A street level clustering analysis of transport options in Oslo

Project Report for
IBM Data Science Professional Certification

Niladri Banerjee, PhD

1. Introduction

1.1 Background

Recently, Oslo won the title of the **European Green Capital:**
https://www.visitoslo.com/en/articles/oslo-european-green-capital-2019/

There is an increased emphasis from the Norwegian government on a greener, more sustainable future. One of the ways in which this is being planned is by curbing the use of fossil fuels. To this end, the government has introduced a slew of measures that make it costlier for an individual to own or drive a car, whilst trying to incentivise the increased use of public transport.

However, not all areas of the country have an equitable distribution of public transport options. To explore this, I decided to investigate the status in Oslo. After all, people are more likely to use public transport if it is readily available.

1.2 Business Problem

As a *hypothetical case study,* the transport department approaches me to help them decide where to build more transport options in Oslo.

My job, as a data science consultant is therefore to help them decide which areas of Oslo are lacking adequate transport infrastructure and where they should invest.

One way to solve this problem is to cluster neighbourhoods/areas based on the number of transport options available and visualise them on a map, that can then help pinpoint areas needing improved connectivity.

The idea here would be- find neighbourhoods that have transport options within 400m, which is a reasonable walking distance that does not take too long to cover. Henceforth, we will search for transport options within a 400m radius of a given geo-location

To accomplish this task, one would need data on *a)* the different types of transport options available *b)* where they are available and *c)* how many are available.

Such information may be obtained through the **Foursquare API.** Additionally, one would need geographic location data of postcode/streets to pin-point the areas that need improvement.

2. Data

2.1 Points to consider

Before we can commence our project, it is important to consider the following points-

- In Oslo kommune/municipality, a given street can belong to different postcodes
- Furthermore, a postcode is not very intuitive to understand/place in one's mind. If someone says "*Oh I live at 0125 postcode*", it is not very easy to mentally think where is that. Instead, if the person says, "*Oh I just live by the Blåsteingata",* then one

can instantly place where they live (assuming of course, that one is familiar with the street names)

- So, in this project, we need to make a decision how deep a level to drill down, w.r.t defining an area around which to find transport options.
- Additionally, having postcode info may lead to redundant street level data: *i.e.* 1 street = several postcodes. *Conversely*, 1 postcode = several streets, with sometimes the same street occurring in different postcodes

Based on the above points, since it is not very intuitive to understand the location of postcodes, and nobody can place them in their head, I will try to find transport options down to the street-level and not postcode level.

But then the question comes, how to define coordinates of street? One possible way may be to use mid points of street start and street end. For this to work, we would need to make an assumption that for a small section of the Earth such as a street, the distance between two points is basically a straight line (*unlike a curved line for large distances)*

2.2 Data availablity

For our project concerning transport options at a street-level, unfortunately **Foursquare API** has no direct way to find all transport options. Instead one would need to manually extract information about:

1. Trikk/Tram
2. Bus Stop/Bus Station
3. T-bane/Metro
4. Train Station

Furthermore, based on my overall experience, Foursquare data is not very good for Oslo, compared to USA/Canada, possibly due to insufficient awareness/use. This made it a considerable challenge to extract relatively clean information regarding various transport modes.

Additionally, to obtain the geo-coordinates of different streets in Oslo will not be an easy task. This information is not readily available. One possible way is with the help of the website created by Erik Bolstad.

- It provides geo-coordinates of all the different bydel/districts of Oslo
- Additionally, within each bydel, it maps out the street addresses which we will need to obtain.

3. Methods

The methods employed primarily fall under the following categories with full details available on GitHub under Part 1 (Web Scraping), Part 2 (Foursquare API), Part 3 (EDA), Part 4 (Clustering):

3.1 Web-scraping- HTML, JSON

The geo-coordinate data necessary for carrying out the project was available on the website of Erik Bolstad , but required significant amounts of parsing. Part of the challenge was the fact that street-addresses were not readily available. They were linked to each postcode.

So, one first needed to find in the HTML where the postcodes were. Not all postcodes could be used either. There were many that one needed to omit e.g. those marked ‘Ikke i bruk’ meaning ‘Not in use’ and others that belonged to postboxes/ service areas/ VIP areas.

Using the beautifulsoup4 and the urllib libraries, 442 postcodes were drilled into and from each of these all the street addresses were extracted. Each street address had its own geo-coordinates, so I determined the mid-point by taking the first and last street-address and assigning that to the street-name. Another challenge was the non-ASCII nature of Norwegian letters- æ, ø, å. This was resolved by the requests library in Python.

The next step required obtaining information about the various transport modes available in Oslo. We were mandated to use the Foursquare API, whose data quality, compared to USA/Canada is much less reliable. There were challenges in extracting information about Buses, Trains, Trams and Metro. Several times the ‘category’ information was missing which made it frustratingly difficult to obtain clean data. As an example, several buses did not have the ‘category’ information. I also found that some bus stops did not have ‘category’ information. Extracting simply by the ‘name’ field in the JSON file led to its own challenges of extracting places that had nothing to do with buses. The other challenge was in extracting ‘Bus’ over ‘Buss’ because in Norwegian it is usually spelled ‘Buss’. Ultimately, I only focused on buses that had route numbers associated with them. In the process, I may have lost information on certain streets that only had bus stop information but no bus route number associated.

Whilst searching for train stations, the Foursquare API also failed to retrieve the ‘Nationaltheatret’ station. This I added manually. Using the mid-point of the street name where it was located, I assigned the station to all streets whose mid-points also lay within 400m from the street where the station was located (Rusløkkveien)

3.2 Data wrangling in Python- Pandas, NumPy

The data from web scraping was converted to pandas data-frames for analysis. I also used the numpy library in conjunction with the pandas library for reshaping data and analysing for example the distribution of various streets in different clusters.

3.3 Machine Learning, using K-Means Clustering to cluster the streets.

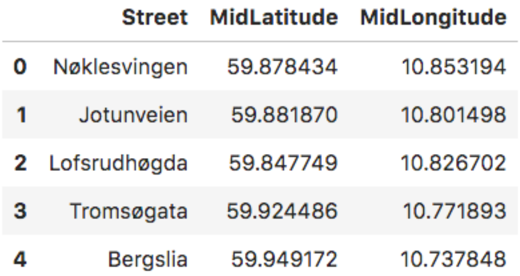
To make the initial visualisation with the help of the Folium map, I assigned the streets to different groups. However, upon performing machine learning with the K-means clustering from the scikit-learn library, I observed different groups. This is one of the benefits of machine learning, in the sense that it gives insight to data in ways that is not manually possible. For example, while I had grouped all streets with 0-3 options into one category, the K-means algorithm actually broke this group into further clusters and it made sense while doing so.

The algorithm broke it into clusters with 0 options, 1 option and then 2-3 options. This makes sense since there is a big difference between streets where there are no options, versus streets having only 1 option. Furthermore, it makes sense to differentiate between streets with just 1 option and slightly more in the range of 2-3 options.

4. Results

4.1 Obtaining street coordinates

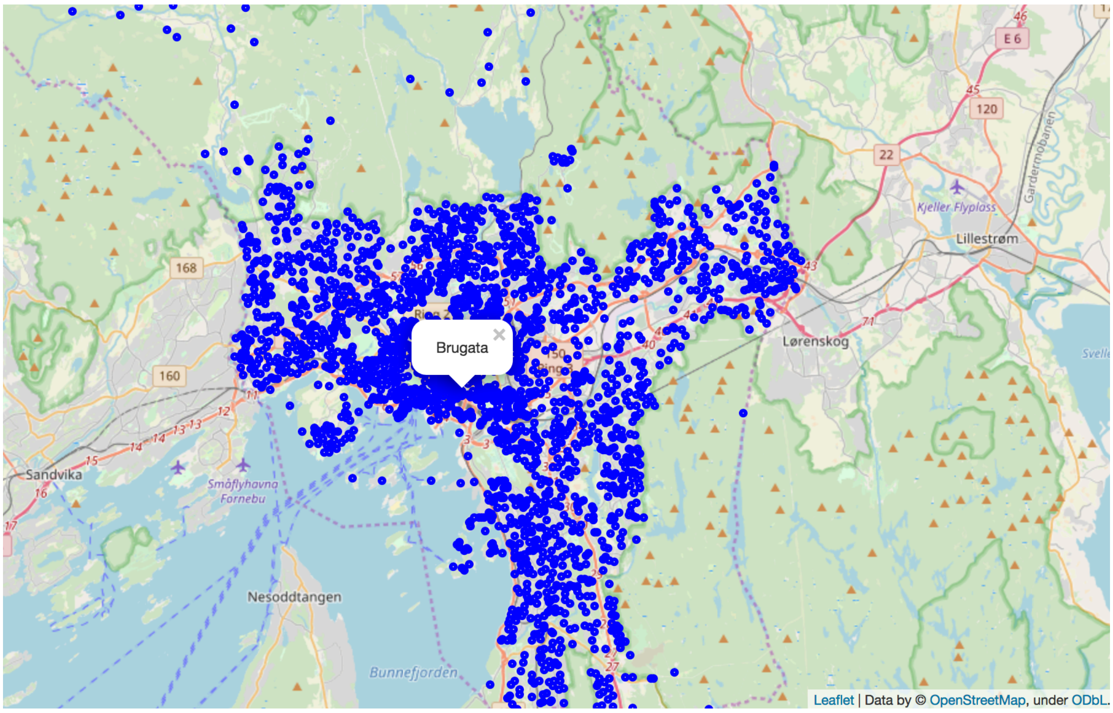
After parsing through **442** postcodes in Oslo, I obtained street coordinates for **2460** streets. In the process, I omitted extracting information from postcodes for postboxes, service postcodes and other postcodes not in use. Here is a snapshot of the first 5 streets from the table:



	Street	MidLatitude	MidLongitude
0	Nøklesvingen	59.878434	10.853194
1	Jotunveien	59.881870	10.801498
2	Lofsrudhøgda	59.847749	10.826702
3	Tromsøgata	59.924486	10.771893
4	Bergslia	59.949172	10.737848

4.2 Basic Visualisation of streets

Using the Folium package, I subsequently generated a basic map of Oslo, plotting all the street names on it. One of the great things about Folium is the ability to generate *point-and-click* markers, as shown below:

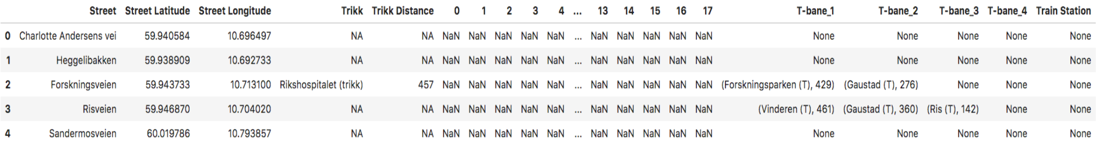


Leaflet | Data by © OpenStreetMap, under ODbL.

4.3 Obtain transport options for each street

After much data wrangling and munging with the extremely messy Foursquare data, including missing labels and categories (*and Foursquare completely missing the*

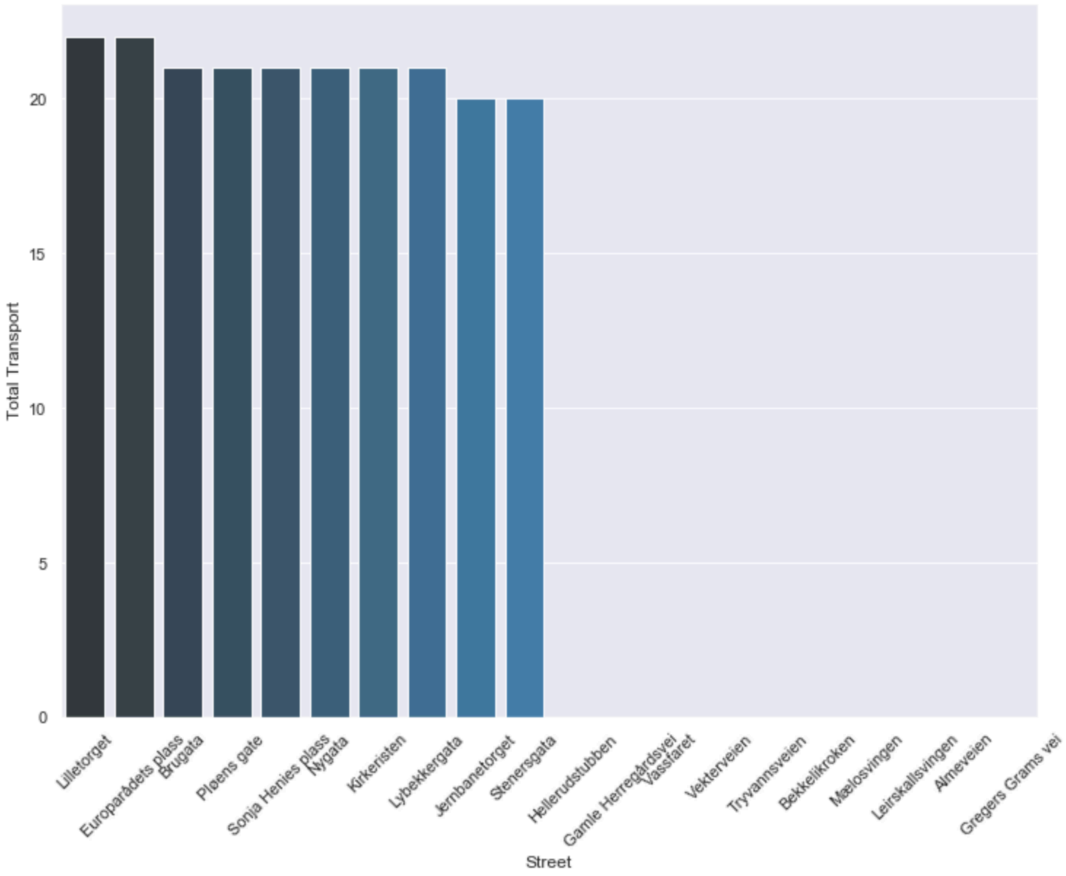
Nationaltheatret train station!), the following table showing the various transport options (**Trikk, Busses, T-bane and Train**) for each street in Oslo was generated.



	Street	Street Latitude	Street Longitude	Trikk	Trikk Distance	0	1	2	3	4	...	13	14	15	16	17	T-bane_1	T-bane_2	T-bane_3	T-bane_4	Train Station
0	Charlotte Andersens vei	59.940584	10.696497	NA	NA	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	None	None	None	None	None
1	Heggelibakken	59.938909	10.692733	NA	NA	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	None	None	None	None	None
2	Forskningsveien	59.943733	10.713100	Rikshospitalet (trikk)	457	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	(Forskningsparken (T), 429)	(Gaustad (T), 276)	None	None	None
3	Risveien	59.946870	10.704020	NA	NA	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	(Vinderen (T), 461)	(Gaustad (T), 360)	(Ris (T), 142)	None	None
4	Sandermosveien	60.019786	10.793857	NA	NA	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	None	None	None	None	None

4.4 Streets with highest and lowest transport options

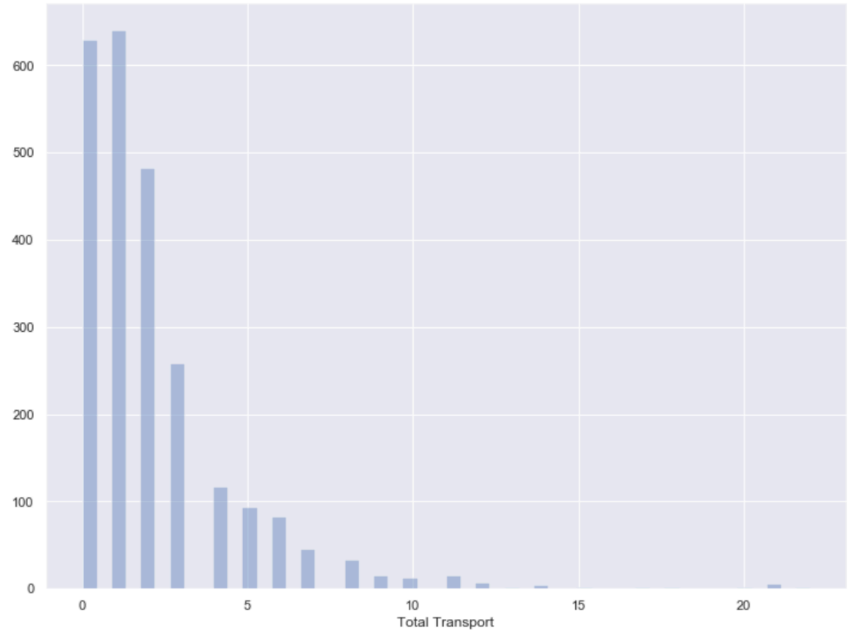
Following the generation of the above table, I computed the total transport options available for each street and then plotted the 10 streets with the most and least number of transport options available.



As we can see, the streets that have the most transport options have as many **22 options available.** On the other hand, there are also streets where there is **No transport** available within a radius of 400m.

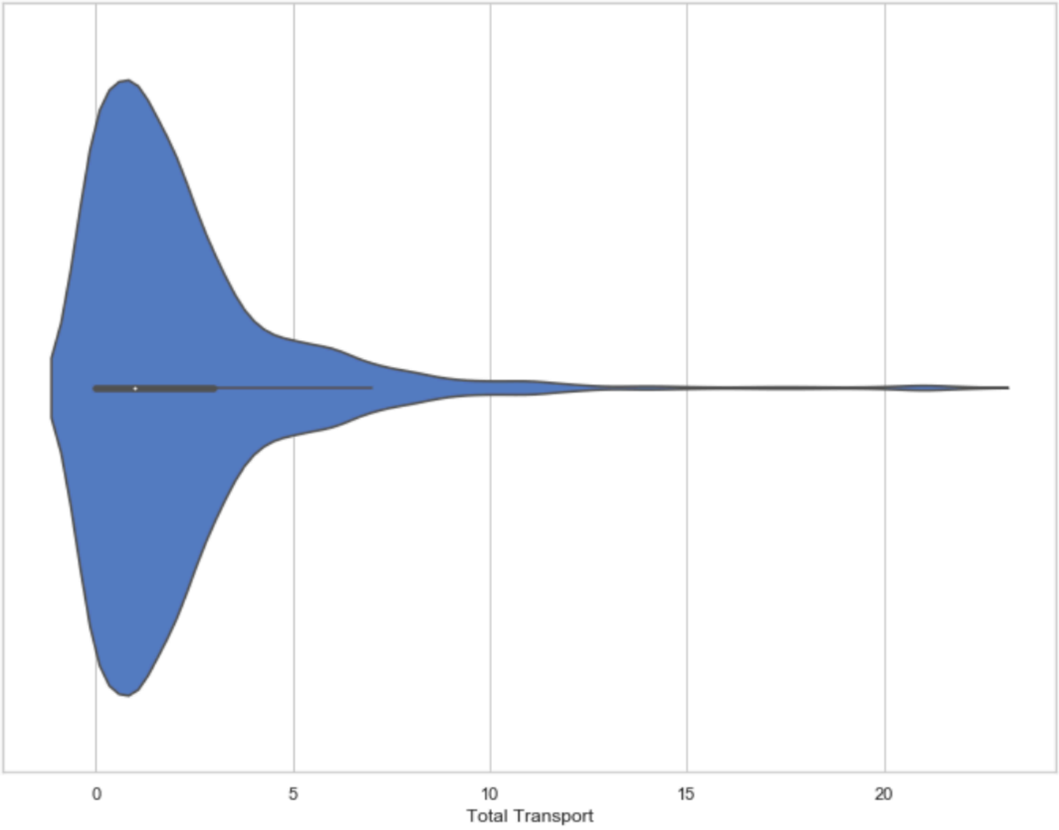
4.5 Distribution of transport options

Let us find out how many streets actually have so many transport options or conversely streets with limited number of transport options.



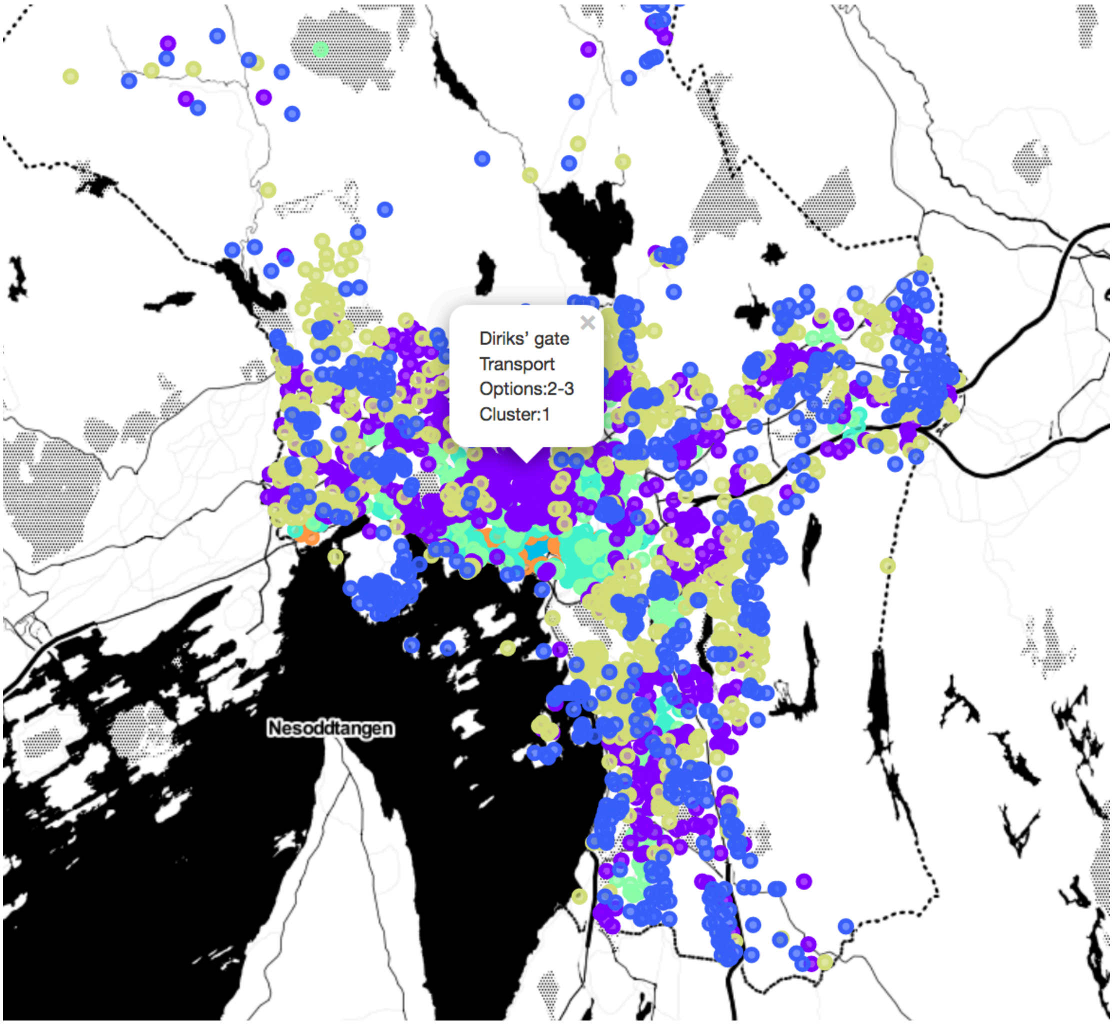
We come to know that an astonishingly high number of streets, in fact, more than **half of all streets** in Oslo, have **only 1 or no transport** available within 400m.

This is confirmed through the violin plot that shows a very high skewness, with the majority of streets having **less than 5 transport options**



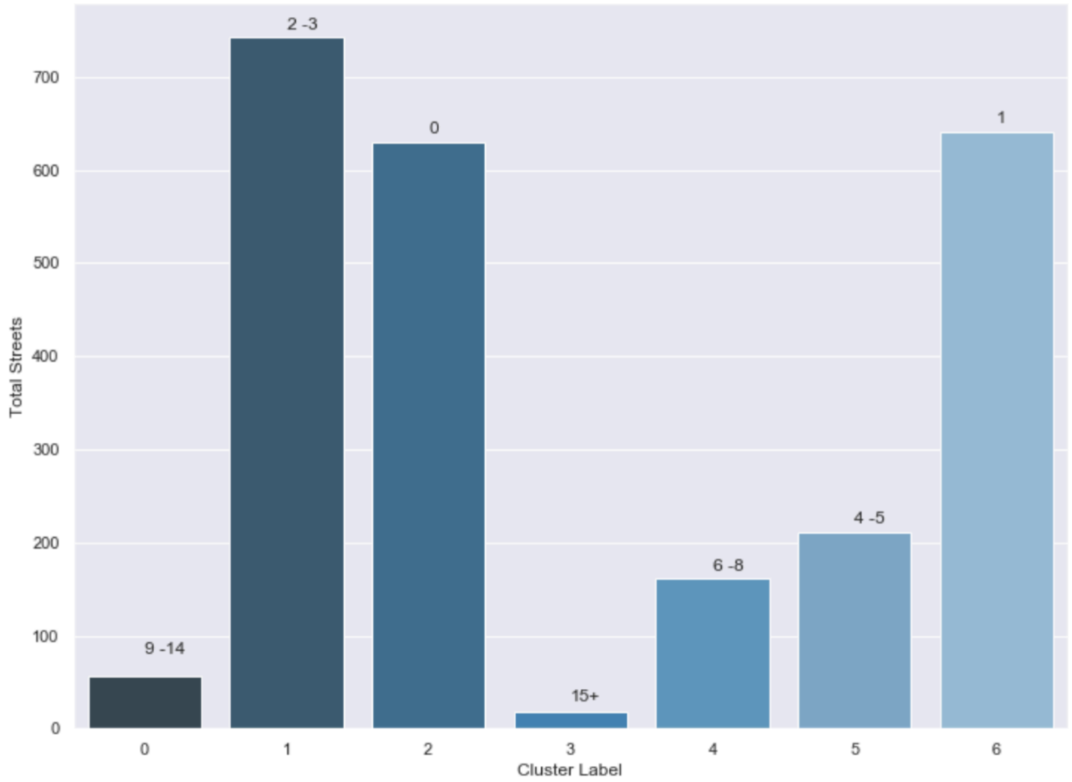
4.6 Clustering and Visualising

Using K-Means clustering and generating 7 clusters, we create the following map that helps group all the streets in Oslo based on the number of transport options available:



From the map, all the purple spots are streets where there are 2-3 transport options available while the blue spots represent areas with the least amount of transport options.

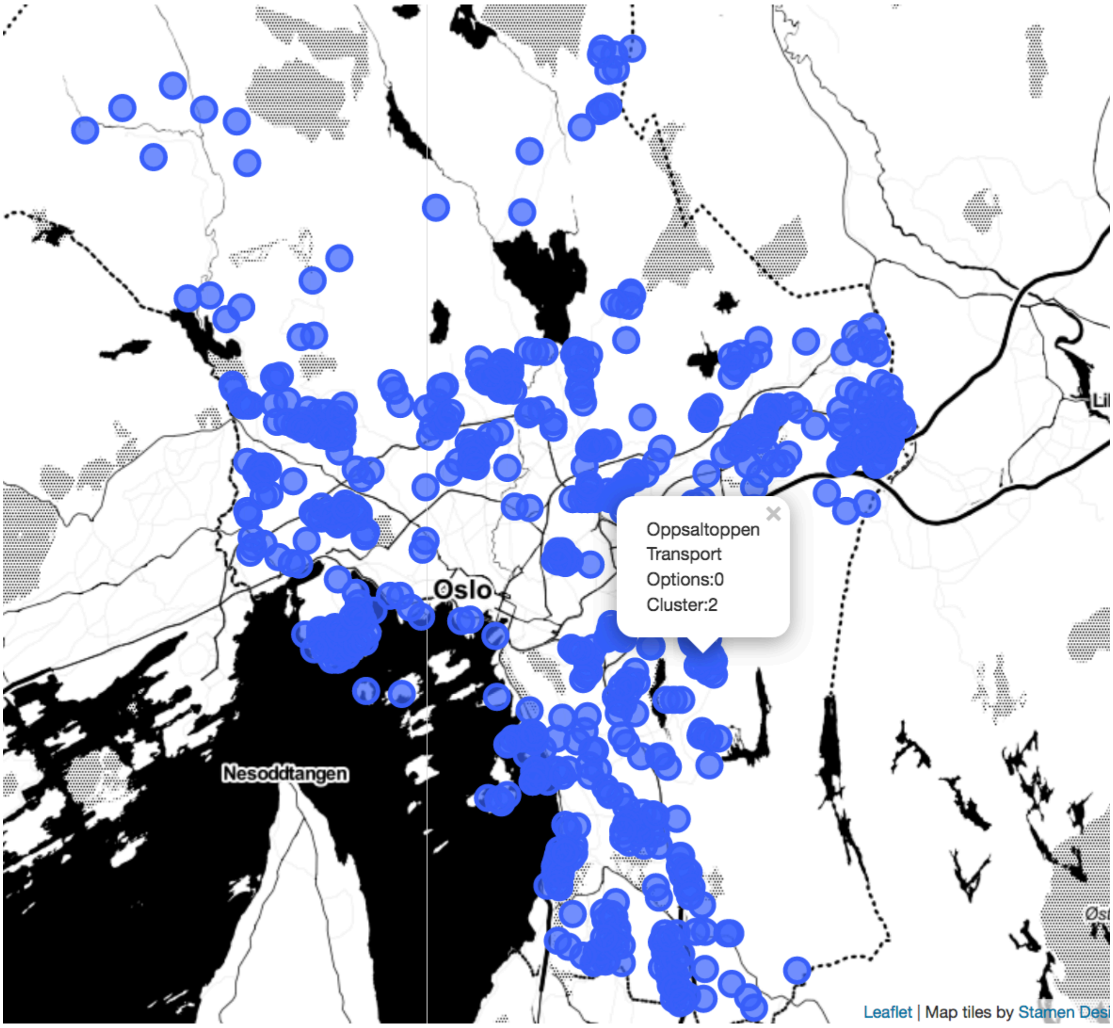
The various clusters, denoting the number of travel options and the total streets belonging to each cluster is depicted below:



5. Discussion

Vast parts of Oslo, measured down to the street level have very few transport options within 400m.

With the help of the map and the clusters, the Transport Department can now easily plan which areas of Oslo to emphasise for building more infrastructure. For example, if they wanted to prioritise areas that have 0 transport options within 400m, then the following map would be very useful:



A major caveat in this work is that mid points of streets were taken. Since we define nearest transport option as those within 400m radius, if the street itself is >800m in length, then the midpoint may not necessarily be a good location measurement. Furthermore, one would obtain different results if one's definition of short walking distance is more than 400m.

Additionally, one can further expand and elaborate this project by:-

- Obtaining population info (folketal) for each bydel (city district) / streets and making choropleth maps showing population density. However, choropleth maps need GeoJSON file containing boundary info for each bydel/city district, so such a source must be found.
- Mapping foot traffic info onto heatmaps, that can help truly emphasize areas with maximum passenger traffic
- Work on ferry data as well
- Use Google Maps API instead of Foursquare API

6. Conclusion

To conclude, we set out to map out areas of Oslo, down to the street level, that can be said to have good transport options. We defined good transport options as having a public transport facility within 400m of the mid-point of the streets in Oslo.

We find that there are several streets that are in need of transport upgrade and if the transport department chooses to focus on the streets with **less than 1 transport option** within 400m, it will have its hands full and will hopefully be a satisfied client :)

7. References

1. Erik Bolstad
2. BeautifulSoup4
3. Foursquare API
4. Folium