Comp20008 Project Phase 3

Education level and personal average income

(2010-2014)

Domain

The domain of the study covers education and labour.

Question

The objective of this study is identifying correlations between education and personal income in Victoria. It highly relates to state governments on education-related policy making and finance planning. The study aims to answer the following questions:

1. How is the personal annual income in Victoria?
2. How the education expenses structure?
3. How government's expenditure on education effects personal annual income?

Datasets

The following three Primary datasets were covered:

- **Estimates of Personal Income for Small Areas**: modelled estimates of annual income details from states, in 2010 to 2014 included attributes *earner(person), median age of earners (years), income ($), median ($) and mean ($)* based on the source from the Australian Tax Office (ATO). It also listed those attributes with detailed local government area level.
  URL: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6524.0.55.0022011-2015?OpenDocument
- **Government Finance Statistics, Education, Australia:** this dataset contained on detailed purpose the education expense invested in (*primary and secondary education, tertiary education, pre-school &education not definable by level, transportation of students, education n.e.c*) started from 2006 to 2015 in the unit of million AUD. The dataset was divided by each state government in separate sheets. It further described more aspect of the educational expense, such as *net acquisition of non-financial asset, gross fixed capital formation, sales of goods and service by purpose*, which would be helpful for study in depth.
  URL: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/5518.0.55.0012015-16?OpenDocument
- **School location in VIC:** this dataset contained the details of all school in Victoria by Department of Education and Training, updated in 2018. Schools details include name, school sector, school type, address, phone, so on and so forth, stored in CSV files. Schools vividly mirrored how government invested expense on the educational system, especially the public schools run by government.
  URL: https://www.data.vic.gov.au/data/dataset/school-locations-2018

Those two chosen datasets were both from reliable sources, the data were recorded have overlapped period, 2010 to 2014, which were critical for successful integration and analysing accuracy.

Pre-processing

The pre-processing stage was paving a straight road for subsequent integration with the help of Pandas library. First of all, we identified which set of data should be imported, since the source xlm file was divided data into multi-sheets in different statistic methods, such as by levels of areas.

Extracted the useful datasets, it still was not ready to be integrated or visualized before data cleaning. The format that used for decorating excel tables on the raw data (header, commentary and footer etc.) were all discarded.

Extraneous data reduction, it aimed on reducing error and noise and screening the useful information. Due to the restriction on excel files, we sliced off the irrelevant data rows/column by Pandas index-based logical selection function, derived '*earner(person), income ($) and mean ($)'* from '*estimates of personal*

income for small areas', derived *'primary and secondary education', 'tertiary education', 'Pre-school and education not definable by level', 'transportation of student', 'education n.e.c.', 'total operating expenses on education, by purpose'* from *'government finance statistics, education, Australia'*, where *'tertiary education'*, we warped up the education same levels (*'university education', technical and further education', 'tertiary education n.e.c.'*) to this instance to match the other instances. Furthermore, we also reduced the time period columns that stood outside the time period, 2010 to 2014 that we were focusing on. Filtered out all non-government school via *'entity type'* in *'all school location'* datasets.

Where there were missing values, the whole columns were removed out of preventing the study outcomes from blindly calculating, making an attempt to fill out the blanked values without sufficient knowledge, which equalled high risk of generating vital errors. Also, the zero recorded values were removed, it could become the outliers skewed the final result. We decided to leave the existing outliers unchanged, as through the datasets were created by authorized institution, we firmly believed the outliers had its own meaning, it might lead our study to an interesting result.

Renamed the rows and columns was a key preparation for the following integration and visualization, we hoped that the datasets listed in temporal consistency when it to be integrated the datasets and the values that visualised consisted to their attributes. More about the renaming, we needed to name them in a clear format, for example, the former year presented in yyyy-yy, now we rewrote them in format yyyy, it's clearer to be identified.

Integration

In this study, we were looking for the relation between education and personal income based on the Victoria and the Victorian local government areas. Two data frames were about to be created in different governments level.

Datasets were integrated challenges thanks to the sufficient preparations have been done at pre-processing stage. The integrated steps were done in the assist with Pandas library, passed the datasets in Pandas data frame, the datasets were presented in matrix. Before integrating, we found the education expense data frame was not in the shape that it could be merged, applied data frame transpose to make it have a consistent number of rows as the other datasets.

To integrate the dataset in Victorian government level, we preformed concatenate datasets along columns with Pandas build-in concatenation function, it's much straight-forward than merge function.
As for the integration of dataset in Victorian local governments level, we extracted the mean annual income from each local government, performed a logical selection in 'all school location' dataset to filter by using local governments name. Then used value count function to count the frequency and place in the number in the corresponding row.

To evaluate if the datasets were ideal for further investigation, if the data randomness was excessively significant, we were hardly able to have a convincing and reliable result at the end, which meant it's not worth investigating. Lag plots were helpful on measuring correlation, which showed us graphically how random the data was (figure 1,2). As we can see, the level of randomness was quite low, and no outlier stood outside the data cluster, which indicated investigation can be proceeded.

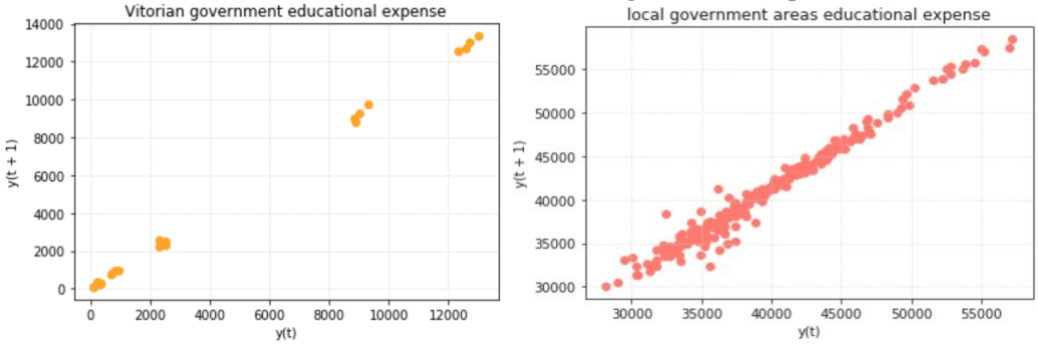

Figure 1-2

Result

In the level of whole Victoria

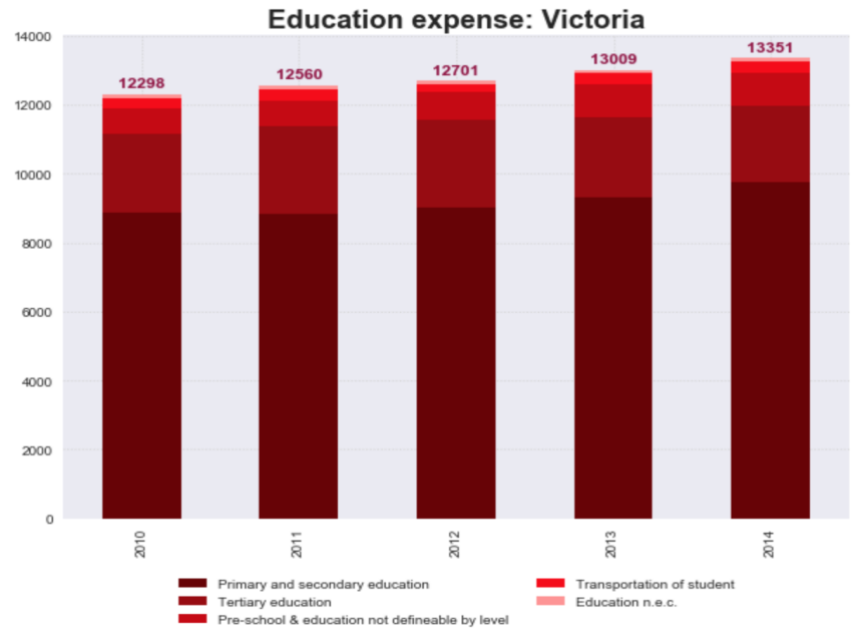

Figure 3 illustrates the education expenditure that government invested in the educational system, the stacked bar plot explicitly shows us the proportion that each section of the expense by purpose. The largest section is primary and secondary education, which means the government think highly of the importance of primary and secondary education and it encouraged us to analyse the relation between the primary and secondary schools and the mean annual income.

Figure 2

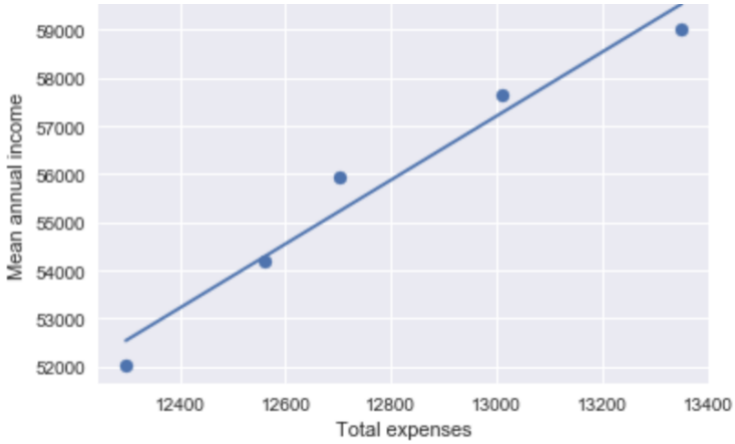

Figure 4 is a scatter plot whose x axis is total operating expenses on education by government, y axis is mean annual income of Victorian, each dot corresponding to the variable changes in five years, 2010 to 2014. We calculated a line of best fine and mapped it on the scatter plot, the correlation coefficient was +6.65, as the definition of Pearson Correlation Coefficient, we have evident to say, the total operating expenses on education put in by government and the mean annual income of Victorian have a perfect uphill linear relationship.

Figure 4

In the level of local government areas

Figure 5 shows the number of government schools in each local government area in descending order. For a clear display, the corresponding number was place on the top of each bar. Yarra Ranges, Casey and Greater Geelong have over 50 government schools, whereas West Wimmera, Mansfield and Queenscliffe have less than 5. It also partially reflects how government allocated their education expense.

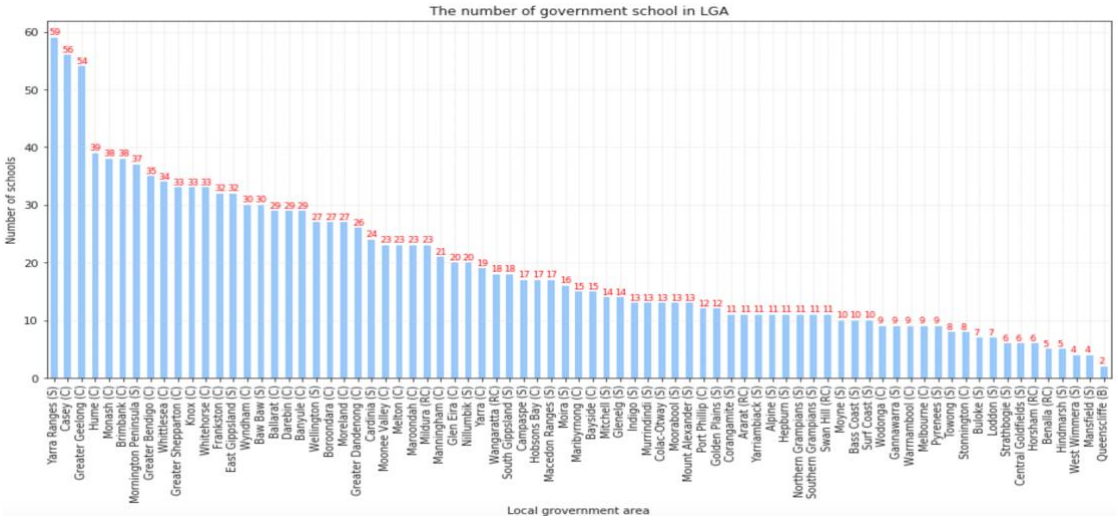

Figure 5

Figure 6 visualised the mean personal annual income from 2010 to 2014 in the order of the number of schools in LGA. By the rough comparison, we didn't have enough evident to tell the relationship between education and income, however, it reveals the most of annual income was growing by years.

It's hard to draw a safe conclusion due to the limited data we had, a further analysis should be taken to dig deeper on the education and income in order to find a convincing result.

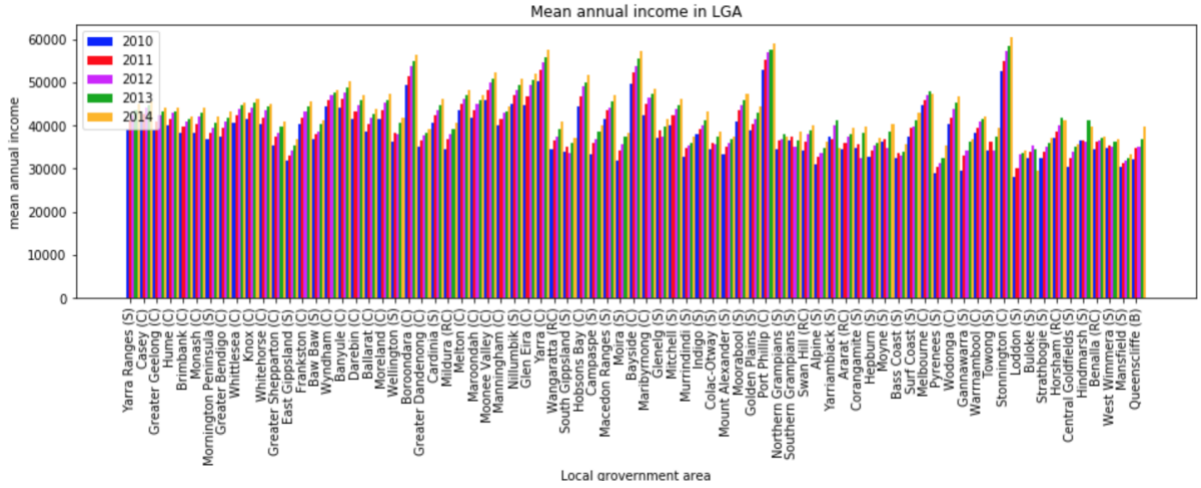

Figure 6

To explore the correlation if existed between the number of government school and mean personal annual income, we visualised the income data from year to year and the number of government school (figure 8-11). We set x axis as the number of government school, y axis as mean annual income, where each dot corresponded to one local government area, five figures corresponded to five years. The figures illustrate there are existing positive relationships between the number of school and the number of government school, even though there are outliers sitting on the top-left of the plots.

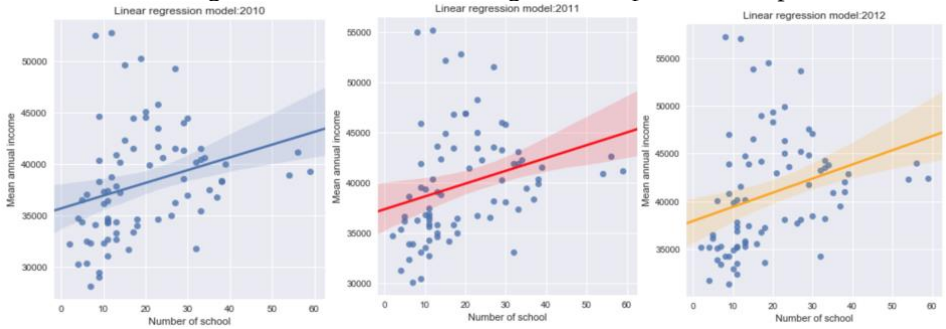

Figure 7-9 Linear regression model:2010-2012

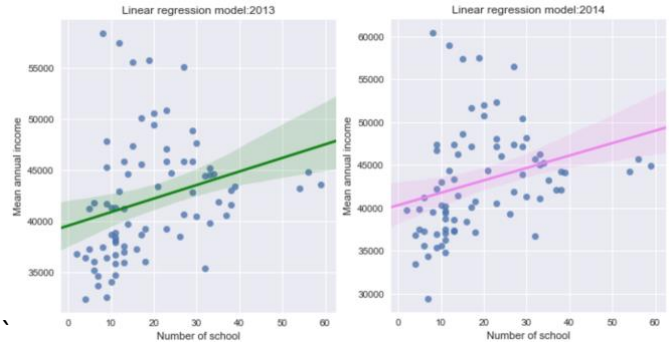

Figure 10-11 Linear regression model: 2013-2014

Value

Although the raw data was chosen out of the study interests, they were not fully having connection to each other. Before pre-processing stage, the raw dataset had large amount of useless data, screened data and removed them was significantly time-consuming, that ensured the following analysis was able to continue successfully. The raw all school data only provide a list of school name, school type, LGA name etc., without sorted. We generated the value which counted the frequency of data by filtering the data by LGA name, the value is the number of schools in each local government, which was a bridge that connect the raw all school data to raw income data. Only that, we could continue processing the datasets and visualised it.

Challenges and Reflection

The study was starting by drawing up a study direction and looking up the relevant datasets, where the dataset searching was most time taken step. The adequate datasets could greatly reduce the workload on all aspects of data wrangling, most importantly, the suitable datasets held highly relevant information to our study. Another important condition on choosing datasets was time consistency, in this case, the analysis result would be meaningless if the datasets weren't at the same period of time.

Question Resolution

The questions have been analysed in this study and here are the answers to the questions posted at the beginning. Firstly, the table 1 is showing the income details about people in Victoria from 2010 to 2014, the mean annual income of Victorian was growing annually. Secondly, there were large proportion of education expense paid on primary and secondary education and the following is tertiary education. However, some local government areas only have few number of government funding schools. Lastly, not only in the level of whole Victoria, also in the level of all local government areas, there are all positive relation on the government education funding and Victorian annual income.

Year	Earner(person)	Total income($)	Mean($)	Growth rate(%)
2010	2980656	155098994007	52035	-
2011	3035381	164497754479	54193	4.0%
2012	3077528	172135438755	55933	3.1%
2013	3152222	181782714358	57668	3.0%
2014	3260682	192443671450	59019	2.3%

Table 1

The result should draw the government's interests, since government is considering enhancing the wellbeing of Victorian all the time. We already known the education relates to income, as the income growing annually, government should keep investing funds in educational system increasingly and try to provide more schooling option to the low number of school areas. Funding the governmental school to improve the educational quality and lowering the fee.

Code

Over 250 lines of code were written for this study, the libraries used for pre-processing and integration were Pandas and Numpy, the libraries used for visualisation were matplotlib and seaborn. The other publicly available code that was used is source from *stack overflow,* have been linked in python code.