Applied Data Science Capstone

- Battle of Neighborhoods -

Urban Farming in Berlin

by Armin Drusko

Introduction

The food industry, like many others, is searching for novel ways to provide more environmentally-friendly products and services. This project was done together with an entrepreneur who is aiming to start an urban farming business in Berlin. His goal is to cover the needs for fresh vegetables and herbs of local restaurants and other food-serving venues with an infrastructure of farming facilities across the city. His motivation is to maintain a fresh source of ingredients for local restaurants, reduce costs and environmental pollution due to logistics and promote a self-sustaining local community.

The client has already conducted previous financial consultations and is now confident about starting the business. The questions that come up now are which neighborhoods in Berlin could be considered as a region for “farming” - encompassing enough food-serving venues that could serve as potential customers. Also, where should the facilities be placed to cover a broad number of venues simultaneously. Furthermore, if a target area is identified, which characteristic cuisines are served there and could the farming be planed accordingly.

The farms (facilities) are big and sophisticated enough to grow and provide plants for several venues (growing in layers with fully optimized conditions), but small enough to fit into an urban area e.g. 6 x 10 x 3 meters. The facilities will either be built into existing infrastructure (renting space around town) or assembled separately (looking like cool high-tech ship containers). The location of the facilities will be determined geographically. Once a potential area is determined, a place for rent has to be found. The transportation of the grown plants will be conducted by bycicle, which limits the transportation up to one kilometer from the facility. Also, plants for up to 20 venues can be grown. These details will be considered while searching for suitable farming locations in Berlin.

Data

To find a good spot for the farming facilities, a list of neighborhoods in Berlin will be extracted and the geolocation of each will be determined. The list of neighborhoods in Berlin was taken from Wikipedia:

https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Ortsteile_Berlins

and their respective geographical location was acquired with the Python library *Nomatim*.
The city of Berlin is divided in 12 boroughs (Bezirke) and 96 localities/neighborhoods (Ortsteile).
The data about venues at certain areas will be extracted from Foursquare.com. Using their API, the recommended venues were acquired for each neighborhood along with their geolocation and venue category.

Methodology

The client is interested in venues serving any sort of food that includes vegetables and herbs which can be grown in the facilities. We will assume that the farms are meeting all the conditions to grow any plants in demand.

To get an overview of areas with a high occurrence of food-serving venues, the extracted data from Foursquare was filtered on the category property for the following keywords:

Restaurant, Place, Joint, Bagel, Food Court, Steakhouse, Diner

A cluster analysis was performed to explore a potential similarity between the neighborhoods based on occurrence of venue categories. This would give a guidance in planing which plants to grow in the individual facilities based on the needs of the venues it is supplying.

Feature engineering

For the cluster analysis, the occurrence of venue categories per neighborhood was one-hot encoded and their frequency was calculated. With this, the neighborhoods can be quantitatively compared by using the venue categories as features for cluster analysis.

To cluster the neighborhoods, a k-means clustering algorithm was used. The number of features from the dataset was initially 77 i.e. each venue category representing one feature. The dataset had 96 entries i.e. neighborhoods.

Due to a high number of features, the k-means algorithm, which is based on the euclidean distance metric, would perform poorly. Also, the low number of data points would diminish the performance of the algorithm. Here, a dimensionality reduction with principal component analysis (PCA) was conducted in an attempt to reduce the number of features. The explained variance form the PCA represents how much variance of the data is covered by each principal component. The dataset shows a variance of 80 % explained by 10 components. The common approach is to proceed with up to five components. Here, due to the spread of variance across components, the first 10 components were used for further analysis, to capture as much variance as possible. Figure 1 shows the ratio of explained variance for each principal component.

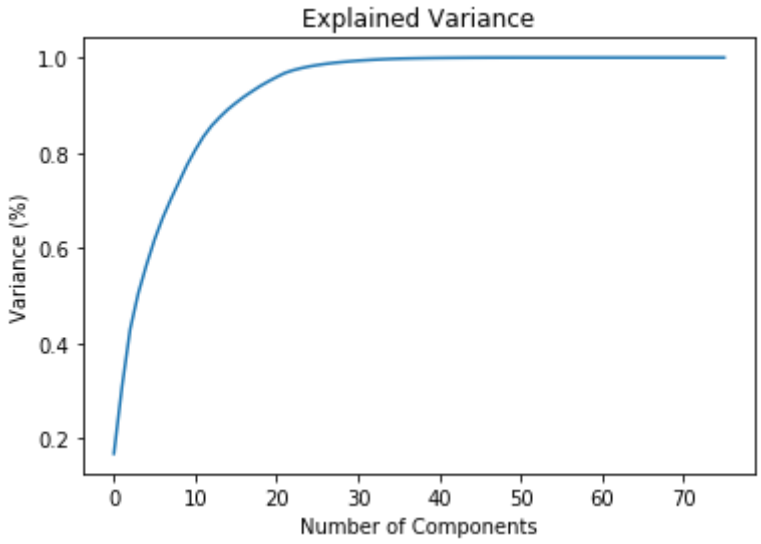

Figure 1: Principal component analysis. The ratio of explained variance for each principal component.

Selecting the number of clusters

The elbow method will be used to determine a suitable number of clusters for k-means. For each *k* in a range from 2 to 30, the intertia of the clusters is calculated. The intertia for each *k* is shown in figure 2.

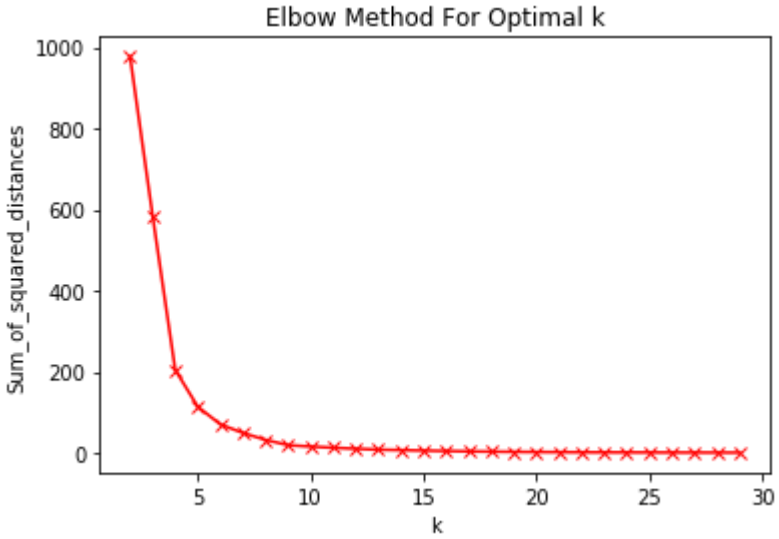

Figure 2: Elbow method for k-means cluster analysis. The sum of squared differences is plotted for each k number of clusters.

A *k* value of 5 was taken as the number of clusters based on the "elbow" position in the curve.

Automated facility placement

The facilities could be placed manually by looking at the geolocation of the target venues and taking the mean latitude and longitude values to calculate a location. This was done in an automated way by utilizing the k-means algorithm. k-means calculates a centroid, based on the the mean value of the data points within the respective clusters.

The geolocation of the venue from the neighborhoods of interest will be used as features to place the centroids, i.e. facilites on the map. A *k* of 10 clusters will be used initially, since 10 facilities can be placed in total. If some clusters are redundant i.e. the coverage over a region is already given by a smaller number of clusters, *k* will be reduced accordingly.

Results

The retrieved dataset from Foursquare was filtered for venues that serve food and sorted for neighborhoods with the highest count. The following table shows the top-five neighborhoods with the count of venues from the selected categories:

	Neighborhood	Venue_Count
0	Moabit	39
1	Schöneberg	38
2	Prenzlauer Berg	38
3	Charlottenburg	37
4	Friedrichshain	37

These 5 neighborhoods contain a total of 189 venues, which is still within the capabilities of 10 venues to supply for their demands. The client wants to identify areas with a high number of potential businesses. Therefore these 5 neighborhoods were taken into account for further analysis.

The map in figure 3 shows a section of Berlin with the top-five neighborhoods (blue markers) and their respective food-serving venues (red markers).

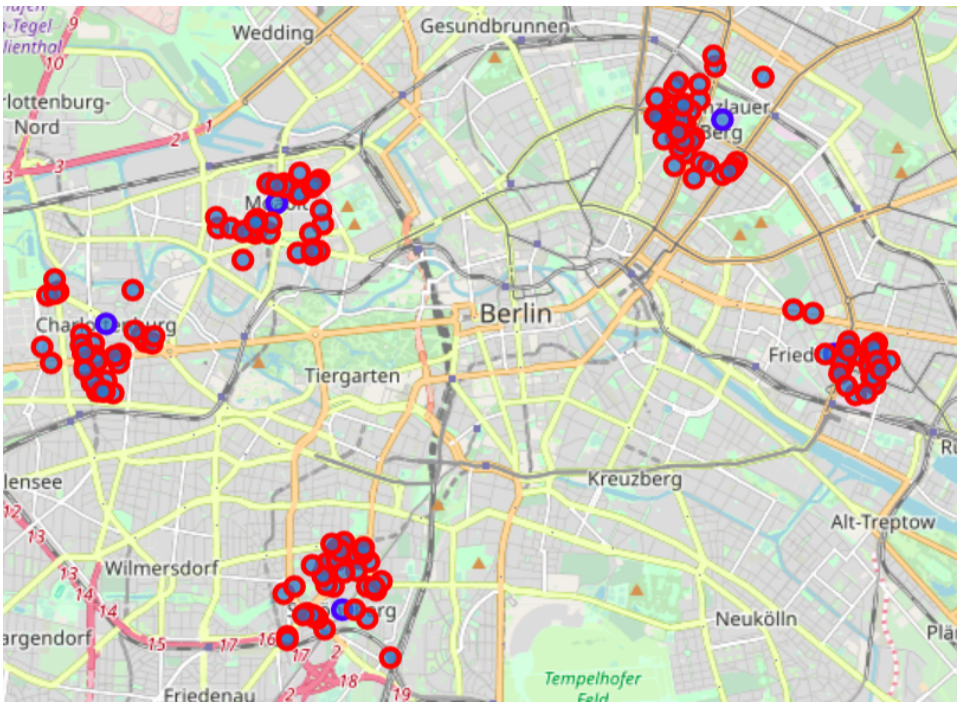

Figure 3: Map of neighborhoods in Berlin with the highest number of food serving venues.

The neighborhoods were clustered based on the frequency of venue categories. Table 2 shows the top five neighborhoods with their respective clustering label. Moabit, Prenzlauer Berg and Friedrichshain are clustered together with a label of 0, whereas Schöneberg and Charlottenburg are within their own cluster with the label 4.

	Neighborhood	Borough	Long	Lat	Cluster_Labels
1	Moabit	Mitte	13.342542	52.530102	0.0
44	Schöneberg	Tempelhof-Schöneberg	13.355190	52.482157	4.0
8	Prenzlauer Berg	Pankow	13.428565	52.539847	0.0
21	Charlottenburg	Charlottenburg-Wilmersdorf	13.309683	52.515747	4.0
6	Friedrichshain	Friedrichshain-Kreuzberg	13.450290	52.512215	0.0

Exploring the clusters 0 and 4, the 1st, 2nd and 3rd most common venues were merged and plotted in figure 4 (left and right, respectively).

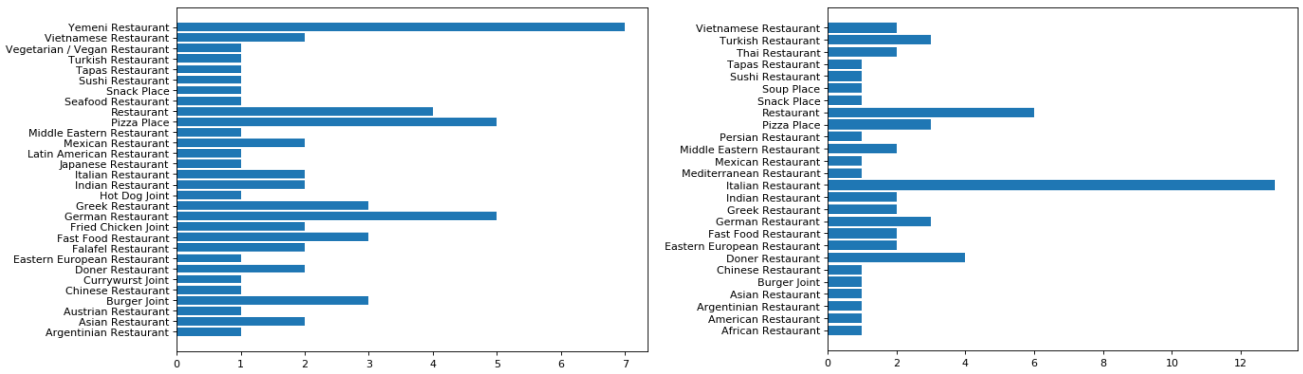

Figure 4: Bar plots of venue occurrences of the 1st, 2nd and 3rd most common categories from the neighborhoods in cluster 0 (left) and 4 (right).

The above category occurrence shows an abundance of Yemeni and German restaurants and pizza places in the cluster 0 and a relatively high number of Italian restaurants in cluster 4.

Since the potential neighborhoods were determined, the automated placement of facilities with the k-means algorithm was performed. The map below shows the placement of facilities with a green circle of 1 km radius, indicating the coverage of each farm. The markers indicate the position of the venues from cluster 0 and cluster 4 colored red and blue, respectively. The facility positions can also be manually adjusted. This was done for the facility 2 and 9, to capture the two outlier venues beyond reach.

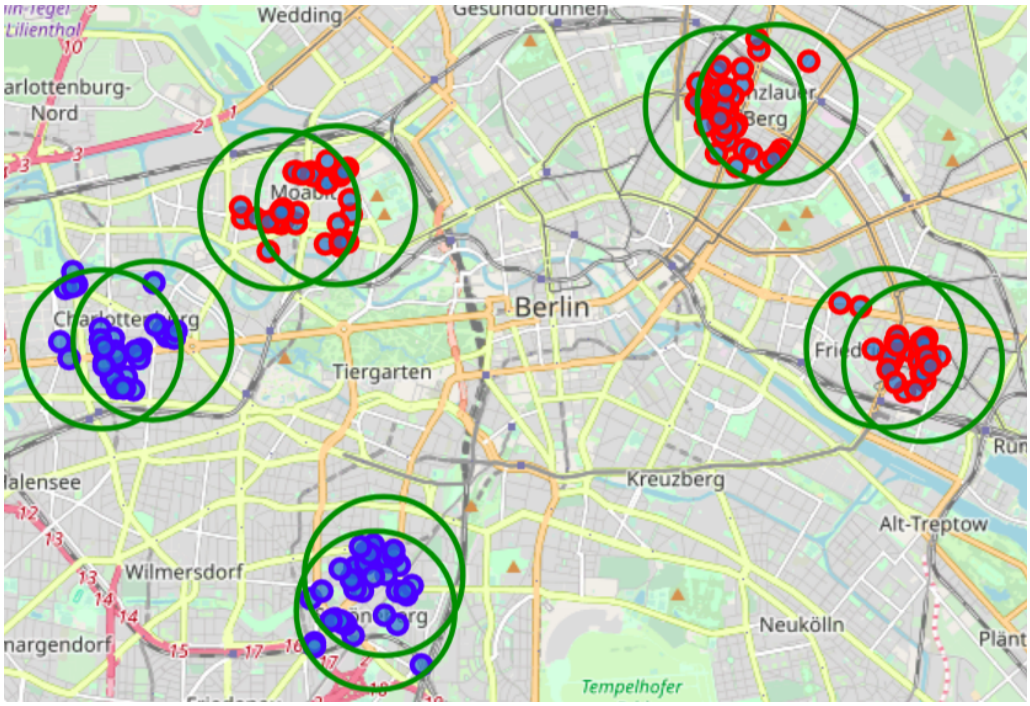

Figure 5: Map of selected neighborhoods in Berlin with a high number of target from the cluster 0 (red markers) and cluster 4 (blue markers). The green circles indicate the supply radius of the placed farming facilities.

Discussion

The analysis resulted in locations for the 10 facilities across 5 neighborhoods with a promising coverage of food-serving venues. Each neighborhood is covered by two facilities. Therefore, the demand would be met. The client can adjust the placement of facilities once the customers were acquired and lower or increase the amount of farms if needed. The spread of the selected neighborhoods around Berlin could also represent an advantageous coverage of the city area, with a lot of potential to spread and merge with neighboring regions. The cluster analysis of the venue categories lead to an interesting distribution of restaurants. This information could now be used to put an emphasis on Italian cuisine in Schöneberg and Charlottenburg and Yemeni and German cuisine in Moabit, Friedrichshain and Prenzlauer Berg. Also the distinction in the frequency of pizza places between clusters 0 and 4 is something to take into consideration. Although the food from both is from the Italian cuisine, cluster 0 could rather contain smaller and fast food-like venues, while cluster 4 includes bigger restaurant, which could have an impact on the demand of ingredients, customer turnover and food quality.

An additional thing to consider is to implement a more sophisticated facility placement algorithm. Although k-means seems to yield satisfactory results, an alternative algorithm could be implemented in the future. Such algorithm would consider maximizing the coverage of surrounding venues within a defined radius of 1 kilometer and sharing venues within centroids, therefore optimizing the placement even further.

Conclusion

The aim of this project was to identify neighborhoods that will provide enough customers for an urban farming business. The farming facilities would be placed strategically around the city to cover the areas of interest and provide fresh vegetables and herbs to food serving venues. The machine learning and data analysis approaches in this project resulted in a list of suggested neighborhoods in Berlin with potential venues that could be further targeted and acquired as customers by a marketing campaign. Additionally, a placement of the farming facilities was suggested in order to meet the demands of the future customers. All in all, an initial consultation and suggestion was given with this project, that could serve as a foundation for further analysis and business development.