Computer Programming Project Report

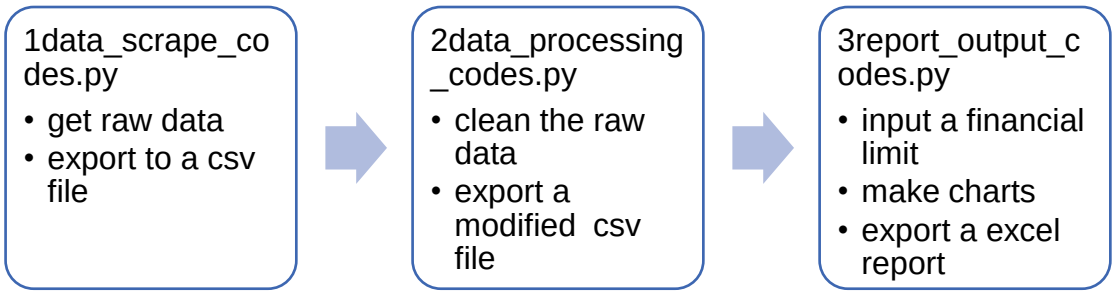
STUDENT NAME: Chen, Yao
STUDENT NO.: 34529638

Summary

PROJECT TITLE: How to Choose A Used Car on Trademe.co.nz Within Your Budget
OVERVIEW OF THE PROJECT: This project is to scrape selling data of used cars from Trademe.co.nz and export a chart report considering a financial limit given by the user. The report can show the correlation between price and different characters of used cars, including districts, number of seats and kilometers and so on. It can be used as a budget guidance when we decide to buy a used car.
LIST OF WORKS IN PORTFOLIO: 1.Python codes, including: • 1data_scrape_codes.py • 2data_processing_codes.py • 3report_output_codes.py 2.Sample data, including: • 1_raw_data_scraped.csv • 2_modified_data.csv • 3_stat_report.xlsx

Design Notes

This project is designed to run three programs and each of them can be used or modified separately for future necessary. Specifically, in this project the codes need to work successively:



Programming description

1. About the scraping part

The main() function contains three steps:

Note: 1.the main website: https://www.trademe.co.nz/motors/used-cars/more-makes
2.each website: the detailed website, e.g. https://www.trademe.co.nz/motors/used-cars/mini/auction-2149243046.htm

Since there are pages on the main website and it is needed to find key data in each car page following the same rules, three functions are defined and called through “for” loops in the main function. “content_of_webpage()” is used to return the content of a given website, “item_url_part()” is to construct the set of car urls, and “item_dict()” will return a dictionary of the characters of a car, with attributes labels as keys and attributes values as values.

“re” module plays an important role in this part. Because the target data are stored in the html form and can be collected by constructing specific patterns and applying the “compile” method in the module. 60% of the workload in this part is to find content that fit the patterns by using the “re.compile” method.

Beside functions and “re” modules, “set” is used to save the target data instead of “list” to save the running time of scanning 50 webpages from 60 minutes to 20 minutes now and to avoid repeatedly analyzing if one record is duplicated as well. It makes the whole program efficient to a great degree.

The scraping codes probably need to be modified when we want to scrape data from a different website, because the patterns of the target data must be different. They can be maximum retained if we only apply them on the “Trade-me used-cars channel”. In addition, scraper can be easily blocked by the website and there are no such codes in this part to alert these conditions, which can be improved in future studying.

2. About the data processing part

There are three steps in the main() function:

Data processing can vary all the time depending on the target you set to achieve, the codes can not be fixed all the time. What we can do to eliminate the variability of the code is to make some small functions to suit for common conditions. “new_column()” is to merge similar columns, for example, there are two columns named 'History' and 'Import history', in fact they represent the same attribute of a car, that is whether this car is imported or not, so they need to be integrated into one column. “remove_character()” is defined to remove the specific character in a column, such as comma, brackets and quotation marks. Data type is another important problem in the data clean process. In the following program, we will use the cleaned data to do some statistical analyzing, so data type of some columns, like price, kilometers and so on, must be float; and some columns should be strings, we need to do some transformation, which can be done well by pandas module.

The process part will also export a csv file, because even though we will get the report in the last part, people tend to do analysis on their own purpose, or to see the detail information, then a file that can be flexible handled is preferred.

The processing part seems some kind complicated and may be improved later. But for this specified project, it can do well.

3. About the report part

The main() function contain three steps:

The main modules used in this part are pandas, matplotlib and xlsxwriter. The first two are applied to generate the charts and the xlsxwriter is to insert graphs into excel. Graphs generated in this program are mainly about the car number and price by using bar and boxplot charts, which can be attained through pandas or matplotlib.pyplot. Graphs always have the same parameters, so defining functions can make it easy. In this program I defined three functions to deal with the repetition of making charts of different variables. “column_plot()” is defined to draw a bar chart and a boxplot chart for the same column at one time, and “group_count_bar_plot()” and “group_price_box_plot()”are defined to produce bar and boxplot chart for a column that needs to be grouped before applied into a statistical calculation. For example, when calculate the price among different kilometers it is better to group the scattered kilometers into intervals, such as 10000-50000, 50000-100000 and so on.

When we run the program, a warning will show up:

```
f:\Anaconda3\lib\site-packages\numpy\core\fromnumeric.py:57: FutureWarning: reshape is deprecated and will raise in a subsequent release. Please use .values.reshape(...) instead
  return getattr(obj, method)(*args, **kwds)
```

But it will not influence the result but may be a warning that the package or module should be updated.

The biggest problem about this part is the quality of charts exported out especially the grouped ones. Plots of grouped objects will clean the title set on the figures, so it is tricky to make the graphs clearly understood. To settle this problem, I used the name of the charts files instead of title to show what these charts mean at the cost of appearances.

Furthermore, on the base of this project, we can do more, for example, we can generate some linear regression to figure out the expected price on different district, different number of seats and kilometers and so on, which can better tackle the question raised at the very beginning of this project.