Clustering Colleges Based on Neighbourhood Similarities

Deepratna Awale

Introduction

- Millions of students pursue master's as higher education.
- They have their own expectations from colleges.
- For Domestic students its comfort zone, for international its security.
- What if you could choose other colleges based on your college of choice?
- You can, by clustering them!
- 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

Data requirement and acquisition

- 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
- Required Data:
  - College Id
  - College Name
  - College Fees
  - Address
  - Geospatial Coordinates

- There are two types of data that we need:
  - College Details (JSON Data): Web Scraping
  - Geospatial Coordinates (JSON Data): Google Places API

We'll be collecting college data in a single JSON file by appending to it all the search results that we obtained.

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'

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

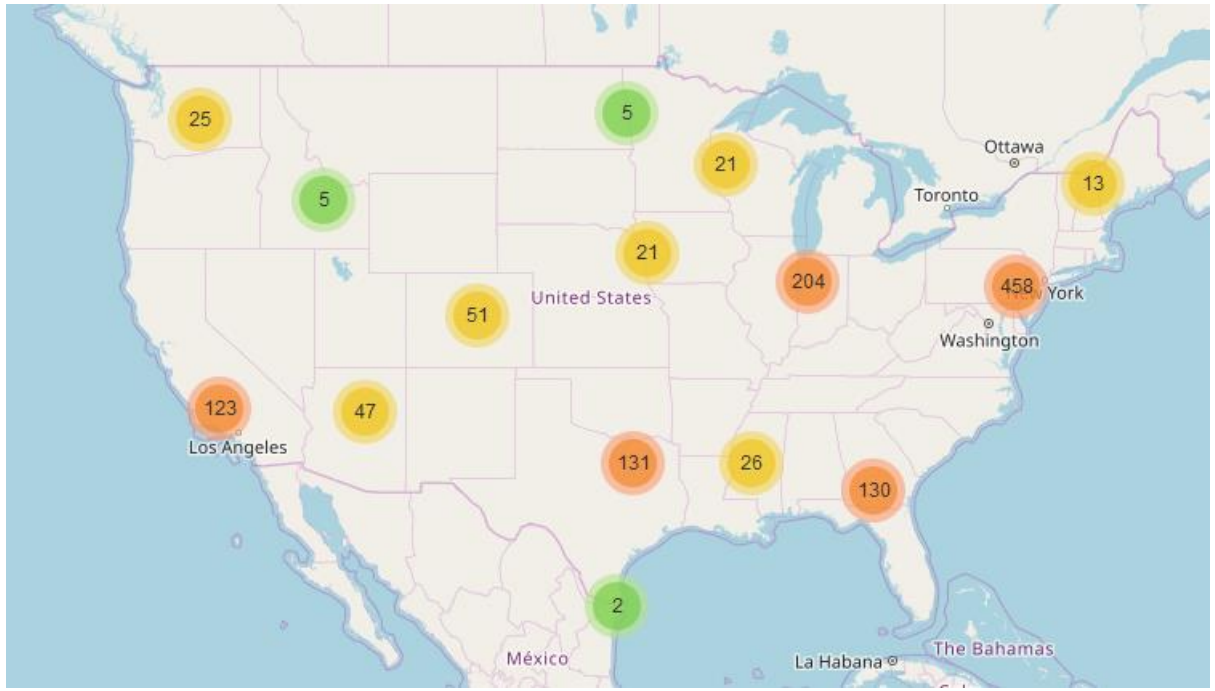
2060 rows × 6 columns

Table 1: Clean data frame after normalizing college data

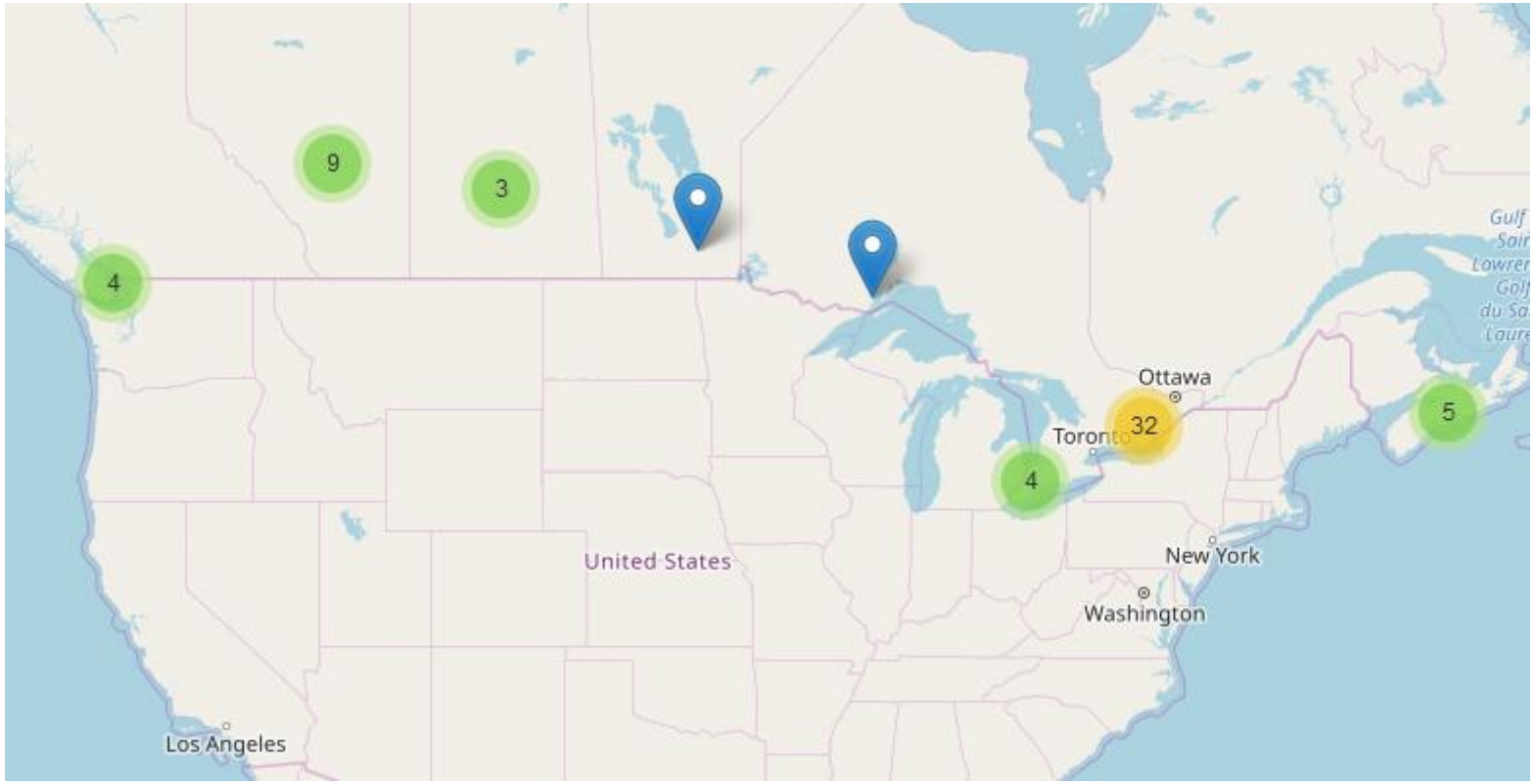
- To give tuition fees a uniform scale, we convert all credit-based fees to yearly fees.
- The formula to convert (approximately) credit based tuition fee to annual tuition fee:
- df.loc[df['tuition_fee.unit'] == 'credit', 'tuition_fee.value'] = (df['tuition_fee.value']*45)/2
- Note: This causes discrepancies 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

EXPLORATORY DATA ANALYSIS
Spatial Data Analysis

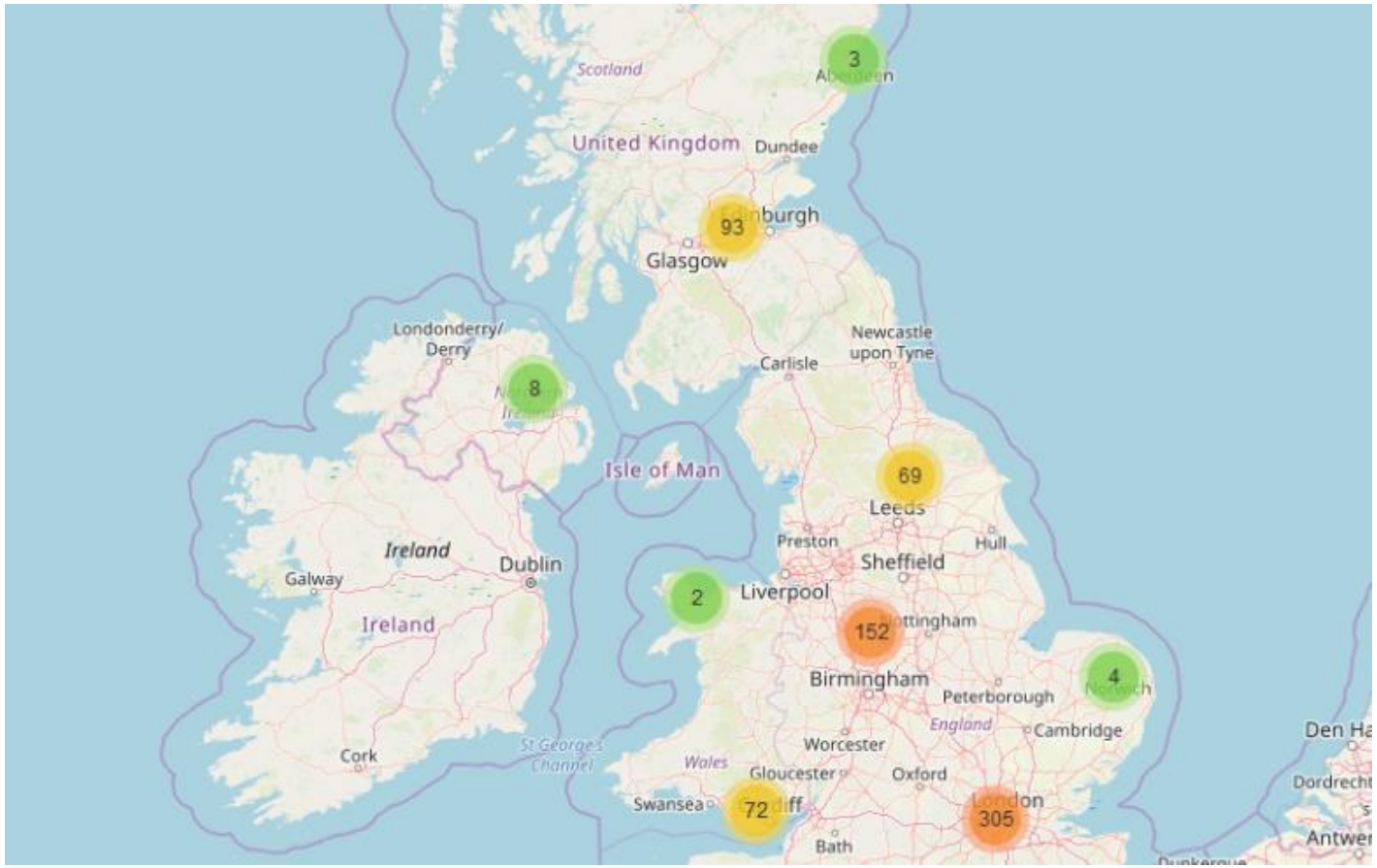
- I formed three more data frames representing colleges in USA, UK and Canada separately.



United States of America


9
3
4
Ottawa
32
5
4
United States
New York
Washington
Los Angeles

Canada

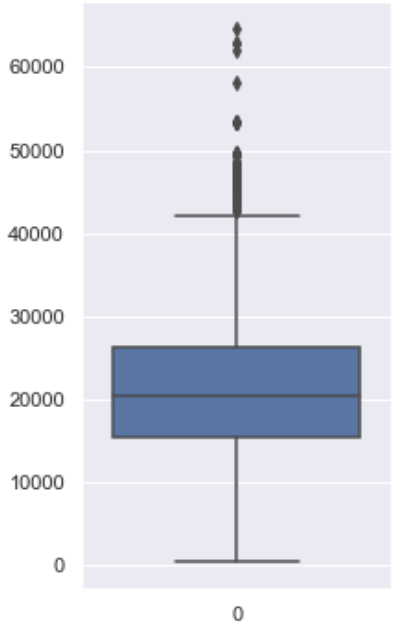


United Kingdoms

- USA has a large number of colleges to the east. Also it has over 1200, the most number of colleges among all the countries.
- Canada has negligible colleges compared to USA and UK. It has only 52 colleges.
- UK has the best distribution of college. It has over 700 colleges.

EXPLORATORY DATA ANALYSIS

Numerical Data Analysis



	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

- 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.

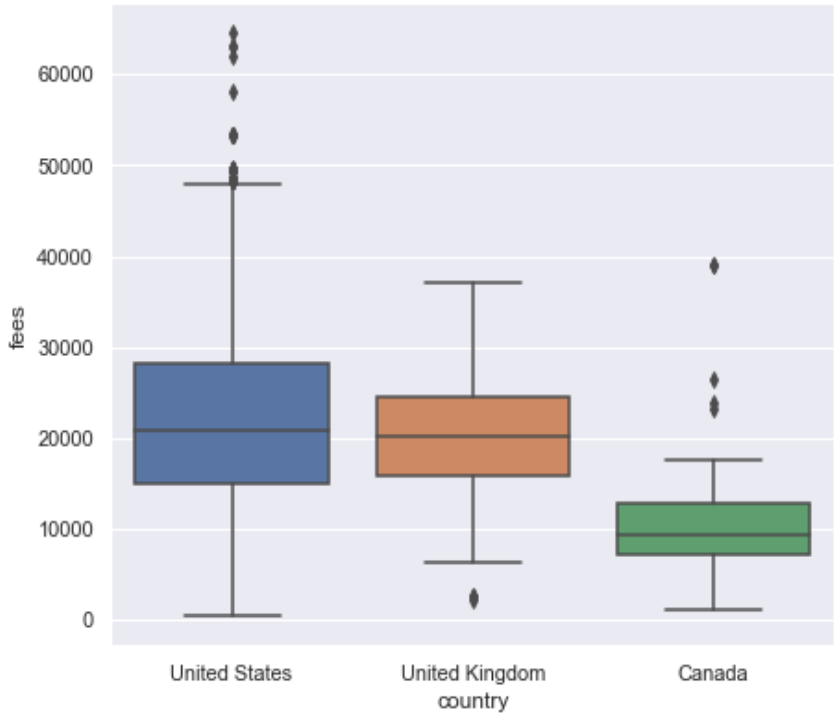


Figure 7: Tuition fee comparison of countries

- Canadian colleges are best 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.

Clustering

- After acquisition of geospatial coordinates and cleaning the dataset, its time to explore each college neighbourhood to cluster them.
- We'll make a data frame where colleges are grouped along with the dummy columns that gives us a data frame of dimensions 629x387.
- I've successfully clustered the college neighbourhood into the following categories:
  - American Eats
  - Exotic Eats
  - Tour/ Outgoing
  - Night Life (Pub) and Fitness
  - Art Prone/ Mature Audience

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.
- We can predict what kind shop should be opened near a college or where should we rent a place to get the most benefit.