**Picking a suitable location for a new restaurant** in Western Singapore

IBM Data Science Capstone Project

Table of Contents

[01. Introduction 2](#_Toc75509629)

[1.1 The problem and background of the study 2](#_Toc75509630)

[1.2 Audience 2](#_Toc75509631)

[02. Data 2](#_Toc75509632)

[03. Methodology 3](#_Toc75509633)

[3. 1 Data Collection and preparation 3](#_Toc75509634)

[3.2 K-means Clustering 4](#_Toc75509635)

[04. Analysis 5](#_Toc75509636)

[05. Results 9](#_Toc75509637)

[06. Discussion and Recommendations 10](#_Toc75509638)

[07. Limitations 10](#_Toc75509639)

[08. Conclusions 10](#_Toc75509640)

# 01. Introduction

A restaurant is a business that prepares and serves food and drinks to customers. Meals are generally served and eaten on the premises, but many restaurants also offer take-out and food delivery services. Restaurants vary greatly in appearance and offerings, including a wide variety of cuisines and service models ranging from inexpensive fast-food restaurants and cafeterias, to mid-priced family restaurants, to high-priced luxury establishments.

Restaurants can be also considered as places where people can get together, have discussions, while eating food. So, restaurants play a major role in the society, and it greatly affects the lifestyle of people.

## 1.1 The problem and background of the study

In the project I have focused on a business scenario where the owner of a high-end restaurant chain is deciding to open a new restaurant in the Western Region of Singapore since he is not having a restaurant in Singapore. But he is uncertain about the city which he will choose to open the restaurant. The intention of the study is to find an optimal location in the western region for his restaurant.

## 1.2 Audience

Here we are assuming a random business owner to whom we are addressing this report, but in the real world this restaurant owner can be treated as a person who is interested in opening a restaurant in the western region of Singapore, thus this analysis could be useful for a group of restaurant owners.

# 02. Data

To perform the analysis, the following data will be used.

1. List of cities in Singapore.
2. Geo-coordinates of each city in Singapore
3. Top venues in the cities in western region.

The list of cities along with their coordinates has been obtained from the web site, <https://simplemaps.com/data/sg-cities>

Top venues have been obtained with the help of Foursquare API.

# 03. Methodology

## 3. 1 Data Collection and preparation

The list of cities in Singapore along with their coordinates has been obtained from the web site, <https://simplemaps.com/data/sg-cities>. Due to the limitations in the website in collecting the data, the details of only 75 cities were obtained. Thus, going forward this list of 75 cities will be used in the analysis.

The data obtained from the csv file were then transformed into a pandas DataFrame in the Jupyter Notebook for the analysis. A snapshot of the DataFrame obtained is as follows.



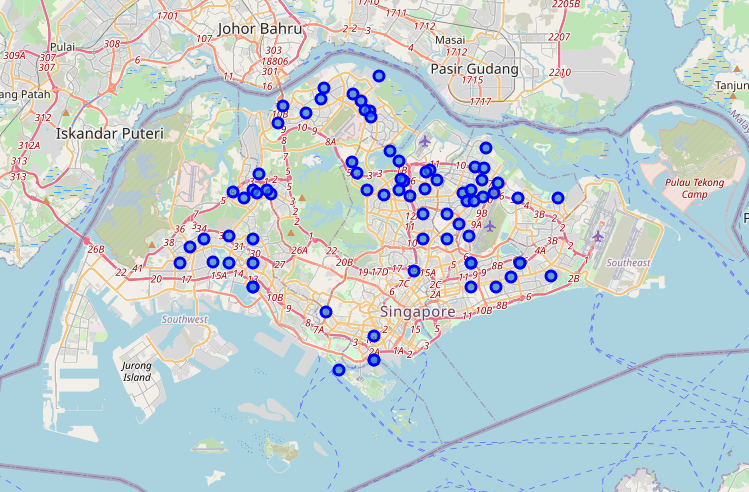
## 3.2 K-means Clustering

K-means clustering algorithm is the most selected technique to cluster data. K-means is a nonhierarchical clustering and use looping to group data into K groups. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

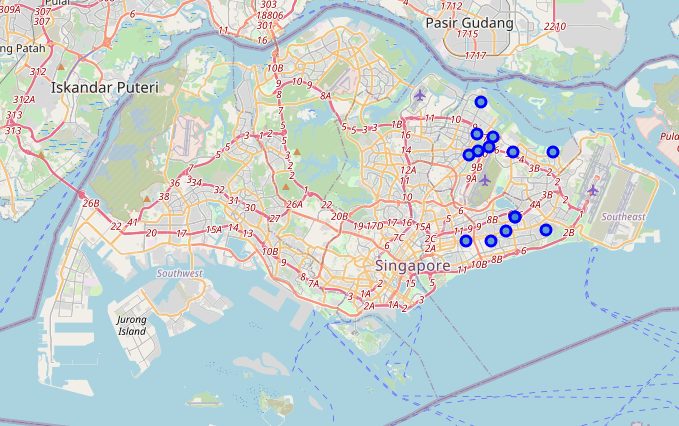
To achieve our objective K-means clustering has been used, with the optimal number of K clusters.

# 04. Analysis

In the first step, a visual representation of how the cities are situated in Singapore has been created.



Since our objective is to find an optimal location in the western region in Singapore for the restaurant, the data has been filtered and only the cities in Western Region has been obtained.



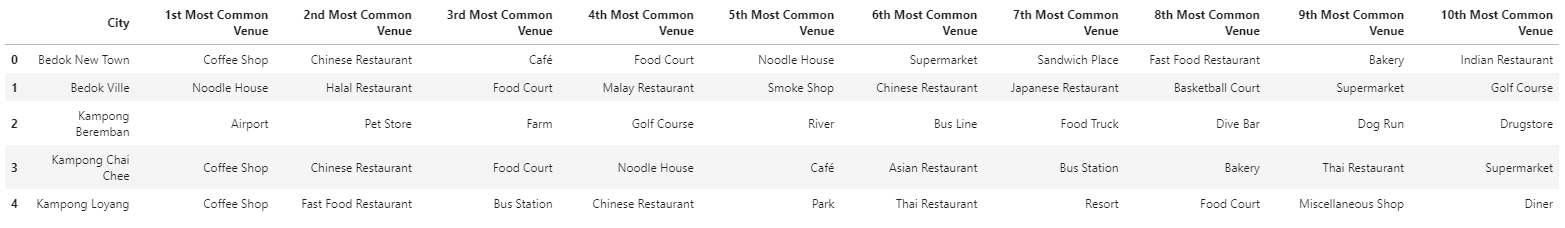
In the next step of the analysis, the cities were explored in greater detail. The venues were collected for each city via Foursquare API. After arranging the data there were up to 100 venues for each city. Venues are collected within a radius of 1000 meters from the point of city coordinates. The following table shows some venues from the first city.



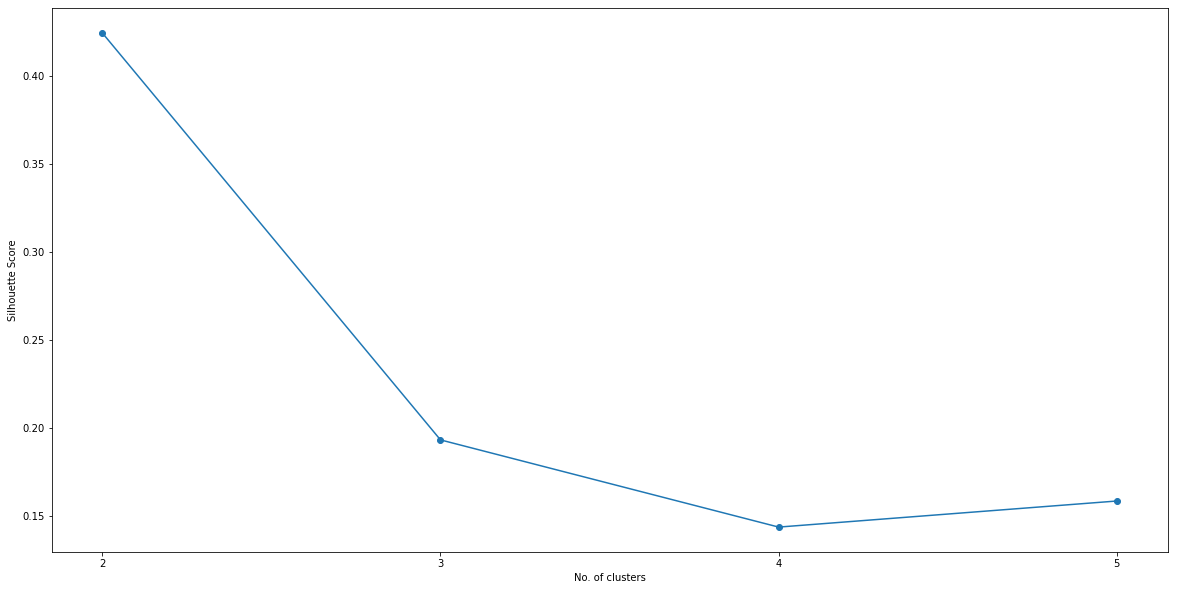
The number of Venues for each city in the western region has also been obtained.

For analysis in the cities, the focus is on venue categories. Therefore, we use one hot encoding to create dummy variables for the categories so the data set could be used to apply machine learning techniques.

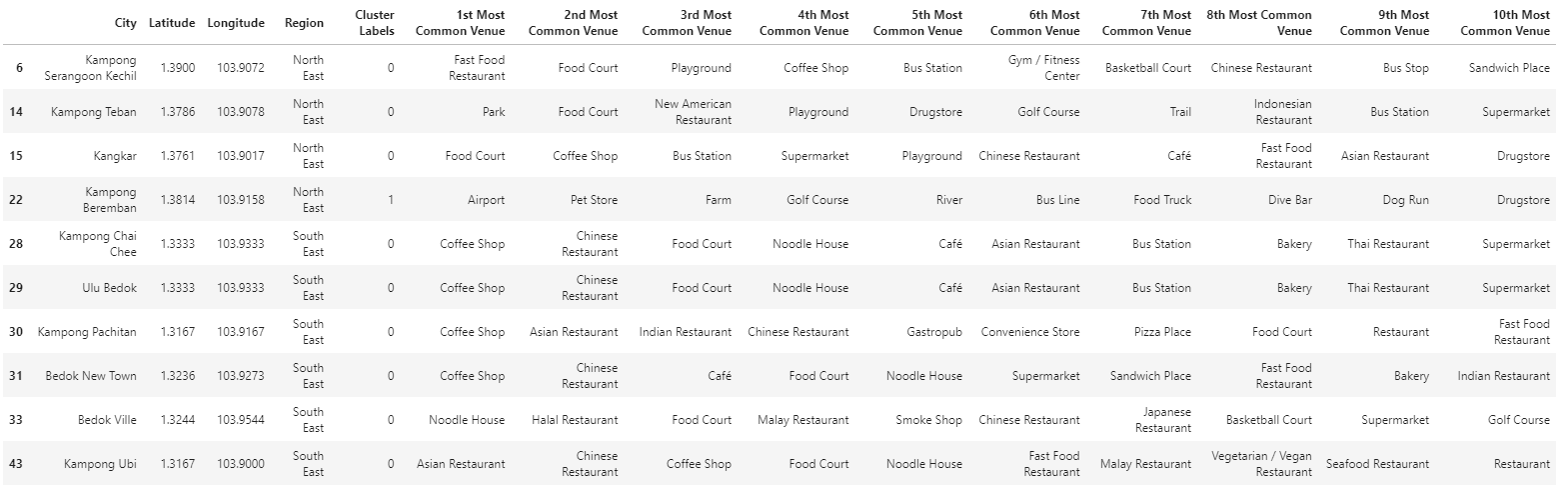
Then the following table with the 10 most common venues in each city has been obtained. (The table only displays the first 5 rows)

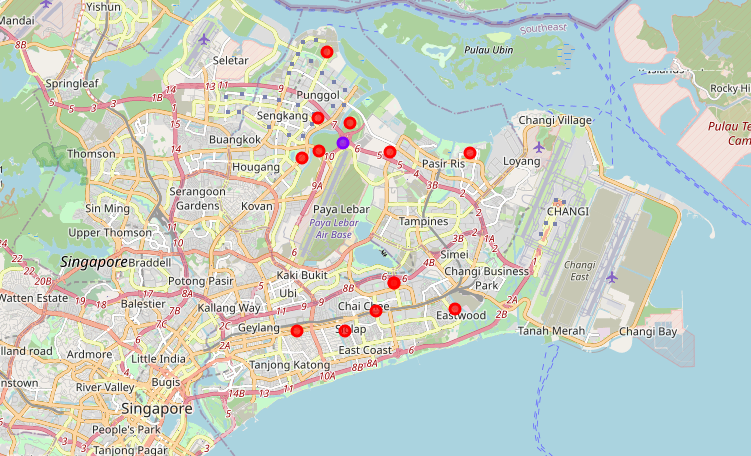


In the clustering process, the K-Means algorithm has been used. To identify the optimal number of K, the Silhouette Scoring method has been used, and the graph obtained were as follows. The optimal number of K’s was 2. Therefore, 2 clusters have been used.



Adding the cluster labels to the dataset we get the following table, and the clusters are also shown in the map.





# 05. Results

By looking at the clustering data, we can clearly identify the cluster which we are interested in.

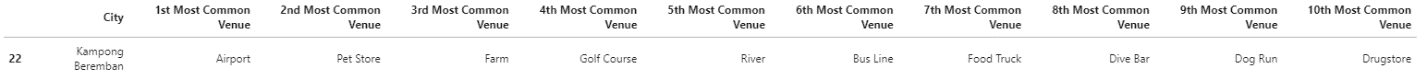
**Cluster 01**

The 1st cluster is exactly what we are looking for. We can observe many coffee shops and restaurants in the cluster 1.



**Cluster 02**

The cities which belong to cluster 2 are the ones the stakeholder should avoid considering to open the restaurant.



# 06. Discussion and Recommendations

Based on the results we’ve obtained, we can advise the restaurant owner to consider the cities in the cluster 1, which are the perfect locations for opening a restaurant. These are the cities where coffee shops, restaurants are very frequent. And he should avoid the cities under cluster 2, where no type of restaurant is preferred going to.

# 07. Limitations

The analysis was performed only on 75 cities in Singapore, due to the limitations of the website of obtaining the data.

The analysis on performed on city level.

When collecting the venues, a 1000-meter radius was considered around the center coordinates of the cities, and the number of collected venues for each city was limited to 100.

In the clustering process only the top 10 venues in each city in western Region of Singapore was considered.

# 08. Conclusions

In the analysis we have discussed the process of coming up with an answer for a real-life business scenario. The analysis was performed based on the tools of data science and relied heavily on the use of Python and Python libraries such s Pandas, Scikit learn, Folium etc. Data were collected from different sources in various formats. For the clustering process the K-means clustering process has been used.

From the results obtained we can clearly identify, which cluster we should consider in opening a new restaurant in the western region of Singapore. From the analysis it was found out that only one city should not be considered, since restaurants are not in the top 10 venues in that city. So, the owner should choose a city from the cluster one.