Final Report

Capstone Project - The Battle of Neighbourhoods (Part-2)

Introduction:

Opening a restaurant is all about location, location, location. However, not every restaurant is suitable for every location, and vice versa. It comes down to a combination of restaurant style, target audience, your competitors. If you can define your restaurant type and identify your target demographic and its most populated areas, you'll be well on your way to choosing a restaurant location that sets your business up for success. For my assignment I have taken a business case of opening an Indian restaurant in United States. And for this my biggest problem or challenge is to find a best suitable location where I can have highest population of Indians, more workingclass people or more earnings and more affordable prices for rent or for land or building purchase and finally less competitors.

Anyone who wants to get into the restaurant business and wants help in finding the best location using Data Science and Machine Learning algorithms will be interested in this project report.

Data:

For our project, we used 3 datas, first one is to find state with highest indian population, next is to find the highest indian populated city in that state and the third dataset is the Neighbourhoods of Boroughs with highest Indian population.

From the First dataset and second Dataset, It cames to know that New York city have the highest Indian population. Hence I continue with Neighbourhoods of Boroughs of New York. New York mainly have 4 Boroughs in which Queens have the highest population of Indians. To solve our problem of finding a best location to start an Indian restaurant in US, we need datasets based on various parameters such as:

1. Population of target audience in all the boroughs of Queens based on their ethnicity
2. Earnings data of the working class living in the target location.
3. We need the data about the required business floorspace and rateable value statistics of each borough.
4. Considering the competitors factor, we also need the data of existing Licensed Restaurants in each borough.

Methodology:

To work on the solution, I have used Pandas library to read the data in CSV format and convert into pandas dataframe. Extensive data exploration analysis is done, where lot of data is cleaned and presented in a suitable format. After cleaning and preprocessing we need to plot them on the geographical maps to get more about the behaviour of the data.

First we need to get the geo-coordinates of the borough and the geo-coordinates of the neighbourhoods of the borough from the web. I have used the Wikipedia pages to get this data.

https://en.wikipedia.org/wiki/Indian_Americans

To read data from these URLs, I have used the requests, urllib and BeautifulSoup libraries of python.

After I have the geo-coordinates information of the borough and its neighbourhoods, I need the other data such as the venues or places of the neighbourhoods, the venue categories, working hours and so on. All this data is called Location data, and to get this data I need a reliable and efficient location data providers and hence I am using Foursquare as the data provider. I have used the Foursquare API to explore the neighbourhoods in London city. I have also used the Explore function to get the most common venue categories in each neighbourhood and then use this feature to group the neighbourhoods into clusters. To cluster the neighbourhoods I am using K-means Clustering algorithm.

Geopy module and Nominatim library is used to convert a given address into the latitude and longitude values.

To visualize the neighbourhoods, the library Folium is used, to display the maps of New York, with the boroughs super imposed on it and to display the map of borough with the neighbourhoods superimposed on it.

A python function getNearbyVenues() is created , to give the venue details like venue name, venue latitude, venue longitude, venue category along with neighbourhood name, latitude and longitude for each neighbourhood.

After the venue data for each neighbourhood of the Queens borough is generated , One-Hot encoding is applied on the venue category data, so that the analysis of the data will be easy in grouping the neighbourhoods based on the frequency of occurrence of each venue category.

Once the neighbourhoods are grouped based on the frequency of occurrence of the venue category, the top 10 venues of each neighbourhood are displayed as a dataframe. After all the above data exploration and analysis and top 10 venues of each neighbourhood are identified, the K-means Clustering algorithm is applied to the resultant dataframe to segment the data into 5 Clusters and all these 5 clusters are visualised in a map using the Folium library and finally the 5 clusters are examined to determine the discriminating venue categories that distinguish each cluster.

Results:

In the Segmenting and Clustering section, the neighbourhoods of Queens borough are explored, and the top 10 venues of each neighbourhood are listed. The neighbourhoods are Clustered into 5 clusters using K-means algorithm and their most common neighbourhoods are identified.

Conclusion:

There is always room for improvement and hence the above solution I have provided can also be improved and the machine learning models can be trained and tested for best results depending upon the data we have.