WRANGLING EFFORTS

Python libraries used: Pandas, NumPy, re, io, Tweepy and os.

Gathering Data

Three datasets required to create a master dataset, they are:

1. twitter_archive - This dataset was obtained manually from Udacity
2. image_predictions - This dataset was downloaded programmatically from Udacity servers using the requests library. The dataset was then read as a StringIO object for scalability and reproducibility.
3. df_json - This dataset was created by requesting additional data for each tweet in the twitter_archive dataset through twitter's API. We used the tweepy library to assess Twitter API. The data collected was the favorite and retweet count for each tweet. The collected data was in JSON format, so we converted it from JSON to CSV file and saved the created dataset as df_json.csv.

Assessing Data

For the twitter_archive, image_predictions, and df_json datasets, we assessed the data programmatically using pandas functions such as info(), head(),value_counts(), sample(). Then we assessed the data visually by observing samples of the data.

We also checked for duplicate values in each dataset. There were no duplicate rows.

CLEANING PROCESS

We made a copy of each dataset and then we proceeded to clean the dataset copy.

Twitter_archive dataset

1. ACCURACY ISSUE: The rating_denominator column has denominators other than 10: This problem occurred due to the following reasons:
   a. The data was captured wrongly from the text column: The regex formula used to obtain the rating from the "text" column obtained the wrong numbers as the rating denominator so we changed the denominator of rows with this issue to 10. And we noted if the rating numerator was affected too.
   b. VALIDITY ISSUE - Some rows have decimal ratings, this does not conform to the rating schema: I fixed this issue by filtering the dataset for rows with decimal ratings and changing them to their rounded down values.
   c. VALIDITY ISSUE: There are some rows with multiple dog stages: I fixed this issue by creating unique values for dogs with multiple stages.

d. Some tweets used different scales to rate the dogs: Some ratings were written in multiples of n/10 where n is a number between 7 and 15. These rows had ratings such as 150/100, 99/90, etc. We solved this issue by converting the affected rows to a rating of n/10.

e. Some tweets aren't ratings at all: We deleted rows whose tweets aren’t ratings.

2. ACCURACY ISSUE: The rating_numerator has values higher than 15 and lower than 7: We solved this issue using the same steps as number one.
3. CONSISTENCY ISSUE - The name column has some values without a capitalized first letter: We solved this issue by using the python capitalize function to change each first letter of the name column to a capital letter.
4. COMPLETENESS ISSUE - Columns {in_reply_to_status_id , in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp} Are not useful columns: We solved this issue by removing the aforementioned columns using the drop function.
5. twitter_archive - Columns 'doggo', 'floofer', 'pupper', 'puppo' in twitter_archive should be a single column: I created a column called ‘dog_stage’ which combined the aforementioned columns into a single column.
6. df_json, twitter_archive,image_predictions - The datasets should be joined to create a master dataset: We used the pandas merge function to join all three datasets together.

df_json dataset

1. df_json - Remove columns that aren't useful for analysis: We used the Pandas drop function to remove unnecessary columns.
2. Change id column to tweet_id: We used the Pandas rename function to change the column name.

image_predictions dataset

1. CONSISTENCY ISSUE - The columns {p1,p2 and p3} has some values without a capitalized first letter: We used the python capitalize function to make each of these columns have a consistent value format.
2. The column names are not descriptive enough: We renamed every column in the dataset to a more descriptive name using the rename function. E.g 'jpg_url' to 'image_url'.