

# Wrangle Report

This report briefly describes the data wrangling exerted in this project.

The dataset that was wrangled (and analyzed and visualized) is the tweet archive of Twitter user @dog rates. This is a Twitter account that rates people's dogs from all over the world. These ratings denominator 10. But all dogs in the account have rated above 10 the reason of that based on the account's owner, is all dogs are good and deserve to be more than 10.

The Wrangling process is divided into three steps:

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

## Gathering Data

I gathered data from three different sources. The first data was WeRateDogs Twitter archive contains basic tweet data, it available CSV file from Udacity, I imported it using pandas.

The second one was additional tweets data it contains Retweets counts and Favorite counts. I took the available data and did for loop and put it in a dictionary after that, I append it in a data frame with name of tweet.

Last one was image predictions, that data tell us if the predictable percentage of the photo whether it is dog or not. It was imported using request library, then I converted it to CSV file.

# **Assessing Data**

## **Tidiness issues**

- 1- All data sets can be structured in one data set
- 2- dogs breeds should be in one column

## **Quality issues**

- 1- Drop unnecessary columns.
- 2- fill the other dogs stages with the name Multible
- 3- drop pictures which are not dogs
- 4- The data type of tweet\_id should be object.
- 5- Change the data type of timestamp
- 6- Missing values in Twitter API
- 7- based on the instructions retweets should be dropped
- 8- Drop rows with missing photos

## **Cleaning Data**

All quality issues and tidiness issues above were solved and cleaned.