IBM Data Science Professional Certificate - Capstone Project

Clustering Neighborhoods in Hamburg

Improving work of renal company

April 2020

Summary

Executive Summary

Executive Summary

The aim of this project is to examined city of Hamburg and explore venues around metro stations. After gathering all necessary data we take a look on different neighborhoods. Data has been divided using K-means clustering into tree clusters. In total, fours clusters were identified. Two of clusters are more common and two are unique. However, when it comes to the most common categories of venues, all the clusters look similar to each other, indicating that Hamburg is quite homogeneous in terms of available venue categories.

1. Introduction

The Hamburg U-Bahn is a rapid transit system serving the cities of Hamburg, Norderstedt and Ahrensburg in Germany. Although technically an underground, most of the system's track length is above ground. The network is interconnected with the city's S-Bahn system, which also has underground sections. With Metro System is easier and faster to get to other parts of the city. Many people looks for their flats situated next to metro. At the same time they want to live close to shops, restaurants, coffeshops or parks.

The company specializing in long-term as well as short-term apartments renting wants to obtain deeper knowledge about city to build housing recommendation system based on clients preferences. There is a lot of recommendation systems on the market now, but while doing research on Germany market we found out that there is none designed for meeting all needs in one place. People looking for housing needs to search among lots of websites to gain information they need. Different people can have their specific preferences such as favorites stores, shopping location etc. In my project I want to analyze different neighborhoods of Hamburg which will help to build final recommendation system in the future.

Center of the city of Hamburg:

2. Data

2.1. Data needed for analysis

For the project its needed to obtain:

- geo-locational information about metro stations in Hamburg - latitude and longitude of every station. Informations will be obtain based on names of stations. Data will be scrapped from wikipedia page: https://en.wikipedia.org/wiki/List_of_Hamburg_U-Bahn_stations and geo-location information will be added using geopy library.

- Forsqure API will be used to find location information about venues. Explore function will be used to get the most common venues categories next to each metro station like restaurants, art galleries, shops.

After collecting data some visualization and statistical analysis will be made. Location of stations will be shown on map prepared with folium library. As the next step this data will help to group metro neighborhoods in clusters. Thanks to collected data we will be able to compare neighborhoods and find differences between them which will help rental company to personalize they offer.

2.2. Data Preparation

The initial data consist of 92 Stations with their lines and Fare zone rings(s). Latitude and longitude needed to be gathered in second step.

	Station	Lines	Fare zone ring(s)
0	Ahrensburg Ost	U1	B
1	Ahrensburg West	U1	B
2	Alsterdorf	U1	A
3	Alter Teichweg	U1	A
4	Barmbek	U3	A
5	Baumwall	U3	A
6	Berliner Tor	U2, U3	A
7	Berne	U1	B
8	Billstedt	U2	A
9	Borgweg	U3	A

As a second step Lines and Fare zones column was deleted and latitude and longitude coordinations were added based on BeautifulSoap Library.

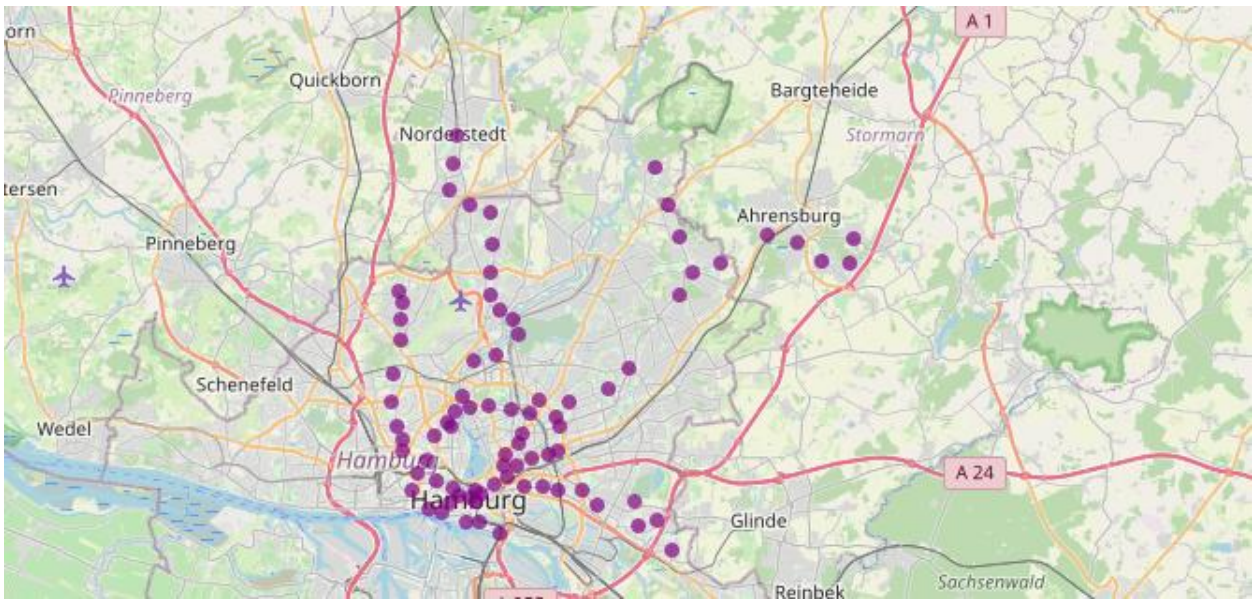
	Station	Latitude	Longitude
0	Ahrensburg Ost	NaN	NaN
1	Ahrensburg West	NaN	NaN
2	Alsterdorf	NaN	NaN
3	Alter Teichweg	NaN	NaN
4	Barmbek	NaN	NaN

	Station	Latitude	Longitude
0	Ahrensburg Ost	53.661347	10.242240
1	Ahrensburg West	53.664639	10.219403
2	Alsterdorf	53.610541	10.003889
3	Alter Teichweg	53.586202	10.064931
4	Barmbek	53.587386	10.044942

Next this data were used to obtain venue tapes from the Forsquere API. The radius from each point was set in 500 meters and the limit of venues retrieved per neighborhood as 100. They venues were merged with stations data frame. New data frame has 1992 rows and 7 columns.

	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ahrensburg Ost	53.661347	10.242240	Ringhotel Ahrensburg	53.661984	10.243314	Hotel
1	Ahrensburg Ost	53.661347	10.242240	U Ahrensburg Ost	53.661227	10.242712	Metro Station
2	Ahrensburg West	53.664639	10.219403	Hansebäckerei Junge	53.663482	10.220533	Bakery
3	Ahrensburg West	53.664639	10.219403	Zum Griechen	53.664910	10.220690	Greek Restaurant
4	Ahrensburg West	53.664639	10.219403	U Ahrensburg West	53.664599	10.220055	Metro Station

Map of Hamburg showing Metro Station points was generated using Folium Library:



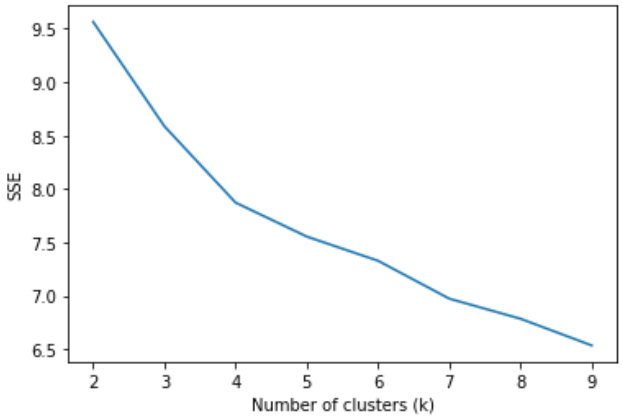
3. Methodology

3.1. Used Methodology

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.

3.2. K-Means

Runing K-means to cluster the neighborhood:

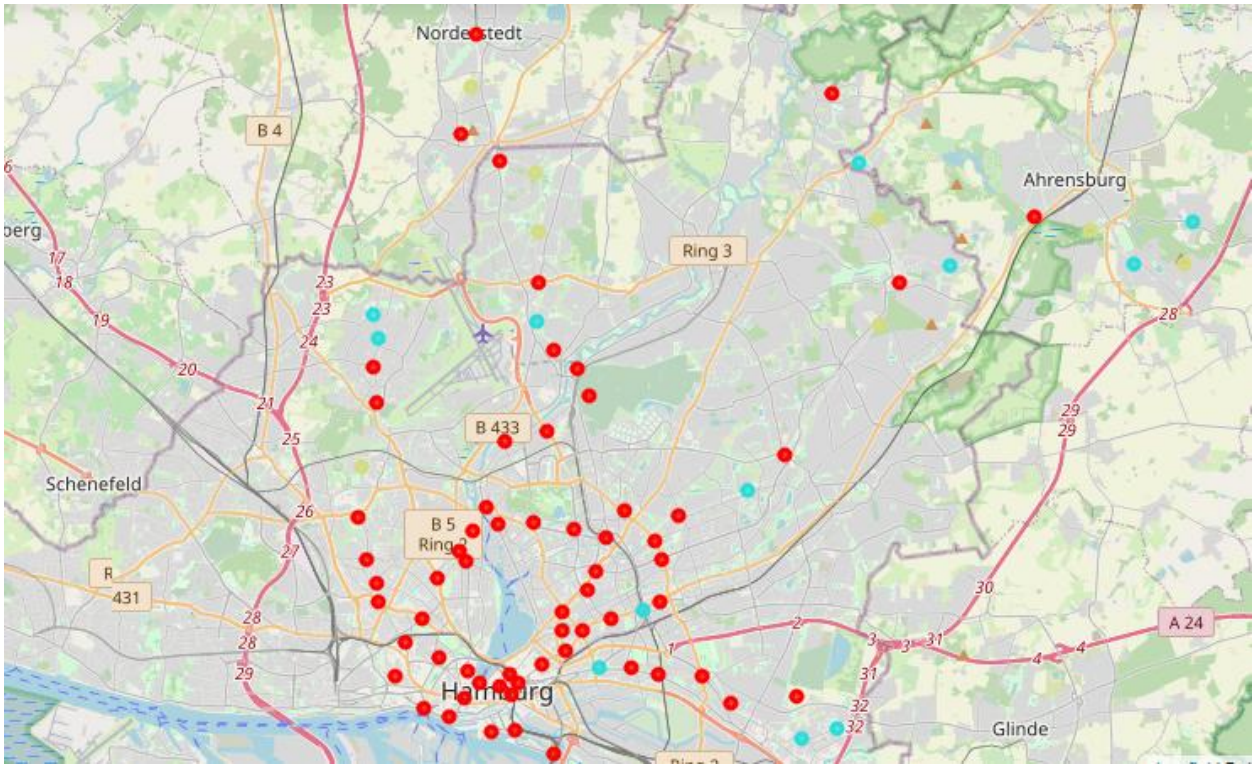


The elbow method applied to the Total Within Sum of Squares of the Silhouette indicates that the optimal number of clusters is $k = 4$.

Clusters were added to dataframe with 10 the most common venues. The most numerous is cluster number 1. The remaining clusters have less neighborhoods:

	Station	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	
87	Wandsbek Markt	53.571664	10.066741	1.0	Drugstore	Asian Restaurant	Bakery	Café	Italian Restaurant	German Restaurant	Clothing Store	Supermarket	
88	Wandsbeker Chaussee	53.569667	10.060021	3.0	Supermarket	Hotel	Middle Eastern Restaurant	Turkish Restaurant	Asian Restaurant	Drugstore	Bank	Gym / Fitness Center	R
89	Wartenau	53.564651	10.035408	1.0	Café	Bakery	Hotel	Pharmacy	Supermarket	Wine Shop	Cosmetics Shop	Comic Shop	
90	Überseequartier	53.540424	9.998431	1.0	Italian Restaurant	Hotel	Bakery	Coffee Shop	German Restaurant	Museum	Asian Restaurant	Bar	R
91	HafenCity Universität	53.540554	10.007963	1.0	Art Gallery	Hotel	Theater	Coffee Shop	Museum	Modern European Restaurant	Scenic Lookout	Food Court	G

On the map results of clustering looks like this:



At the end examination of clusters were made

- cluster no. 0:

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	53.661347	Arts & Entertainment	Hotel	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Flower Shop
11	53.664363	German Restaurant	Event Service	French Restaurant	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop	Flea Market
27	53.604340	Metro Station	Event Service	Electronics Store	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop	Flea Market
40	53.653023	Soccer Field	Italian Restaurant	Metro Station	Fountain	Forest	Food Court	Food & Drink Shop	Event Service	French Restaurant
41	53.674984	Pool	Greek Restaurant	Pub	Event Service	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Eastern European Restaurant
46	53.660976	Ice Cream Shop	Electronics Store	French Restaurant	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop	Flea Market
52	53.638644	Rock Climbing Spot	Electronics Store	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Fish & Chips Shop
68	53.695772	Metro Station	Hotel Bar	Electronics Store	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop	Flea Market

- cluster no.1:

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	53.664639	Bakery	Greek Restaurant	Fast Food Restaurant	Italian Restaurant	Trail	Food & Drink Shop	Flower Shop	Flea Market	Food Court
2	53.610541	Hotel	Pet Store	Asian Restaurant	Bank	Gym / Fitness Center	Salad Place	Café	Greek Restaurant	Supermarket
3	53.586202	Doner Restaurant	Metro Station	Supermarket	Thai Restaurant	Athletics & Sports	Bus Stop	Bakery	Flea Market	Fish & Chips Shop
4	53.587386	Bakery	Drugstore	Bar	Restaurant	Greek Restaurant	Café	Middle Eastern Restaurant	Italian Restaurant	German Restaurant
5	53.544072	Café	German Restaurant	Italian Restaurant	Seafood Restaurant	Hotel	Restaurant	Concert Hall	Hotel Bar	Salad Place
...	...	...	...	...	...	...	...	...	...	...
86	53.592326	Bus Stop	Convenience Store	Volleyball Court	Metro Station	Bank	Falafel Restaurant	Bakery	Event Space	Farmers Market
87	53.571664	Asian Restaurant	Bakery	Café	Italian Restaurant	German Restaurant	Clothing Store	Supermarket	Hotel	Turkish Restaurant
89	53.564651	Bakery	Hotel	Pharmacy	Supermarket	Wine Shop	Cosmetics Shop	Comic Shop	Nightclub	Chinese Restaurant
90	53.540424	Hotel	Bakery	Coffee Shop	German Restaurant	Museum	Asian Restaurant	Bar	Seafood Restaurant	Ice Cream Shop
						Modern				

- cluster no.2:

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	53.183301	Wine Shop	Event Service	French Restaurant	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop	Flea Market

- cluster no.3:

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
10	53.652958	Shipping Store	Supermarket	Greek Restaurant	Event Service	Fountain	Forest	Food Court	Food & Drink Shop	Flower Shop
12	53.555798	Bakery	Asian Restaurant	Metro Station	Café	Doner Restaurant	Laundromat	Park	Flower Shop	Flea Market
21	53.639252	Indian Restaurant	Supermarket	Breakfast Spot	Wine Shop	Flea Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
23	53.663556	Taverna	Italian Restaurant	Metro Station	Bakery	Ice Cream Shop	Pharmacy	Food Court	Food & Drink Shop	Flower Shop
34	53.677703	Pet Store	Metro Station	Snack Place	Food & Drink Shop	Café	Forest	Food Court	Flower Shop	Flea Market
47	54.529794	Pet Store	Drugstore	Auto Garage	Sandwich Place	Big Box Store	Fast Food Restaurant	Falafel Restaurant	Farmers Market	Flea Market
53	53.538636	General Entertainment	Park	Supermarket	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Wine Shop
58	53.528107	Middle Eastern Restaurant	Drugstore	Metro Station	Wine Shop	Event Space	Fountain	Forest	Food Court	Food & Drink Shop
60	53.640853	Convenience Store	Italian Restaurant	Drugstore	Bus Stop	Trail	Bakery	Flower Shop	Flea Market	Fish & Chips Shop
72	53.625206	Drugstore	Insurance Office	Hardware Store	Greek	Italian	Flea Market	Flower Shop	Fish & Chips	Electronics

4. Results

During the analysis, several important statistical features of the boroughs were explored and visualized. Because of not so big data set we have 4 clusters. We can see that some clusters are more common and some are unique.

We can see differences from areas closer to city center and other districts - in city center we can find bigger variety of different categories. However, all the clusters look similar to each other, indicating that Hamburg is quite homogeneous in terms of available venue categories.

5. Discussion

During analysis of project data I had observation which can be base for futher development.

Like in lots of analytical problems we have to made some assumptions and simplifications to understand and present solution in a better way. Analysis was performed on limited data. If more data would be available our results could be more detail. If data about prices in different neighborhoods would be provided we could already make some recommendation. For now we can just assume that centrum of the city is more expensive because of a lot of different businesses around. In center of the city we have more choice and variety of venues.

Forsquere is a good source of data, but to perform more detail analysis we need premium account.

6. Conclusion

Although all of the goals of the project were met there is still a huge potential to improve this research. As a next step it is possible to develop recommendation system which will help people to get accommodation they dream about.