Wrangle Report

Introduction

The purpose of this data wrangling is to put into practice of what I learned in data wrangling classes. The main process can be summarized as gather, assess, and clean the data. The dataset that I wrangled is the tweet archive file from Twitter user WeRateDogs, the JSON file I pulled from Twitter API, and a machine learning image prediction file based on the picture of the tweet archive file. All the data is related to the data from Twitter user WeRateDogs.

Data Gathering

Tweet archive file: this file was provided by Udacity and I downloaded it and import the file to Jupyter Notebook.

Tweet JSON file: I use Python Tweepy to query Twitter's API for additional. And then I write the file line by line into a pandas dataframe with tweet ID, retweet count, and favorite count.

Image prediction file: this is a dataframe full of image predictions alongside each tweet ID, URL, and the image number that corresponded to the most confident prediction this file. This file was ran through a neural network and provided by Udacity.

Data Assessment

I used both Jupyter Notebook and Excel to assess the data visually and programmatically. The data doesn't come clean. I noticed some quality and tidiness problems. Some common issues are erroneous datatypes, inaccurate rating numerators and denominators, inconsistent data values, and missing values.

Data Cleaning

The clean process mainly consists of define, code and test for each data cleaning.

First, a copy of each dataframe was created to keep the original data. There were a couple of cleaning steps which took me some time. For example, I created a nestedif to capture the best prediction of dog breed. Another interesting cleaning is after noticing the outliers of the numerators, I read through the tweet and identified the correct numerator. The reading was fun. The best part is the data merging part, as that's the last step and I can finally merge all the cleaned dataframe into one!

After all, I cleaned the data to ensure the data completeness, validity, accuracy, and consistency. Data wrangling is one of the most important steps and a good data wrangling definitely benefits the further data analysis and visualization.