Countries Cluster Analysis

Clustering Countries by GDP/Capita and by Top Venues using the K-Means method

Table of Content

Introduction

Introduction

⇒This is a simple example of an application of the K-Means clustering method

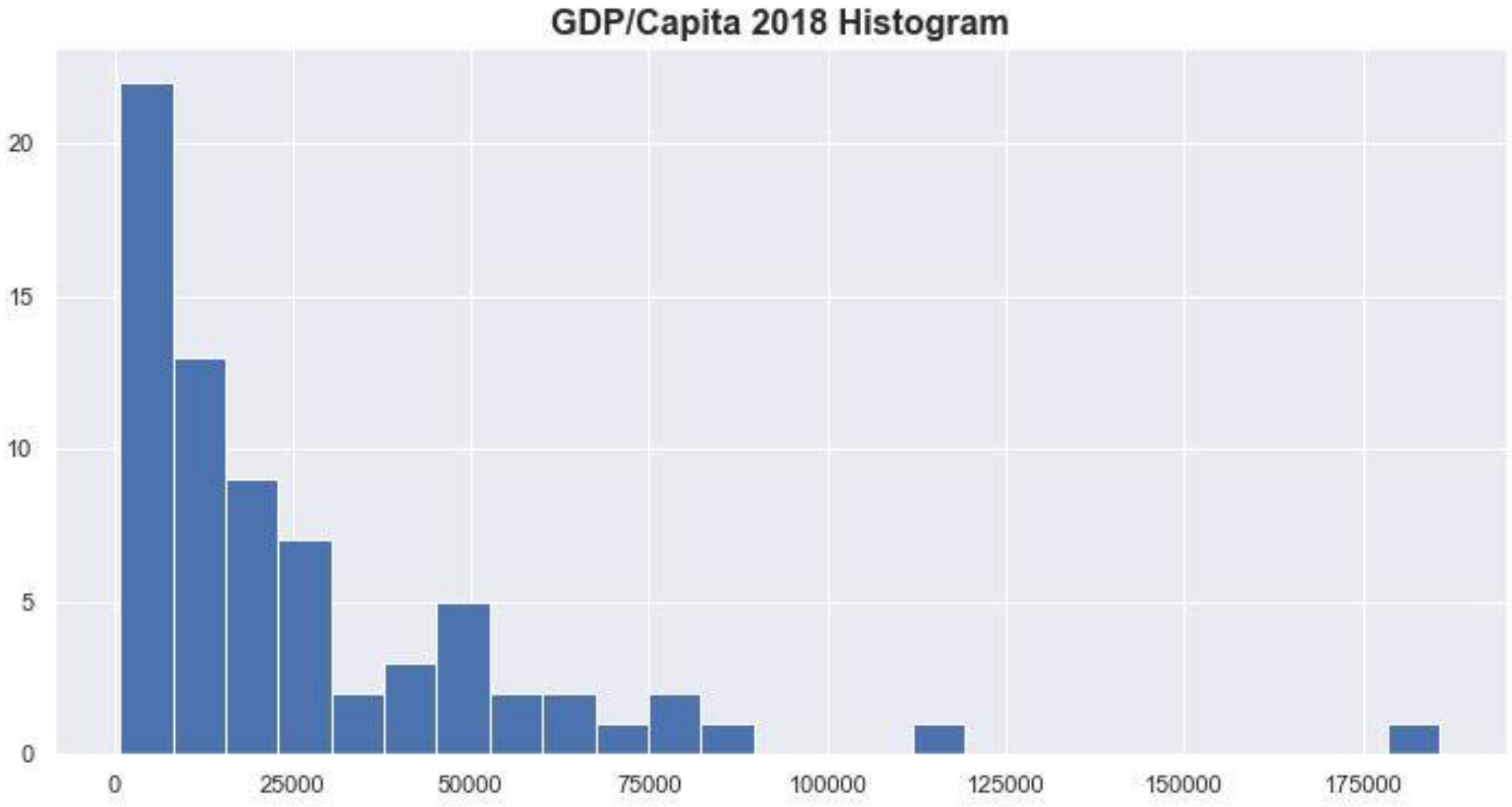
⇒In this analysis countries are clustered into groups

⇒The data set includes countries from a few continents to keep this example simple:

- Europe
- North, Central and South America

⇒Clustering by GDP/Capita from 2018 and by top venues in each country’s capital city

GDP/Capita

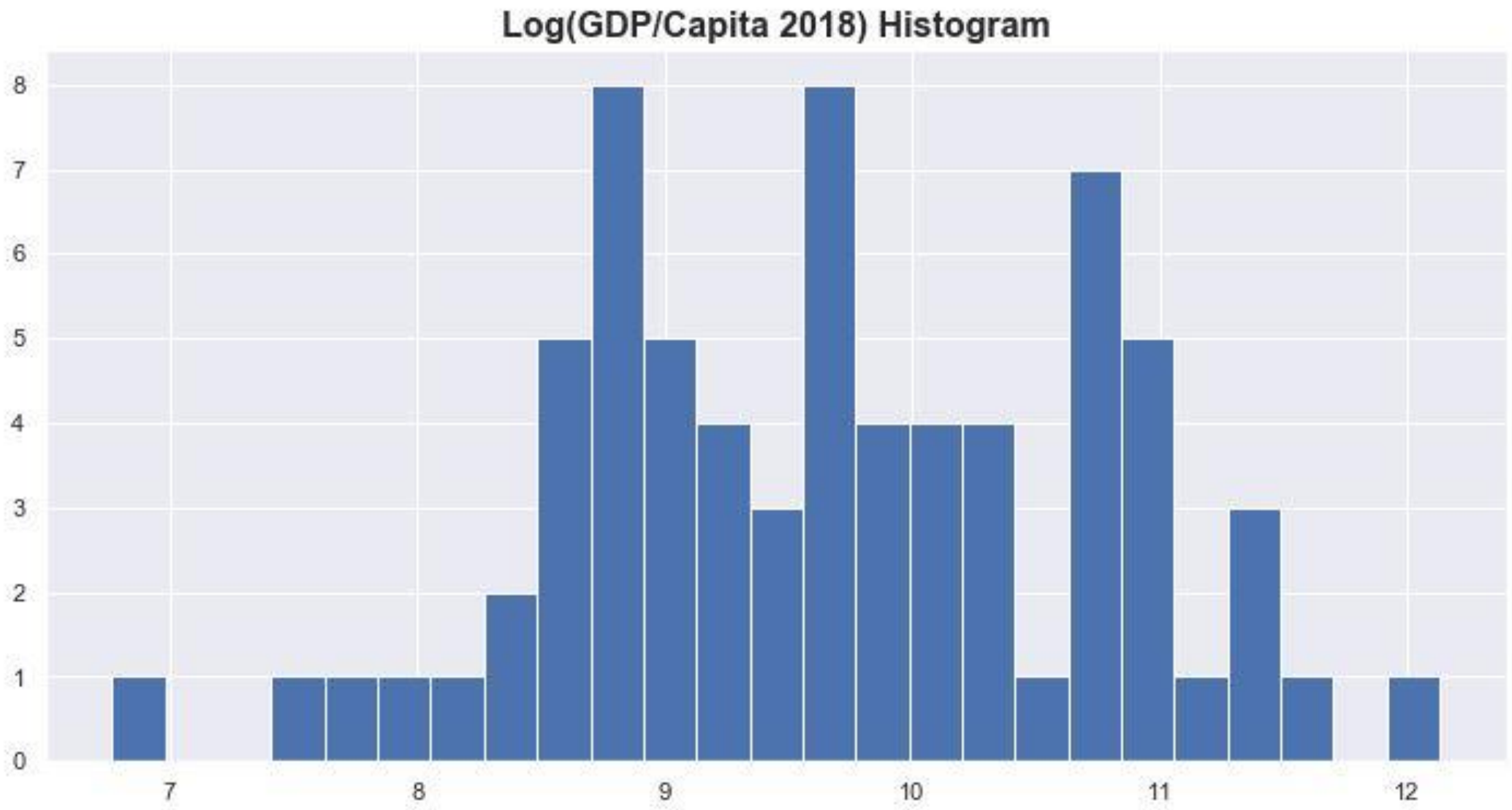

GDP/Capita 2018 Histogram
20
15
10
5
0
0
25000
50000
75000
100000
125000
150000
175000

GDP/Capita Distribution

⇒We can see that GDP/Capita has a logarithmic distribution

⇒A log transformation is required for further processing

Log Transformation


Log(GDP/Capita 2018) Histogram
0
1
2
3
4
5
6
7
8
7
8
9
10
11
12

GDP/Capita Distribution

⇒The distribution of Log GDP/Capita looks more like a normal distribution

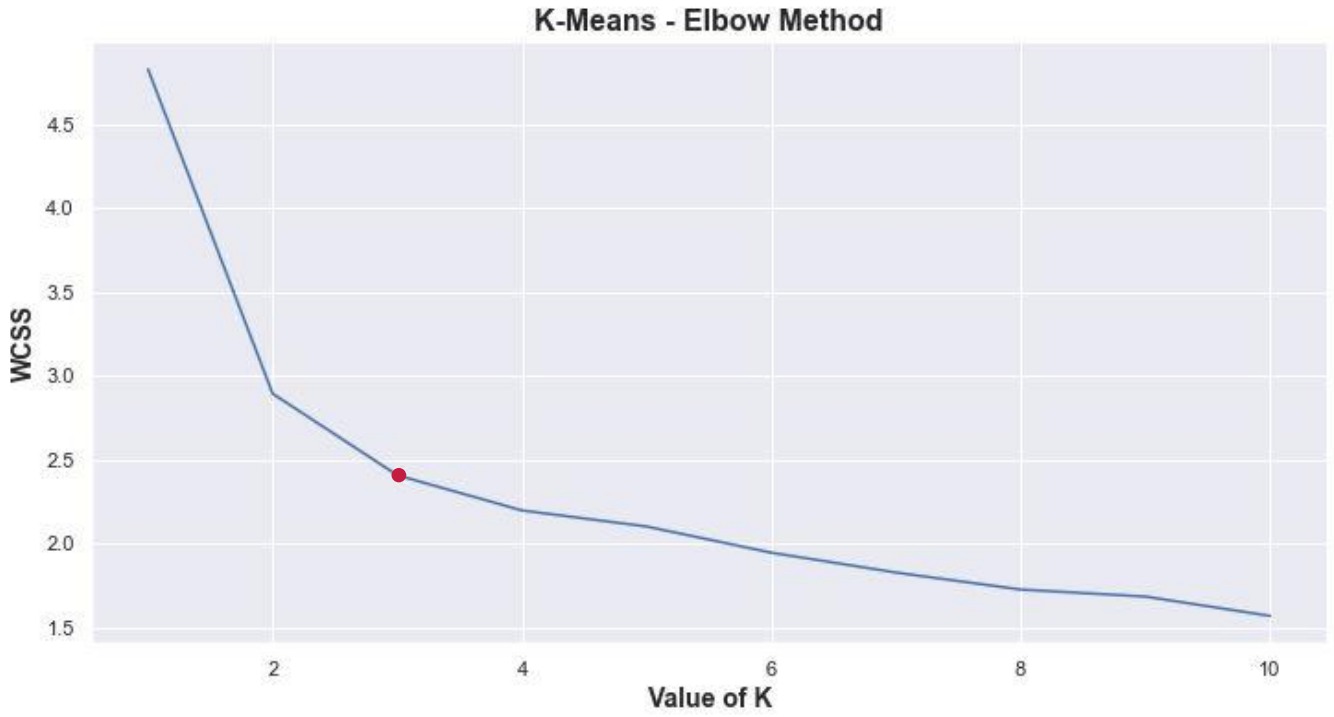
K-Means:

Elbow Method

The Elbow Method

⇒K value is set to 3

⇒This looks like the K that will give the best results



Cluster Properties:

Countries / Cities

Cluster 1

Country	Capital City	GDP/Capita 2018
Puerto Rico	San Juan	$ 31,651
Spain	Madrid	$ 30,371
Cyprus	Nicosia	$ 28,159
Turks and Caicos Islands	Cockburn Town	$ 27,142
Slovenia	Ljubljana	$ 26,124
Portugal	Lisbon	$ 23,408
Estonia	Tallinn	$ 23,266
Czech Republic	Prague	$ 23,079
Greece	Athens	$ 20,324
Lithuania	Vilnius	$ 19,153
Barbados	Bridgetown	$ 17,949
Latvia	Riga	$ 17,861
Uruguay	Montevideo	$ 17,278
Trinidad and Tobago	Port of Spain	$ 17,130
Antigua and Barbuda	St. John's	$ 16,727
Hungary	Budapest	$ 16,162
Chile	Santiago	$ 15,923
Panama	Panama City	$ 15,575
Poland	Warsaw	$ 15,421
Croatia	Zagreb	$ 14,910
Romania	Bucharest	$ 12,301
Costa Rica	San José	$ 12,027
Argentina	Buenos Aires	$ 11,684
Grenada	St. George's	$ 10,640
Mexico	Mexico City	$ 9,673
Turkey	Ankara	$ 9,370
Bulgaria	Sofia	$ 9,273
Brazil	Brasília	$ 8,921

Cluster 2

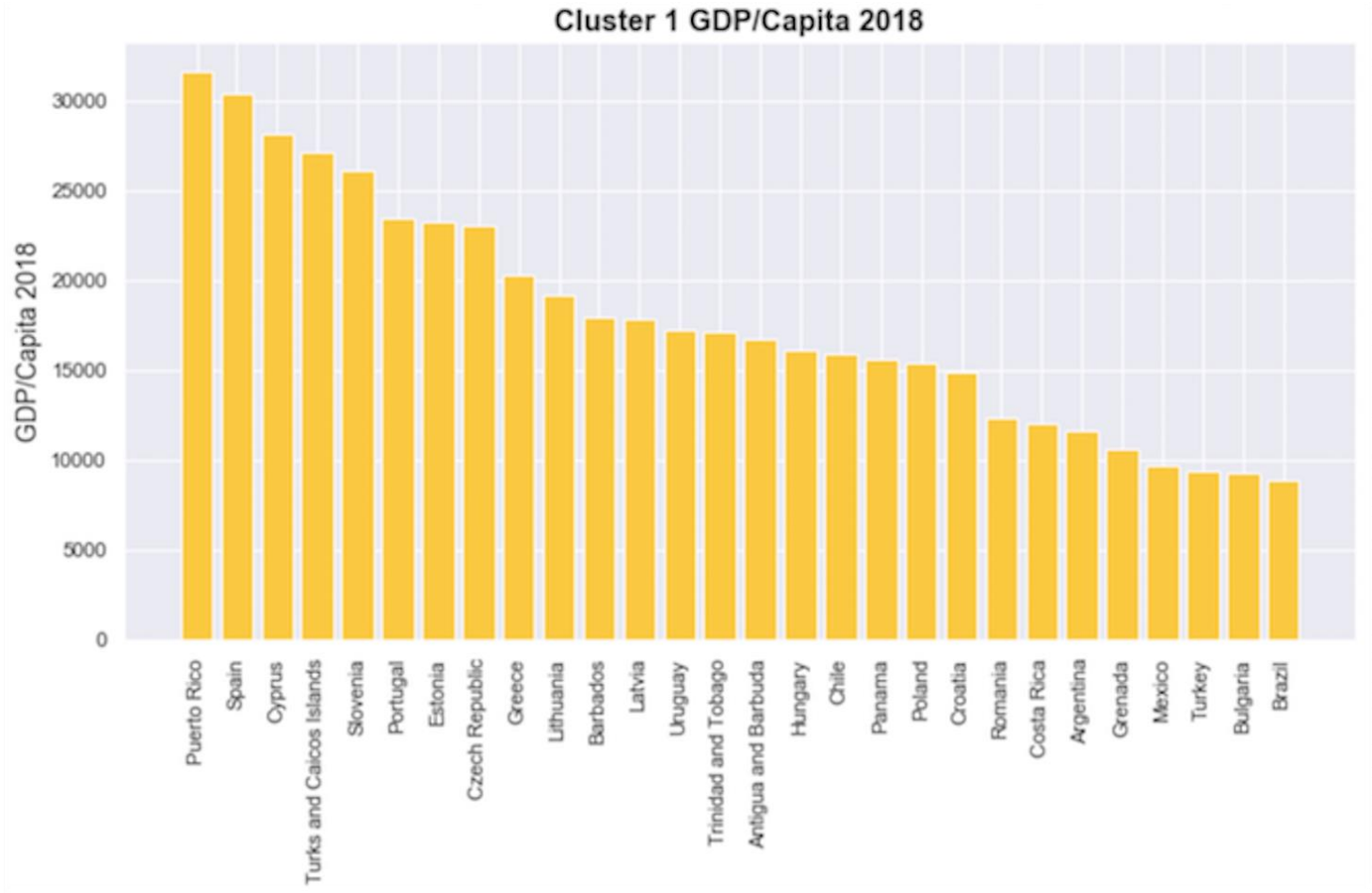
Country	Capital City	GDP/Capita 2018
Montenegro	Podgorica	$ 8,844
Cuba	Havana	$ 8,822
Dominican Republic	Santo Domingo	$ 8,051
Dominica	Roseau	$ 7,691
Serbia	Belgrade	$ 7,247
Peru	Lima	$ 6,941
Colombia	Bogotá	$ 6,668
Ecuador	Quito	$ 6,345
Belarus	Minsk	$ 6,290
Suriname	Paramaribo	$ 6,234
North Macedonia	Skopje	$ 6,084
Bosnia and Herzegovina	Sarajevo	$ 6,066
Paraguay	Asunción	$ 5,822
Jamaica	Kingston	$ 5,354
Albania	Tirana	$ 5,269
Guyana	Georgetown	$ 4,979
Belize	Belmopan	$ 4,885
Guatemala	Guatemala City	$ 4,549
El Salvador	San Salvador	$ 4,058
Moldova	Chișinău	$ 3,227
Ukraine	Kiev	$ 3,095
Honduras	Tegucigalpa	$ 2,500
Nicaragua	Managua	$ 2,029
Haiti	Port-au-Prince	$ 868

Cluster 3			
Country	**Capital City**	**GDP/Capita 2018**	
Monaco	Monaco	$	185,741
Luxembourg	Luxembourg	$	116,640
Switzerland	Bern	$	82,797
Norway	Oslo	$	81,697
Ireland	Dublin	$	78,806
Iceland	Reykjavík	$	73,191
United States	Washington, D.C.	$	62,795
Denmark	Copenhagen	$	61,350
Sweden	Stockholm	$	54,608
Netherlands	Amsterdam	$	53,024
Austria	Vienna	$	51,462
Finland	Helsinki	$	50,152
Germany	Berlin	$	47,603
Belgium	Brussels	$	47,519
Canada	Ottawa	$	46,233
United Kingdom	London	$	42,944
Andorra	Andorra la Vella	$	42,030
France	Paris	$	41,464
Italy	Rome	$	34,483

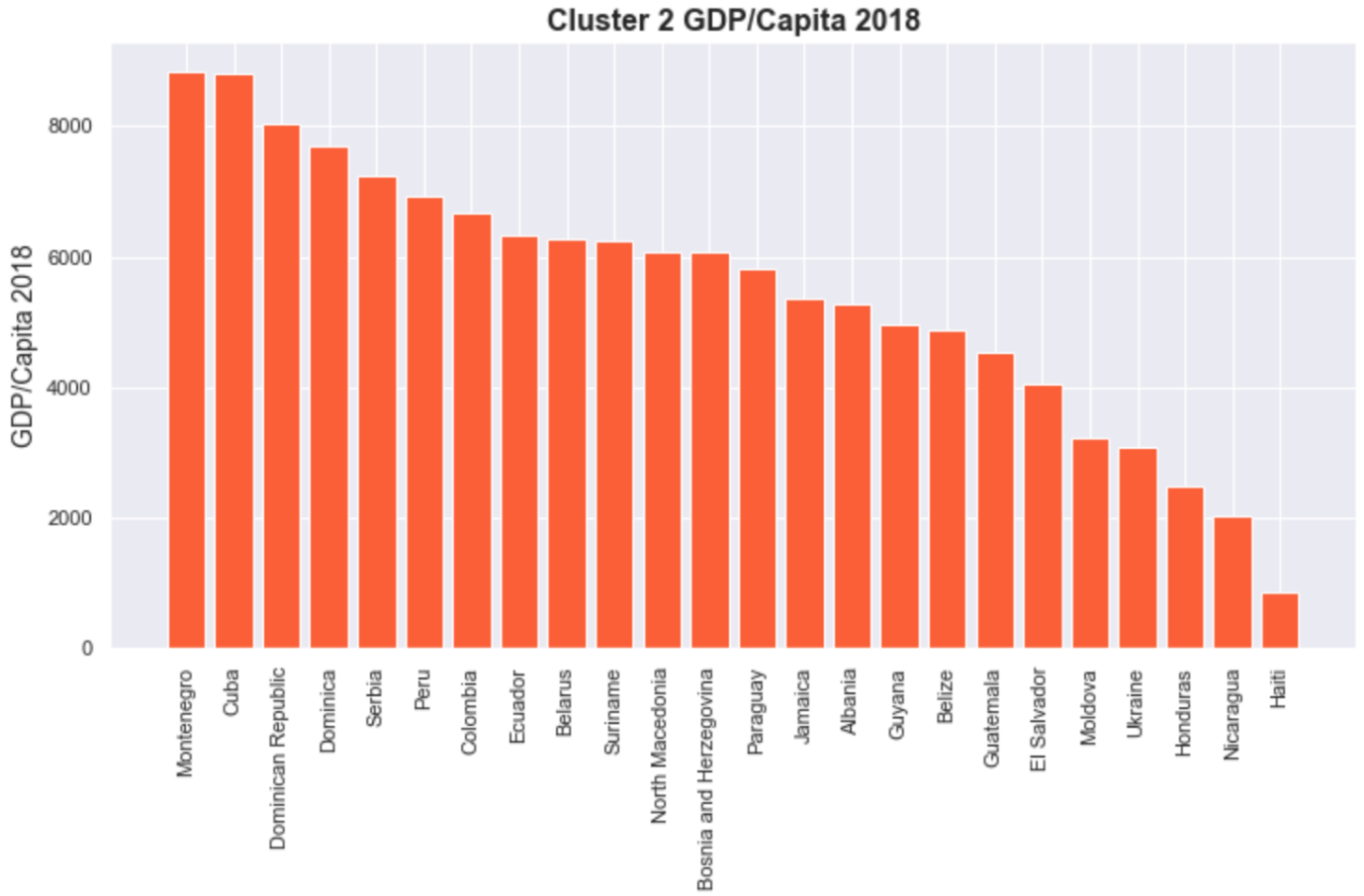
Cluster Properties:

Country GDP/Capita

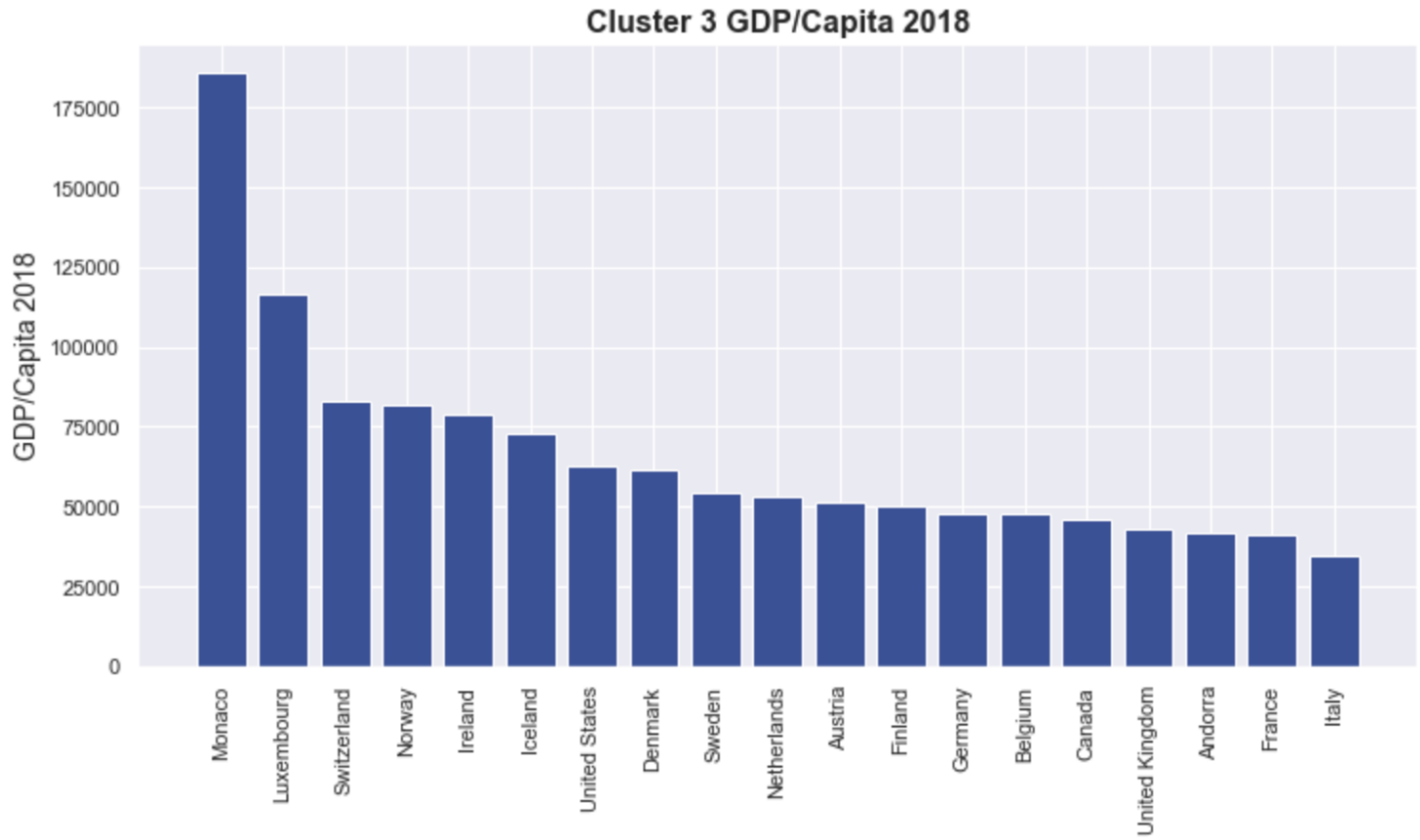
Cluster 1


Cluster 1 GDP/Capita 2018
GDP/Capita 2018
0
5000
10000
15000
20000
25000
30000
Puerto Rico
Spain
Cyprus
Turks and Caicos Islands
Slovenia
Portugal
Estonia
Czech Republic
Greece
Lithuania
Barbados
Latvia
Uruguay
Trinidad and Tobago
Antigua and Barbuda
Hungary
Chile
Panama
Poland
Croatia
Romania
Costa Rica
Argentina
Grenada
Mexico
Turkey
Bulgaria
Brazil

Cluster 2


Cluster 2 GDP/Capita 2018
GDP/Capita 2018
0
2000
4000
6000
8000
Montenegro
Cuba
Dominican Republic
Dominica
Serbia
Peru
Colombia
Ecuador
Belarus
Suriname
North Macedonia
Bosnia and Herzegovina
Paraguay
Jamaica
Albania
Guyana
Belize
Guatemala
El Salvador
Moldova
Ukraine
Honduras
Nicaragua
Haiti

Cluster 3


Cluster 3 GDP/Capita 2018
GDP/Capita 2018
175000
150000
125000
100000
75000
50000
25000
0
Monaco
Luxembourg
Switzerland
Norway
Ireland
Iceland
United States
Denmark
Sweden
Netherlands
Austria
Finland
Germany
Belgium
Canada
United Kingdom
Andorra
France
Italy

Cluster Properties:

GDP/Capita Statistics

Cluster Summary Statistics

Cluster 1	GDP/Capita 2018
Count	28
Mean	$ 17,911
Std.	$ 6,607
Min	$ 8,921
25%	$ 12,233
50%	$ 16,929
75%	$ 23,126
Max	$ 31,651

Cluster 2	GDP/Capita 2018
Count	24
Mean	$ 5,497
Std.	$ 2,083
Min	$ 868
25%	$ 4,426
50%	$ 5,944
75%	$ 6,736
Max	$ 8,844

Cluster 3	GDP/Capita 2018
Count	19
Mean	$ 66,028
Std.	$ 35,128
Min	$ 34,483
25%	$ 46,876
50%	$ 53,024
75%	$ 75,999
Max	$ 185,741

Cluster Summary Statistics

⇒The sizes of the clusters are different

- Cluster 1 – 28 countries
- Cluster 2 – 24 countries
- Cluster 3 – 19 countries

⇒Average GDP/Capita

- Cluster 1 – medium
- Cluster 2 – low
- Cluster 3 – high

Cluster Summary Statistics

⇒Variance of GDP/Capita

- Cluster 1 – moderate
- Cluster 2 – low
- Cluster 3 – high

Cluster Properties:

Venues Statistics

Venue Category Frequencies

Cluster 1	
Average Frequency in Cluster	
Hotel	6.77%
Restaurant	4.47%
Café	4.25%
Coffee Shop	3.67%
Bar	3.06%
Italian Restaurant	2.80%
Bakery	2.12%
Pizza Place	2.00%
Park	1.91%
Eastern European Restaurant	1.66%

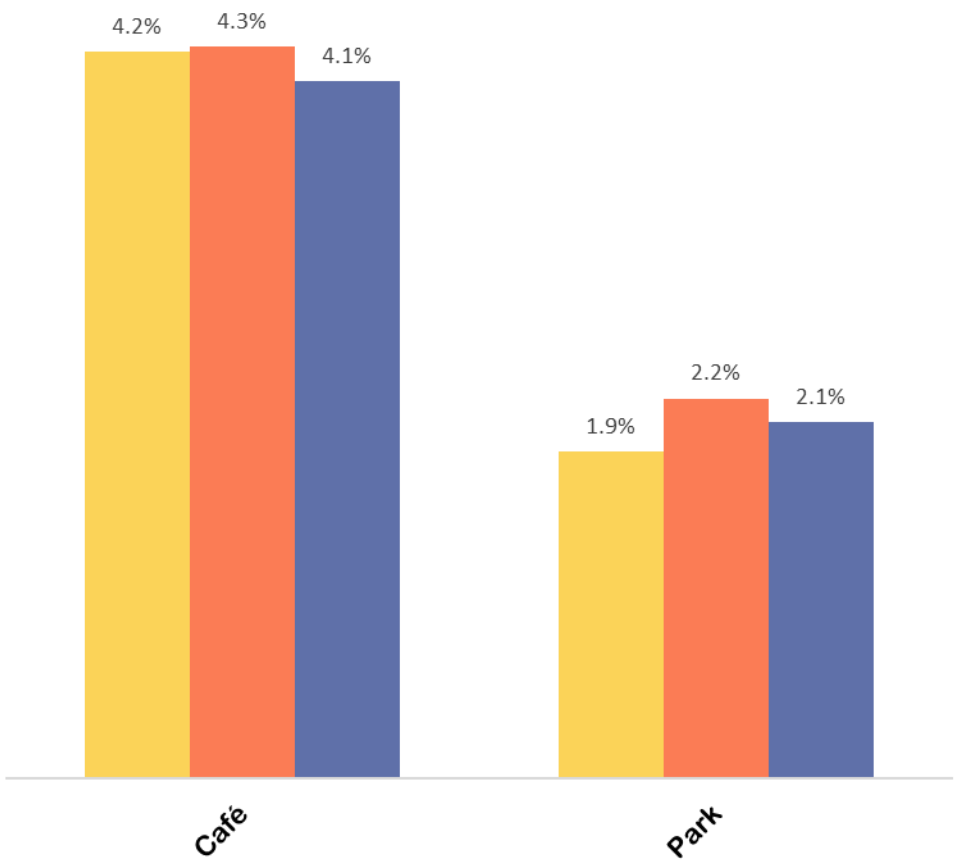
Cluster 2	
Average Frequency in Cluster	
Hotel	5.97%
Restaurant	4.30%
Coffee Shop	4.29%
Café	4.28%
Bar	3.92%
Plaza	2.32%
Park	2.22%
Bakery	1.88%
Caribbean Restaurant	1.87%
Theater	1.71%

Cluster 3	
Average Frequency in Cluster	
Hotel	6.33%
Coffee Shop	4.44%
Café	4.07%
Restaurant	3.59%
Plaza	3.12%
Bar	2.96%
Park	2.08%
Cocktail Bar	1.96%
Historic Site	1.95%
Theater	1.74%

Low variance venues:

⇒These venues have similar frequencies between the clusters per venue category

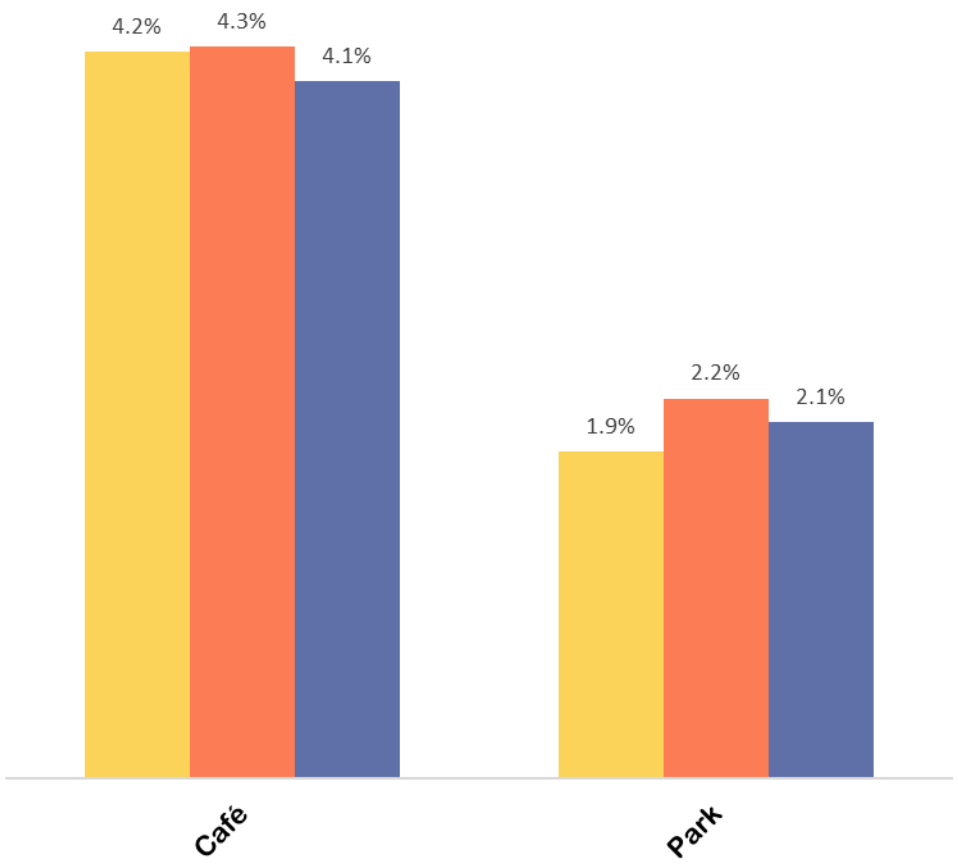
⇒A difference of 0.1%-0.3% between clusters



Low variance venues:

⇒We can see similarity between clusters' ***Cafés*** and ***Parks*** frequencies

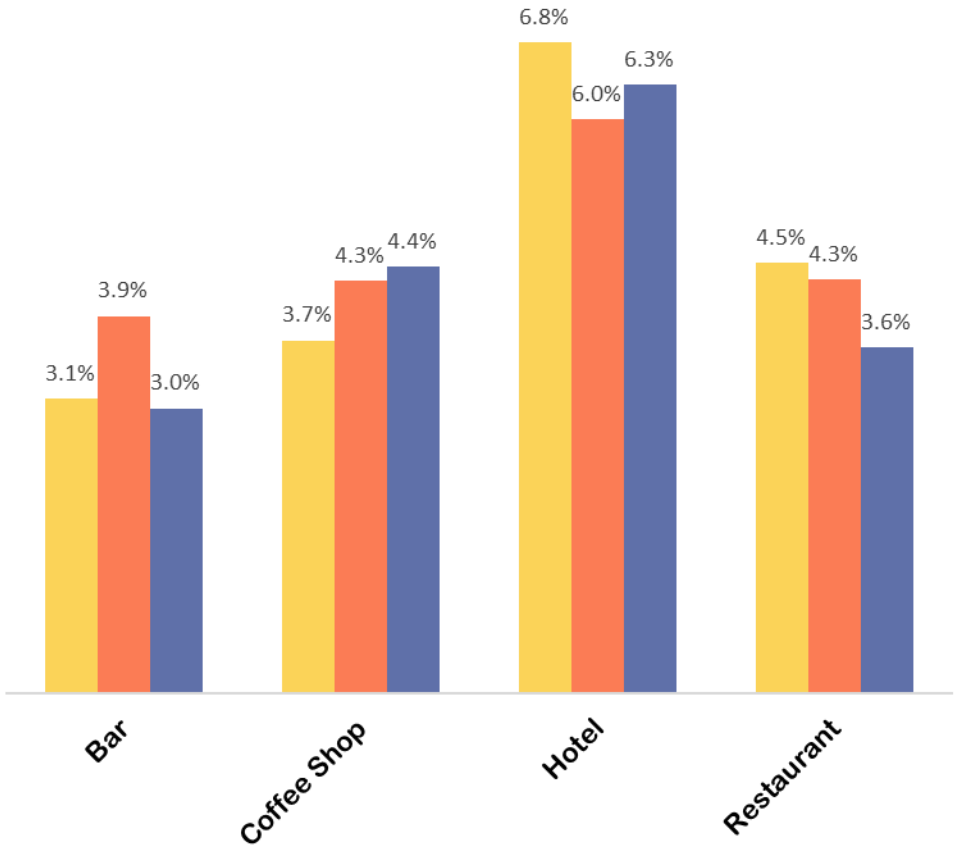
⇒These venues don't contribute much to the difference between clusters



Moderate variance venues:

⇒Frequencies of venues per venue category differ between the clusters

⇒A difference of 0.1%-0.9% between clusters



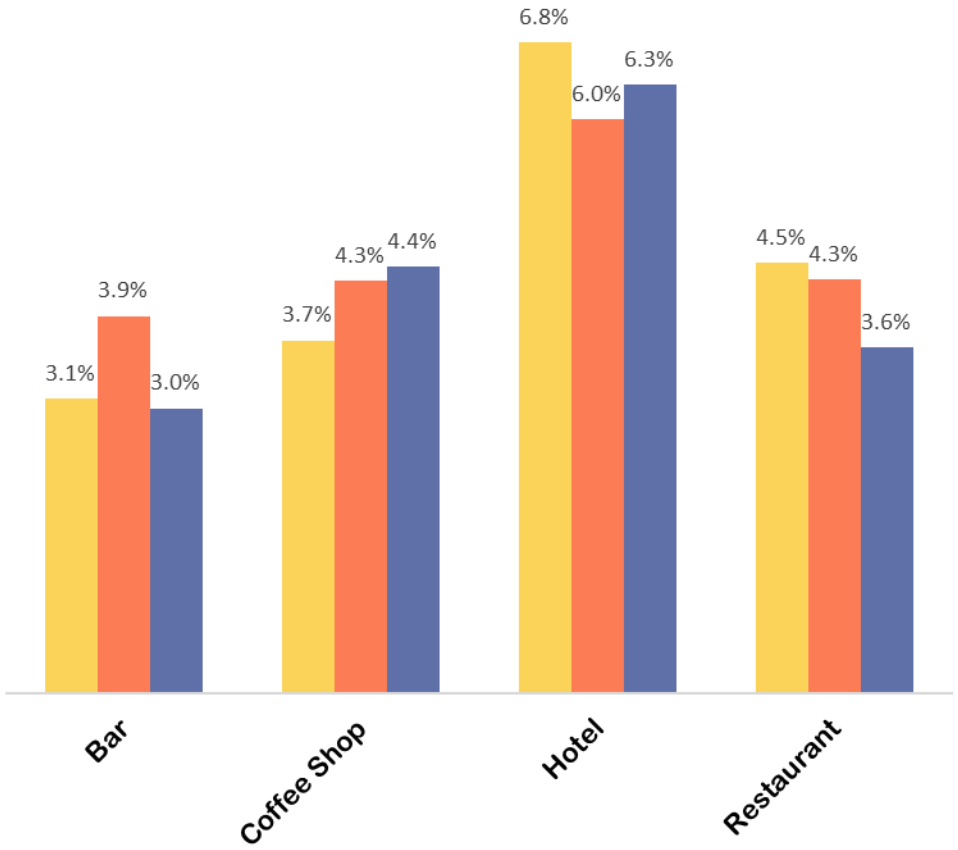
Moderate variance venues:

⇒Some variance between clusters in these venues' categories

⇒Cluster 1 is more ***Hotel*** and ***Restaurant*** oriented

⇒Cluster 2 is high on ***Bars, Coffee Shops*** and ***Restaurants*** but lower on ***Hotels***

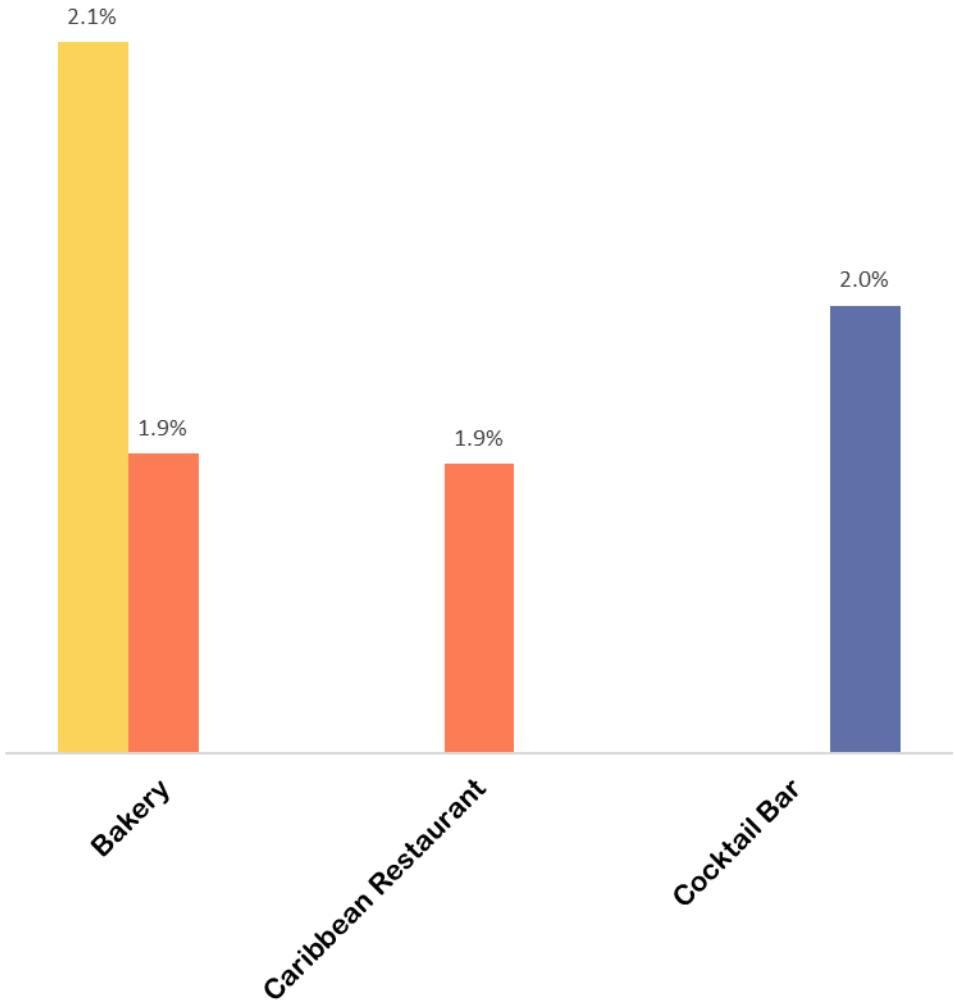
⇒Cluster 3 is low on ***Bars, Hotels*** and ***Restaurants*** but higher on ***Coffee Shops***



Cluster differentiating venues I:

⇒These venues appear only in some of the clusters

⇒These are the venues that give each cluster its uniqueness compared to other clusters

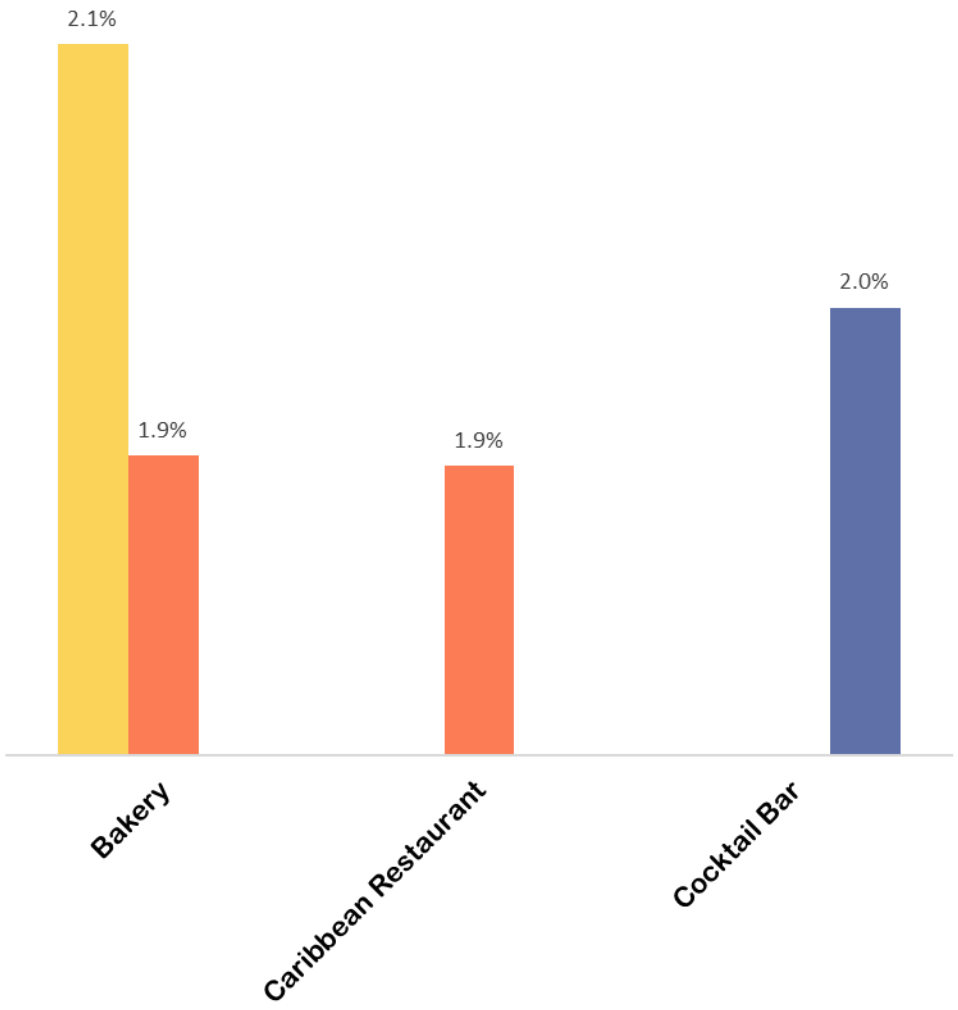


Cluster differentiating venues I:

⇒Cluster 1 has a very high ***Bakery*** frequency compared to other clusters

⇒Cluster 2 has a very high ***Caribbean Restaurant*** frequency and a moderate ***Bakery*** frequency compared to other clusters

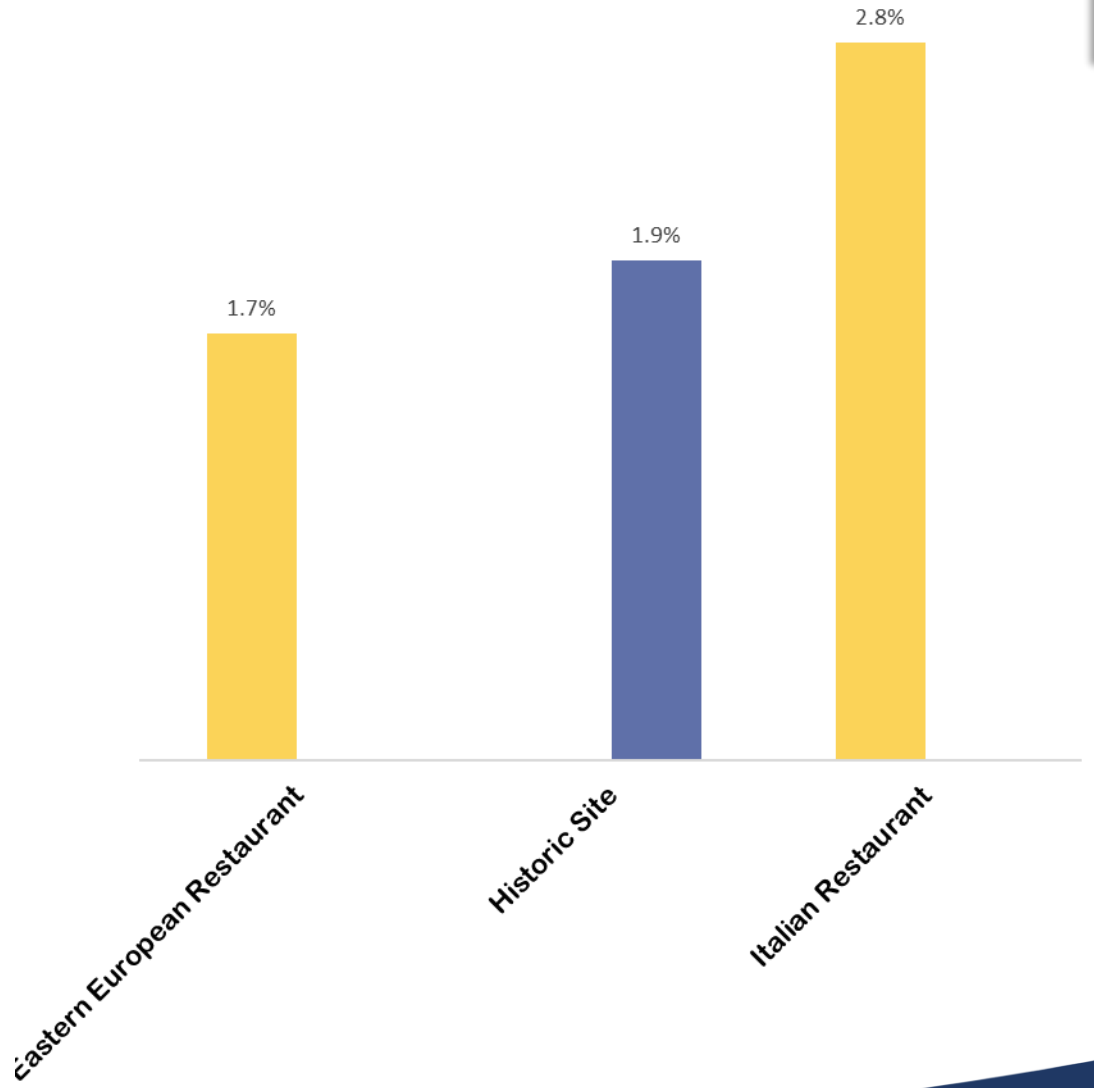
⇒Cluster 3 has a very high ***Cocktail Bar*** frequency compared to other clusters



Cluster differentiating venues II:

⇒These venues appear only in some of the clusters

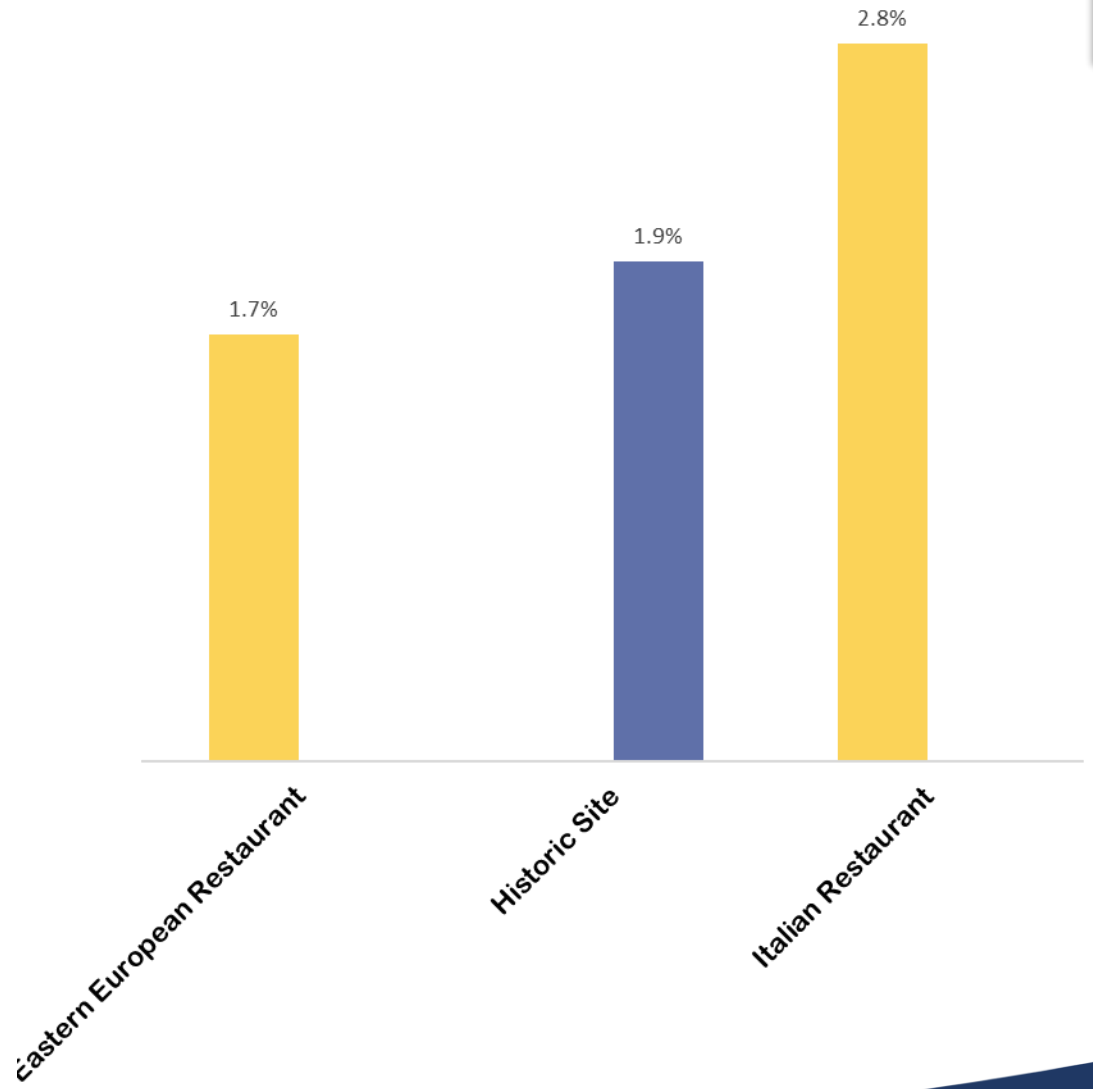
⇒These are the venues that give each cluster its uniqueness compared to other clusters



Cluster differentiating venues II:

⇒Cluster 1 has a very high ***Eastern European Restaurant*** and ***Italian Restaurant*** frequencies compared to other clusters

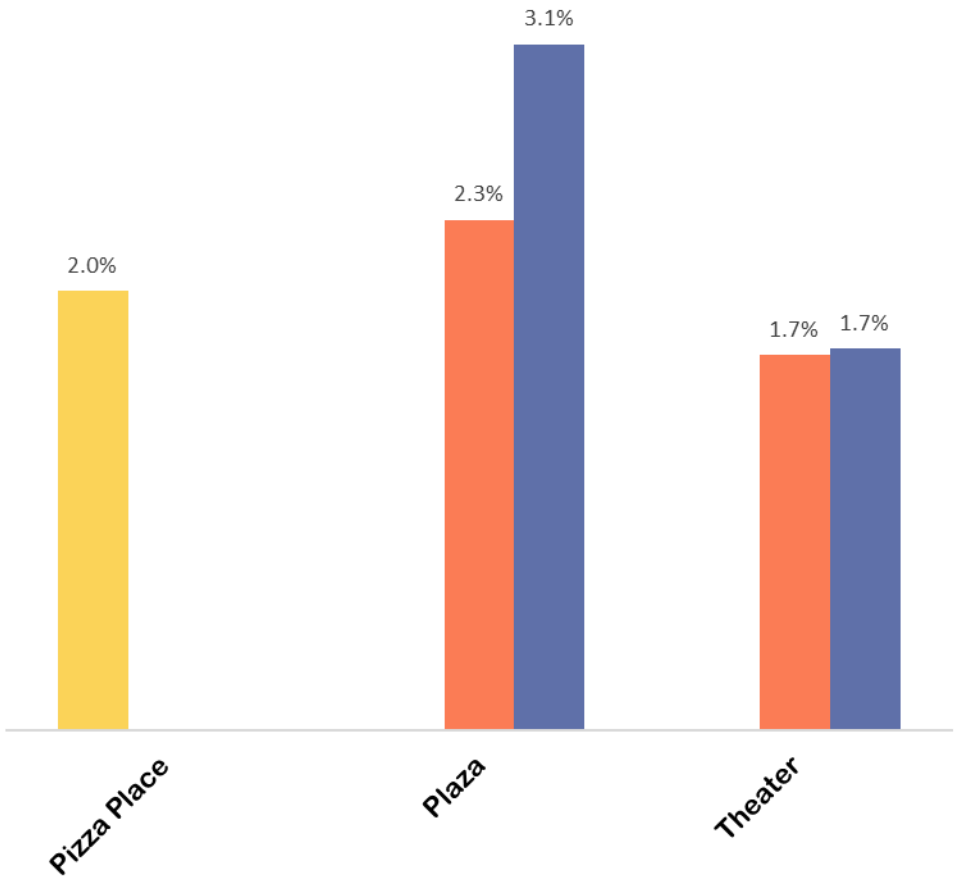
⇒Cluster 3 has a very high ***Historic Site*** frequency compared to other clusters



Cluster differentiating venues III:

⇒These venues appear only in some of the clusters

⇒These are the venues that give each cluster its uniqueness compared to other clusters

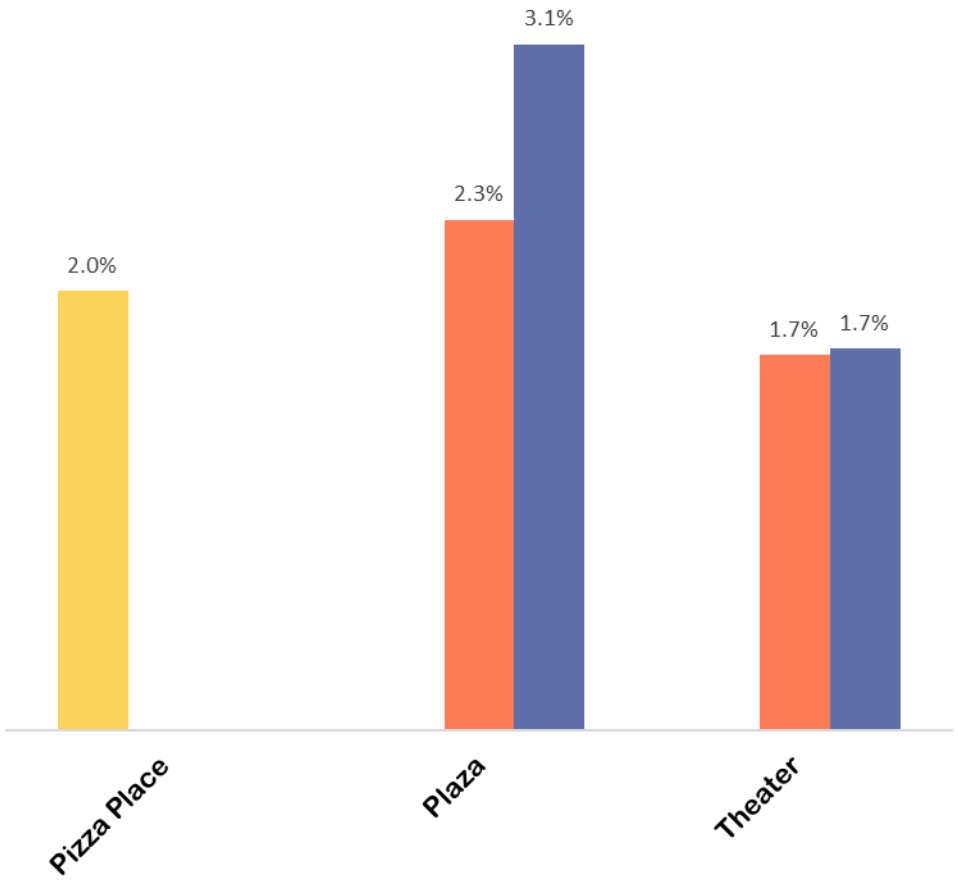


Cluster differentiating venues III:

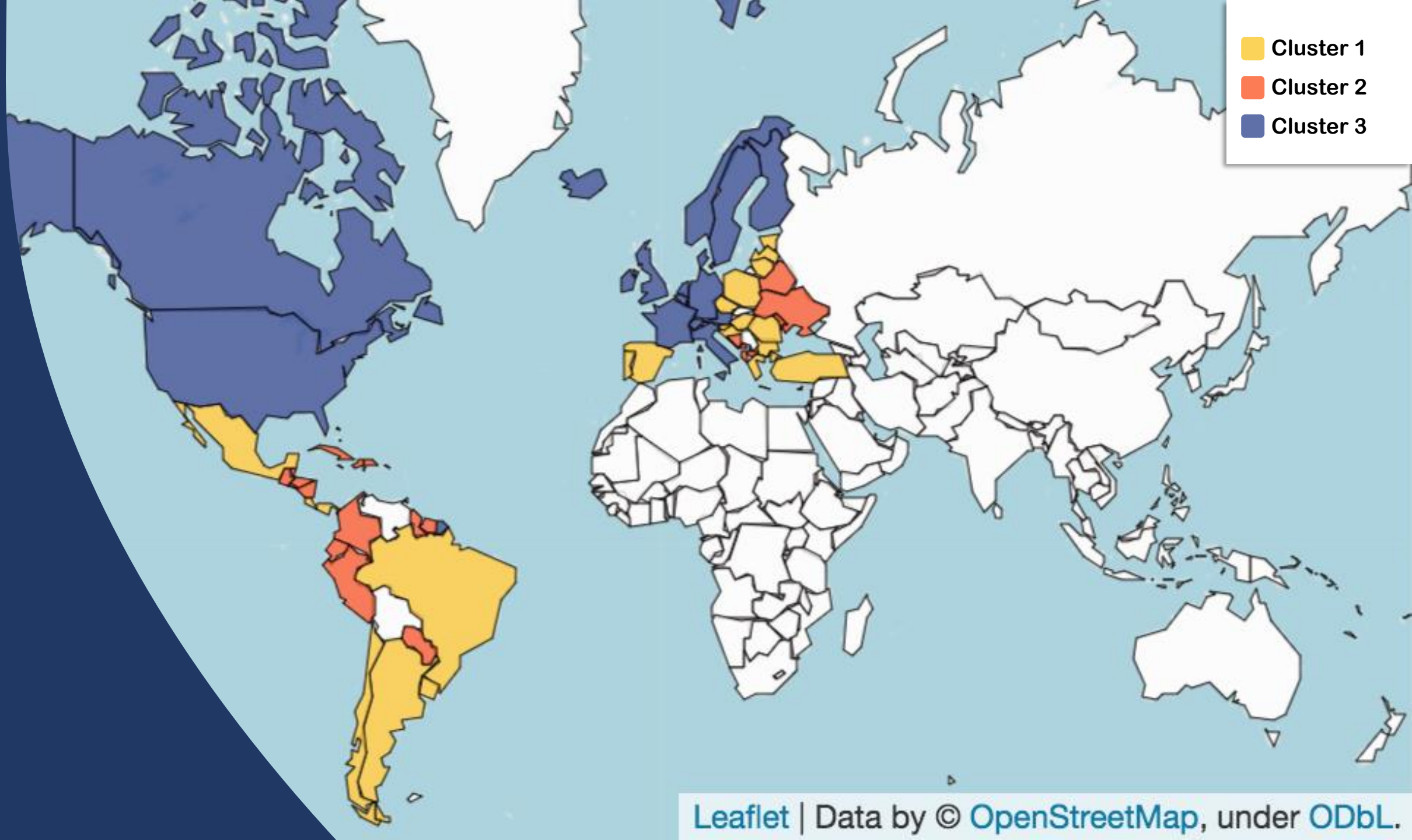
⇒Cluster 1 has a very high ***Pizza Place*** frequency compared to other clusters

⇒Cluster 2 has a high ***Plaza*** and ***Theater*** frequencies compared to other clusters

⇒Cluster 3 has a very high ***Plaza*** frequency and a high ***Theater*** frequencies compared to other clusters



Clusters' Map


Cluster 1
Cluster 2
Cluster 3
Leaflet | Data by © OpenStreetMap, under ODbL.

Discussion

Discussion

⇒There is a total of 3 clusters

⇒Each cluster has countries from more than one continent

⇒Clusters’ sizes are different

Discussion

⇒Clusters are differentiated by their average GDP/Capita:

- Cluster 1 – medium GDP/Capita
- Cluster 2 – low GDP/Capita
- Cluster 3 – high GDP/Capita

Discussion

⇒Clusters are differentiated by their average GDP/Capita:

- Cluster 1 – has more ***Food Diversity***
- Cluster 2 – a **Hybrid** of the other two clusters → has more ***Food Diversity*** than Cluster 3 and more ***Extra Activities*** than Cluster 1
- Cluster 3 – has more options for ***Extra Activities*** like ***Theaters, Plazas and Cocktail Bars***

References

References

Foursquare. (n.d.). *Foursquare Developer*. Retrieved December 28, 2019, from

Foursquare: https://developer.foursquare.com/

Wikipedia contributors. (2019, December). *List of national capitals*. Retrieved March 04, 2020, from

Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_national_capitals&oldid=943802946

World Bank, World Development Indicators. (n.d.-b). *GDP per capita (current US$) [Data File]*. Retrieved December 28, 2019, from

The World Bank: https://databank.worldbank.org/source/world-development-indicators

Thanks!