Vienna - Where to settle for entertainment and recreational purposes as a student or young urban professional

Coursera: Applied Data Science Capstone project
Author: Jörn Grimmer
Date: December 2019/ January 2020

1. Introduction

1. Description & Dicsussion of the Background

According to a recent study, Vienna is the most attractive city worth living for the second year in a row. (Source: Economist Intelligence Unit, Global Livability Ranking 2019). This study ranks 140 cities for their urban quality of life based on assessments of stability, healthcare, culture and environment, education and infrastructure. No surprise, Vienna is attracting a lot of people starting their studies or working life.

2. Problem

In order to decide where to settle down, an individual needs to make a choice among a lot of criteria.
In our project we will try to find an optimal location so settle in Vienna, when your major interest is in recreational & sport facilities, restaurants and night life entertainment close to the university faculty of your choice leaving other factors out of sight.

The analysis is expected to come up with district clusters subject to recreational focus.

3. Target audience

This report is targeted at students & young urban professionals planning to settle in Vienna. We assume this target group is interested in general to settle close to spots of their daily live to minimize distances between every day lcoations.

2. Data aquisition, preparation & cleansing

1. Data sources

In order to perform this analysis we require the following data sources.

- Geospatial data on Vienna, its districts and its universities. Austria publishes this type of data on www.data.gv.at.

Specifically I used datasets on Vienna, its districts, its university and colleges and their respective faculties:

- https://www.data.gv.at/katalog/dataset/stadt-wien_bezirksgrenzenwien/ (BEZIRKSGRENZEOGD.csv)
- https://www.data.gv.at/katalog/dataset/stadtwien_universittenundfachhochschulenstandortewien/ (UNIVERSITAETOGD.csv)

- Venue data for specified districts of Vienna.

Here, we used the **Foursquare data** as required by the Capstone course. On data.gv.at there are distinct sources on recreational spots like sports and swimming facilities. This data could be used to qualify the completeness of Foursquare data and or re-focus such an analysis. However, for scoping and timing reasons of this project, I made use of the Foursquare data only.

2. Data preparation

All relevant data on Vienna, its districts, its universities and colleges is provided in different formats, among them csv. I made use of the csv-file download, read the files with PANDAS and created according dataframes.

The original data set of Vienna's districts had 23 rows and 18 columns.

In order to be able to perform the analysis I prepared Vienna's district data by

- Dropping columns not required
- Sorting districts in ascending order according to the number of the district
- Extracting a latitude/ longitude datapoint from the polygone shape values to have a geospatial value to set a flag on a map for each district. Note: we extract the first datapoint of the Polygon. We do not caclulate the mid-point of the Polygon.

After the preparation, the data set had 23 rows and 8 columns for further analysis.

	index	Borough	BEZNR	DISTRICT_CODE	STATAUSTRIA_BEZ_CODE	STATAUSTRIA_GEM_CODE	Longitude	Latitude
0	3	Innere Stadt	1	1010	901	90101	16.372641	48.216617
1	6	Leopoldstadt	2	1020	902	90201	16.403453	48.231919
2	1	Landstraße	3	1030	903	90301	16.396617	48.207387
3	19	Wieden	4	1040	904	90401	16.369165	48.200713
4	18	Margareten	5	1050	905	90501	16.359449	48.196617
5	20	Mariahilf	6	1060	906	90601	16.363064	48.201827
6	0	Neubau	7	1070	907	90701	16.338725	48.208537
7	2	Josefstadt	8	1080	908	90801	16.349147	48.215158
8	5	Alsergrund	9	1090	909	90901	16.361652	48.231918
9	16	Favoriten	10	1100	910	91001	16.383819	48.185157
10	15	Simmering	11	1110	911	91101	16.425986	48.185575
11	17	Meidling	12	1120	912	91201	16.341743	48.188466
12	21	Hietzing	13	1130	913	91301	16.214234	48.206523
13	10	Penzing	14	1140	914	91401	16.209138	48.264112
14	22	Rudolfsheim-Fünfhaus	15	1150	915	91501	16.327324	48.205005
15	4	Ottakring	16	1160	916	91601	16.276206	48.227037
16	8	Hernals	17	1170	917	91701	16.285159	48.256800
17	7	Währing	18	1180	918	91801	16.295017	48.249609
18	11	Döbling	19	1190	919	91901	16.356813	48.282287
19	9	Brigittenau	20	1200	920	92001	16.373612	48.261269
20	13	Floridsdorf	21	1210	921	92101	16.437762	48.316811
21	12	Donaustadt	22	1220	922	92201	16.507839	48.273446
22	14	Liesing	23	1230	923	92301	16.280553	48.159055

Table 1: Districts of Vienna

A similar preparation was required for the data on Vienna's universities and colleges. Here, the original data set had 158 rows and 6 columns.

In this case we prepared the dataset by

- Dropping columns not required and
- Extracting the latitude/ longitude datapoints

After the preparation, the data set had 158 rows and 3 columns.

	NAME	Longitude	Latitude
0	FH Technikum Wien	16.377856	48.239443
1	FH Technikum Wien	16.426908	48.269503
2	FH Campus Wien	16.382288	48.157733
3	Fachhochschule des bfi Wien	16.403446	48.219132
4	Fachhochschule des bfi Wien	16.426908	48.269503
5	FHWien-Studiengänge der Wirtschaftskammer Wien	16.349201	48.226579
6	Lauder Business School	16.352469	48.242701
7	FH Technikum Wien	16.355891	48.200143
8	Technische Universität Wien	16.363088	48.200171
9	Akademie der bildenden Künste Wien	16.361984	48.199806
10	Universität Wien	16.348993	48.233515

Table 2: Universities and Colleges of Vienna (First 10 entries only)

The venue data was received by using the Foursquare API.

The requested data is provided in a json format. Thus it needed to be cleansed and structured into a PANDAS dataframe.

Below is an example of a venue data pandas dataframe used for further analysis.

Out[40]:

	name	categories	lat	lng
0	ZWE	Jazz Club	48.216341	16.374444
1	Palais Hansen Kempinski Vienna	Hotel	48.216335	16.368463
2	Adria	Beach Bar	48.214945	16.375037
3	Tel Aviv Beach	Beach Bar	48.217081	16.373421
4	Feuerdorf	BBQ Joint	48.215972	16.373495

Table 3: Example of Foursquare Venue Data

3. Data visualization

We used the application Folium to visualize the geospatial data.

First, we draw a map of Vienna to demonstrate the position of the districts in the city of Vienna.

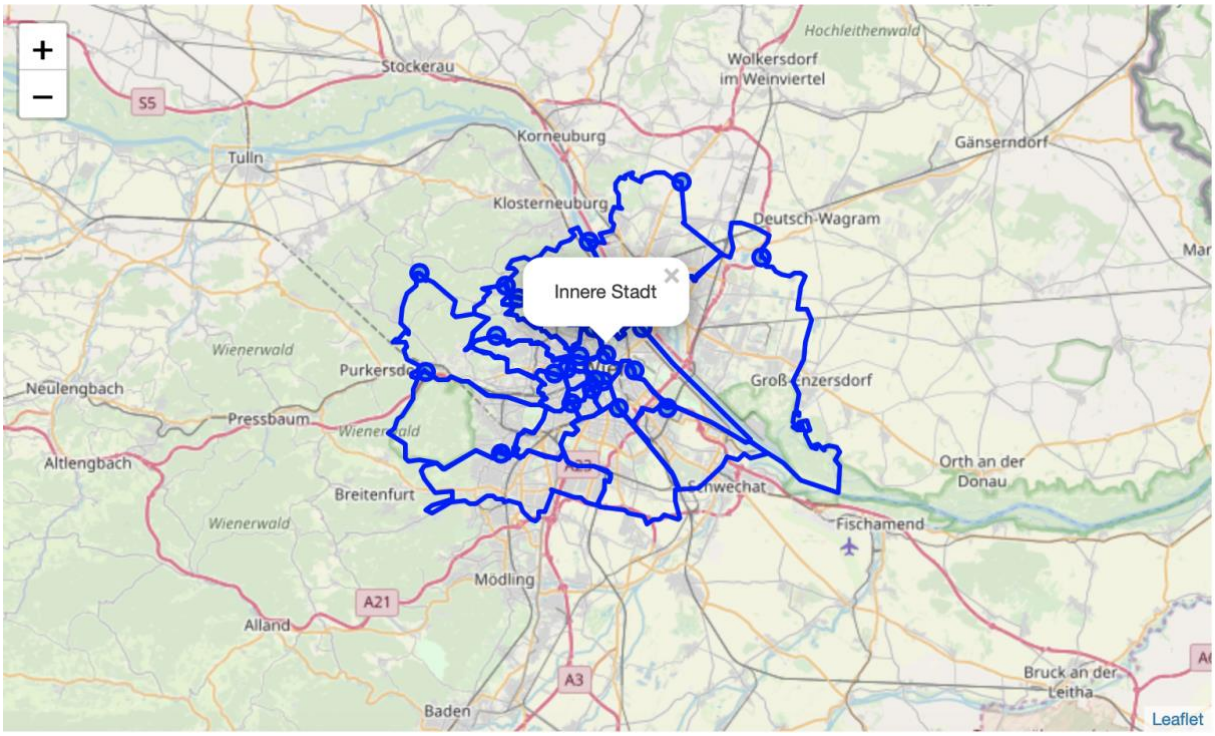

Figure 1: City of Vienna and districts

Second, we identified the distribution of universities and colleges across Vienna's districts.

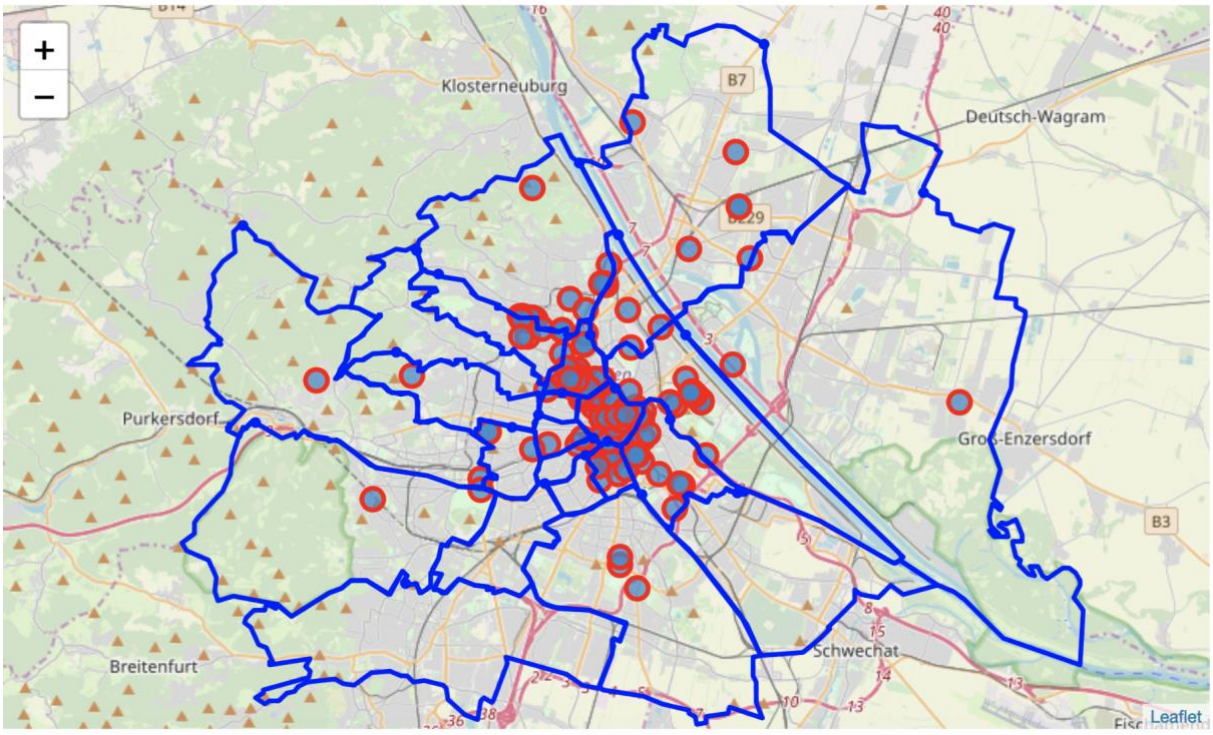

Figure 2: City of Vienna and distribution of universities, faculties and colleges

Universities, faculties and colleges are spread around Vienna city where the majority can be found in the inner districts.

In our study, we will make use of one faculty, namely the ‚Akademie der bildenden Künste. The academy is located on Longitude 16.361984, Latitude 48.199806 which is very close to the first district as shown on the map.

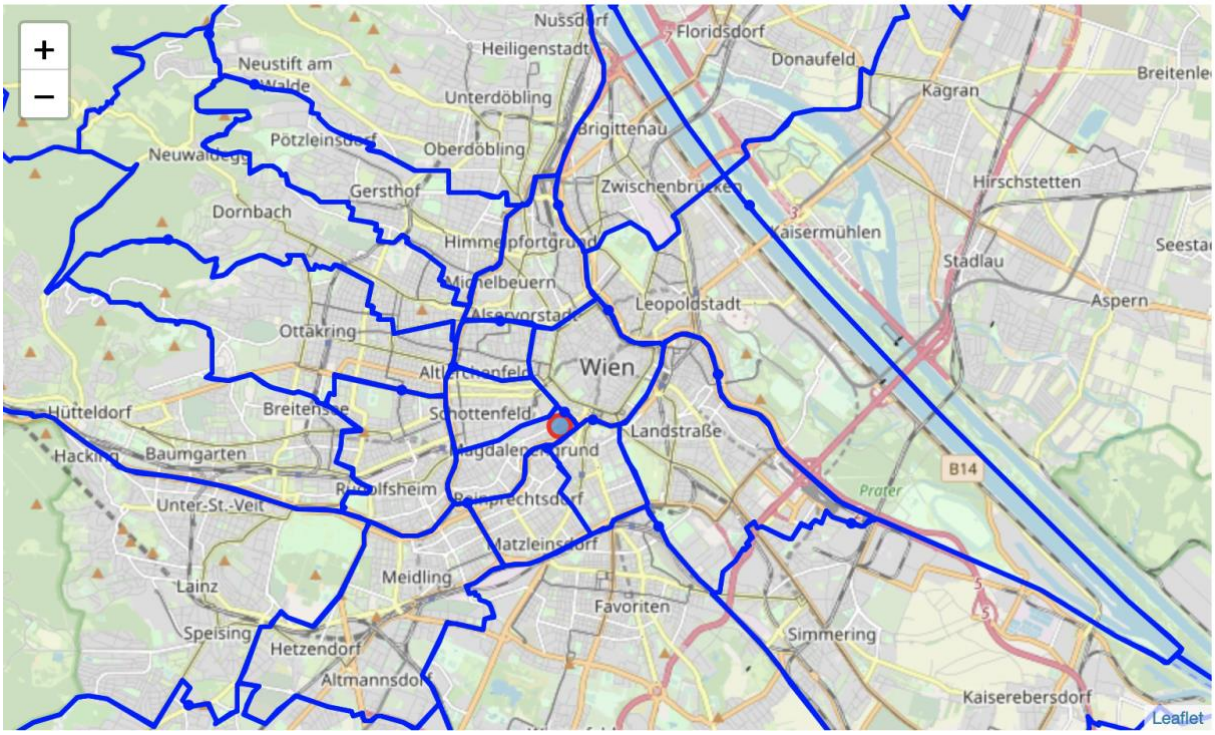

Figure 3: Location of „Akademie der bildenden Künste

3. Methodology

The objective of our study is to identify clusters of recreational focus in Vienna‘s districts close to a faculty of choice.

Thus we need a method of unsupervised learning, specifically a method to cluster the venues within the respective districts according to a pattern. We will use the k-Means algorithm, which is extremely easy to implement and is also computationally very efficient compared to other clustering algorithms. The k-Means algorithm belongs to the category of prototype-based clustering.

Prototype-based clustering means that each cluster is represented by a prototype, which can either be the centroid (average) of similar points with continuous features, or the medoid (the most representative or most frequently occurring point) in the case of categorical features.
We have to specify the number of clusters a priori. In order to verify the optimal number of clusters, we will use the so-called elbow method. We want to verify the value of k where the distortion begins to decrease most rapidly. Intuitively, we can say that, if k increases, the within-cluster SSE (“distortion”) will decrease. This is because the samples will be closer to the medoids they are assigned to.

The analysis is expected to come up with district clusters subject to its recreational focus.
In Vienna, the districts with single digits („BEZNR“) are centered in the middle like a snail shell, while the districts with two digits („BEZNR“) are distributed on the outer shell.

Since we are interested in the districts close to the spots of the daily life of our target group, we selected the districts nearest to the faculty „Akademie der bildenden Künste“. Due to our knowledge we selected the single digit districts for our further analysis.

The below table shows the final selection of districts:

	index	Borough	BEZNR	DISTRICT_CODE	STATAUSTRIA_BEZ_CODE	STATAUSTRIA_GEM_CODE	Longitude	Latitude
0	3	Innere Stadt	1	1010	901	90101	16.372641	48.216617
1	6	Leopoldstadt	2	1020	902	90201	16.403453	48.231919
2	1	Landstraße	3	1030	903	90301	16.396617	48.207387
3	19	Wieden	4	1040	904	90401	16.369165	48.200713
4	18	Margareten	5	1050	905	90501	16.359449	48.196617
5	20	Mariahilf	6	1060	906	90601	16.363064	48.201827
6	0	Neubau	7	1070	907	90701	16.338725	48.208537
7	2	Josefstadt	8	1080	908	90801	16.349147	48.215158
8	5	Alsergrund	9	1090	909	90901	16.361652	48.231918
9	16	Favoriten	10	1100	910	91001	16.383819	48.185157

Table 4: Vienna districts for our analysis

4. Results

In a first iteration, we run a K-Means clustering with 5 clusters.

Further we specify to run the K-means clustering algorithms 10 times independently with different random centroids to choose the final model as the one with the lowest SSE. Via the max_iter parameter, we specify the maximum number of iterations for each single run with 300.

We visualized the results on the map of Vienna with Folium and extracted the most common venues of the identified clusters.

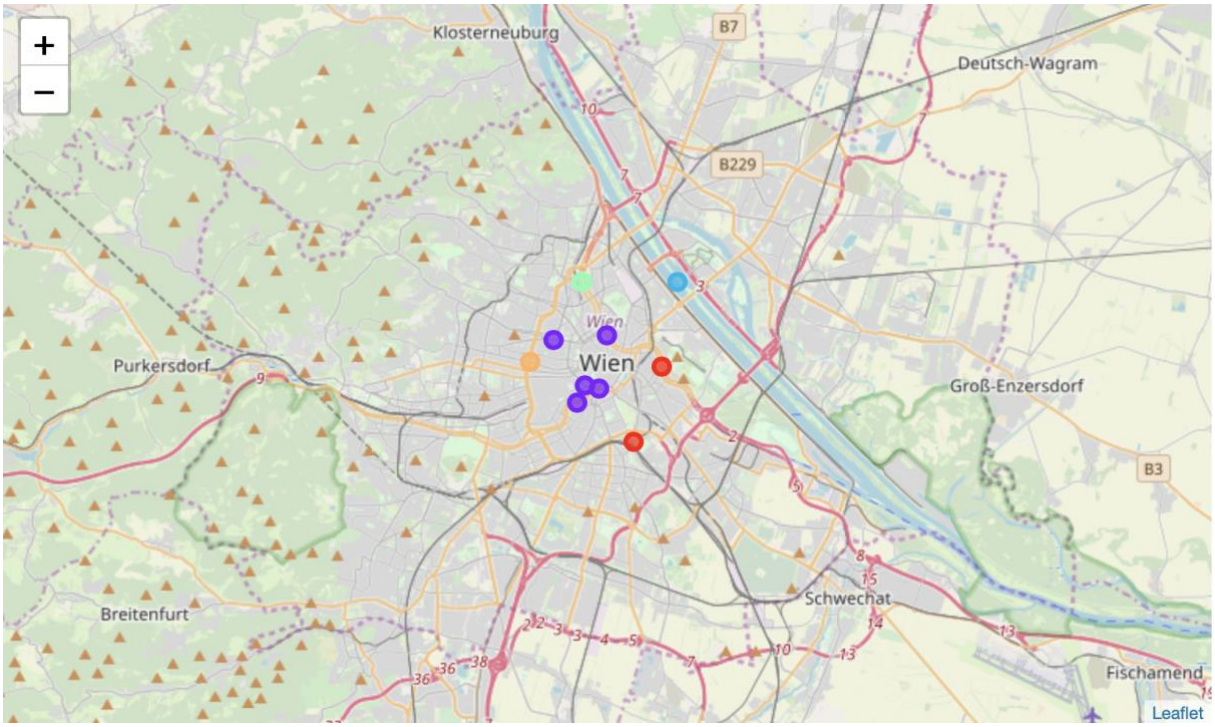

Figure 4: Five District Clusters with various focus of recration, Vienna

		Most Common Venue		
Borough	Cluster Label	1st	2nd	3rd
Landstraße	0	Art Museum	Bus Stop	Hotel
Favoriten	0	Hotel	Cafe	Restaurant
Innere Stadt	1	Restaurant	Cafe	Hotel
Wieden	1	Hotel	Cafe	Plaza
Margareten	1	Cafe	Bar	Asian Restaurant
Mariahilf	1	Cafe	Museum	Plaza
Josefstadt	1	Hotel	Cafe	Restaurant
Leopoldstadt	2	Beach Bar	Latin American Restaurant	Mediterranean Restaurant
Alsergrund	3	Hotel	Austrian Restaurant	Nightclub
Neubau	4	Bakery	Turkish Restaurant	Fast Food Restaurant

Table 5: Five District Clusters and Most Common Venues

From a recreational point of view, we provide the following interpretation to the cluster results:

1) Cluster 0: Museum & Hotel
2) Cluster 1: Cafe & Restaurant
3) Cluster 2: Maritim & Mediterrenean
4) Cluster 3: Domestic Food & Nightclub
5) Cluster 4: Bakery & Fast Food

The algorithm has identified five clusters, however, the clusters display some overlaps.

There are restaurants or at least specific restaurants among the top three venues in all clusters. Further in a couple of districts in clusters labelled ‚0' and ‚1', there is „Cafe" placed on the 2nd venue, for example this is true for the district„Favoriten" cluster ‚0' and the district ‚Innere Stadt', cluster ‚1'.

We want to know, if the number of clusters has been set to an optimum from the very beginning. Therefore we make use of the **elbow method** as a graphical tool to estimate the optimal number of clusters *k*. The idea behind the elbow method is to identify the value of k where the distortion begins to decrease most rapidly.

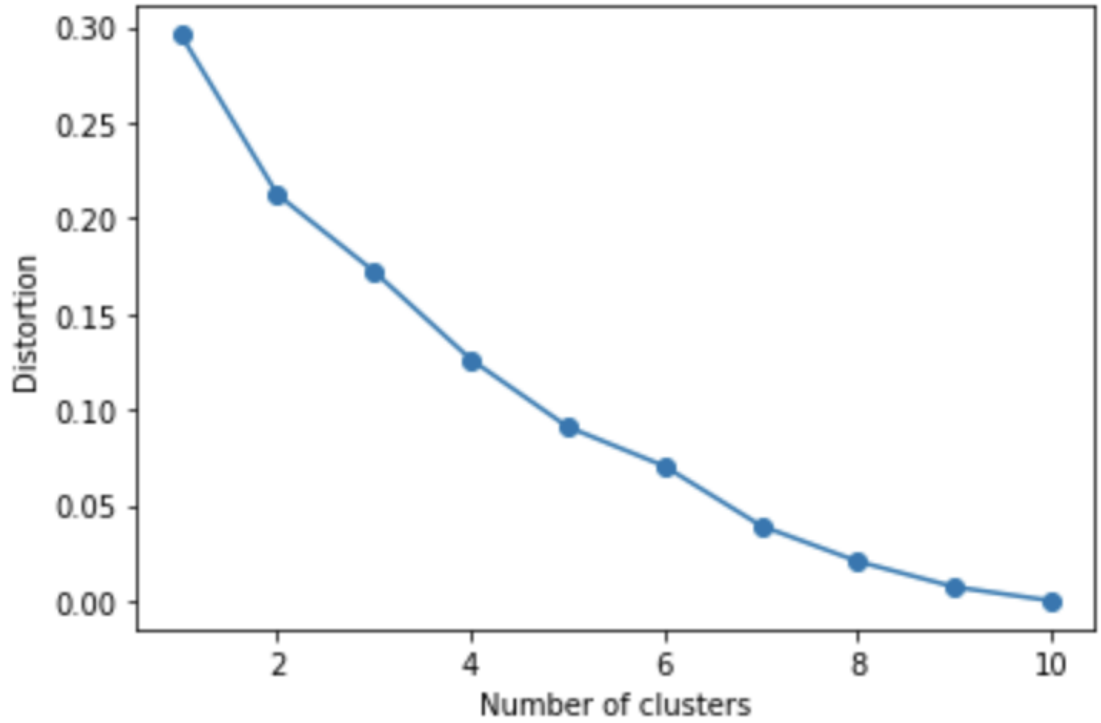

Figure 5: Elbow curve

As we can see in the resulting plot, there is no clear evidence on where the elbow is located. The curve is flattening first at 2 slighlty, then a second time at 5 before it is steepening at 6 again. We run ke-Means a second time with two clusters to see what it comes up with.

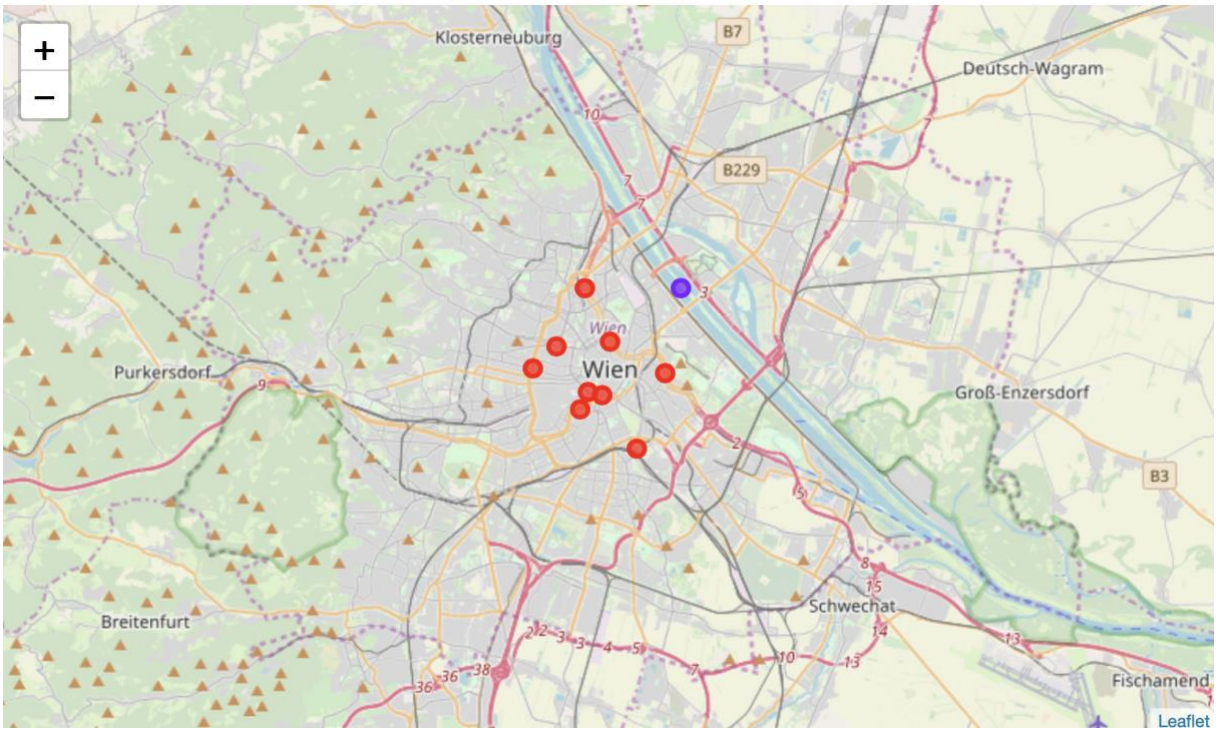

Figure 6: Two District Clusters with various focus of recration, Vienna

		Most Common Venue		
Borough	Cluster Label	1st	2nd	3rd
Alsergrund	0	Hotel	Austrian Restaurant	Nightclub
Favoriten	0	Hotel	Cafe	Restaurant
Innere Stadt	0	Restaurant	Cafe	Hotel
Josefstadt	0	Hotel	Cafe	Restaurant
Landstraße	0	Art Museum	Bus Stop	Hotel
Margareten	0	Cafe	Bar	Asian Restaurant
Mariahilf	0	Cafe	Museum	Plaza
Neubau	0	Bakery	Turkish Restaurant	Fast Food Restaurant
Wieden	0	Hotel	Cafe	Plaza
Leopoldstadt	1	Beach Bar	Latin American Restaurant	Mediterranean Restaurant

Table 6: Two District Clusters and Most Common Venues

From a recreational point of view, we provide the following interpretation to the updated cluster results:

1) Cluster 0: Restaurant/Cafe/Hotel
2) Cluster 1: Maritim/ Mediterrenean

Leopoldstadt remains as a separate cluster.

Combining the data results with the positioning on the city map, we can state the following:

Provided you are studying in a university/ faculty in the centre of Vienna and you have the preference for a place with recreational & sport facilities, restaurants and night life entertainment nearby the faculty of your choice, each district seems to povide a similar broad choice of venues, especially cafes and restaurants. In addition to this broad preference, you can find additional recreational focus in four of the five clusters, namely:

1) Cluster 0: Museum & Hotel
2) Cluster 1: Cafe & Restaurant
3) Cluster 2: Maritim & Mediterrenean
4) Cluster 3: Domestic Food & Nightclub
5) Cluster 4: Bakery & Fast Food

In order to provide evidence, we compare our results with a qualitative study on Vienna, which was published recently.[1] This reports confirms, that there is no shortage of restaurant, food and coffee options in none of this districts.

5. Limitations

The gesopatial shape of Vienna's districts is provided by a Polygon dataset of Longitude/ Latitude datapoints,. We extract the first Longitude/ Latitude datapoint instead of calculating the mid-point of the Polygon for simplification. Thus, the selection of venues in a radius of 500 metres around this datapoint might bias our results.

[1] https://austrianadaptation.com/where-to-live-vienna/ Title: Where to live in Vienna. Your epic guide to Vienna's Districts, In Living Abroad, Vienna, Vienna Local Experiences by CarlyNovember 2, 2017

Second, the Foursquare data limits the radius und number of venues. Again, this limitation might bias our clustering results.

While k-means is very good at identifying clusters with a spherical shape, we observe the distribution of our venue data seems not to provide such a shape within the given districts of Vienna.

As a consequence, our results might be biased and limited.

Nevertheless, our comparison with the external qualitative study confirms our results on a high level.

6. Conclusion

The project assignment was targeted to find an optimal location so settle in Vienna, when the major interest is in recreational & sport facilities, restaurants and night life entertainment close to the university faculty of your choice.

The popular kMeans clustering method delivered cluster results which were limited. The reason for this was found in the apparently even distribution of recreational venues across the districts rather than having a distinct sperical shaped distribution of specific recreational venues.

Nevertheless, the study can serve as a starting point for an additional analysis of other preferences, e.g. cultural diversity or distance to public transportation et.al., to find the most preferred spot in line with your preferences to settle near your university or faculty of choice in Vienna. This type of data is available on www.data.gv.at., too.