# WRANGLE REPORT

## A Report on the wrangling process of a dataset acquired from Twitter

**The dataset wrangled, analyzed, and visualized is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.**

**WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog. The account was started in 2015 by college student, Matt Nelson. WeRateDogs asks people to send photos of their dogs, then tweets selected photos rating and a humorous comment. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "13/10". This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.**

**WeRateDogs has over 4 million followers and has received international media coverage.**

**As a student of Udacity's Nanodegree Data Analysis programme, I was asked to carry out an analysis on the twitter archive of WeRateDogs, downloaded from twitter. WeRateDogs** downloaded their Twitter archive **and sent it to Udacity via email exclusively for my use in this project, as a student of Udacity's Nanodegree Programme.**

**To wrangle the data, The following processes were carried out.**

**First, I had to install and import the various python packages and libraries that I would need to carry out the wrangling and analysis of the data. I installed and imported the following:**

- **Pandas**
- **Numpy**
- **Matplotlib.pyplot and its magic word needed for data visualization**
- **Json**
- **Requests needed to download data programmatically.**

**Then, I had to gather the data, which was made up of three different data sets.**

1. **The enhanced twitter archive, a Comma Separated Values(CSV) file, which was downloaded manually, from Udacity, using Python's Pandas library, and read into a dataframe. The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything.**

2. **The Image prediction data, a Tab Separated Values(TSV) file, which was downloaded programmatically, using Pandas Request library, uploaded into the jupyter notebook or Ipnb file meant for this analysis, and read into a pandas dataframe. The Url used to access the file programmatically was provided by Udacity.**

3. **The Tweet Json file, gathered from Twitter API using Tweepy and read line by line into a pandas dataframe. It contained the Retweet count, and Favourite count for every tweet made. I could not gather it myself, because I could not open a Twitter developer's account due to time constraints. I made use of the already downloaded file provided by Udacity, as an alternative.**

   **Next, I had to clean each data frame. In order to do that, I had to create a copy of each dataset programmatically, so that I would always have access to both the raw data and the cleaned data if need be.**

**I had to assess them visually at first, then, programmatically,using the .info() functions and other functions in the Pandas dataframe, and note the areas where cleaning was required in both quality and tidiness, such as removing unwanted columns(like the retweet rows that were not needed), fixing wrong data types, removing duplicates, etc.**

**Then, I carried out the cleaning programmatically, fixing the issues noted. I ensured that the cleaning process was finalized with merging the datasets, to make the analysis easier to carry out.**

**Then, I carried out a test to prove if the cleaning process was successful. The test proved that it was successful. This brought an end to the wrangling process of WeRateDogs dataset.**