The battle of neighborhoods

Moving to Oslo

02/08/2020

IBM Professional Data Scientist Specialization – Capstone Project

Antonio Sequeros

1.Problem Definition

Problem Background:

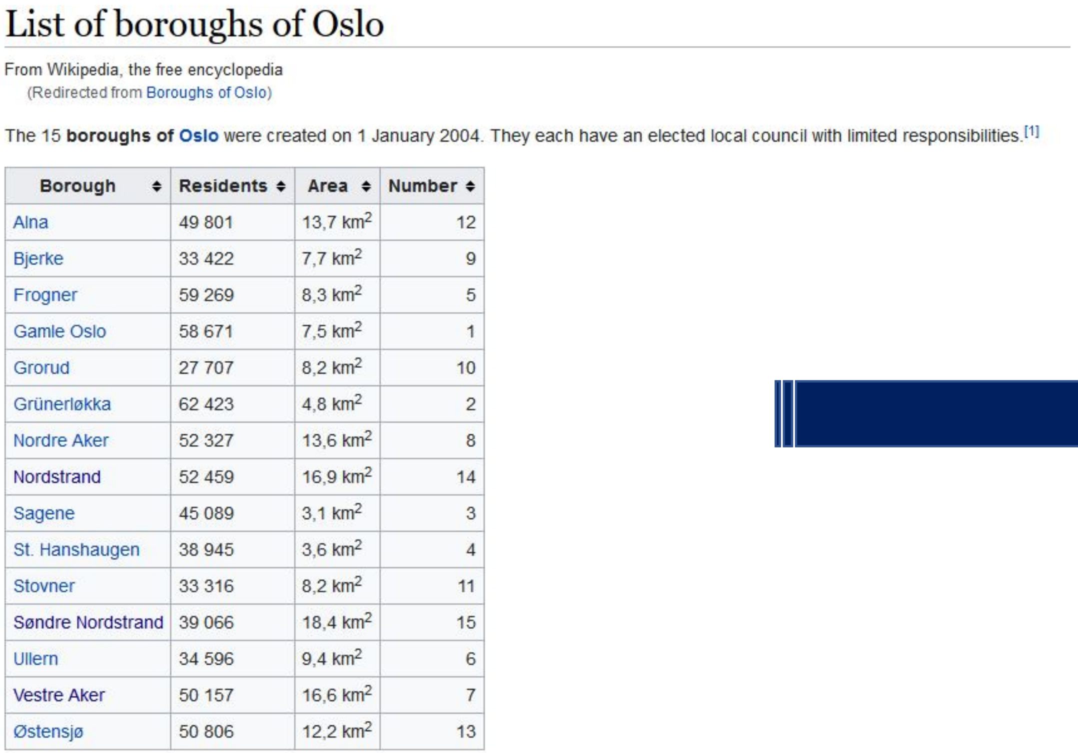
Oslo is relatively small capital, but it has a long history and a charming range of old city quarters which each come with their own distinct character. Oslo is also one of the fastest-growing capitals in Europe, and in recent years, some of its industrial areas have been turned into the most attractive 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. Wikipedia is the first data source

- 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

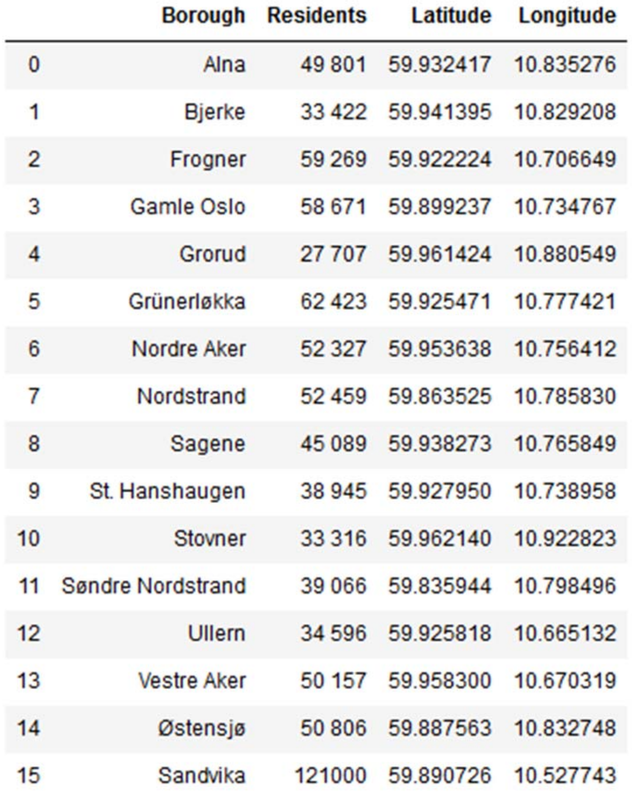


List of boroughs of Oslo

From Wikipedia, the free encyclopedia
(Redirected from Boroughs of Oslo)

The 15 **boroughs of Oslo** were created on 1 January 2004. They each have an elected local council with limited responsibilities.[1]

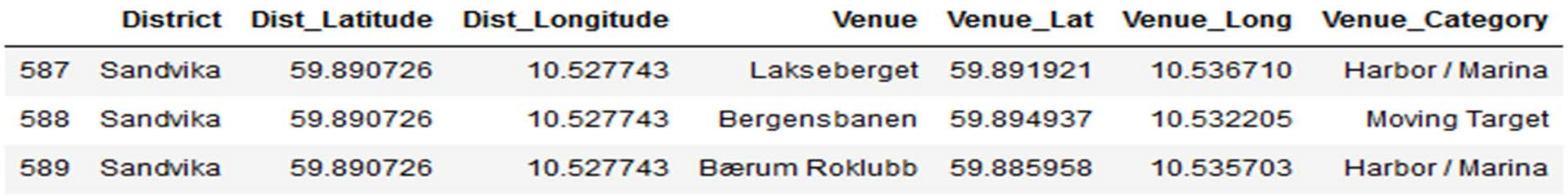
Borough	Residents	Area	Number
Alna	49 801	13,7 km²	12
Bjerke	33 422	7,7 km²	9
Frogner	59 269	8,3 km²	5
Gamle Oslo	58 671	7,5 km²	1
Grorud	27 707	8,2 km²	10
Grünerløkka	62 423	4,8 km²	2
Nordre Aker	52 327	13,6 km²	8
Nordstrand	52 459	16,9 km²	14
Sagene	45 089	3,1 km²	3
St. Hanshaugen	38 945	3,6 km²	4
Stovner	33 316	8,2 km²	11
Søndre Nordstrand	39 066	18,4 km²	15
Ullern	34 596	9,4 km²	6
Vestre Aker	50 157	16,6 km²	7
Østensjø	50 806	12,2 km²	13



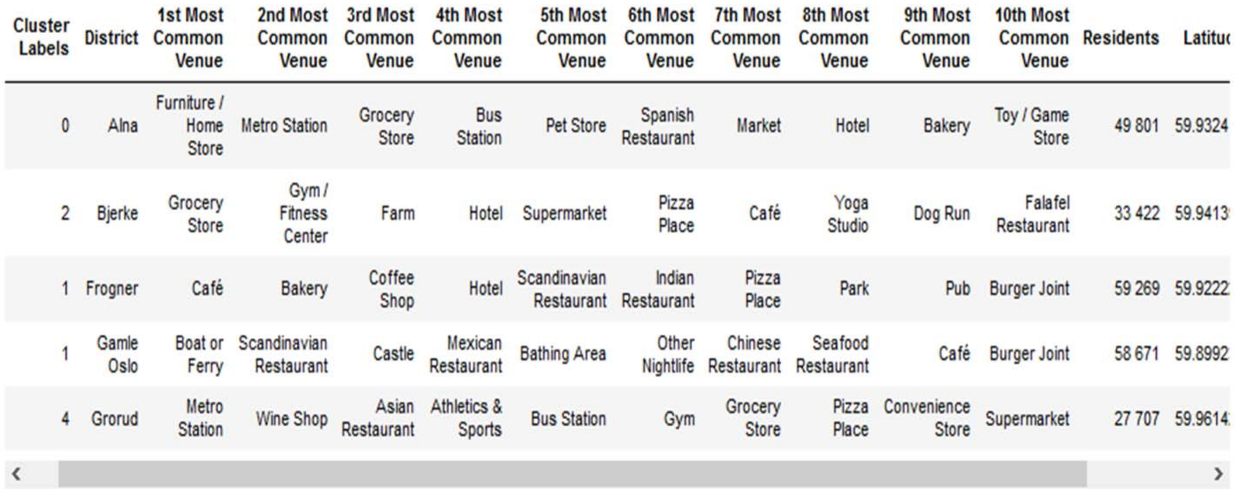
	Borough	Residents	Latitude	Longitude
0	Alna	49 801	59.932417	10.835276
1	Bjerke	33 422	59.941395	10.829208
2	Frogner	59 269	59.922224	10.706649
3	Gamle Oslo	58 671	59.899237	10.734767
4	Grorud	27 707	59.961424	10.880549
5	Grünerløkka	62 423	59.925471	10.777421
6	Nordre Aker	52 327	59.953638	10.756412
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

2. Data. Foursquare venues …

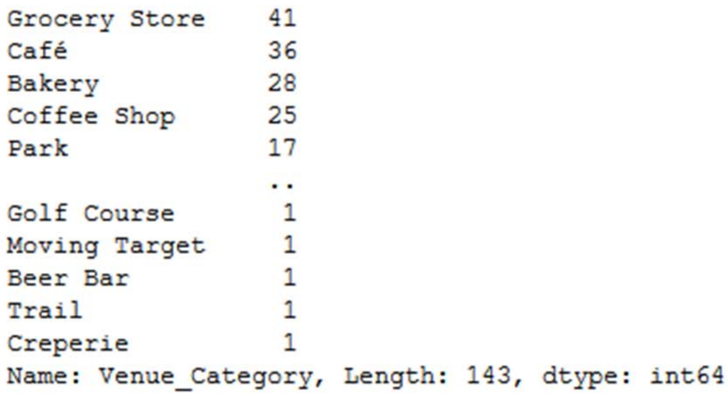
- Foursquare Venues
  - API will be used to explore neighborhoods in Oslo, get the most common venue categories in each neighborhood



	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



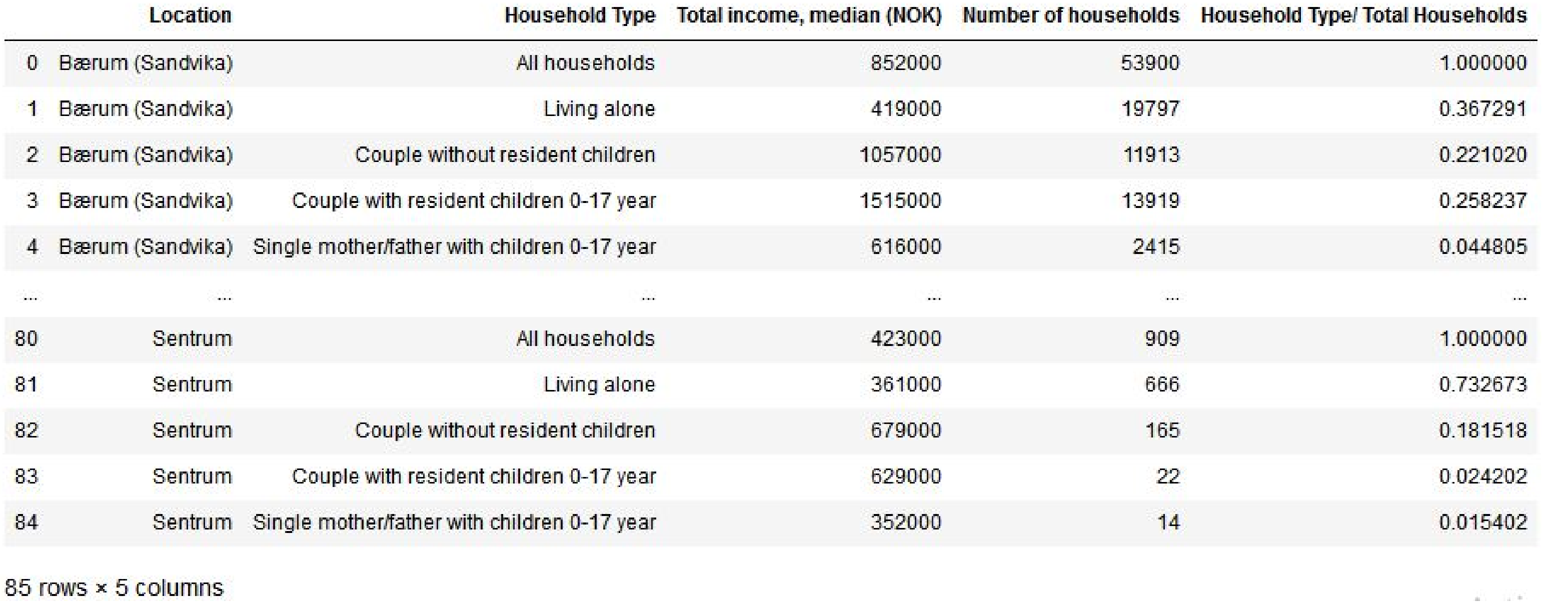
Cluster Labels	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Residents	Latitu
0	Alna	Furniture / Home Store	Metro Station	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



```
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
```

2. Data. ...and Norway statistics data

- Socioeconomic data
  - We will download data 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

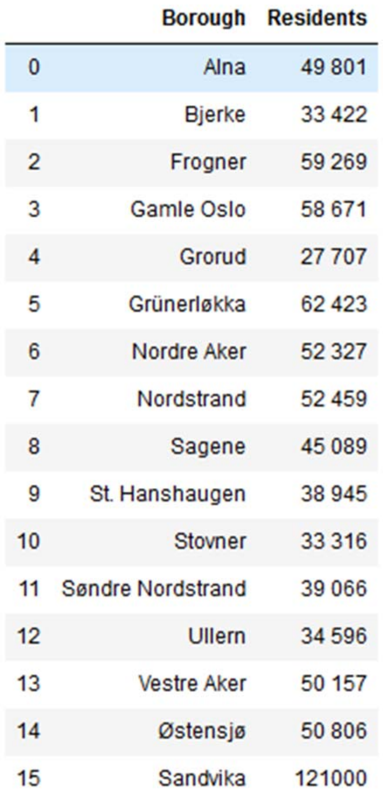


	Location	Household Type	Total income, median (NOK)	Number of households	Household Type/ Total Households
0	Bærum (Sandvika)	All households	852000	53900	1.000000
1	Bærum (Sandvika)	Living alone	419000	19797	0.367291
2	Bærum (Sandvika)	Couple without resident children	1057000	11913	0.221020
3	Bærum (Sandvika)	Couple with resident children 0-17 year	1515000	13919	0.258237
4	Bærum (Sandvika)	Single mother/father with children 0-17 year	616000	2415	0.044805
...	...	...	...	...	...
80	Sentrum	All households	423000	909	1.000000
81	Sentrum	Living alone	361000	666	0.732673
82	Sentrum	Couple without resident children	679000	165	0.181518
83	Sentrum	Couple with resident children 0-17 year	629000	22	0.024202
84	Sentrum	Single mother/father with children 0-17 year	352000	14	0.015402

85 rows × 5 columns

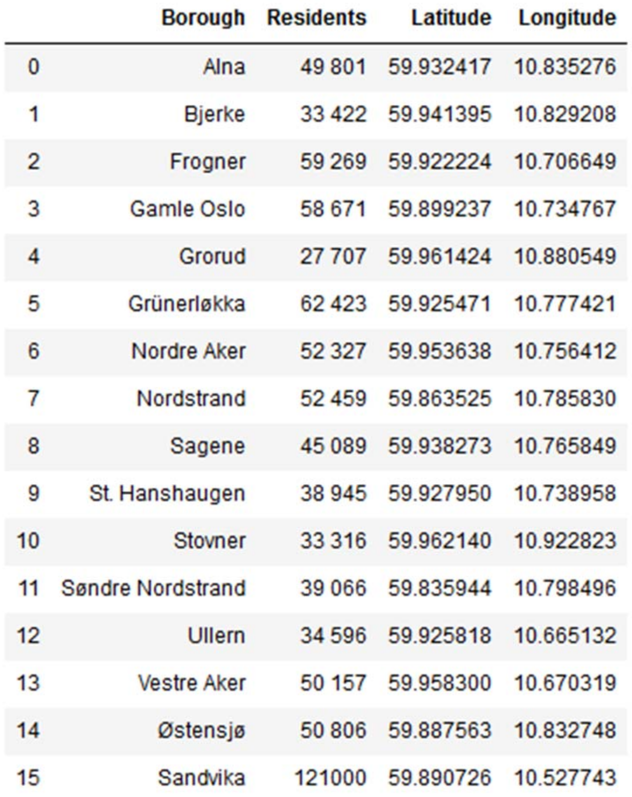
3. Methodology. Scrap borough data and geocode

- Scrap Oslo Neighborhoods from Wikipedia
  - Clean and Add Sandvika



	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

- Geocode using Geopy library
  - Amend mistakes

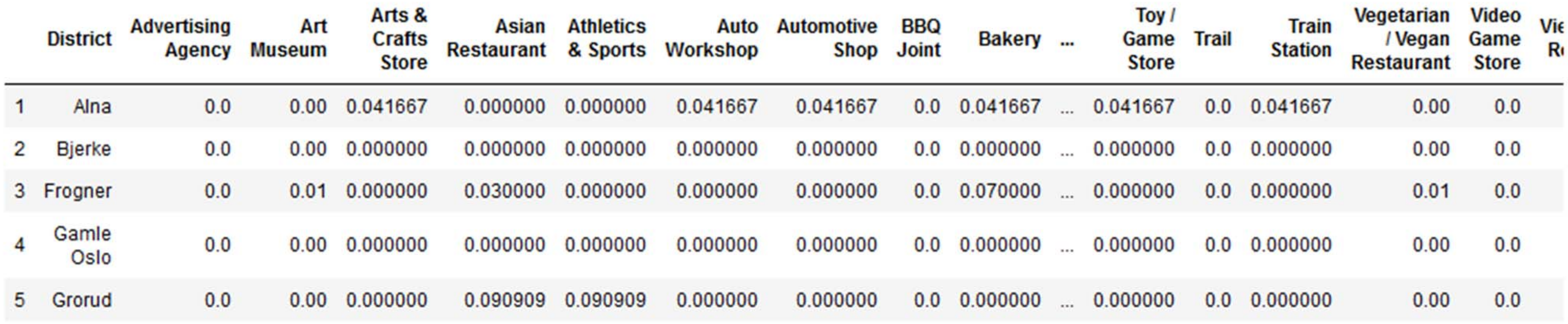


	Borough	Residents	Latitude	Longitude
0	Alna	49 801	59.932417	10.835276
1	Bjerke	33 422	59.941395	10.829208
2	Frogner	59 269	59.922224	10.706649
3	Gamle Oslo	58 671	59.899237	10.734767
4	Grorud	27 707	59.961424	10.880549
5	Grünerløkka	62 423	59.925471	10.777421
6	Nordre Aker	52 327	59.953638	10.756412
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

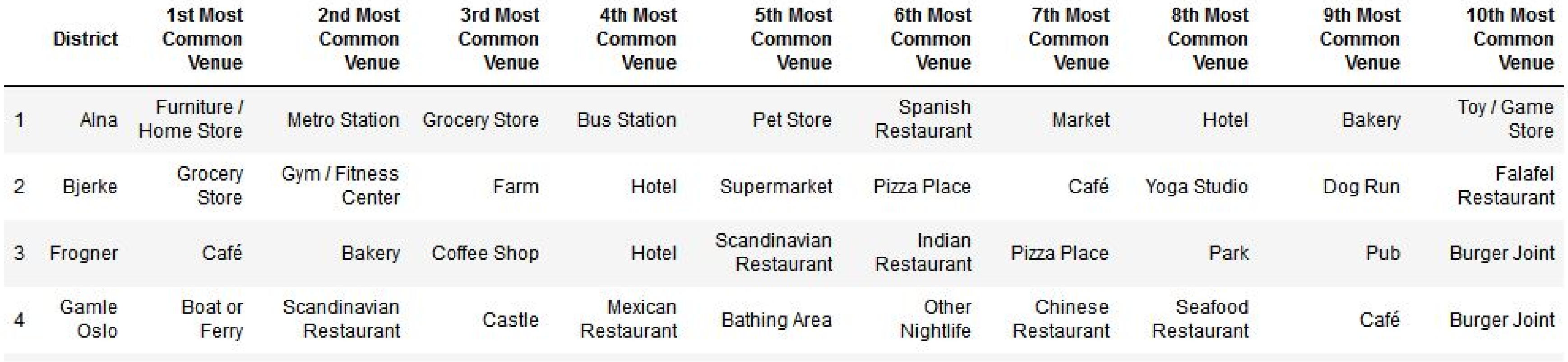
3. Methodology. Obtain venues from Foursquare

Foursquare API is used to look for a maximum of 500 venues within 1000 meters of the geographical coordinates of each neighborhood.

From the venue data acquired we used one hot encoding method to find out what venue categories are most popular. Venues from the same boroughs were grouped by borough names and popular categories were discovered by frequency.



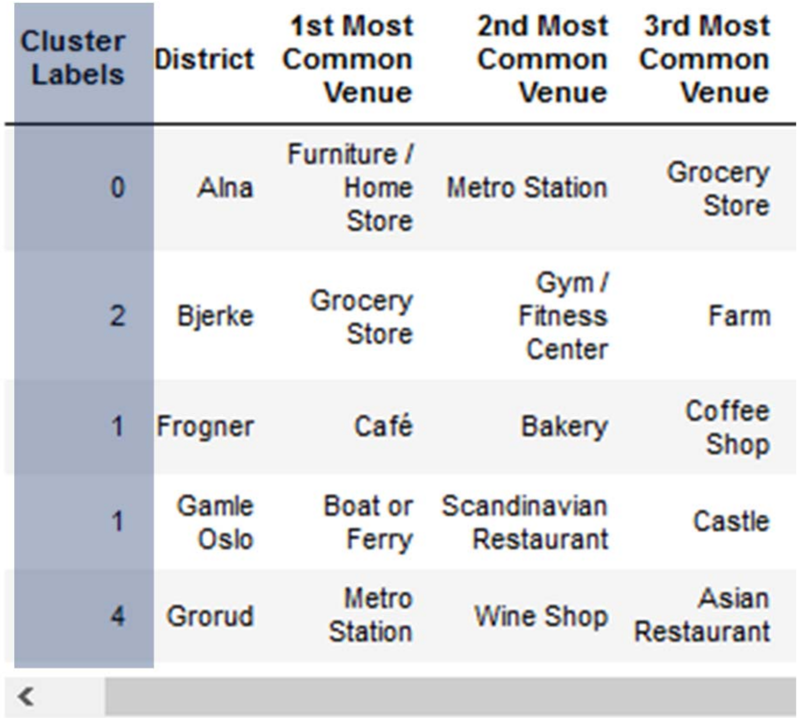
	District	Advertising Agency	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop	Automotive Shop	BBQ Joint	Bakery	...	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Vie R
1	Alna	0.0	0.00	0.041667	0.000000	0.000000	0.041667	0.041667	0.0	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	



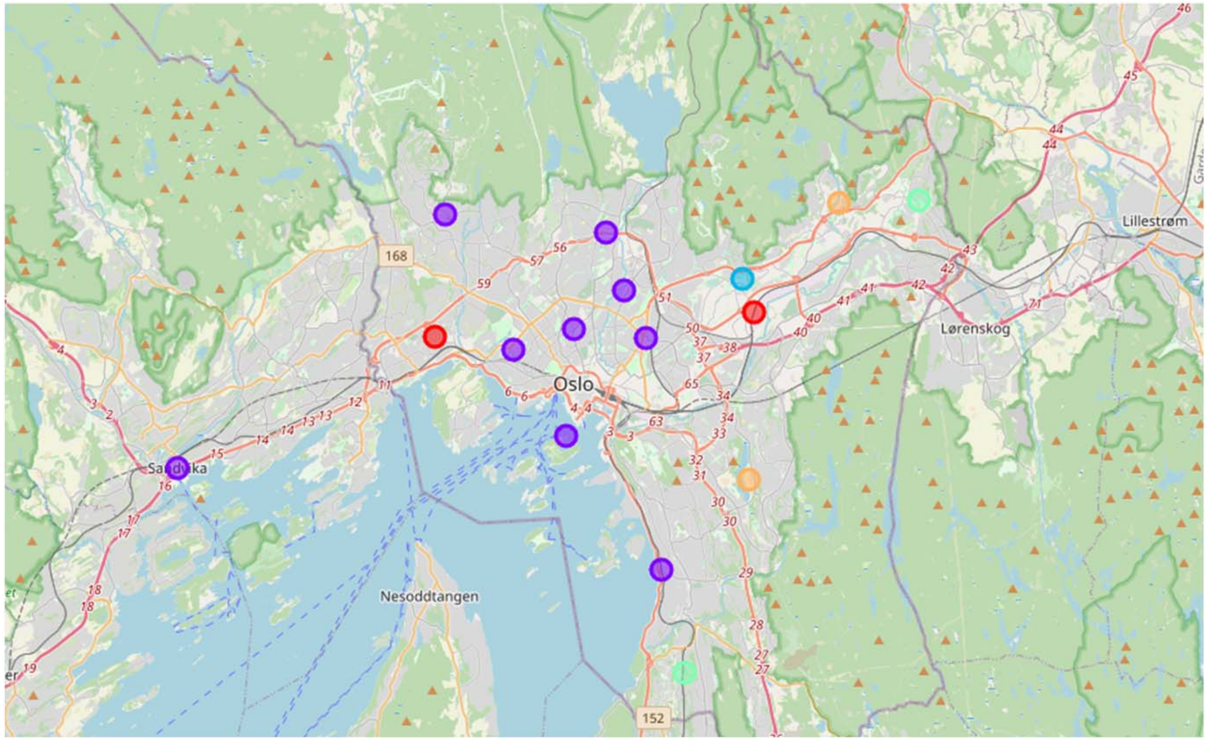
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Alna	Furniture / Home Store	Metro Station	Grocery Store	Bus Station	Pet Store	Spanish Restaurant	Market	Hotel	Bakery	Toy / Game Store
2	Bjerke	Grocery Store	Gym / Fitness Center	Farm	Hotel	Supermarket	Pizza Place	Café	Yoga Studio	Dog Run	Falafel Restaurant
3	Frogner	Café	Bakery	Coffee Shop	Hotel	Scandinavian Restaurant	Indian Restaurant	Pizza Place	Park	Pub	Burger Joint
4	Gamle Oslo	Boat or Ferry	Scandinavian Restaurant	Castle	Mexican Restaurant	Bathing Area	Other Nightlife	Chinese Restaurant	Seafood Restaurant	Café	Burger Joint

3. Methodology. Cluster Analysis: K-Means Clustering

- Based on the common venue categories, boroughs were grouped into five clusters using K-clustering algorithm, and displayed on a map

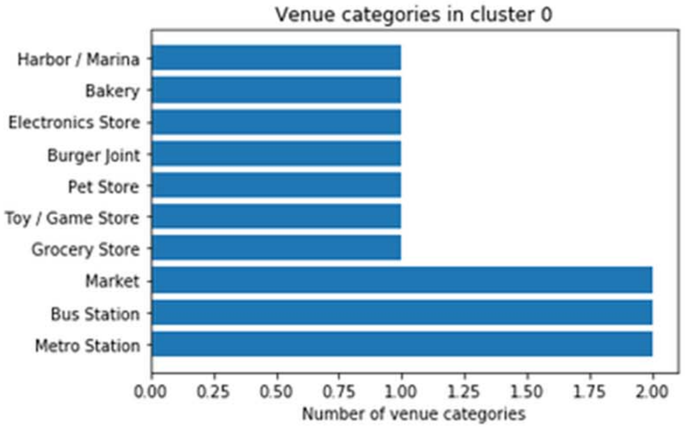


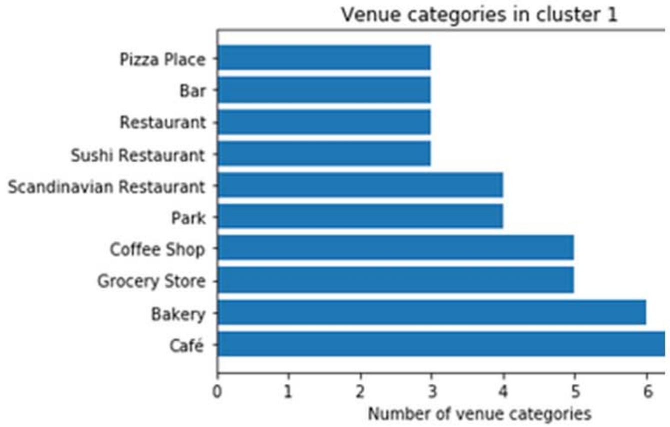
Cluster Labels	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Alna	Furniture / Home Store	Metro Station	Grocery Store
2	Bjerke	Grocery Store	Gym / Fitness Center	Farm
1	Frogner	Café	Bakery	Coffee Shop
1	Gamle Oslo	Boat or Ferry	Scandinavian Restaurant	Castle
4	Grorud	Metro Station	Wine Shop	Asian Restaurant

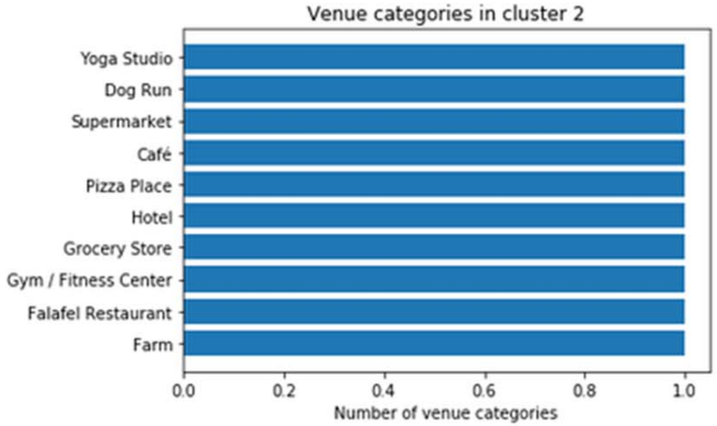


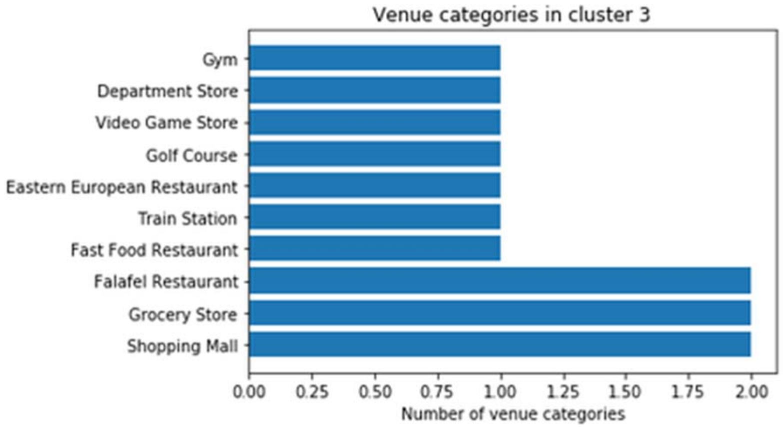
3. Methodology. Cluster characterizations

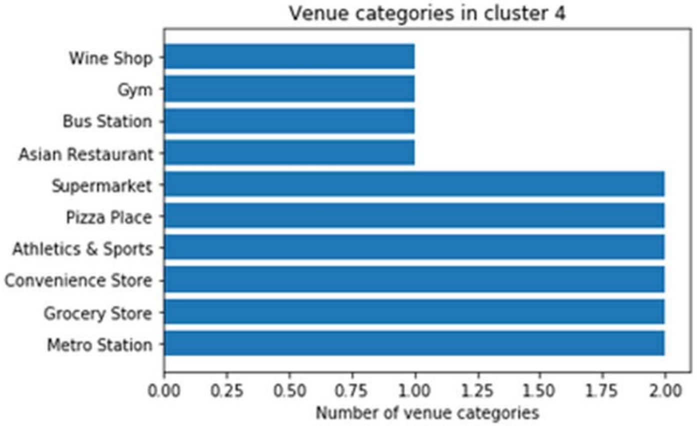
- Each cluster was analyzed by the top categories and named based on the characteristics they display.





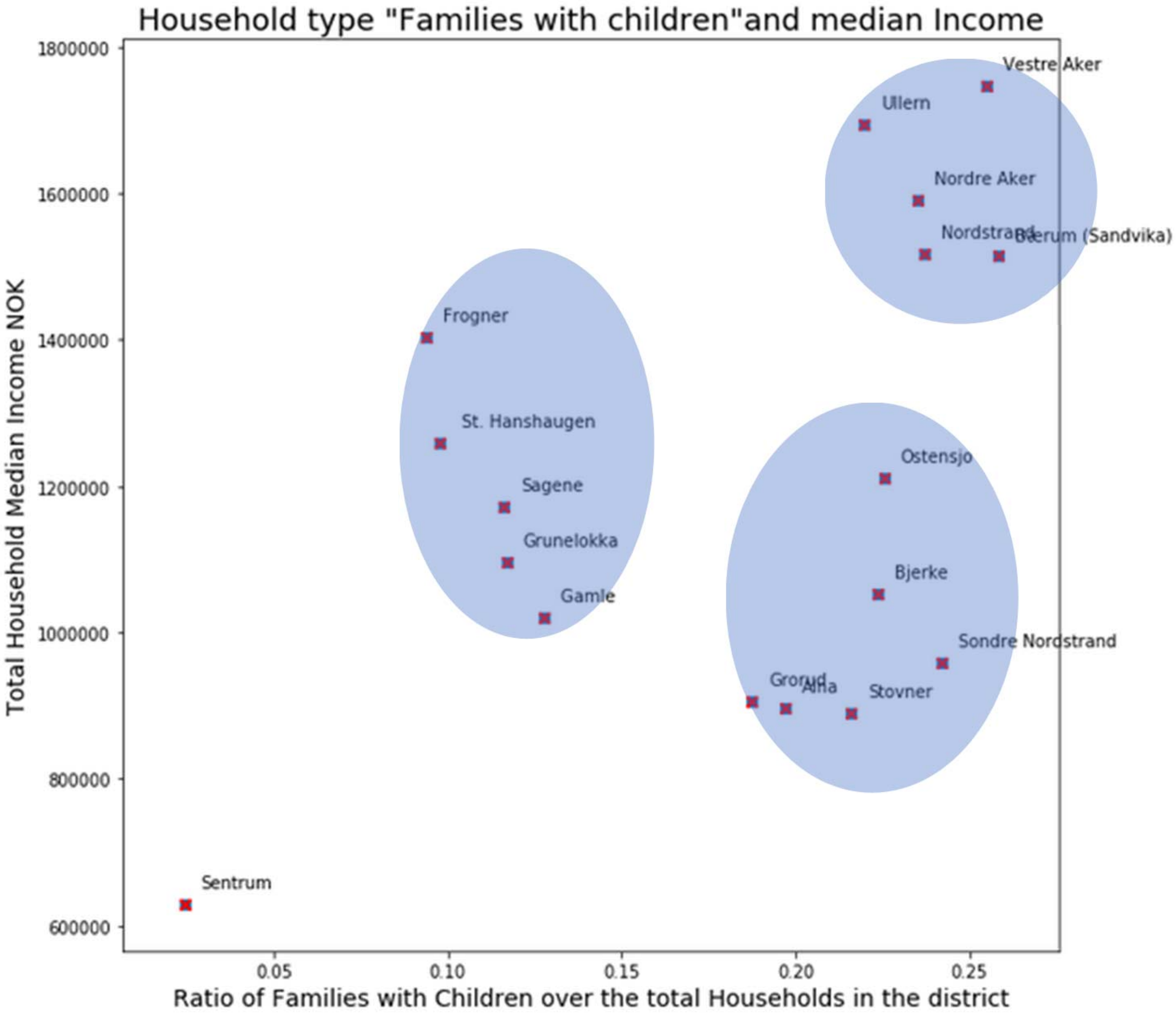






4.Methodology. Socioeconomic data analysis

- Data was downloaded from Norway statistics department and cleaned as needed and plotted in a scatter chart
- This scatter charts shows clearly where families are prevalent and the income level.
  - We can see three clear groups. On the X axis values indicate the ratio of families with children over the total households in the district.
- Group 1 high income and high proportion (>20%) of families with children. Top right corner
- Group 2 middle/low household income and high proportion of families with children. Bottom right corner.
- Group 3. Middle income and low proportion of households with children. Most of the households in these areas are single persons and is also reflected in the average household income.



4.Results

- Through the foursquare location analysis, we identified 3 potential residential clusters one urban and three suburban.
- Considering socioeconomic data presented in a scatter chart with average household income and ratio of families with children over total households we can group Oslo neighborhoods in three groups and related them to the clusters
- **Cluster 0.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.Affluent residential urban**. This group corresponds with groups 1 and 3, middle and high income locations The most attractive locations for families would areas with 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
- **Cluster 2**.Residential suburban. Bjerke seems to be a location of choice for middle income families, but cannot make conclusion based only in venue data
- **Cluster 3 Residential suburban immigration**. 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-ogpublikasjoner/14-per-cent-of-population-are-immigrants
- **Cluster 4 Residential suburban**. Grorud ,Ostenjo

5.Conclusion

- 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