SEGEMENTING THE 100 MOST VISITED CITIES IN THE WORLD

1- BUSINESS PROBLEM:

Each year, A huge number of tourists travel around the world to discover new places. Doing so, they generate lots of money and stakeholders might want to benefit from that tourism. That's why it might be interesting for them to know what makes a city more attractive than another one or if there are any relationships between the most attractive cities in the world. For example, someone owning a business in tourism might want to extend his business in a city similar to the one where he currently works, hoping that would give him the same opportunities and benefits. Also, such a knowledge can benefit not only the stakeholders but also the tourists. Indeed, a tourist might have once travelled to a city and liked it. Therefore, for his next vacation, he would like to go to a similar place.

Consequently, if we try to resolve such a problem, we remark that we can group the most visited cities in the world into clusters and study what features they share. So as a junior data scientist, we will try to uncover why the 100 most visited cities are so attractive to tourists and what do those cities have in common

2- DATA DESCRIPTION:

To tackle this problem, we used the following datasets:

- A ranking of the most visited cities in the world that we found on this Wikipedia page: https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors. We scraped the web page using BeautifulSoup. However, we only considered the data of Euromonitor as they were the most accurate.
- To get the location of each city, we used the geocoder library
- And finally, we used Foursquare to get the 100 top venues of each city

3- METHODOLOGY:

The main purpose of everything we will be doing in this part will be to get the 200 most popular venues of each city and cluster them based on the similarities between those venues.

-DATA ACQUISITION AND WRANGLING:

As we already said, to get the data we need, using BeautifulSoup, we scraped the following Wikipedia page " https://en.wikipedia.org/wiki/List_of_cities_by_international_visitors' " to create a table containing the 100 most visited cities in the world. The table has the following columns :

**City: which is one of the 100 cities*

**Country: the country to which the city belongs to*

**Arrivals 2018: the number of arrivals in the city during 2018*

Here are the five first rows of the table:

[3]:

	City	Country	Arrivals 2018
0	Hong Kong	Hong Kong	29,262,700
1	Bangkok	Thailand	24,177,500
2	London	United Kingdom	19,233,000
3	Macau	Macau	18,931,400
4	Singapore	Singapore	18,551,200

Afterwards we needed the location(latitude, longitude) of each city because we would then need to place a CircleMarker on each city to analyze how they are spread around the world and get the 200 most popular venues of each city. So, to the old dataframe, two columns were added. Here is a picture showing the first five rows:

[7]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018
0	Hong Kong	22.279328	114.162813	Hong Kong	29,262,700
1	Bangkok	13.754253	100.493087	Thailand	24,177,500
2	London	51.507322	-0.127647	United Kingdom	19,233,000
3	Macau	22.175760	113.551414	Macau	18,931,400
4	Singapore	1.357107	103.819499	Singapore	18,551,200

As we have the coordinates of each city, it might be interesting to visualize how the most famous cities are spread around the world before we go further on wrangling data. Using folium, we made a map, placing a CircleMarker on each city. Here is the map:

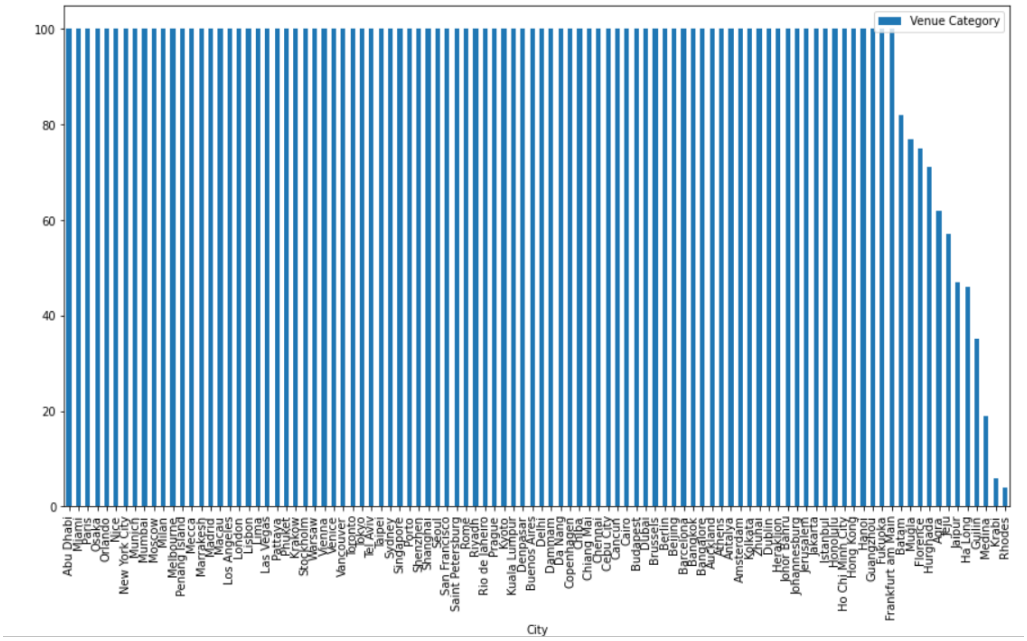
From now on, using the FourSquare API, we will retrieve for each city the 100 most popular venues, within a radius of 10km. Then we will make a new DataFrame containing for each city those venues

and their types as well as their location. Here is the first five rows of that DataFrame called: “world_top_venues”.

[13]:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hong Kong	22.279328	114.162813	Hong Kong Park Aviary (香港公園觀鳥園)	22.277140	114.161399	Zoo
1	Hong Kong	22.279328	114.162813	Hong Kong Park (香港公園)	22.277700	114.161854	Park
2	Hong Kong	22.279328	114.162813	The Upper House (奕居)	22.277499	114.166323	Hotel
3	Hong Kong	22.279328	114.162813	The Asia Society Hong Kong Center (亞洲協會香港中心)	22.276141	114.165263	Non-Profit
4	Hong Kong	22.279328	114.162813	The Murray Hong Kong (香港美利酒店)	22.278127	114.160392	Hotel

It is important to get the count of venues returned by the Foursquare API for each venues because not having enough might spoil the quality of the clustering. The following histogram shows the count of venues for each city.



We see that not 100 venues were retrieved for each city. So, to prevent the cities with low count of venues retrieved, we will drop every city where the count is below 10.

After dropping the countries that can spoil our future clustering, we use the “one hot encoding” to put the data in a form that will be suitable for clustering. That DataFrame is called “world_onehot”.

To get insights of the “world_onehot” Dataframe we will get the 15 most appreciated venues category. Here is the DataFrame showing them:

[17]:

	Venue Category	Count
222	Hotel	671
100	Coffee Shop	535
73	Café	408
337	Park	403
28	Bakery	240
228	Ice Cream Shop	202
355	Plaza	187
241	Italian Restaurant	173
374	Restaurant	162
213	Historic Site	133
351	Pizza Place	131
30	Bar	123
229	Indian Restaurant	111
397	Shopping Mall	106
129	Dessert Shop	102

In the most visited cities in the world, it's normal that hotels are the most common places. So, we think that including them in our clustering will only make it as it can make the algorithm find what truly make them similar. Café and Coffee Shop are also very common in big cities, so we will drop the following three categories:

**Hotel*

**Coffee Shop*

**Café*

Now we are ready to get the 10 most common Venue of each city. The DataFrame containing that information will be called "city_venues_sorted". And here is the head of that DataFrame:

]:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abu Dhabi	Middle Eastern Restaurant	Beach	Restaurant	Chinese Restaurant	Shopping Mall	Park	Steakhouse	Japanese Restaurant	Pizza Place	Resort
1	Agra	Indian Restaurant	Historic Site	Fast Food Restaurant	Resort	Multicuisine Indian Restaurant	Market	Pizza Place	Bed & Breakfast	Airport	Bistro
2	Amsterdam	Bar	Bakery	Bookstore	French Restaurant	Plaza	Park	Cocktail Bar	Restaurant	Ice Cream Shop	Wine Bar
3	Antalya	Seafood Restaurant	Restaurant	Gym	Park	Gym / Fitness Center	Scenic Lookout	Motorcycle Shop	Museum	Beach	Trail
4	Athens	Historic Site	Dessert Shop	Cocktail Bar	Souvlaki Shop	Falafel Restaurant	Meze Restaurant	History Museum	Theater	Boutique	Gourmet Shop

Here end our data wrangling section and our data is finally in a good format for clustering.

-ANALYSIS:

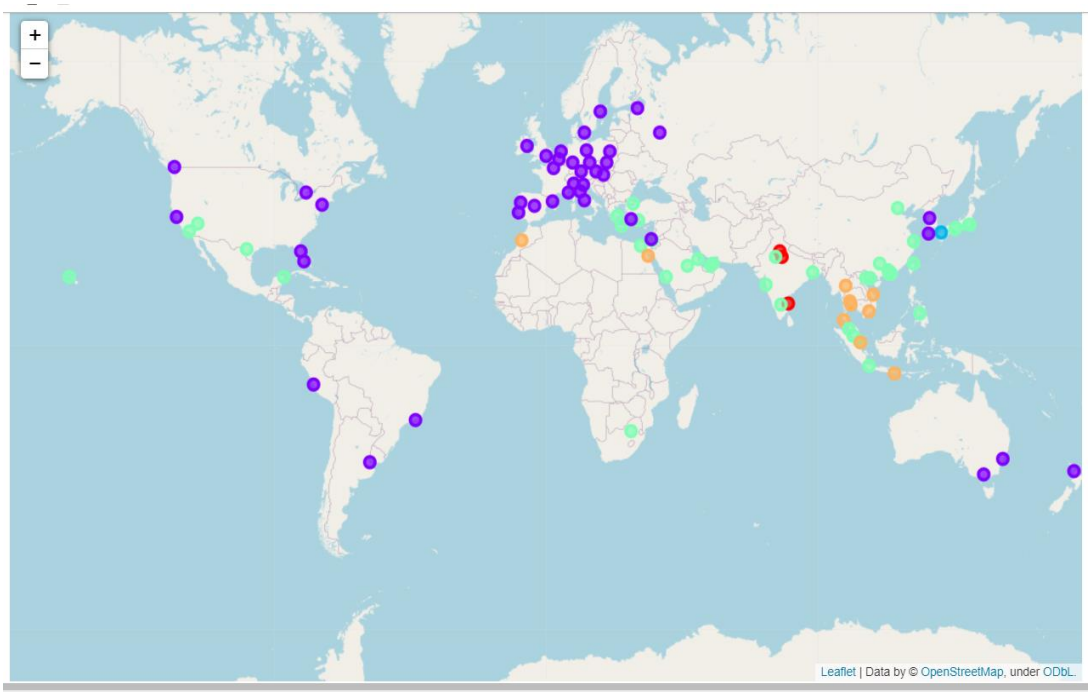
To analyze the data, we will use the K-means algorithm and we will divide the dataset into 5 clusters. We choose such an algorithm because it is the easiest clustering algorithm that can help to uncover the hidden similarities the cities share. Here is the head of the table were each city is assigned to its cluster.

}]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	
0	Hong Kong	22.279328	114.162813	Hong Kong	29,262,700	3	Japanese Restaurant	Italian Restaurant	Thai Restaurant	Steakhouse	Gym / Fitness Center	Bar	Cocktail Bar	Chinese Restaurant	
1	Bangkok	13.754253	100.493087	Thailand	24,177,500	4	Thai Restaurant	Dessert Shop	Palace	Shopping Mall	Noodle House	Park	Bookstore	Bar	F
2	London	51.507322	-0.127647	United Kingdom	19,233,000	1	Park	Garden	Grocery Store	Plaza	Theater	Cocktail Bar	Art Museum	Hotel Bar	
3	Macau	22.175761	113.551414	Macau	18,931,400	3	Portuguese Restaurant	Resort	Lounge	Chinese Restaurant	Cantonese Restaurant	Steakhouse	Historic Site	Italian Restaurant	F
4	Singapore	1.357107	103.819499	Singapore	18,551,200	3	Park	Shopping Mall	Chinese Restaurant	Cocktail Bar	Japanese Restaurant	Bakery	Indian Restaurant	Ice Cream Shop	Su

4- RESULTS

We finally clustered the most visited cities in the world and when placing a CircleMarker identifying each cluster, we get the following map:



If we look at the map, we remark that most of our cities are located near costal areas. In addition, cities from the European continent are mostly similar to those of the American continent as well as Japan. On the other hand, Asian, Arabic and countries from the Caribbean Region are most similar.

Now let's get a deeper analysis of the similarities between the clusters:

CLUSTER 1: (purple color)

```
[24]: cluster_1.head()
```

[24]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
2	London	51.507322	-0.127647	United Kingdom	19233000	1	Park	Garden	Grocery Store	Plaza	Theater	Cocktail Bar	Art Museum	Hotel Bar	Cheese Shop
5	Paris	48.856697	2.351462	France	17560200	1	Wine Bar	Plaza	Italian Restaurant	French Restaurant	Bookstore	Art Museum	Bakery	Garden	Creperie
7	New York City	40.712728	-74.006015	United States	13600000	1	Park	Ice Cream Shop	Bakery	Theater	Movie Theater	Scenic Lookout	Pier	Gourmet Shop	Bookstore
15	Rome	41.893320	12.482932	Italy	10065400	1	Plaza	Ice Cream Shop	Pizza Place	Italian Restaurant	Park	Historic Site	Sandwich Place	Church	Bakery
21	Prague	50.087465	14.421254	Czech Republic	8948600	1	Park	Ice Cream Shop	Scenic Lookout	Garden	Cocktail Bar	Burger Joint	Noodle House	Plaza	Bar

From the head of the DataFrame containing the cluster 1, we remark that what the countries share in common are **restaurant, Park, historic site, Gardens or Theaters.** That is typical to the western world.

CLUSTER 2: (blue color)

```
[25]: cluster_2.head()
```

[25]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
92	Fukuoka	33.625124	130.618002	Japan	2436900	2	Convenience Store	Ramen Restaurant	Train Station	Sushi Restaurant	Discount Store	Hot Spring	Ice Cream Shop	Noodle House	Outdoor Sculpture

The cluster 2 is composed of only one city, Fukoka which lies in Japan. The reason for this is that the specialities in Fukoka does not exist in the other cities. For example, the most common venues are **Sushi Restaurant or Ramen Restaurant.**

CLUSTER 3: (green color)

```
[26]: cluster_3.head()
```

[26]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Hong Kong	22.279328	114.162813	Hong Kong	29262700	3	Japanese Restaurant	Italian Restaurant	Thai Restaurant	Steakhouse	Gym / Fitness Center	Bar	Cocktail Bar	Chinese Restaurant
3	Macau	22.175761	113.551414	Macau	18931400	3	Portuguese Restaurant	Resort	Lounge	Chinese Restaurant	Cantonese Restaurant	Steakhouse	Historic Site	Italian Restaurant
4	Singapore	1.357107	103.819499	Singapore	18551200	3	Park	Shopping Mall	Chinese Restaurant	Cocktail Bar	Japanese Restaurant	Bakery	Indian Restaurant	Ice Cream Shop
6	Dubai	25.065964	55.171340	United Arab Emirates	15920700	3	Shopping Mall	Lounge	Restaurant	Resort	Supermarket	Beach	Food Truck	Multiplex
8	Kuala Lumpur	3.151696	101.694237	Malaysia	13434300	3	Shopping Mall	Hotel Bar	Spa	Malay Restaurant	Grocery Store	Cocktail Bar	Bar	Ice Cream Shop

In the third cluster, we remark that they also share in common Restaurant cooking specialities from the Western World and East Asian World. We find also venues like Beach and Historic Site.

CLUSTER 4: (orange color)

[27]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	
1	Bangkok	13.754253	100.493087	Thailand	24177500	4	Thai Restaurant	Dessert Shop	Palace	Shopping Mall	Noodle House	Park	Bookstore	Bar	R
14	Phuket	7.936602	98.352929	Thailand	10550700	4	Thai Restaurant	Resort	Ice Cream Shop	Dim Sum Restaurant	Department Store	Restaurant	Shrine	Bar	
17	Pattaya	12.931941	100.900953	Thailand	9606400	4	Thai Restaurant	Resort	Restaurant	Spa	Noodle House	Nightclub	Hotel Bar	Massage Studio	R
30	Ho Chi Minh City	10.775844	106.701755	Vietnam	7200000	4	Vietnamese Restaurant	French Restaurant	Vegetarian / Vegan Restaurant	Pizza Place	Thai Restaurant	Spa	Whisky Bar	Bar	R
31	Denpasar	-8.652497	115.219118	Indonesia	7185600	4	Indonesian Restaurant	Bakery	Restaurant	Resort	Chinese Restaurant	Ice Cream Shop	Convenience Store	Asian Restaurant	

In the Fourth cluster, we essentially find Restaurant cooking exclusively asian restaurant. There also places for entertainement.

CLUSTER 5: (red color)

[28]:
```
cluster_5.head()
```

[28]:

	City	CityLatitude	CityLongitude	Country	Arrivals 2018	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th M Comm Ve
10	Delhi	28.651718	77.221939	India	12645300	0	Indian Restaurant	Bar	Lounge	South Indian Restaurant	Park	Fast Food Restaurant	Restaurant	Tibetan Restaurant	Ita Restau
25	Agra	27.175255	78.009816	India	8138200	0	Indian Restaurant	Historic Site	Fast Food Restaurant	Resort	Multicuisine Indian Restaurant	Market	Pizza Place	Bed & Breakfast	Air
35	Chennai	13.080172	80.283833	India	6422800	0	Indian Restaurant	Italian Restaurant	Fast Food Restaurant	Multiplex	Juice Bar	Ice Cream Shop	Sandwich Place	Beach	Sr P

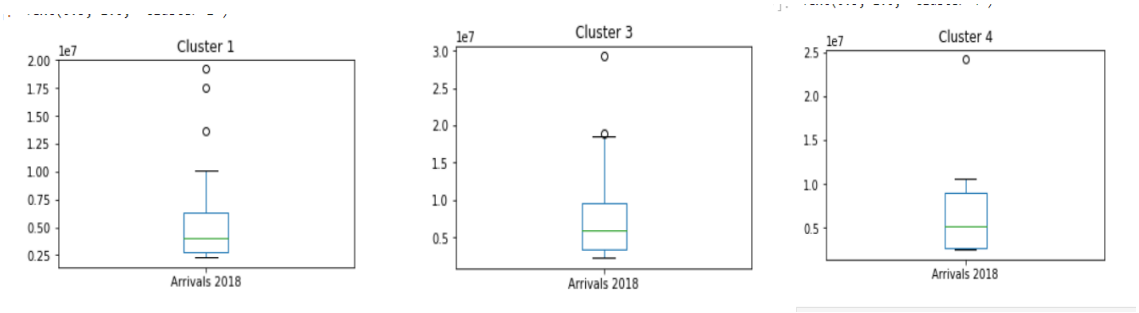
The fifth cluster contains only the indian countries. That's Because those countries aren't similar to any other country.

5- DISCUSSION

In resume, we can say that what the most visited cities in the world share in common are restaurant cooking the specialties of the country and places for entertainment like Historic Site, Parc and Zoos. So, if someone like a city in a given cluster, he will surely like another city belonging to that cluster.

However, for business purposes that might not always be the case. To prove it, we have plotted the box plot of each cluster given the number of persons that visits the cities and spotted that in cluster 1, 3 and 4, there are outliers to which the rules might not apply to.

Here are the box plots:



6- CONCLUSION

To conclude we can say that the most visited cities in the world have in common the fact that they are places where people entertain themselves. They are quite similar as they are places containing historic site or park and even zoos people want to discover. So, if someone ever like one place he might like another one that belongs to the same cluster. However for business purposes that might not always work.