WeRateDogs – Wrangling Report

Introduction

In this paper I will describe my efforts for each wrangling step in *WeRateDogs* twitter archived dataset, that includes (Gathering, Assessing, Cleaning data).
WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering Data

Data was gathered in three different datasets; Twitter archived data, image prediction data, and favorites / retweets counts data.

1. Twitter Archived Data: This file has been provided by Udacity team after they did some enhancement and operations on the data. That includes splitting dog names, ratings, and dog stage.
2. Image Prediction Data: This data frame consists a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).
3. Favorites and retweets counts: I have gathered this missing data using tweepy library by using tweet id column in the first downloaded file, stored the full tweet json in a local text file.

Gathering data is always the first step in each wrangling or analysis processes, this can be done either by downloading a file from internet (Data frame #1 and #2) or gather the data using APIs (Data frame #3).

Assessing Data

Assessing data is the way we use our programmatic and visualizing skills to test the data for any quality or tidiness issues. In this data set there were many different issues with data sets, below is some part of issues observed:

Quality

- make the source as a string categorical variable
- add new column dog gender
- remove invalid tweets_id from enhanced archived data
- change this columns type (tweet_id) to string because We don't want any operations on it
- drop data about retweets (Project requirement) (retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp)
- remove hour:min:seconds and +0000 from timestamp
- divide the date into two columns year, month.
- manipulate name column (None names and wrong names)
- The numerator and denominator columns have invalid values
- tweet_id '810984652412424192'doesn't have a valid rating

- Missing values from prediction images dataset (2075 rows instead of 2356)
- We only want ratings with images. Not all ratings have images. (Project requieremnt)
- retweet and favorite counts to be integer instead of float

Tidiness

- join likes and retweets columns to the same archived data set.
- melt dog type 4 columns into one category column
- rating numerator and denominator should be one variable rating.
- merge prediction table with enhanced archived data (All tables should be part of one dataset)

Cleaning Data

The third step in my wrangling is to apply and work on the identified ussies from the previous step. I have worked on cleaning data manually, programmatically to solve quality and tidiness problems with the data. The process is to define the problem which I'm working on, work on it, and test my results to make sure the issue has been solved as required.

Conclusion

Wrangling data is always iterative. I had to gather more data while working on cleaning some part of the data.

I stored the cleaned data into a new csv file to use further.