Clustering Berlin: Where should I open my coffee shop?

Coursera Capstone Project

November 2020

Contents

Problem: Berlin is the ideal city to open a fancy coffee shop – but where should it be?

Problem description

We would like to open a coffee shop in Berlin, in the higher-end segment. Which neighborhood should we pick as a location?

Main challenges

- Berlin is quite heterogeneous in terms of neighborhoods due to ist historic background
  - Merger of smaller towns
  - Division after WW2
- No one„city center“ but several „Kiez“ centers
- Quite diverse clientele
  - Tech entrepreneurs / startup crowds
  - Students
  - Hardcore clubbers
  - „Poor but sexy“ artists
  - Affluent conservatives
  - …
- No straight answer to the question of optimal spot for a coffee shop, need to find your „sweet spot“, but we don‘t exactly know what we should look for

Approach

- Analyze neighborhoods based on
  - Purchasing power
  - Population (density)
  - Political leanings
  - Number / density of venues – in particular cafés & coffee shops
- Unsupervised learning
- Gain insights from data to support decision-making

So where do we get all the data from?

DATA	*SOURCE*	*FORMAT*
Neighborhoods: ZIP codes - PLZ	wiwald Daten nutzbar machen	JSON
Population	statistik Berlin Brandenburg	XLSX
Purchasing power	Bild	HTML
Political leanings: Election results	Berlin.de	CSV
Venue density: Coffee shop numbers	FOURSQUARE	JSON

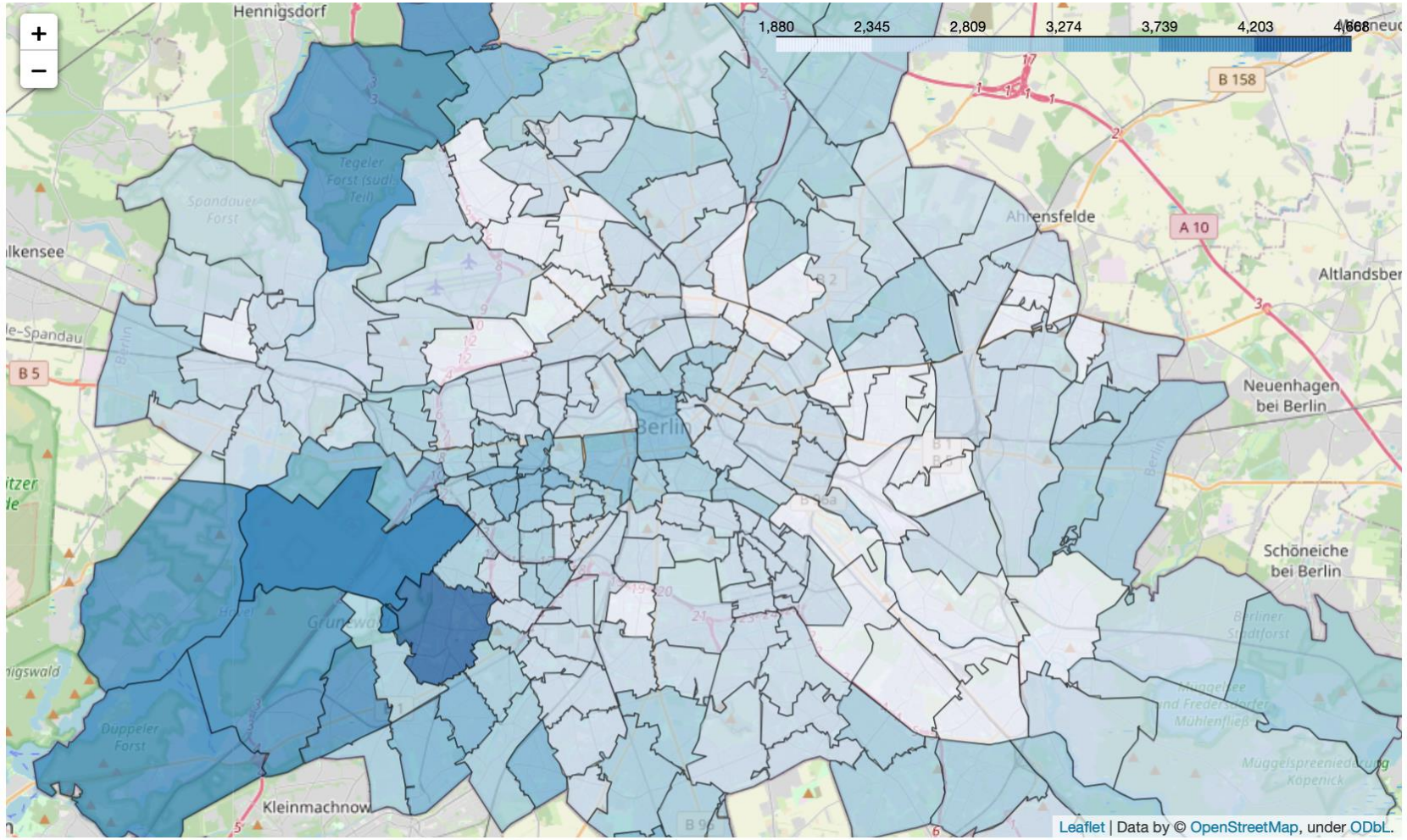
Putting it all together…

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

- PLZ (zip code) as unit of analysis - overall 195 PLZ in dataframe
- Coordinates lat /lng
- Population
- Kiez: Name of associated neighborhood – 1 Kiez has N PLZ
- % votes per political party
  - SPD = social democrat
  - GRÜNE = green
  - CDU: Center-right
  - LINKE: Socialist
  - FDP: Liberal
  - AfD: Far-right
- # of venues fro Foursquare API

Contents

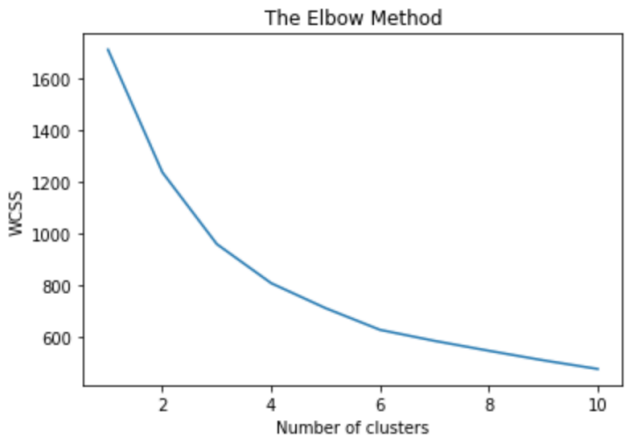
Exploratory analysis – income distribution



- Affluent neighborhoods primarily in the western suburbs
- Low income „ring“ around center in East-Berlin
- City-west more affluent than city-east

We did a K-means clustering to get 6 unique PLZ-clusters

Applied the elbow analysis to get the optimal number of clusters…



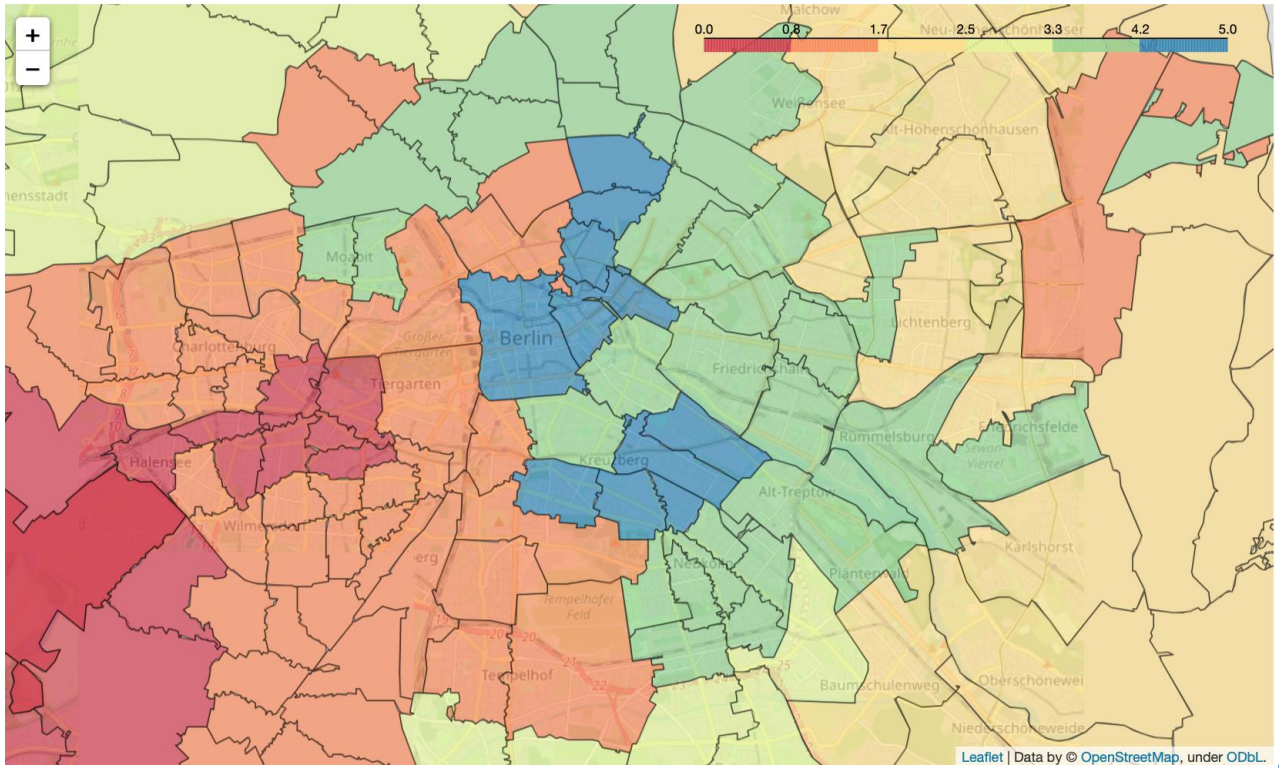
…and summarized data by cluster means to analyze results

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

Cluster descriptions in the next section

Contents

Clustering reveals valuable insights – We go for „*affluent progressive*“!



Cluster 0: High-income / conservative residential

Cluster 1: Average-income "green" residential

Cluster 2: Far-right / low income

Cluster 3: Mixed suburbs

Cluster 4: Poor but sexy

Cluster 5: Affluent progressive

Looks like the most attractive option: Affluent clientele, open, multi-cultural neighborhoods, high density of venues