Project Report: Gathering, Wrangling and Visualization of WeRateDogs Dataset

Introduction:

This project work contains datasets of twitter user @dog_rates, also known as 'WeRateDogs'. This is a twitter handle that rates people's dogs with a humorous comment about the dog. @dog_rates datasets used in this data analysis contains tweet data of over 5000+ which were obtained from three sources.

The aim of this data analysis project is to gather these data from its sources, evaluate and assess these datasets visually and programmatically for its quality and tidiness, then clean and visualize content of the data in a written report.

Task 1: Gathering

The first step in gathering these data was to start with the easiest sub-task - import twitter archive 'WeRateDogs' dataset from the CSV flat file using 'read_csv' pandas function into a panda 'DataFrame' (df1).

The second sub-task was to get 'WeRateDogs' prediction data from an 'tsv' internet source file using python's 'request' library into panda 'DataFrame' (df2). For the third sub-task, efforts to query WeRateDogs twitter archive data directly from Twitter API proved abortive. Nevertheless, data from a Tweet JSON text file was used instead.

Task 2: Assessment Efforts

Before the dataset were merged, each of the imported datasets (df1, df2 and df3) were assessed visually and then programmatically. In all, eight data issues were identified and about 60% of the visually identified data issues were related to poor data quality, with most having invalid data values.

The programmatic assessment provided us with the opportunity to delve deeper and examine the datasets using pandas functions and techniques. Five more data issues were identified in all the datasets, with 80% related to poor data quality. However, many of the identified problems were related to data inaccuracy and consistent issues.

Task 3: Cleaning and Visualization

Highlights of data cleaning includes dropping of the data columns containing retweets, 'NaN', changing timestamp features to the right format, and extracting time-related values like the year, month, and day.

Furthermore, the 'dog_type' columns of df3 were restructured, so that it can be easy to perform data visualization exercises. Lastly, the index of all the dataframes were reset and columns renames as applicable before they were all merged on 'tweet_id' column.

Count plots and point plots visualization techniques were used to understand timeline distribution of WeRateDogs tweets. While the distribution of impact of display test range

across different dog types and whether this feature influences tweet ratings were illustrated visually. The third research question answered was to understand the impact of dog_type on tweet ratings using boxplots and scatterplots.

Conclusion

Limitation:

Data analysis done in project is based on the level of data cleaning and data assessment of identifying 8 data quality issues and 2 data tidiness issues. Also, the merged dataset does not have approximate equal distribution of feature types. Hence, future work entails more data assessments, feature engineering and more data collation

Summary of Results:

- The favourite_count and retweet_counts are positively, linearly correlated. Generally, based on this data analysed, puppo dog_types tweets tend to get more favourite counts than others. While doggo and puppo dog types of tweets have the highest median retweeted counts.

- There is a slight negative correlation between display text chat range and mean dog rating assigned

- Tweets on dog rates are usually issued on Mondays for weekdays and more commonly in December of a year. It is also observed that the latter and early part of the year shows high dog rates tweets.