**Wrangle Report**

The main goal of this project is to master the three main wrangling skills (gathering, assessing and cleaning). In addition to cementing, the knowledge acquired through the courses. The material worked with a is a twitter account by the name of WeRateDogs which a twitter account that rates peoples dogs in a humors comment and assign a rating out of ten.

**GATHERING THE DATA:**

The data for this project came from three separates sources: The Twitter Archive File, The tweet image predictions and Twitter API (using tweepy) with the goal to finish with one master dataset that is ready to manipulate and draw insights from.

1. **Twitter archive file**: the twitter\_archive\_enhanced.csv is to be perceived as file that was handed out in a real life work situation. There for it was downloaded and used local from the machine. The file contains an archive the tweets information
2. **The tweet image predictions:** hosted on Udacity server, aiming to practice either downloading the file programmatically or using the URL in code to create a data frame. File contains a prediction of the dog breed along with a confidence interval of that.
3. **Twitter API:** query the Twitter API to collect additional data based on the tweet ID. And storing the data locally under a Json text file to be converted into a pandas dataframe.

**ASSESSING THE DATA**

Assessing the dataframes visually and programmatically

* Using google sheets to visually inspected data points as pandas can be limiting to view all columns and scrolling can be a bit challenging.
* Programmatically using Pandas built in methods(info, values\_counts ,describe ,head/tail ,groupby )

**CLEANING THE DATA**

In this part, we addressed the issues found in while we were assessing the date:

1. *Tidiness*

* gather all the data points in one dataframe
* retweet data related columns are not needed and has a tremendous amout of missing data and needs to be dropped

1. *Quality*

* Timestamp data type needs to be corrected
* rating denominator != 10 for certain rows
* ranting nominator outliers (>10)
* inconsistent caps in the names of the for the predicition columns p1,p2 and p3
* missing data under name column represented as None instead of Nan
* dog names columns caps are inconsistent
* source column is not easily readable
* rating denominator and nominator date type needs to be changed from int to float to accept decimals
* converting the tweet\_id in the tweet\_json df to float

Several pandas methods were used (i.e. drop\_duplicates(), info(), .drop(), .astype(), .loc(), .sample() .str.title(), str.extract etc.) as well as for loops and functions were used to address the issues.