# TITLE: DOCUMENTATION OF THE WRANGLING PROCESS

**INTRODUCTION**: In this documentation I will be using diagrams and words to explain all the wrangling processes, both the "Tidiness Issue and Quality Issues" **Data Wrangling:**

In the data wrangling process, I imported all the libraries that are necessary for the project to be successful.

I imported pandas using the pd alias, I imported matplotlib as plt, I imported NumPy as np, I imported json and also imported requests.

## Data Gathering:

in this section of data gathering, I gathered three different data using different methods.

- Twitter Archive (WeRateDogs Data)
- Image Predictions Data
- Additional data from Twitter

The twitter archive (weratedogs data) was gathered and imported into the notebook workspace using Pandas Library pd. read_csv('twitterarchiveenhanced.csv').

I also gathered the Image Predictions data by using the request library to pass in the html page of the page where the data is 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv'

I used with open ("image-predictions", mode='wb') as file: file.write(response.content)   to save the html to the directory and then I used pandas library to put the data in a data frame and read it in.

I intended using twitter API to extract the additional data but due to some technical issues I downloaded the available json file made available on Udacity work space and then I used a 'with open' function to read in the json file and wrote a for loop to extract the tweet id, retweet count and favorite count from the json file then created a data frame for the columns by using the append function and finally used the pandas  function  to put the columns in a data frame called 'df'.

## Assess:

Under the Assess section of my data analysis project, I treated at least 8 data quality issues and at least two data tidiness issues

## Quality Issues

**WeRateDogs Data**

- retweeted_status_timestamp is a string instead of datetime
- missing values in some features of the data set
- Incorrect ratings
- Non descriptive naming
- Dog names not well corrected
- Id fields should be in string

**Image Predictions data**

- Id fields not in string datatype
- Unstandardized Dog breeds
- some features are not well descriptive

In the course of trying to asses the data sets, discovered that the data sets have some quality and tidiness issues.

Some features seem to have missing values, some are not descriptive, some are not standardized, some have cases where the values are false, some have incorrect ratings etc.

I was able to do this proper assessment using both programmatic and visual assessment methods and even running summary statistics on all.

**WeRateDogs Dataset:**

This particular data set originally has 2356 observations and 17 features.

From the summary statistics ran on it, it is seen that 25% of 10.000000 are below rating numerator and denominator, 50% are also below 10.000000, 75% are also below 10.000000.

The minimum rating numerator is 0.000000 and the minimum rating denominator is 0.000000. The maximum rating denominator is 170.000000 and the maximum rating numerator is 176.000000.

**Image Predictions Data set:**

This data set has 2075 observations and 12 features.

It has no missing values nor duplicates unlike Weratedogs data set but some of its features are not well descriptive and some features have wrong data types.

**Additional Twitter Data set:**

This data set does not have too many issues, it has no missing values, no duplicates but one of its features “Tweet_id” had to be renamed so that it will be similar to just the same naming method used for the other two data sets though this is not so mandatory.

## Cleaning:

I cleaned every necessary issue in three data sets and also dealt with

- Tidy Issues
- Quality issues

In the cleaning process, I dealt with all the listed data quality issues and the tidiness issues.

**WeRateDogs Data**:

Here I filled the missing values appropriately using several imputation methods like median, mode, ffill etc.

I converted all necessary features to the right data types accordingly using “astype” function.

I renamed some features in order to make every feature more descriptive.

I created a data frame for only ratings whose denominator is not just 10 but based on manual fixing and reading.

I used a melt function to fix all the dog stages into one column so that we can have just a variable for all the values

I also created a rating column where rating = rating_numerator divided by rating_denominator and converted rating to float data type and then I dropped the rating_numerator and rating_denominator

I also dropped all cases where retweet_status_id is null.

**Image Predictions Data:**

Here I also renamed some features to make it more descriptive

I removed all cases where p1_dog is false

## Merge or Combine Dataset:

I used the merge function to join the three data sets on a common column called ‘tweet_id’ and arrived at a data which I named twitter_archive_master.

Finally, I stored the Combined data set in a csv file called “twitter_archive_master”