WRANGLING REPORT

INTRODUCTION:

The dataset that was worked on by me is the tweet archive dataset of twitter user @dog_rate, which is also referred to as WeRateDogs. This twitter account rates so many people's dogs with a lot of likes and comments about dogs. The WeRateDogs has over 9 million followers and they are widely known.

Step 1: Gathering Data

I gathered the first dataset by downloading it via this links; https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archiveenhanced/twitter-archive-enhanced.csv. The second dataset was downloaded programmatically from the web using the python request library. and https://video.udacity-data.com/topher/2018/November/5be5fb7d_tweetjson/tweet-json.txt. The two link were both provided in the classroom. Then I loaded them as pandas DataFrames in My jupyter notebook environment.

Step 2: Accessing the Data

After gathering each of the three datasets, I assessed them visually and programmatically for quality and tidiness issues.

In this process, I found some issues and they are as follows;

Quality issues

1. Dropped the 'tweet_id' column in the new dataset.
2. Removal of rows named 'False' values from the 'p1_dog', 'p2_dog', and 'p3_dog' columns.
3. Converted the 'timestamp' column into its appropriate datatype.
4. Removed created 'index' and 'timestamp' columns after resetting index and extracting texts from the timestamp.
5. Removed or replaced null values.
6. Removed Duplicates.
7. Removed rows with invalid names like 'none', and 'a'.
8. Fixed the datatype issue of in the merged dataframe.

Tidiness issue

1. Removed non-essential columns across the dataframe.

2. Renamed the 'tweet_id' to be the same across all dataframes.

3. Merged the various dog stages present in the 'twitter-archive' dataset.

4. Merged the dataframes to form a new one.

I observed that the rows in the initial columns are over 90% duplicated, there were null values in the source columns. I also observed that there were columns with wrong datatypes and there were empty cells.

Step 3: Cleaning Phase

1. Loaded the three (3) datasets into pandas dataframes.
2. Merged 'twitter-archive-enhanced-2.csv' and 'image-predictions-3.tsv' togetheR to form a new dataframe.
3. Dropped non-essential columns in the new dataframe.
4. Renamed the column to be used to merge on with 'tweet-json.txt' dataframe
5. Merged the various dog stages present in the 'twitter-archive' dataset.
6. Merged all dataframes into one dataframe.
7. Dropped 'tweet_id'.
8. Removed rows containing 'False' values from the 'p1_dog', 'p2_dog', 'p3_dog' columns across the dataframe, so that only dogs are contained in the dataset.
9. Converted the 'timestamp' column into its appropriate datatype.
10. Split the 'timestamp' column into date, year, month, and day columns.
11. Removed created 'index' and 'timestamp' columns after resetting index and extracting texts from the timestamp.
12. Removed or replace null values.
13. Changed Datatypes.
14. Removed Duplicates.