Best city to study in

Joseph M Camilleri

31st July 2020

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:

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 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.

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 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

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/]

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

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:

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 3 columns and the state initials data dropped, to have just city name and country name as per table 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 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.

	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 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.

------- end of data section -------