Capstone Project

Buying a home in Germany - Categorical geospatial study

Applied Data Science Capstone by IBM/Coursera

1. Introduction: Business Problem

In Germany only about 45 percent of households own their main residence. This is the second lowest number among all OECD countries, undercut only by Switzerland. This is driven by housing policies that produce incentives to rent.

Nonetheless buying a home can be a good investment, especially if you are planning to take advantage of good market conditions. One of the most important and impactful decisions to make, when choosing to buy a home, is the location. The key factors of a great location are accessibility, appearance, and amenities available in a neighborhood.

These factors greatly influence the buying price, location desirability and investment value appreciation over time.

Considering the amenities available in a neighborhood, these translate directly to a better living and quality of life, for example: consider not having to travel far to go to a doctor's appointment, doing your shopping or dining or getting food in a good restaurant near your house, these are some of the hallmarks of a great neighborhood.

Here lies **our problem**, regarding the **availability of amenities in an area**, where should you choose to live? Which are the best areas with the most amenities available to purchase a house.

This project is about studying amenities availability when location choosing to buy a house around the **Bad Homburg vor der Höhe** city. Our final objective is clustering and comparing areas to try to find and group the best places to live. We will use exploratory data analysis, data visualization and machine learning algorithms to study this problem.

Our target audience is mainly composed of potential investors interested in purchasing a home in the target area. But it can also be a powerful tool for property developers prospecting where to build a new house or amenity or policymakers to decide where should the government invest to influence regional economic development.

2. Data

Data wise used the following datasets:

1. Locations Coordinates (Latitude and Longitude)

One of the most important steps in our project was choosing the area on which we would do our study, finally we decided to **study all the cities within a ~20 kilometer radius from Bad Homburg vor der Höhe and all of Frankfurt am Main**.

This posed a problem, since our study is based on latitude and longitude coordinates, the larger area cities would be underrepresented. In order to try to solve the aforementioned problem we decided on using postal codes for the larger cities such as Bad Homburg vor der Höhe, Frankfurt am Main and Offenbach am Main, this was further supported by seeing that the smaller cities would usually only have one postal code.

2 - Amenities available in the defined area

After defining the coordinates of the chosen locations and a radius to search, we got all the relevant information about the available amenities around the coordinates. The information we collected and used is: amenity name, amenity latitude, amenity longitude, amenity category.

Data collection

We used the https://www.suche-postleitzahl.org/ website to search all cities within the defined radius afterwards we built a excel file with info **(postal code, name, area, region, residents)**, to support our analysis.

To get the **coordinates (latitude and longitude)** we used the Nominatim API to access the OpenStreetMap database https://www.openstreetmap.org/.

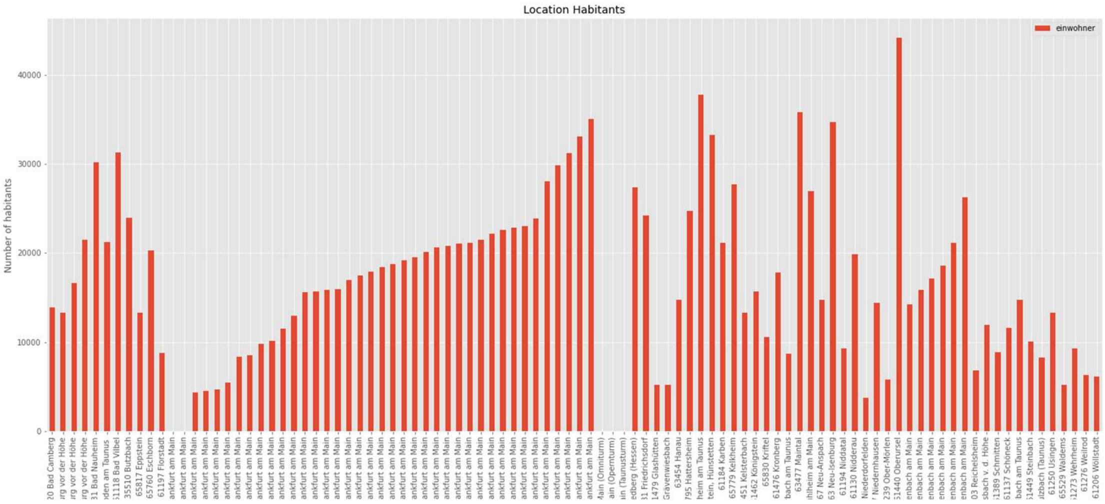
To get the amenities data **(amenity name, amenity latitude, amenity longitude, amenity category)** we used the free tier of Foursquare Developer API https://developer.foursquare.com/.

3. Methodology

Concerning methodology to find the areas with larger concentration of amenities around a ~20Kilometer radius of Bad Homburg vor der Hohë (Postal Code: 61348) we first created our dataset through collecting the available data on location in the studied area, online.

We explored our dataset by checking data types, missing values and describing numerical data the only relevant thing was changing the postal code from integer to string.

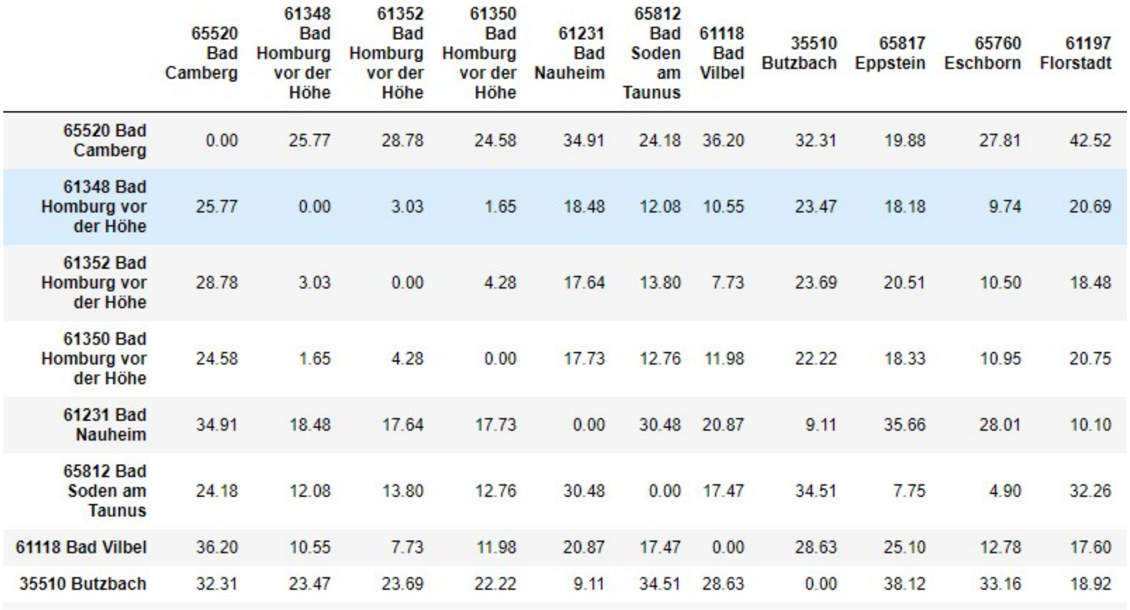
Through studying the residents numerical variable, we decided to drop all row concerning locations where no one lived.



We were left with 45 city centers, 37 Frankfurt am Main postal codes areas, 6 Offenbach am Main postal codes areas and 3 Bad Homburg vor der Höhe postal codes areas to a total of 91 areas to search for amenities.

Having defined which locations to study we used the Nominatim Api to get the coordinates (longitude and latitude).

At this point we took some decisions on locations coordinates and decided it would be interesting to check a distance matrix to see if our main restriction of all locations in a radius of ~20 km from Bad Homburg vor der Höhe would still be true.

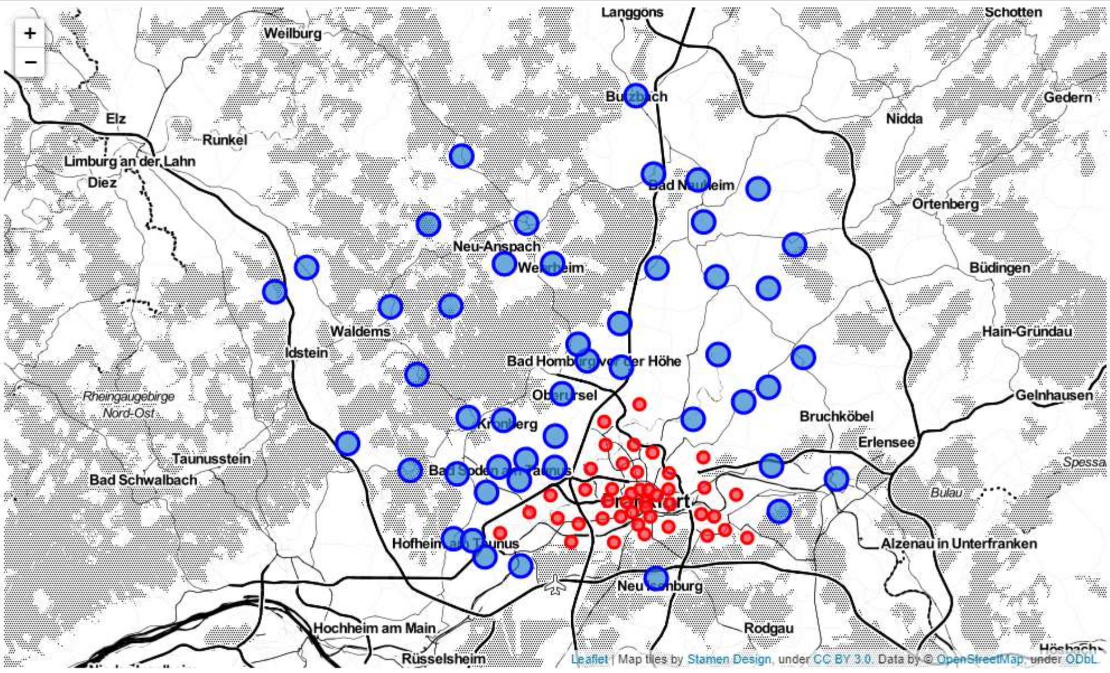


	65520 Bad Camberg	61348 Bad Homburg vor der Höhe	61352 Bad Homburg vor der Höhe	61350 Bad Homburg vor der Höhe	61231 Bad Nauheim	65812 Bad Soden am Taunus	61118 Bad Vilbel	35510 Butzbach	65817 Eppstein	65760 Eschborn	61197 Florstadt
65520 Bad Camberg	0.00	25.77	28.78	24.58	34.91	24.18	36.20	32.31	19.88	27.81	42.52
61348 Bad Homburg vor der Höhe	25.77	0.00	3.03	1.65	18.48	12.08	10.55	23.47	18.18	9.74	20.69
61352 Bad Homburg vor der Höhe	28.78	3.03	0.00	4.28	17.64	13.80	7.73	23.69	20.51	10.50	18.48
61350 Bad Homburg vor der Höhe	24.58	1.65	4.28	0.00	17.73	12.76	11.98	22.22	18.33	10.95	20.75
61231 Bad Nauheim	34.91	18.48	17.64	17.73	0.00	30.48	20.87	9.11	35.66	28.01	10.10
65812 Bad Soden am Taunus	24.18	12.08	13.80	12.76	30.48	0.00	17.47	34.51	7.75	4.90	32.26
61118 Bad Vilbel	36.20	10.55	7.73	11.98	20.87	17.47	0.00	28.63	25.10	12.78	17.60
35510 Butzbach	32.31	23.47	23.69	22.22	9.11	34.51	28.63	0.00	38.12	33.16	18.92

The distance matrix was calculated with the haversine formula, to show us the shortest distance in kilometers between two points on a sphere (Earth) using their latitudes and longitudes measured along the surface. It is important for use in navigation.

What the distance matrix told us was that there were distances greater than 20km from the main location, this merits an explanation. When we were considering the coordinates to study, we first used only postal codes, but we noticed that for some areas particularly the ones where we got one city/one postal code, the postal code coordinates would be far from the city center, and so we would get fewer amenities results. So in order to have a better sense of amenities available we decided that for areas corresponding to one city/one postal code we would get the city center coordinates and this explains why the "61348 Bad Homburg vor der Höhe" distances column are not all under 20 kilometers. The difference in distance may be further explained by the usage of the haversine formula itself since usually the distance is calculated in a straight line.

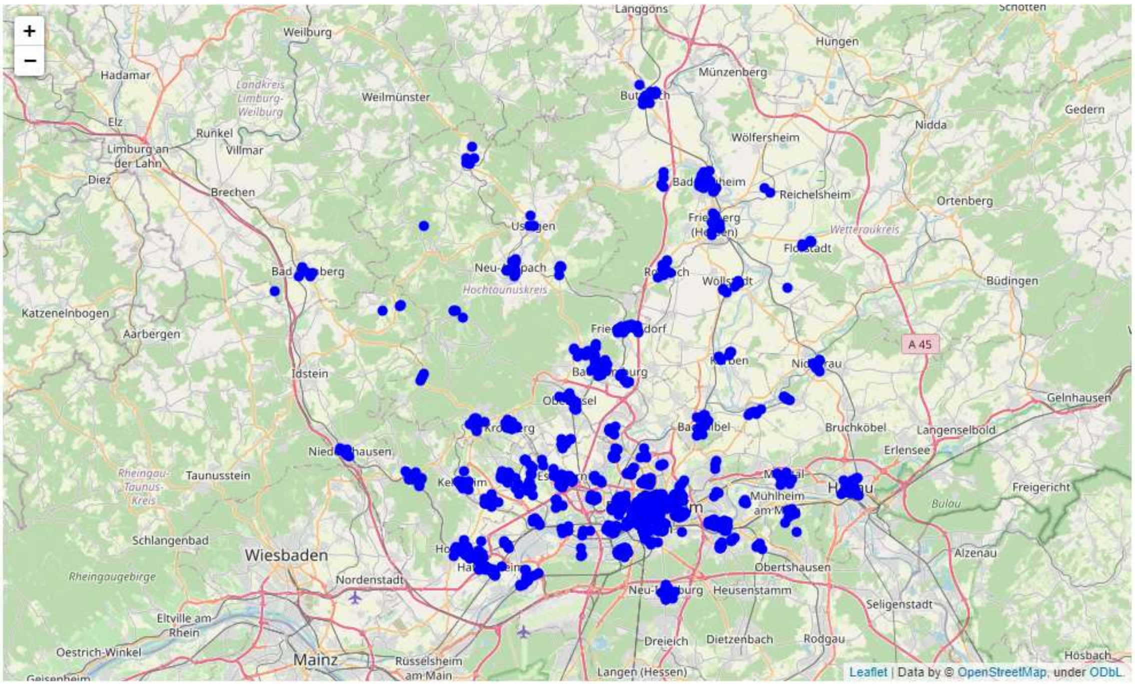
Our final map of locations:



In the map above we can see the areas involved in this study. There you can see blue and red circles; these circles represent the individual areas where we searched for amenities. **A blue circle area is 1000 meters search radius and the red circle area is 500 meters search radius**.

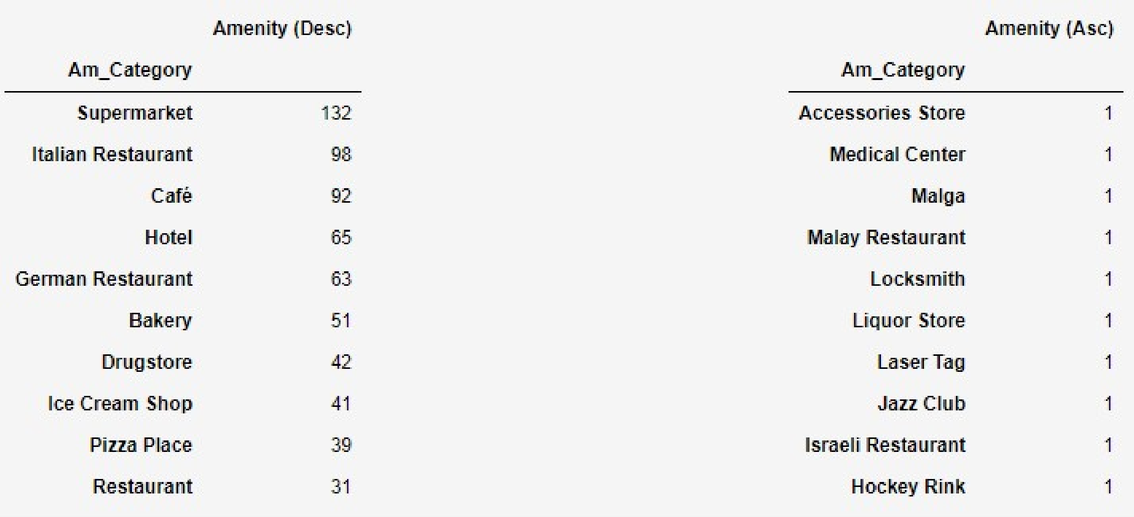
Why did we choose to make this distinction? Because comparing a small city to a big city it's not feasible, the difference in amenities density would be too big. So, we decided that for this study, for clarity and to try to diminish overlapping between areas, particularly in Frankfurt am Main and in Offenbach am Main, we would have to different radius, hence 500 and 1000 meters. In the end we tolerated some overlapping

The next step we collected the amenities data from the foursquare api. We got a dataset of 1673 rows that we put in a map:



With this information our main dataset was complete, and we needed to explore the dataset as a total search locations and search results.

We tried to get a sense of how many categories we were dealing with and we found out that for 1673 amenities we had 234 categories:



	Amenity (Desc)
Am_Category	
Supermarket	132
Italian Restaurant	98
Café	92
Hotel	65
German Restaurant	63
Bakery	51
Drugstore	42
Ice Cream Shop	41
Pizza Place	39
Restaurant	31

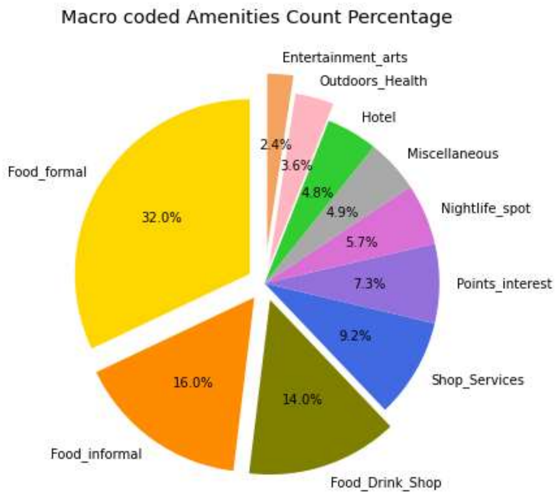
	Amenity (Asc)
Am_Category	
Accessories Store	1
Medical Center	1
Malga	1
Malay Restaurant	1
Locksmith	1
Liquor Store	1
Laser Tag	1
Jazz Club	1
Israeli Restaurant	1
Hockey Rink	1

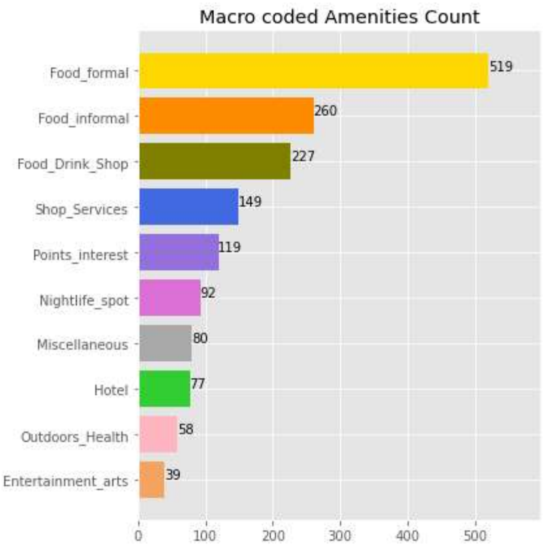
To further analyze this data we decided to recode the categories into a sort of macro categories, converting 234 into 10 manageable categories.

The new categories created:

1. Food_formal - For this category we will consider pretty much all the categories that have the word restaurant in the name
2. Food_informal - This category will have all the places where you can get quick food
3. Food_Drink_Shop - Here we will have groceries and drink shops
4. Hotel - Accommodation spots
5. Nightlife_spot - Drinking spots, bars, etc.
6. Entertainment_arts - Clubs, Museums, Theater, etc.
7. Outdoors_Health - Sports and Health category
8. Shop_Services - Shops and Services
9. Points_interest - Landmarks
10. Miscellaneous - All the other stuff that can't quite fit in any other category

And we got the following information:

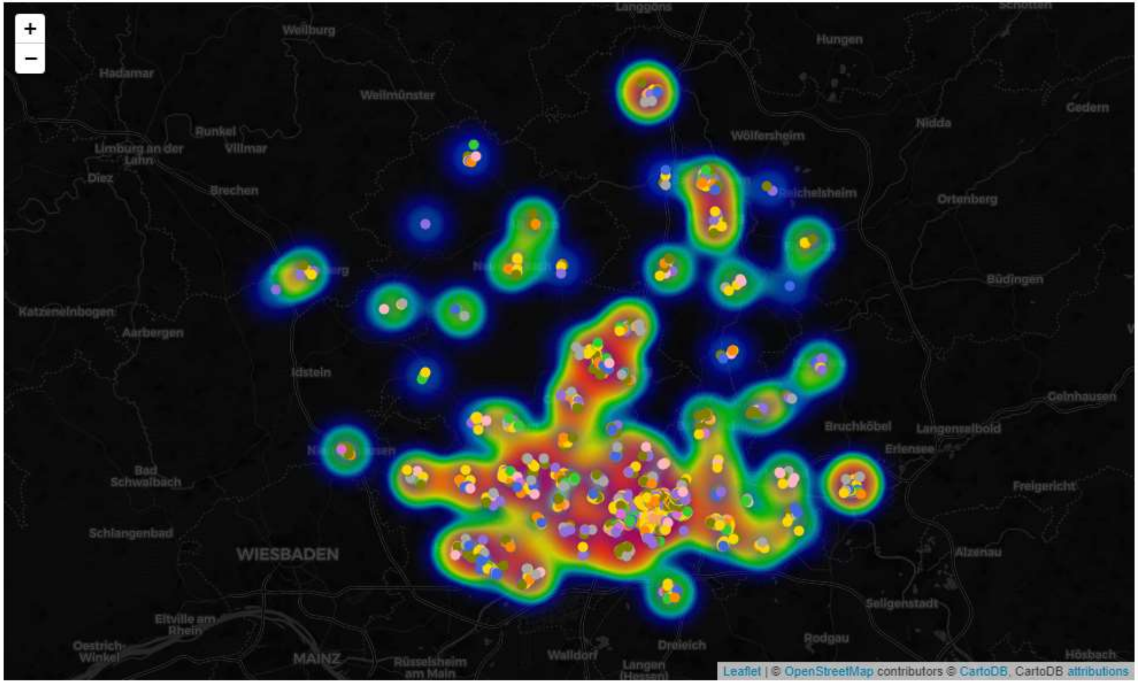




We can see in the graphs:

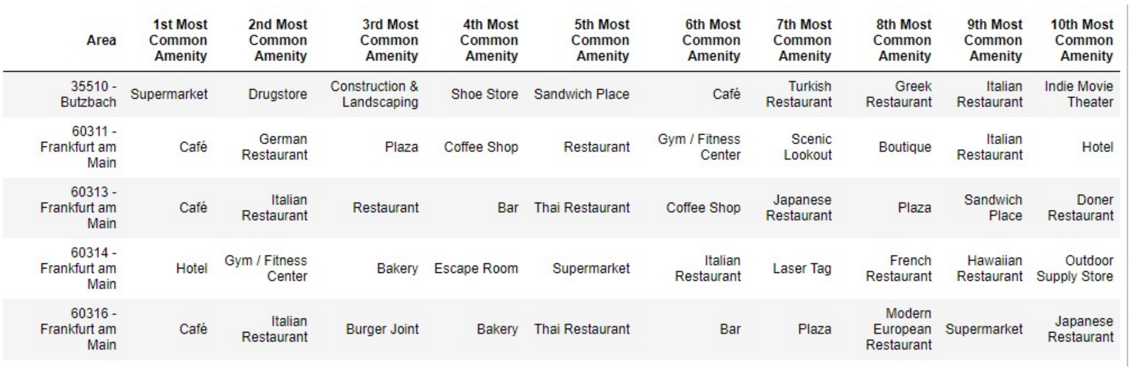
1. We have many Food related amenities.
2. We are poor in Entertainment and Arts amenities.
3. The food category (formal plus informal) makes up almost 50% of all amenities.
4. For fun the entertainment & arts, points of interest, outdoors & health and Nightlife Spots make up 19% of all amenities.

Then we plotted some maps and ended up with a heatmap with the amenities location where we could see the density:



We prepared the data for k-means clustering first by encoding the categories using onehot encoding and then we grouped by area of study and calculated the mean of the frequency of each category.

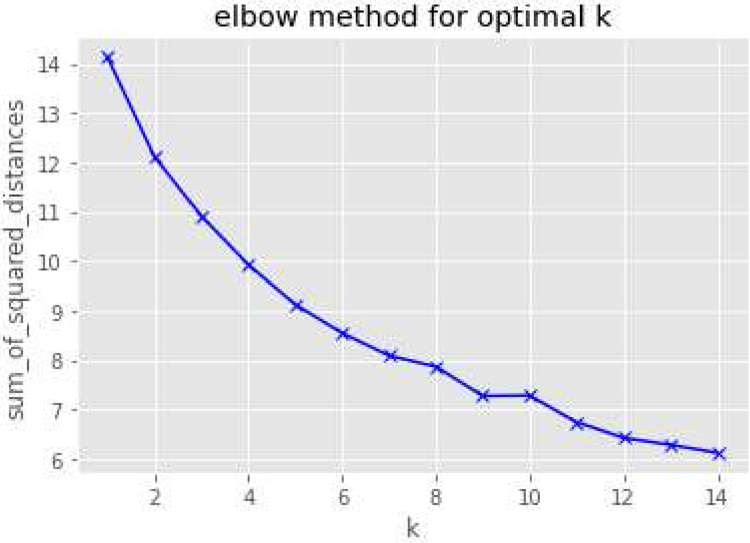
With this step we were able to build a table where we could see the most common amenities in each area:



Area	1st Most Common Amenity	2nd Most Common Amenity	3rd Most Common Amenity	4th Most Common Amenity	5th Most Common Amenity	6th Most Common Amenity	7th Most Common Amenity	8th Most Common Amenity	9th Most Common Amenity	10th Most Common Amenity
35510 - Butzbach	Supermarket	Drugstore	Construction & Landscaping	Shoe Store	Sandwich Place	Café	Turkish Restaurant	Greek Restaurant	Italian Restaurant	Indie Movie Theater
60311 - Frankfurt am Main	Café	German Restaurant	Plaza	Coffee Shop	Restaurant	Gym / Fitness Center	Scenic Lookout	Boutique	Italian Restaurant	Hotel
60313 - Frankfurt am Main	Café	Italian Restaurant	Restaurant	Bar	Thai Restaurant	Coffee Shop	Japanese Restaurant	Plaza	Sandwich Place	Doner Restaurant
60314 - Frankfurt am Main	Hotel	Gym / Fitness Center	Bakery	Escape Room	Supermarket	Italian Restaurant	Laser Tag	French Restaurant	Hawaiian Restaurant	Outdoor Supply Store
60316 - Frankfurt am Main	Café	Italian Restaurant	Burger Joint	Bakery	Thai Restaurant	Bar	Plaza	Modern European Restaurant	Supermarket	Japanese Restaurant

For the remaining analysis we used machine learning clustering algorithms is k-means which allowed us to group the data according to the existing similarities in k clusters, given as input to the algorithm.

First, we needed to decide which number of cluster to use so we decided to use an elbow plot:

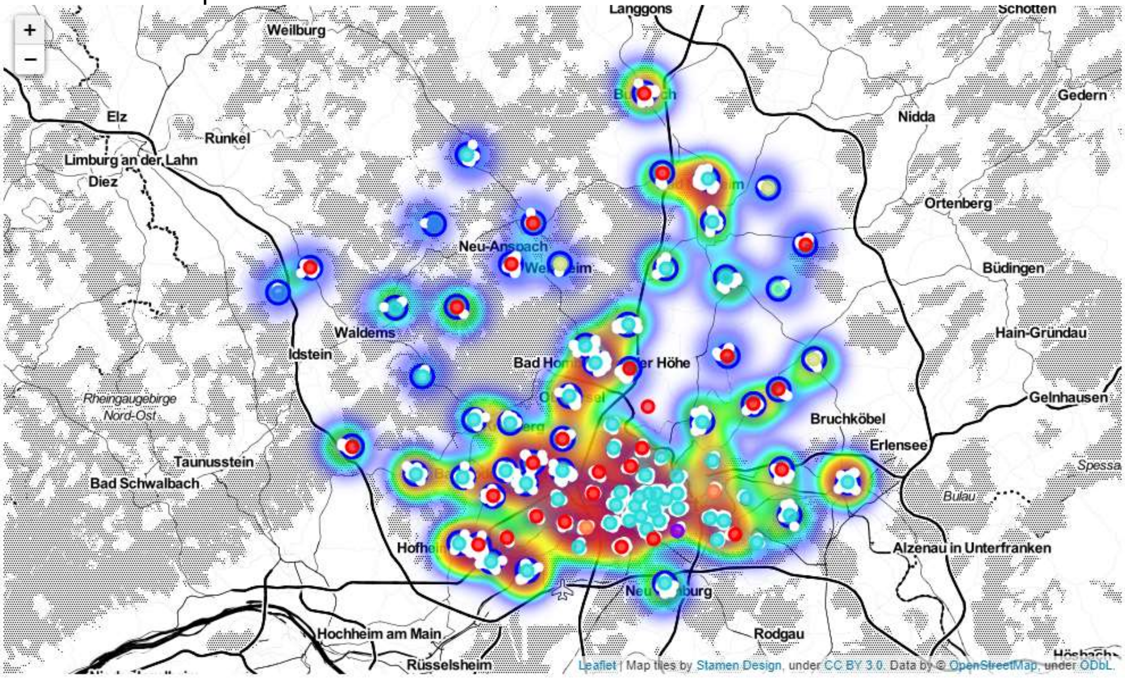


From the graph above we chose k = 7 as the optimal number of clusters for our analysis. We had some doubts whether to choose 7 or 8, but the subsequent metrics analysis proved better for 7.

Our analysis using K Means algorithm, return the following metrics:

```
Silhouette Coefficient:  0.1586
Calinski Harabasz Score:  10.3375
Davies-Bouldin Index:  1.2264
```

And the final map with clusters:



The colored dots represent the clusters.

4. Results

From our analysis we can see that the areas from cluster 0 and 3 have the most amenities while the remaining clusters only have between 1 and 3 areas with very few amenities.

Considering amenities density, cluster 0 areas, are near yellow zones and cluster 3 areas, are near red zones in the heatmap.

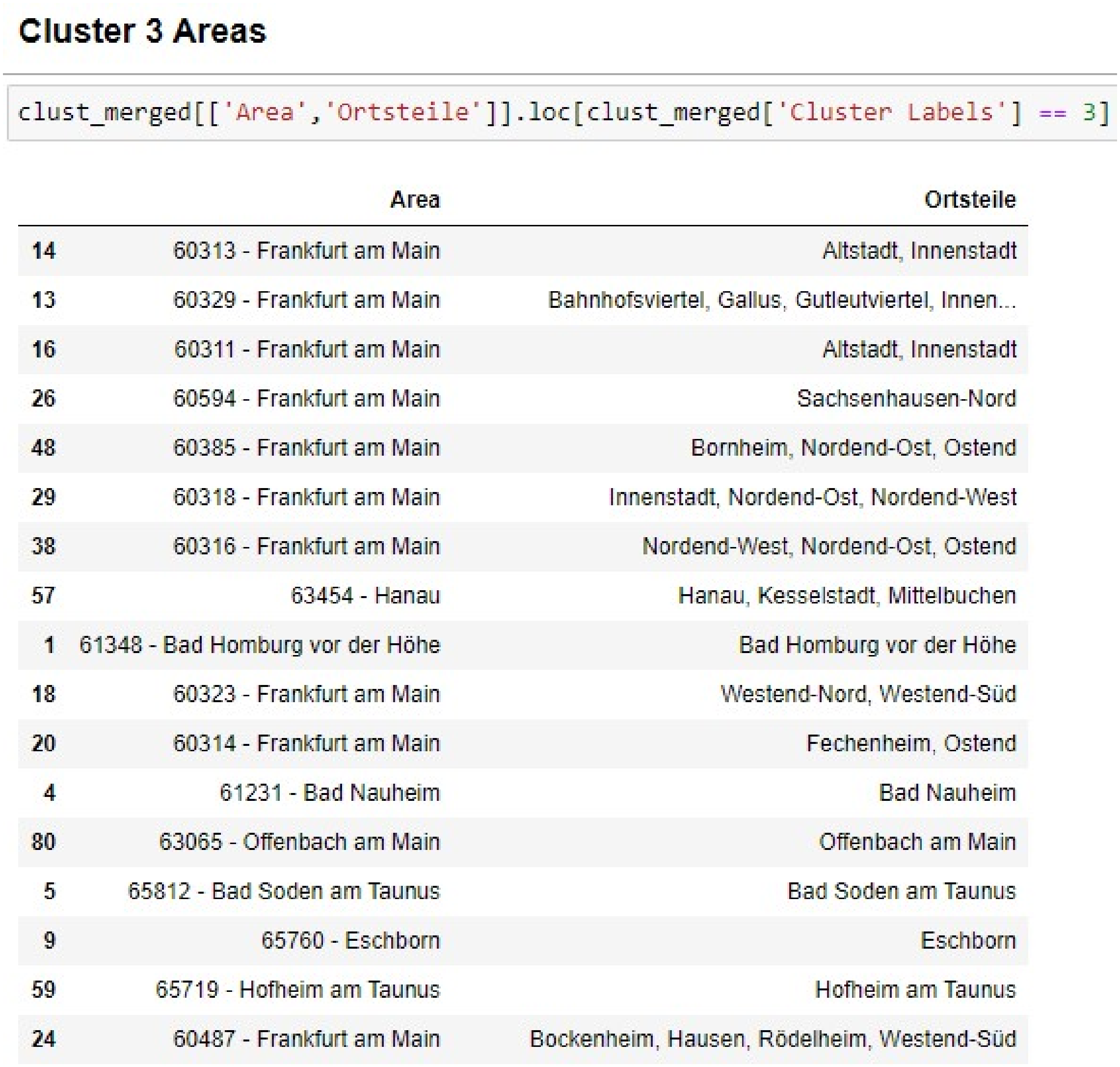
Ranking the clusters by number of amenities available we get:

1. **Cluster 3 with the most amenities of all and greater variety category wise. Food related amenities account for the most common amenities in all areas**.

2. **Cluster 0 contains mostly shopping related amenities. Supermarket is the most common amenity in all clustered areas.**

3. **The remaining clusters 1,2,4,5 and 6 contain areas with low amenities density and little variety.**

Cluster 3 contains 1436 amenities, while cluster 0 has 211 amenities and the remaining 26 belong to the combined clusters 1,2,4,5,6.

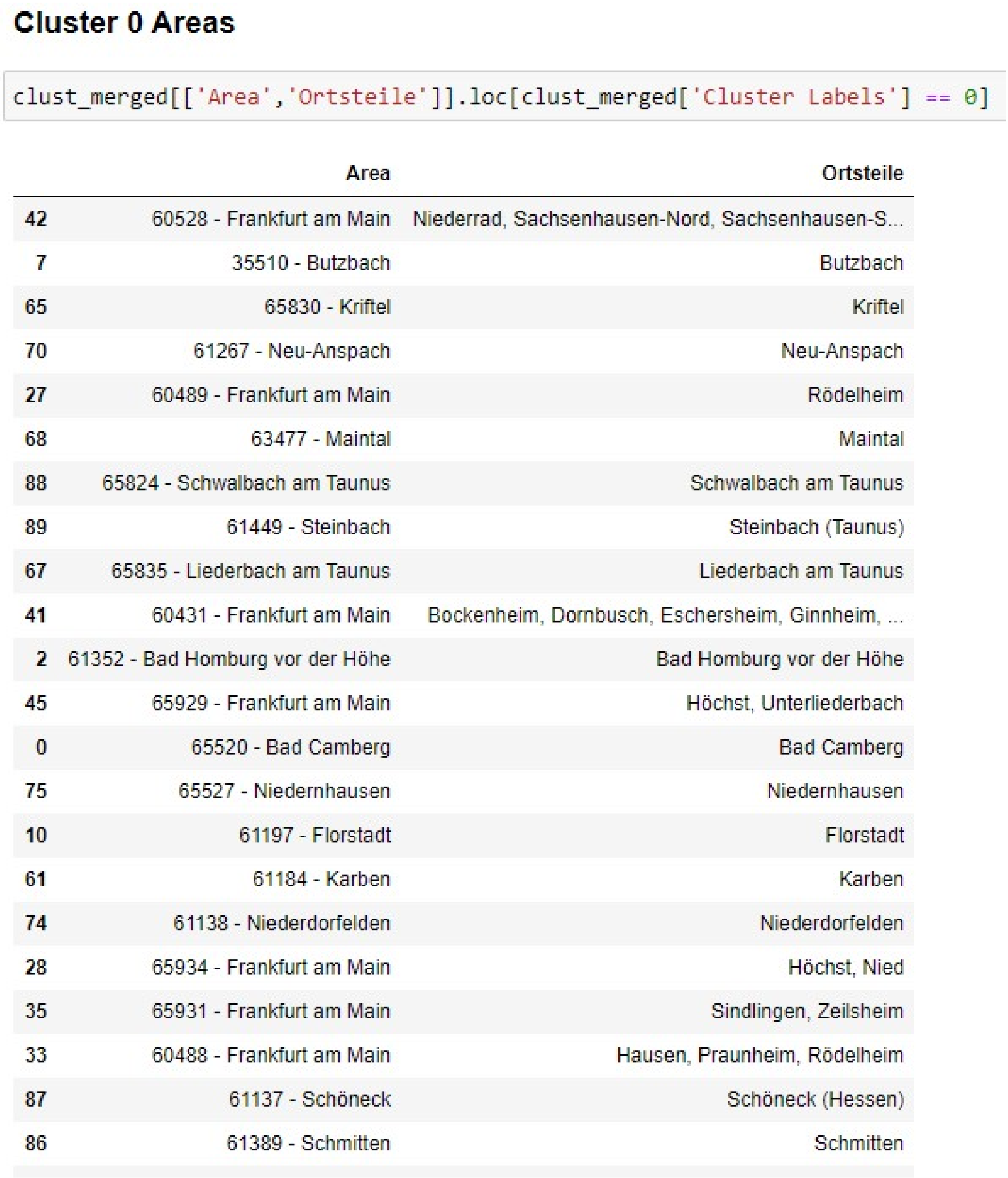


Cluster 3 Areas

```
clust_merged[['Area','Ortsteile']].loc[clust_merged['Cluster Labels'] == 3]
```

	Area	Ortsteile
14	60313 - Frankfurt am Main	Altstadt, Innenstadt
13	60329 - Frankfurt am Main	Bahnhofsviertel, Gallus, Gutleutviertel, Innen...
16	60311 - Frankfurt am Main	Altstadt, Innenstadt
26	60594 - Frankfurt am Main	Sachsenhausen-Nord
48	60385 - Frankfurt am Main	Bornheim, Nordend-Ost, Ostend
29	60318 - Frankfurt am Main	Innenstadt, Nordend-Ost, Nordend-West
38	60316 - Frankfurt am Main	Nordend-West, Nordend-Ost, Ostend
57	63454 - Hanau	Hanau, Kesselstadt, Mittelbuchen
1	61348 - Bad Homburg vor der Höhe	Bad Homburg vor der Höhe
18	60323 - Frankfurt am Main	Westend-Nord, Westend-Süd
20	60314 - Frankfurt am Main	Fechenheim, Ostend
4	61231 - Bad Nauheim	Bad Nauheim
80	63065 - Offenbach am Main	Offenbach am Main
5	65812 - Bad Soden am Taunus	Bad Soden am Taunus
9	65760 - Eschborn	Eschborn
59	65719 - Hofheim am Taunus	Hofheim am Taunus
24	60487 - Frankfurt am Main	Bockenheim, Hausen, Rödelheim, Westend-Süd

This is only a part of the cluster.

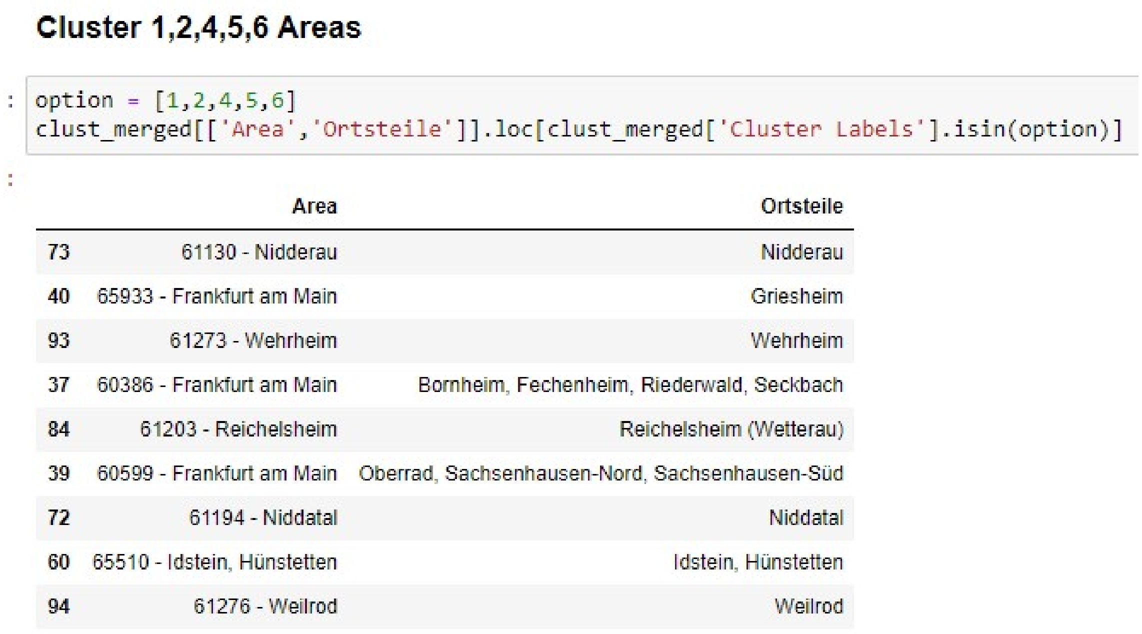


Cluster 0 Areas

```
clust_merged[['Area','Ortsteile']].loc[clust_merged['Cluster Labels'] == 0]
```

	Area	Ortsteile
42	60528 - Frankfurt am Main	Niederrad, Sachsenhausen-Nord, Sachsenhausen-S...
7	35510 - Butzbach	Butzbach
65	65830 - Kriftel	Kriftel
70	61267 - Neu-Anspach	Neu-Anspach
27	60489 - Frankfurt am Main	Rödelheim
68	63477 - Maintal	Maintal
88	65824 - Schwalbach am Taunus	Schwalbach am Taunus
89	61449 - Steinbach	Steinbach (Taunus)
67	65835 - Liederbach am Taunus	Liederbach am Taunus
41	60431 - Frankfurt am Main	Bockenheim, Dornbusch, Eschersheim, Ginnheim, ...
2	61352 - Bad Homburg vor der Höhe	Bad Homburg vor der Höhe
45	65929 - Frankfurt am Main	Höchst, Unterliederbach
0	65520 - Bad Camberg	Bad Camberg
75	65527 - Niedernhausen	Niedernhausen
10	61197 - Florstadt	Florstadt
61	61184 - Karben	Karben
74	61138 - Niederdorfelden	Niederdorfelden
28	65934 - Frankfurt am Main	Höchst, Nied
35	65931 - Frankfurt am Main	Sindlingen, Zeilsheim
33	60488 - Frankfurt am Main	Hausen, Praunheim, Rödelheim
87	61137 - Schöneck	Schöneck (Hessen)
86	61389 - Schmitten	Schmitten

This is only a part of the cluster.



Cluster 1,2,4,5,6 Areas

```
option = [1,2,4,5,6]
clust_merged[['Area','Ortsteile']].loc[clust_merged['Cluster Labels'].isin(option)]
```

	Area	Ortsteile
73	61130 - Nidderau	Nidderau
40	65933 - Frankfurt am Main	Griesheim
93	61273 - Wehrheim	Wehrheim
37	60386 - Frankfurt am Main	Bornheim, Fechenheim, Riederwald, Seckbach
84	61203 - Reichelsheim	Reichelsheim (Wetterau)
39	60599 - Frankfurt am Main	Oberrad, Sachsenhausen-Nord, Sachsenhausen-Süd
72	61194 - Niddatal	Niddatal
60	65510 - Idstein, Hünstetten	Idstein, Hünstetten
94	61276 - Weilrod	Weilrod

5. Discussion

In our study we explored the area surrounding **Bad Homburg vor der Höhe**, to find the locations with the most amenities available, towards a goal of defining the best places to buy a house or live.

After defining the locations, we wished to study and getting all the relevant data about those places, we searched for coordinates (latitude and longitude). When we plotted the different coordinates in a map and defined a radius for the amenities search, we decided that a 500 meters radius was not enough to get a significant, and sometimes any, amenities result, with the exception of Frankfurt am Main and Offenbach am Main. So, we decided that we would do two search radius one 500 meters and the other 1000 meters.

To further increase the amenities search result, we changed the coordinates used. At first we were using postal codes to get all coordinates but we noticed that some postal codes coordinates would be far from the city centers, resulting in not getting any amenities, so we decided that for the cities with only one postal code we would get the coordinates from the city center. This means that apart from **Bad Homburg vor der Höhe**, Frankfurt am Main and Offenbach am Main, the coordinates we got were from the city centers.

After compiling our main dataset, we used a machine learning algorithm - K Means Clustering - to cluster all areas according to their amenities profile.

The results indicate that we have two major clusters 3 and 0.

The first relevant cluster, Cluster 3, show us areas to live that are rich in amenities in number and variety. The main amenities categories present are food related, Italian restaurants, hotels and German restaurants are respectively the first, second and third most common amenities in these areas.

The second relevant cluster, Cluster 0, show us areas to live that easy to shop. The areas in this cluster contain much less amenities than the previous cluster, but they have a good mix. Supermarkets are the first, second and third most common amenities in these areas.

These results matter as a start for a larger study, and also there is room for improvement in this study. We should try to expand our amenities dataset, by pinpointing optimal coordinates or getting other sources. We should also include new datasets like housing market, quality of living indexes, etc. to explore other avenues of research. We should also try other machine learning clustering algorithms and compare results.

6. Conclusion

This research aimed to provide information about areas to buy a house based on amenities availability. For that we used a machine learning algorithm - K means clustering - in a dataset comprised of the mean of the frequency of occurrence of each category in each area. As a result, we found the relevant clusters.

For us it is clear, based on our study, that pertaining to amenities density **one should buy in one of the areas in cluster 3**.

This, of course does not mean that you should be rushing to buy in any of those areas, further developing of this research is needed, this is only a starting point.