Wrangle Report

The data Wrangling project was very challenging and it took me a lot of time and searching in many references like stackoverflow.com, pandas.pydata.org, and of course google.com.

I have learned a lot about data wrangling and became familiar with its process (gathering data, assessing data, and cleaning data), gathering data from different sources**,** dealing with the different file types and different data issues, and learned about Twitter API.

Gathering Data

I gathered data from three different sources, first I downloaded a CSV file that contains the WeRateDogs Twitter archive manually, and then I downloaded a TSV file that contains the tweet image predictions, i.e., what breed of dog programmatically using the requests library, finally by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a txt file then I read this file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

Assessing Data

After gathering each of the above pieces of data, I assessed those visually and programmatically using pandas function for quality and tidiness issues, then I documented these issues.

Cleaning Data

I cleaned each of the issues I documented while assessing. To make high-quality and tidy master pandas DataFrames.

Storing data

I have extracted the columns I needed from the clean DataFrames then merged them into one DataFrame, and then saved it to a CSV file named twitter_archive_master.csv.

Analyzing data

I have done some analyzes like calculating the most common breed in the tweets and the most rated one, the most tweeting hours of the highest interaction, and doggolingo usage percentage in tweets.

Visualizing data

I have done some visualizations like tweeting rate timeline, top 20 breeds in the tweets, and common Doggolingo used in tweets.