wrangle_report

May 27, 2019

1 Udacity Project - We Rate Dogs

by Bilal Hussain

This report shows the steps taken to complete the 'WeRateDogs' project by udacity. The aim of this project is to gather the data and then proceed to cleaning it by identifying quality and tidiness issues. There were three main steps to this project:

- Gather Data
- Assess Data
- Clean Data

The data was gathered three different ways. The first file was provided by udacity which contained the information about the tweets and the dogs. Second, I used twitter API to get the data regarding the tweets. More specifically, how many times the tweet was retweeted and favourited. The last file was a tsv file which was extracted through udacity using the requests module.

Through programmatic and viusal assessment, I made a list of the issues associated witht the data set:

Data Quality

- The three dataframes have different number of tweet_ids
- There are retweeted tweets in twitter_archive_data
- Retweeted status timestamp in twitter_archive is a string
- There are some denominators that are greater than 10 in twitter_archive. We can check the tweets text to see if there is a chance of extracting the correct ratings
- There are outliers in numerators in twitter_archive. We can follow the same process of fixing the denominators here as well
- Timestamp is a string rather than datetime. This is crucial as we will be using this as part of our analysis
- There are missing dog names. We can't go back in time to get the dog names for tweets, so we will have to drop these rows
- P1 has mix of upper case and lower case letters, no standard

Data Tidiness

- doggo, floofer, pupper, puppo should be combined to one column as they all represent dog 'stage' and can only carry one value

- There are three dataframes, while we only need 2 or even just one
- There are three probabilities in image_df, we only need one as that is the biggest probabiity

After addressing these issues one by one, I combined the three dataframes to make one dataframe ready for analysis.