# Project--1 Academic Landscapes: A Comparative Analysis of College Metrics Across Regions and Funding Models

Mihir Thakkar

## TASK-1 Getting Started with the Project

subtask-1

To explain the addition of vertical lines or other plot options within an RMarkdown document,include comments withincode chunk or text outside of it, describing the purpose and effect of these additions:

subtask-2

I have created a .html file using Rmarkdown, Using knit command on the top is used to open a project file and code should be saved in .Rmd file

subtask-3

Difference between .pdf and .RMD is .pdf contains all the data which can be used to present as a project, while .RMD file is a code file.

## TASK-2 : Converting Variables

In preparing the data for analysis, several variables were converted to factors, which are categorical variables that R can handle more effectively for statistical modelling and visualization. Specifically, `FundingModel`, `Region`, and `Geography` were converted to factors to accurately represent the distinct categories within each. Additionally, `HighestDegree` was converted to an ordered factor to reflect the inherent order from "Bachelor's" to "Graduate" degrees. These conversions ensure that subsequent analyses correctly interpret the nature of these variables.

```
# Loading the dataset
data <- read.csv("C:/Users/mihir/OneDrive/Desktop/SDS/Spring'24/STAT-3021/college_sample.csv")

# Converting FundingModel, Region, and Geography to factors
data$FundingModel <- as.factor(data$FundingModel)
data$Region <- as.factor(data$Region)
data$Geography<- as.factor(data$Geography)

# Convert HighestDegree to an ordered factor
data$HighestDegree <- factor(data$HighestDegree, levels = c("Bachelor's", "Graduate"), ordered = TRUE)
```

## TASK-3 : Analyzing Admission Rates

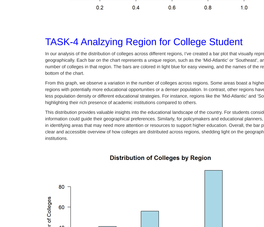
Graph Descriptions:

Histogram: The histogram of admission rates shows how frequently various admission rates occur across colleges. The shape of the histogram—whether it's symmetric, skewed left or right—can give us insights into the overall distribution of admission rates. For example, a right-skewed histogram would indicate that most colleges have lower admission rates, with fewer colleges being highly selective.

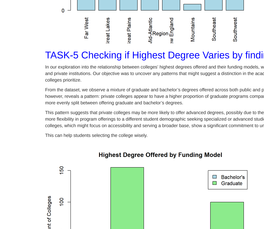
Boxplot: The boxplot provides a visual summary of the admission rates' distribution, highlighting the median, quartiles, and potential outliers. Outliers, represented by dots outside the 'whiskers,' are colleges with exceptionally high or low admission rates compared to the rest.

Statistics: When we explore college admission rates, we find an average rate that gives us a general idea of how selective colleges are overall. The median admission rate helps us see the selectivity level right in the middle, avoiding extreme values. The standard deviation shows us how much these admission rates vary—whether most colleges are similarly selective or if there's a wide range. The interquartile range focuses on the middle chunk of colleges, telling us about the spread of selectivity for the majority, excluding the most and least selective colleges. Together, these figures help us understand both the typical selectivity of colleges and how varied this selectivity is.

**Histogram of Admission Rates**




**Boxplot of Admission Rates**



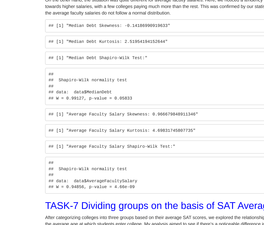

## TASK-4 Analzying Region for College Student

In our analysis of the distribution of colleges across different regions, I've created a bar plot that visually represents how colleges are spread geographically. Each bar on the chart represents a unique region, such as the 'Mid-Atlantic' or 'Southeast,' and the height of the bar shows the number of colleges in that region. The bars are colored in light blue for easy viewing, and the names of the regions are clearly labeled along the bottom of the chart.

From this graph, we observe a variation in the number of colleges across regions. Some areas boast a higher concentration of colleges, indicating regions with potentially more educational opportunities or a denser population. In contrast, other regions have fewer colleges, which might reflect less population density or different educational strategies. For instance, regions like the 'Mid-Atlantic' and 'Southeast' may show taller bars, highlighting their rich presence of academic institutions compared to others.

This distribution provides valuable insights into the educational landscape of the country. For students considering where to apply for college, this information could guide their geographical preferences. Similarly, for policymakers and educational planners, understanding this distribution helps in identifying areas that may need more attention or resources to support higher education. Overall, the bar plot and its underlying data offer a clear and accessible overview of how colleges are distributed across regions, shedding light on the geographical diversity of higher education institutions.

**Distribution of Colleges by Region**




## TASK-5 Checking if Highest Degree Varies by finding Model

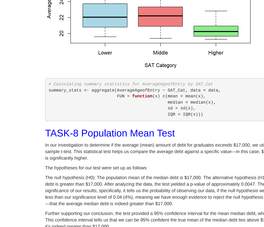
In our exploration into the relationship between colleges' highest degrees offered and their funding models, we delved into data comparing public and private institutions. Our objective was to uncover any patterns that might suggest a distinction in the academic programs these two groups of colleges prioritize.

From the dataset, we observe a mixture of graduate and bachelor's degrees offered across both public and private colleges. A closer look, however, reveals a pattern: private colleges appear to have a higher proportion of graduate programs compared to public institutions, which are more evenly split between offering graduate and bachelor's degrees.

This pattern suggests that private colleges may be more likely to offer advanced degrees, possibly due to their funding structure, which might allow more flexibility in program offerings or to a different student demographic seeking specialized or advanced studies. On the other hand, public colleges, which might focus on accessibility and serving a broader base, show a significant commitment to undergraduate education as well.

This can help students selecting the college wisely.

**Highest Degree Offered by Funding Model**




## TASK-6 Checking how two variables `MedianDebt` and `AveragefacultySalary` are distributed

### Median Debt

The skewness value for median debt among colleges is approximately -0.14, indicating a very slight skew to the left, but it's close enough to 0 to suggest that the distribution is fairly symmetrical. The kurtosis value is approximately -0.92, which is less than the typical kurtosis value of 3 for a normal distribution, suggesting that the distribution of median debt is somewhat flatter across the middle than a normal distribution. This means that there are fewer extreme values (very high or very low debts) than what we might expect in a normal distribution.

The Shapiro-Wilk test for median debt, which checks whether the data are normally distributed, gives a p-value of about 0.058. Since this p-value is greater than our significance level of 0.04 (4%), we do not have enough evidence to reject the null hypothesis that the data are normally distributed. In simpler terms, the distribution of median debt might be close enough to normal for statistical purposes, although it's on the borderline given our chosen significance level.

### Average Faculty Salary

For average faculty salary, the skewness value is approximately 0.96, indicating a moderate right skew. This means that the distribution has a longer tail on the right side, with more colleges having higher-than-average faculty salaries than lower. The kurtosis value is about 1.67, higher than for median debt, indicating a somewhat more "peaked" distribution than a normal distribution but still not excessively so.

The Shapiro-Wilk test for average faculty salary gives a p-value of approximately 4.66e-09, which is much less than 0.04. This very small p-value strongly suggests that the distribution of average faculty salaries is not normal, with the skewness and kurtosis values confirming that the distribution is both skewed to the right and more peaked than a normal distribution.

### Explanation

In our study of colleges, we looked at two key financial aspects: the median debt that students are left with and the average salary that faculty members earn. Our goal was to understand how these values are distributed—whether they follow a typical 'normal' pattern or not.

When we examined the median debt, our findings suggest a distribution that's quite symmetrical and only slightly flatter than the typical bell curve we'd expect if things were perfectly normal. The statistical test we used, known as the Shapiro-Wilk test, further backed this up by indicating that the median debt's distribution might not be significantly different from normal, although it was a close call.

On the other hand, the situation was quite different for average faculty salaries. Here, we noticed a tendency for the distribution to stretch more towards higher salaries, with a few colleges paying much more than the rest. This was confirmed by our statistical test, which clearly showed that the average faculty salaries do not follow a normal distribution.

```
## [1] "Median Debt Skewness: -0.141869991903"
```

```
## [1] "Median Debt Kurtosis: 2.51954194152644"
```

```
## [1] "Median Debt Shapiro-Wilk Test:"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$MedianDebt
## W = 0.99127, p-value = 0.05833
```

```
## [1] "Average Faculty Salary Skewness: 0.966678949011346"
```

```
## [1] "Average Faculty Salary Kurtosis: 4.49831745807735"
```

```
## [1] "Average Faculty Salary Shapiro-Wilk Test:"
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$AveragefacultySalary
## W = 0.94856, p-value = 4.66e-09
```

## TASK-7 Dividing groups on the basis of SAT Average Score

After categorizing colleges into three groups based on their average SAT scores, we explored the relationship between these SAT categories and the average age at which students enter college. My analysis aimed to see if there's a noticeable difference in entry age among students attending colleges with varying SAT score requirements.

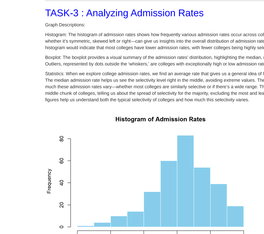
We used side-by-side boxplots to visually compare the Average Age of Entry across the SAT score categories (Lower', 'Middle', 'Higher'). These boxplots allowed us to observe not only median of ages in each category but also how widely, and whether there are any outliers (unusually high or low ages).

The summary statistics provided additional insights, quantifying the average (mean), median, standard deviation (a measure of spread), and interquartile range (middle 50% of ages) for each SAT category. These metrics helped us understand the typical age of entry and the age diversity within each SAT score group.

This explanation interprets the results of the boxplot and summary statistics in a way that highlights potential differences in the average age of entry among colleges with different SAT score requirements. The actual insights and conclusions would depend on the specific patterns observed in your dataset's analysis.

```
# Initialize the SAT_Cat column
data <- read.csv("C:/Users/mihir/OneDrive/Desktop/SDS/Spring'24/STAT-3021/college_sample.csv")
data$SAT_Cat <- NA

# Directly assign categories based on conditions
data$SAT_Cat[data$SATAverage < 970] <- "Lower"
data$SAT_Cat[data$SATAverage >= 970 & data$SATAverage < 1150] <- "Middle"
data$SAT_Cat[data$SATAverage >= 1150] <- "Higher"

# Convert the categories into an ordered factor
data$SAT_Cat <- factor(data$SAT_Cat, levels = c("Lower", "Middle", "Higher"), ordered = TRUE)



# Creating side-by-side boxplots for AverageAgeOfEntry by SAT_Cat
boxplot(data$AverageAgeOfEntry ~ data$SAT_Cat, data = data,
        main = "Average Age of Entry by SAT Category",
        xlab = "SAT Category", ylab = "Average Age of Entry",
        col = c("lightblue", "lightcoral", "lightgreen"))
```

**Average Age of Entry by SAT Category**




```
# Calculating summary statistics for AverageAgeOfEntry by SAT_Cat
summary_stats <- aggregate(AverageAgeOfEntry ~ SAT_Cat, data = data,
                           FUN = function(x) c(mean = mean(x),
                                               median = median(x),
                                               sd = sd(x),
                                               IQR = IQR(x)))
```

## TASK-8 Population Mean Test

In our investigation to determine if the average (mean) amount of debt for graduates exceeds $17,000, we utilized a method known as a one-sample t-test. This statistical test helps us compare the average debt against a specific value—in this case, $17,000—to see if the actual average is significantly higher.

The hypotheses for our test were set up as follows:

The null hypothesis (H0): The population mean of the median debt is $17,000. The alternative hypothesis (H1): The population mean of the median debt is greater than $17,000. After analyzing the data, the test yielded a p-value of approximately 0.0047. The p-value helps us determine the significance of our results; specifically, it tells us the probability of observing our data, if the null hypothesis were true. Here, a p-value of 0.0047 is less than our significance level of 0.04 (4%), meaning we have enough evidence to reject the null hypothesis and accept the alternative hypothesis —that the average median debt is indeed greater than $17,000.

Further supporting our conclusion, the test provided a 95% confidence interval for the mean median debt, which ranges from $17,239.69 to infinity. This confidence interval tells us that we can be 95% confident the true mean of the median debt lies above $17,239.69, reinforcing our finding that it's indeed greater than $17,000.

In summary, our analysis indicates that the average debt burden for graduates exceeds $17,000, a finding that may prompt further discussion on the financial challenges faced by students and the need for policies to alleviate student debt.

```
##
##  One Sample t-test
##
## data:  data$MedianDebt
## t = 2.6169, df = 315, p-value = 0.004651
## alternative hypothesis: true mean is greater than 17000
## 95 percent confidence interval:
##  17239.69      Inf
## sample estimates:
## mean of x
##  17846.32
```

```
## [1] 17239.69     Inf
## attr(,"conf.level")
## [1] 0.95
```

## TASK-9 Difference of Mean of `AverageFacultySalary` and `FundingModel`

Null hypothesis (H0): There is no difference in the mean AverageFacultySalary between private and public colleges. Alternative hypothesis (H1): There is a difference in the mean AverageFacultySalary between private and public colleges.

### Results and Decision

The t-test produced a p-value of 0.001531. In the context of our analysis, the p-value represents the probability of observing the data we have if the null hypothesis were true. Given that our significance level for this test was set at 3% (0.03), and our p-value is well below this threshold, we have strong evidence to reject the null hypothesis.

This means that we found statistically significant evidence suggesting that the average faculty salary is indeed different between private and public colleges. The direction of this difference is indicated by the sample estimates: the mean salary in private colleges is approximately $7,101.71, while it is about $7,810.96 in public colleges.

### Confidence Interval

The 97 percent confidence interval for the difference in means between the two groups ranges from -1182.74 to -225.77. This interval does not include 0, which supports our conclusion that there is a significant difference in average faculty salaries between the two funding models. The negative values of the confidence interval further indicate that public colleges tend to offer higher salaries on average compared to private colleges.

### Conclusion

To put it simply, our analysis provides clear evidence that public colleges offer higher average salaries to their faculty members than private colleges. This difference in compensation could be an essential factor for educators considering their employment options and highlights the impact of the college's funding model on faculty salaries. The confidence interval gives us a range that we are 97% confident contains the true difference in average salaries, reinforcing the reliability of our findings.

```
##
##  Welch Two Sample t-test
##
## data:  AveragefacultySalary by FundingModel
## t = -3.1992, df = 289.94, p-value = 0.001531
## alternative hypothesis: true difference in means between group Private and group Public is not equal to 0
## 97 percent confidence interval:
##  -1182.7431  -225.7695
## sample estimates:
## mean in group Private  mean in group Public
##              7101.710              7810.966
```