Airbnb Prices & Venues Data Analysis of Munich

This article was written as part of final capstone project for IBM Data Science Professional Certification in Coursera. In this article I will share the difficulties I faced and also some concepts that I implemented.

This article will contain the following steps that are necessary for any Data Science project:

1. Problem statement
2. Data Collection
3. Data Preprocessing
4. Machine Learning
5. Visualization
6. Result

Problem Statement

As a turist, people always want to stay in a nice neighborhood with reasonable prices. However, it is hard to assess if you are not a local. In this project, we will look to these neighborhoods with k mean machine learning algorithm, clustering them in to categories. At the end, we will decide which neighborhoods are convenient to stay.

On the other hand, we can look from the Airbnb owner's perspective to show them if their appartments are in the better neighborhood comparing to other ones.

Data Collection

To consider the above problem the data is collected as following:

For Airbnb prices, I searched and found a great website where there is a list included with approxiametly 10000 entry of Airbnb adverts. Each advert has its coordinate, neighborhood, nightly price and type of the apartment. This quality data base provides a lot of quality information so that we reduce steps in preparing database such as finding the coordinates of neighborhoods and matching them with average prices

In addition to Airbnb prices, I used Foursquare API to get the most common venues of given Borough of Munich. Finally, JSON data was available as well with the Airbnb prices so that they were perfect combination for our project

Data Preprocessing

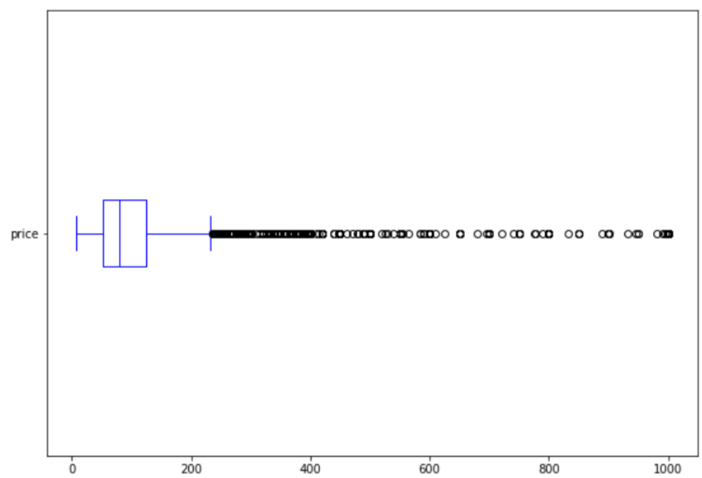
First of all the data scraped from the website has to be clean.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price
0	36720	Beautiful 2 rooms flat, Glockenbach	158413	Gabriela	NaN	Ludwigsvorstadt-Isarvorstadt	48.13057	11.56929	Entire home/apt	95
1	49309	Sublet - Apartment with Balcony in Downtown (1)	224802	Damien	NaN	Ludwigsvorstadt-Isarvorstadt	48.12456	11.55567	Private room	40
2	97945	Deluxw-Apartm. with roof terrace	517685	Angelika	NaN	Hadern	48.11476	11.48782	Entire home/apt	80
3	114695	Apartment Munich/East with sundeck	581737	Stephan	NaN	Berg am Laim	48.11923	11.63726	Entire home/apt	95
4	127383	City apartment next to Pinakothek	630556	Sonja	NaN	Maxvorstadt	48.15198	11.56486	Entire home/apt	120

First of all, I dropped columns like name, host name and id, group etc. Then, we have a cleaner data to work with.

	id	neighbourhood	latitude	longitude	room_type	price
0	36720	Ludwigsvorstadt-Isarvorstadt	48.13057	11.56929	Entire home/apt	95
1	49309	Ludwigsvorstadt-Isarvorstadt	48.12456	11.55567	Private room	40
2	97945	Hadern	48.11476	11.48782	Entire home/apt	80
3	114695	Berg am Laim	48.11923	11.63726	Entire home/apt	95
4	127383	Maxvorstadt	48.15198	11.56486	Entire home/apt	120

However, we need to go few more steps more to make this long data list longer. Groupby is a perfect answer to get averages for each neighborhood, removing some outliers' prices higher than 2000 Euro. Here is the boxplot of the prices.

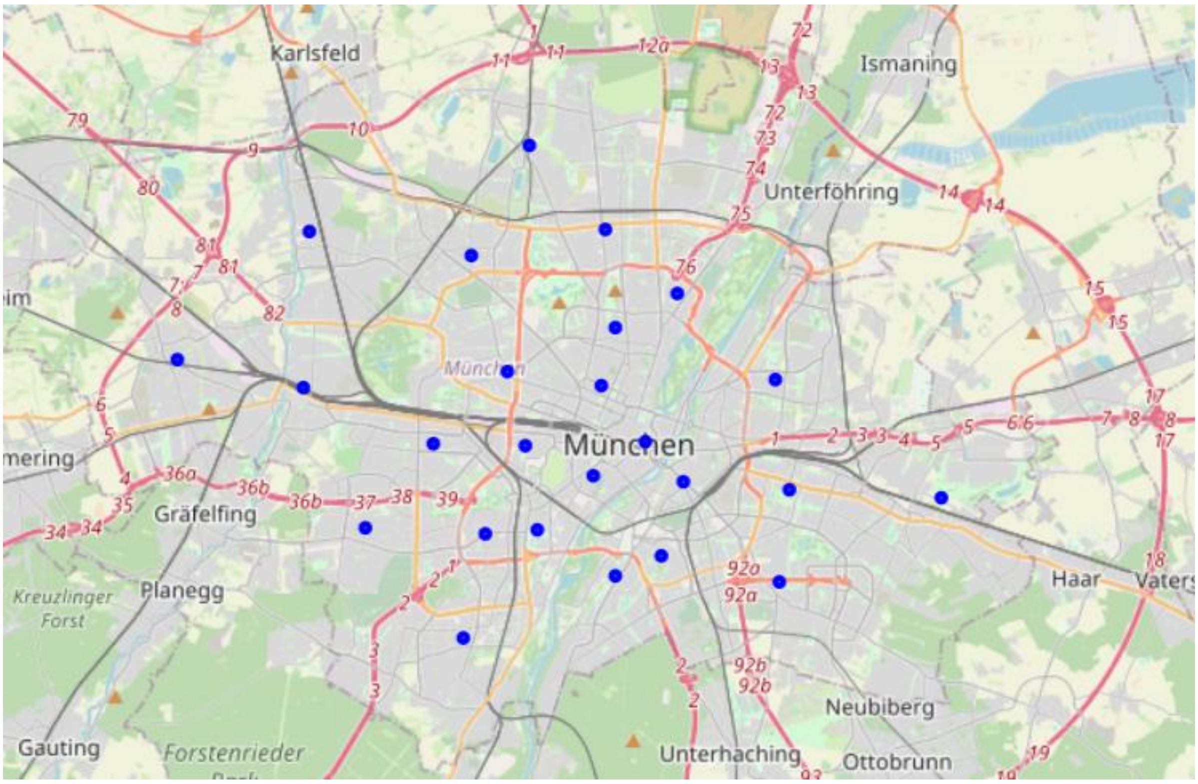


	Area	Avg yield	Avg price	£/sqft	5yr +/-	Explore data
0	BR1	3.5%	£448,605	£461	+21%	Explore data
1	BR2	3.5%	£475,289	£459	+18%	Explore data
2	BR3	3.5%	£432,717	£489	+18%	Explore data
3	BR4	-	£570,326	£461	+18%	Explore data
4	BR5	3.3%	£445,724	£412	+21%	Explore data
5	BR6	3.3%	£505,405	£450	+19%	Explore data
6	BR7	3.0%	£582,654	£483	+21%	Explore data
7	BR8	3.7%	£384,646	£345	+24%	Explore data
8	CR0	4.0%	£363,083	£432	+21%	Explore data
9	CR2	3.5%	£400,905	£429	+21%	Explore data
10	CR4	4.1%	£376,319	£433	+22%	Explore data

Next step is to use **Groupby:**

	id	latitude	longitude	price
neighbourhood				
Allach-Untermenzing	2.485720e+07	48.185855	11.465515	104.214286
Altstadt-Lehel	2.456336e+07	48.137648	11.581302	158.171875
Au-Haidhausen	2.186245e+07	48.128345	11.593697	114.542777
Aubing-Lochhausen-Langwied	2.542403e+07	48.156345	11.420487	96.602273
Berg am Laim	2.196426e+07	48.126613	11.630081	96.606635
Bogenhausen	2.321498e+07	48.151770	11.625256	93.848723
Feldmoching-Hasenbergl	2.405110e+07	48.205644	11.540910	97.292035
Hadern	2.547486e+07	48.117823	11.484957	83.829630
Laim	2.238008e+07	48.137006	11.508123	91.911458
Ludwigsvorstadt-Isarvorstadt	2.271899e+07	48.129925	11.562916	141.768786
Maxvorstadt	2.269014e+07	48.150448	11.565523	116.842007
Milbertshofen-Am Hart	2.415293e+07	48.186282	11.566901	87.632035
Moosach	2.847124e+07	48.180258	11.521115	88.582329
Neuhausen-Nymphenburg	2.247708e+07	48.153873	11.533594	104.026247
Obergiesing	2.296481e+07	48.111455	11.586711	103.824934
Pasing-Obermenzing	2.427333e+07	48.150086	11.463243	100.737778
Ramersdorf-Perlach	2.486991e+07	48.105595	11.626658	82.430622
Schwabing-Freimann	2.346025e+07	48.171757	11.592170	102.291874
Schwabing-West	2.332794e+07	48.163969	11.570419	108.276570
Schwanthalerhöhe	2.221381e+07	48.136602	11.540022	131.410377
Sendling	2.186637e+07	48.117640	11.544175	114.029478
Sendling-Westpark	2.241257e+07	48.116751	11.525741	107.761780
Thalkirchen-Obersendling-Forstenried-Fürstenried-Solln	2.307993e+07	48.092621	11.518454	97.771712
Tudering-Riem	2.561613e+07	48.124694	11.682527	124.843333
Untergiesing-Harlaching	2.119057e+07	48.107136	11.570519	98.902941

Here we see the neighborhoods with location data.



Foursquare Venues

After defining our key data, and we can get the all venues of Munich easily and here we see the list of venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allach-Untermenzing	48.185855	11.465515	Trattoria Olive	48.189905	11.466970	Trattoria/Osteria
1	Allach-Untermenzing	48.185855	11.465515	Würmtalhof	48.188834	11.460680	German Restaurant
2	Allach-Untermenzing	48.185855	11.465515	Sportforum Allach	48.186011	11.468422	Gym
3	Allach-Untermenzing	48.185855	11.465515	Sport Bittl Lagerverkauf	48.186025	11.468463	Sporting Goods Shop
4	Allach-Untermenzing	48.185855	11.465515	Netto Marken-Discount	48.184247	11.461317	Supermarket

However, this data is only the beginning. Next, using categories and machine learning algorithm, we will discover similar and distinct parts of the city.

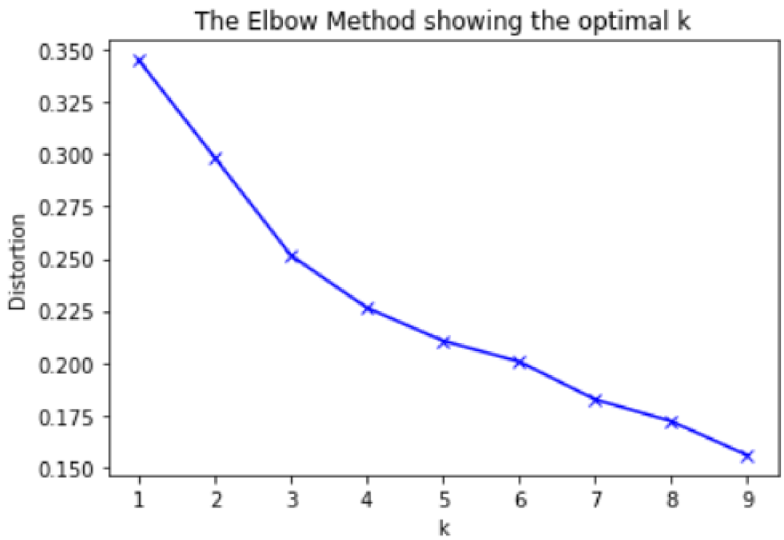
To understand the neighborhoods, we need look to the frequency of categories for each locations. After that, more complex job is to find out the the number of clusters, meaning similar locations will be in same the category.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allach-Untermenzing	Sporting Goods Shop	Light Rail Station	Supermarket	German Restaurant	Bus Stop
1	Altstadt-Lehel	Café	Bavarian Restaurant	Hotel	Coffee Shop	Plaza
2	Au-Haidhausen	Italian Restaurant	Bar	French Restaurant	Cocktail Bar	Bakery
3	Aubing-Lochhausen-Langwied	Bus Stop	Indian Restaurant	Soccer Field	Bakery	German Restaurant
4	Berg am Laim	Hotel	Supermarket	Asian Restaurant	Dog Run	Bus Stop

For example, in Altstadt-Lehel area, there are mostly cafe, restaurants and hotel because it is in the city center.

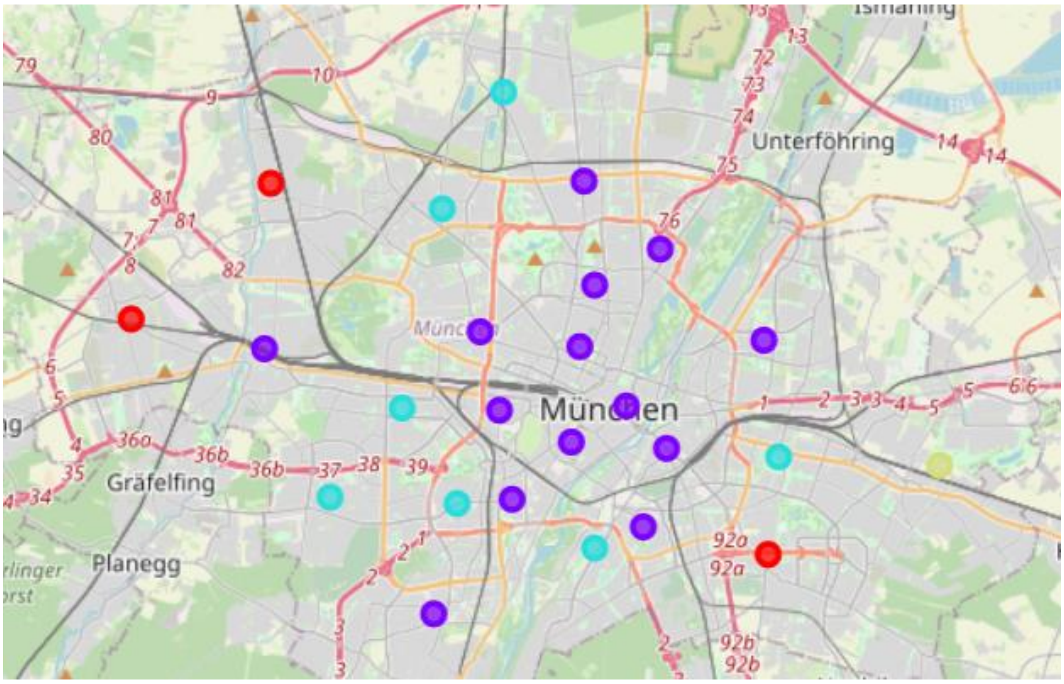
K-Mean Machine Learning Algorithm

Finding out how many categories will suit the city best is another challenge. For that, we need to try different K values with a loop and decide it properly.



In the graph, we see that, after k=4, we have little incremental gain in terms of distortion of data. Thus, we select 4, which means, we categorize neighborhoods in 4 different categories and explore their similarities.

As you can guess, looking the list of venues are not enough, visualizing in the map is a key point to evaluate the similarities. Here is the map of Munich in four categories.



Similarities in Clusters

1. Less areas of attraction – Red

	latitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	48.185855	Sporting Goods Shop	Light Rail Station	Supermarket	German Restaurant	Bus Stop
3	48.156345	Bus Stop	Indian Restaurant	Soccer Field	Bakery	German Restaurant
16	48.105595	Bus Stop	Hotel	Supermarket	Market	Garden Center

2. City center area with cafes, bars and restaurants - Purple

	latitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	48.137648	Café	Bavarian Restaurant	Hotel	Coffee Shop	Plaza
2	48.128345	Italian Restaurant	Bar	French Restaurant	Cocktail Bar	Bakery
5	48.151770	Gym / Fitness Center	Café	Cocktail Bar	Coffee Shop	Chinese Restaurant
9	48.129925	Café	Vietnamese Restaurant	Italian Restaurant	Burger Joint	Bar
10	48.150448	Café	Art Museum	Steakhouse	History Museum	Indian Restaurant
11	48.186282	Bus Stop	Bank	Doner Restaurant	Greek Restaurant	Café
13	48.153873	Italian Restaurant	Bakery	Pizza Place	Sushi Restaurant	Ice Cream Shop
14	48.111455	Hotel	Italian Restaurant	Park	Café	Bus Stop
15	48.150086	Italian Restaurant	Café	Drugstore	Coffee Shop	Supermarket

3. Local areas with mostly supermarkets - Blue

	latitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	48.126613	Hotel	Supermarket	Asian Restaurant	Dog Run	Bus Stop
6	48.205644	Supermarket	Club House	Liquor Store	Italian Restaurant	Drugstore
7	48.117823	Supermarket	German Restaurant	Asian Restaurant	Sushi Restaurant	Soccer Field
8	48.137006	Supermarket	Bank	Bakery	Bookstore	German Restaurant
12	48.180258	Bakery	Supermarket	Plaza	Gym	Light Rail Station
21	48.116751	Drugstore	Supermarket	Bakery	Post Office	Hotel
24	48.107136	Bakery	Café	Hotel	Pharmacy	Indian Restaurant

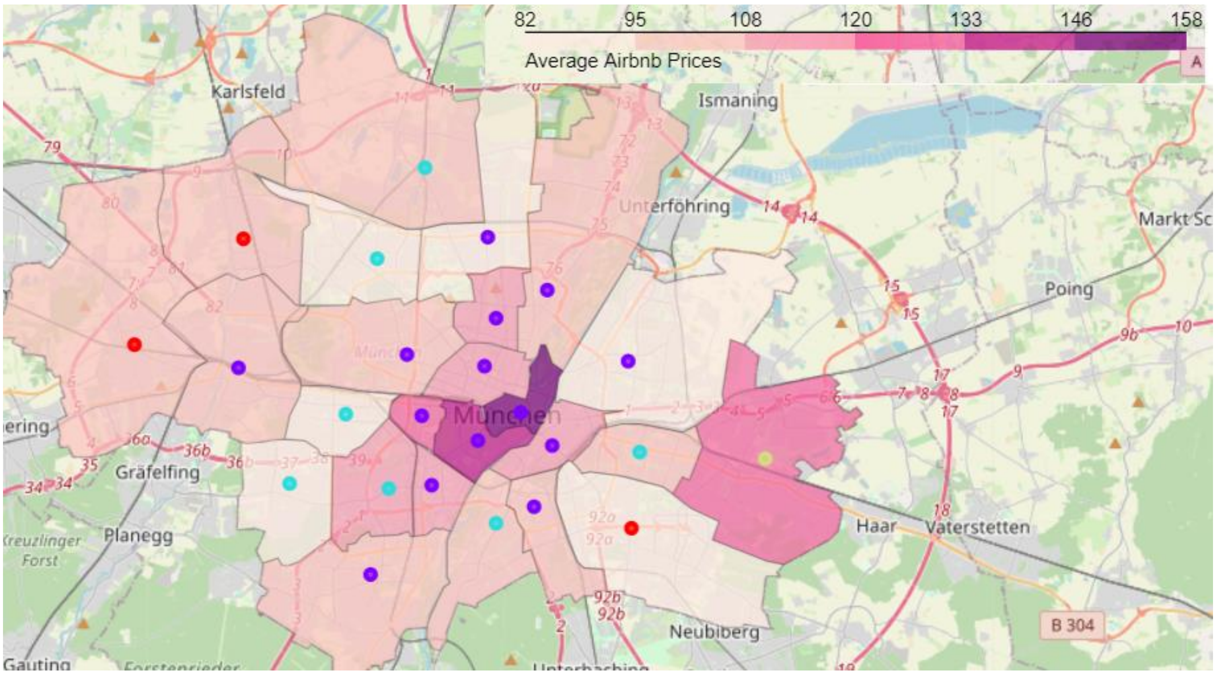
4. Special area of city with big fair halls - Yellow

	latitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
23	48.124694	Auto Workshop	Yoga Studio	Field	Fast Food Restaurant	Farmers Market

Price vs Venues

Visualization

Now, it is come to see if neighborhoods and Airbnb prices are correlated.



Result

As a result, it is obvious to observe that city center (purple) points with bars, cafes and restaurants have higher prices compared to other locations because they offer a lot attraction in the area. One distinct point is the yellow point. When we get more information about Munich, we understand that there is big fair halls there, which attract a lot of people from the world to stay there during the fairs, which effects the prices of Airbnb naturally. In addition, as a tourist in Munich, for example, if you are coming to Oktoberfest, it looks like it is better to stay in blue point areas like Laim, Giesing and Hadern

References

http://data.insideairbnb.com/

https://developer.foursquare.com/

Görkem Kosar, Munich, Germany

https://www.linkedin.com/in/gorkemkosar/