# Wrangle Report

# Gathering

Following instruction from Udacity project detail page first I downloaded Twitter archive file 'twitter_archive_enhanced.csv' which was provided for download, second file 'image_predictions.tsv' download programmatically from Udacity server using request library and third file which casue me all sort of problem as it was tweet_json.txt file using twitter API for @WeRateDogs as I requested for permission from twitter for almost a week and in end contacted mentor and it was provided me by email so I read this file line by line to create data frame for at least three columns id( tweet) , favorite and retweet counts.

After have three dataframe(df) I make copies of them so if I make any changes it won't effect original ones. I called  copied df_clean, image_clean and tweet_clean.

# Assessing

Asses df visually and programmatically and found lots of Quality and Tidiness issue but done minimum requirement for project 8 quality and 2 tidiness issue as below for cleaning

Quality Issues

It mainly include issues like completeness, validity, accuracyand consistency

***df_clean***

1. remove tweet that has been retweet as its not original.

2. combining dog stages to one column

3. remove columns that are not needed for analysis.

4. Change timestamp from string to date time and make separate columns for date and time.

***image_clean***

5. p1,p2 and p3 have inconsisitent capital words

6. drop duplicate jpg_url.

7. p1,p2 and p3 have unnessary underscore instead of space.

***tweet_clean***

8. rename id to tweet_id so can merge later

Tidiness

Happy families are all alike; every unhappy family is unhappy in its own way — Leo Tolstoy.

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In tidy data: 1. Each variable forms a column. 2. Each observation forms a row. 3. Each type of observational unit forms a table.

***Tidiness issues***

1. change tweet_id from number to string.

2. Newly created Date and time column needed to change from object(string) to date time format.

3. perform inner join between three data frame as they all have data for same tweet.

# Cleaning

Used basic python function like duplicates , drop, sort , value_count ,describe , info and others to comply with above mentioned point. I struggle with few issues and had to spend a lot of time to get my understand. As little help was provided its first time I used so many websites for checking syntax and possible solutions.

# Conclusion

I think only after learning thoroughly from Udacity platform I have started to grasp coding mindset but as I am changing career so still a lot more to learn.