Project: Wrangling and Analyze

This project was done using three datasets. These datasets were gathered in different ways.

Data Gathering

The first dataset, weRateDogs was provided by Udacity as a CSV file named twitter_archive_enhanced .csv. It was directly downloaded from the classroom. This dataset contained information on 2356 tweets about dogs and their ratings.

The second dataset, twtImgPred, was downloaded programmatically using the requests library. It contains predictions on the tweeted dog's breed, each prediction with its level of confidence.

The third dataset, twtJson, was queried from the Twitter API, tweepy. It contains 2354 rows of information on weRateDogs tweets' retweet count, favorite count and other information.

Assessing Data

The datasets were first visually assessed in pandas and Excel and then programmatically assessed. After assessment, 8 quality issues and 2 tidiness issues were found.

The tidiness issues which had to do with structure were first cleaned and then the other quality issues were cleaned as well.

Before cleaning, a copy of each dataset was made for the purpose of cleaning.

1. To make the dataset tidy, the dog stages that were in different columns were collapsed into one single column. And then, the four different columns of dog stages which were now irrelevant were dropped.

2. For the second tidiness issue which had to do with having one table for the similar information. The three datasets all contained information on the tweets and the dogs tweeted about and so, having them in different columns was untidy and unnecessary.

They also had a lot of columns that were irrelevant to analysis. To solve this issue, the necessary columns in all three datasets were retained and the three datasets were then merged.

Note that the three datasets were not equal in number of rows. This meant not all dogs had prediction data. But that's okay.

3. The next cleaning was done on the master dataset. This was the issue of timestamp having the wrong datatype and also having extra figures (+0000) which were not necessary. The two issues were resolved by changing the datatype of the timestamp column to datetime.

4. Some entries in the dataset were found to be retweets and not original tweets. To achieve accuracy in analysis, these retweets were dropped.

5. It was also observed that some of the entries had rating denominators other than 10. This was contrary to the information provided about the dataset which stated that the dog rating was /10. For consistency, entries with denominators other than 10 (22 entries) were removed.

6. The primary key of the master dataset which was the tweet_id had the wrong datatype. It was in integer format but was converted to a string so that it won't be affected by summary statistics operations.

7. The other issues such as incomplete data and unnecessary columns were resolved during merging. Additional unnecessary columns were also removed to complete the cleaning process.

After cleaning, the data was stored and analyzed for insight and visualization.