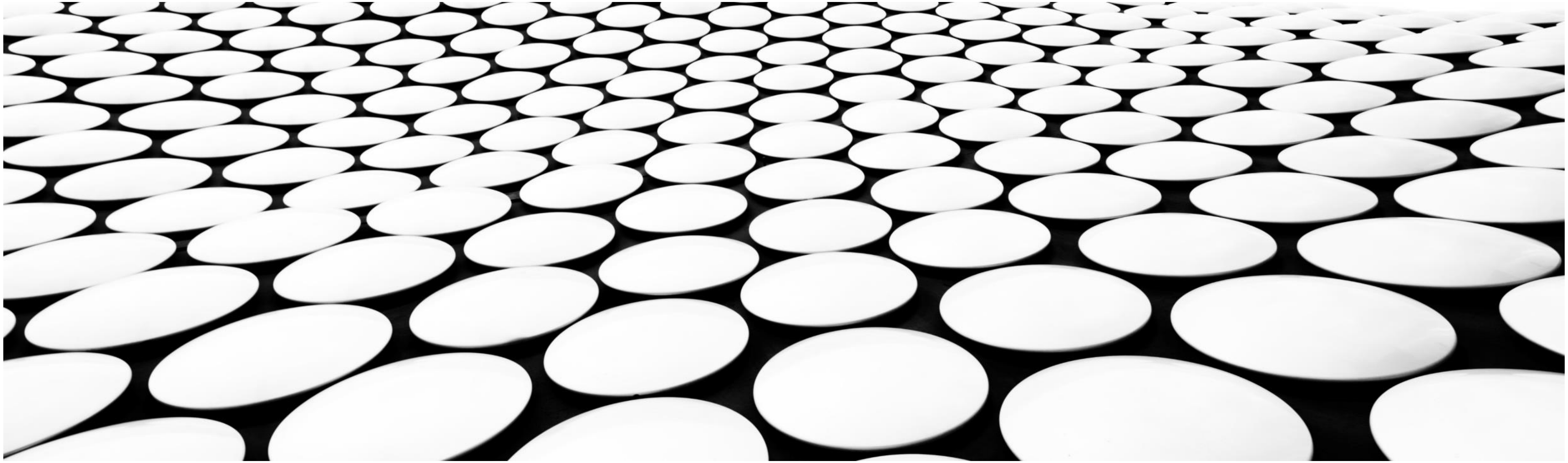
BATTLE OF THE BOROUGHS:
BAR SEGMENTATION OF THE FRANKFURT’S DISTRICTS

ANASTASIA RAGULSKAYA

COURSERA. APPLIED DATA SCIENCE CAPSTONE PROJECT



INTRODUCTION / BUSINESS PROBLEM

- Germany – third in Europe of per-capita consumption
- Frankfurt – most populous city in the German state of Hesse. One of the biggest cities of Germany.

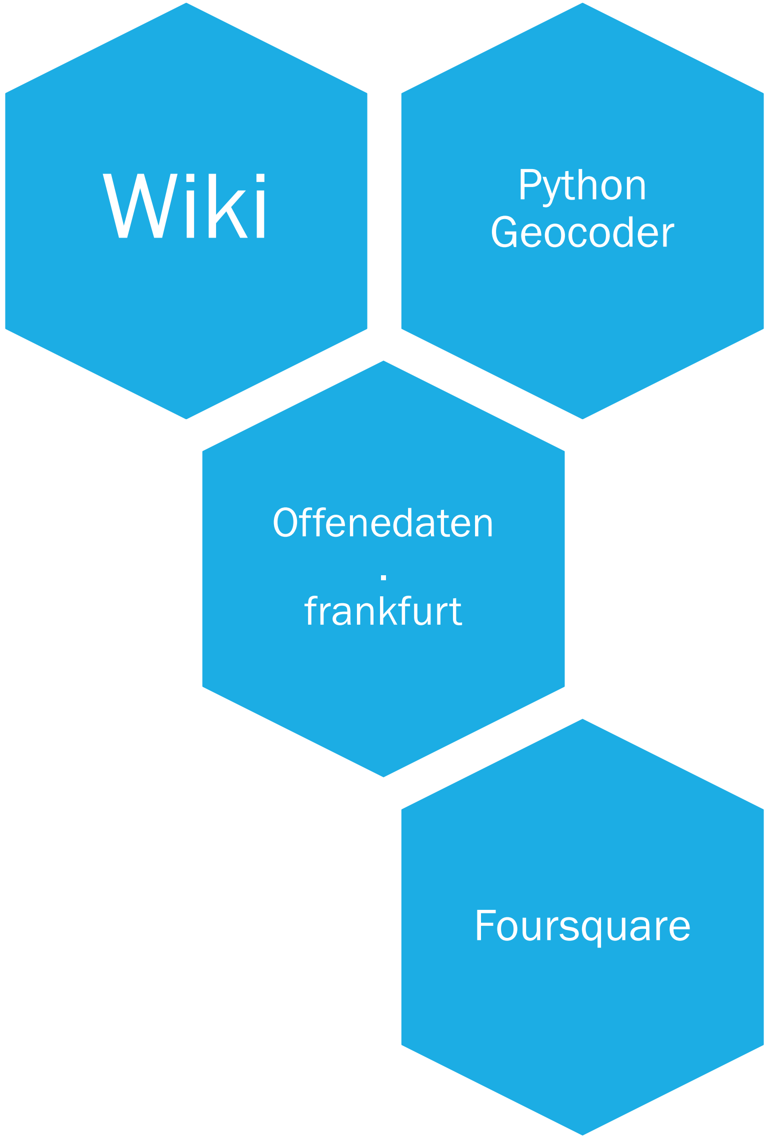
Goal: investigate the "bar"-business in Frankfurt am Main and find top-5 districts for new bar

The hospitality industry may be interested in the results of this project

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)

SERVICES AND SOURCES



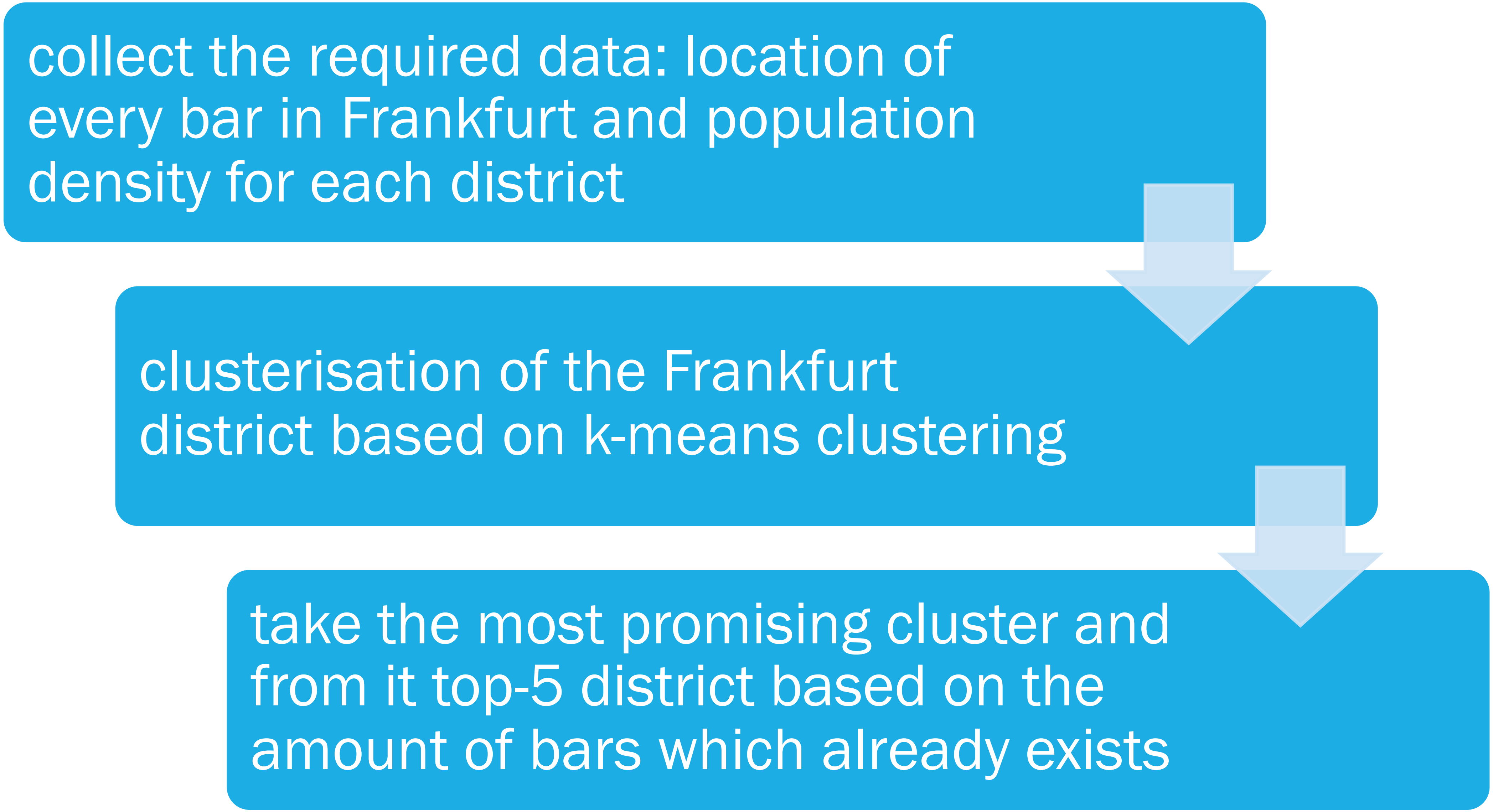
Wiki

Python Geocoder

Offenedaten.frankfurt

Foursquare

METHOLOGY



collect the required data: location of every bar in Frankfurt and population density for each district

clusterisation of the Frankfurt district based on k-means clustering

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).

MACHINE LEARNING 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 Finding best matches

6 Visualization of the clustering results and the best matches

RESULTS AND DISCUSSION

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.

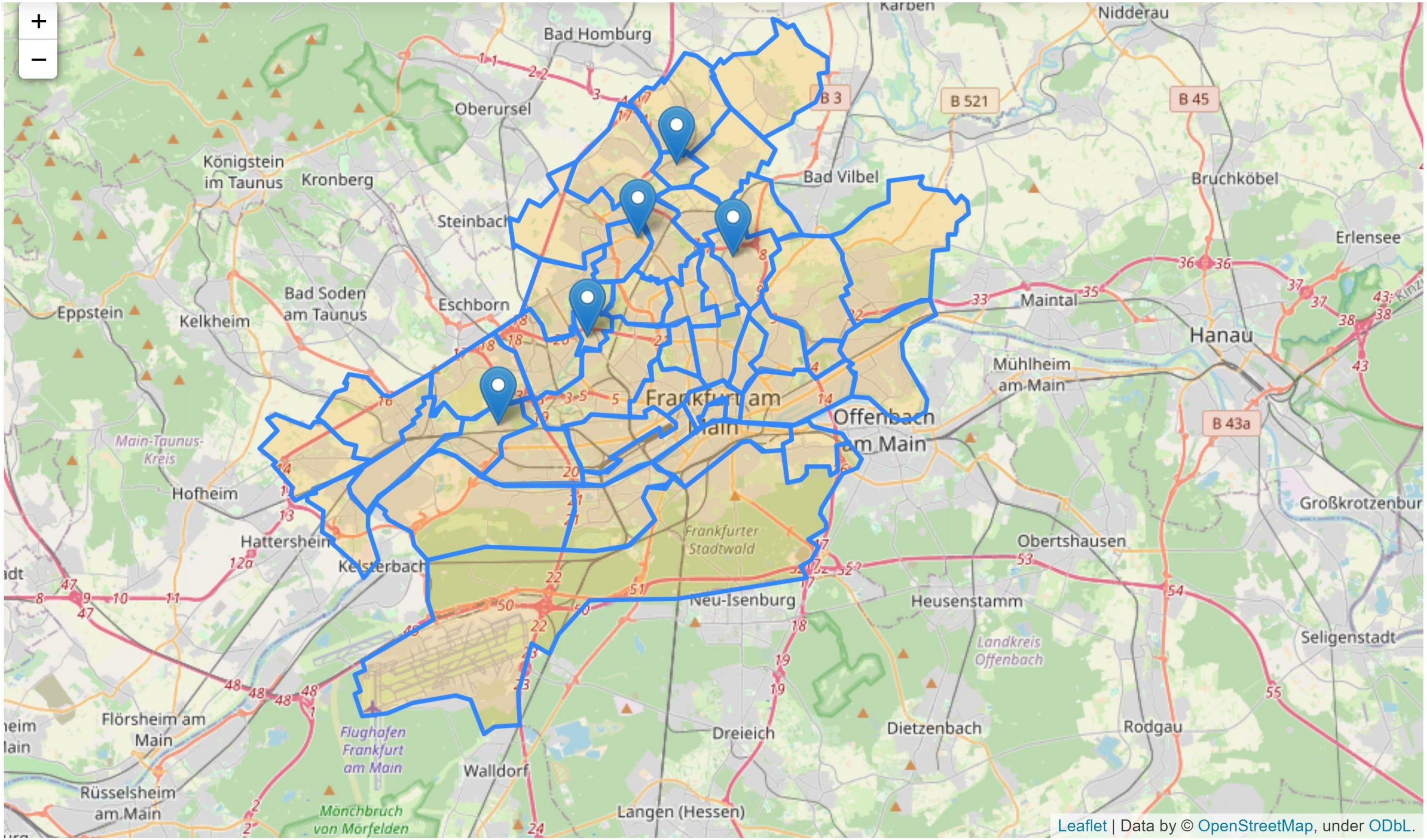
We need to extract the ones (let's say top-5) with the smalles amount of bars. Thus, we will find the districts where the bar may be popular, but will not have a lot of the high competition.

BEST-MATCHES

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



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.

- Extracted the data by parsing Wiki, loading JSON files from the web,using Python Geocoder and Foursquare.
- Clustered the districts of the Frankfurt based on the amount of pub and population density
- Best matches: "Preungesheim", "Nied", "Bonames", "Heddernheim" and "Hausen" districts

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone