IBM Applied Data Science Capstone Project:

Analysing the Neighbourhoods of Vienna (AT)to select a Restaurant Location

ABSTRACT

The prosperous growth of Vienna was identified as an opportunity to open a restaurant in the city. The analysis carried out here using clustering algorithms identify suitable districts based on Foursquare data. Prior to making an investment decision it is recommended to verify the analysis with an alternative data source.

Daniel Kunaver

IBM Applied Data Science Program

Date: 20th February 2021

Table of Contents

List of Figures

List of Tables

1. Introduction

The city of Vienna is perennially ranking as one of the cities with the highest quality of life. Official statistics show the growth of the city in terms of population and tourism. This is reflected in the two charts shown below on population growth and city visitor numbers. On the premise of growth it is enticing to consider an investment into something like a restaurant to harvest some of the growth in population and visitors through a restaurant. Especially, since a city with one of the highest quality of life ratings indicates also that people do like to go out and enjoy music, theatre, food, drink and other spectacles. One would consider that placing the restaurant in the historic city center may be a foregone conclusion, due to the high number of tourists, but it would be best to analyse the data and find the perfect niche. What needs to be considered is the rental cost in the area, which is in the tourist hot spots the highest. This analysis will look where a restaurant can be placed conveniently to ensure a large catchment area which also boasts a large disposable total income of the district or area.

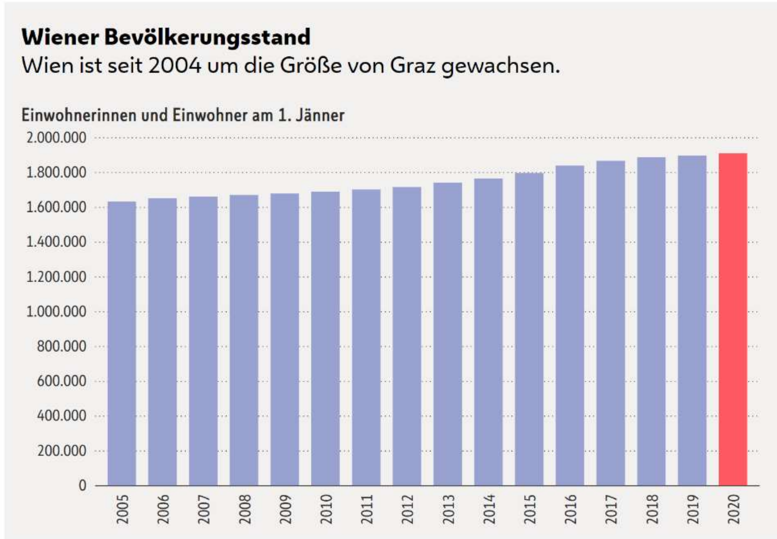

Figure 1: Population growth in the city of Vienna (AT) between 2005 and 2020

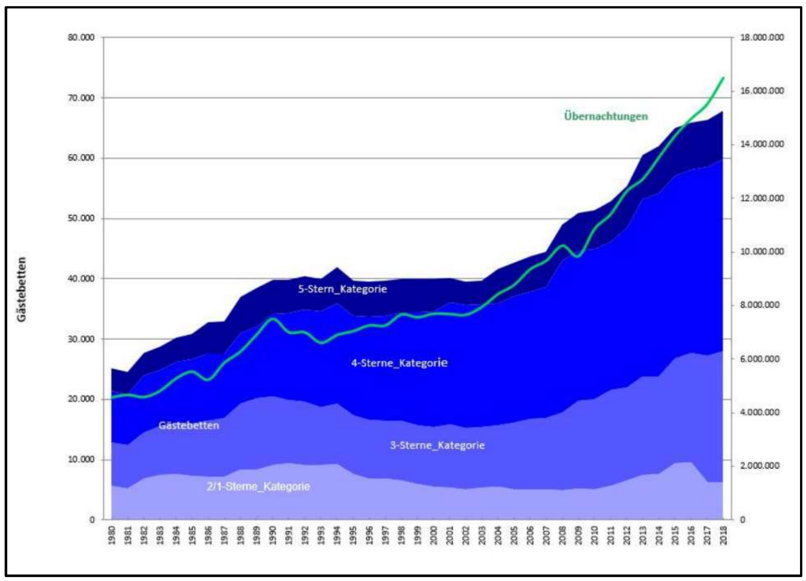

Figure 2: Development of visitor overstay nights (green line – right scale) and total number of guest beds available (blue area by category – left scale) in Vienna (AT) between 1980 and 2018

Based on the described setting the business problem reviewed in this study is as follows:

A) Is the city center the best location for placing the restaurant?

B) What are alternative locations for placing a restaurant?

2. Data

The following data was used for the study:

- List of districts of Vienna (source: https://de.wikipedia.org/wiki/Wiener_Gemeindebezirke)
- Income of people inside the districts (source: https://www.wien.gv.at/statistik/arbeitsmarkt/tabellen/einkommen-gesamt-bez.html)
- Number of overnight stays by visitors in Vienna (source: https://www.wien.gv.at/statistik/wirtschaft/tabellen/uebern-bezirk-zr.html)
- Foursquare data through the Foursquare developer API (see: https://de.foursquare.com/developers/apps)
- Geolocation data through the geopy library (see: https://geopy.readthedocs.io/en/stable/)

The data consists of several tables, which have been scraped from the websites and loaded into a pandas dataframe. Extracting the data from the various sources resulted in a data frame, which is summarized in Table 1 below.

Table 1: Data collected from online sources on the city of Vienna (AT).

Dsitrict Number	District	Area (hectare)	Inhabitants	Employees in District	Population Density per km²	Avg Annual Net Income per Person (€)	Overnight Stays (millions)
1	Innere Stadt	286.9	16,047	108,679	5,593	26,480	3.12
2	Leopoldstadt	1924.2	105,848	66,945	5,501	22,904	2.19
3	Landstraße	739.8	91,680	101,100	12,393	24,172	1.76
4	Wieden	177.5	33,212	28,439	18,711	25,878	0.79
5	Margareten	201.2	55,123	20,567	27,397	20,479	0.61
6	Mariahilf	145.5	31,651	28,676	21,753	23,971	0.73
7	Neubau	160.8	31,961	33,592	19,876	25,100	1.21
8	Josefstadt	109	25,021	15,762	22,955	25,142	0.53
9	Alsergrund	296.7	41,884	49,847	14,117	24,701	0.60
10	Favoriten	3182.8	207,193	76,051	6,510	19,478	1.83
11	Simmering	2325.6	104,434	36,983	4,491	21,606	0.30
12	Meidling	810.3	97,078	38,336	11,981	20,537	0.23
13	Hietzing	3771.5	54,040	24,184	1,433	29,575	0.29
14	Penzing	3376.3	93,634	29,830	2,773	23,755	0.48
15	Rudolfsheim-Fünfhaus	391.8	76,813	29,852	19,605	18,528	0.98
16	Ottakring	867.3	103,117	28,509	11,889	21,168	0.23
17	Hernals	1139.1	57,027	15,070	5,006	22,386	0.35
18	Währing	634.7	51,497	14,364	8,114	26,770	0.03
19	Döbling	2494.4	73,901	31,901	2,963	28,004	0.21
20	Brigittenau	571	86,368	29,541	15,126	18,674	0.31
21	Floridsdorf	4444.3	167,968	55,691	3,779	23,220	0.06
22	Donaustadt	10229.9	195,230	63,126	1,908	25,323	0.70
23	Liesing	3206.2	110,464	53,963	3,445	26,063	0.08

The data shows that just over 1.9 million people are registered to live in Vienna. Just under one million people do work in the city (this includes commuters from outside the city). A total of 17.6 million overnight stays happened in in 2019. The data will be further explored in the Methodology Section.

3. Methodology

The following methodology has been used to analyse the data:

- Data acquisition
- Data exploration
- Plot district data on map
- Find venues in districts via Foursquarte query
- Do cluster analysis
- Identify from clusters optimum solution

The data acquisition was covered in the previous section. The following section is detailing the remaining steps pointed out above.

4. Analysis

The section will go through the data analysis using the various methods pointed out in the previous section.

A. Data Exploration

The statistical details of the data set (from Figure *1Figure 1*) can be seen below in Table 2.

Table 2: Statistics on the Vienna District Data Set.

	Area (hectare)	Population	Employees in District	Population Density (per km²)	Avg Annual Net Income (€)	Overnight Stays by Visitors
count	23	23	23	23	23	23
mean	1803.8	83095	42653	10753	23648	23648
std	2290.0	51538	25857	7835	2954	2954
min	109.0	16047	14364	1433	18528	18528
25%	291.8	46691	28474	4135	21387	21387
50%	810.3	76813	31901	8114	23971	23971
75%	2838.6	103776	54827	16918	25601	25601
max	10229.9	207193	108679	27397	29575	29575

One can see from the statistics that the area of the different districts varies significantly, such that the smallest and largest district areas differ by a factor of 100. Further, the population per district is also ranging significantly. Here the minimum and maximum values only differ by a factor of 12. Another notable set of data is the average net income per person. Here on can see low levels of net income and secondly the variation between to lowest and highest value is only by a factor of 1.6. The reason could be that either the values are suppressed due to the highly mixed nature of the districts with social housing amongst villas or there may be underreporting of real incomes.

Graphically, the data acquired from the various online sources is shown in the panel below.

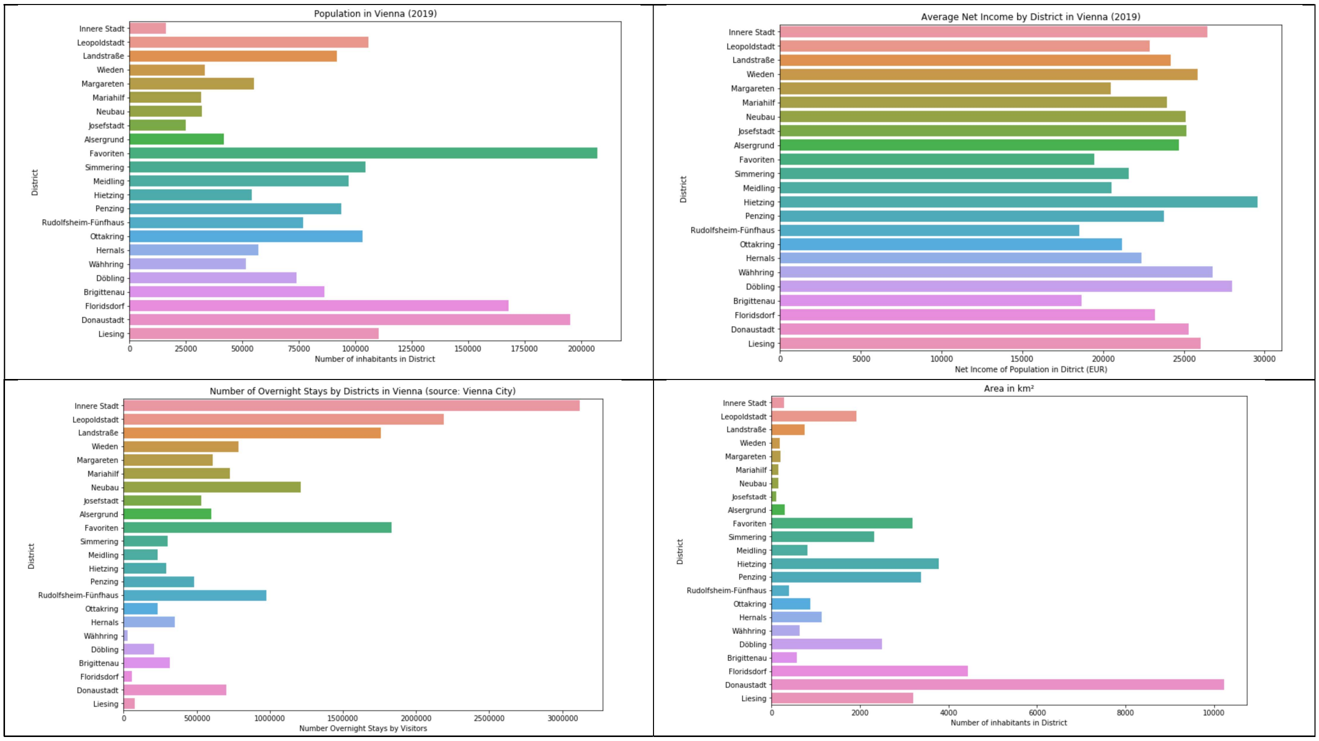

Figure 3: Bar plots on data by district - Top right: Population distribution, Top Left: Average net income (annual), Bottom Right Visitor overnight stays, Bottom Right: Area

B. Putting Data on the Map

To get the data of the districs on the map it was necessary to use a python library to find the latitude and longitude of each district. For that effort the OpenCage library was used. To use this package it is necessary to open a free account on open page and then use the generated key to connect to the API from open cage. The package can find then based on names and other word strings a location and assigns the corrdinates to it. Running the section of code as recommended by the open page guide results in an appended data frame, where for each location entry longitude and latitude data is added. This was then plotted on a map using the folium library. The resulting map with the 23 districts of Vienna is shown below.

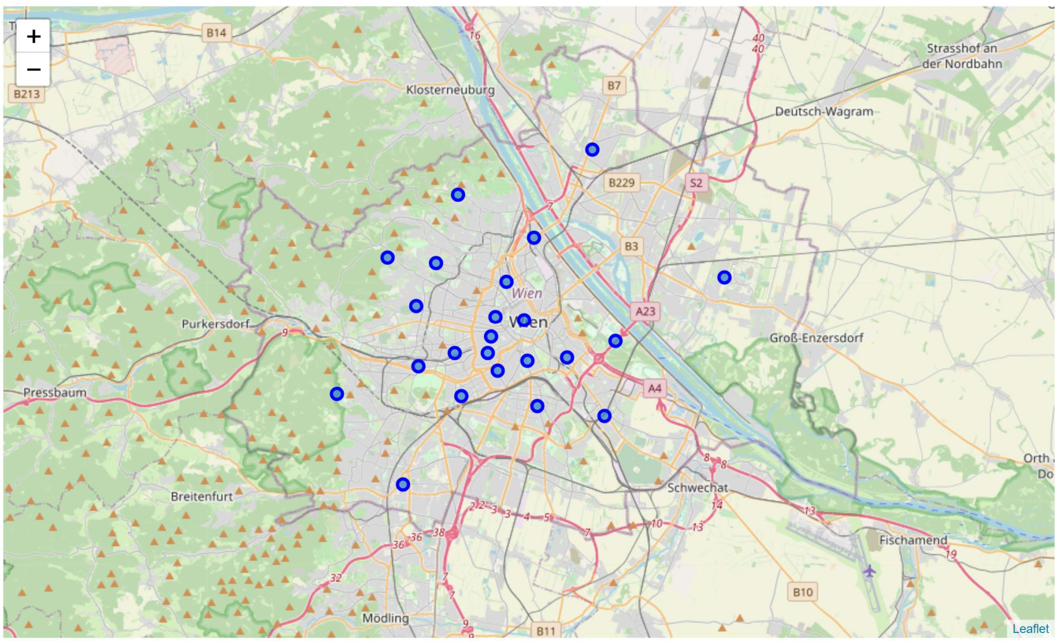

Figure 4: Map of Vienna (AT) showing the center point of each district (blue circle).

C. Find Data on Foursquare

The data was loaded from Foursquare through the API set up with the developer account on Foursquare. It was intended to load a large number of entries to ensure that everything is covered when coming to the analysis. Once the data was loaded the table below shows that most venue counts for the districts appear to max out at 100. It is not clear whether this is the maximum number of data points in each districts, because the setting for the query was 3000 data points, which has not been reached.

Table 3: Show of data points loaded per district.

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alsergrund	100	100	100	100	100	100
Brigittenau	100	100	100	100	100	100
Donaustadt	29	29	29	29	29	29
Döbling	91	91	91	91	91	91
Favoriten	100	100	100	100	100	100
Floridsdorf	76	76	76	76	76	76
Hernals	44	44	44	44	44	44
Hietzing	50	50	50	50	50	50
Innere Stadt	100	100	100	100	100	100
Josefstadt	100	100	100	100	100	100
Landstraße	100	100	100	100	100	100
Leopoldstadt	55	55	55	55	55	55
Liesing	79	79	79	79	79	79
Margareten	100	100	100	100	100	100
Mariahilf	100	100	100	100	100	100
Meidling	100	100	100	100	100	100
Neubau	100	100	100	100	100	100
Ottakring	99	99	99	99	99	99
Penzing	100	100	100	100	100	100
Rudolfsheim-Fünfhaus	100	100	100	100	100	100
Simmering	100	100	100	100	100	100
Wieden	100	100	100	100	100	100
Wähhring	100	100	100	100	100	100

When this was boiled down to the number of restaurants, the number of venue fell significantly (see TTT below). It needs to be pointed out that the number of restaurants found does not agree with the total number of restaurants registered through the chamber of commerce[i]. Here the total number of restaurants is over 2500, whereas the number found in Foursquare is under 700 (i.e. less than a third). It is therefore questionable, whether the data used is sufficient to reduce risk for an investment decision of opening a restaurant in Vienna.

Table 4: List of restaurants by district from Foursquare.

District	Number of restaurants
Alsergrund	22
Brigittenau	33
Donaustadt	9
Döbling	28
Favoriten	26
Floridsdorf	16
Hernals	17
Hietzing	12
Innere Stadt	17
Josefstadt	20
Landstraße	22
Leopoldstadt	10
Liesing	19
Margareten	29
Mariahilf	26
Meidling	24
Neubau	20
Ottakring	36
Penzing	28
Rudolfsheim-Fünfhaus	31
Simmering	21
Wieden	25
Währing	33

D. Cluster Analysis

The cluster analysis was attempted in order to find common properties between districts. The aim was to find clusters which are common on 3 factors, which were deemed to be important for a restaurant business. These factors were:

- The net purchasing power per district
- Density of restaurants per local population
- The number restaurants per overnight stay
- The number of restaurants per number of employees in the district

The first factor was chosen, because there direct measure of net income per person was not good enough to see the whole potential of purchasing power. For example, if a small district with few inhabitants have a high net income, then the number of possible customers is low compared to a large district with a slightly lower average net income. The other factors listed are important, because they show the ratio of restaurants to possible customers from different customer groups.

The cluster analysis was done after normalizing the data using the scikit learn package. The first step was to determine the number of possible clusters that are needed for analysis using the elbow method. Once the k-value for the cluster number was identified, the clustering was performed and then plotted on the city map.

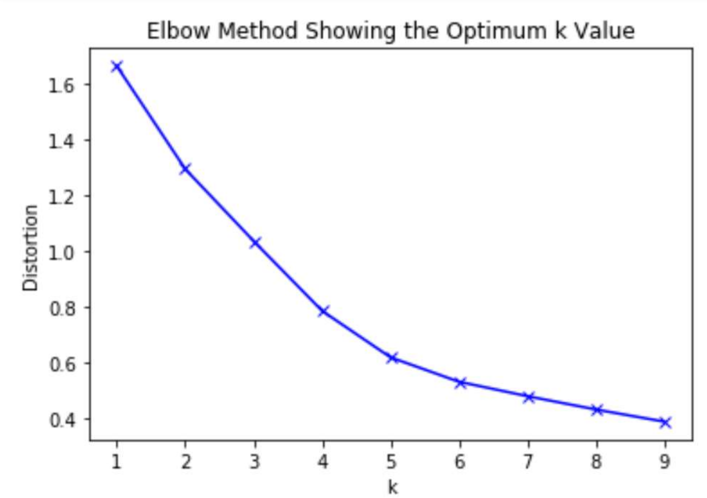

Figure 5: Plot of k-values and corresponding distortion to determine the k-value.

From the elbow analysis it appears that the data should be best clustered into 5 groups. Therefore k was set to 5 for the clustering analysis.

Below the total list of districts and their assigned cluster (second column) is shown.

Table 5: List of Districts with cluster labels.

	Cluster Labels	Dsitrict_#	District	Neighbourhood	Area_sqm	Inhabitants	Emp_District	Inhab_Density	Avg_Net_Inc	Night_Stays
0	2	1	Innere Stadt	Innere Stadt	286.9	16047	108679	5593.238062	26480	3119868
1	0	2	Leopoldstadt	Jägerzeile, Leopoldstadt, Zwischenbrücken	1924.2	105848	66945	5500.883484	22904	2190429
2	0	3	Landstraße	Landstraße, Erdberg, Weißgerberviertel	739.8	91680	101100	12392.538524	24172	1759510
3	2	4	Wieden	Hungelbrunn, Schaumburgergrund, Wieden	177.5	33212	28439	18710.985915	25878	786109
4	3	5	Margareten	Hundsturm, Laurenzergrund, Margareten, Matzlei...	201.2	55123	20567	27397.117296	20479	610601
5	2	6	Mariahilf	Gumpendorf, Laimgrube, Magdalenengrund, Mariah...	145.5	31651	28676	21753.264605	23971	725723
6	2	7	Neubau	Altlerchenfeld, Neubau, Sankt Ulrich, Schotten...	160.8	31961	33592	19876.243781	25100	1209783
7	2	8	Josefstadt	Alservorstadt, Altlerchenfeld, Breitenfeld, Jo...	109.0	25021	15762	22955.045872	25142	531023
8	2	9	Alsergrund	Alservorstadt, Althangrund, Himmelpfortgrund, ...	296.7	41884	49847	14116.616111	24701	597528
9	4	10	Favoriten	Favoriten, Inzersdorf-Stadt, Oberlaa, Rothneus...	3182.8	207193	76051	6509.771271	19478	1833720
10	0	11	Simmering	Albern, Kaiserebersdorf, Simmering	2325.6	104434	36983	4490.626075	21606	302556
11	0	12	Meidling	Altmannsdorf, Gaudenzdorf, Hetzendorf, Obermei...	810.3	97078	38336	11980.501049	20537	230717
12	0	13	Hietzing	Hietzing, Unter-St.-Veit, Ober-St.-Veit, Hacki...	3771.5	54040	24184	1432.851651	29575	289729
13	3	14	Penzing	Baumgarten, Breitensee, Hadersdorf-Weidlingau,...	3376.3	93634	29830	2773.272517	23755	481295
14	3	15	Rudolfsheim-Fünfhaus	Rudolfsheim, Fünfhaus, Sechshaus	391.8	76813	29852	19605.155692	18528	976028
15	3	16	Ottakring	Neulerchenfeld, Ottakring	867.3	103117	28509	11889.426957	21168	232769
16	3	17	Hernals	Hernals, Dornbach, Neuwaldegg	1139.1	57027	15070	5006.320780	22386	349897
17	1	18	Währing	Gersthof, Pötzleinsdorf, Währing, Weinhaus	634.7	51497	14364	8113.596975	26770	26310
18	3	19	Döbling	Grinzing, Heiligenstadt, Josefsdorf, Kahlenber...	2494.4	73901	31901	2962.676395	28004	205745
19	3	20	Brigittenau	Brigittenau, Zwischenbrücken	571.0	86368	29541	15125.744308	18674	314415
20	4	21	Floridsdorf	Donaufeld, Floridsdorf, Großjedlersdorf, Jedle...	4444.3	167968	55691	3779.402831	23220	55738
21	4	22	Donaustadt	Aspern, Breitenlee, Essling, Hirschstetten, Ka...	10229.9	195230	63126	1908.425302	25323	699926
22	0	23	Liesing	Atzgersdorf, Erlaa, Inzersdorf, Kalksburg, Lie...	3206.2	110464	53963	3445.324683	26063	75154

Going through the clustering it appears that the clustering was done as follows:

- Cluster 0: either low density or low income and low number of visitors and employees (a total of 6 districts in here)
- Cluster 1: Outlier, because it consists of only one district, which is middle of the range in most categories, but has the highest number of restaurants per person.
- Cluster 2: The inner city districts with high density and reasonabley high income and good visitor numbers and many employees working in district (a total of 6 districts in here).
- Cluster 3: The poorest districts or and the sprsest populated disctrict lumped together (a total of 7 districts in here).
- Cluster 4: These are 3 of the four largest districts by area, therefore low density of restaurans, but the districts have reasonable income and low visitor numbers (a total of 3 districts in here).

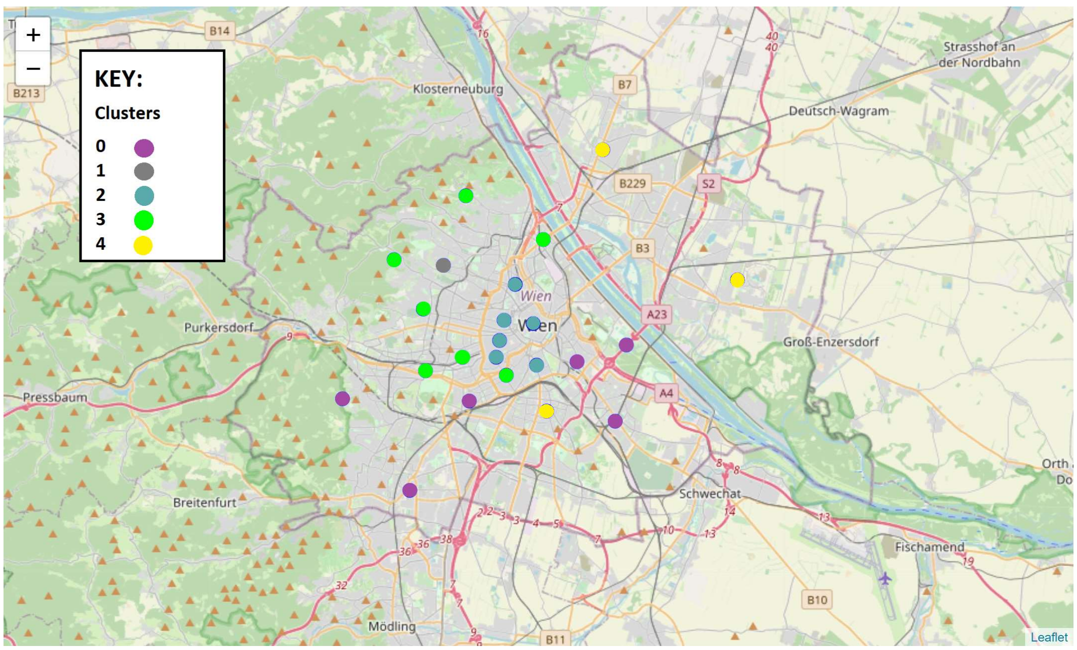

Figure 6: Plot of identified clusters onto Vienna map.

E. Identify Optimum Cluster from Analysis

From the clustering analysis the identification of the optimum cluster had to be found. The objective was to have a low restaurant per customer group ration (i.e. population, visitors and employees) and have a district, where the net purchasing power is also high.

The plots below show the results, which lead to cluster number 2 being selected.

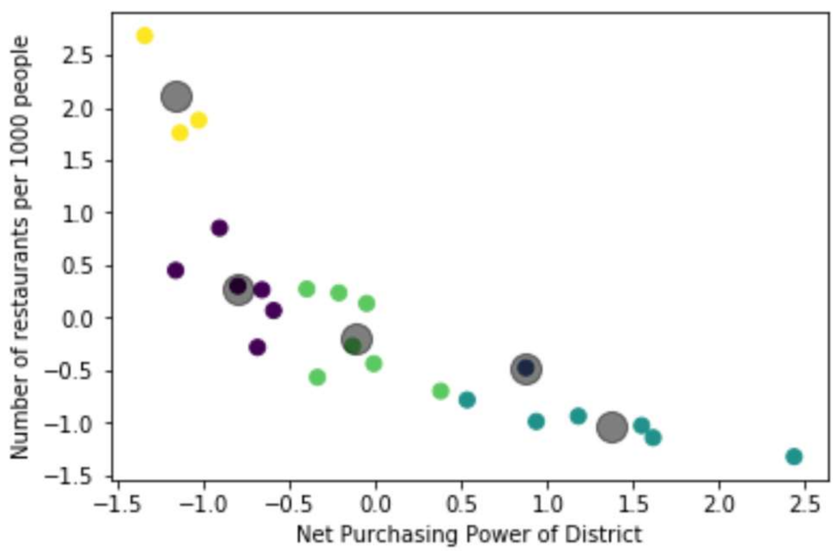

Figure 7: Plot of clustersfor number of poeple per restaurant vs net purchasing power.

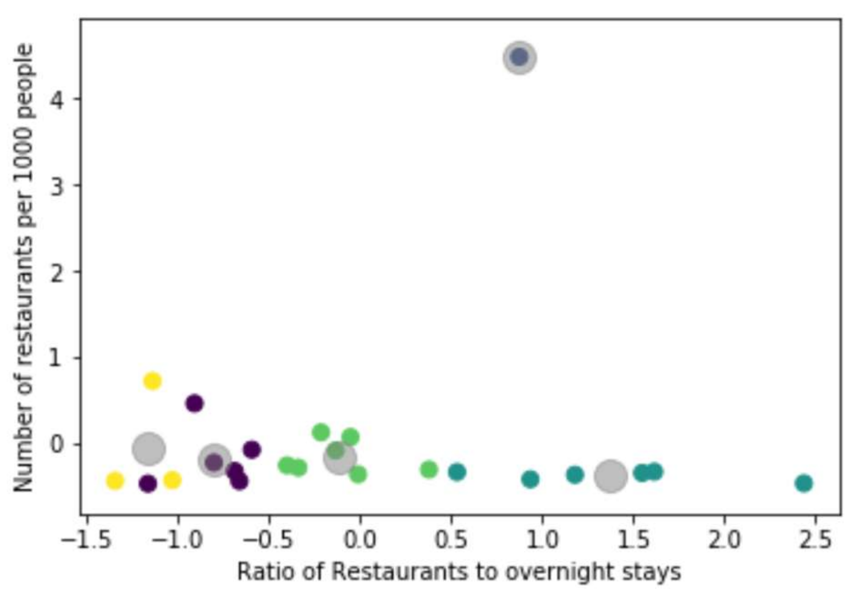

Figure 8: Plot of clustersfor number of poeple per restaurant vs ratio of restaurants to overnight stays.

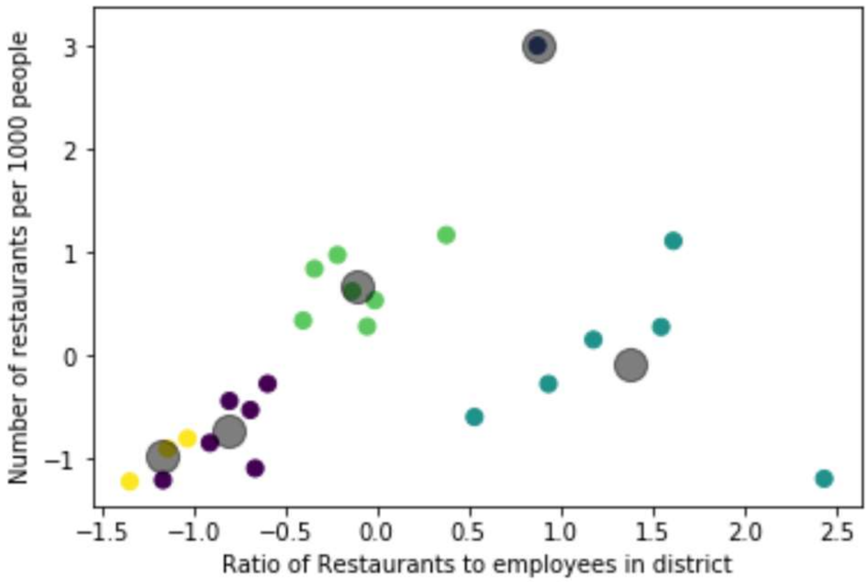

Figure 9: Plot of clustersfor number of poeple per restaurant vs ratio of restaurants to employees.

5. Discusion of Results

The analysis showed that the various districts have been clustered into 5 groups. The subsequent plot of the clusters versus the important parameters of net purchasing power of the districts, number of visitors overnight stays, employees working in the district and number of restaurants per inhabitants in district showed that one cluster is an outlier (cluster number 1) with the most unfavourable conditions to invest in. It became also clear that one of the clusters (cluster number 2) is the most favourable, because it distinguishes itself from clusters number 0, 3 and 4 by the fact that it has highest purchasing power, highest number of visitors and most people working in these districts. These factors indicate that there is a high demand for dining during the day and at night. What was not considered in the data is the fact that most hotels have also restaurants, which may reduce the number of visitors from the restaurants on the street.

6. Conclusion

The aim of the analysis was to find the area with the best conditions to open a restaurant in Vienna. For that matter data on the districts was collected mostly from the city of Vienna authorities. The data was then analysed in conjunction with location information about restaurants in Vienna. The combined data set was then analysed using the clustering method based on the parameters of net purchasing power, number of overnight stays, number of employees and number of restaurants per district population. The result showed that one cluster (number 2) was a clear favourite based on the data provided by Foursquare. This cluster consists of the districts 1, 4, 6, 7, 8 and 9. These are commonly known also as the inner city districts within the city. One final cautionary note needs to be made before starting the investment into a restaurant and location selection: the Foursquare data is not complete as shown in the report, where a link to the Trade Council of Vienna is given indication a factor of 10 difference in the number of restaurants. It would be recommended to compare the information on restaurants to other sources and redo the analysis to gain certainty prior to making a decision on the location.

7. Appendix

A. Footnotes

[i] See https://www.wko.at/branchen/tourismus-freizeitwirtschaft/gastronomie/STATISTIK-UeBER-ALLE-BETRIEBE-BL-2020-Mitglieder.pdf