# Wrangle Report

By Sultan Alharbi

## Introduction:

In this report we will go through the wrangling process of WeRateDogs data, its part of Udacity Data Analyst Nanodegree program.

The wrangling process will include the following:

- Data Gathering
- Data Assessment
- Data Cleaning

## Data Gathering:

The data we have used was gathered from three different sources, and three different gathering methods.

**Enhanced Twitter Archive Dataset:**

This dataset was given to us as a file on hand, and it was a comma-separated values file, aka CSV file.

**Image Prediction Dataset:**

This dataset was extracted using Requests library in Python, we used Requests to retrieve the data from Udacity's servers and the data were in TSV file format.

**Tweets Dataset:**

This dataset was retrieved from Twitter's API using Tweepy library, The data were in JSON format.

## Data Assessment:

After gathering the data, we must assess it programmatically and visually to find any tidiness or quality issues, and that's what we will go through in this section.

**Quality issues:**

- The dataset includes retweets which may cause a duplication in the data.
- Dropping in_reply_to_status_id and in_reply_to_user_id columns.
- Correcting data type in tweet_id (from int into string).
- Correcting data type in timestamp (from string into datetime).
- Creating a column for dog image prediction and another column for dog image prediction confidence.
- Removing p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog, img_num columns.
- Some dogs have a wrong names like 'a' , 'by' ,'an' , etc.
- Converting underscore to space and convert lowercase to uppercase in pred_dog.

**Tidiness issues:**

- The last two columns in TwitterArchive datasets should be merged into one column.
- The three dataframes should be merged into one dataframe.

## Data Cleaning:

After the assessing the issues within the data, we have to fix it using libraries like Pandas and NumPy, and that's what we have done in this part using Define, code and test technique.