Data Wrangling Report

Project Objectives:

- Perform data wrangling
- Analyze, visualize and store the wrangled data
- Report data wrangling effort and insights and visualizations

Step 1: Gathering Data

The data was gathered from three sources:

- The WeRateDogs Twitter Archive: is a file that we download manually and upload to the workspace. Once it is downloaded, we upload it and read it to dataframe
- The Tweet Image Predictions (image_prediction.csv): This is a file produced by running every image in the WeRateDogs Twitter Archive through a neural network. We downloaded this file programmatically using Requests library from a provided URL.
- Additional data from the Twitter API: we gathered each tweet's retweet count and favorite(like) count using tweet IDs in the WeRateDogs Archive by querying Twitter API for each tweet's JSON data and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

Step 2: Assessing Data

After gathering the data, they were assessed visually and programmatically for quality and tidiness issues.

Tidiness:

- Dog stage data is separated into 4 columns.
- All dataset is related but it is separated into 3 dataset

Quality:

Enhanced Twitter Archive Data:

- Some columns have missing values that are more than ten percent of the dataset
- Some dog names are invalid(None, a, an instead of name)
- Invalid Timestamp data type (string instead of datetime)
- Some expanded URLs have two URLs
- Some denominators have values that are not equal to 10
- Row 313 has 0 denominator

- Some columns contain None instead of Nan values
- Text column contains description and URL

Tweet Image Predictions Data

- Missing rows (2075 instead of 2345)
- Some P names start with uppercase while others start with lowercase
- Underscores are used in multi-words names in columns p1, p2 and p3 instead of spaces

Step 3: Cleaning Data

Once we were done assessing data, we cleaned the dataset resulting in high quality and tidy dataframe.

Missing values

- Dropped columns that have missing values up to 10 percent of the dataset
- Drop rows with missing values

Tidiness

- Merge the four columns of dog stage into one column named dog_stage
- Merge all dataset into one dataset called all_df based on tweet_id

Quality Issues

- Converted invalid names to Nan and extract the correct names from the text column
- Changed timestamp datatype to datetime
- Extracted the first URL from all rows with multiple URLs
- Row 313 was most likely dropped while cleaning other issues
- Changed None to Nan values
- Extracted URLs from the text column to a new column and removed the URLs from the text column
- Changed P names to Uppercase
- Changed underscore in P names to spaces