# The Battle of Neighborhoods – Staff Deployment

## 1 Introduction

### 1.1 Background

In an increasingly globalized world, internationally operating companies are required to post their employees to different locations all over the world. There are multiple reasons for staff deployment across all types of industries, e.g. due to a lack of qualified personnel at foreign locations. Further reasons might be knowledge sharing across different locations, international project teams, improving the communication between the headquarter and abroad locations and many more.

### 1.2 Business Problem & Stakeholders

The goal of this project is to provide valuable insights and support for human resources departments that have to deal with staff deployment in terms of the following aspects and questions respectively:

- Based on the personal preferences of an employee, which neighborhood(s) of the target city should be chosen to book a hotel or rent a project apartment for the duration of the deployment abroad?

- An employee might prefer a neighborhood with similar culture and lifestyle compared to where he/she currently lives.
  On the other hand, employees might deliberately be looking for a specific neighborhood that differs from the one they currently live in.

- Enabling employees to live in a neighborhood they prefer, might result in increased willingness to be posted abroad as well as increased employee satisfaction.

- Are there neighborhoods that should be avoided due to very high crime rates?

The issues above will be addressed by segmenting and clustering neighborhoods based on venue data reflecting the culture and lifestyle of the respective area.

In general, the investigations performed in this project might be useful for all kinds of companies sending staff to locations abroad.

In this project, let us assume there is a Munich-based company regularly sending employees to locations in Detroit and Mexico City respectively for one of the reasons outlined in 1.1.

As Munich is known as a safe city with a very low crime rate, the company would like to make sure to post their employees only to areas with a low to moderate crime rate. Thus, prior to the clustering, neighborhoods with significant high crime (rates) are supposed to be excluded from the datasets.

## 2 Data

### 2.1 Data Sources

The following data sources will be used to solve the problem described in section 1.2.

| | |
|---|---|
| City | Munich |
| Description | Dataset containing **Munich districts** |
| Source | Open Data Portal Munich |
| Link | https://www.opengov-muenchen.de/dataset/bevoelkerung-stadtbezirken |
| Format | CSV file |

| | |
|---|---|
| City | Munich |
| Description | Dataset containing **Munich neighborhoods** |
| Source | Open Data Portal Munich |
| Link | https://www.opengov-muenchen.de/dataset/bevoelkerung-stadtbezirksteile-muenchen |
| Format | CSV file |

| | |
|---|---|
| City | Detroit |
| Description | Dataset containing **Detroit districts and neighborhoods** (including shape data) |
| Source | City of Detroit Open Data Portal |
| Link | https://data.detroitmi.gov/datasets/neighborhoods |
| Format | KML file |

| | |
|---|---|
| City | Detroit |
| Description | Dataset containing **crime incidents in Detroit (2019)** |
| Source | City of Detroit Open Data Portal |
| Link | https://data.detroitmi.gov/datasets/rms-crime-incidents |
| Format | CSV file |

| | |
|---|---|
| City | Mexico City |
| Description | Dataset containing **Mexico City districts and neighborhoods** |
| Source | Data Portal of Mexico City |
| Link | https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/ |
| Format | CSV file |

| City | Mexico City |
|---|---|
| Description | Dataset containing **crime incidents in Mexico City (2019)** |
| Source | Data Portal of Mexico City |
| Link | https://datos.cdmx.gob.mx/explore/dataset/carpetas-de-investigacion-pgj-cdmx/table/ |
| Format | CSV file |

In addition to the data sources displayed in the table above, the following sources are used to retrieve data:

- Foursquare API
- geopy library

### 2.2 Data Preparation

#### 2.2.1 Neighborhood Data

The *explore venue* endpoint of the Foursquare API is used to retrieve venue data for the neighborhoods of Munich, Detroit and Mexico City based on geographic coordinates. Thus, the required data fields comprise the following.

- district
- neighborhood
- latitude coordinate of neighborhood
- longitude coordinate of neighborhood

**Munich**

In contrast to Detroit and Mexico City, the district and neighborhood data for Munich needs to be imported from two separate files.
While geographic coordinates for Detroit and Mexico City are basically provided in the files that also contain the district and neighborhood data, the latitude and longitude data for the Munich neighborhoods is retrieved using the *Nominatim* class of the *geopy* library. The geographic coordinates can be obtained via reverse geocoding, i.e. latitude and longitude are returned based on the neighborhood address in string format.

**Detroit**

As can be seen from the table above, data for Detroit is contained in a KML file. This file will be converted to geojson format using *kml2geojson* library.
The required data will be extracted from the geojson file and saved in a data frame. Latitude and longitude values for Detroit will be derived from the polygon shape data utilizing the *shapely* library.

**Mexico City**
Both district and neighborhood data as well as the corresponding geographic latitude and longitude data is available in a single CSV file for Mexico City.

### 2.2.2 Crime Data

In addition to the neighborhood data, crime data is required for Detroit and Mexico City as neighborhoods with significantly high crime (rates) are supposed to be excluded from the datasets prior to the clustering.

The required data for crime incidents in Detroit and Mexico City is obtained from CSV files according to the table above.

Crime data will be treated slightly different for the two cities.
For Detroit, a crime rate will be calculated based on the number of incidents per neighborhood divided by the area (in acres). Thus, the crime information is given on a neighborhood level.
In contrast, when it comes to Mexico City the number of crime incidents will be used to analyze crime on district basis due to unsatisfying data quality.

### 2.2.3 Data Overview

Please see the following examples of the required data for each city.

**Munich**

| district | neighborhood | latitude | longitude |
|---|---|---|---|
| Altstadt – Lehel | Graggenau | 48.139616 | 11.579513 |

**Detroit**

| district | neighborhood | acres | latitude | longitude | incidents | crime rate |
|---|---|---|---|---|---|---|
| 1 | Castle Rouge | 223.66 | 42.381679 | -83.268125 | 207 | 0.93 |

**Mexico City**

| district | neighborhood | latitude | longitude |
|---|---|---|---|
| Azcapotzalco | Petrolera | 19.485156 | -99.199994 |

| district | incidents |
|---|---|
| Álvaro Obregón | 8324 |

# 3 Methodology

## 3.1 Neighborhood Data

As mentioned in the Introduction section, the goal is to segment and cluster neighborhoods of Munich, Detroit and Mexico City in order to find similar areas based on venue information.
It turned out to be important to keep both the neighborhood *and* the district information as there are neighborhoods assigned the same name in multiple districts in Detroit and Mexico City.
Based on the information provided by the data sources shown in section 2, the cities are composed as follows.

| | Districts | Neighborhoods |
|---|---|---|
| Munich | 25 | 106 |
| Detroit | 7 | 207 |
| Mexico City | 16 | 1739 |

The number of neighborhoods incorporated in the clustering however will be reduced due to a variety of reasons, e.g. neighborhoods will be removed from the datasets due to high crime rate (cf. section 3.2) or due to only very little number of venues nearby the respective area.

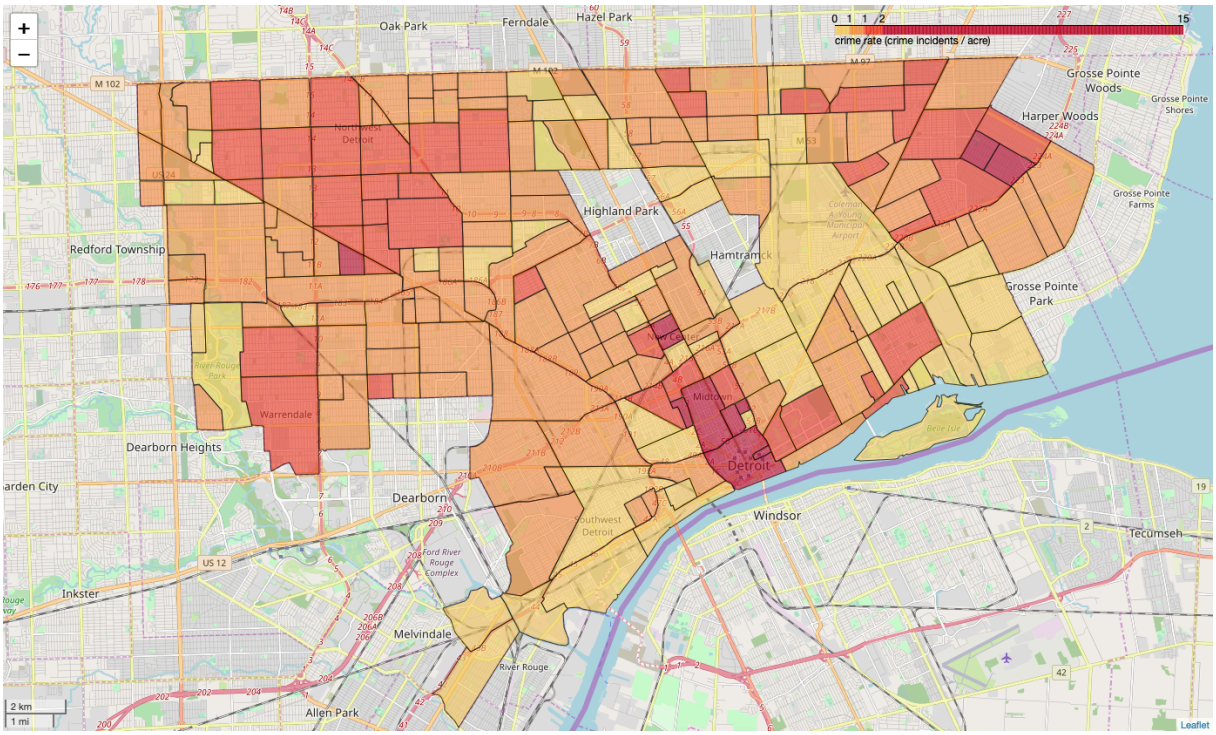
## 3.2 Analysis of Crime Data

Before clustering the neighborhoods, crime (rate) data is to be analyzed in order to identify areas with high crime rates. The idea is to exclude the respective neighborhoods from the datasets and as a result not to consider them in the clustering algorithm. Thus, such highly criminal areas will not be proposed to an employee that is about to be sent abroad.

In a first step, crime incidents can be visualized by creating a choropleth map generated using the *folium* library.

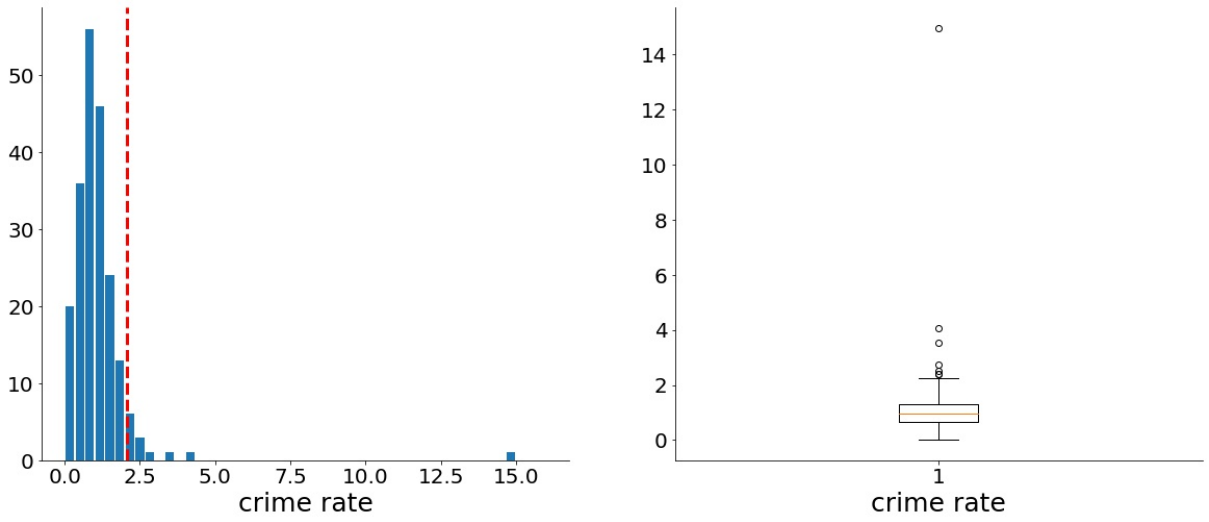
For Detroit, the crime rate calculated as described in section 2.2.2 as data. The polygonial shape data is taken from the geojson file mentioned in section 2.2.1.

The resulting choropleth map looks as follows.



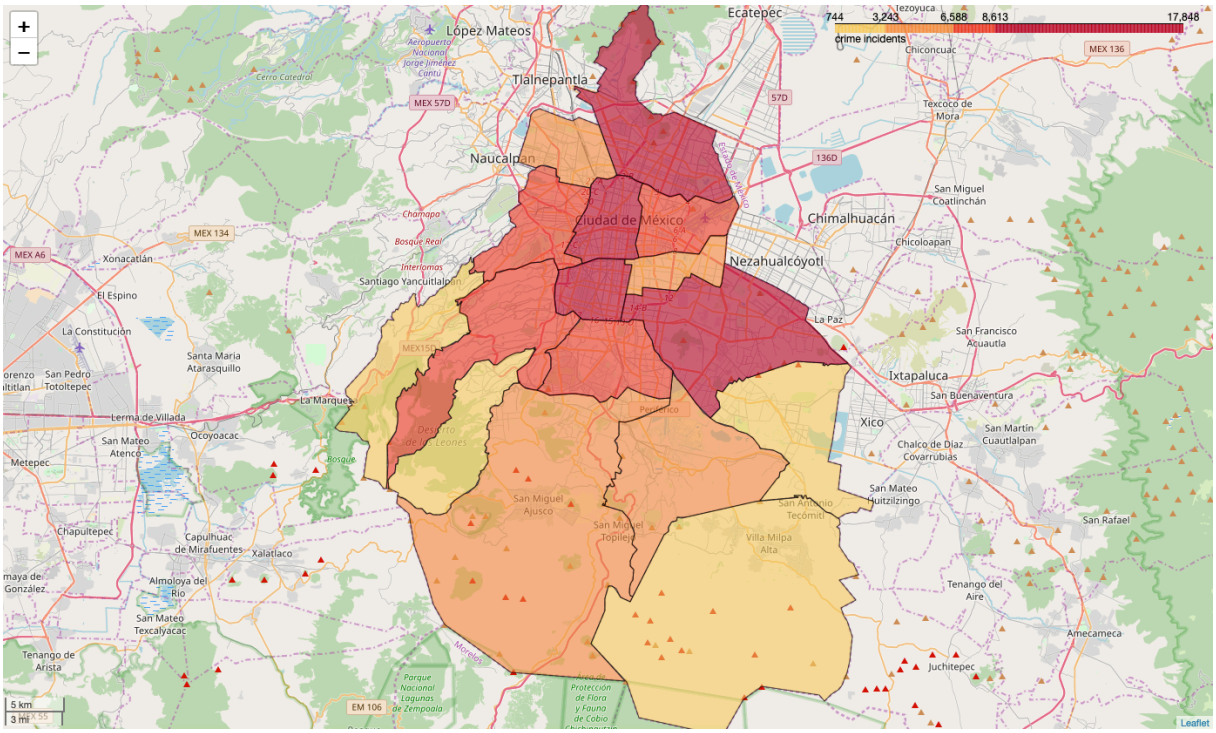

The neighborhoods having a crime rate (total number of crime incidents per acre) that exceeds the 95% quantile are reflected by the dark red color.
There are in total eleven neighborhoods that represent the worst 5% of the data sample with a crime rate higher than 2.07 crimes per acre.

Setting up a histogram and a boxplot yield some more detailed insights with respect to the distribution of the crime rate in Detroit.
As can be seen in the figure below it appears reasonable to exclude neighborhoods above the 95% quantile, which is indicated through the red dotted vertical line in the histogram.



The same analysis is applied to Mexico City. However, the absolute number of crime incidents on a *district level* is used as a measure for crime.

The corresponding choropleth map showing the 16 districts of Mexico City looks as follows.




This time, it seems reasonable to take the 75% quantile as a threshold. Looking at the districts that exceed the 75% quantile we end up with Benito Juárez, Cuathémoc, Gustavo A. Madero and Iztapalapa. As the number of crime incidents in these districts is significantly higher compared to the other districts, all respective neighborhoods that are located within the four identified districts will be excluded from the neighborhood dataset.

### 3.3 Clustering Neighborhoods

Finally, the remaining neighborhoods are supposed to be segmented and clustered based on venue information retrieved via the Foursquare API.
Specifically, the explore venue endpoint is used to make API calls and obtain a response string in json format containing venue information. The relevant information for all neighborhoods in the three cities is then extracted and written into pandas data frames.
Beside the API credentials, the latitude and longitude coordinates of the respective neighborhood as well as the radius and a limit parameter needs to be passed.

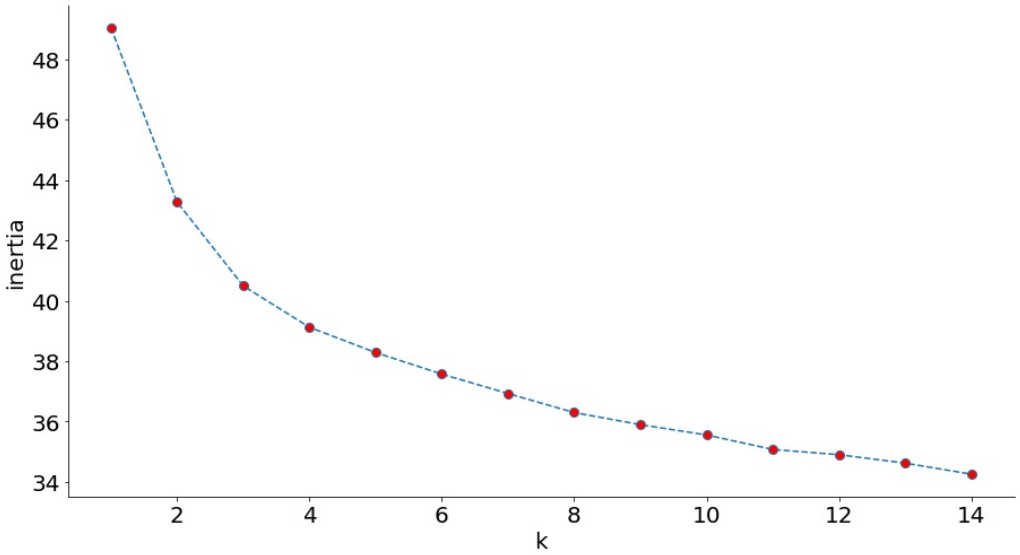
It is particularly important to specify a reasonable value for the radius depending on the medium size of neighborhoods in the respective cities.

It has been decided to use a radius of 500 meters for Munich, 700 meters for Detroit and 400 meters for Mexico City based on the map visualizations.

The most important information, i.e. the field that is used as features for the clustering is the venue category, i.e. Bar, Park or Restaurant.
The procedure is to apply one hot encoding and some further adjustments so that finally a data frame is available indicating the frequency of a specific venue category on a neighborhood level.

A few more adjustments are necessary prior to applying k-means algorithm. All neighborhoods with an extracted number of venues less than 10 are not considered in the clustering.

In order to find a good value for k, the elbow method is applied, i.e. the sum of squared distances is calculated for a range of k values. Plotting the results implies a value of k = 4 as can be seen from the figure below.




Finally, the cluster labels returned from the k-means algorithm are put into a data frame along with the 10 most common venues of the neighborhoods.

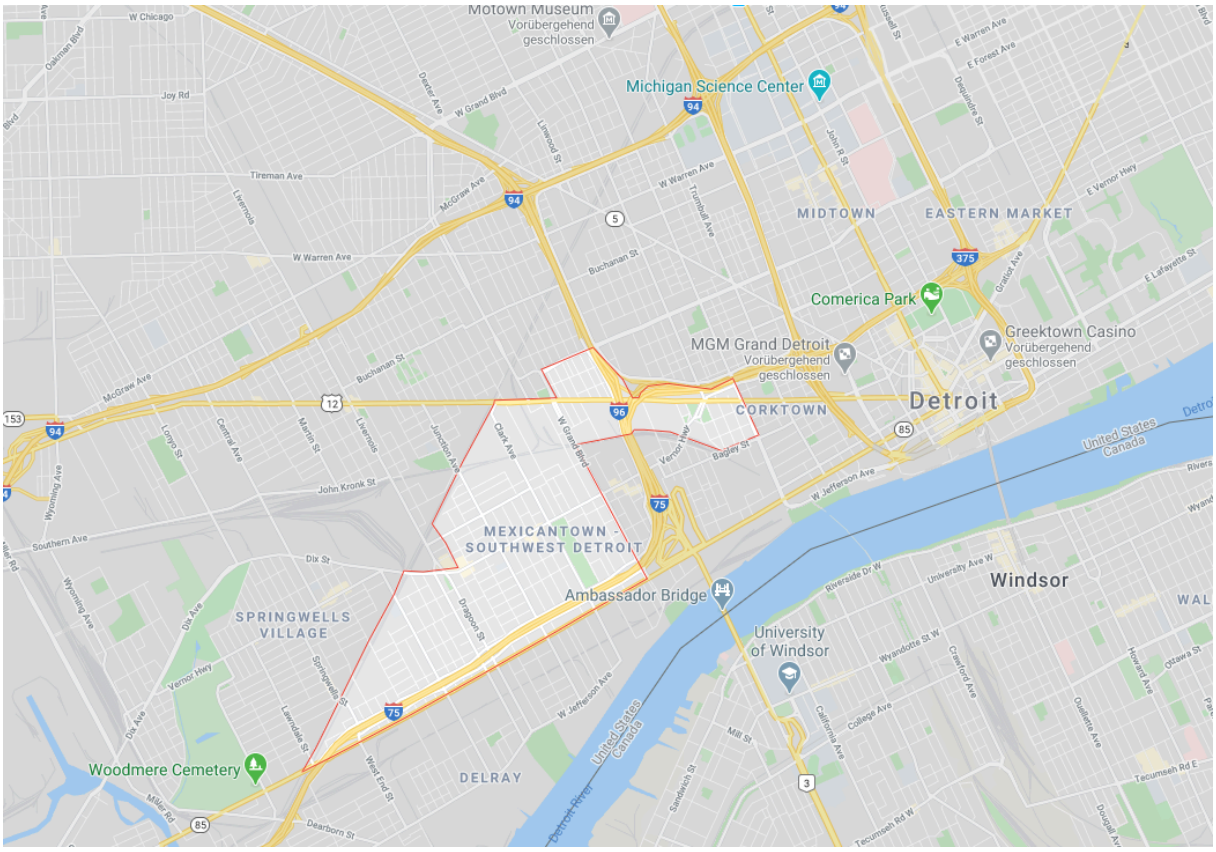
## 4. Results[1]

The resulting clusters can be visualized with a map indicating the neighborhoods in a color that belongs to their cluster.

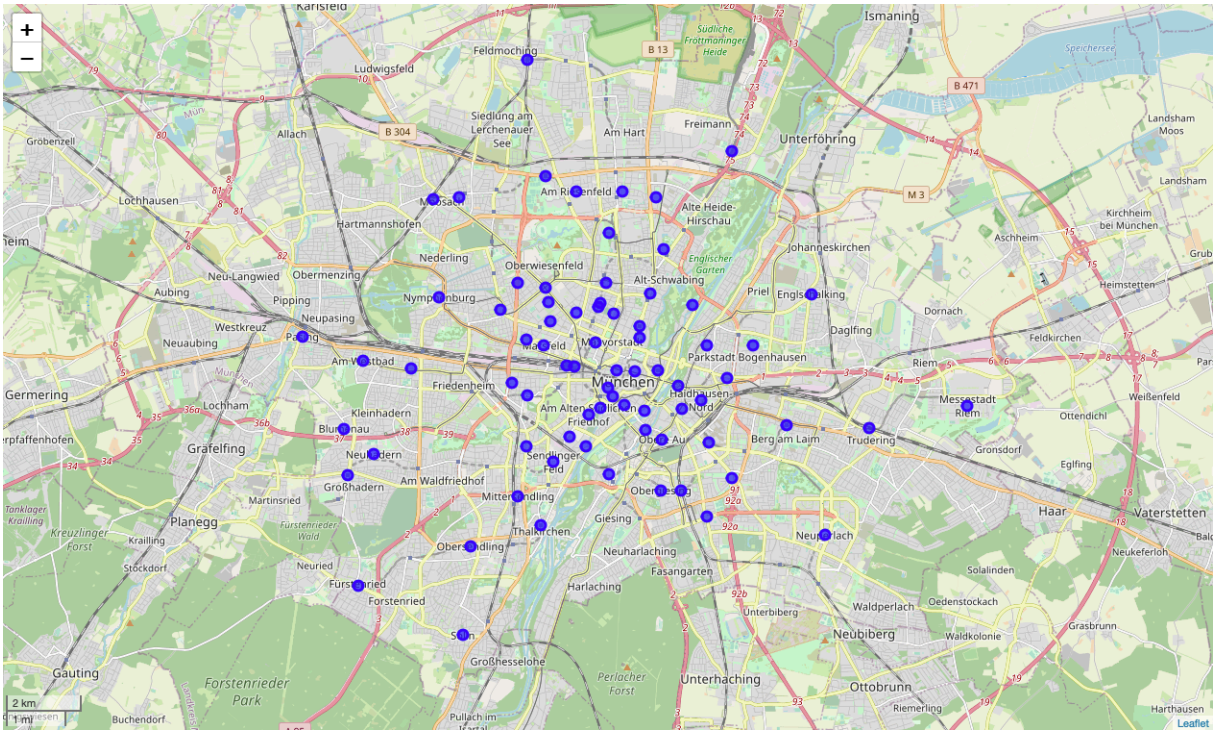
[1] Please note that the results and cluster names/colors differ from the Python notebook as re-running the code might lead to slightly different results.

As mentioned in the previous section, the neighborhoods have been clustered into four separate clusters based on the venue categories retrieved from the Foursquare API.
Looking at the generated maps one can draw the following conclusions:

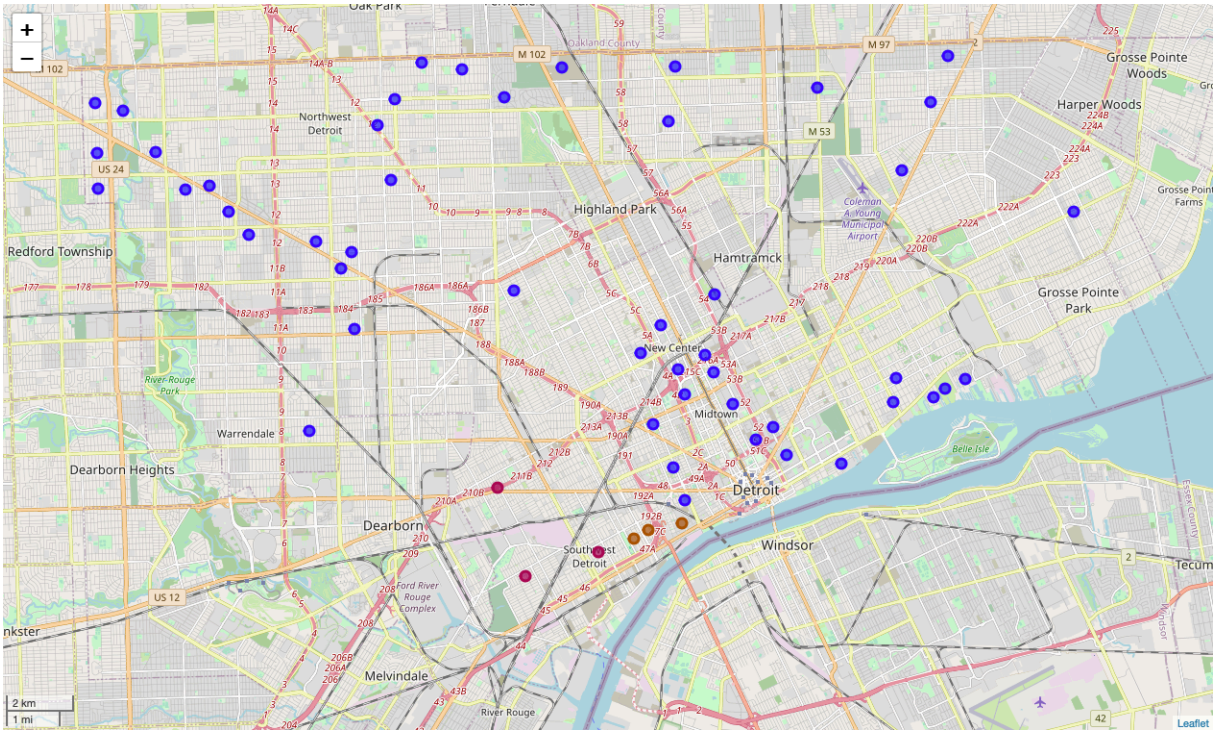
1. In general, Munich seems to be more like Detroit (in comparison to Mexico City)
2. It is remarkable that six neighborhoods in Detroit are clustered in Cluster 3 and 4. Beside these neighborhoods, there are only Mexican City neighborhoods within these clusters. This is plausible as the six Detroit neighborhoods are located close by Detroit's Mexican Town.



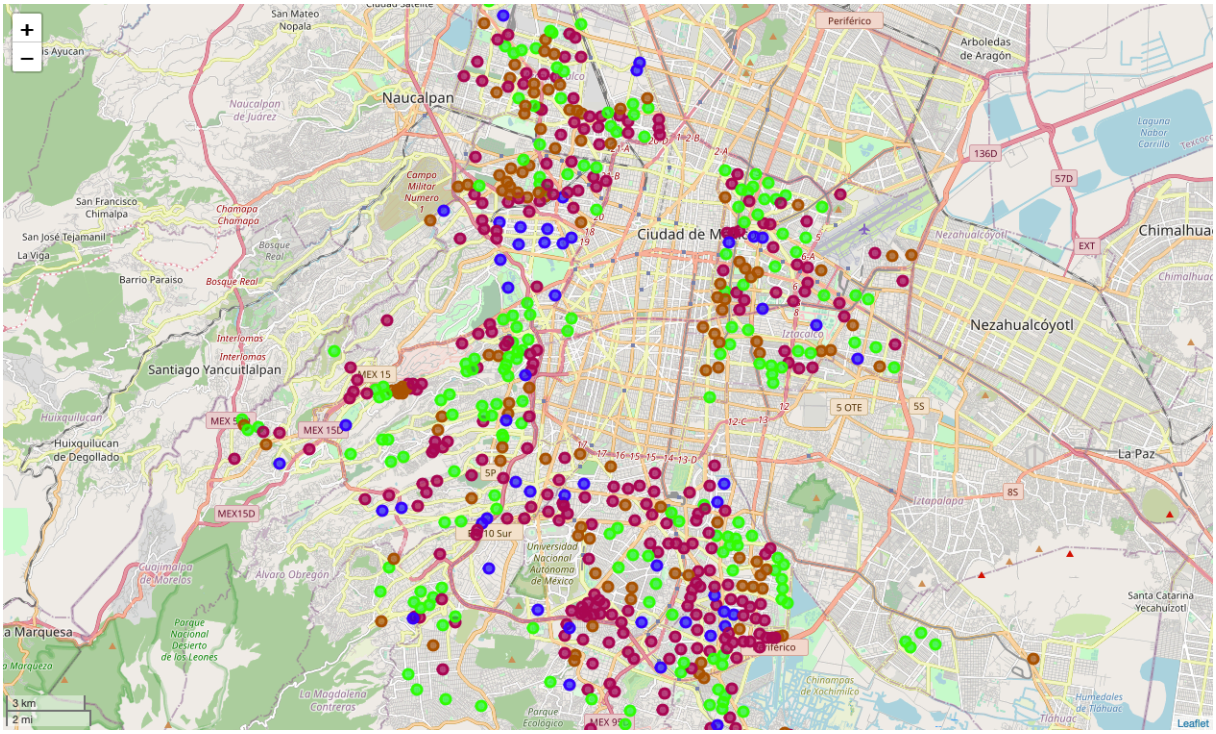
## Munich




## Detroit



**Mexico City**

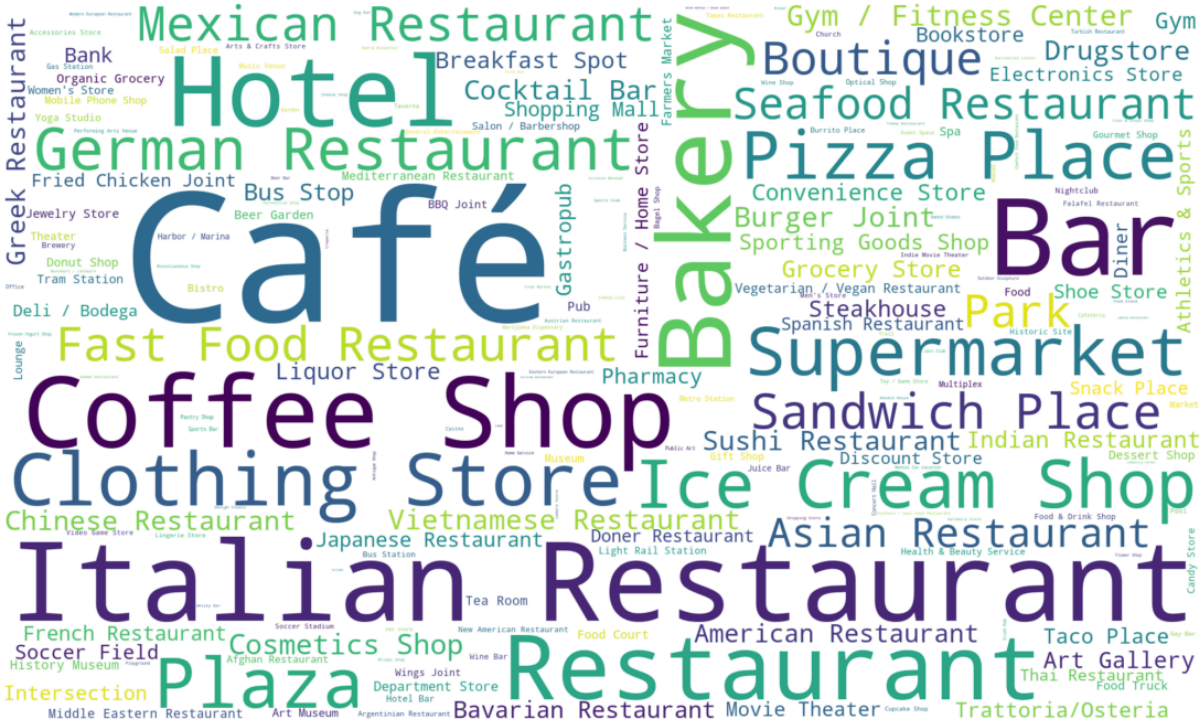



In addition to the maps above, word clouds have been generated for all clusters based on the frequency of venue categories within a specific cluster. Looking at the word clouds helps to get an impression of the neighborhood's characteristics.

**Cluster 1**

Cluster 1 is a ‘pure’ Mexico City cluster as only neighborhoods that belong to Mexico City are clustered within this cluster. Considering the word cloud emphasizes the most dominant venues, i.e. Mexican Restaurant and Taco Place.

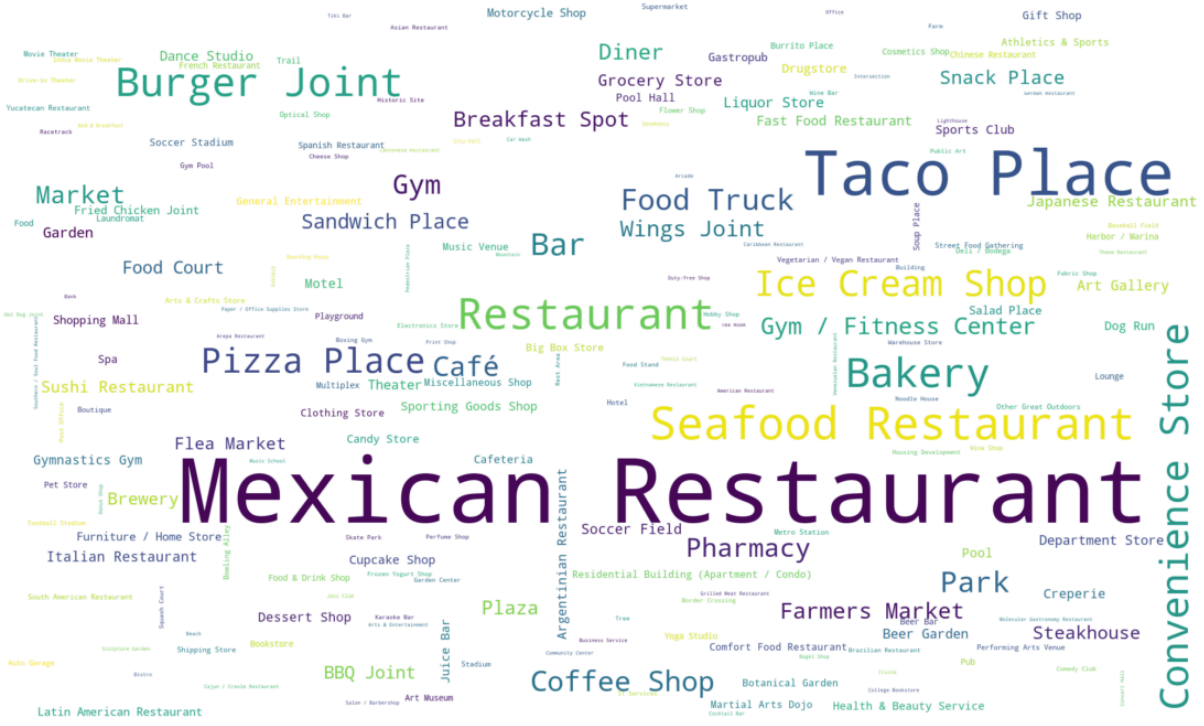
**Cluster 2**




All neighborhoods of Munich and most of Detroit’s neighborhoods are clustered in Cluster 2 reflecting the similarity of these two cities opposed to Mexico City. The word cloud and as a result the most common venues differ significantly from the other three clusters.

**Cluster 3**


**Cluster 4**



## 5. Discussion & Conclusion

It turned out to be quite challenging to cluster neighborhoods of different cities due to a variety of reasons. The structure, i.e. the area and total number of neighborhoods differ significantly for the three cities that have been investigated which might have an impact on the clustering results.

Furthermore, there is a huge number of features (venues categories) retrieved from the Foursquare API. It might be reasonable to apply dimensionality reduction or feature selection methods respectively in order to reduce the large number of features to get better results.

Also, the use of a more advanced clustering algorithm, i.e. density-based clustering might be reasonable as k-means is a quite basic approach with a couple of drawbacks.

Nevertheless, the results presented in the previous section gives good insights in terms of which areas to avoid when posting employees abroad as well as where to find neighborhoods in Mexico City and Detroit that are similar (or dissimilar) to the Munich infrastructure and characteristics respectively.