Wrangle Report

1. Introduction

The purpose of the following report is to outline the main difficulties encountered while attempting to wrangle the data connected to the 'We Rate Dogs' Twitter database. The report is this split into three main parts:

- Data Gathering
- Data Assessment
- Data Cleaning

2. Data Gathering

In order to produce the final dataset used for the analysis, information from the following three sources has been gathered:

- WeRateDogs Twitter archive – the information for this was gathered through a csv file called 'twitter_archive_enhanced.csv'
- Tweet Image Predictions – downloaded programmatically by extracting the information contained at 'image_predictions.tsv' from Udacity's server
- Retweet and Favorite Count for each tweet – extracted using the Twitter API

3. Data Assessment

During this stage a total of 12 issues were identified – 9 data quality issues and 3 data tidiness issues, which would be discussed in the following section.

4. Data Cleaning

This section describes my efforts put into fixing the issues identified at the data assessment stage:

- Since we only want original ratings, I removed all the entries for the 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' columns which are not equal to 0. In addition, I also entirely cut the 3 columns above, as well as the 'in_reply_to_status_id' and 'in_reply_to_user_id' ones which relate to the retweeted data (that we do not need).
- There were a couple of columns in the Twitter archive table which had an incorrect data type. As an example here, I changed the type of the entries in the 'timestamp' column from 'object' to 'datetime'.
- There were some invalid names in the 'name' columns in the Twitter archive data (i.e. 'None', 'the', etc.). I therefore replaced all names starting with a lowercase letter as well as those entries which say 'None' with 'NaN' (I chose to use 'NaN' since that is how the missing entries have been described in most of the other columns)
- There were some invalid values in the rating denominator column (i.e. different than 10). I checked the 'text' entries for all the entries of the 'rating_denominator' column which are greater than 10 in order to see if assigned rating is correct and changed manually the rating where appropriate

- The entries in the ‘text’ column contained the url addresses which in my opinion was unnecessary and polluted the data. I therefore remove the url-s from the 'text' column.
- Most of the column names in the Tweet Image Predictions table were not descriptive enough so I changed the names so that they convey better the message.
- The names of some dg breeds were capitalized while others were not, so I made them all capitalized for the sake of consistency.
- I changed the name of the ‘id’ column to ‘tweet_id’ so that it matches the column names form the other 2 columns
- I combined the 4 columns describing the dogs stages from the Twitter archive table into 1
- I split the ‘timestamp’ column into 2: ‘date’ and ‘time’, which would make the analysis afterwards easier.
- Finally, I combined the 3 tables above into 1 so that we do not need to constantly change tables.

5. Conclusion

To sum up, a number of data quality and tidiness issues have been fixed, however, from my observations there seem to be at least as many more to go. The end result (the ‘twitter_archive_master.csv’ file) combines and prepares the data for the following analysis. The efforts put into wrangling the above data sets really helped me to understand why the cleaning of data is really important in data analysis and also allowed me to greatly improve my Python coding skills.