CAPSTONE PROJECT REPORT

SENTIMENT ANALYSIS USING TWITTER DATA

Bloomiya Kurian

May 29, 2017

Contents

1 INTRODUCTION

Ever since President Trump signed Executive orders on Immigration and travel ban on immigrants from certain countries, people across the globe, especially in the U.S, have started debates in favor and against the move. The basis for these executive orders is national security. But some people believe that the ban is against the US culture and constitution. Others think that the move is essential in the wake of several terrorist attacks happening in the US and across the world. In addition to the bans, people have been voicing their opinions about other actions and views of the President. Many have been expressing their support and calls to protest against these actions and views through social media sites. It will be interesting to find out if the people who supported Mr.Trump during the election still support him after he became the President.

1.1 The Problem

In this project I would like to analyze certain types of tweets posted on Twitter through Sentiment Analysis, and identify Which state supports President Trump's actions more? Is there any relation between these states and the states that supported the President during 2016 election?

2 DATA SET AND SCOPE

Tweets that contain words or hash tags such as ‘immigrationban’, ‘travelban’, ‘muslimban’, ‘BanIslam’, ‘resist’, ‘DeleteUber’, ‘IResist’ and ‘TheResistance’ were collected through Twitter Search API. These search terms and hashtags were selected based on the major actions (executive orders on travel ban, immigration etc.) President Trump implemented immediately after becoming the President. Some articles such as this listed the most common hashtags people used to tweet their opinions about President Trump. The current scope of sentiment analysis is limited to tweets posted from states in the United States.

The Twitter Search API returned the following attributes:

1. Text: - This column contains the text of the tweet
2. Favorited:-“Indicates whether the Tweet has been liked by the authenticating user”
3. Favorite Count:-" Indicates approximately how many times the Tweet has been liked by Twitter users"
4. ReplyToSN:- “If the represented Tweet is a reply, this field will contain the screen name of the original Tweet’s author”
5. Created:- “UTC time when the Tweet was created”
6. Truncated:-“Indicates whether the value of the text parameter was truncated, for example, as a result of a re tweet exceeding the 140 character Tweet length. Truncated text will end in ellipsis, like this . . . Since Twitter now rejects long Tweets vs truncating them, the large majority of Tweets will have this set to false . Note that while native re tweets may have their top level text property shortened, the original text will be available under the re tweeted_status object and the truncated parameter will be set to the value of the original status (in most cases, false).”
7. ReplyToSID:- “If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s ID”
8. ID: “The integer representation of the unique identifier for the Tweet. This number is greater than 53 bits and some programming languages may have difficulty/silent defects in interpreting it”
9. ReplyToUID:- “If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s author ID”
10. StatusSource:- “Utility used to post the Tweet, as an HTML-formatted string. Tweets from the Twitter website have a source value of web”
11. ScreenName:- The user screen name
12. RetweetCount: - “Number of times the Tweet has been retweeted”
13. IsRetweet:-
14. Retweeted
15. Longitude
16. Latitude

3 DATA EXTRACTION AND PREPARATION

3.1 Step 1: Connect to Twitter REST API and download tweets

```
# Install and load required packages

library("twitteR")
library("ROAuth")

# Create a twitter account and obtain required access credentials to connect to
# twitter API

api_key <- ""

api_secret <- ""

access_token <- ""

access_token_secret <- ""

# Set-up twitter authentication using the keys

setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)

# Extract tweets for each US state using searchTwitter function and write to a
# csv file. Few examples are shown here. A table with necessary details is
# provided below for all states. By replacing geocode and radius parameters for
# each city we can obtain tweets for all the locations included in the scope of
# this analysis.

tweets <- searchTwitter("travelban OR immigrationban OR muslimban OR #DeleteUber OR #Bar
#Resist OR #IResist OR #TheResistance",
    n = 1e+05, since = "2017-01-27", until = NULL, lang = "en", geocode = "61.2876,-149.

df_tweets_AK_Alaska <- twListToDF(tweets)
```

Longitude and Latitude values were null for most of the tweets.So location of each tweet was not available from the extracted data. But since the search query was based on a City and State, these attributes can be added to each data frame representing the location of the tweets. The search queries were repeated for all the other States. If the geographic structure of the State was not fitting within a radius, different cities within the state were included in the search with smaller radius. This step was to make sure the search radius does not cover outside the boundary of the State. If a State's boundary was covered in one big radius without splitting into multiple small radius-es, the State name itself was assigned as City

name.

The search process was done in the alphabetical order of States. Steps followed for first few States are included here as examples. The main search attributes for rest of the States are included in the Appendix section.

3.1.1 Alaska (AK)

```
# For Alaska the radius included in the search query covered most of the cities.
# So exact city of each returned tweet is not available, hence adding City Name
# as 'Alaska'

df_tweets_AK_Alaska$City <- "Alaska"

df_tweets_AK_Alaska$State <- "AK"
```

3.1.2 Alabama (AL)

```
# Search for Alabama includes two cities
tweets <- searchTwitter("travelban OR immigrationban OR muslimban OR #DeleteUber OR #Bar
#Resist OR #IResist OR #TheResistance",
    n = 1e+05, since = "2017-01-27", until = NULL, lang = "en", geocode = "33.4564,-86.8

df_tweets_AL_Birmingham <- twListToDF(tweets)

tweets <- searchTwitter("travelban OR immigrationban OR muslimban OR #DeleteUber OR #Bar
#Resist OR #IResist OR #TheResistance",
    n = 1e+05, since = "2017-01-27", until = NULL, lang = "en", geocode = "32.3569,-86.2

df_tweets_AL_Mongomery <- twListToDF(tweets)
```

3.2 Step 2: Add City and State columns:

```
df_tweets_AL_Birmingham$City <- "Birmingham"

df_tweets_AL_Birmingham$State <- "AL"

df_tweets_AL_Mongomery$City <- "Mongomery"

df_tweets_AL_Mongomery$State <- "AL"
```

3.3 Step 3: Merge data frames for each state and remove duplicate tweets:

The search for tweets from Alabama was done in two steps. First with a 100 mile radius centering the city Birmingham and the second with 100 mile radius centering the city Montgomery. Since there can be a radius overlap between these two cities, same tweet can be included in both data frames for the State. The goal was to combine all the data frames for a State into one data frame and remove the duplicate tweets.

```
# Combine data frames for Alabama

df_tweets_AL <- dplyr::bind_rows(df_tweets_AL_Birmingham, df_tweets_AL_Mongomery)

# Remove duplicate tweets

df_tweets_AL_Final <- df_tweets_AL %>% distinct(text, favorited, favoriteCount, replyToS
    created, truncated, replyToSID, id, replyToUID, statusSource, screenName, retweetCou
    isRetweet, retweeted, longitude, latitude, .keep_all = TRUE)
```

Above steps were repeated for all the remaining States in the U.S.

3.3.1 Arkansas (AR)

```
tweets <- searchTwitter("travelban OR immigrationban OR muslimban OR #DeleteUber OR #Bar
#Resist OR #IResist OR #TheResistance",
    n = 1e+05, since = "2017-01-27", until = NULL, lang = "en", geocode = "34.2089,-91.9

df_tweets_AR_Conway <- twListToDF(tweets)

df_tweets_AR_Conway$City <- "Conway"

df_tweets_AR_Conway$State <- "AR"

tweets <- searchTwitter("travelban OR immigrationban OR muslimban OR #DeleteUber OR #Bar
#Resist OR #IResist OR #TheResistance",
    n = 1e+05, since = "2017-01-27", until = NULL, lang = "en", geocode = "35.8358,-90.6

df_tweets_AR_Jonesboro <- twListToDF(tweets)

df_tweets_AR_Jonesboro$City <- "Jonesboro"

df_tweets_AR_Jonesboro$State <- "AR"

df_tweets_AR <- rbind(df_tweets_AR_Conway, df_tweets_AR_Jonesboro)
```

```
df_tweets_AR_Final <- df_tweets_AR %>% distinct(text, favorited, favoriteCount, replyToS
    created, truncated, replyToSID, id, replyToUID, statusSource, screenName, retweetCou
    isRetweet, retweeted, longitude, latitude, .keep_all = TRUE)
```

3.3.2 Arizona (AZ)

```
tweets <- searchTwitter("travelban OR immigrationban OR muslimban OR #DeleteUber OR #Ban
#Resist OR #IResist OR #TheResistance",
    n = 1e+05, since = "2017-01-27", until = NULL, lang = "en", geocode = "33.7039,-112.

df_tweets_AZ_Phoenix <- twListToDF(tweets)

df_tweets_AZ_Phoenix$City <- "Phoenix"

df_tweets_AZ_Phoenix$State <- "AZ"
```

Geocodes and radius used for other States and Cities are provided in the appendix section.

After collecting tweets from all states, all tweets were merged into one file.

3.4 Step 4: Merge all the state level files into one csv file:

```
# In the below code only dataframes for the examples included are included.
# Ideally dataframe for all the states need to be included to get the final
# dataset

df_tweets_US <- dplyr::bind_rows(df_tweets_AK_Alaska, df_tweets_AL_Final, df_tweets_AR_F
    df_tweets_AZ_Phoenix)

# Write to csv file

write.csv(df_tweets_US, file = "Tweets - Presidential actions 2017.csv")
```

4 INITIAL DATA EXPLORATION

The Consolidated Tweets Dataset needs to be read into a dataframe to perform further analysis.

```
# Load library

library(dplyr)

# Read data from CSV file. Link to the dataset:
# https://drive.google.com/open?id=0BxWW5x7AJ4r8OHhueVpKUGkyTWc

# tweets_data = readr::read_csv(file='Tweets - Presidential actions 2017.csv')

tweets_data = read.csv(file = "Tweets - Presidential actions 2017.csv", row.names = NULL
    header = TRUE)

# See structure of the data

str(tweets_data)
```

```
## 'data.frame':    657511 obs. of  18 variables:
##  $ Text         : Factor w/ 152287 levels "'''BATTERED PUNDIT SYNDROME'' LOLOLOLOLOLOL
##  $ Favorited    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ FavoriteCount: int  0 0 1 2 0 0 0 0 0 0 ...
##  $ ReplyToSN    : Factor w/ 10479 levels "__keating","_4P4TH3T1C",..: NA 10149 NA NA
##  $ Created      : Factor w/ 34955 levels "2/17/2017 13:55",..: 24745 24744 24744 2474
##  $ Truncated    : logi  FALSE FALSE TRUE FALSE FALSE TRUE ...
##  $ ReplyToSID   : num  NA 8.37e+17 NA NA NA ...
##  $ ID           : num  8.37e+17 8.37e+17 8.37e+17 8.37e+17 8.37e+17 ...
##  $ ReplyToUID   : num  NA 5120691 NA NA NA ...
##  $ StatusSource : Factor w/ 652 levels "<a href=\"http://10up.com\" rel=\"nofollow\">
##  $ ScreenName   : Factor w/ 189621 levels "______i1i1i1i1i",..: 42095 165558 68091 13
##  $ RetweetCount : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ IsRetweet    : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Retweeted    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Longitude    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Latitude     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ City         : Factor w/ 146 levels "Alaska","Albany",..: 14 14 14 14 14 14 14 14
##  $ State        : Factor w/ 49 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
# Summarize number of tweets collected per State

tweets_data %>% group_by(State) %>% summarise(tweet_count = n()) %>% ungroup() %>%
    arrange(State) %>% knitr::kable()
```

State	tweet_count
AK	2115
AL	5389
AR	1158
AZ	3993
CA	45198
CO	4686
CT	1820
DC	69355
DE	160
FL	10312
GA	11968
IA	1242
IL	15416
IN	4502
KS	1678
KY	1288
LA	2735
MA	16064
MD	73861
ME	1899
MI	1710
MN	3081
MO	24626
MS	749
MT	484
NC	10878
ND	3142
NE	1192
NH	1090
NJ	23276
NM	8537
NV	5499
NY	122622
OH	23140
OK	1838
OR	4342
PA	38733
RI	540
SC	3968
SD	1092
TN	1507
TX	29015
UT	1313

State	tweet_count
VA	53578
VT	310
WA	9002
WI	6124
WV	948
WY	336

The state Delaware had the lowest number of tweets. And New York had the highest number. State Idaho was missed out in the data extraction process. So there are no tweets available for Idaho, and will be excluded from the analysis process. Since the count of tweets per State varied drastically, the population for this data set is not even. States with metro cities usually have more people using social media than remote States. This could explain the reason for the outliers.

5 SENTIMENT ANALYSIS

Once the data set is cleaned and explored the next step is to identify the sentiment of each tweet in the data set. Each tweet will be assigned a score that indicates if the tweet communicates a positive or negative sentiment of the tweeter. For this analysis a positive score will indicate that the user supports President Trump's actions/views and a negative score will indicate that the user is against the current US President and government.

The data set includes many attributes, but only relevant attribute required for calculating sentiment score is the 'text' column. Since the tweet text includes many irrelevant symbols and junks, each text needs to be cleaned to extract the actual tweet itself.

5.1 Data Cleaning

Tweets were cleaned in two different ways: 1) By leaving hashtags in the text 2) By removing hashtags in the text.

```
library("magrittr")
library(stringr)

clean_tweet = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tweets_data[, 1])
clean_tweet = gsub("&amp", "", clean_tweet)
clean_tweet = gsub("@\\w+", "", clean_tweet)
clean_tweet = gsub("[[:punct:]]", "", clean_tweet)
clean_tweet = gsub("[[:digit:]]", "", clean_tweet)
clean_tweet = gsub("http\\w+", "", clean_tweet)
clean_tweet = gsub("[ \t]{2,}", "", clean_tweet)
clean_tweet = gsub("^\\s+|\\s+$", "", clean_tweet)
clean_tweet = gsub("\\\n", " ", clean_tweet)
clean_tweet <- str_replace_all(clean_tweet, "#[a-z,A-Z]*", "")

tweets_data$Clean_Tweet_With_Hashtag <- clean_tweet

# Cleaning round 2 - by removing all words with hashtags

clean_tweet = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tweets_data[, 1])
clean_tweet <- str_replace_all(clean_tweet, "#[a-z,A-Z]*", "")
clean_tweet = gsub("&amp", "", clean_tweet)
clean_tweet = gsub("@\\w+", "", clean_tweet)
clean_tweet = gsub("[[:punct:]]", "", clean_tweet)
clean_tweet = gsub("[[:digit:]]", "", clean_tweet)
clean_tweet = gsub("http\\w+", "", clean_tweet)
clean_tweet = gsub("[ \t]{2,}", "", clean_tweet)
clean_tweet = gsub("^\\s+|\\s+$", "", clean_tweet)
```

```
clean_tweet = gsub("\\\n", " ", clean_tweet)

tweets_data$Clean_Tweet_Without_Hashtag <- clean_tweet
```

5.2 Calculate Sentiment Score

Few sentiment classification packages (sentimentr, doc2vec, syuzhet) were considered for scoring sentiment of each tweet. The selected package was 'sentimentr'. It calculates sentiment on the sentence level rather than counting number of negative and positive words in the sentence. For understanding the opinion of a user it is important to consider the entire sentence and its context when calculating the sentiment score. Sentiment score was calculated for texts with and without hashtags.

```
library(sentimentr)
library(exploratory)

tweets_data$Sentimentscore_withhashtag <- get_sentiment(tweets_data$Clean_Tweet_With_Has

tweets_data$Sentimentscore_withouthashtag <- get_sentiment(tweets_data$Clean_Tweet_Witho
```

5.3 Data Exploration

A correlation analysis was performed between the two sentiment scores. The results showed that the scores have a strong positive correlation. But mean sentiment score on state level showed that texts without hashtag gave more evenly distributed score (similar number of positive/supporting sentiments and negative scores/resisting sentiments). Texts with hashtag gave negative scores for all states except two. **Hence scores generated by texts without hashtag were selected to perform further analysis**.

```
library(ggplot2)

tweets_data <- tweets_data %>% rename(Score_withhashtag = Sentimentscore_withhashtag,
    Score_withouthashtag = Sentimentscore_withouthashtag)

# correlation analysis on two scores

tweets_data %>% select(Score_withhashtag, Score_withouthashtag) %>% cor()
```

```
##                       Score_withhashtag Score_withouthashtag
## Score_withhashtag             1.0000000            0.8766415
## Score_withouthashtag          0.8766415            1.0000000
```

```
tweets_data %>% group_by(State) %>% summarise(Score_withhashtag = mean(Score_withhashtag
    Score_withouthashtag = mean(Score_withouthashtag)) %>% ungroup() %>% arrange(State)
    knitr::kable()
```

State	Score_withhashtag	Score_withouthashtag
AK	-0.0731370	0.0532975
AL	-0.0424748	0.0425829
AR	-0.0221234	0.0514022
AZ	-0.0592239	0.0406680
CA	-0.0933581	-0.0082323
CO	-0.0758562	-0.0105061
CT	-0.0571407	-0.0259095
DC	-0.1158898	-0.0360464
DE	-0.1958914	-0.0465059
FL	-0.0952168	-0.0036908
GA	-0.0416976	0.0145738
IA	-0.1125796	0.0123380
IL	-0.1094912	-0.0209407
IN	-0.0920770	-0.0250201
KS	-0.0531349	-0.0031541
KY	-0.0743011	0.0205606
LA	0.0166136	0.0945410
MA	-0.0732510	-0.1209468
MD	-0.0944237	-0.0454428
ME	-0.1667858	-0.0454154
MI	-0.0231485	0.0666261
MN	-0.0904506	-0.0054573
MO	-0.0549110	-0.0355738
MS	-0.0245226	0.0031115
MT	-0.2950655	-0.2849411
NC	-0.0723026	0.0233658
ND	-0.1254186	-0.0383009
NE	-0.0339460	-0.0021047
NH	-0.1463817	-0.0166604
NJ	-0.0506692	0.0199407
NM	-0.1816457	-0.0347486
NV	-0.0185953	0.0102554
NY	-0.1163298	-0.0774077
OH	-0.0682914	-0.0175786
OK	-0.1349043	-0.0351753
OR	-0.1204114	-0.0063789
PA	-0.2235339	-0.1190241
RI	-0.0190434	0.0029412
SC	0.0380389	0.0850039

State	Score_withhashtag	Score_withouthashtag
SD	-0.0285864	0.0464558
TN	-0.1025077	-0.0274411
TX	-0.0907064	-0.0436198
UT	-0.0102990	0.0473816
VA	-0.0781225	-0.0150474
VT	-0.0831272	0.0054449
WA	-0.0263226	-0.0170811
WI	-0.1955501	-0.1089753
WV	-0.1256993	0.0191438
WY	-0.1208530	-0.0439097

```
library("magrittr")
# State level score distribution analysis
tweets_data %>% group_by(State) %>% summarise(Score_withhashtag = mean(Score_withhashtag
    ggplot(aes(State, Score_withhashtag)) + geom_bar(stat = "identity") + theme(axis.tex
    hjust = 1))
```

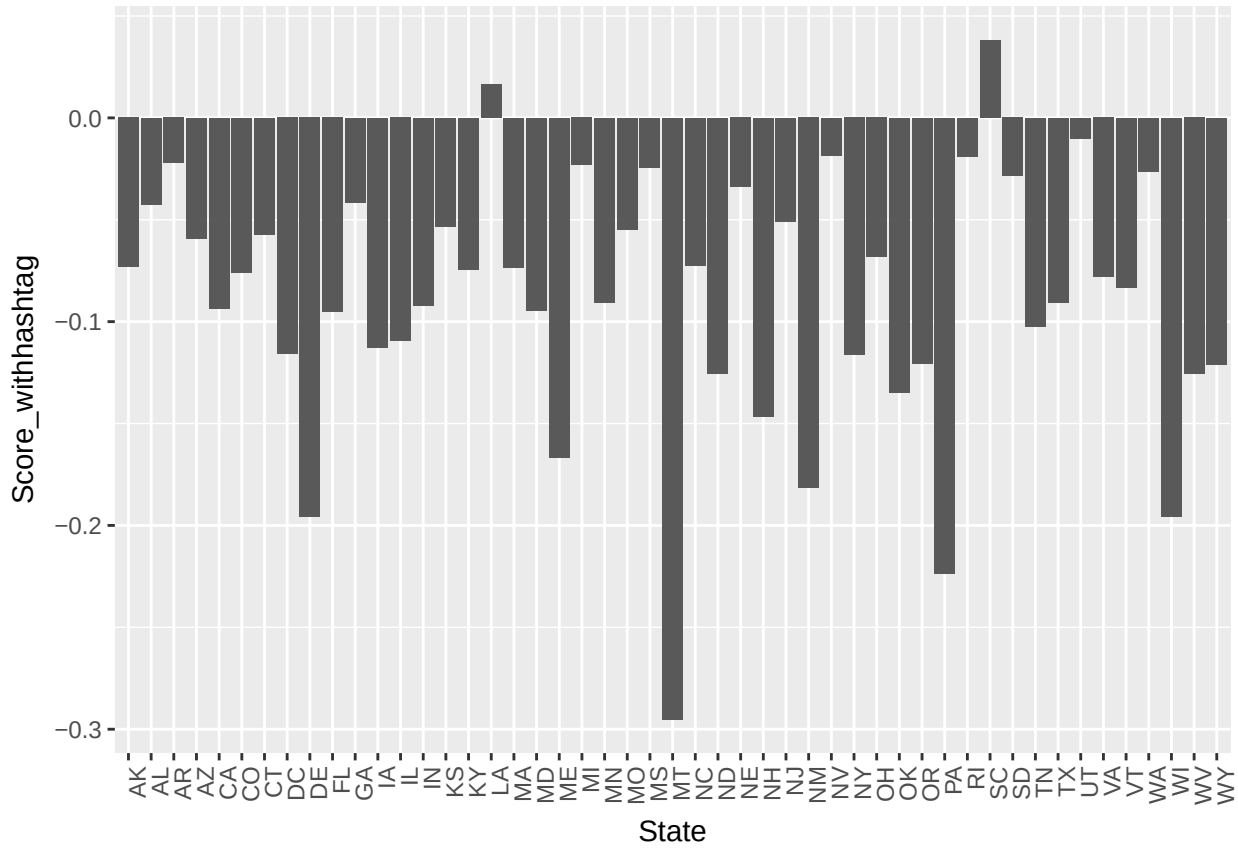

Figure 1: Sentiment Score Distribution - State vs Score_withhashtag

```
tweets_data %>% group_by(State) %>% summarise(Score_withouthashtag = mean(Score_withouth
    ggplot(aes(State, Score_withouthashtag)) + geom_bar(stat = "identity") + theme(axis
    hjust = 1))
```

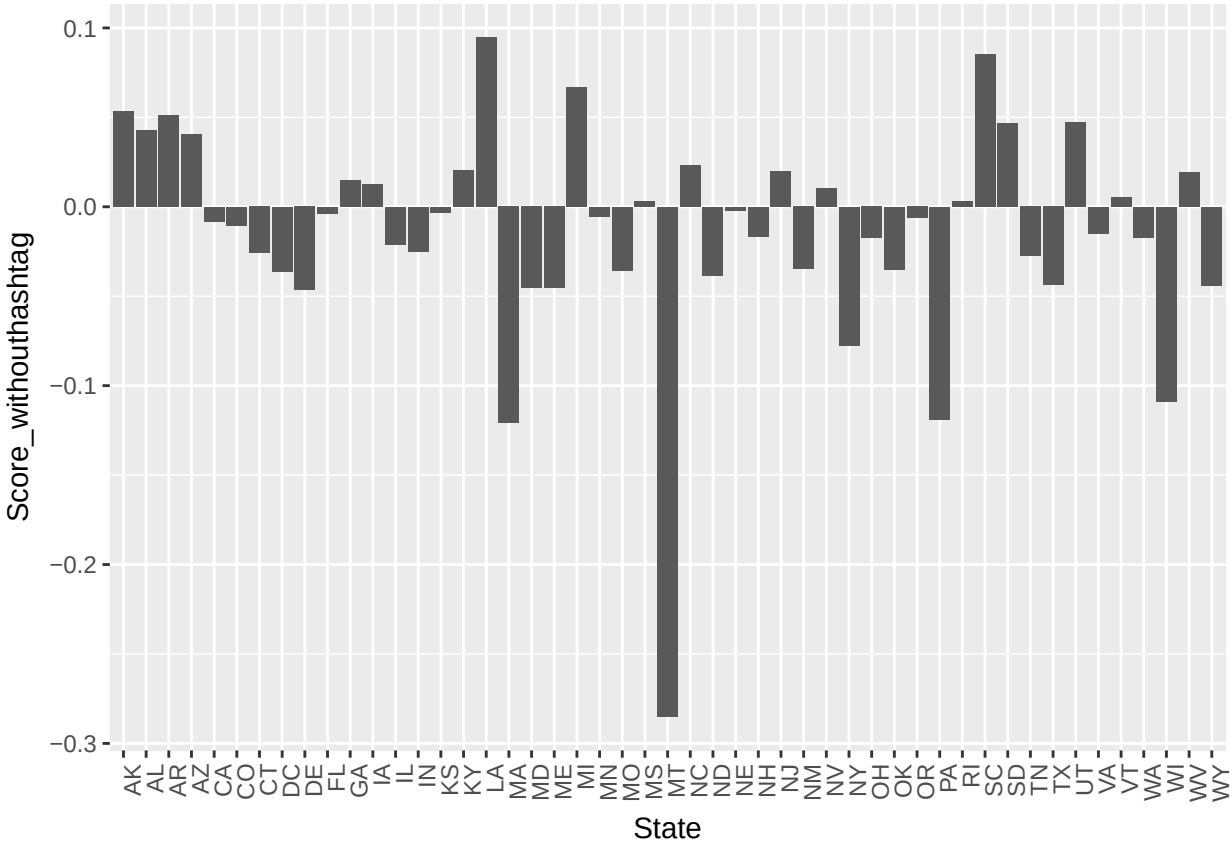

Figure 2: Sentiment Score Distribution - State vs Score_withouthashtag

6 US MAP WITH SENTIMENT SCORE

The selected score (**average of scores generated by texts without hashtag**) for each state needs to be plotted on US Map so that the states can be compared against the election results map. Since State Idaho is excluded from the analysis due to unavailability of tweets, it will be shaded in Gray on the map.

```
library(ggplot2)
library(maps)

# load map data for US States
all_states <- map_data("state")

# calculate state level sentiment score
sentiment_score <- tweets_data %>% group_by(State) %>% summarize(Score_withouthashtag =

tbl_df(sentiment_score)
```

```
## # A tibble: 49 x 2
##     State Score_withouthashtag
##    <fctr>                <dbl>
##  1     AK          0.053297494
##  2     AL          0.042582906
##  3     AR          0.051402166
##  4     AZ          0.040667998
##  5     CA         -0.008232348
##  6     CO         -0.010506056
##  7     CT         -0.025909477
##  8     DC         -0.036046405
##  9     DE         -0.046505948
## 10     FL         -0.003690799
## # ... with 39 more rows
```

```
# get state names for each state code
sentiment_score$State <- state.name[match(sentiment_score$State, state.abb)]

# State name was not retrieved for DC. So it needs to be added seperately
sentiment_score$State[8] <- "District of Columbia"

# view the dataframe
tbl_df(sentiment_score)
```

```
## # A tibble: 49 x 2
##                   State Score_withouthashtag
##                   <chr>                <dbl>
##  1               Alaska          0.053297494
```

```
##  2              Alabama          0.042582906
##  3             Arkansas          0.051402166
##  4              Arizona          0.040667998
##  5           California         -0.008232348
##  6             Colorado         -0.010506056
##  7          Connecticut         -0.025909477
##  8 District of Columbia         -0.036046405
##  9             Delaware         -0.046505948
## 10              Florida         -0.003690799
## # ... with 39 more rows
```

```
# map_data() does not return data for Alaska. Hence it needs to be removed from
# sentiment_score dataframe before it is merged with map_data() results
sentiment_score <- sentiment_score[sentiment_score$State != "Alaska", ]

# Since Idaho was missed in the data extraction process it needs to be added to
# the dataset with a neutral score
sentiment_score <- rbind(sentiment_score, c("Idaho", NA))

sentiment_score <- sentiment_score %>% mutate(Score_withouthashtag = as.numeric(Score_wi

sentiment_score <- sentiment_score[order(sentiment_score$State), ]

sentiment_score_map <- sentiment_score

sentiment_score_map <- sentiment_score_map %>% mutate(State = tolower(State))

sentiment_score_map$region <- sentiment_score_map$State

sentiment_score_map <- merge(all_states, sentiment_score_map, by = "region")
```

```
library(ggplot2)
p <- ggplot()  # plot the data
p <- p + geom_polygon(data = sentiment_score_map, aes(x = long, y = lat, group = group,
    fill = sentiment_score_map$Score_withouthashtag), colour = "white") + scale_fill_cor
    high = "darkred", guide = "colorbar", trans = "reverse")
P1 <- p + theme_bw() + labs(fill = "Sentiment towards President Trump \n(Negative Score
    title = "2017 Presidential Actions - Sentiment Scores for US States (excluding Idaho
    x = "", y = "")
P1 + scale_y_continuous(breaks = c()) + scale_x_continuous(breaks = c()) + theme(panel.
```

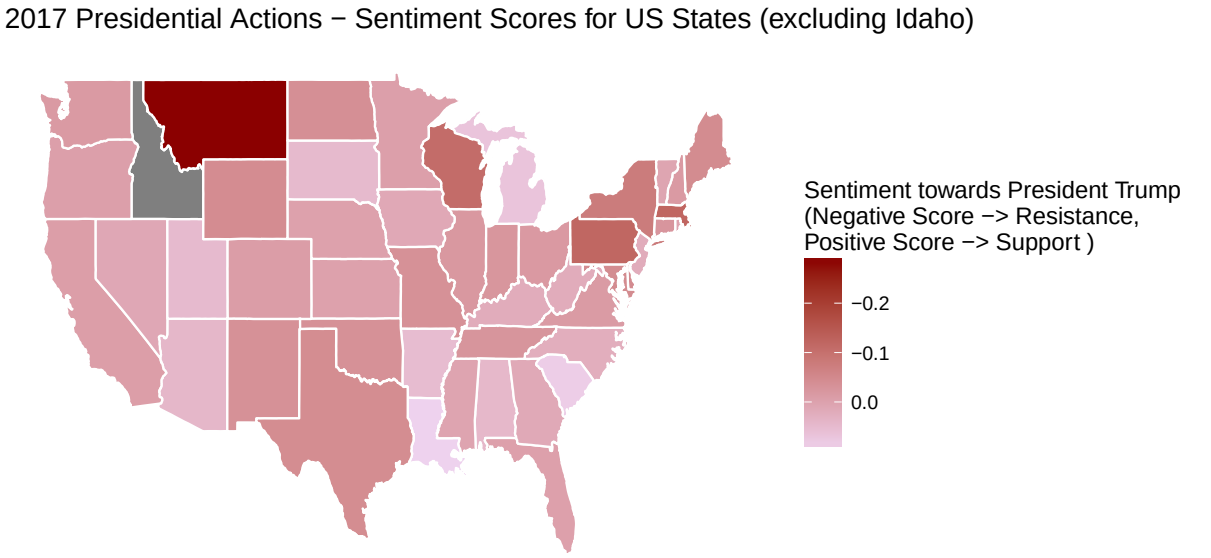

Figure 3: US States Map with Sentiment Score level

7 STATISTICAL ANALYSIS

7.1 Regression Model

Regression analysis on sentiment score for each state and popular vote election results for each state will help explain if there is a relation between a state's people's voting trend and their sentiment towards the president's actions post election. Trump's Victory Margin for each state was taken for the analysis.

The Election results data set contains total number of popular votes received for each state by Clinton, Trump and other presidential candidates of 2016 US Election. Trump's Victory Margin for each state was calculated by taking the difference between total number of votes received by him and total number of votes received by the winning person (or if Trump is the winner for the state, the difference between his votes and the second place candidate's votes).

```
# read popular vote election dataset

ElectionResults_data = readr::read_csv(file = "2016 National Popular Vote Tracker.csv")
```

```
## Parsed with column specification:
## cols(
##   State = col_character(),
##   `Clinton (D)` = col_number(),
##   `Trump (R)` = col_number(),
##   Others = col_number(),
```

```
##   `Clinton %` = col_character(),
##   `Trump %` = col_character(),
##   `Others %` = col_character(),
##   `Total '16 Votes` = col_number(),
##   VictoryMargin = col_character()
## )
```

```
tbl_df(ElectionResults_data) %>% knitr::kable()
```

State	Clinton (D)	Trump (R)	Others	Clinton %	Trump %	Others %	Total '16
Alabama	729547	1318255	75570	34.40%	62.10%	3.60%	212
Arizona	1161167	1252401	159597	45.10%	48.70%	6.20%	257
Arkansas	380494	684872	65269	33.70%	60.60%	5.80%	113
California	8753788	4483810	943997	61.70%	31.60%	6.70%	1418
Colorado	1338870	1202484	238866	48.20%	43.30%	8.60%	278
Connecticut	897572	673215	74133	54.60%	40.90%	4.50%	164
Delaware	235603	185127	20860	53.40%	41.90%	4.70%	44
District of Columbia	282830	12723	15715	90.90%	4.10%	5.00%	31
Florida	4504975	4617886	297178	47.80%	49.00%	3.20%	942
Georgia	1877963	2089104	125306	45.90%	51.00%	3.10%	409
Idaho	189765	409055	91435	27.50%	59.30%	13.20%	69
Illinois	3090729	2146015	299680	55.80%	38.80%	5.40%	553
Indiana	1033126	1557286	144546	37.80%	56.90%	5.30%	273
Iowa	653669	800983	111379	41.70%	51.10%	7.10%	156
Kansas	427005	671018	86379	36.10%	56.70%	7.30%	118
Kentucky	628854	1202971	92324	32.70%	62.50%	4.80%	192
Louisiana	780154	1178638	70240	38.40%	58.10%	3.50%	202
Maine	357735	335593	54599	47.80%	44.90%	7.30%	74
Maryland	1677928	943169	160349	60.30%	33.90%	5.80%	278
Massachusetts	1995196	1090893	238957	60.00%	32.80%	7.20%	332
Michigan	2268839	2279543	250902	47.30%	47.50%	5.20%	479
Minnesota	1367716	1322951	254146	46.40%	44.90%	8.60%	294
Mississippi	485131	700714	23512	40.10%	57.90%	1.90%	120
Missouri	1071068	1594511	143026	38.10%	56.80%	5.10%	280
Montana	177709	279240	40198	35.70%	56.20%	8.10%	49
Nebraska	284494	495961	63772	33.70%	58.70%	7.60%	84
Nevada	539260	512058	74067	47.90%	45.50%	6.60%	112
New Hampshire	348526	345790	49842	46.80%	46.50%	6.70%	74
New Jersey	2148278	1601933	123835	55.50%	41.40%	3.20%	387
New Mexico	385234	319666	93418	48.30%	40.00%	11.70%	79
New York	4556124	2819534	345795	59.00%	36.50%	4.50%	772
North Carolina	2189316	2362631	189617	46.20%	49.80%	4.00%	474
North Dakota	93758	216794	33808	27.20%	63.00%	9.80%	34
Ohio	2394164	2841005	261318	43.60%	51.70%	4.80%	549
Oklahoma	420375	949136	83481	28.90%	65.30%	5.70%	145

State	Clinton (D)	Trump (R)	Others	Clinton %	Trump %	Others %	Total '16
Oregon	1002106	782403	216827	50.10%	39.10%	10.80%	200
Pennsylvania	2926441	2970733	218228	47.90%	48.60%	3.60%	611
Rhode Island	252525	180543	31076	54.40%	38.90%	6.70%	46
South Carolina	855373	1155389	92265	40.70%	54.90%	4.40%	210
South Dakota	117458	227721	24914	31.70%	61.50%	6.70%	37
Tennessee	870695	1522925	114407	34.70%	60.70%	4.60%	250
Texas	3877868	4685047	406311	43.20%	52.20%	4.50%	896
Utah	310676	515231	305523	27.50%	45.50%	27.00%	113
Vermont	178573	95369	41125	56.70%	30.30%	13.10%	31
Virginia	1981473	1769443	231836	49.80%	44.40%	5.80%	398
Washington	1742718	1221747	401179	51.80%	36.30%	11.90%	336
West Virginia	188794	489371	34886	26.50%	68.60%	4.90%	71
Wisconsin	1382536	1405284	188330	46.50%	47.20%	6.30%	297
Wyoming	55973	174419	25457	21.90%	68.20%	10.00%	25

```
ElectionResults_data[, 9] <- as.numeric(sub("%", "", ElectionResults_data[[9]], fixed =

RegressionAnalysis_data <- merge(sentiment_score, ElectionResults_data, by = "State")

# Remove the outliers

# Montana Score = -0.285

# DC VicotryMargin = -0.868

RegressionAnalysis_data_final <- RegressionAnalysis_data %>% filter(!(State %in%
    c("Montana", "District of Columbia")))

# perform regression analysis

Model <- lm(formula = Score_withouthashtag ~ VictoryMargin, data = RegressionAnalysis_da
summary(Model)
```

```
##
## Call:
## lm(formula = Score_withouthashtag ~ VictoryMargin, data = RegressionAnalysis_data_fin
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.107239 -0.026371  0.003121  0.026138  0.092801
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   -0.0122861  0.0067768  -1.813   0.0767 .
## VictoryMargin  0.0007156  0.0003348   2.137   0.0382 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04442 on 44 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.09406,    Adjusted R-squared:  0.07347
## F-statistic: 4.568 on 1 and 44 DF,  p-value: 0.03817
```

```
library(ggplot2)
library("magrittr")
ggplot(RegressionAnalysis_data_final, aes(x = VictoryMargin, y = Score_withouthashtag))
    geom_point() + geom_point(data = RegressionAnalysis_data %>% filter(State ==
    "Montana"), colour = "blue") + geom_point(data = RegressionAnalysis_data %>%
    filter(State == "District of Columbia"), colour = "blue") + stat_smooth(method = "lm
    col = "red")  # plot regression analysis results
```

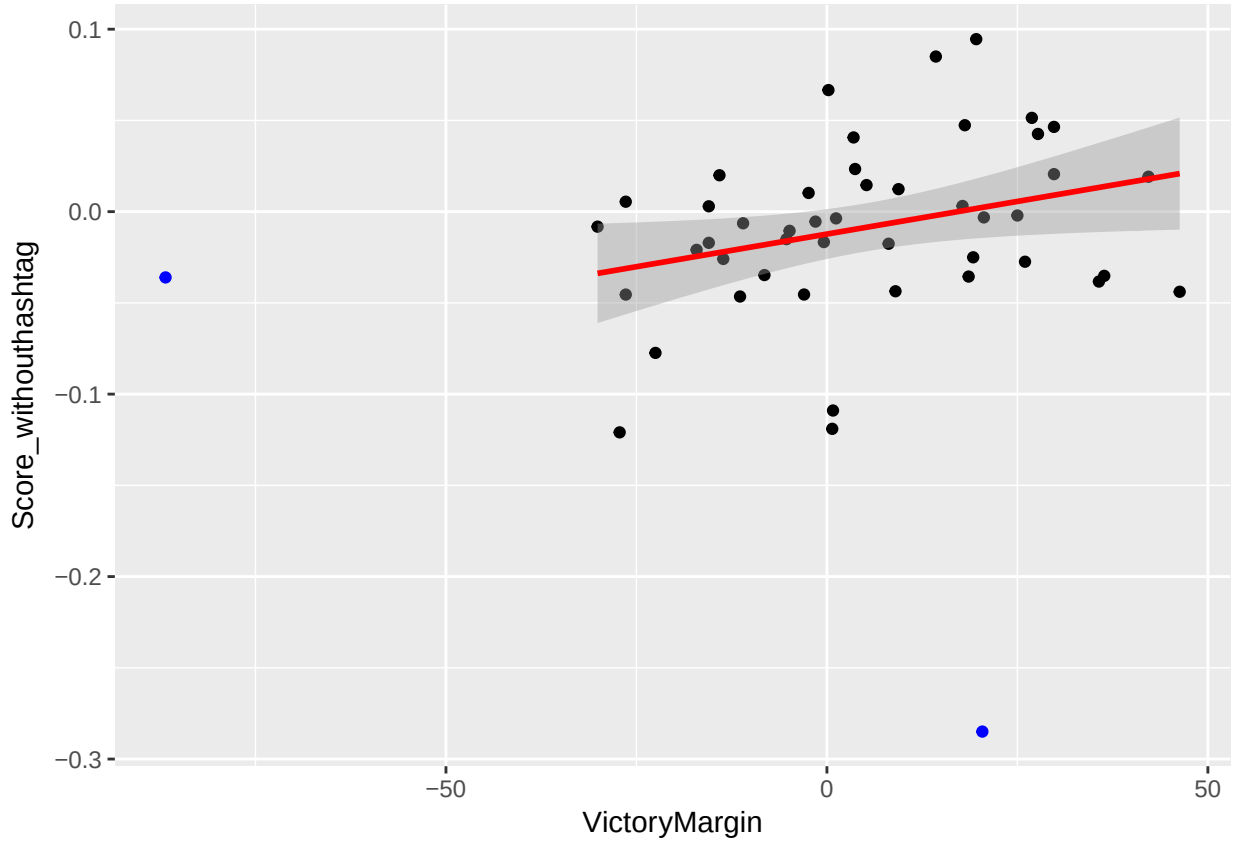

Figure 4: Sentiment Score vs VictoryMargin - Regression Line

The Coefficient Estimate for VictoryMargin is 0.00071. It is the slope of the regression line, indicating the effect VictoryMargin has on Sentiment Score.

Multiple R-squared value is 0.09406 and Adjusted R-squared is 0.07347. The R2 value is a measure of the linear relationship between the predictor variable (Sentiment Score) and the response / target variable (VictoryMargin). It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In multiple regression settings, the R2 will always increase as more variables are included in the model. In that case the adjusted R2 is the preferred measure to consider as it adjusts for the number of variables considered. In analysis performed here only one independent variable (VictoryMargin) is used, hence will be considering Multiple R-squared/R-Squared value. Here R-squared value is 0.09406, indicating that roughly 9.4% of the variance found in the Sentiment Score can be explained by VictoryMargin.

A small p-value indicates that it is unlikely we will observe a relationship between the predictor variable and response variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In the above result the p-value is less than the alpha value .05, indicating the relationship between VictoryMargin and Sentiment Score is significant. The 'signif. Codes' associated to VictoryMargin estimate also indicates significance level. One star (or asterisks) represents a low significant p-value.

8 DISCUSSION

There are few limitations to the data set used in this analysis.

1. Search terms were biased towards resistance against actions and views of President Trump and his government. This could have resulted in more tweets with negative sentiments and thereby more states with overall negative sentiment score. Had there been more support specific search terms included, the results would not have been the same.

2. Twitter Rest API does not return geo locations of all the tweets.Also it returns only tweets from past 6-8 days. The geo attributes are populated only if the user of the tweet opted to disclose the location of the user. 99.9% of extracted tweets didn't have data populated for longitude and latitude attributes. Due to this limitation a manual search process based on radius around a geocode was followed to extract tweets from different states. This manual search resulted in an uneven data population from different states. Hence the overall sentiment calculated for the States would not represent the true opinion of the people in the States. The manual search effort also impacted timeline of tweet extraction. Tweets from all the states were not extracted at the same time frame. Hence the results of analysis done here can not represent the true intent of the actual population.

9 CONCLUSION

Montana state shows the highest resistance and Louisiana shows the highest support to President's actions/views. A regression analysis on these variables explain that only 9.4% of the variation in Sentiment Score can be explained a linear relationship with the VictoryMargin. Consequently, even though small, a p-value less than .05 indicates that we can reject the null hypothesis which allows us to conclude that there is a significant linear relationship between VictoryMargin and Sentiment Score. Hence the tweets collected show that the sentiment towards President Trump has not changed much after the election.

9.1 APPENDIX

9.1.1 Geocodes and radius used for all States and Cities

State	City	Latitude	Longitude	Radius in Miles
AK	Alaska	61.2876	-149.4869	400
AL	Birmingham	33.4564	-86.8019	100
AL	Montgomery	32.3569	-86.2578	100
AR	Conway	34.2089	-91.9859	120
AR	Jonesboro	35.8358	-90.623	20
AZ	Phoenix	33.7039	-112.1871	120
CA	Sacramento	38.3774	-121.4444	100
CA	Redding	40.6244	-122.3076	100
CA	Los Angeles	33.7865	-118.2986	130
CO	Denver	39.6606	-104.7627	70
CO	Grand Junction	39.0891	-108.5665	20
CT	Hartford	41.7663	-72.6746	19
CT	New Haven	41.3657	-72.9275	15
DC	DC	38.9526	-77.0178	5
DE	Newark	39.6147	-75.7012	2
DE	Wilmington	39.7585	-75.5687	2
DE	Port Penn	39.5129	-75.585	2
DE	Dover	39.1086	-75.448	5
DE	Georgetown	38.6328	-75.3342	5
FL	Tampa	27.9466	-82.4272	120
FL	Miami	25.5584	-80.4581	140
FL	Panama	30.2051	-85.6688	75
FL	Jacksonville	30.3375	-81.7686	25
FL	Gainesville	29.6778	-82.4663	90
GA	Atlanta	33.7978	-84.3877	55
GA	Athens	33.9519	-83.3576	35
GA	Macon	32.8279	-83.595	90
GA	Columbus	32.4934	-84.9532	5
GA	Augusta	33.4166	-82.0559	5
GA	Albany	31.5592	-84.1765	50
GA	Brunswick	31.2219	-81.4825	30
GA	Savannah	31.9713	-81.0715	20
IA	Des Moines	41.6433	-93.6213	78
IA	Iowa City	41.6426	-91.5999	50
IA	Mason City	43.1164	-93.2705	25
IA	Le Mars	42.7491	-96.2617	17
IA	Sioux City	42.471	-96.3384	5
IL	Chicago	41.8119	-87.6873	20
IL	Evanston	42.0445	-87.6879	40

State	City	Latitude	Longitude	Radius in Miles
IL	Bloomington	40.5192	-88.8643	80
IL	Rockford	42.284	-89.0162	15
IL	Marion	37.7295	-88.9128	30
IN	Indianapolis	39.7794	-86.1328	80
IN	Fort Wayne	41.0938	-85.1841	20
IN	Westville	41.5499	-86.7429	30
KS	Overland Park	39.0089	-94.7863	5
KS	Manhattan	39.1774	-96.5551	100
KS	Wichita	37.6915	-97.3167	50
KS	Hays	38.8765	-99.3185	80
KY	Lexington	38.0463	-84.4973	75
KY	Florence	39.0003	-84.6251	10
KY	Louisville	38.0227	-85.3368	30
KY	CampbellsVille	37.3341	-85.3602	60
KY	Cave City	37.1344	-85.9727	40
LA	Louisana	30.2248	-91.4902	130
LA	Alexandria	31.2944	-92.5781	55
LA	Ruston	32.3073	-92.5204	100
MA	Boston	42.3637	-71.3626	29
MA	NorthHampton	42.434	-72.7405	30
MA	Rockport	42.6121	-70.6933	30
MA	Plymouth	41.9443	-70.666	30
MA	Worcester	42.383	-71.8098	30
MD	Baltimore	39.2858	-76.6248	34
MD	Frederick	39.5233	-77.4105	25
MD	Ocean City	38.2698	-75.404	30
ME	Bangor	44.6808	-69.1843	100
ME	Portland	43.651	-70.2243	40
MI	Detroit	42.731	-84.5388	90
MI	Grand Rapids	42.9528	-85.2931	80
MI	Petoskey	45.3672	-84.9458	150
MI	Marquette	46.5485	-87.4213	150
MN	Minneapolis	45.1431	-94.6977	100
MN	Rochester	44.0055	-92.4716	42
MN	Marshall	44.4692	-94.9417	100
MN	Brainerd	46.3382	-94.1872	100
MO	Columbia	38.9466	-92.3434	125
MO	Lebanon	37.6609	-92.6495	90
MO	Brookfield	39.7821	-93.0744	95
MS	Jackson	32.3129	-89.6661	45
MS	Hattiesburg	31.4046	-89.1717	40
MS	Wiggins	30.8466	-89.1406	40
MS	Winona	33.4751	-89.7303	90

State	City	Latitude	Longitude	Radius in Miles
MS	Oxford	34.1673	-89.5316	70
MT	Lewistown	47.0249	-109.3877	250
NC	Mount Olive	47.0249	-109.3877	120
NC	Asheboro	35.7078	-79.8074	90
NC	Asheville	35.5887	-82.554	35
ND	McClusky	47.458	-99.9342	200
NE	Ord	41.6041	-98.9217	150
NE	Lincoln	40.7275	-96.8416	60
NE	Alliance	42.0816	-102.8585	80
NH	Concord	43.3016	-71.5764	50
NH	Laconia	43.5277	-71.5142	40
NJ	Newark	40.7338	-74.1656	8
NJ	Vineland	39.4787	-75.0216	35
NJ	Toms River	39.9519	-74.2049	40
NJ	Marlboro	40.3368	-74.2736	25
NJ	Branchburg	40.5848	-74.681	25
NJ	Patterson	40.9129	-74.1802	18
NJ	Jefferson	41.0028	-74.5611	30
NM	Claunch	34.1464	-105.9985	250
NV	Las Vegas	35.9279	-114.972	40
NV	Carson City	39.2025	-119.7526	20
NV	Ely	39.3141	-114.8404	40
NV	Battle Mountain	40.0421	-116.9748	100
NY	Queens	40.7388	-73.79	10
NY	Shirley	40.9223	-72.637	30
NY	New Rochelle	40.9482	-73.7953	30
NY	JFK	40.6645	-73.7559	15
NY	Syracuse	43.0764	-76.1099	73
NY	Rochester	43.1663	-77.6029	73
NY	Watertown	44.0725	-76.0165	38
NY	Albany	42.7197	-73.8206	15
NY	Kansas City	39.1163	-94.688	5
OH	Columbus	40.3721	-82.7587	120
OK	Norman	35.7443	-97.334	140
OR	Portland	45.5194	-122.6901	40
OR	Salem	44.9935	-123.1074	40
OR	Eugene	44.0323	-123.0957	40
OR	Roseburg	43.0535	-123.3608	40
OR	Medford	43.3331	-123.3256	25
OR	Bend	43.843	-121.5764	25
PA	Philadelphia	40.1938	-75.4893	25
PA	Pittsburg	40.8701	-78.7558	110
PA	Harrisburg	41.0592	-77.0805	100

State	City	Latitude	Longitude	Radius in Miles
RI	Providence	41.8643	-71.5649	11
RI	Newport	41.5937	-71.4203	17
SC	Columbia	33.7124	-80.5535	80
SC	Clinton	34.5415	-81.8133	60
TN	Memphis	35.5822	-89.1055	70
TN	Nashville	35.8409	-87.6709	70
TN	Westel	35.8362	-84.6983	70
TN	Greenville	36.2107	-83.0822	30
TX	Austin	30.3263	-97.7712	200
TX	Dallas	32.8338	-96.7715	72
TX	Houston	29.8339	-95.4342	100
TX	Corpus Christi	27.732	-97.3851	100
UT	Salt Lake City	40.7838	-111.8771	50
UT	Brigham City	41.5046	-112.0602	40
UT	Provo	39.8775	-111.716	100
VA	Richmond	37.5303	-77.4448	60
VA	Norfolk	37.0745	-76.3132	40
VA	Lynchburg	37.2896	-79.1074	50
VA	Alexandria	38.8165	-77.3514	20
VA	Culpeper	38.4646	-77.9904	40
VT	Burlington	44.5856	-72.522	50
VT	Granville	43.9945	-72.7839	40
VT	Killington	43.6788	-72.7939	25
WA	Seattle	47.4322	-121.8033	100
WA	Yakima	46.6287	-120.5739	90
WA	Walla Walla	46.1341	-118.2914	90
WA	Spokane	47.7805	-117.4553	20
WA	Kennewick	46.2123	-119.1556	40
WA	Colville	48.6769	117.8093	40
WI	Milwakee	43.505	-88.2182	90
WI	Holcombe	45.1545	-91.1735	70
WI	Spirit Falls	45.3427	-89.6657	65
WI	Iron River	46.4311	-91.2513	50
WV	Charleston	38.1774	-81.3888	70
WV	Philip	39.0928	-80.39	60
WY	Jackson	43.035	-106.8283	180