IBM Data Science Capstone

Opening a bar in Sydney, Australia

Introduction

For many residents, visiting a local bar is always a good way to relax and communicate after going through a high-intense working day. They can talk freely with friends, gather together watching football matches, and also the various parties held by different groups is also unmissable. The bars not only offer a place for people to get drinks but also offer a place to get people connected. As a result, bar culture is of the essence of Australia culture. And opening a nice bar at the right place considerable income. But when comes to the business decision it is more complicated than it seems like. And one of the determiner is its geolocation.

Business Problem

The objective of this project is to analyse the best location to open a new bar in Sydney, Australia. With the help of data science methodology, machine learning methods like clustering and foursquare API. This project aims to give a business solution to the question: Where is appropriate to open a new bar in Sydney, Australia?

Target Audience

This project is aimed to provide suggestions for the investors who is looking to open a new bar in Sydney, Australia. And this is report is timely as it is subjected to the fast-expanding service industry and the influence of COVID-19.

Data

To get the problem solved, the following data is required.

1. Regions of Sydney. This help defines the scope of the project.
2. Latitude and longitude coordinates of those regions. These data are required to plot the map and also get the venue data.
3. Venue data related to the bars. We will use these data to clustering the regions and also calculate the frequency of the bars shown in the venue data.

Methodology

1. First we need to scrap data from wiki page into a data frame, here we use the request function and then parse data from the html(https://en.wikipedia.org/wiki/Category:Regions_of_Sydney) into a beautifulsoup object.
2. We use the unique function to get rid of the duplicated values.
3. Import Geocoder function to get the geographical coordinates in the form of latitude and longitude in order to use Foursquare API.
4. Create temporary data frame to populate the coordinates into Latitude and Longitude and then merge the coordinates into the original data frame
5. Then we visualize the data using Folium package
6. Next we use the Foursquare API to explore the region. We want to top 100 venues within a radius of 3000 meters, here we need to make API calls in a loop. Then we convert the venues list we have into a new Data frame.
7. We use the groupby function to count the returned venues for each region.
8. Lastly, we will perform clustering the data by using k-means clustering. Here we cluster the region into 3 clusters dependent on the occurrence of the bar. The results will help us define which region has higher concentration of bars and in which region there could be more bars being opened.

Results:

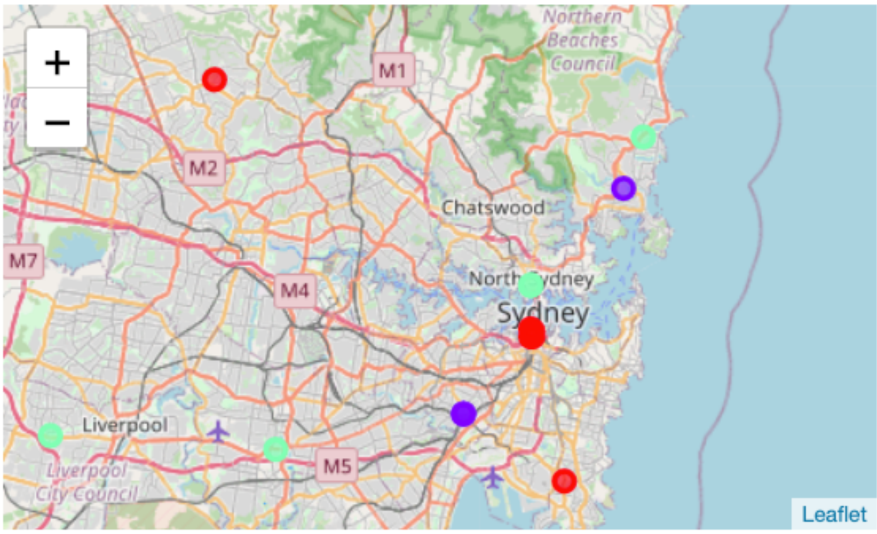
The results from the k-mean clustering showing the we can categorize the regions into 3 clusters.

Cluster 0: regions with low concentration of bars.

Cluster 1: regions with high concentration of bars.

Cluster 3: regions with moderate concentration of bars.

Red: Cluster 0   Purple: Cluster 1.   Green: Cluster 2



Discussion:

As observed from the map the north-east and south-west part of the Sydney has the highest concentration of the bars, which may lead to the oversupply situation and suffering from intense competition. While the central area of Sydney still have a great potential to open some new bars, because their concentration rate are still low. And also the concentration of the bars in the suburb area is moderate. Therefore this project recommends investors to open new bars in the area of cluster 0, which areLong Reef (New South Wales) St George (Sydney) Eastern Suburbs (Sydney) Forest District (Sydney) South-Eastern Sydney. Hills District Macarthur, New South Wales Sydney c entral business district

Conclusion:

In this project, we have been through the process of identifying the business problems and figuring out the potential audience, then we come to the data scraping, data cleaning, data visualization and also machine learning clustering data into three clusters based on their similarities. And we figure out regions in cluster 0 are the most idea place to open the new bars. But here I also have some parts to declare. Here we only consider the occurrence of the bars, there are also many factors to influence the decision, like the rent ,the population density ,the residents age groups ,and etc. Further researches need to be done before make any decisions.