PROJECT 4
Wrangle and Analyze Data

Gathering Data

I followed the instructions given by Udacity to gather data for this project – Wrangle and Analyze Data

- I started my project by downloading the CSV file '**twitter-archive-enhanced.csv'** file manually
- Next, I downloaded the file **'image-predictions. tsv'**
- Then I query the Twitter API for each tweet's JSON data by using the *Tweepy* library and store it to **tweet_json.txt**. After that, I have read the text file line by line, and obtain each tweet's information (tweet_id, retweet_count, favorite_count, followers_count, and friends_count), using the JSON library to append it to the empty list. Finally, I convert the list to pandas DataFrame and Save it into **'twitter_data'**

Assessing and Cleaning Data

Some quality and tidiness issues were apprentice in three tables. Detail of the issues and solutions are the below:

1. **Twitter_archive table:**

- Drop some columns that do not need analysis (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Some name of dog has inaccurate entries like 'a',' an','the'… in the name column. So I replace it with None.
- Text column includes a text and short link. So I removed the hyperlink in tweets

- Change datatype of timestamp column to timestamp

2. **Image_prediction table**
- Some row missing images (only 2075 out of 2356 counts) => Drop rows with missing value
- Change datatype of column tweet_id to string
3. **Twitter_api table**
- Change datatype of column tweet_id to string

Tidiness

1. **Twitter_archive table:**
- Create a function to extract the dog stage from each of four dog stage columns (i.e, doggo, Puppo, pepper and floor). And store it to the column dog_breed
2. **Image_prediction table**
- Merge table to archive table
3. **Twitter_api table**
- Merge table to archive table

Storing Cleaned data

- When the data is fresh and clear for analyzing. I saved it to the dataset master_data.csv