# Best city to study in

## Joseph M Camilleri

### 2nd August 2020

## 1. Introduction

### 1.1 Background

Many students from around the world study in a city other than their home town, many also studying abroad. Their stay in another city is generally of a temporary nature in order to study in an institution that is renowned for their level of academic achievement in a particular subject. Other students choose to follow a multi-year programme in some other city or indeed some other 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 a new language. Other advantages include: making new friends and discovering new interests. Other key advantages are the gaining of new experiences and the related personal development. Studying in a different city also opens up new career opportunities which may otherwise not be available.

### 1.2 Problem

Such student often have a choice to make between two or more cities, as to where to relocate for one or more years to complete their studies. While each student’s case is different in some way, 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.

On the other hand this is not always the case, and often students are required to choose amongst a number of institutions, each located in different cities or even different countries around the world.

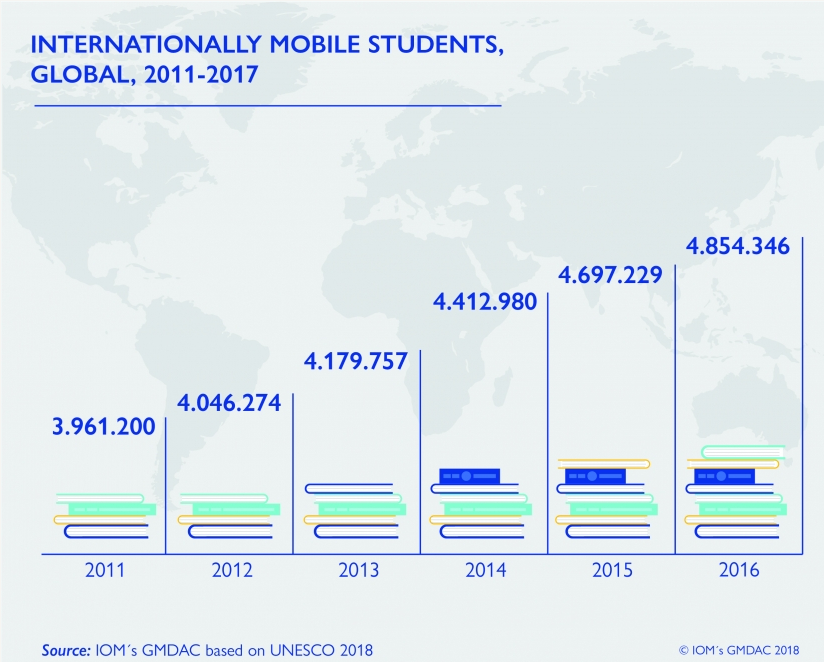
Hence, the final decision will also be influenced by the preferred country and city where that institution is located. Putting it another way, when a student chooses to study in a specific institution they are also choosing to live and study in a particular city.

The objective of this assignment, is to provide some assistance to students in establishing 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 [United States] students studied abroad for credit in 2017-18, representing a 2.7 percent increase from the previous academic year, according to the annual ["Open Doors" report](https://www.iie.org/opendoors), published by the Institute of International Education with funding from the U.S. Department of State. [source: <https://www.insidehighered.com/news/2019/11/18/open-doors-data-show-continued-increase-numbers-americans-studying-abroad>]



The aim of this assignment is to enable prospective students to analyse and compare different cities around the world, in relation to their own personal interests. For example, if a student is also an outdoors activities enthusiast he may give greater consideration to studying in a city which offers more opportunities for such outdoor 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 others students and make new friends.

## 2. Data acquisition and cleaning

### 2.1 Data sources

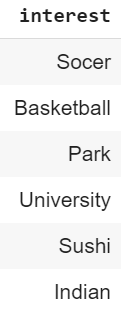
The assignment is design to be used by different students, one at a time. Hence the first step is to capture the set of cities being considered by that student. The cities are introduced into the Jupyter Notes book as a simple dataframe consisting of the city name and country name. I found that the name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



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.



When considering living in another city for this first time, two key considerations will generally be the level of safety and relative cost of living in each city being considered. This data can be obtained by scraping from the following web sites:

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

Below please find a screen shot from this site.



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

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



The longitude and latitude of each city being considered by the student is obtained from 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/>]

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



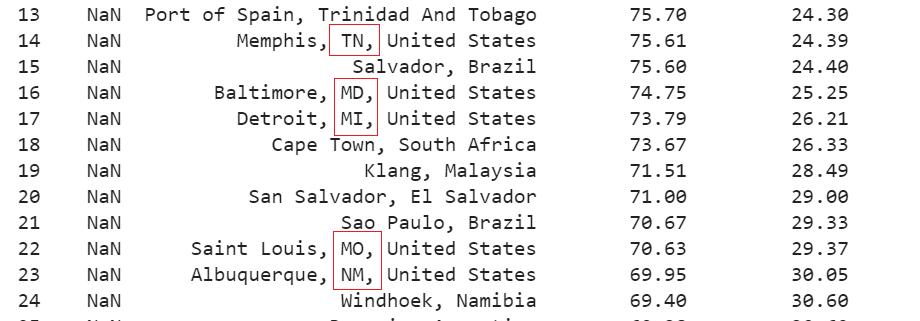
### 2.2 Data cleaning and preparation

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

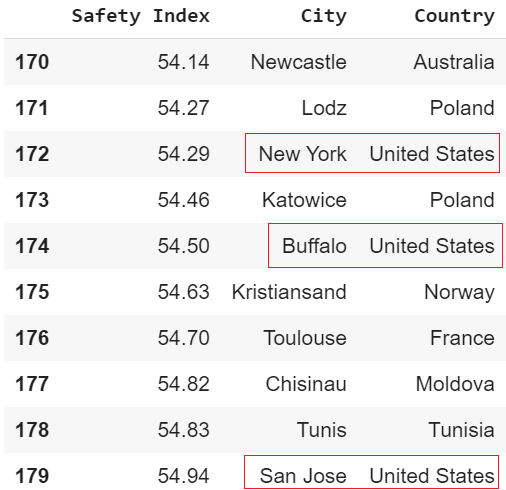
The data scraped from the 2 web sites is initially captured into a number of lists, and the relevant list is converted into a dataframe. In both web sites, the city name was a text string containing the

* Name of the city,
* Name of the Country, and
* in some cases (particularly for United States cities) also the state initials

each separated by a comma. Refer to examples below:



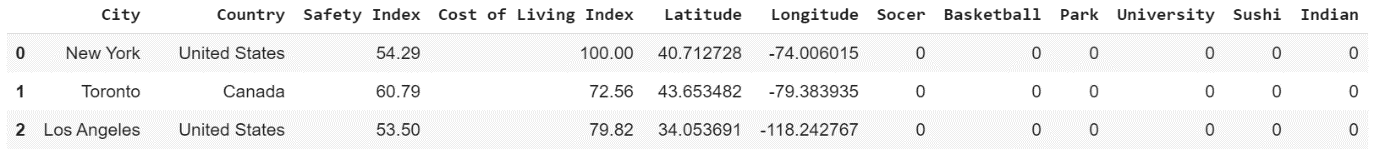
For consistency and in order to match these value with the student’s cities of interest. This had to be split out into separate 3 columns and the state initials data dropped, to have just city name and country name as per table below.



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

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

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



These interest columns were populated with the number of instances of relevant venues for that interest searched within that city using the Foursquare API.

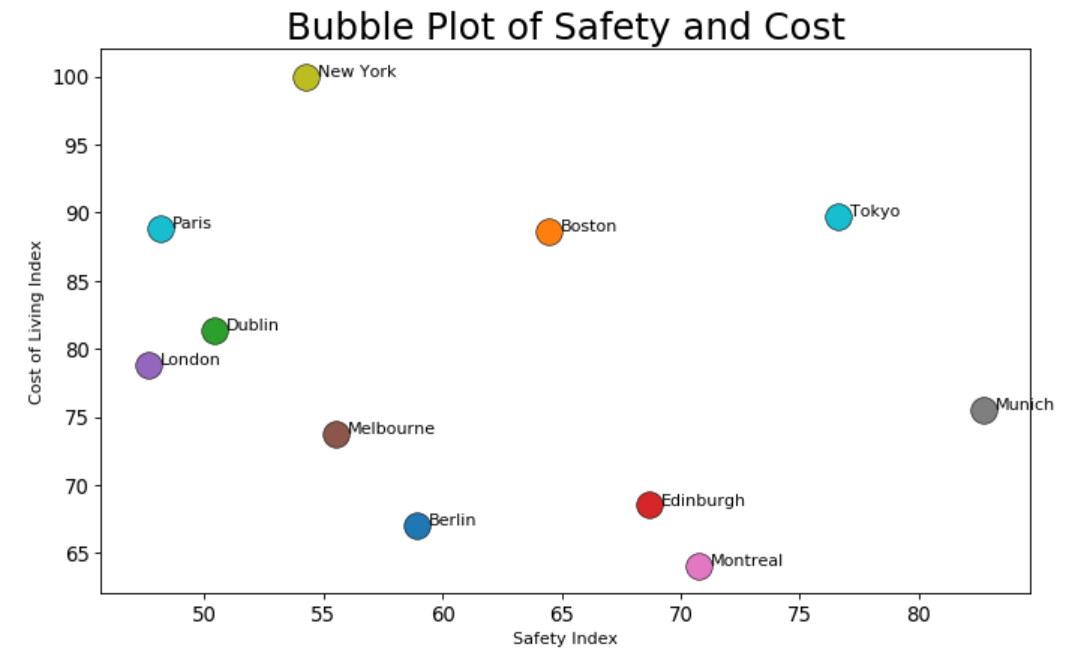
The numbers of relevant venues in each city, the safety index as well as the cost of living index were normalised to allow for visualisation of relative values.

Finally I also obtained “explore” data from Foursquare for each city of interest. From the Foursquare data, I extracted the relevant values, including the category of the venue. This data was transformed into a one hot encoding dataframe, grouped by city. The frequency of each venue category extracted was 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.

## 3. Methodology

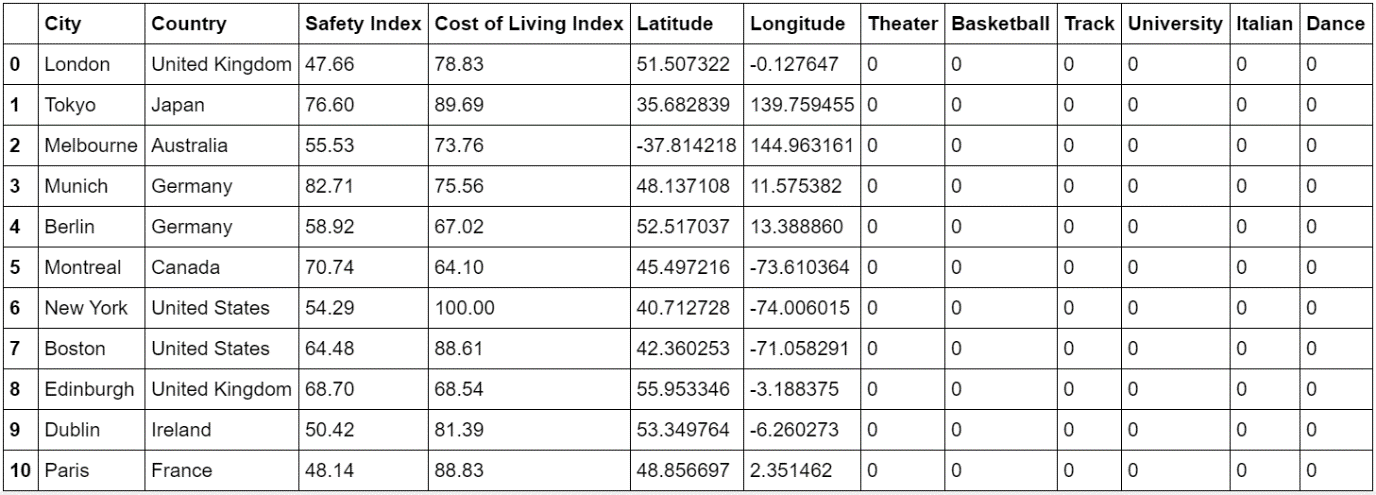
### 3.1 Exploratory data analysis

As the primary considerations will be safety and cost a living, 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.

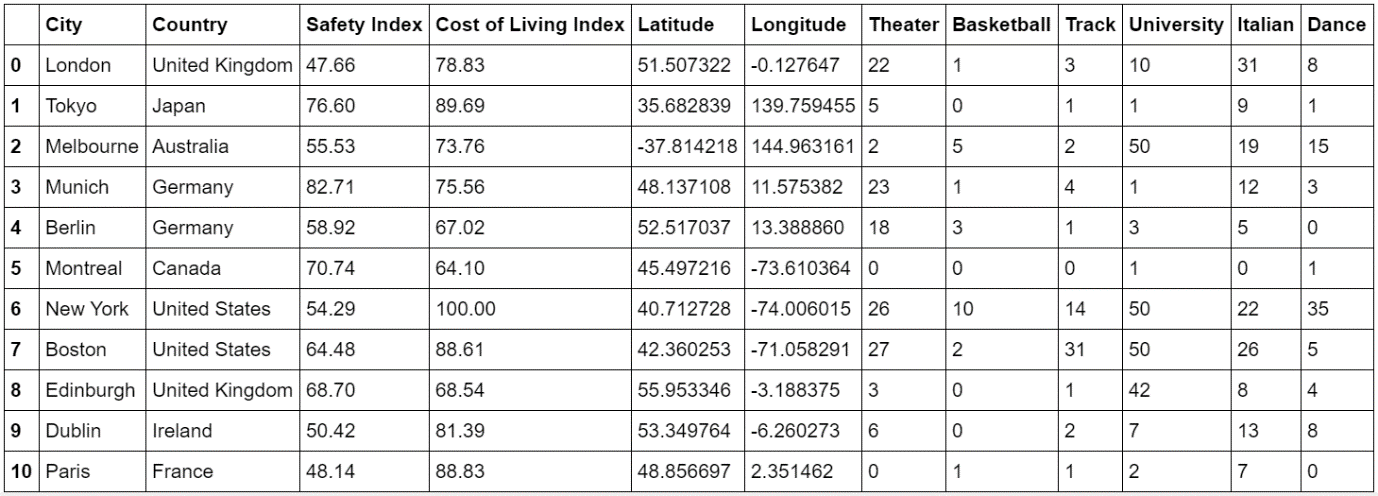


In order to achieve the above I merged the Cities dataframe with that scrapes from the websites for the Safety Index and Cost of Living Index by City. From the above graph it is very easy to see that Munich is the safest while Montreal enjoys the lowest Cost of Living from amongst the selected cities.

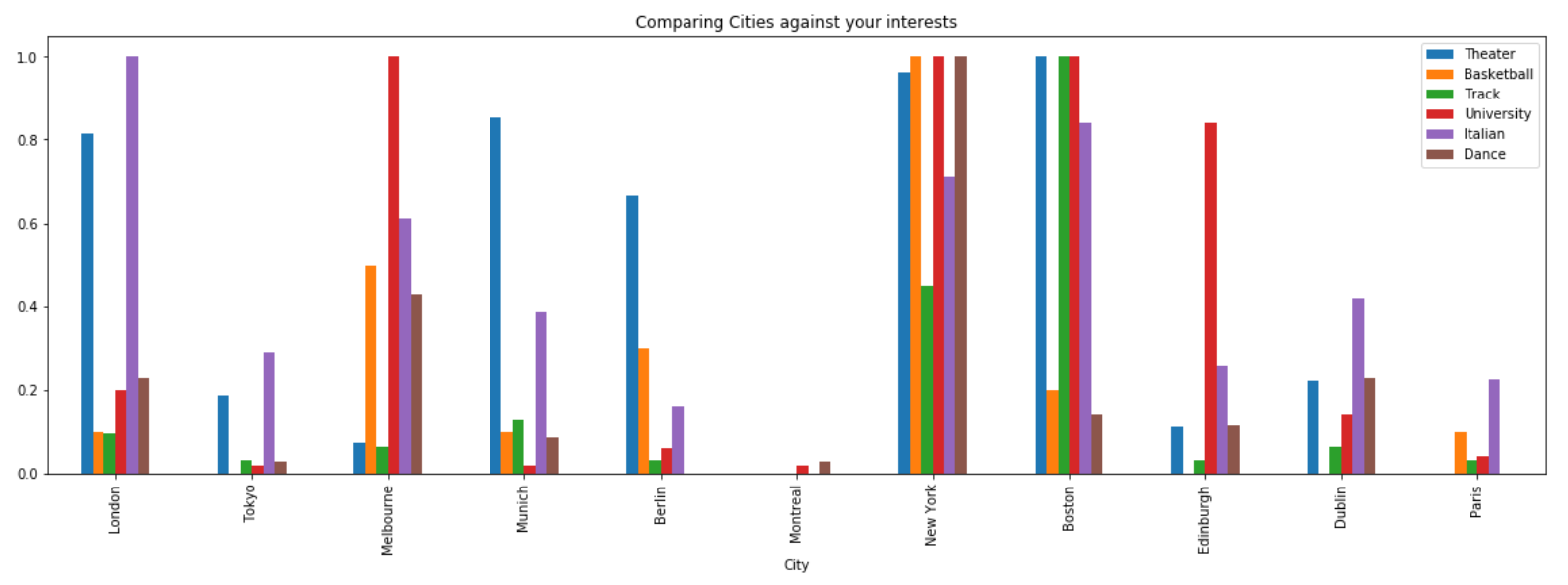
Next, and in order to analyse the cities against the student’s interest, I needed to augment further the table of Cities with the longitude and latitude of each city plus a column for each of the interests specified by the student, as follows:



The student’s interest columns (“Theater” to “Dance” above) where populated by counting the number of matching venues from Foursquare API for each City.



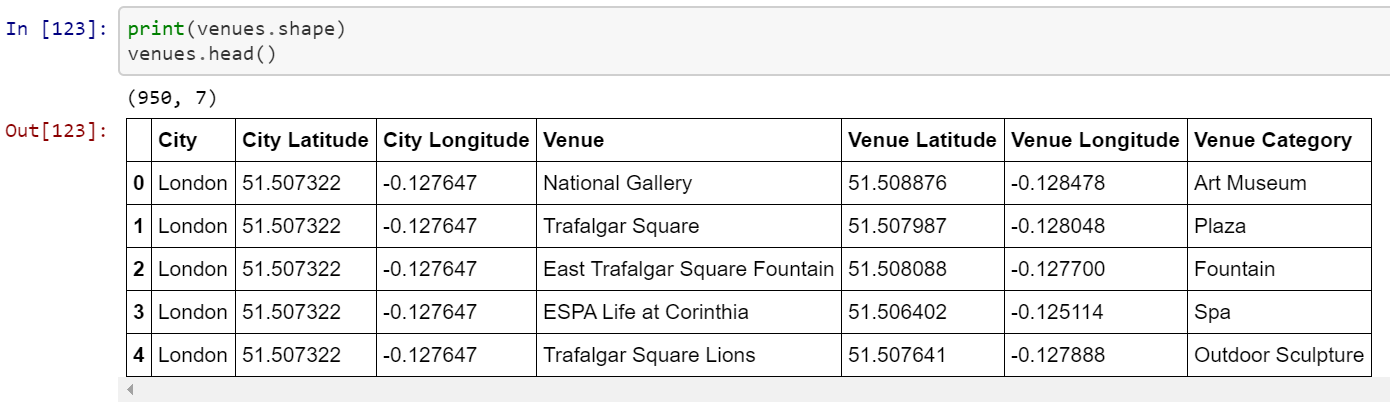
Next I wanted to visualise this data to get a sense of the degree to which each city addressed in interests to the students relative to all the other cities. Firstly the number shown above had to be normalised and subsequently plotted on a bar chart for each city.



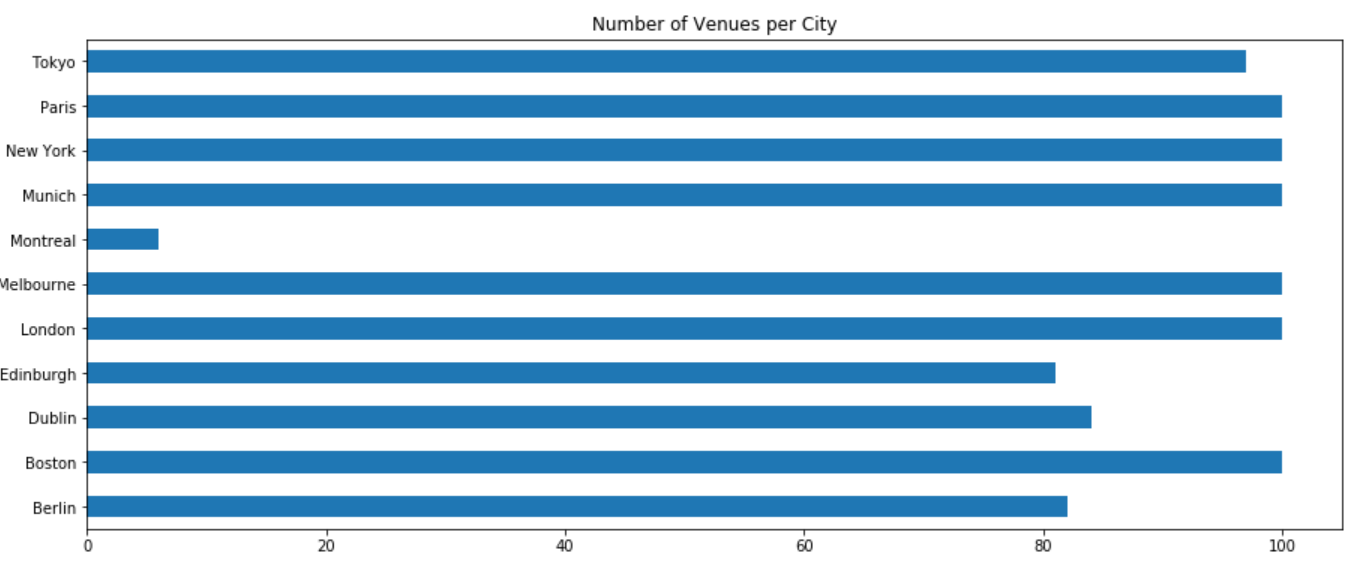
From the above one easily appreciate the New York and Boston are the two cities that most closely match the interests of the student. The student can also consider the relative importance of each interest. For example Boston offer more venues for “Track” activities, whilst New York offers more in terms of “Basket Ball” and “Dance”.

This graph also helped confirm the correctness of the results. For example the city of Mumbai in India obtain a relatively very high score compared to other cities when Indian Food was set as an interest.

The next stage in the analysis of the Cities was to explore and compare the venues in general in 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, as depicted in the sample of rows below.



The following horizontal bar chart also helped me visualise the number of venues extracted per city.



Displaying the unique categories was also helpful in exploring and understanding the data.

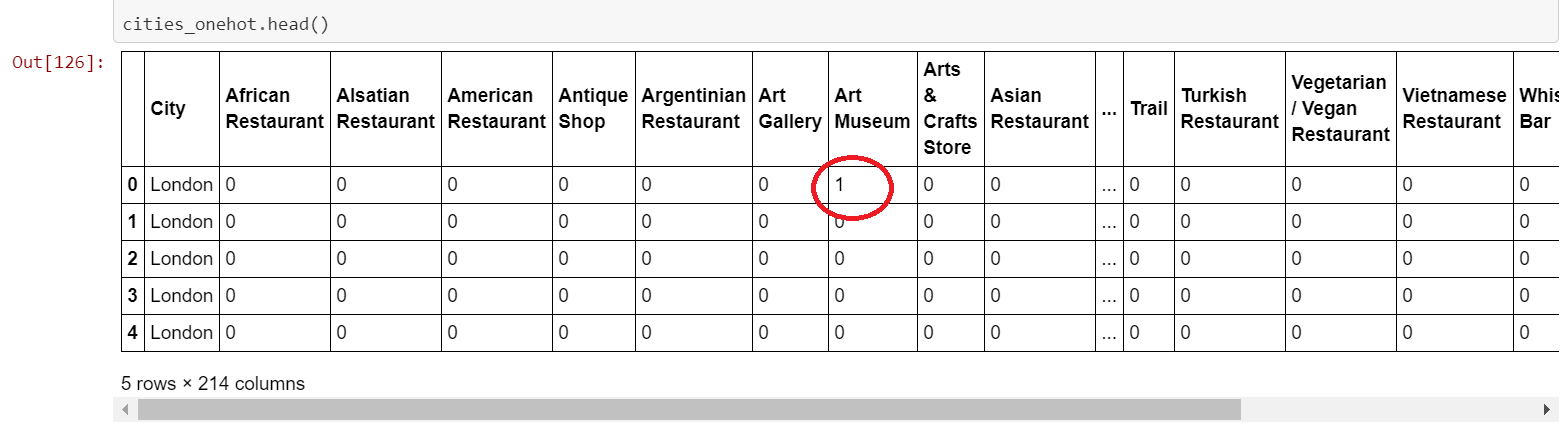


### 3.3 Machine learning

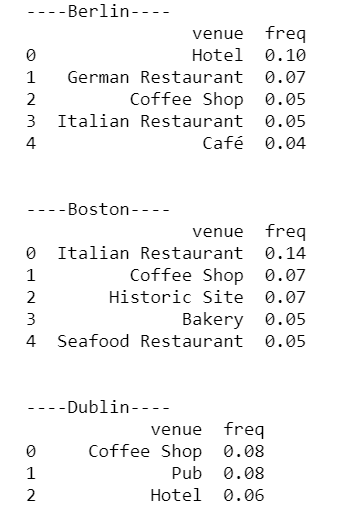
The final stage was to compare the cities of interest to determine which were more alike than others. This comparison was to be done based on the relative frequency of venue categories in each city. Thus the student will know that a group of cities (from those under consideration) are alike and different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

This grouping of similar cities was achieving using the K-means clustering technique. Naturally, this cannot be geographical clustering as the cities may be in different continents, but clustering based on the similarity of their venues. Thus a student can expect a similar experience in any of the cities within a cluster, and a different experience between cities in different clusters.

The first step in preparing the data for this 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.



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



These results were placed in a dataframe with the top 10 most common venue categories for each city, as follows.



The number of clusters for the K-means clustering was set to the number of cities divided by 3. This provided a variable number of clusters based on the number of cities entered. It also ensure that the clustering results in more than one group.

Once the clustering is run the resulting cluster number is added to the dataframe of the top 10 venues per city.

Finally each cluster was displayed listing similar cities in each cluster iteratively and depending on the number of clusters.



## 4. Results

Given the cities and interests used below:

|  |  |
| --- | --- |
| **Cities being considered** | **Student’s Interests** |
|  |  |

### 4.1 Cost of Living and Safety

1. The results obtained in terms of cost of living and safety, are as follows:
   1. the least costly city to live in is Montreal
   2. the safest city to live in is Munich, with a cost of living still below the average of the cities being considered
   3. New York is the most expensive
   4. London is the least safe

### 4.2 The student’s interests

1. The results obtained in terms of matching the students interests, are as follows:
   1. Montreal is the least city likely to meet the interest of the student
   2. New York, and Boston as a close second are the two cities most likely to satisfy the student’s interests
   3. The student subsequently also consider the relative importance of each interest. For example Boston offer more venues for “Track” activities, whilst New York offers more in terms of “Basket Ball” and “Dance”.
   4. Munich, which is the safest city, addresses all of the student’s interest albeit to a lesser extent as there are less venues of each category of interest, however theater is quite well represented in Munich.

### 4.3 Check out similar cities

1. When considering the clustering, the results are:
   1. Most of the cities being considered as similar with some notable exceptions
   2. Montreal in Canada stands out by itself, as having the least frequent venues dedicated to leisure time. While all the other cities have leisure venues are their most frequent venues, within Montreal it is only the 6th most frequent type of venue which is of a leisurely nature. Montreal seems to be more of a quiet residential city.
   3. Tokyo and Boston also standout as have a high frequency of historic venues, so student of history may appreciate these cities for this reason.

## 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 analysis itself.

The model developed here does not provide a simple definitive answer of one ‘best city’, but food for thought for the student. It would be relatively simple to develop this further by implementing a weighted model, giving weights to

* Safety;
* Cost of living; and
* Each of the student’s interests.

It would subsequently be each to mathematically come up with a definitive answer to the one best city to study in for that student.

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

* Colleague / university ranking;
* Cost of tuition;
* Preferred foreign language to learn and proportion of speakers of that language in each city; and
* So on

## 6. Conclusion

In this study I considered a relatively frequent question, the answer to which can have lifelong implications on for a student’s future. I have extracted data

* by scraping web sites;
* the geolocator of Geopy; and
* location venue data from Foursquare.

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

-------------- end of report ------------------------