Predicting where to go on vacation

Sergi

February 14, 2021

1. Introduction

1.1. Background

Holidays are something very important in people's lives. They move a large amount of people, and money. In 2019, more than a trillion dollars were spent on vacations, with an average of $ 1,536 for every household in the US.

But this not only shows the great interest that people have in vacations, but also shows how more and more sites are betting on a tourist economy, marketing themselves, to attract tourists and thus large amounts of money to boost the economy.

1.2. Problem

This is a problem for tourists, since there is more and more to choose from, and it is not easy to choose correctly between so many options, without wasting a lot of time. A time that most people cannot lose. That is why this project aims to provide a tool that helps to plan a vacation quickly, with a great guarantee of success, and without wasting time.

1.3. Interest

Obviously, a large number of people will be interested in such a tool, since as we have said, a large part of the population goes on vacation every year, investing a large sum of money, but has little time to plan them, and search among all possible destinations.

2. Data acquisition and cleaning

2.1. Data sources

For the project we will get the coordinates of the cities through the geocodes API, and the characteristics of the cities from the Foursquare API. We could directly enter the latitude and longitude of the city and ask Foursquare to return the most interesting sites that are nearby, but that would give us bad results.

Why? Well, because the API will return a maximum of 100 sites, but these are organized into more than 500 categories. This would create an underfitting problem in the data. To solve this, we will take two actions.

- The first will be to group all these characteristics into 177 subgroups. By doing this we will group the most similar sites in the same category. For example, we will not distinguish between pasta and pizza restaurants, but they will be simply Italian.

- The second, will be to ask the API how many places of each category there are in each city, so that it will not only return the closest or most important places, but we will be able to know how many Italian restaurants there are, how many Koreans, how many Americans, how many Mediterranean, ... Thus, asking specifically for each of the categories and taking into account that each category can return 50 sites, we will take into account thousands of sites, and not only the first 100.

Once we have all the data of all the cities, we will ask the user for cities that he has visited previously and what grade would he give them, and in this way, we will extract the user profile of this, and we will recommend the cities that best suit his tastes.

2.2. Feature selection

In order to classify the cities and evaluate them, we will have to choose which of their characteristics we will look at. As we have said, we will look at the number of places and places they have for each site. The sites will be organized into the following 100 categories:

- Aquarium
- Arcade & Bowling
- Casino
- Cinema
- Night club
- Disco
- Music
- Art
- Stadium
- Theme Park
- Water Park
- Zoo
- American Restaurant
- African Restaurant
- Italian Restaurant
- Asian Restaurant
- Bistro
- Buffet
- Cafeteria
- Creperie
- Bodega
- Fast Food
- Restaurant
- French Restaurant
- Indian Restaurant

- Irish Pub
- Italian restaurant
- Latin American Restaurant
- Mediterranean Restaurant
- Mexican Restaurant
- Seafood Restaurant
- Steakhouse
- Turkish Restaurant
- Nightlife Spot
- Bar
- Beach Bar
- Cocktail Bar
- Karaoke
- Pub
- Sport bar
- Brewery
- Lounge
- Nightclub
- Golf
- Bay
- Beach
- Surf spot
- Botanical Garden
- Bridge
- Canal
- Castle
- Dive Spot
- Field
- Farm
- Fishing spot
- Forest
- Garden
- Harbor
- Hill
- Island
- Lake
- Lighthouse
- Mountain
- National Park
- Park
- Pedestrian Aera
- Plaza
- River
- Ski Area
- Stables
- Vineyard
- Volcano
- Waterfall

- Windmill
- Government Building
- Library
- Observatory
- Office
- Social Club
- Spiritual Center
- Antique shop
- Arts Store
- Clothing Store
- Gift shop
- Massage Studio
- Music Store
- Outlet
- Airport
- Bike rental
- Boat rental
- Ferry or Boat
- Bus
- Hotel
- Resort
- Motel
- Hostel
- Vacation Rental
- Bed & Breakfast
- Metro station
- Pier
- RV park

2.3. Candidate cities

As possible candidates, we have chosen a total of the 100 most popular destinations to go on vacation, which are:

- Cairo
- Kusadasi, Turkey
- Chamonix
- Beijing
- Cannes
- Amsterdam
- Puerto del Rosario, Canary Islands
- Bodrum
- Iguazu National Park, Argentina
- Courchevel
- Berlin
- Aberdare
- Amritsar
- Edimburgh
- New York

- Orlando
- Sydney
- London
- Paris
- Venice
- Manhattan
- Cape Town
- Las Vegas
- Rome
- Rio de Janeiro
- Maldives
- Hawaii
- South Island, New Zealand
- Grand Canyon
- San Diego
- Niagara Falls
- San Francisco
- Los Angeles
- Dubai
- Auckland
- Singapore
- Seychelles
- Bali
- Durban
- Bangkok
- Iceland
- Whitsunday Islands National Park
- Cairns
- Costa del Sol
- Antigua
- Melbourne
- Mallorca
- Lake District
- Barbados
- Bahamas
- Abu Simbel
- Bora Bora
- Sharm el Sheikh
- Madrid
- Algarve
- Zermatt
- Victoria Falls
- Marbella
- Masai Mara, Kenya
- Chichen Itza
- Disney World
- Florence
- Puerto Banus

- Toronto
- Taj Mahal
- Great Wall of china
- Menorca
- Monaco
- Luxor
- Hong Kong
- Banff National Park
- Sorrento
- Key West
- Koh Samui, Thailand
- Cancun
- Nice
- Machu Picchu
- Yosemite
- Oahu
- Florida Keys
- Guam
- Dublin
- Vancouver
- Ayers Rock
- La Digue Island
- Cayman Islands
- Naples
- St. Pete Beach, Florida
- Barcelona
- Ibiza
- Adelaide
- Airlie Beach Queensland
- Benidorm
- Buenos Aires
- Prague
- Cuba
- Paphos
- Valley of the kings
- Galapagos Islands
- Isle of Man

We can plot them on the map:

3. Methodology

Now we have all the data for all the cities. In this project we will base ourselves on these data to find out which are the most similar cities to each other, and which have the attributes that the client likes the most, to recommend the best possible vacations.

To do this, we will look at the percentage of sites in each category that each city has. That is, we will look for each city what percentage of beach it has, what percentage of mountains, which Italian restaurants, ...

It is important to evaluate cities by percentages of each category, and not by the number of sites they have in each category, since otherwise large cities would always win. That is, if the client was a fan of Italy, and of the beach, they would surely like things like Italian restaurants, art, beaches, beach bars and music stores. However, a city such as Barcelona could have many more Italian restaurants, art galleries, music stores and beaches, and the program would recommend this city rather than a small city in Italy, which is what the user would prefer.

When using the percentages, although Barcelona still has many more places of those than for example Florence, Florence will be recommended much earlier, since the percentages of these things will be much higher than in Barcelona, where there are many Italian restaurants, but many more Mediterranean., Catalan or Spanish, and therefore Italian restaurants are overshadowed.

4. Analysis

4.1. Data preprocessing

The data taken from the API is put into a table where for each of the places we will see the city to which it belongs, the name of the place, the category, its latitude and its longitude. The table will be like the following:

	City	Category	Name	Latitude	Longitude
0	Aberdare	Cinema	Vue	51.738475	-3.377418
1	Aberdare	Night club	Aberdare Constitutional Club	51.713557	-3.447650
2	Aberdare	Night club	Aberdare Rugby Club	51.709252	-3.432933
3	Aberdare	Night club	Aberdare golf club	51.716892	-3.429162
4	Aberdare	Night club	Cwmdare Club	51.715940	-3.469329
...	...	...	...	...	...
150771	Zermatt	Metro station	Taxi Metro	46.068560	7.776482
150772	Zermatt	Metro station	Riffelalp Station	46.004908	7.753993
150773	Zermatt	Metro station	Bahnhof Zermatt	46.023864	7.748048
150774	Zermatt	Metro station	Green Motion Charging Station	46.067318	7.775392
150775	Zermatt	Metro station	Blauherd Station	46.017076	7.785827

The first thing we need to do is to order all the data in cities, so that we have a table, where each row is a city, and each column a category, and we have the number of sites in each category that are in each city.

Once we have it done, we have a table like the following, but with 100 columns, one for each feature:

	Aquarium	Arcade & Bowling	Casino	Cinema	Night club	Disco	Music	Art	Stadium	Theme Park
Aberdare	0	0	0	4	80	0	0	4	0	34
Abu Simbel	0	0	0	0	0	0	0	0	0	0
Adelaide	4	140	40	96	200	92	132	200	36	100
Airlie Beach Queensland	0	0	0	0	44	4	4	0	0	16
Algarve	0	0	0	4	200	16	4	52	0	28
Amritsar	0	0	0	20	24	0	0	16	0	20
Amsterdam	20	28	88	136	200	132	200	200	36	100
Antigua	0	0	0	0	28	0	0	88	4	4
Auckland	8	80	20	64	200	64	124	200	24	100
Ayers Rock	0	0	0	0	0	0	0	8	0	6
Bahamas	0	0	24	4	180	48	28	36	8	62
Bali	0	0	0	0	0	0	0	4	4	0
Banff National Park	0	0	0	4	24	8	8	20	0	42
Bangkok	56	52	48	200	200	200	200	200	200	100
Barbados	0	0	0	0	12	4	0	12	0	6

The next step is to see if there are any cities that should be discarded because it has few sites. This may be because Foursquare has little data about a site, because we put the coordinates wrong (for example, we put the name of a country or region instead of a city) or that the place simply has few sites. An example of a misnomer is for example the place "Hawaii". Hawaii is well known for being one of the most famous states of USA, and it is highly touristic. However, when looking for the coordinates in geocoders, putting the name of Hawaii instead of the name of its capital (Honolulu) returned us a location far from any city.

Keep in mind that we set a 5km radius limit (the red circumference), and that is why Foursquare did not give us a good number of places that allow us to determine what type of destination it is. We could have put a bigger radius, but that would make the number of sites returned in big cities too large.

We then look at all cities that have less than 100 places (our minimum threshold) and eliminate them. The cities that remain after removing the invalid ones are:

Aberdare, Adelaide, Airlie Beach Queensland, Algarve, Amritsar, Amsterdam, Antigua, Auckland, Ayers Rock, Bahamas, Bali, Banff National Park, Bangkok, Barbados, Barcelona, Beijing, Benidorm, Berlin, Bodrum, Bora Bora, Buenos Aires, Cairns, Cairo, Cancun, Cannes, Cape Town, Chamonix, Chichen Itza, Costa del Sol, Courchevel, Disney World, Dubai, Dublin, Durban, Edimburgh, Florence, Florida Keys, Great Wall of china, Guam, Hong Kong, Ibiza, Isle of Man, Key West, Koh Samui, Thailand, Kusadasi, Turkey, La Digue Island, Las Vegas, London, Los Angeles, Luxor, Machu Picchu, Madrid, Maldives, Mallorca, Marbella, Melbourne, Monaco, Naples, New York, Nice, Orlando, Paphos, Paris, Prague, Puerto del Rosario, Canary Islands, Rio de Janeiro, San Diego, San Francisco, Sharm el Sheikh, Singapore, Sorrento, St. Pete Beach, Florida, Sydney, Taj Mahal, Valley of the kings, Vancouver, Victoria Falls and Zermatt.

Next, we have to normalize our data. To normalize the data, what we will do is find the percentage of sites in each category that there are. For example, we will look at the total number of places found, what percentage are beaches, which mountains, which restaurants, ... If instead of normalizing doing the percentage, we would normalize making the maximum number is 1 and the rest the proportional part (for example, if there are 50 beaches and 25 mountains, the number in the beach category is 1 and 0.5 in the mountains category), because the algorithm seeks to maximize what the user prefers, the algorithm will determine what the user likes the most, and will look for the city that has the most in that category. If we did not do the percentages, the big cities would always win, because they are the ones that have the most things, when in truth what we want is not to find a city with many things, but a city of the same style as the ones that the user likes. Then, making the percentage, if the user wants a city that is 70% beach, 5% Italian restaurants, 10% resorts and 15% French restaurants, the algorithm will search for a city similar to this in percentages, and not for example, a city like Barcelona, which may have many more beaches, Italian and French restaurants, and hotels and resorts, but it will not look anything like the city entered by the user.

	Aquarium	Arcade & Bowling	Casino	Cinema	Night club	Disco	Music	Art	Stadium	Theme Park	Water Park
Aberdare	0.000000	0.000000	0.000000	0.002730	0.054608	0.000000	0.000000	0.002730	0.000000	0.023208	0.0
Adelaide	0.000390	0.013643	0.003898	0.009355	0.019489	0.008965	0.012863	0.019489	0.003508	0.009745	0.0
Airlie Beach Queensland	0.000000	0.000000	0.000000	0.000000	0.023170	0.002106	0.002106	0.000000	0.000000	0.008425	0.0
Algarve	0.000000	0.000000	0.000000	0.000816	0.040800	0.003264	0.000816	0.010608	0.000000	0.005712	0.0
Amritsar	0.000000	0.000000	0.000000	0.008617	0.010340	0.000000	0.000000	0.006894	0.000000	0.008617	0.0
Amsterdam	0.001605	0.002247	0.007061	0.010913	0.016049	0.010592	0.016049	0.016049	0.002889	0.008024	0.0
Antigua	0.000000	0.000000	0.000000	0.000000	0.007646	0.000000	0.000000	0.024031	0.001092	0.001092	0.0
Auckland	0.000748	0.007478	0.001870	0.005982	0.018695	0.005982	0.011591	0.018695	0.002243	0.009348	0.0
Ayers Rock	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.023599	0.000000	0.017699	0.0
Bahamas	0.000000	0.000000	0.003306	0.000551	0.024793	0.006612	0.003857	0.004959	0.001102	0.008540	0.0
Bali	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.027397	0.027397	0.000000	0.0
Banff National Park	0.000000	0.000000	0.000000	0.001401	0.008403	0.002801	0.002801	0.007003	0.000000	0.014706	0.0

We can see how now the table no longer shows integer values indicating the number of establishments in each category, but we see the percentage of establishments in each category, with 1 being the total.

4.2. Custom client part

The next thing is defining a user's score to previous cities to see where their next vacation should be.

For this we will invent a user, whose scores will be:

city	rating
New York	2.0
Barcelona	2.5
Bora Bora	5.0
Melbourne	5.0
Bangkok	3.0
Barbados	4.0
Airlie Beach Queensland	5.0
Cancun	5.0
Berlin	2.5
Vancouver	3.0
San Francisco	4.0
Las Vegas	3.5
Cairo	4.0

We can clearly see how this user does not like big cities. Making such a specific profile helps us to validate the recommendations of the program, since if it were a more mixed profile, it will cost us to know if the program has been successful or not. Being such a specific profile, we will see that the program has been correct if it recommends beach and quiet places, and fails if it recommends large cities, such as New York, Barcelona or Berlin.

Next, we match the data from the user input with the cities data.

We create a new data frame, but just with the cities that the user has visited

	Aquarium	Arcade & Bowling	Casino	Cinema	Night club	Disco	Music	Art	Stadium	Theme Park
Airlie Beach Queensland	0.000000	0.000000	0.000000	0.000000	0.023170	0.002106	0.002106	0.000000	0.000000	0.008425
Bangkok	0.003763	0.003494	0.003225	0.013439	0.013439	0.013439	0.013439	0.013439	0.013439	0.006720
Barbados	0.000000	0.000000	0.000000	0.000000	0.008759	0.002920	0.000000	0.008759	0.000000	0.004380
Barcelona	0.001291	0.001291	0.005162	0.009034	0.016132	0.016132	0.016132	0.016132	0.003226	0.008066
Berlin	0.003848	0.003498	0.012244	0.008746	0.017492	0.017492	0.017492	0.017492	0.001050	0.008746
Bora Bora	0.002789	0.000000	0.000000	0.002789	0.008368	0.000000	0.000000	0.000000	0.000000	0.000000
Cairo	0.001799	0.000450	0.010342	0.011691	0.022482	0.001349	0.005845	0.022482	0.003147	0.010567
Cancun	0.000000	0.000527	0.003161	0.002107	0.026344	0.001054	0.006322	0.026344	0.001054	0.002898
Las Vegas	0.000767	0.003452	0.019175	0.001151	0.019175	0.013806	0.006903	0.019175	0.001918	0.009588
Melbourne	0.004042	0.009817	0.009817	0.010683	0.014436	0.008950	0.014436	0.014436	0.008373	0.007218
New York	0.005326	0.006537	0.003874	0.012105	0.012105	0.012105	0.012105	0.012105	0.008474	0.006053
San Francisco	0.003338	0.005006	0.002781	0.010569	0.013906	0.013906	0.013906	0.013906	0.003894	0.006953

From here, we can calculate the user's tastes, seeing which are the greatest attributes of the cities visited, and seeing whether or not they liked each of the cities.

The profile for this user is as follows (We obtain a score for each of the 100 features, but we will show only the 15 most important ones):

```
Beach Bar                   1.804576
Sport bar                   1.531414
Bar                         1.516815
Cocktail Bar                1.516815
Seafood Restaurant          1.006117
Asian Restaurant            0.980840
African Restaurant          0.980840
French Restaurant           0.977399
Mediterranean Restaurant    0.970454
American Restaurant         0.966242
Italian Restaurant          0.966242
Turkish Restaurant          0.966242
Mexican Restaurant          0.966242
Gift shop                   0.850035
Antique shop                0.845822
```

The next step is to remove the user entered cities from the dataset. If the user has already visited them, he/she doesn't need us to tell him/her if he/she will like them.

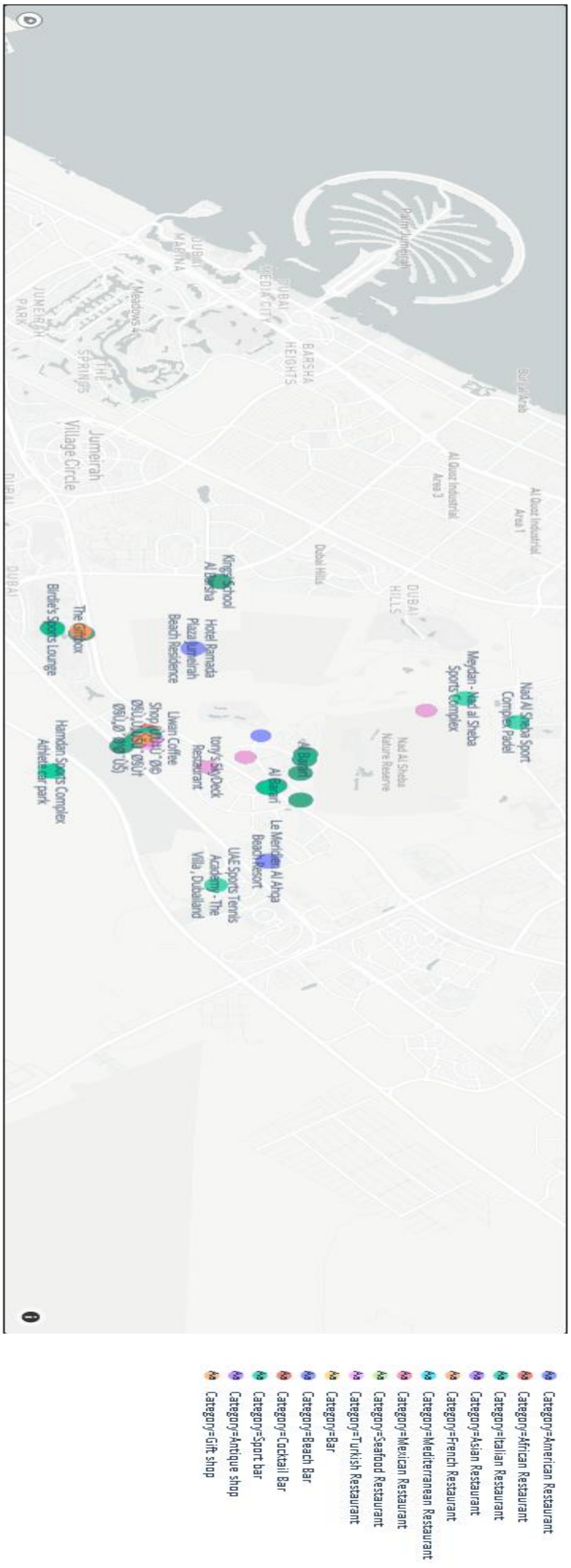
Then we calculate the score for each of the remaining cities, and order them from highest to lowest. Finally, we look at the categories with the greatest weight in each of the cities, and we create a table indicating next to the city, the coincidence with the user and the 10 largest categories. The table is as follows:

	Match	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Maldives	0.024405	Sport bar	Beach Bar	Bar	Antique shop	Cocktail Bar
La Digue Island	0.023480	Beach Bar	Cocktail Bar	Bar	Sport bar	Island
Isle of Man	0.022069	Bar	Beach Bar	Cocktail Bar	Sport bar	Mountain
Puerto del Rosario, Canary Islands	0.021413	Sport bar	Beach Bar	Cocktail Bar	Bar	Cafeteria
Zermatt	0.021052	Asian Restaurant	Sport bar	American Restaurant	Mediterranean Restaurant	African Restaurant
Antigua	0.020885	Turkish Restaurant	Hotel	American Restaurant	African Restaurant	Italian Restaurant
Courchevel	0.020550	Sport bar	Ski Area	Hotel	Cocktail Bar	Beach Bar
Chamonix	0.020474	Hotel	Sport bar	Cocktail Bar	Beach Bar	Bar
Sorrento	0.020099	Cocktail Bar	Sport bar	Bar	Beach Bar	Hotel
Luxor	0.019727	African Restaurant	Hotel	Italian Restaurant	Turkish Restaurant	American Restaurant
Machu Picchu	0.019577	Seafood Restaurant	Hotel	Italian Restaurant	Asian Restaurant	American Restaurant
Victoria Falls	0.019413	Hotel	Beach Bar	African Restaurant	Bar	Cocktail Bar
Algarve	0.019077	French Restaurant	Turkish Restaurant	American Restaurant	African Restaurant	Italian Restaurant
Dubai	0.019050	Sport bar	Turkish Restaurant	Beach Bar	Bar	Cocktail Bar
Chichen Itza	0.018828	Mexican Restaurant	American Restaurant	French Restaurant	Turkish Restaurant	Italian Restaurant

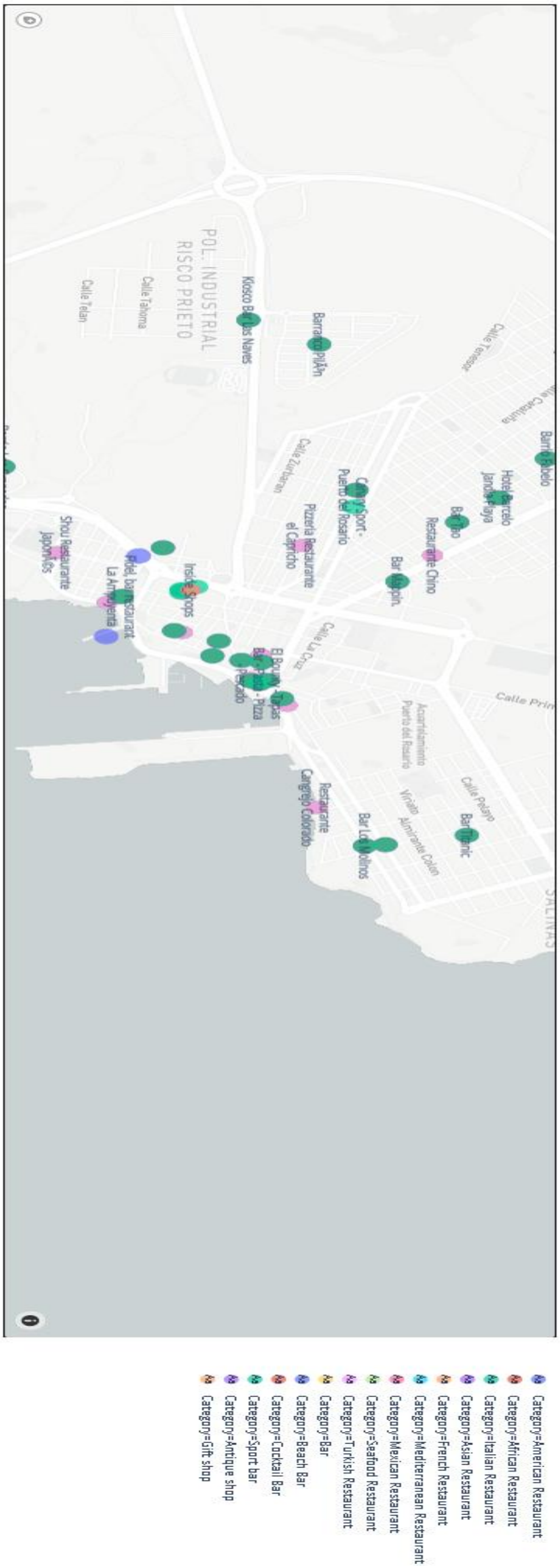
6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Island	Italian Restaurant	African Restaurant	Seafood Restaurant	French Restaurant
Italian Restaurant	American Restaurant	Mexican Restaurant	Mediterranean Restaurant	Turkish Restaurant
Irish Pub	African Restaurant	Mediterranean Restaurant	Waterfall	Turkish Restaurant
French Restaurant	American Restaurant	Turkish Restaurant	Seafood Restaurant	Mexican Restaurant
Mexican Restaurant	Seafood Restaurant	Hotel	Italian Restaurant	Turkish Restaurant
Asian Restaurant	French Restaurant	Mediterranean Restaurant	Sport bar	Mexican Restaurant
Bar	French Restaurant	Asian Restaurant	Mexican Restaurant	Italian Restaurant
French Restaurant	Italian Restaurant	American Restaurant	Mediterranean Restaurant	Asian Restaurant
Italian Restaurant	Mediterranean Restaurant	Night club	African Restaurant	American Restaurant
Asian Restaurant	French Restaurant	Mediterranean Restaurant	Mexican Restaurant	Seafood Restaurant
Turkish Restaurant	French Restaurant	African Restaurant	Mexican Restaurant	Mediterranean Restaurant
Sport bar	American Restaurant	Seafood Restaurant	Asian Restaurant	Italian Restaurant
Mediterranean Restaurant	Asian Restaurant	Bar	Social Club	Beach Bar
Park	Seafood Restaurant	National Park	African Restaurant	American Restaurant
Seafood Restaurant	African Restaurant	Asian Restaurant	Mediterranean Restaurant	Hotel

Finally, we will create a Dash dashboard, where the user can select a city from the recommended 15, and the city and its most relevant places are shown in a map. A few examples of the recommended cities are shown here:

Dubai:


Category=American Restaurant
Category=African Restaurant
Category=Italian Restaurant
Category=Asian Restaurant
Category=French Restaurant
Category=Mediterranean Restaurant
Category=Mexican Restaurant
Category=Seafood Restaurant
Category=Turkish Restaurant
Category=Bar
Category=Beach Bar
Category=Cocktail Bar
Category=Sport bar
Category=Antique shop
Category=Gift shop

Canary Islands



5. Results and Discussion

The result of this project is the recommended sites for a particular user. As has been seen and explained, the scores entered clearly correspond to a person with little interest in big cities, someone who enjoys relaxing vacations much more, in quiet places and especially with the beach. We can see, as of the 15 recommended cities, except Dubai, the other cities are relatively quiet cities, and most are beach, so it seems that the algorithm works quite well, and the recommendations are good.

6. Conclusion

In conclusion, we can say that this program is a good tool when planning a vacation, since due to the wide range of places to go, and how quickly they all change, it is difficult to know where to go, and where you will find what you are looking for. You could choose to manually search for sites that seem good to you, and use applications such as google maps or Foursquare to find out if those sites really have what you are looking for, but it is always better if they can give it to you done, as in this case!

The final decision of where to go will be up to the client, but the recommendations are made.