DATA EXPLORATION & VISUALIZATION REPORT

Exploration、Visualization & Analysis Of Melbourne Property

Feixiang Li
26669900

Table of Contents

1. Introduction

As is well-known, Australia is a popular nation of immigrants. More and more people poured into this country. Julie Tullberg(2007) provides a fact that "Best growth suburbs in 2006 include Tyabb, which had a 99 per cent increase from $226,000 to $450,000 median price, St Kilda West (49 per cent rise from $740,000 to $1.1 million) and Doreen (48 per cent rise from $250,000 to $370,000)". In this article, the author said Victoria's property market in more sustainable growth. Furthermore, the article (Reed & Richard, 2016) have examined the influence of population variables on the level of house prices. Due to these findings, I make an assumption about the property market could be associated with the population and security of suburb. Hence, this report will try to analyse the property trend in Melbourne real estate market and try to reveal its relationship with possibly potential elements associated with it, such as crime data, based on the collected data and implementation of data visualization. Therefore, the target audience would be those persons who intend to invest the Melbourne real estate market or plan to buy a dwelling.

2. Data collection and wrangling

The data is manually collected from the credible government open data website. It includes the population statics data, property data of Melbourne and crime data. For data wrangling, the main tool is R.

2.1 Population statics data

The raw population statics data looks like the following form. It includes a few columns, such as Postcode, Suburb name, age range from 0-4 to above 85, top country of birth and so on.

Postcode	Community	2012 ERP age 0-4	2012 ERP age 5-9	2012 ERP age 10-	2012 ERP age 15-	2012 ERP age 20-	2012 ERP age 25-	2012 ERP age 45-	2012 ERP age 65-	2012 ERP age 70-	2012 ERP age 75-
3040	Aberfeldie	4.2	7.3	8	8.3	7.4	24.3	28.8	3.2	2.1	2
3042	Airport West	6	5	4.3	4.5	6.3	31	22.6	4.6	4.9	4.5
3021	Albanvale	6.5	6.4	6.3	6.2	7.9	28.5	28.5	4.4	2.3	1.5
3206	Albert Park	7.1	4.9	4	2.9	4.9	35.1	25.3	5.4	4.3	2.4
3020	Albion	6.7	4.9	4.2	4.7	10.2	38.9	19.7	3.8	2.5	2.2
3350	Alfredton	7.2	8	8.4	8.9	5.8	26.4	25	3.3	2.9	1.7
3078	Alphington	6.2	6.2	6.7	4.9	6.2	30.9	27.7	3.3	2.2	1.9
3018	Altona	5.8	4.8	4.2	4.4	5.4	29.7	25.6	5.2	3.9	4.4
3028	Altona Meadows	6.2	5.1	5.2	6.5	8	30.7	27	3.8	2.8	2.2
3025	Altona North	7	5.6	5.2	5.4	5.8	29.5	18.9	4.9	5.5	5.2
3022	Ardeer	5.9	5.6	6	4.6	6.4	32.2	21.7	3.6	3.9	3
3143	Armadale	5.5	4.6	3.9	3.5	7.5	36.8	22	5.1	3.9	2.7
3364	Ascot	6.3	6.4	7.5	7.7	7.1	26	27.7	6.2	2.4	1.2

For collection of useful data information, it needs to wrangle the raw data. Firstly, useful columns are manually extracted from the source data. Afterwards, the column name is more precisely and meaningfully renamed. According to Australian Government definition of aging group, the children is under 15 years of age; the adult is between aged 15 and 64 years; the old people is above 65 years of age. Therefore, there are three new columns added into the data by calculation of age range. For example, the value of column “Children” is calculated by adding age range 0-4, 5-9 and 10-14. Then, for the purpose of convenience of data joins, the character values of column “Suburb” is turned into upper case. Eventually, the final data after wrangling is shown as following tabular.

A	B	C	D	E	F	G	H	I	J	K	L	M
POSTCODE	Suburb	Top country	total	Born overseas, %	Born in non-English speaking country, %	Top country	Top country	Top country	AgeRang	percent	AgeGroup	AG_percent
3000	MELBOURNE	China	32927	68.2	58.7	China	6374	12.8	rang0n4	2.3	Children	3.9
3002	EAST MELBO	England	5108	30.8	16.9	England	414	4.9	rang0n4	3.6	Children	6.2
3003	WEST MELBC	China	4135	43.3	34.6	China	394	6.1	rang0n4	3.3	Children	6.5
3006	SOUTHBANK	China	13424	60.6	47.3	China	1796	9	rang0n4	3	Children	4.9
3006	SOUTH WHA	China	75	34.7	28.6	China	8	8.2	rang0n4	0	Children	0
3008	DOCKLANDS	China	6640	56.1	43.2	China	938	9.8	rang0n4	3.6	Children	5.9
3011	FOOTSCRAY	Vietnam	14149	52.9	45.3	Vietnam	2500	10.9	rang0n4	6	Children	12
3011	SEDDON	Vietnam	5077	34	25.1	Vietnam	550	6	rang0n4	9.3	Children	17.7
3012	WEST FOOTS	Vietnam	10946	42.7	36.3	Vietnam	1652	8.7	rang0n4	7.7	Children	16.2
3012	MAIDSTONE	Vietnam	8295	50.5	45.8	Vietnam	2346	16.5	rang0n4	6.2	Children	15.6
3012	BROOKLYN	Vietnam	1728	43.2	37.2	Vietnam	158	5.2	rang0n4	6	Children	12.8
3012	KINGSVILLE	Vietnam	3737	32.3	23.8	Vietnam	230	3.6	rang0n4	10.3	Children	18.9
3013	YARRAVILLE	England	14503	28.4	19.4	England	934	3.6	rang0n4	8.2	Children	18.1
3015	NEWPORT	England	12653	24.1	15	England	832	3.7	rang0n4	9	Children	19.9
3015	SPOTSWOOI	England	2473	27.9	19.3	England	160	3.6	rang0n4	7.1	Children	15.6
3015	SOUTH KING	England	1925	31.1	22.6	England	106	3	rang0n4	8.2	Children	16.4
3016	WILLIAMSTO	England	13830	21.6	11.3	England	1414	5.6	rang0n4	6.5	Children	20.3
3016	WILLIAMSTO	England	1626	28.2	15	England	166	5.5	rang0n4	3.8	Children	14.8
3018	ALTONA	England	10429	32.1	19.7	England	1100	5.8	rang0n4	5.8	Children	14.8
3018	SEAHOLME	England	1957	31.1	19.5	England	202	5.6	rang0n4	5.3	Children	14.8
3019	BRAYBROOK	Vietnam	8760	57.4	54.1	Vietnam	3480	23.7	rang0n4	8	Children	20.2
3020	SUNSHINE W	Vietnam	17878	55.4	51.4	Vietnam	4284	13.7	rang0n4	5.9	Children	17.2
3020	SUNSHINE	Vietnam	9365	51.1	46.3	Vietnam	1978	12.2	rang0n4	7.1	Children	16.2
3020	SUNSHINE N	Vietnam	11284	57.8	55.1	Vietnam	4922	24.8	rang0n4	6	Children	17.8
3020	ALBION	India	4527	54.7	49.4	India	1152	14.5	rang0n4	6.7	Children	15.8

2.2 Property data

The following tabular data mainly shows the property price of each suburb from 2005 to 2015.

Suburb	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Prelim 2016	2014-2015	2005-2015	2005-2015
ABBOTSFORD	427000	448000	602500	600000	638500	700000	704000	700000	773000	862500	925000	771500	7	117	8.0
ABERFELDIE	534000	600000	735000	710000	775000	1046500	1003000	852500	947500	1045000	1185000	1200000	13	122	8.3
AIREYS INLET	459000	440000	430000	517500	512500	606000	680000	634000	664000	625500	677500	730000	8	48	4.0
AIRPORT WEST	311500	320000	381000	421000	455000	570000	555500	495000	530000	573500	630000	689500	10	102	7.3
ALBANVALE	210000	210000	215000	252000	280000	320000	317000	310000	313000	326500	345000	390000	6	64	5.1
ALBERT PARK	681000	777000	986500	1125000	1050000	1155000	1390000	1260000	1360000	1503000	1705000	1547500	13	150	9.6
ALBION	216000	235000	262000	322500	349000	440000	400000	379500	380000	432000	472000	527000	9	119	8.1
ALEXANDRA	200000	190000	207000	210000	220000	230000	255000	239000	261000	247500	285000	248000	15	43	3.6
ALFREDTON	283000	307500	300500	300000	300000	316000	333000	340000	360000	362000	340000	350000	-6	20	1.9
ALLANSFORD	210000	240000	230000	232000	242500	250500	330000	257500	282500	242500	259000	242000	7	23	2.1
ALPHINGTON	552500	687500	740000	804000	832500	1000000	992500	1000000	1107500	1028000	1271500	747500	24	130	8.7
ALTONA	330000	370000	416000	480000	499000	627000	510000	555000	570000	645000	715000	667500	11	117	8.0
ALTONA EAST	295000	310000	391500	412500	435000	560000	528000	550000	552000	607500	656000	NA	8	122	8.3
ALTONA MEADOWS	232000	252000	280000	310000	346500	386000	395000	375000	385000	413000	439000	452500	6	89	6.6
ALTONA NORTH	277000	290000	355000	401000	417000	515000	520000	487500	510000	553500	637500	531000	15	130	8.7
ANGLESEA	391000	410000	445000	455000	516500	580000	600500	595000	550000	612500	637000	652500	4	63	5.0
APOLLO BAY	350000	360000	415000	362500	430000	435000	495000	432500	450000	460000	435000	455000	-5	24	2.2
ARARAT	148000	145000	148500	145000	159500	174500	169000	165000	179500	201000	210000	240000	4	42	3.6
ARDEER	206000	214500	234000	275500	312000	382500	360000	342000	330000	360000	381500	455000	6	85	6.4
ARMADALE	601000	900000	1250000	1415000	1605000	1690000	1546500	1505000	1651000	1750000	2160000	1780000	23	170	10.4
ARMSTRONG CREEK	-	-	-	-	-	-	205000	362500	467500	441500	440000	457000	0	NA	NA
ARTHURS SEAT	370000	275000	385000	442500	410000	470000	662500	715000	480000	705000	820000	NA	16	122	8.3
ASCOT (BENDIGO)	285000	263000	270000	275000	279000	330000	322000	350000	349500	339000	350000	340000	3	23	2.1
ASCOT VALE	425000	480000	600000	667000	622500	750000	734000	700000	742500	809500	880500	880000	9	107	7.6

The raw data file is the .xlsx file. When the data is manually saved as .csv file and then read into RStudio, there are two main problems occurring. The first is the columns name cannot be recognized or shown as wrong format, such as “X2005”. Another problem is the property price should be numerical, but after reading into RStudio, the type of value is factor, even though the result is the same. As the result of wrong data type, there is a disorder problem of Y axes confusing me when implementation of data visualization.

```
> class(raw_propertyData$X2005)
[1] "factor"
> raw_propertyData
                Suburb  X2005  X2006  X2007  X2008  X2009  X2010  X2011  X2012  X2013  X2014  X2015
1           ABBOTSFORD 427000 448000 602500 600000 638500 700000 704000 700000 773000 862500 925000
2           ABERFELDIE 534000 600000 735000 710000 775000 1046500 1003000 852500 947500 1045000 1185000
```

Hence, the result of data wrangling shows as following.

Suburb	year	price
ABBOTSFORI	2005	427000
ABERFELDIE	2005	534000
AIREYS INLET	2005	459000
AIRPORT WE	2005	311500
ALBANVALE	2005	210000
ALBERT PARI	2005	681000
ALBION	2005	216000
ALEXANDRA	2005	200000
ALFREDTON	2005	283000
ALLANSFORD	2005	210000
ALPHINGTON	2005	552500
ALTONA	2005	330000
ALTONA EAS	2005	295000

2.3 Crime data

The crime data is collected to attempt to analyse the relationship with property price. And the following tabular data shows the raw crime data.

Table 9. Offences recorded by postcode - July 2011 to June 2016

	Offences recorded					
	Jul 2011 - Jun 2012	Jul 2012 - Jun 2013	Jul 2013 - Jun 2014	Jul 2014 - Jun 2015	Jul 2015 - Jun 2016	% change 2015 - 2016
Postcode						
3000	21,862	24,674	21,309	22,835	22,732	-0.5%
3002	917	741	1,019	763	833	9.2%
3003	476	639	511	538	661	22.9%
3004	167	163	142	155	119	-23.2%
3006	2,041	2,136	2,344	2,535	3,202	26.3%
3008	941	855	1,002	1,580	1,831	15.9%
3011	3,413	3,178	3,193	2,810	3,328	18.4%

In fact, there are a large amount of data relating to crime with more specific information, such as types of victims, but for comparison between crime and property price by year, the data containing suburb postcode and the amount data of offences from 2011 to 2016 is enough. Here is a data flaw identified. It is the difference of year range in crime

data and property data. However, the crime data with same content from 2005 to 2011 has been not found.

For data wrangling, the first step is to refine the data from the raw data file. The following is the refined crime data file.

postcode	2011	2012	2013	2014	2015
3000	21,862	24,674	21,309	22,835	22,732
3002	917	741	1,019	763	833
3003	476	639	511	538	661
3004	167	163	142	155	119
3006	2,041	2,136	2,344	2,535	3,202
3008	941	855	1,002	1,580	1,831
3011	3,413	3,178	3,193	2,810	3,328
3012	2,390	1,952	1,965	1,978	2,211

Similarly, when the data is read in to RStudio, there is the data formation and column problem coming up. Since similar problems have been coped with, it should be more easy to solve the problems. Hence, here is the wrangled crime data.

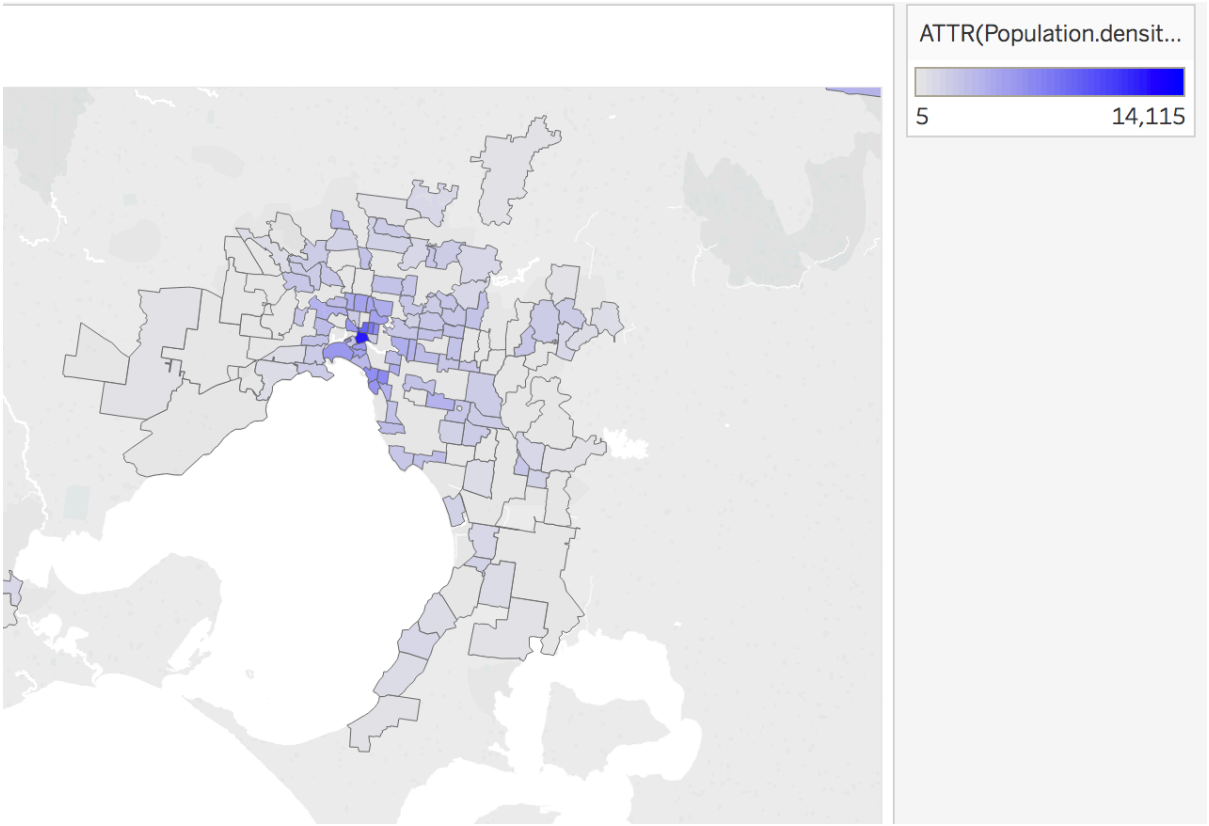
POSTCODE	year	occurence
3000	2011	21,862
3002	2011	917
3003	2011	476
3004	2011	167
3006	2011	2,041
3008	2011	941
3011	2011	3,413
3012	2011	2,390
3013	2011	977
3015	2011	1,098
3016	2011	1,207

3. Data exploration

Currently, Tableau is popular and famous data visualization tool. Thus, it is used for initial data exploration.

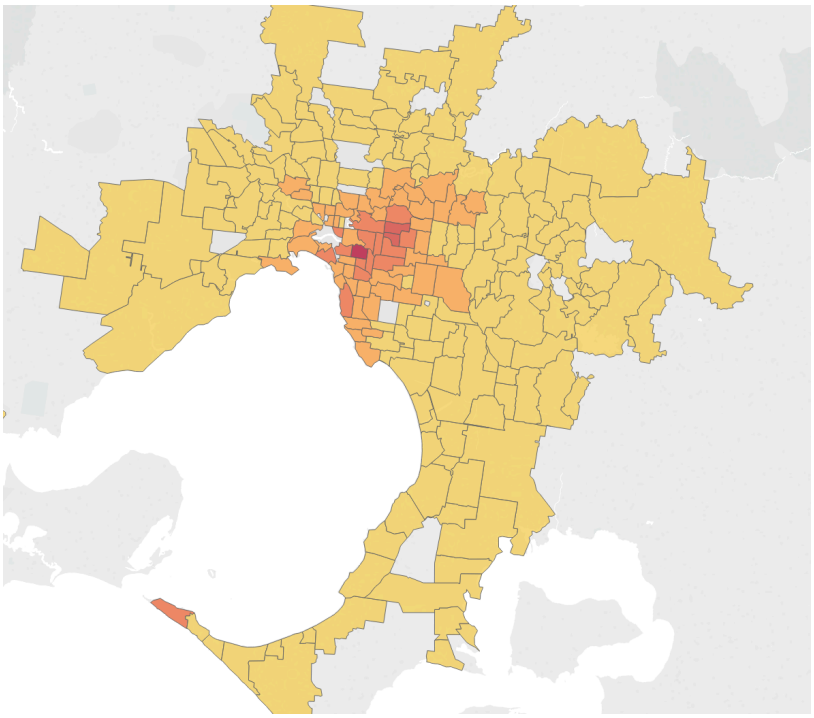
3.1 population geographic distribution

The following figure give the information of population geographic distribution. It can be observed that the population density distribution for each suburb a bit random. However, the colour in the most central area is deeper than the outside. Next, after comparison with other figures, there might be more clues coming out.



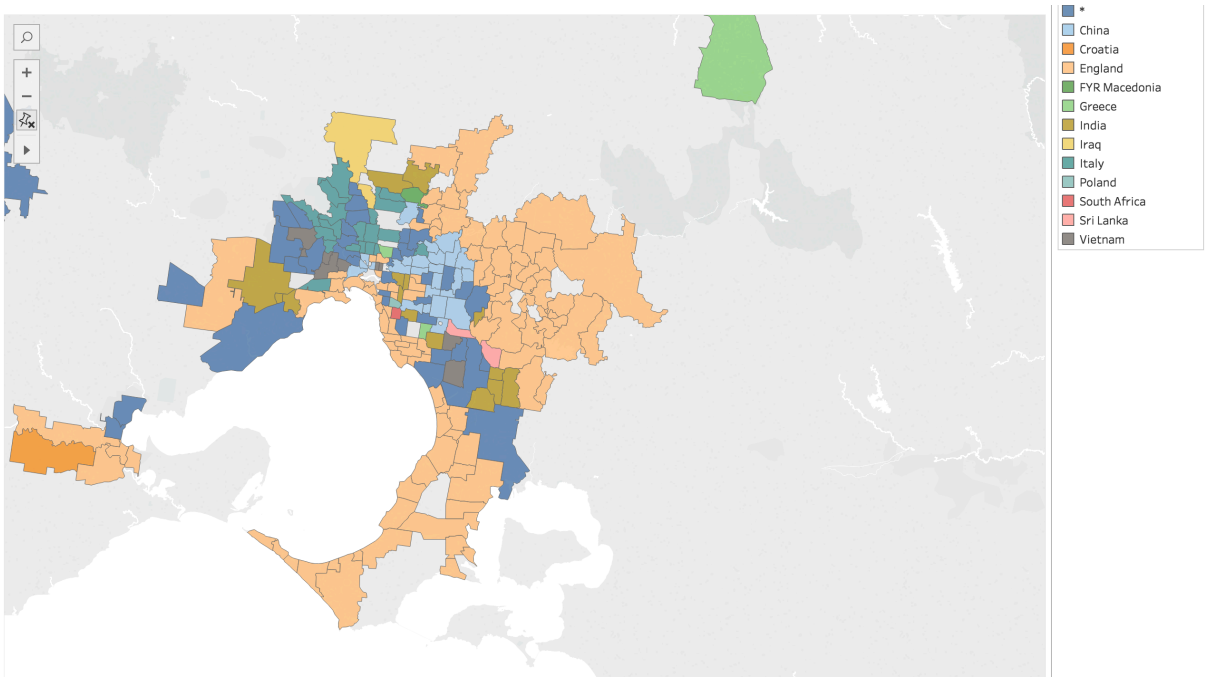
3.2 property geographic distribution

The following figure shows the property geographic distribution for each suburb. The red area stands for the high property price, while the yellow section means lower price. From the figure, it is easily identified that suburbs with high property price concentrate in the centre and the most of outer suburbs have low property price.



3.3 Origin country distribution

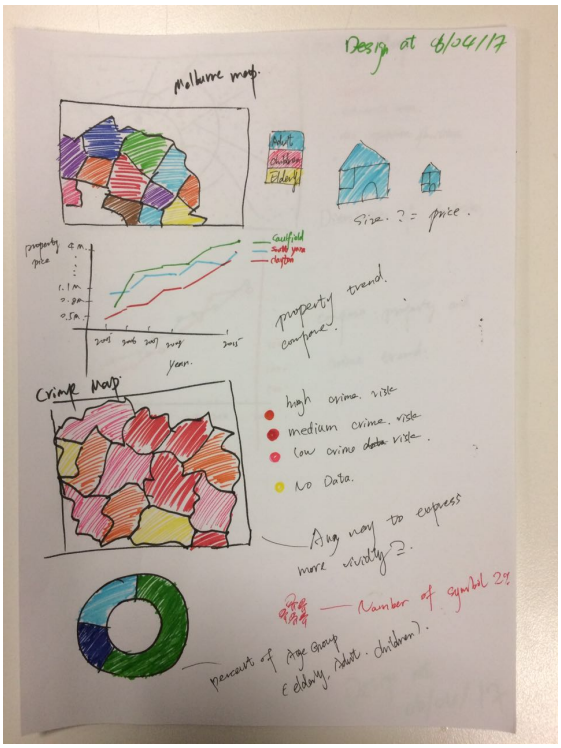
The following figure try to geographically reveal the distribution of origin country for each suburb. There are some results observed from it. Firstly, it can be easily observed from the map that the colour of most area is orange, which means England is the most popular origin country in Melbourne, meanwhile they are distributed outside. Secondly, there is a cluster of light blue area in the centre, so it seems that the people from China like to reside in centre. Thirdly, obviously there are a few of cyan area in the northwest of map, which means that most of Italian gather in the northwest region.

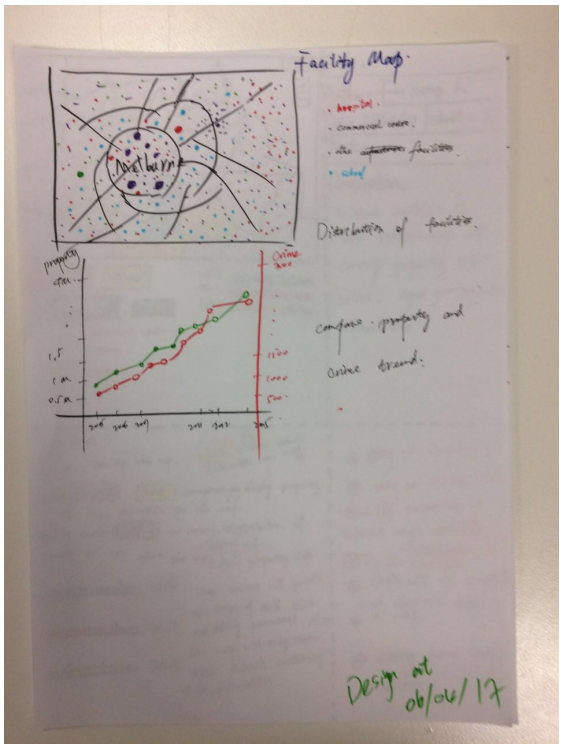


4. 5 sheet design

4.1 Brainstorm Sheet

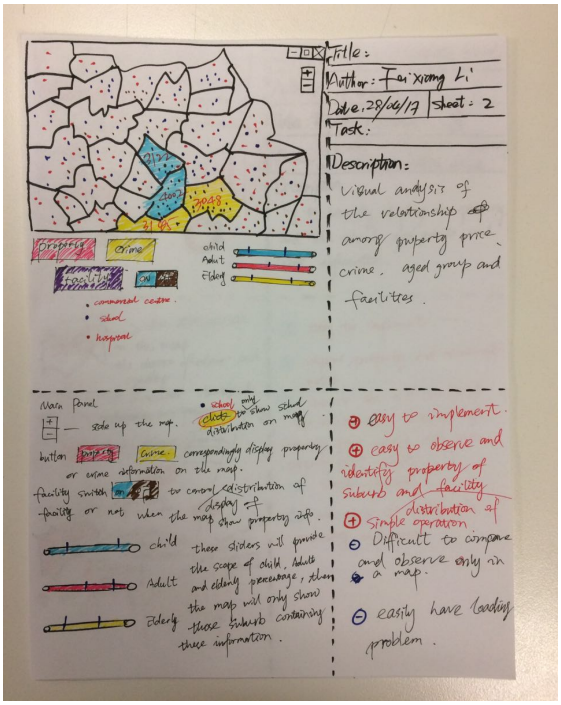
In this stage, based on a range of data that have been gathered, their associated parts are extracted after observation and analysis of those data. Then, some ideas are generated for visual information presentation to reveal their relationship. Coming next is the introduction of brainstorm sheet. The polygon map is deployed to illustrate the any information relating to geographic distribution, such as property price by suburb. The line graph is used to analyse the trend of property price and crime offenses. Moreover, there are some tiny interesting symbols, for example the light blue house symbol whose size stands for the property price.

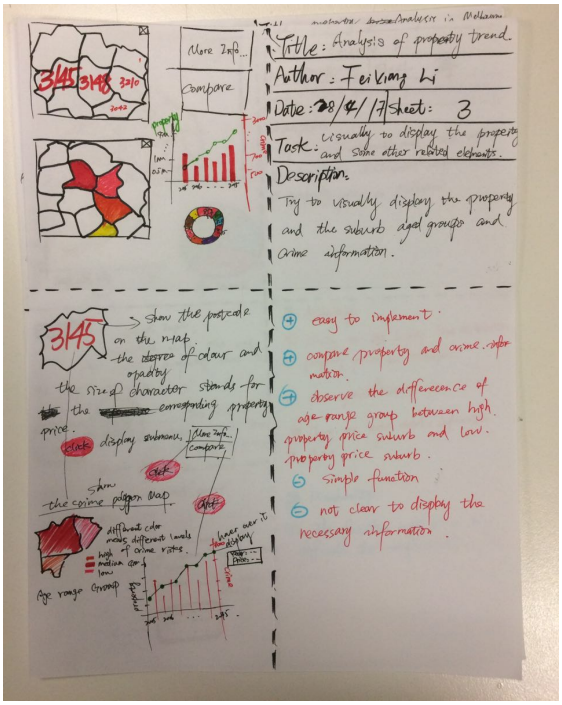


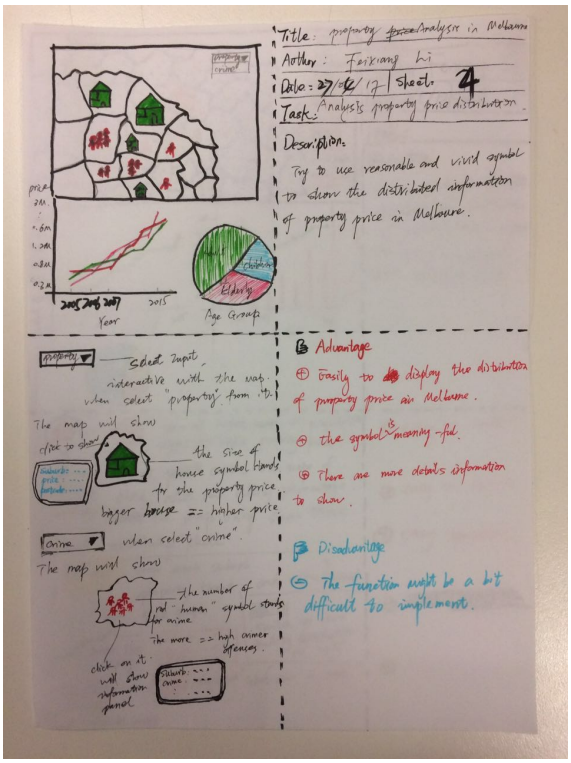


4.2 Initial designs

In terms of initial designs, sheet2, sheet3 and sheet4 respectively provide constructive, special, interesting and specific designs. All of they have their own advantages and disadvantages as well. More specific design information can be found in the following three sheets.

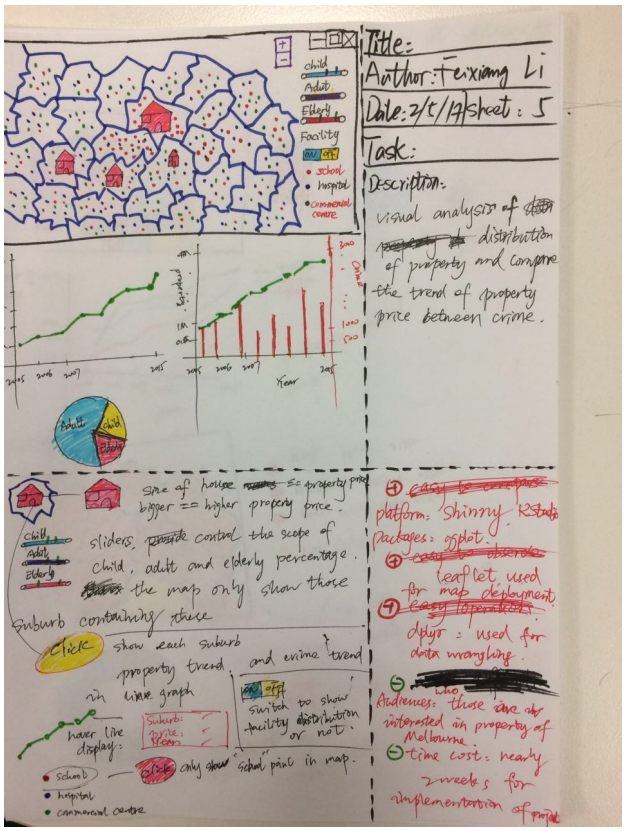






4.3 Sheet 5

In five sheet design methodology, sheet 5 will provide more discussion in detail and make some explanation how to implement the design. Besides that, it also shall give description of algorithms, packages, audiences and analysis of estimates of cost including time cost and human resource cost and so on. The implementation of sheet5 design is based on shinny and RStudio. It provides a few of operations for observation of the alteration in map. Additionally, for each suburb, it implements line-trend graph and pie graph for comparison. As a matter of fact, the sheet 5 design as final design should be completely implemented. However, due to the technological or other problems, the functions are implemented in other ways, which will be descripted in the implementation section.



5. Implementation

5.1 Data Wrangling

Software & Platform	description
Excel	Generally used to gather appropriate data from raw data and for simple data manipulation, such as renaming column names, deleting columns or add new columns and so on
RStudio: R Packages: tidyr、dplyr	RStudio read those raw .xlsx data files, and then filter and reorganize them. At last, these data files are resaved in .csv file.

Tableau	Tableau is a popular data visualization tool. Here, it used to detect the data errors、clean data and for initial data exploration.

5.2 Implementation of project

Development of Environment: RStudio

RStudio is an open-sourced, free and currently the most popular integrated development environment for R language. Its main usage is for statistical computing and data visualization. More significantly, there are a huge amount of libraries providing support for various and creative design. Therefore, in terms of scalability and stability, RStudio is appropriate for my project development.

Shiny: (RStudio, Shiny by RStudio, 2016)

Shiny is a web application frame for R. It is easily understandable and quickly entry, even though those without any background knowledge of web application development are able to turn their inspirations and analysis into an interactive web application. More significantly, there are a lot of excellent examples and amazing sample application online for learning and reference.

Leaflet packages: (RStudio, Leaflet for R, 2014-2016)

Leaflet is the mainstream open-source and light JavaScript library for mobile–friendly interactive maps and it contains all the mapping features most developer ever need. Moreover, it works efficiently across all major platforms and can be extended with lots of plugins.

Plotly packages: (Plotly, 2015)

Plotly is an online data analytics and visualization tool for agile business intelligence and data science. Most importantly, it provides support for the mainstream data analysis and visualization language, such as Python, R and Arduino and so on. Based on its services, a wide variety of characterized and creative graphs can be created for data visualization and analysis.

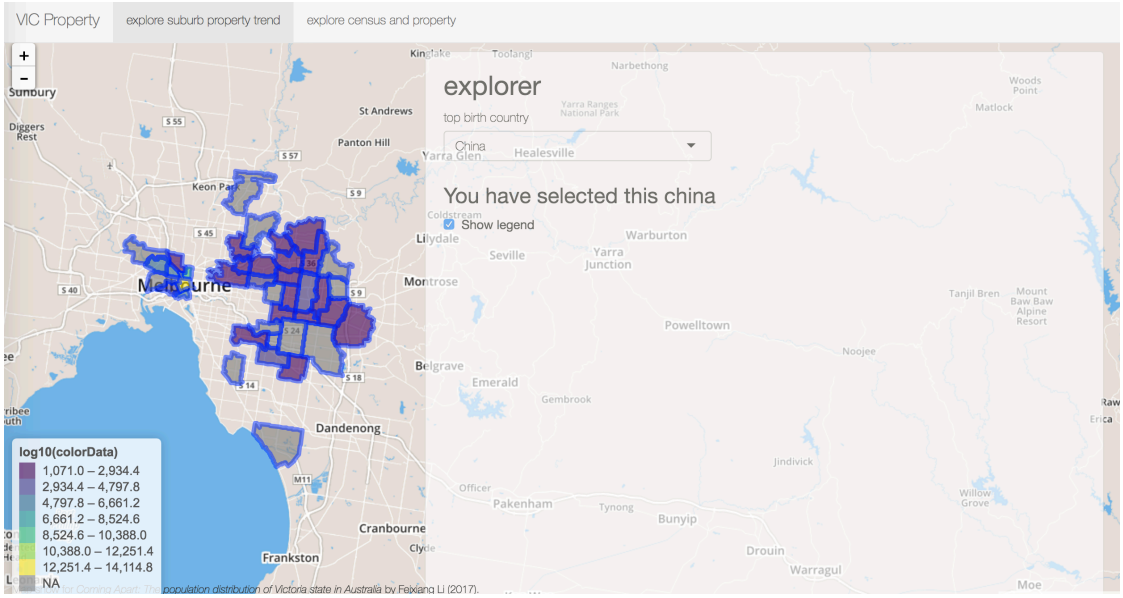
Rgdal and Rmapshaper packages:

The Rgdal and Rmapshaper packages contribute to map the spatial data in R. In this project, usage of both is mainly for polygon of the Melbourne suburbs.

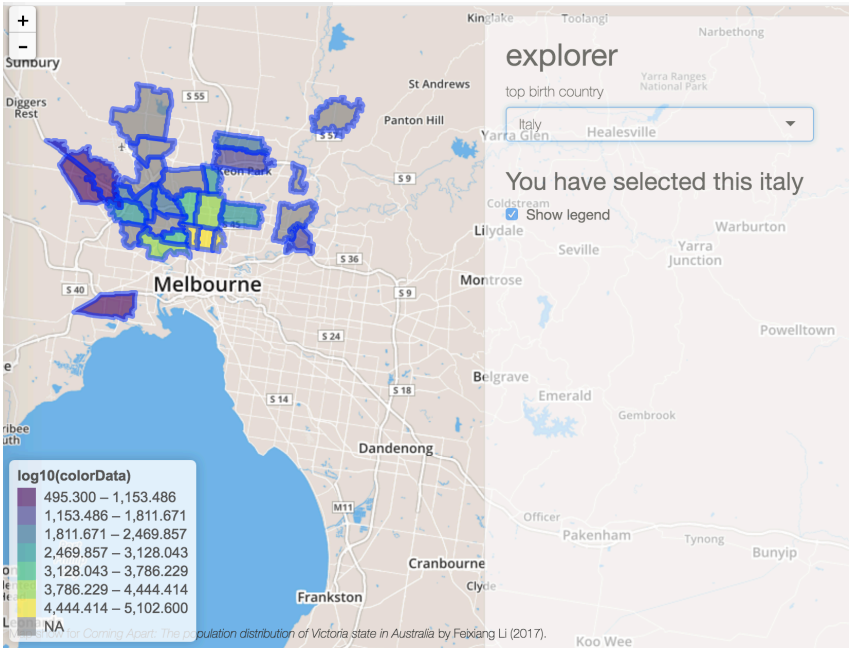
6. Instructions for viewing and exploring

Initial launch of project will show the following page. Please be patient to wait for loading the data until the suburb polygon map showing up. Otherwise, the program cannot run appropriately.

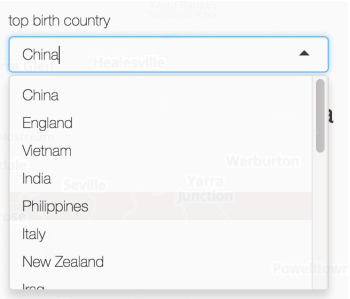
6.1 Tabpanel “explore suburb property trend”

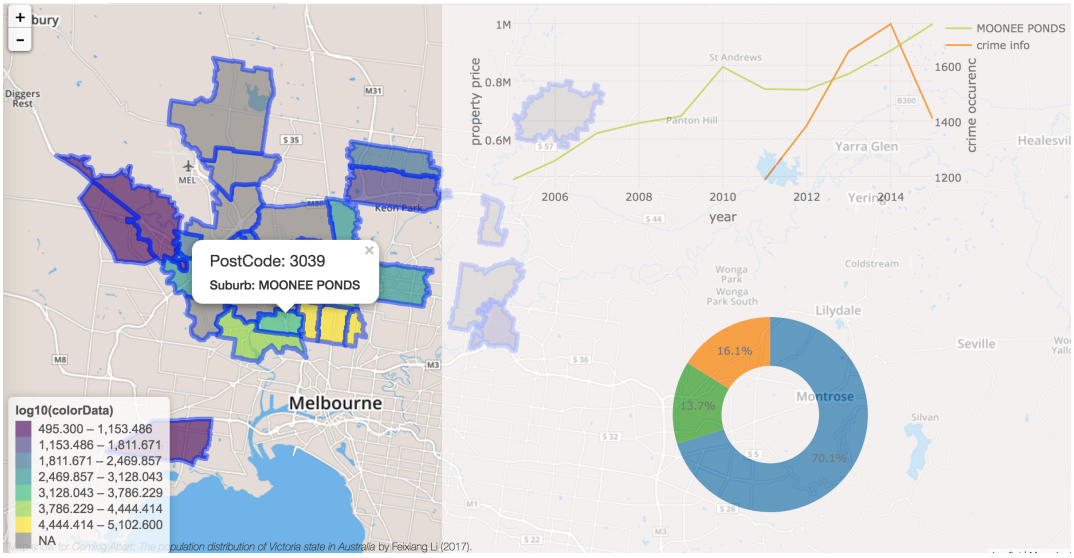


In the explorer panel, the select input is used to filter top origin country and then show corresponding suburbs on the map. For example, by default the map will illustrate the suburbs containing most residents from China and when Italy is selected, the map will show as following.

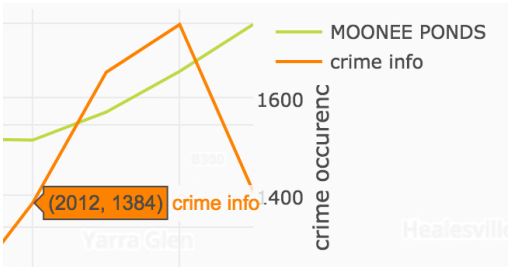


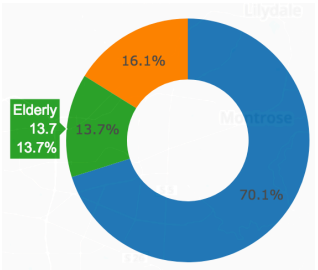
When you zoom in the map, it will display more clear suburb polygon. Then if you casually click one of suburbs, the pop-up will give this suburb postcode and its name, meanwhile there will be a line-trend graph and pie chart coming up in the explore panel. The line-trend graph will show the property price trend of this suburb from year 2005 to year 2015 and its crime trend from year 2011 to 2015. In spite of it illustrate the trend of property and crime for user, it also allows user to observe whether the crime have any influences on the property. In other words, the security of suburb possibly affects its property. Then, the pie chart will illustrate the population composition.





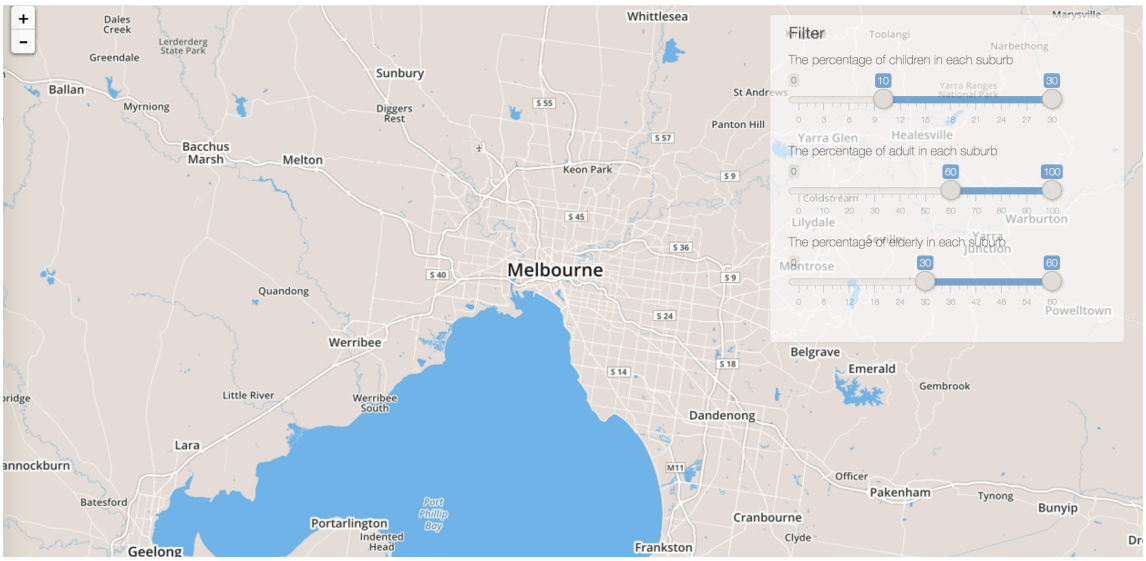
When you hover the line and the section of donut pie chart, the specific details will display as followings.



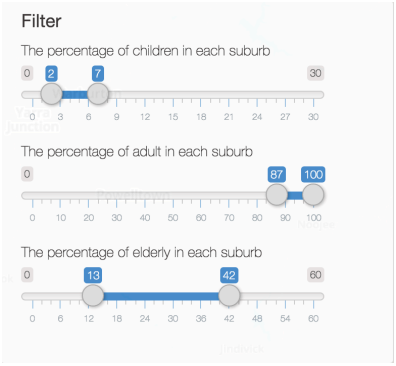


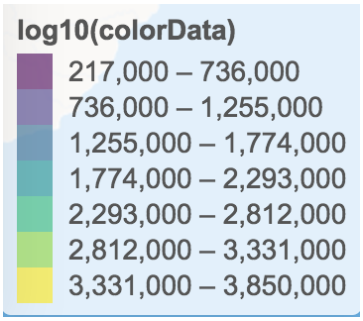
6.2 Tabpanel “explore census and property”

Initial launch of this tab will display nothing, even though the slider has been given the default values.

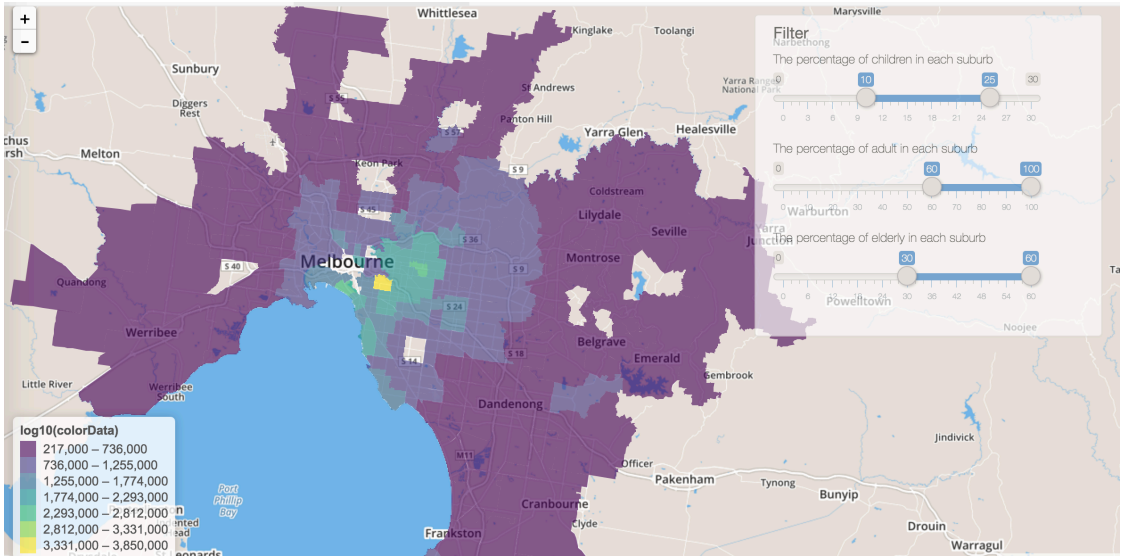


The Filter panel provides three sliders corresponding to the children percentage, adult percentage and elderly percentage. Then, the leaflet map will show suburbs based on the scopes of sliders. The colour legend stands for the degree of property price used for observation.



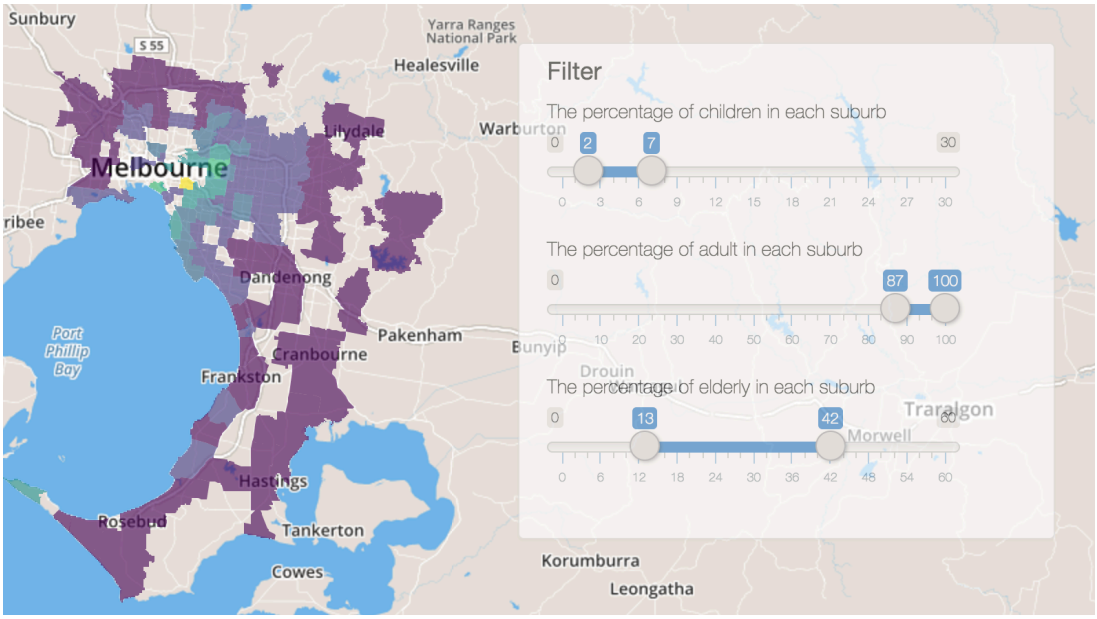


When you adjust one of the sliders, the map will only display appropriate suburbs meeting conditions.



It can be observed that when the adult percentage is between 60 and 100, even if children slider or elderly slider make some changes, the map is nearly unchanged. Therefore, for most of suburbs, its main population is adult and the percentage is between 60 and 100. This can be proved by directly filtering data in R. However, when the percentage

of adult is set between 87 and 100 as following figure shows, then it is obvious that the content of map is different.



7. Conclusion

In this report, we have reviewed the process of development and then looked at a wide variety of different graphics that have been used to visualize multivariate data. In this assignment, the main product is the interactive web application for data visualization with two data exploring function. There are some outcomes and results identified after analysis of data exploration and visualization. These outcomes and results are mentioned above. Although the results and findings are inconsistent with initial expectations, however, it is more important that what we have learnt from this assignment. For myself, after reviewing the program I have learnt that

- clear and profound understanding of data visualization
- how to begin with identification of interested topics
- and then find out the related data source

- analysis of raw data and data wrangling to acquire useful data
- initial data exploration and try to figure out effective and efficient ways for data visualization
- research development and design tools for data visualization
- acquirement of programing skills in R and knowledge of Shinny framework

In summary, through this assignment, I have been familiar with the basic methodology of data visualization and popular data analysis and visualization tools.

References

Plotly. (2015). *Plotly R Library*. Retrieved from https://plot.ly/r/

Reed, & Richard. (2016). The relationship between house prices and demographic variables. *International Journal of Housing Markets and Analysis, 9*(4), 520-537.

RStudio, I. (2014-2016). *Leaflet for R*. Retrieved from https://rstudio.github.io/leaflet/

RStudio, I. (2016). *Shiny by RStudio*. Retrieved from Shiny Gallery: https://shiny.rstudio.com/gallery/

Tullberg, J. (2007). Vic: Melbourne property looking hot. *AAP General News Wire; Sydney*, 1.