Finding the Best Place for a Coffee House in Vienna, Austria

John Oxales

June 7, 2021

1. Introduction

1.1 Background

Vienna’s very first coffee house opened in 1683. Even if Vienna was not the pioneer in coffee house culture, it has - over the centuries - established a coffee house tradition like no other city in the world. Coffee and coffee houses are at their best in Vienna!

1.2 Problem

Since so many coffee houses can be already found in any district in Vienna it would be good to know if there is a strong correlation between the presence of coffee houses in the vicinity of sightseeing places or any other venues like offices or shopping malls. It can be assumed that those places are highly frequented. This is a relevant key-indicator to be considered in order to have a profitable business. Furthermore, the question for a reasonable rent comes up in order to find the ideal spot for the coffee house. So, two main criteria need to be met:

a) area of interest (representing high frequency of passers-by)
b) low rental fee

1.3 Interest

These questions could be of high interest for any new investor or entrepreneur who wants to open one or maybe even more coffee houses in Austria’s capital. This project can also be part of business analytics for larger companies who consider expanding their business.

2. Data acquisition and cleaning

2.1 Data sources

For this project, the following data is going to be used to gain insight and answer the questions:

Data source:

- Foursquare API: "https://developer.foursquare.com/"

Purpose:

- Get all venues in each neighborhood (coffee houses, sightseeing spots, malls etc.)

Data source:

- Vienna rent statistics: “https://www.immopreise.at/Wien/Wohnung/Miete”

Purpose:

- For practice purposes this data is only going to represent a rough overview of mean rent in each district from this year. This should be sufficient in the beginning.

2.2 Data acquisition and cleaning

Data will be scraped from both sources Foursquare and Immopreise and are going to be combined into one table.

2.3 Feature selection

After data cleaning the following features are going to be used:

- Neighborhoods (geolocation coordinates extracted with GeoPy)
- Districts
- Venues (for sightseeing and leisure activity spots)
- Rent prices
- Spatial size (small, medium, big, large)

2.4 Web scraping

I have encountered some problems while scraping ‘Immopreise’ to gather the required rent prices. One issue was that the table was dynamically generated. So, a direct scraping was not possible. I have found a perfect workaround using Selenium, an automation tool for Python. A tutorial on Youtube helped me with all the necessary steps to be taken to make it work.

2.5 Data cleaning

Since the scraped data contained unwanted currency characters, they needed to be deleted in order to get a proper working dataset. Furthermore, the column names and the neighborhood values needed some refinement.

2.6 Extending Data frame

Geospatial data from Vienna’s neighborhoods and their corresponding postal codes were missing and I could not find a fitting dataset for my needs which I manually created in the end. That dataset was additionally merged with my previous data frame.

3. Exploratory Data Analysis

3.1 Map of Vienna

To have a visual overview I started with mapping Vienna first.

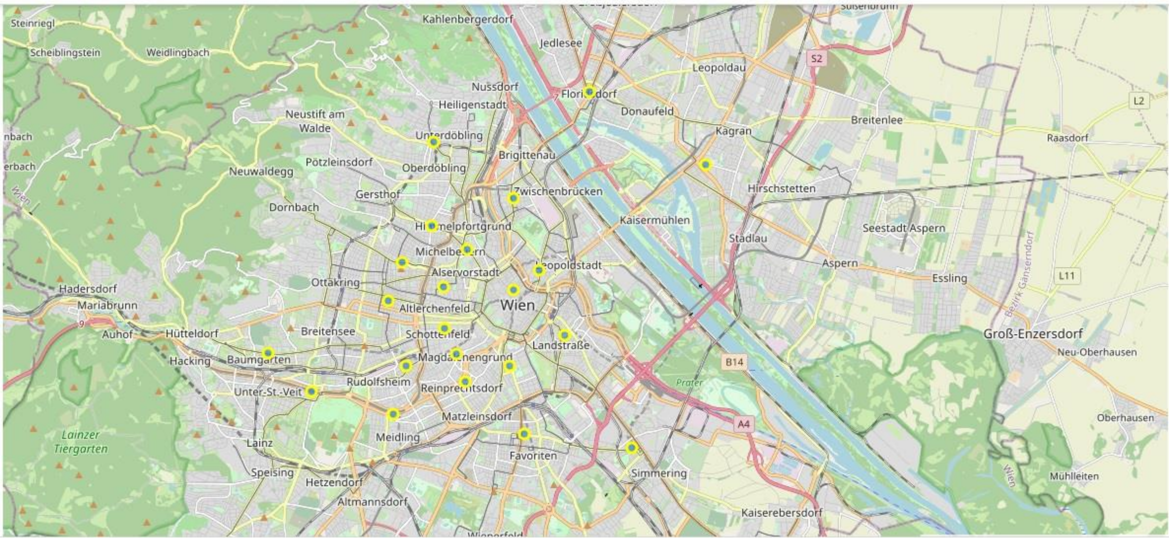


Figure 1. Vienna Map

3.2 Data Visualization

To find out which neighborhoods are the 'cheapest' regarding my question I have utilized the 'meanPrice'-feature which I sorted.

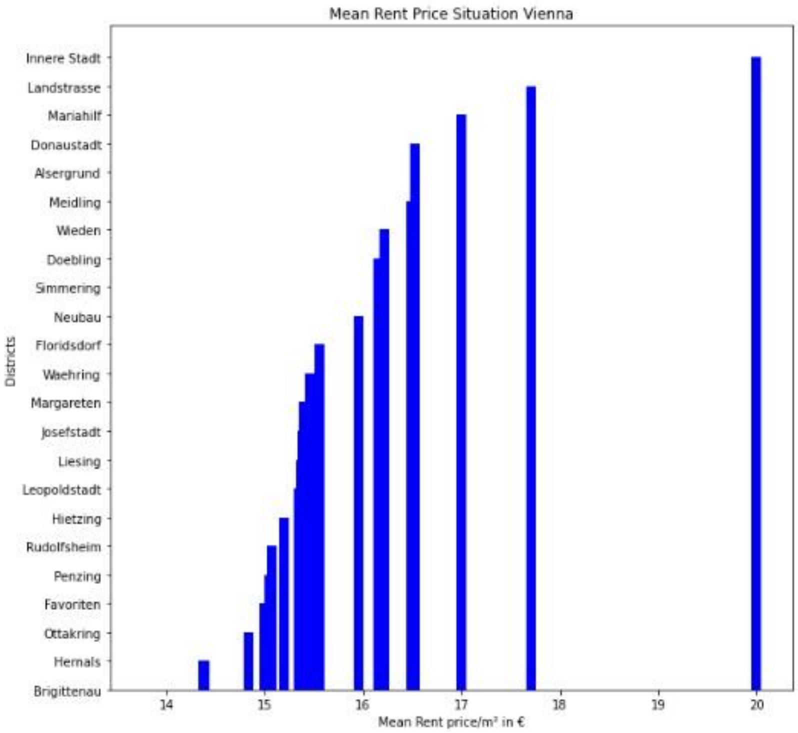
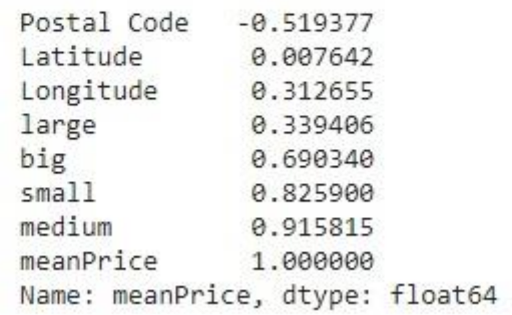


Figure 2. Mean Rent Price Situation in Vienna

3.3 Exploring Neighborhoods

With the Foursquare API I was able to collect all the data to explore the surrounding venues in each neighborhood.
I have decided to establish a price cap with the mean rent which should be my first target to have a starting point. Afterwards I began inspecting each neighborhood. Since I assumed that an ideal spot would be a place with a reasonable price and high frequented, I wanted to find out which neighborhood had the top ten most common venues. My goal was to find a correlation between the density of those venues and compare them with the rent prices. It seemed like ‘Floridsdorf’ is a good point to start with.

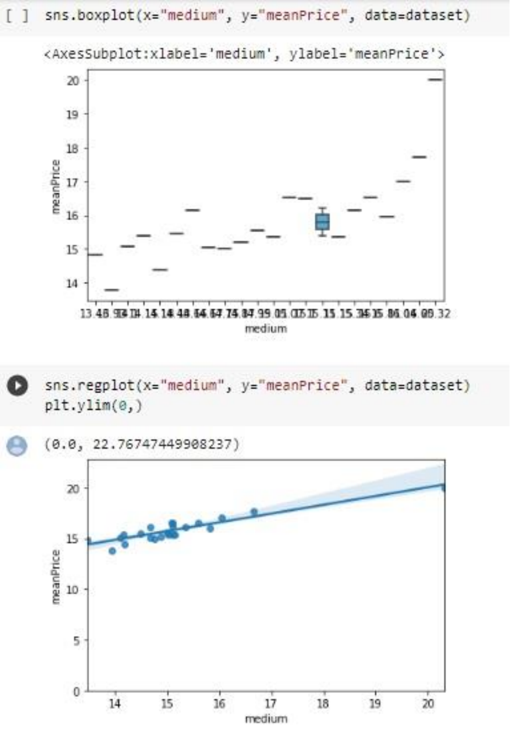
3.4 Find Correlations



```
Postal Code   -0.519377
Latitude       0.007642
Longitude      0.312655
large          0.339406
big            0.690340
small          0.825900
medium         0.915815
meanPrice      1.000000
Name: meanPrice, dtype: float64
```

Figure 3. Correlations of Features

It seems like that there is a high correlation with medium sized objects. Further exploration utilizing this feature was a logical step.



```
[ ]  sns.boxplot(x="medium", y="meanPrice", data=dataset)
```

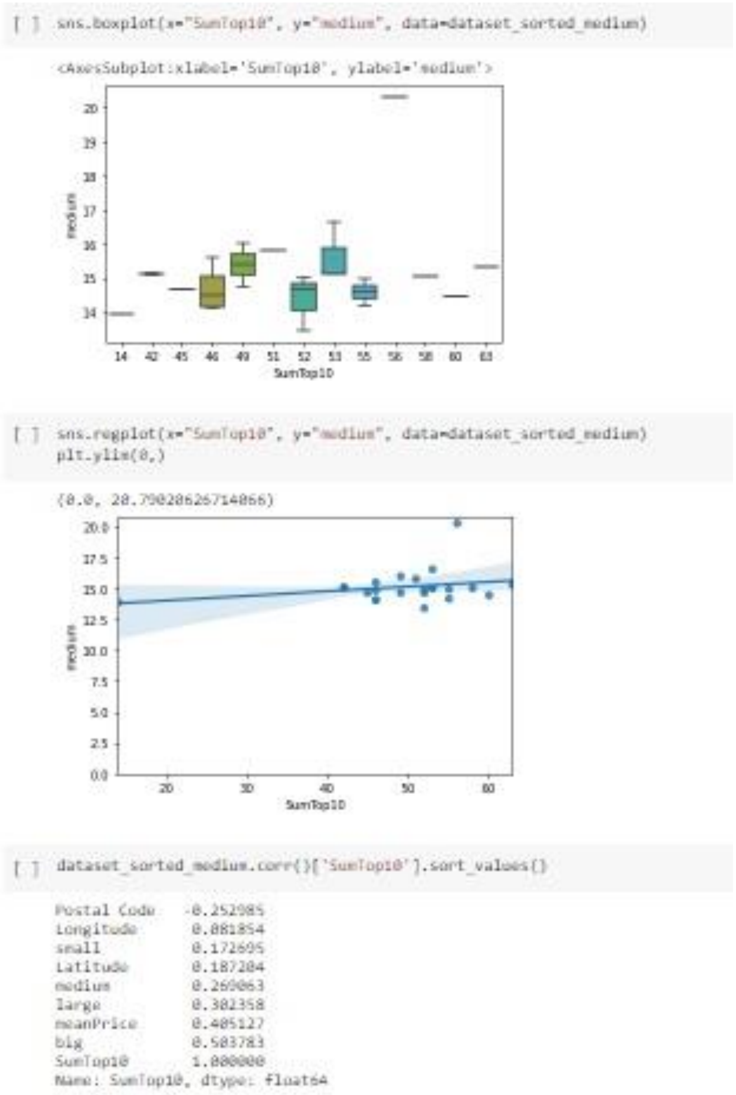
```
<AxesSubplot:xlabel='medium', ylabel='meanPrice'>
```

```
sns.regplot(x="medium", y="meanPrice", data=dataset)
plt.ylim(0,)
```

```
(0.0, 22.76747449908237)
```

Figure 4. Visualizing Correlations

3.5 Correlation with Top 10 common venues



```
[ ]  sns.boxplot(x="SumTop10", y="medium", data=dataset_sorted_medium)
```

```
<AxesSubplot:xlabel='SumTop10', ylabel='medium'>
```

```
[ ]  sns.regplot(x="SumTop10", y="medium", data=dataset_sorted_medium)
     plt.ylim(0,)
```

```
(0.0, 20.79028626714066)
```

```
[ ]  dataset_sorted_medium.corr()['SumTop10'].sort_values()
```

```
Postal Code   -0.252985
Longitude      0.081854
small          0.172695
Latitude       0.187204
medium         0.269063
large          0.302358
meanPrice      0.405127
big            0.503783
SumTop10       1.000000
Name: SumTop10, dtype: float64
```

Figure 5. Visualizing Correlations with Top 10 common venues

4. Model Development

4.1 Finding the best model

After splitting my datasets into training and testing samples and removing some outliers I tried to predict the best model. Figure 6 visualizes the result which model would suit best.

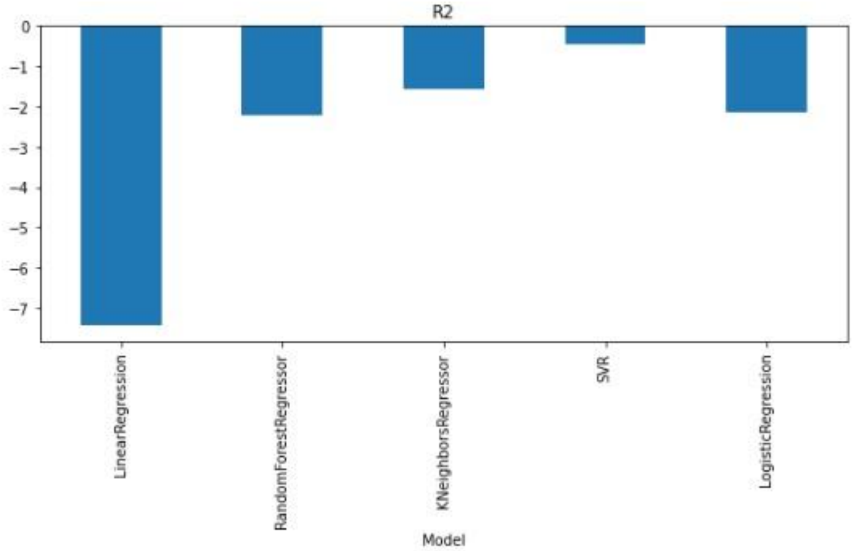


Figure 6. Predict the best Model

The outcome is a problem. SVR (Support Vector Regression) suggests applying a Support Vector Model. But every output is negative. A better dataset is needed.

5. Conclusion

5.1 Summary

After applying different models with the given datasets, a sufficient statement cannot be made at this point. Obviously better and bigger datasets are needed e.g., data with tourist-frequency, offices, passers-by and number of rentable objects from each neighborhood. To have at least an idea where a suitable place could be for a coffee house, comparing the amount of top 10 common venues in relation to the mean rent situation in each district might be raw guideline. Since this is just for practice and studies purpose only and because of time issues, a cut at this point is reasonable. For future and further analysis, a new approach is going to be needed.

6. Future Directions

To have at least an idea I visualized a raw overview comparing the mean rent price situation with the top ten common venues in each district. It can be assumed that in ‘Floridsdorf’ might be an ideal place to open a new coffee house. It is within my declared price cap and appears to have enough frequented venues. Given these parameters this should be examined in a future exploration.