Wrangling Report

Introduction

The objective of this project was to practice the 3-step process that is required when wrangling and analyzing data.

- Gather the data that can come from difference sources.
- Assess the data to see for quality and tidiness.
- Clean data from the observations made in the assess stage.

For this project, the twitter archive WeRateDogs™ was used to practice the skills learned in the data wrangling module.

Gather

For this project I had to gather data from 3 sources.

1) Twitter Archive File- The first was a CSV file, ready to access and I simply had to download it manually.

2) Tweet Image Predictions- This data showed the breed and the image of each breed. It was hosted on Udacity's servers was needed to be downloaded programmatically through Request library and URL information.

3) Twitter API and JSON- For this a twitter account was needed to access the API information but since I chose not to create one, I downloaded the code and used the Json text file, that already had the information I needed. My only task as the run it programmatically
by reading the tweet_json.txt file line by line into a pandas DataFrame with the following columns tweet ID, retweet count, and favorite count."

Assess

The assess portion required careful observation of all 3 data sets. This was done in two ways: 1) Visually and 2) programmatically. On this section I was able to identify 9 issues with quality and 3 on tidiness.

Visual- Visually was easily done by simply seeing the date through the head method.

Programmatic- This required other methods that observed the data structure, information, and numeric values. For this I used values such as value_counts, info, and describe.

Clean

I defined the issue that was I observed on the assess section, then I coded, and finally tested if the changes were made. This was done for all 9 issues pertaining to quality and 3 for tidiness. It is worth noting that for one tidy issue was to combine all 3 datasets into one, with the correct changes made before merging.

Storing, Analyzing, and Visualizing Section

This was the last section of the project where I stored the master dataframe into a CSV file, made 3 insights and one visualization.

Conclusion

This project was by far the most difficult and required me to constantly go over the lessons. The most satisfying part was the observing the final master dataset and understanding the need for each data source and the importance of both good quality and tidy data!