Wrangle and Analyze “WeRateDogs” tweets

Wrangle_report

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. And I’m going to wrangle these data from twitter API and analyze it. This report will focus on wrangling process.

Gather

In this project, I need three information, WeRateDogs Twitter archive, image predictions, and the count of retweet and favorite.

1. WeRateDogs Twitter archive: WeRateDogs downloaded their Twitter archive and sent it to Udacity to use in this project. So I already have the basic tweet data with CSV.file in the beginning. And I use pandas”.read_csv()” function to read the file in Jupyter Notebook.
2. Image predictions: This file was stored in Udacity’s server and can be downloaded with “requests” function.
3. The count of retweet and favorite: This file is the hardest one to acquire among the three. I need to apply an account from Twitter developer. Then I need to generated consumer API, access token, and access secret keys to access twitter API. I used tweet ids from the archive file to programmatically download additional information. However, there still occurs a problem that twitter limits the number of queries you can do in 15-minute windows. So I revised my code, and saved it as a json file.

Assess

After gathering data, I use pandas functions like info, describe, value_counts, sample to assessing the data and summarize the quality and tidness issue as below:

Quality:

Df sheet

1. Retweet id are empty.
2. Too many useless column.
3. Timestamp is object.
4. Tweet_id is integer.
5. The biggest values of ratting denominator seems strange.
6. The biggest values of ratting numerator is stange

7. Numerator value should not under denominator.
8. Lots of the dogs' 'name', 'doogo' , 'floofer' , 'pupper', and 'puppo' is None.

Df_image sheet

1. Tweet_id is integer.

Df_tweet sheet

1. Tweet_id is object.

Tidness:

1. Doogo, floofer, pupper, puppo should be in one column.
2. Df and df_tweet sheet can be combined to one sheet since the information are related.

Clean

I clean my data follow three steps: define, code, and test. While cleaning one issue by one, I made some iterations back to assessing process because there was some other issues occurred. After checking all the issues are solved, I saved the clean sheets to new csv files.

Summary/Conclusion

Data wrangling is a fun but tricky process, each data has its own "feature". However, following the steps: gather, assess, and clean --perhaps sometimes will have some iterations—data will be clean and useable one day!