Wrangle and Analyze Data
Wrangle Report

Introduction:

The dataset that you 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.

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo)

Project Steps Overview:

The tasks in this project are as follows:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 1: Gathering data:

Data for the project was gathered from three sources:

1-Twitter archive file:

The twitter-archive-enhanced.csv was provided by Udacity, downloaded manually hen was loaded from the CSV file into a pandas data frame.

2-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. The content of image-predictions.tsv file is then loaded into the pandas' data frame.

3-Twitter 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.

Step 2: Assessing data

In this step, I have through the datasets to discover any quality or tidiness issues, either visually or programmatically. And I have discovered many quality or tidiness issues:

Quality issues

Twitter Archive Dataset:

1- Missing value None should be NaN.

2- Remove unnecessary columns ("in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", retweeted_status_user_id")

3- In source column remove html a tag.

4- Timestamp should be converted to datetime.

5- Remove retweets (text starting with RT @)

6- In rating_numerator thare are 901 rows less than or equal 10 they should be greater than 10.

7- In rating_denominator thare 23 rows not equal 10 they should always equal 10.

8- In name column there are 55 rows with a value

Image Prediction Dataset:

9- There are too many duplicate tweet 66 rows.

10- Column names are confusing and do not give much information about the content.

Twitter API Dataset:

11- There is one duplicate row.

Tidiness issues

1-Dog Classification (doggo, floofer, pupper or puppo) should be one column.

2-Create master dataframe that merge all the dataframes together

Step 3: Cleaning data

After identified the quality and tidiness issues I will fix some of them.

Step 4: Storing data

As a last step of the cleaning process, we merge all datasets into one and export to twitter_archive_master.csv file. Then store into locally.