Udacity: Data Analyst Nanodegree - Project 4

Wrangle Report: Data Wrangling of WeRateDogs Twitter Archive Dataset

Gurps Rai

Wrangle Report

In this report, the different components of data wrangling will be broken down to describe the various aspects of the processes that was used during the wrangling of the WeRateDogs twitter archive dataset.

Gather

The data for this this project was gathered from three different sources, each of which was in a different format.

The file names of the three sources:

- image-predictions.tsv
- tweet-json.txt
- twitter_archive_master.csv

The data for *image-predictions.tsv*, and *twitter_archive_master.csv,* were read into separate pandas DataFrames, using the **.read_csv()** method. However, the *tweet-json.txt* file was iteratively read into a line by line, and then appended to the DataFrame.

Assess

The three separate DataFrames were both visually and programmatically assessed for quality and tidiness issues, by means of the below pandas functions/methods:

- df.info()
- df[column].value_counts()
- df.head()
- df[column]. duplicated()
- df.describe()

Quality

The quality issues identified during the assessment process:

- Difference in *df_twitter_archive* records to *image_predict_df* records, 2356 to 2075 respectively
- *df_twitter_archive* 'name' column has very likely invalid entries, e.g. 'a', 'an', 'the' etc.
- Only original ratings (no retweets) required, the *df_twitter_archive* has 181 entries of retweeted data
- Dog stages are object (string) types, should be categorical type.
- timestamp in *df_twitter_archive* records are in object type, should be is in 'datetime' type.
- 'tweet_id' columns in 'int' format in all three tables, should be 'object' (string) type.
- Only require ratings that have images, and not all the ratings have images.
- *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_id*, and *retweeted_status_user_id* are all in 'int' format, should be in 'object' type.

Tidiness

The tidiness issues identified during the assessment process:

- *image_predict_df* has three prediction columns, but only single column is required. If we determine the breed of the dog from the predictions, we can drop all of the predict columns.
- Three tables can be merged into single table, with only the necessary columns included.
- Dogs stages are in separate columns, these should be in single 'dog_stages' column, with rows containing the observation data (i.e. what stage the dogs are at).

Clean

The original DataFrames were all copied before the cleaning process.
All issues identified in the assess phase were documented by defining the issues, coding the solution, and testing for success.

The following cleaning solutions were used:

- The DataFrame apply() function was used to confirm if the *image_predict* dataset had a confirmed dog breed, and if so, the breed was appended to the table.
- All other image prediction columns were then dropped, as the useful information had been taken.
- The unneeded columns identified in the *twitter_archive* were dropped.
- Pandas melt function used to 'melt' the various dog stages, into a single 'dog_stage' column.
- The duplicate entries without a 'dog stage', produced from the melt function, were dropped.
- The 'tweed_id' columns from all three tables were converted to object/string type.
- The three tables were merged, on 'tweet_id', to the *twitter_archive* table
- Only valid 'name' entries were found to be capitalised, therefore the apply function was used to convert all invalid entries to 'None'.
- The 'timestamp' column was converted to 'datetime' type
- Convert both 'dog_stage' and 'dog_breed' columns to 'category' type.

Store

The gathered, assessed and cleaned dataset was stored to a CSV formatted file, ‘twitter_archive_master.csv’.

Analysing and Visualise

The cleaned dataset was analysed, and several plots produced to help provide insights into the data, see *act_report.pdf* document.