Data Wrangling Report

Introduction

The purpose of this project was to successfully wrangle various datasets and create some interesting analysis and visualisations about dog ratings on twitter. Data wrangling for this project consisted of 3 steps; gathering, assessing and cleaning data.

Gathering Data

The data came from 3 sources:

1. The WeRateDogs Twitter archive was provided as a csv file which was downloaded manually.

2. The image predictions file was downloaded programmatically using the Requests library.

3. Additional data was queried from the Twitter API to provide some more information than archive provides.

Each of these datasets were loaded into a pandas data frame before cleaning.

Assessing Data

The data can be assessed either visually or programmatically. Best practice ensures both methods are used.

The following visual assessment was carried out

- The archive dataset had tidiness issues in that they were columns that contained values rather variables. These columns were ‘doggo’, ‘floofer’, ‘puppo’ and ‘pupper’.

- There was no rating variable, only 2 columns that contained the denominator and numerator.

- There ‘Source’ column contained urls.

- There were too many prediction columns

- The dataset contains retweets, we only want the original tweets.

- The names column contained values that weren't names

The following programmatic assessment was carried out

- There were columns with the erroneous data types such as 'timestamp' and 'retweeted_status_timestamp' were objects and not dates. 'Tweet_id' and 'source' were also an erroneous datatype across the data sources.

- There were 'None' values in the 'doggo', 'floofer', 'puppo' and 'pupper' columns.

Cleaning Data

Before any further work was done, a copy was taken for each of the data frames, then cleaning was done as follows.

1. In the Twitter Archive dataframe the 'timestamp' and 'retweeted_status_timestamp' columns were changed to the date datatype.

2. In the Twitter Archive data frame, the columns doggo', 'floofer', 'puppo' and 'pupper' contained 'None' values. These were replaced with NaN values.

3. The stage categories of doggo', 'floofer', 'puppo' and 'pupper' should be in one column as they are one variable.

4. The urls were removed from the 'source' column with a function that searches for the source and returns the name of the source.

5. The 'source' column is categorical, therefore the datatype was changed to 'category'

6. Retweets and replies were removed by taking only the rows which have NaN values in the retweeted status id and in reply to status id columns. These columns were dropped also as they only contain NaN values.

7. Tweet Id for all the datasets was changed to type string.

8. There was not one single ratings column, just the numerator and denominator. A new column was created by diving the numerator by the denominator.

9. Some of the names were incorrect, these were lowercase and replaced with null values.

10. There were several columns in the image predictions data frame that were put into the one column. If 2 values existed these were joined together.

11. Columns were removed that were no longer needed.

12. The breed names in the image predictions data frame were tidied up by putting them all to lowercase and replacing a _ with a space.

13. Finally, all 3 datasets were merged into to become tidy. They all contain information about tweets.

Conclusion

Although this is not an exhaustive list of all the cleaning that could be done on the 3 datasets, it was a good start and shows the importance of clean and tidy data using the define, code and test steps.