Project 9 - CSCI 403 Database Management

Margaret Cayce, Karah Han, Jared Schneider, Samson Mammarappallil

A Twitter Investigation of the 2020 Presidential Election

Data Overview:

The data we chose, entitled “US Election 2020 tweets”, consists of two datasets and includes over 1.7 million tweets. The two datasets each contain 21 columns detailing the information of each tweet posted around Election Day in 2020. Each row features a single tweet. One .csv file holds information about every tweet with the hashtag “#joebiden” and the other .csv file holds information for the hashtag “#donaldtrump”. The datasets were comparable in size. These two datasets will henceforth be referred to as Biden dataset and Trump dataset, respectively.

As loaded into the CSCI403 database, the datasets contain the following 13 columns (see Data Cleaning section for data cleaning methods):

- created_at - Time the tweet was created.
- tweet - Actual content of the tweet.
- likes - Number of likes the tweet has.
- retweet_count - Number of retweets the tweet has.
- user_name - User id of the tweet’s author (@<name>).
- user_screen_name - Author’s screen name (their “nickname”).
- user_description - Author’s bio or description.
- user_join_date - Author’s join date.
- user_followers_count - Author’s total number of followers.
- user_location - Author’s location (if provided).
- country - Country the tweet was made in (if provided).
- continent - Continent the tweet was made in (if provided).
- state_code - State code of the state the tweet was made in (if provided).

The dataset was obtained from Kaggle at https://www.kaggle.com/manchunhui/us-election-2020-tweets. The license on the data obtained is a CC0, so there are no copyright issues or license restrictions.

Topic Background:

Tweets have been an increasingly relevant method that reflects the current state of mind of anyone, from news outlets, to individuals, to Presidents themselves. Thus, we determined that analysis of candidate-related tweets would yield relevant and insightful results.

The breadth of information provided in our dataset would allow us to analyze trends across multiple factors. Although there were so many directions we could take this project, after discussion, we landed on the issue of bots on Twitter.

Project 9 - CSCI 403 Database Management

A Twitter bot is an implementation of robot software that uses an account via the Twitter API [1]. Some bots are harmless; for example, there are bots commissioned by government agencies that tweet urgent information about natural disasters or inclement weather. However, there are many concerns about the negative influence of Twitter bots on news and reporting on current events.

Twitter Bots:

According to Twitter, one of the functions of a bot includes “Engaging in bulk or aggressive tweeting, engaging, or following.” Twitter already has limits on tweet frequency and account edits to prevent spam. However, the Twitter algorithm is not perfect. It is unclear how many bots are active on Twitter, but a Carnegie Mellon study estimates that half of all users on the platform are bots [3]. Researchers at the Pew Research Center assert that 66% of all tweeted links were tweeted by bots [2]. In other words: Twitter bots are incredibly active and promote news to a wider audience via activity.

Bots on Twitter can promote false information and serve to further divide an already polarized set of users [4]. So, the issue is clear: malicious bots can be harmful to public discourse. However, the solution is less so: while the onus is on Twitter to remedy this, users should try to determine whether the news they receive on Twitter is trustworthy.

We set out to determine if a Twitter bot could be identified by the average Twitter user. The following questions guided our project: What factors indicate an account is a potential bot? Is there an easy way to identify a bot (a flag or single/set of characteristic(s))? Could Twitter do more to eliminate malicious bots on its platform?

We decided to do what we could with the scope of this class and came up with some queries to investigate tweet trends. It is noticeable that some data that was removed from the final set imported into the database were repeat tweets with differing post times. These, if they were included in the data set, would be suspicious, but had to be removed due to corrupted data.

Data Cleaning:

Because we were initially unsure about the process to upload csv files into the CSCI403 database, we conducted our data cleaning process in a Google Colab notebook using Python. We dropped metadata columns, such as tweet_id, and redundant columns, such as state (since we already had state_code). We also removed all rows with null values for the user_name.

Python script:

```
import numpy as np
import pandas as pd
from google.colab import drive
drive.mount('/content/drive')
```

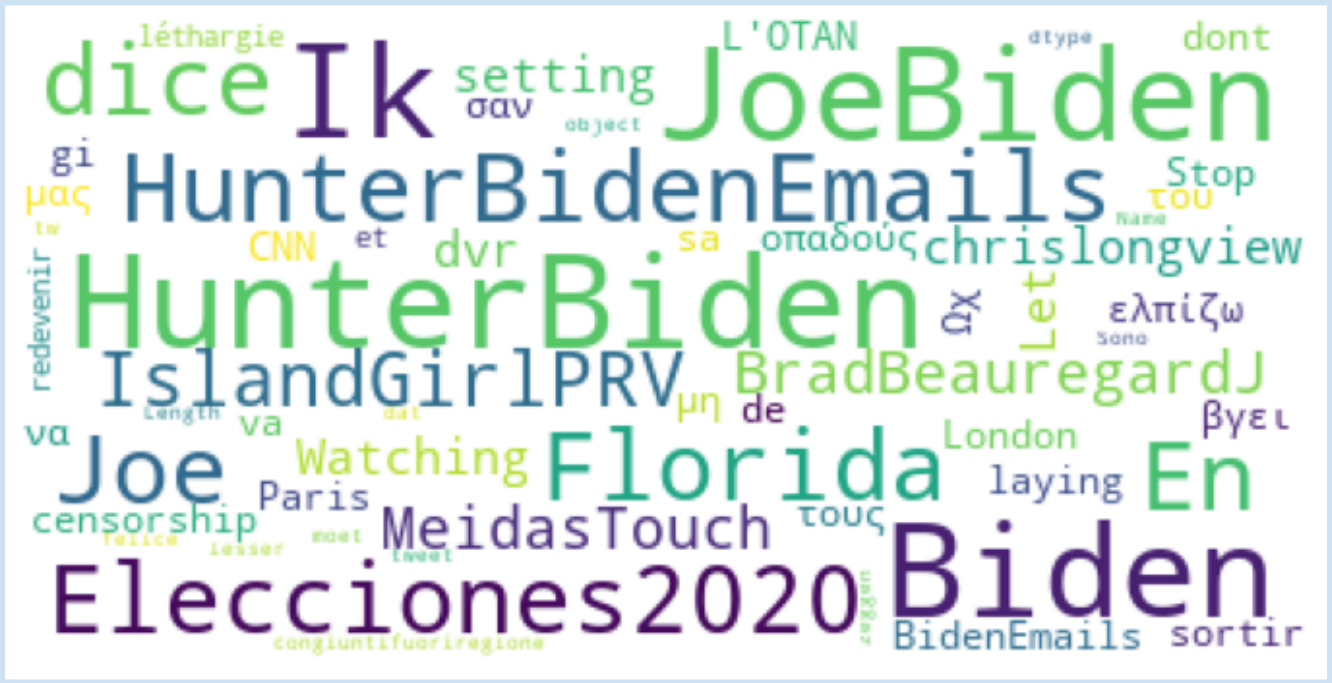
```
path_t = "/content/drive/My Drive/kagglevoting/hashtag_donaldtrump.csv"
path_b = "/content/drive/My Drive/kagglevoting/hashtag_joebiden.csv"
df_t = pd.read_csv(path_t,lineterminator='\n')
df_b = pd.read_csv(path_b,lineterminator='\n')
# Drop unwanted columns: tweet_id, user_id, city, state, collected_at
cols_to_drop = ['tweet_id','user_id','city','state',
'collected_at','lat','long','source']
clean_t=df_t.drop(labels=cols_to_drop, axis=1)
clean_b=df_b.drop(labels=cols_to_drop, axis=1)
# Remove all rows with null values in columns: user_name
clean_t = clean_t[clean_t['user_name'].notna()]
clean_b = clean_b[clean_b['user_name'].notna()]
```

We proceeded to use three different approaches to investigate these tweets:

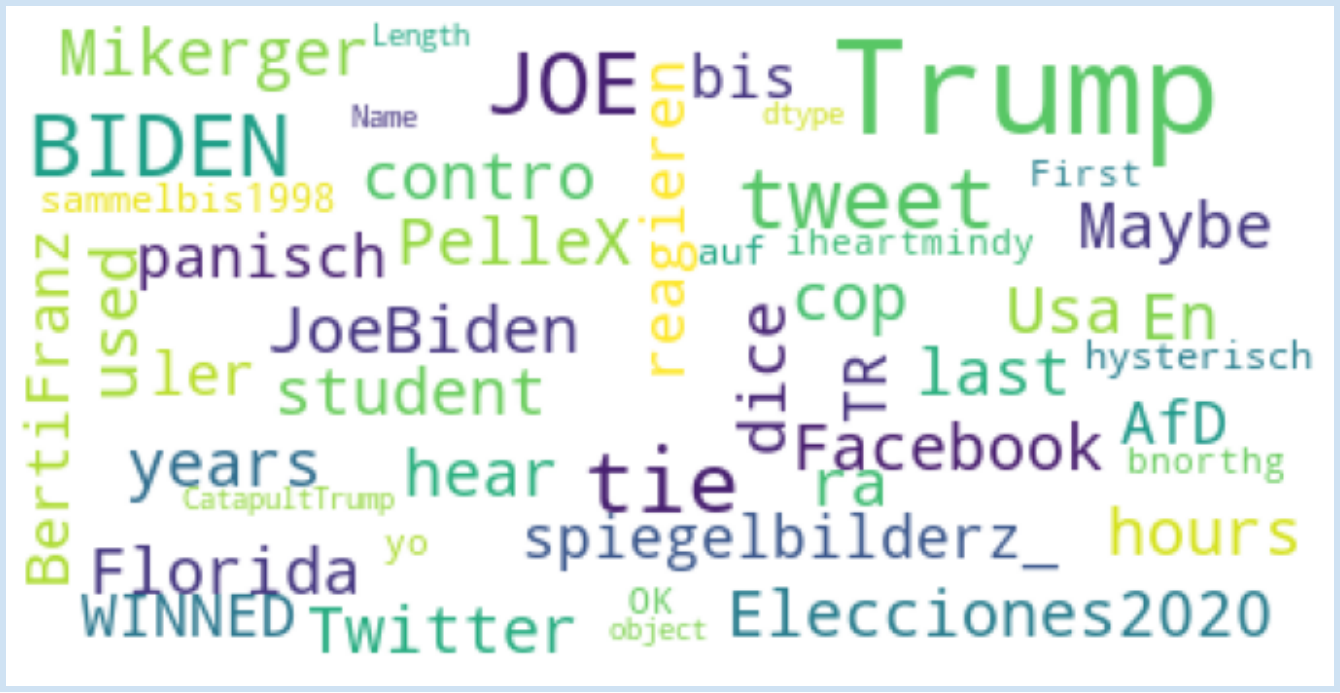
1. Word maps
2. Geographic visualizations
3. Duplicate tweet content

Analysis 1: Word Maps

We created a word cloud for each dataset. We first isolated the tweet content column, then used a Python script to create separate word maps for each candidate.



Word cloud for Biden related tweets.



Word cloud for Trump related tweets.

Analysis 2: Geographic Distribution

The first query we performed aimed to display the tweet count for each state in the United States. This was to:

a) Minimize non-English tweets, for easier analysis later
b) Observe and analyze state trends

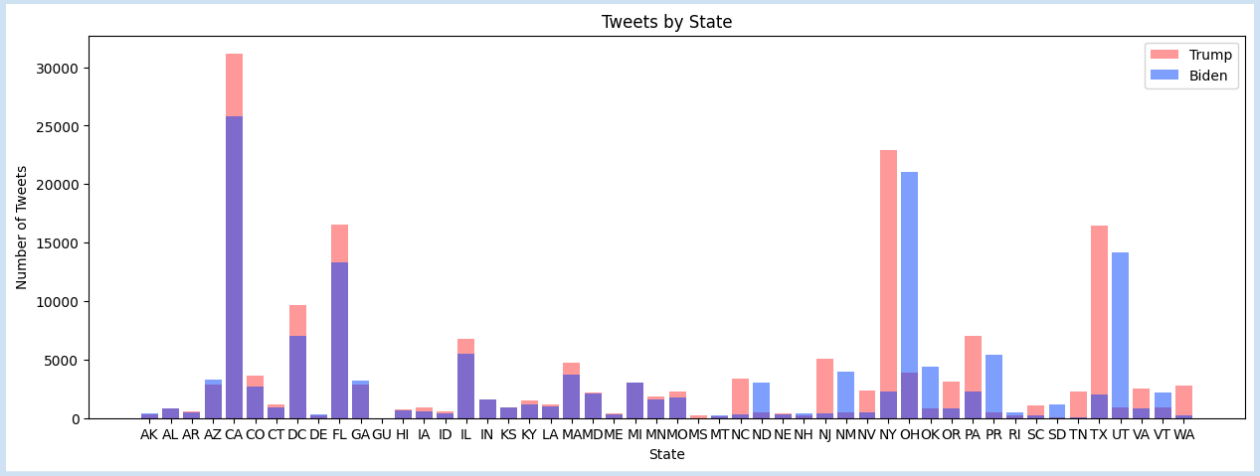
We used the follow queries in SQL:

```
SELECT COUNT(state_code), state_code FROM df_t WHERE country = 'United
States of America' AND state_code NOT NULL GROUP BY state_code LIMIT 50;
SELECT COUNT(state_code), state_code FROM df_b WHERE country = 'United
States of America' AND state_code NOT NULL GROUP BY state_code LIMIT 50;
```

We also used the following Python script, which pulled the query results into a numpy dataframe in order to create a bar graph:

```
import matplotlib.pyplot as plt; plt.rcdefaults()
import matplotlib.pyplot as plt
states = by_state_t['state_code']
num_tweets_t = by_state_t['COUNT(state_code)']
num_tweets_b = by_state_b['COUNT(state_code)']
plt.figure(figsize=(15, 5))
```

```
plt.bar(states, num_tweets_t, align='center', alpha=0.5, color=(1, 0.2,
0.2, 0))
plt.bar(states, num_tweets_b, align='center', alpha=0.5, color=(0, 0.25,
1, 0))
plt.ylabel('Number of Tweets')
plt.xlabel('State')
plt.title('Tweets by State')
plt.legend(['Trump', 'Biden'])
plt.show()
states = by_state_t['state_code']
worth = np.array([.01, .05, .10, .25])
```



Tweets by State.

While nothing here points to direct bot manipulation, there are some interesting notes on this distribution:

- While Ohio is the 7th most populous state [5], it ranked 3rd in tweet volume, where #joebiden heavily outweighs #donaldtrump.
- Notice that the historically blue states of California and New York were red-dominated on Twitter. We determined two possible explanations for this:
    - Larger, more rural areas tend two swing red, while the more densely populated cities lean left. Thus, the population of tweets might appear more right-leaning when compared to the election results of such states.
    - A usage of the hashtag #donaldtrump does not necessarily mean a positive statement was made about the candidate.

We also created the following heat maps using a combination of SQL queries and a map extension from Bing. The first two heat maps represented tweet density by candidate per state. The SQL queries used for the first two tweet density heat maps were the following:

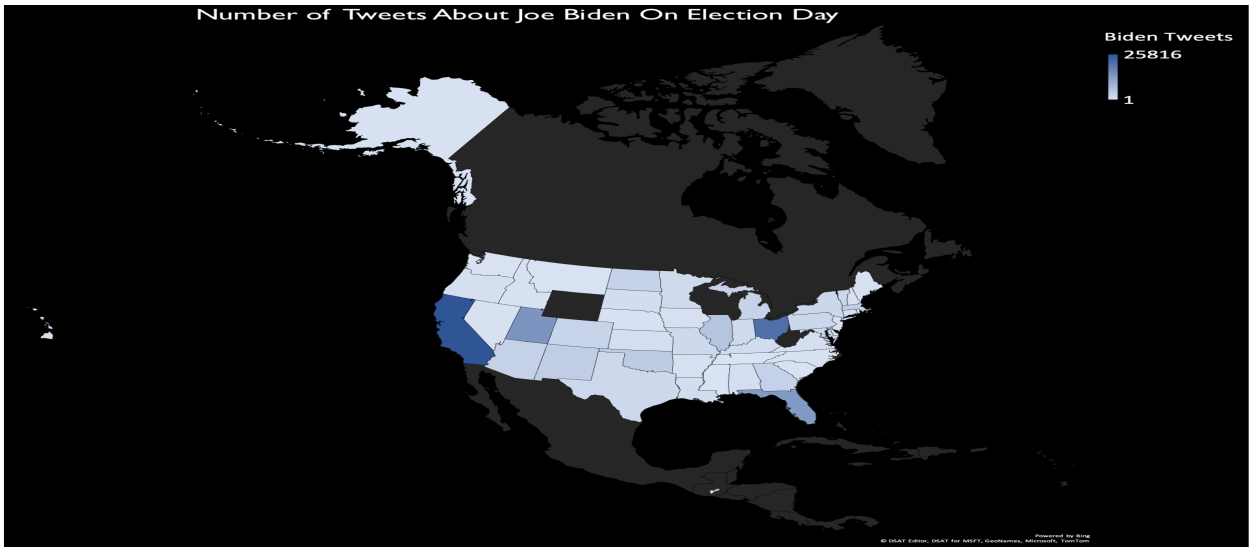
```
SELECT COUNT(state_code), state_code FROM trump WHERE country = 'United
States of America' AND state_code NOT NULL GROUP BY state_code LIMIT 50;
SELECT COUNT(state_code), state_code FROM biden WHERE country = 'United
States of America' AND state_code NOT NULL GROUP BY state_code LIMIT 50;
```

Finally, we created a third heat map that weighed how much representation each candidate had on Twitter and which candidate had more Twitter representation in each state. We used a darker color for Donald Trump and a lighter color for Joe Biden. We used a simple weighting method that gave “one point” to liked tweets, “two points” to retweeted tweets, and “three points” to tweets that were both liked and retweeted. The results were a heat map that essentially showed Twitter popularity of both candidates in each state.

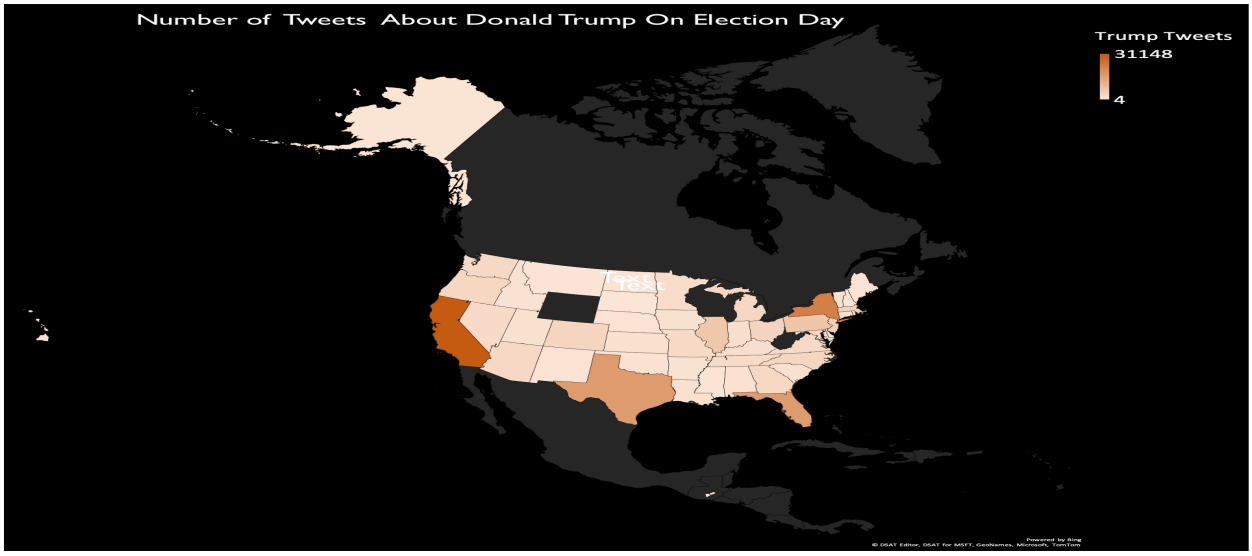
Queries for Heat Map 3:

```
/*biden tweets that were liked: weighted x1*/
SELECT COUNT(state_code), state_code FROM biden WHERE country = 'United
States of America' AND state_code IS NOT NULL AND likes > 0   GROUP BY
state_code LIMIT 50;
/*biden tweets that were retweeted: weighted x2*/
SELECT COUNT(state_code), state_code FROM biden WHERE country = 'United
States of America' AND state_code IS NOT NULL AND retweet_count > 0
GROUP BY state_code LIMIT 50;
/*biden tweets that were retweeted and liked: weighted x3*/
SELECT COUNT(state_code), state_code FROM biden WHERE country = 'United
States of America' AND state_code IS NOT NULL AND likes > 0 AND
retweet_count > 0 GROUP BY state_code LIMIT 50;
/*trump tweets that were liked: weighted x1*/
SELECT COUNT(state_code), state_code FROM trump WHERE country = 'United
States of America' AND state_code IS NOT NULL AND likes > 0   GROUP BY
state_code LIMIT 50;
/*trump tweets that were retweeted weighted x2*/
SELECT COUNT(state_code), state_code FROM trump WHERE country = 'United
States of America' AND state_code IS NOT NULL AND retweet_count > 0
GROUP BY state_code LIMIT 50;
/*trump tweets that were retweeted and liked weighted x3*/
SELECT COUNT(state_code), state_code FROM trump WHERE country = 'United
States of America' AND state_code IS NOT NULL AND likes > 0 AND
retweet_count > 0 GROUP BY state_code LIMIT 50;
```

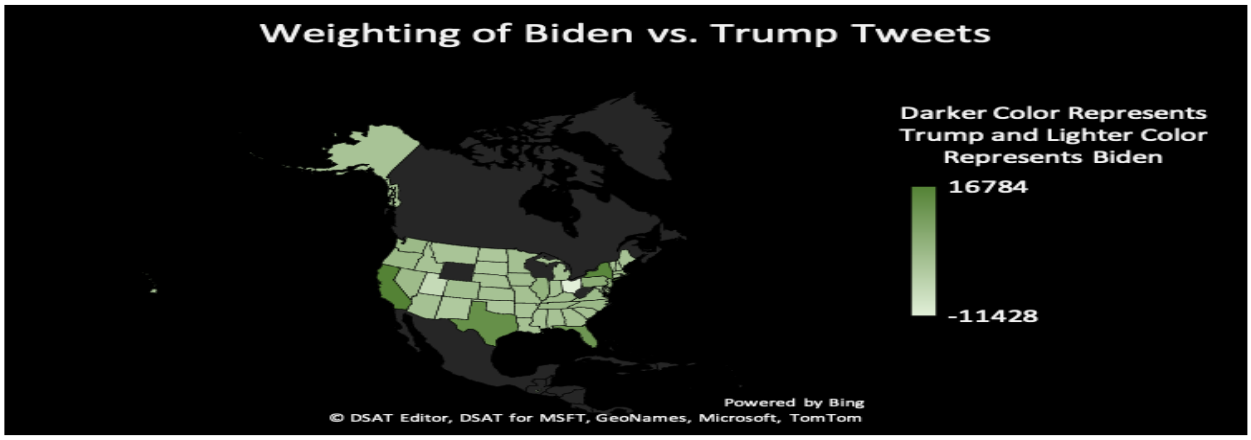
Project 9 - CSCI 403 Database Management



"Joe Biden" Tweet Density Map for the United States



"Donald Trump" Tweet Density Map for the United States



Heat Map Showing Which Candidate Was More Popular In Each State on Twitter

Analysis 3: Duplicate Tweet Content

Another query we performed aimed to find tweets with repeated content, because this can be a hallmark of bot account networks. Some bot networks may have all bots tweet the same content to increase the visibility of their message. Some of the tweets are short and not of relevance, while others were longer and could not be reasonably generated organically. It is hard to imagine 37 different humans tweeting a series of niche hashtags in the same order, for example

```
SELECT tweet, COUNT(*) FROM biden GROUP BY tweet HAVING COUNT(*)>1 ORDER BY COUNT(*)
DESC;
```

tweet	count
#JoeBiden	311
#Biden	216
realDonaldTrump #JoeBiden	51
#biden	47
Congratulations #JoeBiden	43
JoeBiden #BelieveInAmerica #TrumpIsALaughingStock #SpeakUpVoteBiden #StrongerWithBiden #TrumpVirus #TrumpIsPathetic #BidenCares #BidenHarris2020 #BidenHarris #Biden #DitchMitch #DitchMoscowMitch #MoscowMitch #republicansforbiden #FloridaForBiden #TrumpChinaBankAccount #Biden2020	37
realDonaldTrump Emails reveal how Hunter Biden tried to cash in big on behalf of family with Chinese firm Bye Bye #Biden 👺🤦🤡 Enjoy pension 2020 #SmokingGunEmail https://t.co/JIGmRbsASS	33
#JoeBiden होंगे अमेरिका के नए राष्ट्रपति 284 इलेक्टोरल वोट लेकर ट्रंप को पछाड़ा. #USAelection2020	30
JoeBiden 🔴Trump has failed with #COVID19 climate issues and minority rights! No more #Trump! Trump is America's failure! Trump is virus! #Coronavirus #covid_19 #vote #VoteEarly #VoteEarlyDay #VoteBidenHarris #vote2020 #Biden #JoeBiden #Harris #HarrisBiden #LGBT #US #USA #BLM #games	30
realDonaldTrump #Biden	30

```
SELECT tweet, COUNT(*) FROM trump GROUP BY tweet HAVING COUNT(*)>1 ORDER BY COUNT(*)
DESC;
```

tweet	count
#Trump	259
#DonaldTrump	84
realDonaldTrump #Trump	83
#trump	75
realDonaldTrump I agree with this article..No doubt #donaldtrump is the worst president in American history...at least modern history.#TrumpIsANationalDisgrace #TrumpUsARacist#TrumpIsAPoshttps://t.co/MpZpbzJyVJ	51
#Trump Sucks Fatballs in #Russia :D	49
It's OK to be White. #BlackLivesMatter #Feminism #CNN #Trump #Islam #HillaryClinton #BernieSanders #ItsOkToBeWhite #AllLivesMatter	46
realDonaldTrump This is how #donaldtrump gets money in his pockets...steals from kids with cancer. #TrumpIsACriminal#TrumpForPrison#TrumpIsAPoshttps://t.co/YbMVaOgcY2	45
It's still OK to be white in 2017. #BlackLivesMatter #Feminism #CNN #Trump #Islam #HillaryClinton #BernieSanders #ItsOkToBeWhite #AllLivesMatter	44
MAKE EQUESTRIA NEIGH AGAIN #DonaldJHoofForPresident ##trump	43

After looking at the results of these queries, we decided to go one step further: we isolated the content of a single tweet to look at the users. We were curious to see if these tweet authors

would share some characteristics of a suspected bot: no identifying information in user description, generic usernames (a name followed by a series of numbers), and recent join date.

```
SELECT created_at, user_name, user_screen_name, user_description, user_join_date FROM
"jschneider"."biden" WHERE ("tweet"::TEXT LIKE '%JoeBiden #BelieveInAmerica
#TrumpIsALaughingStock #SpeakUpVoteBiden #StrongerWithBiden #TrumpVirus %') LIMIT 600
OFFSET 0;
```

created_at	user_name	user_screen_name	user_description	user_join_date
10/24/2020 22:21	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:22	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:23	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:24	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:24	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:26	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:26	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:26	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45
10/24/2020 22:27	Elias Lefty Caress	eliascaress	Award-winning magician and variety Entertainer. In the middle of the worst recession of my generation I walked away from a secure job to follow my dreams.	10/16/2014 19:45

At first glance this twitter account seemed suspicious because it was tweeting the same content over and over again, in one case 3 times in one minute. But, the user description is not vague, and the user joined 6 years ago. A quick Google search showed that Elias Lefty Caress is indeed a real person.

```
SELECT * FROM trump WHERE ("tweet"::TEXT LIKE '%OK to be White. #BlackLivesMatter #Feminism #CNN
#Trump #Islam%') LIMIT 300 OFFSET 0;
```

id	created_at	tweet	likes	retweet_count	user_name	user_screen_name	user_description	user_join_date
7988	10/15/2020 12:08	It's OK to be White. #BlackLivesMatter #Feminism #CNN #Trump #Islam #HillaryClinton #BernieSanders #ItsOkToBeWhite #AllLivesMatter	0	0	IT'S OK TO BE WHITE	BeingWhiteIsGr8	This is a bot and will not reply to any notifications.	12/20/2017 23:29
45447	10/17/2020 3:08	It's OK to be White. #BlackLivesMatter #Feminism #CNN #Trump #Islam #HillaryClinton #BernieSanders #ItsOkToBeWhite #AllLivesMatter	0	0	IT'S OK TO BE WHITE	BeingWhiteIsGr8	This is a bot and will not reply to any notifications.	12/20/2017 23:29
23221	10/16/2020 2:38	It's OK to be White. #BlackLivesMatter #Feminism #CNN #Trump #Islam #HillaryClinton #BernieSanders #ItsOkToBeWhite #AllLivesMatter	0	0	IT'S OK TO BE WHITE	BeingWhiteIsGr8	This is a bot and will not reply to any notifications.	12/20/2017 23:29

A second query had a more exciting result: we found a bot! The frequency of tweets alone was not questionable, nor was the user join date (almost 3 years ago), but the tweet content was. The user description confirmed our suspicions. Unfortunately, it is not usually this easy.

Technical Challenges:

Throughout our project, we struggled with several technical challenges:

- Loading dataset into course database
- Geographic visualizations
- Challenges with Excel

Our first technical challenge was properly loading our dataset into the course database and ensuring that it fit into a relational schema. We had a tough time converting to proper file formatting that would allow for the sql copy command to work properly. Also, a large number of tweets resulted in the error code “invalid input syntax for integer”, since some tweets had text in place of where integers were supposed to be. These tweets were very tedious to remove since each error code was outputted individually and we resultantly had to remove each one of these tweets individually. Additionally, each team member needed to reset their encoding, since we were all working from Windows rather than Linux. We resolved this with the following command:

```
\encoding UTF8
```

Our second technical challenge was creating a visualization which mapped each tweet in the United States by latitude and longitude. This required the python libraries basemap and

geopandas and usage of .shp objects. After spending a few hours on this approach, we determined it was too time consuming and not relevant enough to our chosen project scope to continue, and opted for the bar graph.

Finally, putting our data in Excel gave us numerous challenges. It unfortunately placed our date and time data into a format which SQL was unable to interpret as a date-time, which limited how well we could query things related to dates and times. Additionally, some characters had to be removed from the data (commas, @ symbols, equal signs) due to challenges of importing the data set as a CSV file. These issues might not have been as prevalent if we were more “Excel savvy” but all of our experience with Excel was fairly limited.

Team Contributions:

Each person attended and actively contributed to each of the 5 regular team meetings. Additionally, each team member was flexible and attended additional meetings to address unexpected issues and attend office hours. All members contributed to data cleaning and report composition.

Individual contributions:

- Margaret: Came up with the scope of the project, implemented word mapping, set up project report, did duplicate tweet analysis.
- Karah: Created state distribution queries and bar graphs, implemented data cleaning in Python and downloaded data, attempted using Python packages basemap and geopandas.
- Samson: Went to additional office hours to troubleshoot .csv upload issues, did sql queries and used a Bing extension in conjunction with the queries to make maps of different kinds showing number of tweets and Twitter popularity (weighting) of each candidate by state.
- Jared: Troubleshot .csv upload issues in Excel and Postgres. Uploading and setting share permissions for datasets.

Bibliography:

[1] Johansen, A. G. (2020, June 16). Whats a Twitter Bot and How to Spot One. *Norton*. https://us.norton.com/internetsecurity-emerging-threats-what-are-twitter-bots-and-how-to-spot-them.html. Accessed 07 December, 2020.

[2] Wojcik, S. (2018, April 9). 5 Things to Know About Bots on Twitter. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2018/04/09/5-things-to-know-about-bots-on-twitter/. Accessed 07 December, 2020.

[3] Twitter. About Twitter Limits. https://help.twitter.com/en/rules-and-policies/twitter-limits

[4] Samuels, E., Akhtar, M.(2019, November 20). Are Bots Manipulating the 2020 conversation? Here's what's changed since 2016. *Washington Post.* https://www.washingtonpost.com/politics/2019/11/20/are-bots-manipulating-conversation-heres-whats-changed-since/. Accessed 07 December, 2020.

[5] https://worldpopulationreview.com/states