Reporting: wragle_report

INTRODUCTION

This is project 2 of udacity nanodegree program and it is to test my feet on the topic learnt so far. After going through the topic called DATA WRANGLING, the project is a dataset from Twitter that was to be gotten either by using Twitter API (Tweepy) or downloaded from Udacity as Json's file. The dataset was from a page named WeRateDogs, this handle takes various dogs images and rate them. the structure of their rating makes them unique, the denominator is usually 10, while the numerator could be higher than 10 i.e. 13/10.

The report is divided into four levels:

- Data gathering
- Data Assessing
- Data Cleaning
- Data Storage
- Data Visulization

Data Gathering

In this level, there are three data uesd and these Data were gotten from different places and through various means.

- The first data, 'Twitter_achieve' was supplied by Udacity through manual download. This file contains coulumns like the rating of the dogs, the tweet id, text, dog stage and some other columns to aid analysis.
- The second, 'Image_prediction' was supplied through programmatic download. This file contains automated predictions of three different stages, level of confidence of this predictions and the breeds of dogs predicted.
- The third was gotten from Udacity as Json file. the file contain the favourite count, retweet counts of each dog posted.

Data Assessing

Under this level, the gathered data were assessed using Jupyter Notebook. also, different means were used to assess this files and detect errors, the quality of the files and the tidiness level. some of the means used are; .info, .describe, .unique, .value-counts. however, this means were through programmatic and visual process. the issues found were pened down as a list, divided into; **Quality and Tidiness Level.**

Data Cleaning

After completing the assessing level with at least **8 Quality, 2 Tidiness issues** detected, progress was made into cleaning of this data to fix the issues. This level consist of Define, COde and Testing to ensure that each issue were fixed. also, before commencement of cleaning process, copy of each data were made to prevent making changes in the original data. Finally, the three data were merged together.

Data Storage

The maerged data called 'Rating_final' was then stored to_csv file.

Data Visualization

Two visualizations were shown using barplot through Seaborn as sns to make comparison between columns of the dataset.

Challenges

- During the cleaning process, it was first difficult to write a spli function code to correct wrong fillings of names especially where names are showing 'a'. this error was due to naming pattern error.
- It was also difficult extracting data using the Twitter API, so i used the alternative.

Conclusion

Three conclusions were drawn;

1. More of golden_retriever dog breed were detected by the image predictor.
2. Among the dogs posted and rated on this handle, pupper has the highest number.
3. Cleaning data using both Visual and Programmatic means is good.
4. Lastly, finding the average mean shows that puppo has more rating than other dog stage names.

References

- https://stackoverflow.com
- http://pandas.pydata.org