Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Mall in Kuala Lumpur, Malaysia

Created By: Abhishek Srivastava

February 2021

Intro – Data Section

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

- List of neighbourhoods 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 neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to malls. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains a list of neighbourhoods in Kuala Lumpur, with a total of 71 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use **Foursquare API** to get the venue data for those neighbourhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, such as

- Web scraping
- Working with API (Foursquare),
- Data cleaning,
- Data wrangling,
- Machine learning (K-means clustering)
- Map visualization/plotting (Folium).

In the final report, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.