Benefit Zones for chain of Indian Restaurants in Helsinki, Finland

Prakirth Govardhanam

Applied Data Science Capstone by IBM/Coursera

Introduction

In a general perspective, Businesses, especially Chain Stores, thrive when the possibility of risk/competition is either low or close to none (Not Applicable to Start-Ups). This perspective (unlike Mark Zuckerberg's strategy to monopolize Social Media) signifies tact and subtlety and not eliminating competition but circumventing the need to compete when possible.
However, competition for every Business is as essential as work-outs for a healthy human body and mind. Hence, I conclude by quoting Jack Ma (Co-founder of Alibaba Group).

> "Forget about Your Competitors. Focus on Your Customers"

Although, Indian restaurants are specific to Indians, the variety of Indian cuisine is enjoyed delightfully by all the Nordic countries. Hence, the Customer base for Indian cuisine in the Nordics is awaiting the arrival of native-authentic restaurants for Business.

(PS: Market research for establishing the above hypothesis is out-of-scope for this project)

In this project, I focus on the possibility of establishing a chain of Indian restaurants in Helsinki, capital city of Finland (one of the Nordic countries). I try to find possible-beneficial locations within the Neighborhoods (Districts) of Helsinki for establishing a chain of Indian restaurants. The primary conditions for the project are as follows:

- Locate Popularity Centre (Major Assumption 1) in the District – for attention
- Identify Districts with:
    - Absence of Indian restaurants – for profitable Business
    - Presence of Indian restaurants – for moderate competition to evolve Business model
    - Presence of Indian restaurants within Popular venues – to circumvent extreme competition

Data

Data sources used to determine the Districts within the city of Helsinki are provided by:

- Wikipedia – for labels of the Districts of Helsinki
- ***geopy*** (Python package) – for coordinates of the Districts of Helsinki
- **Foursquare** (API) – for popular venues, restaurants and their respective geospatial data

Methodology

Hypothesis

Indian restaurants are specific to Indians, the variety of Indian cuisine is enjoyed delightfully by all the Nordic countries. Hence, the Customer base for Indian cuisine in the Nordics is awaiting the arrival of native-authentic restaurants for Business.

Major Assumption

1. Popularity Centre = the centroid (mean position) of the most popular venues from the Top10 Venue Categories (by frequency of occurence) in each District will be considered as the "Popularity Centre" within every District

Clarification #1: Popular Venues from Top10 Venue Categories were ideally planned to be filtered by Ratings of Venues. Unfortunately, I have a Sandbox account & Ratings of Venues at the scale I need would be possible only with Premium accounts

2. In fact, more Indian Restaurants exist than explored Indian Restaurants using FourSquare API. Since, the project is based on "**using FourSquare API for implementation of the Idea**" we will assume the following:
   "*explored Indian Restaurants*" == "*existing Indian Restaurants*"

Minor Assumption

1. From the source of District labels, Swedish-names of Finland could be confused with Swedish-names of Sweden in the FourSquare API (as observed during analysis). Hence, we will extract and work only with Finnish-names of the Districts.

2. Since *explored Indian Restaurants* are very less in number, we will consider both Venue Categories, ***Indian Restaurant*** and ***Himalayan Restaurant*** as *'Indian Restaurant'*

Results & Discussion

Results are separated into two sub-sections based on the work in the project, as follows:

1. Data Preparation – involves Data Extraction, Data Cleaning & Data Organizing
2. Exploratory Data Analaysis – involves Data Analysis & Data Visualization

Data Preparation

The primary source of data, i.e., District labels and coordinates were extracted using **Wikipedia** & **geopy**, respectively.

The Wikipedia page lists the Finnish & Swedish names of the Municipalities (Districts) in Helsinki, Finland. Web scraping was performed using the packages, *requests*, *BeautifulSoup* & *Regular Expressions (re)*. As stated in Minor Assumption 1, to avoid confusion of Districts in Sweden, Finnish names of Districts were chosen. The total relevant districts extracted from the Wikipedia page were 110.

District coordinates or spatial data were extracted using the modules, *geocode* & *Nominatim* in geopy. The none values were circumvented using *try* & *except* loop for *'AttributeError'*. The acquired coordinates were manually scanned through (since it was just 109x2 values) and identified anamolies were investigated, such as the coordinates of 'Pasila' & 'Töölö' were wrongly interpreted by geopy, and 'Kampinmalmi' was not identified at all. Hence, these Districts and their coordinates were omitted from the final data. Hence, total Districts with relevant data are 107 Districts. A dataframe was constructed with information of each District in rows and District, Latitude & Longitude as columns.

Exploratory Data Analysis

Prepared data was visualized as Helsinki city map for further anamolies in the coordinates of Districts using Folium. Five Districts ('Vanhakaupunki', 'Siltasaari', 'Reijola', 'Vironniemi', 'Koivusaari') were identifed as irrelevant Districts, which was probably due to overlapping names of the Districts across regions. Hence, these Districts and their coordinates were omitted from the final data as well, leading to 102 Districts in total.

FourSquare API is used to extract nearby venues within a radius of 500 across all the Districts of Helsinki, considering the coordinates of each District as the point of interest. The function for data extraction from the json file was replicated from 'Neighborhoods-New-York' lab in the course. The acquired data was reconstructed in a dataframe with labels and coordinates of Venues, Venue Categories & Districts as columns.

Clarification #2: Function (*getNearbyVenues*()) might interrupt due to possible errors in .json file structure, which could not be tackled. The code has to be re-run for successful execution.

Clarification #3: Venues data acquired from FourSquare API through *getNearbyVenues*() differs from day-to-day, probably due to satellite movement and communication differences. Reproduciblity cannot be expected in the Top10 Venue Categories on different days of execution. Hence, the values such as the following might differ:

- Total Districts with venues
- Total venues
- Total Venue Categories
- Total Indian venues

However, the code relies on the analysis and not on the values. Hence, hassle-free.

(***With regard to Clarification #3, date of run:*** *December 30th, 2020*)

A total of **1771 venues**, **256 unique venue categories**, across **100 Districts** in Helsinki were extracted. Indian restaurants in Helsinki Districts, are identified as per Major Assumption 2. There are **22 Districts with Indian restaurants** across Helsinki and are visualized as Red Zones (Major-Competition Zones) using *folium*.

The venues were analysed for extracting popular venues based on the Top10 venue categories (by frequency of occurrence across Districts). Popular venues dataframe was reconstructed with detailed information about venues and Districts to identify popular venues and Districts by using geospatial data of the venues. Popular venues were visualized using *ClusterMap* in *folium*.

Total number of Districts were identified based on the presence/absence of Indian restaurants in the Districts as follows:

- Districts **WITH** Indian Restaurants: 22
- Districts **WITH** Indian Restaurants **in Top10 venue categories**: 14

- Districts **WITH** Indian Restaurants **NOT in Top10 venue categories**: 8
- Districts **WITHOUT** Indian Restaurant: 78

Popularity Centres (coordinates) were calculated from popular venues dataframe as the mean of latitudes and longitudes of venues grouped by Districts. (*Centroid is also known as mean (by definition), when a cluster is identified, which is a District, in our case*). *Popularity Centres* were visualized as Helsinki city map using *folium.*

Conclusion

Based on the above analysis, the visualized *Popularity Centres* can be tabulated as different levels of competition for establishing a chain of Indian Restaurants across Helsinki. Green Zones would be ideally considered to be the **"Benefit-Zones"** for the Chain of Indian restaurants.

Zone	Competition	Indian Restaurant in District	Indian Restaurant in Top10 venue categories
RED	Major	Yes	Yes
BLUE	Minor	Yes	No
GREEN	Low/None	No	No

Declaration

All the analysis and the assumptions are based on the data provided by the FourSquare API. Hence, I conclusively declare that these analysis could only be as accurate as the data extracted from the FourSquare API.