Which European Capital for a GetAway?

1. Business Problem

GetAway.com is an online business focused on providing high level weekend trips for busy professionals who want to escape their cities and workloads for the weekend. These trips need to be brief, usually starting Friday night or Saturday afternoon and ending by Sunday night, and directly aimed at the customer interests.

GetAway.com will provide transportation to and accommodation in a city selected based on customer preferences. The customer does not directly select the city, but decides what focus the trip should have, how expensive the location should be, if he likes a populous capital or a less crowded one and the geographical areas they feel most comfortable with.

The customer can provide following parameters:

- Main theme: Food, Nightlife, Freetime and Shopping
- Cost: low, medium, high
- Population: low, medium, high
- Geographic location: south, center-west, center-east, north

Getaway.com is building a network of European capitals and wants to gather data and have a first analysis of those data to be able to further develop their service. The information needed are the following:

- Number of venues for each category that each city has to offer
- Which cities are most suitable for each of the 4 main themes
- the cost of life in a given city
- the population in a given city
- the geographical location of a given city
- a readable dataset that can quickly identify the best match given a set of parameters

The initial analysis should focus on clustering the cities into groups within each of the 4 themes, cost and population levels and geographical location. It should also provide a quick summary of the data and a score system that simplify decision making.

2. Data

In order to perform the analysis required by GetAway.com, we will retrieve and analyze following data:

- A list of the European Capitals with their geographical coordinates, from which we will select some cities to perform the analysis. Such list is readily available and is source-independent (the European Capitals will not change name or position if we change the source of our information). Therefore we will import the data from an available csv file.
- Population data will be retrieved from "worldpopulationreview.com" and imported from there. The population data are from 2021, according to the source website.
- Cost of Life data will be retrieved from "myfunkytravel.com", which is in turn based on data from Numbeo, a well-known database for cost of living data.
- Information regarding the number and type of venues for each city will be retrieved using the FourSquare API. For each Capital a fixed search radius of 5 Km will be used. The data will be retrieved as per 13/03/2021.

All data will be imported, processed, analyze and prepared for the final report using python.

The data will be processed so that at the end we obtain a database containing:

- City
- Area
- Categorical score for: Population, Cost of Living
- Numerical Score for: Food, Nightlife, Freetime, Shopping
- Numerical overall score

The data will focus on the following capitals (by geographical area):

- South: Rome, Madrid, Lisbon
- Center-west: London, Paris, Amsterdam, Brussels
- Center-east: Berlin, Prague, Vienna, Budapest
- North: Copenhagen, Stockholm, Oslo

3. Methodology

3.1 Preparing dataset with population and location data and assign an area

As discussed in the "Data" section, we are retrieving the location and population data from two different sources. In order to facilitate the analysis we will merge them into one dataset. Additionally, we need to divide the selected cities into one of the 4 areas given by GetAway.com. We are selecting a total of 14 capitals. The final dataset is displayed in figure 1:

	Name	2021 Population	CapitalLatitude	CapitalLongitude	Area
0	London	7556900	51.500000	-0.083333	Center_west
1	Berlin	3426354	52.516667	13.400000	Center_east
2	Madrid	3255944	40.400000	-3.683333	South
3	Rome	2318895	41.900000	12.483333	South
4	Paris	2138551	48.866667	2.333333	Center_west
5	Budapest	1741041	47.500000	19.083333	Center_east
6	Vienna	1691468	48.200000	16.366667	Center_east
7	Stockholm	1515017	59.333333	18.050000	North
8	Prague	1165581	50.083333	14.466667	Center_east
9	Copenhagen	1153615	55.666667	12.583333	North
10	Brussels	1019022	50.833333	4.333333	Center_west
11	Amsterdam	741636	52.350000	4.916667	Center_west
12	Oslo	580000	59.916667	10.750000	North
	bon	517802	38.716667	-9.133333	South

Figure 1

3.2 Retrieving the venues data

To classify the different cities and areas into one of the categories (food, nightlife, freetime or shopping) we need to know which type of venues are mostly present in a given city. We use the FourSquare API to retrieve such data.

The function is set with a radius of 5 Km. This is more than enough to ensure that we reach the maximum number of location (limited to 100 per city) even for the smaller cities. After gathering the venues date we processed them into a pandas dataframe.

3.3 Retrieving Cost of Living data

GetAway.com also wants to have some data regarding the costs of each cities. We gathered the data for the cost of living in each city. Even though GetAway.com will focus on tourists and not on locals, the cost of living for the locals highly correlates with the costs of amenities and touristic attractions.

For the city of Oslo there is no available cost of living in the database that we consulted. Searching this specific data point in another database is risky, because of the many parameters and variables that can be considered when calculating the cost of living. Instead, we reasoned that Oslo has a cost of living that can be reasonably approximated with that of Stockholm. Therefore, we assigned the value of Stockholm to Oslo as well.

3.4 Splitting the venues into 4 main types

As per request from GetAway.com, the venues need to be assigned to one of 4 types. The data collected from FourSquare gave 1400 venues divided into 212 categories, for an average of less than 7 venues per category. In order to satisfy the Customer requirement, we divided the venues into the given 4 types using following criteria:

- Each category of venue can only be assigned to one type (i.e. the FourSquare category “Café” can be assigned to only one of the types)
- The same category will always be assigned to the same type (i.e. the FourSquare category “Café” will always be assigned to “Food”)
- Only categories that appear at least 5 times will be assigned (i.e. if the FourSquare category “Flower Shop” appears only 4 times in the whole dataset, then it will not be assigned and it will be removed from the dataset)

As a result, the final dataset only has 4 types of venues and less then 1400 venues (because all those with less than 5 entries were discarded). After the processing the dataset still has 1117 venues, for an average of 279 venues per type.

The type food encloses all kind of restaurants, pizzerias, café, snack bars and places where people mostly eat.

The type nightlife encloses all bars, pubs, lounges, hotels, hostels and places where people go to drink or dance during the night.

The type freetime encloses leisure activities/places such as parks, cinema, festivals and cultural activities such as museums, operas and theaters.

The type shopping encloses all venues that are dedicated to shopping, such us mall, stores and markets.

3.5 One hot encoding

Our goal is to find out if some cities are better than other for certain activity types and if this correlates also with other characteristic such as population or location. We aim to get the information from the venue data without inputting rules or constrains. Therefore, we will perform most of the analysis using k-means clustering. In order to do that, we need to transform the categorical data into numeric values. By

performing the one hot encoding, we obtain a dataset that counts each time that a venue of a specific type is present. This will allow us to easily count the number of recurrences and ultimately to cluster the cities based on their tendency to have more ”food” venues or “shopping” venues.

An extract of the dataset after the one hot encoding can be seen in figure 2:

	City	Food	Freetime	Nightlife	Shopping
0	London	1	0	0	0
1	London	0	1	0	0
2	London	1	0	0	0
3	London	0	0	1	0
4	London	0	0	0	1
...	...	...	...	...	...
1112	Lisbon	0	0	1	0
1113	Lisbon	1	0	0	0
1114	Lisbon	1	0	0	0
1115	Lisbon	1	0	0	0
1116	Lisbon	0	1	0	0

Figure 2

3.6 Preparing the final venue dataset

With some standard pandas processing we created the dataframe on which we will base our analysis. Each city has one row and 9 columns:

- Food: represents the number of venues of type “food” found
- Nightlife: represents the number of venues of type “nightlife” found
- Freetime: represents the number of venues of type “freetime” found
- Shopping: represents the number of venues of type “shopping” found
- Total Venues: represent the total number of venues found
- City Latitude: contains the latitude coordinate
- City Longitude: contains the longitude coordinate
- 2021 Population: contains the city population as per 2021
- Estimated cost of living(euros/month): contains the estimated cost of living for a person resident in that city, including rent.

The final dataset if represented in figure 3:

	City	Food	Nightlife	Freetime	Shopping	Total Venues	City Latitude	City Longitude	2021 Population	ESTIMATED COST OF LIVING(EUROS/MONTH)
0	Amsterdam	25	27	18	4	74	52.350000	4.916667	741636	1702
1	Berlin	25	17	27	3	72	52.516667	13.400000	3426354	1258
2	Brussels	21	22	21	16	80	50.833333	4.333333	1019022	1387
3	Budapest	34	28	13	4	79	47.500000	19.083333	1741041	809
4	Copenhagen	25	24	24	5	78	55.666667	12.583333	1153615	1719
5	Lisbon	32	27	24	1	84	38.716667	-9.133333	517802	1004
6	London	19	32	28	6	85	51.500000	-0.083333	7556900	1926
7	Madrid	32	13	35	4	84	40.400000	-3.683333	3255944	1168
8	Oslo	36	25	14	6	81	59.916667	10.750000	580000	1576
9	Paris	20	15	39	3	77	48.866667	2.333333	2138551	1660
10	Prague	42	14	23	6	85	50.083333	14.466667	1165581	919
11	Rome	29	11	39	4	83	41.900000	12.483333	2318895	1387
12	Stockholm	43	22	9	4	78	59.333333	18.050000	1515017	1576
13	Vienna	31	10	31	5	77	48.200000	16.366667	1691468	1341

Figure 3

3.7 K-means and Clustering

We performed a k-mean analysis in order to obtain clusters of city with similar characteristic. GetAway.com needs to be able, given the type of venues as input, to tell which cities are most likely to satisfy this criteria. Therefore, we analyzed one type of venue per time and clustered the cities according to their affinity with the given type of venue.

We performed the k-means analysis with 4 clusters for the venue types “food”, ”nightlife”, “freetime” and “shopping”. We chose 4 clusters because the cities are clustered into 4 geographical areas and therefore we might be able to spot some correlation between geographical location and some prevalent venue type.

We performed the k-means analysis with 3 clusters for the population and cost of living parameters, because the specification of GetAway.com was based on a 3-level classification (high, medium, low). Therefore it would not make sense to have more than 3 clusters.

While performing cluster analysis we always used the absolute number of venue per type per city. Before deciding this, we checked if it was meaningful to used the relative number of venues per 100.000 inhabitants. Normally, this would be a good way to normalize per population. Unfortunately, the API service only returned the first 100 venues for each city. Therefore, using the number of venues per 100.000 inhabitants would skew the data towards the smaller cities quite decisively. In figure 4 we show the clear correlation between population and number of venues per 100.000 inhabitants (correlation that would not normally be expected). As consequence, we performed the analysis with the absolute number of venues.

	City	2021 Population	Total_p
5	Lisbon	517802	16.222417
8	Oslo	580000	13.965517
0	Amsterdam	741636	9.977941
2	Brussels	1019022	7.850665
4	Copenhagen	1153615	6.761355
10	Prague	1165581	7.292500
12	Stockholm	1515017	5.148457
13	Vienna	1691468	4.552259
3	Budapest	1741041	4.537515
9	Paris	2138551	3.600569
11	Rome	2318895	3.579291
7	Madrid	3255944	2.579897
1	Berlin	3426354	2.101359
6	London	7556900	1.124800

Figure 4. Showing the inverse proportionality between population and venues per 100.000 inhabitants in our dataset.

3.8 Data Visualization

We displayed the data from the k-means analysis in form of maps which geolocalize each city and assign it to a specific cluster through colors.

We additionally displayed the number of Total Venues, Food venues, Nightlife Venues, Freetime Venues and Shopping Venues in a bar chart in order to highlight the contrasts between the different cities.

Finally, we have aggregated the data by area and we show for each venue type which percentage of the venues are located in each area. We used pie charts for this visualization.

3.9 Scoring

For each cluster analysis a score is calculated for each city. The score is 100 for the cities in the top 25%, 75 in the top 50%, 50 in the top 75% and 25 for those in the last quartile. Assigning a score makes it possible to calculate an overall “best” location for the combination of the 4 types of activity. It also makes very easy to identify which cities excel in each type of venues. The score for Population and Cost of living has categorical values of “high”, “medium” and “low”, as the preference for one or the other is quite subjective.

4. Results

The clustering of the cities according to the total number of venues (Figure 5) shows a quite homogeneous division, with clusters that are approximately the same size.

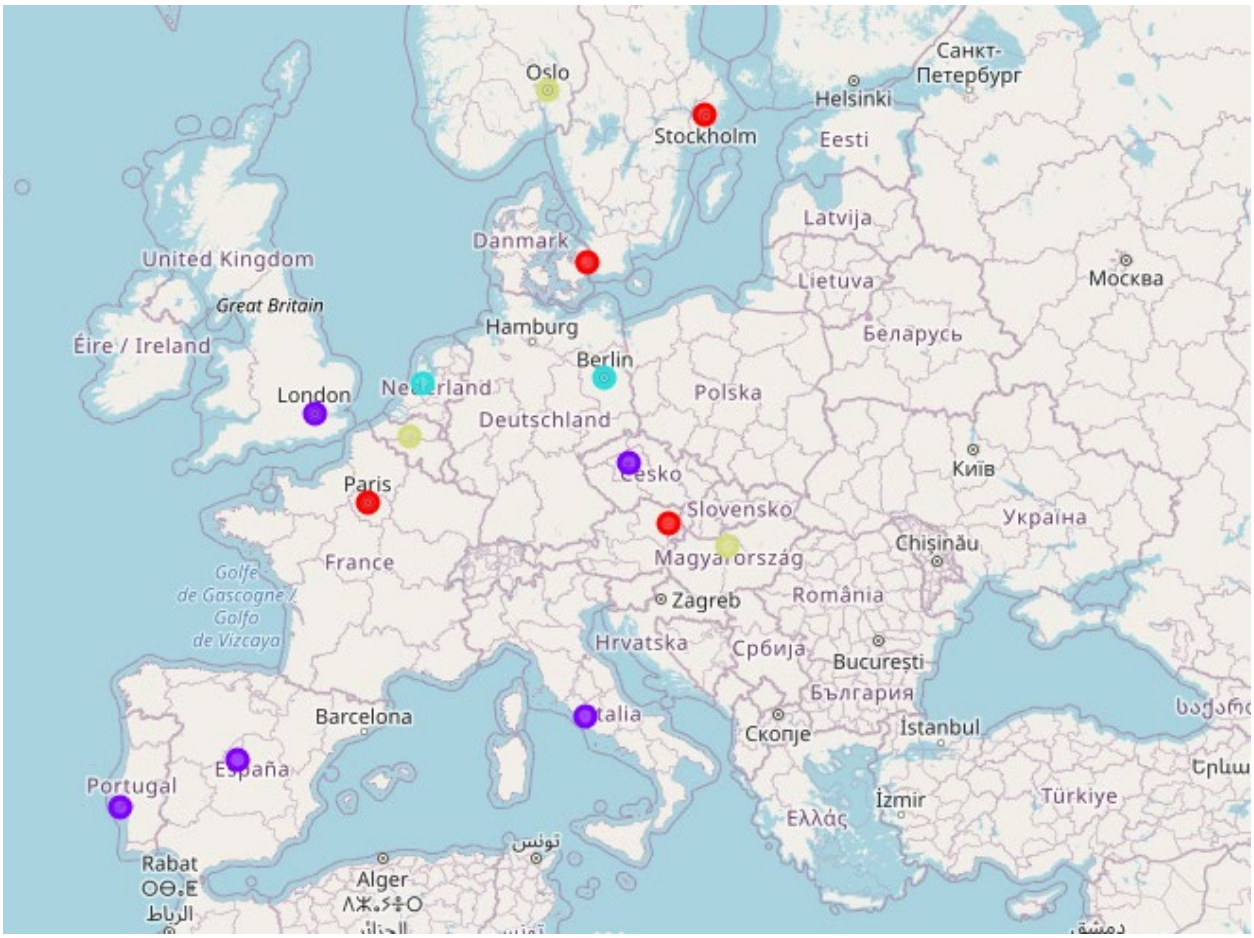

Figure 5:
Purple = very high number of venues
yellow= high number of venues
red= moderate number of venues
light blue=moderate-low number of venues

The clustering according to number of food, nightlife and freetime venues (Figure 6, 7 and 8) also displays homogeneous clusters with very similar size.

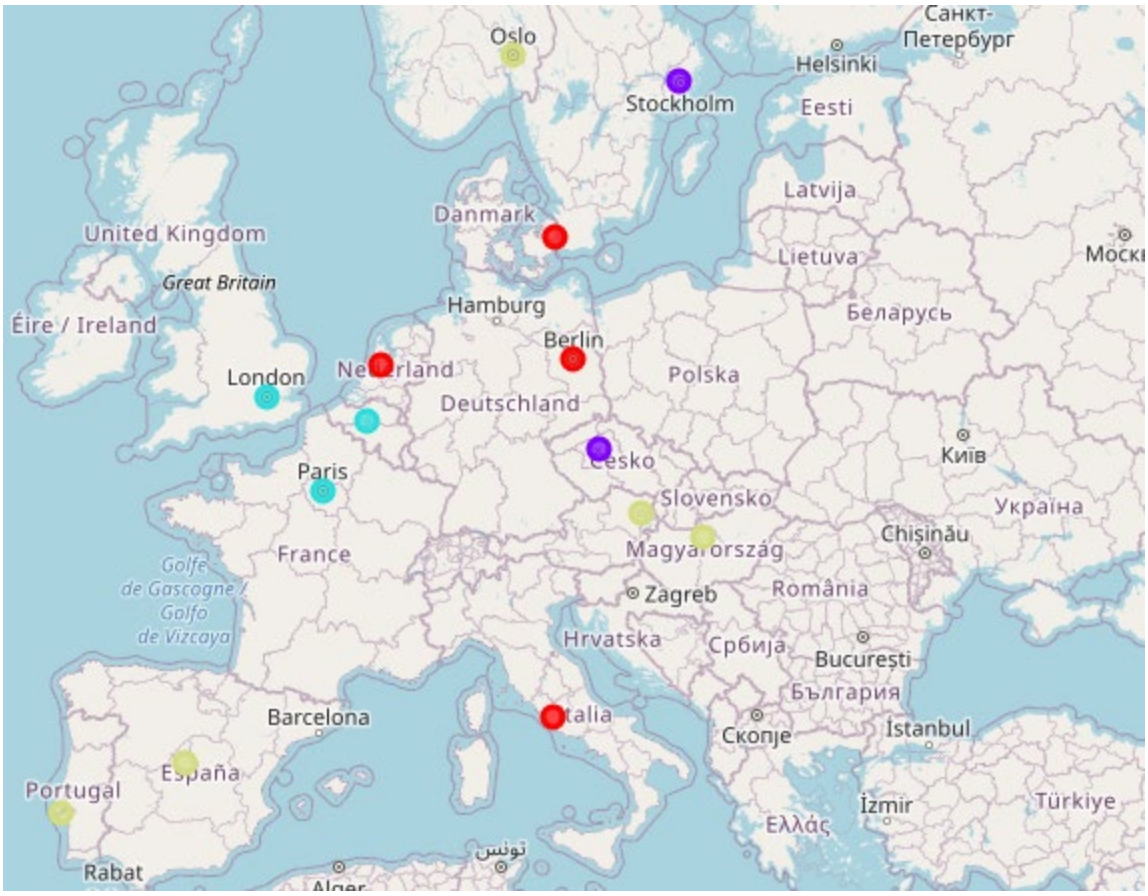

Figure 6
Purple = very high number of food venues
yellow= high number of food venues
red= moderate number of food venues
light blue=moderate-low number of food venues

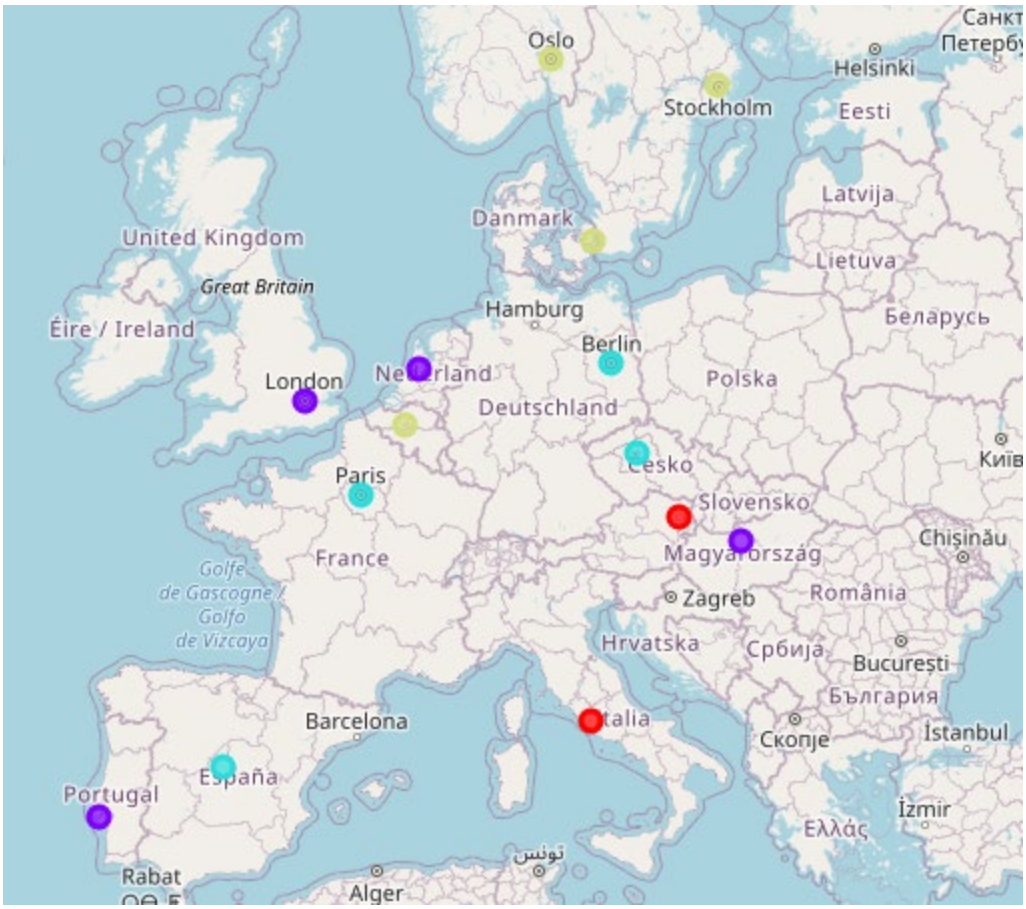

Figure 7
Purple = very high number of nightlife venues
yellow= high number of nightlife venues
light blue= moderate number of nightlife venues
red=moderate-low number of nightlife venues

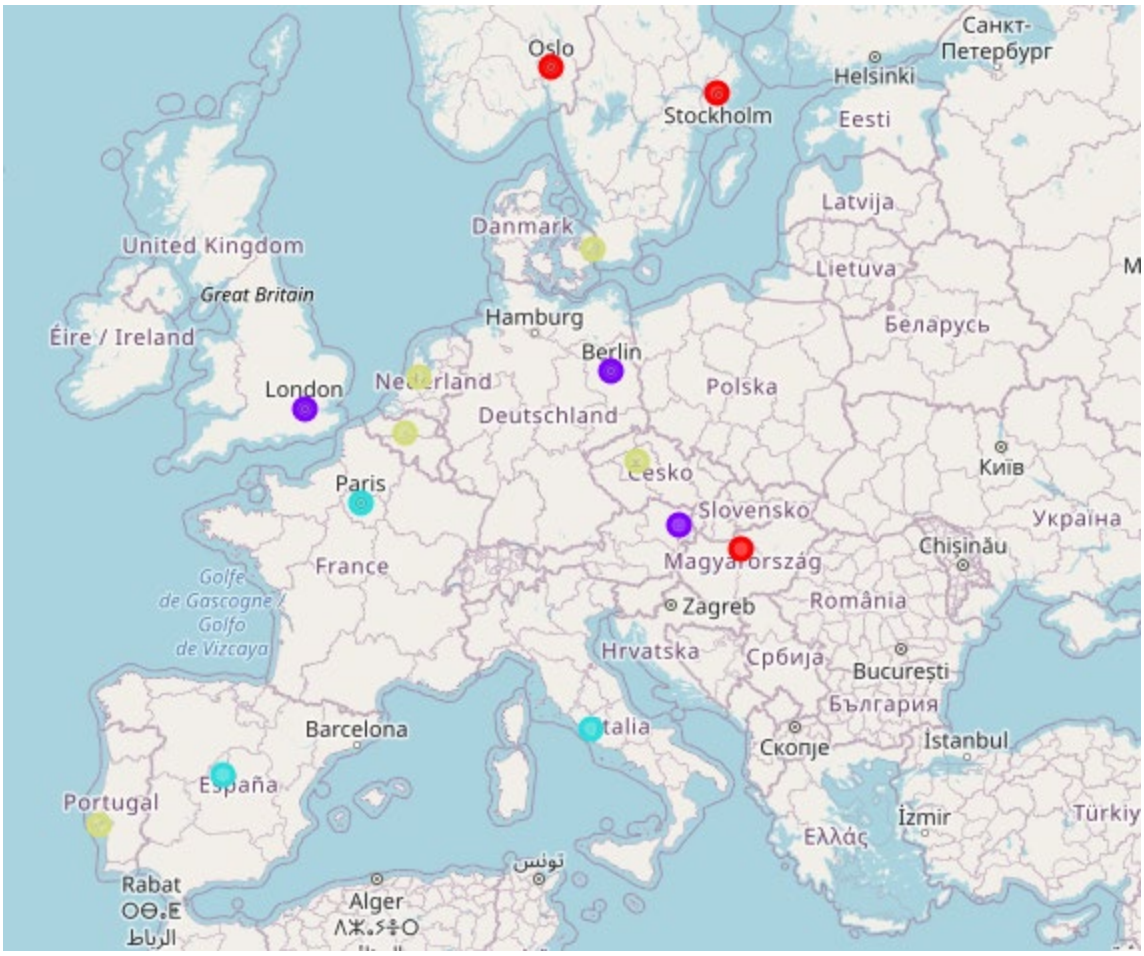

Figure 8
Yellow = very high number of freetime venues
Purple= high number of freetime venues
light blue= moderate number of freetime venues
red=moderate-low number of freetime venues

The clustering according to number of shopping venues (Figure 9) results in 2 large clusters and 2 cluster with only one member. This mean that there are bigger differences among these cities when we only take shopping venues into consideration.

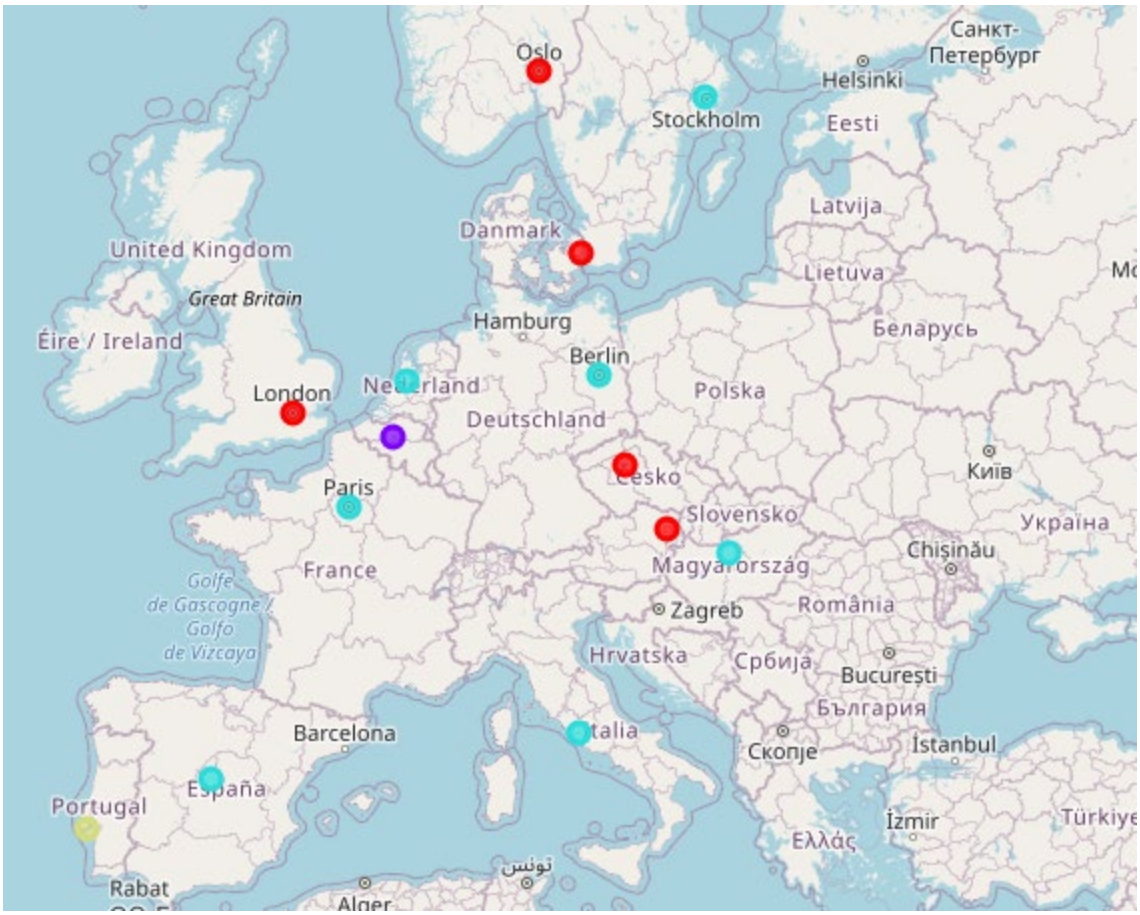

Figure 9
Purple = very high number of shopping venues
Red= high number of shopping venues
Light blue= moderate number shopping venues
Yellow=moderate-low number of shopping venues

When clustering the cities according to the cost of life (Figure 10), we noticed a clear division between the cities in the North and Center-West areas and the rest. The first ones are, with the exception of Brussels, in the cluster of „Highly expensive“. This is the first clear correlation between the geographical areas and the clusters generated by our analysis.

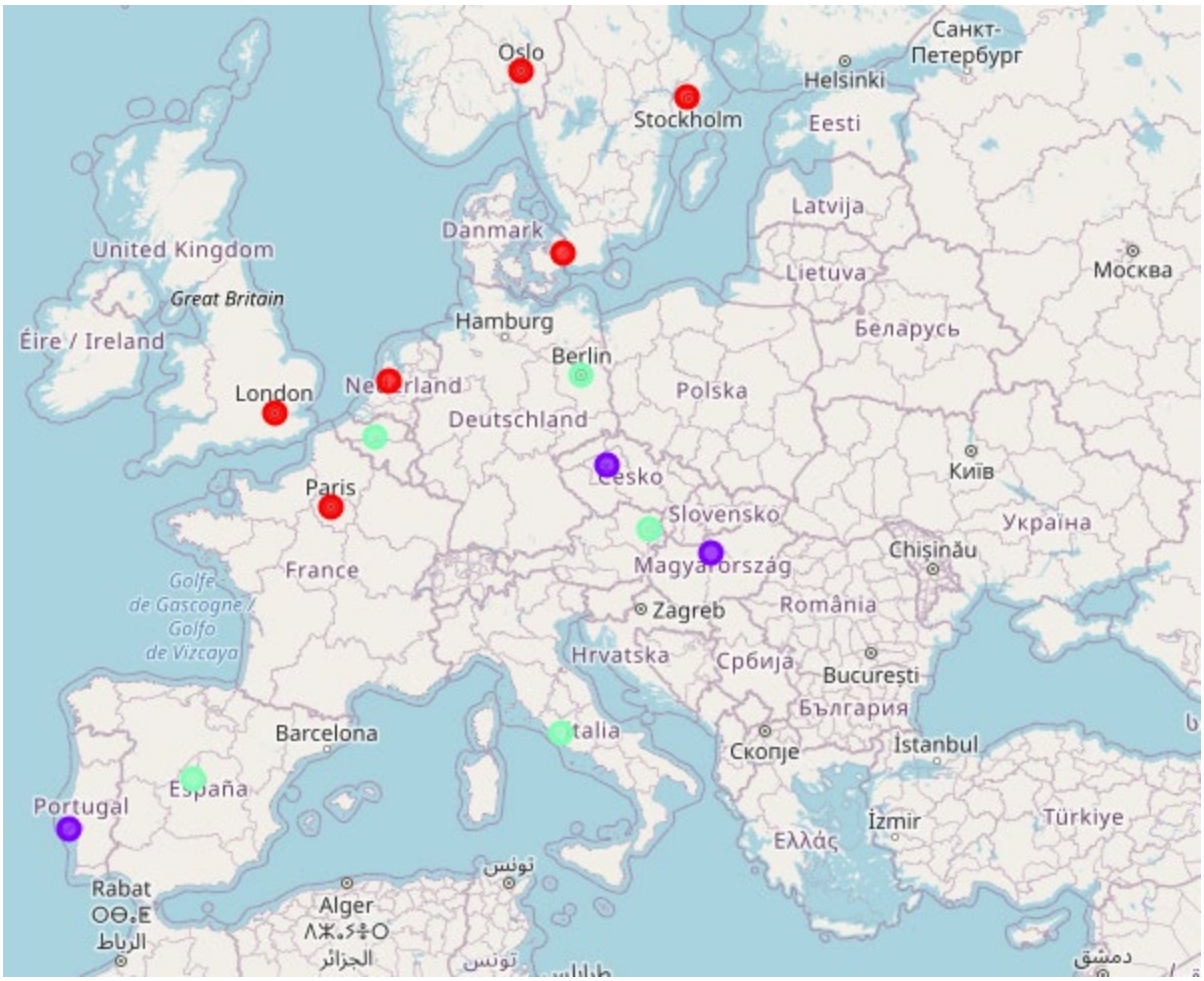

Figure 10
Red = Expensive cities
Light Blue= affordable cities
Purple= chep cities

The clustering according to the population shows that the areas North and Center-east are homogenously low populated, with the exception of Berlin. London is the clear outsider and takes a cluster for itself.

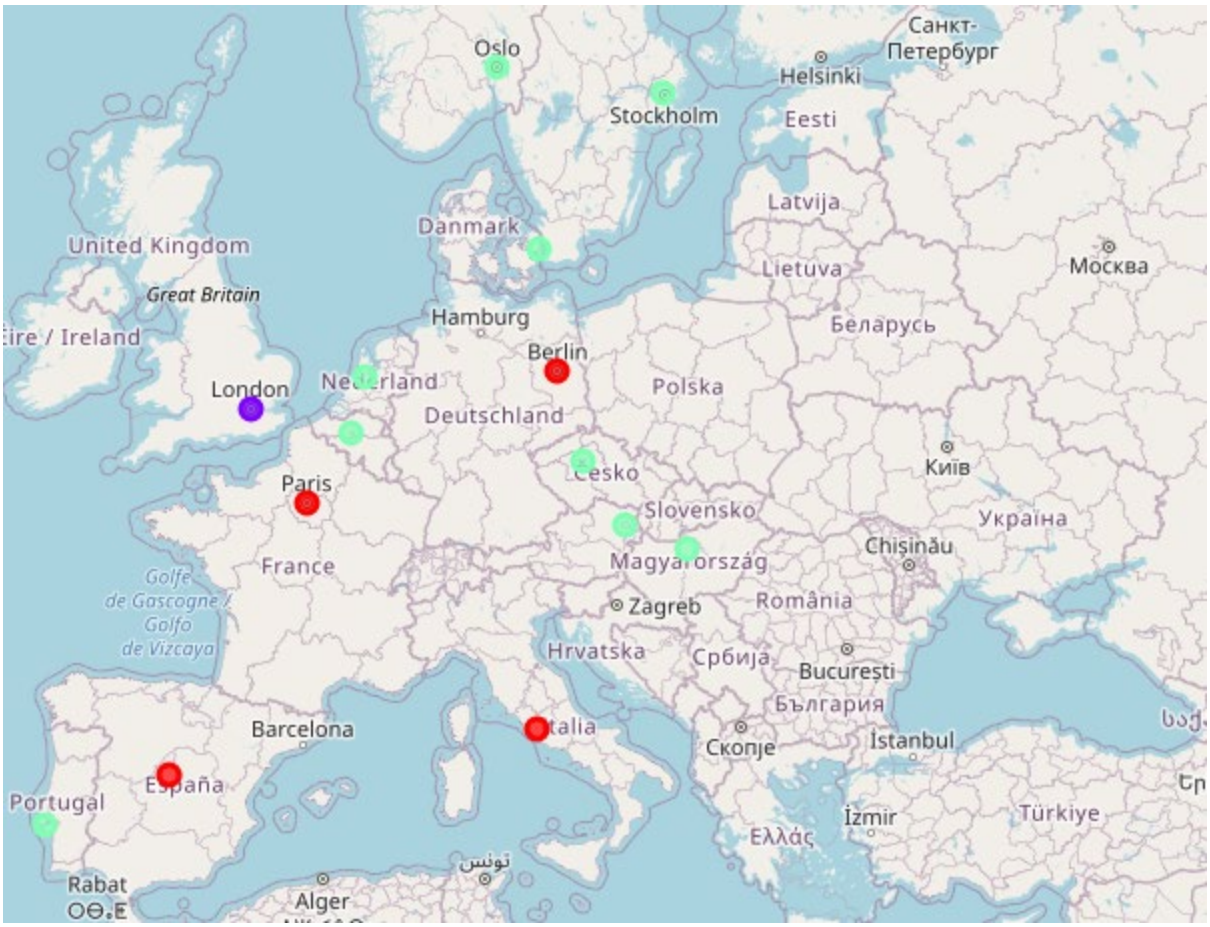

Figure 11
Purple = High population
Red= medium population
Light Blue= Low population

Similar results are shown by the bar charts (Figure 12, 13, 14, 15, 16).

Total number of venues is very homogeneous across the cities. Food, nightlife and freetime venues are less homogeneous and the extremes tend to be further apart.

While the bar charts do not provide many additional information, they do allow for a better overview of the inter- and intra-cluster differences.

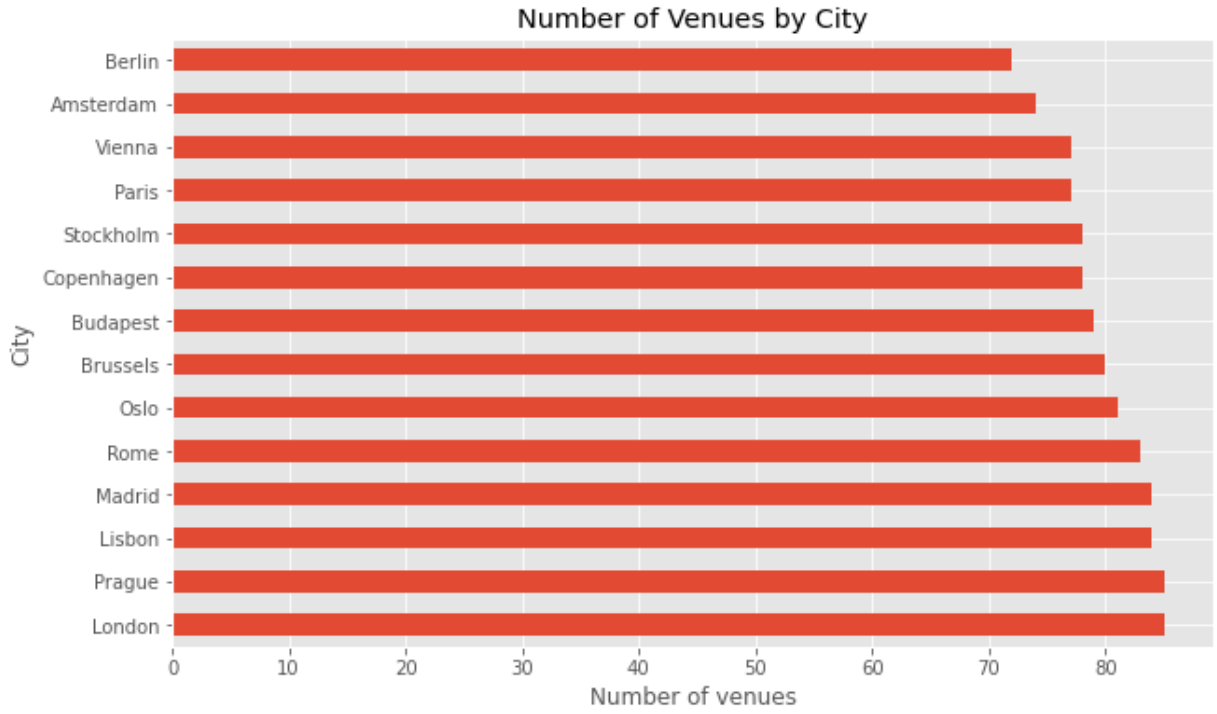

Figure 12. The total number of venues is well distributed, with no outlier.

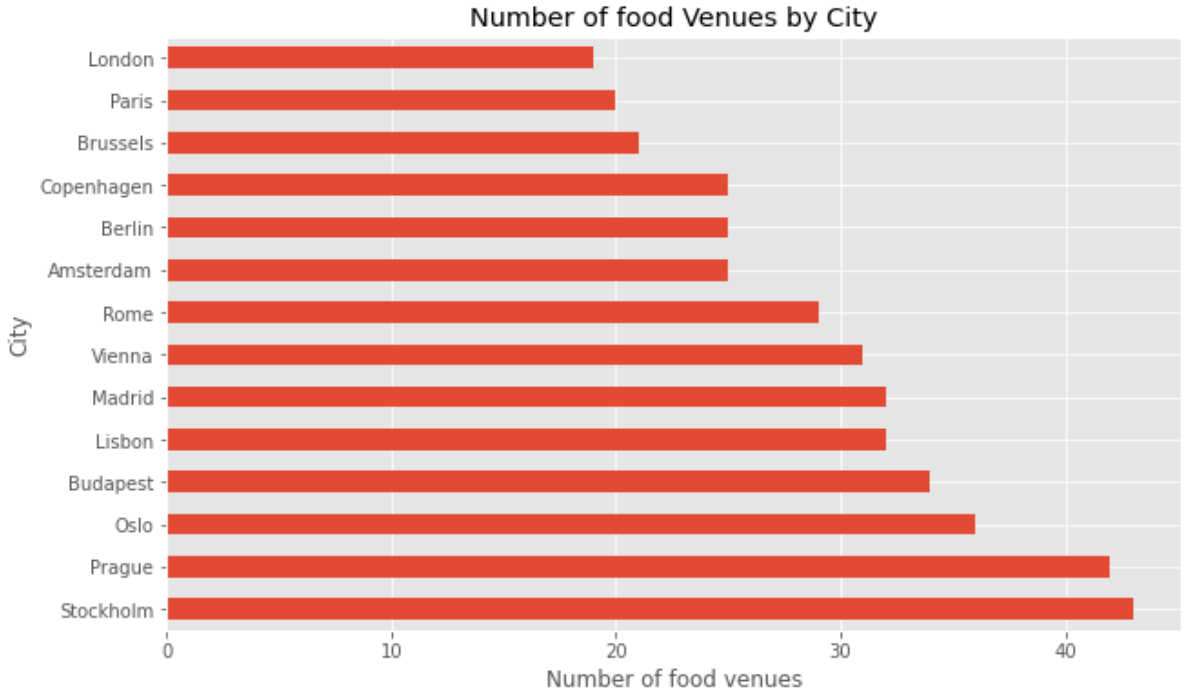

Figure 13. The number of food venues is quite different from one cluster to the other.

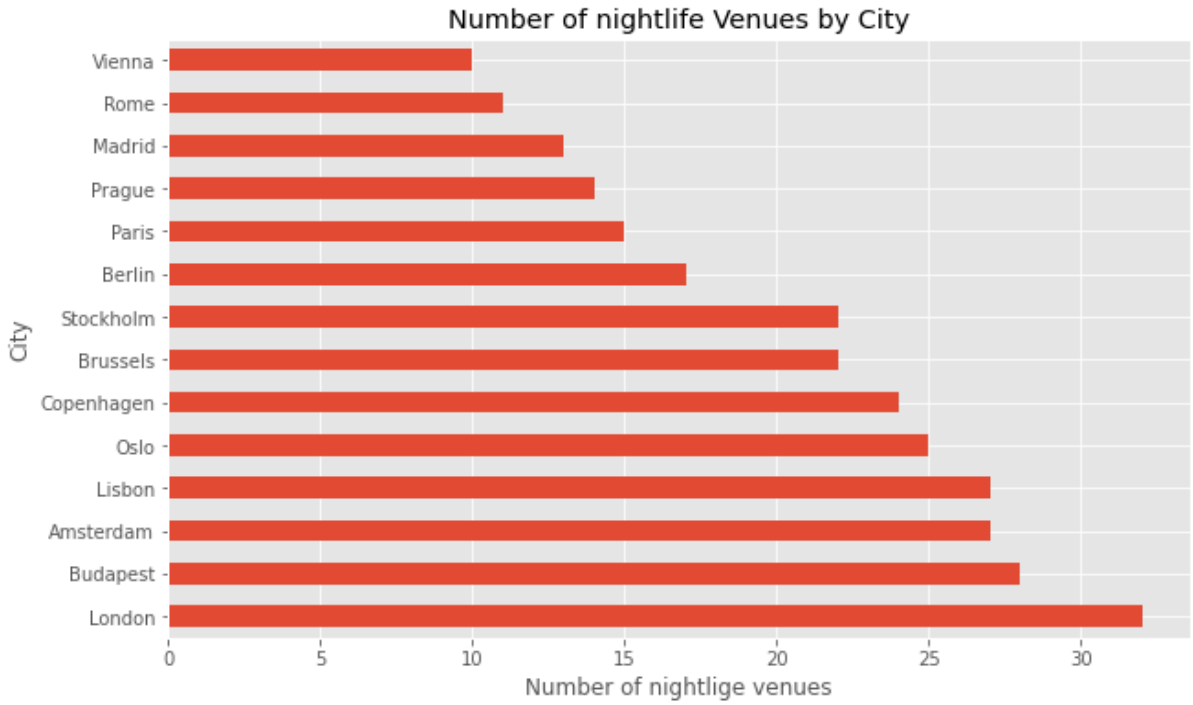

Figure 14. The profile is very similar to the one of food venues

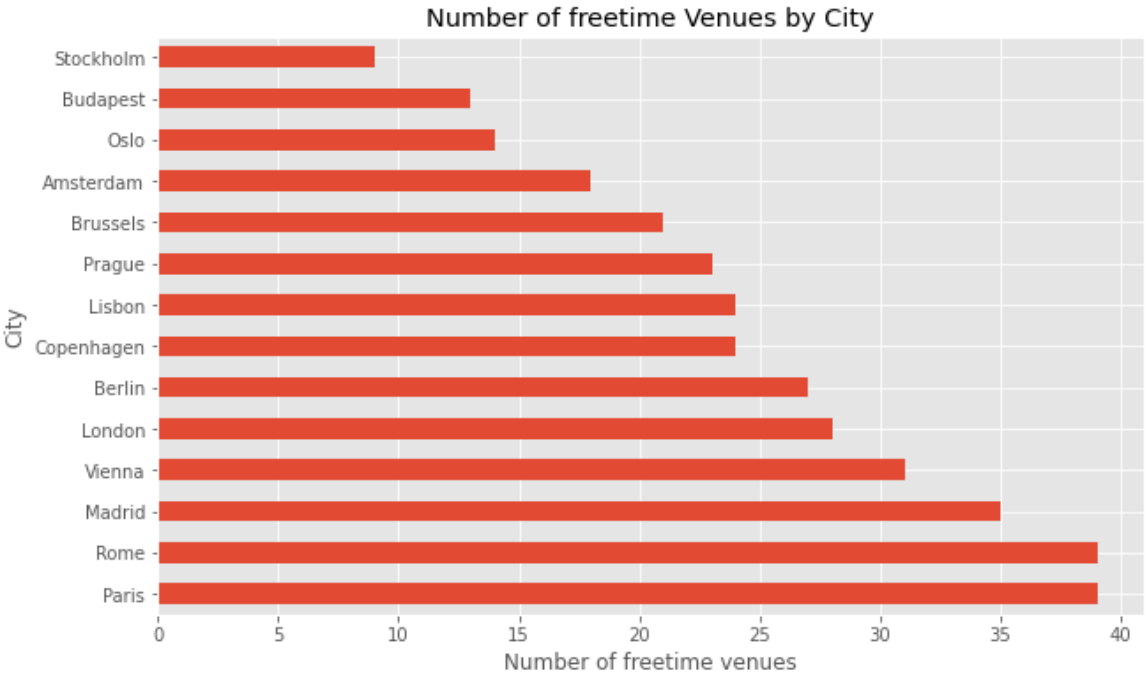

Figure 15. Again, the chart reflects the same profile, although there is an increased sistance between the extremes.

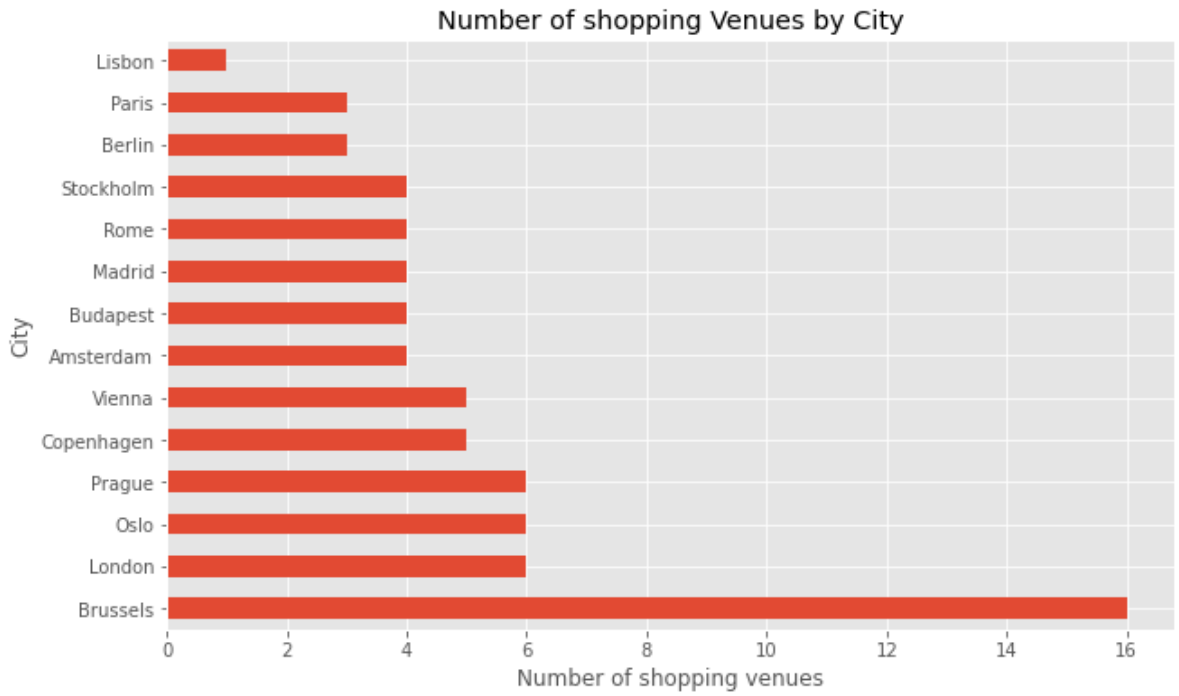

Figure 16. The profile of the shopping venues data shows a clear outsider in Brussels, with more than double the number of venues of the second city ranked.

The cost of living, which as seen in the cluster maps clearly divides the North and Center-west areas from the rest, changes continuously and regularly without big gaps (Figure 17). Although the transition from one extreme to the other is quite smooth the, London does have a cost of living that is more than double that of the three lowest ranked cities.

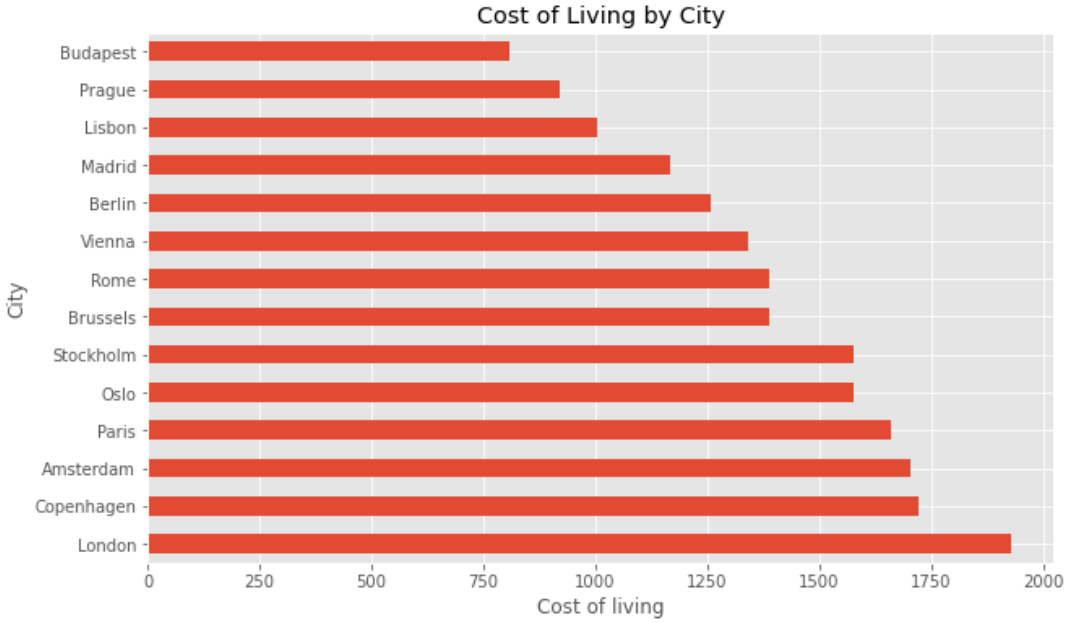

Figure 17. Cost of Living

One closer look at the different geographic areas shows some differences in how the different type of venues are distributed (Fig 18):

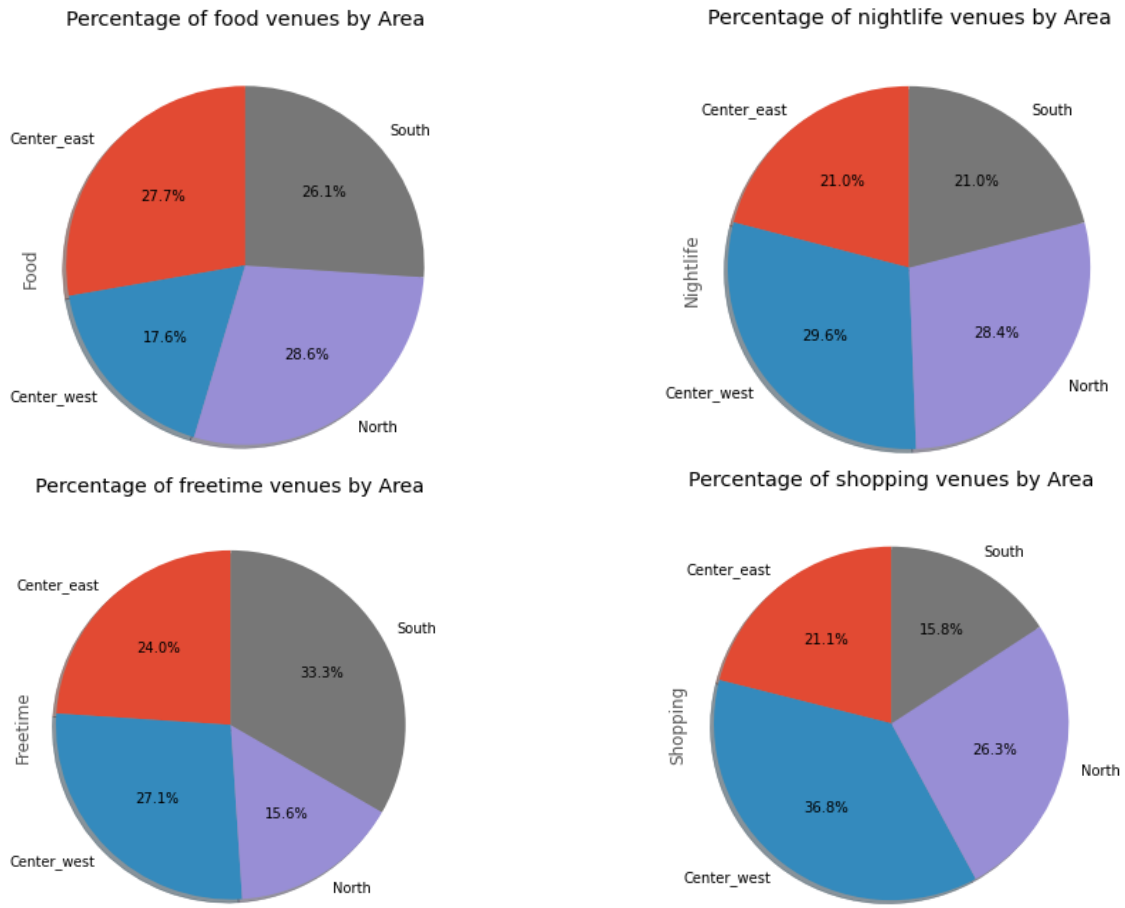

Figure 18

It can be clearly seen that the differences are much less pronounced when looking at the entire area instead of the single cities. For example, the extreme outliers in the shopping venues even out almost completely. The most clear result from the area analysis are a preponderance of freetime venues in the South and of shopping venues in the Center-west (each of them has more than 1/3 of the total).

No area reaches or approaches 50% in any sector. This means that the analyzed areas are mostly very similar with regard to these types of venues.

5. Discussion

The data presented clearly show that it is possible to identify preferred location given the type of venues. It is also possible to refine such location based on preferences such as cost, population of the city and geographical area.

The final data are summarized in Figure 19:

City	Area	food_score	nightlife_score	freetime_score	shopping_score	Population	Cost	Sum
London	Center_west	25	100	75	75	High	High	275
Madrid	South	75	50	100	50	Medium	Medium	275
Prague	Center_east	100	50	50	75	Low	Low	275
Budapest	Center_east	75	100	25	50	Low	Low	250
Vienna	Center_east	75	25	75	75	Low	Medium	250
Stockholm	North	100	75	25	50	Low	High	250
Copenhagen	North	50	75	50	75	Low	High	250
Brussels	Center_west	25	75	50	100	Low	Medium	250
Amsterdam	Center_west	50	100	50	50	Low	High	250
Oslo	North	75	75	25	75	Low	High	250
Lisbon	South	75	100	50	25	Low	Low	250
Berlin	Center_east	50	50	75	50	Medium	Medium	225
Rome	South	50	25	100	50	Medium	Medium	225
Paris	Center_west	25	50	100	50	Medium	High	225

Figure 19

Looking at the “Sum” column on the right of the table, it is possible to clearly recognize 3 clusters of cities.

The best cities overall for a combination of the 4 types of venues are Prague, London and Madrid. They are each in a different area and each has different cost and population. By applying any condition on cost, population or area it is possible to immediately have an indication of the most suitable option.

The best cities for a food experience are, according to this analysis, Prague and Stockholm. For Nightlife, London, Budapest, Amsterdam and Lisbon get the top places. To quietly walk in a park or visit a cultural exhibition, Madrid, Rome and Paris are the best (and they are indeed well known to be among the cities with the biggest and most famous cultural agendas). For shopping, the relatively small Brussels seems to be the better option. Excluding London, all the top cities for shopping have a population classified as

“low”. We need to remember that we are considering the number of shopping venues. Likely, this result does not mean that smaller capitals are better for shopping, but rather than those small cities have a plurality of small shops and less huge shopping centers and therefore might provide a more relaxed and unique shopping experience.

It is also apparent, that the cost of living is not necessarily correlated with the options available. The “low” cost cities seem to score very well with food and nightlife venues and more poorly with free time and cultural offers. This might also hint that culturally rich activities might be more on the expensive side compared to experiencing food or enjoying a few drinks in pubs and bars.

Cities with “medium” population seems to have very high scores with free time and cultural venues, but also seems to be overall ranked towards the bottom. Center-east cities have very good scores for food venues. Capitals in the North area have high scores for nightlife venues.

6. Conclusion

Given the starting requirement of correlating types of venues with different cities across different geographical areas, we can say that it was possible to find some patterns where cities that are correlated by geography or population can also be correlated with specific types of venues. It was possible to create an extremely readable dataset that, given a choice of venues type, geographic area, size or cost, allows to identify one or more cities that fit the given criteria better than other.