PREDICTING THE BEST LOCATION FOR A NEW RESTAURANT.

Vaibhav Pant

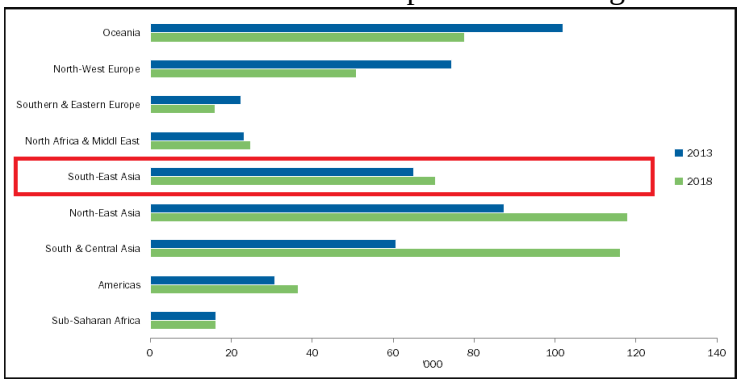
1. Introduction

1.1. Background:

There has been a great increase in the demand for Vietnamese food in Australia.

This is partly due to the increase in the popularity of Vietnamese food but also an increase in the overall Vietnamese population. According to the Australian Bureau of Statistics, there has been a 20% increase in the total migrants coming from South East Asia between the year 2013 and 2018. every growing popularity and demand present an opportunity for new business.

We discuss below the business problem and target audience.



Source: *Australian Bureau of Statistics*

1.2. Business Problem

The objective of this project is to analyse and select the most suitable location for a Vietnamese restaurant in the city of Melbourne. The goal is to come up with the most suitable location using data science and machine learning for the target business audience.

Since there are lots of restaurants in Melbourne, our objective would be to try to find areas/suburbs with no Vietnamese restaurants in the vicinity. We would also prefer locations as close to the CBD as possible.

Keeping the above in mind the following requirements are kept in mind while conducting the analysis and selecting a location for the new restaurant.

- The place should be in close vicinity of CBD. i.e. within 5 km range.
- The place should have no or very few existing Vietnamese restaurants

1.3. Target Audience:

Since the late 1970s when 50,000-odd Vietnamese refugees came to Australia, they have moved up the chain of employment and many have opened restaurants in the capital cities. There are more than 580 Vietnamese restaurants in Melbourne (Source: Zomato). With an inclining population and migration rate fuelled by an increase in the demand of its ethnic cuisines, this is a perfect opportunity for the audience/stakeholders who are interested in opening a Vietnamese restaurant in the city of Melbourne, Australia.

2. Data Acquisition

These data points have been sourced keeping in mind the business problems and factors that will influence our decision.
The information required for the assessment will be sourced from the following areas.

- The number of restaurants and their type and location will be sourced from the Foursquare API. Since there is no exact address of the CBD as it is an area expanding about 6.2 Square kilometres, we have relied on google maps for the address of the CBD which is going to be used as the starting point.

- Coordinates of the Melbourne CBD will be obtained through the ***Nominatim API.***

- Data obtained from ***Corra*** of the suburbs in Australia along with the latitude and longitude. Data cleaned and filtered for the state of Victoria. Original raw data contains 16,080 rows and 7 columns. Click here to access the original data.

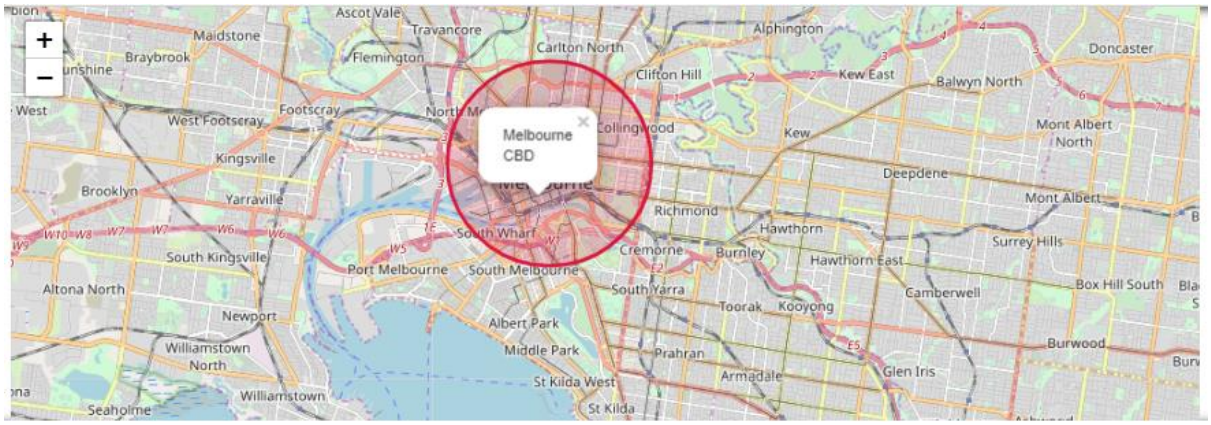
- Due to constraints in the foursquare API and Folium. The data is further filtered by the closest suburbs within a 10 km range from the Melbourne CBD. Click here to access the data.

- Size of the CBD area used for plotting the CBD is obtained from here.

- My Master data is saved in my GitHub Repository. Click here to access the data.

3. Methodology

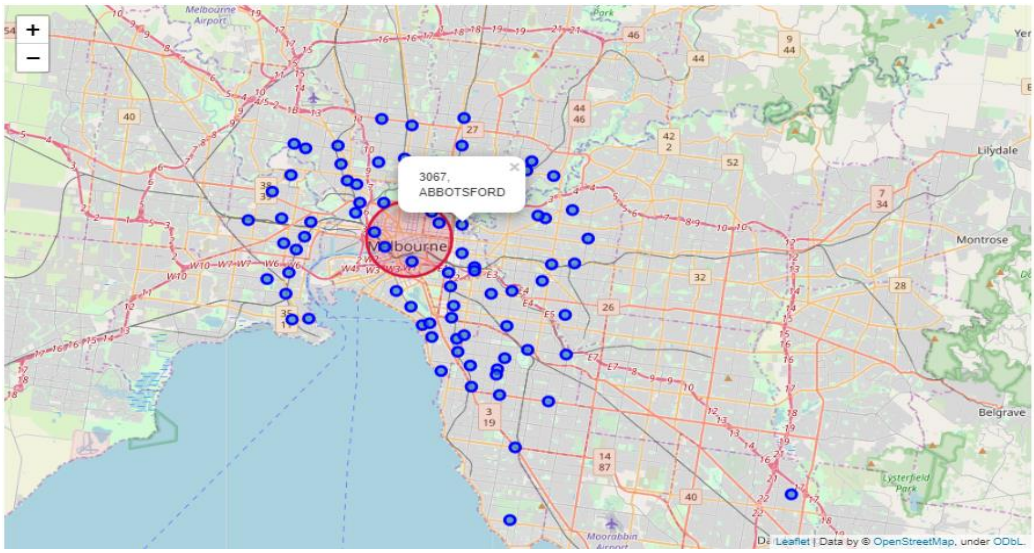
We have used Nominatim to get the latitude and longitude of the CBD. The result is then plotted on a map using python's folium library. The red circle below represents the CBD which is 6.2 Square kilometres or about 2489 meters.



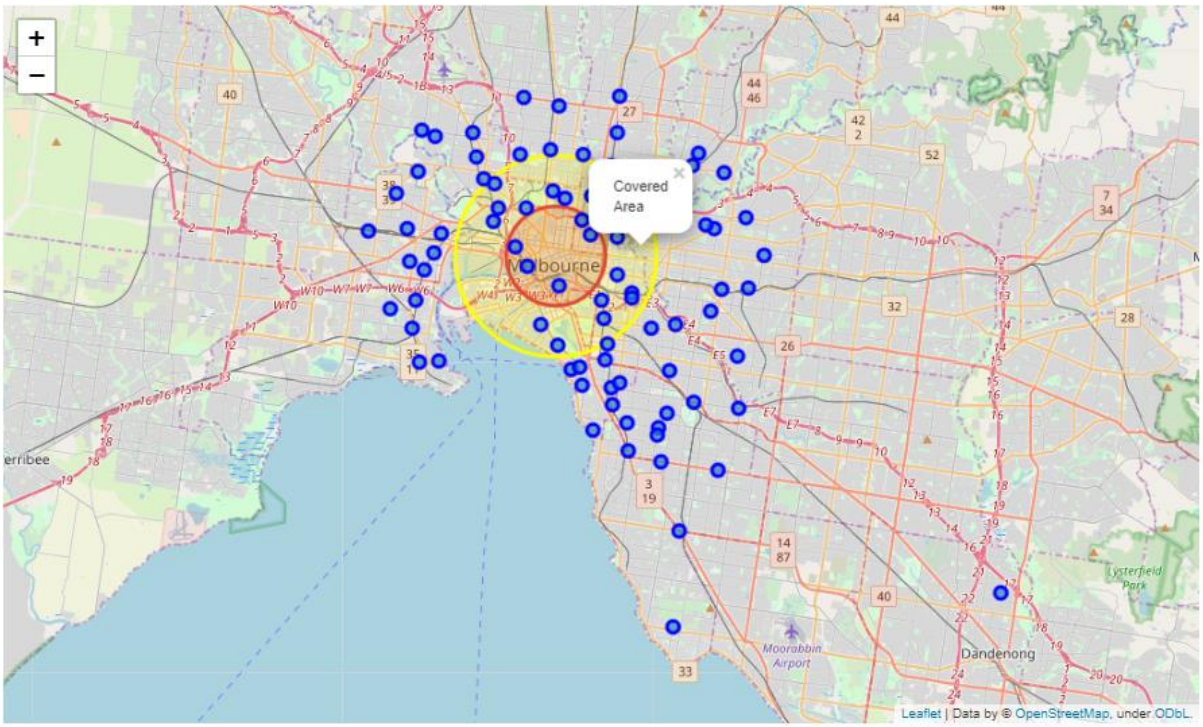
The GitHub repository is used as a database for this study. The main components of the master data are Postcode, Suburb, Lat (Latitude) and Lon (Longitude) of Melbourne.

	Postcode	Suburb	Lat	Lon
0	3003	WEST MELBOURNE	-37.806255	144.941123
1	3006	SOUTHBANK	-37.823258	144.965926
2	3008	DOCKLANDS	-37.814719	144.948039
3	3011	FOOTSCRAY	-37.799770	144.899587
4	3011	SEDDON	-37.808769	144.895486

I used the folium library to visualize the data above. I created a map of Melbourne with Suburb and Postcode superimposed on top. I used latitude and longitude values to get the visual as below:



Now the locations closest to the CBD (Within 5 Kilometres) are taken into consideration. I have used folium again to show the covered area or the area with suburbs that would meet the requirements.



Now I have utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 100 venues and the radius 5000 meters for each suburb from their given latitude and longitude information.

A large radius is used to inculcate all possible venues in and around the suburbs.

Here is a head of the list Venues name, category, latitude and longitude along with the Suburb, Latitude and Longitude from Foursquare API

	Suburb	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	WEST MELBOURNE	-37.806255	144.941123	Twenty & Six Espresso	-37.802773	144.947505	Café
1	WEST MELBOURNE	-37.806255	144.941123	Beatrix	-37.802407	144.944459	Café
2	WEST MELBOURNE	-37.806255	144.941123	Di Bella Roasting Warehouse	-37.804496	144.950684	Café
3	WEST MELBOURNE	-37.806255	144.941123	The French Quarter	-37.802822	144.948299	Café
4	WEST MELBOURNE	-37.806255	144.941123	Auction Rooms	-37.802507	144.949560	Café

The venues are further grouped according to suburbs. Here we can see that most of the Suburbs have reached the 100 limits of venues.

Suburb	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
ABBOTSFORD	100	100	100	100	100	100
ABERFELDIE	100	100	100	100	100	100
ALBERT PARK	100	100	100	100	100	100
ALPHINGTON	100	100	100	100	100	100
ARMADALE NORTH	100	100	100	100	100	100
ASCOT VALE	100	100	100	100	100	100
BALACLAVA	100	100	100	100	100	100
BALWYN	100	100	100	100	100	100
BALWYN NORTH	100	100	100	100	100	100
BRUNSWICK	200	200	200	200	200	200
BRUNSWICK EAST	100	100	100	100	100	100
BRUNSWICK WEST	100	100	100	100	100	100
BURNLEY	100	100	100	100	100	100
CAMBERWELL	100	100	100	100	100	100
CANTERBURY	100	100	100	100	100	100
CARLTON	87	87	87	87	87	87

A list of 213 unique categories is curated from all the returned venues. We have then analysed each suburb by grouping the row by Suburb and taking the mean of the frequency of occurrence of each venue category. By doing so, we have also prepared the date for use in clustering.

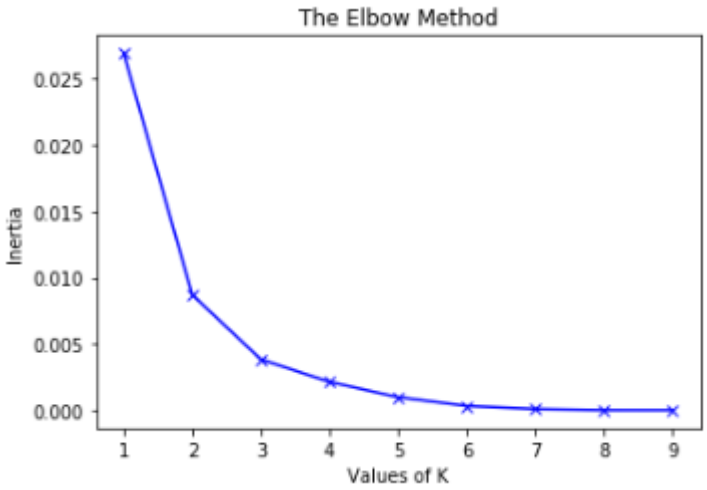
	Suburbs	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Arcade	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	...	Trail	Turkish Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	V
0	ABBOTSFORD	0.000000	0.00	0.00	0.00	0.00	0.010	0.00	0.00	0.020000	...	0.000	0.000	0.04	0.020000	
1	ABERFELDIE	0.000000	0.00	0.01	0.00	0.00	0.000	0.00	0.02	0.010000	...	0.010	0.000	0.00	0.030000	
2	ALBERT PARK	0.000000	0.00	0.00	0.00	0.00	0.000	0.02	0.00	0.030000	...	0.000	0.000	0.00	0.010000	
3	ALPHINGTON	0.000000	0.01	0.00	0.00	0.00	0.000	0.00	0.00	0.000000	...	0.000	0.000	0.04	0.020000	
4	ARMADALE NORTH	0.000000	0.00	0.00	0.00	0.00	0.010	0.01	0.00	0.010000	...	0.000	0.000	0.01	0.030000	
5	ASCOT VALE	0.000000	0.00	0.00	0.00	0.00	0.000	0.00	0.00	0.010000	...	0.010	0.000	0.00	0.040000	
6	BALACLAVA	0.000000	0.00	0.00	0.00	0.00	0.000	0.00	0.00	0.010000	...	0.000	0.000	0.03	0.010000	
7	BALWYN	0.010000	0.00	0.00	0.00	0.00	0.000	0.00	0.00	0.020000	...	0.000	0.000	0.00	0.020000	

Since we are only analysing for “Vietnamese Restaurant” data, we have filtered that venue category for the Suburbs. Below is the head of the list of Suburbs along with the Vietnamese Restaurant data.

	Suburbs	Vietnamese Restaurant
0	ABBOTSFORD	0.02
1	ABERFELDIE	0.03
2	ALBERT PARK	0.01
3	ALPHINGTON	0.02
4	ARMADALE NORTH	0.03

Now since we have a common venue category in Suburbs. Unsupervised learning K-means Algorithm is used to cluster the Suburbs. It is the most common algorithm used for common cluster method of unsupervised learning.

To come up with the optimum degree of K for the K-Means I have used the elbow method before inputting a value of the clusters. Below is the output



To determine the optimal number of clusters, we must select the value of k at the “elbow” i.e. the point after which the inertia start decreasing linearly. Thus, for the given data, we conclude that the optimal number of clusters for the data is three.

We will cluster the suburbs into 3 clusters based on their frequency of occurrence for “Vietnamese Restaurant” The result allows us to identify the Suburbs which have a higher concentration of Vietnamese Restaurant than the others. This will help us further narrow down our search of the best location for the new restaurant.

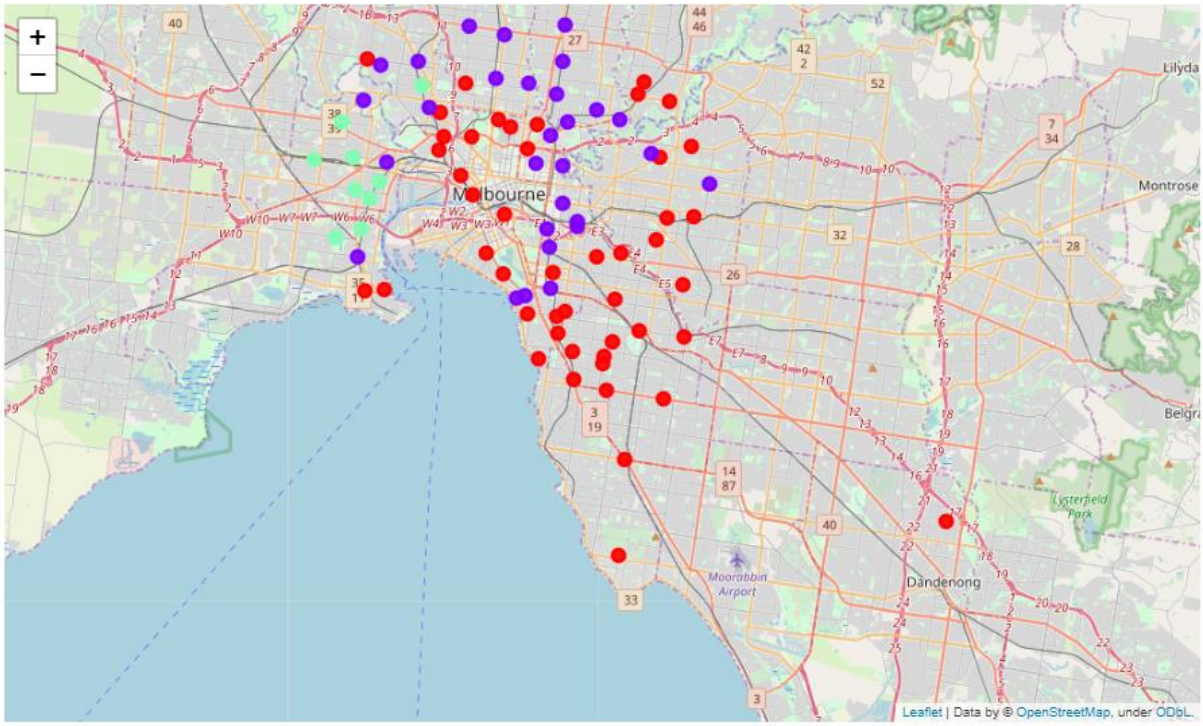
4. Results

The results from the K-means clustering show that we can categories the Suburbs into 3 clusters based on the frequency of occurrence of the “Vietnamese Restaurant”.

- **Cluster 0 –** Suburbs with no to low number of Vietnamese Restaurant.
- **Cluster 1 -** Suburbs with low to moderate number of Vietnamese Restaurant.
- **Cluster 2** – Suburbs with high concentration of Vietnamese Restaurant.

The result of the clustering is visualised in the map below.

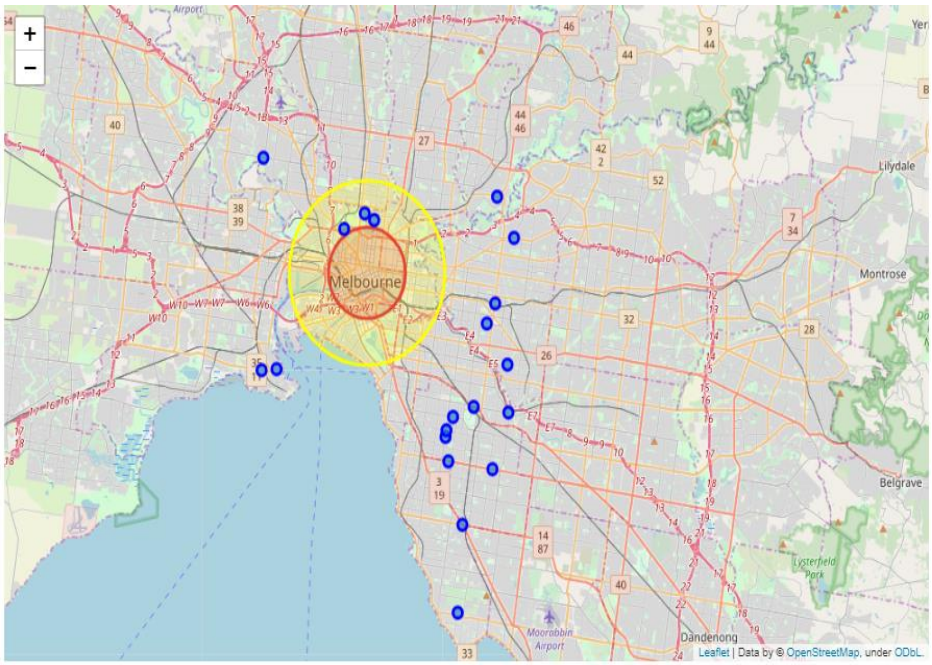
- Red represents cluster 0.
- Violet represents cluster 1.
- Green represents cluster 2.



As our objective was to find a location closest to the city and with the low or few restaurants in the vicinity. We have further refined cluster by filtering it with the frequency of Vietnamese Restaurant set to zero. This would show us the areas where there are no Vietnamese Restaurant. We find all the suburbs with a frequency set to zero in cluster 0.

	Suburb	Vietnamese Restaurant	Cluster Labels	Postcode	Lat	Lon
31	ESSENDON WEST	0.0	0	3040	-37.753802	144.888271
65	PRINCES HILL	0.0	0	3054	-37.780995	144.962792
38	GLEN IRIS	0.0	0	3146	-37.854687	145.067215
39	HAWTHORN	0.0	0	3122	-37.834855	145.052097
40	HAWTHORN	0.0	0	8622	-37.889860	145.021417
62	PORT MELBOURNE	0.0	0	3207	-37.975682	145.030468
44	IVANHOE EAST	0.0	0	3079	-37.772830	145.059401
60	PARKVILLE	0.0	0	3052	-37.788531	144.947731
52	MALVERN EAST	0.0	0	3145	-37.878370	145.067892
58	NORTH MELBOURNE	0.0	0	3051	-37.905996	145.056254
54	MELBOURNE	0.0	0	8001	-38.365017	144.765920
27	EAST MELBOURNE	0.0	0	8002	-38.105449	145.147855
8	BALWYN NORTH	0.0	0	3104	-37.792835	145.071727
83	WILLIAMSTOWN	0.0	0	3016	-37.856902	144.897698
70	SOUTH MELBOURNE	0.0	0	3205	-37.932910	145.033718
14	CAMBERWELL	0.0	0	3124	-37.824818	145.057957
17	CARLTON NORTH	0.0	0	3054	-37.784337	144.969747
18	CAULFIELD	0.0	0	3162	-37.880479	145.026806
19	CAULFIELD EAST	0.0	0	3145	-37.875412	145.041976
84	WILLIAMSTOWN NORTH	0.0	0	3016	-37.857681	144.887041
20	CAULFIELD NORTH	0.0	0	3161	-37.901678	145.023570
21	CAULFIELD SOUTH	0.0	0	3162	-37.886903	145.021979

We then visualise it with folium.



5. Discussion

We can see that 3 suburbs within cluster 0 are inside the covered area. They satisfy both of our required criteria, i.e. They are within 5 km range from the CBD and have no other Vietnamese Restaurant in the vicinity.

This, of course, does not imply that those zones are optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to Melbourne CBD but not crowded with existing Vietnamese restaurants. There may be a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition, but also other factors considered, and all other relevant conditions met.

6. Limitations and Suggestions for Future Research

There are number of limitations in this project. ***Firstly,*** due to processing and API constrains the data has been filtered down from over 3000 locations to about 88 locations. Therefore, this eliminates any possible venue or location that could have been analysed.
Secondly, we have only considered one factor here. i.e. The frequency of occurrence of the restaurant. However, there are other factors such as population, income of resident, proximity to parking and train station, levels of noise / proximity to major roads, real estate cost and social and economic dynamics of the suburbs which might influence the location of the new restaurant.
Finally, there is a lack of the data required for the research and no publicly available data is present. Most data sources are of secondary type.

Future research could make use of a larger database and more complex algorithm and methods such as deep learning. They could use of paid foursquare API or Google API account where there are no constraints to the number of API calls and results to obtain a more accurate result.

7. Conclusion

Purpose of this project was to identify areas in Melbourne close to CBD (within 5km) with low to no Vietnamese restaurants to aid stakeholders in narrowing down the search for optimal location for a new Vietnamese restaurant. By using our primary data and the Foursquare data we have first identified the restaurants and venues according to their frequency of occurrence in each Suburb. Clustering of those locations was then performed to filter the data according to their frequencies and their distance from the CBD.

The filtering of clusters has helped us narrows our search from **88** suburbs down to the **3** suburbs which we believe might be the best location for opening a new Vietnamese restaurant taking all other factors constant and considering the data limitation.

However, final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of suburbs and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to parking and train station), levels of noise / proximity to major roads, real estate availability and prices, social and economic dynamics of every suburb etc.

8. References

1 Foursquare API.

2 Nominatim API.

3 Corra - Suburb data.

4 Wikipedia - Melbourne City Centre.

5 Reiv – Suburbs and their vicinity to CBD data.