STA 220 Project - Data Driven Insights for Informed University Selection: A Holistic Analysis

Mihir Shah *(* mihshah@ucdavis.edu *)*
Piyush Santosh Kulkarni (pkulkarni@ucdavis.edu)
Kriti Kriti (kkriti@ucdavis.edu)

Project Link - https://github.com/Piyush140899/STA220

1. Introduction

In 2021, there were 6.4 million international students globally. According to sources, the total number of international students at U.S. colleges pursuing Bachelor's, Master's, and doctorate degrees was 10,57,588. From personal experience as international students in a foreign land, the most stressful and time-consuming task is college selection and application. International students apply to various universities in different countries by researching multiple resources, websites, consulting numerous individuals, and developing a rudimentary understanding to select dream colleges and countries to apply to. This decision of selecting colleges is very complicated and is driven by multiple factors, including rankings, the country and state of the college, living expenses, job opportunities, tuition fees, and much more. Many international students, including ourselves, paid hundreds of dollars for college consultancy that provided us with this information and helped us with college selection. While going through the process ourselves, we couldn't find a resource that provided a holistic solution, giving us every required piece of information about every deciding factor in one place. Browsing through various websites and then listing each parameter for a university in an Excel sheet seemed to be a very tedious and time-consuming task. Most websites merely showcased the college details and their arbitrary rankings in a tabular format. While these arbitrary rankings are important, several publications have their own rankings, and a lack of a trustworthy unified ranking is evident, hence it cannot be solely trusted to make a decision. Along with information about college, understanding the living expenses is paramount for international/out-of-state students applying for college as it forms the foundation of their financial planning. With this knowledge, students can craft realistic budgets encompassing accommodation, food, transportation, healthcare, and other daily necessities, ensuring they can sustain themselves throughout their studies. Moreover, familiarity with living costs is crucial for meeting visa requirements, as many countries mandate proof of adequate funds to cover these expenses.

The human mind is capable of grasping visualizations or pictorial representations more efficiently and accurately unlike given tabulation of information. We observed a gap where not every website provided all the parameters and information was not pictorially represented which makes it more readable and informative. Visualization enables students to compare different colleges more efficiently and identify key insights about each institution. Hence this project and report aims to provide robust, intuitive, and different forms of visualization like tables, bar plots, slider plots, world maps, choropleth maps, and US country maps.

We embarked on creating a comprehensive analysis to consolidate all relevant data pertaining to colleges and their associated living expenses, catering to ambitious students aspiring for higher education. To accomplish this, our initial step involved scraping data from various relevant websites and conducting thorough data processing and visualization. After extensive research, we sourced college information worldwide from Times Higher Education, scraping data from 2687 colleges across numerous countries. Furthermore, we extracted details regarding essential examinations for international students, such as TOEFL, GRE, and SAT, from QS World Rankings. In addition to college details, a pivotal aspect of our project was to furnish living expenses data. We meticulously gathered living expenses for each city-state pair in the United States from the Numbeo website. Subsequently, we aggregated, manipulated, and integrated data from diverse sources, undertaking multiple pertinent data visualizations to assist students in making one of the most pivotal decisions of their lives: selecting their dream college.

2. Methodology

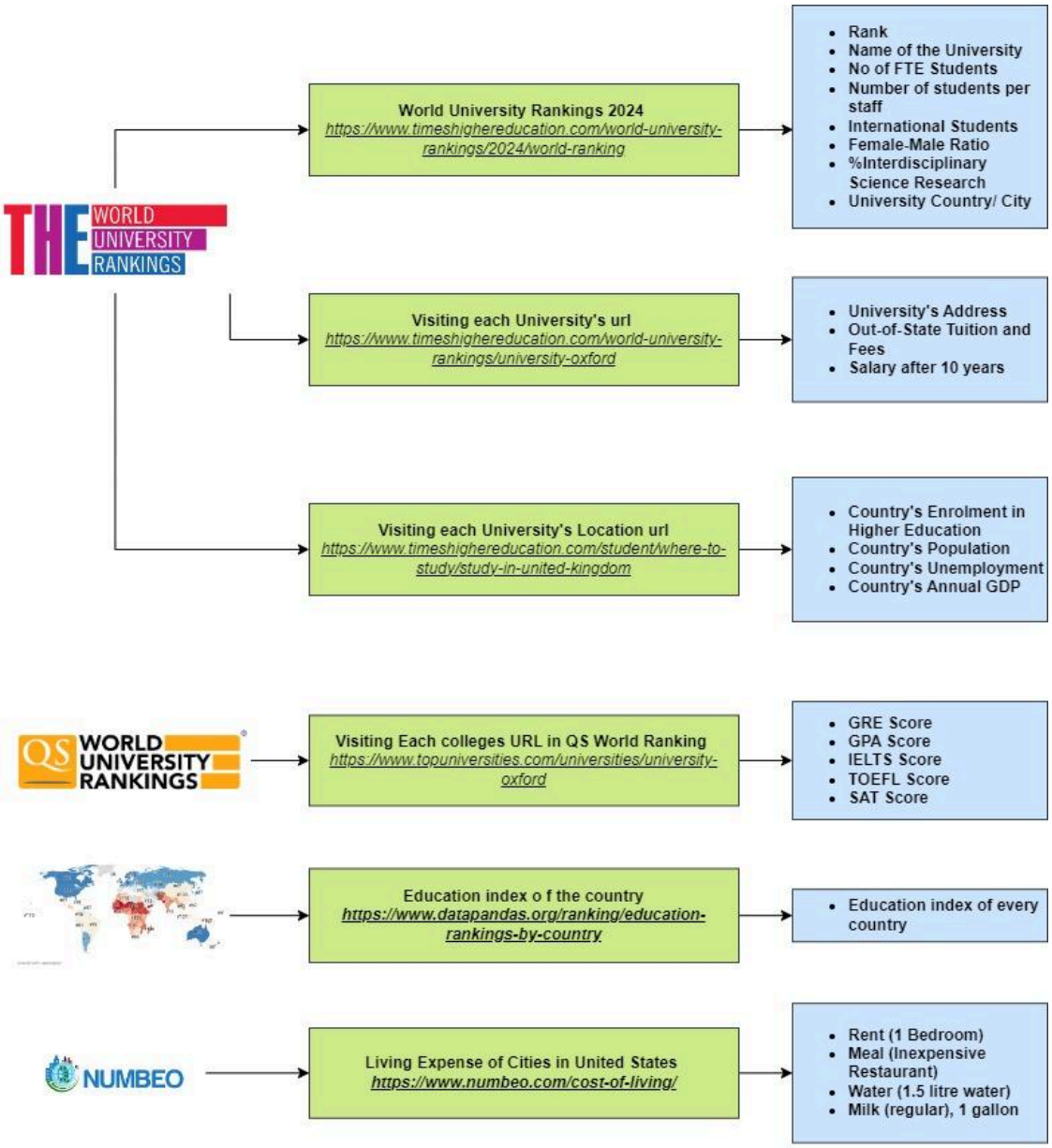
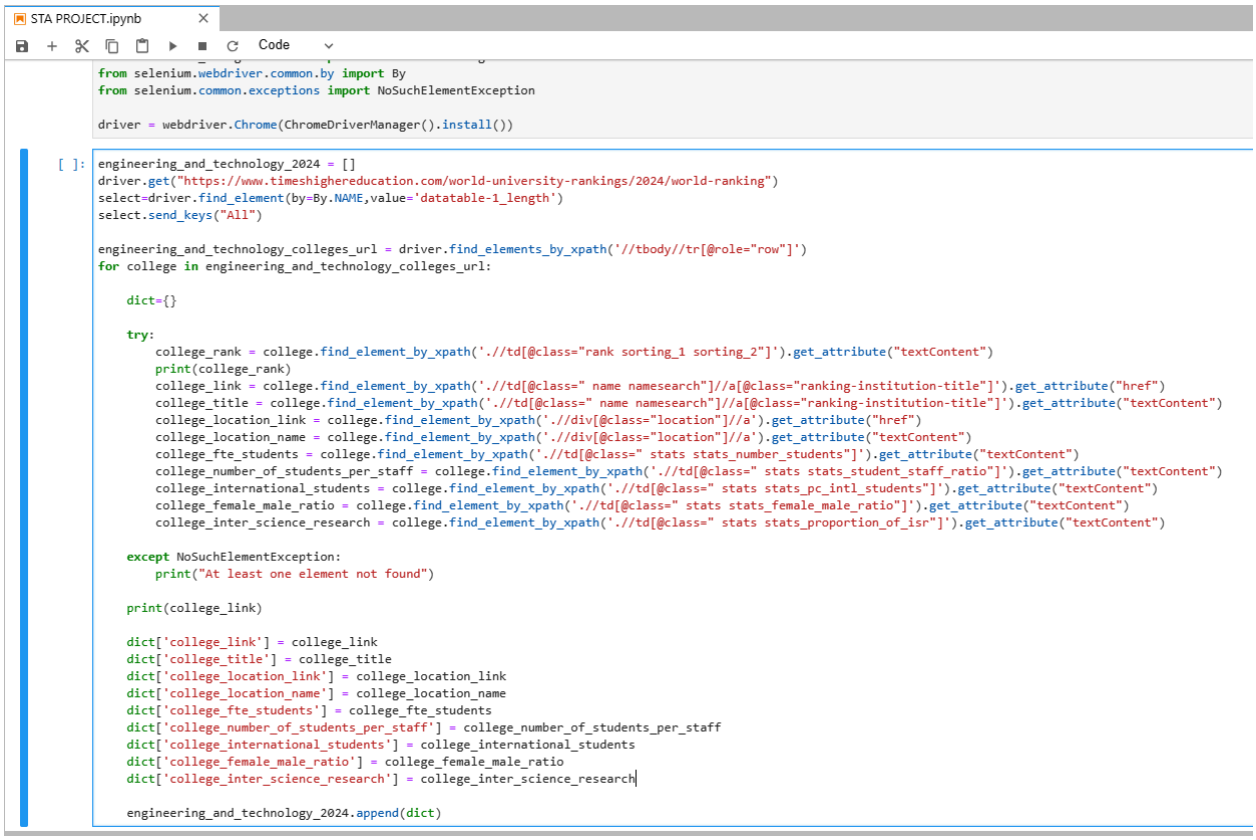


Fig1: Architecture Diagram

1. To begin, we initially researched various websites from which we could gather data about universities. After careful consideration, we settled on the Time Higher Education 2024 World University Rankings (available at https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking). Upon examination, we noted that the website was not static, indicating that the college data was retrieved from backend servers using JavaScript functions, and notably, there were no JSONs present in the response body. Consequently, we opted to use Selenium to scrape the data from this dynamic website. Through our efforts, we identified 2671 colleges and collected their information. From the World University Rankings 2024, we extracted fields that we deemed important for students, namely University name, University rank, University country/city, Number of Full-Time Equivalent (FTE) students, Number of students per staff, Number of International students, Female-to-Male ratio, and the percentage of Interdisciplinary science research. Below is a code snippet:



2. The subsequent step involved visiting the URL of each university that we had obtained from the previous scraping on Times Higher Education and then extracting specific information such as the address, salary after 10 years, and out-of-state tuition fees. We executed this process by extracting each university's link stored in the dictionary as a variable 'college_link'.

```
count = 0
for college in engineering_and_technology_2024[855:]:

    college_address = None
    college_salary_after_10_years = None
    college_out_of_state_tuition_and_fees = None
    driver.get(college.get('college_link'))

    try:
        count+=1
        print(count)
        college_address = driver.find_element_by_xpath('//div[@class="institution-info__contact-detail institution-info__contact-detail--address"]').get_attribute("textContent")
        college_salary_after_10_years_text = driver.find_element_by_xpath('//div[@class="keystats salary_10_years"]')
        college_salary_after_10_years = college_salary_after_10_years_text.find_element_by_xpath('preceding-sibling::div').get_attribute("textContent")
        college_out_of_state_tuition_and_fees_text = driver.find_element_by_xpath('//div[@class="keystats fees_oos"]')
        college_out_of_state_tuition_and_fees = college_out_of_state_tuition_and_fees_text.find_element_by_xpath('preceding-sibling::div').get_attribute("textContent")

    except NoSuchElementException:
        print("At least one element not found")

    college['college_address'] = college_address
    college['college_salary_after_10_years'] = college_salary_after_10_years
    college['college_out_of_state_tuition_and_fees'] = college_out_of_state_tuition_and_fees
```

3. Followed by visiting every University's website, we decided to visit the location link(variable ‘college_location_link’) of every University to gain insights about data of a country in which university is located. We were able to retrieve information such as Country’s Population, Country’s GDP, Country’s unemployment and Country’s enrollment in higher education.

```
            country_enrollment_rate_in_higher_education_text = driver.find_element_by_xpath("//span[contains(text(), 'Enrollment rate in higher education')]")
#             country_enrollment_rate_in_higher_education_text = WebDriverWait(driver, 10).until(EC.visibility_of_element_located((By.XPATH, "//span[contains(text(), 'Enrollment rate
            country_enrollment_rate_in_higher_education = country_enrollment_rate_in_higher_education_text.find_element_by_xpath("following-sibling::span").text

            country_population_text = driver.find_element_by_xpath("//span[contains(text(), 'Population')]")
            country_population = country_population_text.find_element_by_xpath("following-sibling::span").text

            unemployment_text = driver.find_element_by_xpath("//span[contains(text(), 'Unemployment')]")
            unemployment = unemployment_text.find_element_by_xpath("following-sibling::span").text

            annual_gdp_text = driver.find_element_by_xpath("//span[contains(text(), 'Annual GDP')]")
            annual_gdp = annual_gdp_text.find_element_by_xpath("following-sibling::span").text
```

4. Alongside gathering college information, we made the decision to scrape data regarding the scores required in various examinations to gain admission to those specific colleges. We gathered the score requirements for TOEFL, GRE, SAT, and GPA. This was achieved by scraping the QS World Rankings website. Using the university names stored in the college list dictionary, we noticed that to access a university on the QS Rankings website, we only needed to modify the URL. The URL followed a pattern such as: https://www.topuniversities.com/universities/college_name . However, the college names stored in our dictionary had a different format, with spaces and the first letter of each word capitalized. The 'college_name' required to pass in the URL had to be all lowercase, with spaces replaced by dashes. We implemented a string manipulation function to format the university name as required, and then used the 'Requests' library in Python to request the QS World Rankings website for a particular university's page. From there, we extracted GPA, GRE, TOEFL, SAT, and IELTS scores.

```
def modify_college_title(college_title):
    college_title = college_title.lower()
    translator = str.maketrans("", "", string.punctuation)
    # Use the translation table to remove punctuation
    college_title = college_title.translate(translator)
    words_to_remove = ['of','the']
    college_title = remove_words(college_title,words_to_remove)
    college_title = college_title.replace(' ', '-')
    return college_title
```

```
for college in engineering_and_technology_2024:
    college['modified_college_title'] = modify_college_title(college.get('college_title'))
```

```
qs_prefix_url = "https://www.topuniversities.com/universities/"
count = 0
for college in engineering_and_technology_2024:

    driver.get(qs_prefix_url + college.get('modified_college_title'))
    college_gre = None
    college_gpa = None
    college_ielts = None
    college_toefl = None
    college_sat = None

    try:
        count+=1
        print(count)
        college_gre_text = driver.find_element_by_xpath("//label[contains(text(), 'GRE')]")
        college_gre = college_gre_text.find_element_by_xpath("following-sibling::div").text
        college_gpa_text = driver.find_element_by_xpath("//label[contains(text(), 'GPA')]")
        college_gpa = college_gpa_text.find_element_by_xpath("following-sibling::div").text
        college_ielts_text = driver.find_element_by_xpath("//label[contains(text(), 'IELTS')]")
        college_ielts = college_ielts_text.find_element_by_xpath("following-sibling::div").text
        college_toefl_text = driver.find_elements_by_xpath("//label[contains(text(), 'TOEFL')]")
        college_toefl = college_toefl_text[0].find_element_by_xpath("following-sibling::div").text
        college_sat_text = driver.find_element_by_xpath("//label[contains(text(), 'SAT')]")
        college_sat = college_sat_text.find_element_by_xpath("following-sibling::div").text
```

5. As a last metric, we decided to retrieve every country's education index by scraping the website: https://www.datapandas.org/ranking/education-rankings-by-country.

6. For a student along with college curriculum and rankings one of the most important things is the expenses that they are going to incur over the span of 2-4 years. There are various factors influencing the living expenses depending on the city and state the college is situated in. For this, we scrape another website numbeo.com/cost-of-living which provides a detailed cost of living for a particular city and state that includes data like Monthly Rent for a 1 Bedroom, Meal prices for 1 at a restaurant and other grocery costs like Milk, Bread, and Water. One can see how the aggregation of this data along with college information is so valuable and gives a holistic analysis for students pursuing college. It is also important to provide cost for each state and each region. This may sound general knowledge to students residing in that country but according to opendoors data there are ~1 M international students coming to the United States every year. For these students cost of living in many cases will be the most important factor. With no knowledge of the complexities and differences between cost of living of different states, this analysis is very important in making a decision. We first observed that it was a static website that could be scraped to get a list of city-state pairs and change the URL accordingly to get to the page for

the cost of living for that particular city-state. However, upon further investigation, we observed that they had a paid API of 200 USD monthly and had made scraping very very difficult. The city-state url did not follow one single logic; they were different and random which made scraping difficult. After thorough investigation, we found there were 3 logic for websites which were

a. /City-state-country
b. /City-state
c. /city

To make things harder Numbeo had created all three websites on the server so it was not simple if the status was 200 then get the data. But we had to investigate all three websites and identify which website was correct and then get the necessary information. We decided to limit our data dictionary to the following data fields -Rent, Meal, Milk, and Water as we believe these data were important for students to calculate their expenses and also these data were most populated and not null in the websites. The logic we used to traverse each address and traverse 3 urls and get the correct html.

```python
import requests

def get_html(address):
    address = address.strip()
    arr = address.split(',')
    if(len(arr)<2):
        return ''
    arr[0] = arr[0].replace(' ', '-')
    arr[0]=arr[0].strip()
    arr[1]=arr[1].strip()

    prefix_url = "https://www.numbeo.com/cost-of-living/in/"  # Replace this with

    urls_to_try = [
        f"{prefix_url}{arr[0]}-{arr[1]}-United-States",
        f"{prefix_url}{arr[0]}-{arr[1]}",
        f"{prefix_url}{arr[0]}"
    ]

    for url in urls_to_try:
        response = requests.get(url)
        # print(url)
        if response.status_code == 200:
            html_content = lx.fromstring(response.text)
            body=html_content.xpath("//div[@class='innerWidth']/h1/text()")[0]
            # print(body)
            if len(body)==0 or ('Cannot find city id for' not in body):
                print(url)
                return response.text
    return ''  # Or raise an exception if desired
```

```python
import lxml.html as lx

def get_field_from_html1(html_content, field):
    if(len(html_content)==0):
        return '0'
    meal = '0'
    # Parse HTML content using lxml
    tree = lx.fromstring(html_content)

    # Find all <tr> tags
    table_rows = tree.xpath('//tr')

    # Loop through each <tr> tag
    for row in table_rows:
        # Find the first <td> tag in the row
        first_td = row.find('td')
        if first_td is not None:
            # Check if the text of the first <td> tag is equal to the desired field
            if first_td.text_content().strip() == field:
                # Find the next <td> tag which contains the price
                price_td = first_td.getnext()
                if price_td is not None:
                    # Get the text inside the <span> tag
                    price_span = price_td.find('span[@class="first_currency"]')
                    if price_span is not None:
                        # Extract the dollar amount
                        dollar_amount = price_span.text_content().replace(' ', '').replace('$', '').strip()
                        if dollar_amount != '?':
                            meal = dollar_amount
                        break  # Exit loop since we found the desired value
    return meal
```

```python
def get_data(html):
    data={}
    data['Milk']=get_field_from_html1(html,'Milk (regular), (1 gallon)')
    data['Rent']=get_field_from_html1(html,'Apartment (1 bedroom) in City Centre')
    data['Meal']=get_field_from_html1(html,'Meal, Inexpensive Restaurant')
    data['Water']=get_field_from_html1(html,'Water (12 oz small bottle)')
    return data
```

7. We basically get a cost of living dictionary for every city-state pair present in the United states. But the state here is abbreviated and also there is no location that we can use to plot on a world or country map. To make the information we got from the previous website more usable we scrape two other websites a) Nominatim API to get the latitude and longitude for each address(city, state) . Secondly, as the states are abbreviated we use the table present on the website https://www.yourdictionary.com/articles/state-abbreviations to get the state abbreviations of the United States.

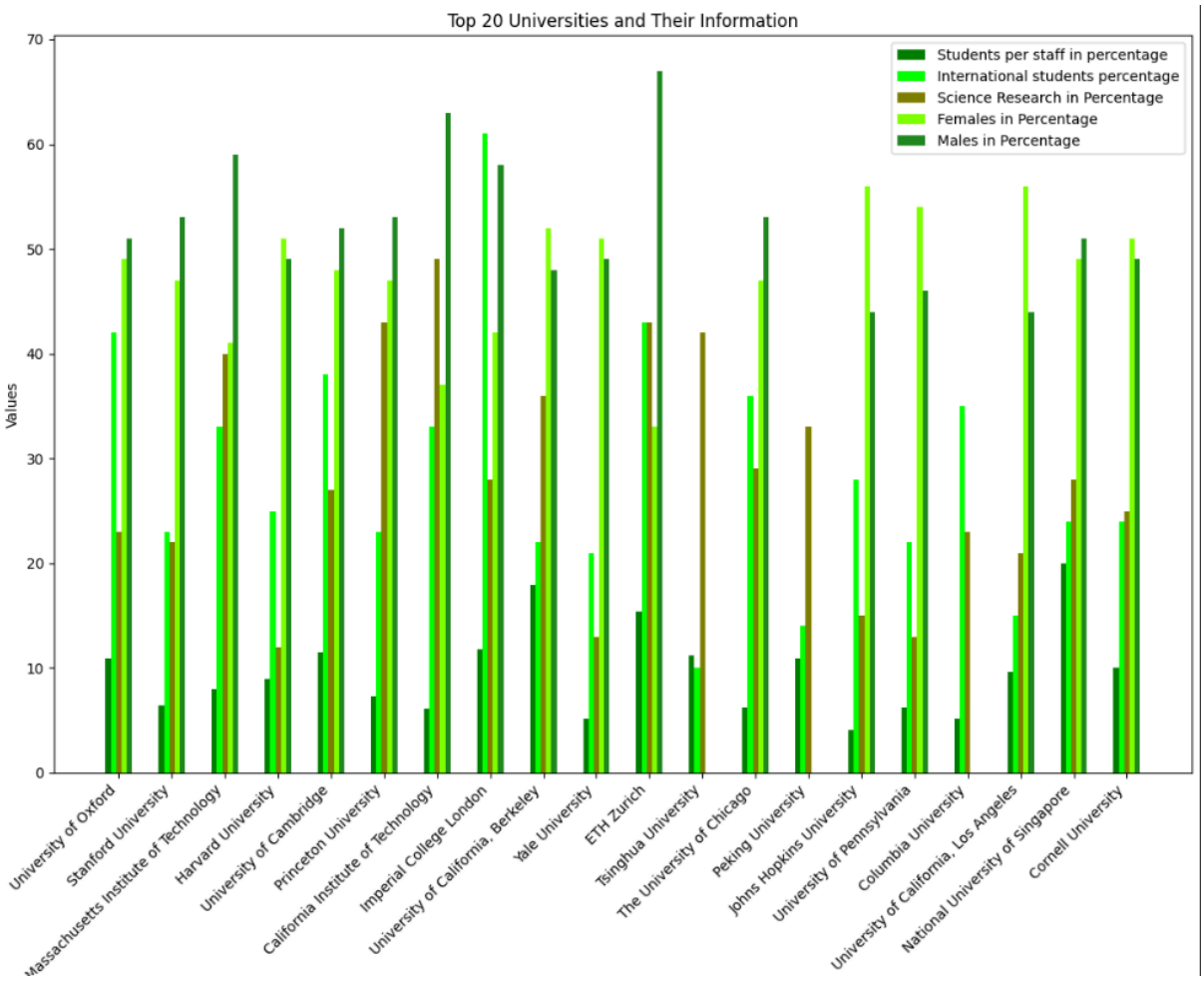
3. Visualizations

1. Displaying Top 20 Universities and comparison based on some of the parameters

a. **Motivation/Relevancy:** As a student, it is important to get into top world colleges. The ambition is always to secure a top 10 top 20 or top 50 college. Along with the ranking, it is also important that students also get their desired factors right such as several students per staff which would help in determining how much attention a student gets from a particular staff, percentage of international students which represents the openness of the university or college towards out of country students, male and female ratio which displays the diversity and many more.

b. **Code snippet/Implementation:** We used matplotlib bar plots with each country consisting of 5 bars representing different factors as well as maintaining the rank of university. The link to the html file for visualization is here.

c. **Visualization screenshot:**



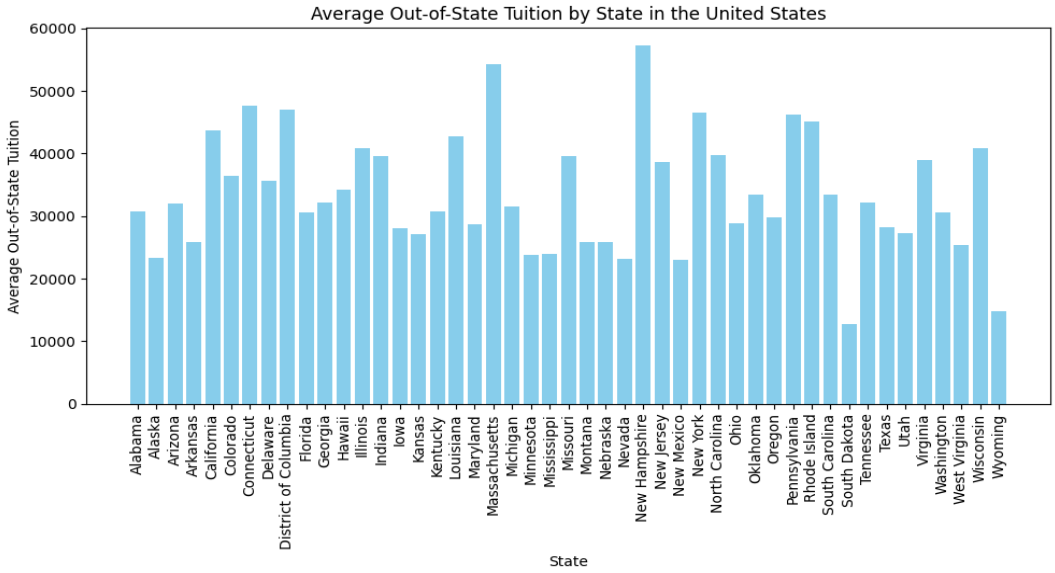
d. Observations: As we can observe in the above plot, although John Hopkins university has a ranking of 15th, it has the lowest students per staff in percentage which means less number of students per staff. If some students want personalized and focused attention from the instructor he would prefer going to John Hopkins instead of top-ranked university. Similarly, based on the individual's choice, all the 5 parameters would play a significant role.

2. Average Out of State Tuition fees for States in USA

a. **Motivation/Relevancy:** For both international and out-of-state students, one significant factor affecting tuition is the out-of-state tuition fees, which vary depending on the state. Since states have different rules and tax systems, their fees also vary accordingly. This overview will offer a more comprehensive analysis. Additionally, the Times website does not provide rankings or categorization based on out-of-state tuition fees and state information.

b. **Code snippet/Implementation:** We implemented a bar plot for the visualization and also displayed the table of states sorted from highest to lowest w.r.t fees. First, we cleaned the data as the fees column had '$' sign and dropped nan values then we aggregated according to states averaged the fees, and displayed bar

c. **Visualization screenshot:**



	state	college_out_of_state_tuition_and_fees
27	New Hampshire	57204.000000
19	Massachusetts	54259.666667
6	Connecticut	47697.000000
8	District of Columbia	47087.400000
30	New York	46593.846154
35	Pennsylvania	46255.285714
36	Rhode Island	45045.000000
4	California	43622.642857
17	Louisiana	42719.500000
45	Wisconsin	40857.500000
12	Illinois	40845.833333
31	North Carolina	39817.800000
23	Missouri	39610.000000
13	Indiana	39531.000000
42	Virginia	38989.833333
28	New Jersey	38611.000000
5	Colorado	36415.500000
7	Delaware	35710.000000
11	Hawaii	34218.000000
37	South Carolina	33488.666667

d. **Observations:** We observe that New Hampshire has the highest out-of-state tuition, it is indeed surprising that not MA NY, or CA has the highest but it is also important to note that this is out-of-state tuition fees and not living expenses or general fees. Also, there are more colleges from the popular states hence the average also might have gone down. Its important to note that this data includes both public and private schools. States like CA, CT, DC, NY MA are very high ranked as well.

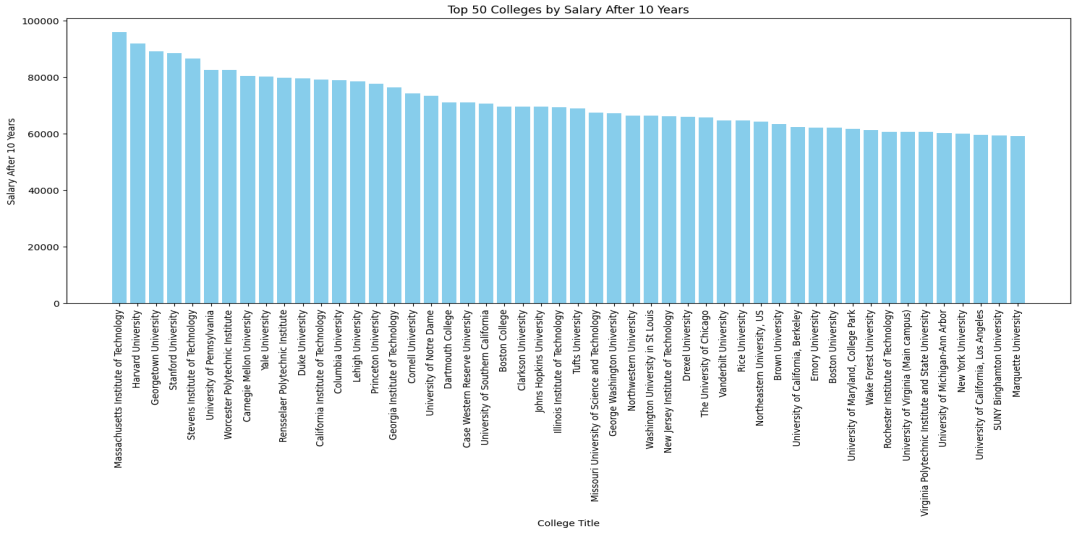
3. Ranking colleges by Highest Salary after 10 years

a. **Motivation/Relevancy:** Visualizing college rankings based on the highest salary after 10 years offers prospective students invaluable insights into the long-term financial outcomes of their educational choices. By highlighting institutions that consistently produce graduates with strong earning potential, this visualization equips individuals with the knowledge needed to make informed decisions about their future

careers and educational investments. It not only empowers students to pursue paths aligned with their professional aspirations but also encourages educational institutions to prioritize programs and initiatives that cultivate real-world skills and enhance graduates' competitiveness in the job market.

b. **Code snippet/Implementation:** We first remove the '$' symbols and other punctuations and then sort the data according to the highest salary and then display the top 50 it in both table and bar plot format

c. **Visualization screenshot:**



d. Observations: We observe that colleges with the highest post-college salaries include MIT, Harvard, Georgetown, and Stanford, as they are prestigious institutions with a long history of academic excellence. These universities attract top students and boast high-quality faculty members. However, it's important to note that the salary range falls below $100k. This could be attributed to the fact that the average salary encompasses various career paths, each with different pay scales. For instance, fields like Computer Science may yield higher salaries compared to other disciplines.

4. Universities displayed according to GRE, GPA and TOEFL score.

a. **Motivation/Relevancy:** As an international student, academic admission depends on various factors but scores play an important role in deciding which universities to apply to and also in classifying universities by level of difficulty to get an admit. Students struggle to find universities filtered on the basis of all three parameters(GRE, GPA, and TOEFL scores). Hence we decided to implement a visualization based on this.

b. **Code snippet/Implementation:** We implemented the following visualization by using sliders where each slider represents the value of one of the exam scores. Change in the value of any one score leads to refreshing output and displaying the relevant output to new scores. We used the **IPython widgets** to implement the sliders.

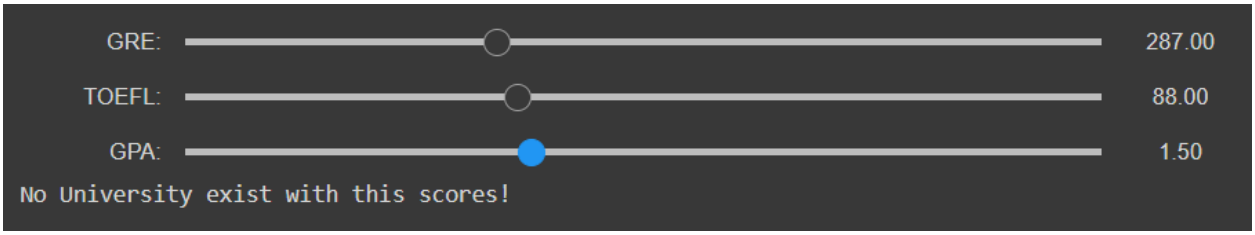
c. **Visualization screenshot:**

GRE: 305.00

TOEFL: 98.00

GPA: 2.80

	college_title	college_location_name	college_number_of_students_per_staff	college_international_students
0	UCL	United Kingdom	11.3	61%
1	Curtin University	Australia	20.8	29%
2	University of Padua	Italy	26.5	7%
3	University of Sussex	United Kingdom	17.3	34%
4	Università Cattolica del Sacro Cuore	Italy	39.1	7%
5	Wayne State University	United States	13	5%
6	Coventry University	United Kingdom	14.9	48%
7	University of Salford	United Kingdom	20.3	23%
8	O.P. Jindal Global University	India	9.1	0%



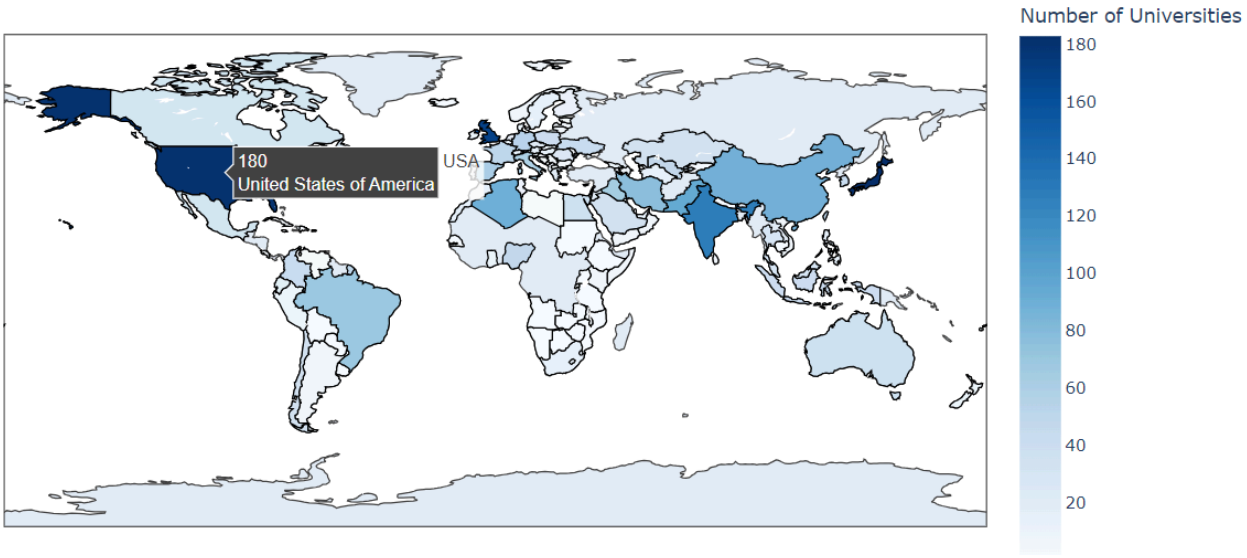
d. **Observations:** As we can see above, we display a message if we are not able to find any universities. If we find universities, we display the university name along with some relevant information about the university in a well-organized tabular form. This would help students to get the university name as well as all other information regarding the university in one place and it can make the process of choosing a university simpler.

5. HotSpot of Universities on a World Map

a. **Motivation/Relevancy:** A lot of times international students are confused regarding the country from which they should pursue higher education. Number of universities in a country not only represents a number but also represents the country's interest in education and financial sponsorship to educational development. Students could shortlist some of the countries using this world map.

b. **Code snippet/Implementation:** We implemented it using the '**plotly'** library. We chose the blue shade for the representation of the number of universities the country has. Lighter shade blue signifies less number of universities in that particular country with respect to a country with darker shade of blue. Number of universities and country names are displayed on hover to a particular area of the world map. The link to the html file for visualization is here.

c. **Visualization screenshot:**



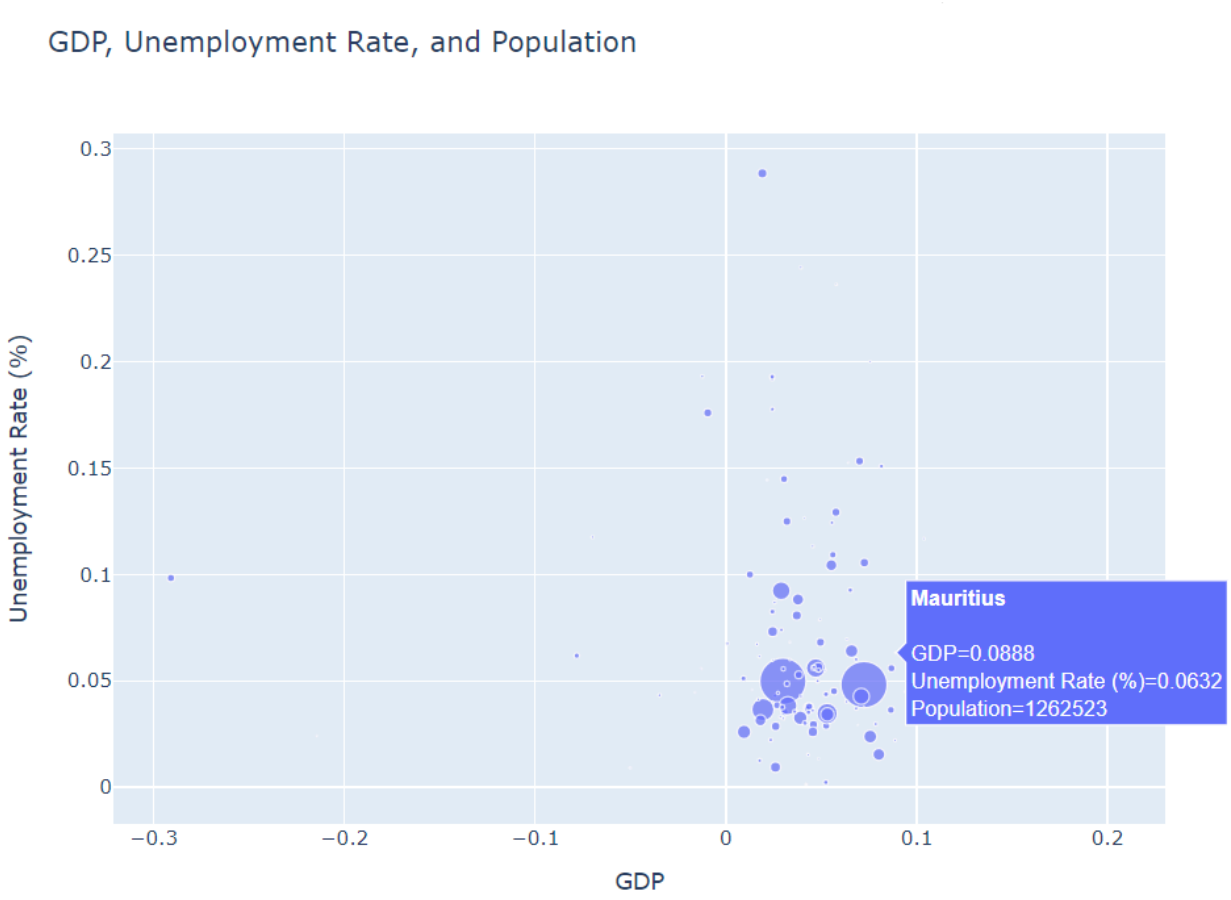
d. **Observations:** As we can see in the above world map, the USA and Japan have the darkest shades of blues denoting the highest number of universities in the world. We can also see the number of universities and country names hovering over the USA.

6. Visualizing GDP, Unemployment, and Population of a country

a. **Motivation/Relevancy:** As academic factors are an important factor for students in choosing the right university, students also look out for the economic conditions of the country. Most students seek job opportunities in countries in which they study after graduation and hence gdp, unemployment and population of a country become of significant importance.

b. **Code snippet/Implementation:** We implemented a bubble scatter plot for this visualization so that we can display three variables in 2D. We represented GDP on the x-axis, unemployment on y-axis and represented the population of a country using the size of the bubble. The

c. **Visualization screenshot:**



d. **Observations:** As we can observe from the plot, we can see details such as gdp, unemployment, and population of the country hovering over a particular bubble. Also, we can observe that the two largest bubbles represent 'India' and 'China' as they are the World's most populated countries. The link to the HTML file for visualization is here.

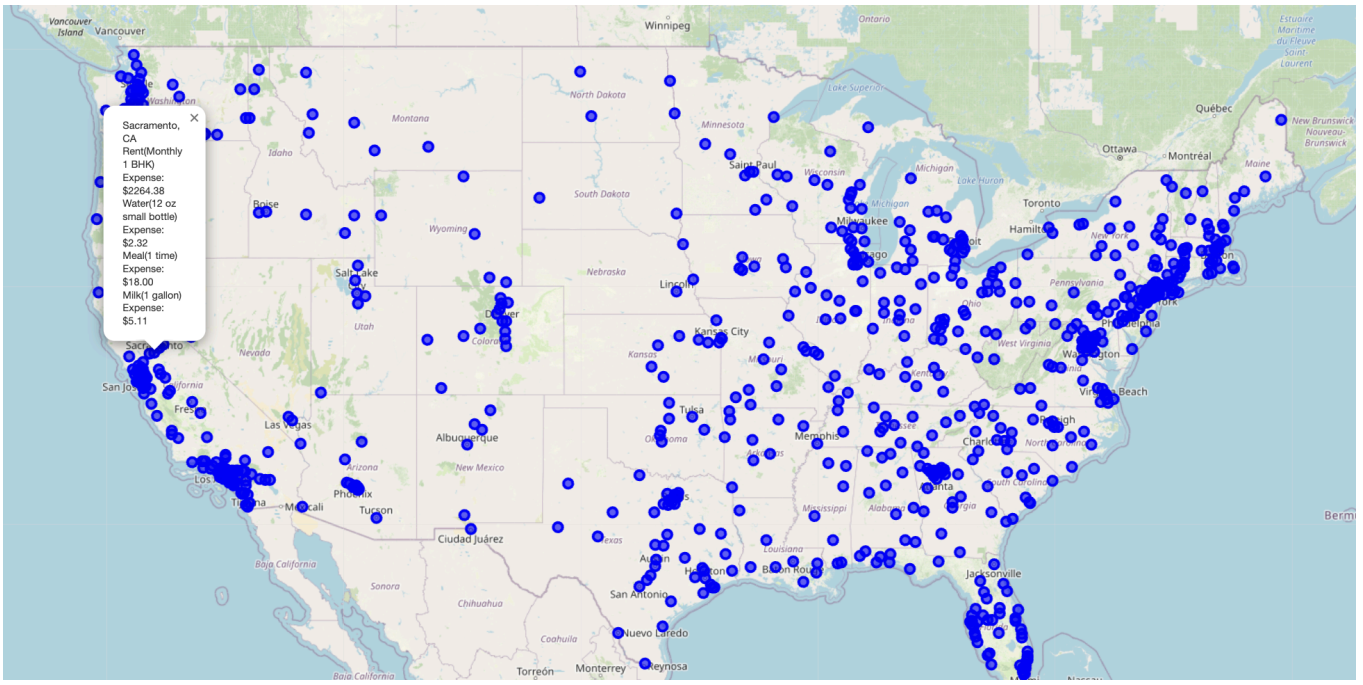
7. US Country map with detailed living expenses for each city and state

a. **Motivation/Relevancy:** Along with college rankings and other information cost of living is also very important, though tuition fees can be accurately estimated college website from For students navigating the complexities of higher education or seeking internships across the United States, a detailed country map showcasing living expenses in each city and state offers invaluable guidance. By visualizing factors like housing costs, transportation expenses, and other essential living expenditures, this tool empowers students to make informed decisions about where to pursue their academic or professional endeavors. Whether assessing the affordability of attending a university or weighing internship opportunities against living costs, this visualization equips students with the financial insights needed to plan effectively and ensure a smoother transition into their educational or career journey.

b. **Code snippet/Implementation:** In order to implement this we use the dataset we get from mining all city-state page cost of living urls from numbeo. We include Meals, Rent, Milk, and Water in our analysis. Now for each address, we get the latitude and longitude using the nominatim API. After getting the

coordinates we plot a static US map using folium and display the address and living expense info when you hover. You can find the link to the html file here .

c. **Visualization screenshot:**



d. **Observations:**This visualization method surpasses conventional tables or searches, offering insight into nearby cities and aiding in estimating comparative living costs. Additionally, it serves as an educational tool for those unfamiliar with US regions, enhancing their understanding. Folium's intuitive interface not only presents cities but also overlays living expenses, providing a comprehensive view for informed decision-making.

8. Ranking Colleges based on their total living expense

a. **Motivation/Relevancy:** While the previous visualization provided the cost of living for each city-state we are interested in the college expenses for each college present in the US. By considering not just tuition fees but also living expenses, students can better understand the total cost of attendance and plan their budgets accordingly. For many students, especially those from low-income backgrounds or reliant on financial aid, the total living expenses can significantly impact their ability to pursue higher education. By providing transparent and comparative data on living expenses, this ranking empowers students to identify institutions that align with their financial circumstances and aspirations.

b. **Code snippet/Implementation:** We have primarily two dataframes till now one is of all colleges and its details and the second is the living expenses according to the city-state pair. First, we observe that the address that is scraped from times ranking websites cannot be used to make a join between two tables. We create a new function to modify the address to a city-state pair. We observe that the address is of format street, city, state,pincode, country (Ex - 450 Jane Stanford Way, Stanford, California, 94305, United States). Hence we filter the required information city, and state that is needed for joining. After that we observe that the state is abbreviated in the living expense table, hence we use the table of state abbreviations we scraped from https://www.yourdictionary.com/articles/state-abbreviations. After this, we join the two tables based on this city-state column(For ex - Davis, CA). After joining we created a new column called total_expenses per month which is basically(15*df'[Meal'] + df['Rent'] + 30*df['Water'] + df['Milk'], Note we researched what is the average consumption of water, milk and number of meals eaten

outside and concluded on this formula)We then sort and rank the colleges according to this column total_expenses and list everything in table.The code for modifying the address of the first table -

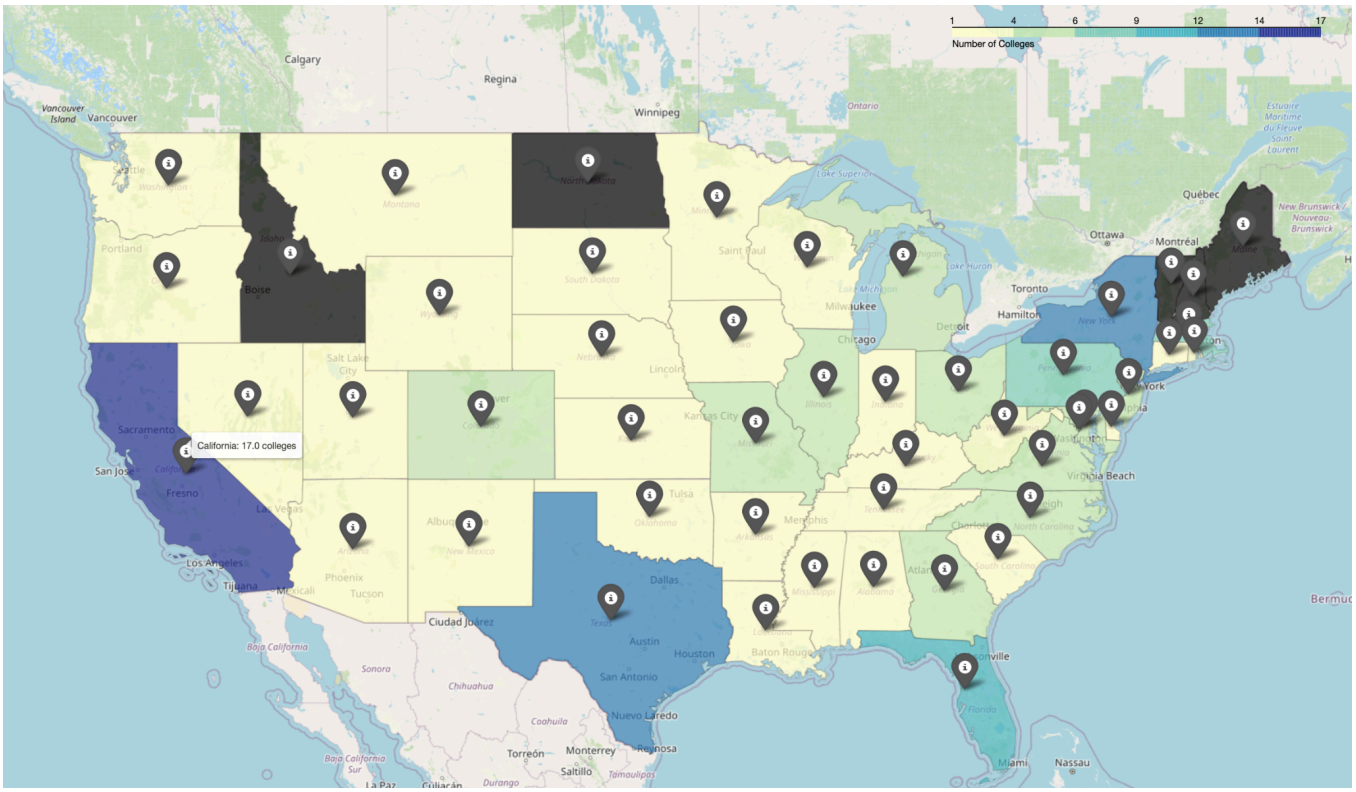
```
def get_modified_address(college_address):
    college_address=str(college_address)
    college_address=college_address.replace('\n','').strip()
    arr=college_address.split(',')
    country=''
    mod_addr=''
    for i,_ in enumerate(arr):
        arr[i]=arr[i].strip()
    if(len(arr)>=4):
        country=arr[-1]
        mod_addr=arr[-4]+', '+state_abbr_dict.get(arr[-3],arr[-3])


    return mod_addr,country
```

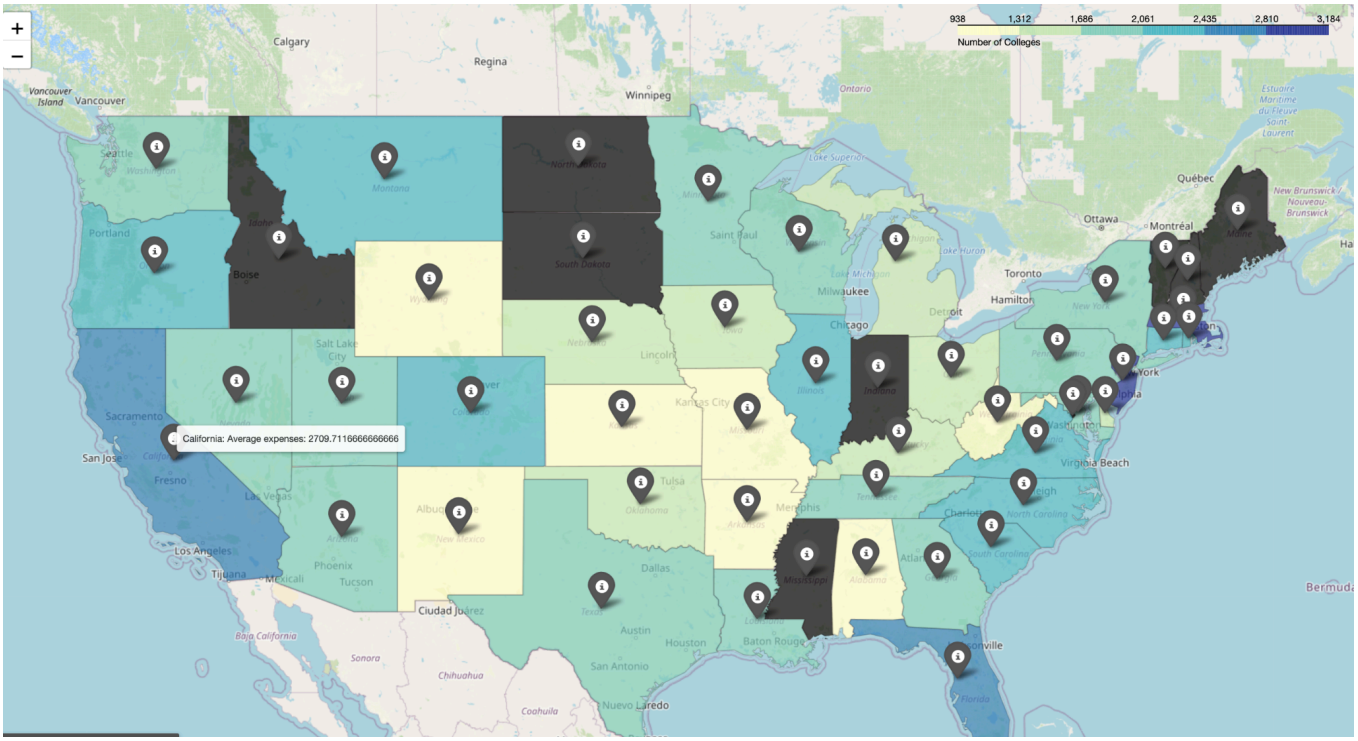
c. **Visualization screenshot:**

college_title	total_expense
University of California, Santa Cruz	3273.75
Massachusetts Institute of Technology	3182.35
Harvard University	3182.35
Boston University	3097.24
University of Massachusetts	3097.24
Northeastern University, US	3097.24
San Diego State University	2972.02
University of Miami	3017.50
University of Miami	3017.50
Florida International University	2878.84
Stevens Institute of Technology	2957.31
William & Mary	2780.36
Princeton University	2790.63
California Institute of Technology	2773.48
Florida Atlantic University	2650.73
University of California, Los Angeles	2501.14
University of Southern California	2501.14
International American University	2501.14
University of California, Irvine	2539.92
Westcliff University	2539.92
University of California, Berkeley	2515.00
College of Charleston	2330.71
College of Charleston	2330.71
The University of Chicago	2385.98
University of Illinois Chicago	2385.98
Illinois Institute of Technology	2385.98

d. **Observations:** As anticipated, we note the presence of colleges in metropolitan cities, where job opportunities are abundant and living standards command higher prices. UC Santa Cruz, MIT, Harvard, and Boston stand out as top-tier institutions with elevated living costs, reaching up to $3,273 USD per month. This correlation underscores the importance of location in determining overall expenses and reflects the premium associated with renowned educational hubs.

9. **ChoroPleth Color Map showing Number of Colleges and Average Total Living Expense for each state in US**
    a. **Motivation/Relevancy:** Visualizing a color map illustrating the number of colleges and average total living expenses per state in the US provides a comprehensive snapshot of educational accessibility and affordability nationwide. By encoding information into color gradients, the visualization efficiently communicates spatial patterns, allowing viewers to discern regions with higher concentrations of educational institutions and varying living costs at a glance. This effectiveness lies in its ability to condense complex data into easily interpretable visual cues, empowering stakeholders—from students and policymakers to researchers and advocacy groups—to identify disparities, make informed decisions, and advocate for equitable access to education and resources across different states. By observing the density of colleges and the average living expenses depicted by colors, students can make informed decisions about where to pursue their higher education based on affordability, available options, and quality of life considerations.
    b. **Code snippet/Implementation:** To implement these two maps we use folium's choropleth map. We first aggregate the rows according to state and count the number of colleges and plot it. We then display the average living expenses according to the states. Folium is primarily a display world map but states cannot be plotted using this. To draw state boundaries on the map we use geopandas and a shape file that is downloaded from the US Census website (https://www.census.gov/geographies/mapping-files/time-series/ geo/carto-boundary-file.html) . The html files can be viewed on github here map2 & map3
    c. **Visualization screenshot:**





d. **Observations:** Firstly, it's important to note that only those US colleges(181) that are in the top 2671 in the world rankings are present on this US country map. We can see how this choropleth map is visually helpful in determining which states have more colleges. In the first map, we observe that California has the highest number (17) of colleges in the rankings, followed by Texas, Florida, and Massachusetts. Secondly, we can see a similar trend in the second map where coastal states have higher living expenses compared to central states. New York, New Jersey, Massachusetts, Florida, and California have higher living expenses for top-ranked colleges. The living expenses range from 938 to 3184 USD. We can also observe that some states like North Dakota and Idaho have no colleges in the top rankings. Additionally, we notice that the living expense calculator isn't able to find expense information for states like Mississippi and South Dakota. Further investigation reveals that this is because the address is not correctly formatted on the Times ranking hmm website for the join to work.

4. Technical Challenges

1. Dynamic vs Static Scrapping - The foremost challenge we faced was scrapping the dynamic websites. We had to figure out and learn the tool to scrape the dynamic websites. We found Selenium to be the most relevant tool for the following task.
2. Computationally Expensive - Secondly, scraping data with Selenium was time expensive as compared to static web scraping tools. Visiting each college link took **2 hours** of computational time.
3. Scrapping multiple unique-styled websites - We wanted to make all the information available related to choosing a university. We found a challenge to find all the important factors and values on a single website. Hence in order to give the user all the significant parameters we had to **scrape data from 4 websites and then aggregate the data.** We had to try different scraping methods for each website and finalize different techniques for each website which we also mentioned in the Methodology section.
4. Address Modification in order to Join - We observed that the address present in the Times ranking website was not of the format that we could join it with the Numbeo website. First, we had to parse and extract city and state, we also observed that some colleges did not follow this logic. Secondly, we had to scrape another website for state abbreviations.
5. Difficulty in scraping Numbeo because of irregular URL formats and Blocked Access - While scraping for living expenses from Numbeo, we meticulously extracted the city and state names from the main page, employing them to construct URLs leading to their respective living expense pages. However, navigating Numbeo's infrastructure proved arduous due to its implementation of numerous and erratic URL logics. Furthermore, rather than maintaining a single webpage per city-state pair, Numbeo hosted multiple pages, each adhering to different formats, with only one harboring accurate data. Therefore, we devised a sophisticated logic to iteratively probe all possible URL structures, meticulously parsing through the content of each webpage to discern the veritable source. To contend with frequent timeouts and access blocks, we judiciously incorporated sleep intervals into our scraping routine. This exhaustive process of data acquisition consumed a total of 1.5 hours.
6. Plotting US states on Folium map - Unlike the previous map where we used static maps where we were plotting a World map without state boundaries, for the last two maps we used choropleth maps for US states. For this, we used the shape files of US states from the US Census website. We first plotted the folium map and then drew the states on top of it using the geopandas and the shape file we downloaded from Census.

5. Conclusion

We successfully scraped a total of 6 websites and collected information about 2671 universities. We scraped the Times Higher Education website to get to know the rankings of the universities as well as some of the common diversity parameters. QS World Rankings website helped to get the data about the academic scores requirement for the universities. Numbeo gave an insight into the living expenses of the various cities in the United States. We not only gathered data but also implemented informative and insightful visualizations that would help the students to make better decisions.

Through visualizations, we observed that the top-ranked university in the World was Oxford University followed by Stanford University in the United States. New Hampshire and Massachusetts were the most expensive college states for an international or out-of-state student whereas South Dakota(12807 USD) was the cheapest state. Even though Massachusetts Institute of Technology was the costliest college, to no surprise the graduates of the college had the highest salary after 10 years as compared to other universities. Japan (183) had the highest number of universities in the 2671 top-ranked universities, followed by the United States of America(180) and India(128) respectively. India and China had the highest populations. According to the data obtained for living expenses, the University of California Santa Cruz(USD 3273.75/month) was the most expensive university followed by MIT(USD 3182.35/month) for monthly living costs. California(17) had the highest number of colleges in the world's top rankings. New Jersey(and Massachusetts colleges have the highest average living expenses

Hence our project organized the scattered information of universities on various websites and by leveraging the technology of web scraping, data processing, and deriving significant insights from the data we simplified the process for students and empowered them to make informed decisions about the selection of the university by analyzing various factors.

6. References

1. https://www.timeshighereducation.com/world-university-rankings
2. https://www.topuniversities.com/university-rankings
3. https://www.datapandas.org/ranking/education-rankings-by-country
4. https://www.numbeo.com/cost-of-living/
5. https://opendoorsdata.org/annual-release/international-students/
6. https://www.migrationdataportal.org/themes/international-students
7. https://www.yourdictionary.com/articles/state-abbreviations
8. https://nominatim.org/
9. https://medium.com/@alex_44314/use-python-geopandas-to-make-a-us-map-with-alaska-and-hawaii-39a9f5c222c6
10. https://www.census.gov/geographies/mapping-files/time-series/ geo/carto-boundary-file.html