Data Wrangling – Wrangle Report

Goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Supporting Files:

- The WeRateDogs Twitter archive: twitter-archive-enhanced-2.csv
- The tweet image predictions (image-predictions-3.tsv)
- JSON datafile of tweet ID, retweet count, and favorite count (tweet_json.txt)

Project Steps:

- Gathering Data
- Assessing Data
- Cleaning Data
- Visualizing Data

I. Gathering Data

Data was gathered from the supporting documents provided by Udacity.

- Twitter archived file was downloaded as CSV file and named as **twitter-archive-enhanced-2.csv**.
- Image prediction files was downloaded in the tsv format and named as **image-prediction-3.tsv**.
- Twitter API file was collected by creating the JSON file by using the Twitter API with Python's Tweepy library. The supporting material provides the result text file named **tweet_json.txt**.

Once I had all the above three files, I created them into 3 different dataframes:

- *df_twitter* - this is converted from the dataset "twitter-archive-enhanced-2.csv".
- *df_image* - this is converted from the dataset "image-prediction-3.tsv"
- *api_data* - This dataset will contain tweet information.

II. Assesing data

For each dateframe described above, I programmatically and visually assess the dataset including the structures, information, columns, duplicated value, null values, etc.

There are quality and tidiness issues presented in the dataframes. The details were described in wrangle_act file. The main issues can be summarized and listed as below:

Quality Issues

- Many null values
- Missing values
- Incorrect data types
- Extreme values in rating_denominator and rating_numerator

Tidiness Issues

- Four columns for dog stages
- Duplicated jpg_url
- Weird dog names

III. Cleaning Data

For cleaning all the 3 dataframes, Here are the steps I followed before after joining the dataframes.

- Dropped the missing values and duplicated values
- Joined the dataframes into one by using the same tweet_id.
- Converted the wrong datatypes into correct datatypes.
- Change the weird dog names to NaN.
- Created dog stage column and convert the dog stage as category datatype

With cleaned dataframe twitter_clean, I was able to continue checking the columns, data types, information, etc. After everything was confirmed as correct, I moved on to the data visualization.

IV. Visualizing Data

The dataframe twitter_clean was analyzed and visualized by plotting to determine the relationships between different variables of interest.