Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Coffee Shop in New Delhi, India

By: Harsh Sharma

October, 2019

This Photo by Unknown Author is licensed under CC BY

Introduction

For many people either it be working professionals or just hanging out buddies a coffee shop provides a soothing place to hang-out or to chill along with your friends. Thus a vast majority of people belonging to age groups 18-35 enjoys coffee breaks from their monotonous and hectic schedule to go for a coffee break. Hence the coffee shops have become an integral part of their lives.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of New Delhi, India to open a new Coffee Shop. 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 Delhi, India, if a property developer is looking to open a new Coffee Shop, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to shopkeepers who want to introduce new ventures such as coffee shops under their trade names. Thus due to vast area of this city a normal businessman or a person willing to open a coffee shop would rely on the other's opinion regarding the best place to open a shop, nut it would be time consuming and cumbersome. Thus This project would help them to identify the best places to set up their shops

Data

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

- List of headquarters in the delhi city. This defines the scope of this project which is confined to the city of New Delhi, the capital city of the country of India in South East Asia.
- Latitude and longitude coordinates of those headquarters. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to coffee shops. We will use this data to perform clustering on the headquarters.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_districts_of_Delhi) contains a list of headquaters in Delhi, with a total of 11 headquaters. 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 headquaters using Python Geocoder package which will give us the latitude and longitude coordinates of the headquaters.

After that, we will use Foursquare API to get the venue data for those headquaters. 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 headquaters in the city of New Delhi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_districts_of_Delhi)

We will do web scraping using Python requests and beautifulsoup packages to extract the list of headquaters 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 headquaters 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 New delhi.

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 headquaters 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 “Coffee shop” data, we will filter the “coffee shop” as venue category for the headquaters.

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 headquaters into 3 clusters based on their frequency of occurrence for “coffee shop”. The results will allow us to identify which headquaters have higher concentration of Coffee shops while which headquaters have fewer number of coffee shops.

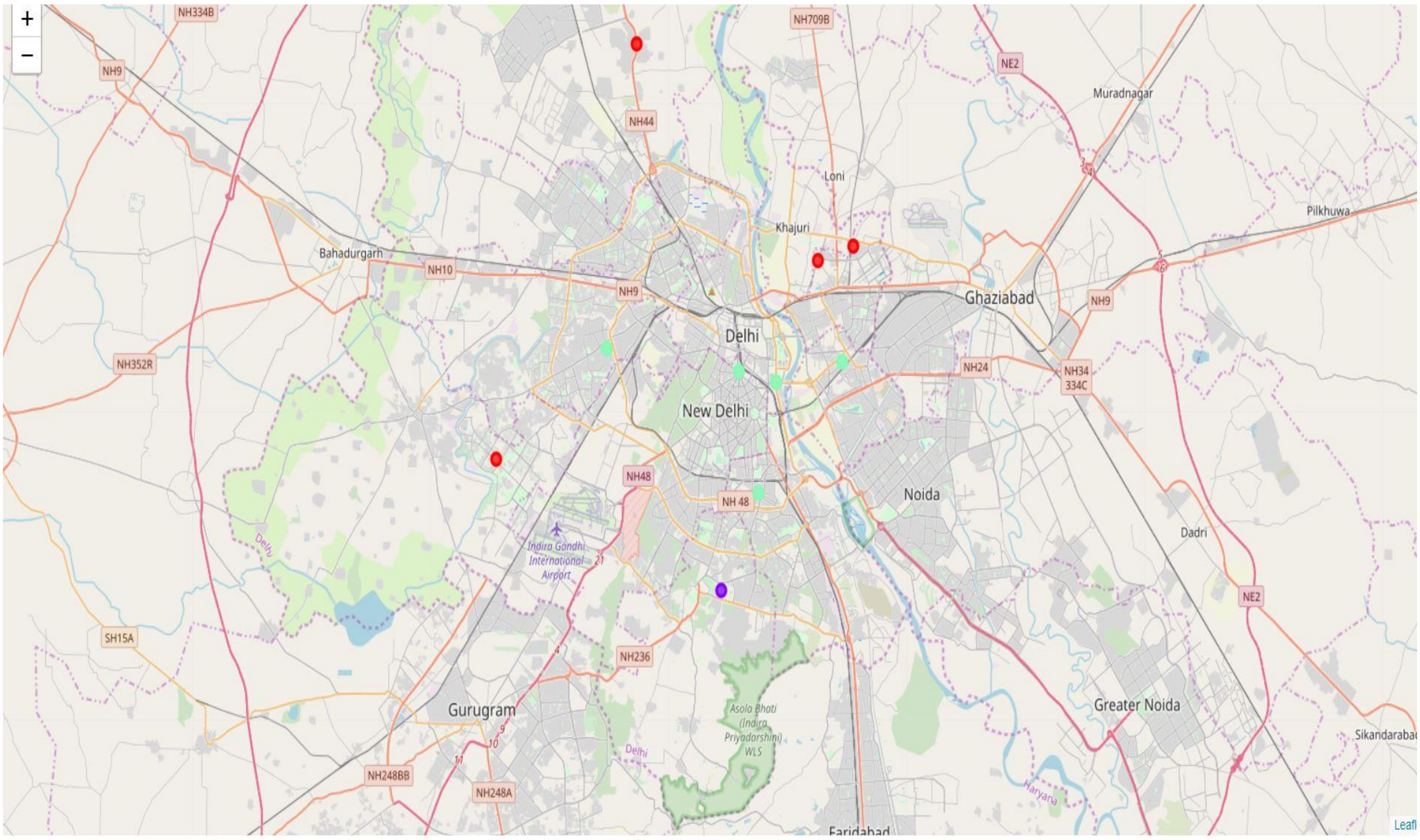
Based on the occurrence of Coffee shops in different headquaters, it will help us to answer the question as to which headquaters are most suitable to open new Coffee shops.

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 "Coffee Shops":

- Cluster 0: Headquaters with moderate number of coffee shops
- Cluster 1: Headquaters with low number to no existence of Coffee Shops
- Cluster 2: Headquaters with high concentration of Coffee Shops

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.



Discussions

Its is seen that only cluster 2 have highest no. of coffee shops as compared to other two, hence it is clear that if anyone wants to open a coffee shop than it is definitely the cluster region 2 than cluster region 0 than 1.

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 coffee shop. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The headquaters in cluster 1 are the most preferred locations to open a new coffee shops. 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 new coffee shops.

References

1. Category: Suburbs in New Delhi. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_districts_of_Delhi
2. Foursquare Developers Documentation. Foursquare. Retrieved from https://foursquare.com/developers

Appendix

Cluster 0

- **Alipur**
- **Dwarka**
- **Nand nagri**
- **shadra**

Cluster 1

- **Saket**

Cluster 2

- **Connaught place**
- **Dariyaganj**
- **Defence Colony**
- **Preet vihar**
- **Rajouri garden**