# Data Wrangling Report

By: Ayodele Anuoluwa

**STEP 1 GATHERING OF DATA**

In this step, I gathered all the three pieces of data using different methods.

1. I manually downloaded the twitter\_archive\_enhanced.csv, uploaded it to my jupyter notebook and then read the data into twitter\_archive dataframe.
2. I programmatically downloaded the image\_predictions.tsv using the requests library with the URL provided and read the data into image\_predictions dataframe.
3. I gathered each tweet’s retweets count and favourite count using the tweets IDs in twitter\_archive dataframe by quering the Twitter API for each tweet's JSON data using Python's [Tweepy](http://www.tweepy.org/) library and store each tweet's entire set of JSON data in a file called tweet\_json.txt file. Each tweet's JSON data was written to its own line. Then I read this .txt file line by line into a tweets\_data DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

**STEP 2 ASSESSING AND CLEANING DATA**

After gathering all three pieces of data, I assess them visually and programmatically for quality and tidiness issues. Below are the results of the assessments.

|  |  |  |
| --- | --- | --- |
| S/N | QUALITY ISSUES | CLEANING PROCESS |
| 1 | It is required to remove retweets and replies | Remove all rows for which retweeted\_status\_id or in\_reply\_to\_status\_id are not null |
| 2 | Some columns holds very low amount of data such as in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id | Remove columns with low amount of data |
| 3 | Timestamp is in string (object) data type | Convert timestamp to be datetime |
| 4 | Source columns are not readable beacuse of the URL attached to it | Remove the URL and replace the source data with readable values e.g iPhone, Twitter, and TweetDeck in df3 dataframe |
| 5 | Some column names are ambigious and not meaningful such as timestamp, p1, p1\_conf, p1\_dog, p2, p2\_conf, p2\_dog, p3, p3\_conf and p3\_dog | Rename the column names for better readability |
| 6 | There are erroneous dog names starting with lowercase characters (e.g. a, an, actually, by). All the erroneous names are in lowercase. | All the erroneous are regarded as no name, thereby all lowercase value should be replaced with None |
| 7 | There are rating\_numerators that are not correct with the text associated with it. Some are way greater than 15 and some are outrightly wrong by picking another values in the text. | Correct the incorrect rating\_numerators values by dividing the numerators by the denominator in tens and also fix some rating\_numerators directly using the tex. |
| 8 | 10 is the default value of ‘rating\_denominator’, there are some denominators with incorrect value | Convert all the rating\_denominators to 10 and correct the rating\_numerator |
| 9 | The prediction dog breeds involve both uppercase and lowercase for the first letter. | Capitalize the First Letters for uniformity |

|  |  |  |
| --- | --- | --- |
| **S/N** | **TIDINESS ISSUES** | **CLEANING PROCESS** |
| 1 | The columns doggo, floofer, pupper, puppo are all referring to the dog stages. | The four columns are merged together into single column by extracting the texts from each column into the stage column. |
| 2 | img\_num contains values ranging from 1 to 4 but only 1 jpg\_url is present in the dataset. This columns do not have strong basis to be included in the datasets | Removed from the dataset as it is not relevant for any analysis. |
| 3 | There are three datasets that could be well tidy by merging into a single dataset. | Merge the datasets into a single dataframe. |

**STEP 3 – STORING DATA**

Data Wrangling Process carried out, although there could be further cleaning. I have stored the wrangled data in twitter\_archive\_master.csv file with a minored number of issues, and ready for a Data Analysis. This file has 1759 observations and 22 features.