Data Wrangling Report

Introduction

This report depicts the practical knowledge we have grasped through the first lesson of Udacity Data Engineering Nanodegree. We were expected to assess the data, clean it if required and provide the insights for the data provided for twitter handle WeRateDogs. **WeRateDogs,** rate dogs in their website with a unique rating system where numerator is mostly greater than the denominator because they believe that every dog deserves at least 10 marks

Expectations

Gathering

We were required to gather data in 2 ways, one by programmatically by downloading Predictions.tsv file from the given website and other by fetching the data through TWITTER. We were expected to use tweepy in order to fetch details such as retweet count and likes for all the dogs referenced with the tweet id provided in the dataset twitter-archive.csv. Spreadsheet was provided as a whole which we manually downloaded and uploaded in Udacity workspace.

Assessing

After gathering we needed to assess at least 8 Quality issues and at least 2 Tidiness Issues. For Quality issues, I used 2 ways to assess it i.e. one by Visually checking the datatypes and gaps in data provided in all the 3 datasets and another programmatically by using functions such as .info(), .describe(), .value_counts(), .sample() etc.
For Tidiness issues which are structural issues, I have re-structured the Dataframes for all the three datasets and then finally merged it based upon principal column (**tweet_id**) and stored it in comma separated file to visualize and present analysis over the data gathered.

Cleaning

Cleaning was done based upon the Key Points mentioned in the Project Overview for Data Wrangling where we needed to reject re-tweets and tweets without photo primarily. There is other cleaning also done to bring consistency throughout and would be meaningful to present it.

Conclusion

Finally the master data was analyzed and insights were provided in the act_report.pdf which contained the details as below through the data which we visualized.

- Most Popular and Unpopular dog within this twitter handle
- Most Tweets contributing to this handle is from which source
- Dog with maximum number of Likes and Retweets
- Most popular dog type predicted through Predictions file provided

Data Wrangling Report

There were few of the review comments that was mentioned at the first submission –

- In Archive_Dataframe – tweet_id, in_reply_to_status_id, in_reply_to_user_id,, retweeted_status_id, retweeted_status_user_id should be of datatype **object** rather than int.
- In Archive_Dataframe – rating_numerator and rating_denominator should be of datatype float to handle decimal ratings.
- In Archive_Dataframe – rows having names as lowercase should be dropped as it will be usefull to identify not names.
- In Retweet_Dataframe – retweet_count and favorite_count should be of datatype int.
- Rating insight was visited again, compared the numerator found within the text of Archive Dataframe and the numerator column in the Archive dataset. Found 4 differences which were corrected.
- Multi Dog Stages for any tweet is handled properly for each and every row
- Master Archive Dataframe while loading it into .csv file with parameter index=False so that unnamed column is avoided.