Coursera Capstone

IBM Applied Data Science Capstone

Developing a commercial building in Canberra, Australia

By: Vijaya Bhusan

August 2019

Introduction

For many business, having buildings at commercial place is a great way to increase and scale their businesses by networking clisents, events and business partnership. They can travel for deals & meetings without hassle, manage any problem occured, host clients, and increase market share. Commercial space are like a one-stop destination for all types of business. For businesses, the central location and ease of transport provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more office spaces for booming startup culture to cater to the demand. As a result, there are many commercial spaces in Canberra and many more are being built. Opening commercial spaces allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new commercial spaces requiring to analyze location, demand and services to be provided. Particularly, the location is most important decisions that will determine whether it will be lucrative to business.

Business Problem

The objective of this capstone project is to analyse and select the best locations in Canberra, Australia to open. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: for location in Canberra.

Target Audience of this project

This project is particularly useful to start-ups, businesses, distributors and investors looking to open or plan to acquire and rent building spaces in . This project is timely as the city is currently suffering right palce and buildings for rental property. Data suggest that there will be modest increase in property and rental prices in 2019 and occupancy rate is also falling due to continued obsession with building more shopping space despite chronic oversupply.

Data

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

- List of neighbourhoods in Canberra. This defines the scope of this project which is confined to suburbs of Canberra, Australia in Oceania.
- 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 building and office. 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_of_Canberra) contains a list of neighbourhoods in Canberra, with a total of 124 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 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

Firstly, we need to get the list of neighbourhoods in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Canberra). 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 on map of Canberra.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 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 “Building” data, we will filter the “Building” as venue category for the neighbourhoods.

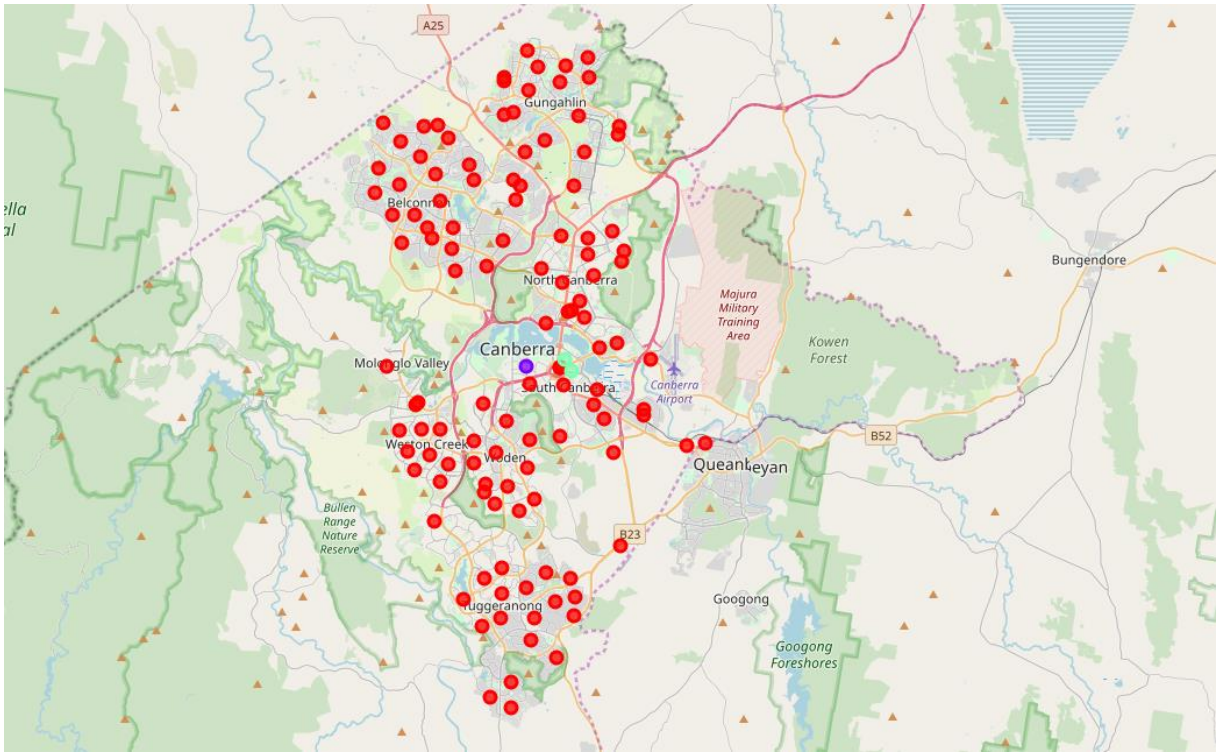
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 “Building”. The results will allow us to identify which neighbourhoods have higher concentration of building while which neighbourhoods have fewer number of building and offices. Based on the occurrence of building in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new building for office space.

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 “Building”:

- Cluster 0: Neighbourhoods with high concentration moderate number of official buildings.
- Cluster 1: Neighbourhoods with low number to no existence of buildings for office space.
- Cluster 2: Neighbourhoods with of moderate number of official buildings.

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 building are concentrated around middle area of Canberra, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 1 has very low number to no shopping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, buildings or office space in cluster 0 are likely suffering from intense competition due to oversupply and high concentration . From another perspective, the results also show that the oversupply of buildings mostly happened in the central area of the city, with the suburb area still have very few buildings naer to airport which provide excellent opportunity ti invest. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open in neighbourhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 which already have high concentration of shopping malls and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of buildings and offices, there are other factors such as population and income of residents that could influence the location decision of a commercial office space. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise 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 3 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 buildings. 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 open a office. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a rental space for businesses

References

Category:Suburbs in Kuala Lumpur. *Wikipedia*.

https://en.wikipedia.org/wiki/Category:Suburbs_of_Canberra

Foursquare Developers Documentation. *Foursquare*. Retrieved from

https://developer.foursquare.com/docs