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 70 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.