Travel adviser tool

This is a presentation of a Coursera Data Science course Capstone Project.

Author: Adam Borowa

Agenda

1. Introduction
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

INTRODUCTION

A description of the problem and a discussion of the background.

Main idea came form a question:

‘I liked this city. Where should I go to be satisfied as much as then?’ This question regards all the places on the world. Having a tool that helps to answer this question one would know where to go for another vacations. Such tool would be also very welcome by every Travel Agency. Just imagine if you propose a vacation detonation for a client and he likes your proposition. This client will come again to your agency.

This project however, is focused on capital cities similarity. This is just due the number of data that should be collected and computed if it focuses on every travel destination.

Main assumption is that appropriate clustering method together with well suited data-set shall combine similar cities in the same cluster.

DATA

A description of the data and how it will be used to solve the problem.

First of all a data with all (or almost all) capital names and its geographical location was collected.
It was taken from this particular web page:
'http://techslides.com/list-of-countries-and-capitals'

Data set contains information like:

- Country Name
- Capital Name
- Capital Latitude
- Capital Longitude
- Country Code
- Continent Name

Another set of data was the Foursquare responses for each geographical location exploration query:

```
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},
{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, lat, lng, radius, LIMIT)

requests.get(url).json()
```

DATA

A description of the data and how it will be used to solve the problem.

Another data-set I incorporated into my project is List of countries by GDP (PPP) per capita taken from:
https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita
I focused on World Bank (2019) data set.

I also incorporated the data regarding the city population taken from:
https://en.wikipedia.org/wiki/List_of_national_capitals_by_population

And the final one is List of countries by life expectancy taken from:
https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy

DATA

View of the table with all gathered data.

Foursquare responses

	Capital Name	Capital Latitude	Capital Longitude	Continent Name	response	Population	Life expectancy	Int$
Afghanistan	Kabul	34.51666667	69.183333	Asia	{'suggestedFilters':...	3.14085e+06	64.5	2293
Algeria	Algiers	36.75	3.05	Africa	{'suggestedFilters':...	3.41581e+06	76.7	11820
Angola	Luanda	-8.833333333	13.216667	Africa	{'suggestedFilters':...	2.45378e+06	60.8	6929
Antigua and Barbuda	Saint John's	17.11666667	-61.85	North America	{'suggestedFilters':...	22679	76.9	22817
Argentina	Buenos Aires	-34.58333333	-58.666667	South America	{'suggestedFilters':...	2.89108e+06	76.5	22947
Armenia	Yerevan	40.16666667	44.5	Europe	{'suggestedFilters':...	1.08049e+06	74.9	14220
Australia	Canberra	-35.26666667	149.133333	Australia	{'suggestedFilters':...	410301	83.3	53320
Austria	Vienna	48.2	16.366667	Europe	{'suggestedFilters':...	1.74967e+06	81.4	59111
Azerbaijan	Baku	40.38333333	49.866667	Europe	{'suggestedFilters':...	2.2042e+06	72.9	15001
Bahamas	Nassau	25.08333333	-77.35	North America	{'suggestedFilters':...	248948	73.8	37266
Bahrain	Manama	26.23333333	50.566667	Asia	{'suggestedFilters':...	140616	77.2	46892
Bangladesh	Dhaka	23.71666667	90.4	Asia	{'suggestedFilters':...	8.90604e+06	72.3	4950
Barbados	Bridgetown	13.1	-59.616667	North America	{'suggestedFilters':...	110000	79.1	16287
Belarus	Minsk	53.9	27.566667	Europe	{'suggestedFilters':...	1.70206e+06	74.6	19943
Belgium	Brussels	50.83333333	4.333333	Europe	{'suggestedFilters':...	148873	81.5	54545
Belize	Belmopan	17.25	-88.766667	Central America	{'headerLocation': '...	16451	74.5	7295
Benin	Porto-Novo	6.483333333	2.616667	Africa	{'headerLocation': '...	223552	61.5	3423
Bhutan	Thimphu	27.46666667	89.633333	Asia	{'suggestedFilters':...	101259	71.5	11613
Bosnia and Herzegovina	Sarajevo	43.86666667	18.416667	Europe	{'suggestedFilters':...	395133	77.3	15792
Botswana	Gaborone	-24.63333333	25.9	Africa	{'suggestedFilters':...	225656	69.3	18503
Brazil	Brasilia	-15.78333333	-47.916667	South America	{'suggestedFilters':...	2.64853e+06	75.7	15259

DATA

View of the table with venues data added

Foursquare responses

	Capital Name	Capital Latitude	Capital Longitude	Population	GDP per capita [int$]	Life expectancy	Venue Name	Venue Category
0	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Kabul Serena Hotel	Hotel
1	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Ciano ISAF	Pizza Place
2	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Park Star Hotel	Hotel
3	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Barg Continental	Afghan Restaurant
4	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	İstanbul Restaurant	Turkish Restaurant
5	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	NKC New Kabul Compou	Gym
6	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Kabul Star Hotel	Hotel
7	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Gulbahar Center	Shopping Mall
8	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Tora Bora Bar ISAFHQ	Food
9	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Blue Cafe - ISAF HQ	Coffee Shop
10	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Sufi Restaurant	Afghan Restaurant
11	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Majid Mall	Shopping Mall
12	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Table Talk Coffee Sh	Café
13	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Afghan Fried Chicken	Fried Chicken Joint
14	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Afghanistan Football	Soccer Field
15	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Flower Street Cafe	Café
16	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	park mall	Shopping Mall
17	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Istanbul Restaurant	Restaurant
18	Kabul	34.51666667	69.183333	3.14085e+06	2293	64.5	Makroyan 3 Market	Flea Market

METHODOLOGY
Venue categories simplification

Foursquare
Venue categories

As a result of foursquare response there are 497 venue categories. Nevertheless some of them are pretty similar For example: Soccer Field or Soccer Stadium. Some others are similar in its functionality: Metro Station, Train Station, Bus Station etc.

Therefore, I decided to ***merge some categories***. As a result I get 28 unique categories such as:

- Café,
- Culture,
- Food,
- Hotel,
- Nightlife,
- Sport,
- and more...

Venue Name	Venue Category
Kabul Serena Hotel	Hotel
Ciano ISAF	Pizza Place
Park Star Hotel	Hotel
Barg Continental	Afghan Restaurant
İstanbul Restaurant	Turkish Restaurant
NKC New Kabul Compou	Gym
Kabul Star Hotel	Hotel
Gulbahar Center	Shopping Mall
Tora Bora Bar ISAFHQ	Food
Blue Cafe - ISAF HQ	Coffee Shop
Sufi Restaurant	Afghan Restaurant
Majid Mall	Shopping Mall
Table Talk Coffee Sh	Café
Afghan Fried Chicken	Fried Chicken Joint
Afghanistan Football	Soccer Field
Flower Street Cafe	Café
park mall	Shopping Mall
Istanbul Restaurant	Restaurant
Makroyan 3 Market	Flea Market

METHODOLOGY

Visualization of final dataset for clustering.

After:

- ***venue categories aggregation*** [*One hot Venue Category transform*],
- ***merging it*** to the data gathered in the first place and
- ***normalization***

I get such a dataset for clustering purposes:

	Population	Life expectancy	Int$	Airport	Attraction	Café	Culture	Food	Garden	General Entertainment	...	Pool	Relax	Restaurant	Road	Sport	Store	Tea Room	Temple	Transport	number of venues
Abu Dhabi	0.026870	0.788644	0.572833	0.0	0.000000	0.270000	0.010000	0.130000	0.0	0.0	...	0.00	0.05	0.290000	0.0	0.010000	0.110000	0.010000	0.0	0.000000	1.000000
Abuja	0.035853	0.047319	0.036273	0.0	0.000000	0.025641	0.000000	0.179487	0.0	0.0	...	0.00	0.00	0.333333	0.0	0.076923	0.051282	0.000000	0.0	0.000000	0.383838
Accra	0.075877	0.347003	0.038675	0.0	0.010000	0.020000	0.050000	0.170000	0.0	0.0	...	0.01	0.01	0.350000	0.0	0.010000	0.130000	0.000000	0.0	0.000000	1.000000
Addis Ababa	0.140897	0.422713	0.011030	0.0	0.000000	0.142857	0.000000	0.000000	0.0	0.0	...	0.00	0.00	0.428571	0.0	0.142857	0.285714	0.000000	0.0	0.000000	0.060606
Algiers	0.158313	0.753943	0.090068	0.0	0.012658	0.139241	0.063291	0.126582	0.0	0.0	...	0.00	0.00	0.392405	0.0	0.012658	0.012658	0.012658	0.0	0.012658	0.787879

METHODOLOGY
Kmeans clustering

For clustering purposes I used a k-means algorithm from sklearn library.

```
kmeans = KMeans(n_clusters=kclusters,
                random_state=0,
                algorithm='elkan',
                max_iter=1000).fit(capital_grouped_clustering)
```

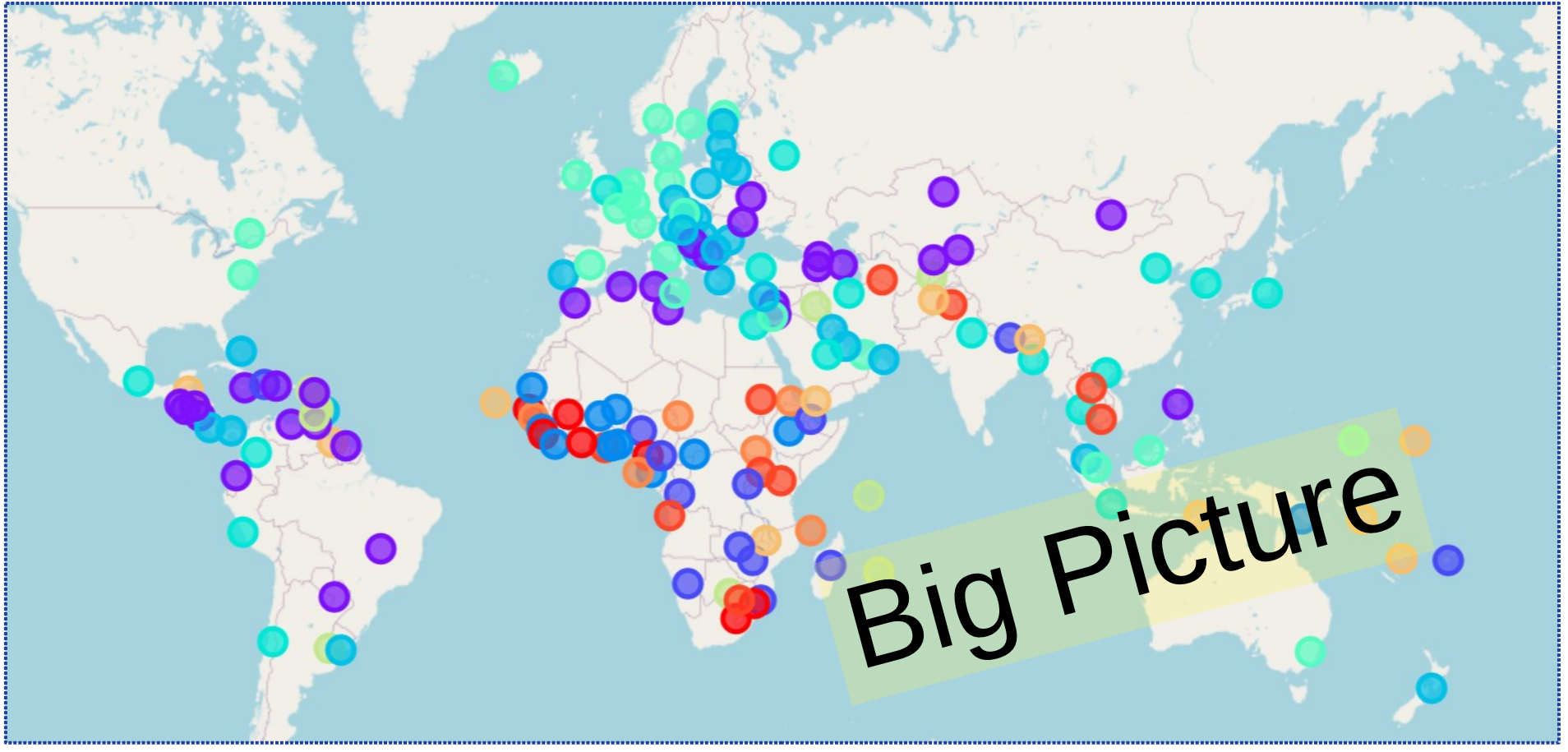
I run number of runs in order to find a proper number of clusters. I was searching such a number of clusters that gives reasonable results.

My assumptions:

- minimize number of one-element clusters
- avoid clusters with huge number of elements [in comparison to other clusters]

Final number of clusters was 12.

RESULTS



RESULTS

Single cluster example

Each cluster gets its name. It was the name of the country which capital was closest to the center of the cluster. Let us look closer to the Netherlands cluster:

```
CLUSTER NAME: Netherlands
```

	Capital Name	Cluster Labels	Country Name	Capital Latitude	Capital Longitude	Continent Name
0	Abu Dhabi	6	United Arab Emirates	24.466667	54.366667	Asia
142	Singapore	6	Singapore	1.283333	103.850000	Asia
18	Bandar Seri Begawan	6	Brunei Darussalam	4.883333	114.933333	Asia
61	Jerusalem	6	Israel	31.766667	35.233333	Asia
40	Canberra	6	Australia	-35.266667	149.133333	Australia
109	Ottawa	6	Canada	45.416667	-75.700000	Central America
163	Washington, D.C.	6	United States	38.883333	-77.000000	Central America
45	Copenhagen	6	Denmark	55.666667	12.583333	Europe
49	Dublin	6	Ireland	53.316667	-6.233333	Europe
57	Helsinki	6	Finland	60.166667	24.933333	Europe
28	Bern	6	Switzerland	46.916667	7.466667	Europe
159	Vienna	6	Austria	48.200000	16.366667	Europe
84	Madrid	6	Spain	40.400000	-3.683333	Europe
108	Oslo	6	Norway	59.916667	10.750000	Europe
27	Berlin	6	Germany	52.516667	13.400000	Europe
114	Paris	6	France	48.866667	2.333333	Europe
128	Reykjavik	6	Iceland	64.150000	-21.950000	Europe
131	Rome	6	Italy	41.900000	12.483333	Europe
6	Amsterdam	6	Netherlands	52.350000	4.916667	Europe
145	Stockholm	6	Sweden	59.333333	18.050000	Europe
157	Valletta	6	Malta	35.883333	14.500000	Europe
35	Brussels	6	Belgium	50.833333	4.333333	Europe
83	Luxembourg	6	Luxembourg	49.600000	6.116667	Europe

As you may see, this cluster contains capitals from western Europe, few from Asia, Washington,Ottawa and Canberra.
Almost all those cities are similar to each other in terms of culture, comfort of living etc…

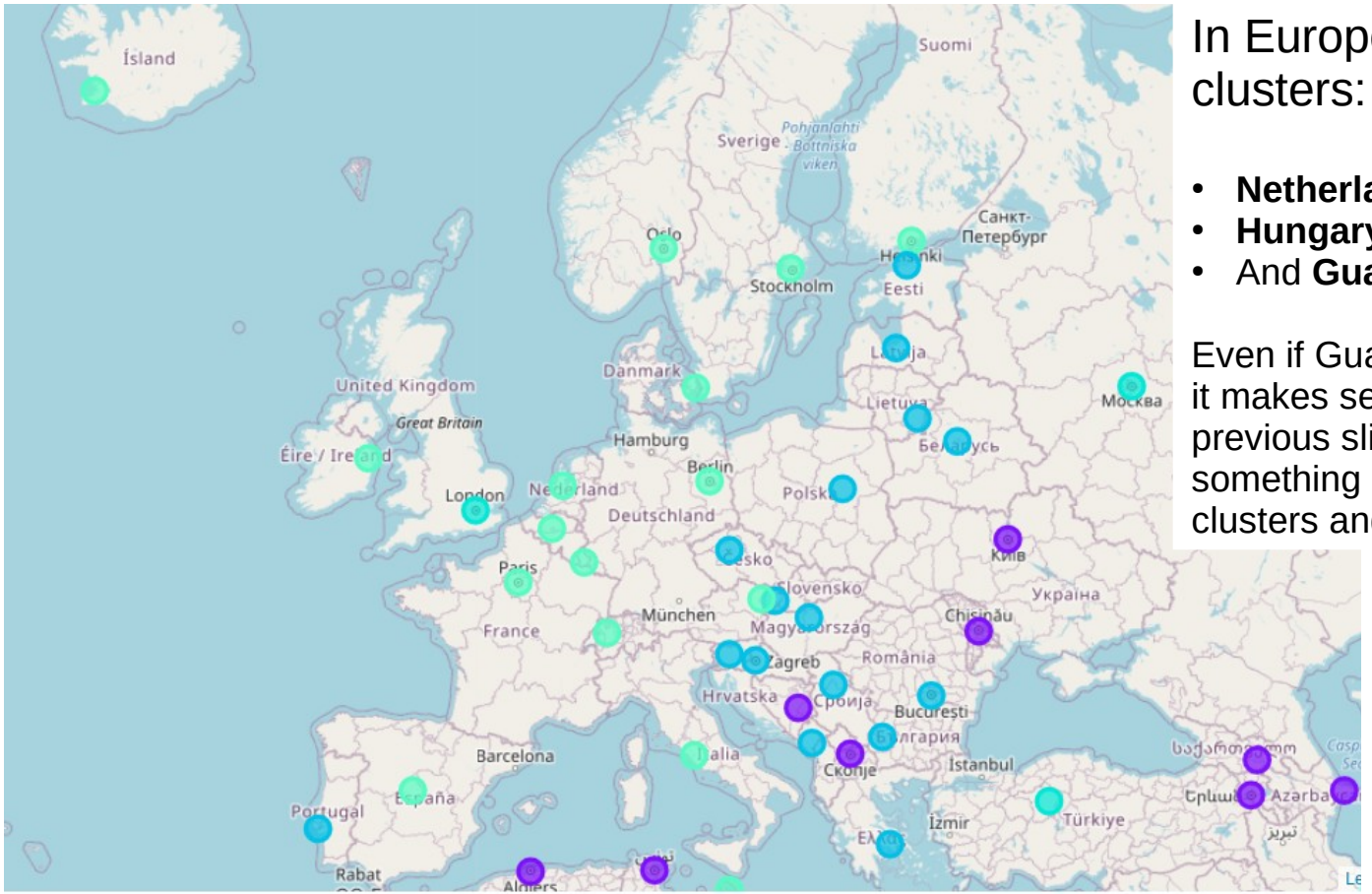
In order to make this presentation clear, detailed results are given in excel file. You may download from here:
CapstoneProjectResults.xlsx

RESULTS
Distances between clusters

Besides clustering one may be also interested in which luster is the closest to the one I really like.
Therefore, a table with the closest cluster is also given as a result:

	cluster_name	nearest_claster	second_nearest_claster
0	Equatorial Guinea	Niger	Timor-Leste
1	Guatemala	Hungary	Kenya
2	Mozambique	Saint Lucia	Niger
3	Niger	Timor-Leste	Equatorial Guinea
4	Hungary	Guatemala	Netherlands
5	Mexico	Guatemala	Hungary
6	Netherlands	Hungary	Guatemala
7	Federated States of Micronesia	Timor-Leste	Niger
8	Saint Lucia	Mozambique	Timor-Leste
9	Timor-Leste	Niger	Saint Lucia
10	South Sudan	Equatorial Guinea	Timor-Leste
11	Kenya	Guatemala	Mozambique

DISCUSSION
Europe clusters



In Europe one may see three different clusters:

- **Netherlands cluster** [western Europe]
- **Hungary cluster** [middle and eastern Europe]
- And **Guatemala cluster** [purple one]

Even if Guatemala cluster name sounds strange, it makes sense. Take a look at the table on the previous slide. Guatemala cluster seems to be something like a bridge between ‘western culture’ clusters and others.

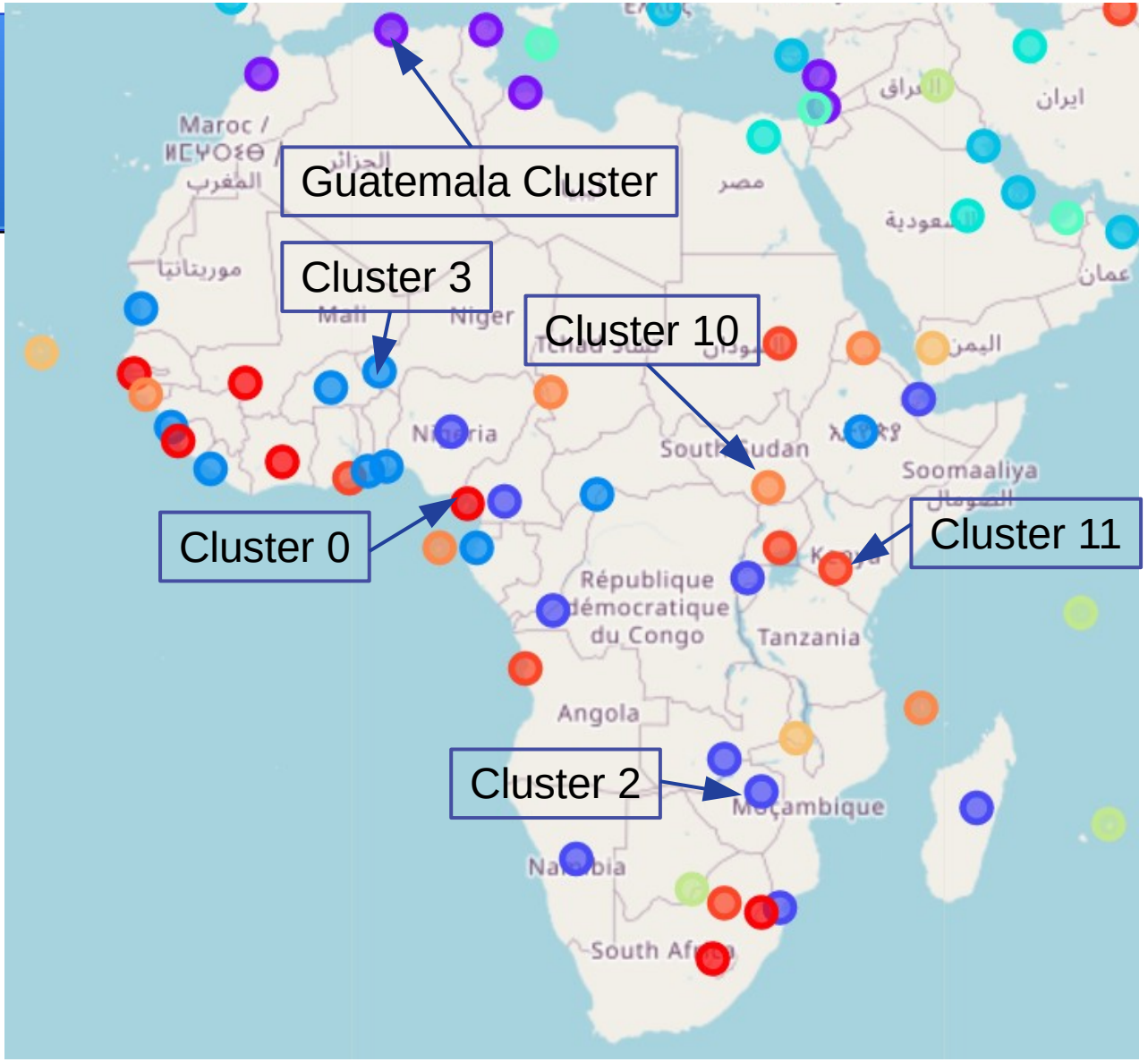
DISCUSSION
Africa clusters

Most numerous cluster in Africa are:

- Cluster 3: Niger
- Cluster 2: Mozambique
- Cluster 0: Equatorial Guinea
- Cluster 11: Kenya
- Cluster 10: South Sudan

Moreover, one may see Guatemala cluster in the north Africa – closest countries to Europe [*Morocco, Algeria, Tunisia, Libya*]

It seems to be clustered in a proper way.



DISCUSSION
Guatemala cluster

Capital Name	Country Name	Continent Name
Algiers	Algeria	Africa
Tripoli	Libya	Africa
Tunis	Tunisia	Africa
Rabat	Morocco	Africa
Amman	Jordan	Asia
Beirut	Lebanon	Asia
Bishkek	Kyrgyzstan	Asia
Tashkent	Uzbekistan	Asia
Manila	Philippines	Asia
Ulaanbaatar	Mongolia	Asia
Astana	Kazakhstan	Asia
Tegucigalpa	Honduras	Central America
San Salvador	El Salvador	Central America
Guatemala City	Guatemala	Central America
Managua	Nicaragua	Central America
Skopje	Macedonia	Europe
Sarajevo	Bosnia and Herzegovina	Europe
Tbilisi	Georgia	Europe
Yerevan	Armenia	Europe
Kyiv	Ukraine	Europe
Chisinau	Moldova	Europe
Baku	Azerbaijan	Europe
Saint John's	Antigua and Barbuda	North America
Port of Spain	Trinidad and Tobago	North America
Santo Domingo	Dominican Republic	North America
Kingston	Jamaica	North America
Caracas	Venezuela	South America
Brasilia	Brazil	South America
Asuncion	Paraguay	South America
Quito	Ecuador	South America
Paramaribo	Suriname	South America

Guatemala cluster is the biggest cluster. It seems to be not the best one.

However, if you look closer and focus on the cities from the same continent, you may see that it looks quite fine.

Moreover, it is quite possible that all those cities are somehow similar to each other not in the obvious way.

CONCLUSION

Analysis presented in this presentation shows that it is possible to have a helpful travel advisory tool.

Of course a real implementation would have focus on bigger number of cities across the world.

I believe that having unlimited access to Foursquare data as well as to some other available data such as hotel prices etc., one would be able to develop much better travel advisory tool.