Wrangle report – TAIWO Ridwan

This report details the steps employed for wrangling the data used in the "WeRateDogs" project.

The three steps in the wrangling phase of data analysis – gathering, assessing, and cleaning – were strictly followed.

Gathering

The data used in the project were gathered from three different sources, including a CSV file named "twitter_archive_enhanced.csv" that was manually downloaded from the Udacity classroom, a tsv file that was programmatically downloaded from a given URL using the Requests library, and a JSON file that was downloaded with the aid of Tweepy library. It should be noted that the latest version of the Tweepy library was not working to retrieve the desired information. Thus, the lower version 3.10.0 was employed for this purpose. Additionally, a Twitter developer account was created for the authentication purpose.

Assessing

The downloaded files were assessed visually and programmatically in order to identify any tidiness or quality issues. For visual assessment, the three files were opened in Excel and were visually assessed by checking the columns and rows of the files. The three files were assessed programmatically in Pandas using various functions such as.info,. describe, .shape, and amongst others. As a result of the two assessments, two tidiness and eight quality issues were identified and highlighted below.

Tidiness issues:

- The three dataframes can be combined to form one neat dataframe
- The various stages of dogs: doggo, pupper, puppo, and floof(er), represent only one information. Thus, we can have one column instead of four columns for such information.

Quality issues:

- The three datasets do not have the same number of roles due to missing photos and maybe retweets
- There are columns that consist of values other than "10" in the denominator.
- There are unimaginable numbers in the rating numerator column.
- There are 181 retweets and 78 replies in the datasets, which are not needed.
- The names of some of the dogs are wrongly spelled.
- The timestamp column contains unnecessary characters.
- The data types of some of the columns are not appropriate.
- There are three sources of the data which are not clearly displayed.

Cleaning

Before cleaning, a copy of each file was kept in case of future reference. All the identified issues were cleaned carefully. The tidiness issues were first dealt with as they are structural problems that need to be tackled first to avoid repetition. Subsequently, the eight quality issues were cleaned. For each issue, the define-code-test format was adopted for clarity purposes.