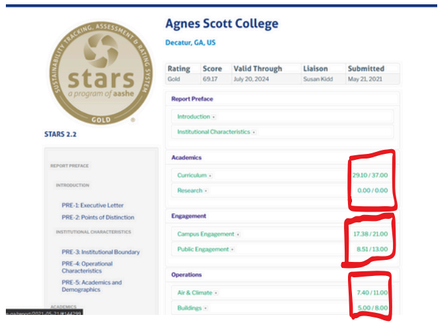
Eco-Scripts

Python-driven Insights into AASHE's Sustainability Data

(An end-to-end ETL project where data was extracted from various web pages using BeautifulSoup, transformed, and loaded for data visualization through Python scripts.)

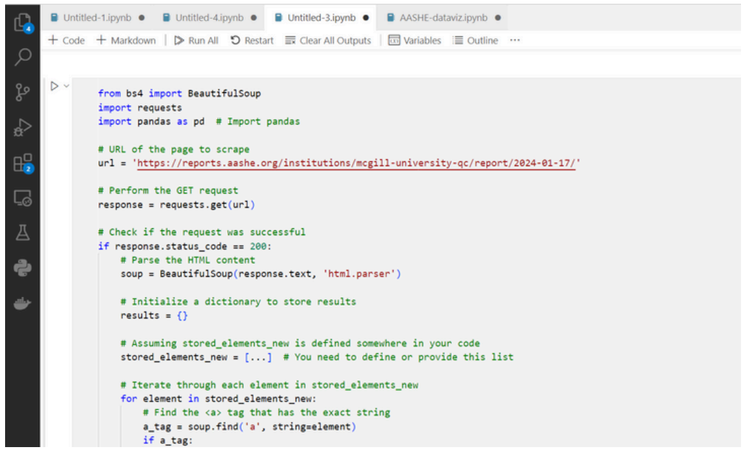
Data pipeline

Data spread across 350 Web pages



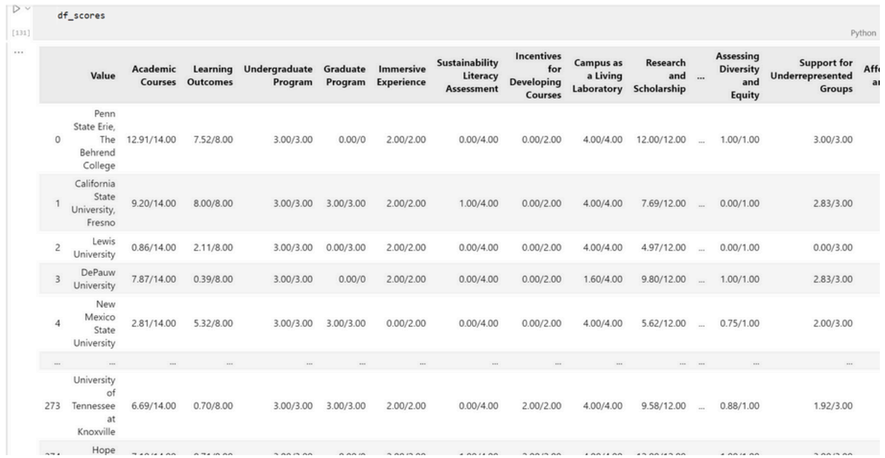
1

Python scripts to automate the web scraping



2

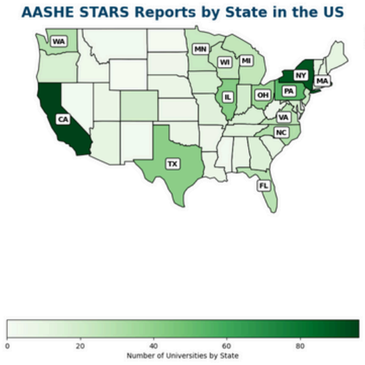
Output stored in a data frame

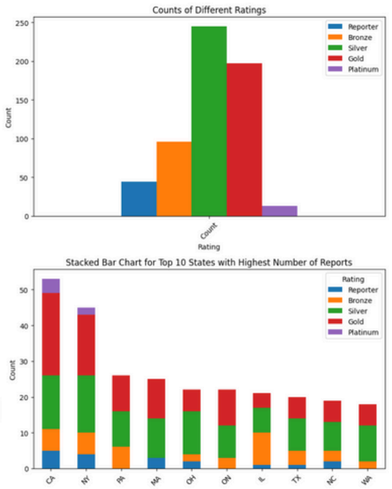


	Value	Academic Courses	Learning Outcomes	Undergraduate Program	Graduate Program	Immersive Experience	Sustainability Literacy Assessment	Incentives for Developing Courses	Campus as a Living Laboratory	Research and Scholarship	...	Assessing Diversity and Equity	Support for Underrepresented Groups	Aff... a
0	Penn State Erie, The Behrend College	12.91/14.00	7.52/8.00	3.00/3.00	0.00/0	2.00/2.00	0.00/4.00	0.00/2.00	4.00/4.00	12.00/12.00	...	1.00/1.00	3.00/3.00	
1	California State University, Fresno	9.20/14.00	8.00/8.00	3.00/3.00	3.00/3.00	2.00/2.00	1.00/4.00	0.00/2.00	4.00/4.00	7.69/12.00	...	0.00/1.00	2.83/3.00	
2	Lewis University	0.86/14.00	2.11/8.00	3.00/3.00	0.00/3.00	2.00/2.00	0.00/4.00	0.00/2.00	4.00/4.00	4.97/12.00	...	0.00/1.00	0.00/3.00	
3	DePauw University	7.87/14.00	0.39/8.00	3.00/3.00	0.00/0	2.00/2.00	0.00/4.00	0.00/2.00	1.60/4.00	9.80/12.00	...	1.00/1.00	2.83/3.00	
4	New Mexico State University	2.81/14.00	5.32/8.00	3.00/3.00	3.00/3.00	0.00/2.00	0.00/4.00	0.00/2.00	4.00/4.00	5.62/12.00	...	0.75/1.00	2.00/3.00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
273	University of Tennessee at Knoxville	6.69/14.00	0.70/8.00	3.00/3.00	3.00/3.00	2.00/2.00	0.00/4.00	2.00/2.00	4.00/4.00	9.58/12.00	...	0.88/1.00	1.92/3.00	
	Hope													

3

Visualizing the data





Data source: University report links

STARS Participants & Reports

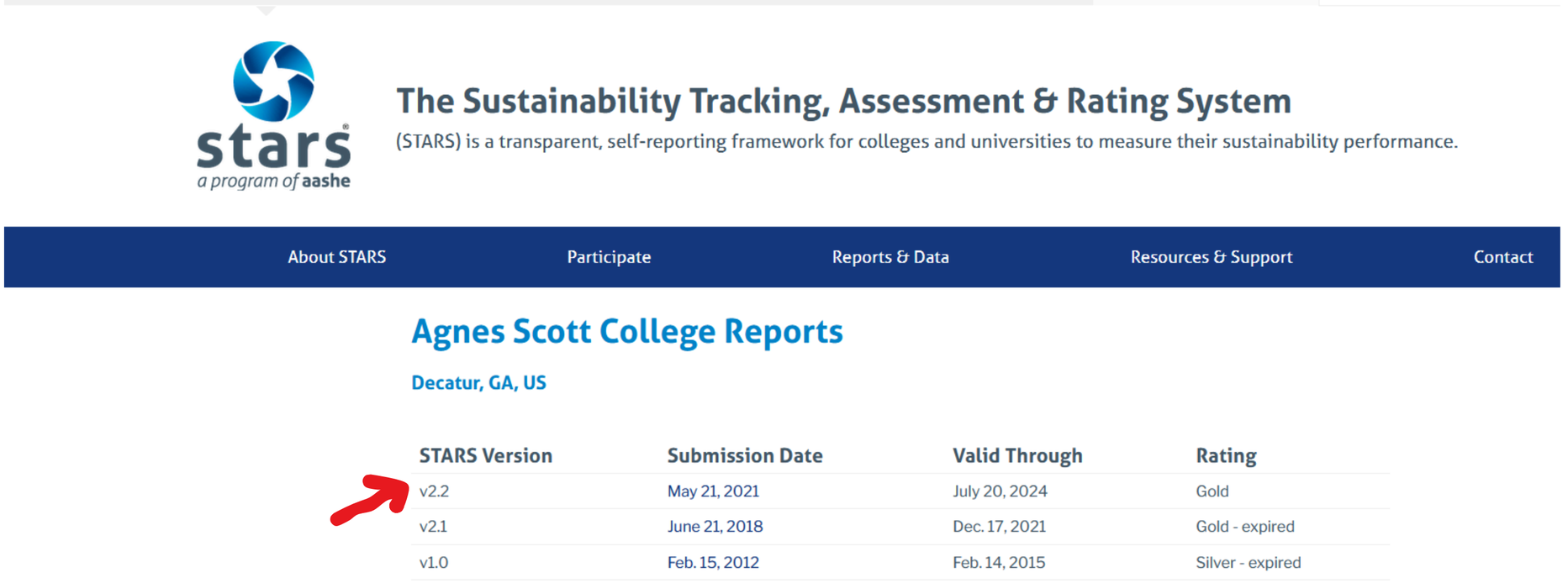
Current Ratings | Report Archive | All Registrants

There are currently 340 institutions with a valid STARS rating. Click on an institution's rating to access its current report.

Institution	Location	Rating	Valid Through
Agnes Scott College	United States , GA	Gold	July 20, 2024
American University of Sharjah	United Arab Emirates , Ash Shariqah	Silver	June 21, 2024
Appalachian State University	United States , NC	Gold	Feb. 28, 2027
Arizona State University	United States , AZ	Platinum	April 18, 2026
Auburn University	United States , AL	Silver	March 31, 2025
Austin College	United States , TX	Bronze	April 28, 2025
Austin Peay State University	United States , TN	Bronze	April 5, 2026
Baldwin Wallace University	United States , OH	Silver	July 13, 2025
Ball State University	United States , IN	Gold	Sept. 10, 2026
Barnard College	United States , NY	Gold	Feb. 1, 2027
Bates College	United States , ME	Gold	Jan. 17, 2027
Bennington College	United States , VT	Gold	May 4, 2025
Bentley University	United States , MA	Gold	Nov. 15, 2026
Binghamton University	United States , NY	Gold	April 2, 2026
Black Hills State University	United States , SD	Silver	July 9, 2026
Boise State University	United States , ID	Silver	Sept. 22, 2024
Boston College	United States , MA	Gold	April 7, 2025
Boston University	United States , MA	Gold	July 18, 2026
British Columbia Institute of Technology	Canada , BC	Silver	June 28, 2024
Bryant University	United States , RI	Silver	June 12, 2025
Bryn Mawr College	United States , PA	Silver	Jan. 7, 2027
Bucknell University	United States , PA	Gold	May 17, 2026

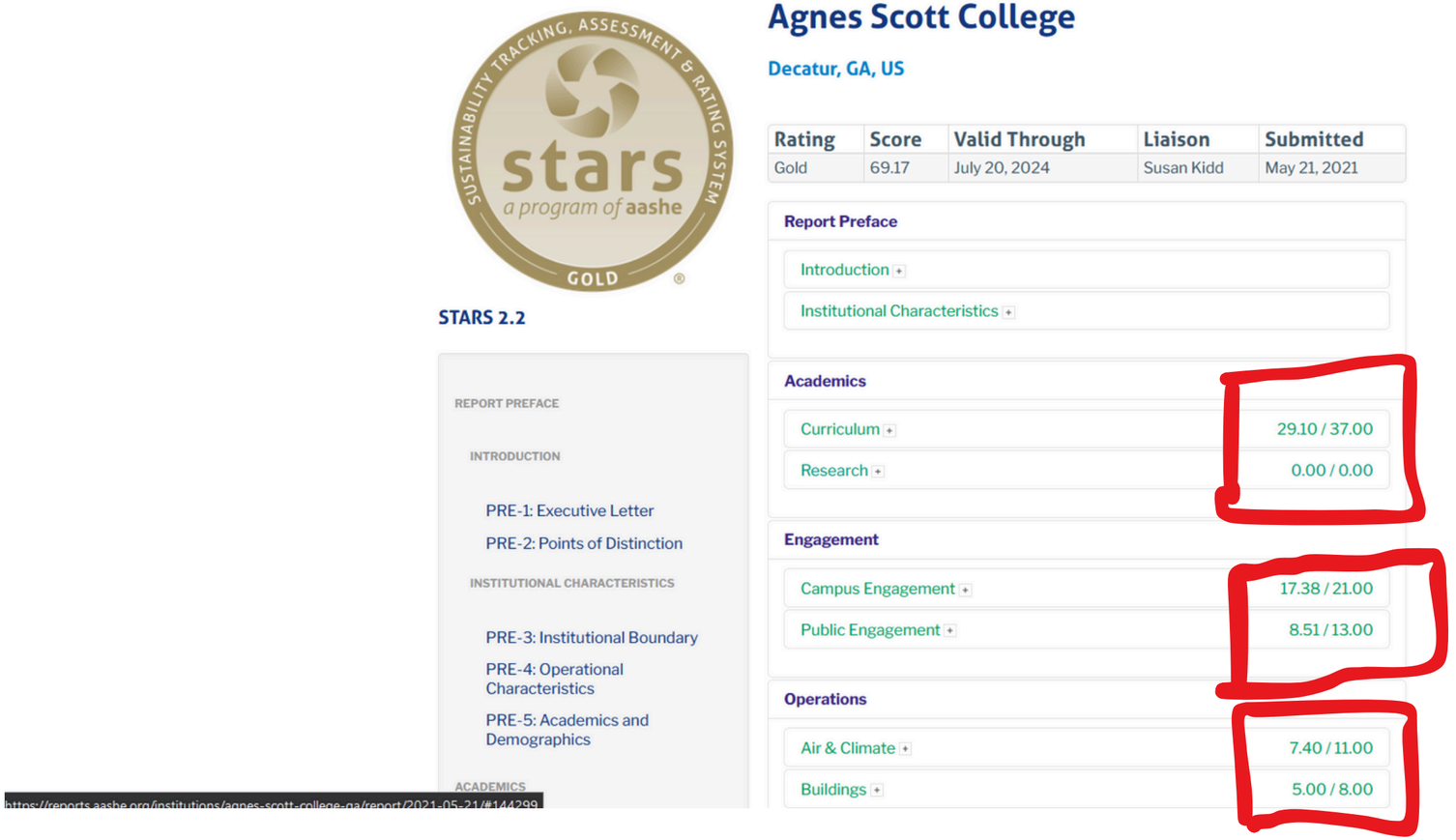
(The sustainability data for each university is accessed through the (AASHE) STARS Participants & Reports portal, where each institution's current rating and report details are listed.)

Data source:All reports submitted by a University



(For the purpose of this project, I only looked at the current report for each University.So my script just accessed the first report for each University)

Actual data of interest



(For each individual University, their data was stored on its respective report page.)

Data extracted using beautiful soup

```python
from bs4 import BeautifulSoup
import requests

# URL of the page to scrape
url = 'https://reports.aashe.org/institutions/mcgill-university-qc/report/2024-01-17/'

# Perform the GET request
response = requests.get(url)

# Check if the request was successful
if response.status_code == 200:
    # Parse the HTML content
    soup = BeautifulSoup(response.text, 'html.parser')

    # Find the <a> tag that contains the text 'Academic Courses'
    a_tag = soup.find('a', string='Academic Courses')

    # Check if the <a> tag was found
    if a_tag:
        # Navigate to the parent <td> tag of the <a> tag
        parent_td_tag = a_tag.find_parent('td')

        # Find the next <td> tag, which should be the first one following the parent
        first_td_tag = parent_td_tag.find_next_sibling('td')

        # Now find the second <td> tag following the first one
        second_td_tag = first_td_tag.find_next_sibling('td')

        # Extract and print the text from the second <td> tag
        if second_td_tag:
            value = second_td_tag.get_text(strip=True)
            print(f'The value in the second <td> tag is: {value}')
        else:
            print('Second <td> tag not found')
    else:
        print('No <a> tag with the text "Academic Courses" was found.')
else:
    print("Failed to retrieve the webpage")
```

```
    <title>
  Penn State Erie, The Behrend College | Scorecard | Institutions | STARS Reports
</title>



    <meta name="viewport" content="width=device-width, initial-scale=1">
    <link rel="profile" href="http://gmpg.org/xfn/11">



    <script type="text/javascript">
      document.documentElement.className = 'js';
    </script>

    <link href="https://fonts.googleapis.com/css?family=Libre+Franklin:400,400i,700,700i" rel="stylesheet">



    <style type="text/css">
     img.wp-smiley,
     img.emoji {
          display: inline !important;
          border: none !important;
            box-shadow: none !important;
            height: 1em !important;
            width: 1em !important;
            margin: 0 .07em !important;
            vertical-align: -0.1em !important;
            background: none !important;
            padding: 0 !important;
     }
      a.disabled {
        pointer-events: none;
          cursor: default;
      }
    </style>
    <link rel='stylesheet' id='aashe-theme-style-css'  href='https://stars.aashe.org/wp-content/themes/aashe
    <link rel='stylesheet' id='js_composer_front-css'  href='https://stars.aashe.org/wp-content/uploads/js_c
```

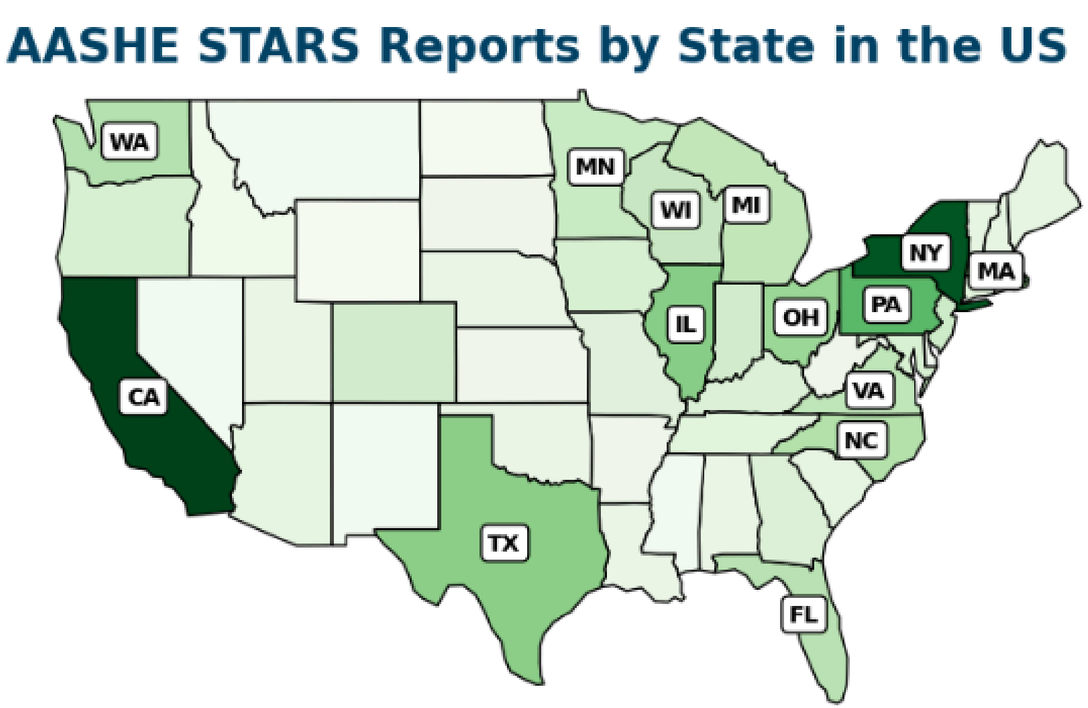
(Here I passed the list of tags to fetch the data from.)

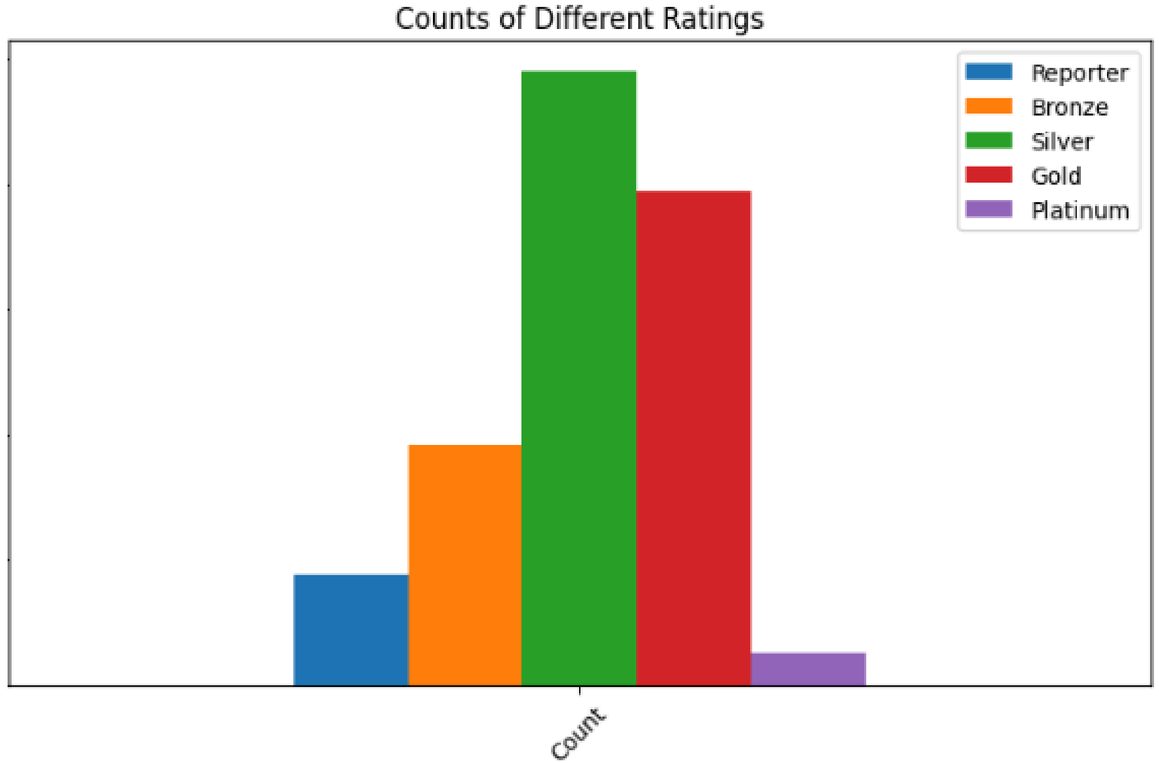
Output dataframe with performance scores for all Universities

	Value	Academic Courses	Learning Outcomes	Undergraduate Program	Graduate Program	Immersive Experience	Sustainability Literacy Assessment	Incentives for Developing Courses	Campus as a Living Laboratory	Research and Scholarship	...	Assessing Diversity and Equity	Support for Underrepresented Groups	Affo ar
0	Penn State Erie, The Behrend College	12.91/14.00	7.52/8.00	3.00/3.00	0.00/0	2.00/2.00	0.00/4.00	0.00/2.00	4.00/4.00	12.00/12.00	...	1.00/1.00	3.00/3.00	
1	California State University, Fresno	9.20/14.00	8.00/8.00	3.00/3.00	3.00/3.00	2.00/2.00	1.00/4.00	0.00/2.00	4.00/4.00	7.69/12.00	...	0.00/1.00	2.83/3.00	
2	Lewis University	0.86/14.00	2.11/8.00	3.00/3.00	0.00/3.00	2.00/2.00	0.00/4.00	0.00/2.00	4.00/4.00	4.97/12.00	...	0.00/1.00	0.00/3.00	
3	DePauw University	7.87/14.00	0.39/8.00	3.00/3.00	0.00/0	2.00/2.00	0.00/4.00	0.00/2.00	1.60/4.00	9.80/12.00	...	1.00/1.00	2.83/3.00	
4	New Mexico State University	2.81/14.00	5.32/8.00	3.00/3.00	3.00/3.00	0.00/2.00	0.00/4.00	0.00/2.00	4.00/4.00	5.62/12.00	...	0.75/1.00	2.00/3.00	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
273	University of Tennessee at Knoxville	6.69/14.00	0.70/8.00	3.00/3.00	3.00/3.00	2.00/2.00	0.00/4.00	2.00/2.00	4.00/4.00	9.58/12.00	...	0.88/1.00	1.92/3.00	
274	Hope	7.10/14.00	0.71/8.00	3.00/3.00	0.00/0	2.00/2.00	1.00/4.00	2.00/2.00	4.00/4.00	12.00/12.00	...	1.00/1.00	3.00/3.00	

(The final Python script iterated through each report URL, opened the first report, and subsequently stored the data in an output dataframe.)

Data visualization using various python libraries





(**I am currently working on the data visualization aspect of this project and looking for someone with Python visualization skills to collaborate. If you're interested, please feel free to reach out!)

Summary

1)Defining the data sources
As the first step of the project, I extracted URLs for the sustainability reports for over 350 Universities from the AASHE STARS Participants & Reports portal.

2)Data Extraction & Transformation 🔀
I then wrote a Python script that opened each report URL, accessed the first report for each university, and then finally extracted the necessary data. This extraction was done using the BeautifulSoup package in python.

3)Data Loading 📊
The extracted data was then organized into an output dataframe, prepared for analysis and visualization.

4)Data Visualization 📈 (This part is still work in progress)
Currently, I am focusing on the visualization of this data. I'm looking for someone with Python visualization skills to collaborate.

(The code for the project is shared on my github link which is also attached to this post.)