Capstone Project - The Battle of Melbourne Neighbourhoods

1. Problem Background

As a business starting off, finding the perfect location is very important. Whether yours is a restaurant or a manufacturing company, where you locate your business plays a pivotal role in the chances of it succeeding. Location may not feature in the top five to-do list items of a new business, because seemingly more important things may be on the forefront. But being in a bad location can be disastrous for many enterprises; it can cause an otherwise great business and management to be suffocated and fail from the start.

Before you start looking for a business location, you should have a clear picture of what you have and what you want to have in future. Coming up with that picture is a time-consuming process, which is both tedious and exciting – but you need to give it the attention that it deserves. Although many business mistakes can be corrected later, a bad location is sometimes impossible to repair.

This is why businesses must conduct a **business location analysis**. To help businesses, we have carefully selected the **most crucial factors** to consider, when choosing a business location for your business.Here are some other factors that you should consider when choosing the best business location:

1.1 Demographics

When considering demographics, you should think about two important angles. First, you should think about who your customers are and how close they are to your location. This is critical for some service providers and retailers but not so for other businesses. The demographic profile that you have for your target audience will allow you to make this decision.

Secondly, you should consider your community. Is your customer base local, and does a percentage of it support your business or match your customer profile? When choosing communities that are largely dependent on a specific industry, you need to be careful because a slump can be bad for business.

1.2 Parking and Accessibility

Consider the accessibility of the location for every person who will be coming there. If you are on a busy street, is it easy for cars to get in and out of your parking lot? Your facility also needs to be accessible to people with disabilities. Which sort of deliveries are you likely to receive, and will your suppliers be able to access the facility easily?

If you are considering an office building, ask yourself whether you need the keys for periods when the main doors are locked. If the building closes on weekends and you would like to work then, you should look elsewhere. Make sure that there is sufficient parking for employees and customers.

Just as with foot traffic, you should monitor the facility and see how the parking demand fluctuates. Moreover, you should make sure that the parking lot is adequately lit and well maintained.

1.3 Competition

Are competing companies close by? In some instances, this can be advantageous if comparison shopping is popular. You might end up catching the excess from nearby businesses if you are situated near an entertainment area or restaurant. However, if you are selling CJ aviation fuel pumps and there is a competitor nearby that sells the same thing, start looking elsewhere. When consumers are looking for very specific products, they understand that their choices may be limited, so they will probably only visit one location.

1.4 Summary

The closer the products are to your customers, the higher your market value. From planning future expansions, relocating to newer offices, or opening new shops in the right location can mean many things for your business.

2. Problem Description

Melbourne is the coastal capital of the southeastern Australian state of Victoria. At the city's centre is the modern Federation Square development, with plazas, bars, and restaurants by the Yarra River. In the Southbank area, the Melbourne Arts Precinct is the site of Arts Centre Melbourne – a performing arts complex – and the National Gallery of Victoria, with Australian and indigenous art.

Given the this background, A businessman from Italy is thinking of open a new restaurant in Melbourne city in Australia and has approached a data scientist to help determine the best location for his restaurant to kick start his chain of restaurants in the country

2.1 Interest

Almost everything about business exists at a particular time and location. It could be objects like raw materials, products, facilities, people like employees, agents, customers or events like deliveries, purchases, production runs. By understanding how these elements relate to one another through locational analytics, businesses can make more informed decisions that can improve both efficiency and effectiveness. Location analytics helps in understanding and targeting customers and understanding and optimising business processes.

3. Data sources

3.1 Australian Postcodes

The project requires a dataset that has suburb or postal codes with latitude and longitude and the dataset should include all the major suburbs in Melbourne. We are going to use data on https://www.matthewproctor.com/australian_postcodes which has Australian Post Codes + Latitude/Longitude with more than 16K entries. The data files have eight fields as show below:

Data Fields

The files have eight fields in each:

Field	Description	Example	Updated
id	Primary Key from source database	1	Regularly
postcode	The postcode in numerical format - 0000 to 9999	3000	Regularly
locality	The locality of the postcode - typically the city/suburb or postal distribution centre	Melbourne	Regularly
state	The Australian state in which the locality is situated	VIC	Regularly
long	The longitude of the locality - defaults to 0 when not available	144.956776	Regularly
lat	The latitude of the locality - defaults to 0 when not available	-37.817403	Regularly
dc1	The Australia Post distribution Centre servicing this postcode - defaults to blank when not available	MELBOURNE	Infrequently
type1	The type of locality, such as a delivery area, post office or a "Large Volume Recipient" such as a GPO, defaults to blank when not available	LVR	Regularly

3.2 Four Square

Foursquare is a places API that offers real-time access to Foursquare's global database of rich venue data and user content to power your location-based experiences in applications or websites. We are going to use regular endpoints for this project that include basic venue firmographic data, category, and ID.

The dataset obtained for Melbourne will provide the project with the different latitude and longitude of neighbourhoods and will be used to make calls to the Foursquare API for different purposes. We will construct a URL to send a request to the API to search for a specific type of venues, to explore a particular venue, to explore a Foursquare user, to explore a geographical location, and to get trending venues around a location.

3.3 K Square Clustering

You will use the explore function from Foursquare to get the most common venue categories in each neighbourhood , and then use this feature to group the neighbourhood into clusters. You will use the k-means clustering algorithm to complete this task

K Means Clustering Algorithm

- Specify number of clusters K.
- Initialise centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

4. Methodology

This section represents the main component of the report where I discuss and describe exploratory data analysis that was carried, inferential statistical testing that was performed and machine learnings used and why.

4.1 Downloading Dependencies

The first exercise after having determined the data sources is to create a notebook to write machine learning code. Before writing any useful code we would need some python libraries

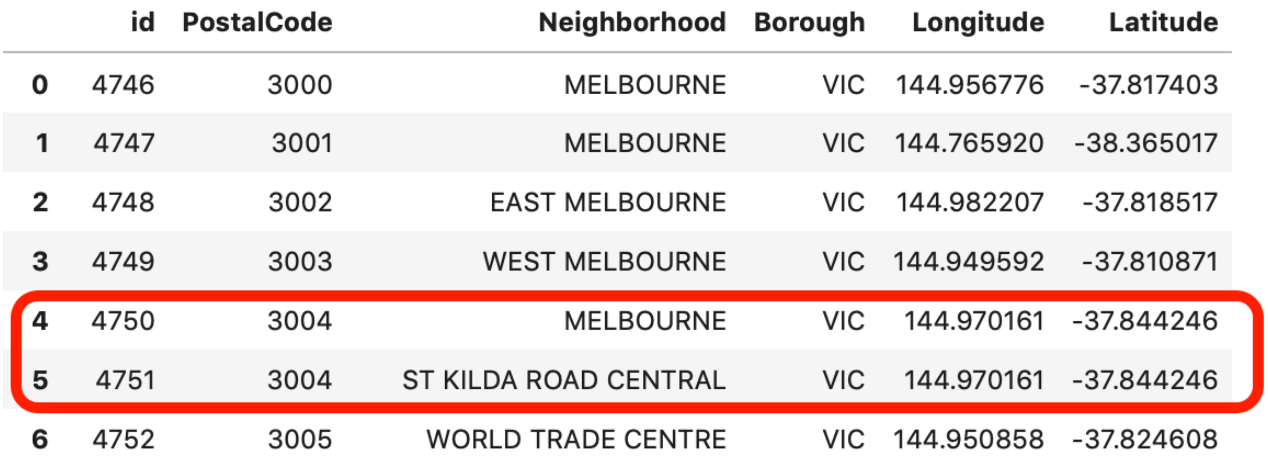
Python library is a collection of functions and methods that allows you to perform many actions without writing your code. Each library in Python contains a huge number of useful modules that you can import to make RESTful API calls to the Foursquare API to retrieve data about venues in different neighbourhoods around the world. Some of the imported libraries will help navigate in creative situations where data is not readily available and scraping web data and parsing HTML code would be required. Python and its pandas library will be used to manipulate data, which will help refine your skills for exploring and analysing data. Finally, we will be require to use the Folium library to great maps of geospatial data and to communicate results and findings.

4.2 Dataset Cleaning

Data cleaning is the process of ensuring that your data is correct, consistent and useable. The following techniques were used to clean the Melbourne Postal codes dataset

4.2.1 Data Cleaning Techniques-Remove Duplicates

More than one neighbourhood can exist in one postal code area. For example, in the Melbourne postal codes dataset, you will notice that 3004 is listed twice and has two neighborhoods: MELBOURNE and ST KILDA ROAD CENTRAL



	id	PostalCode	Neighborhood	Borough	Longitude	Latitude
0	4746	3000	MELBOURNE	VIC	144.956776	-37.817403
1	4747	3001	MELBOURNE	VIC	144.765920	-38.365017
2	4748	3002	EAST MELBOURNE	VIC	144.982207	-37.818517
3	4749	3003	WEST MELBOURNE	VIC	144.949592	-37.810871
4	4750	3004	MELBOURNE	VIC	144.970161	-37.844246
5	4751	3004	ST KILDA ROAD CENTRAL	VIC	144.970161	-37.844246
6	4752	3005	WORLD TRADE CENTRE	VIC	144.950858	-37.824608

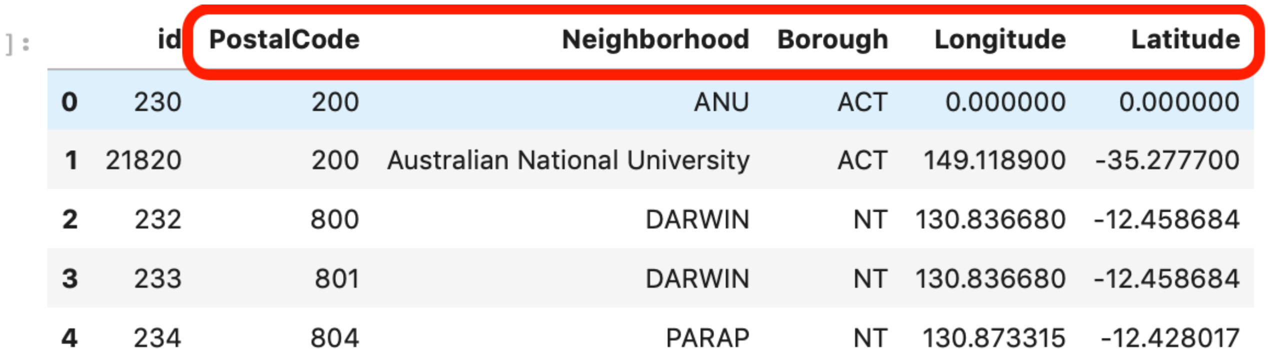
These two rows will be combined into one row with the neighbourhoods separated with a comma as shown in row 11 in the above table.

4.2.2 Data Cleaning Techniques-Select and Treat All Blank Cells

Only process the cells that have an assigned borough. Ignore cells with a borough or coordinates that are **Not assigned or blank or unreasonable values**

4.2.3 Data Cleaning Techniques-Remove unnecessary columns

Only process columns that will provided answers to the research questions and rename some columns to more reflective names

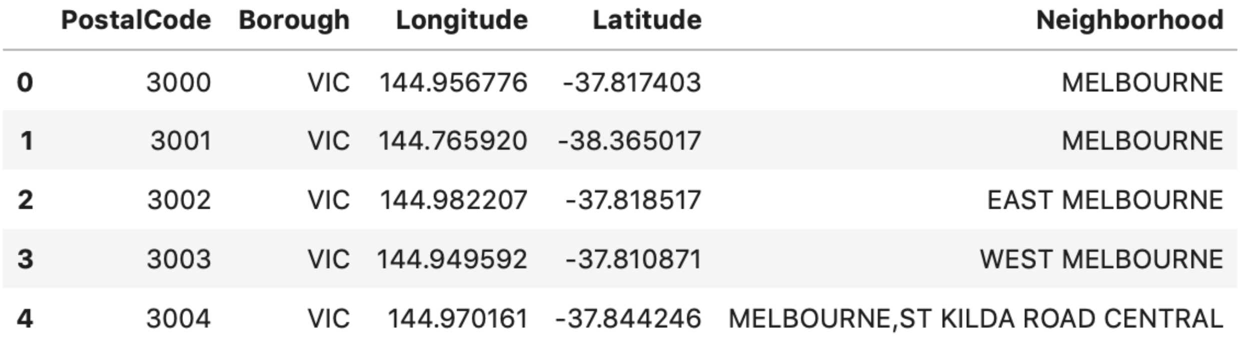


]:

	id	PostalCode	Neighborhood	Borough	Longitude	Latitude
0	230	200	ANU	ACT	0.000000	0.000000
1	21820	200	Australian National University	ACT	149.118900	-35.277700
2	232	800	DARWIN	NT	130.836680	-12.458684
3	233	801	DARWIN	NT	130.836680	-12.458684
4	234	804	PARAP	NT	130.873315	-12.428017

4.2.3 Data Cleaning Techniques-Reduce the number of Boroughs

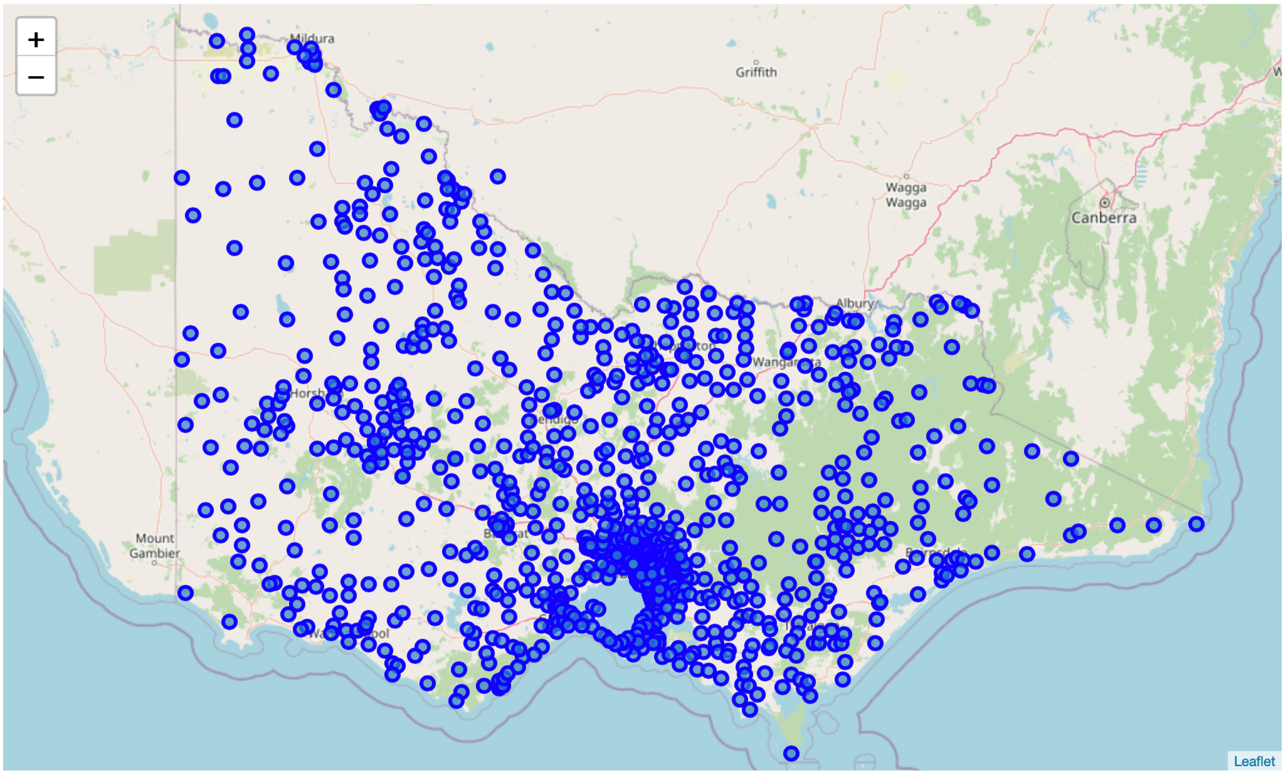
Only process a dataset the focuses on Melbourne instead of the rest of Australia. So the analysis will be based on neighbourhoods in and around Melbourne (Borough == “VIC”). In addition to that, the coordinates that have null values will be dropped as well. Finally the final dataset will look as below with 1 Borough and 1011 Neighbourhoods.



	PostalCode	Borough	Longitude	Latitude	Neighborhood
0	3000	VIC	144.956776	-37.817403	MELBOURNE
1	3001	VIC	144.765920	-38.365017	MELBOURNE
2	3002	VIC	144.982207	-37.818517	EAST MELBOURNE
3	3003	VIC	144.949592	-37.810871	WEST MELBOURNE
4	3004	VIC	144.970161	-37.844246	MELBOURNE,ST KILDA ROAD CENTRAL

4.3 Explore Dataset

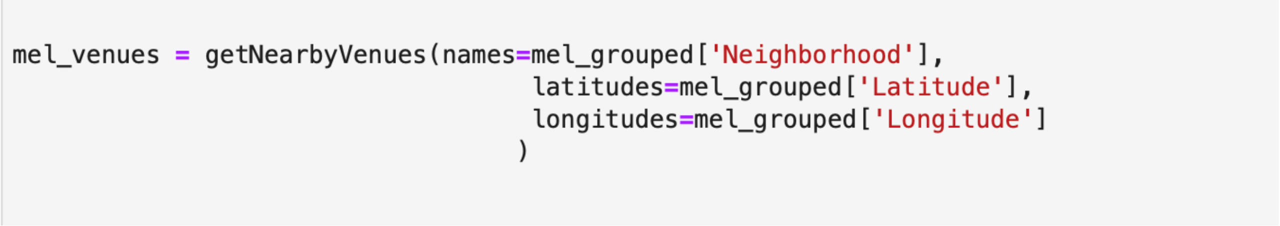
Used **geopy** library to get the latitude and longitude values of Melbourne and created a map of Melbourne with neighbourhoods superimposed on top. For displaying neighbourhoods on the map purposes, Python folium library was used.



4.4 Explore Neighbourhoods in Melbourne

Also, i will use the Foursquare API to explore neighbourhoods in Toronto. i will use the **explore** function to get the most common venue categories in each neighbourhood, and then use this feature to group the neighbourhoods into clusters.

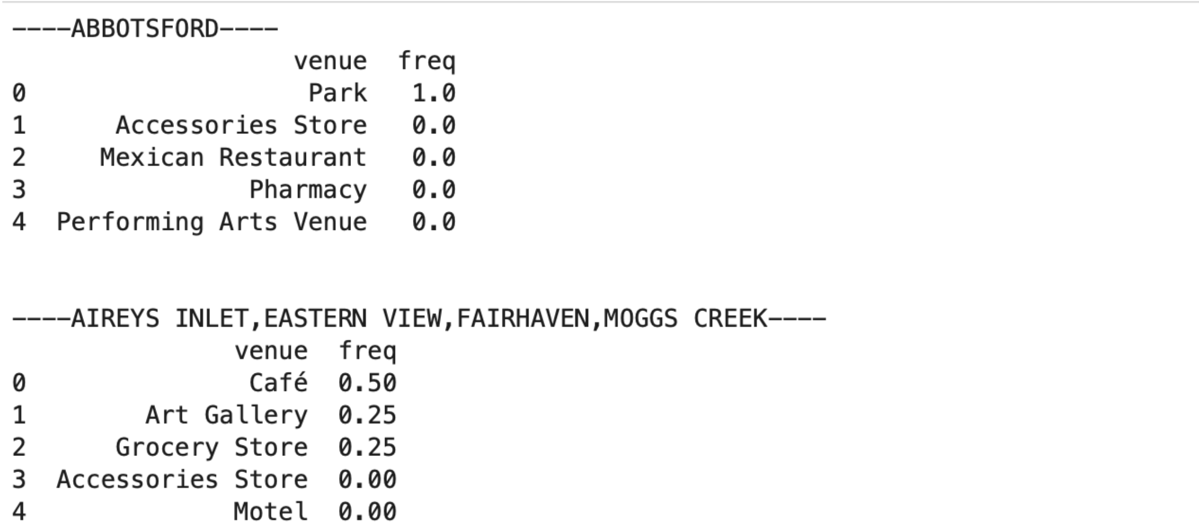
Now, let's get the top 40 venues that are in all Melbourne neighbourhoods are within a radius of 500 meters using a user defined function called getNearbyVenues



```
mel_venues = getNearbyVenues(names=mel_grouped['Neighborhood'],
                                   latitudes=mel_grouped['Latitude'],
                                   longitudes=mel_grouped['Longitude']
                                  )
```

4.5 Analyse Each Neighbourhood

Once we have all the different venues (exactly 1531) I will group them by their Neighbourhood to make it more readable and more understandable, and once it is grouped, I will apply One Hot Encoding to extract the dummies variables of all of those buildings. I will do that to help me after to know what is the most frequency thing in each neighbourhood. Next, let's group rows by neighbourhood and by taking the mean of the frequency of occurrence of each category and printing each neighbourhood along with the top 5 most common venues for example a park is the most venue in ABBOTSFORD



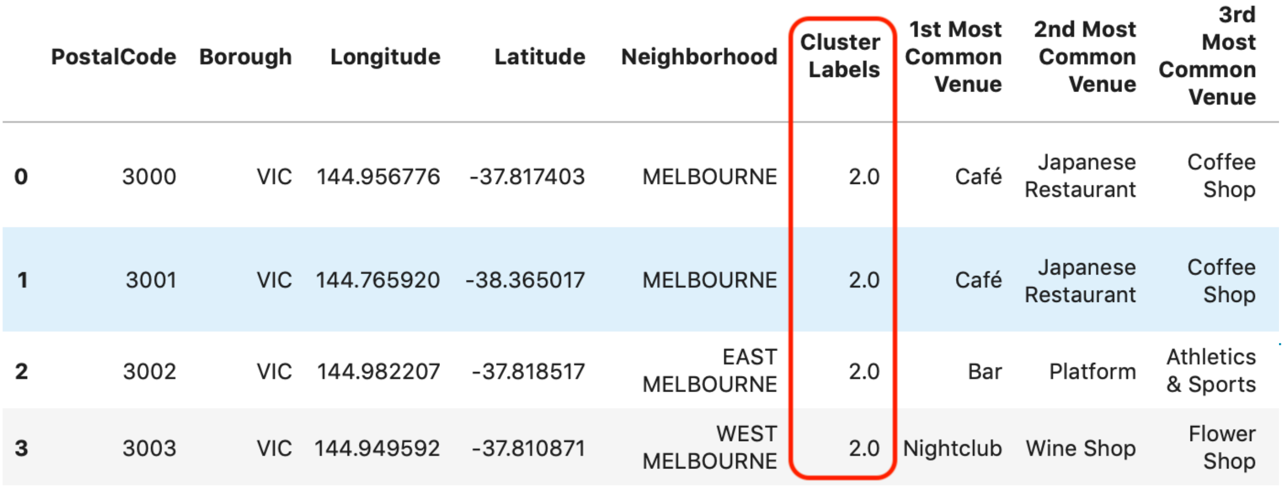
```
----ABBOTSFORD----
                  venue  freq
0                  Park   1.0
1     Accessories Store   0.0
2    Mexican Restaurant   0.0
3              Pharmacy   0.0
4  Performing Arts Venue  0.0


----AIREYS INLET,EASTERN VIEW,FAIRHAVEN,MOGGS CREEK----
              venue  freq
0              Café  0.50
1       Art Gallery  0.25
2     Grocery Store  0.25
3 Accessories Store  0.00
4             Motel  0.00
```

4.6 Clustering Neighbourhoods

There are many models for clustering out there. In this project, we will be presenting the model that is considered one of the simplest models amongst them. Despite its simplicity, the K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabelled data. In this project, I will use k-Means for neighbourhood clustering/segmentation.

Some real-world applications of k-means:

- Customer segmentation
- Understand what the visitors of a website are trying to accomplish
- Pattern recognition
- Machine learning
- Data compression



	PostalCode	Borough	Longitude	Latitude	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	3000	VIC	144.956776	-37.817403	MELBOURNE	2.0	Café	Japanese Restaurant	Coffee Shop
1	3001	VIC	144.765920	-38.365017	MELBOURNE	2.0	Café	Japanese Restaurant	Coffee Shop
2	3002	VIC	144.982207	-37.818517	EAST MELBOURNE	2.0	Bar	Platform	Athletics & Sports
3	3003	VIC	144.949592	-37.810871	WEST MELBOURNE	2.0	Nightclub	Wine Shop	Flower Shop

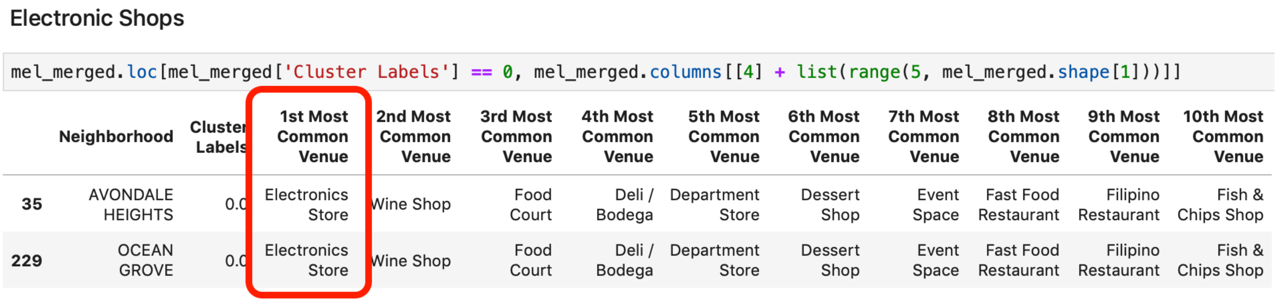
Note: As you can see, the cluster labels are floats, I will need to make them int for be able to represent them with different colours in the map with each city separated by their cluster.

5. Results and Discussion

Now, each cluster can be examined and determine the discriminating neighbourhood or business location categories that distinguish each cluster. Based on the defining categories, you can then assign a name to each cluster according to suitability for a new business

Cluster 0 = Electronics Shops

People are more likely to come here for electronic goods but could be a good site for new restaurant.



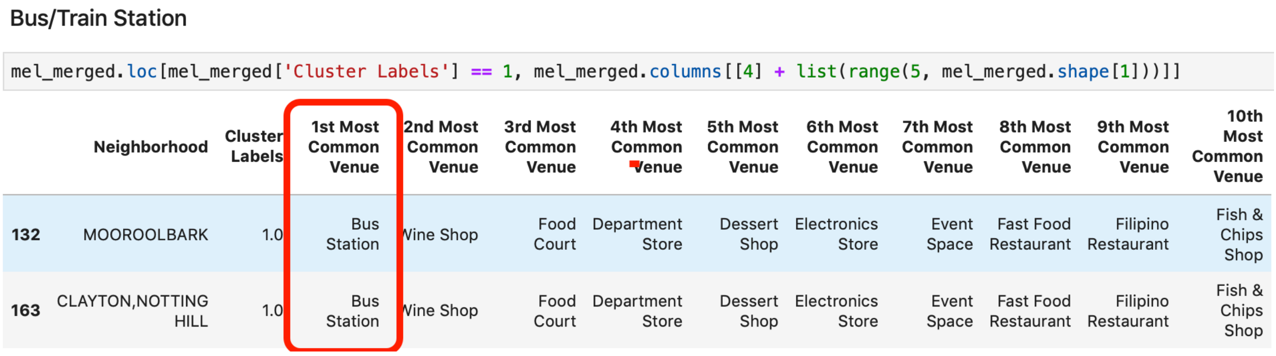
Electronic Shops

```
mel_merged.loc[mel_merged['Cluster Labels'] == 0, mel_merged.columns[[4] + list(range(5, mel_merged.shape[1]))]]
```

	Neighborhood	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
35	AVONDALE HEIGHTS	0.0	Electronics Store	Wine Shop	Food Court	Deli / Bodega	Department Store	Dessert Shop	Event Space	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop
229	OCEAN GROVE	0.0	Electronics Store	Wine Shop	Food Court	Deli / Bodega	Department Store	Dessert Shop	Event Space	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop

Cluster 1 = is Bus or Train stations

Usually people don't have time to sit down and eat because they are on the move and might not a good site for restaurant



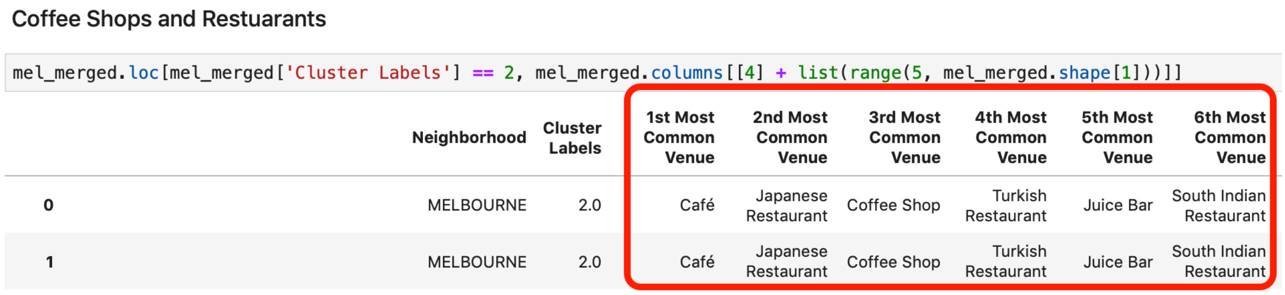
Bus/Train Station

```
mel_merged.loc[mel_merged['Cluster Labels'] == 1, mel_merged.columns[[4] + list(range(5, mel_merged.shape[1]))]]
```

	Neighborhood	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
132	MOOROOLBARK	1.0	Bus Station	Wine Shop	Food Court	Department Store	Dessert Shop	Electronics Store	Event Space	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop
163	CLAYTON,NOTTING HILL	1.0	Bus Station	Wine Shop	Food Court	Department Store	Dessert Shop	Electronics Store	Event Space	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop

Cluster 2 = Coffee Shops and Restaurants

People would come here mainly to have food and its first choice for a restaurant but services must be competitive, more commercial area where there are plenty of restaurants, and places to spend money



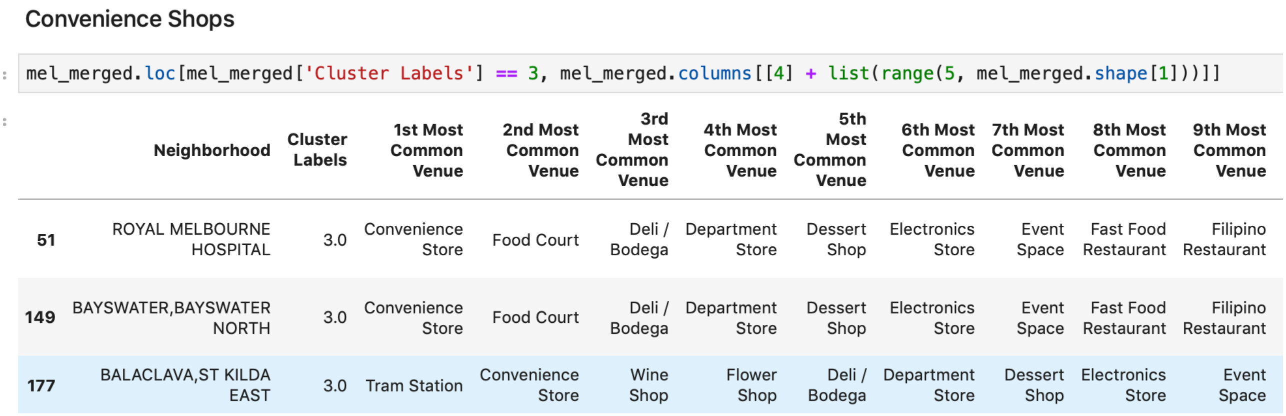
Coffee Shops and Restuarants

```
mel_merged.loc[mel_merged['Cluster Labels'] == 2, mel_merged.columns[[4] + list(range(5, mel_merged.shape[1]))]]
```

	Neighborhood	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
0	MELBOURNE	2.0	Café	Japanese Restaurant	Coffee Shop	Turkish Restaurant	Juice Bar	South Indian Restaurant
1	MELBOURNE	2.0	Café	Japanese Restaurant	Coffee Shop	Turkish Restaurant	Juice Bar	South Indian Restaurant

Cluster 3 = Convenience shops

Most likely close to residential areas or a service station



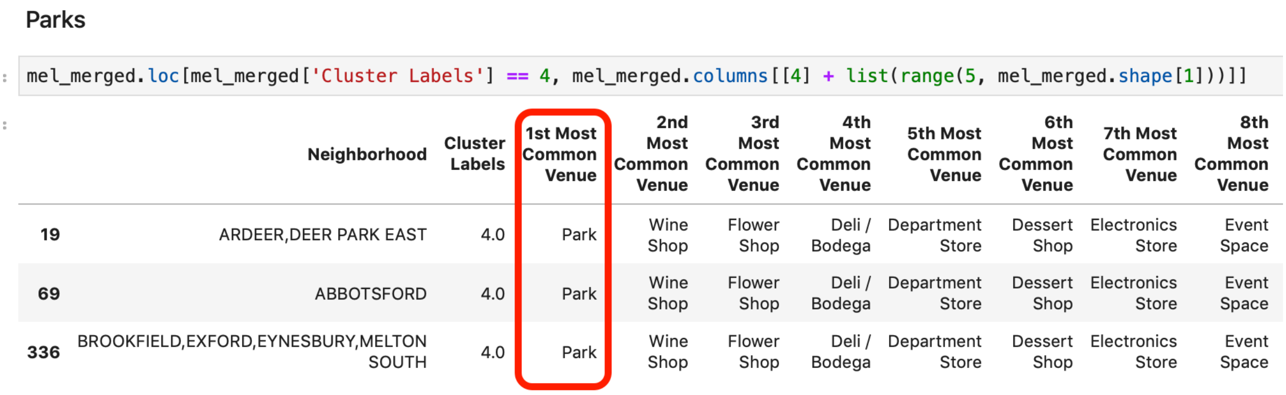
Convenience Shops

```
mel_merged.loc[mel_merged['Cluster Labels'] == 3, mel_merged.columns[[4] + list(range(5, mel_merged.shape[1]))]]
```

	Neighborhood	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
51	ROYAL MELBOURNE HOSPITAL	3.0	Convenience Store	Food Court	Deli / Bodega	Department Store	Dessert Shop	Electronics Store	Event Space	Fast Food Restaurant	Filipino Restaurant
149	BAYSWATER,BAYSWATER NORTH	3.0	Convenience Store	Food Court	Deli / Bodega	Department Store	Dessert Shop	Electronics Store	Event Space	Fast Food Restaurant	Filipino Restaurant
177	BALACLAVA,ST KILDA EAST	3.0	Tram Station	Convenience Store	Wine Shop	Flower Shop	Deli / Bodega	Department Store	Dessert Shop	Electronics Store	Event Space

Cluster 4 = Parks

People usually carry picnic bags this could be another best area for a restaurant as a substitute to picnic bags



Parks

```
mel_merged.loc[mel_merged['Cluster Labels'] == 4, mel_merged.columns[[4] + list(range(5, mel_merged.shape[1]))]]
```

	Neighborhood	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
19	ARDEER,DEER PARK EAST	4.0	Park	Wine Shop	Flower Shop	Deli / Bodega	Department Store	Dessert Shop	Electronics Store	Event Space
69	ABBOTSFORD	4.0	Park	Wine Shop	Flower Shop	Deli / Bodega	Department Store	Dessert Shop	Electronics Store	Event Space
336	BROOKFIELD,EXFORD,EYNESBURY,MELTON SOUTH	4.0	Park	Wine Shop	Flower Shop	Deli / Bodega	Department Store	Dessert Shop	Electronics Store	Event Space

6. Observations and Recommendation

6.1 Foursquare limitation to returned data

Because of the daily limit set by Foursquare as defined here https://developer.foursquare.com/docs/places-api/rate-limits the model kept returning the

Foursquare API will return a 429 error until the time of reset. This meant that the we had to limit the number of returned venues and the radius that might make the model less accurate

6.2 Future improvements

The model on location is not complete and will need more datasets to help give a good indication on the best site as highlighted in the following case studies

F**ood chain Whole Foods**, now owned by Amazon, picks their locations based on many factors, not just population density in a neighbourhood. They found that one of the key drivers that determines whether patrons will shop at their grocery stores is their level of education. As a result, their site selection process looks at locations with a higher per capita level of college degrees.

Costco takes into account population trends to ensure that the neighborhoods in which they locate their stores can sustain sales of their bulk-packaged products.

Walmart uses advertisements to see how far people will go to buy products at their stores. They track usage of mobile advertisements and create a geofence boundary to identify who goes where to buy what. This analysis helps them with their site selection for new stores.

7. Conclusion

The overall objective of site analysis is to select the optimal location in terms of feasibility, economy, and future sustainability so that you are in the best position to achieve your strategic goals. This entails studying and evaluating a large number of factors, always taking individual customer requirements into account.

Now we can try to answer the question: So, if you would like to **establish a new restaurant**, where should you go? The following is the ranking of the clusters where a restaurant can be placed

1. Areas with Parks (Cluster 4 = Parks)
2. Areas with existing restaurants (Cluster 2 = Coffee Shops and Restaurants)