Data Wrangling Report for the WeRateDogs Data

Gathering Data

1. twitter_archive: The WeRateDogs Twitter file, which is given by the Udacity Course and I use pd.read_csv() to import them into dataframe.

2. image_predictions: The tweet image prediction, i.e., what breed of dog (or different objects, animal, and so on.) is available in each tweet as per a neural system. This record ('image_predictions.tsv') is facilitated on Udacity's servers and downloaded programmatically utilizing the python library and the provided URL.

3. tweet_data: Using the tweet IDs in the WeRateDogs Twitter file, query the Twitter API for each tweet's JSON data utilizing Python's Tweepy library and store each tweet's whole set of JSON data in a file called 'tweet_json.txt'. Each tweet's JSON data is kept in touch with its line.

I started by downloading ‘twitter-archive-enhanced.csv’ manually. Afterwards, I downloaded ‘image-predictions.tsv’ programmatically from Udacity's server using the requests library. ‘tweet_json.csv’ was created by accessing and downloading Twitter’s JSON data using the tweepy library. To do that, I obtained a list of tweet ID from the ‘twitter-archive-enhanced.csv’ file, looped through each ID and query Twitter’s API with the ID to get each tweet’s JSON data. Subsequently, I recorded the data in a text file named ‘tweet_json.txt’, with each tweet’s data written in a new line. After the query was completed and all the data was written in the text file, I read the text file line by line, obtained each tweet’s information (tweet ID, favorite count, and retweet count) using the JSON library, and appended the information into an empty list.

Assessing and Cleaning Data

The CSV files for each dataset were downloaded and opened in Excel. The dataset with around 2500 rows was manageable in Excel and using the filters function gave a good feel of the data inside each of the three datasets. From Excel, it was quickly identified the many incorrect names in the dataset for the dogs and the strange rating scores being used for both denominators when you would only expect 10 and for the numerator.

- After gathering all the necessary data sets, I read them into pandas DataFrames and assessed them visually and programmatically.

- In the first assessment, I focused on the general structure of the data sets and noticed five tidiness issues and four quality issues, mainly about incorrect data types, missing data, and duplicates.
- After cleaning the three data sets accordingly and combining them into one data set, I assessed it for a second time, this time paying more attention to the values in the dataset. I also

Storing Cleaned Data

After the conversion, the data set is clean enough for analysis. Therefore, I saved the data set into a CSV file named ‘twitter_archive_clean.csv’ which was the combined using all the 3 datasets and used it for my analysis.