Wrangle _report

This report briefly describes the effort of wrangling in this project. The dataset that was analyzed and visualized is the tweet archive of Tweeter user @dog_rate better known as WeRateDogs. This is a Tweeter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have denominator of 10. The numerators, though, almost always greater than 10, because they are good dogs Brent. The account has over 4 million followers and has received international media coverage.

The project aims to gather data from Twitter API and Udacity provided tweet data, to create analysis about the tweets and the predicted dog's breed. The entirety of this project was completed on the Udacity Project Workspace, however the reports were completed and exported as PDF files using Microsoft Word.

The Wrangling process is divided into three steps.

- Gathering data
- Assessing data
- Cleaning data

Data Gathering

I have gathered data from Enhanced Tweeter Archive and Image Prediction, which are provided by Udacity using the requests package.

The Enhanced Tweeter Archive file contains basic tweet data like rating, dog names, stages and some other related information.

The gathered data are loaded into three different DataFrame,

- t_archive : Loaded data from twitter_archive_enhanced.csv
- i_predictions : Loaded data from image_predictions.tsv
- favorite_count : Loaded data from retweet_favorite_count.csv

Additional data via Tweeter API was provided by Udacity for those who faced challenges creating a tweeter developer account.

Assessing Data

After data was gathered, it was assessed visually and programmatically for quality and tidiness issues. The following issues were concluded;

Quality

1. Invalid timestamp datatype (string not date time) which needed to be converted to the valid datatype.
2. Remove from table retweets and replies keeping only original tweets
3. Missing values in columns and unnecessary columns

4. Dog breeds uniformity
5. Clean up text column
6. Removing doubles
7. Erroneous datatype for tweet_id
8. Non-descriptive columns

Tidiness

1. Dog stages are separated into 4 columns

Data Cleaning

The issues were cleaned and saved to a master DataFrame. This were some of the issues that was cleaned;

- Correct invalid data type by converting timestamp to datetime.
- Find the retweets and replies using the retweeted_status_id and in_reply_to_status_id columns and remove from the DataFrame
- Removed columns with missing values using dropna() method. Also, use the drop() method to drop source column from table as well.
- In the image_clean table, the dogbreeds in the p1, p2, and p3 are converting all the names to a uniform way.
- In the archive_clean table, change the html ampersand code from "&amp;" to "&" in the text column Remove the "/n " the newline symbol Remove ending url link.
- In the archive_clean table, there were some tweets with two dogs being rated at the same time they were droped.
- Took both the clean_t_archive and tweet_clean tables and merge into one table using the join() method on the columns tweet_id.

Conclusion

I only managed to clean 8 quality issues and 1 tidiness but the DataFrame still has plenty of work to be done to it.