Capstone Project – The Battle of Neighbourhood

One Stop Immigration Services by Smart Living Sydney

Background

Smart Living is an immigration consultant company in Sydney. Its main business is to provide services to clients overseas who wish to move to Australia for various reasons such as migration, education or works. It advises clients on immigration related issues and helps to prepare and lodge visa applications on behalf of the clients.

Expanding of Business Model

The company has often faced number of questions from the clients related to housing and schooling. Clients have often interested to find out information about the city in which they are planning to live. The company is currently looking at expanding its business model to include housing and schooling services. As majority of its clients are mainly applying to live in Sydney, it has decided to start launching its new services in this metropolitan city.

Problems Description

Over the years, the company has seen people from different parts of the world moving to Sydney for various reasons. In fact, Sydney is the most populous city in Australia. The city has welcomed many people from many different cultural backgrounds. As the results, businesses of different kinds have blossomed all over the metropolitan area. For example there are many different kinds of authentic restaurants opened across different suburbs of the city. Often certain suburbs are popular of certain types of cuisines for example Middle Eastern, South East Asian, Indian, Chinese, Japanese and Italian etc. Naturally, clients are interested in living in a suburb where they can easily assess to the food they are familiar with. The company is interested to find out in general what restaurant types are popular in which parts of the metropolitan city.

For clients who move into the city with young children, they are certainly interested in schools for their children. Therefore, Smart Living will extend its services to include schooling advises and

applications. It wants to find out locations of primary and high schools in different suburbs of the city.

Another popular venue information the company is aiming to provide to the clients is grocery shopping. For example, clients from India will be interested in places where they easily can buy their Indian groceries.

Finally, one important factor of choosing where to live is proximity to public transport. Clients will be advised on convenience of taking public transport on different suburbs.

In short, we need to provide an overview of Sydney suburbs based the following information:

- Restaurants (of different cuisines)
- Grocery Shops (of different countries e.g. Indian, Italian, Malaysian, Chinese etc.)
- Schools (Primary and Secondary)
- Public transport (Train, bus and ferry)

Target Audience

Our target audience of this project is Smart Living Sydney Immigration Agent, the company is expanding the business to include housing advises to their clients.

Understanding our data: Sydney Background Information

Sydney is located on the east coast of Australia. It is the state capital city of New South Wales. The population of Sydney is estimated to be around 5 millions and it is considered the most populous city in Australia and Oceania. Sydney is made up of over 650 suburbs, 40 local government areas and 15 contiguous regions. The regions consist of Blue Mountain, City, Eastern Suburbs, Forest Districts, Greater Western Sydney, Inner West, Northern Beaches, Upper North Shore, Lower North Shore, Northern Suburbs, South-Eastern Sydney, Southern Sydney and St. George.

Data Source

The first data that we need is a complete list of suburbs of Sydney. This list can be attained from the following website: https://www.intosydneydirectory.com.au/sydney-postcodes.php

In this website, we will find a list of Sydney suburb names, state (which is NSW) and postcodes.

Example of the list found in this website is showed as below:

Sydney Postcodes - Sydney Suburbs

Abbotsbury	NSW	2176
Abbotsford	NSW	2046
Agnes Banks	NSW	2753
Airds	NSW	2560
Alexandria	NSW	2015
Alfords Point	NSW	2234
Allambie Heights	NSW	2100
Allawah	NSW	2218
Ambarvale	NSW	2560
Annandale	NSW	2038
Annangrove	NSW	2156
Appin	NSW	2560
Arcadia	NSW	2159
Arncliffe	NSW	2205
Arndell Park	NSW	2148
Artarmon	NSW	2064

Data Preparation and Cleaning

The table/list that we found from the website has the column, 'State'. As this column has constant value of NSW, it does not provide any valuable information, therefore we will discard this column.
From this final list, we will use Python library, Geocoder to retrieve the location data (i.e. Latitude and Longitude) for all the suburbs.

Our dataframe will look like:

	Suburb	Postcode	Longitude	Latitude
0	Abbotsbury	2176	-33.869285	150.866703
1	Abbotsford	2046	-33.850553	151.129759
2	Agnes Banks	2153	-33.614508	150.711448
3	Airds	2560	-33.909157	151.192128

While running the Python Geocoder to retrieve location data, we have encountered error. The error was due to two entries in the list, i.e. The Sydney Domestic Airport and Sydney International Airport. Geocoder is not able to return longitude and latitude information for these two entries and program hit into error. As these two entries are clearly not a suburb of Sydney, we will remove them from our dataframe.

Once we have the location, FourSquare is used to retrieve venue information to perform our data analysis. The venue categories that we interested are restaurant, schools, public transport and grocery store. We will further examine the types of restaurant and grocery store available in the suburbs of Sydney.

For example we can use the following FourSquare API to retrieve venues of Indian restaurant in a suburb with the following code :

```
search_query = 'Indian'
radius = 500

url =
'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll
={},{}&v={}&query={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET,
latitude, longitude, VERSION, search_query, radius, LIMIT)

results = requests.get(url).json()
```

The result returned by the request is in JSON format, so we need to write a function to retrieve the relevant venues information from the JSON data and create dataframe contains venue information columns.

Exception Handling

When FourSquare is called with the list of suburbs, error is encountered because FourSquare fails to return venues for some suburbs. Therefore we have to include python code: ”Try and Except” to handle this situation.

Finding The Venue Categories

From the results returned by Foursquare, we use Python One hot encoding to convert the result into the following format:

[19]:

	Suburb	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art: & Craft: Stor
0	Abbotsbury	0	0	0	0	0	0	0	0	0	0	0	0	(
1	Abbotsbury	0	0	0	0	0	0	0	0	0	0	0	0	
2	Abbotsford	0	0	0	0	0	0	0	0	0	0	0	0	
3	Abbotsford	0	0	0	0	0	0	0	0	0	0	0	0	
4	Abbotsford	0	0	0	0	0	0	0	0	0	0	0	0	(

Support

We inspect all the venue categories from the above dataframe and consolidate all the categories useful to our analysis and set-up our dataframe for further investigation.

From the sydney_onehot dataframe, we extract venues and create separate dataframe for Indian, Chinese, Malay and Italian venues

[20]:

```
sydney_indian = sydney_onehot[["Suburb","Suburb Latitude","Suburb Longitude","Indian Restaurant",
                               "South Indian Restaurant","Indie Movie Theater"]]
sydney_chinese = sydney_onehot[["Suburb","Suburb Latitude","Suburb Longitude", "Chinese Restaurant","Szechuan Restaurant",
                                "Cantonese Restaurant","Dim Sum Restaurant","Dumpling Restaurant","Shanghai Restaurant",
                                "Taiwanese Restaurant"]]
sydney_italian = sydney_onehot[["Suburb","Suburb Latitude","Suburb Longitude","Italian Restaurant"]]
sydney_malay = sydney_onehot[["Suburb","Suburb Latitude","Suburb Longitude", "Malay Restaurant"]]
sydney_transport = sydney_onehot[["Suburb","Airport","Bus Station","Bus Stop","Light Rail Station", "Metro Station","Train Station"]]
```

Missing Venue Information

Unfortunately, FourSquare does not return school venues when called. Therefore we have no choice but to remove the school analysis from the project. We can in future provide school information from the following link: https://en.wikipedia.org/wiki/List_of_schools_in_Greater_Western_Sydney

Methodology – Data Analysis with table and graphs

From the dataframe we have created by using Python one hot encoding, we retrieve venues from the suburbs of type Indian, Malaysian, Chinese and Italian. From the results obtained, we will list out the top 10 suburbs contain the highest number of Indian, Malaysian, Chinese and Italian venues respectively. Finally, we use bar chart to visualize our findings.

For example, here are the steps for our analysis of Indian venues:

- Set-up indian venues dataframe
- Count the venus for each suburb having Indian venues

- Sort the list and display the top 10 Suburbs
- Present our results in bar chart

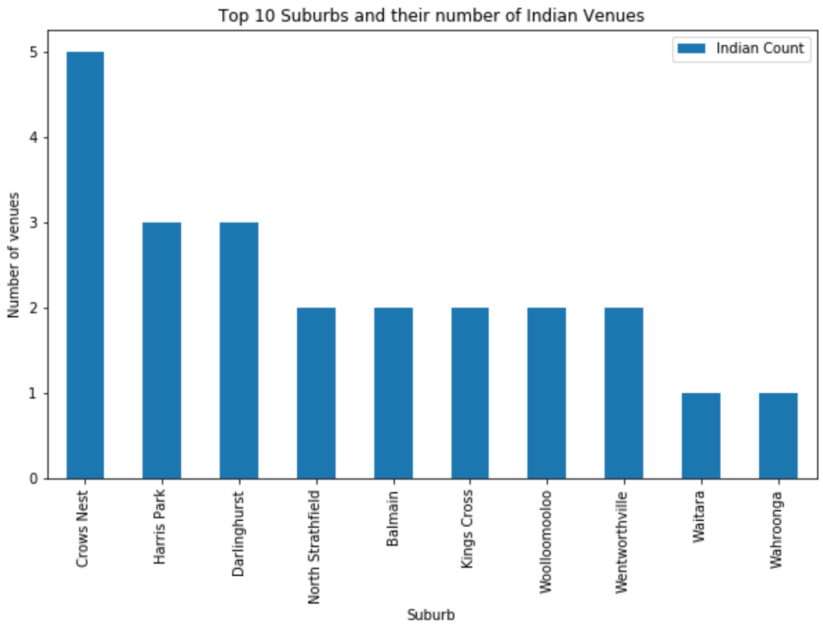
We will repeat the above process for Chinese, Italian, Malay and public transport venues.

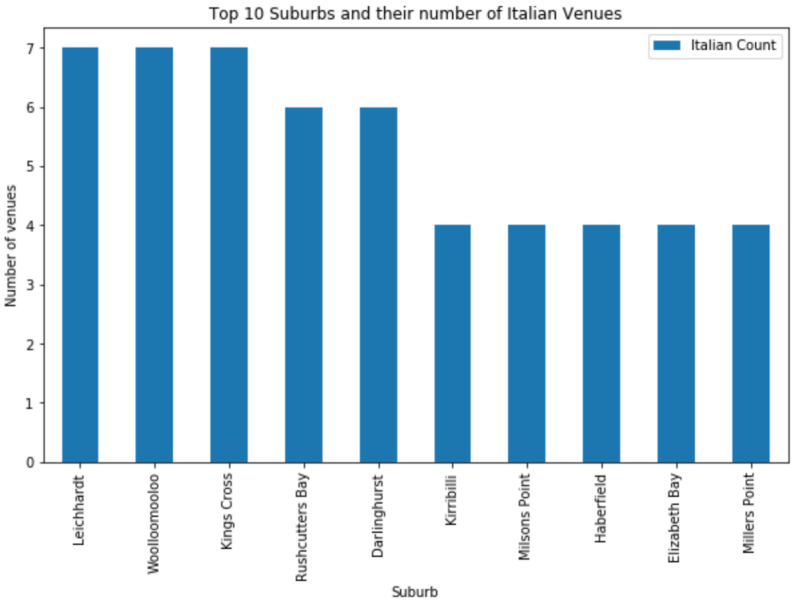
Modelling

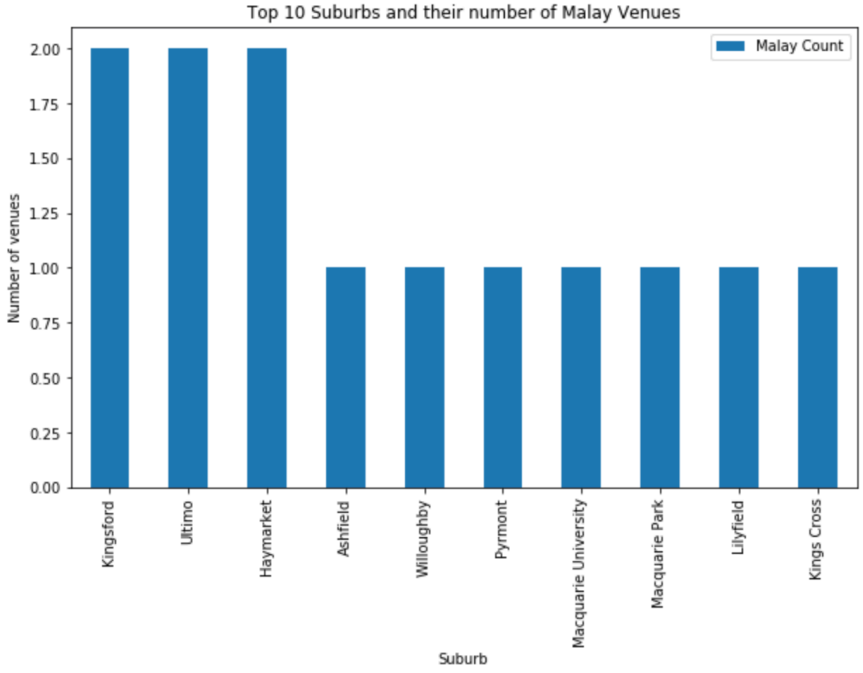
We will use K-means to perform clustering of Sydney suburbs according to their similarities by venues. The one hot encoding dataframe is used as input for the K-means model. We have set the number of cluster to be 5 in this exercise.
The resulting information is useful as housing cost of some suburbs is high even though they are ideal place to live. With the finding of the K-means clustering, we can propose other similar suburbs to live in with lower housing cost in future.

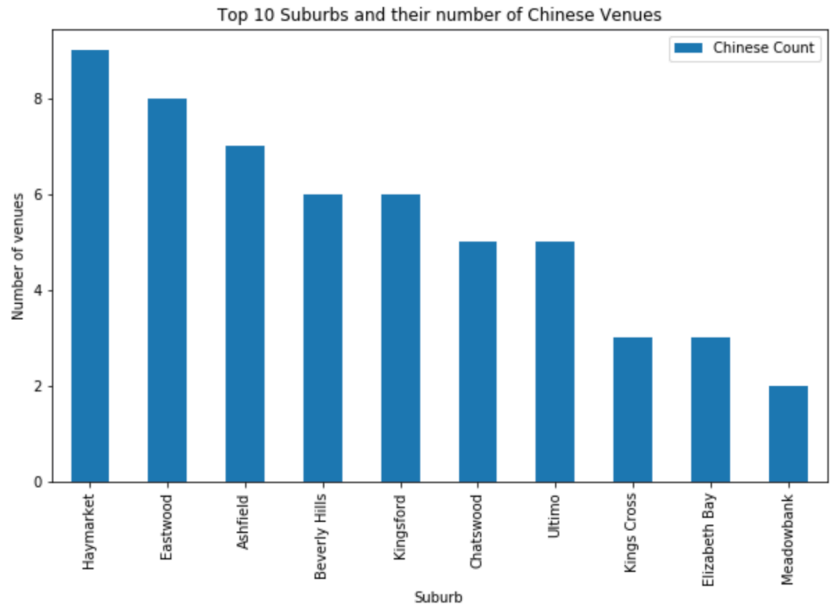
Results

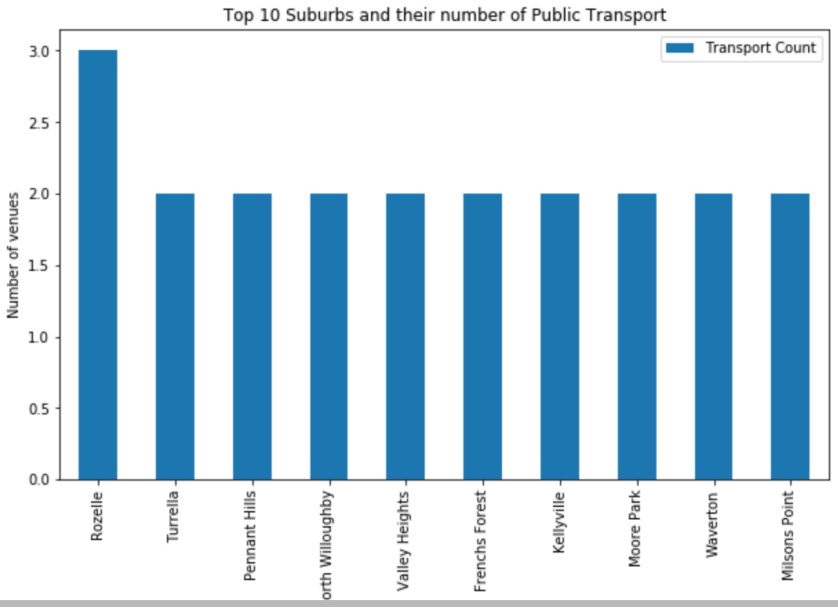
Based on the information we retrieved from FourSquare, we have discovered the top 10 Suburbs with Indian, Malay, Chinese, Italian and Public Transport venues as shown below:




Top 10 Suburbs and their number of Italian Venues
Italian Count
Number of venues
0
1
2
3
4
5
6
7
Leichhardt
Woolloomooloo
Kings Cross
Rushcutters Bay
Darlinghurst
Kirribilli
Milsons Point
Haberfield
Elizabeth Bay
Millers Point
Suburb

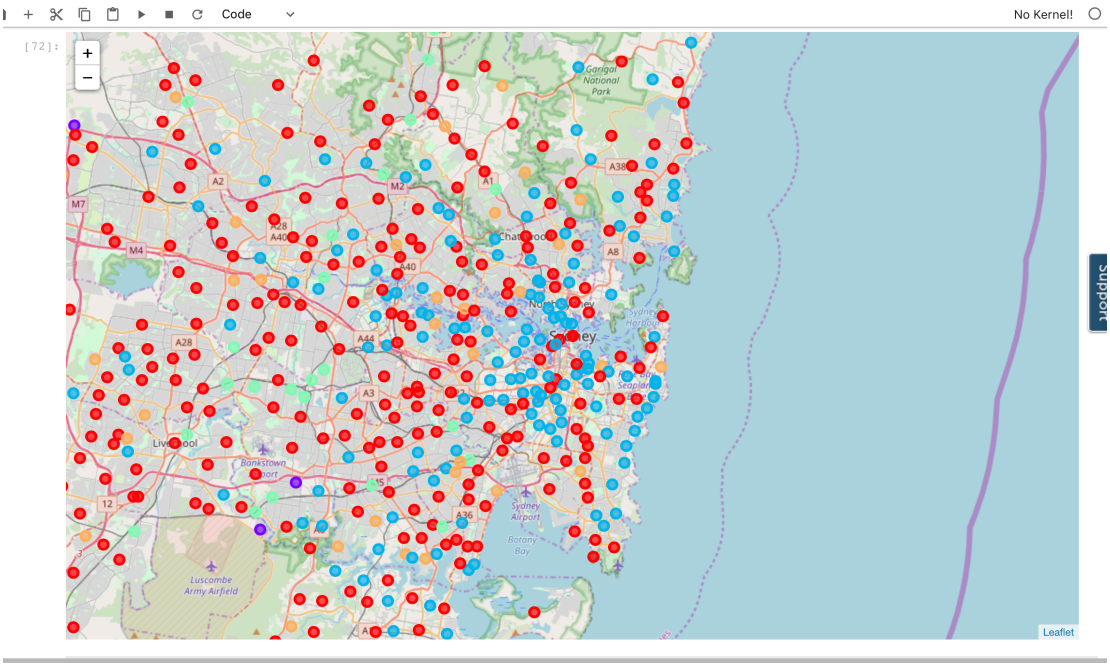

Top 10 Suburbs and their number of Malay Venues
Malay Count
Number of venues
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Kingsford
Ultimo
Haymarket
Ashfield
Willoughby
Pyrmont
Macquarie University
Macquarie Park
Lilyfield
Kings Cross
Suburb





Clustering of Suburbs

After running K-means clustering with 5 clusters, the result is shown in the following map:



Discussion of Results

Our analysis is heavily based on the information obtained by running FourSquare. Firstly, we need to investigate whether FourSquare has returned accurate set of venues for our suburbs. The results is also depends on the LIMIT and Radius that we have to set when passing parameters to API. We have set the LIMIT to 100 and Radius to 500 metres. We need to consider the following questions:

If we change the value of these two parameters, will FourSquare return the similar results?

Which are the best parameters to use in order to give the most accurate results?

Secondly, we have observed the venues returned by FourSquare related to Indian, Chinese etc. are mainly restaurants. We have no information of other important venues such as grocery stores.

Thirdly, we have observed FourSquare failed to return any School venues.

Finally, the public transport venues returned by FourSquare are limited. We have found the top suburb only have 3 venues and further down the top suburb list, all suburbs are having 2 venues. Therefore based on the limited data we have, it is hard for us to make conclusion whether which suburbs have better public transport

facilities. We might have to search for other data source for this part of the project.

Conclusion

From our analysis we will recommend the best suburbs for Indian migrants are: Crows Nest, Harris Park, Darlinghurst, North Strathfield, Balmain, King Cross, Woolloomoolo, Wentworthvile, Waitara and Wahroonga.

For Malay migrants are: Kingsford, Ultimo, Haymarket, Ashfield, Willoughby, Prymont, Macquarie University, Macquarie Park, Lilyfield, Kings Cross.

For Italian migrants are: Leichhardt, Woolloomooloo, Kings Cross, Rushcutters Bay, Darlinghurst, Kirribirri, Milsons Point, Harberfield, Elizabeth Bay, Millers Point.

For Chinese migrants are: Haymarket, Eastwood, Ashfield, Beverly Hills, Kingsford, Chatswood, Ultimo, Kings Cross, Elizaberth Bay, Meadowbank.

Future Extension of Project

There are certainly a few parts of our results are not satisfying and should be researched on in future. The following are suggestions of future works:

- Schools information
- Public transport
- Venues to include grocery stores
- Find out the housing cost for similar suburbs in order to provide housing options for the customers