Wrangling Report
By: Rawan Alghamdi - February 27, 2019

Data wrangling project was very challenging, but I have learned a lot while working on it. Data wrangling process contains of three main steps, gathering , assessing and finally cleaning. I did all these three steps in the project then, I did some analysis to provide some insights of the data and it will be presented in details in other sepertaed report “act_report”. The dataset that used in this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

The wrangling process is documented and explained below:

First, the dataset was gathered and loaded to be ready for assessment.

Second, the data was assessed for finding two types of issues, quality and tidiness issues. The following are the issues found in the data while assessing the data visually and programmatically:

1- Quality Issues:

- twitter_archive Table
  - missing values: expanded_urls
  - Erroneous Datatypes: timestamp, tweet_id
  - name column contains values written as "None" instead of NaN
  - Duplicated "expanded_urls"
  - There are some useless columns which will not be needed later (in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp)
- df_image_predictions Table
  - Duplicated "jpg_url"
  - Some p1,p2 and p3 are uppercase, others are lowercase
  - Erroneous Datatype: tweet_id
  - missing data: 2075 length instead of 2356
- df_tweet_info Table
  - a column called "id" has another name "tweet_id" in other two tables

- Erroneous Datatype: id
- missing data: 2354 length instead of 2356

2- Tidiness Issues:

- twitter_archive Table
  - The four columns (doggo, floofer, pupper, and puppo) must be merged into one column (one variable)
  - All three dataframes contain shared information (About Tweets) so, it should be joind to be one table

Finally, the issues detectes in the assessment step were cleaned (fixed) during the cleaning step as the following:

- Fill null expanded_urls by using tweet_id in twitter_archive_clean
- Change timestamp data type to datetime in twitter_archive_clean
- Change tweet_id in all three tables to be object instead of integer
- Rename the column "id" to "tweet_id" in df_tweet_info_clean
- Replace 'None' values with a NaN in "name" column in twitter_archive_clean
- Drop duplicated "expanded_urls" in twitter_archive_clean
- Drop duplicated "jpg_url" in df_image_predictions_clean
- p1,p2 and p3 will be changed to be lowercase in df_image_predictions_clean
- Drop useless columns from twitter_archive_clean after extracting the null cells in retweet columns.
- Create new column "dog stage" (doggo, floofer, pupper, and puppo) in twitter_archive_clean by saving value "None" if there is no dog stage and indicates if there are multiple dog stages by seperating values using comma.
- The issue of having missing records in tables and shared tweet information among all three tables will be solved and fixed by joining all tables in one table called "twitter_archive_master"