**Homework 2**

**IST 652**

**David Doman**

**Data Source & Preprocessing**

The program and results described in this analysis utilized data from the well-known social media platform, Twitter. Tweets from users were pulled down from Twitter using API Keys obtained by applying for Developer access. One does not have to have a twitter account to obtain information contained in tweets distributed from Twitter users. The “Tweepy” package was used to accomplish pulling in the tweets. Code provided from class was used to bring in the tweets and to clean it up so that it was able to be transformed into a format provides understanding. Ultimately, the resulting format will be in the JSON format. The first search initializes a cursor that is stored in the metadata results and allows additional searches to return more tweets. Additional preprocessing of the tweets once they are pulled in includes changing names to lowercase and removing any special characters, such as hashtags or spaces. The final step in preprocessing is saving the tweets to a database.

I will not include any snippets of the preprocessing code in this report but will submit the code in addition to this report.

**Question: How do the two names “Wentz” and “Brady” compare when analyzing Twitter data?**

**Analysis**

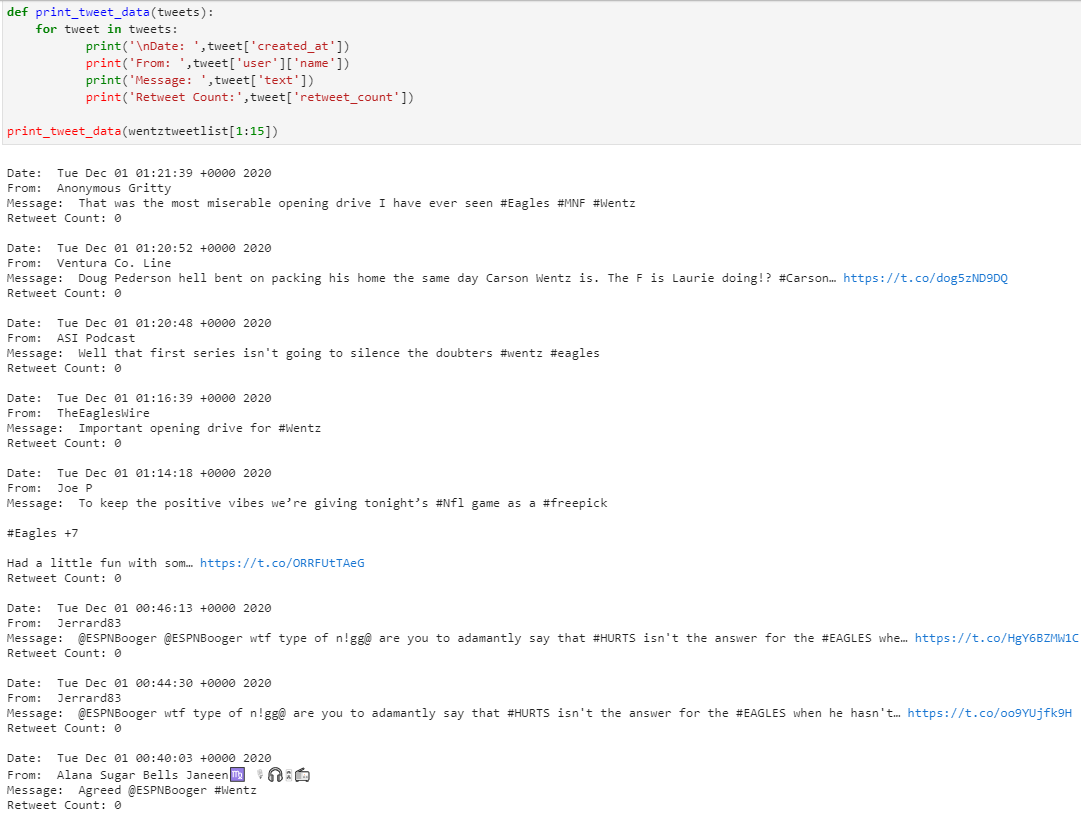
The data for “Wentz” and “Brady” was analyzed in two separate files. Only snippets of code for “Wentz” will be shown throughout the rest of this report since the code for “Brady” is identical outside of replacing the name.

First step was to import the “pymongo” package in order to access the Mongodb, where the twitter data resides. This is shown in the code below:

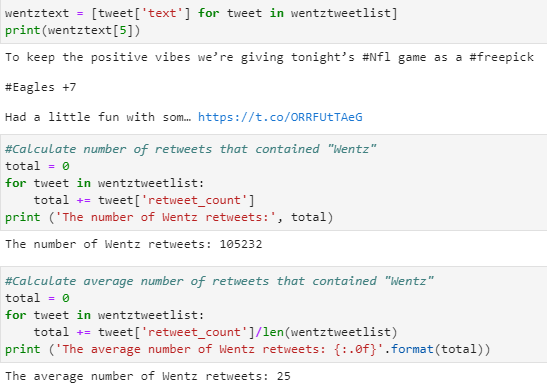


The last section of code above shows the collection of the number of tweets that contained “Wentz”. “Wentz’s” 4,289 was drastically higher than “Brady’s” 940.

Next, I printed out various fields from the database including when the date was created, the username of who sent the tweet, the message sent and the retweet count.



Next, the number of retweets and average number of retweets was compared between the two. This is shown below:



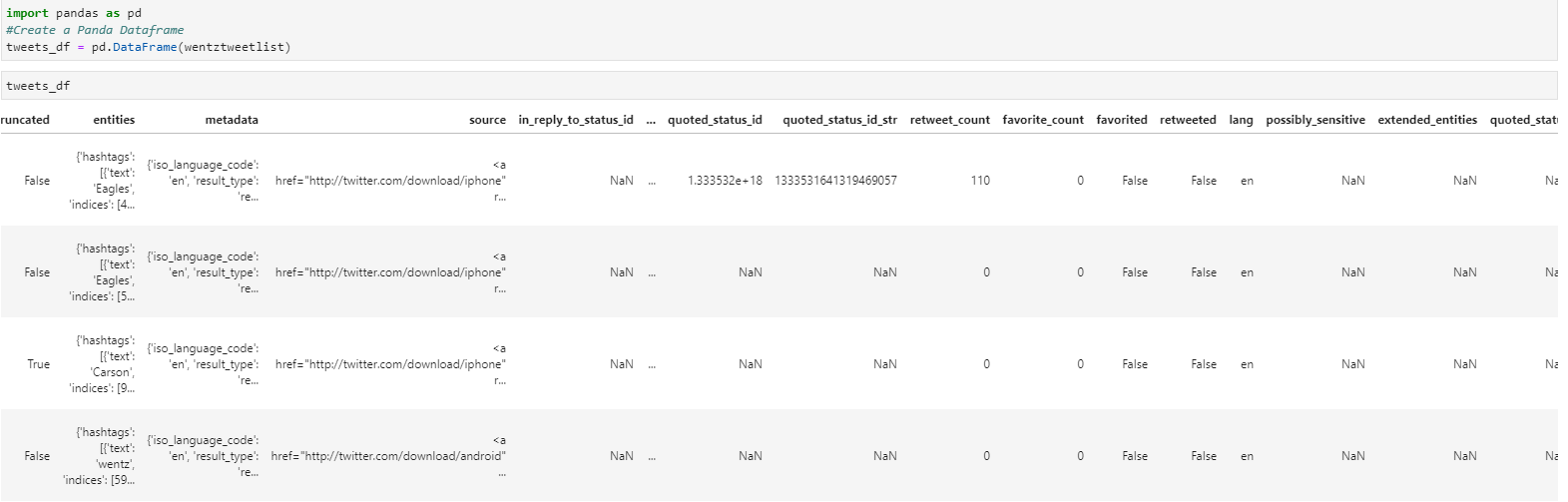
Wentz Retweets: 105,232

Brady Retweets: 1,551

Wentz Avg Number of Retweets: 25

Brady Avg Number of Retweets: 2

Next, a Panda Dataframe was created to allow for further analysis of the data. A snippet is shown below:



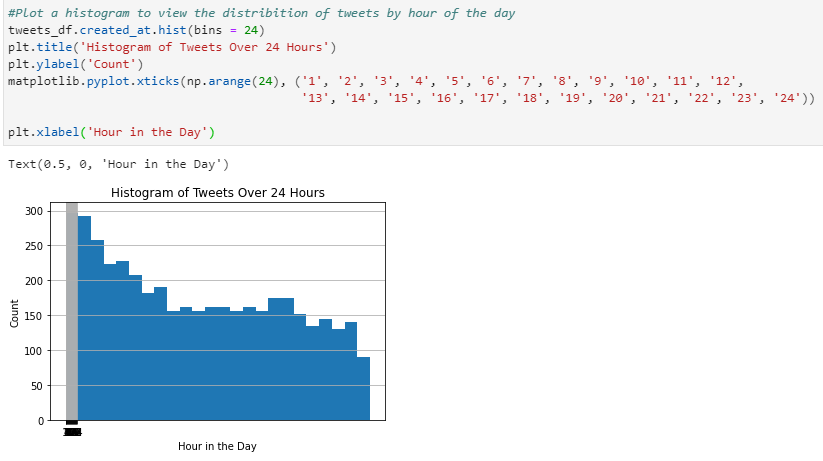
The number of likes were calculated below:

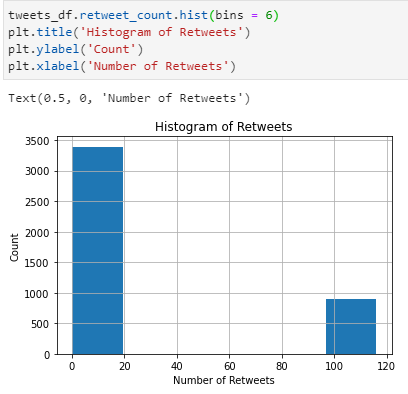


Wentz likes: 5,197

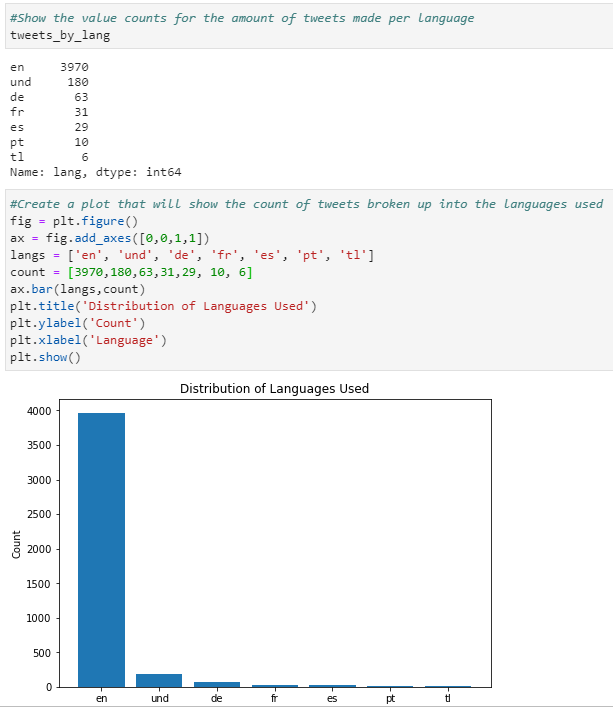
Brady likes: 1,317

“matplotlib” and “numpy” were next used to create visuals of the data.





Lastly, the count of tweets was calculated by language:



**Results**

It was shown from above that “Wentz” had many more tweets with his name than did “Brady”. Looking at the first 15 records from Pandas, there were some November tweets for “Brady” while for “Wentz” there were only December tweets. This may be attributed to the fact that this analysis was completed during a game that “Wentz” was playing on national tv. He was a current interest at the time so that may have contributed to higher tweet and retweet numbers. There was a large skew to the left for “Wentz” as compared to “Brady” for tweets by hour in the day also probably due to the fact that “Wentz” was current at the time. The language comparison was similar between the two with English having by far the most tweets by language. This also makes sense since both “Wentz” and “Brady” reside in the U.S. Overall, this analysis was very interesting in pulling information that ordinary individuals or businesses tweet. Further analysis could present more accurate and interesting findings.