Project – Wrangle and Analyze Data

Data Wrangling Report

By Somya Bharti

Data Gathering-

In this process, we gathered the data from three different sources. The dataset that I gathered is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

First Source - Local file
Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets but here we have filtered around 2000+ tweets with ratings.

Second Source – URL

Image predictions

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 we will be downloading it programmatically using the Requests library and the given URL- https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Third Source - Twitter API
Gathered retweet count and favorite count which are two of the notable column omissions from Twitter's API.

Data Assessing-

Assessed the three datasets visually and programmatically and made the following observations-

Quality Issues-

- Many rows in the twitter enhanced dataset did not mention the stage of dog that is all the four stages in many rows are None.
- There are 1976 rows with no definition of the dog's stage.

- DataType of columns in the twitter enhanced dataset such as 'timestamp','retweeted_status_timestamp' are defined as String whereas it should be datetime.
- There are missing expanded urls in the twitter enhanced dataset.
- There are 181 retweeted_status_id which means that our dataset contains retweets as well.
- We do not need retweets in out dataset for analysis so we need to remove retweet_user_id and other columns related to retweets.
- Some of the names are 'a', 'an', 'the' which are not invalid.
- Source names need to be redefined without tags.
- The common numerator ratings given by @weratedogs are 11,12,13,16 so on. But,here we find that most of the ratings are too high such as 1776,960,666 etc.
- We know that @WeRateDogs keep their denominator as 10 always while rating dogs but here some of the ratings are 11,50,2,7,0,110 etc.
- After assessing the image prediction dataset visually, we find that for the last row, all the predictions of dog breed are false, which means, some images are not dogs.
- Some of the names of dog breed are not defined, like 'bookshop','bakery','book_jacket', 'orange'.
- The Image Urls are same for some images.
- The names of dog in Image prediction Dataset are separated by underscore instead of space.

<u>Tidiness Issues-</u>

- There are four columns namely doggo, floofer,puppo, pupper for the stages of a particular dog. We don't need four columns for the stage, only one column will be enough.
- We only need one master dataset for our analysis and visualizations, so we will merge all the three datasets collected from different sources.

Data Cleaning –

For this, the first step is to make dataset copies of the original datasets. There are three steps in this. First we define the process to clean the data , then convert it in code and finally test it.

The processes involved in cleaning are-

- Select the rows with null retweeted_status_id and remove the non-null retweets from the dataset.

- Select the columns related to retweets and drop them as it is of no use further.

- Select the four columns of stages and make a new dataframe.
- Add a new column 'Stage' to the new dataframe.
- Append the non-null values to column Stage.
- Add the new column 'Stage' to our original dataset.
- Drop the four columns 'Doggo', 'Floofer', 'Pupper', 'Puppo' from original dataset.
- Select the column 'timestamp' and change the DataType of timestamp from string to datetime.

- Select the column 'timestamp' and change the DataType of timestamp from string to datetime.

- Select invalid Names, which most probably starts with lower case letter and set those cells to None.

- Set the numerator rating in terms of denominator as most of the times denominator is 10 and then remove the denominator column with ratings not equal to 10.

- Select the source column and extract the text between anchor tags.

- Select the columns for which dog breed classifier is true and remove the images which are not dogs.

- Select the dog breed prediction columns that is p1, p2 and p3 and then replace underscore in dog breed's name with space.

- Merging all the datasets using join and make tweet_id as main key as it unique for everyone.
- Merge two datasets first and then merge the third dataset in the master dataset.