Reporting: Wrangle report

The wrangle process was divided into three which are
Gathering of data
Accessing of data
Cleaning of data

GATHERING OF DATA

The datasets used are gathered from three different sources. The sources are enlisted below

1. CSV file
2. Website link provided for downloading
3. API file saved in txt format.

ACCESSING OF DATA

Each dataset was checked visually and programmatically. In the aspect of visual, the whole file was loaded for a visual check, also the datasets were saved in CSV file format for more visual assessment with the use of Microsoft Excel.

While accessing the file visually, the below issues were observed

- ❖ retweeted columns have values and do not suppose to have values
- ❖ floofer, doggo, pupper, and puppo have values (none and their name) each instead of one each
- ❖ in_reply columns have many missing values
- ❖ img_num has 4 unique values instead of 3 values
- ❖ P(x)_dog columns have two unique values each (True and False) instead of only True each

While accessing the data programmatically, with the use of info(), isnull(), and nunique() methods, the below issues were observed

df_weratedog:

- ❖ tweet_id is an int, not a string
- ❖ retweeted_status_id is a float dtype instead of being a string data type
- ❖ retweeted_statust_timestamp is an object dtype instead of being a time data type
- ❖ timestamp is an object, not a DateTime

df_image_predictions:

- ❖ tweet_id is an integer dtype instead of being a string data type

df_tweets:

- ❖ tweet_id is an integer dtype instead of being a string data type

On the tidiness issues below issues were observed

df_WeRateDogs:

- ❖ floofer, doggo, pupper, and puppo columns suppose to be a value under a variable (columns)

df_image_predictions:

- ❖ P(x)_dog columns have to be in a columns

CLEANING OF DATA

Before the commencement of the cleaning process, copies of the three datasets were created to avoid losing that data.

The first issue treated on the cleaning path is the datatype of the wrong columns while the correct data type was assigned to those columns. The dropping of the rows that contain 4 as a value in img_num column of the df_image_prediction table, since our image prediction focuses on the first three images serves as the second issue that was addressed.

Sub tables were created from df_image_prediction table for the filtering of false predictions of each dog image. The sub-tables were created by filtering the img_num and filtering out false predictions. While filtering columns that do not incline with the img_num were deleted. After all this process, the three sub-tables created were merged together where each prediction dog column for each of the subtables was renamed to prediction.

Based on the information, retweeted columns that have values are insignificant to the df_WeRateDog table, so the table was filtered using retweeted_status_id. The result led to the deleting of some columns that do not have enough values and will not be used for analysis.

In df_WeRateDog table, floofer, doggo, pupper, and puppo were un-pivoted to form a new column due to the reason that the four columns are supposed to be values which led to the creation of a new column named dog_rate, and the former columns were dropped. More so, due to the former columns having two values each, during the un-pivoting process, the second value changed to **0** which was later replaced with **None**

The three datasets were merged using the inner type of merging dataset and tweet id serves as the primary key which will enable us to have all the data with the same tweet_ids in the new table created.