Income, Welfare and Fertility

Domain: Demographics and Wlefare Design

Question:

The Report seeks the answer to two question:

1 what is the relationship between peoples' income and their fertility.

2How government welfare affect people's choice on fertility

DataSet(Both included in the code submission):

- SA2 Income support.
  https://portal.aurin.org.au/

  This dataset is achieved from AURIN Portal which records the number of people that claim government welfare in each suburban. Welfare include welfare to children, unemployed person which are used in current report. This is included in the Code Submission.

- Local Government Area Basic Community Profile:
  http://www.abs.gov.au/websitedbs/censushome.nsf/home/map

  This dataset is achieved from ABS. There are 27 LGA profile being used, each is a collection of census data that conducted in 2011.

  Two data sheets in each Basic Community Profile were used in this project.

  One is datasheet B24, which records the **number of children ever born by age of parents** in each LGA.

  Another one is datasheet B17B, which records **total personal income(weekly) by age** in each LGA.

Pre-Processing:

- SA2 Income support.
  This dataset has missing values, which are all replaced by 0.
  However, the use of this data is a problem. Because the dataset is based on suburban level, yet another dataset is based on LGA level. Therefore, to convert the SA2 income Support Dataset, I had to first copy the LGA and their suburban level list from website. Then using python to classify each suburban into 27 LGAs. After getting the LGA list with their suburbans, I merge suburban income support data to the LGA level then being used. It also take time to do the string matching between two dataset
- Local Government Area Basic Community Profile:
  The data is clean. However, there are 27 LGA's datafile, therefore two script were used to extract the data in worksheet B24 and B17a, then stored in the CSV file "mother_by_children.csv" and "people_by_income_level.csv". The challenge for this dataset is to figure out one or more dataframes to store the multii-dimension data.

After getting row data, this data is convert by normalizing them to value in [0,1]. So that each factor is equally weighted.

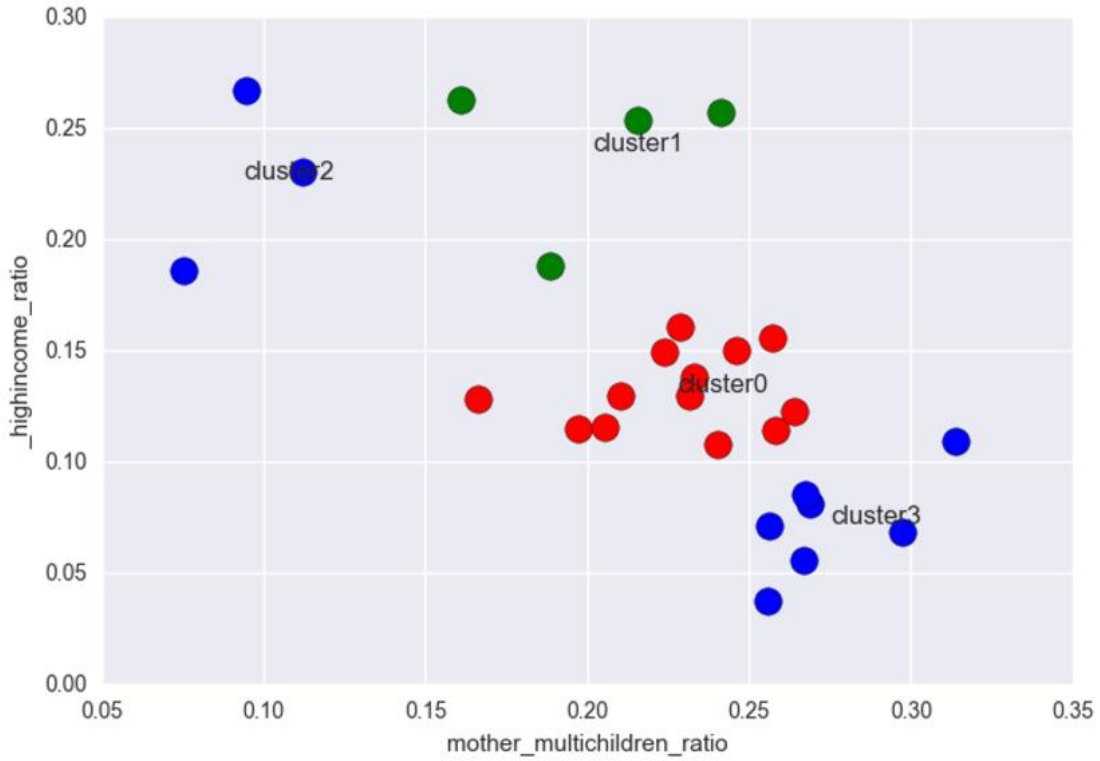
Data Integration (Classifying of LGA group using k-means algorithm):

To group LGA in a rough way, we define**:**

Mother Multi-children ratio: number of female that born >= 3 children / total female that report their children ever born in LGA

High income_ratio: number of people weekly income >= 1500 (By ABS definition) / total people that report their income in LGA

At the Integration part, since we concerned how welfare will influence people's fertility. K-means algorithm were used to classify 27 LGA into 4 clusters. This is because there might be LGA that has high income, high fertility; high income, low fertility; low income, high fertility and low income, low fertility. Therefore we classified LGAs into 4 clusters so that similar LGA with similar characteristic can be grouped together. And this will make their characteristic more obvious to help the research. This yield the following:



With Grouping of:

Cluster0	'Banyule ','Darebin ','Hobsons Bay ','Kingston ','Knox ','Manningham ', 'Maribyrnong ','Maroondah ','Monash ', 'Moonee Valley ','Moreland ','Whitehorse ','Wyndham '
Cluster1	'Bayside ', 'Boroondara ', 'Glen Eira ', 'Stonnington '
Cluster2	'Melbourne ', 'Port Phillip ', 'Yarra '
Cluster3	'Brimbank ', 'Casey ', 'Frankston ', 'Greater Dandenong ', 'Hume ', 'Mornington Peninsula ', 'Whittlesea '

And the following research is based on tis 4 grouping.

However, the limitation of such grouping, in calculating Pearson coefficient, there are limited data set in the cluster. Therefore the linear relationship between two factors may be over-estimate.

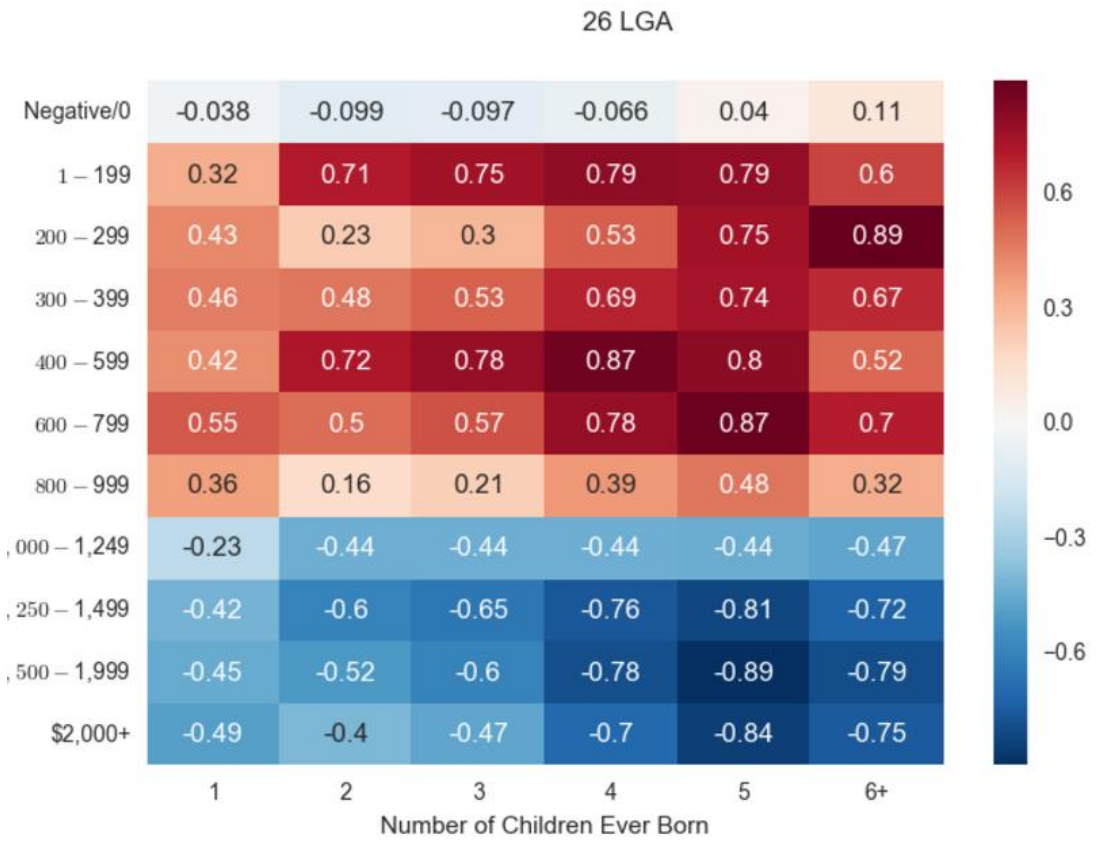
Value:

The merge of SA2 income support enable me to used two dataset at the same LGA level. The frequent use of heatmap help me identify the stronger linear correlation between different factors, and how correlated they are.

Also k-means algorithm was also used to separate 27 LGA groups in greater Melbourne. This reduce the scope of my research as well as reduced noise data so that, the particular characteristic of similar LGA's can be more obvious. This helped me easier to find out the underlying relationship between welfare and people's fertility.

Result:

Overall: Higher Income, Lower Fertility(Cluster0):



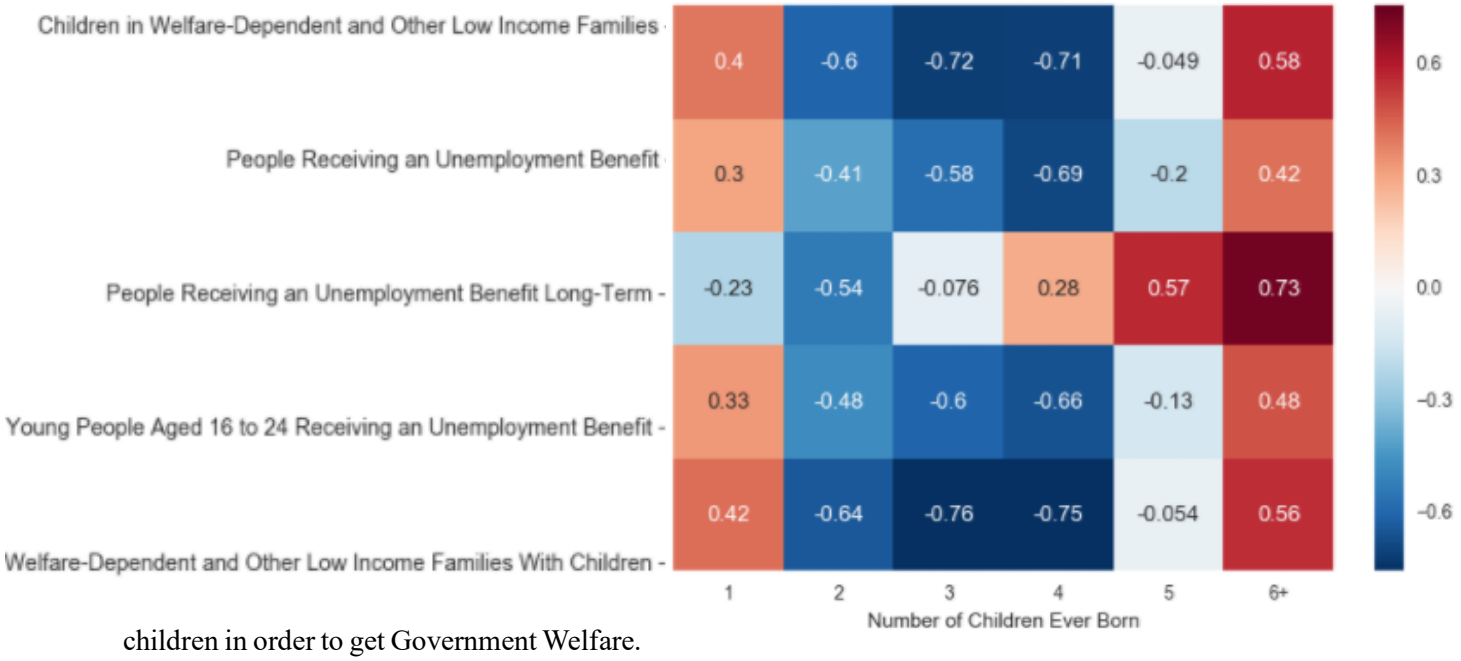
The color of heatmap is the Pearson Coefficient of parents with number of children Ever Born and the number of people at a particular income level.

From the heat map, we can see that, for Number of Children greater than 4, there is a strong positive relation with people that weekly income lower than 800, and a strong negative relationship with high income people whose weekly income is greater than 1250.

Therefore, we can deduce that, for high income people, they tend to have less children, that is why the more children, the lower negative correlation with the income level. While low weekly income people tend to have more children, evident by the Strong correlation at right top of the heatmap.

Poor Peoble Born Children to Get Government Welfare instead of Working:

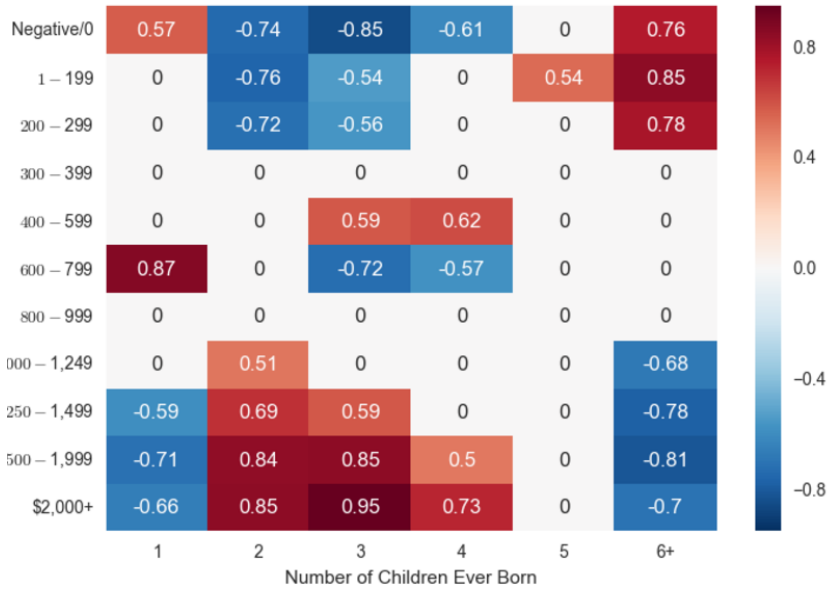
The result shows that, for **low income, high fertility area(Cluster3)**, there is a strong positive relationship between number of children ever born and the people that claim government welfare. In particular, this result tells us, at Cluster3 Grouping's area, poor people may just born



children in order to get Government Welfare.

This is support by the strong positive correlation between family with 6+ children and People Receiving An Unemployment Benefit Long-Term, which indicates that many people with more than 6 children might also unemployed. That is why in the 6+ column, all relationship with government welfare are positive and strong.

This is also supported by an income heatmap for cluster 0:



The income here suggests that, family with 6+ children has a strong positive relationship with weekly income lower than 300.Therefore it is consistent with previous argument that, many in this cluster3 group, many people born children to get government welfare.

Challenge:

The biggest challenge is that the topic set up in the Phase2A is too narrow and obvious-, as well as dealing with multi-dimension data. Dealing with multi-dimension data is the most time-consuming task. Because while I was trying to expand my previous topic to a more border topic, I have to extract 27 dataset Several program were made, in total of more than 500 lines. The complexity of multi-dimensional data also made modifying old topic hard and time-consuming. I tried to consider factors of occupation, industry of employment, different way of visualization. The complexity of multi-dimension data, fantasy of making a big project, lack of practice for pandas and the late decision on topic jointly result this late submit. The biggest mistake was that I did not choose a topic carefully.

Question Resolution:

From result, we have found out that higher income people tend to have lower fertility. Yet poor people tend to have higher fertility to get government welfare as an alternative way to get income.The designer of government welfare system would interest in the result. It is true and necessary that we should provide incentive to people to increase our fertility, because once a country's fertility has gone down, it can hardly raise again. However, it is also import to what kind of fertility we want. Do we want children being used as a way to get government welfare so that parenst do not need to work? Or we want children born and being educated to become a part of society, to help society make progress. This report has identified that some parents are using children to get welfare, and it is up to government that whether to indulge such thing or not.

Code:

There are 3 scripts around 300 lines of code were used for data wrangling

ChildrenVsIncome.py was used to extract a rough data from Community Basic profile. In particular, it collect the number of people with weekly income greater than 1500 and smaller than 1500, as well as parents born more than 3 children.

CoPYIncomeAndChildren.py was used to copy income data by age from 27 LGA's basic community profile as well and number of children ever born by age of parent

welfare.py was used to read the SA2 Income Support dataset. In particular, it first generate a LGA list with their Suburban, extract from Wikipedia. Then Based on this LGA list, the program classify each suburban data into the dataframe. Eventually it convert the SA2 Income data from suburban level to LGA level

Several Library were used, the function of them is listed as a table:

OS	System library of Python, used to open the dataset files
xlrd	An python Library that used to read excel file
Pandas	Mainly data tool that used to data integration and wrangling
Matplotlib	This libraray is used to do all visualization in the program
numpy	This is an assistance library that help data visualization with Matplotliv
sklean	This library is purely used to implement K-Mean algorithm
Seaborn	This is a visualization library that draw heatmap

Also hundred's of code were us in Jupyter for normalizing the data, creating dataframe for visualization and finding correlation with the help of different library in the above table.