COURSERA APPLIED DATA SCIENCE CAPSTONE PROJECT

The Battle of boroughs in Oslo, Norway

Report of the Applied Data Science Capstone by IBM/Coursera

Table of content

This document is conceived as a part of the final capstone project for IBM Data Science Professional Certification in Coursera. The objectives of this final assignments is to define a business problem, search for data and leverage the Foursquare location data and compare locations to solve a problem or provide insights.

Introduction Business Problem

Background

Oslo is Norway's capital and by far the largest city, and also one of the oldest cities in the country. Oslo is both a municipality and a county and lies at the heart of the Oslo Fjord, with the city center at the heart of the two bays Pipervika and Bjørvika. The town center also includes the area between these coves and the adjacent area just to the north. The city of Oslo (which also includes areas in the neighboring municipalities) has more than one million inhabitants as of 2018.

As of 2004, Oslo has been divided into 15 districts each with it's own administration and with district committees appointed by the city council. Sentrum and Marka are outside the district scheme; these are considered as "common areas" for the municipality's inhabitants. The inhabitants here get their services offered by the nearest district.

Oslo is by far the country's most populous county and municipality. Oslo is a relatively compact capital. The city center is not so big, and from there one can visit many of the main sights by foot. In addition, the city's public transport is excellent, making it easy to get around when it comes to making long distances.

The city center being relatively compact gives first time visitors who chose to live near the city center the possibility to see as much as possible; both the city center and the neighboring districts. For someone planning to live, buy a property and work in the city, the situation will be different. In this paper I am going to segment Oslo's neighborhood and find the best one to live in that offers much in regard to property prices and access to popular venues.

Problem Definition

A newly wed couple is going to relocate in Oslo as the wife has got a new job in the city and the man is going to attend his last year of college in Oslo. Not knowing the city/county very much they want some help in finding the best place for them.

Criterium were defined the following way:

- An "affordable" apartment with proximity to popular venues offering a diverse cultural experience
- Nearby public transportation

Target Audience

- To people interested in investing in living and work in Oslo
- First time real-estate buyer in Oslo
- People planning to move to the capital and acquire a real-estate
- Anyone curious to find more about Oslo in terms of venues and affordable places

Data Collection

Data will be collected from:

- Source : https://no.wikipedia.org/wiki/Liste_over_Oslos_bydeler
  Description : This data set contains the required information. And we will use this data set to explore various boroughs of Oslo.

- Data Source : http://statistikkbanken.oslo.kommune.no/webview/ This site provides the possibility to build datasets and export them in different formats.
  - Kvadratmeterpris_og_omsatte_boliger.csv real-estate price/sqm from 2009 to 2018 in each Borough

  - Distrikter_i_Oslo.csv containing a list over the sub-districts of Oslo: The boroughs or districts of Oslo were in 2007 each divided into 4-8 sub-boroughs as geographical areas within the administrative districts. This was to meet the need for statistics at a lower level than the district /borough. Each sub-district contains several basic units. In many ways, the sub-districts often overlap with school districts, constituencies, districts and parishes, but the intention is, as the Development and Competence Agency says, to have local areas with specific boundaries, in order for statistical measurements to be in a more local area than the administrative districts are. I believe the same reason will also be valid for anyone considering relocating on the basis of numbers.

- Data source : Foursquare API
  Description : By using this API we will get all the venues in each borough/sub-borough

Approach

- Collect data Oslo
- Collect population data
- Collect sqm price
- Visualize data (borough, population, property price, education level)
- Use Foursquare API to get venues in each borough
- Analyze clusters with K-Means
- Inference from the results and related conclusions

Data Wrangling

The data preparation for each of the three sources of data is done separately. The Oslo data containing a list of neighborhoods with their belonging borough is grouped by boroughs. Giving us the table below (*see fig. 1*).

	Borough	Neighborhoods	lat	long
0	Alna	Ellingsrud, Furuset, Teisen, Hellerudtoppen, L...	59.932417	10.835276
1	Bjerke	Linderud, Årvoll, Veitvet, Økern	59.941395	10.829209
2	Gamle Oslo	011 Lodalen, 012 Grønland, 013 Enerhaugen, 014...	59.899237	10.734767
3	Grorud	Ammerud, Grorud, Rødtvet, Nordtvet , Romsås	59.961424	10.880549
4	Vestre Aker	Holmenkollen, Hovseter, Holmen , Vinderen, Røa...	59.958300	10.670319
5	Grünerløkka	Grünerløkka vest, Grünerløkka øst, Dælenenga, ...	59.925471	10.777421
6	Sagene	Iladalen, Sagene, Bjølsen, Sanaker, Torshov	59.938273	10.765849
7	St. Hanshaugen	Hammersborg, Bislett, ila , Fagenborg, Lindern	59.927950	10.738958
8	Frogner	Bygdøy, Frogner, Frognerparken, Majorstuem nor...	59.922224	10.706649
9	Ullern	Ullernåsen, Lilleaker, Ullern, Montebello_Hoff...	59.925818	10.665132
10	Nordre Aker	Disen, Myrer, Grefsen, Kjelsås, Korsvoll, Tåse...	59.953638	10.756412
11	Stovner	Vestli, Fossum, Rommen, Haugenstua, Stovner, H...	59.962140	10.922823
12	Østensjø	Manglerud, Godlia, Oppsal, Bøler, Skullerud	59.887563	10.832748
13	Nordstrand	Ljan, Nordstrand, Bekkelaget, Simensbråten, La...	54.487378	8.865286
14	Sentrum	Sentrum	59.909960	10.743164
15	Maarka	Marka	12.966670	27.816670

Fig. 1 Boroughs in Oslo with coordinates

Cleaning this data I renamed subdistrict to neighborhoods then with OpenCageGeocode add geospatial data to the boroughs.

The second data is scraped from Wikipedia using the Beautiful Soup in Python (*see fig. 1.1*). This data contains the number of residents in each borough. I dropped 'Area' and 'Number' columns as they were of no use.

<title>List of boroughs of Oslo – Wikipedia</title>

Borough	Residents	Area	Number
Alna	49 801	13,7 km^2	12
Bjerke	33 422	7,7 km^2	9
Frogner	59 269	8,3 km^2	5
Gamle Oslo	58 671	7,5 km^2	1
Grorud	27 707	8,2 km^2	10

Fig. 1.1 Scraped data from Wikipedia.

The two datasets are merged on the Borough names to form a new dataset that combines the needed information (*see fig. 1.*2). This will help visualize population in each borough.

	Borough	Neighborhoods	lat	long	Population
0	Alna	Ellingsrud, Furuset, Teisen, Hellerudtoppen, L...	59.932417	10.835276	49834
1	Bjerke	Linderud, Årvoll, Veitvet, Økern	59.941395	10.829209	33491
2	Gamle Oslo	Lodalen, Grønland, Enerhaugen, Nedre Tøyen, Ka...	59.899237	10.734767	58713
3	Grorud	Ammerud, Grorud, Rødtvet, Nordtvet , Romsås	59.961424	10.880549	27630
4	Vestre Aker	Holmenkollen, Hovseter, Holmen , Vinderen, Røa...	59.958300	10.670319	50876
5	Grünerløkka	Grünerløkka vest, Grünerløkka øst, Dælenenga, ...	59.925471	10.777421	62409
6	Sagene	Iladalen, Sagene, Bjølsen, Sanaker, Torshov	59.938273	10.765849	45053
7	St. Hanshaugen	Hammersborg, Bislett, ila , Fagenborg, Lindern	59.927950	10.738958	40321
8	Frogner	Bygdøy, Frogner, Frognerparken, Majorstuem nor...	59.922224	10.706649	59292
9	Ullern	Ullernåsen, Lilleaker, Ullern, Montebello_Hoff...	59.925818	10.665132	34500

Fig. 1.2 First combined dataset

I then use **folium** library to visualize geographic details of Oslo and its boroughs superimposed on top (*see fig. 1*.3). I used latitude and longitude values to get the visual as below:

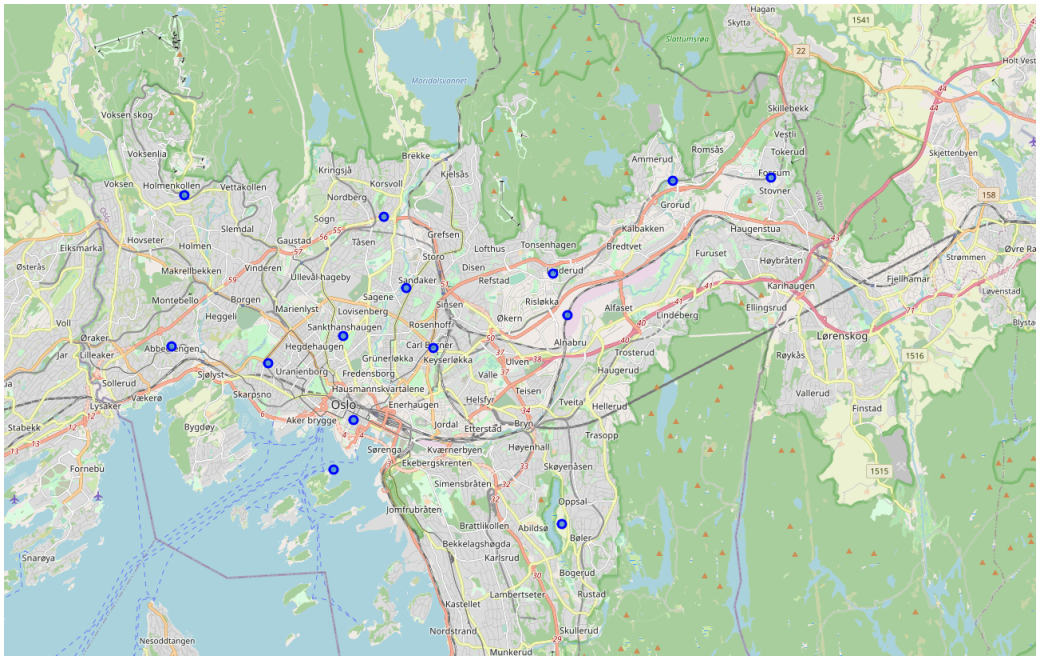


Fig. 1.3 Map of Oslo, with markers of different boroughs

The last table is a dataset containing square meter price pr. borough In Oslo. Visualizing this data will help get an idea of price trend from 2008 to 2018. For our couple this can be important in terms of real estate acquisition which is an investment (*see fig. 1*.4). For the final dataset we will use only the last available price information; 2008, and rename that column to 'Sqm_price'.

	Borough	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	Gamle Oslo	34845	35452	37584	42045	46181	48421	50255	57329	68730	73181	74605
1	Grünerløkka	36440	37314	40084	44456	48990	51078	53173	60548	71505	75909	76413
2	Sagene	37390	38812	41273	46450	51625	53733	55362	63422	74206	79348	80766
3	St. Hanshaugen	40942	41759	44969	50247	54898	57328	58065	65631	76307	82436	81649
4	Frogner	44180	45150	49273	54552	58995	61100	61839	68158	79721	87326	87923

Fig. 1.4 Square meter price pr. borough from 2008 -2018

Merging this table with the former gives us our final dataset with the 16 boroughs with their coordinates, their respective population, their neighborhoods and sqm. price (*see fig. 1*.5).

	Borough	lat	long	Neighborhoods	Population	Sqm_price
0	Alna	59.932417	10.835276	Ellingsrud, Furuset, Teisen, Hellerudtoppen, L...	49834	49042
1	Bjerke	59.941395	10.829209	Linderud, Årvoll, Veitvet, Økern	33491	58834
2	Gamle Oslo	59.899237	10.734767	Lodalen, Grønland, Enerhaugen, Nedre Tøyen, Ka...	58713	74605
3	Grorud	59.961424	10.880549	Ammerud, Grorud, Rødtvet, Nordtvet , Romsås	27630	49224
4	Vestre Aker	59.958300	10.670319	Holmenkollen, Hovseter, Holmen , Vinderen, Røa...	50876	67186

Fig. 1.5 Final dataset

Methodology

Exploratory Data Analysis

Statistical summary

Using the python describe function I got the statistics for my dataset (*see fig. 1*.6)., returning the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime.

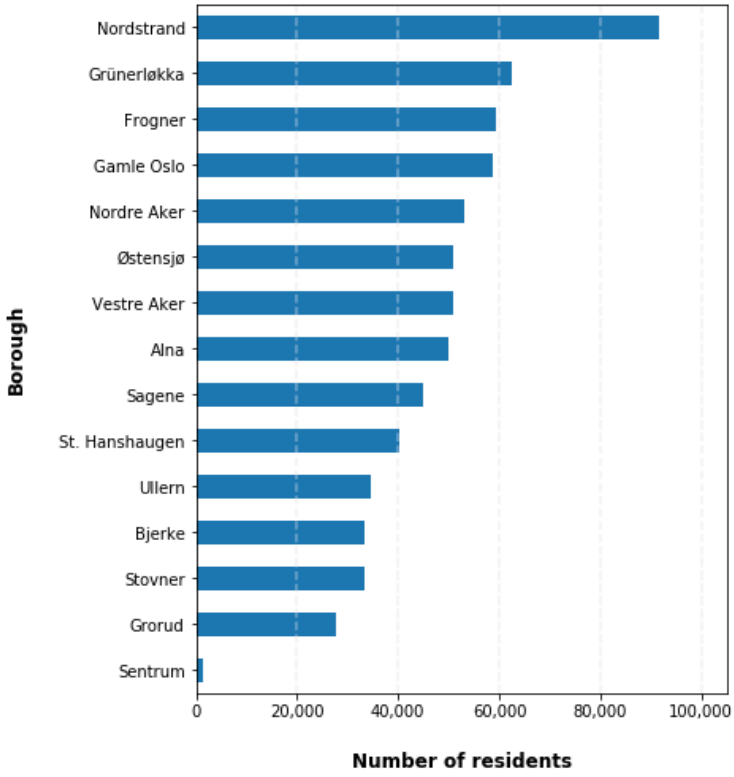
	lat	long	Population	Sqm_price
count	15.000000	15.000000	15.000000	15.000000
mean	59.568879	10.648304	46176.200000	67518.400000
std	1.405928	0.498655	20023.163109	15051.000061
min	54.487378	8.865286	1471.000000	42394.000000
25%	59.916092	10.720708	33995.500000	53674.000000
50%	59.927950	10.756412	49834.000000	74109.000000
75%	59.947516	10.830978	55959.500000	79419.500000
max	59.962140	10.922823	91683.000000	87923.000000

Fig. 1.6 Statistical description of the final dataset

The count for each of the major categories in the dataset returns 15 which is the number of boroughs in our dataset. The most populated borough has 91683 residents while the least one has 1471. Regarding square meter price the most expensive is as twice as the least expensive.

Most populated area

Nordstrand is a combination of 2 former boroughs resulting in being the most populated borough in Oslo. Sentrum being the least populated can be explained by the fact that it's a typical are for commerce, administration and so on.



Borough with most expensive sqm.price

The borough with the index 4 is where the square meter price is the highest. This borough, Frogner, has the index 4. We can also see that in other expensive boroughs prices are slightly increasing or decreasing. In Frogner the tendency has remain the same with a larger leap from 2016 (*see fig. 2.1).*

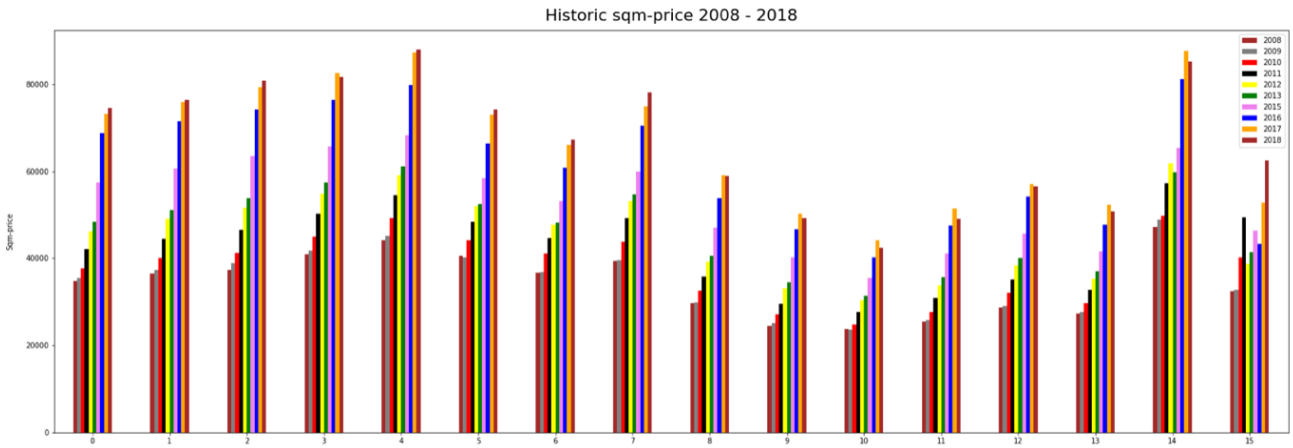


Fig. 2.1. Square meter price trend pr. borough

Modelling

I used the Foursquare API to explore the boroughs and segment them. With the final dataset I then used the Foursquare API to explore the first borough and repeated the process for all the other boroughs. I sat the limit to 100 venues and the radius 500 meter for each borough from their given latitude and longitude information. Here (*see fig. 3*) is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

	Borough	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alna	59.932417	10.835276	BROBEKK STORCASH AS	59.932201	10.831548	Grocery Store
1	Alna	59.932417	10.835276	Vollebekk (B)	59.935928	10.831141	Bus Station
2	Alna	59.932417	10.835276	Vollebekk (T)	59.935934	10.831086	Metro Station
3	Bjerke	59.941395	10.829209	SATS	59.943087	10.835261	Gym / Fitness Center
4	Bjerke	59.941395	10.829209	Thon Hotel Linne	59.939471	10.829001	Hotel

Fig. 3. Venues details of each borough

One Hot Encoding

Because we have strings as labels for each borough and need a way to digitizing them so that we can use them in our classification algorithm. We need to parse our labels and assigns dummy values to each as well as creates new columns per each label and using 1 or 0 to determine whether that row of table has that feature or not.

I then grouped the data by boroughs and calculate the mean of the venues (*see fig. 3.1*) and got the 5 most common most common venues in each borough. Finally by using the Foursquare API in conjunction with the created datasets, a table of most common visited venues in Oslo's neighborhoods is generated.

	Borough	American Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Bakery	Bar	Beer Bar	Bike Rental / Bike Share	Boarding House	Boat or Ferry	Bookstore	Burrito Place	Bus Station	Bus Stop	
0	Alna	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.333333	0.000000	0
1	Bjerke	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0
2	Frogner	0.00	0.00	0.035714	0.00	0.000000	0.00	0.035714	0.000000	0.000000	0.00	0.035714	0.000000	0.00	0.00	0.035714	0.000000	(
3	Gamle Oslo	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.363636	0.00	0.00	0.000000	0.000000	0
4	Grorud	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.125000	0.000000	0
5	Grünerløkka	0.00	0.00	0.000000	0.00	0.066667	0.00	0.066667	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.066667	0.000000	0
6	Nordre Aker	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.285714	0.285714	(
7	Nordstrand	0.00	0.00	0.000000	0.00	0.000000	0.00	1.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0
8	Sagene	0.00	0.00	0.000000	0.00	0.000000	0.00	0.103448	0.034483	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0
9	Sentrum	0.01	0.02	0.000000	0.01	0.000000	0.00	0.000000	0.040000	0.010000	0.01	0.000000	0.000000	0.01	0.01	0.000000	0.000000	0
10	St. Hanshaugen	0.00	0.00	0.000000	0.00	0.000000	0.00	0.153846	0.000000	0.038462	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0
11	Stovner	0.00	0.00	0.000000	0.00	0.000000	0.25	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0
12	Ullern	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.200000	0.000000	0
13	Vestre Aker	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0

Fig. 3.1 One hot grouped

To help our couple find similar boroughs in Oslo, we will cluster boroughs using K-means clustering. This is an unsupervised machine learning algorithm for clustering data on the basis of predefine cluster size. In regard to the goal of this exercise, we want to help our couple shortlist their area of interests based on venues around each borough. For this project I will choose a cluster size of 3.

Results and discussion

K-means clustering gives us here access to the different clusters in Oslo where we can see the most common venues.

Borough	Neighborhoods	lat	long	Population	Cluster Labels	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
Alna	Ellingsrud, Furuset, Teisen, Hellerudtoppen, L...	59.932417	10.835276	49834	0.0	Grocery Store	Metro Station	Bus Station	Wine Shop	Fish Market	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dog Run
Bjerke	Linderud, Årvoll, Veitvet, Økern	59.941395	10.829209	33491	2.0	Grocery Store	Café	Hotel	Gym / Fitness Center	Convenience Store	Fast Food Restaurant	Fish Market	Department Store	Dessert Shop	Dim Sum Restaurant
Gamle Oslo	Lodalen, Grønland, Enerhaugen, Nedre Tøyen, Ka...	59.899237	10.734767	58713	2.0	Boat or Ferry	Pier	Market	Scandinavian Restaurant	Café	Other Nightlife	Historic Site	Gastropub	Furniture / Home Store	Gym / Fitness Center
Grorud	Ammerud, Grorud, Rødtvet, Nordtvet , Romsås	59.961424	10.880549	27630	0.0	Grocery Store	Wine Shop	Supermarket	Pizza Place	Bus Station	Soccer Field	Convenience Store	Furniture / Home Store	Gastropub	Gym / Fitness Center
Vestre Aker	Holmenkollen, Hovseter, Holmen , Vinderen, Røa...	59.958300	10.670319	50876	0.0	Grocery Store	Restaurant	Metro Station	Ski Area	Wine Shop	Fast Food Restaurant	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant

Fig. 4. Cluster labelling

Looking closely at each cluster, we can also see what neighborhoods from different boroughs are similar (*see fig. 4.1, 4.2 & 4.3*)

Cluster 1

```
boroughs.loc[boroughs['Cluster Labels'] == 0, boroughs.columns[[1] + list(range(3, boroughs.shape[1]))]]
```

	Neighborhoods	long	Population	Cluster Labels	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
0	Ellingsrud, Furuset, Teisen, Hellerudtoppen, L...	10.835276	49834	0.0	Grocery Store	Metro Station	Bus Station	Wine Shop	Fish Market	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dog Run
3	Ammerud, Grorud, Rødtvet, Nordtvet , Romsås	10.880549	27630	0.0	Grocery Store	Wine Shop	Supermarket	Pizza Place	Bus Station	Soccer Field	Convenience Store	Furniture / Home Store	Gastropub	Gym / Fitness Center
4	Holmenkollen, Hovseter, Holmen , Vinderen, Røa...	10.670319	50876	0.0	Grocery Store	Restaurant	Metro Station	Ski Area	Wine Shop	Fast Food Restaurant	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant
10	Disen, Myrer, Grefsen, Kjelsås, Korsvoll, Tåse...	10.756412	53206	0.0	Bus Stop	Bus Station	Grocery Store	Café	Shopping Mall	Wine Shop	Fast Food Restaurant	Department Store	Dessert Shop	Dim Sum Restaurant
11	Vestli, Fossum, Rommen, Haugenstua, Stovner, H...	10.922823	33259	0.0	Video Game Store	Grocery Store	Athletics & Sports	Shopping Mall	Wine Shop	Fish Market	Deli / Bodega	Department Store	Dessert Shop	Dim Sum Restaurant

Fig. 4.1 Cluster 1

Cluster 2

```
boroughs.loc[boroughs['Cluster Labels'] == 1, boroughs.columns[[1] + list(range(3, boroughs.shape[1]))]]
```

	Neighborhoods	long	Population	Cluster Labels	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
13	Ljan, Nordstrand, Bekkelaget, Simensbråten, La...	8.865286	91683	1.0	Bakery	Wine Shop	Food Court	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dog Run	Exhibit	Fast Food Restaurant

Fig. 4.2 Cluster 2

Cluster 3

```
boroughs.loc[boroughs['Cluster Labels'] == 2, boroughs.columns[[1] + list(range(3, boroughs.shape[1]))]]
```

	Neighborhoods	long	Population	Cluster Labels	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	Linderud, Årvoll, Veitvet, Økern	10.829209	33491	2.0	Grocery Store	Café	Hotel	Gym / Fitness Center	Convenience Store	Fast Food Restaurant	Fish Market	Department Store	Dessert Shop	Dim Sum Restaurant
2	Lodalen, Grønland, Enerhaugen, Nedre Tøyen, Ka...	10.734767	58713	2.0	Boat or Ferry	Pier	Market	Scandinavian Restaurant	Café	Other Nightlife	Historic Site	Gastropub	Furniture / Home Store	Gym / Fitness Center
5	Grünerløkka vest, Grünerløkka øst, Dælenenga, ...	10.777421	62409	2.0	Gym / Fitness Center	Sushi Restaurant	Coffee Shop	Dog Run	Chinese Restaurant	Sports Bar	Bus Station	Bakery	Plaza	Asian Restaurant
6	Iladalen, Sagene, Bjølsen, Sanaker, Torshov	10.765849	45053	2.0	Coffee Shop	Park	Bakery	Pizza Place	Sushi Restaurant	Concert Hall	Performing Arts Venue	Pet Store	Deli / Bodega	Pub
7	Hammersborg, Bislett, ila , Fagenborg, Lindern	10.738958	40321	2.0	Bakery	Park	Coffee Shop	Café	Wine Shop	Beer Bar	Gourmet Shop	Gastropub	Italian Restaurant	French Restaurant
8	Bygdøy, Frogner, Frognerparken, Majorstuem nor...	10.706649	59292	2.0	Hotel	Café	Sculpture Garden	Italian Restaurant	Gourmet Shop	Scandinavian Restaurant	Bus Station	Coffee Shop	French Restaurant	Restaurant
9	Ullernåsen, Lilleaker, Ullern, Montebello_Hoff...	10.665132	34500	2.0	Light Rail Station	Bus Station	Convenience Store	Flower Shop	Wine Shop	Fish Market	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
14	Sentrum	10.743164	1471	2.0	Hotel	Coffee Shop	Restaurant	Scandinavian Restaurant	Bar	Café	Italian Restaurant	Indian Restaurant	Sushi Restaurant	Clothing Store

Fig. 4.3 Cluster 3

Helping a couple make a decision about where to live in Oslo was the objective in this project. This objective was met by making use of different sets of data covering boroughs and their associated neighborhoods, population data in each borough and finally sqm price to find out the best place to invest for their housing. We grouped those neighborhoods into clusters and this will contribute in shortlisting their decision.

Conclusion

During the analysis, several important statistical features of the boroughs were explored and visualized. Clustering helped highlight optimal areas in terms of venues, and population number. Knowing Oslo I found out that Foursquare doesn't represent the full picture, since many of the venues weren't fetched.