**WeRateDogs – Twitter Data**

1. **Gather Data**

We download the **twitter-archive-enhanced.csv** manually, the **prediction.tsv** file programmatically and the **tweet\_josn.txt** using api.

Once we have all the 3 files, we create a dataframe for each of these files:

* 1. twitter\_df
  2. predict\_df
  3. api\_df

1. **Assessing the data**

We assess this data by going through each of the dataframes and looking for issues like missing values or incorrect datatypes, first visually and then programmatically.

We use many pandas functions to find the issues, like value\_counts(), duplicated() or info().

We find the main issues like:

### **Quality**

##### *twitter\_df table*

* The datatype of the tweet\_id column is integer and should be str
* The datatype of columns such as timestamp and retweeted\_status\_timestamp are incorrect
* The source column contains html code
* Some of the names of the dogs are incorrect.
* The columns which have missing values in doggo, floofer, pupper , puppo - has "None" instead of NaN
* Rating\_numerator and rating\_denominator have some inconsistent values

##### *predict\_df table*

* The datatype of the tweet\_id column is integer and should be str
* Contains retweets(multiple rows in column 'jpg\_url')
* Sometimes the pictures do not contain dogs at all
* The predictions are sometimes uppercase, sometimes lowercase with an ‘\_’ present I
* Also there is a "\_" instead of a whitespace in the predictions

##### *api\_df table*

* The datatype of the tweet\_id column is
* column is integer and should be str

### **Tidiness**

##### *twitter\_df table*

* The dog stages are in multiple columns and must be put into one column

##### *df\_predict table*

* The prediction must be reduced to one column

##### *all tables*

* All the tables share a common tweet\_id and must be merged together

## Data Cleaning

#### **Cleaning steps:**

**1.** Merge the tables together (Tidiness)  
**2.** Drop the replies, retweets and the corresponding columns (Quality)  
**3.** Drop the tweets without an image or with images which don't display doggos (Quality)  
**4.** Clean the datatypes of the columns (Quality)  
**5.** Extract the source from html code (Quality)  
**6.** Remove the incorrect names from the name column (Quality)  
**7.** Reduce the prediction columns into one column: breed (Tidiness)  
**8.** Clean the new breed column by replacing the "\_" with a whitespace and make them all lowercase (Quality)  
**9.** Convert Nones in all the 4 dog states to Nans (Quality)  
**10.** Merge the dog state columns 'doggo', 'floofer', 'pupper', 'puppo' into a single column (Tidiness)  
**11.** Clean the wrong numerators and denominators (Quality)