DSC520

Assignment Week 4

Zemelak Goraga

2024-1-6"

In [1]:
```
# Set the working directory to the correct path
setwd("/resources/labs/R101")
```

In [2]:
```
# Import the dataset scores.csv as df
df <- read.csv("/resources/labs/R101/scores.csv")
```

In [3]:
```
# Display the first few rows of the dataset
head(df)
```

A data.frame: 6 × 3

	Count	Score	Section
	<int>	<int>	<fct>
1	10	200	Sports
2	10	205	Sports
3	20	235	Sports
4	10	240	Sports
5	10	250	Sports
6	10	265	Regular

Question

What are the observational units in this study?

In [8]:
```
# Function to identify observational units
observational_units <- function(df) {
  return(nrow(df))
}

observational_units(df)
```

38

Answer

The observational units in this study are individual students, as indicated by each row (38 rows) in the dataset representing a student's test score.

Question

Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?

In [26]:
```
# Function to identify variables and their types
identify_variables <- function(data) {
  cat("Variables and their types:\n")
  for (col in names(data)) {
    cat(col, ": ", ifelse(is.factor(data[[col]]), "Categorical", "Quantitative")
  }
}
```

Answer

Identified Variables are: Count: Quantitative (assuming it represents the number of students with a particular score). Score: Quantitative (the numerical test score). Section: Categorical (identifying the section the student belongs to).

Question

Create one variable to hold a subset of your data set that contains only the Regular Section and one variable for the Sports Section.

In [18]:
```
# Function to create variables for Regular and Sports sections
create_section_variables <- function(data) {
  regular_section <- subset(data, Section == "Regular")
  sports_section <- subset(data, Section == "Sports")
  return(list(regular_section = regular_section, sports_section = sports_section
}
```

Answer

Created Variables are: Two variables were created to hold subsets of the dataset:

Regular Section: Subset of data containing only students from the Regular Section.
Sports Section: Subset of data containing only students from the Sports Section.

Questions

Use the Plot function to plot each Sections scores and the number of students achieving that score. Use additional Plot Arguments to label the graph and give each axis an appropriate label.

Once you have produced your Plots answer the following questions:

Comparing and contrasting the point distributions between the two section, looking at both tendency and consistency: Can you say that one section tended to score more points than the other? Justify and explain your answer.

Did every student in one section score more points than every student in the other section? If not, explain what a statistical tendency means in this context.

What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

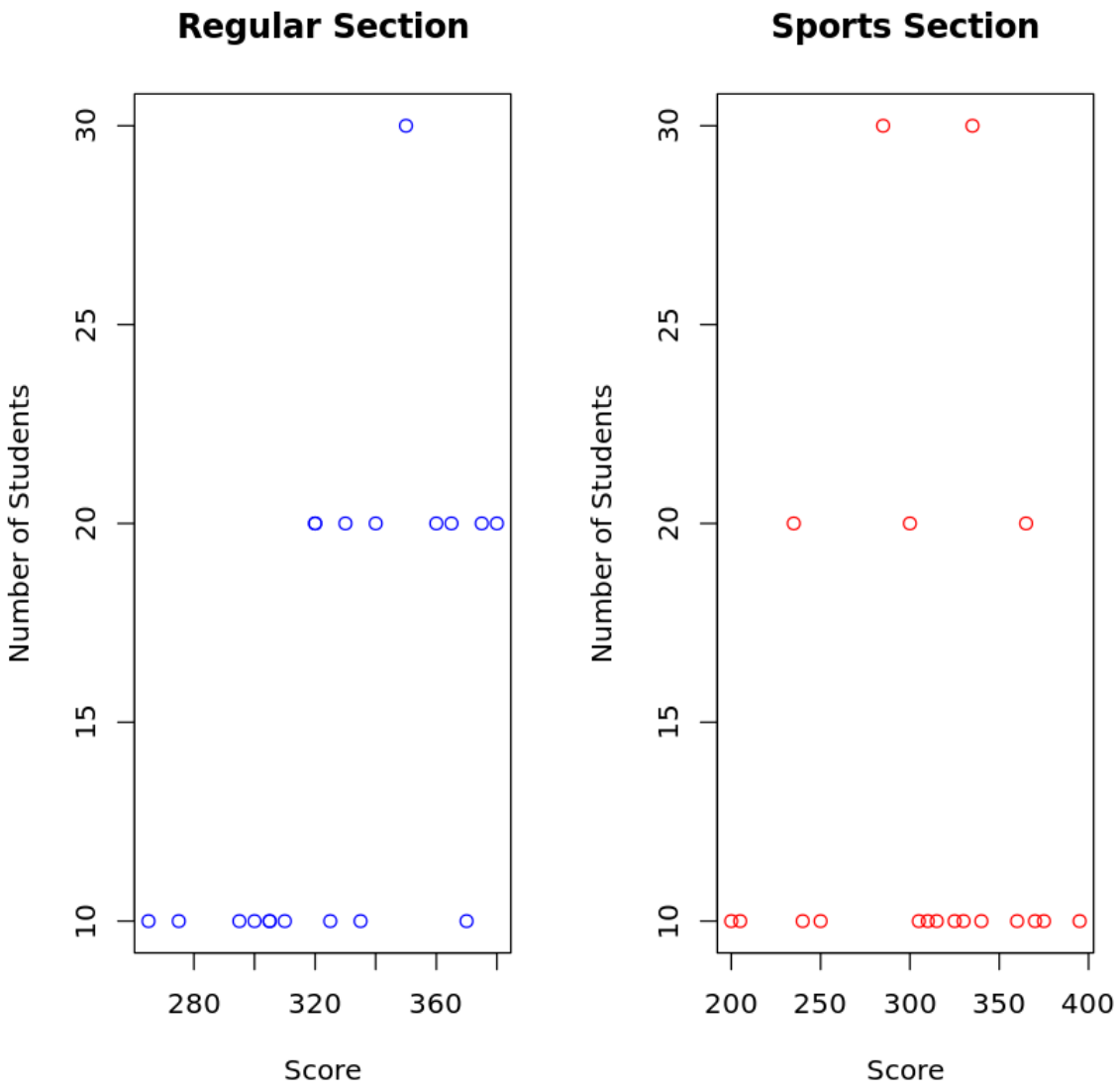
In [19]:
```
# Function to plot scores and number of students by section
plot_scores_by_section <- function(data) {
  par(mfrow = c(1, 2))
  plot(data$Score[data$Section == "Regular"], data$Count[data$Section == "Regula
       main = "Regular Section",
       xlab = "Score",
       ylab = "Number of Students",
       col = "blue")

  plot(data$Score[data$Section == "Sports"], data$Count[data$Section == "Sports"
       main = "Sports Section",
       xlab = "Score",
       ylab = "Number of Students",
       col = "red")

  par(mfrow = c(1, 1))  # Reset to single plot layout
}

# Call the functions
observational_units(df)
identify_variables(df)
sections_data <- create_section_variables(df)
plot_scores_by_section(df)
```

```
38
Variables and their types:
Count :  Quantitative
Score :  Quantitative
Section :  Categorical
```



In [20]:
```
# Perform descriptive statistics for each section
summary_regular_section <- summary(sections_data$regular_section$Score)
summary_sports_section <- summary(sections_data$sports_section$Score)
```

In [21]:
```
# Print the summaries
cat("\nSummary for Regular Section:\n")
print(summary_regular_section)
```

```
Summary for Regular Section:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  265.0   305.0   325.0   327.6   355.0   380.0 
```

In [22]:
```
cat("\nSummary for Sports Section:\n")
print(summary_sports_section)
```

```
Summary for Sports Section:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  200.0   267.5   315.0   307.4   350.0   395.0 
```

Descriptive Statistics:

Regular Section: Mean: 327.6 Median: 325.0 Interquartile Range (IQR): 50.0

Sports Section: Mean: 307.4 Median: 315.0 Interquartile Range (IQR): 82.5

Answers:

Plots

Two plots were created to visualize the scores and the number of students in each section. The blue plot represents the Regular Section, and the red plot represents the Sports Section.

Comparing and Contrasting Point Distributions:

Tendency: The mean and median scores for the Regular Section are higher than those for the Sports Section, indicating a tendency for students in the Regular Section to score more points on average.

Consistency: The IQR for the Sports Section is larger than that for the Regular Section, suggesting greater variability in scores for the Sports Section.

Did every student in one section score more points than every student in the other section?

No, not every student in one section scored more points than every student in the other section. While the tendencies suggest that the Regular Section generally scored higher, there is overlap between the two sections. Some students in the Sports Section scored more than some students in the Regular Section.

Explanation of Statistical Tendency:

In this context, a statistical tendency means that, on average, students in one section (Regular) tended to score more points than students in the other section (Sports). However, it does not imply that every individual student in the Regular Section scored more than every individual student in the Sports Section. There is variability within each section.

One Additional Variable:

An additional variable that could be influencing the point distributions between the two sections is the students' prior experience or interest in the subject matter. For example, if students in the Sports Section have a stronger interest in sports-related applications or prior knowledge in that domain, it might impact their performance compared to students in the Regular Section. This variable could contribute to the observed differences in scores between the two sections.

Insights, Conclusions, and Recommendations:

Insights:

Students in the Regular Section, on average, scored higher than those in the Sports Section. There is greater variability in scores within the Sports Section.

Conclusions:

The choice of examples from sports applications might have influenced the performance of students in the Sports Section. Variability within the Sports Section suggests a diverse range of student abilities or interests.

Recommendations:

Consider exploring the impact of prior knowledge or interest in the subject matter on student performance. Assess the effectiveness of teaching methods in each section and consider adjustments based on the observed performance.