Applied Data Science

week 2

Capstone Project — IBM Data Science

THE BATTLE OF NEIGHBORHOODS

An analysis of neighborhoods in Munich using data science methodologies

Abstract

There are different reasons why people move to another place. In this report we identify the most promising borough in Munich for John who needs to move to a new place as his landlord needs the apartment he lives in for himself. John would like to use this as a chance to explore new neighborhoods in Munich. But how to find a suitable neighborhood? Using machine learning algorithms the most promising borough for John is identified, taking into account his specific needs and preferences. This kind of analysis could help landlords to identify the best possible place to live for their customers.

Marie Tritschel

June 04, 2020

List of Figures

List of Tables

Contents

1 Introduction

1.1 Background Information

Almost everybody is moving to another place - at least once in a lifetime. Starting to study in a new city, during a semester abroad, after changing the job, or moving together with a new partner are only some of the reasons why people would find a new place to live. Besides the price of the housing, another important factor is its surrounding. When moving to another place, people most likely would like to find a place that is comparable to the current home to feel comfortable. Even though different people put emphasis on different venues: For example, for some it might be interesting to have a park nearby, while other prefer good restaurants or grocery stores nearby - they all have the same problem in common: How to find a place that meets the personal requirements? In this report we will answer this question using Data Science Methods.

1.2 Problem statement: Where should John move within Munich

In this report, we are trying to find a borough in Munich for John, where he most likely would feel comfortable. John is 27 years old and is currently living in Schwanthalerhöhe, 80335. He really enjoys living there. Unfortunately, his landlord needs the apartment for himself - that is why John needs to move to another place. He would like to use this as a chance to explore new neighborhoods in Munich. Schwanthalerhöhe is directly located at the *Theresienwiese*, a park where John starts his day with running and grabbing a coffee afterwards. That is why it would be nice to have a park and coffee store close to the new home. John works for a big consulting company, that is why he comes home pretty late. It would therefore be good to have a grocery store nearby. When it is getting too late at work, he likes to grab dinner

in a restaurant. That is why it would be a plus to have bars or restaurants close to his home. The price per m^2 from his current apartment is: $34.29e/m^2$. Finding one with lower price per m^2 would be perfect. Summarized, John would most likely enjoy to have the following venues nearby:

- park nearby
- coffee store
- restaurants bars
- grocery store

If possible, cheaper price per m^2 than the current apartment ($34.29e/m^2$) All in all, John really loves his current borough, that is why we need to find out in addition, which of the other boroughs is most similar to the one he is currently living in. We will use data science methods to identify the most promising neighborhoods based on these criteria.

1.3 Target audience

This report is an analysis of the boroughs in Munich customized to the needs of John (park, coffee store... nearby). However, the approach is applicable to all people with any specific needs. The information gathered from Foursquare [1] in combination with data science methods are a good basis to derive data driven decisions regarding boroughs that best fit the specific needs at hand. It would even be possible for real estate agents to use some similar approaches to find the perfect home for their customers.

2 Data

2.1 Description of the Data

In order to find the most promising borough for John the following data is needed:

1. **Average price per m^2 of the apartments in Munich**

   This information is gathered through web scraping the internet [2]. Some modifications have been performed to transform the data frame in a format that can be used for further analysis. Detailed steps can be found in the Data preparation section 2.3.

2. **Information about the venues in all boroughs of Munich (including those around Johns home)**

   This information is gathered through web scraping from the internet [3]. As the format did not fit our further data processing requirements, it has been transformed. The Geocoder Python package [4] was used to receive the latitude

and longitude coordinate for all of the boroughs. The boroughs and their corresponding latitude and longitude are used as input for Foursquare [1] to source information about the boroughs.

2.2 How the data will be used to solve the problem

We will start with an exploratory data analysis, where we intend to understand the underlying data.

To get a first impression about the distribution of venues in Munich, they are visualized using Folium map [5]. The chosen color code will give immediate yet superficial insight, how John's favourite venues are distributed across Munich and how the surrounding of his actual location looks like.

For the further analysis, the venues will be divided into two types: Firstly, we have a data frame containing all venues of John's personal interest (i.e. parks, coffee stores, restaurants, bars and grocery stores) and one containing all other venues. This subdivision will let us analyse which boroughs are most similar to John's current neighborhood in terms of his personal preferences but also tells us the most common venues within the boroughs. This ensures to find the top borough in terms of John's interests - which is of course the most important criterion - but also provides a list for John with the most common venues in the neighborhoods that he could use for his final decision.

One hot encoding and k-means will narrow the list of the most promising boroughs to three. Combining these three with the pricing analysis lets us recommend the best borough match for John.

2.3 Data preparation

Several steps needed to be performed to use the data and derive meaningful recommendations from it. First of all we gathered the boroughs of Munich and their corresponding postal code through web scraping of the web page [3] in a data frame. The result is shown in figure 2.1(a). If you look at the column *Postleitzahl*, you'll notice, that there are many postal codes in the PostalCode column, all separated by a comma and each referring to a different area of the borough. As we would like to compare all of them, we need to get all of these postal codes in a separate row. The result can be seen in figure 2.1(b) which also already contains the renamed column names. After that, information about the average price per m^2 of the boroughs of Munich is needed. This is why we used web scraping [2] to gather the data in a data frame. Here, only smaller modifications were necessary. First of all the columns that were not needed for the investigation were dropped. After that, the columns were renamed and the price was displayed in €/m^2 instead of Cents/m^2. The result is shown in 2.1(c).

The next step was to merge the data frames what turned out to be challenging, as we did not have a price per m^2 for every borough. That is why we used a common

	Stadtteil	Postleitzahl
0	Allach-Untermenzing	80995, 80997, 80999, 81247, 81249
1	Altstadt-Lehel	80331, 80333, 80335, 80336, 80469, 80538, 80539
2	Au-Haidhausen	81541, 81543, 81667, 81669, 81671, 81675, 81677
3	Aubing-Lochhausen-Langwied	81243, 81245, 81249
4	Berg am Laim	81671, 81673, 81735, 81825

(a) Boroughs in Munich. Data frame before data cleansing

	Borough	PostalCode
0	Allach-Untermenzing	80995
1	Allach-Untermenzing	80997
2	Allach-Untermenzing	80999
3	Allach-Untermenzing	81247
4	Allach-Untermenzing	81249

(b) Boroughs in Munich. Data frame after data cleansing

	PostalCode	PricePerm2
1	80995	14.10
2	80997	13.25
3	80999	13.05
4	81247	14.55
5	81249	13.25

(c) Data frame after data cleansing containing information about the price.

Figure 2.1: Data cleansing of data frame containing boroughs in Munich.

data science method: The NANs were replaced by the means of the corresponding borough. In one borough, Thalkirchen-Obersendling-Fürstenried-Forstenried-Solln, was not a single price. That is why we were not able to apply the "mean-method" to this borough. Instead, we excluded these boroughs from the further investigation.

The Geocoder Python package [4] was used to get the coordinates (latitude and longitude) for all of the neighborhoods of Munich. They were added as extra column in the merged data frame. Therefore a function was defined, that takes as input parameters the postal code and borough and gives back the latitude and longitude of the specific postal code and borough. The latitude and longitude of Johns current home were received the same way. Details about the Python code can be found in the Jupyter notebook of this report [6].

The next step was to gather the venues within the boroughs in a separate data frame using the Foursquare API [1]. We chose a radius of 750m and 50 as limit of number of venues returned by Foursquare API[1]. The data frame received from this approach is shown in figure 2.2.

This data frame was used as input to gather the venues in the boroughs in a separate data frame. Summarized, the following data(frames) were created which were used for further analysis:

[1]Several radius were tried. It turned out that 750 is the best choice, as accessing the API for lower radius was kind of unstable and returned back error messages.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Altstadt-Lehel	48.13487	11.581988	Globetrotter	48.134611	11.581879	Sporting Goods Shop
1	Altstadt-Lehel	48.13487	11.581988	Little London	48.135562	11.580961	Steakhouse
2	Altstadt-Lehel	48.13487	11.581988	OOH BABY I LIKE IT RAW	48.134023	11.580400	Café
3	Altstadt-Lehel	48.13487	11.581988	Literatur Moths	48.133762	11.582408	Bookstore
4	Altstadt-Lehel	48.13487	11.581988	Item Shop	48.133460	11.581509	Hobby Shop

Figure 2.2: Venues in the borough of Munich. This data was received using the Foursquare API [1].

- `df_muc`

  Contains the average price per m of the apartments in Munich. This data frame is used to compare the average price per square meter of the boroughs in Munich with the actual price per square meter of Johns apartment.

- `muc_venues`

  Contains all the venues in the boroughs of Munich. This data frame is used to find boroughs that fits most Johns requirements.

In addition to that, we know that Johns home is in Gollierstr 6, 80339 with latitude and longitude of 48.1358166 and 11.5466346.

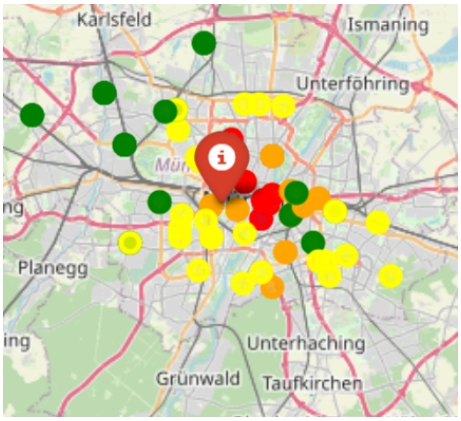
3 Methodology

This section represents the main component of the report. It starts with an exploratory data analysis before we dig deeper into solving the problem and applying machine learning algorithms. For the analysis, the venues are divided into two different data frames: Those who are Johns interests (i.e. parks, coffee stores, restaurants, bars and grocery stores) and all the other ones. We will perform one hot encoding to narrow the list of the most promising boroughs for both of the venue data frames. Combining the results with a k-mean cluster analysis of all venues in Munich will provide us the most promising neighborhoods for John.

3.1 Exploratory Data Analysis

We started with an exploratory data analysis of the average price per square meter of apartments in all boroughs of Munich which is shown in figure **??**. Figure **??** shows the frequency of prices in Munich. It is obvious that Johns home has by far the highest price per m^2 compared to all other boroughs. Figure 3.1(b) shows the distribution of the prices across Munich and the corresponding color codes are specified in table 3.1.

(a) Frequency of price per m^2 in Munich. The red line at $34.29e/m^2$ is the price of Johns current apartment.



(b) Price distribution across Munich. The color codes are shown in table 3.1. Johns current home is indicated by the red icon.

Figure 3.1: Distribution of price per m^2 across Munich.

Price range	color	Nr of boroughs
(2.04, 4.59]	purple	3
(4.59, 7.12]	blue	0
(7.12, 9.65]	black	0
(9.65, 12.18]	brown	0
(12.18, 14.71]	green	40
(14.71, 17.24]	yellow	49
(17.24, 19.77]	orange	54
(19.77, 22.3]	red	19

Table 3.1: Price range and color code corresponding to figure 3.1 and number of boroughs within the specific price range.

Apartments in the center of Munich tend to be more expensive compared to those with higher distance to the center. Johns current home is within the center of Munich as indicated by the red icon in figure 3.1(b). If John would like to have a cheaper apartment he needs to move away from the center of Munich.

In a further exploitative analysis the distribution of venues in Munich was investigated. The result is shown in figure 3.1 with the corresponding color code explained in table 3.1. It can be seen that all venues of Johns interest are located around Johns home, which is indicated with the red icon as before. However, even though John lives in the perfect area to meet all of his requirements, there a a lot of other promising neighborhoods that do so as well[1].

[1]Please take a look at the Folium Map and its zoom function in the Jupyter notebook [6]

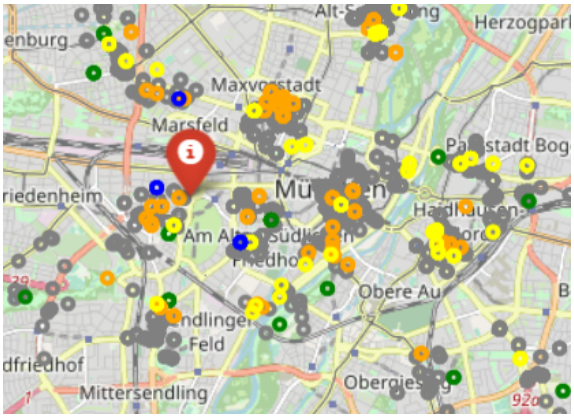


Figure 3.2: Venues across boroughs in Munich. The color code can seen in table 3.1.

Venue	color
parks	green
coffee stores	orange
bars and restaurants	yellow
grocery stores	blue
other	grey

Table 3.2: Color code of venues in Munich corresponding to figure 3.1.

3.2 One hot encoding

To analyse each neighborhood, one hot encoding is used for both data frames:

- Johns favourite venues: Parks, coffee, bars, restaurants, grocery stores
- All venues in Munich except of Johns favourite venues

3.2.1 Johns favourite venues

As we already know, that John prefers to have parks, coffee stores, bars, restaurants and grocery stores nearby. In this analysis, we used one hot encoding to identify how many of the favourite venues are located in the different neighborhoods. Looking at our data frame, that contains only Johns favourite venues (`muc_venueFAV`) we had several columns with sub-categories, e.g. Italien Restaurant, African Restaurant etc. As we were just interested in the category but not the specific subcategory, we combined all restaurants in one column. The same was done for all of the other venues, as we would like to know the number of bars, irrespective if it is a wine bar or a beer bar. A "Favorite score" was defined for the corresponding data frame building the normalized sum of favourite venues in the neighborhood:

$$\text{Favourite Score} = \frac{\sum_i \text{Venues in this borough}}{\sum_i \text{All venues}} \tag{3.1}$$

This way we receive the result shown in figure 3.3(a). Based on that list, we tend to say that all neighborhoods above Schwanthalerhöhe fit Johns needs the best. Looking at "Maxvorstadt", for example which has the highest "Favourite Score" it is obvious, that there are a ton of cafes and restaurants, however, no bar, grocery store or park. Therefore lines that have a zero were drooped, to ensure that at least one of all of Johns favourite venues is within this neighborhood. The result is shown in 3.3(b).

	Neighborhood	Bars	Parks	Cafés	Grocery Stores	Restaurants	FavoriteScore
0	Maxvorstadt	0	0	112	0	210	0.155932
1	Au-Haidhausen	44	2	20	6	164	0.114286
2	Altstadt-Lehel	42	0	31	9	139	0.107022
3	Ludwigsvorstadt-Isarvorstadt	25	9	26	7	152	0.106053
4	Neuhausen-Nymphenburg	5	11	12	4	148	0.087167
5	Sendling	0	8	16	8	120	0.073608
6	Schwabing-Freimann	24	1	20	0	105	0.072639
7	Schwabing-West	25	2	18	0	103	0.071671
8	Bogenhausen	2	2	0	0	102	0.051332
9	Schwanthalerhöhe	16	4	24	4	52	0.048426

(a) One hot encoding and favourite score of all boroughs in Munich.

	Neighborhood	Bars	Parks	Cafés	Grocery Stores	Restaurants	FavoriteScore
2	Au-Haidhausen	44	2	20	6	164	0.114286
9	Ludwigsvorstadt-Isarvorstadt	25	9	26	7	152	0.106053
13	Neuhausen-Nymphenburg	5	11	12	4	148	0.087167
19	Schwanthalerhöhe	16	4	24	4	52	0.048426
23	Untergiesing-Harlaching	4	1	4	1	18	0.013559

(b) One hot encoding and favourite score of all boroughs in Munich, that contain at least one of the venues of Johns interest. Compared to 3.3(a), lines with zeros were dropped.

Figure 3.3: One hot encoding and favourite score of all boroughs in Munich.

3.2.2 All venues in Munich except Johns favourite venues

In this analysis we identified the ten most common venues in all boroughs of Munich without taking into account the venues of Johns interest. John could use this list to explore the boroughs further before he makes a final decision.

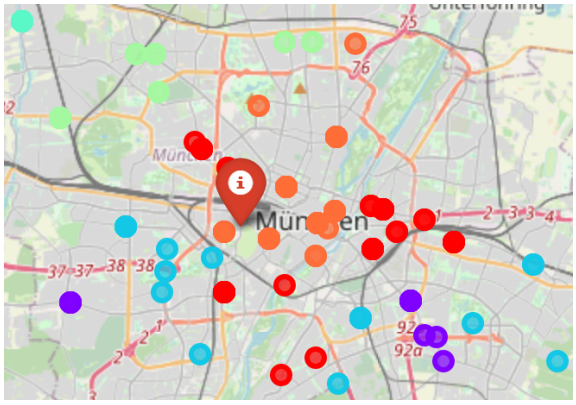
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allach-Untermenzing	Hotel	Sporting Goods Shop	Bakery	Trattoria/Osteria	Bus Stop
1	Altstadt-Lehel	Hotel	Coffee Shop	Plaza	Opera House	Department Store
2	Au-Haidhausen	Plaza	Bakery	Gourmet Shop	Ice Cream Shop	Pub
3	Aubing-Lochhausen-Langwied	Pharmacy	Bus Stop	Light Rail Station	Zoo Exhibit	Farm
4	Berg am Laim	Tram Station	Smoke Shop	Hotel	Gym / Fitness Center	Big Box Store

Figure 3.4: Except of most common venues in the boroughs of Munich.

To generate this information we used one hot encoding and grouped the data frame based on the neighborhood. A mean value of the specific neighborhood is calculated. An except of the resulting data frame is shown in figure 3.4. Further information can be found in the Jupyter Notebook [6].

3.2.3 Clustering of the neighborhoods

The k-means clustering was used to cluster the neighborhood into 8 clusters. In this analysis all venues in the boroughs of Munich were taken into account. To do that, we first used one hot encoding, then calculated the mean of each venue in each neighborhood and finally grouped the data frame based on the neighborhood. The distribution of the clusters is shown in figure 3.5(a), where the orange one, belonging to Cluster 7, are most similar to Schwanthalerhöhe and thus Johns current home.



(a) Distribution of Cluster identified through k-means clustering method.

	Neighborhood
0	Altstadt-Lehel
1	Maxvorstadt
2	Ludwigsvorstadt-Isarvorstadt
3	Schwabing-Freimann
4	Schwabing-West
5	Schwanthalerhöhe

(b) Boroughs that belong to cluster 7 (orange in3.5(a). These boroughs are most similar to Johns current home Schwanthalerhöhe, as they are in the same cluster.

Figure 3.5: Results from k-means clustering.

The corresponding boroughs of cluster 7 are shown in figure 3.5(b). Based on the clustering algorithm, we would recommend John to move in one of the boroughs shown in figure 3.5(b) as tehse are most similar to Johns actual neighborhood.

4 Results and Discussion

Different analysis have been performed to find the most suitable new neighborhood for John. Considering the price per m John can move anywhere, as all of the boroughs have a lower price per m.

The venues within the different neighborhoods have been divided in two different data frames: a) Johns favourite venues and b) All venues in Munich except of Johns favourite venues.

For the analysis where just Johns favourite venues have been considered (a), we defined a favourite score, taking into account, how often Johns favourite venues occurred in the specific neighborhoods. This analysis narrowed down the results to the following three neighborhoods:

- Au-Haidhausen
- Ludwigsvorstadt-Isarvorstadt
- Neuhausen-Nymphenburg

For the analysis, where all venues in Munich except of Johns favourite venues have been considered (b), we identified the 10 most common venues within all of the boroughs which could be helpful for John, to further explore other neighborhoods.

Based on the k-means cluster analysis, the following neighborhoods are most similar to the one John currently lives in:

- Altstadt-Lehel
- Maxvorstadt
- Ludwigsvorstadt-Isarvorstadt
- Schwabing-Freimann
- Schwabing-West

If you combine these findings, the most promising borough for John is identified: **Ludwigsvorstadt-Isarvorstadt** as this is the one with the highest favourite score, but also the one which is most similar to his current place according to the k-means algorithm. This recommendation can even be further specified, if you take into account the different pricing across Ludwigsvorstadt-Isarvorstadt visualized in figure 4.1.

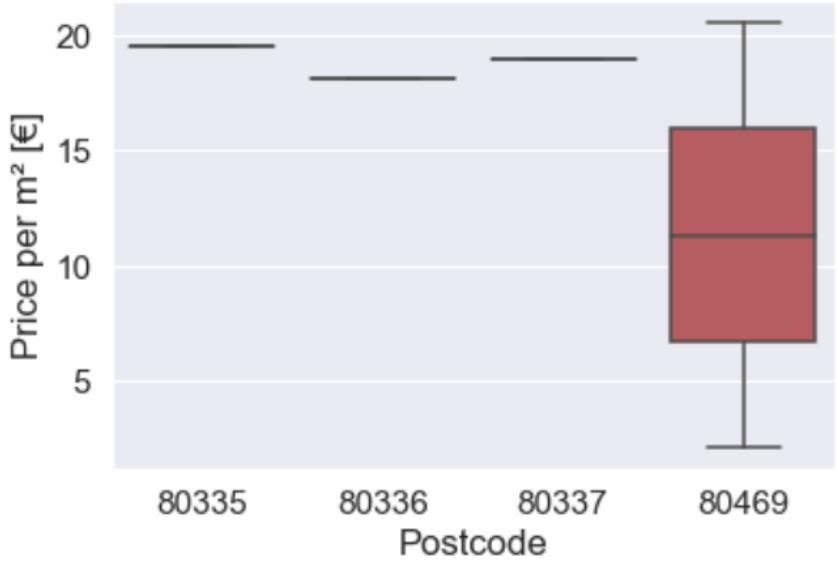


Figure 4.1: Pricing in the different areas of Ludwigsvorstadt-Isarvorstadt.

As can be seen, the price across the boroughs is pretty constant around 19 €per m^2. The only exception is the borough with postal code 80469, where the mean average price per m^2 is with around 11€per m^2 clearly lower. Based on this analysis Johns perfect match would be Ludwigsvorstadt-Isarvorstadt in 80469 Munich.

5 Conclusion

The purpose of this project was to identify a borough that is similar to Johns current one and has venues, that are important for John (i.e. arks, coffee, bars, restaurants, grocery stores) in order to narrowing down the search for optimal new borough as a place to live. For this report different analysis have been performed. Considering the average prices per m for apartments in munich showed, that no matter were John is moving he will most likely find a cheaper apartment.

By just taking into account the venues that are important for John, we could identify three boroughs, which seemed to be most promising as new neighborhood for Johns home. The k-means provided an insight into similar neighborhoods, compared to Johns actual one. After combining these results, we identified one single borough, that is most likely the best choice for Johns new area to live: Ludwigsvorstadt-Isarvorstadt. Further dividing this neighborhood into different postal codes, showed, that the one Ludwigsvorstadt-Isarvorstadt in 80469 would be the best choice, as it is the cheapest area within this neighborhood.

Of course, the price and points of his personal interest are not the only criteria how John should make his final decision, as additional factors like availability of apartments, noise, proximity to friends also matter. However, it serves as an orientation and good neighborhood to start searching.

Bibliography

[1] *Foursquare*. URL: `https://foursquare.com/` (visited on 06/03/2020).

[2] *Price of Boroughs in Munich*. URL: `https://de.statista.com/statistik/daten/studie/260438/umfrage/mietpreise-in-muenchen-nach-bezirken/` (visited on 06/03/2020).

[3] *Boroughs in Munich*. URL: `https://www.muenchen.de/leben/service/postleitzahlen.html` (visited on 06/03/2020).

[4] *Geocoder*. URL: `https://geocoder.readthedocs.io/index.html` (visited on 06/03/2020).

[5] *Folium Map*. URL: `https://pypi.org/project/folium/` (visited on 06/03/2020).