Wrangling report for WeRateDogs

By Sarah Ali

The Twitter Analysis Project task was to perform analysis from WeRateDog twitter profile from 2015 till August 1, 2017. The dataset was gathered from Twitter and provided by Udacity before the cleaning process begun.

Gather:

The first dataset WeRateDogs **Enhanced Twitter Archive** contains basic tweet data for all 5000+ of their tweets, but not everything, it was downloaded pragmatically. The second dataset was **Tweet-Json** has additional information about the tweets, which was read directly from a txt file using python, then stored into a data frame. The last dataset **Image Prediction** contained images about the dogs and was stored in an URL, so I had to obtain the texts from the web address and store the data in a data frame.

Accessing and Cleaning:

The data was accesses pragmatically using python, Quality issues and Tidyness issues were detected as followed. Copys of each dataframe was created so that every trail and error will not affect the original data set.

Quality issues:

1. Some columns had retweets associated with them in the Twitter archive csv, since we are working with original tweets, I removed records with retweet status ID, and i dropped the columns because they are not needed

2. Some texts have been mistaken for dog names, such names carry texts like a, an, like, the, that, etc, this is data inconsistency and missing data, a lot of dogs do not have names. First, I cleaned the colums by removing the wrong names and replacing it with "none", eventually every dog without a name was replaced with "none".
3. Irrelevant columns twitter_archive_cleaned.text, ImImage_prediction.img_num and twitter_archive_cleaned.expanded_urls after being accessed was dropped because I'll not be using it for analysis.
4. Redundant information surrounding the source text, i only need the actual values of the source and not the urls, this column was cleaned, I extracted the information we needed using pandas replace function to strip the html codes away.
5. Some rating_denominator do not equal 10, this was unusual, i dropped the rows with extremely high outliers, i.e., rows with denominator values of 100 and above, i then found the mean of the other denominators and replaced the high value with the mean.
   For the numerator, numbers that are higher than 10 is normal, as that is the unique way the dogs are rated. But it still had outliers, similarly I dropped rows with extreme outliers, 100 and above.

6. Null values recorded as None and empty string was replaced with NaN in twitter archive data frame.
7. Data type for tweet id was recorded as integer, and was changed to string datatype, and time stamp was saved as object, i changed it to datetime data type.

Tidiness issues

1. The Image prediction dataframe, some images were not that of dogs, which in turn means the tweet associated with it was not for dogs, so I filtered to select only images that has confirmed prediction as dogs. After that I combined the 3 colums into 1 using select conditions as the column with the highest confidence was selected.
2. The columns (doggo, floofer, pupper and puppo) do not need to be separated. Each dog is classified as one of these, i created one column using the pandas concatenate to combine values in the columns together.
3. Finally, the data set was merged into 1, using pandas merge and join in tweet ID.

The file was then saved in a master file as csv ready for export. This project was done using Python programming language. And properly documented with comments using the Define-Code-Test Method for data wrangling. The overall state of the data at this point was clean with