Capstone Project - The Battle of Neighborhoods

Finding the optimal location for a supermarket in Kiel, Germany

1. Introduktion

1.1. Background

The geographic location of a supermarket has a great impact on sales and earnings. Therefore it is important to evaluate all possible areas and find the best location before opening a new store. Walk-in traffic has a positive effect on earnings, while competitors nearby will have a negative effect.

1.2. Problem

In this project we will try to find the optimal site for a new supermarket in Kiel, Germany. The new supermarket should be in an area which is not already crowded with other supermarkets to minimize the effect of competitors. At the same time it should be close to a pharmacy or drugstore. The latter could cause a higher walk- in traffic and increase the revenue. The aim of the project is to have a map showing the best areas for an opening of a new supermarket based on the given conditions by using data science.

1.3. Interest

This study will be of interest to anyone who wants to open a supermarket in Kiel, for example supermarket chains. Possible stakeholders might not be physically present to search for the right place. An agent might be needed to evaluate the full area, which is expensive and time consuming. Data science provides a possibility to evaluate the full area remotely and time efficient.

2. Data acquisition and preparation

2.1. Data sources

For the study all supermarket, drugstore and pharmacy locations within the area of interest are needed. The information is gathered by using Foursqaure API.

2.2. Data formatting

The first step was to define the city center of Kiel. According to ‘www.gps-latitude-longitude.com’ the latitude is 54.3232927 and the longitude is 10.1227652.

The area of interest is 5km x 5km around the city center. A regular spaced grid was created with a cell size of 250m x 250m around the city center. For accuracy distances will be converted to a Cartesian 2D coordinate system before latter calculation and transformed back to latitude/longitude for displaying.

The retrieved API data is transformed into python dataframes. Post data import data was cleaned and formatted. Only the important information such as 'name' of the store, address (street, postcode, city and state), latitude and longitude were kept. In the end three dataframes exits, one for all supermarkets, one for all drugstores and one for all pharmacies.

	name	address	lat	lng	postalCode	city	state
0	REWE	Weißenburgstraße 15	54.325208	10.117955	24116	Kiel	Schleswig-Holstein
1	sky	Knooper Weg 41-43	54.323887	10.127047	24103	Kiel	Schleswig-Holstein
2	REWE	Knooper Weg 41-43	54.324142	10.126881	24103	Kiel	Schleswig-Holstein
3	REWE	Holstenstr. 1-11	54.322404	10.139209	24103	Kiel	Schleswig-Holstein
4	ALDI NORD	Torfmoorkamp	54.354451	10.104890	24106	Kiel	Schleswig-Holstein

Table 1: Top of supermarket dataframe.

	name	address	lat	lng	postalCode	city	state
0	Rossmann	Weißenburgstr. 21-29	54.325340	10.118420	24116	Kiel	Schleswig-Holstein
1	Rossmann	Schönberger Straße 32-34	54.326501	10.183425	24148	Kiel	Schleswig-Holstein
2	Rossmann	Am Ihlberg 4	54.301307	10.047539	24109	Melsdorf	Schleswig-Holstein
3	dm-drogerie markt	Herzog-Friedrich-Str. 30-42	54.320519	10.134703	24103	Kiel	Schleswig-Holstein
4	Rossmann	Kurt-Schumacher-Platz 15	54.322130	10.052540	24109	Kiel	Schleswig-Holstein

Table 2: Top of drugstore dataframe.

	name	address	lat	lng	postalCode	city	state
0	Forellen-Apotheke	Hamburger Landstr. 26a	54.286980	10.085543	24113	Molfsee	Schleswig-Holstein
1	West-Apotheke	Ringstr. 64	54.317645	10.124075	24103	Kiel	Schleswig-Holstein
2	Kronen-Apotheke	NaN	54.328792	10.133755	24105	Kiel	Schleswig-Holstein
3	Doc Morris Apotheke	NaN	54.360407	10.133042	24106	Kiel	Schleswig-Holstein
4	Impuls Apotheke	Sophienblatt 36	54.315578	10.130846	24103	Kiel	Schleswig-Holstein

Table 3: Top of pharmacy dataframe.

By plotting all stores (supermarkets, pharmacies and drugstores) it was observed that 3 stores (1 pharmacy, 1 drugstore and 1 supermarket) where outside the defined area of interest. They were discarded for further analysis. Finally 49 supermarkets, 37 pharmacies and 27 drugstores were found and displayed by using folium (see figure 1).

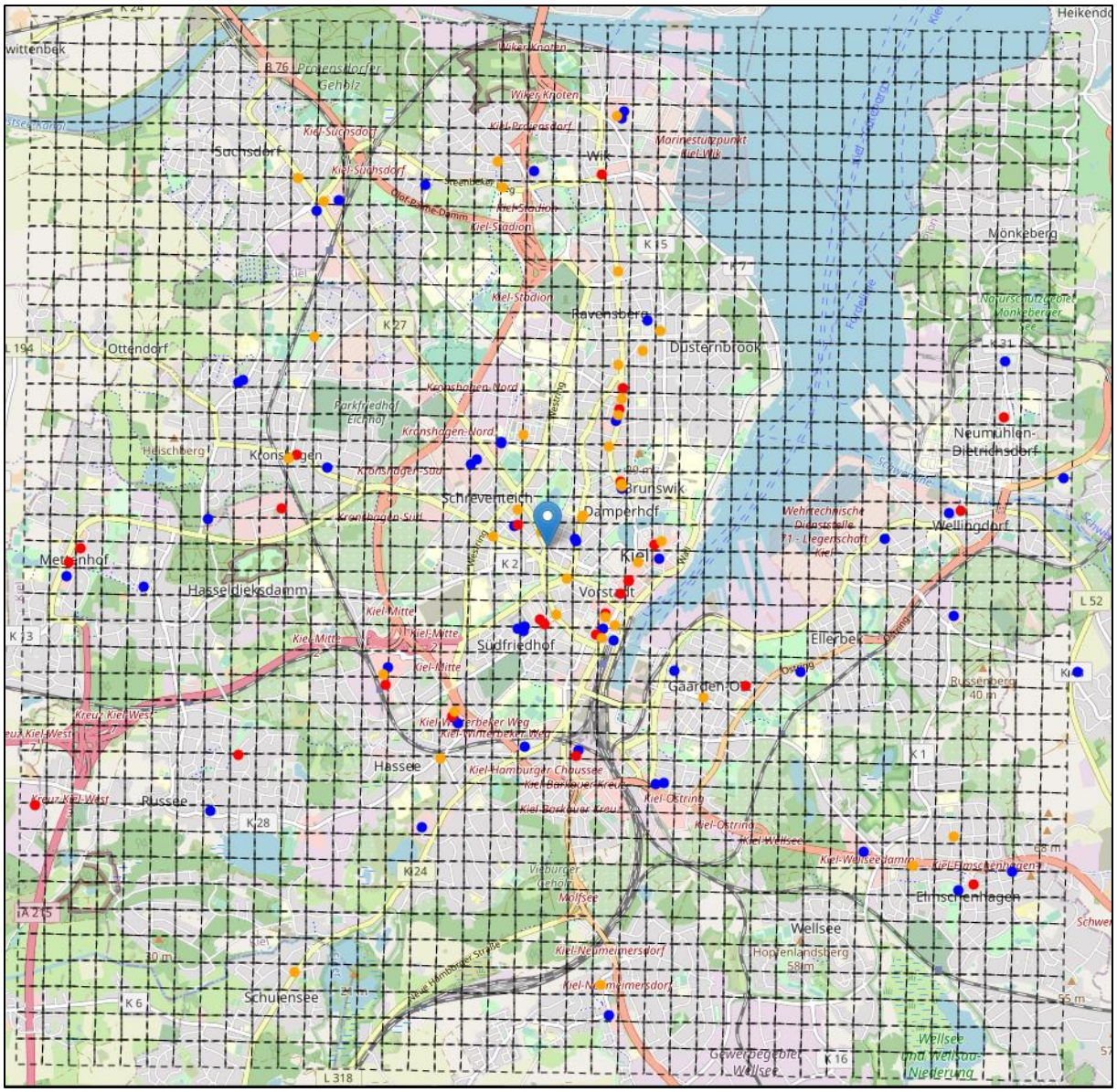

Figure 1: The Grid shows the area of interest around the city center of Kiel. Blue: supermarket, red: drugstore and orange: pharmacy.

3. Exploratory Data Analysis

3.1. Relationship between supermarket and competitors.

The right spot for a new store was defined by two criteria. The first criterion had the aim to limit the effect of competitors. Therefore the distance between a supermarket and its nearest other supermarket (competitor) was analyzed. The distance was calculated in Cartesian 2D coordinate system for more accuracy. Figure 2 shows the result as a histogram. A distance between 250m is most common followed by a frequency gap to approx. 500m. Distances above 1500m are less frequent. The calculated average distance is 566m, the maximum distance is 2256m and the minimal distance is 10m.

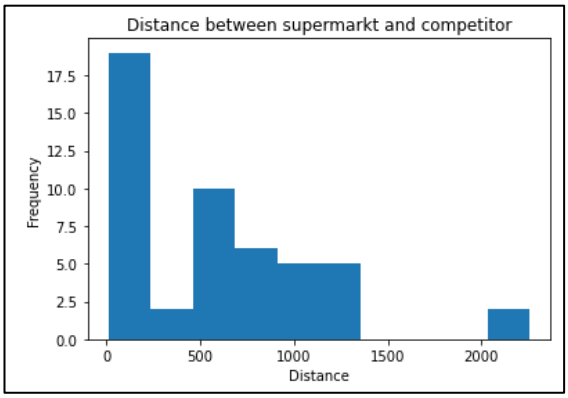

Figure 2: Distance between Supermarket and Competitor.

3.2. Relationship between supermarket and drugstore/ pharmacy

The second criterion for a good spot was defined by the distance to the next drugstore or pharmacy. Therefore the distance between the existing supermarkets and the nearest drugstore or pharmacy was analyzed. An average distance of 392m, a minimum distance of 17m and a maximum distance of 1898m was calculated.

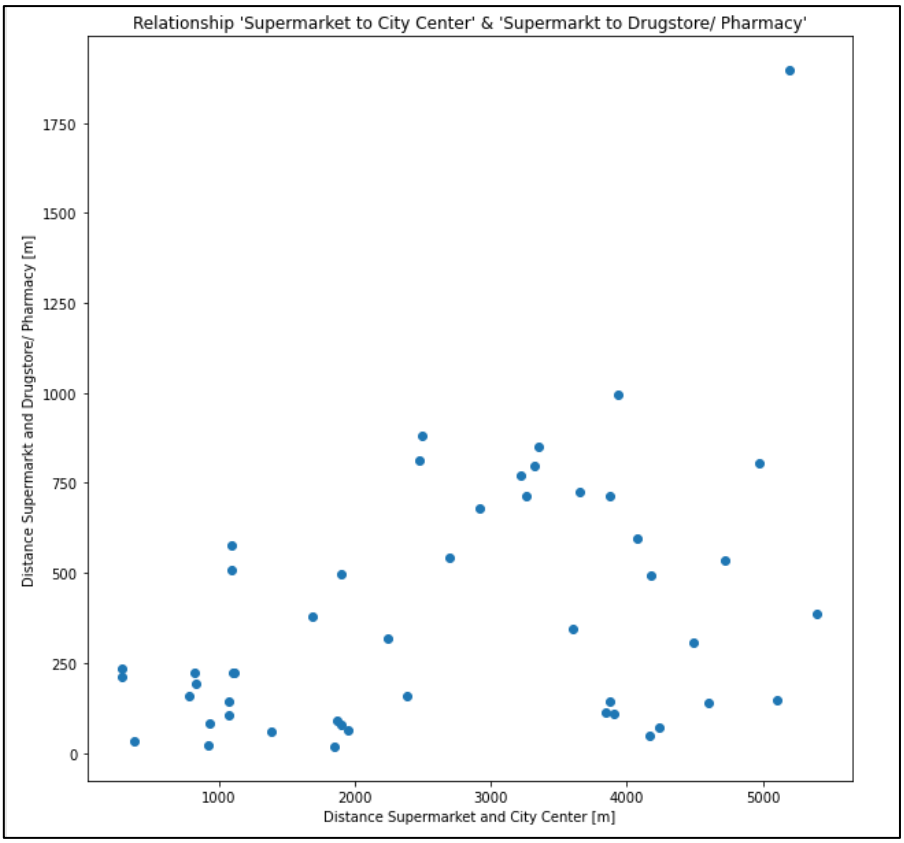

Figure 3: Relationship between 'Supermarket to City Center' and 'Supermarket to Drugstore or Pharmacy'.

People living further away from a city center might be more used to travel longer distances than those closer by. Therefore the relationship between ‘distance supermarket to drugstore/pharmacy’ and ‘supermarket to city center’ was analyzed. Figure 3 shows the result. The relationship seems to follow a polynomial function. For distances under 3000m to the city center the distance between supermarket and drugstore/pharmacy increases with increasing distance to the city center. Between 3000m and 4200m to the city center the ‘distance supermarket to drugstore/pharmacy’ decreases with increasing distance to the city center. Above 4200m with increasing distance to the city center the distance between supermarket and drugstore/pharmacy seems to increase again.

4. Predictive Modeling

4.1. Preparation for Clustering

The final result should be a map showing clusters of areas which meat both criteria. In the beginning each cell of the grid could be the right spot for a new supermarket. Next cells were evaluated if they meet the criteria. Afterwards the remaining cells will be clustered.

Criterion 1: Less Competitors

As described in during data analysis the average distance between supermarket and drugstore or pharmacy was 566m. Quite often the nearest supermarket was within the same cell (<250m). To limit the effect of completion it was decided to drop all cells with a distance of 300m to the next store. This value is under the common (average) distance. Figure 4 shows all cells which were dropped due to this criterion.

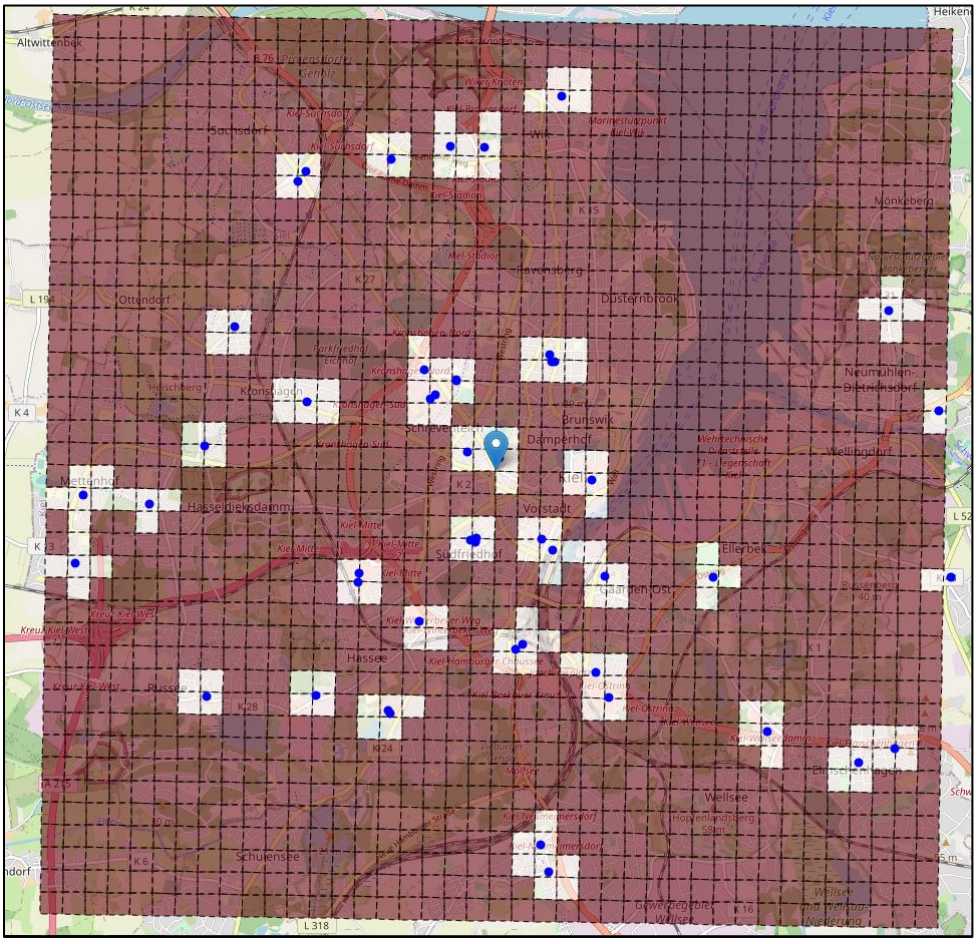

Figure 4: Dropped cells due to criterion 1 (Competitors). All red shaded areas meet the criteria. All others were dropped. The blue dots show location of existing supermarkets.

Criterion 2: Nearest Drugstore/Pharmacy – Polynomial Regression

During data analysis a polynomial relationship between distance to city center and the nearest drugstore or pharmacy was observed. To evaluate the cells a polynomial regression is used to investigate the relationship further. The distance between cell and city center was the dependent variable while the distance between nearest drugstore/ pharmacy was the independent variable.

Firstly the optimal polynomial degree was calculated by calculating the RMSE for different degrees. The result is plotted in figure 5. The optimum calculated degree is 10. Figure 6 (bottom- right) shows the best polynomial regression with a degree of 10. For distances to the city center higher 4400m the model seems to over fit the data. For a degree of 7 the overfitting is less but the accuracy for distances below 4400m is bad (see figure 6, top- left). It was decided to use the model with a degree of 10 up to 4400m to the city center and a linear relationship for larger distances above 5000m (figure 7). Figure 8 shows all cells which were dropped due to this second criterion.

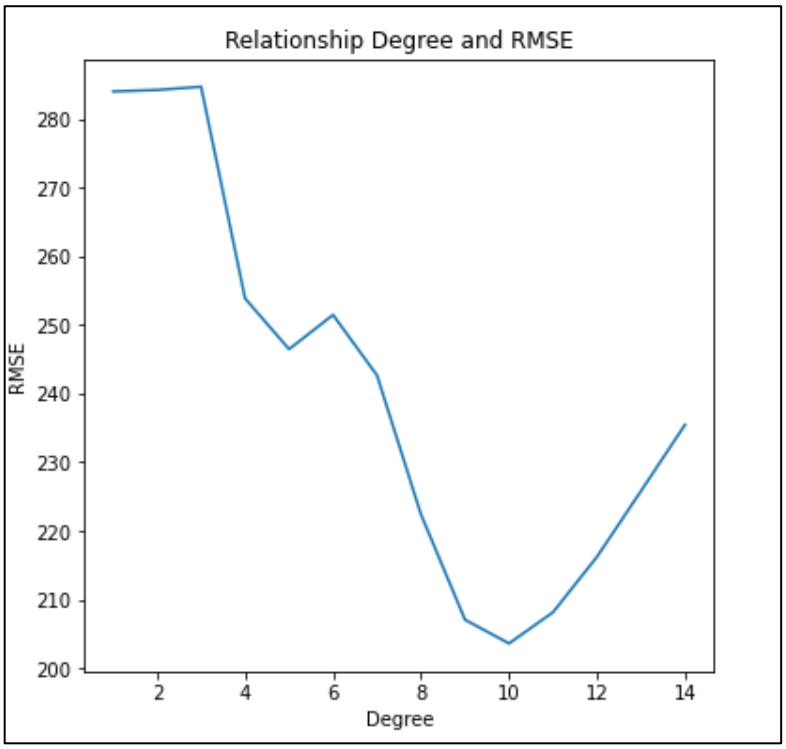

Figure 5: Relationship between polynomial degree and RMSE.

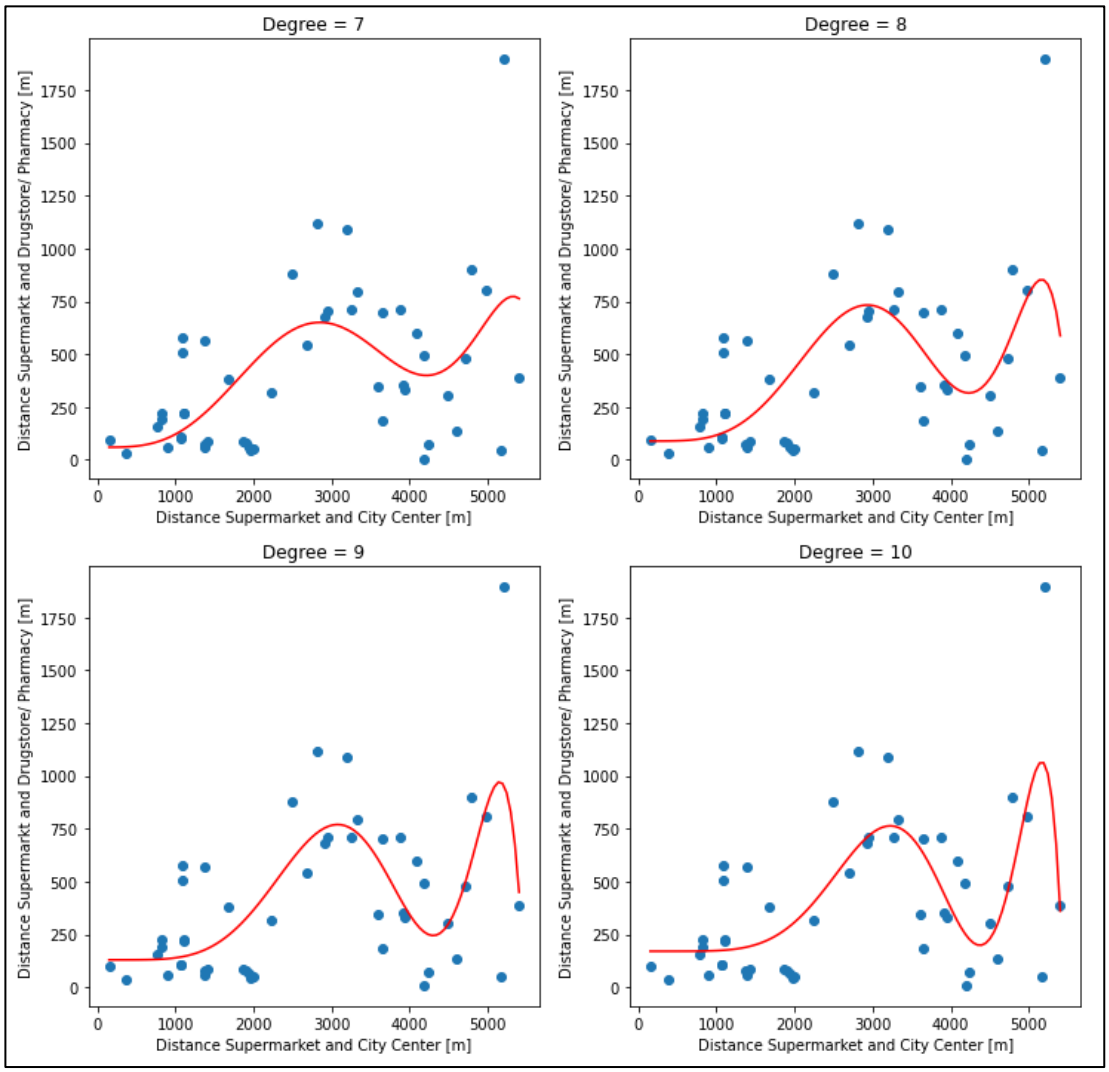

Figure 7: Different polynomial regressions for different degrees.

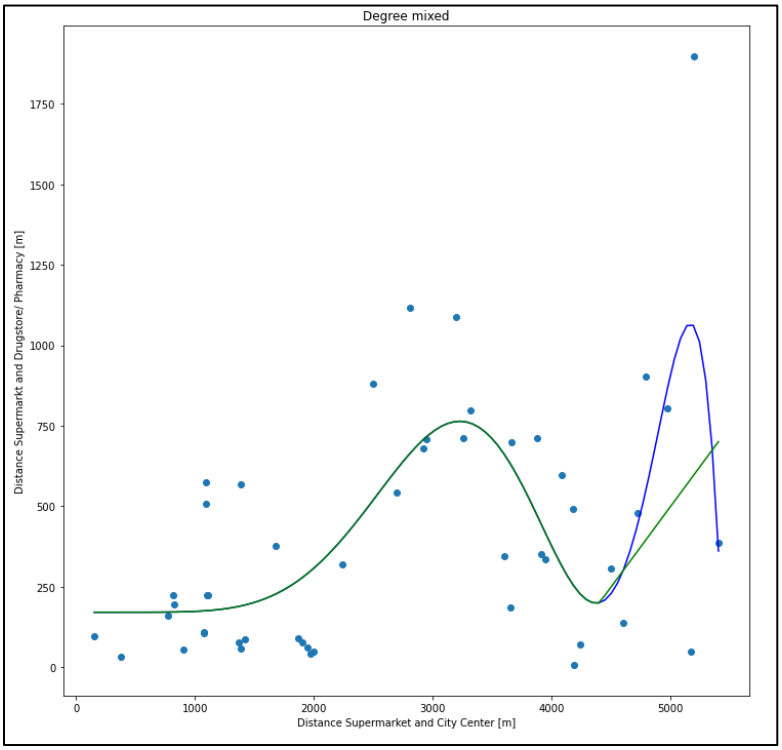

Figure 6: Final function (green). Below 4400m to the city center polynomial regression (degree=10, blue) above linear function.

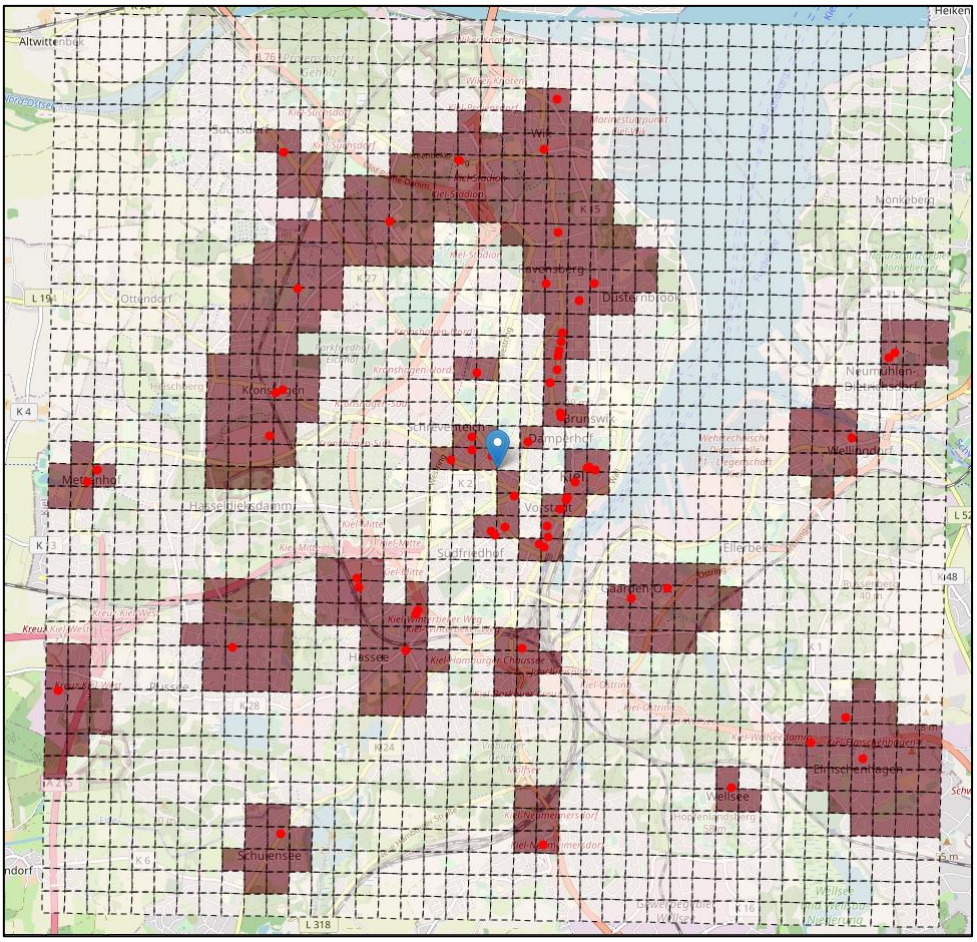

Figure 8: Dropped cells due to criterion 2 (Walk in effect). All red shaded areas meet the criteria. All others were dropped. The red dots show location of existing pharmacies/ drugstores.

4.2. Classification models

For clustering DBSCAN (Density – Based Clustering of Application with Noise) was used. A 400m radius of neighborhood and a minimum of 5 data points within a neighborhood was used to define a cluster. The advantage of DBSCAN over other clustering algorithm is the ability to separate outliers. Especially directly around the city center of Kiel neighborhoods which meet both criteria were incoherently displaced. The final result with 11 clusters is displayed in figure 9. In total 26 neighborhoods were defined as noise.

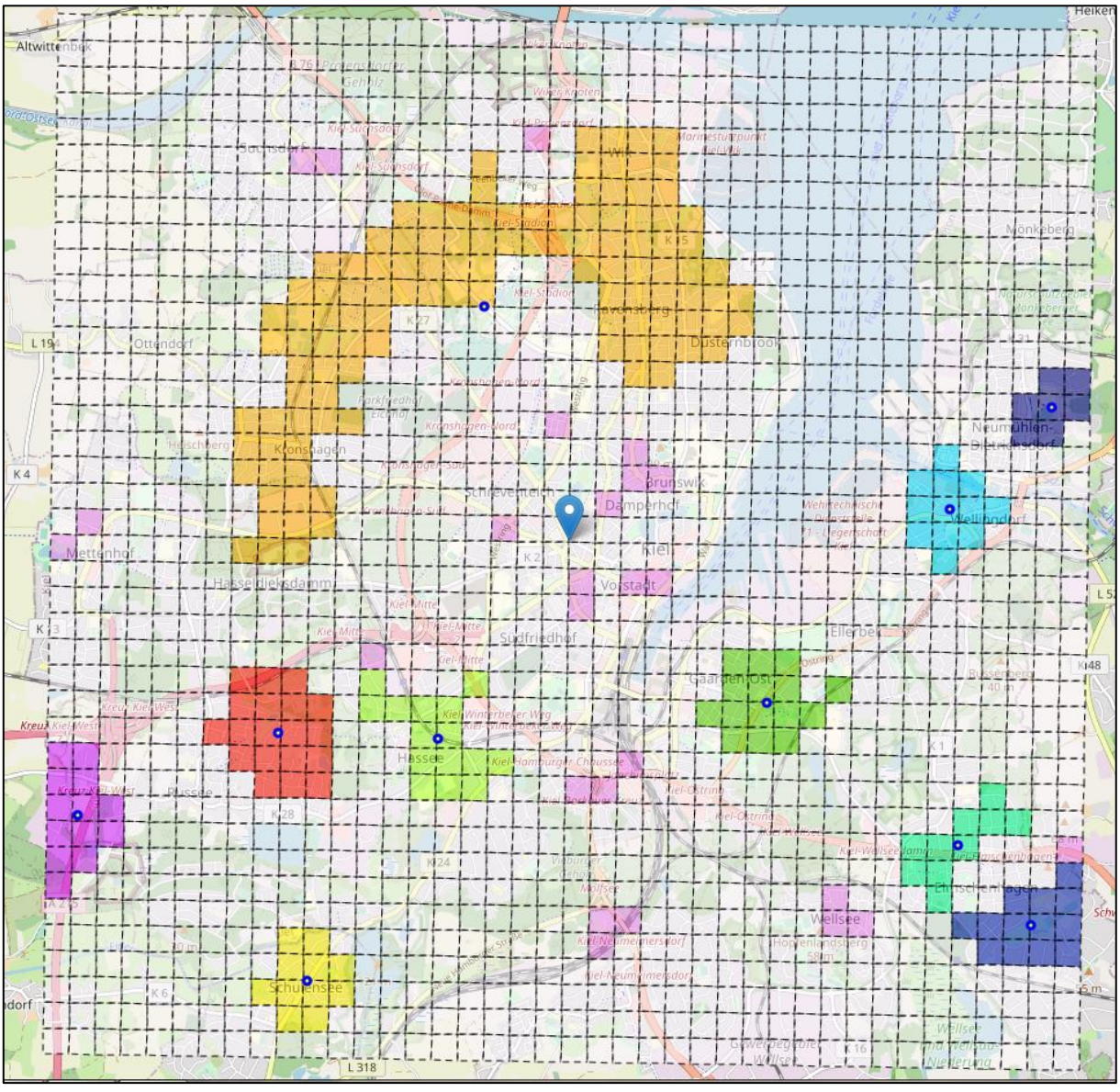

Figure 9: Final map showing 11 clusters inclusive their centers (blue dots). All cells in rosa are classified as noise.

5. Result

The analysis shows that there are areas where both criteria meet. In total 11 clusters were defined. The biggest one is situated north of the city center. It covers an area of approx. 7,1km². It has a flipped u- shape and covers Kronshagen in the west and Ravensberg in the north/east. The remaining 10 clusters are smaller (between 0,25km² and 1,2km²). They can mainly be assigned to the center of districts. For example Schulensee in the south, Hassee, Garden Ost, Wellsee, Elmschhagen (two clusters) and Wellingdorf, Neumühlen Dietrichsdorf in the north east. In the west two areas are around the center of Russee, where one is situated close to the Autobahn.

6. Discussion

The 11 found clusters depend strongly on the two criteria. Changing them has an impact on the neighborhoods which go into clustering. By using DBSCAN for clustering the effect on the input parameters is minimized due to the reason, that a few neighborhoods were classified as noise.

It was observed that one cluster is quite large compared to the other clusters. It might be useful to split this cluster. One possibility would be to drop all neighborhoods which were defined as noise by using DBSCAN and run K-Means afterwards. The advantage would be to define the numbers of clusters upfront.

Also other criteria to define the optimal location for a new store could be useful. For example the amount of people living in each neighborhood, the average income, the type of neighborhood (residential or industrial) and if there is a station close by. One possibility would be to calculate the score for all these criteria (as it has been done for the two used). A density map showing the amount of meet criteria could give a more detailed map.

7. Conclusion

An area of 5km x 5km around the city center was analyzed to find the ideal location for a new supermarket. Two criteria were used to define potential areas. Polynomial regression was used to determine the relationship between city center and the next drugstore/pharmacy for the optimal walk in traffic. To minimize the effect of competitors the distance between existing supermarkets and the next competitor was analyzed. Finally all remaining neighborhoods were clustered by using DBSCAN. The resulting map shows eleven potential areas for a new supermarket. The described method shows how data science can help to evaluate the full area remotely and time efficient.