WeRateDog Twitter Analysis Project
Data Wrangling Report

The Twitter Analysis Project task was to perform analysis from WeRateDog twitter profile from 2015 till August 1, 2017. This project was done using Python.

Initially, the dataset were gathered before the cleaning process begun. One is the Twitter archived Data which was pre-gathered for this project. The second dataset was to be downloaded programmatically using python then the content were stored in a text file. The last dataset which is needed to provide additional information on the tweets such as its number of retweets and number of favorites (Likes) was gotten from Twitter API using Tweepy.

Data Wrangling Process

The datasets were assessed visually with the aid of Microsoft Excel and Programmatically using Jupyter Notebook. Quality Issues and Tidiness Issues were spotted with the data

Quality Issues

For the Twitter archived data, the task was to carry out analysis on “Original” tweets. With the information, I drooped the records for retweeted data. Some of the fields (Columns) had inaccurately computed datatypes which was changed to their correct datatypes.

The names of the Dogs in the Name columns used “None” to represent dogs with no names and some of the names had phrases such as his, my, a, an etc. After Further investigation, the names were changed respectively to their accurate names and records without names were given “None”. Afterwards I changed all the None names to Null which is the perfect representation for “no name”.

For the Scraped tweets data, I had to drop some of the fields (columns) which I was not going to make us of in my analysis. It was missing two records containing 2354 records instead of 2356 records.

For the Image prediction dataset, all the images were from WeRateDogs timeline were recorded and not all the images are dogs. I then created two filtering conditions which I used to select/filter out pictures that are less likely to be dogs. The fist condition was that p1_dog & p2_dog should be true since p1 and p2 has a higher confidence level than p3. The second condition was that, since p1 has the highest confidence level, I then considered values above its 75^{th} percentile (0.843855) and compared it with p1_dog, filtering for only values that are True. I combined the two filters together to filter for only records with a high possibility of being dogs.

After this, I changed the datatype of the tweet ID column in the image prediction dataset.

After filtering for Dog images only in the image prediction dataset, I used to dog images to filter for only dog related tweets in the twitter archive table leaving behind only records of tweets for dogs.

I noticed the presence of outliers in the rating numerator and rating denominator in the Twitter archived data and after observing them in a boxplot and identifying them using statistical methods, I replaced them with the mean values of the ratings numerator and ratings denominator respectively.

Tidiness Issues

After cleaning the data and fixing all the Data Quality issues, it is remaining to fix the tidiness issues.

I had to join the twitter archived data with the scraped tweet data using the tweet id to merge them both to become one table.

Then I noticed that multiple columns were used for dog stages and I had to combine them to be one and dropping all the “none” dog types.

And that concludes the data wrangling process