JUNE 3, 2020

PREDICTING THE BEST WESTERN SUBURBS IN MELBOURNE, AUSTRALIA TO OPEN A NEW FRANCHISE FOR AN INDIAN RESTAURANT

PRADEEP REDDY NAKKALA
Data Enthusiast
Melbourne

1. Introduction

As part of the IBM Data Science professional program project, I have reached out to one of my close acquaintance who is the restaurant owner of INKA Australia **Inkarestaurant Australia** and will be doing analysis on the best western suburbs in Melbourne, Australia chosen by the management to expand their business. This is a real time example for which I will be performing analysis and advising the management as to which would be the best suburb from the 3 councils (**Hobsons Bay, Brimbank, Wyndham).**

1.1. Background

INKA Australia is one of the Indian restaurants which is located in the inner suburbs (Hawthorn) of Melbourne, Australia. During one of the conversations, the restaurant management have expressed their plans of expansion of their business to Western Suburbs in Melbourne.

1.2. Business Problem

Since expanding their restaurant business to other suburbs would be a cost and risk-based plan for the management, they have selected 3 councils i.e, Hobsons Bay, Brimbank and Wyndham but again there is challenge for them to select out of 60 suburbs as to which would be the best suburb for the setup.

We will analyse the localities in the western suburbs in Melbourne to identify the most profitable suburb since the success of the restaurant depends on the nearby venues and categories.

In this project, I will go through all the process and will provide a conclusion whether the analysis can be leveraged by the business stakeholders to make their decisions.

2. Data Requirements and Cleaning

Few Data components are deemed as key factors in selecting the restaurant location We need to analyse the councils' data, geo-location about the 3 chosen councils as the management has already made up their mind about the councils. Throughout the assignment, I will be using missing value imputation, Foursquare API, Folium map and k-mean clustering.

2.1. Data Sources

- I will be using the XLS document downloaded from site https://www.matthewproctor.com/full_australian_postcodes_vic which will make my analysis handy, as it has all the relevant information for the project. Geo-locational information (latitude and longitude) about that specific locality and the suburbs
- Data about different venues in different localities based on the suburb under the local councils.
- Suburbs Population, household income was extracted from the https://itt.abs.gov.au/ to an excel file and filtered out based on the 3 suburbs

- Foursquare API locational information to be used. (basic and advanced information about that venue)

2.2. Data Cleaning

- The data preparation for each of the sources of data is done separately.
- Australian post codes and the suburbs population and household income was filtered out based on the 3 chosen councils

ID	Postcode	Locality	State	Long	Lat	DC	Type	SA3	SA3 Name	SA4	SA4 Name	Region	Status
230	200	ANU	ACT	0.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	R1	NaN
21820	200	Australian National University	ACT	149.118900	-35.277700	NaN	NaN	NaN	NaN	NaN	NaN	R1	Added 19-Jan-2020
232	800	DARWIN	NT	130.836680	-12.458684	NaN	NaN	70101.0	Darwin City	701.0	Darwin	R1	Updated 6-Feb-2020
233	801	DARWIN	NT	130.836680	-12.458684	NaN	NaN	70101.0	Darwin City	701.0	Darwin	R1	Updated 25-Mar-2020 SA3
234	804	PARAP	NT	130.873315	-12.428017	NaN	NaN	70102.0	Darwin Suburbs	701.0	Darwin	R1	Updated 25-Mar-2020 SA3

	Postcode	Locality	Long	Lat	SA3 Name
0	800	DARWIN	130.836680	-12.458684	Darwin City
1	801	DARWIN	130.836680	-12.458684	Darwin City
2	804	PARAP	130.873315	-12.428017	Darwin Suburbs
3	810	ALAWA	130.866242	-12.381806	Darwin Suburbs
4	810	BRINKIN	130.866242	-12.381806	Darwin Suburbs

- Used missing value imputation for values which have NAN

```
# Observed based on the dataframe there are some missing values and shows NAN dropping the values.  Dropping those rows.

#Deleting the columns which we do not need for analysis

data_df.drop(data_df.columns[[2, 5, 6,7,9,10,11,12]], axis = 1, inplace = True)
data_df = data_df.dropna()
data_df = data_df.reset_index(drop=True)
```

```
data_df.head()
```

	Postcode	Locality	Long	Lat	SA3 Name
0	800	DARWIN	130.836680	-12.458684	Darwin City
1	801	DARWIN	130.836680	-12.458684	Darwin City
2	804	PARAP	130.873315	-12.428017	Darwin Suburbs
3	810	ALAWA	130.866242	-12.381806	Darwin Suburbs
4	810	BRINKIN	130.866242	-12.381806	Darwin Suburbs

- Rename the column SAE Name to Council name to recognise the data frame based on the councils.

```
# Renaming the suburb column name SA3 Name to Council_Name

data_df.rename(columns = {'SA3 Name':'Council_Name','Long':'Longitude','Lat':'Latitude'}, inplace = True)
data_df.head()
```

	Postcode	Locality	Longitude	Latitude	Council_Name
0	800	DARWIN	130.836680	-12.458684	Darwin City
1	801	DARWIN	130.836680	-12.458684	Darwin City
2	804	PARAP	130.873315	-12.428017	Darwin Suburbs
3	810	ALAWA	130.866242	-12.381806	Darwin Suburbs
4	810	BRINKIN	130.866242	-12.381806	Darwin Suburbs

- Base on the data we have retrieved 18019 rows and 5 columns but again we need to filter the data based on the 3 councils.

```
In [9]: # Total number of rows and columns

        data_df.shape

Out[9]: (18019, 5)
```

before filtering out with councils names

```
In [10]: # Filtering the suburbs based on the 3 councils

         temp_df =data_df[(data_df.Council_Name == 'Hobsons Bay') | (data_df.Council_Name == 'Brimbank')| (data_df.Council_Name == 'Wyndha

In [11]: temp_df.head()

Out[11]:
```

	Postcode	Locality	Longitude	Latitude	Council_Name
6011	3015	NEWPORT	144.880556	-37.838242	Hobsons Bay
6012	3015	SOUTH KINGSVILLE	144.880556	-37.838242	Hobsons Bay
6013	3015	SPOTSWOOD	144.880556	-37.838242	Hobsons Bay
6014	3016	WILLIAMSTOWN	144.888461	-37.863743	Hobsons Bay
6015	3016	WILLIAMSTOWN NORTH	144.888461	-37.863743	Hobsons Bay

```
In [12]: temp_df.shape

Out[12]: (60, 5)
```

After filtering out with councils names

- The coordinates of the locality and venues to be obtained using Foursquare Maps API geocoding
to get the final dataset.

	Locality	Locality Latitude	Locality Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	NEWPORT	-37.838242	144.880556	7-Eleven	-37.840988	144.883114	Convenience Store
1	NEWPORT	-37.838242	144.880556	Newport IGA Plus Liquor	-37.842439	144.882312	Grocery Store
2	NEWPORT	-37.838242	144.880556	Mamma Teresa Woodfired Restaurant	-37.841520	144.882800	Pizza Place
3	NEWPORT	-37.838242	144.880556	The Backyard Est.2016	-37.842660	144.881590	Café
4	SOUTH KINGSVILLE	-37.838242	144.880556	7-Eleven	-37.840988	144.883114	Convenience Store

- Grouped the venues, category, Lat, long by Locality

Locality	Locality Latitude	Locality Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
ALBANVALE	10	10	10	10	10	10
ALBION	19	19	19	19	19	19
ALTONA	4	4	4	4	4	4
ALTONA EAST	2	2	2	2	2	2
ALTONA GATE	2	2	2	2	2	2
ALTONA MEADOWS	11	11	11	11	11	11
ALTONA NORTH	2	2	2	2	2	2
ARDEER	1	1	1	1	1	1
BROOKFIELD	4	4	4	4	4	4
CHARTWELL	1	1	1	1	1	1
COCOROC	1	1	1	1	1	1
DEER PARK EAST	1	1	1	1	1	1
DERRIMUT	1	1	1	1	1	1
EXFORD	4	4	4	4	4	4
EYNESBURY	4	4	4	4	4	4
GARD[illegible]ITY	5	5	5	5	5	

3. Methodology

3.1 Exploratory Data Analysis

- Getting the data based on the 3 councils from the list

	Postcode	Locality	Longitude	Latitude	Council_Name
6011	3015	NEWPORT	144.880556	-37.838242	Hobsons Bay
6012	3015	SOUTH KINGSVILLE	144.880556	-37.838242	Hobsons Bay
6013	3015	SPOTSWOOD	144.880556	-37.838242	Hobsons Bay
6014	3016	WILLIAMSTOWN	144.888461	-37.863743	Hobsons Bay
6015	3016	WILLIAMSTOWN NORTH	144.888461	-37.863743	Hobsons Bay

- Grouping the data based on the locality and the counts

```
West_venues.groupby('Locality').count()
```

Locality	Locality Latitude	Locality Longitude	Venue	Venue Latitude	Venue Longitude	Venue Catego
ALBANVALE	10	10	10	10	10	1
ALBION	19	19	19	19	19	
ALTONA	4	4	4	4	4	
ALTONA EAST	2	2	2	2	2	
ALTONA GATE	2	2	2	2	2	
ALTONA MEADOWS	11	11	11	11	11	
ALTONA NORTH	2	2	2	2	2	
ARDEER	1	1	1	1	1	
BROOKFIELD	4	4	4	4	4	
CHARTWELL	1	1	1	1	1	
COCOROC	1	1	1	1	1	
DEER PARK EAST	1	1	1	1	1	
DERRIMUT	1	1	1	1	1	
EXFORD	4	4	4	4	4	
EYNESBURY	4	4	4	4	4	4
GARDEN CITY	5	5	5	5	5	
GLENGALA	19	19	19	19	19	19
HOPPERS CROSSING	1	1	1	1	1	
[illegible]A		10	10		[illegible]0	

3.2 Modelling

Using the final dataset containing the localities in 3 western suburbs in Melbourne along with the latitude and longitude, we can find all the venues within a 500-meter radius of each locality by connecting to the Foursquare API. This returns a json file containing all the venues in each locality which is converted to a pandas data frame. This data This data frame contains all the venues along with their coordinates and category.

```python
# one hot encoding

West_onehot = pd.get_dummies(West_venues[['Venue Category']], prefix="", prefix_sep="")
West_onehot.insert(loc=0, column='Locality', value=West_venues['Locality'] )

West_onehot.head(20)
```

	Locality	Asian Restaurant	Athletics & Sports	Badminton Court	Bakery	Basketball Court	Beach	Bus Station	Business Service	Café	...	Restaurant	Sandwich Place	Shopping Mall	Skating Rink	Su
0	NEWPORT	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	NEWPORT	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	NEWPORT	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	NEWPORT	0	0	0	0	0	0	0	0	1	...	0	0	0	0	
4	SOUTH KINGSVILLE	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
5	SOUTH KINGSVILLE	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
6	SOUTH KINGSVILLE	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
7	SOUTH KINGSVILLE	0	0	0	0	0	0	0	0	1	...	0	0	0	0	
8	SPOTSWOOD	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
9	SPOTSWOOD	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
10	SPOTSWOOD	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
11	SPOTSWOOD	0	0	0	0	0	0	0	0	1	...	0	0	0	0	
12	WILLIAMSTOWN	0	0	0	0	0	1	0	0	0	...	0	0	0	0	
13	WILLIAMSTOWN	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
14	WILLIAMSTOWN	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
15	WILLIAMSTOWN	0	0	0	0	0	1	0	0	0	...	0	0	0	0	
16	WILLIAMSTOWN	0	0	0	0	0	0	0	0	1	...	0	0	0	0	
17	WILLIAMSTOWN	0	0	0	0	0	0	0	0	0	...	1	0	0	0	
18	WILLIAMSTOWN	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the locality and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the locality.

To help people find similar locality in the safest borough we will be clustering similar locality using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the 3 localities into 5 clusters. The reason to conduct a K- means clustering is to cluster locality with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each locality.

```
import numpy as np
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Locality']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
west_Sub_venues_sorted = pd.DataFrame(columns=columns)
west_Sub_venues_sorted['Locality'] = west_grouped['Locality']

for ind in np.arange(west_grouped.shape[0]):
    west_Sub_venues_sorted.iloc[ind, 1:] = return_most_common_venues(west_grouped.iloc[ind, :], num_top_venues)

west_Sub_venues_sorted.head()
```

	Locality	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Commo Venue
0	ALBANVALE	Asian Restaurant	Bus Station	Vietnamese Restaurant	Pharmacy	Portuguese Restaurant	Restaurant	Grocery Store	Bakery	Beach	Business Servic
1	ALBION	Gym	Pizza Place	Café	Department Store	Filipino Restaurant	Furniture / Home Store	General Entertainment	Grocery Store	Vietnamese Restaurant	Convenience Stor
2	ALTONA	Café	Train Station	Thai Restaurant	Convenience Store	Wine Shop	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	Clothing Store
3	ALTONA EAST	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	Convenience Store
4	ALTONA GATE	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	Convenience Stor

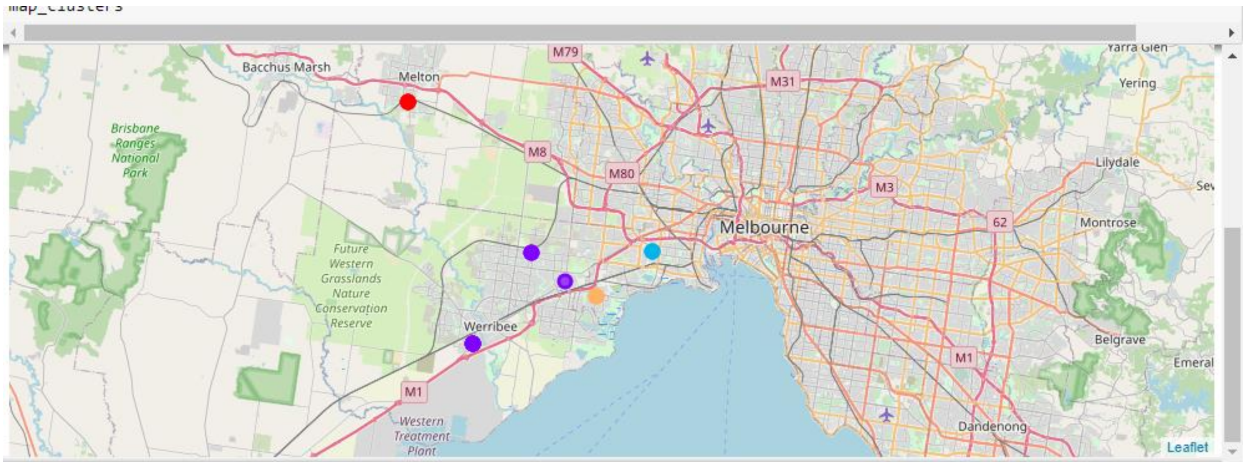
4 Results

After running the K-means clustering we can access each cluster created to see which locality was assigned to each of the five clusters.

```
#Checked there were missing values and dropped those rows .

west_merged = west_merged.dropna()

#Checking the status
west_merged .isnull().values.any()

False

west_merged.head()
```

	Postcode	Locality	Longitude	Latitude	Council_Name	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th M Comm Ven
6044	3025	ALTONA EAST	144.839704	-37.835562	Wyndham	2.0	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Filip Restaur
6045	3025	ALTONA GATE	144.839704	-37.835562	Wyndham	2.0	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Fil Restau
6046	3025	ALTONA NORTH	144.839704	-37.835562	Wyndham	2.0	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Fil Restaur
6049	3027	WILLIAMS LANDING	144.743016	-37.861998	Wyndham	1.0	Playground	Wine Shop	Gym	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast F Restaura
6051	3028	ALTONA MEADOWS	144.777165	-37.875066	Wyndham	4.0	Fast Food Restaurant	Bakery	Asian Restaurant	Pharmacy	Supermarket	Italian Restaurant	Shop

- Checking the Melbourne coordinates using geopy.geocoders and then creating a map using folium maps.
- After running K-means clustering we can access each cluster created to see which locality were assigned to each of the 5 clusters. Visualizing the clustered locality on the map using folium library. Each Cluster is color coded for the ease of presentation.
- Purple cluster dominated which has a smaller number of clusters and which is the least desirable location for setup the business followed by blue and red colors
- The orange cluster which shows on the map is more desirable suburb to setup a new restaurant

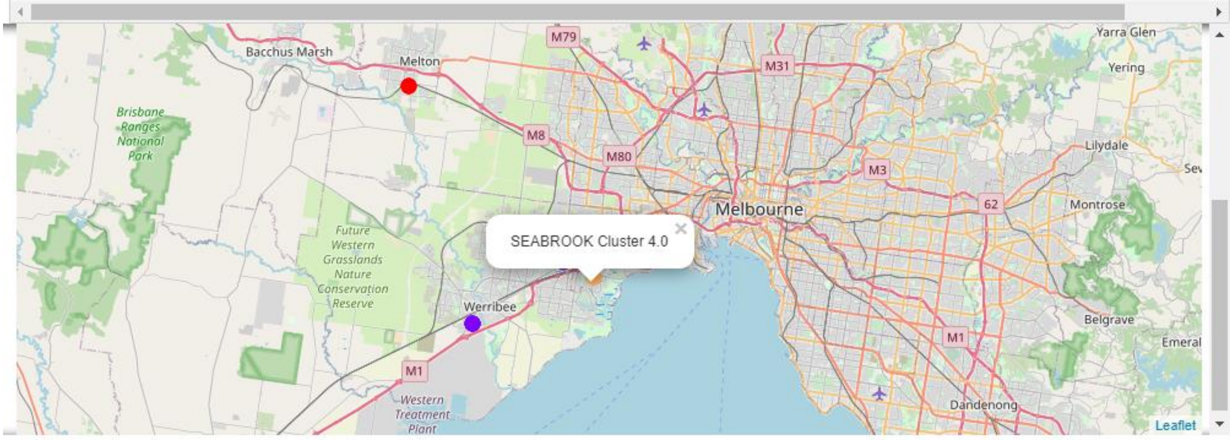


- Getting the list of the Cluster labels which has highest number 5

```
west_merged[west_merged['Cluster Labels']==4]
```

	Postcode	Locality	Longitude	Latitude	Council_Name	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
6051	3028	ALTONA MEADOWS	144.777165	-37.875066	Wyndham	4.0	Fast Food Restaurant	Bakery	Asian Restaurant	Pharmacy	Supermarket	Italian Restaurant	Shopping Mall
6052	3028	LAVERTON	144.777165	-37.875066	Wyndham	4.0	Fast Food Restaurant	Bakery	Asian Restaurant	Pharmacy	Supermarket	Italian Restaurant	Shopping Mall
6053	3028	SEABROOK	144.777165	-37.875066	Wyndham	4.0	Fast Food Restaurant	Bakery	Asian Restaurant	Pharmacy	Supermarket	Italian Restaurant	Shopping Mall

```
map_clusters
```



- Getting the list of the Cluster labels which has number 4. No venues were found

```
west_merged[west_merged['Cluster Labels']==3]
```

Postcode	Locality	Longitude	Latitude	Council_Name	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue

- Getting the list of the Cluster labels which has number 3.

```
west_merged[west_merged['Cluster Labels']==2]
```

	Postcode	Locality	Longitude	Latitude	Council_Name	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th M Comi Ve
6044	3025	ALTONA EAST	144.839704	-37.835562	Wyndham	2.0	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast F Restau
6045	3025	ALTONA GATE	144.839704	-37.835562	Wyndham	2.0	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast F Restau
6046	3025	ALTONA NORTH	144.839704	-37.835562	Wyndham	2.0	Badminton Court	Business Service	Wine Shop	Clothing Store	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast F Restau

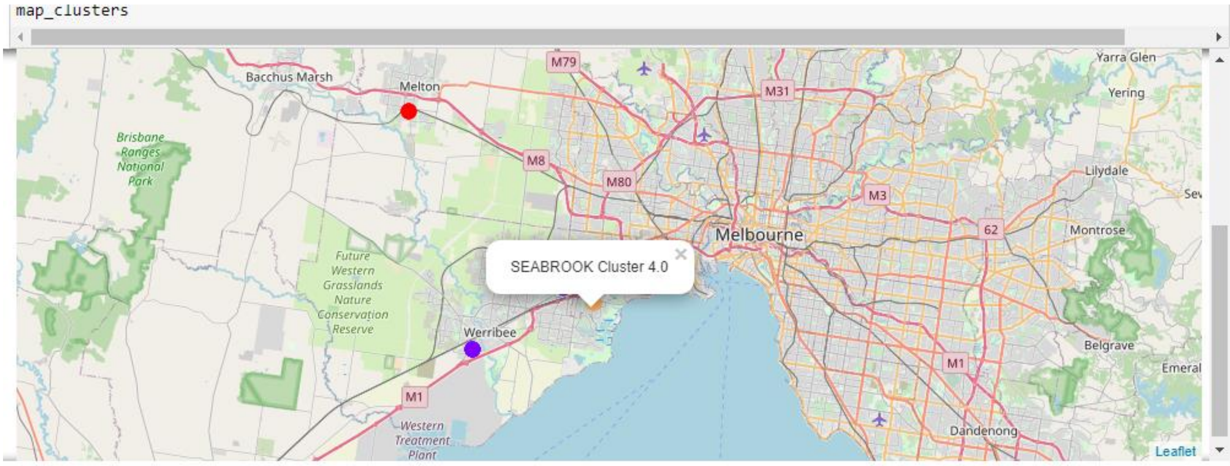
- Getting the list of the Cluster labels which has number 2.

```
In [226]: west_merged[west_merged['Cluster Labels']==1]
```

	Postcode	Locality	Longitude	Latitude	Council_Name	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	
6049	3027	WILLIAMS LANDING	144.743016	-37.861998	Wyndham	1.0	Playground	Wine Shop	Gym	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	C
6054	3029	HOPPERS CROSSING	144.705831	-37.837165	Wyndham	1.0	Playground	Wine Shop	Gym	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	C
6055	3029	TARNEIT	144.705831	-37.837165	Wyndham	1.0	Playground	Wine Shop	Gym	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	C
6056	3029	TRUGANINA	144.705831	-37.837165	Wyndham	1.0	Playground	Wine Shop	Gym	General Entertainment	Furniture / Home Store	Filipino Restaurant	Fast Food Restaurant	Department Store	C

- Getting the list of the Cluster labels which has number 1.
-

- Based on the above map and the data retrieved based on the top venues, categories and locality the Seabrook cluster shows the most desirable out of 3.

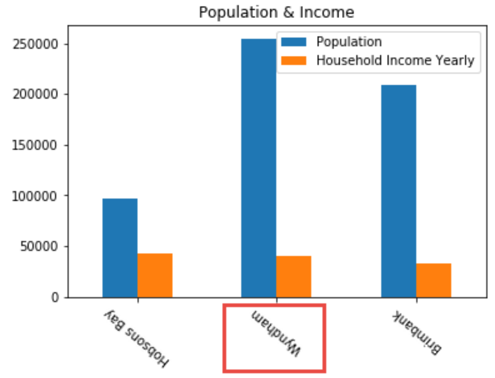


Now checking whether the Council -Wyndham we have chosen is the desirable based on the income and population. I have created a compounded bar chart to show based on the population and the Household income $/year the analysis is appropriate and accurate.

	Suburb	Population	Working Age Population (aged 15-64 years) (%)	Median total household income (yearly) ($)
0	Hobsons Bay	96470	66.4	42482
1	Wyndham	255322	67.0	40060
2	Brimbank	208714	67.7	32914

Used the matplotlib.pyplot library and will be plotting a chart to show the variance. Clearly shows that Wyndham council has the more population and the household income.

```python
import matplotlib.pyplot as plot

 # A python dictionary

data = {"Population":[96470,255322,208714],"Household Income Yearly":[42482, 40060, 32914]

        };

index = ["Hobsons Bay", "Wyndham", "Brimbank"];

# Dictionary loaded into a DataFrame

dataFrame = pd.DataFrame(data=data, index = index);

# Draw a vertical bar chart

dataFrame.plot.bar(rot=500, title="Population & Income");

plot.show(block=True);
```



5. Discussion

The aim of this project is to help the restaurant management to make a decision to setup a location choosing the best council and locality based on the venues, categories, population, household income. Based on the data Cluster 4 are more suitable due to the common venues in that cluster, these localities to have common venues such as Parks, Gym/Fitness centres, Bus Stops, Restaurants, Electronics Stores and Soccer fields which is ideal for a to setup a restaurant so that the foot traffic can be increased and can be more profitable.

6.Conclusion

This project gives a high-level documentation for the restaurant management team to get a better understanding of the localities under 3 councils with respect to the most common venues, population and household income in those localities.

It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before setting up a restaurant in a particular area. The ultimate investment and decision of this project would require consideration of other factors such as cost of living in the suburbs, ethnicity, median house prices which would give more in depth analysis.