Data Appendix

I. Dataset 1: “hashtag_donaldtrump.csv”

In the “hashtag_donaldtrump.csv” derived from Kaggle, there were 9 variables remaining after the dataset was parsed through and cleaned. Each row in the dataset represents a tweet towards Donald Trump and the unit of observation is “tweet.”

The columns of the data set, uncleaned:

Column	Description	Example Responses
created_at	Date and time of tweet creation	2020-10-15 00:00:01
tweet_id	Unique ID of the tweet	1.316529221557252e+18
tweet	Full tweet text	2 hours since last tweet from #Trump! Maybe he is VERY busy. Tremendously busy.
likes	Number of likes	2.0
retweet_count	Number of retweets	1.0
source	Utility used to post tweet	Twitter Web App
user_id	User ID of tweet creator	8436472.0
user_name	Username of tweet creator	snarke
user_screen_name	Screen name of tweet creator	snarke
user_description	Description of self by tweet creator	Will mock for food! Freelance writer, blogger, commentator. Civics nerd. She/Her
user_join_date	Join date of tweet creator	2007-08-26 05:56:11
user_followers_count	Followers count on tweet creator	1185.0
user_location	Location given on tweet creator's profile	Portland
lat	Latitude parsed from user_location	45.5202471
long	Longitude parsed from user_location	-122.6741949

city	City parsed from user_location	Portland
country	Country parsed from user_location	United States of America
state	State parsed from user_location	Oregon
state_code	State code parsed from user_location	OR
collected_at	Date and time tweet data was mined from twitter	2020-10-21 00:00:00.746433060

The data set, after cleaning:

Column	Description	Example Responses
created_at	Date and time of tweet creation	2020-10-15 00:00:01
tweet	Full tweet text	2 hours since last tweet from #Trump! Maybe he is VERY busy. Tremendously busy.
likes	Number of likes	2.0
retweet_count	Number of retweets	1.0
user_name	Username of tweet creator	snarke
user_screen_name	Screen name of tweet creator	snarke
user_join_date	Join date of tweet creator	2007-08-26 05:56:11
user_followers_count	Followers count on tweet creator	1185.0
state	State parsed from user_location	Oregon
state_code	State code parsed from user_location	OR
collected_at	Date and time tweet data was mined from twitter	2020-10-21 00:00:00.746433060

II. Dataset 2: “hashtag_joebiden.csv”

Next, the dataset titled “hashtag_joebiden.csv” also derived from Kaggle underwent the same cleaning process. Each row of the data file represents a tweet directed towards the presidential candidate, Joe Biden. The unit of observation is “tweet.”

The data set, after cleaning:

Column	Description	Example Responses
created_at	Date and time of tweet creation	2020-10-15 00:00:01
tweet	Full tweet text	2 hours since last tweet from #Trump! Maybe he is VERY busy. Tremendously busy.
likes	Number of likes	2.0
retweet_count	Number of retweets	1.0
user_name	Username of tweet creator	snarke
user_screen_name	Screen name of tweet creator	snarke
user_join_date	Join date of tweet creator	2007-08-26 05:56:11
user_followers_count	Followers count on tweet creator	1185.0
state	State parsed from user_location	Oregon
state_code	State code parsed from user_location	OR
collected_at	Date and time tweet data was mined from twitter	2020-10-21 00:00:00.746433060

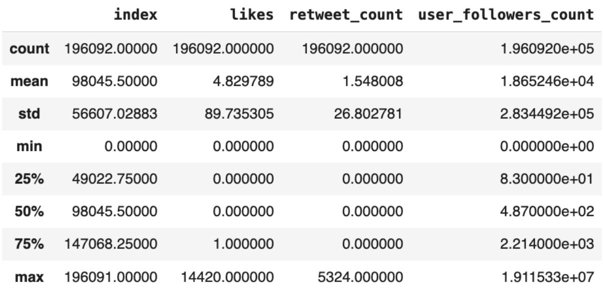
III. Variables in the Analyzed Datasets

The “hashtag_donaldtrump.csv” and “hashtag_joebiden.csv” have the same variables:

- created_at - The date and time of tweet creation
- tweet- The full tweet text (This is the text data analyzed for the purpose of this project)
- likes - The number of likes
- retweet_count - The number of retweets
- user_name - The username of tweet creator
- user_screen_name - The screen name of tweet creator
- user_followers_count - The followers count on tweet creator
- state - The state parsed from user_location
- state_code - The state code parsed from user_location

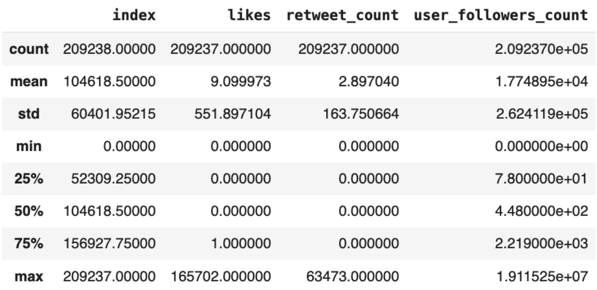
IV. Summary Statistics on Numerical Variables

For Trump cleaned dataframe:



	index	likes	retweet_count	user_followers_count
count	196092.00000	196092.000000	196092.000000	1.960920e+05
mean	98045.50000	4.829789	1.548008	1.865246e+04
std	56607.02883	89.735305	26.802781	2.834492e+05
min	0.00000	0.000000	0.000000	0.000000e+00
25%	49022.75000	0.000000	0.000000	8.300000e+01
50%	98045.50000	0.000000	0.000000	4.870000e+02
75%	147068.25000	1.000000	0.000000	2.214000e+03
max	196091.00000	14420.000000	5324.000000	1.911533e+07

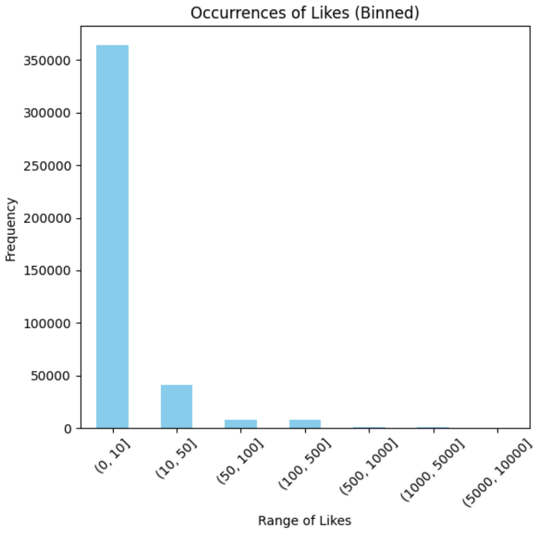
Biden cleaned dataframe:

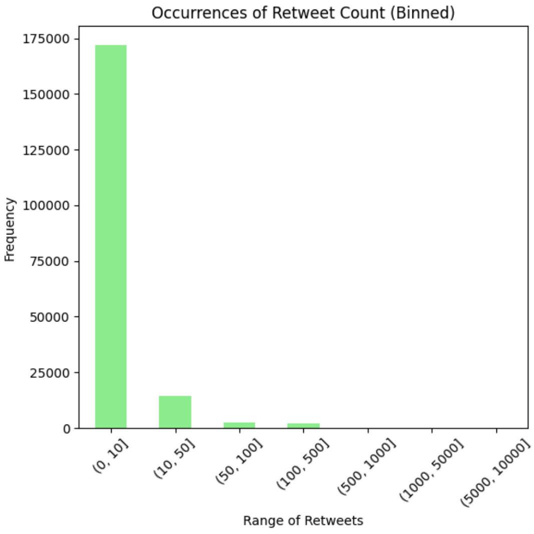


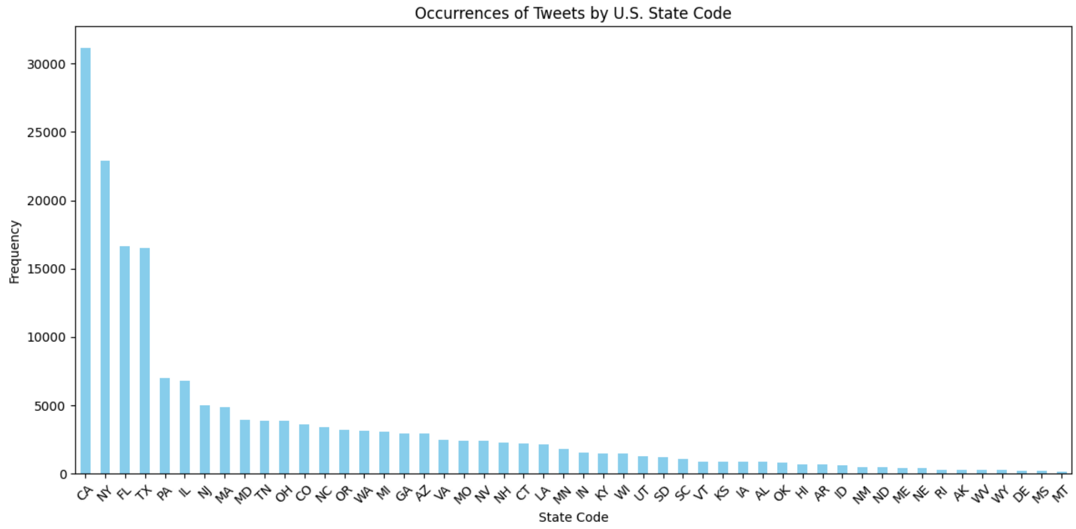
	index	likes	retweet_count	user_followers_count
count	209238.00000	209237.000000	209237.000000	2.092370e+05
mean	104618.50000	9.099973	2.897040	1.774895e+04
std	60401.95215	551.897104	163.750664	2.624119e+05
min	0.00000	0.000000	0.000000	0.000000e+00
25%	52309.25000	0.000000	0.000000	7.800000e+01
50%	104618.50000	0.000000	0.000000	4.480000e+02
75%	156927.75000	1.000000	0.000000	2.219000e+03
max	209237.00000	165702.000000	63473.000000	1.911525e+07

V. Frequency Tables on Categorical and Numerical Variables

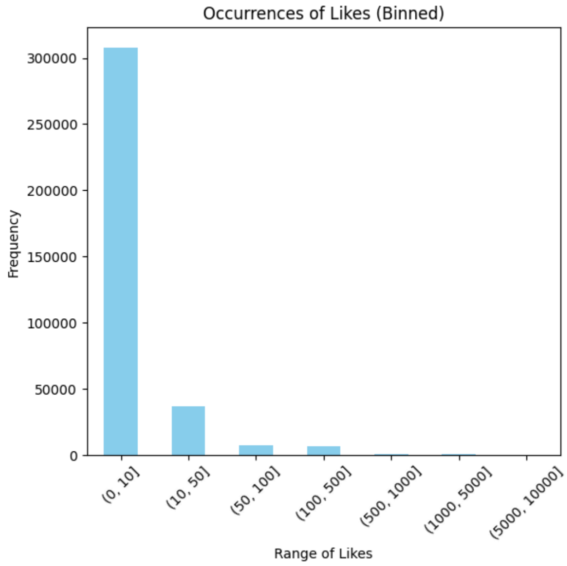
For Trump cleaned dataframe:

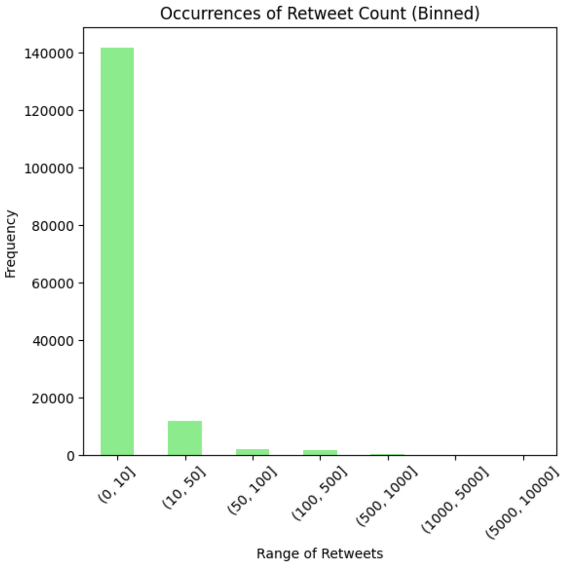


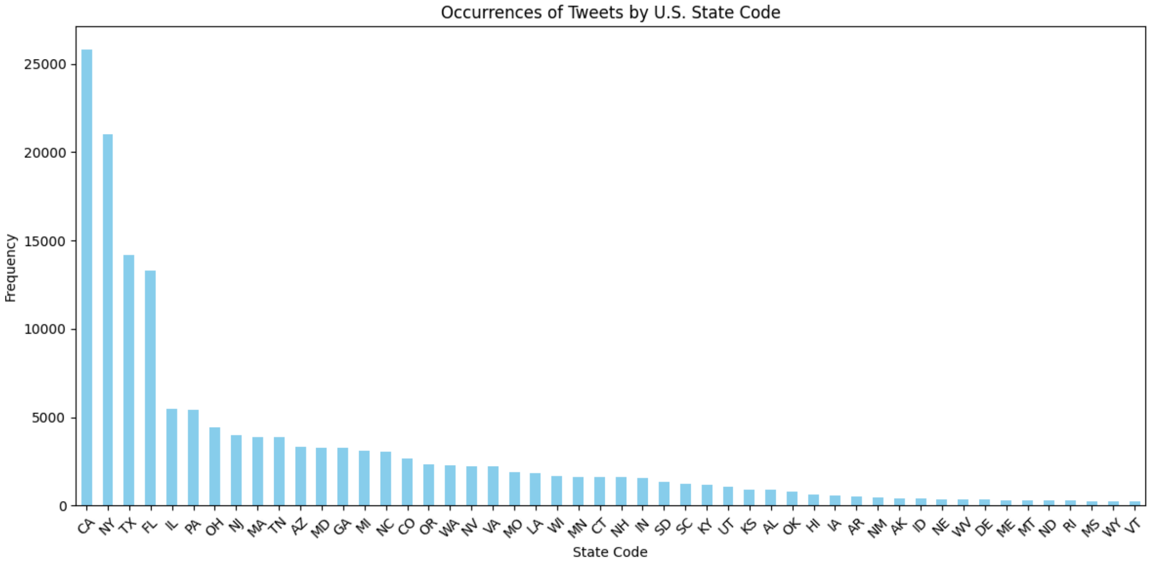




For Biden cleaned dataframe:







VI. Coding Scheme

Data preprocessing began by importing the raw CSV's and cleaning the data by dropping unnecessary columns and variables with significant portions of null values. Then, the tweets were filtered to only include those from U.S. states. Preliminary exploratory data analysis was conducted to help gain a better understanding of major patterns and anomalies in the dataset. Then using VADER, sentiment scores (negative, neutral, positive, and compound) were calculated for each tweet with the distributions visualized through various graphs. For text analysis, tweets were further cleaned by removing special characters and stopwords, and then lemmatized. TF-IDF vectorization was then used to identify the most significant words, and graphing the unique high-ranking words for each candidate.

VII. Exploratory Data Analysis

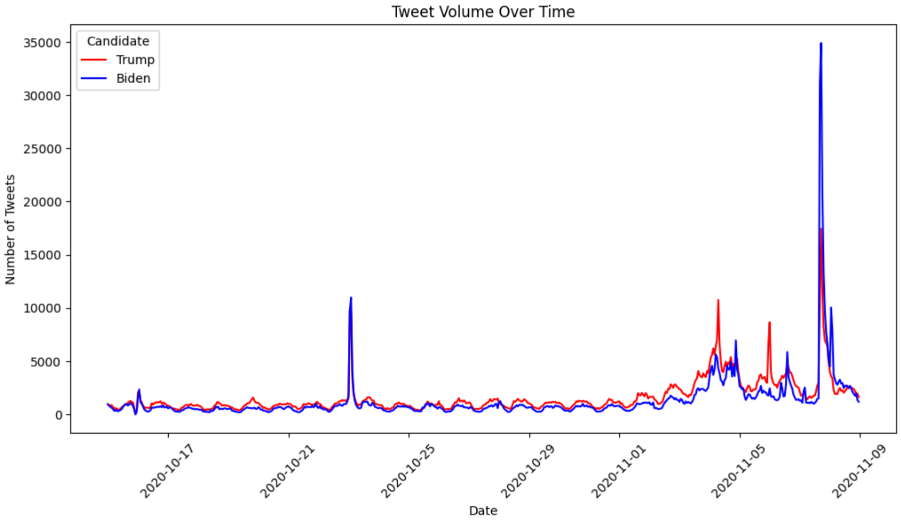


Figure 1. Tweet Volume Over Time

Figure 1 depicts the number of tweets for Biden and Trump in the weeks prior to Election Day. From 10/17/2020 to 11/1/2020, there is not a big difference in the number of tweets towards each candidate. However, after 11/1/2020, there is a small increase in the number of tweets per candidate and the candidates teeter back and forth on who has the most tweets. The graph features a notable peak on 11/08/2020 with a significant number of tweets towards Biden, essentially doubling the number of tweets towards Trump.

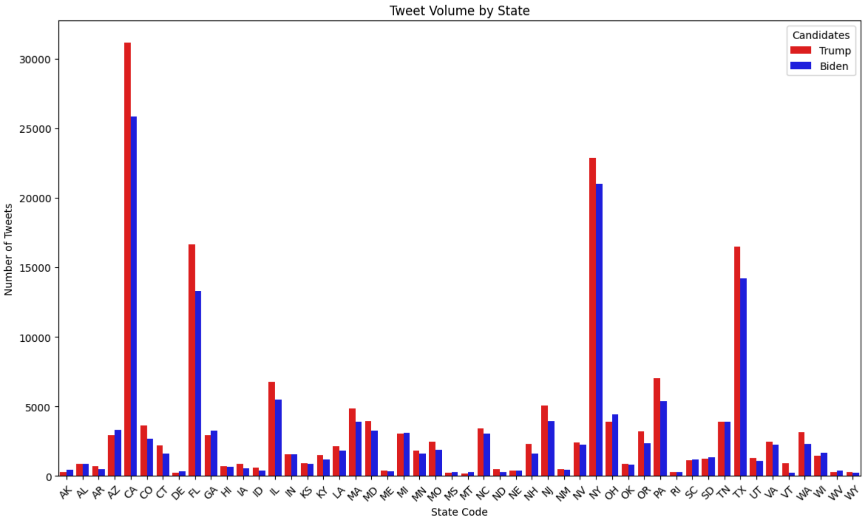


Figure 2. Tweet Volume by State

Figure 2 demonstrates the number of tweets for both candidates again but separated by state. This figure is useful in understanding which states are more involved in the political conversation and voicing their political opinion. Namely, California, Florida, New York, and

Texas stand out and have the highest volume of tweets. It is important to note that these states also happen to be the states with the most electoral votes in the Electoral College.

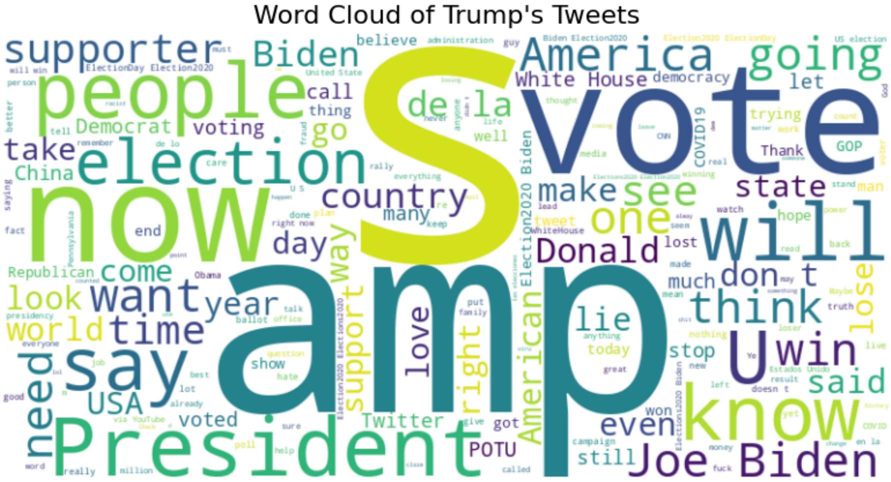


Figure 3. Word Cloud of Trump's Tweets

In figure 3, the words "President", "amp", "s" "vote" , "people" and "now" stand out the most . Many of the words on this word cloud are expected and so it is unsurprising why they are populating as the most common words.

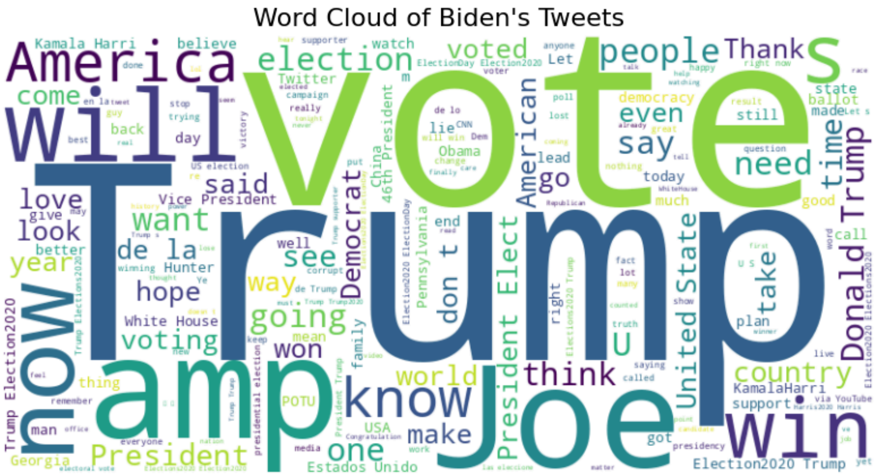


Figure 4. Word Cloud of Biden's Tweets

In figure 4, the words that stand out the most in this word cloud are "Trump", "vote", "will", "amp", "Joe" and "America." The word "Trump" jumps out and remains the largest word in this visualization. It is surprising that "Trump" appears most commonly in tweets directed towards Biden.