Applied Data Science

Capstone Project  
The Battle of Neighborhoods/Boroughs

Relocation to Munich based on crime, housing prices and venues of Munichs boroughs

Final report by Alexander Fuchs | 25th May, 2020

*Figure 1: Munich, alps in the background (https://upload.wikimedia.org/wikipedia/commons/4/40/Vista\_panorámica\_desde\_Olympiapark%2C\_Múnich%2C\_Alemania\_2012-04-28%2C\_DD\_03.JPG)*

# INTRODUCTION

## 1.1 Background

|  |  |
| --- | --- |
| [Munich](https://en.wikipedia.org/wiki/Munich) (German: München (can also be written as Muenchen)) is the capital and most populous city of Bavaria, the second most populous German state. With a population of around 1.5 million, it is the third-largest city in Germany, after Berlin and Hamburg, and thus the largest which does not constitute its own state, as well as the 11th-largest city in the European Union. The city's metropolitan region is home to 6 million people [cite from [Wikipedia.org](https://en.wikipedia.org/wiki/Munich)] | *Figure 2: Coat of arms of Munich (https://upload.wikimedia.org/ wikipedia/commons/1/17/ Muenchen\_Kleines\_Stadtwappen.svg)* |

Munich ranked the [most livable city of the world](https://www.dw.com/en/munich-tops-list-of-worlds-most-livable-cities-berlin-and-hamburg-also-in-top-10/a-44478511) back in 2018, where "[t]he analysts acknowledged the high cost of living in the Bavarian capital but insisted residents get their money's worth, citing the city's 'excellent infrastructure with a booming economy' and 'proximity to nature with appreciation for culture.'" [cite from [DW.com](https://www.dw.com/en/munich-tops-list-of-worlds-most-livable-cities-berlin-and-hamburg-also-in-top-10/a-44478511)]. Especially the "high cost of living" is fueled by the high costs for housing. Just recently Munich lost its 'top spot' as most expensive city in Germany after 20 years on top. [cite from [DW.com](https://www.dw.com/en/stuttgart-unseats-munich-as-germanys-most-expensive-city-for-renters/a-51374468)]. Furthermore, not only is Munich one of the most livable cities of the world based on economy and nature, but also based on the fact, that Munich is one of the safest cities in Europe [cite from [DW.com](https://www.dw.com/en/munich-tops-list-of-worlds-most-livable-cities-berlin-and-hamburg-also-in-top-10/a-44478511)].

## 1.2 Business problem

The project aims to select a borough (not neighborhood; explanation see below) for people who want to relocate to Munich. The selection process is driven by **three main factors**:

1. **Safety**, or in other words the **number of crimes**: Although one of the safest cities in Europe, Munich also has its "dark spots" and therefore safety is a major factor.
2. **Housing prices**: Since Munich is one of the most expensive cities in Germany finding affordable housing is crucial. Especially when the costumer intendeds to stay for a longer period of time and wants to buy an apartment, the cost per square meter of an apartment is a major factor.
3. **Venues**: Imagine living in the safest and most affordable borough of a city. But no venues of interest around. Based on that, venues are considered a major factor as well.

Since this is a multi-factor problem a way to determine a borough has to be obtained:

1. First get a overview of the city and its boroughs.
2. Combine the insights with relevant data for crime per borough.
3. Extend the crime date by general population data to come to a more reasonable solution.
4. Extend and combine the data with data for housing prices.
5. Get a factor for 'best combination' of crime data and housing prices
6. Search for venues for each borough.
7. Cluster boroughs based on their top venues.
8. Determine the safest and most affordable borough for each cluster.

## 1.3 Interest

The interest in this business problem stems from the fact, that in the last few decades the population of Munich steadily increased. Based on that, finding an affordable living space which is located in a safe borough, which also has a certain amount of venues of interest for the customer, is on the one hand side a demanding but on the other hand side also very rewarding task and business idea. Especially for international customers who are not familiar with the somewhat uncommon ways of German rental, this business stems a crucial support in finding a place to call home.

## DATA

## 2.1 Data acquisition

Based on the definition of the business problem different data sets are needed.

1. [Crime statistics](https://www.muenchen.de/rathaus/dam/jcr:6291ac42-463d-4267-b436-c4b1a3313454/jt190904.pdf): The city of Munich and its police department publish data for crimes only on a borough level (see Figure 3 a), **not** on a neighborhood level (see Figure 3 b). Therefore, all following investigations are performed on a borough level since no date on a smaller scale (neighborhood) is available [Data ported from a pdf-file of [Muenchen.de](https://www.muenchen.de/rathaus/dam/jcr:6291ac42-463d-4267-b436-c4b1a3313454/jt190904.pdf) into an csv- file. csv-file used for further investigations].

|  |  |
| --- | --- |
|  |  |
| *Figure 3 a. Boroughs of Munich (https://upload.wikimedia.org/wikipedia/commons/2/2d/ Karte\_der\_Stadtbezirke\_in\_München.png) Stadtbezirke von München = German for Boroughs of Munich* | *Figure 3 b: Neighborhoods of Munich (https://upload.wikimedia.org/wikipedia/commons/e/e8/ Karte\_der\_Stadtbezirksteile\_in\_München.png) Stadtbezirksteile von München = German for Neighborhoods of Munich* |

1. [General population data](https://de.wikipedia.org/wiki/Stadtbezirke_Münchens): General population data as population per borough and/or neighborhood, population density per borough, area of the borough, percentage of foreigners per borough to find a diverse borough are needed [Data imported from [Wikipedia.org](https://de.wikipedia.org/wiki/Stadtbezirke_Münchens)].
2. [Housing prices](https://www.immobilienscout24.de/neubau/ratgeber/aktuelle-neubau-themen/neubau-kauf-map-wohnungen-muenchen-2019.html): Housing prices are taken and imported from one of the leading and most prominent German real estate sites [Immobilienscout24.de](https://www.immobilienscout24.de/). Since this site has a vast number of apartments offers in all boroughs and also its prices, it is a good bases for housing prices in Munich. The prices are Euro per square meter, when buying a three-room apartment with about 80 square meters of area [Data imported from [Immobilienscout24.de](https://www.immobilienscout24.de/neubau/ratgeber/aktuelle-neubau-themen/neubau-kauf-map-wohnungen-muenchen-2019.html)].
3. Venues: Venues are taken/imported from [FourSquare.com](https://de.foursquare.com/) using the API of FourSquare.

## 2.2 Data preview, pre-processing, and cleaning

As mentioned in section 2.1, the used data is not coherent. First of all, the data about crimes in Munich is only available on a borough level and **not** on a neighborhood level. Therefore, no statements can be made on the smaller scale of neighborhoods. It was decided, that for this project the data on a borough scale is sufficient enough and will be used for further investigations.

Also, the latest available crime statistics are from 2018. No newer data regarding crimes is publicly available at the moment of this investigation/report (May 2020). Since information on media outlets suggest that the overall number of [crimes kept constant or fell](https://www.sueddeutsche.de/muenchen/muenchen-polizei-kriminalitaetsstatistik-2019-1.4843394), the crime data of 2018 is used in this investigation and seen as appropriate, because the data from 2018 is either at the same level as nowadays or represents a ‘worst-case’ scenario.

Another challenge is the way the crime statistics are published, or to be more precise the data format. The crime data is published in a [pdf-file](https://www.muenchen.de/rathaus/dam/jcr:6291ac42-463d-4267-b436-c4b1a3313454/jt190904.pdf) which is not immediately readable by Python/Pandas. Therefore, the data was pre-processed in a way that the containing rows and columns are translated to a csv-file for further investigation. This csv-file is imported using Pandas. Afterwards a first and preliminary factor of interest of this study is generated: the ‘*Total crimes*’. Generated because the initial crime statistics only provides ‘*Total major crimes*’ (which in itself is the sum of a multitude of smaller crime categories also listed in the data set; see Figure 4) and ‘*Total other crimes*’ but not the sum of both.

|  |
| --- |
|  |
| *Figure 4: Pre-processed and cleand crime statistics of Munich in 2018 on a borough level. The ‘Total crimes’ as sum of ‘Total major crimes’ and ‘Total other crimes’ was added to the data set.* |

Therefore, and consequently, the factor of ‘*Total crimes*’ is generated by summing up both of the beforementioned categories. This way a statement on a borough level regarding all crimes committed in 2018 can be made.

After the crime data set is available is has to be merged with geographical data scraped for each borough and Munich as a whole (see Figure 5). This geographical data is provide by queries to the *geopy* library for Python (see Figure 6). This data set is especially necessary for the later performed task of using the FourSquare API to get venues for each borough. For this task for each and every borough the geographical position has to be known to be able to perform a ‘nearby’ search, which in itself uses the ‘center-position’ of each borough. This ‘center-position’ is defined by latitude and longitude data and readable by *folium* (a Python library for visualization) as well as by the FourSquare API.

|  |
| --- |
|  |
| *Figure 5: Crime data set enhanced by geographical data of latitude and longitude of each borough.* |

Next general population data as population per borough and/or neighborhood, population density per borough, area of the borough, percentage of foreigners are scraped from its corresponding [Wikipedia.org](https://de.wikipedia.org/wiki/Stadtbezirke_Münchens) page.

|  |
| --- |
|  |
| *Figure 6: The city of Munich with its boroughs marked (shown is the ‘center-point’ position of each borough).* |

This data set (see Figure 7) is used to provide further insight into each borough as well as normalize certain other parameters by an appropriate factor like population to end up with parameters on a ‘*per-capita*’ basis.

|  |
| --- |
|  |
| *Figure 7: General population data from* [Wikipedia.org](https://de.wikipedia.org/wiki/Stadtbezirke_Münchens). |

After pre-processing this data set, it is merged with the previously shown data set to form a more comprehensive data set.

[Housing prices](https://www.immobilienscout24.de/neubau/ratgeber/aktuelle-neubau-themen/neubau-kauf-map-wohnungen-muenchen-2019.html) are scraped from a real estate site called [Immobilienscout24.de](https://www.immobilienscout24.de/), which is of the leading and most prominent German real estate sites. Since this site has a vast number of apartments offers in all boroughs and also its prices, it is a good bases for housing prices in Munich. The prices are Euro per square meter, when buying a three-room apartment with about 80 square meters of area. Unfortunately, the housing prices are listed on a neighborhood level (see Figure 8 b and Figure 9) and not on a borough level (see Figure 8 a). Therefore, the housing prices are imported into a Pandas data frame and pre-processed for further use in later investigations. Especially the housing prices on a borough level are generated. By computing the average of the housing prices of all neighborhoods in a borough the housing prices on a borough level are provided.

|  |  |
| --- | --- |
|  |  |
| *Figure 8 a. Boroughs of Munich (https://upload.wikimedia.org/ wikipedia/commons/2/2d/ Karte\_der\_Stadtbezirke\_in\_München.png)* | *Figure 8 b: Neighborhoods of Munich and corresponding housing prices (https://www.immobilienscout24.de /content/dam/is24/neubau/bilder/neubau-kauf-maps/muenchen-2019-gross.jpg)* |

After pre-processing this data set, it is merged with the previously shown data set to form a more comprehensive data set.

|  |
| --- |
|  |
| *Figure 9: Head of housing prices spraced from the real estate site.* |

Last but not least venues data (e. g. Figure 10) is used and imported/scraped from FourSquare. For each borough within a radius of 100 meters all venues are imported using a FourSquare query using their API. This data set on venues is sorted afterwards and

|  |
| --- |
|  |
| *Figure 10: Head of venues on a borough basis proviede by FourSquare and their API.* |

# METHODOLOGY

## 3.1. Exploratory data analysis

Before the crime data set is further processed an initial exploratory data analysis is performed. Within this analysis the ‘Top 5’ and ‘Bottom 5’ boroughs regarding crimes (see Figure 11 for an overview) are concluded. Afterwards the crime data is normalized with the population per borough to get a parameter ‘*Total crimes per capita*’ on a borough level to make the boroughs and the corresponding crime data more comparable. As an additional side note the distribution of crimes in major categories is investigated for the ‘safest’ borough based on the parameter ‘*Total crimes per capita*’.

|  |
| --- |
|  |
| *Figure 11: ‘Total crimes’ for all boroughs.* |

The housing prices, after imported and pre-processed to fit the borough level of detail also used in the crime data analysis, are investigated on both a neighborhood and borough level. The initial and additional investigation on a neighborhood level of detail mainly aims toward an initial preview of the housing prices data as well as a form of a plausible check for the later used and generated housing prices data on a borough level.

### 3.1.1.’Top 5’ boroughs regarding the total number of crimes

Sorting the crime data set descending for the ‘Total crimes’ leads to the ‘Top 5’ boroughs of Munich (see Figure 12) with the highest absolute number of crimes committed in 2018. The borough of ‘Ludiwigsvorstadt-Isarvorstadt’ shows to be the borough of Munich with the highest number of total crimes.

|  |
| --- |
|  |
| *Figure 12: ‘Top 5’ boroughs of Munich considering the total number of crimes.* |

### 3.1.2.’Bottom 5’ boroughs regarding the total number of crimes

Sorting the crime data set ascending for the ‘Total crimes’ leads to the ‘Bottom 5’ boroughs of Munich (see Figure 13) with the lowest absolute number of crimes committed in 2018. The borough of ‘Allach-Untermenzing’ shows to be the borough of Munich with the lowest number of total crimes.

|  |
| --- |
|  |
| *Figure 13: ‘Bottom 5’ boroughs of Munich considering the total number of crimes.* |

## 3.2. Normalize crime data by population/capita

To take the different population sizes of each borough into account when discussing the crimes of each borough, a new parameter, the ‘*Total crimes per capita*’, is computed. Using the previously discussed ‘*Total crimes*’ data for each borough, this parameter is put into relation to the population size of each borough. By dividing the ‘*Total crimes*’ by the ‘*Population*’ and afterwards multiplying it with 100’000 a parameter ‘*Total crimes per capita*’ (here per 100’000 citizens) is computed. Figure 14 provides an overview of this parameter for all boroughs of Munich.

|  |
| --- |
|  |
| *Figure 14: ‘Total crimes per capita’ for all boroughs of Munich. ATTENTION: the y-scale is logarithmical.* |

Analog to the absolute values of ‘*Total crimes*’ the ‘Bottom 5’ boroughs of Munich with the lowest number of ‘*Total crimes per capita*’ are concluded by sorting this parameter in an ascending order. Figure 15 shows, that the former safest borough of Munich, ‘Allach-Untermenzing’ is still in the ‘Bottom 5’ of safest boroughs, but not the safest borough of them all. This spot takes now the borough of ‘Hadern’. A closer look into the crime statistics of the borough of ‘Hadern’ reveals, that the majority of crimes committed in 2018 in this borough are of a minor scale, and only very few are major crimes, as Figure 16 clearly states.

|  |
| --- |
|  |
| *Figure 15: ‘Bottom 5’ boroughs of Munich considering ‘Total crimes per capita’.* |

|  |
| --- |
|  |
| *Figure 16: Crimes committed as well as their category in the Munich borough of ‘Hadern’ in the year 2018.* |

## 3.3 Housing prices

[Housing prices](https://www.immobilienscout24.de/neubau/ratgeber/aktuelle-neubau-themen/neubau-kauf-map-wohnungen-muenchen-2019.html) are taken and imported from one of the leading and most prominent German real estate sites [Immobilienscout24.de](https://www.immobilienscout24.de/). As mentioned in the data section, the housing prices are published on a neighborhood level of detail (see. Figure 17) and range from 5335 Euro per square meter on the outskirts of Munich up to 13095 Euro per square meter in the city center.

Unfortunately, the date for housing prices is given based on neighborhoods of Munich not based on boroughs. Since the focus lies on boroughs (crime data is only available on a boroughs level) and not on neighborhoods, a transition from neighborhood level of housing prices to a borough level of housing prices has to be performed. This is done by averaging the housing prices in a borough by the values of the neighborhoods of this borough (see Figure 18). As additional information and side note also the increase of housing price in the neighborhoods/boroughs is provide and given in per cent; but not further taken into account. After the transition to housing prices on a borough level of detail, this data is merged with the previously shown and used Pandas data frame to get a coherent and consistent data set.

|  |
| --- |
|  |
| *Figure 17: Housing prices on a neighborhood scale for the city of Munich*. |

|  |
| --- |
|  |
| *Figure 18: Housing prices on a borough scale for the city of Munich*. |

## 3.4. ‘Best combination’ parameter based on crime and housing prices

Since there is no simple minimum for the two-factor problem of lowest total crimes per capita and lowest housing prices, a factor has to be computed. The methodology to get this 'best combination' factor is as follows:

1. Normalize 'Total crimes per capita' by the parameters maximum
2. Normalize 'Price' by the parameters maximum
3. Add both normalized values to an 'Optimum of crimes and prices'
4. Search for the minimum of this newly generated '*Optimum of crimes and prices*'

Following this methodology the ‘best’ borough in Munich with the lowest ‘*Optimum of crimes and prices*’ is ‘Aubing-Lochhausen-Langwied’ (see Figure 19).

|  |
| --- |
|  |
| *Figure 19: ‘Best’ borough of Munich based on lowest ‘Optimum of crimes and prices’*. |

All three new parameters, ‘Total crimes per capita normalized’, ‘Prices normalized’, and ‘Optimum of crimes and prices’, are added to the previously shown and general Pandas data frame (see Figure 19).

## 3.5 Venues

Using the FourSquare API and the previously added geographical data for each borough 100 venues are crawled, based on the boroughs ‘center-position’ and a radius of 500 meters around this position (see Figure 10). The FourSquare API returns a JSON file which is then ported to a Pandas data frame for further data handling and manipulation.

The venues are than grouped by borough, thereafter an investigation regarding the number of unique venues is carried out with result of 176 unique venues in Munich.

Using a ‘one-hot’ encoding the venues are processed further, sorted and normalized by the mean value. Afterwards the 10 most common venues of each borough (see Figure 20) are computed by further sorting and manipulating the underlying Pandas data frame by a user-defined function.

|  |
| --- |
|  |
| *Figure 20: 10 most commen venues of each borough in Munich (displayed only the first four entries – head – of the Pandas data frame).* |

After the above-mentioned process the generated Pandas data frame, of 10 most common venues for each borough, is merged with the previously shown and general Pandas data frame.

## 3.6. Machine learning approach for clustering boroughs of Munich – k-means-algorithm

Using the k-means- algorithm the boroughs of Munich are clustered into five different clusters (see Figure 21) based on their most common venues.

|  |
| --- |
|  |
| *Figure 21: The city of Munich with its boroughs marked (shown is the ‘center-point’ position of each borough) and color-coded as mark of the corresponding cluster. Cluster 1: red, Cluster 2: purple, Cluster 3: dark green, Cluster 4: light green, Cluster 5: orange.* |

The k-means approach was chosen because of ease of use and the categorial nature of this problem to cluster boroughs of a city based on other categorial values.

There are three clusters (cluster 1, 3, and 4; see Figure 22, 24, and 25) with only one borough in it. The other two clusters contain the majority of boroughs of Munich with 5 (cluster 2, see Figure 23) and 17 boroughs (cluster 5, see Figure 26) respectively. Cluster 1 (see Figure 22) has one borough in it. The most common venue are restaurants.

|  |
| --- |
|  |
| *Figure 22: Cluster 1* |

Cluster 2 (see Figure 23) has 5 boroughs in it. The most common venues are bus stops and bakeries.

|  |
| --- |
|  |
| *Figure 23: Cluster 2* |

Cluster 3 (Figure 24) has 1 borough in it. The most common venues are supermarkets and discount stores.

|  |
| --- |
|  |
| *Figure 24: Cluster 3* |

Cluster 4 (see Figure 25) has one borough in it. The most common venues are restaurants.

|  |
| --- |
|  |
| *Figure 25: Cluster 4* |

Cluster 5 (see Figure 26) has 17 boroughs in it. The most common venues are cafes and restaurants.

|  |
| --- |
|  |
| *Figure 26: Cluster 5* |

# 4. Results

Using the cluster just generated and the optimum parameter of crimes and housing prices, the 'best option' in each cluster, regarding the total crimes per capita and borough as well as housing prices in this borough can be concluded.

For the three clusters with each only one borough in it (cluster 1, 3, and 4) nothing changes. They’re the only borough in their cluster and therefore the ‘best-option’ within this cluster.

In cluster 2, which holds 5 boroughs, the ‘best-option’ is the borough of ‘Feldmoching-Hasenbergl’ (see Figure 27), because it has the lowest value for ‘Optimum of crimes and prices’.

|  |
| --- |
|  |
| *Figure 27: ‘Best-option’ in cluster 2.* |

Cluster 5, which holds the majority of boroughs with a number of 17, the same methodology applies. Using it analogous, the borough of ‘Aubing-Lochhausen-Langwied’ proves to be the ‘best-option’ within this cluster.

|  |
| --- |
|  |
| *Figure 28: ‘Best-option’ in cluster 5.* |

# 5. Discussion

As mentioned in the results section, based on key features and certain number of clusters people who want to relocate to Munich can be advised which borough to choose.

Since the used data changes over time, and the links to the data keeps changing as well, one could invest more work into making this notebook more 'general' for future changes in data, links, cities, more factors to be considered, etc.

Especially the generalization to other cities is a major task, since this would grand additional business opportunities, and therefore additional customers.

Furthermore, the integration of more and more detailed key features seems to be a reasonable and valuable task. This would grant the opportunity to support the customers even more and give them a deeper and more customized place to relocate to.

# 6. Conclusion

Based on pre-defined main features and certain number of clusters customers can choose from a borough which fits their needs best regarding crime statistics, housing prices and venues.

# 7. References

[1] Pictures of Munich and general information from Wikipedia.org: https://en.wikipedia.org/wiki/Munich

[2] Most livable cities from DW.com: https://www.dw.com/en/munich-tops-list-of-worlds-most-livable-cities-berlin-and-hamburg-also-in-top-10/a-44478511

[3] Most expensive cities from DW.com: https://www.dw.com/en/stuttgart-unseats-munich-as-germanys-most-expensive-city-for-renters/a-51374468

[4] Crime statistics form Muenchen.de: https://www.muenchen.de/rathaus/dam/jcr:6291ac42-463d-4267-b436-c4b1a3313454/jt190904.pdf

[5] General population data from Wikipedia.org: https://de.wikipedia.org/wiki/Stadtbezirke\_Münchens)

[6] Housing prices and picture with housing prices in boroughs from Immobilienscout24.de: https://www.immobilienscout24.de/neubau/ratgeber/aktuelle-neubau-themen/neubau-kauf-map-wohnungen-muenchen-2019.html

[7] Venues from FourSquare.com: https://de.foursquare.com