# wrangle_report

June 27, 2022

## 0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

## 0.2 PROJECT 2: TWITTER DATA WRANGLING

**Introduction:** The project is to introduce the learner to the real world of messy data as data is real world is rarely clean. The analyst is usually needed to clean quality and tidiness issues of the data - a process known as **data wrangling** before it can be analysed. In this project, the real world data used are gotten from various sources in different formats as provided by a twitter user (@dog_rates) also known as WeRateDogs. The account rates people's dogs with humorous comments about the dogs with the ratings almost always having a denominator of 10 while the numerator is almost always greater than 10. The timestamp of the archive provided by WeRate-Dogs did not go beyond August 1, 2017.

**Project Motivation:** The motivation behind the project is to wrangle the datasets effectively and produce beautiful, trustworthy and meaningful insights and visualizations.

**Data** There are three datasets to be sourced and used. These Datasets are:

1. ***Enhanced Twitter Archive***: The file provided by *WeRateDogs* contains *2356 rows and 17 columns* of tweets about rated dogs. The rows show the number of tweets while the columns show other information about the tweets. The information include but not limited to reply, retweet, dog name, dog ratings, dog stages etc.

2. ***Image Prediction File***: This file is as a result of neural network developed by some students in Udacity and ran against the enhanced twitter archive provided by *WeRateDogs* to classify breeds of dog. The file contains *2075 rows and 12 columns* The rows show the number of predictions made while the columns show information about the prediction. The information include but not limited to image url, image prediction confidence, tweet ID etc.

3. ***Twitter Json File***: This file was scrapped from Twitter using the Tweepy API to extract the retweet and favourite counts about each tweet presented in *Enhanced Twitter Archive*. The file contains 2354 rows and 3 columns.

**Project Steps Overview** The step followed in this project are:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Storing Data
5. Analyzing and visualizing Data
6. Reporting

The tools used for this project are:

1. Jupyter Notebooks
2. Pandas Library
3. Numpy Library
4. Matplotlib Library
5. Request Library
6. OS Library
7. Seaborn Library
8. Json Library

**Gathering Data:**

1. **Twitter_Enhanced_Archive.csv**: This data was provided by Udacity and was manually downloaded and uploaded for use in the notebook. The uploaded file was read into a dataframe using the ***pd.read_csv('file_name')*** of the Pandas Library
2. **image-prediction.tsv**: This data was hosted on Udacity server and programmatically downloaded for use in the notebook using ***get()*** function of the Request Library. After it was downloaded successfully, a response message of *200* was displayed and the ***with open()*** of the OS Library was used to save it into the notebook and **_pd.read_csv('file_name')__** of the Pandas Library was used to read it into a dataframe.
3. **twitter-json.txt**: This data was supposed to be scrapped from Twitter but I could not do this as I do not have the elevated access to the Twitter Developer account as at the time of doing this project. So, the already scrapped json file provided by Udacity was manually downloaded and uploaded into the notebook. The tweet_id, retweet_count and favourite_count were extracted from the file using ***with open()*** of the OS Library, ***for loop*** of Pandas Library and loaded the lines from the file as ***json***.

**Assessing Data** The data was assessed in two ways namely:

1. **Visual Assessment**: This was done by printing each dataframe and going through the datasets.
2. **Programmatic Assessment**: This was done using the methods from the Pandas Library. The methods used are: *head(), tail(), sample(), describe(), info(), isnull(), isduplicated(), shape, value_counts()*

**Cleaning Data** To clean the data, the '***Define-Code-Test*** methodology was used. This methodology was used to clean the quality and tidiness issues found in the datasets.

The following was done during the data cleansing stage:

1. The three original dataframes were duplicated with the copy cleaned.
2. I used the ***Define-Code-Test*** methodology to clean the following quality and tidiness issues.

**Quality issues**

1. The datatype of the column named '*timestamp*' in Twitter Archive Data (df1) is a '**string**' which is a wrong datatype for such column. The correct datatype should be '**datetime**'.
2. The datatype of the column named '*tweet_id*'in Twitter Archive Data (df1), Tweet Image Prediction Data (df2) and Twitter Json Data (df3) should be a **string** not an **integer** as seen.
3. There is are missing values in '*expanded_url*' column in Twitter Archive Data.
4. There are invalid values in the '*name*' column in Twitter Archive Data. The values in lowercase are '**articles**' not names of dogs
5. '**tweet_id**' with values in '**retweeted_status_id**', '**retweeted_status_user_id**', '**retweeted_status_timestamp**' in the Twitter Archive Data are not needed for this report
6. There are inconstitencies in the values in the columns '**p1**', '**p2**', '**p3**' in Tweet Image Prediction Data
7. The following columns: '**retweeted_status_id**', '**retweeted_status_user_id**', '**retweeted_status_timestamp**', '**in_reply_to_status_id**' and '**in_reply_to_status_user_id**' in the Twitter Archive Data consist of mainly missing values
8. The number of decimal places of the values in the columns '**p1_conf**', '**p2_conf**', '**p3_conf**' in Tweet Image Prediction Data es

##### Tidiness Issues 1. Four columns in the Twitter Archive Data are supposed to be combined into one column with the column names as values in the new column. The affected columns are: '*doggo*', '*floofer*', '*pupper*' and '*puppo*'

```
2. The three tables should be merged into on Dataset
```

**Storing Data** After succesfull executing the gathering, assessing and cleaning stages, the file was merged using the **pd.merge()** method of Pandas Library

**Analyzing and Visualization** In this section, the merged file was loaded into the notebook and six insights were drawn and two visualization were made using the Pandas and Seaborn Libraries respectively.

**Reporting** In this section, a report was written on the process using the notebooks provided. The notebook povided are: **wrangle_report.ipynb** and **act_report.ipynb**. The notebooks were then downloaded as ***html*** and ***pdf*** and uploaded into the workbook environment for submission.