**COURSERA: IBM DATA SCIENCE PROFESSIONAL CERTIFICATE**

**Applied data science capstone project: Opening a new shopping mall in Singapore**

By: Eugene Loh Wen Hong

Dec 2020

**Introduction**

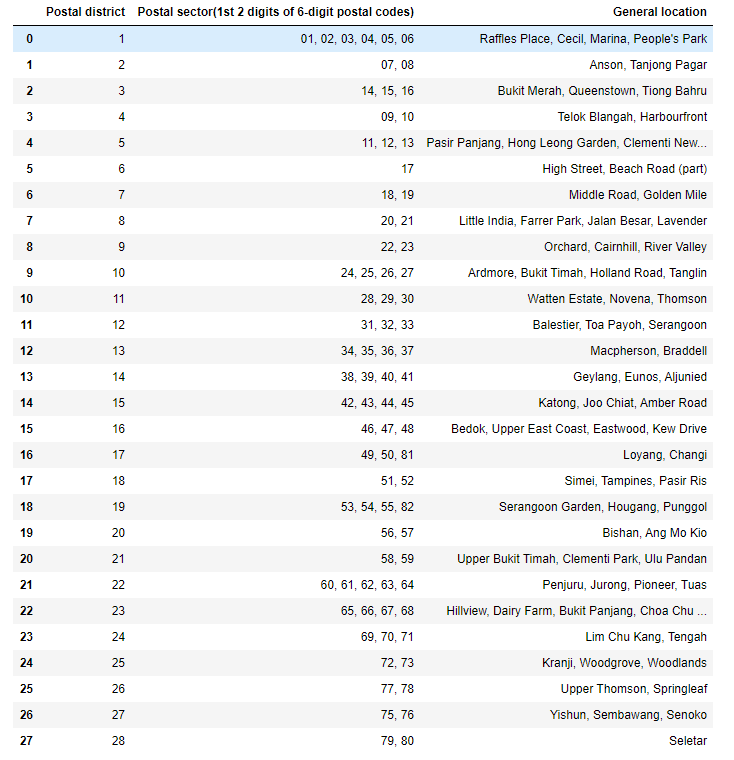
1. **Background.** Singapore is a country where eating has become part of our culture. There are endless selections of food ranging from street snacks, steamboat, buffets and fine dining from all over the world. Being a global financial trading hub and a country that uses the tourism industry as a vital economic pillar, it is no wonder people from all around the world want to come to Singapore either for work or leisure. With that said, this makes Singapore an ideal country to open up a food business.
2. **Problem.** The business problem of this project is to analyze and provide insights into discovering ideal locations for opening a food business in Singapore. By using location data, we are going to use data science to address the problem and answer the following business questions: If someone wants to open a food business in Singapore, which locations will be recommended?
3. **Target audience.** The target audiences in this capstone project are Singaporeans or foreigners who intend to open a food business in Singapore. This report can also be relevant for angel investors looking to invest in restaurants and want to know where are the possible prime locations. Lastly, this is also relevant for food business owners who are looking to expand their business and open more outlets. Therefore, this project is not only targetted at a particular food business owner but can be a good analyze for the F&B industry players too.

**Data**

1. **Data sources.** To start off, we will need to know what kind of data we need and where we can gather the data.
   1. The postal districts, sector and the general location of Singapore are required and can be obtained [here](https://www.iproperty.com.sg/news/know-which-district-you-are-in-based-on-postal-code/) and [here](https://en.wikipedia.org/wiki/Postal_codes_in_Singapore). Singapore’s postal codes are in 6 digits format and the first 2 digits represent the postal districts.
   2. The latitude and longitude of the neighbourhoods for visualization and plotting of data which can be obtained from the Python Geocode package.
   3. The Foursquare Application Programming Interface(API) is essential to explore the venues in the different districts of Singapore. The API can allow us to search for venues, get venue categories, get recommended venues and can further be explored to help work on this capstone project.

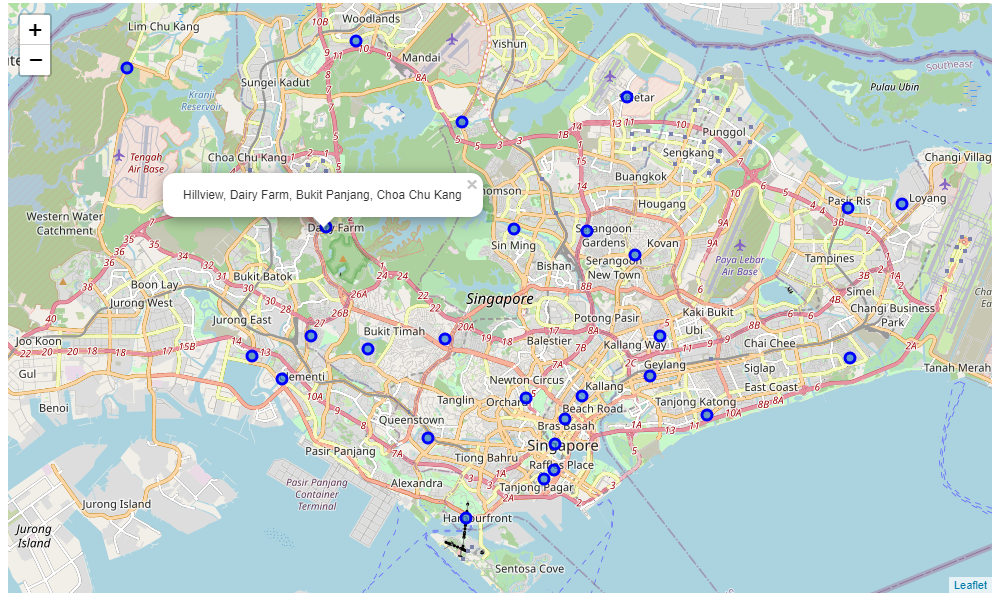
**Methodology**

1. **Data preparation.** The data source of the districts in Singapore is from a wiki page and web scraping is used as a method to acquire information from websites to our computer or databases and then we can change the data into a table format where further processing can be done. I use Python requests packages and the BeautifulSoup library for this project.



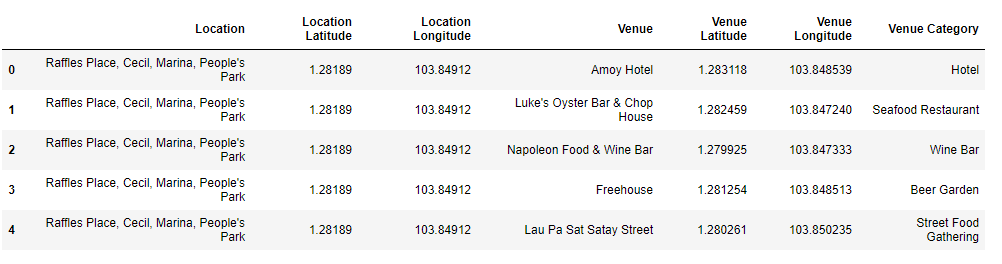
There are a total of 28 districts in Singapore. I found out that the postal districts were first implemented in 1950 and then replaced by the postal sectors in 1955. The 2 digits postal sectors are actually the same as the postal districts. Although the old districts were not used by the Singapore Post anymore, they are still widely used for buying and selling of properties. Also, due to the land size of the country, one postal district can refer to a few locations or neighbourhoods in the vicinity. For example, postal district 1 refers to Raffles Place, Cecil, Marina and People’s Park.

The coordinates were then retrieved through the Geocoder package and then visualize it on a folium map.

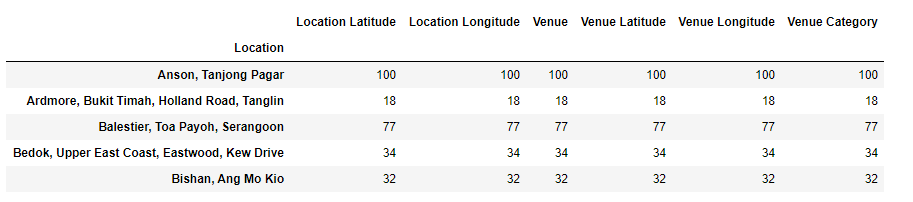


In the initial analysis, we can see from the folium map that many of the districts were located around the central and north-south region.

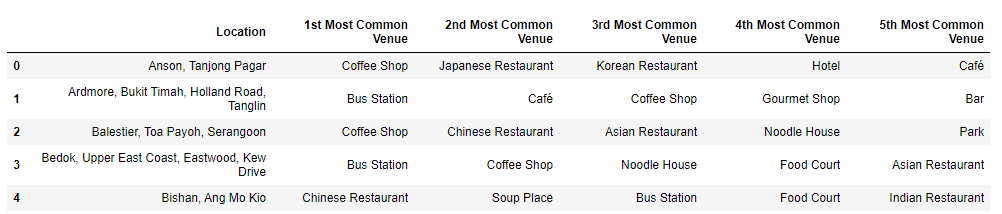
In the next step of the analysis, we will explore the Foursquare API and understand more about the districts in Singapore. Using the get request, the 100 venues of a district in a 1km radius from Foursquare were obtained. The get request will then return a list of recommended venues near the current location. Below shows some of the venues that are retrieved for a district.



We can also view the total numbers for each district capped at 100. From this, we can see how many venues are in each location.



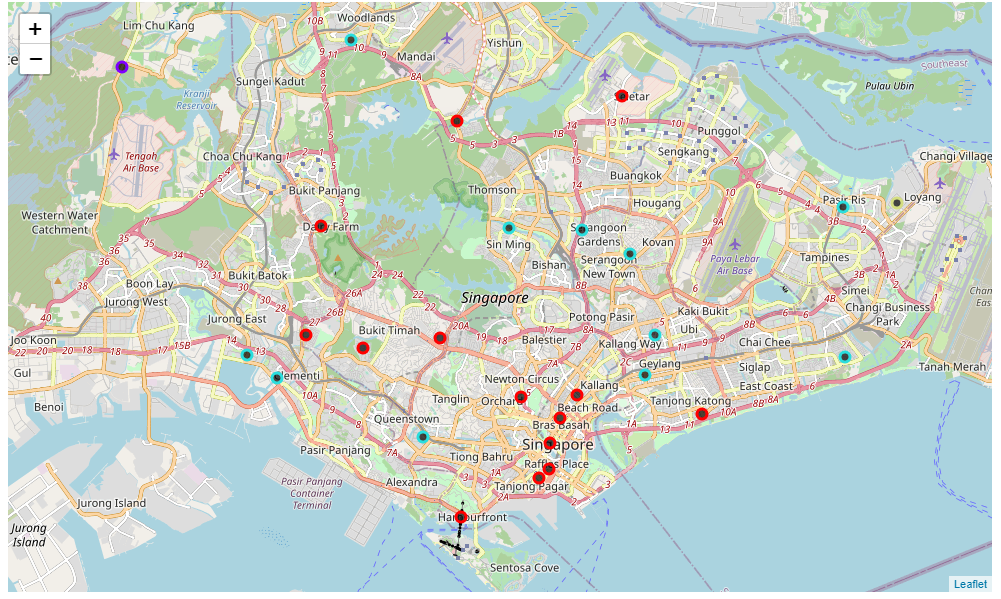
We can group the venues by the district and show the top 5 most common venues so that we can better understand the data set.



1. **Data Modelling.** After getting some understanding of the data we have, we will begin to perform machine learning on the data. As there is no labelled data available, an unsupervised learning model will be used for this project and the technique is by K-means clustering. K-means clustering will form clusters and find groups within the data. K-means clustering groups data points that are close together and at the same time, maximise the inter clusters’ distance.

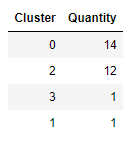
After which, we can start to derive certain assumptions from those clusters. In order to perform clustering, numerical data is required and one-hot encoding is used to change our data variables to numerical ones.

Once done, we can perform clustering and visualize the clusters in the map below.



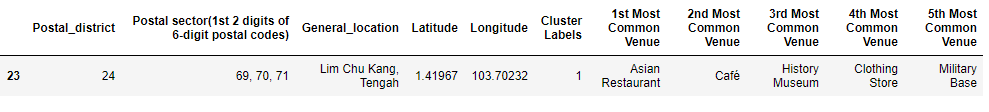
**Results**

1. **Understanding the clusters.** Cluster 0 and 2 are the ones that we are more interested in as that is where the 2 clusters have the densest results of 14 and 12 respectively. For cluster 3 and 1, there is only 1 data point each.

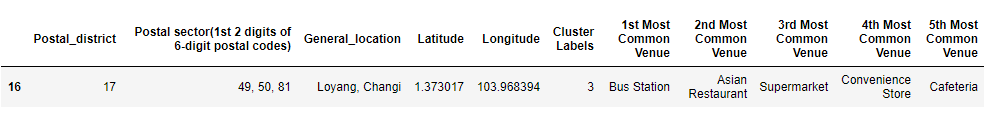


1. **Cluster 1 and 3.** We can observe that cluster 1's fifth most common venue is a military base and we will not be expecting much crowds and we can assume that most of the time will be spent in the base instead. Furthermore, both cluster 1 and 3 are located around the outskirts of the country and many tourists would usually hang out around the central region as that is a more convenient part of the country for travelling wise. Hence, these 2 clusters are not highly recommended to open a food business.

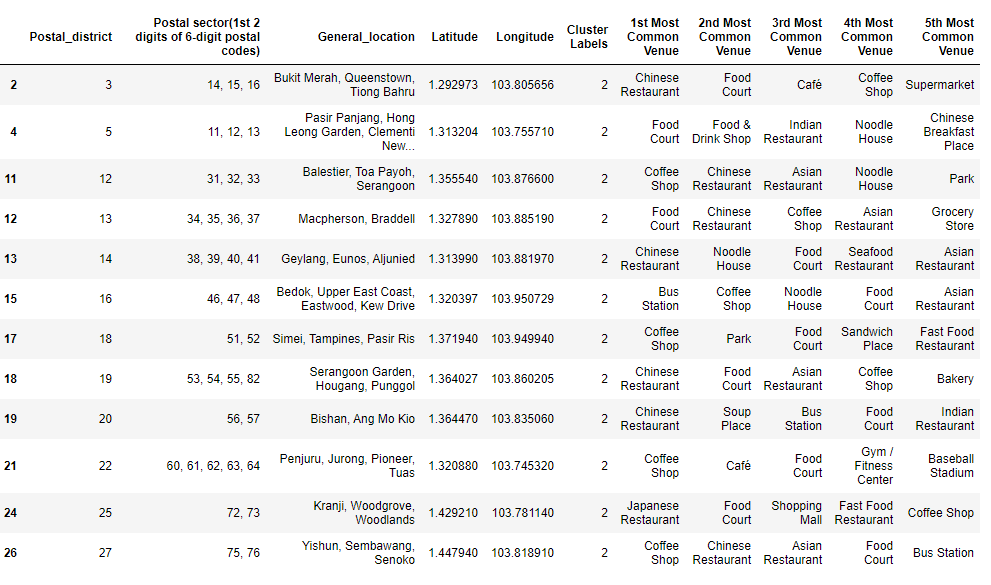
Cluster 1



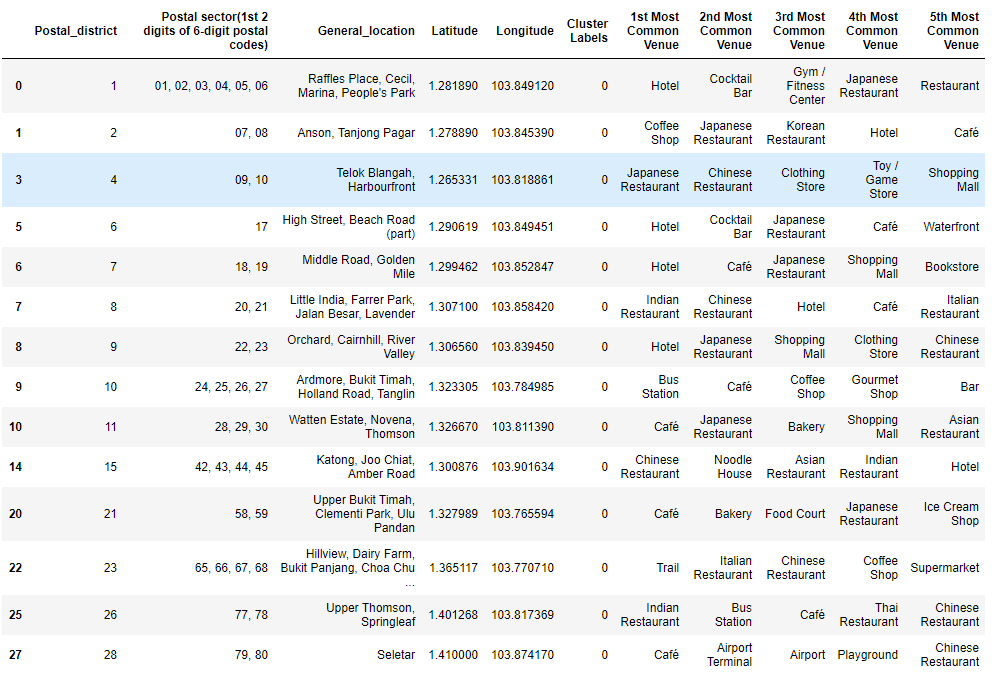
Cluster 2



1. **Cluster 2.** For cluster 2, there is an overwhelming amount of restaurants, bars and coffee shops which makes this a great cluster for opening a food business. This could also mean that those locations are where the crowds gather which reflects the high number of food businesses there. As we look into more details, the food businesses are mainly Asian and Chinese restaurants there which I made the assumption that the restaurants are targeting locals and base on research, 76% of Singapore's population are Chinese.

****

1. **Cluster 0**. Lastly, in cluster 0, there is a unique venue that only cluster 0 has, hotels. There are also a couple of Italian restaurants too. With this, I made the assumption that this cluster caters more towards the foreigner's area. Hotels mean that there will definitely be many people around the area be it tourists or locals going for a short staycation holiday. It is a great place to consider when opening a food business.



**Discussion and recommendation**

1. Based on the machine learning model that is being applied and after evaluating the maps and datasets based on location data only, I would recommend a food business to be opened at cluster 0 as this cluster is populated with not just Singaporeans, but also foreigners and this could mean more customers and more profit for the business. The fact that only cluster 0 has hotels in their venue categories is the winning edge among the remaining clusters.

**Conclusion**

1. There are some limitations and assumptions taken for this capstone project. There is only one factor to perform our recommendation and that that is from the location data derived from Foursquare. A sandbox tier Foursquare account was used for this project that further limits certain calls and requests from Foursquare itself. There are other possible factors to account for like the human traffic flow in that area, income of the residents residing there, rental rates and many more. However, only the district data is available and required for this project. Lastly, the assumption was made by taking the 100 venues within a 1km radius only.
2. This project goes through the data methodology process from understanding the business problem, data collection & preparation, exploring of data, data modelling and evaluation. The main programming language used was Python to do web scraping, retrieval of geographical coordinates, visualizing maps and applying k-means clustering. The output provides us with a recommendation for the target audience.

**References**

**Singapore’s postal district:**

<https://en.wikipedia.org/wiki/Postal_codes_in_Singapore>

**Foursquare API**

[**https://developer.foursquare.com/docs/places-api/endpoints/**](https://developer.foursquare.com/docs/places-api/endpoints/)

**Github repository of the code:**

[**https://github.com/eugeneloh1994/Coursera\_Captstone/blob/main/IBM%20Data%20Science%20Capstone%20Project.ipynb**](https://github.com/eugeneloh1994/Coursera_Captstone/blob/main/IBM%20Data%20Science%20Capstone%20Project.ipynb)

**Presentation in Medium:**

[**https://eugenelohh.medium.com/my-ibm-data-science-capstone-project-74ddcf1163dc**](https://eugenelohh.medium.com/my-ibm-data-science-capstone-project-74ddcf1163dc)