Project Report

CS661: BIG DATA VISUAL ANALYTICS

2024-2025 Semester II

THE World University Ranking Analysis 2016-2025

Team Members

Deepak Soni (241110018)
deepaksoni24@iitk.ac.in

Divyanshu (241110023)
divyanshu24@iitk.ac.in

Hritik Chauhan (241110030)
hritikc24@iitk.ac.in

Khuswant Kaswan (241110035)
khuswantk24@iitk.ac.in

Rajan Kumar (241110087)
rajank24@iitk.ac.in

Richik Majumder (241040068)
richik24@iitk.ac.in

Rishit Kumar (241110056)
rishitk24@iitk.ac.in

Senthil Ganesh P (241110089)
senthil24@iitk.ac.in

Contents

List of Figures

1 Introduction

University rankings serve as a crucial benchmark for evaluating academic performance, assisting students, policymakers, and researchers in assessing the quality of institutions worldwide. Over the past decade (2016–2025), higher education has undergone substantial transformations, driven by technological advancements, globalization, and policy shifts.

This project aims to analyze and visualize trends in university rankings from 2016 to 2025, focusing on key factors such as teaching quality, research impact, student diversity, industry collaboration, and international outlook. To support this objective, we propose the development of an interactive web-based visual analytics system utilizing Python-based tools, including Matplotlib, Seaborn, Plotly, Dash, and Streamlit.

Our visual analytics journey reveals several notable trends. While the United States and Europe continue to dominate global rankings, Asian institutions, particularly from countries like China and Singapore, are significantly improving, especially in terms of research quality and international appeal. This shift highlights the growing competitiveness of higher education across different regions.

Gender diversity in higher education also demonstrates progressive improvement. Female student participation has consistently risen across many institutions, indicating the success of global efforts and policies aimed at promoting inclusivity. However, certain regions still lag behind, emphasizing the need for more targeted diversity strategies.

Additionally, resource allocation emerges as a critical performance driver. Institutions exhibiting lower student-to-staff ratios tend to demonstrate superior teaching quality. This observation suggests that policymakers should strategically invest in faculty development to enhance educational outcomes.

1.1 Dataset Description

The dataset utilized in this project is sourced from Times Higher Education(THE) and is also available on Kaggle. It provides detailed information on university rankings from 2016 to 2025, covering multiple dimensions of academic and institutional performance. Table 1 summarizes the main attributes contained in the dataset.

Table 1: Dataset Attributes Description

SNo	Attributes	Descriptive Information
1	University Rank & Name	Global ranking and institution name
2	Country	Country where the university is located
3	Student Population	Total number of enrolled students
4	Student-to-Staff Ratio	Availability of faculty per student
5	International Students	Percentage of international student enrollment
6	Female-to-Male Ratio	Gender distribution among students
7	Overall Score	Composite score determining overall rank
8	Teaching Score	Rating based on teaching quality
9	Research Environment Score	Assessment of research facilities and output
10	Research Quality Score	Research impact measured through citations
11	Industry Impact Score	University collaborations with industries
12	International Outlook Score	Diversity of faculty and student body
13	Year (2016–2025)	Academic year of the ranking data

1.2 Problem Solved by Visualization

Through our visual analytics system, we aim to address several critical analytical challenges:

- Identifying global and regional trends in higher education performance.
- Analyzing gender diversity and international student enrollment patterns.
- Comparing university performance across multiple criteria over time.

- Understanding the impact of resource allocation (student-to-staff ratios) on teaching quality.

2 Tasks

2.1 Top 20 Universities Bar Chart

In this task, we aimed to visualize the top 20 universities based on their rank for any selected country or globally, depending on user filter settings. The task's objective was to create an intuitive view where users could instantly identify the best-ranked institutions. .

Proposed Solution: To achieve this, we first implemented sidebar controls allowing users to filter the dataset dynamically by Year, Country, Rank range, and Overall Score range. After applying these filters, we extracted the top 20 universities using the Pandas .nsmallest(20, 'Rank') function, which efficiently selects the lowest 20 ranks — corresponding to the best universities. For visualization, we employed Plotly Express's px.bar function to generate a **horizontal bar chart**, where the X-axis represented the rank and the Y-axis displayed the university name. The rank axis was reversed (autorange='reversed') so that the 1 university appeared at the top, ensuring natural reading order from best to less-best downward. This method leverages an easy-to-understand and space-efficient design, helping users immediately grasp which universities dominate in the selected subset. Using a horizontal bar also improves label readability for university names compared to vertical bars. By structuring it this way, the user gets a clear, dynamic, and visually clean overview of the leading academic performers in any selected region or globally.

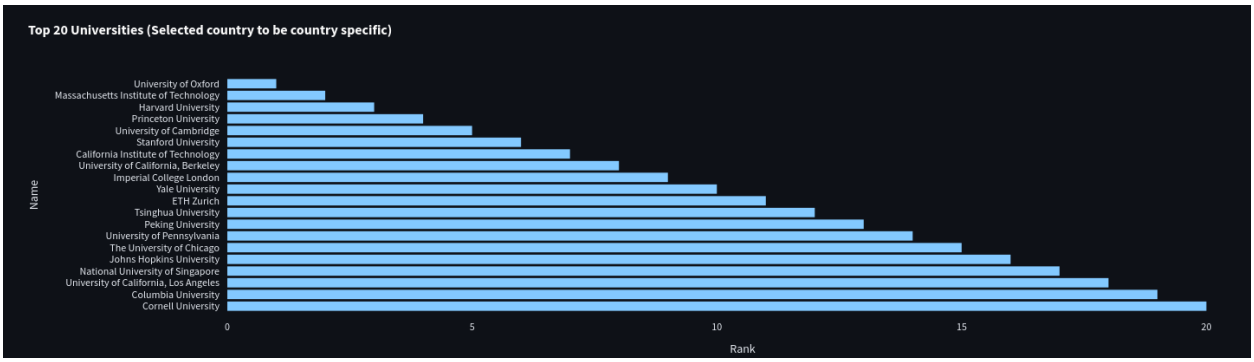


Figure 1: Top 20 Universities by rank

2.2 Bottom 10 Universities Bar Chart

In this task, we focused on identifying and visualizing the bottom 10 universities based on their rank, again filtered dynamically according to the user's selections for year, country, and scoring ranges. The main objective was to complement the Top 20 Universities plot by providing a contrasting view, allowing stakeholders to see which institutions are currently underperforming relative to others in the dataset.

Proposed Solution: After applying the necessary sidebar filters, we selected the bottom 10 universities using the Pandas .nlargest(10, 'Rank') function, which extracts universities with the highest rank numbers — i.e., the least favorable positions in the rankings. We visualized this data using a **horizontal bar plot** through Plotly Express's px.bar, where the X-axis represented Overall Score and the Y-axis displayed the university names. The bars were oriented horizontally to maximize label readability and to allow a clear, side-by-side comparison of performance. The university names were arranged top-to-bottom, while the scores increased from left to right. This method was chosen to highlight the lower end of the ranking spectrum, enabling users to spot institutions struggling with academic performance. It helps in full-spectrum analysis because both high performers (Top 20) and low performers (Bottom 10) are visible within the dashboard in an intuitive, consistent manner.

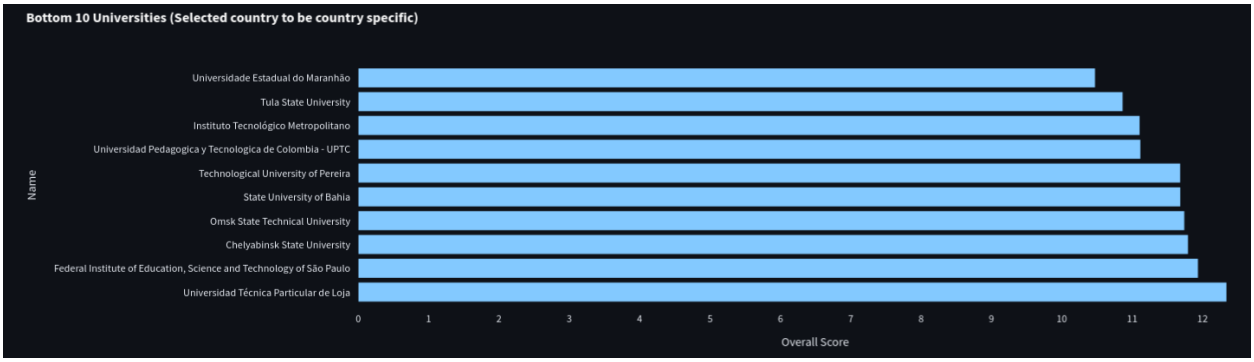


Figure 2: Bottom 10 Universities by rank

2.3 Rank Trajectories (Line Plot)

This task aimed to illustrate the rank trajectory of selected universities across the years 2016 to 2025. The goal was to give users a time-series perspective on how a university's standing has evolved rather than showing a static snapshot for a single year.

Proposed Solution: To achieve this, users were provided with a multiselect dropdown (with top 20 universities as default values) allowing them to choose one or multiple universities from the dataset. We then filtered the data accordingly and used Plotly Express's px.line function to create a **line plot** where the X-axis represented the Year and the Y-axis represented the University Rank. Importantly, we applied autorange='reversed' to the Y-axis so that higher rankings (lower rank numbers) appeared higher up in the plot — this natural orientation makes it easy to understand whether a university's performance is improving (going up) or deteriorating (going down) over time. The line plot used markers to emphasize year-to-year points distinctly. This time-series analysis was crucial because rankings often fluctuate; seeing the pattern over multiple years helps users identify trends, stability, or sudden changes in performance, offering a dynamic, longitudinal perspective instead of a flat ranking.

Figure 3: Rank Trajectories using Line Plot

2.4 Animated Top 20 Universities Over Years

This task extended the static Top 20 analysis into an animated visual representation that evolves across time, year by year, from 2016 to 2025. The objective was to create a lively, storytelling animation that dynamically showcases how the competition among universities changes each year.

Proposed Solution: After sorting the data by year and selecting the top 20 universities per year using grouping and head(20), we constructed an **animated horizontal bar plot** using Plotly Express's px.bar with the parameter animation_frame='Year'. Each frame in the animation represents a different year, allowing the viewer to watch universities rise and fall over time as the bars move and reshuffle. The Y-axis displayed university names, and the X-axis displayed their Overall Scores, scaled from 0 to 100 for clarity. The animation feature, combined with smooth transitions, helps highlight year-over-year competition dynamics — making it clear which institutions are consistently strong and which ones surge or fall during specific periods. This approach turns static ranking data into a compelling, movie-like experience that greatly improves user engagement and insight discovery.

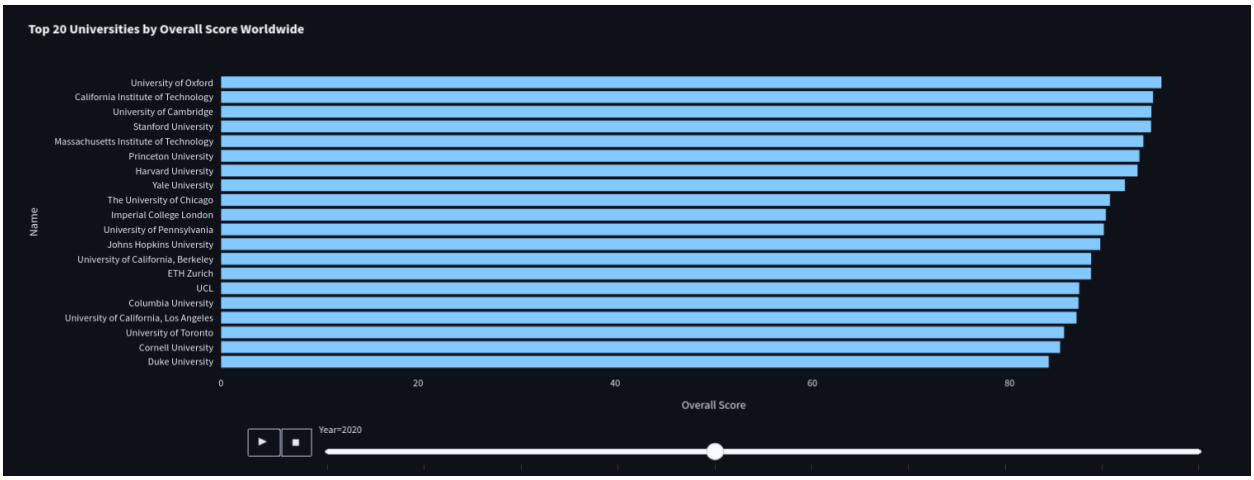


Figure 4: Animated Top 20 Universities Over Years

2.5 Continent Wise Score (Sunburst and Treemap Plots)

We wanted to summarize educational performance hierarchically across continents, countries, and top N universities using a sunburst chart as well as Treemap. The goal was to present multilevel aggregated performance in a visually intuitive, compact form.

Proposed Solution: We aggregated the data first at the Continent level, then at the Country level, and finally at the Top N Universities level within each country using Overall Scores. We utilized Plotly Express's px.sunburst, a radial hierarchical visualization where each level represents a deeper layer: the center represents continents, the middle ring represents countries, and the outer ring represents universities. The values plotted were Overall Scores aggregated appropriately at each level. Sunburst charts were chosen because they efficiently display hierarchical relationships while preserving the relative contribution of each segment via area size. This plot allows users to quickly grasp which continents, countries, and institutions dominate academically, offering both macro and micro perspectives within a single visual. Similary the px.treemap was used.

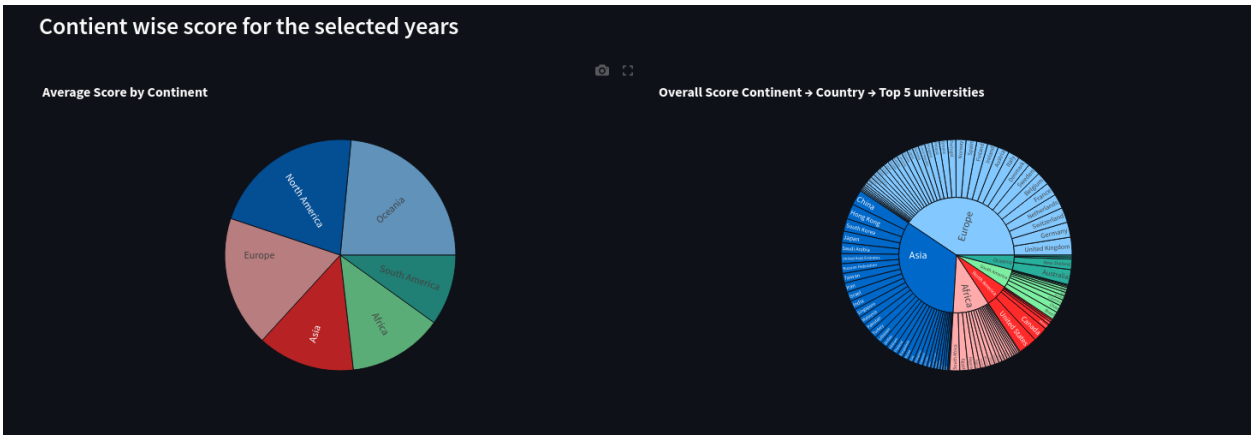


Figure 5: Continent Wise Score using Sunburst plot

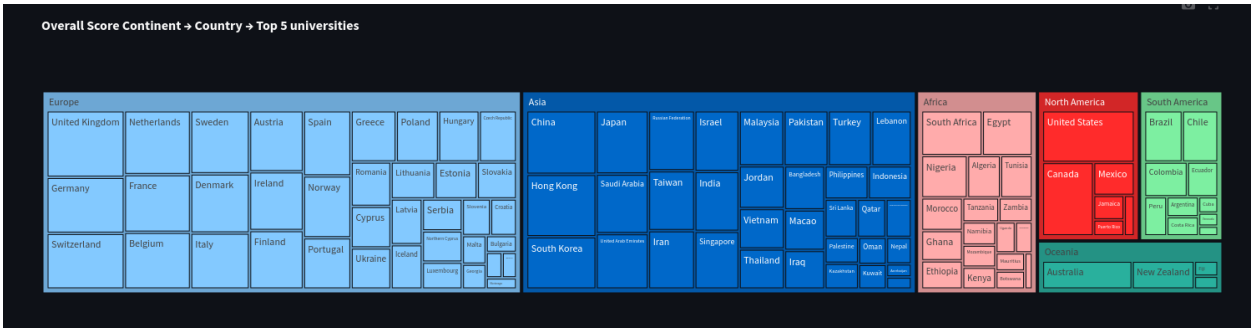


Figure 6: Continent Wise Score using Treemap plot

2.6 Top 10 Countries by University Count (Bar Chart)

For this task, our goal was to highlight countries contributing the largest number of universities to the dataset.

Proposed Solution: We computed the total number of universities per country using Pandas' .value_counts() function and selected the Top 10. We plotted the results using a horizontal bar chart via px.bar, where the X-axis represented the number of universities and the Y-axis listed countries. The Y-axis was ordered top-to-bottom according to the number of universities. This plot serves as a measure of educational breadth — countries with a high number of participating universities show broader higher-education ecosystems, while those with fewer entries may have fewer internationally recognized institutions. It provides critical context when interpreting aggregate scores and rankings at the country level.

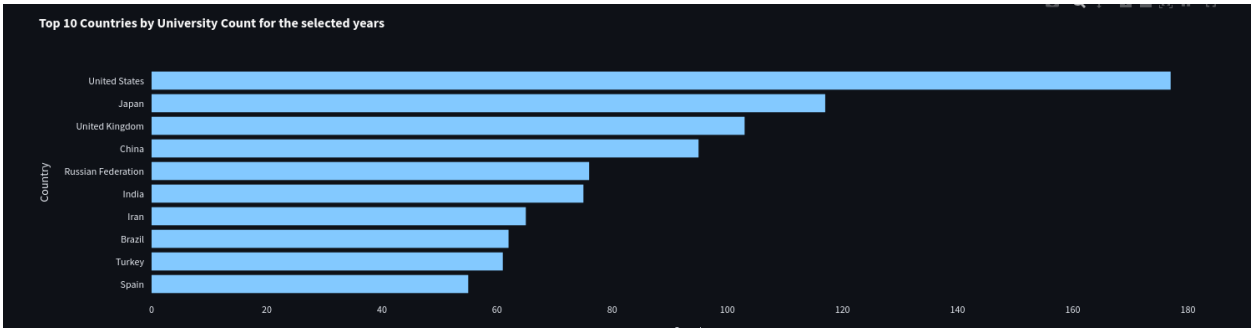


Figure 7: Top 10 Countries by University Count using Bar Chart

2.7 Average Scores by Country (Bar and Scatter Plots)

We focused on measuring the academic quality of countries by calculating the mean Overall Score of universities within each country.

Proposed Solution: After grouping the data by Country and averaging their Overall Scores, we plotted two visuals: a horizontal bar plot and a scatter plot. Both were created using Plotly Express. The bar chart made it easier to quickly see rankings, while the scatter plot offered a more distributional view of scores. The reasoning behind using two types of plots was to address different cognitive styles: some users grasp sorted rankings faster with bars; others prefer seeing dispersion or clustering visually. This analysis is crucial for policymakers and academic analysts to understand not just who has the most universities, but which countries have the highest average quality.

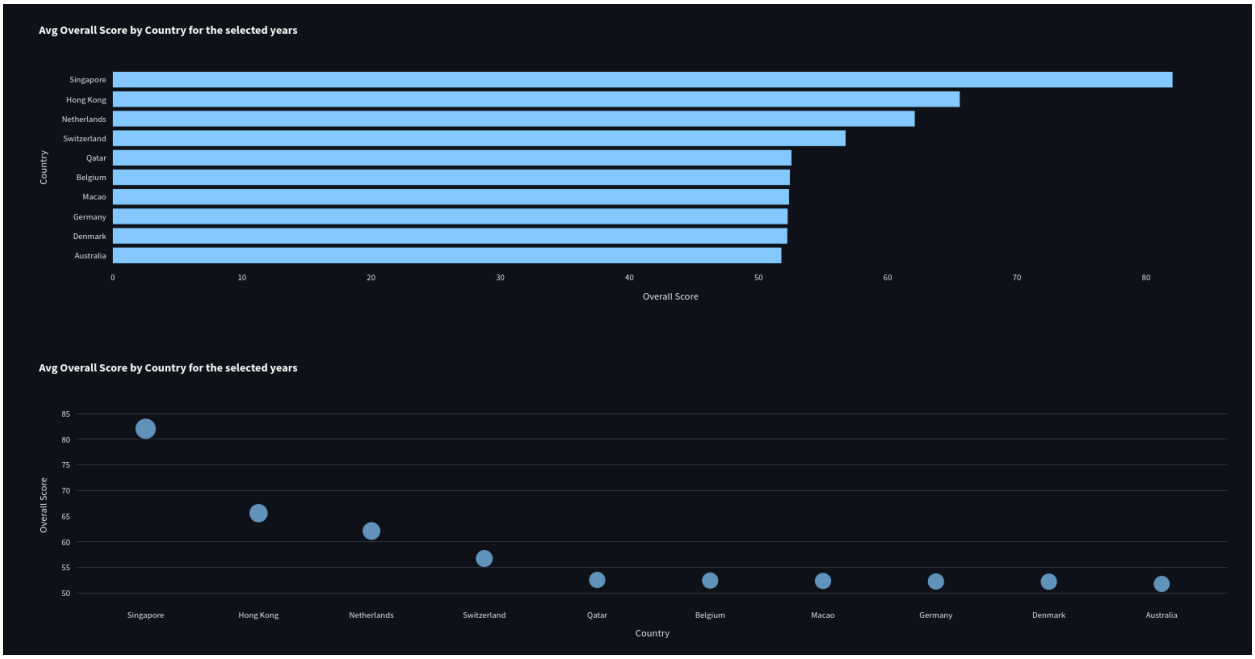


Figure 8: Average overall scores by Country using Bar and Scatter Plots

2.8 Animated Choropleth Maps (World Map)

In this task, we aimed to animate global educational trends across multiple dimensions using choropleth maps. The task was broken down into five animated maps showing: the count of universities per country, the average percentage of international students, the average percentage of female students, the overall student population, and the distribution based on Overall Score and Industry Impact.

Proposed Solution: To achieve this, we first aggregated the necessary metrics year-wise and country-wise using group-by operations. Then, we used Plotly Express's px.choropleth with animation_frame='Year' to generate dynamic world maps that update over time from 2016 to 2025. The natural earth projection was chosen for clarity and familiarity to the global audience. In each map, countries were shaded based on the metric being analyzed, with darker or lighter colors indicating higher or lower values. This method was selected because choropleth maps are one of the most effective ways to depict geographically-distributed data, and adding animation provides a temporal layer of insight. These maps allow users to visually explore where educational growth is happening, how internationalization is spreading, how gender diversity is evolving, and how overall scores and industry engagement are changing across the world — without needing to read complex tables or charts.

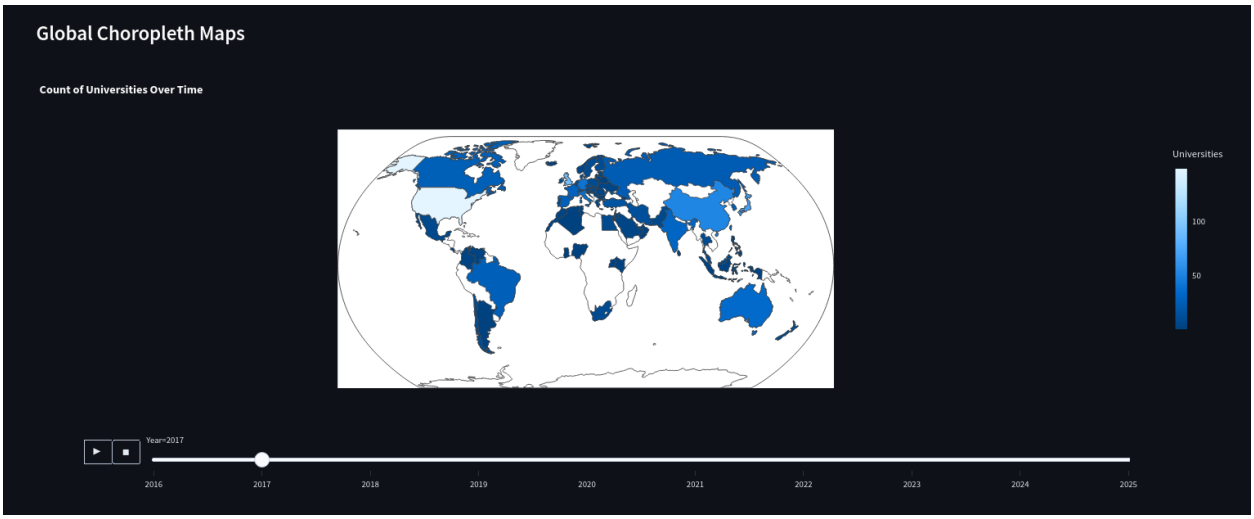


Figure 9: Animated Choropleth Maps (World Map)

2.9 Gender Balance Over Time (Line Plot)

Our objective for this task was to analyze the global shift in gender balance among university students over the 2016-2025 period. Specifically, we tracked the evolution of average Female % and Male % across all universities year-by-year.

Proposed Solution: To solve this, we aggregated the dataset using groupby('Year') and calculated the mean values for Female % and Male % separately. We then plotted these trends using Plotly Express's px.line, with one line representing the female percentage and another for male percentage, both plotted against the year. The Y-axis range was specifically controlled (from 40% to 52%) to focus the viewer's attention on subtle but important shifts. This visualization helps illuminate progress made globally towards gender equity in higher education. Even minor percentage changes are significant when scaled to millions of students worldwide. By using a clean line plot, trends become immediately obvious to the reader without overwhelming them with numbers.

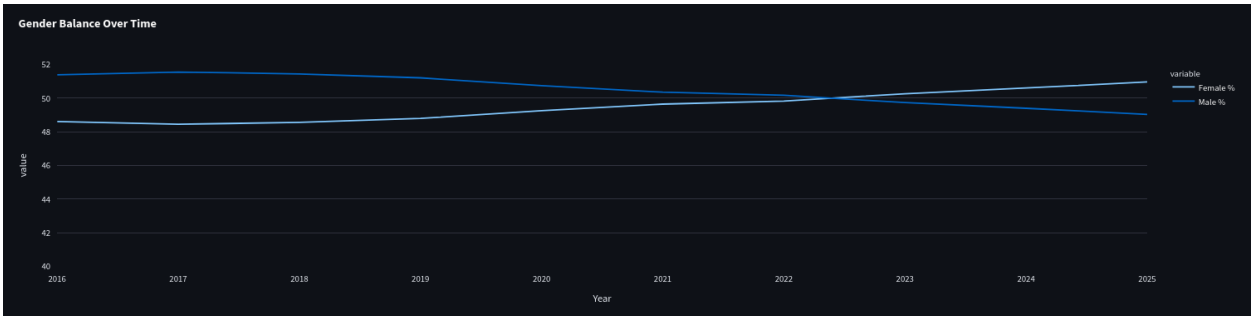


Figure 10: Gender Balance Over Time using Line Plot

2.10 Evolution of International Students (Line Plot)

In the tenth task, the aim was to identify which countries have been most successful in attracting international students over time.

Proposed Solution: We first determined the Top 10 countries based on cumulative international student totals across all years. We then filtered the dataset to only include these countries and calculated year-wise totals of international students hosted. A multi-line plot was created using px.line, where each

country had its own line tracing international student numbers across years. Markers were added to each year's data point for better readability. This approach allows for direct visual comparison of trends — users can quickly see whether a country like the United States, Australia, or China has consistently grown, plateaued, or declined as a preferred destination for international education. Such an analysis is crucial for policymakers aiming to understand their country's competitiveness in the global academic market.

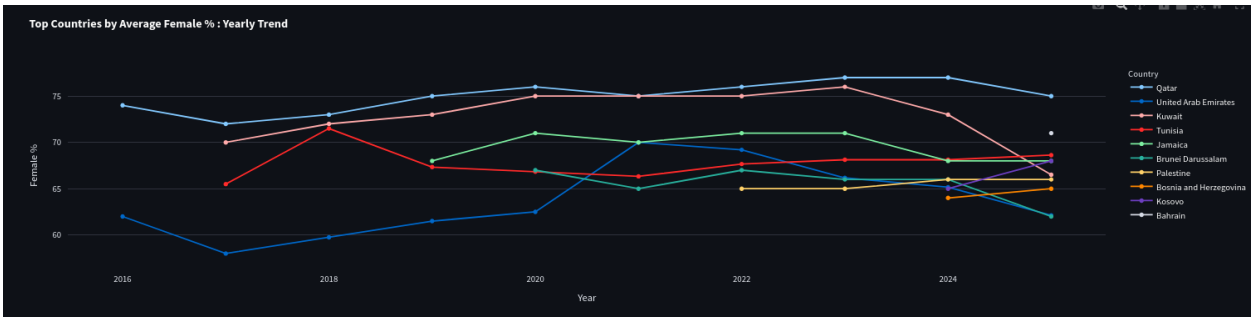


Figure 11: Evolution of International Students using Line Plot

2.11 Scatter Plots for Research vs Industry Impact Metrics

For this task, we created scatter plots to explore relationships between core academic metrics, notably Research Quality vs Overall Score and Industry Impact vs Research Environment. The goal was to determine how research excellence and industry collaboration relate to overall institutional performance.

Proposed Solution: We used px.scatter from Plotly Express, plotting each university as a point in a 2D plane. For some plots, we included a trendline (trendline='ols') to show regression patterns more clearly. The Research Quality vs Overall Score scatter shows whether high research standards directly correlate with top rankings (which they generally do). The Industry Impact vs Research Environment plot was chosen to see if practical engagement with industry complements a strong academic environment. Scatter plots were ideal here because they reveal correlations, clusters, outliers, and overall patterns at a glance — far more intuitively than a table or a simple statistical correlation coefficient could.

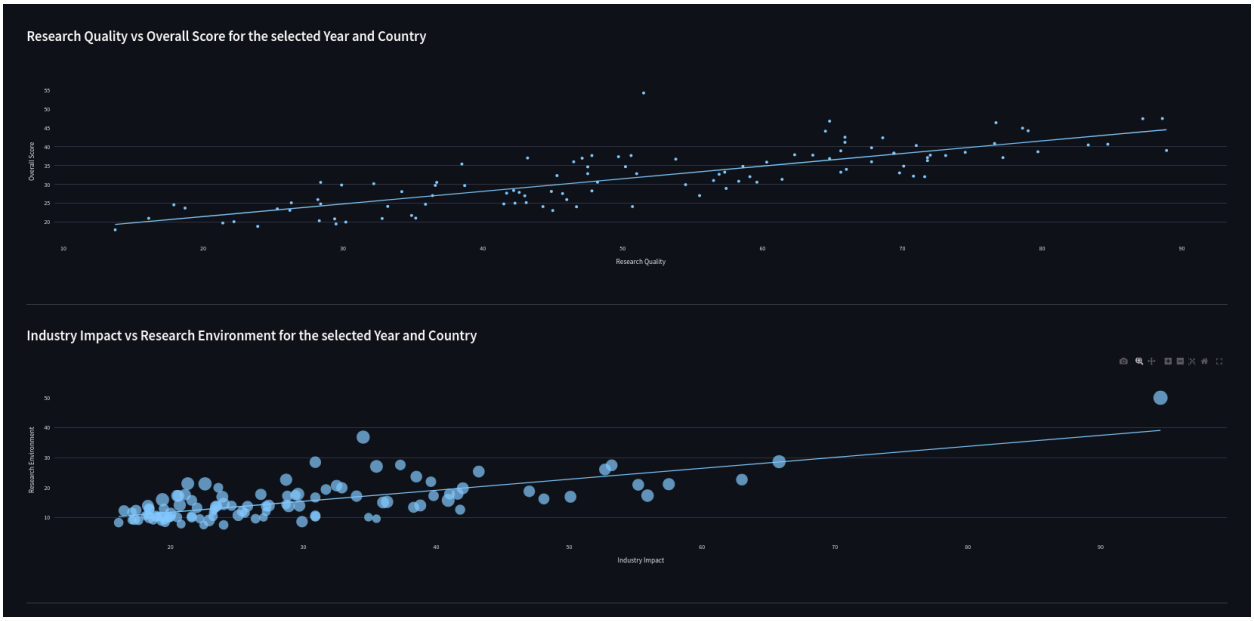


Figure 12: Scatter Plots for Research vs Industry Impact Metrics

2.12 Box Plots for Distribution of Core metrics

In this task, the goal is to visualize the distribution and variability of key university performance metrics to enable deeper insights into how different attributes (like Overall Score, Teaching, Research Quality, etc.) vary across institutions, countries, and years.

Proposed Solution: The proposed solution uses a combination of box plots and violin plots built with Plotly inside a Streamlit app. First, the selected variables are reshaped using the melt function to organize the data into a long format, allowing for an efficient comparison across different metrics. A box plot is generated to show the statistical distribution (median, quartiles, outliers) of core metrics, providing a straightforward overview. To offer a more detailed and interactive analysis, users can dynamically select specific metrics through a multiselect widget, and a corresponding violin plot is displayed, combining box plot elements with the full distribution shape and individual data points. This approach is used because it provides both summary statistics and distribution shape at a glance, allowing users to quickly identify patterns, outliers, and differences across metrics, which is essential for understanding the underlying diversity and trends in university performance data.

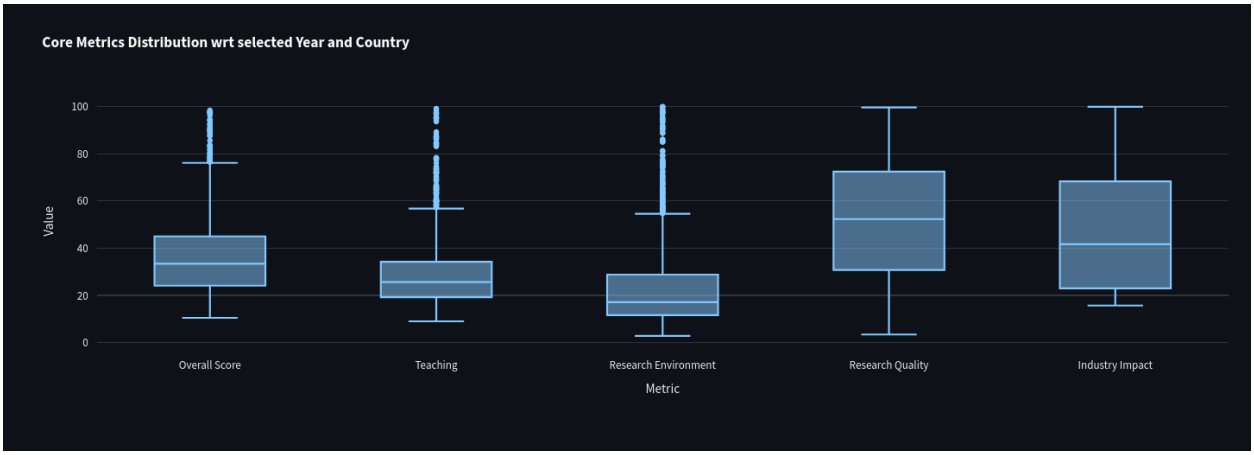


Figure 13: Box Plots for Distribution of Core metrics

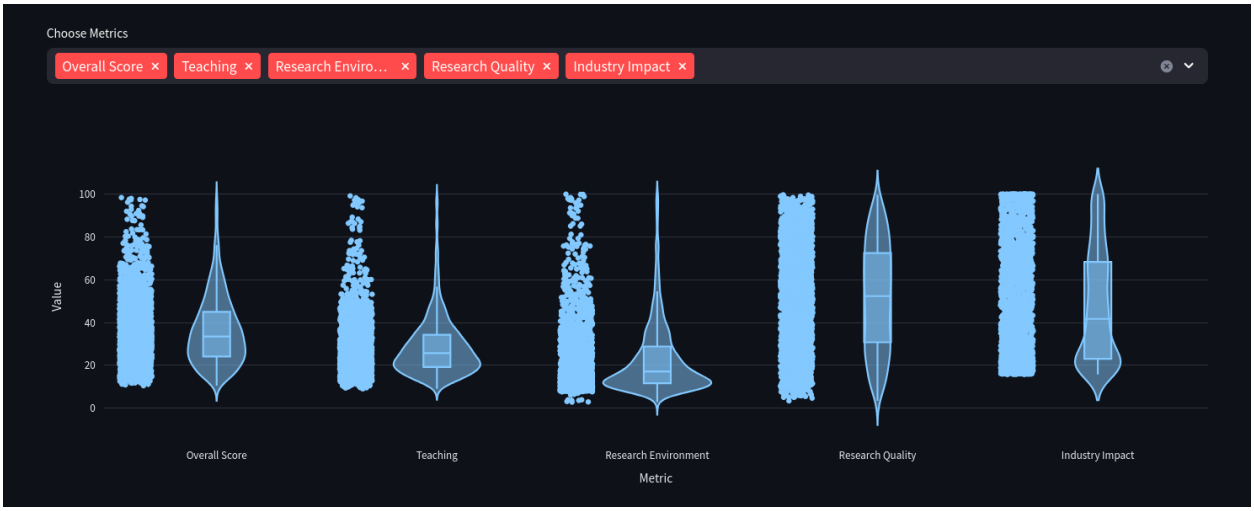


Figure 14: Box and Violin Plots for Distribution of selected metrics

2.13 Scatter Matrix and Correlation Heatmap

In this task, we aimed to provide a multivariate view of how different selected metrics interact with one another.

Proposed Solution: We first created a scatter matrix using Plotly Express's px.scatter_matrix, where each cell shows a scatter plot between two different metrics. Simultaneously, we computed a Pearson correlation matrix between all selected numeric columns, visualized using px.imshow as a heatmap with color gradients and numeric values. The scatter matrix helps users visually identify patterns like positive or negative linear relationships across all combinations of metrics. The correlation heatmap quantifies the strength of these relationships numerically, showing for instance that Overall Score strongly correlates with Research Quality but less so with Students to Staff Ratio. These combined visuals empower the reader to quickly understand internal structures and redundancies within the dataset, helping drive feature selection, policy focus, or academic investment strategies.

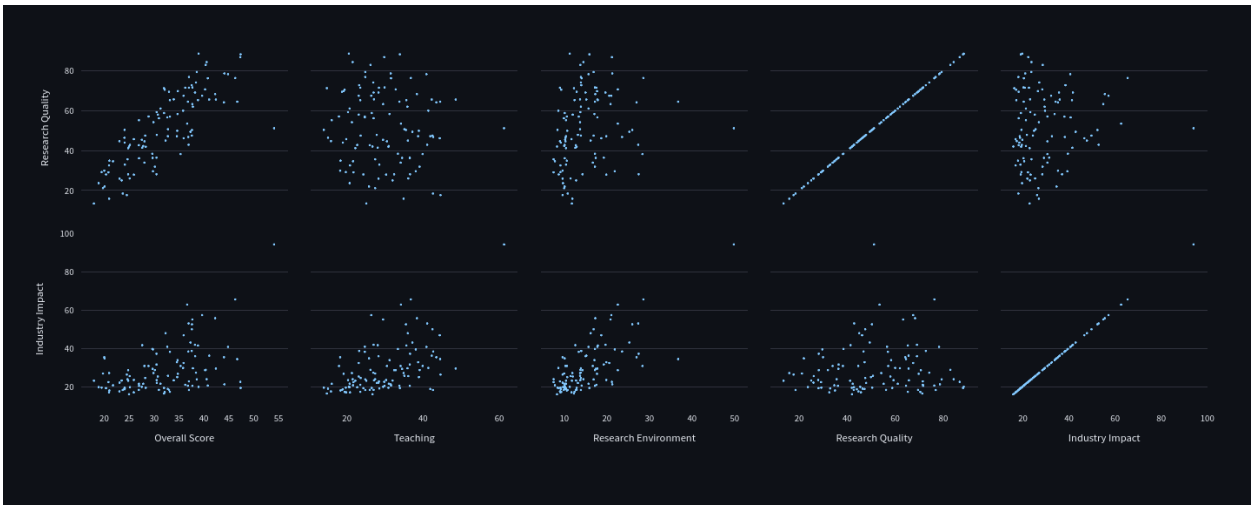


Figure 15: Scatter Matrix for different metrics

2.14 PCA-based KMeans Clustering

The goal of this task was to cluster universities into groups based on their performance across selected metrics.

Proposed Solution: We used Principal Component Analysis (PCA) to reduce high-dimensional feature space into two or three principal components, capturing the majority of variance in the data. PCA transformation was performed after standardizing features via StandardScaler. We then ran KMeans clustering on the PCA-transformed data, letting users control the number of clusters (k) through a sidebar slider. To visualize the clusters, we plotted:

- 2D PCA scatter colored by cluster ID,
- 3D PCA scatter (for deeper insights),
- Elbow plot of inertia vs. number of clusters to help users choose optimal k.

This method helps users discover natural groupings among universities — for example, grouping research-intensive universities separately from teaching-focused or industry-strong institutions. Clustering allows segmentation for strategic planning, benchmarking, or research collaboration targeting.

Through PCA Loadings, we discover that PC1 acts as a 'University Strength' axis blending teaching, research, and industry outreach, while PC2 distinguishes universities highly focused on research impact. This

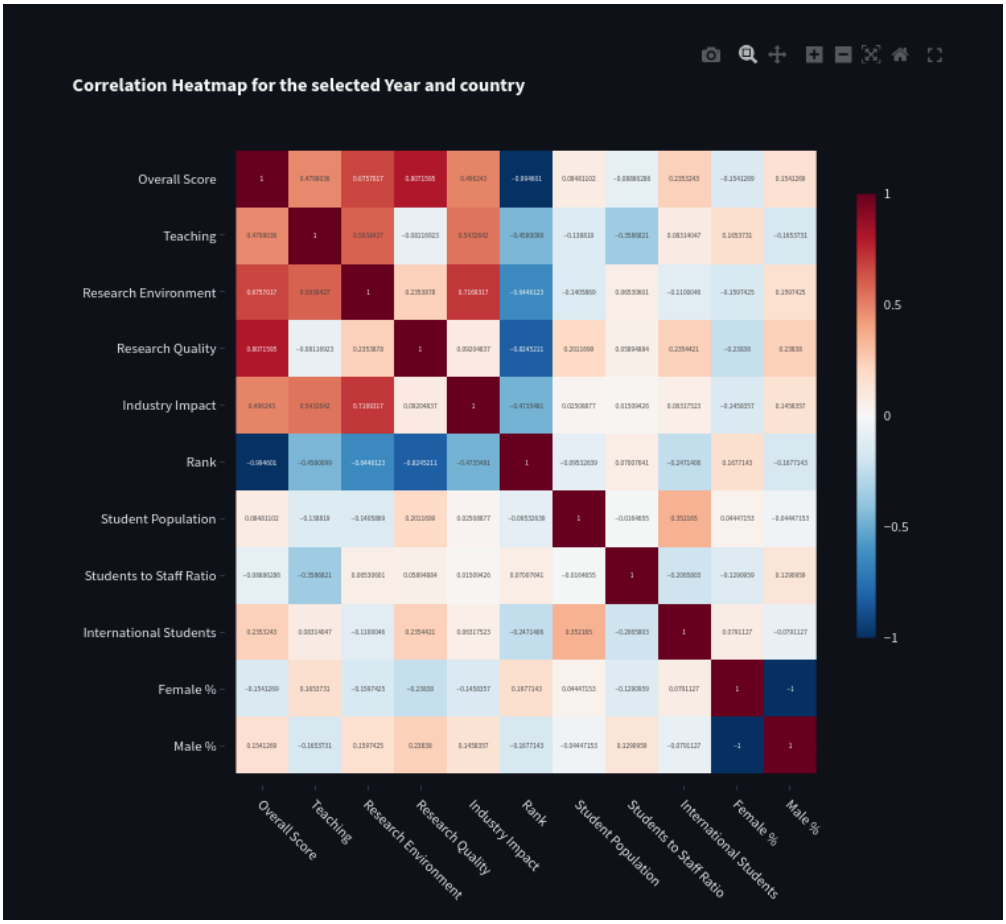


Figure 16: Correlation Heatmap for the selected year and country

two-axis structure explains over 90% of the variance, validating the use of PCA scatter plots for meaningful clustering and interpretation of global university profiles.

Our clustering analysis notably reveals six distinct university profiles: Cluster 5 represents globally elite institutions excelling across all metrics. Cluster 1 universities demonstrate powerful research and industry engagement but moderate teaching. Meanwhile, clusters 0, 2, and 3 consist of emerging universities striving for global prominence but facing challenges in teaching, research, and internationalization.

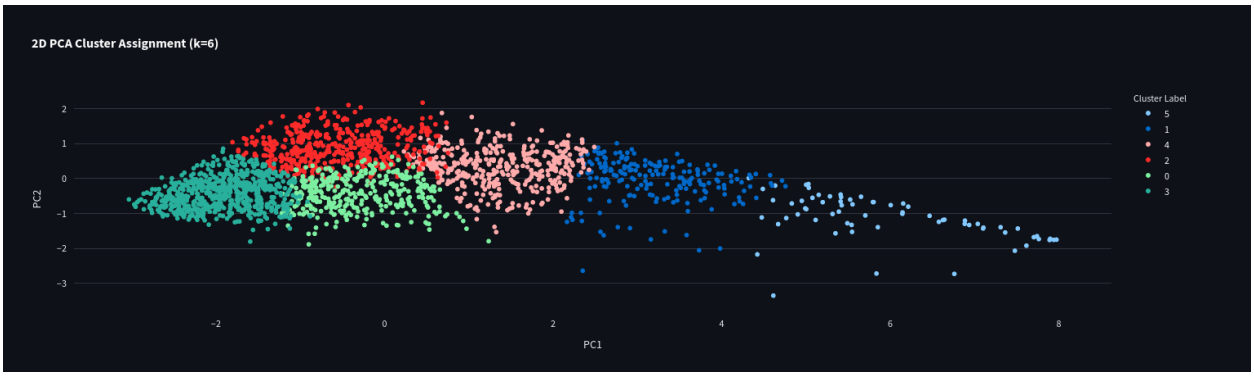


Figure 17: 2D PCA Cluster Assignment with k=6

Cluster Profiles: Cluster Centers Metrics

Cluster	Overall Score	Teaching	Research Quality	Industry Impact	International Students	Student Population	Students to Staff Ratio	Female %
0	31.42	28.59	40.1	54.45	7.86	17866.22	16.05	48.52
1	61	46.29	81.85	86.91	20.56	25876.34	18.73	49.71
2	35.15	22.1	63.91	29.12	9.52	21203.34	19.77	51.6
3	20.25	20.11	24.33	24.97	5.09	23339.58	18.78	51.48
4	47.09	32.36	71	71.52	15.54	20313.74	18.74	52.73
5	82.22	75.82	89.75	91.08	28.22	29741	13.49	48.24

Figure 18: Cluster profile

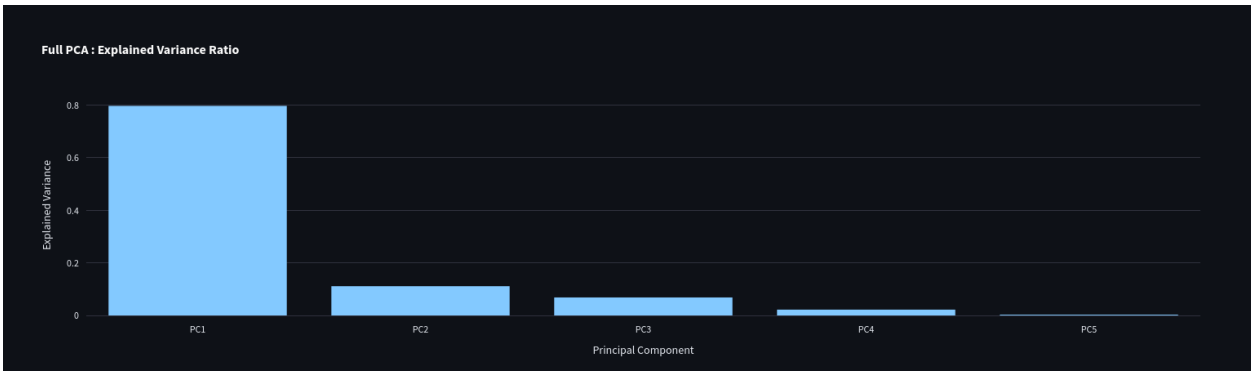


Figure 19: Bar chart for variances captured by different PCAs

2.15 UMAP + HDBSCAN Clustering

This task tackled the clustering problem using a non-linear manifold learning method (UMAP) combined with density-based clustering (HDBSCAN).

Proposed Solution: We first projected the original selected metrics into a 2D latent space using UMAP (umap.UMAP(random_state=42)). Unlike PCA, UMAP preserves local neighbor relationships even when data lies on a complex manifold. We then applied HDBSCAN, a hierarchical clustering algorithm that automatically detects clusters without predefining the number of clusters (unlike KMeans). HDBSCAN labels outliers as -1, highlighting institutions that do not belong to any dense group. We visualized clusters on the 2D UMAP plot using px.scatter, coloring points by cluster ID. UMAP + HDBSCAN is ideal for this dataset because educational institutions often exhibit non-linear relationships, and forcing linear projections would mask those structures. This technique uncovers hidden clusters like elite research universities, rising regional players, or industry-driven universities, offering profound insights into academic ecosystems.

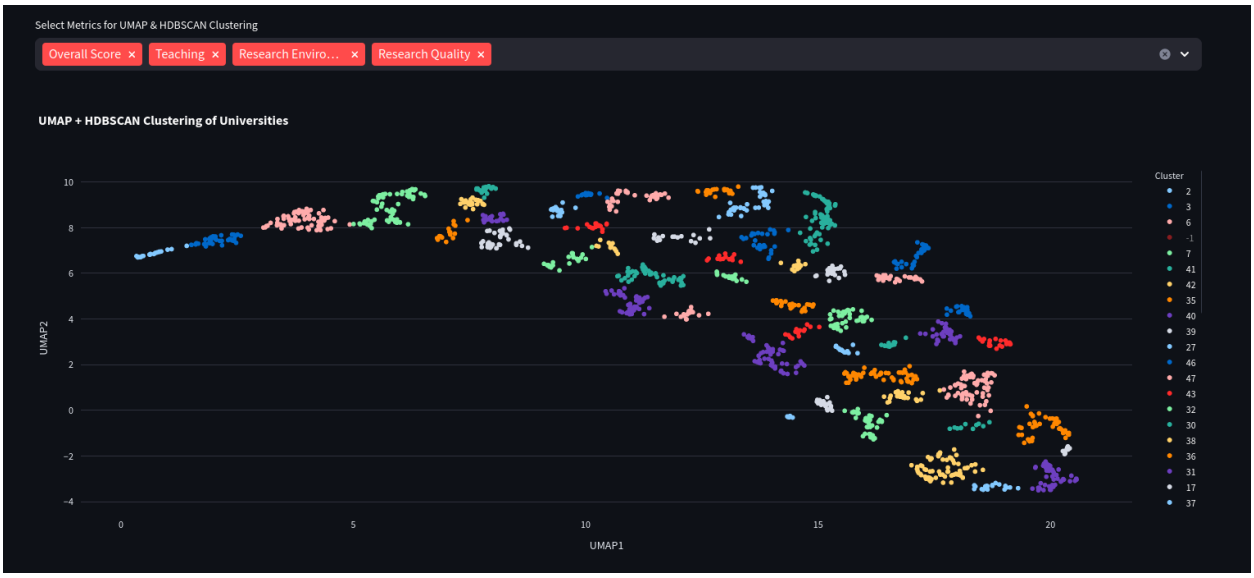


Figure 20: UMAP + HDBSCAN Clustering

2.16 University Similarity Network

In this task, we built an advanced University Similarity Network where users select one university and one country, and the graph shows edges only if the university is highly similar to universities from the selected country..

Proposed Solution: Critically, we used the similarity calculation to be based on raw selected metrics (Teaching, Research Environment, Research Quality, Industry Impact, Overall Score), rather than UMAP-compressed values. Cosine similarity was computed using cosine_similarity after standardizing metrics with StandardScaler. If similarity exceeded a user-set threshold, an edge was drawn in the network graph using NetworkX's spring_layout for positioning. Each node represented a university, colored by Overall Score. This method allows users to precisely explore cross-country academic neighbors based on true academic profiles, ensuring full analytical rigor while maintaining visual clarity and engagement.

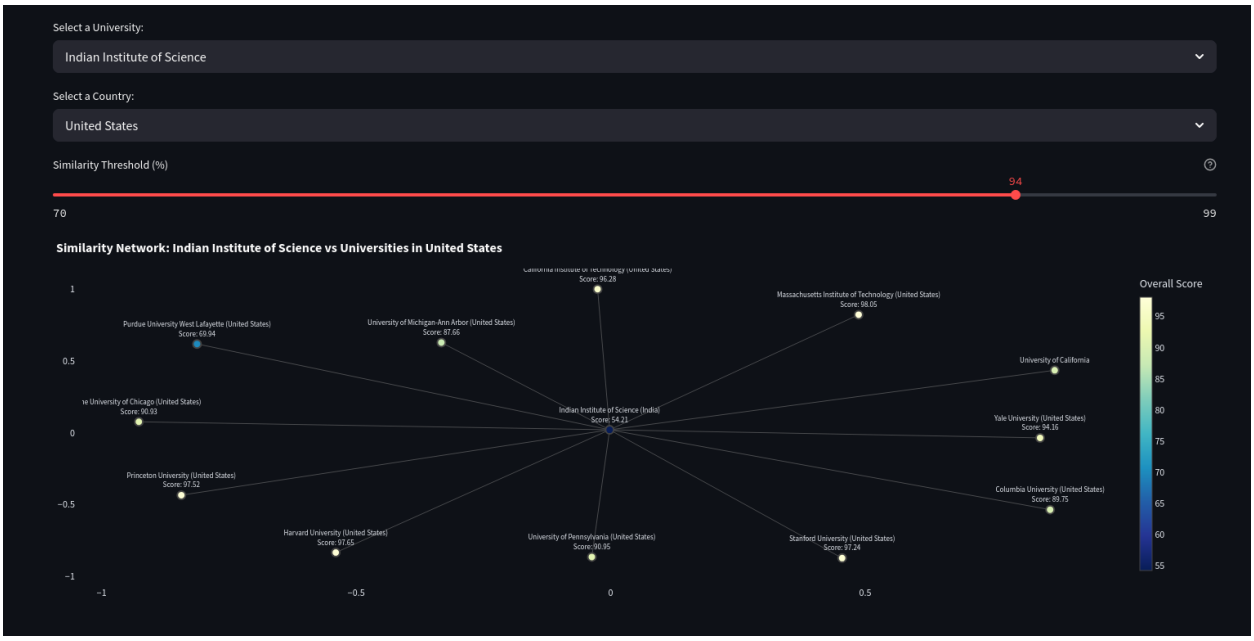


Figure 21: University similarity using Network Graph

2.17 University Comparer Tab (Radar and Line Charts)

For this task, we enabled direct head-to-head comparison between any two selected universities across all major performance metrics.

Proposed Solution: For the latest available year for each university, we built a Radar Chart (also known as Spider Chart) using Plotly Graph Objects to plot all selected metrics around a circle for each university. Additionally, we provided time-series line charts comparing the universities' Rank, Overall Score, Student Population, and Diversity indicators across years. The radar chart offers an intuitive 360° performance snapshot, while the trend lines highlight the evolution and divergence of performance over time. This module is essential for students, researchers, and policymakers conducting institutional benchmarking or competitive analysis

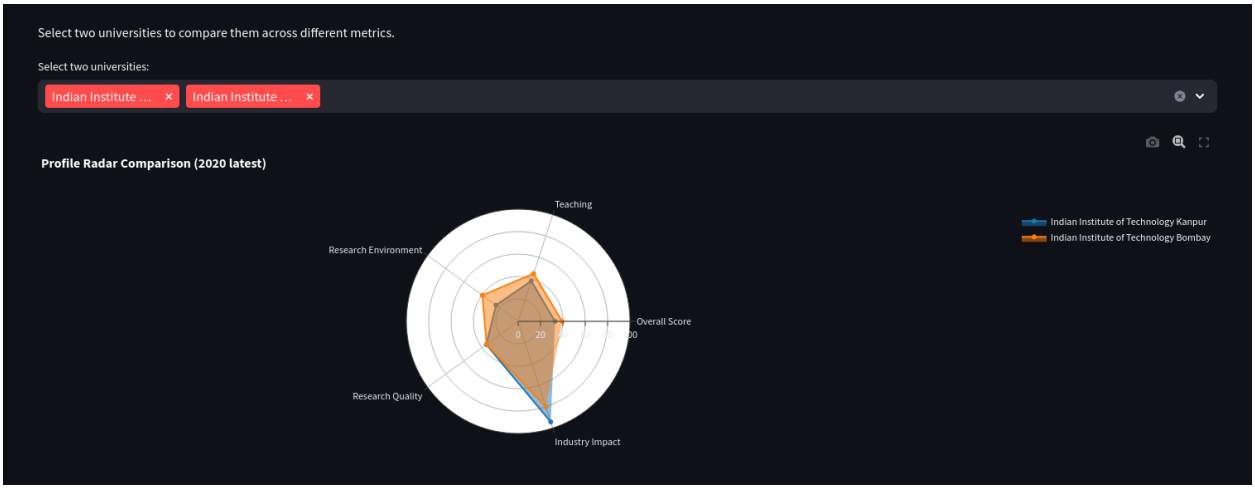


Figure 22: University Comparer using Radar Chart

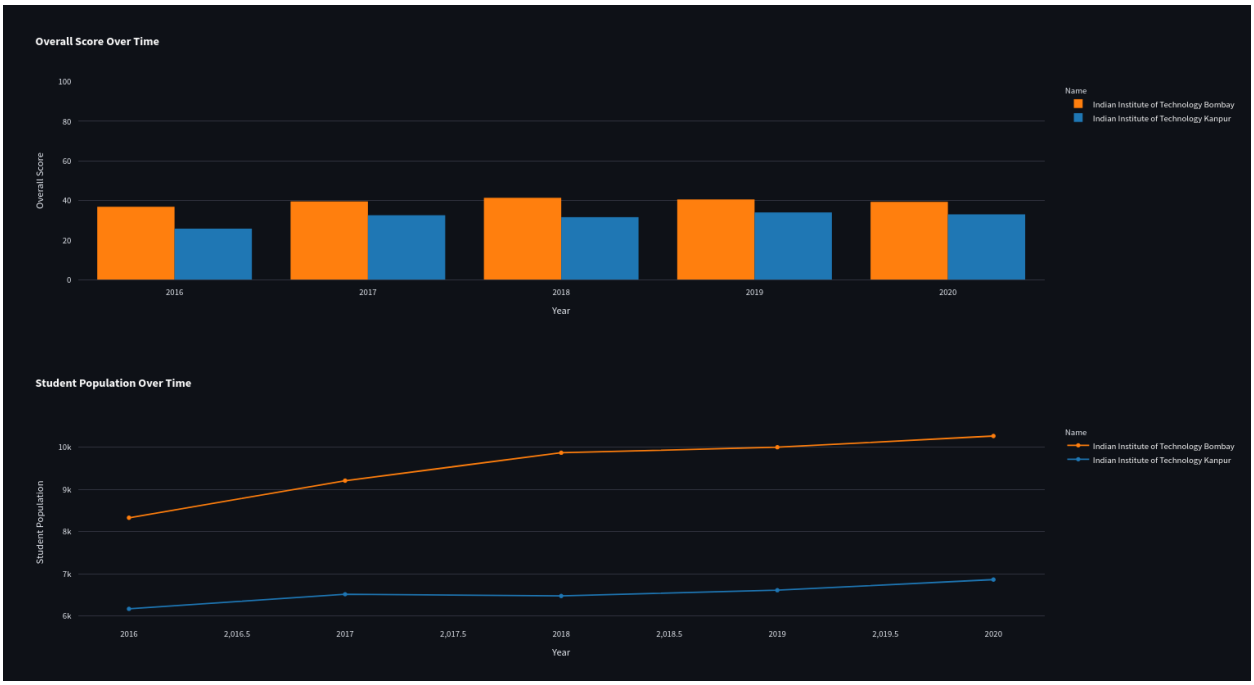
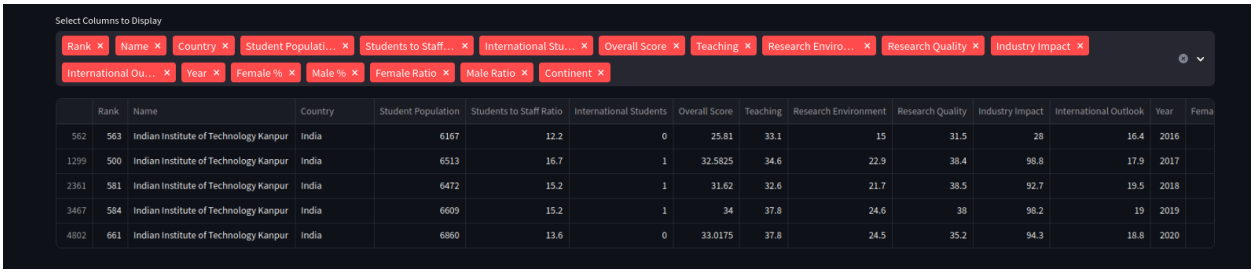


Figure 23: University Comparer using bar chart and line chart

2.18 View Data and EDA

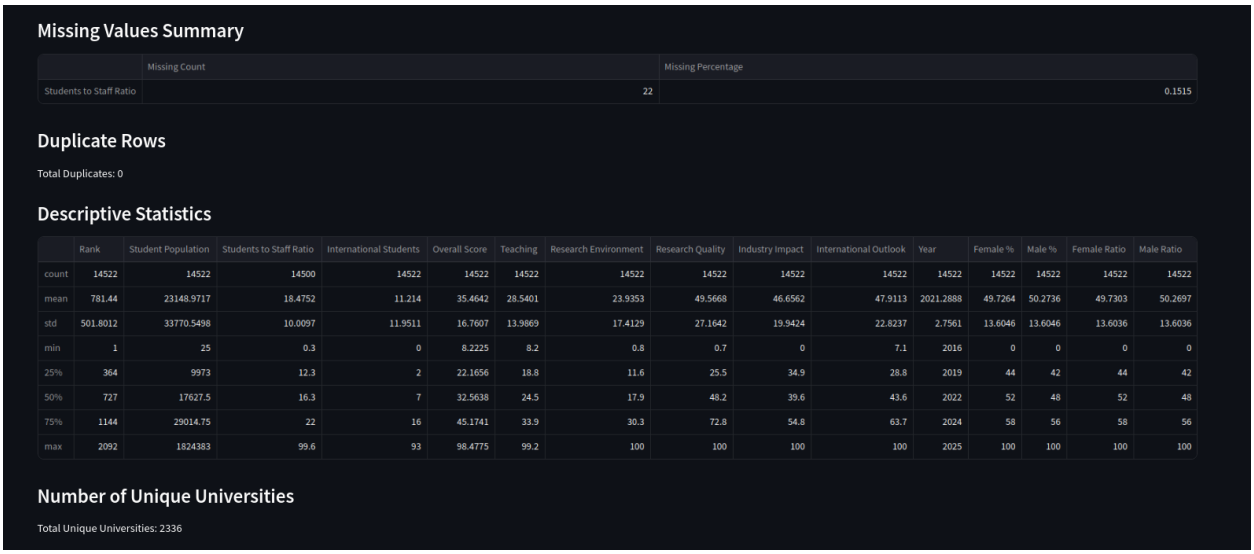
Here the objective was to allow users to explore raw data and exploratory data analysis (EDA) interactively by having exploratory tools to make the dashboard both interactive and fully transparent, encouraging open-ended exploration.

Proposed Solution: In this data-oriented task, we created a View Data Explorer and EDA (Exploratory Data Analysis) module. The View tab allows users to filter universities by year, country, or university name substring and explore any subset of data in a clean table. The EDA tab displays statistics about missing data, duplicate rows, and descriptive summaries (mean, standard deviation, etc.). Additionally, we built histograms for core metrics like Student Population, Students to Staff Ratio, Overall Score, and International Students % using Seaborn and Matplotlib, ensuring deeper understanding of dataset distributions.



	Rank	Name	Country	Student Population	Students to Staff Ratio	International Students	Overall Score	Teaching	Research Environment	Research Quality	Industry Impact	International Outlook	Year	Fema
562	563	Indian Institute of Technology Kanpur	India	6167	12.2	0	25.81	33.1	15	31.5	28	16.4	2016	
1299	500	Indian Institute of Technology Kanpur	India	6513	16.7	1	32.5825	34.6	22.9	38.4	98.8	17.9	2017	
2361	581	Indian Institute of Technology Kanpur	India	6472	15.2	1	31.62	32.6	21.7	38.5	92.7	19.5	2018	
3467	584	Indian Institute of Technology Kanpur	India	6609	15.2	1	34	37.8	24.6	38	98.2	19	2019	
4802	661	Indian Institute of Technology Kanpur	India	6860	13.6	0	33.0175	37.8	24.5	35.2	94.3	18.8	2020	

Figure 24: Bottom 10 Universities by rank



Missing Values Summary

	Missing Count	Missing Percentage
Students to Staff Ratio	22	0.1515

Duplicate Rows

Total Duplicates: 0

Descriptive Statistics

	Rank	Student Population	Students to Staff Ratio	International Students	Overall Score	Teaching	Research Environment	Research Quality	Industry Impact	International Outlook	Year	Female %	Male %	Female Ratio	Male Ratio
count	14522	14522	14500	14522	14522	14522	14522	14522	14522	14522	14522	14522	14522	14522	14522
mean	781.44	23148.9717	18.4752	11.214	35.4642	28.5401	23.9353	49.5668	46.6562	47.9113	2021.2888	49.7264	50.2736	49.7303	50.2697
std	501.8012	33770.5498	10.0097	11.9511	16.7607	13.9869	17.4129	27.1642	19.9424	22.8237	2.7561	13.6046	13.6046	13.6036	13.6036
min	1	25	0.3	0	8.2225	8.2	0.8	0.7	0	7.1	2016	0	0	0	0
25%	364	9973	12.3	2	22.1656	18.8	11.6	25.5	34.9	28.8	2019	44	42	44	42
50%	727	17627.5	16.3	7	32.5638	24.5	17.9	48.2	39.6	43.6	2022	52	48	52	48
75%	1144	29014.75	22	16	45.1741	33.9	30.3	72.8	54.8	63.7	2024	58	56	58	56
max	2092	1824383	99.6	93	98.4775	99.2	100	100	100	100	2025	100	100	100	100

Number of Unique Universities

Total Unique Universities: 2336

Figure 25: Bottom 10 Universities by rank

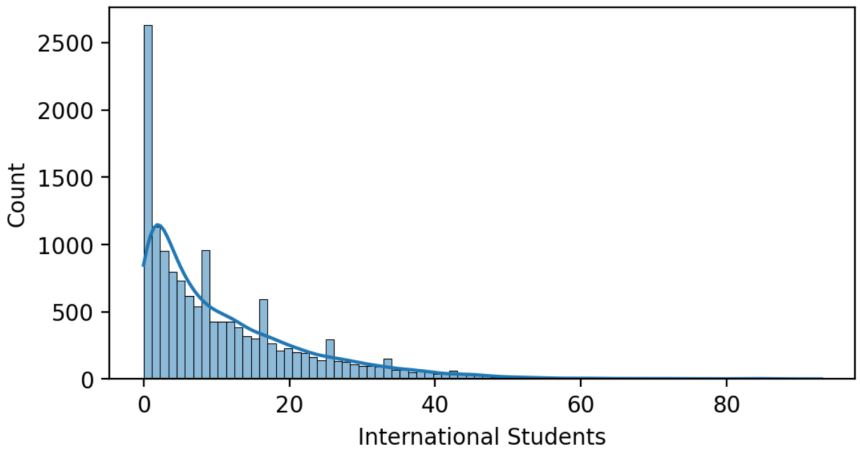


Figure 26: Bottom 10 Universities by rank

3 Results

Through our visual analytics system, we address several critical analytical problems:

- Identifying global and regional trends in higher education performance.
- Analyzing gender diversity and international student enrollment trends.
- Comparing university performances across multiple criteria over time.
- Understanding resource allocation impact (student-to-staff ratios) on teaching quality.
- Uncovering clusters of universities with similar characteristics.

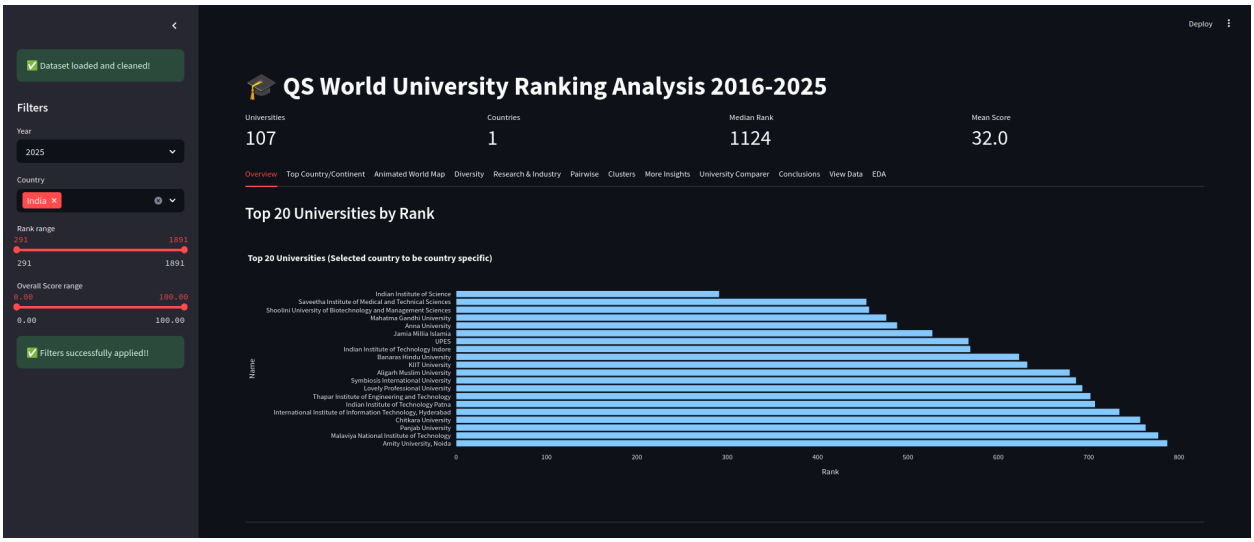


Figure 27: Homepage of Visualization Dashboard

3.1 Interactive Visualizations and their Significance

3.1.1 Overview Tab

- **Bar Charts (Top and Bottom Universities):** Bar charts are ideal for comparing discrete categories. Here, they effectively highlight the highest and lowest-ranking universities clearly. It is inferred that institutions like MIT and Oxford consistently dominate the top ranks, while certain smaller or emerging universities populate the bottom ranks.

- **Line Charts (Rank Trajectories):** Line charts are excellent for illustrating trends over time. They convey stability and volatility in rankings. We see stability among top universities like Stanford and Harvard, but volatility for mid-ranked institutions, indicating competitive shifts.

3.1.2 Top Country/Continent Tab

- **Sunburst Chart:** This hierarchical visualization effectively displays nested categorical data (continent → country → university), showcasing contributions at each hierarchy level clearly.

- **Treemap:** Similar to sunbursts, treemaps represent hierarchical data as nested rectangles. It visually emphasizes relative size differences clearly. U.S. universities form a large portion of global top scorers. Europe and North America dominate in high scores, but Asia's inner layers are thickening, indicating rising competitiveness.

- **Bar Chart (Country Comparisons):** Offers easy direct comparison of overall scores across selected countries. It confirms that the United States and United Kingdom lead in overall scores, followed by China and Japan.

3.1.3 Animated World Map Tab

- **Choropleth Maps:** Geographic visualizations color-coded to represent data values, effectively conveying global trends dynamically over time. They highlight geographical patterns clearly. University counts have increased globally, especially in Asia. International student percentages are highest in Australia, U.K., and Canada. Female student ratios have slightly increased across all continents.

3.1.4 Diversity Tab

- **Line Charts (Gender Diversity Trends):** Clearly illustrate the evolution of gender representation. They show steady improvement in gender balance globally.
- **Scatter Plot (Student-to-Staff Ratio vs Teaching Score):** Shows individual university data points, highlighting correlations and outliers. The trend line indicates a negative correlation: as student-to-staff ratio increases (worsens), teaching scores decrease, emphasizing the importance of faculty availability.
- **Line Charts (International Students Trends):** Australia, Canada, and U.K. consistently attract high numbers of international students.

3.1.5 Research & Industry Tab

- **Scatter Plots (Research vs. Industry):** Clearly demonstrate relationships and trends between two variables, highlighting performance relationships effectively. Show positive correlations between industry collaboration and research strength, notably in technical universities.
- **Box and Violin Plots (Metrics Distribution):** Offer clear statistical insights into data spread and central tendencies. Indicate that research quality scores are more skewed toward higher values, while teaching scores have a wider spread, suggesting inconsistency in teaching quality globally.

3.1.6 Pairwise Tab

- **Scatter Matrix:** Series of scatter plots clearly illustrating relationships between all selected variables simultaneously. Reveals that Overall Score is strongly correlated with Research Quality and Teaching.
- **Correlation Heatmap:** Visually encodes correlation coefficients, clearly highlighting strong and weak metric relationships. Shows that Teaching and Research Quality have the highest positive correlation with Overall Score. Student-to-Staff Ratio negatively correlates with Teaching.

3.1.7 Clusters & PCA Tab

- **K-Means Clustering:** Groups similar data points clearly to identify meaningful clusters within universities. It was found that research-intensive institutions form a distinct group, separate from teaching-focused universities.
- **PCA Scatter Plots:** Reduce dimensionality clearly, illustrating underlying data structures in 2D and 3D formats effectively. Most variance is captured in the first two principal components, confirming that Overall Score, Research Quality, and Teaching dominate variance.

3.1.8 More Insights Tab

- **Scatter Matrix (Global Pair Plot):** Clearly shows comprehensive relationships and trends among all variables over the full dataset. Global trends show improvement in Research Environment and Industry Impact over the years.
- **UMAP + HDBSCAN Clustering** : . UMAP + HDBSCAN is ideal for this dataset because educational institutions often exhibit non-linear relationships, and forcing linear projections would mask those structures. This technique uncovers hidden clusters like elite research universities, rising regional players, or industry-driven universities, offering profound insights into academic ecosystems.
- **University Similarity Network** : This method allows users to precisely explore cross-country academic neighbors based on true academic profiles, ensuring full analytical rigor while maintaining visual clarity and engagement.

3.1.9 University Comparer Tab

- **Radar Chart:** Clearly compares multiple dimensions simultaneously between two universities. Visually compares universities across multiple metrics. For instance, IIT Bombay and IIT Delhi show higher Industry Impact compared to Teaching scores.
- **Timeline Charts (Line and Bar):** Clearly illustrate historical performance trends across selected metrics. Show performance consistency or improvements over time, critical for strategic analysis.

3.1.10 10. Exploratory Data Analysis Tab

- **Missing Values & Duplicates Analysis:** Data quality was validated ensuring robust results.
- **Histograms:** Clearly display distribution and frequency of numeric data, useful for identifying central tendencies and variability. Distributions confirm that while a majority of universities have moderate Overall Scores, only a few institutions achieve very high scores.

3.2 Additional Insights

- **Internationalization Trends:** Increased international student percentages in smaller nations like Singapore and UAE show shifts in global educational hubs.
- **Gender Disparity Reduction:** Policies aiming for diversity are succeeding gradually; however, focused efforts are still necessary in parts of Asia and Africa.
- **Industry Collaboration Impact:** Universities with high Industry Impact scores generally have strong Research Quality scores too, suggesting synergistic effects.
- **Student Population Distribution:** High variance in student population sizes highlights that size alone does not determine rank or quality.
- **Continental Comparisons:** Europe maintains teaching quality; Asia improves research quality; Oceania leads in international outlook.

4 Conclusion & Story

In this project, we explored how combining machine learning techniques with rich data visualizations can significantly deepen our understanding of complex datasets such as global university rankings. By building an interactive, multidimensional dashboard, we moved beyond simple ranking tables and enabled a dynamic exploration of various dimensions like teaching quality, research impact, student diversity, and international competitiveness.

Techniques like Principal Component Analysis (PCA) and KMeans helped us capture broader patterns and clusters among institutions worldwide. To uncover deeper, non-linear relationships, we applied UMAP and HDBSCAN, which allowed us to isolate hidden clusters and detect outliers with greater accuracy. By constructing similarity networks based on true cosine similarities, we also introduced a new perspective: identifying academic "twins" across different countries and continents, something traditional ranking lists cannot easily reveal.

Our analysis shows that while traditional academic leaders such as the United States and the United Kingdom maintain strong positions, universities in the Asia-Pacific region are rapidly closing the gap. Focused investments in research quality and international engagement are driving this change. We also observed encouraging improvements in gender diversity, though regional gaps still exist and highlight the need for targeted policy interventions. Moreover, the growing link between industry collaboration and research environments suggests that future academic excellence will increasingly depend on practical, real-world engagement alongside pure academic metrics.

Every method and visualization in this project was carefully chosen based on the type of insight it could provide, rather than for aesthetic appeal. From animated world maps to UMAP projections that preserve

manifold structures, each visualization was designed to highlight underlying movements, relationships, and patterns that traditional views might miss.

Overall, this project goes beyond simply reporting which universities rank highest. It sheds light on how and why certain patterns emerge, evolve, and sometimes diverge over time. The final system offers an educational intelligence framework that is not only informative but also actionable — serving students, researchers, policymakers, and institutional leaders aiming to better understand and shape the future of higher education.

5 Link to Source Code

The complete project source code is available at: `https://github.com/khushwantk/THE_uni_vizualization`

6 References

- **Times Higher Education World University Rankings Dataset (2016–2025)**: https://www.kaggle.com/datasets/raymondtoo/the-world-university-rankings-2016-2024
- **Streamlit**: https://streamlit.io/
- **Pandas**: https://pandas.pydata.org/
- **NumPy**: https://numpy.org/
- **Plotly**: https://plotly.com/python/
- **Scikit-learn**: https://scikit-learn.org/stable/
- **NetworkX**: https://networkx.org/
- **UMAP-learn**: https://umap-learn.readthedocs.io/en/latest/
- **HDBSCAN**: https://hdbscan.readthedocs.io/en/latest/
- **Matplotlib**: https://matplotlib.org/
- **Seaborn**: https://seaborn.pydata.org/