Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 1)¶

NEW HOME

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 Gym nearby, while other prefer good restaurants or coffee 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 miami

In this report, we are trying to find a borough in miami for John, where he most likely would feel comfortable. John is 27 years old and is currently living in The Asia Brickell Key. He really enjoys living there. Unfortunately, he lost his wealth - that is why John needs to move to another place. He would like to use this as a chance to explore new neighborhoods in miami.Brickell Key is is an artificial triangular island at the mouth of the Miami River in Miami, Florida, United States. John like to see modern Art and grab coffee afterwards.That is why it would be nice to have a Art Gallery and coffee store close to the new home ,john likes to keep his body fit, 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.Summarized, John would most likely enjoy to have the following venues nearby:

- Art Gallery
- restaurants & bars
- Gym
- coffee store

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 for this report

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 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 m² of the apartments in Miami:This information is gathered through web scraping from this webpage Miami Average price.
2. Information about the venues in all boroughs of Miami (including those around Johns home): This information is gathered through web scraping from wikipedia List of neighborhoods in Miami.the Geocoder Python package (https://geocoder.readthedocs.io/index.html) is used to receive the latitude and logitude coordinate for all of the boroughs. The boroughs and their corresponding latitude and longitude are used as input for FourSquare to source information about the boroughs.

2.2 How will the data be used to solve the problem

We will start with an exploratory data analysis, where we intend to understand the underlying data. The describe method provides valuable insights for the "average price per m² of the apartments in miami" investigations.

To get a first impression about the distribution of venues in miami, they are visualized using Folium map. The chosen color code will give immediate yet superficial insight, how John's favourite venues are distributed across miami 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 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.

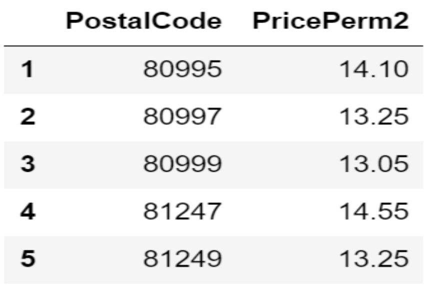
One hot encoding and k-means will narrow the list of the most promising boroughs to three. Combining these three with the pricing analysis lets us recommend the best borough match for John.

2 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 Miami, 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 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

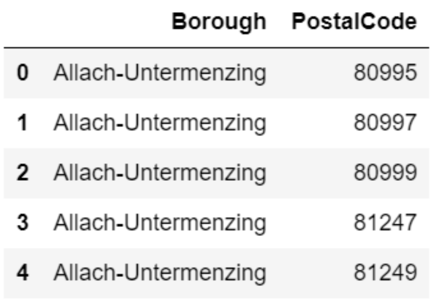
list of the most promising boroughs to three. Combining these three with the pricing analysis lets us recommend the best borough match for John.

2.3 Data preparation

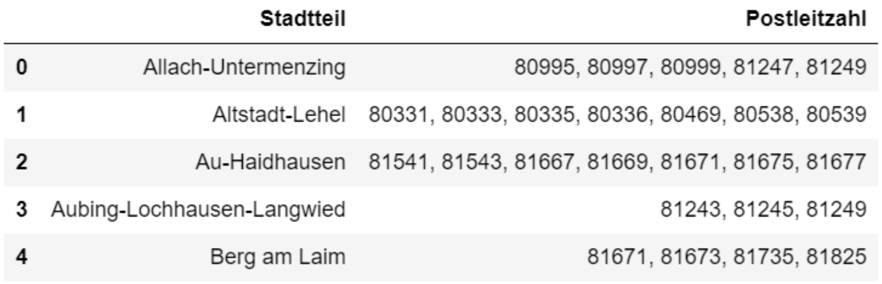
Several steps needed to be performed to use the data and derive meaningful recommendations from it. First of all we gathered the boroughs of Miami and their corresponding postal code through web scraping of the web page [3] in a data frame. The result is shown in figer2.1(a). If you look at the column 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 m2 of the boroughs of Miami 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 e/m2 instead of Cents/m2. 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 m2 for every borough. That is why we used a common data science method: The NANs were replaced by the means of the corresponding borough. In one borough, Thalkirchen-Obersendling-Furstenried-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 Miami. 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



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



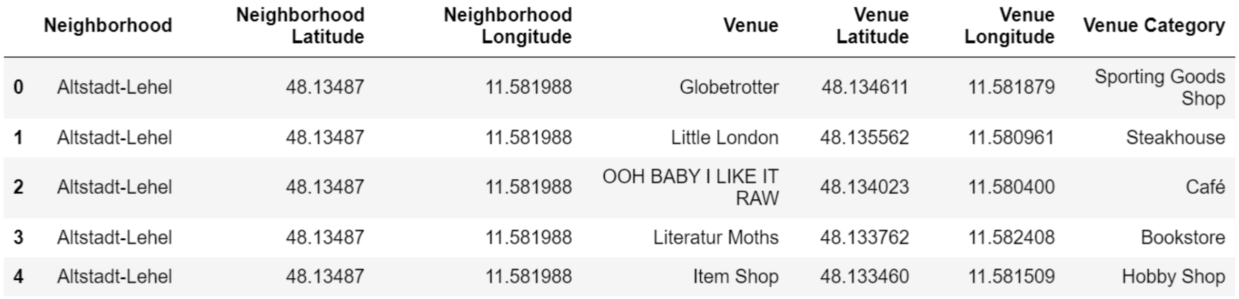
	Borough	PostalCode
0	Allach-Untermenzing	80995
1	Allach-Untermenzing	80997
2	Allach-Untermenzing	80999
3	Allach-Untermenzing	81247
4	Allach-Untermenzing	81249



	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

limit of number of venues returned by Foursquare API1. 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:



	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

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

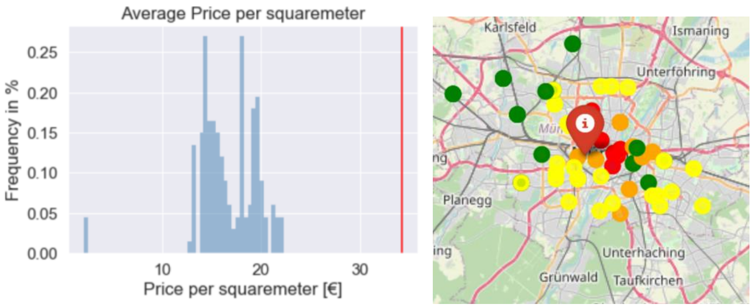
2- mis_venues: Contains all the venues in the boroughs of Munich. This data frame is used to find boroughs that fits most Johns requirements.

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 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 miami which is shown in Figure ?? shows the frequency of prices in miami. It is obvious that Johns home has by far the highest price per m2 compared to all other boroughs. Figure 3.1(b) shows the distribution of the prices across miami and the corresponding color codes are specified in table 3.1.



3.2 One hot encoding

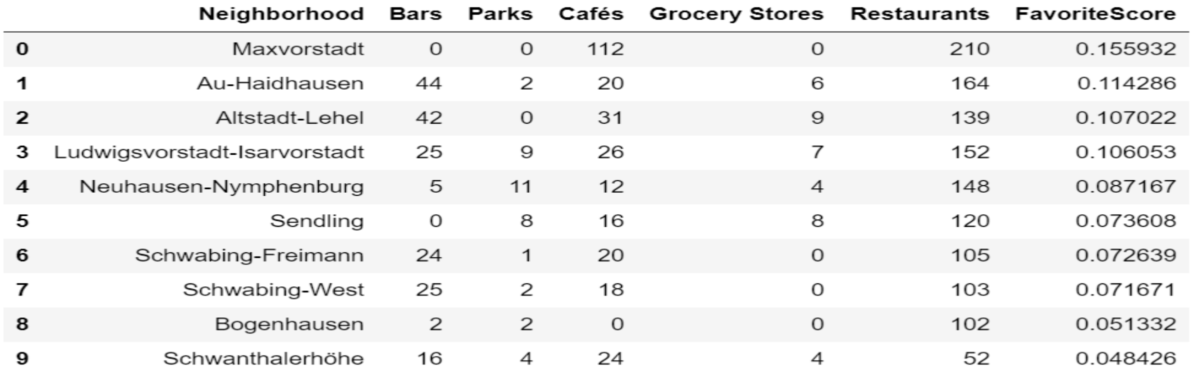
To analyse each neighborhood, one hot encoding is used for both data frames:

1- Johns favourite venues:

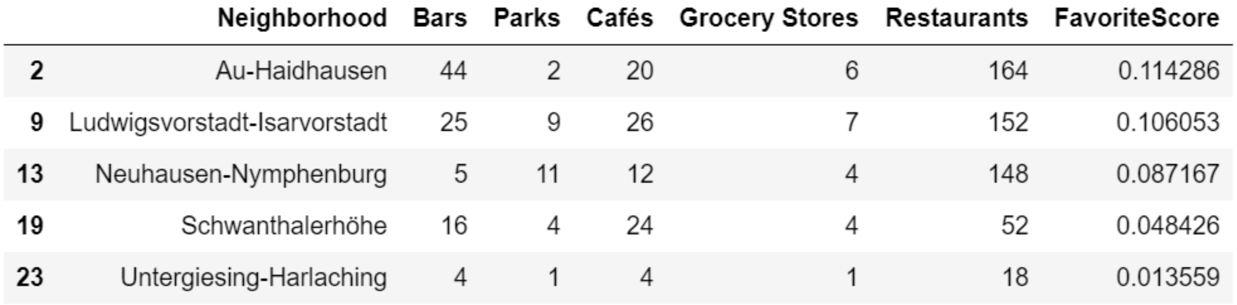
2- All venues in miami except of Johns favourite venues

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.



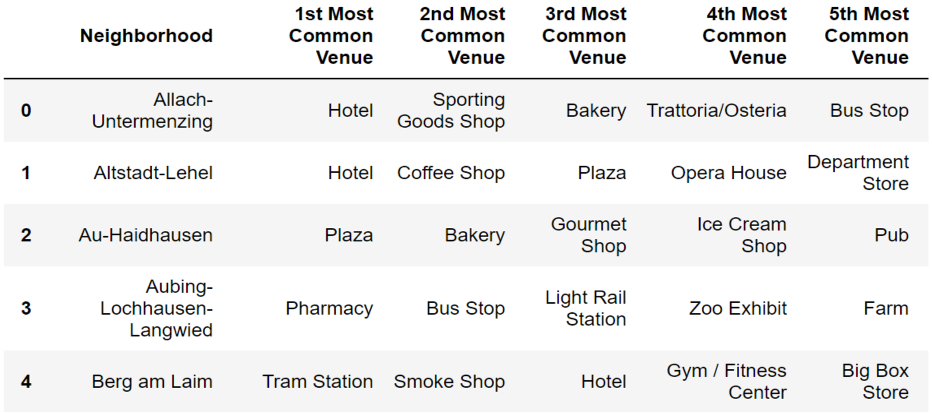
	Neighborhood	Bars	Parks	Cafés	Grocery Stores	Restaurants	FavoriteScore
0	Maxvorstadt	0	0	112	0	210	0.155932
1	Au-Haidhausen	44	2	20	6	164	0.114286
2	Altstadt-Lehel	42	0	31	9	139	0.107022
3	Ludwigsvorstadt-Isarvorstadt	25	9	26	7	152	0.106053
4	Neuhausen-Nymphenburg	5	11	12	4	148	0.087167
5	Sendling	0	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



	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

All venues in miami except Johns favourite venues

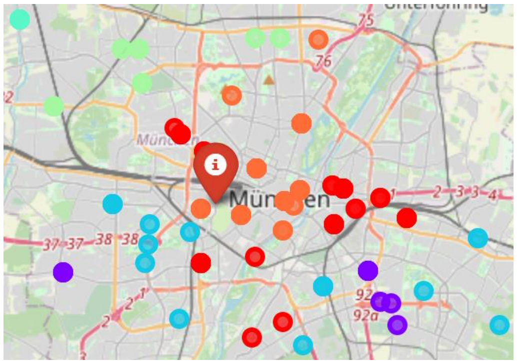
In this analysis we identified the ten most common venues in all boroughs of miami without taking into account the venues of Johns interest. John could use this list to explore the boroughs further before he makes a final decision.

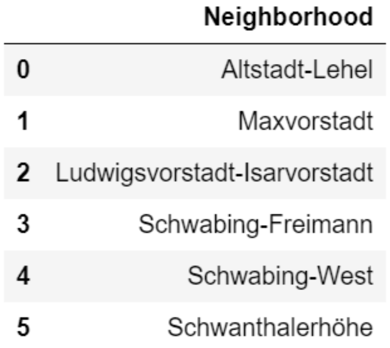


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allach-Untermenzing	Hotel	Sporting Goods Shop	Bakery	Trattoria/Osteria	Bus Stop
1	Altstadt-Lehel	Hotel	Coffee Shop	Plaza	Opera House	Department Store
2	Au-Haidhausen	Plaza	Bakery	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

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 Schwanthalerhohe and thus Johns current home.





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

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.9 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 tothe following three neighborhoods:

1- Au-Haidhausen

2- Ludwigsvorstadt-Isarvorstadt

3- 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:

1-Altstadt-Lehel

2- Maxvorstadt

3-Ludwigsvorstadt-Isarvorstadt

4-Schwabing-Freimann

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 kmeans algorithm. This recommendation can even be further specified, if you take into account the different pricing across Ludwigsvorstadt-Isarvorstadt visualized in figure 4.1.

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 in order to narrowing down the search for optimal new borough as a place to live. For this report diffanalysis have been performed. Considering the average prices per m for apartments in miami 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.