My Data Wrangling Experience

Wrangling data is no small task. The data formats and host requirements vary greatly across the web. My experience was wrought with obstacles. First, Twitter denied me a developer account to wrangle my own data, so I had to use the Mentor section of Udacity to access the API data. Then, the tweet data for the WeRateDogs handle archive was in a text format with each tweet separated by {"":"":, etc.}. My workspace went into queue mode and wouldn't load at times. Ultimately, this was a fun experience and I was able to access all the data required for this project.

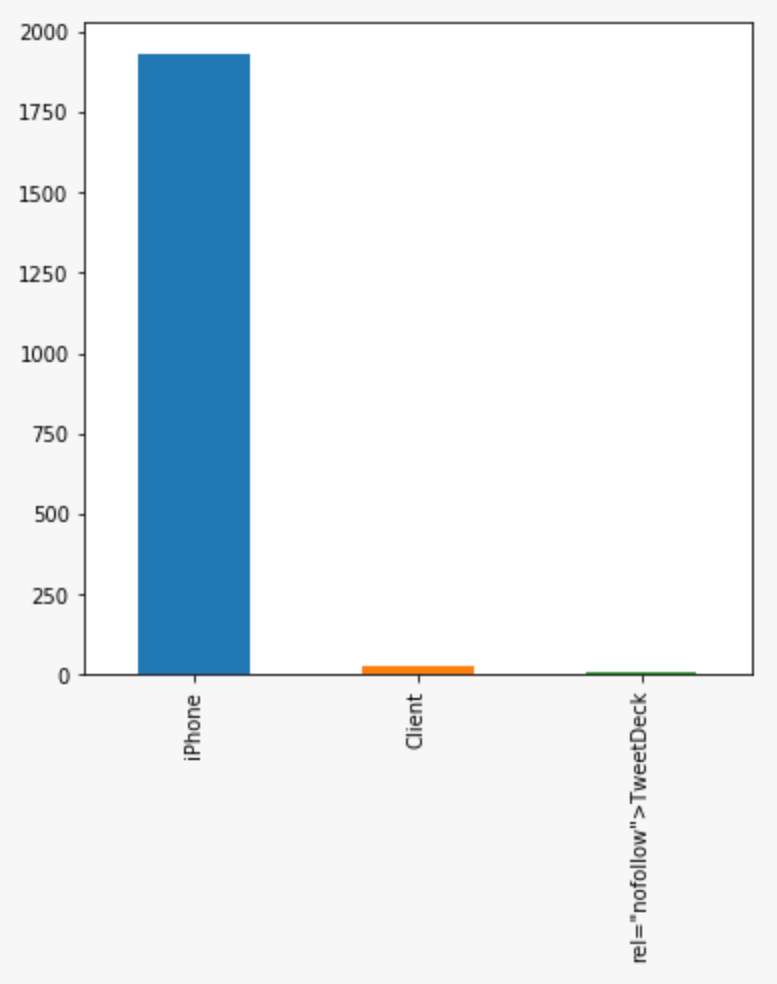
The data accessed through the Twitter API was the last data I added to my jupyter notebook for the project. The easier to gather data was added first, and I explored the data using .head() and .info(). The first file loaded to the project workspace was 'twitter-archive-enhanced.csv'. I could immediately see there were NaN values throughout some columns. I used .info() to see the number of rows for each column, noting the retweet columns only had 181 rows versus the total of 2356 in the 'tweet_id' column. The instructions included removing these rows, so I added the notes to do this to the 'Assess' section.

The 'weratedogs-image-predictions.tsv' was interesting because it was using an algorithm to predict the dog type based on the photos. There were three prediction columns. I noticed there were instances where the dog type was a thing which isn't a dog, such as a bagel. The dog type predictor even rated itself for accuracy, which was impressive. Each dog type prediction column had a self-rating column. Should I remove the rows where the image was predicted to not be a dog? Should I merge the prediction column into one column where the self-rated accuracy was above a certain percentage? There were many ways to think about this data.

The Twitter data was the final bit of data I added to the project. I needed the most help here, as I said, the wrangling was wrought with obstacles! This data was restricted to the 'favorite_count' and 'retweet_count' columns. I merged all of this data into one file, 'twitter_archive_master.csv', which I am able to open in Google sheets. Hopefully I've given readers a good idea of my data wrangling experience.

Iphone was the most popular device format. This was interesting because I wondered if the size of the screen people were viewing had any impact on the ratings. Upon investigation there wasn't much variance in device types, so the image size didn't appear to be a relevant factor in ratings.

By Angie Koepcke, koepcke.angie@gmail.com or angiek99908@protonmail.com
https://github.com/Angie99908/WeRtDgs_Tweets



Hopefully I've given readers a good idea of my data wrangling experience. Data wrangling can be fun and interesting!

By Angie Koepcke, koepcke.angie@gmail.com or angiek99908@protonmail.com
https://github.com/Angie99908/WeRtDgs_Tweets