

Coursera Capstone

IBM Applied Data Science Capstone
Opening a New private language center

Ahmed ZIANE, October 2021

[NOM DE LA SOCIETE] [Adresse de la société]

Business Problem

- Analyze and select the best locations in the city Delhi, India to open a new shopping mall.
- Used data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question
- In the city of Delhi, India, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

Data

- List of neighborhood's in Delhi. This defines the scope of this project which is confined to the city of Oslo, the capital city of the country of Norway.
- 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, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Sources of Data & Methodology

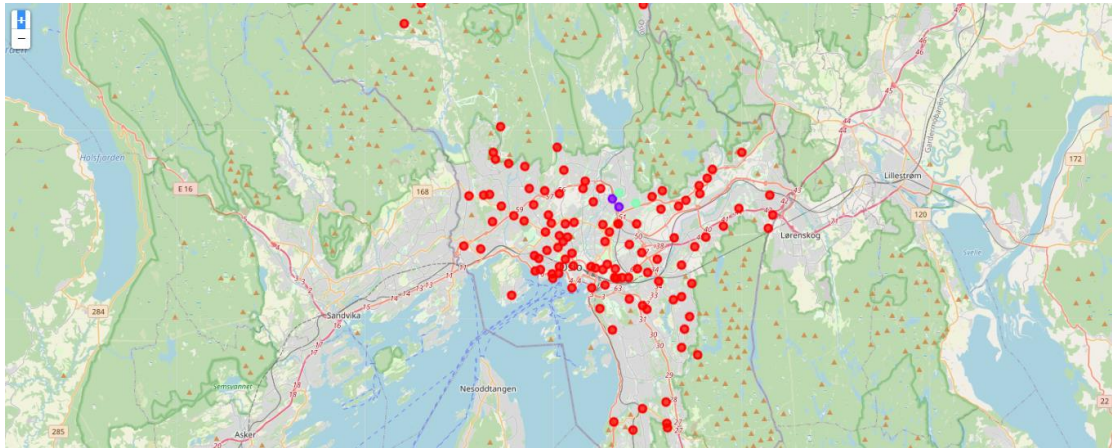
- This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Oslo) contains a list of neighborhoods in Oslo.
- Used web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and BeautifulSoup packages.
- Geographical coordinates of the neighborhood's using Python Geocoder package
- Used Foursquare API to get the top 100 venues that are within a radius of 2000 meters.
- Analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.
- Preparing the data for use in clustering.
- Analyzing the "Language center" data, we will filter the "Community/language center" as venue category for the neighborhoods.
- Performed 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.
- 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.

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

Most of the shopping malls are concentrated in the just some areas of Oslo.

- Highest number in cluster 2.
- Moderate number in cluster 1.
- On the other hand, cluster 0 has very low number to no language centers in the neighborhoods.
- Shopping malls in neighborhoods in cluster 1 with moderate competition.
- 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 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 new shopping mall. 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.