Clustering Colleges Based on Neighbourhood Similarities

Deepratna Awale

3rd November, 2019

1. Introduction

1.1 Background

Every year millions of students pursue master's as higher education. In this highly competitive era, western colleges are preferred by many students. Both domestic and international students try to pursue their dream to get in their college of choice. However, this is not always possible, the most elite and sharp students, or those with a proven track record, or simply those who submitted the application first, get in first (à-la-carte basis). So, students who didn't get in would have to go to other colleges. When it comes to domestic students, they have their own comfort zone and preferences, say a student in New York would prefer to stay in the vicinity. As for international students, they want a good cultural demo graph, a safe neighbourhood, affordable city and a similar area to their college of choice. What if you could choose other colleges based on your college of choice?

1.2 Problem

Colleges, if chosen randomly can have unfavourable conditions for studies. However, they can be chosen so that they have the same ‘characteristics’ as our dream college. The goal of this project is to cluster colleges, based on neighbourhood their similarity.

1.3 Target Audience / Interest

My main target audience is students. Specifically, students like me, dreaming to pursue a Master’s in ‘Computer Science & Information Technology’ in Data Science / Analytics. This would help students to decide colleges while completing their Application-process. It would be easier to choose colleges without worrying about the how the city / living experience would be, because this project will do so for them.

2. The Data

2.1 Data Source and Requirement

We are using www.mastersportal.com to mine the data. We are searching for Canada, United States, United Kingdom based colleges/ universities that provide full time Master's Degree Programme in the field of Computer Science and IT. The method is simple, we just search and get the results, the returned results can have multiple pages, each page has 10 colleges, depending on the number of colleges returned by search, the number of web pages to scrape increases.

We have the following details available:

Degree, Density, Full Time Duration, ID, level, listing_type, logo, organization, organization_id, summary, tution fee, Address: area, city, country.

2.1.1 Required Data:

- College Id
- College Name
- College Fees
- Address
- Geospatial Coordinates

2.1.2 Purpose of Data:

- College Id: To have distinct college despite the name.
- College Name: To get its geospatial data.
- College Fees: To segregate colleges based on tuition.
- Address: To get geospatial data of college and plot maps
- Geospatial Coordinates: To get places nearby, pinpoint college on map

2.2 Data Acquisition and Cleaning

2.2.1 There are two types of data that we need:

- College Details (JSON Data)
- Geospatial Coordinates (JSON Data)

2.2.2 Methods used of obtaining data:

- College Details: Web Scraping
- Geospatial Coordinates: Google Places API

2.2.3 Getting the College Details JSON

We have to intercept the webpage JSON traffic. We are looking for a json that fills up search requests. So, the column Type would be ‘json’, and since we are requesting for data, the network fetches data so cause is ‘fetch’. One of the Domains fulfils the needs, the 'search.prtl.co' domain. Voila! We have the link of a json string that generates the result. We can save the json from there but well good luck with 214 files. We need to concatenate the files into a single one (receive data in a single file and append it).

Since the data is coming from 213 webpages, the json file strings ends 213 times i.e. there’s '][' (end and start of json) in between which should be ',' (comma to continue the json). So, I manually replaced them using the editor. A simple find and replace would do.

2.2.4 Cleaning the College Details JSON

The College Details JSON is loaded into a pandas data frame using pandas native json normalization method. Since there’s heavy nesting in the venues area, we’ll make a separate data frame for it.

We join the main data frame and venues data frame to form a complete dataset. We keep only the required columns:

- College Id: ‘id’
- College Name: ‘college_name’
- College Fees: ‘fees’
- Address: ‘City’, ‘Area’, ‘Country’

To give tuition fees a uniform scale, we convert all credit-based fees to yearly fees. We remove any entry where currency is not euros. Since all fees are per year in euros, we don’t need its unit (annual / credit) and currency (Euro).

Note: This causes discrepancies in data which we’ll solve later. Also, the number of credit-based fees were too few compared to annual fees, thus there is no significant impact.

The data frame looked like this after cleaning:

	id	college_name	fees	area	city	country
0	173294	School of Nursing and Health Professions	29047.5	California	San Francisco	United States
1	127982	Kogod School of Business, American University ...	35730.0	Washington, D.C.	Washington, D. C.	United States
2	152893	University at Buffalo, The State University of...	20857.0	New York	Buffalo	United States
3	262046	School of Management	21419.0	England	London	United Kingdom
4	104991	University of South Wales	15514.0	Wales	Pontypridd	United Kingdom
...	...	...	...	...	...	...
2135	61571	Florida International University	20362.5	Florida	West Miami	United States
2136	107225	University of Cincinnati	5774.0	Ohio	Cincinnati	United States
2137	125952	San Diego State University	8075.0	California	San Diego	United States
2138	261320	Rdi Uk	6638.0	England	Coventry	United Kingdom
2139	98012	Rochester Institute of Technology	41100.0	New York	Rochester	United States

2060 rows × 6 columns

Table 1: Clean data frame after normalizing college data

2.2.5 Getting the Geospatial Data

We create two functions, one to get the latitude and longitude of the college and another to repeat the process for all colleges in the data frame. We use the Google Places API's 'findplacefromtext' endpoint to search for the colleges.

The data is written to a json file called 'G-lat-long.json' and can be accessed later.

The return fields are geometry/location, which will give us the latitude and longitude of the colleges. We are trying to find the college with its name and city first, if not found then use college name only. We successfully obtained 2057 records of colleges and 3 were not obtained.

id	college_name	fees	area	city	country	latitude	longitude
173294	School of Nursing and Health Professions	29047.5	California	San Francisco	United States	37.776567	-122.450309
127982	Kogod School of Business, American University ...	35730.0	Washington, D.C.	Washington, D. C.	United States	38.938340	-77.087635
152893	University at Buffalo, The State University of...	20857.0	New York	Buffalo	United States	43.002837	-78.787595
262046	School of Management	21419.0	England	London	United Kingdom	51.522407	-0.131678
104991	University of South Wales	15514.0	Wales	Pontypridd	United Kingdom	51.589239	-3.330827

Table 2: Pandas data frame after inserting geospatial coordinates

The data is saved into 'college_dataset.csv' for further use. Next, we make sure that the colleges we have only belong to USA, UK and Canada. I performed some manual descriptive statistics on the data and found abnormal fee values, caused due to the credit-based calculations, after manually searching the web, I replaced them with actual values made available by college websites. Similarly, I checked the geospatial

coordinates by plotting them on map using folium, I replaced the faulty coordinates with appropriate coordinates by manually searching on google maps. The colleges that were no where to be found on google maps were removed from the data frame. After all the cleaning and manual scraping of data, it was saved to ‘final_college_dataset.csv’. This data is ready for further analysis.

3. Exploratory Data Analysis

3.1 Spatial Data Analysis

I formed three more data frames representing colleges in USA, UK and Canada separately. The data was plotted on map using folium and the following maps were obtained.

- ❖ USA Map

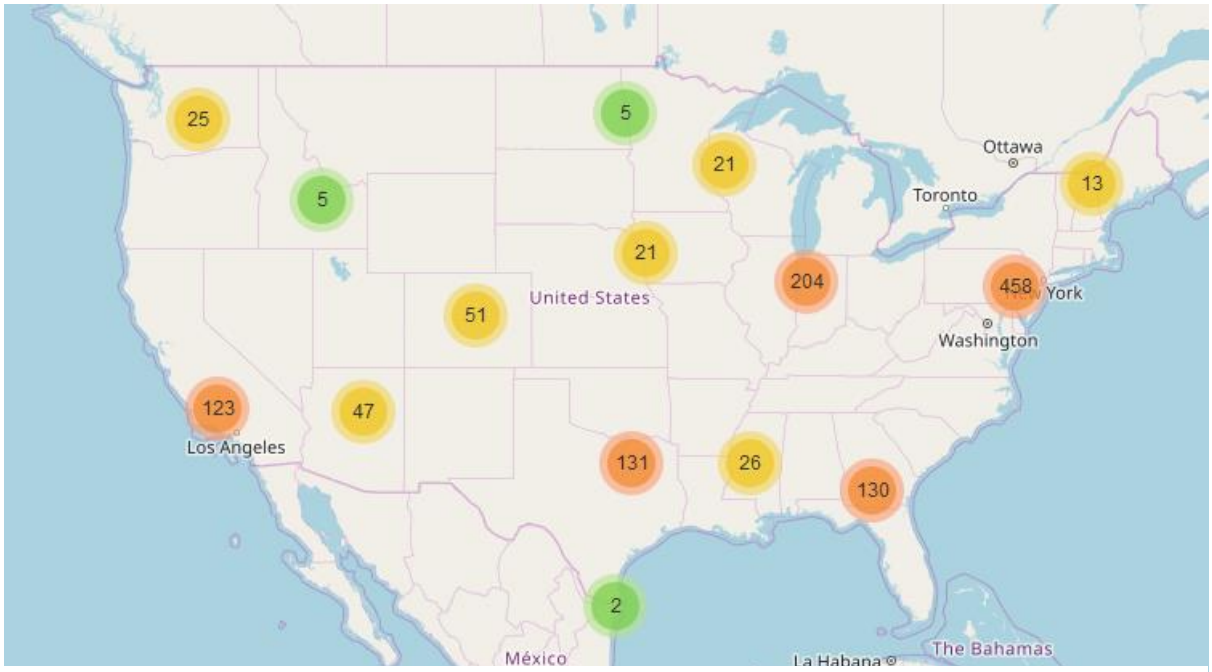


Figure 3: USA map representing college count

I’ve used marker clusters than using markers or circle markers to reduce the clutter.

The right side of map has a very high density of colleges while the top and bottom parts don’t, it wouldn’t be a surprise if the college clusters are heavy in the east.

- Canada Map

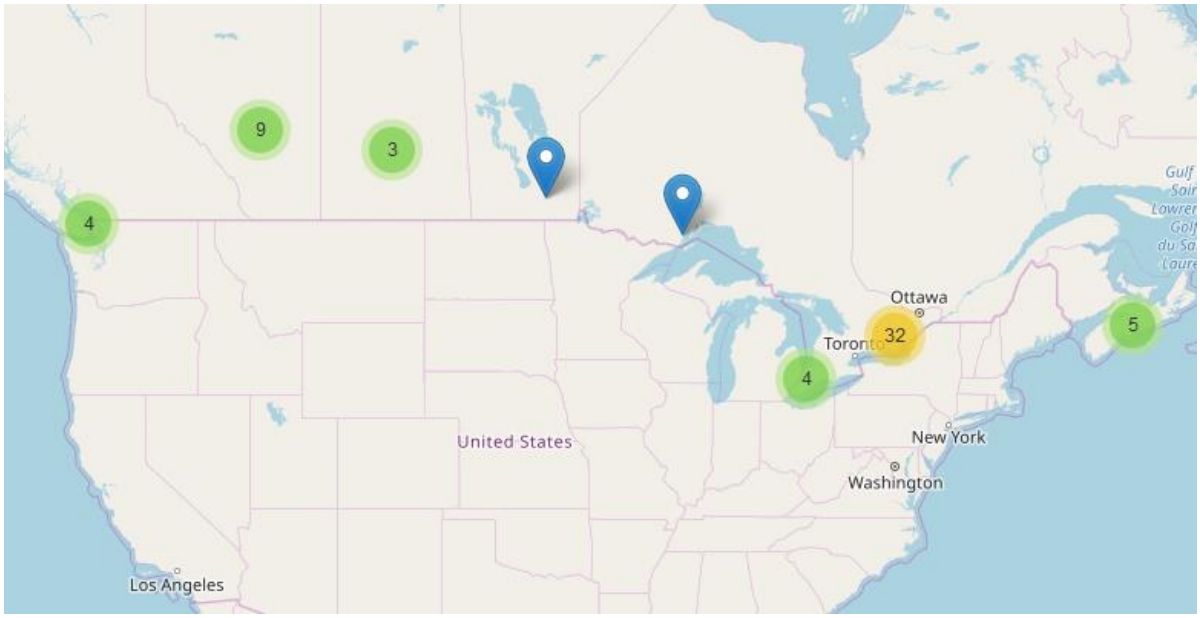


Figure 4: Canada representing college count

The colleges in Canada are heavily concentrated near its financial capital, Toronto. Since there are only a few colleges in Canada, its contribution to changing the environment setting is very low. Canada has a total of 52 colleges in our data set, compared to USA and UK having more than 1200 and 700 colleges respectively, it’s too low.

❖ UK Map

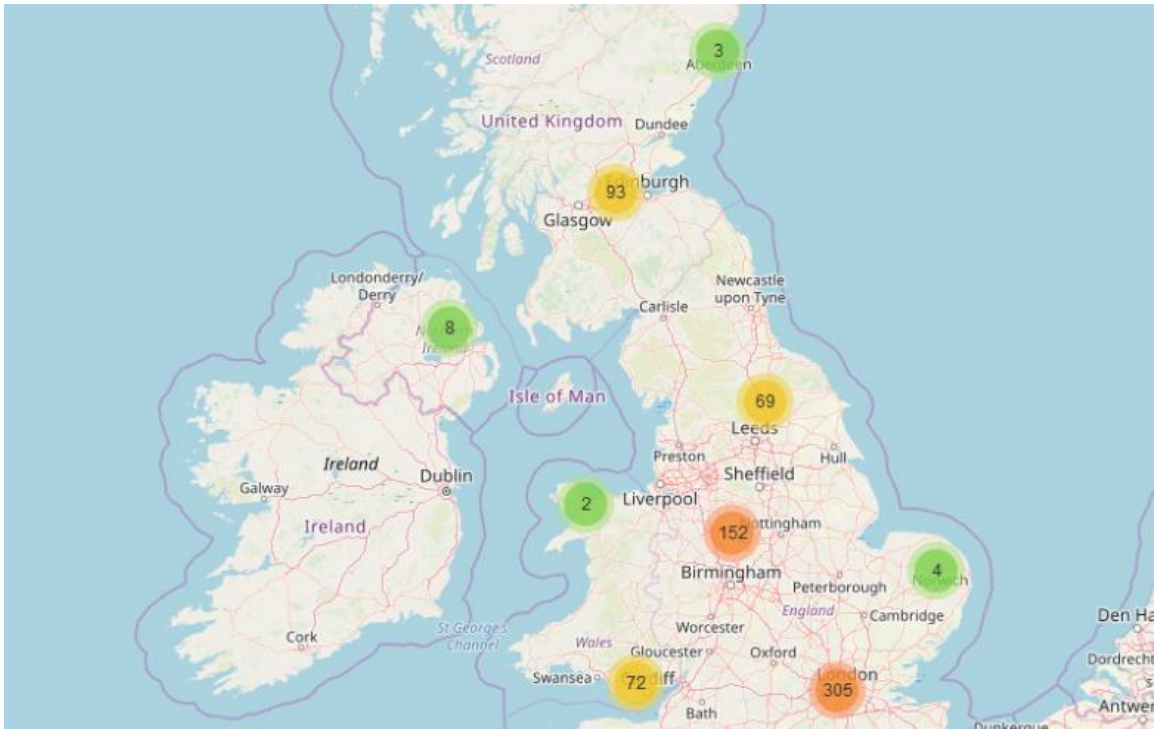


Figure 5: UK map representing college count

United Kingdom has a high concentration of colleges to the south. But overall, UK has the best distribution among all three countries (according to the dataset we obtained).

Obviously, the dataset to begin with isn't perfect as there are just eight colleges in Ireland, even in USA the concentration of colleges to the west is very poor, this maybe because of the course we selected, but anyhow, moving on with the data we have.

3.2 Numerical Data

There's only one kind of numerical data with us, tuition fees. If we look at the distribution of our fees data, the approximation would stop us from concluding anything with confidence,

however, if we assume that the calculated fees is correct / approximately near to the original tuition, then the fees can be accurately represented by a box chart.

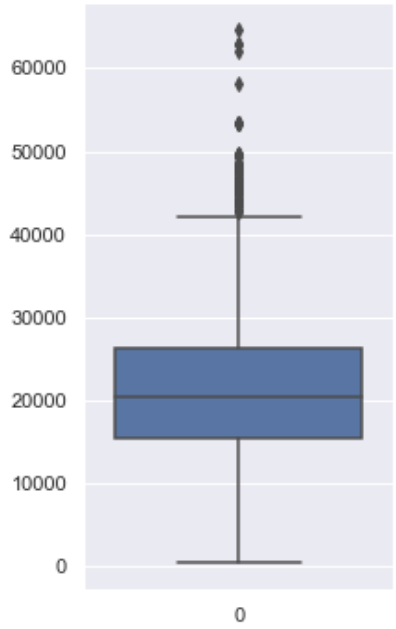


Figure 6: Tuition Fees Box Plot

	id	fees	latitude	longitude
count	2034.000000	2034.000000	2034.000000	2034.000000
mean	158242.061455	21496.101032	43.465296	-58.660623
std	87511.264137	9907.860290	7.933261	43.554852
min	597.000000	371.000000	21.296939	-157.817112
25%	81069.500000	15342.000000	38.765159	-89.252071
50%	152929.000000	20331.000000	41.924756	-76.603588
75%	247536.000000	26166.000000	51.521970	-2.465706
max	287854.000000	68640.000000	64.855757	1.239176

Table 3: Described df

The box plot and the table show us the following:

25% colleges have fees up to 15342.

50% colleges have fees up to 20331.

75% colleges have fees up to 26166.

The highest reported tuition fee is 68640.

Now, let’s compare the countries with respect to tuition.

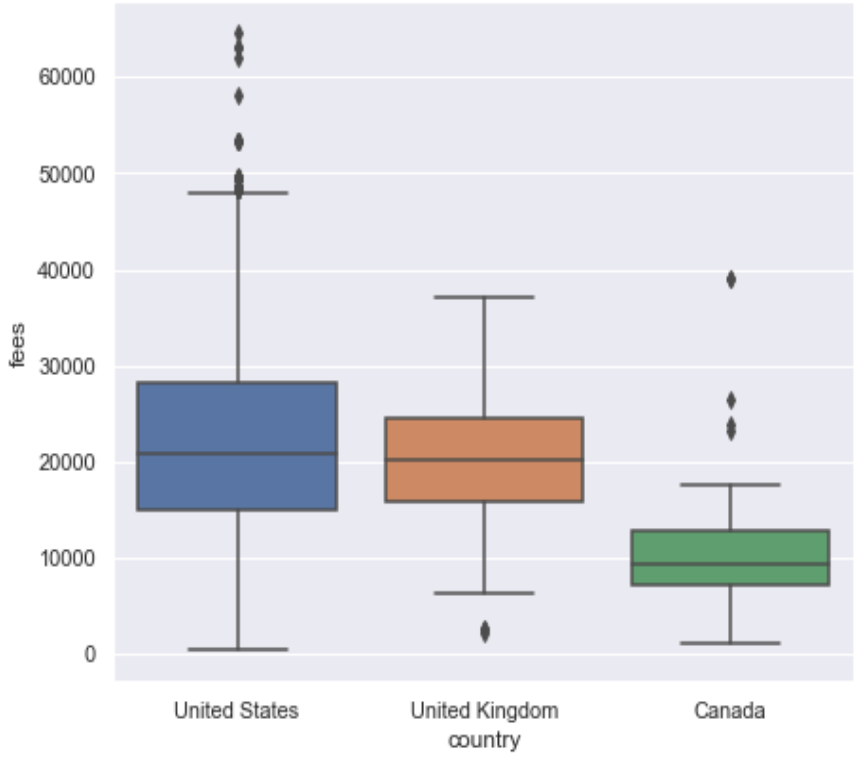


Figure 7: Tuition fee comparison of countries

The box plot of Canada lies within the 50% Quartile range of USA, it won't be surprising if we get more Canadian colleges when looking for similar low fees colleges when holding money as a criterion.

Similarly, UK would be a preferred choice when choosing mid to high fees colleges when compared to USA when holding money as a criterion.

It is obvious, but let me specifically point out, USA has the highest fees in all 3 nations.

4. Four Square Places API

After acquisition of geospatial coordinates and cleaning the dataset, its time to explore each college neighbourhood to cluster them. We use four square api's explore endpoint to get the amenities near the college.

We need the 'type' of the place near the college, like café, bar, restaurant, club, etc. To do this we'll use the 'shortName' key in the response of the request. After getting all the venues, its time to remove the null entries from the obtained list and the previous data frame as well.

If we view the venues obtained grouped by college, we observe the following:

college_name	c_latitude	c_longitude	venue	v_latitude	v_longitude	v_category
Aberystwyth University	120	120	120	120	120	120
Acadia University	20	20	20	20	20	20
Albany College of Pharmacy and Health Sciences	20	20	20	20	20	20
Alliant International University	20	20	20	20	20	20
American National University	40	40	40	40	40	40
...	...	...	...	...	...	...
Wrexham Glyndwr University	60	60	60	60	60	60
Wright State University	60	60	60	60	60	60
Yale University	80	80	80	80	80	80
Yeshiva University	40	40	40	40	40	40
York St John University	20	20	20	20	20	20

629 rows × 6 columns

Table 4: Venue numbers grouped by college

5. Clustering

Before we can cluster the groups, we need to introduce dummy variables (one hot encoding), so as to represent 386 unique categories separately in a column of its own.

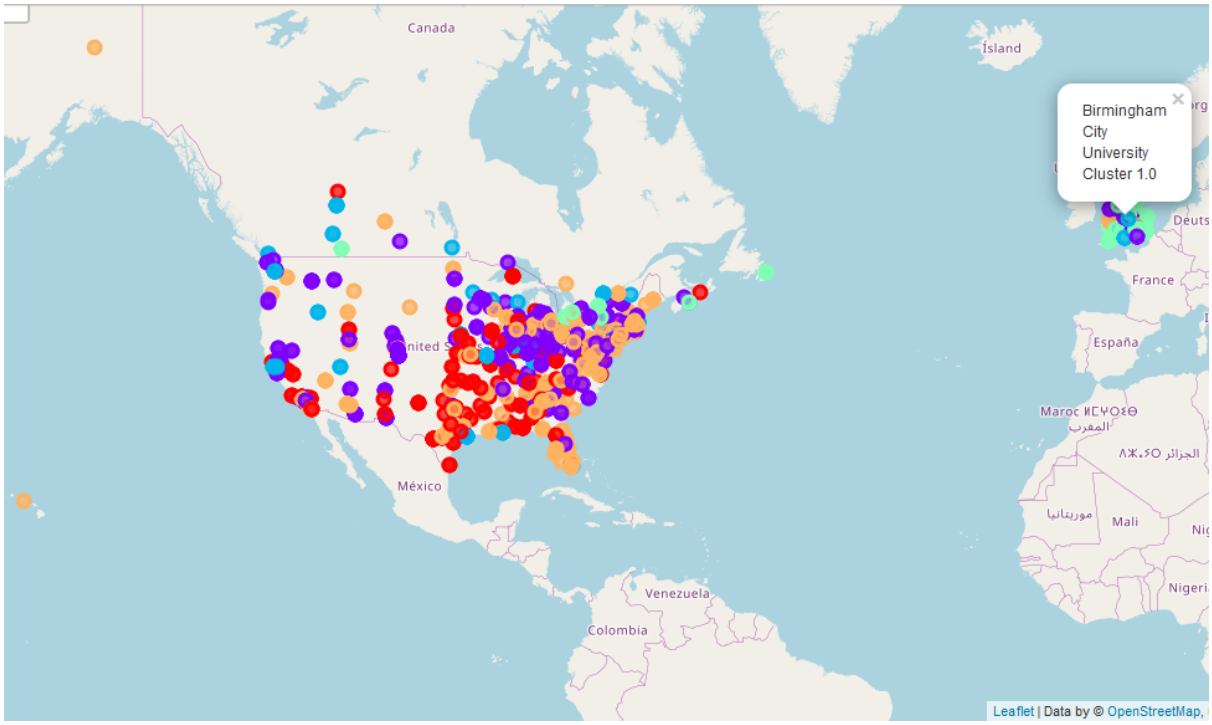
All we need right now is the college name and the one hot encoded columns. We'll make a data frame where colleges are grouped along with the dummy columns that gives us a data frame of dimensions 629x387. What we really need is the top ten most occurring names, so that we can have a data frame that looks like this:

	college_name	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	Aberystwyth University	Hotel	Bookstore	Italian	Fast Food	Mediterranean	Café	Surf Spot	Beach	Bar	Library
1	Acadia University	Café	Coffee Shop	Brewery	Hotel	Farmer's Market	Sandwiches	Movie Theater	Pharmacy	Ice Cream	Restaurant
2	Albany College of Pharmacy and Health Sciences	Café	Deli / Bodega	Sandwiches	New American	Park	Burgers	Bar	Vietnamese	Mexican	Ice Cream
3	Alliant International University	Sandwiches	Brewery	Sushi	Vietnamese	Lake	Burgers	Juice Bar	Korean	Gym / Fitness	Falafel
4	American National University	American	Supermarket	Coffee Shop	Brewery	Fast Food	Pizza	Cosmetics	Steakhouse	Donuts	Bakery

Table 5: Top 10 common venue near colleges

We'll be using K-means clustering of degree 5 to cluster the colleges. After Clustering, we get a map with 5 coloured dots/ circles, notice how the cyan circle exists only in UK (majorly).

Canadian and American colleges are almost undistinguishable based on cluster, so they are not clustered on the basis on nation either.



6. Conclusion

The colleges have been genuinely clustered on the basis of their neighbourhood, and have an indefinite trend. From the clusters it is apparent that the neighbourhood of UK will majorly defer from that of USA or Canada. Hence, I've successfully clustered the college neighbourhood into the following categories:

1. American Eats
2. Exotic Eats
3. Tour/ Outgoing
4. Night Life (Pub) and Fitness
5. Art Prone/ Mature Audience

7. Future Possibilities

By clustering such colleges, we can sort colleges according to our whim by putting constraints like, want café or want gym. This will help us shortlist colleges for the application process.

Also, by introducing money as an independent variable, we can predict what kind shop should be opened near a college or where should we rent a place to get the most benefit.