IBM Capstone Project

March 9, 2021

1. Problem Description

Our customer signed a long-term contract as a media manager with a successful football club in Hamburg and wants to move with his family close to the job location (max. 7 km distance). He hasn't been happy with the 2 houses his real-estate agent had shown him before because he did not put much emphasis on the surrounding area of the house. But as a family person, he wants to make sure that there are good venues for his children to play and parks to walk the dog. Therefore, we have been given the order to make a preselection for him.

Hamburg is the second-largest city in Germany and has 5.3 million inhabitants. Many well known industrial and media companies such as Beiersdorf, Der Spiegel and Die Zeit have their HQ's there and it is home to the famous football clubs Hamburger SV and FC St. Pauli. Hamburg has a good infrastructure, it has much to offer and is really diverse in its venues. All this leads to relatively high rental prices.

Our aim is to cluster the relevant quarters of Hamburg, find a cluster which stands for family friendliness with parks and playgrounds and select the quarter which has the lowest rental prices in the family-friendly cluster.

2. Data

As there are many components which we take into consideration to make a reliable selection, we also need to use multiple datasets.

First, we need a dataset with the different quarters in Hamburg. These information we scrape from Wikipedia[1], turn it into pandas data frame and make a query to get the

[1] https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Stadtteile_Hamburgs.

latitude and longitude data by GeoPy, a Python client for geocoding web services. We assign the coordinates to the data frame with quarters.

On top of that, we select only the quarters that are within a 7 km radius for the job location. We do that, by again using GeoPy and give it also the coordinates for the job location, which we also get from Wikipedia. We drop the quarters which are not inside the radius.

Next, the rental prices[2] will be scraped, modified and merged with the other dataset. The rental price is important in order to know which quarter in the family-friendly cluster has the lowest rental prices.

	Quarter	Borough	Latitude	Longitude	Price per m^2 in EURO
0	Neustadt	Hamburg-Mitte	53.549881	9.979048	10,69
1	Finkenwerder	Hamburg-Mitte	53.530882	9.858523	7,85
2	Altona-Altstadt	Altona	53.549660	9.945352	10,60
3	Sternschanze	Altona	53.561768	9.963282	11,73
4	Altona-Nord	Altona	53.561400	9.944720	10,61

We use the Foursquare API to get the different venues of the quarters (within a radius of 5 km). Then we categorize them and assign the 10 most common venue categories to the respective quarter in order to have a reasonable data basis for the k-means algorithm which we will use to make the clusters. For the visualization, we merge the latter dataset with the dataset that contains the quarters with the latitude and longitude data.

The recommendation for our customer: The quarter with the lowest rental prices in the family-friendly cluster.

[2] https://www.hamburgportal.de/immobilien/mietwohnungen/mietenspiegel/.

3. Methodology

The original data set has 104 entries (=104 quarters) and 8 columns. We dropped 4 columns and and kept 3 columns (quarter, borough and population density). The reason to keep population density will be mentioned later.

In total, we have 104 quarter which distribute across the 7 boroughs as follows:

```
Hamburg-Mitte    0.182692
Wandsbek         0.173077
Harburg          0.163462
Bergedorf        0.134615
Altona           0.134615
Hamburg-Nord     0.125000
Eimsbüttel       0.086538
```

Hamburg-Mitte has the most boroughs (~18%) and Eimsbüttel is way below the average. The remaining borough are close to each other in term of quarters affiliation.

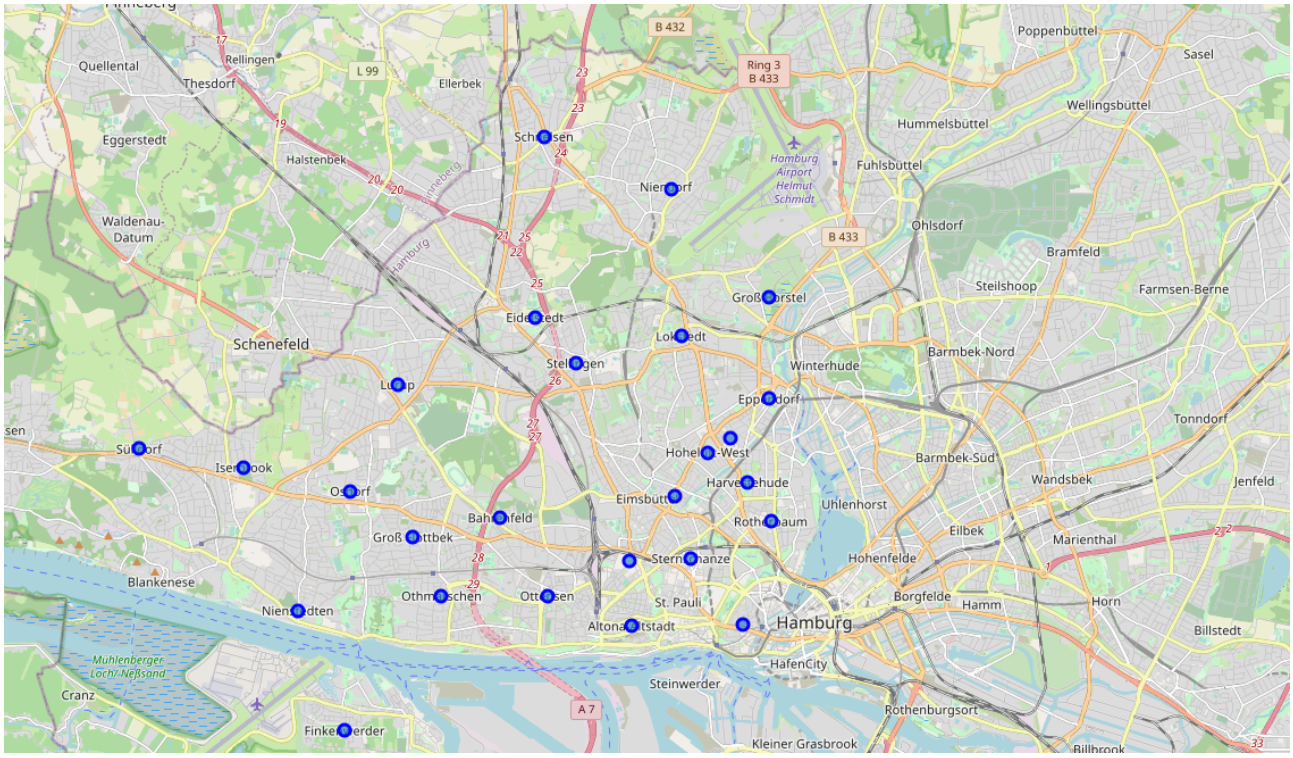
Next, we retrieved the coordinates for the quarters and what catched our eyes was that the quarter „Neuwerk“ completely dropped off the line when comparing its longitude coordinates to those of the others:

14	Kleiner Grasbrook	Hamburg-Mitte	246	53.528549	9.997694
15	Steinwerder	Hamburg-Mitte	4	53.539324	9.961541
16	Waltershof	Hamburg-Mitte	0	53.519654	9.910932
17	Finkenwerder	Hamburg-Mitte	597	53.530882	9.858523
18	Neuwerk	Hamburg-Mitte	4	53.922422	8.502859
19	Altona-Altstadt	Altona	10418	53.549660	9.945352
20	Sternschanze	Altona	16184	53.561768	9.963282
21	Altona-Nord	Altona	11153	53.561400	9.944720
22	Ottensen	Altona	12709	53.555066	9.919819

This led to the assumption that is is not located in Hamburg. A quick search on Wikipedia shows why. Neuwerk is an enclave of Hamburg and is approx. 100 km apart from the city borders. We don't need to drop the entry now, because we will make further selections which will lead the dropout anyway.

After retrieving the coordinates of the job location, we dropped all quarters that are not within a 7 km radius which made up approx. 73% of the quarters. Thus, we went by approx. 27% of the quarters.

On the map, you can see the quarters which are mostly inside the north and west axis of Hamburg.



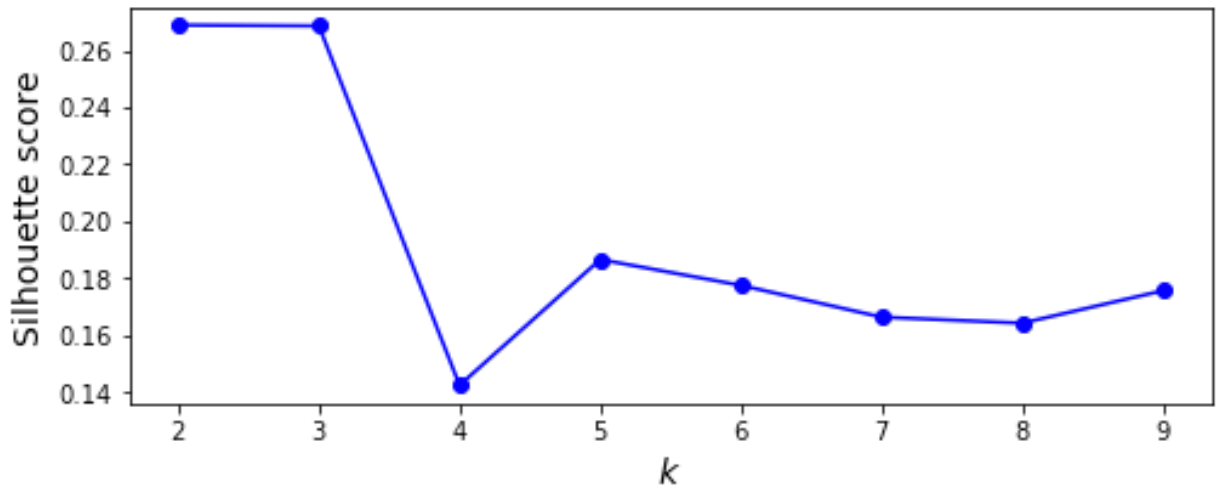
From the Foursquare data we categorized the 15th most common venues of each quarter in order to be able to cluster the quarters. Here you can see and extract:

	Quarter	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Altona-Altstadt	Café	Park	Coffee Shop	Bakery	Seafood Restaurant	Pizza Place
1	Altona-Nord	Café	Bakery	Coffee Shop	Park	Seafood Restaurant	Pizza Place
2	Bahrenfeld	Café	Bakery	Park	Coffee Shop	Farmers Market	Ice Cream Shop
3	Eidelstedt	Café	Supermarket	Bakery	Zoo Exhibit	Park	Ice Cream Shop

In this project we have unlabeled data therefore need an algorithm from the unsupervised learning section of machine learning. We chose the kmeans algorithm due its reliability and ease of use.

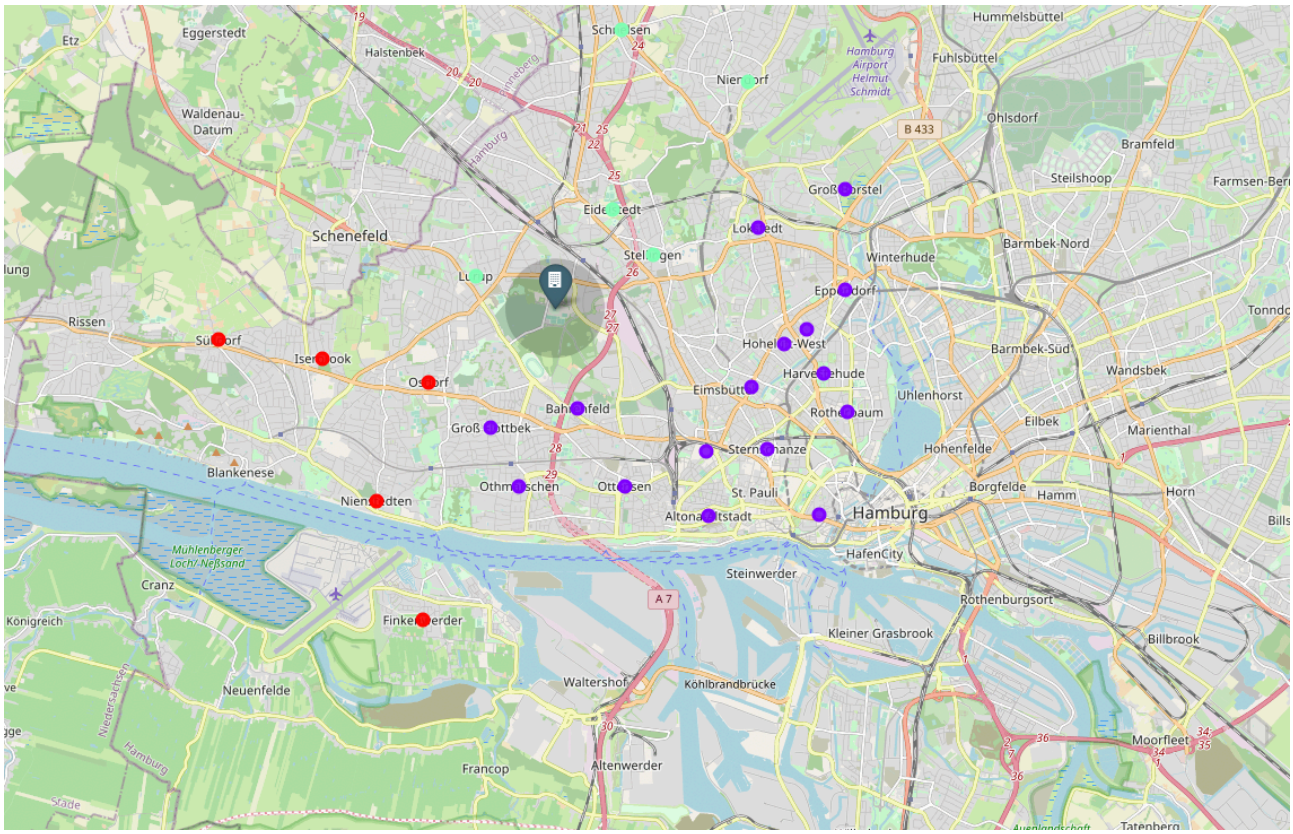
Before running the kmeans algorithm, we calculated the silhouette scoring which tells us which k (number of clusters) to pick for the algorithm.

So the decision is between k=2 and k=3. We went for k=3, hoping to get 3 clusters that strongly differentiate from each other and to get one clean cluster with many family-friendly venues.



4. Results

The result - as you can see on the map below - happen to be clustered geographically which leads to conclusion that neighboring quarters are supposedly alike. Now, interesting will be which cluster is now the one we were looking for.



Cluster 1 (red dots on the map) seems not to be the cluster we were looking for. It has one quarter which could be interesting (i.e. Osdorf) but the other quarters don't show a clear tendency towards parks and/or playgrounds.

	Quarter	Price per m^2 in EURO	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
1	Finkenwerder	7,85	0	Café	Ice Cream Shop	Seafood Restaurant	Restaurant	Park	Hotel	Supermarket	German Restaurant	Beach	Art Museum
10	Osdorf	9,82	0	Park	Bakery	Café	Supermarket	Seafood Restaurant	Ice Cream Shop	Beach	Hotel	German Restaurant	Coffee Shop
11	Nienstedten	11,93	0	Supermarket	Café	Seafood Restaurant	German Restaurant	Park	Ice Cream Shop	Beach	Bakery	Restaurant	Pool
12	Iserbrook	9,34	0	Supermarket	Restaurant	Café	Seafood Restaurant	Hotel	Ice Cream Shop	Beach	Clothing Store	Bakery	Shopping Mall
13	Sülldorf	9,62	0	Supermarket	Café	Ice Cream Shop	Drugstore	Seafood Restaurant	Clothing Store	German Restaurant	Restaurant	Shopping Mall	Beach

Cluster 2 is exactly what we were looking for. We almost have in every quarter a park either as the most common venue or as the 2nd most common. Playgrounds seem to be underrepresented across all clusters. Here, we still have a clear tendency towards our customer's expectations and wishes. The last task was to determine the quarter with the lowest rental prices. A quick look at table reveals: Groß Borstel is the way to go

	Quarter	Price per m^2 in EURO	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
0	Neustadt	10,69	1	Park	Café	Coffee Shop	Bakery	Seafood Restaurant	Plaza	Pizza Place	Ice Cream Shop	Pub	Nightclub
2	Altona-Altstadt	10,60	1	Café	Park	Coffee Shop	Bakery	Seafood Restaurant	Pizza Place	Ice Cream Shop	Hotel	Nightclub	Scenic Lookout
3	Sternschanze	11,73	1	Café	Bakery	Coffee Shop	Plaza	Seafood Restaurant	Pizza Place	Park	Wine Bar	Falafel Restaurant	Austrian Restaurant
4	Altona-Nord	10,61	1	Café	Bakery	Coffee Shop	Park	Seafood Restaurant	Pizza Place	Ice Cream Shop	Nightclub	Bistro	Supermarket
5	Ottensen	11,60	1	Café	Park	Bakery	Coffee Shop	Pizza Place	Ice Cream Shop	Seafood Restaurant	Supermarket	Beach Bar	Nightclub
6	Bahrenfeld	10,09	1	Café	Bakery	Park	Coffee Shop	Farmers Market	Ice Cream Shop	Seafood Restaurant	Supermarket	Pizza Place	German Restaurant
7	Groß Flottbek	11,38	1	Park	Bakery	Café	Seafood Restaurant	Ice Cream Shop	Supermarket	Falafel Restaurant	Restaurant	Harbor / Marina	Shopping Mall
8	Othmarschen	11,81	1	Park	Bakery	Café	Ice Cream Shop	Seafood Restaurant	Farmers Market	Wine Bar	Pizza Place	Supermarket	Bistro
14	Eimsbüttel	11,36	1	Bakery	Café	Coffee Shop	Park	Pizza Place	Wine Bar	Ice Cream Shop	French Restaurant	Plaza	Hotel
15	Rotherbaum	12,41	1	Coffee Shop	Café	Ice Cream Shop	Wine Bar	Restaurant	Park	Cocktail Bar	Plaza	Hotel	Farmers Market
16	Harvestehude	12,72	1	Café	Bakery	Coffee Shop	Park	Plaza	Ice Cream Shop	Restaurant	Wine Bar	Farmers Market	Seafood Restaurant
17	Hoheluft-West	11,68	1	Café	Bakery	Coffee Shop	Park	Wine Bar	Plaza	Ice Cream Shop	Farmers Market	German Restaurant	Supermarket
18	Lokstedt	10,28	1	Café	Park	Bakery	Supermarket	Zoo Exhibit	Coffee Shop	Wine Bar	Farmers Market	Hotel	Beer Store
23	Hoheluft-Ost	12,28	1	Café	Park	Bakery	Coffee Shop	Wine Bar	Farmers Market	Supermarket	Ice Cream Shop	French Restaurant	Falafel Restaurant
24	Eppendorf	11,83	1	Café	Park	Bakery	Coffee Shop	Wine Bar	Farmers Market	Supermarket	Ice Cream Shop	Tapas Restaurant	Bistro
25	Groß Borstel	9,59	1	Café	Park	Supermarket	Bakery	Coffee Shop	Italian Restaurant	Ice Cream Shop	Wine Bar	Hotel	Bistro

Cluster 3 looks like cluster 1 with supermarkets on top of the ranking. Therefore, k=2 would be sufficient.

	Quarter	Price per m^2 in EURO	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
9	Lurup	8,40	2	Supermarket	Bakery	Zoo Exhibit	Gym / Fitness Center	Drugstore	Café	Hotel	Park	Clothing Store	Shopping Mall
19	Niendorf	9,55	2	Supermarket	Hotel	Bakery	Café	Greek Restaurant	Park	Zoo Exhibit	Airport Service	Ice Cream Shop	Pool
20	Schnelsen	8,88	2	Zoo Exhibit	Supermarket	Ice Cream Shop	Furniture / Home Store	Indoor Play Area	Bakery	German Restaurant	Italian Restaurant	Hotel	Greek Restaurant
21	Eidelstedt	8,92	2	Café	Supermarket	Bakery	Zoo Exhibit	Park	Ice Cream Shop	German Restaurant	Italian Restaurant	Gym / Fitness Center	Wine Bar
22	Stellingen	9,81	2	Café	Bakery	Zoo Exhibit	Supermarket	Italian Restaurant	Ice Cream Shop	German Restaurant	French Restaurant	Coffee Shop	Park

5. Discussion

We can be satisfied with the result, however it would be interesting to compare the results with other clustering algorithms. That is an idea for further analysis. Also the amount of categories (15) could be reduced in order to put the focus on e.g. 3 or 4 venue categories. That way, you could eventually make better distinctions and the algorithm would probably lead to even better results. However, the way we prepared the data, made a preselection of quarters that are within a 7 km radius, and also took the rental prices into consideration made it able for us to make a solid recommendation.

6. Conclusion

The result is quite good. We were able to determine a cluster which shows clear patterns of family friendliness and the customer is happy to tell his real estate agent that he should focus in his search only on Groß Borstel. That way, he can save his time and make sure the next house will not be discarded because of the location.