Report: act_report

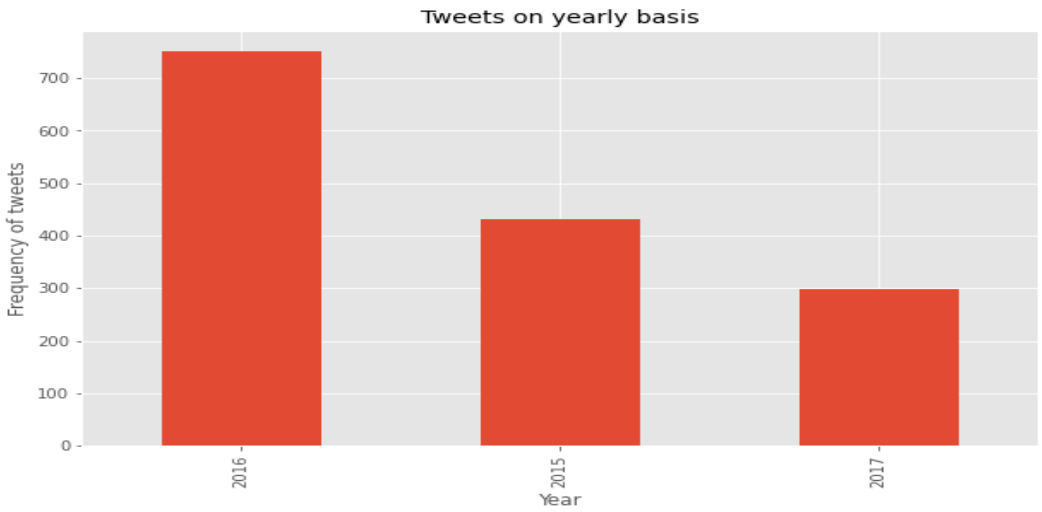
Introduction:

In this project, we gather, assess, and clean data and then act on it by wrangling the data.

- WeRateDogs Twitter archive was downloaded from Ucadity (given to us by downloading it from the classroom)
- Twitter Image Prediction (Image_predictions.tsv) was downloaded programmatically from the URL using the Requests library
- Additional data from Twitter was downloaded using Twitter API (Tweepy) and stored as tweets_data.txt
- Data was then assessed visually and programmatically before being cleaned for exploratory analysis.

Insight and visualizations from the data

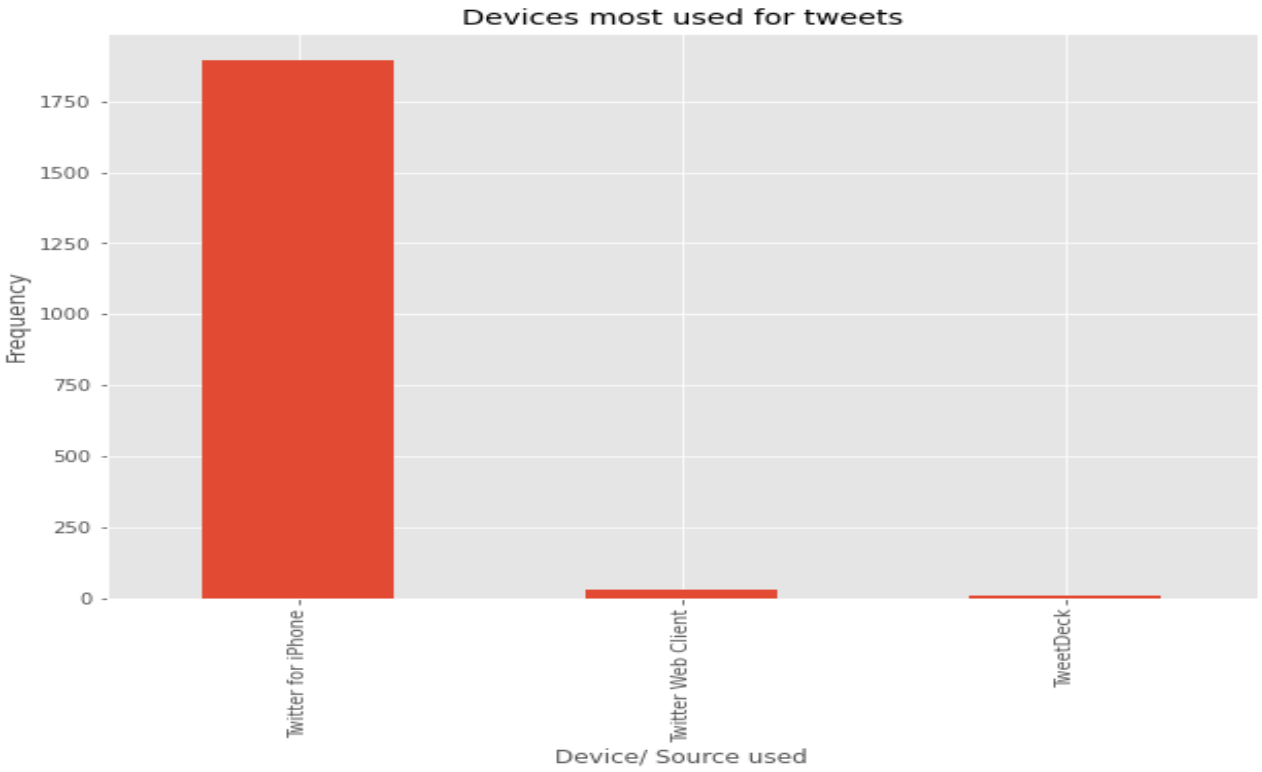
Question 1: How many tweets are posted yearly?



From the bar chart, it is discovered that the year 2016 recorded the greatest number of tweets, followed by the years 2015 and 2017 respectively.

The findings and insight produced from wrangling the data to answer the question above, we found that the years 2016, 2015 and 2017 recorded 963 tweets, 599 tweets and 368 tweets respectively. From the bar chart visualized based on these figures, it is discovered that the year 2016 recorded the greatest number of tweets, followed by the years 2015 and 2017 respectively.

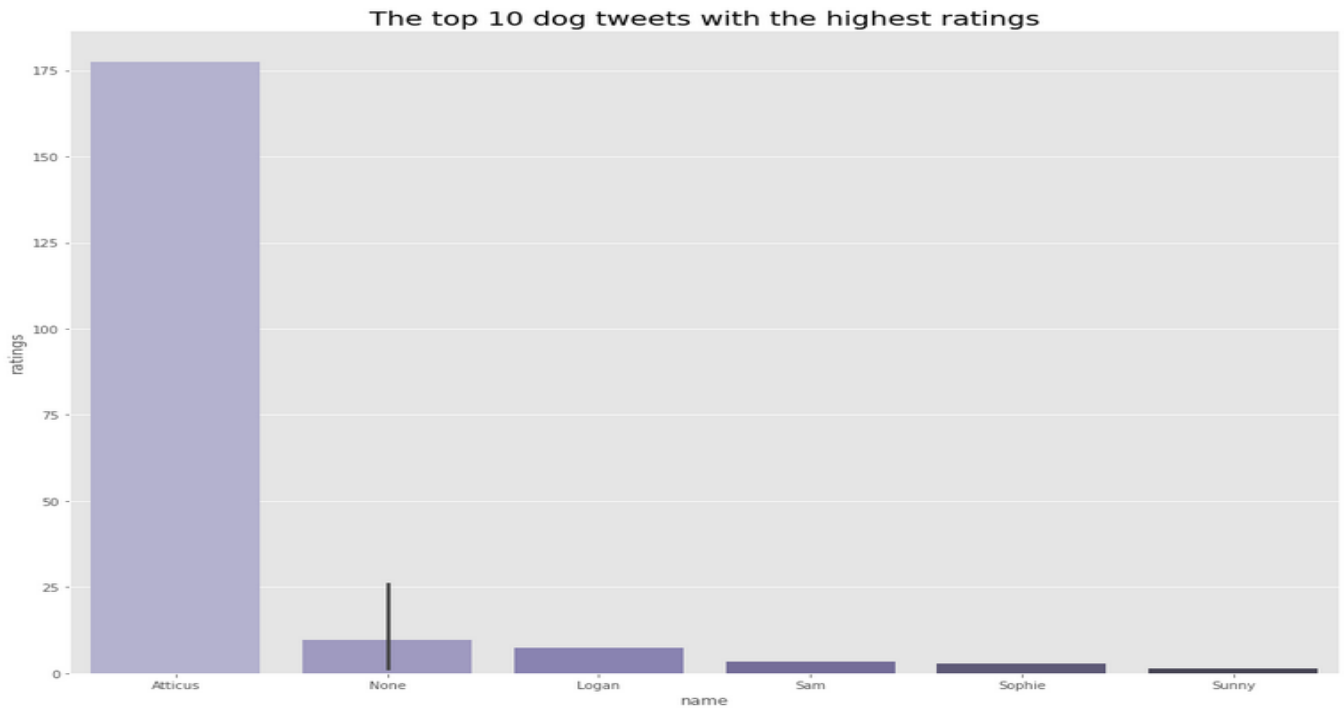
Question 2: What kind of devices were the tweets posted from?



From the bar chart, it can be deduced that tweets from Twitter for iPhone were the highest, followed by Twitter web clients and TweetDeck respectively.

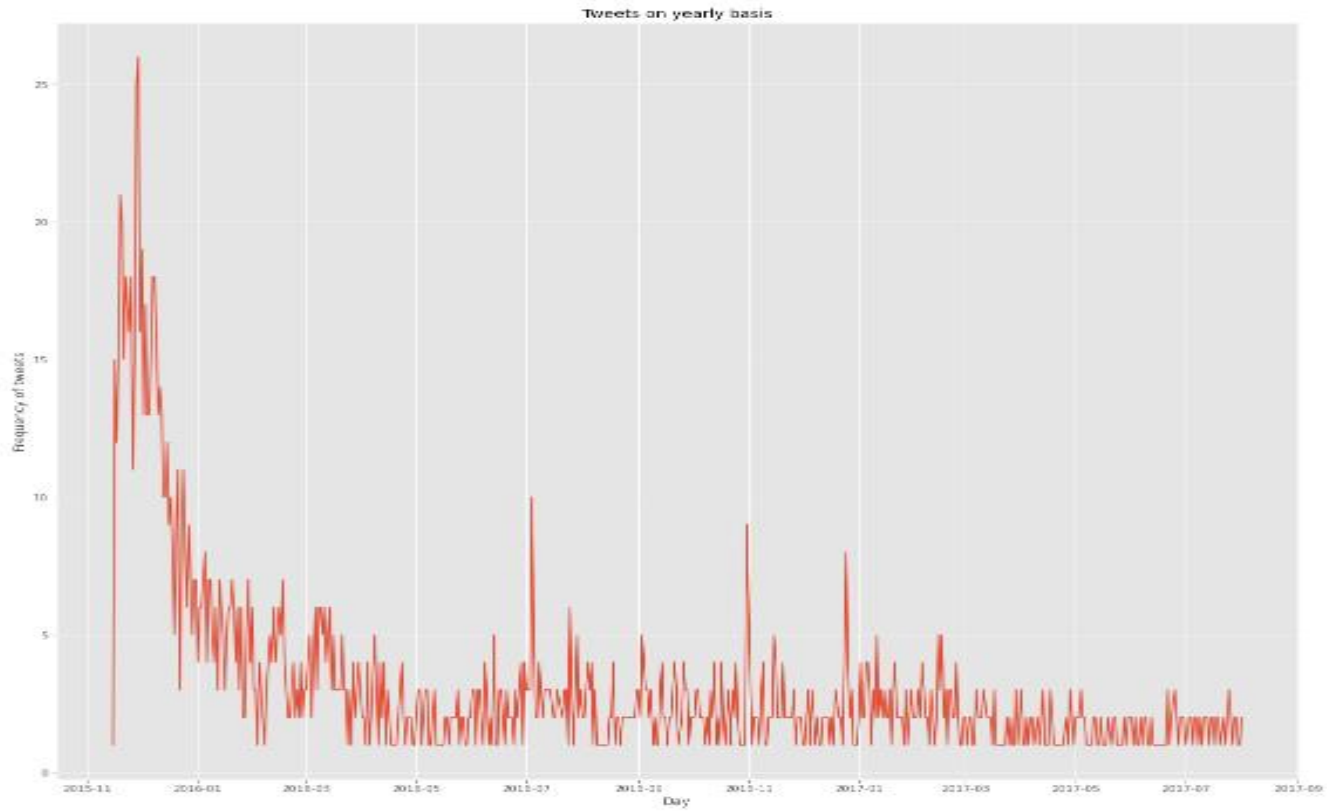
Figuratively, sources of tweets from Twitter for iPhone was 1894, followed by Twitter Web Client which also was 27 and tweets from TweetDeck recorded was 9. From the bar chart visualized based on these figures, it can be deduced that tweets from Twitter for iPhone were the highest, followed by Twitter web clients and TweetDeck respectively.

Question 3: Which Dogs had the highest tweet ratings (Top 10)?



From the bar chart, the top 10 dog tweets were Atticus, None (dogs without name in the tweet), Logan, Sam, and Sophie, to mention but a few respectively. However, from 5th place to 6th place are names of dogs which was probably a typing mistake or rare names people call their dogs.

Question 4: What is the nature of tweets daily from 2015 to date?



From this line chart, it can be deduced that there is a sharp drop in the number of tweets from 2015-11 to 2026-01 and a gradual rise and drop in the tweets from 2016-03 to 2017-09.

The line chart visualized based on the question stated above indicates that there is a sharp drop in the number of tweets from 2015-11 to 2026-01 and a gradual "rise and drop" in the tweets from 2016-03 to 2017-01. It can also be noticed that there were occasional peaks between "2016-05 and 2016-09" and “2016-03 to 2017-09" respectively.

Limitations:

I struggled with fully grasping what some of the columns in the datasets represented; for example, the p1_conf column, and p2_conf column, to mention but a few. Again, there were 0 non-null values in the favourite_count column in tweets data downloaded using tweepy (twitter API) which was a vital part of the data but was not present because it had NaN values. Another limitation was that the datasets were very huge (numerous records) hence assessing and cleaning were very difficult to carry out.

References

- https://stackoverflow.com/questions/28384588/twitter-api-get-tweets-with-specific-id
- https://www.youtube.com/watch?v=bNDRiaFyLrs&t=28s
- https://www.youtube.com/watch?v=Lu1nskBkPJU
- https://medium.com/analytics-vidhya/fetch-tweets-using-their-ids-with-tweepy-twitter-api-and-python-ee7a22dcb845
- https://stackoverflow.com/questions/54407027/use-api-to-write-to-json-file
- https://www.geeksforgeeks.org/python-convert-tsv-to-csv-file/
- https://sparkbyexamples.com/pandas/pandas-change-string-object-to-date-in-dataframe/
- https://stackoverflow.com/questions/44116194/import-a-function-from-another-ipynb-file