JAMIL ABDULAI

PROJECT 4:
WRANGLE AND ANALYZE DATA
WRANGLE REPORT – WE RATE DOGS DATA
AUGUST 7, 2022

OVERVIEW

GATHERING DATA:

The project was initialized by downloading the **'twitter-archive-enhanced.csv'** from the Udacity website and installing all the python packages before importing them into Jupyter Python Notebook. I subsequently uploaded the **'twitter-archive-enhanced.csv'** into the notebook followed by programmatically downloading the **'image-predictions.tsv'** file from the Udacity servers using the request library. Furthermore, I used the Tweepy library to create the **'twitter_data'** by accessing and downloading Twitter's JSON data. I initiated the process by extracting a list of tweet IDs from the **'twitter-archive-enhanced.csv'** and looped through each ID to query Twitter's API with the respective IDs to extract the necessary tweets JSON data.

Furthermore, the data was transposed into a text file created, **'tweet-json.txt'** whereby each data is corresponding to a new line. Upon the completion of the query into the text file, the text file was read per line to capture each tweet's information i.e., tweet_ID, retweet_count, favorite_count, and follow_count using the JSON library which was appended into an empty list. Lastly, the list of dictionaries was converted to Pandas DataFrame and saved into the **'twitter_data' file.**

ACCESSING AND CLEANING DATA:

A great deal of emphasis was implemented into the accessing and cleaning phase to ensure the right measures were incorporated to achieve e the desired results.

Under the **quality category** with the **twitter_df_archive** table, the following manipulations were made.

- The columns not needed were dropped.
- The **tweet_id** was converted to a string from the **twitter_df_archive.**
- The timestamp was converted to datetime
- The source was converted to a category datatype.
- Retweets were deleted by filtering the NAN of **retweeted_status_user_id**.
- The hyperlinks were removed from the tweets.
- A change was made in the error name from the dog's name to NAN
- An extraction of HTML values was initiated from the source
- A separation of the dog stages was conducted to examine which stages have the dogs been through.
- A change of the missing values in the dog's name to unnamed.

Under the **image_prediction** table, the tweet_id was converted to strings and missing rows were dropped whereas one change was made to the **twitter_api** table, a conversion of the **tweet_id** to string.

Likewise, the tidiness category provided more manipulations to achieve the desired results. With the **twitter_df_archive** table, I merged the multiple columns into a single column, **dog_stage,** and the **twitter_api** and **image_prediction** tables were merged with the **twitter_df_archive** table.

STORING THE CLEANED DATA:

After the cleaning and assessment were completed, the above data was saved in the **twitter_archive_master.csv**.