Selecting a Suitable Neighborhood
In
San Jose, California
For
Opening a New Indian Restaurant

S.Vithiyashankar

30 December 2019

1. Introduction

Background

San Jose California is said to be the capital of Silicon Valley. It is home to many High tech and Information Technology companies. A diverse population lives in San Jose who belongs to many ethnicities. Therefore San Jose have many kind of restaurants where food lovers can taste ethnic food from any part of the world. Restaurant business is a profitable venture here.

Business Problem

As San Jose has a high concentration of Indian and Asian population, there is a demand and market for Indian food. Not only people from India, people from other ethnic groups also enjoy Indian food. This project analyses the best neighborhood where a new Indian Restaurant can be opened so that it will have less competition.

Interest

This analysis will be of great interest for Business entrepreneurs who want to invest in an indian restaurant and looking for the most suitable venues for opening such a restaurant.

2. Data

Data Acquisition

The scope of the project was limited within the San Jose city Limits. Following data were needed for this project:

1. Neighborhood data of San Jose city.
2. Geo locations of the neighborhoods
3. Venue data of the Indian restaurants currently exist in San Jose

For the San Jose Neighborhood data, the following San Jose Wikipedia page was used: https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Jose,_California

For geolocation and geocoding, Nominatim API was used. FourSqure API was used for obtaining the current venues of the Indian restaurants.

Data cleaning

The San Jose neighborhood data obtained from Wikipedia page and web scraping techniques used to extract the data using Beautiful Soap library. Geo coordinates were obtained from Nominatim API. Venue data was obtained from FourSqure API. Data acquired from various sources were combined into one table. There were some rows have missing values, which were removed.

Feature Selection

The neighborhood data, coordinates of the neighborhoods, and the categories of the venues were used for the analysis.

3. Methodology

The Methodology is comprised of the following steps:

1. Build a data frame of neighborhoods in San Jose, California by web scraping the data from Wikipedia page
   As the first step, the neighborhood data of the San Jose city is needed. The data available in the Wikipedia page was used. The web scraping was done using Python requests and Beautiful Soap library.

2. Get the geographical coordinates of the neighborhoods
   Next, the geographical coordinates of the neighborhoods need to be obtained in order to use the FourSqure API. Therefore, the Nominatim API was used. It converts the Neighborhood names to Geographical coordinates in the form of latitude and longitude. After gathering the data, the data was populated in to a Pandas data frame and the neighborhoods were visualized in a map using Folium package. This allows us to visualize the data and verify that the geographical coordinates returned by Geocoder are correct and correctly plotted in the San Jose city map.

3. Obtain the venue data for the neighborhoods from Foursquare API
   Next, the FourSqure API was used to obtain the venue data. In each neighborhood, the top 100 venues within the 2000 m radius of the geo coordinate of a neighborhood. As San Jose is a densely populated city, the neighborhoods were situated close by, that is why a 2000 m radius was selected. FourSquare returned the venue data in JSON format. The

data needed were extracted from the response JSON data, which include venue name, category latitude and longitude. With this data, the information on how many venues were returned for each neighborhood was obtained. From this, how many unique categories could be curated from the returned venues was determined. Then each row was grouped based on neighborhood and the mean of the frequency of occurrence of each category was taken. This helps in preparing the data for clustering as well. As we are analyzing “Indian Restaurant” data, the venue category “Indian Restaurant” was used to filter the venue categories.

4. Explore and cluster the neighborhoods
   As the next step, the clustering was performed on the data by k-means clustering. K Means clustering algorithm identifies k number of centroids and then allocates every data point to nearest cluster, while keeping the centroids small as possible. It is one of the simplest and popular unsupervised algorithm and is particularly suited to solve the problem for this project. The neighborhoods were clustered into 3 clusters based on their frequency of occurrence for “Indian Restaurant”. The results will allow to identify which neighborhoods have higher concentration of Indian Restaurants while which neighborhoods have fewer number of Indian Restaurants.

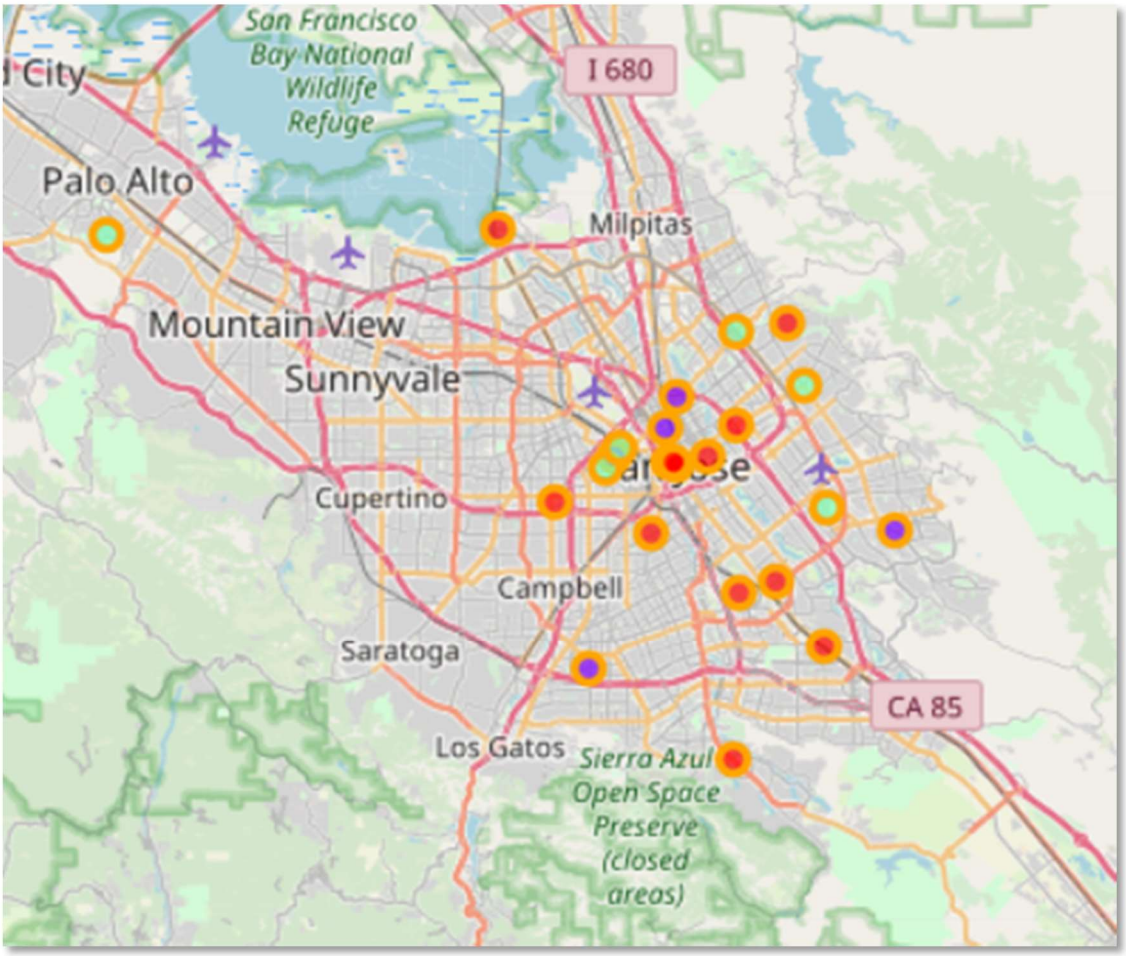
5. Select the best cluster to open a new Indian Restaurant
   Based on the occurrence of Indian restaurants in different neighborhoods, it will help to answer the question as to which neighborhoods are most suitable for opening a new Indian restaurant. The neighborhood with the minimum occurrences of Indian restaurants will be the best candidates for opening an Indian restaurant.

6. Results

The results from the K-means clustering show that the neighborhoods can be categorized into 3 clusters based on the frequency of occurrences for “Indian Restaurant”:

1. Cluster 0: Neighborhoods with low number to no existence of Indian Restaurants
2. Cluster 1: Neighborhoods with high concentration of Indian Restaurants
3. Cluster 2: Neighborhoods with moderate number of Indian Restaurants

The results of the clustering are visualized in the following map:



7. Discussion

Most of the Indian Restaurants are concentrated in the northern part of the San Jose city, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number or no Indian restaurants in the neighborhoods. This represents a great opportunity and high potential to open new Indian restaurants in cluster 0 areas as there is very little to no competition from existing restaurants. Meanwhile, Indian restaurants in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of restaurants. From another perspective, this also shows that the oversupply of Indian restaurants mostly happened in the northern and central area of the city, with the southern area still have very few or no Indian restaurants. Therefore, this project recommends businessmen to capitalize on these findings to open new Indian restaurants in neighborhoods in cluster 0 with little to no

competition. Businessmen are advised to avoid neighborhoods in cluster 1 which already have high concentration of Indian restaurants.

There is a shortcoming in this analysis; In this study, only one aspect is being considered, which is the frequency of occurrence in each neighborhood. However, other factors such as the population density of each neighborhood, the number of Indian and Asian population in each neighborhood and the income levels of each neighborhood also will affect the results. However, there were no data available in those aspects in neighborhood level.

8. Conclusion

In this project, we have gone through the process of identifying a 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 providing recommendations to the relevant stakeholders regarding the neighborhood to open a new Indian restaurant. It was found that neighborhoods given by Cluster 0 are the most recommended.