Stats382-Proj1

Erica Castillo

2024-03-01

Load R Libraries

```
library(readr)
library(e1071)
```

Load Dataset

```
college_data <- read.csv("college_sample.csv")
```

TASK 1

Explanation 1

In RMarkdown, you can use chunk options to customize the behavior of the R code chunks, including plotting options. Here is an example of how you can use chunk options for plots, such as adding an additional vertical line:

Example: ```{r, echo=FALSE, fig.cap="Customized Plot", out.width="70%", fig.align="center", fig.height=5, fig.width=8} plot(pressure) abline(v = 300, col = "red", lty = 2)

In the example, I used chunk options:

- `echo=FALSE` : Prevents the code from displaying.
- `fig.cap="Customized Plot"` : Adds a caption to the plot.
- `out.width="70%"` : Sets the width of the plot to 70% of the text width.
- `fig.align="center"` : Aligns the plot to the center of the document.
- `fig.height=5` : Sets the height of the plot to 5 inches.
- `fig.width=8` : Sets the width of the plot to 8 inches.

These options give you control over the appearance and behavior of the code chunks and the resulting plots in your RMarkdown document. There are more options based on what display you want for your plot.

Explanation 2

In this project, I will be creating a PDF from an HTML file. The steps for this process are:

- Open the HTML file (from your computer files) in a web browser.
- Use the browser's "Print" option (usually found in the browser menu or by pressing Ctrl+P).
- In the print dialog, choose "Save as PDF" as the printer option.
- Then save the document as a PDF file.

Explanation 3

- The .RMD file (.Rmarkdown) is the source file containing both your written content and the R code chunks. It is essentially a plain text file that includes the narrative, code, and instructions on how the document should be rendered.
- The .PDF file is the output generated by knitting the .RMD file. It is a formatted document containing the final report with text, code outputs, and any visualizations or plots. The PDF serves as the polished and presentable version of your analysis.
- While the .RMD file is the editable source, the .PDF file is the static, shareable output that reflects the finalized report. The PDF is what others will see and read, while the .RMD file allows for further editing and modification.

TASK 2

```
college_data$FundingModel <- as.factor(college_data$FundingModel)
college_data$Region <- as.factor(college_data$Region)
college_data$Geography <- as.factor(college_data$Geography)

college_data$HighestDegree <- factor(college_data$HighestDegree,
                                     levels = c("Bachelor's", "Graduate"), ordered = TRU
E)
```

TASK 3

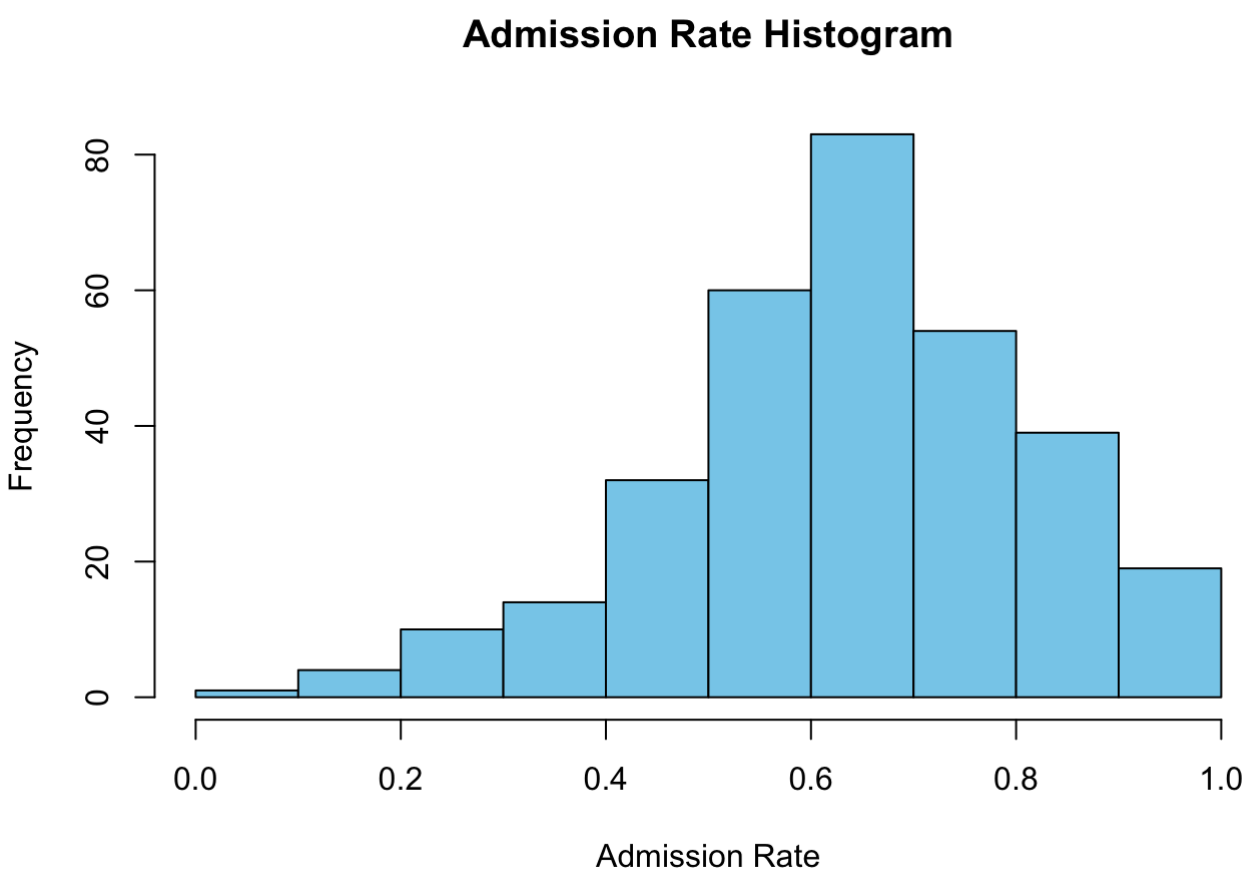
In our dataset, college_sample, the variable AdmissionRate reflects the number of applicants accepted by various colleges. In other words, the variable shows us the percentage of how many students applied and how many were accepted for specific colleges in the dataset. So, if a college has a high AdmissionRate, it means a lot of the students who apply there get accepted. On the other hand, if a college has a low AdmissionRate, it means that only a small percentage of students who apply actually get accepted. It measures how easy or difficult it is to get into a particular college.

```
# Check for NA values in AdmissionRate
sum(is.na(college_data$AdmissionRate))
```

```
## [1] 0
```

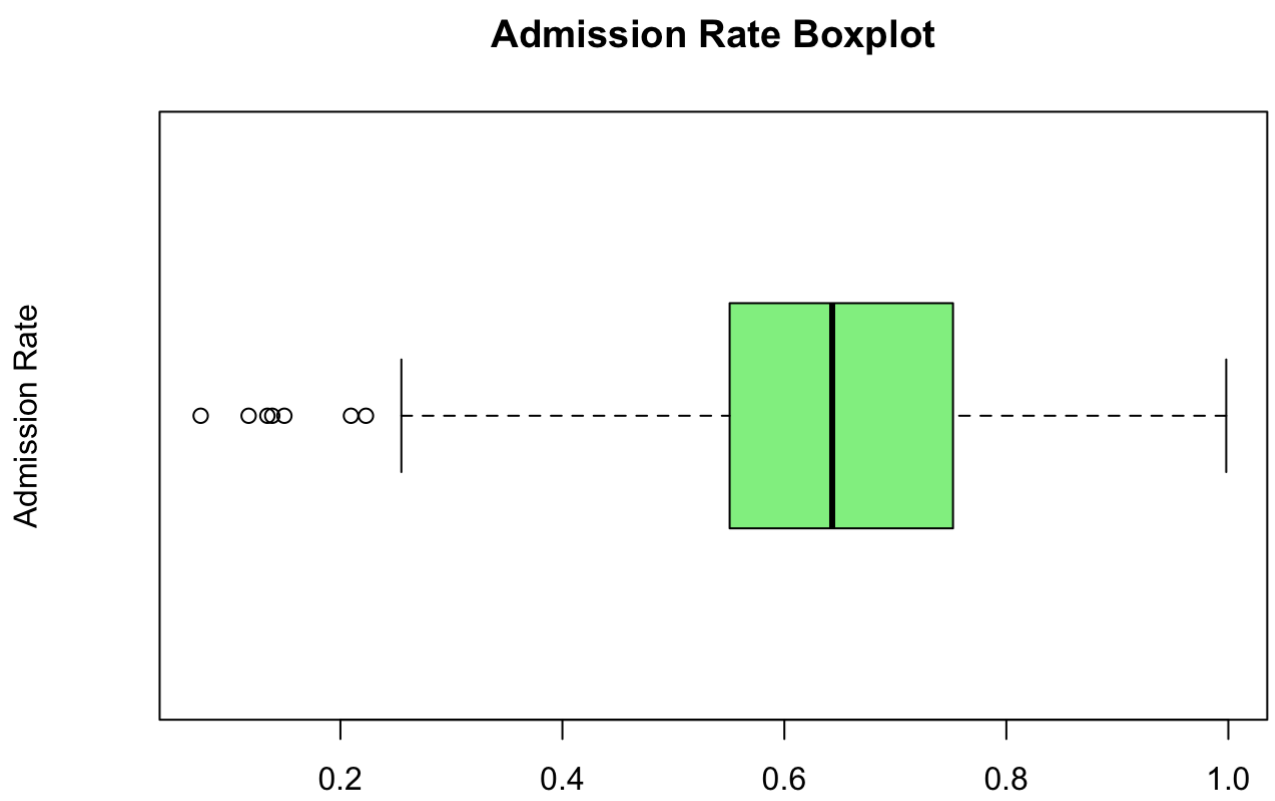
Relevant Graph: Histogram and Boxplot for AdmissionRate

```
hist(college_data$AdmissionRate, main = "Admission Rate Histogram",
     xlab = "Admission Rate", ylab = "Frequency", col = "skyblue", border = "black")
```



Based on our Admission Rate histogram, a left-skewed distribution implies that a substantial number of colleges are inclusive in their admissions. In other words, most colleges admit a larger percentage of applicants.

```
boxplot(college_data$AdmissionRate, main = "Admission Rate Boxplot",
        ylab = "Admission Rate", col = "lightgreen", border = "black", horizontal = TRU
E)
```



The boxplot complements the left-skewed histogram by providing a visual representation of the distribution's central tendency, spread, and the presence of outliers. The outliers represent colleges with extremely low admission rates which suggests that those colleges are highly selective in their admissions process. Otherwise, the central tendency indicates that most colleges are more inclusive, such as our histogram portrayed.

Calculate descriptive statistics for AdmissionRate

```
mean_admission_rate <- mean(college_data$AdmissionRate, na.rm = TRUE)
sd_admission_rate <- sd(college_data$AdmissionRate, na.rm = TRUE)

cat("Mean Admission Rate:", mean_admission_rate, "\n")
```

```
## Mean Admission Rate: 0.6385231
```

```
cat("Standard Deviation of Admission Rate:", sd_admission_rate, "\n")
```

```
## Standard Deviation of Admission Rate: 0.1722819
```

The mean admission rate being approximately 0.6385 indicates that, on average, colleges in the dataset admit around 63.85% of their applicants.

The standard deviation measures the spread or variability of admission rates across colleges. The standard deviation being approximately 0.1723 suggests that while there is some variability, the majority of colleges have admission rates within a reasonable range of the mean.

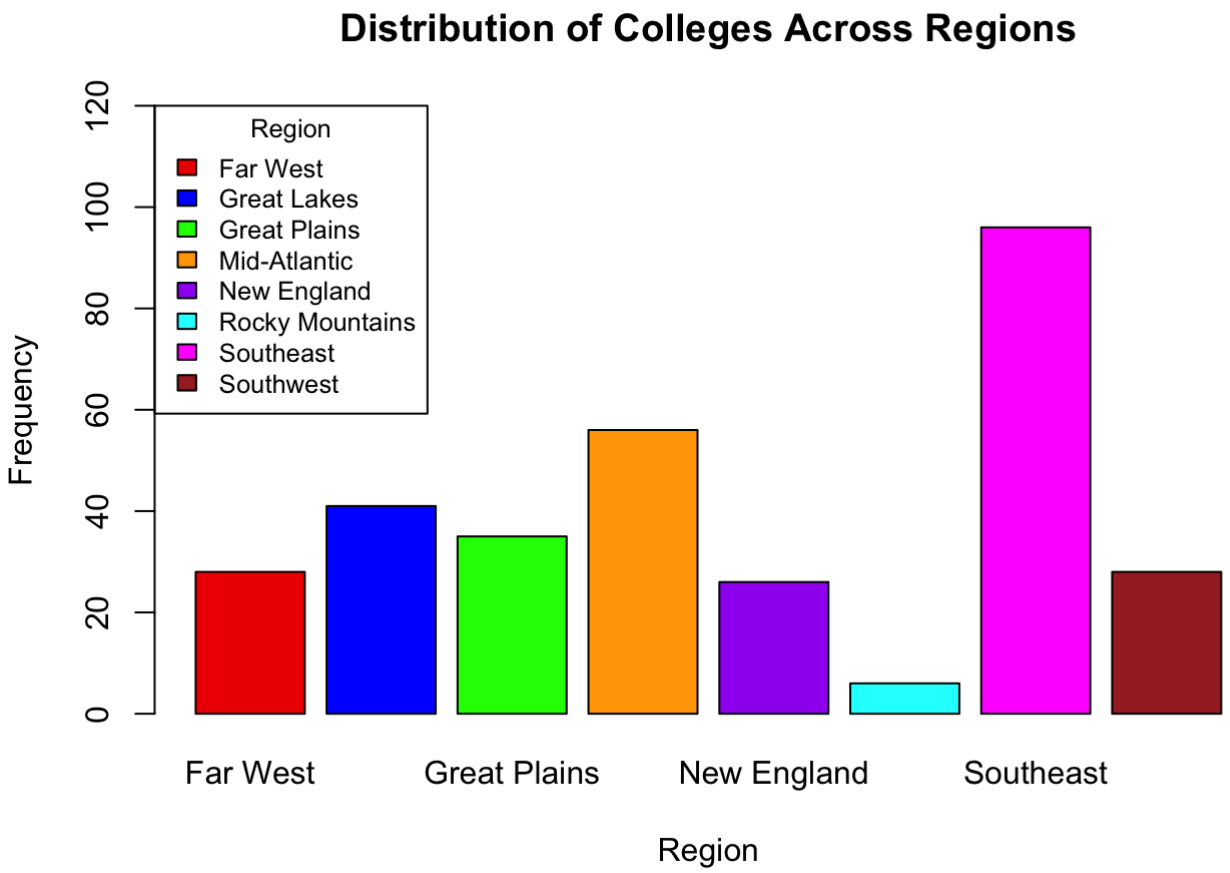
TASK 4

The “Region” variable categorizes colleges based on their geographic location within the college_sample dataset. The variable showcases the distribution of colleges across regions such as the West, Midwest, South, and Northeast. This information is really useful because it gives us a way to see how diverse the dataset is in terms of where the colleges are. It helps us see if there are any patterns or differences in the characteristics of colleges based on their location in different regions.

Relevant Graph: Barplot for Region

```
region_colors <- c("red2", "blue", "green", "orange", "purple", "cyan", "magenta", "brow
n")

barplot(table(college_data$Region),
        main = "Distribution of Colleges Across Regions",
        xlab = "Region",
        ylab = "Frequency",
        ylim = c(0,120),
        col = region_colors,
        border = "black",
        legend.text = levels(college_data$Region),
        args.legend = list(title = "Region", x = "topleft", cex = 0.8))
```




```
region_table <- table(college_data$Region)
print(region_table)
```

```
##
##        Far West     Great Lakes    Great Plains    Mid-Atlantic     New England
##              28              41              35              56              26
## Rocky Mountains       Southeast       Southwest
##               6              96              28
```

The barplot and frequency counts reveal the number of colleges within each region. This provides the dataset's geographical distribution such as the Southeast region which has the highest number of colleges (96), while the Rocky Mountains region has the lowest (6). Understanding these counts is essential for assessing the representation of colleges across different parts of the country. This is important because it gives us a sense of whether our information about colleges is balanced across the country. It allows us to see the spread of our college locations and how other parts of the data is linked to it. For example, most of our dataset information is coming from the Southeast region whereas the least amount of college information is coming from the Rocky Mountain region.

TASK 5

The HighestDegree variable is the academic degree offered by a college, categorized as either "Bachelor's" or "Graduate." The FundingModel is the funding structure of a college, categorized as either "Private" or "Public." To examine if the HighestDegree variable is varied by the FundingModel variable, I will create a side-by-side barplot

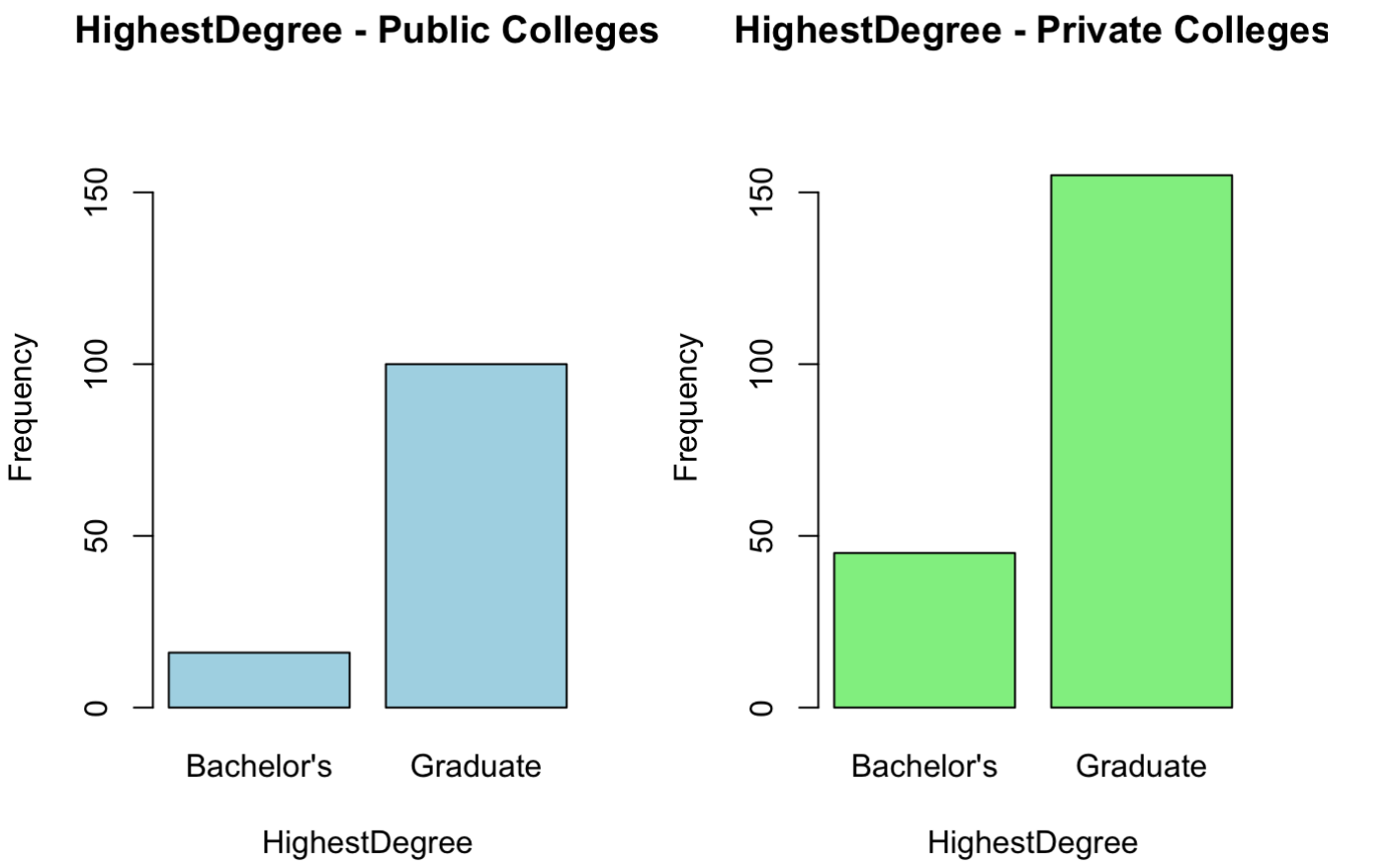
and computed summary statistics to bring a conclusion.

Relevant Data: Barplot / Contingency Table / Chi-Square Test

```
par(mfrow = c(1, 2))

barplot(table(college_data$HighestDegree[college_data$FundingModel == "Public"]),
        main = "HighestDegree - Public Colleges", ylim = c(0,175),
        xlab = "HighestDegree", ylab = "Frequency", col = "lightblue", border = "black")

barplot(table(college_data$HighestDegree[college_data$FundingModel == "Private"]),
        main = "HighestDegree - Private Colleges", ylim = c(0,175),
        xlab = "HighestDegree", ylab = "Frequency", col = "lightgreen", border = "blac
k")
```




```
degree_funding_table <- table(college_data$HighestDegree, college_data$FundingModel)
cat("Contingency Table:\n")
```

```
## Contingency Table:
```

```
print(degree_funding_table)
```

```
##
##              Private Public
##   Bachelor's      45     16
##   Graduate       155    100
```

Based on our barplot and contingency table, private colleges outnumber public colleges in both Bachelor's and Graduate degrees. In other words, private colleges have more Graduate programs. Although, from our barplot there isn't a signigicant difference when comparing private and public in their respective categories. To test for independence, I will conduct a Chi-square test.

```
chi_square_test <- chisq.test(degree_funding_table)
cat("\nChi-square Test for Independence:\n")
```

```
##
## Chi-square Test for Independence:
```

```
print(chi_square_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  degree_funding_table
## X-squared = 3.0359, df = 1, p-value = 0.08144
```

This test checks if there's a significant connection between the variables HighestDegree and FundingModel. The test result (X-squared) is 3.0359, with 1 degree of freedom and the p-value is 0.08144. Since the p-value (0.08144) is above 0.05, so there's no strong proof to say the differences are significant. Based on our plots and statistics, there is no strong evidence showing a big link between the highest degree offered and funding model. Therefore, I conclude that the variable FundingModel does not vary the variable HighestDegree.

TASK 6

The MedianDebt variable represents the median amount of student debt for graduates from a given college, providing a measure of the central tendency for student indebtedness. The AverageFacultySalary variable denotes the mean salary of faculty members at a college, serving as a measure of the typical compensation for academic staff. To find out if the variables are normally distributed, I will provide a histogram and QQ Plot and calculate the skew and kurtosis.

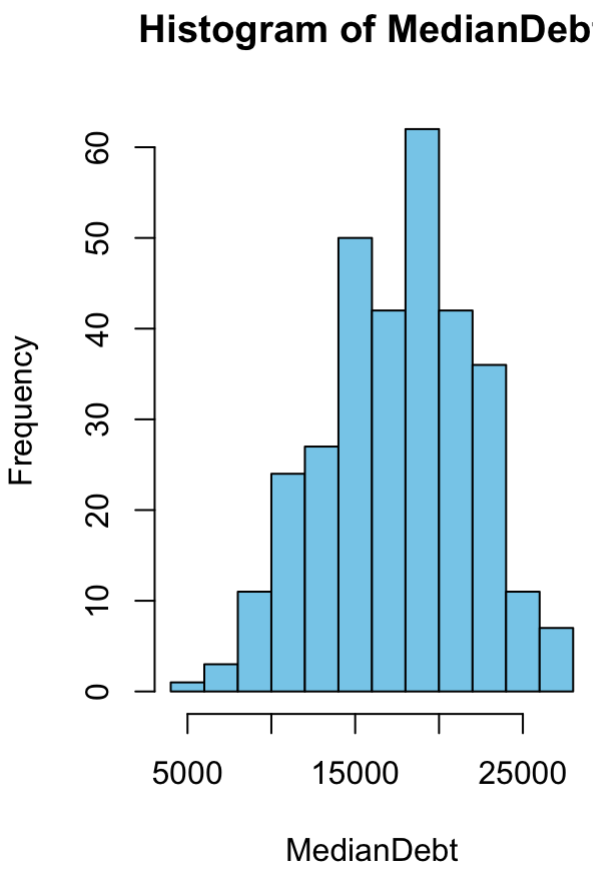
- Skewness measures the asymmetry of a distribution.
- A skewness value of 0 indicates a perfectly symmetrical distribution.
- A negative skewness suggests that the left tail of the distribution is longer or fatter than the right tail.
- A positive skewness suggests that the right tail of the distribution is longer or fatter than the left tail.
- Kurtosis measures the "tailedness" of a distribution.
- A kurtosis value of 0 indicates that the distribution has the same tail risk as a normal distribution.
- Negative kurtosis (platykurtic) indicates thinner tails than a normal distribution.

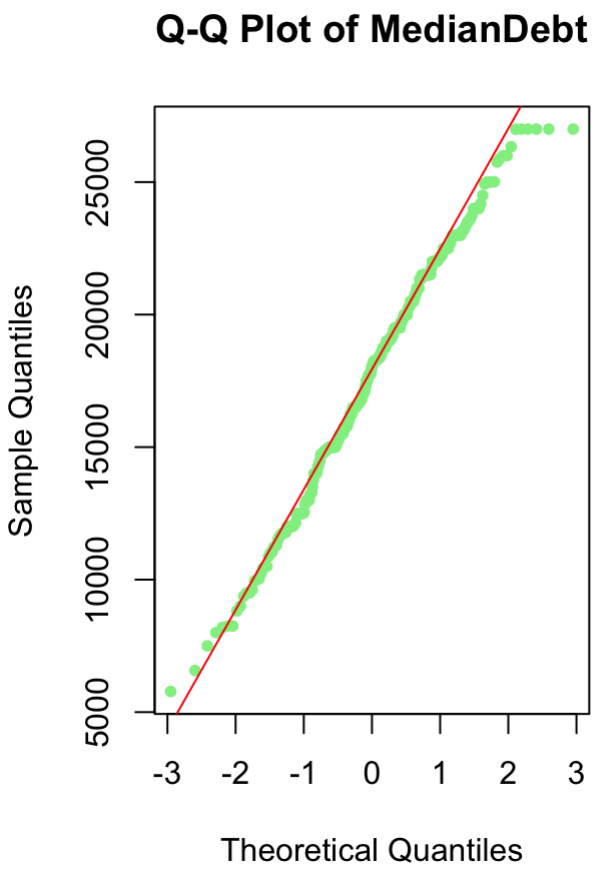
- A kurtosis value of 0 indicates that the distribution has the same tail risk as a normal distribution.
- Positive kurtosis (leptokurtic) indicates fatter tails than a normal distribution.

```
par(mfrow = c(1, 2))

hist(college_data$MedianDebt, main = "Histogram of MedianDebt",
     xlab = "MedianDebt", ylab = "Frequency", col = "skyblue", border = "black")

qqnorm(college_data$MedianDebt, main = "Q-Q Plot of MedianDebt", col = "lightgreen", pch
= 20)
qqline(college_data$MedianDebt, col = "red")
```






```
skew_median_debt <- skewness(college_data$MedianDebt)
kurt_median_debt <- kurtosis(college_data$MedianDebt)
cat("Skew for median debt: ", skew_median_debt, "\n")
```

```
## Skew for median debt:  -0.141197
```

```
cat("Kurt for median debt: ", kurt_median_debt, "\n")
```

```
## Kurt for median debt:  -0.4963793
```

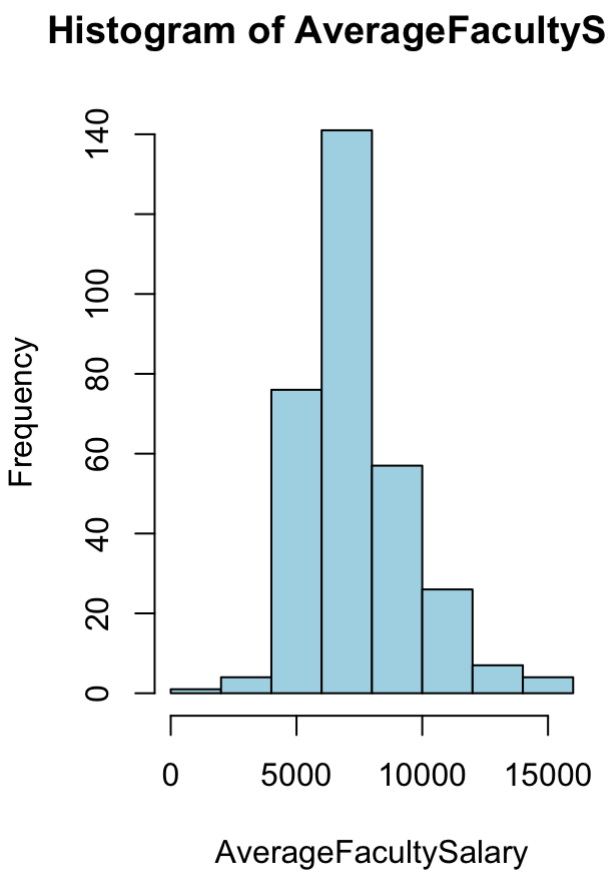
Median Debt Distribution Analysis:

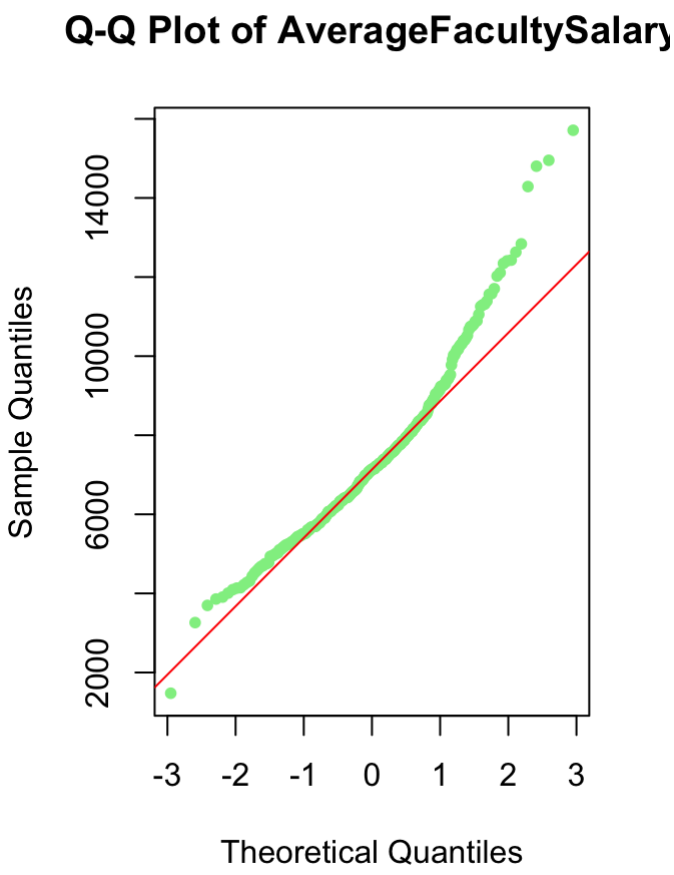
- Visual Assessments:
- Histogram: Slightly left-skewed but appears bell-shaped (implies normal).
- QQ Plot: Overall normal with slight abnormality towards the end, hinting at potential outliers.
- Skewness (-0.141197):
- Interpretation: Slightly left skew, matching the left-skewed histogram.
- Implication: The distribution is approximately normal, with a few colleges having less-than-average median debt, pulling the distribution to the left.
- Kurtosis (-0.4963793):
- Interpretation: Negative kurtosis indicates thinner tails than a normal distribution.
- Implication: The distribution of median debt has slightly thinner tails than normal, suggesting fewer extreme values or outliers.
- Overall Implications:
- The distribution of Median Debt is approximately normal, with a few potential outliers.
- Negative kurtosis implies fewer extreme values than a perfectly normal distribution.
- Summary:
- Median Debt distribution is generally normal with a slight left skew, indicating a concentration of colleges with lower median debt. The negative skewness and kurtosis values affirm the absence of extreme values, although outliers may be present. Visual assessments, including histograms and QQ plots, complement numerical measures for a comprehensive understanding.

```
par(mfrow = c(1, 2))

hist(college_data$AverageFacultySalary, main = "Histogram of AverageFacultySalary",
     xlab = "AverageFacultySalary", ylab = "Frequency", col = "lightblue", border = "bla
ck")

qqnorm(college_data$AverageFacultySalary, main = "Q-Q Plot of AverageFacultySalary", col
= "lightgreen", pch = 20)
qqline(college_data$AverageFacultySalary, col = "red")
```






```
skew_avg_faculty_salary <- skewness(college_data$AverageFacultySalary)
kurt_avg_faculty_salary <- kurtosis(college_data$AverageFacultySalary)
cat("Skew for average falculty salary: ", skew_avg_faculty_salary, "\n")
```

```
## Skew for average falculty salary:  0.9620948
```

```
cat("Kurt for average falculty salary: ", kurt_avg_faculty_salary, "\n")
```

```
## Kurt for average falculty salary:  1.668628
```

Average Faculty Salary Distribution Analysis:

- Visual Assessments:
- Histogram: Slightly left-skewed for MedianDebt and right-skewed with a heavier right tail for AverageFacultySalary.
- QQ Plot: While generally normal, some deviations suggest potential outliers in the higher salary range for AverageFacultySalary.
- Skewness (0.96):
- Interpretation: Positive skewness indicates a rightward shift in the distribution.

- Implication: Some colleges pay professors significantly more than the majority, creating a rightward tail.
- Kurtosis (1.67):
- Interpretation: Positive kurtosis suggests heavier tails and more extreme values.
- Implication: There are colleges with faculty salaries noticeably higher than the average, resulting in a distribution with more extreme values.
- Overall Implications:
- The skewness and kurtosis collectively suggest that Average Faculty Salary distribution deviates from perfect normality.
- Positive skewness indicates a concentration of higher salaries, and positive kurtosis points to more extreme values.
- Summary:
- The distribution of Average Faculty Salary is not perfectly normal, showing deviations, especially in higher salary ranges. The positive skewness and kurtosis highlight the presence of colleges with significantly different faculty salaries than the majority. While the distribution may not be ideal for certain analyses, further assessments are needed for a comprehensive understanding, considering visual tools and formal tests.

```
shapiro_median_debt <- shapiro.test(college_data$MedianDebt)
shapiro_median_debt
```

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

```
shapiro_avg_faculty_salary <- shapiro.test(college_data$AverageFacultySalary)
shapiro_avg_faculty_salary
```

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

Shapiro-Wilk Test Results:

For MedianDebt:

Hypotheses: H0: The distribution of MedianDebt is normal H1: The distribution of MedianDebt is not normal Test Statistic: 0.9912691 P-value: 0.05833025 Decision at 2% signigicance level: Fail to reject H0 (normal) since the p-value is greater than alpha. Conclusion: The distribution of MedianDebt is normal since the p-value is greater than alpha.

For AverageFacultySalary:

Hypotheses: H0: The distribution of AverageFacultySalary is normal H1: The distribution of AverageFacultySalary is not normal Test Statistic: 0.948556 P-value: 4.660238e-09 Decision at 2% signigicance level: Reject H0 (not normal) since the p-value is less than alpha. Conclusion: The distribution of AverageFacultySalary is not normal since the p-value is less than alpha.

TASK 7

```
college_data$SAT_Cat <- cut(college_data$SATAverage,
                            breaks = c(-Inf, 970, 1150, Inf),
                            labels = c("Lower", "Middle", "Higher"),
                            include.lowest = TRUE,
                            ordered_result = TRUE)

par(mfrow = c(1, 2))

boxplot(AverageAgeofEntry ~ SAT_Cat, data = college_data,
        main = "Boxplot of AverageAgeofEntry by SAT_Cat",
        xlab = "SAT Category", ylab = "Average Age of Entry",
        col = c("skyblue", "lightgreen", "lightcoral"),
        border = "black")

summary_stats <- tapply(college_data$AverageAgeofEntry, college_data$SAT_Cat,
                        function(x) c(mean = mean(x), sd = sd(x)))

lower_mean <- summary_stats$Lower["mean"]
lower_sd <- summary_stats$Lower["sd"]

middle_mean <- summary_stats$Middle["mean"]
middle_sd <- summary_stats$Middle["sd"]

higher_mean <- summary_stats$Higher["mean"]
higher_sd <- summary_stats$Higher["sd"]

cat("Lower SAT Category:\n",
    "- Mean Age of Entry:", lower_mean, "\n",
    "- Standard Deviation:", lower_sd, "\n\n")
```

```
## Lower SAT Category:
##  - Mean Age of Entry: 22.59459
##  - Standard Deviation: 2.395391
```

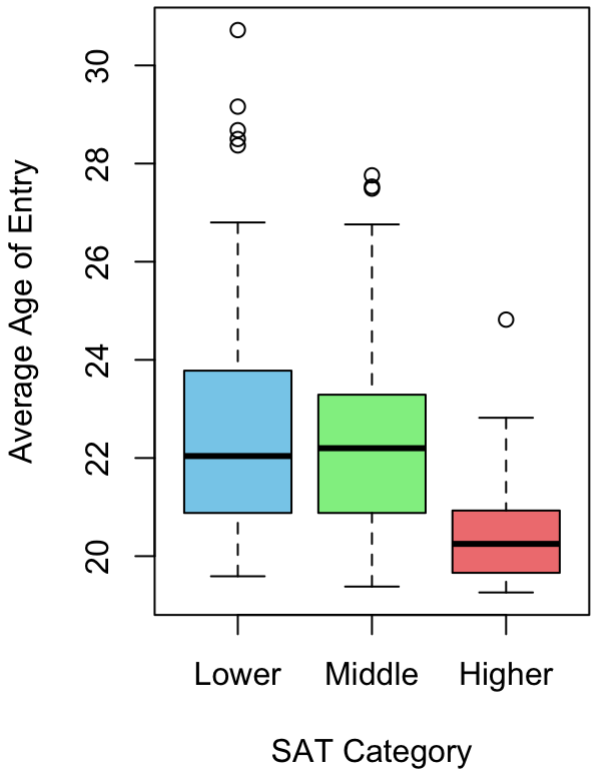
```
cat("Middle SAT Category:\n",
    "- Mean Age of Entry:", middle_mean, "\n",
    "- Standard Deviation:", middle_sd, "\n\n")
```

```
## Middle SAT Category:
##  - Mean Age of Entry: 22.27224
##  - Standard Deviation: 1.847986
```

```
cat("Higher SAT Category:\n",
    "- Mean Age of Entry:", higher_mean, "\n",
    "- Standard Deviation:", higher_sd, "\n")
```

```
## Higher SAT Category:
##  - Mean Age of Entry: 20.4982
##  - Standard Deviation: 1.075107
```

oxplot of AverageAgeofEntry by SA



The SAT category helps sort colleges based on how well students did on a standardized test, giving us an idea of the students academic performance. On the other hand, the Average Age of Entry tells us the average age when students usually start going to college, helping us understand how old most students are when they begin their college journey. To explore whether AverageAgeofEntry varies by SAT_Cat, I divided the schools into three categories based on SAT scores: “Lower” (SAT < 970), “Middle” (970 <= SAT < 1150), and “Higher” (SAT >= 1150).

The boxplots depict the distribution of AverageAgeofEntry within each SAT_Cat group. The graph shows potential variations in the center and spread of AverageAgeofEntry across the SAT categories. The “Lower” SAT group seems to have a slightly lower median age of entry compared to the other two categories.

Summary Statistics:

Lower SAT Category:

- Mean Age of Entry: 22.59
- Standard Deviation: 2.40
- Interpretation: The mean age of entry for colleges in the Lower SAT category is approximately 22.59, with a moderate standard deviation of 2.40. This suggests that there is some variability in the age of entry, but the majority of colleges in this category have an average entry age around 22.59.

Middle SAT Category:

- Mean Age of Entry: 22.27
- Standard Deviation: 1.85
- Interpretation: Colleges in the Middle SAT category have a mean age of entry of approximately 22.27, with a slightly lower standard deviation of 1.85. This indicates that the age of entry is relatively consistent across colleges in this category, with less variability compared to the Lower SAT category.

Higher SAT Category:

- Mean Age of Entry: 20.50
- Standard Deviation: 1.08
- Interpretation: The mean age of entry for colleges in the Higher SAT category is noticeably lower at 20.50, and the standard deviation is relatively low at 1.08. This suggests that colleges in this category tend to have a lower average age of entry, and there is less variability compared to the other categories.

Overall Conclusion:

The data provides evidence that there is a relationship between SAT categories and the Average Age of Entry. Colleges in the Higher SAT category tend to have a lower average age of entry with less variability, while colleges in the Middle and Lower SAT categories have higher average entry ages with varying levels of variability. The observed mean and standard deviation values support the conclusion that the SAT category is associated with differences in the age at which students enter college.

TASK 8

```
median_debt_test <- t.test(college_data$MedianDebt, mu = 17000, alternative = "greater",
conf.level = 0.96)

p_value_median_debt <- median_debt_test$p.value
conf_interval_median_debt <- median_debt_test$conf.int
decision_median_debt <- ifelse(p_value_median_debt < 0.04, "reject", "fail to reject")

cat("Test Statistic:", median_debt_test$statistic, "\n")
```

```
## Test Statistic: 2.616878
```

```
cat("P-value:", p_value_median_debt, "\n")
```

```
## P-value: 0.004650883
```

```
cat("Decision at 4% significance level:", decision_median_debt, "\n")
```

```
## Decision at 4% significance level: reject
```

```
cat("Confidence Interval:", conf_interval_median_debt, "\n\n")
```

```
## Confidence Interval: 17213.26 Inf
```

To investigate whether the population mean MedianDebt is greater than $17,000, a significance test was conducted at a 4% significance level. The hypotheses tested are as follows: - Null Hypothesis (H0): The population mean MedianDebt is equal to or less than $17,000. - Alternative Hypothesis (H1): The population mean MedianDebt is greater than $17,000.

Summary Statistics

Test Statistic: 2.616878 P-value: 0.004650883 Decision at 4% significance level: reject Confidence Interval: 17213.26 Inf

Conclusion:

The p-value supports the idea that the population mean MedianDebt is greater than $17,000. Since the p-value (0.004650883) is less than the significance level (0.04), the decision is to reject the null hypothesis. This means that there is enough evidence to support the idea that the population mean MedianDebt is greater than $17,000. In other words, there is enough evidence to claim that the median student debt exceeds $17,000.

Additionally, the confidence interval was calculated. The confidence interval was [17213.26, Inf], and since it includes values greater than $17,000, it aligns with the result of the significance test, reinforcing the idea that the median student debt is likely higher than $17,000.

Therefore, the evidence from the test and confidence interval suggests that, on average, students may have a median debt that surpasses $17,000.

TASK 9

```
faculty_salary_test <- t.test(AverageFacultySalary ~ FundingModel, data = college_data,
conf.level = 0.97)

p_value_faculty_salary <- faculty_salary_test$p.value
conf_interval_faculty_salary <- faculty_salary_test$conf.int
decision_faculty_salary <- ifelse(p_value_faculty_salary < 0.03, "reject", "fail to reje
ct")

cat("Test Statistic:", faculty_salary_test$statistic, "\n")
```

```
## Test Statistic: -3.199235
```

```
cat("P-value:", p_value_faculty_salary, "\n")
```

```
## P-value: 0.001531169
```

```
cat("Decision at 3% significance level:", decision_faculty_salary, "\n")
```

```
## Decision at 3% significance level: reject
```

```
cat("Confidence Interval:", conf_interval_faculty_salary, "\n\n")
```

```
## Confidence Interval: -1192.742 -225.7695
```

To explore whether there is enough evidence to suggest that the mean of AverageFacultySalary varies based on the FundingModel of a college. A significance test was conducted at a 3% significance level to compare the means among different funding models. The hypotheses tested were as follows: - H0 (Null Hypothesis): The mean of AverageFacultySalary is the same for all FundingModel categories. - H1 (Alternative Hypothesis): The mean of AverageFacultySalary differs by FundingModel.

Summary Statistics

Test Statistic: -3.199235 P-value: 0.001531169 Decision at 3% significance level: reject Confidence Interval: -1192.742 -225.7695

Conclusion:

The p-value supports the idea that the mean of AverageFacultySalary is different by FundingModel, providing evidence that faculty salaries vary among different funding models. Since the p-value (0.001531169) is less than the significance level (0.03), the decision is to reject the null hypothesis. This suggests that there is enough evidence to claim that the mean of AverageFacultySalary differs by FundingModel. In other words, there is enough evidence to claim that the average faculty salary differs based on the funding model of the college.

Additionally, a confidence interval was calculated. The confidence interval does not include zero (-1192.742 to -225.7695), supporting the result of the significance test and indicating that there are significant differences in average faculty salaries based on FundingModel.

Therefore, the evidence from the test and confidence interval suggests that, on average, faculty salaries differ depending on the funding model of the college.