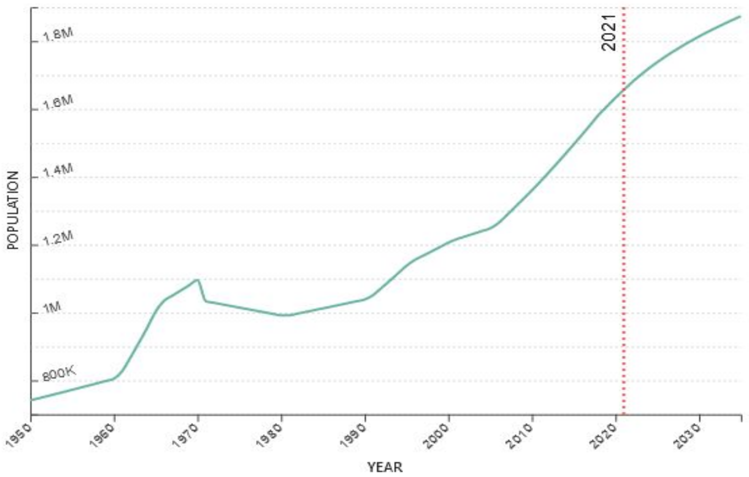
Segmentation and clustering of neighborhoods in Stockholm

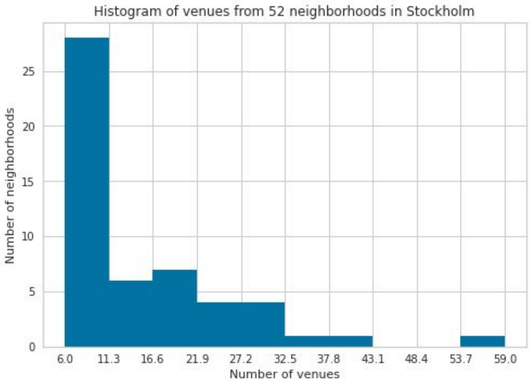
Daniel Pustan

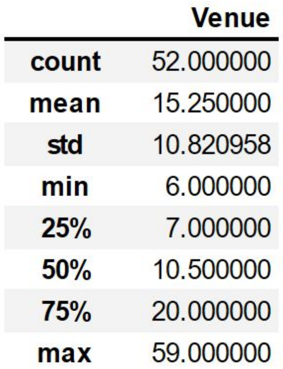
Introduction

- Stockholm is the biggest city in Scandinavia
- >25% of those who move to Sweden choose Stockholm
- Results offer insights into specificity of neighborhoods => can serve newcomers, locals and businesses

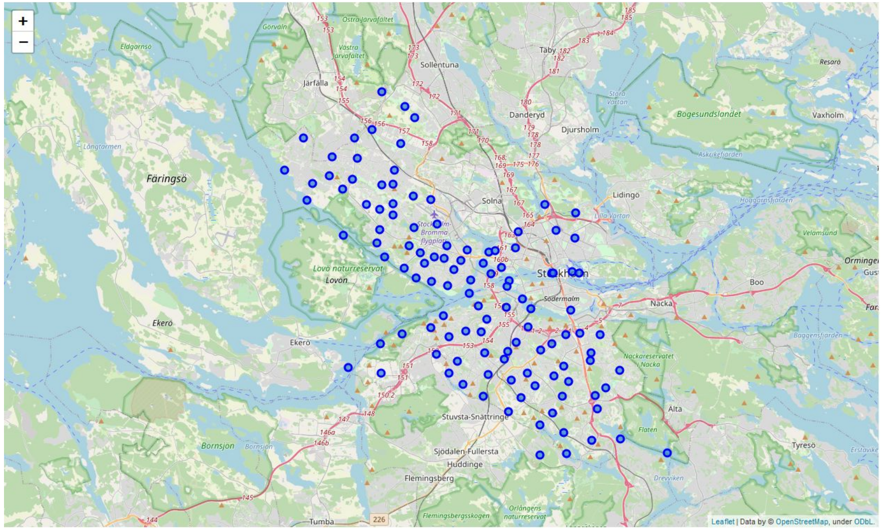


Data



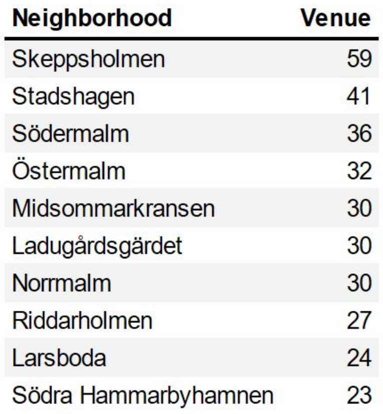


	Venue
count	52.000000
mean	15.250000
std	10.820958
min	6.000000
25%	7.000000
50%	10.500000
75%	20.000000
max	59.000000

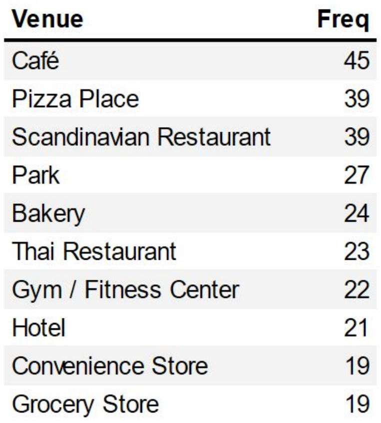


Data

- 14 boroughs and 116 neighborhoods
- 75% of neighborhoods have <20 venues
- Most venues are categorized as food



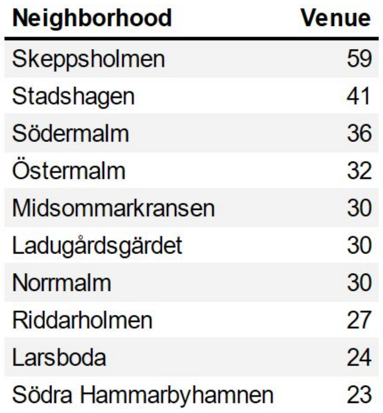
Neighborhood	Venue
Skeppsholmen	59
Stadshagen	41
Södermalm	36
Östermalm	32
Midsommarkransen	30
Ladugårdsgärdet	30
Norrmalm	30
Riddarholmen	27
Larsboda	24
Södra Hammarbyhamnen	23



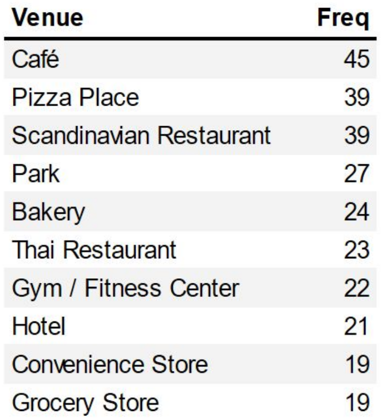
Venue	Freq
Café	45
Pizza Place	39
Scandinavian Restaurant	39
Park	27
Bakery	24
Thai Restaurant	23
Gym / Fitness Center	22
Hotel	21
Convenience Store	19
Grocery Store	19

Data

- 14 boroughs and 116 neighborhoods
- 75% of neighborhoods have <20 venues
- Most venues are categorized as food
- Data restricted to venues within 500 m from neighborhood’s center
- Removed neighborhoods with <6 venues



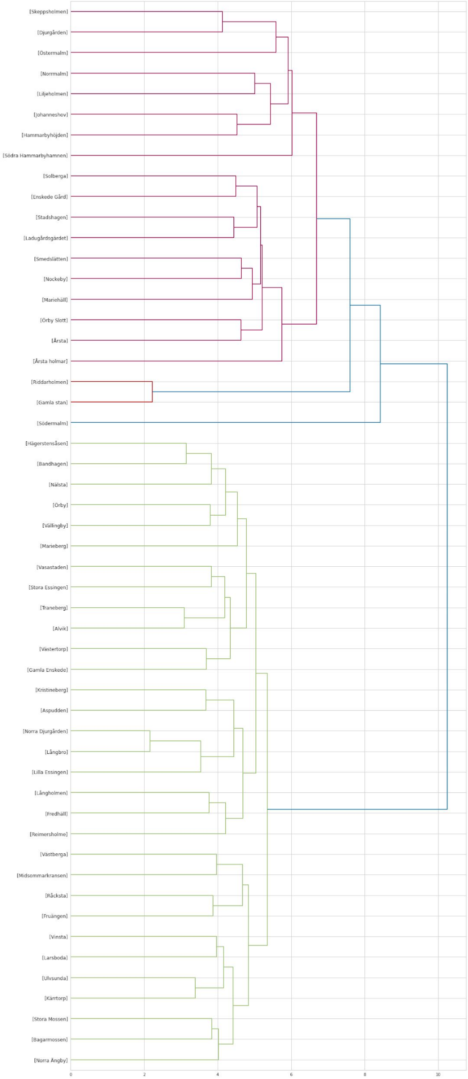
Neighborhood	Venue
Skeppsholmen	59
Stadshagen	41
Södermalm	36
Östermalm	32
Midsommarkransen	30
Ladugårdsgärdet	30
Norrmalm	30
Riddarholmen	27
Larsboda	24
Södra Hammarbyhamnen	23



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

Methodology

Hierarchical clustering

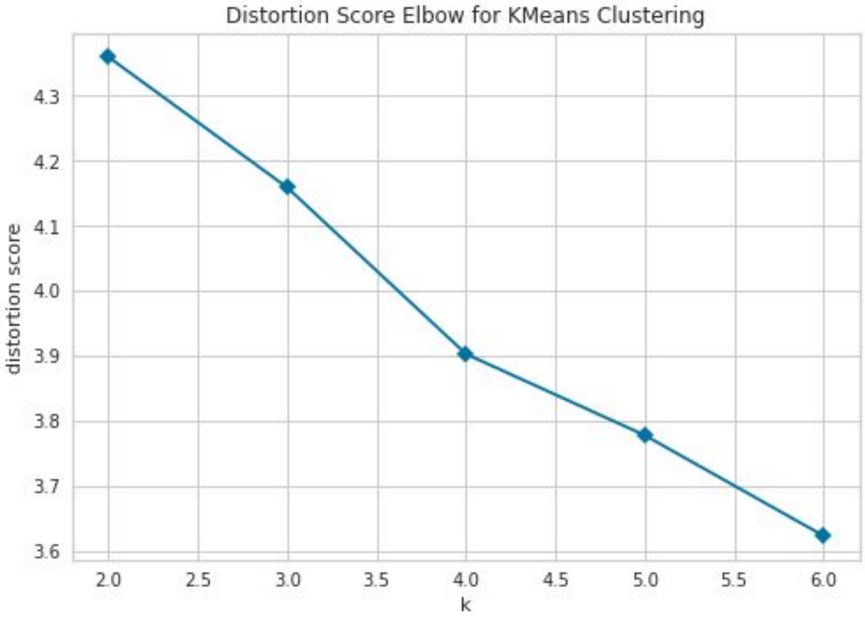


- Agglomerative strategy was used
- Complete-Linkage Clustering criteria to determine distance between clusters
- It indicates that dividing data into three or four clusters is optimal

Optimal k

The Elbow method

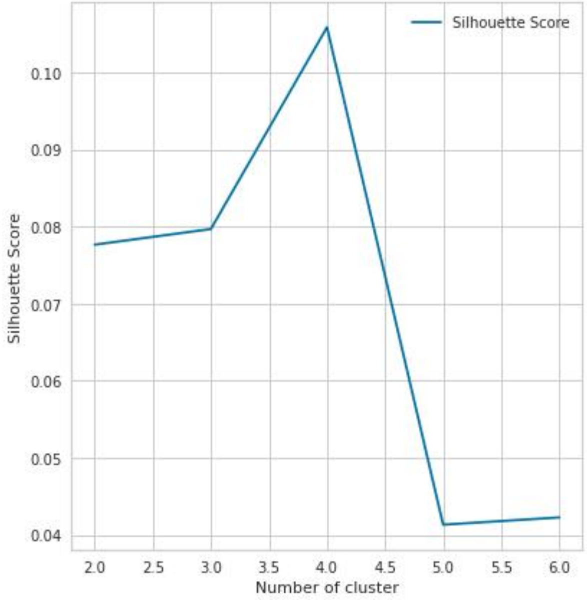
- No elbow point was detected
- Slight bent at k = 4
- Smooth curve => data is not very clustered



Optimal k

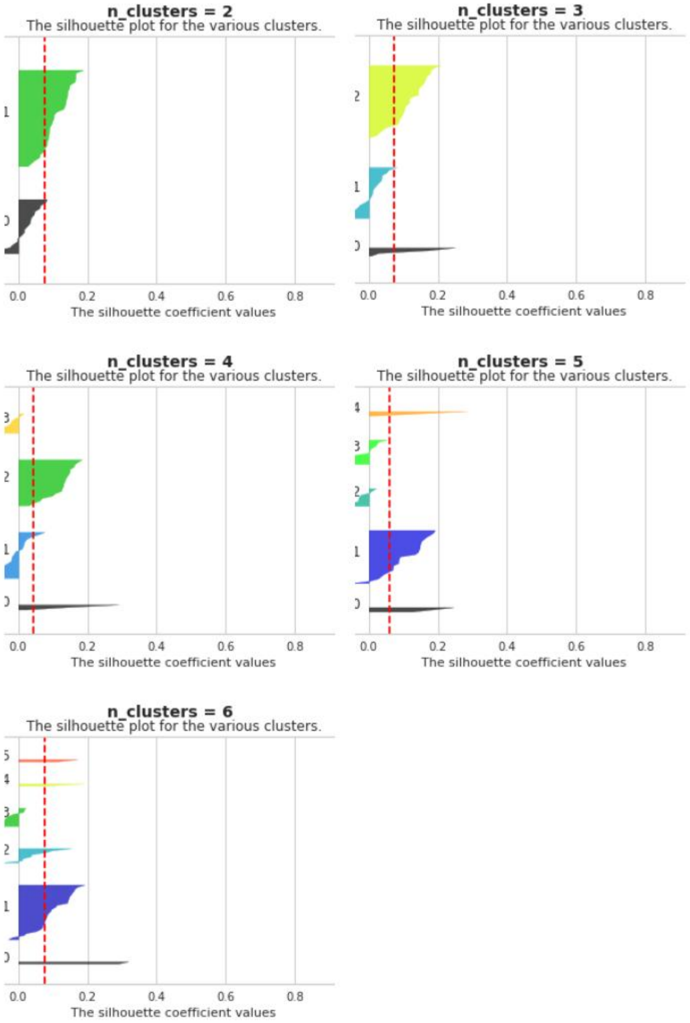
The Silhouette method

The average score suggests that the optimal number of clusters is k = 4



Optimal k

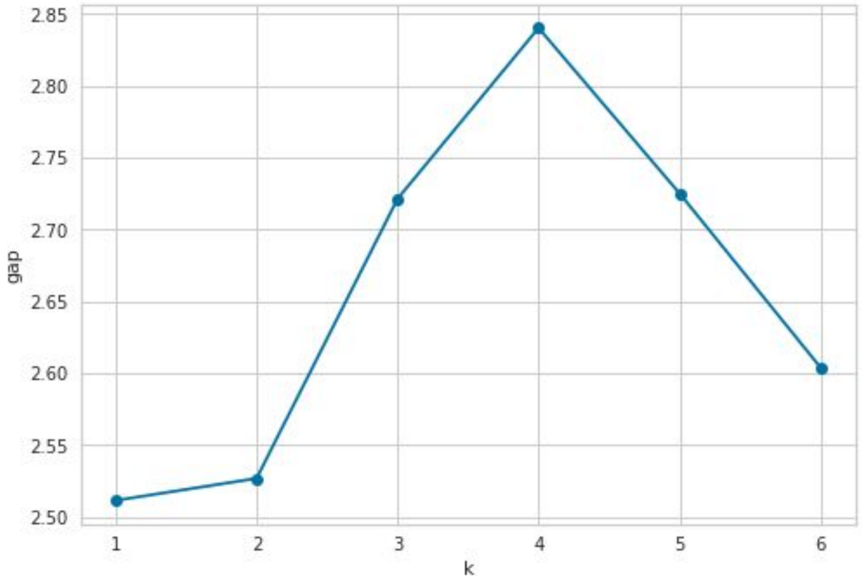
The Silhouette Analysis



Optimal k

The Gap statistic

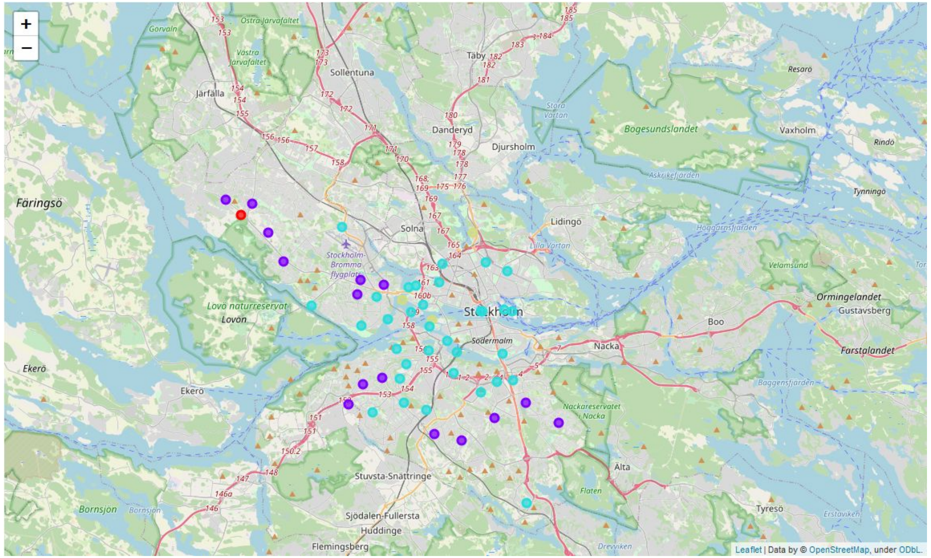
- The Gap statistic is maximized with 4 clusters
- The method confirms the previous results



Methodology

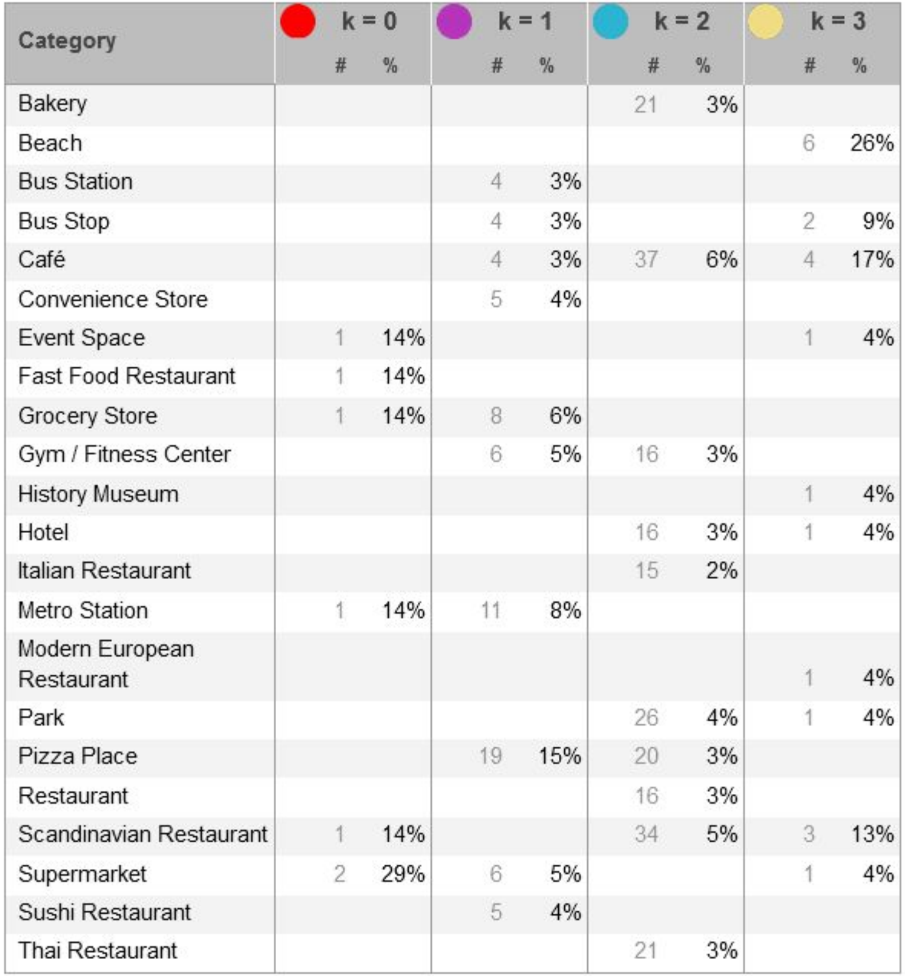
K-means

- k = 4
- For the description of clusters, number of venues’ categories and percentages were calculated



Methodology

K-means



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%		

Results

- 🔴 **Cluster 1 (k = 0)**: suitable for the needs of the elderly
- 🟣 **Cluster 2 (k=1)**: adapted for medium income households
- 🔵 **Cluster 3 (k=2)**: ideal for business people and higher income households
- 🟡 **Cluster 4 (k=3)**: higher income households with children and youngsters

Conclusion

- Clustered various neighborhoods in Stockholm to help newcomers make informed decision when moving
- Identified four clusters with specific venue types for each:
  - 🔴 **Cluster 1**: Supermarket
  - 🟣 **Cluster 2**: Bus and metro stations
  - 🔵 **Cluster 3**: Scandinavian restaurants
  - 🟡 **Cluster 4**: Beaches
- Number of venues is highest for downtown and decreases towards the outskirts of town

Future directions

- Expand the dataset with information on amenities that are currently not included in the dataset such as: hospitals, elderly homes, kindergartens, etc.
- Include the housing price and household income in the analysis
- Explore further using data on crime rates by neighborhood

Thank you!