Wrangle Report

In this project, we wrangled, analyzed, and visualized the data of WeRateDogs Twitter account. Wrangling data is consists of multiple processes: gathering data from different resources, assessing data, and cleaning data. In the next sections, we will describe the wrangling efforts briefly.

Gathering Data

We gathered data from various sources. The first piece of data was given as a file on hand, so we downloaded the 'twitter_archive_enhanced.csv' file manually, and then we used pd.read_csv() function to read the existing file and store it as a dataframe. The second file was the tweet image predictions file, and it was hosted on Udacity's servers. So, we downloaded it programmatically using the Requests library, store it in a text file, and read it as a CSV file. The last piece of data was supposed to be extracted using Twitter API and store each tweet's entire set of JSON data in tweet_json.txt file. But, I applied for a Twitter developer account to access Twitter API and got rejected. And because of the time limitation, I used the provided code by Udacity and accessed the project data by reading the given JSON file line by line.

Assessing Data

The three dataframes were assessed visually and programmatically using different pandas functions, such as info(),describe(),value_counts,head(), and sample(). Below are the identified issues for each dataframe:

tw_archive_df:

Quality issues:

- name column has None values instead of NaN (The None value won't be counted as a null value)
- tweet_id datatype is incorrect
- timestamp datatype is incorrect
- some rows are retweets or replies
- erroneous numerator values (for example, the number 75 was assigned to the numerator variable while it was a fraction, and the number 960 was mentioned as an invalid rating and it also was extracted as a numerator)
- erroneous denominator values (for example, dates were extracted as a denominator)
- 'a' was extrated as a dog name more than 50 times

Tidiness issues:

- the dog categories distributed between four columns (Doggo, Floor, Pupper, and Puppo)
- 'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp','in_reply_to_status_id', and 'in_reply_to_user_id' columns are useless since the retweets and replies will be dropped

image_predictions_df:

Quality issues:

- tweet_id datatype is incorrect
- some values in p1,p2, and p3 columns start with a capital letter, and some values start with a small letter.

Tidiness issues:

- data values about the same observations spread out over image_predictions_df table and tw_archive_df table (we should combine the two dataframes)

tweet_df :

Quality issues:

- tweet_id datatype is incorrect

Tidiness issues:

- total number of retweets and favorites in a separate dataframe (we should combine the two columns with tw_archive_df)

Cleaning Data

After taking a copy of each dataframe, we cleaned data using different Pandas and NumPy functions, such as replace(), astype(), drop(), merge(), and melt(). The cleaning processes followed standard cleaning practices, each quality or tidiness issue was defined, coded, and tested individually.