Data Wrangling Report - WeRateDogs Twitter dataset.

Data Gathering:

The project required data to be downloaded from three different sources.

- First dataset twitter archive data was downloaded from a link provided in the project overview section.
- Second data about image prediction was downloaded from url provided in the project overview programmatically.
- The third data set was downloaded by setting up a twitter developer account and programmatically downloading the data. Initially i was not able to set up the twitter developer account and i wrote to udacity team for help. But eventually i set up a new twitter account , then a twitter developer account and managed to generate the required api keys for access and also download the data.

Data Assessing:

The gathered data was assessed both visually and programatically. For visual assessment i opened the files in google doc and tried to look for quality and tidiness issues. It was not easy to look for problems in the data visually , programatically i found many more issues which my eye could not catch. Some of the quality and tidiness issue are listed below.

twitter-archived-enhanced data

- The dataset contains a total of 2356 unique twitter id data.
- The columns in df dataframe in_reply_to_status_id , in_reply_to_user_id ,retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp , expanded_urls have missing values.
- The timestamp column datatype must to datetime instead of object.
- The null values in 'name', 'doggo', 'floofer', 'pupper','puppo' columns are represented as 'None'.
- The dog name columns consist (109)values such as 'a', 'an','the' etc. These names could be a typo and does not look like a legit name.
- The source column in df dataframe consist of four values 'Twitter for iPhone','Vine - Make a Scene', 'Twitter Web Client', 'TweetDeck'. The links to these source are hard coded html links which are repetitive of of no significance.
- The ratings_denominator column consist of values other than 10.
- The ratings_numerator columns contain both values less than 10 and some value greater than twenty. They also contain 3 and four digit ratings.

image-predictions.csv data

- The dataset consists 2075 unique twitter image data. It is less than the twitter archive dataset.

twitter api data

- All the twitter id did not return data using Twitter api. There are 2342 rows in the dataset. The data is a subset of the main twitter data.
- The id column name in the twitter api data is not consistent with the tweet_id column name in the other 2 dataset.
- There are column in the twitter api dataframe which are not required for this analysis and can be removed.

Tideness

- The column doggo, floofer, pupper, puppo can be combined into one column
- The column rating_denominator and rating_numerator can be condensed into one column.
- All 3 dataframe can be merged into for easy of analysis.

Data Cleaning

The above listed quality and tidiness issues were cleaned programatically in the data cleaning section. Not all the quality and tidiness issues were captured during the initial assessment . Some of the problems were identified during the data cleaning process and the above list was updated accordingly.

Saving the data:

On cleaning the data i felt it was easy to analyse by merging all the 3 data into one single file .