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. 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. Fortunately the Mentor section was extremely useful in this project. 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 project. I added the easier to gather first and explored the data using .head() and .info(). The first file loaded to the project workspace 'twitter-archive-enhanced.csv'. I could immediately see there was data missing and 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 that to the 'Assess' section. This file had the most amount of obvious cleaning and tidiness issues, so I was able to fill the 'Assess' section easily with just this first file.

The 'weratedogs-image-predictions.tsv' was interesting because it was using an algorithm to predict the dog type based on the photos. 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. I think I could have spent more time just playing with this data image than I feasibly have to spend. I get caught up in the fun of analyzing 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. I could go on, but I have data to analyze!

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