SQL for Data Science Capstone Project

Proposal

Selected client/dataset:

Lobbyists4America was selected to find insights on key topics, members, and relationships within Congress.

Steps took to import and clean the data:

The Tweets.json file was uploaded to the database on Databricks, then read into PySpark DataFrame. The PySpark DataFrame was converted to pandas DataFrame. The created_at date was converted to datetime format.

Description:

Congressional Tweets Dataset from 2008 to 2017 will be analyzed to help to find insights on key topics, members, and relationship within Congress. This will help customers of Lobbyist4America to better influence the legislation within the US. By identifying relevant topics, they can focus on topics that will get more engagements. Additionally, popular members can be contacted to help spread messages more effectively. These analyses will provide intuition for lobbyists to identify actions with higher impact on lobbying efforts.

Questions:

1. Who tweets and how often?
   a. This can show important members within the congress

2. What topics are discussed and how often?
   a. This will show what the congress likes to talk about
   b. Show key topics that was focused on

3. What's the "sentiment" or "tone" of what was discussed?
   a. Show if tweets sound positive or negative
   b. And the change of "sentiment" over time

Hypothesis:

1. Congressional tweets will be highly connected to politicians
   a. As the Congress is associated with politics

b. Will appear in tweets, mentions, RTs, etc.

2. Topics will be about politics and lobbying
   a. As the Congress is associated with politics
   b. Will appear in tweets, mentions, RTs, etc.

3. The sentiment will be mostly negative
   a. Politics is mostly depressing

Approach:

1. The focus will be on the frequency metrics (connections) such as:
   a. Retweets by finding tweets that starts with RT
   b. Timing of tweets by looking the time the tweets was created
   c. Mentions from the entities column
2. Counts / word frequencies will also be looked at. Examples:
   a. Sentiment Analysis using text blob
   b. Topic Analysis using word clouds

Graphics / Visualizations:

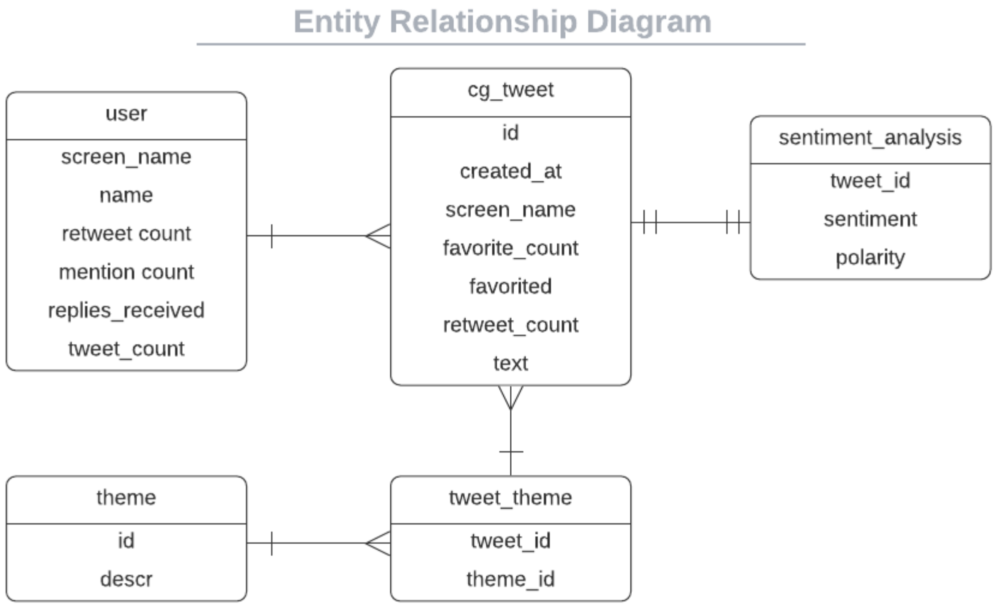
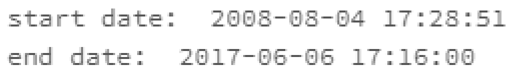


Figure 1: Entity relationship diagram



```
start date:  2008-08-04 17:28:51
end date:  2017-06-06 17:16:00
```

Figure 2: Start and end dates

The first start date of the collected tweets is on August 4, 2008 and the last date collected is on June 6, 2017

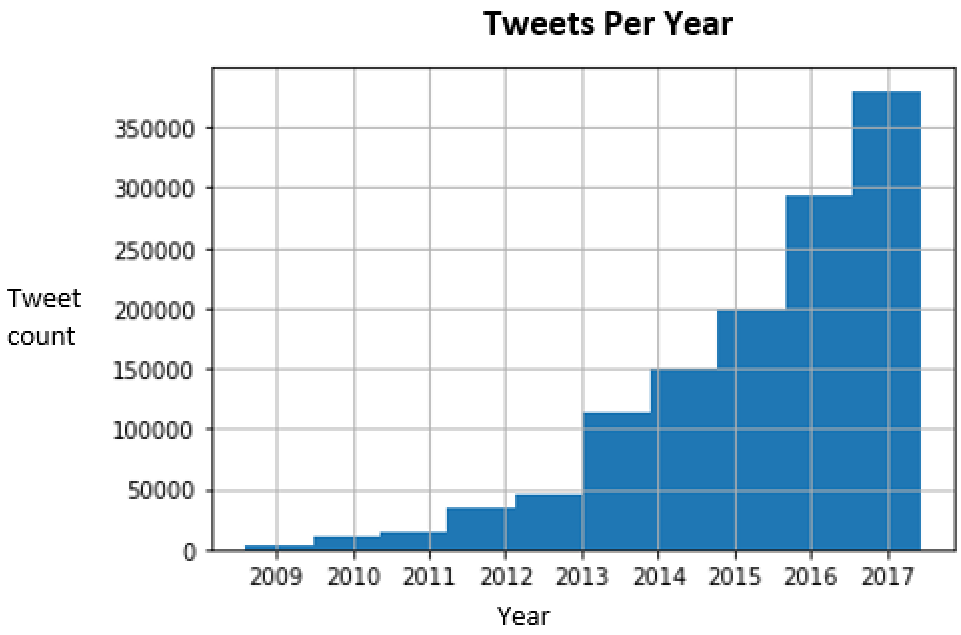
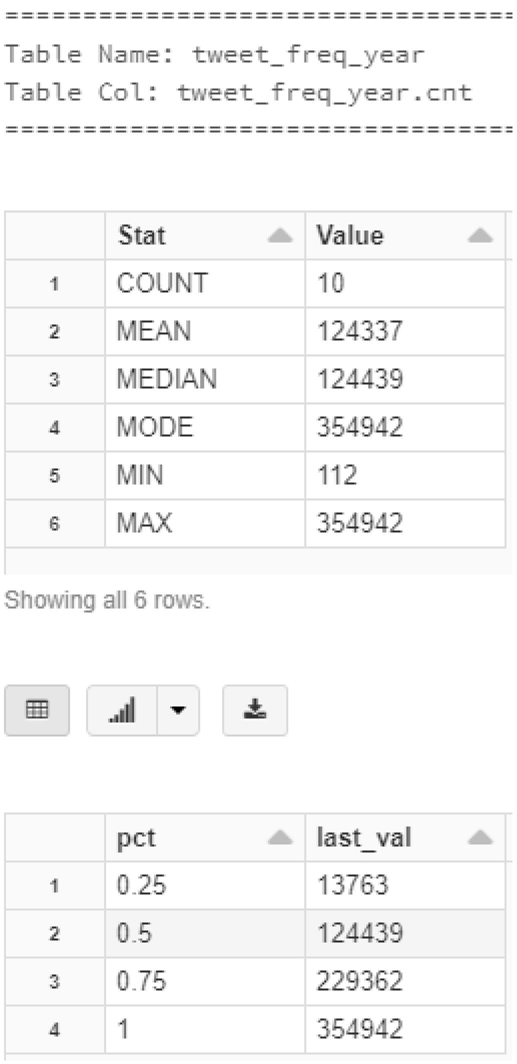


Figure 3: Tweets per year

The number of tweets is increasing exponentially every year.



```
==================================
Table Name: tweet_freq_year
Table Col: tweet_freq_year.cnt
==================================
```

	Stat	Value
1	COUNT	10
2	MEAN	124337
3	MEDIAN	124439
4	MODE	354942
5	MIN	112
6	MAX	354942

Showing all 6 rows.

	pct	last_val
1	0.25	13763
2	0.5	124439
3	0.75	229362
4	1	354942

Figure 4: Descriptive statistics of tweet frequency by year

Interquartile percentage is represented with pct. The mean and median of tweets count by year are about the same. The minimum tweets per year is significantly lower than the maximum.

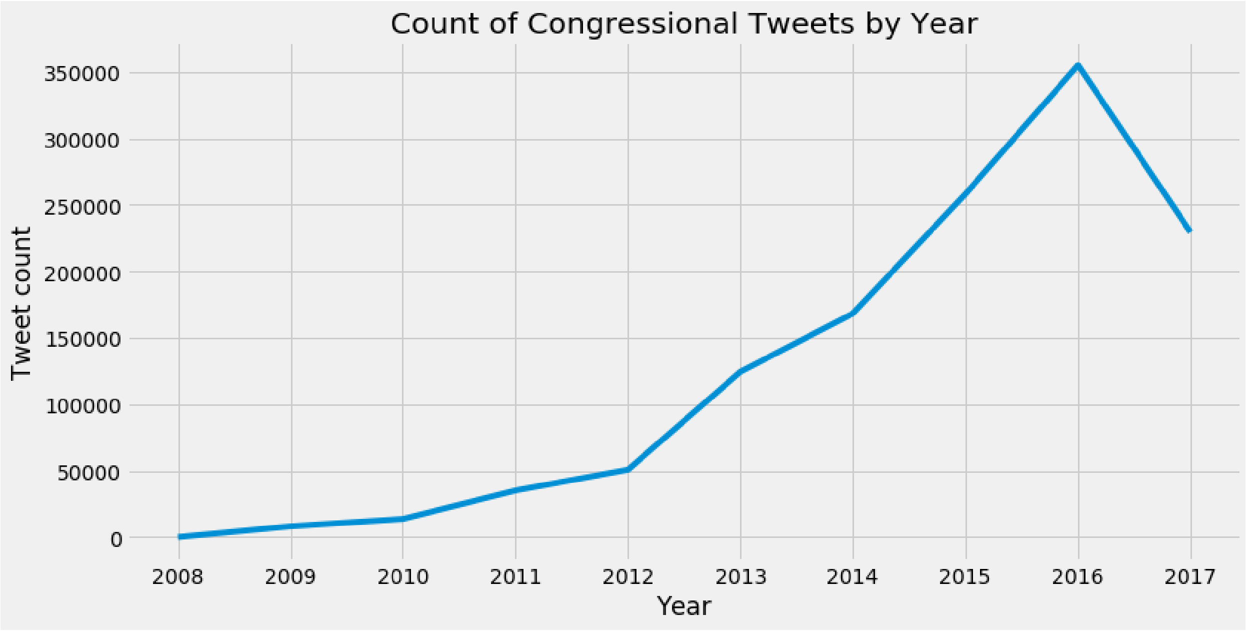
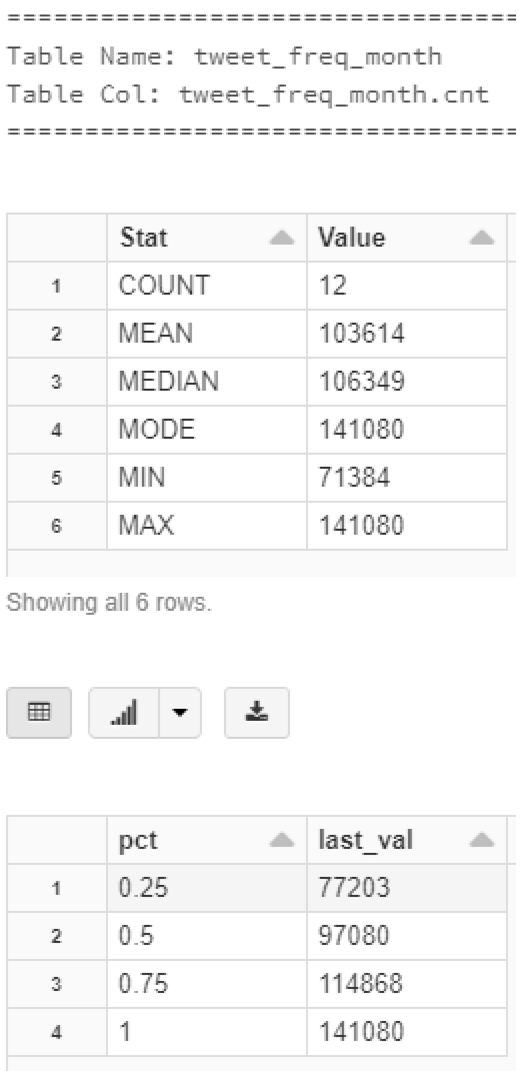


Figure 5: Count of congressional tweets by year

Tweets frequency is increasing every year except for 2017. This is due to data being collected only for half a year in 2017.



```
==================================
Table Name: tweet_freq_month
Table Col: tweet_freq_month.cnt
==================================
```

	Stat	Value
1	COUNT	12
2	MEAN	103614
3	MEDIAN	106349
4	MODE	141080
5	MIN	71384
6	MAX	141080

Showing all 6 rows.

	pct	last_val
1	0.25	77203
2	0.5	97080
3	0.75	114868
4	1	141080

Figure 6: Descriptive stats of tweets frequency by month

Interquartile percentage is represented with pct.

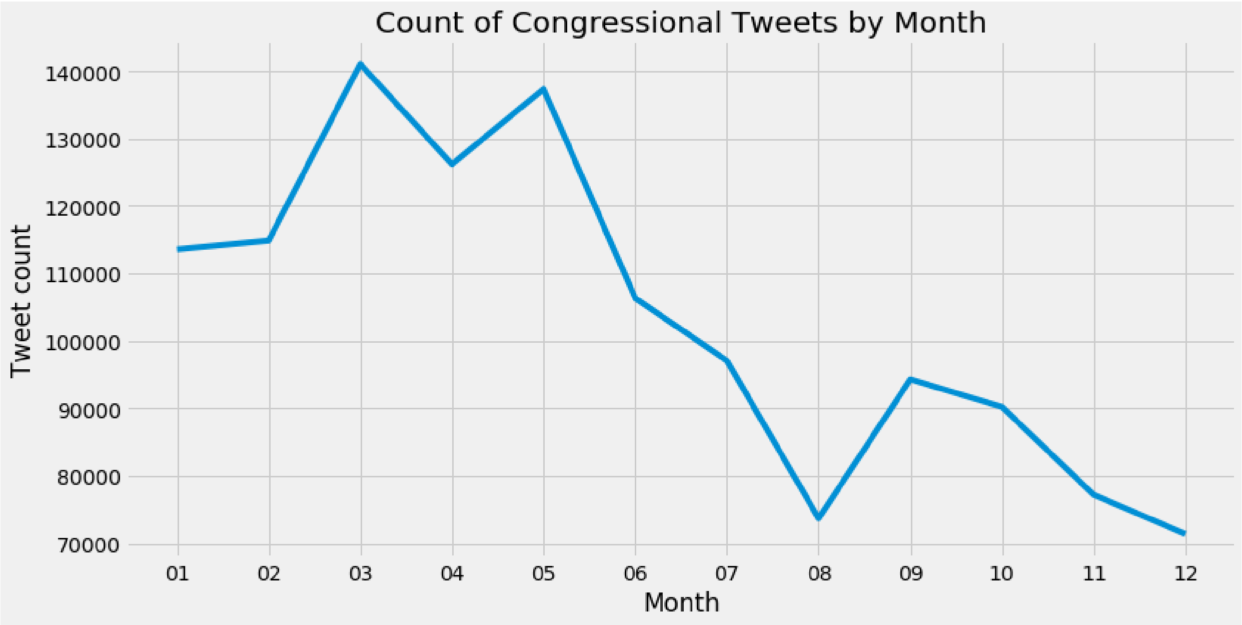
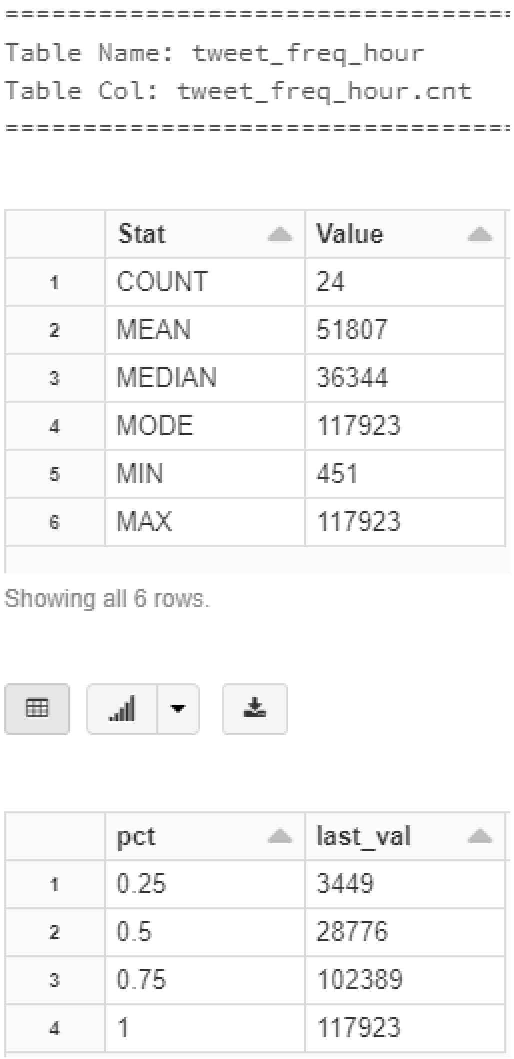


Figure 7: Count of congressional tweets by month

Month 1 to 6 have more tweets than any month from 7 to 12. With fewest tweets in December and highest tweets in March.



```
==================================
Table Name: tweet_freq_hour
Table Col: tweet_freq_hour.cnt
==================================
```

	Stat	Value
1	COUNT	24
2	MEAN	51807
3	MEDIAN	36344
4	MODE	117923
5	MIN	451
6	MAX	117923

Showing all 6 rows.

	pct	last_val
1	0.25	3449
2	0.5	28776
3	0.75	102389
4	1	117923

Figure 8: Descriptive stats of tweets frequency by hour

Interquartile percentage is represented with pct. There is a significant difference between the mean and median of tweets frequency by hour. The minimum tweets is very small compared to the maximum tweets count by hour.

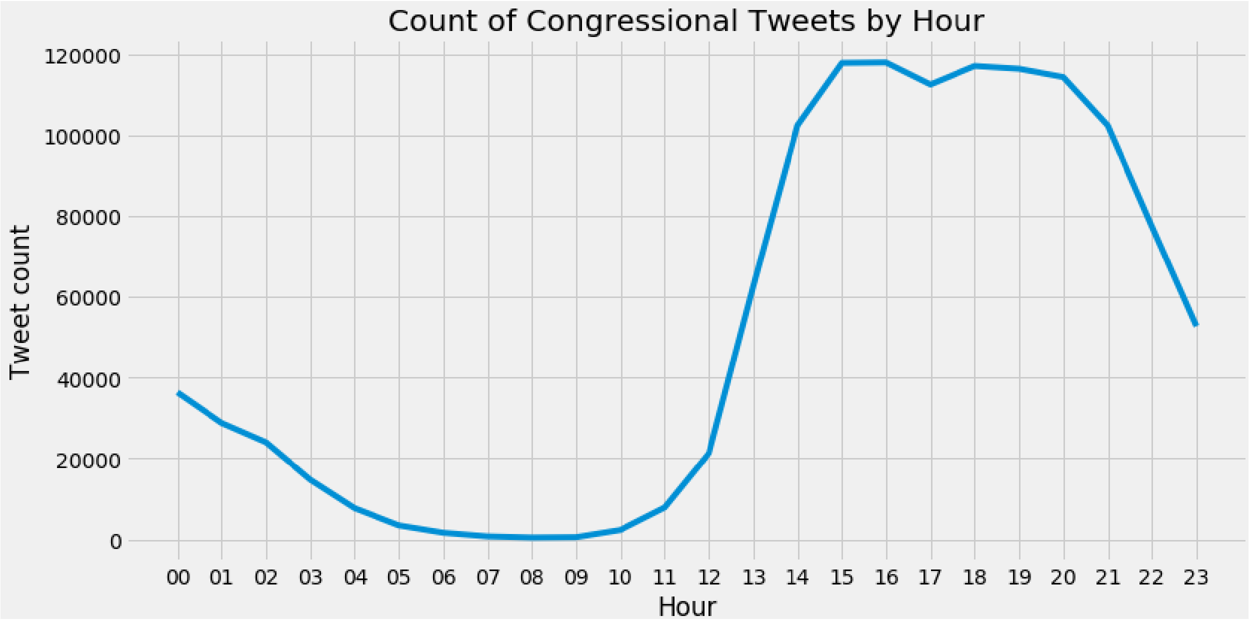
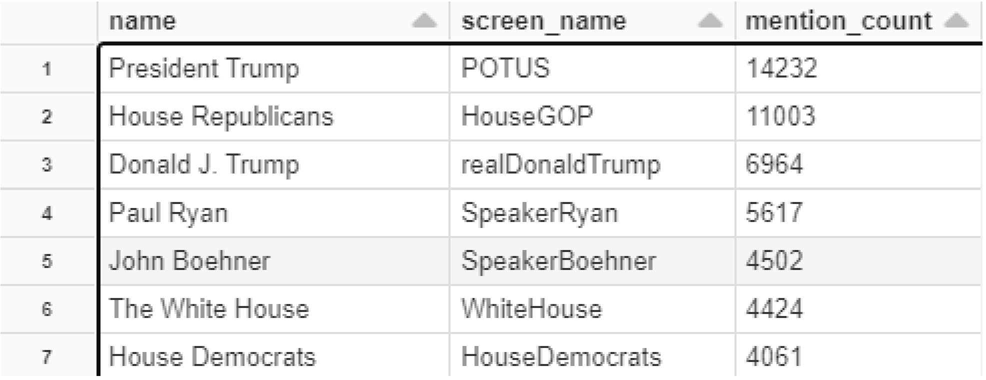


Figure 9: Count of congressional tweets by hour

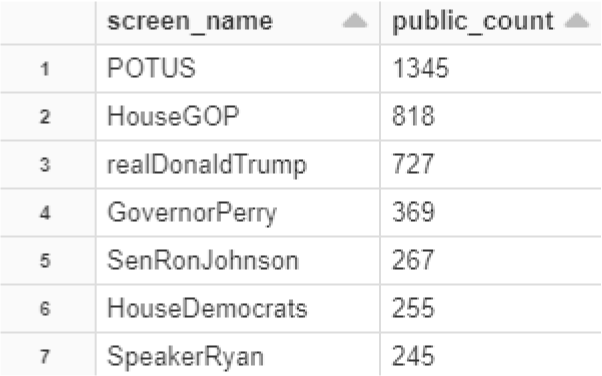
Most tweets happen between 3-8pm. Tweets count is lowest between 5-10am.



	name	screen_name	mention_count
1	President Trump	POTUS	14232
2	House Republicans	HouseGOP	11003
3	Donald J. Trump	realDonaldTrump	6964
4	Paul Ryan	SpeakerRyan	5617
5	John Boehner	SpeakerBoehner	4502
6	The White House	WhiteHouse	4424
7	House Democrats	HouseDemocrats	4061

Figure 10: Mention count of users

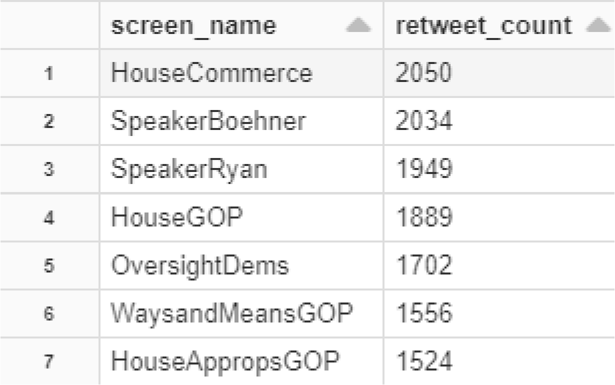
The top 5 users with the most mentions are republicans.



	screen_name	public_count
1	POTUS	1345
2	HouseGOP	818
3	realDonaldTrump	727
4	GovernorPerry	369
5	SenRonJohnson	267
6	HouseDemocrats	255
7	SpeakerRyan	245

Figure 11: Public counts of users

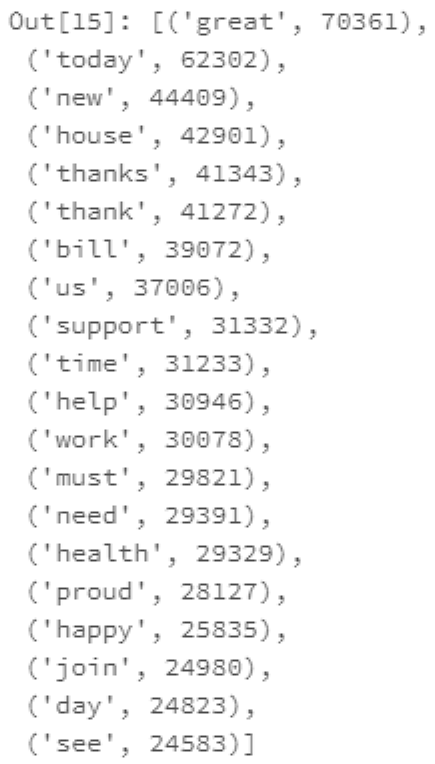
Other than the top 6 user the rest of the top 7 users with the highest public counts are republicans.



	screen_name	retweet_count
1	HouseCommerce	2050
2	SpeakerBoehner	2034
3	SpeakerRyan	1949
4	HouseGOP	1889
5	OversightDems	1702
6	WaysandMeansGOP	1556
7	HouseAppropsGOP	1524

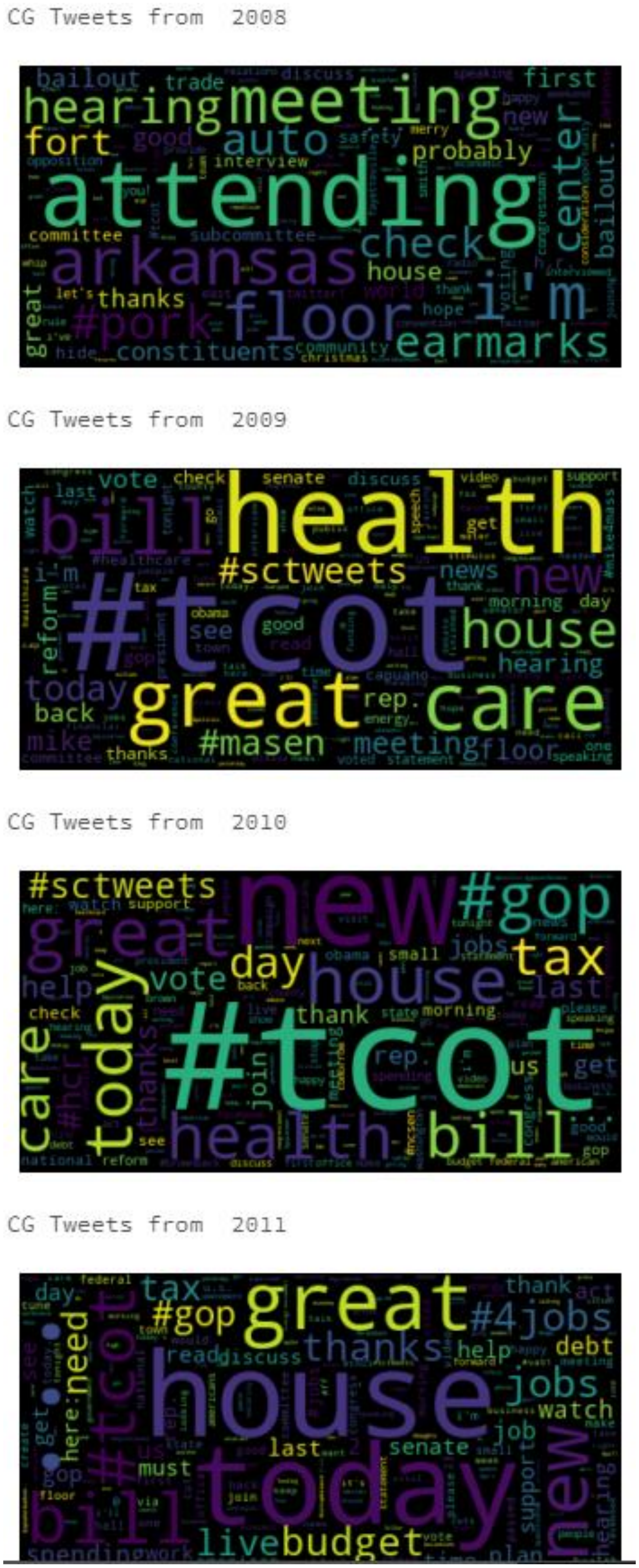
Figure 12: Retweet counts of users

The top 7 users are republican except for the 5th user is a democrat.



```
Out[15]: [('great', 70361),
 ('today', 62302),
 ('new', 44409),
 ('house', 42901),
 ('thanks', 41343),
 ('thank', 41272),
 ('bill', 39072),
 ('us', 37006),
 ('support', 31332),
 ('time', 31233),
 ('help', 30946),
 ('work', 30078),
 ('must', 29821),
 ('need', 29391),
 ('health', 29329),
 ('proud', 28127),
 ('happy', 25835),
 ('join', 24980),
 ('day', 24823),
 ('see', 24583)]
```

Figure 13: Word counts of most used words



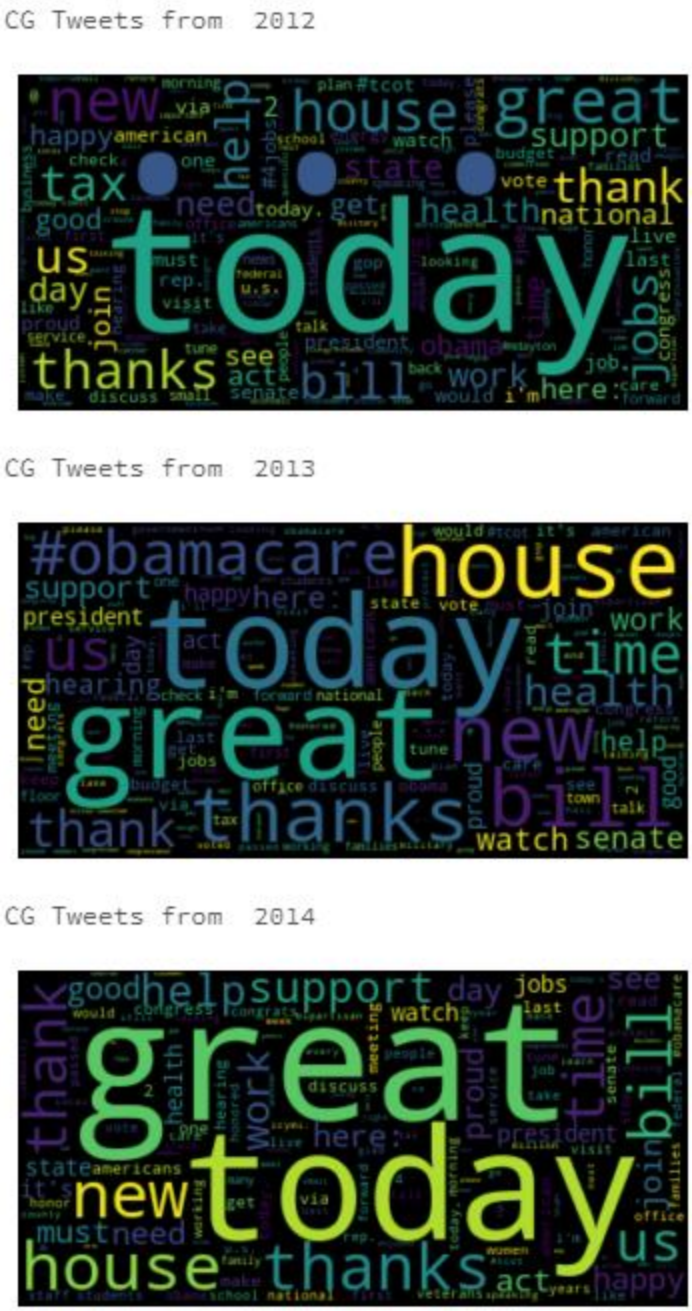
CG Tweets from 2008

CG Tweets from 2009

CG Tweets from 2010

CG Tweets from 2011

Figure 14: Word clouds from 2008-2011



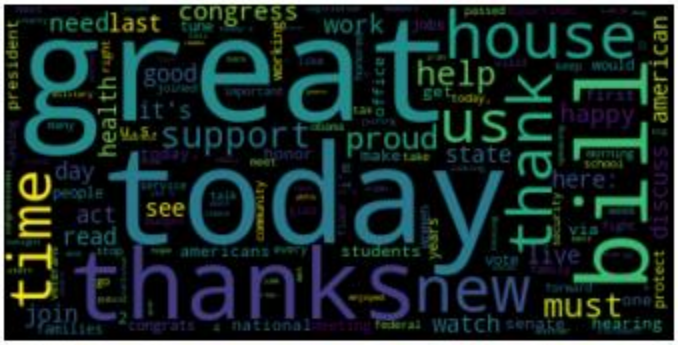
CG Tweets from 2012

CG Tweets from 2013

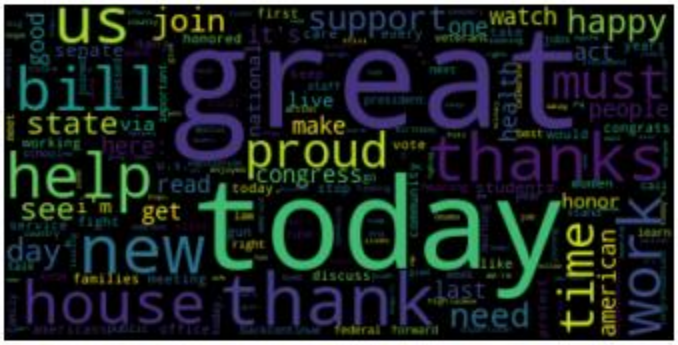
CG Tweets from 2014

Figure 15: Word clouds from 2012-2014

CG Tweets from 2015



CG Tweets from 2016



CG Tweets from 2017

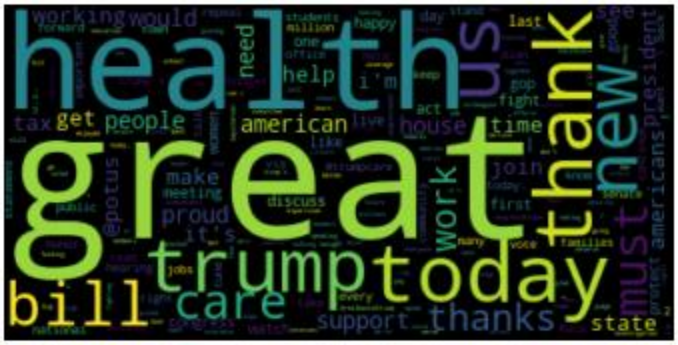


Figure 16: Word clouds from 2015-2017

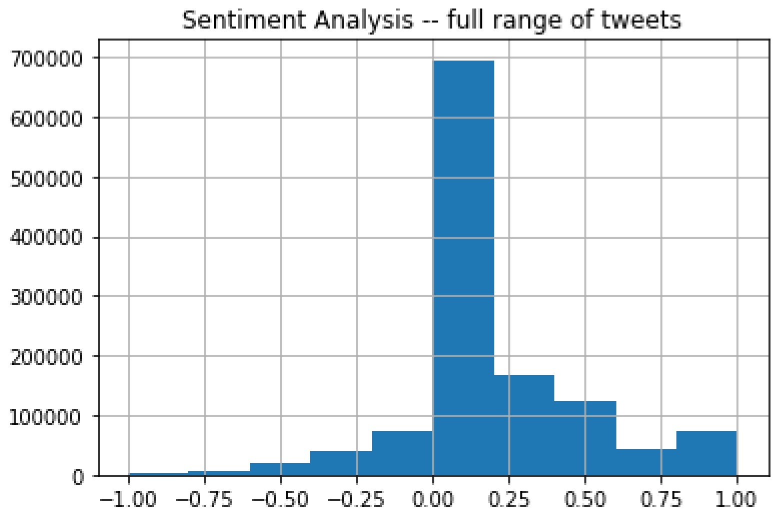


Figure 17: Sentiment Analysis

Discuss Insights Discovered

Key Points Discovered About the Data:

1. Republicans are very popular on twitter, followed by democrats.
1. The most popular topics in 2009 and 2010 are tcot and one of the most popular topic in 2017 is trump.
2. The sentiment analysis shows that tweets have mostly neutral sentiment and slightly more positive than negative.

Proved and Disproved Hypotheses:

1. The hypotheses that users are involved in politics is true as the top 7 users are politicians.
2. The most popular topics are not all associated with politics. However, top political topics are associated with the conservative party more than the liberal party.
3. The hypothesis that political tweets will have negative sentiment is disproven.

Additional Questions:

1. What topics are discussed most often?
2. Is the sentiment of tweets mostly positive or negative?

Relationships / Correlation and Textual Analysis:

1. There is a positive correlation between year and tweet counts. The number of tweets increases every year from 2008 to 2016. The tweet counts are predicted to continue increasing for a few more years.
2. Doing Textual Analysis for Term Frequency-Inverse Document Frequency of the tweets, it was found that 'great' is the most used word, which was used 70361 times, followed by 'today' being used 62302 times. The third being 'new' was used 44409 times.

Go Broader:

1. The top 20 words used in tweets are either positive or neutral. The top 20 words in descending order are great, today, new, house, thanks, thank, bill, us, support, time, help, work, must, need, health, proud, happy, join, day, and see.
2. In the word cloud of 2009 and 2010, '#tcot' which stands for Top Conservative On Twitter, was the most used word in tweets. This shows the conservative party focused on using Twitter to influence the population significantly more than the liberal party in 2009 and 2010.
3. In 2017, trump was in the top 10 of the most frequently used words. This coincides with the election of Trump.
4. Sentiment analysis of tweets was considered to obtain insight on whether most tweets are positive or negative.

New Metric:

1. Mention count, public count, and retweet count was created to find the most popular users of congressional tweets.
2. Word count metric was used to track popular topics.
3. The sentiment was created as a new metric to track how positive or negative tweets are and how the sentiment change over time.

Recommendations and Actions

1. As the most famous users are mostly republicans, democrats can put more effort in finding ways to get engagement on Twitter. On the other hand, republicans can more easily spread their influence by working with one of the top republican users.
2. The word great was heavily used in tweets, thus replacing it with words such as exceptional, extraordinary, or prominent can probably make your tweets more memorable.
3. Since the sentiment of tweets are mostly neutral, sending tweets that are very positive or very negative can make your tweet stand out more.