Opening an Indian Restaurant in Atlanta, Georgia

IBM Data Science Capstone Project

RITU JINDAL

March 2021

INTRODUCTION

Atlanta is home to the largest Indian population among metropolitan areas. The 10-county Atlanta region is home to approximately 42 percent of Georgia's total population but is home to 77 percent of the state's Asian-Indian population.

For this Capstone Project, I am using the hypothetical scenario for opening an Indian Restaurant in Atlanta, Georgia. As Indian cuisine is quite popular with Americans and Indian alike, it might present a good opportunity for an Indian already living in Atlanta.

Ambience, menu, hygiene and of course taste are all important factors to be kept in mind before getting into the Hospitality Industry, but these are all problems that can be tackled internally by the person(s) in charge. The location of a restaurant is also of utmost importance regardless of the history of a business or the taste of the food.

PROBLEM STATEMENT

The objective is to find a suitable location(s) to open an Indian Restaurant in Atlanta, GA. This project makes use of various Data Science and Machine Learning methodologies (k-means Clustering) to provide a Solution to the client. The project aims to provide a Solution to the Question: 'Where should we consider opening an Indian Restaurant in Atlanta?'

Data

I have used the following Data for the completion of the project:

- List of top suburbs in Atlanta - This gives the coordinates of all the neighborhoods and is used to call the Foursquare API.
- List of Venues in Atlanta- This contains data about all the nearby venues like Restaurants, Bars, Gym etc.
- Demographics of Indians in Atlanta – This is important to understand the distribution of the target audience in Atlanta.

Data Sources

- List of Atlanta suburbs Data from (https://metroatlantasuburbs.com/communities)
- Nearby Venues Data created using Foursquare API (https://api.foursquare.com/v2/venues/)
- The Demographics Data is scraped from (http://zipatlas.com/us/ga/city-comparison/percentage-indian- population.htm)
- Latitude and Longitude values are obtained using the Geocoder package in python.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Atlanta,Georgia which is available in (https://metroatlantasuburbs.com/communities). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data.We also 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 neighborhoods in a map using Folium package. 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. 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood 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 "Restaurant" data, we will filter the "Indian Restaurant" as venue category for the neighborhoods. 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 neighborhoods into 3 clusters based on their frequency of occurrence for “Indian Restaurant”. The results will allow us to identify which neighborhoods have higher concentration of Indian Restaurant while which neighborhoods have fewer number of Indian Restaurant. Based on the occurrence of Indian restaurant in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Indian Restaurant.

Results

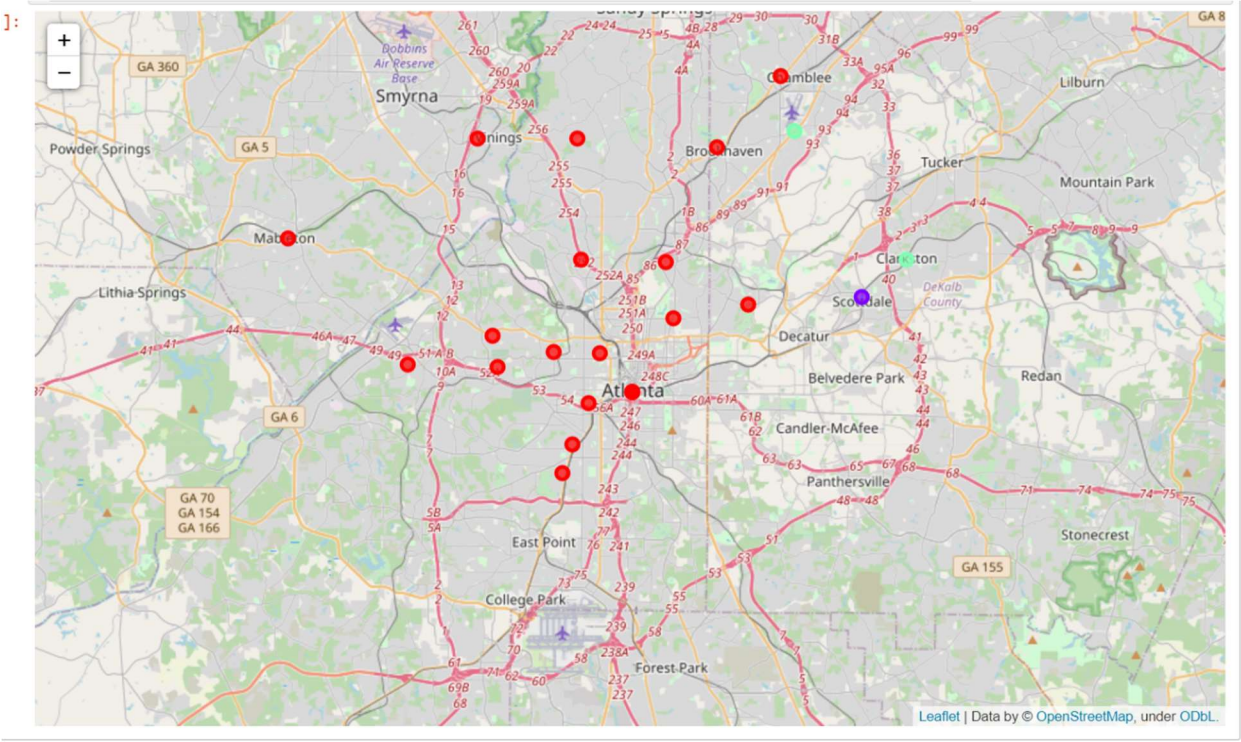
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Indian Restaurant”:

Cluster 0: places having no Indian Restaurant

Cluster 1: places having moderate no. of Indian Restaurant

Cluster 2: places having maximum no. of Indian Restaurant

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 observations noted from the map in the Results section, most of the Indian Restaurants are at places with cluster 2 and low number in cluster 1. On the other hand, cluster 0 has no Indian Restaurant in the neighborhoods. This represents a great opportunity and high potential areas to open an Indian Restaurant as there is very little to almost no competition. Therefore, this project recommends property developers to capitalize on these findings to open an Indian Restaurant in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the

competition can also open new shopping malls in neighborhoods in cluster 0 with low competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of Indian Restaurant and suffering from intense competition.

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 an Indian Restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open an Indian Restaurant.