Data Gathering

The first step that I took was the data gathering phase. In this phase, I first uploaded a file on hand (**twitter_archive_enhanced.csv**) to the Jupyter Notebook directly for the twitter archive. From there, I then download the tweet image predictions file (**image_predictions.tsv**), that was generated via a neural network, from the Udacity servers. I did this programmatically using the Requests library. Following this, I then created a Twitter developer account in order to access the Twitter API to query it for each tweet in the twitter archive file that was already on hand. I queried for each tweets information via the tweet ID's available in the twitter archive file. The purpose of this was to store each tweet's entire set of JSON data in a file called **tweet_json.txt**. Following all of this, each of the three obtained files were read into a separate pandas Dataframe within the Jupyter Notebook **wrangle_act.ipynb**. For the **tweet_json** dataframe, I only included the '**id**', '**retweet_count**', and '**favorite_count**' columns, as these were the only data columns specified as needed by the project requirements.

Data Assessment

During the data assessment phase, I ran a series of code statements to both visually and programmatically check for quality and tidiness issues. This included displaying the entire tables inline in the Notebook, gaining a general overview of the count and datatypes for each column, and diving deeper into each table and the values within specific columns to find any issues. During this phase, I identified 8 data quality issues and 2 tidiness issues.

Data Cleaning

During the cleaning phase, I created copies of the original data to clean. After this, I tackled the 2 tidiness issues first, which were: each of the dog stages had their own column in the **tweet_json** table, all 3 tables were better off as one table (**twitter_archive_master**). Following this, I cleaned the 8 data quality issues below:

- Table contains some tweets that are just retweets, which are not supposed to be included
- Table contains some tweets that are just replies, which are not supposed to be included
- Table contains some tweets that are not original ratings with images, which are not supposed to be included
- 'tweet_id' isdatatype int instead of string
- 'timestamp' is not datatype datetime
- Created 'dog_stage' column non-null "None" values that should be null, as well as messy entries such as 'doggoNoneNoneNone', 'NoneflooferNoneNone', etc
- There are non-names (such as *a, an, the, this*, etc) in the 'name' column as well as non-null 'None' values.
- Denominator column contains incorrect values

After completing my cleaning tasks, I saved the master cleaned file as a CSV called **twitter_archive_master.csv**.

Data Analysis

In the data analysis phase, I answered 3 main questions: what are the top 5 most common dog names, what is the average retweet and favorite count per tweet, and what are the most common dog stages. After completing this, I found the following 3 insights:

1. **The top five most common dog names are 'Charlie', 'Tucker', 'Cooper', 'Lucy', and 'Penny'/'Oliver' (tied for 5th).**
2. **The average WeRateDogs retweet count is roughly 2618 retweets, and the average favorite count is approximately 8636 favorites.**
    a. These numbers are only referring to the retweet count and favorite count for tweets that contain original ratings. Additionally, this is averaged out over the entire lifespan of the WeRateDogs account, including before it reached it currently high levels of popularity.
3. **By a fairly large margin, dogs in the ‘pupper’ stage (small and young) are overwhelmingly shown with 201 instances, compared to the next closest dog stage of ‘doggo’ (big and older) at 63.**
    a. Only around 15.3% of WeRateDog original rating tweets included a dog stage.