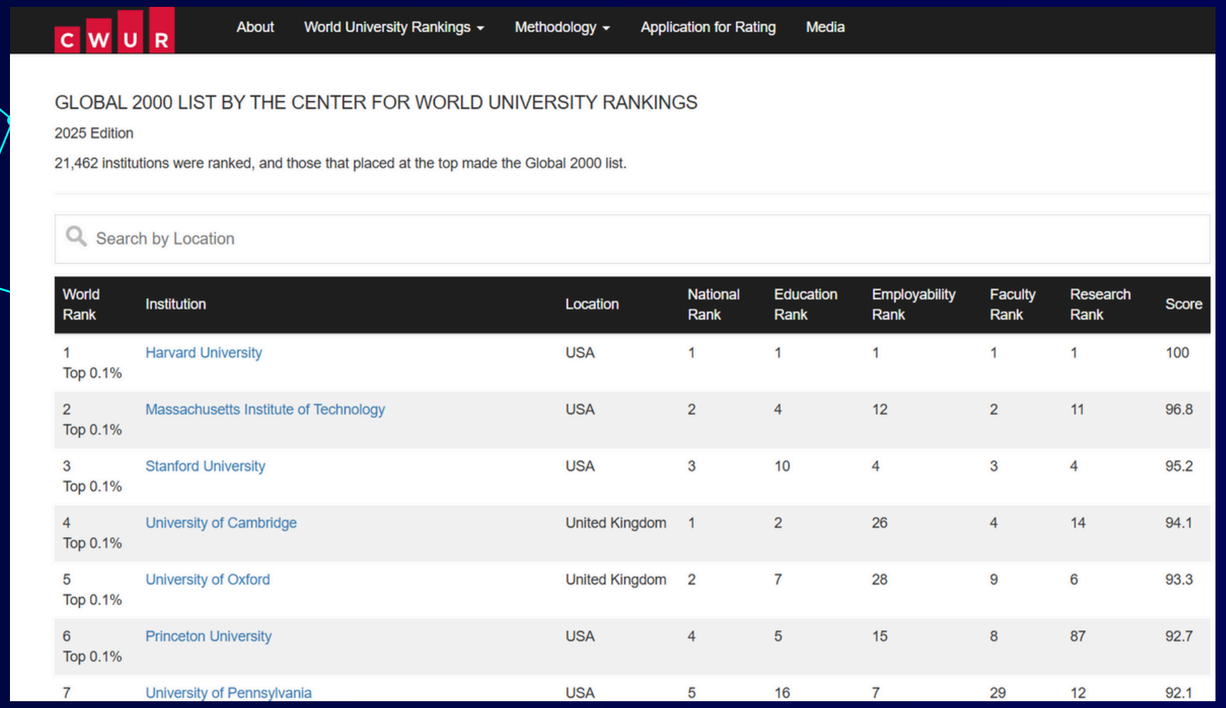
GLOBAL UNIVERSITY RANKINGS ANALYSIS (2020–2025)

Team Members:

- Sakina Saifuddin - 2512137 (Team Manager)
- Sandra Ann Thomas - 2457188
- Bathini Shanmukha priya - 2490778
- Sanju Shree Chandrasekaran - 2447501
- Satish korada - 2400066



CWUR About World University Rankings Methodology Application for Rating Media

GLOBAL 2000 LIST BY THE CENTER FOR WORLD UNIVERSITY RANKINGS

2025 Edition

21,462 institutions were ranked, and those that placed at the top made the Global 2000 list.

Search by Location

World Rank	Institution	Location	National Rank	Education Rank	Employability Rank	Faculty Rank	Research Rank	Score
1 Top 0.1%	Harvard University	USA	1	1	1	1	1	100
2 Top 0.1%	Massachusetts Institute of Technology	USA	2	4	12	2	11	96.8
3 Top 0.1%	Stanford University	USA	3	10	4	3	4	95.2
4 Top 0.1%	University of Cambridge	United Kingdom	1	2	26	4	14	94.1
5 Top 0.1%	University of Oxford	United Kingdom	2	7	28	9	6	93.3
6 Top 0.1%	Princeton University	USA	4	5	15	8	87	92.7
7	University of Pennsylvania	USA	5	16	7	29	12	92.1

Website scraped: https://cwur.org/

ABOUT DATASET

That dataset contains global university rankings (2020-2025) from CWUR, including variables such as:

- World & National Rank
- Education Rank
- Employability Rank
- Faculty Rank
- Research Rank
- Score
- University

OBJECTIVES:

- Collect CWUR university ranking data from 2020 to 2025 using web scraping.
- Combine data from all years into one clean and structured dataset.
- Analyze and compare the rankings of universities and countries over time.
- Identify the top-performing universities and countries each year.
- Explore how different factors (education, employability, faculty, research) affect overall rankings.
- Visualize trends and patterns in global higher education.

```
Requirement already satisfied: tqdm<5.0.0,>=4.42.1 in /usr/local/lib/python3.12/dist-packages (from pyppeteer>=0.0.14->requests-html) (4.67.1)
Collecting urllib3<2.0.0,>=1.25.8 (from pyppeteer>=0.0.14->requests-html)
  Downloading urllib3-1.26.20-py2.py3-none-any.whl.metadata (50 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.1/50.1 kB 2.4 MB/s eta 0:00:00
Collecting websockets<11.0,>=10.0 (from pyppeteer>=0.0.14->requests-html)
  Downloading websockets-10.4.tar.gz (84 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 84.9/84.9 kB 6.1 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.12/dist-packages (from bs4->requests-html) (4.13.5)
Collecting cssselect>=1.2.0 (from pyquery->requests-html)
  Downloading cssselect-1.3.0-py3-none-any.whl.metadata (2.6 kB)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests->requests-html) (3.4.4)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests->requests-html) (3.11)
Requirement already satisfied: zipp>=3.20 in /usr/local/lib/python3.12/dist-packages (from importlib-metadata>=1.4->pyppeteer>=0.0.14->requests-html) (3.23.0)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.12/dist-packages (from pyee<12.0.0,>=11.0.0->pyppeteer>=0.0.14->requests-html) (4.15.0)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.12/dist-packages (from beautifulsoup4->bs4->requests-html) (2.8)
Downloading requests_html-0.10.0-py3-none-any.whl (13 kB)
Downloading lxml_html_clean-0.4.3-py3-none-any.whl (14 kB)
Downloading pyppeteer-2.0.0-py3-none-any.whl (82 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 82.9/82.9 kB 5.7 MB/s eta 0:00:00
Downloading bs4-0.0.2-py2.py3-none-any.whl (1.2 kB)
Downloading fake_useragent-2.2.0-py3-none-any.whl (161 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 161.7/161.7 kB 12.8 MB/s eta 0:00:00
Downloading parse-1.20.2-py2.py3-none-any.whl (20 kB)
Downloading pyquery-2.0.1-py3-none-any.whl (22 kB)
Downloading w3lib-2.3.1-py3-none-any.whl (21 kB)
Downloading appdirs-1.4.4-py2.py3-none-any.whl (9.6 kB)
Downloading cssselect-1.3.0-py3-none-any.whl (18 kB)
Downloading pyee-11.1.1-py3-none-any.whl (15 kB)
Downloading urllib3-1.26.20-py2.py3-none-any.whl (144 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 144.2/144.2 kB 10.7 MB/s eta 0:00:00
Building wheels for collected packages: websockets
  Building wheel for websockets (setup.py) ... done
  Created wheel for websockets: filename=websockets-10.4-cp312-cp312-linux_x86_64.whl size=107325 sha256=e6ef0590a02cae5809decca638af873423e906b9ff4023698ec9519c17dc059b
  Stored in directory: /root/.cache/pip/wheels/80/cf/6d/5d7e4c920cb41925a178b2d2621889c520d648bab487b1d7fd
Successfully built websockets
Installing collected packages: parse, appdirs, websockets, w3lib, urllib3, pyee, lxml_html_clean, fake-useragent, cssselect, pyquery, pyppeteer, bs4, requests-html
  Attempting uninstall: websockets
    Found existing installation: websockets 15.0.1
    Uninstalling websockets-15.0.1:
      Successfully uninstalled websockets-15.0.1
  Attempting uninstall: urllib3
    Found existing installation: urllib3 2.5.0
    Uninstalling urllib3-2.5.0:
      Successfully uninstalled urllib3-2.5.0
```

LIBRARIES USED:

1. requests-html
2. lxml_html_clean
3. nest_asyncio.apply()

All these used for ascynchronous web scraping and operations on collab.

DATA SCRAPING METHODOLOGY

Start
↓
Define fetch_cwur(year)
↓
Build dynamic URL based on year
↓
Load webpage using AsyncHTMLSession
↓
Render JavaScript content
↓
Find HTML table (#cwurTable)
↓
Extract table rows & columns
↓ →

Attach header + add year column
↓
Return data for the year
↓
Define scrape_all_years()
↓
Loop through years 2020–2025
↓
Call fetch_cwur(year) for each year
↓
Combine all rows into master list
↓
Export/save final dataset (CSV/DataFrame)
↓
End

```python
async def fetch_cwur(year):
    # Dynamically construct the URL based on the year
    if year >= 2023: # For 2023, 2024, 2025 and onwards
        url = f"https://cwur.org/{year}.php"
    else: # For 2020, 2021, 2022 use YYYY-YY.php format
        # Example: 2020 becomes 2020-21
        end_year_suffix = str(year + 1)[-2:]
        url = f"https://cwur.org/{year}-{end_year_suffix}.php"

    session = AsyncHTMLSession()
    r = await session.get(url)
    await r.html.arender(sleep=2, timeout=30)  # async rendering

    # Find the table
    table = r.html.find('table#cwurTable', first=True)

    all_rows_for_year = []

    # Define the header row (needed by scrape_all_years logic)
    header_row = [
        "World Rank",
        "University",
        "Country",
        "National Rank",
        "Education Rank",
        "Employability Rank",
        "Faculty Rank",
        "Research Rank",
        "Score",
        "Year"
    ]

    # If table is not found, print a warning and return an empty list of data rows.
    if not table:
        print(f"⚠️ Warning: Table 'cwurTable' not found for year {year} at {url}. Returning no data for this year.")
        # Return an empty list of data rows, the caller will handle the header separately.
        return [header_row] # Returning just header to maintain consistency for `scrape_all_years`

    # Iterate through table rows and extract data
    # Skip the first `tr` as it's typically the header in the HTML table.
    for row in table.find('tr')[1:]:
        cols = [c.text for c in row.find('td')]
        if cols:
            cols.append(str(year)) # Add the year as a string
            all_rows_for_year.append(cols)

    # Prepend the header row only if this fetch_cwur call is the one that's expected to provide the header.
    # However, the `scrape_all_years` function already handles adding the header once.
    # So, we just return the data rows, and `scrape_all_years` will manage the header from the first successful year's fetch.
    # To fit the current `scrape_all_years` logic (which expects data[0] to be header for first_year, and then data[1:] for subsequent), we always include the header here.
    all_rows_for_year.insert(0, header_row) # Insert header at the beginning

    return all_rows_for_year
```

DEFINE FETCH_CWUR(YEAR) FUNCTION

This function handles scraping one specific year.
Inside this function:

📌 a. Construct Dynamic URL

- If year ≥ 2023 → format is https://cwur.org/{year}.php
- If year ≤ 2022 → format uses "YY-YY" format
- Example: 2020 → 2020-21 → https://cwur.org/2020-21.php

📌 b. Create an AsyncHTMLSession

- Opens an asynchronous browser session.

📌 c. Load the Webpage

- Uses session.get(url) to fetch HTML

```python
all_years_data = []
years_to_scrape = range(2020, 2026) # Years from 2020 to 2025

async def scrape_all_years():
    global all_years_data # Declare all_years_data as global to modify it
    first_year = True

    for year in years_to_scrape:
        print(f"Scraping data for year: {year}")
        data = await fetch_cwur(year)

        if first_year:
            all_years_data.extend(data) # Include header for the first year
            first_year = False
        else:
            all_years_data.extend(data[1:]) # Exclude header for subsequent years
        print(f"✅ Data for year {year} collected.")

    print("✅ All years data combined successfully.")
```

DEFINE SCRAPE_ALL_YEARS() FUNCTION

This function collects data for every year you're scraping (2020–2025).

📌 a. Create an Empty List: all_years_data = []

📌 b. Loop Through Each Year

For every year inside range(2020, 2026):

- Print status message
- Call data = await fetch_cwur(year)

📌 **c. Combine Data**

- First year → include header
- Remaining years → skip header
- Append scrapped rows into all_years_data

📌 **d. Print Success Messages**

```
Scraping data for year: 2020
[INFO] Starting Chromium download.
INFO:pyppeteer.chromium_downloader:Starting Chromium download.
100%|██████████| 183M/183M [00:00<00:00, 229Mb/s]
[INFO] Beginning extraction
INFO:pyppeteer.chromium_downloader:Beginning extraction
[INFO] Chromium extracted to: /root/.local/share/pyppeteer/local-chromium/1181205
INFO:pyppeteer.chromium_downloader:Chromium extracted to: /root/.local/share/pyppeteer/local-chromium/1181205
✅ Data for year 2020 collected.
Scraping data for year: 2021
✅ Data for year 2021 collected.
Scraping data for year: 2022
✅ Data for year 2022 collected.
Scraping data for year: 2023
✅ Data for year 2023 collected.
Scraping data for year: 2024
✅ Data for year 2024 collected.
Scraping data for year: 2025
✅ Data for year 2025 collected.
✅ All years data combined successfully.
```

Total Rows scraped: 12001 rows

2020

World Rank	University	Country	National Rank	Education Rank	Employability Rank	Faculty Rank	Research Rank	Score	Year
1	Harvard University	USA	1	3	1	1	1	100	2020
2	Massachusetts Ins	USA	2	4	11	2	7	96.7	2020
3	Stanford Universit	USA	3	10	4	3	2	95.2	2020
4	University of Caml	United Kin	1	1	21	4	11	94.1	2020
5	University of Oxfo	United Kin	2	7	26	9	4	93.3	2020
6	Columbia Universi	USA	4	11	14	10	15	92.6	2020
7	Princeton Universi	USA	5	6	15	7	71	92	2020
8	University of Penn	USA	6	14	9	43	12	91.6	2020
9	University of Chica	USA	7	8	18	29	22	91.1	2020
10	Yale University	USA	8	5	35	11	20	90.7	2020
11	California Institute	USA	9	2	128	5	74	90.4	2020
12	University of Calif	USA	10	9	57	8	14	90.1	2020
13	Cornell University	USA	11	13	28	17	24	89.8	2020
14	University of Toky	Japan	1	37	7	105	27	89.5	2020
15	Johns Hopkins Uni	USA	12	20	188	22	6	89.3	2020
16	University of Mich	USA	13	32	37	82	5	89	2020
17	Northwestern Uni	USA	14	139	17	46	29	88.8	2020
18	University of Calif	USA	15	49	63	19	10	88.6	2020
19	University College	United Kin	3	33	666	14	8	88.4	2020
20	Duke University	USA	16	61	51	16	21	88.2	2020
21	PSL University	France	1	16	27	100	121	88	2020
22	University of Illino	USA	17	23	133	12	61	87.8	2020
23	University of Wasl	USA	18	60	216	23	9	87.7	2020
24	University of Toro	Canada	1	225	159	21	3	87.5	2020
25	New York Universi	USA	19	43	42	40	45	87.3	2020
26	University of Wisc	USA	20	34	93	34	31	87.2	2020

BASIC DATA EXPLORATION (EDA)

1. Dataset Structure and Dimensions

- **Shape:** The dataset contains 12,000 rows and 10 columns.
- **Columns:** The dataset includes key information like World Rank, University Name, Country, and various other Rank categories (National, Education, Employability, Faculty, Research), Score, and Year (2020 to 2025).
- **Countries:** The data covers universities from 99 unique countries.

2. Missing Values and Data Types

- **Missing Values:** Several key ranking columns have a high number of missing values (represented by '–'), indicating incomplete data:
  - **Faculty Rank:** 10,342 missing.
  - **Education Rank:** 9,388 missing.
  - **Employability Rank:** 5,974 missing.
- **Data Types:** Most columns are initially stored as the Object type (text), even the ranking columns.

3. Outlier Analysis

- **World Rank & Education Rank:** Ranks are widely spread (from 0 to 2,000 for World Rank, 0 to 400 for Education Rank), showing a big difference in performance among universities.
- **National Rank:** Most ranks are below 100, but there are many outliers (above 200), meaning a few universities are ranked significantly lower nationally.
- **Faculty Rank:** Values fall between 50 and 250, suggesting more consistent faculty performance compared to other ranks.
- **Score:** Most scores are clustered tightly between 65 and 75, but there are many significant outliers above 90.

B) UNIVARIATE, BIVARIATE, AND MULTIVARIATE ANALYSES

1. DATA CLEANING

[27]:
```
# Replace '-' with NaN for consistency
cwur_copy.replace('-', np.nan, inplace=True)
```

[28]:
```
cwur_copy.tail()
```

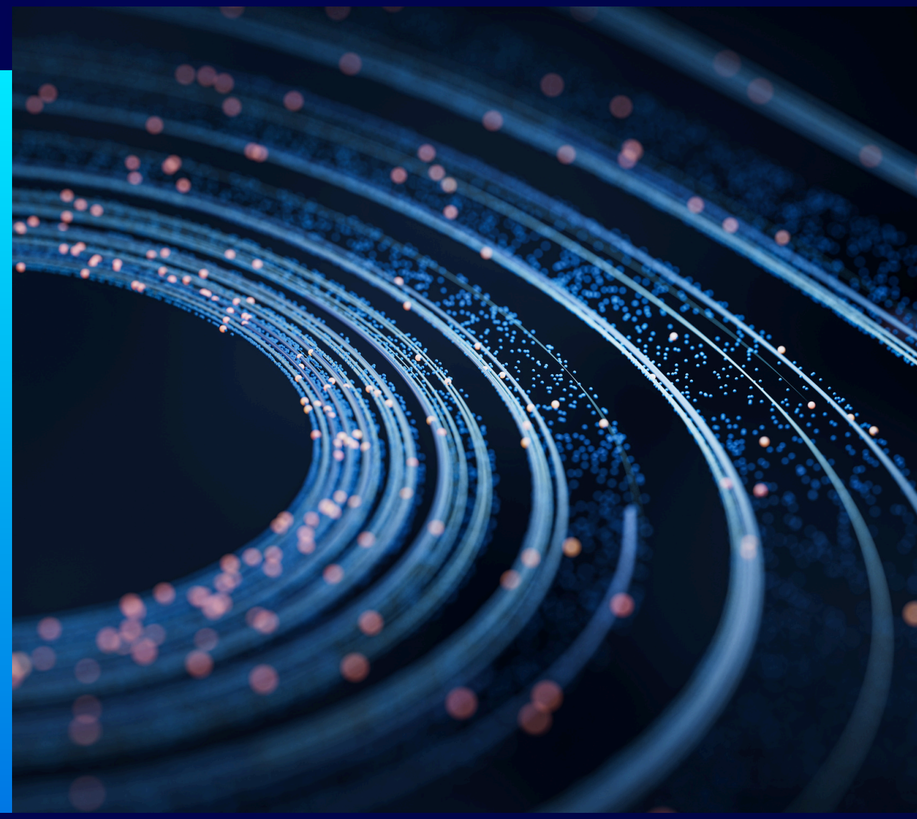
[28]:

	World Rank	University	Country	National Rank	Education Rank	Employability Rank	Faculty Rank	Research Rank	Score	Year
11995	1996\nTop 9.4%	Hunan University of Technology	China	344	NaN	NaN	NaN	1920	66.2	2025
11996	1997\nTop 9.4%	Guizhou Normal University	China	345	NaN	NaN	NaN	1922	66.2	2025
11997	1998\nTop 9.4%	Bengbu Medical University	China	346	NaN	NaN	NaN	1924	66.2	2025
11998	1999\nTop 9.4%	Federal University of Amazonas	Brazil	53	NaN	NaN	NaN	1925	66.2	2025
11999	2000\nTop 9.4%	Başkent University	Turkey	52	NaN	1223	NaN	1930	66.2	2025

Most scores are clustered tightly between 65 and 75, but there are many significant outliers above 90.

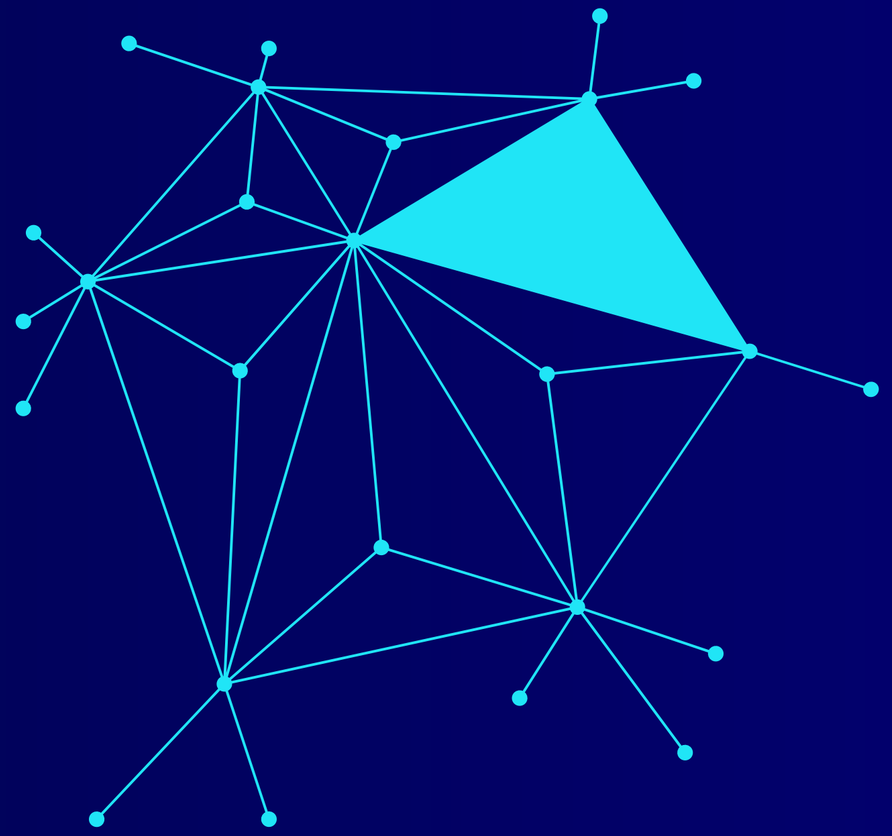
LIBRARIES USED:

Thynk Unlimited combines cutting-edge tools with expert insights to deliver precise and impactful data analysis. From data cleaning and visualization to advanced modeling, we ensure accuracy and relevance in every project.



TOOLS AND TECHNIQUES

We utilize state-of-the-art software and algorithms, including AI-driven analytics, machine learning, and big data technologies. Our approach ensures efficiency, scalability, and the ability to handle complex datasets with ease.



REAL-WORLD APPLICATIONS

Data analysis is essential across various sectors. In healthcare, it predicts patient outcomes; in retail, it enhances customer targeting; in finance, it mitigates risks. Thynk Unlimited adapts solutions to meet the unique needs of every industry.

KEY BENEFITS OF DATA ANALYSIS

Enhanced Decision-Making:

Gain actionable insights to inform strategies.

Risk Mitigation:

Predict and prevent potential issues with data-driven foresight.

Improved Efficiency:

Streamline processes and identify bottlenecks.

Competitive Edge:

Stay ahead of trends and market shifts.

CASE STUDIES

Our data analysis services have transformed businesses. For example, we helped a retail client increase sales by 25% through predictive analytics. Another project enabled a logistics firm to optimize delivery routes, saving both time and costs.

CHALLENGES IN DATA ANALYSIS

Data analysis comes with its own set of challenges, including data quality issues, integration complexities, and security concerns. Thynk Unlimited addresses these challenges with robust strategies, ensuring reliable and secure analysis.

FUTURE OF DATA ANALYSIS

As technology advances, data analysis will become even more integral. AI, IoT, and predictive analytics are paving the way for real-time decision-making and deeper insights. Thynk Unlimited stays ahead of the curve to offer cutting-edge solutions.

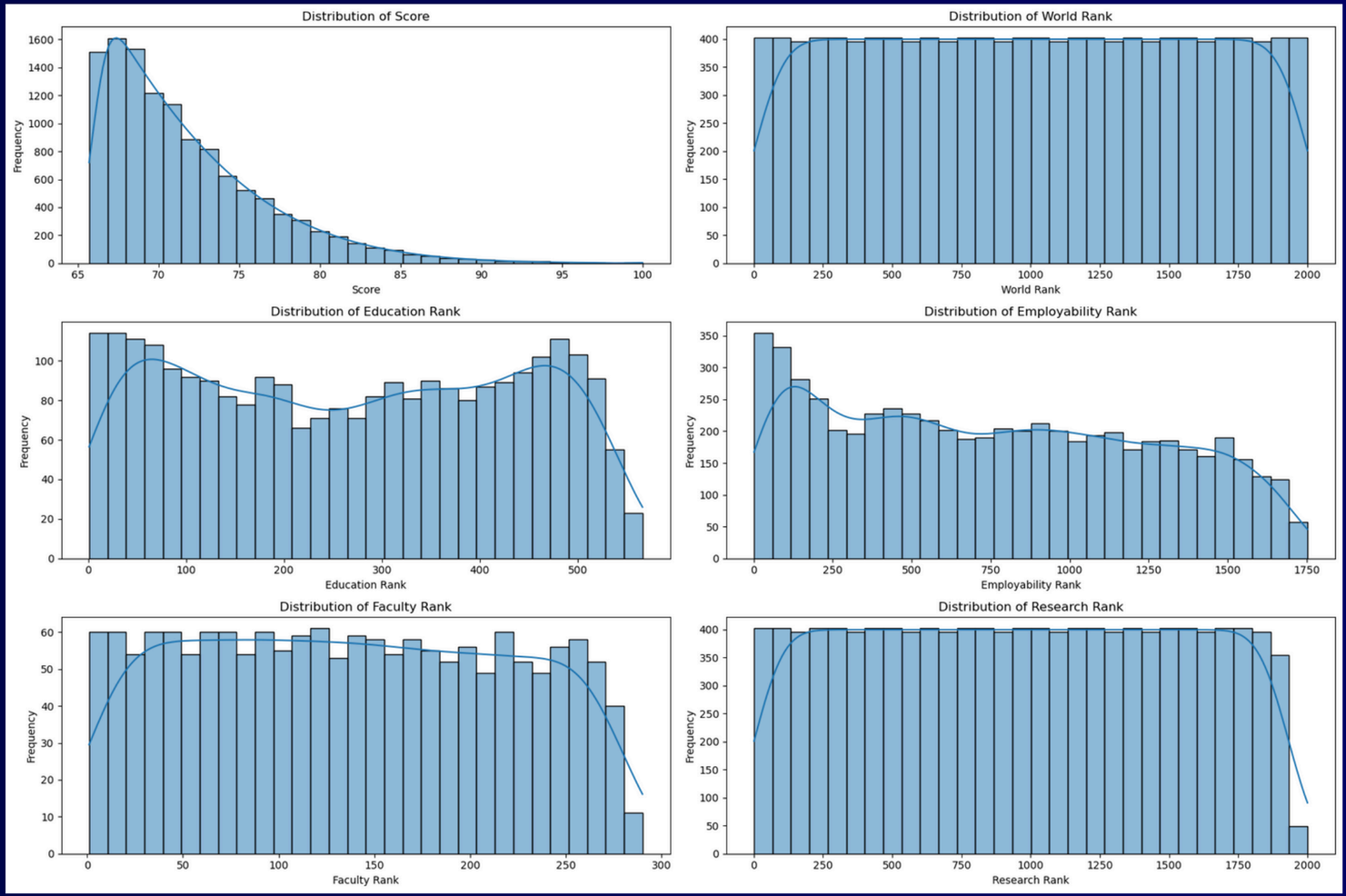
Data Visualization

Deep Cleaning

- Converted ranking columns and Year column to numeric
- Invalid values automatically converted to NaN to avoid errors
- Removed rows with missing Year values
- Year converted to integer for time-based graphs

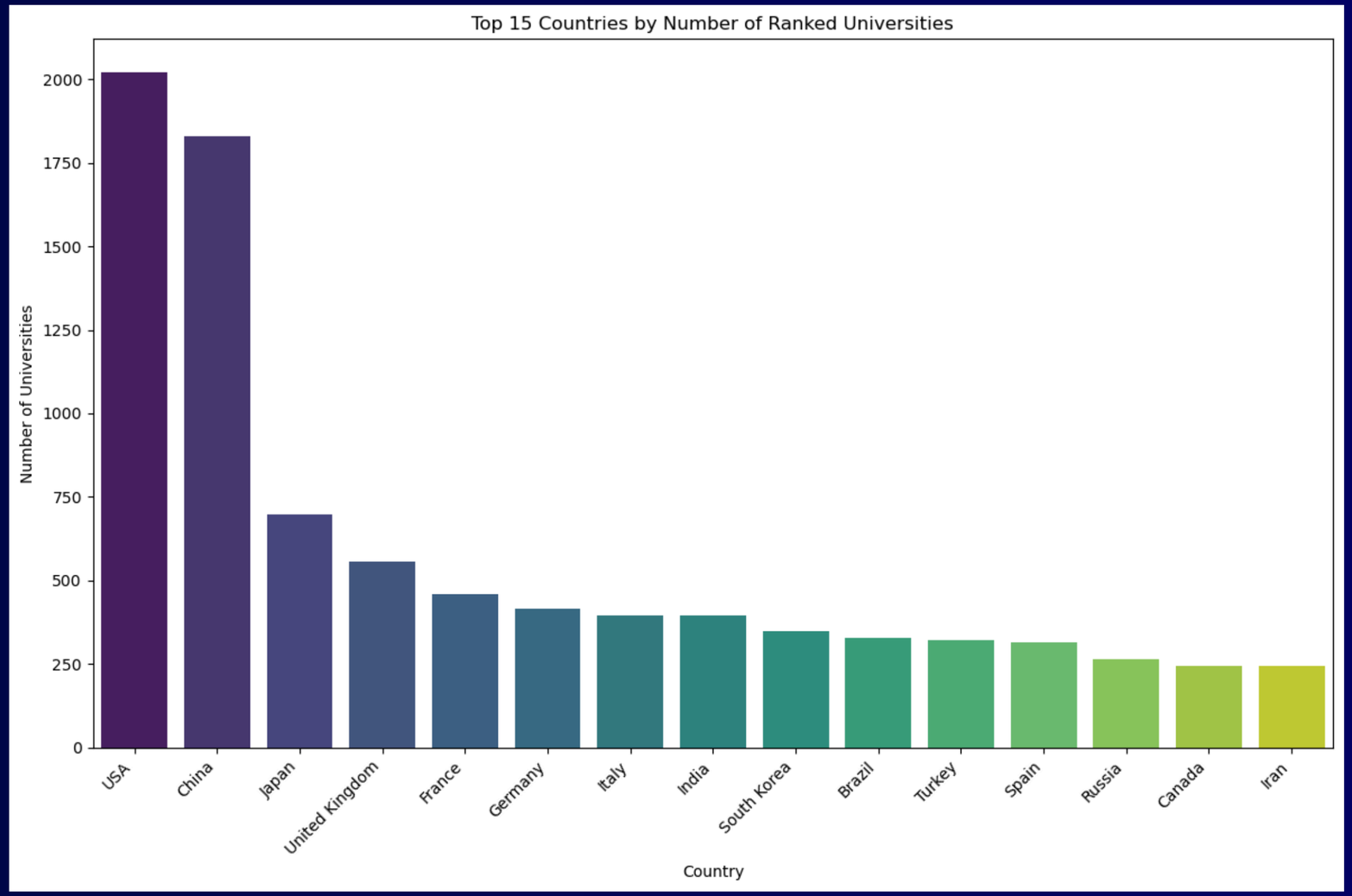
Score & Rank Distributions

- Most universities have high Scores
- Employability Rank is right-skewed → only a few universities have excellent employability outcomes
- Other Rank metrics are more evenly spread across their ranges


Distribution of Score
Frequency
Score
Distribution of World Rank
Frequency
World Rank
Distribution of Education Rank
Frequency
Education Rank
Distribution of Employability Rank
Frequency
Employability Rank
Distribution of Faculty Rank
Frequency
Faculty Rank
Distribution of Research Rank
Frequency
Research Rank

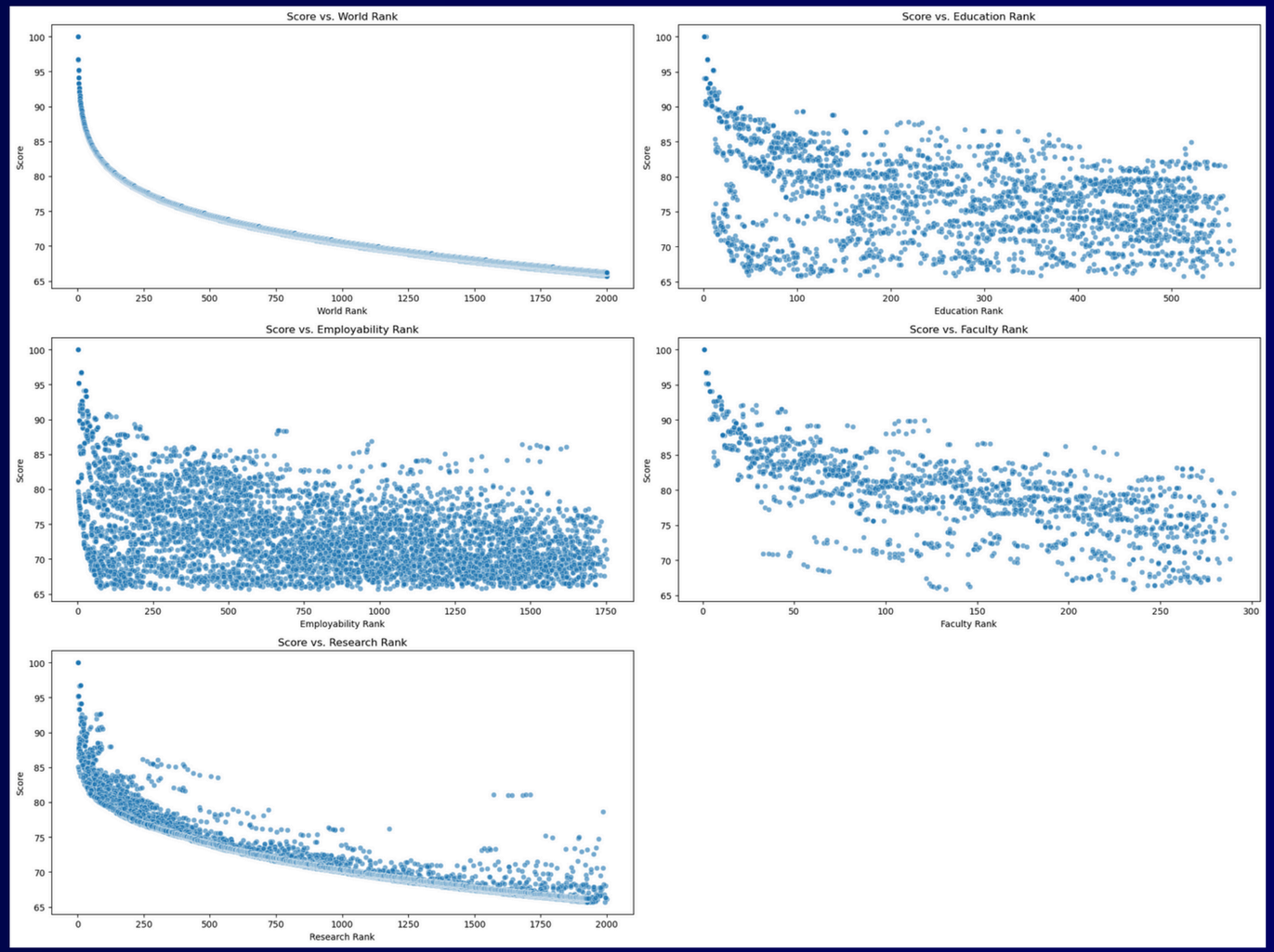
Top 15 Countries by Representation

- USA has the highest number of ranked universities
- Followed by the China, Japan and United Kingdom,


Top 15 Countries by Number of Ranked Universities
Number of Universities
0
250
500
750
1000
1250
1500
1750
2000
USA
China
Japan
United Kingdom
France
Germany
Italy
India
South Korea
Brazil
Turkey
Spain
Russia
Canada
Iran
Country

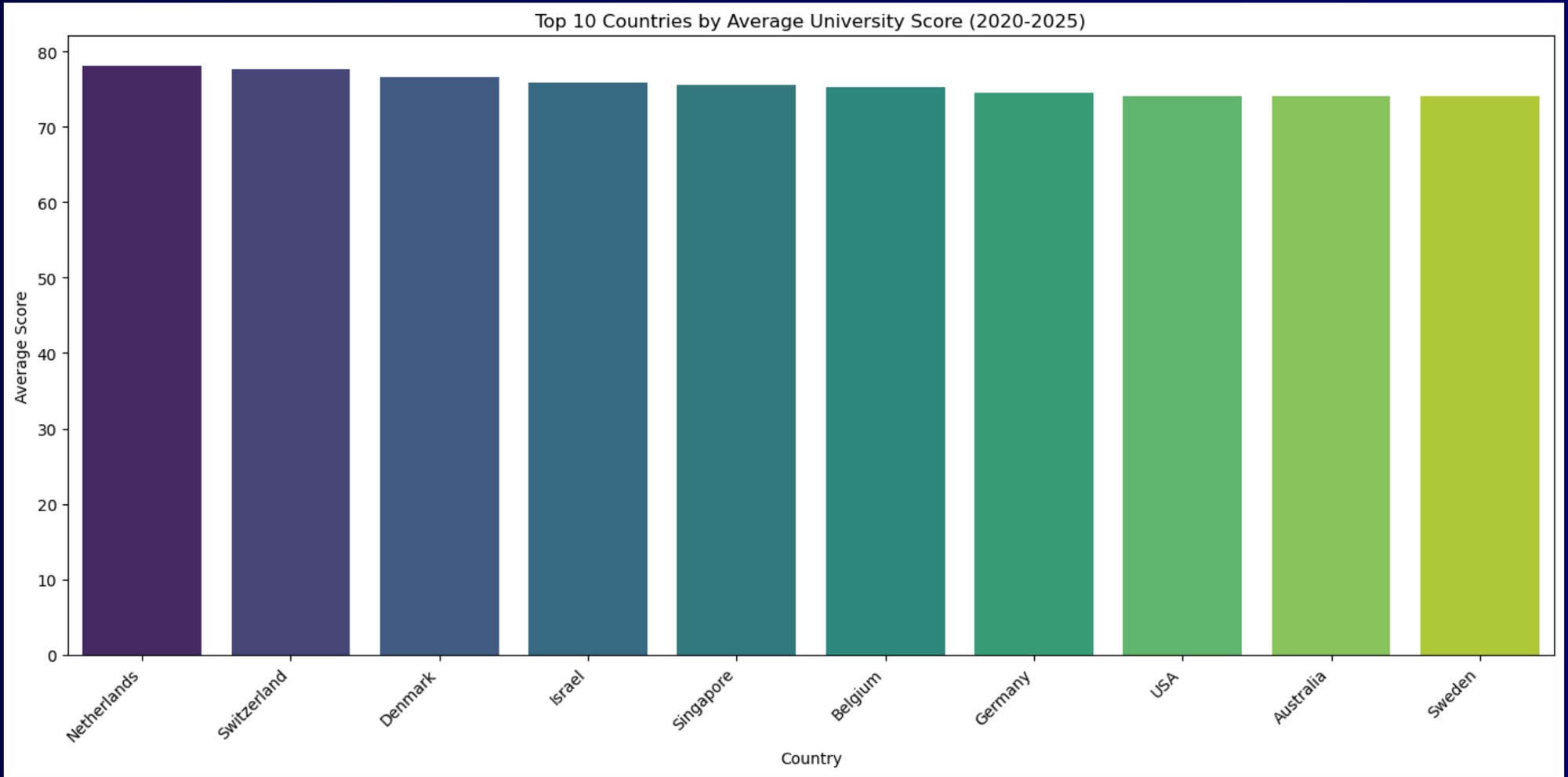
Score vs Ranking Metrics

- Better ranking position → higher score
- Research Rank has the strongest impact on Score among all metrics


Score vs. World Rank
Score
World Rank
Score vs. Education Rank
Score
Education Rank
Score vs. Employability Rank
Score
Employability Rank
Score vs. Faculty Rank
Score
Faculty Rank
Score vs. Research Rank
Score
Research Rank

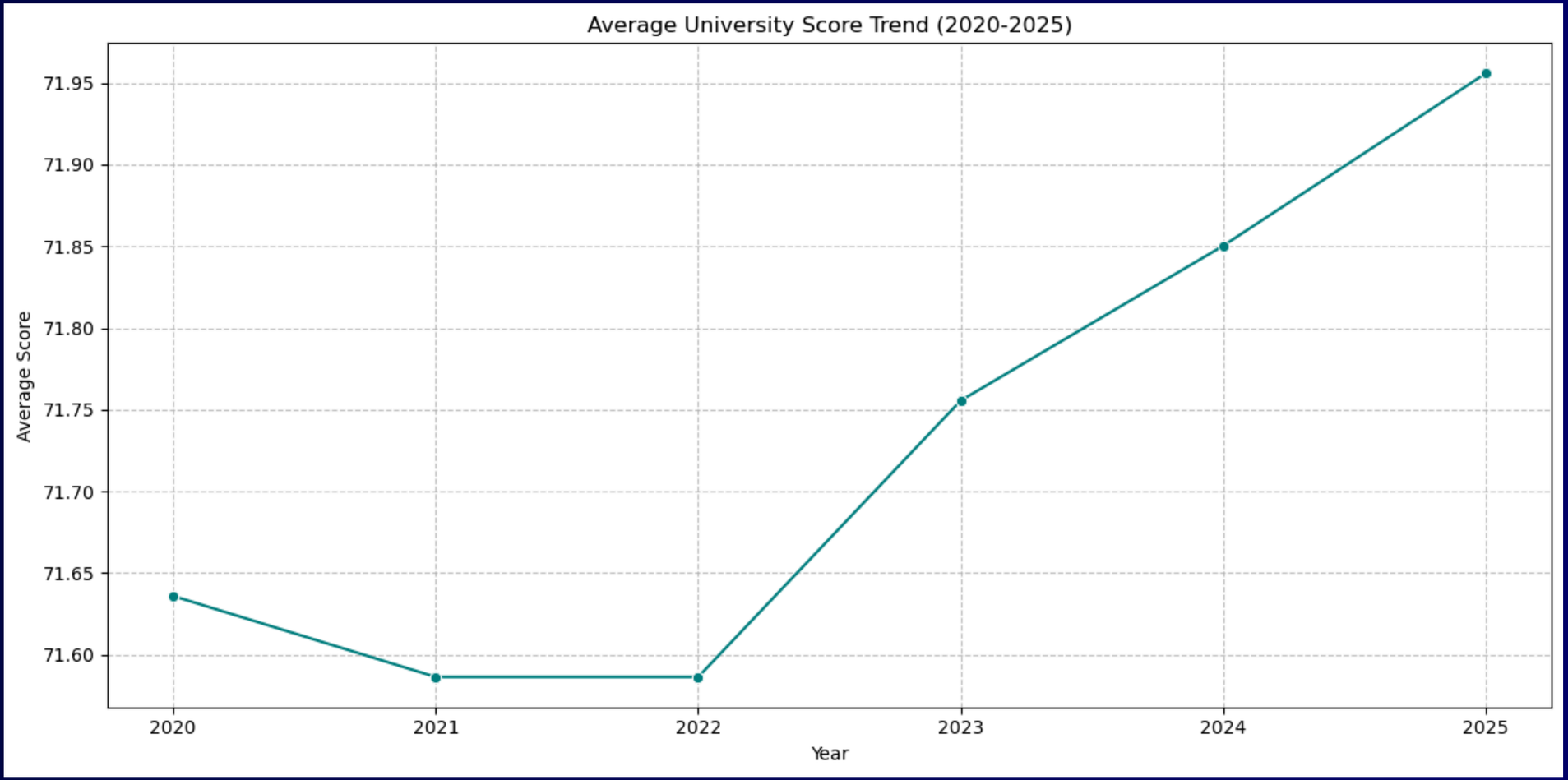
Top 10 Countries by Average Score

- Netherlands, Switzerland, and Denmark have the highest average university scores
- All top 10 countries show consistently strong academic performance aragraph text


Top 10 Countries by Average University Score (2020-2025)
Average Score
0
10
20
30
40
50
60
70
80
Netherlands
Switzerland
Denmark
Israel
Singapore
Belgium
Germany
USA
Australia
Sweden
Country

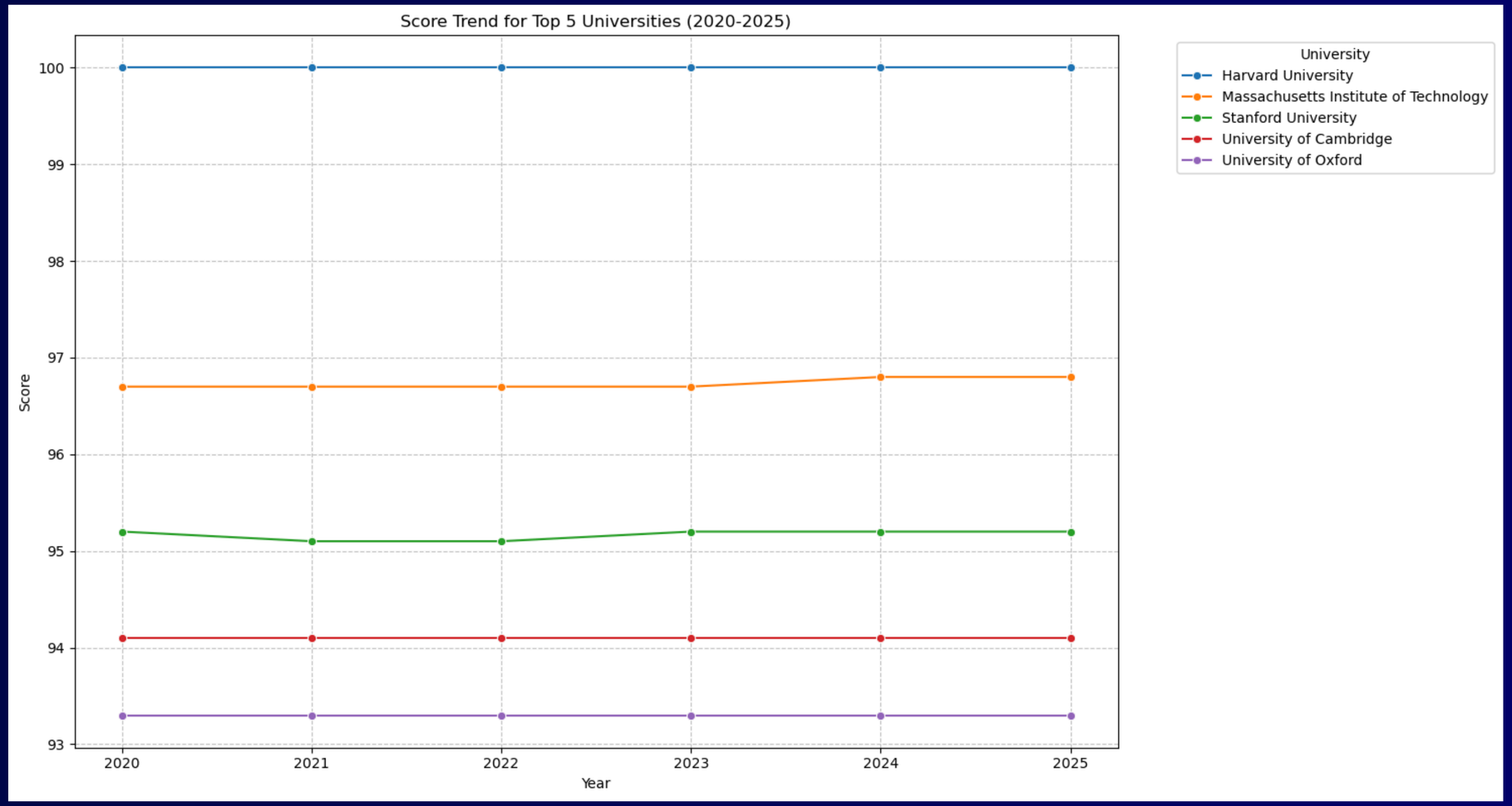
Score Trend Over Time

- Slight drop in average score between 2020 and 2022
- Scores steadily increase from 2023 onward
- Highest average score reached in 2025


Average University Score Trend (2020-2025)
Average Score
71.95
71.90
71.85
71.80
71.75
71.70
71.65
71.60
2020
2021
2022
2023
2024
2025
Year

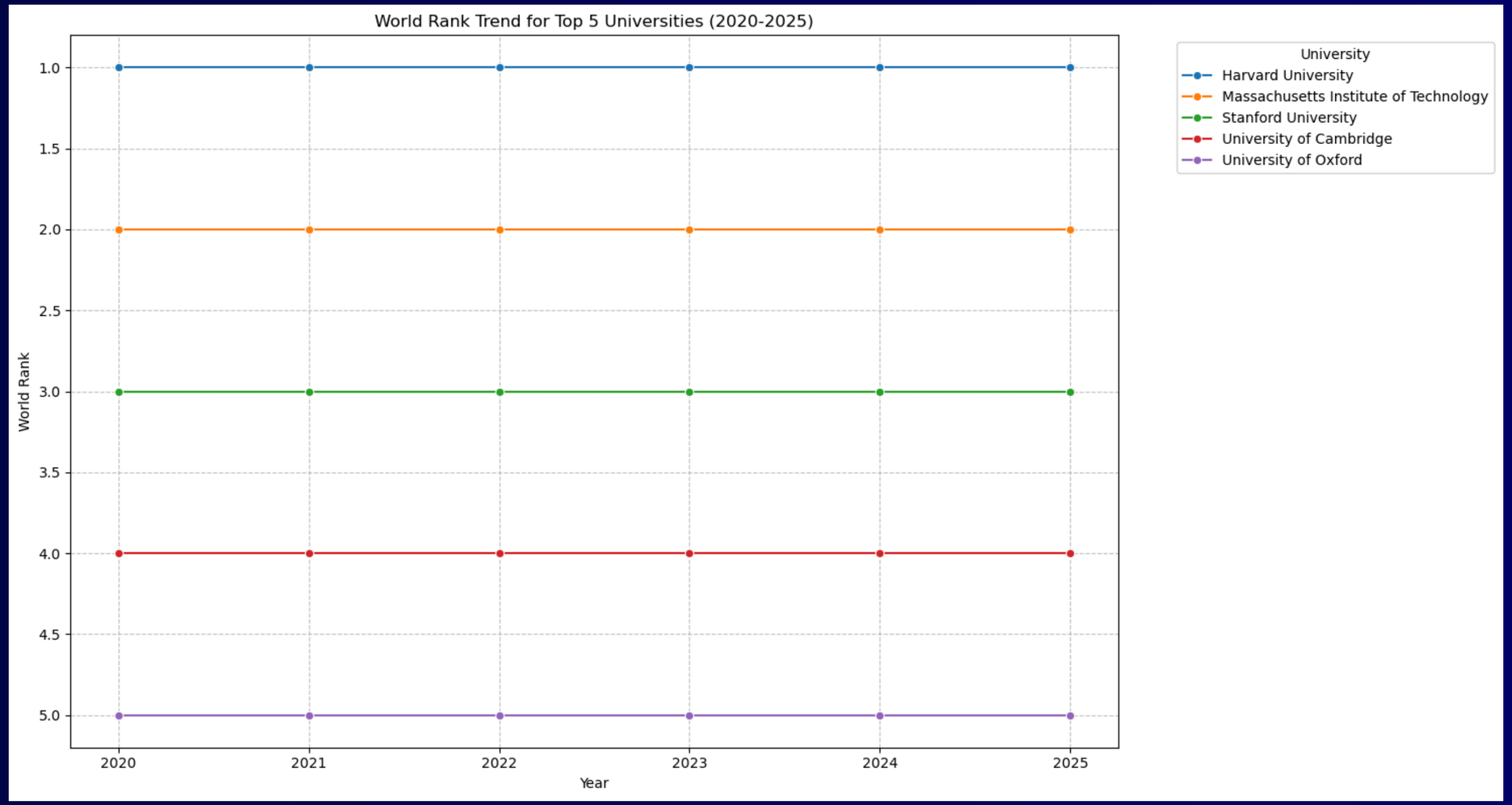
Top 5 Universities

- Harvard, MIT, Stanford, Cambridge, and Oxford are the top five institutions
- Harvard maintains a perfect score every year
- The other four universities remain stable with only minor variations


Score Trend for Top 5 Universities (2020-2025)
Score
Year
100
99
98
97
96
95
94
93
2020
2021
2022
2023
2024
2025
University
Harvard University
Massachusetts Institute of Technology
Stanford University
University of Cambridge
University of Oxford

World Rank Stability

- World Rank remains unchanged for all top five universities from 2020 to 2025
- Indicates long-term and consistent global dominance

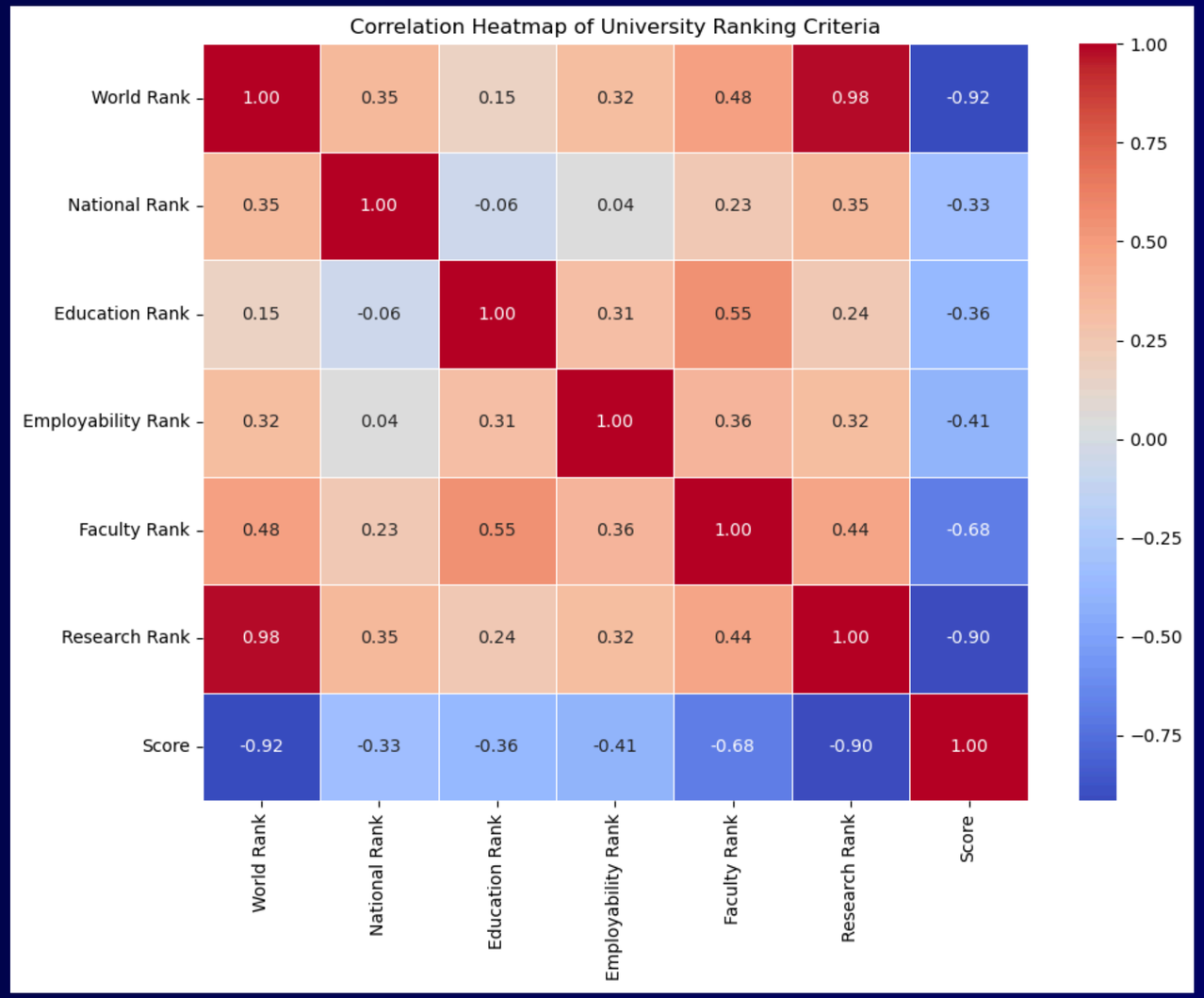

World Rank Trend for Top 5 Universities (2020-2025)
World Rank
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
2020
2021
2022
2023
2024
2025
Year
University
Harvard University
Massachusetts Institute of Technology
Stanford University
University of Cambridge
University of Oxford

Correlation Results

Research Rank most strongly influences Score and World Rank

Score is strongly negative with all ranking metrics

Education Rank and Faculty Rank are moderately related

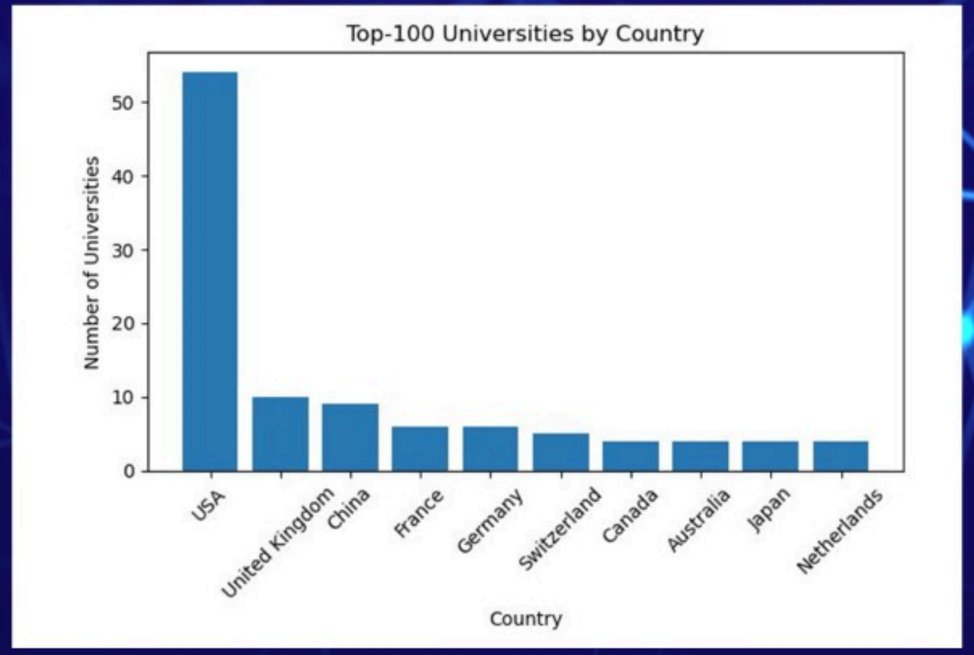

Correlation Heatmap of University Ranking Criteria
World Rank
National Rank
Education Rank
Employability Rank
Faculty Rank
Research Rank
Score
1.00 0.35 0.15 0.32 0.48 0.98 -0.92
0.35 1.00 -0.06 0.04 0.23 0.35 -0.33
0.15 -0.06 1.00 0.31 0.55 0.24 -0.36
0.32 0.04 0.31 1.00 0.36 0.32 -0.41
0.48 0.23 0.55 0.36 1.00 0.44 -0.68
0.98 0.35 0.24 0.32 0.44 1.00 -0.90
-0.92 -0.33 -0.36 -0.41 -0.68 -0.90 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Summary

- Research strength is the strongest driver of global university ranking
- Leading universities and countries maintain consistent long-term performance
- Visualizations reveal clear global patterns in academic outcomes

Data Mining

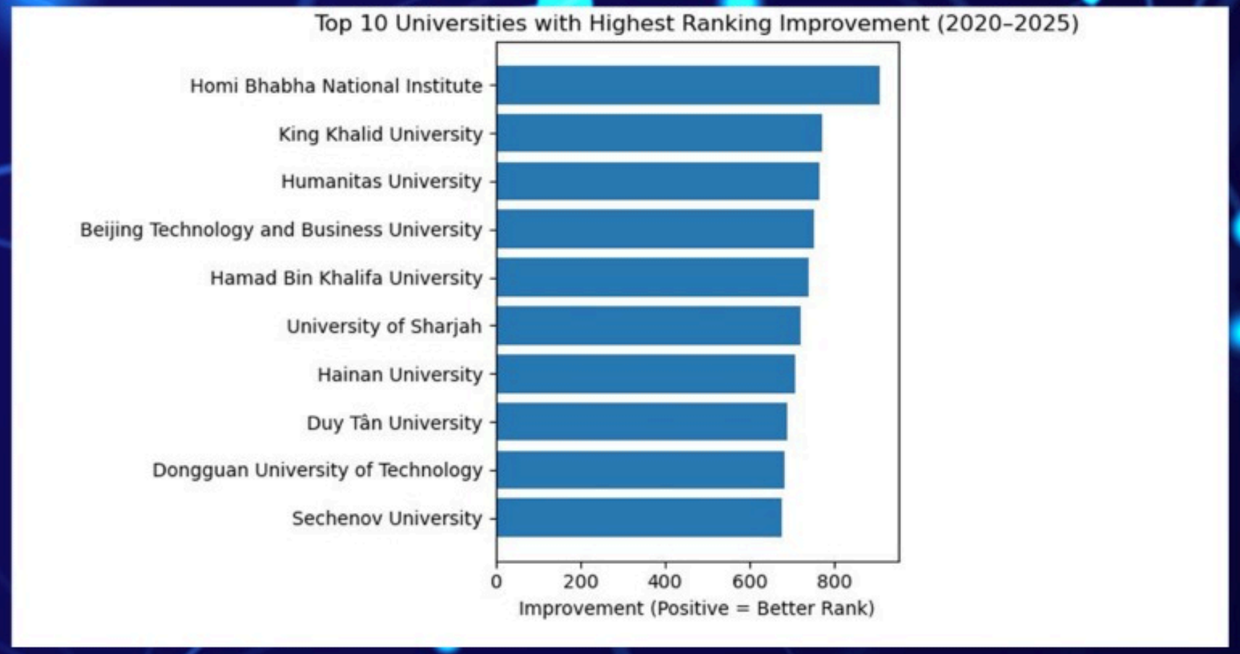
Q1. Countries with highest number of Top-100 universities



USA overwhelmingly dominates with 50+ universities in the top 100

Shows strong academic infrastructure concentration in the US

Q2. Universities that improved the most (2020 → 2025)



Large improvements seen in universities from India, Middle East, and Asia

Emerging countries are rapidly climbing global rankings

3. Correlation between Research Rank & World Rank

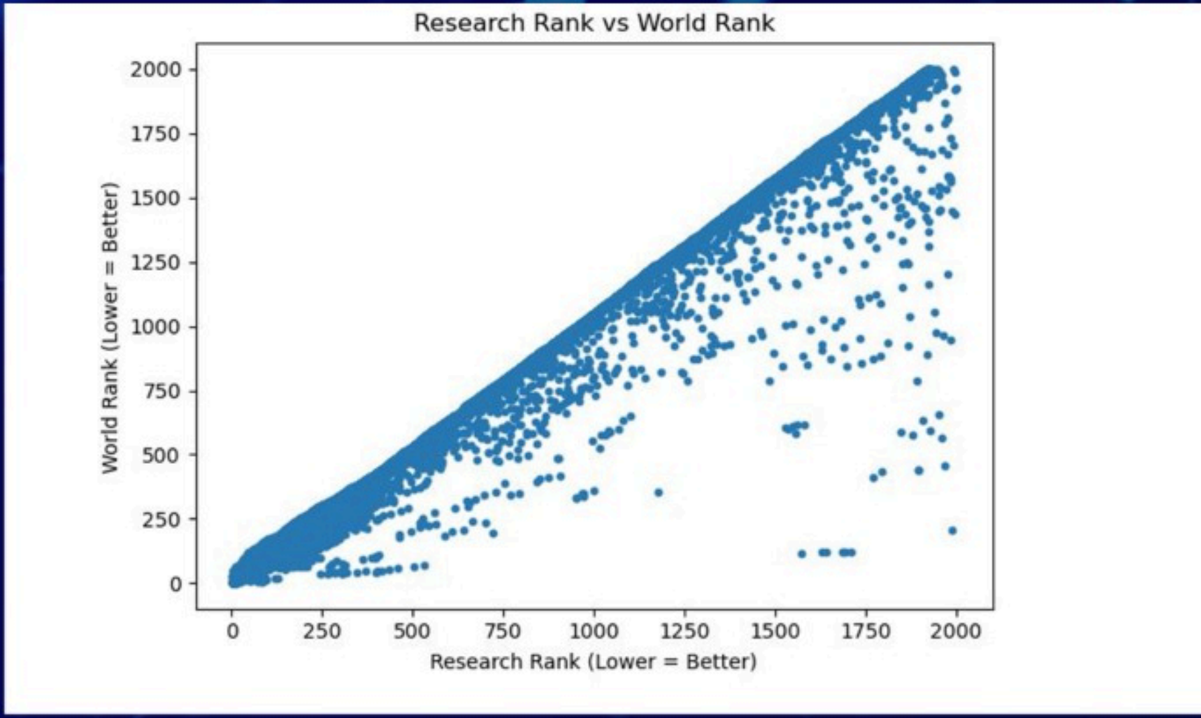
```
def research_vs_world(df):
    df2 = df.dropna(subset=['World Rank', 'Research Rank'])
    corr, pval = spearmanr(df2['Research Rank'], df2['World Rank'])
    return corr, pval
```

```
correlation = research_vs_world(df_clean)
correlation
```

```
(0.9838575310530999, 0.0)
```

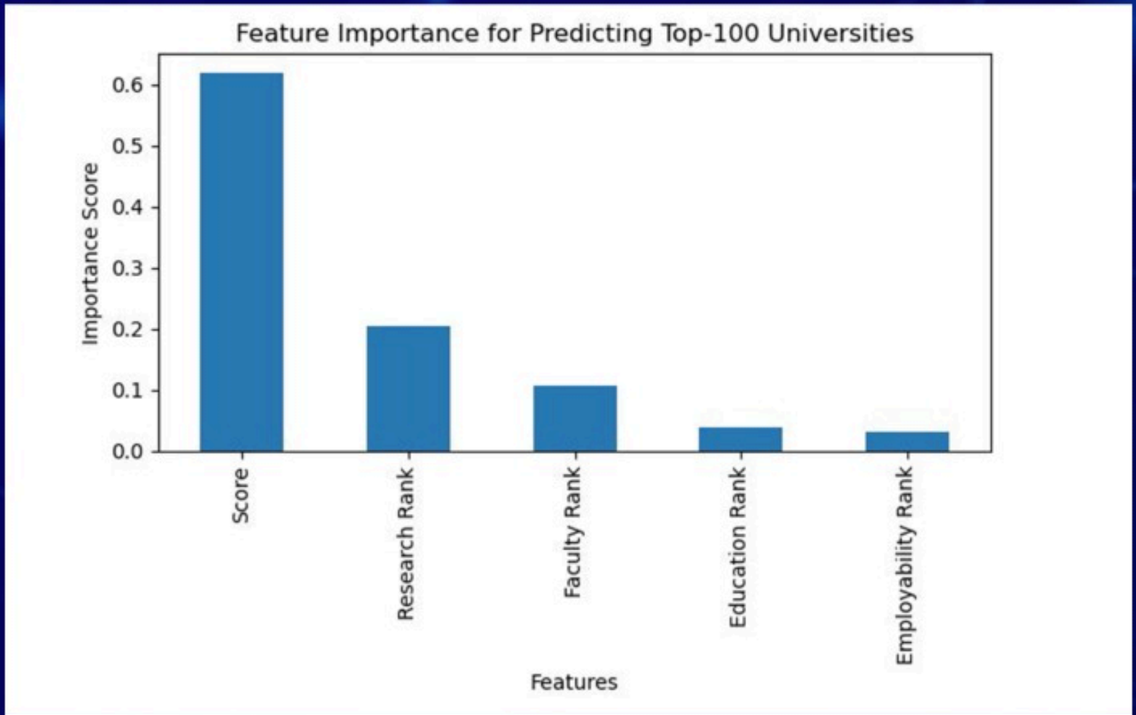
Very strong positive correlation (Spearman ρ ≈ 0.98)

Q4. Future leader countries (ranking improvement trend)



Better research rank almost always leads to a better world rank

Q5. Which ranking factors matter most?

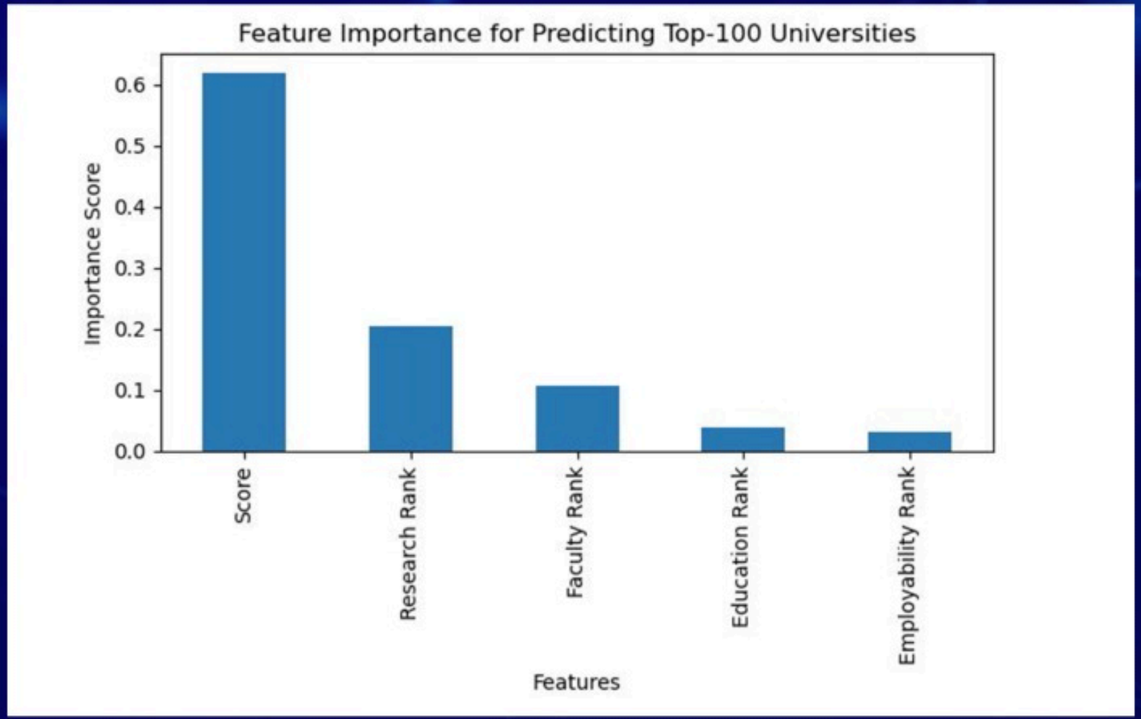


Score is the most important predictor (~0.62 importance)

Research rank also strongly influences prediction

Faculty, Education, and Employability have smaller but relevant r

Q5. Which ranking factors matter most?



Score is the most important predictor (~0.62 importance)

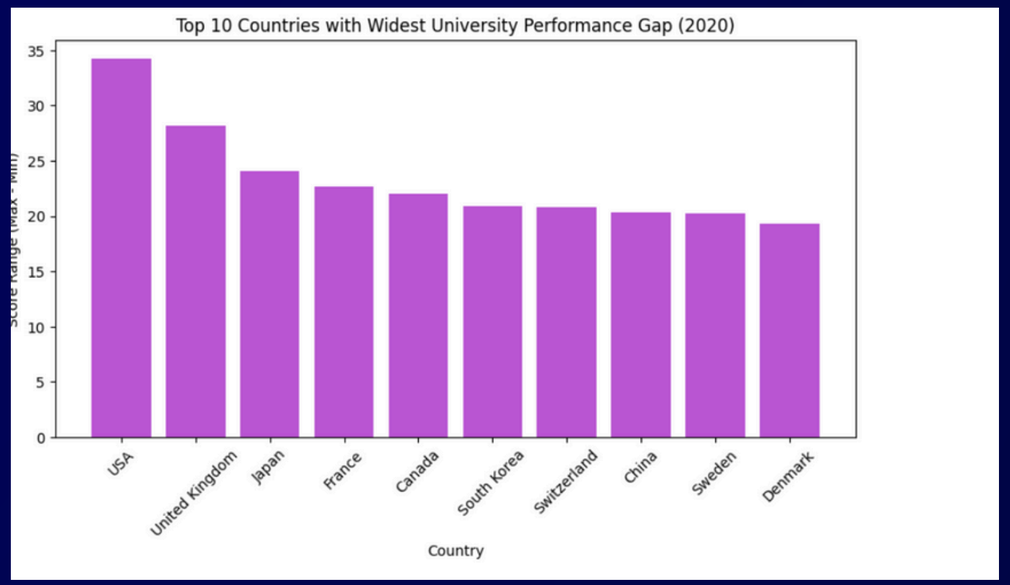
Research rank also strongly influences prediction

Faculty, Education, and Employability have smaller but relevant rol

Q6: Which countries have the widest performance gap between their best and worst universities?

```python
country_score_gap = (
    df.groupby('Country')['Score']
    .agg(['max', 'min'])
    .assign(Gap=lambda x: x['max'] - x['min'])
    .sort_values('Gap', ascending=False)
    .head(10)
)
```

```python
plt.figure(figsize=(10,5))
plt.bar(country_score_gap.index, country_score_gap['Gap'], color='mediumorchid')
plt.title("Top 10 Countries with Widest University Performance Gap (2020)")
plt.xlabel("Country")
plt.ylabel("Score Range (Max - Min)")
plt.xticks(rotation=45)
plt.show()
```



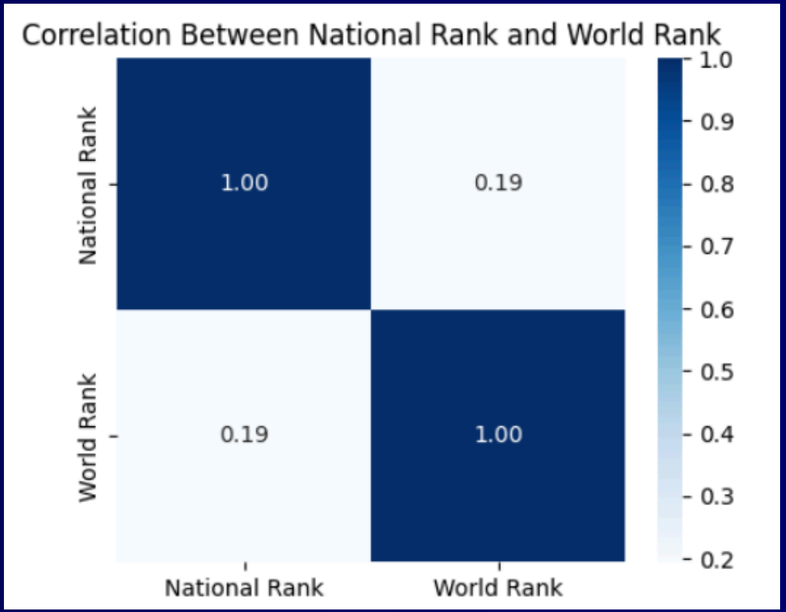
Q7: How does National Rank relate to World Rank within each country?

```python
corr = df[['World Rank', 'National Rank']].astype(float).corr().iloc[0,1]
```

```python
print(f"Correlation between National and World Rank: {corr:.2f}")
```

```
Correlation between National and World Rank: 0.19
```

```python
plt.figure(figsize=(5,4))
sns.heatmap(
    df[['National Rank', 'World Rank']].corr(),
    annot=True,
    cmap='Blues',
    fmt=".2f"
)
plt.title("Correlation Between National Rank and World Rank")
plt.show()
```



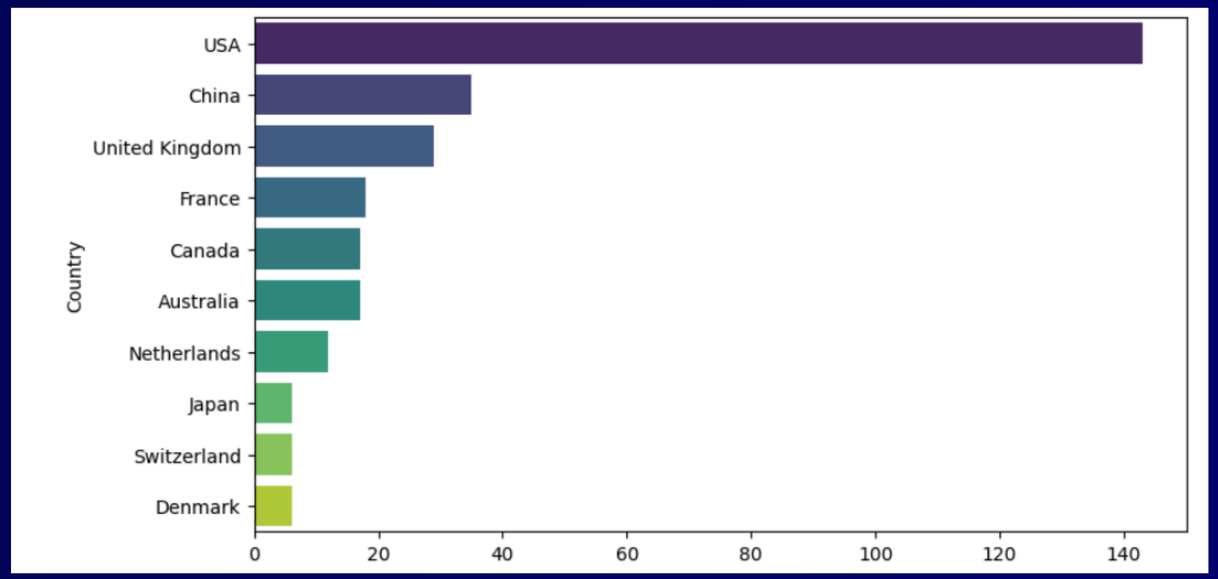
Q8) countries consistently produce top research universities (Top 50 Research Rank)?

```
df['Research Rank'] = pd.to_numeric(df['Research Rank'], errors='coerce')


top_research = df[df['Research Rank'] <= 50]


research_country = top_research['Country'].value_counts().head(10)
```

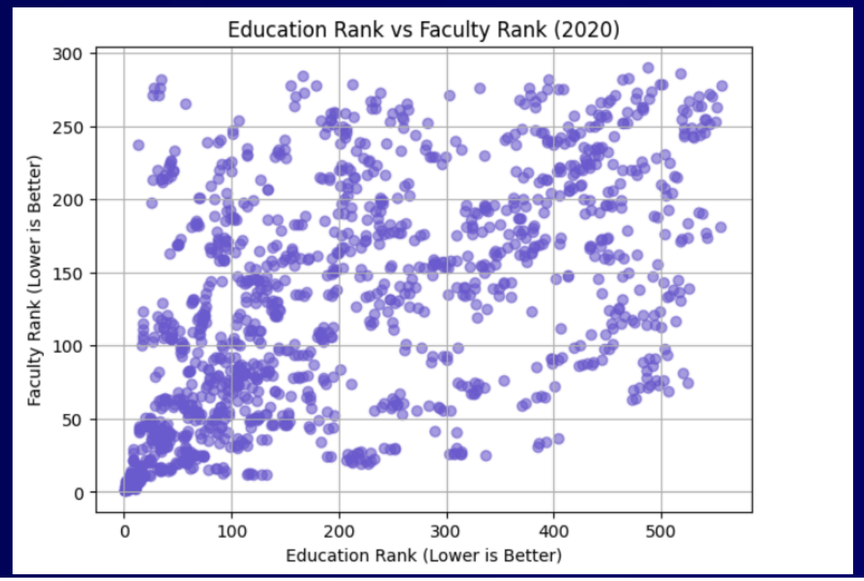
```
plt.figure(figsize=(9,5))
sns.barplot(
    y=research_country.index,
    x=research_country.values,
    hue=research_country.index,  # add hue for color mapping
    dodge=False,
    palette='viridis',
```



Q9: Are countries with higher education quality also strong in faculty rankings?

```python
# Correlation check
corr = edu_faculty[['Education Rank', 'Faculty Rank']].corr().iloc[0,1]
print(f"Correlation between Education Rank and Faculty Rank: {corr:.2f}")
```

```python
# Scatter plot
plt.figure(figsize=(7,5))
plt.scatter(edu_faculty['Education Rank'], edu_faculty['Faculty Rank'],
            color='slateblue', alpha=0.6)
plt.title("Education Rank vs Faculty Rank (2020)")
plt.xlabel("Education Rank (Lower is Better)")
plt.ylabel("Faculty Rank (Lower is Better)")
plt.grid(True)
plt.show()
```




```
Correlation between Education Rank and Faculty Rank: 0.55
```

Q10: Which universities have the best Employability Rank?

```python
top_employability = df[['University', 'Country', 'Year', 'Employability Rank']]
    .sort_values(by='Employability Rank', ascending=True) \
    .head(50) \
    .reset_index(drop=True)
```

```python
print("Top 10 Universities with the Best (Lowest) Employability Rank:")
print(top_employability)
```

```
Top 10 Universities with the Best (Lowest) Employability Rank:
                         University Country  Year  Employability Rank
0                Harvard University     USA  2020                 1.0
1                Harvard University     USA  2022                 1.0
2                Harvard University     USA  2023                 1.0
3                Harvard University     USA  2024                 1.0
4                Harvard University     USA  2025                 1.0
5                Harvard University     USA  2021                 1.0
6                           INSEAD  France  2021                 2.0
7                           INSEAD  France  2022                 2.0
8                           INSEAD  France  2024                 2.0
9  École nationale d'administration  France  2020                 2.0
```

CONCLUSION

- Collected CWUR university ranking data (2020–2025) using web scraping, then cleaned and standardized it to ensure accuracy and consistency.
- Performed descriptive statistics and data quality checks to understand dataset size, structure, and reliability before analysis.
- Used visualizations such as line and trend charts to clearly display ranking patterns and regional performance differences.
- Found that Asian universities showed the highest average improvement, and research score had the strongest correlation with global rank.
- Concluded that consistent academic growth and strong research performance are the main drivers of improved university rankings.