Coursera / IBM Applied Datascience

Capstone Project

„The Battle of Neighborhoods“

Week 4/5

New Business opportunities:

Which City in Germany is Best Suited for a new Opening of a Vegetarian Restaurant?

A Spatial Suitability Analysis of the 700 Largest Cities in Germany

April, 19, 2020

Written by: Hubert Kohler, Germany

Table of Content

Spatial Suitability analysis for the opening and operation of a vegetarian restaurant in Germany

PROJECT WORK FOR THE COURSERA/IBM CAPSTONE PROJECT "APPLIED DATASCIENCE"

- **I MERGED WEEK 4 AND 5 TOGETHER AS THE TOPICS OVERLAP THE CHAPTERS FOR WEEK 4 CAN BE FOUND UP TO PAGE 9**
- **THIS DOC IS STORED AT GITHUB: HTTPS://GITHUB.COM/HUKO-DATASCIENTIST/COURSERA_CAPSTONE/BLOB/MASTER/COURSERA-IBM-APPLIED_DATASCIENCE_WEEK5%20EN.PDF**
- **THE JUPYTER NOTEBOOK WITH THE CODE CAN BE FOUND AT GITHUB:HTTPS://GITHUB.COM/HUKO-DATASCIENTIST/COURSERA_CAPSTONE/BLOB/MASTER/CAPSTONE_WEEK4AND5_PROJECT.IPYNB**
- **BLOGPOST:** HTTPS://GEOSCANALYTICS.EU/2020/04/22/HOT-TREND-OPENING-A-VEGETARIAN-RESTAURANT-IN-GERMANY-CASH-COW-OR-HIGH-RISK/

Aim of this project work and background:

The following question should be answered:

"In Which of the 700 largest Cities in Germany is the opening of a vegetarian restaurant recommended"

In general, the calculation model used here answers the following questions:

"What is the Relative Risk of Success of a Vegetarian Restaurant in a City in Germany?

A study is to be carried out to assess the relative risks of opening and operating a vegetarian restaurant in Germany at a certain location.
The results are to be divided into so-called risk classes and presented per city.
The relative success risk is a qualitative estimate of the individual risk classes.
This study is limited to the 700 largest cities in Germany with a population of 20000 or more.
No distinction should be made between the individual vegetarian intensity levels. Everything that can be considered vegetarian is considered in this paper without indicating or considering the degrees or intensity.

Initial situation / description of the problem and its background:

Vegetarian gastronomy is still relatively rare in Central Europe and especially in Germany. Most people so far prefer a gastronomy in which meat dishes are offered. Especially in rural areas vegetarian gastronomy is partly not to be found at all.
This means that up to now there is little experience to assess the chances and risks of opening a vegetarian restaurant that offers only vegetarian dishes.
The spatial conditions in the surrounding area play an absolutely central role in the assessment.
Due to the limited freely accessible data, only a small part of the desired parameters can be used as input for the calculation model.

"Decline in meat consumption"

„The 512 million EU citizens account for 6.8 percent of the world's population, but are responsible for 16 percent of the world's total meat consumption . The current per capita amount of meat eaten by Europeans stood at 69.3 kg in 2018 but that figure is expected to fall to 68.6 kg in 2030, according to the European Union agricultural outlook for 2018-2030 report , though dairy product consumption will climb." [1]

The motivation to eat a vegetarian diet is manifold.

While some focus on animal welfare and protection, others are concerned about climate or health. In addition, religion plays an important role, especially in Asia.

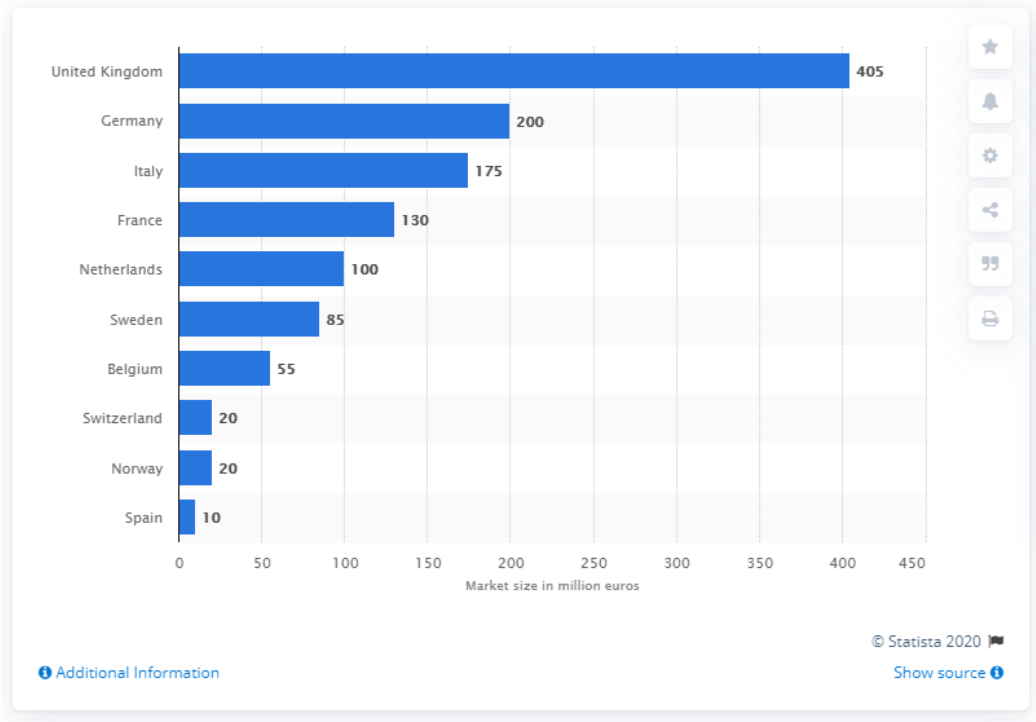


Figure 1: Market sizes for meat substitutes 2018 [2]. UK is currently the largest market, followed by Germany

The market for meat substitutes can be seen as a clear trend towards a vegetarian diet, because meat substitutes are supposed to imitate the taste and consistency of meat. They are therefore aimed primarily at people who originally ate meat, but in future want to consume less meat or even eat a meat-free diet.
Figure 1 and Figure 2 show the current market situation in Europe and worldwide.
Currently, the United Kingdom is the country with the largest market for meat substitutes, followed by Germany.

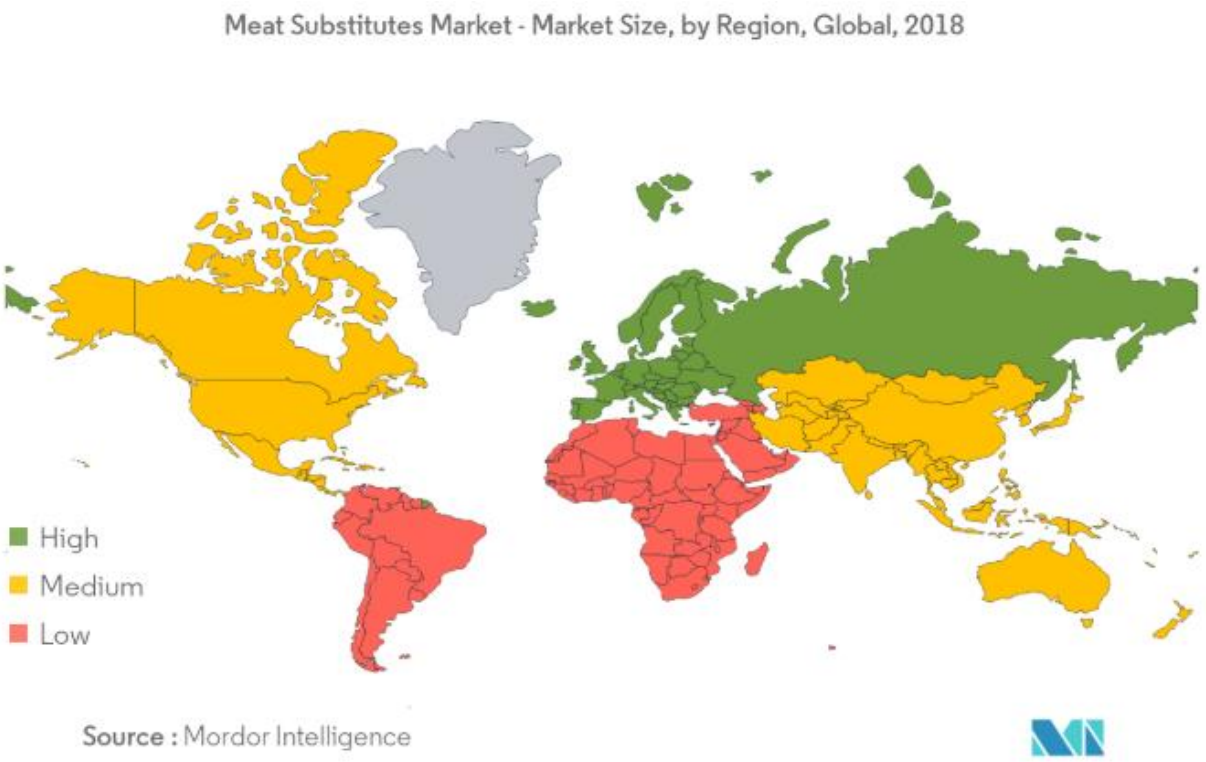


Figure 2: Distribution of the market for meat substitutes in 2018 [3]. The greatest market potential lies in Europe and Russia

From a global perspective, Europe and Russia seem to be the leader in meat substitution.
As this trend towards a meat-free diet will continue, vegetarian gastronomy will also become a clear focus of potential investors.

Who is interested in this study, who could be the potential client?

- **Investors who follow the new vegetarian trend and want to realize new business ideas.**
- **Businessmen who have been running traditional restaurants or restaurant chains but are looking for new business opportunities.**
- **Local politicians who want to increase cultural diversity in the cities and set new trends.**
- **Operators of ecologically oriented businesses who wish to expand their portfolio on a horizontal level.**

Which data sources are used and how:

In principle, data is used which is made available free of charge on the Internet. The following data is used for this work:

- **Objects of investigation / Cities:**
  The 700 largest cities, with a minimum population of 20,000 inhabitants.
  ➔ Data source Wikipedia:
  https://de.wikipedia.org/wiki/Liste_der_Gro%C3%9F-_und_Mittelst%C3%A4dte_in_DeutschlandMethode/Modul: Pandas read_html Function
- **Existing vegetarian Restaurants:**
  ➔ Data source FourSquare:
  https://developer.foursquare.com/docs/build-with-foursquare/categories
  Category: '4bf58dd8d48988d1d3941735' => Vegetarien/Vegan Restaurants
  Methode/Modul: requests.get(url).json()
- **Existing organic food stores:**
  ➔ Data source FourSquare:
  https://developer.foursquare.com/docs/build-with-foursquare/categories
  Category: '52f2ab2ebcbc57f1066b8b45' => Organic Grocery Stores
  Method/module: requests.get(url). json()
- **Coordinates of the cities (WGS84 / GPS data):**
  Method/Module: Geolocator.Nominatim Openstreetmap➔
- **Vector polygon of the borders of Germany:**
  ➔ Data source: https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/verwaltungsgebiete.html

Data and methods

Data used:

- ⇨ List with the 700 largest cities in Germany and their locations (Centroid, WGS84)
- ⇨ Number of current vegetarian restaurants per city (V_c)
- ⇨ Number of current organic food shops per city (B_c)
- ⇨ Number of inhabitants per city (P_c)
- ⇨ Foursquare search radius 3000m maximum (r_{max})

Calculation method:

- In a simplified way, the number of existing organic food stores is used to draw conclusions about how well suited a city is for a vegetarian restaurant. This means that we assume that a city with many organic food shops is suitable for opening and running vegetarian restaurants.
- The indicator ("Risk Score") for the suitability of a city for a new vegetarian restaurant is represented by the ratio $\left(\frac{V_c}{B_c}\right)$:
  - ⇨ Low value (minimum 0) : recommended city to open a vegetarian restaurant
  - ⇨ The higher the value, the higher the risk and therefore less recommendable.
- As proof of the suitability of the ratio $\frac{V_c}{B_c}$ the Pearson correlation coefficient and R Square are used. The linear regression analysis also serves this purpose. A subdivision into training and test data does not make sense here, because the linear regression is not intended to make predictions in this case.
- The risk classes are classified using the K-MEANS clustering procedure. 5 clustered risk classes are created, taking into account V_c , B_c and also the population size (P_c).
- Additionally, DBSCAN Clustering results are compared to K-Means
- The application of K-Means can lead to an overlap of risk scores per separated class.
- A further risk class "6" is added manually. This contains cities that do not contain any organic food stores and whose indicator $\left(\frac{V_c}{B_c}\right)$ therefore cannot not be defined.
- The risk classes do not give absolute values but only relations to each other.

Correlation between number of existing Vegetarian Restaurants and Organic Grocery Stores

- Pearson Standard correlation: 0.9, see also Figure 3
- R Squared: 0.77
- The population as input parameter does show also strong correlation to V_c and B_c

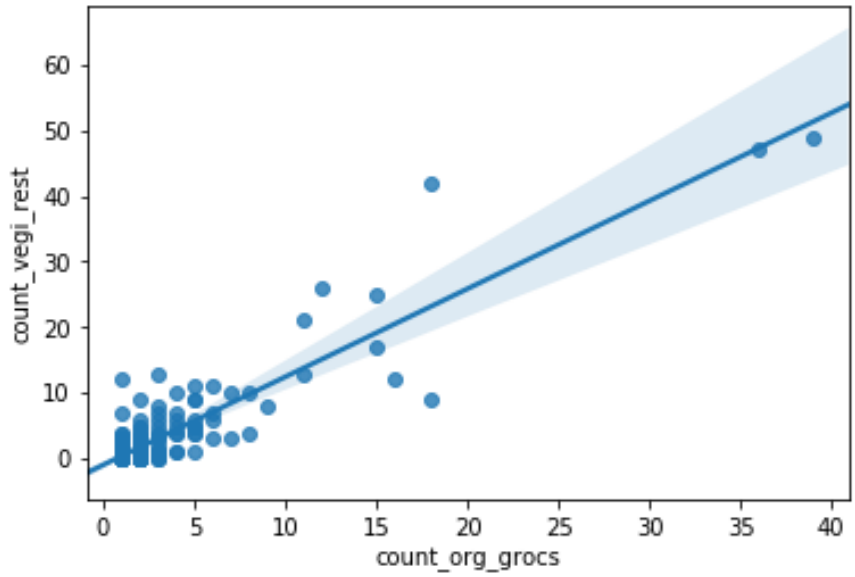


Figure 3: Scatter Plot of number of existing organic grocery stores (x) vs vegetarian restaurants(y)

Applied machine learning algorithms:

- **Linear Regression / Correlation Score.**
  - This is applied to just demonstrate the linear relationship between V_c and B_c
  - Generally spoken the best risk_levels are given by outliers, very high B_c and low V_c

- **K-Means Clustering.**
  - K-means is suitable to cluster unlabeled data (unsupervised)
  - The K-Value was chosen from the results by the Elbow algorithm
  - The input features are limited to $\mathrm{V_c}$, $\mathrm{B_c}$ and also the population size $\mathrm{P_c}$.
  - The produced K-means clusters are assigned to risk levels according the mean Risk Score ratio. The cluster with lowest Risk Score is assigned to Risk Level 0 . Accordingly, the cluster with highest Risk Score is assigned to Risk Level 5.
  - A Boxplot is used to check how the Risk Scores do overlap between the Risk Levels.

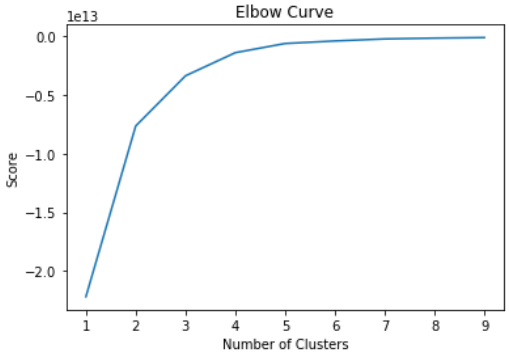


Figure 4: Determination of K for the K-Means. Value 5 is chosen.

- **DBSCAN Clustering**
  - A calculation is run just to compare the results of K-Means to DBSCAN algorithm applying the same Input data as for K-means.

Results

Cities considered

Figure 5 shows an overview of the cities studied.
As shown in the Figure 6, this study covers 60% of the German population.
The highest number of cities but also of population is represented by category 1 (20,000 to 100,000).

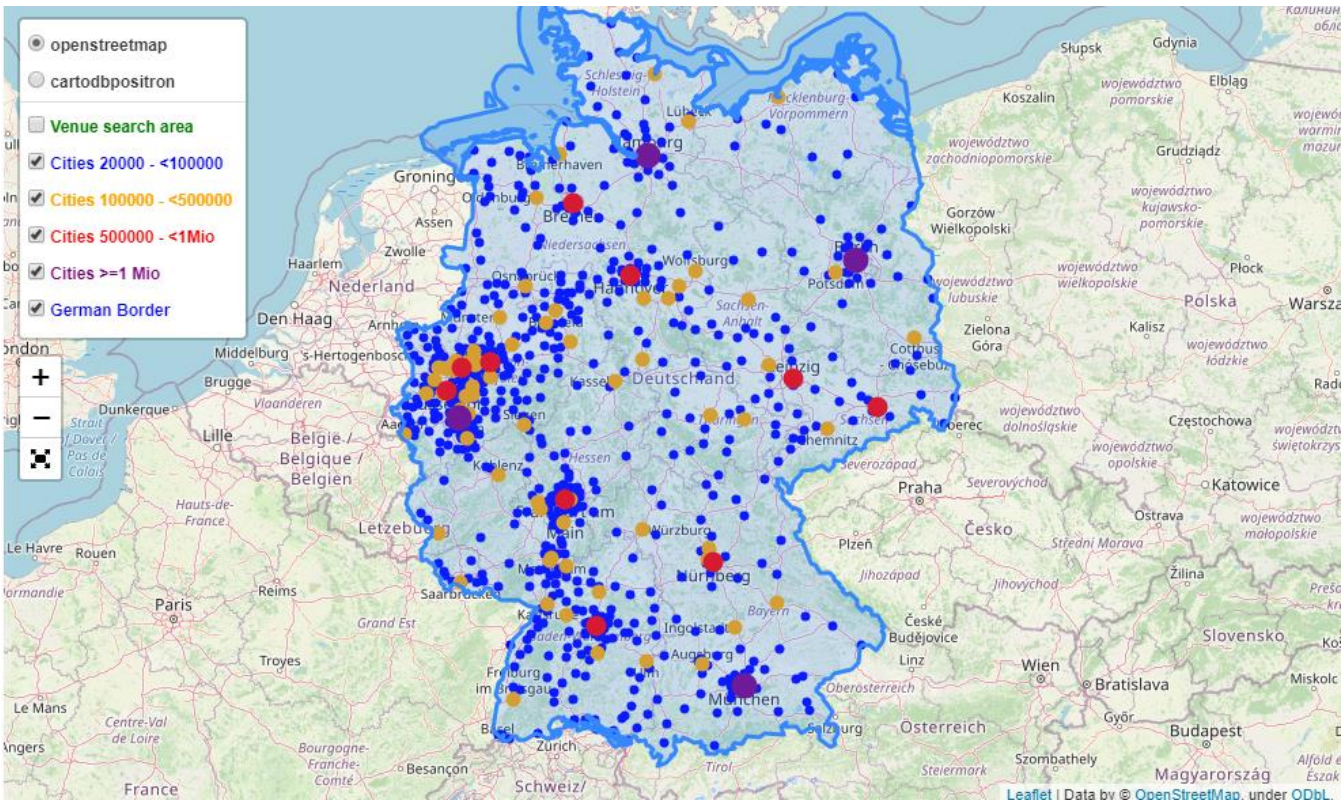


Figure 5: Cities considered with a minimum population of 20000

city size category	Population (2018)	Count
1: > = 20000, <100000	22748652	619
2: >=100000, <500000	12651058	67
3: >=500000, < 1 Mio	5945590	10
4: >= 1 Mio	8043177	4
SUM	49388477 ⇨ 60% of the total population	700

Figure 6: Size distribution of the 700 largest cities by number and population

Figure 5 shows the most dense area is at middle-west ("Ruhrgebiet") and south-west. Berlin as the Capital and biggest City also represents a high density area.

Distribution of the existing vegetarian restaurants:

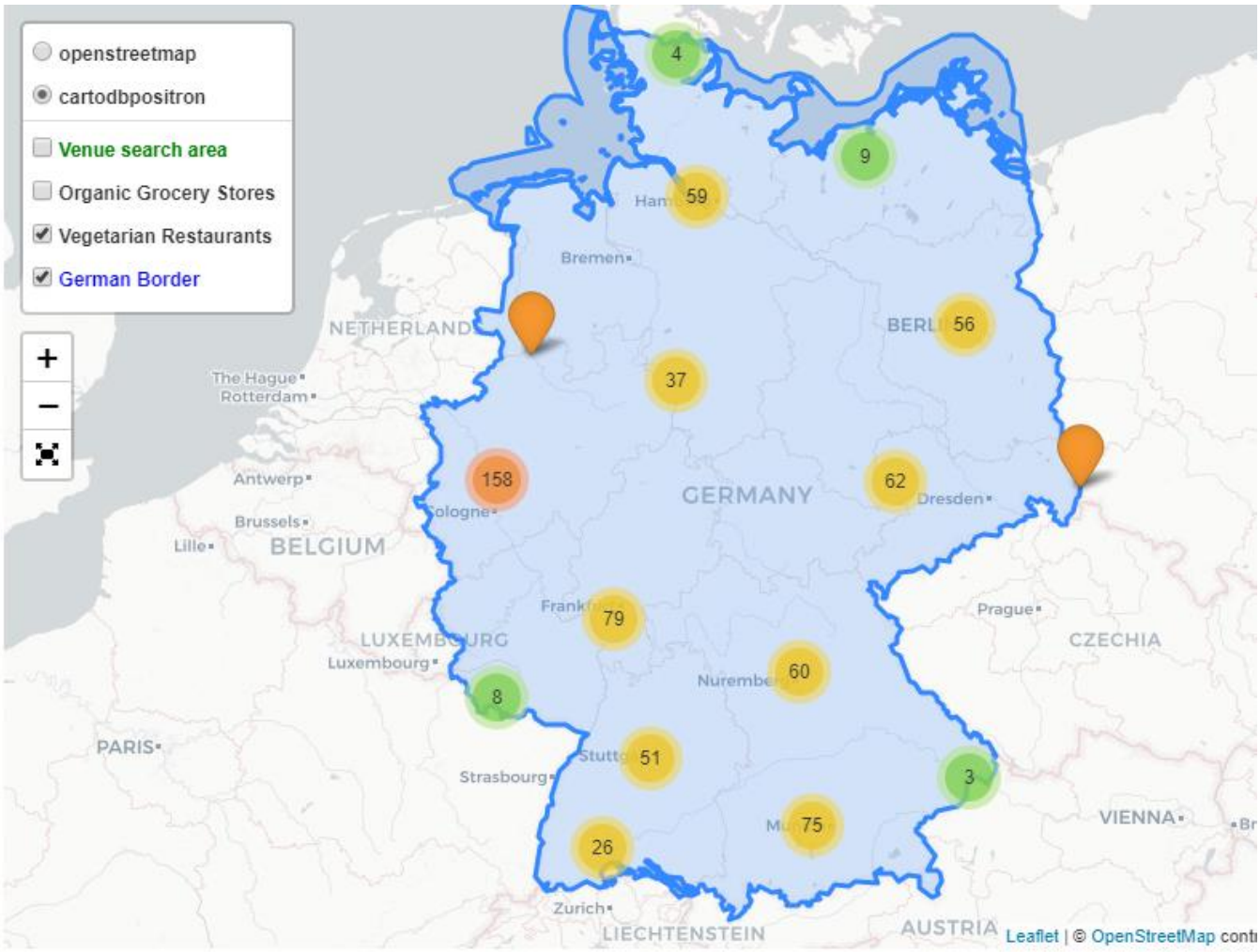


Figure 7: Distribution of existing vegetarian restaurants. The number in each region was automatically summarized by Folium Cluster Markers. The color was also determined by Folium and is dependent on the number

The spatial distribution of the vegetarian restaurants goes hand in hand with the distribution of the dense city areas. Middle-west and south-west operate the most vegetarian restaurants. There is also a hot spot at the area at and around Munich.

Distribution of existing organic food stores

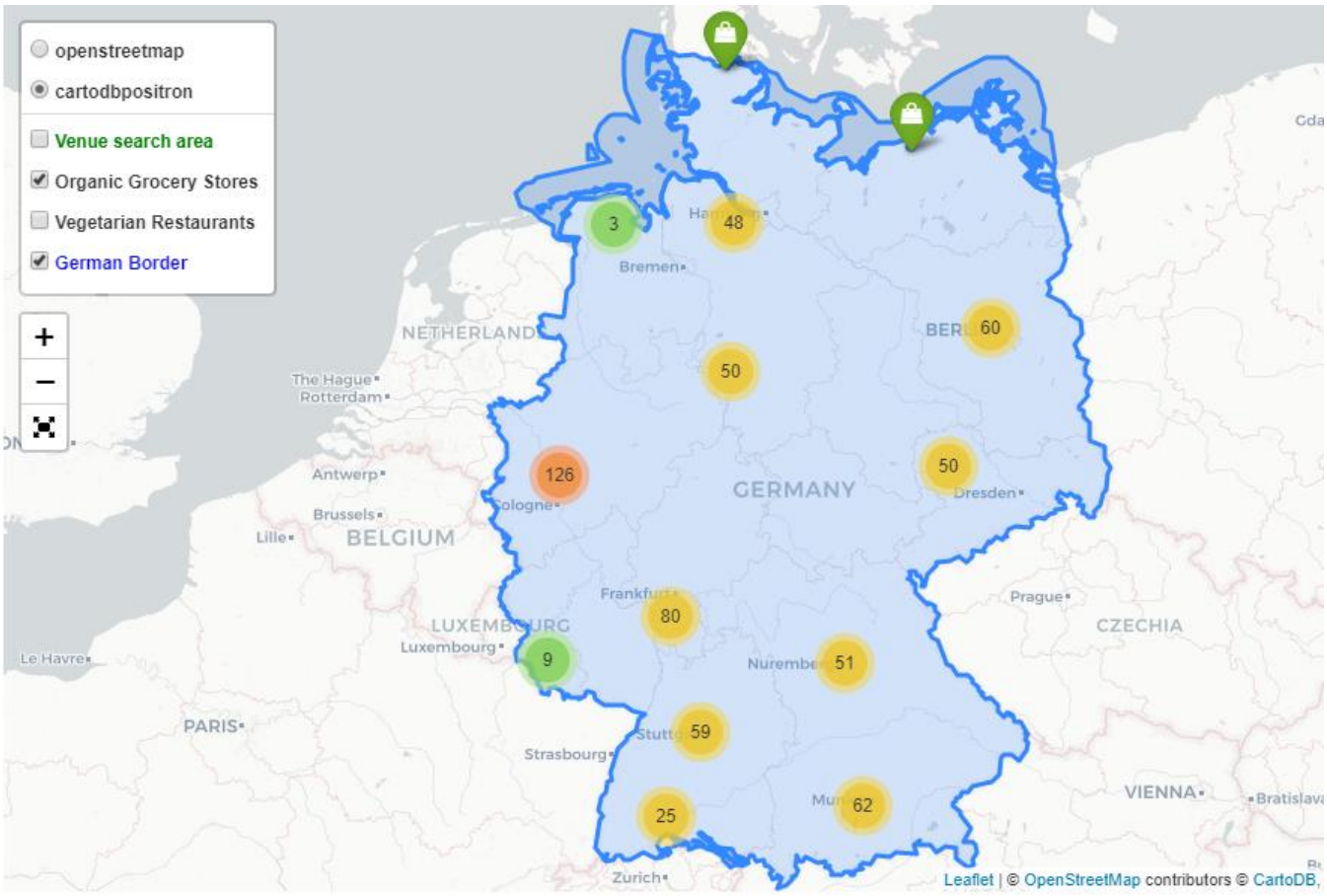


Figure 8: Distribution of existing organic food stores. The number in each region was automatically summarized by Folium Cluster Markers. The color was also determined by Folium and is dependent on the number

The distribution of organic grocery stores show similar behaviour as the distribution of the vegetarian restaurants do.

The Formation of the Risk Classes

As described in the chapter "Data and methods", a risk score is calculated which is divided into 5 clustered classes by the K-Means algorithm applied.
The result of these classes is ordered by their mean Risk Score and assigned to Risk levels. The cluster with the least Risk Score is assigned to Risk Level 1, the cluster with the highest Risk Score is assigned to Risk Level 5, the others are in between.

Assessment of the risk level:
By dividing the risk scores into 5 risk classes, the overlapping of the classes was shown by means of a boxplot. It can be seen very clearly in the Figure 9 that risk classes 1 and 5 do not overlap. The separation of the classes by the K-means algorithm seems to be valid. Thus, cities assigned to risk

class 1 offer the best possibilities for opening a vegetarian restaurant, while risk class 5 shows a very high risk in this respect.

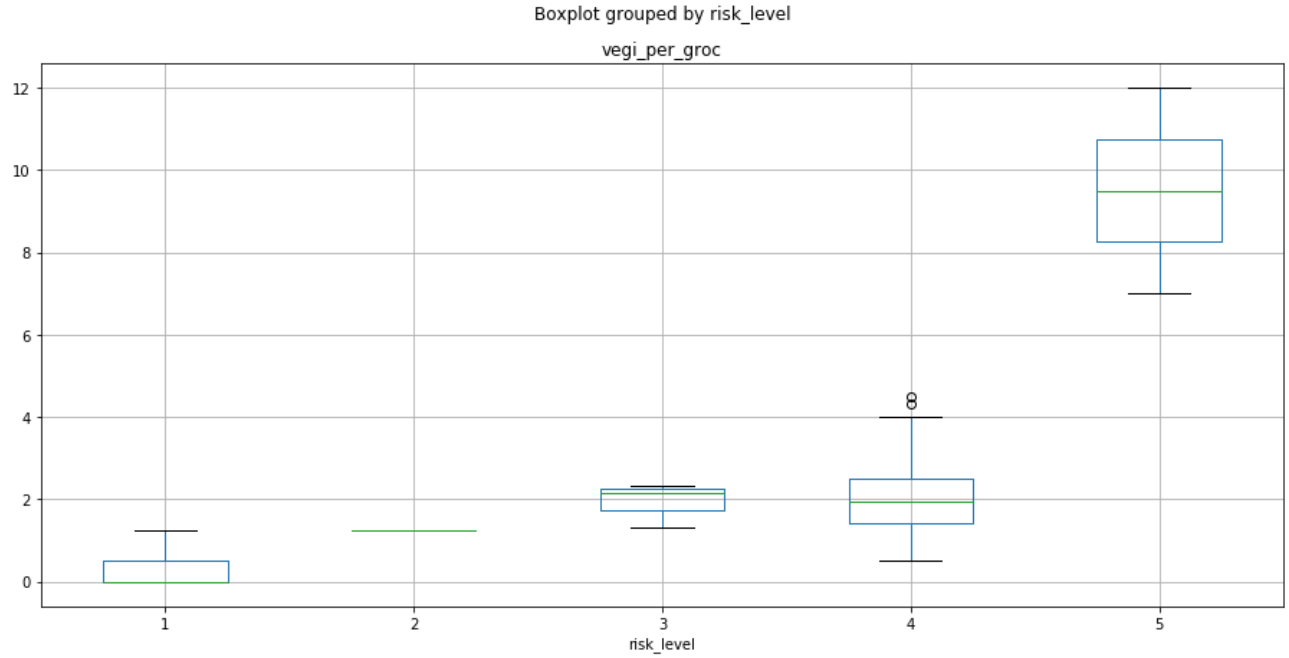


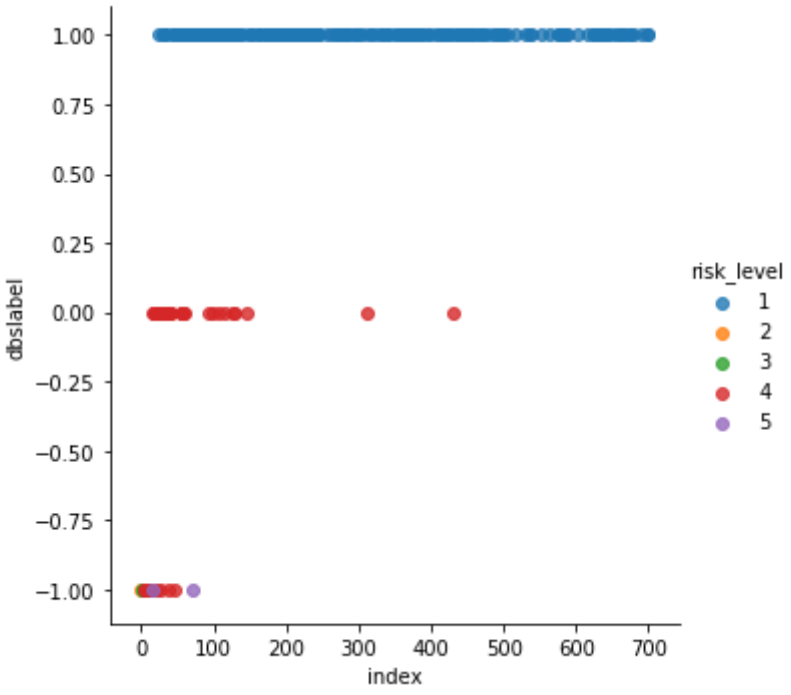
Figure 9: Boxplot of the risk levels vs risk scores to check the overlaps. There is almost no overlap of Risk level 1 (best) and 5 (worst) to the other ones. This shows a good separation job of K-Means

DBSCAN vs K-means:

DBSCAN shows very similar results, however the amount of clusters , including outlier class is less than K-Means. The chart below shows the cities, represented as "index", and the labels assigned by DBSCAN. Risk Level 1 as produced by K-Means is actually assigned to the same cities as the DBSCAN did ("dbslabel" == 1.0).

Except Risk Level 4 all other risk levels are treated as "outliers" (-1) by DBSCAN.

This seems absolutely valid as the amount is very low.



risk_level	dbslabel	index
1	1	194
2	-1	1
3	-1	3
4	-1	14
	0	30
5	-1	2

Final result

The K-means algorithm produced 5 very unevenly distributed clusters , see Figure 13.
However, there is a clear result: Risk Level 1 has a high count of near 28% of the cities.
Additionally, the cities with Risk Level 1 are distributed over the whole country. Areas with high amount of Risk Level 1 cities are located at middle-west and south-west while north-west locations seems offer quite less opportunities.
Furthermore there is a huge number of cities which we cannot rate as there is no grocery store and therefore we cannot determine the risk score indicator (see Figure 11)

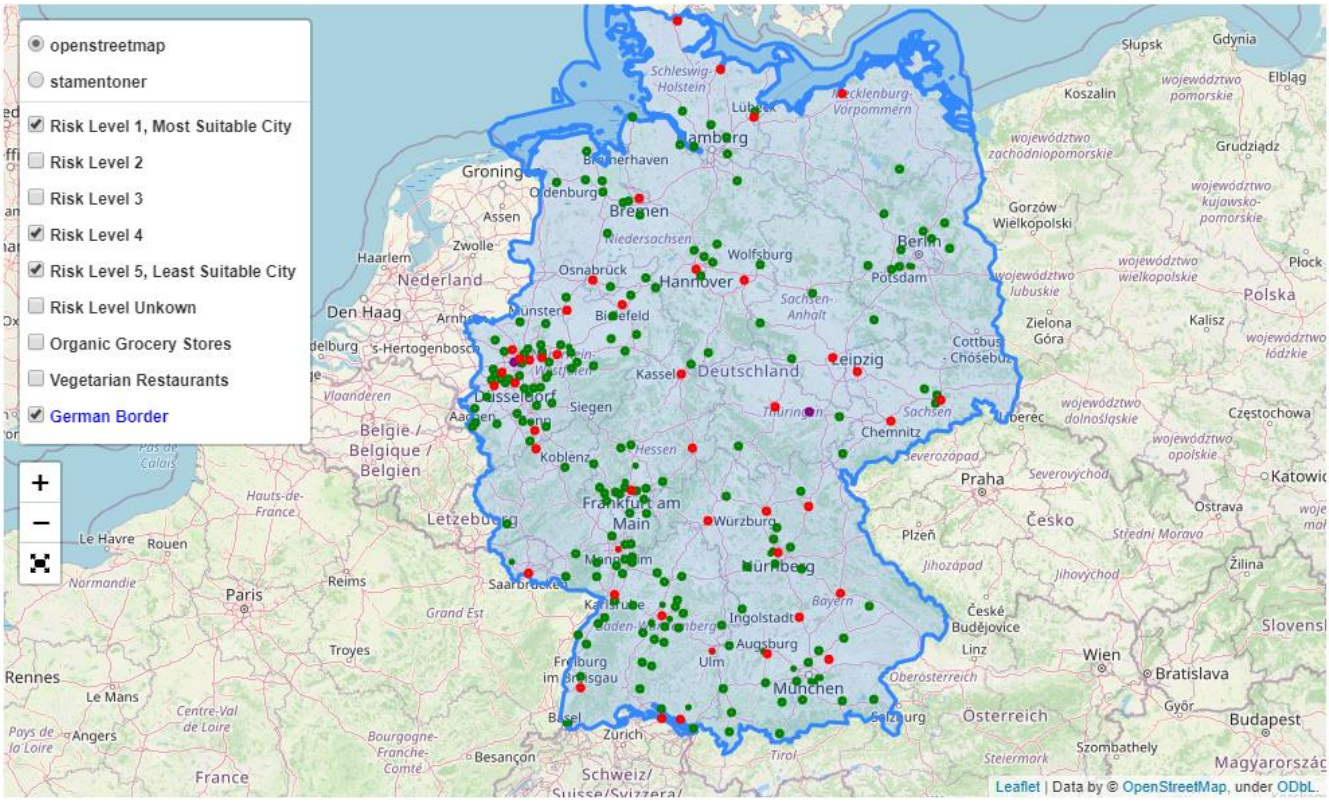


Figure 10: Suitability for opening and running a vegetarian restaurant.
Green: lowest risk class, most suitable.
Red / Violet: high risk, less suitable

Figure 11: Locations with unknown risk

Summary of the final result for the stakeholders

This study provides information on where in Germany there are the best chances or the lowest risks of opening a vegetarian restaurant and operating it successfully.
There is a strict linear relationship between existing organic food stores and the number of vegetarian restaurants.
Therefore, the ratio of the number per city was used to form a risk classification.
The risk classes indicate the relative suitability of a city for a vegetarian restaurant to be successful.
There is a class division from 1-5, plus a manual class 6 where the risk class cannot be determined due to a lack of organic food stores.
The complete report can be made available on demand by Jupyter notebook.
The Top 20 of the most recommended cities is illustrated at Figure 12.
The first candidate in the ranking where to look first is Herne, followed by Wolfsburg and Reutlingen. They are Cities with population above 100,000 and therefore offer higher amount of customers than other with same risk level do.

Furthermore it needs to emphasized, that there is no big City >500000 in the top candidates list with risk_level 0. Those Cities are mostly already oversaturated with vegetarian restaurants. However a deeper analysis would be recommended to get a deeper insight in the big cities, see also chapter "Limitation, possibilities for improvement:"

Top 20 best suited cities with low risk level and high population:

Name	Population	Federal State	latitude	longitude	count_vegi_rest	count_org_grocs	vegi_per_groc	risk_level
Herne	156374	Nordrhein-Westfalen	51.538039	7.219985	0.0	1.0	0.0	1.0
Wolfsburg	124151	Niedersachsen	52.420559	10.786168	0.0	1.0	0.0	1.0
Reutlingen	115966	Baden-Württemberg	48.491951	9.211414	0.0	2.0	0.0	1.0
Bergisch Gladbach	111966	Nordrhein-Westfalen	50.992930	7.127738	0.0	2.0	0.0	1.0
Remscheid	110994	Nordrhein-Westfalen	51.179871	7.194354	0.0	2.0	0.0	1.0
Hanau	96023	Hessen	50.133554	8.916818	0.0	1.0	0.0	1.0
Gera	94152	Thüringen	50.877230	12.079621	0.0	2.0	0.0	1.0
Iserlohn	92666	Nordrhein-Westfalen	51.374678	7.699971	0.0	1.0	0.0	1.0
Düren	90733	Nordrhein-Westfalen	50.803168	6.482081	0.0	1.0	0.0	1.0
Lünen	86449	Nordrhein-Westfalen	51.614248	7.522809	0.0	1.0	0.0	1.0
Worms	83330	Rheinland-Pfalz	49.630262	8.362090	0.0	1.0	0.0	1.0
Minden	81682	Nordrhein-Westfalen	52.288105	8.916885	0.0	1.0	0.0	1.0
Delmenhorst	77607	Niedersachsen	53.048095	8.628607	0.0	1.0	0.0	1.0
Viersen	76905	Nordrhein-Westfalen	51.256212	6.390548	0.0	2.0	0.0	1.0
Gladbeck	75687	Nordrhein-Westfalen	51.571866	6.987734	0.0	1.0	0.0	1.0
Arnsberg	73628	Nordrhein-Westfalen	51.400238	8.060591	0.0	1.0	0.0	1.0
Brandenburg an der Havel	72124	Brandenburg	52.410826	12.549793	0.0	2.0	0.0	1.0
Celle	69602	Niedersachsen	52.624056	10.081052	0.0	2.0	0.0	1.0
Lippstadt	67901	Nordrhein-Westfalen	51.674707	8.347194	0.0	1.0	0.0	1.0
Herford	66608	Nordrhein-Westfalen	52.115224	8.671112	0.0	2.0	0.0	1.0

Figure 12: Top 20 cities suitable for opening a vegetarian restaurant

Risk level - Number of cities:

The risk levels are very unevenly distributed. While risk level1 accounts for over 27%, the rest are very poorly represented, with the exception of the manually added risk level 6, which accounts for the vast majority (over 65%). For risk level 6 ("unknown risk level"), no clear statement can be made due to the lack of indicators.

Risk level	Count	Percent
1	194	27,7%
2	1	0,1%
3	3	0,4%
4	44	6,3%
5	2	0,3%
6	456	65,1%
Sum	700	100,0%

Figure 13: Risk level Number of cities

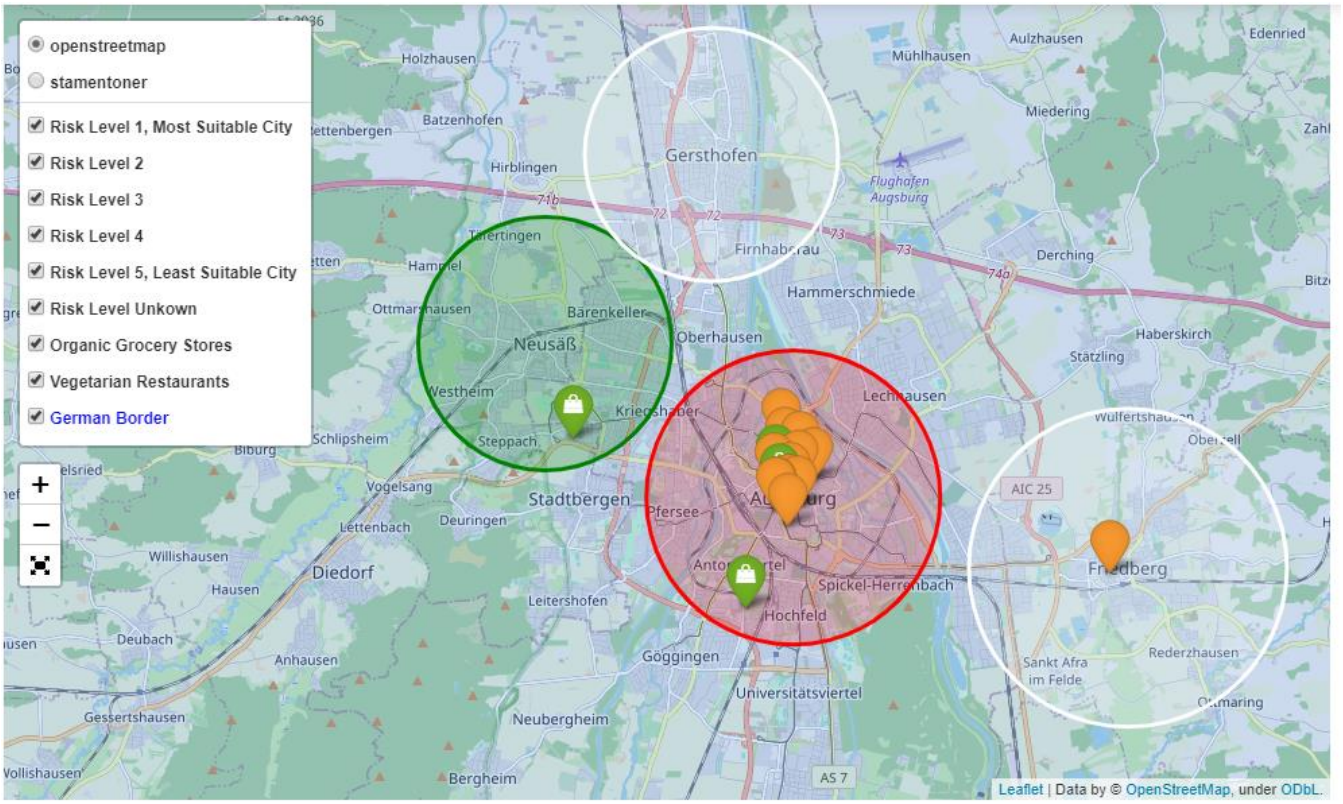


Figure 14: Example of several cities with various risk levels located close together. Green: well suited, red: oversaturated, no good choice. white: unknown as there is no organic grocery store to build a risk score with

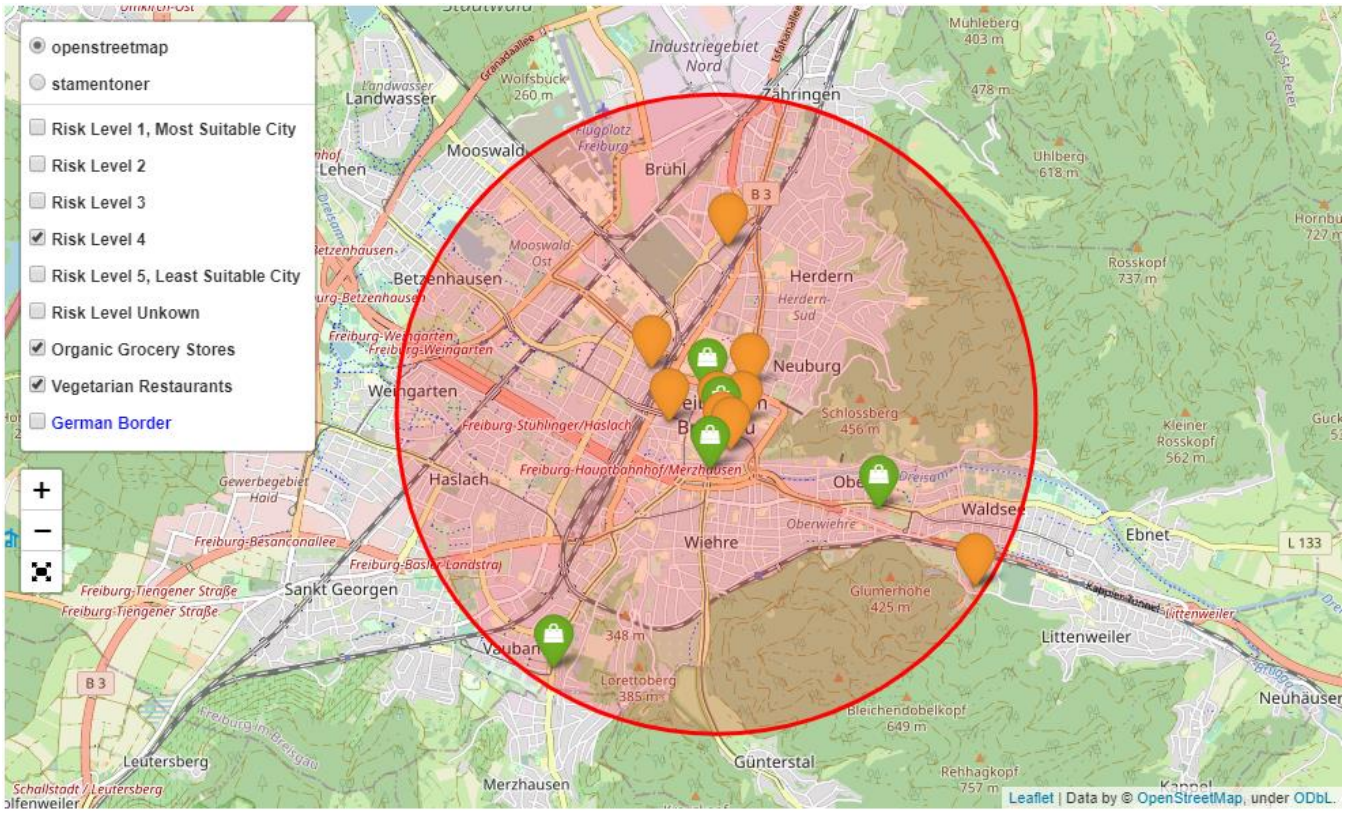


Figure 15: High Risk Level (Freiburg), saturated with vegetarian restaurants, even though a high amount of grocery stores indicate good market opportunities

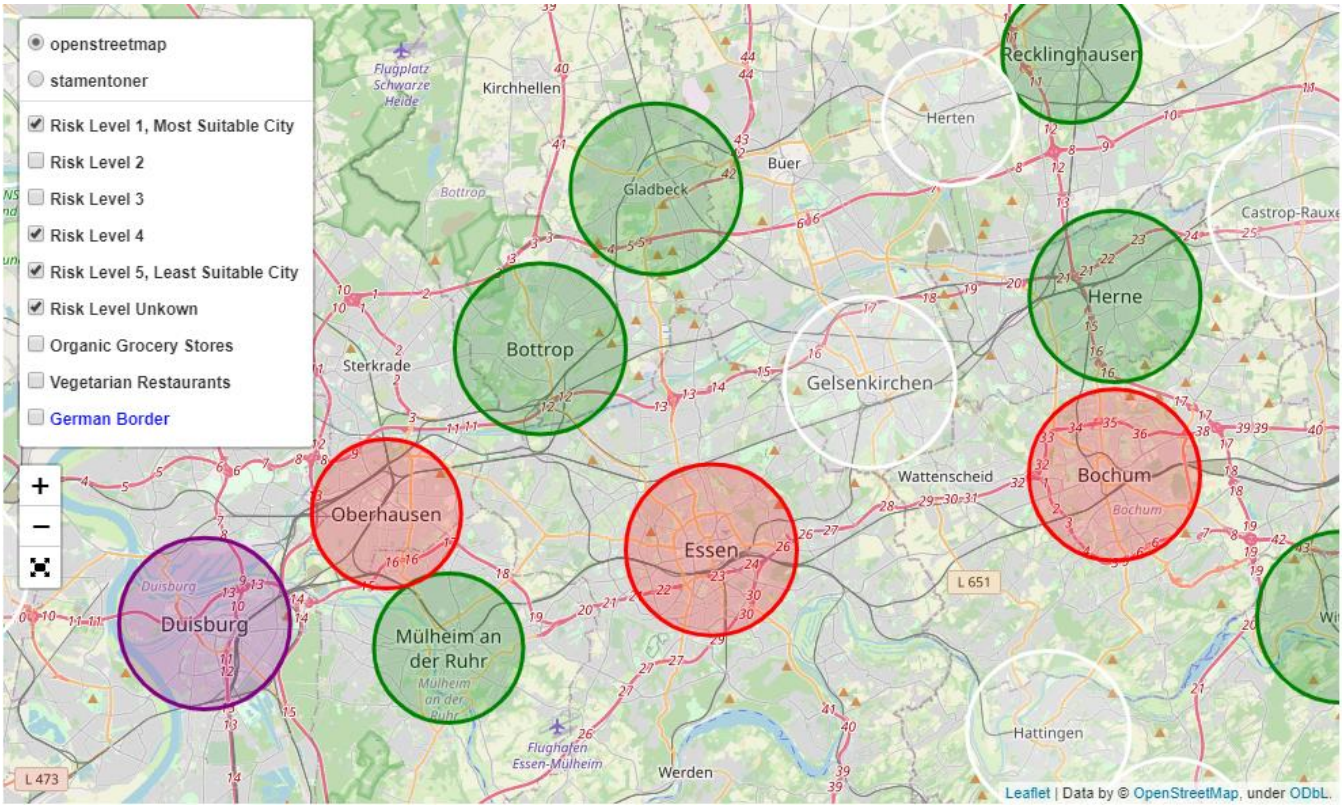


Figure 16: A view of the most populated area in Germany shows clearly the recommended cities (green) as Herne, Gladbeck, Bottrop. The "no go " areas (Purple and red) are Duisburg, Oberhausen, Essen and Bochum. No result can be provided on Herten, Gelsenkirchen etc.

Discussion of the results

Based on the results, it can be concluded that the market opportunities for vegetarian restaurants are very good at for almost 28% of the cities in scope. As the charts show those cities are located mainly in the Middle-West and South-West in areas of high population density.
Mainly medium-sized cities with less than 100000 inhabitants in the Middle-West and South-West are particularly suitable. However, larger cities at a low risk level appear most attractive.
In principle, however, the results can only be a simple indicator due to the existing limitations .
Ultimately, the most promising condition is given if a very low amount of vegetarian restaurants is combined with a high amount of existing organic grocery stores.
It would be very helpful if many more parameters were included, such as age structure, income, political preferences. Surveys would also provide valuable information. Furthermore, historical data on the development over time would provide valuable information.
The more parameters are supplied as input, the more valuable the results are through machine learning.

Limitation, possibilities for improvement:

- **Precise location information within a city is not given, but the result only lists whole cities as possible locations.**
- **The result is a classified assessment of the risk based on the indicators used, which were calculated using K-means clustering. The result of DBSCAN showed less clusters but the most important one (lowest risk level) showed exactly the same.**
- **For their part, the indicators used have weaknesses, e.g. there is no indication of turnover and size in organic food shops. Both factors could be used as input parameters to get a better result**
- **It is not clear whether Foursquare has actually registered all restaurants and shops. Incorrect classification can also influence the result.**
- **The search radius for the Foursquare data is limited to a maximum of 3 km. If a neighboring city is closer than 6 km, the search radius is reduced to half the distance between the two cities to avoid overlapping.**
- **In the case of cities near the border, greater inaccuracies may occur, as foreign cities are not taken into account (e.g. Basel / Switzerland)**
- **Historical data: Trend data could provide valuable information on whether and how the market situation is changing in order to determine a positive or negative trend. Unfortunately, this data is not available free of charge.**
- **In principle, the following parameters should also be used for a more in-depth investigation:**
  - Age structure-
  - Does the city home a university
  - Income structure
  - Political orientation ...

Final statements and further information

This work produced a so-called "Suitability Analysis" where the spatial reference plays a central role. For this purpose, there are also special professional tools like ESRI ARCGIS PRO , which were originally developed for the area of "Geographical Information Systems", but are now very intensively using artificial intelligence / machine / deep learning algorithms [4].
The tool “Jupyter” notebook has its advantages but also its weaknesses. As a “swiss knife tool” it is capable of doing almost everything one can do with Python. On the other hand it is sometimes a little hard and very time consuming to program the charts and each and every tick on the axis. For some tasks Excel might be a better friend, especially if you are not that much experienced.

Bibliography

[1] A. Vou, „Europe is going veg,“ 12 03 2019. [Online]. Available: https://www.europeandatajournalism.eu/eng/News/Data-news/Europe-is-going-veg.

[2] N.-G. Wunsch, „Meat substitutes market size in selected countries Europe in 2018,“ 29 11 2019. [Online]. Available: https://www.statista.com/statistics/1056227/meat-substitutes-market-size-in-europe-by-country/.

[3] Mordor Intelligence, „MEAT SUBSTITUTES MARKET - GROWTH, TRENDS, AND FORECAST (2020 - 2025),“ 2019. [Online]. Available: https://www.mordorintelligence.com/industry-reports/meat-substitute-market.

[4] ESRI, „ARCGIS PRO Suitability Analysis,“ ESRI, 2020. [Online]. Available: https://pro.arcgis.com/de/pro-app/tool-reference/business-analyst/understanding-suitability-analysis.htm.

[5] C. Horseman, „No meat today: The rise of vegetarianism and veganism,“ 01 10 2019. [Online]. Available: https://iegpolicy.agribusinessintelligence.informa.com/PL221686/No-meat-today-The-rise-of-vegetarianism-and-veganism.

[6] European Vegetarian Union, „European Vegetarian Union,“ 2020. [Online]. Available: https://www.euroveg.eu/publications/events/.

[7] Wikipedia, „Vegetarianism,“ 2020. [Online]. Available: https://en.wikipedia.org/wiki/Vegetarianism.