Wrangle_Report

In this project I have worked on Jupiter notebooks and used Python and its libraries such as: pandas, numpy, matplotlib, seaborn, os, requests and io to gather data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it.

The dataset I wrangled (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

For **Gathering** I have downloaded the WeRateDogs Twitter archive given to me as a csv file, used the request library to request the Image_Predections dataset from the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and finally I have read the tweet_json.txt file line by line into a pandas DataFrame.

After gathering my data I started **Assessing** every data set programmatically and visually and documenting my assessment as Quality issues and Tidiness Issues

Then I started **Cleaning** my data starting by copying datasets into cleaned to be data frames to keep the original data sources from change, then started the cleaning with solving the null values issues, moving to the tidiness issues and finally cleaning the quality issues, all following the format of defining the cleaning decision, coding and then testing.

Finally after I cleaned my data I have **Stored** them in a csv file and **Analyzed** them and extracted three insights from them, supporting the insights by visualizing them and finally creating a document delivering those insights.