# Introduction

The purpose of the Project #4 in the DAND was to put in practice the data wrangling skills I learned in the previous sections. The used dataset is the tweet archive from the Twitter user @dog_rates, also known as WeRateDogs. Purpose of this twitter account is to rate people's dogs with a funny comment. These ratings almost always have a denominator of 10.

The project consist of the following steps, according to the steps learned in the data wrangling course:

1. Gathering data
2. Assessing data
3. Cleaning data
    a. Define
    b. Code
    c. Test

## Gathering Data

To gather the data we needed to use three different methods: csv import, URL import and via API Import

- **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
- **The tweet image predictions**, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
- **Twitter API & JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, retweeted status and url.

## Assessing Data

Once the data was gathered, I began to assess the data on both quality and tidiness issues. I used visual as well as programmatic assessing technics in jupyter notebooks to find the following issues:

**Quality**

***twitter_archive***

- There are original ratings and retweets
- Delete columns that won't be used for analysis
- dog stages are not correct
- timestamp is not a datetime

- missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- Rating denominator is getting higher than 10
- Numerator as float only gets the value after period
- Result of the Rating should have a column in float
- Sometimes Name of the dog is None

***image_prediction***

- p1, p2, p3 dog races are sometimes capital letter sometimes small letter
- p1, p2 and p3 columns have invalid data...like a birdhouse, can_opener, or breastplate etc.

***tweet_json***

- tweet_id is a string not a int
- missing data for the tweet_ids
- Retweets in this df

**Tidiness**

***twitter_archive***

- dog stages are in 4 columns

***image_prediction***

- needs to be included into one big dataframe

***tweet_json***

- needs to be included into one big dataframe

## Cleaning data

Cleaning the data followed always the same process: Define, Code, Test. The following issues were cleaned to have a cleaner dataframe.

**Quality**

1. Keep original ratings (no retweets) that have images
2. Delete Columns in `tweet_archive` 'source','in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls'
3. Convert timestamp as date in `tweet_archive`
4. Convert `tweet_json` tweet_id into int
5. Fix Numerator in `tweet_archive`
6. Delete all rows in `tweet_archive`, which have a denominator which is not 10
7. Dog races all small letters in `image_predictions`
8. Delete 66 duplicated rows with picture in `image_predictions`
9. Delete Retweets in `tweet_json`

**Tidiness**

1. Dog Stages in `twitter_archive` needs to be one column
2. Creating one dataframe out of the three dfs called `df`