Udacity

Data Analysis Nanodegree

Wrangle Report

By: Mohamed Gabr

Introduction

This report is based on the data wrangling project created by Udacity for the Data Analysis Nanodegree program. The project's main task is to wrangle and analyze the tweets archive of the @dog_rates Twitter account which goes under the name WeRateDogs.

Several datasets were gathered in the process and after assessing them I found several issues that needed some adjustments.

Quality issues:

twitter-archive-enhanced.csv

- 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

tweet-json.txt

- (friends_count, source, retweeted_status, url) columns should be dropped

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’ and was a result of merging the three datasets together. Those datasets are: twitter-archive-enhanced.csv, tweet-json.txt, and image-predictions.tsv.