# Data Wrangling Report

# Overview about Dataset

The dataset that I used in this project is the tweet archive of Twitter user @dog\_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

# Data Gathering

* The WeRateDogs Twitter archive. I download twitter\_archive\_enhanced.csv file manually to the Jupiter and then create archive\_df by using pd.read\_csv.
* **The tweet image predictions. I download 'image-predictions’' programmatically using the Requests library and the following URL:**[https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad\_image-predictions/image-predictions.tsv](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv" \t "_blank)**, and then i use it to create 'image\_df' Data frame.**

# Assessing Data

## Type of assessment:

* Visual Assessment
* Programmarly Assessment

**Quality**: issues with content. Low quality data is also known as dirty data.

**Tidiness**: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:

* Each variable forms a column.
* Each observation forms a row.
* Each type of observational unit forms a table.

## Quality

* **There are some records in archive\_df which are retweets**
* **Tweet\_id \* there is some tweet\_id data in arrchive\_df not exist in image\_df**
* **Text column \* contains untruncated text instead of displayable text**
* **In\_reply\_to\_status\_id, in\_reply\_to\_user\_id columns and timestamp has wrong datatype**
* **Source unnecessary html tags in source column in place of utility name e.g., <ahref=""[http://twitter.com/download/iphone""](http://twitter.com/download/iphone%22%22" \t "_blank) rel=""nofollow"">Twitter for iPhone**
* **Names \* dog names starting with lowercase characters Define**
* **Expanded\_urls missing data in the column**
* **Rating\_denominator: some data is not equal 10**
* **Rating\_numerator: some data is less than 10**
* **Missing data from doggo, floofer, pupper and puppo columns**
* **Text: Contain link of tweet which also in expanded\_urls column Define**

## ****Tidiness:****

* **Archive\_df table without any duplicates (i.e., retweets) have empty retweeted\_status\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp columns**
* **Doggo, floofer, pupper and puppo columns in archive\_df table should be merged into one column named "stage"**
* **Timestamp columns have both time and date data**
* **Retweet\_count and favorite\_count columns from status\_df (tweet status) table should be joined with archive\_df table**
* **The names of columns p1, p2, p3, p1\_conf, p2\_conf, p3\_conf, p1\_dog, p2\_dog, p3\_dog does not describe it**
* **In column Image\_df Every algorithm has three columns**

# **Cleaning**

## Quality Problem

1. 1.there are some records in archive\_df which are retweets
   * + 1. Solution: Keep only tweet data instead of retweet data
2. 2.tweet\_id \* there is some Tweet\_id data in archive\_df not exist in image\_df
   * + 1. Solution: save only data which is in archive\_df and image\_df and drop anther data
3. 3.text column \* contains untruncated text instead of displayable text
   * + 1. Solution: remove untruncated text from text column
4. 4. In\_reply\_to\_status\_id, in\_reply\_to\_user\_id columns and timestamp: has wrong datatype
   * + 1. Solution:

In\_reply\_to\_status\_id, in\_reply\_to\_user\_id columns: convert from float to integer

* + - * 1. Timestamp: convert datatype from string to datetime

1. 5.source unnecessary html tags in source column in place of utility name e.g., <ahref=""http://twitter.com/download/iPhone"" rel=""nofollow"">Twitter for iPhone
   * + 1. Solution: Strip all html anchor tags (i.e., <a..>) in source column and retain just the text in between the tags. Convert the datatype from string to categorical.
2. 6.names dog names starting with lowercase characters
   * + 1. Solution: Convert lowercase characters to higher case
3. 7.expanded\_urls missing data in the column
   * + 1. Solution: Drop Null data from the Data
4. 8.rating\_denominator: some data is not equal 10
   * + 1. Solution:

For Data which has rating\_denominator is greater than 10 and divisible by 10, use the quotient as the divisor to divide the rating\_numerator. If the remainder=0, assign this quotient as the rating\_numerator.

For the rest of data, check if the text column contains any fraction whose denominator is 10. If it does, change the rating\_denominator to 10. Then, change the rating\_numerator with the value of numerator in it.

1. 9.rating\_numerator: some data is less than 10
   * + 1. Solution: For data in which rating numerator is less than or equal to 10 or greater than 10 but has a very high value, check if the text has any fraction whose denominator is 10. If it does, change the rating\_numerator with the numerator of it.
2. 10. missing data from doggo, floofer, pupper and puppo columns
   * + 1. Solution: Delete no data from doggo, floofer, pupper and puppo columns to repere them for tenderness cleaning Code
3. 11. Text: Contain link of tweet which also in expanded\_urls column
   * + 1. Solution: As link of tweet is in expanded\_urls column so we will remove it from text column

## Tendency Problem:

* 1. 1.archive\_df table without any duplicates (i.e., retweets) have empty retweeted\_status\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp columns
     + 1. Solution: Drop the retweet data retweeted\_status\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp columns
  2. 2.doggo, floofer, pupper and puppo columns in archive\_df table should be merged into one column named "stage"
     + 1. Solution: Merge the three columns in one column
  3. 3.timestamp columns have both time and date data
     + 1. Solution: Separate timestamp data in two columns time and Date.
  4. 4.retweet\_count and favorite\_count columns from status\_df (tweet status) table should be joined with archive\_df table
     + 1. Solution: Separate timestamp data in two columns time and Date.
  5. 5.The names of columns p1, p2, p3, p1\_conf, p2\_conf, p3\_conf, p1\_dog, p2\_dog,p3\_dog doesn’t describe it
     + 1. Solution: change columns names.
  6. Every algorithm has three columns
     + 1. Solution: Add algorithms in one column name breed, add the confidence of these algorithms in column breed\_confedance, and add the type of doges in Dog\_type column
  7. Merge breed column in archive\_cleaning
     + 1. Solution: Merge breed column, breed\_confedance, and Dog\_type in archive\_cleaning

# Store Data

Store data frame in csv files by using to\_csv

# Analyzing Data

* Check Datatype of all data which in both archive\_cleaning and image\_Copy tables
* Display dogs’ names which is popular use
* Check frequency of each stage for dogs
* Display Counts of each source for tweet
* Display Description of Rating\_numerator data
* Display Description of Retweet\_count, favorite\_count
* Find mean value of rating\_numerator, retweet\_count and favorite\_count according to stage
* Display how many times each algorithm was correct

# Visualizing Data

* Number of tweets per year and month
* The algorithm how is successful most
* Most frequent doges type
* Common source for tweets
* Common dogs’ names