Finding the Best Location to Live

for new migrant families

coursera APPLIED DATA SCIENCE

Capstone Project

Kamin Arifin

10 June 2019

Agenda

- Introduction
- Data Sources
- Data Analysis
- Methodology
- Result
- Conclusion

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).

Introduction (Continued….)

Australia is a very big country consisting of 6 states, with immigrants usually choosing to reside in Victoria or NSW. There are 3665 suburbs in Victoria, 41 of which are popular for fresh migrants (as these suburbs are not too far from Melbourne City).

In this project, we will look at the population of Victoria according to 2 criteria: Religion and country of origin.

Data Sources

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).

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’.

Data Analysis (Continued...)

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.

Religions (Data Analysis Continued)

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

- Christian
- Buddhist
- Hindu
- Islam
- Judaism

Country of Origin (Data Analysis Continued)

These are the countries of origin provided by the ABS:

- 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

Methodology

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

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)

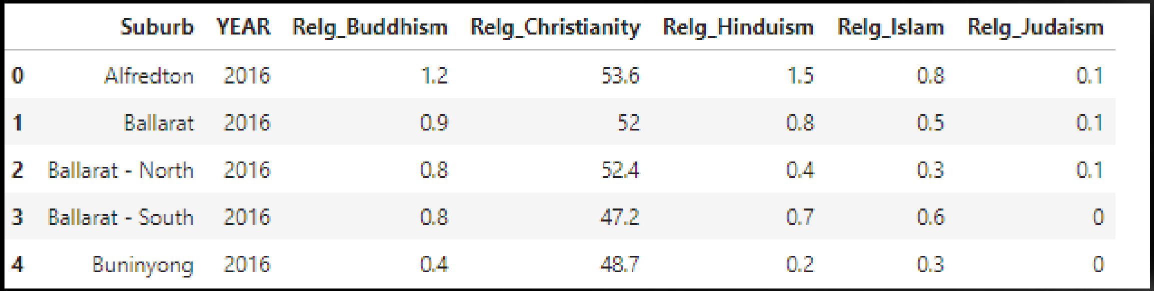
Folium is a good library to visualize data in Python on an interactive leaflet map. I will use this to display a map of Victoria or Melbourne.

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.

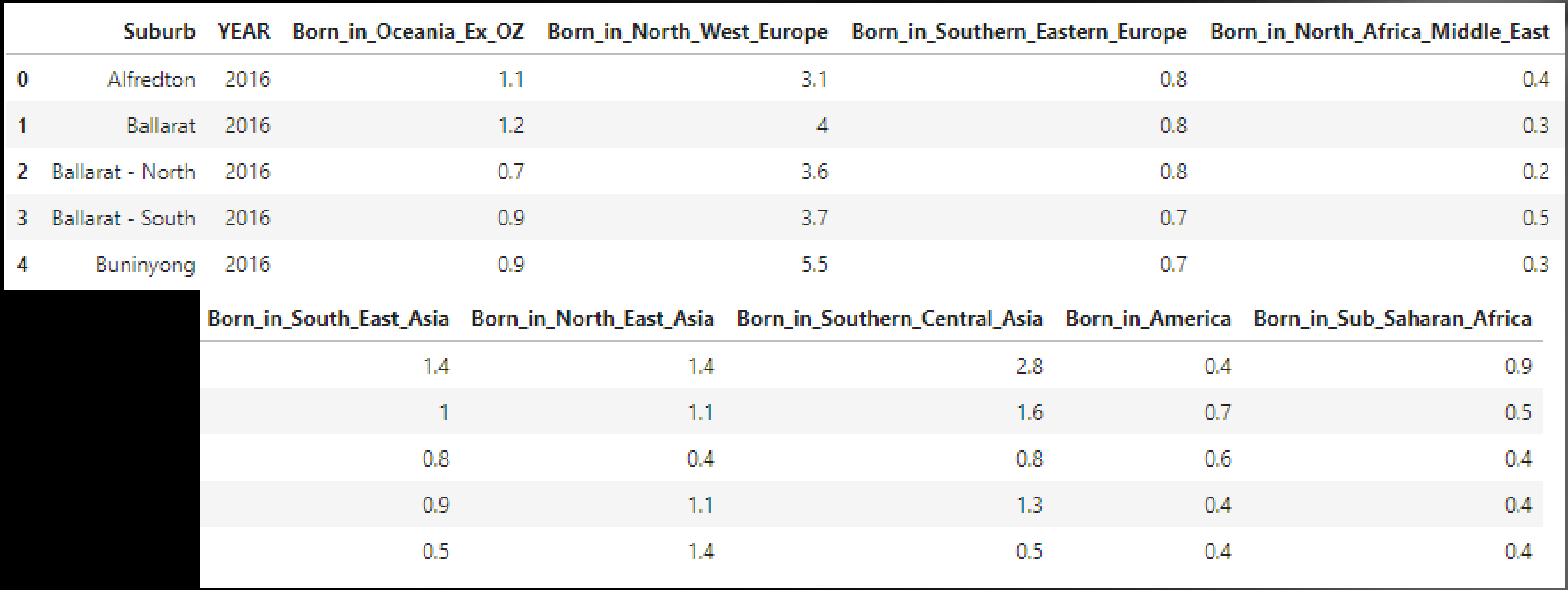
Sample Data – List of Columns

```
['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']
```

Sample Data



	Suburb	YEAR	Relg_Buddhism	Relg_Christianity	Relg_Hinduism	Relg_Islam	Relg_Judaism
0	Alfredton	2016	1.2	53.6	1.5	0.8	0.1
1	Ballarat	2016	0.9	52	0.8	0.5	0.1
2	Ballarat - North	2016	0.8	52.4	0.4	0.3	0.1
3	Ballarat - South	2016	0.8	47.2	0.7	0.6	0
4	Buninyong	2016	0.4	48.7	0.2	0.3	0



	Suburb	YEAR	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
0	Alfredton	2016	1.1	3.1	0.8	0.4	1.4	1.4	2.8	0.4	0.9
1	Ballarat	2016	1.2	4	0.8	0.3	1	1.1	1.6	0.7	0.5
2	Ballarat - North	2016	0.7	3.6	0.8	0.2	0.8	0.4	0.8	0.6	0.4
3	Ballarat - South	2016	0.9	3.7	0.7	0.5	0.9	1.1	1.3	0.4	0.4
4	Buninyong	2016	0.9	5.5	0.7	0.3	0.5	1.4	0.5	0.4	0.4

Methodology (Continued)

As Religious Data and Country of Origin comes from different spreadsheets, we need to merge the 2 datasets and only use the relevant columns and rows.

Here are the steps:

- Download the data from the ABS for those 2 datasets.
- 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.

Methodology (Continued)

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.

Clustering - Religion Dataset

(Methodology, Continued)

Regarding the Religion dataset, I used the K-Means method for clustering.

To understand the distribution better, I did the analysis using the k values: k=4, k=6 and k=7.

The results, though surprising, is quite understandable. They are shown in the following slides.

Clustering - Country of Origin
(Methodology, Continued)

In this dataset, to determine *k*, I used the Elbow Method to obtain an optimal *k* value.

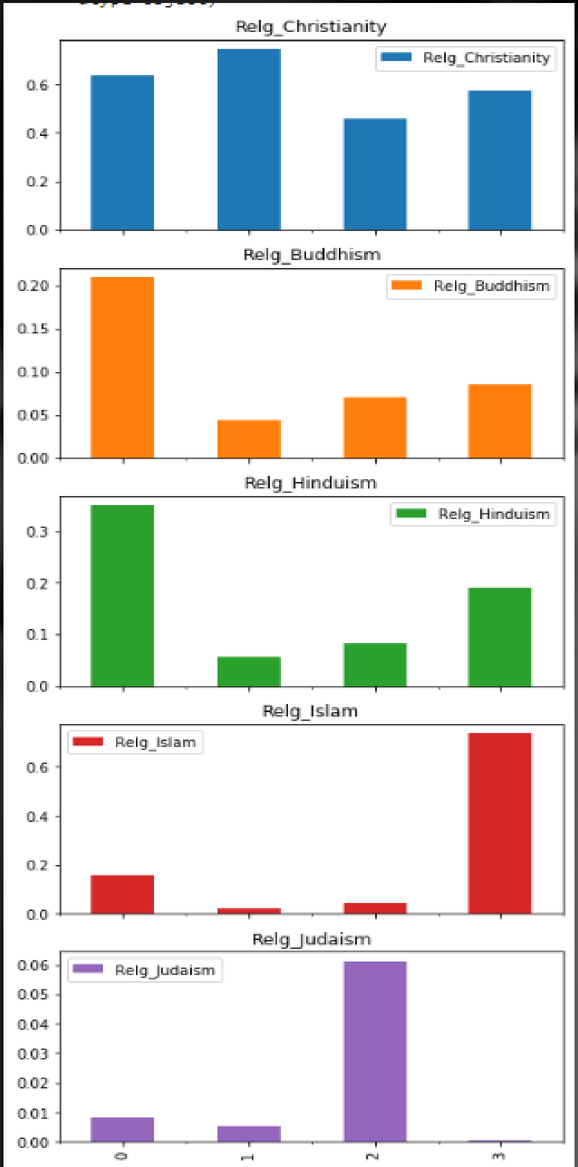
According to the chart, the optimal k value is 16.

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 other religions.

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



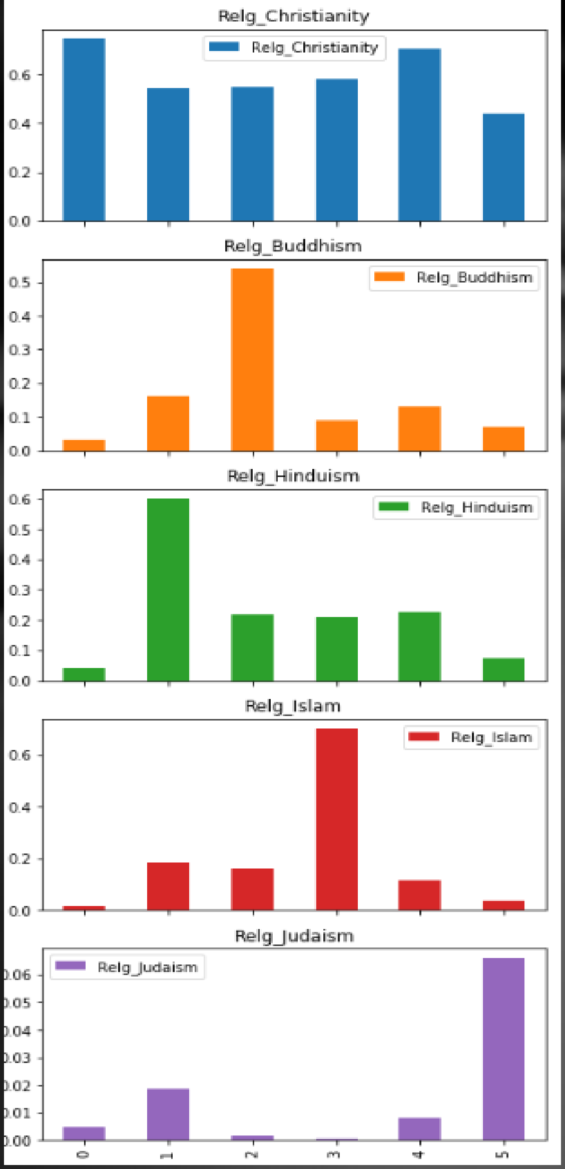
Religious Distribution – 6 Clusters

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.



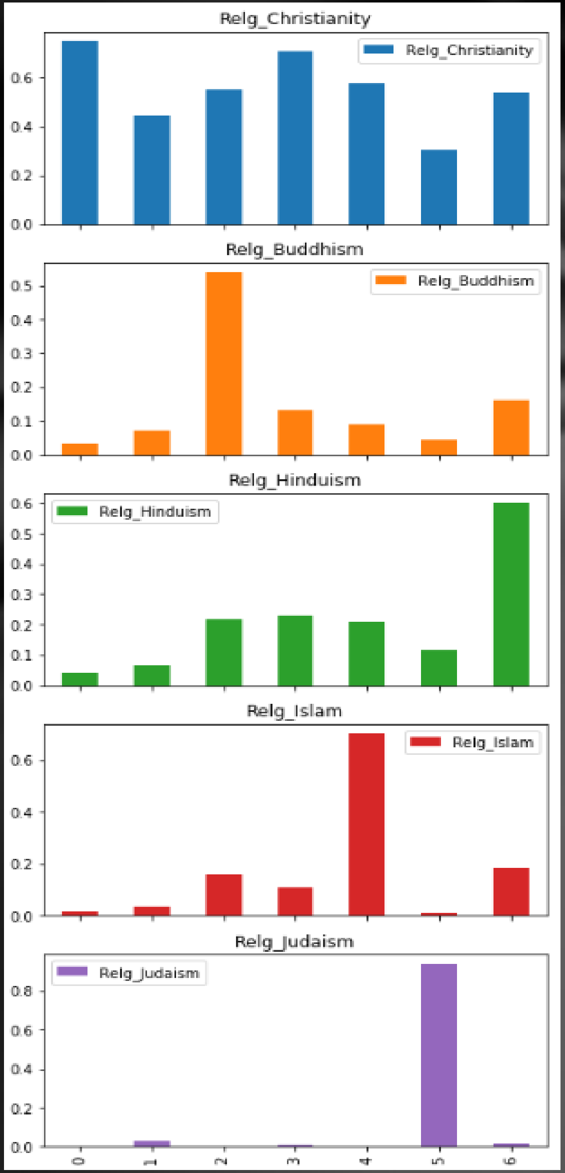
Religious Distribution – 7 Clusters

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).

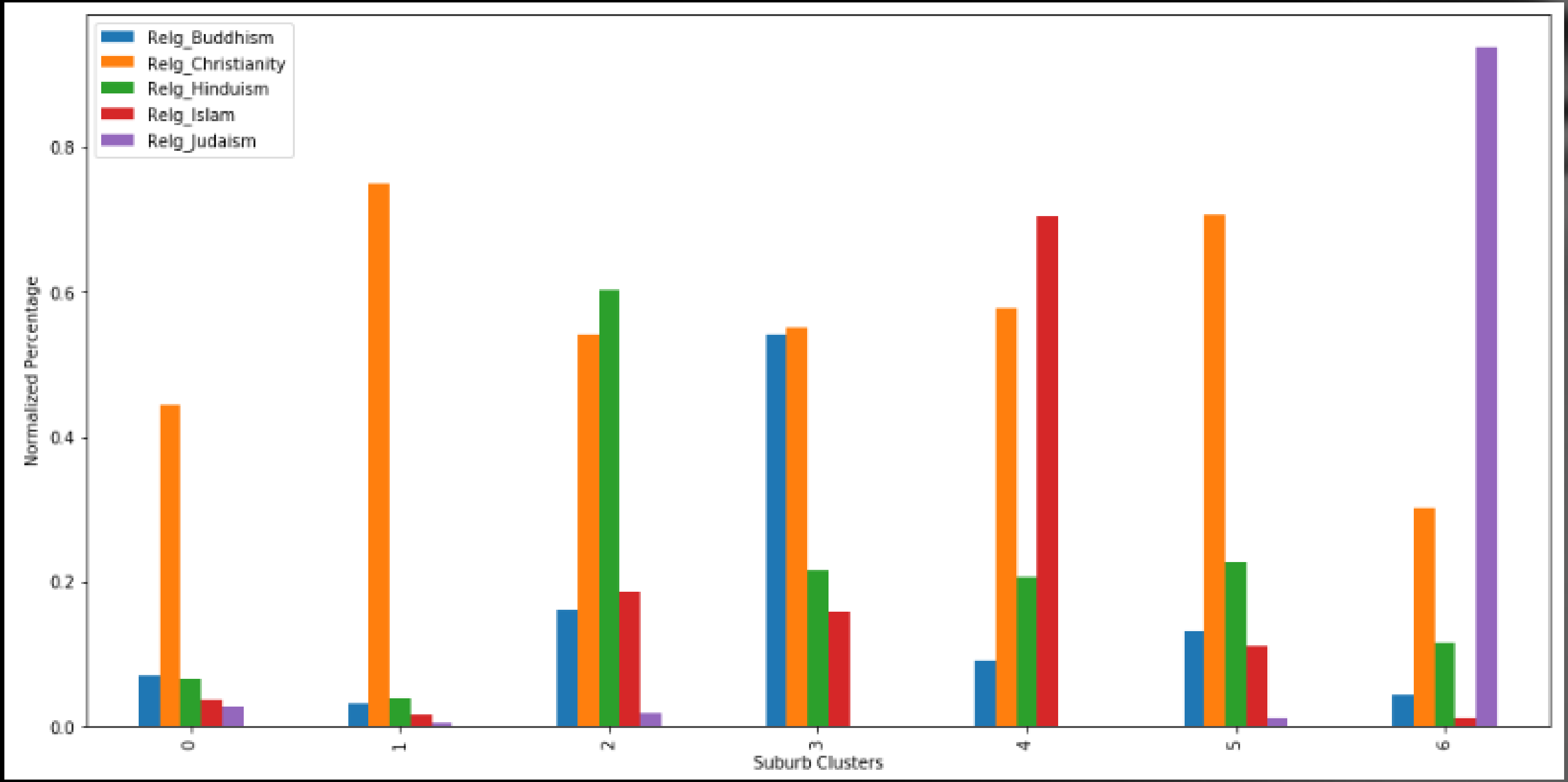
Aside from housing requirements, It will be interesting to understand this phenomena.

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.

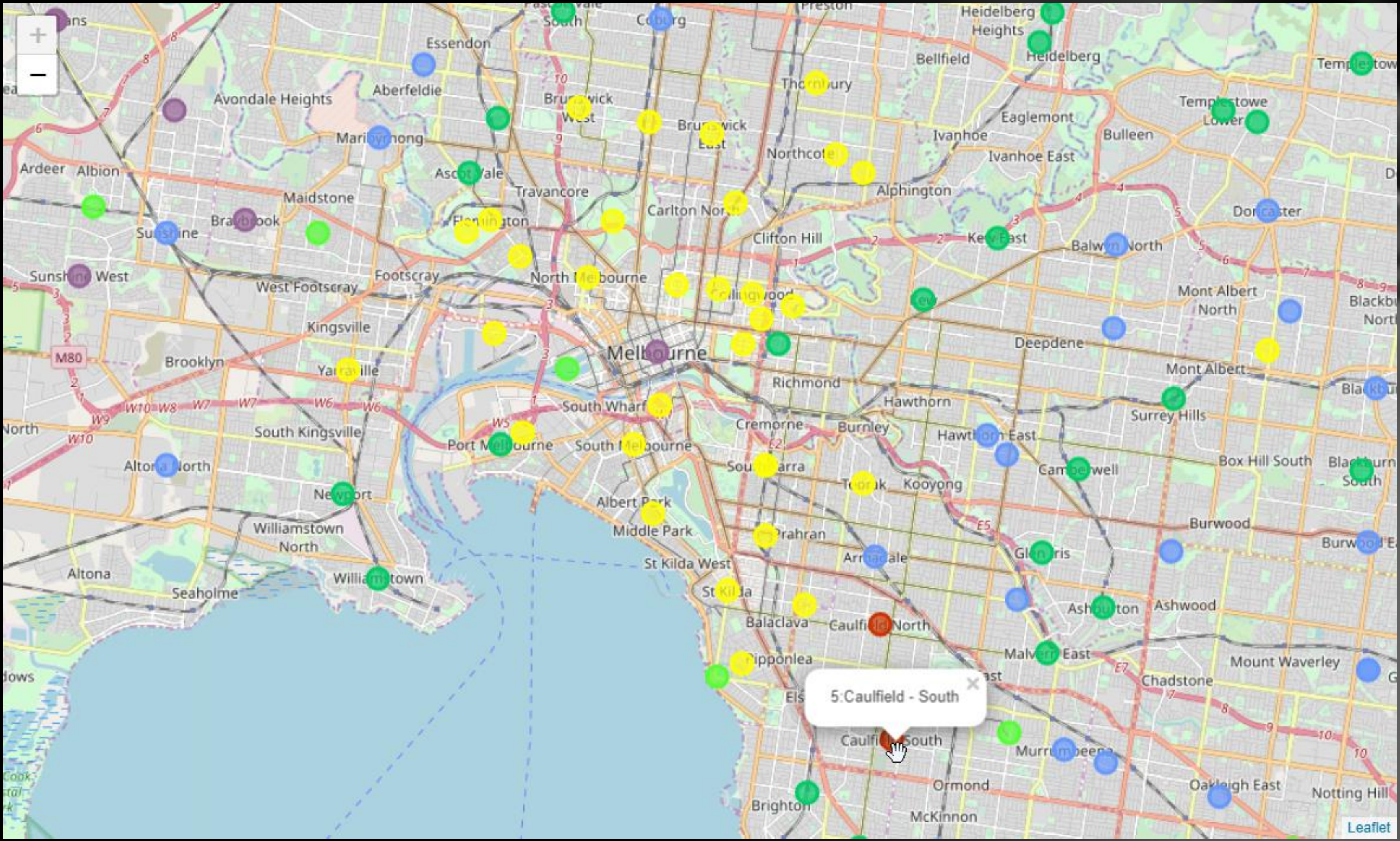


Religious Distribution – 7 Clusters
(Continued)

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

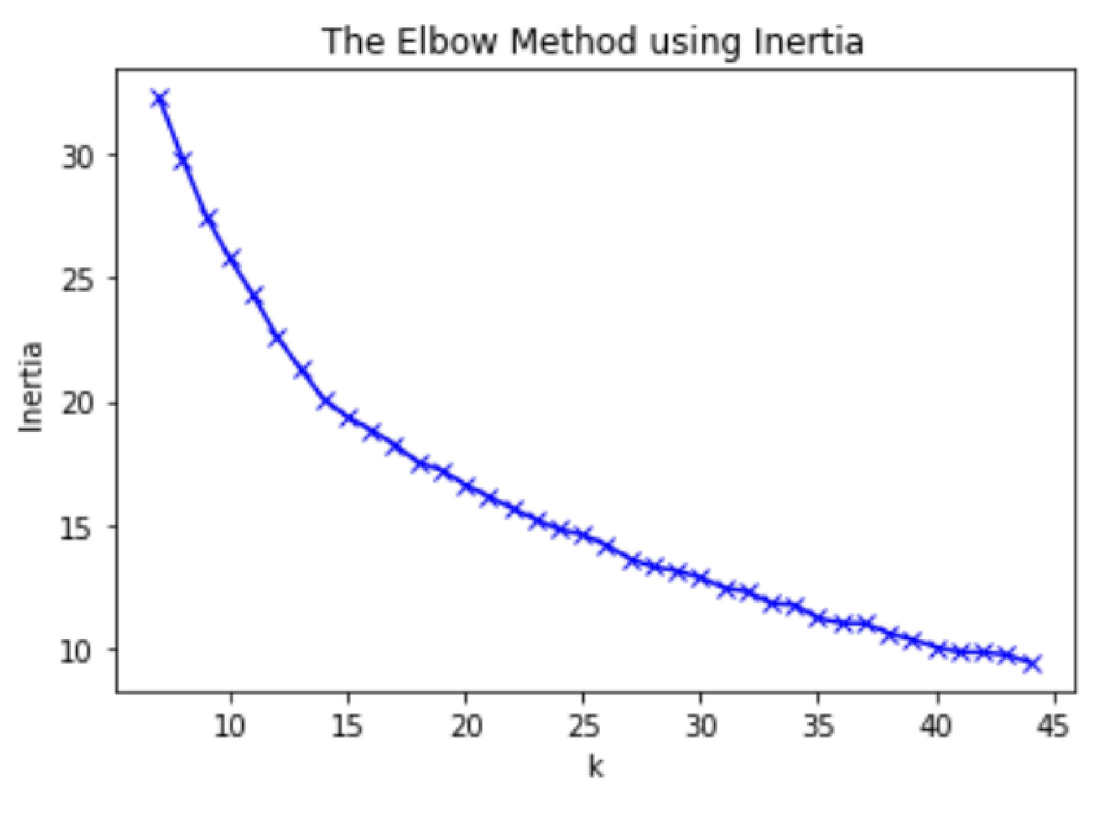


Melbourne Religious Clusters



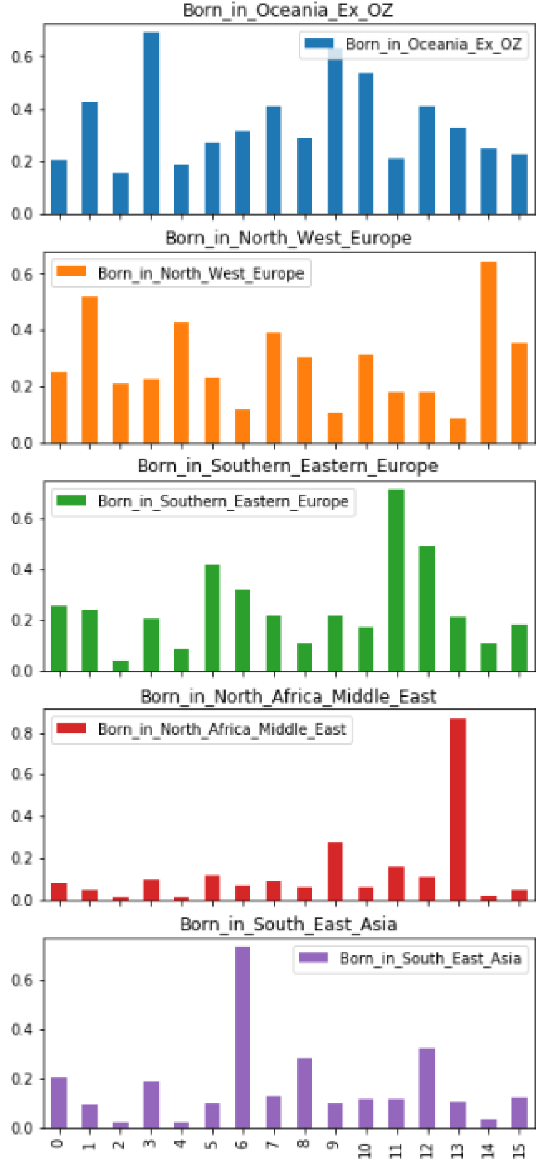
Clustering - Country of Origin

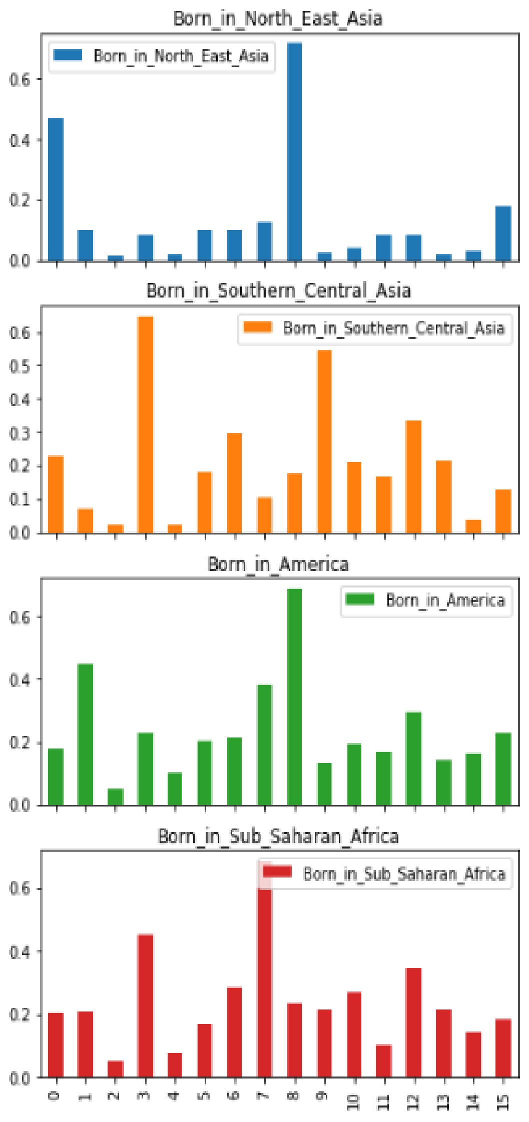
In this dataset, I use the Elbow Method to obtain an optimal k value of K-Means. Based on the chart below, I decide to use k=16.



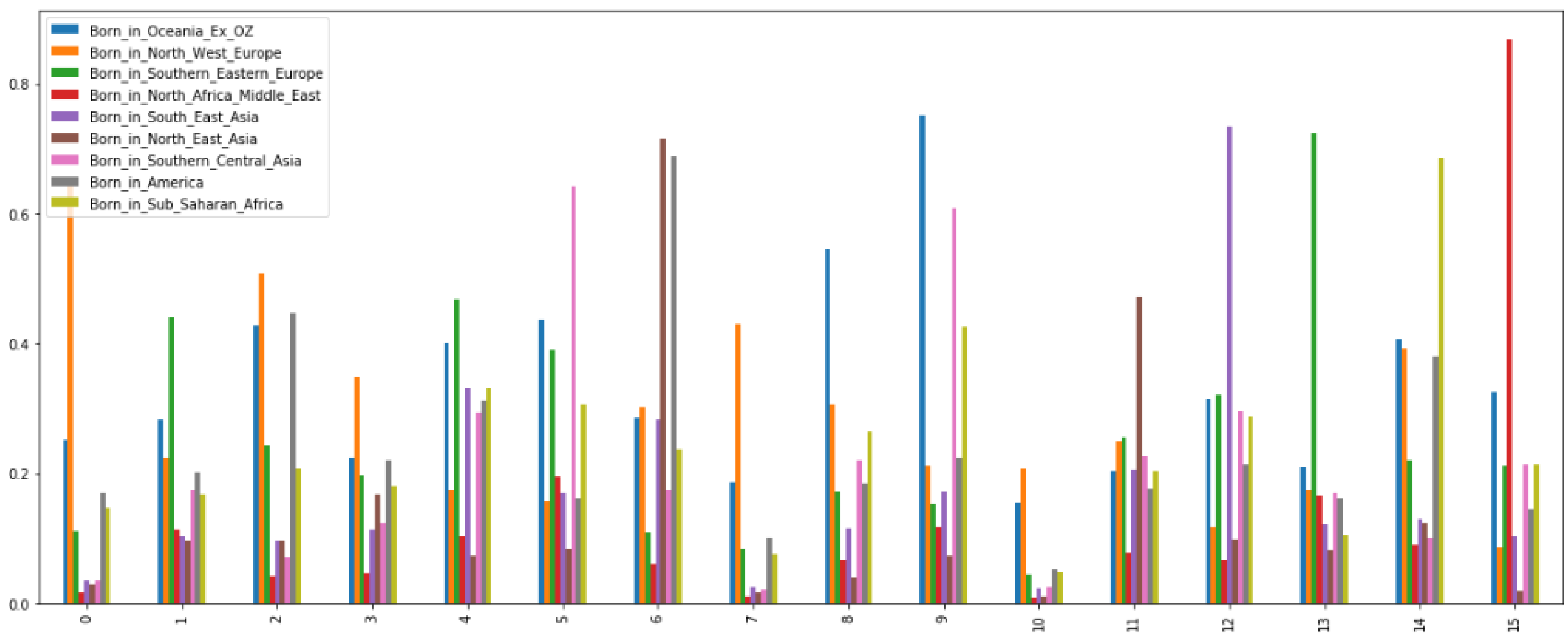
Clustering - Country of Origin

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, and East Asia.

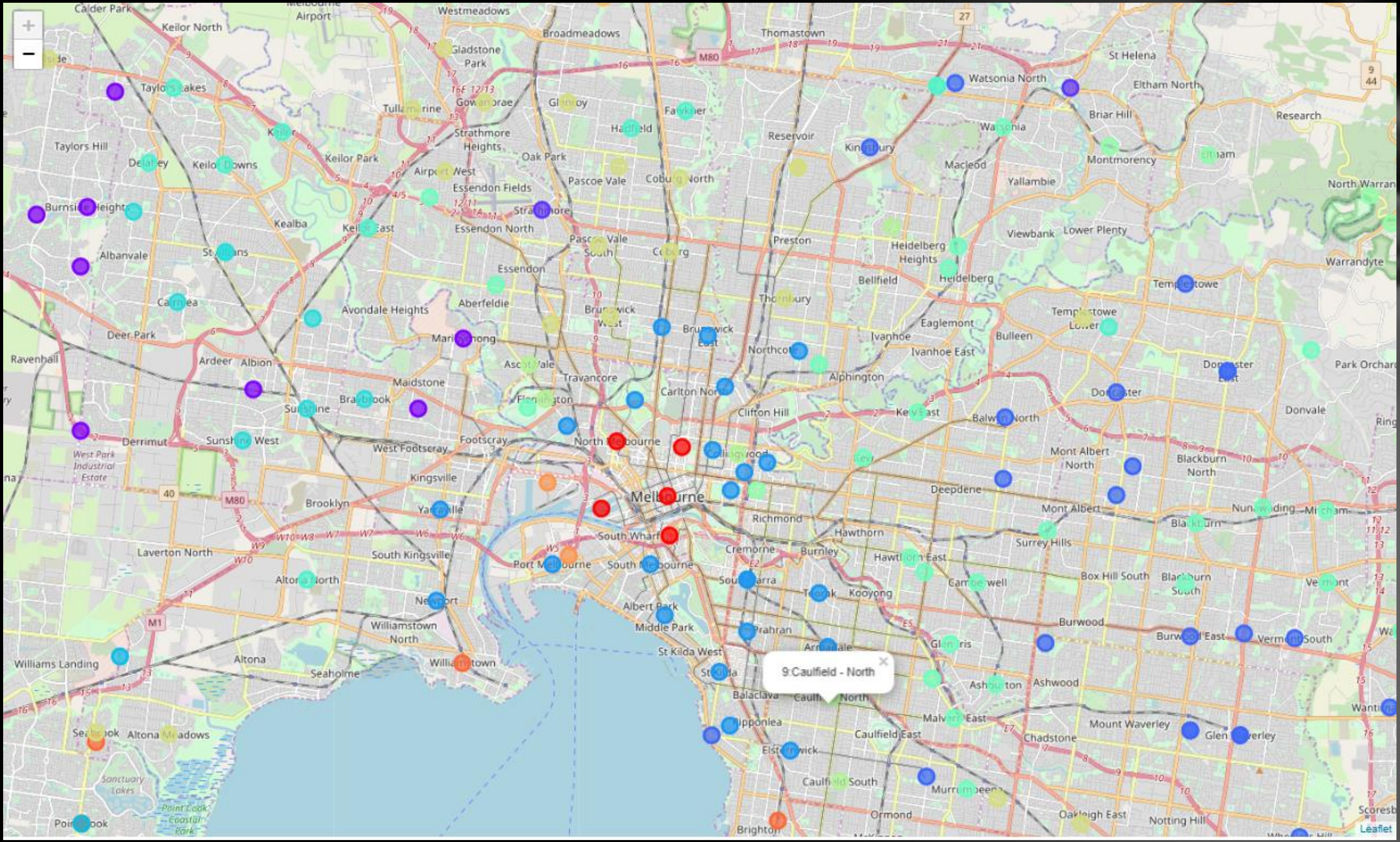




Clustering - Country of Origin



Clustering - Country of Origin



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.

End – of – Slides

Thanks