WeRateDogs – Twitter Data

1. Gather Data

We download the **twitter-archive-enhanced.csv** manually, the **prediction.tsv** file programmatically and the **tweet_josn.txt** using api.
Once we have all the 3 files, we create a dataframe for each of these files:

a. twitter_df
b. predict_df
c. api_df

2. Assessing the data

We assess this data by going through each of the dataframes and looking for issues like missing values or incorrect datatypes, first visually and then programmatically.

We use many pandas functions to find the issues, like value_counts(), duplicated() or info().

We find the main issues like:

Quality

twitter_df table

- The datatype of the tweet_id column is integer and should be str
- The datatype of columns such as timestamp and retweeted_status_timestamp are incorrect
- The source column contains html code
- Some of the names of the dogs are incorrect.
- The columns which have missing values in doggo, floofer, pupper , puppo - has "None" instead of NaN
- Rating_numerator and rating_denominator have some inconsistent values

predict_df table

- The datatype of the tweet_id column is integer and should be str
- Contains retweets(multiple rows in column 'jpg_url')
- Sometimes the pictures do not contain dogs at all
- The predictions are sometimes uppercase, sometimes lowercase with an '_' present I
- Also there is a "_" instead of a whitespace in the predictions

api_df table

- The datatype of the tweet_id column is
- column is integer and should be str

Tidiness

twitter_df table

- The dog stages are in multiple columns and must be put into one column

df_predict table

- The prediction must be reduced to one column

all tables

- All the tables share a common tweet_id and must be merged together

3. Data Cleaning

Cleaning steps:

1. Merge the tables together (Tidiness)
2. Drop the replies, retweets and the corresponding columns (Quality)
3. Drop the tweets without an image or with images which don't display doggos (Quality)
4. Clean the datatypes of the columns (Quality)
5. Extract the source from html code (Quality)
6. Remove the incorrect names from the name column (Quality)
7. Reduce the prediction columns into one column: breed (Tidiness)
8. Clean the new breed column by replacing the "_" with a whitespace and make them all lowercase (Quality)
9. Convert Nones in all the 4 dog states to Nans (Quality)
10. Merge the dog state columns 'doggo', 'floofer', 'pupper', 'puppo' into a single column (Tidiness)
11. Clean the wrong numerators and denominators (Quality)