Expanding business within Berlin

Similarity analysis of Berlin's neighborhoods

elvijsm

April 15, 2020

1 Introduction

Berlin is the capital city of the Federal Republic of Germany. With its 3.8 million inhabitants it is the largest city within the European Union and thus a city of crucial economic, political and cultural importance. Administratively, the city is divided into 12 districts (boroughs, German *Bezirke*) and 96 localities (neighborhoods, German *Ortsteile*).

In this project we explore the similarity of the localities of Berlin. Such analysis could be of interest to the following groups of people:

- Owners of small or medium-sized businesses, when considering to expand or relocate their business, would be interested in knowing which parts of the city would provide similar conditions for their business.
- People considering to change their residence within city limits. People being forced to move would most likely prefer moving to a similar locality. People dissatisfied with their current environment, on the other hand, might prefer to move to a different type of locality.

Measuring the similarity of two neighborhoods is a multi-faceted problem. There are many variables one could justifiably consider. Here we focus on only two aspects: the types and frequency of social venues present and demographics.

2 Data

Three main sources of data were used: data on localities, social venues within the localities, and demographic makeup of the localities.

2.1 Localities

A table of the districts and localities of Berlin was obtained from the German wikipedia page on the "Administrative structure of Berlin" [1]. The data was retrieved using the `read_html` method of the `pandas` library, and processed to yield a table of the number,

district, area (in km^2) and population of each of the 96 localities (here and elsewhere a sample of 5 rows are shown):

	Nr.	District	Locality	Area	Population
6	201	Friedrichshain-Kreuzberg	Friedrichshain	9.78	134900
22	402	Charlottenburg-Wilmersdorf	Wilmersdorf	7.16	102240
46	703	Tempelhof-Schöneberg	Tempelhof	12.20	62442
72	1003	Marzahn-Hellersdorf	Kaulsdorf	8.81	19366
75	1101	Lichtenberg	Friedrichsfelde	5.55	53411

The geographical latitude and longitude of each locality were obtained using the `geocode` method of the `geopy` library and appended to the table:

	Nr.	District	Locality	Area	Population	Latitude	Longitude
19	312	Pankow	Rosenthal	4.90	9484	52.598319	13.375519
37	601	Steglitz-Zehlendorf	Steglitz	6.79	75428	52.457257	13.322287
48	705	Tempelhof-Schöneberg	Marienfelde	9.15	32463	52.412577	13.366592
52	803	Neukölln	Buckow	6.35	40888	52.418662	13.428950
69	915	Treptow-Köpenick	Schmöckwitz	17.10	4394	52.375665	13.648855

2.2 Social venues

The venues of each locality were obtained by communicating with the Foursquare API [2], mainly with the `venues/explore` endpoint.

2.3 Demographic information

Demographic information was obtained from the Berlin-Brandenburg Department of Statistics [3]. The data, formatted as a csv file of 7194 rows and 8 columns, consists of the number of people in each locality at the end of 2018, broken down according to gender, citizenship, and age group (every 5 years). The data was retrieved using the `read_csv` method of the `pandas` library.

3 Methodology

3.1 Visualizing neighborhoods

Since the geographical coordinates of each locality were obtained independently of the data on localities, we first visualized the localities on a map using the `folium` library, color-coding them according to districts (Fig. 1). The continuity of each district is readily visible.

It should be noted that the coordinates of some localities depend on which source they are obtained from. The differences can be fairly substantial (Fig. 2). For example, the latitude and longitude of "Tiergarten, Mitte" according to google maps are (52.514534, 13.35010), whereas according to `geopy` they are (52.509778, 13.35726). Generally, it

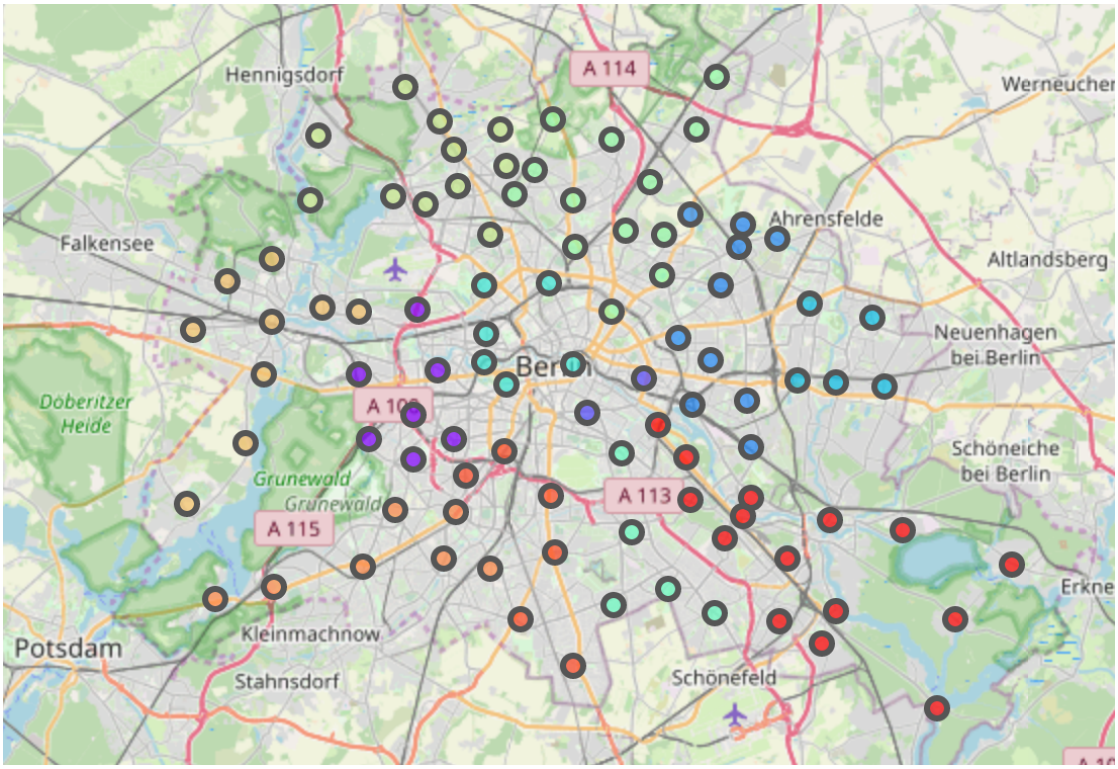

Figure 1: Localities of Berlin color-coded according to their districts.

seems that the coordinates returned by `geopy` are more geographically central, which is preferable when looking up venues in each neighborhood.

3.2 Retrieving and processing venues

The venues of each locality were retrieved by querying the `venues/explore` endpoint of the Foursquare API [2] with the latitude and longitude of each locality:

```
https://api.foursquare.com/v2/venues/explore?&client_id=***
&client_secret=***&v=20200410&ll=lat,lng&radius=R&limit=100&offset=0
```

The endpoint returns a json object containing extensive information about the venues in the surroundings of the specified location. The following attributes of each venue were kept and processed to arrive at a data frame like shown in Fig. 3:

- `['name']`: name of the venue ("Venue")
- `['location']['lat']`: latitude of the venue
- `['location']['lng']`: longitude of the venue
- `['location']['distance']`: distance from the venue to the location specified by the query

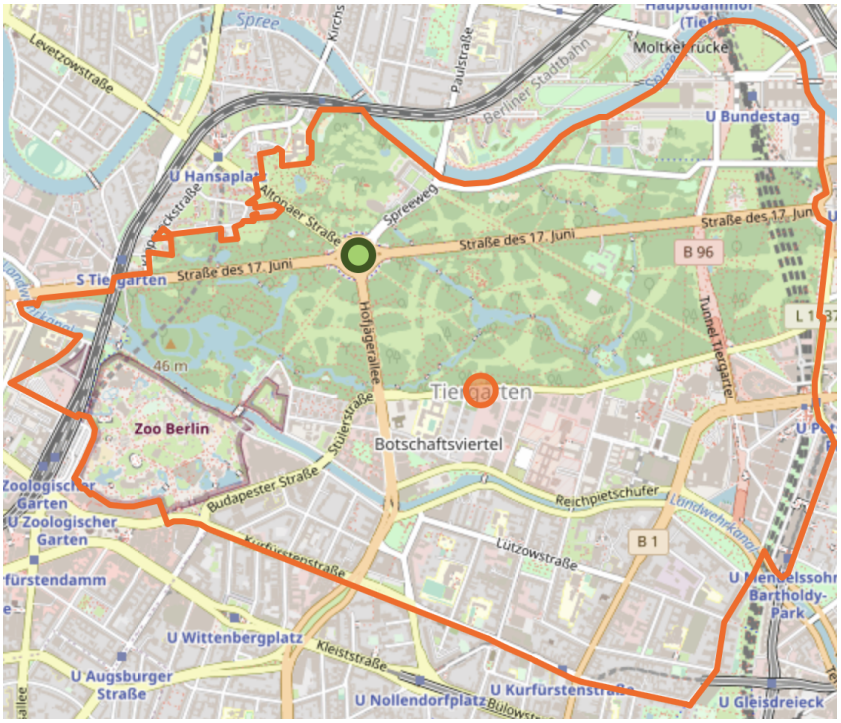

Figure 2: Geographical location of "Tiergarten, Mitte" according to google maps (green circle) and `geopy` (orange circle).

- `['categories'][0]['name']`: category of the venue ("CategoryFull")
- `['categories'][0]['icon']['prefix']`: a link to the icon of the venue category ("Type", "Category")
- `['id']`: unique identifier of the venue ("Id")

Some remarks regarding the use of the `venues/explore` endpoint:

1. If the radius around the point of origin is too large, the same venue (Id) can be returned by multiple queries. In such cases the venue was assigned to the locality for which the distance was smallest. In practice only four venues were returned more than once.

2. There are a large number of detailed categories in the Foursquare database ("CategoryFull"). For Berlin, there are 323 unique categories. Meanwhile, few localities have more than a few tens of venues. This means that we are dealing with fairly sparse data. For this reason we made use of `['categories'][0]['icon']['prefix']`, which looks something like this:
   `https://ss3.4sqi.net/img/categories_v2/food/italian_`
   `https://ss3.4sqi.net/img/categories_v2/food/cafe_`
   `https://ss3.4sqi.net/img/categories_v2/nightlife/pub_`

	Locality	Venue	Latitude	Longitude	Distance	CategoryFull	Type	Category	Id
0	Lichterfelde	Schloßpark-Grill	52.435208	13.313280	235	Eastern European Restaurant	food	default	4dbd79e4815439392fb01b20
1	Lichterfelde	La Malga	52.439900	13.315969	323	Italian Restaurant	food	italian	4f1ffa49e4b057da8c539ec9
2	Lichterfelde	Da Remo	52.441284	13.314669	447	Italian Restaurant	food	italian	4ea8263d9adfcc4eb0d791f7
3	Lichterfelde	Bao Vietnamese Cooking	52.446250	13.315415	1002	Vietnamese Restaurant	food	vietnamese	4dcfe032b0fb25f6e36710f6
4	Lichterfelde	Tomasa	52.433795	13.317572	463	Café	food	cafe	513b2fc4e4b0809a1e3736b9
5	Lichterfelde	Loch Ness Scottish Pub	52.446975	13.311626	1088	Whisky Bar	nightlife	whiskey	4cb8aa687148f04da1c3d1ab

Figure 3: Sample of venues.

`https://ss3.4sqi.net/img/categories_v2/shops/bookstore_`
Foursquare appears to categorize the venues more broadly when displaying venues as icons on a map. These links were parsed to categorize each venue in 8 "Types" – food, shops, travel, parks_outdoors, arts_entertainment, nightlife, education, building – and 217 "Categories" which results in a Bavarian and a Schnitzel restaurant both being categorized as "Food, German" instead of as "Bavarian Restaurant" and "Schnitzel Restaurant", respectively.

3. Since the localities are of very different sizes, ranging from $0.5\,\mathrm{km}^2$ (Hansaviertel, Mitte) to $35\,\mathrm{km}^2$ (Köpenick, Treptow-Köpenick), using the same radius in the query for every locality is bound to yield poor results. We thus computed a conservative radius for each locality according to

$$R = \mathrm{int}\left(\min\left(\mathrm{round}\left(1000\sqrt{\tfrac{A}{\pi}}, -2\right)/2, 1500\right)\right)\ \mathrm{m}.$$

In words, the radius R of a circle, whose area equals the area A of the locality, is rounded to 50 m and capped at 1500 m. The additional division by a factor of 2 significantly reduces the radii to account for the fact that localities are not circular. Effectively, only the the inner quarter area of the locality is assumed to be circular. The distribution of the resulting radii, which were used as parameters in the query, are shown in Fig. 4.

4. Since the endpoint is limited to returning at most 100 venues, we used the `offset` parameter in the query to return the next set of venues within the specified radius of the given location. In other words, if 100 results were returned by the query, the query was performed again with an `offset=100` to return the remaining venues.

3.3 Processing demographic data

The raw demographic data was pivoted by age, gender, and citizenship to yield a more compact representation of the demographics. In particular, the age groups were reduced from the original 20 to just four: children (ages 0-15), students (ages 15-25), adults (ages 25-65), and seniors (ages 65+). Finally, the number of children, students, adults, seniors,

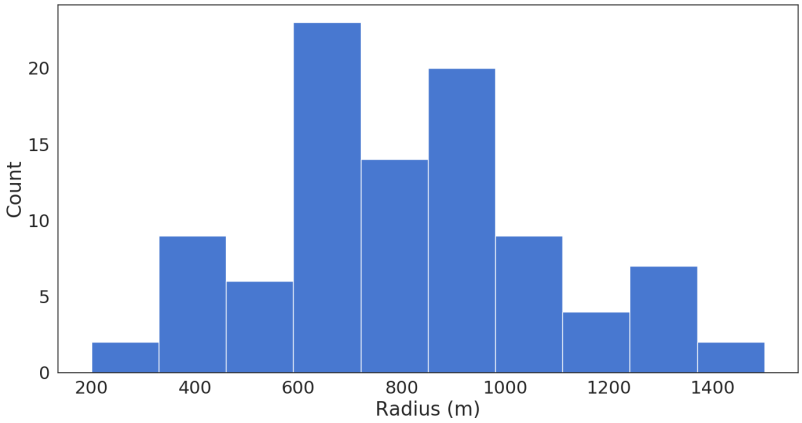

Figure 4: Distribution of locality radii.

men, women, foreigners and citizens in each locality were expressed as a fraction of the total number of people in the locality. The result was the following table:

Altersgr	District	Locality	total	children	students	adults	seniors	men	women	foreigners	citizens
34	Neukölln	Rudow	42497	0.133162	0.101254	0.506600	0.258983	0.486787	0.513213	0.104525	0.895475
87	Treptow-Köpenick	Grünau	6530	0.117917	0.066769	0.552527	0.262787	0.488974	0.511026	0.087136	0.912864
32	Neukölln	Gropiusstadt	37533	0.140516	0.097541	0.488503	0.273439	0.474089	0.525911	0.201636	0.798364
4	Charlottenburg-Wilmersdorf	Schmargendorf	22090	0.116523	0.085967	0.511000	0.286510	0.460933	0.539067	0.190448	0.809552
12	Lichtenberg	Friedrichsfelde	53059	0.117228	0.085603	0.526998	0.270171	0.482840	0.517160	0.147213	0.852787

Other than dividing by the population size, no further feature scaling was done on these columns. The reason this makes sense can be understood when considering the fractions of gender. The ratio of women-to-men varies across the localities, but the differences, with one exception, are minor: the ratio is always close to unity. Scaling the features further would exaggerate differences that are for most practical purposes irrelevant.

Note that there are minor differences between the total population in this table and the population listed in the first table because the data on wikipedia are dated with June 2019. For the sake of consistency, the total population from the demographic data was used.

Finally, two other features were added: population of the locality as a fraction of maximum population across all localities, and population density (inhabitants per km^2), also normalized by dividing by the maximum density across all localities.

3.4 Measuring the similarity of neighborhoods

The similarity of localities was considered on two accounts: social venues and demographics. The localities were clustered using the K-means clustering algorithm to find localities that are similar in terms of available social venues (both in terms of type and category), as well as in terms of demographics. When clustering the localities according to venues, the values of the venue types and categories were one-hot encoded, and the

Locality	1st Category	2nd Category	3rd Category	4th Category	5th Category
Blankenfelde	shops_default	shops_automotive	parks_outdoors_hikingtrail	food_cafe	NaN
Bohnsdorf	shops_financial	shops_hardware	shops_food_grocery	food_italian	food_gastropub
Buckow	shops_food_grocery	shops_realestate	shops_discountstore	food_indian	food_german
Charlottenburg-Nord	travel_subway	shops_food_grocery	parks_outdoors_rockclimbing	parks_outdoors_plaza	parks_outdoors_park
Falkenhagener Feld	shops_pharmacy	shops_food_grocery	food_snacks	arts_entertainment_stadium_soccer	NaN
Johannisthal	travel_trainstation	travel_busstation	shops_food_grocery	parks_outdoors_park	nightlife_pub
Karow	shops_food_grocery	travel_busstation	food_german	food_default	NaN
Neu-Hohenschönhausen	shops_food_grocery	travel_trainstation	shops_technology	shops_pharmacy	shops_mall
Pankow	food_cafe	shops_food_grocery	building_gym	shops_pharmacy	food_italian
Waidmannslust	building_gym	travel_lightrail	shops_technology	shops_pharmacy	shops_pet_store

Figure 5: A sample of most common venue categories in different localities.

Locality	1st Type	2nd Type	3rd Type	4th Type	5th Type	6th Type	7th Type
Blankenfelde	shops	parks_outdoors	food	NaN	NaN	NaN	NaN
Bohnsdorf	shops	food	NaN	NaN	NaN	NaN	NaN
Buckow	shops	food	NaN	NaN	NaN	NaN	NaN
Charlottenburg-Nord	parks_outdoors	building	travel	shops	nightlife	food	arts_entertainment
Falkenhagener Feld	shops	food	arts_entertainment	NaN	NaN	NaN	NaN
Johannisthal	food	travel	arts_entertainment	shops	parks_outdoors	nightlife	NaN
Karow	shops	food	travel	NaN	NaN	NaN	NaN
Neu-Hohenschönhausen	shops	food	travel	parks_outdoors	building	arts_entertainment	NaN
Pankow	food	shops	building	nightlife	travel	parks_outdoors	NaN
Waidmannslust	shops	food	building	travel	NaN	NaN	NaN

Figure 6: A sample of most common venue types in different localities.

mean of the occurrence of each type/category was computed. This yielded as a byproduct the most common venue categories (Fig. 5) and types (Fig. 6) for each locality. Note, a missing value starting at the "N-th Category/Type" means that there are only $N-1$ unique venue categories or types in this locality.

The optimal value of K for the K-means algorithm was sought using the Elbow method. The error was computed according to

$$e = \sum_{i=1}^{96} \min_j \left(\|x_i - c_j\| \right),$$

i.e. as the sum of the Euclidian distances between each locality with features x_i and the nearest cluster centroid c_j. To avoid local minima, the algorithm was run with 100 different centroid seeds for each value of K.

Finally, the features used for the two clustering analyses were considered together. For each locality the cosine similarity was taken between its feature vector and the feature vector of every other locality, resulting in an $N \times N$ matrix of the type shown in Fig. 7. For each locality, the most (least) similar locality was taken to be the locality with which

Locality	Mitte	Moabit	Hansaviertel	Tiergarten	Wedding	Gesundbrunnen	Friedrichshain	Kreuzberg
Mitte	-99.000000	0.946401	0.797653	0.767006	0.969561	0.963663	0.963996	0.967381
Moabit	0.946401	-99.000000	0.765721	0.751292	0.991550	0.963482	0.961131	0.953263
Hansaviertel	0.797653	0.765721	-99.000000	0.750164	0.794265	0.836496	0.717623	0.712789
Tiergarten	0.767006	0.751292	0.750164	-99.000000	0.771433	0.721269	0.648565	0.638414
Wedding	0.969561	0.991550	0.794265	0.771433	-99.000000	0.966439	0.961116	0.959653
Gesundbrunnen	0.963663	0.963482	0.836496	0.721269	0.966439	-99.000000	0.969148	0.969133
Friedrichshain	0.963996	0.961131	0.717623	0.648565	0.961116	0.969148	-99.000000	0.996734
Kreuzberg	0.967381	0.953263	0.712789	0.638414	0.959653	0.969133	0.996734	-99.000000

Figure 7: Locality-similarity matrix.

the cosine similarity is greatest (smallest).

4 Results

4.1 Venues

In total 2852[1] venues were retrieved. Fig. 8 shows the localities with 50 or more venues. Unsurprisingly, the central localities dominate the figure. The only more or less remote localities with more than 50 venues are Spandau (district Spandau), Köpenick (Treptow-Köpenick), Tegel (Reinickendorf), and Zehlendorf (Steglitz-Zehlendorf). Note how retrieving only 100 venues for the busiest localities would have left a lot of venues out.

The most frequent venue category across all localities is "shops_food_grocery" (240 venues), followed by "food_cafe" (178), "food_italian" (123), and "nightlife_pub" (93). Meanwhile, 49 categories only have a single venue to represent them. The most frequent venue type across all localities is, unsurprisingly, "food" (1254 venues), followed by "shops" (702) and "travel" (249).

4.2 Clustering based on venues

Localities were clustered according to venue categories and types for different values of K. Figure 9 shows that the clustering results are rather ambiguous. Particularly when using categories of venues, there is no hint of an elbow-like feature. Instead, the error simply decreases slowly with increasing number of clusters. The clustering by venue type is somewhat more successful: three clusters are clearly better than two, which are clearly better than one. Increasing the number of clusters further does not, however, help much. The clustering by type is more successful because in terms of venue type the data is not sparse – there are only 8 venue types. On the other hand, the data is very sparse in

[1] Since the venues on Foursquare are continuously updated, executing the notebook on different days yielded slightly different numbers of venues. Therefore, the figures and tables of this report, as well as the quoted figures can have slight differences with what is in the notebook.

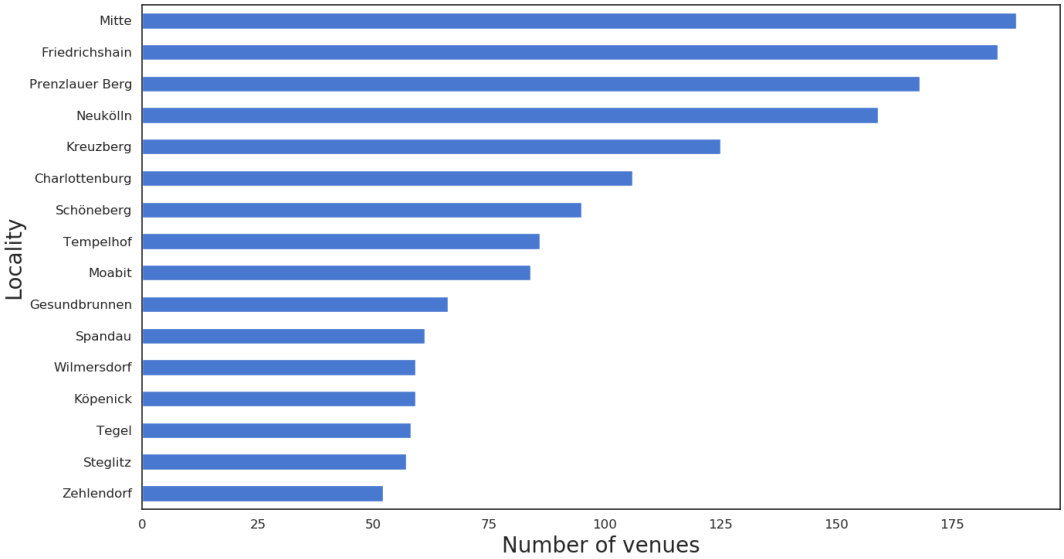

Figure 8: Localities with 50 or more venues.

terms of categories: there are more than 200 categories, yet most localities have only a few (tens) of unique venue categories.

We thus proceed with clustering the localities by venue type into three clusters. The resulting clusters have the mean frequencies of each venue type as listed in Fig. 10. The clusters actually separate quite well. Cluster 1 is dominated by venues associated with parks and outdoor activities, as well as travel venues. Cluster 2 is dominated by localities with a very high fraction of food-related venues. Finally, cluster 3 is dominated by localities with a high relative fraction of shopping venues.

Figure 11 shows the map of Berlin's localities color-coded according to these clusters. One may expect that Cluster 3 localities, a large fraction of whose venues are related to shopping, would be located near the city centre, but actually the opposite is the case.

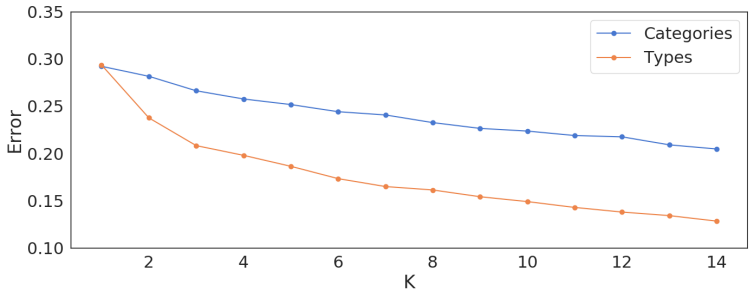

Figure 9: K-means clustering by venue category and type with different values of K.

Cluster	arts_entertainment	building	education	food	nightlife	parks_outdoors	shops	travel
1	0.051193	0.015014	0.001623	0.215857	0.016764	0.246429	0.183052	0.270068
2	0.051492	0.027365	0.000242	0.549394	0.039199	0.076204	0.198390	0.057714
3	0.044593	0.027354	0.000000	0.268161	0.007492	0.060174	0.473055	0.119171

Figure 10: Clusters when clustering by venue type.

The reason is that we are considering the relative fractions of venue types. A locality with only two venues, one of which is a grocery store, has a very high frequency of shopping-related venues. That is actually what the cluster is mostly capturing – localities in which the main attraction is the local grocery store.

4.3 Clustering based on demographics

Clustering based on demographics also presents a mixed picture (Fig. 12). Here too, no clear elbow-like feature is present. The choice appears to be between $K = 2$ and $K = 4$. We choose $K = 4$ here.

The resulting clusters have the mean demographics as listed in Fig. 13. The clusters that stand out the most are clusters 2 and 3. Cluster 2 contains localities with the highest relative fractions of the elderly, fewest foreigners, and overall low population. Cluster 3 is dominated by localities with large and dense populations with few seniors – the main working force. Clusters 1 and 4 fall between these two extremes.

Figure 14 shows the map of Berlin's localities color-coded according to these clusters. As expected, localities of cluster 2 are mostly located in the suburbs of Berlin. Localities of cluster 3 are, on the other hand, exclusively found in the central region of Berlin.

4.4 Most (dis)similar neighborhoods

Table 15 shows a sample (see notebook for the full table) of 32 localities, and the localities that are most and least similar to them. The locality that appears as similar to most other localities, i.e. in some sense the most typical locality, is Lichtenrade (district Tempelhof-Schöneberg; Fig. 16). The locality appearing to be by far the most distinct is Stadtrandsiedlung Malchow (district Pankow; Fig. 17).

5 Discussion

This work attempts to identify neighborhoods of Berlin that are similar along two aspects: their social venues and demographics. The results with respect to social venues indicate that significant additional investment in data cleaning and preparation is necessary to obtain high-quality results. In particular, further aggregating the venue categories listed on Foursquare should significantly improve the quality of the results. Here each venue category was encoded as a dummy variable. Ideally, the encoded categories would retain some measure of their similarity, such that different food venues are more similar to each

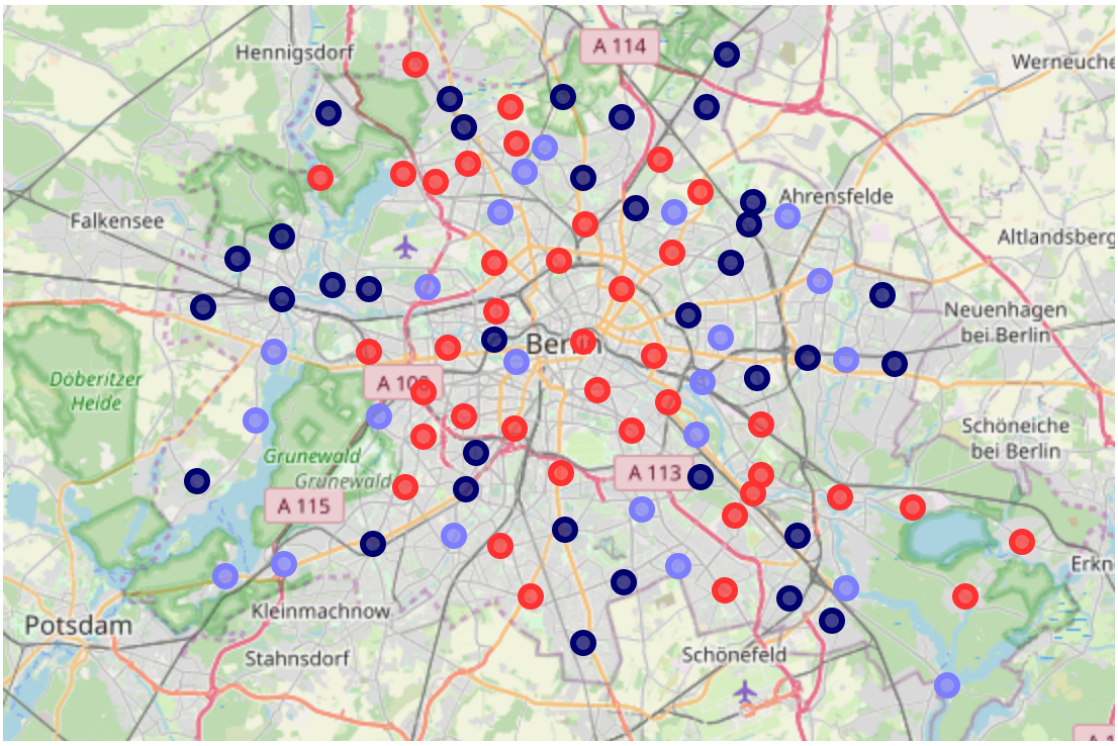

Figure 11: Localities of Berlin color-coded according to venue types. Light blue, red, and dark blue localities belong to clusters 1, 2, and 3 (see Fig. 10), respectively.

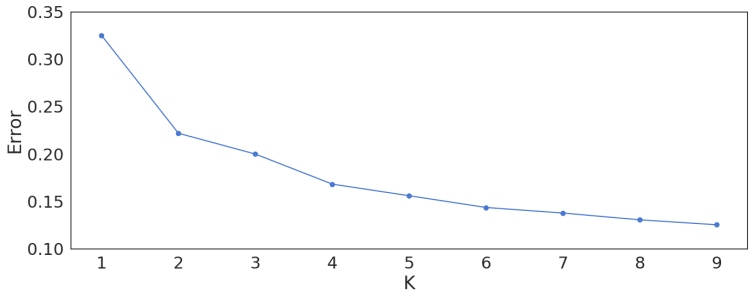

Figure 12: K-means clustering by demographics with different values of K.

Cluster	children	students	adults	seniors	men	women	foreigners	citizens	population	density
1	0.134015	0.100210	0.565230	0.200545	0.491465	0.508535	0.214015	0.785985	0.316882	0.680146
2	0.136134	0.085367	0.540726	0.237774	0.495925	0.504075	0.096325	0.903675	0.080748	0.105068
3	0.130668	0.093817	0.636547	0.138968	0.505625	0.494375	0.284216	0.715784	0.779866	0.782274
4	0.139882	0.096328	0.547936	0.215854	0.490507	0.509493	0.174710	0.825290	0.250328	0.269663

Figure 13: Clusters when clustering by demographics.

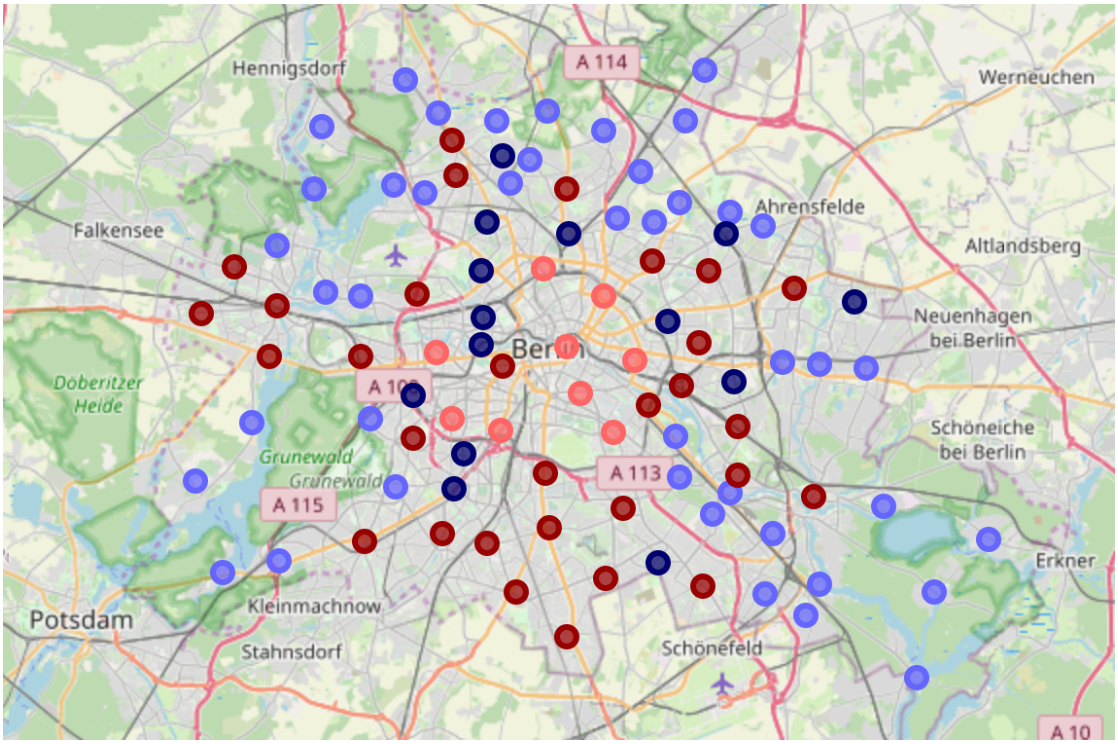

Figure 14: Localities of Berlin color-coded according to demographics. Dark blue, light blue, orange, and brown localities belong to clusters 1, 2, 3, and 4 (see Fig. 13), respectively.

other than a food venue and a travel venue, and restaurants are more similar to each other than a restaurant and an ice cream stand. This would also make use of categories like "building_religious_church" or "nightlife_secretbar" that appear only once but are obviously related to other churches and bars, respectively.

It would be interesting to consider not only the relative number of different venue types and categories, but also incorporate their ratings (`venues/{venueId}` endpoint) or likes (`venues/{venueId}/likes` endpoint). One could, for example, attempt to identify venue categories that are present and well liked in one locality, but missing in an another otherwise similar locality. Unfortunately, the number of venues in Berlin is too large to retrieve the number of likes and ratings in the allocated limit of free API calls.

As discussed previously, the characteristics of the clusters obtained by K-means clustering of venue types and of demographics could be readily interpreted. There does not, however appear to be a significant overlap between the two groups of clusters (Fig. 18). For example the demographics cluster 2, comprising older and less populous localities, is split fairly evenly between all three venues clusters. The only exception is the demographics cluster 3, comprising localities with large and relatively young populations: all of them belong to the venues cluster characterized by a very high fraction of food-related venues (cluster 2).

Locality	1st Similar	2nd Similar	3rd Similar	1st Dissimilar	2nd Dissimilar	3rd Dissimilar
Lichterfelde	Marzahn	Weißensee	Köpenick	Stadtrandsiedlung Malchow	Malchow	Wartenberg
Prenzlauer Berg	Friedrichshain	Kreuzberg	Neukölln	Stadtrandsiedlung Malchow	Plänterwald	Heiligensee
Pankow	Steglitz	Wilmersdorf	Märkisches Viertel	Stadtrandsiedlung Malchow	Plänterwald	Wilhelmsruh
Kreuzberg	Friedrichshain	Prenzlauer Berg	Neukölln	Stadtrandsiedlung Malchow	Plänterwald	Heiligensee
Französisch Buchholz	Heiligensee	Heinersdorf	Altglienicke	Stadtrandsiedlung Malchow	Malchow	Falkenberg
Rahnsdorf	Friedrichshagen	Müggelheim	Wannsee	Plänterwald	Neukölln	Wilhelmsruh
Reinickendorf	Lichtenberg	Hellersdorf	Mitte	Malchow	Stadtrandsiedlung Malchow	Lübars
Altglienicke	Heinersdorf	Mahlsdorf	Karow	Stadtrandsiedlung Malchow	Falkenberg	Malchow
Zehlendorf	Lichtenrade	Mariendorf	Tempelhof	Stadtrandsiedlung Malchow	Wilhelmsruh	Plänterwald
Karow	Baumschulenweg	Mahlsdorf	Niederschönhausen	Stadtrandsiedlung Malchow	Falkenberg	Neukölln
Buch	Heinersdorf	Biesdorf	Hakenfelde	Stadtrandsiedlung Malchow	Neukölln	Wilhelmsruh
Adlershof	Biesdorf	Hermsdorf	Baumschulenweg	Stadtrandsiedlung Malchow	Wilhelmsruh	Plänterwald
Konradshöhe	Blankenburg	Frohnau	Friedrichshagen	Stadtrandsiedlung Malchow	Neukölln	Wilhelmsruh
Johannisthal	Dahlem	Westend	Niederschöneweide	Stadtrandsiedlung Malchow	Heiligensee	Plänterwald
Tegel	Rudow	Friedrichshagen	Zehlendorf	Stadtrandsiedlung Malchow	Wilhelmsruh	Plänterwald
Staaken	Niederschönhausen	Alt-Hohenschönhausen	Britz	Stadtrandsiedlung Malchow	Falkenberg	Malchow
Mariendorf	Lichtenrade	Alt-Hohenschönhausen	Tempelhof	Stadtrandsiedlung Malchow	Wilhelmsruh	Plänterwald
Köpenick	Zehlendorf	Tegel	Westend	Stadtrandsiedlung Malchow	Wilhelmsruh	Plänterwald
Lankwitz	Wittenau	Marienfelde	Schmargendorf	Stadtrandsiedlung Malchow	Plänterwald	Wilhelmsruh
Schmargendorf	Karlshorst	Lankwitz	Alt-Treptow	Stadtrandsiedlung Malchow	Plänterwald	Wilhelmsruh
Dahlem	Wittenau	Niederschöneweide	Oberschöneweide	Stadtrandsiedlung Malchow	Plänterwald	Heiligensee
Schmöckwitz	Gatow	Grünau	Nikolassee	Neukölln	Prenzlauer Berg	Kreuzberg
Lichtenberg	Britz	Friedrichsfelde	Staaken	Stadtrandsiedlung Malchow	Malchow	Falkenberg
Marzahn	Lichterfelde	Hellersdorf	Mitte	Malchow	Stadtrandsiedlung Malchow	Plänterwald
Charlottenburg-Nord	Tiergarten	Wilhelmstadt	Nikolassee	Malchow	Prenzlauer Berg	Neukölln
Malchow	Lübars	Oberschöneweide	Borsigwalde	Stadtrandsiedlung Malchow	Plänterwald	Heiligensee
Tempelhof	Mariendorf	Spandau	Lichtenrade	Stadtrandsiedlung Malchow	Wilhelmsruh	Plänterwald
Charlottenburg	Schöneberg	Kreuzberg	Friedrichshain	Stadtrandsiedlung Malchow	Plänterwald	Wilhelmsruh
Rosenthal	Kaulsdorf	Französisch Buchholz	Plänterwald	Stadtrandsiedlung Malchow	Malchow	Neukölln
Rummelsburg	Weißensee	Alt-Treptow	Johannisthal	Stadtrandsiedlung Malchow	Heiligensee	Wartenberg
Weißensee	Rummelsburg	Lichterfelde	Niederschönhausen	Stadtrandsiedlung Malchow	Heiligensee	Wartenberg
Haselhorst	Französisch Buchholz	Hakenfelde	Waidmannslust	Malchow	Stadtrandsiedlung Malchow	Lübars

Figure 15: The most and least similar localities for a sample of 32 localities.

	Count	Similar Localities
Lichtenrade	8	Alt-Hohenschönhausen, Buckow, Falkenhagener Feld, Mariendorf, Niederschönhausen, Rudow, Tempelhof, Zehlendorf
Mariendorf	7	Alt-Hohenschönhausen, Buckow, Falkenhagener Feld, Lichtenrade, Spandau, Tempelhof, Zehlendorf
Kreuzberg	7	Charlottenburg, Friedrichshain, Gesundbrunnen, Mitte, Neukölln, Prenzlauer Berg, Schöneberg
Heinersdorf	6	Altglienicke, Buch, Französisch Buchholz, Heiligensee, Siemensstadt, Waidmannslust
Nikolassee	6	Charlottenburg-Nord, Gatow, Grunewald, Grünau, Schmöckwitz, Wannsee
Niederschönhausen	6	Baumschulenweg, Karlshorst, Karow, Marienfelde, Staaken, Weißensee
Frohnau	6	Blankenburg, Borsigwalde, Friedrichshagen, Konradshöhe, Lübars, Müggelheim
Friedrichshagen	6	Blankenburg, Frohnau, Konradshöhe, Müggelheim, Rahnsdorf, Tegel
Alt-Hohenschönhausen	6	Buckow, Falkenhagener Feld, Lichtenrade, Mariendorf, Spandau, Staaken
Friedrichshain	6	Charlottenburg, Gesundbrunnen, Kreuzberg, Neukölln, Prenzlauer Berg, Schöneberg
Hellersdorf	5	Friedrichsfelde, Marzahn, Neu-Hohenschönhausen, Reinickendorf, Steglitz
Adlershof	5	Baumschulenweg, Biesdorf, Bohnsdorf, Hermsdorf, Waidmannslust
Dahlem	5	Johannisthal, Lübars, Niederschöneweide, Oberschöneweide, Wittenau
Wittenau	5	Dahlem, Lankwitz, Niederschöneweide, Oberschöneweide, Westend
Grünau	5	Gatow, Kladow, Nikolassee, Schmöckwitz, Wannsee
Gatow	5	Grunewald, Kaulsdorf, Nikolassee, Schmöckwitz, Wilhelmsruh
Altglienicke	5	Französisch Buchholz, Heiligensee, Heinersdorf, Mahlsdorf, Siemensstadt

Figure 16: The most typical localities.

	Number of dissimilar localities
Stadtrandsiedlung Malchow	79
Plänterwald	46
Malchow	35
Wilhelmsruh	34
Neukölln	23
Heiligensee	20
Falkenberg	20
Prenzlauer Berg	11
Kreuzberg	7
Wartenberg	6

Figure 17: The most unique localities.

The most typical localities of Berlin are Lichtenrade (district Tempelhof-Schöneberg), Mariendorf (Tempelhof-Schöneberg) and Kreuzberg (Friedrichshain-Kreuzberg). Of these, the suburbs Lichtenrade and Mariendorf are also most similar to each other, have primarily venues related to shopping and food, and an older, mostly German population. Kreuzberg is on the other hand one of the most populous localities with a relatively large fraction of foreigners, and venues related to food and nightlife activities. All of its most similar localities are located close to the city centre.

The most unique locality of Berlin is found to be Stadtrandsiedlung Malchow (district Pankow). This seems to be a result of the fact that the locality contains only a single venue, which is of type "parks_outdoors". Hence, no wonder that the locality consistently shows up as the most dissimilar. On the other hand, denoting a locality as most unique based on a single venue is rather dubious. Either discarding localities with very few venues, or accounting for the absolute number of each venue category, instead of only

Venues Cluster	Demographics Cluster	Count
1	1	2
	2	12
	4	8
2	1	5
	2	13
	3	9
	4	12
3	1	7
	2	18
	4	10

Figure 18: Number of localities for each K-means cluster combination.

their mean frequency, should eliminate such artificial results.

6 Conclusion

The analysis of social venues is a promising approach to identifying similar localities (neighborhoods) within a city or even across different cities. Identifying similar localities would be of interest to businesses, allowing them to identify best locations for expansion or, in case business is not going so well, relocation, as well as people considering a change of residence.

This work explores which localities of Berlin are similar according to their social venues and demographics. These are of course only two of the aspects that a person should consider when moving their business or changing their residence. Some important additional examples are rent prices, median income level, and crime statistics. Including these metrics would allow for a more accurate identification of similar neighborhoods.

References

1. Wikipedia: Administrative structure of Berlin
   https://de.wikipedia.org/w/index.php?title=Verwaltungsgliederung_Berlins&oldid=198162445
2. Foursquare API
   https://developer.foursquare.com/
3. Amt für Statistik Berlin-Brandenburg - Einwohnerregisterstatistik
   https://www.statistik-berlin-brandenburg.de/opendata/EWR_Ortsteile_2018-12-31.csv