# Wrangle Report

I create this report to describe all of the steps of wrangling data that I uses in my project: **Wrangle and Analyze Data**

In the begging, The first step that for every Wrangle is the Gathering Data in the Gathering Data I used the WeRateDogs Twitter archive, **WeRateDogs** is a Twitter account that rates people's dogs with a humorous comment about the dog. their twitter user is @dog_rates you could check it if you want.

The entire process and project was done in Udacity. and they managed and reviewed all my projects and this on one of them to make it's good and to finish my entire Nano degree.

So what is Wrangling?
We can say wrangling in 3 steps:
1-Gathering
2-Assesing
3-Cleaning

1. **Gathering Data**

The data I gathered in this project:
**A. Enhanced Twitter Archive**
**twitter_archive_enhanced.csv** and it's a Data frame with 2356 rows × 17 columns and it has every thing that you want to know about the tweet ( text, the time that the tweet tweeted, id,url,rating,name of the dog … etc).
and you could download it from above if you want to know more about it.

B. **The tweet image predictions**
image_predictions.tsv
**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**
This is the second Dataframe with 2075 rows × 12 columns
And it's about the full of images predictions such as ( tweet id , image url, image number …etx)
And you can download it from the above link to know more about it .

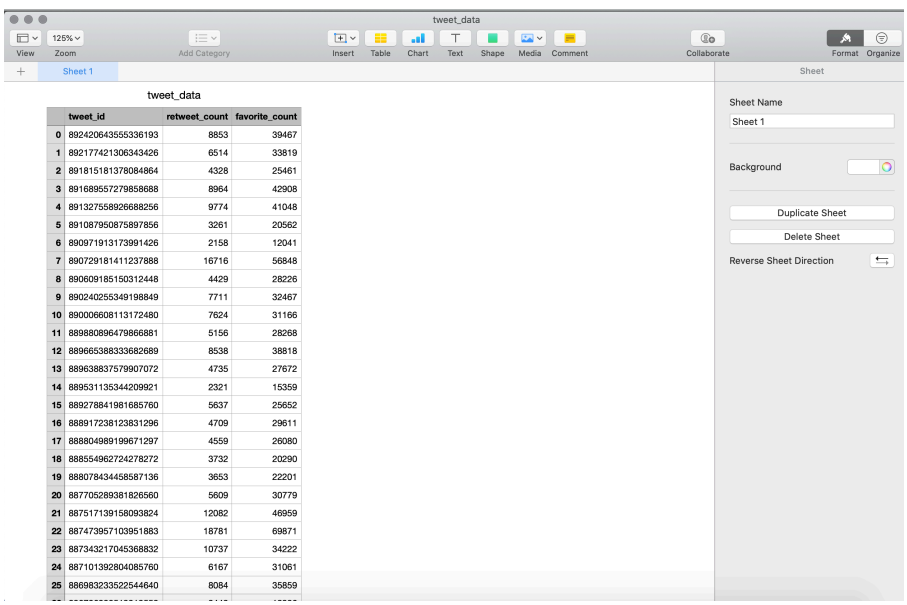
C. **Additional data from the Twitter API**

Here you have 2 way of gathering this file one of them is creating a twitter developer account and query the Twitter API for each tweet's JSON data using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called tweet_json.txt file
The other way is if don't want to create a twitter developer account the file is ready to use from udacity without querying and bring it from twitter dev account.
The Dataframe contains a ( tweet_id,retweet_count,favorive_count) and it's a 2354 rows and 3 columns.

## 2. Assessing Data

in this part I must Detect and document at least eight (8) quality issues and two (2) tidiness issues
in two ways:
**(Visual assessment, Programmatic assessment)**

I used to the visual **assessment** a mac "Number applications ":




tweet_data

| | tweet_id | retweet_count | favorite_count |
|---|---|---|---|
| 0 | 892420643555336193 | 8853 | 39467 |
| 1 | 892177421306343426 | 6514 | 33819 |
| 2 | 891815181378084864 | 4328 | 25461 |
| 3 | 891689557279858688 | 8964 | 42908 |
| 4 | 891327558926688256 | 9774 | 41048 |
| 5 | 891087950875897856 | 3261 | 20562 |
| 6 | 890971913173991426 | 2158 | 12041 |
| 7 | 890729181411237888 | 16716 | 56848 |
| 8 | 890609185150312448 | 4429 | 28226 |
| 9 | 890240255349198849 | 7711 | 32467 |
| 10 | 890006608113172480 | 7624 | 31166 |
| 11 | 889880896479866881 | 5156 | 28268 |
| 12 | 889665388333682689 | 8538 | 38818 |
| 13 | 889638837579907072 | 4735 | 27672 |
| 14 | 889531135344209921 | 2321 | 15359 |
| 15 | 889278841981685760 | 5637 | 25652 |
| 16 | 888917238123831296 | 4709 | 29611 |
| 17 | 888804989199671297 | 4559 | 26080 |
| 18 | 888554962724278272 | 3732 | 20290 |
| 19 | 888078434458587136 | 3653 | 22201 |
| 20 | 887705289381826560 | 5609 | 30779 |
| 21 | 887517139158093824 | 12082 | 46959 |
| 22 | 887473957103951883 | 18781 | 69871 |
| 23 | 887343217045368832 | 10737 | 34222 |
| 24 | 887101392804085760 | 6167 | 31061 |
| 25 | 886983233522544640 | 8084 | 35859 |

And for the **programmatic assessment** I used jupyter notebook (python libraries) and queries to find every issues that I could find and these is what I found:

## Quality

### `twitter_archive` *table issues*

- tweet_id format from int to string
- source format (<.a> href) it's for another language (html) we don't need it so after removing it we can make it categorial
- some dogs name are (missing , wrong)
- invaild datimestamp datatype is a string not a object
- four columns (doggo, floofer, pupper and puppo) have 'None' for missing
- we have in some of the dogs text the rating and link

### `Image_predictions` *table issues*

- we have to change tweet_id type to string so we can combine the dataframes
- difference in the upper and lower case in P names
- we have some columns that we will not use

### `tweet_data` *table issues*

- we have to change tweet_id type to string so we can combine the dataframes

## Tidiness

- `twitter_archive table` four columns (doggo, floofer, pupper and puppo) must be in 1 column .
- all datasets should be 1 dataset.

## 3. Cleaning data

I cleaned the previous issues for the quality and Tidiness and save it to tidy masted pandas DataFrame.