**Data**:

'president\_state.csv', 'hashtag\_joebiden.csv', 'hashtag\_donaldtrump.csv'

**Source:**

[US Election 2020 | Kaggle](https://www.kaggle.com/unanimad/us-election-2020)

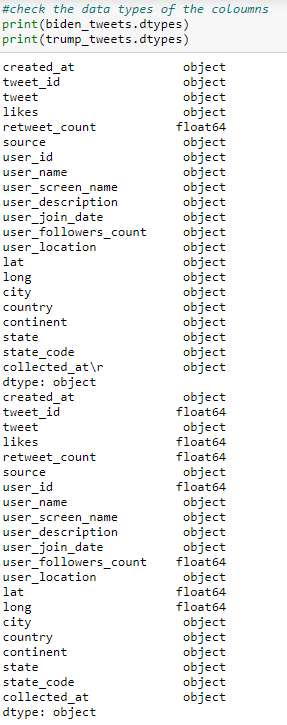
[US Election 2020 Tweets | Kaggle](https://www.kaggle.com/manchunhui/us-election-2020-tweets)

**Description of your data exploration**:

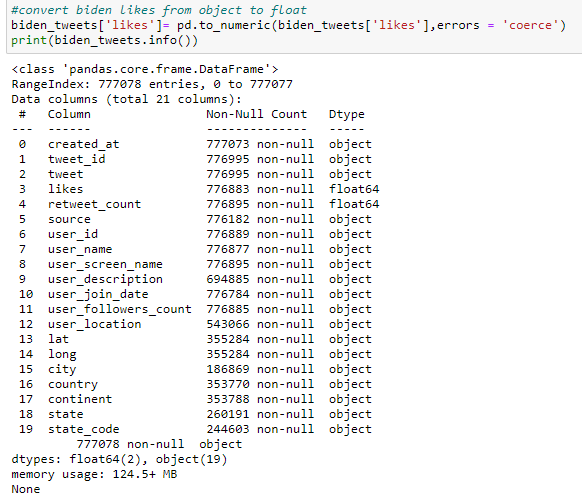
The president\_state data frame holds the winning candidate and vote total for each state. The hashtag\_joebiden and hashtag\_donaldtrump datasets contain tweets about the respective candidate. The votes in addition to the total engagement with and sentiment for the main candidates via twitter will be analyzed to see if there is a possible association between the two. All three files were read in with pd.read\_csv.

**data cleaning steps**:

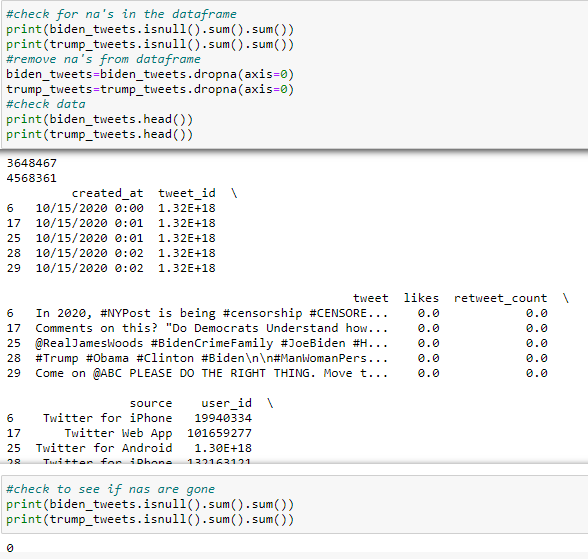
Once the datasets were loaded in, the tweet data frames data types were called to see if any columns had to be coerced into another data type.



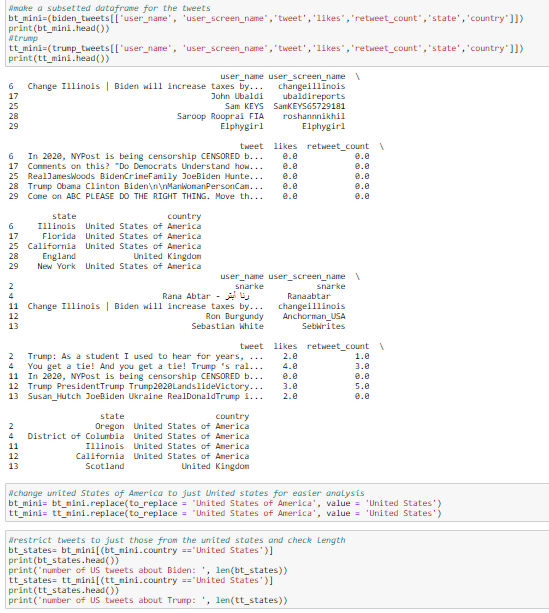
The data types showed that Biden’s data frame had ‘likes’ as an object which was then coerced into float to allow for comparisons to retweet numbers.



After changing the likes to float, the nan values were removed to get more accurate numbers. The resulting data frame was half that of the original file but there was enough data for results to be statistically significant.



The new datasets then had @ and # removed from the ‘tweet’ column so that only the most meaningful symbols remain. Those clean tweets were then extracted with: ‘user\_screen\_name’, ‘user\_name’, ‘likes’, retweet\_count’, ‘state’, ‘county’. These columns were used because the name served as an identifier for the tweets that will later be grouped by regions of the United states.



The new dataset’s length(len) was measured along with the layout of the table to ensure that there was a substantial number of tweets and that the filtering was completed correctly.

The new united states twitter data was filtered once again by counting the number of tweets by state in one date frame with the average number of likes and retweets by state in another. The datasets were then split into the four census regions (northeast, south, Midwest, and west) to prepare for analysis and comparisons.

To prepare for sentiment analysis the following was cleaned from the tweet data of each presidential candidate to ensure a more accurate reading of the senses; punctuation removal, lower casing, number removal, trailing space removal, URL removal, inter word spaces, and stop words from the English nltk stop words corpus.

The cleaned tweets were then processed through TextBlob and plotted using seaborn factorplot.

Text

Description automatically generated with medium confidence

**Comparison questions with the unit of analysis, the comparison values and how they are computed**:

1. Compare the descriptive statistics of tweet engagement by region (retweets, likes, and number of tweets)
2. Compare the sentiment of Biden and Trump tweet
3. Compare the descriptive statistics of the b (followers, friends, status)

**Brief description of the program**:

The programs used are in three parts: twitter descriptive statistics, twitter sentiment, and regional voting blocks.

**Description of the output files**:

The twitter statistics that were analyzed were likes, tweets, and retweets to gauge engagement by region. Each barplot took the two data frames and plotted them. The first graph only shows tweet totals while the second set compares averages between likes and retweets in a region.

Graphical user interface, application

Description automatically generated

Chart, waterfall chart, box and whisker chart

Description automatically generated



This process is repeated for the Midwest, West, and South.

The output for the sentiment analysis were two barcharts that reveal both candidates had mostly neutral tweets with more positive than negative tweets. The factorplot use to plot these graphs are deprecated so they would need to be recoded as a catplot when the new version of python is released.

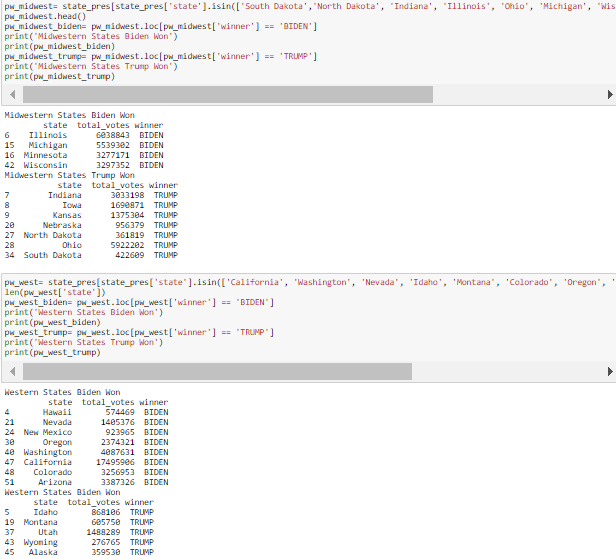
Waterfall chart

Description automatically generated with low confidence

State voting totals were regional subsets of the president\_state dataset. The tables then reveal which states fell to which candidate.

Graphical user interface, application

Description automatically generated



**Conclusion about your overall result**:

The results show that there are some clear regional voting blocs emerging from the 2020 election. The south and Midwest were predominantly for Trump while the West and Northeast gravitated toward Biden. This trend can be explained by the presence of more metro areas in the northeast and west which typically translates to more liberal thought while the rural, blue collar areas of the South and Midwest are strictly conservative with a bit of Trumpism mixed. Trumpism is the cult of personality surround his voter base. Speculation is leading some to believe that if Trump ran as third-party candidate, he would garner more than 17% of the popular vote, potentially taking electoral votes in the process.

The overall excitement behind trumps political movement shows in the overall higher positive and neutral tweets as media outlets are drawn to what’s popular. The negative tweets are also more prevalent due to his controversial stances on key issue along with his laundry list of sexual assault and tax fraud allegations. Biden overall is a boring career politician who promised no fundamental change which, compared to Trumps four years, seems like a radical shift to normal for some.

When taking this information and comparing it to the twitter engagement. There was generally more engagement with Trump. This could be due to factors such as his incumbency which presents a lot of talking points regarding his first four years in office. Other public issues like his delegitimization of mail-in voting which has caused the media and therefore, discourse to pivot around that. Lastly, Trump’s based was way more excited for trump and the individual campaigning efforts were much more enthusiastic than Biden’s. All these factors explain the strong showing that trump would go on to have despite the losing the election. In the future, sentiment analysis and polarity by region should be used in tandem with regression models to get a more concrete outlook at the correlation between tweets and election results.