1. Introduction

I currently live in a neighborhood that I really like. If I have to move to a new city, I want to find a similar neighborhood. In this project I am writing a program that will evaluate Neighborhoods close to a user-defined destination and compares them to an (also user-defined) reference neighborhood. The scope of the project is limited to Germany.

In order to get the location data for German neighborhoods, data from opendatasoft.com is used.

The neighborhoods are evaluated based on venues listed in foursquare.

2. Data

2.1 Data Sources

Two main data sources are used in this project: opendatasoft.com and foursquare. opendatasoft.com provides the "postleitzahlen Deutschland" dataset which includes coordinate data for each German postal code area. In total there is data for 8697 postal codes

Foursquare is used to get a list of venues for each neighborhood that is included in the comparison.

2.2 Data cleaning and feature selection

The localization data from opendatasoft.com needs very little cleaning. The current data set includes 3 rows with missing data which are dropped. Since I want to limit the search of a candidate neigborhoods to a small area around the target address, the *geopy* function *distance* is used to calculate the distance from the reference address for each postal code in the dataset. Then only the neighborhoods with a distance below a selected threshold (e.g. 10 km) are retained. If for example the target address in located in central Berlin, this reduces the number of possible candidate neighborhoods to 186.

Attaining useful data from foursquare requires more attention. The first step is to aquire a large enough data set. For that purpose, for each neighborhood the 50 most popular venues are requested from foursquare.

In the end, I want to categorize the neighborhoods based on the type of venues that are prevalent there. Therefore, the category data is extracted for all venues of a neighborhood. Since foursquare has few regulations concerning the categorization, this leads to highly fragmented data. Coming back to the example of the 186 neighborhoods surrounding central Berlin, foursquare currently return 6044 venues in 344 categories. This is not suitable for good clustering. In a first step in order to reduce the dimension of the data, I attempt to allocate the extracted categories to one of the 5 main categories defined by foursquare (https://developer.foursquare.com/docs/build-with-foursquare/categories/). When this matching is not possible, it retain the original category. This approach reduced the number of categories from 344 originally to 126. Still a bit to much for clustering.

Hence, on more step for data reduction is performed. Since I want to center the selection of candidate neighborhoods around one chosen reference neighborhood, I eliminate all venues of categories that do not occur in the reference neighborhood. For the chosen example, this brings the number of categories down to 19 which is a sufficiently small number to feed into a clustering algorithm.

3. Methodology

3.1 Characterization of reference and destination neighborhoods

First, I want to find neighborhoods close to the government district in Berlin (Platz der Republik) (Marked by the red dot) which are similar to a suburb of Stuttgart(Vaihingen) (marked by the blue dot)

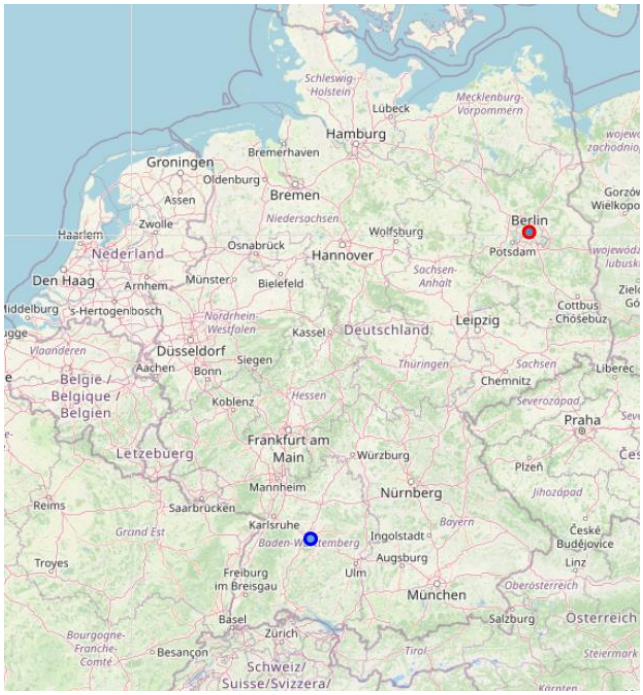

Fig.1: Example reference and destination

With the approach detailed in the data section, 6077 venues are gathered for the clustering of the 186 neighborhoods. After the clean-up procedure, 19 categories remain:

```
'Arts & Entertainment', 'Chinese Restaurant',
       'Climbing Gym', 'Farmers Market', 'Food', 'Gym / Fitness Center',
       'Hotel Pool', 'Ice Cream Shop', 'Nightlife Spot',
       'Outdoors & Recreation', 'Professional & Other Places',
       'Shop & Service', 'Supermarket', 'Sushi Restaurant',
       'Swabian Restaurant', 'Thai Restaurant', 'Trattoria/Osteria',
       'Travel & Transport', 'Wine Shop'
```

Looking at the mean occurrence frequency of each category over all neighborhoods gives a good first insight into the data:

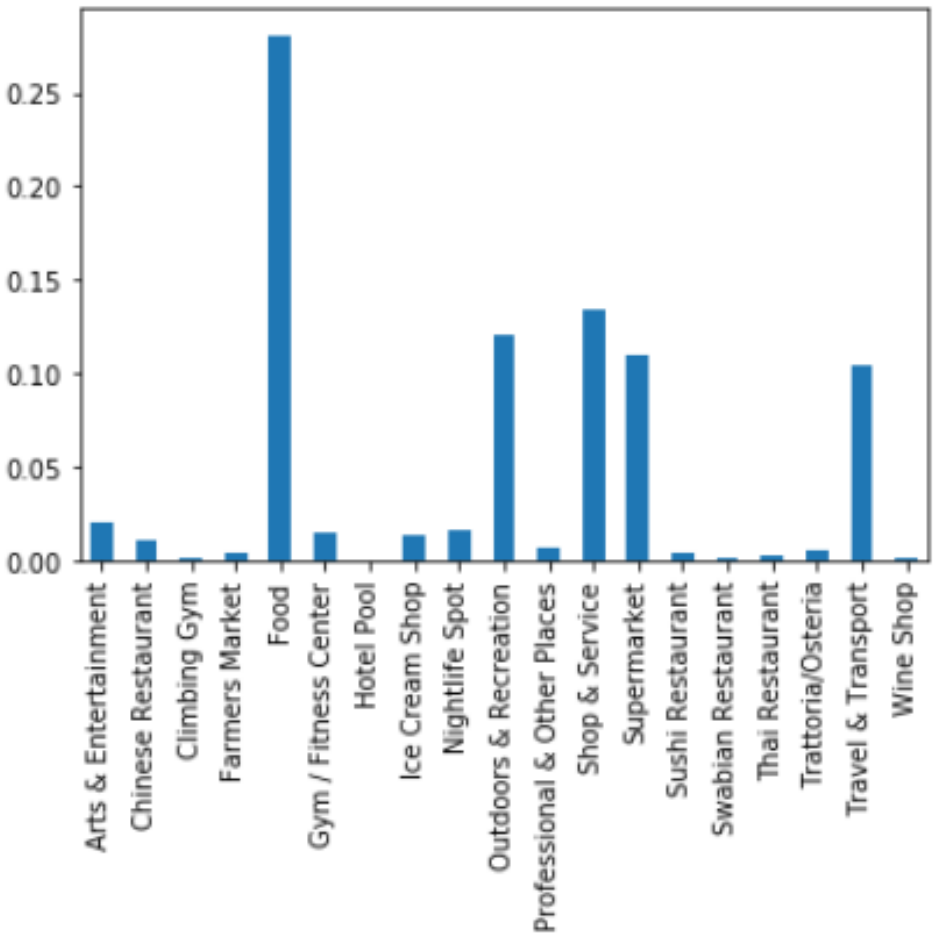

Fig.2 Mean venues frequency in destination area

Food is obviously the most frequently occurring category. This is most likely due to the inner city location chosen as the destination neighborhood with an abundance of restaurants. But it may also show a skewed data base from foursquare. Other frequently occurring categories are outdoors & recreation, shops & services, supermarkets and Travel and Transport. The remaining categories are retained nonetheless, in order to give the clustering algorithm the chance to actually find some specific neighborhood characteristics.

When we compare this mean occurrence of venues to the distribution of the reference negihborhood, it can be seen that the reference neighborhood has an even higher dominance of food venues. The second and third most frequent venues are of the types Travel & Transport and Shops & Services. Therefore it is to be expected that the clustering pairs the reference neighborhood with those neighborhoods at the destination that have a large number of food venues.

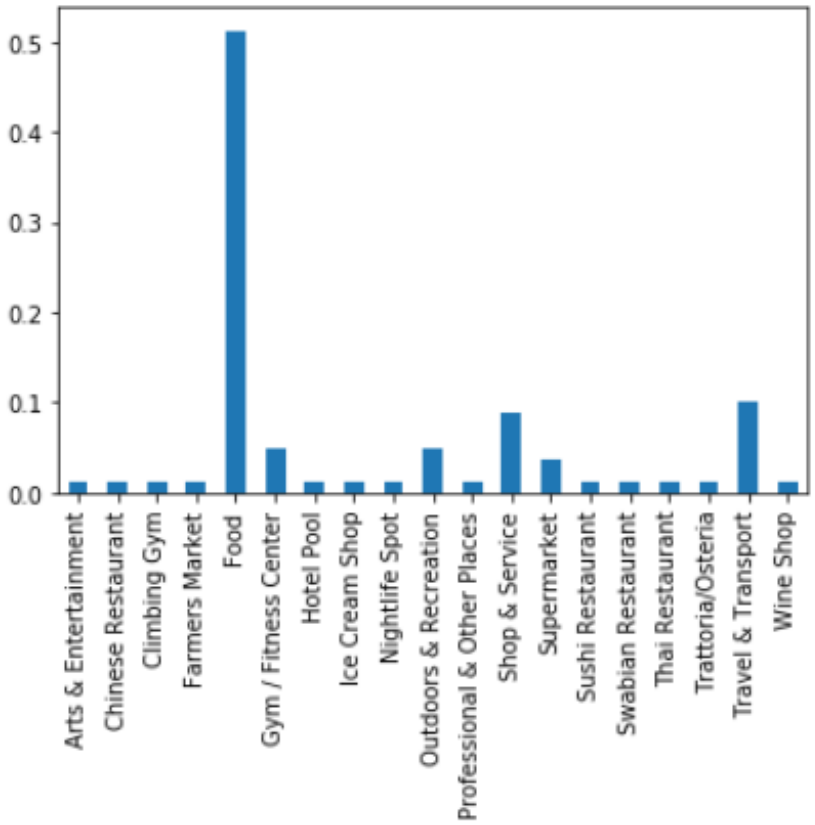

Fig. 3: Venues frequency in reference neighborhood

3.2 Clustering

Based on the extracted features, K-Means Clustering as implementing sklearn is applied to the dataset. The number of cluster is chosen to be 5. In order to use this approach to find neighborhoods similar to the reference neighborhood, the reference neighborhood in included into the clustering data set. After the clustering is finished, the cluster which contains the reference neighborhood is chosen as candidate cluster.

The resulting five centroids are characterized by the following venues frequencies

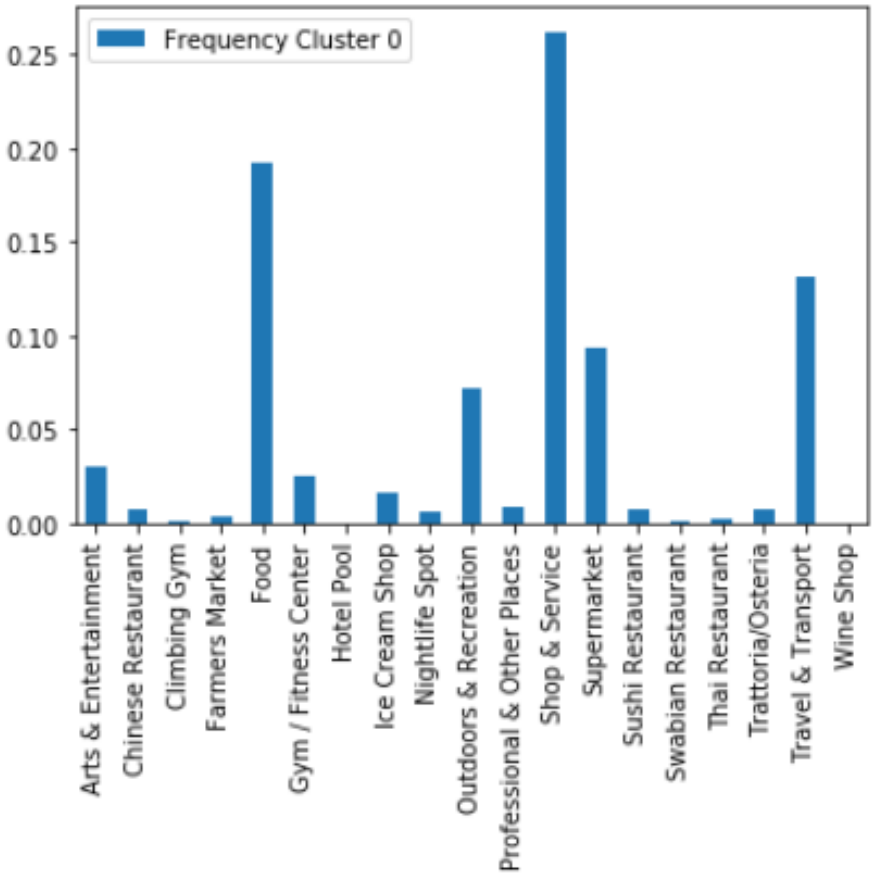

Fig. 4: Venues frequency of centroid of cluster 0

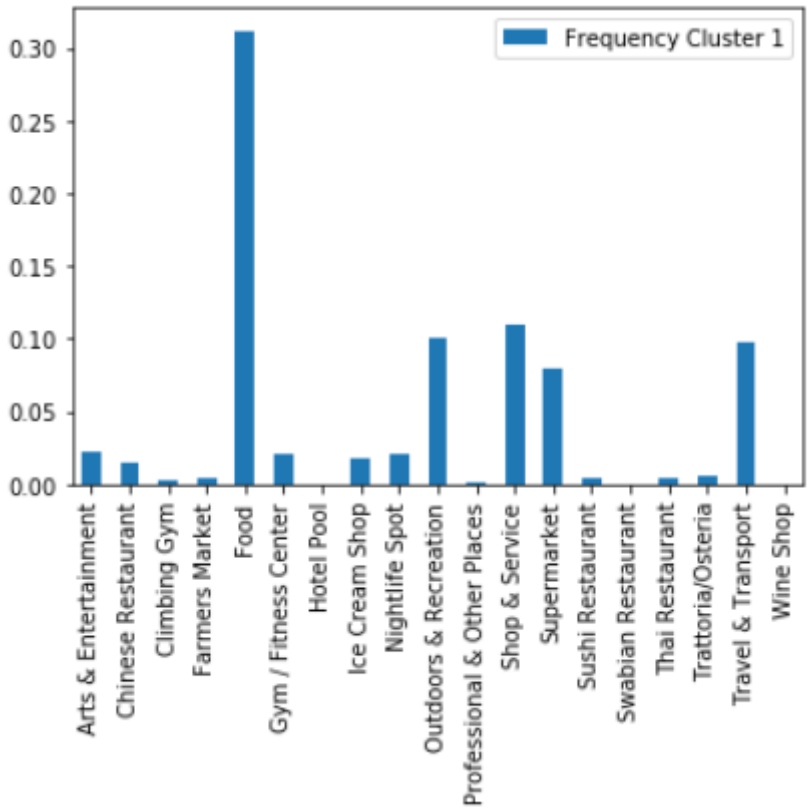

Fig. 5: Venues frequency of centroid of cluster 1

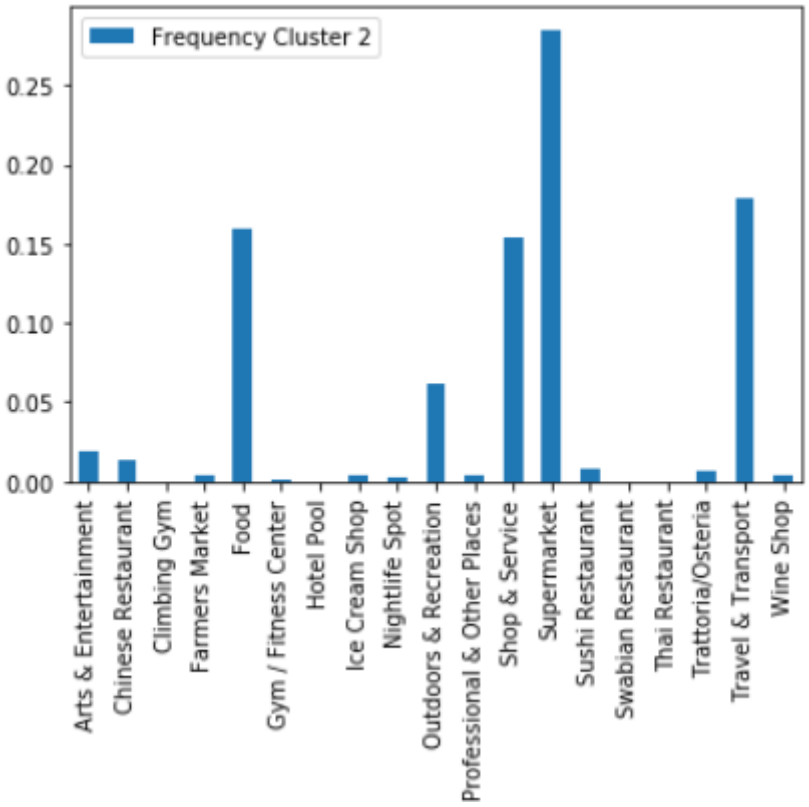

Fig. 6: Venues frequency of centroid of cluster 2

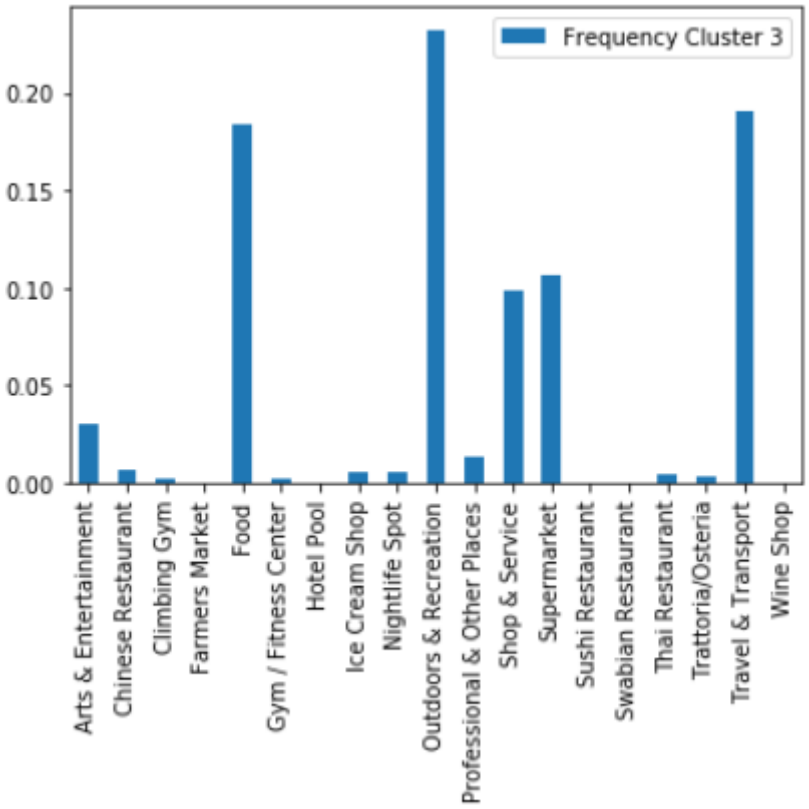

Fig. 7: Venues frequency of centroid of cluster 3

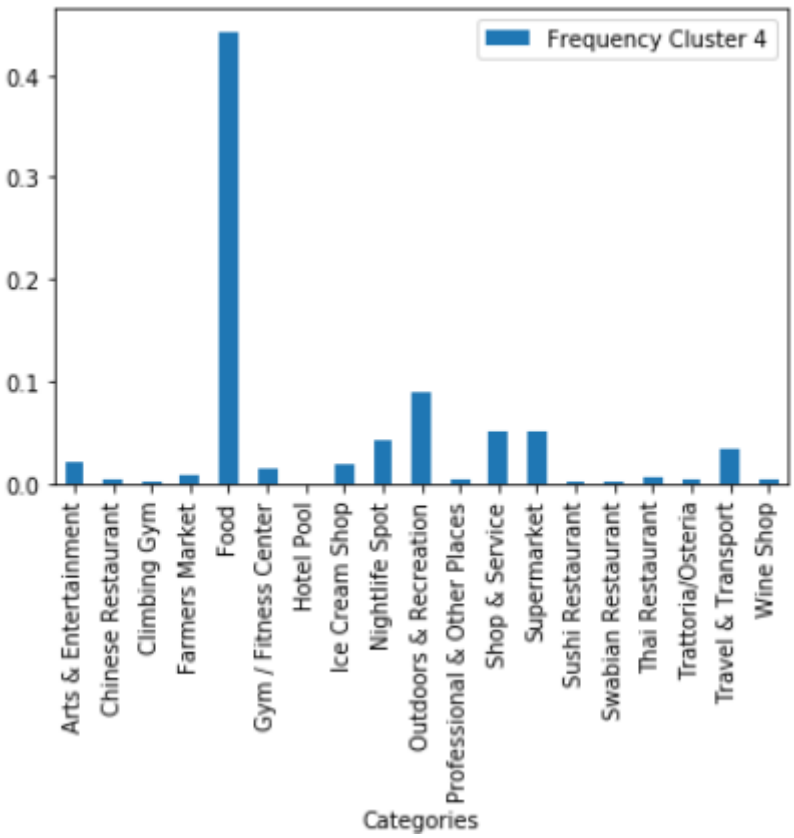

Fig. 8: Venues frequency of centroid of cluster 4

Cluster 4 includes the reference neighborhood and therefore represents the candidate cluster. As expected, it shows a very high frequency of foord venues compared to all other categories. All other cluster show a much more even distribution. Even cluster 1, where the frequency of food venues is also quite pronounced does not show a similar level of bias.

3.3 Choosing a neighborhood

The candidate cluster still contains a total of 35 neighborhoods. First, let's look where those 35 neighborhoods are located:

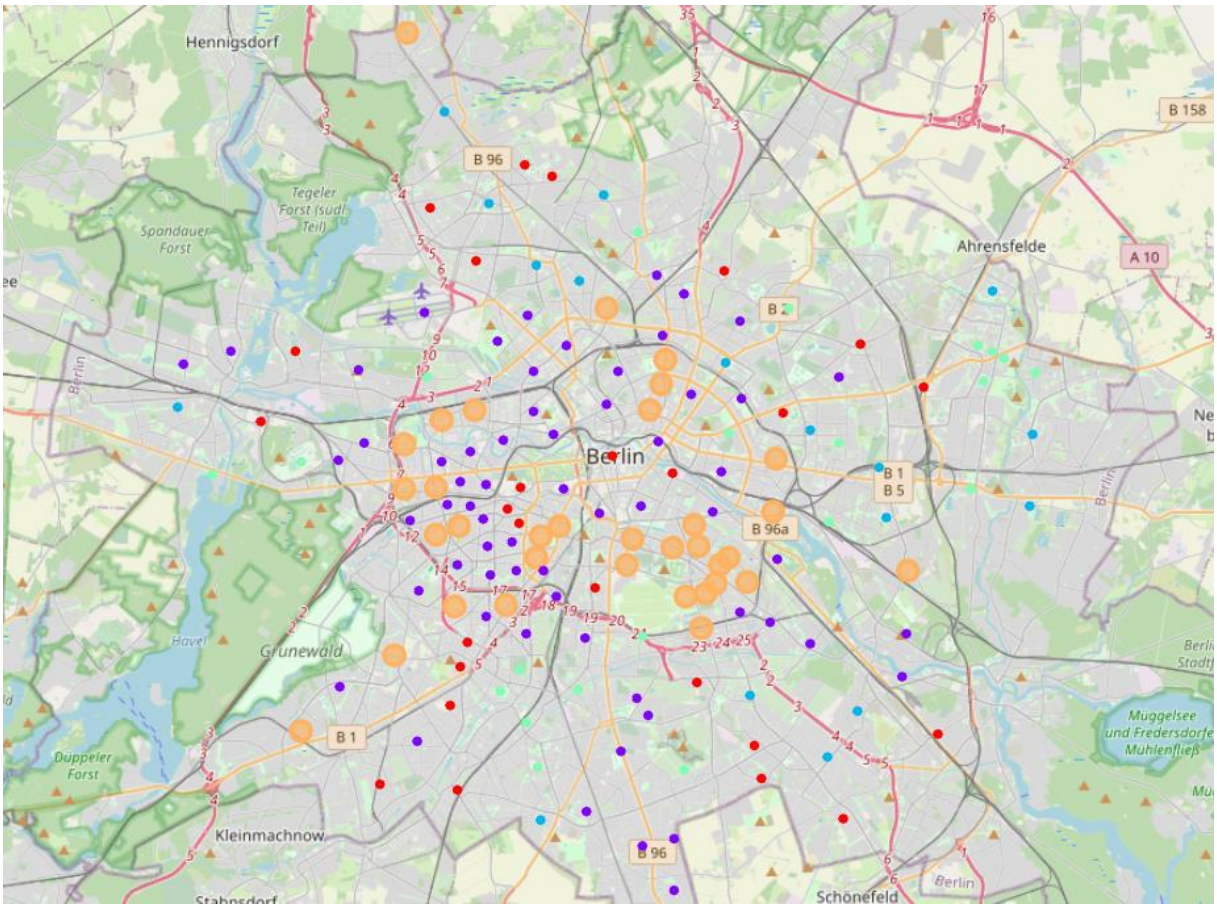

Fig. 9: Candidate neighborhoods at destination (marked by large circles)

While they are spread out all over the Berlin city limits, it can be seen that they are mostly on the outer limits of the inner city. This is very interesting as the chosen reference neighborhood is also on the outskirts of Stuttgart. To further narrow down the candidate areas, I go back to using the location data. By sorting the neighborhoods ascending by their distance from the chosen destination point, we can chose the 5 closest neighborhoods which are similar to the reference neighborhood:

	Neighborhood	Post Code	Distance
22	Berlin Mitte	10119	2.380342
11	Berlin Schöneberg	10783	2.644311
32	Berlin Prenzlauer Berg	10435	3.190693
17	Berlin Schöneberg	10781	3.206165
15	Berlin Kreuzberg	10961	3.239111

Table 1: Final candidate neighborhoods

4. Results

Summing up the results, one can make the following observations:

- The reference neighborhood Stuttgart Vaihingen is mainly characterized by an abundance of food venues
- The average neighborhood at the destination (Berlin) has a much more even distribution of venues
- Using K-Means Clustering can be used to find neighborhoods at the destination that are very similar to the reference neighborhood
- In total 35 neighborhoods in Berlin would be a suitable new place to live at for someone who enjoyed living in Stuttgart Vaihingen
- To further reduce the number of candidate neighborhoods, the distance from the chosen destination point is used

5. Discussion

Based on the results, any of the five neighborhoods mentioned above would be a good choice. To really make a decision, one should likely consider additional aspects. E.g. one could now look at average real estate prices in order to find the cheapest good option. The real insight of the data analysis lies within the data of the reference neighborhood. Here we can see, that the main point of attraction is the large number of food venues.

6. Conclusion

Looking at the data, the conclusion can be made extremely short: I like food and if I ever move, I should move to a neighborhood with many food options. If Berlin was to be my next destination, luckily there would be many good options. Interestingly, similar to the neighborhood I chose as a reference, these neighborhoods are also mainly on the outskirts of the inner city.