# wrangle_report

July 19, 2024

## 1 Project: Wrangling Report

### 1.1 Table of Contents

**1.Gatharing Data**
**2.Assessing Data**
**3.Cleaning Data**
**4.Storing Data**

### 1.2 1 - Gatharing Data

Before I start in Gatharing Data I import the libraries, after that i start in gatharing data, my gatharing is seprated to three parts,

1- Download the WeRateDogs Twitter archive
2- Download the image predictions file
3- Additional Data from Provided Files OR Twitter API access

**1- Download the WeRateDogs Twitter archive**

This the easiest step in gatharing i only download the file from udacity classroom and read it in jupyter notebook using pandas,

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

**2- Download the image predictions file**

This file i download it using request libraiy by the link in the udacity classroom,

This file (image_predictions.tsv) is present in each tweet according to a neural network.

**3-Additional Data from Provided Files OR Twitter API access**

This step is the hardest in gatharing in this step we have two options Create a twitter developer account and access to API (This option is premium) the second option is download the file called `tweet_json.txt` using requests library, this file in the classroom, I choose to use the provided file in udacity classroom, after i download it using request library i convert it to data frame,

I gather each tweet's retweet count and favorite ("like") count from `tweet_json`

### 1.3 Assessing Data

1 - Visual assessment

2 - Programmatic assessment

**1- Visual assessment**

In visual assessment I look at the three data sets `twitter_archive`, `image_predictions` and `tweet_json`, in jupyter notebook and excel

**2- Programmatic assessment**

I separate the programmatic assessment to 5 steps,

1 - get summary of each data frame using `.info()`

2 - check for missing values

3 - check for duplicates

4 - get summary of statistics of each dataframe using `.descripe()`

5 - assessing and verify columns and rows, in this step i check for specific columns and rows to know more about theme

After all of this I start to detect the quality issues and tidiness issues,

I get about 8 quality issues and 3 tidiness issues

### 1.4 Cleaning Data

First i take a copy of each dataframe to clean theme, i start to clean the issues i found in assessing data step, In cleaning Data i use this method, Define, code and Test,

First Define: i define the issue and talk little about it,

Second Code: In this step i start to make a code clean this issue,

Third Test: In this step i start to test the code such as show the data after clean part from it

### 1.5 Storing Data

After all of the last steps i merge the dataframes in one dataframe and convert it to csv file called `twitter_archive_master.csv`

```
In [ ]:
```