Introduction

This project describes our wrangling work that was made in the section of wrangling WeRateDog project as part of the Data Science Course by Udacity. The main objective of the project is to gather data from Twitter combining it with a third party data frame, creating a complete analysis about all the tweets and the predicted dogs breed.

I will split this document in the following parts:

Data Gathering

I've started the project by gathering the files twitter_archive_enhanced.csv and image_predictions.tsv using the requests package provided by Udacity. It is important to note that a python request library was used to extract both files mentioned.

The tweet image predictions what breed of dog is present in each tweet according to a neural network.

Additionally, I've extracted data from the Twitter API using Python's tweepy library. It is possible to check all this process on the Jupyter Notebook file. The file tweet_json.txt is in json format and contain each tweet's tweet id, favorite count and retweet count.

So, Gathering the data was the first step in all the data wrangling process.

Data Assessing

The next step was to gather each piece of data mentioned before, assessing them programatically and visually so we could identify quality and tidiness issues. After a thorough analysis it was possible to identify the following:

Tidiness and Quality Issues

- Replace 'None' with 'NaN' in the columns doogo, floofer, pupper and puppo
- The columns doggo, floofer, pupper, puppo should be values in one column, not columns
- Rename names with lower case that are not actually names
- Replace words with the '&amp' in the column text
- Convert Denominator and Numerator from int to float
- The timestamp column is an object. It has to be a datetime object.
- The column text can be splitted in two columns, text and url
- Convert the tweet_id to string
- Combine Denominator and Numerator in one column
- Remove columns in_reply_to_user_id and in_reply_to_status_id

Data Cleaning

Finally the next step was to clean the data. Here is where we can fix all the quality and tidiness issues that were identified in the previous step. Basically, data cleaning is the process of detecting and correcting corrupt or inaccurate data from the data set. So, after proceeding with the data cleaning and solved the tidiness and quality issues, I've generate the final version of the file that is twitter_archive_master.csv.