Housing rent prices and venues data analysis of Frankfurt

Rocco Incardona

Description of the problem

Frankfurt am Main is the largest financial centre in continental Europe and the fifth-most populous city in Germany with over 763,380 inhabitants and a population density of 3,100 people per square kilometre. The city has experienced rising housing prices despite the impact of Covid19. Last year UBS ranked Frankfurt as having the second-most overvalued housing market of any major city in the world, behind Munich, according to its Global Real Estate Bubble Index. Living in this city, I decided to use Frankfurt for my IBM Capstone project.

People moving to Frankfurt may want to choose districts where rental prices are lower but can offer entertainment, shops, bars, restaurants and are not too far from the workplace. At the same time, recent restrictions and the possibility to work from home have pushed more and more people to leave central districts in order to find cheaper and bigger apartments in the suburbs.

Especially for expats like me, it might be difficult to find this information. For this reason, I decided to dedicate the final project of my **IBM Data Science course** to this topic.

Exploiting the information available on Foursquare and real estate websites, we can create an information chart where the rental price index by district is displayed on the map of Frankfurt and each district is clustered according to the venue density. In this way, it is possible to find districts, which have similar venues composition but different rental prices.

Data description

The main data sources, which were used in this project, are:

- **Foursquare API** to get the most common venues of given district of Frankfurt;
- **Rental price index** as available in the following page: https://www.miet-check.de/stadtteile_uebersicht.php?stadt=Frankfurt%20am%20Main;
- **List of Frankfurt's districts** as available in the following Wikipedia page: https://de.wikipedia.org/wiki/Liste_der_Stadtteile_von_Frankfurt_am_Main;
- **Nominatim – OpenStreetMap** services to retrieve the geo coordinates of the districts of Frankfurt: https://nominatim.org/release-docs/develop/api/Overview/.

Methodology section

The first step of the data retrieval consisted in getting the list of Frankfurt's district (plus some additional demographic information) from the corresponding Wikipedia page via web scraping. I then used the **Open Street Map** services and the **geopandas** library to retrieve the geo coordinates of the centre and borders of each of the district.

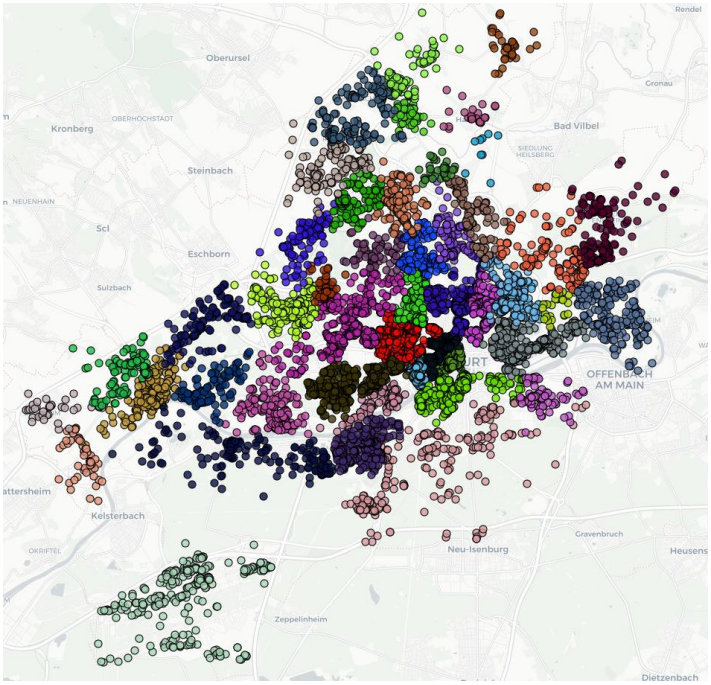
	District	Surface_km2	population	female_pop	male_pop	german_pop	foreign_pop	Latitude	Longitude	Coordinates	geometry
0	Altstadt	506	4218	2065	2153	2669	1549	50.11044	8.68235	[[8.6748666, 50.1092226], [8.6757349, 50.10751...	POLYGON ((8.67487 50.10922, 8.67573 50.10752, ...
1	Innenstadt	1491	6599	3112	3487	3539	3060	50.11456	8.68359	[[8.6685139, 50.1132235], [8.6690588, 50.11253...	POLYGON ((8.66851 50.11322, 8.66906 50.11253, ...
2	Bahnhofsviertel	542	3552	1321	2231	1706	1846	50.10841	8.66815	[[8.6605492, 50.1095394], [8.6625518, 50.10855...	POLYGON ((8.66055 50.10954, 8.66255 50.10856, ...
3	Westend-Süd	2497	19314	9839	9475	14006	5308	50.11524	8.66227	[[8.6433346, 50.1140438], [8.6434858, 50.10917...	POLYGON ((8.64333 50.11404, 8.64349 50.10918, ...
4	Westend-Nord	1632	10373	5391	4982	7366	3007	50.12636	8.66792	[[8.6558043, 50.1266272], [8.6564224, 50.12597...	POLYGON ((8.65580 50.12663, 8.65642 50.12597, ...
...	...	...	...	...	...	...	...	...	...	...	...
41	Kalbach-Riedberg	6580	21795	11047	10748	16683	5112	50.18628	8.63905	[[8.6140631, 50.1767957], [8.6172061, 50.17465...	POLYGON ((8.61406 50.17680, 8.61721 50.17466, ...
42	Harheim	4837	5234	2664	2570	4396	838	50.18229	8.69297	[[8.6711429, 50.1865714], [8.6727365, 50.18564...	POLYGON ((8.67114 50.18657, 8.67274 50.18565, ...
43	Nieder-Eschbach	6348	11518	5932	5586	8756	2762	50.20173	8.66671	[[8.6436283, 50.2023429], [8.6446848, 50.20191...	POLYGON ((8.64363 50.20234, 8.64468 50.20192, ...
44	Bergen-Enkheim	12601	17941	9240	8701	14321	3620	50.15801	8.76204	[[8.735863, 50.1629854], [8.7394536, 50.159920...	POLYGON ((8.73586 50.16299, 8.73945 50.15992, ...
45	Frankfurter Berg	2400	8168	4110	4058	5987	2181	50.16802	8.67569	[[8.6546814, 50.1735908], [8.6553804, 50.17321...	POLYGON ((8.65468 50.17359, 8.65538 50.17322, ...

46 rows × 11 columns

For each of the districts, I also retrieved information on the rental price index from www.miet-check.de, taking care of correctly merging the information by district name with the original data frame. The rental price index is measured as Euro per square meter.

The most important data collection step consisted in getting the list of venues from **Foursquare API** service. For this purpose, I used the "*search*" endpoint to get the list of all the possible venues around the specified coordinates. Using the "*explore*" endpoint could have been an alternative, but I would have retrieved only a list of the best venues per district. For the purpose of this project however, I needed to have access to the largest and most representative sample of venues for each district. In addition to this, to overcome Foursquare's limitation of 50 venues per query, I randomly generated 10 points within the boundaries of each district and used each of the points to run my queries. I used a very big radius (5km) to retrieve as many venues possible and I ensured that each retrieved venue was assigned to the correct district.

At the end of this process, I removed the duplicated venues and I was left with around 11000 venues (results are affected by the random generation of the geo coordinates) spread around 45 districts. The venues were correctly allocated to each district as show in the map below, which was realized using **folium** (venues coloured by district).

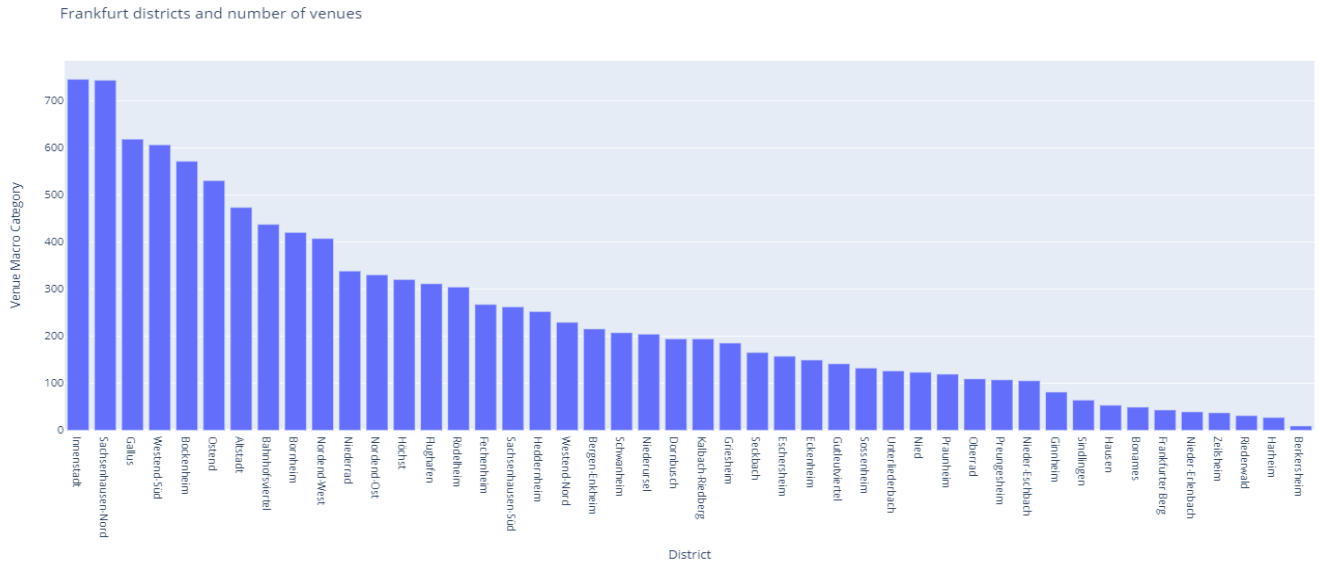


For each of the venues, I collected information on the geo coordinates, the venue category and the **macro category**.

:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Macro Category
0	Altstadt	50.11044	8.68235	Hotel am Dom	50.110951	8.685683	Hotel	travel
1	Altstadt	50.11044	8.68235	Dortmunder Pils-Treff	50.111178	8.686056	Dive Bar	nightlife
2	Altstadt	50.11044	8.68235	Kaiserdom St. Bartholomäus	50.110640	8.685411	Church	building
3	Altstadt	50.11044	8.68235	Paulaner am Dom	50.110876	8.685925	German Restaurant	food
4	Altstadt	50.11044	8.68235	Haus am Dom	50.110922	8.684831	Museum	arts_entertainment

The **venues density** appears to be very different across districts, with the city centres including as expected the largest share of venues.



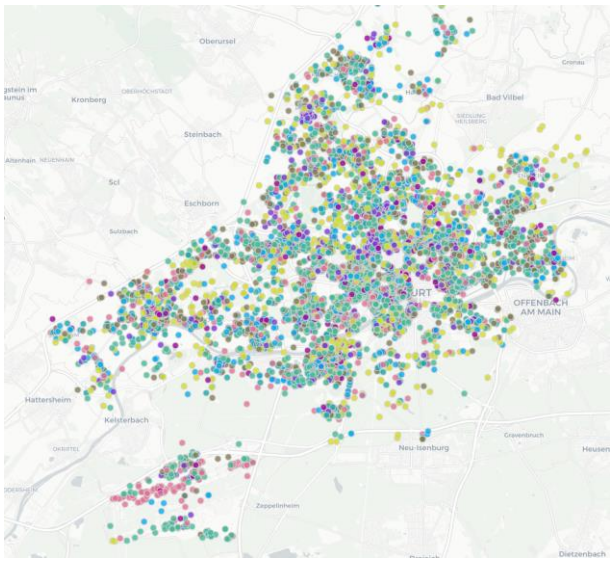
Around 560 unique venue categories were retrieved distributed across eight macro categories (shops, building, food, travel, parks_outdoors, arts_entertainment, nightlife and education). I did a quick inspection of the venues per macro categories and decided to reallocate some the venues from the “building” group to other macro categories in order to make the differences between categories more clear. For this reason, I also generated a new “health” group, containing venues related to doctors and hospital activities. The final distribution across the nine categories was the following:

shops	2562
building	2412
food	2221
travel	972
parks_outdoors	743
arts_entertainment	716
health	551
nightlife	548
education	503

The four most common venues per macro category are displayed below:

```
Macro category - shops ---> Salon / Barbershop, Miscellaneous Shop, Pharmacy, Automotive Shop
Macro category - food ---> Café, Italian Restaurant, German Restaurant, Bakery
Macro category - building ---> Office, Building, Residential Building (Apartment / Condo), Event Space
Macro category - arts_entertainment ---> Art Gallery, General Entertainment, Gym / Fitness Center, Soccer Field
Macro category - parks_outdoors ---> Park, Playground, Garden, Other Great Outdoors
Macro category - nightlife ---> Bar, Nightclub, Pub, Lounge
Macro category - travel ---> Hotel, Bus Stop, Tram Station, Airport Gate
Macro category - health ---> Doctor's Office, Dentist's Office, Medical Center, Hospital
Macro category - education ---> General College & University, College Classroom, School, College Academic Building
```

On a first glance the different categories of venues seemed to be distributed equally across Frankfurt. Each venue in the map below is coloured by macro category.

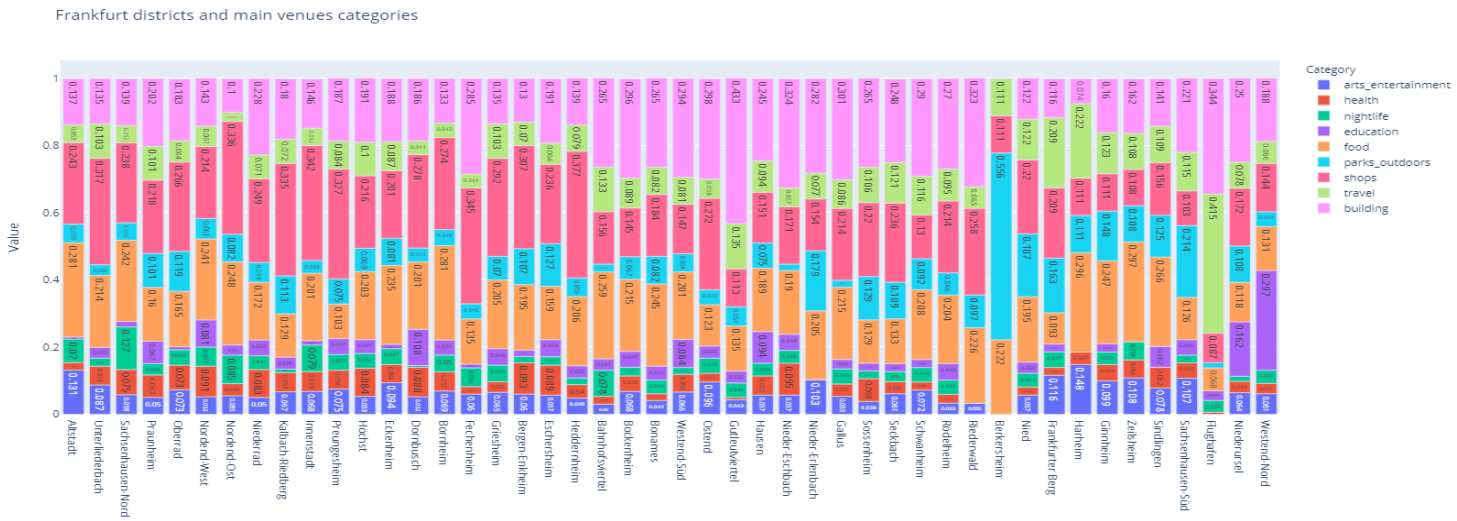


Since we wanted to find similar districts and move to an area where rent prices are lower (or higher, according to our preferences), the next step consisted in clustering the districts using **K-means algorithm**, one of the most common unsupervised machine-learning algorithm.

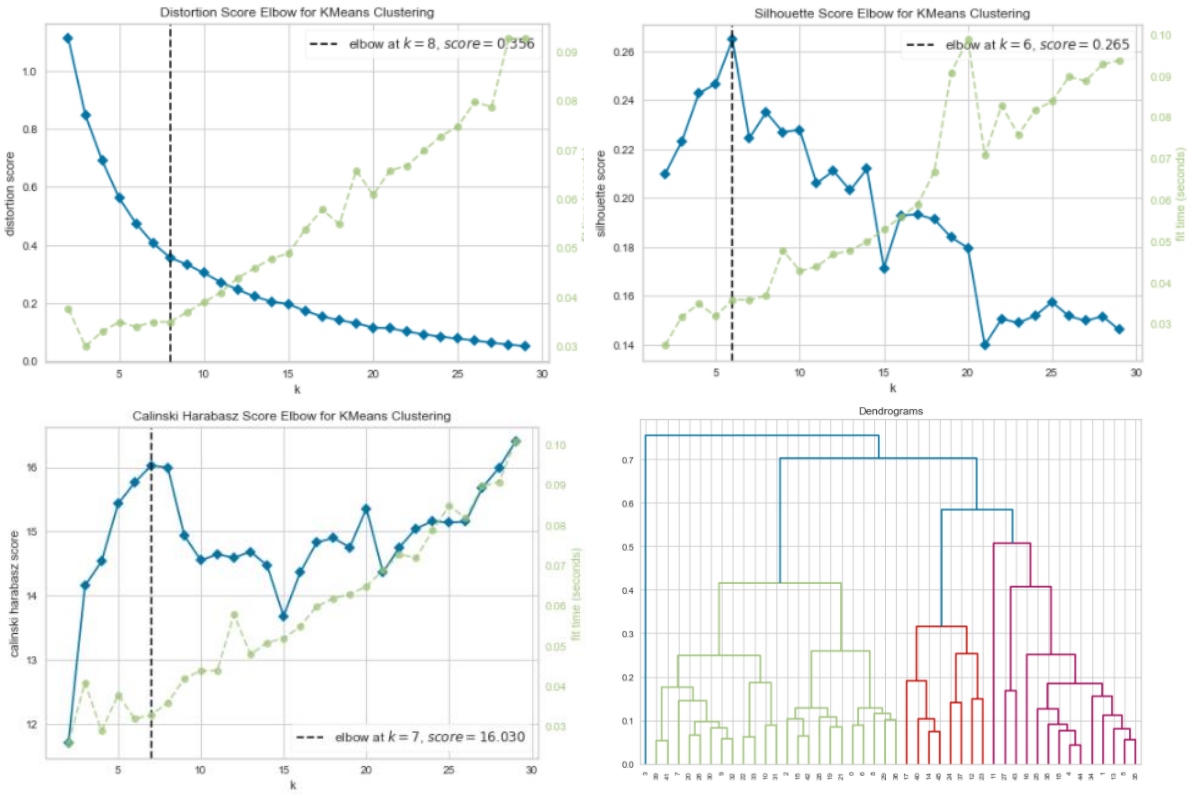
The general idea is to count how many times certain venues occur in a specific district and then find districts which present similar venues pattern. For this project, I used the distribution of venues by macro categories to compute the similarity between districts. I performed a one hot encoding of the retrieved venues by macro categories and then computed the mean frequency of occurrence of each category per district. To have an idea of the result, the extract below shows for instance that “Altstadt” has a score of 0.28 in the food category, meaning that 28% of the retrieved venues located in this district are on average restaurants, cafes or bakeries.

	District	arts_entertainment	building	education	food	health	nightlife	parks_outdoors	shops	travel
0	Altstadt	0.131078	0.137421	0.006342	0.281184	0.023256	0.069767	0.054968	0.243129	0.052854
1	Bahnhofsviertel	0.029748	0.265446	0.034325	0.258581	0.022883	0.077803	0.022883	0.155606	0.132723

The K-means algorithm would try to group together districts, which present similar composition of venues. The following charts gives a first glance on the venues composition of the separate Frankfurt’s districts, before clustering:

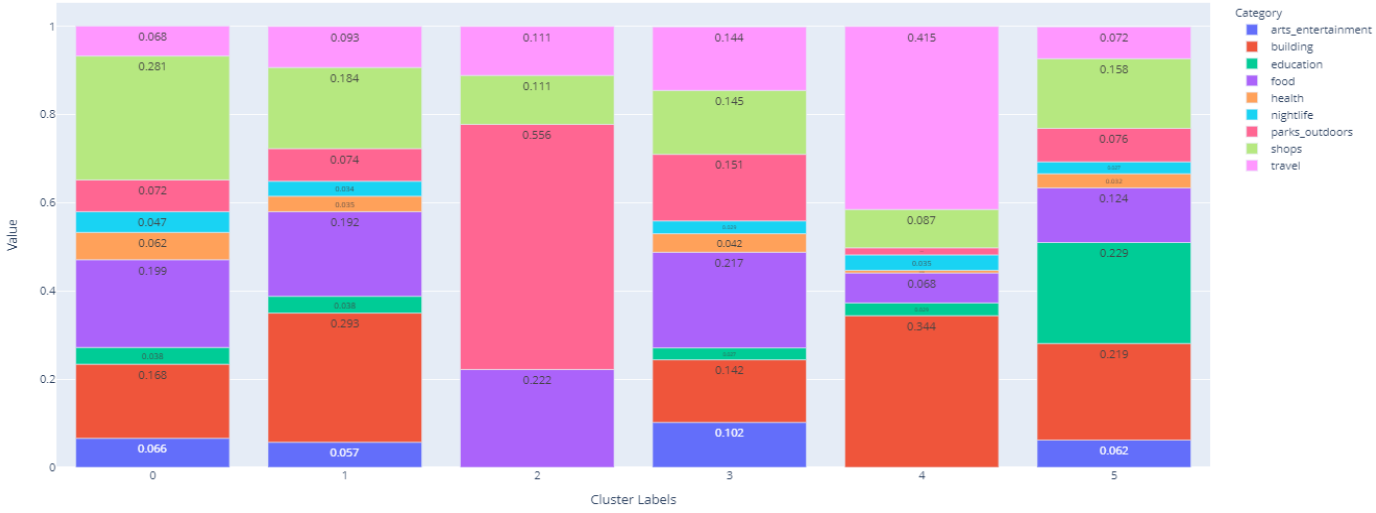


In order to choose the appropriate number of clusters I used several statistical indicators, namely the Elbow method, the Silhouette score, the Calinski Harabasz score and a Dendogram[1].



The Elbow method, which is the most common approach, seemed to indicate an optimal number of clusters k=8, but the chart did not to show a sharp decrease of distortion for that value of k. At the same time the Silhouette and the Calinski Harabasz score seemed to point to a lower value of k, respectively 6 and 7. I decided to opt for a k=6 since the Dendogram suggested as well a lower number of clusters and considering that the venues distribution across Frankfurt seemed quite homogeneous in our previous map.

After applying the algorithm, I calculated the average venue composition for each cluster:



[1] Nicely explained in the following article: https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad

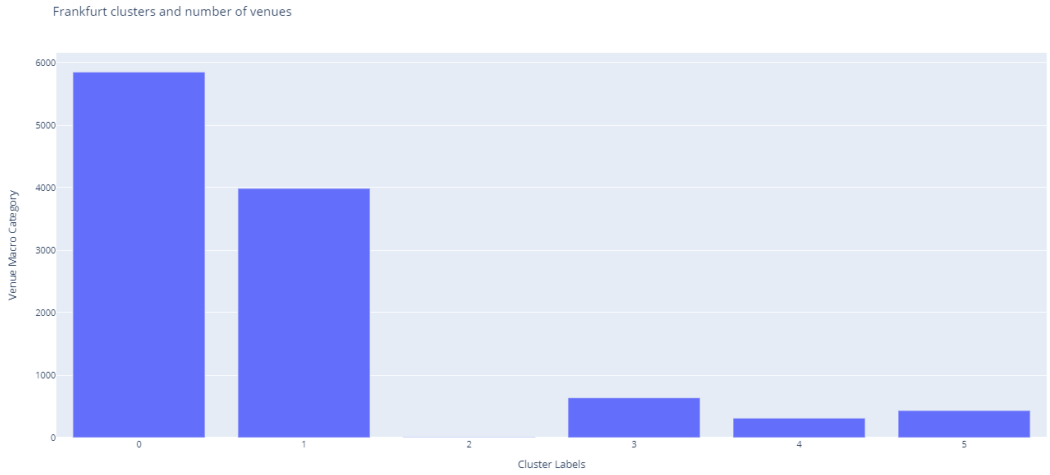
The districts associated to the different groups are:

- **Cluster 0**: Altstadt, Bergen-Enkheim, Bornheim, Dornbusch, Eckenheim, Eschersheim, Fechenheim, Griesheim, Heddernheim, Höchst, Innenstadt, Kalbach-Riedberg, Niederrad, Nordend-Ost, Nordend-West, Oberrad, Praunheim, Preungesheim, Sachsenhausen-Nord, Unterliederbach.
- **Cluster 1**: Bahnhofsviertel, Bockenheim, Bonames, Gallus, Gutleutviertel, Hausen, Nieder-Erlenbach, Nieder-Eschbach, Ostend, Riederwald, Rödelheim, Schwanheim, Seckbach, Sossenheim, Westend-Süd.
- **Cluster 2**: Berkersheim.
- **Cluster 3**: Frankfurter Berg, Ginnheim, Harheim, Nied, Sachsenhausen-Süd, Sindlingen, Zeilsheim.
- **Cluster 4**:Flughafen.
- **Cluster 5**: Niederursel, Westend-Nord.

Looking at the most common venues inside each cluster, we could label the groups as follow:

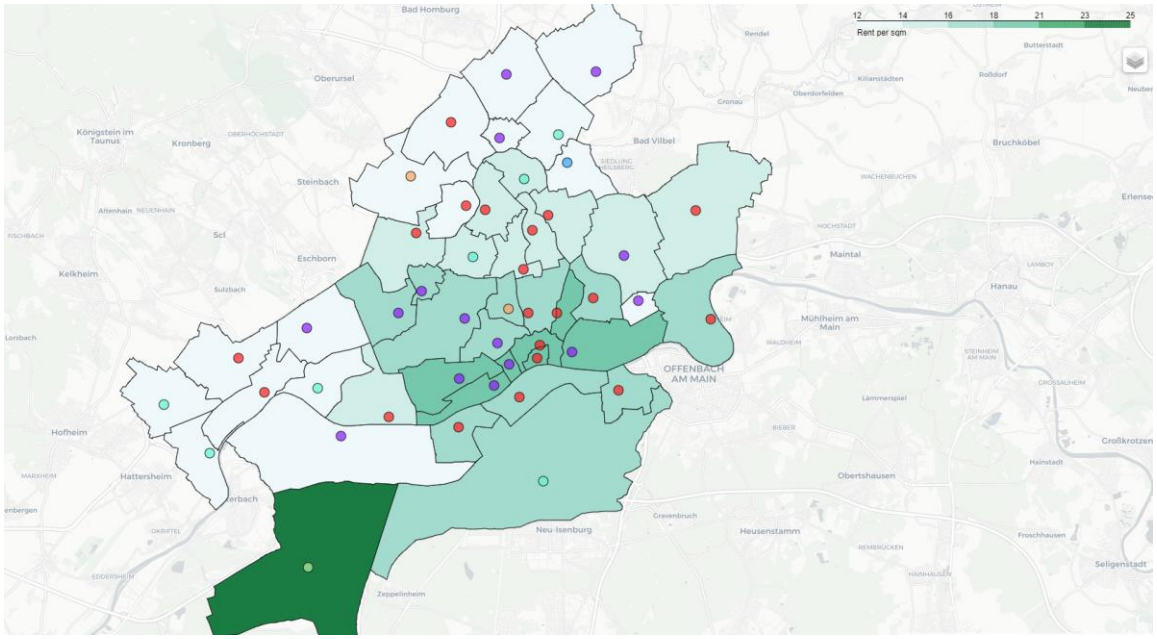
- **Cluster 0**: “Shops , nightlife and health”
- **Cluster 1**: “Offices and restaurants”
- **Cluster 2**: “Berkersheim parks”
- **Cluster 3**: “Art and entertainment and outdoors”
- **Cluster 4**: “Airport”
- **Cluster 5**: “University campus and offices”

The clusters with the most remarkable differences are cluster 2 and 4, which contain respectively only one district. They can be considered outliers in venues composition: for Berkersheim only few venues were retrieved predominantly in the park_outdoors category, mostly due to its peripheral position and its vicinity to the Nidda River; for the Airport several venues were retrieved but its composition is by definition skewed toward travel activities (hotels, airport gates and tram/bus stops). For number of districts, also Cluster 5 is relatively small and mostly influenced by the presence of the Goethe University campuses, while cluster 3 has a relatively larger share of entertainment and outdoor venues. Cluster 0 and 1 are the biggest ones as revealed by a quick inspection of the venue distribution across clusters.

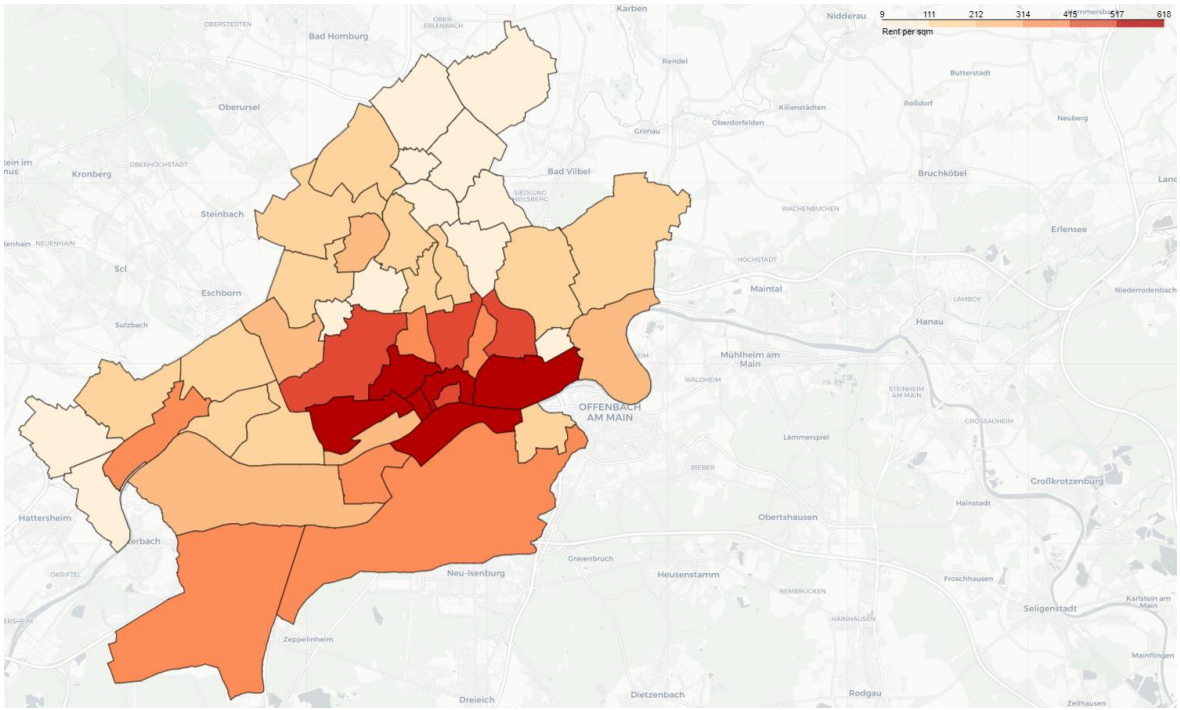


Results

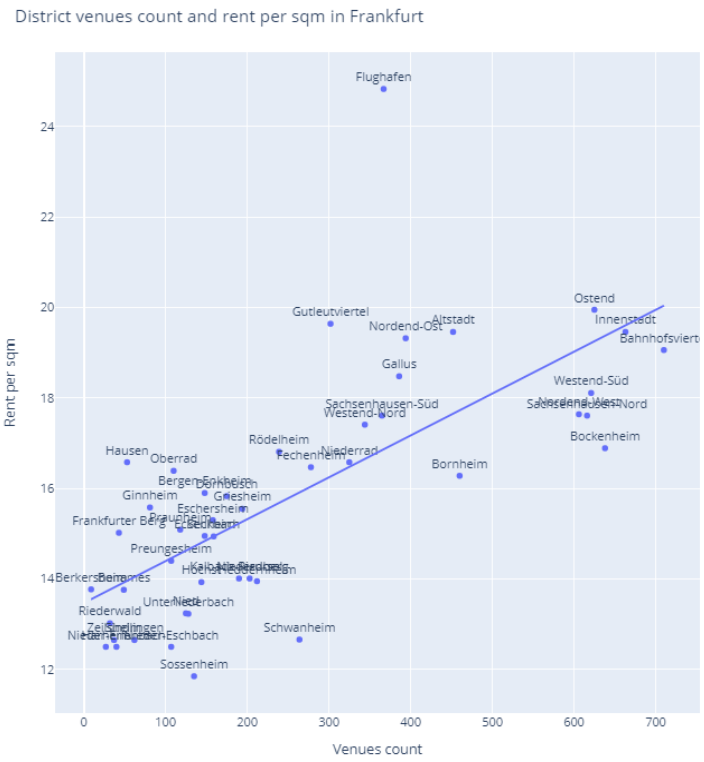
The final step of the analysis was then to plot the rental price index by district on the map of Frankfurt together with the district clustered according to the venue density. The map below shows the different districts of Frankfurt coloured by rent price per sqm; the circles inside the districts are the centre of the areas and are coloured according to the cluster found in the previous section.

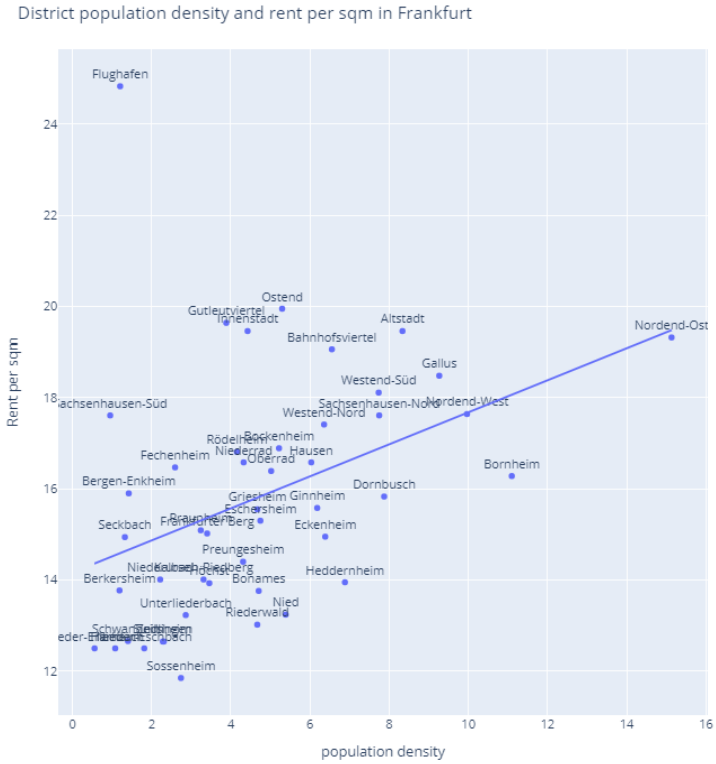


As expected, rent prices in the city centre are higher than in the periphery (with the exception of the Airport district). In addition, cluster 0 and 1 seems to define an east-west split in the city, but overall the venue density analysis highlighted a quite balanced distribution of venues categories across Frankfurt. Therefore, someone looking for a district with several shops and pubs (red dots) for instance could decide to relocate to the north-centre of the city where prices are lower and venue density is not too low.



The venue density is definitely one of the factors explaining the higher prices in the city centre as shown in the chart below on the left: areas with more venues are associated with higher rental prices. However, several other factor are definitely playing a role in price dynamics: from buildings supply to share of foreign people or population density. The chart below on the right shows for instance that venues with higher population density (defined as population per square kilometre) have also higher rent per square meter.





Discussion

As I mentioned before, Frankfurt is a big city with a high population density in a narrow area. The total number of venues retrieved in this project is limited and does not fully capture the complexity of the venues structure of the city. Information on the rental prices by district is also quite limited.

One way to improve the data availability on the rental prices could be to use the RestAPI services provided by online renting portals like immowelt.de or Immobilien Scout 24. Despite requiring special authorizations, these type of services can offer detailed up-to-date information on apartments demand across the city. Regarding the venues retrieved, data that are more accurate could be retrieved using a premium Foursquare account or trying to query the API with equally distanced points within the district, rather than using random ones.

Furthermore, other factors in addition to venues composition might be useful for future tenants to understand which districts are more attractive and convenient. In this sense, additional public sources[2] can be used to add new perspectives to this analysis.

Finally, from a methodological point of view, other models other than K-means clustering could be used (e.g. Mean-Shift Clustering, DBSCAN, Agglomerative Hierarchical Clustering etc.) to optimize the classification of districts and find better similarity patterns.

[2] See for instance https://statistik.stadt-frankfurt.de/strukturdatenatlas/stadtteile/html/atlas.html

Conclusion

Use of geolocation data offers endless applications when combined with other data sources, like socio-economic data. Machine learning also comes at hand when processing large amount of data like the number of venues and no a-priori knowledge is available. With this project, I was successfully able to combine the two and obtain some practical suggestions for future tenants of Frankfurt. Further work can be dedicated to refine the data availability and the clustering methodology and to include additional variables explaining districts similarities and the interaction with rental prices.

Sources

- https://statistik.stadt-frankfurt.de/strukturdatenatlas/stadtteile/html/atlas.html
- https://www.miet-check.de/stadtteile_uebersicht.php?stadt=Frankfurt%20am%20Main
- https://en.wikipedia.org/wiki/Frankfurt
- https://www.ft.com/content/3e4f8c40-1dca-447e-a3c4-69911cfc162a
- https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad