2021

IBM Applied Data Science Capstone

RECOMMENDING A SUBURB TO START AN INDIAN RESTURANT

ANEESH RAJ

Introduction

Hospitality industry in Sydney is one of the highly competitive industries. Often it is daunting to make a decision on the location to start a restaurant without considering all the facts. Most of the time these decisions are made subjectively. With the advancement in data science technology, these type of business decisions can be made objectively through data thereby improving the chance of success considerably for the business.

Business Problem

The objective of this project is to analyse the two local government areas (LGA) in Sydney (Parramatta and Sutherland Shire) and recommend the three best suburbs in one of these LGAs to start an Indian restaurant. The target audience for this project is entrepreneurs who is thinking about opening an Indian restaurant in any of the two LGAs in Sydney.

Data

In order to find a solution for the business problem, the following datasets are required.

- List of suburbs in the two LGAs
- Latitude and Longitude of all the suburbs
- Population density of the two LGAs to gauge potential customer traffic
- Median household income of the two LGAs
- Number of business in accommodation and food services in the two LGAs
    - Business entry vs business exit
- Number of people employed in accommodation and food services in the two LGAs

We can use Australian bureau of statistics website to retrieve the population density, median household income, number of business in accommodation and food services and its entry vs exit overtime, number of people employed in the accommodation and food services industry to gauge availability of skilled workforce. Based on this information, we can select the LGA that is best suited to open the Indian restaurant. This can be done via the read_html function in pandas.

The Wiki page 'https://en.wikipedia.org/wiki/Template:Sydney_Sutherland_suburbs' and 'https://en.wikipedia.org/wiki/Template:Sydney_Parramatta_suburbs' provides the list of suburbs in

the two LGAs. We can then use read_html function in pandas to extract the suburb tables in the wiki and convert it to pandas dataframe. Then we can use Python's geopy package to obtain the latitude and longitude of all the suburbs present in the dataframe. With this we should be able to retrieve the available venue details of the suburbs in the selected LGA using the Foursquare API. This can be used to select the suburbs suited to open an Indian restaurant. The example below shows the 10 common venues in suburbs of Parramatta LGA.

	Suburb	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 Common Venue
0	Camellia,NSW	Pier	Electronics Store	Climbing Gym	Fast Food Restaurant	Café	Middle Eastern Restaurant	Waterfront	Fried Chicken Joint	Discount Store	Donut Shop
1	Carlingford,NSW	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Café	Supermarket	Shopping Mall	Korean BBQ Restaurant	Soccer Field	Music Store	Donut Shop
2	Clyde,NSW	Lebanese Restaurant	Dessert Shop	Platform	Fast Food Restaurant	Furniture / Home Store	Convenience Store	Train Station	Pub	Racetrack	Buffet
3	Dundas Valley,NSW	Park	Athletics & Sports	Pub	Café	Burger Joint	Farmers Market	Football Stadium	Food & Drink Shop	Food	Fast Food Restaurant
4	Dundas,NSW	Train Station	Pizza Place	Park	Café	Sports Club	Library	Grocery Store	Seafood Restaurant	Convenience Store	Home Service

Methodology

Initial step is to collect the data for both Parramatta and Sutherland LGAs which includes the population density, number of business and jobs related to the hospitability industry, media weekly household income and unemployment rate. These data will tell us which LGA is best suited to start the Indian restaurant business.

This can be achieved by scraping the following websites from Australian Bureau of statistics.

'https://itt.abs.gov.au/itt/query.jsp?method=GetGenericData&datasetid=ABS_REGIONAL_LGA2019&or=MEASURE&and=LGA_2019.16260,FREQUENCY.A&TIME_FORMAT=P1Y&periods=2014,2015,2016,2017,2018,2019&format=csv&order=chunked&filename=Parramatta%20(C)'

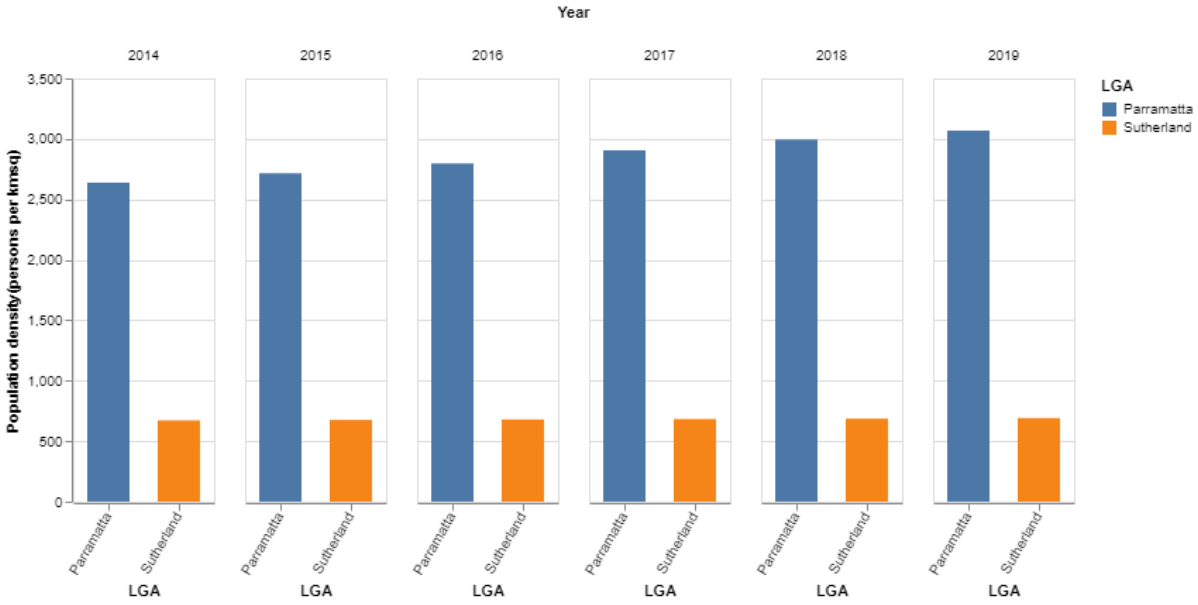
'https://itt.abs.gov.au/itt/query.jsp?method=GetGenericData&datasetid=ABS_REGIONAL_LGA2019&or=MEASURE&and=LGA_2019.17150,FREQUENCY.A&TIME_FORMAT=P1Y&periods=2014,2015,2016,2017,2018,2019&format=csv&order=chunked&filename=Sutherland%20Shire%20(A)'

The filtered data retrieved will be as follows.

	LGA	Parent Description	Description	2011	2014	2015	2016	2017	2018	2019
0	Parramatta	Population Density - As at 30 June	Persons (persons/km2)	NaN	2637.6	2715.6	2796.7	2905.4	2995.0	3068.1
1	Parramatta	Number of Businesses by Industry - As at 30 June	Accommodation and food services (no.)	NaN	NaN	956.0	1002.0	1021.0	1068.0	1079.0
3	Parramatta	Equivalised Total Household Income - Census	Median equivalised total household income (wee...	871.0	NaN	NaN	1012.0	NaN	NaN	NaN
4	Parramatta	Jobs In Australia - Year ended 30 June	Number of Employee Jobs - Accommodation and fo...	NaN	10148.0	10871.0	10861.0	12354.0	NaN	NaN
5	Parramatta	Labour Force Status - Persons aged 15 years an...	Unemployment rate (%)	5.9	NaN	NaN	7.0	NaN	NaN	NaN
7	Sutherland	Population Density - As at 30 June	Persons (persons/km2)	NaN	671.2	675.1	678.9	682.6	686.4	691.3
8	Sutherland	Number of Businesses by Industry - As at 30 June	Accommodation and food services (no.)	NaN	NaN	709.0	721.0	747.0	740.0	722.0
10	Sutherland	Equivalised Total Household Income - Census	Median equivalised total household income (wee...	977.0	NaN	NaN	1136.0	NaN	NaN	NaN
11	Sutherland	Jobs In Australia - Year ended 30 June	Number of Employee Jobs - Accommodation and fo...	NaN	9920.0	10200.0	9640.0	9932.0	NaN	NaN
12	Sutherland	Labour Force Status - Persons aged 15 years an...	Unemployment rate (%)	3.5	NaN	NaN	3.5	NaN	NaN	NaN

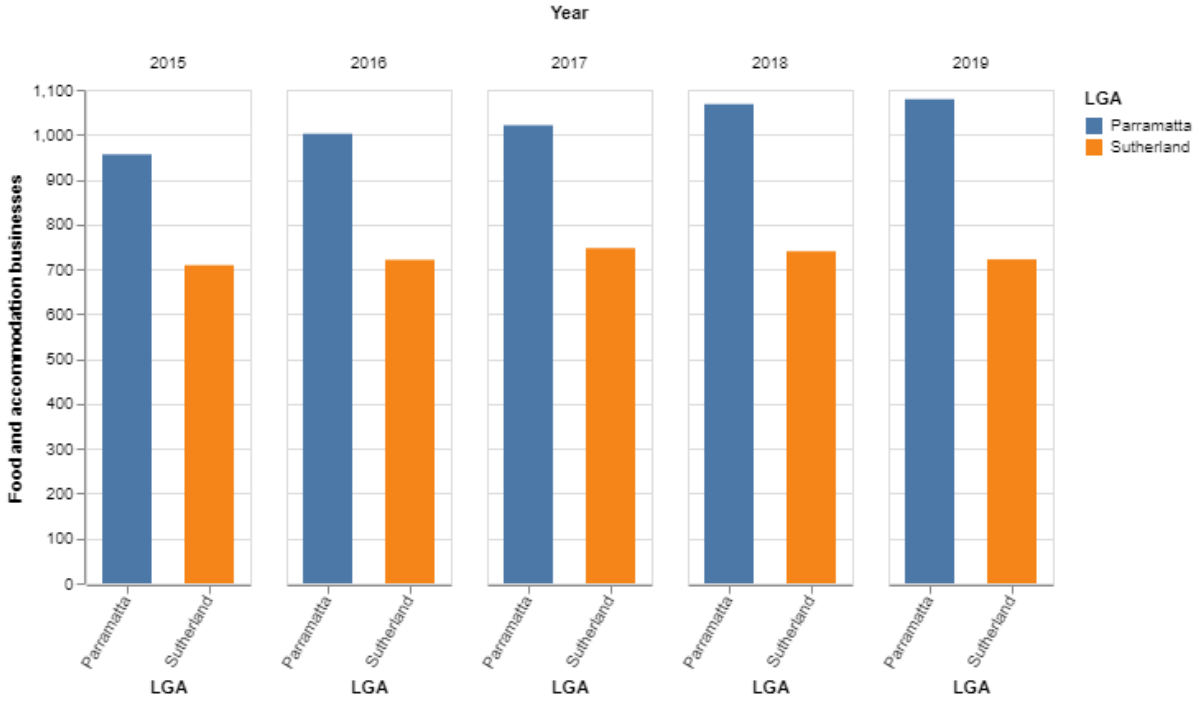
Then some exploratory data analysis is performed to understand the trends. I used the grouped bar chart for comparison of key factors between the two LGAs.

Population density



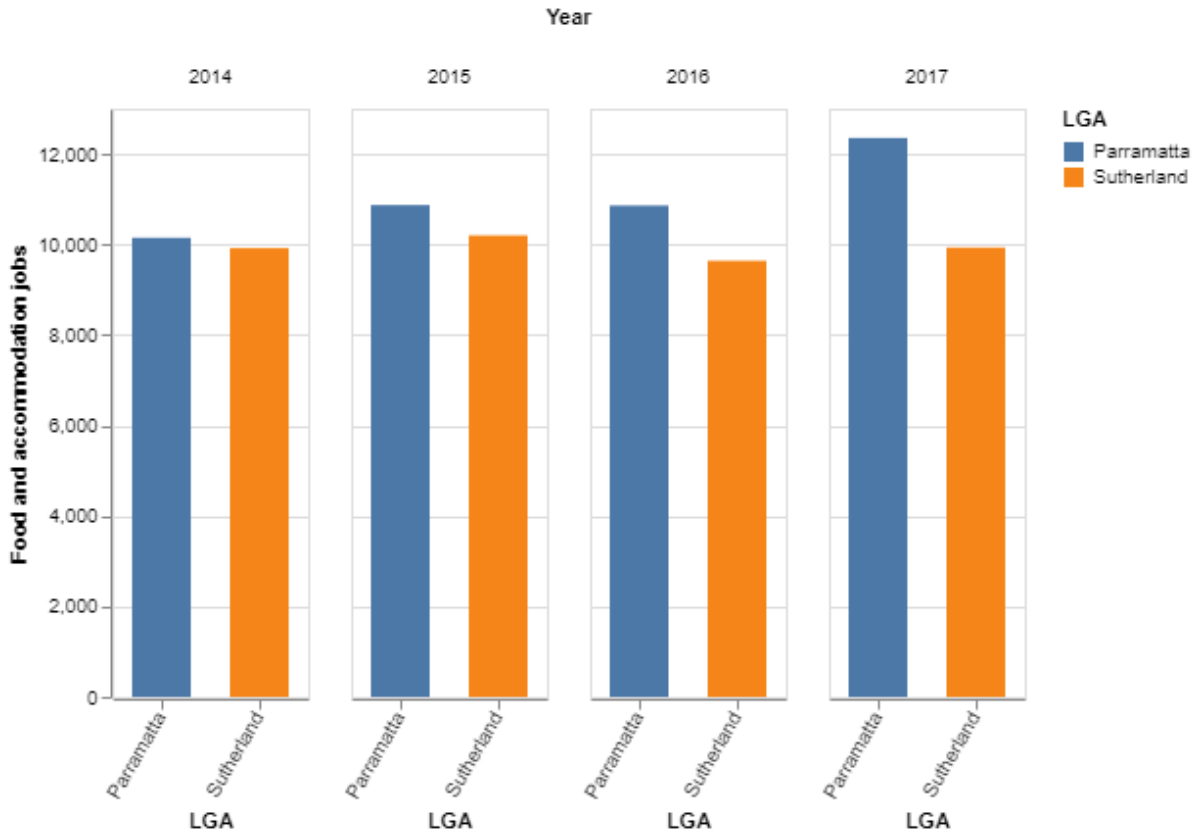
As show in the chart above, the population density of Parramatta LGA is much greater than Sutherland. This shows that the potential customer foot traffic will be much greater in Parramatta LGA than Sutherland LGA.

Businesses in food and accommodation sector



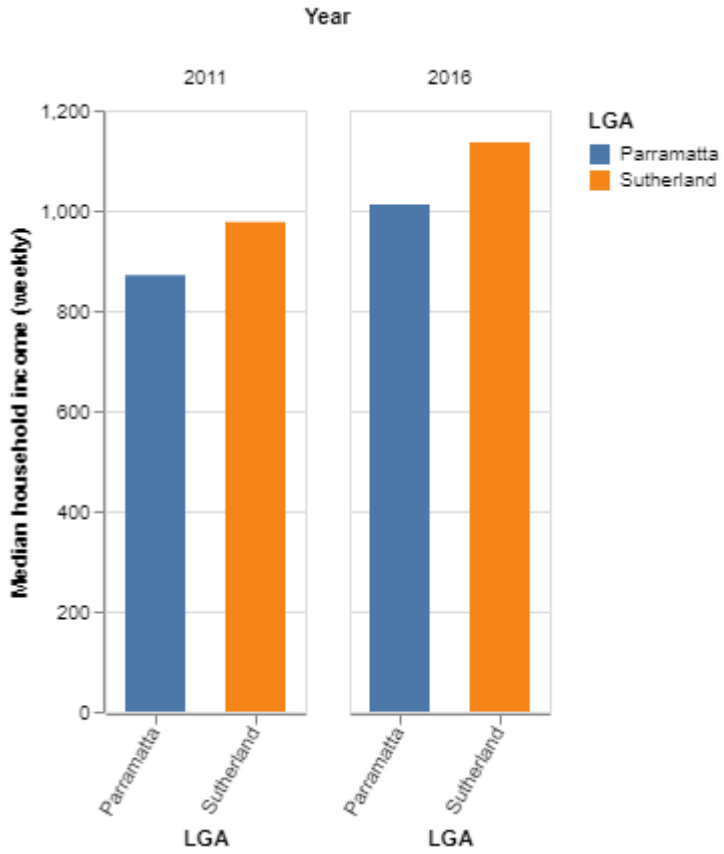
The growth in the number of food and accommodation business is very evident in paramatta LGA as shown above when compared to Sutherland LGA which is pretty flat over the course of four years.

Jobs in food and accommodation business



There seems to be growth in number of jobs in food and accommodation sector in Parramatta LGA which suggests the availability of skilled workforce as well as growth food and accommodation business.

Median household weekly income



Median household weekly income is a bit higher in Sutherland LGA. But this factor is overridden by the population density which is much higher in Parramatta.

The next step is to retrieve the list of suburbs in both Parramatta and Sutherland LGAs. This is achieved by scraping the websites 'https://en.wikipedia.org/wiki/Template:Sydney_Sutherland_suburbs' and

'https://en.wikipedia.org/wiki/Template:Sydney_Parramatta_suburbs'.

After the list of suburbs are retrieved, its corresponding coordinates are gathered using the geopy library.

	Suburb	Longitude	Latitude
0	Camellia,NSW	151.034649	-33.819780
1	Carlingford,NSW	151.047521	-33.774495
2	Clyde,NSW	151.017066	-33.835975
3	Constitution Hill,NSW	151.246320	-31.908768
4	Dundas,NSW	151.044059	-33.802949

Then the venues in these suburbs are collected using the Foursquare API. One hot encoding is performed to get the count of the venues in these suburbs. Then a list of five common venues is retrieved because the focus of this report is to identify only the restaurant venues.

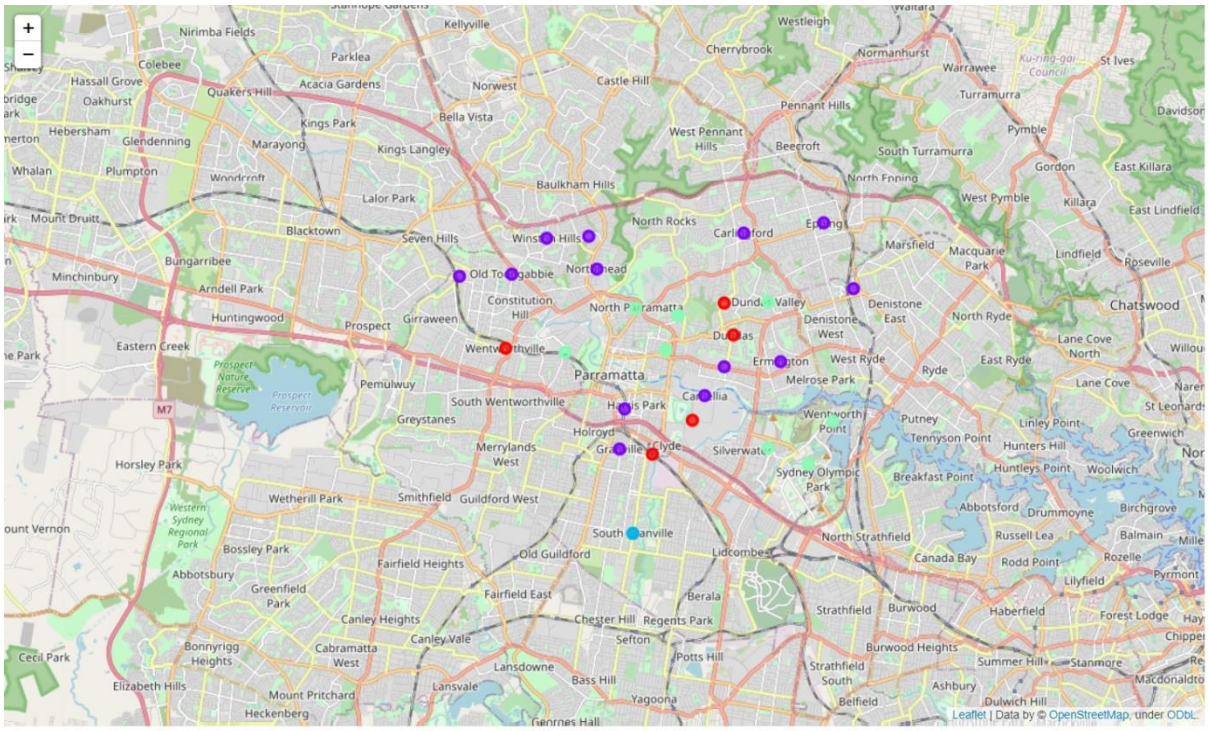
	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Camellia,NSW	Electronics Store	Pier	Gym	Café	Middle Eastern Restaurant
1	Carlingford,NSW	Chinese Restaurant	Fast Food Restaurant	Shopping Mall	Pizza Place	Sandwich Place
2	Clyde,NSW	Lebanese Restaurant	Dessert Shop	Train Station	Convenience Store	Furniture / Home Store
3	Dundas Valley,NSW	Pub	Athletics & Sports	Café	Park	Burger Joint
4	Dundas,NSW	Train Station	Sports Club	Library	Seafood Restaurant	Grocery Store

After that it is K means clustering is performed to identify suburbs that has at least restaurant as 1st , 2nd or 3rd common venue. Then it is merged with the suburb list data frame that includes the coordinates.

	Suburb	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Camellia,NSW	151.034649	-33.819780	0.0	Electronics Store	Pier	Gym	Café	Middle Eastern Restaurant
1	Carlingford,NSW	151.047521	-33.774495	0.0	Chinese Restaurant	Fast Food Restaurant	Shopping Mall	Pizza Place	Sandwich Place
2	Clyde,NSW	151.017066	-33.835975	0.0	Lebanese Restaurant	Dessert Shop	Train Station	Convenience Store	Furniture / Home Store
3	Constitution Hill,NSW	151.246320	-31.908768	NaN	NaN	NaN	NaN	NaN	NaN
4	Dundas,NSW	151.044059	-33.802949	0.0	Train Station	Sports Club	Library	Seafood Restaurant	Grocery Store

Results & Discussion

Based on the EDA performed on the LGAs, its clearly evident that Parramatta LGA is best suited for starting an Indian restaurant than Sutherland LGA. So, the cluster analysis is performed on the venues in Parramatta LGA to narrow done on the potential suburbs to start the Indian restaurant.



The colours red, purple, and green represents cluster 0, 1, and 3 respectively. What this shows is that clusters in green have the common venues as café as shown below.

	Suburb	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Dundas Valley,NSW	151.055907	-33.793674	3.0	Burger Joint	Athletics & Sports	Pub	Park	Café
12	Newington,NSW	151.055872	-33.834230	3.0	Café	Sandwich Place	Japanese Restaurant	Seafood Restaurant	Shopping Mall
14	Oatlands,NSW	151.025653	-33.797339	3.0	Café	Grocery Store	Golf Course	Bus Stop	Waterfront
15	Oatlands,NSW	151.025653	-33.797339	3.0	Café	Grocery Store	Golf Course	Bus Stop	Waterfront
17	Parramatta,NSW	151.021363	-33.806970	3.0	Café	Pizza Place	Arts & Crafts Store	Bakery	Sandwich Place
18	Parramatta,NSW	151.021363	-33.806970	3.0	Café	Pizza Place	Arts & Crafts Store	Bakery	Sandwich Place
19	North Parramatta,NSW	151.011665	-33.795275	3.0	Café	Lake	Gym	Bus Stop	Pet Store
26	Sydney Olympic Park,NSW	151.069092	-33.838740	3.0	Café	Italian Restaurant	Stadium	Scenic Lookout	Athletics & Sports
29	Wentworth Point,NSW	151.077435	-33.826896	3.0	Café	Park	Waterfront	Japanese Restaurant	Shopping Mall
32	Westmead,NSW	150.987727	-33.807650	3.0	Café	Bus Station	Steakhouse	Platform	Australian Restaurant

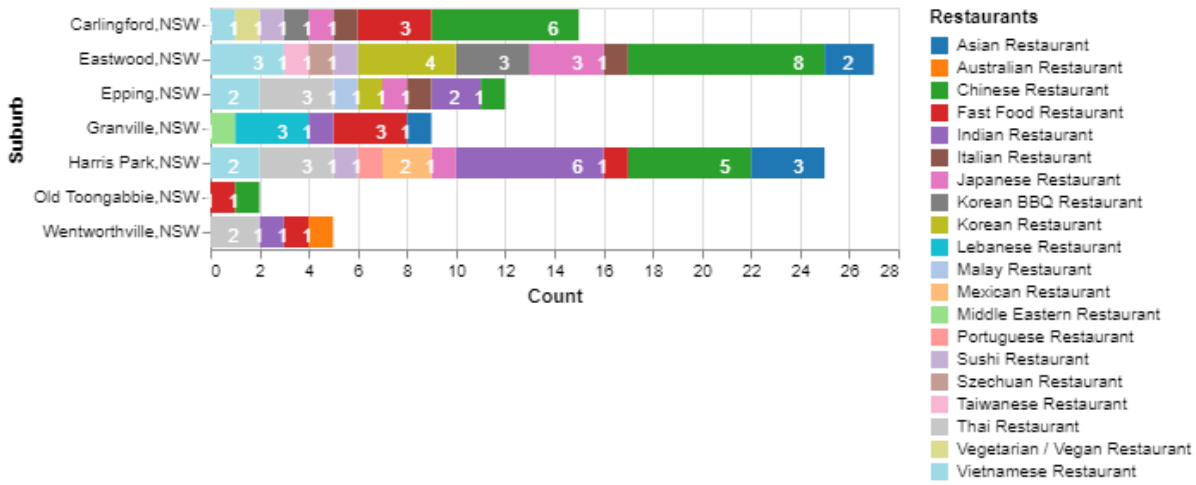
whereas the cluster in red has common venues as train stations as shown below.

	Suburb	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Clyde,NSW	151.017066	-33.835975	0.0	Dessert Shop	Lebanese Restaurant	Furniture / Home Store	Fast Food Restaurant	Convenience Store
4	Dundas,NSW	151.044059	-33.802949	0.0	Train Station	Sports Club	Home Service	Australian Restaurant	Café
21	Rosehill,NSW	151.030556	-33.826389	0.0	Hotel	Platform	Train Station	Stadium	Racetrack
27	Telopea,NSW	151.040944	-33.793922	0.0	Gas Station	Train Station	Convenience Store	Grocery Store	Soccer Field
30	Wentworthville,NSW	150.967778	-33.806667	0.0	Thai Restaurant	Imported Food Shop	Grocery Store	Pizza Place	Platform

Cluster in green has what is required, the most common venues as restaurants. So that means it is the suburbs listed below that is best suited to start an Indian restaurant.

	Suburb	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Camellia,NSW	151.034649	-33.819780	1.0	Café	Electronics Store	Dentist's Office	Pier	Shipping Store
1	Carlingford,NSW	151.047521	-33.774495	1.0	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Café	Sandwich Place
6	Eastwood,NSW	151.084444	-33.790000	1.0	Chinese Restaurant	Korean Restaurant	Café	Vietnamese Restaurant	Korean BBQ Restaurant
7	Epping,NSW	151.074537	-33.771855	1.0	Thai Restaurant	Platform	Indian Restaurant	Vietnamese Restaurant	Pizza Place
8	Ermington,NSW	151.060056	-33.810116	1.0	Café	Park	Fast Food Restaurant	Liquor Store	Italian Restaurant
9	Granville,NSW	151.006011	-33.834510	1.0	Lebanese Restaurant	Fast Food Restaurant	Dessert Shop	Convenience Store	Platform
10	Harris Park,NSW	151.007654	-33.823338	1.0	Indian Restaurant	Chinese Restaurant	Café	Sandwich Place	Asian Restaurant
11	Model Farms,NSW	150.995659	-33.775558	1.0	Pizza Place	Coffee Shop	Bus Stop	Bowling Green	Café
13	Northmead,NSW	150.998329	-33.784442	1.0	Gym	Bowling Green	Shopping Mall	Farmers Market	Park
16	Old Toongabbie,NSW	150.969953	-33.785855	1.0	Park	Fast Food Restaurant	Chinese Restaurant	Gym	Grocery Store
22	Rydalmere,NSW	151.041170	-33.811709	1.0	Electronics Store	Café	Arts & Crafts Store	Supermarket	Park
28	Toongabbie,NSW	150.952650	-33.786689	1.0	Gym	Convenience Store	Bar	Platform	Fast Food Restaurant
31	Winston Hills,NSW	150.981661	-33.776110	1.0	Bus Stop	Italian Restaurant	Pub	Park	Chinese Restaurant

Now this has to be narrowed down to few suburbs. A filter is applied to choose only the suburbs that has restaurants in at least two of the first three common venues columns. Then only the restaurant venues are included and all other venues are filtered out. After that the data frame can be visualized as below.



From the above chart, it can be seen that the top 3 suburbs to start an Indian restaurant is Eastwood, Carlingford and Epping. This is because these suburbs have the greatest number of restaurants. Harris Park is ignored because it has the highest number of Indian restaurant and therefore it will be very highly competitive. Eastwood has 27 restaurants but there are no Indian

restaurants. Carlingford and 15 restaurants and no Indian restaurant. Epping has 12 restaurants and only 2 Indian restaurants. So, it is recommended to open an Indian restaurant in any of these three suburbs.

Conclusion

This project has made use of Australian Bureau of Statistics website to get data related to the food and accommodation industry in Parramatta and Sutherland LGA. Exploratory data analysis is performed on that data to choose the LGA that is best suited to start an Indian restaurant. Then Foursquare API is used to get the restaurant locations situated in Parramatta LGA as it seems to be the best to start an Indian restaurant. K-means clustering algorithm has been used to cluster the suburbs in Parramatta LGA to find out the suburbs that has restaurants as the 1st, 2nd and 3rd most common venue. Then best 3 suburbs are recommended to start an Indian restaurant based on the customer foot traffic (Based on highest concentration of restaurants) as well as least amount of direct competition (By eliminating suburbs with high number of Indian restaurants).