Coursera Capstone Project

# IBM Applied Data Science Capstone

Opening a New Shopping Mall in Kuala Lumpur, Malaysia

## By: Arjun Thottappillil

A view of a city with tall buildings in the background

Description automatically generated

**Introduction**

For many people, shopping malls are locations for them to relax and unwind on the weekends and holidays. Malls are an all-in-one destination for both residents and tourists, by allowing them a multitude of options to spend their time with. For the retail store owners, the central location and large number of potential visitors that most shopping malls attract serves as an ideal distribution channel for them to market their products and services. Property developers can receive a steady stream of rental income from retailer stores who join them. However, the success of a shopping mall is greatly dependent on its location. Some of the problems that can arise from choosing bad locations range from oversupply to intense competition between shopping malls that are in proximity of one another.

**Business Problem**

The main objective of this capstone project is to analyze and then select the ideal locations in the city of Kuala Lumpur, Malaysia in which to build a new shopping mall. Through the use of data science methodology and a variety of machine learning techniques like clustering an k-means, the end goal of this project is to find the best solution to the business question: In the city of Kuala Lumpur, Malaysia, for a property develop seeking to open a new shopping mall, where is the ideal location to build the mall?

**Target Audience of this project**

The main benefiters of this project would be the property developers and investors who are interested in opening or investing in the opening of new shopping malls in the capital city of Malaysia, Kuala Lumpur. In recent years Kuala Lumpur has had issues with oversupply of their shopping malls in one location rather than being more distributed and spread out. This project can serve as a guide to opening malls in more strategic locations to avoid the oversupply issue and increase the profitability of the mall.

**Data**

**To find an optimal solution to this issue, we will be using the following data:**

* A list of all the neighborhoods in Kuala Lumpur. This will help use determine the scope of the project which will be restricted to the capital city of Malaysia in South East Asia, Kuala Lumpur.
* Latitude and Longitude coordinates from these neighborhoods. For creating a plotted map and to get the venue data.
* Venue data, specifically data retaining to shopping malls. We will make use of this data to perform clustering on the neighborhoods.

**Data Sources and Tools**

Wikipedia provides the list of all 71 neighborhoods in Kuala Lumpur, Malaysia. (<https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur>)

The Python Geocoder package provides the means to calculate the latitude and longitude for all neighborhoods.

The Foursquare API provides the venue data for these neighborhoods. Foursquare is one of the largest databases of over 105 million locations and is used by over 150,000 developers. While the Foursquare API will give us all the venues near the neighborhoods, we will focus on the Shopping Mall category for our use case.

We use k-means and clustering to categorize the frequency of the venues and then filter them for only the shopping malls near by and then record it for the next step.

Folium is used for the final map visualization of the data.

**Methodology**

The first step is to retrieve the list of neighborhoods in Kuala Lumpur, Malaysia, using the Wikipedia page. We will then perform web scraping using Python requests and beautifulsoup packages to create a new list of all the neighborhoods in Kuala Lumpur. The second step would be to use the geocoder package to get the geographical coordinates of the neighborhoods in the form of latitude and longitude to use Foursquare API. After that we will populate the data into a panda Dataframe and then create a folium visualization of a map of Kuala Lumpur with markers on all the neighborhood locations. This will allow us to ensure that all the geographical coordinates data retrieved by Geocoder are correctly plotted in the city of Kuala Lumpur.

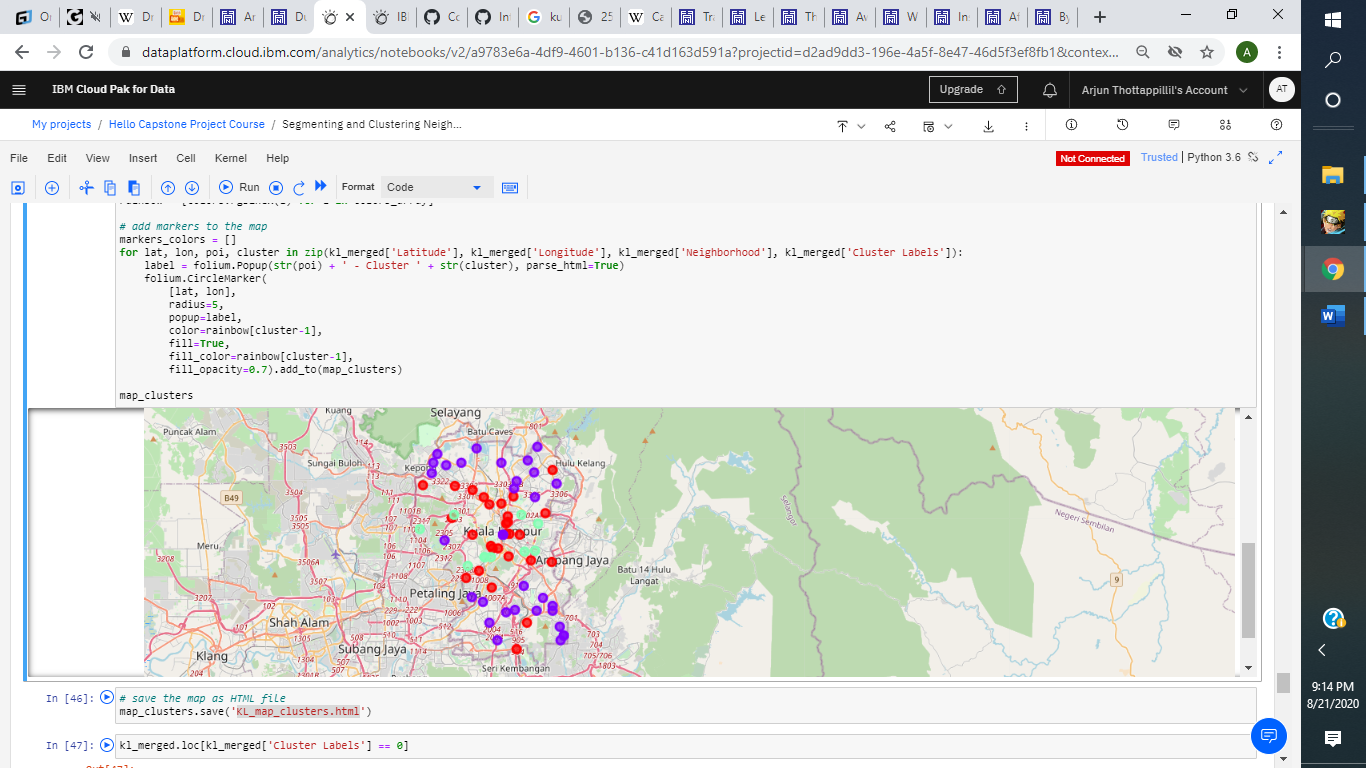
From there we start using the Foursquare API to find the top 100 venues within a 2000 meters of each neighborhood using the registered Foursquare developer account’s Foursquare ID and secret key. Then using the API calls to Foursquare passing in the geographical coordinates of the neighborhood using a Python loop. Foursquare will return the venue data in JSON format which will then be able to extract the venue name, venue category, venue latitude and longitude. Using this data, we can then examine how many venues were near each neighborhood and then check for unique categories. From that we will group all the rows by neighborhood and then check for the mean frequency of each venue category. With that the data is prepared for clustering. Then we filter the venues to only include “Shopping Mall” as the venue category and exclude all the other categories from the neighborhoods.

We then perform the clustering on the data using the k-means clustering method. This method identifies k number of centroids, and then allocates all the data points to the cluster closest to it while keeping then centroids as small as possible. We will then divide the clusters basted on the mean frequency of the occurrence of the “Shopping Mall” venue category. This will give us a clear idea of which neighborhoods are the most shopping malls nearby and which ones have fewer shopping malls nearby. This will help give us a clear idea of what neighborhoods will be most suitable to open shopping malls nearby.

**Results**

The results from the k-means clustering shows us how we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

* Cluster 0: Neighborhoods with moderate number of shopping malls
* Cluster 1: Neighborhoods with very few to zero shopping malls
* Cluster 2: Neighborhoods with a high concentration of shopping malls

The results of the clustering are displayed on the map below cluster 0 is represented with the red color, cluster 1 is the purple color and cluster 2 is mint green color.

**Discussion**

Based on the observations from the map in the Results section, most shopping malls in Kuala Lumpur are concentrated near central Kuala Lumpur. Most shopping malls are found in cluster 2 and cluster 0, with cluster 2 containing around 150% more shopping malls compared to cluster 0 on average. In contrast, cluster 1 has only a very low number of shopping malls or no shopping malls at all in their neighborhoods. In comparison, cluster 2 shopping malls likely have an issue with over saturation and intense competition due to have too many shopping malls in such a proximity from one another.

This map indicates that there is a high concentration of shopping malls in the central part of the city and a low concentration of shopping malls in the suburban part of the city. Therefore cluster 1 is the best region for property developers to consider for new malls due to minimal competition. Cluster 0 could also be considered, so long as the mall to be built has a unique enough selling proposition to stand out amongst the competitors. Finally, neighborhoods in cluster 2 should be avoided as they already have a high saturation of shopping malls that are most likely suffering from intense competition which will affect profit margins.

**Limitations and Suggestions for Future Research**

In this project, we only considered one factor: the frequency of shopping malls, already present in the neighborhoods. There are other factors like population density and average income of nearby residents which could be used to influence not just the location but also the products and services that the shopping mall may focus on. However, this information is not collected on the neighborhood level required for this project. Future research could develop a methodology to estimate such data using a clustering algorithm to determine the best locations to open a new shopping mall. The Foursquare free sandbox tier account that was used for this project had a limited number of API calls and results returned. Future research done by property developers could invest in the paid account to bypass these limitations and therefore obtain more optimized results.

**Conclusion**

Throughout the course of this project, we went through the process of identifying the business problem, specifying the data required, extracting and preparing data, performing machine learning by clustering the data into 3 clusters based on their similarities, and finally we provided recommendations to the property developers and investors who would benefit from this information pertaining to the best locations to open a new shopping mall to minimize competition. To answer this question: The neighborhoods in cluster 1 are the most optimal location to open a new shopping mall. These findings will allow the property developers and investors to capitalize on the opportunities on these high potential locations while also avoiding the overcrowding areas that suffer from intense competition when they make their decision on where to open a new shopping mall.

**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/>