The Battle of Students

Bernhard Neumayer

October 11, 2019

1 Introduction

1.1 Background

The city of Graz is the second biggest city in Austria with almost 330,000 residents [1]. It is a typical student city with roughly 60,000 students who mainly live in the city. The housing of students therefore is a significant economic factor in Graz. Accordingly, apartment-hunting is omnipresent and a time-consuming issue.
When looking for an apartment preferences may vary from person to person but when people start to study, close proximity to the university and affordable rent tend to be important factors. Furthermore, many prefer lively neighborhoods to get in contact with other students and to socialise. The social environment, however, cannot be put into numbers that easily and when moving to a new city one has to rely on assumptions and recommendations.
Additionally, many companies in and around Graz are offering employment to graduates of the different universities and therefore, many move to Graz permanently after having finished their studies. Since interest rates are currently at a low, people planning to study and having an intention to stay afterwards may consider to invest into the future, take out a loan and buy property.

1.2 Problem

When considering to rent an apartment or to buy property, preferences will differ significantly. An apartment can be chosen considering the current situation in a certain borough or neighbourhood but for the acquisition of property the outlook is more important. Questions to be answered are:

- Where can I find affordable property?
- What is the current state of the borough?
- Will the value increase, stay or decrease over the years?

This requires a comprehensive analysis of rents, prices, the structure of the different boroughs and the envrionmental circumstances.

1.3 Target Audience

This data analysis project aims at upcoming students planning to move to Graz with little or no knowledge on the structural and environmental circumstances of the different boroughs. The outcome is a classification of all available boroughs into groups with similar characteristics regarding rents, property prices and demographics. Ultimately, the analysis provides conclusions on recommended boroughs for renting an apartment or buying property.

2 Data

2.1 Data Sources

This project will use four data sources:

1. Information on the different boroughs of Graz including environmental and demographic characteristics can be obtaind from Wikipedia [2].
2. The website *www.immobilienscout24.at* provides rent and property price information for Graz down to borough-level [3]. This information is based on advertisements posted on and purchases made through this site.
3. The website *www.immowert123.at* is owned by a real estate agency evaluating and selling realty. The agency provides estimates for selected objects in Graz, including object type, size and recommended price [4].
4. The Foursquare database [5] is used to evaluate the vibrancy of a borough, which is based on the availability of venues in a certain area.

The location data, i.e. latitude and longitude, for Graz and its boroughs is acquired using *geopy* with *Nominatim* and*ArcGIS*.

2.2 Data Acquisition

The Wikipedia site for Graz provides information in tabular form, the same applies for the second data source, *www.immobilienscout24.at*. This means the data can easily be obtained using *pandas* and *read_html()*.

The acquisition of the data provided by *www.immowert123.at* requires a certain amount of creativity. The URL in [4] serves as the base URL for all data acquisitions, the data can then be obtained by adding a district to that URL. Scraping the data on the respective website requires *bs4* and *BeautifulSoup*, since it is not directly packed in a table. Furthermore, the districts do not necessarily coincide with the boroughs of Graz; therefore, the locations of the districts will be compared to the borough locations to merge the information accordingly.

Combining the data retrieved from [3] and [4] provides a data set with rent and realty prices for all boroughs of Graz. This data set is then combined with the demographic data from [2] and venue information from Foursquare. Since Graz is still a rather small city, the absolute amount of available venues will be sufficient for the time being.

2.3 Data Cleaning, Data Preparation and Feature Selection

The data in [2] provides names of the boroughs, number of residents, area, population density, date of foundation and postal codes. The number of residents can be dropped since its information is inherently included in area and population density, also the date of foundation is not of interest in this context. Postal codes turned out to show too much overlap for the boroughs and therefore do not bear useful information.

The tables from [3] can be imported directly. Care has to be taken of the different encoding of numbers in the German language; missing values are set to *NaN*.

The data published in [4] comprises all sorts of realty. This project focuses on apartments; therefore, the data will be filtered for entries with the German word for apartment: “wohnung”. The price

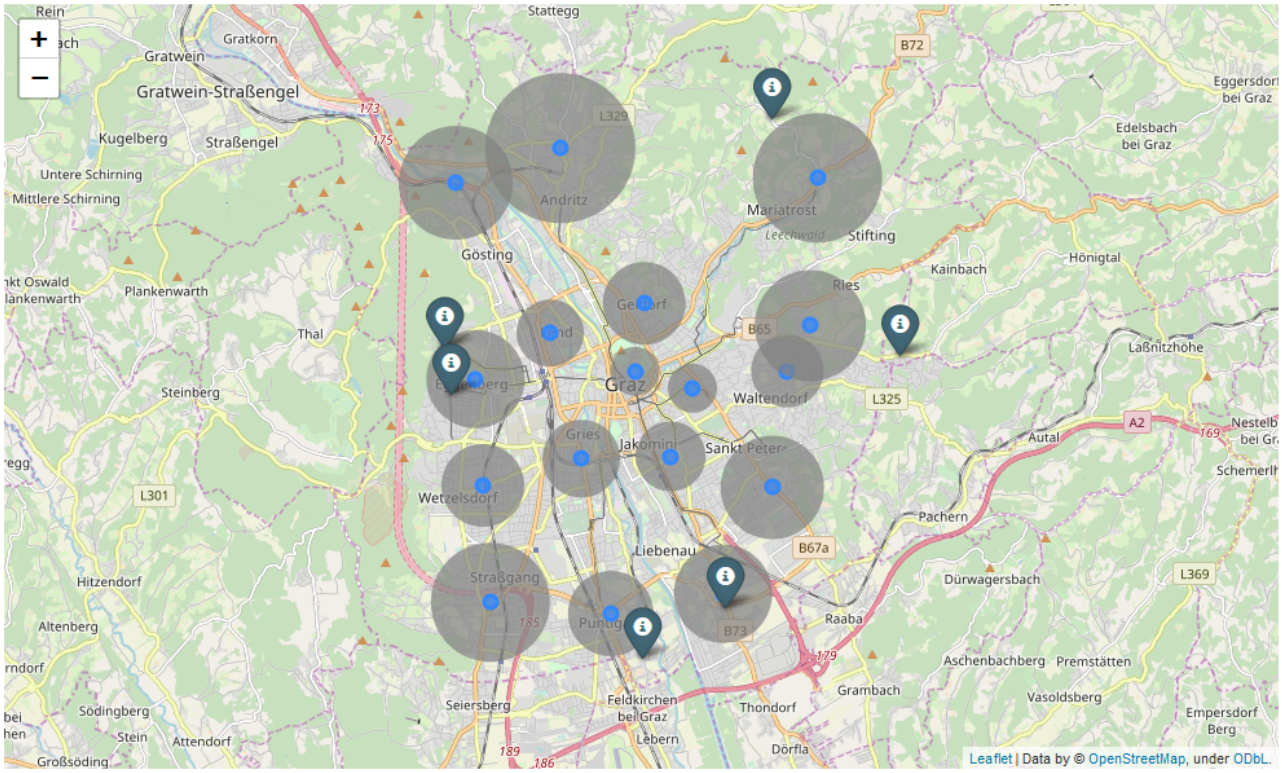

Figure 1: Starting point of the analysis. Blue circles mark the 17 boroughs of Graz, grey circles show the area used for venue search and info markers show the positions of additional data available from [4].

per square metre is calculated from the recommended price and the apartment size. The results are averaged per district or borough.

We know from [2] that the areas of the boroughs differ significantly. Therefore, the radius for venue search using Foursquare is weighted by the square root of the area to account for this occurence. The minimum radius applied for the search is set to 500.

The final data set used for clustering comprises rent per square metre, price per square metre, borough area, population density and venue count. Prior to clustering the data is normalized using *sklearn*'s *StandardScaler()*.

3 Methodology

3.1 Starting Point

As mentioned in chapter 2, the data from [4] is partially provided for boroughs and partially for certain disctricts that are parts of a borough. Therefore, the boroughs and districts are displayed on a folium map to assign the district data to the correct boroughs. This could also be done automatically by calculating minimum difference between boroughs and districts but doing it manually allows to integrate knowledge on the locations, which is more reliable in this case.
Furthermore, it was also mentioned in chapter 2 that the boroughs differed in size and that this should be considered for the analysis of the environment. This is done by weighting the radius for the Foursquare request by the square root of each borough's area. A map showing the locations of all boroughs of Graz, the location of the additional district data and the weighted areas for venue

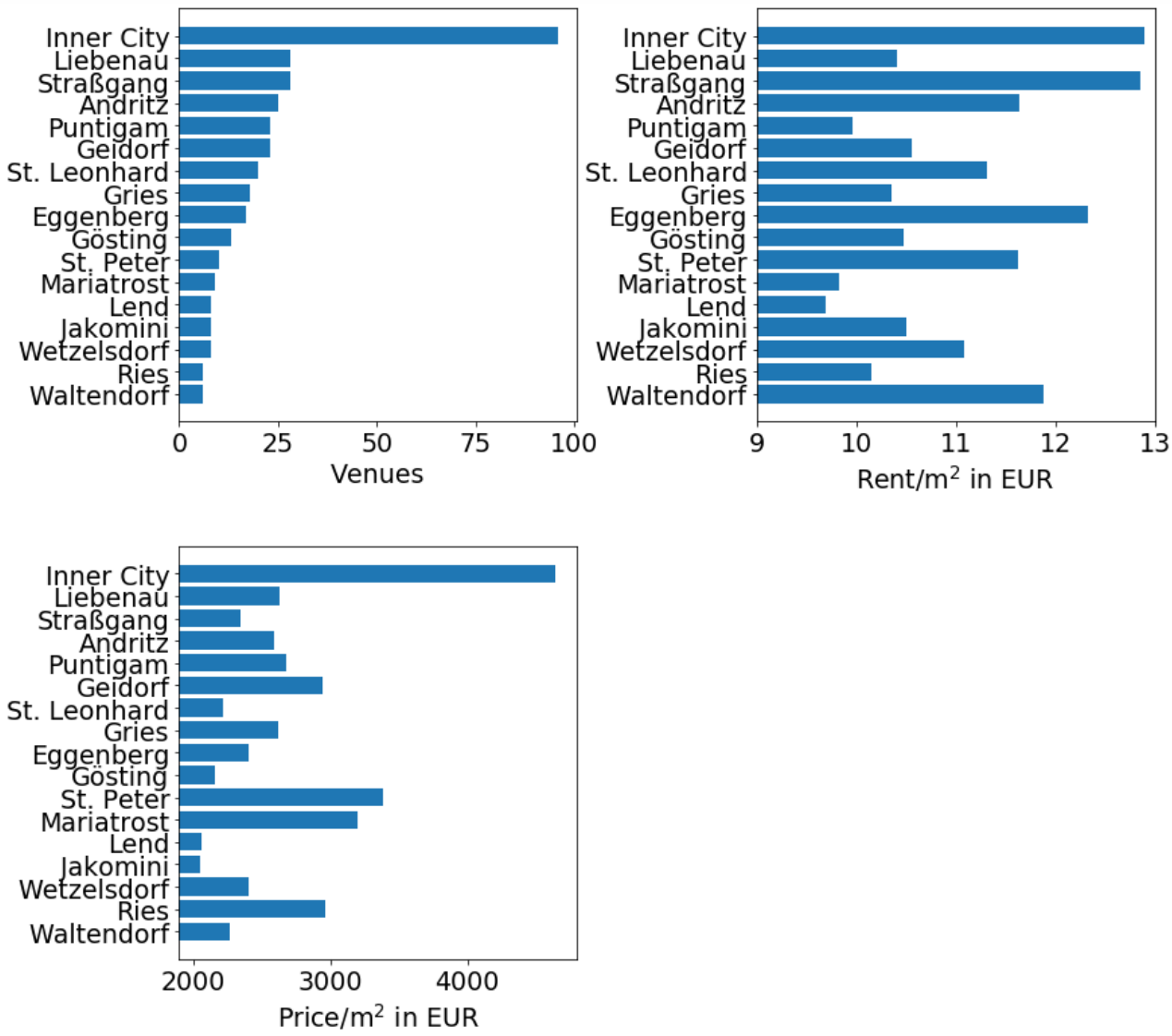

Figure 2: Comparison of vibrancy (upper left), rent (upper right) and realty prices (lower left).

search are shown in figure 1. The data from the additional districts were added to the closest borough (further details which district was added to which borough can be found in the jupyter notebook[1]).

3.2 Exploratory Data Analysis

3.2.1 Vibrancy, Rent and Realty Price

A very interesting aspect is the relationship between rent, price and the vibrancy of a borough, since this is a main pillar of the entire analysis. As mentioned before, the vibrancy is derived directly from the absolute number of venues in a region of interest. This approach was chosen since Graz is still a rather small city and the number of venues is relatively reduced compared to bigger cities.
Figure 2 shows the number of venues sorted by frequency and the rent and realty prices in the corresponding boroughs. These bar plots are very promising, since there is no obvious correlation between vibrancy and either rent or realty prices. This means that our data can potentially be

[1] https://github.com/bneumayer/Coursera_Capstone/blob/master/Assignment_The_Battle_of_Students.ipynb

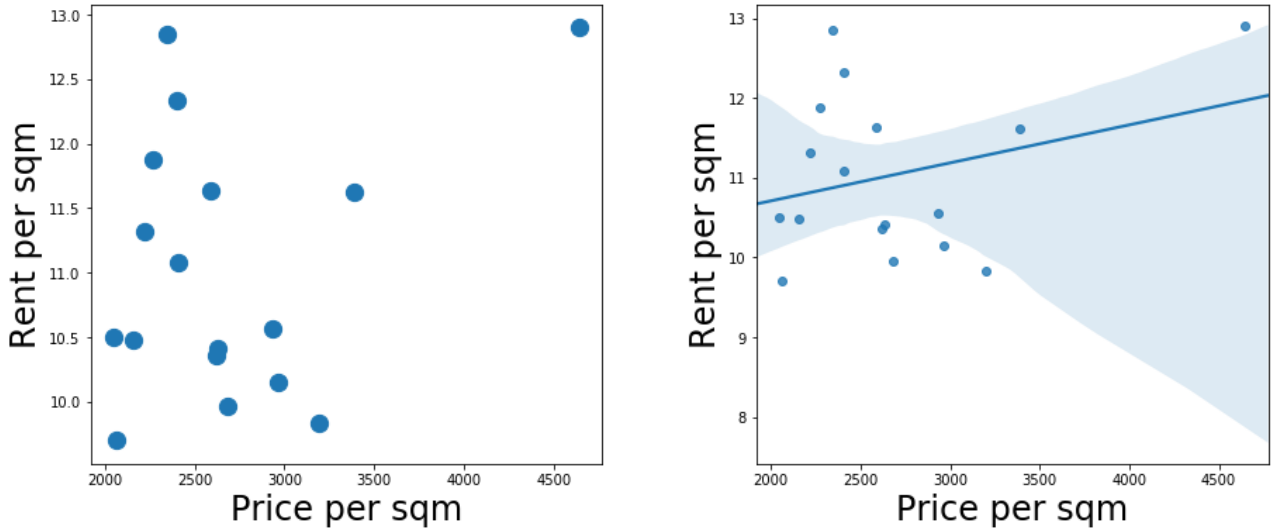

Figure 3: Correlation between realty price and rent per square metre. Left: scatter plot, right: regression plot.

clustered based on the additional information that we acquired. The general approach appears to be feasible.

3.2.2 Rent and Realty Price

Since we want to make recommendations for renting and bying an apartment we should look at a possible correlation between rent and realty price. This relationship is shown in figure 3. It can be seen that extreme values correlate, i.e. the lowest realty price correlates with the lowest rent and the highest price also shows the highest rent. This is not unusual since rent will depend on several factors, such as the size or the state of an apartment and not only the worth of the realty. We can, however, see that the overall correlation of these two entities is rather weak. This independece may lead to different clustering for rent and realty prices.

3.2.3 Population Density and Vibrancy

The fewer data we have, the more it is essential that our variables are independent. Therefore, we should also look at a potential correlation between population density and vibrancy, since densely populated areas could likely be areas with a wider range of entertainment. This analysis is shown in figure 4 and we can see that there is virtually no connection between these two variables. Without the outlier of the inner city, we would most likely even see a downward trent. Again, this is a good sign because it means that our variables can contribute independently to the outcome of the analysis.

3.3 Clustering

Clustering is performed using K-means clustering. As already mentioned in chapter 2, the data used for clustering comprises rent per square metre, price per square metre, borough area, population density and venue count. While rather compact, this data set is capable of describing important aspects of each borough:

1. Larger boroughs leave room for new buildings.

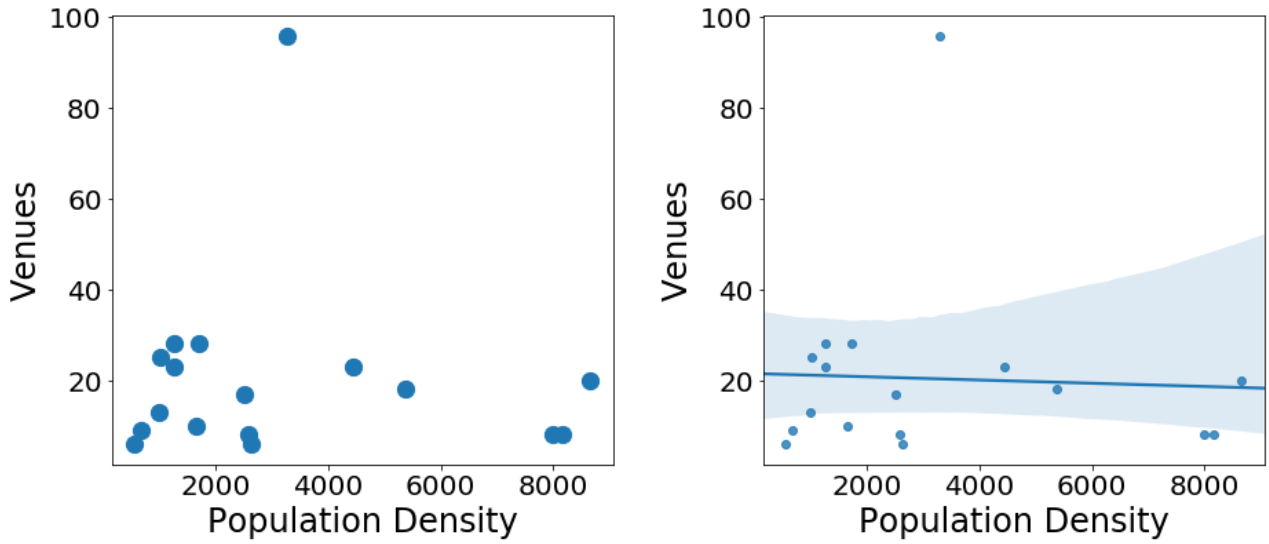

Figure 4: Correlation between population density and vibrancy. Left: scatter plot, right: regression plot.

2. Areas with high population density can be too crowded.
3. Very few venues suggest that areas are calm and provide only little entertainment.

Therefore, the clustering algorithm can be expected to provide a result that will allow to identify areas best suited for renting and for buying property. Generally, three different clusters are expected, namely one recommended for buying, one for renting and one not recommended at all. From figure 2 we can see that the inner city will most probably form a cluster on its own and finally it could be seen that 5 clusters provide the most useful result (see chapter 5 for further details).

Since we could see in figure 3 that the correlation between rent and price is low, clustering will be performed using rent and price information separately to investigate, whether this results in different clusters.

3.4 Assumptions for Cluster Interpretation

The aim of this project is to identify boroughs that can be recommended for renting or buying an apartment based on an objective analysis. Since the clusters are formed based on similarity, it can also be expected that the clusters can be interpreted by looking at their average values. The interpretation of the clusters will be performed as follows:

1. A recommendation for buying requires low or moderate realty prices. Furthermore, the borough has to show moderate vibrancy as a sign for a developing borough.
   The combination of these two characteristics suggests that an apartment will increase its value.
2. To be recommended for renting a borough has to show cheap or at least moderate rent and high vibrancy to provide ample opportunities to socialise.
3. Densely populated areas are not preferred.
4. Little vibrancy is a general drawback.
5. Expensive areas are not recommended at all.

The threshold values for the differentiation will be set empirically based on the results.

Cluster	Rent/m^2 EUR/m^2	Average Price/m^2 EUR/m^2	Area km^2	Population Density $1/km^2$	Venue Count
0	10.94	2566.15	6.11	2941.29	17.57
1	12.90	4642.89	1.16	3287.00	96.00
2	11.66	2363.28	13.68	1097.00	22.00
3	10.53	3184.44	11.00	968.67	8.33
4	10.51	2110.92	3.20	8263.67	12.00

Table 1: Average Values of the clusters formed by the K-means algorithm.

4 Results

Clustering provided the same clusters for both approaches (using either rent or price as a variable). The average results of the clustering are shown in table 1. The clusters can easily be differentiated by their properties and can quite well be interpreted easily by applying the assumptions listed in section 3.4.

- Cluster 0 consits of 7 boroughs, provides moderate rent and a lively neighbourhood.
- Cluster 1 consists of 1 borough and is very expensive.
- Cluster 2 consists of 3 boroughs, shows low rent, high realty prices and low vibrancy in the environment.
- Cluster 3 consists of 3 boroughs and is cheap but very densely populated.
- Cluster 4 consists of 3 boroughs, shows rather high rent, moderate realty prices and a rather vibrant neighbourhood.

The clustering can also be shown on a map (see figure 5) with colour-coded recommendations (an interactive map can be found at the end of the notebook):

1. Dark green: “Buy here! Affordable prices, moderate entertainment.”
2. Light green: “Rent here! Affordable rent, lively environment.”
3. Orange: “Cheap but densely populated.”
4. Light red: “Cheap rent but little entertainment.”
5. Dard red: “Simply expensive.”

5 Discussion

This project aimed at clustering demographic data combined with rent and realty price information to ultimately provide recommendations in which borough of Graz an upcoming student should look for an apartment to buy or to rent. The result is a map with recommendations, which easily allows to further filter for certain regions, for example, to find an apartment closer to a certain university.
The final results look reasonable for a typical student city: the cluster of boroughs recommended for renting is the largest one consisting of 7 boroughs. Furthermore, the inner city is not recommended for students, since it is too expensive and boroughs including or being close to transport hubs are cheap but densely populated.

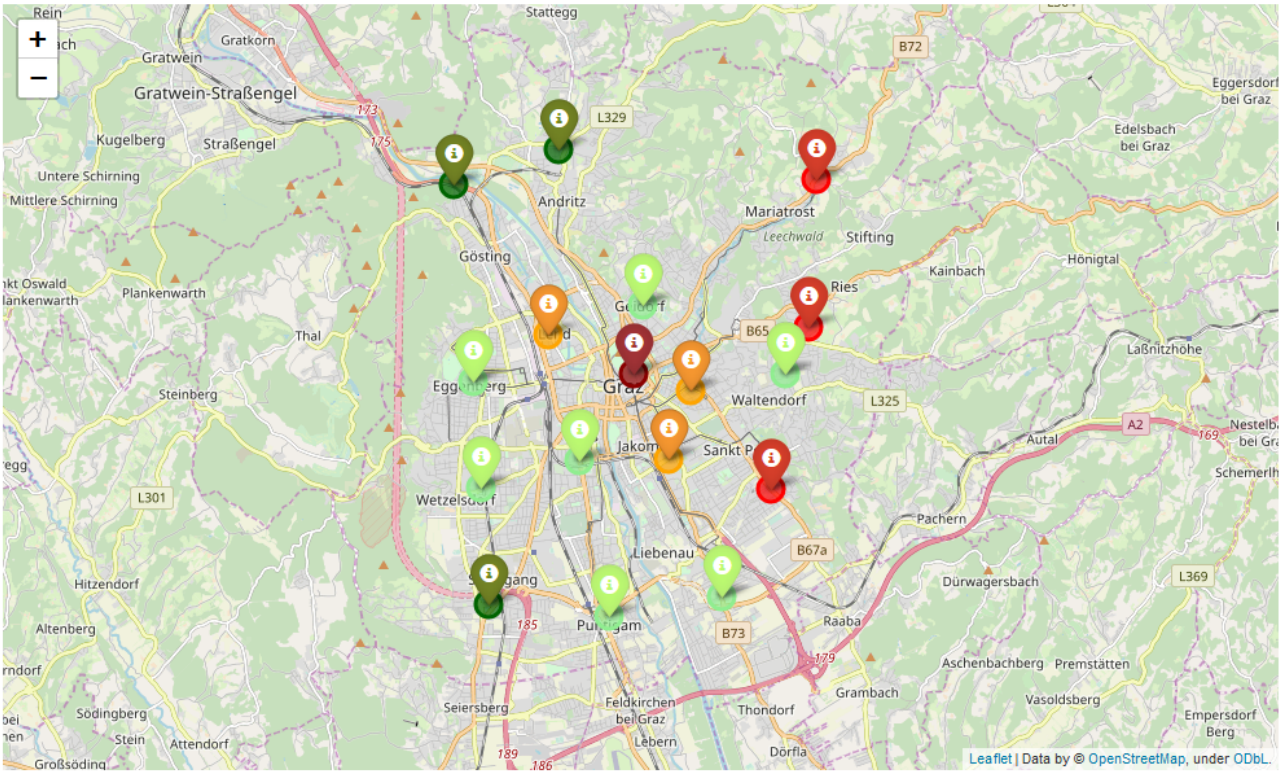

Figure 5: Final result of the clustering. The recommendations are color-coded in the map (dark green for buying, light green for renting, orange for cheap but densely populated areas, light red for little vibrancy and dark red for expensive.)

The recommendation where to buy an apartment was the reason to eventually increase the number of clusters to 5. Using only 4 clusters, cluster 2 and 4 would be merged; however, they definitely show different characteristics: cluster 2 is more expensive, less populated and only shows little vibrancy – characteristics of the periphery. For cluster 4, on the other hand, realty prices are lower and it provides more entertainment – a developing borough promising an increase in value.

6 Conclusion and Outlook

This data analysis project showed that a recommendation system for renting and buying apartments in Graz is generally feasible. The results are derived from an objective analysis, i.e. a clustering of similar characteristics, look reasonable and are based on freely available data. It is furthermore noteworthy that this approach can simply be applied to any other location provided the required information is available.

The approach has still potential for improvement. The analysis could be expanded to even smaller districts, which would most probably identify hotspots in a borough; however, such information is currently not freely avaiable. The data provided by Foursquare also appears limited. Especially for the target group of students a more detailed analysis of the available venues would surely be appreciated but this requires more data. Therefore, for the time being the absolute number of venues is an adequate estimator.

References

[1] URL: https://www.graz.at/cms/beitrag/10034466/7772565/Zahlen_Fakten_Bevoelkerung_Bezirke_Wirtschaft.html.

[2] URL: https://de.wikipedia.org/wiki/Liste_der_Stadtbezirke_von_Graz.

[3] URL: https://www.immobilienscout24.at/immobilienpreise/graz.html.

[4] URL: https://www.immowert123.at/grundstueckspreise/steiermark/grazstadt/.

[5] URL: https://foursquare.com/developers/apps.