1. Title

A Brief Study of High Income Resident distribution in Victoria

2.Domain

Communities, Finance, Urban Planning

3. Question this report seeks to answer

What is the population in Victoria? Is population and wealthy (having higher income population) correlated?

Which regions are resided with high income residents?

What types of people are resided in those high-income regions? Are they Foreigner or Australian?

What are their occupations? What can we summarize from all the analysis?

4. Datasets

The Dataset I used in the report are:

→ Set I: Industry of Employment by Occupation-Census 2016

This Dataset is about the occupation of different individuals within Victoria in 2016.

Categories: SA3 Code, Different Occupation titles (by Industry), the total of personnel in that industry

URL: https://data.aurin.org.au/dataset/au-govt-abs-sa3-p42a-industry-of-employment-by-occupation-census-2016-sa3-2016

→ Set II: Total Household Income (Weekly) by Household Composition-Census 2016

This Dataset is about the income generated by a household weekly within Victoria in 2016

Categories: SA3 Code, Household and non-household total, their income range.

URL: https://data.aurin.org.au/dataset/au-govt-abs-sa3-p28-total-hsehold-income-by-hseholdcensus-2016-sa3-2016

→ Set III: Data by Region - Population & People 2011-2016h

This Dataset is about the characteristics of the resident of each region within Victoria in 2016

Categories: SA3 Code, Residents age, district population and destiny, overseas born %(sort by countries)

URL: https://data.aurin.org.au/dataset/au-govt-abs-sa3-dbr-pop-and-people-2011-2016-sa3-2016

5. and 6. Pre-processing and Integration

Here is how I pre-process and integrate the raw data:

→Checking if there are any missing values

Luckily, all the datasets I acquired from Aurin are complete and without any missing data. They all have proper formats, like SA3 Code is in integer, and SA3 Name is in string format…etc.

→Sorting Irrelevant Data

However, there are over 600 categories in total for three datasets, most of their names have high similarities which make the analysis difficult. Therefore, the first step is to sort out some of the less relevant categories and group them into useful attribute.

1)I Remove most of them and educed the total attribute into some key attributes

- region names (For all three Set of Data),
- total population in that area (For all three Set of Data),
- Combined the **Weekly Income** into 5 Attributes (Low, mid_low, mid, mid_high, high)

Low: 0 - 499	Mid-Low: 500-999	Mid: 1000-1999	Mid-High: 2000-3499	High: 3500 or Above

- Combined the **Age** into 6 Attributes (Children, Young_Adult, Adult, Old_Adult, Young_Senior, Senior)

Children: 0 - 14	Young_Adult: 15-24	Adult: 25-39
Old_Adult: 40-54	Young_Senior: 55-64	Senior: 65 or Above

-Calculate the percentage with respect to the total population in that regions (For all three Set of Data)

-For Dataset III, I didn't import all the attributes, but only focus on the data of top 1 region.

→Wrong or Mismatch Value or Name in the Datasets

I) Some of the raw data like percentage of different ages add up more than 100%, which is not correct. I have also put the code in my Python zip file. II). Name Mismatch, for example, one of the dataset have code **"Creswick - Daylesford - Ballan",** however, another one put **"Creswick Daylesford Ballan"** which miss the hyphen, it will generate wrong results, and therefore I corrected it.

→ Normalization:

Many of the data are within huge range, like the maximum value is few thousands and the min is in single digit, we should normalize them to make the chart easier to see the relationships. **(e.g. the population)**

→ Previous Mistakes and Questions Modified:

I have done a lot of mistakes in Phase 2, some of the algorithms are not correct, therefore I acquired the wrong analysis. Such as the region with high percentage of high income residents. In this detailed analyzed report, I would correct most of them and slightly adjust the Aim I wrote on Phase 2.

→ Limitations:

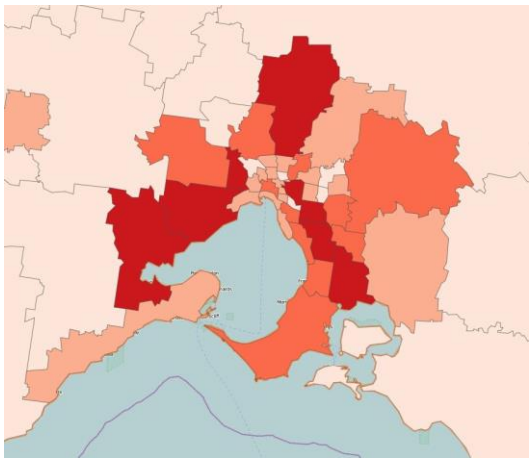
Map visualization is the best method of visualizing the results of this topic, However, I failed. The Python Modules I tried to use were **Basemap** and **GeoPandas,** because of my skill level and time limit, I failed to compile any code. Secondly**,** I did some researches, and I have no clue how to acquire **GIS** data of the SA3 code from online free sources. Fortunately, I found a way to visualize my data on Map by uploading processed dataset into Aurin.

7. Results

Before the analysis: Visualizing Population in Victoria

Since I do not have such knowledge to create Maps on Python, I use the mini app on Aurin's Portal to make such visualization. In the map, we can see where are the most highly populated areas.

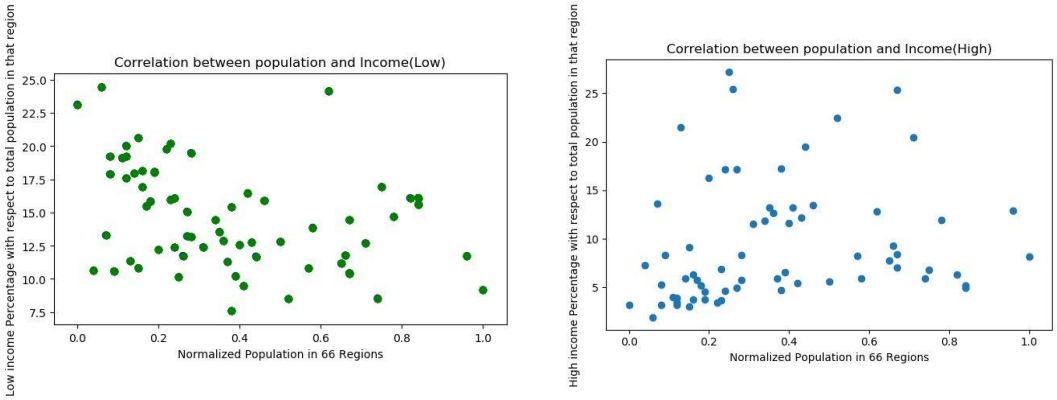
<Figure1: Total Population Distribution in Victoria and Melbourne, Tools used: Aurin.org>



Here is a small table to show their total population (Regions in red):

1.	Wyndham	233138	5.	Geelong	192393
2.	Whittlesea Wallan	224581	6.	Monash	184848
3.	Dandenong	197453	7.	Boroondara	177361
4.	Brimbank	196858	8.	Casey South	176002

Analysis 1 Is population and wealthy (having higher income population) correlated?



<Figure2&3: Correlation between population and income, Tools used: Python>

From the above graph, we can conclude that correlation between population and low-income resident have a slightly obvious relationship; regions with less population tend to have higher population of low-income group; While medium populations around 0.4~0.6 are likely to have more high-income resident; And for high population area tend to have more middle-income resident. This is in fact True, according to common sense: less populated regions most likely are far from the CBD, therefore not favored by the rich people; while regions with high population are likely reside most of the middle-class income group people.

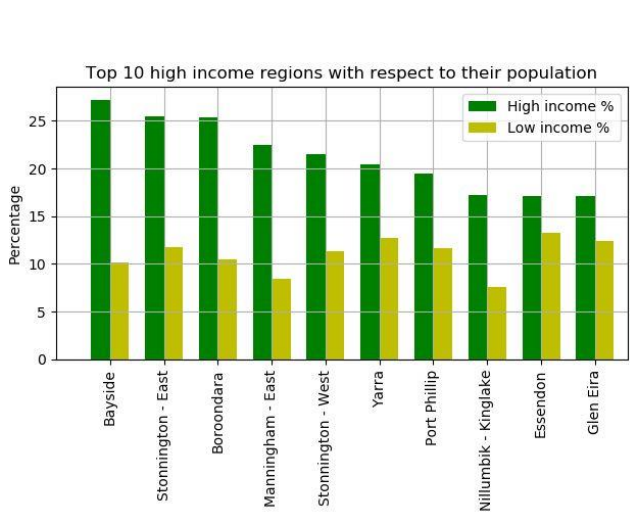
Analysis 2: How is the wealth distributed? Which regions are resided with high income residents?

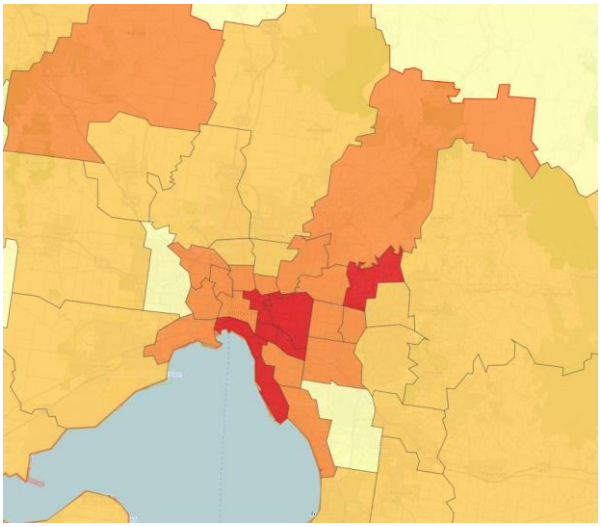
Here is the Top 10 Regions with highest percentage of high income residents with respect to their population and I also plot them into the Aurin map to have a better visualization.

Region Name	Percentage	Region Name	Percentage
1. Bayside	27.18	**6. Yarra**	20.48
2. Stonnington - East	25.46	**7. Port Phillip**	19.49
3. Boroondara	25.38	**8. Nillumbik - Kinglake**	17.25
4. Manningham - East	22.46	**9. Essendon**	17.14
5. Stonnington - West	21.51	**10. Glen Eira**	17.13

In average, there are around 81.01% people who are Australian Citizens in the above top 10 high-in regions.

Below will show their bar chart and their distribution around Melbourne (Deeper the color, higher the value) *(*Note: Area further away from Melbourne are less populated and less relevant)*





<Figure4&5: Total 10 high income regions in chart and their distribution on Map, Tools used: Python(Chart), Aurin.org(Map)>

Analysis 3: What types of people are resided in those high-income regions? What is the age distribution? What are their occupations? Are they Foreigner or Australian?

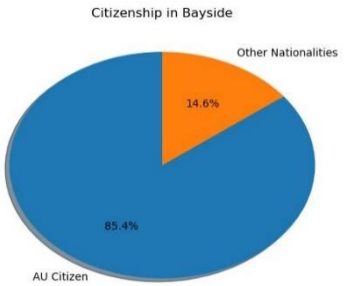
In this report, due to the length limit, I will only pick the top region to analyze. **The Bayside**

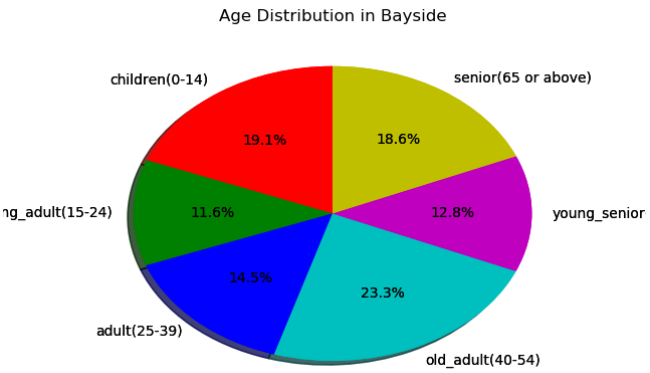
→ Where is Bayside?

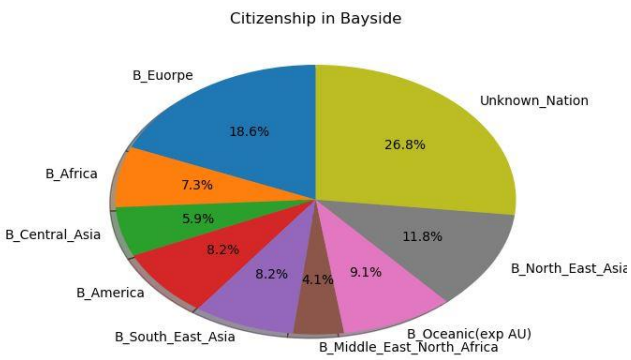
Bayside is a statistical area which contains the suburb Brighton, Brighton East, Hampton, Hampton East, Sandringham, Highett, Black Rock and Beaumaris. It is an area below Port Philip, South of St Kilda, this area is near the coastline. It has a population of **102737**. SA3 Area Code **20801**.

<Figure 5: Bayside, Tools used: Aurin.org>

And here is some chart on the resident's characteristics in Bayside:





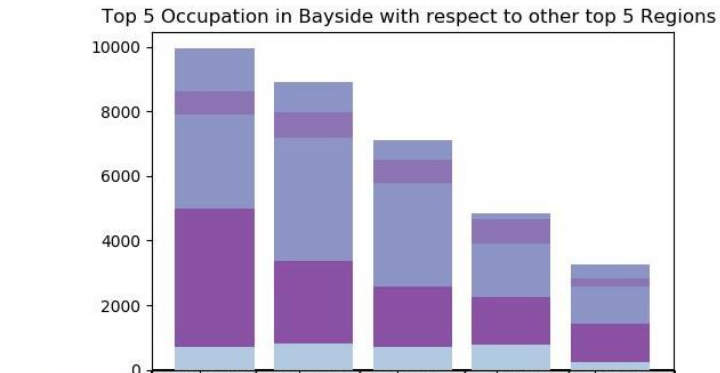


<Figure 6,7& 8: Bayside Characteristics Tools used: Python>

In this pie chart, we can see Old Adults (40-54) occupy **23.3%** of the total population in Bayside. And there is around **14.6% Non-Australian**, which there were **26.8% unknown nationality**, we can conclude from this graph that most senior and old adults with high income would reside there.

Analysis 4 and Conclusion: What are their occupations? What can we summarize from all the analysis?

Top 5 Occupations In Bayside (By Industry)	No of Res
Professional scientific and technical services Occupation Professionals (psc_tech_pfn)	3914
Health care and social assistance Occupation Professionals (hlth_sass_pfn)	2839
Education and training Occupation Professionals (edtr_pfn)	2836
Retail trade Occupation Professionals (retr_sal)	2130
Financial and insurance services Occupation Professionals (finin_pfn)	1352



	psc_tech_pfn	hlth_sass_pfn	edtr_pfn	retr_sal	finin_pfn
	0	1	2	3	4
Bayside	4251	2550	1858	1491	1185
Stonnington - East	724	802	720	780	248
Boroondara	7884	7186	5794	3895	2570
Manningham - East	2065	1710	1305	954	692
Stonnington - West	3914	2839	2836	2130	1352

From the table and graph, we can see the amount of top 5 occupations in Bayside are similar to the other top 5 high income regions. A small conclusion could be made is, these professions, especially **scientific and medical experts**, they are likely to have a higher income compare to other occupations. And they are likely to reside at the **East Side** of Melbourne CBD and close to the coastline. They are most likely Australian Resident, and most of their ages are around **40~54, or 65 above**. They tend to live in the suburb with **medium** population.

8. & 9 Value and Challenge and Reflection

Some reasons my integration and analysis bring value compare to raw data: I) Those Raw datasets didn't show which area is populated with high income people, with respect to region population and percentage. II) My Map visualization do clearly show where those regions are distributed geographically. III) The Characteristics of those resident giving some brief insight of what those people are doing and their age and ethnicity**. Challenge:** The most challenging or time-consuming task was to filter out those similar attributes and group them into useful attributes, they have 600 attributes in total, they spent me at least 6~7 hours to sort out. How to visualize it on map troubles me for several days, sadly I couldn't figure it out how to use **Basemap** and **Geopandas.** Fortunately, I realize there is mini tools on Aurin I could make use of. So, I create my own dataframe, export it as CSV and import it to Aurin and Visualize them. Some of data wrangling problem I have described them earlier in Part 5 & 6. They wasted me a lot of time.

10 & 11 Question Resolution and Code Used in this analysis report

I do answer all the questions in this report, however it is very broad topic, therefore a 5 pages report won't be enough to express all findings. But most of the primary and essential questions were answered through map visualization and graphing. (I have made a conclusion in Part 7) Those who would like to target high income customer group or hi-end customer; those who study wealth distribution in Victoria; for urban planning development; Real Estate company or agents might be interested in this **Small Findings.** They could target their customer easily by knowing where they live, their age group, their ethnicity…etc. Honestly, this is just a very short report, I don't think anyone in reality would interested in such primary work.

→ For the Code I use:

I mainly use python as the wrangling and processing tools, except for map visualization on Aurin. I wrote around 500 lines of code, with around 15~30 hours thinking and searching some of the methods online. (Google and Youtube) The library I mainly use: **pandas** and **matplotlib**. And also, some small functions that require libraries like **heapq** and **difflib**.

I mainly use pandas to organize the Dataframe and Series in the CSV file, also use it to export the CSV file. Some of the functions I used in pandas: **csv_read, DataFrame, Series, to_csv, nlargest, nsmallest...etc**

Matplotlib are used for Bar Chart, Pie Chart, Table plotting. Function used: **Plot.Bar, plt.pie, table() …etc**

I use **Heapq**'s function **nlargest** to find the largest values in a list.

Difflib is for its **SequenceMatcher** function to find similarity rates between two strings.

12 Bibliography

Nil