Data Wrangling Report

The Dataset(s)

The dataset I used is the archive data from WeRateDogs twitter account, whose daily tweet is about showing images and giving a rating from the dogs' pictures. The original dataset contains 2356 tweets data. The data is in CSV format. The Python can use the Pandas library to read the CSV format into a data frame. Each data in this data frame contains a unique tweet ID which can be the primary key to join other dataframe.

I also got an image predication dataset based on the image from WeRateDogs Twitter archive. This dataset consists of the top three image predictions with tweet ID. I would use this prediction result to find which dogs people like the most. I used a request library to download the data from the URL link. After saving the TSV file in the local, I used the read_csv function to read the TSV file with separate equal to '\t.'

Then, I used the tweet IDs from the Titter archive dataframe and Twitter API to query the tweet status from the original twitter account. I used a for loop to query the tweets and set up a timer. It took 1887 seconds to download the data. I dumped the tweets JSON file in a text file. Then, I used the JSON library to read the JSON data and stored the tweet ID, retweets number, and favorites number in a dataframe.

Assessing the Data

Visual assessment

Quality: The source data from the archive tweets dataframe is too long to read. It contains an HTML format which should be removed.

Quality: The data from doggo, floofer, pupper, and puppo used the string "None" as the null value. In the Pandas dataframe, the null value should be Nan.

Tidiness: The tweet dataframe contains retweet data, which is a kind of duplicates can cause problems.

Programmatic assessment

Tidiness: These three data frames share different numbers of data.

Quality: The rating denominator has values other than 10.

Quality: The rating numerator has some huge numbers.

Tidiness: This three dataframe should join into a large dataframe using tweet id as their primary key.

Quality: One of the dataset's tweet ID is string data types, while others are int64.

Cleaning the Data

I used the Pandas function to clean the data. For example, I drop the retweets data by filtering the dataframe. I also used the merge function to combine all three datasets.