Neighborhood Segmentation and Clustering – Setting up a restaurant in Oslo

1. Introduction

For many people, dining outside is an important part of their lifestyle – be it on lunch break at work, a causal meal with friends and family, or a formal dinner to celebrate something special. It is therefore important that restaurant owners reflect upon their customers' needs and adjust their services accordingly to offer a better customer experience and at the same time make a profitable business.

As with any business decision, opening a new restaurant requires serious consideration and there is more to it than meets the eye. In particular, the location of the restaurant is one of the most important decisions that are potentially going to determine its success or failure.

The purpose of this projects is to select the best locations in the city of Oslo, Norway, to open a new restaurant, and also recommend types of restaurants that the city lacks. The project will mainly focus on geospatial analysis of the city of Oslo to understand which would be the best place to start a new restaurant business. Using machine learning techniques like clustering, this projects aims to recommend location where restaurant owners should open new restaurants as well as possible types of restaurants.

2. Description of Data and Sources

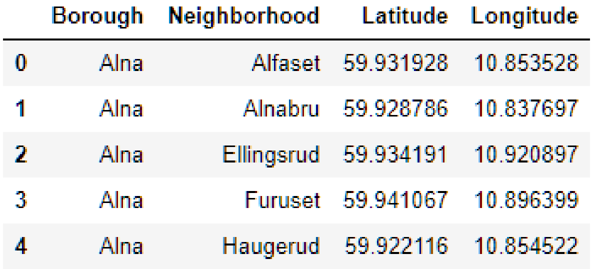
To approach the problem, we will use the following data:

- Using Wikipedia we have created a list of all boroughs and neighborhoods in the city of Oslo. Then, we have retrieved the respective latitude and longitude of all neighborhoods using geopy, a Python client for several popular geocoding web services.
- Having a complete list of locations including names and coordinates, we utilize the Foursquare API to explore the neighborhoods and segment them by getting the most common venues of a given neighborhood of Oslo.

3. Methodology & Results

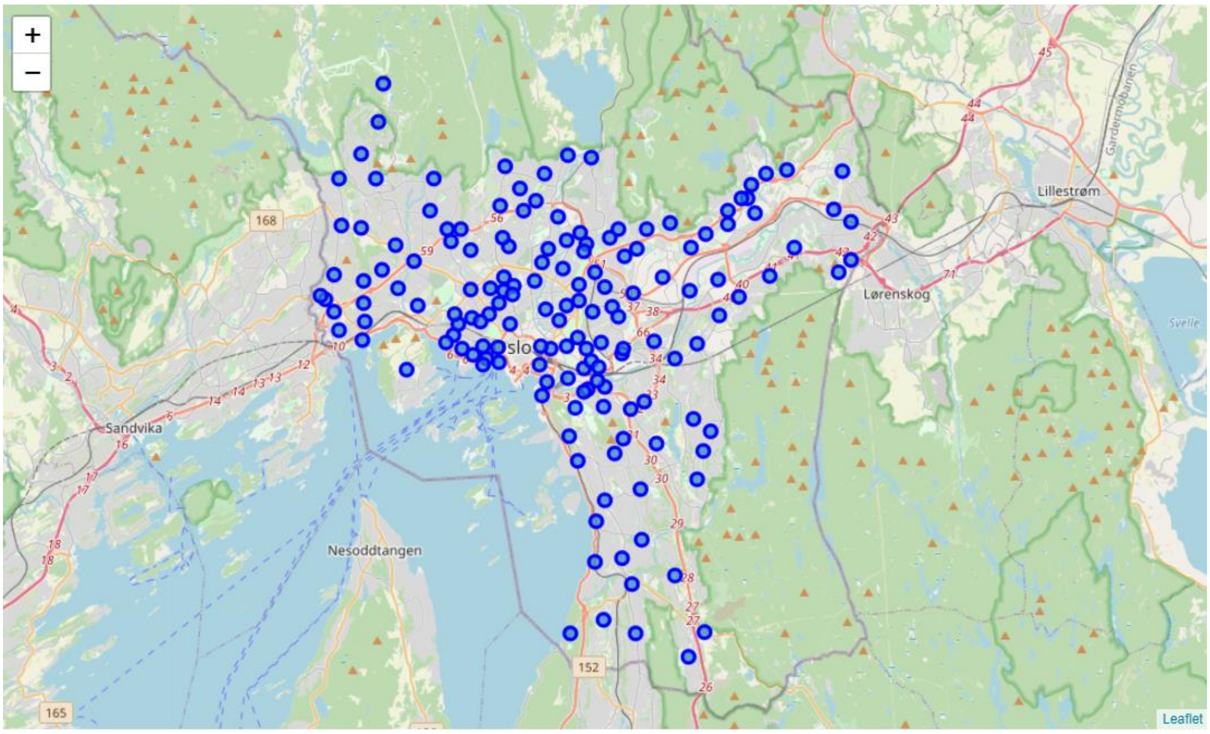
3.1. Exploratory Data Analysis

The main components of the data in this project are *Borough, Neighborhood, Latitude* and *Longitude*. The initial neighborhood list consisted of 157 neighborhood. However, one of the neighborhoods was not registered in the geolocator database and we could not retrieve its coordinates. The final data set consists of 156 entries, the head of which is shown below.

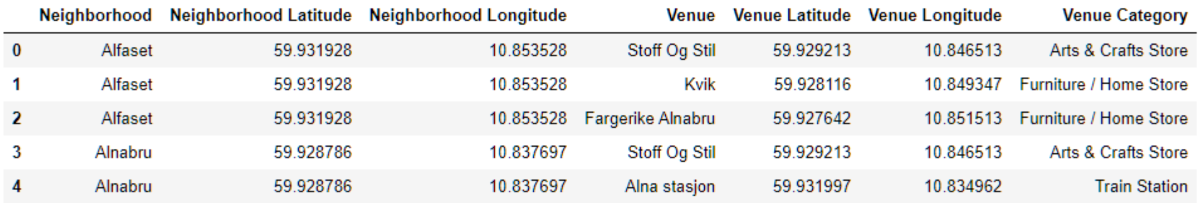


	Borough	Neighborhood	Latitude	Longitude
0	Alna	Alfaset	59.931928	10.853528
1	Alna	Alnabru	59.928786	10.837697
2	Alna	Ellingsrud	59.934191	10.920897
3	Alna	Furuset	59.941067	10.896399
4	Alna	Haugerud	59.922116	10.854522

Next, we visualize the neighborhoods of Oslo in the map. We utilize Python's folium library and the coordinates of each neighborhood to get the following visual.



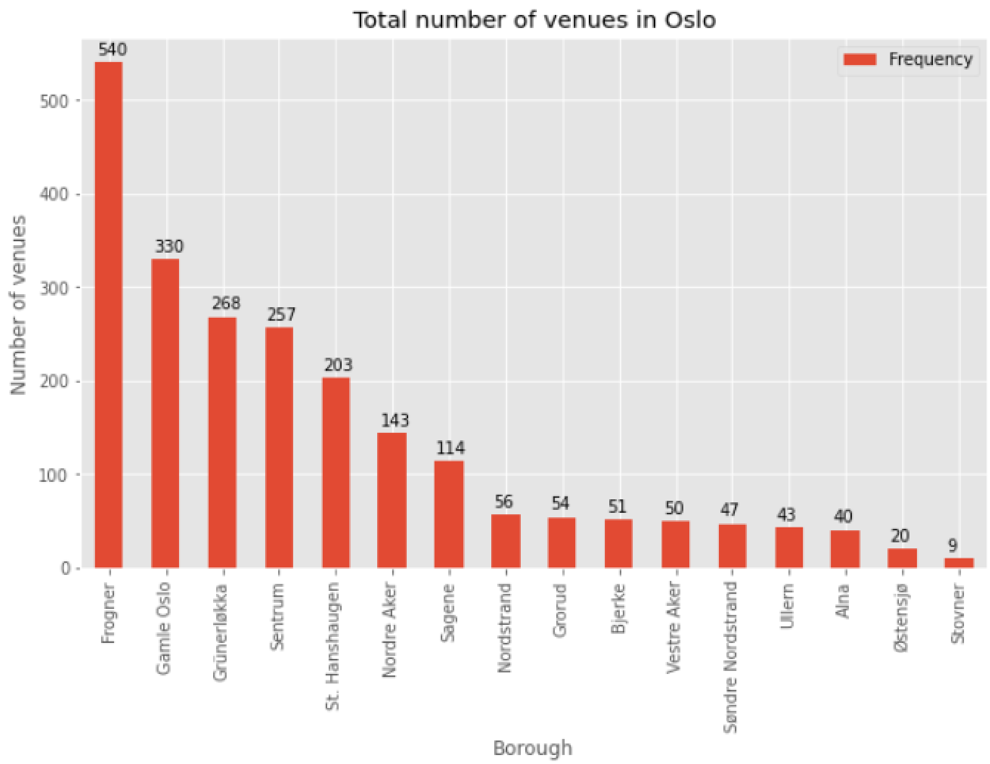
Moving on with the analysis of the neighborhoods, we start by making use of the Foursquare API to get a list of venues for each neighborhood. The default limit of 100 venues is applied and a radius of 500 meters from each neighborhood's given latitude and longitude values. The head of the Venues data frame which includes the name of the venue, its coordinates and category, as well as the neighborhood it belongs to and its coordinates is shown below. The Foursquare API has returned in total 2401 venues that belong to 227 unique categories.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alfaset	59.931928	10.853528	Stoff Og Stil	59.929213	10.846513	Arts & Crafts Store
1	Alfaset	59.931928	10.853528	Kvik	59.928116	10.849347	Furniture / Home Store
2	Alfaset	59.931928	10.853528	Fargerike Alnabru	59.927642	10.851513	Furniture / Home Store
3	Alnabru	59.928786	10.837697	Stoff Og Stil	59.929213	10.846513	Arts & Crafts Store
4	Alnabru	59.928786	10.837697	Alna stasjon	59.931997	10.834962	Train Station

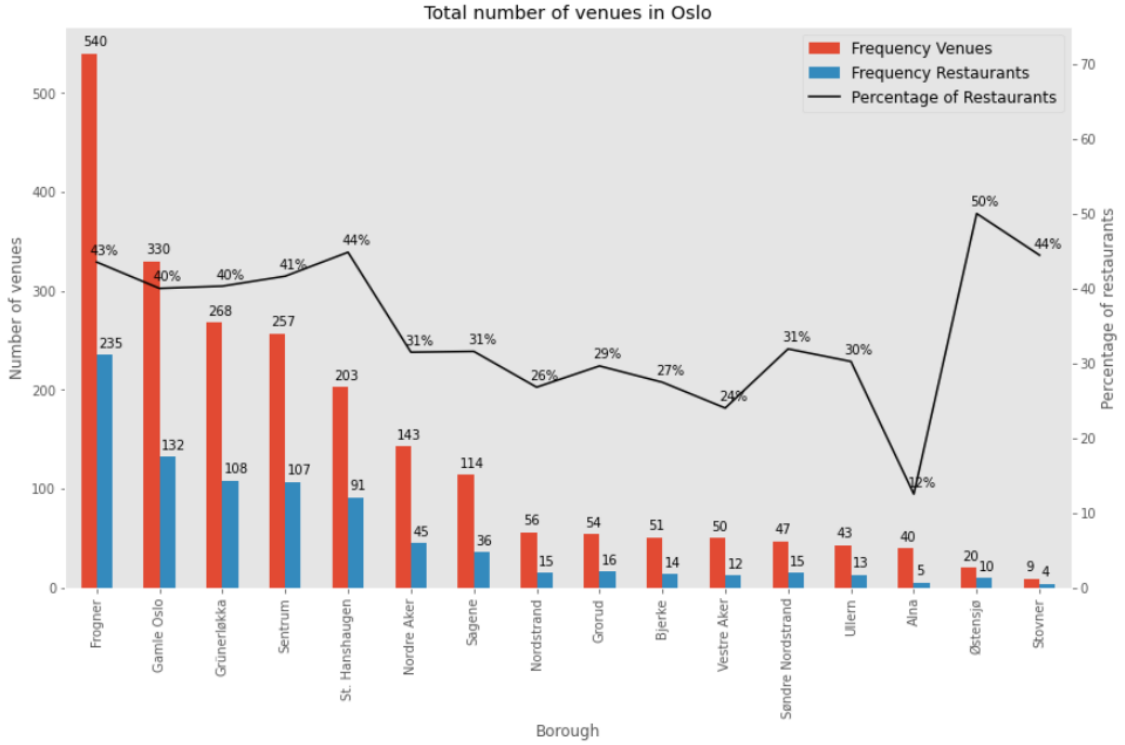
Some venue categories include stations like train, bus, metro, etc. These transportation venues are more or less evenly distributed across the boroughs. In addition, they are not really venues in the broader sense of venues, which consist of restaurants, shopping stores, malls, entertainment facilities etc. Thus, we have decided to remove them from the analysis. After removing the transportation venues the number of unique venue categories drops from 220 and the total number of venues to 2225.

The majority of the venues occur in the borough of Frogner followed by the boroughs of Gamle Oslo, Grunnerløkka, Sentrium and St. Hanshaugen. The boroughs of Østensjø and Stovner, on the other hand, experience the least number of venues. The following bar plot depicts the frequency of venues in the different boroughs of Oslo.

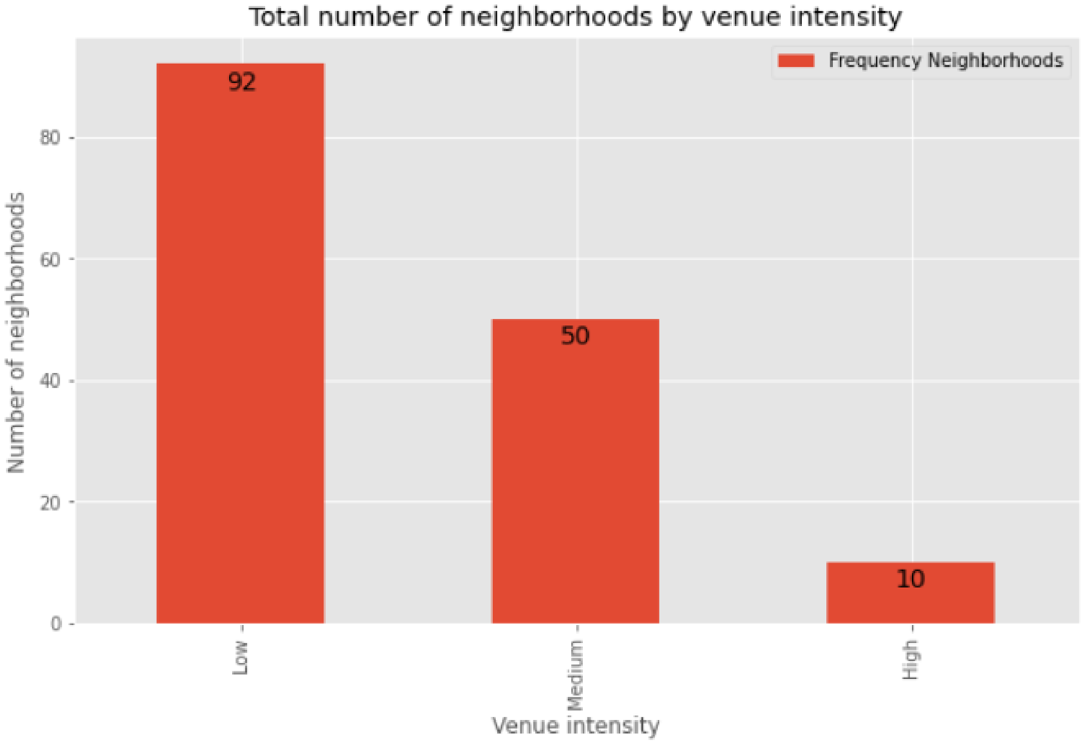


To get a feeling of how many venues are associated with food and drink venues, we create a restaurant category that compromises 76 unique venue categories. The venues in the restaurant category contain key words like cuisine, restaurant, etc.

Then, we computed the frequency of restaurants in each borough as well as the respective ratio between them and the total number of venues. Our results are visualized in the following figure.



In the analysis, the number of venues that are returned for each neighborhood varies significantly. To visualize these differences we generate a new variable called *Venue Intensity* that takes the values low, medium and high, and captures the number of venues within each neighborhood. Neighborhoods with less than 10 venues are categorized as low and with more than 50 as high. The figure below visualizes the frequency of neighborhoods in the data set based on their venue intensity.

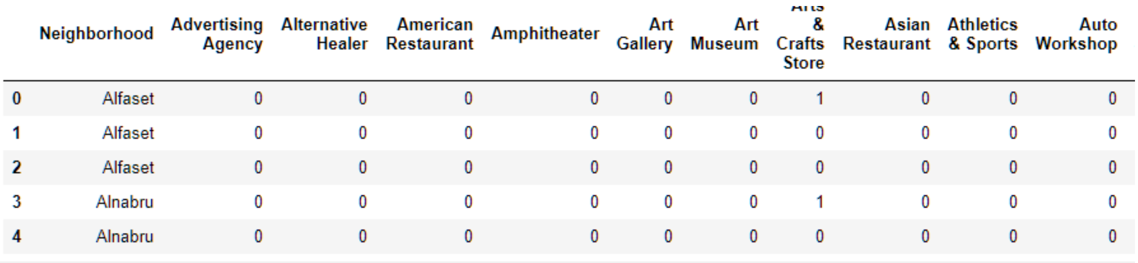


At this point it is important to mention that not all the venues that exist in the different boroughs of Oslo are returned by the Foursquare API inquiry. The inquiry depends on the latitude and longitude information for a given neighborhood and in this analysis we run a single coordinate pair for each neighborhood. It is possible to increase the possibilities by including information like popular streets and areas.

3.2. Clustering neighborhoods with the k-means algorithm

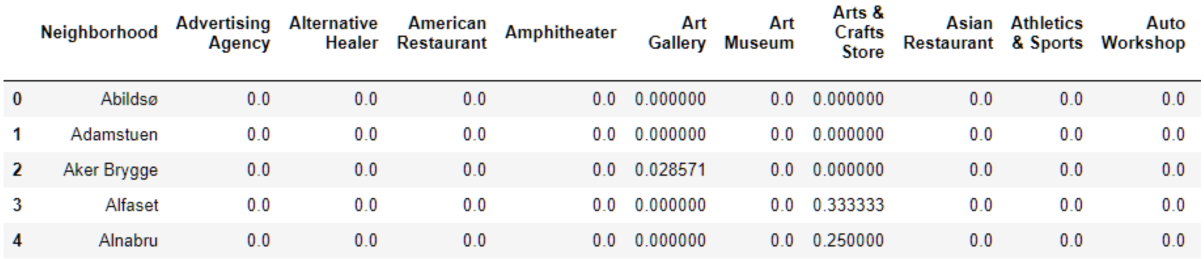
In order to cluster the neighborhoods we will apply the k-means algorithm. This algorithm is one of the most common cluster methods of unsupervised learning.

Before applying it to our data set, we transform the different venue categories to categorical variables using one hot encoding. The head of the new data frame that also includes the name of the neighborhoods is as follows.



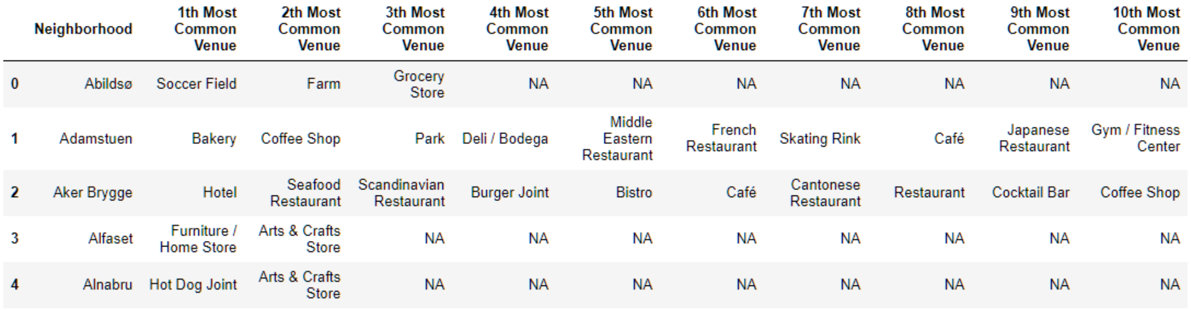
	Neighborhood	Advertising Agency	Alternative Healer	American Restaurant	Amphitheater	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop
0	Alfaset	0	0	0	0	0	0	1	0	0	0
1	Alfaset	0	0	0	0	0	0	0	0	0	0
2	Alfaset	0	0	0	0	0	0	0	0	0	0
3	Alnabru	0	0	0	0	0	0	1	0	0	0
4	Alnabru	0	0	0	0	0	0	0	0	0	0

This data frame consists of 2225 entries or rows and 221 columns. To make more sense of the data, we group the rows by neighborhood and take the mean of the frequency of occurrence of each category. This leads to a reduced in size data frame of 152 rows and 221 columns. The head of the new data frame is shown below.



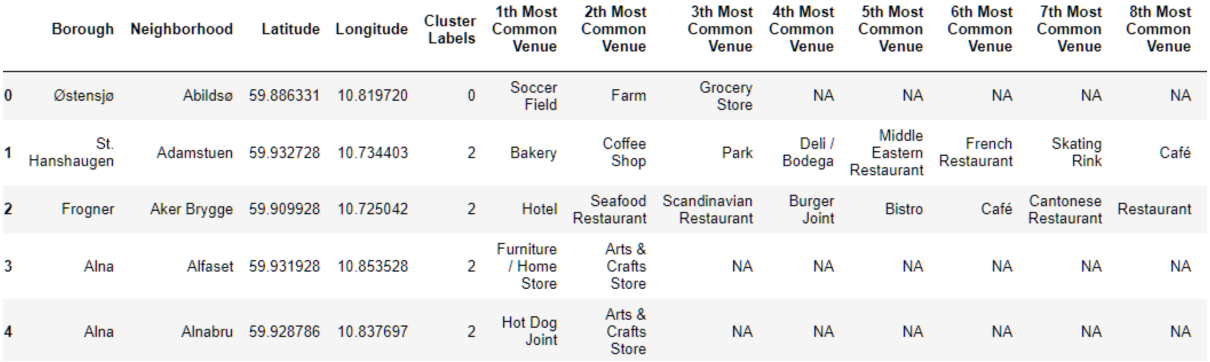
	Neighborhood	Advertising Agency	Alternative Healer	American Restaurant	Amphitheater	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop
0	Abildsø	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0
1	Adamstuen	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0
2	Aker Brygge	0.0	0.0	0.0	0.0	0.028571	0.0	0.000000	0.0	0.0	0.0
3	Alfaset	0.0	0.0	0.0	0.0	0.000000	0.0	0.333333	0.0	0.0	0.0
4	Alnabru	0.0	0.0	0.0	0.0	0.000000	0.0	0.250000	0.0	0.0	0.0

Before proceeding with clustering, we visualize the top 10 venues of each neighborhood. Remember that for the majority of neighborhoods the Foursquare API has return less than 10 venues. The head of the new data frame is as follows.



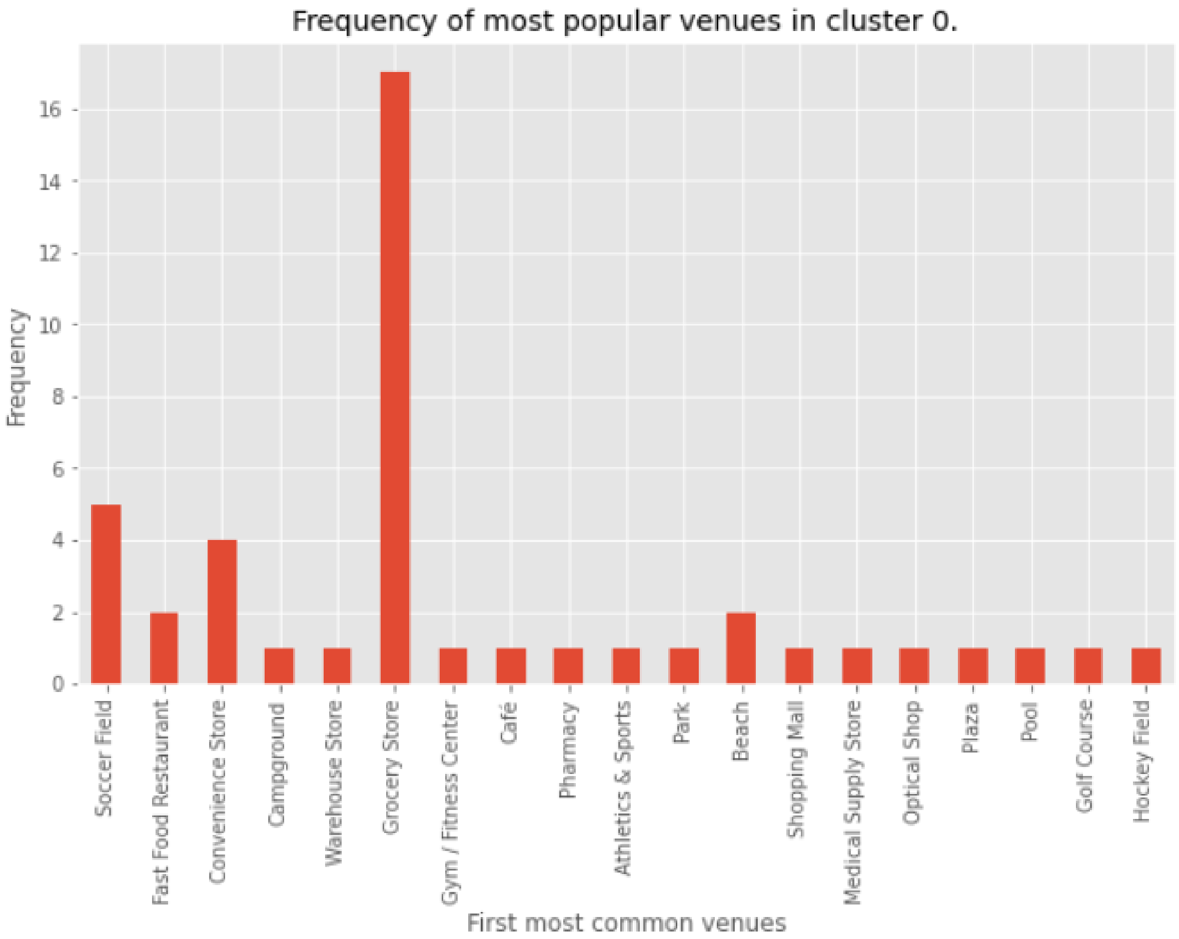
	Neighborhood	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abildsø	Soccer Field	Farm	Grocery Store	NA	NA	NA	NA	NA	NA	NA
1	Adamstuen	Bakery	Coffee Shop	Park	Deli / Bodega	Middle Eastern Restaurant	French Restaurant	Skating Rink	Café	Japanese Restaurant	Gym / Fitness Center
2	Aker Brygge	Hotel	Seafood Restaurant	Scandinavian Restaurant	Burger Joint	Bistro	Café	Cantonese Restaurant	Restaurant	Cocktail Bar	Coffee Shop
3	Alfaset	Furniture / Home Store	Arts & Crafts Store	NA	NA	NA	NA	NA	NA	NA	NA
4	Alnabru	Hot Dog Joint	Arts & Crafts Store	NA	NA	NA	NA	NA	NA	NA	NA

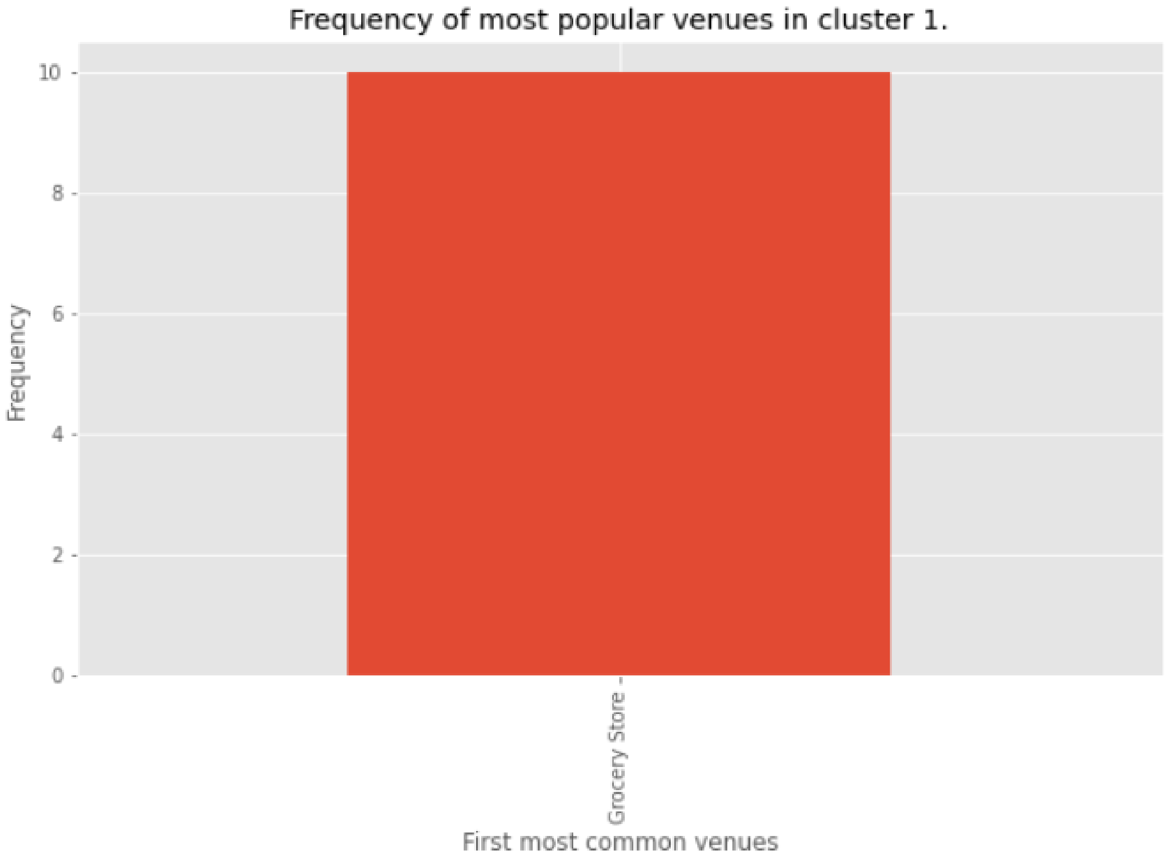
Finally, we run the k-means algorithm to cluster the venues into three clusters. The head of the resulting data frame including the cluster labels for each neighborhood is below.

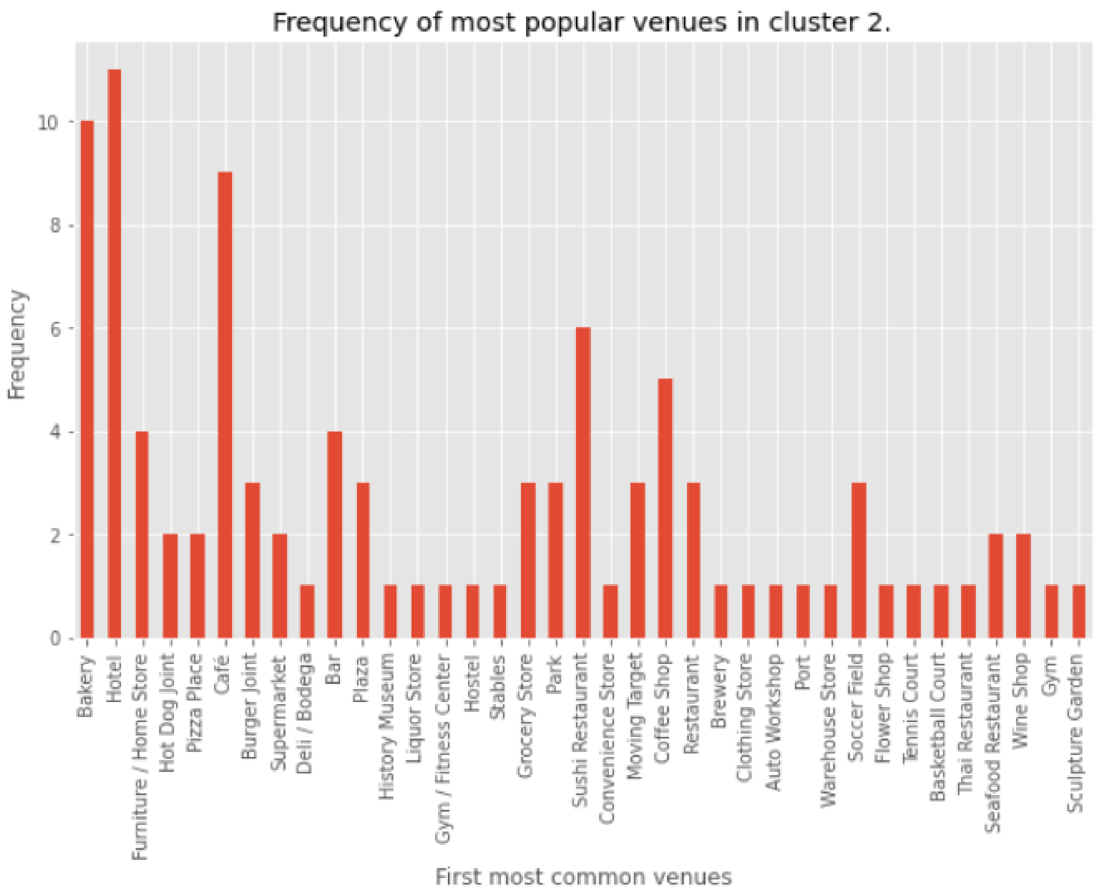


	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Østensjø	Abildsø	59.886331	10.819720	0	Soccer Field	Farm	Grocery Store	NA	NA	NA	NA	NA
1	St. Hanshaugen	Adamstuen	59.932728	10.734403	2	Bakery	Coffee Shop	Park	Deli / Bodega	Middle Eastern Restaurant	French Restaurant	Skating Rink	Café
2	Frogner	Aker Brygge	59.909928	10.725042	2	Hotel	Seafood Restaurant	Scandinavian Restaurant	Burger Joint	Bistro	Café	Cantonese Restaurant	Restaurant
3	Alna	Alfaset	59.931928	10.853528	2	Furniture / Home Store	Arts & Crafts Store	NA	NA	NA	NA	NA	NA
4	Alna	Alnabru	59.928786	10.837697	2	Hot Dog Joint	Arts & Crafts Store	NA	NA	NA	NA	NA	NA

In order to label our clusters, we make a list of the most common venue in each neighborhood and measure the within the cluster frequency. The following plots display our results.


Frequency of most popular venues in cluster 0.
Frequency
16
14
12
10
8
6
4
2
0
Soccer Field
Fast Food Restaurant
Convenience Store
Campground
Warehouse Store
Grocery Store
Gym / Fitness Center
Café
Pharmacy
Athletics & Sports
Park
Beach
Shopping Mall
Medical Supply Store
Optical Shop
Plaza
Pool
Golf Course
Hockey Field
First most common venues

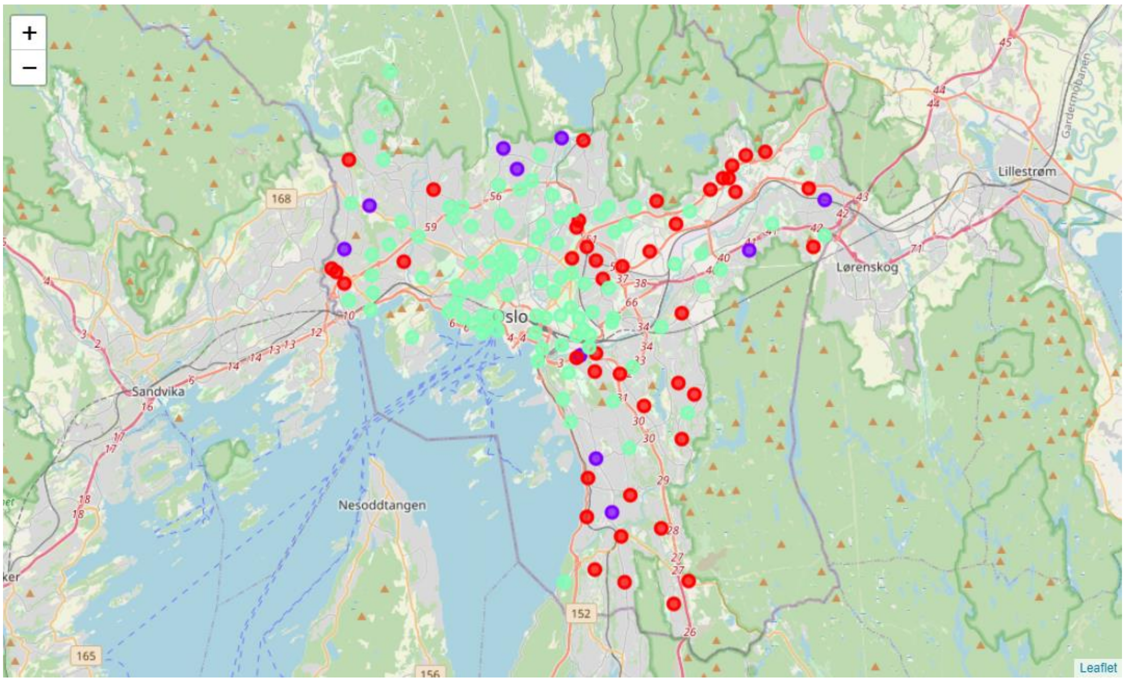

Frequency of most popular venues in cluster 1.
Frequency
10
8
6
4
2
0
Grocery Store
First most common venues



Based on the above results, we can label the three clusters as follows:

- Cluster 0: “Multiple social venues”
- Cluster 1: “Grocery venues”
- Cluster 2: “Accommodation and restaurant venues”

A clustered map of the neighborhoods in Oslo is shown below.



4. Discussion

The purpose of this analysis is to recommend possible places to start up a new restaurant business in the city of Oslo. Oslo is a relative big cities with 15 boroughs and more than 150 neighborhoods. Because of the complexity, different methods can be used for clustering and segmentation. In addition, not all clustering methods can yield the same results for this case.

Here we applied the k-means clustering algorithm with three clusters. We used 152 distinct locations and obtained their venues through the Foursquare API. After examining our data set and segmenting the neighborhoods we find the following:

- In cluster 0, there exist multiple social venues. From the 213 venues in total 62 are restaurants. This means that that the restaurant penetration in that cluster defined as the share of restaurants is 29.11%.
- In cluster 1, there exist mainly grocery shops and zero restaurants.
- In cluster 2, there exist many accommodation and restaurant venues. From the 687 venues in that cluster, 258 are restaurants yielding a restaurant penetration equal to 37.55%.

5. Conclusion

Starting a new business is by no means an easy decision, opening a new restaurant requires serious consideration. In this project, we tried to determine the best locations in the city of Oslo for opening a new restaurant.

The analysis was based on geospatial analysis of the neighborhoods in Oslo. We applied machine learning techniques and in particular unsupervised learning to cluster and segment the neighborhoods in Oslo. Our results indicate the presence of three major clusters with different restaurant penetration rates.