Udacity

Data Analysis Nanodegree

Wrangle Report

By: Mohamed Gabr

After assessing the dataset I found several issues that needed some adjustments

twitter-archive-enhanced.csv

Quality issues:

- rating_denominator column has values other than 10
- Dog names in name column have (none, the, an, a, this, None) as values
- there are 181 different retweets which are not original rog ratings
- (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) columns should be dropped after removing the retweets
- There are 78 replies to original dog rating tweets
- (in_reply_to _status, in_reply_to_user_id) columns should be dropped after the replies are removed
- timestamp is a string not datetime
- source column is not informative

Tidiness issues:

- timestamp consists of one column only (date + time) and not two separate columns
- twitter-archive-enhanced.csv and image-predictions.tsv are not merged together in one dataframe or csv file

After going through the issues mentioned above, I created a final dataset which included the modifications of the issues. This dataset was exported as 'twitter_archive_master.csv'