COURSERA CAPSTONE
IBM APPLIED DATA SCIENCE CAPSTONE
OPENING A NEW INDIAN RESTAURANT IN SAN DIEGO, CALIFORNA

By Vishvajit Indurkar

Introduction

For many residents in San Diego, visiting Indian restaurants is a great way to relax and enjoy themselves or even to grab a gift for a friend. They have a wide variety of desserts to choose from. Some Indian restaurants are like a one-stop destination for all types of foodies. For Indian restaurant owners, the central location and the large crowds near Indian Restaurants provide a great distribution channel to market their products and services. Business developers are also taking advantage of this trend and are building more bars to cater to the demand. As a result, there are many Indian restaurants in San Diego and many more are being built. Opening a Indian restaurant allows the business owner to earn a consistent income. Of course, as with any business decision, opening a new location requires serious consideration and is a lot more complicated with a lot of moving parts. Particularly, the location of the Indian restaurant is one of the most important decisions that will determine whether the business will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in San Diego California to open a new Indian restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In San Diego California if someone is looking to open a new Indian restaurant, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to business owners and investors looking to open or invest in new Indian restaurant in San Diego California. This project is timely as San Diego is saturated with Indian restaurants. Data from San Diego .gov showed that Indian restaurants are expected to grow by an additional 8 percent.

Data

To solve this problem, we will need the following data: • List of neighborhoods in San Diego. This defines the scope of this project, which is confined to the city of San Diego, which is in California.

• Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and to get the venue data.

• Venue data, particularly data related to Indian restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This wiki page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Diego) contains a list of neighborhoods in San Diego, with a total of 170 neighborhoods. I will be using web scraping techniques to extract the data from the wiki page, using Python requests and beautifulsoup packages. For the San Diego neighborhoods I will get the geographical coordinates of the neighborhoods using the Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. The Foursquare API will provide many categories of the venue data, we are particularly interested in the desert bar category which will help us solve the business problem

stated above. This is a project that will make use of the data science skills covered in this course. Skills such as working with APIs (Foursquare), data cleaning, data wrangling, machine learning (K-means clustering) and visualizing maps with 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 techniques that were used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of San Diego. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Diego). We will use web scraping with Python requests and the beautifulsoup package to extract a list of the neighborhood's 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 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 neighborhoods on a map using the Folium package. This allows us to perform a stare and compare to make sure that the geographical coordinates returned by the Geocoder package are correctly plotted in the city of San Diego.

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 neighborhoods 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 validate how many venues were returned for each neighborhood and examine how many unique categories can be created 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. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "Indian Restaurant" data, we will filter the "Indian Restaurant" as a venue category for the neighborhoods.

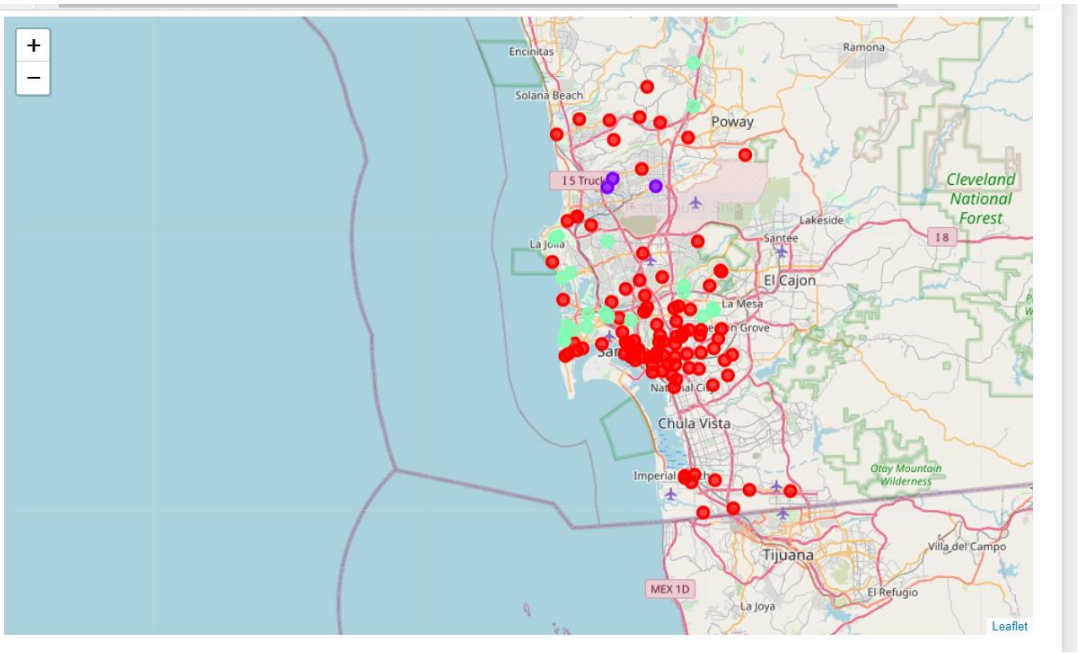
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and most popular unsupervised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

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 Restaurant": • Cluster 0: Neighborhoods with low number to no existence of Indian Restaurants • Cluster 1: Neighborhoods with moderate number of Indian Restaurants • Cluster 2: Neighborhoods with high concentration of Indian Restaurants

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

Out[43]:



Discussion:

As observations noted from the map in the Results section, Most of the Indian Restaurants are concentrated in the central area of San Diego, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to totally no Indian restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new Indian restaurants as there is very little to no competition from existing Indian restaurants. Meanwhile, Indian restaurants in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Indian restaurant. From another perspective, this also shows that the oversupply of Indian restaurants mostly happened in the central area of San Diego, with the suburbs only have a few restaurants. Therefore, this research recommends that restaurant owners capitalize on these findings to open new Indian restaurants in neighborhoods in cluster 0 with little to no competition. Restaurant owners with unique selling points that stand out from the competition can also open new Indian restaurants in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of Indian restaurant and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Indian Restaurants, there are other factors such as population and income of residents that could influence the location decision of a new Indian Restaurant. However, such level of detail was not required for this project. Future endeavors could create a methodology to estimate the population and income to be used in the clustering algorithm to determine the preferred locations to open a new Indian 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:

Most of the Indian Restaurants are concentrated in the central area of San Diego, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to totally no Indian restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new Indian restaurants as there is very little to no competition from existing Indian restaurants. Meanwhile, Indian restaurants in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of Indian restaurant. From another perspective, this also shows that the oversupply of Indian restaurants mostly happened in the central area of San Diego, with the suburbs only have a few restaurants. Therefore, this research recommends that restaurant owners capitalize on these findings to open new Indian restaurants in neighborhoods in cluster 1 with little to no competition. Restaurant owners with unique selling points that stand out from the competition can also open new Indain restaurants in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of Indian restaurant and suffering from intense competition