Wrangle report

Introduction:

This report is part of the Wrangle and Analyze Data project from Udacity. The purpose of this report is to document the wrangling efforts of the project. The dataset used in this project is from Twitter archive of user@dog_rates. It is also known as WeRateDogs. This report documents three step of data wrangling: gathering, assessing, and cleaning.

Gathering Data

In this project, I gathered data from several sources, and they were in different format.

1. The WeRateDogsTwitter archive. The twitter_archive_enhanced.csv file was provided to student. The file could be downloaded from the project downloadable resources section. It is also included in jupyter notebook with Udacity.
2. The Tweet Image Predictions was provided by Udacity as well. The file could be downloaded from the project downloadable resources section. It could also be downloaded directly on jupyter notebook by using the link provided by Udacity.
3. Twitter API provided by Udacity didn't work because Twitter changed their policy, I had to download the json file from Udacity to get favorite counts.

Assessing Data and Cleaning Data

I assessed the data visually by excel and programmatically by jupyter notebook to identify any issues with my data. The most issues I found were with the archive file and the image predict file. I write assessing and cleaning together to make it clear and simple to understand. I make copies of my data before I clean. That way, I have backups if I accidentally change or remove something.

Archive file: I validated data type of my data frame. Timestamp column had an incorrect data type, so I had to change data type from string to datetime. Rating numerator column had some super high values. I had to filter and take out the abnormal values out of the table. Rating denominator had values of 20, 15, 11, 10, 2. I only kept value 10 and took out the rest of them. The source column had the url in html format. I must remove the markup just to have a normal url. In this project, I only analyze the original tweet, I don't need retweet values, so I remove retweet columns. Column p1, p2, and p3 contain a lot of values and they are not dog breed. I have to remove those values out of the table. Column p1, p2, p3 also have inconsistent capitalization. I convert all values into lower case to have uniformed values.

Image predict file: the file has some duplicate url. I remove the duplicate.

There are two tidiness issues that I have to clean is the retweet columns and dog stage columns. I only use the original tweet so I remove any data that has retweet values from those columns first. After I remove those values, I remove retweet columns because I don't need them anymore. Dog stage has 4 columns but they only contain one value for each row. I combine the data of 4 columns into 1 new column called dog_stage.

The data set has more issues to clean. However, cleaning data is very time consuming. So, I only focus on the areas that important for my project. I store the clean data in a csv file for my analyze step.