Analyzing AirBNB and Foursquare data in Berlin

Introduction

As part of the lodging sector, you want to invest in a new hotel, B&B or just want to rent a flat through Airbnb. Before investing, you need to know what neighborhood would be a nice choice and a good price would be in the neighborhood you choose. It is also a good idea to invest in a neighborhood with good facilities, restaurants, cafes, parks and tourist locations, so people find your place convenient when they're traveling.

The goal of this project is to provide some guidance about where it would be a good idea to invest, what neighborhoods of a city are better than others, and give you an idea of the average price for an Airbnb in the area.

Data

To solve this problem, I am going to use the data provided by the Foursquare API, and data from the Airbnb database for Berlin. This data is part of the project Inside AirBNB, and is publicly available in http://insideairbnb.com/get-the-data.html, along with information from other major cities in the world.

The dataset that includes the details of the available has a total of 106 columns, including information about the neighborhood, burough, coordinates, price per night, total accommodates, description of the place, facilities around, cancellation policies, among many others.

In particular, I have used the data about neighborhood, burough, location, prices and accommodates. Here is a sample of the kind of information we can obtain for one place:

```python
airbnb[['host_id',
        'neighbourhood_cleansed',
        'neighbourhood_group_cleansed',
        'price',
        'accommodates',
        'latitude',
        'longitude']].loc[0]
```

```
host_id                                        3718
neighbourhood_cleansed         Prenzlauer Berg Südwest
neighbourhood_group_cleansed                 Pankow
price                                        $90.00
accommodates                                      4
latitude                                     52.535
longitude                                   13.4176
Name: 0, dtype: object
```

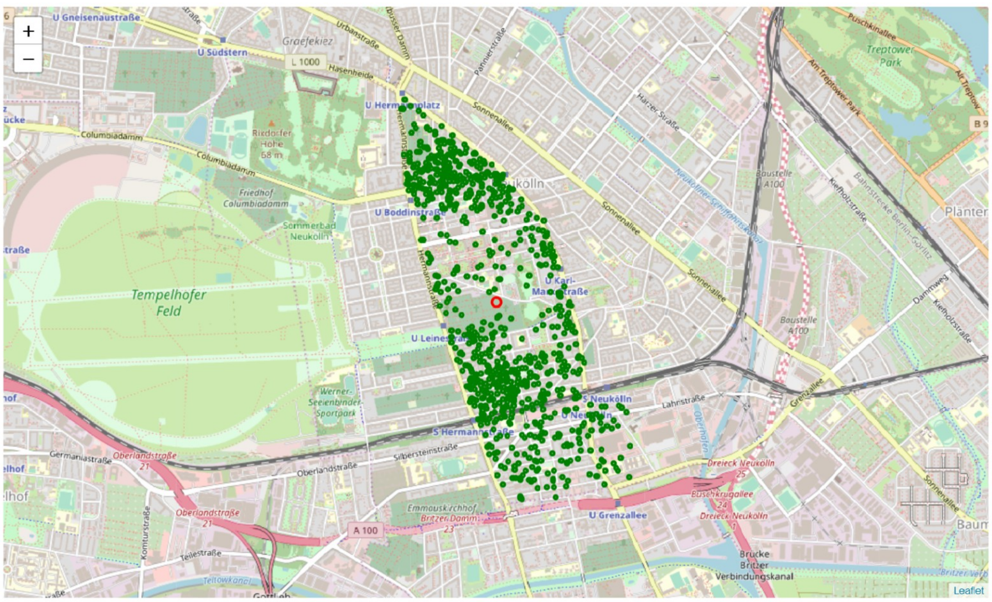
There is a total of **25,197** registered hosts like this one. I have used this information to calculate general data for each one of the neighborhoods in Berlin.

Methodology

Calculating the coordinates for each neighborhood

Because the Airbnb information doesn't include coordinates for each one of the neighborhoods, I have calculated them using the mean of all the latitudes and longitudes in each one.

The following picture shows the calculation made for Neuköllner Mitte, a very popular neighborhood in Berlin. The green dots show each one of the Airbnb hosts, and the red circle shows the calculated center using the coordinates of those hosts.



I have calculated similar coordinates for each one of a total of 137 neighborhoods in Berlin, and I've also calculated the total amount of Airbnb hosts, and the average price per person in every neighborhood. Here is a sample of this data:

	Neighborhood	Borough	Latitude	Longitude	Avg Price	Total Count
0	Adlershof	Treptow - Köpenick	52.436802	13.547116	21.480108	31
1	Albrechtstr.	Steglitz - Zehlendorf	52.455627	13.337145	21.646389	119
2	Alexanderplatz	Mitte	52.522387	13.404324	61.899000	1255
3	Allende-Viertel	Treptow - Köpenick	52.447843	13.598447	22.087302	3
4	Alt Treptow	Treptow - Köpenick	52.490437	13.450552	23.093964	185
...	...	...	...	...	...	...
132	Wilhelmstadt	Spandau	52.525758	13.189837	25.102564	39
133	Zehlendorf Nord	Steglitz - Zehlendorf	52.447411	13.261869	31.175571	73
134	Zehlendorf Südwest	Steglitz - Zehlendorf	52.421631	13.172213	27.789003	59
135	nördliche Luisenstadt	Friedrichshain-Kreuzberg	52.501651	13.427058	26.680173	475
136	südliche Luisenstadt	Friedrichshain-Kreuzberg	52.496579	13.435710	26.276013	686

Exploring one of the neighborhoods

In order to check the information obtained from the Foursquare API, I have first analyzed one of the neighborhoods to check the relative amount of hotels and restaurants that are located in the area. It is important to have in mind that the Foursquare API can only give 100 venues in each request, so the amount of hotels and restaurants is really the percentage of those venues in the area.

This is a sample result of the values obtained for Alexanderplatz:

	name	categories	lat	lng
0	19grams	Coffee Shop	52.522697	13.407440
1	Die Hackeschen Höfe	Monument / Landmark	52.524094	13.402157
2	Barrio Weine Berlin	Wine Shop	52.523531	13.405946
3	Hackescher Markt	Plaza	52.522993	13.402378
4	Waffel oder Becher	Ice Cream Shop	52.521007	13.403815
...	...	...	...	...
95	Hundt Hammer Stein	Bookstore	52.525790	13.407068
96	Ace & Tate	Optical Shop	52.526328	13.407714
97	Atrium Lobby Lounge & Bar	Hotel Bar	52.519597	13.402774
98	Berliner Fernsehturm	Scenic Lookout	52.520936	13.410007
99	Altes Museum	History Museum	52.519537	13.398803

100 rows × 4 columns

Using this information, I have counted how many hotels in restaurants are listed for this neighborhood:

```
## Number of venues identified as "Hotel" by Foursquare.

venues[venues['categories'].str.contains('Hotel')].shape[0]
```

```
8
```

```
## Number of venues identified as "Restaurant" by Foursquare.

venues[venues['categories'].str.contains('Restaurant')].shape[0]
```

```
12
```

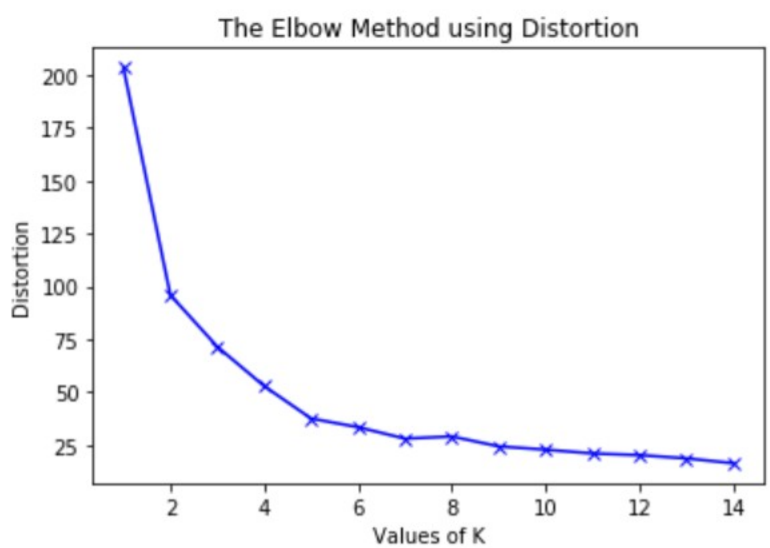
Getting information for all the neighborhoods

I have used a function to iterate this process through all of the neighborhoods, and appended the number of hotels and restaurants to the neighborhoods data.

	Neighborhood	Borough	Latitude	Longitude	Avg Price	Total Count	Hotels	Restaurants
0	Adlershof	Treptow - Köpenick	52.436802	13.547116	21.480108	31	0	2
1	Albrechtstr.	Steglitz - Zehlendorf	52.455627	13.337145	21.646389	119	0	3
2	Alexanderplatz	Mitte	52.522387	13.404324	61.899000	1255	8	12
3	Allende-Viertel	Treptow - Köpenick	52.447843	13.598447	22.087302	3	0	1
4	Alt Treptow	Treptow - Köpenick	52.490437	13.450552	23.093964	185	0	7
...	...	...	...	...	...	...	...	...
132	Wilhelmstadt	Spandau	52.525758	13.189837	25.102564	39	0	2
133	Zehlendorf Nord	Steglitz - Zehlendorf	52.447411	13.261869	31.175571	73	0	0
134	Zehlendorf Südwest	Steglitz - Zehlendorf	52.421631	13.172213	27.789003	59	1	1
135	nördliche Luisenstadt	Friedrichshain-Kreuzberg	52.501651	13.427058	26.680173	475	0	38
136	südliche Luisenstadt	Friedrichshain-Kreuzberg	52.496579	13.435710	26.276013	686	0	24

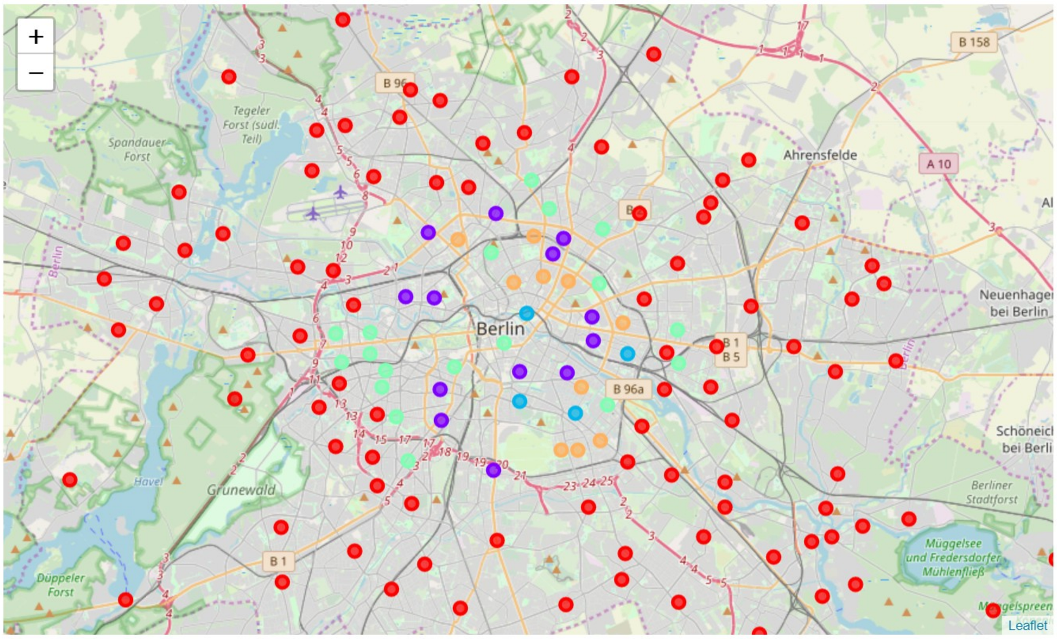
Clustering

I have used the K-means clustering algorithm to group the neighborhoods in clusters. First, I have tested the method with a range of 1 to 15 clusters, and used the elbow method to check the optimum number of clusters for the classifier:



We can see that there is a pronunciated elbow around *k = 5*, so let's use that number of clusters.

Then, using the value *k* = *5*, I have plotted the clusters in a Berlin map:



Results

We can now inspect closely the results for each cluster. We are specially interested in looking at the Average Price, Total Count, Hotels, and Restaurant counts.

Cluster 0:

	Neighborhood	Borough	Avg Price	Total Count	Hotels	Restaurants
0	Adlershof	Treptow - Köpenick	21.480108	31	0	2
1	Albrechtstr.	Steglitz - Zehlendorf	21.646389	119	0	3
3	Allende-Viertel	Treptow - Köpenick	22.087302	3	0	1
5	Alt-Hohenschönhausen Nord	Lichtenberg	21.358844	14	0	1
6	Alt-Hohenschönhausen Süd	Lichtenberg	21.082846	44	1	3
...	...	...	...	...	...	...
130	Westend	Charlottenburg-Wilm.	27.670811	117	1	5
131	Wiesbadener Straße	Charlottenburg-Wilm.	27.909770	58	0	5
132	Wilhelmstadt	Spandau	25.102564	39	0	2
133	Zehlendorf Nord	Steglitz - Zehlendorf	31.175571	73	0	0
134	Zehlendorf Südwest	Steglitz - Zehlendorf	27.789003	59	1	1

92 rows × 6 columns

Cluster 1:

	Neighborhood	Borough	Avg Price	Total Count	Hotels	Restaurants
50	Helmholtzplatz	Pankow	32.877516	488	0	28
53	Karl-Marx-Allee-Nord	Friedrichshain-Kreuzberg	25.281627	334	1	1
54	Karl-Marx-Allee-Süd	Friedrichshain-Kreuzberg	25.859384	383	2	3
74	Moabit Ost	Mitte	23.804617	429	1	10
75	Moabit West	Mitte	22.326451	525	0	24
86	Osloer Straße	Mitte	22.037519	414	2	4
93	Parkviertel	Mitte	18.875517	384	0	8
95	Prenzlauer Berg Nord	Pankow	24.741871	438	0	12
111	Schöneberg-Nord	Tempelhof - Schöneberg	89.094846	577	3	20
112	Schöneberg-Süd	Tempelhof - Schöneberg	24.637210	485	0	34
116	Südliche Friedrichstadt	Friedrichshain-Kreuzberg	29.896479	417	1	5
118	Tempelhof	Tempelhof - Schöneberg	22.241366	327	1	9
135	nördliche Luisenstadt	Friedrichshain-Kreuzberg	26.680173	475	0	38

Cluster 2:

	Neighborhood	Borough	Avg Price	Total Count	Hotels	Restaurants
2	Alexanderplatz	Mitte	61.899000	1255	8	12
33	Frankfurter Allee Süd FK	Friedrichshain-Kreuzberg	25.071870	1466	0	32
102	Reuterstraße	Neukölln	24.118911	1091	0	25
119	Tempelhofer Vorstadt	Friedrichshain-Kreuzberg	27.671989	1368	1	25

Cluster 3:

	Neighborhood	Borough	Avg Price	Total Count	Hotels	Restaurants
4	Alt Treptow	Treptow - Köpenick	23.093964	185	0	7
7	Alt-Lichtenberg	Lichtenberg	22.104933	174	1	1
17	Brunnenstr. Nord	Mitte	27.055848	301	0	0
27	Düsseldorfer Straße	Charlottenburg-Wilm.	39.050163	205	4	27
34	Friedenau	Tempelhof - Schöneberg	24.685377	212	1	17
52	Kantstraße	Charlottenburg-Wilm.	28.362777	156	1	34
58	Kurfürstendamm	Charlottenburg-Wilm.	44.305928	160	8	47
77	Neu Lichtenberg	Lichtenberg	21.327316	266	2	1
80	Neue Kantstraße	Charlottenburg-Wilm.	44.545485	169	0	17
90	Otto-Suhr-Allee	Charlottenburg-Wilm.	26.488284	185	1	10
91	Pankow Süd	Pankow	20.841618	188	0	3
92	Pankow Zentrum	Pankow	25.717430	142	0	6
97	Prenzlauer Berg Ost	Pankow	26.145327	267	0	0
101	Regierungsviertel	Mitte	37.390766	244	12	21
107	Schloß Charlottenburg	Charlottenburg-Wilm.	26.376301	173	0	13
120	Tiergarten Süd	Mitte	30.205483	231	9	16
121	Volkspark Wilmersdorf	Charlottenburg-Wilm.	189.903238	218	2	10
123	Weißensee	Pankow	39.349147	230	0	6

Cluster 4:

	Neighborhood	Borough	Avg Price	Total Count	Hotels	Restaurants
18	Brunnenstr. Süd	Mitte	33.863272	861	1	20
31	Frankfurter Allee Nord	Friedrichshain-Kreuzberg	29.774940	757	0	12
81	Neuköllner Mitte/Zentrum	Neukölln	21.770421	842	0	14
96	Prenzlauer Berg Nordwest	Pankow	25.984741	702	0	14
98	Prenzlauer Berg Süd	Pankow	30.758178	634	0	22
99	Prenzlauer Berg Südwest	Pankow	31.523641	679	1	29
103	Rixdorf	Neukölln	21.350045	920	0	6
106	Schillerpromenade	Neukölln	23.455187	705	1	21
122	Wedding Zentrum	Mitte	24.447208	603	0	19
136	südliche Luisenstadt	Friedrichshain-Kreuzberg	26.276013	686	0	24

Discussion

We can identify certain tendencies and patterins in the clusters, for example:

1. In **Cluster 0**, the amount of venues, restaurants and hotels is quite low, and so is the number of Airbnb hosts. The prices are also slightly lower compared to other clusters, so it might be a bad idea to invest in these neighborhoods. These are the neighborhoodss in the outer city.

2. All the neighborhoods in **Cluster 1** are well located in the city center, or close to the more crowded neighborhoods. Also, they have a nice amount of venues and facilities, and the Airbnb offer is not as big as in clusters 2 and 4. This is definitely a market worth looking at.

3. **Cluster 2** contains the most crowded neighborhoods, including Alexanderplatz and Frankfurter Allee Süd. However, the prices for Alexanderplatz are considerably higher compared to other neighborhoods in the same cluster and with more or less the same offers. It might be interesting to invest in one of these neighborhoods, depending on the housing costs.

4. **Cluster 3** is more disperse across the city, and seem to include more residential areas. These neighborhoods are probably not so interesting for an investment.

5. **Cluster 4** is slightly more crowded than Cluster 2, and doesn't seem to have as many restaurants. However, some of the neighborhoods are very well located in touristic places, like the ones in Friedrichsain-Kreuzberg or Neukölln.

Conclusion

Although we could get some useful insight from the Berlin neighborhoods, this is only one of the analysis that we can do with the data sources we got. We could include more information like the ratings of the Airbnb hosts, other kinds of venues from Foursquare, geographical information about places like museums, airports and landmarks and so on.