# wrangle_report

February 6, 2018

## 1 Objective

To have all WeRateDogs tweets data gathered and cleaned for analysis

## 2 Rundown

The whole wrangling process consists of three sections:

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

#### 2.0.1 Data Gathering

This section includes the steps to gather the data needed for the analysis. The goal is to have the data ready for data assessing afterwards. Three datasets need to be gathered:

1. WeRateDogs Twitter archive - Downloaded directly from the classroom as instructed
2. Tweet image predictions - Downloaded programmatically with the help of Requests
3. Tweet JSON - Downloaded programmatically with the help of Tweepy

To caveat, querying all of the tweet IDs in the WeRateDogs Twitter archive would take about 20-30 minutes, and Twitter does have a rate limit. Therefore, a few tunes on the parameters, `wait_on_rate_limit` and `wait_on_rate_limit_notify` in `tweepy.api` class would be nice. Also, a progress tracker like tqdm would come in handy.

#### 2.0.2 Data Assessing

This section includes the steps to assess data. The goal is to pinpoint and summarize the issues (Both quality and tidiness) that need to be tackled during data cleaning afterwards.

***Quality Issues***

Twitter Archive Enhanced data:

1. 181 observations of retweets are found
2. 59 observations of missing expanded url (NaN)
3. Inapproproiate datatypes for timestamp and dog stages
4. Erroneous dog names (a, an, the etc.)

5. Inappropriate string representation of missing dog names ('None')
6. Redundant column (Denominator), given the fact that the rating is on a out-of-10 scale

Image Predictions data:

1. Inappropriate datatype for dog breed (all p columns)
2. Letters inconsistency in which words sometimes start with lower case letter/ upper case

Tweet JSON data:

1. 172 observations of retweets are found
2. Unclear/ confusing column naming (eg. id)
3. Redundant columns are included (May depend on what sort of analyses need to be performed)

***Tidiness Issues***
Twitter Archive Enhanced data:

1. Dog stages as a single variable (Categorical) are represented by four separate columns

Image Predictions data:

1. Image URLs, predictions, and results should be part of the archive data

Tweet JSON data:

1. Retweet and favorite count should be part of the archive data
2. Text column duplicated in the archive data

Although this is not a super huge dataset, there are numbers of variables. It is highly recommended to assess the datasets programmatically with pandas functions like: 1. info() 2. head() 3. tail() 4. sample()

#### 2.0.3 Data Cleaning

This section includes the steps to clean the data. The goal is to create one well consolidated dataset for analysis afterwards. Issues can most be tackled by the pandas library.

1. Observations of retweets - Subset datasets to exclude unwanted observations
2. Missing expanded urls - Subset datasets to exclude unwanted observations
3. Inappropriate data types - pandas astype() and to_datetime() function
4. Erroneous dog names and inappropriate representations - pandas rename() function
5. Letters inconsistency - pandas str.lower() function
6. Confusing column name - pandas rename() function
7. Redundant columns - pandas drop()

After the quality issues are tackled, the three separate datasets can be merged by pandas merge() function. Some final refinements may be needed for the tidiness, but pandas functions like the ones above will mostly suffice.