mutate(sentiment_value=as.numeric(as.character(sentiment_value)),
           sentiment_value=ifelse(sentiment_value < -10,-10,ifelse(sentiment_value>10,10,sentiment_value))) %>%
    group_by(candidate,sentiment_value) %>%
    summarise(Freq=sum(Freq)) %>%
    mutate(ratio=Freq/sum(Freq)) %>%
    data.frame()

## merge rnc dataset
x <- data.frame(candidate='clinton',table(clinton_sentiment_nrc))
names(x)[2] <- 'sentiment_value'
y <- data.frame(candidate='trump',table(trump_sentiment_nrc))
names(y)[2] <- 'sentiment_value'
nrc <- rbind(x,y) %>%
    mutate(sentiment_value=as.numeric(as.character(sentiment_value)),
           sentiment_value=ifelse(sentiment_value < -4,-4,ifelse(sentiment_value>4,4,sentiment_value))) %>%
    group_by(candidate,sentiment_value) %>%
    summarise(Freq=sum(Freq)) %>%
    mutate(ratio=Freq/sum(Freq)) %>%
    data.frame()
```

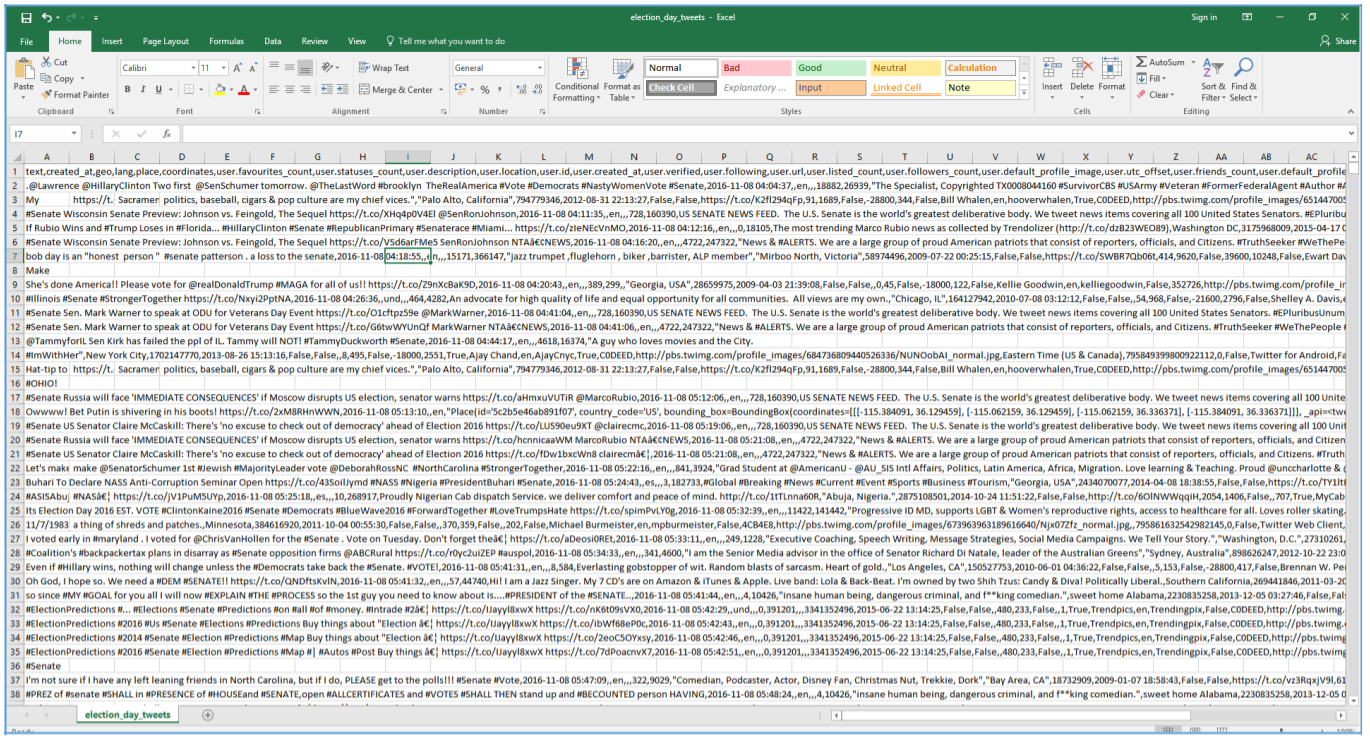
NRC sentiment analysis graph.

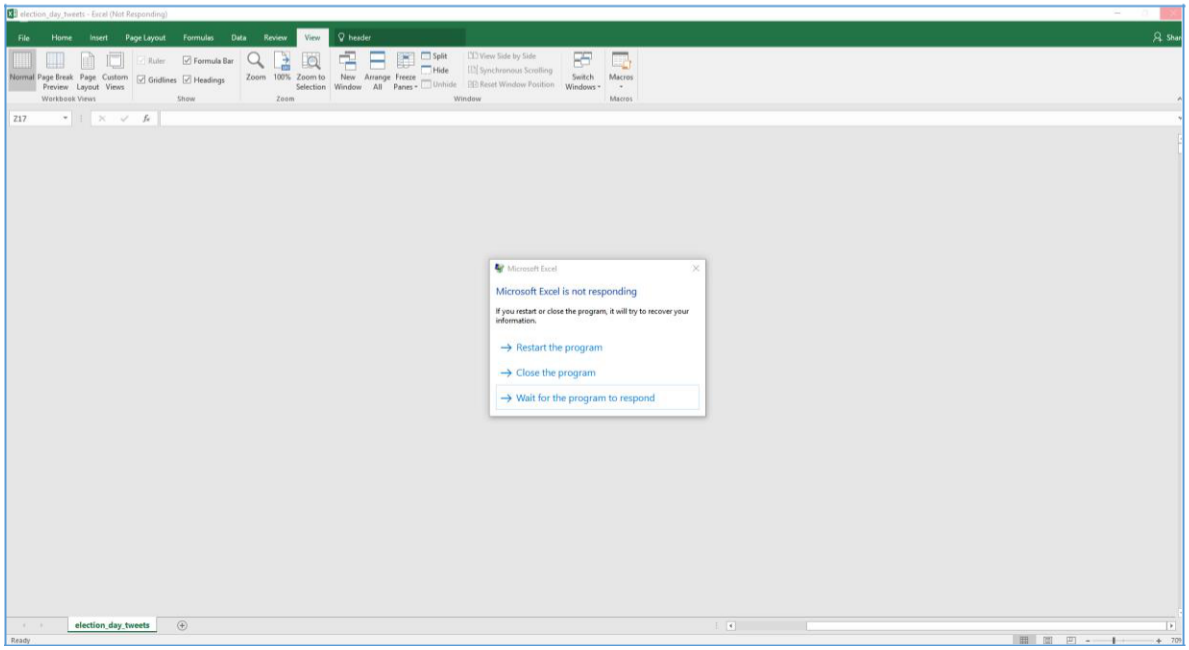


Running sentiment analysis. Downloading the word banks. Takes time.



Excel Screenshots. Why you need to set header as true when importing CSV files.





Appendix C: Personal Experience

This project all started out back in October when the ongoing phenomenon of the US Election was circulating. When given the task of analysing a large dataset for our major project we thought to ourselves what would be the best route to pursue. We wanted to choose a topic that was relevant to us and impacted our everyday lives. We discussed as a group what we felt was the most dominant medium in our lives today – social media. Social media has become a huge part in our lives with nearly all of the population having a smartphone and a laptop. As Donald Trump was elected president, we all questioned how did that happen? Hillary seemed such a clear winner, we wondered who would even elect Donald Trump as President? We wanted to look into why and how Trump managed to be elected. We found that Trumps fanbase on Twitter was massive with Hillary's not far behind. The US Election essentially turned into a social media war. All the opinion polls showed Hillary as prevailing triumphant, it made no sense. We felt the main way to move forward with this project is to investigate is Twitter a better predictor of the election than traditional opinion polls. To conduct analysis we would have to create our own dataset. This became an increasingly difficult task due to the millions of tweets related to the US Election, it would have been near impossible to cut it all down and would of taken months on end. The 14 day Twitter API limitation also stood in our way, disabling us from scraping Tweets older than 14 days. We were then presented with the online public dataset – Kaggle. Kaggle had suitable, up-to-date, publicly available datasets related to the US Election. We found three datasets that we decided we would proceed with. They were as follows:

1. Our first dataset is a Collection of Presidential Election Polls from 2015-2016.
2. Tweets from Hillary Clinton and Donald Trump
3. Tweets during the final day of the US election

With the wide range of datasets on Kaggle and the number of angles we could have taken this project from, it took time to carefully choose what datasets we thought were fitting. In the end these datasets allowed us to significantly develop our programming skills in R and Tableau. It was very frustrating at time with exams, study, assignment and SAP courses getting in the way but in the end we prevailed with an excellent project with insightful findings and results.

We all discussed what the most important part of the project would be. We all agreed that managing time would be a crucial aspect in the completion of our project so we put project management at the forefront of our objectives. We focused on becoming increasingly familiar with Google Docs, Lean Thinking, Trello boards and The Kanban philosophy. These project management techniques provided us with the flexibility we needed in order to complete this project. Week in week out the number of tasks and objectives needed to be completed was increasing. Lean thinking, Kanban and Trello boards allowed us to organise our workflow, ensuring it didn't pile up too much or leave

us overwhelmed. Stress levels increased at times but these techniques reminded us to keep or focus and not let our minds go astray.

The analysis stage proved to be the most challenging part of our project. We each explored possible areas of analysis with the datasets using R, Tableau and a variety of visualisation tools.

R Programming proved to be a huge aspect of our project. We conducted extensive analysis for weeks on end with the use of this tool. We created endless amounts of graphs comparing Trump vs Hillary along with eye-catching WordClouds with contributed greatly to the penultimate of our project. It took time to figure out how to execute certain functions properly. Witnessing error message after error message can be draining at times and sometimes led us to grow frustrated with our project. Everything worked out well in the end and we can safely say we implemented R to the fullest possible extent.

Tableau aided us greatly throughout the duration of our project. Tableau was a whole new tool to us with none of us having prior knowledge of how to use it. We were informed that it is an excellent visualisation tool and were encouraged to use it repeatedly throughout our project. We learned the basics of Tableau on tutorialspoint.com. They provided us with tutorials in developing our skills over time. After hours of tutorials videos on TutorialsPoint and on YouTube we eventually grasped an understanding of the tool.

We can safely say this project was a steep learning curve. The large scope that this project entailed allowed is the freedom to choose what we wanted to base the project on and conduct extensive analysis on our chosen topic. There were no rules or strict guidelines as to what programs, tools and techniques we could use and not use. We were essentially left on our own to develop and further our analytical skills. It was an experience that we will never forget. It was like no other project that we were faced with before. We worked vigorously throughout the project and bonded well as a team. We all worked simultaneously towards our common goal aided by our determined work ethic. Overall, this project has been an enjoyable, valuable experience.

10. Bibliography

[A] Credera 2017. Twitter Analytics Using R Part 1: Extract Tweets - www.credera.com. [ONLINE] Available at: https://www.credera.com/blog/business-intelligence/twitter-analytics-using-r-part-1-extract-tweets/. [Accessed 2017].

[B] 2016 Election Polls| Kaggle. 2017. [ONLINE] Available at: https://www.kaggle.com/fivethirtyeight/2016-election-polls. [Accessed 2017].

[C] Hillary Clinton and Donald Trump Tweets| Kaggle. 2017. [ONLINE] Available at: https://www.kaggle.com/benhamner/clinton-trump-tweets. [Accessed 2017].

[D] Business Insider. 2017. Polls: Hillary Clinton beats Trump, Cruz - Business Insider. [ONLINE] Available at: http://uk.businessinsider.com/polls-hillary-clinton-trump-ted-cruz-john-kasich-map-2016-4?r=US&IR=T. [Accessed 2017].

[E] Statista. 2017. • 2016 U.S. election: Twitter followers of candidates, September 2016 | Statistic. [ONLINE] Available at: https://www.statista.com/statistics/509579/twitter-followers-of-2016-us-presidential-candidates/. [Accessed 2017].

[F] Jon Keegan. 2017. Clinton vs. Trump: How They Used Twitter - WSJ.com. [ONLINE] Available at: http://graphics.wsj.com/clinton-trump-twitter/. [Accessed 2017].

[G] Statista. 2017. • 2012 election: Twitter followers of Obama and Romney | Statistic. [ONLINE] Available at: https://www.statista.com/statistics/243305/number-of-twitter-followers-of-barack-obama-and-mitt-romney/. [Accessed 2017].

[H] Statista. 2017. • Twitter: number of active users 2010-2017 | Statista. [ONLINE] Available at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/. [Accessed 2017].

[I] Kaggle. 2017. Election Day Tweets | Kaggle. [ONLINE] Available at: https://www.kaggle.com/kinguistics/election-day-tweets. [Accessed 2017].

[J] Kaggle. 2017. Hillary Clinton and Donald Trump Tweets | Kaggle. [ONLINE] Available at: https://www.kaggle.com/benhamner/clinton-trump-tweets. [Accessed 2017].

[K] Kaggle. 2017. 2016 Election Polls | Kaggle. [ONLINE] Available at: https://www.kaggle.com/fivethirtyeight/2016-election-polls. [Accessed 2017].

[L] Tableau Software. 2017. Free Tableau Download and Trial. [ONLINE] Available at: https://www.tableau.com/products/trial. [Accessed 2017].

[M] Internet Live Stats. 2017. Twitter Usage Statistics - Internet Live Stats. [ONLINE] Available at: http://www.internetlivestats.com/twitter-statistics/. [Accessed 2017].

[N] ggplot2. 2017. ggplot2. [ONLINE] Available at: http://ggplot2.org/. [Accessed 2017].

[O] ggrepel. 2017. . [ONLINE] Available at: https://cran.r-project.org/web/packages/ggrepel/README.html. [Accessed 2017].

[P] Garth Tarr. 2017. Parallel computation in R – Garth Tarr. [ONLINE] Available at: http://garthtarr.com/parallel-computation-in-r/. [Accessed 2017].

[Q] Matthew Jockers. 2017. Introduction to the Syuzhet Package. [ONLINE] Available at: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html. [Accessed 2017].

[R] *Election 2016 - Live Results - President Map*. [ONLINE] Available at: https://www.realclearpolitics.com/elections/live_results/2016_general/president/map.htm l [Accessed 2017].

[S] Statmethods.net. (2017). *Quick-R: ggplot2 Graphs*. [ONLINE] Available at: http://www.statmethods.net/advgraphs/ggplot2.html [Accessed 2017].

[T] *Predicting the 2016 Presidential Election - Experfy Insights*. [ONLINE] Available at: https://www.experfy.com/blog/predicting-the-2016-presidential-election-and-2025-unemployment-rate-d45dac23-8b41-4004-903f-e0fc73c97d44 [Accessed 2017].

[U] *Text Mining: Sentiment Analysis · UC Business Analytics R Programming Guide*. [ONLINE] Available at: http://uc-r.github.io/sentiment_analysis [Accessed 2017].

[V] Pier Lorenzo Paracchini. 2017. NLP - Sentiment Analysis using the tidytext package. [ONLINE] Available at: https://rstudio-pubs-static.s3.amazonaws.com/236096_2ef4566f995e48c1964013310bf197f1.html. [Accessed 2017].

[W] LeanKit. 2017. What is a Kanban Board? | LeanKit. [ONLINE] Available at: https://leankit.com/learn/kanban/kanban-board/. [Accessed 2017]

[X] 101 Ways. 2017. 7 Key Principles of Lean Software Development | 101 Ways. [ONLINE] Available at: http://www.101ways.com/7-key-principles-of-lean-software-development-2/. [Accessed 2017].

[Y] Trello. 2017. Trello. [ONLINE] Available at: https://trello.com/. [Accessed 2017].

[Z] Election 2016 - Live Results - President. 2017. Election 2016 - Live Results - President. [ONLINE] Available at: https://www.realclearpolitics.com/elections/live_results/2016_general/president/. [Accessed 29 June 2017].