2020

Tastes of Norway

DATA SCIENCE CAPSTONE PROJECT

KRISTOFFER JAN ZIEBA

Contents

Introduction

Although Norway is a country of relatively small population (~ 5 million) it has a well-developed culinary scene. The scene is however very different through the entire country – from very sophisticated and posh Michelin-starred restaurants in the biggest cities to simple food places in smaller towns. Each year many restaurants get opened but also many get closed due to mismatch between the restaurant type and specific culinary tastes in the Norwegian towns. I aim to close the gap between the tastes and types of newly opened restaurants – this work will show **which food venue type has the best chances to succeed in the biggest 20 Norwegian cities and towns.** I hope this work will be a directly useful for these who want to establish a new venue and for municipalities who grant restaurant permits. In addition, the study will be interesting for broader public to understand the culinary preferences of different regions and infer what may it say about the culture. In addition, the study may serve as a guideline for travelers to set right culinary expectations.

Data

Data composes of 3 sources:

- Data from Statistics Norway about population (http://www.ssb.no)
- Data from Foursquare API (https://developer.foursquare.com/docs). Foursquare API, a location data provider, will be used to make RESTful API calls to retrieve data about venues in different neighborhoods. This is the link to Foursquare Venue Category Hierarchy. Venues retrieved from all the neighborhoods are categorized broadly into "Arts & Entertainment", "College & University", "Event", "Food", "Nightlife Spot", "Outdoors & Recreation", etc. The data was used to get food venues around centers of towns.
- Data from Google Geocoding API (https://developers.google.com/maps/documentation/geocoding/start). Geocoding is the process of converting addresses (like a street address) into geographic coordinates (like

latitude and longitude), which you can use to place markers on a map or position the map. The data was used to get geographical coordinates of the towns' centers.

Methodology

The workflow

The workflow for the entire study is described in Figure 1. It consists of 1) acquiring population data and geographical coordinates about Norwegian towns and food venues in these towns, 2) initial analysis of food venue frequencies, 3) clustering of the towns according to venue category importance, 4) a recommendation system.

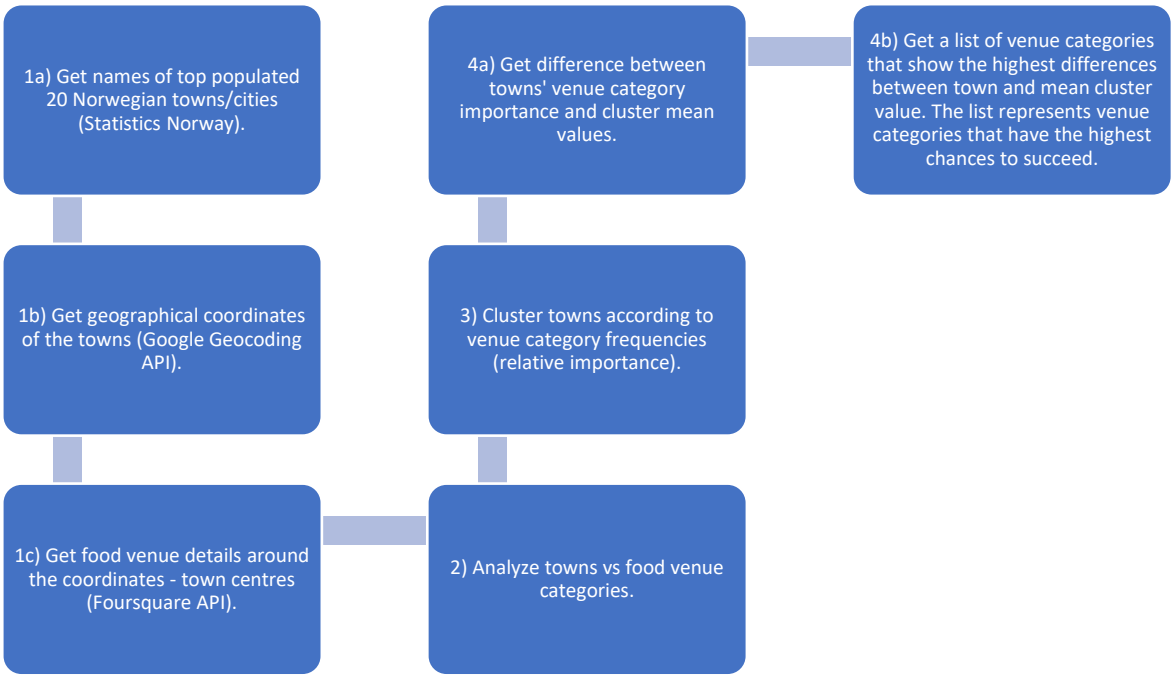

Figure 1. A general workflow of the study.

Data acquisition approach

Table 1 shows details all top 20 populated Norwegian towns obtained from both Statistics Norway and Google Geocoding API. Please note that some of the towns have double names and represent double towns (conurbation). These are however often located apart of each other and require additional geographical location. For location of the towns on map of Norway see Figure 5.

Table 1. 20 most populated towns and cities in Norway including additional information used in the study.

	Town	Population	Area_kmsq	Population_p_kmsq	lat1	lon1	lat2	lon2
0	Oslo	1019513	270.68	3766	59.913869	10.752245	NaN	NaN
1	Bergen	257087	87.34	2944	60.391263	5.322054	NaN	NaN
2	Stavanger/Sandnes	225020	79.31	2837	58.969976	5.733107	58.853258	5.732946
3	Trondheim	186364	58.21	3202	63.430515	10.395053	NaN	NaN
4	Fredrikstad/Sarpsborg	113622	58.08	1956	59.220537	10.934701	59.284073	11.109403
5	Drammen	107930	46.94	2299	59.744074	10.204457	NaN	NaN
6	Porsgrunn/Skien	93255	53.12	1756	59.138557	9.655515	59.208913	9.605753
7	Kristiansand	64057	24.89	2574	58.159912	8.018206	NaN	NaN
8	Ålesund	53234	28.36	1877	62.472228	6.149482	NaN	NaN
9	Tønsberg	52419	26.46	1981	59.267570	10.407561	NaN	NaN
10	Moss	47135	21.51	2191	59.434091	10.658383	NaN	NaN
11	Haugesund	45040	20.74	2172	59.413581	5.267987	NaN	NaN
12	Sandefjord	44368	23.89	1857	59.131310	10.216595	NaN	NaN
13	Arendal	43515	31.15	1397	58.461757	8.772450	NaN	NaN
14	Bodø	41720	14.17	2944	67.280356	14.404916	NaN	NaN
15	Tromsø	40471	13.45	3009	69.649205	18.955324	NaN	NaN
16	Hamar	27947	13.89	2012	60.794533	11.067998	NaN	NaN
17	Halden	25708	14.13	1819	59.132996	11.387457	NaN	NaN
18	Larvik	24647	13.62	1810	59.053836	10.029546	NaN	NaN
19	Kongsberg	22219	13.64	1629	59.668878	9.650189	NaN	NaN

Table 2 shows first and last rows of the all the venue data downloaded from Foursquare API for all the towns analyzed in this study. Despite Foursquare provides details about different venue categories such as parks, museums, in this study only food venues were considered.

Table 2. Part of the dataset of the food venue data acquired from Foursquare API.

	Town	Town Latitude	Town Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Oslo	59.913869	10.752245	Oslo Street Food	59.915819	10.750810	Food Court
1	Oslo	59.913869	10.752245	Arakataka	59.916381	10.750594	Scandinavian Restaurant
2	Oslo	59.913869	10.752245	Lulu	59.914837	10.750802	Japanese Restaurant
3	Oslo	59.913869	10.752245	Le Benjamin Bar & Bistro	59.918648	10.757985	French Restaurant
4	Oslo	59.913869	10.752245	Koie	59.917087	10.752708	Ramen Restaurant
...	...	...	...	...	...	...	...
651	Fredrikstad/Sarpsborg	59.284073	11.109403	Grålum Grill	59.294693	11.064784	Burger Joint
652	Porsgrunn/Skien	59.208913	9.605753	Cafe Tullis (Lykke)	59.205539	9.611729	Café
653	Porsgrunn/Skien	59.208913	9.605753	Burger King - Myren	59.201901	9.588023	Burger Joint
654	Porsgrunn/Skien	59.208913	9.605753	Stockmand	59.208249	9.608274	Café
655	Porsgrunn/Skien	59.208913	9.605753	Peppes Pizza	59.196610	9.600939	Pizza Place

Initially the dataset comprised 52 unique venue categories. The least common categories (below 3 venues in a category) were grouped into one larger category. In addition, some of the categories were merged together (e.g. BBQ and steakhouse).

Clustering approach

K-means clustering method was used to group the towns according to the venue category importance. The importance is a ratio between number of venues in a certain category to total number of venues in a town and it varies between 0 (the lowest) and 1 (the highest). This particular clustering method has been chosen because of relative implementation ease.

"K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, Better Euclidean solutions can be found using k-medians and k-medoids." Source: https://en.wikipedia.org/wiki/K-means_clustering

One of the most important parameters for k-means clustering is selection of cluster number. This can be set subjectively, or the number might be chosen by using one of optimization methods. The most popular ones are the Elbow and Silhouette methods/scores. In the Elbow method the optimum k (cluster number) is the one where Sum of Squared Distances first starts to diminish. In the plot of SSD-versus-k, this is visible as an elbow. For the Silhouette Method, its score reaches its global maximum at the optimal k. This should ideally appear as a peak in the Silhouette Value-versus-k plot. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters.

In this study both optimization methods were used (see Figure 2). Based on them number of clusters = 4 was set.

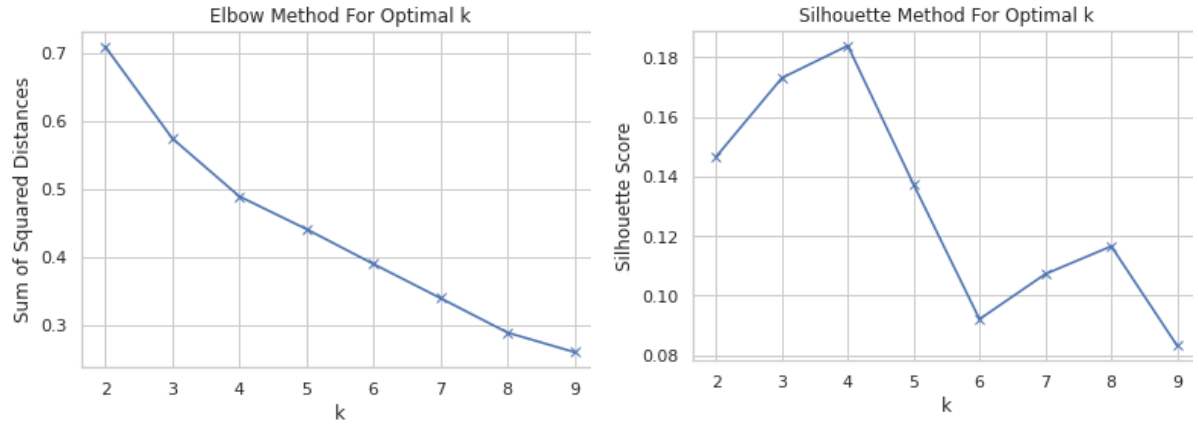

Figure 2. Optimization of cluster number. Number of clusters vs Elbow and Silhouette scores.

The recommender system approach

The goal of the system is to provide information what type of new restaurant has the best chances to succeed in the towns. The calculations were performed by comparing number of venues in a town and mean value for a relevant cluster. If a town has less venues in a certain category then a cluster mean value, then most likely a new venue in this category can succeed. The subtraction gives a certain number that represents a relative 'need' for a certain food venue category. It might be interpreted that the higher the value, the higher the 'need'.

Results

Top venue categories

The top 10 most popular food venue categories in the top 20 populated Norwegian towns are shown in Figure 3. Sizes of the squares represent importance of the categories. Café is the most popular type, followed by classic restaurant, fast food and pizza place.

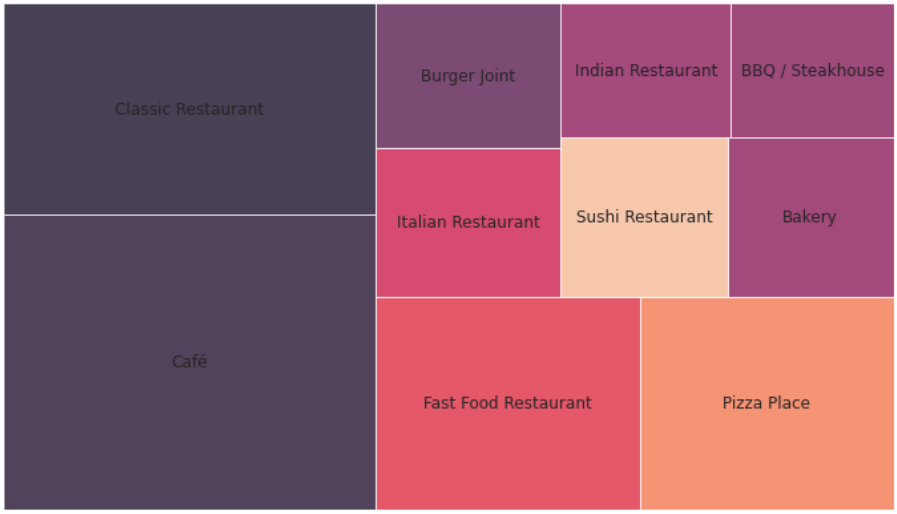

Figure 3. The most common food venue types in top 20 populated Norwegian towns.

Figure 4 shows importance of top 10 food venue categories in all the towns. The importance is a ratio between number of venues in a certain category to total number of venues in a town and it varies between 0 (the lowest) and 1 (the highest). As shown, some of the venue types such as café are common in (almost) all towns, while some categories, such as sushi, cannot be found everywhere.

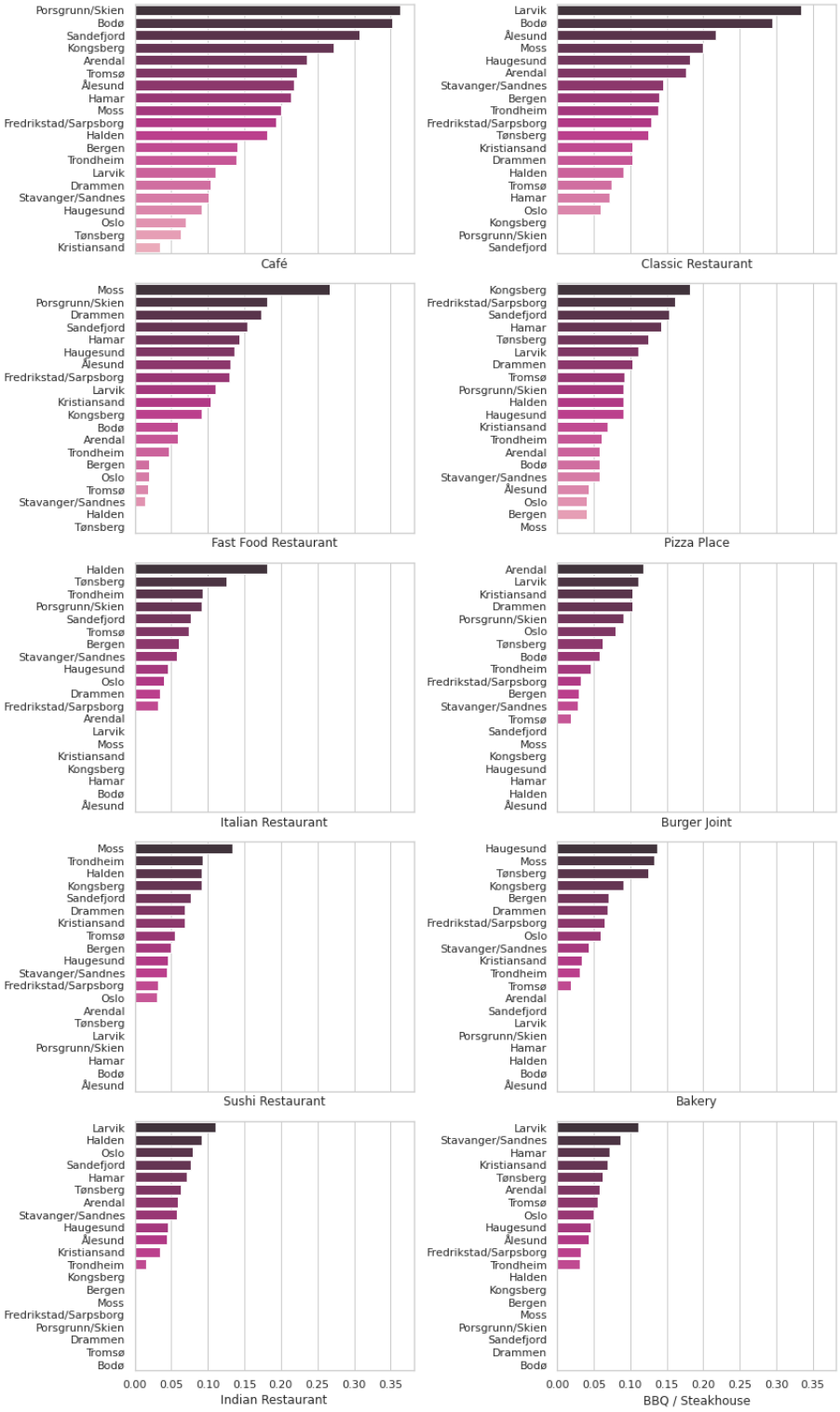

Figure 4. Relative importance of top 10 food venue types in Norwegian towns. The importance is a ratio between number of venues in a certain category to total number of venues in a town and it varies between 0 (the lowest) and 1 (the highest).

Clusters

The towns were grouped into 4 clusters: cluster 0, 1, 2 and 3 according to food venue similarity and their relative importance in the towns. An overview of cluster characteristics is shown in Figure 5. Cluster 0 has large share of cafés and classic restaurants and relatively low share of other types.

Cluster 1 is the best-balanced cluster according to variety of venue types. Cluster 2 includes mostly cafés, pizza places and fast food restaurants while cluster 3 cafés, fast food bakery and sushi places.

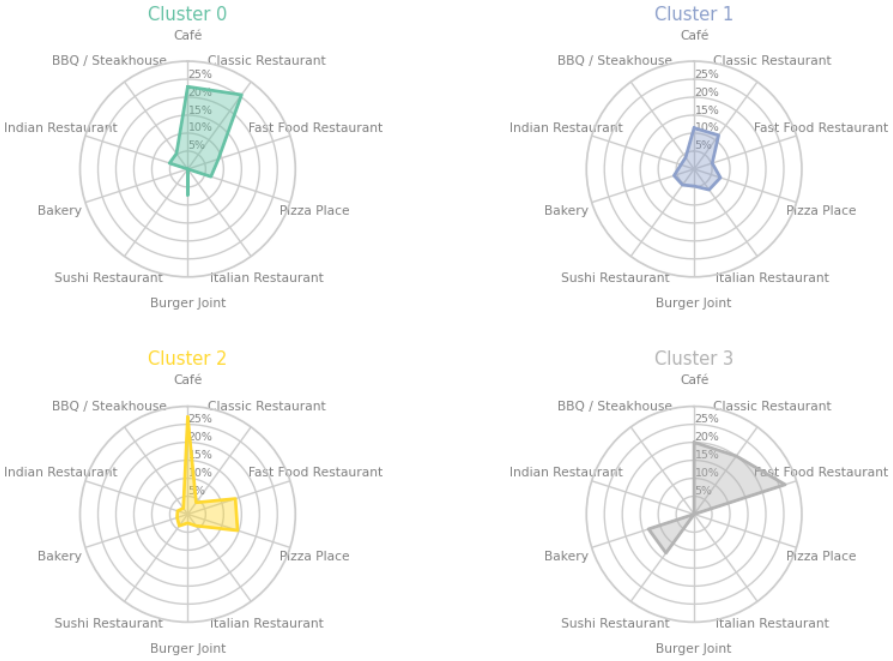

Table 3 shows names of towns and 5 most common venue types in each cluster.

Table 3. Resulting town clusters according to food venue similarity and their relative importance in the towns. 5 most common venue types in each cluster are also shown in the table.

Cluster: 0 | Towns: Arendal, Bodø, Larvik, Ålesund
The most popular venue categories (fraction):
Classic Restaurant 0.255328
Café 0.229184
Fast Food Restaurant 0.089798
Burger Joint 0.071895
Pizza Place 0.068059
Cluster: 1 | Towns: Bergen, Drammen, Halden, Haugesund, Kristiansand, Oslo, Stavanger/Sandnes, Tromsø, Trondheim, Tønsberg
The most popular venue categories (fraction):
Classic Restaurant 0.116209
Café 0.114529
Pizza Place 0.077133
Italian Restaurant 0.071111
Bakery 0.058758
Cluster: 2 | Towns: Fredrikstad/Sarpsborg, Hamar, Kongsberg, Porsgrunn/Skien, Sandefjord
The most popular venue categories (fraction):
Café 0.270378
Pizza Place 0.146144
Fast Food Restaurant 0.139693
Gastropub 0.051822
Classic Restaurant 0.040092
Cluster: 3 | Towns: Moss
The most popular venue categories (fraction):

Fast Food Restaurant	0.266667
Café	0.200000
Classic Restaurant	0.200000
Bakery	0.133333
Sushi Restaurant	0.133333

Figure 5 shows location of the towns including their cluster labels. Please note that cluster 1 incorporates the largest 4 Norwegian cities: Oslo, Bergen, Stavanger/Sandnes and Trondheim. Apart from this relationship no clear relationship between population and cluster label was found.

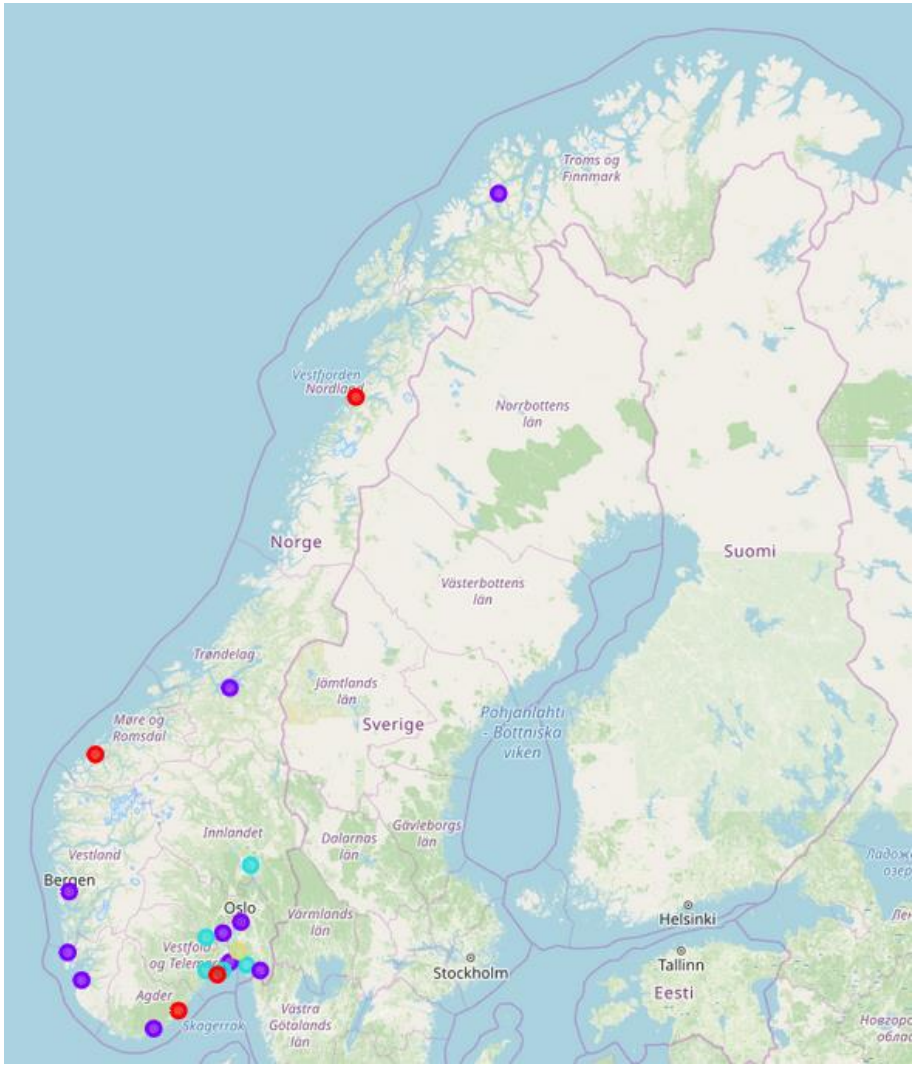

Figure 5. Location of the analyzed towns including their cluster labels (colors). Cluster 0: red, cluster 1: purple, cluster 2: blue, cluster 3: yellow.

The recommender system

To provide information on what type of new food venue has the best chances to succeed in the towns a difference of number of venues in a category in a town and mean category value for a relevant cluster was made. Figure 6 shows part of the dataset used for this calculation (first 9 venue categories). Please note that cluster 3 includes a single town – Moss due to its specific character. Recommendations cannot be therefore made for this town.

	Cluster Labels	Town	American Restaurant	Asian Restaurant	Bakery	Bistro	Burger Joint	Café	Chinese Restaurant	Diner	Fast Food Restaurant
0	0	Arendal	0	0	0	1	2	4	1	0	1
1	1	Bergen	4	4	7	1	3	14	1	0	2
2	0	Bodø	0	1	0	0	1	6	0	0	1
3	1	Drammen	1	0	2	1	3	3	1	0	5
4	2	Fredrikstad/Sarpsborg	2	0	2	0	1	6	0	0	4
5	1	Halden	1	1	0	0	0	2	0	0	0
6	2	Hamar	1	0	0	0	0	3	0	0	2
7	1	Haugesund	0	1	3	0	0	2	0	0	3
8	2	Kongsberg	0	0	1	0	0	3	0	1	1
9	1	Kristiansand	0	0	1	0	3	1	2	0	3
10	0	Larvik	0	0	0	0	1	1	0	0	1
11	3	Moss	0	0	2	0	0	3	1	0	4
12	1	Oslo	1	3	6	4	8	7	1	1	2
13	2	Porsgrunn/Skien	0	1	0	0	1	4	0	1	2
14	2	Sandefjord	0	0	0	0	0	4	0	0	2
15	1	Stavanger/Sandnes	0	4	3	0	2	7	3	3	1
16	1	Tromsø	4	1	1	2	1	12	2	0	1
17	1	Trondheim	0	2	2	2	3	9	1	2	3
18	1	Tønsberg	1	1	2	0	1	1	0	0	0
19	0	Ålesund	0	0	0	0	0	5	0	0	3

	American Restaurant	Asian Restaurant	Bakery	Bistro	Burger Joint	Café	Chinese Restaurant	Diner	Fast Food Restaurant
Cluster Labels									
0	0.000000	0.014706	0.000000	0.014706	0.071895	0.229184	0.014706	0.000000	0.089798
1	0.031197	0.037612	0.058758	0.015229	0.047305	0.114529	0.021935	0.008425	0.053139
2	0.027189	0.018182	0.031085	0.000000	0.024633	0.270378	0.000000	0.036364	0.139693
3	0.000000	0.000000	0.133333	0.000000	0.000000	0.200000	0.066667	0.000000	0.266667

Figure 6. Upper figure: total number of venues in each category (first 9 categories) in each town. Lower figure: mean value of each category (first 9 categories) in each cluster. Difference between these two values in each town was used as a base for the recommender system.

The resulting recommendations are shown in Figure 7. The figure shows the top 3 food venue categories that have the largest chances to succeed.

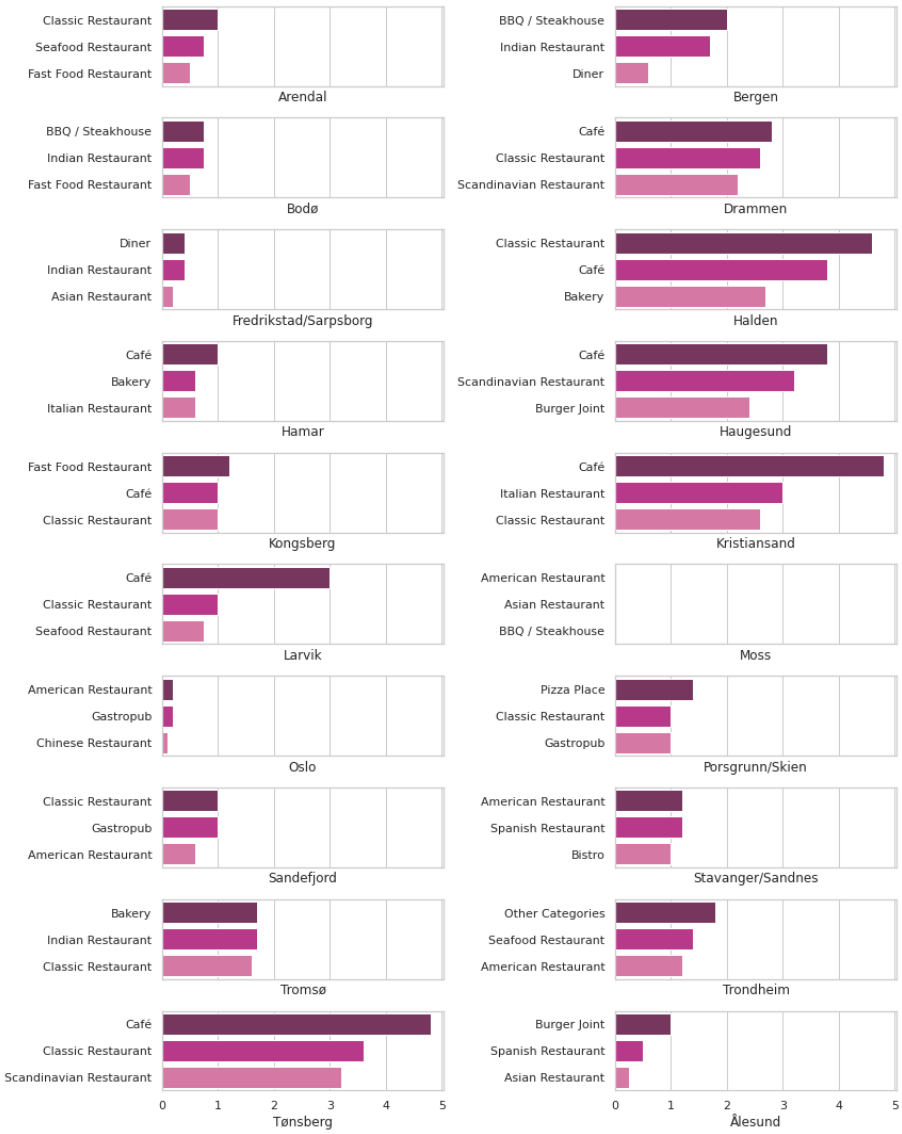

Figure 7. Food venue types that have the largest chances to succeed (top 3). The higher the number the higher chances for success according to the model. For model limitations see discussion chapter.

Discussion

The model has its own limitations that should be included in a follow-up study. Firstly, the model does not consider all venues in the tows: the number is limited both by the radius – 10 km, and maximum number of venues that are acquired from Foursquare API for one call – 100 venues. Some of the towns exceed these limits. Secondly, the clustering results using k-means algorithm are, to some degree, random. Therefore, we might expect slightly different results if more iterations were performed. Thirdly, in order to provide a better-quality recommendations about most likely successful food venues, one needs to take into account socio-economic factors such as: age and income distribution and understand habits of the towns' inhabitants.

Conclusions

1. The study gives an overview of the most common food venue types in the 20 most populated Norwegian towns and cities. The 4 most common are: cafés, classic restaurants, fast food and pizza places.

2. Some of the venue types such as café are common in (almost) all towns, while some categories, such as sushi, cannot be found everywhere.
3. The study grouped towns into clusters according to food venue category importance and thus to tastes of inhabitants. Cluster 0 incorporates: Arendal, Bodø, Larvik, Ålesund. Cluster 1 incorporates: Bergen, Drammen, Halden, Haugesund, Kristiansand, Oslo, Stavanger/Sandnes, Tromsø, Trondheim, Tønsberg. Cluster 2 incorporates: Fredrikstad/Sarpsborg, Hamar, Kongsberg, Porsgrunn/Skien, Sandefjord. Cluster 3 incorporates Moss only.
4. A recommender model was developed to predict the venue types that have the best chances to succeed. For specific recommendations see Figure 7.