Coursera Capstone Project:

Introduction

Opening up a restaurant is an entrepreneur’s dream or passion. To share the joy with patrons who end up relishing the ambience, food and services is fulfilment of its own. However, only the owner knows what all efforts they have put in before opening up a restaurant. Developing and executing a solid business strategy for the restaurant is extremely important in order to make the business successful. So We will use our data analytics powers to generate a few most promising neighborhoods based on this criteria.Advantages of each area will then be clearly expressed so that best possible final locations can be chosen by stakeholders.

In this Project, we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening a **Chinese restaurant** in **Bangalore**, India.

Since there are lot of restaurants in Bangalore, we will try to detect **locations that are not already crowded with Chinese restaurants**. We are also particularly interested in **areas with number of Chinese restaurants in vicinity**. We would also prefer locations **as close to city center as possible**, assuming that first two conditions are met.

Problem Statement

The main objective of this capstone project is to analyze and select the best locations in the city of Bangalore, India to open a new Chinese Restaurant. 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 Bangalore, India, if a property developer were looking to open a new Chinese restaurant, where would he/she be recommended to open his new venture precisely.

Target Audience of this project

This project is particularly useful to investors looking to open or invest in new Chinese Restaurants in the Bangalore city, India

Data

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

- List of neighbourhoods in Bangalore. This defines the scope of this project which is confined to the city of Bangalore, India.
- 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 Chinese Restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore) contains a list of neighbourhoods in Bangalore, with a total of 58 neighbourhoods. I will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful-soup 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 Chinese Restaurants 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).

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Bangalore. Fortunately, the list is available in the page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore). I will do web scraping using Python requests and beautiful-soup packages to extract the list of neighbourhoods data. However, this is just a list of names. I 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 Bangalore.

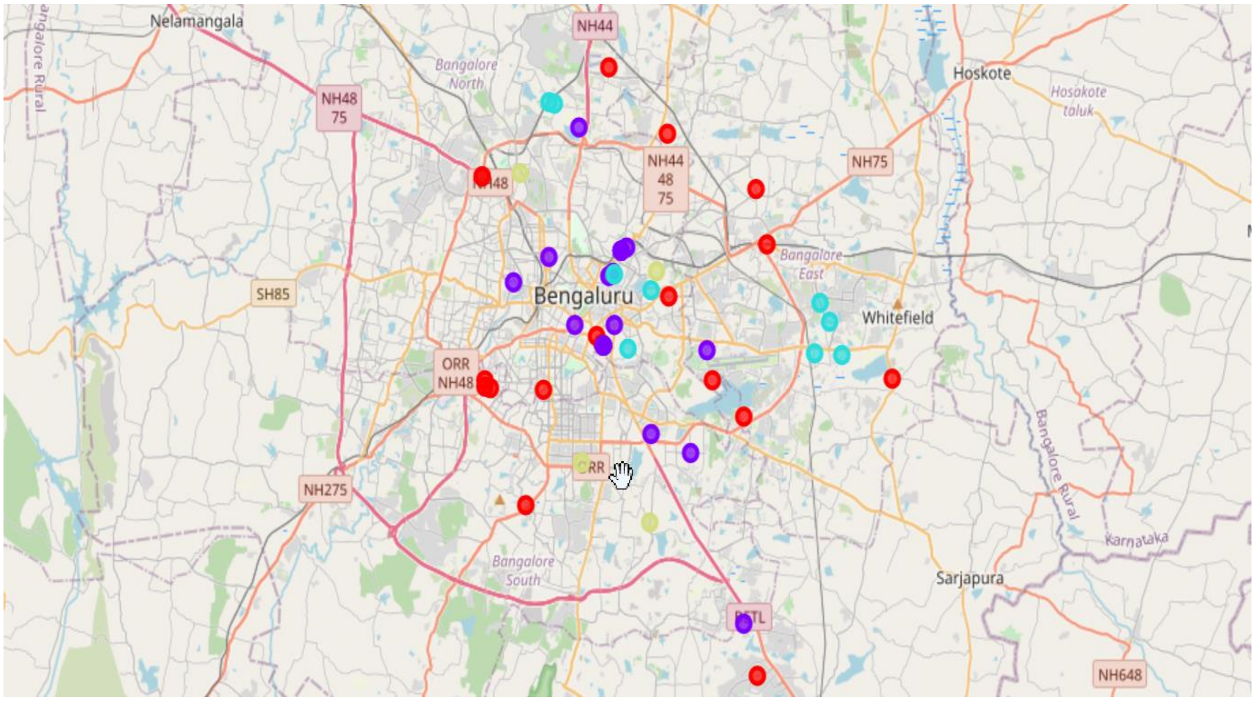
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods 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 "Chinese Restaurants" data, we will filter the "Chinese Restaurants" 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 "4" clusters based on their frequency of occurrence for "Chinese Restaurants".

The results will allow us to identify which neighbourhoods have higher concentration of Chinese restaurants while which neighbourhoods have fewer number of Chinese Restaurants. Based on the occurrence of Chinese Restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Chinese Restaurants.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 4 clusters based on the frequency of occurrence for "Chinese Restaurants":



- Cluster 0:Neighbourhoods with low number to no existence of Chinese Restaurants
- Cluster 1:Neighbourhoods with almost equal concentration of Chinese Restaurants
- Cluster 2:Neighbourhoods with moderate concentration Chinese Restaurants
- Cluster 3:Neighbourhoods with high concentration of Chinese Restaurants

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in mint green colour,and cluster 3 whitish yellow.

Discussion

As observations noted from the map in the Results section, most of the Chinese Restaurants are concentrated in the central area of Bangalore city, with the highest number in cluster 4 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to no Chinese restaurants in the neighbourhoods. This represents a great opportunity and high potential areas to open new Chinese restaurants as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends investors to capitalize on these findings to open new Chinese Restaurants in neighbourhoods in cluster 0 with little to no competition. Restaurants Investors with unique selling propositions to stand out from the competition can also open new Chinese Restaurants in neighbourhoods in cluster 0 with moderate competition. Lastly, Restaurants Investors are advised to avoid neighbourhoods in cluster 3 which already have high concentration of Chinese Restaurants 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 Chinese Restaurants, there are other factors such as population and income of residents that could influence the location decision of a new Chinese Restaurants. 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 Chinese Restaurant.

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.