WeRateDog Twitter Analysis Project

Data Wrangling Report

This project was aimed at analyzing Twitter data from the WeRateDog Twitter profile from 2015 till August 1, 2017, using the Python language.

The datasets used for this analysis were mostly pre-gathered. The first dataset was an archive file pre-gathered for this analysis. The second was another pre-gathered dataset with more information on the dataset like the images of the tweets. Lastly was a JSON file containing more information on the tweets like the retweets count, favorite count and followers count.

Data Wrangling Process

Quality Issues

The datasets were assessed visually using both Microsoft Excel and Programmatically using Jupyter Notebook. Some Quality Issues and Tidiness issues were discovered from our assessments.

Starting with the first dataset which is the Archive dataset, It had some Html tags included in the source name when they were being scrapped from the webpage. Same with the text column which had a short link embedded in the Text which wasn't needed
Next, I had some records which were just retweets so I had so to delete them from the dataset. Looking at the dataset I found out that most of the data types were wrong which could affect our analysis. The right datatypes were applied to the columns that were useful to our analysis like the tweet Id and the timestamp which were changed to string and DateTime respectively.
Lastly, on the Archive dataset, I noticed that some of the dog names were wrong which needed to be probably identified so I went with the none to show that there were unidentified from the collection of our dataset.

Looking at our second dataset “Image” I had to change the Tweet Id column which had a wrong data type which was changed to a string. Some columns were also needed for our analysis so they were dropped too

The numerator and denominator ratings were incorrectly gathered so I went ahead to extract them from the text column and making the corrections

Tidiness Issues

On the last dataset, I had to drop a column because it was not needed for my analysis

After the quality issues were handled I moved on to the tidiness issues.
Firstly I made to merge all my datasets so I could have a unified table for my analysis.
Lastly, there were multiple columns for dog stages and using my rational thinking skills to name the observations to make the following conclusion

From my study of the dataset, I made the following conclusions

- The none statements meant the dog stage was "undecided" yet meaning the dog was either in transition
- PupperPupper is equal to pupper
- doggopupperpupper is equla to pupper
- doggofloofer is equal to floofer

After that, I was done with the cleaning process and went along to save it