WRANGLE REPORT

Rofhiwa Kgomo

September 2022

This short report describes the wrangling efforts involved in the project (data-wrangling) of twitter account “WeRateDogs”, as part of Udacity’s Data Analysis Nanodegree.

1. **Data Gathering:**
   Gathering Data for this Project involved obtaining three different datasets from three different sources. Each one testing a different way of obtaining a dataset. The data was gathered from three different resources in three different formats.

   a. twitter_archive_enhanced.csv file - This file was downloaded manually and then imported into our working environment using Pandas library.
   b. image_prediction.tsv file - This file was downloaded programmatically using Requests library from a provided URL. This file has image predictions results for the dogs’ breeds obtained through a neural network on most of the tweets in the archive file.
   c. api_json.txt file - This file was gathered from twitter’s API via the Tweepy library by querying the API to obtain extra information pertinent to the tweets’ ids in the archive file, e.g. retweets count and favourite count.

2. Data Assessment

The three saved data frames were assessed visually and programmatically. In this stage, we inspect our datasets to make sure it is eligible for our next analysis stage. The datasets were assessed under two criteria, quality and tidiness. When an issue was detected it was documented under one of these two criteria.

Quality refers to issues related to the content of the data, sometimes called dirty data. The standard criteria of completeness, validity, accuracy, and consistency of the data were used to identify quality issues. These issues were varied and are listed in the assessment section of the “wrangle_act.ipynb” jupyter notebook.

Tidiness refers to issues related to the structure of the data, sometimes called messy data. The basis for assessment is that each variable forms a column, each observation forms a row and each type of observational unit forms a table. After assessing, an assessment summary was created to list all discovered issues for the next stage.

3. Data Cleaning

The final step in the wrangling process is cleaning the data for quality and tidiness issues. The cleaning followed the standard process of define, code and test for each of the issues and they were tackled in a logical order, which is reflected in the numbering order in the “wrangle_act.ipynb”. Most of the cleaning was performed programmatically, such as defining functions, developing regular expressions to capture right records or using pandas built-in functions (merge, melt, extract, etc.) Also, some manual cleaning was performed to correct ratings error values.

4. Conclusion

Data wrangling provides a clean data frame for future analysis and visualisation. In my case I exported two CSV files: “twitter_archive_master.csv’ and “image_predictions_master.csv”. These files can also be shared with others without having to wrangle the data. Also, I exported SQL database file. “master_df.db”