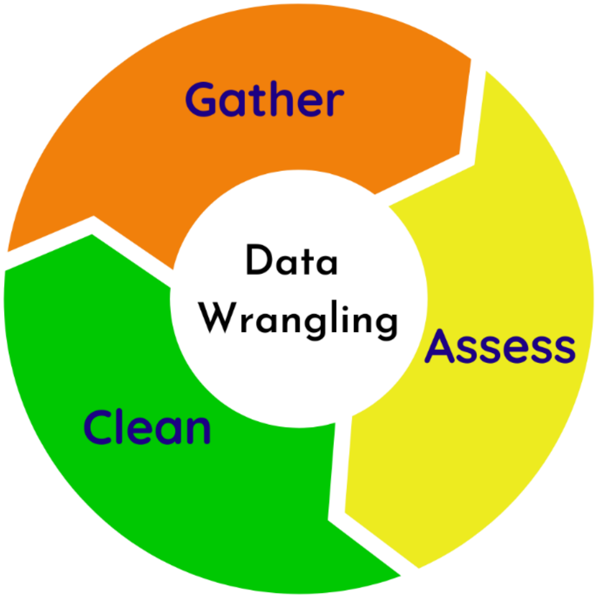
Data Wrangling

Reported by:

Mohammed Ezzat Yassin



A. Overview

The dataset to be wrangling 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.

WeRateDogs™
@dog_rates

This is Winnie. She lost her body saving a children's hospital from an avalanche. 13/10 what a h*ckin hero

There are three data sources:

1- Enhanced Twitter Archive:
The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything.

2- Additional Data via the Twitter API:
Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions.

3- Image Predictions File:
a neural network that can classify breeds of dogs. The results: a table full of image predictions (the top three only)

Data wrangling includes three parts:

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

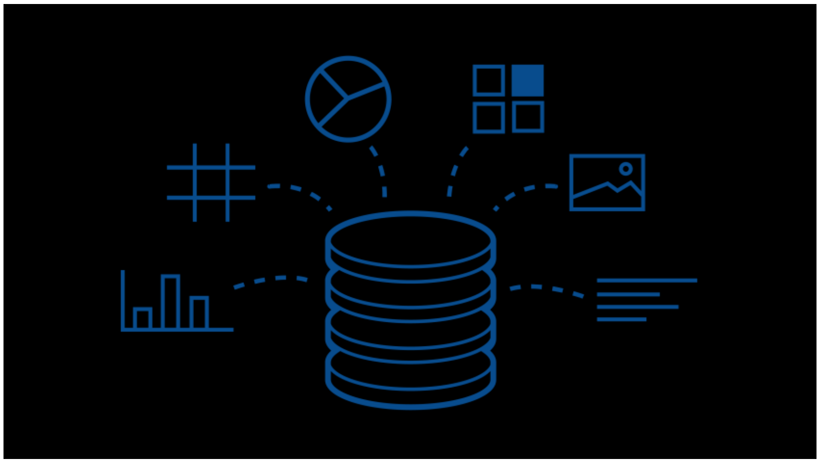
B. Gathering Data

The data have been gathered from the sources with three different methods:

1- Enhanced Twitter Archive:
The csv file was manually downloaded from the Udacity website and converted to a data-frame using the panda read method.

2- Additional Data via the Twitter API:
I found difficulties to create a developer Twitter account. So that the txt file was manually downloaded from the Udacity website and a code was used to read the file line by line and convert it to a data-frame.

3- Image Predictions File:
The tsv file was progmatically downloaded from the given url using Requests library and converted to a data-frame using the panda read method.



C. Assessing Data

The data assessed visually and programmatically taking into account the key points of the project to detect the following quality and tidiness errors:

Quality issues

For the twitter archive enhanced data

1- There are retweets corresponding to rows where the value of 'in_reply_to_status_id' column is non-null.
2- There are original tweets without pictures although the analysis is about dog's picture rating.
3- There are columns that will be not used in the analysis most of them belongs to the retweet case plus the expanded urls column.
4- There are a lot of wrong names.
5. The rating_denominator should be equal 10.
6- There are some observation with decimal ratings
7- tweet id is integer while timestamp is string.
8 - the rating smaller than 5 is illogical, this is a dog lovers society. So that these may be not pictures of dogs.

For the image prediction data
9- tweet id is integer and number of images is float.
10- There are duplicated values in the image url column.
11- The names of columns are confusing.
12- The number images values are confusing while using in visualization

For the additional data

13- The tweet id is integer.

Tidiness issues

For the twitter archive enhanced data

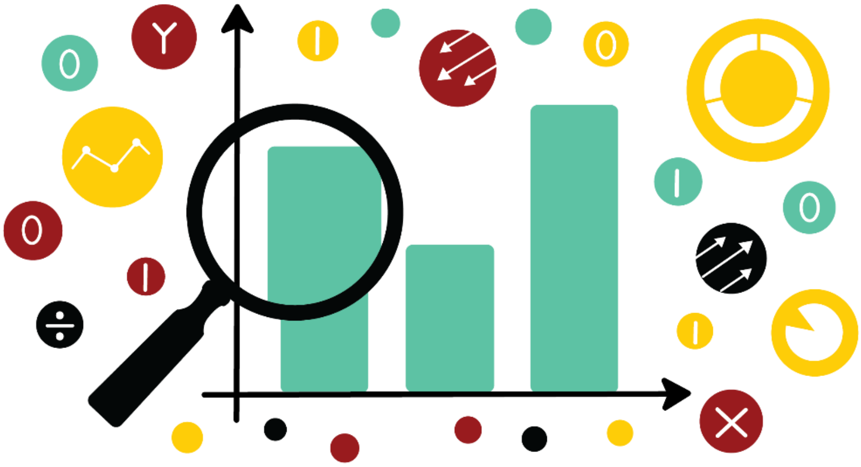
1- The two columns of rating_numerator and rating_denominator are representing the same rating variable.

2- The four columns of dog's stages are representing the same variable.

For image prediction and additional data

3- The six columns containg algorthm types and confidence levels which give predictions of dog type are so confused.

4- The additional data gives some extra attributes for the twitter archive data Also the image prediction contains extra information about the same tweets in twitter archive data.



D. Cleaning Data

The cleaning data has three steps:

1- The gathered data-frames was copied, so that if any errors occur in the cleaning process don’t affect the original data.

2- The quality and tidiness errors, one by one, was defined and cleaned using the suitable functions as shown in wrangl_act.ipynb file and finally tested to ensure that the cleaning process succeeded.

3- Finally, the cleaned data-frames had been merged and saved to a csv file.