IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

SETTING UP A NEW VEGAN RESTAURANT IN BENGALURU, INDIA

By

HRISHIKESH BHANDARKAR

Introduction

The restaurants have been a most profitable now a days. To mention specifically, the metropolitan city like Bengaluru always makes a top place when it comes to breakfast items like Idly, Dosa etc. A Bengalurian is always aware of the crowd at restaurants for morning coffee at 6. So, a Vegetarian restaurant makes a good sound about profitable business.

Business Problem

A person is willing to open a new Vegetarian Restaurant in Bengaluru city and is in a confusion on which is the right place to invest so it would be a profitable business for him. Now, when we speak about the good place to start, it is known that the city is already having many such restaurants and in that locality the estimation of profit would not be as expected. So, the best way to resolve this issue is by looking for a location/neighborhood that has less similar cuisines.

Target Audience:

Since the selection of a profitable place is given the uttermost importance while starting a business, in this project the main focus is in helping people to find a location where a vegetarian restaurant could gain business.

Data Description

To tackle the above-mentioned problem, we need to have the dataset that contains

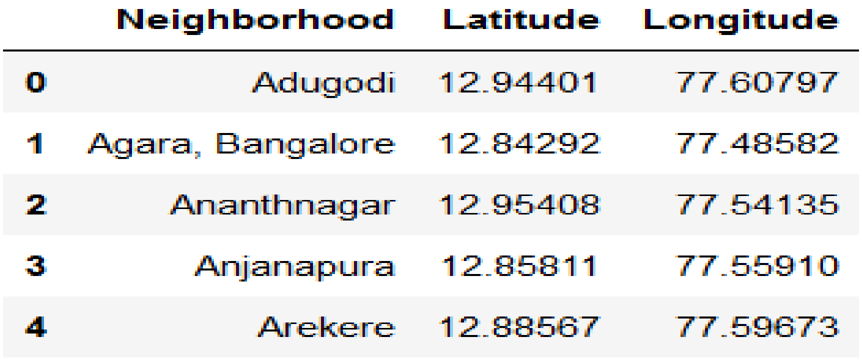
i. All the neighbourhoods of Bengaluru city.
ii. Latitude and longitudes of all the neighbourhoods in city.

The page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore is the major source of data that is being used to obtain the neighbourhoods of Bengaluru. We then use beautifulsoup4 package, a Python module that helps to scrape information from the web pages to extract all the tables from this Wikipedia page and convert it into a pandas data frame.

Then the data is cleaned by removing the unwanted cells and then we use Python's geopy package to obtain the latitude and longitude of all the neighbourhoods present in the data frame. The obtained coordinates are then merged with the main data frame with list comprehension operation.

Once the coordinates are obtained, we use Foursquare API to extract the venues using client credentials along with the version, radius and limit values and to cluster them based on preferences we use kmeans clustering. We also use folium map to visualize the clusters geographically.

The following is the data frame sample used for analysing the locations geographically:



	Neighborhood	Latitude	Longitude
0	Adugodi	12.94401	77.60797
1	Agara, Bangalore	12.84292	77.48582
2	Ananthnagar	12.95408	77.54135
3	Anjanapura	12.85811	77.55910
4	Arekere	12.88567	77.59673

Methodology

Data Exploration-

Firstly, we need to get the list of neighbourhoods in Bengaluru. For this the data is available at wiki page from where we have to extract and clean. The page link is (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Bangalore). Now we need to use Web scraping to extract the data using BeautifulSoup module. And then convert the extracted data to pandas dataframe and then the data frame is cleaned by removing unnecessary data. However, this data frame contains only neighbourhood information.

Data Geocoding-

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 the address into geographical coordinates in form of latitude and longitude. After gathering the data, we will populate the data into a pandas Data Frame.

Data Visualization-

Visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Bengaluru.

Finding top 100 venues exploration-

Next, we make use of Foursquare API to get the top 100 venues within 2000 meters radius. For this we would require the client credentials such as ID and secret key from our Foursquare developer account. We then make the API calls to Foursquare passing in the geographical coordinates of neighborhoods in a loop of Python program. 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.

Data Wrangling-

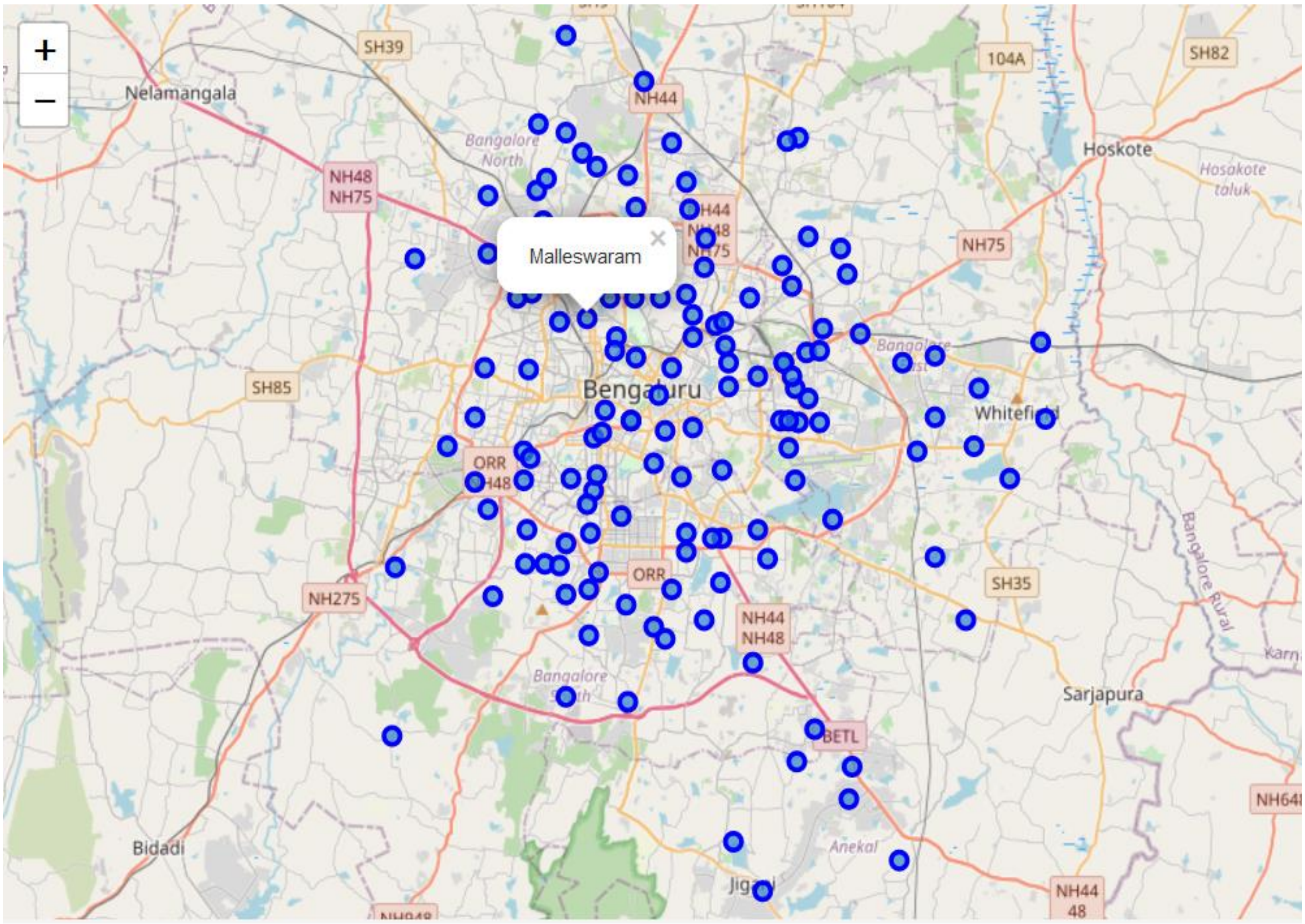
We are also preparing the data for use in selection. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open a new restaurant.

Data Clustering-

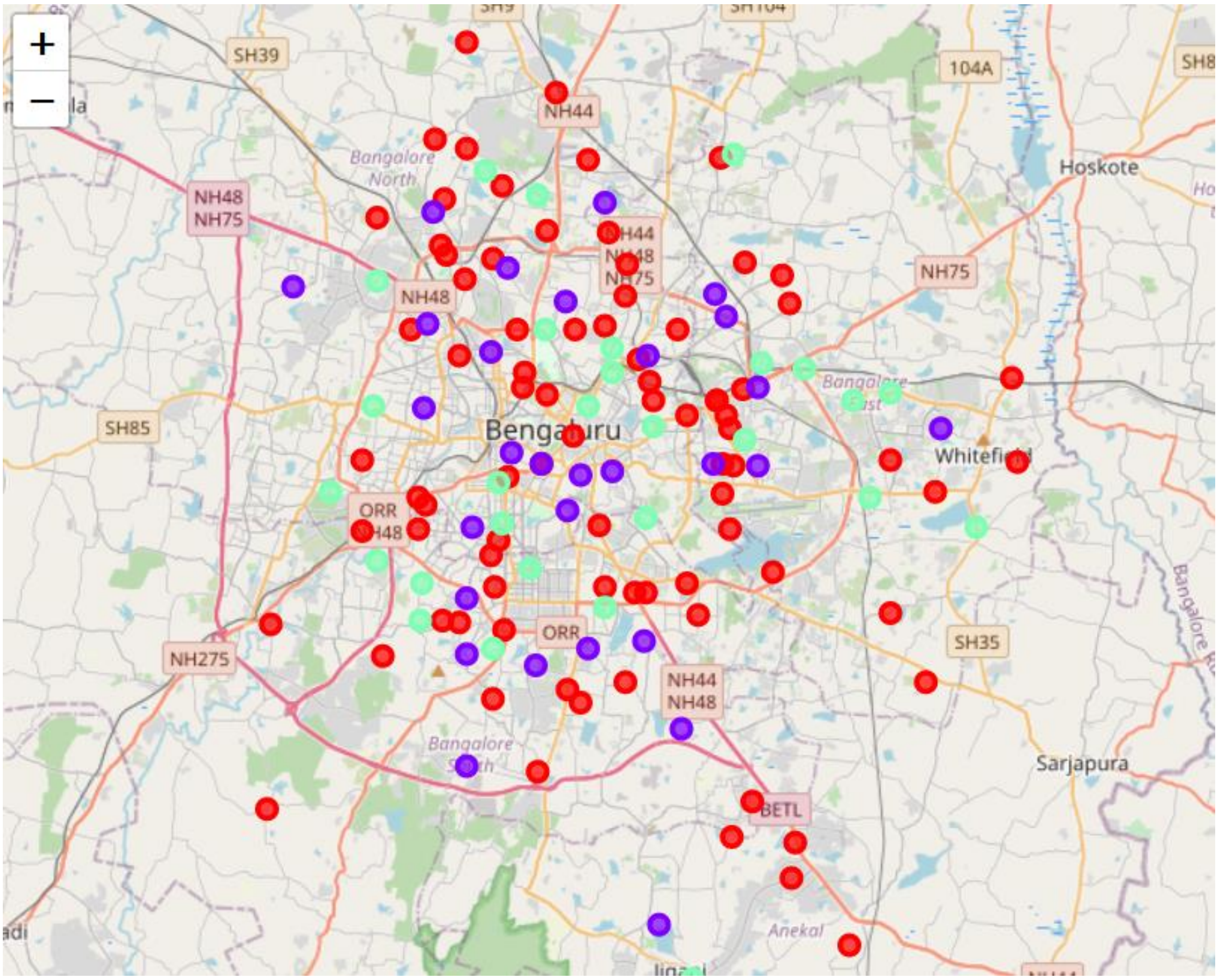
Finally, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised 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 "Number of existing vegetarian restaurants". The results will allow us to identify which neighborhoods have higher, medium and lower concentration of restaurants. Based on the occurrence of hotels in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new restaurant.

Result

1. Bengaluru neighborhoods on map



2. The map after the clustering into 3 groups.



The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Vegetarian Restaurant":

- Cluster 0: Neighbourhoods with a moderate concentration of Veg. Restaurants
- Cluster 1: Neighbourhoods with a very less or no concentration of Veg. Restaurants
- Cluster 2: Neighbourhoods with a high concentration of Veg. Restaurants

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 observation noted from the map of clusters, it can be seen that the concentration of veg restaurants are dense in the areas come under cluster 2 which means the competition is very high and for a new opening restaurant trying to get the customers, the chances would be much lesser and that would lead to decline / loss in business. Now, we have two other clusters to look up. So, say we are going with the cluster 1 with very less or no restaurants at all. Here, we might have to rethink before suggesting to open the restaurant because, the very less means there is a chance that it would be hard to find customers in those regions. So, this might result in failure as well. So, when we look for a location to open up a restaurant, it should have a little competition as it helps in keeping up the standards time to time. The customers will also come eventually to explore the cuisine and that could help in getting the popularity. So, when we look at such a condition, cluster 0 matches with our requirement.

Limitations and Future Suggestions:

In this project, we only consider one factor i.e., frequency of occurrence of veg restaurant, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher, such data are not available to the neighbourhood 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 to open a new restaurant. 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 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., investors regarding the best locations to open a new Vegetarian Restaurant. 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 vegetarian restaurant. The findings of this project will help the relevant people to invest on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.