The Best District in Berlin

Analysis carried out by Jussi Mattila

9.8.2020

Contents

1. Introduction

A client has approached our data science team and has requested us to do an analysis of the best district to move in Berlin as she plans to relocate herself into the city. She has the following criteria with respect to the new apartment location:

- The district should have good amount of venues as she likes to go to restaurants and bars during her freetime

- The crime rate in the area should be relatively low

Consequently, the client is looking for is a district where the ratio of venues/crimes is as high as possible, venues here referring to both restaurants and bars. The exact location or price of the apartment is not relevant to the client and therefor the analysis can be carried out directly on District-to-district comparison. In order to execute the task, datasets on the crimes on Berlin districts are sought for, along with data on the restaurants and bars on each district.

2. Methodology

Once the data is available and prepared, we can start assigning venue data and crime rates for each district and then cluster the districst base on these. For the approach, we have chosen first to look at the simple descriptive statistics of the venue number/crime rate -ratio and then by using K-means clustering, cluster the districts based on the available dataset and make a recommendation of the best district for the client.

3. Data

3.1 Original data

Data on the geospatial location of the districts on Berlin were extracted from the following webpage:

https://raw.githubusercontent.com/funkeinteraktiv/Berlin-Geodaten/master/berlin_bezirke.geojson

The districts are shown in the map below (Figure 1) and the districts are as follows:

- Treptow-Köpenick
- Steglitz-Zehlendorf
- Neukölln
- Tempelhof-Schöneberg
- Charlottenburg-Wilmersdorf
- Friedrichshain-Kreuzberg
- Marzahn-Hellersdorf
- Spandau
- Mitte
- Lichtenberg
- Reinickendorf
- Pankow

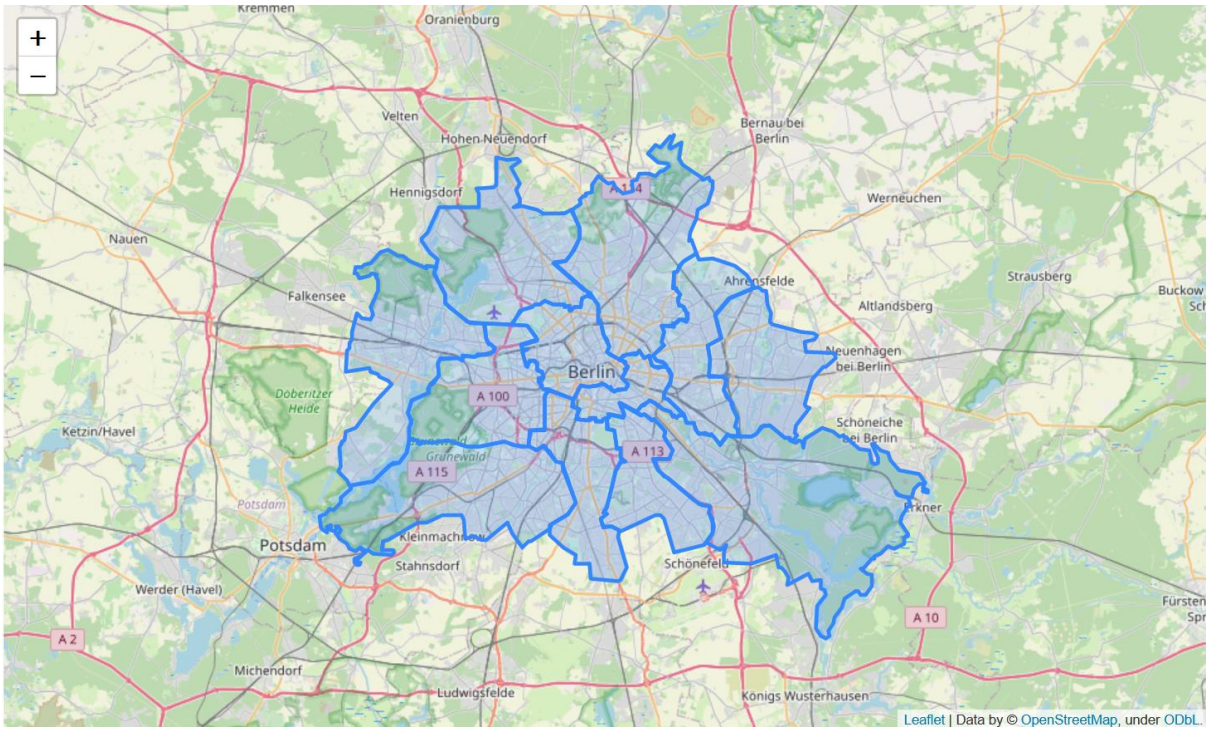

Figure 1. Districts of Berlin

Crime data for Berlin districts were acquired from Kaggle:

https://www.kaggle.com/danilzyryanov/crime-in-berlin-2012-2019/version/4

The data contains crime counts for different districts for the years 2012-2019 and the crimes are reported as counts per crime types.

	Year	District	Code	Location	Robbery	Street_robbery	Injury	Agg_assault	Threat	Theft	Car	From_car	Bike	Burglary	Fire	Arson	Damage
0	2012	Mitte	10111	Tiergarten Süd	70	46	586	194	118	2263	18	328	120	68	16	4	273
1	2012	Mitte	10112	Regierungsviertel	65	29	474	123	142	3203	10	307	170	37	10	4	380
2	2012	Mitte	10113	Alexanderplatz	242	136	1541	454	304	8988	81	792	822	275	49	27	1538
3	2012	Mitte	10114	Brunnenstraße Süd	52	25	254	60	66	1916	86	192	396	131	14	5	428
4	2012	Mitte	10221	Moabit West	130	51	629	185	199	2470	94	410	325	161	42	22	516

Foursquare data were extracted for each district using Foursquare API and the data incorporated into tables as shown below, with Venue location and category included. The extracted datapoints are shown Figure 2.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Treptow-Köpenick	52.429584	13.61121	SpreeArche	52.443846	13.618813	Seafood Restaurant
1	Treptow-Köpenick	52.429584	13.61121	Da Dalt	52.453217	13.624990	Ice Cream Shop
2	Treptow-Köpenick	52.429584	13.61121	Mauna Kea	52.452572	13.624984	Café
3	Treptow-Köpenick	52.429584	13.61121	Der GRIECHE	52.446453	13.627280	Greek Restaurant
4	Treptow-Köpenick	52.429584	13.61121	Pizza Dorado	52.424761	13.574858	Pizza Place

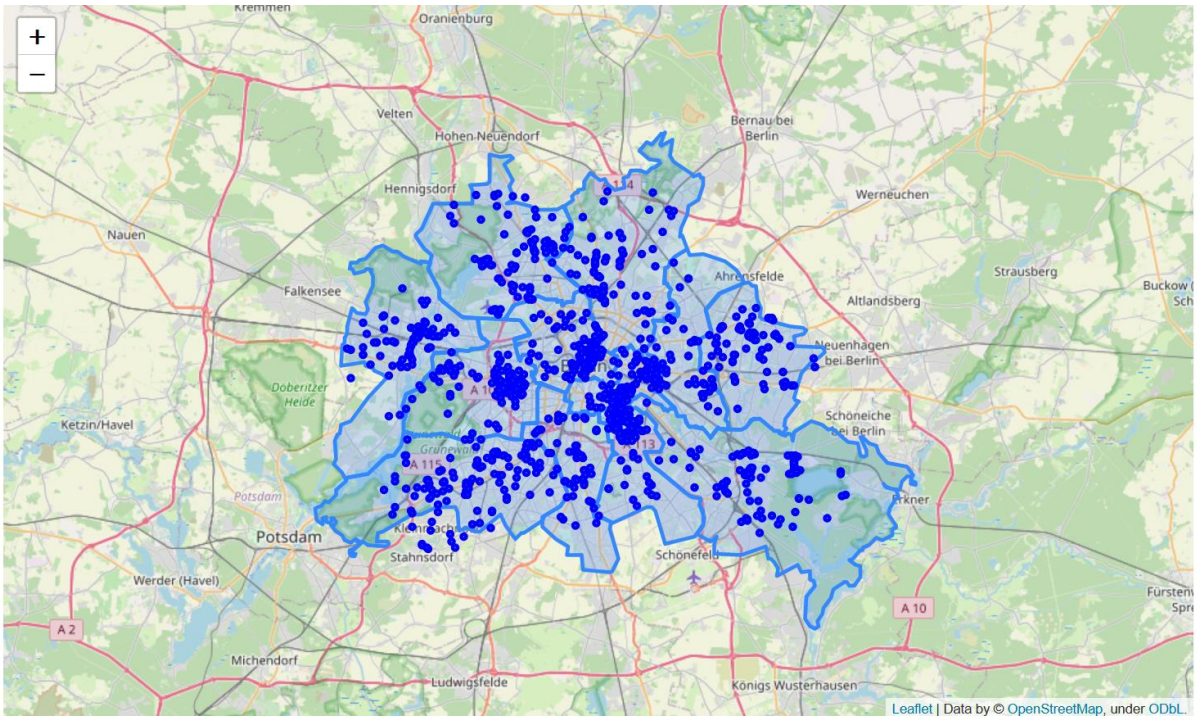

Figure 2. Foursquare venue data extracted from the Berlin districts

3.2 Data preparation

As our customer is interested mostly on the restaurant and bar venues of the Districts, the venues from the Forsquare data, which contain the words "Restaurant" or "Bar" were extracted. Furthermore, the counts of crimes of each district were appended into the data set and a column of suitability index was added. The suitability index is defined as:

Suitability index= (Number of bars and restaurants)/(Number of crimes)

The suitability index values were further normalized with the highest suitability index value, thus getting a dataset where the suitability index varies from 0 to 1, allowing more easier comparison of the districts. And extract of the generated dataset is shown below.

	Year	District	crime_count	venue_count	suitability_index
0	2012	Charlottenburg-Wilmersdorf	57053	34	0.518577
1	2013	Charlottenburg-Wilmersdorf	59329	34	0.498683
2	2014	Charlottenburg-Wilmersdorf	59990	34	0.493188
3	2015	Charlottenburg-Wilmersdorf	63567	34	0.465436
4	2016	Charlottenburg-Wilmersdorf	60363	34	0.490141
...	...	...	...	...	...
91	2015	Treptow-Köpenick	28395	24	0.735499
92	2016	Treptow-Köpenick	30469	24	0.685434
93	2017	Treptow-Köpenick	30006	24	0.696011
94	2018	Treptow-Köpenick	29597	24	0.705629
95	2019	Treptow-Köpenick	30422	24	0.686493

4. Results

4.1 Exploratory analysis and data visualization

As a first part of the analysis, we first looked at how suitability indices have evolved within each district in order to see whether this could have significant effect on the results. The data is shown in Figure 3, which gives already much insight onto the suitability index for each district. It is evident that the Renickendorf-district has the highest suitability index and the Mitte the lowest. In addition, the index-values seem to have been rather constant for each of the districts, thus there is no need to assess the changes of the index-values any furhter and we can focus only on the year 2019 in the following analysis. A bar blot of the suitability index values for each district for the year 2019 are shown in Figure 4. In the figure we see the same effect for the year 2019 as already shown in Figure

3: the Renickendorf-district has clerly the highest suitability index and the Mitte the lowest. The suitability indices are shown on map in Figure 5.

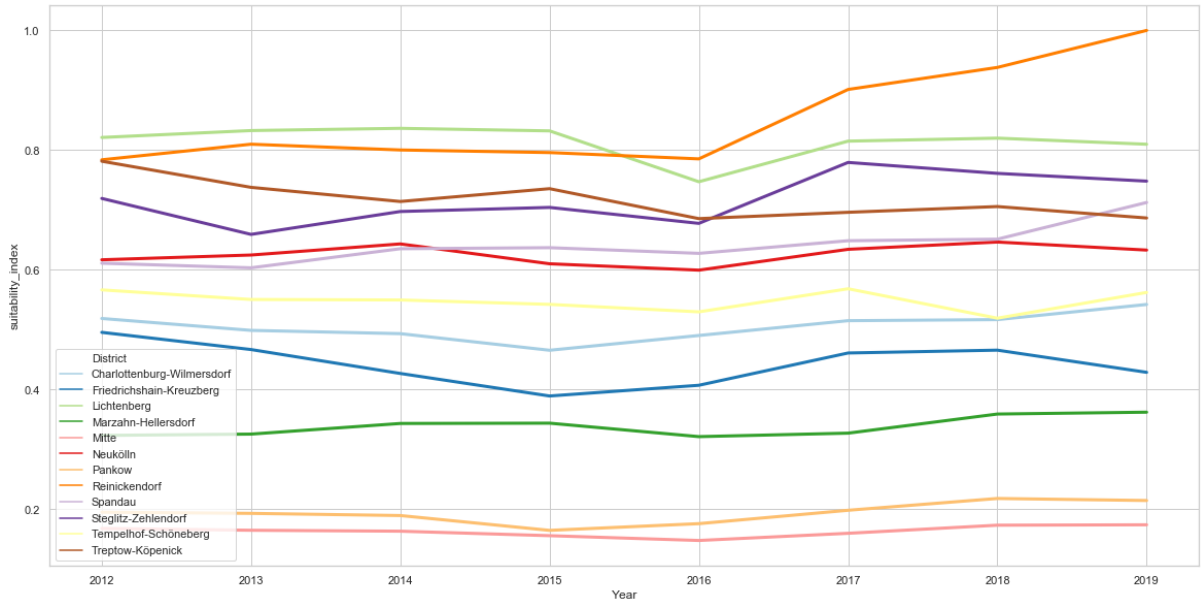

Figure 3. Suitability index per year for each district

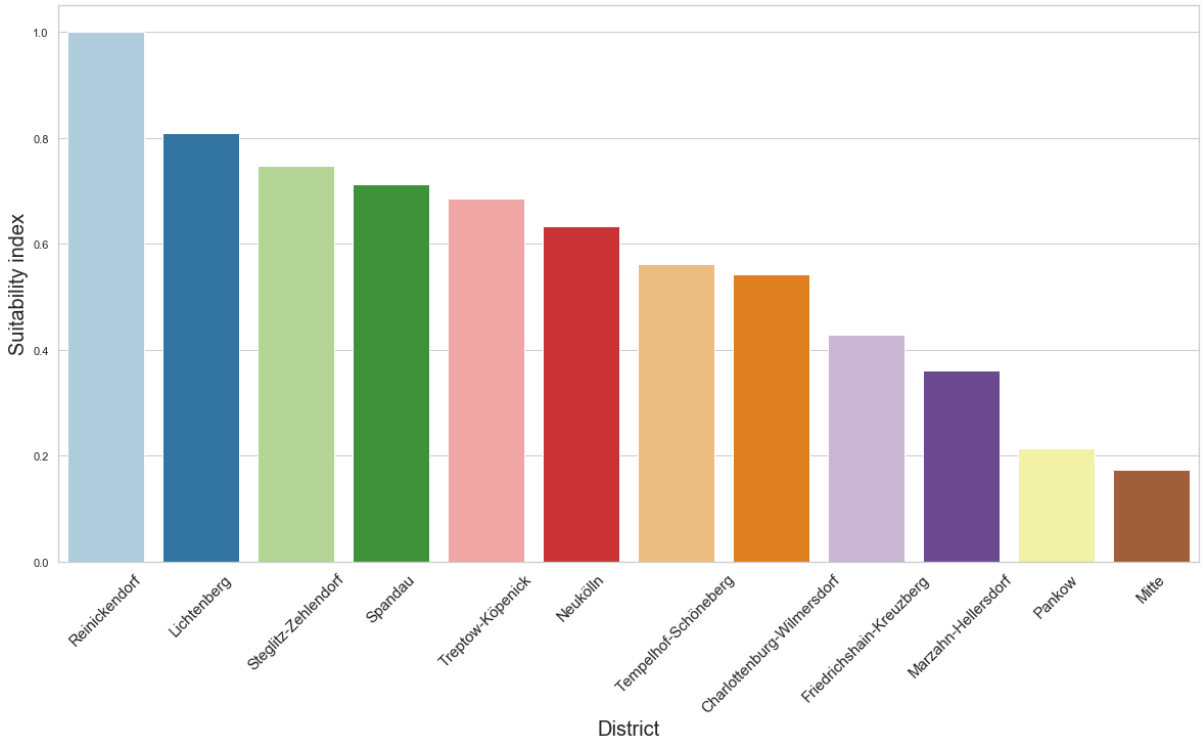

Figure 4. Suitability index for each district for the year 2019.

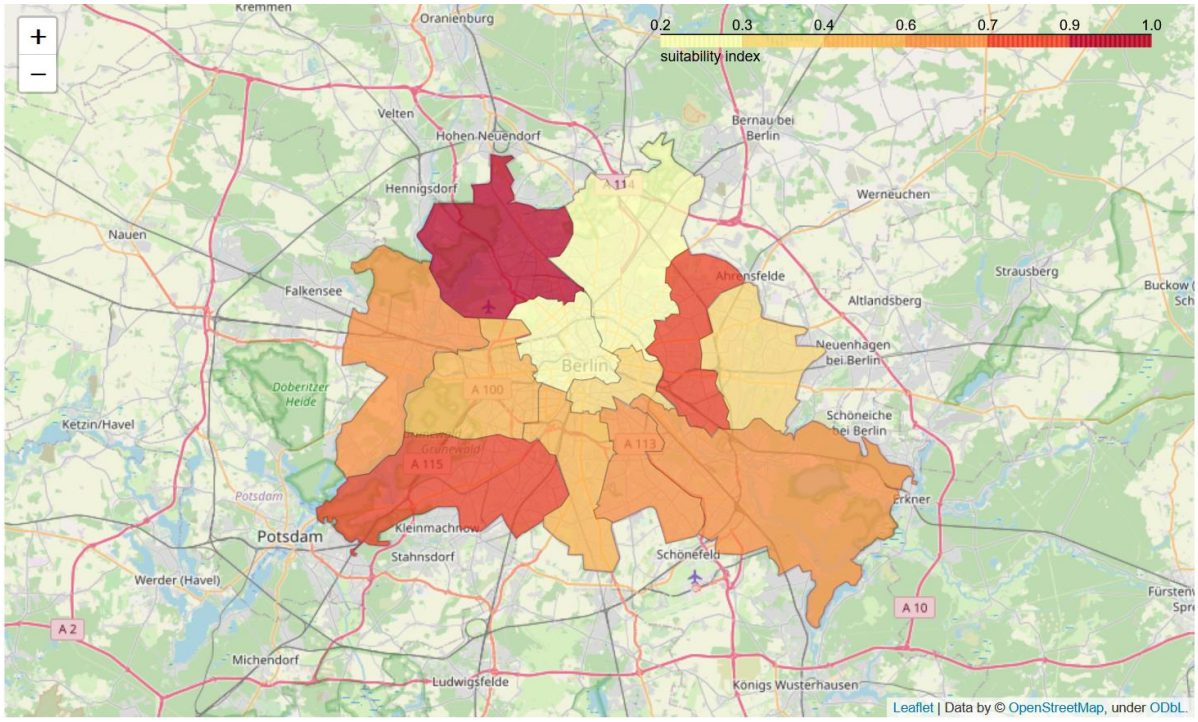

Figure 5. Suitability index for each district for the year 2019 shown on map.

4.2 K-means clustering

The results presented in the previous section already indicated, which districts have the highest suitability indices but in order to gain further insight on which districts are similar with respect to crime and venue counts, we can ran K-means clustering on the dataset. The analysis was carried out with 4 clusters this is suitable number with respect to number of districts and the scope of the task. The four clusters are shown in the scatterplot in Figure 6. Based on the clustering, the following classes could be identified:

0. Low venue count and hight crime count
1. Modest to high venue count and low crime count
2. Modest to high venue count and modest crime count
3. High venue count and modest crime count

The clusters are also shown on map in Figure 7. With respect to the client, cluster number 1 is the most relevant as these districts have modest to high venue count and low crime count.

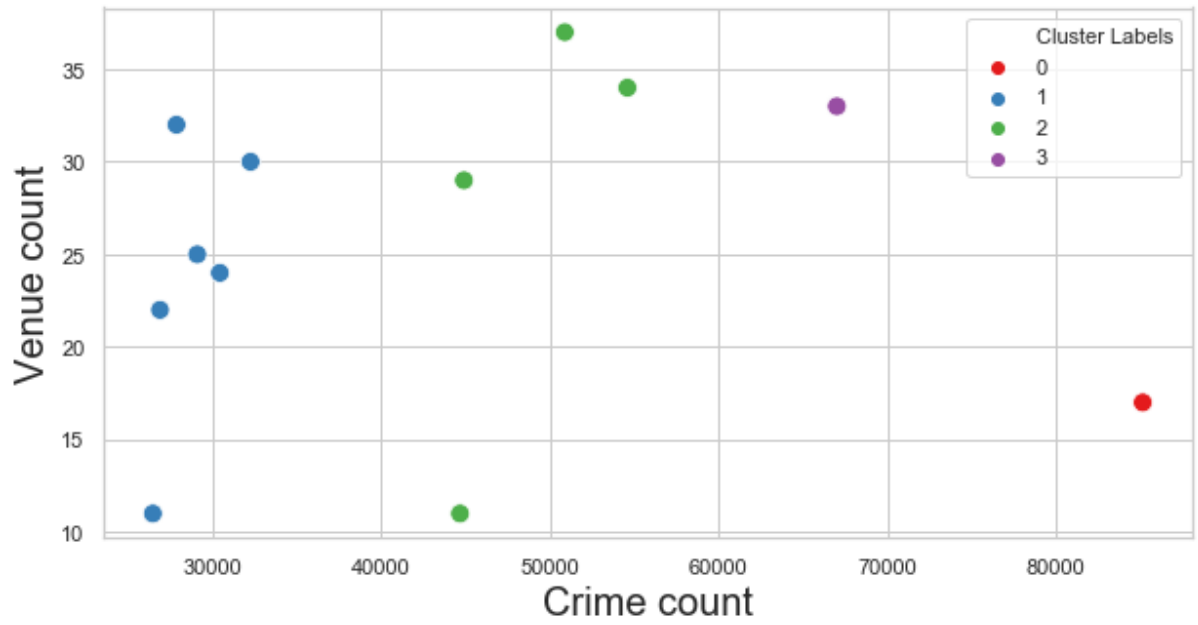

Figure 6. District cluster based on venue and crime counts

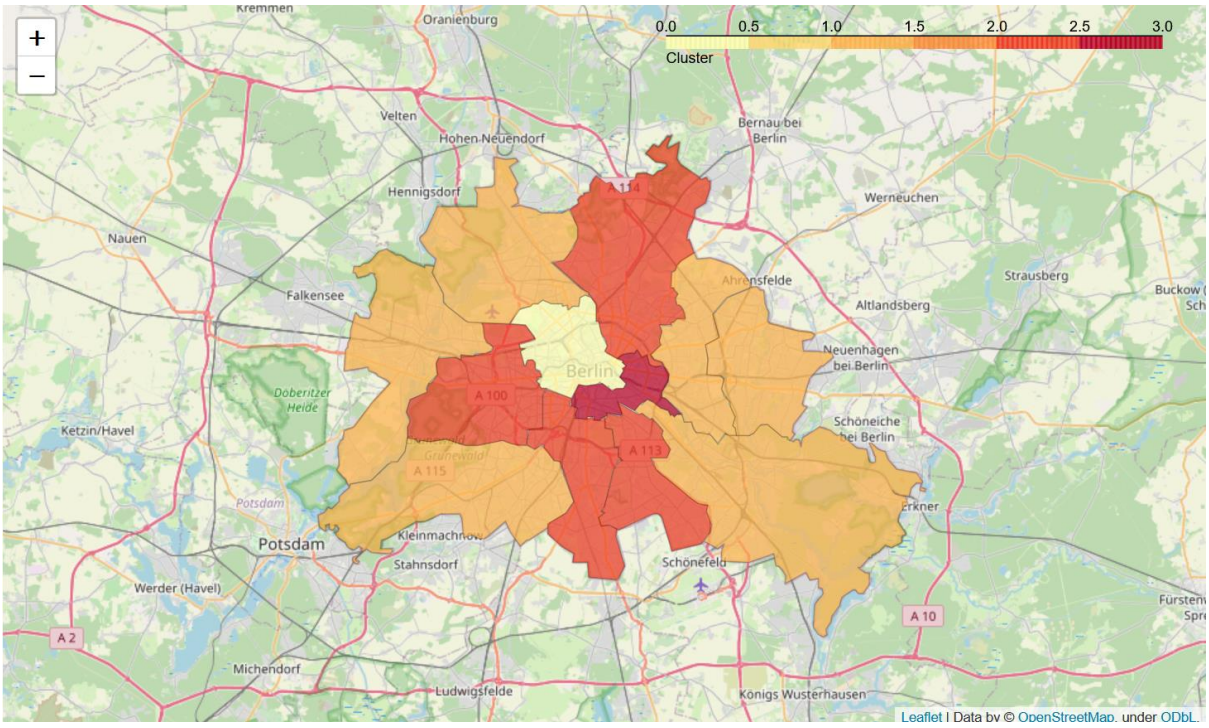

Figure 7. District cluster based on venue and crime counts shown on map

5. Discussion

Based on the analysis, the suitability index, which is a measure of the count of restaurants and bars in a district divided by the number of crimes in the district gives an index value which can be used to asses which are the best districts for the client to relocate herself within Berlin. Based on the analysis, the highest suitability index is for the Renickendorf-district and the lowest for the Mitte. Using a K-means clustering, we could also classify the districts into four different classes, which share

similar venue and crime counts and could therefore considered as potential district as well. Renickendorf-district belongs to the cluster 1 and the districts which belong to the same cluster are given in the table below. The cluster includes a total of 6 districts and all of these, yet excluding Marzahn-Hellersdorf which has anomalously low suitability index, could be considered as potential districts for the client as well.

	Cluster Labels	Year	District	crime_count	venue_count	suitability_index
63	1	2019	Reinickendorf	27846	32	1.000000
23	1	2019	Lichtenberg	32239	30	0.809753
79	1	2019	Steglitz-Zehlendorf	29080	25	0.748098
71	1	2019	Spandau	26872	22	0.712419
95	1	2019	Treptow-Köpenick	30422	24	0.686493
31	1	2019	Marzahn-Hellersdorf	26447	11	0.361934

It is noted that the analysis is based on crime data from 2019 and as such contains some uncertainties on what the current status with respect to crime counts is. But it is expected that these are similar in the present day as the crime count trends were rather constant. The venue count is also dependent on what venues are registered into the Foursquare database, yielding further uncertainty on the results, but the Foursquare data is considered as the best available at this time and the results therefore indicative of the true situation in Berlin.

6. Conclusions

Based on the analysis and on the numbers of reported crimes and restaurants and bars in Berlin, we recommend that the client should look for apartment in the Renickendorf-district as this has the highest suitability index (ratio of venues to crimes). The Renickendorf-district shares however similar suitability index with the following district, which could also be consider as potential districts for relocation, depending on client's other potential preferences:

- Lichtenberg
- Steglitz-Zehlendorf
- Spandau
- Treptow-Köpenick