Coursera Data Science Capstone Project

Hendrik Welsch, January 2021

Project: Analysis of Hamburg Neighbourhoods

1. Introduction: Business Problem and Stakeholder Analysis

This data science project will focus on the city of Hamburg, Germany. Hamburg is one of the largest cities in Germany, an industry hub in the north of the country, and a desired place to live and work for many, especially young people such as students. With these advantages, however, come comparably high rents. Looking for a place to live in Hamburg requires one to look at several variables to make a reasonable choice:

- First of all, where can one afford to rent an apartment?
- Second, which areas cover ones personal interests best? A nature-loving person might like to live in the vincinity of parks and trails, while others might enjoy a neighbourhood with plenty of restaurants and bars.

In this project, I will analyze neighbourhoods in Hamburg in regards to rent per square meter and popular venues. Primary stakeholders who would be interested in this data are:

- people house-/apartment-hunting
- realtors, who can use the data to define areas best matching the demands of their clients

Providing a compact and easy-to-understand dataset characterizing Hamburg neighbourhoods would provide these stakeholders with a foundation to quickly determine target neighbourhoods for their search and avoid them wasting time searching through the whole city.

Furthermore, there are secondary stakeholders who also might find the analysis interesting: for example, city officials could use the visualization to find areas that would benefit the most from additional housing projects or additional public recreational areas.

Entrepreneurs that plan to open a certain type of service could use the data to find areas where this type is underrepresented. This will be exercised in this project using the example of nightlife locations in Hamburg.

2. Data

This project will work with a dataset provided by the city of Hamburg. It can be found here: https://mietspiegeltabelle.de/mietspiegel-hamburg/. The dataset contains the names of the neighbourhoods of Hamburg and their respective mean rents per square meter from 2017.

Mietpreis Hamburg

Durchschnittliche Miete in Euro je Quadratmeter des Jahres 2017.

sehr niedrig niedrig durchschnittlich sehr hoch

Stadtteil	Miete pro m^2
Allermöhe	8,70 Euro
Alsterdorf	10,72 Euro
Altengamme	8,00 Euro
Altenwerder	9,08 Euro
Altona-Altstadt	10,60 Euro

Figure 1: Head of table containing Hamburg neighbourhoods and associated mean rent per square meter in €. Retrieved from https://mietspiegeltabelle.de/mietspiegel-hamburg/ (German)

Furthermore, the neighbourhood names in the above dataset will be used to determine geolocation data (latitude, longitude) and subsequently retrieve venue data using the Foursquare API. Venue data retrieved from Foursquare will be further disassembled to retrieve relevant information such as venue category types at different levels of precision. As an example, a restaurant of type "Italian Restaurant" can be more broadly categorized into "Italian" and even more generalized, be categorized as "food" (see figure 2 below).


```
'venue': {'id': '54200eb6498e5af295bdd77c',
       'name': 'cantinetta ristorante & bar',
       'location': {'address': 'Pickhuben 3',
        'lat': 53.54411350571698,
        'lng': 9.994533061981201,
        'labeledLatLngs': [{'label': 'display',
          'lat': 53.54411350571698,
          'lng': 9.994533061981201}],
        'distance': 158,
        'postalCode': '20457',
        'cc': 'DE',
        'city': 'Hamburg',
        'state': 'Hamburg',
        'country': 'Deutschland',
        'formattedAddress': ['Pickhuben 3', '20457 Hamburg', 'Deutschland']},
       'categories': [{'id': '4bf58dd8d48988d110941735',
        'name': 'Italian Restaurant',
        'pluralName': 'Italian Restaurants',
        'shortName': 'Italian',
        'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/italian_',
         'suffix': '.png'},
        'primary': True}],
       'photos': {'count': 0, 'groups': []}},
      'referralId': 'e-0-54200eb6498e5af295bdd77c-2'},
     {'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]}
```


Figure 2: Example of venue data retrieved from Foursquare. Extractable category types are marked red.

3. Methodology & Exploratory Data Analysis

3.1. Data Acquisition and Preparation

Data of Hamburg neighbourhood names and mean rent was downloaded from https://mietspiegeltabelle.de/mietspiegel-hamburg/ and saved to a data frame. Column labels were translated to English and rent in € was converted to data type "float". The island of "Neuwerk", which belongs to Hamburg, was removed from the list of neighbourhoods as it is not located within or near the city.

	Neighbourhood	rent per m2
0	Allermöhe	8.70
1	Alsterdorf	10.72
2	Altengamme	8.00
3	Altenwerder	9.08
4	Altona-Altstadt	10.60

Figure 3: Dataframe containing Hamburg neighbourhoods and mean rent after clean-up.

Coordinates for each neighbourhood, consisting of latitude and longitude, were retrieved using the `geopy` package.

	Neighbourhood	rent per m2	Latitude	Longitude
0	Allermöhe	8.70	53.483600	10.125000
1	Alsterdorf	10.72	53.610541	10.003889
2	Altengamme	8.00	53.443796	10.273921
3	Altenwerder	9.08	53.504700	9.920560
4	Altona-Altstadt	10.60	53.549660	9.945352

Figure 4: Dataframe containing Hamburg neighbourhoods with mean rent and coordinates. First 5 rows shown.

Based on these coordinates, location data of popular venues in a radius of 600 meters were retrieved from foursquare, with a limit of 100 venues per location. From the retrieved data, venue category types of different detail levels were extracted from the 'categories' dictionary:

```
'categories': [{'id': '4bf58dd8d48988d1e0931735',
          'name': 'Coffee Shop',
          'pluralName': 'Coffee Shops',
          'shortName': 'Coffee Shop',
          'icon':                                                {'prefix':
'https://ss3.4sqi.net/img/categories_v2/food/coffeeshop_',
           'suffix': '.png'},
          'primary': True}]
```

- "Venue Category", retrieved from "name"
- "Higher Category", the last part of the URL (i.e., coffeeshop_)
- "Highest Category", the second-to-last part of the URL (i.e., sfood)

Most of these categories are venues of interest for someone to have in the vincinity of their apartment. However, some categories might skew with the clustering approach later on: Mainly, these categories are transport locations ("Bus Stop", "Bridge", "Intersection", "Light Rail Station", "Metro Station", "Road", "Train Station", "Bus Station"). Although it is nice to have a good connection

to public transport, it is often not the deciding factor when choosing a neighbourhood since public transport connection in German cities is generally very good. Entries of these categories were therefore removed from the dataset.

After clean-up, there were 259 unique venue categories, 184 unique higher-level categories, and 7 highest-level categories.

	Neighbourhood	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Higher Category	Highest Category
0	Alsterdorf	53.610541	10.003889	Eppendorfer Moor	53.613315	10.002277	Nature Preserve	naturepreserve_	parks_outdoors
1	Alsterdorf	53.610541	10.003889	REWE	53.607647	10.005769	Supermarket	food_grocery_	shops
2	Alsterdorf	53.610541	10.003889	Braband	53.613330	10.002281	Café	cafe_	food
3	Alsterdorf	53.610541	10.003889	Best Western Premier Alsterkrug Hotel	53.613080	9.999037	Hotel	hotel_	travel
4	Alsterdorf	53.610541	10.003889	Eiskaffee Eis Perle	53.608354	10.009394	Ice Cream Shop	icecream_	food

Figure 5: Example of a dataframe containing different detail levels of venue categories retrieved from Foursquare. First five rows shown.

For the analysis of nightlife locations in Hamburg (see Python Notebook part 5), data was further trimmed to only include entries from the highest category “nightlife”.

3.2. Data Visualization

With `folium`, maps were generated and populated with coloured markers. Colour legends were generated using `branca`.

Results of elbow method to determine best k for k-Means clustering as well as mean rent per cluster were plotted using `mathplotlib`.

3.3. Machine-Learning based Neighbourhood Clustering

To prepare data for clustering, the data was transformed into data frames containing frequencies of venue categories for each neighbourhood using one-hot-encoding.

	Neighbourhood	ATM	Accessories Store	Afghan Restaurant	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Austrian Restaurant	Auto Dealership	
0	Alsterdorf	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.105263	0.000000	0.0	0.0	0.(
1	Altona-Altstadt	0.0	0.0	0.0	0.0	0.023256	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.(
2	Altona-Nord	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.026316	0.026316	0.0	0.0	0.(
3	Bahrenfeld	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.(
4	Barmbek-Nord	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.045455	0.000000	0.0	0.0	0.(
5	Barmbek-Süd	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.(
6	Bergedorf	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.(
7	Bergstedt	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.(
8	Billbrook	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.:
9	Billstedt	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.(

Figure 6: Example of a dataframe showing frequencies of venue categories for each neighbourhood. First ten rows shown.

3.4. Exploratory Data Analysis

3.4.1. Visualization of mean rent distribution of Hamburg neighbourhoods

To help someone find a suitable neighbourhood to live in, the mean rent per square meter is a useful metric. The aim was therefore to visualize the mean rent in different neighbourhoods in Hamburg to give the user a good overview of rent prices. To make this understandable at a quick glance, neighbourhoods were coloured based on the rent price. To reduce the number of colours used, several threshold values were defined:

- “min rent”: lowest mean rent found in Hamburg
- “max rent”: highest mean rent found in Hamburg
- “cheap”: mean rents located below the 25 % quantile
- “expensive”: mean rents located above the 75 % quantile

The threshold values were determined using pandas functions .min(), .max() and .quantile(n), where n is 0.25 or 0.75, respectively.

Markers for neighbourhoods were then plotted onto a map of Hamburg based on their coordinates (lat, long) and coloured based on the above listed threshold values (Figure 7). Neighbourhoods closer to the city center around north/northwest of river Elbe show higher rents (orange markers, > 9.88 €/m2), with the highest rent for HafenCity (= harbour city center district, 14.4 €/m2). Neighbourhoods farther out show medium (gray markers, between 8.02 and 9.88 €/m2) rent north of the river, while parts south of the river contain a lot of neighbourhoods with lower rent (blue, < 8.02 €/m2), including the neighbourhood with lowest rent in Hamburg (Moorburg, green, 5.79 €/m2).

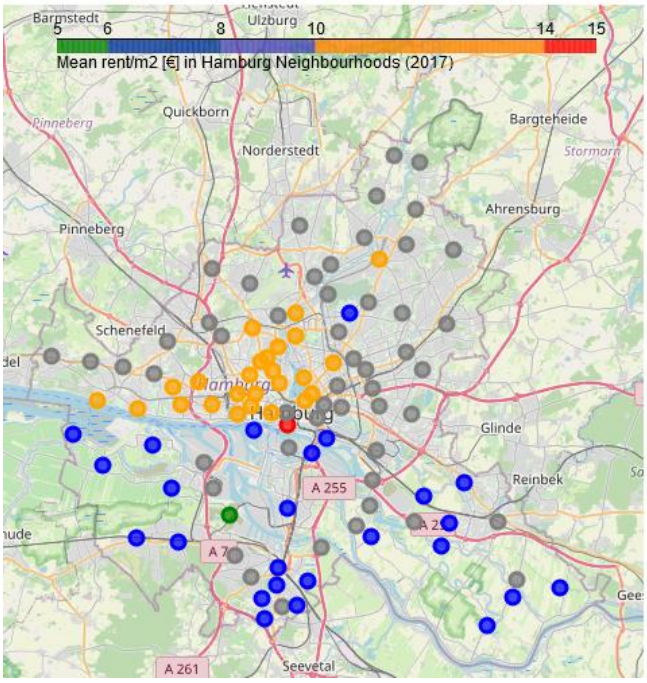


Figure 7: Map of Hamburg with neighbourhoods coloured based on mean rent/m2 [€].

3.4.2. Neighbourhood Clustering

People not only choose a neighbourhood to live in based on the rent, but also for venues of personal interest located in this neighbourhood. While one person might look for lively neighbourhoods with plenty of cafes and restaurants, some other person might prefer a neighbourhood with lots of nature or sport/fitness venues.

Foursquare provides a substantial database of venue data for most regions in the world. This data can be used to define the most frequent venue types of each neighbourhood, and subsequently cluster these neighbourhoods based on their similarity in popular venue types. This can be very helpful to choose a neighbourhood most fitting to one's personal needs.

From Foursquare, the top 100 most popular venues for each neighbourhood were retrieved. As described in more detail in section 3.1, venue category type data at three different levels of detail was selected from each venue data set and saved into a data frame. After clean-up of this dataset and one-hot-encoding, three data frames with the frequency of each venue type per neighbourhood were generated.

Because we do not have a defined parameter on which we could separate and group the neighbourhoods, an unsupervised clustering method should be employed. Of several ones available, I chose to use the **k-means** approach implemented in the scikit learn package. With k-means, the machine learning algorithm will cluster neighbourhoods based on similarity within a distance matrix into k clusters. K is defined by the user, and optimal k can be determined using the elbow method (https://www.scikit-yb.org/en/latest/api/cluster/elbow.html). Here, the modeling using the k-means algorithm is run with the dataset for ks in a given range (here, 1 to 15), and for each model a distortion value is calculated and plotted. A sharp bend in the distortion curve then marks the optimal k.

Although the distortion decreases with increasing k (= the goodness of fit increases), no real bend is visible for the detailed venue categories as well as the higher-level categories (Figure 8, top and middle). Thus, a very high k seems to preferential, however, to many clusters would at one point nullify the aim of grouping neighbourhoods to decrease the complexity of the data. Thus, k=6 was chosen, since this was deemed a good compromise between low distortion and low complexity. For the highest category level, which consisted only of 7 unique types, the result was more apparent, and a k=3 was chosen (Figure 8, bottom).

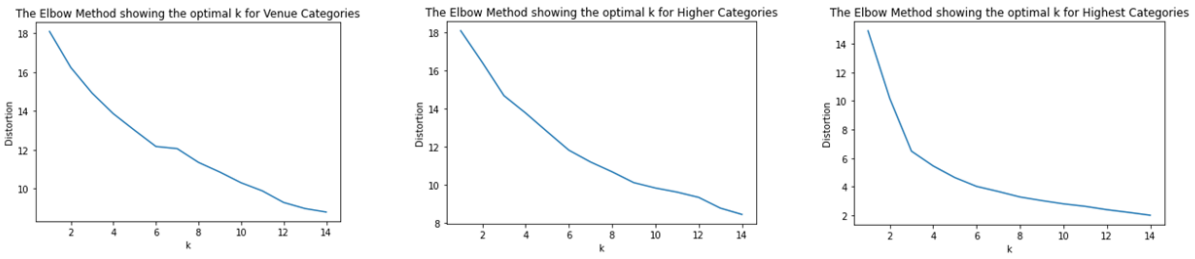


Figure 8: Results of elbow-method to determine optimal k for k-means clustering. From left to right: Distortion curve for detailed venue categories, higher-level categories and highest-level categories.

Neighbourhood data was clustered with k-means and above determined optimal k's, and cluster labels were added to the data frames containing neighbourhood and venue type data. After removing neighbourhoods without venue types retrieved from Foursquare, markers for

neighbourhoods were plotted onto a map of Hamburg, and markers were coloured based on cluster affiliation. Cluster colours were allocated based on cluster size to have the biggest cluster have the same colour in all three approaches.

To reduce the complexity of the clustered data, which is presented in form of a data frame, a function for a meta analysis was generated:

```
def topVenueCategories(df, cluster, n=5):
    v = ["1st", "2nd", "3rd", "4th", "5th"]
    df_1 = df["1st Most Common Venue"].loc[df["Cluster Labels"]==cluster].value_counts()
    df_2 = df["2nd Most Common Venue"].loc[df["Cluster Labels"]==cluster].value_counts()
    df_3 = df["3rd Most Common Venue"].loc[df["Cluster Labels"]==cluster].value_counts()
    print("Top categories are:\n","\n\n1st Most Common Venue \n",
          df_1[0:n],
          "\n\n2nd Most Common Venue \n",
          df_2[0:n-1],"\n\n3rd Most Common Venue \n",
          df_3[0:n-2])
```

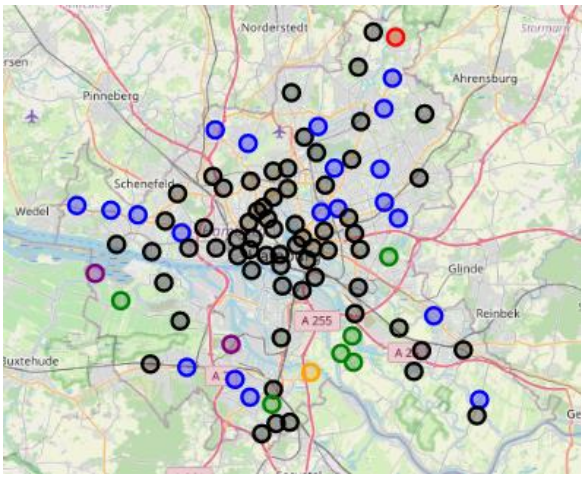
Figure 9: Meta analysis function to retrieve overall top venue categories within a cluster.

It returns up to five most frequent venue categories in the first column, that already lists the most frequent category per neighbourhood and likely had the strongest impact on the clustering process. This is repeated for the second and third column, returning the top four and three results, respectively. Below, an example output is shown:

```
Top categories are:


1st Most Common Venue
 Café              10
Hotel               8
Bakery              7
Ice Cream Shop      4
Supermarket         3
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Supermarket           7
Café                   7
Italian Restaurant     5
Hotel                  4
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Hotel                   5
Ice Cream Shop           4
Fast Food Restaurant     4
Name: 3rd Most Common Venue, dtype: int64
```

Figure 10: Example of return from the meta analysis function.

3.4.2.1 Cluster Analysis for Detailed Venue Categories




```
Mean rent of cluster 5.0 (65 neighbourhoods):  9.63 +/- 1.71 [€]
Mean rent of cluster 0.0 (20 neighbourhoods):  8.8 +/- 1.02 [€]
Mean rent of cluster 4.0 (6 neighbourhoods):  7.87 +/- 0.47 [€]
Mean rent of cluster 2.0 (2 neighbourhoods):  6.07 +/- 0.4 [€]
Mean rent of cluster 1.0 (1 neighbourhoods):  8.88 +/- nan [€]
Mean rent of cluster 3.0 (1 neighbourhoods):  8.95 +/- nan [€]
```

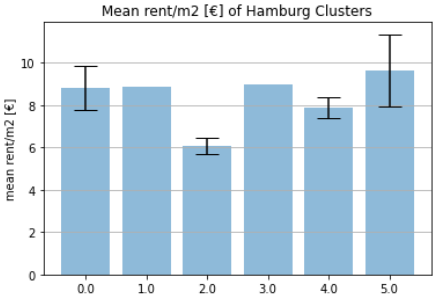


Figure 11: Left: Map of Hamburg showing neighbourhoods clustered based on most frequent venue category types. Clusters are coloured based on cluster size (black > blue > green > purple > orange > red). Marker radius equals 600 m on map. Right: Mean rent per square meter in € +/- standard deviation for each cluster.

Cluster 5:

```
Top categories are:


1st Most Common Venue
 Café               10
Hotel               8
Bakery              7
Ice Cream Shop      4
Supermarket         3
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Supermarket         7
Café                 7
Italian Restaurant   5
Hotel                4
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Hotel                5
Ice Cream Shop        4
Fast Food Restaurant  4
Name: 3rd Most Common Venue, dtype: int64
```

Figure 12: Top venue categories in Cluster 5.

The largest cluster (5, black, 69 neighbourhoods) contains all neighbourhoods in the city center, and is spread all over Hamburg. There is a strong variety in top venue categories. Overall, restaurants, hotels, cafes and various shopping locations dominate the image. These neighbourhoods can therefore be defined as "inner city"-like neighbourhoods. They show the overall highest mean rent, but there is also a wide spread (high standard deviation).

Cluster 0:

```
Top categories are:


1st Most Common Venue
 Supermarket              14
Fast Food Restaurant      1
Bank                      1
Asian Restaurant          1
Bakery                    1
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Greek Restaurant     2
Supermarket           2
Bakery                2
Shopping Mall         2
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Hotel                4
Supermarket           3
Asian Restaurant      2
Name: 3rd Most Common Venue, dtype: int64
```

Figure 13: Top venue categories for Cluster 0.

The second-largest cluster (0, blue, 20 neighbourhoods) shows neighbourhoods more present at the outskirts of the city. Supermarkets are mostly the top venues here, and other services for the daily life such as bakeries, grocery and department stores and banks can be found among the top venues. The mean rent is lower compared to cluster 5. These neighbourhoods can therefore be defined as "suburb living"-like neighbourhoods.

Clusters 4 and 2:

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
12	Billstedt	8.04	53.548899	10.122625	4.0	Lighting Store	Cosmetics Shop	Soccer Field	German Restaurant	Fish & Chips Shop
61	Neuenfelde	7.05	53.518241	9.807916	4.0	German Restaurant	Taverna	Pier	Tennis Court	Zoo Exhibit
68	Ochsenwerder	7.68	53.475203	10.081077	4.0	IT Services	Hotel	German Restaurant	Fish & Chips Shop	Falafel Restaurant
83	Spadenland	8.37	53.481400	10.066400	4.0	German Restaurant	Vegetarian / Vegan Restaurant	Breakfast Spot	Farmers Market	Food Court
91	Tatenberg	8.21	53.493600	10.079700	4.0	German Restaurant	Boat or Ferry	Lake	Canal Lock	Zoo Exhibit
100	Wilstorf	7.89	53.445867	9.985275	4.0	German Restaurant	Lake	Supermarket	Water Park	Farmers Market

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
17	Cranz	6.35	53.537285	9.777929	2.0	German Restaurant	Zoo Exhibit	French Restaurant	Forest	Food Court
58	Moorburg	5.79	53.487800	9.937477	2.0	German Restaurant	Zoo Exhibit	French Restaurant	Forest	Food Court

Figure 14: Top venue types for neighbourhoods in Cluster 4 (top) and Cluster 2 (bottom).

Cluster 4 (green) and 2 (purple) both show German Restaurants to be defining top venues. However, neighbourhoods in Cluster 2 are separated from neighbourhoods in Cluster 4 by the presence of Zoo Exhibits and French restaurants. Interestingly, even though both neighbourhoods in Cluster 2 show the exact same venue signature, they are located in different areas of Hamburg. Cluster 2 shows the lowest mean rent in Hamburg.

Cluster 1 and 3:

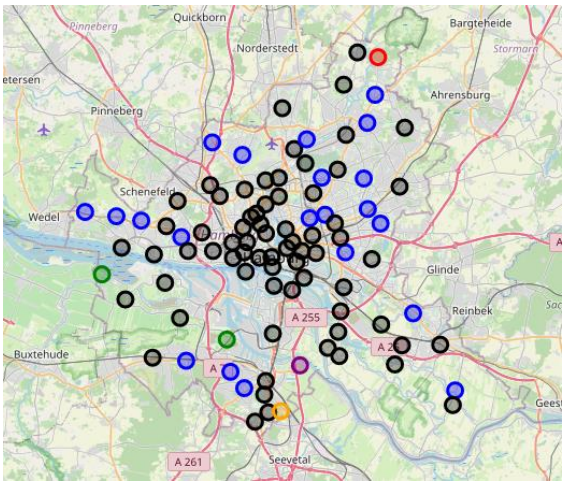
	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
64	Neuland	8.88	53.468155	10.030961	1.0	Coffee Shop	Zoo Exhibit	Exhibit	Forest	Food Court

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
102	Wohldorf-Ohlstedt	8.95	53.7019	10.1306	3.0	Forest	Zoo Exhibit	Dance Studio	Fountain	Food Court

Figure 15: Top venue types for neighbourhoods in Cluster 1 (top) and Cluster 3 (bottom).

The last two clusters both contain only one neighbourhood, making an assessment difficult.

3.4.2.2 Cluster Analysis for Higher-Level Venue Categories




```
Mean rent of cluster 4.0 (69 neighbourhoods):  9.52 +/- 1.72 [€]
Mean rent of cluster 0.0 (21 neighbourhoods):  8.79 +/- 1.0 [€]
Mean rent of cluster 2.0 (2 neighbourhoods):  6.07 +/- 0.4 [€]
Mean rent of cluster 1.0 (1 neighbourhoods):  8.88 +/- nan [€]
Mean rent of cluster 5.0 (1 neighbourhoods):  7.3 +/- nan [€]
Mean rent of cluster 3.0 (1 neighbourhoods):  8.95 +/- nan [€]
```

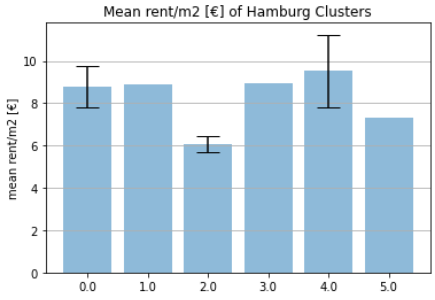


Figure 16: Left: Map of Hamburg showing neighbourhoods clustered based on most frequent higher-level venue category types. Clusters are coloured based on cluster size (black > blue > green > purple > orange > red). Marker radius equals 600 m on map. Right: Mean rent per square meter in € +/- standard deviation for each cluster.

We can observe a distribution very similar to the approach in 4.1.

```
Top categories are:


1st Most Common Venue
 cafe_            10
default_           8
food_grocery_      7
german_            5
hotel_             5
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 default_          9
hotel_             7
bakery_            7
food_grocery_      5
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 food_grocery_     8
cafe_              3
fastfood_          3
Name: 3rd Most Common Venue, dtype: int64
```

Figure 17: Top venue categories for Cluster 4.

The biggest cluster (4, black, 69 venues) contains all inner-city neighbourhoods and also spreads out all over Hamburg (Figure 17). Cafes, food-grocery stores, bakeries and hotels are frequently present. A problem with the higher-level categories retrieved from the icon URL is the category "default", which does not hold any information about what type of venue this is.

```
Top categories are:


1st Most Common Venue
 food_grocery_    14
fishandchips_      1
hotel_             1
greek_             1
hikingtrail_       1
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 gym_          2
asian_        2
pharmacy_     2
park_         1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 food_grocery_    4
bakery_           3
default_          2
Name: 3rd Most Common Venue, dtype: int64
```

Figure 18: Top venue categories for Cluster 0.

The second-largest cluster (2, blue, 21 neighbourhoods) is predominantly defined by the food_grocery category (Figure 18).

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
17	Cranz	6.35	53.537285	9.777929	2.0	german_	zoo_	financial_	food_foodcourt_	food_fishmarket_
58	Moorburg	5.79	53.487800	9.937477	2.0	german_	zoo_	financial_	food_foodcourt_	food_fishmarket_

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
64	Neuland	8.88	53.468155	10.030961	1.0	coffeeshop_	zoo_	financial_	food_gourmet_	food_foodcourt_

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
79	Rönneburg	7.3	53.433214	10.006395	5.0	greek_	zoo_	field_	food_foodcourt_	food_fishmarket_

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
102	Wohldorf-Ohlstedt	8.95	53.7019	10.1306	3.0	default_	zoo_	financial_	food_gourmet_	food_foodcourt_

Figure 19: Top venue types for neighbourhoods in Clusters 2, 1, 5 and 3 (from top to bottom).

The last four clusters contain 2 or 1 neighbourhoods (Figure 19), and the cluster in 3.4.2.1 that was dominated by german restaurants is not present anymore. Neighbourhoods Cranz and Moorburg again have been clustered together. While neighbourhoods Neuland and Wohldorf-Ohlstedt are again each their own cluster, now also Rönneburg has been put into a separate cluster.

3.4.2.2 Cluster Analysis for Highest-Level Venue Categories

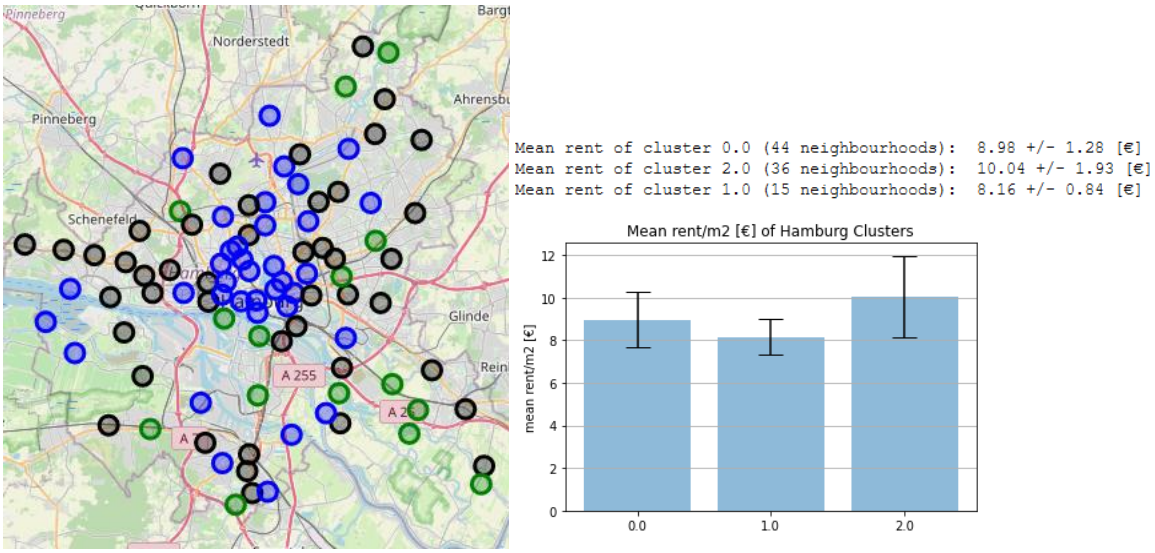


Figure 20: Map of Hamburg showing neighbourhoods clustered based on most frequent highest-level venue category types. Clusters are coloured based on cluster size (black > blue > green). Marker radius equals 600 m on map. Right: Mean rent per square meter in € +/- standard deviation for each cluster.

```
Top categories are:


1st Most Common Venue
 shops      29
food        13
travel       2
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 food                   25
shops                   14
travel                   2
arts_entertainment       1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 travel                 15
parks_outdoors          13
arts_entertainment       7
Name: 3rd Most Common Venue, dtype: int64
```

Figure 21: Top venue categories for Cluster 0.

For clustering based on highest-level venue categories, we chose only three clusters (k=3). The largest cluster (0, black, 44 neighbourhoods) in this case covers neighbourhoods spread all over Hamburg, but predominantly more at the outskirts. Neighbourhoods are defined by frequency of locations for shopping and eating (shops, food).

```
Top categories are:


1st Most Common Venue
 food    36
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 shops                 20
travel                10
parks_outdoors         3
arts_entertainment     2
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 travel       12
shops         9
nightlife     5
Name: 3rd Most Common Venue, dtype: int64
```

Figure 22: Top venue categories for Cluster 2.

The second-largest cluster (2, blue, 36 neighbourhoods) is also spread over Hamburg, but exclusively covers the complete inner city. Here, "food" locations (restaurants, fast food shops) dominate the scene, followed by "travel" locations (such as hotels, hostels), which fits to the image of typical inner-city neighbourhoods with more tourists.

```
Top categories are:


1st Most Common Venue
 parks_outdoors        10
travel                 3
shops                  1
arts_entertainment     1
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 shops                 6
travel                5
food                  2
arts_entertainment    1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 food                  5
shops                 3
arts_entertainment    3
Name: 3rd Most Common Venue, dtype: int64
```

Figure 23: Top venue categories for Cluster 1.

The third cluster (1, green, 15 neighbourhoods) is defined by the "parks_outdoors" category. Most of the neighbourhoods can be found farther away from the city center, and there is a higher frequency south of river Elbe.

3.4.3 Distribution of Nightlife Locations

The venue type data retrieved from Foursquare can also be used to answer more specialized questions. In this section, the aim was to cluster neighbourhoods based on frequency of nightlife location types. This can be interesting for stakeholders who are planning to open a nightclub or bar and ask questions such as:

- where is a certain type of nightlife location over- or underrepresented? Information on this can be useful to avoid too much competition.
- how is the mean rent in the different locations? It would not be a good idea to open a high-prised cocktail bar in a neighbourhood predominantly occupied by people with low-to-medium income.

Venue data was filtered for highest-level category “nightlife”. Remaining venue categories were: 'Pub', 'Bar', 'Nightclub', 'Cocktail Bar', 'Dive Bar', 'Hookah Bar', 'Lounge', 'Irish Pub', 'Beer Bar', 'Beer Garden', 'Bar', 'Whisky Bar', 'Beer Store', 'Beach Bar'. Venue Categories are not too detailed, but also not to general. However, there are two problematic categories:

- Beer Store: this usually means a store where you can by alcoholic drinks late at night (like a 24/7 supermarket), but this is not a spot to spend your time when going out.
- Other nightlife: this category encompasses a variety of more niche nightlife spots. Since our clustering algorithm does not know this, this category could skew the analysis.

Data containing one of these two categories were therefore removed. The remaining dataset was preprocessed for k-means clustering as described earlier.

Using the elbow-method, optimal k for k-means clustering was determined to be k=4, since this seemed to be a good compromise between low distortion and low complexity.

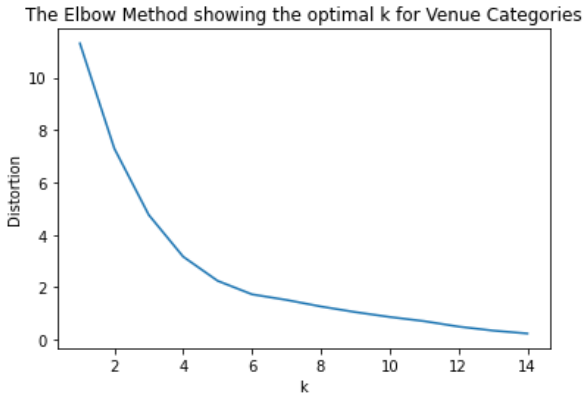
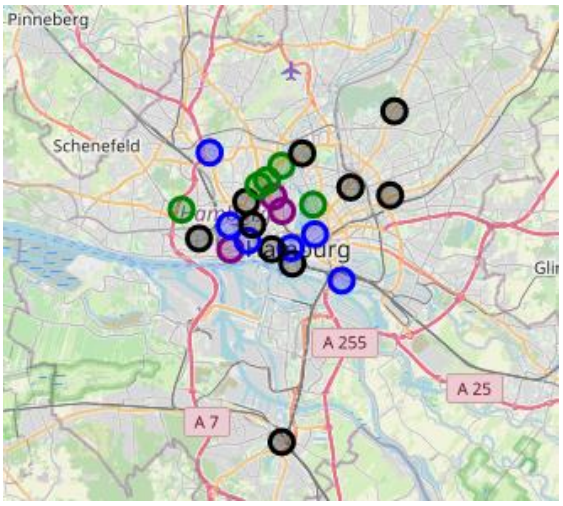


Figure 24: Results of elbow-method to determine optimal k for k-means clustering.

As described before, data was clustered using k-means and k=4, and neighbourhoods were plotted onto a map of Hamburg and coloured based on cluster affiliation.

A lot of Hamburg neighbourhoods did not return any nightlife locations from Foursquare and were therefore removed from clustering. Most of the neighbourhoods which contained nightlife locations among the top 100 popular venues retrieved from Foursquare are located close to the City Center north of river Elbe (Figure 25).




```
Mean rent of cluster 1.0 (10 neighbourhoods):  10.78 +/- 1.85 [€]
Mean rent of cluster 2.0 (6 neighbourhoods):  10.03 +/- 1.3 [€]
Mean rent of cluster 0.0 (5 neighbourhoods):  11.59 +/- 0.87 [€]
Mean rent of cluster 3.0 (3 neighbourhoods):  11.91 +/- 1.15 [€]
```

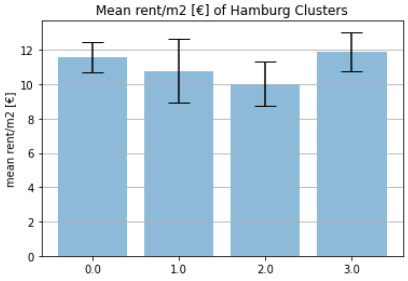


Figure 25: Map of Hamburg showing neighbourhoods clustered based on nightlife location clusters. Clusters are coloured based on cluster size (black > blue > green > purple). Marker radius equals 600 m on map. Right: Mean rent per square meter in € +/- standard deviation for each cluster.

Compared to clustering of neighbourhoods based on popular venues in general (see part 3.4.2), clustering based on nightlife locations reveals a relatively clear image:

Cluster 1:

```
Top categories are:


1st Most Common Venue
 Bar           9
Hookah Bar     1
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Whisky Bar       4
Pub               2
Cocktail Bar      1
Irish Pub         1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Pub           4
Whisky Bar     3
Lounge         1
Name: 3rd Most Common Venue, dtype: int64
```

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
8	Barmbek-Süd	10.26	53.579885	10.043251	1.0	Bar	Whisky Bar	Pub
16	Bramfeld	8.96	53.616809	10.079001	1.0	Bar	Whisky Bar	Pub
23	Eimsbüttel	11.36	53.572937	9.958261	1.0	Bar	Pub	Dive Bar
33	HafenCity	14.40	53.542913	9.995835	1.0	Bar	Cocktail Bar	Whisky Bar
37	Harburg	7.91	53.456222	9.987211	1.0	Bar	Whisky Bar	Pub
65	Neustadt	10.69	53.549881	9.979048	1.0	Bar	Irish Pub	Beer Bar
72	Ottensen	11.60	53.555066	9.919819	1.0	Bar	Pub	Whisky Bar
89	Sternschanze	11.73	53.561768	9.963282	1.0	Bar	Nightclub	Lounge
97	Wandsbek	9.05	53.576003	10.075535	1.0	Hookah Bar	Bar	Whisky Bar
101	Winterhude	11.83	53.596390	10.003832	1.0	Bar	Whisky Bar	Pub

In Cluster 1 (black), bars (general, whiskey, and hookah) dominate the scene. Neighbourhoods can be found both in the city center and farther away.

Cluster 2:

```
Top categories are:


1st Most Common Venue
 Nightclub    4
Gay Bar       1
Lounge        1
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Bar            2
Beer Garden     1
Lounge          1
Pub             1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Whisky Bar     2
Pub             2
Hookah Bar      2
Name: 3rd Most Common Venue, dtype: int64
```

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
5	Altona-Nord	10.61	53.561400	9.944720	2.0	Nightclub	Bar	Whisky Bar
34	Hamburg-Altstadt	9.57	53.550468	9.994640	2.0	Nightclub	Lounge	Hookah Bar
77	Rothenburgsort	7.79	53.534658	10.036005	2.0	Nightclub	Beer Garden	Whisky Bar
84	St.Georg	11.46	53.557149	10.014256	2.0	Gay Bar	Pub	Hookah Bar
85	St.Pauli	10.93	53.553935	9.959432	2.0	Nightclub	Bar	Pub
88	Stellingen	9.81	53.596777	9.928410	2.0	Lounge	Whisky Bar	Pub

Cluster 2 (blue) is defined by a high frequency of nightclubs and most of the neighbourhoods are close to the city center and to the Elbe riverside. This cluster contains among others “St. Pauli” with the (in)famous red-light district “Reeperbahn”.

Cluster 0:

```
Top categories are:


1st Most Common Venue
 Cocktail Bar    5
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Whisky Bar    3
Hookah Bar     1
Pub            1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Pub           3
Whisky Bar     2
Name: 3rd Most Common Venue, dtype: int64
```

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
6	Bahrenfeld	10.09	53.569070	9.905583	0.0	Cocktail Bar	Whisky Bar	Pub
25	Eppendorf	11.83	53.590391	9.986877	0.0	Cocktail Bar	Whisky Bar	Pub
41	Hoheluft-Ost	12.28	53.583434	9.975220	0.0	Cocktail Bar	Pub	Whisky Bar
42	Hoheluft-West	11.68	53.580754	9.968331	0.0	Cocktail Bar	Hookah Bar	Whisky Bar
93	Uhlenhorst	12.07	53.571509	10.012736	0.0	Cocktail Bar	Whisky Bar	Pub

Cluster 0 (green) is defined by a high frequency of cocktail bars and three out of five neighbourhoods are located next to each other north-west of the city center.

Cluster 3:

```
Top categories are:


1st Most Common Venue
 Pub    3
Name: 1st Most Common Venue, dtype: int64

2nd Most Common Venue
 Whisky Bar      2
Cocktail Bar     1
Name: 2nd Most Common Venue, dtype: int64

3rd Most Common Venue
 Nightclub    2
Bar           1
Name: 3rd Most Common Venue, dtype: int64
```

	Neighbourhood	rent per m2	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
4	Altona-Altstadt	10.60	53.549660	9.945352	3.0	Pub	Whisky Bar	Nightclub
38	Harvestehude	12.72	53.575433	9.980440	3.0	Pub	Whisky Bar	Nightclub
78	Rotherbaum	12.41	53.568407	9.987432	3.0	Pub	Cocktail Bar	Bar

Cluster 3 (purple) is defined by high frequency of pubs.

4. Results & Discussion

In this section, I will summarize and discuss the results from section 3.

Neighbourhoods of Hamburg were grouped based on mean rent and statistical thresholds and plotted on a map of Hamburg. The visualization revealed that living in the city center or north-west of it is pricier than living farther away or south of river Elbe. This visualization can be helpful to stakeholders who are interested in the distribution of rent in the city of Hamburg. Both people looking for apartments as well as businesses looking for a place to open a store or service focusing on a certain income bracket can profit from this data visualization. Future improvements of this visualization approach should include among others data about whether a neighbourhood is dominated by living area or business/industry, since it is not desirable to live in an industry-dominated neighbourhood.

The different clustering approaches yielded somewhat satisfying results. Using detailed or higher-level venue categories for clustering both resulted in rather similar cluster composition. Mainly, most neighbourhoods can be separated into “inner-city”-like neighbourhoods, which are close to the city center and show high frequencies of restaurants, cafes and hotels, or “suburb”-like neighbourhoods, which are located more at the outskirts and show high frequencies of supermarkets and other services relevant to the daily life. The definition of venue categories by Foursquare can be problematic, especially the higher-level category type “default”. To return a more useful analysis from this data, it might be helpful to group venue type categories into higher-level categories manually; however, this can be a tedious amount of work and might be induce a bias due to a single person doing the grouping.

Using the highest-level category types, we can again show the difference between inner-city and suburb. Interestingly, this approach also finds a cluster of neighbourhoods defined by high frequency of parks and other outdoor activities. The three clusters determined this way therefore can be deemed most useful to give someone a general overview of Hamburg and a good starting point when looking for a place to live.

Generally, one has to consider the effects of two parameters when evaluating the clustering approach. First of all, only 100 venues are returned per neighbourhood. This can skew the image we get for a neighbourhood, especially when comparing neighbourhoods which contain significantly more popular venues. Furthermore, the radius significantly influences the result. Here, 600 m was chosen to avoid too much overlap of neighbourhoods. However, the given coordinates of a neighbourhood might be in a location that, paired with a small radius, does not cover the main living area of a neighbourhood, thus skewing the data.

The power of the location data is best exploited when defining a more precise question. In this work, neighbourhoods were clustered based on nightlife venues. It became apparent that nightlife in Hamburg is predominantly happening in or closer to the city center. However, there are also neighbourhoods such as Harburg, which are farther away and contain predominantly bars. Those neighbourhoods could for example be typical student neighbourhoods, but we lack data to make such a safe conclusion. The clustering approach was able to separate neighbourhoods and show where nightclubs, pubs and cocktail bars, respectively, define the image of a neighbourhood. This data, combined with the mean rent data, can be very useful for someone planning to open a nightlife location to determine the best area for a certain venue type and clientele. For example, opening a

nightclub at or near the northern waterfront around the city center can result in a lot of competition with other nightclubs; on the other hand, a new nightclub in this area can profit from the high number of potential customers going from club to club.

6. Conclusion

This work analyzed Hamburg neighbourhoods for mean rent and defining venue types. The approach can be further improved in the future by adding more parameters such as mean age, mean income, and distribution of education status.

Furthermore, popular venues from Foursquare are dominated by places for shopping, eating & drinking, and other free time activities. Underrepresented are categories such as schools and medical facilities, which are of equal interest to people looking for apartments. Those venues could be added to the data set, however, other rating metrics should be applied.

Overall, the returned data, visualized for easy comprehension, can serve as a starting point for stakeholders interested in finding a place to live or to open a business in Hamburg.