Data Wrangling Report

About the Dataset(s)

The dataset I had to wrangle and analyze was the tweet archive for Twitter user @dog_rates (https://twitter.com/dog_rates), also known as WeRateDogs. This dataset consisted of 2356 basic tweet data from November 2015 to August 2017.

According to Wikipedia, WeRateDogs is a Twitter account started in 2015 by college student Matt Nelson. Today the account has 7.94million followers and over 142,000likes.Matt has been able to achieve this mass appeal and followership by rating people's dogs humorously.

1. Gathering Data

To carry out this project, I had to gather data from 3 different data sources

a. Gather Twitter archive CSV file

Using the link provided by Udacity, I downloaded the WeRateDogs Twitter archive manually from the link provided by Udacity and imported this file into a dataframe (ted_df).

b. Gather tweet image predictions

Having done the above, I programmatically downloaded the image predictions file using the Python Requests library. The file contained image predictions on the breed of the dogs stored on a neural network for some of the tweets already downloaded in the archive file. The file was in tsv format and I imported into the dataframe (img_predictions_df)

c. Gather data from Twitter API

For the final dataset, I queried the twitter API and used a Python library called Tweepy to obtain further data on the tweets in the archive file using the tweet id. The Tweepy library returned the data in json format, from which I was able to iterate through and append data to a panda dataframe.

2. Assessing

I assessed the imported data visually by opening the CSV files in excel to understand what needed to be done in terms of cleaning. This I achieved using row filters. Following this, I also programmatically assessed the data in jupyter notebook with pandas using the following functions, df.info (), df.head (). During assessment, I watched out for Quality (completeness, validity, accuracy, and consistency issues) & tidiness. For issues I noticed, I put them under the 'Assessing Data" section of the "wrangle_act.ipynb" jupyter notebook.

Quality refers to issues related to the content of the data, sometimes called dirty data while Tidiness refers to issues related to the structure of the data, sometimes called messy data.

3. **Cleaning**
   Leveraging programmatic tools in python like pandas, melt etc. I tried to resolve the issues I was able to detect during the assessing data stage. Details of this can be found in the “Cleaning the Data” section of the “wrangle_act.ipynb” jupyter notebook.

4. **Storing Data**
   When I was done wrangling the data, I stored in “twitter_archive_master.csv’ for future use. The Data wrangling process enables the provision of clean data for future analysis and visualization.