IBM Data Science Capstone – Coursera

Where to open a new pizza place in Melbourne, Australia?

Introduction

Melbourne is one of the largest cities in Australia. It happens to be a sports and culture capital of the country as well. Melbourne is rated as one of the most livable and student friendly cities in the world. This attracts a lot of foreign investors planning to gain residency in the city by investing a large sum of money in it.

Pizza is a universally loved food item. So if a pizza place is opened and the operation is executed well, it can lead to a great amount of profit create a good return for the investor. So if a foreign investor wants to open a pizza place in Melbourne, in which Suburb should he/she open it? This report will attempt to answer this question by using data science methods.

Business problem

The aim of this project is to analyze and recommend the best suburbs/areas in Melbourne, Australia to open up a pizza place. This will be done utilizing data science methodology and machine learning techniques such as clustering. The question this report attempts to answer is: Where would you recommend a new investor to open a new pizza place in the city of Melbourne?

This project will be useful to… (target audience)

This project is useful for any investors who are willing to open a new pizza place in the city of Melbourne.

Data

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

- List of Melbourne suburbs.
- Lat/Longs of the Melbourne suburbs.
- Venue data. To find the concentration of pizza places in different suburbs.

Sources of data and methods to extract them

The Wikipedia page (https://en.wikipedia.org/wiki/List_of_Melbourne_suburbs) contains a list of 539 inner and outer city suburbs of Melbourne. Web scraping will be used to extract the data from the Wikipedia page. This will be done using the Pandas package. Then we will use Python Geocoder to get geographical coordinates (lat/longs) of each suburb. Then we will use Foursquare API to get the venue data for those suburbs.

Methodology

Before anything else, the first thing to do would be to find a list of Melbourne suburbs, which is available on Wikipedia. Using the web scraping of Python through Pandas package, all the names of suburbs are extracted. The list is then compared to Wikipedia to confirm all suburbs are correctly scraped.

Next step is to find out the geographical latitude/longitude coordinates to use the Foursquare API. Therefore, the Geocoder package is used to convert neighborhood address into geographical coordinates in the form of geographical coordinates. The collected data is then put into a pandas DataFrame. The DataFrame is then visualized using the Folium package so the suburbs are seen on the city map. This helps us in performing a check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted on the Melbourne map.

A list of top 50 venues that are within a radius of 5000 meters was captured using the FourSquare API.

Then, each suburb was analyzed by grouping the rows by suburb and taking the average/mean of the occurrence frequency of each venue category. So, we are also preparing the data for clustering. Since we are analyzing the data of “Pizza Places”, we will filter the “Pizza Places” as venue category for the suburbs.

Finally, we will perform clustering on data by using k-means clustering method, which is an unsupervised machine learning method. The suburbs will be clustered into 3 clusters based on their frequency of “Pizza Places”. The results will allow us to identify which suburbs have higher concentration of Pizza places while which suburbs have fewer number of Pizza places. Based on the occurrence of Pizza places in different suburbs, it will help us to answer the question as to which suburbs are best to start new a pizza place.

Results

The results from the k-means clustering showing that we can categorize the Suburbs into 3 clusters based on the frequency of occurrence for “Pizza Places”:

- Cluster 0: Suburbs with low concentration of pizza places; around 46%
- Cluster 1: Suburbs with high concentration of pizza places; around 19%
- Cluster 2: Suburbs with medium pizza places; around 35%

Discussion

Analysis showing that most of the pizza places are concentrated in cluster 1 which includes some popular areas including Brunswick, where there is the Lygon street, famous for Italian restaurants (around 19%). Cluster 0 have around 46% concentration and cluster 2 have around 35%.

Analysis shows that most pizza places are clustered around mid-city suburbs. This is just where the commercial areas end and the residential areas begin. One exception to that is Brunswick, where there is the famous Lygon street full of Italian restaurants and pizza places.

Conclusion

In this project, all steps in a data science analysis is applied like 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. investors regarding the best locations to open a new pizza places.

If a person were to open a pizza place in Melbourne, according to me the place should be somewhere in cluster 1, where there is a high concentration on pizza places. A high concentration of pizza places suggests that there is a high demand for pizza in those aread. However, a high concentration also means high competition. So one must open a pizza place where there is a high demand and then differentiate the product offering in a way that is able to beat the competition and stand out. So, I would suggest that a person should open a pizza place on Lygon street in Brunswick with a highly differentiated and competitive product offering.