IBM Data Science Capstone Project

by Nicolas Dinh

1. Introduction

This project concludes of the IBM Data Science Certification Specialization. This goal of this last module is about to apply the knowledge and skills acquired during the course to a real-life problem. This project will be about solving a hypothetical business problem with real data (geolocation data from Foursquare using RESTful API calls as well as data scrapped directly from the web using libraries like BeautifulSoup) and analyzing it using Python in a Jupyter Notebook.

Similar to the approach that was taken during the course with New York and Toronto, the focus of this project will be about finding an optimal neighborhood (suburb) in the city of Melbourne in order to open a French restaurant.

2. Business Problem

Melbourne is a city with a very multicultural population without over 200 nationalities and over 230 languages spoken. The city is the nest of a large amounts of communities and hosts numerous cultural events from music, international festivals to shows, expositions and galleries. And one of the big results of this multicultural aspect is the diversity of restaurants and food that the city has to offer.

My client, a French chef and entrepreneur asked me to take a data-driven approach to find a good place to open a French restaurant. Ideally, it would close to the city center and in a popular suburb outside of the CBD that already offer a large choice of restaurants, bars and cultural events/gathering.

3. Data gathering

In order to solve the above business problem, the following data would be required:

1. A list of the suburbs in Melbourne close to the CBD: this data can be scrapped from Wikipedia (https://en.wikipedia.org/wiki/List_of_Melbourne_suburbs).
2. The Geo-coordinates (latitude, longitude) associated to the Melbourne suburbs that can be obtained using the Geocoder library.
3. The popular venues data of each suburb from Foursquare using RESTful API calls.

4. Methodology

DATA COLLECTION AND PREPROCESSING

Firstly, in order to get the data from Wikipedia, we need to use the *BeautifulSoup* library in order to scrap the list of suburbs in the inner city of Melbourne.

Figure 1 - Wikipedia page layout

We would then store the data in a DataFrame and clean it in order to remove the duplicated suburbs (e.g., the Melbourne suburbs sometimes cross about different municipality and we only used the main suburb in this analysis).

	Postal Code	Municipality	Suburb
0	3053	City of Melbourne	Carlton
1	3054	City of Melbourne	Carlton North
2	3008	City of Melbourne	Docklands
3	3002	City of Melbourne	East Melbourne
4	3031	City of Melbourne	Flemington

Figure 2 - Header of the list of suburbs

Secondly, we would gather the geo-coordinates of each suburbs using the *Nominatim* function from the *geopy* library. We merge the data obtained from it in the main DataFrame as per below.

	Postal Code	Municipality	Suburb	Latitude	Longitude
0	3053	City of Melbourne	Carlton	-37.800144	144.969268
1	3054	City of Melbourne	Carlton North	-37.784559	144.972855
2	3008	City of Melbourne	Docklands	-37.817870	144.938654
3	3002	City of Melbourne	East Melbourne	-37.815588	144.982555
4	3031	City of Melbourne	Flemington	-37.787277	144.923242

Figure 3 - Header of the list of suburbs with their respective geo-coordinates (latitude, longitude)

This then allows us to visualize each suburbs on a map using the *Folium* library in Python:

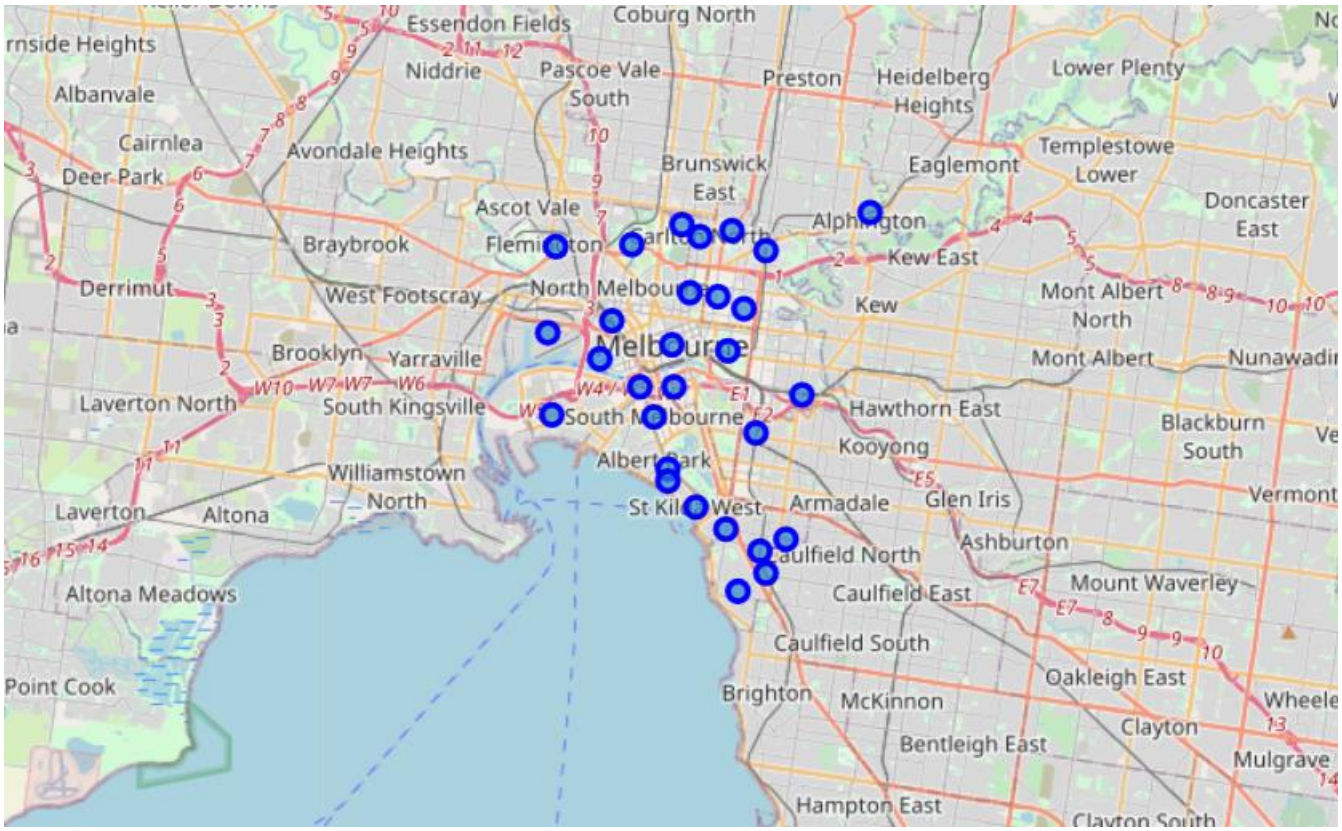

Figure 4 - Visualisation of each suburbs in a map of Melbourne

ANALYSIS OF THE SUBURBS USING THE FOURSQUARE DATA

Finally, we would gather of the top 100 trending venues in a radius of 500m of that particular suburb using the Foursquare data, and collecting it via RESTful API calls.

Using a developer account on the Foursquare platform, we can collect the data in a json format using the Foursquare API.

As a first step, let's analyze one suburb (Carlton) in order to look at the output of its top venues. We collect 4 information from each venue: name, category, latitude and longitude.

	name	categories	lat	lng
0	Carlton Wine Room	Wine Bar	-37.798584	144.968610
1	D.O.C. Pizza & Mozzarella Bar	Pizza Place	-37.798954	144.968490
2	Yo-Chi	Frozen Yogurt Shop	-37.798659	144.967849
3	Gewürzhaus	Gourmet Shop	-37.799050	144.967480
4	Baker D. Chirico	Bakery	-37.798788	144.968499

Figure 5 - Top 5 venues in Carlton

We then apply the same logic but to collect all venues from each suburb and we look at which suburb contains the most venues in the 500m radius:

Suburb	Venue
Melbourne	100
Fitzroy	100
South Yarra	91
South Melbourne	60
Collingwood	57
Balaclava	53
Southbank	47
Carlton	47
Docklands	44
Fitzroy North	39

Figure 6 - Top 10 suburbs with the most venues

Secondly, we run another analysis to look at the frequency of occurrence of each category of venues. This would allow us to find the top 10 most frequent category of venues in each suburb as per below:

	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	Abbotsford	Café	Wine Shop	Italian Restaurant	Pizza Place	Park	Thai Restaurant	Zoo Exhibit	Flea Market	Fruit & Vegetable Store	Frozen Yogurt Shop
1	Albert Park	Café	Metro Station	Tennis Court	Golf Course	Seafood Restaurant	Food & Drink Shop	Athletics & Sports	Hotel	Indian Restaurant	Racetrack
2	Alphington	Liquor Store	Gym / Fitness Center	Convenience Store	Farmers Market	Fast Food Restaurant	Park	Train Station	Thai Restaurant	Flower Shop	Fruit & Vegetable Store
3	Balaclava	Café	Coffee Shop	Breakfast Spot	Bar	Pharmacy	Pizza Place	Vietnamese Restaurant	Tram Station	Salad Place	Japanese Restaurant
4	Burnley	Café	Pub	Furniture / Home Store	Breakfast Spot	Convenience Store	Park	Cocktail Bar	Shop & Service	Liquor Store	Food & Drink Shop

Figure 7 - Most popular venue categories in each suburb

CLUSTERING

Now that we've collected, wrangled and pre-processed the data we need, we can use a clustering algorithm. In our case, we use a popular unsupervised machine learning technique named K-means clustering.

We need to first find the optimal number of cluster by calculating the distortion in function of the number of cluster k and visualizing it (see below).

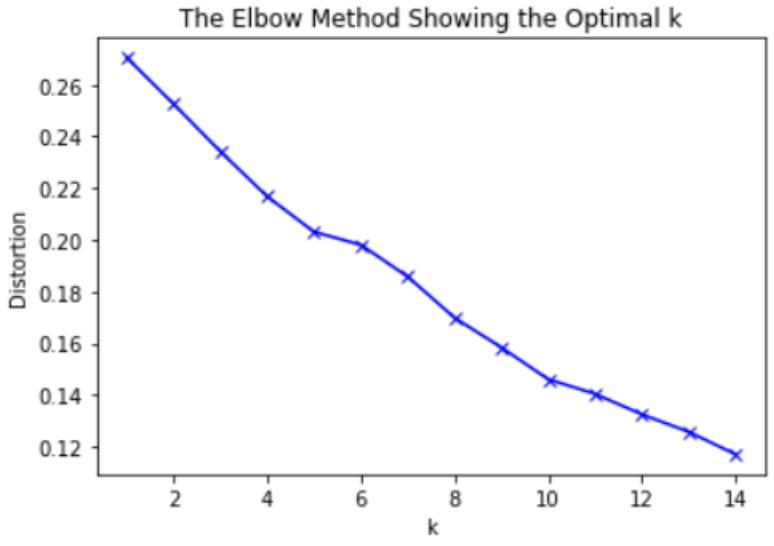

Figure 8 - Distortion in function of the number of cluster k

Using the elbow method, we choose to use 5 clusters for our clustering analysis.

The *scikit-learn* library allows us to cluster each suburbs and label it with a number from 0 to 4. We obtain the following DataFrame:

	Postal Code	Municipality	Suburb	Latitude	Longitude	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	10th Most Common Venue
0	3053	City of Melbourne	Carlton	-37.800144	144.969268	0	Italian Restaurant	Café	Coffee Shop	Ice Cream Shop	Gourmet Shop	French Restaurant	Indie Theater	Park	Cheese Shop	Pub
1	3054	City of Melbourne	Carlton North	-37.784559	144.972855	3	Café	Bakery	Tram Station	Wine Bar	Flower Shop	Pub	Deli / Bodega	Italian Restaurant	Grocery Store	Liquor Store
2	3008	City of Melbourne	Docklands	-37.817870	144.938654	0	Italian Restaurant	Middle Eastern Restaurant	Café	Restaurant	Pier	Chinese Restaurant	Seafood Restaurant	Steakhouse	Sandwich Place	Tapas Restaurant
3	3002	City of Melbourne	East Melbourne	-37.815588	144.982555	0	Café	Cricket Ground	Hotel	Restaurant	Bar	Sushi Restaurant	Museum	Stadium	Coffee Shop	Tourist Information Center
4	3031	City of Melbourne	Flemington	-37.787277	144.923242	0	Hotel	Racecourse	Pizza Place	Tram Station	Park	Bowling Green	Pharmacy	Supermarket	Café	French Restaurant
5	3031	City of Melbourne	Kensington	-33.911719	151.219506	0	Pizza Place	Sushi Restaurant	Chinese Restaurant	Convenience Store	Kids Store	Gym / Fitness Center	Cosmetics Shop	Malay Restaurant	Liquor Store	Burger Joint
6	3000	City of Melbourne	Melbourne	-37.814218	144.963161	0	Coffee Shop	Bar	Café	Cocktail Bar	Dessert Shop	Shopping Mall	Clothing Store	Italian Restaurant	Cosmetics Shop	Candy Store
7	3051	City of Melbourne	North Melbourne	-37.807609	144.942351	3	Café	Sandwich Place	Theater	Zoo Exhibit	Flea Market	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant
8	3052	City of Melbourne	Parkville	-37.787250	144.949596	2	Zoo Exhibit	Park	Hockey Arena	Gift Shop	Food & Drink Shop	Sculpture Garden	Basketball Court	Fast Food Restaurant	BBQ Joint	Sports Club
9	3207	City of Melbourne	Port Melbourne	-37.833361	144.921920	0	Climbing Gym	Café	Go Kart Track	Latin American Restaurant	Beach	Zoo Exhibit	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint

Figure 9 - List of all suburbs and its associated label

5. Results

That allows us to then visualize each cluster on the Melbourne map.

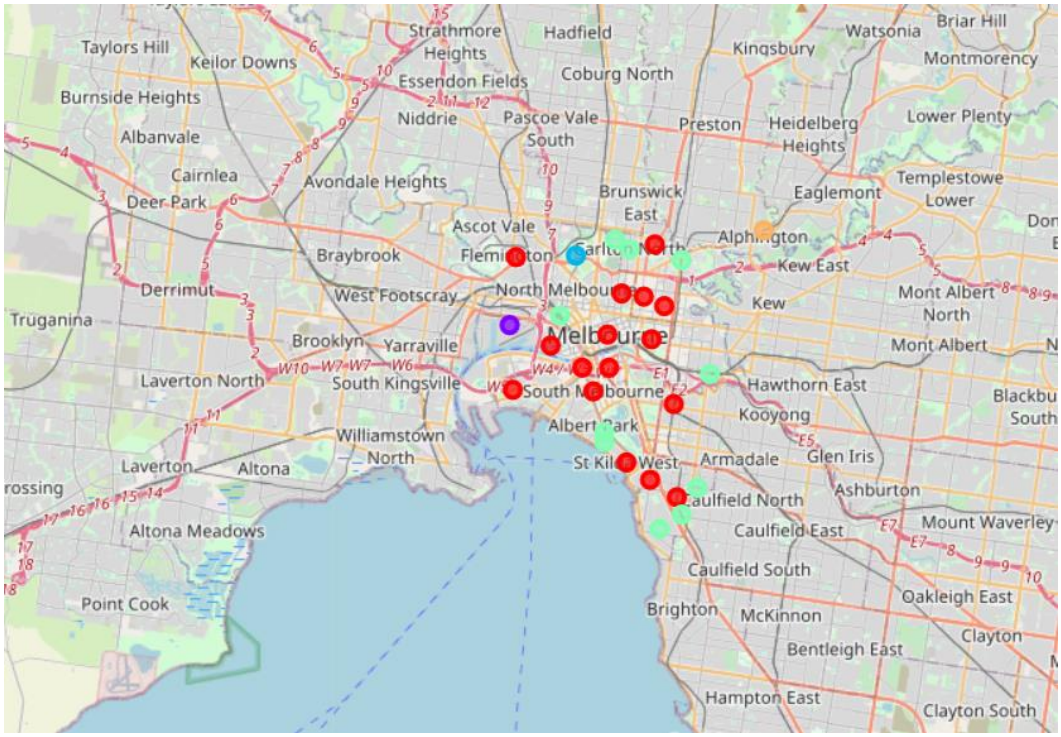

Figure 10 - Visualization of the 5 clusters of suburbs in Melbourne

By looking at the cluster data (see below), we can see that the cluster 1 seems to be the most relevant in order to open a restaurant which is close to a gastronomical center and also close to bars and other popular venues.

The 1st cluster (label 0) contains a lot of Restaurant, cafes and bars and is the biggest cluster.

	Municipality	Suburb	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	10th Most Common Venue
0	City of Melbourne	Carlton	0	Italian Restaurant	Café	Coffee Shop	Ice Cream Shop	Gourmet Shop	French Restaurant	Indie Theater	Park	Cheese Shop	Pub
2	City of Melbourne	Docklands	0	Italian Restaurant	Middle Eastern Restaurant	Café	Restaurant	Pier	Chinese Restaurant	Seafood Restaurant	Steakhouse	Sandwich Place	Tapas Restaurant
3	City of Melbourne	East Melbourne	0	Café	Cricket Ground	Hotel	Restaurant	Bar	Sushi Restaurant	Museum	Stadium	Coffee Shop	Tourist Information Center
4	City of Melbourne	Flemington	0	Hotel	Racecourse	Pizza Place	Tram Station	Park	Bowling Green	Pharmacy	Supermarket	Café	French Restaurant
5	City of Melbourne	Kensington	0	Pizza Place	Sushi Restaurant	Chinese Restaurant	Convenience Store	Kids Store	Gym / Fitness Center	Cosmetics Shop	Malay Restaurant	Liquor Store	Burger Joint
6	City of Melbourne	Melbourne	0	Coffee Shop	Bar	Café	Cocktail Bar	Dessert Shop	Shopping Mall	Clothing Store	Italian Restaurant	Cosmetics Shop	Candy Store
9	City of Melbourne	Port Melbourne	0	Climbing Gym	Café	Go Kart Track	Latin American Restaurant	Beach	Zoo Exhibit	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint
10	City of Melbourne	Southbank	0	Café	Hotel	Grocery Store	Bar	Performing Arts Venue	Italian Restaurant	Restaurant	Coffee Shop	Bakery	Australian Restaurant
11	City of Melbourne	South Wharf	0	Hotel	Bar	Clothing Store	Sporting Goods Shop	Australian Restaurant	Multiplex	Restaurant	Plaza	Seafood Restaurant	Shoe Store
12	City of Melbourne	South Yarra	0	Café	Italian Restaurant	Hotel	Japanese Restaurant	Grocery Store	Bakery	Convenience Store	Coffee Shop	Pizza Place	Dessert Shop
15	City of Port Phillip	Balaclava	0	Café	Coffee Shop	Breakfast Spot	Bar	Pharmacy	Pizza Place	Vietnamese Restaurant	Tram Station	Salad Place	Japanese Restaurant
19	City of Port Phillip	St Kilda	0	Café	Fast Food Restaurant	Convenience Store	Japanese Restaurant	Pub	Tram Station	Australian Restaurant	Thai Restaurant	Pizza Place	Theater
21	City of Port Phillip	St Kilda West	0	Café	Tram Station	Italian Restaurant	Hotel Bar	Australian Restaurant	Garden	Music Venue	Beach	Restaurant	Juice Bar
22	City of Port Phillip	South Melbourne	0	Café	Bar	Coffee Shop	Gastropub	Mexican Restaurant	Bakery	Spa	Fish & Chips Shop	Breakfast Spot	Pub
27	City of Yarra	Collingwood	0	Café	Bar	Japanese Restaurant	Gay Bar	Cocktail Bar	Bakery	Kebab Restaurant	Pub	Coffee Shop	Adult Boutique
28	City of Yarra	Cremorne	0	Café	Bar	Park	Shopping Mall	Sushi Restaurant	Korean Restaurant	Grocery Store	Cocktail Bar	Fast Food Restaurant	Movie Theater
30	City of Yarra	Fitzroy	0	Café	Bar	Cocktail Bar	Vietnamese Restaurant	Pub	Bakery	Bookstore	Japanese Restaurant	Wine Bar	Vegetarian / Vegan Restaurant
31	City of Yarra	Fitzroy North	0	Coffee Shop	Café	Tram Station	Pub	Ice Cream Shop	Bakery	Italian Restaurant	Park	Veterinarian	Beer Garden

Figure 11 - Cluster 1

The 2nd cluster only contains 1 suburb with a flea market as its top most popular venue category.

	Municipality	Suburb	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	10th Most Common Venue
13	City of Melbourne	West Melbourne	1	Flea Market	Asian Restaurant	Farmers Market	Flower Shop	Garden	Gaming Cafe	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint

Figure 12 - Cluster 2

The 3rd cluster is located near a Zoo, a Park and a Hockey Arena.

	Municipality	Suburb	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	10th Most Common Venue
8	City of Melbourne	Parkville	2	Zoo Exhibit	Park	Hockey Arena	Gift Shop	Food & Drink Shop	Sculpture Garden	Basketball Court	Fast Food Restaurant	BBQ Joint	Sports Club

Figure 13 - Cluster 3

The 4th cluster is the second biggest cluster and also seems to be a popular place for cafes, pubs and restaurants.

	Municipality	Suburb	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	10th Most Common Venue
1	City of Melbourne	Carlton North	3	Café	Bakery	Tram Station	Wine Bar	Flower Shop	Pub	Deli / Bodega	Italian Restaurant	Grocery Store	Liquor Store
7	City of Melbourne	North Melbourne	3	Café	Sandwich Place	Theater	Zoo Exhibit	Flea Market	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant
14	City of Port Phillip	Albert Park	3	Café	Metro Station	Tennis Court	Golf Course	Seafood Restaurant	Food & Drink Shop	Athletics & Sports	Hotel	Indian Restaurant	Racetrack
16	City of Port Phillip	Elwood	3	Café	Indian Restaurant	Convenience Store	Bakery	Fish & Chips Shop	Bar	River	Event Space	Food & Drink Shop	Furniture / Home Store
17	City of Port Phillip	Middle Park	3	Café	Tram Station	Indian Restaurant	Food & Drink Shop	Seafood Restaurant	Metro Station	Grocery Store	Thai Restaurant	Beach	Playground
18	City of Port Phillip	Ripponlea	3	Café	Molecular Gastronomy Restaurant	Tram Station	Fish & Chips Shop	Grocery Store	Coffee Shop	Park	Train Station	Jewish Restaurant	Pharmacy
20	City of Port Phillip	St Kilda East	3	Pub	Pizza Place	Convenience Store	Café	Tram Station	Zoo Exhibit	Flea Market	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint
23	City of Yarra	Abbotsford	3	Café	Wine Shop	Italian Restaurant	Pizza Place	Park	Thai Restaurant	Zoo Exhibit	Flea Market	Fruit & Vegetable Store	Frozen Yogurt Shop
25	City of Yarra	Burnley	3	Café	Pub	Furniture / Home Store	Breakfast Spot	Convenience Store	Park	Cocktail Bar	Shop & Service	Liquor Store	Food & Drink Shop
26	City of Yarra	Clifton Hill	3	Café	Pizza Place	Seafood Restaurant	Park	Convenience Store	Bakery	Stadium	Garden	Flower Shop	Gastropub
29	City of Yarra	Fairfield	3	Pub	Café	Soccer Field	Park	Grocery Store	Zoo Exhibit	Flea Market	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant
32	City of Yarra	Princes Hill	3	Flower Shop	Breakfast Spot	Café	Park	Tram Station	Zoo Exhibit	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fried Chicken Joint

Figure 14 - Cluster 4

Finally, the 5th cluster contains shops or gyms as its most popular venues.

	Municipality	Suburb	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	10th Most Common Venue
24	City of Yarra	Alphington	4	Liquor Store	Gym / Fitness Center	Convenience Store	Farmers Market	Fast Food Restaurant	Park	Train Station	Thai Restaurant	Flower Shop	Fruit & Vegetable Store

Figure 15 - Cluster 5

6. Discussion

Based on the above cluster, we can definitely recommend the 1st and 4th cluster to open a restaurant. However, based on a previous analysis we could see that Melbourne, Fitzroy and South Yarra are the suburbs with the highest density of restaurants, cafes and bars in a 500m radius. These 3 suburbs all belong to the 1st cluster. As the client wants to avoid opening a restaurant directly in the CBD but still wants proximity to it, **the best suburbs to open a French restaurant would be Fitzroy or South Yarra**.

7. Conclusion

In this project, we used different concepts and python libraries to come up with an answer to a real-life like business problem and use data from renown sources like Wikipedia and Foursquare. We've been able to use data science concepts that we learn along the specialization and apply it for this capstone project:

- Data collection (using web-scraping techniques and RESTful API calls)
- Data cleaning and preprocessing using pandas
- Geo-coordinates data Visualization using Folium
- Unsupervised machine learning to perform clustering using scikit-learn

All of these contributed to provide a recommendation to the business problem.