Applied Data Science Capstone

by Alizamin Jafarli

&

Capstone Project

Coursera

June, 2019

Table of Contents

Figures

1 Introduction[1]

1.1 Problem

Migration has recently been one of the hottest topics in politics. In the book "Statup Migrants" by Nicolai Strøm-Olsen and Maria Amelie[2], it is claimed that the first-generation migrants usually launch conventional businesses like restaurant/café, grocery stores to provide their families when they cannot find relevant jobs and those businesses usually require less capital relative to tech startups. So, the problem I will try to shed some light on is if immigrants actually contribute to small businesses in the country where they settle down. In our example, this will be Oslo, Norway.

The purpose here is not to argue in favor of or against the book. Instead, I will try to find out if there is any relationship between immigrants and less capital-heavy businesses in Oslo, Norway. Clustering method of unsupervised machine learning has been deployed to group boroughs in Oslo. Then I have benefited visualization tools Python offers to find any pattern in data.

[1] Introduction where you discuss the business problem and who would be interested in this project.
[2] Read more here: https://www.amazon.com/Startup-Migrants-Nicolai-Str%C3%B8m-Olsen/dp/8293097574

2 Data description[3]

2.1 Data sources

Data that have been used for this analysis freely available on the internet. The overview over boroughs in Oslo scraped from the Wikipedia[4]. The latitude and longitude of Oslo boroughs have been collected via GeoPy[5]. The data on immigration population have been taken from Statistics Bank of Oslo Municipality[6]. The numbers on social security receivers can be found on the website of Norwegian Statistics Bureau[7], The borough coordinates in json-format for folium maps are easily accessible on the internet as well[8]. Finally, Foursquare API has been used to fetch venues in Oslo boroughs[9]. Though the data is available on the internet, the quality of data has been an issue. First of all, it was challenging to collect data from different sources and merge them in an efficient way. Secondly, you the only json data I found did not return right coordinates for some boroughs. Moreover, with one borough I had to change the coordinates manually that GeoPy returned.

The following variables constitute analysis:

- "Borough": The boroughs in Oslo
- "Latitude": Latitude coordinates for boroughs
- "Longitude": Longitude coordinates for boroughs
- "AreaInKm2": Total area of the boroughs
- "ImmigrantsTotal": Number of people with immigrant background
- "SocialSecurity": Number of social security receivers
- "Venue": Venue data fetched from Foursquare API

[3] Data where you describe the data that will be used to solve the problem and the source of the data.
[4] Read more here: https://no.wikipedia.org/wiki/Liste_over_Oslos_bydeler
[5] Read more here: https://geopy.readthedocs.io/en/stable/
[6] Read more here: http://bit.ly/2F2ozb8
[7] Read more here: https://www.ssb.no/statbank/table/12404/
[8] Read more here: https://hannemelling.carto.com/tables/bydeler_xls/public/map
[9] Read more here: https://developer.foursquare.com/places-api

3 Methodology[10]

To explore the problem, numerous visualizations and unsupervised machine learning algorithm have been deployed. I have actively used histograms and folium library for choropleth maps.

Though the problem to be discussed is a bit more econometrics problem and so, maybe the regression might be a better method, we are supposed to use Foursquare API. Consequently, I went for clustering K-Means method, and see if it is any pattern around, i.e. how densely small businesses are located where immigrant population abounds.

There were several challenges that I was aware of even if they might have implications for my analysis:

- Most small businesses started by immigrants are usually not present online, especially on Foursquare. My subjective observation tells that this target group is less tech-savvy and lacks online marketing.

- Among the venues that Foursquare spits out, there are many common outside areas like parks, bus / tram / railway stations. Nevertheless, those type of areas usually have plenty of businesses, especially seasonal that are non-existent online.

[10] Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

4 Results[11]

I started first scraping data from Wikipedia to get the following table:

	Borough	Population	AreaInKm2	BoroughNumber
0	Alna	49 358	137	12
1	Bjerke	31 973	77	9
2	Frogner	58 283	83	5
3	Gamle Oslo	54 575	75	1
4	Grorud	27 525	82	10
5	Grünerløkka	58 906	48	2
6	Nordre Aker	50 724	136	8
7	Nordstrand	51 169	169	14
8	Sagene	43 131	31	3
9	St. Hanshaugen	38 109	36	4
10	Stovner	32 850	82	11
11	Søndre Nordstrand	38 925	184	15
12	Ullern	33 463	94	6
13	Vestre Aker	48 605	166	7
14	Østensjø	49 968	122	13

Figure 1. Oslo boroughs

Then via GeoPy, I managed to get geocoordinates for boroughs. Please, notice that I had to make amendments to the geocoordinates of *Alna*, since the returned information by GeoPy was wrong. Furthermore, a column on immigrant population is added:

[11] Results section where you discuss the results.

	Borough	Population	AreaInKm2	Latitude	Longitude	ImmigrantsTotal
0	Alna	49 358	137	59.929854	59.929854	10.817046
1	Bjerke	31 973	77	59.823297	10.851059	14405.000000
2	Frogner	58 283	83	59.909640	10.687961	16498.000000
3	Gamle Oslo	54 575	75	59.899237	10.734767	21839.000000
4	Grorud	27 525	82	59.962343	10.875290	14011.000000
5	Grünerløkka	58 906	48	59.925471	10.777421	21141.000000
6	Nordre Aker	50 724	136	59.953638	10.756412	9518.000000
7	Nordstrand	51 169	169	59.870880	10.780353	9201.000000
8	Sagene	43 131	31	59.938273	10.765849	11332.000000
9	St. Hanshaugen	38 109	36	59.927950	10.738958	10666.000000
10	Stovner	32 850	82	59.959292	10.924499	19021.000000
11	Søndre Nordstrand	38 925	184	59.835944	10.798496	21779.000000
12	Ullern	33 463	94	59.925818	10.665132	6717.000000
13	Vestre Aker	48 605	166	59.958300	10.670319	8726.000000
14	Østensjø	49 968	122	59.887563	10.832748	13106.000000

Figure 2. Geocoordinates and immigrant population

Where immigrant population in Oslo mainly lives is given in the following histogram:

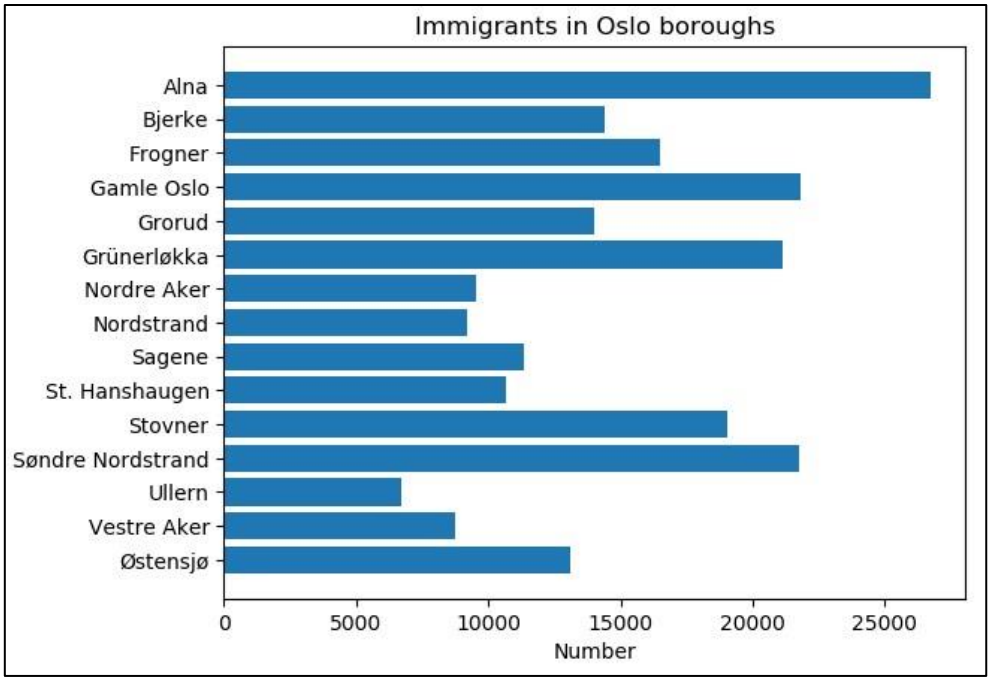

Figure 3. Histogram visualisation of immigrant population living in Oslo boroughs

To check if geocoordinates of Oslo. I visualised them on a folium map (that is how I found one mistake as well):

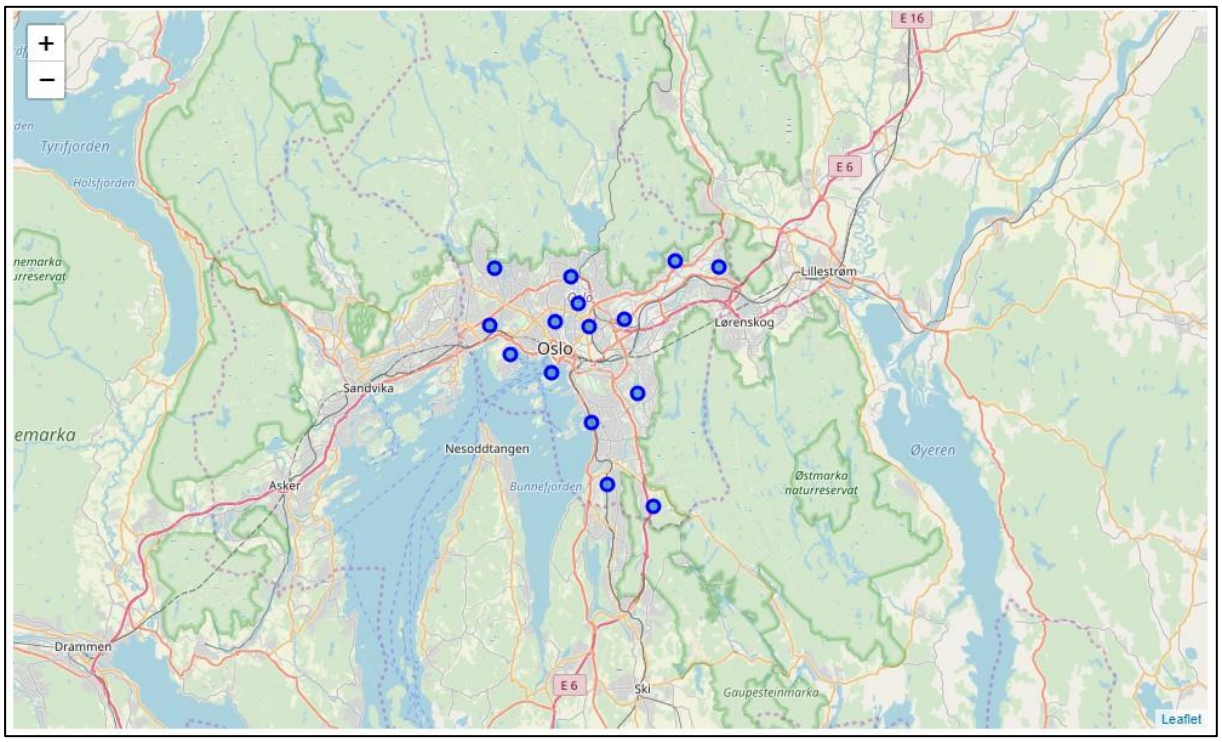

Figure 4. Borough centres visualised on the map

Just initial exploration of top venues in Alna let me know that some boroughs have very few data which is problematic:

	name	categories	lat	lng
0	Holmlia stasjon	Train Station	59.835016	10.796475
1	kiwi	Grocery Store	59.836004	10.796451
2	Vitus Apotek	Pharmacy	59.834435	10.794882
3	Meny	Grocery Store	59.834244	10.794893
4	ACTIC Norge	Gym	59.834470	10.792411
5	Rema 1000	Grocery Store	59.832007	10.796556
6	Holmliahallen	Athletics & Sports	59.834329	10.791173

```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
7 venues were returned by Foursquare.
```

Figure 5. Some boroughs have very few data / venue

Notice that when you try to apply the same function to fetch venues in all boroughs, sometimes Foursquare API does not return any result for *Bjerke*. The total number of venues should be 141:

	Borough	Population	AreaInKm2	Latitude	Longitude	ImmigrantsTotal	NumberVenues
0	Alna	49 358	137	59.929854	10.817046	26723	7
1	Bjerke	31 973	77	59.941395	10.829208	14405	10
2	Frogner	58 283	83	59.909640	10.687961	16498	5
3	Gamle Oslo	54 575	75	59.899237	10.734767	21839	9
4	Grorud	27 525	82	59.962343	10.875290	14011	5
5	Grünerløkka	58 906	48	59.925471	10.777421	21141	14
6	Nordre Aker	50 724	136	59.953638	10.756412	9518	7
7	Nordstrand	51 169	169	59.870880	10.780353	9201	6
8	Sagene	43 131	31	59.938273	10.765849	11332	27
9	St. Hanshaugen	38 109	36	59.927950	10.738958	10666	30
10	Stovner	32 850	82	59.959292	10.924499	19021	4
11	Søndre Nordstrand	38 925	184	59.835944	10.798496	21779	7
12	Ullern	33 463	94	59.925818	10.665132	6717	4
13	Vestre Aker	48 605	166	59.958300	10.670319	8726	5
14	Østensjø	49 968	122	59.887563	10.832748	13106	1

Figure 6. Total venues in Oslo are 141

When I tried to visualised immigrant population and number of venues on folium, I noticed that the only json file I found for Oslo boroughs delivers wrong geocoordinates for *Grunerløkka, Søndre Nordstrand and Østensjø*. From maps below man may see the exact same “white spots” of both boroughs on both maps.

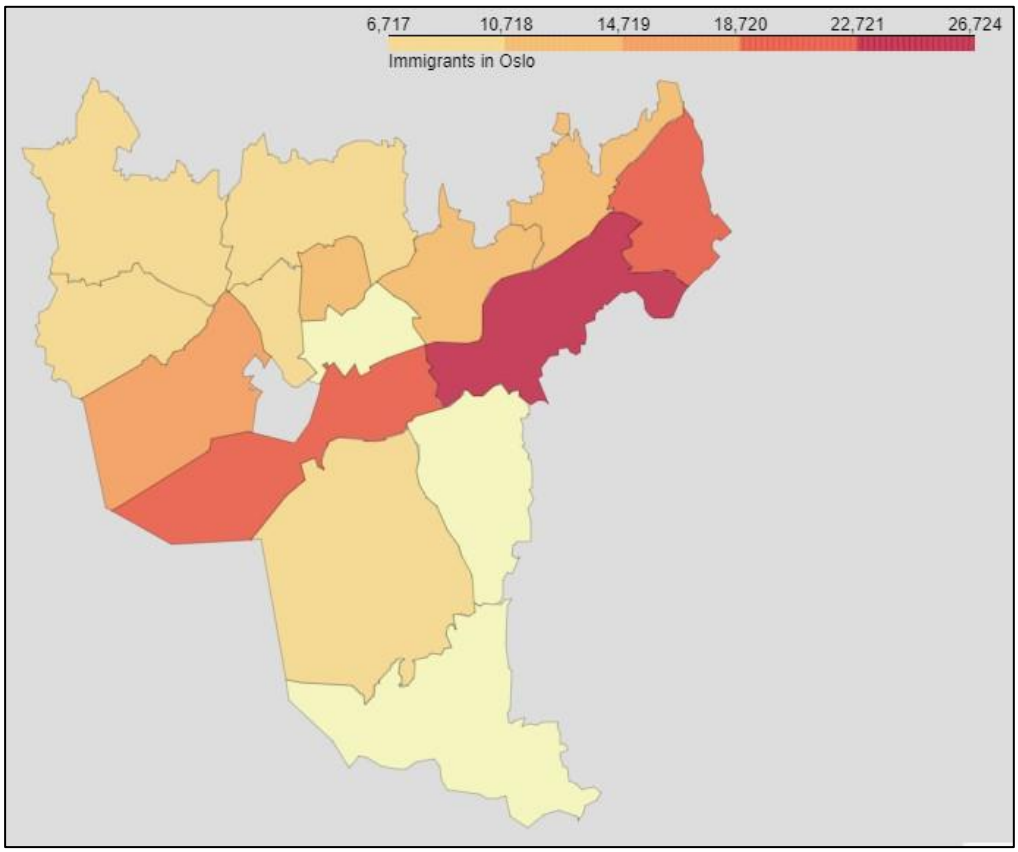

Figure 7. Map visualisation of immigrant population

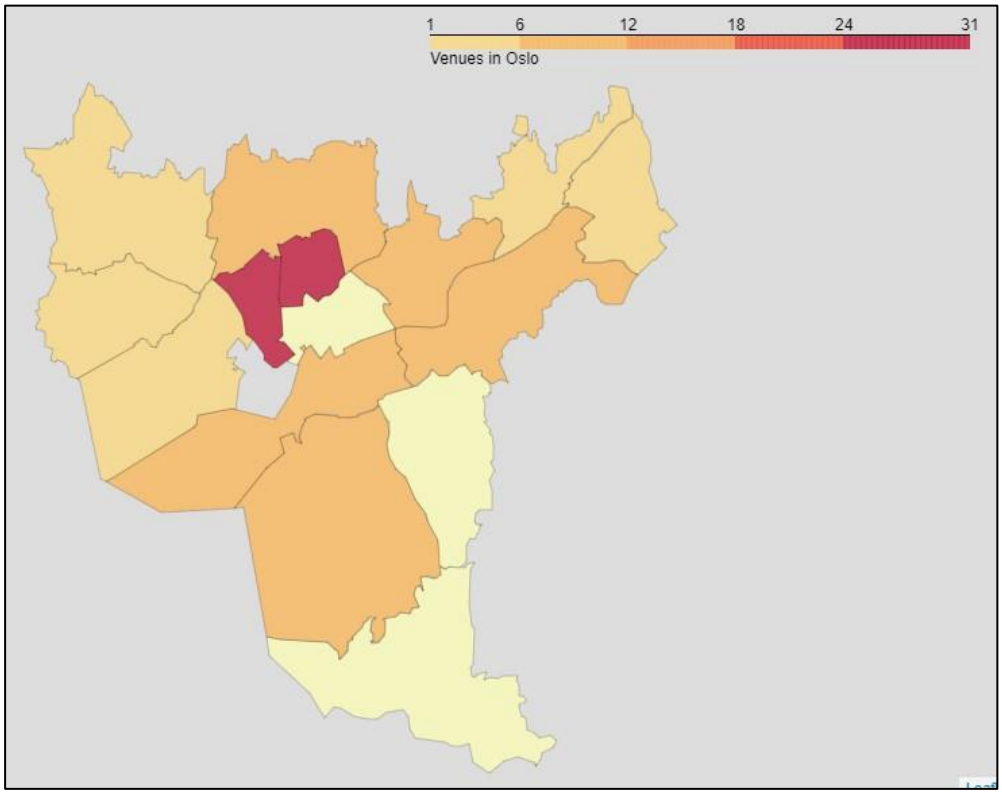

Figure 8. Map visualisation of total venues

As mentioned above, I applied clustering to see venues with most immigrant population will produce any result. 5 clusters after choosing top 5 venues for each borough, and got the following results:

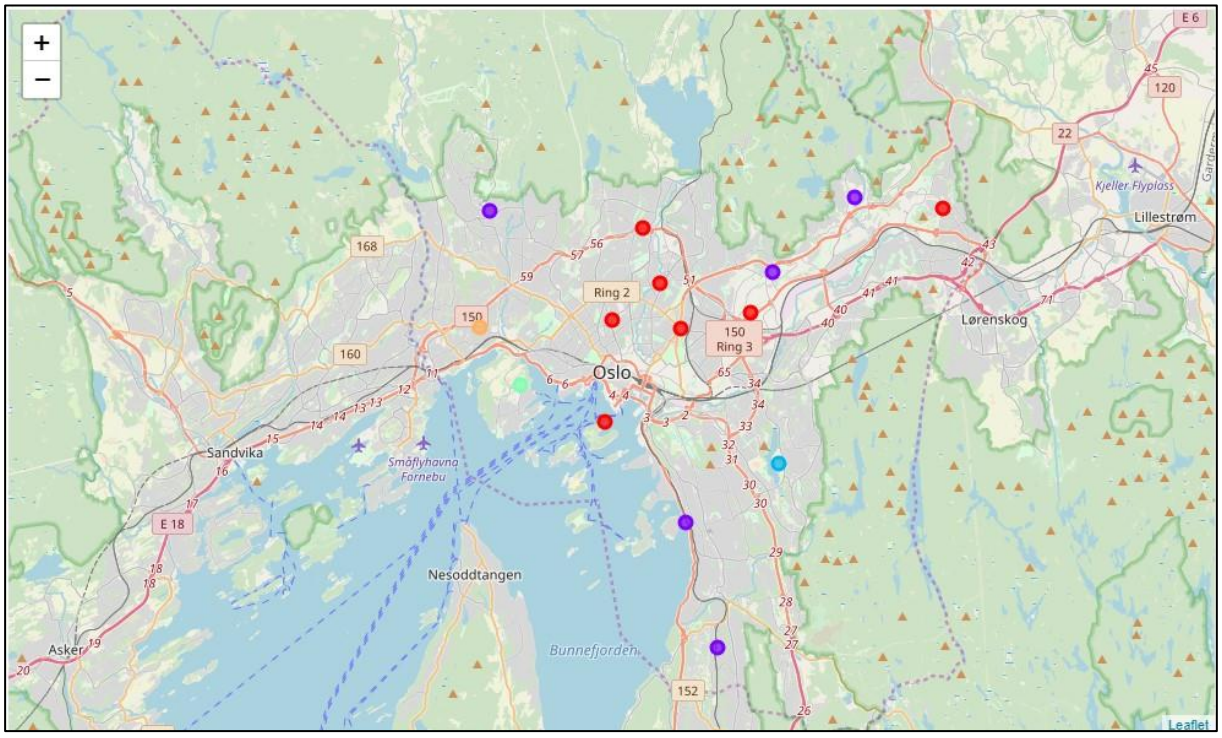

Figure 9. Clustering of venues

5 Discussion[12]

So, is there any pattern? In general, small sample is a evidently problem. But let's have a look at each cluster:

Cluster 1

```
oslo_merged.loc[oslo_merged['Cluster Labels'] == 0, oslo_merged.columns[[0] + list(range(5, oslo_merged.shape[1]))]]
```

	Borough	ImmigrantsTotal	NumberVenues	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alna	26723	7	0	Park	Auto Workshop	Hotel	Hotel Bar	Nature Preserve
3	Gamle Oslo	21839	9	0	Boat or Ferry	Café	Historic Site	Other Nightlife	Pier
5	Grünerløkka	21141	14	0	Gym / Fitness Center	Sushi Restaurant	Asian Restaurant	Indian Restaurant	Coffee Shop
6	Nordre Aker	9518	7	0	Bus Stop	Bus Station	Park	Grocery Store	Metro Station
8	Sagene	11332	27	0	Bakery	Park	Pizza Place	Sushi Restaurant	Coffee Shop
9	St. Hanshaugen	10666	30	0	Bakery	Park	Coffee Shop	Grocery Store	Wine Shop
10	Stovner	19021	4	0	Athletics & Sports	Fast Food Restaurant	Scenic Lookout	Metro Station	Wine Shop

Cluster 2

```
oslo_merged.loc[oslo_merged['Cluster Labels'] == 1, oslo_merged.columns[[0] + list(range(5, oslo_merged.shape[1]))]]
```

	Borough	ImmigrantsTotal	NumberVenues	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Bjerke	14405	10	1	Grocery Store	Café	Hotel	Convenience Store	Bakery
4	Grorud	14011	5	1	Wine Shop	Bus Station	Grocery Store	Supermarket	Metro Station
7	Nordstrand	9201	6	1	Grocery Store	Health & Beauty Service	Harbor / Marina	Beach	Bus Station
11	Søndre Nordstrand	21779	7	1	Grocery Store	Athletics & Sports	Train Station	Gym	Pharmacy
13	Vestre Aker	8726	5	1	Grocery Store	Restaurant	Metro Station	Ski Area	Wine Shop

Figure 10. Cluster 1 and 2

[12] Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

Cluster 3

```
oslo_merged.loc[oslo_merged['Cluster Labels'] == 2, oslo_merged.columns[[0] + list(range(5, oslo_merged.shape[1]))]]
```

	Borough	ImmigrantsTotal	NumberVenues	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
14	Østensjø	13106	1	2	Bar	Wine Shop	French Restaurant	Diner	Dog Run

Cluster 4

```
oslo_merged.loc[oslo_merged['Cluster Labels'] == 3, oslo_merged.columns[[0] + list(range(5, oslo_merged.shape[1]))]]
```

	Borough	ImmigrantsTotal	NumberVenues	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Frogner	16498	5	3	History Museum	Harbor / Marina	Scandinavian Restaurant	Farm	Flower Shop

Cluster 5

```
oslo_merged.loc[oslo_merged['Cluster Labels'] == 4, oslo_merged.columns[[0] + list(range(5, oslo_merged.shape[1]))]]
```

	Borough	ImmigrantsTotal	NumberVenues	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	Ullern	6717	4	4	Light Rail Station	Bus Station	Flower Shop	Dog Run	Farm

Figure 11. Cluster 3,4 ,5

Cluster 5: We had only 4 venues returned: 2 rail stations, 1 bus station and flower shop. Dog run and farm are just random outputs which is not present in Ullern. So, it should be no surprise why stations were clustered together.

Cluster 4: 5 venues for this borough. The borough is rich with museums and nature which might be reason for a separate clustering.

Cluster 3: We had only one venue for Østensjø, and that one venue weighs heaviest as a result. This is small sample problem.

Cluster 2: Groceries abound in this cluster but mix of immigrant heavy and light boroughs.

Cluster 1: Bakeries rule in this cluster, and mainly immigrant heavy eastern boroughs (Alna, Stovner) and central boroughs (Gamle Oslo, Grunerløkka, Sagene, St. Hanshaugen) in this cluster except Nordre Aker.

In Figure 3, we see that immigrant population has settled down mainly in east and centre. When it comes to venues, they are located mainly in the centre:

Figure 12. Immigrant population and total venues

N	Immigrant Population	Most Venues
1	Alna (east)	St. Hanshaugen (centre west)
2	Gamle Oslo (centre)	Sagene (centre)
3	Søndre Nordstrand (east south)	Grunerløkka (centre)
4	Grunerløkka (centre)	Bjerke (centre east)
5	Stovner (east)	Gamle Oslo (centre)
6	Frogner (west, many foreign companies)	Alna (east)
7	Bjerke (east)	Nordre Aker (west)
8	Grorud (east)	Søndre Nordstrand (east south)

So, we see that the intersection immigrant population and venues is the central part of Oslo. But how much of these small businesses are due to natural tendency for businesses to be in the city center or the engagement of immigrant population is hard to examine. But those eastern regions like Alna, Østensjø, Grorud and Stovner do not have centrality bias, but they have either few venues or data. Let's have a look at social security receivers in Oslo to examine the economic situation in Oslo. Mind that the data cover all irrespective of cultural background:

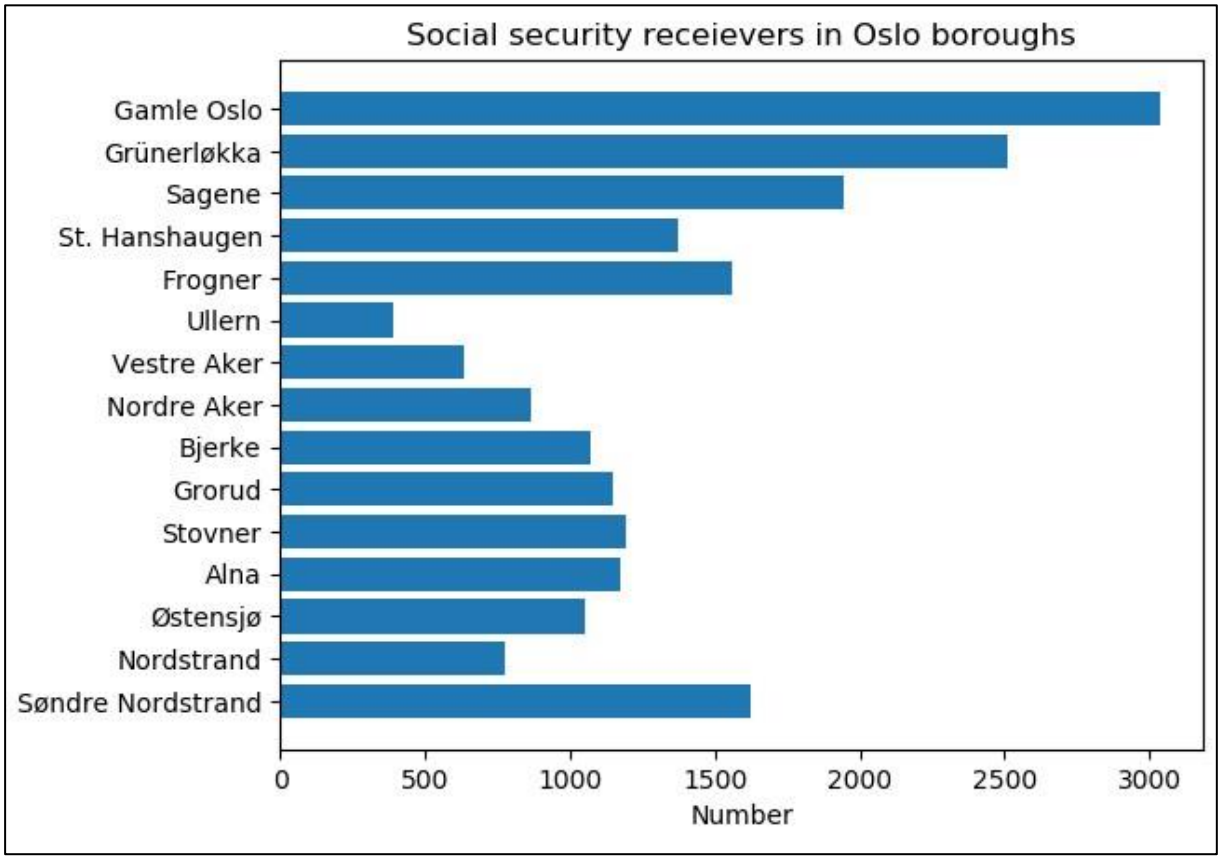

Figure 13. Total social security receivers in Oslo boroughs

N	Immigrant Population	Top Social Security Receivers
1	Alna (east)	Gamle Oslo (centre)
2	Gamle Oslo (centre)	Grunerløkka (centre)
3	Søndre Nordstrand (east south)	Sagene (centre)
4	Grunerløkka (centre)	Søndre Nordstrand (east south)
5	Stovner (east)	Frogner (west)
6	Frogner[13] (west)	St. Hanshaugen (centre west)
7	Bjerke (east)	Stovner (east)
8	Grorud (east)	Alna (east)

[13] Many foreign companies have offices in Frogner

6 out of 8 top rows under Immigrant Population and Top Social Security Receivers coincide.

	AreaInKm2	Latitude	Longitude	ImmigrantsTotal	NumberVenues	SocialSecurity
AreaInKm2	1.000000	-0.429327	-0.045028	-0.006455	-0.632242	-0.469866
Latitude	-0.429327	1.000000	0.096902	-0.235440	0.123390	-0.218143
Longitude	-0.045028	0.096902	1.000000	0.402237	-0.149014	0.027611
ImmigrantsTotal	-0.006455	-0.235440	0.402237	1.000000	-0.109468	0.601718
NumberVenues	-0.632242	0.123390	-0.149014	-0.109468	1.000000	0.366228
SocialSecurity	-0.469866	-0.218143	0.027611	0.601718	0.366228	1.000000

Figure 14. Correlation table

The correlation between social security receivers and immigrant population is high enough: 0.60 which may explain risk aversion to start private business. This is one of the inferences in the book, i.e. the higher welfare state, the less incentive to start a private business.

6 Conclusion[14]

In this analysis, I tried to find a pattern between immigrant population and small businesses in Oslo boroughs. Given the dataset, there are plenty of businesses in central Oslo where immigrant population densely populated. But that might be a natural tendency for all small business to be in city center, something that we might observe everywhere. So, it is hard to name it as causality. However, in the eastern part of Oslo, far from the city center and immigrant population dominates, it is hard to find an evidence that supports our initial theory. Moreover, Norway is a welfare state, and it looks like it is a demotivator start private business for immigrant population as it is mentioned in the book.

Bear in mind that it was hard to find the relevant data, and it is totally plausible that our analysis suffers from small sample bias. The challenges I have described in chapter 3 may also have consequences for the analysis.

[14] Conclusion section where you conclude the report.