Data Wrangling

Dogs Rating

Introduction

The dataset that will be wrangling (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 8 million followers and has received international media coverage.

Gathering Data

We gathered data from different sources

First:

Twitter Archive File that got from resources from Udacity

This file have 5000+ tweets details related to the WeRateDogs account for users rating dogs on the account

Second:

The tweet image predictions

This file contains the top three predictions of dog breed for each dog image from the WeRateDogs Enhanced Twitter Archive. Data is downloaded programmatically using the Requests library from the URL address into a tsv file.

Third

Twiter API File

Twitter API file contains tweet id, favorite count and retweet count. Data was provided by Udacity, downloaded manually then was loaded from the tweet-json.txt file into a pandas data frame. Data frame size is 2354 rows and 2 columns. The tweeter ID column is used as an index.

Data Assessing

- After Data Gathering we started to assessing data for any Quality and Tidiness issue.
- Removed incorrect data.
- Delete any retweets and only keeps the unique tweets by deleting the duplicates on the data sets.
- Delete some Columns that will not be used in the Data wrangling and analysis.
- Merge Data sets in one File that is `twitter_archive_master.csv.`

Data Cleaning

- Fixed The Time frames by changing the type to datetime
- Corrected the wrong URLs and changed by the correct ones
- Removed the denominators that was more than 10