1. Introduction

This ‘Wrangle Report’ is one of documents that I have to submit to complete the ‘Wrangle and Analyze Data’ project offered by Udacity. The dataset I was wrangling, analyzing and visualizing is the tweet archive of Twitter user @dog_rates. I completed data analysis by using project workspace offered by Udacity.

2. Data Gathering

I had to gather ‘tweet_json.txt’ file from the WeRateDogs™ tweets by using the tweepy package. However, I didn’t have any Twitter account and I preferred not to create a Twitter account for this project. Therefore, I accessed the Twitter data without actually creating a Twitter account by downloading the ‘tweet_json.txt’ file from Udacity. I named it as a ‘tweet_info’. I also gathered the files ‘twitter_archive_enhanced.csv’ and ‘image_predictions.tsv’ by reading these files. I named them as ‘twitter_archive’ and ‘image_predictions’.

3. Data Assessing

There are three data frames. First data frame is ‘twitter_archive’ which is loaded from ‘twitter_archive_enhanced.csv’ file. Second data frame is ‘image_predictions’ which is loaded from ‘image_prediction.tsv’ file. Third data frame is ‘tweet_info’ which is loaded from ‘tweet_json.txt’ file.

I read information of these data frames and checked duplication. In addition, I checked the number of values in specific columns and confirmed descriptive information. As a result, I could find several issues through this data assessing process. There are ten (10) quality issues to be fixed and two (2) tidiness issues to be fixed.

✓ Tidiness

- These tables are separated.
- Variables related to dog’s stage are spread in different columns.

✓ Quality

- Rating denominators are not consistent.
- 'tweet_id' columns should be string types.
- There are missing data in 'expanded_urls'.
- Dog's stages should be categorical data.
- There is inconsistency in dog breeds.
- There are missing dog names. (recorded as 'None')
- Several rating numerators are relatively high.

- Many entries are retweets or replies.
- timestamp column should be of datatime type
- There are dog names with 'a', 'an', 'the'

4. Data Cleaning

I solved the tidiness issues by combining the tables ‘twitter_archive_enhanced.csv’ and ‘image_predictions.tsv’ in one table called ‘twitter_archive_master.csv’. I also merged four columns (doggo, pupper, puppo and floofer) into one categorical column.

I solved the quality issues by using diverse programmatic methods. For example, I solved duplication problem by removing all retweets and reply. Also, I used the ‘.astype()’ and ‘.loc[]’ to fix data type problems. In addition, I used ‘.replace()’ to fix problems related to dog names.

5. Conclusion

I made the ‘twitter_archive_master.csv’ file which is the final file version of these data frames. Although, I found 12 issues and fixed them, this file is not totally free of issues. However, I’m sure that this file has a minored number of issues, and it will be helpful for data analysis.