Coursera Capstone Project

IBM Applied Data Science Course

Opening a new Shopping Mall in Sydney

By: Akash M

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and breaks. 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 terminus for all types of shoppers. For vendors, the central location and the large crowd at the shopping malls provides a great delivery 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 world now and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. As with any commercial decision, opening a new shopping mall requires serious thought and is a lot more complicated than it seems. Mainly, the location of the shopping mall is one of the most important decision that will determine whether the mall will be a success or a failure.

Business Problem

The aim of this capstone project is to analyse and choose the best places in the city of Sydney to open a new shopping mall. Using data science methodology and machine learning methods like clustering, this project aims to provide solutions to answer the business question: If a property developer is looking to open a new shopping mall in the busy city of Sydney, where would you recommend that they open it?

Target Audience

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the city of Sydney.

Data

- List of neighbourhoods in Sydney. This defines the scope of this project which is confined to the city of Sydney, the state capital of New South Wales and the most populous city in Australia.

- 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.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_Sydney_suburbs)contains a list of neighbourhoods in city of Sydney. 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 or using csv file of the coordinates.

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 Shopping Mall category in order to help us to solve the business problem put forward. The project makes 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).

Methodology

Initially, we need to get the list of neighbourhoods in the city of Sydney. The list is available in the Wikipedia page(https://en.wikipedia.org/wiki/List_of_Sydney_suburbs). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. Next 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 Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude or any csv file available already having the coordinates. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2500 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 found out 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 “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighbourhoods.

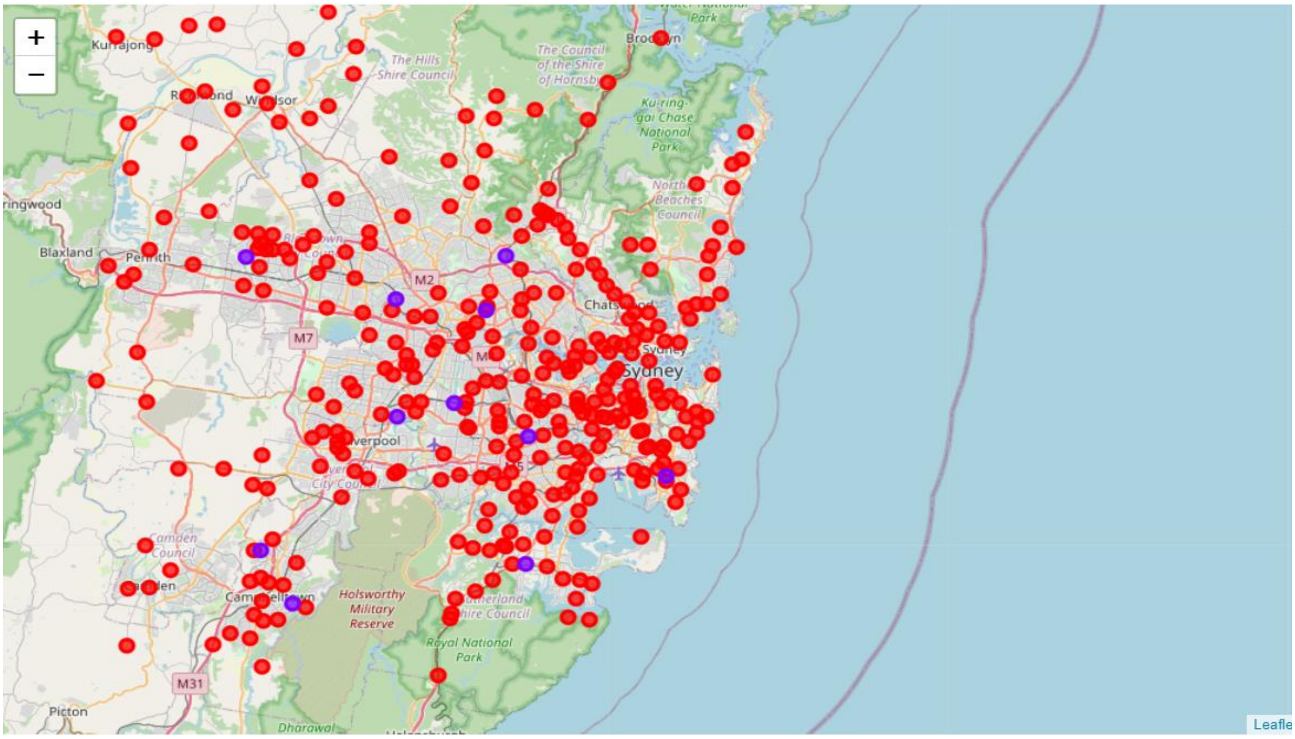
Lastly, we will perform clustering on the data by using k-means clustering. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 2 clusters based on the frequency of occurrence for "Shopping Mall":

- Cluster 0: Neighbourhoods with no shopping malls
- Cluster 1: Neighbourhoods with shopping malls

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



Observations

The shopping malls are all concentrated in the cluster 1. On the other hand, cluster 0 has no shopping mall in the neighbourhoods. This presents a great opportunity and high potential areas to open new shopping malls as there is no competition from existing malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 0 with no competition. From these clusters we can see that even though the shopping malls are all concentrated in cluster 1 the total number of them in the cluster still provides an opportunity for opening malls. Also, we see more areas in the neighbourhood of Sydney lack shopping malls as cluster 0 has more neighbourhoods which provides a great scope for the property developers.

Limitations

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could create a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

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 2 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders 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 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations in their decisions to open a new shopping mall.

References

Category: Suburbs in Sydney. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/List_of_Sydney_suburbs

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs