Wrangle and analyze data

Wrangle report

5 TH PROJECT

Udacity Data Analysis Nanodegree

Udacity Project - We Rate Dogs

by Taif Alghamdi

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 favorited. The last file was a tsv file which was extracted through udacity using the requests module.

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

Data Quality

- The original tweets all we need so remove retweets
- Timestamp is a string rather than datetime
- in column name there are dogs named (a) it should be None
- p1,p2 and p3 have unnecessary underscore instead of space
- There is (59) row that does not have expanded_url, and they are not useful so, we should drop them
- There are some denominators that are greater than 10 in twitter archive
- There are outliers in numerators in twitter archive
- The tweet_id is should be str instead of int

Data Tidiness

- Unclear Column names(p1,p2,p3)
- in twitter archive doggo, pupper, puppo, and floofer are dog stage.
- unused column img_num