Final report: Classification of Stockholm neighborhoods using Foursquare data

Oxana Falk, July 2019

Summary

This assignment aims to analyze different neighborhoods of Stockholm by exploring venues surrounding the metro stations and describe different clusters of neighborhoods. Unsupervised machine learning method, K-means clustering, is used to carry out the analysis. In total, three clusters were identified and analyzed. The main conclusion is that there are significant differences between the clusters in number of venues surrounding metro stations and what venues are the most common ones in the city center compared to the other areas in Stockholm. However, when it comes to the most common categories of venues, all the clusters look similar to each other, indicating that Stockholm is quite homogeneous in terms of available venue categories.

1. Introduction

Stockholm is one of the five fastest growing regions in Europe. The city population has grown by almost 30 percent during last 20 years and in 2020, Stockholm is expected to have one million inhabitants.

This project aims to describe different parts of Stockholm by studying neighborhoods surrounding metro stations. There are 100 stations in use of which 82 belong to the city of Stockholm and the rest in neighboring municipalities that are relatively close to the city center.

The result can be useful for someone who is planning to move to Stockholm and would like to find a place that fits their lifestyle and interests.

2. Data

2.1 Data sources

The data sources used in this assignment:

1. List of the metro stations in Stockholm and their geographical coordinates web scraped from Wikipedia.
2. Foursquare venue data:
   - In order to explore and classify the neighborhoods surrounding each metro station, Foursquare Venue Data was used. The average distance between two metro stations in Stockholm is approximately 1 km (both the arithmetic mean and the median). In order to avoid overlapping and to cover reasonable walking distance around each station, the radius of the query was limited to 500 meters.
   - List of the venue categories on the main level and on lower levels was downloaded from Foursquare Venue Categories and used in the exploratory data analysis section.

2.2 Data cleaning

For the metro station data, the following cleaning was performed:

- Only opened and functioning stations were used.

- Duplicates were removed. In the original table there are some stations that are represented on multiple metro lines but since the fact that those stations have the same name and coordinates and metro lines are not of interest for this project, only unique combinations of station names and coordinates were kept.
- Information about previous station names was removed since it's not relevant for this assignment.

For the Foursquare data, the following cleaning was performed:

- Venue category "Metro station" was removed from the requested data since it's our entry variable and would only mislead the result of clustering.

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Stockholm metro

The metro station data from Wikipedia has information about longitude, latitude and distance from the city center, i.e. the central train station. Every station is visualized with a dark blue marker on the map of Stockholm (see Figure 1). Of 100 stations in total, 82 belong to the municipality of Stockholm but all the stations are kept for this project since the remaining 18 stations are still quite close to the city center and contribute to a better coverage of the area.

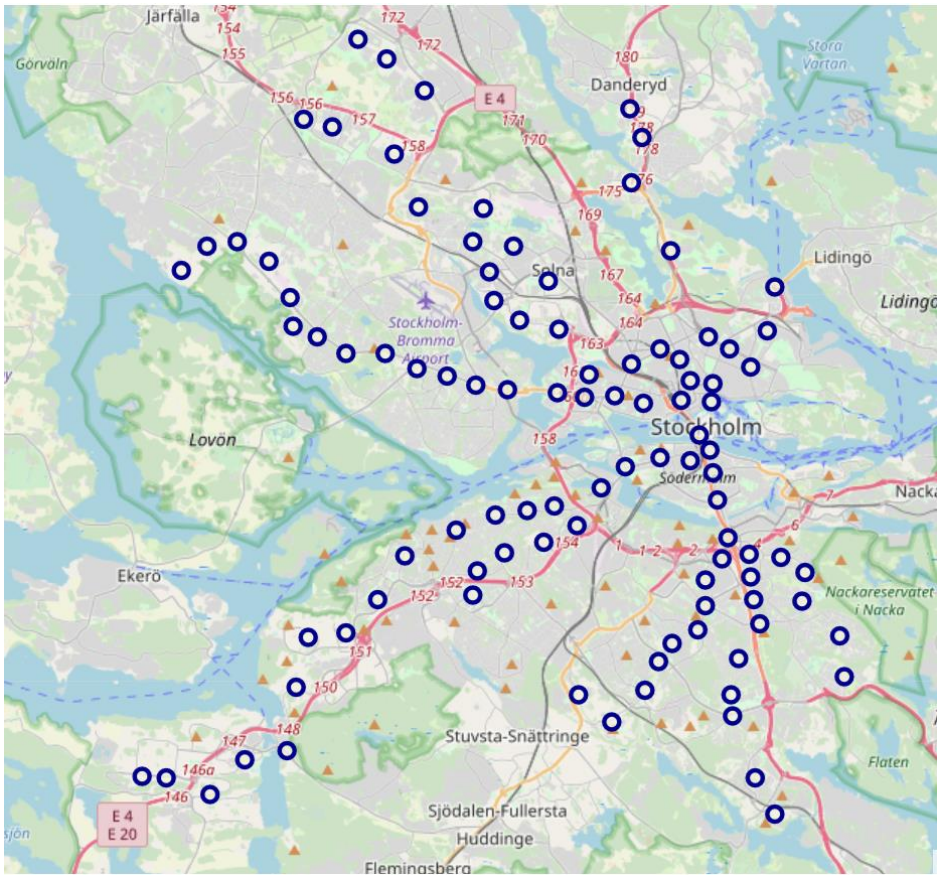

Figure 1: Map of Stockholm with metro stations (dark blue markers)

The distance of the metro stations from the city center (the central train station) is as following:

	Distance (km)
mean	8.0
std	4.9
min	0.0
25%	4.5
50%	7.3
75%	10.9
max	22.3

Table 1: Descriptive statistics – the distance between the Stockholm metro stations and the city center

The length of the metro lines varies between 15.1 and 28.6 kilometers.

3.1.2 Foursquare data

The venue data from Foursquare with limit of up to 100 venues and 500 meters surrounding each station has resulted in a dataframe with 2332 rows[1]. Slightly more than a half of the dataset belongs to the Food category while some of the main categories are not represented at all (see Table 2).

Main category	No of venues	Percent
Food	1175	50.4
Shop & Service	429	18.4
Outdoors & Recreation	265	11.4
Travel & Transport	165	7.1
Nightlife Spot	155	6.6
Arts & Entertainment	113	4.8
Professional & Other Places	26	1.1
College & University	4	0.2
Event	0	0.0
Residence	0	0.0
Total	**2332**	**100**

Table 2: Main categories for Stockholm venues, amount and percentage

Not surprisingly, the 10 largest categories on the lower level are largely part of the Food category, with the exception of grocery stores, hotels and gyms (see Table 3).

Category	No of venues

[1] As mentioned in the data cleaning section, the venue “Metro stations” was removed from the dataset as it is the entry variable in this project.

Café	122
Scandinavian Restaurant	113
Pizza Place	101
Grocery Store	80
Bakery	76
Hotel	69
Gym / Fitness Center	66
Italian Restaurant	65
Sushi Restaurant	55
Thai Restaurant	51

Table 3: Top 10 venue categories and number of venues

Number of venues per neighborhood varies a lot (see Figure 2), from 1 to 100 with the average of approximately 23 venues and the median of 12 venues. Half of the neighborhoods has between 5 and 22 venues surrounding them.

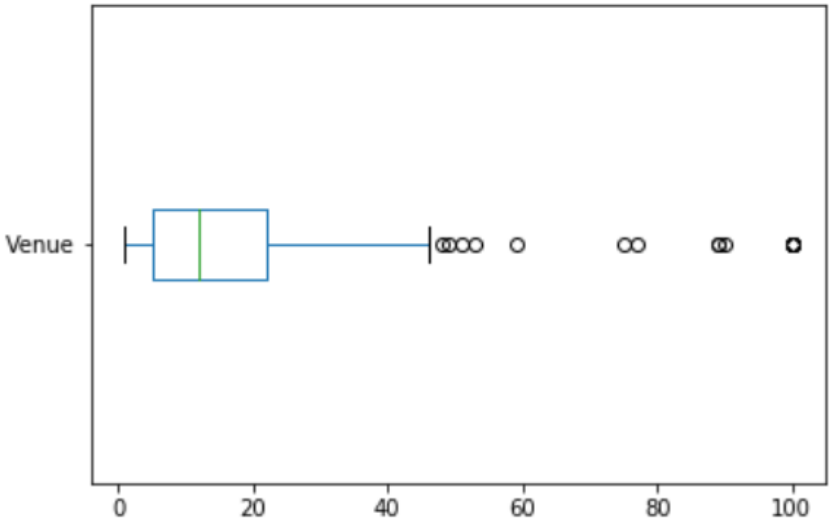

Figure 2: Boxplot for number of venues per neighborhood (metro station) in Stockholm

Note that the maximum of 100 venues depends on the limit when requesting Foursquare data. In reality those stations are most likely to have more than 100 venues surrounding them.

As expected, the number of venues decreases the further away the metro station is from the city center (see Figure 3).

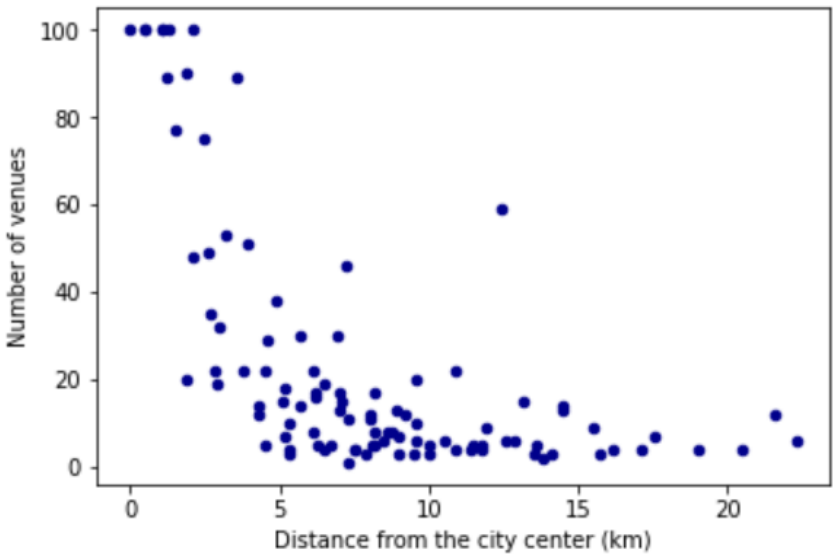

Figure 3: Scatter plot, no of venues per neighborhood vs. distance from the center of Stockholm

3.2 Data preparation

As shown in the previous section, some of the stations have too few venues to get reliable results. Stations with 5 or less venues were therefore removed from the analysis and formed their own "cluster" in the result section. I believe that some people would prefer to move to a calm area with almost no venues, so I found it important to keep those 30 neighborhoods in the final result section.

After removing all stations with 5 or less venues, 2216 rows are left in the dataset, which means that the deletion reduced the original dataset by approximately 5 percent. The dataset used for the clustering contains 248 unique categories and 70 metro stations.

In order to prepare the dataset for clustering and deal with categorical values (venue categories), one hot encoding was performed. After that, the data was normalized resulting in a new dataframe with 70 rows representing every metro station and 248 categories as columns where sum of the share of venues per row is equal to 1 (i.e. normalized).

3.3 Clustering

3.3.1 Finding the optimal number of clusters

In order to try to determine the optimal number of clusters (k), two different methods were used – the elbow method and the silhouette score. Both of the analyses were run on a range between 2 and 9 clusters.

According to the elbow method, the sum of squared errors (SSE) has been calculated for every *k* and plotted against number of k (see Figure 4). The “elbow” is not very clear but a slight bend in the curve at *k* = 3 can be an indication that this might be the best number of clusters.

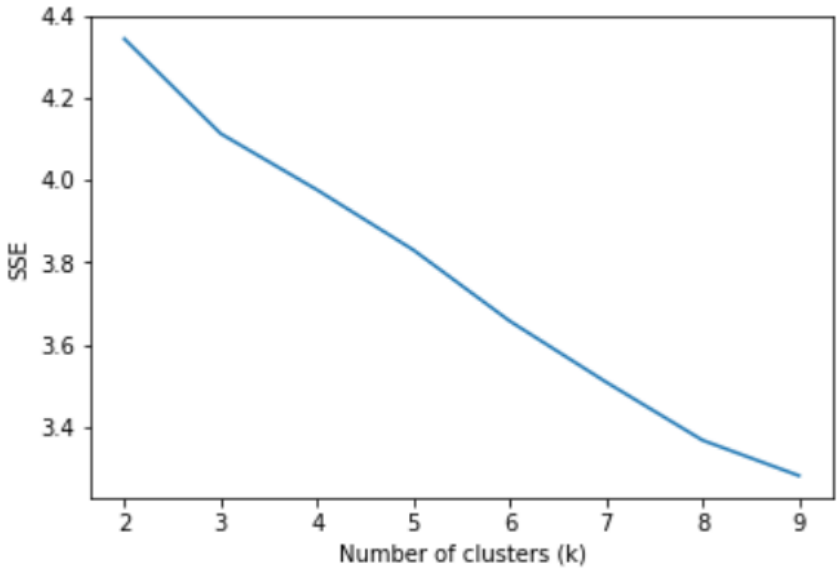

Figure 4: The elbow method

The silhouette score can take values between -1 and 1 and measures how similar a neighborhood is to the assigned cluster compared to the other clusters. Values close to 1 indicate that the neighborhood is well matched to its own cluster and poorly matched to the rest of the clusters. Values around 0 indicate overlapping between the clusters, while negative values indicate that the neighborhood is assigned to the wrong cluster.

The silhouette method indicates that the optimal number of clusters for the dataset is k = 3 since it has the highest score (see Table 4).

No of clusters (*k*)	Silhouette score
2	0.1232
3	0.1523
4	0.1061
5	0.0117
6	0.0192
7	0.1298
8	0.0833
9	0.1310

Table 4: Top 10 venue categories and number of venues

Worth mentioning that the highest silhouette score (0.1523) is closer to 0 than 1, which indicates that the clusters are close to each other in content.

3.3.2 K-Means clustering

K-Means clustering with 3 clusters was performed. See Figure 5 for the visualization of the result on the map of Stockholm. As mentioned above, 30 stations that have 5 or less venues were excluded from the clustering but kept on the map as their own group of neighborhoods. Those are visualized as black circles. The red markers represent cluster 1, the green markers – cluster 2 and the blue markers – cluster 3.

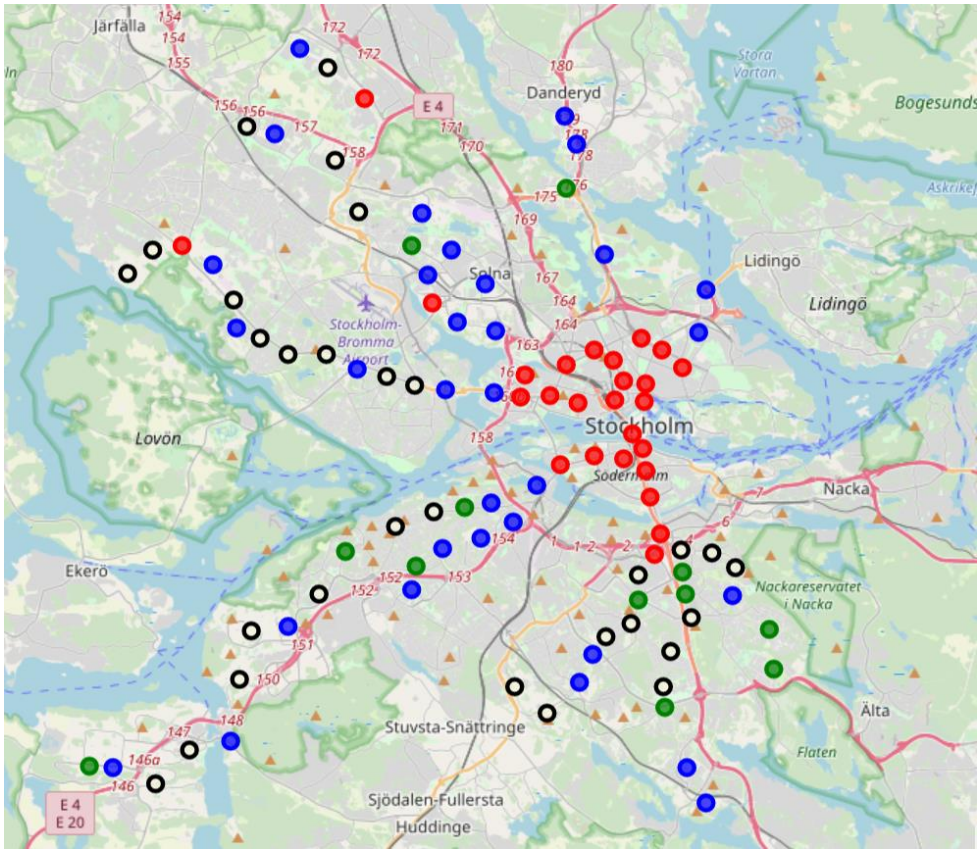

Figure 5: Stockholm neighborhoods divided into 3 clusters (colored) and neighborhoods that weren't included in the clustering due to lack of sufficient number of venues (black circles)

To describe the clusters, number and percentage of venues belonging to the main categories per cluster were calculated (see Table 5). The color scheme is similar to the map above where the first cluster has red color, the second – green etc.

Main category	*k* = 1 (26 metro stations)		*k* = 2 (12 metro stations)		*k* = 3 (32 metro stations)	
	No of venues	Percent (%)	No of venues	Percent (%)	No of venues	Percent (%)
Arts & Entertainment	85	5.4	7	5.3	16	3.2
College & University	3	0.2	0	0.0	1	0.2
Event	0	0.0	0	0.0	0	0.0
Food	854	53.8	64	48.1	214	43.1
Nightlife Spot	142	8.9	4	3.0	9	1.8
Outdoors & Recreation	143	9.0	23	17.3	76	15.3
Professional & Other Places	22	1.4	1	0.8	2	0.4
Residence	0	0	0	0.0	0	0.0
Shop & Service	247	15.6	24	18.0	129	26.0
Travel & Transport	91	5.7	10	7.5	49	9.9
Total	**1587**	**100**	**133**	**100**	**496**	**100**

Table 5: Number and percentage of venues by main category and cluster

The total number of venues is largest in the central parts of the city (k = 1), followed by the neighborhoods in k = 3. The total number of venues in k = 2 is relatively small despite the fact

that many of the metro stations belonging to this cluster are as close to the city center as the metro stations within k = 3.

The most common category is Food, followed by Shop & Service and then by Outdoor & Recreation throughout all of the clusters. Proportion of Nightlife spot is by far highest in k = 1 while the proportion of Outdoor & Recreation is approximately twice as high in k = 2 and in k = 3 compared to k = 1.

The top 7 most common venues on the lower level of classification for each cluster are:

***k* = 1, the most common venues and their occurrence (in brackets)**	***k* = 2, the most common venues and their occurrence (in brackets)**	***k* = 3, the most common venues and their occurrence (in brackets)**
Scandinavian Restaurant (96)	Pizza Place (25)	Café (32)
Café (84)	Grocery Store (17)	Grocery Store (30)
Hotel (55)	Bakery (8)	Pizza Place (28)
Bakery (54)	Scandinavian Restaurant (7)	Convenience Store (21)
Italian Restaurant (53)	Bus Stop (7)	Gym / Fitness Center (19)
Coffee Shop (45)	Gym / Fitness Center (6)	Sushi Restaurant (19)
Gym / Fitness Center (37)	Thai Restaurant (5)	Thai Restaurant (17)

Table 6: Top 7 most common venues per cluster and their occurrence within the cluster

According to the result in Table 6, the central parts of the city are associated mostly with places to eat and drink while the other clusters include more mixed combination of the most common venues.

4. Results

Different clusters can be categorized as followed:

- Cluster 1 (*k* = 1, red) has the highest number of venues and is mainly located in the city center. On average, there are more than 60 venues surrounding every metro station within this cluster. The most common categories of venues beginning with the largest category are Food, Shop & Service, Outdoors & Recreation, Nightlife Spot, Travel & Transport and Arts & Entertainment. Within category Food, the most common venues are Scandinavian restaurants, cafés, bakeries, Italian restaurants and coffee shops.
- Cluster 2 (*k* = 2, green) has the lowest number of venues among the clusters, with approximately only 11 venues per metro station on average. This cluster is located outside the central part of Stockholm. The most common categories of venues are Food, Shop & Service and Outdoors & Recreation. Within category Food, the most common venues are pizza places, bakeries, Scandinavian and Thai restaurants.
- Cluster 3 (*k* = 3, blue) has an average of approximately 16 venues per metro station and is located mainly outside of the city center. The most common categories of venues are Food, Shop & Service, Outdoors & Recreation and Travel & Transport. Within category Food, the most common venues are cafés, pizza places, Sushi and Thai restaurants.

For someone who prefers to move to calm areas with very few venues (5 or less), there are 30 neighborhoods in Stockholm that might fit. Please see Appendix for the list of the metro stations and information about their cluster.

5. Discussion

The purpose of this study was to cluster different neighborhoods in Stockholm based on areas surrounding every metro station. For that, Foursquare venue data was used. Foursquare is mainly focused on food places and for someone who is interested in other kind of parameters, e.g. housing prices or what schools are available, it would be interesting to add this kind of information to the analysis.

Another possible development is to include a larger area than just Stockholm municipality and find another way of grouping neighborhoods. This could potentially be valuable for someone who prefers riding a bike or using a car instead of public transportation and would like to discover other neighborhoods than just around the metro stations.

6. Conclusion

In this study, I tried to cluster different neighborhoods of Stockholm in order to help someone who are moving to Stockholm to find a place that would fit their lifestyle and preferences.

Three clusters were identified along with a group of metro stations that don’t have that many venues around them. The main differences between the clusters is the average number of venues per neighborhood and what venues are the most common ones in the city center compared to the other areas in Stockholm. Although, when it comes to the most common categories of venues, all the clusters look similar to each other, indicating that Stockholm is quite homogeneous in terms of available venue categories.

Appendix

List of the metro stations in Stockholm and their belonging to the clusters.

***k* = 1 (26 metro stations)**	***k* = 2 (12 metro stations)**	***k* = 3 (32 metro stations)**	**Neighborhoods with 5 or less venues, excluded from the clustering (30 metro stations)**
Fridhemsplan	Bagarmossen	Akalla	Abrahamsberg
Gamla stan	Bergshamra	Alvik	Alby
Globen	Blåsut	Aspudden	Axelsberg
Gullmarsplan	Bredäng	Bandhagen	Björkhagen
Hornstull	Duvbo	Blackeberg	Enskede gård
Hötorget	Hökarängen	Brommaplan	Fittja
Johannelund	Norsborg	Danderyds sjukhus	Gubbängen
Karlaplan	Sandsborg	Farsta	Hagsätra
Kista	Skarpnäck	Farsta strand	Hammarbyhöjden
Kungsträdgården	Sockenplan	Fruängen	Hjulsta
Mariatorget	Västertorp	Gärdet	Husby
Medborgarplatsen	Örnsberg	Hallonbergen	Hässelby gård
Odenplan		Hallunda	Hässelby strand
Rådhuset		Huvudsta	Islandstorget
Rådmansgatan		Hägerstensåsen	Mälarhöjden
Sankt Eriksplan		Högdalen	Rinkeby
Skanstull		Kristineberg	Rissne
Slussen		Kärrtorp	Råcksta
Solna strand		Liljeholmen	Rågsved
Stadion		Masmo	Skogskyrkogården
Stadshagen		Midsommarkransen	Skärmarbrink
T-Centralen		Mörby centrum	Stora mossen
Tekniska högskolan		Näckrosen	Stureby
Thorildsplan		Ropsten	Svedmyra
Zinkensdamm		Skärholmen	Sätra
Östermalmstorg		Solna centrum	Tallkrogen
		Sundbybergs centrum	Vårberg
		Telefonplan	Vårby gård
		Tensta	Ängbyplan
		Universitetet	Åkeshov
		Vällingby	
		Västra skogen	