SERVICE BASED FEATURES ANALYSIS OF CITIES SURROUNDING BALTIC SEA

APPLIED DATA SCIENCE CAPSTONE PROJECT

BY ALINA ŠERIENĖ

2020.05.31

EXECUTIVE SUMMARY

- THE **BALTIC SEA** IS A MEDITERRANEAN SEA OF THE ATLANTIC OCEAN ENCLOSED BY CITIES FROM **9 DIFFERENT COUNTRIES**. COVID-19 AS WELL AS GLOBAL WARMING MAY INCREASE INTEREST IN THIS AREA.
- USING **K-MEANS CLUSTERING** ALGORITHM, CITIES WERE DIVIDED IN **3 CLUSTERS** BASED ON VISUAL ANALYSIS AND COMPOSITION PARAMETERS OF EACH CLUSTER:

Clusters	0	1	2
All venues mean within cluster	70.33	123.54	21.88
All venues count in cluster	2321.0	3212.0	897.0
Cities count in cluster	33	26	41
Summary	Average sized cities with orientation towards outdoor & recreation activities and sufficient supply of other type of venues	Biggest cities with highest diversity of venues. Choose for wide selection of food venues or if you have preference on nightlife spots	Small cities with low number of venues and focus on shop & service venues

1. INTRODUCTION

BALTIC SEA

- THE BALTIC SEA IS A MEDITERRANEAN SEA OF THE ATLANTIC OCEAN, **ENCLOSED BY NINE DIFFERENT COUNTRIES** - DENMARK, ESTONIA, FINLAND, LATVIA, LITHUANIA, SWEDEN, NORTHEAST GERMANY, POLAND AND RUSSIA.
- DUE TO IT'S COLD CLIMATE BALTIC SEA IS RARELY CHOSEN AS HOLIDAY DESTINATION, BUT **COVID-19 AND GLOBAL WARMING** MIGHT SOON CHANGE THE PICTURE. AS A RESULT,
- **RESEARCH OBJECTIVE** – IDENTIFY DIFFERENCES BETWEEN BALTIC SEA SURROUNDING CITIES AND MAKE A TRAVELLING ADVISE BASED ON ANALYSIS
- **AUDIENCE** – ANYONE INTERESTED IN TRAVELLING IN THE AREA

MAIN DATA SOURCES

- LIST 100 OF CITIES SURROUNDING BALTIC SEA FROM WIKIPEDIA ARTICLE. DATA EXTRACT:

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	1236	12,686	53°51'N 13°41'E
Baltijsk	Russia	1725	32,667	54°39'N 19°55'E
Copenhagen	Denmark	1254	1,295,686	55°40'N 12°34'E
Darłowo	Poland	1310	14,911	54°26'N 16°25'E
Elbląg	Poland	1246	124,257	54°9'N 19°24'E
Flensburg	Germany	1284	87,402	54°46'N 09°26'E
Frombork	Poland	1310	2,416	54°21'N 19°41'E
Gdańsk	Poland	1263	468,754	54°21'N 18°38'E
Gdynia	Poland	1926	247,799	54°30'N 18°32'E

- FOURSQUARE VENUES INFORMATION PER CITY WITH RADIUS OF 10 KM.
- FOURSQUARE DATA GROUPED TILL THE HIGHEST LEVEL CATEGORIES.
- FINAL MAJOR CATEGORIES COUNT IN DATASET:

```
City                                  AnklamBaltijskCopenhagenDarłowoElblągFlensburg...
Venue_Arts & Entertainment                                                          635
Venue_College & University                                                            3
Venue_Food                                                                         2292
Venue_Nightlife Spot                                                                503
Venue_Outdoors & Recreation                                                        1291
Venue_Professional & Other Places                                                    59
Venue_Residence                                                                       2
Venue_Shop & Service                                                               1343
Venue_Travel & Transport                                                            568
dtype: object
```

CHOOSING CORRECT COORDINATES

2. METHODOLOGY AND MODELLING

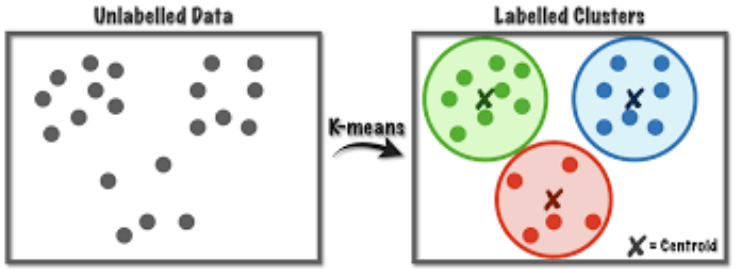
METHODOLOGY

MAIN MODEL APPLIED:

- **K-MEANS CLUSTERING:**
  - MODEL SPLITS DATA INTO NON-OVERLAPPING CLUSTERS WITH STRONG RELATION WITHIN CLUSTER AND WEEK RELATION BETWEEN CLUSTERS.
  - MODEL WORKS ON NON-LABELED DATA AND IT IS UNSUPERVISED MACHINE LEARNING METHOD.

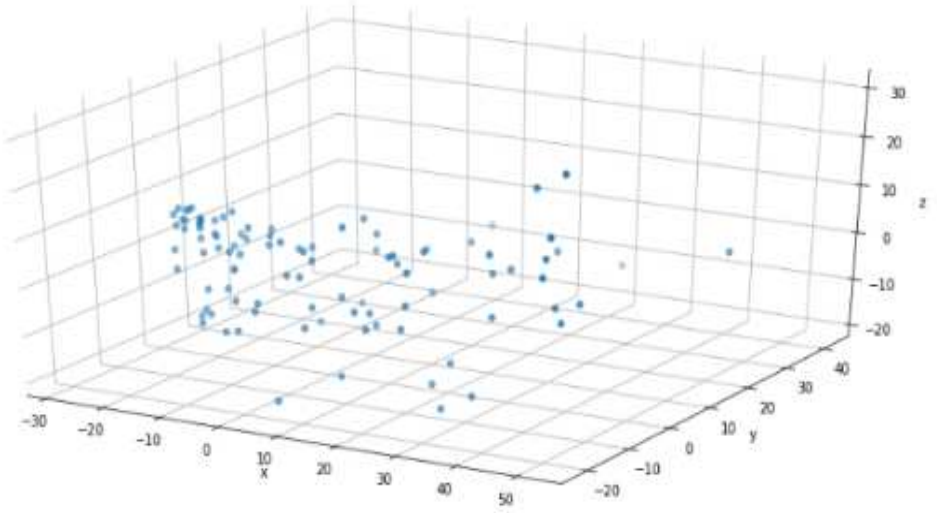
OTHER METHODS APPLIED:

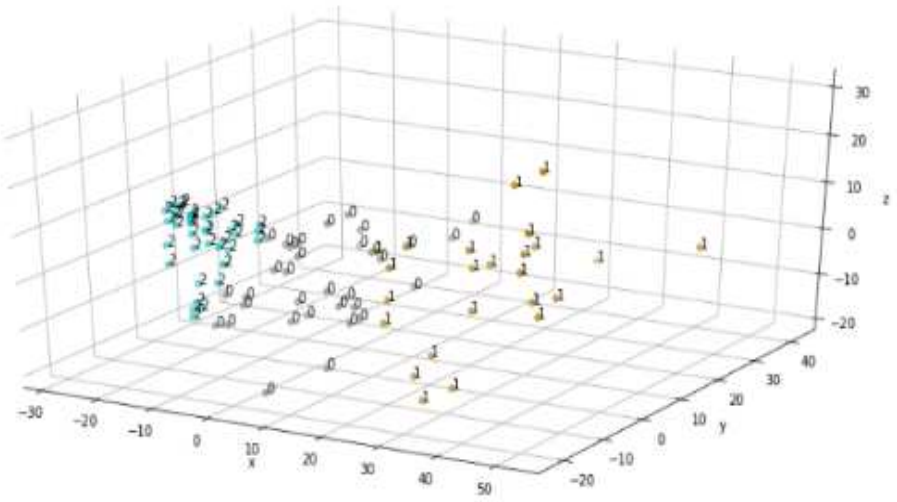
- CORRELATION
- DESCRIPTIVE STATISTICS
- VISUAL ANALYSIS
- PRINCIPLE COMPONENTS ANALYSIS (PCA)



Picture source: https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c

TRANSFORMED DATASET DOES NOT SUGGEST NUMBER OF CLUSTERS TO BE APPLIED





CLEAR NUMBER OF CLUSTERS CANNOT BE IDENTIFIED FROM INITIAL VISUAL ANALYSIS AFTER DATA IS BEING TRANSFORMED INTO 3D USING PCA METHOD.

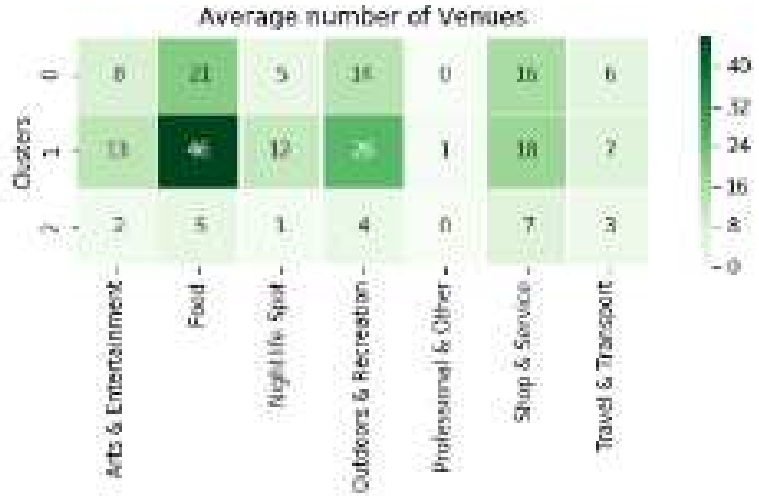
AFTER SEVERAL TRIALS OF DIFFERENT CLUSTERS IT WAS DECIDED TO STAY WITH **3 CLUSTERS** AS TRANSFORMED DATA WAS NOT VISUALLY OVERLAPPING AND NUMBER OF OBSERVATIONS IN EACH WAS COMPARABLE AND SOUND.

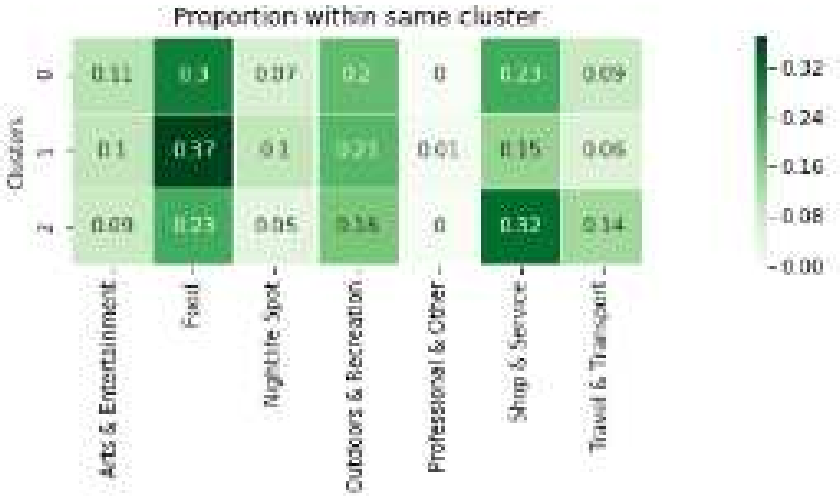
ANALYSIS AND RESULTS

CLUSTERS DIVIDED BY NUMBER OF VENUES

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

- CLUSTERS ARE DIFFERENCED BY SIZE WHERE CLUSTER 1 HOLDS BIGGEST CITIES WITH ~ 124 VENUES PER CITY AND CLUSTER 2 HOLDS THE SMALLEST ONES WITH ~22 VENUES.
- CLUSTER 1 IS MOST ORIENTED TOWARDS FOOD VENUES, WHILE CLUSTER 2 ALONG SHOP & SERVICE

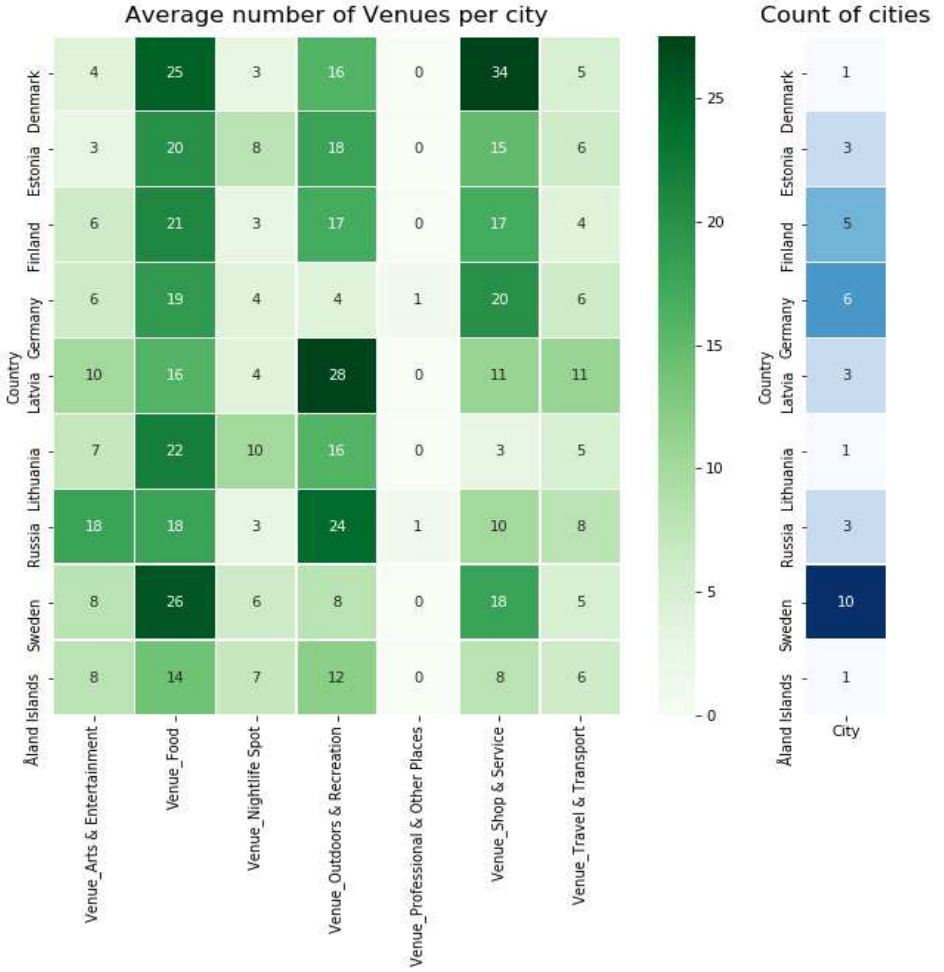




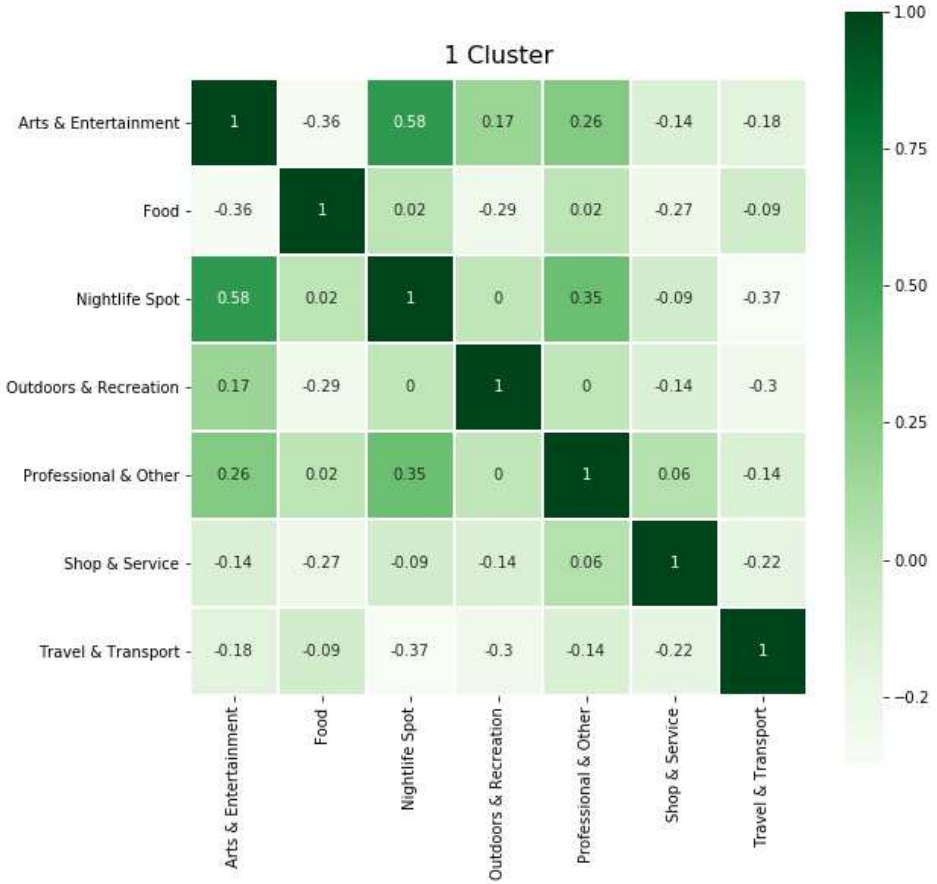
CITIES OF DIFFERENT CLUSTERS DISTRIBUTED WIDELY ALONG THE COAST

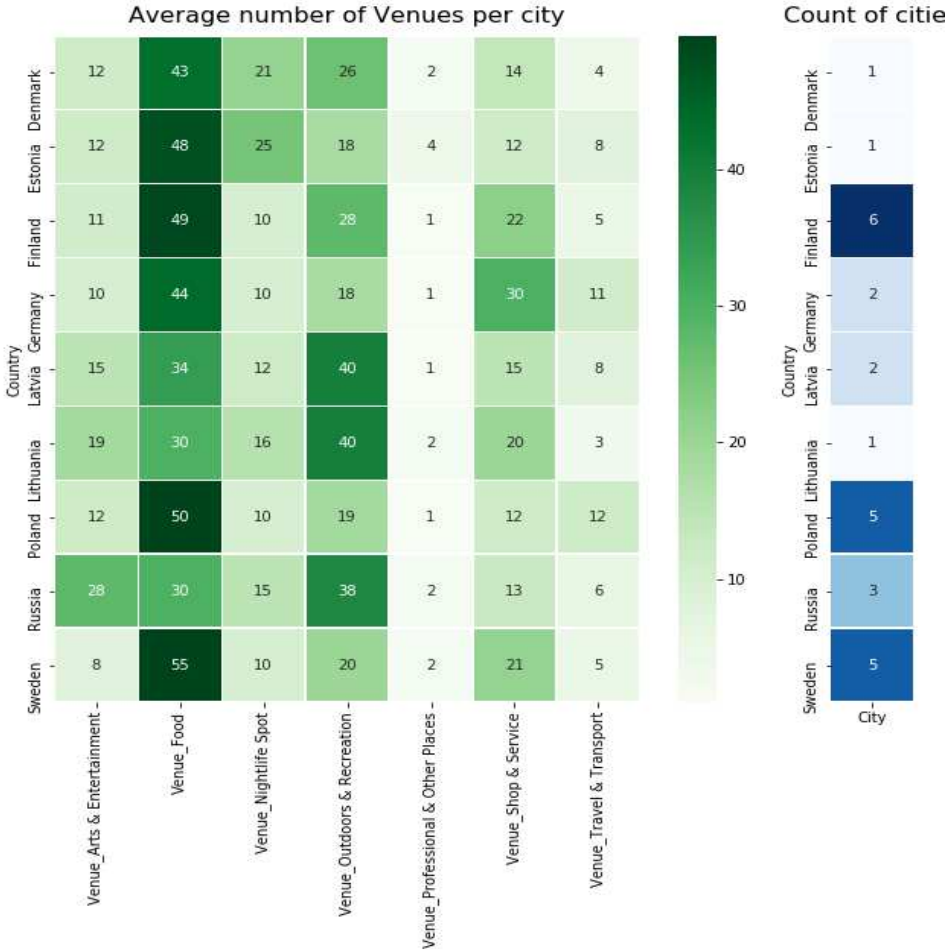
- CLUSTER 1 HOLDS MAJORITY OF CAPITALS OF COUNTRIES SURROUNDING BALTIC SEA UNLESS THOSE COUNTRIES HAVE CAPITALS NOT ALONG THE SEA.
- POLAND HAS A LOT OF CLUSTER 2 CITIES LOCATED IN THE SOUTH OF BALTIC SEA.
- SOUTH COAST IN GENERAL HAS HIGHER CONCENTRATION OF CITIES.

CLUSTER 0: AVERAGE SIZED CITIES WITH ORIENTATION TOWARDS OUTDOOR & RECREATION ACTIVITIES AND SUFFICIENT SUPPLY OF OTHER TYPE OF VENUES

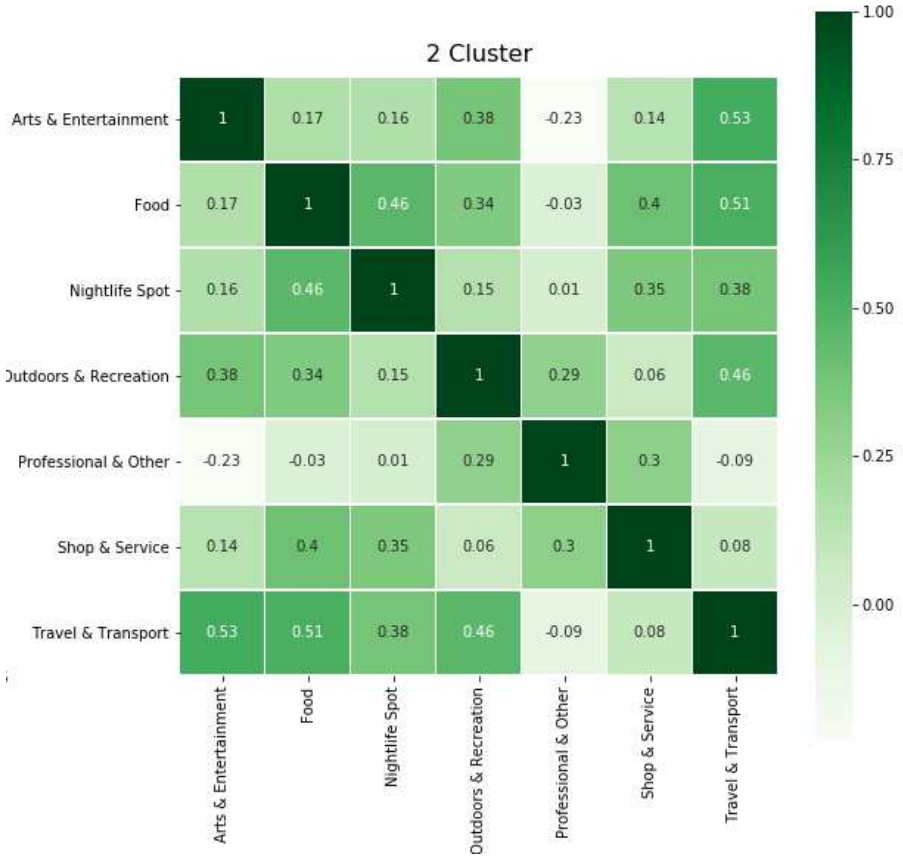


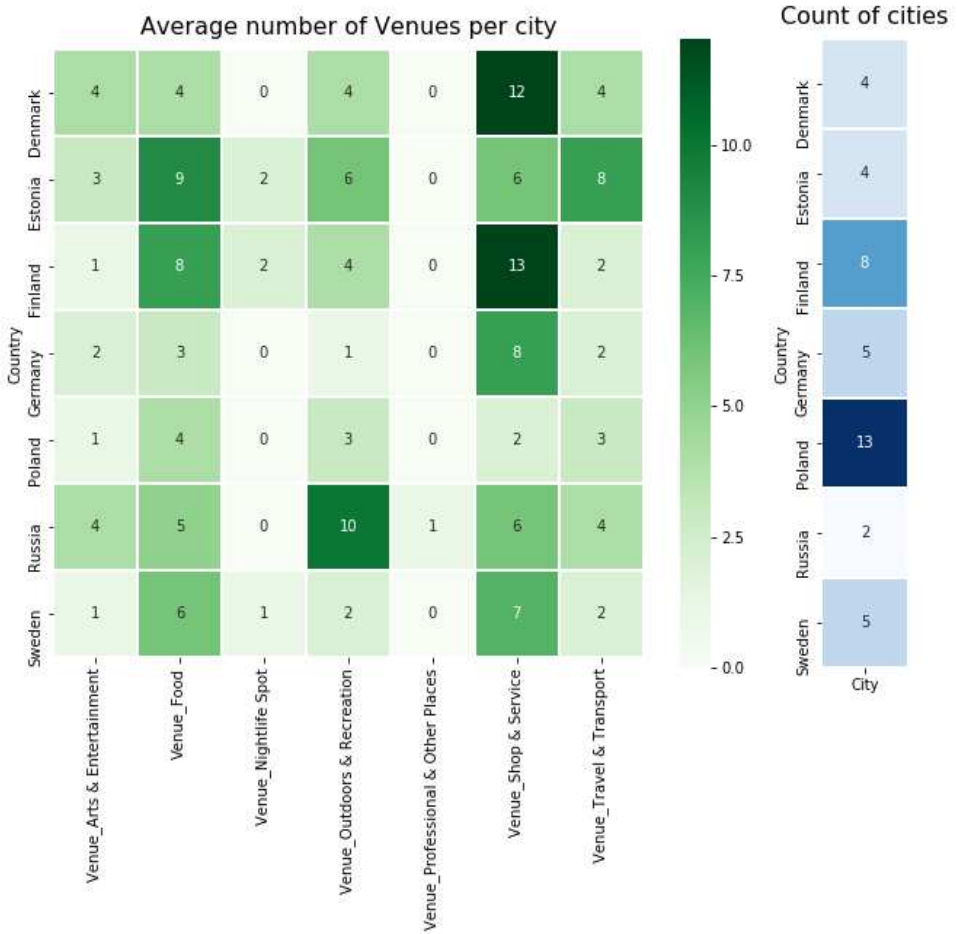
CLUSTER 1: BIGGEST CITIES WITH HIGHEST DIVERSITY OF VENUES. CHOOSE FOR WIDE SELECTION OF FOOD VENUES OR IF YOU HAVE PREFERENCE ON NIGHTLIFE SPOTS





CLUSTER 2: SMALL CITIES WITH LOW NUMBER OF VENUES AND FOCUS ON SHOP & SERVICE VENUES





CONCLUSION AND FUTURE DIRECTIONS

CONCLUSION:

- MODEL HELPED TO CLUSTER THE CITIES AND PROVIDE TRAVELLING RECOMMENDATIONS FOR INTERESTED PARTIES.
- AFTER COMPARING RESULTS VERSUS INTERNET TRAVELLING ADVISE, MENTIONED CITIES LIED UNDER THE SAME CLUSTERS AS EXPECTED SO MODEL RESULTS ARE SUPPORTED ALSO BY LITERATURE.

FUTURE IMPLICATIONS:

- ADDITIONAL DATA ON TOURISM STATISTICS WOULD HELP TO IDENTIFY AREAS FOR BUSINESS TO OPEN NEW VENUES.
- ADDING SIZE OF CITIES WOULD PROVIDE ADDITIONAL INSIGHT ON CORRELATION BETWEEN CITY ACTUAL SIZE AND NUMBER OF VENUES IN IT.
- INVOLVING MORE VENUES PER CITY WOULD PROVIDE BETTER PICTURE ON VENUES COMPOSITION IN LARGEST CITIES.