This Photo by Unknown Author is licensed under CC BY-SA

Vienna Cafés District Map

Finding the best location for opening a new coffee place in Vienna

Coursera IBM Data Science Capstone Project

Katerina Mincheva, April 2019

1. Introduction

Vienna is known for its strong coffee culture, which can be also recognized by the solid number of cafés and roasteries in the city, as well as the good quality of the coffee, that the Viennese highly value. It is a renowned cultural trait as well as a habit of the people to go out and enjoy fresh brewed coffee in local cafés. Furthermore, the vibrant city atmosphere, the big number of students, as well as that of tourists that are looking to experience the coffee culture, secure the constant demand for such places. This is the reason why both established café owners seek to expand their businesses in new locations, as well as new entrants see the market as an attractive and never too saturated to enter.

Thus, the aim of this report is to create an overview of the 'Café map' of Vienna that will have many uses. It will provide a cluster analysis and a valuable overview of the amount of coffee places in each of Vienna's so-called districts. This could, on one hand, be used by entrepreneurs looking for opening a new café in the right location. However, it is also a tool that could help students or Vienna's visitors to choose, for example, an accommodation in an area with bigger saturation of renowned coffee places.

In a business perspective, we will look for an answer to the question: **Which districts of Vienna offer the most attractive location for opening a new café?**

2. Data

The data to be used in the scope of the project comes from few different sources:

i. A Wikipedia article containing the needed information about the 23 different districts in Vienna, called 'Bezirke' (German): https://de.wikipedia.org/wiki/Wiener_Gemeindebezirke. In order to extract the required information - the number of the district and its name, that is contained in a table, web scrapping technique will be applied.

ii. Location data - the coordinates of each district, will be then collected through the Python Geocoder. Having each district's coordinates will then be used in the next step.

iii. Foursquare - after having each district’s coordinates, analysis of all Café venues will be done by querying data and gathering all relevant information through the Foursquare API. Foursquare provides information related to venues in a specific area, such as location, venue category, reviews and tips. Data will be extracted by using the latitude and longitude coordinates of each district and saved in a json object, which will then be transformed in a data frame with the corresponding venue information. Based on that, a clustering of the districts will be done.

iv. In addition, from Statistics Austria we will extract information regarding the population at the end of 2018 of each district, to later see if there are districts with bigger population, but smaller amount of coffee places: http://statistik.at/web_en/statistics/index.html, a potentially attractive location for opening up new cafés.

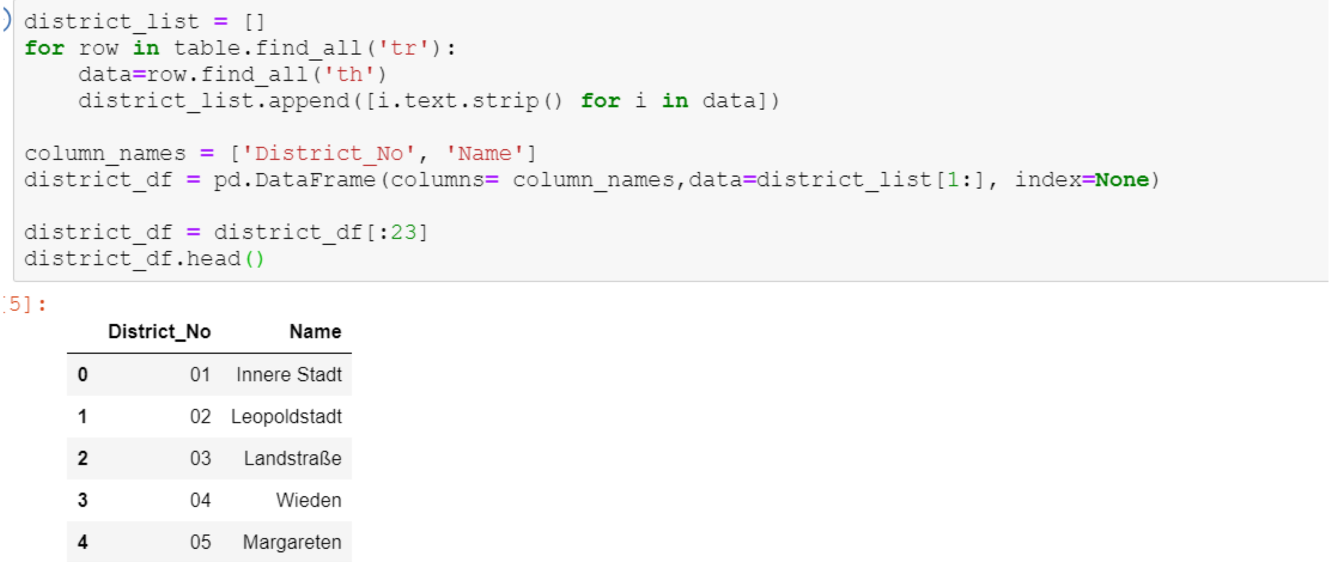
3. Methodology

3.1. Data preprocessing

In this section, all sources of data will be consolidated into one data frame that will be used for the analysis of the problem.

3.1.1. District information

The city of Vienna is separated in the so-called districts (German: 'Bezirke'). They are 23 in total. The information therefore was gathered from a Wikipedia article that lists all relevant information for each district. In order to get the information, a web scraping technique is applied with the BeautifulSoup library for Python. From that source (see the link in p.2) was extracted data regarding the districts' names and their numbers (1 to 23). The source contains further information, i.e. regarding the population of each district, but we will not use it, since the information is for 2017. Later on, we extract this data for 2018 from another source.

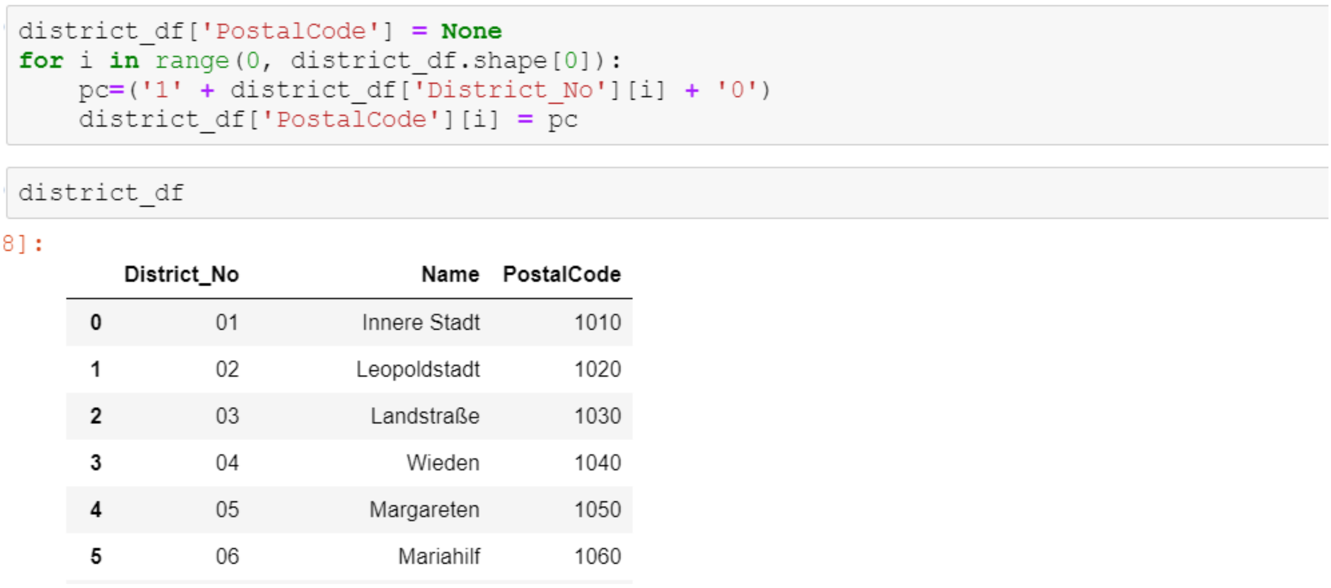


```
district_list = []
for row in table.find_all('tr'):
    data=row.find_all('th')
    district_list.append([i.text.strip() for i in data])

column_names = ['District_No', 'Name']
district_df = pd.DataFrame(columns= column_names,data=district_list[1:], index=None)

district_df = district_df[:23]
district_df.head()
```

5]:

	District_No	Name
0	01	Innere Stadt
1	02	Leopoldstadt
2	03	Landstraße
3	04	Wieden
4	05	Margareten

Part of the resulting data frame is displayed above. In addition to that, for determining location coordinates, it is always useful to have the corresponding postal codes. This information can be again extracted from and online source, however, the postal codes in Vienna are easily formed by using the district number and adding a '1' in the beginning and '0' in the end. That's what we're doing in the next step, and simultaneously adding this to our data frame.



```
district_df['PostalCode'] = None
for i in range(0, district_df.shape[0]):
    pc=('1' + district_df['District_No'][i] + '0')
    district_df['PostalCode'][i] = pc
```

```
district_df
```

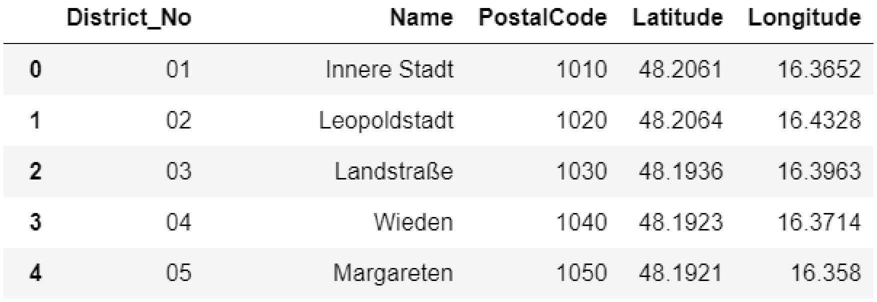
8]:

	District_No	Name	PostalCode
0	01	Innere Stadt	1010
1	02	Leopoldstadt	1020
2	03	Landstraße	1030
3	04	Wieden	1040
4	05	Margareten	1050
5	06	Mariahilf	1060

Again, we see a part of the resulting data frame, with the postal code added to it.

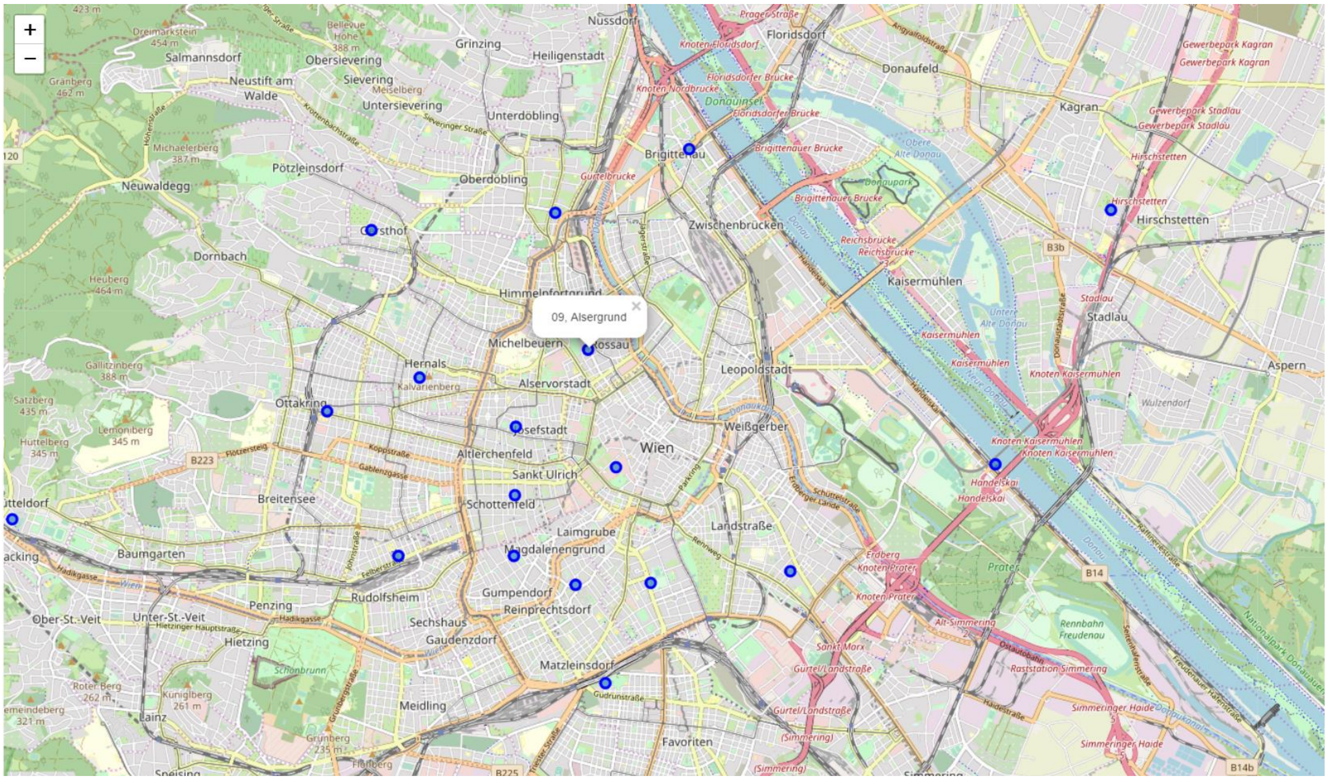
3.1.2. Districts’ location data

Having obtained the postal codes for each district, those are used to extract the latitude and longitude coordinates of each district using the GeoPy's Nominatum service.



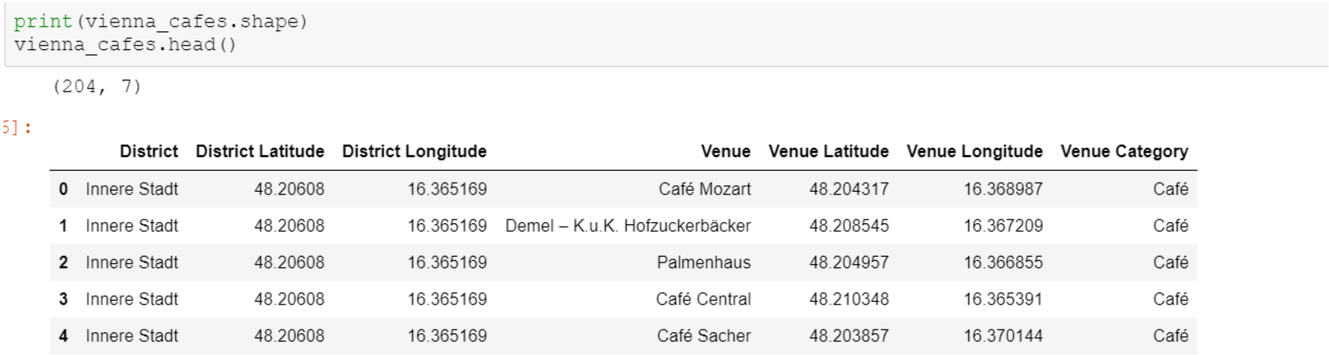
	District_No	Name	PostalCode	Latitude	Longitude
0	01	Innere Stadt	1010	48.2061	16.3652
1	02	Leopoldstadt	1020	48.2064	16.4328
2	03	Landstraße	1030	48.1936	16.3963
3	04	Wieden	1040	48.1923	16.3714
4	05	Margareten	1050	48.1921	16.358

As a result, the dataset thus contains the latitudes and longitudes of each of Vienna’s 23 districts. To get a better overview of the spatial location of each district, the information is visualized using the Folium library:



3.1.3. Venue data

Next step is to identify all venues in the 23 districts using the Foursquare API. Since we are only interested in finding cafés, the search is limited by a search query respectively. The data is obtained and recorded in a JSON object. From that we extract only the relevant data for each location: its name, category, district and location (latitude and longitude). As a result, we form a new dataset that looks as follows:

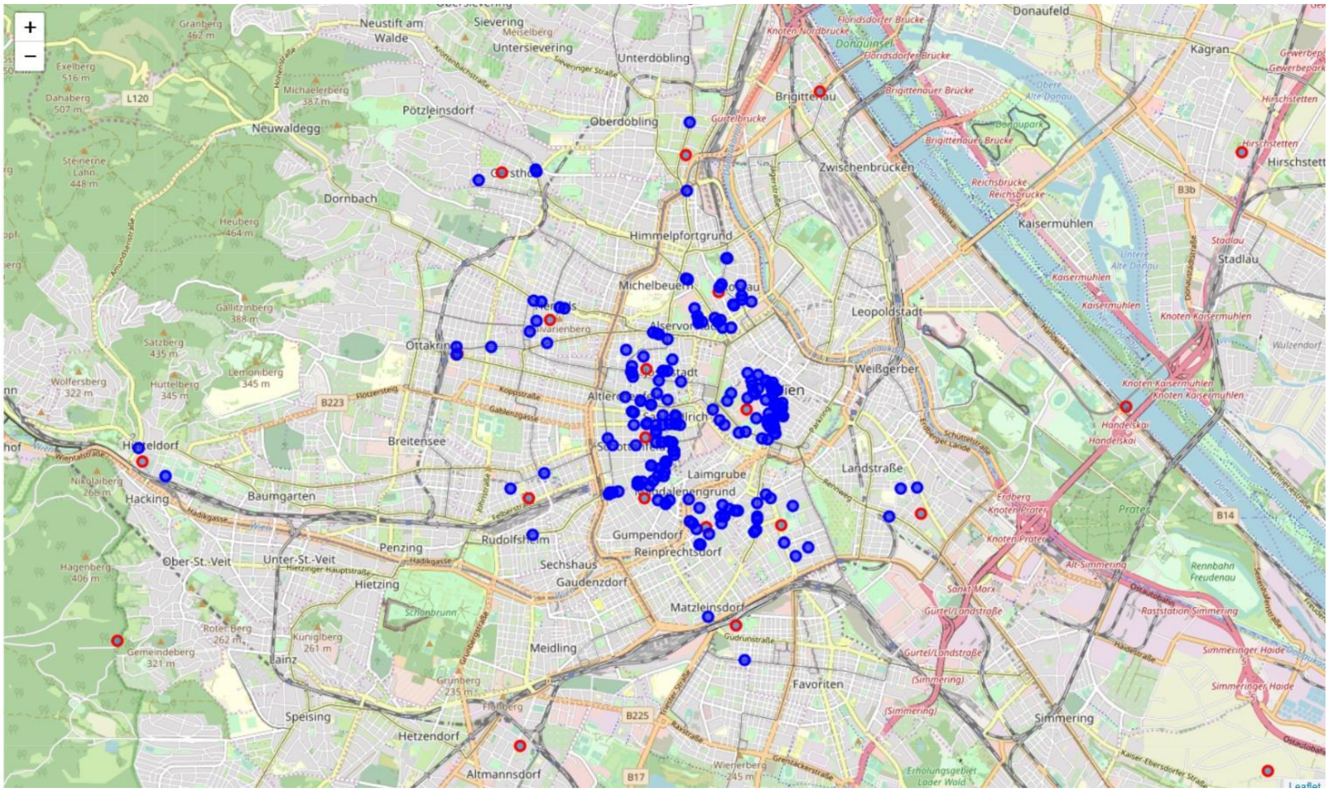

```
print(vienna_cafes.shape)
vienna_cafes.head()
```

```
(204, 7)
```

5]:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Innere Stadt	48.20608	16.365169	Café Mozart	48.204317	16.368987	Café
1	Innere Stadt	48.20608	16.365169	Demel – K.u.K. Hofzuckerbäcker	48.208545	16.367209	Café
2	Innere Stadt	48.20608	16.365169	Palmenhaus	48.204957	16.366855	Café
3	Innere Stadt	48.20608	16.365169	Café Central	48.210348	16.365391	Café
4	Innere Stadt	48.20608	16.365169	Café Sacher	48.203857	16.370144	Café

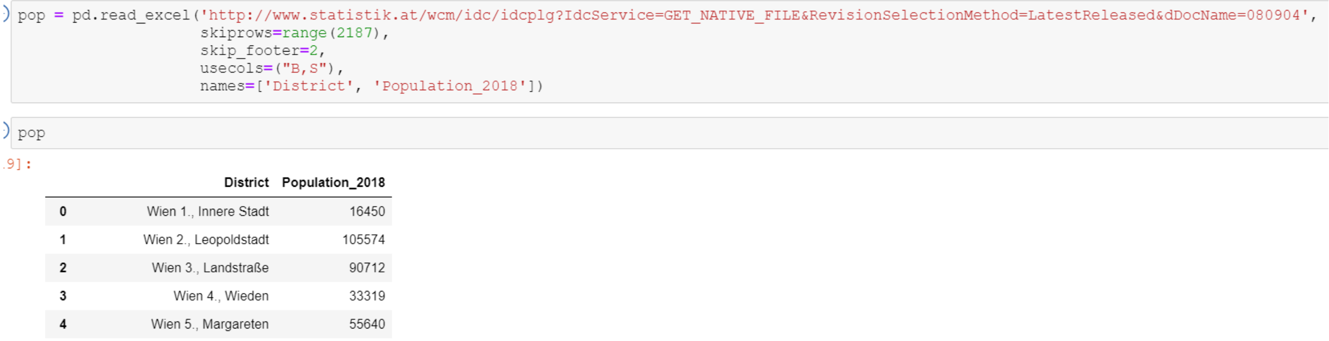
Again, we plot the venues on a map using the Folium library to get an overview of their spatial dissemination:



This time, the districts are marked by a red circle, while each venue is marked by a blue one. Already here we can observe concentration of venues in certain areas of the city.

3.1.4. Population

To complete the data collection, we obtain population data for the city of Vienna from Statistics Austria. We use data for the year 2018:

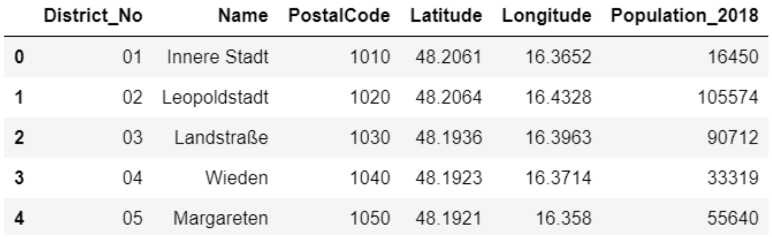


```
pop = pd.read_excel('http://www.statistik.at/wcm/idc/idcplg?IdcService=GET_NATIVE_FILE&RevisionSelectionMethod=LatestReleased&dDocName=080904',
                    skiprows=range(2187),
                    skip_footer=2,
                    usecols=("B,S"),
                    names=['District', 'Population_2018'])
```

```
pop
```

	District	Population_2018
0	Wien 1., Innere Stadt	16450
1	Wien 2., Leopoldstadt	105574
2	Wien 3., Landstraße	90712
3	Wien 4., Wieden	33319
4	Wien 5., Margareten	55640

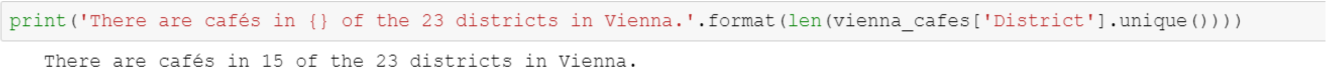
Next, we add this to the dataset we already have, in order to complete it. This is the final dataset that we be used for further analysis:



	District_No	Name	PostalCode	Latitude	Longitude	Population_2018
0	01	Innere Stadt	1010	48.2061	16.3652	16450
1	02	Leopoldstadt	1020	48.2064	16.4328	105574
2	03	Landstraße	1030	48.1936	16.3963	90712
3	04	Wieden	1040	48.1923	16.3714	33319
4	05	Margareten	1050	48.1921	16.358	55640

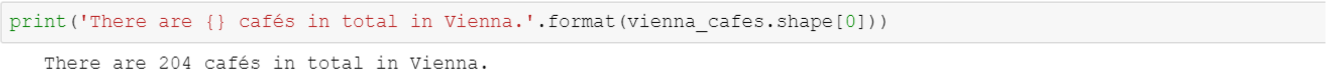
3.2. Exploratory data analysis

Next step is to get more familiar with the data we have obtained and processed so far. Going back to the venue data set we se that there are 204 venues (cafés) in Vienna’s 23 districts. However, those are located only in 15 of the districts, while in 8 there are no cafés at all:



```
print('There are cafés in {} of the 23 districts in Vienna.'.format(len(vienna_cafes['District'].unique())))
```

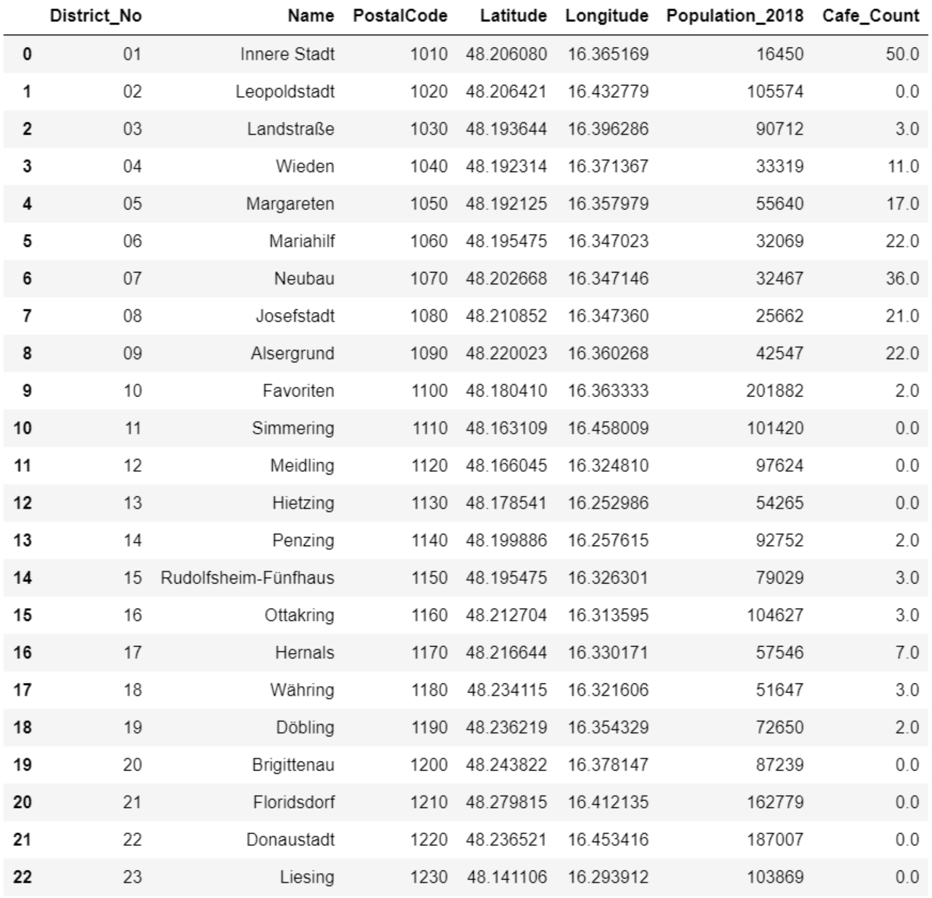
```
    There are cafés in 15 of the 23 districts in Vienna.
```



```
print('There are {} cafés in total in Vienna.'.format(vienna_cafes.shape[0]))
```

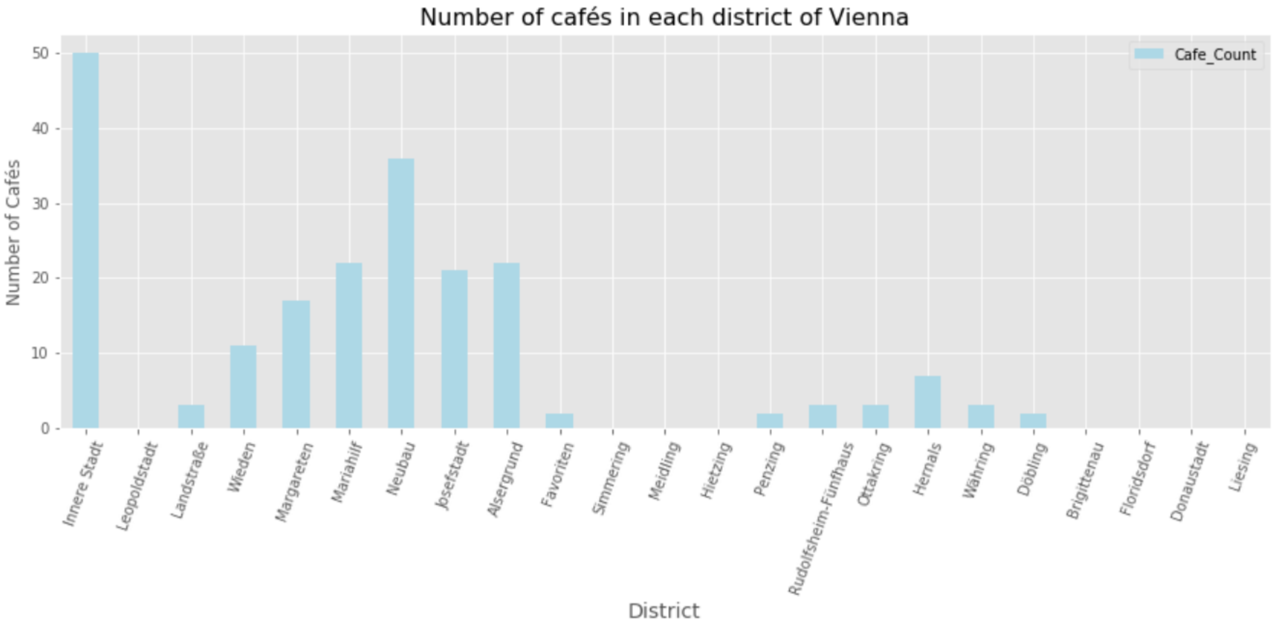
```
    There are 204 cafés in total in Vienna.
```

Therefore, we look to see in which districts the venues are, and how many per district exist. We then add this information to our district dataset from before. When we merge the two data frames, we use the 'left' merge option to not leave out the districts with no venues. We fill out the blanks with '0's:



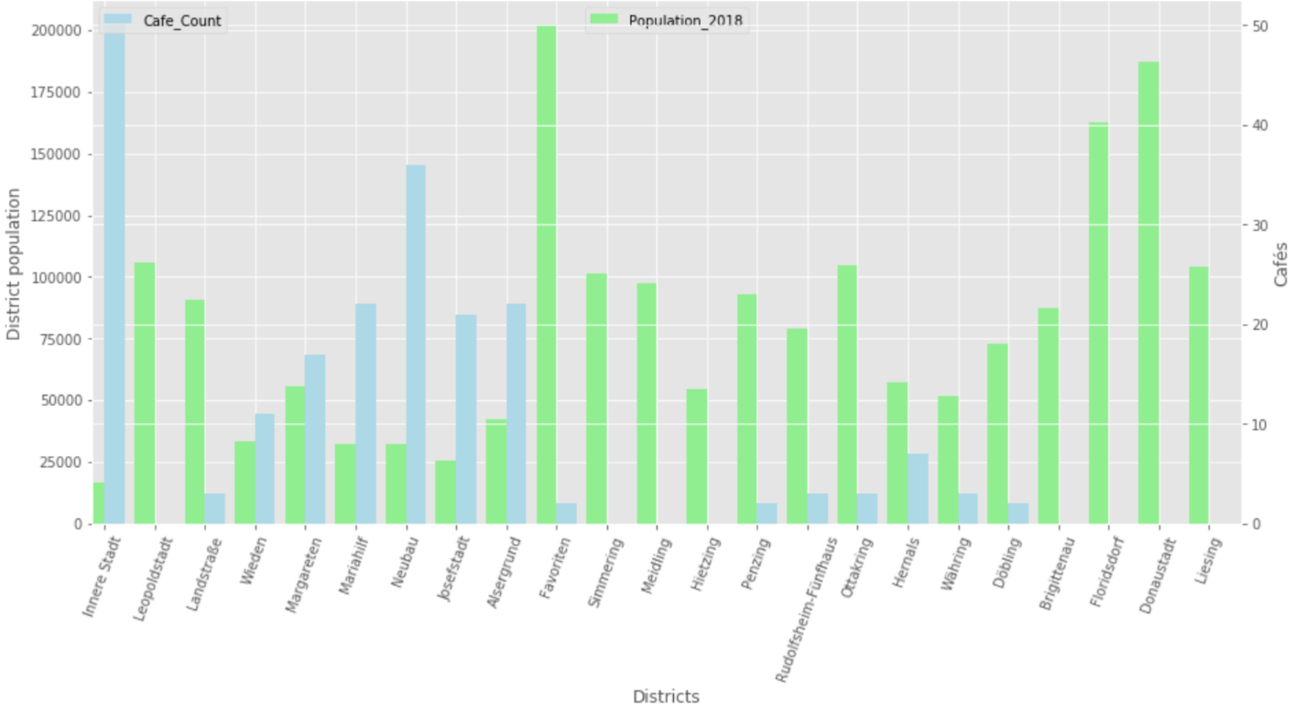
	District_No	Name	PostalCode	Latitude	Longitude	Population_2018	Cafe_Count
0	01	Innere Stadt	1010	48.206080	16.365169	16450	50.0
1	02	Leopoldstadt	1020	48.206421	16.432779	105574	0.0
2	03	Landstraße	1030	48.193644	16.396286	90712	3.0
3	04	Wieden	1040	48.192314	16.371367	33319	11.0
4	05	Margareten	1050	48.192125	16.357979	55640	17.0
5	06	Mariahilf	1060	48.195475	16.347023	32069	22.0
6	07	Neubau	1070	48.202668	16.347146	32467	36.0
7	08	Josefstadt	1080	48.210852	16.347360	25662	21.0
8	09	Alsergrund	1090	48.220023	16.360268	42547	22.0
9	10	Favoriten	1100	48.180410	16.363333	201882	2.0
10	11	Simmering	1110	48.163109	16.458009	101420	0.0
11	12	Meidling	1120	48.166045	16.324810	97624	0.0
12	13	Hietzing	1130	48.178541	16.252986	54265	0.0
13	14	Penzing	1140	48.199886	16.257615	92752	2.0
14	15	Rudolfsheim-Fünfhaus	1150	48.195475	16.326301	79029	3.0
15	16	Ottakring	1160	48.212704	16.313595	104627	3.0
16	17	Hernals	1170	48.216644	16.330171	57546	7.0
17	18	Währing	1180	48.234115	16.321606	51647	3.0
18	19	Döbling	1190	48.236219	16.354329	72650	2.0
19	20	Brigittenau	1200	48.243822	16.378147	87239	0.0
20	21	Floridsdorf	1210	48.279815	16.412135	162779	0.0
21	22	Donaustadt	1220	48.236521	16.453416	187007	0.0
22	23	Liesing	1230	48.141106	16.293912	103869	0.0

To get a better idea of the distribution, we visualize the information using a matplotlib's bar chart:



We see that there is an uneven distribution of café locations between all districts. The inner city has as many as 50 venues, while, as we have already observed, there are 8 districts that have no cafés. There are six other districts with more than 10 venues, Neubau being on 2nd place with 36.

However, since we don't know that much about the districts, next step is to plot together the districts' population and the number of coffee places and to make a comparative analysis.

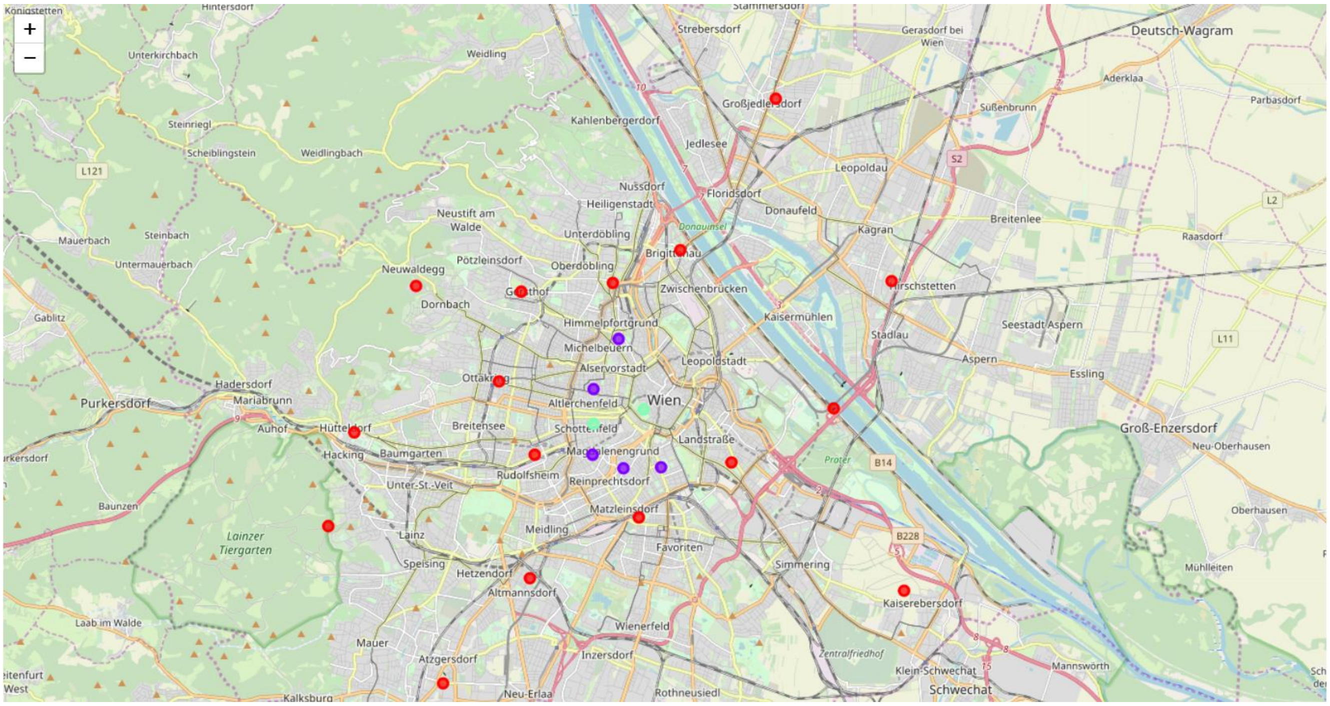


Although there are two y-axis showing us the information for both the population and the café count, we can see that in some neighborhoods the proportion of coffee places is much bigger compared to the population inhabiting the district. At the same time, there are several districts with more than 175 000 inhabitants, but with no cafés in them. This hints important implications for our conclusions later.

3.3. Clustering

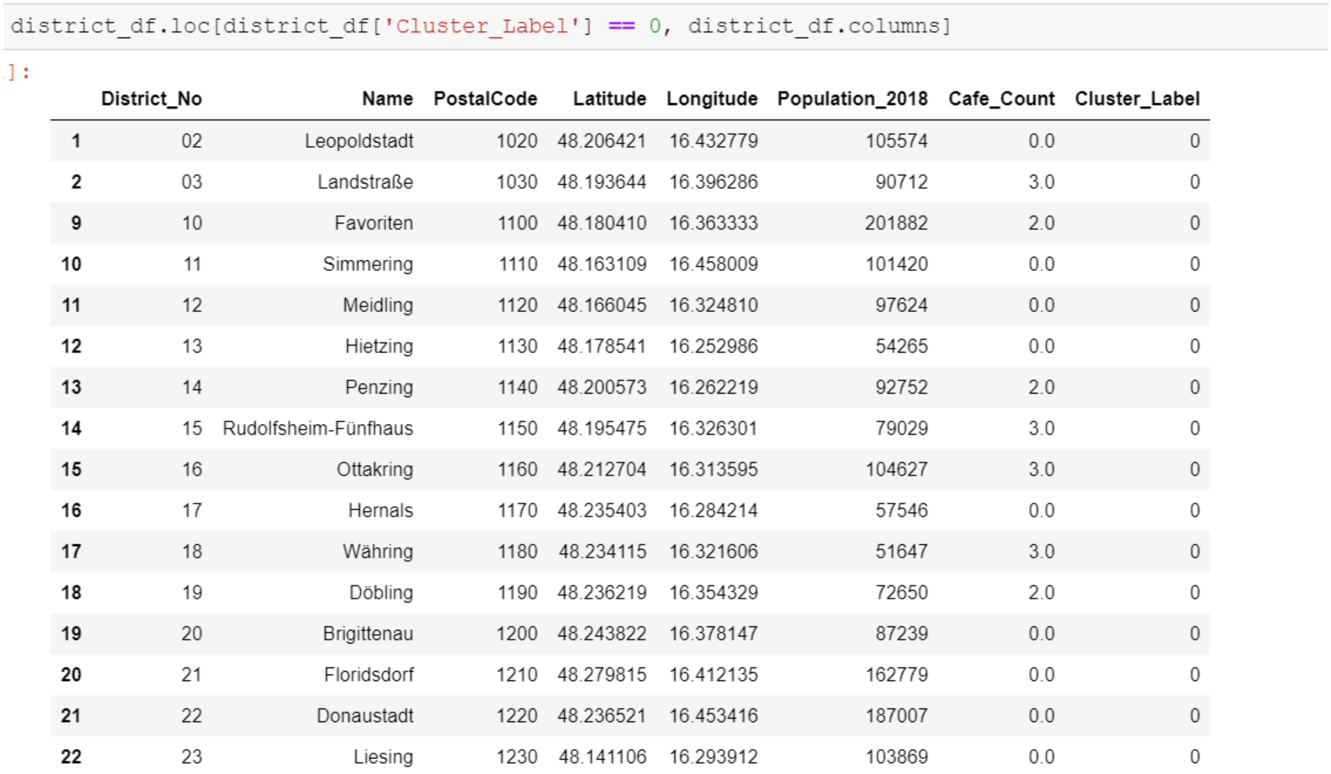
Next, using the machine learning KMeans clustering technique, the districts will be segmented into three clusters based on similar characteristics, and will give us more insights into the density of café locations in each of them, as well as their attractiveness for opening a new café. After setting up the KMeans model, it is fitted with the district dataset and the generated labels are added to our data frame in the same step. Note: since we're only looking at one type of venues, doing the one-hot encoding and determining the frequency of types of locations is unnecessary.

Having done that, we will visualize the created clusters in our model on the map:



We clearly notice the different clusters. To gather more insights from the segmentation the model has done, we will observe each cluster with the district it has been assigned with.

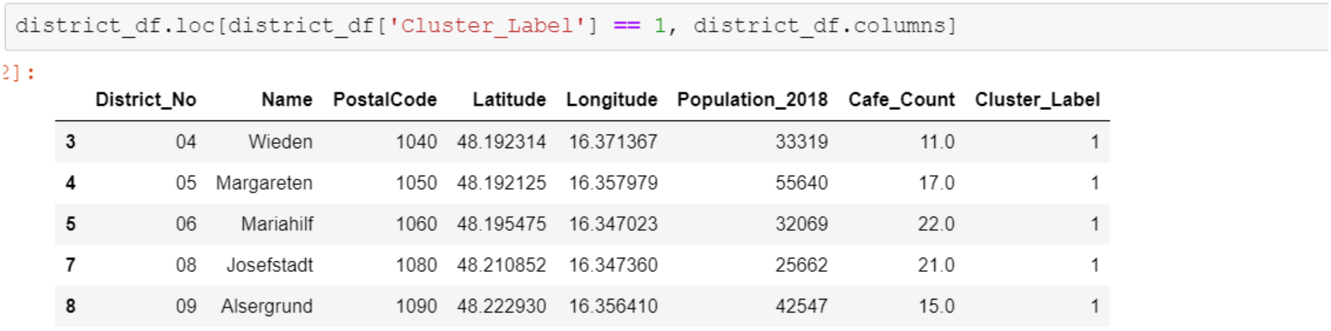
Cluster 1:



```
district_df.loc[district_df['Cluster_Label'] == 0, district_df.columns]
```

	District_No	Name	PostalCode	Latitude	Longitude	Population_2018	Cafe_Count	Cluster_Label
1	02	Leopoldstadt	1020	48.206421	16.432779	105574	0.0	0
2	03	Landstraße	1030	48.193644	16.396286	90712	3.0	0
9	10	Favoriten	1100	48.180410	16.363333	201882	2.0	0
10	11	Simmering	1110	48.163109	16.458009	101420	0.0	0
11	12	Meidling	1120	48.166045	16.324810	97624	0.0	0
12	13	Hietzing	1130	48.178541	16.252986	54265	0.0	0
13	14	Penzing	1140	48.200573	16.262219	92752	2.0	0
14	15	Rudolfsheim-Fünfhaus	1150	48.195475	16.326301	79029	3.0	0
15	16	Ottakring	1160	48.212704	16.313595	104627	3.0	0
16	17	Hernals	1170	48.235403	16.284214	57546	0.0	0
17	18	Währing	1180	48.234115	16.321606	51647	3.0	0
18	19	Döbling	1190	48.236219	16.354329	72650	2.0	0
19	20	Brigittenau	1200	48.243822	16.378147	87239	0.0	0
20	21	Floridsdorf	1210	48.279815	16.412135	162779	0.0	0
21	22	Donaustadt	1220	48.236521	16.453416	187007	0.0	0
22	23	Liesing	1230	48.141106	16.293912	103869	0.0	0

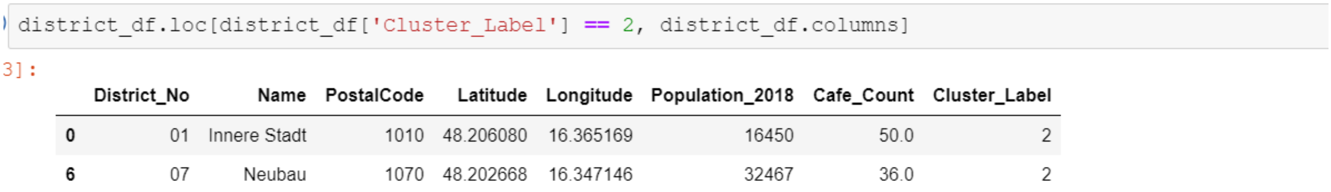
Cluster 2:



```
district_df.loc[district_df['Cluster_Label'] == 1, district_df.columns]
```

	District_No	Name	PostalCode	Latitude	Longitude	Population_2018	Cafe_Count	Cluster_Label
3	04	Wieden	1040	48.192314	16.371367	33319	11.0	1
4	05	Margareten	1050	48.192125	16.357979	55640	17.0	1
5	06	Mariahilf	1060	48.195475	16.347023	32069	22.0	1
7	08	Josefstadt	1080	48.210852	16.347360	25662	21.0	1
8	09	Alsergrund	1090	48.222930	16.356410	42547	15.0	1

Cluster 3:



```
district_df.loc[district_df['Cluster_Label'] == 2, district_df.columns]
```

	District_No	Name	PostalCode	Latitude	Longitude	Population_2018	Cafe_Count	Cluster_Label
0	01	Innere Stadt	1010	48.206080	16.365169	16450	50.0	2
6	07	Neubau	1070	48.202668	16.347146	32467	36.0	2

4. Results

We can see that the model has segmented the districts and by looking at the density of coffee places in each, we observe that the districts in the Cluster 3 have the biggest count on cafés, the ones in Cluster 2 have between 11 and 21 locations, while the districts in Cluster 1 have a maximum of 3 cafés.

In addition, we have before observed the distribution of population among all districts and the results indicate that the 5 districts with the most population have either no, or a maximum of two coffee places. This speaks for a certain disproportionality between people and café locations. All of them were place by the model in Cluster 1.

This allows for the conclusion, that namely the districts in Cluster 1 offer the most attractive location for opening a new café. Although, by looking at the map, one notices that they are further away from the city center, the lack of cafés and the big number of inhabitants speak for the potentially unmet demand, that early entrants can make a big use of.

5. Discussion

Although we manage to get good insights in the scope of this study, it is accompanied by several limitations.

First, we should consider that the Foursquare database does not hold information regarding every coffee place that exists in the city of Vienna. Furthermore, there might be cafés that were not categorized in the 'Café' venue category that we have applied here and thus were not included in our venue map. However, we assume that the venues that are characterized with what is considered necessary to meet the needs mentioned in the beginning of this report (experiencing the renowned Viennese coffee culture, including surrounding, atmosphere etc. and not simply being able to buy a cup of coffee) are only fully present in this venue category.

Moreover, there are different factors except the population in a district and the number of cafés already available that play essential role in deciding about opening a new venue, i.e. is the district inhabited by families, students, tourists etc. For example, there is a university in the second district of Vienna: 'Leopoldstadt' with more than 20 000 students, who would raise demand for coffee places in this district, however there are none venues on Foursquare to be found. Such factors could be considered within a more comprehensive analysis on the topic.

6. Conclusion

Analyzing the districts of Vienna based on the number of café venues available in each of them according to the Foursquare database, gave us some important insights regarding their attractiveness for opening a new venue in districts with potential unmet demand of such. Clustering the districts based on their similarities in terms of density of the venues and their population, although quite uncomplicated, hints that all districts in Cluster 1 are potentially attractive for either established café owners to expand their businesses from other districts there, or for new entrants to start a business in this sector to make use of the highly likely unmet demand in such districts with big population, but no local coffee places.

To get an even better understanding of the level of that attractiveness, a further analysis including more (also) subjective factors is recommended, a different algorithm could be applied, as well as a different venue database could be included to ensure capturing more of the current status quo of the Vienna café map.