Wrangle Report

This project consisted of 3 stages that included gathering ,assessing & cleaning data

Gathering Data

This was the first step in this project which involved gathering data of three different sources as listed below. Each of the them had a different method to be collected with, these methods are :

- Importing data via csv
- Using requests to download data off internet
- Scrape data from an API

Three data sources

Enhanced Twitter Archive

The WeRateDogs Twitter archive provided by Udacity. I downloaded this file manually then imported it into dataframe using "read_csv" method, this file played a big role in this project.

Image Predictions File

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: image_predictions.tsv

Data via the Twitter API

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing data

After gathering data this data I had to asses it to decide the step should be taken to clean it & make it ready to use it in analyzation

The data issues are Quality & Tidiness issues :

Quality issues :

1-only original tweets are required

2-expanded_urls column is useless(need to be dropped)

3- Many dog names are incorrect-starts with lowerscase letters-

4-The rating_numerator column should be of type float .

5-timestamp's datatype should be converted to "datatime".

6-Remove "_" in image predictions.(p1, p2, p3 column names)

7-convert tweet id's data type into String

8-The rating_numerator column decimal values should be correctly extracted.

Tidiness issues :

1-merging 3 data frames

2-dog's breed needs to be in one column

3-tweet text & dog's photo url are included in one column

4-dog's life Stage needs to be in one column

Cleaning Data

Cleaning these data issues took some effort but after finishing I think this dataset is finally ready to be used in real world analyzing as I did in my 3 insights