Segmenting and Clustering Neighborhoods in Berlin
Is Berlin still a divided City?

by Benjamin Bachstein, February 2020

Introduction

Berlin, the capital of Germany, celebrates its 30th anniversary of reunification this year. After almost thirty years of division into a capitalistic western and a socialistic eastern part, the quarters and neighborhoods of Berlin had developed quite differently. In the years after the reunification, the convergence of the eastern and western parts of the city towards a more homogenous, united city, was one of the important political goals.

Therefore, now is a good moment to analyze the similarities and differences of Berlin's boroughs and neighborhoods. There are many approaches to do so. While it is relatively easy to compare figures such as the average income per person or the unemployment rate per borough (and in these terms, the differences between the eastern and western parts of the city are still obvious), it is a more difficult task to compare the "look and feel" of the boroughs and neighborhoods today.

Data

My approach to do this will be by using the Foursquare API to find out about the most common venue categories in the respective boroughs and neighborhoods. This analysis, supported by the official latitude / longitude data of the Berlin districts, is supposed to lead into a k-means-clustering of the neighborhoods. Is there something like a "typical group" of western and eastern districts? Or will we see a mixed picture of similar boroughs and neighborhoods on both sides of the former Berlin wall? Finally, the answer to this question is going to be illustrated on a Folium map of Berlin.

Methodology

The whole analysis was conducted using the programming language “Python 3” and a so-called Jupyter Notebook, which can be accessed on GitHub. Within the notebook, a number of modules were imported, some of which were:

- Pandas and Numpy: libraries for data handling and analysis
- JSON: a library for handling (in this case geographic) data in the JSON format
- GeoPy: a library for converting address information into coordinates
- Matplotlib: a library to create graphics and plots
- Sklearn: a library that contains machine learning algorithms
- Folium: a library that lets the user create animated maps

After importing the libraries, the first task was to find and download GeoJSON data on the boroughs and neighborhoods of Berlin. Fortunately, these data was easy to

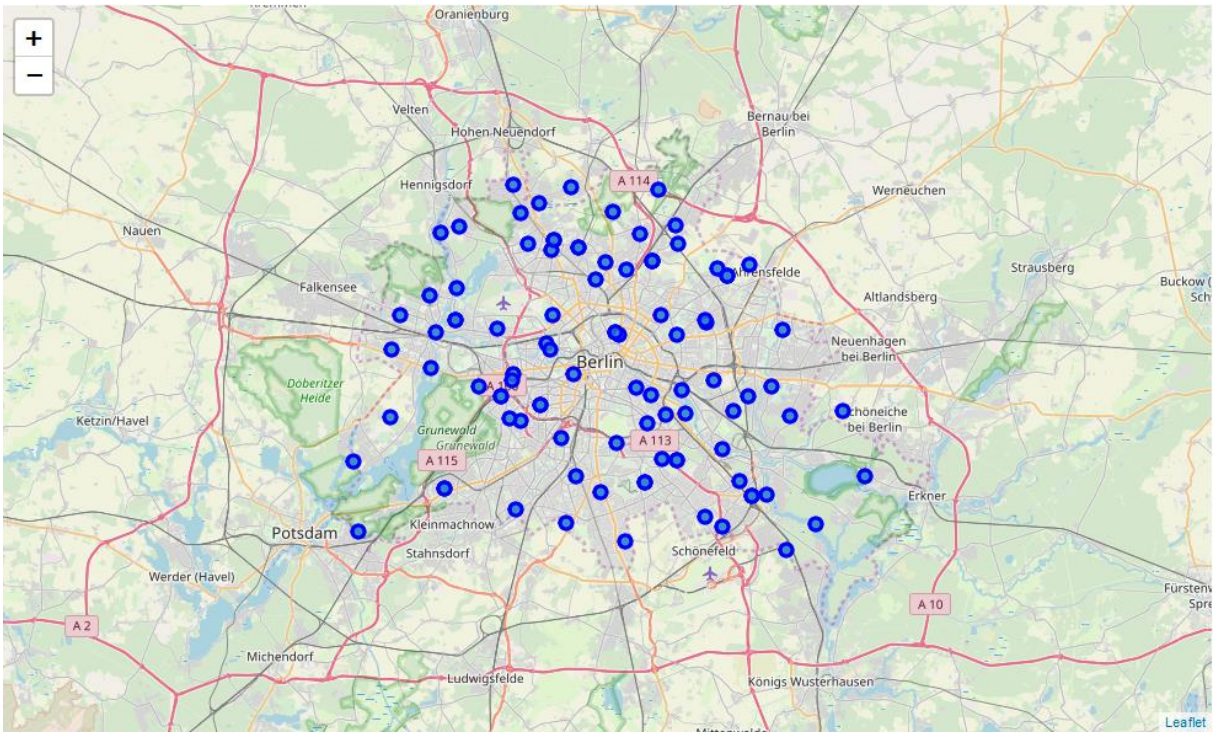
locate and was taken from: https://data.technologiestiftung-berlin.de/dataset/lor_bezirksregionen

In further steps, this data was processed and converted into a so-called Pandas dataframe, which is basically the gold-standard relational table for data analysis in Python. This dataframe looked as follows:

	Borough	Neighborhood	Latitude	Longitude
0	Mitte	Mitte	52.536962	13.406490
1	Mitte	Moabit	52.529735	13.328836
2	Mitte	Hansaviertel	52.525565	13.333217
3	Mitte	Tiergarten	52.508781	13.358794
4	Mitte	Wedding	52.548789	13.336565
5	Mitte	Gesundbrunnen	52.573386	13.384491
6	Friedrichshain-Kreuzberg	Friedrichshain	52.535546	13.409753
7	Friedrichshain-Kreuzberg	Kreuzberg	52.499610	13.429264
8	Pankow	Prenzlauer Berg	52.536962	13.406490
9	Pankow	WeiÃŸensee	52.548495	13.457310

and contained, in its first version, borough and neighborhood names and the respective coordinates. In total, the dataframe consisted of 12 boroughs and 96 neighborhoods.

The second part of the analysis comprised the creation of a Berlin map with its neighborhoods superimposed on top. For this task, to coordinates of the center of Berlin were determined using GeoPy as being 52.5170365, 13.3888599. These coordinates served as the center point for a folium map to which the neighborhoods from the pandas dataframe were added as circle markers. The resulting map looked as follows:

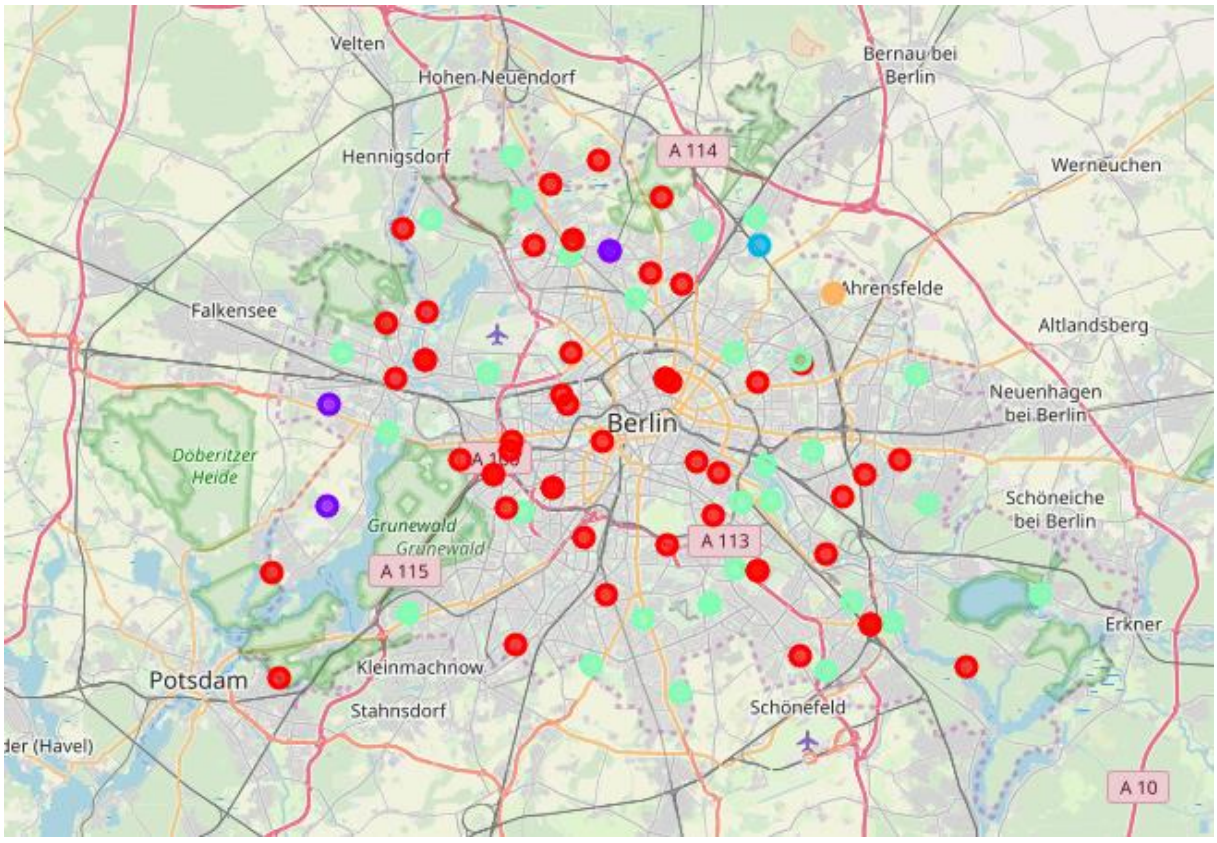


After this preparatory work was finished, the FourSquare API was used to find up to 100 venues within a distance of up to 1.000 meters for all the 96 neighborhoods. Actually, the limit of 100 venues was never reached. The requests yielded 1.114 hits in total, averaging to 11.6 venues per neighborhood, ranging from as little as only one venue up to 92 venues for a single neighborhood.

The found venue data was processed in multiple steps, finally leading into a table with the then most common venues per neighborhood. This table looked as follows:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adlershof	Greek Restaurant	Italian Restaurant	Trattoria/Osteria	Supermarket	Bank	Light Rail Station	Drugstore	Electronics Store	Food & Drink Shop	Flower Shop
1	Alt-HohenschÃ¶nhausen	Tram Station	Coffee Shop	Pharmacy	Drugstore	Doner Restaurant	Discount Store	Hardware Store	Asian Restaurant	Supermarket	Fast Food Restaurant
2	Alt-Treptow	Café	German Restaurant	Park	Bar	Pizza Place	Beer Garden	Sporting Goods Shop	Speakeasy	Soup Place	Lounge
3	Altglienicke	Train Station	Harbor / Marina	Auto Workshop	Yoga Studio	Eastern European Restaurant	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Fish Market
4	Baumschulenweg	Garden Center	Bakery	Café	Organic Grocery	Yoga Studio	Electronics Store	Food & Drink Shop	Flower Shop	Flea Market	Fish Market

Based on this table, a k-means clustering was conducted to group the 96 neighborhoods into similar clusters. After numerous attempts, a k of 5 clusters was determined to lead to reasonable results. With the cluster-information added to the previously mentioned table, the folium map was re-created with different colors for the respective neighborhood clusters. The new map looked as follows:



Results and Discussion

The results of the k-means clustering are to some extent surprising. Although the venue types that were found do not seem all too different to me, the algorithm managed to find clusters that make sense - mostly. I decided to name the clusters as follows:

Cluster 1 - The "big" city center and some minor centers around it

This is, where the life is - the world-famous Berlin center with all the places and venues that attract tourists as well as people who live there. These neighborhoods are strongly concentrated in the city center, but, as Berlin is a city of many centers, also occur in other places around the inner city. However, it is worth to note, that these "red dots" are less frequent not only the further one gets away from the middle of the city, but also the further one gets to the east. In the north-east there are none of them at all, which could be taken as a hint that this is not the most attractive region of the city.

Cluster 2 - "villages" on the outer city borders

This little number of neighborhoods were originally villages which were, at some point, politically integrated into the city of Berlin. Nevertheless, they seem to have preserved their village-like character to some extent, at least enabling the venue-comparing algorithm to identify them as a group that's slightly different from the rest.

Cluster 3 - an outlier?

This one-neighborhood cluster has no obvious justification of existence to me. Nevertheless, reducing the number of clusters led to completely different results, so I decided to stick to k=5.

Cluster 4 - a place to live - not for party

This second really big cluster represents to me neighborhoods where usually high numbers of people live. These places are dominated by venues that are usually needed and used primarily by the inhabitants of the neighborhoods (like supermarkets, traffic infrastructure, local restaurants, etc.) and do not aim too much at tourists as customers. The neighborhoods of cluster 4 are never in the city center but can be found in the sub-urban areas in all cardinal directions. Additionally, it can be said that the eastern part of the city has a higher number of cluster 4 neighborhoods than the west.

Cluster 5 - socialistic mass-housing

The two Neighborhoods contained in this cluster represent areas where there are many examples of former socialistic mass-housing, meaning buildings that provided (and still provide) an affordable place to live for many people on limited space. To me, it is not quite clear why exactly these two neighborhoods stand out from the rest so much that they "deserve" their own cluster, but for the fundamental questions of this analysis, this doesn't seem two important. Just as with cluster 3, I will accept it without further examination.

Conclusion

The initial question, that is still waiting to be answered, was: Is Berlin still a divided city? I think, that from what can be seen in this analysis, there is no absolute yes or no as an answer.

Yes, Berlin is divided in a sense, that the hip and world-famous parts in the center of the city stand in contrast to less touristically attractive neighborhoods in the sub-urban areas, which are home to a high percentage of the Berlin inhabitants.

Yes, Berlin is also divided in a way, that some of the eastern parts of the city still cannot offer the same level of quality of life as many western parts can - at least from a venues point of view.

But: No, the Berlin wall is no longer clearly visible on the map we created here. Berlins attractive center today consists of former eastern and western parts and there are attractive and rather boring neighborhoods on both sides of the former Berlin wall. Thirty years of (at least partial and politically enforced) convergence in a re-united city have clearly left their footprint, especially in the cityscape. The factors that are still dividing the city usually lie deeper and consist of less wealth and income in some of the eastern parts, a higher dependency on social transfers, less well-paid jobs, etc., just to mention a few. But those kinds of statistics were not in the scope of this analysis.