US College Ranking Analysis

Moni Nguyen

2023-07-06

Let read some data:

df <- read.csv("C:/Users/nguye/Downloads/2022 US College Rankings.csv")

Let take a quick look at the data set:

head(df)

## College.Name Adjusted.Rank Tuition  
## 1 Princeton University 1 56010  
## 2 Columbia University 2 63530  
## 3 Harvard University 2 55587  
## 4 Massachusetts Institute of Technology 2 55878  
## 5 Yale University 5 59950  
## 6 Stanford University 6 56169  
## Enrollment.Numbers  
## 1 4773  
## 2 6170  
## 3 5222  
## 4 4361  
## 5 4703  
## 6 6366

Let find out if the data set have any missing value:

num\_missing\_college.name <- sum(is.na(df$College.Name))  
print(num\_missing\_college.name)

## [1] 0

num\_missing\_rank <- sum(is.na(df$Adjusted.Rank))  
print(num\_missing\_rank)

## [1] 0

num\_missing\_tuition <- sum(is.na(df$Tuition))  
print(num\_missing\_tuition)

## [1] 0

num\_missing\_enrollment <- sum(is.na(df$Enrollment.Numbers))  
print(num\_missing\_enrollment)

## [1] 0

Hence, there are no missing values.

Here is the summary of the data:

summary(df)

## College.Name Adjusted.Rank Tuition Enrollment.Numbers  
## Length:161 Min. : 1.00 Min. : 6120 Min. : 901   
## Class :character 1st Qu.: 38.00 1st Qu.:33928 1st Qu.: 6114   
## Mode :character Median : 79.00 Median :44196 Median :11612   
## Mean : 77.96 Mean :44110 Mean :16057   
## 3rd Qu.:117.00 3rd Qu.:55776 3rd Qu.:24505   
## Max. :148.00 Max. :63530 Max. :63124

The top 5 highest ranked college and the top 5 lowest ranked:

asc\_ranked <- head(df[order(df$Adjusted.Rank),],5)  
print(asc\_ranked)

## College.Name Adjusted.Rank Tuition  
## 1 Princeton University 1 56010  
## 2 Columbia University 2 63530  
## 3 Harvard University 2 55587  
## 4 Massachusetts Institute of Technology 2 55878  
## 5 Yale University 5 59950  
## Enrollment.Numbers  
## 1 4773  
## 2 6170  
## 3 5222  
## 4 4361  
## 5 4703

desc\_ranked <- head(df[order(-df$Adjusted.Rank),],5)  
print(desc\_ranked)

## College.Name Adjusted.Rank Tuition Enrollment.Numbers  
## 148 Colorado State University 148 31540 25186  
## 149 CUNY--City College 148 19010 12587  
## 150 Duquesne University 148 43526 5383  
## 151 George Mason University 148 36579 27105  
## 152 Michigan Technological University 148 38112 5642

The college with the highest, median and lowest tuition are:

highest\_tuition <- df$College.Name[which.max(df$Tuition)]  
college\_median <- df$College.Name[df$Tuition == median(df$Tuition)]  
lowest\_tuition <- df$College.Name[which.min(df$Tuition)]  
  
print(highest\_tuition)

## [1] "Columbia University"

print(college\_median)

## [1] "University of California--Santa Barbara"

print(lowest\_tuition)

## [1] "Brigham Young University--Provo"

The college with the highest, median and lowest enrollment are:

highest\_enrollment <- df$College.Name[which.max(df$Enrollment.Numbers)]  
median\_enrollment <- df$College.Name[df$Enrollment.Numbers == median(df$Enrollment.Numbers)]  
lowest\_enrollment <- df$College.Name[which.min(df$Enrollment.Numbers)]  
  
print(highest\_enrollment)

## [1] "Arizona State University"

print(median\_enrollment)

## [1] "Loyola University Chicago"

print(lowest\_enrollment)

## [1] "California Institute of Technology"

The correlation between tuition and enrollment:

correlation <- cor(df$Tuition,df$Enrollment.Numbers)  
print(correlation)

## [1] -0.5558214

Interpreting the correlation coefficient:

The negative sign indicates an inverse relationship, meaning that as tuition increases, enrollment tends to decrease, and vice versa.

Conclusions:

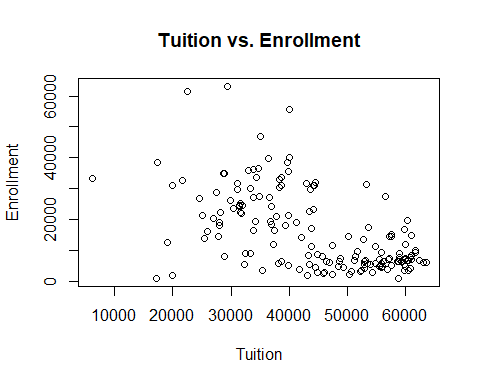
Colleges with higher tuition fees tend to have lower enrollment numbers.As tuition increases, the demand for enrollment may decrease, possibly due to affordability concerns.However, it’s important to note that correlation does not imply causation. Other factors not considered in the analysis could influence enrollment numbers.

Let install the ggplot2 for visualization:

library(ggplot2)

Let create a scatter plot for the correlation between tuition and enrollment:

plot(df$Tuition, df$Enrollment.Numbers,xlab = "Tuition",ylab = "Enrollment",main = "Tuition vs. Enrollment")



By observing the scatter plot, we can draw the conclusion that tuition is not an ultimate factor that influence enrollment number but it can be other factors such as reputation, scholarship and location.