DATA WRANGLING REPORT

Project Name: 'We rate dogs'

In this project. It focus on data wrangling process on the data set from "we rate dogs" twitter page. The @dog_rates twitter account is an account that rates peoples dogs. The entire wrangling process includes gathering the data, Assessing the data, Cleaning the data

GATHERING DATA

The first phase of data wrangling is the gathering of data. In this project, data were gather from 3 different sources. The direct download of the 'twitter_archive_enhanced.csv file. This download is done manually. The second data set was gathered using the request library. The URL to the file was use by means of the request library. The final dataset 'tweet_json.txt was gathered by means of the tweeter api. Although I was unable to get the twitter authorization for the api. I use the data provided by the udacity class while I await my twitter approval for my API.

ASSESSING DATA

In this stage of the data wrangling stage, I read data gathered into the Jupyter library. This is to enable proper assessment of the data. It was done both manually and programmatically. This helps me to check for quality issues that the data contains which includes datatypes, missing values, duplicated data and many more. Likewise the structure of the data (Tidiness) is also checked. Tidiness issues might include merging tables together. For this project, all 3 tables are merge together.

Below are the list of quality and tidiness issues dealt with during the project.

QUALITY DATA

Twitter Archive

1. Delete retweet
2. Drop columns that are not needed
   (in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,
   retweeted_status_user_id, retweeted_status_timestamp)
3. datatype error in tweet_id, source, timestamp, rating numerator, denominator
4. Dog names with errors Eg. a which is not actually a dog name

5. Correct numerator with decimals values. (Correct values in text
6. Name columns with missing values
7. Dog stages with missing values and represented with none
8. Some dogs record include more than 1 dog stage
9. Source column includes links

Image predication table

1. 2075 images against 2356

Twitter API Table

1. Missing tweets

Tidiness issues

1. Doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into a single column named "dog_stage"

2. Image prediction table should be added to the twitter archive table.

3. Retweet_count, favorite_count, followers_count column should be added to the archive table

CLEANING

After the quality issues are determined and the tidiness issues determined, action are been carried out to ensure the data are of best quality. Pandas function like

a. Astype – use to change datatype
b. Drop – to drop unwanted colunms
c. Dropna – to drop null data rows
d. Merge – to join all 3 tables together.

After the entire cleaning is done, the data was analysed and visualize