Clustering Neighborhoods in Frankfurt am Main

COURSERA CAPSTONE PROJECT

Name: Rohan Desai

Date:

Course of study: IBM Data Science Professional Certificate

1. Introduction

1.1 Business problem

A client is interested in opening a franchise of their Asian restaurant chain in the city of Frankfurt am Main, preferably close to the city center. It will be their first restaurant in the city, and they want us to find out which would be the best neighborhood/district to open an Asian restaurant in the city. Additionally, the results of the clustering algorithm in this project can also be used by someone interested in moving to Frankfurt and wanting to know about the cuisines available in the various districts.

2. Data

2.1 Data sources

Following datasets have been used in this project:

1. **Street Directory of the city of Frankfurt am Main: https://offenedaten.frankfurt.de/dataset/strassenverzeichnis-der-stadt-frankfurt-am-main**
2. Foursquare API to get the most common venues in Frankfurt districts.
3. Demographics of Frankfurt am Main Neighborhoods : https://offenedaten.frankfurt.de/dataset/stadtteilprofile-bevoelkerung

2.2 Data gathering and cleaning

The districts of the city of Frankfurt am Main will be analyzed in this project.

- **Data 1** : Street directory of Frankfurt am Main:
  This dataset will be used to extract the district names and postcodes in Frankfurt. It is available as a CSV file and can be accessed via the link given above. Frankfurt contains 46 city districts. This is a huge dataset containing 4540 rows and 15 columns. Therefore, it was necessary to shorten and clean it by keeping only the data that is required. It is a street directory, which is why the dataset is so big. It was shortened to extract only the district names and postcodes. The resultant dataset contained 46 rows (one for each district) and 3 columns.

- **Data 2 :**
  The geographical coordinates of the districts will be utilized as input for Foursquare API that will be leveraged to extract information for each district respectively. We will use the Foursquare API to explore the districts in Frankfurt. We use Foursquare API to get the most common venues for each district. Foursquare returns a json file, from which required data needs to be extracted. We only extract the venue name, category and geographical coordinates for each venue. These are then stored in a separate dataframe, for use in clustering.

- **Data 3 :** Frankfurt Demographics:
  This dataset contains district-wise distribution of population for the city of Frankfurt. It also contains useful data about the percentage of foreigners, and specifically population of various ethnicities in the districts. It contains 46 rows (one for each district) and 164 columns. It needs to be shortened to analyze. Only the required columns were picked from this dataset, which contained information about the total population of each district, population of foreigners and so on. Moreover, the column names are in German. These were translated to English for easy understanding.

- **Data 4 :** Frankfurt neighborhoods GeoJSON:
  The geoJSON file is required for plotting the Choropleth maps to analyze the demographics of Frankfurt districts. The district names in this file must match the district names in the dataset which is intended to be plotted. After checking, it was found that the districts of Bahnhofsviertel and Gutleutviertel are combined into a single district in the geoJSON file. Thus, the 2 district rows were merged in the demographics dataset. Also, there was an issue with the German letters containing umlauts, i.e. ü, ä, ö. Hence, districts containing these letters were also renamed as per the characters found in their equivalent names in the geoJSON file.

	Sozialrathaus Name	Stadtteil Name	Postleitzahl,
0	Sachsenhausen	Flughafen	60549,
1	Sachsenhausen	Flughafen	60549,
2	Sachsenhausen	Sachsenhausen-N	60594,
3	Sachsenhausen	Sachsenhausen-N	60596,
4	Gallus	Gallus	60326,
5	Nord	Ginnheim	60431,
6	Dornbusch	Dornbusch	60431,
7	Bockenheim	Bockenheim	60486,
8	Bockenheim	Bockenheim	60486,
9	Nord	Nieder-Eschbach	60437,

Fig.1. Overview of the Districts and Postal codes in Frankfurt

	District	Latitude	Longitude	Postal Code
0	Altstadt	50.110442	8.682901	60311
1	Bahnhofsviertel	50.108411	8.668151	60329
2	Bergen-Enkheim	50.158015	8.762039	60388
3	Berkersheim	50.173289	8.697312	60435
4	Bockenheim	50.123311	8.646056	60486
5	Bonames	50.181347	8.663331	60437
6	Bornheim	50.129731	8.710612	60385
7	Dornbusch	50.139046	8.675271	60431
8	Eckenheim	50.151710	8.679746	60435
9	Eschersheim	50.158203	8.656212	60439

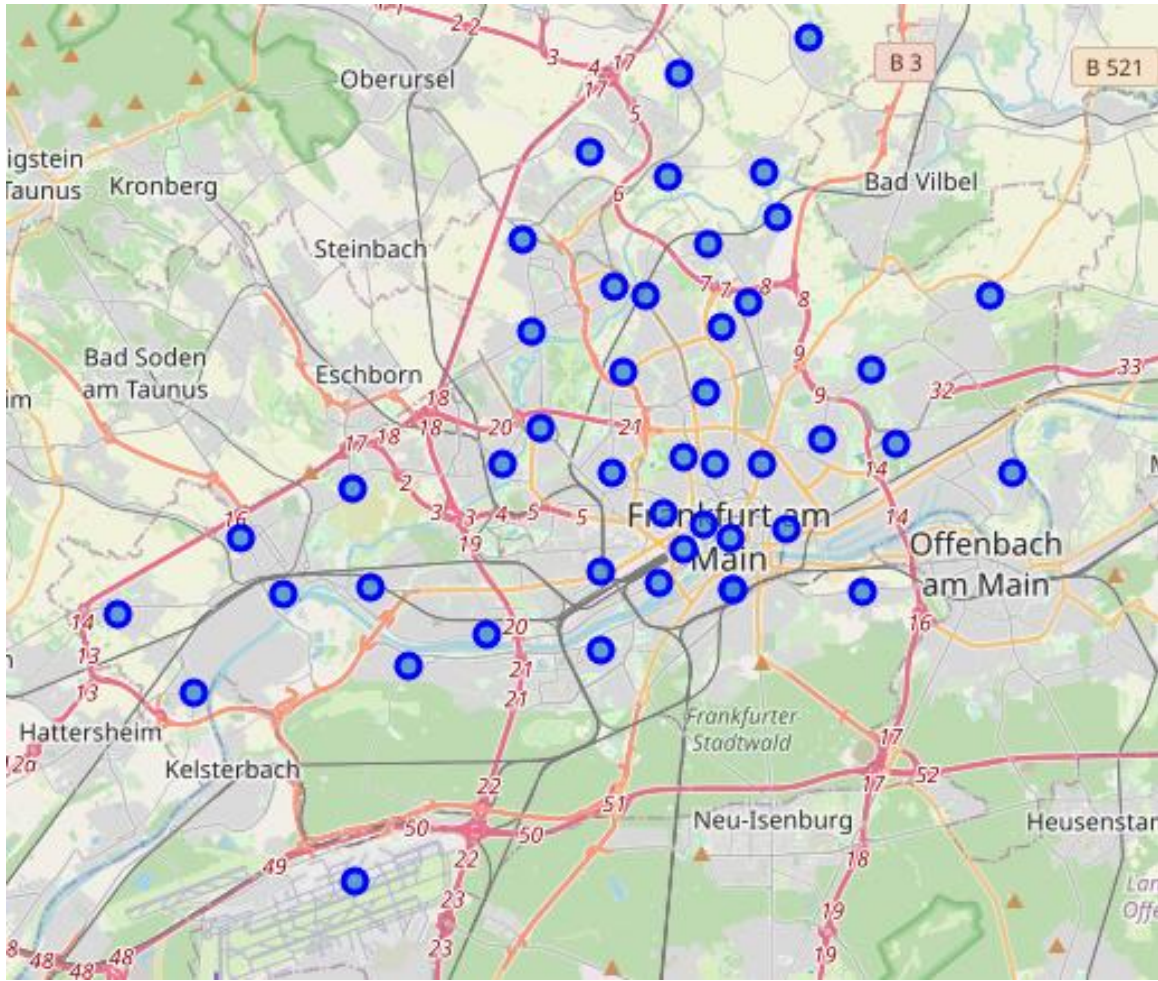Fig.2. Dataset containing the latitude and longitude for each District

Fig. 3. Map of Frankfurt am Main with its Districts

3. Methodology

3.1 Analytical Approach

In this project, we shall first use k-means clustering to cluster the neighborhoods in Frankfurt. Frankfurt has 46 districts. We shall use the geocoder to get the geographical coordinates for each of these districts. We will use Foursquare API to explore the districts using their coordinates, and get the most common venues in each district. Based on this information, we shall cluster the districts using k-means and take a look at each cluster. We need to look at clusters with a greater number of Asian and similar cuisine restaurants, as that indicates that there is demand for Asian cuisine in that cluster.

Then we shall use the demographics data to find the districts with a greater population and compare that with the cluster data. We shall find districts which have more Asian restaurants as well as a sizeable Asian population, as these will be ideal for opening a new Asian restaurant. Additionally, we shall also look at closeby districts with lesser Asian restaurants but a sizeable Asian population, as this is also a good prospect, due to less competition in the area.

3.2 Foursquare API

This section describes the algorithm further. As stated in section 2 the venue data for each district gets fetched from the Foursquare API and is then used to cluster each neighborhood according to the available venues in a radius within 500 meters from the fetched latitude and longitude values. Then the data is explored further. The top five most common venues are printed for each district. After this, the venue categories are one-hot encoded and grouped by the district names. This enables the kmeans algorithm to work with the data, because a Machine Learning algorithm is not able to work on textual data. The results can be seen in the Jupyter notebook. In the next step, the ten most common venues for each district are stored in a new data frame. This data frame helps to get an idea of the venues in each district. Figure 4 shows the first five rows of the resulting data frame.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Altstadt	Café	Plaza	Restaurant	German Restaurant	Burger Joint	Art Museum	Gym / Fitness Center	Coffee Shop	Pastry Shop	Bar
1	Bahnhofsviertel	Hotel	Bar	Indian Restaurant	Café	Vietnamese Restaurant	Thai Restaurant	Chinese Restaurant	Seafood Restaurant	Theater	Drugstore
2	Bergen-Enkheim	Plaza	Bus Stop	Drugstore	Farmers Market	Farm	Falafel Restaurant	Exhibit	Event Space	Ethiopian Restaurant	Electronics Store
3	Berkersheim	Light Rail Station	Farm	Shipping Store	German Restaurant	Zoo	Farmers Market	Falafel Restaurant	Exhibit	Event Space	Ethiopian Restaurant
4	Bockenheim	Italian Restaurant	Café	Asian Restaurant	Supermarket	Turkish Restaurant	Ice Cream Shop	Japanese Restaurant	Wine Bar	Drugstore	Spanish Restaurant

Fig.4. Overview of top venues for each District in Frankfurt am Main

3.2 Clustering using K-means

The one hot encoded and grouped data is the input to the kmeans algorithm and the number of clusters is set to five. The resulting cluster labels are then additionally stored in the data frame containing the ten most common venues for each district.

	District	Latitude	Longitude	Postal Code	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mos Commo Venu
0	Altstadt	50.110442	8.682901	60311	0	Café	Plaza	Restaurant	German Restaurant	Burger Joint	Art Museum	Gym / Fitness Center	Coffee Shop	Pastr Sho
1	Bahnhofsviertel	50.108411	8.668151	60329	0	Hotel	Bar	Indian Restaurant	Café	Vietnamese Restaurant	Thai Restaurant	Chinese Restaurant	Seafood Restaurant	Theate
2	Bergen-Enkheim	50.158015	8.762039	60388	3	Plaza	Bus Stop	Drugstore	Farmers Market	Farm	Falafel Restaurant	Exhibit	Event Space	Ethiopia Restaurar
3	Berkersheim	50.173289	8.697312	60435	4	Light Rail Station	Farm	Shipping Store	German Restaurant	Zoo	Farmers Market	Falafel Restaurant	Exhibit	Ever Spac
4	Bockenheim	50.123311	8.646056	60486	0	Italian Restaurant	Café	Asian Restaurant	Supermarket	Turkish Restaurant	Ice Cream Shop	Japanese Restaurant	Wine Bar	Drugstor

Fig.5. Cluster labels added to dataset after kmeans clustering (Overview)

The dataframe containing the cluster labels and top venues was then merged with the dataframe containing latitude and longitude as seen in Figure 5. This data was then used to visualize the clusters on a map using Folium.
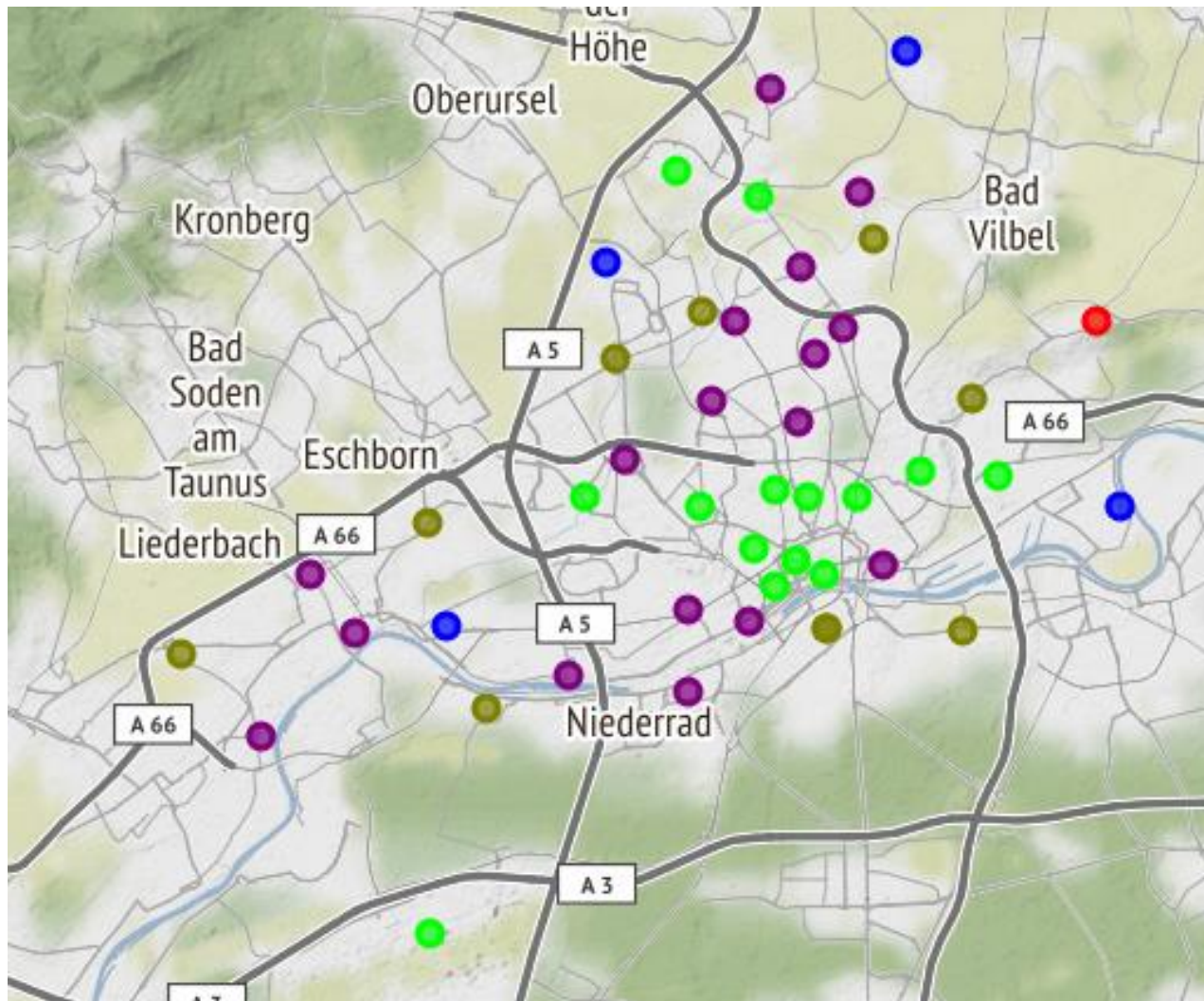
Fig.6. Map of clusters of Frankfurt districts

3.3 Examining the clusters

Cluster 0 (Light green):

This cluster contains Italian restaurants as its most common venue. But it also contains German, Indian and Thai restaurants. Hence we shall name this the Multicuisine cluster. Note that the districts in this cluster are close to the city center. That will have an effect on the decision to the business problem.

Cluster 1 (Blue):

This cluster contains 4 districts, all of which are on the outskirts of the city. There aren't many venues to be found in this cluster, although it does contain German restaurants as it's most common venue. We shall name this the outskirts cluster.

The fact that the districts in this cluster are on the outskirts of the city far from the city center, automatically eliminates this cluster from contention, as the client requires a neighborhood close to the city center.

Cluster 2 (Purple):

This cluster contains mostly hotels and supermarkets. Hence we shall name this the Hotels cluster. The abundance of hotels indicates that there will be more tourists in the areas in this cluster. Also, we can see that the districts in this cluster are not too far from the city center. Hence, this is also an important cluster for making our decision.

Cluster 3 (Red):

This cluster contains only 1 district, which does not have many popular venues as per Foursquare. Hence we shall name this the Bergen-Enkheim cluster.

Cluster 4 (Olive Green):

This cluster contains German restaurants as its most common venue. We shall name this the German restaurants cluster. Additionally, there are Italian and Thai restaurants as well in this cluster, but not many.

3.4 Observations

We observe that the purple and light green clusters contain the most districts and the most number of venues. While the light green cluster contains more restaurants, the purple cluster contains more hotels, which indicates tourists. We can see that a variety of cuisines are offered in the light green cluster, indicating that they cater to a variety of customers. Most of the districts are located close to the city center. These factors make this cluster the most eligible for opening a new Asian restaurant. But, the fact that there are many restaurants, could also be a negative factor as that means more competition - not a good thing for a new restaurant to go up against established eateries catering to customers for years.

The purple cluster, on the other hand, although it does not contain many restaurants, has a lot of hotels and is pretty close to the city center. Fewer restaurants means lesser competition, and more tourists, some of them Asian, means more prospective customers and if one finds a location not too far from the city center, an Asian restaurant here could flourish.

To know which district specifically would be perfect for opening an Asian restaurant, we look at the district-wise demographics of Frankfurt am Main, and then explore districts from both the light green and purple clusters.

3.5 Data Exploration – Frankfurt demographics

The demographics dataset contains district-wise distribution of population for the city of Frankfurt. It also contains useful data about the percentage of foreigners, and specifically population of various ethnicities in the districts. Only the required columns were picked from this dataset, which contained information about the total population of each district, population of foreigners and so on. This dataset was then merged with the dataset containing the latitude and longitudes of the districts. The resulting dataset is as seen in Figure 7.

	District	Population structure Residents 2012	Population structure of foreigners in 2012	Population structure of foreigners in % 2012	Population structure average age 2012	Population structure Young adults aged 18 to 29 in 2012	Population structure Foreign nationals from 18 to 29 years 2012	Population structure Residents aged 30 to 64 in 2012	Population structure Foreigners aged 30 to 64 in 2012	Population structure Foreigners from Asia and Australia 2012	Latitude	Longitude	Postal Code
0	Altstadt	3601	1254	34.823660	43.4	647	280	1981	761	272	50.110442	8.682901	60311
1	Bahnhofsviertel	3117	1630	52.293872	37.5	865	443	1804	960	329	50.108411	8.668151	60329
2	Bergen-Enkheim	17563	2842	16.181746	44.3	2229	532	8942	1763	254	50.158015	8.762039	60388
3	Berkersheim	3643	611	16.771891	38.9	463	101	1752	376	127	50.173289	8.697312	60435
4	Bockenheim	35789	10170	28.416553	38.9	7489	2546	19357	6048	1586	50.123311	8.646056	60486

Fig.7. Frankfurt demographics data overview

3.6 Data visualization: Choropleth maps

The data from the demographics dataset is then plotted on a Choropleth map to visualize the population distribution across the city of Frankfurt. This data will then be used to select districts based on the earlier clustering results to explore further.
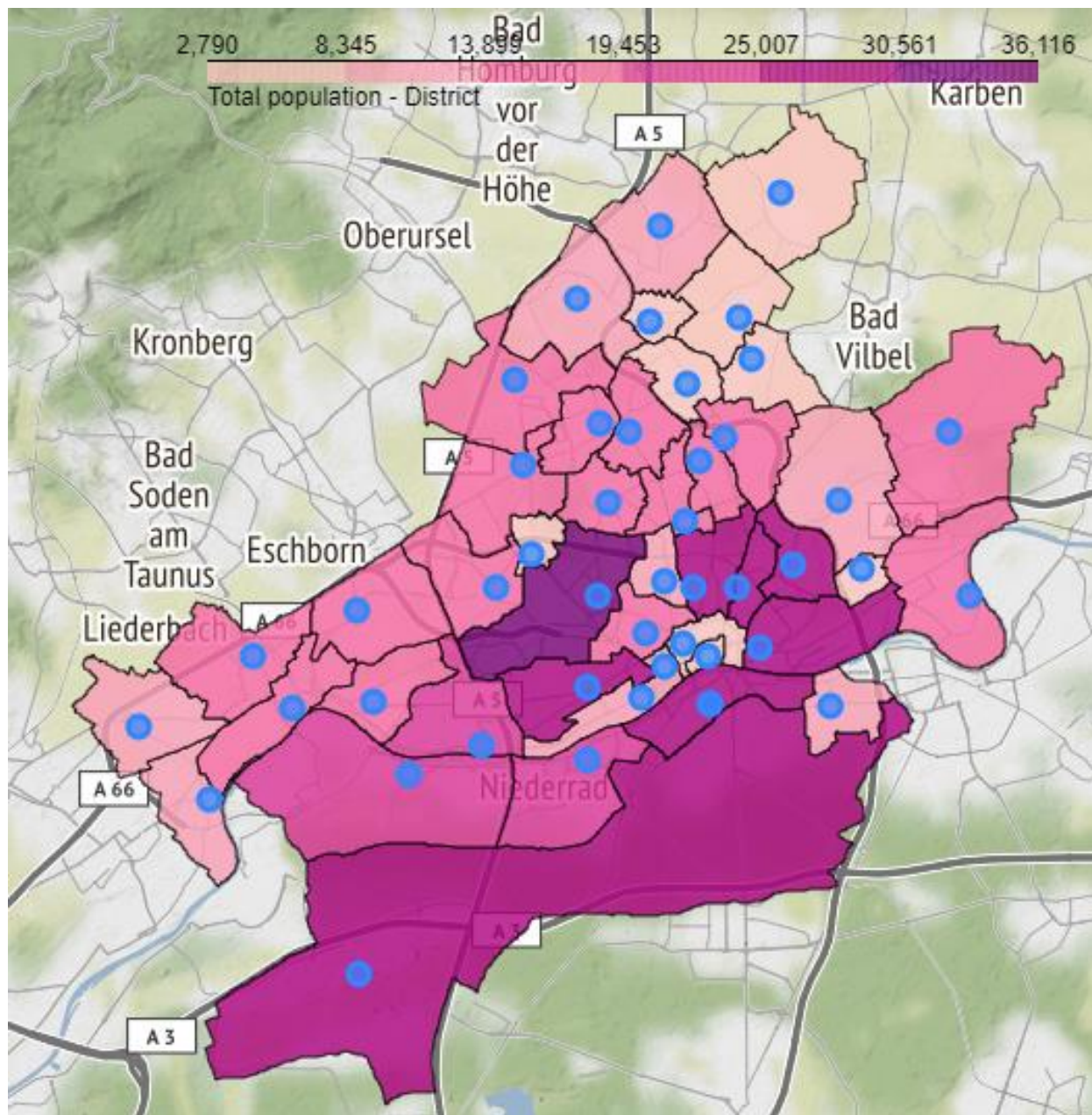
Fig.8. District-wise population distribution – Frankfurt am Main

From this map, we observe that the central districts have the highest populations in Frankfurt, along with the district of Flughafen on the outskirts.

Next we take a look at the distribution of foreigner population in Frankfurt.
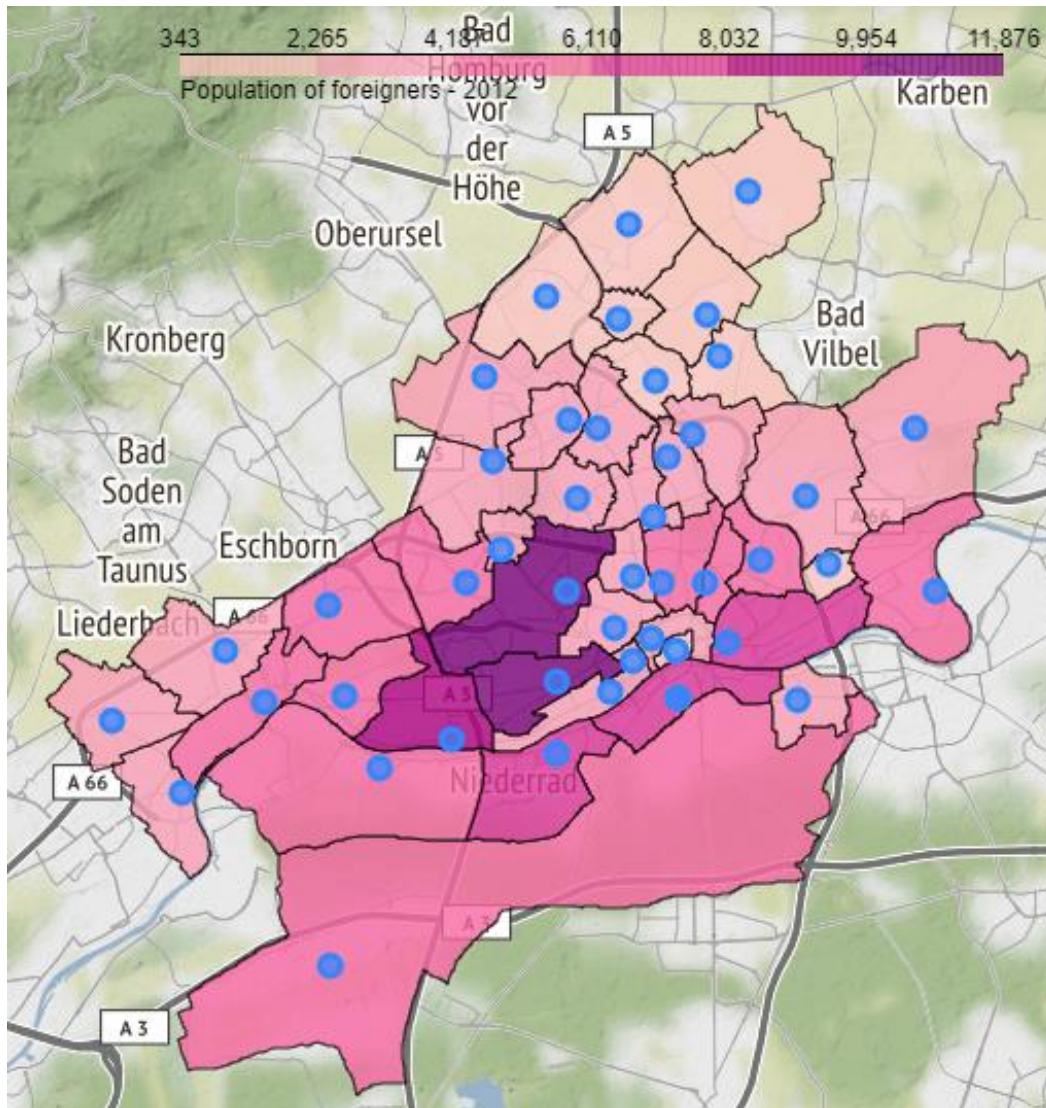
Fig.9. District-wise distribution of foreigner population – Frankfurt am Main

It can be observed from the map that the 2 central districts of Gallus and Bockenheim have the highest populations of foreigners.

We shall now take a look at the population distribution of Asians and Australians in Frankfurt.
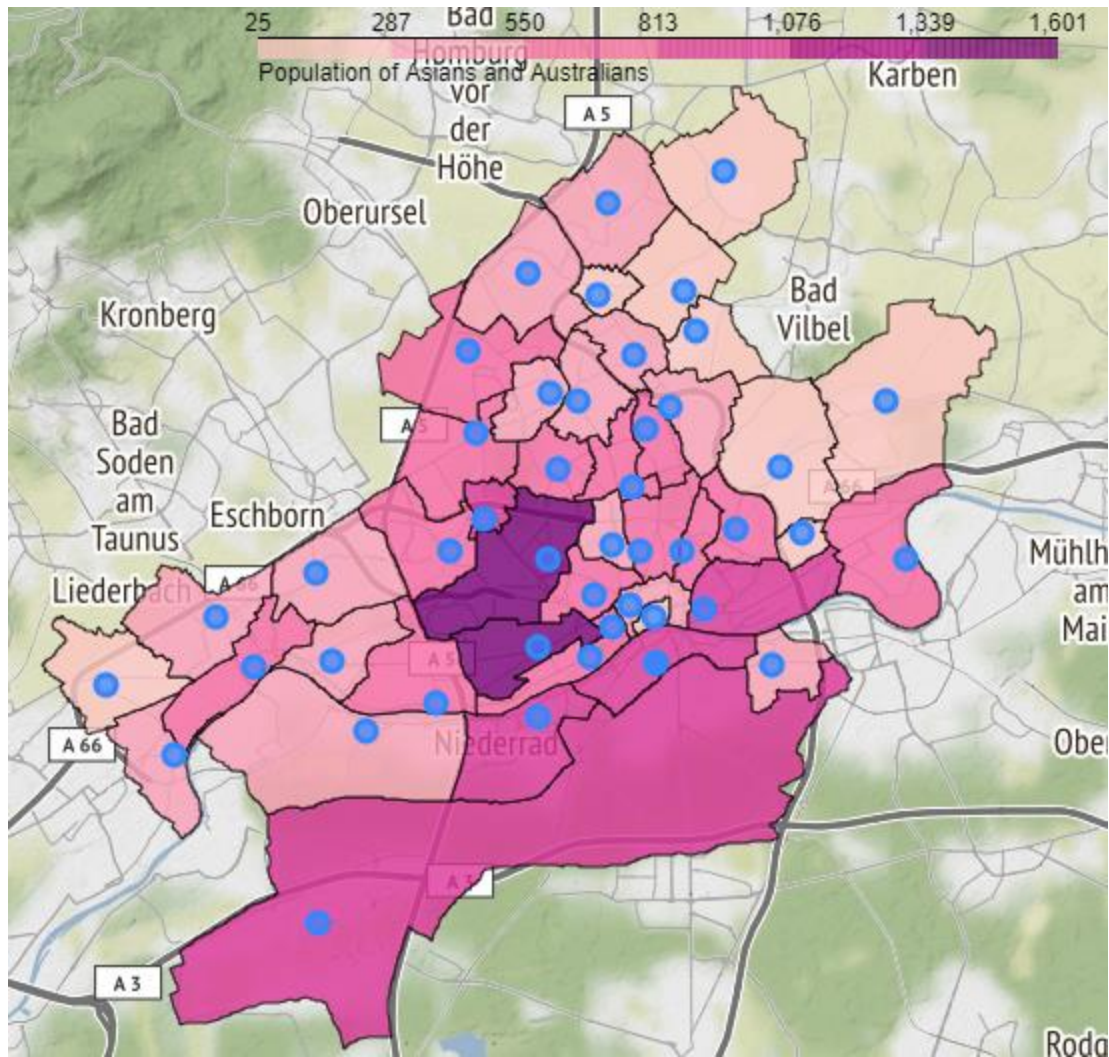
Fig. 10. District-wise distribution of Asian and Australian population – Frankfurt am Main

We can see from the above maps, that the districts of Bockenheim and Gallus have the highest population of Asians and Australians. Out of these, Bockenheim comes under the light green cluster, and Gallus comes under the purple cluster. These 2 neighborhoods are then explored to find out the number of Asian or similar cuisine restaurants in these districts.

	name	categories	lat	lng
11	T.Style	Japanese Restaurant	50.119270	8.644980
12	Lhamo Bistro	Asian Restaurant	50.123901	8.641923
17	Ban Thai	Thai Restaurant	50.121440	8.648446
19	Mai Vien	Asian Restaurant	50.118236	8.644628
21	Hama Sushi	Japanese Restaurant	50.119954	8.648582
34	Mangetsu	Japanese Restaurant	50.115902	8.646909
43	Wayang Indonesische Spezialitäten	Asian Restaurant	50.123712	8.641863

Fig. 11. Asian or similar cuisine restaurants in Bockenheim

We can see that there are 7 Asian or similar cuisine restaurants in Bockenheim. This indicates that as expected from the demographics, there is good demand for Asian food in this district. But so many restaurants also means more competition.

	name	categories	lat	lng
21	Konamon	Japanese Restaurant	50.103886	8.636540
35	Mr. Lee	Korean Restaurant	50.102042	8.660384
38	Coa	Asian Restaurant	50.108350	8.654117
47	Mangetsu	Japanese Restaurant	50.115902	8.646909
49	Mei Mei Chinapoint	Asian Restaurant	50.102821	8.641377

Fig.12. Asian or similar cuisine restaurants in Gallus

Gallus also has a good number of Asian or similar restaurants, and has a high demand for Asian cuisine as expected. But even here, since the number is high, it means more competition. But, being in the purple cluster, we know that there are more hotels in this area as well, so that is a plus point. Nevertheless we shall explore another district nearby which is in the purple cluster and also has a sizeable Asian population - Niederrad.

	name	categories	lat	lng
36	Bambus Haus	Asian Restaurant	50.086535	8.642155

Fig.13. Asian or similar cuisine restaurants in Niederrad

We can see that there is only 1 Asian restaurant in Niederrad. Next, we check the number of hotels in the district.

	name	categories	lat	lng
32	Dorint Hotel Frankfurt Niederrad	Hotel	50.085524	8.632752
33	INNSIDE Frankfurt Niederrad	Hotel	50.080489	8.628484
35	Hotel NH Frankfurt Niederrad	Hotel	50.084805	8.629488

Fig.14. Hotels in Niederrad

There are 3 hotels in Niederrad. 3 hotels and only 1 Asian restaurant in a district with a sizeable population of Asians. It is in close proximity to the city center as well. Therefore, it seems that Niederrad is also a good prospect for opening an Asian restaurant.

4. Results and Discussion

By clustering the districts in Frankfurt and subsequently analysing the district-wise demographics of the city, and then merging the two findings, we could arrive at 3 prospective neighborhoods that would be ideal for opening an Asian restaurant in the city.

1. Bockenheim:

Bockenheim falls in the light green cluster and is very close to the city center. It has 7 Asian restaurants which shows that there is a lot of demand for Asian cuisine in the area. It also has the highest population of Asians in the city at 1586. But there is a lot of competition too. Hence, it is a good option for opening an Asian restaurant provided that the client is ready to accept the competition and establish themselves.

2. Gallus:

Gallus is in the purple cluster containing a greater number of hotels. It is not far from the city center and has 5 Asian restaurants indicating that there is demand here as well. It has the second highest population of Asians in the city at 1512. Five restaurants would be a considerable amount of competition but this district will also see a lot of tourists, thus indicating more customers. Hence, this seems like a better option than Bockenheim for opening an Asian restaurant owing to lesser

competition, similar Asian population and more prospective customers in the form of tourists.

3. Niederrad:

Niederrad is also in the purple cluster having more hotels. It is also not far from the city center, but has only 1 Asian restaurant - much less than both Bockenheim and Gallus. Niederrad also has a sizeable Asian population at 929, although a bit less than the other 2 districts in contention. Since it is in the purple cluster, we can expect more tourists in this district. We see that there are 3 hotels in the area. This translates to more prospective customers. Hence, this also seems like a good alternative to Gallus owing to much lesser competition, proximity to the city center and more tourists.

5. Conclusion

The neighborhoods in Frankfurt am Main were clustered and displayed on a map containing the results. The demographics were studied and based on the findings, 3 districts were found to be ideal as a solution to the Business problem of opening an Asian restaurant. The client can choose any of the 3 neighborhoods to open an Asian restaurant, based on their preferences, confidence and affinity to risk-taking.

6. References:

[1]. Street Directory of the city of Frankfurt amMain: https://offenedaten.frankfurt.de/dataset/strassenverzeichnis-der-stadt-frankfurt-am-main

[2]. Foursquare API: https://developer.foursquare.com

[3]. Demographics of Frankfurt am Main Neighborhoods : https://offenedaten.frankfurt.de/dataset/stadtteilprofile-bevoelkerung