Exploring Opportunities for Upcoming Café Chains in Central Singapore

Marcus Chua

April 25, 2020

1. Introduction

1.1 Background

Singapore, a city defined by its world-leading education and trade efficiency, is without a doubt among the most advanced cities in Asia. Its rapid modernisation has transformed the capital into a metropolitan area, making it a perfect entry location for businesses. In a region where cafés have become a part of every city dwellers' lifestyle, there is a huge demand for more of such businesses to open. As Singapore is a city split into 28 districts, I will aim to analyse the problem by looking into existing café venues within the individual districts. More specifically, I will be focusing my research on the 9 central districts in Singapore.

1.2 Target Audience

The following clients/groups of people would take interest in my research:

1. Businesses/Entrepreneurs - Any café business considering their opportunities in Singapore. This could either be a small boutique cafés looking to enter the market in Singapore, or an international chain looking to add new outlets and expand its market share within the city.

2. Consumers - The multitude of coffee-lovers dwelling in the urban areas of Singapore, or the streams of international tourists who view Singapore as an exciting holiday destination.

3. Data Scientists – Any individual with interest in exploratory analysis and would like some to glean some project ideas/techniques from my work.

2. Data

2.1 Data Sources

2.1.1 Lists of Districts in Singapore

I will be using a page from the Urban Redevelopment Authority (URA) of Singapore to obtain basic information of the 28 districts.

2.1.2 Location Data (Coordinates)

I will then use the Geopy client to obtain location data of the districts. The Nominatim.geocode function will enable me to assess the exact coordinates of the districts.

2.1.3 Location Data (Venues)

Following this, I will primarily be using the Foursquare API to extract any remaining data to be used. The Foursquare platform provides users with the option to request different types of data. For this project, I will using two types of data from the platform: (1) List of venues within a radius of the district centre, and (2) List of food venue categories.

2.2 Data Mining

2.1.1 Lists of Districts in Singapore

The URA page has split the district information into their respective regions in Singapore, so I must write a loop to concatenate the 5 tables on the webpage for the full list.

2.1.2 Location Data (Coordinates)

I extracted the ‘District’ column as string values into a list, so that I could use the geocoder function on them. To add specificity to the location search, I appended

‘Singapore’ to the district names. Afterwards, I used the geocoder function in a loop so that I could obtain coordinates for each separate district. These coordinates were then added to the dataframe of districts.

2.1.3 Location Data (Venues)

To use the Foursquare location data to sieve out venues, I created a Foursquare Developer account to gain access credentials. Following which, I crafted two separate request URLs. The first one was done using the explore tool to return the list of venues near district centres. The second one was done using the categories tool to return a list of food venue types within the food category.

2.3 Data Cleaning & Selection

2.1.1 Lists of Districts in Singapore

As the list was extremely concise, there were no columns which I dropped. However, I renamed the columns for ease of reference. Additionally, some of the District Names were not given formally. As such, the Geopy client will not be able to fetch location data on them. To solve this, I identified the informally named Districts and replaced them with their formal names using the .replace function.

Subsequently, I identified the 9 districts which constituted Central Singapore and scoped the dataframe down to these 9 rows.

2.1.2 Location Data (Coordinates)

The coordinates were assigned to the districts extracted from the URA tables. Figure 1 below shows the cleaned dataframe of the 9 districts, along with their respective latitudes and longitudes. For further clarity, I plotted these districts on a map of Singapore using the Folium library.

District Number	District	Localities	Latitude	Longitude
1	Marina Bay	Boat Quay, Chinatown, Havelock Road, Marina Sq...	1.275682	103.854617
2	Downtown	Anson Road, Chinatown, Neil Road, Raffles Plac...	1.279395	103.852993
6	City Hall	City Hall, High Street, North Bridge Road	1.293199	103.852582
7	Beach Road	Beach Road, Bencoolen Road, Bugis, Rochor	1.303052	103.862267
8	Little India	Little India, Farrer Park, Serangoon Road	1.306648	103.849269
9	Orchard	Cairnhill, Killiney, Leonie Hill, Orchard, Oxley	1.305272	103.832876
10	Tanglin	Balmoral, Bukit Timah, Grange Road, Holland, O...	1.306044	103.815280
11	Newton	Chancery, Bukit Timah, Dunearn Road, Newton	1.313183	103.838040
12	Toa Payoh	Balestier, Moulmein, Novena, Toa Payoh	1.335391	103.849741

Figure 1: Dataframe of Districts in Central Singapore and their Location

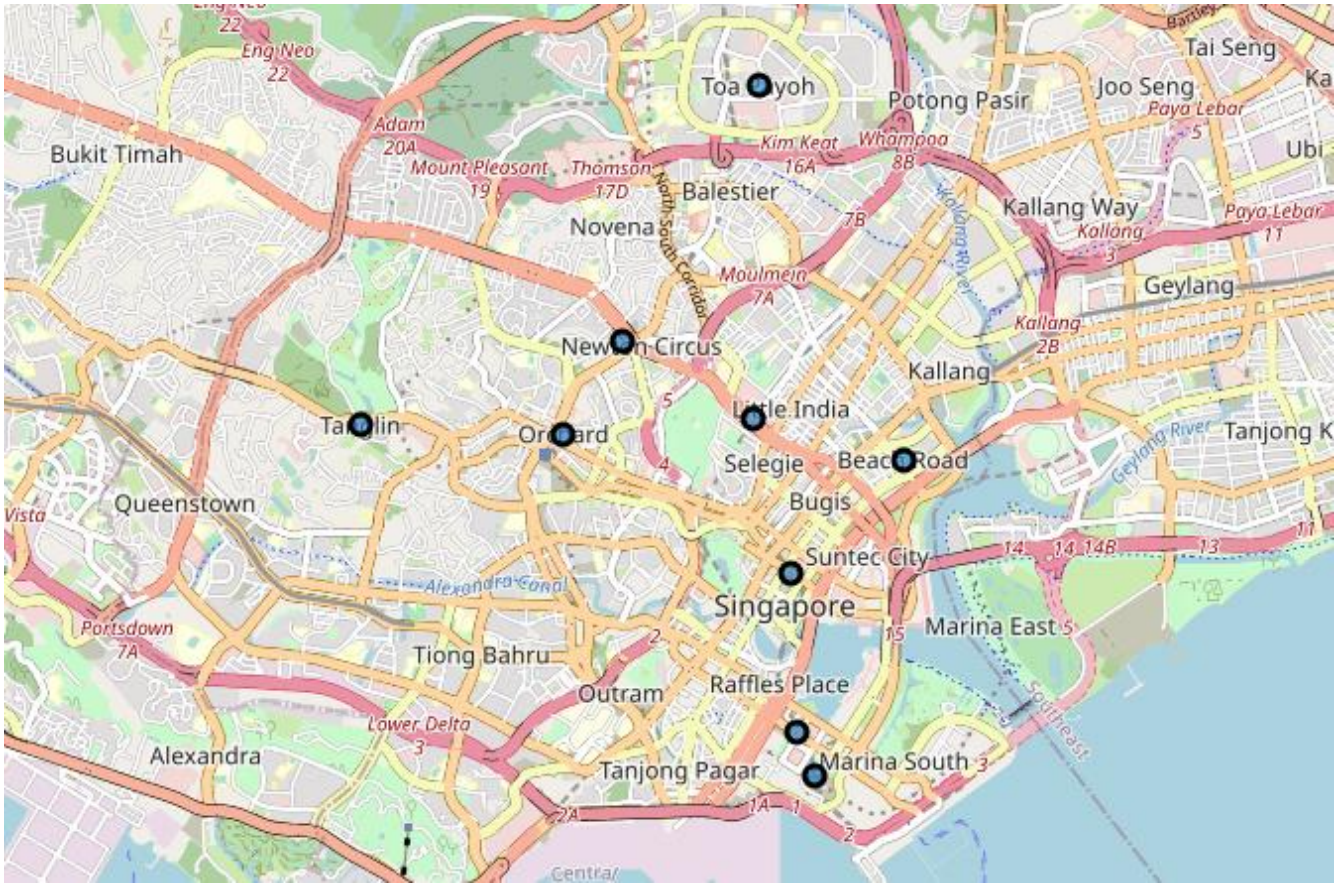

Figure 2: Map of the Districts in Central Singapore

2.1.3 Location Data (Venues)

For the side query, in order to identify the food venues from the list of venues I will obtain, I needed to get all the possible food venue types from the subcategory of 'food' in Foursquare. I wrote a query and then cleaned the responses to obtain a list of these venues.

As for the main query, I defined a function to carry it out in a loop. This function would clean through the responses by:

1. Looping through all the districts in my list, and searching the vicinity of each district for venues
2. Retrieving the essential information from each venue response, storing them in a dataframe

After which, I applied the function to the 9 districts.

With this cleaned dataset, I had to perform another task to further narrow down my venues. Using the side query from the previous part, I had obtained a list of food venues. The dataframe of venues were trimmed using this list. I found that 203 of the 479 venues were food venues, which is a huge proportion.

3. Methodology

3.1 One-Hot Encoding

To prepare the data for further analysis, I used one-hot encoding to assign binary values. For each row of a venue, a binary value was assigned to each column (which represented a food venue category). This dataframe format allowed me to easily calculate the frequency of each venue appearing in the individual districts.

3.2 Study of Food Venue Frequencies within each District

After I had completed the frequency calculations, I displayed the top food venues (by their frequency) for each district. This provided me with the information for each of the 9 districts in an easily readable format.

I then defined a function to collect and convert the information in my display into a dataframe. This would allow me to move onto the next phase of clustering.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Beach Road	Italian Restaurant	Coffee Shop	Café	Bakery	Restaurant
1	City Hall	Coffee Shop	Café	Asian Restaurant	Bakery	French Restaurant
2	Downtown	Coffee Shop	Café	Sandwich Place	Food Court	Mexican Restaurant
3	Little India	Indian Restaurant	Vegetarian / Vegan Restaurant	Restaurant	Bakery	Coffee Shop
4	Marina Bay	Spanish Restaurant	Seafood Restaurant	Gastropub	Wings Joint	Café
5	Newton	Seafood Restaurant	Italian Restaurant	American Restaurant	Café	Asian Restaurant
6	Orchard	Bakery	Café	Bubble Tea Shop	Coffee Shop	Asian Restaurant
7	Tanglin	French Restaurant	Modern European Restaurant	Indian Restaurant	Bakery	Seafood Restaurant
8	Toa Payoh	Food Court	Snack Place	Coffee Shop	Asian Restaurant	Steakhouse

Figure 3: Dataframe Ranking the Most Common Food Venues by District

3.3 Clustering

Now that each district had been clearly profiled, I used KMeans Clustering to group the districts into distinct clusters. I set the number of clusters as 5, and dropped the District column so that only the venue columns remained. I then ran the KMeans function to fit the data.

In order to view the results of the clustering, I included a column for the Cluster Label to the resulting data. I also re-inserted the district names for ease of reference, by joining the resulting data with the column of district names from an earlier dataframe.

I then used color codes to visualise the districts in their respective clusters.

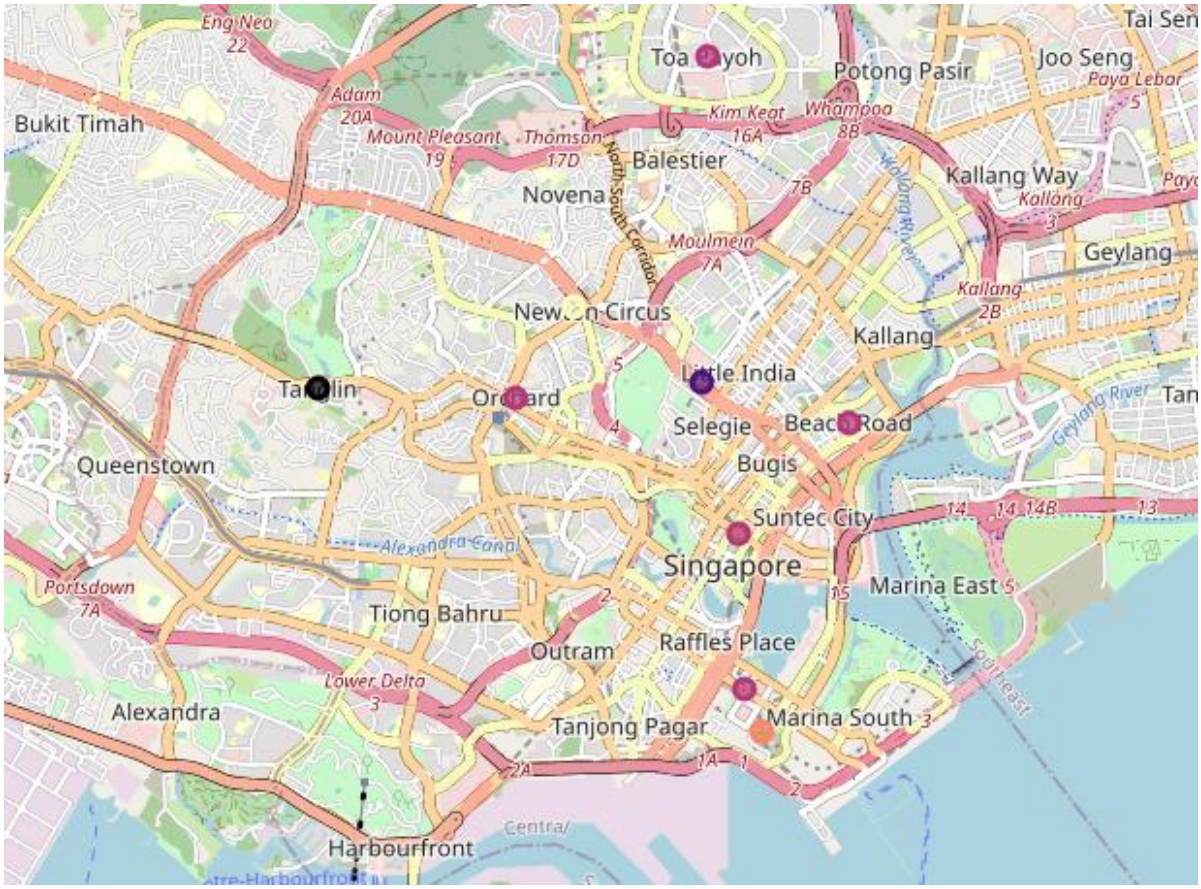

Figure 4: Districts in Central Singapore,

Color-coded based on Food Venues Clusters

4. Results

4.1 Popularity of Food Venues in Singapore

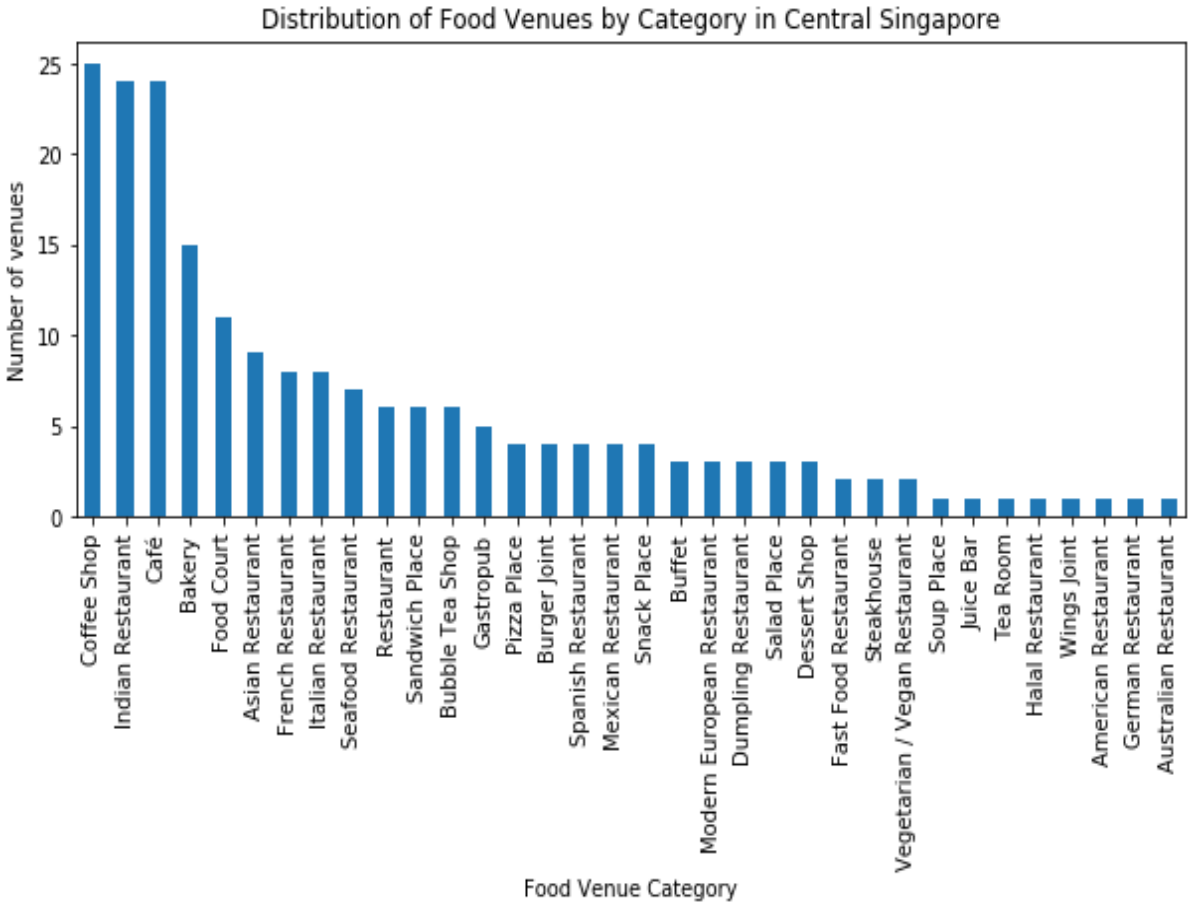

Figure 5: Food Venues in Central Singapore by Category

There are 34 food venue categories in Central Singapore, with the majority of them already being coffee shops/cafes. There is high demand for cafés and coffee shops in Central Singapore, as seen from Figure 5. Two of the top three venues are cafés and coffee shops, showing that these two popular food venues have a huge market share.

With this finding in mind, we will continue our analysis to see if the prevalence of coffee shops/cafes is well-balanced across the districts. If there are districts

which do not fall under this trend, we can identify them as potential districts for new outlets.

4.2 Clustering

4.2.1 Cluster 1

	District	Cluster Number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	Newton	0	Seafood Restaurant	Italian Restaurant	American Restaurant	Food Court	Café

Only one district falls under Cluster 1, where restaurants are commonplace.

4.2.2 Cluster 2

	District	Cluster Number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	Tanglin	1	French Restaurant	Seafood Restaurant	Indian Restaurant	Modern European Restaurant	Gastropub

Extremely similar to Cluster 1, Cluster 2 seems ideal for exotic cuisine and might possibly be an area active in the evenings (which is unideal for cafés and coffee shops).

4.2.3 Cluster 3

	District	Cluster Number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
7	Little India	2	Indian Restaurant	Vegetarian / Vegan Restaurant	Café	Restaurant	Coffee Shop

Cluster 3 comprises of an extremely niche location, where the overwhelming majority of food venues fall under Indian Vegetarian cuisine (due to religious and racial reasons).

4.2.4 Cluster 4

	District	Cluster Number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Downtown	3	Coffee Shop	Café	Sandwich Place	Food Court	Mexican Restaurant
5	City Hall	3	Café	Coffee Shop	Asian Restaurant	French Restaurant	Bakery
6	Beach Road	3	Italian Restaurant	Coffee Shop	Café	Bakery	Restaurant
8	Orchard	3	Bakery	Café	Coffee Shop	Bubble Tea Shop	Asian Restaurant
11	Toa Payoh	3	Snack Place	Food Court	Coffee Shop	Asian Restaurant	Steakhouse

5 of the 9 districts were clustered together. These 5 districts were notably littered with cafés and coffee shops, with these 2 venues ranking in the top 5 for all districts (except Toa Payoh). Furthermore, the trend in this cluster was a low number of restaurants, with a larger proportion of grab-and-go food venues like bakeries.

4.2.5 Cluster 5

	District	Cluster Number	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Marina Bay	4	Gastropub	Spanish Restaurant	Seafood Restaurant	Café	French Restaurant

Cluster 5 is likely a suitable cluster for fine dining and expensive cuisine, with the common venues being Restaurants. Its profile bears huge resemblance to those of Clusters 1 and 2.

5. Discussion

5.1 Potential Locations for a Café/Coffee Shop

Upon analysis of the clusters, it can be easily identified that Cluster 4 is abundant with cafés and coffee shops. As the accessibility of a coffee place is already extremely high in these districts, it might be fairly difficult for a new business to break into the market here.

On the other hand, the other clusters serve as great options for potential cafés/coffee shops. Cluster 1 and 2, being in close proximity to the districts in Cluster 4, will provide alternative spots close by for regular coffee hunters. Cluster 3 has the most promising location, but it might be tougher to create a demand for coffee in an area considered to be an enclave. Cluster 5 will be a potential area for cafés/coffee shops too, but there are a number of existing cafés which could pose as some competition.

In all, Clusters 1, 2 and 5 provide the greatest potential. Cluster 3 and 4 can definitely be considered also, but other factors will have to be considered too. For example, the demographics/rental price of a location in these districts.

5.2 Limitations

Some possible limitations of my research will include the currency of the Singapore Districts data from URA. While the profiling of districts by the District Numbers are somewhat accurate, the classification method using the District Numbers has been slowly phased out by the government in an attempt to urbanise the entire Singapore as a whole. As such, some of the district's profiles might continue to evolve and make the analysis less relevant than it can be right now as compared to another 3-5 years in the future.

Another limitation linked to the data used could be the accuracy of the Foursquare information. While Foursquare is constantly updated, new cafés and coffee shops are constantly popping up every other week. The data might take a while to be collected and uploaded.

The use of KMeans clustering has its own limitations too. Even if the optimal number of clusters for the 9 districts is indeed 5, the density of the districts in my clusters differ greatly. If further districts were to be added to the current model, the number of clusters will likely need to be adjusted to ensure that the results of clustering remain useful. Extensions to this research could make use of other techniques like DBSCAN.

6. Conclusion

This project has provided deep insight on the food venues in Central Singapore, and can also be used to analyse the potential of setting up various types of food shops in the area. Furthermore, the project has demonstrated how skills in data mining, data cleaning, data visualisation and machine learning can be combined and harnessed to solve a Data Science problem.

On a personal note, the findings of this project reflect the profiles of the districts rather well. Being a local in Singapore, the demographics and amenities of each district did serve as a good reference point for us to learn more about the districts. I hope that the findings will be of interest to my different target audience.