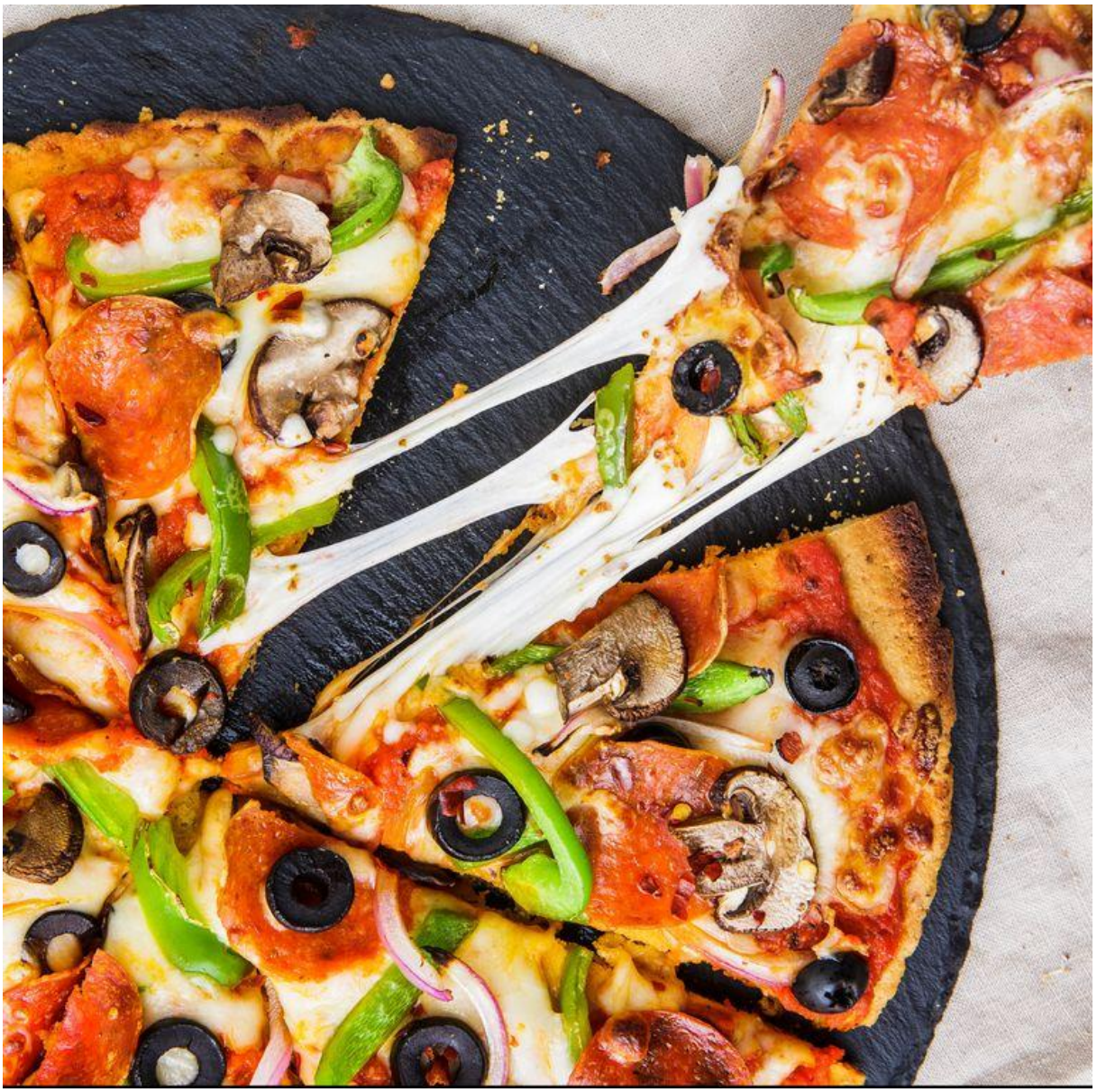
Coursera IBM Data Science Capstone Project Report



Report by :
Shitij Roy Chaudhary

Introduction

For this Capstone project, I have created a hypothetical scenario of a Italian restaurateur who wants to open an authentic Italian restaurant in Delhi. The idea behind this project is that there may not be enough Italian restaurants in Delhi and it might present a great opportunity for this entrepreneur.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Italian restaurant in Delhi, India. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Delhi, if an entrepreneur wants to open a Italian restaurant, where should they consider opening it?

Target Audience

The entrepreneur who wants to find a location to open an authentic Italian restaurant.

Data

To solve this problem, I will need below data:

- List of districts in Delhi, India.
- Latitude and Longitude of these districts.
- Venue data related to food, shop & service and travel & transport. This will help us find districts that are most suitable to open a Italian restaurant.

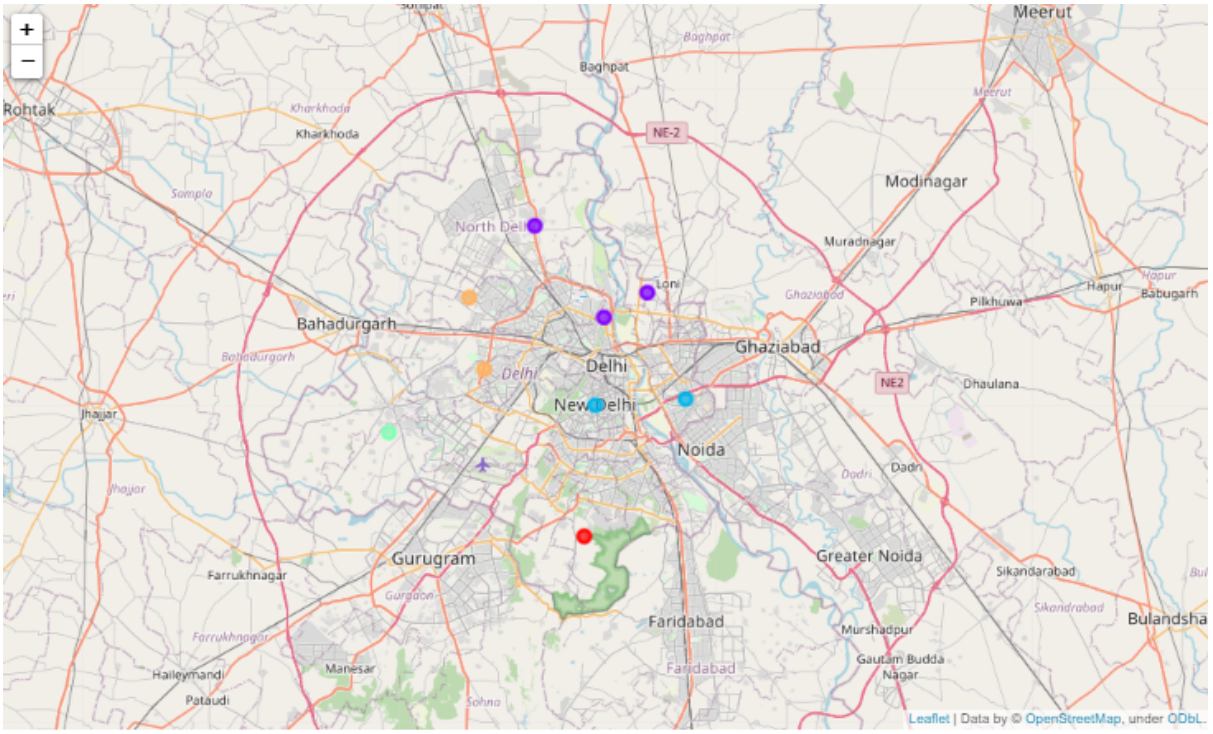
Extracting the Data

- Scrapping of Delhi districts via Wikipedia
- Getting Latitude and Longitude data of these districts via Geocoder package.
- Using Foursquare API to get venue data related to these districts.

Methodology

First, I need to get the list of districts of Delhi, India. This is possible by extracting the list of districts from the wikipedia page (“https://en.wikipedia.org/wiki/List_of_districts_in_India”). I did web scraping by utilizing pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into a dataframe. To get the coordinates, I used Geocoder. After gathering all these coordinates, I visualized the map of Delhi using the Folium package to verify whether these are correct coordinates.
Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain an account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each district by grouping the rows by district and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for "Food, shop & service and travel & transport". Lastly, I performed the clustering method by using k-means clustering. K means clustering algorithm identifies k number of centeriods, 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 it is highly suited for this project as well. I have clustered the districts in Delhi into 5 clusters based on their frequency of occurrence. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Result



The above image shows 5 clusters :

Cluster 0 (Purple) : shows a very high number of Italian restaurants.
Cluster 1 (Blue) : shows a high number of Italian restaurants.
Cluster 2 (Green) : shows low number of Italian restaurants.
Cluster 3 (Orange) : shows 0 to low number of Italian restaurants.
Cluster 4 (Red) : shows a medium number of Italian restaurants.

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 the machine learning by utilizing k-means clustering and providing recommendations to the stakeholder.

References

https://en.wikipedia.org/wiki/List_of_districts_in_India
https://developer.foursquare.com/docs