IDENTIFY RESIDENTIAL AREAS IN OSLO

CAPSTONE PROJECT.THE BATTLE OF NEIGHBORHOODS

Antonio Sequeros

July 2020

1.Introduction

Problem Background:

Oslo is relatively small capital, but it has a long history and a charming range of old city quarters which each come with their own distinct character. Oslo is also one of the the fastest-growing capitals in Europe, and in recent years, some of its industrial areas have been turned into the most attractive and modern neighborhoods for its ever-expanding population. The Akerselva River splits Oslo into the western and eastern districts. Officially, the city is divided into 15 boroughs or municipalities, which are largely self-governed. Each is responsible for its own clinics, kindergartens and other public services. The west is where established Norwegian families, the wealthy and most expats live, especially diplomats.

Problem Definition

The goal of this exercise is to identify suitable areas to live in Oslo for a family with children and characterize them in terms of their socioeconomic features.

Target Audience

- The main audience would be people planning or relocating to Oslo specially families as the analysis will be focused in that segment.
- Anybody interested in understanding Oslo
- Data Scientists, who want to implement some of the most used Exploratory Data Analysis techniques to obtain necessary data, analyze it, and, finally be able to tell a story out of it.

2.Data

- We will get The Names of Major Districts and Population from Wikipedia; https://en.wikipedia.org/wiki/List_of_boroughs_of_Oslo'
- Geopy will be used to geolocate these districts
- Foursquare API will be used to explore neighborhoods in Oslo, get the most common venue categories in each neighborhood, use the k-means clustering algorithm to find similar neighborhoods, use the Folium library to visualize the neighborhoods in Oslo and their emerging clusters.
- Finally we will used data downloaded from Norway statistics department to further characterize the districts of interest. https://data.ssb.no/api/v0/dataset?lang=en. the data will be downloaded and manipulated separately as it goes beyond this exercise the use of the provided API

3.Methodology

First I scrapped a list of neighborhoods in Oslo from wikipedia and added an additional neighborhood that though is not strictly in Oslo municipality its in practical terms part of the bigger Oslo area as many people live there. I did some cleaning on the data and saved to a dataframe.

	Borough	Residents
0	Alna	49 801
1	Bjerke	33 422
2	Frogner	59 269
3	Gamle Oslo	58 671
4	Grorud	27 707
5	Grünerløkka	62 423
6	Nordre Aker	52 327
7	Nordstrand	52 459
8	Sagene	45 089
9	St. Hanshaugen	38 945
10	Stovner	33 316
11	Søndre Nordstrand	39 066
12	Ullern	34 596
13	Vestre Aker	50 157
14	Østensjø	50 806
15	Sandvika	121000

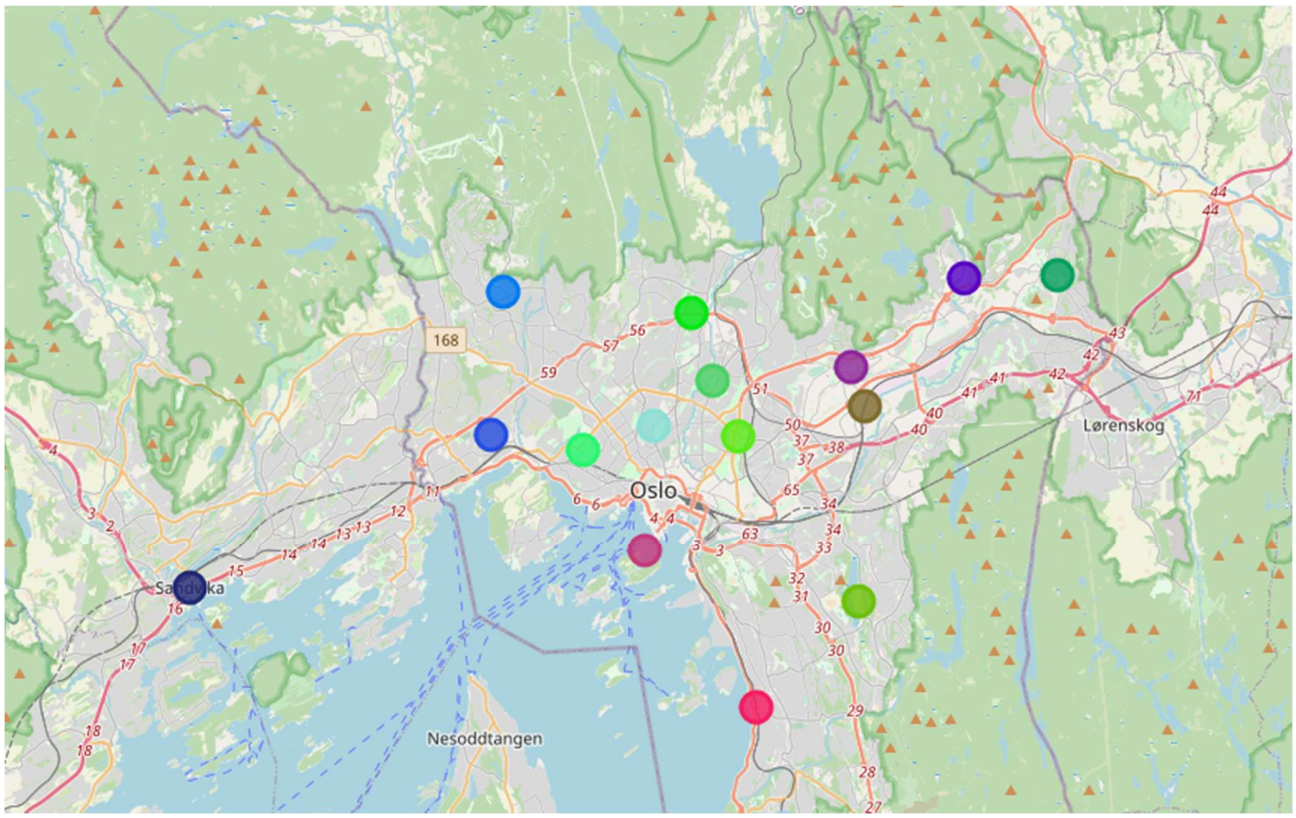
The second step was to add geographical coordinates to these neighborhoods using geopy library. The coordinates of Nordstrand were not correct so that was amended modifying the dataframe and inserting the right ones

	Borough	Residents	Latitude	Longitude
0	Alna	49 801	59.932417	10.835276
1	Bjerke	33 422	59.941395	10.829208
2	Frogner	59 269	59.922224	10.706649
3	Gamle Oslo	58 671	59.899237	10.734767
4	Grorud	27 707	59.961424	10.880549
5	Grünerløkka	62 423	59.925471	10.777421
6	Nordre Aker	52 327	59.953638	10.756412
7	Nordstrand	52 459	54.487378	8.865286
8	Sagene	45 089	59.938273	10.765849
9	St. Hanshaugen	38 945	59.927950	10.738958
10	Stovner	33 316	59.962140	10.922823
11	Søndre Nordstrand	39 066	59.835944	10.798496
12	Ullern	34 596	59.925818	10.665132
13	Vestre Aker	50 157	59.958300	10.670319
14	Østensjø	50 806	59.887563	10.832748

	Borough	Residents	Latitude	Longitude
0	Alna	49 801	59.932417	10.835276
1	Bjerke	33 422	59.941395	10.829208
2	Frogner	59 269	59.922224	10.706649
3	Gamle Oslo	58 671	59.899237	10.734767
4	Grorud	27 707	59.961424	10.880549
5	Grünerløkka	62 423	59.925471	10.777421
6	Nordre Aker	52 327	59.953638	10.756412
7	Nordstrand	52 459	59.863525	10.785830
8	Sagene	45 089	59.938273	10.765849
9	St. Hanshaugen	38 945	59.927950	10.738958
10	Stovner	33 316	59.962140	10.922823
11	Søndre Nordstrand	39 066	59.835944	10.798496
12	Ullern	34 596	59.925818	10.665132
13	Vestre Aker	50 157	59.958300	10.670319
14	Østensjø	50 806	59.887563	10.832748
15	Sandvika	121000	59.890726	10.527743

For a better understanding of Oslo Geography and confirming everything was correct I visualize the neighborhoods in a Map using folium library.

OSLO NEIGHBORHOODS



Next, I used the Foursquare API to explore the neighborhoods. After creating Foursquare credentials, created a get request url using Foursquare ID to look for a maximum of 500 venues within 1000 meters of the geographical coordinates of each neighborhood.

The Foursquare API call resulted in 590 venues that were saved in a dataframe.

	District	Dist_Latitude	Dist_Longitude	Venue	Venue_Lat	Venue_Long	Venue_Category
587	Sandvika	59.890726	10.527743	Lakseberget	59.891921	10.536710	Harbor / Marina
588	Sandvika	59.890726	10.527743	Bergensbanen	59.894937	10.532205	Moving Target
589	Sandvika	59.890726	10.527743	Bærum Roklubb	59.885958	10.535703	Harbor / Marina

Overall there were 143 different venue categories and most of the venues were in a few districts and some neighborhoods had very few datapoints as illustrated below.

	District	Venue
0	Alna	24
1	Bjerke	10
2	Frogner	100
3	Gamle Oslo	24
4	Grorud	11
5	Grünerløkka	76
6	Nordre Aker	27
7	Nordstrand	22
8	Sagene	83
9	Sandvika	29
10	St. Hanshaugen	100
11	Stovner	6
12	Søndre Nordstrand	13
13	Ullern	20
14	Vestre Aker	27
15	Østensjø	18

```
Grocery Store     41
Café              36
Bakery            28
Coffee Shop       25
Park              17
                  ..
Golf Course        1
Moving Target      1
Beer Bar           1
Trail              1
Creperie           1
Name: Venue_Category, Length: 143, dtype: int64
```

Dataset preparation for analysis

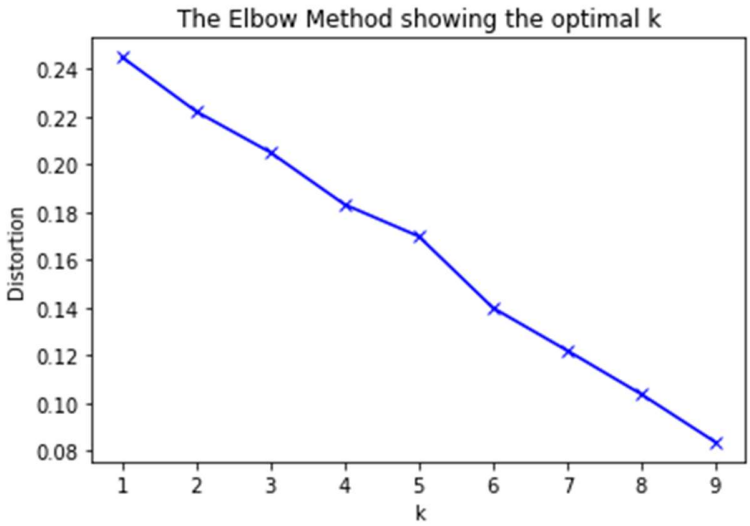
I used one hot encoding to move venue categories from rows to columns grouped rows by neighborhoods and calculated the mean of the frequency of occurrence of each category.

	District	Advertising Agency	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop	Automotive Shop	BBQ Joint	Bakery	...	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vie R
1	Alna	0.0	0.00	0.041667	0.000000	0.000000	0.041667	0.041667	0.0	0.041667	...	0.041667	0.0	0.041667	0.00	0.0	
2	Bjerke	0.0	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.00	0.0	
3	Frogner	0.0	0.01	0.000000	0.030000	0.000000	0.000000	0.000000	0.0	0.070000	...	0.000000	0.0	0.000000	0.01	0.0	
4	Gamle Oslo	0.0	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.00	0.0	
5	Grorud	0.0	0.00	0.000000	0.090909	0.090909	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.00	0.0	

5 rows × 144 columns

Clustering.

In order to identify an optimal number of clusters I used the Elbow method but the results were not conclusive, indicating that maybe K-means was not the best suited algorithm for this data. Nevertheless and for the sake of the learning exercise I chose K=5.

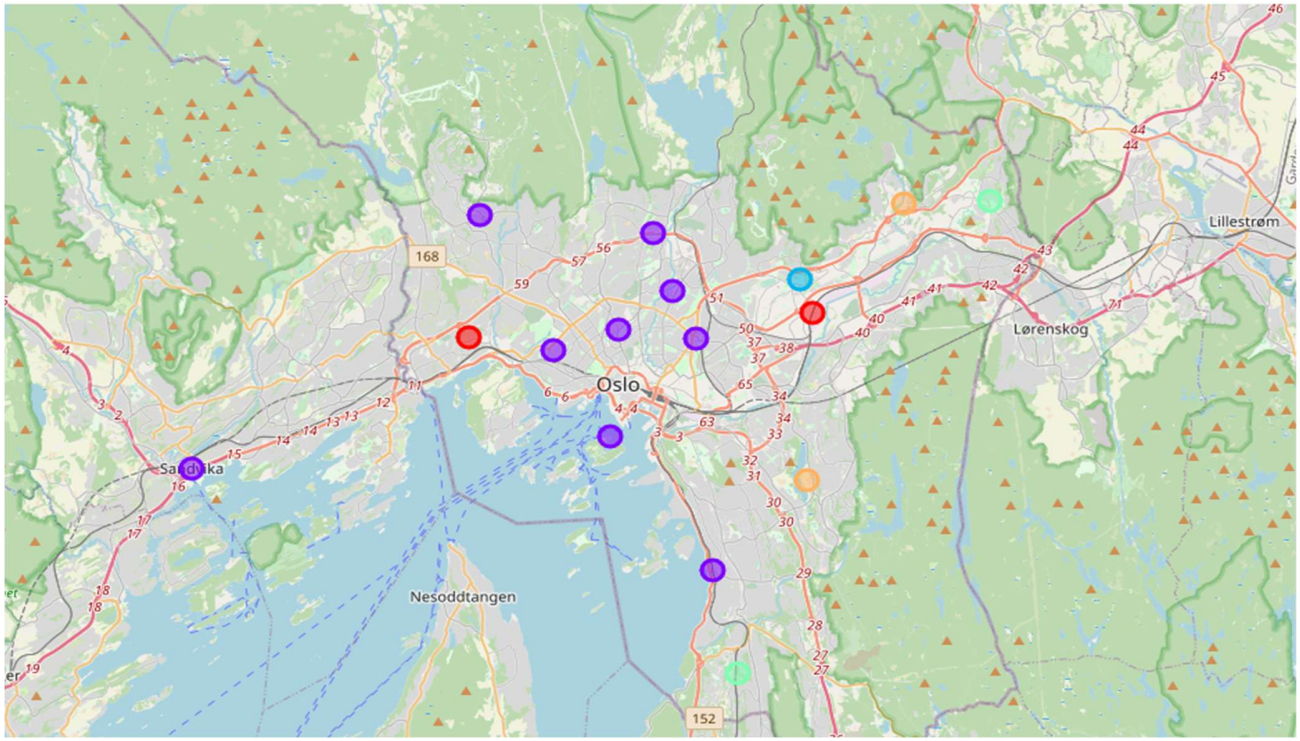


I created a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood and merged it with our original df with districts and coordinates

Cluster Labels	District	1st Most Common Venue	2nd Most Common Venue	3rd 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	Residents	Latitu
0	Alna	Furniture / Home Store	Metro Station	Grocery Store	Bus Station	Pet Store	Spanish Restaurant	Market	Hotel	Bakery	Toy / Game Store	49 801	59.9324
2	Bjerke	Grocery Store	Gym / Fitness Center	Farm	Hotel	Supermarket	Pizza Place	Café	Yoga Studio	Dog Run	Falafel Restaurant	33 422	59.9413
1	Frogner	Café	Bakery	Coffee Shop	Hotel	Scandinavian Restaurant	Indian Restaurant	Pizza Place	Park	Pub	Burger Joint	59 269	59.9222
1	Gamle Oslo	Boat or Ferry	Scandinavian Restaurant	Castle	Mexican Restaurant	Bathing Area	Other Nightlife	Chinese Restaurant	Seafood Restaurant	Café	Burger Joint	58 671	59.8992
4	Grorud	Metro Station	Wine Shop	Asian Restaurant	Athletics & Sports	Bus Station	Gym	Grocery Store	Pizza Place	Convenience Store	Supermarket	27 707	59.9614

I visualized the clusters in a map

OSLO CLUSTERS



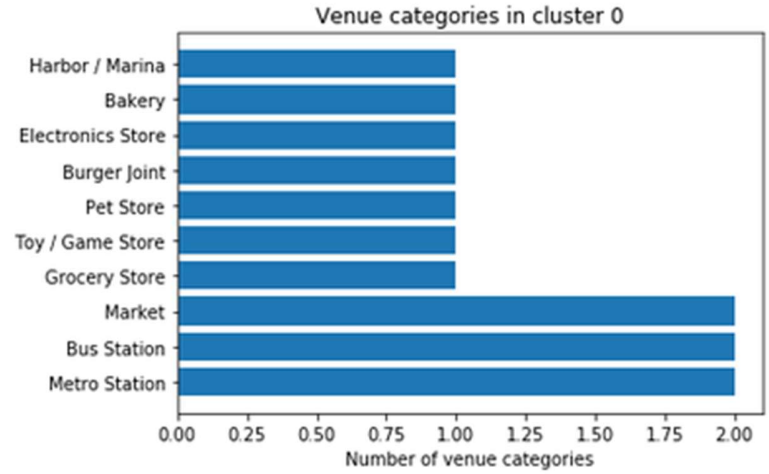
4. Results

In order to understand what type of neighborhoods there was in each cluster I analyzed the prevalence of venues and tried to reach some conclusions.

Cluster 0. Transportation hub

Cluster 0 includes 2 neighborhoods and seems to have a prevalence of transportation infrastructure like bus or metro station plus services for residents like markets and groceries

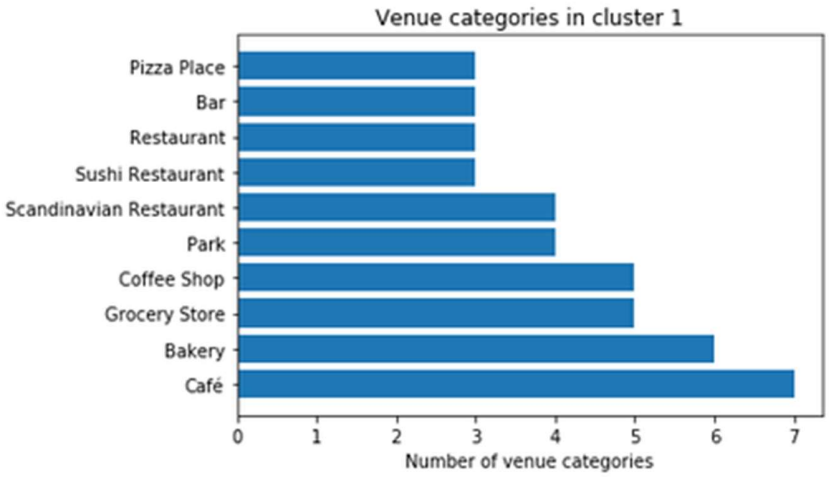
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Residents	Latitude	Longitude
1	Alna	Furniture / Home Store	Metro Station	Grocery Store	Bus Station	Pet Store	Spanish Restaurant	Market	Hotel	Bakery	Toy / Game Store	49 801	59.932417	10.835276
14	Ullern	Bus Station	Metro Station	Market	Light Rail Station	Harbor / Marina	Coffee Shop	Flower Shop	Electronics Store	Italian Restaurant	Burger Joint	34 596	59.925818	10.665132



Cluster 1. Residential

Cluster 1 includes 9 neighborhoods. This cluster comprises districts relatively close to the urban centers and having a good variety of services and with prevalence of Coffee shops, bakeries and grocery stores that can be indicative of affluent urban residential areas

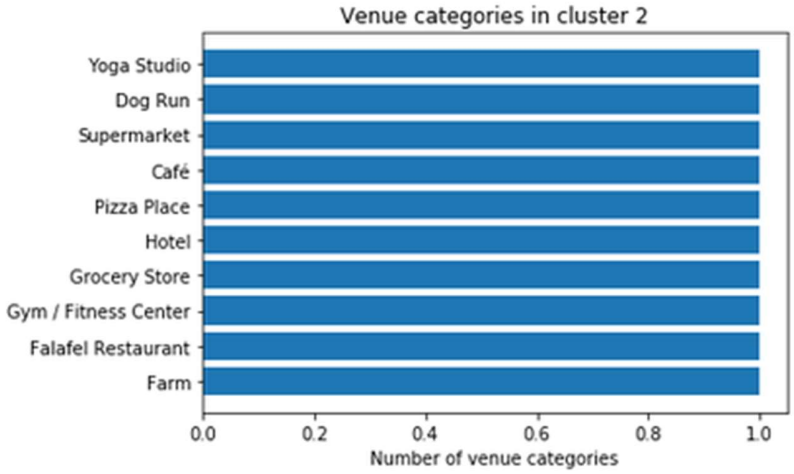
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Residents	La
3	Frogner	Café	Bakery	Coffee Shop	Hotel	Scandinavian Restaurant	Indian Restaurant	Pizza Place	Park	Pub	Burger Joint	59 269	59.9
4	Gamle Oslo	Boat or Ferry	Scandinavian Restaurant	Castle	Mexican Restaurant	Bathing Area	Other Nightlife	Chinese Restaurant	Seafood Restaurant	Café	Burger Joint	58 671	59.8
6	Grünerløkka	Grocery Store	Café	Bus Station	Park	Bakery	Bar	Gym / Fitness Center	Coffee Shop	Asian Restaurant	Botanical Garden	62 423	59.9
7	Nordre Aker	Bakery	Bus Stop	Gym	Grocery Store	Metro Station	Shopping Mall	Bus Station	Advertising Agency	Sushi Restaurant	Hotel	52 327	59.9
8	Nordstrand	Grocery Store	Beach	Fast Food Restaurant	Shopping Mall	Bakery	Supermarket	Sushi Restaurant	Juice Bar	Convenience Store	Restaurant	52 459	59.8
9	Sagene	Café	Sushi Restaurant	Coffee Shop	Park	Pizza Place	Bar	Bakery	Brewery	Gym	Grocery Store	45 089	59.9
10	Sandvika	Coffee Shop	Café	Restaurant	Harbor / Marina	Electronics Store	Fast Food Restaurant	Movie Theater	Moving Target	Sporting Goods Shop	Beach	121000	59.8
11	St. Hanshaugen	Bakery	Café	Scandinavian Restaurant	Coffee Shop	Park	Pizza Place	Indian Restaurant	Bar	Gym / Fitness Center	Thai Restaurant	38 945	59.9
15	Vestre Aker	Grocery Store	Ski Area	Metro Station	Restaurant	Café	Soccer Field	Lake	Scandinavian Restaurant	Disc Golf	Museum	50 157	59.9



Cluster 2. Residential Suburban

Cluster 2 only includes one neighborhood with few so we cannot really infer much about it

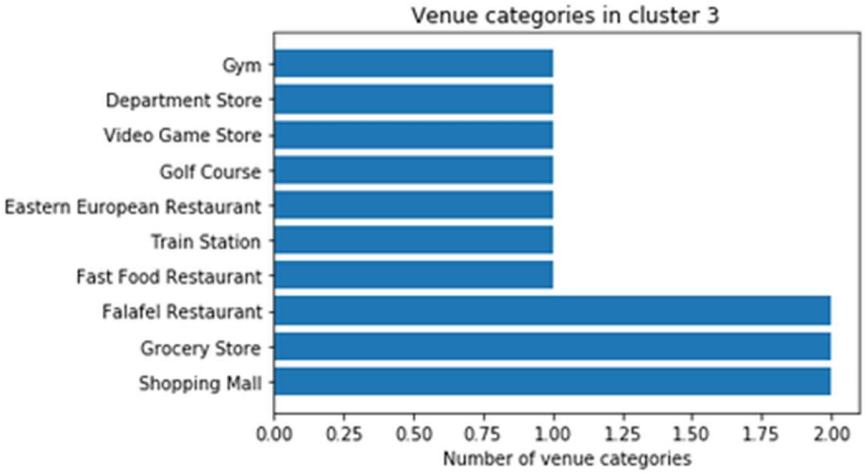
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Residents	Latitude	Longitude
2	Bjerke	Grocery Store	Gym / Fitness Center	Farm	Hotel	Supermarket	Pizza Place	Café	Yoga Studio	Dog Run	Falafel Restaurant	33 422	59.941395	10.829208



Cluster 3. Residential suburban Immigrants

Including two suburban districts include suburban districts with prevalence of grocery stores and shopping mall. Presence of foreign restaurants might be indicative of immigrant population.

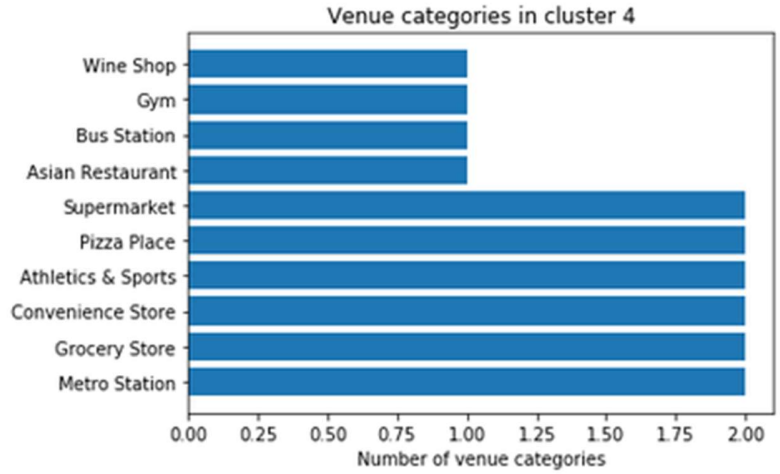
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Residents	Latitude	Longit
12	Stovner	Gas Station	Video Game Store	Grocery Store	Department Store	Golf Course	Shopping Mall	Yoga Studio	Falafel Restaurant	Electronics Store	Eastern European Restaurant	33 316	59.962140	10.922
13	Søndre Nordstrand	Grocery Store	Shopping Mall	Fast Food Restaurant	Gym	Athletics & Sports	Train Station	Stadium	Pharmacy	Farm	Falafel Restaurant	39 066	59.835944	10.798



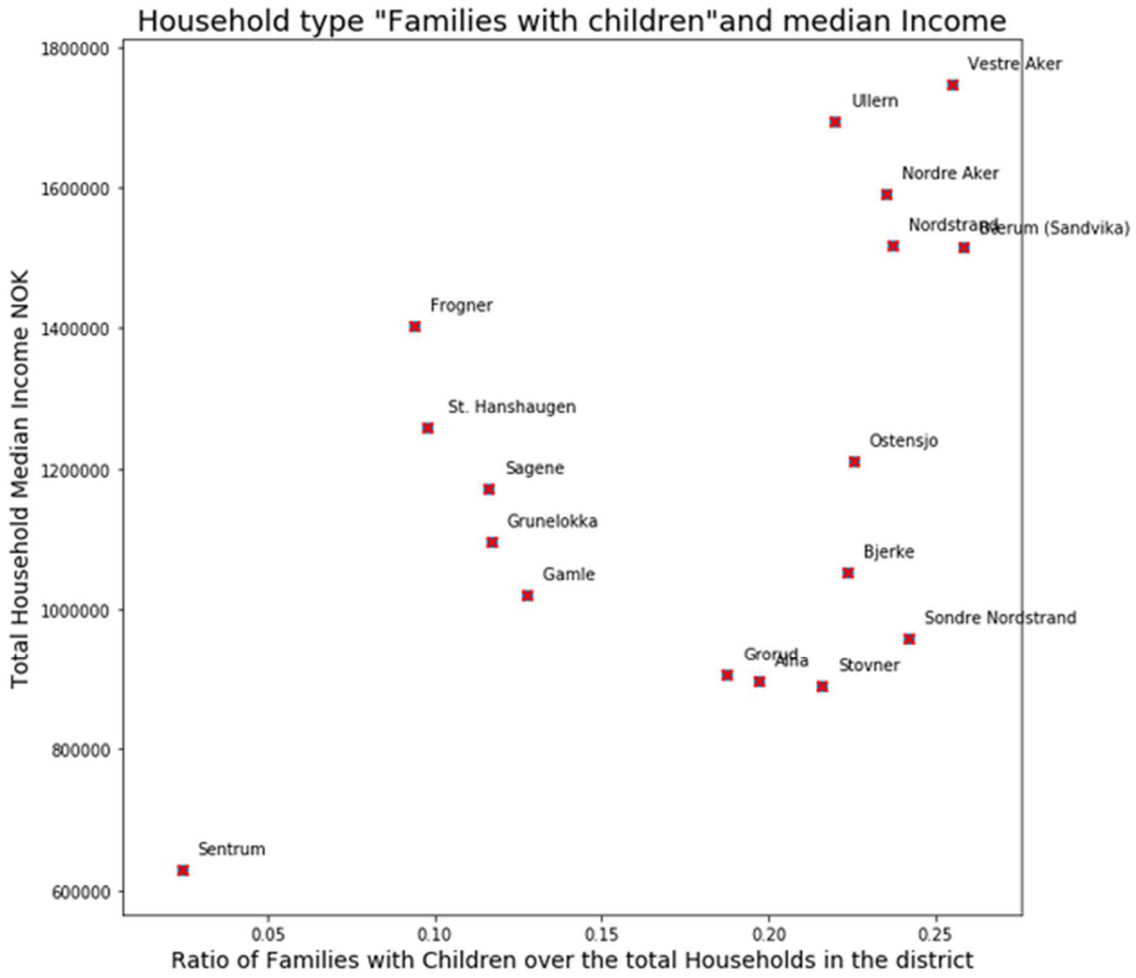
Cluster 4. Residential Suburban

Cluster 4 Prevalence of grocery shops, supermarkets, convenience store indicating residential areas

	District	1st Most Common Venue	2nd Most Common Venue	3rd 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	Residents	Latitude	Long
5	Grorud	Metro Station	Wine Shop	Asian Restaurant	Athletics & Sports	Bus Station	Gym	Grocery Store	Pizza Place	Convenience Store	Supermarket	27 707	59.961424	10.8
16	Østensjø	Metro Station	Athletics & Sports	Shopping Mall	Yoga Studio	Supermarket	Grocery Store	Lake	Pizza Place	Convenience Store	Burger Joint	50 806	59.887563	10.8



Additionally, I downloaded socioeconomic data from Norway statistics department converted it to a pandas data frame and after some cleaning I created a scatter chart to visualize average household income and ration of families with children over total households of the district and related it to the clustering results.


Household type "Families with children"and median Income
Total Household Median Income NOK
Ratio of Families with Children over the total Households in the district
1800000
1600000
1400000
1200000
1000000
800000
600000
0.05
0.10
0.15
0.20
0.25
Vestre Aker
Ullern
Nordre Aker
Nordstrand
Bærum (Sandvika)
Frogner
St. Hanshaugen
Ostensjo
Sagene
Grunelokka
Bjerke
Gamle
Sondre Nordstrand
Grorud
Alna
Stovner
Sentrum

4. Results

Through the foursquare location analysis, we identified 3 potential residential clusters one urban and three suburban. When looking at socioeconomic data we can further characterize the districts within the clusters

Cluster 0. This is what i called Transportation center due to the presence in transportation infrastructure however the two districts included here are reasonable locations for families though in opposite sides of the income scale while Ullern is in the high-income side and similar to the districts in cluster 0, Alna would be in the lowest income group similar to location in cluster 2. In my opinion these are examples of the limits of k-clustering with the limited data we have

Cluster 1. This group corresponds with groups 1 and 3, middle and high income locations The most attractive locations for families would be as they present the highest proportions of families with children. These districts also correspond to highest household income

- Nordre Aker
- Nordstrand
- Sandvika
- Vestre Aker
- Ostenjo

The rest of the districts in the cluster would be more interesting for households with one person as most of the households are single persons which probably corresponds to students and young professionals

Clusters 2,3 and 4 correspond almost perfectly with group 3 middle and low income.

Cluster 2. Also identified as residential suburban areas. Bjerke seems to be a location of choice for middle income families

Cluster 3. Residential suburban with immigrants. The conclusion obtained from the location analysis is confirmed when looking at socioeconomic data. Stovner and Sondre Nordstrand are indeed suburban residential areas with high % of families with children. These are also districts with the lowest household income and the highest percentage of immigrant population as stated in: https://www.ssb.no/en/befolkning/artikler-og-publikasjoner/14-per-cent-of-population-are-immigrants

Cluster 4 Residential suburban. Grorud ,Ostenjo

5. Discussion

This was a good project to complete the Coursera Capstone course and was definitely an excellent learning exercise as to complete it i had to deal and figure out how to solve a number of issues. From the results standpoint there are some positives and some negatives.

- The data points from Foursquare API are limited and thus its is not really possible to reach strong conclusions out of it and should be taken with a pinch of salt. Another approach would be to do the same exercise with different data sources like google maps API. In other cities where Foursquare is more popular results might be better
- K means clustering is not the most relevant approach for clustering given the available data as the Elbow method didn't show any differences in the number of clusters. Number 5 was chosen for the sake of the learning exercise and though some clusters identified really had some consistency like Cluster 1, it also resulted in confusing results like with cluster 0
- The clustering is done based on the presence of similar categories, however a more focused analysis choosing only specific categories of interest might render better results.
- It was not possible for me to obtain a geojson dataset with Oslo districts boundaries. Should it be available it could be used to better show the socioeconomic features together with the clusters
- Interestingly the most affluent districts in Oslo have the highest number of datapoints mainly corresponding to Cluster0.
- Clustering did a decent job if we consider income criteria. It did separate higher income areas from middle and lower income

6. Conclusion

To conclude this project i will say that it was definitely a very good learning exercise were i had to use different libraries to scrape websites, manipulate data frames, visualize maps, use an external API, conduct clustering analysis using k-means method and plot charts. There are significant gaps on the quality of the data and how defensible the conclusions would be using only Foursquare data. Nevertheless, as mentioned in the previous section the results are not completely off the reality of Oslo