22.7.2021

Battle of the Boroughs:

Bar segmentation of the Frankfurt's Districts

Anastasia Ragulskaya

COURSERA. APPLIED DATA SCIENCE CAPSTONE PROJECT

Introduction / Business Problem

Beer is a major part of German culture. In 2012, Germany ranked third in Europe in terms of per-capita beer consumption, behind the Czech Republic and Austria. Germany is also famous for its beer festivals (e.g. Octoberfest in Munchen or Cannstatter Volksfest in Stuttgart). Thus, taking a glass of beer in the bar the evening seems to be a popular activity in Germany.

Goal: For this project, I decided to investigate the "bar"-business in Frankfurt am Main and find top-5 districts where it may be worth making a new bar.

Frankfurt is the most populous city in the German state of Hesse. Frankfurt is a global hub for commerce, culture, education, tourism and transportation, and rated as an "alpha world city" according to GaWC. Sounds like a good candidate for the research!
The hospitality industry may be interested in the results of this project

Data

For this research we need the following data:

1. Frankfurt districts (Stadtteile)
2. Their longitude and latitude
3. Their geometry (optional)
4. Information about the population density for each district
5. Information about the bars in Frankfurt (amount, longitude, latitude)

I will use the following services to extract the data:

1. Parsing Wikipedia to get Frankfurt districts and information about the area and population which is used to calculated the population density ("https://de.wikipedia.org/wiki/Liste_der_Stadtteile_von_Frankfurt_am_Main").
2. Python Geocoder is used to get the latitude and longitude for each district. This coordinates will be later used in Foursquare API search/expolore query to find bars for each district.
3. Geometry of the districts will be downloaded as the .geojson file (https://offenedaten.frankfurt.de/dataset/85b38876-729c-4a78-910c-a52d5c6df8d2/resource/84dff094-ab75-431f-8c64-39606672f1da/download/ffmstadtteilewahlen.geojson)

4. Foursquare search/explore query API is used to find bars and their coordinates for each neighbourhood. The categoryId='4bf58dd8d48988d116941735' ('bars')

Plan: Cluster Frankfurt districts based on the amount of bars and population density. Based on the clusterization, define districts-candidates for bars

Methology

In this project we will direct our efforts on detecting areas of Frankfurt that have small amount of bars and a high population density.

In first step we collect the required data: location of every bar in Frankfurt and population density for each district.

Second step in our analysis is clusterisation of the Frankfurt district based on k-means clustering.

In the final step we will take the most promising cluster and from it top-5 district based on the amount of bars which already exists.

Analysis steps:

1. Import Frankfurt district data from the Wiki web page
2. Import geometry data of Frankfurt districts from the offenedaten.frankfurt.de
3. Add latitude and longitude to districts using Geocoder Python package
4. Some coordinates were found to be the same. So, these districts were merged
5. Calculation of the population density based on the information of the population and the area of the final districts
6. Get Frankfurt bars information using Foursquare venues/explore API request using neighbourhood latitude and longitude coordinates. The radius was not specify.
7. Remove duplicates from the resulting dataframe by calculating distance from the bar to the coordinates of the district. Then, the

dataframe was assigned to the closest district and the other records for the same bar were deleted.

8. Amount of bars for each district is calculated
9. Combining two separate dataframes (geographic and demographic dataframe + the information about the amount of pubs in districts).

Analysis resulted in 208 bars in Frankfurt.

Machine Learning (ML):

To cluster the districts based on the above information the “k-means” cluster analysis model was used. ML steps:

1. Data preprocessing and normalization
2. Finding the optimal k using the ‘elbow method’
3. Running the model with the optimal k
4. Interpretation of the resulted cluster datasets
5. Taking the top-5 districts based with the smallest amount of bars from the group with the average population density
6. Visualization of the clustering results and the best matches.

Results

The final dataframe after the clustering is shown below.

Neighborhood	Cluster	Bars	Population density (Population/km^2)
Heddernheim	1	2	6791.169451
Ostend	1	11	5345.794393
Niederrad	1	5	4415.904637
Nied	1	1	5335.760518
Innenstadt	1	40	4393.024816
Hausen	1	2	5898.073836
Preungesheim	1	1	4286.684783
Ginnheim	1	2	6243.413729
Griesheim	1	2	4621.372549
Bonames	1	1	4701.166181

Bahnhofsviertel	1	31	6832.103321
Bockenheim	1	9	5247.167227
Rödelheim	1	2	4131.545064
Eschersheim	1	2	4741.027228
Eckenheim	1	2	6390.860692
Gutleutviertel	2	3	3786.830357
Sossenheim	2	1	2732.725122
Unterliederbach	2	1	2848.530144
Höchst	2	3	3502.936698
Fechenheim	2	1	2554.839633
Nieder-Eschbach	2	1	1805.608066
Bergen-Enkheim	2	1	1427.505754
Niederursel	2	1	2238.075990
Schwanheim	2	4	1392.811210
Seckbach	2	2	1309.288661
Frankfurter Berg	2	1	3427.083333
Altstadt	3	26	8409.090909
Nordend-Ost & Nordend-West	3	7	9991.935484
Gallus	3	5	9300.863405
Bornheim	3	5	11041.277818
Sachsenhausen-Nord	3	28	7750.649351
Westend-Nord & Westend-Süd	3	5	7736.483781

Thus, the data showed 3 groups of districts:

High population density (>7000) and average-high amount of bars (>5).

Low population density (<3502) and low amount of bars.

Average population density (4000-7000) and mixed amount of bars.

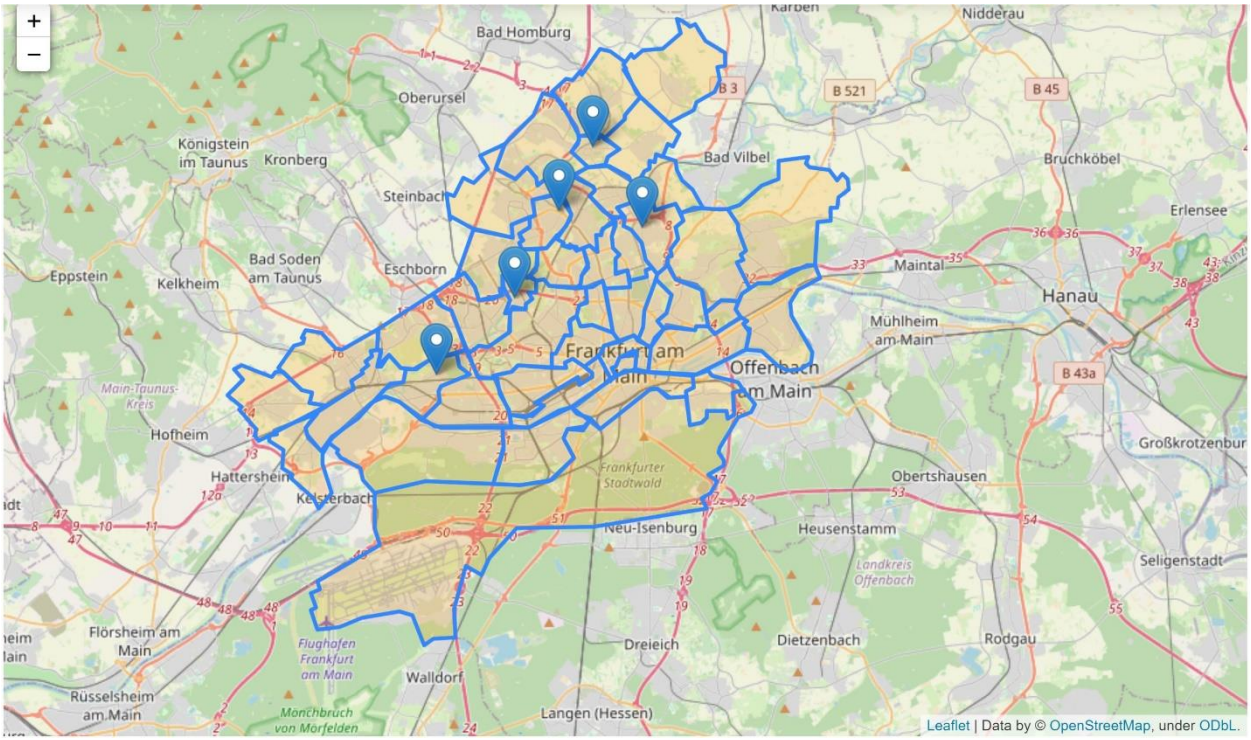
This last cluster is the one we are interested in. The population density is enough to make the bar popular, however, not for all districts there is a lot of the already existing bars. So we need to extract the ones (let's say top-5) with the smalles amount of bars. Thus, we will find the districts where the bar may be popular, but will not have a lot of the high competition.

Keep in mind: restarting the code may result in the different correlation between the exact number of the cluster and the description shown above.

Extracting top-5 bar districts from the interested group results in the following dataframe.

Neighborhood	Cluster	Bars	Population density (Population/km^2)
Heddernheim	1	2	6791.169451
Ostend	1	11	5345.794393
Niederrad	1	5	4415.904637
Nied	1	1	5335.760518
Innenstadt	1	40	4393.024816

Visualisation of the found best-matches on the Frankfurt map with the polygon-type district marks is shown below.



Discussion

Our analysis shows that surprisingly there is a relative small amount of bars in Frankfurt: only 208 in the whole area, baring in mind that Frankfurt is one

of the largest cities in Germany. The districts can be devided in 3 categories.

- High population density (>7000) and average-high amount of bars (>5).
- Low population density (<3502) and low amount of bars.
- Average population density (4000-7000) and mixed amount of bars.

Investigation of the map shows that the first category disrticts are located near the city center. Thus, the amount of pubs is high due to the tourist attraction to this location and other advantages of the city center.

The second custer (low density and low amount of bars) is located in the city perephery.

The final cluster is the most promising. The population density is enough to make the bar popular, however, not for all districts there is a lot of the already existing bars.

Exctracting top-5 districts with the smallest amount of bars from this cluster allows us to find perfect candidated for the consideration for the new bars. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Conclusions¶

Purpose of this project was to identify Frankfurt areas with low number of bars in order to aid stakeholders in narrowing down the search for optimal location for a new bar. By calculating amount of bars from Foursquare data and merging this information with the density population data we prepared the dataset for further consideration. Clustering of those locations was then performed in order to create major zones of interest. Thus, we found the "Preungesheim", "Nied", "Bonames", "Heddernheim" and "Hausen" districts are worth consideration for making a new bar there.

Final decission on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.