WRANGLE REPORT

1. **Definition**

   Data wrangling is a crucial step in the data analysis pipeline as it helps ensure that the data is accurate, complete, and ready for analysis. It requires a combination of technical skills, domain knowledge, and critical thinking to effectively handle the challenges associated with real-world data.

   Data Wrangling process is divided into three main steps:

   • Gathering.
   • Assessing.
   • Cleaning.

2. **Gathering Data**

   Data was gathered from three different sources:

   a. The data we used for this project is a collection of tweets from the WeRateDogs Twitter account. We got this data from Udacity in a file format called CSV ("twitter-archive-enhanced.csv"). We used a tool called Pandas to read the data from the file and store it in a structured format called a DataFrame. We named this DataFrame *"df_arc".*

   b. We used the 'get' method from the Requests library to download the file and then used Pandas to read the content and store it in a DataFrame called *"df_img"*.

   c. In order to get more data from Twitter as an alternative, we used the open function to read row by row and translate into pandas dataframe and store it in a DataFrame called *"df_tweet"*.

3. **Assessing Data**

   The data we collect is often not in the formats we need or want. Once we gathered the necessary data, I discovered the following problems with it.

   a. Quality Issue for twitter_archive_enhanced

   Quality

   Visual Method

1. Invalid names or "None" names (Column: name)

   Programmatic

2. Invalid ratings, Max rate should be 10 not 1776 (Column: rating_numerator)
3. Invalid denominator, I expected fixed 10. (Column: rating_denominator)

4. Convert to date format. (Column: timestamp)
5. It is similar to zip code, it must be a string. (Column: tweet_id)
6. Same dog retweeted so it can be dublicated records. (Column: retweeted_status_id)
7. Same dog replied so it can be dublicated records. (Column: in_reply_to_status_id)

Tidiness

Visual Method

8. HTML tags, URL, and content in a single column.

Programmatic

9. Categorical variable **translate into one column** as shown drug example shown in Udacity. (Column: doggo, floofer, pupper, and puppo)
10. There is two information in a single column. It should be two column as **text** and **URL**. (Column: text)

b. Quality Issue for image_predictions.tsv

Quality

Visual Method

11. Dog's breed is not standard, Capital Letter Issue. (Column: p1, p2, p3)

Programmatic

12. ID must be string. (Column: tweet_id)
13. There is dublicated entries which are belongs to retweet or replies. (Column: jpg_url)
14. Merging these two tables (df_arc, df_twitter and df_img) into one. (Table: twitter_master)
15. **My Suggestion**: If there is any False prediction image is not relevent to dog. If all predictions are True, image is relevent to dog.

c. Quality Issue for tweet-json.json

Quality

Programmatic

16. ID must be string. (Column: tweet_id)

4. Cleaning Data

The dog's names issue was solved evaluating if it starts with a capital letter it was a name if not it was an ordinary word and I have converted to "None". Most of the issues involving non-usual values to rating_numerator and rating_denominator were solved as Denomiator should be constant and max nominator should be equal to max denomiator to reach higest rate 1.00.

In regard to the duplicated information, I decided to remove all retweets and reply to avoid double entries of the same dog.

The most challange was cleaning of image prediction table. Because the predictions are not exact information. Therefore, to use these data for insisghts were not easy as shared on dataframe.

Picture -1: Prediction: “Car Mirror”

Once the data was prepared, I analyzed it using visualizations, as documented in the act_report.pdf file.