WORLD CAPITALS: HOW DIFFERENT ARE WE?

Introduction 1
Data 1
Methodology 3
Results 7
Discussion 7
Conclusion 8

Introduction

The idea is to get data on most of the World's Capitals, namely coordinates, population and venues and then create some clusters and try to find out how different we really are around the World.

I think this topic is interesting not only from a social-cultural point of view, but also for anyone planning to visit or starting a business in a different country, because this analysis will generate clusters of similar countries and also the type of businesses that have more success in each country/cluster, at least when it comes to businesses that show up on social networks.

Data

I found a list of cities with their coordinates at https://simplemaps.com/data/world-cities.

There are thousands of cities in the list, so I'll just work with official capitals.

Here's a sample of the content (filtered by capital equal to "primary"):

city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
Tokyo	Tokyo	35,685	139,7514	Japan	JP	JPN	Tōkyō	primary	35676000	1392685764
Mexico City	Mexico City	19,4424	-99,131	Mexico	MX	MEX	Ciudad de México	primary	19028000	1484247881
Dhaka	Dhaka	23,7231	90,4086	Bangladesh	BD	BGD	Dhaka	primary	12797394	1050529279
Buenos Aires	Buenos Aires	-34,6025	-58,3975	Argentina	AR	ARG	Buenos Aires, Ciud	primary	12795000	1032717330
Cairo	Cairo	30,05	31,25	Egypt	EG	EGY	Al Qāhirah	primary	11893000	1818253931
Beijing	Beijing	39,9289	116,3883	China	CN	CHN	Beijing	primary	11106000	1156228865
Manila	Manila	14,6042	120,9822	Philippines	PH	PHL	Manila	primary	11100000	1608618140
Moscow	Moscow	55,7522	37,6155	Russia	RU	RUS	Moskva	primary	10452000	1643318494
Paris	Paris	48,8667	2,3333	France	FR	FRA	Île-de-France	primary	9904000	1250015082
Seoul	Seoul	37,5663	126,9997	Korea, South	KR	KOR	Seoul	primary	9796000	1410836482
Jakarta	Jakarta	-6,1744	106,8294	Indonesia	ID	IDN	Jakarta	primary	9125000	1360771077
London	London	51,5	-0,1167	United Kingdom	GB	GBR	London, City of	primary	8567000	1826645935

From this list I want to keep the following columns:

- city_ascii, readable City name (I will rename the column to Capital)
- lat, lng, the coordinates that I will use to get venues from FourSquare
- country, the Country that the City belongs to (I will rename the column to Country)
- population, the City's population

Since population values are off for several cities, I went to Wikipedia and found a list of reasonably updated population values for the World's Capitals here: https://en.wikipedia.org/wiki/List_of_national_capitals_by_population

Here's a sample of the content:

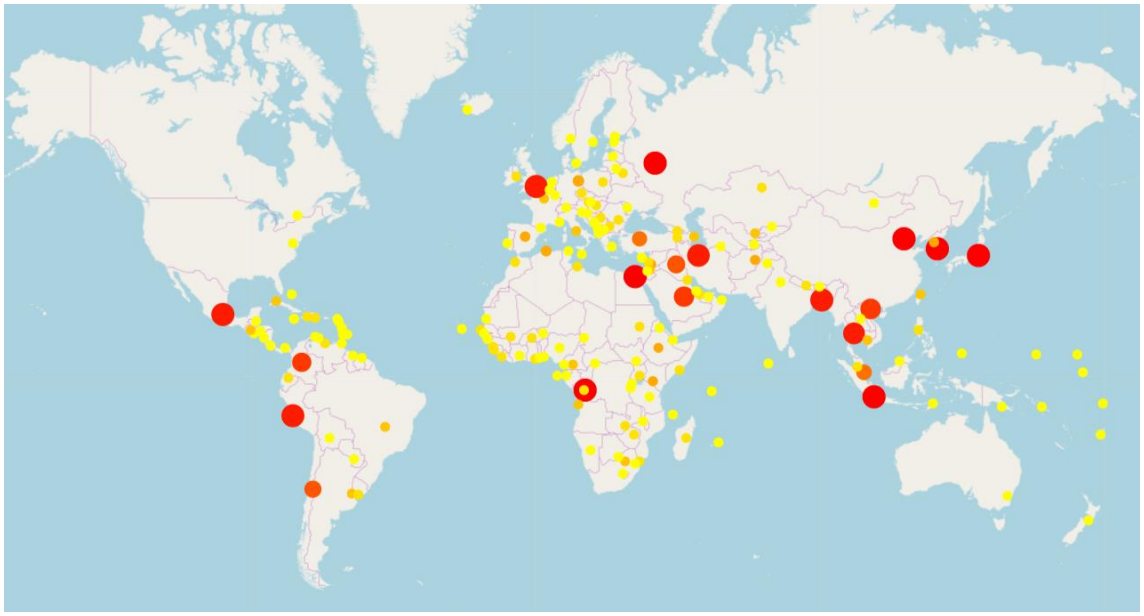
Rank	Country/Territory	Capital	Population	Year	% of country's population
1	China PR	Beijing	21,542,000[1]	2010	1.5%
2	Japan	Tokyo	13,929,286[2]	2017	11.03%
3	Russia	Moscow	12,506,468[3]	2011	8.52%
4	DR Congo	Kinshasa	11,855,000[4]	2012	12.9%
5	Indonesia	Jakarta	10,075,310[5]	2011	3.76%
6	South Korea	Seoul	9,838,892[6]	2015	19.03%
7	Egypt	Cairo	9,500,000	2012	9.54%
8	Mexico	Mexico City	8,918,653[7]	2015	7.05%
9	United Kingdom *England*	London	8,908,081[8]	2015	13.19%
10	Bangladesh	Dhaka	8,906,039 [9]	2011	5.52%
11	Peru	Lima	8,852,000[10]	2012	26.74%
12	Iran	Tehran	8,693,706	2014	10.53%

From this list I want to keep the following columns:

- Capital, used to merge the two tables
- Population, the Capital's population

Both columns required some processing due to additional content like references and also special characters.

Finally I merged the two tables and got a couple of hundred Capitals that have both the coordinates and the population, and I decided to proceed with that. The following map depicts the Capitals that we'll be working with, with colors ranging from yellow to red as population grows, and also bigger circles.



Now I can go on and fetch from Foursquare the venues around those Capitals. I'm using the "explore" request to obtain the recommended venues. Since the population is quite different from Capital to Capital, I'm defining a different radius for each of them, starting at 5000 meters and then adding 500 meters for each 1 million people.

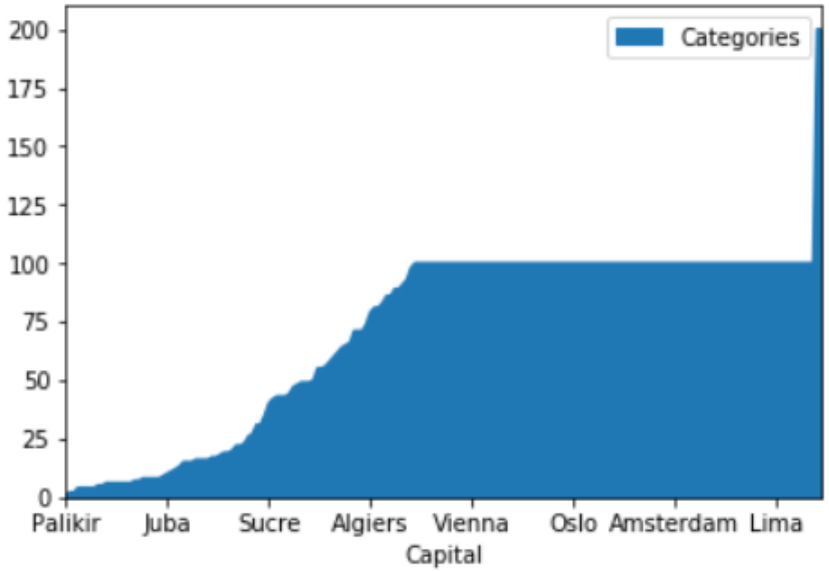
The query returned over 13 thousand venues and over 500 categories, here's a sample of the content:

	Capital	Venue	lat	lng	Category
3387	Caracas	Kabuki Sushi + Salads La Campiña	10.497430	-66.873549	Sushi Restaurant
13222	Philipsburg	Beau Beau's Restaurant	18.052367	-63.015744	Caribbean Restaurant
7308	Yerevan	History Museum of Armenia | Հայաստանի Պատմութ...	40.178449	44.513588	History Museum
4072	Santo Domingo	Don Nestor Parrillada	18.477242	-69.883639	Steakhouse
11373	Windhoek	Craft Cafe	-22.572002	17.083820	Café
7808	Dublin	Mad Egg	53.333668	-6.264568	Fried Chicken Joint
732	Moscow	Клуб Алексея Козлова	55.757724	37.633843	Music Venue
6211	San Salvador	i-shi cha	13.678636	-89.238081	Bubble Tea Shop
7917	Amsterdam	Generator Amsterdam	52.360802	4.918966	Hostel
12466	Luxembourg	Chemin de la Corniche	49.610389	6.134496	Trail
8414	Kingston	Shipme (Exec Direct Aviation)	17.997328	-76.787991	Shipping Store
7143	Helsinki	Kaffecentralen	60.167580	24.932526	Coffee Shop
4738	Beirut	Kampai (كامباي)	33.899152	35.500536	Japanese Restaurant
9459	Ashgabat	Köpetdag Club	37.929570	58.412445	Nightclub
8627	Managua	Ruta Maya	12.148480	-86.286591	Food
843	Paris	Papacionu	48.874792	2.342917	Corsican Restaurant
225	Dhaka	Baburchi Restaurant	23.740544	90.375109	Restaurant
10024	Manama	Fraser Suites Diplomatic Area	26.242840	50.589755	Hotel
12780	Andorra la Vella	Tequilando	42.541667	1.518174	Mexican Restaurant
4316	Accra	Kwame Nkrumah Memorial Park	5.544634	-0.203227	Park

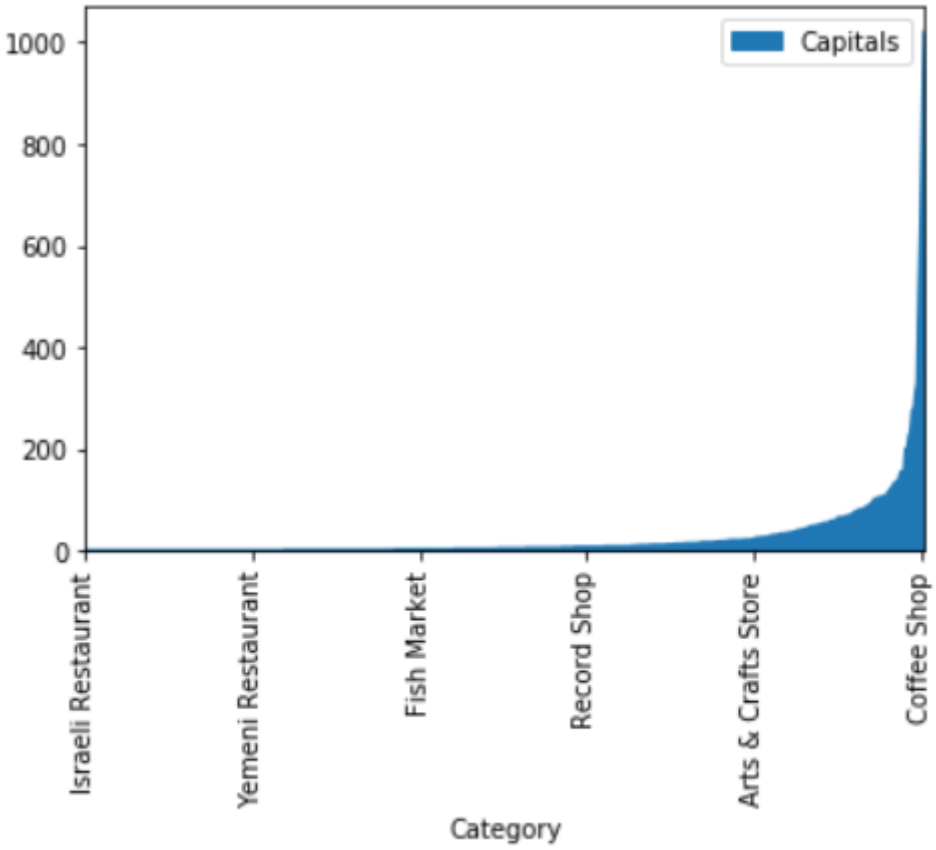
We should now have enough data to proceed to the analysis.

Methodology

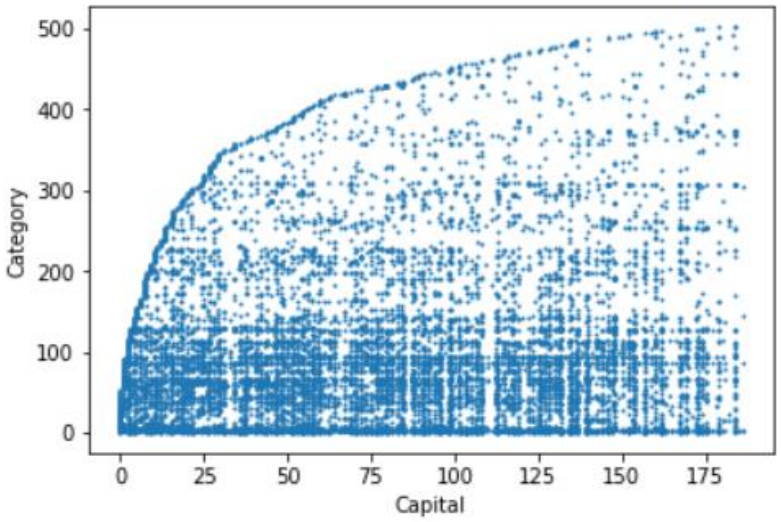
Let's start by plotting Capitals and categories against each other. Starting with categories per Capital, we can see that we have at least a handful of categories in all Capitals, with half of them in the hundreds (Foursquare has a limit of 100, that for some reason was ignored for a couple of queries).



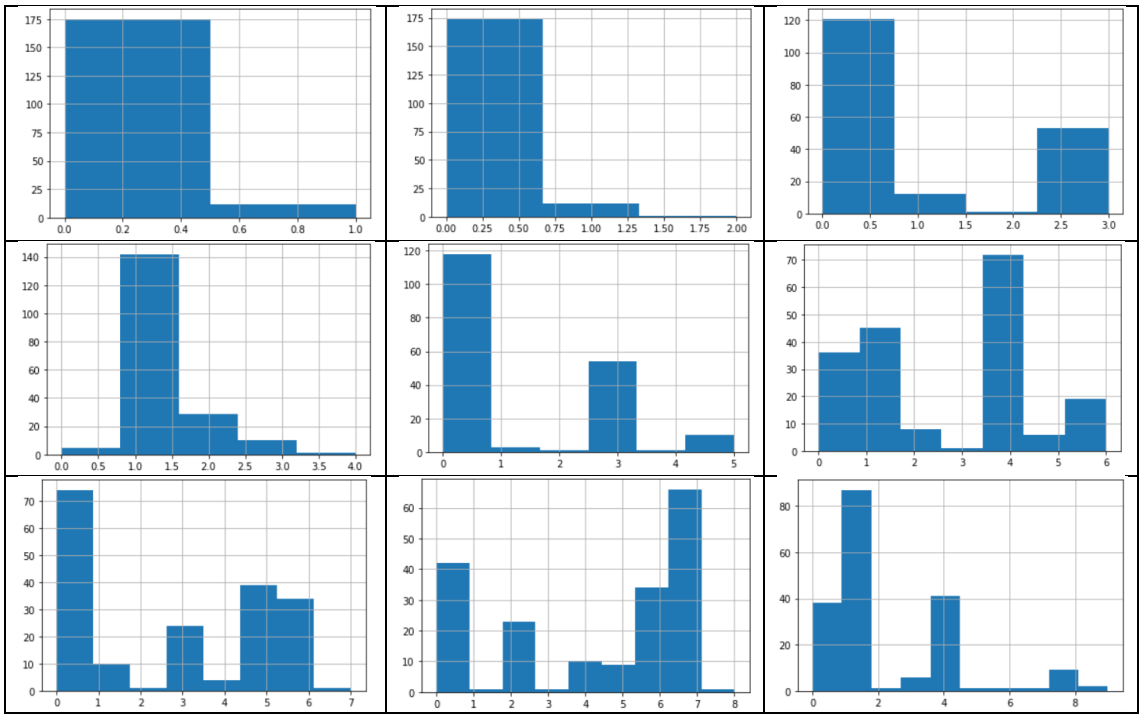
Now Capitals per category, we can see that most categories don't show up very frequently, with an average under 1 venue per Capital (recall that we have close to 200 Capitals), but a few of the categories are quite common, including multiple venues in the same Capital, with Coffee Shops leading the pack.



Finally a scatter plot of Capitals and categories (the actual names were factorized, the numbers aren't important, we just want an idea of the spread). We can see that the data is reasonably spread out, without any obvious tendencies so far.

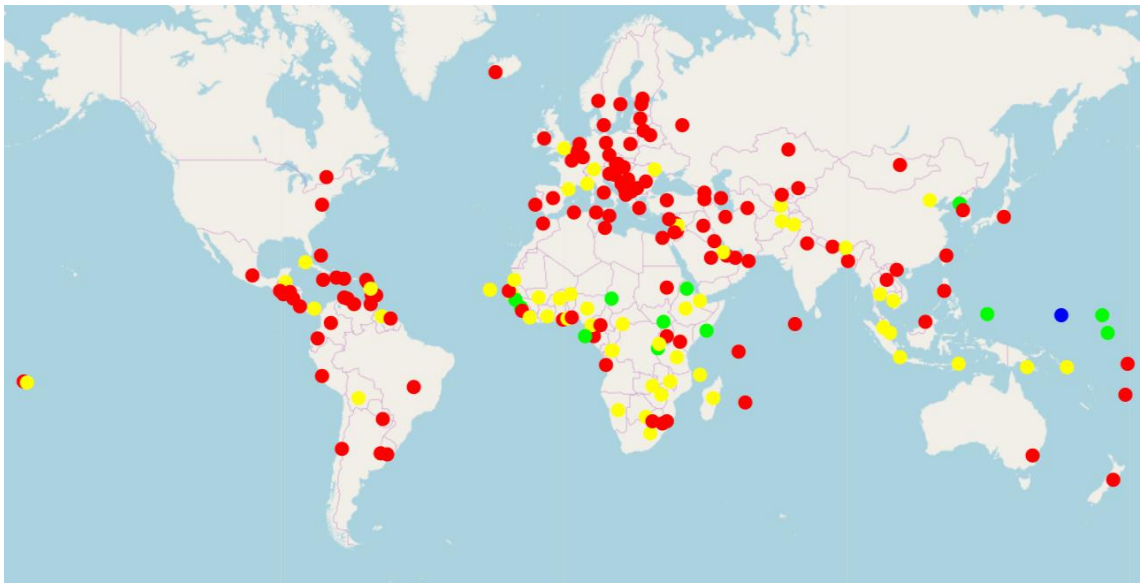


Next step is to group the Capitals based on the occurrences of the categories. I used the K-Means algorithm for that, ran it for values between 2 and 10 clusters, and got these distributions of Capitals per cluster (number of clusters grows left to right, top to bottom):



Finally I looked at the results for some of the cases, I selected the situations where we have 4, 7 and 10 clusters.

For 4 clusters, the results plotted on a World map give the following:

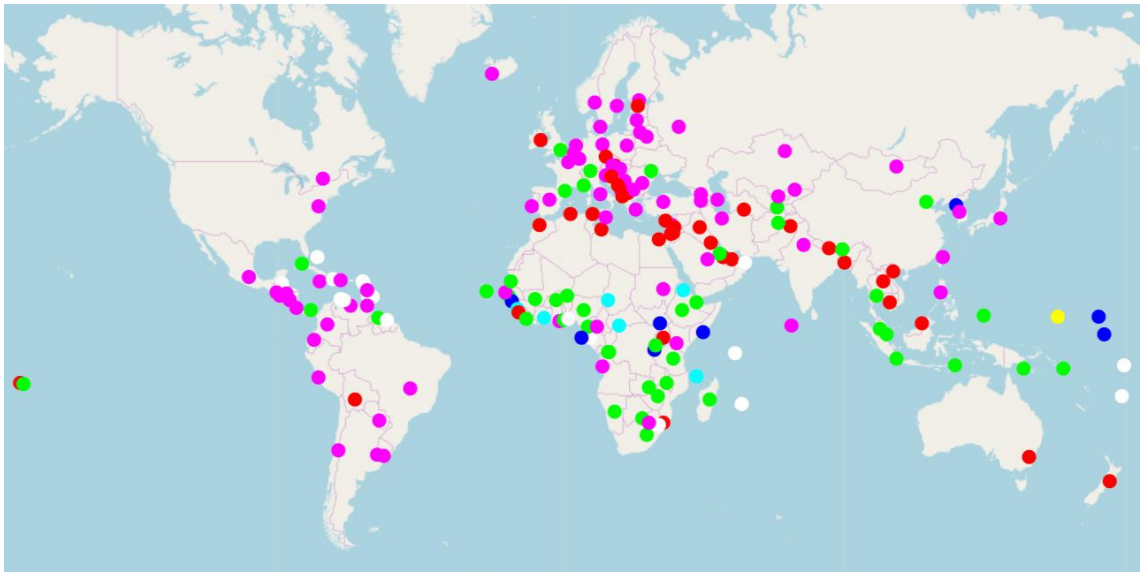


The most common categories in the clusters are as follows:

Cluster	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
0	Café	Hotel	Coffee Shop	Restaurant	Bar	Bakery	Park	Italian Restaurant	Pizza Place	Plaza
1	Hotel	Beach	Café	Bar	Plaza	Hotel Bar	Airport	French Restaurant	Brewery	National Park
2	Dive Shop	Zoo	Fish & Chips Shop	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant
3	Hotel	Café	Restaurant	Shopping Mall	Coffee Shop	Italian Restaurant	Pizza Place	French Restaurant	Resort	Airport

Please use the following color code to map cluster colors and numbers in the images:

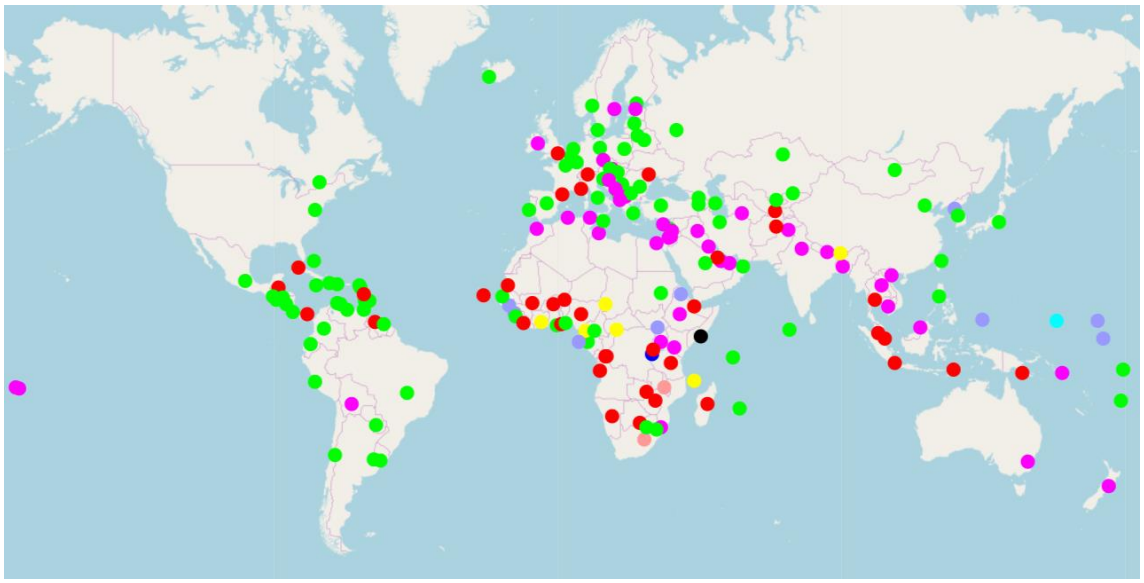
For 7 clusters, the results plotted on a World map give the following:



The most common categories in the clusters are as follows:

Cluster	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
0	Café	Hotel	Coffee Shop	Restaurant	Italian Restaurant	Bakery	Bar	Pizza Place	Park	Burger Joint
1	Hotel	Café	Restaurant	Shopping Mall	Coffee Shop	Italian Restaurant	Pizza Place	Fast Food Restaurant	French Restaurant	Grocery Store
2	Hotel	Beach	Bar	Hotel Bar	Plaza	Café	Restaurant	Noodle House	Motel	Shopping Mall
3	Dive Shop	Zoo	Fish & Chips Shop	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant
4	Hotel	Coffee Shop	Café	Restaurant	Park	Bar	Bakery	Plaza	Italian Restaurant	Pizza Place
5	Hotel	Resort	Airport	African Restaurant	French Restaurant	Hotel Pool	Movie Theater	Lake	Museum	Electronics Store
6	Hotel	Fast Food Restaurant	Chinese Restaurant	Beach	Shopping Mall	Restaurant	Bar	Caribbean Restaurant	Resort	Seafood Restaurant

For 10 clusters, the results plotted on a World map give the following:



The most common categories in the clusters are as follows:

Cluster	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category	6th Most Common Category	7th Most Common Category	8th Most Common Category	9th Most Common Category	10th Most Common Category
0	Hotel	Restaurant	Café	Shopping Mall	Pizza Place	Coffee Shop	Italian Restaurant	Fast Food Restaurant	Grocery Store	Bar
1	Hotel	Coffee Shop	Café	Restaurant	Bar	Park	Bakery	Italian Restaurant	Plaza	Ice Cream Shop
2	Hotel	Zoo	Eastern European Restaurant	Electronics Store	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop
3	Hotel	Resort	Airport	French Restaurant	African Restaurant	Hotel Pool	Café	Museum	Lake	Irish Pub
4	Café	Hotel	Coffee Shop	Restaurant	Italian Restaurant	Bar	Pizza Place	Bakery	Park	Plaza
5	Dive Shop	Zoo	Fish & Chips Shop	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant
6	Harbor / Marina	Hotel	River	Clothing Store	Nature Preserve	Fabric Shop	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant
7	Beach	Hotel	Zoo	Egyptian Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop
8	Hotel	Café	Bar	Plaza	Hotel Bar	National Park	Brewery	Restaurant	Movie Theater	Noodle House
9	Shopping Mall	Hotel	Restaurant	Border Crossing	Gas Station	Café	Supermarket	Pub	Hotel Bar	Indian Restaurant

Results

Looking at the results we can see that Europe and America always have lots of similarities, with many of the Capitals showing up the in the same cluster, either with a few clusters or with a lot of them.

Africa seems to always have its own cluster, even as the number of clusters increases.

Asia and Oceania are aligned with the rest of the continents when a small number of cluster is calculated, but, as the number of clusters increases, multiple specific clusters start appearing, denoting higher cultural differences.

Discussion

Looking at the tables of the most common categories it is obvious that the Capitals are not that different, with the primary venues focusing on lodging and food, namely Hotels and restaurants or similar. The biggest differences come from the specific type of cuisine, for example, denoting that it is in the cultural differentiation that any new business should probably invest.

Conclusion

Although there are some associations that should be taken into account, the comparison of the World Capitals shows that venues are not that different around the World, and that it is probably in the different cultural experiences that one should bet when considering a similar business to those included in the analysis.