CAPSTONE REPORT

IBM Applied Data Science Course

Summary

For my final project, I chose the fictional business problem of a person with the idea to open a bar in Vienna, the capital of Austria. Location data from Foursquare, a GeoJSON file, and public data about apartment sales within Vienna are jointly employed to create a visual map to guide the decision where to open the bar. Among the tools used to achieve this are the Foursquare API, Folium, and Pandas.

Philipp Schardax

1. Introduction

Fictional character Mr. Smith is very keen to open a bar in Vienna. Due to proximity to universities and areas where a lot of university students live, he knows that locating his bar southeast of Vienna's centre shall be beneficial. However, he wants to avoid direct competition with other bars and clubs in the area, not locatating his bar too close to these. Further, he wants to buy the venue instead of renting it, which for him results in the need to look for a promising venue in an area as cheap as possible. Therefore, he approaches me to create a data-driven map to support his decision.

2. Data

Data of the competition is collected by making a call to the Foursquare API of the area in question. Therefore, the API call was centred around the address Rochusgasse 1 southeast of Vienna's center, with a radius of 1500m. The Foursquare IDs for the categories "Bars" and "Nightclubs" were used to retrieve all relevant existing venues in the area.
In the beginning it also was planned to use a loop in order to retrieve and append each venue's rating to the dataframe, but Foursquare did not allow to do so.

For the data about real estate prices in Vienna, publicly available data was retrieved from https://www.data.gv.at/auftritte/?organisation=stadt-wien. The dataset includes data of around 52.000 sales of real estate in 2019, but turned out to be quite messy.

For drawing the borders of Vienna's districts, a GeoJSON file from this Github depository was employed: https://github.com/codeforamerica/click_that_hood/tree/master/public/data

3. Data Analysis

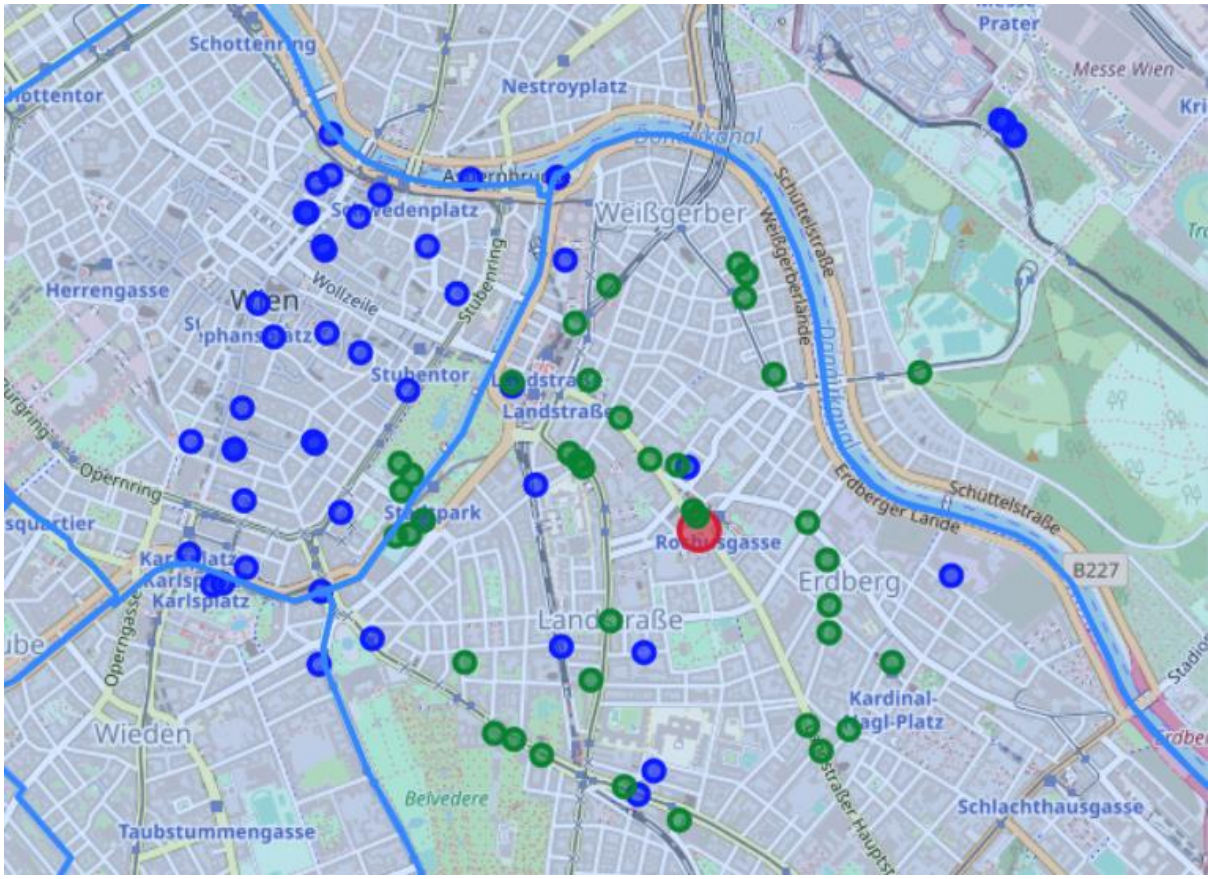
As expected for each data science issue, cleaning and pre-processing the data before analysing it was very time-consuming. This was especially true for the public data about real estate transactions, which was unfortunately quite messy. The 45 columns were quickly reduced to only 6, and the data filtered for only relevant districts and categories, and rows with missing values were dropped. The initial idea was to use prices per m^2 as the basis for analysis only for the category “Mietwohnhaus voll/tw. Vermietet”. However, the dataset did not provide the data for exactly this category. Calculating it with Python neither was possible, as the data set did not include any total area for each property sold. Therefore, the plan was to base the analysis on prices per total property instead (which of course is quite biased, but should do the job for learning purposes).

So, a loop was written to request each property’s latitude and longitude data using GeoPy and append it in a new column to the dataframe. In principle, the loop worked, but timed out quickly and persistently. Unable to request the latitudes and longitudes for thousands of addresses manually, instead I grouped the properties by district and calculated averages for each.

Initially, the plan was to additionally populate the dataframes of bars and nightclubs retrieved from Foursquare with their ratings. A loop was set up to request each venue’s rating using it’s ID. However, this did not work as the service rejected the numerous requests. So, their locations and IDs are now simply visualized in the Folium map. A layer displaying the borders of each districts was created by including the GeoJSON.

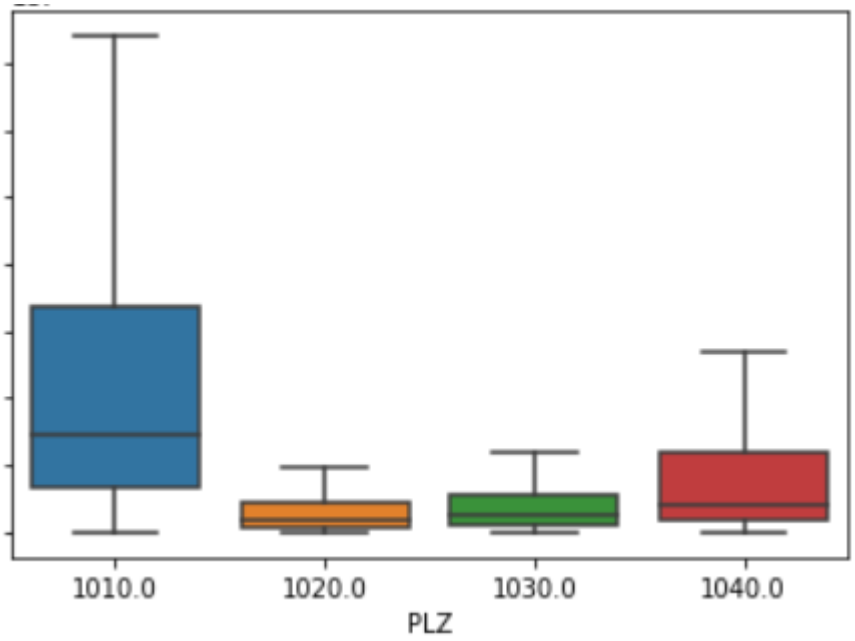
4. Results & Conclusion

The result was a map of all competing venues in the area in question, which also showed the borders of each district. In red, we see the center of the researched area southeast of Vienna's center. In green, other bars within 1000 meters are displayed. In blue, nightclubs within 1500 meters are shown. In lighter blue, we can identify the borders of Vienna's districts.



Further, averages for each district, as well as box plots comparing the distribution of prices for each district, provide guidance on where to buy. The box plots not only display the median prices, but also give a glance how likely it is to find an especially cheap venue.

```
PLZ
1010.0    6474874
1020.0    1051841
1030.0    1371054
1040.0    2190004
Name: Kaufpreis €, dtype: int32
```



As expected, property in the 1st district (1010) is by far the most expensive, followed by the 4th district (1040). In the second and third district (1020 and 1030) property is significantly cheaper, and apparently there is a very low chance to find cheap outliers on the low end. It is worth mentioning that, as the prices are calculated per property and not per m², the outliers are likely caused by extraordinarily small or large property, respectively, and do not necessarily represent good or bad opportunities to buy.

Taking into consideration competition and property prices for each district, Mr. Smith is advised to locate his bar in one of the less competitive areas in the 2nd district shown in yellow below: