# DAND Wrangle And analyze Data Project
# Wrangling Report

**By Naif Abdulrazaq Alsofyani**

## Contents

# Introduction

This project is aimed at wrangling and analyzing data from WeRateDogs Twitter account to apply the skills we gained from the Data Analyst Nanodegree so far. Data wrangling steps are:

- Gathering Data
- Assessing Data
- Cleaning Data

I will be going through each in this report.

# Gathering Data

In this project, we had Three data sources

1- WeRateDogs twitter archive (downloaded manually from Udacity).
2- The tweet image prediction built on neural network model ( downloaded using Requests library from Udacity's server).
3- Downloaded Tweets from WeRateDogs account using tweet IDs, by querying Twitter API using Tweepy library and stored the data on a JSON file.

# Assessing Data

The following Quality and Data issues was found and fixed.

## Quality Issues

- Change `Timestamp` from string to Datetime format.
- Change `tweet_id` from integer to string format in all datasets.
- `Source` contains HTML (a href) tag, should be removed, and contain twitter agent only.
- There are dogs named "a" in the data set and should be replaced to None and then changed to NaN.
- Drop all NaN rows in `name` column

- Drop unnecessary `in_reply_to_status_id`,`in_reply_to_user_id`,`retweeted_status_id`,`retweeted_status_user_id`, `retweeted_status_timestamp` columns
- `Source` data type should be changed to category.
- There are 59 null rows in `expanded_url` and should be dropped.
- Drop the 66 duplicated `jpg_url`

### Tidiness Issues

- Create `rating_ratio` column by dividing `rating_numerator` over `rating_denominator` which makes more sense
- Create column named `type` for each dog type on the Dataset instead of each type having its own column
- - Create a master DataFrame that merges all the DataFrames together

## Cleaning Data

All data quality and tidiness issues were cleaned and fixed programmatically using pandas' methods

## Storing Data

After cleaning the three DataFrames, a master DataFrame was created to merge all three datasets into one to visualize and analyze.