IBM CAPSTONE PROJECT WEEK 5

Introduction

This is a capstone venture for IBM Data Science Professional Certificate. In this task, I am making a speculative situation for an idea that there may not be sufficient Indian Restaurants in Los Angeles Area. In this manner it may be an extraordinary open door for a business visionary who is situated in USA. As the Indian food is well known among Asian people group, so this business visionary may consider starting its business in territories where Asian network resides. With the reason at the top of the priority list, finding the area to open such an eatery is one of the most significant choices for this business visionary and I am planning this task to assist him with finding the most appropriate area.

Business Problem

The goal of this capstone project is to locate the most appropriate area for the business person to open a new Indian Restaurant in Los Angeles, USA. By utilizing Datascience techniques and devices alongside Machine Learning algorithms, for example, Clustering, this project expects to give response to answer the business question :
In Los Angeles, if a entrepreneur needs to open an Indian Restaurant, where would it be advisable for them to think about opening it?

Target Audience

The entrepreneur who wants to find the location to open authentic Indian restaurant.

Data

To solve this problem, we will need below data:

- List of neighborhoods in Los Angeles, USA
- Latitude and Longitude of these neighborhoods
- Venue data related to Indian restaurants. This will help us find the neighborhoods that are more suitable to open an Indian Restaurant.

Extracting The Data

- Scrapping of Los Angeles neighborhoods via online website.
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

Methodology

First, I need to get the list of neighborhoods in Los Angeles, USA. This is possible by extracting the list of neighborhoods from https://familypedia.wikia.org/wiki/List_of_incorporated_cities_and_towns_in_California

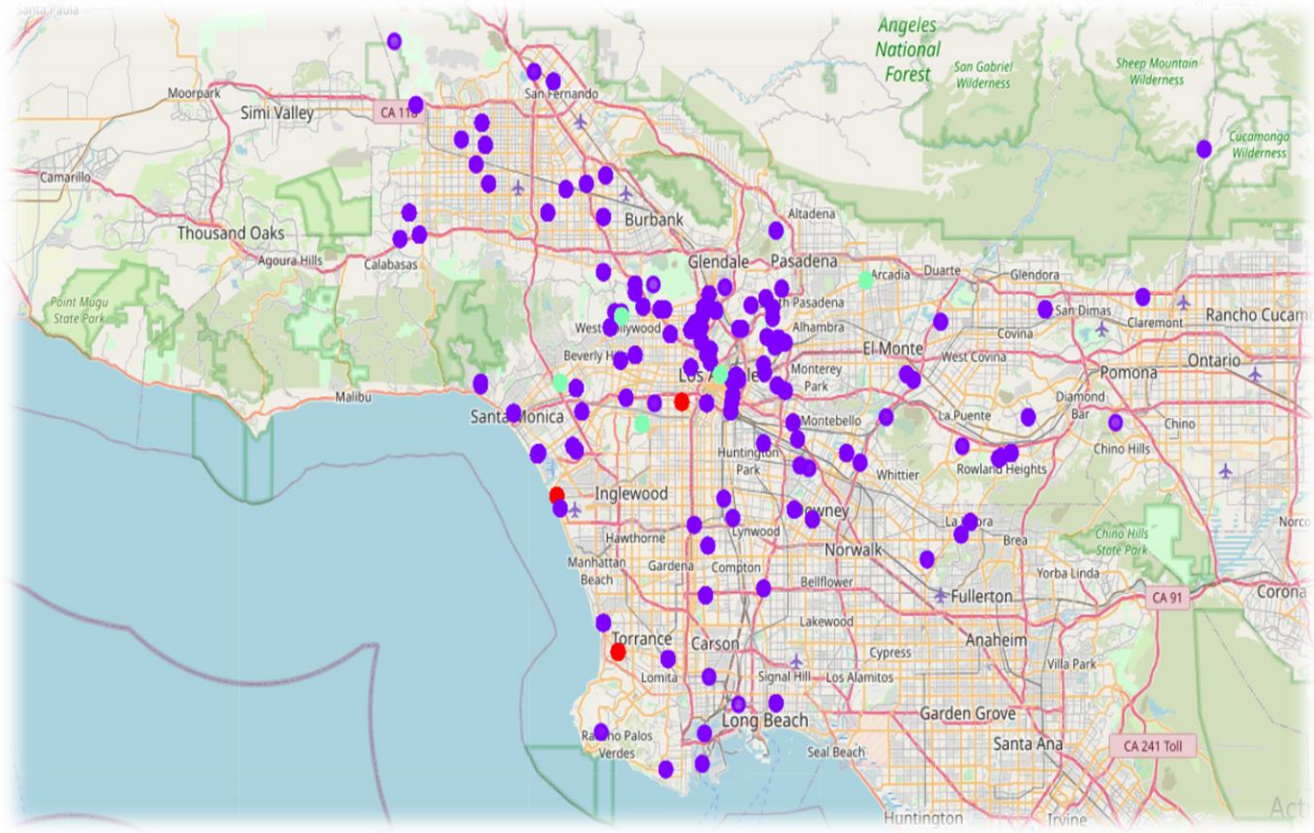
I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

However, it is only a list of neighborhood names. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I used Geocoder Package which gives us the coordinates of Los Angeles neighborhoods. After gathering these coordinates, I visualize the map using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Indian restaurants”. Lastly, I performed the clustering method 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 it is highly suited for this project as well. I have clustered the neighborhoods in Los Angeles into 3 clusters based on their frequency of occurrence for “Indian food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

RESULT

CLUSTERS



The results from k-means clustering show that we can categorize Los Angeles neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood:

- **Cluster 0: Neighborhoods with the less number of Indian restaurants.**
- **Cluster 1: Neighborhoods with no Indian restaurants.**
- **Cluster 2: Neighborhoods with a more number of Indian restaurants**

The results are visualized in the above map with Cluster 0 in Violet, Cluster 1 in Cyan , Cluster 2 in Red.

Discussions

- Most of the Indian Restaurants are concentrated in the central area of the city
- Highest number in cluster 2 and moderate number in cluster 0
- Cluster 1 has no Indian Restaurants in the neighbourhoods

Recommendations

Most of the Indian restaurants are in cluster 2 which is around Green Field , Huntington Beach, Rocklin, Cupertino and Nevada City lowest in Cluster 1 areas which are in Auburn, Apple Valley, Adelanto and Barstow areas. Looking at nearby venues it seems cluster 0 might be a good location as it is a business area and there are not a lot of Indian restaurants in the areas like Redlands, Roseville, Seaside and Dana Point.

Conclusion

- The neighbourhoods in cluster 0 are the most preferred locations to open a new Indian Restaurant.
- Findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Indian Restaurant.