IBM – Coursera

Data Science Specialization

Capstone project - Final report

The Battle of Neighbourhoods – Geolocating ideal venues for Indian Restaurants in San Diego

Divyaprakash Dhurandhar

Table of Contents

1. Introduction

Think of India, and one of the first things that come to mind is its diversity. A large country, its population is second only to China. Its languages are numerous, and every state (of which there are 28 and seven Union territories) is unique in its traditions and, very importantly, its food. Food from one region may be alien to a person from another province! The common thread that runs through most Indian food is the use of numerous spices to create flavor and aroma.

Indian restaurants have come a long way from the mid-1960s when the first significant wave of immigrants arrived. While still not entirely as assimilated as Italian and Mexican, Indian food is one of the fastest-growing segments in the culinary scene and is gaining popularity within the American mainstream. And as Indians have spread throughout America, so has their food. Restaurants with cult-like followings are no longer limited to New York City.

2. Business Problem

Concerning the growing demand for Indian food and restaurants across the United States, the need for intelligent business solutions regarding opening Indian restaurants arises. For this project, we assume that a client in the city of San Diego wants to open an Indian restaurant. To maximize revenue and business success, insights into existing Indian restaurants and peer competition are required. Thus, the project's main objective is to find ideal spots in the city where Indian restaurants can be set up.

3. Data

This project's data has been retrieved and processed through multiple sources, giving careful considerations to the methods' accuracy.

3.1. Neighborhoods

The data of the neighborhoods in San Diego can be extracted out by web scraping using the BeautifulSoup library for Python. The neighborhood data is scraped from a Wikipedia webpage https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Diego.

3.2. Geocoding

The file contents from San-Diego.csv is retrieved into a Pandas DataFrame. The latitude and longitude of the neighborhoods are retrieved using the geocoder API. The geometric location values are then stored in the initial DataFrame.

3.3. Venue Data

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another DataFrame to contain all the venue details along with the respective neighborhoods.

4. Methodology

A thorough analysis of the principles of methods, rules, and postulates have been made to ensure the inferences to be as accurate as possible.

4.1. Folium

Folium builds on the Python ecosystem's data wrangling strengths and the mapping strengths of the leaflet.js library. All cluster visualization is done with Folium's help, which generates a Leaflet map made using OpenStreetMap technology. Creating a map of San Diego with neighborhoods superimposed on top

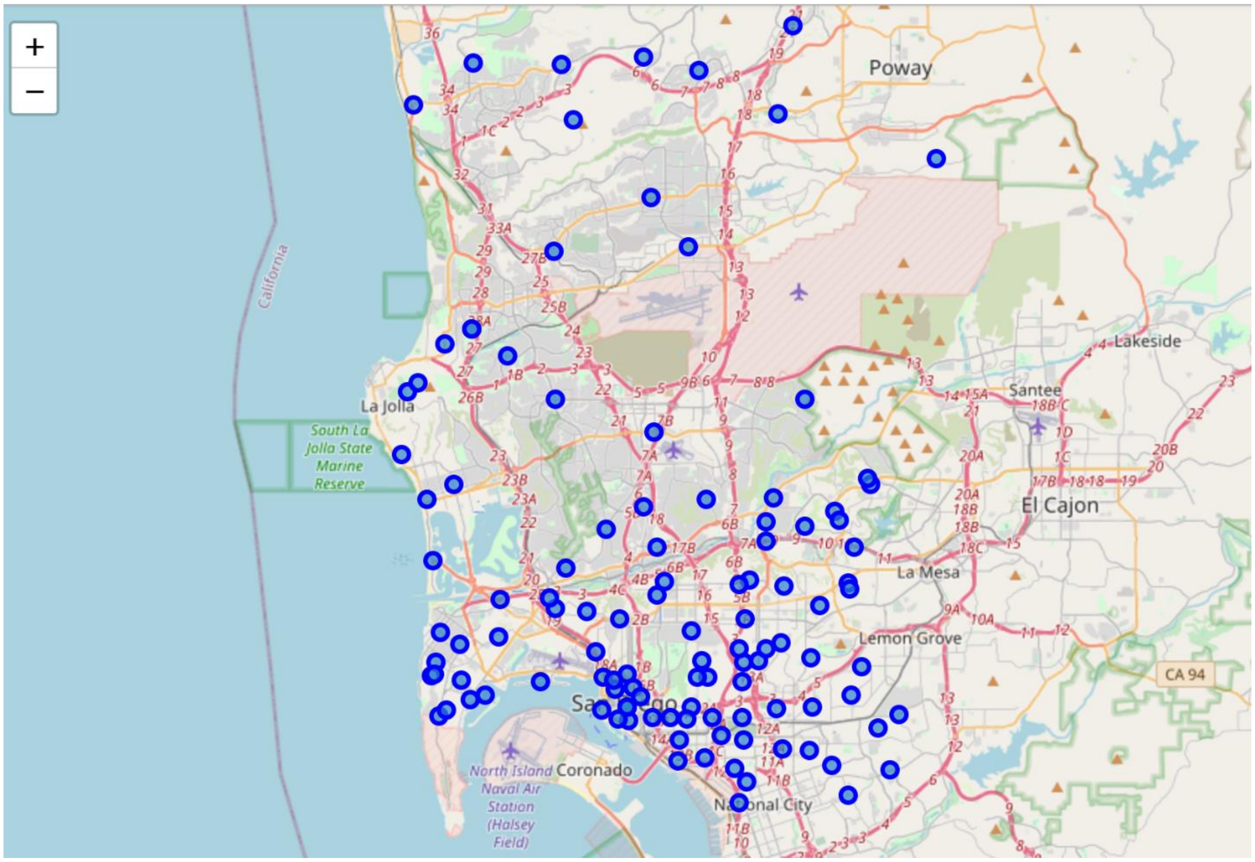
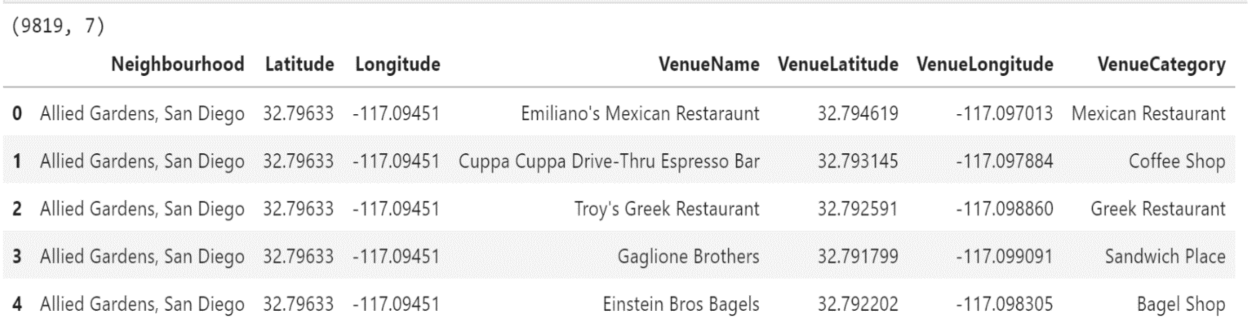


Figure 1: Neighbourhoods of San Diego.

4.2. Top 100 most common venues

We are using Foursquare API to get the top 100 venues within a radius of 2000 meters. We need to register a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare, passing in the neighborhoods' geographical coordinates 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. We can check how many venues were returned for each neighborhood with the data and examine how many unique categories can be curated from all the returned venues.



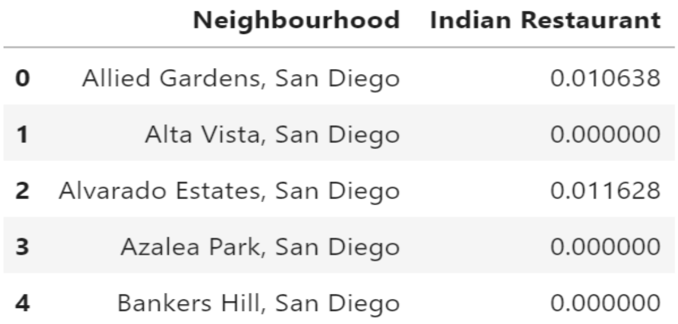
```
(9819, 7)
```

	Neighbourhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Allied Gardens, San Diego	32.79633	-117.09451	Emiliano's Mexican Restaraunt	32.794619	-117.097013	Mexican Restaurant
1	Allied Gardens, San Diego	32.79633	-117.09451	Cuppa Cuppa Drive-Thru Espresso Bar	32.793145	-117.097884	Coffee Shop
2	Allied Gardens, San Diego	32.79633	-117.09451	Troy's Greek Restaurant	32.792591	-117.098860	Greek Restaurant
3	Allied Gardens, San Diego	32.79633	-117.09451	Gaglione Brothers	32.791799	-117.099091	Sandwich Place
4	Allied Gardens, San Diego	32.79633	-117.09451	Einstein Bros Bagels	32.792202	-117.098305	Bagel Shop

Table 1: Top 100 venues that are within a radius of 2000 meters of each neighborhood

4.3. One-hot Encoding

One hot encoding is when categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are one-hot encoded. We will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of each venue category's frequency of occurrence. By doing so, we are also preparing the data for use in clustering. Since we analyze the “Indian Restaurant” data, we will filter the “Indian Restaurant” as a venue category for the neighborhoods.



	Neighbourhood	Indian Restaurant
0	Allied Gardens, San Diego	0.010638
1	Alta Vista, San Diego	0.000000
2	Alvarado Estates, San Diego	0.011628
3	Azalea Park, San Diego	0.000000
4	Bankers Hill, San Diego	0.000000

Table 2: Indian Restaurant data for each Neighborhood

4.4. K-Means Clustering

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on. K-means was chosen as the variables (Venue Categories) are huge, and in such situations, K-means will be computationally faster than other clustering algorithms.

We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Indian Restaurants.” The results will allow us to identify which neighborhoods have a higher concentration of Indian Restaurants and which neighborhoods have fewer. Based on Indian restaurants' occurrence in different neighbourhoods, it will help us answer the question as to which neighborhoods are most suitable to open new Indian restaurants.

5. 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 Restaurants:

- Cluster 0: Neighbourhoods with a low number of Indian Restaurants
- Cluster 1: Neighbourhoods with a high number of Indian Restaurants
- Cluster 2: Neighbourhoods with a moderate concentration of Indian Restaurants

The clustering results are visualized in the map below with cluster 0 in red color, cluster 1 in purple, and cluster 2 in mint green color.

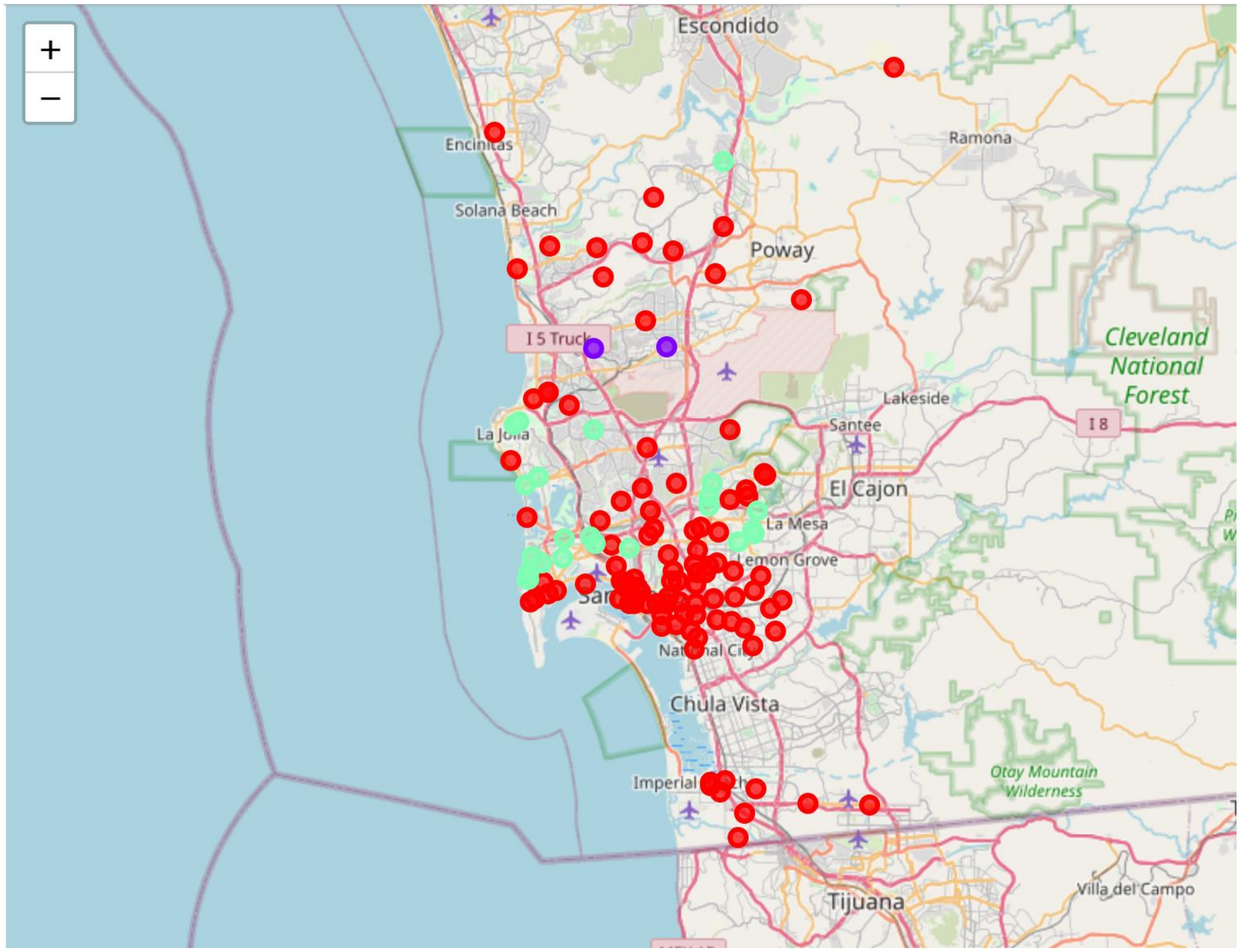


Figure 2: Clusters representing the density of Indian Restaurants in neighborhoods.

6. Discussion

As observations noted from the map in the Results section, most Indian restaurants are concentrated in cluster 1 area of San Diego city, and moderate number in cluster 2. On the other hand, cluster 0 has a deficient number of Indian restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new Indian restaurants as it is very little to no competition from existing

Indian restaurants. Meanwhile, Indian restaurants in cluster 1 are likely suffering from intense competition due to oversupply and high Indian restaurants' concentration. The results also show that the plethora of Indian restaurants mostly happened in the city's central area, with the suburb area still have very few Indian restaurants. Therefore, this project recommends people in business to capitalize on these findings to open new Indian restaurants in neighborhoods in cluster 0 with little to no competition. Business people with unique selling propositions to stand out from the game can also open new Indian restaurants in neighborhoods in cluster 2 with moderate competition. Lastly, people in business are advised to avoid neighborhoods in cluster 1 that already have high Indian restaurants' concentration and suffer intense competition.

7. 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., people in the business of restaurants and investors regarding the best locations to open a new Indian restaurant. To answer the business question raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new Indian restaurant. This project's findings will help the relevant stakeholders capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Indian restaurant.