# Coursera Capstone

# IBM Applied Data Science Capstone

## Buying a House in Kuala Lumpur

Muhammad Syahril

## Introduction

House is one of the essential requirements for humans. In this fast-moving world, house prices increased year by year. Purchasing a house is one of the largest investments to be made. Nowadays, demand for housing increases as the population of Malaysia increases but the supply of housing is overpriced that people cannot afford. Housing affordability have been the major concerns issues for housing market. Location plays a main role of choosing a house.

Getting information for housing market is a hassle since the information sometimes uses terms that is too complicated to understand. Due to the fast growing of the housing market, it is hard to keep up with new information. In correlation to this giving information on the housing market is one of important factor for the community to gain knowledge easily on housing news.

## Business Problems

The objective of this capstone project is to analyse and select the best locations in the city of Kuala Lumpur, Malaysia to buy a house. 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 homebuyer is looking to buy a house, where would you recommend that they buy it?

## Target and Significance

Homebuyers will be one of the benefiters from this system. They can gather the information given for the housing market and also suitable houses for each location. With this system, homebuyers can make decision on buying houses that location wise or just for getting knowledge on housing in Kuala Lumpur, Malaysia.

## Data Acquisition

Data 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 shopping malls. We will use this data to perform clustering on the neighbourhoods.

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 has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping 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, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, 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.

# Methodology

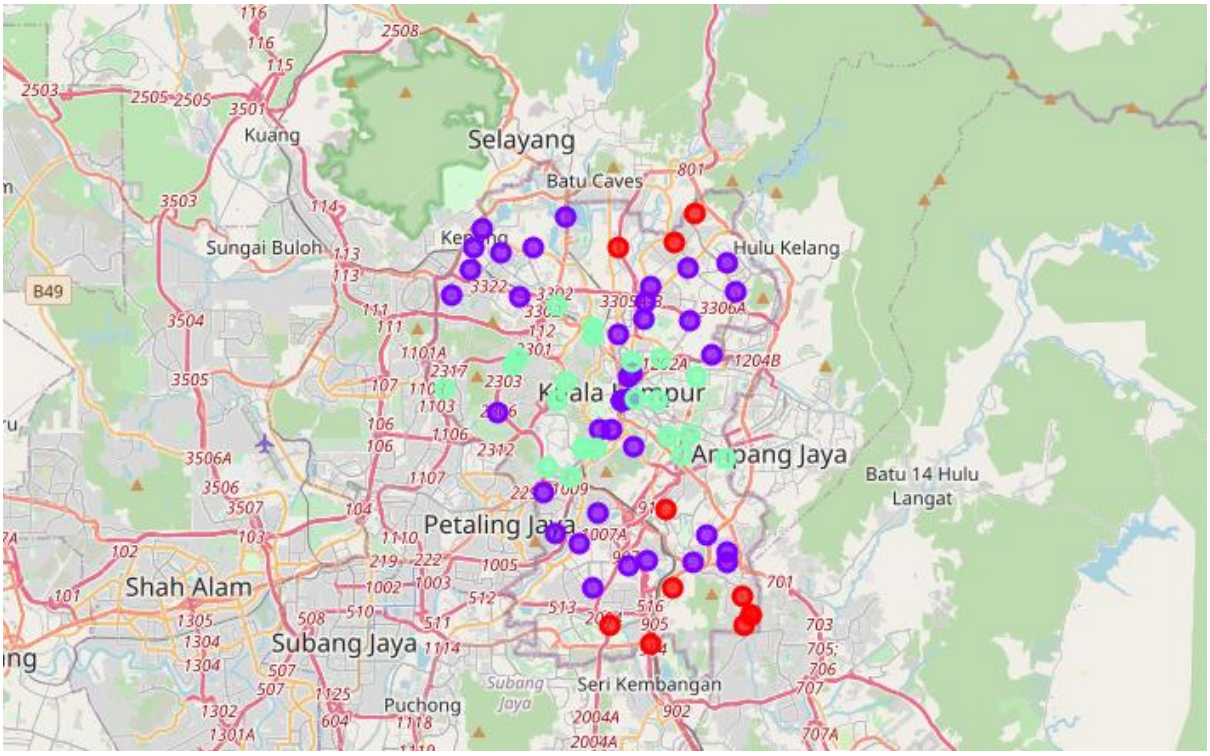
1. Data Acquisition and Wrangling
    - Using Python Request and beautifulsoup.
    - Extract data of neighbourhoods in Kuala Lumpur at Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur).
    - Using Geocoder to gather geographical coordinates in form of latitude and longitude.
    - Using Folium to visualize map to check if the coordinates are correct.
    - Using Foursquare to get top 100 venues that are within a radius of 2000 meters.
    - Check all unique categories and analyse by taking mean of the frequency each of category.
    - Collect and create data frame for categories that have been chosen (Shopping Mall, Supermarket, Convenience Store, Halal Restaurant, Grocery Store, Gas Station).
2. Clustering using K-means
    - K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
    - Cluster the neighbourhoods into 3 clusters based on all of categories frequency.
    - The results will allow us to identify which neighbourhoods have higher concentration of categories.
3. Visualizations
    - By using Folium, visualize map for the 3 cluster.

## Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for Categories (Shopping Mall, Supermarket, Convenience Store, Halal Restaurant, Grocery Store, Gas Station).

- Cluster 0: Neighbourhoods with no presence of Supermarket and majority of Shopping Mall is not exist.
- Cluster 1: Neighbourhoods with low number in majority of all the categories.
- Cluster 2: Neighbourhoods with high frequency of Shopping Mall and majority decent value of other categories.

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



## Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Kuala Lumpur city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no shopping mall in the neighbourhoods. This cluster 2 represents high potential area to buy a house as there is very many eases of access of necessities and convenience. Meanwhile, neighbourhood in cluster 0 are likely suffering from shortage of available access to necessities.

From another perspective, the results also show that the oversupply of necessities building mostly happened in the central area of the city, with the suburb area still have very few necessities building.

Therefore, this project recommends homebuyers to capitalize on these findings to buy a house in cluster 2 that could give benefits according to lifestyle and necessities. Cluster 1 is also an option for homebuyers who prefer to avoid bustle place but still wanted to have necessities at reach. However, avoid cluster 0 because of shortage on many things that only could be suitable for certain type of lifestyle that does not depends too much necessities.

## Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant homebuyers i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to buy a house. The findings of this project will help the relevant homebuyers on high potential locations for houses by their needs and lifestyle.