

Segmentation and clustering of neighborhoods in Stockholm

Daniel Pustan

February 21, 2021

1. Introduction

Stockholm is the biggest city in Scandinavia estimated to reach a population of 3 million by 2045. More than 25% of those who move to Sweden chose to live in Stockholm. This project aims to describe Stockholm neighborhoods by studying their venues. There are 14 boroughs and 116 neighborhoods in Stockholm. The result can serve people planning to move to Stockholm, locals and businesses interested in the specificity of various neighborhoods in the city.

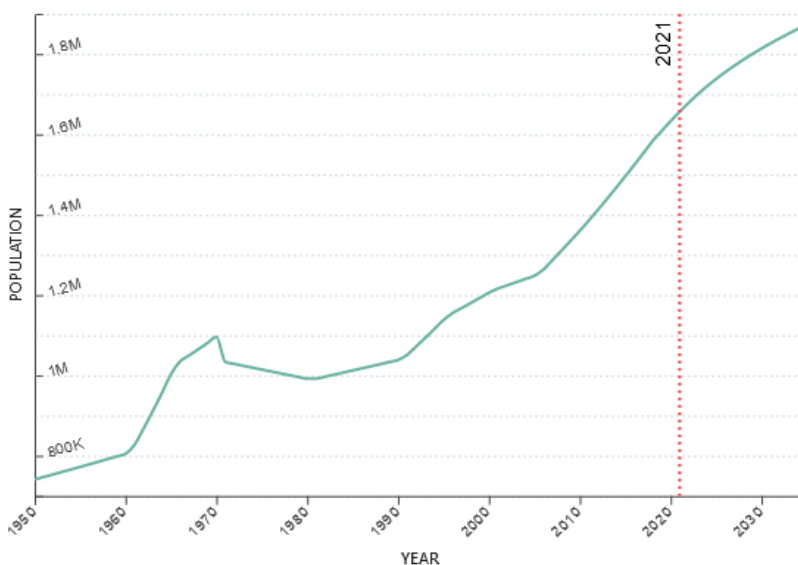


Figure 1 Stockholm population growth

2. Data

The data for this project comes mainly from the Foursquare database. Specifically, that related to neighborhoods and their coordinates, venues with the coordinates and venue category. In addition, a dataset comprising Stockholm boroughs and the corresponding neighborhoods stored in my Github repository was used.

The initial dataset contained 14 boroughs and 116 neighborhoods. However, only neighborhoods with more than five venues were considered to be representative for the study and kept in the final dataset. This resulted in a decrease of the sample size to 52 neighborhoods.

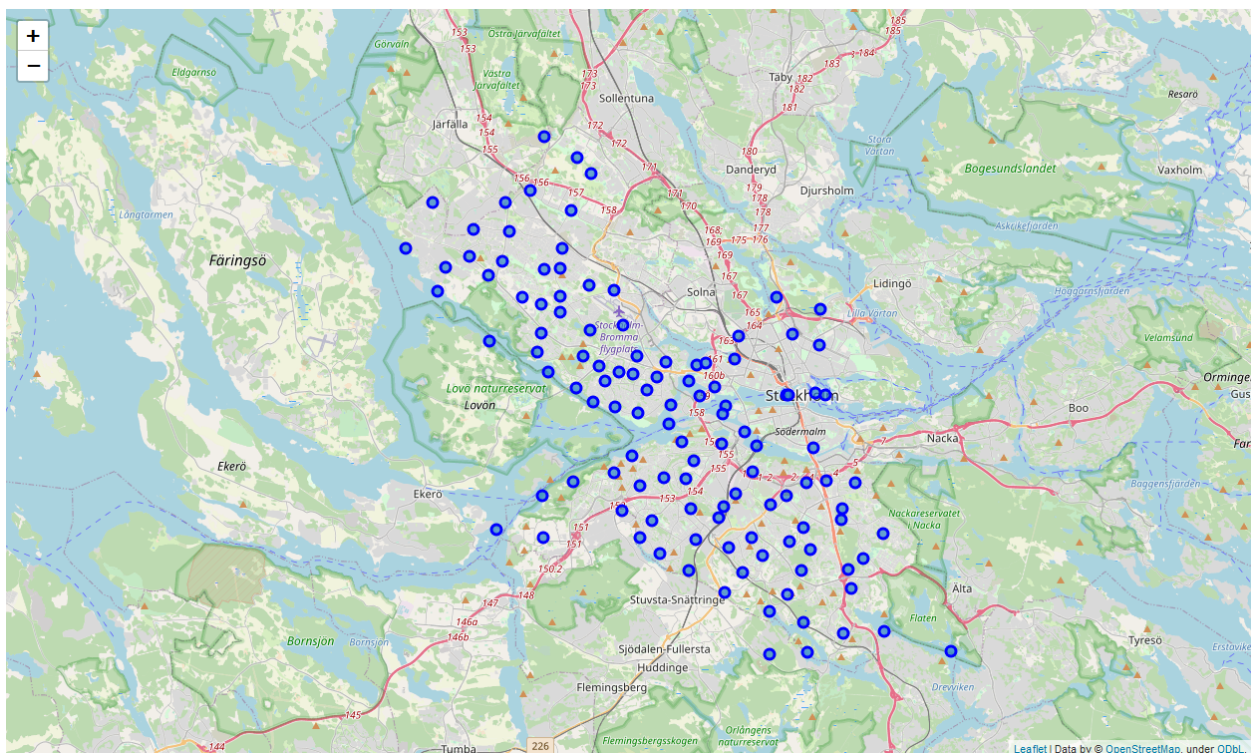


Figure 2 Map of Stockholm neighborhoods

Furthermore, the dataset was restricted to venues within a radius of 500 meters with respect to the center of their neighborhoods in order to avoid overlapping.

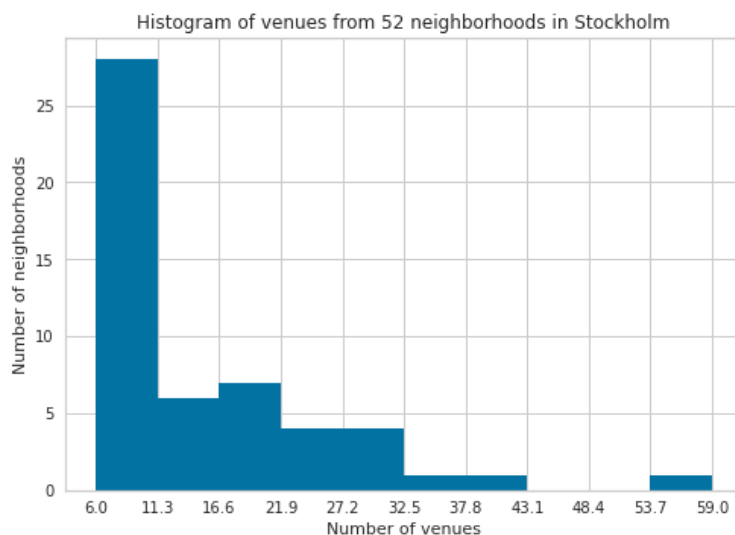
Neighborhood	Venue
Skeppsholmen	59
Stadshagen	41
Södermalm	36
Östermalm	32
Midsommarkransen	30
Ladugårdsgärdet	30
Normalm	30
Riddarholmen	27
Larsboda	24
Södra Hammarbyhamnen	23

Table 1 Neighborhoods by number of venues

Venue	Freq
Café	45
Pizza Place	39
Scandinavian Restaurant	39
Park	27
Bakery	24
Thai Restaurant	23
Gym / Fitness Center	22
Hotel	21
Convenience Store	19
Grocery Store	19

Table 2 Venues by category

Not surprisingly, the neighborhoods with the highest number of venues are located in the central area of Stockholm. 75% of the neighborhoods have 20 venues or less. The great majority of the venues belong to food related categories.

**Figure 3** Histogram of venues from 52 neighborhoods

	Venue
count	52.000000
mean	15.250000
std	10.820958
min	6.000000
25%	7.000000
50%	10.500000
75%	20.000000
max	59.000000

Table 3 Descriptive statistics

3. Methodology

Unsupervised learning is applied considering the unlabeled character of the data and the purpose of the study to group similar neighborhoods in Stockholm based on whether they share similar attributes such as venues category. Specifically, the clustering method is considered suitable for this project as it allows to discover structure based on the similarity of the neighborhoods to each other.

In this paper, two clustering methods are used, namely: *decision tree* (hierarchical clustering) and *k-means*. In addition, three approaches were taken to determine the optimal number of clusters (k), specifically: the Elbow, Silhouette and Gap statistic. All three analyses were run on a range between two and five clusters based on the knowledge acquired during the exploration stage of the dataset which indicated a relatively high level of similarity between neighborhoods.

3.1. Hierarchical clustering

Hierarchical clustering was chosen first not only because it does not require to specify the number of clusters but also because it can serve as an indicator for the optimal number of clusters (k) for the k-means method. Furthermore, the resulting dendrogram from hierarchical clustering is very useful in understanding the data.

The algorithm has been implemented using the agglomerative strategy. The criteria Complete-Linkage Clustering was used to determine the distance between clusters. This criteria identifies the longest distance between the points in each cluster.

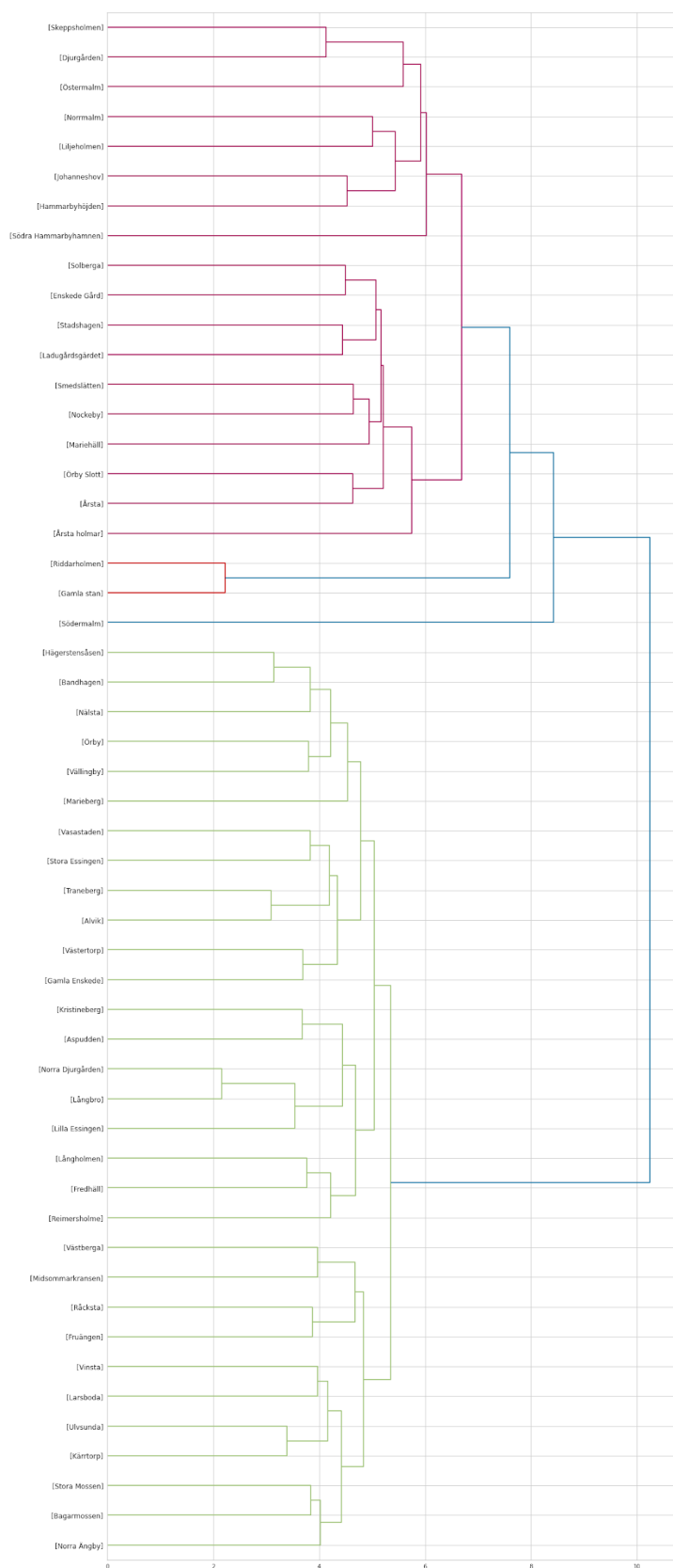


Figure 4 Dendrogram of Stockholm neighborhoods

The dendrogram indicates that dividing the data into three or four clusters would be optimal. This initial clue will be verified in the following section.

3.2. Determine the optimal number of clusters

Three clustering validation metrics were selected out of the many options based on their popularity and accuracy.

The Elbow method

It is one of the most popular methods for determining k . The idea behind is that the explained variation changes rapidly for a limited number of clusters and slows down after that forming an “elbow” like figure. The “elbow” or the point of inflection on the curve is an indicator that the model fits best at that point and that number represents the number of clusters to use for the clustering algorithm.

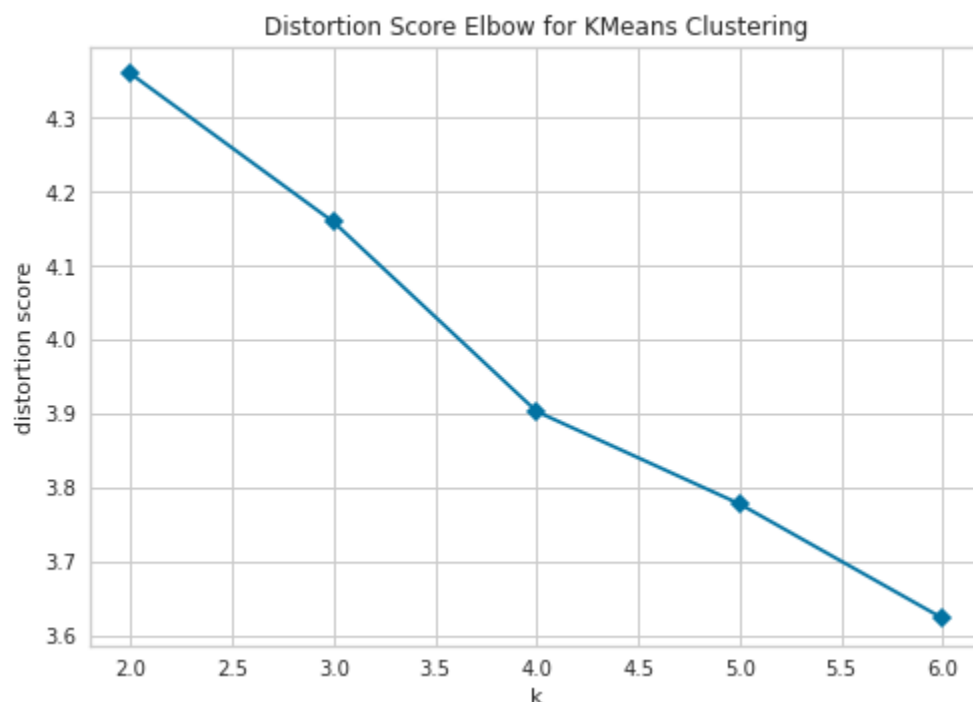


Figure 5 The Elbow method

Unfortunately, no elbow point was detected although there is a slight bent at $k = 4$. The smooth curve in the figure above shows that the data is not very clustered and therefore the optimal number of clusters remains unclear although there is a weak indicator to use four clusters.

The Silhouette method

The Silhouette method is considered a better alternative to the Elbow method. This method indicates if the neighborhoods are correctly assigned to their clusters. The score ranges between -1 and 1 where values close to 0 indicate that the neighborhood is between two clusters. Values closer to 1 shows that the neighborhood belongs to the correct cluster while values towards -1 indicates that the neighborhood would be better-off assigned to another cluster.

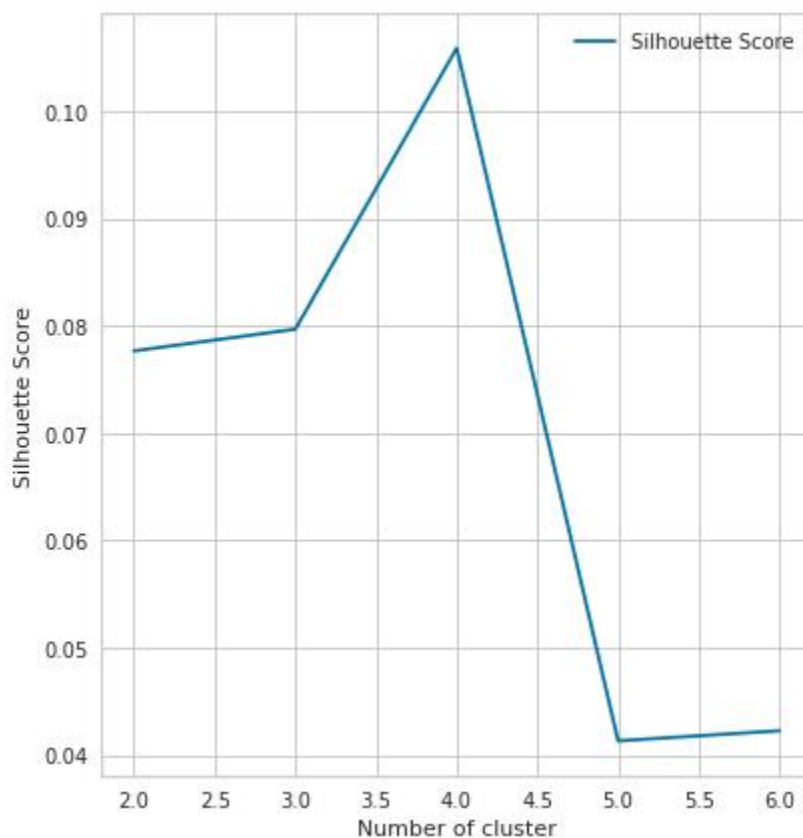


Figure 6 Average Silhouette score

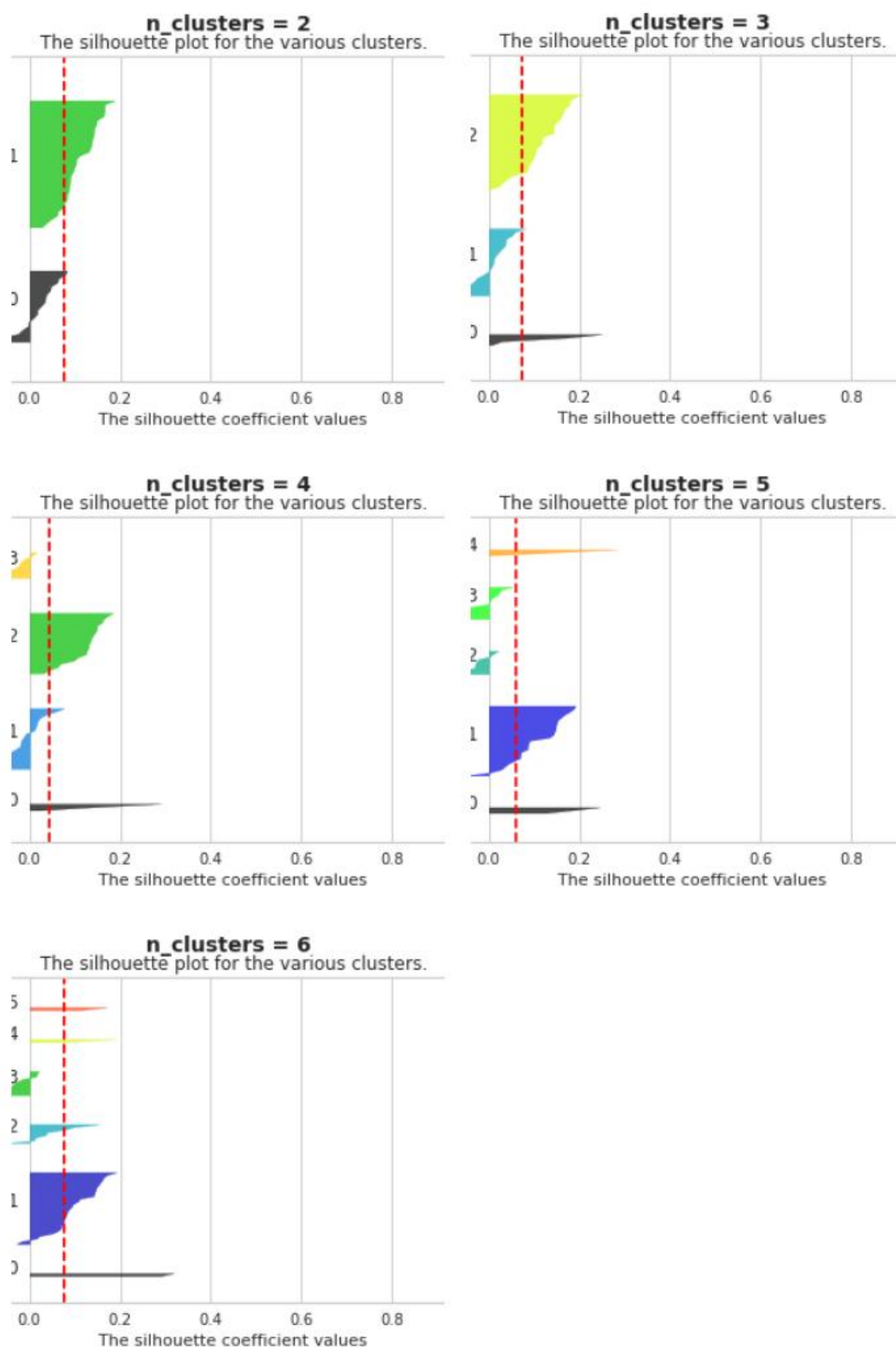


Figure 7 Silhouette Analysis for each cluster in KMeans with $n_cluster=[2,3,4,5,6]$

The Silhouette method suggests that the optimal number of clusters is $k = 4$ because it has the highest average score. However, from Silhouette analysis we can observe the size of each cluster from the thickness of the silhouette plot. We can further infer that clusters are relatively similar in content given their values closer to zero.

The Gap Statistic

The Gap statistic compares the total within intra-cluster variation for different values of k with their expected value under null reference distribution of the data. The estimate of the optimal clusters is the value that yields the largest gap statistic.

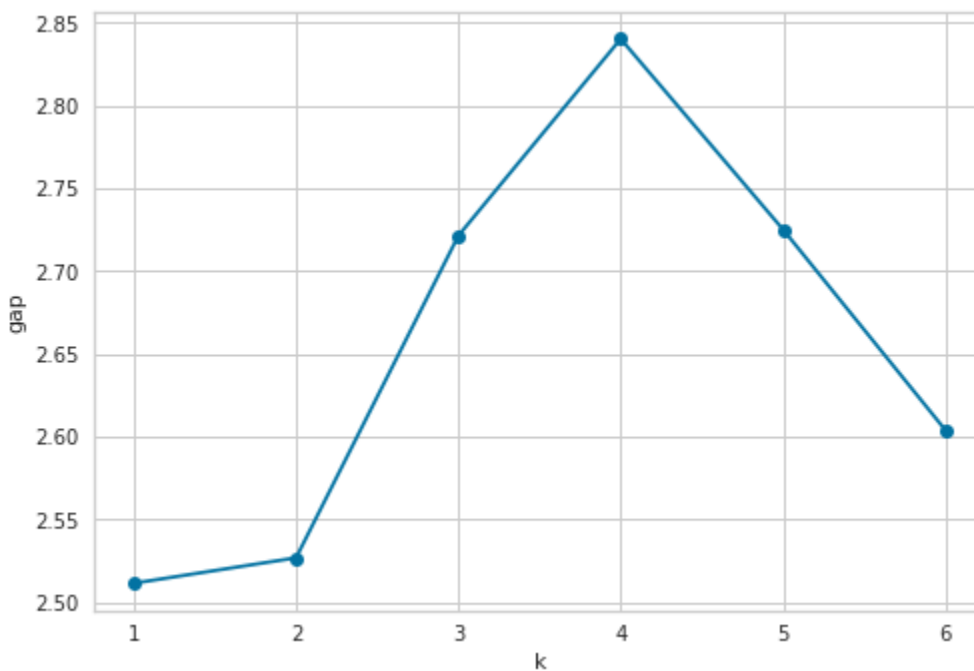


Figure 8 Gap statistics for various values of clusters

As seen in the figure above, the gap statistics is maximized with 4 clusters for our data. Hence, this method confirms the result of the previous two methods used to identify the optimal number of clusters.

3.3. K-means

Considering the coincidence in outcome of $k=4$ for all three methods in the previous section, K-means clustering was performed with four clusters. Moreover - as mentioned in the “Data” section - it should be noted that neighborhoods with less than six venues were excluded from the analysis.

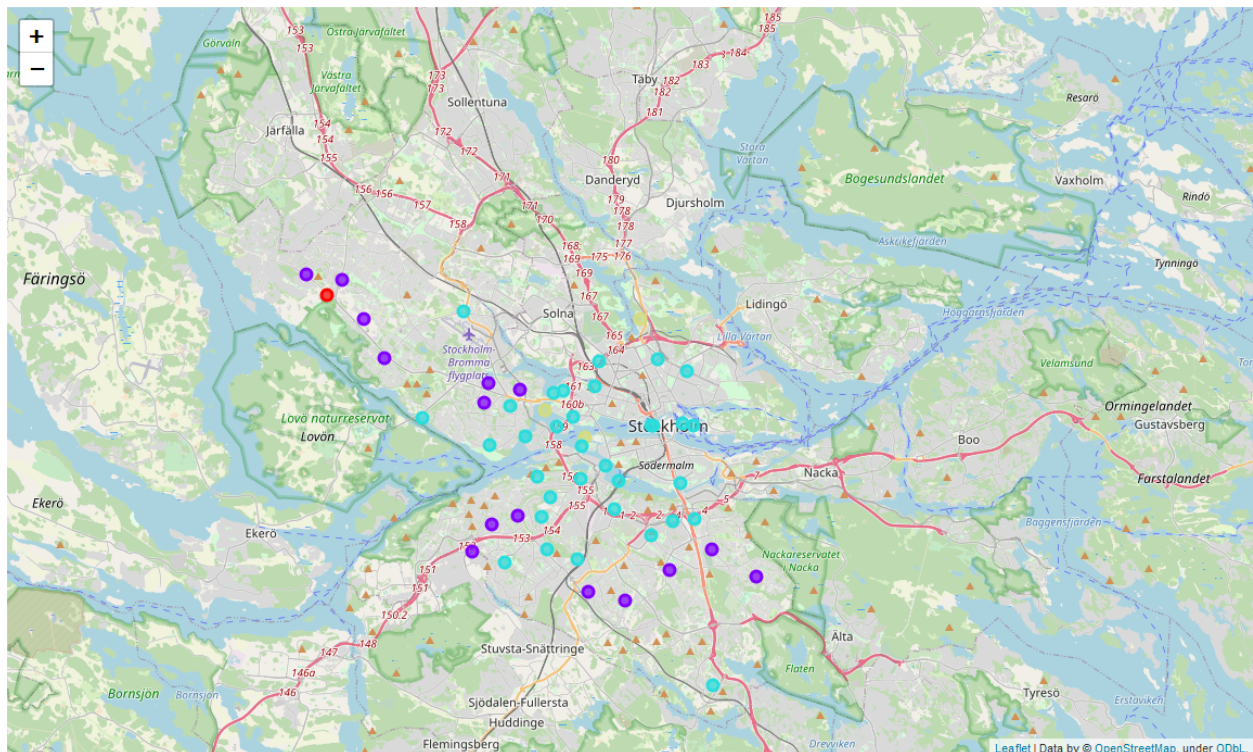


Figure 9 Stockholm neighborhoods divided into clusters

In order to describe the clusters, the number of venues' categories and percentages were calculated. The color of the clusters is represented with colored circles both on the map as well as in the text.






Category	 k = 0	 k = 1	 k = 2	 k = 3
	# %	# %	# %	# %
Bakery			21 3%	
Beach				6 26%
Bus Station		4 3%		
Bus Stop		4 3%		2 9%
Café		4 3%	37 6%	4 17%
Convenience Store		5 4%		
Event Space	1 14%			1 4%
Fast Food Restaurant	1 14%			
Grocery Store	1 14%	8 6%		
Gym / Fitness Center		6 5%	16 3%	
History Museum				1 4%
Hotel			16 3%	1 4%
Italian Restaurant			15 2%	
Metro Station	1 14%	11 8%		
Modern European Restaurant				1 4%
Park			26 4%	1 4%
Pizza Place		19 15%	20 3%	
Restaurant			16 3%	
Scandinavian Restaurant	1 14%		34 5%	3 13%
Supermarket	2 29%	6 5%		1 4%
Sushi Restaurant		5 4%		
Thai Restaurant			21 3%	

Table 4 Top ten venue categories by cluster

Interesting to note that while pizza places are present in both cluster 2 ($k = 1$) and 3 ($k = 2$) in about the same number, it has a significantly higher weight (15%) in the former than in the second (3%). On the other side, there are plenty of Scandinavian restaurants in the third cluster ($k = 2$) while in cluster 1 and 3 their presence is merely symbolic. Bus and metro stations have been assigned to cluster 2 as well as convenience stores and supermarkets.

4. Results

The four clusters can be categorized as follow:

-  **Cluster 1 ($k = 0$)** contains only one neighborhood (Vällingby) and seven venues which is the lowest number of venues in the sample. In addition to being a quiet neighborhood, the predominance of nearby supermarkets, events space, grocery stores and a metro station might very well make it suitable for the needs of the elderly.
-  **Cluster 2 ($k = 1$)** includes 15 neighborhoods located mostly on the outskirts of Stockholm. It is characterized by pizza places, metro and bus stations as well as grocery stores and gyms centers. It might be a good choice for medium income households.
-  **Cluster 3 ($k = 2$)** consist of 33 neighborhoods situated mainly in the central area of the city. Cafés, restaurants, parks and hotels abound in this cluster which makes it ideal for business people and higher income households.
-  **Cluster 4 ($k = 3$)** comprises three neighborhoods: Norra Djurgården, Fredhäll and Långholmen. The presence of beaches differentiate this cluster from the others as well as that of a museum. Both, the location of neighborhoods and the venue types suggest that it might be well adapted to the needs of higher income households with children and youngsters.

5. Discussion

The purpose of this study was to cluster various neighborhoods in Stockholm based on the venues found in a radius of 500 m from their center. Foursquare database was used to achieve this purpose. Considering that the Foursquare database is mostly based on

user generated data, it does not represent a comprehensive database and tends to overrepresent certain areas.

That is, it tends to capture popular places thus explaining the overwhelming presence of downtown venues in the dataset relative to the number of those situated on the outskirts. Second, it tends to limit to trendy places (such as cafes and restaurants) and it does not include all amenities that a neighborhood has to offer. Hence, it is understandable the lack of hospitals, elderly homes, kindergartens, etc. Further studies could explore this possibility further.

Another direction of exploration could be the addition of housing prices and household income to this data. Information about crime rates could further foster the analysis of neighborhoods and shed more light on the topic.

6. Conclusion

In this study, I clustered various neighborhoods in Stockholm to meet newcomers' need to easily grasp what sets apart different neighborhoods so they can make informed decisions when moving to this city.

Four clusters were identified with a significant difference in the total number of venues between them. The number of venues is highest for downtown neighborhoods decreasing progressively as we move towards the outskirts of town. Most venue categories are variations of the food category but this does not necessarily represent the reality but rather the effect of the way data is gathered.

In spite of this homogeneity, each cluster presents a representative category. In the first cluster ($k = 0$) we have the supermarket and in the second cluster ($k = 1$) it is the bus and metro stations. The third cluster ($k = 2$) concentrates most of the Scandinavian restaurants. Finally, the fourth cluster ($k = 3$) distinguishes by the presence of the beaches.