Coursera Capstone

Famous Food Places in New Delhi, India

By

Megha Dua

1.Introduction

1.1 Background

There is so much more to Delhi than what meets the eye and that is certainly true for the food it serves! One can find snack serving eateries or hawkers in almost every street here and this is probably the reason why foodies love this city so much. But what about those who are new in their culinary journey or who simply want to explore all the delicacies the city has to offer.

Delhi is composed of number of food places which attracts foodies from the world. There are many venues (like cafes, restaurants, etc.) which can be explored. Delhi is composed of number of food places which attracts foodies from the world. This project involves data from both the Foursquare API and the Zomato API to fetch complete information of various places. Further, a map of the venues will be plotted to highlight their location, and information about these places. Such plots assimilate plentiful data as their shaded portraits and area on the guide. This empowers any foodie to take a quick look and chose what spot to visit.

1.2 Audience that can be interested

The target audience for such a project is of two types.

1. Any person who is visiting Delhi, can use the plots and maps from this project to quickly select places that suit their budget and user's ratings.
2. A person who is interested in exploring different parts of Delhi and taste different delicacies offered like the food bloggers and foodies.

This information can be useful for any entrepreneur to open a new hotel, restaurant, café etc, by identifying the potential areas of growth of their business using these plots.

2.Data

2.1 Data Sources

To get location and other information about various food places in Delhi, I used two APIs and decided to combine the data from both of them.

Using the Foursquare's explore API (which gives places recommendation), I fetched places to range of 15 kms from the center of Delhi and collected their names, categories and locations (longitudes and latitudes).

Using the name, latitude and longitude values obtained from the Foursquare API, I used the Zomato search API to fetch venues from its database. This API allows to find venues based on search criteria (usually the name), latitude and longitude values and more. Given that the data from the two APIs did not align completely, I had to use data cleaning to combine the two datasets properly.

From Foursquare API (https://foursquare.com/developers/api), I retrieved the following information for each venue:

- Name: The name of the venue.

- Category: The category type as defined by the API.

- Latitude: The latitude value of the venue.

- Longitude: The longitude value of the venue.

From Zomato API (https://developers.zomato.com/api), I retrieved the following information for each venue:

- Name: The name of the venue.

- Address: The complete address of the venue.

- Rating: The ratings as provided by many users.

- Price range: The price ranges the venue belongs to as defined by Zomato.

- Price for two: The average cost for two people dining at the place. I later convert it to the average price per person by dividing it by 2.

- Latitude: The latitude value of the venue.

- Longitude: The longitude value of the venue.

2.2 Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/Special:Search?search=famous+food+places+in+Delhi&go=Go&ns0=1) is a list of locations in New Delhi, with a total of 5473 localities. 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 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 localities. 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 cuisines data, we are particularly interested in the Famous food places category in order to help us to solve the 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.

2.3 Methodology

Firstly, we need to get the list of localities in the city of New Delhi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Special:Search?search=famous+food+places+in+Delhi&go=Go&ns0=1).We will do web scraping using Python requests and beautifulsoup packages to extract the list of localities 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 localities 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 localities 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 cuisines were returned for each locality and examine how many unique categories can be curated from all the returned cuisines. Then, we will analyze each locality by grouping the rows by locality and taking the mean of the

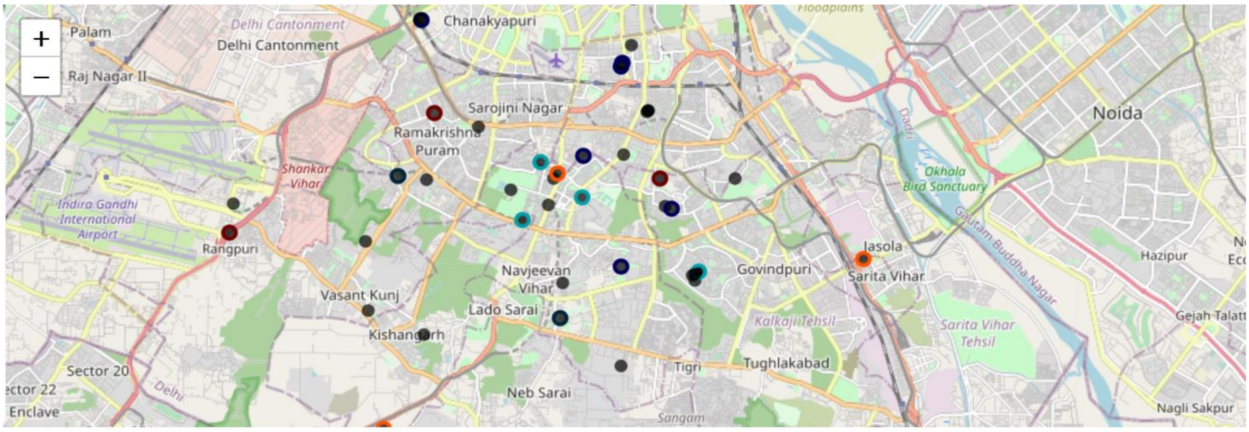
frequency of occurrence of each cuisine category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Famous Food Places” data, we will filter the “Food Places” as cuisines category for the localities.
Lastly, we will perform clustering on the data by using 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 localities into clusters based on their frequency of occurrence for “Famous Food Places”. The results will allow us to identify which localities have higher concentration of food places while which localities have fewer number of food places.

2.4 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 “Famous Food Places”:

- Cluster 0: Localities with moderate number of Cuisines.
- Cluster 1: Localities with low number to no existence of Cuisines.
- Cluster 2: Localities with high concentration of Cuisines.

The results of the clustering are visualized in the map below with cluster 0 , cluster 1 and cluster 2 .



3.Discussion

As observations noted from the map in the Results section, most of the food places are concentrated in the central area of New Delhi city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no shopping mall in the localities.

4. Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of famous food places, there are other factors such as cuisines and average cost that could influence the attraction for the foodies to explore. However, to the best knowledge of this researcher such data are not available to the locality 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. 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.

5.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 for the foodies based on different cuisines and Aggregate ratings of the cuisines. The Locality in cluster 2 is the most preferred locations for the people to explore different cuisines.