# Wrangle Report

## Data Analyst Nanodegree

by **Mohammed Barakh**

## Introduction:

The aim of this project is to apply the knowledge gained from the data wrangling segment of the Udacity Data Analysis Nanodegree program. The dataset subjected to wrangling pertains to a collection of tweets from the Twitter account @dog_rates, also recognized as WeRateDogs. WeRateDogs is a Twitter account renowned for assessing individuals' dogs while adding witty annotations about each dog.

Typically, these assessments involve a denominator of 10. This report provides a succinct overview of the data wrangling endeavors undertaken to bring the project to fruition.

## Data Wrangling steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

## Project step by step

### A. First Gathering Data:

To Gather The project data, we need fist to get 3 dataset from three different ways:-

1. Twitter archive: the ‘twitter_archive_enhanced.csv’ is a dataset provided by Udacity to be downloaded manually
2. Twitter Image Prediction: the file ‘image_predictions.tsv’ is downloaded from Udacity serve using Requests library and URL
3. Twitter API and JSON file: the file ‘tweet-json.txt’ we get it by querying Twitter's API to gather this valuable data.

### B. Assessing Data:

Once the three data sets were obtained. I started to assess the data as following:

- Visually by using methods such as ‘sample’ or ‘head’ and ‘tail’
- Programmatically by using different pandas methods such as (info, describe, value_counts, duplicated… etc)

Then I write down everything that seems off or wrong in the data or needs to be edited, by classifying them into two different problems, the first is quality issues and the second is tidiness issues.

## C. Cleaning data:

The first step is to take a copy of every data set and name them at the end '_clean' to avoid any damage to the original datasets. After that I was ready to start cleaning the dataset.

First let's talk about Quality issues:

a. Twitter_archive:
    i. retweeted_status_timestamp, timestamp needs to be datetime
    ii. dogtionary should be date time
    iii. all IDs - object
    iv. in several columns null object are non-null
    v. rating_denominator and rating_numerator have invalid values
    vi. incorrect names or missing names in name column & 'None' Values
b. Image_prediction
    i. IDs should be strings
c. tweet_json
    i. IDs should be strings
    ii. display_text_range should be single character count

Second, Tidiness issues:

a. Twitter_archive:
    i. separate replies and retweets into different tables
    ii. doggo, floofer pupper and puppo should be a single variable
b. Image prediction:
    i. p1, p2 and p3 should be categorical datatype
    ii. p1_conf, p2_conf and p3_conf columns should be merged, the same for p1_dog, p2_dog and p3_dog
c. Tweet_json
    i. Drop all columns in common with archive