Summary

This report describes the wrangling and data analysis process for project # 4 of Udacity Data Analyst Nanodegree Program

Gathering Data

In this section, the data was gathered from 3 different sources as follow:

1. The WeRateDog twitter archive file was manually downloaded from Udacity website as a CSV file and imported into Pandas called df1.
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network was downloaded programmatically using the Requests library. This file is hosted on Udacity's servers as a flat tsv file. This file also was imported to Pandas dataframe called df2 after downloading.
3. Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a .txt file. Each tweet's JSON data was written to its own line. This .txt file is then read line by line into a pandas DataFrame called df3 with tweet ID, retweet count, and favorite count

Assessing Data

In this section, we assessed the 3 data frames we obtained above, visually, and programmatically for quality and tidiness issues and took notes on any possible issue for addressing later.

As required by the project, the goal was to detect and document at least eight (8) quality issues and two (2) tidiness issues to address key requirements for this project as:

- We only wanted original ratings (no retweets) that had images. Though there were 2356 tweets in the dataset, not all were dog ratings, and some were retweets.
- Merging individual pieces of data according to the rules of tidy data.

After assessing all the 3 dataframe gathered above and are called df1, df2 and df3, I came up with the following list:

Quality issue observations

- *rating_numerator* has outliers with min = 0 and max = 1776
- *rating_denumerator* has outliers with min = 0 and max = 170
- Dog stage can be categorial
- *Null* represent as *None* in dog stage, also the 5 columns of dog stage could be melt into one column
- 745 dog names are missing
- 55 dogs named as *a*
- Potential for duplicate names such as *Charlie*, *Lucy*, *Cooper* or more (to be investigated)
- *timestamp* datatype is incorrect
- 181 of the tweets are retweeted as there is value for *retweeted_status_id*
- df1 has 281 more tweets with images than df2
- df1, df2 and df3 need to be merged at the end

Tidiness issue observations

- Convert the dog stage 4 columns to one column representing the stage in df1
- Predicted dog names inconsistent on the first letter in df2
- Only the dog information for the highest probability needed in df2

Cleaning Data

As per our observation above, I have defined the cleaning process as below:

Tidiness issues

1. Fix the dog stage columns by putting them all into one column
2. Make the dog names consistent in df2
3. Drop the columns related to the second and third rank predictions in df2

Quality issues

1. Fix outliers in *rating_numerator* in df1
2. fix outliers in *rating_denumerator* in conjunction with *rating_numerator* by calculating the actual rating
3. Make dog stage categorial
4. *Null* represent as None in dog stage
5. 745 dog names are missing, 55 dogs named as *a* and the rest as *None*
6. Fix dog names listed as *None*
7. Potential for duplicate names such as *Charlie*, *Lucy*, *Cooper* or more (to be investigated)
8. Fix *timestamp* datatype
9. Drop 181 *retweeted_status_id*
10. Merge df1, df2 and df3
11. Drop the tweets that do not have images and retweeted

The above issues are addressed one by one by coding and testing as shown in ***wrangle_act.ipynb***

Storing, Analyzing and Visualizing Data

Once the cleaning process is finished, the cleaned single dataframe is saved into a CSV file then I thought about some interesting questions to ask about this dataframe that could be used for analysing and visualization. As the result, the following analyses were performed on this dataframe:

1. Which names are the most popular dog name?
2. Which breeds are the most popular breed?
3. What are the 5 top rated breeds?
4. What are the 5 top retweeted breeds?
5. What are the 5 top favorite breeds?
6. Which dog had the most retweet counts?
7. Which dog had the most favorite counts?
8. Which dog had the highest rating?
9. What is the dog stage distribution?