Service based features analysis of cities surrounding Baltic Sea

Applied Data Science Capstone project

Alina Šerienė

2020-05-25

Table of Contents

1. Introduction

1.1 Background

The Baltic Sea is a Mediterranean sea of the Atlantic Ocean. It is enclosed by nine different countries - Denmark, Estonia, Finland, Latvia, Lithuania, Sweden, northeast Germany, Poland and Russia. Countries surrounding the sea are quite different by their size, economy, political system and history. Due to it‘s location, sea and surrounding territories were always in area of conflict i.e. during World War I and II.

Nowadays Baltic Sea is main seafood provider for surrounded territories due to developed fishing in area. Also oil is extracted in small-scale from the Baltic Sea as well as amber is being mined on the coast of Kaliningrad region. Sea is an important waterway for export and import of surrounded countries.

Climate around the sea is usually quite cold, compared to the most popular holiday destination seas, such as Caribbean, Mediterranean or Black seas. Water surface temperature varies depending on the season from 0 to 20 degrees Celsius with annual average of 9-10 °C.

1.2 Problem

Baltic Sea, due to its comparably cold climate was not usually a top holiday destination. Even citizens of surrounding countries usually preferred different places for holidays. As a result Baltic Sea does not have extensive analysis of tourism environment in the area.

Recently Global Climate Change is making surrounding territories more and more attractive for tourism. Weather in surrounding territories is getting warmer but is not too hot as in i.e. Egypt or Turkey, where during hot days temperature may even cause illnesses. Covid-19 crisis also made everyone to turn back to closer holiday spots.

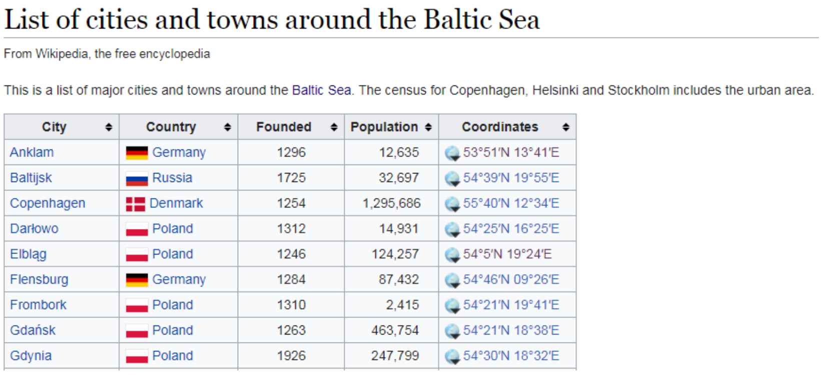
1.3 Interest

Shortage of analysis and information on Baltic Sea region tourism opportunities creates a need both for travelling people to know the area better and to find most attractive areas in the region as well as for business people to find investment opportunities in area having increasing potential.

2. Data acquisition and cleaning

2.1 Data sources

To reach research objective, it is needed to look into areas surrounding Baltic Sea. To do this, article from Wikipedia containing list of cities with some additional information was chosen as an initial source of data. Extract of table provided below as well as web address of the table.



List of cities and towns around the Baltic Sea

From Wikipedia, the free encyclopedia

This is a list of major cities and towns around the Baltic Sea. The census for Copenhagen, Helsinki and Stockholm includes the urban area.

City	Country	Founded	Population	Coordinates
Anklam	Germany	1296	12,635	53°51′N 13°41′E
Baltijsk	Russia	1725	32,697	54°39′N 19°55′E
Copenhagen	Denmark	1254	1,295,686	55°40′N 12°34′E
Darłowo	Poland	1312	14,931	54°25′N 16°25′E
Elbląg	Poland	1246	124,257	54°5′N 19°24′E
Flensburg	Germany	1284	87,432	54°46′N 09°26′E
Frombork	Poland	1310	2,415	54°21′N 19°41′E
Gdańsk	Poland	1263	463,754	54°21′N 18°38′E
Gdynia	Poland	1926	247,799	54°30′N 18°32′E

Table 1. Extract of initial dataset in webpage: https://en.wikipedia.org/wiki/List_of_cities_and_towns_around_the_Baltic_Sea

Coordinates of cities provided in Wikipedia‘s article list are compared to Geopy library outcomes. As a result, this library can also be included into the list of data sources.

Another large portion of data is taken from Foursquare API to get cities‘ venues information and group them into major categories for further clustering.

2.2 Data pre-processing

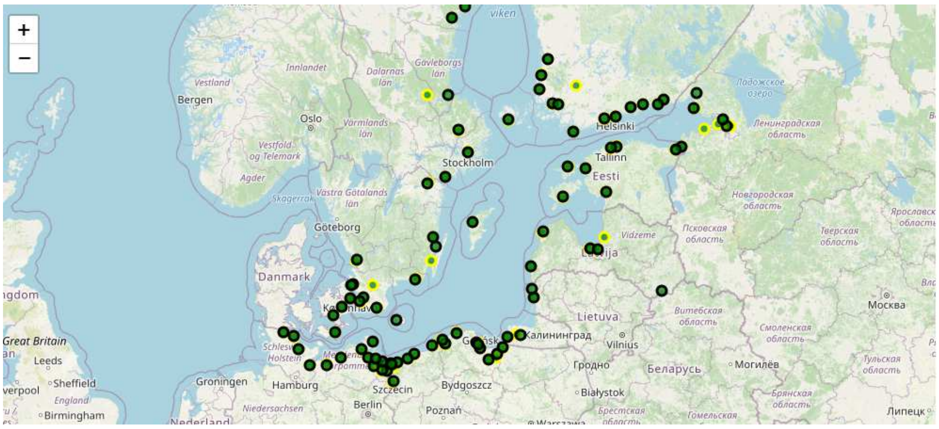
Initial data list with cities had several problems to be solved. First of all it had two types of coordinates in the same column. As a result, these had to be split and GPS coordinates were dropped. Next issue was to make Founded, Population and Coordinates fields' numeric. While it was quite strait forward with first two columns, coordinates required more iterations. The only issue with Founded column, was that some values had text elements inside and had to be replaced:

```
df['Founded']=df['Founded'].str.replace("1865 (city 2002)","1865")
df['Founded']=df['Founded'].str.replace("15th century","1450")
df['Founded']=df['Founded'].str.replace("14th century","1350")
df['Founded']=df['Founded'].str.replace("1150s","1150")
df['Founded']=df['Founded'].str.replace("13th century","1250")
```

The code above represent the changes made. As this field is not used for modelling and will be involved only in after-clustering analysis, extreme accuracy is not required in this field. As a result, mid of centuries is selected as value for replacing century.

What is uncommon to Decimal coordinates, the ones provided in Wikipedia list had letters representing direction involved. As a result, these had to be removed from dataset. In addition, Latitude values had hidden string „\ufeff53.850" which also had to be replaced before transforming column to numeric values.

Next issue with data quality was accuracy of coordinates provided in initial list. Despite of strange formatting, coordinates seemed to be not fully correct after checking them against the ones provided in Geopy library. After plotting both coordinates in one map, it was clear, that coordinates from initial list had to be dropped and Geopy ones left for further analysis.



Picture 1. Map representing the mismatch of coordinates in initial list and Geopy information. Geopy coordinates represented in yellow bordered circles.

After initial data was in place, Foursquare venues information came into picture. Request url was made to Explore Venues in cities provided in the list with 10km radius. Using wider radius was not required as for small cities it is fully enough and for big ones it the maximum limit on number of venues was still limiting the total count of venues.

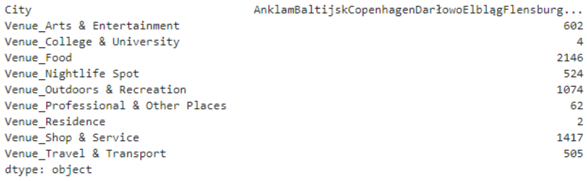
Following fields were extracted from the dataset: 'City', 'City Latitude', 'City Longitude', 'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category'. Venues information was the one, which was used for modelling. As the main purpose if this analysis was to segment the cities around Baltic Sea it is not so important to look into the detailed categories of venues. To understand the type of the city, I‘ve decided to use parent categories of all venues categories and do modelling on summarized data.

Foursquare categories dictionary is made recursively. Meaning, that you have to go step by step from the most detailed category to the most general one. After initial mapping with extracted two level dictionary in data frame format, it was clear, that at least three iterations will be needed, deciding on levels identified.

Foursquare has 10 general categories:

> Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport

After mapping to major categories, nine of them were found in analysis dataset – there were no Event category entries identified with following total counts:



```
City                              AnklamBaltijskCopenhagenDarłowoElblągFlensburg...
Venue_Arts & Entertainment                                                      602
Venue_College & University                                                        4
Venue_Food                                                                     2146
Venue_Nightlife Spot                                                            524
Venue_Outdoors & Recreation                                                    1074
Venue_Professional & Other Places                                                62
Venue_Residence                                                                   2
Venue_Shop & Service                                                           1417
Venue_Travel & Transport                                                        505
dtype: object
```

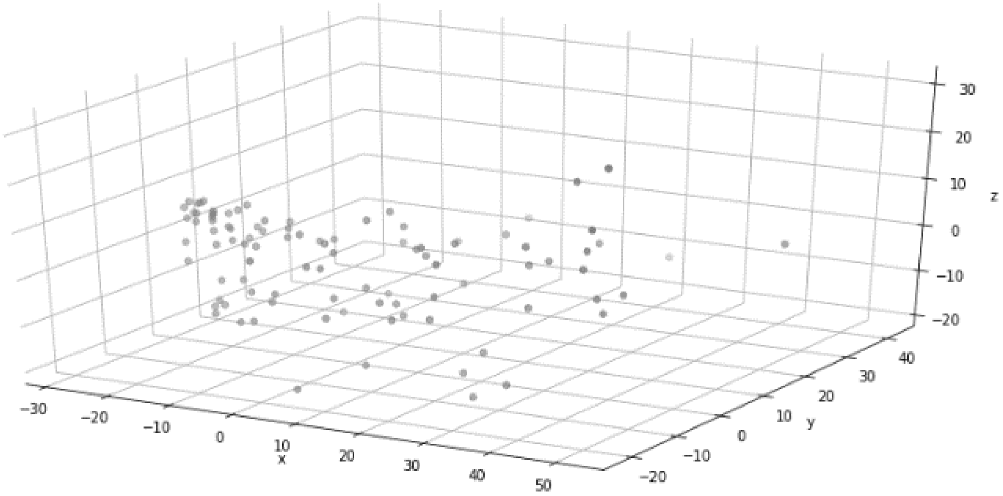
Table 2. Count of different Venue parent categories.

3. Methodology

In methodology section i will describe machine learning algorithm chosen for modelling in analysis of cities surrounding Baltic Sea as well as go through other statistics applied on related data. As the main purpose of this analysis is to identify differences of these cities, the best suitable model for it is K-Means clustering. This is unsupervised machine learning model, which splits dataset into k non-overlapping clusters. The algorithm finds patterns within data and makes clusters.

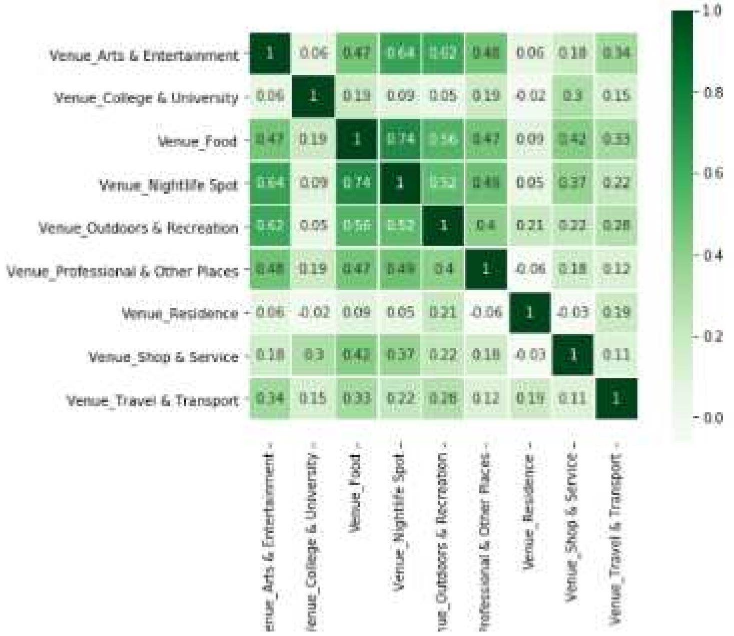
3.1. Setting number of clusters

To set correct number of clusters for multidimensional data is quite complicated task. In order to decide on most appropriate number of clusters, I‘ve decided to analyse data visually and identify the best option of clusters. Plotting multi-dimensional data without transformation is impossible, so Principle Components Analysis (PCA) is used to transform dataset into 3D format.



Picture 2. PCA transformed nonclustered data contribution

Visually it is quite hard to identify number of clusters to use in the model. So it is worth to look into correlation matrix of the data.



Picture 3. Correlation matrix on initial clustering dataset

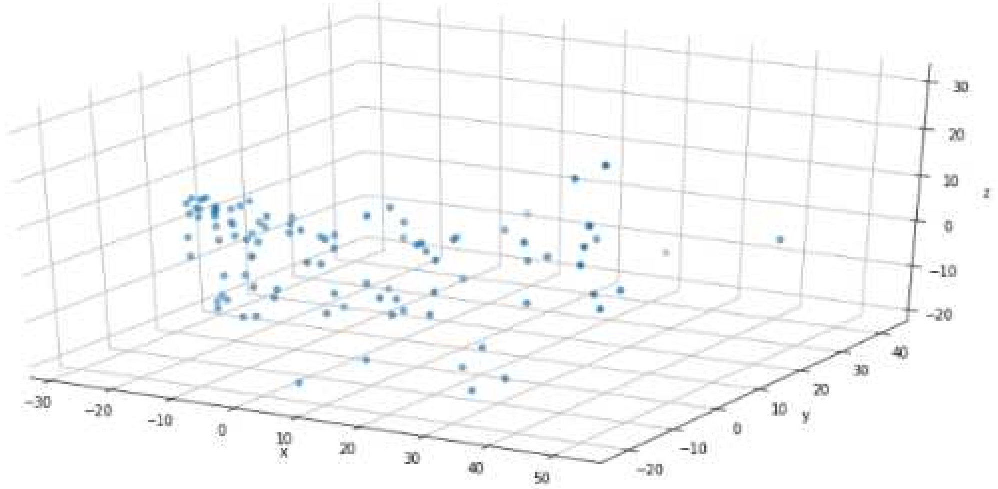
Correlation matrix shows that correlations are very weak in Residence and College & University categories. Looking into summary table below, it can be seen that these categories also have very low number of entries, which is why we can simply drop the out.

Parent_cat	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport

Venue	675	3	2108	498	1269	58	1	1302	520

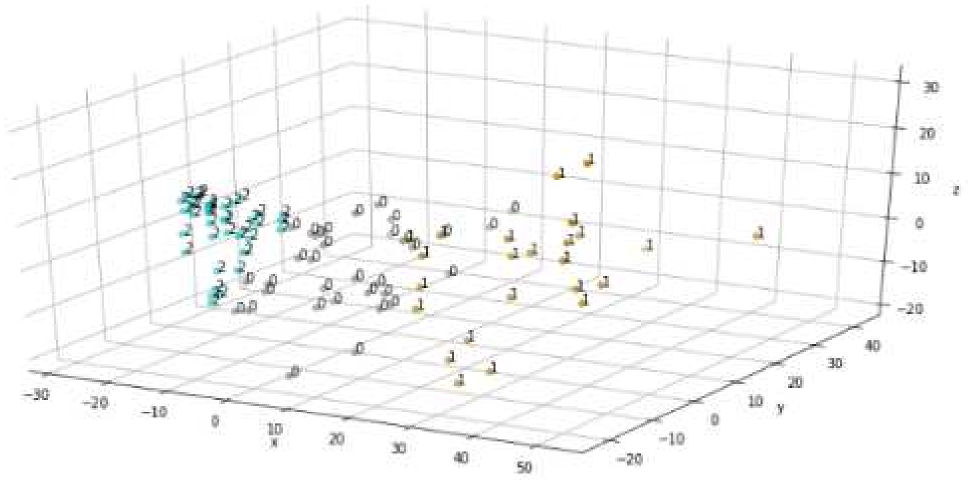
Table 3. count of total venues in dataset by category

After dropping mentioned columns, overall picture does not change too much. Updated picture presented below almost did not change compared to previous one.



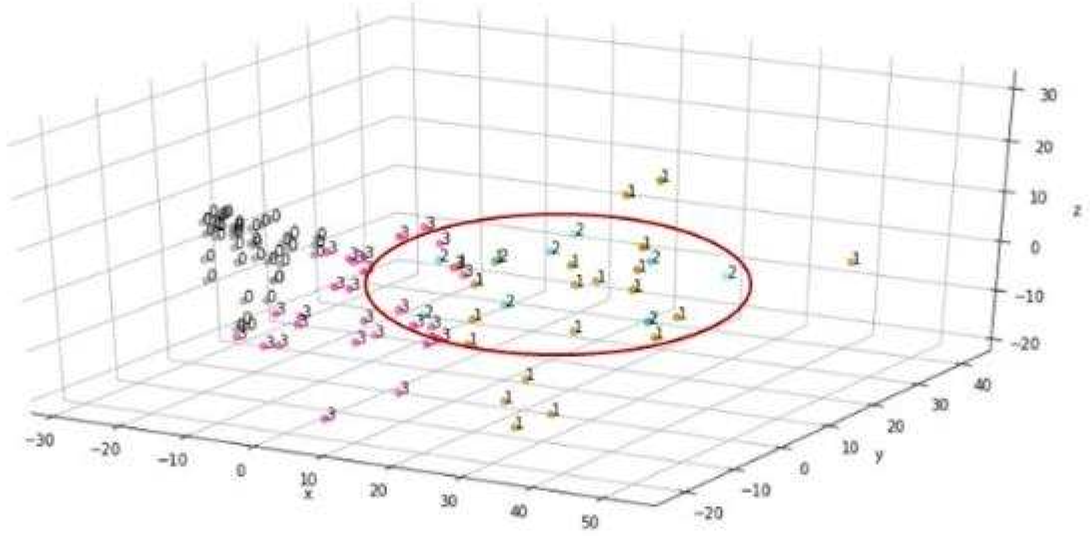
Picture 4. Transformed dataset after dropping low value columns

After initial causal analysis of data exact number of clusters cannot be clarified. So I‘ve decided to start with 3 clusters. The picture below represent an outcome of clustering. Cluster 2 seems to be quite good. While cluster 0 and 1 are quite scattered.



Picture 5. PCA restructured data clustered in 3 clusters

In order to search for other clustering options, higher number of clusters was examined. Though, applying already 4 clusters seem to create mixed view at least in reconstructed data as it can be seen in picture below. Despite the fact, that k-means clustering create non-overlapping clusters, still picture is suggesting an opinion that clustered cities parameters might overlap and this would create a difficulty to describe each cluster specifics.

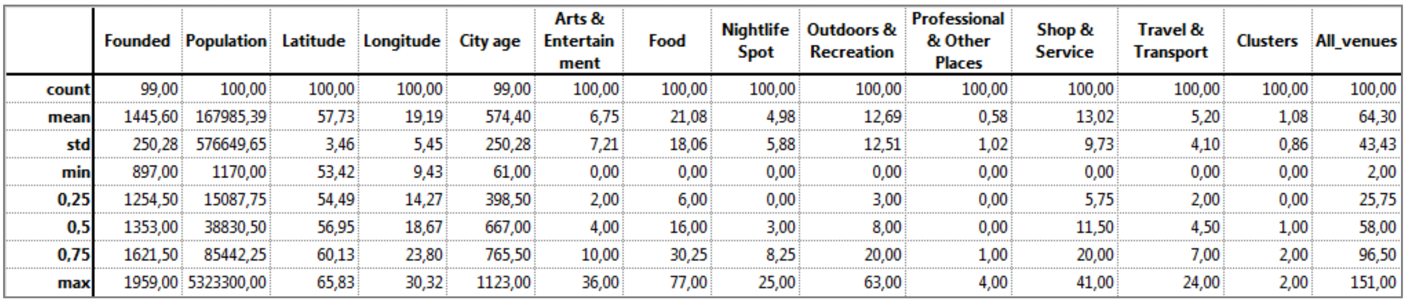


Picture 6. 4 level clustering result

As a result, 3 level clustering was selected for modelling as it seems to be providing the best representation in this case.

3.2. Statistical analysis

In data preparation phase, some exploratory analysis was already covered. Main types of analysis used were descriptive statistics, describing data set means, standard deviations, count, min, max and similar aggregates. In some cases I was looking into shapes of datasets to follow the changes in dataset length.



	Founded	Population	Latitude	Longitude	City age	Arts & Entertain ment	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport	Clusters	All_venues
count	99,00	100,00	100,00	100,00	99,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
mean	1445,60	167985,39	57,73	19,19	574,40	6,75	21,08	4,98	12,69	0,58	13,02	5,20	1,08	64,30
std	250,28	576649,65	3,46	5,45	250,28	7,21	18,06	5,88	12,51	1,02	9,73	4,10	0,86	43,43
min	897,00	1170,00	53,42	9,43	61,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,00
0,25	1254,50	15087,75	54,49	14,27	398,50	2,00	6,00	0,00	3,00	0,00	5,75	2,00	0,00	25,75
0,5	1353,00	38830,50	56,95	18,67	667,00	4,00	16,00	3,00	8,00	0,00	11,50	4,50	1,00	58,00
0,75	1621,50	85442,25	60,13	23,80	765,50	10,00	30,25	8,25	20,00	1,00	20,00	7,00	2,00	96,50
max	1959,00	5323300,00	65,83	30,32	1123,00	36,00	77,00	25,00	63,00	4,00	41,00	24,00	2,00	151,00

Table 4. Descriptive statistics on clustered dataset

For summarization reasons I was using pivot tables which provide an opportunity to summarize data in easy way. Pivot table was also used for creating main analysis dataset. I had to transpose venues data within the dataset and make aggregated count function on it to be able to use for clustering. Pivot was the easiest option to reach the result.

Quite extensive part of analysis is based on visual analysis. Key visual representation was made by representing cities and afterwards same cities clustered in world map. Though also regular charts were used both in understanding data area as well as in analysing results. All outcomes will be discussed wider in results section.

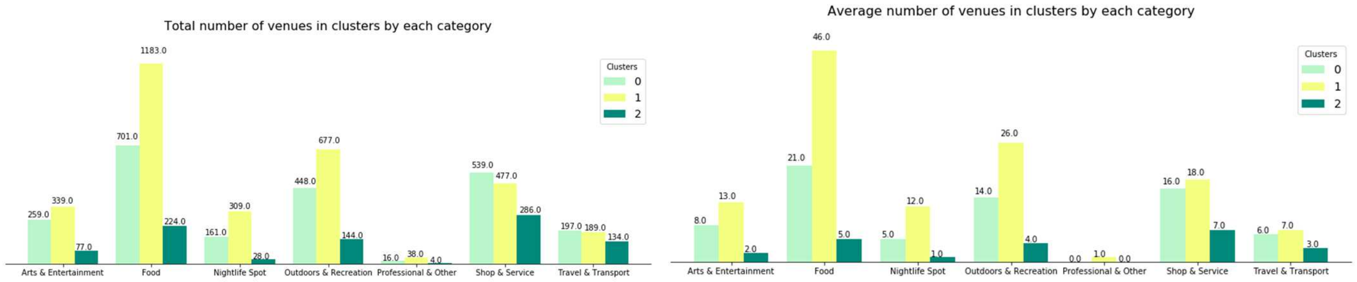
4. Results

In results section I will cover analysis outcomes. To begin with, let's review the clusters. Based on table provided below, we can say that clusters are separated by number of venues in the city. Both sum and mean show the same trend – number of venues is the highest in cluster 1 and the lowest in cluster 2. Let's call those cities the biggest and smallest respectively in terms of venues number. By comparing the size of clusters, we can note, that cluster 2 (combining the smallest cities) is the largest one. Cluster 1, combining the largest cities has the lowest number of cities within.

Clusters	All venues mean within cluster	All venues count in cluster	Cities count in cluster
0	70.33	2321.0	33
1	123.54	3212.0	26
2	21.88	897.0	41

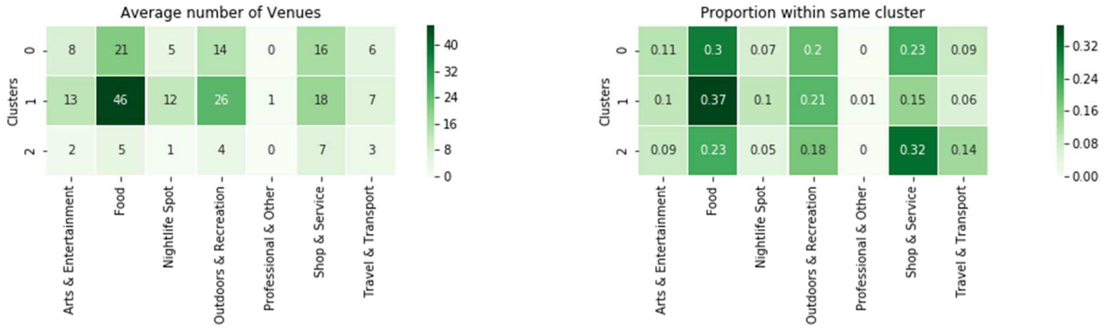
Table 5. Overall summary of clustered data

Plots below visually show separation between venue categories and clusters. As proposed by aggregates, the picture support the idea of clustering based on total number of venues. Cluster 1 has the highest columns in all categories (based on averages) and other clusters follow the same logic.



Picture 7. Total and average number of venues by cluster, by category.

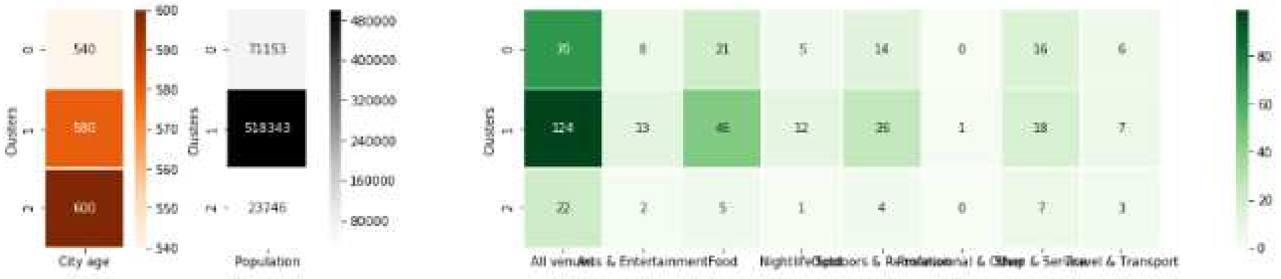
In order to look deeper into clusters and to find out whether there are any other description we could make on cities clustered together, let's see how contribution looks in each cluster. The left heatmap represent average count of venues per city in cluster identifying, how many venues of each category you should expect to find in cities clustered in one or another cluster. While the right heatmap shows the proportion of venues categories within the cluster. By looking into it we can state, that smallest cities, clustered under cluster 2 are more focused on shops versus restaurants. So if you would be travelling in such cities, most probably you might consider cooking your own meal as Shop&service category include also food shops, while Food category holds restaurants, cafes and similar type of venues within.



Picture 8. Within cluster contribution by categories

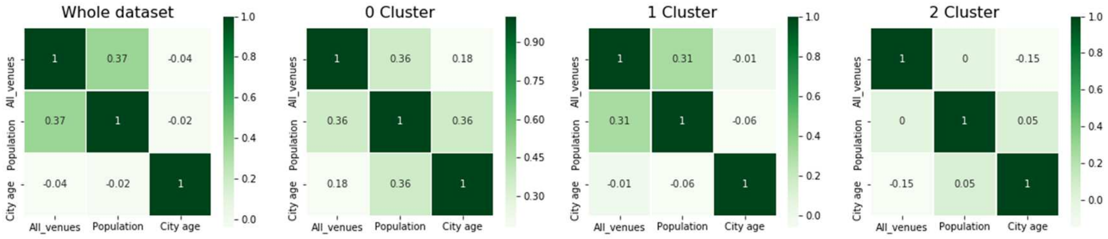
Initial dataset had more numeric variables, which were not included into clustering. These were cities parameters – population and age (calculated from foundation year). During result analysis it is worth to check whether these parameters have any relation towards venues information. Heatmaps below are based on average metrics per city. City ages are quite comparable for all cities and the widest gap between cluster averages is just 60 years. Any relation cannot be seen as well.

On the other hand, population seems to have some relationship towards clusters as the highest average population can be found in cluster 1 and the lowest on in cluster 2. Therefore correlation should be examined to confirm or deny this hypothesis.



Picture 9. Clusters comparison to out-of-scope variables - population and age of the city

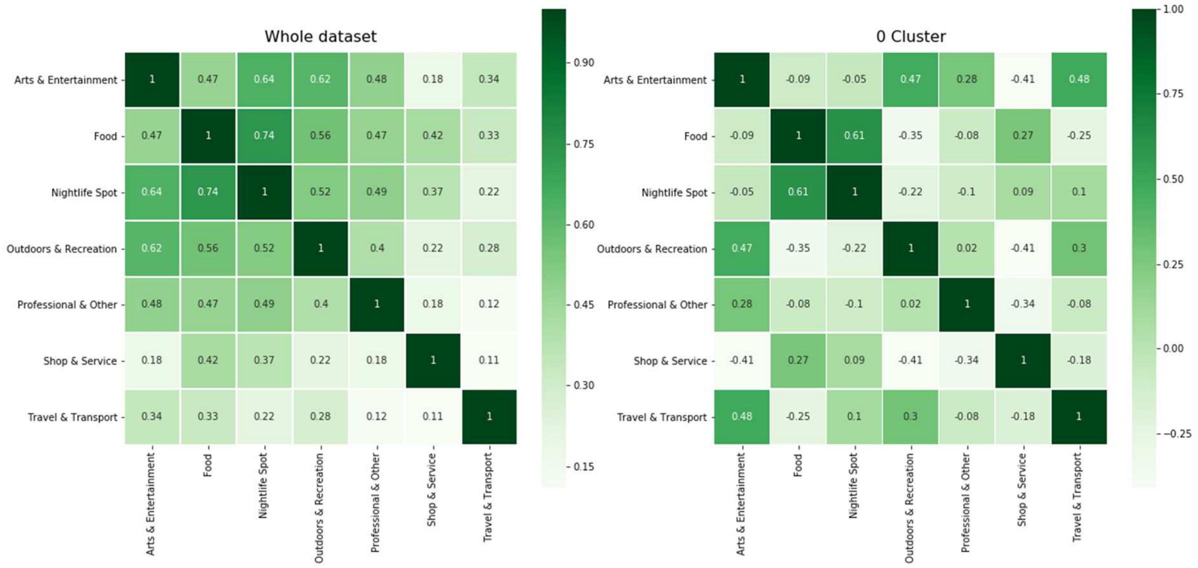
As clusters are based on size logic, for correlation total venues are taken. As it can be seen in tables below, the correlation is very low with these out-of-scope metrics. So despite of logic supporting picture on aggregates view, we should deny the hypothesis on population relation to clustering outcome.



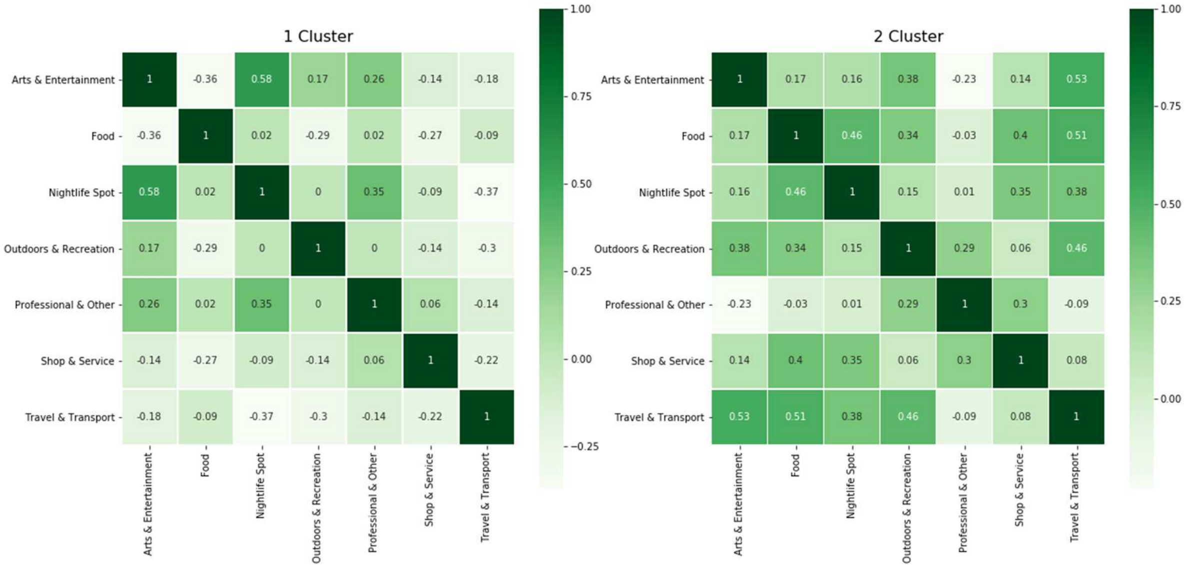
Picture 10. Correlation between out-of-scope parameters

The other option to specify clusters is to look into inner correlation. Tables below represent within cluster correlation between different types of venues.

If you are interested into Nightlife, you should choose 0 clustered cities as here the correlation with Food venues, which are the most common in all clusters, is the highest. Cluster 1 in general has very week relationship with it while cluster 2 seems to be much more correlated. As a result, the expected outcome in cities within 2 cluster, should be the most correct.

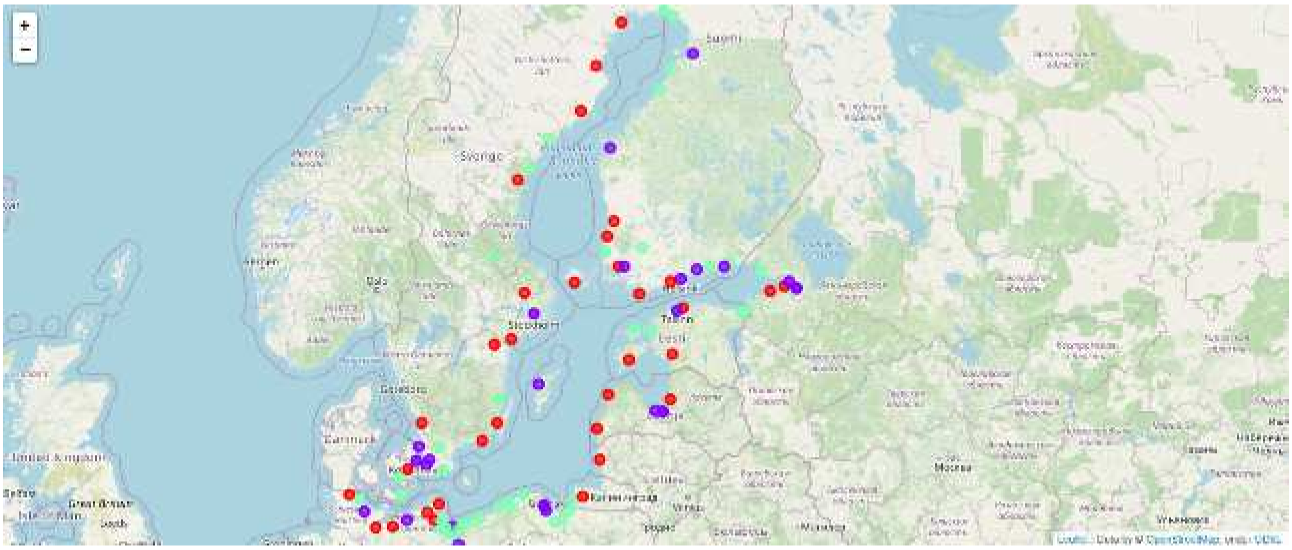


Picture 11. Inner correlations whole dataset and cluster 0



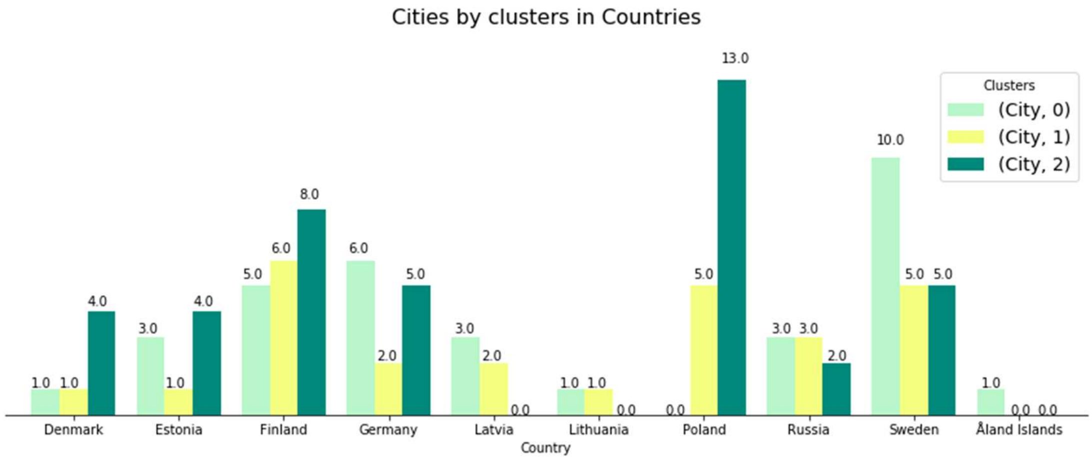
Picture 12. Inner correlation cluster 1 and 2

Now we can look into distribution in map. We can see that cities of different clusters are quite distributed. Just by glancing though, we can see that violet cities (cluster 1) are usually capitals of the countries or just big cities.



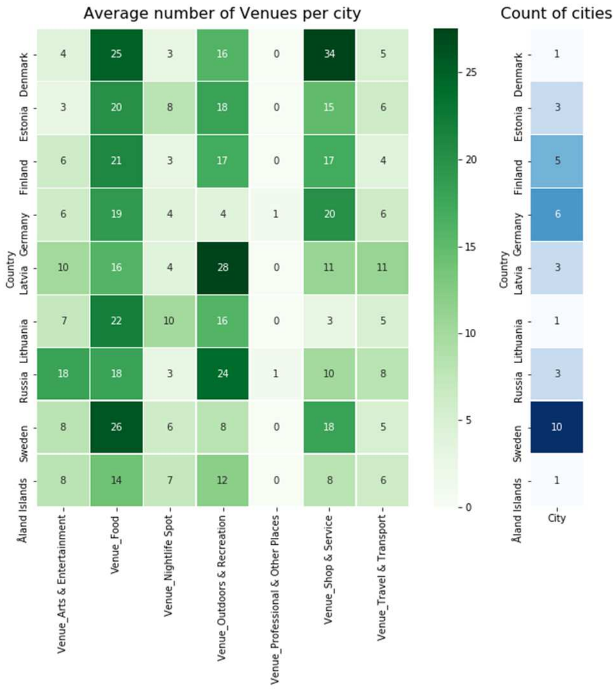
Picture 13. Colours by clusters: red – 0; violet – 1; green - 2

After analysing clusters and cities, let’s look into countries. The chart below shows distribution of cities around the Baltic Sea within countries they belong to in term of clusters. i.e. if you prefer to visit only one country and travel along the coast you might consider visiting Poland for 2 clustered cities, where infrastructure is the least extensive. If you prefer to see all kind of cities in one country, you should choose Finland coast. And if you prefer visiting more extensive cities, you might consider going to Sweden, Latvia or Lithuania depending on number of cities you would like to visit.



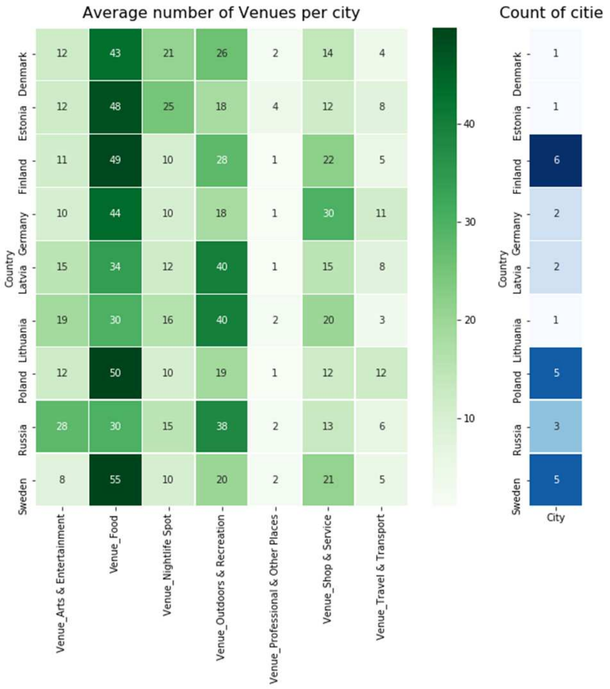
Picture 14. Number of cities in each country by cluster

Let’s look into country specifics in each cluster more specifically. Cluster 0 combining mid-size cities contain cities in 9 countries: Denmark, Estonia, Finland, Germany, Latvia, Lithuania, Russia, Sweden and Aland Islands. Sweden has even 10 cities of this clusters along the coast and those cities on average are the most equipped by Food and Shop&Service venues. Germany with 6 cities suggest similar infrastructure as well, while Finland with 5 cities in cluster 0, suggest also more Outdoor&recreation venues. If you are interested in Outdoor&recreation venues you should also consider visiting Latvia, Russia or Estonia as well as these countries in the cluster have the most of such activities.



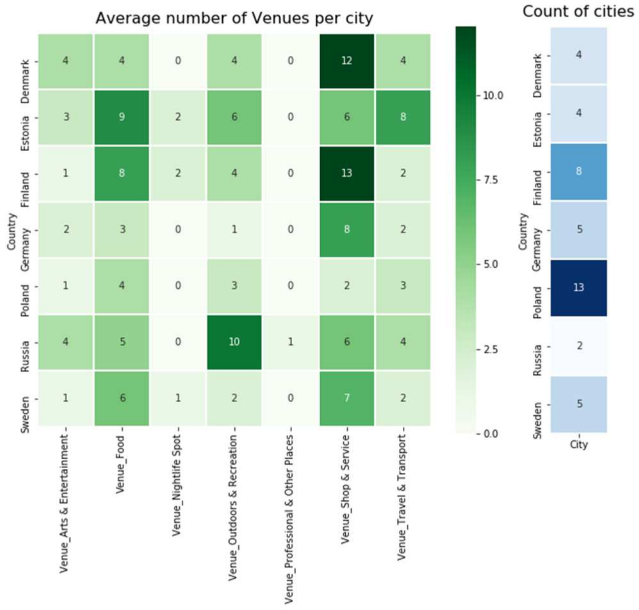
Picture 15. Cluster 0 by country

Cluster 1, which hold the biggest cities in it, contain cities in nine countries as well. Denmark, Estonia, Finland, Germany, Latvia, Lithuania, Russia, Sweden as in cluster 0 and Poland. By visiting these type of cities, you should visit Finland, Poland or Sweden for wide selection of Food venues and Latvia or Russia for Outdoor and Recreation activities. Cities within Germany would suggest you more shops and services.



Picture 16. Cluster 1 by country

Last cluster 2, combining smallest cities has low number of all types of venues, though, it can be reconfirmed, that Shop and Service venues dominate in this cluster, especially in Denmark and Finland. Even 13 cities from Poland lie under this cluster and these have very low number of any type of venues. These cities might be really small town.



Picture 17. cluster 2 by country

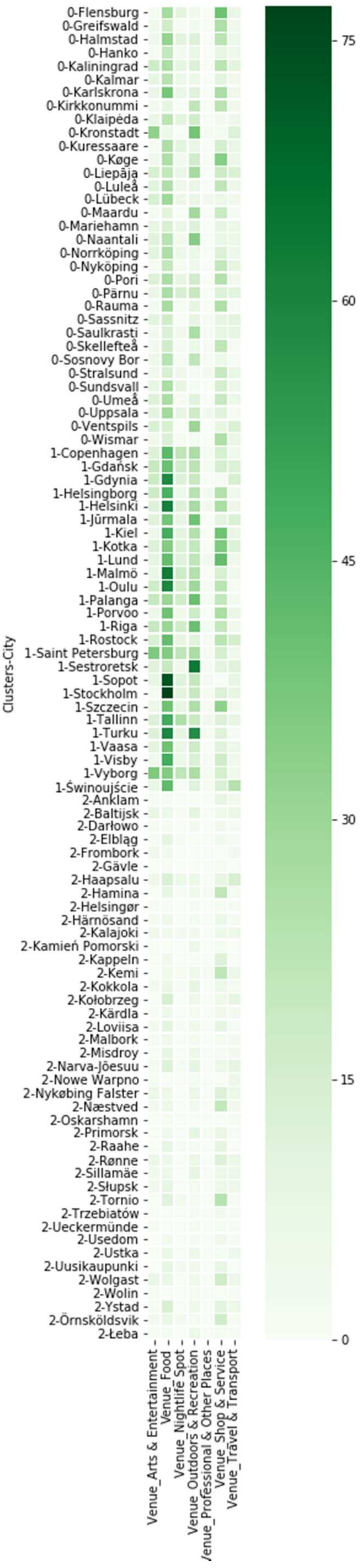
5. Discussion

In literature, you may find cities worth to visit along the Baltic Sea. i.e. U.S.News[i] suggest to visit Gdansk in Poland, Tallinn in Estonia and St. Petersburg in Russia. As you can see in the picture on the right hand side, all these three cities are clustered under cluster 1. As a result, if you have visited those cities and you liked them, you might consider visiting also other cities in the same cluster.

AB Poland travel[ii] suggest to visit Gdynia and Sopot together with Gdansk. All three cities are also clustered together in the same cluster in our analysis. Same article refers to Kolobrzeg as a resort worth to visit in Poland. This city lies under cluster 2 in our analysis, identifying that you might find calm and not overloaded rest there.

ILP (International Language Programs)[iii] also review Baltic countries and provide travelling guidance. Here we will find Tallinn in Estonia, Riga in Latvia and Klaipeda in Lithuania. Also some other Lithuanian cities are mentioned, but those are not along the Baltic Sea. So both Tallinn and Riga which are also capitals are clustered together in our analysis. Klaipeda is under cluster number 1 which also is supported by article, as it is referred as „cute little town".

After comparing to literature outcomes, we might agree, that clustering analysis came off well and the outcome might be used for recommendations depending on what kind of holiday traveller wants – explorers might consider cluster 1 cities, while the ones having calm rest preference should glance through clusters 0 or 2 for their holiday plans.



Picture 18. List of cities and venues by cluster

5.1. Limitations and future directions

The project definitely had its limitations. The most important limitation was max number of venues given per city by Foursquare, which was equal to 100 and had an impact on biggest cities. This limitation would not have affected the list of biggest cities as well as clustering outcome, but might have changed the inner view of venue types in those cities, because different combination might have been seen if whole picture would have been represented. As a result, to overcome this limitations, more points within cities have to be taken. Though, each country has its own divisions within cities and getting such an extensive list for all 100 cities, might have been quite challenging. As a result, extensive analysis might have been carried out on some selected cities instead of looking into whole set.

Other important limitation is initial source of the data – list of cities provided in Wikipedia article. It could be that not all cities were covered in the list or some of them might have been divided into regions of cities. Though by looking into map, we can ensure, that there are no skipped areas and distribution seems quite even. So this limitation should not have very big impact.

[i] https://travel.usnews.com/features/3-cant-miss-cities-along-the-baltic-sea
[ii] https://abpoland.com/top-6-places-to-visit-at-the-baltic-sea/
[iii] http://blog.ilp.org/top-cities-to-travel-to-in-the-baltic-states

In future analysis it would be really interesting to add tourism statistics as well as size of the cities in square meters to see if there is any relation between venues in cities and mentioned statistics. This might give also business view into it and might help to identify places, where it would be worth to open one or another type of venue.

6. Conclusion

The main objective of this project was to analyse cities located by the coast of Baltic Sea in order to provide information to travellers how to choose preferred city. We ended up with three clusters combining cities by their size. Cluster 0 contained the most venues equipped cities, while cluster 2 contained the least venue filled cities. Cluster 1 landed in the middle of those two. Cluster 0 holds the biggest variety of venues as a result, might be more crowded compared to cities in other clusters. Cluster 1 has slight preference on Outdoor and Recreation venues while Cluster 2 holding the smallest cities within, has some preference on Shop and Service venues so are suitable for more autonomous and calm rest.

To summarize, Baltic Sea region seem to be quite diverse by number of countries, by type of cities and everyone might find the place they would like there.