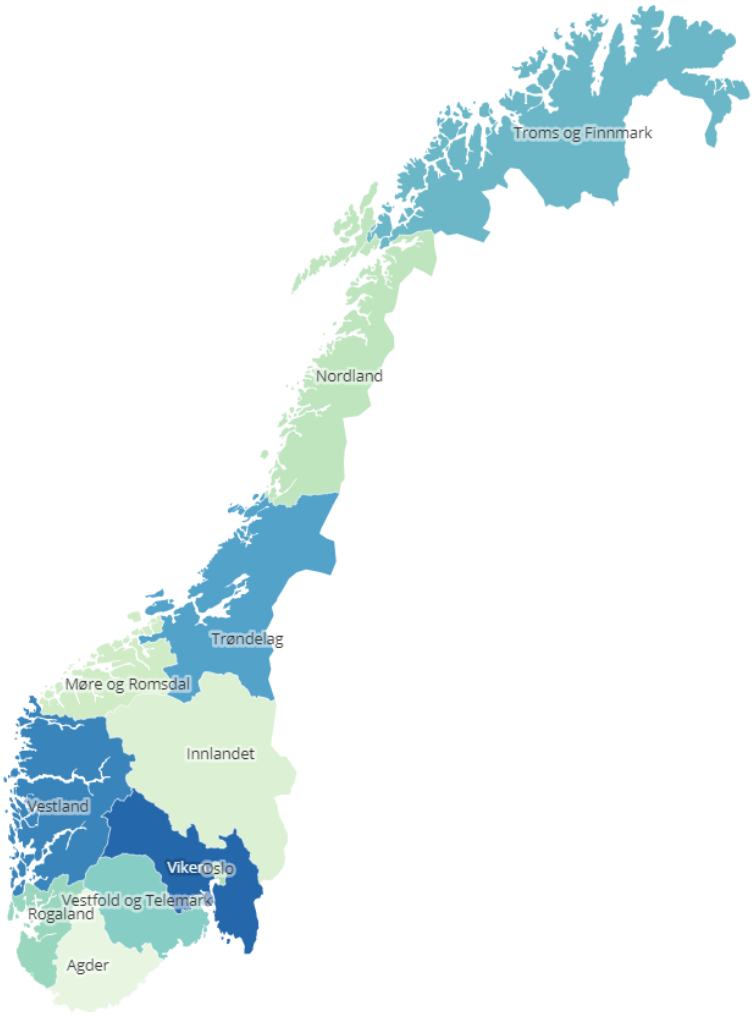


Kommunekampen

A STUDY OF THE NEW 2020 MUNICIPALITIES OF NORWAY

Sindre Misund Dahl | Applied Data Science Capstone | 26.04.2020

Introduction

Norway is a nordic country in northwestern Europe with a total area of 385,207 square kilometres and a population of approximately 5 368 000. Norway is divided into administrative regions, called counties and municipalities. The capital city Oslo is considered both a county and a municipality.

On January 1st 2020 the number of counties in Norway was redused from 19 to 11, while the number of municipalities was reduced from 428 to 354.

With the relatively small population of Norway, the amount of venues in the different municipalities varies a lot. In the municipalities with the largest cities in Norway, urban venues are expected to be present, e.g. cafes and restaurants. While in the smaller municipalities a larger difference in venues is expected to be seen.

In this project I would like to find and scrape the coordinates and population data of the municipalities of Norway. I would also like to use data from Foursquare to compare the municipalities. I will cluster the locations using the data. During the analysis I will consider whether all municipalities are to be studied or the total number of municipalities will be reduced.

Data

The Norwegian statistics bureau SSB have population data for each municipality in Norway (https://www.ssb.no/statbank/table/11342/).

The python package Geopy with Nominatim will be used to find the location data for the municipalities.

Finally, Foursquare will be used to find the top venues of each location.

Methodology

The name and population data was downloaded from the Norwegian statistics bureau in a .csv format. The .csv-file was used to create a dataframe containing all 354 municipalities of Norway. The data in the dataframe was cleaned and the 200 largest municipalities based on population was chosen for further study.

The location data for each municipality was found using the geopy package in python with Nominatim as geolocator. In order to end up with the location of the centre of the largest village/town/city in the municipality, a search query of bus stop together with the municipality names was used. This search worked for all municipalities, except for 7 small ones. The remaining 193 municipalities was used for the rest of the study.

With the locations of each municipality, Foursquare was used to scrape venue data in each area. First, the largest municipality based on population was studied. The 100 most popular venues in the municipality was fetched of the city center with a radius of 2000 meters. A dataframe was created from the data and the data was studied.

The process of fetching venue data was repeated for all municipalities. The data was stored in a dataframe. The data from all municipalities was studied before it was used to create clusters.

The venues of each municipality were used to create clusters of all municipalities. The K-Means method from the python package sklearn was used. Only the venue data from Foursquare was used when creating the clusters. The number of clusters was set to 11, which equals to the number of counties in Norway. Finally, the clusters were studied.

Results

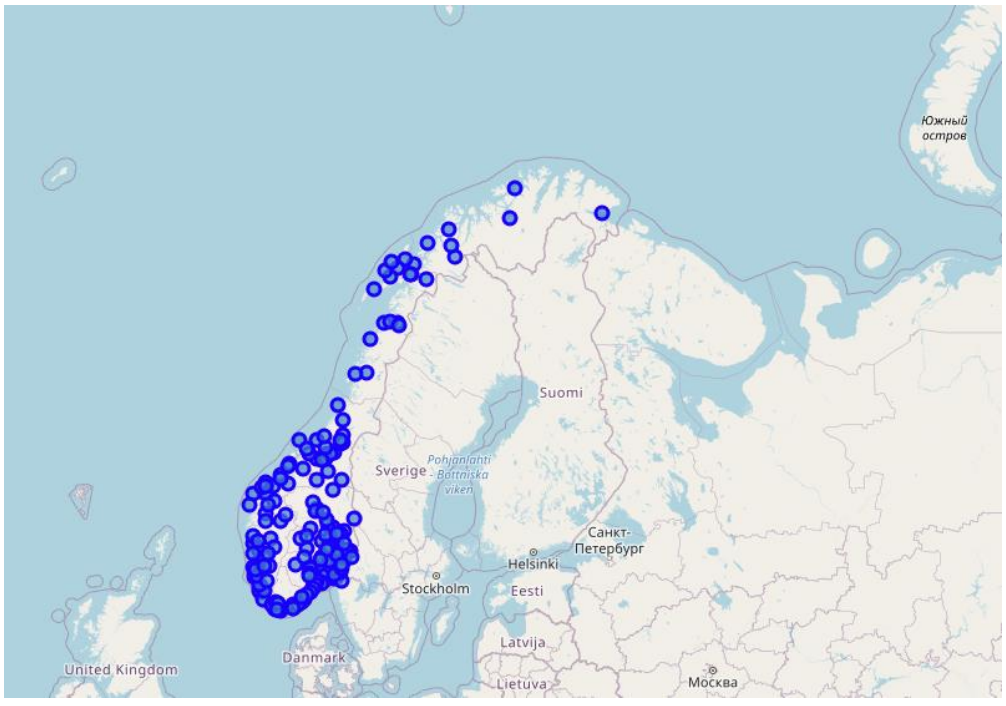
The head of the dataframe created from the SSB-data are shown below.

	region	Befolkning per 1.1. (personer) 2020	Areal (km2) 2020	Landareal (km2) 2020	Innbyggere per km2 landareal 2020
0	3001 Halden	31373	642	595	53
1	3002 Moss	49273	138	128	385
2	3003 Sarpsborg	56732	406	370	153
3	3004 Fredrikstad	82385	293	284	290
4	3005 Drammen	101386	318	305	332

The cleaned data sorted by largest population are as follows:

	Region	Befolkning	Areal	Landareal	Innbyggere per km2 landareal
51	Oslo	693494	454	426	1628
169	Bergen	283929	465	445	638
238	Trondheim	205163	529	496	414
147	Stavanger	143574	263	257	559
20	Bærum	127731	192	189	676

The locations of the 193 largest municipalities were plotted in the python folium package, as shown below:



The location data for the following 7 municipalities were not found and were not used for the rest of the analysis:

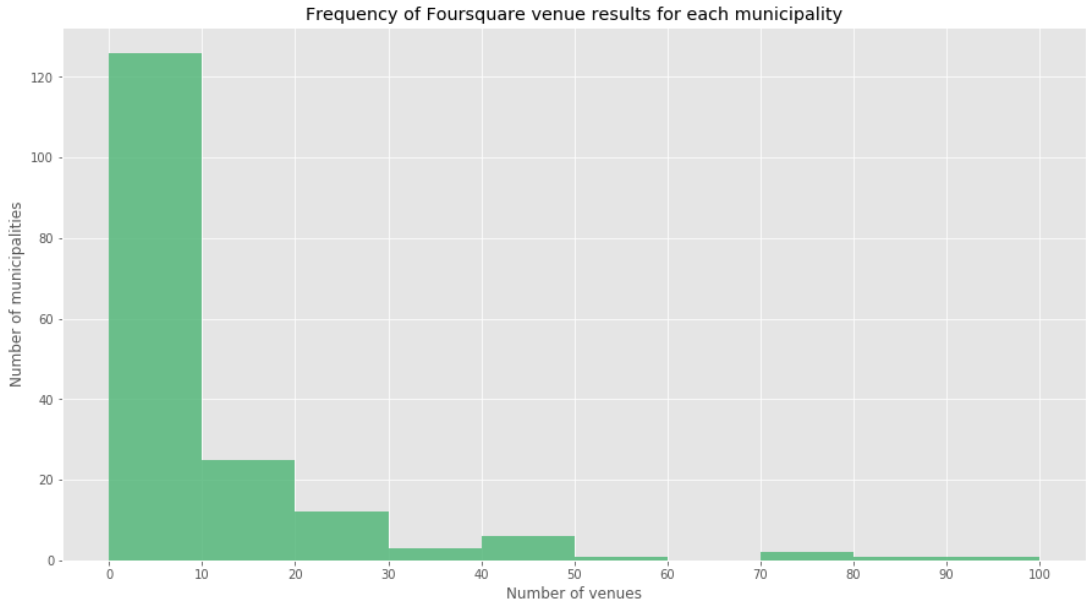
Municipality missing location
Færder
Hustadvika
Sør-Varanger
Herøy (Møre og Romsdal)
Nordreisa
Hvaler
Grue

The largest municipality based on population was found to be Oslo, the capital of Norway, with a population of 693494.

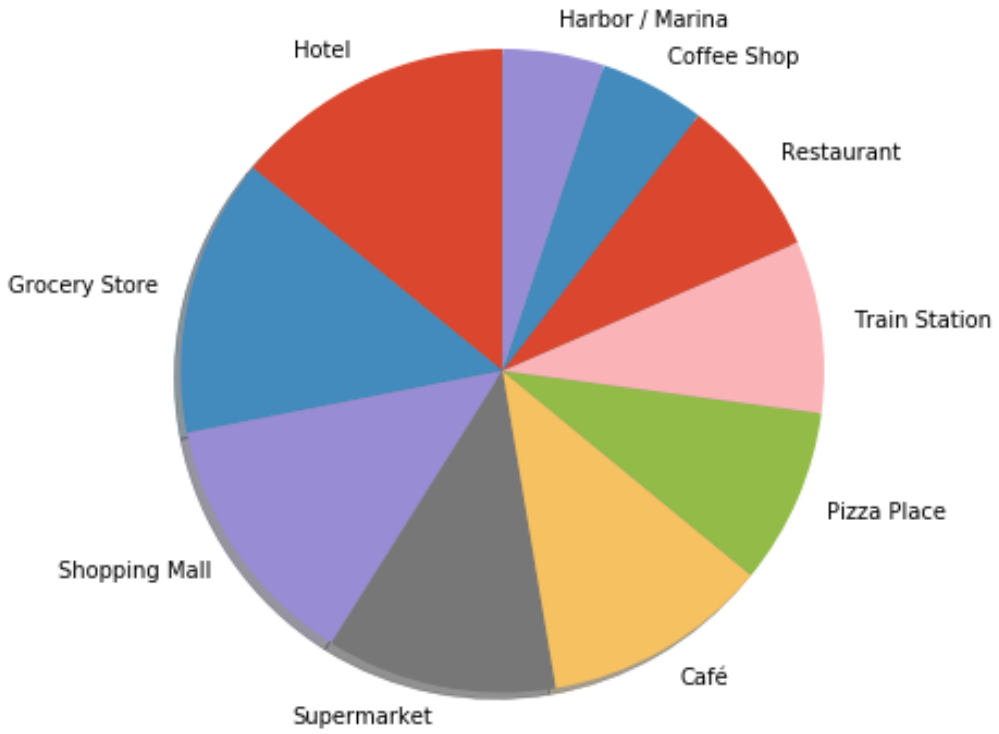
The top 100 Foursquare venues for Oslo was found to be:

	name	categories	lat	lng
0	Ben & Jerry's	Ice Cream Shop	59.914360	10.737109
1	Stockfleths	Coffee Shop	59.913656	10.741206
2	Det Norske Teatret	Theater	59.915360	10.738657
3	Nordvegan	Vegetarian / Vegan Restaurant	59.915591	10.737863
4	Der Peppern Gror	Indian Restaurant	59.912403	10.735058

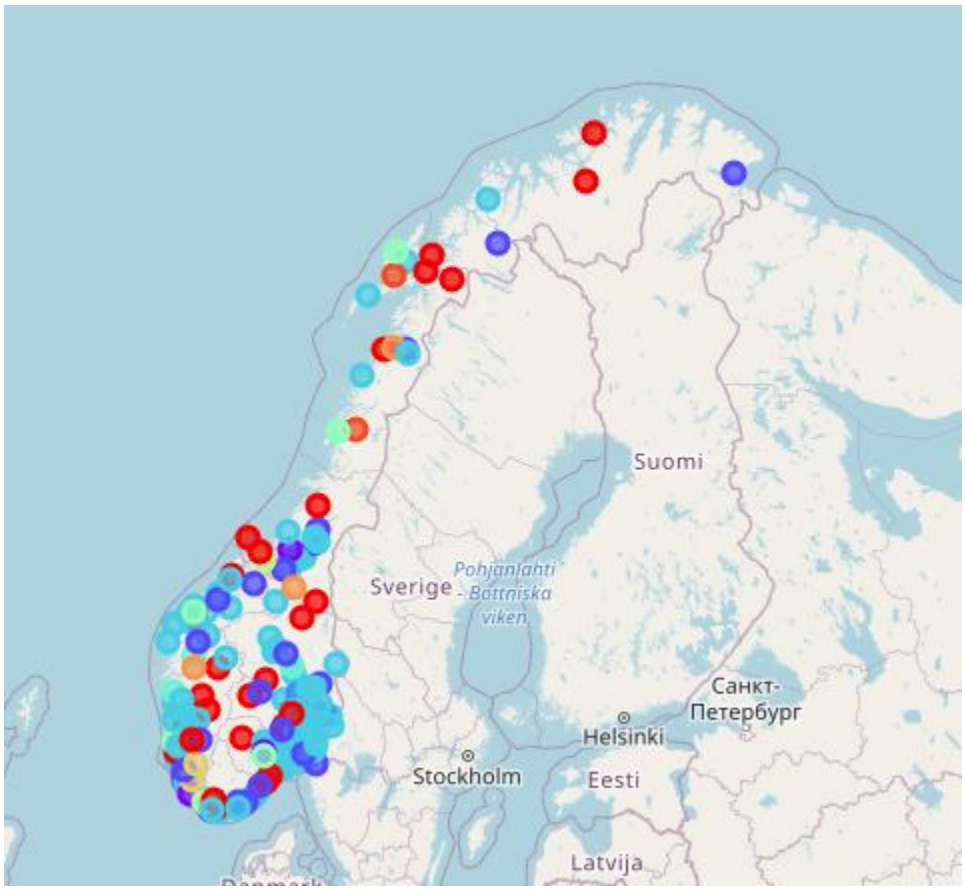
Repeating the process for the 193 largest municipalities of Norway gave a dataframe with 1899 entries from 176 municipalities. The remaining 17 municipalities gave no venue results in Foursquare. This averages to 9.8 venues for each municipality. The number of municipalities giving the maximum 100 venues results was found to be only Oslo. The plot on the next page shows a histogram of the frequency of Foursquare venues for each municipality. The horizontal axis is the amount of venues for each municipality, while the vertical axis is the number of municipalities that falls into each bin. As seen from the plot a large number of municipalities have between 0 to 10 venue results from Foursquare.



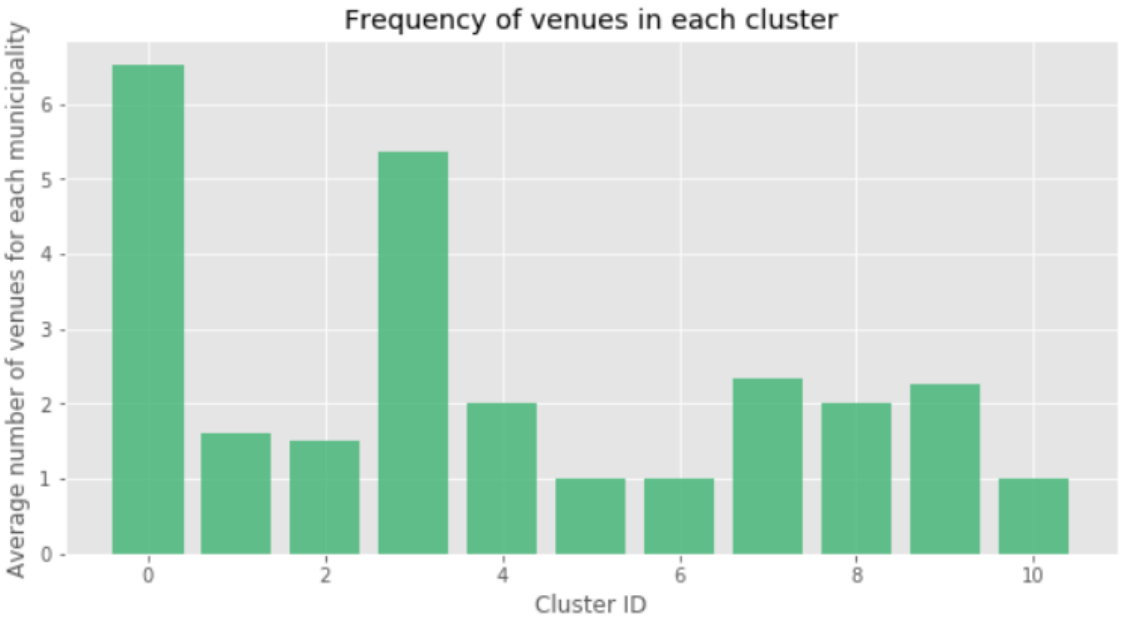
There are 216 unique venues in total. The most common venue in Norway is Hotel, with Grocery store and Shopping mall following. The pie chart below shows the distribution among the 10 most occurring venues in Norway.



A folium plot of the clustering is shown in the figure below.



The average amount of venues for each cluster are shown in the bar chart below.



Discussion

As seen in the results, several municipalities were to small to have any venue results from Foursquare. In addition, the average number of venues for each municipality was 9.8, with the largest number of municipalities having between 0-10 venues each. The number of municipalities giving the maximum 100 venues results was found to be only 1.

As a result, it was expected that clustering of the venues data was highly affected by the large number of municipalities with a small amount of venues each. This was confirmed by the several clusters sharing few of municipalities, with clustering data showing that the municipalities of most clusters had less than 3 venues in average.

All the municipalities in each cluster was found to have similarities in the data, where especially the top venue was typically shared for each municipality in a cluster.

A description of the trend of each cluster is given in the table below.

Cluster number	Number of municipalitites	Characteristics
1	113	Hotel as top venue (sometimes second)
2	5	Campground as top venue, few other venues
3	2	Grocery store as top venue
4	35	Train station as top venue, no other venues
5	3	Café, coffee shop and restaurants as top venues
6	1	Campground in top 3 venues, few other venues
7	3	Boat or ferry as top venue, few other venues
8	6	Harbor / Marina as top venue, no other venues
9	3	Lake in top 2 venues, few other venues
10	4	Mountain as top venue, few others
11	2	Bed & Breakfast as top venue, no other venues

Conclusion

As seen in this study, with a large variation of the number of data, where most entries have few results, will highly influence the clustering. Most clusters contained few municipalities with few venues, where the top venue was shared between the municipalities in the cluster. Using machine learning it is important to have good data and in addition to have enough data to work with.