Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Chain of Super Markets in Melbourne, Australia

~By: Nisha Devendra Kumar

Introduction

For many shoppers, visiting Super Markets is a great way to do grocery shopping and other daily essential needs. Super Markets are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the Super Markets provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more Super Markets to cater to the demand. As a result, there are many Super Markets in the city of Melbourne and many more are being built. Opening Super Markets allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new Super Market requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the Super Market is one of the most important decisions that will determine whether the Super Market 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 Melbourne, Australia to open a new Super Market. 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 Melbourne, Australia, if a property developer is looking to open a new Super Market, 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 Super Markets in the capital city of Australia i.e. Melbourne. This project is timely as the city is currently suffering from oversupply of Super Markets.

Data

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

- List of neighbourhoods in Melbourne. This defines the scope of this project which is confined to the city of Melbourne, the capital city of the country of Australia 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 Super Markets. We will use this data to perform clustering on the neighbourhoods.
- Please note that we have considered 50 Suburbs of Melbourne for our analysis for faster processing of API calls

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 Melbourne, 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 database 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 Super Market 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

Firstly, we need to get the list of neighbourhoods in the city of Melbourne. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Melbourne.

Next, we will use Foursquare API to get the top 50 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Super Market” data, we will filter the “Super Market” as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as Super Market as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Super Market”. The results will allow us to identify which neighbourhoods have higher concentration of Super Markets while which neighbourhoods have fewer number of Super Markets. Based on the occurrence of Super Markets in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Super Markets.