**Coursera Capstone**

**IBM Applied Data Science Capstone**

**Opening a New Shopping Mall in Kuala Lumpur, Malaysia**

**Introduction**

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Kuala Lumpur and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

**Business Problem**

The objective of this capstone project is to analyse and select the best locations in the city of Kuala Lumpur, Malaysia to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Kuala Lumpur, Malaysia, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

**Target Audience of this project**

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the capital city of Malaysia i.e. Kuala Lumpur. This project is timely as the city is currently suffering from oversupply of shopping malls. Data from the National Property Information Centre (NAPIC) released last year showed that an additional 15 per cent will be added to existing mall space, and the agency predicted that total occupancy may dip below 86 per cent. The local newspaper The Malay Mail also reported in March last year that the true occupancy rates in malls may be as low as 40 per cent in some areas, quoting a Financial Times (FT) article cataloguing the country's continued obsession with building more shopping space despite chronic oversupply.

**Data**

**To solve the problem, we will need the following data:**

• List of neighborhoods in Kuala Lumpur. This defines the scope of this project which is confined to the city of Kuala Lumpur, the capital city of the country of Malaysia in South East Asia.

• Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

**Sources of data**

* Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs\_in\_Kuala\_Lumpur) contains a list of neighborhoods in Kuala Lumpur, with a total of 70 neighborhoods.
* Geographical coordinates of the neighborhoods using Python Geocoder package
* Foursquare API provide many categories of the venue data

**Methodology**

* Web scraping Wikipedia page for neighborhoods list
* Get latitude and longitude coordinates using geocoder
* Foursquare API to get venue data
* Filter venue category by shopping mall
* Perform Clustering on the data using k-means clustering
* Visualize the clusters using folium map

**Results**

The results from the k-means clustering show that 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 low number to no existence of shopping malls

• Cluster 2: Neighborhoods with high concentration of shopping malls. The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color