DATA WRANGLING REPORT ON WERATEDOGS PROJECT

Introduction

The WeRateDogs project is one of the projects in the Udacity Nanodegree Program. In this report, I will be documenting a summary of the efforts and steps taken in wrangling the dataset. The Data Wrangling process consists of three different steps, namely:

1. Data Gathering
2. Data Assessing
3. Data Cleaning

Data Gathering

For the success of this project, and in order to make insightful analysis, I made use of data from three sources which were gathered in different ways.

The first dataset I gathered was the 'Enhanced Twitter Archive' which is a .csv file. This was the easiest dataset to gather, as I only downloaded manually.

Next was the tweet image predictions dataset, a .tsv file. I downloaded this file programmatically from Udacity's server using the Requests library.

Lastly, because the data in the twitter archive was missing some significant features such as the retweet count and favorite count, I had to query Twitter's API for each tweet's JSON data using the Tweepy library. This step in particular was quite daunting as I just became familiar with API in this course, as a result, I had to spend some more time understanding the concept and going through various examples of querying data from an API.

Data Assessing

In the assessment stage, I used Microsoft Excel to carry out the visual assessment, after which I moved on to the programmatic assessment on Jupyter Notebook. I checked the basic features of the datasets using some of Pandas library methods such as:

- .shape to know the number of rows and columns
- .info() to get the summary information of the datasets
- .duplicated() to ensure there are no duplicated rows
- .isnull().sum() to get the number of values missing in each columns

I also conducted some other assessments that helped me take note of the quality and tidiness issues present in the datasets, which include the following.

Quality Issues

1. Some tweets are retweets
2. Erroneous datatypes(tweet_id(str), timestamp(datetime))
3. Missing values in columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)
4. Nulls represented as None in name column
5. Source column still in html format
6. Wrong names assigned to dogs (eg. a, an, the, very, quite). It can be observed that all improper names are in lower case apart from 'None'

7. Some ratings have wrong values
8. Dog breeds are not standardized

Tidiness issues

1. Dog stages in archive table should be in a single column instead of four different columns
2. extra_tweet_data and image_pred should be part of archive table

Data Cleaning

I cleaned the quality and tidiness issues in the dataset with the use of the define-code-test structure which I found really helpful. It made me to first carefully think about how I would like to address each problem, explicitly define it, and then code it out, before finally testing to confirm that the code worked. The more clearly the define portion of this process is written, the easier it is to transform to code.

I cleaned the tidiness issues by concatenating the different dog stage variables to one column, then used .merge() to merge the extra tweet data and image predictions to the archive table.

For the quality issues, I corrected the columns with erroneous datatypes, replaced wrong values with the right ones, dropped missing values and retweet rows, used regex to extract the URL that was still in html format, and standardized the dog breeds to non-capitalized letters.

Conclusion

After gathering, assessing, and cleaning the datasets, I saved the cleaned data to a .csv file called twitter_archive_master.csv.