Applied Data Science Capstone by IBM/Coursera

“The Battle of Hamburg’s Districts for a new Study Café“

Capstone project by Karima Chakroun
March 9, 2019

Table of contents

1. Introduction

1.1. Background

Hamburg is the second-largest city in Germany with more than 1.8 million inhabitants and one of the major science, research, and education metropoles in Europe [1,2]. The city has several public and private universities, offering a diverse range of excellent degree programs, with more than 100 000 students in total [2,3]. Notably, the number of Hamburg's students rises every year [3-6] and will continue to grow, since projects like the "Hamburg excellence strategy" [8,9] and the "Science City Bahrenfeld" [10-12] will draw even more young people from all over the world to Hamburg [1]. However, the city's university buildings and campuses offer only very limited space and are often overcrowded, old, and decrepit [7]. Hence, they are not a very attractive place for students to pass the time, learn, or socialize inbetween and after courses. With more students to come in the following years, there is clearly a need for more student venues like cafés and bars nearby these universities. For example, a new venue type called "study café" might be of particular interest for most students, which could offer free WLAN/electricity, low prices for drinks & foods, long opening hours, and enough space/tables for students to work with books and laptops alone or in groups. However, such off-campus study cafés are yet entirely lacking in Hamburg. Private or public investors should size this opportunity and initiate the opening of such new types of student venues.

1.2. Business problem

Before opening a new study café in Hamburg, investors are faced with the question where to best open such a place. To approach this problem, several aspects should be considered:

a) **Where are the universities located?** Since the new venue should be located nearby one or more universities.

b) **What are the characteristics of different districts?** Since the new venue should be located in a lively district with other food & drink venues, but not with too many cafés yet (< 5).

c) **How high are the rental prices for different districts?** Since the new venue should be located in a district with low to moderate rental prices (< 15 €/m²) in order to be able to offer low-priced food & drinks to students.

The present project addresses this problem by using different data about Hamburg (e.g. rental prices, locations data of universities, venues, and districts) in order to aid investors in finding the optimal location for opening a new study café.

1.3. Interested parties

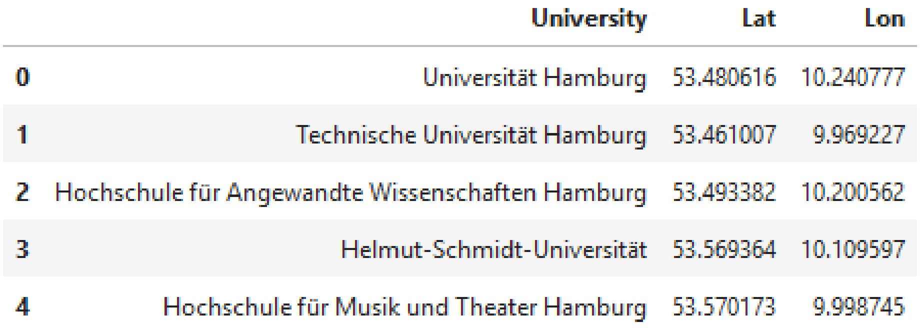
This report will be of particular interest to any investors, private or public, who might want to open a new student venue in Hamburg and are yet undecided about the best location for such a place. Such investors could even include the universities themselves, which might decide to rent nearby off-campus places to offer more space for their students. Especially non-local investors who are not familiar with the city of Hamburg might profit from this report, as it will give them a nice visual overview of Hamburg's districts and their relevant characteristics.

2. Data acquisition and cleaning

To solve the problem, datasets from several sources were combined to answer the three above questions.

2.1. Where are the universities located?

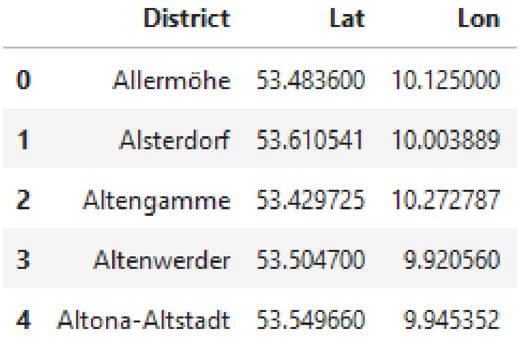
To obtain data on this question, a list of Hamburg's main universities was scraped from the website https://www.4icu.org/de/hamburg/, containing 18 name entries. Based on these names, the geopy client (https://pypi.org/project/geopy/) was then used with the Nominatim geolocator service to request the geographical coordinates (latitude and longitude) of each university. Since the request returned no result for 6 of the 18 universities, the latitude and longitude of these 6 universities were searched and added manually using google maps (https://maps.google.de/). All data were combined into a dataframe including the name, latitude, and longitude for each of the 18 universities (shown here are the first 5 rows):



	University	Lat	Lon
0	Universität Hamburg	53.480616	10.240777
1	Technische Universität Hamburg	53.461007	9.969227
2	Hochschule für Angewandte Wissenschaften Hamburg	53.493382	10.200562
3	Helmut-Schmidt-Universität	53.569364	10.109597
4	Hochschule für Musik und Theater Hamburg	53.570173	9.998745

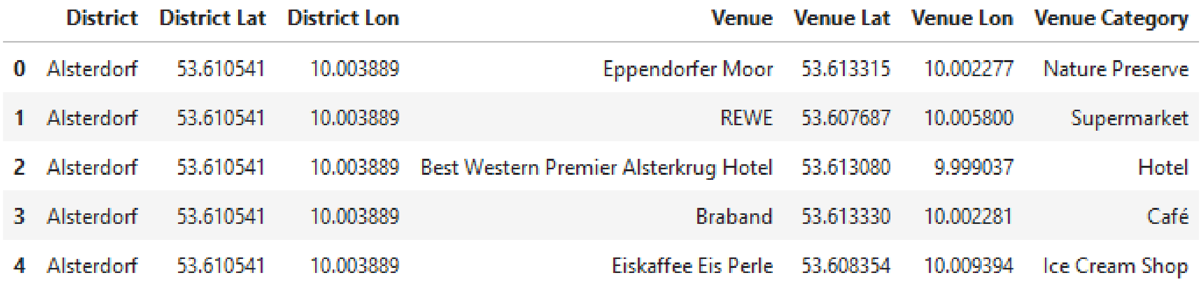
2.2. What are the characteristics of different districts?

To obtain data on this question, a geojson file of Hamburg's districts was first downloaded from https://gist.github.com/webtobesocial/935759ba975ffd9f6df6d1059fe5ad82/raw, containing the names and geographical borders of 104 districts. On one hand, this geojson file was used to create a choropleth map of rental prices in Hamburg's districts (see 2.3.). On the other hand, the data in this file were also used to obtain a list of Hamburg's district names. Based on these names, the geopy client was used with the Nominatim service to request the central geographical coordinates (latitude and longitude) of each district. These data were combined into a dataframe containing the name, latitude, and longitude of Hamburg's 104 districts (shown here are the first 5 rows):



	District	Lat	Lon
0	Allermöhe	53.483600	10.125000
1	Alsterdorf	53.610541	10.003889
2	Altengamme	53.429725	10.272787
3	Altenwerder	53.504700	9.920560
4	Altona-Altstadt	53.549660	9.945352

Next, a list of venues for each district was obtained via the Foursquare API, using the "explore" endpoint with a limit of 100 and a radius of 500 meter around a district's given latitude and longitude. The returned information were combined with the district data into a dataframe showing for each venue the name, latitude, longitude, and venue category (shown here are the first 5 rows):

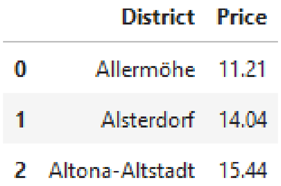


	District	District Lat	District Lon	Venue	Venue Lat	Venue Lon	Venue Category
0	Alsterdorf	53.610541	10.003889	Eppendorfer Moor	53.613315	10.002277	Nature Preserve
1	Alsterdorf	53.610541	10.003889	REWE	53.607687	10.005800	Supermarket
2	Alsterdorf	53.610541	10.003889	Best Western Premier Alsterkrug Hotel	53.613080	9.999037	Hotel
3	Alsterdorf	53.610541	10.003889	Braband	53.613330	10.002281	Café
4	Alsterdorf	53.610541	10.003889	Eiskaffee Eis Perle	53.608354	10.009394	Ice Cream Shop

These venue data were then used in a k-means clustering analysis to cluster Hamburg's districts based on their venue characteristics (see methodology section).

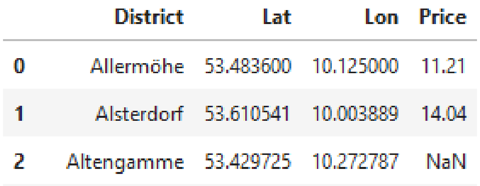
2.3. How high are the rental prices for different districts?

To obtain data on this question, a list of rental prices for Hamurg's districts was scraped from the website https://www.4icu.org/de/hamburg/, containing the names and the average rental prices (€/m²) of 86 districts for the year 2019. Rental prices were in a string format (e.g. "11,21 €/m²") and were converted to floats (e.g. 11.21) for further analysis. Shown here are the first 3 rows:



	District	Price
0	Allermöhe	11.21
1	Alsterdorf	14.04
2	Altona-Altstadt	15.44

To integrate these data with the district's location and venue data from above, it was checked if the spelling of all 86 districts included in this dataframe was identical to the districts' spelling in the above dataframe, i.e. the one based on the geojson file. Three district names needed to be changed to match the spelling in the above dataframe ("Hamb.-Altstadt" to "Hamburg-Altstadt", "St. Georg" to "St.Georg", and "St. Pauli" to "St.Pauli"). Furthermore, the three subdistricts "Hamm-Nord", "Hamm-Süd", and "Hamm-Mitte" needed to be combined to the district "Hamm" and the rental price for this district was calculated as the average rental price of the three combined subdistricts. After these changes, both the districts' location and rental price data were combined into a dataframe containing the name, latitude, longitude, and rental price (if available) of the 104 districts (shown here are the first 3 rows):



	District	Lat	Lon	Price
0	Allermöhe	53.483600	10.125000	11.21
1	Alsterdorf	53.610541	10.003889	14.04
2	Altengamme	53.429725	10.272787	NaN

These data were then used in combination with the geojson file from above to create a choropleth map of rental prices in Hamburg's districts (see methodology section).

3. Methodology

This project collected and combined several data about Hamburg in order to recommend good locations for opening a new study café to potential investors. In a first step, data from different sources were aquired and cleaned (see section 2), resulting in the following datasets:

- a dataframe including the name, latitude, and longitude of Hamburg's universities (18 rows)
- a dataframe including the name, latitude, longitude, and rental price (if available) of Hamburg's districts (104 rows)
- a dataframe including the name, latitude, longitude, and category of nearby venues for each of Hamburg's districts (1537 rows).

We will now use these data for further analysis, including data exploration (3.1), k-means clustering (3.2), and data visualization (3.2).

3.1. Data exploration

Hamburg's venue data were further explored before using them for clustering. First, it was checked how many unique venue categories were returned in total (=239). Also, the number of returned venues per district was plotted in a histogram (Figure 1) in order to get a first idea about the venue richness of Hamburg's districts.

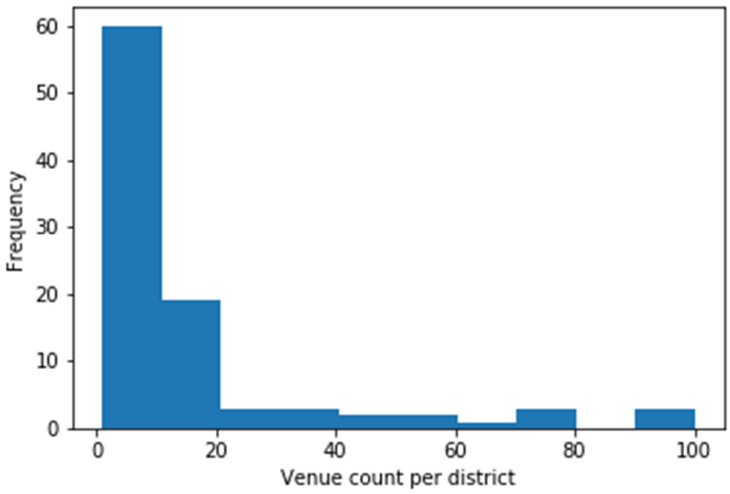
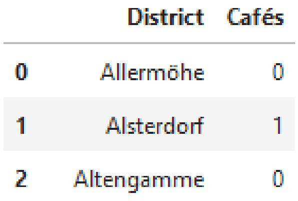


***Figure 1**. Histogram of the number of venues per district.*

Furthermore, the names of all venue categories were explored to identify the categories that describe coffee shops. These included "café" and "coffee shop", for which the combined total number per district was calculated and saved in a new dataframe for later data visualization (shown here are the first 3 rows):



	District	Cafés
0	Allermöhe	0
1	Alsterdorf	1
2	Altengamme	0

From the histogram in Figure 2, it is obvious that most districts have either no or only very few cafés yet. However, some districts seem to have already many cafés and would thus be less suitable for opening the new study café. For example, districts with ≥ 10 cafés include Hamburg-Altstadt, Hoheluft-Ost, Rotherbaum, St.Pauli, and Sternschanze.

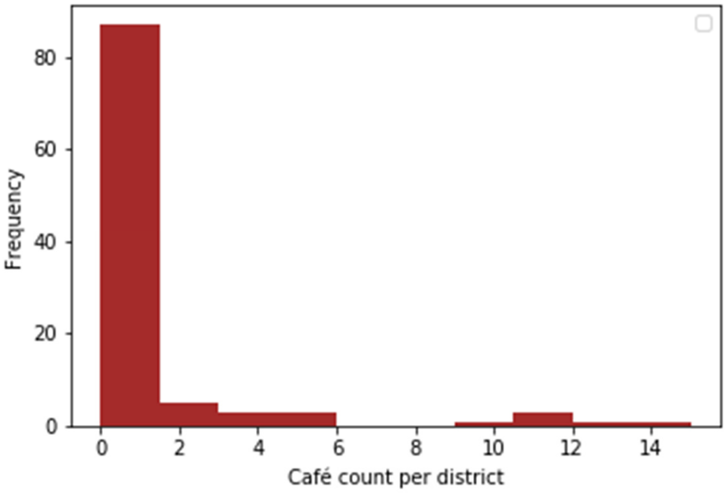


Figure 2. *Histogram of the number of cafés per district.*

Finally, a histogram of the rental prices is shown in Figure 3 to get an idea about the rental price distribution in Hamburg's districts. Some districts seem to have very high rental prices and would thus be less suitable for opening the new study café. For example, districts with an extremely high rental price (≥ 18 €/m²) are HafenCity, Hamburg-Altstadt, Hammerbrook, and Harvestehude.

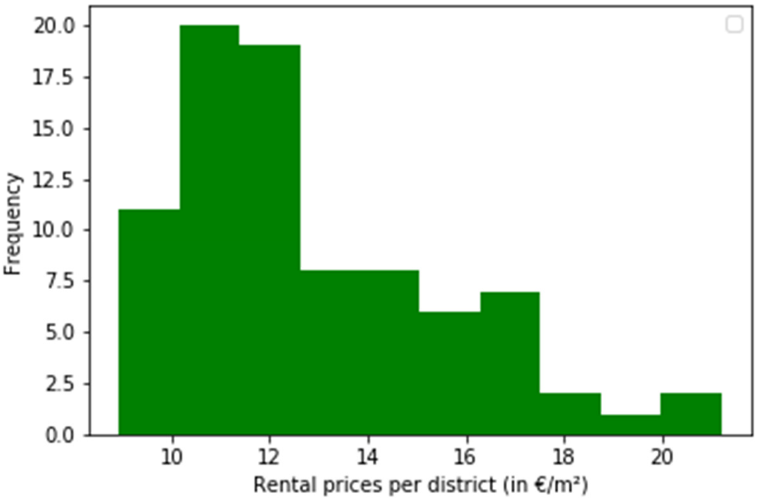
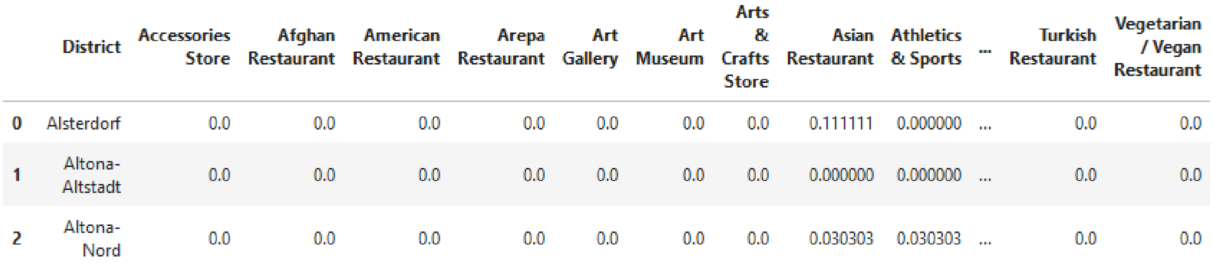


Figure 3. *Histogram of the district's rental prices.*

3.2. Clustering analysis of Hamburg's districts by venues

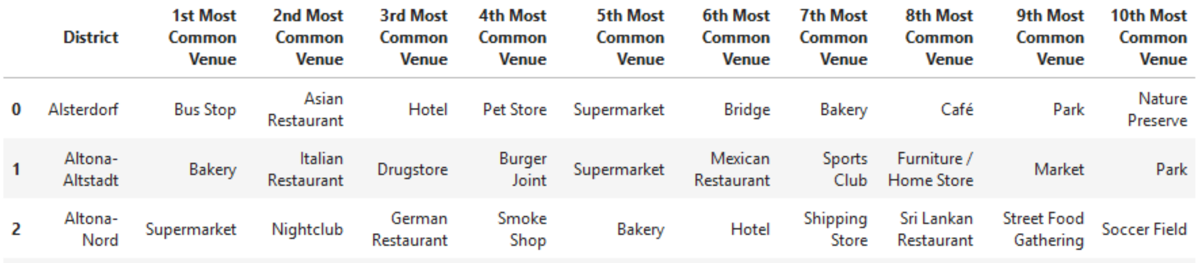
One problem that needed to be addressed by this project was to identify districts that are most suitable for opening a new study café based on their current venue characteristics. In particular, we were looking for lively districts in which several food & drink venues are already located and which might be attractive places for students to meet and socialize inbetween or after courses. To approach this problem, a k-means clustering algorithm was used on Hamurg's venue data. Specifically, we first used the venue data acquired from Foursquare (see above) to obtain for each district the mean frequencies of the venue categories, and then used this feature to cluster the districts via the k-means clustering algorithm.

To prepare the venue data for clustering, the mean frequencies of the different venue categories in each district were calculated and stored in a new dataframe (shown here are the first 3 rows and a subset of the columns):



	District	Accessories Store	Afghan Restaurant	American Restaurant	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Turkish Restaurant	Vegetarian / Vegan Restaurant
0	Alsterdorf	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.111111	0.000000	...	0.0	0.0
1	Altona-Altstadt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0
2	Altona-Nord	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.030303	0.030303	...	0.0	0.0

Based on these data, another dataframe was created that shows the top 10 venue categories for each district (shown here are the first 3 rows):



	District	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	Alsterdorf	Bus Stop	Asian Restaurant	Hotel	Pet Store	Supermarket	Bridge	Bakery	Café	Park	Nature Preserve
1	Altona-Altstadt	Bakery	Italian Restaurant	Drugstore	Burger Joint	Supermarket	Mexican Restaurant	Sports Club	Furniture / Home Store	Market	Park
2	Altona-Nord	Supermarket	Nightclub	German Restaurant	Smoke Shop	Bakery	Hotel	Shipping Store	Sri Lankan Restaurant	Street Food Gathering	Soccer Field

Next, the districts were clustered by their venue frequencies using the *k*-means clustering algorithm from the sklearn library (https://scikit-learn.org/). As the optimal *k* for clustering is unknown, the *k*-means clustering algorithm was first run with different values for *k* and an elbow plot (Figure 4) was created that shows the winthin-cluster sum of squares (i.e. inertia or distortion) for each value of *k*.

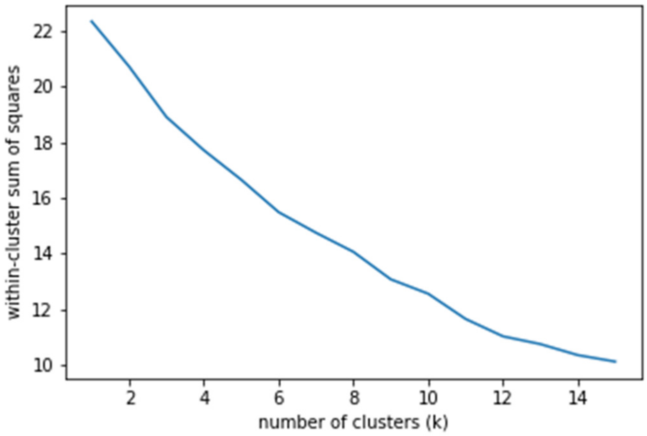
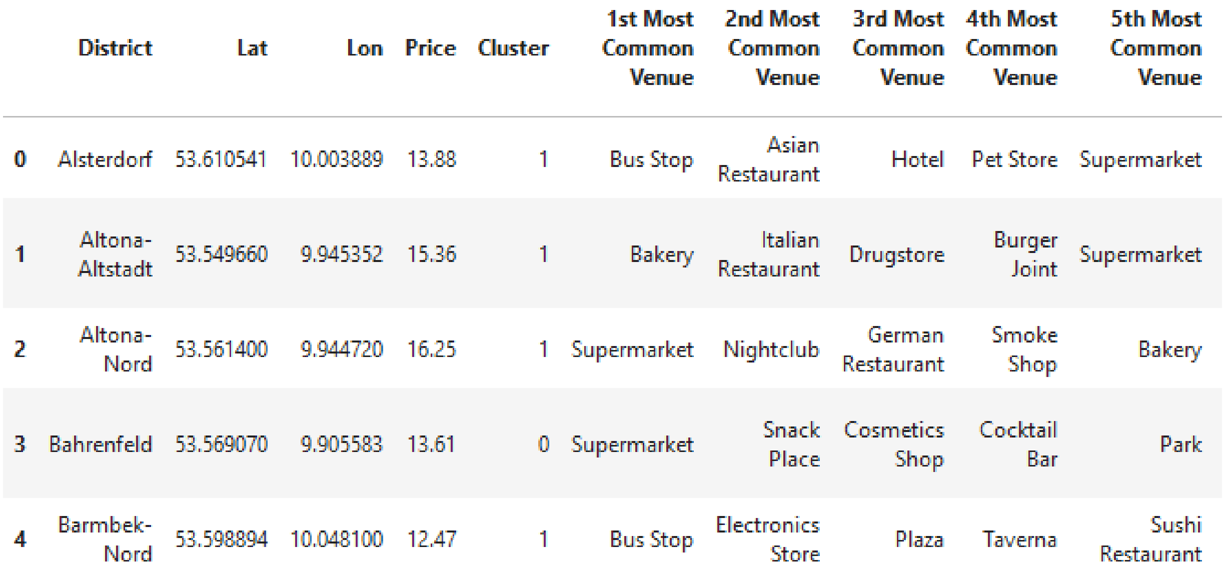


***Figure 4**. Elbow plot to identify the optimal k value for k-means clustering.*

As we see in Figure 4, there is no clear "elbow" that reflects the optimal value of *k* for clustering. Note that I also tried alternative methods to determine the optimal *k*, e.g. the silhouette coefficient [13] or the gap statistic [14], for which results are not shown here to keep the report short and clean. However, these alternative methods also yielded no clear answer about the optimal value for *k*. Hence, the value for clustering was set to ***k* = 5**, based on the following rationale: On the one hand, we want more than 1-2 clusters to better distinguish the large number of districts based on their different venue characteristics. Yet, we also do not want too many clusters, since we want to keep the resulting cluster structure easily understandable (e.g. we do not want many unique clusters that only contain 1-2 districts each).

Next, the resulting cluster labels were included in the above dataframe containing the top 10 venues for each district. Note that 8 of the 104 districts showed an NaN value as a cluster label, since Foursquare apparently returned no venues within the 500 m radius for these districts. Hence, the 8 disctricts with no returned venues and no cluster labels were dropped from this dataframe, as they are anyways not of interest for opening a study café. Note also that the district's rental price information (if available) was inserted as a new column into this dataframe (shown here are the first 5 rows and first 10 columns):



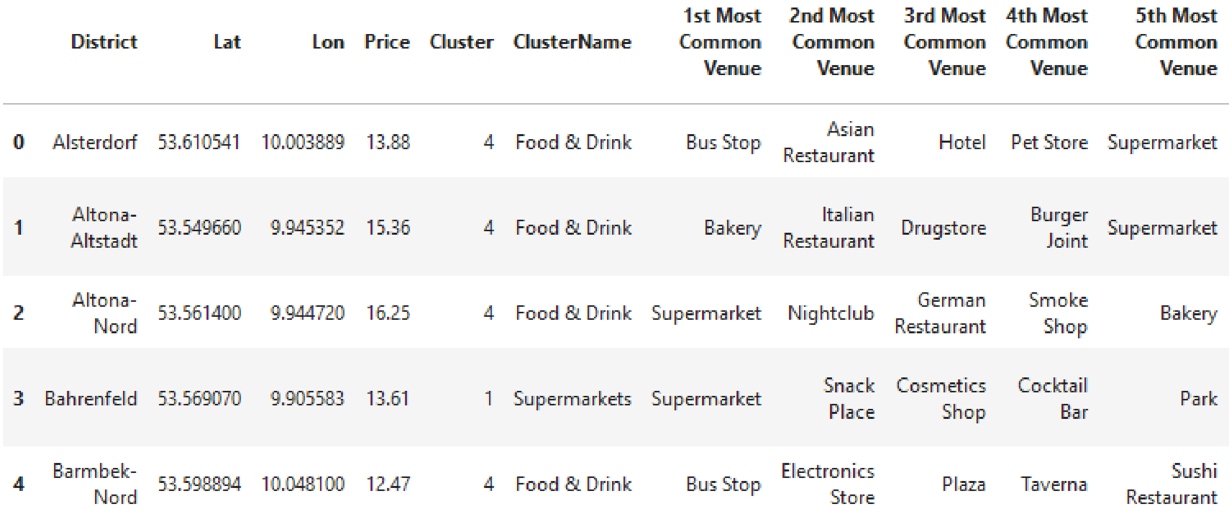
	District	Lat	Lon	Price	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alsterdorf	53.610541	10.003889	13.88	1	Bus Stop	Asian Restaurant	Hotel	Pet Store	Supermarket
1	Altona-Altstadt	53.549660	9.945352	15.36	1	Bakery	Italian Restaurant	Drugstore	Burger Joint	Supermarket
2	Altona-Nord	53.561400	9.944720	16.25	1	Supermarket	Nightclub	German Restaurant	Smoke Shop	Bakery
3	Bahrenfeld	53.569070	9.905583	13.61	0	Supermarket	Snack Place	Cosmetics Shop	Cocktail Bar	Park
4	Barmbek-Nord	53.598894	10.048100	12.47	1	Bus Stop	Electronics Store	Plaza	Taverna	Sushi Restaurant

After obtaining the different cluster labels, the resulting clusters were further examined to determine the discriminating venue categories that distinguish each cluster. Based on the 5 most common venue categories, a name was assigned to each cluster resulting in the following cluster names:

- **Cluster 0**: "Memorials" (1 district)
- **Cluster 1**: "Supermarkets" (24 districts)
- **Cluster 2**: "Outdoor & Recreation" (7 districts)
- **Cluster 3**: "German Restaurants & Zoo" (4 districts)
- **Cluster 4**: "Food & Drink" (60 districts)

Note that the cluster colors introduced here will also be used for later visualization (see 3.3) in order to better distinguish the five clusters on the map.

Finally, these cluster names were also added to the above data frame (shown here are the first 5 rows and first 11 columns):



	District	Lat	Lon	Price	Cluster	ClusterName	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alsterdorf	53.610541	10.003889	13.88	4	Food & Drink	Bus Stop	Asian Restaurant	Hotel	Pet Store	Supermarket
1	Altona-Altstadt	53.549660	9.945352	15.36	4	Food & Drink	Bakery	Italian Restaurant	Drugstore	Burger Joint	Supermarket
2	Altona-Nord	53.561400	9.944720	16.25	4	Food & Drink	Supermarket	Nightclub	German Restaurant	Smoke Shop	Bakery
3	Bahrenfeld	53.569070	9.905583	13.61	1	Supermarkets	Supermarket	Snack Place	Cosmetics Shop	Cocktail Bar	Park
4	Barmbek-Nord	53.598894	10.048100	12.47	4	Food & Drink	Bus Stop	Electronics Store	Plaza	Taverna	Sushi Restaurant

3.3. Data visualization using Folium

In the final step, an interactive map of Hamburg was created that combines all relevant data to solve the initial problem of finding an optimal location for the new study café. The map was created using the Folium library (https://pypi.org/project/folium/) and included the following data:

- The locations of Hamburg's universities, each marked by a small yellow circle on the map with a popup label showing the university's name.
- The locations of Hamburg's districts, each marked by a larger circle on the map with a popup label showing the following information for each district: the district's name, its cluster label (0-4) and cluster name, its exact rental price (if available, otherwise NaN), and the number of cafés already located in that district.
- The available rental prices were additionally used to build a choropleth map on which the above markers were superimposed, in order to get a visual impression on where rental prices are low to moderate (appropriate for the study café) and where they are very high (inappropriate for opening the study café).

As this map represents the main result(s) of this project, a figure of this map will be presented in the following results section.

4. Results

In this project, several data about Hamburg (e.g. rental prices, location data of districts, universities, and venues) were combined and visualized in an interactive Folium map in order to find good locations for opening a new study café. Figure 5 gives a (static) impression of the resulting Folium map. To briefly summarize (see 3.3 for details): This map consists of a choropleth map of the district's rental prices, on which two types of markers are superimposed: (1) small yellow cirles showing Hamburg's universities, and (2) larger circles in several colors showing Hamburg's districts, whereby the color reflects the district's cluster label. The district's markers also hold information about the cluster name, the exact rental price (if available), and the number of cafés already located in that district.

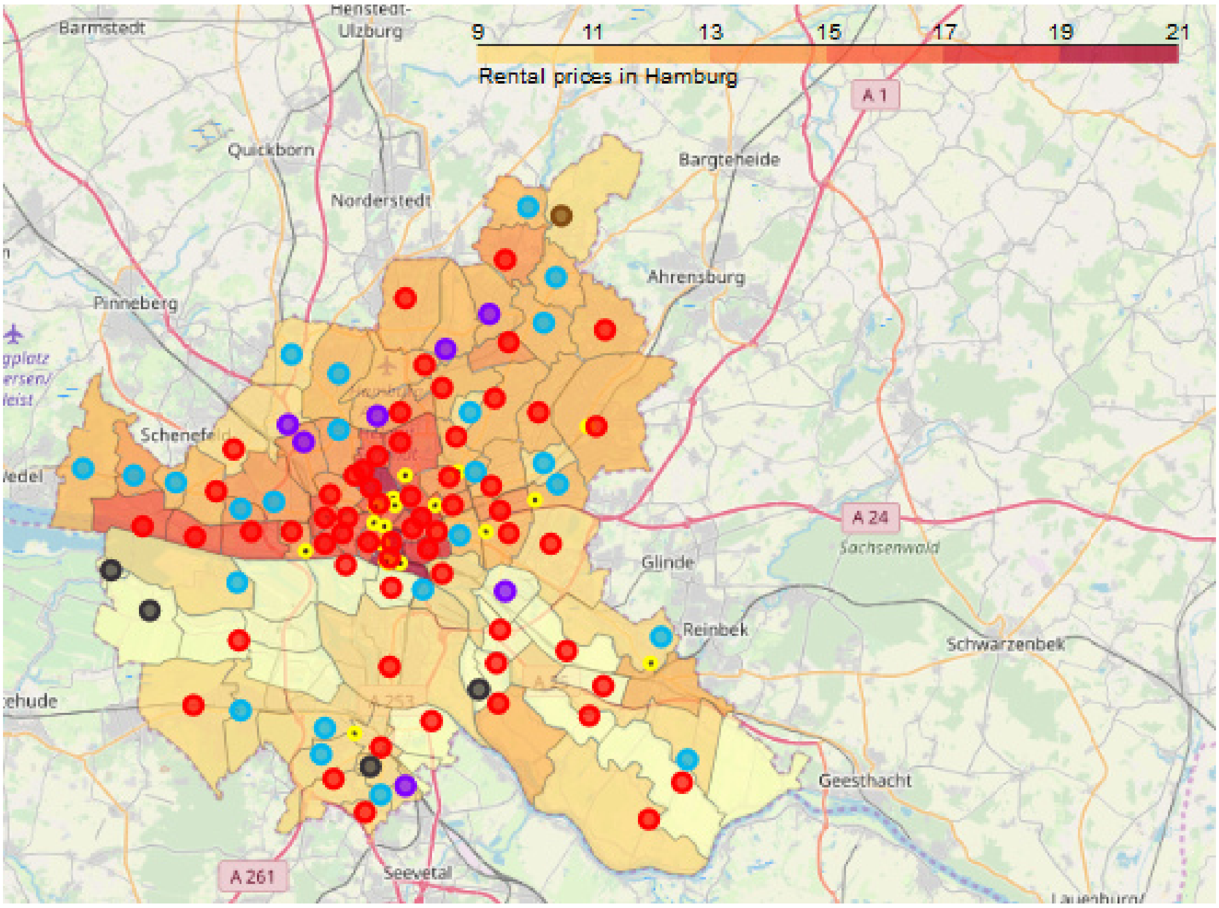


Figure 5*: Folium map of Hamburg. The map shows Hamburg's universities (small yellow circles) and districts (large circles) along with the district's rental price (choropleth map). Note that different venue characteristics (clusters) are marked by different colors. See text for further details.*

From the map in Figure 5, we can already make three important observations: First, most universities are located in Hamburg's center, especially in the areas north of the river Elbe. Second, we can see some pattern in the resulting clusters: All central districts belong to the **Cluster 4 "Food & Drink"** (in red), while the districts located with some distance to the center mostly belong to the **Cluster 1 "Supermarkets"** (in blue) or **Cluster 2 "Outdoor & Recreation"** (in violet). The other two clusters, i.e. **Cluster 0 "Memorials"** (in brown) and **Cluster 3 "German restaurants & Zoo"** (in grey), only contain district in Hamburg's periphery. Third, we can see that rental prices are highest in the central districts, while they are lower in Hamburg's periphery, especially in the districts south of the river Elbe. Interestingly, all these observations fit perfectly to my own experience, having lived in Hamburg for more than 15 years now. As most universities and "Food & Drink" venues are located in the central districts, we can zoom in and only show the central part of the map in Figure 6.

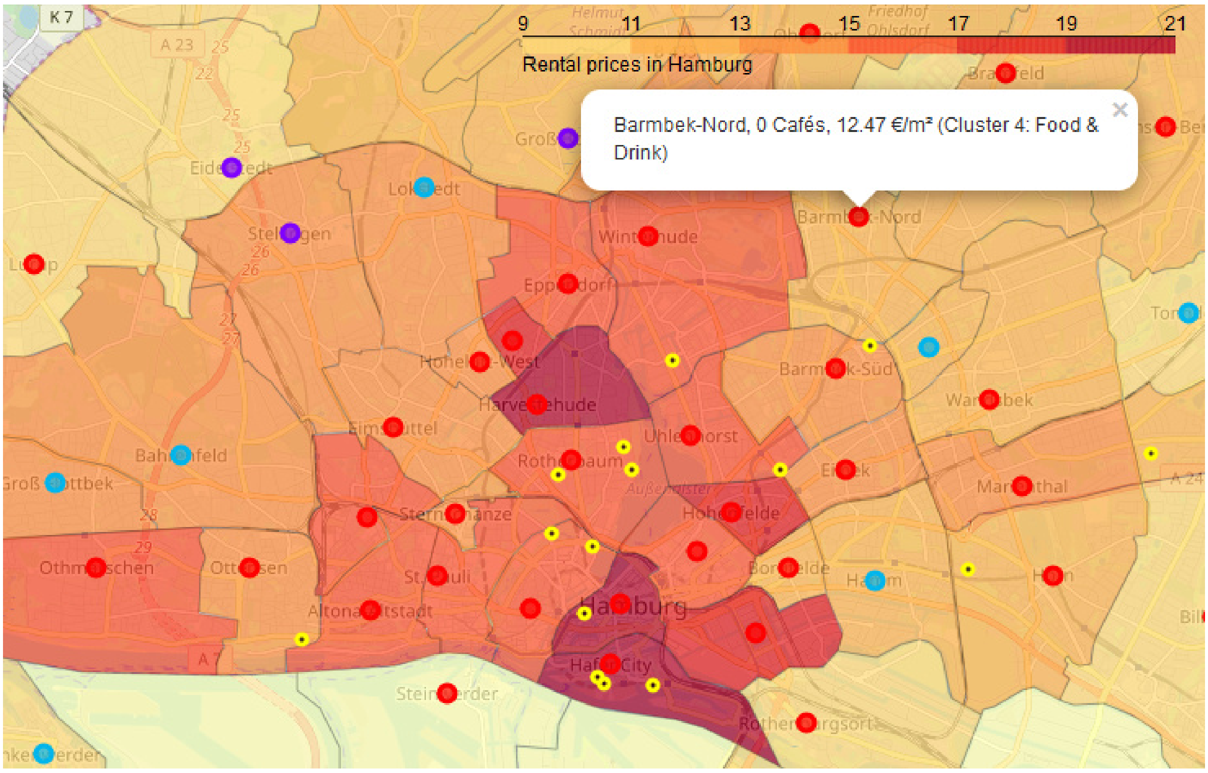


***Figure 6**: Folium map of central Hamburg. The map shows Hamburg's universities (small yellow circles) and districts (large circles) along with the district's rental price (choropleth map). Note that different venue characteristics (clusters) are marked by different colors. For demonstration, one of the markers is shown with its popup label. See text for further details.*

In the next section, we will come back to the business problem and use these results to address the initial question of where to best open the new study café.

5. Discussion

This project addressed the problem of finding the optimal district for opening a new study café in Hamburg. Specifically, we were looking for a district with the following characteristics:

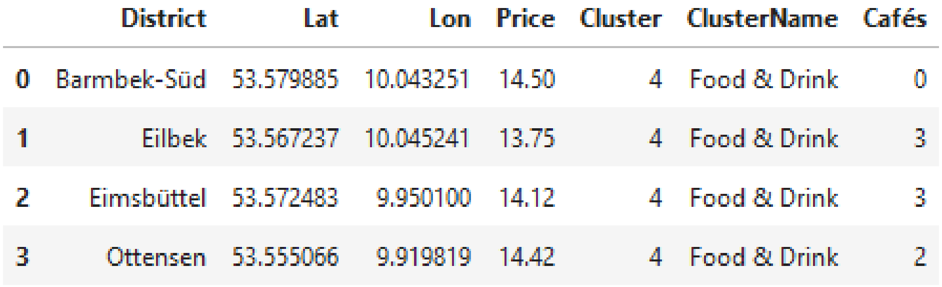
a) The study café should be located nearby one or more universities.

b) The study café should be located in a lively district with other food & drink venues, but not with too many cafés yet (< 5).

c) The study café should be located in a district with low to moderate rental prices (< 15 €/m^2) in order to be able to offer low-priced food & drinks to students.

By zooming in on the map and exploring the central districts (see Figure 4), we can identify some districts that fullfill the above criteria: **Barmbek-Süd**, **Eilbek**, **Eimsbüttel**, and **Ottensen**. The basic characteristics of these selected districts are also summarized in Table 1.

***Table 1**: Basic characteristics of the selected districts for the new study café.*



	District	Lat	Lon	Price	Cluster	ClusterName	Cafés
0	Barmbek-Süd	53.579885	10.043251	14.50	4	Food & Drink	0
1	Eilbek	53.567237	10.045241	13.75	4	Food & Drink	3
2	Eimsbüttel	53.572483	9.950100	14.12	4	Food & Drink	3
3	Ottensen	53.555066	9.919819	14.42	4	Food & Drink	2

We can now present the above map and the list of selected districts to make recommendations to potential investors about where to open the new study café. Based on this list and my own experiences from living in Hamburg for more than 15 years, I would specifically recommend the district **Eimsbüttel**, as it is a very lively and nice area, has good traffic connections and good proximity to Hamburg's center and several universities.

6. Conclusion

The aim of this project was to help investors in finding the optimal location for opening a new study café by giving them a nice visual overview of Hamburg's districts and their relevant characteristics. To this end, the project collected and combined several data about Hamburg (e.g. rental prices, locations data of universities, venues, and districts) and visualized these data on a geographical map of Hamburg. Based on this map, this project identified four districts (**Barmbek-Süd**, **Eilbek**, **Eimsbüttel**, **Ottensen**) that meet the required criteria for the new venue, i.e. close proximity to one or more univerisites, low to moderate rental prices, lively area with other food & drink venues, but not too many other cafés yet. These districts should be considered by potential investors as optimal places to open a new study café or other off-campus student venues, which may greatly enrich student life in Hamburg.

7. References

[1] https://www.hamburg.de/buergermeisterreden-2017/9969574/hamburg-eine-metropole-der-wissenschaft-im-norden/
[2] https://en.wikipedia.org/wiki/Hamburg
[3] https://www.abendblatt.de/hamburg/article208791261/Hamburg-hat-so-viele-Studenten-wie-nie-zuvor.html
[4] https://www.abendblatt.de/hamburg/article212674037/Rekordzahl-an-Studenten-in-Deutschland.html
[5] https://de.statista.com/statistik/daten/studie/255207/umfrage/studierende-an-hochschulen-in-hamburg/
[6] https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Hochschulen/StudierendeHochschulenSommersemester2110410187314.pdf?__blob=publicationFile
[7] https://www.welt.de/regionales/hamburg/article153979282/Die-Universitaet-ist-zum-Sanierungsfall-geworden.html
[8] https://www.uni-hamburg.de/forschung/forschungsprofil/exzellenzcluster.html
[9] https://www.uni-hamburg.de/newsroom/forschung/2018/0927-exzellenzstrategie.html
[10] https://www.uni-hamburg.de/newsroom/campus/2019/0122-sciencecitybahrenfeld.html
[11] https://fiona.uni-hamburg.de/a0e139bf/broschuereSciencecitybahrenfeld.pdf
[12] https://www.hamburg.de/pressearchiv-fhh/12097952/2019-01-22-bwfg-science-city-bahrenfeld/
[13] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[14] https://statweb.stanford.edu/~gwalther/gap