Verifying Berlin neighborhoods

1. Introduction

1.1. Background

Berlin the capital of Germany is a fast growing city. A lot of money is invested to construct new living and office buildings, attractions, shopping malls, etc. Unfortunately this is mainly done in areas where the life is already very attractive and expensive. But what about the other boroughs? In this case study the Berlin government has a budget to support investors for a project to progress undeveloped neighborhoods in terms of building new flats, playgrounds, family centers, parks, etc. So, whatever helps people feeling better in their neighborhood or raising the neighborhoods attraction for people moving there. Therefore in this analysis project all 96 neighborhoods of Berlin will be taken into account mapped to their boroughs.

1.2. Problem

The problem for the government is to identify the boroughs having neighborhoods which benefit most from this financial support.

1.3. Interest

The Berlin government is interested in developing all neighborhoods to a certain standard to satisfy people living and working there. Furthermore it is better to have an similar number of inhabitants per km² all over the city. Currently living and working areas are mostly split. Combining those when having attractive life standards may also lead to a reduction of the traffic pollution when more people want to live and work in the same neighborhood.

2. Data acquisition and selection

2.1. Data sources

The first thing which is required is a list of Berlins neighborhoods. This can be retrieved by the following Wikipage: https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins
In the middle of this page one can find a table which provides the required information (date: 07/2019).
Extract [number | neighborhood | borough | area km² | inhabitants | inhabitants per km²]:

Nr.	Ortsteil	Bezirk	Fläche (km²)	Einwohner[2] (30. Juni 2019)	Einwohner pro km²
101	Mitte	Mitte	10,70	101.932	9526
102	Moabit	Mitte	7,72	79.512	10.299
103	Hansaviertel	Mitte	0,53	5.894	11.121
104	Tiergarten	Mitte	5,17	14.753	2854
105	Wedding	Mitte	9,23	86.688	9392
106	Gesundbrunnen	Mitte	6,13	95.393	15.562
201	Friedrichshain	Friedrichshain-Kreuzberg	9,78	134.900	13.793
202	Kreuzberg	Friedrichshain-Kreuzberg	10,40	154.862	14.891
301	Prenzlauer Berg	Pankow	11,00	164.593	14.963
302	Weißensee	Pankow	7,93	53.737	6776
303	Blankenburg	Pankow	6,03	6.865	1138

Resulting dataframe

	NH-number	Neighborhood	Borough	Area (km²)	Inhabitants	Inhabitants per km²
0	0101	Mitte	Mitte	10,70	101.932	9526
1	0102	Moabit	Mitte	7,72	79.512	10.299
2	0103	Hansaviertel	Mitte	0,53	5.894	11.121
3	0104	Tiergarten	Mitte	5,17	14.753	2854
4	0105	Wedding	Mitte	9,23	86.688	9392

```
neighborhoods_Berlin.shape
```

```
(96, 9)
```

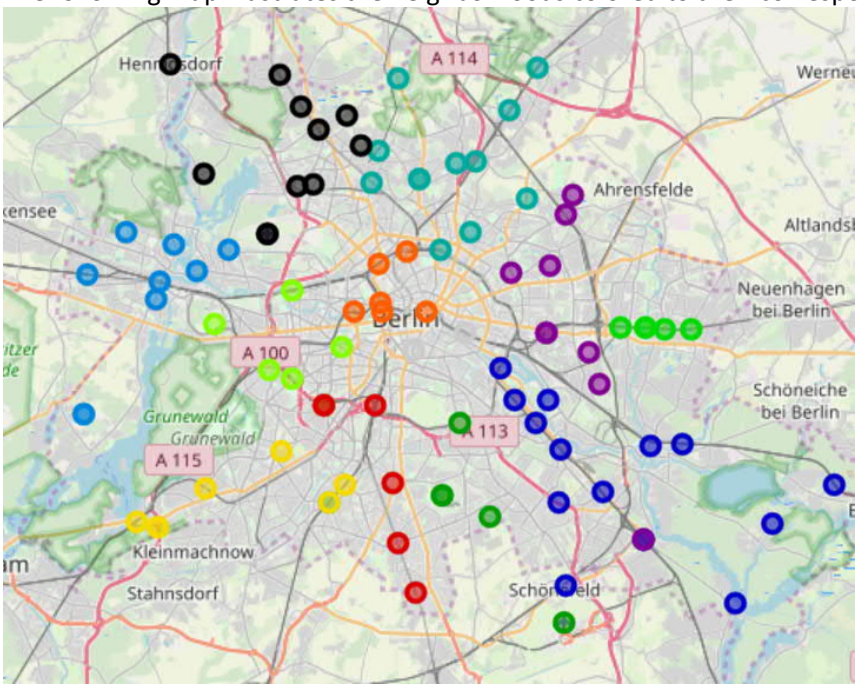
For judging the current attractiveness of a neighborhood the foursquare dataset is used. There the number and categories of venues being in a certain radius are collected. Therefore it is also required to have the geographic coordinates of the neighborhoods centers. This will be retrieved by using the geolocator package of Python.

	NH-number	Neighborhood	Borough	Area (km²)	Inhabitants	Inhabitants per km²	Latitude	Longitude
0	0101	Mitte	Mitte	10,70	101.932	9526	52.519982	13.404159
1	0102	Moabit	Mitte	7,72	79.512	10.299	52.524945	13.369661
2	0103	Hansaviertel	Mitte	0,53	5.894	11.121	52.519985	13.348070
3	0104	Tiergarten	Mitte	5,17	14.753	2854	52.520226	13.370487
4	0105	Wedding	Mitte	9,23	86.688	9392	52.542787	13.367000

```
neighborhoods_Berlin.shape
```

```
(96, 8)
```

The following map illustrates the neighborhoods colored to their corresponding boroughs.



2.2. Data manipulation

The plain data set of neighborhoods is already convenient for a direct usage.
So there are 96 neighborhoods in the dataset available having the six columns coming out of the wiki table (excl. index) including the geographic coordinates.

In my case study I additionally define a dynamic radius in dependency of the area size of a neighborhood to retrieve the foursquare data. So small sized neighborhoods get a small radius and large neighborhoods get a large radius. This is done to avoid an overlapping of small neighborhoods but also to explore almost the complete neighborhood of very large sized neighborhoods.
Therefore the ('Area (km²)' column needs to be redefined as float (was object) to get a dot separated value instead of a comma separated one.

	NH-number	Neighborhood	Borough	Area (km²)	Inhabitants	Inhabitants per km²	Latitude	Longitude
0	101	Mitte	Mitte	10.70	101932	9526	52.519982	13.404159
1	102	Moabit	Mitte	7.72	79512	10299	52.524945	13.369661
2	103	Hansaviertel	Mitte	0.53	5894	11121	52.519985	13.348070
3	104	Tiergarten	Mitte	5.17	14753	2854	52.520226	13.370487
4	105	Wedding	Mitte	9.23	86688	9392	52.542787	13.367000

2.3. Data creation

The calculation of the individual radius is done via the min-max normalization process of the area size ('Area (km²)'), times the mean of all areas, times 1000 to transfer the result to a km – value. Finally I divide this outcome by two to receive a neighborhood individual radius.

Formula: Radius = MinMaxNormalization['Area (km²)'] * mean(['Area (km²)']) * 1000 / 2

	NH-number	Neighborhood	Borough	Area (km²)	Inhabitants	Inhabitants per km²	Latitude	Longitude	Radius
91	1207	Waidmannslust	Reinickendorf	2.3	10958	4764	52.606272	13.321194	238.0
92	1208	Lübars	Reinickendorf	5.0	5174	1035	52.612997	13.342222	602.0
93	1209	Wittenau	Reinickendorf	5.9	24306	4120	52.580149	13.316386	723.0
94	1210	Märkisches Viertel	Reinickendorf	3.2	40258	12581	52.598875	13.354212	359.0
95	1211	Borsigwalde	Reinickendorf	2.0	6826	3413	52.579655	13.304926	197.0

2.4. Foursquare data acquisition

So with the help of the neighborhood coordinates and the individual radius value the foursquare dataset is called to explore the area and get a list of all available venues limited to 100. The received information is the neighborhood with its coordinates combined with the venue names, coordinates and categories.

Resulting dataset

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mitte	52.519982	13.404159	Buchhandlung Walther KÃ¶nig	52.521301	13.400758	Bookstore
1	Mitte	52.519982	13.404159	Kuppelumgang Berliner Dom	52.518966	13.400981	Scenic Lookout
2	Mitte	52.519982	13.404159	Radisson Blu	52.519561	13.402857	Hotel
3	Mitte	52.519982	13.404159	Fat Tire Bike Tours	52.521233	13.409110	Bike Rental / Bike Share
4	Mitte	52.519982	13.404159	Lustgarten	52.518469	13.399454	Garden

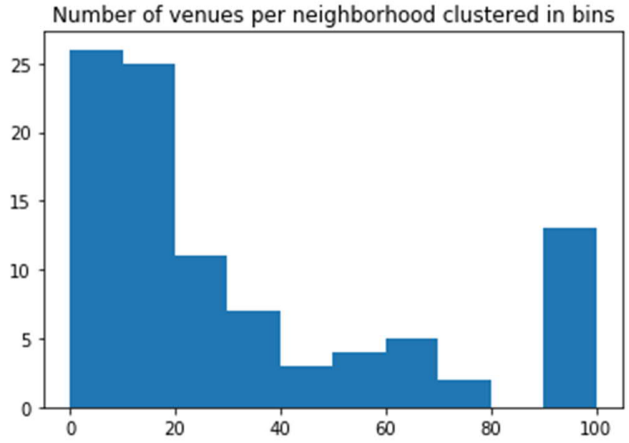
This dataset contains of 3071 venues.

3. Data exploration and analysis

3.1. Number of venues clustered in bins

First of all the number of venues per neighborhood is counted. For neighborhoods which do not have any venue a 0 is entered instead of NaN because they are statistically significant and need to be considered in the further analysis.

```
(array([26., 25., 11.,  7.,  3.,  4.,  5.,  2.,  0., 13.])
<function matplotlib.pyplot.show(*args, **kw)>
```

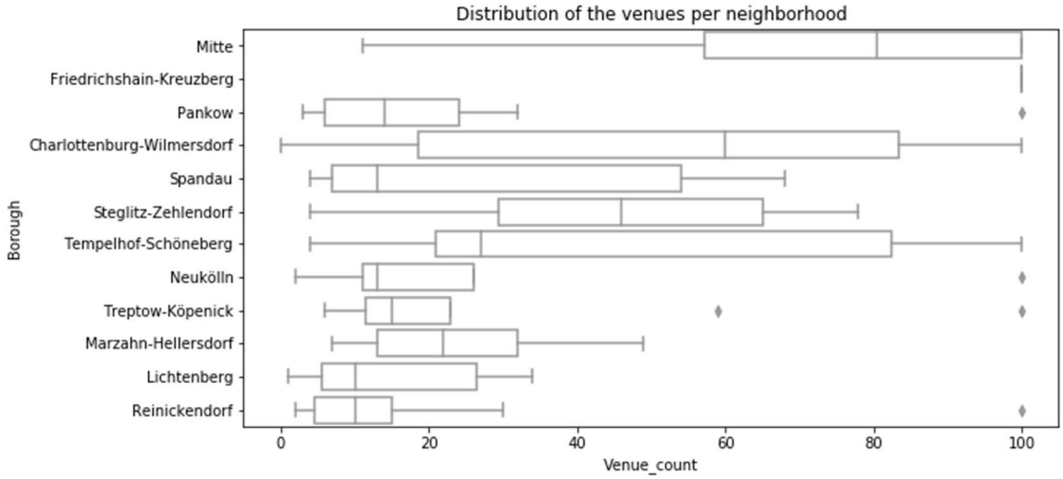


In the illustration can be seen that more than the half of all neighborhoods (26 + 25 = 51 of 96) have less than 20 venues in their area. This is one indication which is worth to analyze more deep.

3.2. Number of neighborhood venues per borough

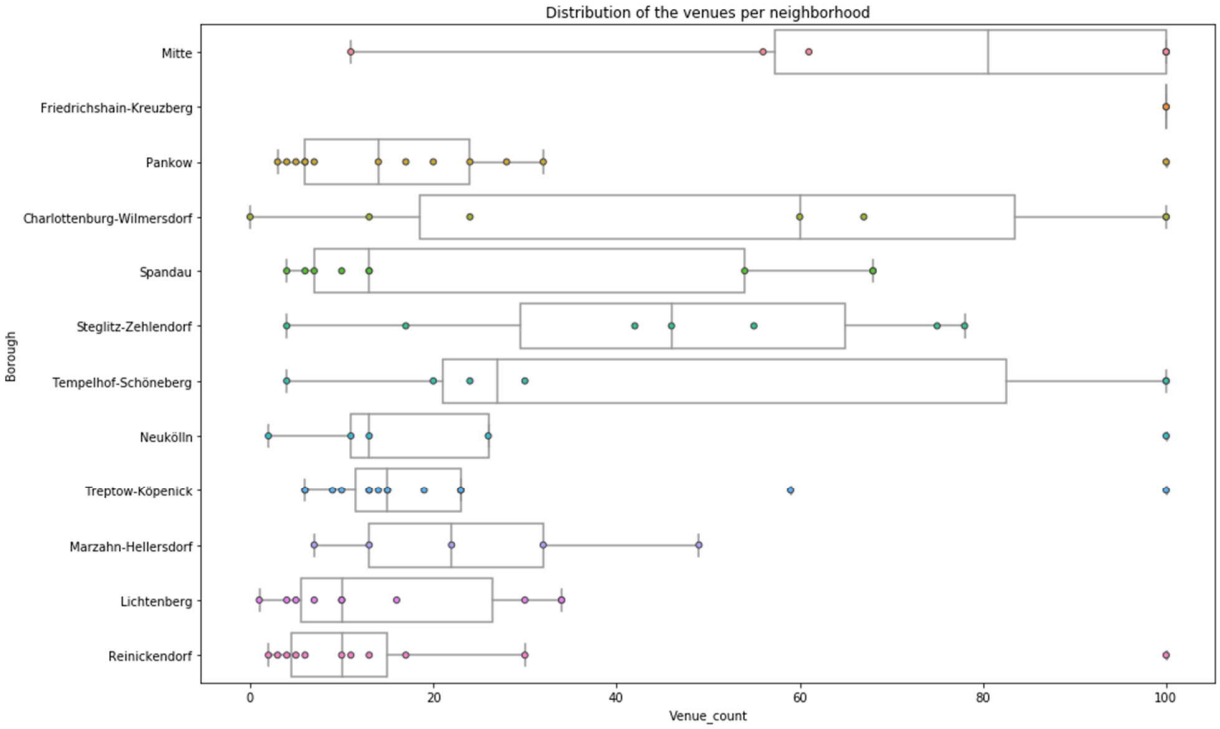
A related interesting point is to get the number of venues per neighborhood grouped by the borough with the help of a boxplot and stripplot.

Boxplot:



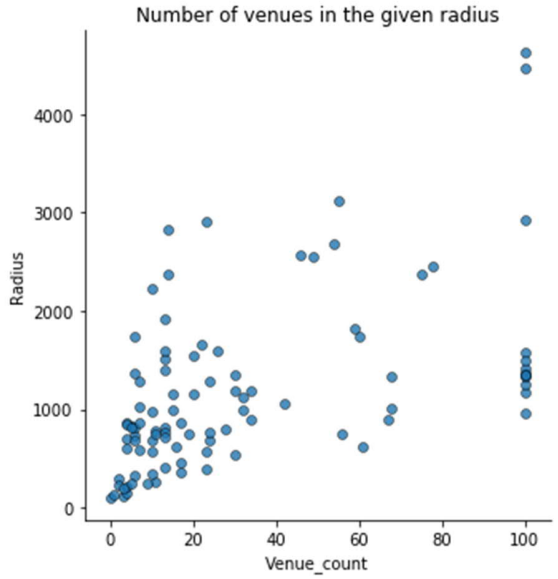
Here one can see that the upper quartile only of Reinickendorf is within the range of 20 venues per neighborhood except the outliers. This means almost all neighborhoods of Reinickendorf have less than 20 venues. Still five boxes have at least their median within this range which are Lichtenberg, Treptow-Köpenick, Neukölln, Pankow and Spandau. So for these boroughs the average venues per neighborhood is less than 20.

To make the previous analysis easier to follow a stripplot is added to the box plot. So the single data points of the boxes are made visible and confirm the previous analysis.

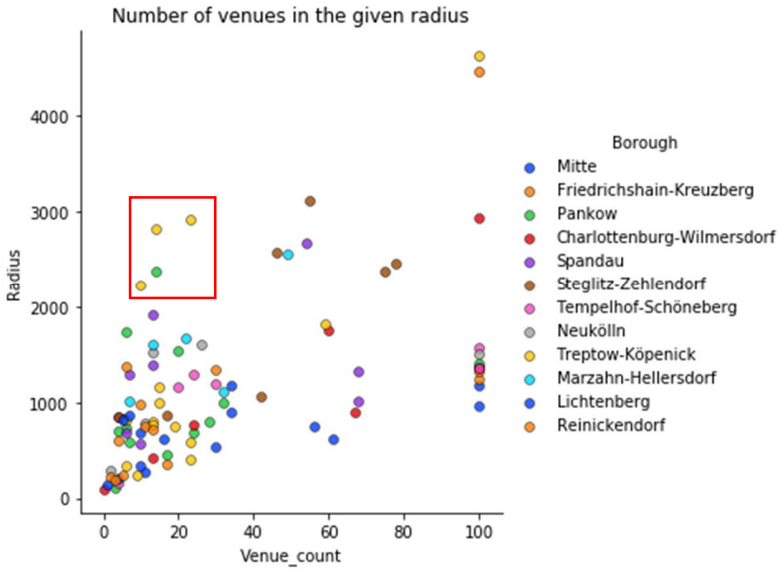


3.3. Number of venues vs radius

One relation which should be explored is if there is a relationship between the number of venues and the given individual radius per neighborhood. The following illustration is quite what could have expected. The larger the radius the more venues are listed. But there are outliers having a big radius but only less venues.



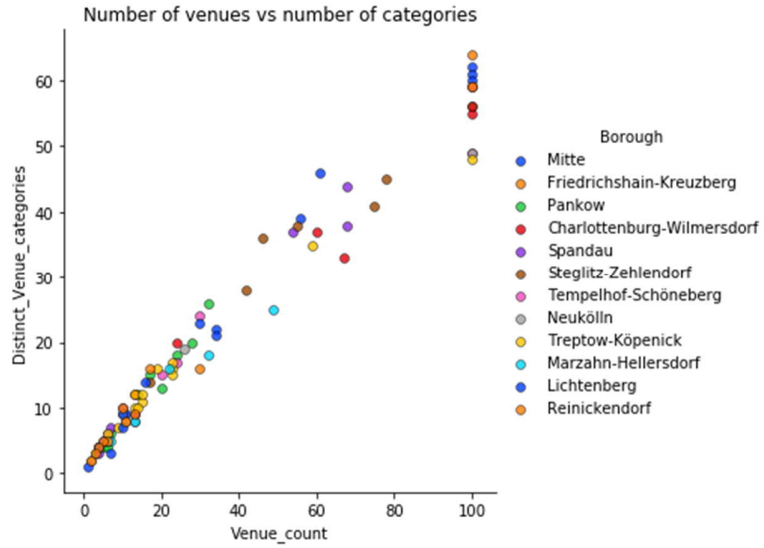
Let us investigate to which boroughs these outliers belong to by coloring the scatter plot.



One can see the neighborhoods having a radius of more than 2000m but only approximately 20 venues belong to Pankow (one) and Treptow-Köpenick (three).

3.4. Check diversity of the neighborhoods

Surely it nice to have a lot of venues being available. But an interesting point to judge the attractiveness is the diversity of the existing venues.



This picture is quite clear. There is a linear relationship between the number of venues and the number of venue categories being available. So it can be said the more venues a neighborhood has the more diverse the neighborhood is.

3.5. Ratio of neighborhoods

The following table illustrates the actual numbers of neighborhoods having less than 20 venues.

	Borough	Num_of_less_20_venues
10	Treptow-Köpenick	10
6	Reinickendorf	9
5	Pankow	8
1	Lichtenberg	7
7	Spandau	6
4	Neukölln	3
0	Charlottenburg-Wilmersdorf	2
2	Marzahn-Hellersdorf	2
8	Steglitz-Zehlendorf	2
3	Mitte	1
9	Tempelhof-Schöneberg	1

Treptow-Köpenick is obviously having the most neighborhoods (10) with less than 20 venues followed by Reinickendorf (9). To correctly interpret this number it is necessary to also have the total number of neighborhoods in relation to that.

Therefore the following table additionally shows the total number of neighborhoods per borough.

	Borough	Num_of_neighborhoods	Num_of_less_20_venues	Ratio: tot NH / <20 NH
7	Reinickendorf	11	9.0	0.818182
2	Lichtenberg	10	7.0	0.700000
11	Treptow-Köpenick	15	10.0	0.666667
8	Spandau	9	6.0	0.666667
6	Pankow	13	8.0	0.615385
5	Neukölln	5	3.0	0.600000
3	Marzahn-Hellersdorf	5	2.0	0.400000
0	Charlottenburg-Wilmersdorf	7	2.0	0.285714
9	Steglitz-Zehlendorf	7	2.0	0.285714
4	Mitte	6	1.0	0.166667
10	Tempelhof-Schöneberg	6	1.0	0.166667
1	Friedrichshain-Kreuzberg	2	NaN	NaN

Here the ratio is calculated which leads to the point that it is again Reinickendorf having the most neighborhoods with less than 20 venues compared to the total number of their neighborhoods, namely 81,8%. This number is also far ahead of second place Lichtenberg with 70%.

3.6. Remember the area size

If we see Reinickendorf having the worst ratio we also need to consider the total area of each borough. Because if Reinickendorf is a very small borough it would be acceptable if the number of venues is correspondingly small. So in the following table the total area of the boroughs is summed up and multiplied with the ratio determined before.

	Borough	Num_of_neighborhoods	Num_of_less_20_venues	Ratio: tot NH / <20 NH	Area (km^2)	Ratio * <20 NH
0	Reinickendorf	11	9.0	0.818182	89.40	7.363636
2	Treptow-Köpenick	15	10.0	0.666667	165.70	6.666667
4	Pankow	13	8.0	0.615385	103.26	4.923077
1	Lichtenberg	10	7.0	0.700000	52.02	4.900000
3	Spandau	9	6.0	0.666667	91.90	4.000000
5	Neukölln	5	3.0	0.600000	44.91	1.800000
6	Marzahn-Hellersdorf	5	2.0	0.400000	61.71	0.800000
7	Charlottenburg-Wilmersdorf	7	2.0	0.285714	64.62	0.571429
8	Steglitz-Zehlendorf	7	2.0	0.285714	102.47	0.571429
9	Mitte	6	1.0	0.166667	39.48	0.166667
10	Tempelhof-Schöneberg	6	1.0	0.166667	53.08	0.166667
11	Friedrichshain-Kreuzberg	2	NaN	NaN	20.18	NaN

But this illustrations also confirms that even depending on the area Reinickendorf has most less venues.

4. Clustering models

The data analysis gives already a very interesting insight which borough requires the financial support at most. In this chapter a clustering model with the help of kmeans is applied to verify the resulting clusters if they match the previous analysis.

4.1. Data preparation

For the modelling procedure we use the venue categories belonging to each neighborhood including the inhabitants per km². The categories are one hot encoded which lead to the following data frame having 3071 rows but 329 columns.

Extract:

	Neighborhood	Inhabitants per km²	ATM	Adult Boutique	African Restaurant	Airport	Airport Lounge
0	Mitte	9526.0	0	0	0	0	0
1	Mitte	9526.0	0	0	0	0	0
2	Mitte	9526.0	0	0	0	0	0
3	Mitte	9526.0	0	0	0	0	0
4	Mitte	9526.0	0	0	0	0	0
...	...	...	...	...	...	...	...
3066	Märkisches Viertel	12581.0	0	0	0	0	0
3067	Märkisches Viertel	12581.0	0	0	0	0	0
3068	Borsigwalde	3413.0	0	0	0	0	0
3069	Borsigwalde	3413.0	0	0	0	0	0
3070	Borsigwalde	3413.0	0	0	0	0	0

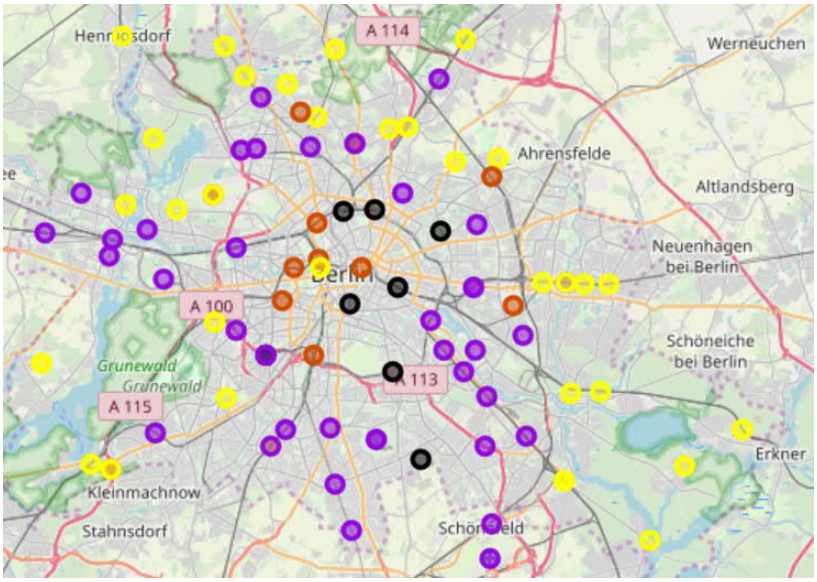
Afterwards the mean values are calculated which serve as basis for the kmeans clustering method. In this case study 4 different clusters are taken for the kmeans procedure.

4.2. Clustering results

The following clusters result from the classification model.

```
2.0    38
0.0    36
3.0    12
1.0     9
```

The following illustration shows the clusters within the map of Berlin.



4.3. Cluster analysis

The map shows the four resulting clusters. Two clusters are mostly centered within Berlin (black and brown). If one assume that the center is the most attractive area the focus should be on the remaining two clusters (cluster 0 and cluster 2) which can be defined as ‘support required’.

The following table shows how many neighborhoods of cluster 0 and cluster 2 belong to each borough sorted descending.

	Borough	Cluster_Labels_Count	Num_of_neighborhoods
0	Treptow-Köpenick	15	15
3	Spandau	9	9
6	Marzahn-Hellersdorf	5	5
5	Steglitz-Zehlendorf	6	7
1	Pankow	11	13
2	Reinickendorf	9	11
4	Lichtenberg	7	10
8	Tempelhof-Schöneberg	4	6
9	Neukölln	3	5
7	Charlottenburg-Wilmersdorf	4	7
10	Mitte	1	6

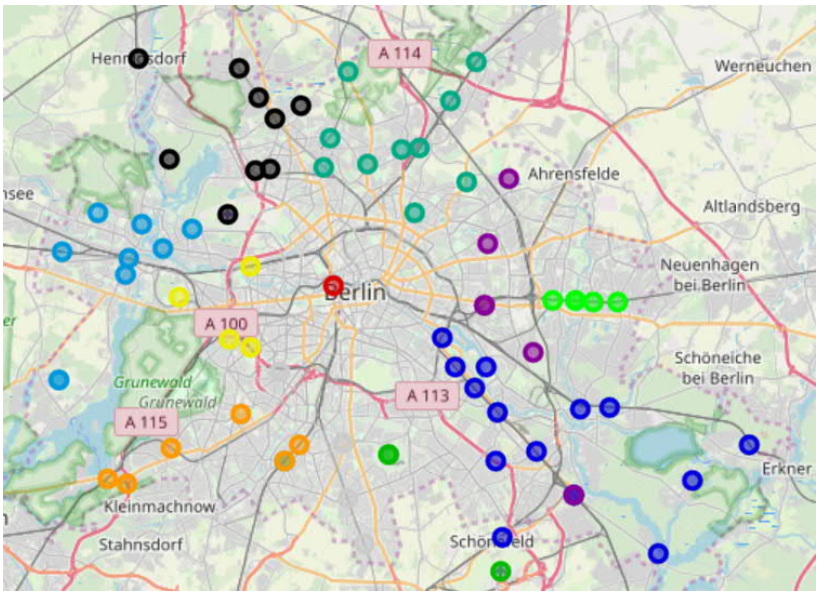
So again the ratio can be calculated with the help of the total number of neighborhoods sorted descending.

	Borough	Cluster_Labels_Count	Num_of_neighborhoods	Ratio
0	Treptow-Köpenick	15	15	1.000000
3	Spandau	9	9	1.000000
6	Marzahn-Hellersdorf	5	5	1.000000
5	Steglitz-Zehlendorf	6	7	0.857143
1	Pankow	11	13	0.846154
2	Reinickendorf	9	11	0.818182
4	Lichtenberg	7	10	0.700000
8	Tempelhof-Schöneberg	4	6	0.666667
9	Neukölln	3	5	0.600000
7	Charlottenburg-Wilmersdorf	4	7	0.571429
10	Mitte	1	6	0.166667

It can be seen that all neighborhoods of the boroughs Treptow-Köpenick, Spandau, and Marzahn-Hellersdorf belong to cluster 0 and 2. So these boroughs are highly interesting for the financial support by the government.

4.4. Geographic judgement

The following map illustrates all neighborhoods belonging to cluster 0 or 2 colored to their corresponding borough.



The deep blue circles belong to Treptow – Köpenick, the light green circles to Marzahn – Hellersdorf and the light blue ones to Spandau. One can see that Treptow – Köpenick has almost six neighborhoods which are located within natural environments. This can be an explanation why those are set to cluster 0 or 2. Spandau is well known for its working and industrial area including the Tegel

airport. So this might also be one explanation for belonging only to cluster 0 and 2. The neighborhoods of Marzahn – Hellersdorf are very close to each other along the S-Bahn track. This area is known as a living area. So it is quite surprisingly that this borough seems to be under developed in terms of social attractiveness.

5. Conclusion

The data exploration and analysis results focus on the borough of Reinickendorf having the most demand on developing support. Thus it has the most neighborhoods with less than 20 venues in comparison to the total number of neighborhoods and total area size.

As a contrast the clustering model shows that Treptow – Köpenick, Marzahn – Hellersdorf and Spandau have all of their neighborhoods belonging only to cluster 0 and 2 which are defined as 'support required'. As some neighborhoods of Treptow – Köpenick can be seen as countryside areas this borough should not longer be in focus. Spandau is a working area and the plan is to increase the attractiveness of working areas to also live there. But the recommendation for the financial support would be Marzhan - Hellersdorf. Because it is more important to satisfy the peoples needs already living in this borough instead of attracting new people to Spandau. As the belonging neighborhoods are close together the benefit of a supporting project would might have side effects to each of the neighborhoods.

So finally it can be said that Reinickendorf has the most demand for financial support. But in Marzahn – Hellersdorf the financial support might have bigger effects or can probably be reached with lower financial resources. If this can be realized both of the boroughs could be supported which leads to the most benefits.