Capstone - Berlin Venues with crimes and neighborhood aggregated

Author: Javier Inocente

Publication: November, 2020

Introduction

Background

In the last years, Berlin has become more interesting for tech start-ups investors[1]. To satisfy the demand for an increasing number of young people, the capital of Germany has attracted many other diverse businesses. The idea of this research is to cluster venues in Berlin to make recommendations to investors interested to start a new business in Berlin. This project is based on the capsote Toronto project with the addition of two more data sources, namely the Berlins crimes in 2019 and the Berlin population density by the same year. The data is described below.

Observations: this essay may be strongly affected after the corona lockdown at the finish of 2020.

Problem

At the time to invest in a fiscal shop, investors looks for the best places.
I made three assumptions to face this research

1. Like Hotelling's Law[2] suggest, competing firms are located close each other. The Reason why is like this is out of scope of this research. This agglomeration is named clustering, therefore it is possible to elucidate that the methodology used to find a solution to this problem is Cauterization, but the method will be covered in detail in the corresponding section.
2. The density of population play an important role. Either if the shop what to attract many people as possible or if the business looks for peace and calm as differentiating point.
3. Security is a last but absolutely not the least factor to be taken into consideration at the time to find a plate to run a business, especially in large cities like Berlin, this factor will be determinant at the moment to settle down.

Audience Interest

This project may be interest to:

1. Small entrepreneurs of any kind who are interested in settled done in Berlin.
2. Financiers, aiming to invest in the City.
3. Governmental entities willing to facilitate the funding of many from private investors
4. Public looking for radiate in Berlin and wants to know more where to live.

---

[1] https://www.berlin.de/en/business-and-economy/business-developement/5615902-5886496-for-startups-and-founders.en.html

[2] https://en.wikipedia.org/wiki/Hotelling%27s_law

Data

Source

As described above the three main factors playing significant role at the moment to choose a place to mount a business are, similar venues' closeness, density of population and crime rates in the zone. This research will use three sources of data:

- Population: Berlin's neighborhood with the corresponding zip codes and population density are can be found in Wikipedia[3]
- Crimes: Classification of crimes in Berlin, sorted by Neighborhood, can be found in a CSV on Klagge[4].
- Venues: it will be used Foursquare.

Cleaning

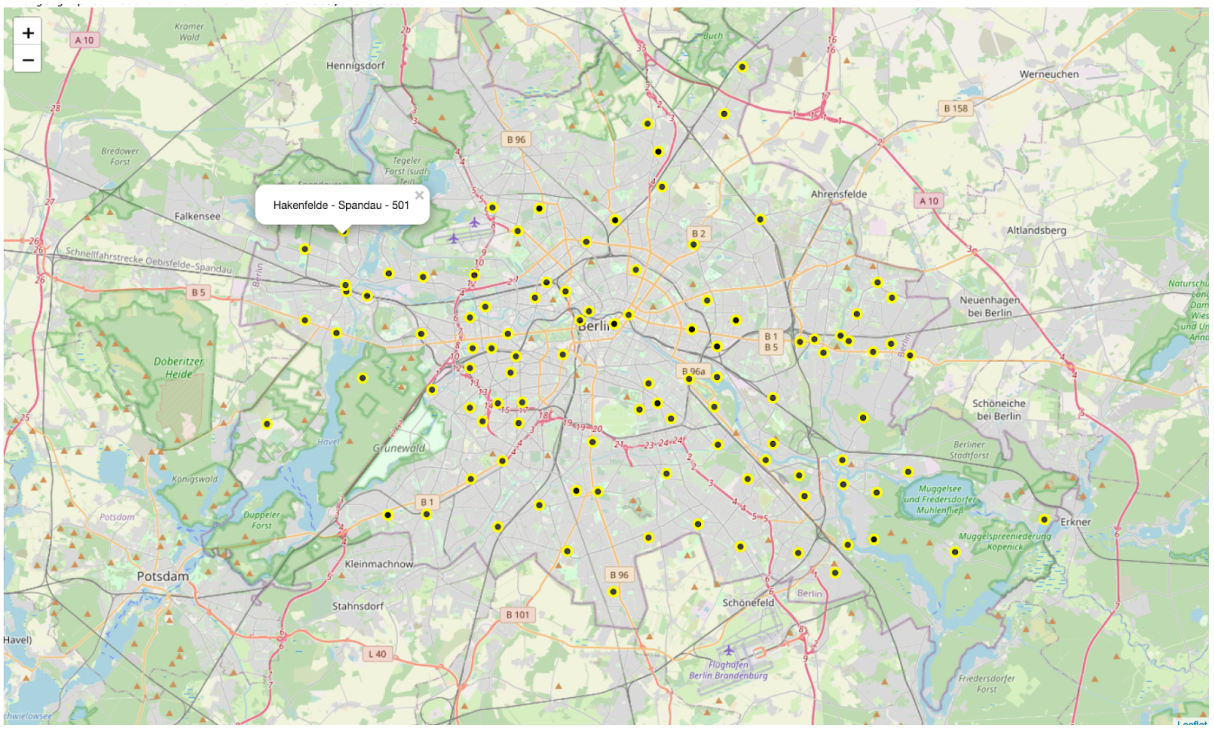
I got first the data from the Berlin population. This table was clean and it was not necessary to do anything more.
The second table was the Berlin crimes. The data contains information from 2016 to 2019, however I took only data from 2019.
I joined these two tables over the zip code resulting a table with zip code and Berlin district, then I used this information to enter in geolocator getting the latitudes and longitudes for each venue. Geolocator was not able to find all entries, for this reason I rearrange the data of the missing values to get the location information.
As a result, I got a table like this:

Table 1. Berlin Data with crimes, Density and location

	District	Code	Location	Robbery	Street_robbery	Injury	Agg_assault	Threat	Theft	Car	From_car	Bike	Burglary	Fire	Arson	Damage	Graffiti	Drugs	Local	Density	Latitud	Longitud
0	Mitte	101	Tiergarten Süd	60	35	365	92	128	2271	15	198	296	55	13	6	347	77	231	980	9576.000	52.5038	13.3634
1	Mitte	101	Regierungsviertel	42	20	554	136	152	3692	13	172	352	22	19	4	497	162	170	1057	9576.000	52.5195	13.3765
2	Mitte	101	Alexanderplatz	173	102	1966	500	420	11233	63	587	940	137	43	12	1307	381	1133	3813	9576.000	52.522	13.4136
3	Mitte	101	Brunnenstraße Süd	40	29	268	64	79	1859	39	182	361	64	18	7	424	172	86	902	9576.000	52.5177	13.4024
4	Mitte	102	Moabit West	66	29	685	210	202	2107	47	322	326	93	28	15	641	91	618	1409	10.427	52.5301	13.3425

With this info it is possible to draw a map:

[3] https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins
[4] https://www.kaggle.com/danilzyryanov/crime-in-berlin-2012-2019/version/4



Map 1: Berlin Neighborhoods.

I considered necessary to concentrate the clustering using not all the information regarding crime, but only the Robbery and Drugs.

With the help of the Forsquare API, I got all the venues passing as parameter the latitude and longitude.
Using onehot technique, I converted the venues categories into numbers.

The final table containing, Berlin Robbery and Drugs, Density and venues per zip code:

Table 2: Final table with: Venue Robbery and Drugs, Venue Density and all venues

	ZipCode	Venue Robbery	Venue Drugs	Venue Density	ATM	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum
0	101	97.502347	541.600939	9576.000	0.0	0.004695	0.0	0.0	0.0	0.0	0.0	0.0	0.042254	0.00939
1	102	61.764706	603.176471	10.427	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000
2	103	61.562500	355.500000	11.181	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.037500	0.01250
3	104	75.956522	382.630435	2878.000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.021739	0.00000
4	201	125.000000	622.000000	13.910	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00000

Data Analysis

Population Analysis

From the population table, I grouped the neighborhoods and sorted by density (Einwohnerpro Km2). The result is the following table

Table 3: Berlin neighborhoods sorted by density

Bezirk	Nr.	Fläche(km²)	Einwohner[2](31. Dezember 2019)	Einwohnerpro km²
Treptow-Köpenick	13620	16570	273.689	46657.000
Lichtenberg	11065	5202	928.566	44584.104
Pankow	3991	10326	409.335	38831.474
Reinickendorf	13266	8940	266.408	31780.618
Spandau	4545	9190	245.197	30638.000
Steglitz-Zehlendorf	4228	10247	310.071	25173.164
Mitte	621	3948	385.748	21896.134
Tempelhof-Schöneberg	4221	5308	350.984	19469.740
Charlottenburg-Wilmersdorf	2828	6462	343.592	17418.964
Neukölln	4015	4491	329.917	13502.348
Marzahn-Hellersdorf	5015	6171	269.967	12446.137
Friedrichshain-Kreuzberg	403	2018	290.386	28.751

From table 3 it is possible to see the most dense Berlin Neighborhoods, which are: Treptow-Köpenick, Lichtemberg and Pankow.

Berlin Crimes analysis

Relationship between Robbery and Drugs

The same is made for the Crimes table.

Table 4 and 5 : Robbery per District ordered, Drugs per District ordered

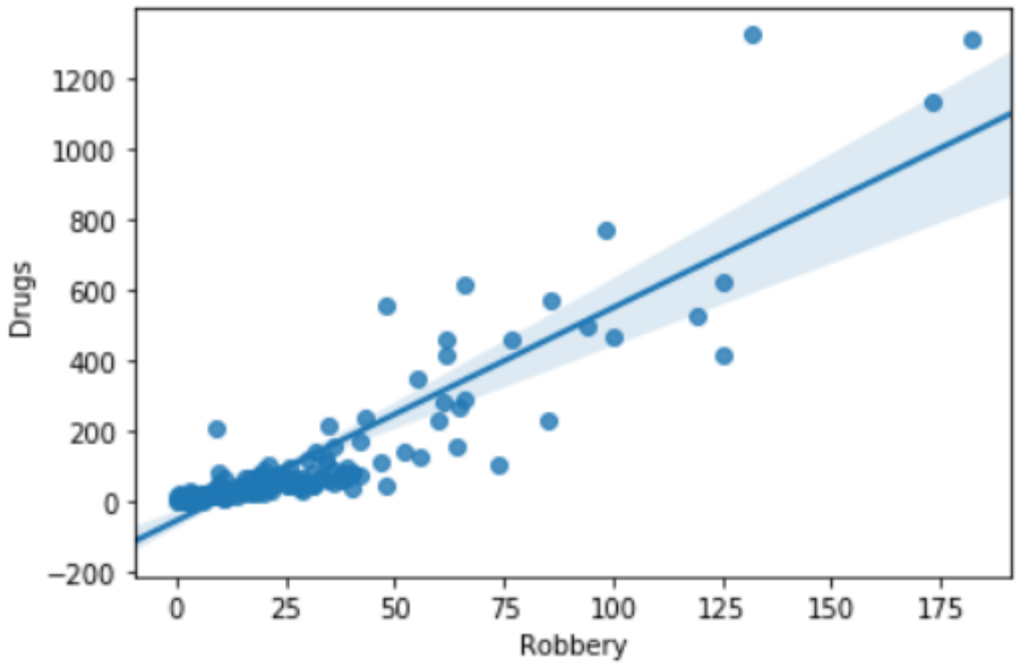
District	Robbery	Drugs
Friedrichshain-Kreuzberg	820	5232
Mitte	707	4233
Neukölln	480	2126
Tempelhof-Schöneberg	352	1209
Charlottenburg-Wilmersdorf	420	1174

District	Robbery	Drugs
Friedrichshain-Kreuzberg	820	5232
Mitte	707	4233
Neukölln	480	2126
Charlottenburg-Wilmersdorf	420	1174
Tempelhof-Schöneberg	352	1209

We can see from the next two tables, that the leading neighborhoods in terms of Robbery are Friedrichheim-Kreuzberg and Mitte and Neukön, for both Robbery and Drugs.

With this information I created a correlation graph showing the correlation between Robbery and Drugs

Figure1: regression relationship between Drugs and Robbery



Methodology

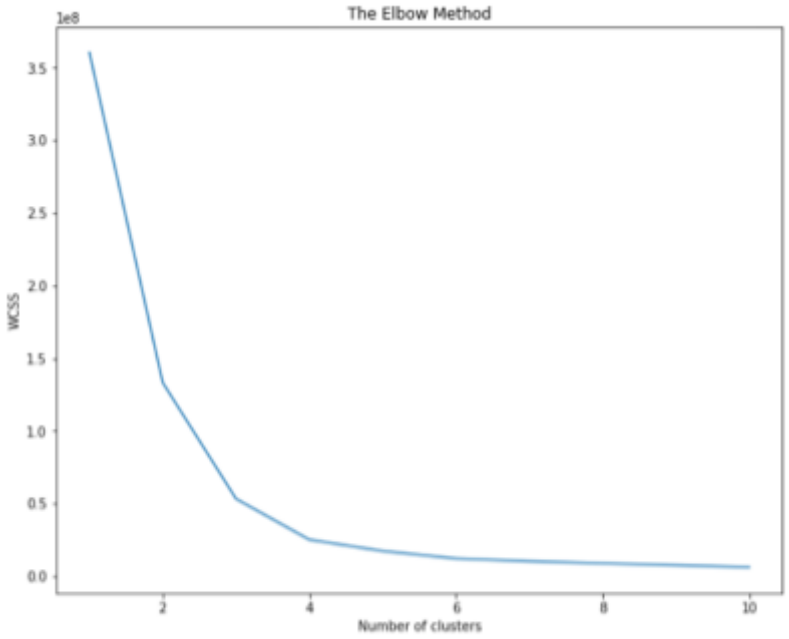
I mentioned before that the methodology to be applied to this problem is clustering.

The justification for choosing this method is based in the fact that there is no data to predict, therefore I have to use a non-supervised learning method.

Of all the non-supervised learning methods, the most suitable to fit this problem is “clustering”, being **KMEANS** the preferred algorithm.

Clustering

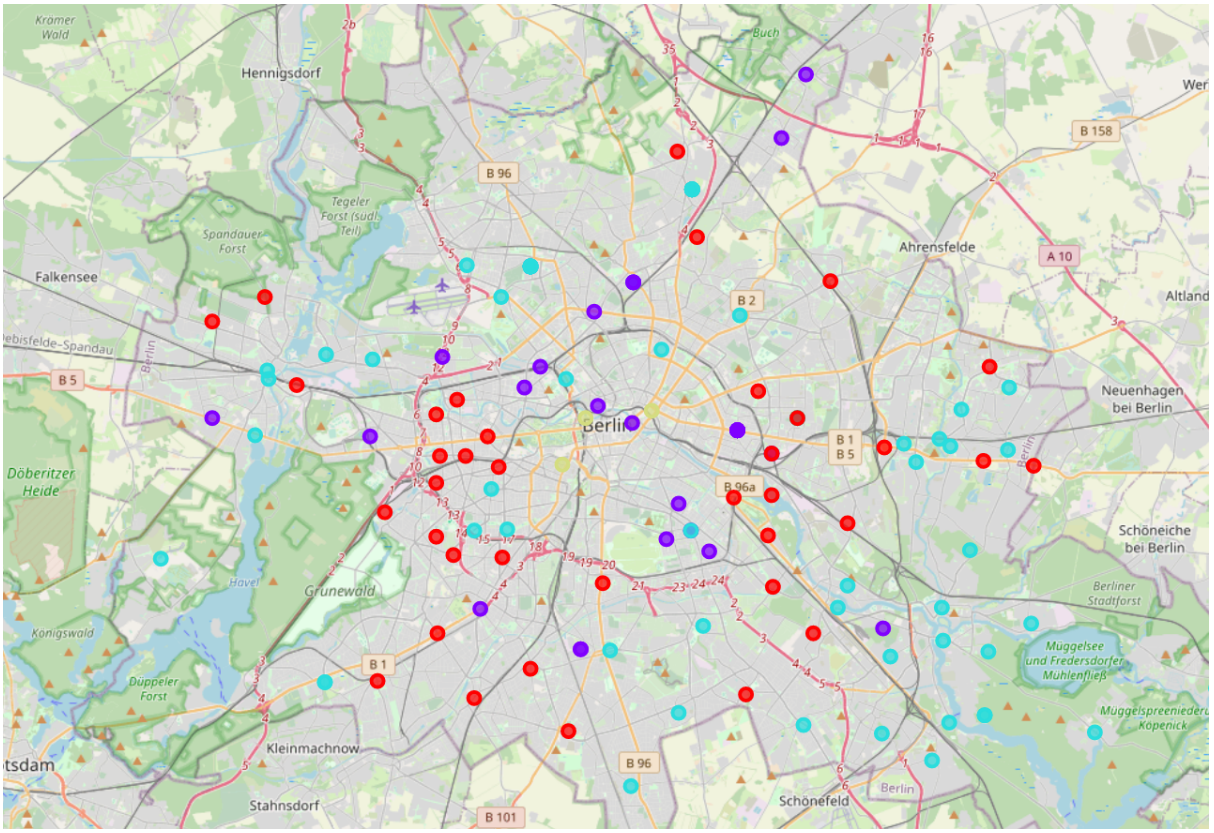
Kmean need to define this K parameter in upfront. Therefore, I found the best K for this problem using the elbow model:



Graph 2: Elbow Method.

I take a K = 4 for this analysis.

After fitting the model and rearranged the table, a map with the labeled neighbors is draw.



Map 2: Berlin with labeled districts.

Results

As expected there was created 4 clusters, lets explore each of them to make some conclusions.

Analyzing cluster labeled 0

Table 7: **Labeled 0**, districts with crimes, density and most commons venues

Location	District	Longitud	Latitud	Robbery	Drugs	Density	1st Most Common Venue	2nd Most Common Venue
Blankenfelde/Niederschönhausen	Pankow	13.4364	52.5976	7	22	6814	Turkish Restaurant	Garden Center
Buchholz	Pankow	13.4281	52.6105	3	9	6814	Turkish Restaurant	Garden Center
Schloß Charlottenburg	Charlottenburg-Wilmersdorf	13.2927	52.5206	26	54	6185	Hotel	Café
Mierendorffplatz	Charlottenburg-Wilmersdorf	13.3044	52.5257	26	48	6185	Hotel	Café
Otto-Suhr-Allee	Charlottenburg-Wilmersdorf	13.3215	52.5132	31	84	6185	Hotel	Café

Table 8: Labeled 0, sorted by Robbery

District	Robbery	Drugs	Density
Charlottenburg-Wilmersdorf	258	650	58091
Lichtenberg	190	371	44601
Marzahn-Hellersdorf	117	253	17079
Tempelhof-Schöneberg	109	395	15906
Spandau	97	359	14781
Steglitz-Zehlendorf	97	202	21940
Treptow-Köpenick	63	210	22300
Neukölln	48	47	6410
Pankow	10	31	13628

Analyzing cluster labeled 1

Table 9: **Labeled 1**, districts with crimes, density and most commons venues

Location	District	Longitud	Latitud	Robbery	Drugs	Density	1st Most Common Venue	2nd Most Common Venue
Moabit West	Mitte	13.3425	52.5301	66	618	10	German Restaurant	Café
Moabit Ost	Mitte	13.351	52.537	48	555	10	German Restaurant	Café
Osloer Straße	Mitte	13.3814	52.5558	61	279	11	German Restaurant	Hotel
Brunnenstraße Nord	Mitte	13.4024	52.5177	62	415	11	German Restaurant	Hotel
Südliche Friedrichstadt	Friedrichshain-Kreuzberg	13.4616	52.5153	125	622	13	Bar	Café

Table 10: Labeled 1, sorted by Robbery

District	Robbery	Drugs	Density
Friedrichshain-Kreuzberg	818	5223	111
Neukölln	318	1820	70
Mitte	237	1867	42
Tempelhof-Schöneberg	168	654	28
Charlottenburg-Wilmersdorf	82	304	40
Steglitz-Zehlendorf	74	129	22
Pankow	65	179	2317
Reinickendorf	55	223	24

Analyzing cluster labeled 2

Table 11: **Labeled 2**, districts with crimes, density and most commons venues

Location	District	Longitud	Latitud	Robbery	Drugs	Density	1st Most Common Venue	2nd Most Common Venue
Parkviertel	Mitte	13.3652	52.533	64	156	2878	Bakery	Supermarket
Wedding Zentrum	Mitte	13.3295	52.5607	86	573	2878	Bakery	Supermarket
Schönholz/Wilhelmsruh /Rosenthal	Pankow	13.4364	52.5976	11	8	1992	Tram Station	Asian Restaurant
Pankow Zentrum	Pankow	13.4364	52.5976	40	41	1992	Tram Station	Asian Restaurant
Pankow Süd	Pankow	13.4364	52.5976	26	46	1992	Tram Station	Asian Restaurant

Table 12: **Labeled 2**, Robbery vs Drugs sorted.

District	Robbery	Drugs	Density
Pankow	208	565	19782
Reinickendorf	178	611	26001
Mitte	150	729	5756
Marzahn-Hellersdorf	120	281	13547
Neukölln	112	252	14128
Spandau	111	242	18743
Treptow-Köpenick	105	421	57612
Charlottenburg-Wilmersdorf	73	215	9306
Tempelhof-Schöneberg	73	155	6779
Steglitz-Zehlendorf	46	79	6440
Lichtenberg	35	62	8472

Analyzing cluster labeled 3

Table 13: **Labeled 3**, districts with crimes, density and most commons venues

Location	District	Longitud	Latitud	Robbery	Drugs	Density	1st Most Common Venue	2nd Most Common Venue
Tiergarten Süd	Mitte	13.3634	52.5038	60	231	9576	Hotel	German Restaurant
Regierungsviertel	Mitte	13.3765	52.5195	42	170	9576	Hotel	German Restaurant
Alexanderplatz	Mitte	13.4136	52.522	173	1133	9576	Hotel	German Restaurant
Brunnenstraße Süd	Mitte	13.4024	52.5177	40	86	9576	Hotel	German Restaurant
Malchow, Wartenberg und Falkenberg	Lichtenberg	13.495	52.5197	1	6	9696	Tram Station	Park

Table 14: **Labeled 3**, Robbery vs Drugs sorted.

District	Robbery	Drugs	Density
Mitte	315	1620	38304
Lichtenberg	32	100	29088

Venues vs labels

Now I will show the tables of the 4 labeled scenarios sorted by the three most common venues in each label.

Table 15: label 0, three most common venues, sorted from most to less

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Turkish Restaurant	Garden Center	Supermarket
Tram Station	Supermarket	Nightclub
Supermarket	Movie Theater	Shoe Store
	Fast Food Restaurant	Drugstore
Pool	Restaurant	Trail
Park	Doner Restaurant	Fried Chicken Joint
	Chinese Restaurant	Bus Stop
Italian Restaurant	Supermarket	German Restaurant
Ice Cream Shop	Italian Restaurant	Bank
Hotel	Furniture / Home Store	Supermarket
	Fast Food Restaurant	Furniture / Home Store
	Café	Italian Restaurant
Café	Korean Restaurant	Ice Cream Shop
Bus Stop	Supermarket	Asian Restaurant
Bakery	Café	Organic Grocery

Table 16: label 1, three most common venues, sorted from most to less

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Tram Station	Park	Bowling Alley
Hotel	German Restaurant	Art Gallery

Table 17: label 2, three most common venues, sorted from most to less

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Tram Station	Park	Bowling Alley
	Asian Restaurant	Turkish Restaurant
Supermarket	Tram Station	Platform
	Plaza	Trail
	Italian Restaurant	History Museum
	Historic Site	Miscellaneous Shop
Steakhouse	Park	Greek Restaurant
Soccer Field	Supermarket	Bus Stop
Pool	Italian Restaurant	Bus Stop
Metro Station	Pool	Bus Stop
Italian Restaurant	Indian Restaurant	Supermarket
	Bus Stop	Liquor Store
Historic Site	History Museum	Athletics & Sports
Furniture / Home Store	Historic Site	Supermarket
Clothing Store	Bus Stop	Bakery
Café	Lake	Dessert Shop
	German Restaurant	Plaza
	Drugstore	Supermarket
		Bakery
	Bar	Bakery
Bar	Café	Italian Restaurant
Bakery	Supermarket	Restaurant
	Miscellaneous Shop	Palace

Table 18: label 3, three most common venues, sorted from most to less

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Tram Station	Park	Bowling Alley
Hotel	German Restaurant	Art Gallery

Analysis of the four cluster

Labeled 0 cluster analysis

Analyzing tables 7,8 and 15, it is possible to see that this cluster offer a variety of different venues and shops. High density and moderate crimes, make this cluster the top to be preferred to install a shop

Labeled 1 cluster analysis

Tables 9,10 and 16 gather the information about cluster 1. This cluster is far away where is concentrated the most crimes. Unfortunately, there is also no relevant business offer.

Labeled 2 cluster analysis

Paying attention to cluster 11,12 and 17, we can get the information about this interesting cluster. Moderate criminality and very high number of opportunities characterize this cluster.

Labeled 3 cluster analysis

This cluster share characteristics with cluster 1. Tables 13,14 and 18 shows that the level of criminality in this cluster is a bit high and no many opportunities are displayed around these locations.

Discussion

At the time to make any kind of recommendation about where a fiscal business can proper in Berlin city, without any doubt, locations placed in cluster 2 and 0 are the most attractive. In particular cluster 2 is the most lively, with a Hughes number of shops, gallery arts, public transport and so on.
In the other hand cluster 3 and 1 are either dangerous and not attractive from the business point of view.

Conclusion

This project can help to understand the venues' opportunity panorama in Berlin city at the time to make any kind of investment.

I paid attention in three pillars, population density, criminality and diversity of venues around some location in Berlin. Future direction should include rent pricing. This issue is high relevant.