
Capstone Project
The Battle of the Neighborhoods

INTRODUCTION
STAFF DEPLOYMENT – POSTING EMPLOYEES ABROAD

- 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

  - Lack of qualified personnel at foreign locations
  - Knowledge sharing across different locations
  - International project teams
  - Improving the communication between headquarter and abroad locations
  - …

Which neighborhood of the target city should be chosen to book a hotel or rent a project apartment for the duration of the deployment abroad based on the personal preferences of the employee?

An employee might prefer a neighborhood with similar culture and infrastructure compared to where he/she currently lives ?!

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

Enabling employees to live in a neighborhood they prefer might result in increased motivation to work abroad

PROCEDURE
DATA JOURNEY

1 Import and Prepare Neighborhood Data

District and Neighborhood data (including geographic coordinates) for Munich, Detroit and Mexico City is imported and prepared

2 Import and Prepare Crime Data

Crime (rate) data is imported and prepared for Detroit and Mexico City

3 Visualize and Analyze Neighborhoods

Prepared Neighborhood Data for Munich, Detroit and Mexico City is analyzed and visualized using *folium* maps

4 Analyze Crime

Crime (rates) are analyzed for Detroit (on neighborhood level) and Mexico City (on district level)

5 Clustering Neighborhoods

Venue Data retrieved via Foursquare API is used to segment and cluster neighborhoods of the three cities using k-means algorithm

6 Examine Clusters

Generated clusters are analysed using visualization tools like maps and word clouds

DATA

MAIN DATA SOURCES

CITY OF DETROIT OPEN DATA PORTAL

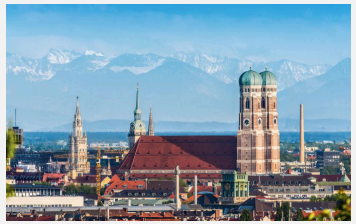
CITY OF MUNICH DATA PORTAL

DATA PORTAL MEXICO CITY

FOURSQUARE API

DATA

DATA PREPARATION



MUNICH

- District and neighborhood data needs to be imported from two separate files
- Geographic coordinates are retrieved using the *Nominatim* class of the *geopy* library (reverse geocoding based on the neighborhood address in string format)
- No crime data is analyzed as considered to be a safe city

DETROIT

- District and neighborhood data is provided in a KML file that is converted to geojson format using *kml2geojson* library
- Latitude and longitude coordinates will be derived from the polygon shape data utilizing the *shapely* library
- Crime data (number of incidents) is provided in a separate CSV file and available on neighborhood level
- Calculation of crime rate (incidents per acre)

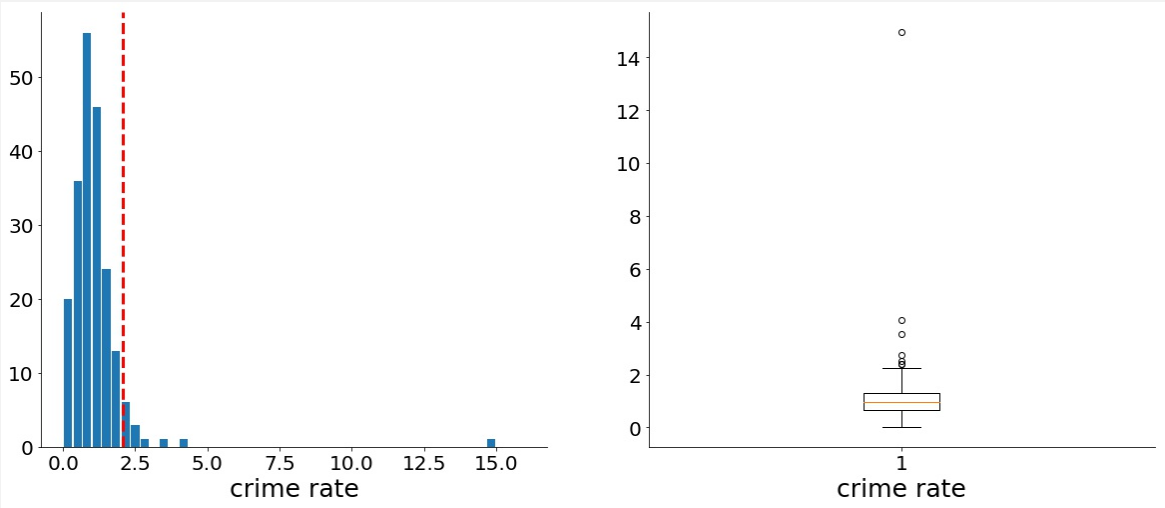
MEXICO CITY

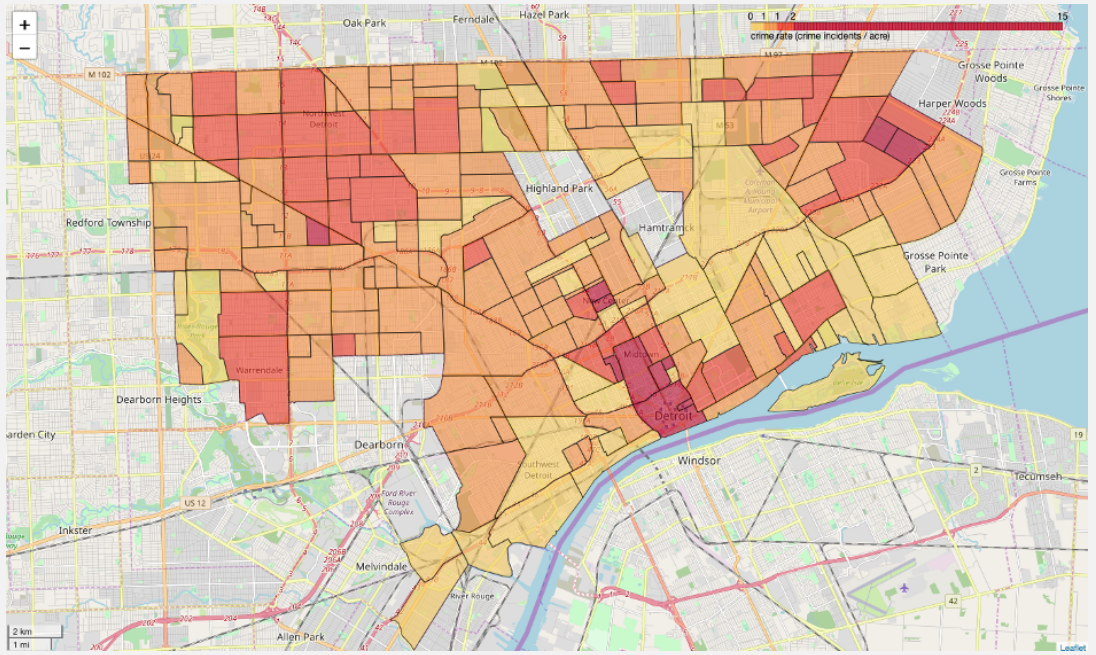
- Both district and neighborhood data as well as the corresponding geographic coordinates are provided in a single CSV file
- Crime data only available on district level
- In comparison to Detroit, the absolute number of crime incidents is used to analyze crime

CRIME ANALYSIS (1)
DETROIT

Statistical Analysis of Crime Rate in Detroit

The figure is showing a histogram and boxplot of the crime rate data of Detroit. Looking at the histogram it appears reasonable to exclude neighborhoods above the 95% quantile, which is indicated through the red dotted vertical line.





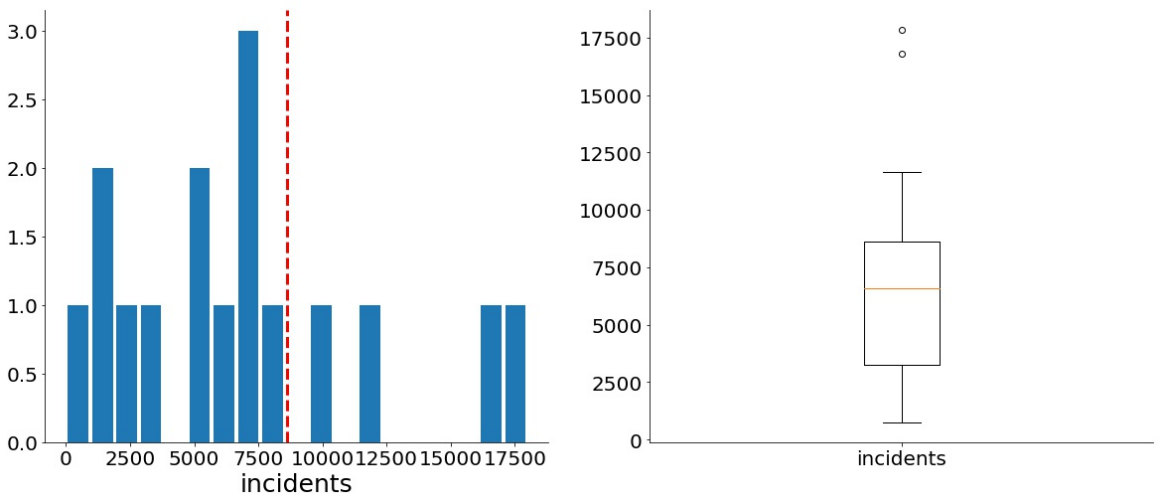
Choropleth Map of Crime Rate in Detroit

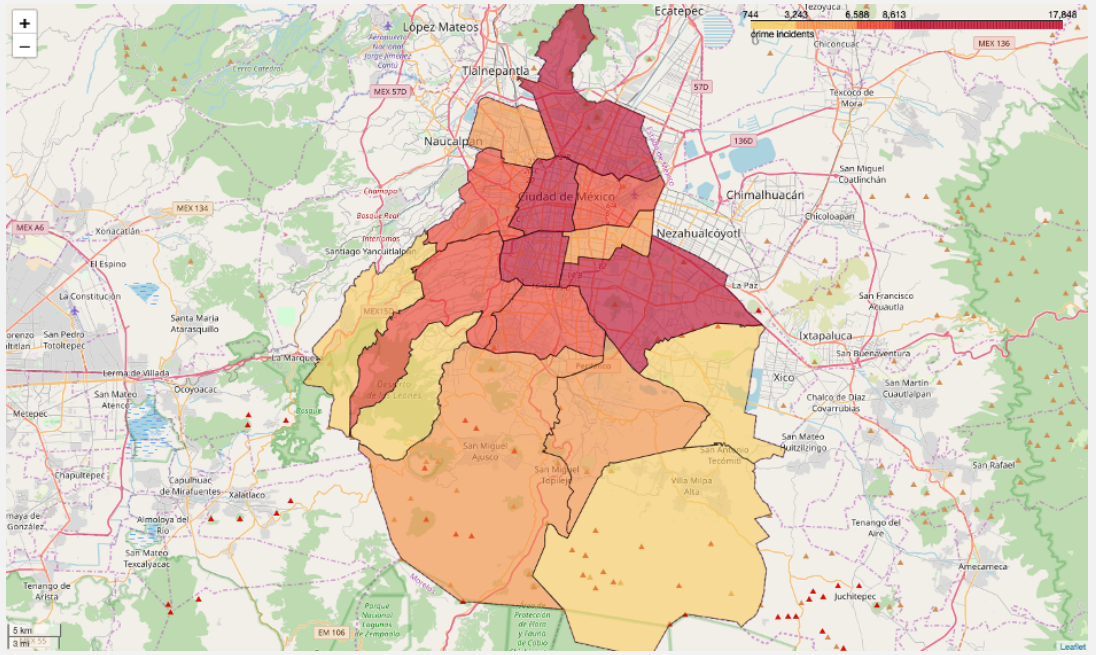
Neighborhoods having a crime rate that exceeds the 95% quantile are reflected by the dark red color. There are in total eleven neighborhoods that represent the worst 5% of the data sample with a crime rate higher than 2.07 crime incidents per acre

CRIME ANALYSIS (2)
MEXICO CITY

Statistical Analysis of Crime Rate in Mexico City

The figure is showing a histogram and boxplot of the crime incident data of Mexico City.
In contrast to Detroit, it seems reasonable to take the 75% quantile as a threshold. This threshold is indicated through the red dotted vertical line.





Choropleth Map of Crime in Mexico City

In the case of Mexico City, one can see the 16 districts rather than the neighborhoods. Crime analysis has been performed on district level due to data quality issues. There are in total four districts exceeding this threshold, namely Benito Juárez, Cuathémoc, Gustavo A. Madero and Iztapalapa.

METHODOLOGY
CLUSTERING NEIGHBORHOODS

GET VENUE DATA

- Venue data is retrieved via the Foursquare API using the explore venue endpoint to make API calls
- The data is obtained as a response string in json format and the relevant data is stored in pandas data frames

DATA MANIPULATION

- In order to guarantee satisfying clustering results, neighborhoods with less than ten venues are removed from the data sets
- Application of one hot encoding and calculation of the frequency for a specific venue on a neighborhood level
- Calculation of the top 10 venues per neighborhood

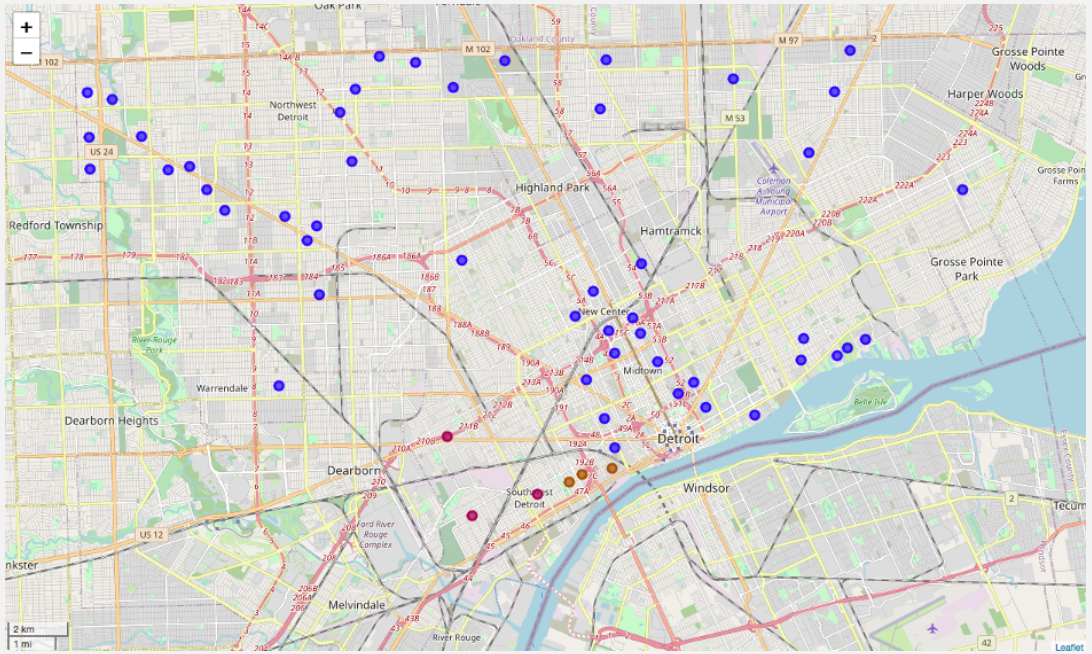
APPLY K-MEANS

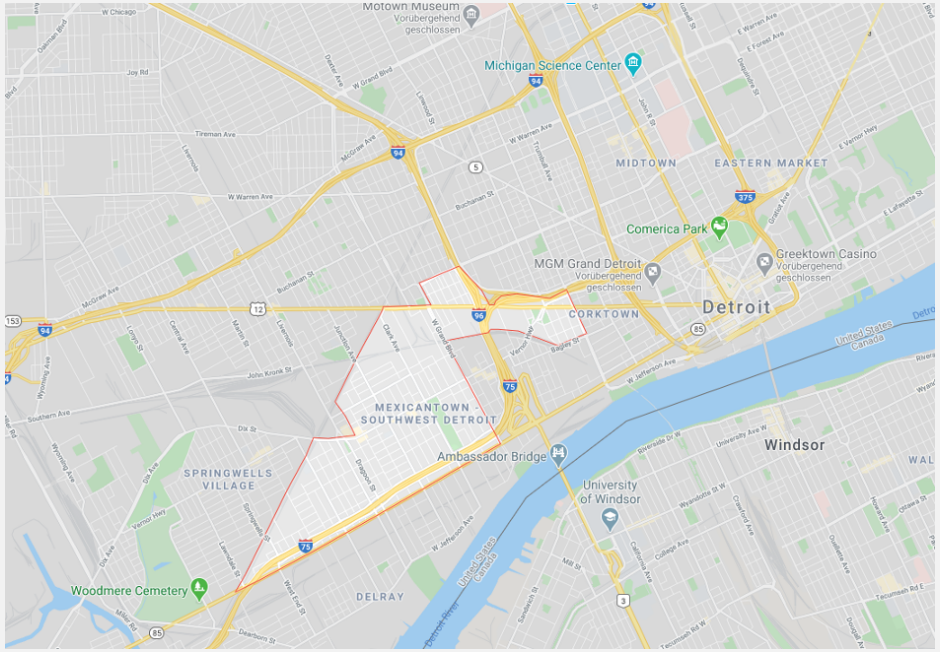
- K-means algorithm is applied in order to cluster neighborhoods based on venue frequencies
- Inertia, i.e. the sum of squared distances is calculated for a range of k values to figure out the optimal value for k
- The optimal value is k = 4 which has been used for the clustering of the neighborhoods

RESULTS (1)
CLUSTER MAP - DETROIT

Cluster Distribution Detroit

The neighborhoods of Detroit have been clustered into three clusters, Cluster 2, Cluster 3 and Cluster 4. However, note that the great majority of neighborhoods are within Cluster 2 which is also the cluster where all the Munich districts are in (see next slide).





Mexican Town in Detroit (Source: Google Maps)

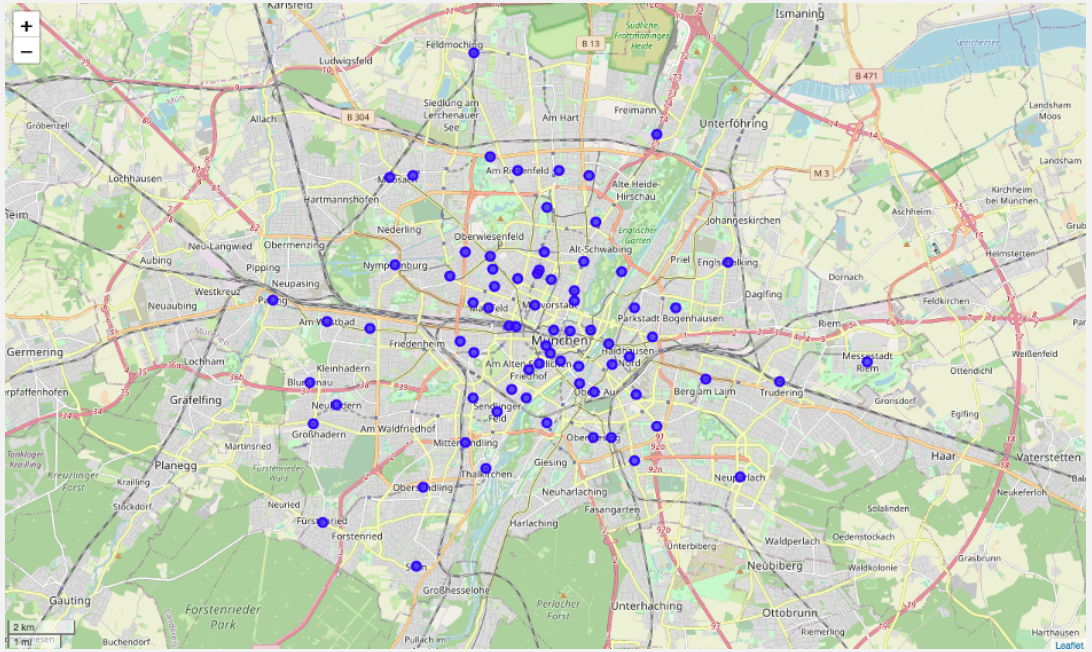
One can perfectly see, that the six neighborhoods that have been clustered in Cluster 3 and Cluster 4 are located in or close by the Mexican town district in Detroit.

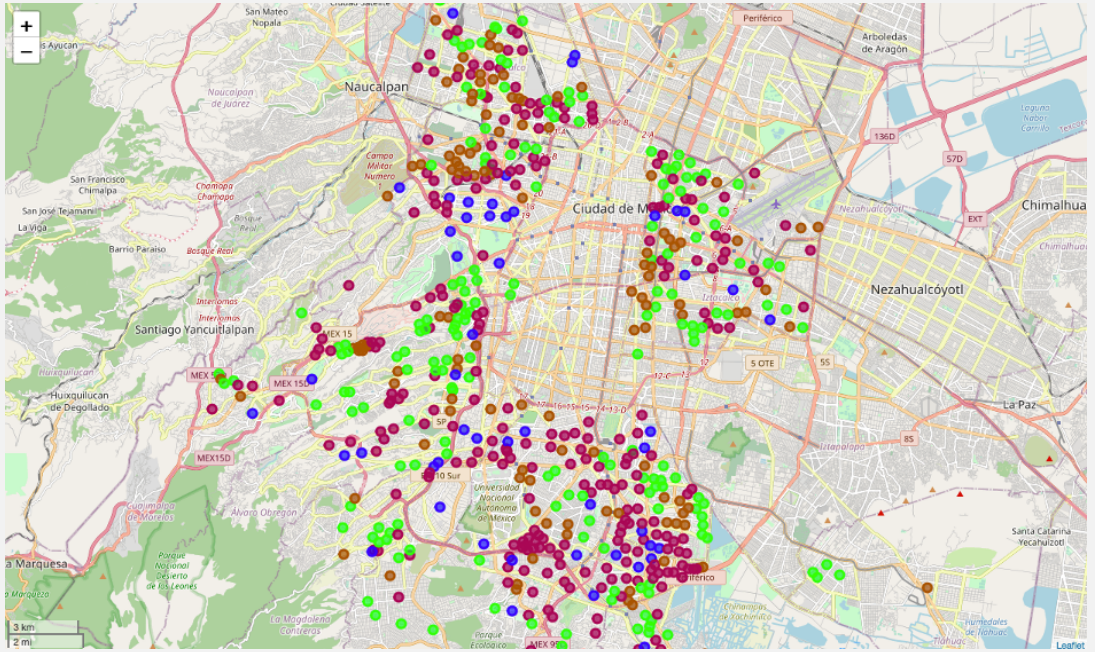
* Please note that the results and cluster names / colors differ from the Python notebook as re-running the code might lead to slightly different results

RESULTS (2)
CLUSTER MAP - MUNICH & MEXICO CITY

Cluster Distribution Munich

All neighborhoods in Munich have been clustered together in Cluster 2. Neighborhoods within this cluster can also be found in Mexico City as well as Detroit. From the next slide, one can see that most of Detroit's neighborhoods are in the same cluster. Thus, we can draw the conclusion that in general, Munich is more like Detroit than to Mexico City.





Cluster Distribution Mexico City

The neighborhoods of Mexico City are spread over all four clusters.

RESULTS (3)

CLUSTER – WORD CLOUDS

– The word clouds have been generated 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 characteristic

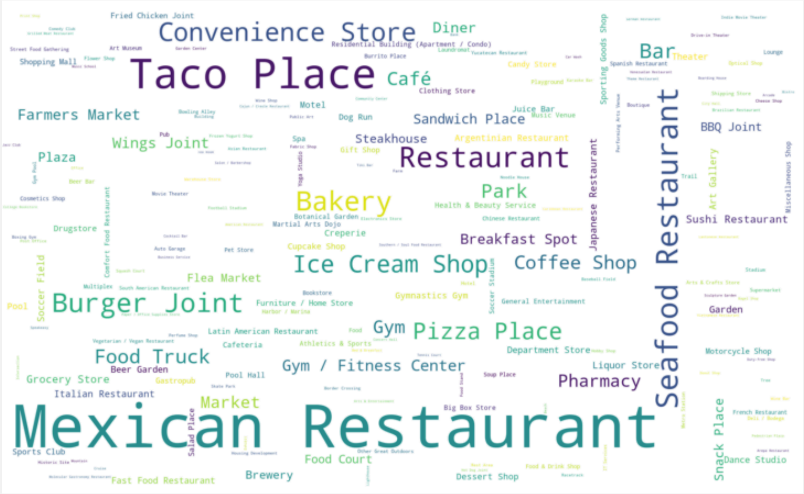
– It is remarkable that the word cloud of Cluster 2 (where all Munich and most of Detroit's neighborhoods are clustered in) differs significantly from the other three word clouds

CLUSTER 1

CLUSTER 2

CLUSTER 3

CLUSTER 4



DISCUSSION & CONCLUSION

GENERAL USE CASE

- 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
- In general, the presented results give good insights in terms of which areas to avoid when posting employees abroad
- From the results, one can identify neighborhoods in Detroit and Mexico City that are similar (or dissimilar) to the Munich infrastructure and characteristics respectively

FEATURE SELECTION

- A huge number of features (venue categories) has been retrieved via 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 and reduce complexity

CLUSTERING ALGORITHM

- 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