Best city to study in

Coursera - Applied Data Science Capstone
Peer-graded Assignment: Capstone Project - The Battle of Neighbourhoods (week 2)

By: Joseph M Camilleri
6th August 2020

1. Introduction

1.1 Background

Many students from around the world study in a city other than their home town, many also study abroad. Their stay in another city is generally temporary in order to study in an institution that is renowned for its academic record in a particular subject. Other students choose to follow a multi-year programme in some other city or even in another country.

This offers many benefits to the students, over and above the education itself. These benefits include the experience of living in a different part of the world, and absorbing an entirely new culture. Experiencing life in a cosmopolitan city, is probably the best preparation for work life within a multi-cultural international organisation. Living in a city where the student is immersed in a language other than his/her mother tongue is highly effective in accelerating the learning of that new language. Other advantages include: making new friends, discovering new interests, gaining of new experiences and the related personal development. Studying in a different city also opens up career opportunities which may otherwise not be available.

1.2 Problem

As a first step student often have a choice to make, as to where to relocate for one or more years in order to complete their studies. While each student's case is different, the decision is sometimes driven by:

- having received a scholarship to a specific institution; or
- a desire to specialise in a field of study at a particular educational institution.

In such cases the decision of where to relocate to non-discretionary as it depends solely on the location of the particular institution.

However this is not always the case, and sometimes students are required to choose amongst a number of institutions, each located in different cities or even in different countries around the world. Hence, the student's final decision will also be influenced by the preferred country and city where the various institutions are located. Putting it another way, when a student chooses to study at a specific colleague or university they are also choosing to live in a particular city.

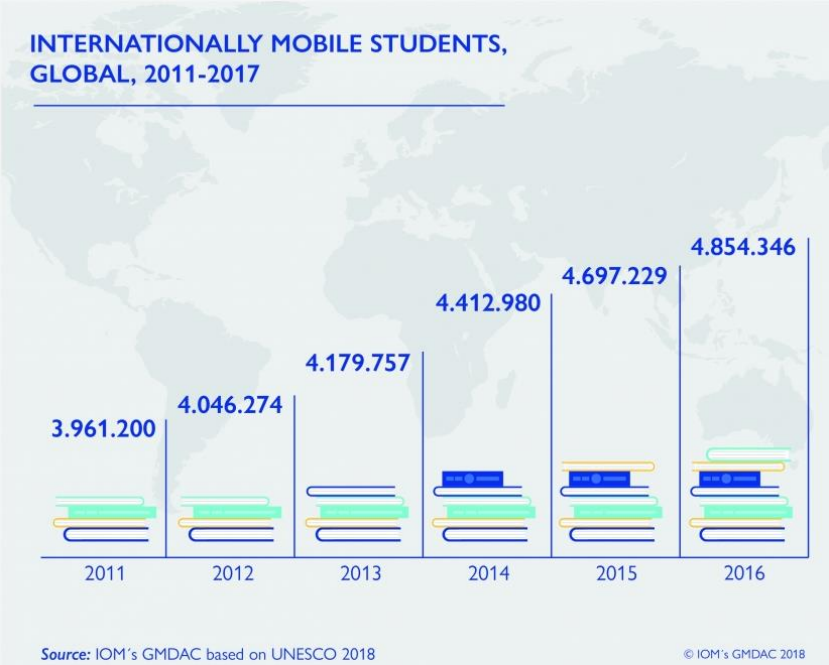
The objective of this assignment, is to provide some assistance to students in establishing their preferences and choosing from amongst a number of viable options in relation to their preferred city and country.

1.3 Interest

There were 1.7 million students from abroad who were undertaking tertiary level studies across the EU-28 in 2017. Across the EU-28 in 2017, some 436,000 students from abroad (25.5% of the total) were studying in the United Kingdom, far more than in any other EU Member State. More than one third (37.8 %) of the students from abroad who were undertaking tertiary level studies across the EU in 2017 were from Europe, 30.1% were from Asia and 13.0% were from Africa. [source: https://ec.europa.eu/eurostat/statistics-explained/index.php/Learning_mobility_statistics]

A total of 341,751 US students studied abroad for credits in 2017-18, representing a 2.7% increase from the previous academic year, according to the annual "Open Doors" report, published by the Institute of International Education with funding from the US Department of State. [source: https://www.insidehighered.com/news/2019/11/18/open-doors-data-show-continued-increase-numbers-americans-studying-abroad]

The graph below shows that on average between 2011 and 2016 the growth in student mobility increase by an average of 4.2% year-on-year.



The aim of this assignment is to enable prospective students to analyse and compare different cities around the world, also in relation to their own personal interests. For example, if a student is an outdoors activities enthusiast he/she may give greater consideration to studying in a city which offers more opportunities for such activities. Similarly a student may choose one city over another having given consideration that one city hosts a greater number of colleagues and universities, and thus offering greater opportunities to interact with likeminded students and make new friends.

2. Data acquisition and cleansing

2.1 Data sources

The aim of this assignment is to produce a model design to be used by different students, one at a time. Hence the first step is to capture the set of cities being considered by a student. The cities are introduced into the Jupyter Notes book as a simple dataframe consisting of the city name and country name. I found that the name of the city by itself is insufficient and must be qualified by the country as there are a few city names which have the same name, for example London UK and London Canada. Without the country name looking up data for such cities would be a problem. A student can enter any number of cities to be considered as there is no fixed number of cities that must be entered. An example is given below:

City	Country
New York	United States
Toronto	Canada
Los Angeles	United States
Houston	United States
Melbourne	Australia
London	United Kingdom
Dublin	Ireland
Rome	Italy
Paris	France
Barcelona	Spain

In order to identify the city that is best suited for a student, the students must also enter a set of interests to search for venue categories in each city under consideration. For example

- if the student likes Italian food, one interest may be “Italian”;
- if a student is particularly excited about making new friends and meeting other students, another search string may be “University” to highlight cities have more University venues.
- if a student is interested in a particular sport such as “football”, “baseball” or “basketball” these can also be search strings.

Again the student can specify any number of interests in a simple dataframe as per example below.

interest
Socer
Basketball
Park
University
Sushi
Indian

When considering living in another city for this first time, two other key considerations will generally be the level of safety and cost of living in that city. This information can be obtained by scraping data from the following two web pages:

https://www.numbeo.com/crime/rankings.jsp

Below please find a screen shot from this site containing the safety index for each city.

Select date: 2020 Mid-Year

Select Region: Africa America Asia Europe Oceania

Select display column: ---All columns---

Search:

Rank	City	Crime Index	Safety Index
1	Caracas, Venezuela	84.57	15.43
2	Pretoria, South Africa	81.89	18.11
3	San Pedro Sula, Honduras	81.35	18.65
4	Port Moresby, Papua New Guinea	81.05	18.95

And the cost of living index data from the same web site:

https://www.numbeo.com/cost-of-living/rankings.jsp

Select date: 2020 Mid-Year

Select Region: Africa America Asia Europe Oceania

Select display column: ---All columns---

Search:

Rank	City	Cost of Living Index	Rent Index	Cost of Living Plus Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index
1	Zurich, Switzerland	131.49	64.37	98.80	131.23	120.39	121.12
2	Lugano, Switzerland	130.75	40.20	86.66	134.80	115.19	101.15
3	Basel, Switzerland	130.65	46.61	89.72	126.05	131.92	109.90
4	Geneva, Switzerland	126.08	66.56	97.10	123.84	119.47	108.09
5	Lausanne, Switzerland	125.03	51.99	89.46	125.56	118.13	110.86
6	Bern, Switzerland	116.53	41.60	80.04	106.32	110.46	125.72
7	New York, NY, United States	100.00	100.00	100.00	100.00	100.00	100.00
8	Trondheim, Norway	98.23	34.32	67.11	93.13	94.83	83.32
9	Stavanger, Norway	97.91	32.75	66.18	90.42	99.35	83.99
10	Oslo, Norway	95.78	40.23	68.72	88.40	96.20	86.05
11	Bergen, Norway	94.78	31.59	64.01	85.87	96.17	93.49
12	San Francisco, CA, United States	92.13	109.76	100.72	89.79	88.26	139.00

Data about the venues in each city is obtained using the Foursquare API. This would be used to check for the relative number of venues matching each of the interests of a students, as well as in order to explore what category of venues exit in a city and thus be able compare the cities. Below please find a sample of the type of data extracted using Foursquare API. [source: https://developer.foursquare.com/]

City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
New York	40.712728	-74.006015	The Bar Room at Temple Court	40.711448	-74.006802	Hotel Bar
New York	40.712728	-74.006015	The Beekman, A Thompson Hotel	40.711173	-74.006702	Hotel
New York	40.712728	-74.006015	Alba Dry Cleaner & Tailor	40.711434	-74.006272	Laundry Service
New York	40.712728	-74.006015	Gibney Dance Center Downtown	40.713923	-74.005661	Dance Studio
New York	40.712728	-74.006015	City Hall Park	40.712415	-74.006724	Park

Finally the longitude and latitude of each city being considered by the student is required to find venues in the vicinity of the city centre via the Foursquare API. These geographic coordinates are obtained using the geolocator of geopy. Geopy is a Python client for several popular geocoding web services. Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources. [source: https://geopy.readthedocs.io/en/stable/]

2.2 Data cleansing and preparation

All the data from the various sources mentioned above will need to be merged into one dataframe.

The data in tables scraped from the two web sites mentioned above is initially captured into a number of lists, and the relevant lists are converted into a dataframe. In both web sites, the city name was a text string containing the:

- name of the city,
- name of the Country, and
- in some cases (particularly for US cities) also the state initials in the middle.

Each of the 2 or 3 elements would be separated by a comma, as per sample below:

```
13    NaN  Port of Spain, Trinidad And Tobago        75.70        24.30
14    NaN          Memphis, TN, United States        75.61        24.39
15    NaN                    Salvador, Brazil        75.60        24.40
16    NaN        Baltimore, MD, United States        74.75        25.25
17    NaN          Detroit, MI, United States        73.79        26.21
18    NaN             Cape Town, South Africa        73.67        26.33
19    NaN                      Klang, Malaysia        71.51        28.49
20    NaN           San Salvador, El Salvador        71.00        29.00
21    NaN                    Sao Paulo, Brazil        70.67        29.33
22    NaN      Saint Louis, MO, United States        70.63        29.37
23    NaN      Albuquerque, NM, United States        69.95        30.05
24    NaN                    Windhoek, Namibia        69.40        30.60
```

For consistency and in order to match these value with the student's cities of interest, this had to be split out into separate 2 or 3 columns and the state initials data is dropped, to have just city name and country name as per dataframe slice below.

	Safety Index	City	Country
170	54.14	Newcastle	Australia
171	54.27	Lodz	Poland
172	54.29	New York	United States
173	54.46	Katowice	Poland
174	54.50	Buffalo	United States
175	54.63	Kristiansand	Norway
176	54.70	Toulouse	France
177	54.82	Chisinau	Moldova
178	54.83	Tunis	Tunisia
179	54.94	San Jose	United States

Other columns which were not of interest were dropped from the dataframe scraped from each web site. The resulting two dataframes, holding the safety index and the cost of living index, were then merged with the initial table holding the cities being considered.

Furthermore, the cities dataframe was augmented with the longitude and latitude data from Geopy.

The dataframe was further augmented by adding one column for each interest specified by the student, as per example below.

	City	Country	Safety Index	Cost of Living Index	Latitude	Longitude	Socer	Basketball	Park	University	Sushi	Indian
0	New York	United States	54.29	100.00	40.712728	-74.006015	0	0	0	0	0	0
1	Toronto	Canada	60.79	72.56	43.653482	-79.383935	0	0	0	0	0	0
2	Los Angeles	United States	53.50	79.82	34.053691	-118.242767	0	0	0	0	0	0

These interest columns were populated with the number of instances (the numeric count) of matching category venues for that interest searched within that city using the Foursquare API. The numbers of matching venues in each city were then normalised to allow for visualisation of relative values between cities.

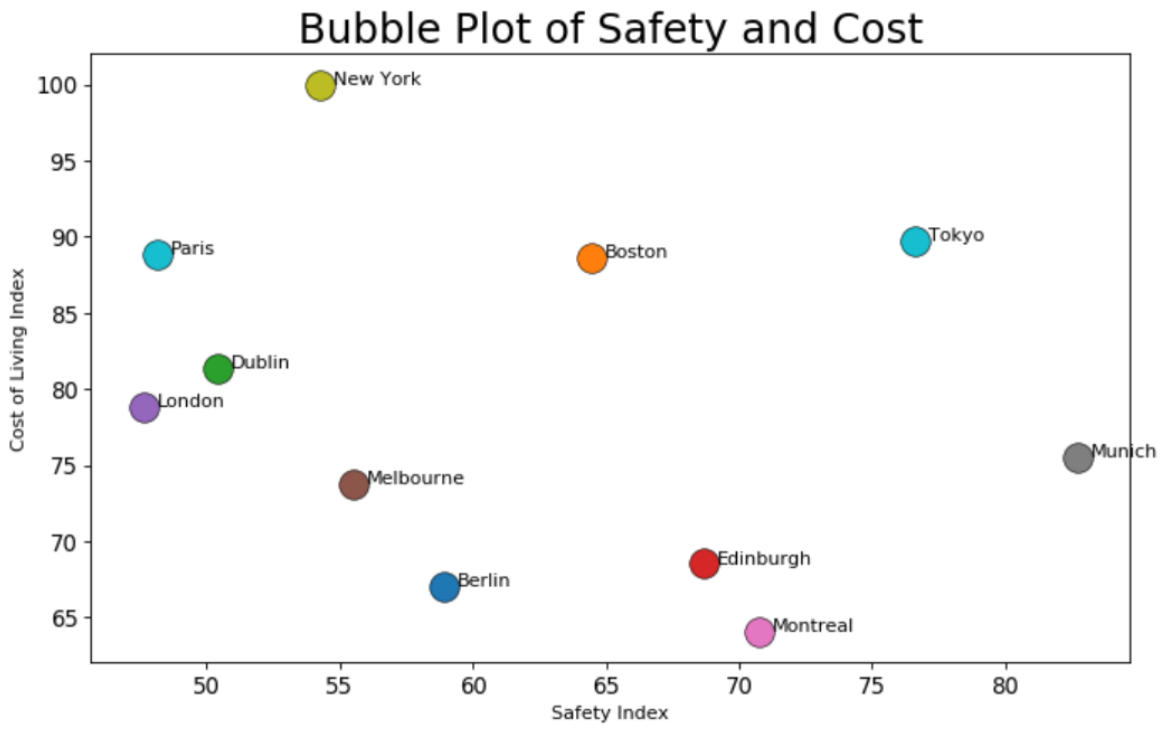
Finally I also obtained "explore" data from Foursquare for each city being considered by the student. From the Foursquare data retrieved, I extracted the relevant venues, including the category of each

venue. This data was transformed into a one hot encoding dataframe, grouped by city. The frequency of each venue category extracted was also normalised and placed into a new dataframe containing the top 10 venues by category for each city. This will enable comparison of the cities under consideration by the student to determine which cities are relatively similar to others within the consideration set of cities. This will assist the student in determining which group of cities may be a better for them.

3. Methodology

3.1 Exploratory data analysis

As mentioned above, the primary considerations will generally be safety and cost a living in each city, and therefore the first step was to visualise the cities under consideration against these two dimensions. After trying out a couple of visualisation tools I decided that bubble plot (based on matplotlib scatter plot) was the ideal graphical format, as depicted below.



In order to achieve the above visualisation I merged the Cities dataframe with data scraped from the websites for the Safety Index and Cost of Living Index by city. From the above graph it is very easy to see that Munich is the safest city while Montreal enjoys the lowest Cost of Living from amongst the selected cities. Note: in this case only the positioning of the bubbles has relevance while their size is

the same for each city. As a further improvement we could set the size of the bubble to represent a third dimension in our analysis, such as the size of the city's population, however this may not be of particular relevance to the student's decision.

Next, and in order to analyse the cities against the student's interest, I needed to augment further the dataframe of Cities with the longitude and latitude of each city plus a column for each of the interests specified by the student, resulting in a dataframe as follows:

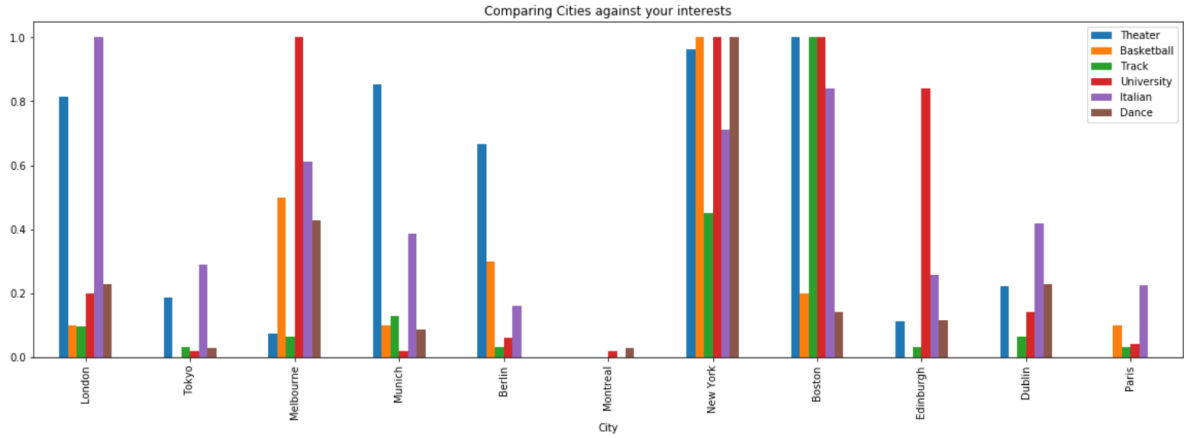
	City	Country	Safety Index	Cost of Living Index	Latitude	Longitude	Theater	Basketball	Track	University	Italian	Dance
0	London	United Kingdom	47.66	78.83	51.507322	-0.127647	0	0	0	0	0	0
1	Tokyo	Japan	76.60	89.69	35.682839	139.759455	0	0	0	0	0	0
2	Melbourne	Australia	55.53	73.76	-37.814218	144.963161	0	0	0	0	0	0
3	Munich	Germany	82.71	75.56	48.137108	11.575382	0	0	0	0	0	0
4	Berlin	Germany	58.92	67.02	52.517037	13.388860	0	0	0	0	0	0
5	Montreal	Canada	70.74	64.10	45.497216	-73.610364	0	0	0	0	0	0
6	New York	United States	54.29	100.00	40.712728	-74.006015	0	0	0	0	0	0
7	Boston	United States	64.48	88.61	42.360253	-71.058291	0	0	0	0	0	0
8	Edinburgh	United Kingdom	68.70	68.54	55.953346	-3.188375	0	0	0	0	0	0
9	Dublin	Ireland	50.42	81.39	53.349764	-6.260273	0	0	0	0	0	0
10	Paris	France	48.14	88.83	48.856697	2.351462	0	0	0	0	0	0

The student's interest columns ("Theater" to "Dance" above) were next populated by counting the number of matching venue categories retrieved using Foursquare API for each city and for each interest using nested loops.

	City	Country	Safety Index	Cost of Living Index	Latitude	Longitude	Theater	Basketball	Track	University	Italian	Dance
0	London	United Kingdom	47.66	78.83	51.507322	-0.127647	22	1	3	10	31	8
1	Tokyo	Japan	76.60	89.69	35.682839	139.759455	5	0	1	1	9	1
2	Melbourne	Australia	55.53	73.76	-37.814218	144.963161	2	5	2	50	19	15
3	Munich	Germany	82.71	75.56	48.137108	11.575382	23	1	4	1	12	3
4	Berlin	Germany	58.92	67.02	52.517037	13.388860	18	3	1	3	5	0
5	Montreal	Canada	70.74	64.10	45.497216	-73.610364	0	0	0	1	0	1
6	New York	United States	54.29	100.00	40.712728	-74.006015	26	10	14	50	22	35
7	Boston	United States	64.48	88.61	42.360253	-71.058291	27	2	31	50	26	5
8	Edinburgh	United Kingdom	68.70	68.54	55.953346	-3.188375	3	0	1	42	8	4
9	Dublin	Ireland	50.42	81.39	53.349764	-6.260273	6	0	2	7	13	8
10	Paris	France	48.14	88.83	48.856697	2.351462	0	1	1	2	7	0

Next I wanted to visualise this data to get a sense of the degree to which each city addresses the interests of the student, relative to the other cities. Firstly the numbers under each interest column shown above were normalised and subsequently plotted on a bar chart, as depicted below.

Normalisation was necessary because we may be viewing venues occurring in relatively high frequencies against others occurring in relatively low frequencies. Without normalisation differences in the low frequency vertical bars may be too small to discern. We are not interested in absolute numbers here, but to be also to compare relative values across cities.



From the above one can very easily appreciate that New York and Boston are the two cities that most closely matching the interests of this student. In order to decide between New York and Boston, the student may next consider the relative importance of each interest. For example Boston offer more venues for “Track” activities, whilst New York offers more in terms of “Basketball” and “Dance” venues. Is Track by itself of higher value to this student than Basketball and Dance together?

This graph has also helped confirm the correctness of the results. For example the city of Mumbai India (not depicted in the above example) obtained the highest score by far when compared to other cities for ‘Indian Food’ as an interest.

The next stage in the analysis of the Cities was to explore and compare the venues in general for each city. For this part of the analysis I used the “Explore” function of the Foursquare API, to obtain a number of venues in each city, which were placed in a dataframe as depicted in the small sample of rows below.

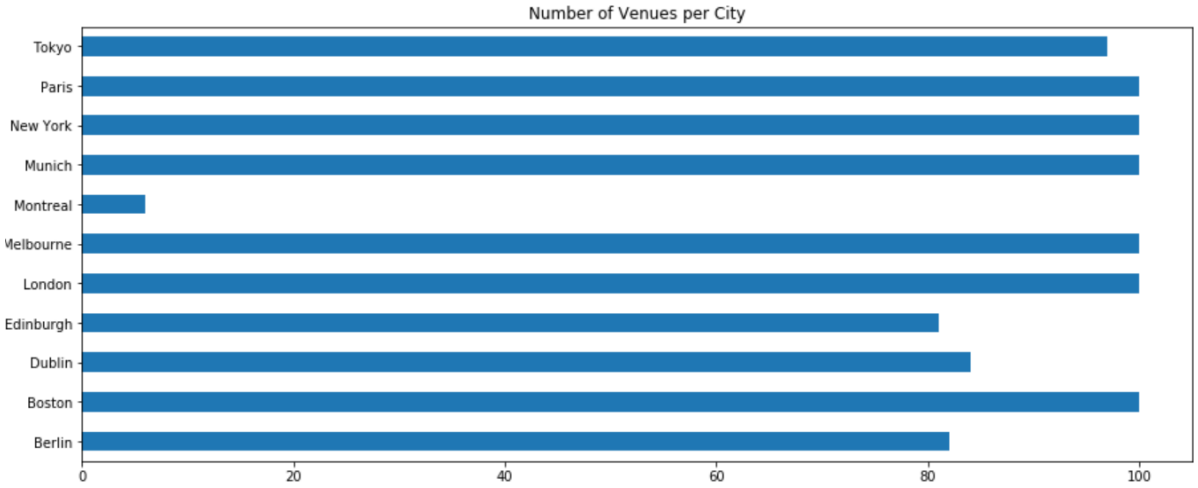
In [123]:
```
print(venues.shape)
venues.head()
```

```
(950, 7)
```

Out[123]:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	London	51.507322	-0.127647	National Gallery	51.508876	-0.128478	Art Museum
1	London	51.507322	-0.127647	Trafalgar Square	51.507987	-0.128048	Plaza
2	London	51.507322	-0.127647	East Trafalgar Square Fountain	51.508088	-0.127700	Fountain
3	London	51.507322	-0.127647	ESPA Life at Corinthia	51.506402	-0.125114	Spa
4	London	51.507322	-0.127647	Trafalgar Square Lions	51.507641	-0.127888	Outdoor Sculpture

The following horizontal bar chart also helped me to visualise the number of venues extracted per city. It seems that for a number of cities such as New York and Paris the API returned the maximum number of venues (i.e. 100). Even when the variable was set above 100 at 120 the maximum number of venues returned was still 100. It seems that the API is hitting a limit of the free Foursquare API licence.



Displaying the unique categories was also helpful in understanding the data that was retieved.

```
In [56]: print('There are {} uniques categories listed below:\n'.format(len(venues['Venue Category'].unique())))
         print(venues['Venue Category'].unique())
```

```
There are 211 uniques categories listed below:

['Art Museum' 'Plaza' 'Fountain' 'Spa' 'Outdoor Sculpture' 'Hotel'
 'Church' 'Monument / Landmark' 'North Indian Restaurant' 'Tea Room'
 'Spanish Restaurant' 'Art Gallery' 'Bookstore' 'Restaurant' 'Wine Bar'
 'Italian Restaurant' 'Thai Restaurant' 'Pub' 'Theater' 'Coffee Shop'
 'Garden' 'Burger Joint' 'Café' 'Ice Cream Shop' 'Boutique'
 'Indie Movie Theater' 'Japanese Restaurant' 'Steakhouse' 'Park'
 'Pharmacy' 'Cocktail Bar' 'Liquor Store' 'Greek Restaurant'
 'Toy / Game Store' 'Sandwich Place' 'Gay Bar' 'Bakery'
 'Pakistani Restaurant' 'Bar' 'Lounge' 'Tour Provider' 'French Restaurant'
 'Comedy Club' 'Chinese Restaurant' 'Ramen Restaurant' 'Sushi Restaurant'
 'Candy Store' 'Lebanese Restaurant' 'Noodle House' 'Irish Pub'
 'Seafood Restaurant' 'English Restaurant' 'Multiplex' 'Event Space'
 'Modern European Restaurant' 'American Restaurant' 'Road' 'Historic Site'
 'Dessert Shop' 'Brazilian Restaurant' 'Paper / Office Supplies Store'
 'Mediterranean Restaurant' 'Sports Club' 'Electronics Store'
```

3.3 Machine learning

The final stage was to compare the cities to determine which are more alike than others. This comparison was done based on the most frequent venue categories in each city. Thus the student will know that a group of cities (from those cities under consideration) are alike and different from the other group(s) of cities.

This grouping of similar cities was achieving using the K-means clustering technique. This clustering was based on the similarity in terms of the most 10 frequent venues in each city. Thus a student can expect a similar experience in any of the cities within a given cluster, and a different experience in cities in different clusters.

The first step in preparing the data for the K-means clustering machine learning technique was to reformat the data into a one-hot encoding dataframe, as per example below. Such a dataframe is very sparsely populated with 1s.

```
cities_onehot.head()
```

Out[126]:

	City	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Trail	Turkish Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Whis Bar
0	London	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0
1	London	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	London	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
3	London	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	London	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

5 rows × 214 columns

This dataframe was grouped by city and the values set by taking the mean of the frequency of occurrence. The results were confirmed by displaying the top 5 categories by frequency for each city, as follows:

```
----Berlin----
                venue  freq
0               Hotel  0.10
1   German Restaurant  0.07
2         Coffee Shop  0.05
3  Italian Restaurant  0.05
4               Café  0.04


----Boston----
                venue  freq
0  Italian Restaurant  0.14
1         Coffee Shop  0.07
2       Historic Site  0.07
3              Bakery  0.05
4  Seafood Restaurant  0.05


----Dublin----
            venue  freq
0     Coffee Shop  0.08
1             Pub  0.08
2           Hotel  0.06
```

These results were placed in a dataframe with the top 10 most common venue categories for each city, as per table below:

	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	Berlin	Hotel	German Restaurant	Coffee Shop	Italian Restaurant	Wine Bar	Café	Clothing Store	Boutique	Bookstore	Exhibit
1	Boston	Italian Restaurant	Historic Site	Coffee Shop	Seafood Restaurant	Bakery	Sandwich Place	Pub	Park	Hotel	Salad Place
2	Dublin	Pub	Coffee Shop	Hotel	Restaurant	Bed & Breakfast	Clothing Store	Café	Theater	Department Store	Italian Restaurant
3	Edinburgh	Hotel	Bar	Restaurant	Pub	Café	Indian Restaurant	Whisky Bar	Comedy Club	Coffee Shop	Cocktail Bar
4	London	Theater	Hotel	Cocktail Bar	Plaza	Pub	Japanese Restaurant	Ice Cream Shop	Steakhouse	Monument / Landmark	Bakery
5	Melbourne	Coffee Shop	Café	Bar	Dessert Shop	Cocktail Bar	Shopping Mall	Clothing Store	Sushi Restaurant	Burger Joint	Juice Bar
6	Montreal	Bus Station	Convenience Store	Residential Building (Apartment / Condo)	Furniture / Home Store	Business Service	Yoga Studio	Event Space	Food Court	Food & Drink Shop	Flower Shop
7	Munich	Café	Bavarian Restaurant	Plaza	Clothing Store	Hotel	Boutique	German Restaurant	Coffee Shop	Italian Restaurant	Seafood Restaurant
8	New York	Coffee Shop	Hotel	Café	Sandwich Place	Clothing Store	Cocktail Bar	Park	Burger Joint	Salad Place	Pizza Place
9	Paris	French Restaurant	Ice Cream Shop	Plaza	Art Gallery	Gay Bar	Cocktail Bar	Wine Bar	Pub	Coffee Shop	Tea Room
10	Tokyo	Historic Site	Café	Convenience Store	Italian Restaurant	French Restaurant	Japanese Restaurant	Park	Chinese Restaurant	Lounge	Sake Bar

Since the number of cities may vary, the number of clusters for the K-means clustering was set depending on the number of cities divided by 3. Due to this division by 3 ideally a student enters at 6 cities or more. In any case, if the integer division results in 1 this is automatically forced to 2 to ensure that we get at least 2 clusters.

Once the clustering is run the resulting cluster number is added as a new column to the dataframe of the top 10 venue categories per city. Finally each cluster was displayed listing similar cities in each cluster iteratively, depending on the number of clusters as shown below.

Examine Cluster 0 of similar cities by most common venues

	City	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	10th Most Common Venue
0	London	0	Theater	Hotel	Cocktail Bar	Plaza	Pub	Japanese Restaurant	Ice Cream Shop	Steakhouse	Monument / Landmark	Bakery
2	Melbourne	0	Coffee Shop	Café	Bar	Dessert Shop	Cocktail Bar	Shopping Mall	Clothing Store	Sushi Restaurant	Burger Joint	Juice Bar
3	Munich	0	Café	Bavarian Restaurant	Plaza	Clothing Store	Hotel	Boutique	German Restaurant	Coffee Shop	Italian Restaurant	Seafood Restaurant
4	Berlin	0	Hotel	German Restaurant	Coffee Shop	Italian Restaurant	Wine Bar	Café	Clothing Store	Boutique	Bookstore	Exhibit
6	New York	0	Coffee Shop	Hotel	Café	Sandwich Place	Clothing Store	Cocktail Bar	Park	Burger Joint	Salad Place	Pizza Place
8	Edinburgh	0	Hotel	Bar	Restaurant	Pub	Café	Indian Restaurant	Whisky Bar	Comedy Club	Coffee Shop	Cocktail Bar
9	Dublin	0	Pub	Coffee Shop	Hotel	Restaurant	Bed & Breakfast	Clothing Store	Café	Theater	Department Store	Italian Restaurant
10	Paris	0	French Restaurant	Ice Cream Shop	Plaza	Art Gallery	Gay Bar	Cocktail Bar	Wine Bar	Pub	Coffee Shop	Tea Room

Examine Cluster 1 of similar cities by most common venues

	City	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	10th Most Common Venue
5	Montreal	1	Bus Station	Convenience Store	Residential Building (Apartment / Condo)	Furniture / Home Store	Business Service	Yoga Studio	Event Space	Food Court	Food & Drink Shop	Flower Shop

Examine Cluster 2 of similar cities by most common venues

	City	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	10th Most Common Venue
1	Tokyo	2	Historic Site	Café	Convenience Store	Italian Restaurant	French Restaurant	Japanese Restaurant	Park	Chinese Restaurant	Lounge	Sake Bar
7	Boston	2	Italian Restaurant	Historic Site	Coffee Shop	Seafood Restaurant	Bakery	Sandwich Place	Pub	Park	Hotel	Salad Place

4. The Results

As I did not have a specific student in mind, but developed a generic model the results will depend on the cities being considered by the student and the particular student's interests. In order to discuss possible results I will assume that the student is considering the 10 cities below and also that the main interests of the student are the 5 listed here:

Cities being considered

	City	Country
0	London	United Kingdom
1	Tokyo	Japan
2	Melbourne	Australia
3	Munich	Germany
4	Berlin	Germany
5	Montreal	Canada
6	New York	United States
7	Boston	United States
8	Edinburgh	United Kingdom
9	Dublin	Ireland
10	Paris	France

Student's Interests

	interest
0	Theater
1	Basketball
2	Track
3	University
4	Italian
5	Dance

4.1 Cost of Living and Safety

1. The results obtained in terms of cost of living and safety, are as follows:
   1.1. the least costly city to live in is Montreal, which also enjoys a good level of safety;
   1.2. the safest city to live in is Munich, with a cost of living below the average of the cities being considered;
   1.3. New York is the most expensive and amongst the less safe cities; while
   1.4. London is the least safe city enjoying average cost of living.

4.2 The student's interests

2. The results obtained in terms of matching the students interests, are as follows:
   2.1. Montreal, and to a lesser degree Tokyo, are the least cities likely to meet the expectations of this student;

2.2. New York, and Boston as a close second, are the two cities most likely to satisfy the student's interests;

2.3. The student may next consider the relative importance of his/her interests. For example Boston offer more venues for "Track" activities, whilst New York offers more in terms of "Basketball" and "Dance". Is Track by itself of higher value to this student than Basketball and Dance combined?

2.4. Munich, which is the safest city, addresses all of the student's interest albeit to a lesser extent (compared to say Boston and New York) as there are less venues of each category of interest, however "theatre" is quite well represented in Munich. So again the question of relative importance of each interest to the student is raised.

4.3 Which cities are similar to each other

3. When considering the clustering of the cities, the results are:

3.1. Most of the cities being considered are in fact similar to each other, with a couple of notable exceptions.

3.2. Montreal Canada is one such exception and stands out by itself. Montreal has the least frequent venues dedicated to leisure time. While all the other cities have leisure venues as their most frequent venue. For Montreal it is only the 6th most frequent type of venue which is of a leisurely nature. Montreal seems to be more of a relatively quiet residential city when compared to the other 9 cities being considered.

3.3. Tokyo and Boston also standout for a different reason. Tokyo and Boston have a predominance of historic venues, so students of history may appreciate these two cities for this reason. In our case "history" was not specified as an interest nor do we know if the student will be studying history.

5. Discussion

As the objective was to build a generic model which can be used by different students each considering a different set of cities and each having a different set of interests, the results will vary from student to student.

Given the nature of the analysis, the results will be also extracted during the data exploration itself, and these results also help the student consider his/her options in terms of choosing a city to study from.

The model developed here does not provide a simple definitive answer of one 'best city', but provides food for thought and leaves the final decision to the student. However, it would be relatively simple to develop this model further say by implementing a weighted model, giving weights of importance to:

- Safety;
- Cost of living; and
- Each of the student's interests.

It would subsequently be easy to mathematically come up with a definitive answer to the one best city to study in for that student. However, the correctness of such an answer very much depends on whether the model incorporates all other dimensions of question, and whether the student have specified and correctly weighted all his/her interests.

The model can also be improved further by including other factors such as:

- Colleague / university rankings in that city;
- Cost of tuition;
- Preferred foreign language to learn and proportion of speakers of that language in each city;
- Classification of venues of interest using Foursquare venue category hierarchy; and
- so on.

However, it is not possible to be exhaustive and one must ultimately strike a balance between the adding more complexity to the model and the value derived from the added complexity.

Finally, one highly practical enhancement would be to implement an HTML front end for this project where a student would enter the list of cities and his/her interests and get the results of the analysis, within seconds with just one click.

6. Conclusion

In this study I have considered a relatively common question, but one which is not easy to answer and where the answer has lifelong implications for a student's future.

I have extracted data:

- by scraping 2 web pages;
- using the geolocator of Geopy; and
- using the API to obtain location venue data from Foursquare, both searching for specific categories of venues and also to explore venues around the centre of each city.

The data has been visualised using different graph types and finally analysed using the K-means clustering machine learning technique. While the results are crude, the model is useful as is, up to a point. It is also quite possible and to develop this model further into a fully operational commercial application as discussed above.

End of report