WRANGLE REPORT

1. GATHERING DATA

Data is gathered from three resources and saved as three DataFrames**: df1, df2, df3**.

1.1 GATHER DATA FROM FILE ON HAND

I gather first data from **twitter-archive-enhanced.cs** file**.** I read the file using **pd.read_csv()** then store it in **df1**.

1.2 DOWNLOAD FILE USING REQUESTS LIBRARY AND URL

Download file **image_prediction.tsv** programmatically from the Internet and store data in **df2**.

1.3 GATHER DATA FROM TWITTER API USING PYTHON'S TWEEPY LIBRARY

Get **retweet_count** and **favorite_count** from twitter API for records. Then save the data as text file **tweet_json.txt**, then read the file and store data in **df3**.

2. ASSESSING DATA

FIRST DATAFRAME

QUALITY ISSUES

- 'None' in the dataset should be replaced by 'NaN' in columns: name, doggo, floofer, pupper, and puppo.
- Replace dog name with NaN for not corrected dog name like 'a' and 'an'
- Missing data in columns: name, doggo, floofer, pupper, and puppo.
- Wrong data types in timestamp column.
- Drop rows that contain null value in tweet_id

TIDINESS ISSUES

- doggo, floofer, pupper, and puppo should be in one column.

SECOND DATAFRAME

QUALITY ISSUES:

- Some predictions are not dogs, like seat_belt, web_site, and remote_control.
- Data Type is wrong: **tweet_id**.

TIDINESS ISSUES:

- Change column name: like **p1_conf** and **p1_dog**.
- Capitalize prediction dog type and remove underscore.

THIRD DATAFRAME

QUALITY ISSUES:

- Remove duplicated rows

TIDINESS ISSUES:

- Merge the three **dataframes** with **tweet_id** column

3. CLEANING DATA

Copy **df1**, **df2**, **df3** as **df1_clean**, **df2_clean**, **df3_clean**.

DEFINITION 1

- 'None' values in the dataset replaced by 'NaN' in columns: name, doggo, floofer, pupper, and puppo. then store it in **df1_clean**
- Replace dog name with NaN for not corrected dog name like 'a' and 'an'
- Create new column for columns doggo, floofer, pupper, and puppo and label it as DogStage

DEFINITION 2

- Correct the wrong data type in **df1_clean** (timestamp).

DEFINITION 3

- Drop rows that contain null value in tweet_id in **df1_clean.**

DEFINITION 4

- Correct the wrong data type in **df2_clean** (tweet_id).

DEFINITION 5

- Change column labeling. Ex: **p1_conf** to FirstPrediction , **p1_dog** to **IsFirstPredictionConfidentBreedDog, and p1_conf** to **FirstPredictionConfident**

DEFINITION 6

- Capitalize prediction dog type and remove underscore.

DEFINITION 7

- Remove duplicated rows then store the data in **df3_clean.**

DEFINITION 8

- Remove rows that **rating_denominator** is not equal to 10 then store the data in **twitter_archive_master.**

DEFINITION 9

- remove outliers' rows from **rating_numerator** then store the data in **twitter_archive_master.**

DEFINITION 10

- Replace numerator and denominator columns with **DogRate**

4. STORING DATA

Store the clean DataFrame **df1_clean,df2_clean,and df2_clean** in a CSV file named **'twitter_archive_master.csv'**