# Data Wrangling Project Report

**Bin Mohaya, Turki**

**June 9, 2019**

## INTRODUCTION

This report aims to discuss the process of data wrangling and how it was applied on the WeRateDogs project. Data wrangling process consists of three main stages; gathering is the first and most important stage where different files and data types are collected either from a database or from the internet. The second stage in the process of wrangling is assessing. At this particular stage all gathered data are investigated to detect errors and typos. After done with exploring and detecting problems in the data, we began the process of cleaning which is the last stage in the wrangling process. Then, analyzing the data set can be much easier and smoother.

## WeRateDogs WRANGLING

In this project we have gathered three files in different ways. The first file has been downloaded manually and is called tweet_archive and contain all tweets data for the account WeRateDog such as time, ID number, URL, and dog name. The second file is called image_predictions and has been downloaded through the requests library in Python 3. The last file is a txt file that has been manually downloaded and treated as a json file to extract extra tweets information.

At the second stage, assessing data, we mainly assessed programmatically using pandas commands such as df.head(), df.info(), df.describe(), and df.deplicated(). These commands were applied to all three files to detect quality and tidiness problems. Quality problems are those issues related to completeness, validity, accuracy, and consistency. The detected quality issues are:

1- In tweet_archive, some names are written as 'a'.

2- In tweet_archive, five columns mostly have NaN values.

3- In tweet_archive, the last four columns variables should be boolean type; either 0 or 1.

4- In twitter_archive, some rating denominators exceed 10.

5- In image_predictions, all predictions columns (p1, p2, p3) should start with capital letter.

6- In image_predictions, jpg_url column is useless for the analysis.

7- In image_predictions, p1_conf that is less than 0.5 should not be considered for the analysis.

8- In image_predictions, the first predictions will be used and the other two will be dropped.

Where are tidiness problems are documented as following:

1- In tweet_archive the last four columns should be one column since they represent one variable.

2- All three tables should be merged on the tweet_id column.

The third stage in the wrangling process is related to cleaning the above documented issues programmatically using pandas, numpy, and OS libraries. After cleaning each file separately, they all were merged as a one table based on tweet_id column and then the assessing stage comes again for the whole data set as a second iteration. Also, another cleaning process is done on the final master data set to remove any row with missing values.