CAPSTONE PROJECT: CLUSTERING BERLIN – WHERE SHOULD I OPEN MY FANCY COFFEE SHOP?

Contents

1. Problem description: Coffee shop in Berlin – but where?

Berlin, the capital of Germany, is one of the most interesting towns in Europe. Once dubbed “poor but sexy” by one of its mayors, it is a capital that is famous for its nightlife, freewheeling alternative crowds, art galleries and, especially recently, its vibrant tech and startup scene crowded by expats.

Our problem is simple and challenging at the same time: We would like to open a coffee shop in Berlin. It should be high-end in terms of the pricing and interior design: A place where people enjoy having their coffee and relaxing, but also a place known for its good coffee, catering to people who buy take-away. We are looking for a good spot to open the coffee shop and that’s where the matter gets a bit complicated.

Berlin -historically- is a merger between smaller towns that outgrew their vicinity and melted into one large metropolitan region. As such, there is no one city center as in many other towns. Moreover, each neighborhood has its own unique micro-culture. Berlin also was a divided town after the second world war until 1990 and the iron curtain went straight through it: East Berlin belonged to the socialist German Democratic Republic whereas the western part was divided into allied zones of US, British and French administration.

There is no straight answer to our coffee shop question as the city is very diverse. That’s why we would like to analyze the city neighborhoods in terms of:

- Population
- Household purchasing power / income
- Political views
- Number of coffee shops

We don’t have an hypothesis yet as to what is desirable so we’ll stick with unsupervised learning: We would like to cluster the neighborhoods based on the above variables to see if there are any patterns in the clusters we find. Hopefully the insights we gain from will help us decide on the location. As a unit for neighborhood, we’ll use PLZ (Abbreviation for Postleitzahl, German word for zip code):

- PLZ

In the next section, we’ll talk about the sources of data for the above-mentioned variables.

2. Data: Description and sources

For the cluster analysis of Berlin neighborhoods based on their PLZ codes, we will need the following data, all coming from different sources:

2.1. PLZ-data with respective geo-coordinates:

PLZ data for Germany as a whole is available as a JSON file under the following link: https://www.wiwald.com/ds/kostenlose-liste-deutscher-postleitzahlen-und-zugehoeriger-orte/id/ww-german-postal-codes

We download this and put it into a pandas dataframe, drop everything that is not Berlin. This is what we get:

	primary_key	zipcode	city	state	community	latitude	longitude
1288	1288	10115	Berlin	Berlin	Berlin, Stadt	52.5323	13.3846
1289	1289	10117	Berlin	Berlin	Berlin, Stadt	52.5170	13.3872
1290	1290	10119	Berlin	Berlin	Berlin, Stadt	52.5305	13.4053
1291	1291	10178	Berlin	Berlin	Berlin, Stadt	52.5213	13.4096
1292	1292	10179	Berlin	Berlin	Berlin, Stadt	52.5122	13.4164

2.2. Population data by PLZ code:

This is available on the website of the regional statistics agency, Statistik Berlin-Brandenburg: https://www.statistik-berlin-brandenburg.de/publikationen/Stat_Berichte/2018/SB_A01-05-00_2017h02_BE.xlsx

This is a great source, because it not only gives us the population by PLZ, but also the neighborhood ('Bezirk' or 'Kiez' in German) which the PLZ belongs to. The dataframe, once read with pandas, looks like this:

PLZ	Einwohner	Kiez
10115	26274	Mitte
10117	15531	Mitte
10119	19670	MittePankow
10178	14466	Friedrichsh.-Kreuzb.Mitte
10179	23970	Friedrichsh.-Kreuzb.Mitte

2.3. Purchasing power / household income data:

Now this is where we have to be a bit more creative as data is not easily available. Luckily enough, the German tabloid newspaper Bild had a study which is available on its website: https://www.bild.de/regional/berlin/einkommen/kaufkraft-liste-berliner-kieze-28373092.bild.html

It shows household income by PLZ – we scrape it and after some cleaning-up and data type conversion, it looks like this:

	PLZ, Lage	Kaufkraft pro Haushalt*	Rang**	PLZ	Kaufkraft
0	10585 Deutsche Oper	3015 Euro	45	10585	3015
1	10587 TU/Otto-Suhr-Allee	2753 Euro	81	10587	2753
2	10589 Mierendorffplatz	2718 Euro	90	10589	2718
3	10623 Savignyplatz	3545 Euro	15	10623	3545
4	10625 Karl-August-Platz	3048 Euro	40	10625	3048

2.4. Election results:

As a proxy for political leanings of neighborhood populations, we take the election results from the local elections in 2016 for Berlin senate & municipalities. The data is available under the local govt election website, https://www.wahlen-berlin.de/ as a csv. However, there is a small problem: Election zone splits do not correspond 1:1 with PLZ so that some manual mapping was necessary, also based on data available on the same website- when we do the work and clean-up the data and group by PLZ, we get what we want: % votes of the 6 largest parties per PLZ. SPD are the social democrats, CDU is conservative center-right, FDP are the liberals, GRÜNE are the Greens, DIE LINKE are socialists and, AfD is the far-right party. It will be interesting to see if there will be a relationship between coffee shop density and political leanings. Once all cleaned up, the dataframe looks like this:

PLZ	Bezirksname	Gültige Stimmen	SPD	CDU	GRÜNE	DIE LINKE	FDP	AfD
10115	Mitte	2958	0.244760	0.168357	0.262677	0.125085	0.090264	0.078431
10117	Mitte	8971	0.230521	0.156170	0.177461	0.216475	0.072902	0.113700
10119	Pankow	3563	0.213865	0.122930	0.336514	0.174011	0.074937	0.044906
10178	Mitte	1544	0.253238	0.120466	0.102332	0.312824	0.040803	0.139896
10179	Mitte	7142	0.259311	0.113974	0.145477	0.279194	0.041025	0.125175

2.5. Foursquare data: Cafés and coffee shop by PLZ:

This is where we utilize the Foursquare API. We make an API request for each PLZ to get the nearby coffee shops and cafés within a certain radius of the PLZ central coordinates. We get 2 large JSON files (1 for cafés and 1 for coffee shops), put them into pandas dataframes, merge, remove duplicates and group by number of cafés to finally get what we want: The number of cafés and coffe shops per PLZ.

PLZ	Venue
10115	21
10117	69
10119	68
10178	103
10179	20

2.6. Putting it all together:

Now we have 5 dataframes that we have to join based on PLZ. Once we do this task of doing several inner joins of the dataframes, we get a nice-looking structured dataset that we can finally work with:

PLZ	latitude	longitude	income	population	Kiez	SPD	CDU	GRÜNE	DIE LINKE	FDP	AfD	number_of_cafes
10115	52.5323	13.3846	3118.0	26274.0	Mitte	0.244760	0.168357	0.262677	0.125085	0.090264	0.078431	21.0
10117	52.5170	13.3872	3673.0	15531.0	Mitte	0.230521	0.156170	0.177461	0.216475	0.072902	0.113700	69.0
10119	52.5305	13.4053	3018.0	19670.0	MittePankow	0.213865	0.122930	0.336514	0.174011	0.074937	0.044906	68.0
10178	52.5213	13.4096	2717.0	14466.0	Friedrichsh.-Kreuzb.Mitte	0.253238	0.120466	0.102332	0.312824	0.040803	0.139896	103.0
10179	52.5122	13.4164	2583.0	23970.0	Friedrichsh.-Kreuzb.Mitte	0.259311	0.113974	0.145477	0.279194	0.041025	0.125175	20.0
10243	52.5123	13.4394	2428.0	30655.0	Friedrichsh.-Kreuzb.	0.185139	0.080490	0.259760	0.245244	0.031075	0.087722	6.0
10245	52.5007	13.4647	2439.0	33509.0	Friedrichsh.-Kreuzb.	0.171369	0.077022	0.312147	0.216711	0.035244	0.067221	6.0
10247	52.5161	13.4656	2463.0	39491.0	Friedrichsh.-Kreuzb.Pankow	0.156329	0.065162	0.327305	0.221043	0.025407	0.058736	25.0
10249	52.5238	13.4428	2436.0	28885.0	Friedrichsh.-Kreuzb.Pankow	0.205592	0.092083	0.210302	0.274578	0.037836	0.112293	6.0
10315	52.5132	13.5148	2151.0	33424.0	Lichtenberg	0.227868	0.111517	0.071133	0.309392	0.024862	0.204617	4.0
10317	52.4979	13.4908	2342.0	23027.0	Friedrichsh.-Kreuzb.Lichtenberg	0.222026	0.102065	0.189577	0.270600	0.034022	0.121731	3.0
10318	52.4835	13.5287	2690.0	27217.0	Lichtenberg	0.269375	0.160280	0.101124	0.258043	0.033228	0.145587	8.0
10319	52.4992	13.5188	2104.0	24481.0	Lichtenberg	0.248000	0.198000	0.158000	0.154000	0.056000	0.141000	0.0
10365	52.5206	13.4969	2278.0	27052.0	Lichtenberg	0.213463	0.092585	0.106927	0.321659	0.026829	0.185659	0.0
10367	52.5246	13.4821	2192.0	21735.0	Lichtenberg	0.225220	0.113551	0.136763	0.274153	0.039523	0.156211	1.0
10369	52.5295	13.4695	2286.0	20386.0	Lichtenberg	0.234551	0.102902	0.112843	0.260881	0.030091	0.196131	4.0
10405	52.5352	13.4257	2773.0	32065.0	Pankow	0.257936	0.110728	0.245075	0.208185	0.046244	0.067992	39.0
10407	52.5336	13.4492	2555.0	25254.0	Pankow	0.180162	0.097166	0.281377	0.232794	0.046559	0.087045	4.0

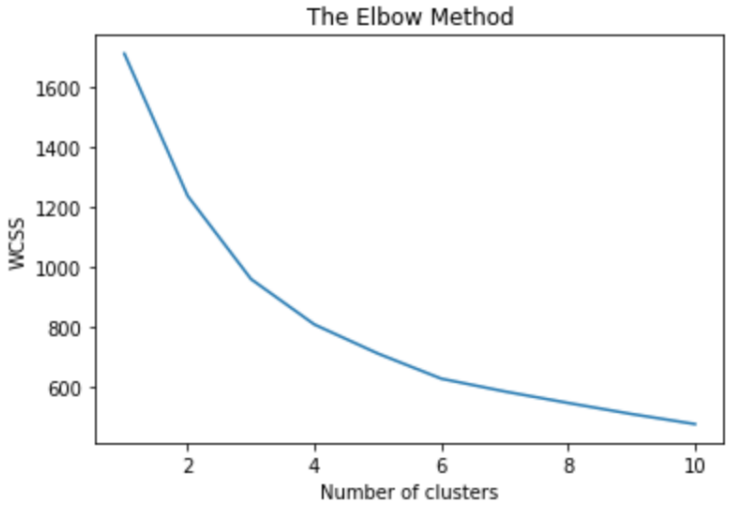
Finally, before we start the clustering analysis, we standardize the relecant variables using the scikit learn preprocessing library's StandardScaler. Now we have a dataset that is ready to be clustered:

PLZ	latitude	longitude	income	population	Kiez	SPD	CDU	GRÜNE	DIE LINKE	FDP	AfD	number_of_cafes
10115	52.5323	13.3846	0.795168	0.952178	Mitte	-0.210863	-0.235586	1.075738	-0.303655	1.138978	-0.982463	0.833516
10117	52.5170	13.3872	2.028190	-0.566072	Mitte	-0.543981	-0.388277	0.120688	0.889122	0.565641	-0.430709	3.957918
10119	52.5305	13.4053	0.573001	0.018871	MittePankow	-0.933630	-0.804735	1.903248	0.334899	0.632851	-1.506950	3.892826
10178	52.5213	13.4096	-0.095719	-0.716582	Friedrichsh.-Kreuzb.Mitte	-0.012519	-0.835603	-0.721307	2.146605	-0.494324	-0.020876	6.171036
10179	52.5122	13.4164	-0.393421	0.626566	Friedrichsh.-Kreuzb.Mitte	0.129548	-0.916949	-0.237758	1.707682	-0.486999	-0.251184	0.768425

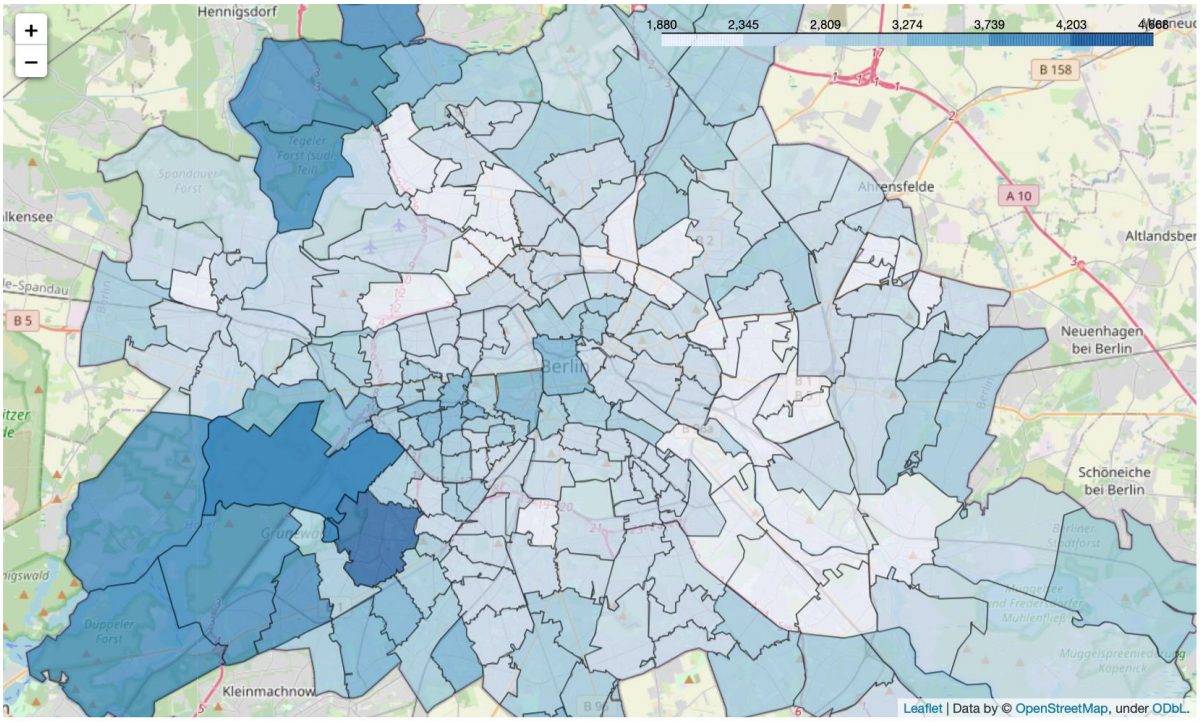
3. Methodology

As we do not have a labeled dataset with a clear target variable, we go for a cluster analysis and decide to do a K-means clustering of PLZ areas. The variables we'll be using are population, household income, % votes for each of the 6 parties and number of cafés. For that, we use the scikit-learn library and specifically K-Means.

In order to get the optimal number of clusters (the "right" K in K-means, we utilize the so-called elbow method and depict WCSS (within cluster sum of squares) against number of clusters. We also do this using scikit-learn, for each K we calculate the WCSS and depict it against the number of clusters. We choose the number where the curve makes an "elbow" i.e. when the change in slope of the curve with increasing K significantly decreases after a certain K – this appears to be at K=6 in our case:



An exploratory analysis that we can do before we do the clustering is to show purchasing power by PLZ on a map to see if there is a concentration of high purchasing power areas. The map is below, and there indeed is a concentration: West Berlin suburbs appear to be affluent, especially to the south-west. Also, there are 3 neighborhoods to the north of Tegel airport with high purchasing power. In the central parts, West Berlin appears to be richer whereas there is a "ring" of poorer areas to the east.



4. Results

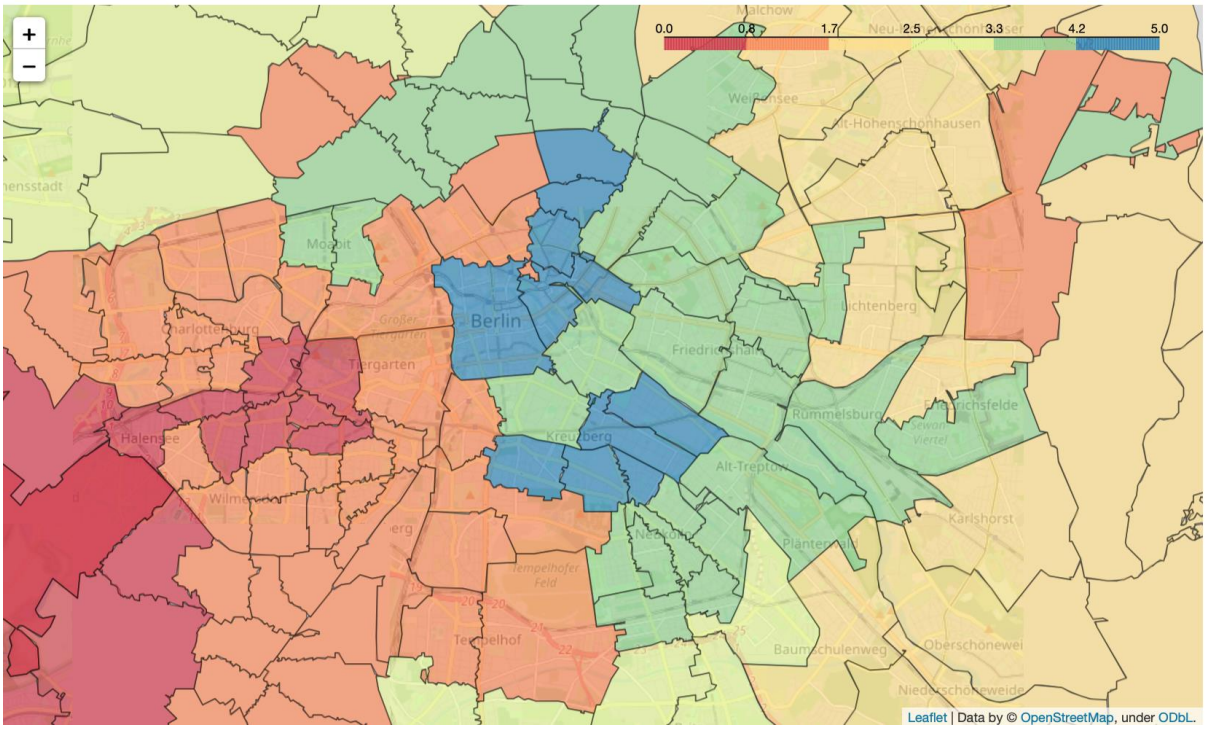
The K means clustering approach assigns each PLZ area to one of the 6 overall clusters that the machine has discovered for us. When we group the PLZ codes based on their cluster and compare averages for population, income, number of coffee shops and % party votes, we indeed see that the clusters that we found are unique. They help us make sense of the different characteristics of neighborhoods to

support us in our decision as to where to open our new fancy coffe shop. Here is what the clusters look like:

cluster	latitude	longitude	income	population	SPD	CDU	GRÜNE	DIE LINKE	FDP	AfD	number_of_cafes	PLZ
0	52.484915	13.284307	3543.296296	14144.000000	0.250494	0.282043	0.164814	0.066754	0.105535	0.111677	6.962963	12753.370370
1	52.497458	13.344867	2763.250000	16075.187500	0.275193	0.175457	0.225062	0.120458	0.062646	0.101202	7.354167	11824.541667
2	52.513995	13.532224	2564.736842	21381.421053	0.222846	0.164129	0.072669	0.234640	0.023014	0.231119	1.263158	12515.368421
3	52.507223	13.318334	2623.857143	20346.542857	0.295699	0.259662	0.090003	0.072123	0.066926	0.181405	0.742857	13017.342857
4	52.515734	13.432853	2451.250000	25743.406250	0.230387	0.103602	0.225818	0.219064	0.034831	0.104523	10.437500	11581.406250
5	52.512290	13.413050	2837.600000	21002.100000	0.205435	0.088303	0.328085	0.215162	0.040880	0.048443	60.800000	10725.700000

5. Discussion

In addition to the dataframe summarizing cluster averages I have visualized the clusters on a Choropleth map using a Geojson file available online for PLZ geo-data in Berlin. This is How it looks like:



The cluster analysis reveals some interesting insights:

Cluster 0: High-income / conseravtive neighborhood with low population (map: dark red): The PLZ areas in this cluster are parts of town with detached houses with high-income households who tend to elect conservative / center-right. Could potentially work for a high-end cafe if the design & interior is a bit more on the traditional side.

Cluster 1: Average-income "green" residential areas (map: orange) Similar to cluster 0 in terms of population and coffe shop density, while the political opinion is a bit more tilted toward the green party and the income is in-line with city average. A representative neighborhood for this segment would be Schöneberg / Tempelhof. Probably equally attractive as cluster 0, but the clientele will be completely different - need more vegan stuff!

Cluster 2: Far-right / low income neighborhoods (map: darker-yellow in the East): Parts of town with low household income, tendency to elect far-right, very low number of coffee shops. Also high % for far-left. Former East-Berlin neighborhoods outside the central areas. Not interesting as a spot for our coffee shop idea .

Cluster 3: Mixed neighborhoods (map: light-yellow): These neighborhoods are quite heterogeneous in that there is a large social democratic crowd but also a high number of far-right people. Quite far away from the commercial centers of Berlin, these neighborhoods are clustered to the north-west and further south to the city ring. Number of cafes quite low, so also not really interesting for our coffee shop idea.

Cluster 4: "Poor, but sexy" (map: green): Alternative cluster: This is interesting. Relatively lower purchasing power, but highly cosmopolitan, progressive and high-population. Probably student neighborhoods. Definitely interesting, but we would need to investigate the potential clientele more in terms of their "wallet size".

Cluster 5: Affluent porgressive (map: blue): This is the ideal cluster for our coffee shop idea. Former leftie / green students who now are affluent and love to spend time and money in alternative but expensive coffe shops! Concentrated in upper-middle class areas in terms of income, these neighborhoods boast a young, open-minded population that works hard and enjoys life.

6. Conclusion

The cluster analysis reveals very interesting insights that aid us in our decision around where to open a coffee shop. We seem to have a clear recommendation - we have found a cluster -**Affluent porgressive (map: blue)-** with relatively high purchasing power, alternative-left leaning population with potentially a large budget for good, probably organic coffee. The high density of coffee shops in these neighborhoods is also an indication that there is a market for us here – albeit with fierce competition. Neighborhoods in our favorite cluster are quite concentrated, too: 4 of them in Kreuzberg, in the artsy-alternative neighborhoods around Görlitzer Park and Schlesisches Tor, whereas the rest are in another very popular part of town: Former center of East Berlin around Mitte and Prenzlauer Berg. Incidentally, these are also the same neighborhoods that a local Berliner would recommend for a good night out.