wrangle_report

October 21, 2021

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

The aim of the project was to gather insights about dogs that are posted and rated from a twitter account called @dog_rates, also known as WeRateDogs. The tweets usually contain dog namees and a cute/humorous tweet about the dogs accompanied by a rating out of 10 and the dog's picture. The numerators are usually more than 10 as seen on the insights and the rating is justified because "they're good dogs Brent." To get the insights I needed gather, assess and clean the data.

0.1.1 Gathering data

There were 3 data sources that I needed to work with: one was given to use in an excel format called twitter-archive-enhanced, the other one we needed to download from a URL address. It contained the tweeted dogs' pictures and name predictions of the dog type with a percentage of accuracy in 3 trials. The last one I needed extract the data from tweepy which is a twitter API which had retweet and "like" favorite counts. After that was succesfully done I moved to the next step.

0.1.2 Assessing data

This step is to ensure that that I came out with quality and tidy data from visual to checking for approprite data type, programtic assesment. I found different issues from the 3 data sources like missing data in most columns- dog names, stages, imagges, etc, inaccurate data types on the dates and ids, mispelled column names, column data that was inaccurately extracted from the text, split coulmns that could be read better in one and data in split tables rather than 1, inaccurate, grammatic errors, non plausible rates and many more. I found and documented about 10 data quality issues and 2 data cleanliness issues. After finding the issues I moved to the next stage.

0.1.3 Cleaning data

After finding the various issues with the data, I needed to clean it so its useable for analysis. I applied the Define-Code-test frame work where I went through all the issues identified on the assessing stage, used the appropriate code and tested if my code worked. So that I dont mess up my originally extracted data while I clean, I made a copy of all the data frames before I started this process. Once I was satisfied, the last step I did was to combine the 3 data frames into one woorking data frame using the column tweet_ids as that's what the tables had in common and stored the data. Then I moved to what I would like to call the fun part.

0.1.4 Analyzing and Visualizing Data

There is so much insight to extract from the clean data. I found that it is indeed true that most ratings are greater than the denominator with the most frequent rating being 12/10. The most common dog type to be featured on the tweets is the golden_retriever. There is a positive correlation between retweet counts and the 'like" count. I also included some visualization on a pie chart showing that dogs at a pupper stage were the most featured on the tweets, taking up 64.6% of the pie share.

```
In [ ]:
```