Data Wrangling Report

This project is about wrangling a dataset from WeRateDogs' twitter account. Wrangling involves collecting and cleaning data from various sources by checking for accuracy and tidiness problems. The data analysis can lead to misleading results without cleaning the data. So it is very important to wrangle data correction using correct techniques. I have written the data provided by the Twitter API for the WeRateDogs account for this project.This account rate for dogs on a very different scale, with numerator being higher than denominator most commonly because dogs deserve it In this process, there are a total of three steps invoked, including data gathering, accessing data and cleaning data.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:
• The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archiveenhanced.csv')
• The tweet image predictions ('image-predictions.tsv'). This file was be downloaded programmatically using the Requests library from a provided URL.
• Each tweet's entire set of JSON data (with at minimum tweet ID, retweet count, and favorite count) in a file called 'tweet_json.txt' were stored using Twitter API and Python's Tweepy library. Each tweet's JSON data was written to its own line.

Step 2 and 3: Assessing and Cleaning Data

Quality

archive table

- tweet_id has to be a string dataype.
- Timestamp is not a datetime variable.
- The source format is incorrect.
- The text of a tweet includes urls.
- We only want the original Tweets with images, so the columns related to Retweet should be dropped.

predictions table

- tweet_id has to be a string dataype.
- Inconsistent writing in p1, p2 and p3.
- Drop duplicate jpg_url.
- Column names are not clearly identified.

tweet_ids table

- tweet_id has to be a string dataype.

Tidiness

- Four variables (doggo, floofer, pupper, puppo) should be merged in one column.
- Three dataframes join the inner.

Then we copied the data and cleanup (Define, code ,test)