Wrangle Report

The data wrangling project was very challenging, and I learned a lot about the data gathering process and the Twitter API.

Data Gathering

I gathered data from three different sources for this data analysis. WeRateDogs gave Udacity exclusive access to their Twitter archive for this project in the form of a csv file.

The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

This archive contains basic tweet data (tweet ID, timestamp, text, etc.) . I queried the Twitter API for each tweet's JSON data using the Python's Tweepy library I stored each tweet's entire set of JSON data, which I would later use to analyze the tweet's retweet and favorite (i.e. "like") counts.

Data Assessing:

The data assessing was divided into two-part Quality and Tidiness Issues.

Quality Issues:

As I have investigated the dataset I have fine lots of quality issues, that I must correct before making our analysis and visualization.

- As per the project requirements; only original ratings (no retweets) that have images should be included.
- Columns of retweeted_status_id and its related data have entries which not part of our analysis.
- Inaccurate data in name 'a' represented 55 times.
- Inaccurate data in rating_nominator and df_denominator.
- Null values reprsened as "None" in columns doggo, floofer, pupper, puppo.
- Invalid dtype in timestamp represented as object type.
- Column tweet_id represented as int64 across all data set.
- Columns of predictions have lables represented as letters.
- Column 'p1_dog' have 543 false prediction, p2_dog have 522 false prediction, p3_dog have 576 false prediction.
- Column 'p1' have invalid data like(china_cabinet, shield, orange, walking_stick,...etc).
- Column 'p2' have invalid data like(ice_lolly, Japanese_spaniel, china_cabinet, necklace,...etc).
- Column 'p3' have invalid data like(kimono, cab, axolotl, passenger_car, tripod, grocery_store,...etc).

Tidiness Issues:

- Tidiness issue that had been taken care of were as follows:
- Variable dog stages have four columns. df_2:
- Variable prediction have three columns. df_3:
- Table df3_clean have same observation unit that in df1_clean.
- Column name id in stead of tweet_i.

Data Cleaning

All cleaning was done programmatically, firstly I have made copies from the data frames to keep our original dataset unmodified. After each issue was defined, solution was coded and tested, and then data was stored as requested in a file name twitter_achive_master.csv and then I have taken all the insights using python pandas library and for the viz I have used matplotlib.