# WRANGLE AND ANALYZE DATA

## Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The goal is gathering data from a variety of sources and formats like csv,tsv, assessing its quality and tidiness, then cleaning it. We eventually get to showcaseour wrangling efforts through analyses and visualizations.

The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have as a denominator of 10.

Steps that I have followed in this project goes into 3 main steps

1. Gathering data
2. Assessing data
3. Cleaning data

## Gathering data

The data that have been gathered came from 3 sources :

**Enhanced Twitter Archive**
The twitter_archive_enhanced.csv was provided by
Udacity and downloaded manually. That contains basic tweets data.

**Twitter API & JSON**

This file contains retweet count and favorite count are two of the notable column omissions. Unfortunately, this data should be gathered from twitter API but they did not approved my developer account. So I have decided to use the supporting file from Udacity page.

**Image Predictions File**
What breed of is present in each tweet according to a neural network. This file is hosted on Udacity's servers and was downloaded programmatically using the requests library and URL information.

## Assessing data

This step allows us to identify quality and tidiness issues. Once the three tables were obtained I assessed the data as following:

• Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.

• Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Then I separated the issues that I have faced in quality issues and tidiness issues. Key points to keep in mind for this process was that original ratings with images were wanted.

## Cleaning data

Here we refined our data. The data have been cleaned and wrangled by the programmatic method. With this method I have defined the function or cleaning task. Then, we code the issue to get it cleaned (drop, extract, is lower, etc., methods). And the end we tested the dataset to assure that the cleaning operations work correctly.

## Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with. Through the data wrangling and analysis, we used many libraries such as pandas, NumPy, requests and JSON, which allow us to gather, assess, and clean the data.
Finally, I outputted everything together.