

# Website Scrapping Project

By

**Etieneabasi Kingsley Effiong**

**Topic: World University Rankings of The Top Recognized Higher-Education Institution for 2022**

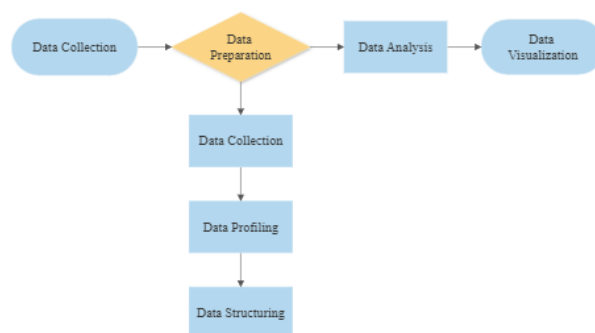
Many times, despite having access to a wealth of information online, people are unable to identify the world's top university. By ranking a list of institutions from across the world using factors like license or accreditation, degrees provided, modes of learning, and educational format as selection criteria, UniRank has made that simple.

## Aim and Objectives

The aim of this project is to scrape the UniRank website, analyse the data obtained and further visualize the data.

## Flow Process

The steps taken are illustrated with the flowchart below:



## Data Sourcing

The website consists of the top universities in the world and their ranking for the year 2022. The data was extracted from the link below:

[\[Top 200 Universities in the World | 2022 World University Ranking \(4icu.org\)\]](https://www.4icu.org/top200.php)

## Data Preparation

The following steps were followed in the data preparation process:

- **Data collection:** The Hypertext Markup Language (Html) tags and elements handling the data to be extracted were identified using the page inspection elements on the Google Chrome browser application, to help me know which tags to call out when extracting data from the website.
- **Data discovery and profiling:** The `<img>` html tag was within the `<td>` tag and the `<td>` tag value was the country abbreviation. The `<img>` tag had the alt element which contained the country name. I got the value of the alt element that is, the country

name and created an extra column to hold the country names as it was not visible for frontend users on the website as shown in the figures below.

2022 Top 200 Universities in the world


Rank	University	Country
1	Massachusetts Institute of Technology	 us
2	Harvard University	 us
3	Stanford University	 us

Figure 1: Frontend users' view

```
<td>  
    
  us  
</td>
```

Figure 2: Alt element value

- **Data structuring:** The data extracted using BeautifulSoup and was imported to an excel workbook using a python library known as the 'openpyxl'. The excel workbook had four columns namely: Rank, University, Country, and Country Abbrev.

## Data Analysis and Visualisation

After the analysing process was done, visualisation was carried out to comprehend the prepared dataset using plots with visual representations including bar charts, scatter plots, tree maps, and pie charts.

```
data.describe()
```

	RANK
count	200.000000
mean	100.500000
std	57.879185
min	1.000000
25%	50.750000
50%	100.500000
75%	150.250000
max	200.000000

Figure 3: Data Description

The figure above shows a description of the scraped data that was analysed. It takes into account the count - number of not empty values, mean - average mean value, std - standard deviation, min - minimum value, the 25%, 50% & 75% percentile and the max - maximum value.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype
---  -
0   RANK             200 non-null    int16
1   UNIVERSITY       200 non-null    object
2   COUNTRY          200 non-null    object
3   COUNTRY ABBRV    200 non-null    object
dtypes: int16(1), object(3)
memory usage: 5.2+ KB
```

*Figure 4: Data Information*

The figure above shows the data type in each of the columns and their non-null counts.

```
# grouping of Countries and their Universities
CountryGroup=data.groupby('COUNTRY').value_counts(ascending=False).sort_index(ascending=False)
print(CountryGroup)
```

COUNTRY	RANK	UNIVERSITY	COUNTRY	ABBRV
United States	196	Clemson University	us	1
	195	The University of Oklahoma	us	1
	190	University of Vermont	us	1
	189	Virginia Commonwealth University	us	1
	188	San Diego State University	us	1
				..
Australia	67	The University of Melbourne	au	1
	65	The University of Sydney	au	1
	56	Monash University	au	1
	54	The University of New South Wales	au	1
Argentina	96	Universidad de Buenos Aires	ar	1

Length: 200, dtype: int64

*Figure 5: Grouping of universities by countries*

The figure above shows the grouping of universities according to their various countries. This was derived using the groupby () function.

```
# total number of times a country occurred in the country column
valueCount=data['COUNTRY'].value_counts()
print(valueCount)
```

```
✓ 0.6s
```

United States	105
United Kingdom	18
Germany	15
Canada	14
Australia	7
Netherlands	5
Spain	4
Switzerland	4
Sweden	3
Czech Republic	2
Italy	2
Belgium	2
Austria	2
China	2
Japan	2
Norway	2
Argentina	1
Mexico	1
Singapore	1
Denmark	1
Finland	1
New Zealand	1

Figure 6: Value Count

The figure above shows the total number of times a country occurred in the country column.

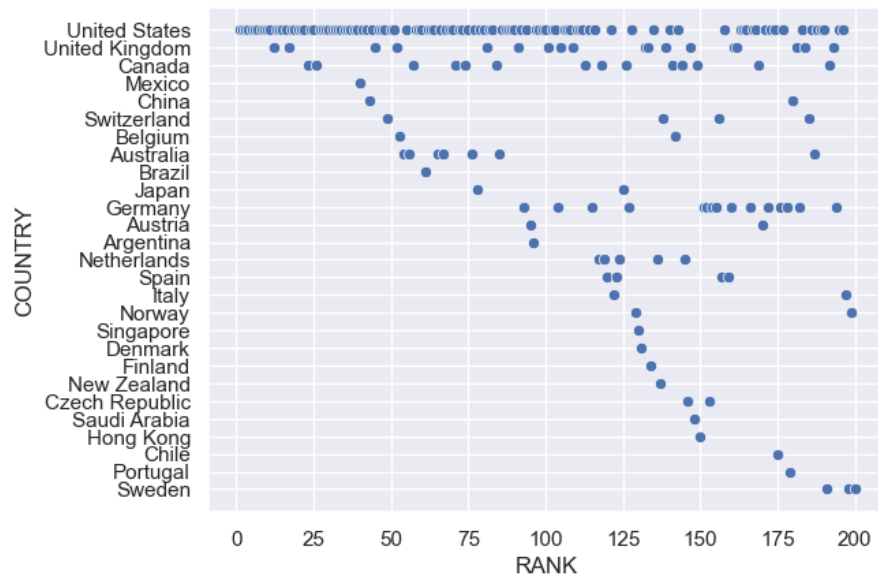


Figure 7: Country Vs Rank using scatter plot

The scatter plot above illustrates the relationship between the ranking and the country where ranking ranges from 0-200. Each dot on the scatterplot represents one country from the data set. The location of each point on the graph depends on both the rank and country.

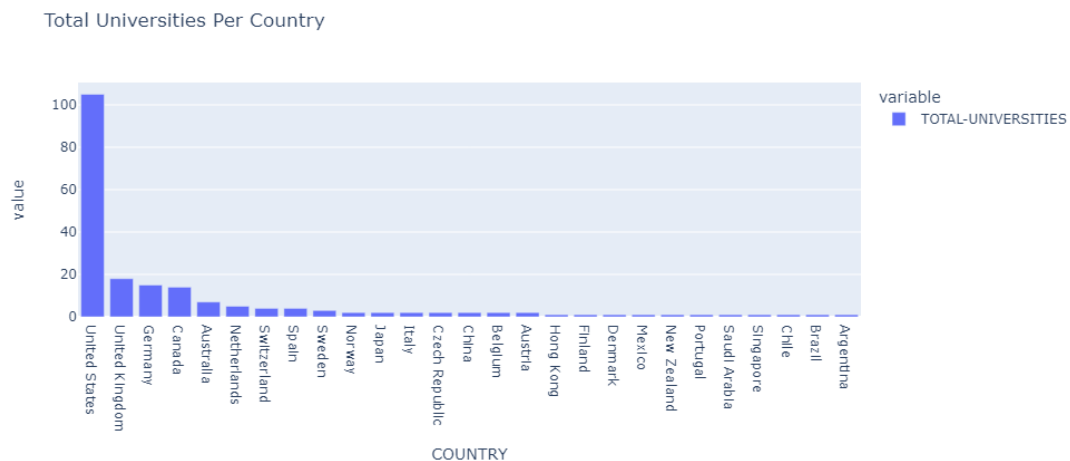


Figure 8: Total Universities Per Country using bar chart

The chart above shows that United States is the country with the highest number with a total of 105 universities. It is followed by the United Kingdom that had 18 universities included in the world ranking, and so on.

## Countries And Their Universities

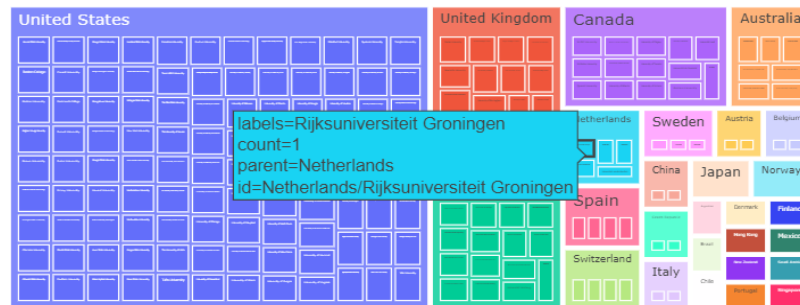


Figure 9: Countries and Their Universities using treemap

The above figure shows the names of the universities mentioned in a particular country. If the cursor is hovered around each of the boxes, it reveals the content of each box.

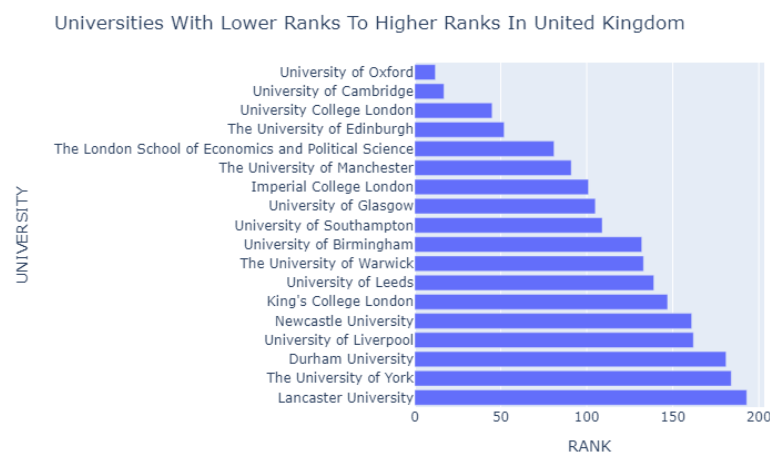


Figure 10: Universities in the United Kingdom from lower ranks to higher ranks

The above chart shows the names of the universities in the United Kingdom and their ranking.

## Results

The data extraction process revealed that the total number of countries ranked was 27, and the total number of universities ranked was 200. The United States had the most universities, followed by the United Kingdom. Some countries had approximately the same number of universities.

The number one university was Massachusetts Institute of Technology in the United States, and the number 200 university was Kungliga Tekniska högskolan in Sweden.

## Conclusion and Recommendation

The results of this web scraping process demonstrates that data may be used to fulfil the highest expectations and provide answers to a variety of issues as long as it is properly extracted. Further information on all analysed and visualised data including the codes to execute the

project, can be accessed with the link - GitHub: <https://github.com/etyking/Hamoye-Web-Scraping-.git>

## Glossary

### A

**Alt Text:** Alternative text is a word or phrase that can be used as an attribute to describe the nature or contents of an image.

### B

**Beautiful Soup:** A Python library used for web scraping to extract data from HTML and XML files. It generates a parse tree from the page source code, which can be used to extract data in a more hierarchical and readable format.

### E

**Elements:** A part of an HTML file that instructs a web browser how to organize and interpret a specific section of the HTML file.

### G

**Groupby () Function:** Used to divide data into groups based on certain criteria.

### H

**HTML:** Also known as Hypertext Markup Language, is a standardized system for tagging text files to achieve font, colour, graphic, and hyperlink effects on World Wide Web pages.

### I

**<img>:** Img or IMG is an abbreviation for image.

### P

**Python Library:** A collection of useful functions that eliminates the need to write code from scratch.

### T

**Tags:** Similar to keywords in that they define how a web browser will format and display content.

**<td>:** The <td> HTML element defines a cell of a table that contains data.

### W

**Website Scraping:** The automated method of gathering structured web data. Additionally known as web data extraction.