We Rate Dogs Wrangle Report

This report briefly describes the wrangling efforts in this project as part of Udacity Nano degree Data analysis. The wrangled dataset is the tweet archive of Twitter user @dog rates or WeRateDogs, who rates people's dogs with humorous content about the dog. The ratings have a denominator of 10.

The report is divided into three sections.

- Data gathering
- Assessing
- Cleaning

Data gathering

The we rate project involves gathering three data frames form three different sources. The data were all about the Weratedogs experiment on twitter. Each tasted a different way of obtaining data ie. Csv, API and tsv. The data was gathered from three different sources in different formats

1. *Twitter_archive_enhanced.csv* file. The twitter archive file was provided by Udacity and was downloaded manually.
2. *Image_predictions.tsv* file. The tweet image prediction gives a prediction of the dos breeds according to a neural network. The file is hosted in Udacity servers and is downloaded programmatically using *request* library and URL.
3. *Tweet_json.txt* file. Using tweet id, twitter API and json, I queried tweeter API for each tweet json data using *tweepy* library and store all data sent into a file called tweet_json.txt file. There after I selected the needed columns into a panda data frame.

Data assessment

The three datasets were assessed visually and programmatically under two areas: quality and tidiness.

- Visually, I examined the entire frames on Jupiter notebook and checked the file externally using MS Excel.
- Programmatically used different in-built functions like info, head, duplicated, head etc.

8 issues were detected under quality and two under tidiness.

Under quality issues pertaining content of data was assessed. This is sometimes called dirty data. We checked out the completeness, validity accuracy and content of the data to determine the quality issues. the quality issues were documented in the assessment section.

Under tidiness criterion issues related to structure of data was looked at. This also called messy data was assessed by observing variable forms of a column, rows and tables. after

the assessment the issues were documented, the assessment summary created a list of issues to be addressed in the data cleaning stage

Data cleaning

in the third stage, the data is processed to address the quality and tidiness issues. the data cleaning started with merging of all the tree datasets. Then the issues were tackled through, the systematic define-code-test framework. This involved documentation of issues which translated to code. The code was executed then tested to verify if the issue was resolved. If there was another pertinent issue, that still affected data analysis, the three-step framework was repeated to address the issue. The issues were tackled in logical order. most of the issues were issues talked programmatically using function, or built-in functions like merge, melt.

An Interesting cleaning was to melt the dog stages into one column instead of four (puppo, flooper, pupper and doggo)

Conclusion

At the end of the data wrangling phase, cleaned datasets were merged and stored as a separate file and made available for the next part of this project. Data wrangling steps created clean data frame for future analyses and visualization.

All of the data wrangling processes must be redone if the collected data is insufficient to solve these queries or issues. Thus, a large portion of data analysis is undoubtedly defined by data wrangling, which is probably iterated over the course of the analysis to at the very least make it possible to analyze and visualize data in order to draw insights.

I used Python and some of these packages which is more advanced than excel. This case created a master file that can be used for future visualization it master.csv