Capstone Project - The Battle of Neighborhoods

Where to open a Veggie Center

Thomas Pfau

April 13, 2020

Contents

1 Introduction

1.1 Background

We are a group of people dreaming of a wholesome nutrition for all citizens of Dresden (Germany). This nutrition should not only save the health of residents but the environment too.

Therfore we bought a farm near the city to cultivate regional vegetables. So we are able to avoid long shipping distances. As we farm ourselvs it is possible to prevent the use of too much plant protection.

Of course in the first step we can't feed the whole city. We want to start with one shop and hopefuly becoming more and more popular after some time to successfully grow at the end.

An important key for this process will be to find the right place to start the business. The first shop has to become a well-known place to bye wholesome nutrition. It should be located in the neighborhood of venues for people that are looking for such a shop.

1.2 Problem

To find the optimal place for the new shop we have to know which groups of people like organic products and at which places we can find them within the city. In this way it should be possible to be visible as shop to the core customer groups to get started very fast.

1.3 Interest

Other people that have the same dream can benefit from the result of our analysis. Furthermore it would be very easy to adopt the same technics to other kind of shops.

2 Data acquisition and cleaning

To solve the problem sufficient data sources are needed.

For the first part there is a study of Georg-August-Universität Göttingen ("Target groups for organic food: an overview") that describes these groups.

To find the venues that these people like the Foursquare location data set is used. It is a very big resource for many kinds of analysis using many data about other shops, several points of interest and ratings. Correlations between these data should construct clusters of areas with high attraction.

Some work will be needed to clean and restructure the aquired data.

2.1 Data sources

As mentioned before Foursquare has been used to retrieve location data. Foursquare is a great source to get interesting information about venues in an area. Unfortunately there are some limitations to beat:

- in a bigger area it doesn't return as many venues as expected
  So it was necessary to stride the area in smaler steps. 300 meters seems to be an optimal distance.
- I only got a sandbox account for Foursquare
  So I had to 'optimize' this stride to 600 meters for 841 calls of the 950 daily available ones.

The solution of this problem was a function that iterates the area in strides of some distance. In this case 600 meters. After the successful process the result was saved to a CSV file for later processing.

It contains 10.280 entries and eleven columns of useful information. The most important column is 'categories' where we find the meaning of each venue.

2.2 Cleaning data

After some exploration of the resulting venues list it was very clear that it had to be preprocessed:

- drop wrong countries
  It is not understandable why there were venues from Turkey, Nederlands and Czech Republic in the list. One for each country: I removed them.
- a similar problem with venues from other federal state
  Foursquare lists venues from Berlin, Hessen, Brandenburg and Bayern: I removed this 29 entries too. Additional the right spelling of 'Sachsen' must be corrected.
- correct spelling of Dresden
  The name of the city had to be corrected in some cases.

At the end there were 8.490 venues left.

2.3 Feature selection

The main feature for selecting and rating venues are there categories. Since we had 537 unique categories it seems to be a very hard dealing with such a variety.

Therfore the decision was made to summarize them into a lower but also meaningful set of higher level categories found in column 'domains':

Market: To summarize all market venues: **1923**.

Restaurant: To summarize all restaurant like venues: **814**.

Café: To summarize all café like venues: **198**.

Pub: To summarize all pub like venues: **69**.

Snack: To summarize all snack like venues: **77**.

Amusement: To summarize all amusement venues: **336**.

Sights: To summarize all sights venues: **191**.

Education: To summarize all educational venues: **486**.

Culture: To summarize all culture related venues: **265**.

Medical: To summarize all medical venues: **495**.

Sports: To summarize all sports venues: **282**.

Recreation: To summarize all recreation like venues: **276**.

Infrastructure: To summarize venues beloning to the city infrastructure: **1819**.

Trafic: To summarize all trafic related venues: **749**.

Hotel: To summarize all hotel like venues: **253**.

Unknown: As first step all venues became this master category. So we mark all venues not in an interesting domain as 'Unknown'. At the end we had **265** in this special main category.

3 Exploratory Data Analysis

Figure 1 on page 4 shows the overall distibution of the venues on the map of Dresden and its surrounding environment. What we can't see are the best places to start with our new shop. These location have to be determined using an algorithm.

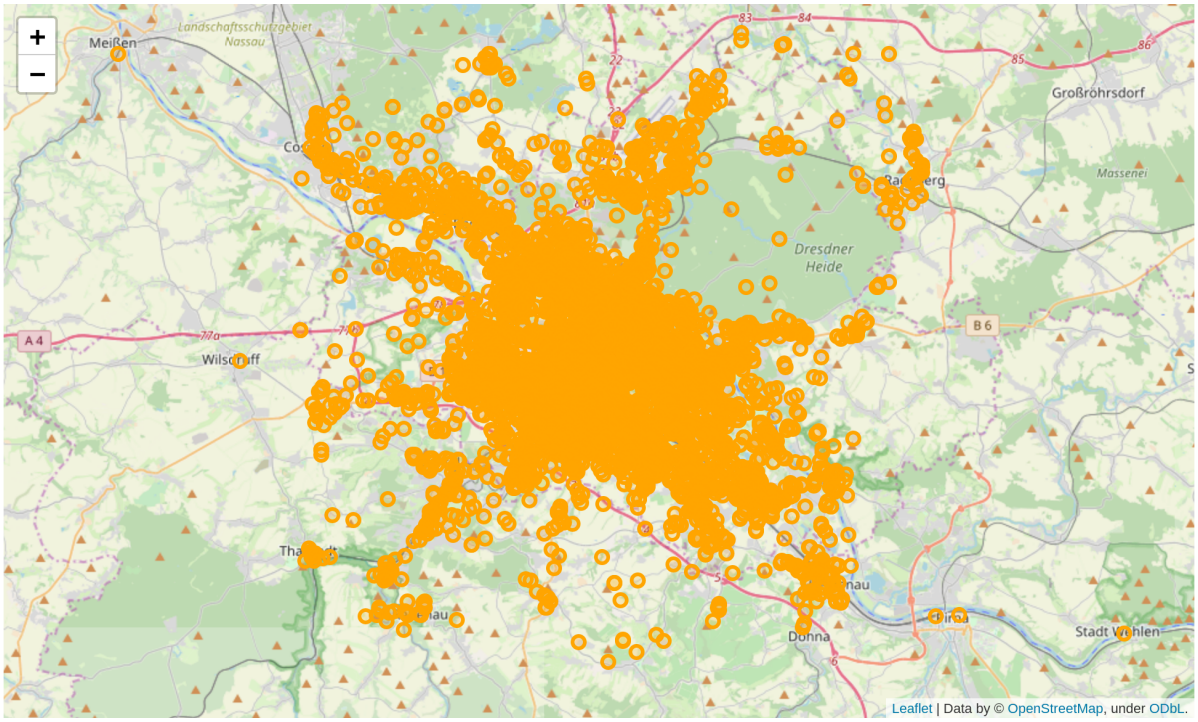

Figure 1: Resulting map

3.1 Definition of the target

In this case we are looking for a list of weighted geographic coordinates as potential sites of our new shop. It will be assumed that the quality of a site depends on the kind of the surrounding venues in an appropriate distance.

The algorithm has to find coordinates as center points of a bunch of interesting near located venues. After that the quality of each found center can be weighted using the weigths of the cluster members.

3.2 Relationship between the venue categories and the best site for the shop

It is very important to find the right weights for the different kinds of venues. A study of Georg-August-Universität Göttingen ("Target groups for organic food: an overview") has described different groups of people and how they buy organic food. On this foundation we defined the different weights of the venue domains. To show positive and negative impact of venue types the weight can be negative too. This way some kind of venue can increase or decrease the summarized weight of a site.

The following table shows the weights of each master category (domain) used by the algorithm:

Domain (master category)	**Weight**
Market	100
Infrastructure	50
Trafic	40
Education	30
Culture	30
Hotel	30
Restaurant, Café, Snack, Pub	20
Sights	20
Medical	10
Sports	10
Recreation	10
Amusement	-50

4 Modeling

4.1 Choosing the model

Since this problem is related to explore clusters of venues the K-Means clustering algorithm seems to be a very good choice. It is fast and doesn't need labeled data.

The only challenge would be to find the right 'K'.

4.2 Find the right 'K' of K-Means

After some experimenting with the Elbow method we found a K by defining how many venues should surround one site. From the real distribution of venues at the map we

decided to find clusters with about twenty members. This should scan the whole area of Dresden fine grained enough.

As long as we found 8490 venues the value of 'K' was set to 424. The calculation of this amount of clusters runs surprisingly fast: under a minute.

The result of the clustering using K-Means can be found in figure 2 on page 6.

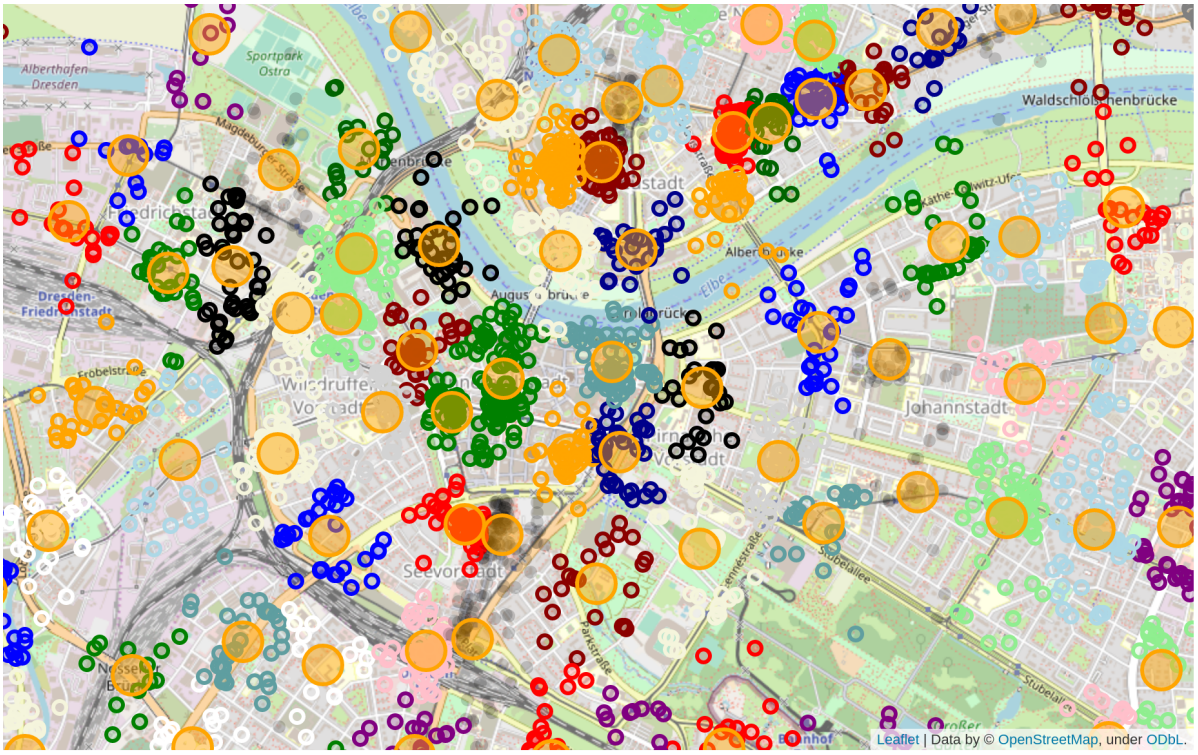

Figure 2: Overview of the clustered city center

It shows the bigger orange colored centers and the surrounding cluster members using distinct colors.

4.3 Solving the problem

Using the list of found clusters and its centers the next step was to weight and compare all clusters. This has been done by sum-up all weights of venues of a cluster to its centers.

Figure 3 shows the distribution of the summarized weights of the centers. It was also necessary to find some classification stages of the weights.

	lat	lng	weights
count	424.000000	424.000000	424.000000
mean	51.056172	13.739545	858.584906
std	0.081601	0.103691	915.093235
min	50.854359	12.556950	-50.000000
25%	51.024545	13.695424	230.000000
50%	51.051021	13.740821	615.000000
75%	51.081385	13.789458	1145.000000
max	52.405352	14.120188	7190.000000

Figure 3: Overview of the clustered city center

Classification	**Color**	**Weight distribution**
Excellent places (4)	Green	Bigger than max. weights/2
Good places (3)	Blue	Between max. weights/4 and max. weights/2
Neutral places (2)	Dark gray	Between mean weights and max. weights/4
Bad places (1)	Yellow	Between min. weights/4 and mean weights
Worst places (0)	Read	Less than min. weights/4

Additional we defined some useful colors of this classifications for visualization purposes.

In figure 4 on page 8 you can see the resulting map showing all interesting and weighted site centers.

5 Conclusions

In this study, I analyzed the distribution of venues in relation to find promising sites for our new Veggie Center. I identified all interesting venues in the area using Foursquare as data source.

After some preprocessing and cleaning I prepared the data to be useful. So it was needed to build a higher level category of the Foursquare categories and to weight them in relation to the effect for our new shop.

Using K-Means as cluster algorithm it was possible to process the whole area at once with very high performance finding 424 centers of interest. Nine of them were reported

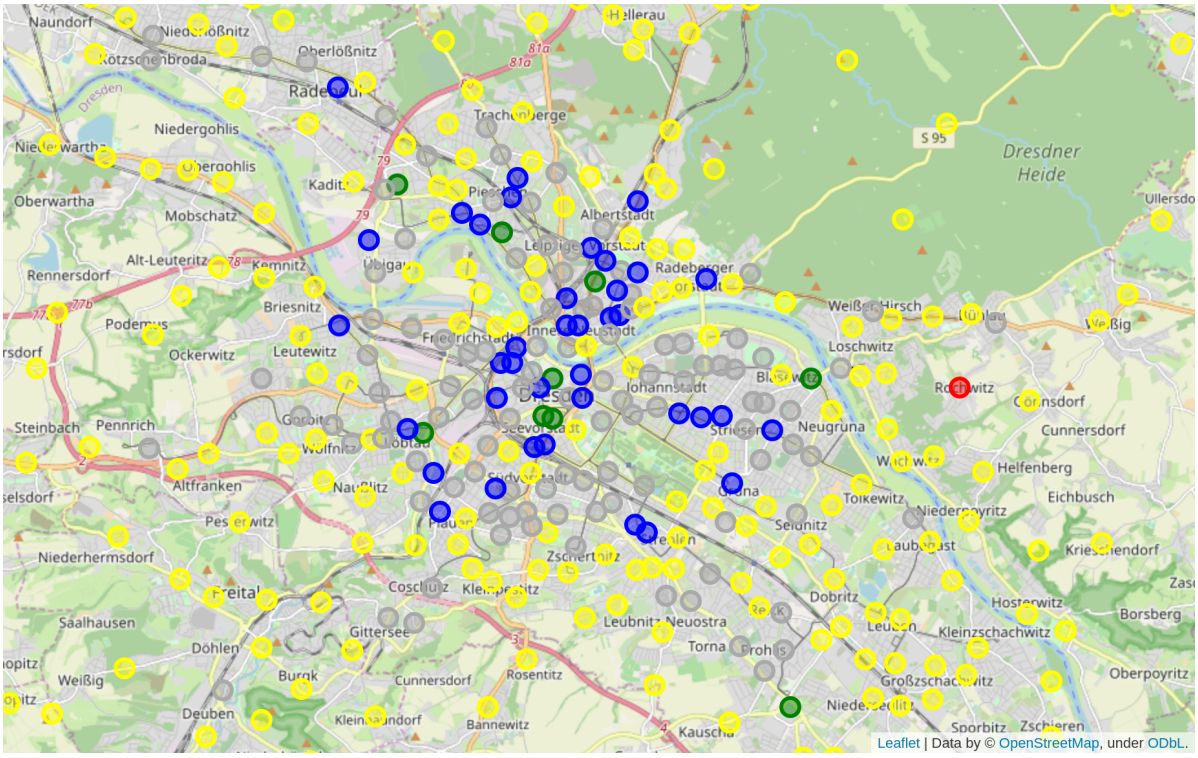

Figure 4: Overview of the weighted centers of interest

as excellent places after the evaluation of the weights. They can be found in figure 4: the green marks. However, if it isn't possible to rent some rooms at the green sites there are enogh of the blue ones as good alternatives. Only the red and yellow sites should be avoided.

6 Future directions

I found nine excellent and many good places. That seems to be sufficient to find a site. When exploring the excellent sites using Google Maps I recognized that the study has delivered useful results.

For future improvements I think about extenting the weights with more data. For instance, the rental prices are as important to the decision as the number of residents in the areas of the sites.