WRANGLE REPORT

By Njeri Macharia

Introducion

The objective of this project can be described in 3 steps as followed:

1. The main goal of this project is to wrangle (gather, assess and clean) WeRateDogs Twitter Data set.
2. The next step is to store, analyze and visualize the wrangled dataset.
3. The final step is to come up with two reports:
    a. Wrangle report- A brief report describing the wrangling efforts made on the datasets.
    b. Act report - A report communicating the insights and visualizations produced from the wrangled data.

Data Wrangling

Gathering Data

The data for this project was required to be gathered from 3 different sources as follows:

1. The WeRateDogs Twitter archive - ('twitter-archive-enhanced.csv'). This is a file on hand, provided by the Udacity Course. It required a direct and simple download to access. The contents of this dataset were the tweets themselves alongside their ID’s, the ratings of the dogs, and other information like the names of the dogs and their stages/ages.
2. The tweet image predictions - ('image_predictions.tsv'). This file had to be accessed by downloading it programmatically via the Request Library using a provided URL. These datasets contained images of the dogs, associated with the tweets and three predictions of the dog types, together with the confidence of those predictions.
3. The tweet data - ('tweet_json.txt'). This file was to be accessed by querying the Twitter API for each tweets’ JSON data using Python’s Tweepy Library. In this dataset, the data relevant to the tweets such as favourites count and retweets count is found.

Assessing

The data has to be assessed both visually and programmatically. When assessing data, there are two aspects to look out for:

1. The quality of the data. This means, looking out for issues such as completeness, validity, accuracy and consistency.
2. The tidiness of the data. This consists of structural issues.

Quality

a. twitter-archive-enhanced

1. Timestamp column is string instead of datetime.
2. There is inconsistency in the dog name column, some names start with uppercase, some with lowercase.
3. Some of the data contains retweets and replies information, while we only want original data information.
4. Inconsistency in denominators. Some ratings are not out of 10 as they should be.
5. Some dog names are invalid such as: 'a', 'an', 'the'.
6. 'None' instead of NaN for missing values.

b. image_predictions

1. There is inconsistency in the names in the predictions columns, some names start with uppercase, some with lowercase.

c. general

1. The tweet_id for all tables is int data type but should be string data type instead.

Tidiness

a. twitter-archive-enhanced

1. The stages of dog ages are in four separate columns instead of one row.

b. general

1. There are 3 different tables, making the data bulky, repetitive and disorganized.

Cleaning

The first step involved creating copies of all three datasets in order not to lose the original datasets.

The cleaning of the data involved taking 3 steps to fix each issue found in the assess stage. These steps were: define, code and test. This meant that it was necessary to properly state the issue, write code to fix it and then write code to ensure the issue was fixed.

Conclusion

Once all the data had been cleaned, merging it into one file was simple and it was now stored into a master CSV file (‘twitter_archive_master’). I was then able to do analysis and visualizations on it, meaning the data wrangling was successful.