Wrangle Report

There were a lot of issues and problems that need to be solved for analyzing and visualizing the data. Data wrangling is important process to make the data ready for answer your questions. Data wrangling process consists of three stages, which are Gather, assess and cleaning. In the next section I will discuss what is done in each stage.

Gathering Data:

As we know, we have two CSV files, first one is **twitter-archive-enhanced.csv** and **image-predictions.tsv**, **twitter-archive-enhanced.csv** has information about tweet, but it is not all what we need, we need to gather information about the number of retweets(retweet_count) and the number of likes(favorite_count).

Tweepy is an easy-to-use Python library for accessing the Twitter API, it is required a permission from Twitter to use the API, after the permission has arrived, the ids of the tweets were used to gather the needed information, and make new Dataframe called **twitter_info** that contains three column, which are (id, retweet_count, favorite_count). Not all the tweets exist, some of them was deleted. 2331 tweets are collected correctly out of 2356.

Assessing & Cleaning Data:

In this section, I will mention the problems and issues in each dataset, and what is the solution for those problems.

Quality Issues:

- ***twitter-archive-enhanced.csv***
  1. Erroneous datatypes at **tweet_id** and **timestamp** columns.
     - Since the **tweet_id** will be not used for any numerical analysis, I should convert it from **int** datatype to **String** datatype.
     - **timestamp** column contains the date in string form, so we can use **pd.to_datetime()** function to convert it to datetime.

  2. There are duplicated data because of retweet.
     - They are just retweets of the same tweets exist in our dataset, so we should delete them from the dataset.

  3. Some retweets are irrelevant to what we need. Examples, the rate of a dog is 1773/10, it was a celebration of American occasion.
     - They should be deleted.

  4. There is a tweet that announces the start date of the account.
     - It should be deleted.

  5. There are a lot of miss extractions of dog rating. Most of them extract date rather than the rate.
     - So I collected the id of all of them, and manually extract the correct rate, then the correct rate replaced the wrong rate.

- ***image_predictions.tsv***
  1. Erroneous datatypes at **tweet_id** and **timestamp** columns.
     - Since the **tweet_id** will be not used for any numerical analysis, I should convert it from **int** datatype to **String** datatype.

  2. Breeds exist in lowercase and uppercase at **p1, p2, p3** columns.
     - Change all these columns values to lowercase string.

- ***twitter_info table***
  1. The name of id column is not expressive.
     - Change the name of **id** column to **tweet_id**.

  2. Erroneous datatypes at **tweet_id** and **timestamp** columns.
     - Since the **tweet_id** will be not used for any numerical analysis, I should convert it from **int** datatype to **String** datatype.

Tidiness Issues:

- ***twitter-archive-enhanced.csv***

  1. We could make column describe the normalized rating because there are some tweets have more than one dog and the rating is very large because they add all of them, example 5 dogs rating is 55/50.
     - Make new column called **rating_frac** by divide **rating_numerator** by **rating_denominator**.

  2. One variable in four columns (**doggo, floofer, pupper, and puppo columns**).
     - Melt the columns to one column called **dog_stage**.

  3. Un-needed columns:
     - Delete columns:
       - **source**
       - **in_reply_to_status_id**
       - **in_reply_to_user_id**
       - **expanded_urls**
       - **retweeted_status_id**
       - **retweeted_status_user_id**
       - **retweeted_status_timestamp**

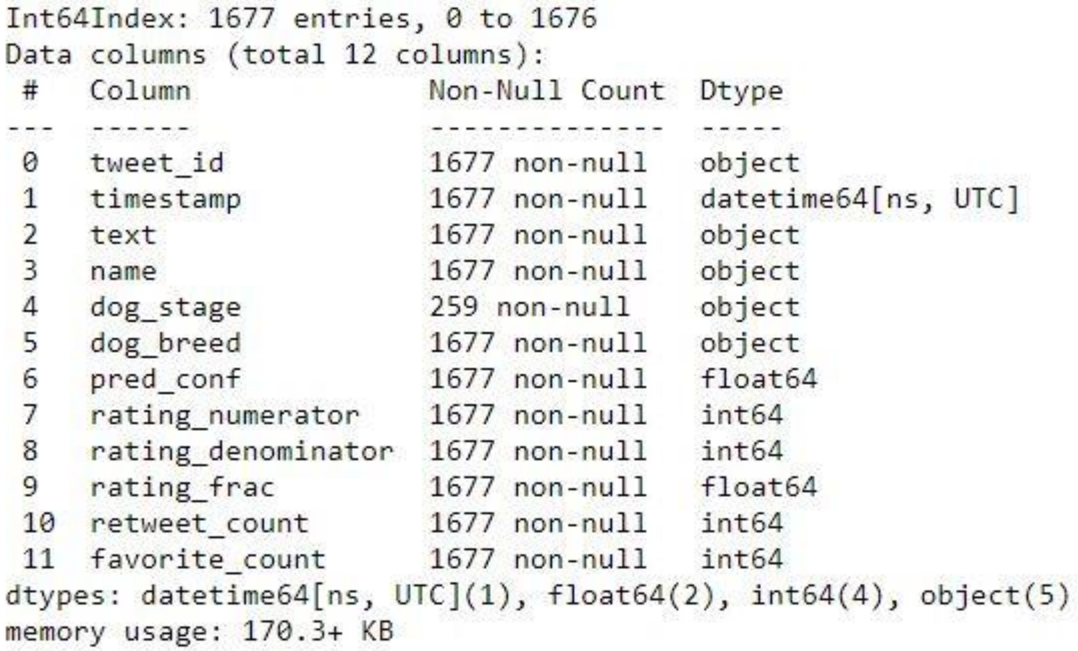
- ***image_predictions.tsv***
    1. Predictions should be in one column and, the confidence also.
        - Choose the correct predictions and its confidence and make two columns called **dog_breed** and **pred_conf**.

    2. It should be part of **twitter_archive** table.
        - Merge **twitter_archive** table with **image_predictions** table.

- ***twitter_info table***
    1. **retweet_count** and **favorite_count** should be part of **twitter_archive** table.
        - Merge **twitter_archive** table with **twitter_info** table.

The Final Dataset:

- Records: **1677**



```
Int64Index: 1677 entries, 0 to 1676
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1677 non-null   object
 1   timestamp           1677 non-null   datetime64[ns, UTC]
 2   text                1677 non-null   object
 3   name                1677 non-null   object
 4   dog_stage           259 non-null    object
 5   dog_breed           1677 non-null   object
 6   pred_conf           1677 non-null   float64
 7   rating_numerator    1677 non-null   int64
 8   rating_denominator  1677 non-null   int64
 9   rating_frac         1677 non-null   float64
 10  retweet_count       1677 non-null   int64
 11  favorite_count      1677 non-null   int64
dtypes: datetime64[ns, UTC](1), float64(2), int64(4), object(5)
memory usage: 170.3+ KB
```