Coursera Capstone

IBM Applied Data Science Capstone

Opening a New BAKERY shop in Coimbatore, India

By: Anand Shanmugam

July 2020

Introduction

Coimbatore is the second largest city in the Indian state of Tamilnadu and is known for variety of bakeries and pastries. You will find bakeries in each and every corner of the city. I am passionate to open a bakery chain. The success of the chain depends on the location and competitors. If I choose a location that already has lot of bakeries, then my chain is likely to fail.

Business Problem

The objective of this capstone project is to analyse and select the best locations in Coimbatore to open a new bakery. 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 Coimbatore, India, if one has to open a bakery shop , where would you recommend to start?

Data description

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

- **List of neighbourhoods** in Coimbatore. This defines the scope of this project which is confined to the city of Coimbatore, second largest city of Tamilnadu, 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 **Bakeries**. We will use this data to perform clustering on the neighbourhoods.

Methodology:

Data Source:

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Coimbatore) contains a list of neighbourhoods in Coimbatore, with a total of 36 neighbourhoods.

Data extraction:

We will use **web scraping** techniques to extract the data from the Wikipedia page, with the help of Python requests and **beautifulsoup** packages. We will wrangle the data, clean it, and then read it into a *pandas* dataframe so that it is in a structured format.

Latitude and Longitude:

Next, we will use the Geocoder package to get the latitude and longitude of the neighborhoods. After gathering the data, we will populate the data into a pandas Data Frame 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 Coimbatore.

Venue Data:

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We will use Foursquare ID and Foursquare secret key obtained by registering in Foursquare. 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 venue categories are available. Then, we will analyse by taking mean of the frequency of occurrence of each venue category per neighborhood. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “BAKERY” data, we will filter the “BAKERY” as venue category for the neighbourhoods.

Clustering

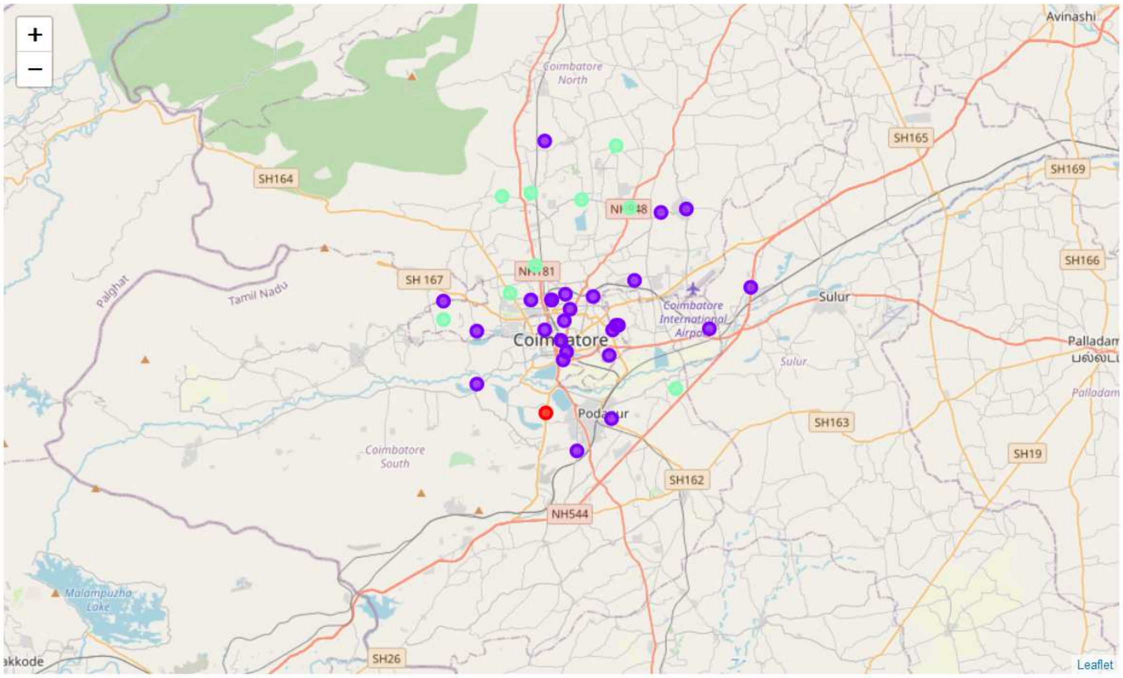
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods has more and less Bakeries. Based on the occurrence of Bakeries , it will help us to answer the question as to which neighbourhoods are most suitable to open new Bakeries.

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 “Bakery”:

- Cluster 0: Neighbourhoods with zero to less number of bakeries
- Cluster 1: Neighbourhoods with High number of bakeries.
- Cluster 2: Neighbourhoods with Medium concentration of bakeries .

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 observed from the above map, most of the bakeries are found in the city centre and in the outskirts , with the highest number in cluster 1 and medium in cluster 2.cluster 0 has very low number of bakeries in the neighbourhoods. This represents a great opportunity to open new bakeries in cluster 0 as there is very little to no competition from existing bakeries.

Meanwhile, cluster 2 already facing competition from other bakeries and may not be likely location for a new bakery. Therefore, this project recommends opening new bakery in cluster 0, where there is little to no competition.

Recommendations for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of bakeries, there are other factors such as population , Purchasing power of people, Taste of the local etc that could influence our decision.

However, to the best of my knowledge such data are not available. 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 Bakery. 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 defining a business problem, specifying the data source, data wrangling and preparation, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e **suggesting the best locations to open a new Bakery**. 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 Bakery.**