Analyzing the Neighborhoods in Bengaluru for Starting a Restaurant

Applied Data Science Capstone Project

By: Jeevan Joyce

Table of Contents

Introduction

Bengaluru, often referred to as the Silicon Valley of India, is renowned for its vibrant tech industry and cosmopolitan culture. Located in the southern part of India, it is one of the fastest-growing cities in the country and attracts a significant number of tourists and professionals from around the world each year. Personally, I have developed a deep admiration for Bengaluru, appreciating its blend of modernity and tradition. The city is a major hub for technology and innovation, and its diverse population includes people from various ethnicities and cultures. This multicultural environment has led to a rich culinary scene, offering a plethora of cuisines from all over the globe. The people of India generally have a deep love for food, and I am no exception, with a passion for exploring different cuisines and flavors.

Therefore, the aim of this project is to study the neighborhoods in Bengaluru to determine possible locations for starting a restaurant. This project can be highly valuable for business owners and entrepreneurs who are considering investing in and opening a restaurant in Bengaluru. The main objective of this project is to carefully analyze relevant data and provide recommendations for stakeholders. By leveraging this data-

driven approach, we can identify the most promising areas for new dining establishments, ensuring their success in the vibrant and competitive market of Bengaluru.

The system aims to provide comprehensive recommendations by:

1. Identifying the types of restaurants in a specific area.
2. Locating similar restaurants based on food preferences.
3. Ranking different restaurants according to personal preferences.

The target audience includes everyone, not just frequent travelers. People may seek similar restaurants due to a preference for specific cuisines, while others might want highly-rated restaurants nearby. Thus, this project caters to anyone exploring new or familiar dining options.

As the restaurant scene evolves with new food categories and hybrid dishes, a system that accesses a wide variety of foods becomes indispensable. It's impractical for people to rely on word-of-mouth for recommendations. However, computers excel at remembering and utilizing data. With machine learning at its peak, this technology can act as a personal guide, enhancing success rates over time by aligning with individual likes and dislikes. This project, therefore, promises to be a valuable personal assistant for food enthusiasts.

Problem

Bengaluru's culinary diversity mirrors its social and economic variety, with roadside vendors, tea stalls, South Indian, North Indian, Muslim food, Chinese, and Western fast food all being popular. Udupi restaurants, known for their vegetarian fare, are particularly favored. The city's Chinese and Thai food can be tailored to suit Indian tastes, making Bengaluru a foodie's paradise with a unique blend of local tradition and vast culinary options.

Frequent travel and constant relocation can be exhausting, especially when navigating unfamiliar environments. Food becomes a crucial factor in rating trips and making recommendations. Good food can draw people from around the world. Thus, finding the right place, at a reasonable cost, is essential for an optimal dining experience. Several questions arise:

1. How many types of foods are available in the restaurant?
2. Which nearby restaurant has the best ratings?
3. How many similar restaurants are in the vicinity?
4. Are similar restaurants more expensive? If so, what makes them special?

To address these questions, I've been tasked with a project to develop a system that not only answers these questions but also recommends new places based on previous experiences.

Data Collection

The following data is required for the project:

1) Neighborhood data of Bangalore
2) Geographical coordinates of Bengaluru and its various neighborhoods.
3) Venue data for neighborhoods in Bengaluru

Neighborhoods Data

The data for the neighborhoods in Bengaluru was scraped from this Kaggle article. The data is read into a pandas DataFrame using the read_html() method. The primary reason for this approach is that the Wikipedia page offers a comprehensive and detailed table of data, which can be easily scraped using the read_html() method of pandas. The top 13 rows of the DataFrame are displayed in Figure 1.

Adugodi	560 030	12.9716	77.5946
Agaram	560 007	12.8431	77.4863
Air Force Stn. Yelahanka	560 063	13.1048	77.5763
Banashankari	560 050	12.925453	77.546761
Banashankari 2nd Stage	560 070	12.9249	77.5662
Banashankari 3rd Stage	560 085	12.9271	77.5548
Bangalore City H.O.	560 002	12.972442	77.580643
Bangalore G.P.O.	560 001	12.972442	77.580643
Bangalore University	560 056	12.9462	77.5103
Bannerghatta	560 083	12.9426	77.6027
Bannerghatta Road	560 076	12.9426	77.6027
Bansawadi	560 043	13.0108	77.6493
Basavangudi	560 004	12.9422	77.5748

Figure 1: Top 13 rows of Bengaluru neighborhoods data scraped from Kaggle.

Geographical Coordinates

The geographical coordinates for Bengaluru have been obtained using the GeoPy library in Python. This data is essential for plotting a map of

Bengaluru using the Folium library in Python. The code to obtain the geographical coordinates of Bengaluru is as follows:

```python
from geopy.geocoders import Nominatim # type: ignore

# Initialize the geolocator
geolocator = Nominatim(user_agent="geoapiExercises")

# Get the location of Bengaluru
location = geolocator.geocode("Bengaluru")

# Extract the latitude and longitude
latitude = location.latitude
longitude = location.longitude

print(f"The geographical coordinates of Bengaluru are: Latitude = {latitude}, Longitude = {longitude}")
```

Figure 2: Obtaining geographical coordinates of Mumbai.

The Geocoder library in Python has been utilized to obtain latitude and longitude data for various neighborhoods in Bengaluru. These coordinates are used to verify the accuracy of the coordinates provided on Kaggle, and they are replaced in our DataFrame if the absolute difference exceeds 0.001. These refined coordinates are then used to plot neighborhoods using the Folium library in Python. The result then displays the coordinates of neighborhoods in Bengaluru from Kaggle as ‘Latitude’ and ‘Longitude’, and those obtained from Geocoder as ‘Latitude1’ and ‘Longitude1’. It also shows the absolute differences between the two sets of latitude and longitude as ‘Latdiff’ and ‘Longdiff’, respectively.

Venue Data

The venue data has been extracted using the Foursquare API. This data includes venue recommendations for all neighborhoods in Bengaluru and is used to analyze the popular venues in different neighborhoods, as well as to build an unsupervised learning model to cluster these neighborhoods. Venue recommendations for all neighborhoods were obtained with a limit of 200, meaning a maximum of 200 venue recommendations per neighborhood, within a 1 km radius around each neighborhood's geographical coordinates.

	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cantonment area	Central	12.972442	77.580643	Hotel Fishland	12.975569	77.578592	Seafood Restaurant
1	Cantonment area	Central	12.972442	77.580643	Sapna Book House	12.976355	77.578461	Bookstore
2	Cantonment area	Central	12.972442	77.580643	Vasudev Adigas	12.973707	77.579257	Indian Restaurant
3	Cantonment area	Central	12.972442	77.580643	Adigas Hotel	12.973554	77.579161	Restaurant
4	Cantonment area	Central	12.972442	77.580643	Kamat Yatrinivas	12.975985	77.578125	Indian Restaurant

Figure 3: Data obtained from Foursquare API after cleaning.

Methodology

This section provides details for the methodology used in the project.

Data Visualization

In order to understand the data obtained for Bengaluru neighborhoods, basic visualization was carried out. Figure 4 shows a bar plot depicting the number ofvenues in each respective neighborhood.

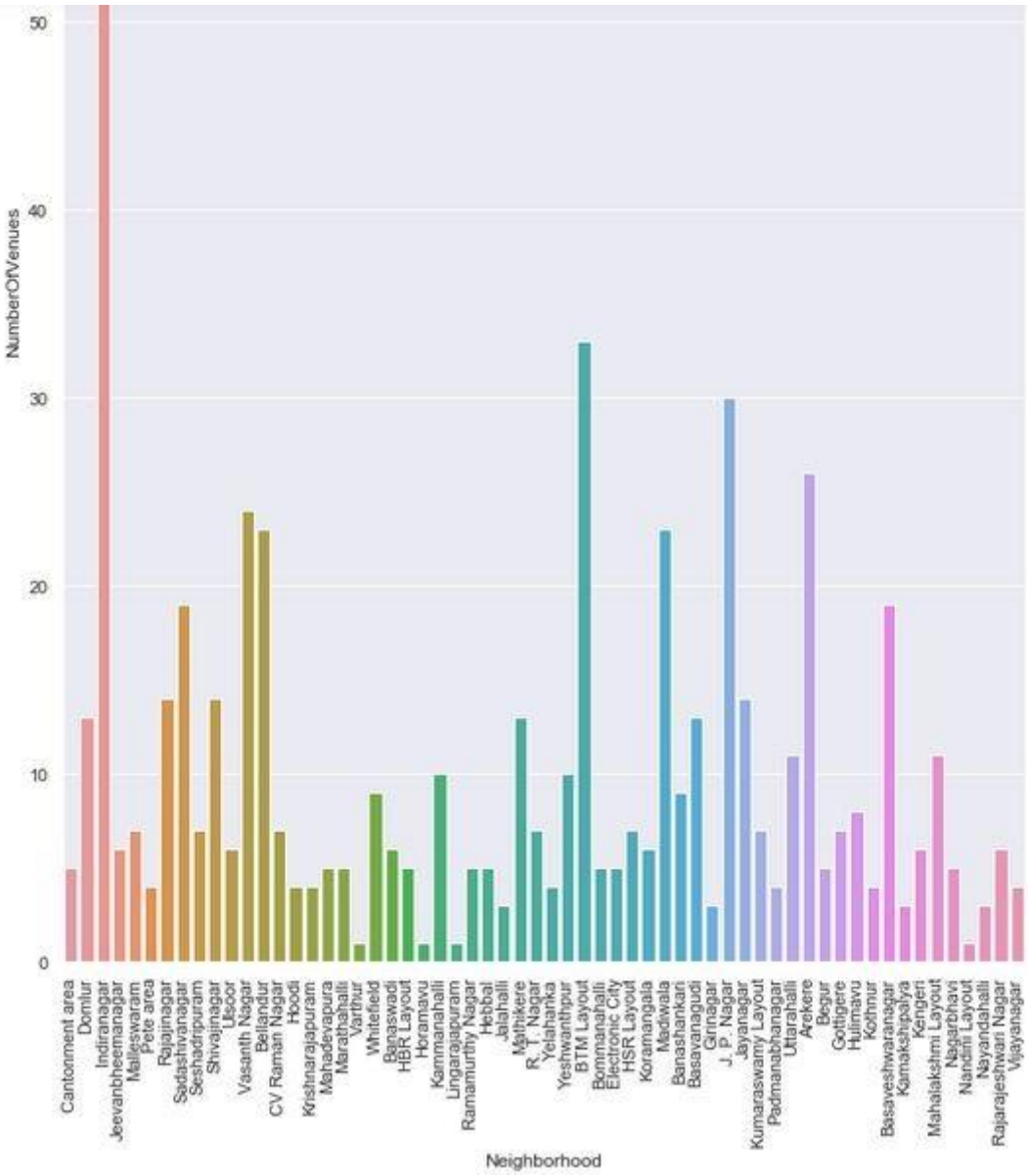

Figure 4: Number of venues in each neighborhood.

Further, we created a map depicting the various neighborhoods of Bengaluru used in this report, using Folium.

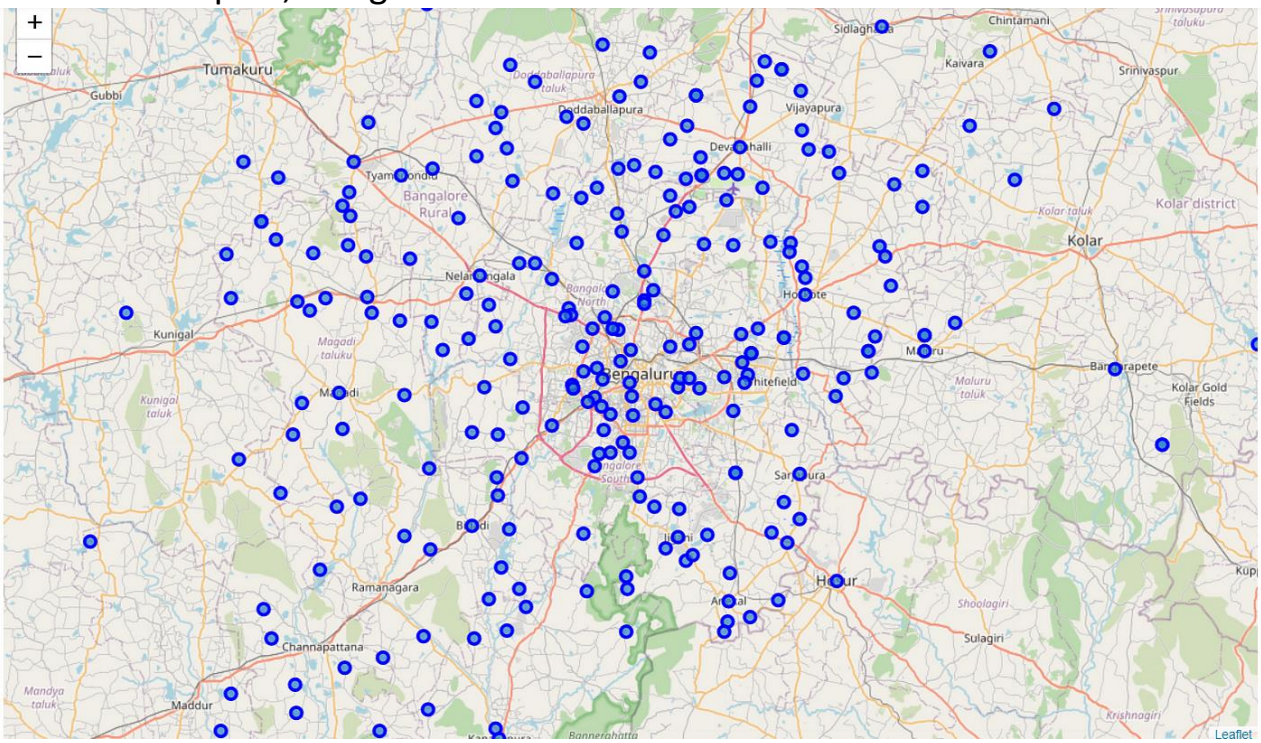

Figure 5: Depicting the neighborhood spread across Bengaluru.

Feature Extraction

Feature extraction was carried out to obtain features from the Foursquare API data (as shown in Figure 3) which was used for building the unsupervised learning model. In order to achieve this, the "Venue Category" column had to be converted to some form of numeric value to be used for building the model. This was

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	...	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Co
0	Achitnagar	13.091176	77.483482	0	Indian Restaurant	Bakery	Asian Restaurant	Falafel Restaurant	Food Court	Food & Drink Shop	...	Fast Food Restaurant	Farmers Market	Farm	Elec
1	Adugodi	12.942847	77.610416	3	Indian Restaurant	Café	Dessert Shop	Chinese Restaurant	Ice Cream Shop	Bookstore	...	Multiplex	Bar	Bakery	B
2	Amruthahalli	13.066513	77.596624	3	Indian Restaurant	Ice Cream Shop	Café	Department Store	Flea Market	Dhaba	...	Chinese Restaurant	Building	Bubble Tea Shop	Rest
3	Bagalgunte	13.056649	77.504822	0	Pizza Place	Indian Restaurant	Hobby Shop	Gas Station	Yoga Studio	Eastern European Restaurant	...	Farmers Market	Farm	Falafel Restaurant	Elec
4	Bagalur S.O (Bangalore)	13.133187	77.668709	3	Food Truck	Sports Club	Memorial Site	Yoga Studio	Food & Drink Shop	Food	...	Farmers Market	Farm	Falafel Restaurant	Elec

achieved by the One-hot Encoding method which takes all the unique categories and creates a column for each category. Then, if a neighborhood venue belongs to that category, it would get a value of 1 for that row in that specific category column and if a neighborhood venue does not belong to the particular category, the value would be 0. This process was repeated for all venues in all neighborhoods and the result was a sparse matrix containing the neighborhood name and all unique category columns with either 1 or 0 based on whether the neighborhood venue belonged to that category or not. This dataframe was then grouped by the neighborhood name and the average value was taken for all categories. The result is shown in Figure 6 which shows only the top 4 rows.

Feature extraction was performed on the Foursquare API data (as shown in Figure 3) to build the unsupervised learning model. To accomplish this, the “Venue Category” column had to be converted into numeric values suitable for the model. This was done using the One-Hot Encoding method, which takes all the unique categories and creates a column for each one. If a neighborhood

venue belongs to a category, it receives a value of 1 in that category column for that row; if it does not, it receives a 0. This process was repeated for all venues in all neighborhoods, resulting in a sparse matrix containing the neighborhood names and all unique category columns with either 1 or 0 based on the presence of the category.

The DataFrame was then grouped by neighborhood name, and the average value was calculated for all categories. Figure 6 shows the top 4 rows of this result. Note that most of the values are 0, as there were many unique categories, and not all neighborhoods had venues in every category. This data was used for the unsupervised learning model, with the neighborhood names removed. The unsupervised learning model is explained in the next section.

Unsupervised Learning

K-means unsupervised learning technique was used to cluster the neighborhoods based on the category of venues near the neighborhoods. One important aspect of the k-means model is to determine the number of clusters to use in model development. This was determined by the Silhouette score which was calculated for a range of clusters from 2 to 16. The resulting number of clusters and their respective Silhouette scores are shown in Figure 7.

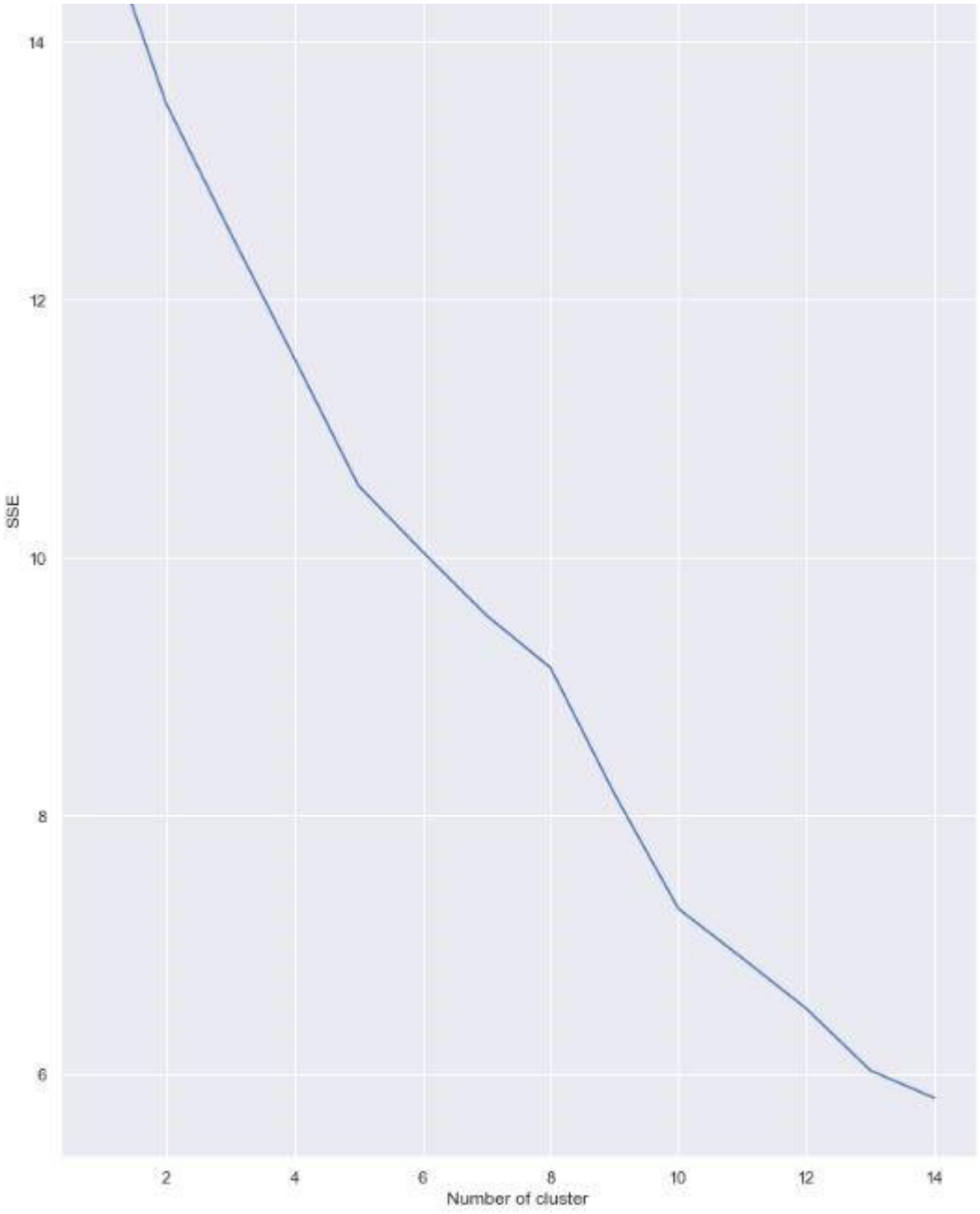

Figure 7: Silhouette scores for different number of clusters.

It is evident that the Silhouette scores are not very high even as the number of clusters increases. This means that the inter-cluster distance is not very high over the range of k-values. Despite this, the data will be clustered to the best possible extent. For this, 5 clusters will be used for the k-means clustering

model since itprovides the highest silhouette score as seen in Figure 7.

Results

The recommender system generates a list of top restaurants and popular venue types that users can enjoy. During runtime, a simulation was conducted using ‘Whitefield’ as the neighborhood input to the model. The system then recommended neighborhoods with similar characteristics to ‘Whitefield’.

The resulting image illustrates this process:

Out[211]:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Basaveshwaranagar	Venue Category_Ice Cream Shop	Venue Category_Indian Restaurant	Venue Category_Fast Food Restaurant	[0.6426377807870477]
1	Begur	Venue Category_Indian Restaurant	Venue Category_Indian Sweet Shop	Venue Category_Food Court	[0.7361321887351776]
2	Electronic City	Venue Category_Outlet Store	Venue Category_Furniture / Home Store	Venue Category_Bus Stop	[0.5423513638809381]

Figure 8

Given the nonlinear relationship between income and population, it's essential to employ an inferential approach to understand relationships among different sets of features. During clustering, neighborhoods with similar attributes should be grouped into appropriate clusters.

The graph below depicts the clusters:

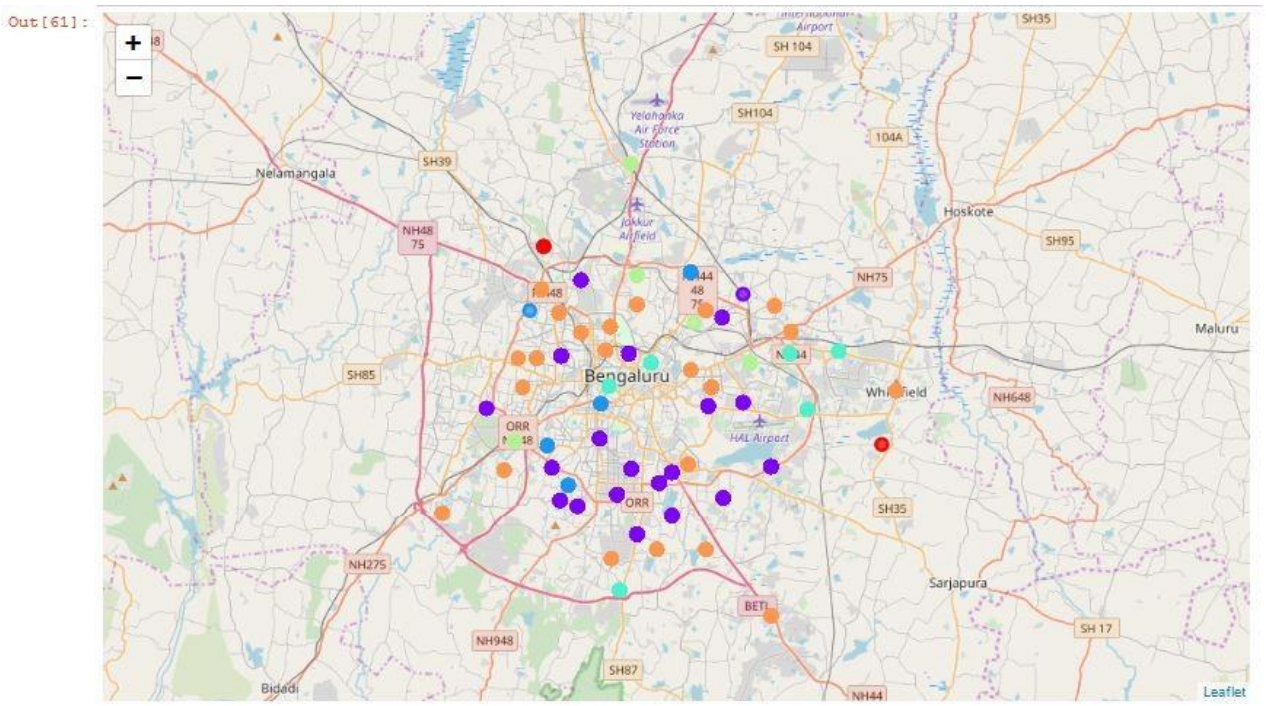

Figure 9 - Final Result of Clustering in Bengaluru

An important observation is that the choice of the number of clusters can significantly impact the results. Some clusters may be overly specific (overfitted), while others may be too general (underfitted). Therefore, careful analysis and evaluation of the number of clusters are crucial for optimal clustering results.

Discussion

Upon analyzing the clusters obtained for Bengaluru, it is evident that certain clusters are more conducive to hosting restaurants and hotels, while others are less favorable. From the result we can see that some clusters are less suitable for opening new restaurants.

Conversely, the other clusters exhibit a higher prevalence of restaurants, hotels, multiplexes, cafes, bars, and other food joints. Therefore, neighborhoods within these clusters are more promising for launching a new restaurant venture.

When comparing these clusters, neighborhoods in some appear particularly suitable for restaurant establishments due to a larger percentage of food joints among their top 10 common venues compared to the others. The neighborhoods in these clusters boast a diverse array of dining options including Japanese, Indian, Chinese, Italian, and seafood restaurants, alongside cafes, bakeries, steakhouses, and pubs.

While the other suitable clusters also feature Indian restaurants as the most common venue, further analysis reveals a mix of other venues such as soccer fields, flea markets, smoke shops, gyms, train stations, dance studios, music stores, and cosmetics shops. Thus, it is recommended to consider opening the new restaurant in neighborhoods belonging to the other clusters.

Conclusion

In this project, the neighborhoods in Bengaluru, India have been effectively analyzed to identify optimal locations for opening a new restaurant.

According to the analysis, neighborhoods in cluster 1 have been recommended as ideal locations. This recommendation has been visualized on the map. Stakeholders and investors can refine this recommendation by taking into account additional factors such as transportation accessibility, legal regulations, and associated costs. These factors, while important, were beyond the scope of this project and were not considered in the current analysis.