Project Report

The goal of this Project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. WeRateDogs is a Twitter account that rates people's dog in a humorous way. The steps taken to complete this project were: Gathering data, Assessing data, Cleaning data, Storing data and Analysing and visualizing data and reporting.

Gathering Data

The data used for this project was gathered from three datasets, namely the ***twitter_acrhive_enhanced.csv***, ***image_prediction.tsv*** which was downloaded programmatically from one of Udacity's servers and ***text-json.txt,*** Using Pandas all the file s were read into a DataFrame. The first two file were read directly with exception of the text-json.txt which had its individual lines read as JSON and then converted into a DataFrame.

Assessing data

The datasets were assessed programmatically and visually. Here are some quality and tidy issues associated with them.

Quality issues

1. The expanded_url column should be dropped for this analysis
2. There are missing values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id columns
3. Wrong rating numerators and rating denominators
4. The timestamp column had a wrong datatype
5. The actual source values should be extracted from the anchor tags
6. Some of the tweets are not about dogs
7. The data had retweet and reply data.
8. The p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog column should be dropped.
9. There are names like "None", "a", "quite", "not", "one", "incredibly", "an", "very", "not", "actually", "just", "getting", "mad", "unacceptable", "all", "infuriating", "such", "by", "the", "life" which are not proper names.
10. The tweet_id should be changed from int to string
11. There were inconsistencies in the breed names under the breed column

Tidy issues

1. doggo, floffer, pupper and puppo columns should collapsed into a single column since they are rather observations
2. The three dog prediction algorithms (p1, p2, p3) should be collapsed into a single column.

Cleaning data

The quality and tidy issues were cleaned programmatically using the Define-Code-Test format.

I started working on missing values first. The third DataFrame didn't really have issues. Most the issues were associated with the first DataFrame(twitter_archive_enhanced.csv).

Storing data

The datasets were merged into a one DataFrame and stored as ***twitter_archive_master.csv***. This was used for the analysing and visualization section of the project

Analysing and Visualizing

The merged data was used to answer, create insights and visualization for the following questions.

- Which breed had the highest average rating, favorite count and retweet count?
- Do dogs with higher ratings have higher retweets counts?
- Do dogs with higher ratings have higher favorite counts?
- What are the top 10 most tweeted breeds.
- What is the relationship between retweet_count and favorite_count?
- What was the highest rating attained by a dog and what was the impact of the number of images on the retweet and favorite counts on high rated dogs?

While answering these questions realised there were outliers and hence at some points the data needed to be trimmed. The findings made from the analysis were put under Conclusions.