Travel Suggestion Scorecard

Leveraging on Foursquare Data and Individual Preferences to Create Travel Recommendations

Introduction

Many businesses have been affected by the pandemic. Lockdowns and quarantines have been required to flatten the curve on new cases and as such, travel plans have been put on hold. While this may be the least of our concerns now, several businesses that rely on travel and tourism have been severely negatively affected. Among the hardest hit are travel agencies. As such, my target audience for this project are travel agencies. They can use the resulting simple engine or a variant of it to make recommendations to future vacationers which will hopefully help bolster the economy of the target countries.

Problem

The first travel after the pandemic will be important as this could impact the desire for any succeeding travels. To facilitate a pleasant experience once a vaccine is created and the world is open again, travel agencies can implement a recommendation engine that considers people's preferences in a location.

Data

Data Sources

To create this engine in Python, we would need the following data:

Data	Source
List of Countries, Capital Cities	https://lab.lmnixon.org/4th/worldcapitals.html
Coordinates of Capital Cities	https://lab.lmnixon.org/4th/worldcapitals.html
Top Venues per Location	Foursquare API
Happiness Index of each country	https://en.wikipedia.org/wiki/World_Happiness_Report
Preferred types of venues	User Input
Relative importance for each venue type	User Input

The happiness index of each country is where I will do initial testing and confirmation that the environment impacts happiness. And the user inputs will be used to generate a 'score' for each place. This is ideally customizable depending on the client's preferences.

Data Cleansing and Transformation

Using Python in JupyterLab, I had to perform the following transformations:

Data	Transformation
List of Countries, Capital Cities	Table from webpage to Pandas data frame
Coordinates of Capital Cities	Coordinates to correct format for Foursquare query, e.g. from (34.28N, 69.11E) to (34.28,69.11)

Top Venues per Location	Manually created new dictionary (using list) to reduce number of categories. Use pd.get_dummies to create a matrix format containing count of each category, and then group the results by Country.
Happiness Index of each country	Table from webpage to Pandas data frame
Preferred types of venues	Input as List in Python
Relative importance for each venue type	Input as List in Python

The table below shows the words that were grouped together. For example, a venue category name that contains “pizza” or “burger” will just be grouped into restaurant.

Final Category Name	**Words to Associate with Category Name**
restaurant	['pizza','burger','sandwich','bbq','noodle','soup','hotdog','buffet','diner','osteria','fish','burrito','taco','restaurant','steak','chicken','wings','salad','food','snack','breakfast','brasserie','cafeteria']
cafe_desserts	['cafe','café','bakery','coffee','ice cream','pastry','juice','tea','creperie','candy','bagel','pastry,"chocolate','pie','yogurt','cupcake','cake','donut','dessert']
alc_bev	['pub','wine','vin','speakeasy','nightlife','liquor','club','cocktail','lounge','bistro','whisky','beer','brewery','bar','distillery']
grocery	['market','grocery','gourmet','deli','fruit & vegetable']
fitness	['gym','fitness','pool','tennis','basketball','golf','soccer','track','bowling','hockey']
amusement_cultural	['theater','arts','historic','museum','opera','art','cultural','monument']
amusement_nature	['ski','trail','park','garden','zoo','mountain','beach','resort','forest','surf','harbor','scenic','outdoors','tree','dive','outdoor','waterfall']
entertainment	['movie','game','music','shopping','plaza','stadium','entertainment','recreation','concert','video']

Methodology

As part of exploratory data analysis, I initially tried to see if the kind of venues within a location impacted happiness (e.g. does having constant access to a cup of coffee in the morning in your neighborhood improve your mood?) . This required (1) location/coordinates data for the capital cities of each country to query the top venues of each country in Foursquare, and the (2) happiness index for each country. I am operating under the assumption then that the happiness index, while composed of answers of locals, can be used as a response variable for tourists as well in this project.

In the hopes of identifying if there are specific venue categories that might contribute to a happy trip and regressing it somehow to determine the likelihood that one will have a good trip, I initially attempted univariate analysis using correlations and data visualizations, i.e. plots where y-axis is the Happiness Index and x-axis is %venue category in the area where venue categories could be restaurants, cafes, sources of alcoholic beverages in the area, etc. However, these were inconclusive. I could not observe any linear relationships.

I followed with an unsupervised machine learning method, K-means clustering, to check if it would be able to group the countries in any meaningful way using (only) the top venues of each country. Once the clusters were obtained, I computed the average happiness index of each cluster and observed that these were on different levels. I proceeded to do a t-test to check if there any significant difference in the means and it concluded that there was indeed a significant difference in average Happiness Index levels for some clusters. Without reading through the entire methodology of the happiness index, we can conclude that the top venues, or the environment, impacts happiness to some degree.[1] I then checked the resulting clusters and noted some differences in top venues which might have contributed to happiness. While I cannot assign a score for a specific venue from these results, these clusters can be used for marketing materials which we shall see later.

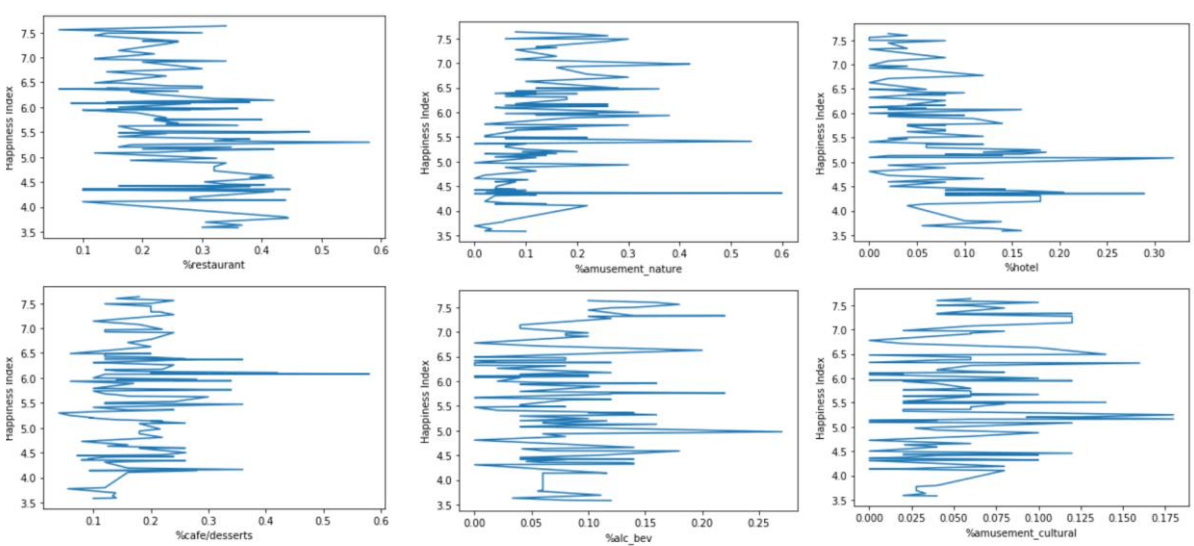
Finally, acknowledging then that each person has different preferences, I wanted to be able to consider user inputs. While it may be simplistic, I proceeded with a scorecard that was user-based. First, the user must input two things: (1) the types of venues that are important for him/her, and (2) the relative importance of each where the relative importance is a % and all of these weights sum to 100%. The score for each country is then:

$$Score = \sum_{i=1}^{n\ categories\ listed} weight_i \times number\ of\ venues_i$$

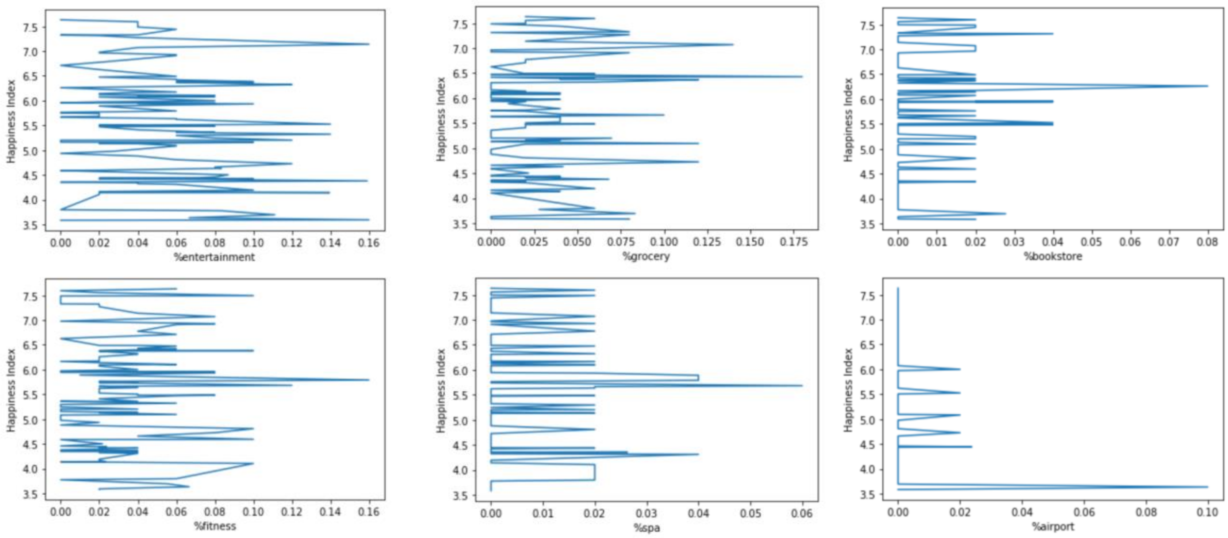
Where $\sum_{i=1}^{n} weight_i = 100\%$ and number of venues are the results from the Foursquare data extraction.

Results

The results of the univariate analysis are inconclusive. The plots for the top categories against the Happiness Index are shown below. If there were an increasing trend line for %restaurants against Happiness Index, it would have indicated that the more restaurants there are in the area, the happier people are. However, such is not the case for any venue category.



[1] Note that we are using Foursquare data and from what I observed when I initially attempted to do a different study on local Philippine data and hospitals, the entries are mainly restaurants and cafes.



The correlations portray the same message:

Venue Category	Correlation with HI
restaurant	-0.406501
cafe/desserts	0.145805
amusement_nature	0.349484
alc_bev	0.028303
hotel	-0.490189
amusement_cultural	0.226268
entertainment	-0.146013

Venue Category	Correlation with HI
fitness	0.011613
grocery	0.112568
bookstore	0.177417
spa	0.030326
airport	-0.200702
multiplex	-0.110114
castle	0.241804
hostel	-0.097842

Using k-means clustering on the top venue categories per capital city with k=4,[2] and merging the resulting clusters (Country, Cluster Label) with the happiness index dataset (Country, Happiness Index), we can see that in some clusters, the resulting average happiness index are significantly different from each other based on the p-values of a t-test.

	1	2	3	4
1	1.000000	0.941399	0.004050	0.016212
2	0.941399	1.000000	0.003849	0.016724
3	0.004050	0.003849	1.000000	0.354098
4	0.016212	0.016724	0.354098	1.000000

[2] Changing the k higher will result to NaN values. K=3 will still result to at least 1 cluster pair with average happiness indices that are not significantly different from each other.

Cluster 1 is significanly different from clusters 3 and 4. Cluster 2 is significantly different from clusters 3 and 4. Clusters 3 is significantly different from clusters 1 and 2. Cluster 4 is significantly different from clusters 1 and 2.

Let's inspect the members of the clusters now. Cluster 1 has the lowest overall average happiness at 5.29 and majority of the venues are composed of restaurants, cafes and dessert places and places that serve alcoholic beverages.

Cluster 1

```
cluster1=group[group['Cluster Labels']==0]
print('# of members:', cluster1.shape[0])
print('average happiness index:', cluster1['HI'].mean())
cluster1
# average happiness, lots of restaurants, cafes/dessert places, and places that serve alcohol
```

```
# of members: 33
average happiness index: 5.2920606060606055
```

	Country	HI	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Finland	7.632	0	restaurant	cafe/desserts	alc_bev	amusement_nature	amusement_cultural
3	Iceland	7.495	0	restaurant	cafe/desserts	alc_bev	fitness	amusement_nature
6	Canada	7.328	0	restaurant	cafe/desserts	amusement_nature	alc_bev	grocery
14	Belgium	6.927	0	restaurant	amusement_nature	cafe/desserts	fitness	alc_bev
31	Slovakia	6.173	0	restaurant	cafe/desserts	alc_bev	amusement_nature	hotel
32	El Salvador	6.167	0	restaurant	amusement_nature	cafe/desserts	hotel	alc_bev
33	Nicaragua	6.141	0	restaurant	cafe/desserts	amusement_cultural	amusement_nature	clothing store
36	Uzbekistan	6.096	0	restaurant	cafe/desserts	alc_bev	hotel	entertainment
40	Ecuador	5.973	0	restaurant	cafe/desserts	hotel	alc_bev	amusement_nature
43	Slovenia	5.948	0	restaurant	cafe/desserts	amusement_nature	entertainment	alc_bev
49	Bolivia	5.752	0	restaurant	cafe/desserts	alc_bev	hotel	amusement_cultural

Cluster 2 appears to contain countries with overall slighty better happiness level than cluster 1 at 5.31 average and majority of the venues are composed of restaurants, amusement_nature and hotels.

Cluster 2

```
cluster2=group[group['Cluster Labels']==1]
print('# of members:', cluster2.shape[0])
print('average happiness index:', cluster2['HI'].mean())
cluster2

# average happiness index, lots of restaurants and hotels
```

```
# of members: 29
average happiness index: 5.310999999999999
```

	Country	HI	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
10	Austria	7.139	1	amusement_nature	restaurant	entertainment	amusement_cultural	hotel
16	United Arab Emirates	6.774	1	restaurant	amusement_nature	cafe/desserts	hotel	fitness
22	Panama	6.430	1	restaurant	grocery	cafe/desserts	amusement_nature	alc_bev
23	Brazil	6.419	1	restaurant	cafe/desserts	amusement_nature	hotel	fitness
25	Uruguay	6.379	1	amusement_nature	restaurant	cafe/desserts	fitness	alc_bev
28	Malaysia	6.322	1	restaurant	cafe/desserts	entertainment	grocery	amusement_nature
29	Spain	6.310	1	restaurant	amusement_nature	amusement_cultural	entertainment	cafe/desserts
38	Thailand	6.072	1	restaurant	hotel	amusement_nature	cafe/desserts	buddhist temple

Cluster 3 countries have the highest average happines level at 6.22 and majority of the venues are of the amusement_nature category. These include beaches, trail parks, gardens, mountains, surfing sports, watefalls, and other scenic viewpoints.

Cluster 3

```
cluster3=group[group['Cluster Labels']==2]
print('# of members:', cluster3.shape[0])
print('average happiness index:', cluster3['HI'].mean())
cluster3

#highest happiness, lots of amusement_nature venues
```

```
# of members: 18
average happiness index: 6.223055555555556
```

	Country	HI	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Norway	7.594	2	amusement_nature	alc_bev	cafe/desserts	restaurant	grocery
2	Denmark	7.555	2	amusement_nature	cafe/desserts	alc_bev	amusement_cultural	restaurant
4	Switzerland	7.487	2	amusement_nature	restaurant	cafe/desserts	hotel	alc_bev
5	Netherlands	7.441	2	amusement_nature	cafe/desserts	restaurant	alc_bev	amusement_cultural
12	Ireland	6.977	2	amusement_nature	cafe/desserts	restaurant	alc_bev	grocery
13	Germany	6.965	2	amusement_nature	restaurant	cafe/desserts	amusement_cultural	palace
17	Czech Republic	6.711	2	amusement_nature	cafe/desserts	restaurant	fitness	alc_bev
19	France	6.489	2	amusement_nature	castle	amusement_cultural	restaurant	hotel
21	Chile	6.476	2	amusement_nature	restaurant	cafe/desserts	fitness	alc_bev
39	Italy	6.000	2	amusement_nature	restaurant	amusement_cultural	cafe/desserts	entertainment

Cluster 4 countries have overall higher happiness levels than clusters 1 and 2. The primary difference appears to be that the top category is cafes and dessert places as opposed to restaurants.

Cluster 4

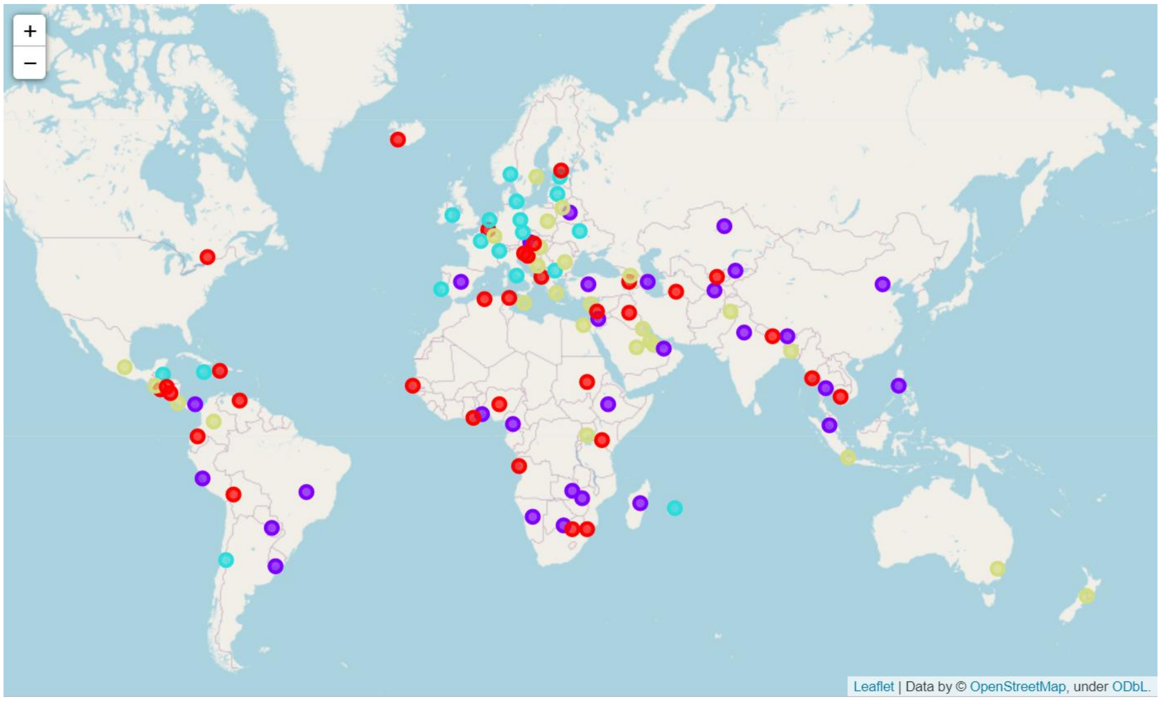
```
cluster4=group[group['Cluster Labels']==3]
print('# of members:', cluster4.shape[0])
print('average happiness index:', cluster4['HI'].mean())
cluster4

# high happiness
# primarily cafes/dessert places as opposed to restaurants from Cluster1
```

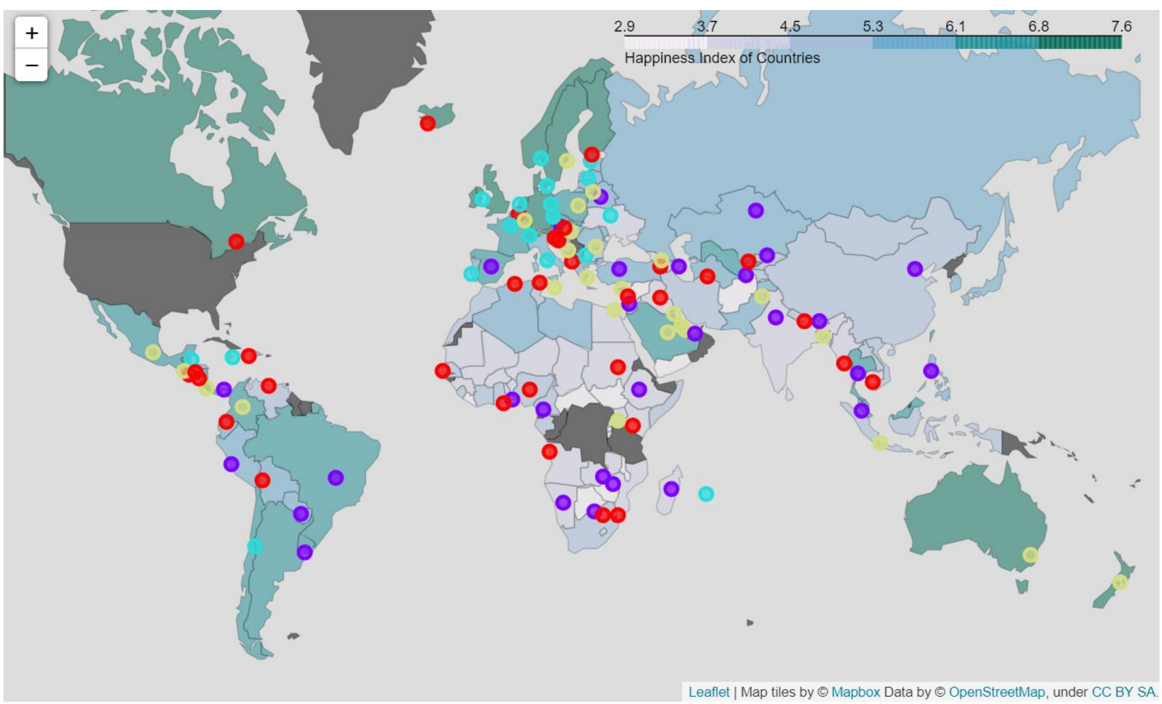
```
# of members: 26
average happiness index: 5.940615384615385
```

	Country	HI	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
7	New Zealand	7.324	3	cafe/desserts	restaurant	alc_bev	amusement_nature	amusement_cultural
8	Sweden	7.314	3	restaurant	cafe/desserts	amusement_nature	alc_bev	bookstore
9	Australia	7.272	3	restaurant	cafe/desserts	amusement_cultural	alc_bev	amusement_nature
11	Costa Rica	7.072	3	cafe/desserts	restaurant	grocery	fitness	amusement_nature
15	Luxembourg	6.910	3	cafe/desserts	restaurant	amusement_nature	alc_bev	grocery
18	Malta	6.627	3	restaurant	alc_bev	cafe/desserts	amusement_cultural	amusement_nature
20	Mexico	6.488	3	cafe/desserts	restaurant	amusement_cultural	amusement_nature	grocery
24	Guatemala	6.382	3	restaurant	cafe/desserts	entertainment	grocery	amusement_cultural
26	Qatar	6.374	3	cafe/desserts	restaurant	grocery	entertainment	hotel
27	Saudi Arabia	6.371	3	cafe/desserts	grocery	amusement_nature	farm	amusement_cultural

Using folium and matplotlib packages, a visualization of the clusters around the world is shown below.



Against a choropleth map with the Happiness Index scale:



Cluster 1 countries with the lowest average happiness level are those marked red while Cluster 3 countries with the highest average happiness level are those marked in blue. Most of the happiest countries are clustered in Europe.

For the-user specified output, first recall that we require two inputs with examples shown below:

(1) Important venue categories for the user

```
Choose 5 categories from the dropdown list below.
Press Ctrl key to select multiple.
```

Venue
- restaurant
- cafe/desserts
- amusement_nature
- alc_bev
- hotel
- amusement_cultural
- entertainment
- fitness
- grocery
- spa

(2) Relative importance of each

```
Enter the weights for each category chosen above. The weights must sum to 100.
20 30 10 20 20
```

If the weights entered are incorrect, the code will output a message that it doesn't sum up to 100. Otherwise it will confirm with a message like below:

```
Enter the weights for each category chosen above. The weights must sum to 100. 20 30 10 20 20

weight for restaurant is 20 %
weight for amusement_nature is 30 %
weight for alc_bev is 10 %
weight for amusement_cultural is 20 %
weight for entertainment is 20 %
```

After which, we can get a score column using the formula:

$$Score = \sum_{i=1}^{n\ categories\ listed} weight_i \times number\ of\ venues_i$$

Country	restaurant	amusement_nature	alc_bev	amusement_cultural	entertainment	score
Jamaica	19	20	11	4	2	12.1
Norfolk Island	19	20	11	4	2	12.1
Saint Lucia	10	28	3	0	2	11.1
Portugal	9	26	1	3	2	10.7
Aruba	17	22	4	1	0	10.6
Mauritania	5	30	3	0	1	10.5
Andorra	18	19	4	1	0	9.9
United States of Virgin Islands	12	20	12	0	0	9.6
Martinique	10	21	5	1	2	9.4
British Virgin Islands	11	20	12	0	0	9.4

And these are the countries that we can recommend to our client. We also have visibility of the specific categories so we can explain why those are the returned recommended countries.

Discussion and Recommendations

From the results, we can see that majority of the top venues will lean towards the restaurants and cafes or dessert places categories. I also note that the results are dependent on how active people in a certain country are in using Foursquare.

The results indicate that the environment does play a factor into happiness. In particular, the existence of **nature** as a top venue in a country appears to have the strongest positive impact. These include beaches, mountains, surfing spots, skiing spots, forests, among others.

Assuming people are still hesitant to travel and it will be unlikely that you can get user inputs soon, initially, the outputs can be used to generate marketing material. We can filter the top venue categories and create a list from there. Examples are shown below[3]:

[3] Image sources:
Nature: https://www.activeme.ie/useful-info/top-10-best-scenic-drives-in-ireland-great-european-road-trips-tourist-driving-routes-holiday-vacation/
Food:

When people are ready to travel again, we can use our simple tool to create scores and rank countries for each client depending on their preferences.

Conclusion

Despite the simplistic scorecard approach, one can create something of use because we were able to leverage on free Foursquare data.

While the initial univariate analysis did not reap anything useful and one can initially prematurely conclude that the environment does not factor into happiness (assuming we did not read the methodology for this index), an unsupervised machine learning method (k-means clustering) was able to somehow group the countries based on their top venues and lead to select clusters with significantly different happiness index levels.