Coursera IBM Applied Data Science Capstone

Opening a New Private Language center in Oslo, Norway

Introduction

For many people who are wondering to visit new countries for studies, tourism, the first thing that it comes in mind, is what is the language of the destination country. So a lot of them start taking courses to improve their skills in that language. To do so, joining a language center may be a good choice. As with any business decision, opening a new private language center requires serious consideration and is a lot complicated than it seems. Particularly, the location of the language center is one of the most important decisions that will determine whether the project will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of Oslo, Norway to open a new private center language. Using data science methodology and machine learning techniques like clustering this project aims to provide solutions to answer the business question : What would be the best place, neighborhood to open a new private center language ?

Data

To solve the problem, we will need the following data :

- List of neighborhoods in Oslo. This defines the scope of this project which is confined to the city of Oslo
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particulary data related to centers of languages. We will use this data to perform clustering on the neighborhoods.

Source of Data

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Oslo) contains a list of neighborhoods in Oslo. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhood's using 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. Foursquare API will provide many categories of the venue data; we are particularly interested in the Commumity/Language center category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (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 technique that was used

Methodology

Firstly, we need to get the list of neighborhoods in the city of Delhi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Oslo). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods 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 Data Frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Oslo. Removing and adding any data manually. 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 check how many venues were returned for each neighborhood and examine how many unique categories can be curated 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 "Language center" data, we will filter the "Language

center” as venue category for the neighborhoods. Lastly, we will perform clustering on the data by using k-means clustering. Kmeans 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 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 “Language center”. The results will allow us to identify which neighborhoods have higher concentration of language centers while which neighborhoods have fewer number of language centers. Based on the occurrence of centers languages in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new language center.

Results

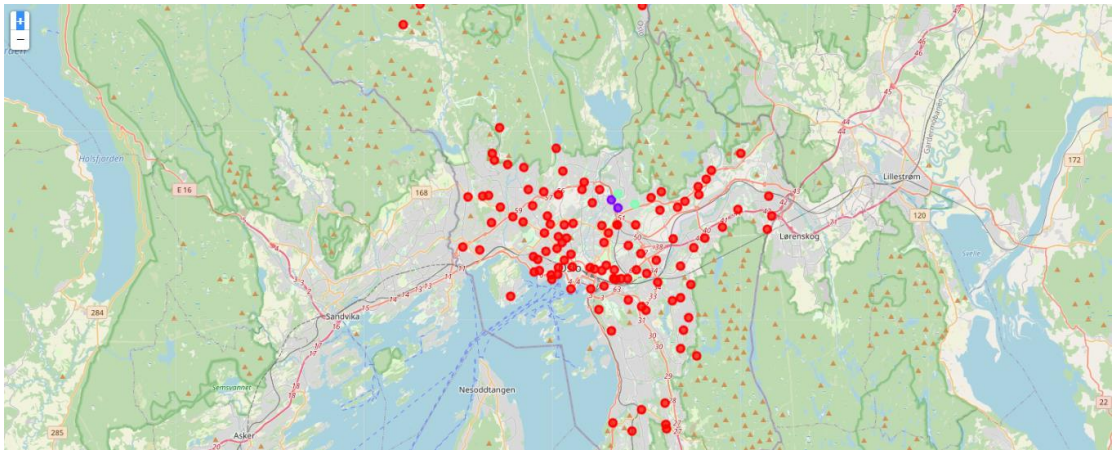
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Language center”:

• Cluster 0: Neighborhoods with low number to no existence of language centers.

• Cluster 1: Neighborhoods with moderate number of language centers.

• Cluster 2: Neighborhoods with high concentration of language centers.

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



Discussion

As observations noted from the map in the Results section, most of the language centers are concentrated in a few areas of Oslo, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no language center in the neighborhoods.

This represents a great opportunity and high potential areas to open new language center as there is very little to no competition from existing language centers. Meanwhile, language

center in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of language centers. From another perspective. Therefore, this project recommends property developers to capitalize on these findings to open new language center in neighborhoods in cluster 0 with little to no competition. Property developers with unique offers propositions to stand out from the competition can also open new language center 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 language centers and suffering from intense competition.

Conclusion

In this project, we have gone through the process of identifying the 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 lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new private language center. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a language center. The 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 language center.

References

Category: Neighborhoods in Delhi Retrieved from Wikipedia

https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Oslo

Foursquare Developers.

https://developer.foursquare.com/docs