Report: Berlin - a city to live... but where?

Inhalt

1 Introduction:

Berlin is not only the capital city of Germany it is also a city with an interesting history. After the splitting into 4 sectors by the end of the Second World War, the Berlin Wall (construction began on August 13, 1961) was later built as a symbol of the great differences between the East and West powers and of the Cold War. With the fall of the Berlin Wall (November 9, 1989) the reunion of East and West was initiated. And after the German reunification of the GDR and the Federal Republic of Germany (on October 3, 1990), Berlin was simultaneously named the capital city of the Federal Republic of Germany. Today Berlin is a city with many different aspects in art, culture, politics, technology and science. With about 3.7 million inhabitants and an area of about 892 square kilometres, it is the largest city in Germany.

1.1 Problem:

With many inhabitants, however, aspects such as crime and its influence on normal city life are becoming increasingly important. Nowadays, information defines our everyday life. But can they also help us to make decisions that are related to everyday situations? What about moving to a foreign city, for example. Where should you settle down? What are the localities? How do I get around? And above all, am I safe? All these factors can have a strong influence on our decision where to settle down. In the following sections I have tried, taking Berlin as an example, to work out a possibility based on data that should make it easier to decide for a certain district. The question that arises is therefore: *"Is it possible to discover safe and yet attractive and lively areas of the city based on data like district areas positions, population, crime and venues?"*

1.2 Data and Data Processing

The data for the analysis presented here originates from various data sources.

- Technologiestiftung Berlin 2020
- Amt für Statistik Berlin-Brandenburg
- Kriminalitätsatlas Berlin
- Foursquare

In the first part of the Data Processing Part all the necessary datasets are loaded and afterwards cleaned. These datasets mainly consist of crime, population and geographical data. In the data Analysis section all the needed data is analysed and investigated. The regions of Berlin were then further investigated by obtaining interesting venues via the Foursquare API and further clustering of the data.

1.3 Data Cleaning

The obtained data was cleaned and all necessary columns were renamed for a better data handling.

Geojson-Data: The Geojson data (Technologiestiftung Berlin 2020) could be directly accessed from the source and only the column renaming was necessary. For the Geojson data it was most suitable to use the *geopandas* package since the geographical informations were stored as polygon informations. Through this polygon information the borders and areas of the boroughs, neighboorhoods and district areas could be displayed via the *folium* package. With the *geopandas* package it was also possible to easy access the centroids from the polygon information.

```
gdf_Borough.head(2)
```

	gml_id	Gemeinde_name	Gemeinde_schluessel	Land_name	Land_schluessel	Schluessel_gesamt	geometry
0	s_wfs_alkis_bezirk.F176__1	Reinickendorf	012	Berlin	11	11000012	MULTIPOLYGON (((13.32074 52.62660, 13.32045 52...
1	s_wfs_alkis_bezirk.F176__2	Charlottenburg-Wilmersdorf	004	Berlin	11	11000004	MULTIPOLYGON (((13.32111 52.52446, 13.32103 52...

```
gdf_Neighborhood.head(2)
```

	gml_id	spatial_name	spatial_alias	spatial_type	OTEIL	BEZIRK	FLAECHE_HA	geometry
0	re_ortsteil.0101	0101	Mitte	Polygon	Mitte	Mitte	1063.8748	POLYGON ((13.41649 52.52696, 13.41635 52.52702...
1	re_ortsteil.0102	0102	Moabit	Polygon	Moabit	Mitte	768.7909	POLYGON ((13.33884 52.51974, 13.33884 52.51974...

```
gdf_District_Area.head(2)
```

	gml_id	spatial_name	spatial_alias	spatial_type	BZR_NAME	PGR_NAME	BEZNAME	DATUM	SHAPE_AREA	geometry
0	re_bezirksregion.010111	010111	Tiergarten Süd	Polygon	Tiergarten Süd	Zentrum	Mitte	2007-10-26T00:00:00	5.171462e+06	POLYGON ((13.36953 52.49887, 13.37036 52.49951...
1	re_bezirksregion.010112	010112	Regierungsviertel	Polygon	Regierungsviertel	Zentrum	Mitte	2007-10-26T00:00:00	2.691032e+06	POLYGON ((13.37686 52.51964, 13.37686 52.51947...

Figure 1: Imported Geojson Data via the geopandas package.

Crime, Population-Data: The crime data and also the population data (Kriminalitätsatlas Berlin and Amt für Statistik Berlin-Brandenburg) needed some pre-processing via Notepad++ before, because the data from both sources had to be converted to the UTF-8 unicode format for easier handling.

Foursquare-Data: The Foursquare data was called via the Foursquare API.

1.4 Feature Selection

All necessary features were provided inside the data itself. Since the crime data only contained the total number of crimes, it was necessary to calculate an HZ-Index.

$$HZ - Index = \frac{Crimes}{Population} * 10^5$$

This HZ-Index describes the amount of crimes per 100000 inhabitants of the respective district area and is therefore a more suitable factor for the definition of higher crime activity in a district.

2 Data Analysis / Methodology

Berlin is a city with two kinds of region descriptions (neighborhoods and district areas) besides the boroughs. The following figure shows both regions plotted via their longitude and latitude values. From the figure itself it es visible that the neighborhood and the districts are in some areas the same but they differ in other ones. Generally Berlin consists of 12 boroughs, 96 neighborhoods and 138 district areas. Since the crime and population data is provided with respect to the district-areas, the following data evaluation and description is related towards these.

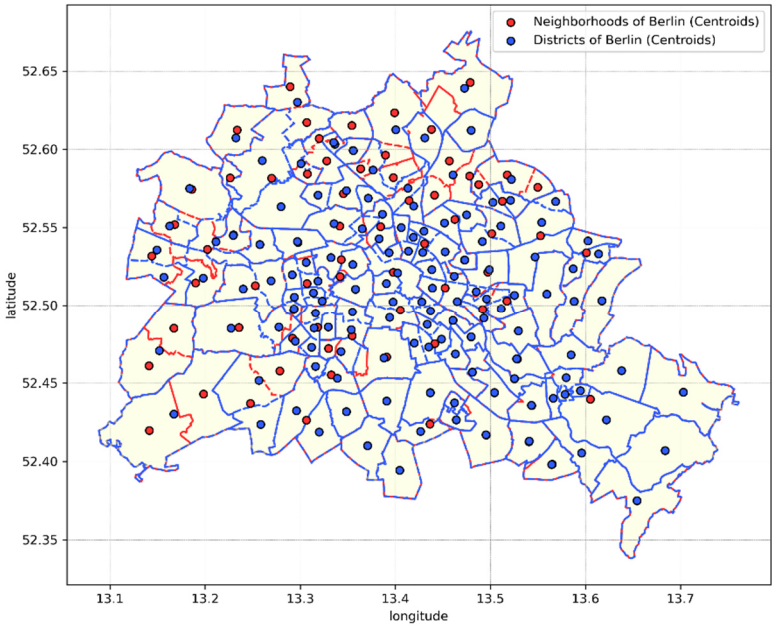

Figure 2: Map of Berlin with neighborhoods and district-areas.

2.1 Average Income of the Boroughs

First the net household income of the different boroughs of Berlin was checked. As indicated in Figure 3 the income differs in a range of 2554 to 3762 €. The top 3 boroughs with the highest household incomes are Steglitz-Zehlendorf (3763 €), Pankow (3520 €) and Charlottenburg-Wilmersdorf (3484 €) and the boroughs with the lowest average income are Spandau (2687 €), Lichtenberg (2640 €) and Neukölln (2554 €). Although the data itself shows that there are districts with a higher household income, it should be noted that the household income is not differentiated into couples, families or even singles. The household income plays therefore only a subordinate role as a description of a good residential area.

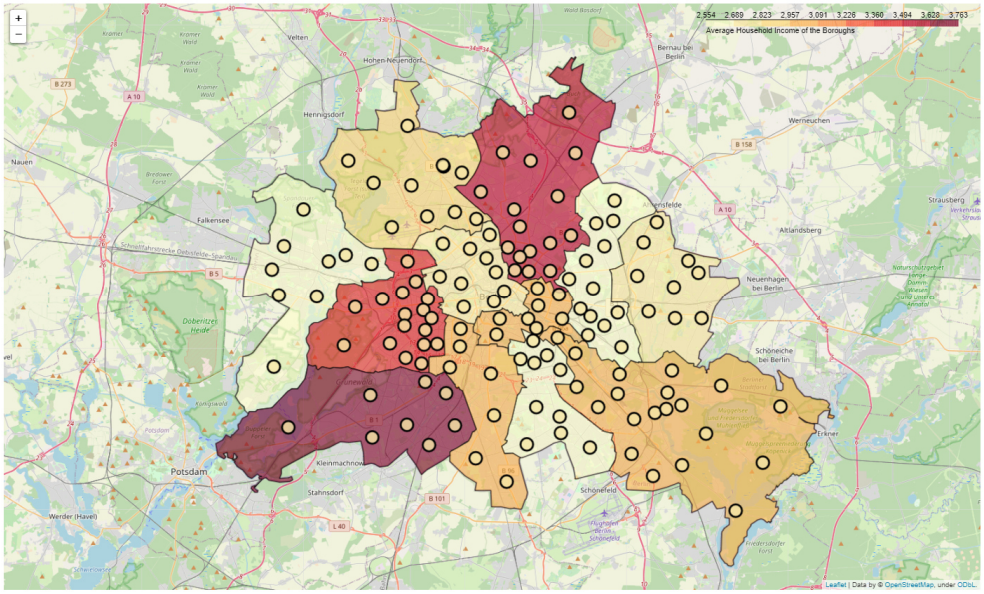

Figure 3: Borough net household income.

2.2 Population of the different District Areas

One main factor of the investigation is to decide if a region is lively or not. So, the amount of inhabitants seems to be a good factor. A region with a high number of people is more likely to be vibrant and active than one with a low number. The analysis of the different amounts of inhabitants is therefore shown in the figure below. As was to be expected, especially regions near the centre of Berlin have a large number of inhabitants. Furthermore, it can also be seen that the boroughs Tempelhof-Schöneberg (south), Steglitz Zehlendorf (south-west) and Neukölln (south-east) have districts with a large numbers of inhabitants. The district areas with the most inhabitants are located close to the center of Berlin and are mainly Tempelhofer Vorstadt (borough: Tempelhof-Schöneberg), Tempelhof (borough: Tempelhof-Schöneberg) and Alexanderplatz (borough: Mitte).

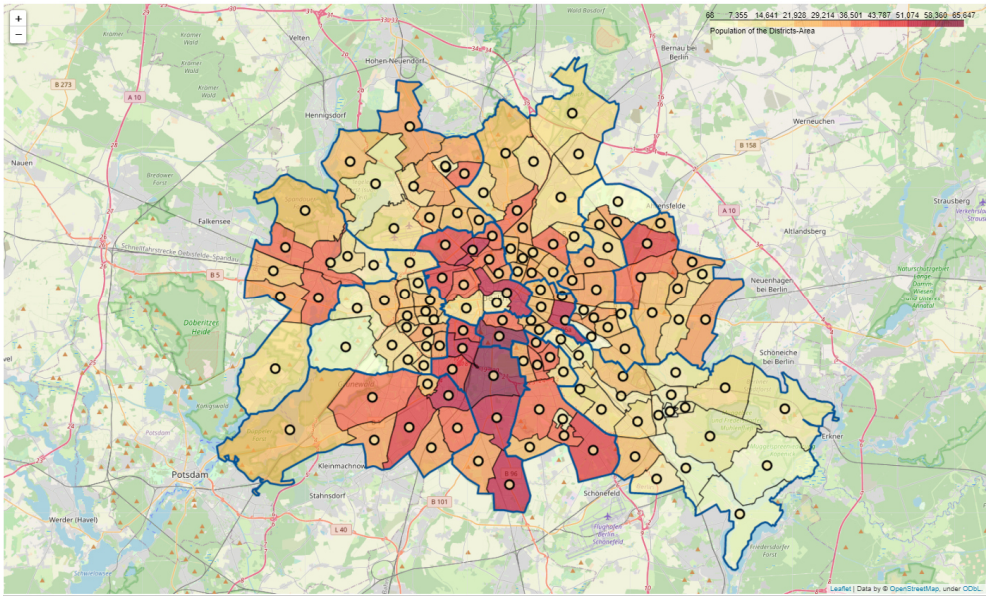

Figure 4: Mapping of the population of Berlin.

2.3 Crime Count of the City

The crime rate of a region is best represented by HZ Index. The following illustration therefore shows the distribution of the HZ-Index among the different districts within Berlin.

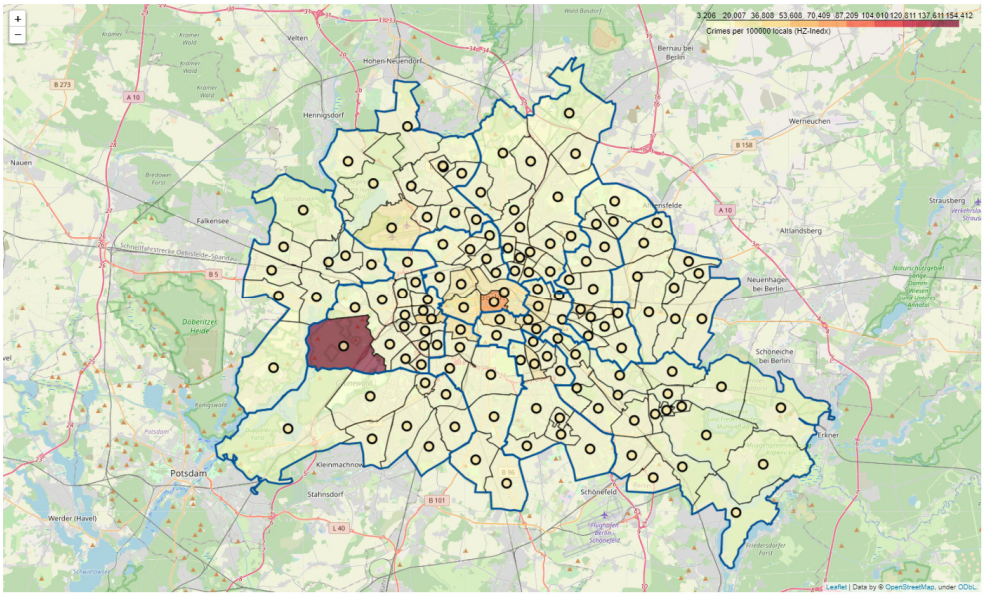

Figure 5: HZ-index of the districts.

It seems that one region has a really high crime-index, but with a closer look the data shows that this region has only 65 inhabitants and a crime count of 105. With an even more closer look it is also visible that this district area mostly covers only Green spaces and parks. Therefore, it is more suitable to skip this region from the mapping in order to provide a clearer result.

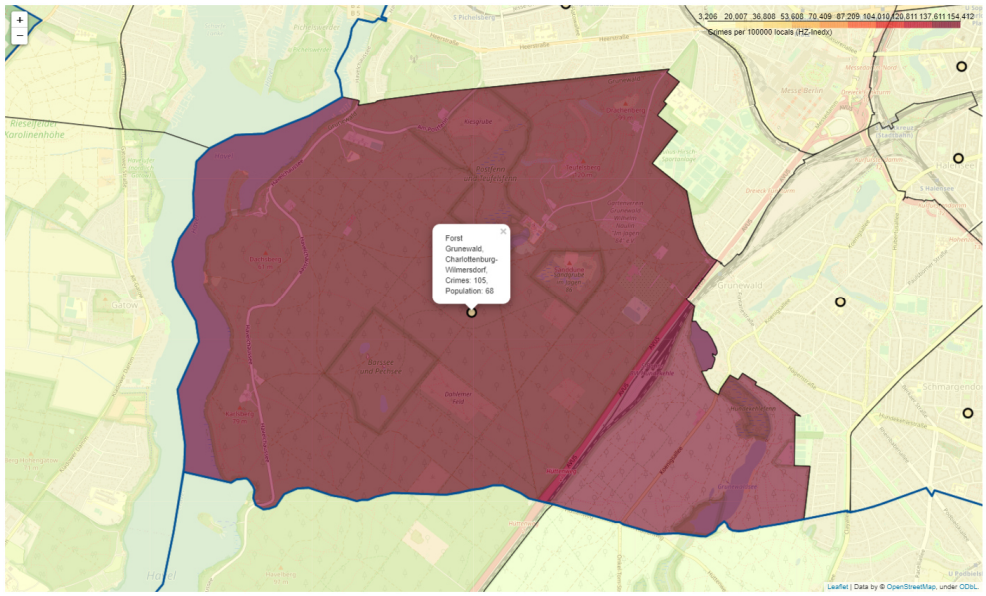

Figure 6: HZ-Index of Fort Grunewald.

After setting the HZ-index of Fort-Grunewald to an “Nan” value the following figure shows a better picture of the situation.

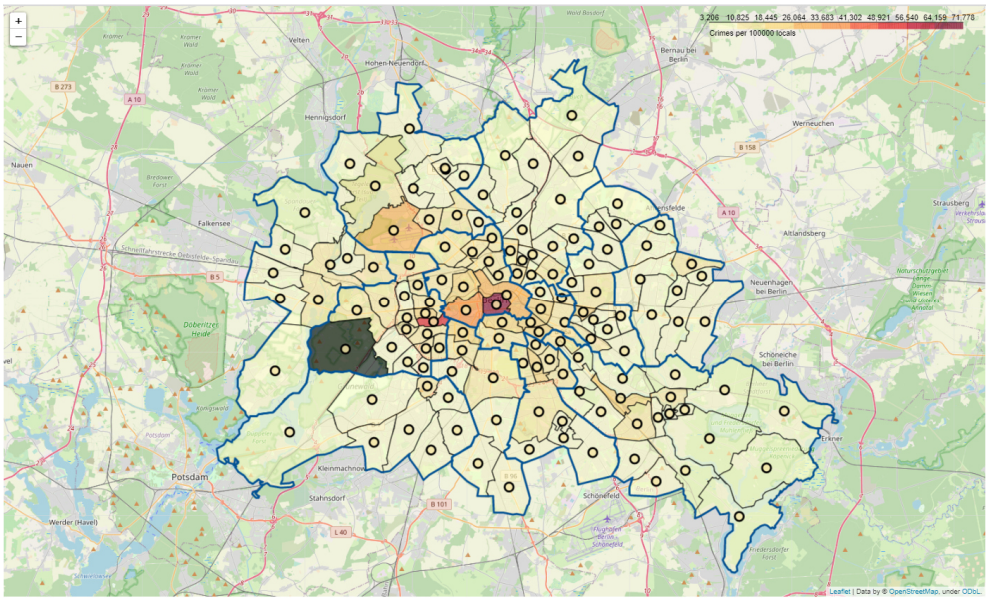

Figure 7: HZ-index of the district areas (without Fort Grunewald).

It is clear that the center of Berlin shows the highest HZ-index. The areas Regierungsviertel (HZ: 71778), Kurfürstendamm (HZ: 52312) and Tiergarten Süd (HZ: 34507) are the districts with a high crime-index. Also, to mention is that the district West 1 – Tegel-Süd (HZ: 31293) located north-west of Berlin has a high index. But the main reason here will be the airport (Berlin-Tegel), which is located in this region.

2.4 Crime Index and Population Normalization

In order to make an initial pre-selection of the areas for the question of a good living environment, the HZ-index and also the district population were normalised (min-max normalisation). The idea behind this was to select the regions based on both factors.

DISTRICT_AREA	BOROUGH	CRIME_COUNT	POPULATION	CRIME_per_100000_POP	normPOPULATION	normCRIME_per_100000_POP
Regierungsviertel	Mitte	9146	12742	71778.370742	0.193263	1.000000
Kurfuerstendamm	Charlottenburg-Wilmersdorf	8120	15522	52312.846283	0.235655	0.716130
Tiergarten Sued	Mitte	5171	14985	34507.841175	0.227466	0.456476

Figure 8: Normalization of the HZ-index and Population.

By normalising both factors, it was then possible to describe and identify the first areas. For this purpose, a limit was set for both factors which was then used to make the selection. In this case all areas with a normalized crime rate of < 0.4 and a population > 0.6 were selected as interesting areas. With these indicated and extracted areas the following step was to search for the venues via the Foursquare API.

```
crime_pop_reduced = gdf_District_crime_pop_District_Area
crime_factor=0.4
population_factor=0.6
crime_pop_reduced = crime_pop_reduced[(crime_pop_reduced['normCRIME_per_100000_POP']<=crime_factor) & (crime_pop_reduced['normPOPULATION']>=population_factor)]
print("{} Districts are matching a lower crime factor of {} and a higher population factor of {} ".format(crime_pop_reduced.shape[0],crime_factor,population_factor))
crime_pop_reduced.sort_values(by="CRIME_per_100000_POP").head(3)
```

```
25 Districts are matching a lower crime factor of 0.4 and a higher population factor of 0.6
```

Figure 9: Pre-selection of low crime high population areas.

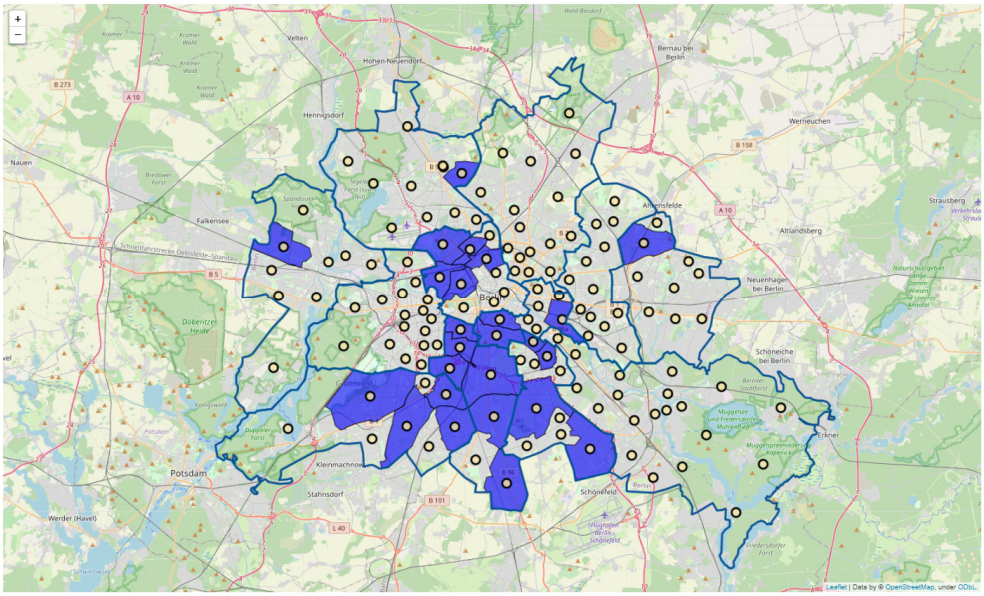

Figure 10: Pre-selected areas (crime factor < 0.4 / population > 0.6)

2.5 Foursquare Data for Determination of a Lively and Save District Area

Using the Foursquare API, the venues within a radius of 1000 m and a limit of 100 numbers of venues from the district centers were selected. As expected, there are more venues near the centre of Berlin.

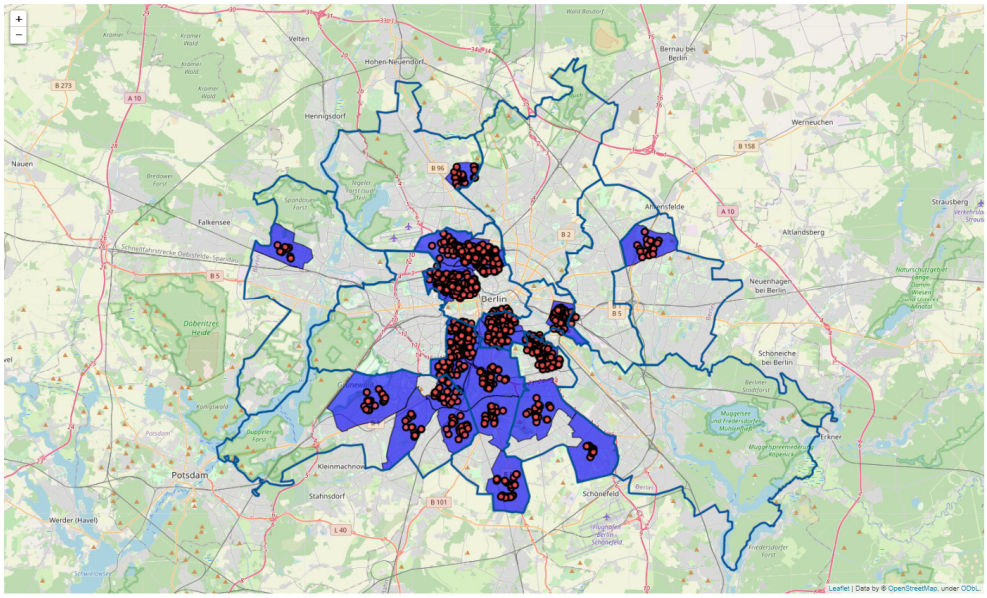

Figure 11: Venues of the district regions.

After the one-hot encoding of the individual locations, a first list covering the most frequently preserved locations for the respective regions was created.

	DISTRICT_AREA	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Albrechtstrasse	Supermarket	Café	Chinese Restaurant	Park	Trattoria/Osteria
1	Britz	Supermarket	Bakery	Liquor Store	Fast Food Restaurant	Park
2	Brunnenstr. Nord	Bakery	Supermarket	Vietnamese Restaurant	Café	Hotel

Figure 12: Most common venues near the district centers (first three districts and their 5 most common venues are shown).

2.6 k-means Clustering

The determined locations served as a basis for the k-mean cluster formation. At the beginning an elbow plot was made to analyse the cluster behaviour of different sets of cluster points and to get a basis for the choice of the optimal amount of cluster points.

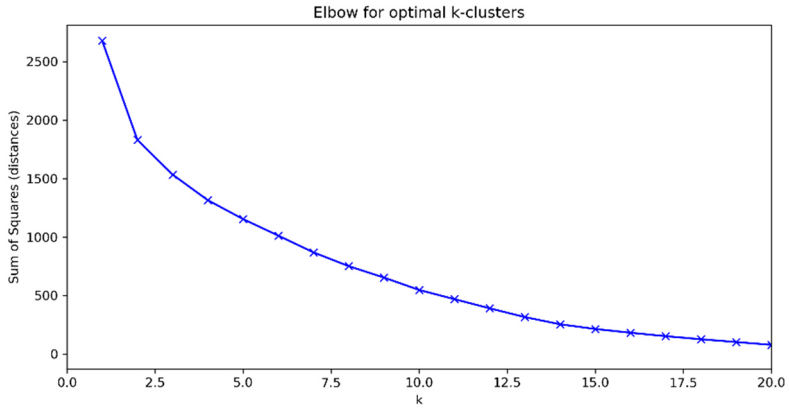

Figure 13: Elbow-plot.

Based on the plot, it is difficult to decide how many clusters to take. Most likely the range k=4-6 seems to be a good choice. Therefore, 5 clusters were chosen for the further data evaluation. The calculation of the clusters then provided the location of the respective cluster and thus an allocation to the districts which are similar.

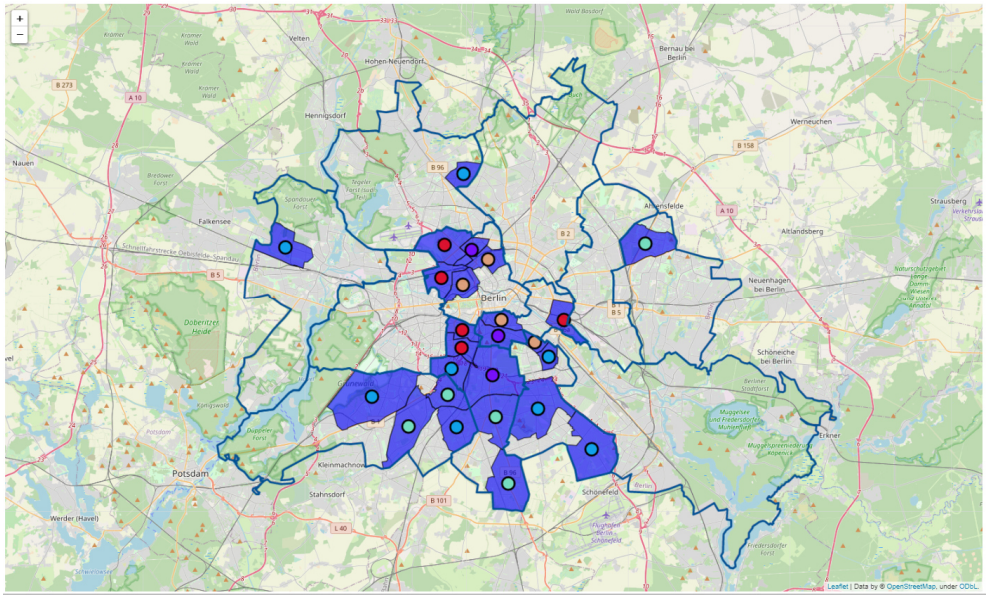

Figure 14: Cluster of the districts.

In a further step, the dependence of the population and crime on cluster indexing as box plots was investigated. The respective factors were plotted against the cluster index. The following illustration

shows that the population and the crime rate also show a dependency with respect to the indexed clusters.

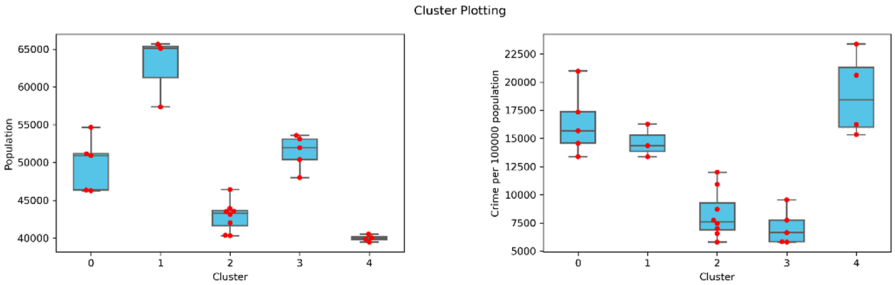

Figure 15: Boxplots of the population and crime-index.

On the basis of this representation, it was possible to make a further selection according to the crime-index and also population size. Since a low crime rate is beneficial for a good living environment, cluster 3 was chosen. It is visible that in cluster 3 the median for crime is lowest one of the dataset and yet there is a medium population size too.

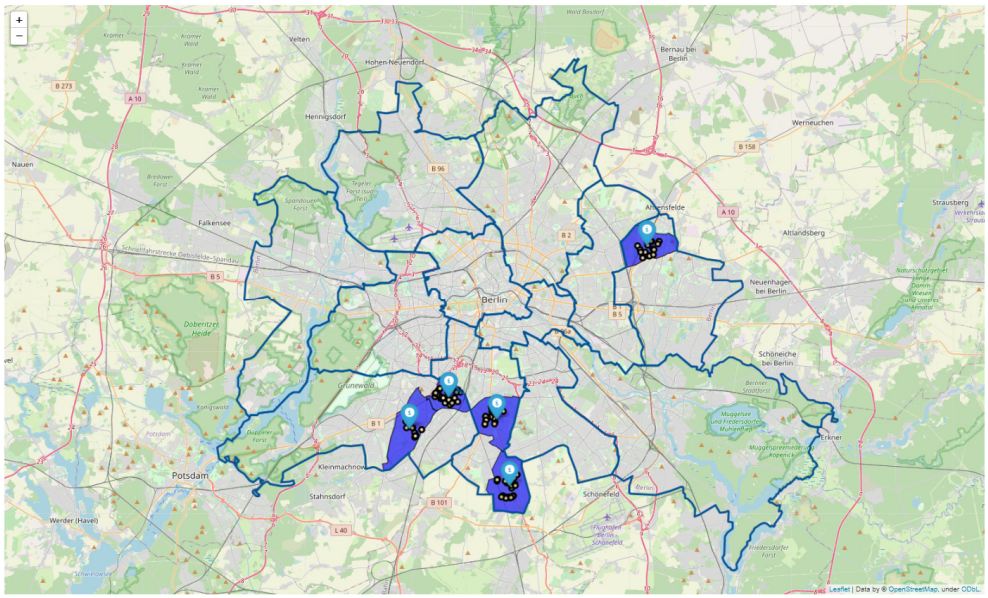

Figure 16: District areas choosen from the k-mean clustering.

	Cluster_Labels	DISTRICT_AREA	POPULATION	CRIME_per_100000_POP	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	3	Albrechtstrasse	53582	5792.990183	Supermarket	Café	Chinese Restaurant	Park	Trattoria/Osteria
3	3	Drakestrasse	50385	6628.957031	Supermarket	Bakery	Bus Stop	Café	BBQ Joint
8	3	Lichtenrade	51955	5824.271004	Supermarket	Soccer Field	Bakery	Doner Restaurant	Light Rail Station
10	3	Mariendorf	53106	7739.238504	Supermarket	Bank	German Restaurant	Pool	Bakery
11	3	Marzahn Mitte	48008	9550.491585	Tram Station	Supermarket	Cultural Center	Shopping Mall	Park

Figure 17: Most common venues of the cluster.

3 Results: Sum of Venues

After the most interesting areas were indexed, a further classification was made based on the frequency of the venues. Since a particularly lively environment is preferred by the business question, the total sum of the locations can play a key role. With the help of the total sum a further selection can be made to answer the question which of these regions would be the most vibrant one. The three top relevant district areas are therefore Albrechtsstrasse (borough: Steglitz-Zehlendorf), Lichtenrade (borough: Tempelhof-Schöneberg) and Mariendorf (borough: Tempelhof-Schöneberg).

```
The top three relevant district areas are Albrechtstrasse, Lichtenrade and Mariendorf
```

	Cluster_Labels	DISTRICT_AREA	POPULATION	CRIME_per_100000_POP	SUM_VENUES	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	3	Albrechtstrasse	53582	5792.990183	31	Supermarket	Café	Chinese Restaurant	Park	Trattoria/Osteria
1	3	Lichtenrade	51955	5824.271004	28	Supermarket	Soccer Field	Bakery	Doner Restaurant	Light Rail Station
2	3	Mariendorf	53106	7739.238504	28	Supermarket	Bank	German Restaurant	Pool	Bakery

Figure 18: Most relevant regions after summing the venues.

A second option, however, is to set preferences, because maybe you don't need a football field (Mariendorf) or a park (Albrechtsstrasse) (you can't understand, but maybe you do) nearby. For this reason it can also be useful to define the necessary places in advance. Therefore in this example a list of places was given (Chinese Restaurant, Café, Shopping Mall, Gas Station, Drugstore, Pharmacy, Grocery Store, Organic Grocery, Gym / Fitness Center, Supermarket, Bakery, German Restaurant). The new summation of all places then represents the personally best fitting regions. The dataframe is showing the same regions as before, but as can be seen the preferences of the regions hat changed.

```
The top three relevant district areas are Lichtenrade, Albrechtstrasse and Mariendorf
```

	DISTRICT_AREA	POPULATION	CRIME_per_100000_POP	SUM_VENUES	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Lichtenrade	51955	5824.271004	12	Supermarket	Bakery	German Restaurant	Gym / Fitness Center	Pharmacy
1	Albrechtstrasse	53582	5792.990183	11	Supermarket	Café	Chinese Restaurant	Bakery	Organic Grocery
2	Mariendorf	53106	7739.238504	9	Supermarket	German Restaurant	Bakery	Drugstore	Chinese Restaurant

Figure 19: Top 3 of the best fitting regions to live in upon personal interests.

4 Discussion and Conclusion

The results of this data investigation show that it is possible to determine the most relevant district areas of Berlin for a relatively save and lively neighborhood to live in. After predefining areas with a low (normalized) crime index and high (normalized) population index, these regions could be used to obtain the necessary venues out of Foursquare. After clustering the data via the *KMeans* method from *sklearn* the clusters could be assigned. Plotting the crime index and the population against these clusters helped to investigate both factors and to make a decision which cluster would be the most interesting one. Finally the sum of venues and its reduced version showed which district area would be preferable to live in. In generel this method can help to identify the most relevent regions in a city, but since it is strongly dependent on the quality of the venue data, this data plays a crucial role. Upon investigation of the venue positions it is clear that for Berlin maybe many positions and datapoints are missing and therefore the reliability of the process is influenced. Nevertheless, this procedure is maybe a good starting point to investigate other cities as well. This investigation showed that it is possible to search for good areas to live inside a city with a focus upon crime, population and venues.