The Battle of Students

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 [1].
2. The website *www.immobilienscout24.at* provides rent and property price information for Graz down to borough-level [2]. 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 [3].
4. The Foursquare database [4] 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 [3] 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 [2] and [3] 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 [1] 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 [1] 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 [2] 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 [3] 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 per square metre is calculated from the recommended price and the apartment size. The results are averaged per district or borough.

We know from [1] 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()*.

References

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

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

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

[4] URL: `https://foursquare.com/developers/apps`.