# wrangle_report

June 21, 2022

## 1 Wrangle Report

Wrangling was done in three phases:

1. Gathering.
2. Assessment.
3. Cleaning.

### 1.1 Gathering

The data needed for this project was spread across three different data sets;

- Files 1 Was download manually by clicking the link to it the project resources. Once it was downloaded, upload it and read the data into a pandas DataFrame as twitter_archive_enhanced.csv.
- file 2 was downloaded programmatically using the Requests library from Udacity's servers. A folder was created and the response object of the request which was a TSV file was written into the folder. This was then read into a pandas Datframe object using pd.read_csv() setting the "sep" parameter to ''as image_predictions.tsv.
- File 3 contained each tweet's retweet count and favorite ("like") count and other additional data. This was gathered Using the tweet IDs in the twitter_archive_enhanced dataframe. The Twitter API for each tweet's JSON data was queried using Python's Tweepy library after which each tweet's entire set of JSON data was written line by line in a .txt file called tweet_json.txt. The .txt file was then read line by line using pd.read_JSON.

### 1.2 Assessment

- The Dataframes were first assessed visually using MS Excel.
- The Dataframes were then assessed programmatically using the info(), Describe(), Query() among others.

A summary of the findings from the assessment of the three datasets are as follows;

#### 1.2.1 Quality issues

1. 181 non-null values in the retweet_id column indicating retweets on the tweet_archive table.
2. Timestamp is a string on the tweet_archive table.
3. 78 non-null values in the in_reply_to_status_id column indicating a reply tweets on the tweet_archive table.
4. Some ratings Denominators are invalid e.g 170 on the tweet_archive table.
5. Some ratings numerators are invalid eg 1,776 on the tweet_archive table.
6. non descript column name on the tweet IDs table.
7. 324 predictions that aren't dog breeds on the image prediction table.
8. Non descript column names on the image prediction table.
9. some tweets were beyond Aug. 1st 2017 on the twitter archove Table.

#### 1.2.2 Tidiness issues

1. The dog stage being stored with the texts in the text column of the twitter enhanced table.
2. Ratings is a single variable but it was spread into two columns on the twitter enhanced Table.
3. 29 columns on the tweets IDs table are not needed because some are repeated columns or they donot help the current analysis in any way.
4. 13 columns not neccesary on the archive table(including retweet counts because we wouldnt be making use of retweets for this analysis)
5. three predictions with different confidence levels on the image prediction table.
6. Drop unneccesary columns on the image prediction table.
7. only one table is neccesary.

### 1.3 Cleaning

Before cleaning the datframes, a copy of all the original data was made. After this, all of the Quality and Tidiness isuess identified in the three dataframes were then cleaned sequentially in the following steps:

**Quality issues**

- Dropped all retweets using the retweet status id column from the twitter archive table.
- Changed timestamp to datetime format on the twitter archive table.
- Dropped all replies using the reply status id column from the twitter archive table.
- Dropped all rows with rating denominators with 3 digits (>99)on the twitter archive table.
- Dropped all rows with rating Numerators with 3 digits (>99) on the twitter archive table.
- The favorite_count column was renamed to like_count and ID changed to tweets ID on the tweets id table.

- Removed all predictions without atleast one dog breed prediction on the image prediction table.
- Non descript column names on the image prediction table were cleaned by changing all "p" to "prediction".
- Dropped all tweets beyond Aug. 1st 2017.

**Tidiness issues**

- Extracted the dog stage from the text column into a new "stage" column on the twitter_enhanced table.
- Divided the numerator rating by the denominator rating and stored it in a "ratings" column on the twitter archive table.
- Dropped all columns except tweet_ids, like_count and retweet_count on the tweets id table.
- Dropped all coumns except; tweet_id, timestamp, text, name, stage, ratings twitter archive table.
- Applied a function that select the breed prediction with the highest confidence level among all three predictions and dropped the rest on the image prediction table.
- Dropped all other columns apart from tweet id and highest conf breed on the image prediction table.
- All three tables were merged using the tweets id column.

**The merged dataframe was then stored in a CSV file named twitter_archive_Master.csv**

```
In [ ]:
```