Coursera Capstone

IBM Applied Data Science Capstone

Opening a Bar in Oslo, Norway

By: Iulian Catalin Costache
August 2020

Introduction

A friend of mine wants to open a bar in Oslo. Oslo is a vibrant city and people like to hang around in bars with friends. Since he worked as a bartender for many years, he would like to have his own bar. The question is where the new bar should be open?

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Oslo, Norway to open a successful new bar. Using data science methodology and machine learning techniques like clustering, this project aims to provide a solution to the most important question: in what location should be the bar open?

Data

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

- List of neighbourhoods in Oslo.
- Latitude and longitude coordinates of those neighbourhoods. This is required to plot the map and to get the venue data.
- Venue data, particularly data related to bars. We will use this data to perform clustering on the neighbourhoods.

Sources: Wikipedia and Forthsquare API

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Oslo. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_boroughs_of_Oslo). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. Then, 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 wonderful 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 DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Oslo.

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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the “Bars” data, we will filter the “Bars” as venue category for the neighbourhoods.

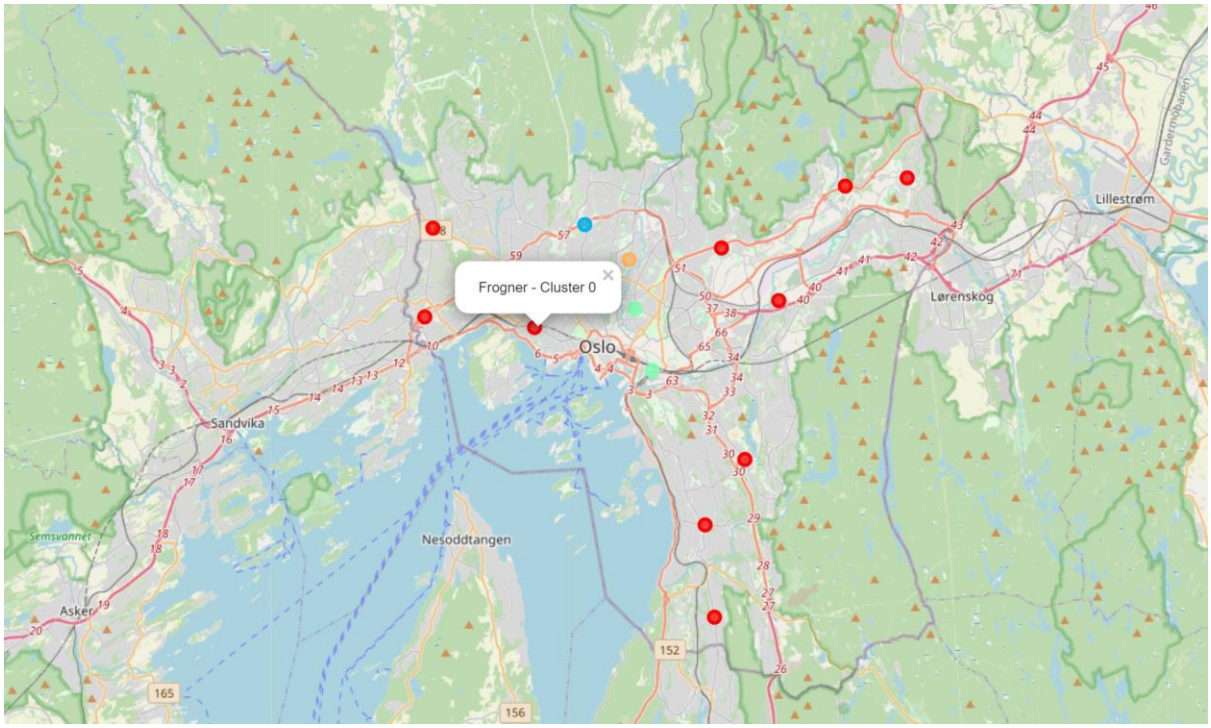
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for “Bar”. The results will allow us to identify which neighbourhoods have higher concentration of Bars and which not. Based on the occurrence of bars in different neighbourhoods, it will help us to answer the question as in which neighbourhoods are most suitable to open new bar.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters based on the frequency of occurrence for “Bar”:

- Cluster 0: Neighbourhoods with few or none bars, the biggest cluster
- Cluster 1: Neighbourhoods with high number of bars
- Cluster 2: Neighbourhoods with few bars
- Cluster 3: Neighbourhoods with the highest number of bars
- Cluster 4: Neighbourhoods with many bars

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in blue colour, cluster 3 in mint green, cluster 4 in Orange.



Discussion

As observations noted from the map in the Results section, most of the bars are concentrated in the central area of Oslo city, with the highest number in cluster 3 and almost none in cluster 0. However, there is a exception with Frogner neighbourhoods. Frogner is locate in the centre of Oslo, but it has very few bars, compared with the other central neighbourhoods. This represents a great opportunity and high potential area to open new bar as there is very little to no competition from ones. Meanwhile, bars in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of bars. From another perspective, the results also show that the oversupply of bars mostly happened in the central area of the city. Therefore, this project recommends that a new bar to be open in Frogner.

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 5 clusters based on their similarities, and lastly providing recommendations for where will be the best to open a new bar in Oslo.