Wrangle report

Introduction

This is a wrangling report for the WeRateDogs Twitter page for the Udacity Data Analysis Nanodegree program. The Twitter user's tag is @dog_rates and it rates people's dogs and leaves humorous comments. These ratings mostly are out of 10, but they can be inconsistent. In this, I will briefly report the wrangling efforts on the data retrieved from this account.

The main tasks

1. Gathering the data
2. Assessing the data
3. Cleaning the data

Gathering the data

The data was obtained from 3 sources:

1. Twitter archive file: twitter_enhanced.csv was provided by Udacity.
2. Image predictions on the images retrieved from the Twitter page: Provided by Udacity. It is hosted on Udacity servers, and it was downloaded using the request library. The file name is image_predictions.tsv.
3. JSON data: This data was retrieved by using the tweet IDs provided in the Twitter archive file. It was pulled using the Twitter API using Tweepy library. The data was stored in a .txt file called tweet_json.txt. It was then read into a Pandas dataframe.

Assessing the data

There are 2 key assessment methods used: visual and programmatic.

As for the visual assessment, it was done by printing out the data in the Jupyter notebook and by looking at the files manually.

As for the programmatic assessment, things like counting the values, looking at the information of the dataframe, checking duplicates, and so on, were applied.

The issues found were categorised as either quality issues or tidiness issues.

Cleaning the data

Twitter archive

1-Delete retweets

2-Delete unneeded columns

3-Create a column from the dog stage columns

4-Convert the timestamp into day, month, and year.

5-Convert them numerators and denominators type to float

6-Update some numerators manually

7-Further manual updates

8-Delete the tweets for which there are no ratings

9-Create a new column for rating, which is numerator divided by denomerator, multiplied by 10.

Image predictions

10-Delete duplicates in jpg_url

11-Use first prediction to be true as dog type and capture the confidence level in the prediction.

12-Drop rows that contain errors in prediction

13-Get rid of unneeded columns

Tweet_json

14- Tidiness - Only keep original tweets

17- Tidiness - Change tweet_id column to int64 to match the format of the df_twitter1 dataframe

Merged twitter archive and image predictions

15- Tidiness - Merge twitter_archive_clean and image_prediction_clean into 1 dataframe

16-Keep rows that contain a picture only

Merged twitter archive, image predictions, and tweet_json

18- Tidiness - Merge df_twitter and tweet_json_clean into 1 dataframe

19 - Tidiness - Return the tweet_id to string

Conclusion

Using Python and a couple of libraries, I have managed to wrangle this dataset to be able to produce some useful results and be able to manage and store it properly. Alongside that, the tools provided with these libraries enable us to scrape data off the web, assess the data, clean the data, and store the data. It is much more capable of handling large data than Excel and the other software tools commonly used.