Wrangle Report

Introduction

The purpose of this project is to do data wrangling. The dataset used for this activity is Twitter user @dog_rates, also known

as WeRateDogs. WeRateDogs is a Twitter account that rates dogs with comment & ratings about the dog.

Project details

Major tasks in the project were:

• Gathering data

• Assessing data

• Cleaning data

Gathering data

The data used for this project was broken into three different dataset that were obtained as following:

- Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity online & manually downloadable file.
- The tweet image predictions as breed, the file named as (image_predictions.tsv)
- Using request account from twitter API: by using the tweet IDs in the WeRateDogs Twitter archive, Twitter API was queried for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. This file contains tweet ID, favorite count, retweet count, followers count, source, retweeted status and url.

Assessing data

After obtaining 3 datasets,

- Visually checked using head or df commands to see the data.

- Programmatically checked by using commands like info, value_counts, sample, duplicated.

Listed the issues in separate by quality or tidiness.

Cleaning data

There are 3 parts to this approach - Define the statement, write the code and verify by test.

- Original data copies were stored in case we need to recover the original datasets.
- I Had major challenge in accessing twitter enhanced csv file, so i had to convert to txt file to keep the tweet id same.
- Joined multiple datasets on tweet id joiner. and combined dataset was store as combined csv file.
- There were few challenges faced to clean up the data and used trend line to check visually trends between 2 variables.

Conclusion

- Data wrangling is a heart of the any project which makes you familiar with data and get better insights and findings errors in the datasets. Used various python pip packages to clean up the data for better visual.
- This project helps me familiar with big data approaches and twitter API access and libraries and how to read json unstructured data into panda's data frame.
- I have used visual and programming approach to clean up the data.
- Comments or labels or statements were made using headline to separate out the parts of the analysis.