Finding the Best Location to Live

for new migrant families

Coursera Applied Data Science

Capstone Project

Kamin Arifin

11 June 2019

Table of Contents

1. Introduction

As an immigrant myself, I still remember very clear the struggle that I was experiencing to find a house when I migrated to Australia. My family and I chose Melbourne as a new place to call home.

As a new immigrant, we preferred to live in an area that we were more familiar/comfortable with, e.g. in an area where we have the same background, ethnicity, religion or language.

The data used in this project is obtained from an Australian government website (The Australian Bureau of Statistics).

The main sources of data obtained from the Australian Bureau of Statistics which will be used in this project are:

- Population and People, ASGS, 2011 to 2018
- Education and Employment, ASGS, 2011 to 2018

The data obtained is presented in an Excel Format.

Optionally, a list of venues could be obtained through Foursquare (for further enhancement).

2. Data Analysis

The ABS provides a great amount of data regarding the Australian population. Using this data, we will be able to review the ages, education, religion(s) and country of origins of the people in each suburb (or areas).

Aside from census data from the ABS, we can also enrich our data by using a list of venues in those suburbs.

By combining all this data, we will be able to see which suburb is the most suitable for a new immigrant.

As a result, new immigrants can feel more secure, adapting more easily to their new place that they can now call 'home'.

The data will be grouped (clustered) into 2 categories.

Firstly, according to religion(s), and next, based on countries of origin.

'K-Means Clustering' will be the method used to find the list of suburbs based on the above criteria.

K-Means is one of the unsupervised clustering algorithms used to group a set of data based on their similarity.

The 'Elbow method' will be used to find the right 'K' in K-Means clustering.

2.1 Religions Data

The ABS only provides 5 major religions in their data which are the following:

- Christian
- Buddhist
- Hindu
- Islam
- Judaism

2.2 Country of Origin

- Oceania (Except Australia)
- North West Europe
- Southern Eastern Europe
- North Africa and Middle East
- South East Asia
- North East Asia
- Southern Central Asia
- America
- Sub-Saharan Africa

3. Methodology

Our main goals is to group (cluster) Victoria (especially Melbourne) suburbs based on their religions and countries of origin.

The following libraries are used for pre-processing, clustering and displaying Victoria map.

Library	Functions
	Scikit is a machine learning library with many functions such as pre-processing, Classification and Regression. I will use this library for clustering (K-Means)
	Folium is a good library to visualize data in Python on an interactive leaflet map. I will use this to display a map of Victoria or Melbourne.
	GeoPy will be used to find the coordinates of addresses, cities, or to measure distance from one location to the another. It has a simple interface and is easy to use.

3.1 Data Source and Columns

Religious Data and Country of Origin comes from different spreadsheets. The following Excel spreadsheets from the ABS were downloaded to local data spaces:

- Population and People, ASGS, 2011 to 2018
- Education and Employment, ASGS, 2011 to 2018

We need to merge these 2 datasets and only use the relevant columns and rows.

- Merge them as a single Python Dataframe and use "Suburb" to join those 2 datasets.
- Data from the year 2016 is used as part of this analysis, as it has the most complete data.

Below is the list of columns that I used after the merging:

['Suburb', 'YEAR', 'Median Age', 'Born_in_Oceania_Ex_OZ', 'Born_in_North_West_Europe', 'Born_in_Southern_Eastern_Europe', 'Born_in_North_Africa_Middle_East', 'Born_in_South_East_Asia', 'Born_in_North_East_Asia', 'Born_in_Southern_Central_Asia', 'Born_in_America', 'Born_in_Sub_Saharan_Africa', 'Born_Overseas', 'Relg_Buddhism', 'Relg_Christianity', 'Relg_Hinduism', 'Relg_Islam', 'Relg_Judaism', 'Relg_Other', 'Relg_Secular', 'Relg_not_stated', 'Residency_Citizen', 'Residency_not_Citizen', 'Residency_not_Stated', 'Edu_PostGraduate', 'Edu_Diploma', 'Edu_Bachelor', 'Edu_Adv_Dipl', 'Unempl_Rate']

3.2 Pre-processing:

For better results, I did the following pre-processing before clustering.

- I standardised the data using sklearn.preprocessing.scale function.
- I then scaled back the data so that it would have values from 0 to 1.00 (or from 0 to 100%).

I used GeoPy to find the coordinates of all suburbs.

This data is required when working on the map of Victoria/Melbourne.

3.3 Clustering by Religion

Only 5 religions are included in this analysis: Christian, Buddhism, Hinduism, Islam and Judaism. The rest of religions is too small to be included.

The actual data is in percentage as comparison to other religions. To get a better presentation of minority of other regions aside from Christianity, we need to standardise the data. Following that I scale the data back so its value will be a range from 0 to 1 (zero to one).

As we involve 5 religions, I would think at least there are 3 clusters. So finally I use three k values that will be reviewed, k=4, k=6 and k=7.

The clustering method that's being used is K-Means from Scikit library. Please refer to section "4.1 Religious Distribution" for the output from each individual k value.

Output from K-Means clustering by Religion with k = 7

Cluster	Relg_Christianity	Relg_Buddhism	Relg_Hinduism	Relg_Islam	Relg_Judaism
0	0.750826	0.032608	0.039183	0.017365	0.004800
1	0.443708	0.070617	0.067281	0.037145	0.027416
2	0.541608	0.162846	0.602525	0.185838	0.018618
3	0.706978	0.131553	0.228068	0.112476	0.012542
4	0.550288	0.541725	0.216377	0.158887	0.001684
5	0.578279	0.090572	0.207885	0.704962	0.000541
6	0.301128	0.043939	0.116129	0.012136	0.937956

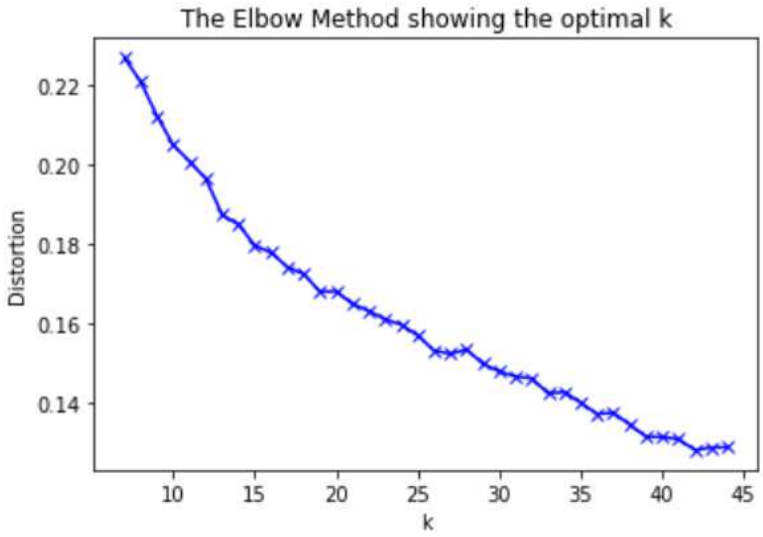
We can see from the chart above, almost all Judaism believer, 94% of them live in Cluster-6, which are suburbs Caulfield North and Caulfield South.

3.4 Clustering by where Country of Origin

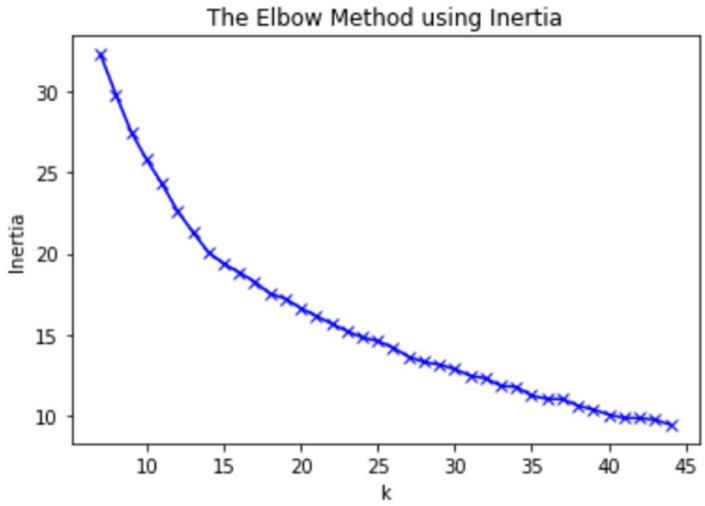
Different from the previous dataset, for country of origin dataset, I used the Elbow Method to obtain an optimal k value.

To get a better result, I compare the following 2 of Elbow models:

Using Distortion model:



Using Inertia model:



According to the chart, the optimal k value is somewhere between 16 to 18. I am going to use 16 as my clustering (country of origin).

Output from K-Means clustering with k = 16

Country of Origin Cluster	Oceania	North West Europe	Southern Eastern Europe	North Africa / Middle East	South East Asia	North East Asia	Southern Central Asia	America	Sub Saharan Africa
0	0.389925	0.187066	0.422722	0.107955	0.279762	0.096085	0.351685	0.317308	0.345766
1	0.182937	0.32714	0.066769	0.009119	0.021879	0.011727	0.020419	0.080042	0.056452
2	0.203358	0.248843	0.255337	0.077214	0.204696	0.471504	0.226489	0.176282	0.202957
3	0.455224	0.486111	0.231351	0.043082	0.102891	0.095443	0.070421	0.46978	0.216014
4	0.672388	0.192708	0.209606	0.142133	0.167196	0.067241	0.631818	0.201923	0.398387
5	0.353234	0.123264	0.409278	0.07838	0.679894	0.085728	0.276385	0.219551	0.291667
6	0.21791	0.168519	0.721511	0.160373	0.142504	0.079119	0.171996	0.169231	0.109677
7	0.222258	0.349034	0.199293	0.046063	0.112664	0.165605	0.124029	0.216973	0.178822
8	0.202822	0.523592	0.082597	0.010585	0.031057	0.02102	0.022888	0.114067	0.094344
9	0.385928	0.358135	0.24349	0.100899	0.13681	0.140805	0.110166	0.39011	0.732719
10	0.282183	0.222873	0.439655	0.1191	0.1021	0.095097	0.184757	0.199519	0.165827
11	0.324627	0.085069	0.211823	0.868881	0.103175	0.018678	0.214734	0.144231	0.21371
12	0.129077	0.168853	0.040777	0.009907	0.023908	0.012346	0.027575	0.045228	0.051075
13	0.300719	0.674383	0.164933	0.026159	0.038213	0.042784	0.048299	0.235043	0.197133
14	0.542999	0.31746	0.173352	0.067766	0.117536	0.041598	0.222571	0.173993	0.277266
15	0.28607	0.303241	0.110016	0.061772	0.283951	0.715517	0.175026	0.689103	0.236559

4. Result

In this project, I am interesting to know how the distribution of population based on their religions and countries of origin. I will analysed both aspects and review the result.

4.1 Religious Distribution

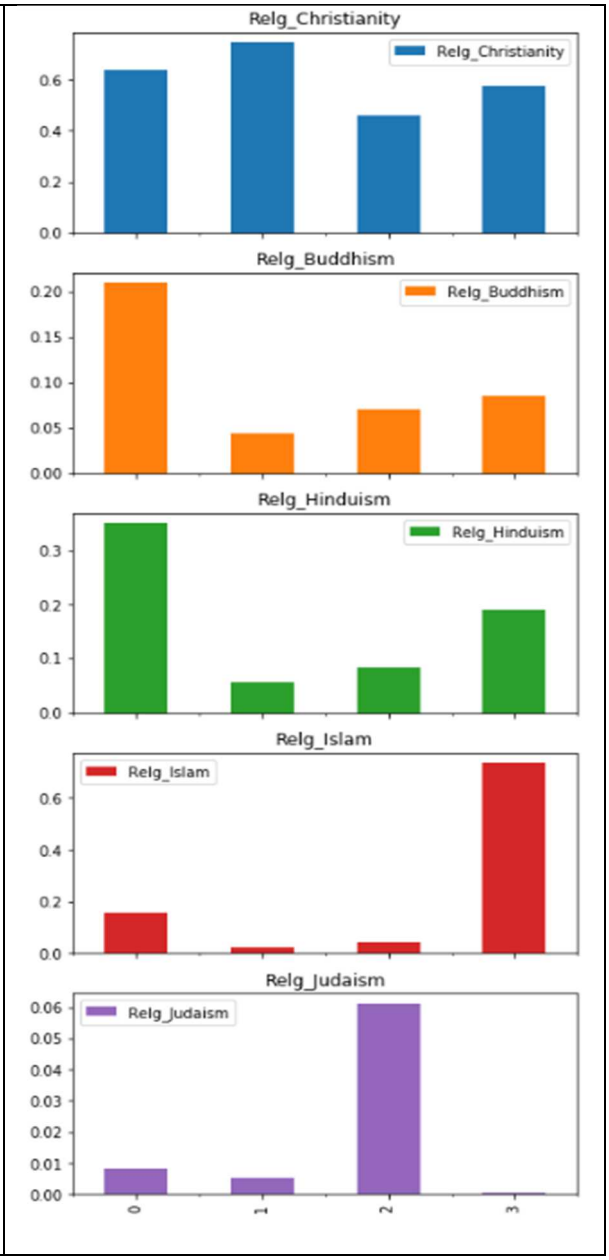
For better understanding of how population is distributed in Victoria, I am going to use a few different values of k: k=4, k=6 and k=7

Religious Distribution – 4 Clusters

The charts on the right show Victoria's Religious distribution, with four Clusters.

We can clearly see that the Christian population is evenly distributed in almost all clusters, but not with other religions.

The Muslim and Hindu population fall under the same cluster, cluster-0.



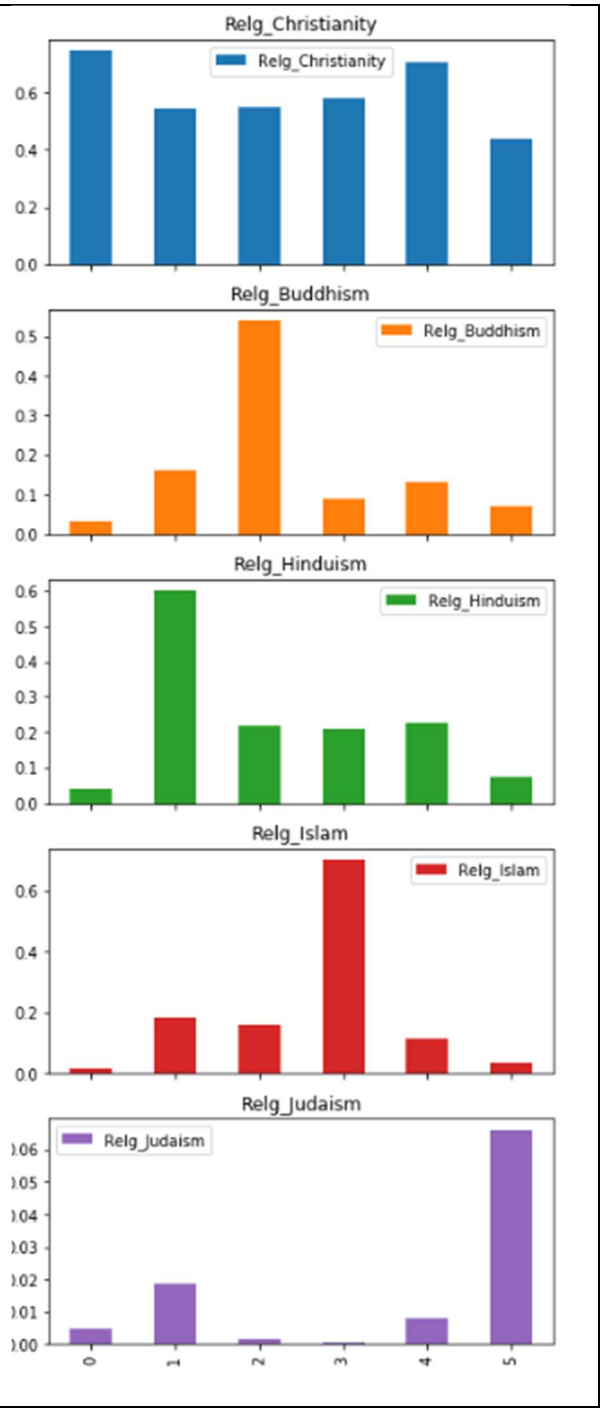
Religious Distribution – 6 Clusters

With 6 Clusters, we get a better picture of Victoria's Religious population distribution.

Contrary to our previous conclusion, we can now see that Hindu and Muslim (Islam) are actually in different clusters.

Christians still maintain the same result with the population being evenly distributed across all suburbs.

Most Judaism believers live in cluster 5, with a few living in cluster 1.



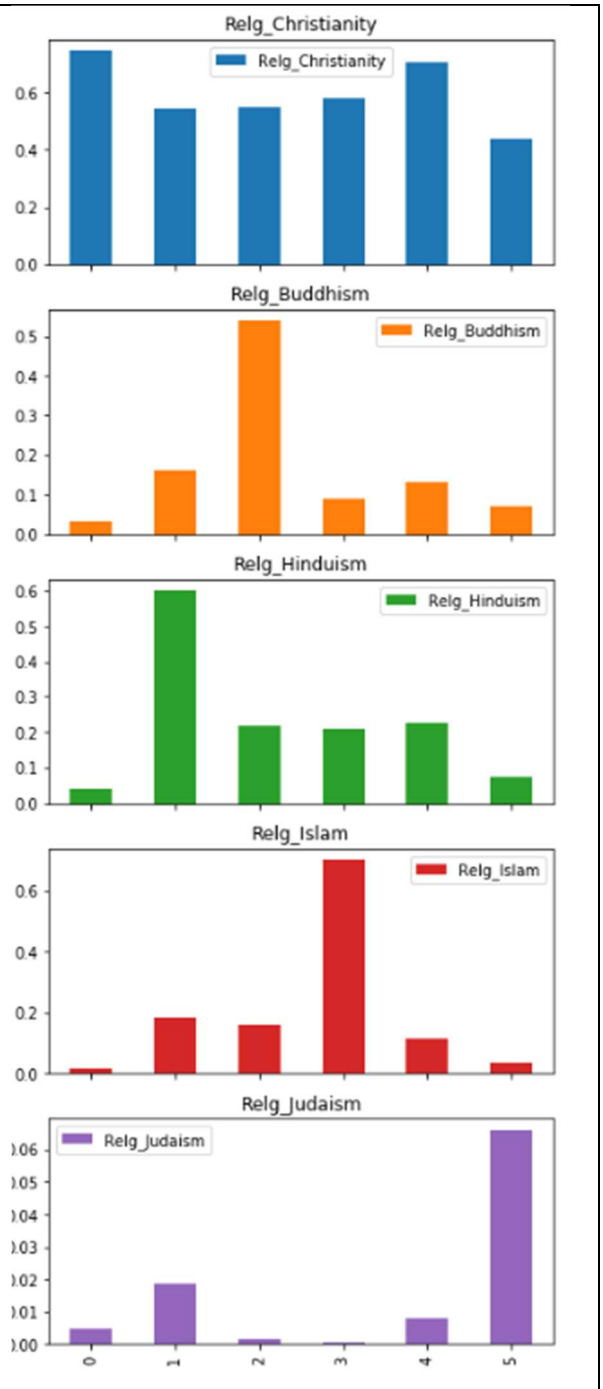
Religious Distribution – 7 Clusters

When we produce 7 Clusters, one religion stands out from the rest: Judaism.

Almost 100% of Judaism believers live in the same cluster (Cluster-5).

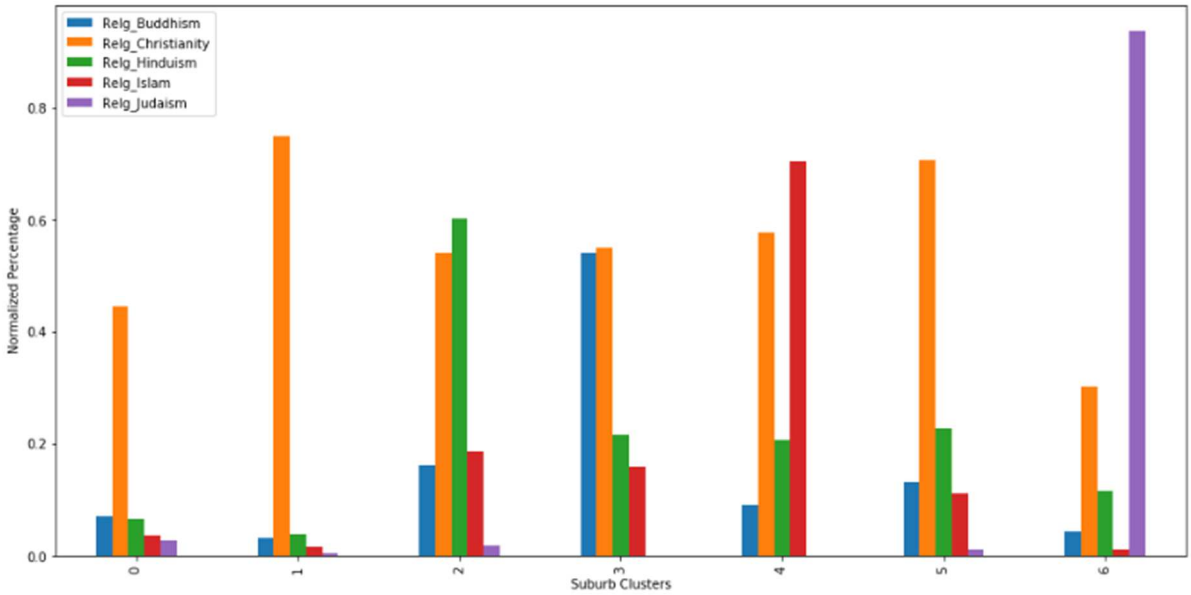
Aside from housing requirements, It will be interesting to understand this phenomena.

I would think it may have something to do with their religion, e.g. it will be easier for them to live in the same cluster (suburbs) so they can find kosher food, or it's easier for them when they need to observe Sabbath in their synagogue.

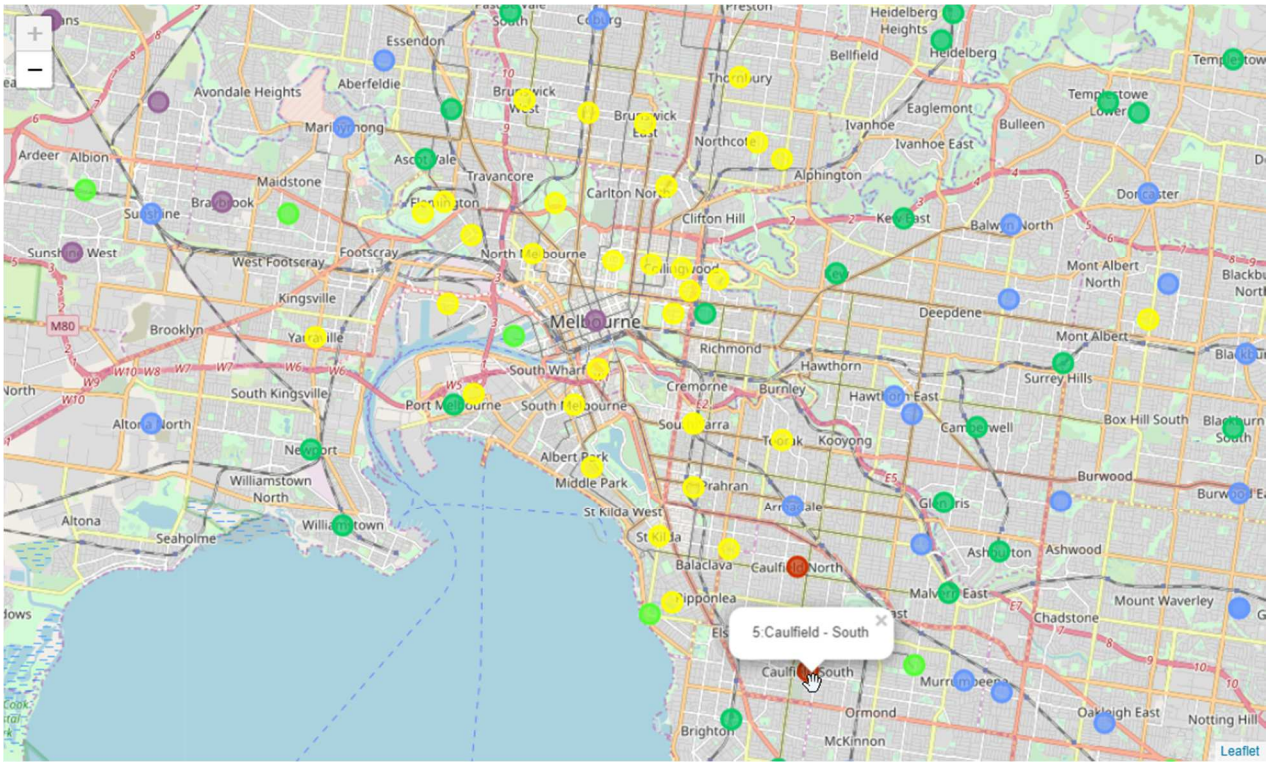


Religious Distribution – 7 Clusters in one chart

Below is religious population distribution with 7 clusters for all major religions displayed in 1 chart.



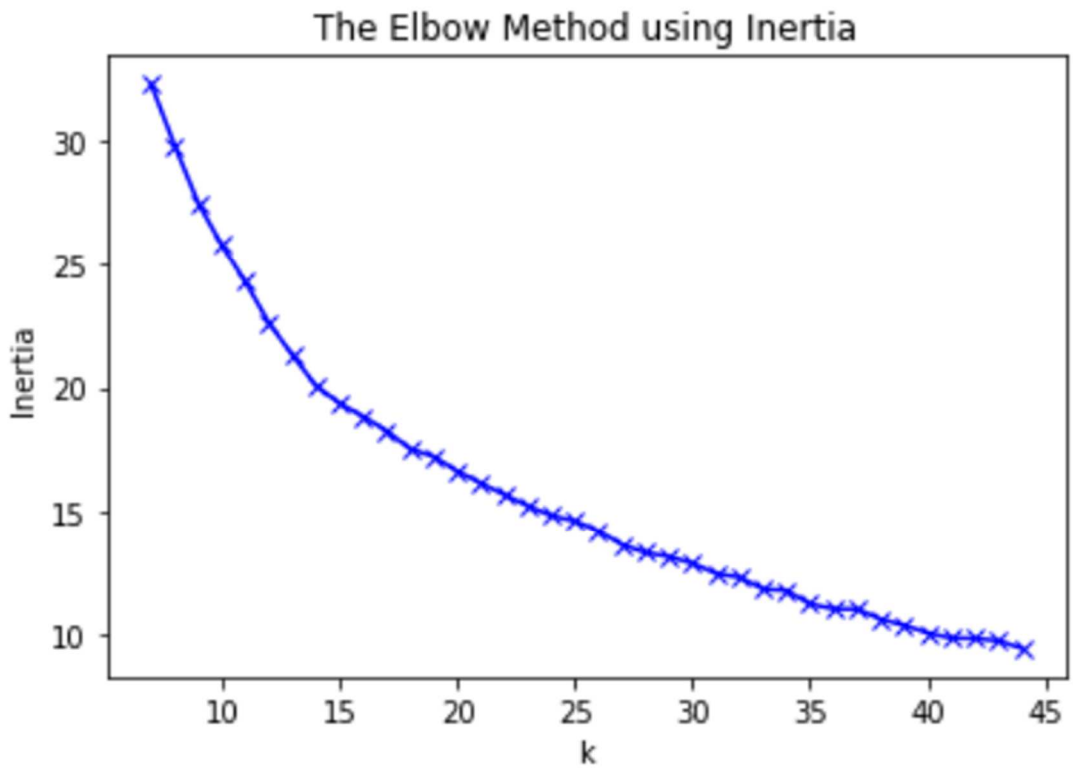
4.2 Map of Melbourne Religious Clusters



4.3 Country of Origin Distribution

4.3.1 Finding an optimal k of K-Means

I use the Elbow Method to obtain an optimal k value of K-Means.

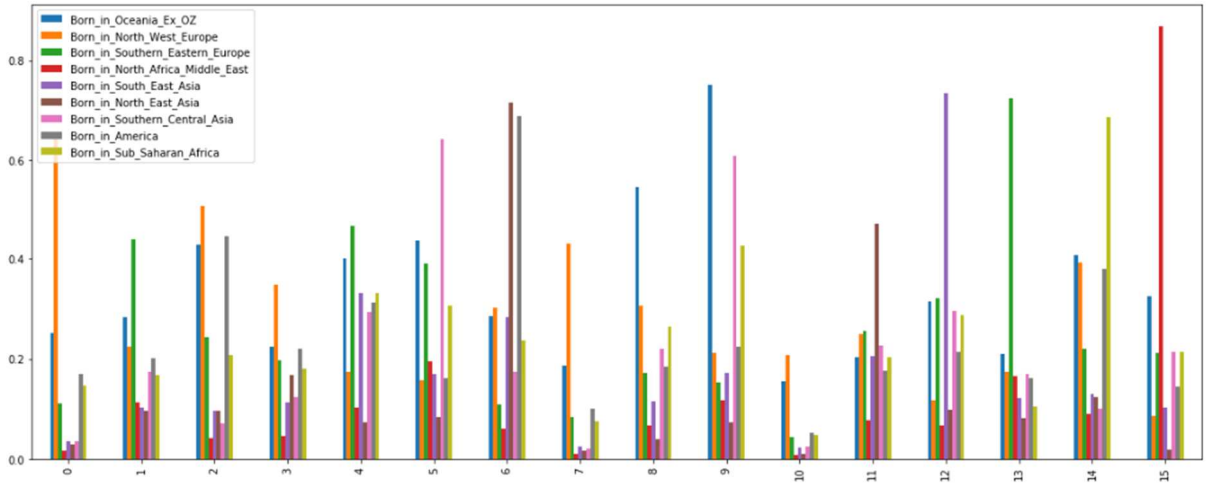


4.3.2 Clustering - Country of Origin

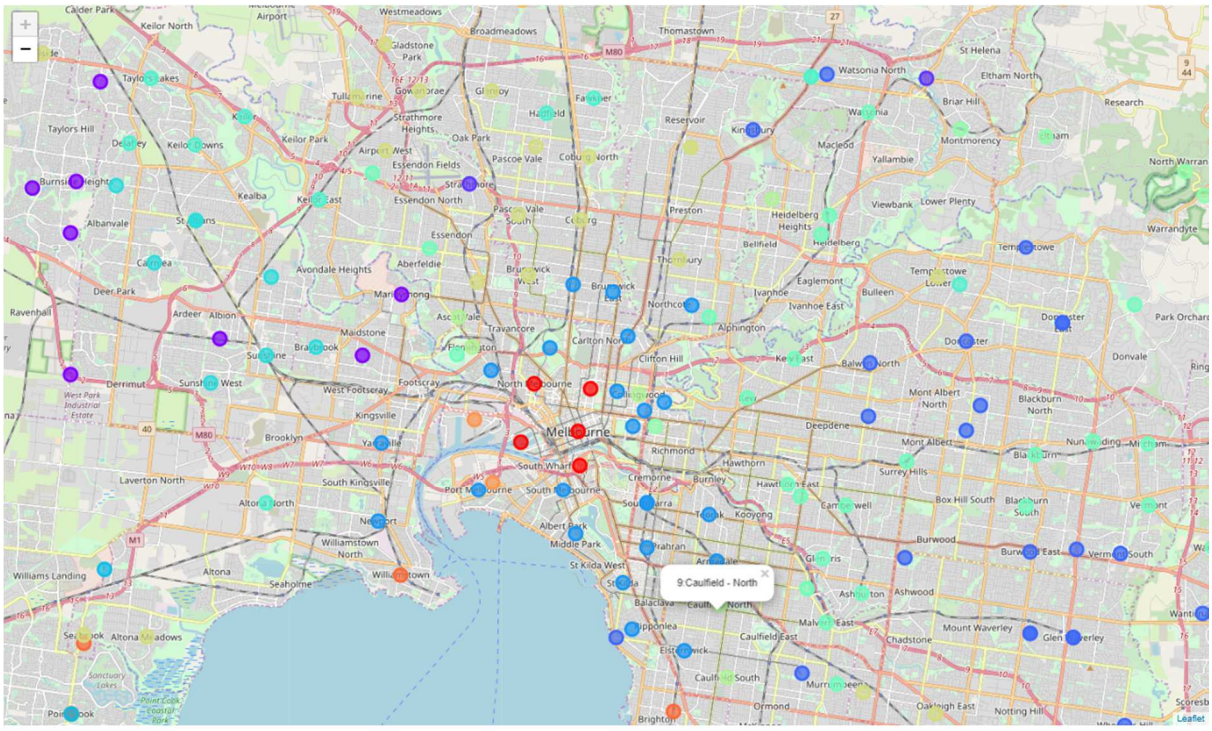
Similar to the clustering of religion, we can see that the distribution of Victoria's population based on country of origin is also not distributed evenly, especially North Africa + Middle east, also East Asia.

All Countries of Origin Distribution in one chart

Below is religious population distribution fall all clusters based on where they come from when they came to Victoria, Australia.



4.4 Map of Melbourne Religious Clusters



5. Conclusion

It is evident that new migrants have a tendency to live in an area where people share the same background, ethnicity, culture or religion.

With clustering, we can help them to find a place to call home.

We can expand the scope of this research by adding more features, such as the median age of population, their education, or facilities found in each suburb (schools, hospitals, shopping centres, religious centres, etc.).

In addition, the distance of the suburb from the city will also greatly influence their decision when buying a house.