Exploration of the café in Melbourne, Australia

Applied Data Science Caption

Jun Yong Chan

Submitted on 14-August-2020

Contents

List of figures

1.0 Introduction

The café culture has had long history in Melbourne since 1890s. It has created a significant impact on the food scene. It started with traditional style café and street coffee stalls between 1900s and 1950s. The coffee related café culture started gaining attention in the 1970s. It moved towards a youth focused in 1980s which expanded the targeted consumers to younger generations while they could enjoy coffee, food and music in one place. It continuously evolved to a venue where people could have enjoyable dining experience in a less formal dining style. In the 1990s, the café culture was encouraged by government which was believed to repopulate of local areas and stimulate a service-based economy. [1]

The strong café culture in Melbourne has connection to the Italian style coffee due to the incoming of Italian immigrants after World War II. This is the reason that each café pays attention to their own specialty coffee and makes some well-known chained coffee brand suffering from business. Up to date, this café culture still attaches to a strong focus on coffee blending with lots of innovative and creative elements. Every café uses their specialty coffee and unique dish to attract customers while some of them have become tourism attractions. [1]

With all selections given, how do you make decision? Imagine you're a tourist, how will you pick the top cafés from the sea of coffee?

1.1 Aim

This project aims to serve several objectives:

- Explore a great variety of cafés in Melbourne
- Investigate the general affordability of dining in cafés
- Understand the rating and reviews of cafés
- Provide an easier way to choose by categorizing the cafés

1.2 Stakeholders

This project can provide an insight for different categories of people such as

- coffee lovers
- foodie
- food bloggers
- tourists

2.0 Data

2.1 Data Source

Two datasets were collected in this project, a restaurants-based dataset and a location-based dataset.

The first dataset was retrieved through Zomato API. Zomato is a famous restaurant aggregator which provides information of restaurants, user reviews, rating, price range and others. It is a useful tool in this project as the rating and price range features were adopted for analysis. Zomato API is free and one can make up to 1000 calls with basic package. Information of cafés were requested by setting 10 km radius around the center of Melbourne CBD. To specify the data to be only café, a category parameter should be added on the API call.

The second data set came from Foursquare API. Foursquare is a high-quality location data platform. By using the information of 100 cafés retrieved from Zomato, the corresponding geographical coordinated were collected. This location dataset will be plotted for visualization.

2.2 Data Processing

Data processing in this project was straightforward. The data scrapped from two sources were processed separately at first and combined for the clustering in the end.

The first dataset was named as 'cafe_df' which was the information restaurant from Zomato API. As Zomato API returned 20 results per call, a for-loop was coded to run for five iterations and summed up to a total amount of 100 cafés. It is important to reset the offset after every iteration in order to avoid receiving repeating data. The incoming data was converted to a flat table using json_normalize function. With all features provided, the cafe_df did filtration and only displayed the necessary columns. In this case, cafe_df consists of 9 columns, including the name, cuisines, address, latitude, longitude, user rating, votes, price range and average cost for two. By sorting the latitude and longitude in order, an outlier was identified with incorrect coordinates. This café was eliminated from the dataset. Hence, cafe_df displayed 99 rows and 9 columns.

Although latitude and longitude were achieved from Zomato API, it was worth mentioning that the coordinates were slightly different from the actual locations. Therefore, the latitude, longitude and address were collected through Foursquare API to provide high accuracy and most up-to-date locations. This dataset was named as 'foursquare_cafe_df'. To execute this, a for-loop was coded to make calls for each café. For users with basic free package, they should note that this could make up to 99 calls. Using the similar approach mentioned above, the dataset only displayed three columns, namely ‘location.lat’, ‘location.lng’ and ‘location.address’. The ‘foursquare_cafe_df’ comprises 99 rows and 3 columns.

In the section of clustering, location coordinates were based on ‘foursquare_cafe_df’ while the rating and price_range features were based on cafe_df. Hence, both datasets were combined, followed by filtering the necessary columns to form the last dataset called ‘combine_df’. This dataset consists of 99 rows and 6 columns. For detailed display of each dataset, please refer to the notebook attached.

3.0 Methodology

In this project, several exploratory data analysis techniques were performed with main focuses on the impact of price range, ratings, magnitude of reviews. In the last part, KMeans clustering was executed for the categorization of cafés.

Exploratory Data Analysis

Firstly, the dining cost was investigated. To find out a more precise average dining cost, the feature ‘average cost for two’ was plotted against ‘price_range’ in a **scatter plot**. The contrast of colors corresponded to the number of cafés that fell on the specific dining cost. Next, the price range was shown in a **bar chart** with the respective number of cafes which gave a better understanding of how much to pay when dining in cafés.

From the perspective of café’s quality, it was reflected through the ratings given by users. Hence, this could be presented by another **bar char** with the rating with its respective number of cafés. It would give the customers an idea of the general quality of cafés in Melbourne.

Although rating could be indicator of the cafés, it should be supported by a considerable number of votes in order to avoid biases. Hence, the number of reviews was plotted in a **histogram** to find out the general number of votes received by each café.

KMeans Clustering

K-means clustering is a method of vector quantization that aims to partition the cafés into cluster by taking dining cost and rating into accounts. In this project, three cluster centroids were generated to categorize the cafés.

4.0 Results and Discussions

The average dining cost in café

From the dataset 'cafe_df", there was a feature called 'price range'. It was a rating scale of 1 to 5 where 5 indicated 'most expensive' and vice versa. On the other hand, another feature called 'average_cost_for_two' provided a general dining cost but it was also indicative only. In another word, a price range of 2 could mean the average dining cost was between $20 and $50. Figure 1 showed the scatter plot of average cost for two versus the price range. According to Figure 1, most cafés fell in the price range of 2 and 3. The darker spots indicated the likelihood of cafés having the specific average cost in certain price range. For example, one should expect to pay $40 and $50 when dining in a café with price range of 2 and 3 respectively.

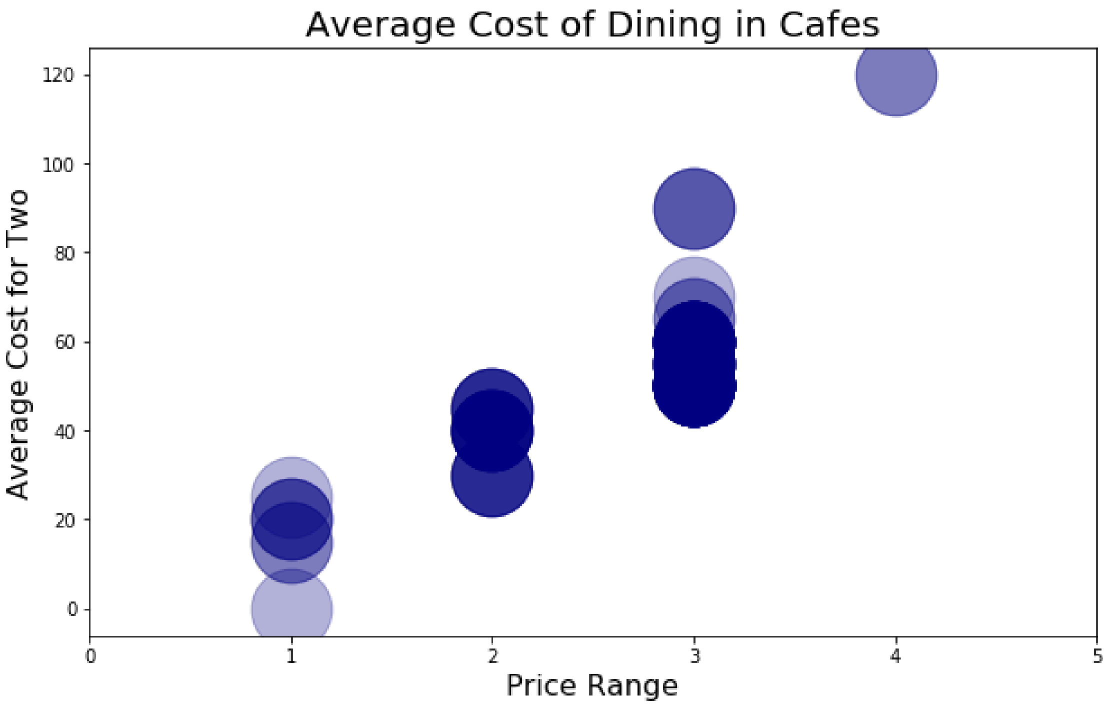

Figure 1: Average cost of dining in café

The price range distribution of cafés

With the idea of how much cost to expect in a café, a price range distribution was plotted in Figure 2. Based on Figure 2, 70 cafés were in the price range of 3. It was a good indication that the pricing of cafés in Melbourne were similar. Apart from that, there were around 27 cafés with lower price range which offered better affordability for consumers.

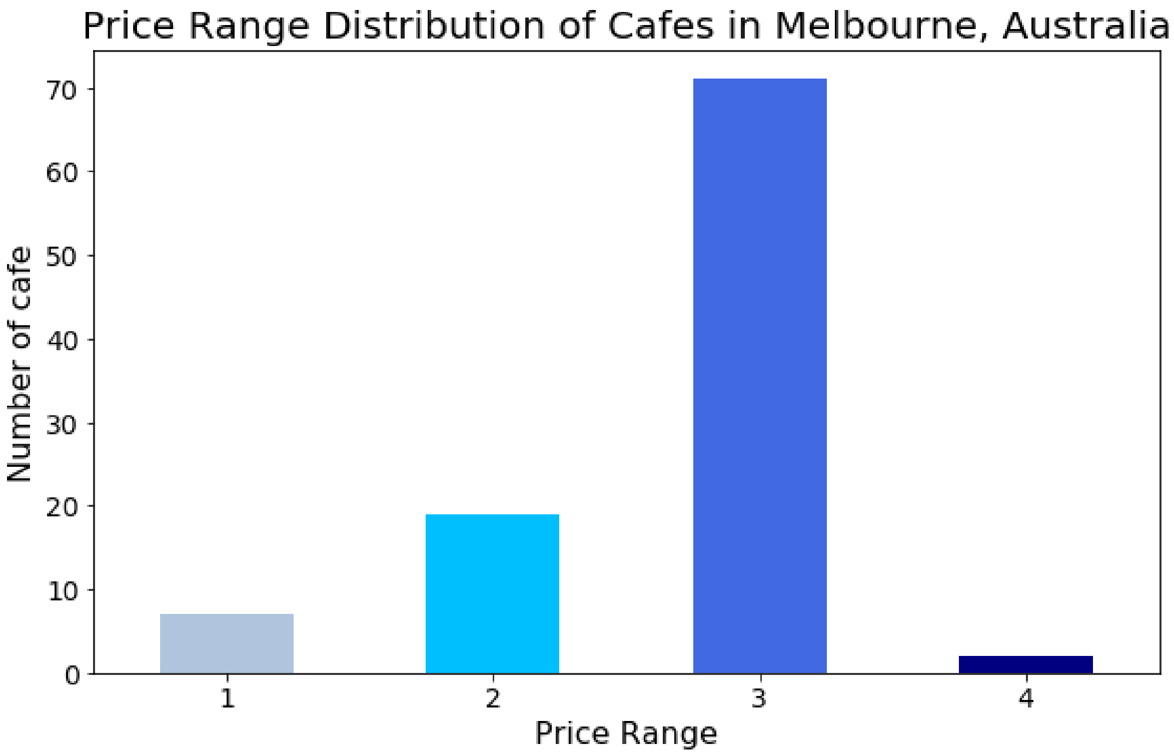

Figure 2: Price Range Distribution of Cafés in Melbourne, Australia

Rating of cafés in Melbourne, Australia

The food quality and service of a café can be conveniently reflected by the rating and reviews given by users. Hence, the ‘rating’ feature from dataset was plotted in Figure 3. On a scale from 1 to 5 where 5 indicating the best, most cafés received rating between 4.4 and 4.6.

By calculating the average of rating, a value of 4.38 was obtained. This value could be used to the line to distinguish between good and poor cafés. This value was used as a standard for the clustering in the later section.

It was not hard to identify that there was a minority lying at the lower end of the chart. This could be due to very few to no review received from users. Nonetheless, the same theory applied on those cafés on the higher end where fewer number of reviews led them to higher ratings. This will be further discussed in the coming section.

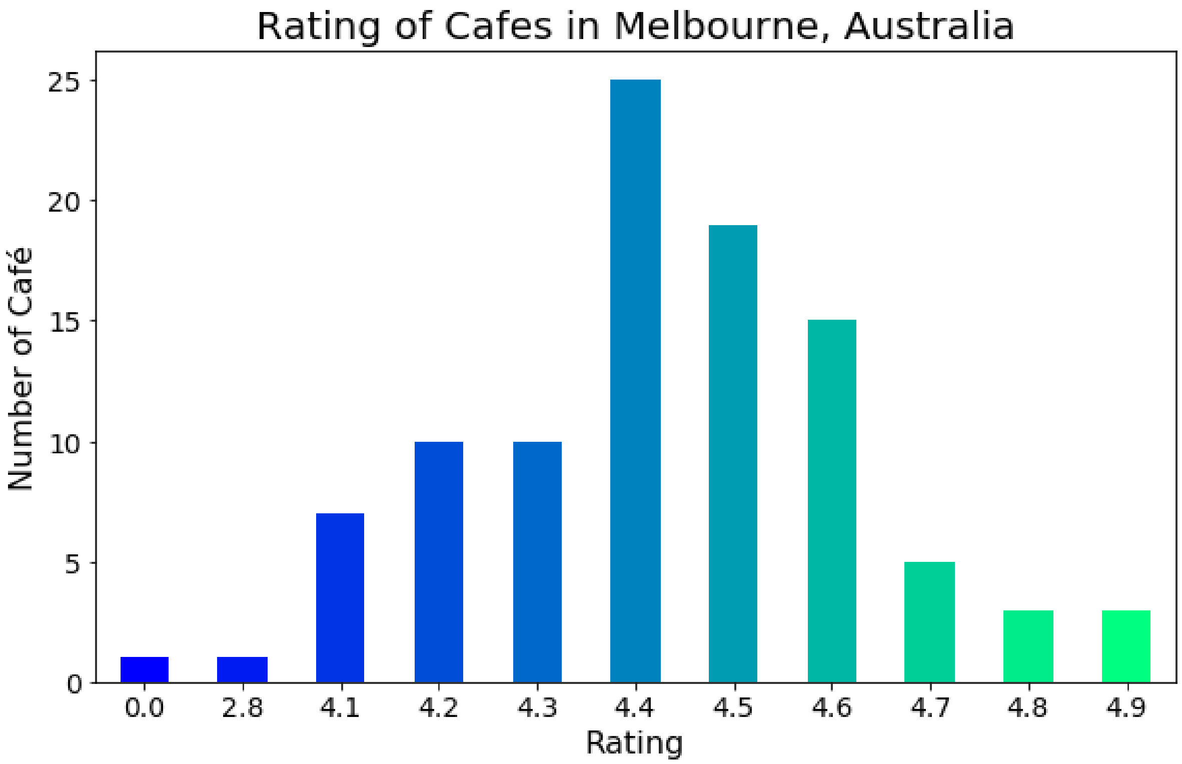

Figure 3: Rating of Cafés in Melbourne, Australia

Magnitude of rating

Although rating could be indicator of the cafés, it could be subject to personal opinions and sometimes it could be far from reality. Therefore, the reputation of a café should not be

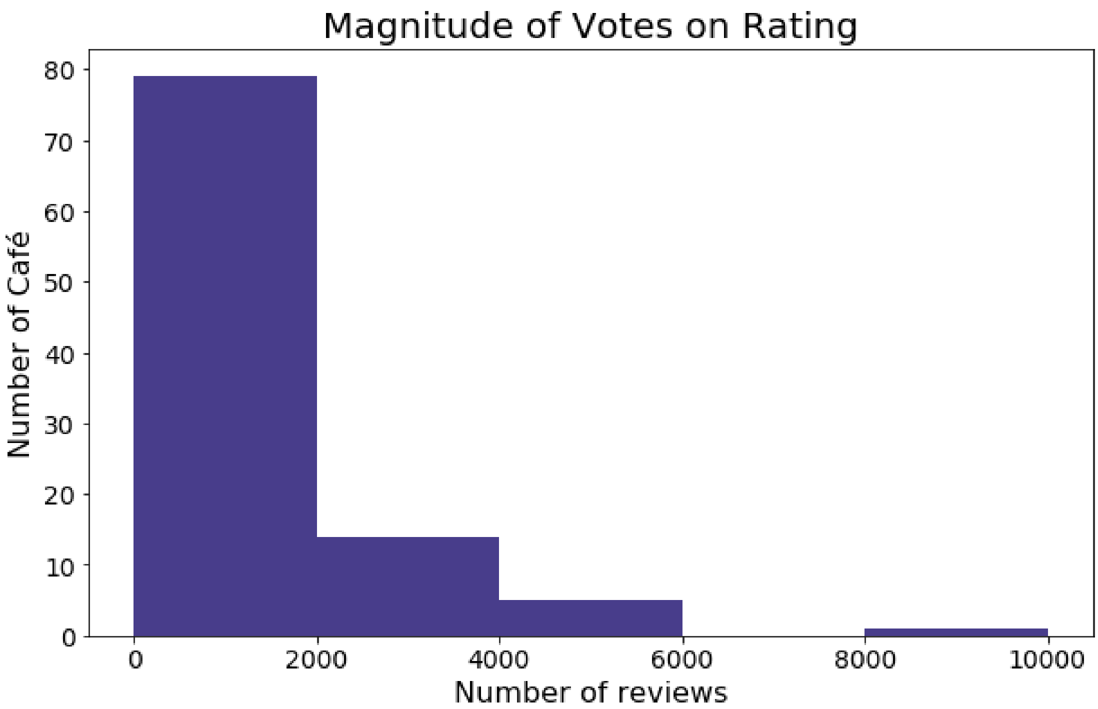

Figure 4: Magnitude of Votes on Rating

purely justified by the rating but also the number of votes. In Figure 4, a histogram was plotted based on the number of votes to represent the reliability of rating. It was obvious that most cafés have received less than 2000 votes. To further investigate it, a breakdown of votes was plotted in Figure 5. It was clearer that about 70 cafés had received more than 500 votes. In another word, around 70% of the cafés received more than 500 reviews by users which generated the corresponding ratings. Hence, it was safe to say that the reliability of rating was high to be used as a reference.

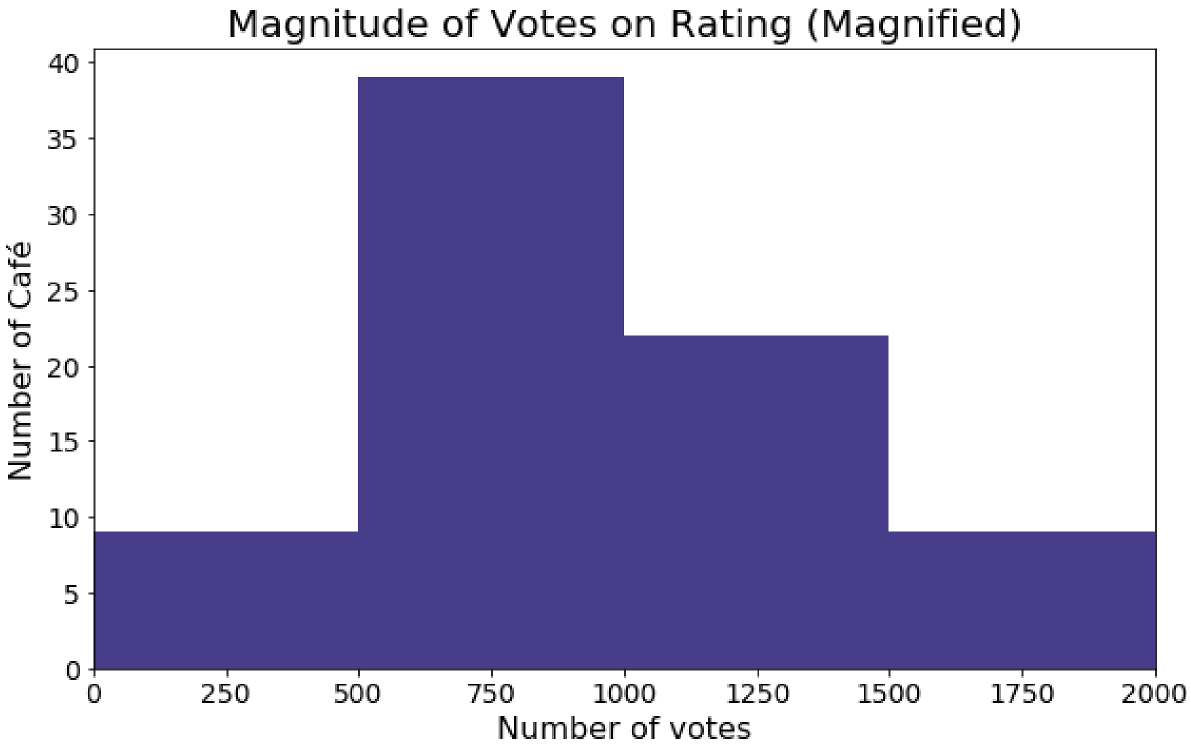

Figure 5: Magnitude of Votes on Rating (Magnified)

KMeans Clustering

After exploring the cafés from the perspective of dining cost and rating, the cafés were grouped to different clusters to give customers a guidance in choosing the right café according to their personal preference. Firstly, all locations were plotted on a map using the geographical coordinated achieved from Foursquare API (Figure 6).

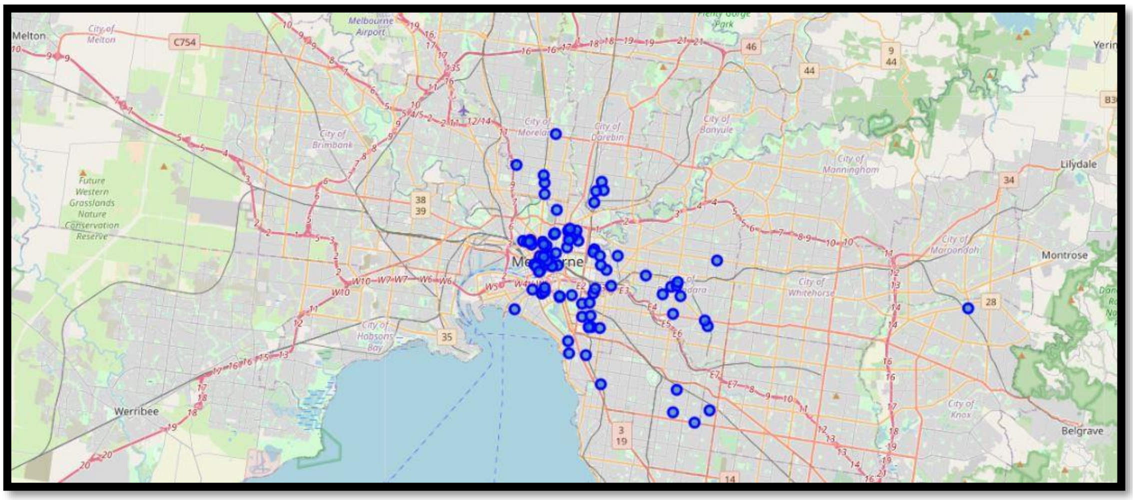

Figure 6: Locations of selected cafes in Melbourne, Australia

Three centroids were generated for clustering. By using K-means clustering, the result was shown in Figure 7. The resulting clusters are colored in red (Cluster 1), yellow (Cluster 2) and navy (Cluster 3). The characteristics of each cluster was summarized below:

	Average dining cost for two persons	Average rating	Number of Cafés
Cluster 1	$52.68	4.39	82
Cluster 2	$102.00	4.38	5
Cluster 3	$22.08	4.30	12

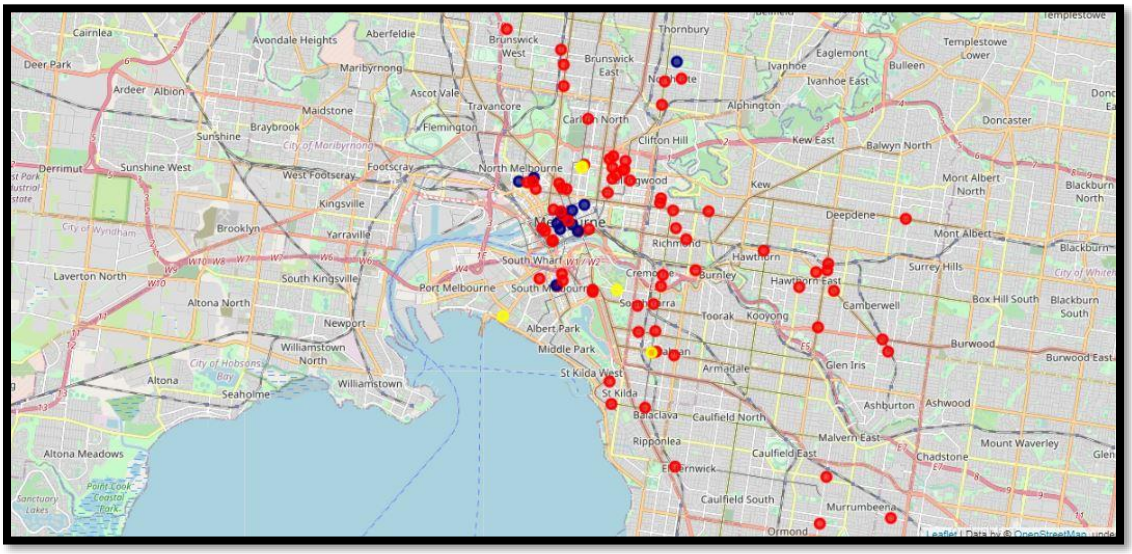

Figure 7: Clustering of cafes. Cluster 1(red), Cluster 2(yellow) and Cluster 3(navy)

5.0 Conclusion and recommendation

In conclusion, this project explored the cafés in Melbourne, Australia in terms of dining cost and rating. Firstly, the average dining cost spread in the range and $20 and $120. 70% of the selected cafés had a price range of 3 which indicated the average cost for two persons to be $50. On the other hand, the selected cafés received an average rating of 4.38/5.0 which indicated the good quality of food and coffee as well as the service. According to data, 70% of cafés received more than 500 votes in generating the average rating which simply reflected that the reliability of rating is high enough to be used as a selecting criterion. Lastly, the cafés were clustered into three categories for the stakeholders which provided them a guideline to select the right café based on their preferences. To briefly summarize, Cluster 1 included cafés with average dining cost and good rating, Cluster 2 showed high dining cost and average rating and Cluster 3 presented low dining cost and low rating.

To further improve the analysis, several recommendations were suggested:

- Increase the size of data set
- Take more parameters into consideration such as opening hours, vegan options, takeaway availability etc.
- Expand the range of cuisines or be more specific on certain part of cafés (only coffee etc.)

References

1. Hanscombe, R. (2009) Cafe Culture in Museums Victoria Collections *https://collections.museumsvictoria.com.au/articles/2933*