FINDING A PLACE TO LIVE

The power of analytics

FEBRUARY 20, 2021

JOACHIM ERIKSSON

Original work created for the Coursera Data Science Capstone assignment

Contents

Oslo, Norway[4)]

Introduction:

One of the most sought-after places to live in Norway is Oslo with its cultural diversity, great business opportunities and wide offers in different activities and parks. Before I moved here, I was quite uncertain of where to settle down in Oslo and heard that there were some differences to the boroughs. However, I was not able to find any information about this before moving and having this information could have been helpful in deciding where to live. In addition to having information about activity offers and parks, the population density is also a factor that plays a role in the atmosphere where you live, as some of us like to be surrounded more by people than others. I therefore included population density to help highlight this and make it possible to do decisions based on multiple variables and place this information on a map.

In short, I believe this analysis could be helpful for people moving to Oslo, so I wanted to use the Foursquare API to show venues, activities, parks, and people density to make the difference of these boroughs more distinguishable.

Problem statement

How do we distinguish the boroughs in Oslo so that a person moving to this city may get an understanding of activity offers and population density that differs each borough to aid him or her when deciding where to live?

Data

The dataset I used for this project is generated using GeoPy[1)] to retrieve coordinates for each borough, the Foursquare API[2)] for its venues and parks, and a table of boroughs with population in Oslo from Wikipedia[3)]. *Please visit references for details.*

Fetching Boroughs from Wikipedia

All 15 boroughs where simply retrieved from Wikipedia using pandas read_html method and required renaming to English column headers and the population field to be converted into float values.

	Borough	Population	Square KM	Borough ID
0	Alna	49 801	137	12
1	Bjerke	33 422	77	9
2	Frogner	59 269	83	5
3	Gamle Oslo	58 671	75	1
4	Grorud	27 707	82	10
5	Grünerløkka	62 423	48	2
6	Nordre Aker	52 327	136	8
7	Nordstrand	52 459	169	14
8	Sagene	45 089	31	3
9	St. Hanshaugen	38 945	36	4
10	Stovner	33 316	82	11
11	Søndre Nordstrand	39 066	184	15
12	Ullern	34 569	94	6
13	Vestre Aker	50 157	166	7
14	Østensjø	50 806	122	13

Coordinates with GeoPy

With GeoPy I fetched all 15 Boroughs' corresponding latitude and longitude values to be used for the Foursqare API calls.

	Borough	Population	Square KM	Borough ID	lat	lng
0	Alna	49801.0	137.0	12	59.932417	10.835276
1	Bjerke	33422.0	77.0	9	59.940668	10.808725
2	Frogner	59269.0	83.0	5	59.922224	10.706649
3	Gamle Oslo	58671.0	75.0	1	59.899237	10.734767
4	Grorud	27707.0	82.0	10	59.961424	10.880549
5	Grünerløkka	62423.0	48.0	2	59.923856	10.757889
6	Nordre Aker	52327.0	136.0	8	59.953638	10.756412
7	Nordstrand	52459.0	169.0	14	59.864561	10.786143
8	Sagene	45089.0	31.0	3	59.936887	10.755306
9	St. Hanshaugen	38945.0	36.0	4	59.927950	10.738958
10	Stovner	33316.0	82.0	11	59.962140	10.922823
11	Søndre Nordstrand	39066.0	184.0	15	59.835944	10.798496
12	Ullern	34569.0	94.0	6	59.925567	10.655798
13	Vestre Aker	50157.0	166.0	7	59.958300	10.670319
14	Østensjø	50806.0	122.0	13	59.887563	10.832748

Venues from Foursquare

I used the /explore endpoint of the Foursquare API in order to retrieve max 50 venue names and categories per borough as seen below:

	index_col	Borough	venue_name	venue_category	venue_lat	venue_lng
423	8	Sagene	Bombay Cuisine	Indian	59.929423	10.760781
331	6	Nordre Aker	Mat & Mer	Deli / Bodega	59.940046	10.759173
12	0	Alna	en til! pizza	Pizza	59.942593	10.814542
593	11	Søndre Nordstrand	Godt Brød	Bakery	59.907029	10.757393
716	14	Østensjø	Ulvøya	Beach	59.869724	10.772059
596	11	Søndre Nordstrand	Østre Greverud Idrettshall	Athletics & Sports	59.773655	10.815591
28	0	Alna	Harald Huysman Karting	Racetrack	59.919503	10.836280
357	7	Nordstrand	Fiskevollbukta	Beach	59.842372	10.777234
711	14	Østensjø	Fuglen Coffee Roasters Oslo	Coffee Shop	59.906185	10.774646
279	5	Grünerløkka	East Kitchen	Asian	59.920755	10.757422

The shape of this dataframe is 750 rows long and 6 columns wide.

Methodology

Feature engineering

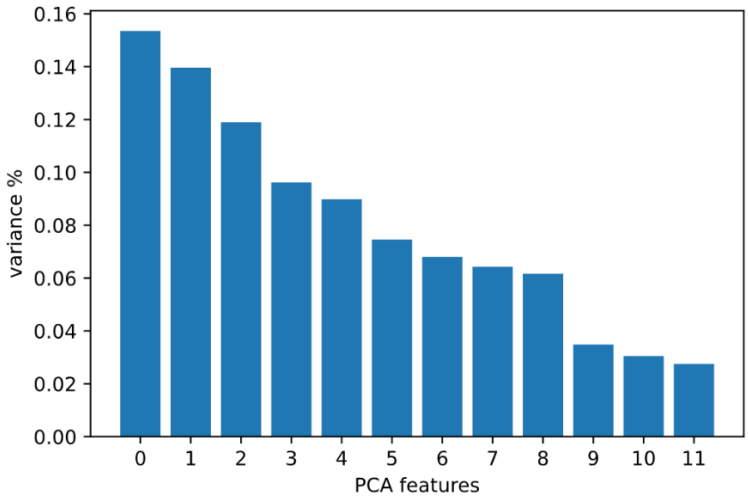
With the current data on different venues I decided to one-hot encode venue categories to create features and find the frequency of all features per borough.

	Borough	index_col	venue_lat	venue_lng	Advertising Agency	Amphitheater	Apparel	Art Gallery	Art Museum	Arts & Crafts	...	Theme Park	Theme Restaurant	Track	Trail	Train Station	Wat
0	Alna	0	59.930589	10.821124	0.0	0.0	0.00	0.00	0.00	0.02	...	0.0	0.00	0.0	0.0	0.0	
1	Bjerke	1	59.939794	10.787494	0.0	0.0	0.00	0.00	0.00	0.00	...	0.0	0.00	0.0	0.0	0.0	
2	Frogner	2	59.920573	10.714704	0.0	0.0	0.00	0.00	0.02	0.00	...	0.0	0.00	0.0	0.0	0.0	
3	Gamle Oslo	3	59.908831	10.739617	0.0	0.0	0.02	0.04	0.00	0.00	...	0.0	0.02	0.0	0.0	0.0	
4	Grorud	4	59.944661	10.858087	0.0	0.0	0.00	0.00	0.00	0.02	...	0.0	0.00	0.0	0.0	0.0	

The dataset then became 15 rows long with 127 features.

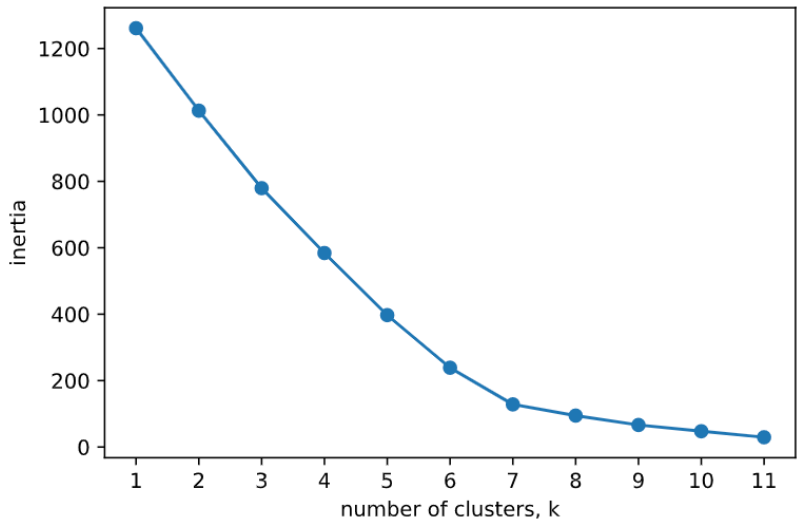
Dimensionality reduction using Principal Component Analysis (PCA)

As the dataset became quite rich in features, but with a low number of samples I wanted to see if I could extract the most important features from the data so the model would perform better. PCA therefore became a choice in reducing the dimensionality and number of features was reduced from 127 to 12 where 95% of the variance was explained. The data was also scaled using scikit learn StandardScaler beforehand to normalize values for a more accurate model. However, the total features chosen from this was 7 as the K-Means algorithm does not work too well with very many features. More on this below.



K-Means Clustering

For the analysis I choose to use K-Means from scikit learn to create clusters based on the venues and population density. As explained above I used 7 features to be able to create clusters that could separate and make better distinctions. One of the reasons was also to be able to choose the K, or number of clusters using the knee method – more features makes this difficult and doesn't work well with Euclidian distance used by the K-Means algorithm.



After the clustering was run, I was left with these 7 clusters:

	Cluster Labels	Borough	Population	Square KM	Borough ID	lat	lng
0	1	Alna	49801.0	137.0	12	59.932417	10.835276
1	4	Bjerke	33422.0	77.0	9	59.940668	10.808725
2	2	Frogner	59269.0	83.0	5	59.922224	10.706649
3	6	Gamle Oslo	58671.0	75.0	1	59.907349	10.773927
4	1	Grorud	27707.0	82.0	10	59.961424	10.880549
5	0	Grünerløkka	62423.0	48.0	2	59.923856	10.757889
6	4	Nordre Aker	52327.0	136.0	8	59.953638	10.756412
7	3	Nordstrand	52459.0	169.0	14	59.864561	10.786143
8	4	Sagene	45089.0	31.0	3	59.936887	10.755306
9	4	St. Hanshaugen	38945.0	36.0	4	59.927950	10.738958
10	1	Stovner	33316.0	82.0	11	59.962140	10.922823
11	3	Søndre Nordstrand	39066.0	184.0	15	59.835944	10.798496
12	2	Ullern	34569.0	94.0	6	59.925567	10.655798
13	5	Vestre Aker	50157.0	166.0	7	59.958300	10.670319
14	3	Østensjø	50806.0	122.0	13	59.887563	10.832748

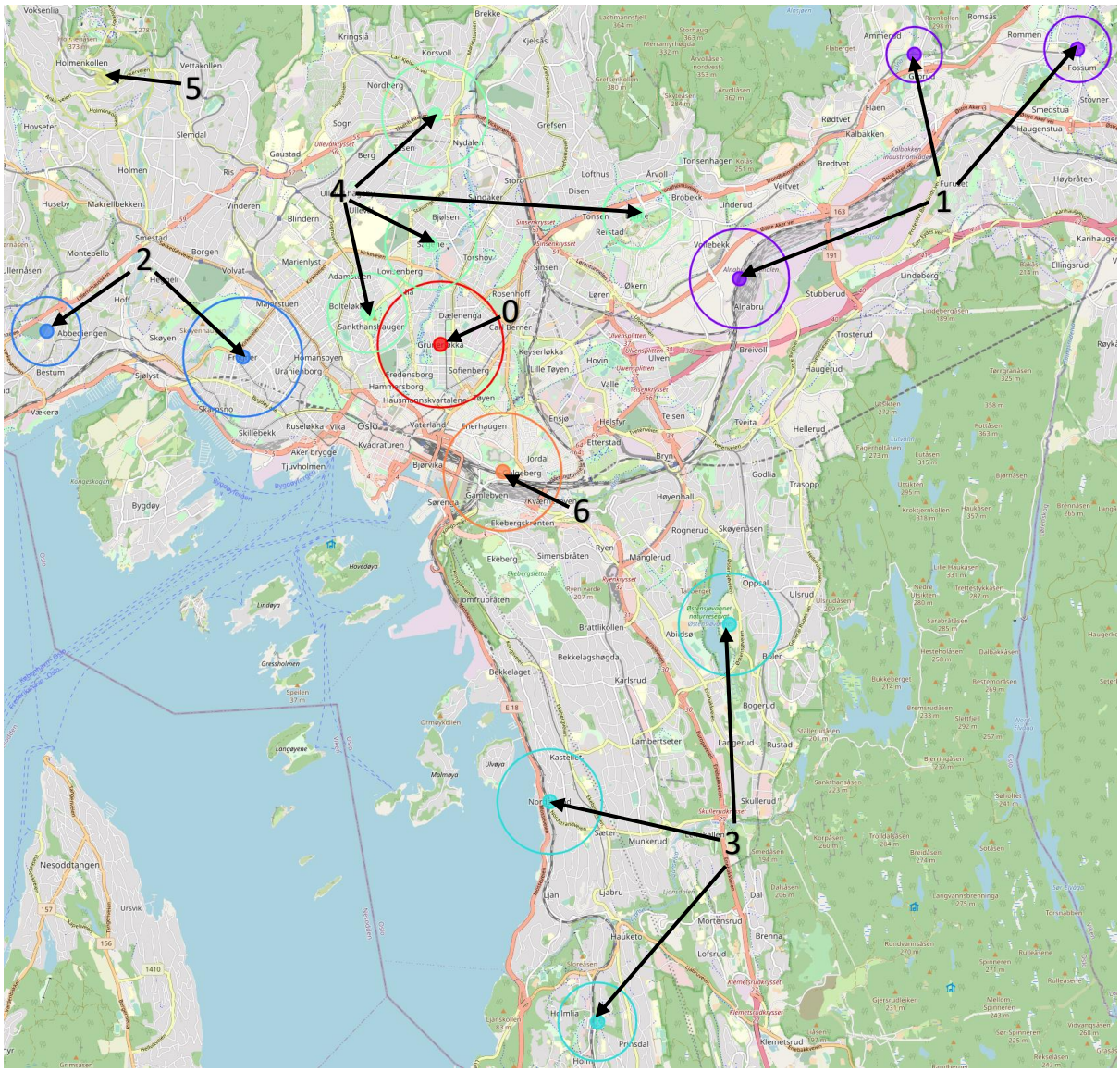
I wanted to look at what features where more prominent within each cluster. As you can see here in cluster 0 grocery stores are highly represented compared to the rest of the clusters. I have appended all cluster results in the appendix at the end of the report for those who are interested. These are also used for the tooltips in the Folium map.

Cluster Labels	0	1	2	3	4	5	6
Coffee Shop	0.080000	0.040000	0.060000	0.120000	0.320000	0.020000	0.120000
Bar	0.080000	0.000000	0.020000	0.100000	0.160000	0.000000	0.140000
Park	0.080000	0.020000	0.060000	0.160000	0.220000	0.020000	0.080000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Burgers	0.040000	0.040000	0.060000	0.000000	0.020000	0.020000	0.000000
Brewery	0.040000	0.000000	0.000000	0.000000	0.040000	0.000000	0.000000
Tapas	0.040000	0.000000	0.020000	0.000000	0.000000	0.000000	0.000000
French	0.040000	0.000000	0.040000	0.000000	0.040000	0.020000	0.000000
Cocktail	0.040000	0.000000	0.000000	0.000000	0.120000	0.000000	0.000000
Italian	0.040000	0.000000	0.100000	0.020000	0.040000	0.000000	0.020000

Analysis

Clusters applied to map

Here you can see all 15 different boroughs which are colored after cluster ID. The size of the ring around the center points shows the relative population density of all boroughs in Oslo i.e. small ring = less people. These distinctions are discussed in detail in the results beneath.



Data used for the map.

	Category Freq	Cluster Labels	Borough	Population	Square KM	Borough ID	lat	lng	Density pct of total
0	Coffee Shop 8.0 Bar 8.0 Park ...	0	Grünerløkka	62423.0	48.0	2	59.923856	10.757889	9.1
1	Grocery Store 15.0 Hotel ...	1	Alna	49801.0	137.0	12	59.932417	10.835276	7.2
2	Grocery Store 15.0 Hotel ...	1	Grorud	27707.0	82.0	10	59.961424	10.880549	4.0
3	Grocery Store 15.0 Hotel ...	1	Stovner	33316.0	82.0	11	59.962140	10.922823	4.8
4	Gym / Fitness 9.0 Hotel 9.0 Ca...	2	Frogner	59269.0	83.0	5	59.922224	10.706649	8.6
5	Gym / Fitness 9.0 Hotel 9.0 Ca...	2	Ullern	34569.0	94.0	6	59.925567	10.655798	5.0
6	Beach 10.0 Grocery Store 7.0 Su...	3	Nordstrand	52459.0	169.0	14	59.864561	10.786143	7.6
7	Beach 10.0 Grocery Store 7.0 Su...	3	Søndre Nordstrand	39066.0	184.0	15	59.835944	10.798496	5.7
8	Beach 10.0 Grocery Store 7.0 Su...	3	Østensjø	50806.0	122.0	13	59.887563	10.832748	7.4
9	Coffee Shop 8.0 Bakery 7.0 Gym ...	4	Bjerke	33422.0	77.0	9	59.940668	10.808725	4.9
10	Coffee Shop 8.0 Bakery 7.0 Gym ...	4	Nordre Aker	52327.0	136.0	8	59.953638	10.756412	7.6
11	Coffee Shop 8.0 Bakery 7.0 Gym ...	4	Sagene	45089.0	31.0	3	59.936887	10.755306	6.6
12	Coffee Shop 8.0 Bakery 7.0 Gym ...	4	St. Hanshaugen	38945.0	36.0	4	59.927950	10.738958	5.7
13	Hotel 10.0 Ski Area ...	5	Vestre Aker	50157.0	166.0	7	59.958300	10.670319	7.3
14	Bar 14.0 Coffee Shop 12.0 Park ...	6	Gamle Oslo	58671.0	75.0	1	59.907349	10.773927	8.5

Results

The differences in the boroughs becomes more transparent as we apply clustering and show us the following details:

[n] *= cluster ID*

- **[0]** Grünerløkka has bars, coffee shops, parks, cafés, Italian, French and other exotic food types. Overall, this borough is not tilted towards any specific category, and seems to have many different options.
- **[1]** The boroughs Alna, Grorud and Stovner are areas with the most grocery stores and supermarkets, hotels and cafés. Grorud and Stovner are one of the areas with the lowest population densities with 4,0 and 4,8 percent of total respectively. It is likely that these burrows are more quiet places to live.
- **[2]** Frogner and Ullern are the highest in Gyms and hotels, Frogner having a high population density and lower for Ullern.
- **[3]** Nordstrand, Søndre Nordstrand and Østensjø have access to beaches and lodges. There are also parks nearby. The density is high for Norsdatand and Østensjø around 7,5% and 5,7 for Søndre Nordstrand.
- **[4]** Bjerke, Nordre Aker, Sagene and St. Hanshaugen are all dense in coffee shops, bakeries, restaurants, parks and gyms. They are also middle tier in population density – so this has something for all.
- **[5]** Vestre Aker has hotels, ski areas and golf courses and medium population density.
- **[6]** Gamle Oslo is the 3rd most population dense area and the highest in number of bars. It also has parks and record shops.

Conclusion

So, how do we distinguish the boroughs in Oslo so that a person moving to this city may get an understanding of activity offers and people density that differs each borough to aid him or her choosing where to live?

As we can see on the map clustering is great for making distinctions and helps us separate the boroughs. By analyzing the venues, activities, population density, parks and beaches it makes finding a new place to live a lot easier.

Additional data that could be used to improve this model is median and mean income, education levels, age distributions and housing cost. This would give us more information about the demographic that lives in each area and expected prices for either buying or renting an apartment or house.

Appendix

Frequency tables on all clusters

The following tables are sorted descending per cluster label to show the highest frequencies of all cluster categories:

Cluster 0:

Cluster Labels	0	1	2	3	4	5	6
Coffee Shop	0.080000	0.040000	0.060000	0.120000	0.320000	0.020000	0.120000
Bar	0.080000	0.000000	0.020000	0.100000	0.160000	0.000000	0.140000
Park	0.080000	0.020000	0.060000	0.160000	0.220000	0.020000	0.080000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Burgers	0.040000	0.040000	0.060000	0.000000	0.020000	0.020000	0.000000
Brewery	0.040000	0.000000	0.000000	0.000000	0.040000	0.000000	0.000000
Tapas	0.040000	0.000000	0.020000	0.000000	0.000000	0.000000	0.000000
French	0.040000	0.000000	0.040000	0.000000	0.040000	0.020000	0.000000
Cocktail	0.040000	0.000000	0.000000	0.000000	0.120000	0.000000	0.000000
Italian	0.040000	0.000000	0.100000	0.020000	0.040000	0.000000	0.020000

Cluster 1:

Cluster Labels	0	1	2	3	4	5	6
Grocery Store	0.000000	0.440000	0.040000	0.200000	0.120000	0.040000	0.020000
Hotel	0.020000	0.220000	0.180000	0.020000	0.080000	0.100000	0.040000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Gym / Fitness	0.000000	0.180000	0.180000	0.100000	0.220000	0.040000	0.020000
Supermarket	0.000000	0.120000	0.080000	0.080000	0.020000	0.040000	0.000000
Furniture / Home	0.000000	0.100000	0.000000	0.020000	0.000000	0.000000	0.000000
Pizza	0.020000	0.100000	0.020000	0.020000	0.180000	0.000000	0.000000
Lodge	0.000000	0.100000	0.000000	0.160000	0.020000	0.000000	0.000000
Electronics	0.000000	0.080000	0.000000	0.000000	0.000000	0.000000	0.000000
Sporting Goods	0.000000	0.080000	0.000000	0.000000	0.000000	0.000000	0.000000

Cluster 2:

Cluster Labels	0	1	2	3	4	5	6
Gym / Fitness	0.000000	0.180000	0.180000	0.100000	0.220000	0.040000	0.020000
Hotel	0.020000	0.220000	0.180000	0.020000	0.080000	0.100000	0.040000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Italian	0.040000	0.000000	0.100000	0.020000	0.040000	0.000000	0.020000
Bakery	0.040000	0.060000	0.100000	0.080000	0.280000	0.040000	0.000000
Wine Shop	0.000000	0.060000	0.080000	0.020000	0.140000	0.020000	0.000000
Supermarket	0.000000	0.120000	0.080000	0.080000	0.020000	0.040000	0.000000
Park	0.080000	0.020000	0.060000	0.160000	0.220000	0.020000	0.080000
History Museum	0.000000	0.000000	0.060000	0.020000	0.020000	0.000000	0.000000
Coffee Shop	0.080000	0.040000	0.060000	0.120000	0.320000	0.020000	0.120000

Cluster 3:

Cluster Labels	0	1	2	3	4	5	6
Beach	0.000000	0.000000	0.060000	0.300000	0.000000	0.020000	0.000000
Grocery Store	0.000000	0.440000	0.040000	0.200000	0.120000	0.040000	0.020000
Sushi	0.000000	0.060000	0.040000	0.160000	0.140000	0.020000	0.040000
Lodge	0.000000	0.100000	0.000000	0.160000	0.020000	0.000000	0.000000
Park	0.080000	0.020000	0.060000	0.160000	0.220000	0.020000	0.080000
Asian	0.020000	0.060000	0.020000	0.120000	0.020000	0.000000	0.040000
Coffee Shop	0.080000	0.040000	0.060000	0.120000	0.320000	0.020000	0.120000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Gym / Fitness	0.000000	0.180000	0.180000	0.100000	0.220000	0.040000	0.020000
Bar	0.080000	0.000000	0.020000	0.100000	0.160000	0.000000	0.140000

Cluster 4:

Cluster Labels	0	1	2	3	4	5	6
Coffee Shop	0.080000	0.040000	0.060000	0.120000	0.320000	0.020000	0.120000
Bakery	0.040000	0.060000	0.100000	0.080000	0.280000	0.040000	0.000000
Gym / Fitness	0.000000	0.180000	0.180000	0.100000	0.220000	0.040000	0.020000
Park	0.080000	0.020000	0.060000	0.160000	0.220000	0.020000	0.080000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Restaurant	0.000000	0.020000	0.040000	0.080000	0.180000	0.040000	0.040000
Gym	0.000000	0.040000	0.000000	0.000000	0.180000	0.000000	0.000000
Pizza	0.020000	0.100000	0.020000	0.020000	0.180000	0.000000	0.000000
Bar	0.080000	0.000000	0.020000	0.100000	0.160000	0.000000	0.140000
Sushi	0.000000	0.060000	0.040000	0.160000	0.140000	0.020000	0.040000

Cluster 5:

Cluster Labels	0	1	2	3	4	5	6
Hotel	0.020000	0.220000	0.180000	0.020000	0.080000	0.100000	0.040000
Ski Area	0.000000	0.040000	0.000000	0.000000	0.000000	0.080000	0.000000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Grocery Store	0.000000	0.440000	0.040000	0.200000	0.120000	0.040000	0.020000
Restaurant	0.000000	0.020000	0.040000	0.080000	0.180000	0.040000	0.040000
Golf Course	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000
Athletics & Sports	0.000000	0.060000	0.000000	0.020000	0.020000	0.040000	0.000000
Trail	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000
Bakery	0.040000	0.060000	0.100000	0.080000	0.280000	0.040000	0.000000
Supermarket	0.000000	0.120000	0.080000	0.080000	0.020000	0.040000	0.000000

Cluster 6:

Cluster Labels	0	1	2	3	4	5	6
Bar	0.080000	0.000000	0.020000	0.100000	0.160000	0.000000	0.140000
Coffee Shop	0.080000	0.040000	0.060000	0.120000	0.320000	0.020000	0.120000
Park	0.080000	0.020000	0.060000	0.160000	0.220000	0.020000	0.080000
Indian	0.040000	0.000000	0.040000	0.000000	0.120000	0.000000	0.060000
Café	0.060000	0.180000	0.140000	0.120000	0.200000	0.060000	0.060000
Asian	0.020000	0.060000	0.020000	0.120000	0.020000	0.000000	0.040000
Hotel	0.020000	0.220000	0.180000	0.020000	0.080000	0.100000	0.040000
Sushi	0.000000	0.060000	0.040000	0.160000	0.140000	0.020000	0.040000
Record Shop	0.000000	0.000000	0.000000	0.040000	0.000000	0.000000	0.040000
Restaurant	0.000000	0.020000	0.040000	0.080000	0.180000	0.040000	0.040000

References

1) **Geopy Python Library:** https://geopy.readthedocs.io/en/stable/#
2) **Foursquare API:** https://developer.foursquare.com/docs/api-reference/venues/explore/
3) **Wikipedia, List of Boroughs in Oslo:** https://no.wikipedia.org/wiki/Liste_over_Oslos_bydeler
4) **Photo of Oslo:** https://www.google.com/search?q=oslo&sxsrf=ALeKk00NVsLogc2aFmwSC_E7_4SMouOV2w:1613828060396&source=lnms&tbm=isch&sa=X&ved=2ahUKEwixhpKCyvjuAhUjlosKHfJnBTYQ_AUoAnoECB4QBA&biw=1065&bih=1504#imgrc=5_LFi2GQzreMXM