WORLD CAPITALS: HOW DIFFERENT ARE WE?

Introduction

(Discuss the business problem and who would be interested in this project)

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.

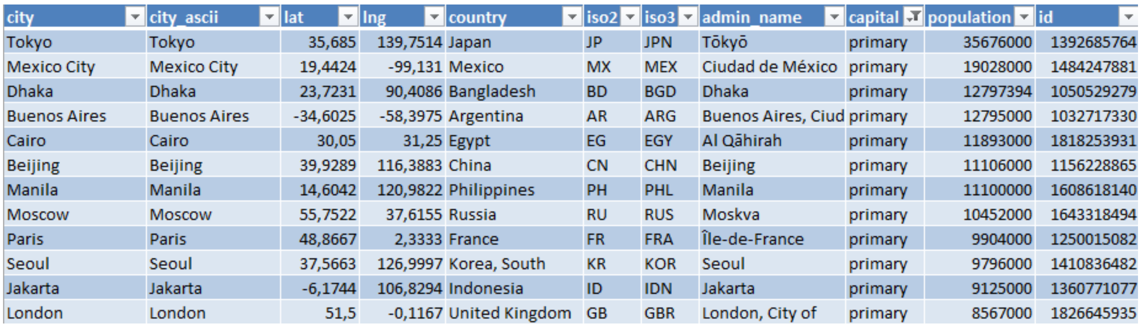
Data

(Describe the data that will be used to solve the problem and the source of the 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"):



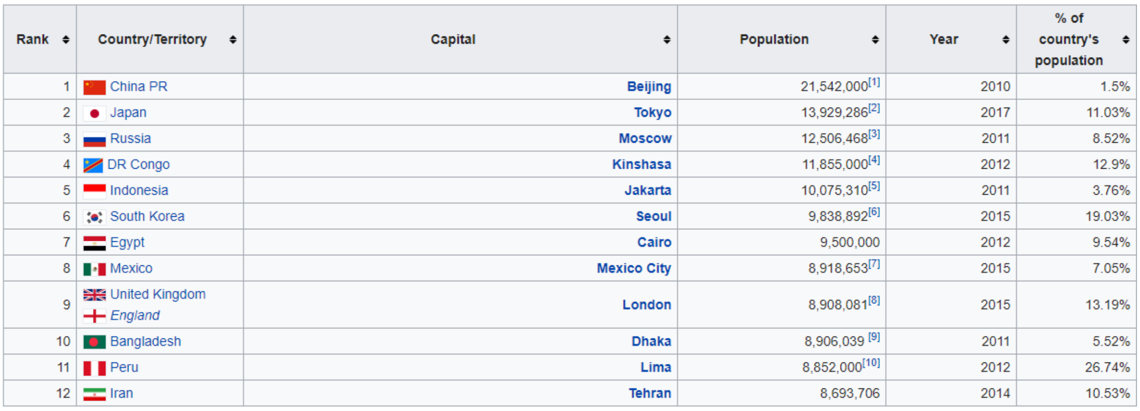
city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
Tokyo	Tokyo	35,685	139,7514	Japan	JP	JPN	Tōkyō	primary	35676000	1392685764
Mexico City	Mexico City	19,4424	-99,131	Mexico	MX	MEX	Ciudad de México	primary	19028000	1484247881
Dhaka	Dhaka	23,7231	90,4086	Bangladesh	BD	BGD	Dhaka	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:

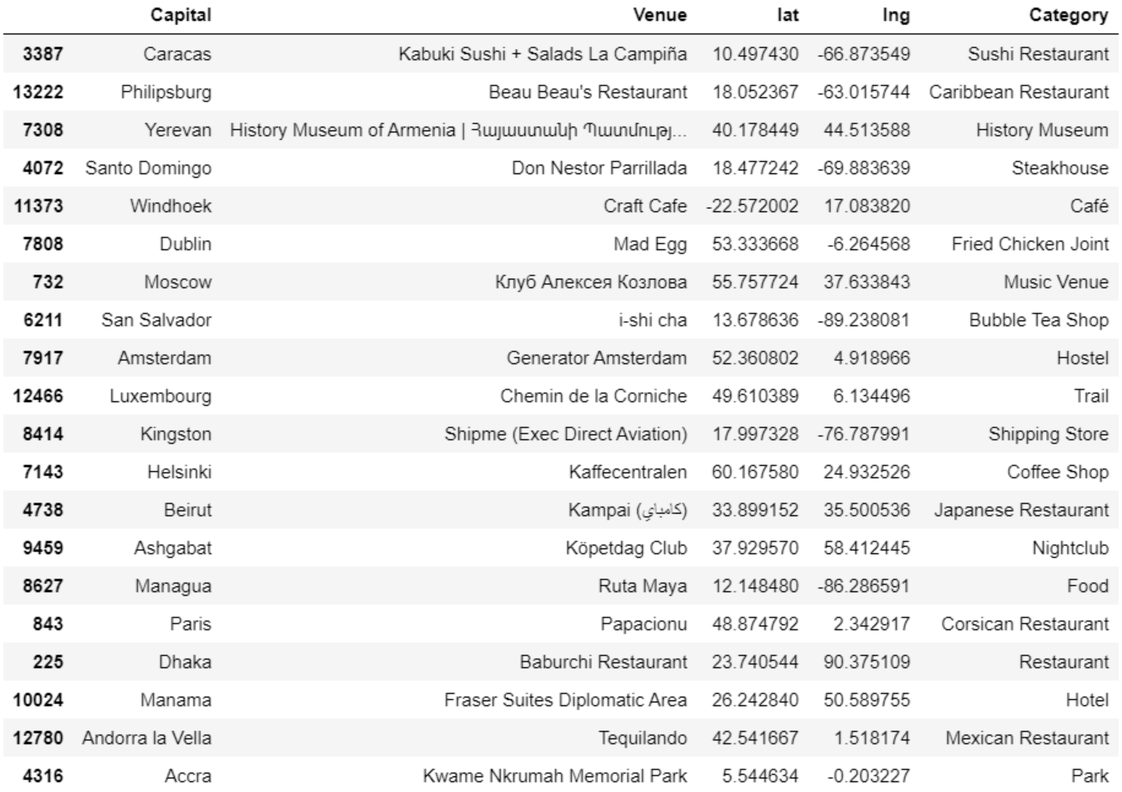
- Capital, used to merge the two tables
- Population, the Capital's population

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 with both the coordinates and the population, and I decided to proceed with that.

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 Պատմությ...	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

Methodology

(Discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why)

Results

(Discuss the results)

Discussion

(Discuss any observations you noted and any recommendations you can make based on the results)

Conclusion

(Conclude the report)