Coursera Capstone

IBM Data Science Professional Certificate

Planning to open a New Shopping Mall in Kuala Lumpur, Malaysia

Week 5

Prepared By: Divin Divakar
Dated: 23 May 2021

Introduction

Shopping malls are a popular way for many people to unwind and enjoy themselves on weekends and holidays. They can shop for groceries, eat at restaurants, shop at various fashion outlets, watch movies, and engage in a variety of other activities. Shopping malls provide as a one-stop shop for a variety of shoppers.

The shopping malls' central location and vast crowds provide an excellent distribution route for shops to sell their products and services. Property developers are also capitalizing on this trend by constructing more shopping malls to meet demand. As a result, Kuala Lumpur has a plethora of retail malls, with more being erected all the time.

Property developers can earn continuous rental income by developing shopping malls. Of course, opening a new retail mall, like any other business decision, necessitates careful analysis and is far more complicated than it appears. The site of the shopping mall is one of the most critical factors that will determine whether the mall succeeds or fails.

Business Problem / Problem Statement

The goal of this capstone project is to research and pick the finest locations for a new retail mall in Kuala Lumpur, Malaysia. This project seeks to deliver solutions to the business question using data science methodology and machine learning techniques such as clustering.

If a property developer wanted to build a new retail mall in Kuala Lumpur, Malaysia, where would you suggest they build it?

Target Audience of this project

Property developers and investors planning to open or invest in new shopping malls in Malaysia's capital city, Kuala Lumpur, will find this project particularly valuable. This initiative comes at a good moment because the city is now experiencing an overabundance of shopping complexes.

According to data issued last year by the National Property Information Centre (NAPIC), existing mall space will be expanded by 15%, and total occupancy might fall below 86%.

In March of last year, the local newspaper The Malay Mail stated that genuine mall occupancy rates in some locations could be as low as 40%, citing a Financial Times (FT) piece detailing the country's persistent infatuation with creating more shopping space amid chronic oversupplies.

Data

We will need the following information to solve the problem:

- A list of Kuala Lumpur neighborhoods. This outlines the project's scope, which is limited to the city of Kuala Lumpur, Malaysia's capital, and largest metropolis in Southeast Asia.
- Coordinates of those neighborhoods' latitude and longitude. This is essential for both plotting the map and retrieving the venue information.
- Statistics about venues, notably data on commercial malls. This information will be used to cluster the neighborhoods.

Data Sources and Methods of Extraction

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) offers a list of 70 different neighborhoods in Kuala Lumpur. With the help of Python requests and beautiful soup packages, we will extract data from the Wikipedia page using web scraping techniques. Then, using the Python Geocoder library, we will get the geographical coordinates of the neighborhoods, which will give us their latitude and longitude coordinates.

Following that, we will use the Foursquare API to gather venue data for those areas. Foursquare is used by over 125,000 developers and has one of the largest databases of 105+ million places.
The Foursquare API will provide a variety of venue data categories, but we are particularly interested in the Shopping Mall category to help us address the business challenge we have presented.

This project will require a wide range of data science abilities, including web scraping (Wikipedia), API work (Foursquare), data cleaning, data wrangling, machine learning (K-means clustering), and map visualization (Folium).

The Methodology part will cover the procedures performed in this project, the data analysis that was done, and the machine learning methodology that was utilized.

Methodology

To begin, we must obtain a list of the city of Kuala Lumpur's neighborhoods. Fortunately, the list can be found on Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur). To extract the list of neighborhoods data, we will use web scraping with Python requests and beautiful soup tools. This is, however, only a list of names. To use the Foursquare API, we need to obtain geographical coordinates in the form of latitude and longitude.

To do so, we will use the fantastic Geocoder library, which allows us to translate addresses into geographical coordinates (latitude and longitude). We will collect the data, load it into a panda Data Frame, and then use the Folium package to show the neighborhoods on a map. This allows us to run a sanity check to ensure that the geographical coordinates data supplied by Geocoder is plotted accurately in Kuala Lumpur.

We will then use the Foursquare API to get the top 100 venues within a 2000-meter radius. To access the Foursquare ID and Foursquare secret key, we must first create a Foursquare Developer Account. In a Python loop, we then make API requests to Foursquare, passing in the geographical coordinates of the neighborhoods.

Foursquare will supply the venue data in JSON format, from which we will extract the name, category, latitude, and longitude of the venue. We can use the data to see how many venues were returned for each neighborhood and how many distinct categories can be curated from all the venues that were returned.

Then, by grouping the rows by neighborhood and calculating the mean of the frequency of occurrence of each venue category, we will analyze each neighborhood. We are also prepping the data for clustering by doing so. We will filter the "Shopping Mall" as a venue category for the neighborhoods because we are analyzing the "Shopping Mall" data.

Finally, we will use k-means clustering to achieve data clustering. The K-means clustering algorithm finds k centroids and then assigns each data point to the closest cluster, keeping the centroids as tiny as possible. It is one of the most basic and widely used unsupervised machine learning methods, and it is well suited to the task at hand.

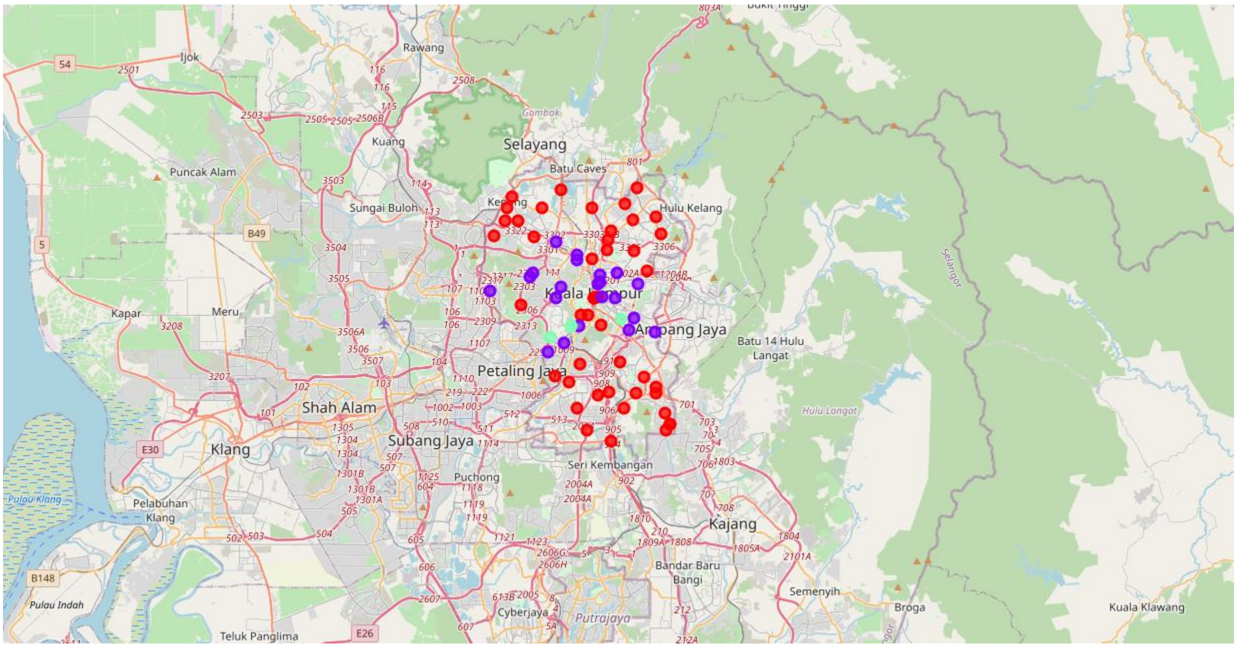
We will divide the neighborhoods into three groups based on the frequency with which "Shopping Mall" appears. The findings will help us determine which areas have a larger concentration of shopping malls and which areas have a lower number of shopping malls. According to the number of retail centers in various neighborhoods, that will enable us to convey the best neighborhood for the shopping mall.

Results

Based on the frequency of occurrence for “Shopping Mall,” the k-means clustering findings reveal that we can group the neighborhoods into three clusters:

- Cluster 0: Neighborhoods with a moderate number of shopping malls
- Cluster 1: Neighborhoods with a low number of shopping malls or none
- Cluster 2: Neighborhoods with a high density of shopping malls

The clustering findings are depicted in the map below, with cluster 0 being red, cluster 1 being purple, and cluster 2 being mint green.



Discussion

Most shopping malls are concentrated in the central region of Kuala Lumpur city, with the largest number in cluster 2 and a reasonable number in cluster 0. As seen in the map in the Results section, most shopping malls are in the central region of Kuala Lumpur city, with the highest number in cluster 2 and a moderate number in cluster 0. Cluster 1 on the other hand, has an exceedingly small number of retail malls in its surrounding areas, if any at all.

There is little to no competition from current malls, making this a terrific chance and high-potential area for new shopping malls. Meanwhile, due to overstock and high concentration of retail malls, shopping malls in cluster 2 are likely to face fierce rivalry. From a different perspective, the findings reveal that the overstock of retail malls occurred mostly in the city's center area, with the suburbs having very few malls. As a result, this initiative advises property developers to use these data to establish new shopping malls in cluster 1 neighborhoods where there is little to no competition.

Property developers who want to differentiate themselves from the competition can open new shopping malls in cluster 0 neighborhoods with moderate competitiveness. Finally, property developers should avoid cluster 2 neighborhoods, which already have a high concentration of shopping malls and are subject to fierce rivalry.

Limitations and Suggestions

We only consider one aspect in this project, which is the frequency of occurrence of shopping malls; nevertheless, other factors such as population and household income could impact the location choices of a new shopping mall. However, to the best of this researcher's knowledge, such data are not available at the neighborhood level that this project requires. Future study could develop a system for estimating such data, which could then be used in the clustering process to select the best places for a new retail mall to open. In addition, this project took use of the Foursquare API's free Sandbox Tier Account, which has constraints on the amount of API calls and results produced. Future study could use a premium account to get around these restrictions and get greater results.

Conclusion

In this project, we identified the business challenge, specified the data required, extracted, and prepared the data, performed machine learning by clustering the data into three clusters based on their similarities, and finally provided suggestions to the relevant stakeholders regarding the optimal locations for a new retail mall for the property developers and investors.

The response proposed by this project to the business issue presented in the introduction section is: Cluster 1 neighborhoods are the most favored sites to develop a new retail mall. The outcomes of this project will aid key parties in their decisions to establish a new retail mall by allowing them to capitalize on opportunities in high-potential sites while avoiding overcrowding.

References

Category: Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

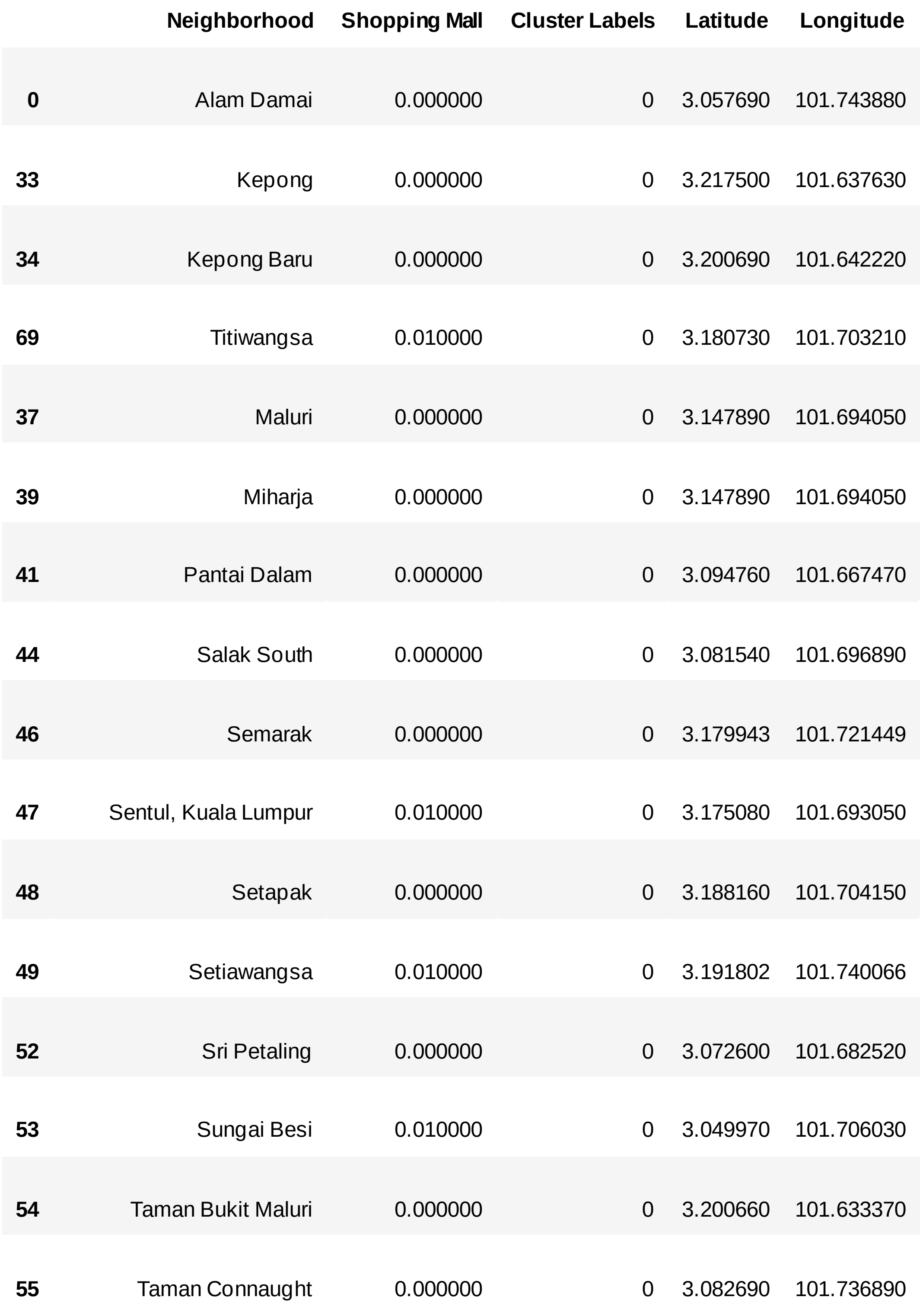
Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs

Malay Mail. (2018, March 14). Malls facing meltdown as glut continues. *Malay Mail*. Retrieved from https://www.malaymail.com/s/1597735/malls-facing-meltdown-as-glut-continues

Tan, H. H. (2018, April 5). An oversupply of retail space in Malaysia. *StarProperty.my*. Retrieved from http://www.starproperty.my/index.php/articles/property-news/an-oversupply-of-retail-space-in-Malaysia/

Appendix

Cluster 0



	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
0	Alam Damai	0.000000	0	3.057690	101.743880
33	Kepong	0.000000	0	3.217500	101.637630
34	Kepong Baru	0.000000	0	3.200690	101.642220
69	Titiwangsa	0.010000	0	3.180730	101.703210
37	Maluri	0.000000	0	3.147890	101.694050
39	Miharja	0.000000	0	3.147890	101.694050
41	Pantai Dalam	0.000000	0	3.094760	101.667470
44	Salak South	0.000000	0	3.081540	101.696890
46	Semarak	0.000000	0	3.179943	101.721449
47	Sentul, Kuala Lumpur	0.010000	0	3.175080	101.693050
48	Setapak	0.000000	0	3.188160	101.704150
49	Setiawangsa	0.010000	0	3.191802	101.740066
52	Sri Petaling	0.000000	0	3.072600	101.682520
53	Sungai Besi	0.010000	0	3.049970	101.706030
54	Taman Bukit Maluri	0.000000	0	3.200660	101.633370
55	Taman Connaught	0.000000	0	3.082690	101.736890

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
56	Taman Desa	0.010000	0	3.102970	101.684710
58	Taman Ibukota	0.000000	0	3.212310	101.715250
59	Taman Len Seng	0.000000	0	3.069080	101.742870
60	Taman Melati	0.010000	0	3.223570	101.723990
61	Taman Midah	0.000000	0	3.093590	101.728370
62	Taman OUG	0.000000	0	3.210051	101.634508
63	Taman P. Ramlee	0.000000	0	3.193940	101.705730
64	Taman Sri Sinar	0.010204	0	3.190070	101.652930
65	Taman Taynton View	0.000000	0	3.087070	101.736810
68	Taman Wahyu	0.000000	0	3.222400	101.671730
32	Kampung Padang Balang	0.000000	0	3.209430	101.693180
30	Kampung Datuk Keramat	0.010000	0	3.166400	101.730460
35	Kuchai Lama	0.000000	0	3.090690	101.677320
70	Wangsa Maju	0.010000	0	3.203870	101.737150
16	Bukit Petaling	0.010000	0	3.129290	101.698960
18	Cheras, Kuala Lumpur	0.000000	0	3.061870	101.746750
10	Batu, Kuala Lumpur	0.000000	0	3.147890	101.694050
9	Batu 11 Cheras	0.000000	0	3.061870	101.746750
5	Bandar Tun Razak	0.000000	0	3.082760	101.722810

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
13	Bukit Jalil	0.000000	0	3.057810	101.689650
21	Damansara Town Centre	0.010000	0	3.136444	101.690294
3	Bandar Sri Permaisuri	0.000000	0	3.103910	101.712260
4	Bandar Tasik Selatan	0.010204	0	3.072750	101.714610
23	Desa Petaling	0.000000	0	3.083300	101.704380
24	Federal Hill, Kuala Lumpur	0.010000	0	3.136370	101.685640
25	Happy Garden	0.000000	0	3.201630	101.721070
1	Ampang, Kuala Lumpur	0.010000	0	3.148499	101.696728
27	Jinjang	0.000000	0	3.209500	101.658740
2	Bandar Menjalara	0.010000	0	3.190350	101.625450
14	Bukit Kiara	0.000000	0	3.143480	101.644330

Cluster 1

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
66	Taman Tun Dr Ismail	0.03	1	3.152830	101.622710
8	Bangsar South	0.02	1	3.111020	101.662830
57	Taman Duta	0.02	1	3.155620	101.671840
67	Taman U-Thant	0.03	1	3.157700	101.724520
11	Brickfields	0.03	1	3.129160	101.684060

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
12	Bukit Bintang	0.03	1	3.147770	101.708550
29	Kampung Baru, Kuala Lumpur	0.03	1	3.165460	101.710280
50	Shamelin	0.02	1	3.124570	101.735970
31	Kampung Kasipillay	0.02	1	3.177760	101.682400
26	Jalan Cochrane, Kuala Lumpur	0.03	1	3.134620	101.721705
22	Dang Wangi	0.02	1	3.157825	101.697280
38	Medan Tuanku	0.02	1	3.159260	101.698340
40	Mont Kiara	0.03	1	3.165290	101.652420
20	Damansara Heights	0.02	1	3.147970	101.667950
51	Sri Hartamas	0.02	1	3.162200	101.650360
15	Bukit Nanas	0.02	1	3.148609	101.699854
43	Putrajaya	0.02	1	3.125860	101.718624
19	Chow Kit	0.02	1	3.163780	101.698140
45	Segambut	0.02	1	3.186500	101.667950
17	Bukit Tunku	0.02	1	3.173810	101.682760
28	KL Eco City	0.03	1	3.117130	101.673840

Cluster 2

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
42	Pudu, Kuala Lumpur	0.04	2	3.133540	101.713070
36	Lembah Pantai	0.04	2	3.121189	101.663889
7	Bangsar Park	0.05	2	3.129200	101.678440
6	Bangsar	0.05	2	3.129200	101.678440

-- End of Document --