Data Wrangling Report for the WeRateDogs Data

Gathering Data

I started by downloading 'twitter-archive-enhanced.csv' manually. Afterwards, I downloaded 'image-predictions.tsv' programmatically from Udacity's server using the requests library.

'fav_ret_count.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. Finally, I created a pandas DataFrame using the list and saved the DataFrame into a csv file named 'fav_ret_count.csv'.

Assessing and Cleaning Data

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. As a result, I found five additional quality issues, which likely stemmed from incorrect extraction of values from the 'text' column. These issues were also addressed and the data set was cleaned for the second time.

Subsequently, I assessed the data for the third time and found that the newly created 'dog_stage' column should be converted into categorical format. So, I cleaned the data set for the third time by converting the column's data type.

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_master.csv' and used it for my analysis.