Wrangle Report - WeRateDogs Dataset

Wrangling is about Gathering the data, Assessing it and Cleaning the unwanted data also fixing some errors in data. In this WeRateDogs Dataset, the same process has been followed.

Gathering Data for this Project

This project involved gathering of data from three different sources as listed below. For each of the data source a different method of data gathering was used namely:

- Importing data via csv
- Using requests to download data off internet
- Using Twitter API data for analysis

Three data sources

Enhanced Twitter Archive

The WeRateDogs Twitter archive provided by Udacity. This contains basic tweet data for all 5000+ of their tweets. I manually downloaded this file by clicking twitter_archive_enhanced.csv

Image Predictions File

The tweet image predictions, i.e., what breed of dog (other objects, 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

Twitter API Data

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

In this step, Assessing data was made both visually using spreadsheets and programatically. This step is all about finding errors in dataset and noting down to correct it in the next steps. The issues found in the datasets include Tidiness issues and Quality issues.

1. Removing Null values.
2. Fixing Datatypes of various columns.
3. Remove unwanted columns (retweeted_status_timestamp, retweeted_status_user_id, retweeted_status_id) and clean up duplicate rows and NaNs.
4. Fix numerator and denominators.
5. Drop columns with one low values or similar kind of values.
6. Combine and clean different dog stages (eg: pupper,doggo) columns into one.

7. Making Column headers more descriptive.
8. Combine three different dataframes into one master data set.

Cleaning Data for this Project

I used my knowledge of python and searching over the internet i.e. google, stackoverflow, w3schools, took Udacity’s mentor help etc. for references and possible guidance to resolve the above mentioned issues to the best of my knowledge.

Overall, I learned a lot about how to use python effectively and efficiently to clean data and store it.

Finally, once the problems were fixed and data was ready I analyzed it using visualizations as document in act_report.