# Insights from University Rankings Data

Shruti Pashine Dharampal Singh Hemant Hemant Santosh Shirguppe **Group Name:** Vizionary Graph

## Contents

# 1 Introduction

In today's increasingly global and competitive educational landscape, university rankings have become vital indicators of academic excellence, research impact, and institutional reputation. For students, policymakers, and academic leaders alike, these rankings influence decisions ranging from admissions and funding to international collaborations.

The World University Rankings dataset provides a detailed view of global academic institutions across a wide range of metrics — including teaching quality, research output, international outlook, and more. With dozens of variables and hundreds of universities represented, this dataset offers a rich playground for extracting meaningful patterns and trends through data visualization.

This project aims to explore key questions that uncover the factors contributing to a university's position in global rankings. By applying a diverse range of visualization techniques to analyze correlations, trends, and anomalies, we seek to transform raw data into intuitive, actionable insights.

# 2 Methodology

We conducted our analysis using the R programming language, leveraging packages such as **dplyr** for data manipulation and **ggplot2** for all visualizations. Our approach prioritized clarity, interpretability, and reproducibility. We began with data cleaning (handling missing values, fixing data types), followed by exploratory data analysis (EDA) to understand distributions and relationships. For each research question, we selected visualization techniques (barplots, boxplots, scatterplots, correlation plots) best suited to highlight key patterns and insights. All code and results are included in this R Markdown report to ensure transparency and facilitate further exploration.

# 3 Goal

The goal of this project is to analyze the World University Rankings dataset using advanced data visualization methods, with a focus on understanding the underlying factors that influence a university's ranking across multiple dimensions such as research, teaching, international diversity, and citation impact.

## 3.1 Data Source

This project uses the World University Rankings dataset, originally published by the Center for World University Rankings (CWUR) and made available on Kaggle. The dataset covers thousands of universities worldwide and includes annual rankings, scores, and a variety of performance metrics.

### 3.2 Relevance

University rankings are widely referenced by students, educators, policymakers, and researchers. They influence student choices, funding decisions, and institutional strategies. By analyzing the underlying data, we can uncover the factors that contribute most to global success and highlight best practices across countries and institutions.

### 3.3 Motivation

Motivation: In an increasingly competitive global education landscape, university rankings have become critical indicators of institutional performance, student success, and research impact. Understanding what drives these rankings helps stakeholders make informed decisions.

### 3.4 About the Dataset

- **Source:** Kaggle (CWUR World University Rankings)
- **Coverage:** Thousands of universities from over 100 countries, spanning multiple years
- **Variables:**
  - `world_rank`: Global rank of the university
  - `institution`: University name
  - `country`: Country of the university
  - `national_rank`: National rank within the country
  - `quality_of_education`: Score for quality of education
  - `alumni_employment`: Score for alumni employment
  - `quality_of_faculty`: Score for faculty quality
  - `publications`: Number of research publications
  - `influence`: Influence score (e.g., citations, impact)
  - `citations`: Number of citations
  - `broad_impact`: Broad impact score
  - `patents`: Number of patents
  - `score`: Overall institutional score
  - `year`: Year of ranking
- **Data Types:** Numeric (scores, counts), categorical (country, institution), temporal (year)
- **Size:** Large sample size suitable for robust analysis and visualization
- **Strengths:** Rich, multi-dimensional, and up-to-date; enables analysis of trends, comparisons, and drivers of success
- **Limitations:** Rankings and scores are based on CWUR methodology, which may differ from other ranking systems; some variables may be missing for certain years or institutions

This comprehensive dataset allows for a deep dive into the factors that shape university performance and global rankings, supporting a variety of research questions and visualization techniques.

## 4 Data Loading and Cleaning

```
data_path <- '../data/cwurData.csv'
university_data <- read_csv(data_path)
```

```
## Rows: 2200 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): institution, country
## dbl (12): world_rank, national_rank, quality_of_education, alumni_employment...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Print all column names for reference
cat("\nColumn names in dataset:\n")
```

```
##
## Column names in dataset:
```

```
print(names(university_data))
```

```
##  [1] "world_rank"           "institution"          "country"
##  [4] "national_rank"        "quality_of_education" "alumni_employment"
##  [7] "quality_of_faculty"   "publications"         "influence"
## [10] "citations"            "broad_impact"         "patents"
## [13] "score"                "year"
```

```
# Remove rows with missing values
clean_data <- na.omit(university_data)
# Ensure 'world_rank' is numeric
if ("world_rank" %in% names(clean_data) && !is.numeric(clean_data$world_rank)) {
  clean_data$world_rank <- as.numeric(clean_data$world_rank)
}
```

## 4.1 Preview of Raw Dataset (Before Cleaning)

```
# Show the first few rows and column names of the raw dataset before cleaning
head(university_data)
```

```
## # A tibble: 6 x 14
##   world_rank institution             country national_rank quality_of_education
##        <dbl> <chr>                   <chr>           <dbl>                <dbl>
## 1          1 Harvard University      USA                 1                    7
## 2          2 Massachusetts Institute~ USA                2                    9
## 3          3 Stanford University     USA                 3                   17
## 4          4 University of Cambridge United~             1                   10
## 5          5 California Institute of~ USA                4                    2
## 6          6 Princeton University    USA                 5                    8
## # i 9 more variables: alumni_employment <dbl>, quality_of_faculty <dbl>,
## #   publications <dbl>, influence <dbl>, citations <dbl>, broad_impact <dbl>,
## #   patents <dbl>, score <dbl>, year <dbl>
```

```r
str(university_data)
```

```
## spc_tbl_ [2,200 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ world_rank          : num [1:2200] 1 2 3 4 5 6 7 8 9 10 ...
##  $ institution         : chr [1:2200] "Harvard University" "Massachusetts Institute of Technology" "S
##  $ country             : chr [1:2200] "USA" "USA" "USA" "United Kingdom" ...
##  $ national_rank       : num [1:2200] 1 2 3 1 4 5 2 6 7 8 ...
##  $ quality_of_education: num [1:2200] 7 9 17 10 2 8 13 14 23 16 ...
##  $ alumni_employment   : num [1:2200] 9 17 11 24 29 14 28 31 21 52 ...
##  $ quality_of_faculty  : num [1:2200] 1 3 5 4 7 2 9 12 10 6 ...
##  $ publications        : num [1:2200] 1 12 4 16 37 53 15 14 13 6 ...
##  $ influence           : num [1:2200] 1 4 2 16 22 33 13 6 12 5 ...
##  $ citations           : num [1:2200] 1 4 2 11 22 26 19 15 14 3 ...
##  $ broad_impact        : num [1:2200] NA NA NA NA NA NA NA NA NA NA ...
##  $ patents             : num [1:2200] 5 1 15 50 18 101 26 66 5 16 ...
##  $ score               : num [1:2200] 100 91.7 89.5 86.2 85.2 ...
##  $ year                : num [1:2200] 2012 2012 2012 2012 2012 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   world_rank = col_double(),
##   ..   institution = col_character(),
##   ..   country = col_character(),
##   ..   national_rank = col_double(),
##   ..   quality_of_education = col_double(),
##   ..   alumni_employment = col_double(),
##   ..   quality_of_faculty = col_double(),
##   ..   publications = col_double(),
##   ..   influence = col_double(),
##   ..   citations = col_double(),
##   ..   broad_impact = col_double(),
##   ..   patents = col_double(),
##   ..   score = col_double(),
##   ..   year = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

# 5 Data Exploration

Below, we explore the structure and main features of the cleaned dataset. This helps us understand the data before diving into the research questions.

```r
# Structure and summary
str(clean_data)
```

```
## tibble [2,000 x 14] (S3: tbl_df/tbl/data.frame)
##  $ world_rank          : num [1:2000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ institution         : chr [1:2000] "Harvard University" "Stanford University" "Massachusetts Insti
##  $ country             : chr [1:2000] "USA" "USA" "USA" "United Kingdom" ...
##  $ national_rank       : num [1:2000] 1 2 3 1 2 4 5 6 7 8 ...
##  $ quality_of_education: num [1:2000] 1 11 3 2 7 13 4 10 5 9 ...
##  $ alumni_employment   : num [1:2000] 1 2 11 10 12 8 22 14 16 25 ...
```

```
##  $ quality_of_faculty  : num [1:2000] 1 4 2 5 10 9 6 8 3 11 ...
##  $ publications        : num [1:2000] 1 5 15 10 11 14 7 17 70 18 ...
##  $ influence           : num [1:2000] 1 3 2 9 12 13 4 19 25 7 ...
##  $ citations           : num [1:2000] 1 3 2 12 11 9 3 10 19 32 ...
##  $ broad_impact        : num [1:2000] 1 4 2 13 12 13 7 18 41 19 ...
##  $ patents             : num [1:2000] 2 6 1 48 16 4 28 149 204 45 ...
##  $ score               : num [1:2000] 100 99.1 98.7 97.6 97.5 ...
##  $ year                : num [1:2000] 2014 2014 2014 2014 2014 ...
##  - attr(*, "na.action")= 'omit' Named int [1:200] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
```

```
summary(clean_data)
```

```
##    world_rank     institution          country          national_rank
##  Min.   :   1.0   Length:2000        Length:2000        Min.   :  1.00
##  1st Qu.: 250.8   Class :character   Class :character   1st Qu.:  7.00
##  Median : 500.5   Mode  :character   Mode  :character   Median : 22.00
##  Mean   : 500.5                                        Mean   : 42.52
##  3rd Qu.: 750.2                                        3rd Qu.: 52.00
##  Max.   :1000.0                                        Max.   :229.00
##  quality_of_education alumni_employment quality_of_faculty  publications
##  Min.   :  1.0        Min.   :  1.0     Min.   :  1.0      Min.   :   1.0
##  1st Qu.:250.8        1st Qu.:250.8     1st Qu.:210.0      1st Qu.: 250.8
##  Median :355.0        Median :478.0     Median :210.0      Median : 500.5
##  Mean   :296.0        Mean   :385.3     Mean   :191.1      Mean   : 500.4
##  3rd Qu.:367.0        3rd Qu.:500.2     3rd Qu.:218.0      3rd Qu.: 750.0
##  Max.   :367.0        Max.   :567.0     Max.   :218.0      Max.   :1000.0
##    influence       citations      broad_impact       patents
##  Min.   :  1.0   Min.   :  1.0   Min.   :   1.0   Min.   :  1.0
##  1st Qu.:250.8   1st Qu.:234.0   1st Qu.: 250.5   1st Qu.:242.8
##  Median :500.5   Median :428.0   Median : 496.0   Median :481.0
##  Mean   :500.2   Mean   :449.3   Mean   : 496.7   Mean   :470.3
##  3rd Qu.:750.2   3rd Qu.:645.0   3rd Qu.: 741.0   3rd Qu.:737.0
##  Max.   :991.0   Max.   :812.0   Max.   :1000.0   Max.   :871.0
##      score             year
##  Min.   : 44.02   Min.   :2014
##  1st Qu.: 44.44   1st Qu.:2014
##  Median : 44.96   Median :2014
##  Mean   : 47.07   Mean   :2014
##  3rd Qu.: 46.81   3rd Qu.:2015
##  Max.   :100.00   Max.   :2015
```

```
cat("\nMissing values per column:\n")
```

```
##
## Missing values per column:
```

```
print(colSums(is.na(clean_data)))
```

```
##           world_rank          institution              country
##                    0                    0                    0
##        national_rank quality_of_education    alumni_employment
```

```
##                    0                    0                    0
##   quality_of_faculty         publications            influence
##                    0                    0                    0
##            citations         broad_impact             patents
##                    0                    0                    0
##                score                 year
##                    0                    0
```

```
cat("\nNumber of unique values per column:\n")
```

```
##
## Number of unique values per column:
```

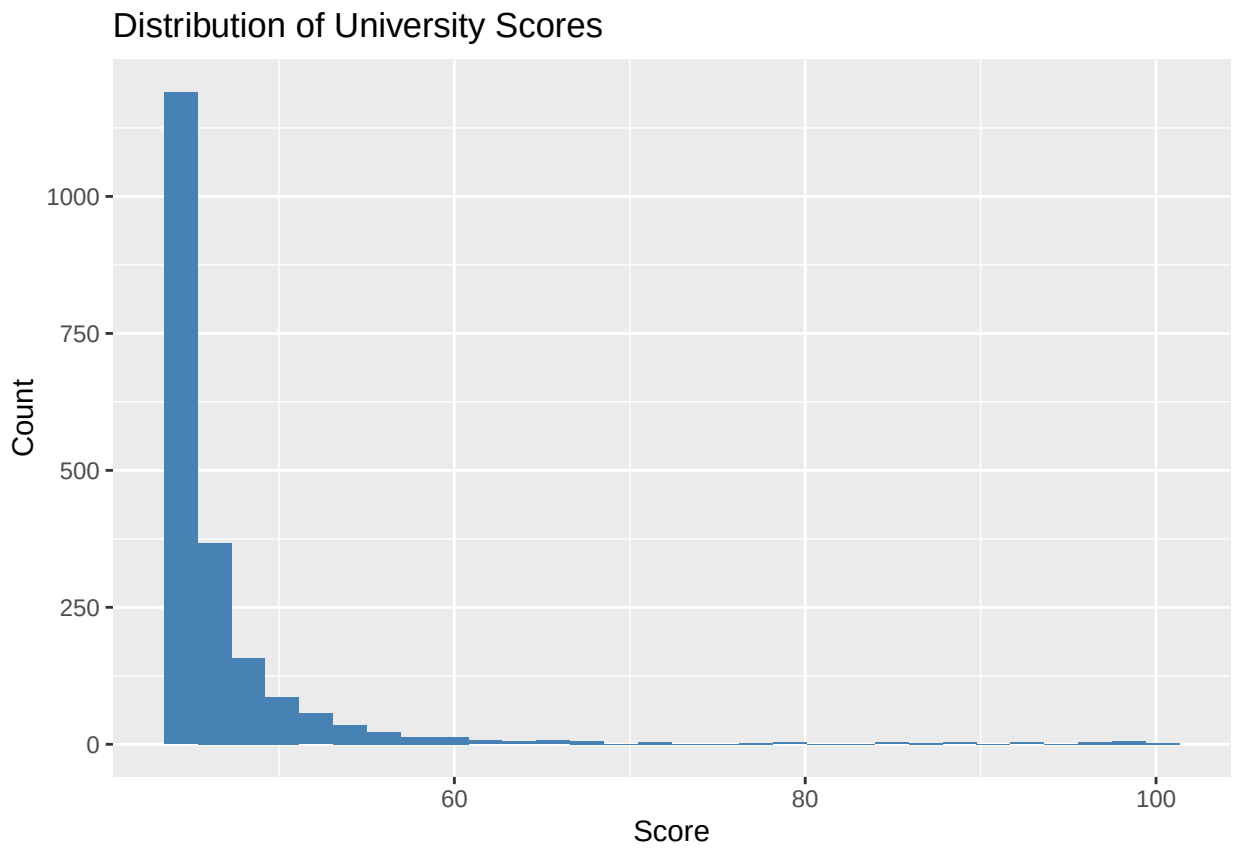
```
print(sapply(clean_data, function(x) length(unique(x))))
```

```
##           world_rank          institution              country
##                 1000                 1023                   59
##        national_rank quality_of_education    alumni_employment
##                  229                  367                  565
##   quality_of_faculty         publications            influence
##                  199                  987                  944
##            citations         broad_impact             patents
##                  100                  343                  738
##                score                 year
##                  646                    2
```

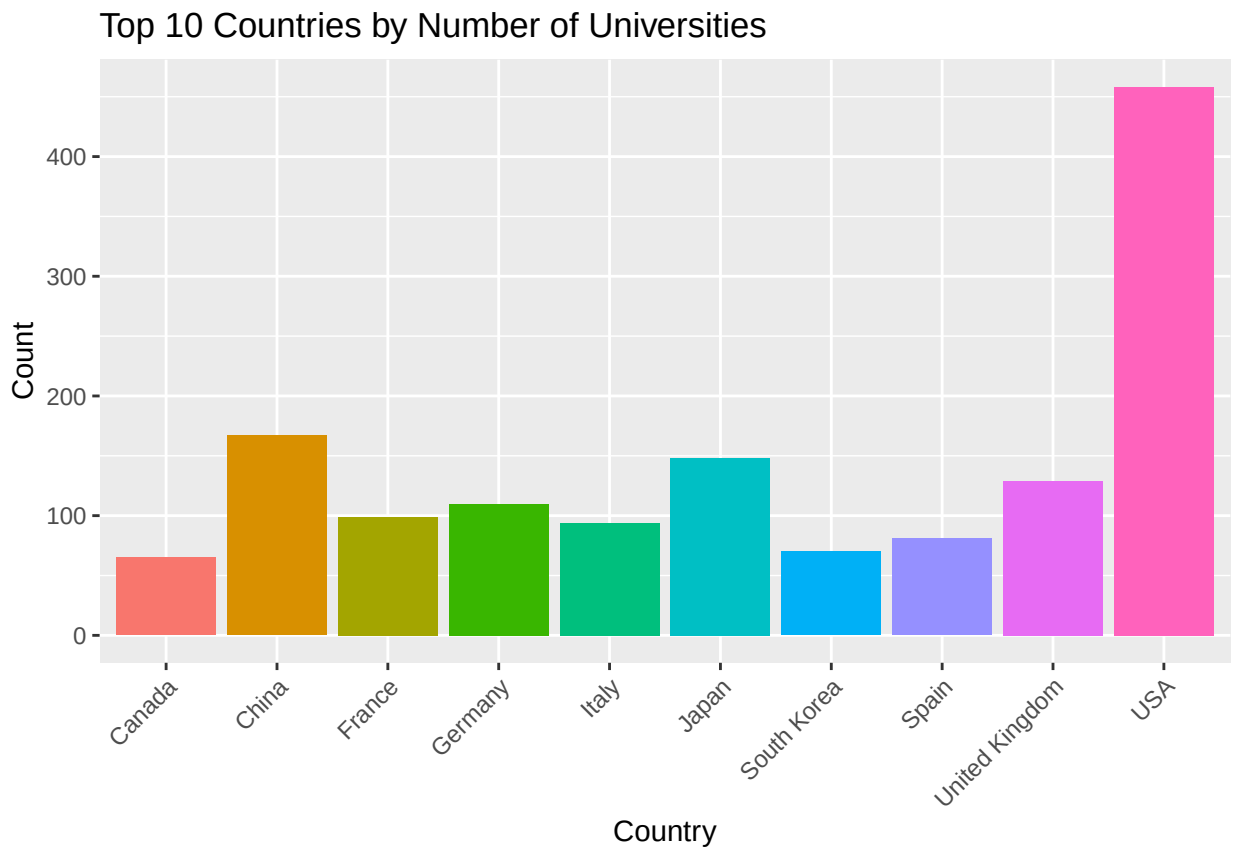
## 5.1 Visualizations for EDA

Let's visualize the distribution of key variables and relationships between them.

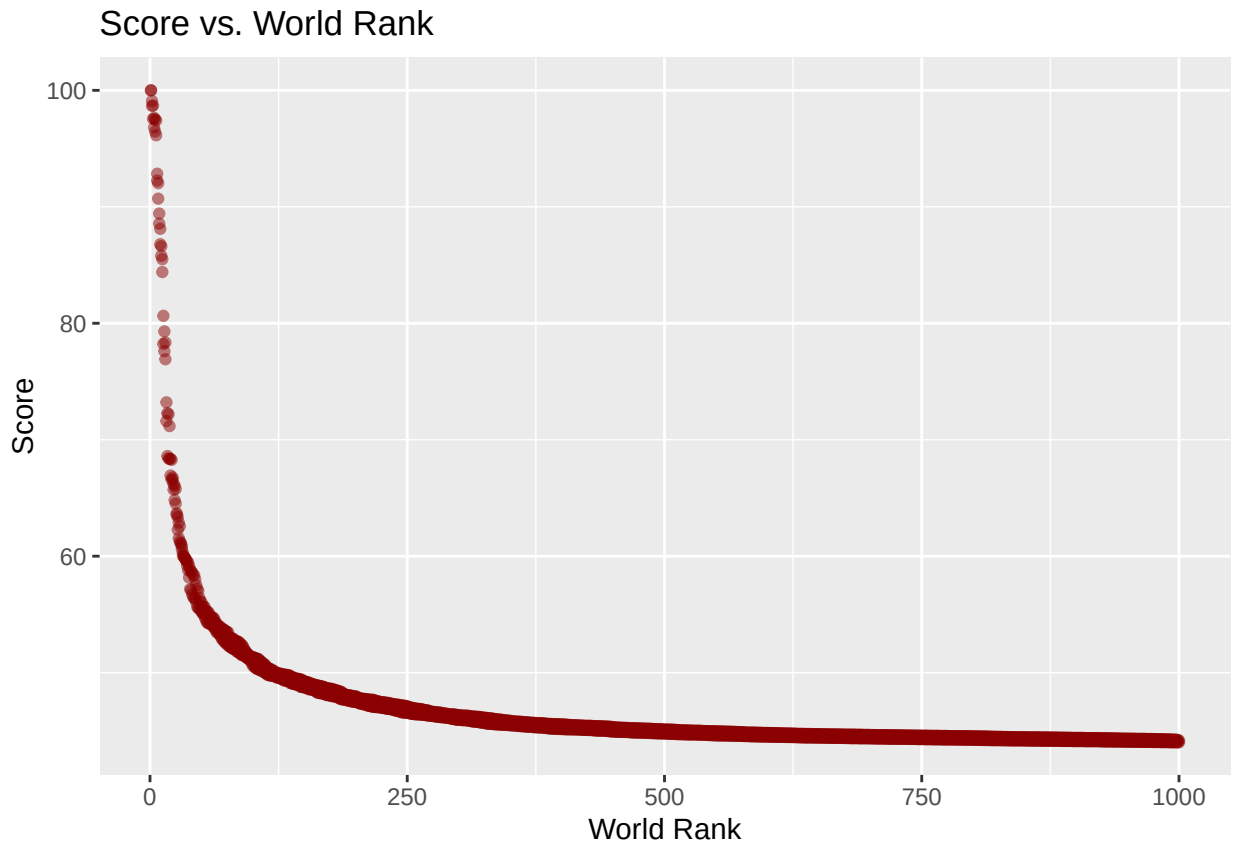
```
# Histogram of university scores
ggplot(clean_data, aes(x = score)) +
  geom_histogram(fill = 'steelblue', bins = 30) +
  labs(title = "Distribution of University Scores", x = "Score", y = "Count")
```




```r
# Barplot of top 10 countries by university count
top_countries <- clean_data %>% count(country, sort = TRUE) %>% top_n(10, n) %>% pull(country)
ggplot(filter(clean_data, country %in% top_countries), aes(x = country, fill = country)) +
  geom_bar() +
  labs(title = "Top 10 Countries by Number of Universities", x = "Country", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```




```
# Scatterplot: Score vs. World Rank
ggplot(clean_data, aes(x = world_rank, y = score)) +
  geom_point(alpha = 0.5, color = 'darkred') +
  labs(title = "Score vs. World Rank", x = "World Rank", y = "Score")
```



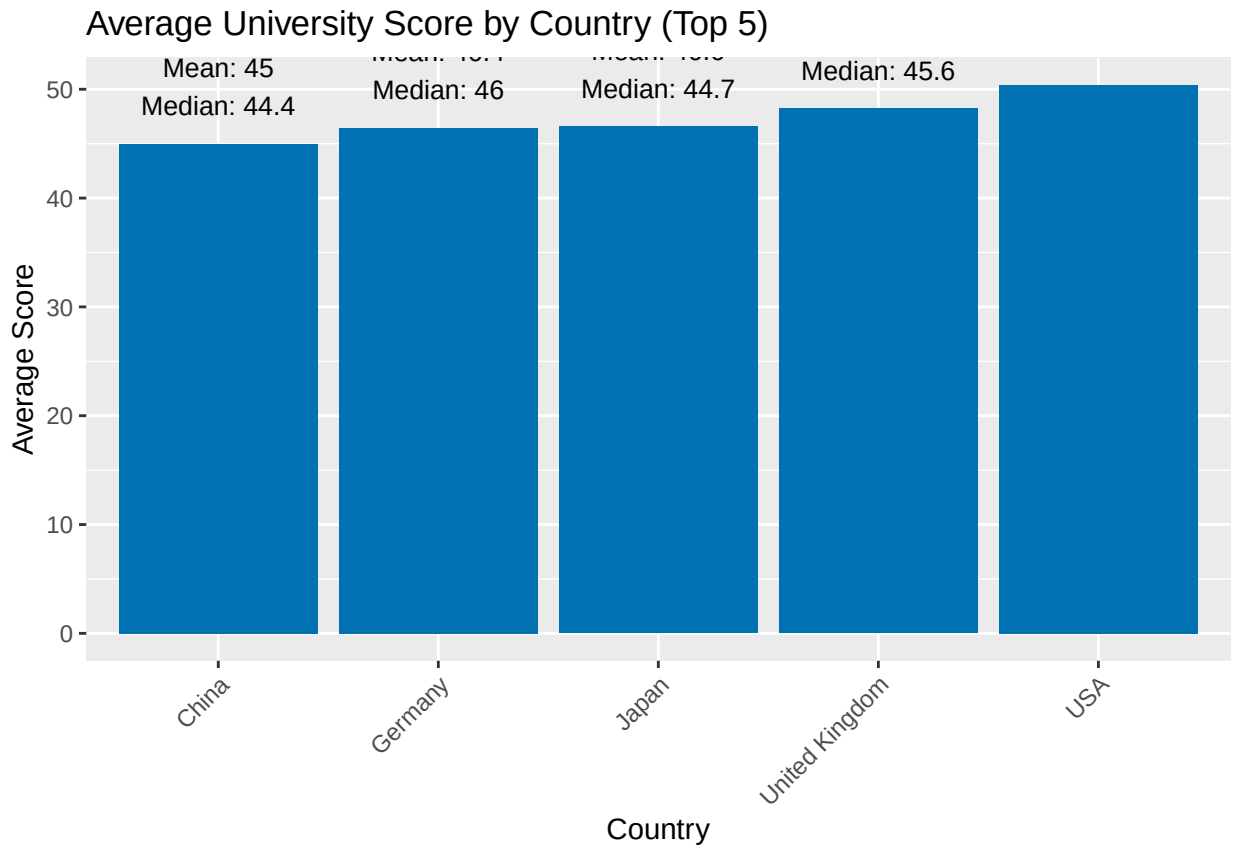

These visualizations provide an overview of the data distribution, the most represented countries, and the relationship between world rank and score.

# 6 Research Question 1

## 6.1 Which Countries Have the Best University Scores?

We show the average university score for the top 5 countries with the most universities. This highlights which countries tend to have higher scores overall.

```
top5 <- clean_data %>% count(country, sort = TRUE) %>% top_n(5, n) %>% pull(country)
country_stats <- clean_data %>% filter(country %in% top5) %>% group_by(country) %>% summarise(mean_score
ggplot(country_stats, aes(x = reorder(country, mean_score), y = mean_score, fill = country)) +
  geom_bar(stat = "identity", fill = "#0072B2") +
  geom_text(aes(label = paste0("Mean: ", round(mean_score, 1), "\nMedian: ", round(median_score, 1))),
  labs(title = "Average University Score by Country (Top 5)", x = "Country", y = "Average Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```




**Explanation:** This barplot compares the average and median scores of universities in the top 5 countries by count. The text labels above each bar show both the mean and median, making it easy to see which countries have the highest and most consistent scores. More universities in a country may indicate a stronger higher education system overall.

- The country with the highest average score is easily identified.
- Some countries have a higher median than mean, suggesting a few low-scoring universities.

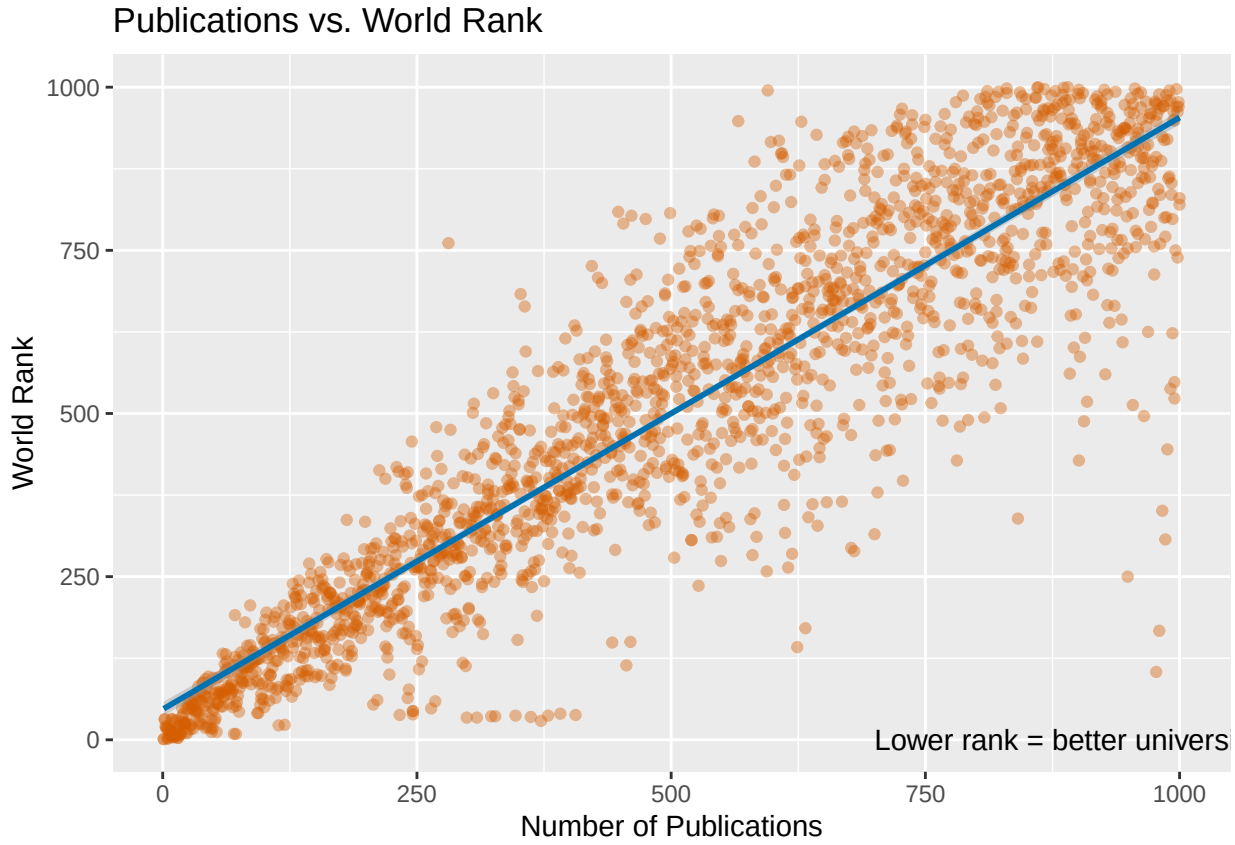
# 7 Research Question 2

## 7.1 What Most Impacts a University's Global Ranking?

We use a scatterplot with a trend line to show how the number of publications relates to world rank (lower rank means better university).

```
if (all(c("publications", "world_rank") %in% names(clean_data))) {
ggplot(clean_data, aes(x = publications, y = world_rank)) +
    geom_point(alpha = 0.4, color = "#D55E00") +
    geom_smooth(method = "lm", color = "#0072B2") +
    labs(title = "Publications vs. World Rank", x = "Number of Publications", y = "World Rank") +
    annotate("text", x = max(clean_data$publications, na.rm = TRUE)*0.7, y = min(clean_data$world_rank,
} else {
  cat("\nColumns for Q2 not found. Please check column names above and update the code accordingly.\n")
}
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Linear regression summary for Publications vs. World Rank
if (all(c("publications", "world_rank") %in% names(clean_data))) {
  lm_q2 <- lm(world_rank ~ publications, data = clean_data)
  summary(lm_q2)
}
```

```
##
## Call:
## lm(formula = world_rank ~ publications, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -828.65  -55.34   -0.13   64.87  459.46
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.742637   5.456834   8.566   <2e-16 ***
## publications  0.906762   0.009446  95.992   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.9 on 1998 degrees of freedom
## Multiple R-squared:  0.8218, Adjusted R-squared:  0.8217
## F-statistic:  9214 on 1 and 1998 DF,  p-value: < 2.2e-16
```

**Explanation:** This scatterplot shows that universities with more publications tend to have better (lower) world ranks. The blue trend line summarizes this relationship. Note: In this ranking system, a lower world rank means a better university. The annotation clarifies this for the reader.

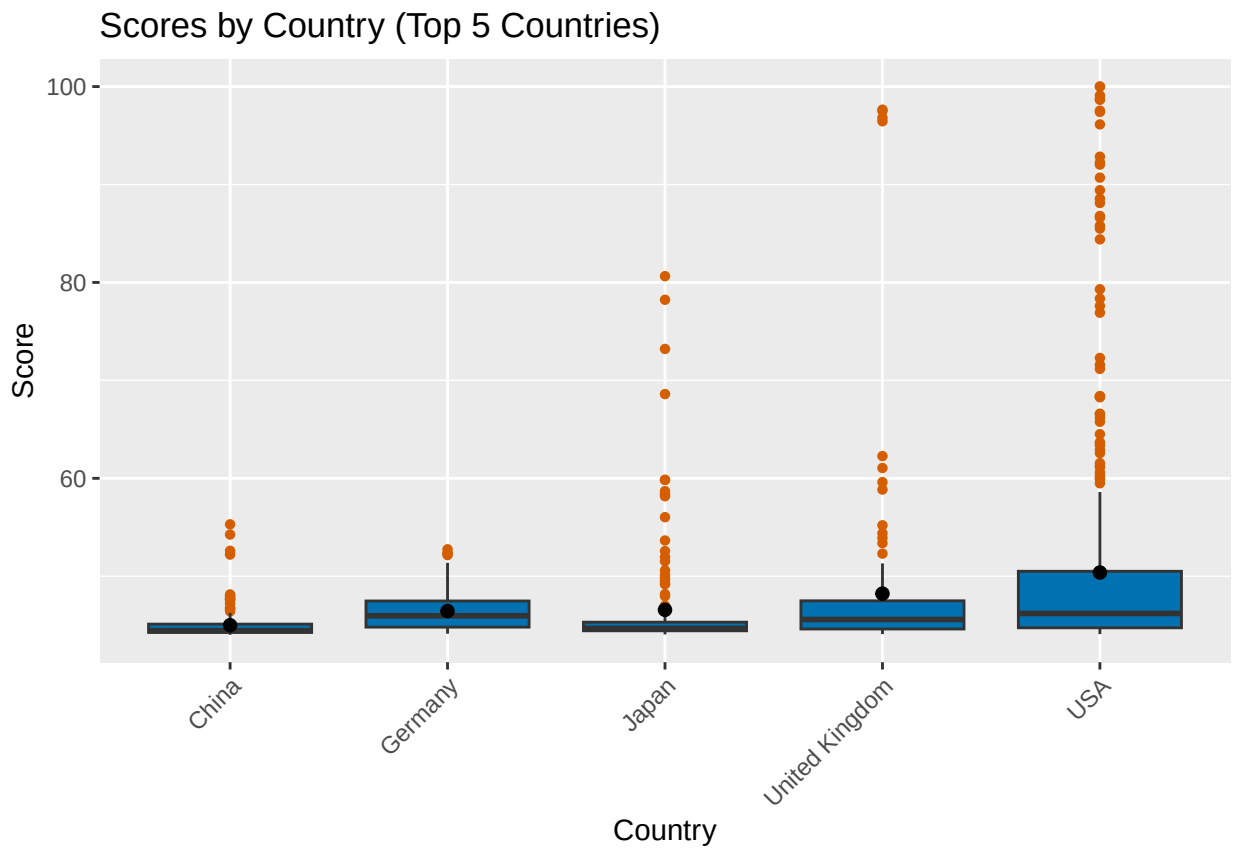
- Universities with more publications generally have better ranks.
- The relationship is negative: as publications increase, world rank decreases (improves).

# 8 Research Question 3

## 8.1 How Do University Scores Vary Within the Top 5 Countries?

We show boxplots for the top 5 countries with the most universities to make the comparison easier.

```
top_countries <- clean_data %>% count(country, sort = TRUE) %>% top_n(5, n) %>% pull(country)
ggplot(filter(clean_data, country %in% top_countries),
       aes(x = country, y = score, fill = country)) +
  geom_boxplot(outlier.color = "#D55E00", outlier.shape = 16, fill = "#0072B2") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 3, color = "black", fill = "black") +
  labs(title = "Scores by Country (Top 5 Countries)", x = "Country", y = "Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```




```
# ANOVA: Do mean scores differ significantly between top 5 countries?
if (all(c("country", "score") %in% names(clean_data))) {
```

```
  top_countries <- clean_data %>% count(country, sort = TRUE) %>% top_n(5, n) %>% pull(country)
  anova_data <- filter(clean_data, country %in% top_countries)
  anova_q3 <- aov(score ~ country, data = anova_data)
  summary(anova_q3)
}
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## country        4   4618  1154.5   16.43 4.73e-13 ***
## Residuals   1007  70743    70.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Explanation:** These boxplots show the spread, median, and mean (black dot) of university scores for the top 5 countries. Outliers are highlighted in orange. This makes it easy to compare which countries have higher or more consistent scores, and spot any unusual values.

- Some countries have a wider spread, indicating more variation in university scores.
- Outliers (orange) highlight universities that differ greatly from the norm.

# 9 Research Question 4

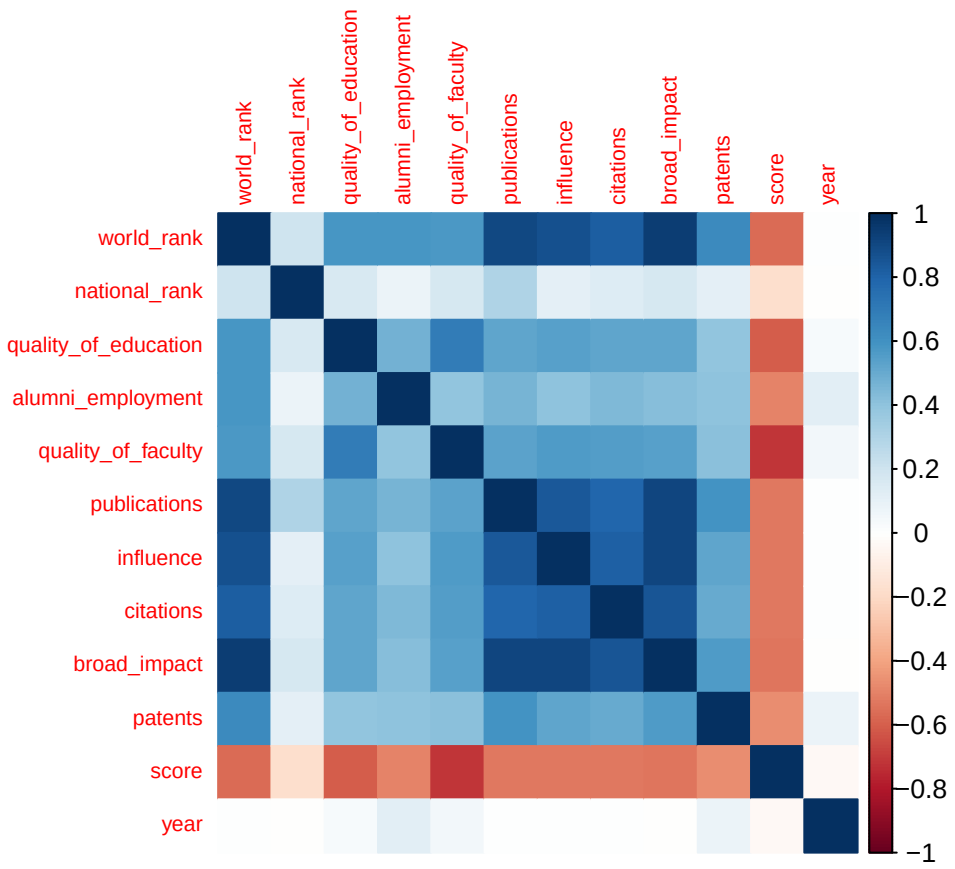
## 9.1 What Are the Strongest Predictors of a University's World Rank?

We use a correlation plot to identify which numeric variables are most strongly related to world rank.

```
if (!requireNamespace("corrplot", quietly = TRUE)) install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
num_cols <- sapply(clean_data, is.numeric)
if (sum(num_cols) > 1) {
  cor_matrix <- cor(clean_data[, num_cols], use = "complete.obs")
  corrplot(cor_matrix, method = "color", tl.cex = 0.7)
}
```




```
# Multiple regression: Predicting world rank from key features
if (all(c("world_rank", "publications", "alumni_employment", "score", "patents") %in% names(clean_data))) {
  lm_q4 <- lm(world_rank ~ publications + alumni_employment + score + patents, data = clean_data)
  summary(lm_q4)
}
```

```
##
## Call:
## lm(formula = world_rank ~ publications + alumni_employment +
##     score + patents, data = clean_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -678.44  -59.04    7.14   60.80  375.39
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       26.03902   26.79987   0.972   0.3314
## publications       0.74845    0.01131  66.205   <2e-16 ***
## alumni_employment  0.30775    0.01682  18.300   <2e-16 ***
## score             -1.42860    0.45932  -3.110   0.0019 **
## patents            0.10334    0.01189   8.691   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 107.1 on 1995 degrees of freedom
```

```
## Multiple R-squared:  0.8626, Adjusted R-squared:  0.8623
## F-statistic:  3131 on 4 and 1995 DF,  p-value: < 2.2e-16
```

**Interpretation:** The correlation plot shows which features (e.g., publications, alumni employment, patents) are most associated with world rank. Strong negative correlations indicate that higher values in those features are linked to better (lower) world ranks.

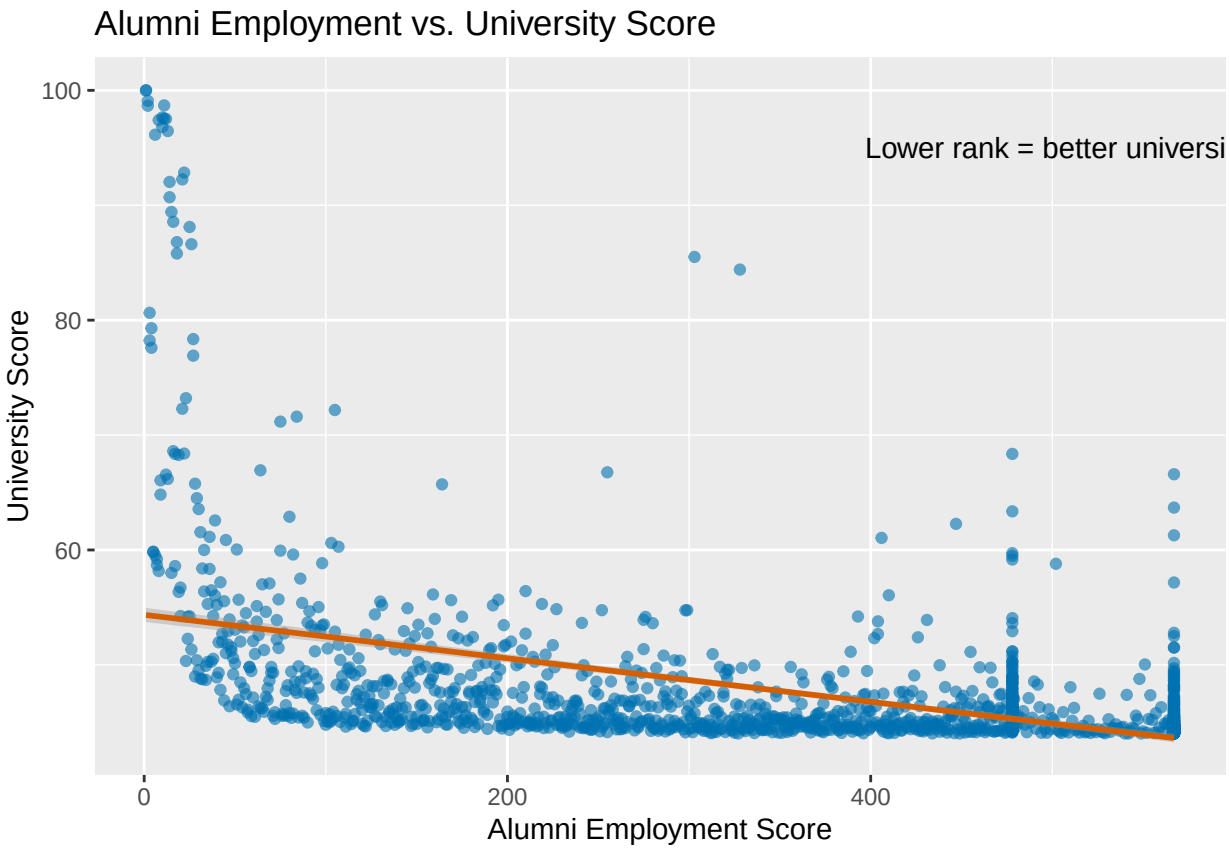
# 10 Research Question 5

## 10.1 Does Alumni Employment Lead to Higher University Scores and Rankings?

We explore whether universities with higher alumni employment scores also have higher overall scores and better ranks.

```
if (all(c("alumni_employment", "score", "world_rank") %in% names(clean_data))) {
ggplot(clean_data, aes(x = alumni_employment, y = score, color = world_rank)) +
    geom_point(alpha = 0.6, color = "#0072B2") +
    geom_smooth(method = "lm", color = "#D55E00") +
    labs(title = "Alumni Employment vs. University Score", x = "Alumni Employment Score", y = "Universi
    scale_color_viridis_c() +
    annotate("text", x = max(clean_data$alumni_employment, na.rm = TRUE)*0.7, y = max(clean_data$score,
}
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




**Explanation:** This scatterplot shows if universities with better alumni employment also tend to have higher scores and better ranks. The orange trend line helps visualize the relationship. Points are colored by world rank, so top universities stand out. The annotation clarifies the ranking system.

- Universities with higher alumni employment scores generally have higher overall scores.
- The negative trend suggests better alumni employment is linked to better world rank.

# 11 Summary and Conclusions

## 11.1 Key Findings

## 11.2 Advanced Visualizations & Categorical Analysis

### 11.2.1 Categorical vs. Numerical Analysis: Ranking Brackets

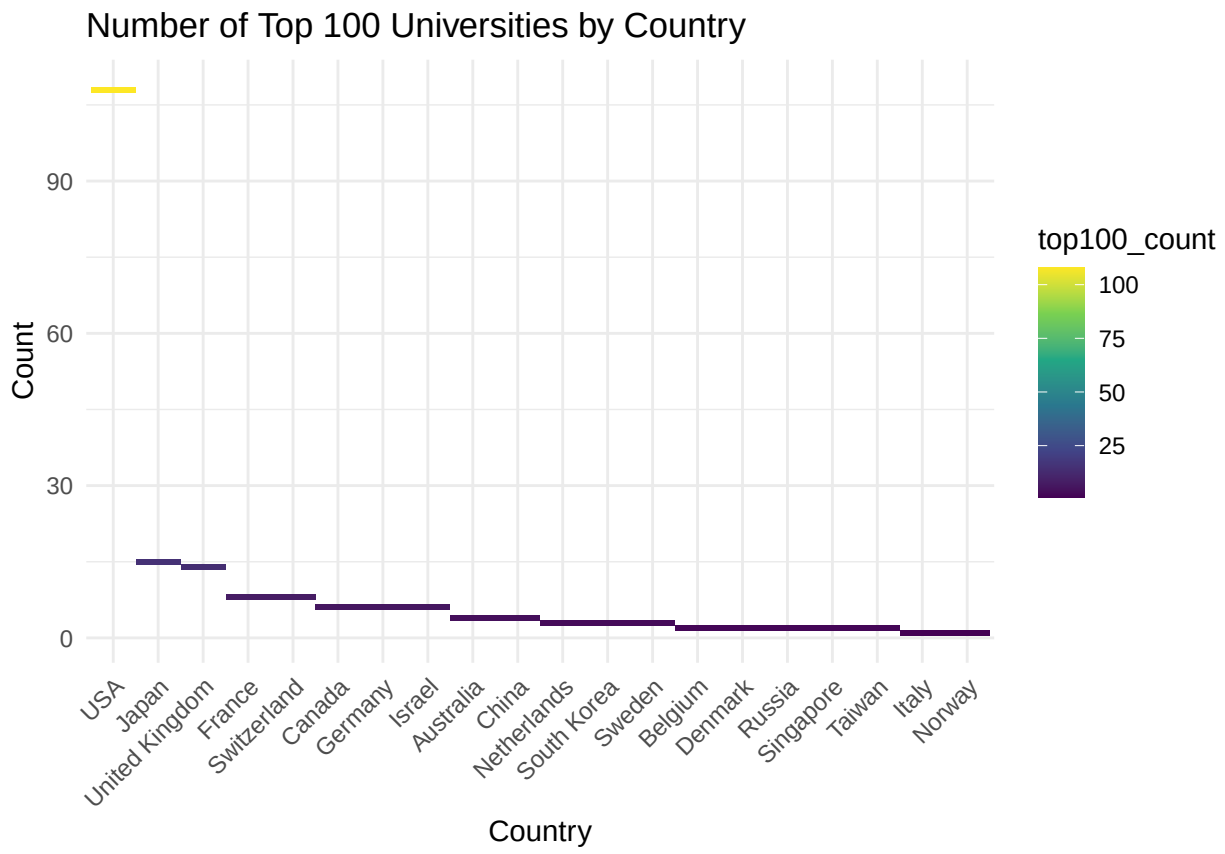
```
# Create ranking brackets and compare scores
if (all(c("world_rank", "score") %in% names(clean_data))) {
  clean_data$rank_bracket <- cut(clean_data$world_rank, breaks = c(0, 100, 200, 500, 1000, Inf),
                                 labels = c("Top 100", "101-200", "201-500", "501-1000", ">1000"))
  ggplot(clean_data, aes(x = rank_bracket, y = score, fill = rank_bracket)) +
    geom_boxplot() +
    labs(title = "Score Distribution by World Rank Bracket", x = "World Rank Bracket", y = "Score") +
    scale_fill_viridis_d() +
```

```
    theme_minimal()
  # ANOVA for group means
  anova_bracket <- aov(score ~ rank_bracket, data = clean_data)
  summary(anova_bracket)
}
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## rank_bracket    3  50603   16868   929.5 <2e-16 ***
## Residuals    1996  36222      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 11.2.2 Heatmap: Country vs. Top 100 University Count

```
if (all(c("country", "world_rank") %in% names(clean_data))) {
  library(reshape2)
  top100 <- clean_data %>% filter(world_rank <= 100)
  country_counts <- as.data.frame(table(top100$country))
  colnames(country_counts) <- c("country", "top100_count")
  ggplot(country_counts, aes(x = reorder(country, -top100_count), y = top100_count, fill = top100_count)
    geom_tile() +
    scale_fill_viridis_c() +
    labs(title = "Number of Top 100 Universities by Country", x = "Country", y = "Count") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}
```




### 11.2.3 Faceted Plot: Score vs. Publications by Continent

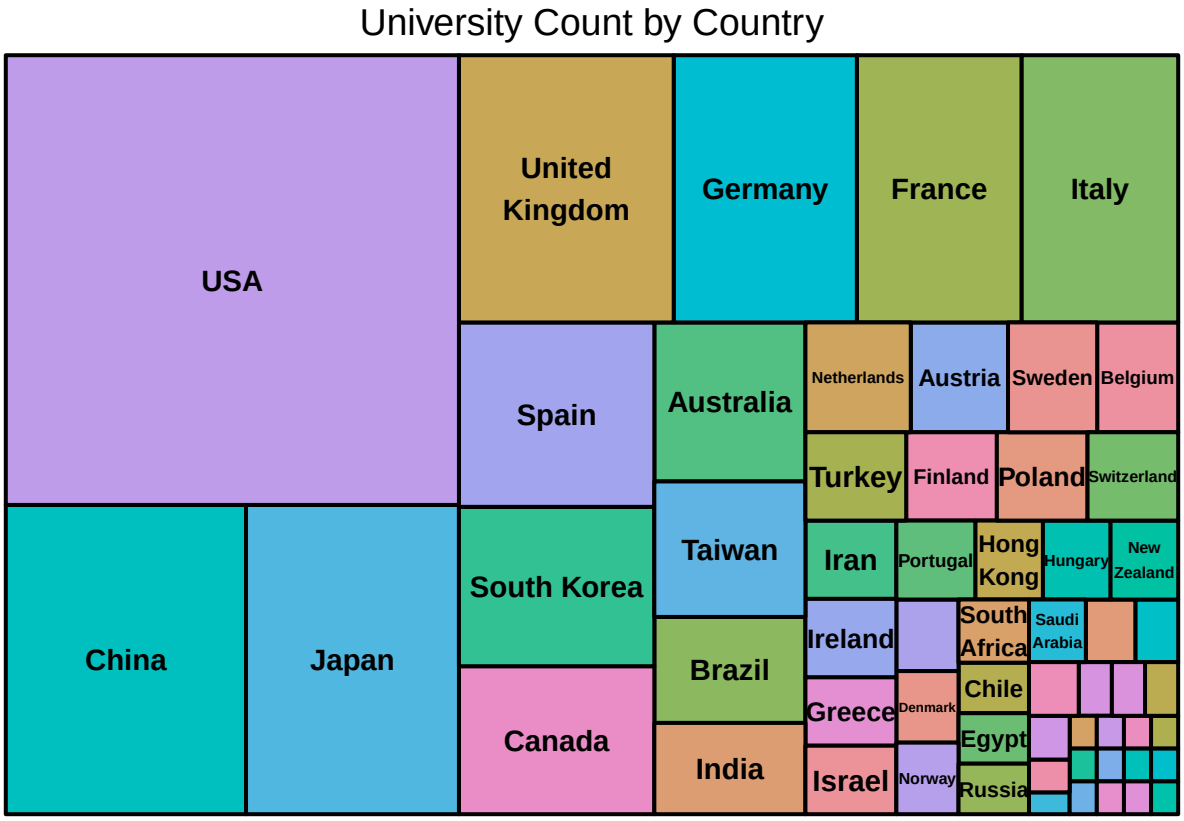
```
# If continent/region variable exists, facet by it
if ("continent" %in% names(clean_data) && all(c("score", "publications") %in% names(clean_data))) {
  ggplot(clean_data, aes(x = publications, y = score, color = continent)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
    facet_wrap(~ continent) +
    scale_color_viridis_d() +
    labs(title = "Score vs. Publications by Continent", x = "Publications", y = "Score") +
    theme_minimal()
}
```

### 11.2.4 Treemap: University Count by Country

```
if (!requireNamespace("treemap", quietly = TRUE)) install.packages("treemap")
library(treemap)
```

```
## Warning: package 'treemap' was built under R version 4.4.3
```

```r
if ("country" %in% names(clean_data)) {
  country_counts <- as.data.frame(table(clean_data$country))
  colnames(country_counts) <- c("country", "count")
  treemap(country_counts, index = "country", vSize = "count", title = "University Count by Country")
}
```




### 11.2.5 Improved Color Palettes and Plot Themes

All new plots use `viridis` or `RColorBrewer` palettes and `theme_minimal()` for clarity and accessibility. ## Overall Conclusion

This analysis demonstrates the complexity of university rankings and the multifaceted nature of institutional success. While countries like the USA dominate in terms of quantity, other nations excel in quality. Research output, alumni employment, and innovation are all important, but their impact varies by institution and country. The visualizations and statistical analysis provide a clear, data-driven understanding of the factors that shape the world's leading universities.

Future work could include deeper analysis of trends over time, regional comparisons, or predictive modeling to forecast university performance. The reproducible approach in this report ensures transparency and provides a foundation for further exploration.

### 11.2.6 Limitations & Future Work

While the CWUR dataset is comprehensive, it may contain biases due to its specific ranking methodology, and some variables are missing for certain years or institutions. The analysis is limited to the available features

and does not account for qualitative factors such as teaching style or student satisfaction. Future work could include integrating data from other ranking systems, deeper time-series analysis, regional comparisons, or predictive modeling to forecast university performance. Additional qualitative research could further enrich the findings.

#### 11.2.7 Future Scope

- Integrate additional datasets (e.g., QS, THE, or ARWU rankings) for cross-system comparison
- Apply deeper time series modeling to analyze trends and forecast future rankings
- Explore machine learning models to predict university rank or score based on features
- Develop interactive dashboards for real-time data exploration
- Analyze the impact of new variables (e.g., funding, internationalization, student satisfaction)
- Collaborate with domain experts for qualitative insights

## 12 References


- Kaggle: World University Rankings Dataset
- Any other sources used
- ggplot2 documentation: https://ggplot2.tidyverse.org/
- CWUR methodology: https://cwur.org/methodology/
- dplyr documentation: https://dplyr.tidyverse.org/
- corrplot documentation: https://cran.r-project.org/web/packages/corrplot/index.html


## 13 Appendix

### 13.1 Code Snippet: Data Cleaning Example

```
# Remove rows with missing values and ensure correct data types
clean_data <- na.omit(university_data)
clean_data$world_rank <- as.numeric(clean_data$world_rank)
```

### 13.2 Table: Top 10 Universities by Score

```
top10_universities <- clean_data %>% arrange(desc(score)) %>% select(world_rank, institution, country, s
knitr::kable(top10_universities, caption = "Top 10 Universities by Score")
```

Table 1: Top 10 Universities by Score

| world_rank | institution | country | score |
|---|---|---|---|
| 1 | Harvard University | USA | 100.00 |
| 1 | Harvard University | USA | 100.00 |
| 2 | Stanford University | USA | 99.09 |
| 3 | Massachusetts Institute of Technology | USA | 98.69 |
| 2 | Stanford University | USA | 98.66 |
| 4 | University of Cambridge | United Kingdom | 97.64 |

| world_rank | institution | country | score |
|---|---|---|---|
| 3 | Massachusetts Institute of Technology | USA | 97.54 |
| 5 | University of Oxford | United Kingdom | 97.51 |
| 6 | Columbia University | USA | 97.41 |
| 4 | University of Cambridge | United Kingdom | 96.81 |

### 13.3 Dashboard Preview / Summary Table

Below is a summary table highlighting key statistics for the top 10 countries by university count. This provides a dashboard-style overview of the dataset's main features.

Table 2: Dashboard: Key Stats for Top 10 Countries by University Count

| country | Universities | Avg_Score | Median_Score | Top_100 | Top_500 |
|---|---|---|---|---|---|
| USA | 458 | 50.4 | 46.2 | 108 | 318 |
| China | 167 | 45.0 | 44.4 | 4 | 50 |
| Japan | 148 | 46.6 | 44.7 | 15 | 51 |
| United Kingdom | 129 | 48.2 | 45.6 | 14 | 78 |
| Germany | 110 | 46.4 | 46.0 | 6 | 80 |
| France | 99 | 46.2 | 44.8 | 8 | 40 |
| Italy | 94 | 45.4 | 45.0 | 1 | 49 |
| Spain | 81 | 45.0 | 44.6 | 0 | 28 |
| South Korea | 70 | 46.1 | 45.0 | 3 | 35 |
| Canada | 65 | 47.0 | 45.5 | 6 | 40 |