Wrangle Report

Introduction:

The purpose of this project was to learn about the skill of wrangling data in all its stages. It was to work on a dataset for the WeRateDogs Twitter account. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Project steps:

- Gathering Data
- Assessing Data
- Cleaning Data
- Storing Data

Gathering Data:

This step is the first in the wrangling process. I collected three data sources in three different ways.

- The direct method: I downloaded the CSV file, which is provided by the Udacity platform on the website. Then read the file inside the DataFrame called tw_archive.
- The second method: Image Predictions File, which was downloaded programmatic by the Request library, and then read inside the DataFrame called img_prediction
- The third Method: First of all, I tried to get a Twitter developer account, but my request was rejected. After that, I used the code available on the Udacity page to query the data from Twitter.
    - Query Twitter API for each tweet in the Twitter archive and save JSON in a text file.
    - Save each tweet's returned JSON as a new line in a .txt file -- After that I made the file readable.
    - From the file, extracted the tweet_id, followers_count, favorite count and retweet count.

Assessing Data:

In this section, assessed the data by two ways:

- Visual: Display the data of the three tables by using head() method
- Programmatic: by using several methods such as (info, value_count, sample)

Some of the results of the assessment:

Type of issue	Asses type	Description
Quality	Visual	Missing values in many of columns
Quality	Visual	Source culomn have a string between html tages
Quality	Visual	Unacceptable rating of some types of dogs.
Quality	Visual	Unnecessary columns.
Quality	Programmatic	incorrect datatype.
Quality	Programmatic	Abnormal names, some of names starting with uppercase and names starting with lower.
Quality	Visual	Columns: p1, p2, and p3. have names starting with uppercase and names with lowercase letters.
Tidiness	Visual	Columns: doggo, floofer, pupper and puppo. Have the same categorical classification, shoud be Merging in one column
Tidiness	Programmatic	Text column had multiple variables, such as text and url.
Tidiness	Programmatic	All three tables can be merged for manageability.

Cleaning Data:

This section was divided into three steps:

- Define - Code - Test

These steps have been applied to most of the problems identified in the assess section

- At first, I made a copy of the original data. Then I dropped all the unnecessary columns, so it would be easier to work with the data.
- Then Fixed all wrong data types for columns in the three tables
- One of the challenges in the cleaning process is dealing with the rating issue, some of the rating values were very high, and some of the rating denominator not equal 10. I modified the values by using a mathematical method to convert the rating denominator to 10, which is the basis for the rating. Also, decimal ratings issue has been fixed.
- Some of abnormal dog names such as (a, an, all, such, my, ...) that seem to have been extracted from the context of the text. converted these names and standardize the letters of the names as uppercase.

Storing Data:

After cleaning the data and merging the three tables together, the data was stored in a CSV file called twitter_archive_master.csv

Conclusion:

It is important that the analysis be done on clean data, so data wrangling is a core skill that whoever handles data should be familiar with.