IBM Applied Data Science Capstone

Opening a new hotel in Singapore

Introduction

The tourism industry is an integral part for any country where hotels play an important role in ensuring that tourists have a proper and safe place to rest. Many developers are taking advantage of this trend to see where they can build more hotels to generate more revenue. Furthermore, opening hotels offer developers a somewhat steady stream of income especially in a country such as Singapore which is reliant on its tourism industry. However, it is important to weigh the pros and cons before making a business decision considering the high costs. Therefore, the importance of choosing an ideal location to set up a hotel cannot be understated.

Business Problem

The objective of this project is to select the best locations in the city of Singapore where it is ideal to open a new hotel This can be achieved by using data science and machine learning techniques such as clustering. This is especially prevalent in a country such as Singapore where land is scarce

Data

To tackle this issue, the data that is required is:

- A list of districts in Singapore, where the capital is Singapore
- The latitude and longitude coordinates of the districts to plot the map
- Proximity data, to see where the hotels are in each district

Sources of data

The district data (https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore) is obtained from Wikipedia where we can use web scrapping packages and techniques from beautifulsoup to extract data from this page.
Following that, we can use the geocoder package to obtain the exact latitude and longitude coordinates before visualising it on a map using folium. We will also be using Foursquare's API to get the venue data to check where the hotels are in each district.

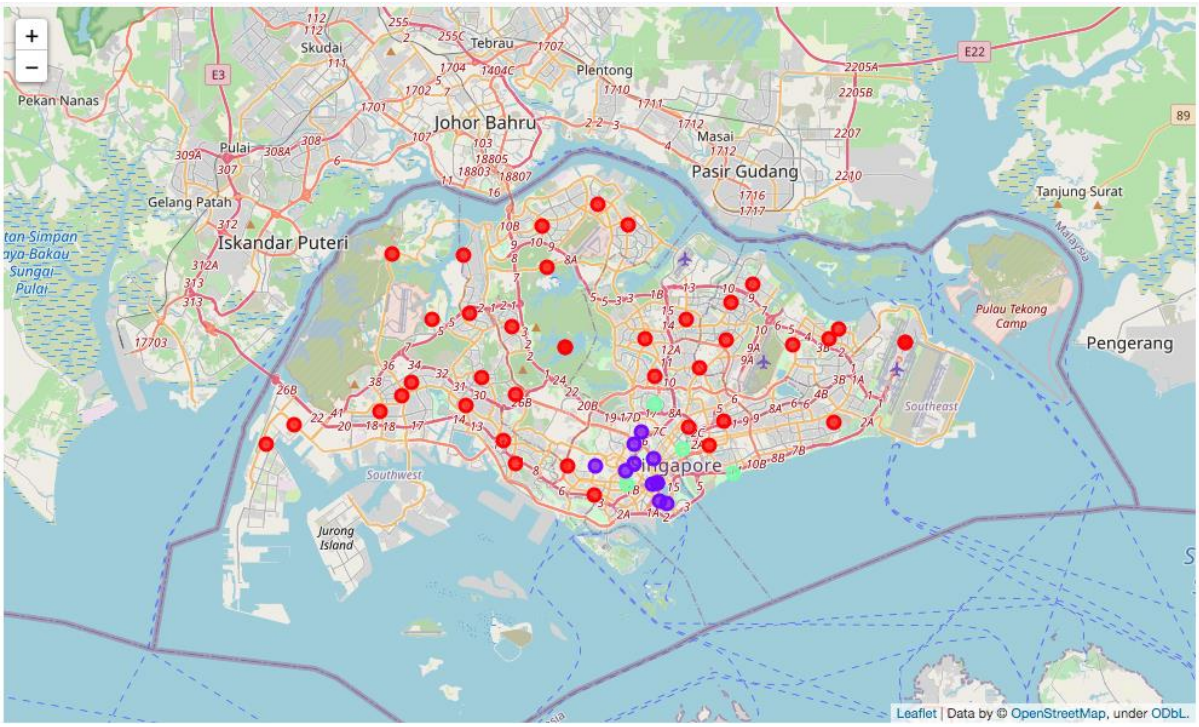
Methodology

Firstly, we have to extract data from the Wikipedia page stated in the previous section. We applied web scraping techniques using Python and the beautifulsoup package. This gives us a whole list of names but does not allow us to visualise anything on a map.
Following that, I applied the Geocoder package on Python to obtain the exact latitude and longitude coordinates for the various districts that I have scraped out. After obtaining the various coordinates, I was able to fill up the dataframe and visualise the neighbourhoods using the Folium package. This was done to ensure that I was on the right track and that I obtained accurate data.
Next, I tapped onto the Foursquare API to get the top 100 venues within a certain radius. I made the calls into the API with my ID and secret key. I then made API calls to Foursquare passing in the coordinates of the districts in a loop which was returned in a JSON format. This allowed me to obtain venue data and I examined the various categories that I obtained before finding the mean of the frequency of occurrence of each venue category. Lastly I just filtered it down to just "Hotels" since it is the scope of the project.

Finally, we are able to perform clustering on the data by using k-means clustering. This algorithm identifies k number of centroids before allocating every data point to the nearest cluster, keeping the centroids as small as possible. In my case, I decided to cluster the neighbourhoods into 3 clusters which is based on the frequency of “Hotels” in the area. This allows us to see which neighbourhoods have a high concentration of hotels and allows use to suggest to a developer where they should build a hotel to maximise profits.

Results

The results from the clustering based on frequency of hotels can be visualized and be seen below:



There are 3 clusters:

- Cluster 1 (Red): Districts with a high concentration of hotels
- Cluster 2 (Purple): Districts with a moderate concentration of hotels
- Cluster 3 (Light Green): Districts with a low concentration of hotels

Discussion

As observed from the map, majority of the districts have a high concentration of hotels around the area especially in Cluster 1. This can be explained due to the small land area of Singapore where it is very densely populated. Hotels in Cluster 1 are possibly suffering from fierce competition due to the oversupply and high concentration of hotels in that area. On the other hand, Cluster 2 and Cluster 3 are relatively around the same area and hotel developers can consider opening a hotel in the South side of Singapore where there is little competition. Considering that Singapore does not really have suburbs or rural areas, it does not really matter which area has the higher propensity to spend. Therefore, my suggestion is for the developer to avoid Cluster 1 and focus on finding suitable locations in either Cluster 2 or 3.

Limitations and Areas to Improve on

In this capstone project, only one factor was considered which is the frequency of hotels around the district. However, it should be noted that other factors can influence this decision such as the distance of the hotel to the various tourist attractions. Future research can be done to help obtain data on this to mitigate this issue. Additionally, the data that was obtained is limited to the number of API calls to Foursquare which could be improved as a collection of a wider set of data would lead to a better decision in this aspect.

Conclusion

This project has shed light on the process of identifying the problem, understanding the data that is needed, extracting the data, preparing the data and eventually performing machine learning techniques to solve issues. After clustering the districts, a more concise and confident recommendation can be made to developers with regards to building a hotel.

In conclusion, to answer the question that was raised in the beginning, a developer should choose to build in cluster 2 or 3 to minimize the competition in the area.

References

Planning Areas of Singapore, *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore

Foursquare Developers, *Foursquare*. Retrieved from https://foursquare.com/developers/apps

Appendix

Cluster 1

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
0	Ang Mo Kio	0.000000	0	1.371610	103.845460
1	Bedok	0.020000	0	1.324250	103.952970
2	Bishan	0.000000	0	1.350790	103.851100
3	Boon Lay	0.000000	0	1.346958	103.712757
4	Bukit Batok	0.000000	0	1.349520	103.752770
5	Bukit Merah	0.020000	0	1.283220	103.816760
6	Bukit Panjang	0.000000	0	1.378770	103.769770
7	Bukit Timah	0.000000	0	1.340410	103.772210
9	Changi	0.017544	0	1.369960	103.993110
10	Changi Bay	0.017544	0	1.369960	103.993110
11	Choa Chu Kang	0.000000	0	1.386160	103.746180
12	Clementi	0.000000	0	1.314380	103.765370
13	Downtown Core	0.000000	0	1.377160	103.955530
14	Geylang	0.000000	0	1.311470	103.882180
15	Hougang	0.000000	0	1.371240	103.891620
16	Jurong East	0.000000	0	1.334370	103.743670
17	Jurong West	0.000000	0	1.339490	103.707390
19	Lim Chu Kang	0.000000	0	1.419670	103.702320
20	Mandai	0.000000	0	1.412590	103.789687
23	Marine Parade	0.010000	0	1.321480	103.870480
24	Museum	0.000000	0	1.301160	103.771950
26	North-Eastern Islands	0.000000	0	1.366670	103.800000
30	Pasir Ris	0.000000	0	1.371940	103.949940
31	Paya Lebar	0.000000	0	1.325030	103.890490
32	Pioneer	0.000000	0	1.323297	103.646664
33	Punggol	0.000000	0	1.402460	103.906860
34	Queenstown	0.010000	0	1.299660	103.801720
37	Seletar	0.000000	0	1.382508	103.868626
38	Sembawang	0.000000	0	1.447940	103.818910
39	Sengkang	0.000000	0	1.392460	103.894590
40	Serangoon	0.000000	0	1.355540	103.876600
43	Southern Islands	0.000000	0	1.366670	103.800000
45	Sungei Kadut	0.000000	0	1.419098	103.742711
46	Tampines	0.000000	0	1.368190	103.929480
48	Tengah	0.000000	0	1.382499	103.724822
50	Tuas	0.000000	0	1.311820	103.630520
51	Western Islands	0.000000	0	1.330580	103.695220
53	Woodlands	0.011111	0	1.435850	103.786980
54	Yishun	0.000000	0	1.436210	103.835820

Cluster 2

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
8	Central Water Catchment	0.16	1	1.290410	103.852110
22	Marina South	0.12	1	1.278570	103.857620
25	Newton	0.15	1	1.312180	103.839120
27	Novena	0.10	1	1.319100	103.843720
28	Orchard	0.11	1	1.301090	103.839650
35	River Valley	0.11	1	1.296855	103.834348
36	Rochor	0.12	1	1.304130	103.850290
41	Simpang	0.16	1	1.290410	103.852110
42	Singapore River	0.13	1	1.289710	103.849640
44	Straits View	0.10	1	1.279863	103.853595
47	Tanglin	0.11	1	1.299751	103.817357
52	Western Water Catchment	0.16	1	1.290410	103.852110

Cluster 3

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
18	Kallang	0.07	2	1.309415	103.866730
21	Marina East	0.05	2	1.295790	103.895440
29	Outram	0.07	2	1.289241	103.835002
49	Toa Payoh	0.05	2	1.334480	103.851080