@WeRateDogs Twitter Archive Project - Wrangle Report

By Manuel Cabrera

For this project we used three data sets:

* Enhanced Twitter Archive.
* Tweet Image Prediction.
* Retweets and Likes (Twitter API).

# Enhanced Twitter Archive – Gathering and Pre-Assessment

## Gathering

This data set was given to us by Udacity in the form of a comma-separated value file, to gather the data, it was necessary to simply use pandas *“read\_csv”*.

## Assessment

These were the findings during the assessment of the data set:

* Replies seem to be part of a different observational unit
* Source seems to have unnecessary characters and information
* Retweets seem to be part of a different observational unit
* Variables “doggo”, “floofer”, “pupper” and “puppo” seem to be values of a single variable, “Internet\_Nickname”.
* Tweet ID is categorized as an integer.
* “in\_reply\_to\_status\_id”, “in\_reply\_to\_user\_id”, “retweeted\_status\_id” and “retweeted\_status\_user\_id” are categorized as floats
* “Timestamp” and “retweeted\_status\_timestamp” are a combination of variables in a single variable.
* Rating denominator is not constant, hence the actual rating needs to be calculated.
* Several dog names are “such”, “the”, “this”, “unacceptable”…

# Tweet Image Predictions – Gathering and Pre-Assessment

## Gathering

This data set was generated by a team of Udacity data scientists however, to download the file, it was necessary to do so programmatically. For this, we used the *Requests* library and the URL: <https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv>

## Assessment

These were the findings during the assessment of the data set:

* Tweet ID is categorised as an integer.
* This dataset should be combined with the main tweets table.

# Retweets and Likes (Twitter API) – Gathering and Pre-Assessment

## Gathering

This data set was downloaded from Twitter’s API and to do so it was necessary to apply for a developer account and request several keys. Once approved by Twitter, we extracted the data as a text file and translated it into a dataframe.

## Assessment

These were the findings during the assessment of the data set:

* Tweet ID is categorised as an integer.
* This table should be combined with the main tweets table (and retweets).

# Final Assessment

Once the initial assessments were completed, then these were categorised according to the type of issue (quality or tidiness)

## Visual and Programmatic Assessment - Quality

* Tweet ID is categorised as an integer.
* “in\_reply\_to\_status\_id”, “in\_reply\_to\_user\_id”, “retweeted\_status\_id” and “retweeted\_status\_user\_id” are categorized as floats.
* Several dog names are not actual dog names.
* “Timestamp” and “retweeted\_status\_timestamp” are a combination of variables in a single variable.
* Source seems to have unnecessary characters and information.
* Rating Denominator is not constant.

## Visual and Programmatic Assessment - Tidiness

* Variables “doggo”, “floofer”, “pupper” and “puppo” seem to be values of a single variable, “Internet\_Nickname”.
* Replies seem to be part of a different observational unit
* Retweets seem to be part of a different observational unit
* RT\_L should be combined with the main tweets table (and retweets).

# Cleaning

* Once finalized, the issues were resolved as it follows:
* Variables Categorised as integers – these were converted into strings.
* Timestamp having multiple variables – these were split into day, months, year and time.
* Source column requires stripping – the main relevant information (source) was extracted
* "Internet\_Nickname" variable – The columns of “doggo”, “floofer”, “pupper” and “puppo” were removed and a single column called “Internet\_Nickname” was created in which the tweets were categorised according to the previously deleted columns.
* Invalid Dog Names – Observations with invalid dog names were replaced with None.
* Combine Tables – All of the tables were combined into a single big Dataframe.
* Calculating actual Rating – A new column was created based on the rating numerator and denominator.
* Separate Observational Units – The big dataframe was then split into three smaller tables, “tweets”, “retweets” and “replies”.