Data Wrangling report
By Aravind Gowthaman

Introduction

The WeRateDogs twitter account rates dogs based on the cuteness with a humorous comment and unique rating system having the numerator greater than denominator if the dog's cuteness is overloaded.
This twitter account data is gathered, assessed,cleaned and explored in this project.

Objective

The objective of the project is to gather data from different sources of a Twitter account ''WeRateDogs", wrangle the gathered data using the techniques learned during the program and make the data suitable for exploratory analysis and modelling.

Data Gathering

The data was gathered from the following sources:
1. Twitter_archive_enhanced.csv file which was downloaded programatically by udacity.
2. Gather additional information of the twitter account Twitter API named tweepy.
3. From url "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv" which was downloaded programatically and imported as a tsv file into the jupyter notebook, which contains prediction data about dogs breed.

Data Wrangling report
By Aravind Gowthaman

Data Assessing

The three datasets gathered were mainly assessed visually and programatically, the dimensions (number of rows and columns), data types, presence of null/none values, duplicated values and other quality issues were identified from programatic assessment and Tidyness issues such as merging the doggo,puppo,pupper and floofer columns into one column and merging Twitter API data with the archive data were identified when assessed visually.

Data Cleaning

This part of the process is mainly addressing the issues defined in the data assessing part. Each problem was divided into three sub-stages:Define,Code and Test.

And before starting the process, it is important to work on the copy of the datasets so that original datasets doesn’t get affected.

In the twitter archive data, the most interesting quality issue I found was that the ratings following the unique rating system (which the WeRateDogs are popular for),i.e., the ratings_denominator was not equal to 10. So I sorted this issue by filtering out the rows having value more than 10.
Also the last 4 columns mentioning about the dog types were scattered which could have been in 1 column, so I have solved the same. These column also had mutliple

clasess and None values. The Multiple classes is due to rating of two dogs from one picture and one owner which I chose to leave like that and the None values were removed.

For the API data the tweet_id format was in string so I chose to change it to integer to maintain uniformity among other datasets.

The image prediction data, certain posts were not related to dogs so had to filter out that and also the dog names were in upper and lower case, I chose to change those to lower case to maintain uniformity.

Storing the data

Finally the cleaned data was stored to a csv file. I merged the Archive and API data into one file 'Twitter_data.csv' and keeping the prediction data as a separate file 'prediction_data.csv'.

Conclusion

The data wrangling process is the most significant for any Data analyst to only understand the data structure but also the source of it as well. Gathering data using API has been the highlight of this project and I have made extensive use of the knowledge given from Udacity to clean and explore the data using powerful Python libaries.