Data Wrangling Report

Project Name: Udacity Data Wrangling Project

In this project, I completed the data wrangling process on a data set from "We rate Dogs" twitter page. The process included:

- Gathering the data
- Assessing the data
- Cleaning the data.

GATHERING: This is the first phase of the data wrangling process and the datasets were spread across 3 different sources. The first data set "***twitter_archive_enhanced.csv***" was available for direct download, hence I downloaded it with just a click. The second dataset "***image_predictions.csv***" was saved in a URL, this was downloaded using pandas ***request library***. The final dataset "***tweet_json.txt***" was extracted from twitter using the tweeter API, and read line by line to create a pandas dataframe.

ASSESSING: In this phase of the data wrangling process, I read all the datasets into the Jupyter notebook using the pandas read function i.e*. pd.read_csv* , and then performed assessment both visually and programmatically to check for quality and tidiness issues. **Quality issues** are issues that has to do with the data content, for example missing values, misspelled words, duplicates or incorrect data, while **Tidiness issues** has to do with the structure of the data that can so you down when cleaning, analyzing or visualizing the data, for example having multiple variables in one column or having 3 tables that can easily merged to form one table. For a dataset to become tidy it must meet Hadley Wickham's rules of tidy data which states that:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

After assessing the datasets, the following were issues discovered, I have grouped them under Quality Issues and Tidiness Issues for easy comprehension.

Quality issues

- Archive table
    1. The data contains retweets and replies (We only want tweets of original ratings)
    2. Some rating denominators are not equal to 10 (inconsistency)
    3. Timestamp column has datatype as "object" instead of "datetime"
    4. The column name "floofer" should be "floof" according to the dogtionary
    5. Invalid Dog names such as "None", "a", "the" and "an"
    6. Tweet_id column datatype should be object since we are not performing calculations on it
    7. Some rating numerators are way too high, they usually fall within 0-15, others above 15 are probably outliers

- Predictions table
    8. Tweet_id column datatype should be object since we are not performing calculations on it
    9. p1,p2 and p3 have some dog names beginning with lower case while others beginning with upper case

Tidiness issues

1. All data can be used to form one table for easy analysis (they have "tweet_id" in common)
2. Dog stage name should be in one column instead of 4.

CLEANING: After assessing, I cleaned the all issues detected by first defining how or what to correct, writing and running the code and finally testing if the code worked. During the cleaning process, I used tools and functions such as:

- Drop- To drop irrelevant rows and columns
- Replace function- To manually replace wrong entries which the codes might made errors inputting
- As type- To change column types to their appropriate types.
- Merge- To merge the 3 datasets to form a master dataset and so on.

After cleaning and merging to form a master dataset. I saved as a csv file read for analyzing and visualizing .