Data Wrangling Report

This project reports the data wrangling process conducted on the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

Three major steps were conducted to achieve the wrangling process. The steps are as follows:

- Data Gathering
- Data Assessing
- Data Cleaning

Data Gathering

The dataset was gathered from three different data sources

- WeRateDogs Twitter archive (a CSV file provided by WeRateDogs via Udacity). The archive contained basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- Tweet Image prediction file (tsv file). This file contained the confidence level predictions of the dogs using a neural network.
- Additional data was obtained from Twitter API (containing the retweet and favorite counts of each tweet ID). Unfortunately, my request for Twitter API was rejected; therefore, Udacity provided a tweet_json.txt file which contained the retweet and favorite counts for each tweet ID.

Data Assessing

The gathered dataset was assessed using two methods: visual and programmatic assessment.

- Visual Assessment: Each gathered data was displayed in the jupyter Notebook. Once displayed, the data was additionally assessed using an external application (Ms.Excel).
- Programmatic Assessment: Pandas' functions were used to assess the data.

After assessing the data both visually and programmatically, the following observations were noted:

Tidiness issues

- In the Twitter archive dataset, column headers such as doggo, floofer, pupper, and puppo are values, not variable names.
- favorite_count, retweet_count, and the Tweet Image prediction dataset should be part of the Twitter archive dataset.

Quality issues

Twitter archive dataset

- rating_numerator and rating_denominator have extremely large and inconsistent values
- The name feature has invalid records such as "a", "an", "the", and "bo."

- Delete columns with null values.
- Data in the source column includes a URL.
- Delete the null values in the expanded_urls column.
- Remove retweet, i.e., all records where retweeted_status_id is not null.
- Extract the ratings from the text column.

Tweet Image prediction dataset

- Columns p1, p2, and p3 should be merged as a single column.
- Some breed name begins with the upper case while others with lower case.

Data Cleaning

After the data was assessed, it was cleaned to solve the identified issues. The following steps were taken to clean the dataset.

- Merged all four columns (doggo, pupper, puppo, and floffer) into one column named dog_stage. Also merged the favorite_count, retweet_count, and prediction table to the archive_data table using tweet_id.
- Inconsistency in the rating was solved by performing feature engineering (where: rating = rating_numerator / rating_denominator).
- Invalid names starting with words such as "a," "an," "the," and "bo" were replaced with "None."
- Empty columns were deleted.
- URL was deleted from the source names.
- Null values were filtered and dropped in the expanded_urls column.
- Correct ratings for each tweet were filtered in the text column.
- All breed names were revised to start with upper case.
- All records where retweeted_status_id is not null were removed.
- Finally, the cleaned dataset was stored in the master DataFrame.

Conclusion

In the first iteration, eight issues have been documented about the dataset. However, the master dataset is not free of issues, as Data Wrangling is an iterative process. The wrangled data was stored in the twitter_archive_master.csv file with minor issues and was ready for Data Analysis.