WRANGLE REPORT

DATA GATHERING:

The first step of the wrangling process is “Data Gathering” where all the datasets needed for the analysis are pulled from the respective databases/website/server where it lives. In this project I gathered three pieces of data.

- The first one being to manually read in the twitter-archive-enhanced.csv data downloaded on my workstation
- Programmatically downloading the image-predictions.tsv data from the link
- And the third being to query Twitter API using the tweet ID in the twitter-archive-enhanced.csv data to gather each tweet's JSON data using Python's tweepy library and store each tweets entire set of JSON data in a file called tweet_json.txt file.

After downloading the dataset I needed to extract variables of interest like tweet_id, retweet_count and favorite_count from the JSON file downloaded from twitter which I did and read into a dataframe.

DATA ASSESSING:

This is the second and penultimate stage in our wrangling process, here we will be detecting quality and tidiness issues in our data, either visually or programmatically.

For the purpose of this project I documented at least 8 (eight) quality issues and 2(two) tidiness issues.

I went on to read in all three datasets, visually and programmatically inspected them and came up with the following issues in all 3 datasets respectively.

QUALITY ISSUES

Listed below are some of the quality issues found in the data frames

df_archive

1. The name column contains values like "a", "very", "the" e.t.c which start with lowercase and are suspected not to be dog names, also we found 745 null values represented as "None".
2. Null objects in columns (in_reply_to_status_id, in_reply_to_user_id) e.t.c represented as "None"
3. The data type for the timestamp column is object when it should be a datetime.
4. The tweet source should be extracted from the source column to reflect (Twitter for iphone, Twitter Web Client e.t.c)

5. The columns in_reply_to_status_id and in_reply_to_user_id are missing 2278 values each **which might not need cleaning** as some tweets were tweeted directly with images from WeRateDogs account and they were not in reply to any tweet. Moreover we are only interested in original tweets so we will be dropping the observations in this column.
6. Likewise the columns retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp are missing 2175 values each **which might also not need cleaning** will be treated as mentioned above, hence they will be dropped.

df_image

7. This dataframe is missing 281 observations present in the df_archive dataframe.
8. Rename p1, p1_conf and the other column headers to be more reflective column headers.
9. Values in the p1, p2 and p3 columns sometimes start with uppercase and sometimes lowercase

df_json

10. This dataframe is missing 29 observations which are present in the df_archive dataframe.

TIDINESS ISSUES

df_archive

1. The dogstages(doggo,floofer,pupper,puppo) should be melted into one column

df_image and df_json

2. This dataframes should be merged with the df_archive dataframe to ensure each observational unit forms a separate table and there is no duplication of information in various tables.

DATA CLEANING:

This is the final stage of the wrangling process where all the issues raised in the assessing phase are addressed and cleaned up.

I cleaned all the issues raised above using functions from the pandas and numpy libraries.

Before proceeding to draw insights and create visualizations as would be addressed in the act report.