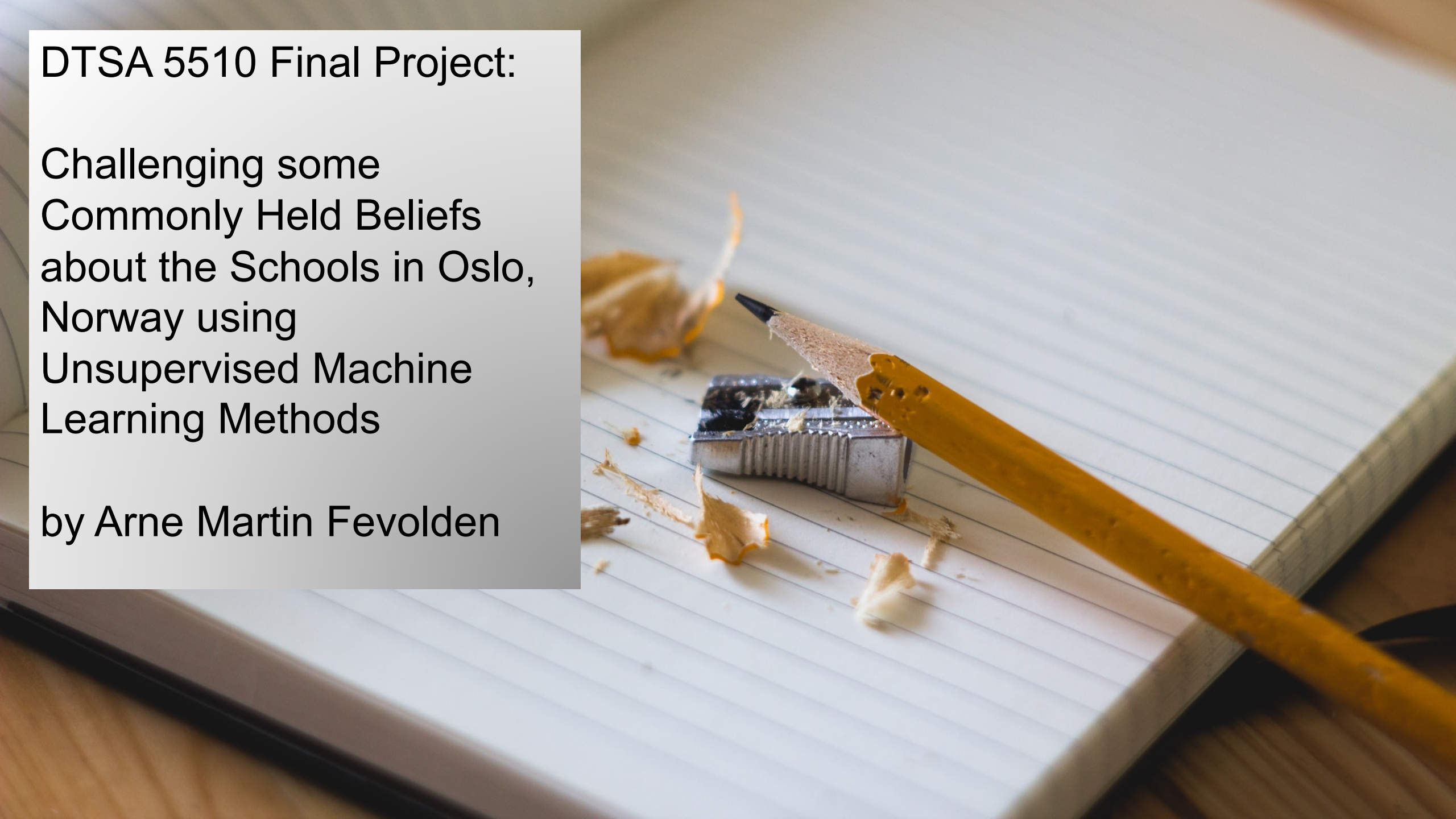
DTSA 5510 Final Project:

Challenging some Commonly Held Beliefs about the Schools in Oslo, Norway using Unsupervised Machine Learning Methods

by Arne Martin Fevolden



Overview

Testing two hypotheses about the school system in Oslo, Norway:

- Socio-economic hypothesis (rich areas have better schools)
- Immigration hypothesis (immigrant areas excel at math and sciences)

Data:

Average final assessment grade received by pupils attending secondary school at the 10^{th} grade level (according to the Norwegian system) in six different subjects.

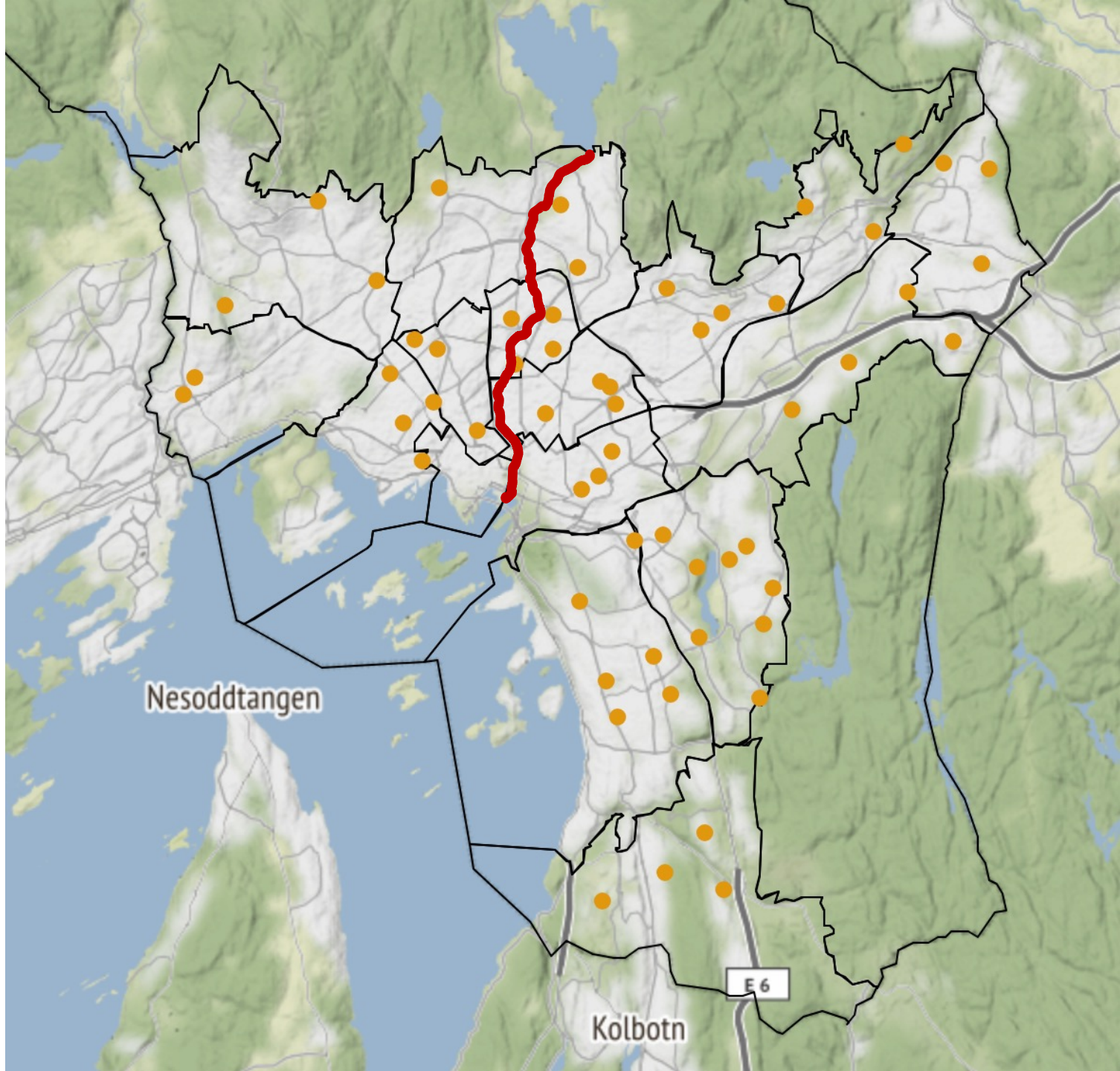
Methods:

- kMeans
- Hierarchical Clustering

Socio-economic hypothesis

Hypothesis I: Schools located at historically richer west side of Oslo provides a better education than schools on the east side.

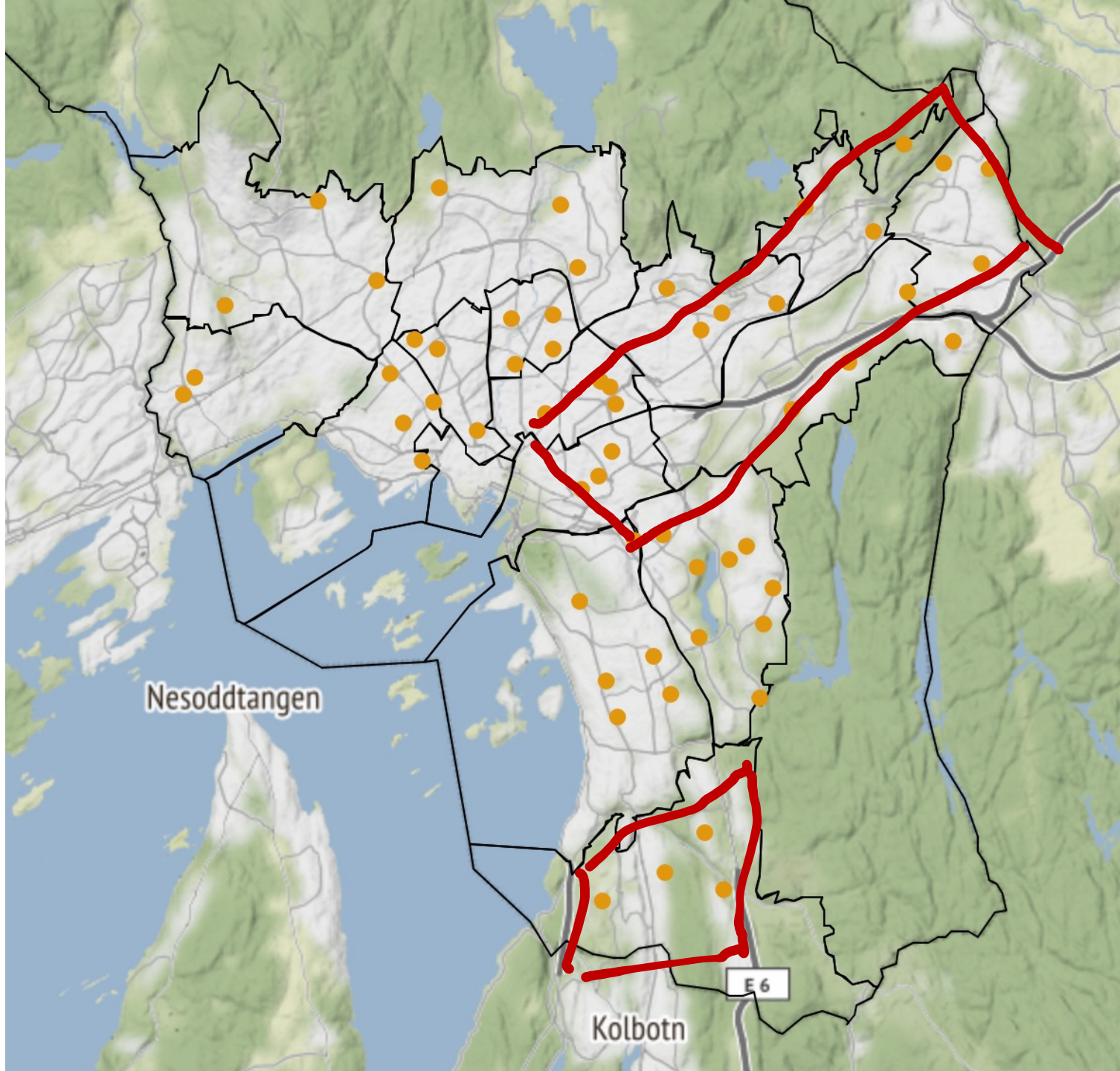
- The dividing line between the west and the east has traditionally followed the Akerselva river, which is marked with a red line on the map
- The part of the city to the left of the line is the traditionally rich side of Oslo.
- It is a strong indication that Hypothesis I is true if the cluster analysis place schools on the west and east into different clusters



Immigration hypothesis

Hypothesis II: The second belief is that schools located in more immigrant dense areas of Oslo provides a relatively better environment for learning Math, Natural Sciences and English than Norwegian.

- It is a strong indication that the immigration hypothesis is true if we find one cluster consisting of a rectangular shape stretching from the city center to the north-east and a circle in the far south.

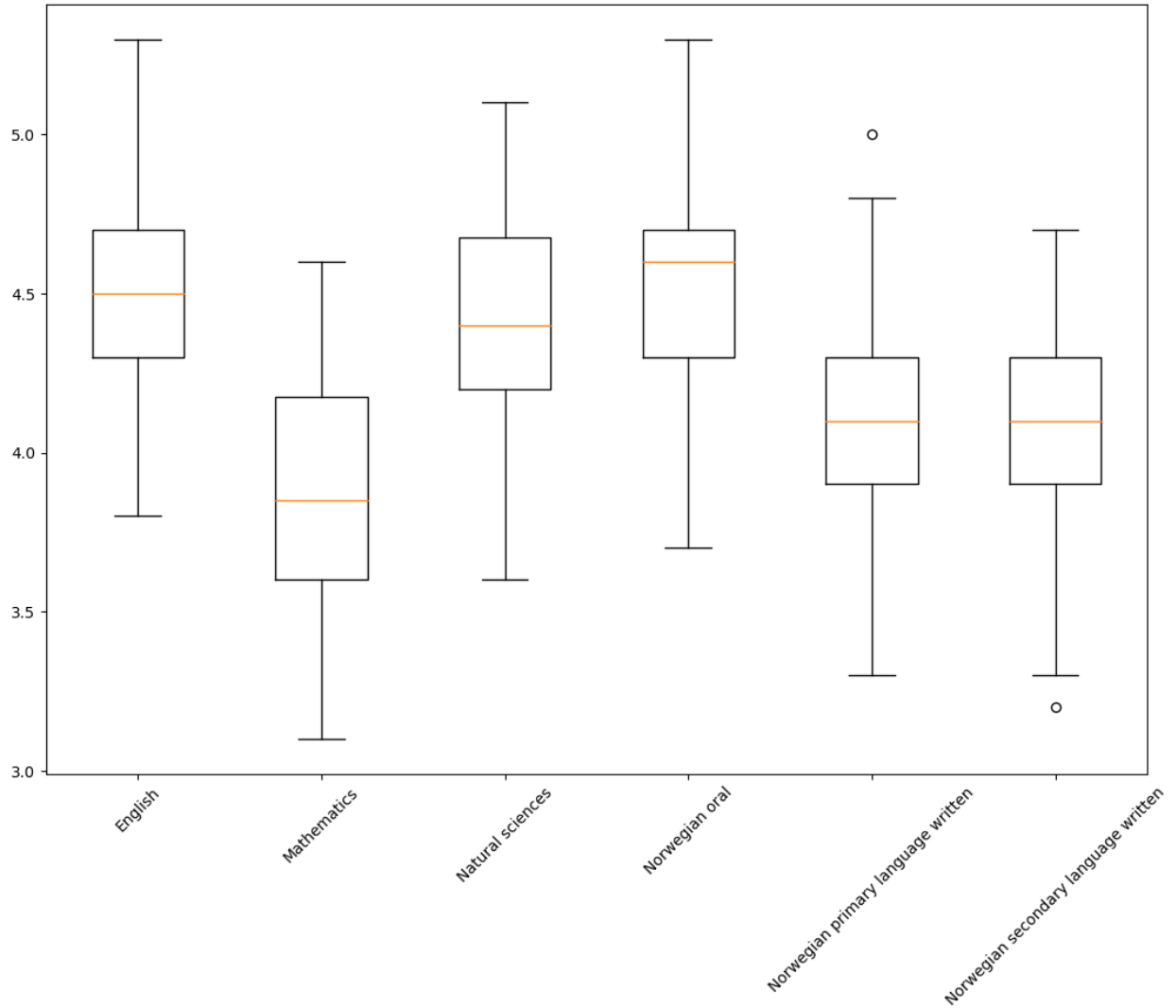


Data

- The dataset is made available by the Norwegian Directorate for Education and Training and contains information about the final assessment grade received by pupils attending secondary school at the 10th grade level (according to the Norwegian system).
- The final assessment grade is given as an average of the grades received by the pupils at a particular school following a particular subject.
- This dataset is a subset of a wider dataset that contains only information about pupils attending school in the capital of Norway, Oslo, during the school year 2021-22.
- The dataset contains information about six subjects: English, Mathematics, Natural Science, Norwegian oral, Norwegian primary language written and Norwegian secondary language written.
- The three Norwegian subjects test different aspects of Norwegian mastery: Norwegian oral tests the ability to explain linguistical concepts verbally, Norwegian primary language tests the students' ability to express themselves in writing using the Norwegian standard “Bokmål” and Norwegian secondary language tests the students' ability to express themselves using the written standard “Nynorsk” (at least for most pupils).

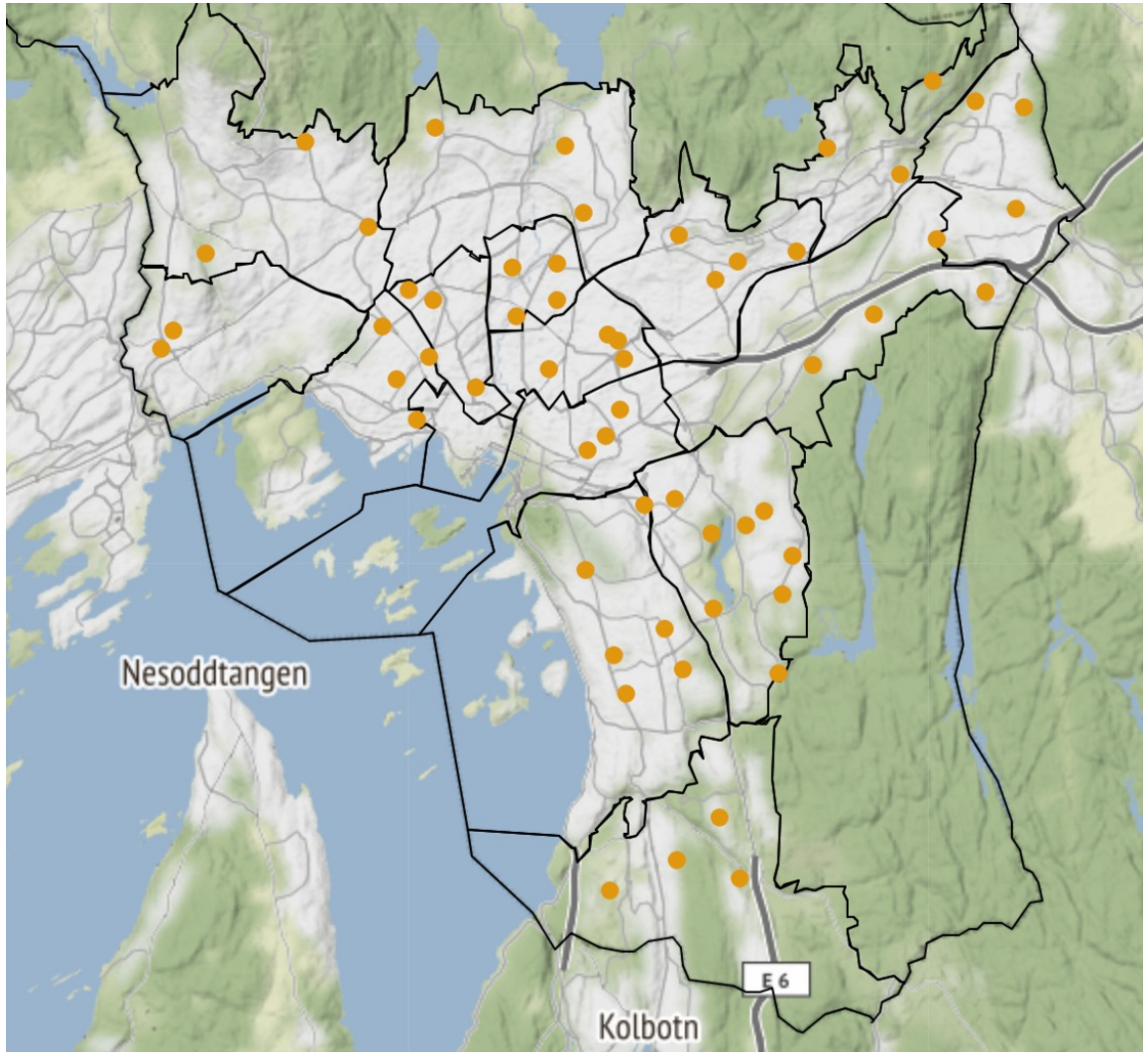
Exploratory Analysis I

- From the box plot we can see that the median is somewhere between 3.8 and 4.6 on a grade scale that goes between 1 and 6. This implies that there are not big differences between the subjects and that they are comparable. This is helpful when we carry out a cluster analysis.
- In addition, the box plot show that most of the school have grade averages that are fairly close to the median.
- We can also see that there are only a few outliers.



Exploratory Analysis II

- From the map we can see that all our schools are located within the city boundaries
- We can also see that the schools are evenly spaced across the whole city.
- Both these aspects indicate that the quality of our data is good.



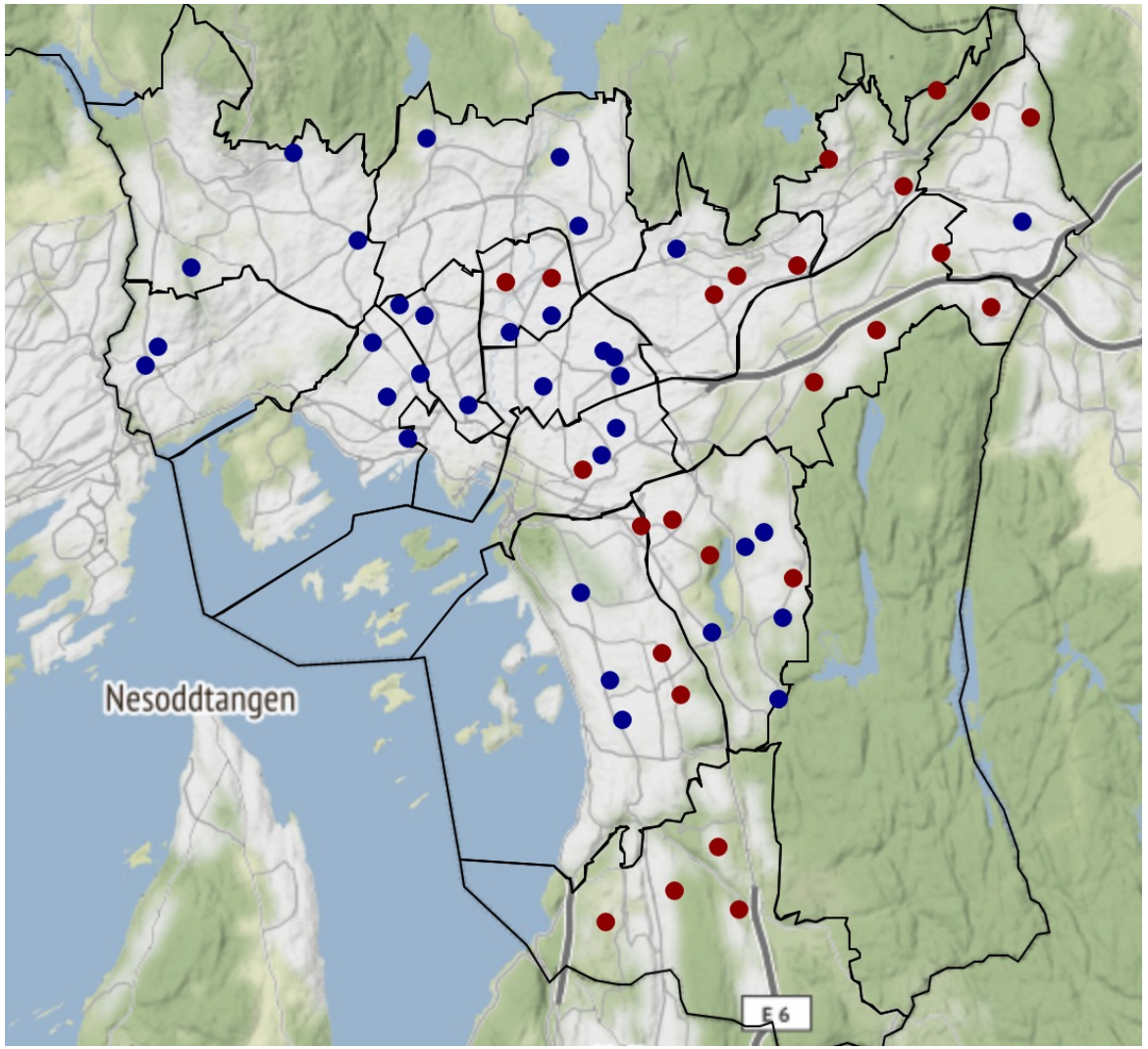
Unsupervised methods - Clustering

- We will try out two types of clustering algorithms – kMeans and Hierarchical clustering.
- For both of these algorithms, we will try out models with two and three clusters. Based on our theories, we would expect that there are two to three clusters: two clusters if only one of the hypotheses are true and three clusters if both hypotheses are true.
- The hypotheses are true if the clusters follow the geographical pattern discussed above: The socio-economic hypothesis is true if we find one cluster in the east and one in the west and the immigration hypothesis is true if we find one cluster consisting of an oval shape stretching from the city center to the north-east and a circle in the far south and another consisting of the remaining schools.
- Since there are no reason to expect more than three clusters, we will not try out four or more clusters. The cluster algorithms would surely find four or more clusters, but we would not be able to interpret them in any meaningful way.
- We employ both kMeans and Hierarchical clustering to ensure that our results are robust.

Results I:

- A tendency towards blue on the west-side and red on the east-side
- The blue cluster has a higher-grade centroid than red
- Weak support for the socio-economic hypothesis

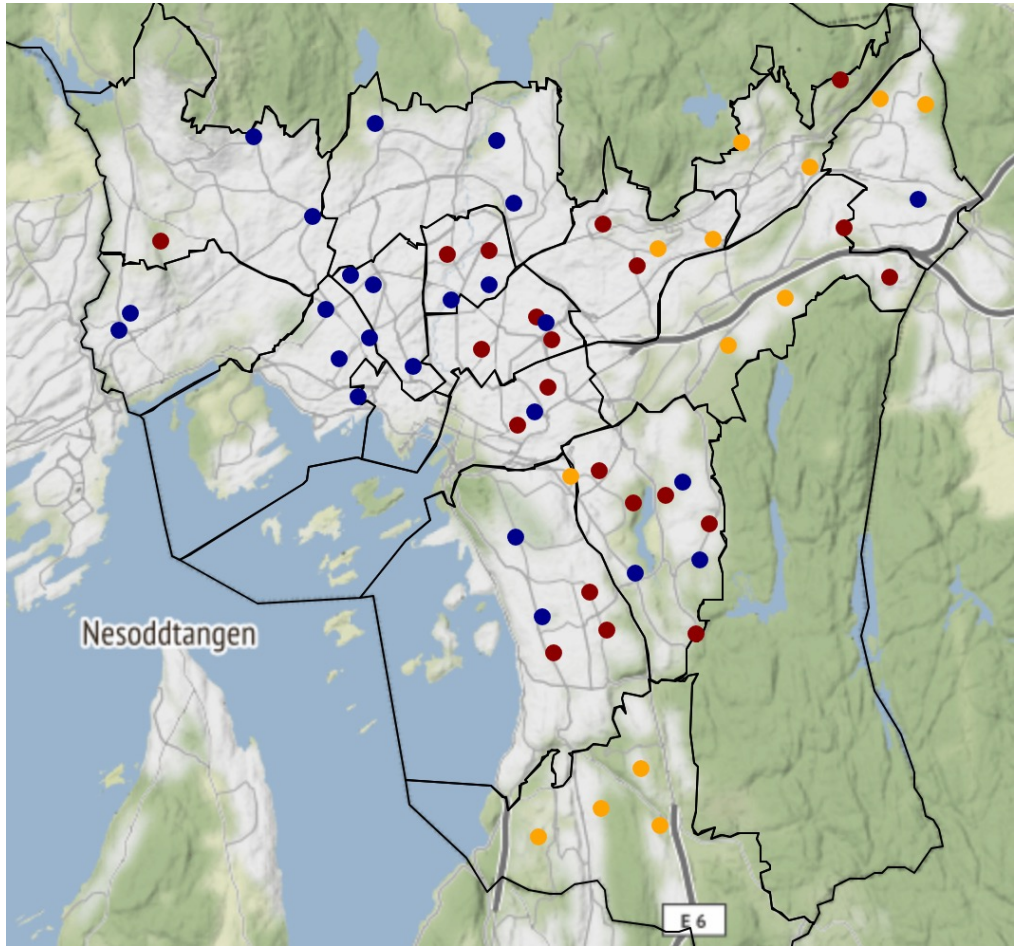
	Blue	Red
English	4.648485	4.268
Mathematics	4.124242	3.488
Natural sciences	4.593939	4.156
Norwegian oral	4.718182	4.252
Norwegian primary language written	4.309091	3.800
Norwegian secondary language written	4.236364	3.844



Results II:

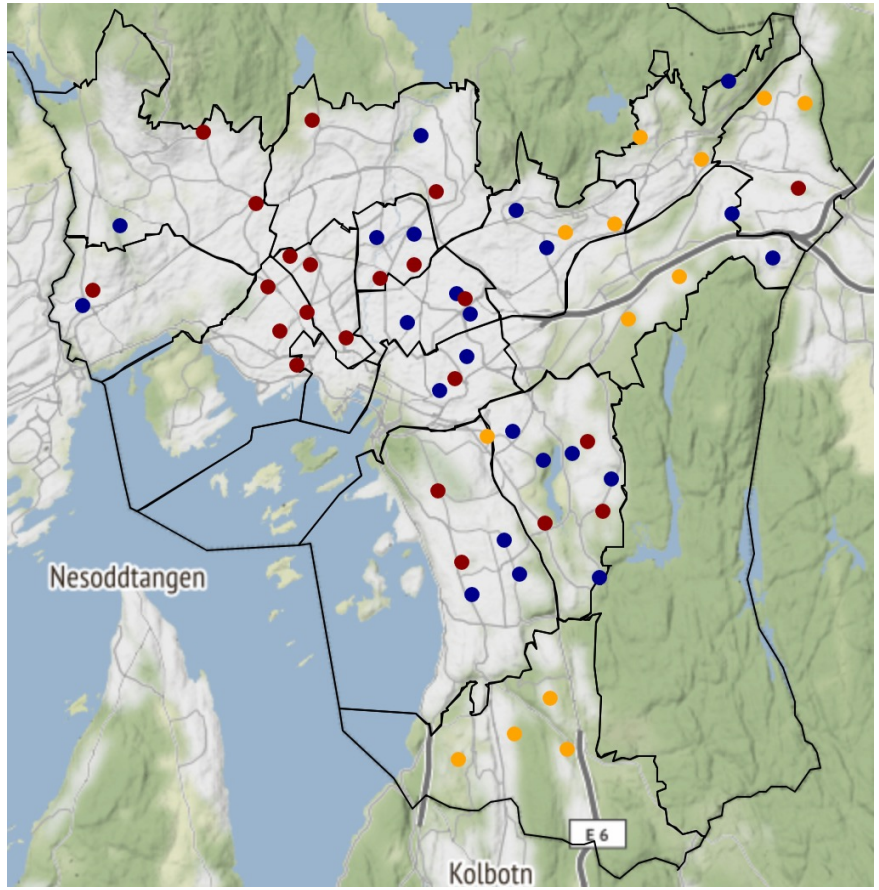
- Cluster result provide some support for both hypothesis I and II
- Centroid also partially confirms the two hypotheses
- However, there are many cluster labels that are inconsistent with either one or both hypotheses

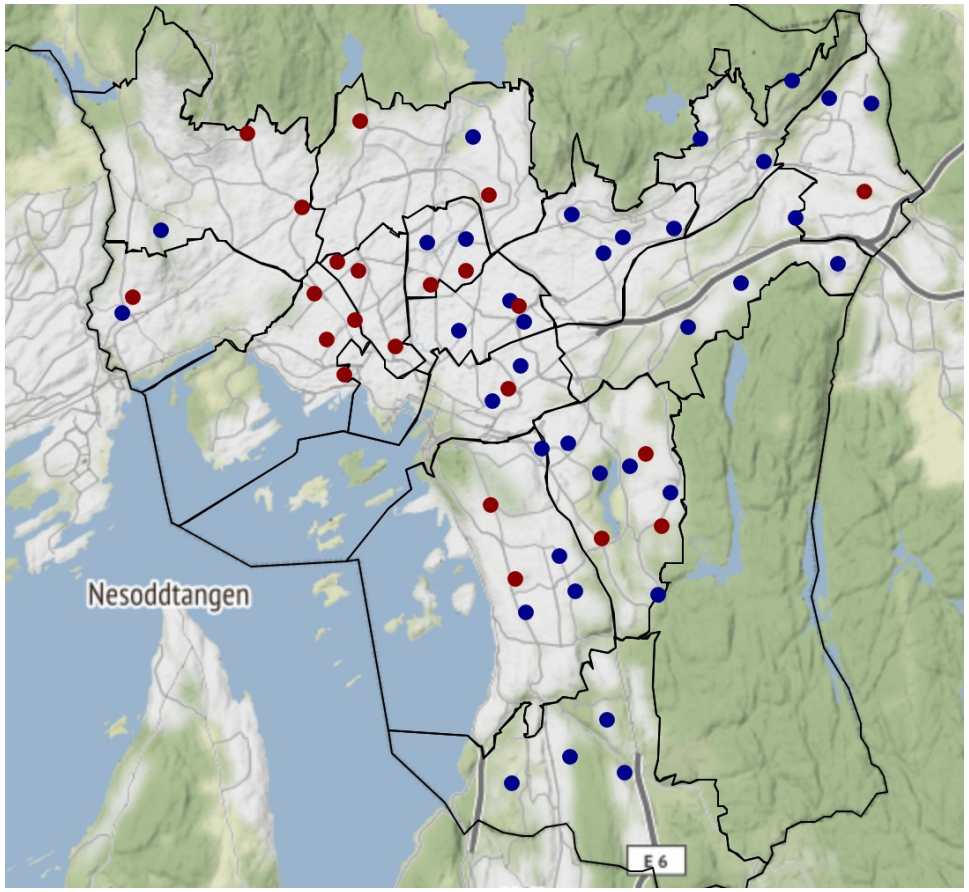
	Blue	Red	Orange
English	4.72	4.40	4.18
Mathematics	4.20	3.75	3.38
Natural sciences	4.66	4.34	4.03
Norwegian oral	4.78	4.50	4.05
Norwegian primary language written	4.41	4.02	3.62
Norwegian secondary language written	4.30	4.08	3.63



Results III:

Hierarchical clustering confirms the results from kMean clustering





Conclusion

- The analysis found that there are some support for the two hypotheses
- However, the reality is much more nuanced and complex, and the differences between the schools are in general quite small.