# Wrangle Report

**Introduction**

The aim of the project is to wrangle, analyze and visualize a dataset. The dataset, which is wrangled in this project, is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account, which rates people's dogs with a humorous comment about the dog. A brief description of the wrangling effort is depicted here in this report.

**Project details**

The python libraries used in this project were: pandas, numpy, requests, tweepy, json, matplotlib and seaborn

The methodology of the project can be divided into three steps:

- Gathering data
- Assessing data
- Cleaning data

*Gathering data*

The data for this project was gathered from three different sources

1. The WeRateDogs Twitter archive: The twitter_archive_enhances.csv file was provided by Udacity and downloaded manually.
2. The tweet image predictions: The file image_predictions.csv contains the information about what breed of dog is present in which tweet according to a neural network. It was hosted on Udacity's server and was downloaded programmatically using the Requests library from the URL provided by Udacity.
3. Twitter API and JSON: The Twitter API for each tweet's JSON data was queried using Python's Tweepy library and store each tweet's entire set of JSON data in the tweet_json.txt file. Then this tweet_json.txt file was then read line by line into pandas DataFrame.

*Assessing data*

The dataset was assessed both visually and programmatically.

1. Visual assessment: The twitter_archive_enhances.csv and image_predictions.tsv files were assessed visually by opening them in Excel. The tweet_json.txt file was opened with Jupyter Notebook to assess visually.
2. Programmatic assessment: The datasets were assessed programmatically using different functions like sample, info, describe, value_counts etc.

A number of issues with the three dataframes were observed in the assessment process like: the names of the dogs are not always actual names, wrong datatype, missing data, ratings were incorrectly reported, the dog stage variables can be reported in one column instead of four etc. The issues observed while assessments of the datasets were divided into quality issues and tidiness issues.

*Cleaning data*

Before the beginning of the cleaning process, copies were created for the three datasets. The cleaning part of the data wrangling process was divided into three parts: Define, Code and Test. In the cleaning process, we got rid of all the retweet entries as only original ratings were supposed to be counted in the analysis. The three dataframes were merged into a single dataframe and the unnecessary columns were deleted.

**Storing data**

The cleaned dataset was stored in a new file called twitter_archive_master.csv for analyzing and visualization.