Wrangle Report

1. Introduction

Data preparation is always the most difficult aspect of a data analyst's job; in this project, we will use our data wrangling skills to pull real-world data from Twitter, clean it, and analyze it. To build our analysis, we will use the original Twitter data from Twitter user @dog rates, as well as an image prediction dataset. Data analyst makes most of his time cleaning data for making more accurate analysis.

WeRateDogs is a popular Twitter hash tag in which people rate dogs with a denominator of 10 and a numerator that is usually greater than 10 to show how adorable the dog is. Those rates are countable.

2. Gathering data

The data was gathered from three sources:

- **Enhanced Twitter archive**: Udacity provided the WeRateDogs Twitter archive. This includes parts of the tweets, which accounts for more than 2000. This file was manually downloaded by clicking the following link: twitter archive enhanced.csv
- **Image prediction**: the results of neural network model for dog prediction using images. It is downloadable through links provided by Udacity.
- **Twitter API**: This part of dataset which counts the number of tweets and favorites or likes on the we rate dogs is accessible only through Twitter API.

3. Assessing Data

After evaluating the datasets, I summarized several quality and tidiness issues:

Quality problems:

- **rating denominator in archive table is with mean and 3 quartiles of 10, but with a max of 1770 and minimum 0. Alongside the dispersion of the data in the `rating_numerator`:** Some of the denominators are less than ten. One reason for this is that some text contains multiple number/number formats, and the ratings only convert the first number/number into the rating numerator and rating denominators, which is not the rating but something like date/time. Another reason is that some of the posted images contain multiple dogs, so they will rate 10 dogs with a denominator of 100. rating numerator: Some of the numerator is excessively large. Aside from the reasons listed in the denominator section, there is another reason: people simply adore the dog and give it such a high rating. So, if we solved the problem in the denominator, we don't have to worry about the numerators.

- **Dogs are with two or more category : puppoer, floofer etc ...**
- **Images with the same url are predicted for many times, but with different id**
- **All the twit_id are in integer, tweet_id for scrapped data is just id**
- **retweeted columns are full of NA values**
- **retweeted rows are not necessary for the analysis of data : we have removed all the rows of retweets.**
- **`timestamp` and `retweeted_status_timestamp` are in string type**
- **`Names` of dogs which start with a or the are not compatible with the common known name, alongside the names which starts with a lower case.**
- **59 tweets without `expanded urls`**

Tidiness problems:

- **Table of image prediction could be reformulated into 5 columns : id, image url, number of images, predicted and confidence interval. We merge all of p1_conf, p2_conf and p3_conf into confidence interval.**
- **All of the types of dogs could be merged to one column**
- Merge the scrapped datasets and archive data

4. Cleaning data

Multiple methods were used to clean the data quality and tidiness issues mentioned above, including pandas join, regular expression, combining multiple columns, pandas subsetting, removing missing values, and so on. We have built function to detect regex patterns alongside making new columns to summarize existing data.

I saved the cleaned version of the data into a csv file at the end of this section for future use.

5. Conclusion

This data wrangling project allows me to put what I've learned in the course into practice; I've struggled to find subtle data problems and cleaned those difficult, time-consuming data quality problems. The most difficult aspect of this project is dealing with strings using regular expressions.

Finally, I'm able to construct a well-structured wrangling processing notebook with details in every section: exhaustive steps in how I gathered, assessed, and cleaned the data, with illustrations of each problem and its resolution process.