ASSIGNMENT (4.2 Exercise)

Maxim Bilenkin

2024-12-17

Assignment: ASSIGNMENT (4.2 Exercise)

Name: Bilenkin, Maxim

Date: 2024-12-17

```
# Loading the necessary packages
library(readr)
library(dplyr)
```

```
# Importing data file with scores
scores_data_file <- read.csv("C:\\Users\\maxim\\Downloads\\scores.csv")

# Displaying the file with all the data including fields to understand it
str(scores_data_file)
```

```
## 'data.frame':    38 obs. of  3 variables:
##  $ Count  : int  10 10 20 10 10 10 10 30 10 10 ...
##  $ Score  : int  200 205 235 240 250 265 275 285 295 300 ...
##  $ Section: chr  "Sports" "Sports" "Sports" "Sports" ...
```

```
summary(scores_data_file)
```

```
##      Count           Score         Section
##  Min.   :10.00   Min.   :200.0   Length:38
##  1st Qu.:10.00   1st Qu.:300.0   Class :character
##  Median :10.00   Median :322.5   Mode  :character
##  Mean   :14.47   Mean   :317.5
##  3rd Qu.:20.00   3rd Qu.:357.5
##  Max.   :30.00   Max.   :395.0
```

```
colnames(scores_data_file)
```

```
## [1] "Count"   "Score"   "Section"
```

```
head(scores_data_file)
```

```
##   Count Score Section
## 1    10   200  Sports
## 2    10   205  Sports
## 3    20   235  Sports
## 4    10   240  Sports
## 5    10   250  Sports
```

```
## 6    10   265 Regular
```

```
# Separating the sections
sports_section <- scores_data_file %>% filter(Section == "Sports")
regular_section <- scores_data_file %>% filter(Section == "Regular")
```

```
summary(regular_section$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   265.0   305.0   325.0   327.6   355.0   380.0
```

```
# Summary for the Sports section.
summary(sports_section$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   200.0   267.5   315.0   307.4   350.0   395.0
```

```
sum(is.na(regular_section$Score))
```

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

```
head(regular_section)
```

```
##   Count Score Section
## 1    10   265 Regular
## 2    10   275 Regular
## 3    10   295 Regular
## 4    10   300 Regular
## 5    10   305 Regular
## 6    10   310 Regular
```

```
# Removing rows with NA values in the Score column
cleaned_regular_section <- regular_section %>% filter(!is.na(Score))

# Descriptive statistics for the cleaned variety section
regular_variety_summary <- cleaned_regular_section %>% summarize(
  mean_score = mean(Score),
  standard_deviation_score = sd(Score)
)

# Printing cleaned variety summary
print(regular_variety_summary)
```

```
##   mean_score standard_deviation_score
## 1   327.6316                 33.26528
```

```
# Descriptive statistics for the sports section
sports_summary <- sports_section %>% summarize(
  mean_score = mean(Score),
  standard_deviation_score = sd(Score)
)

# Printing sports summary
print(sports_summary)
```

```
##   mean_score standard_deviation_score
## 1   307.3684                  58.0318
```

```r
# Descriptive statistics for the variety section
variety_summary <- regular_section %>% summarize(
  mean_score = mean(Score),
  standard_deviation_score = sd(Score)
)

# Printing variety summary
print(variety_summary)
```

```
##   mean_score standard_deviation_score
## 1   327.6316                 33.26528
```

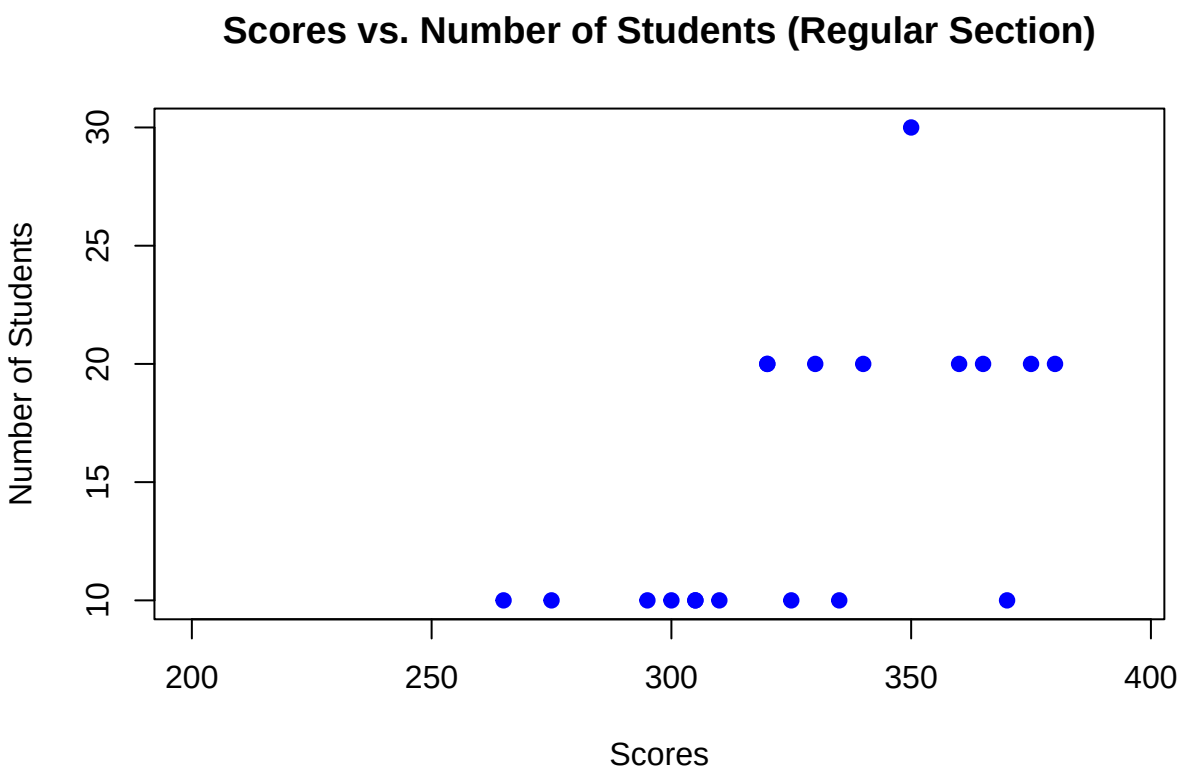
```r
# 1) What are the observational units in this study?

# Answer: The observational units in this study are the student representing
#         a data point. Each student has it's own fields such as "Count",
#         "Score" and "Section".

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

# Answer: There are three variables, "Count", "Score" and "Section". Both, Count
#         and Score represent numbers and are therefore quantitative variables.
#         On the other hand, Section represents a category of each student
#         making it a categorical variable because it describes the section each
#         student belongs to.

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

# Answer:

# Creating a subset for Regular Section.
regular_section <- scores_data_file %>% filter(Section == "Regular")
head(regular_section)
```

```
##   Count Score Section
## 1    10   265 Regular
## 2    10   275 Regular
## 3    10   295 Regular
## 4    10   300 Regular
## 5    10   305 Regular
## 6    10   310 Regular
```

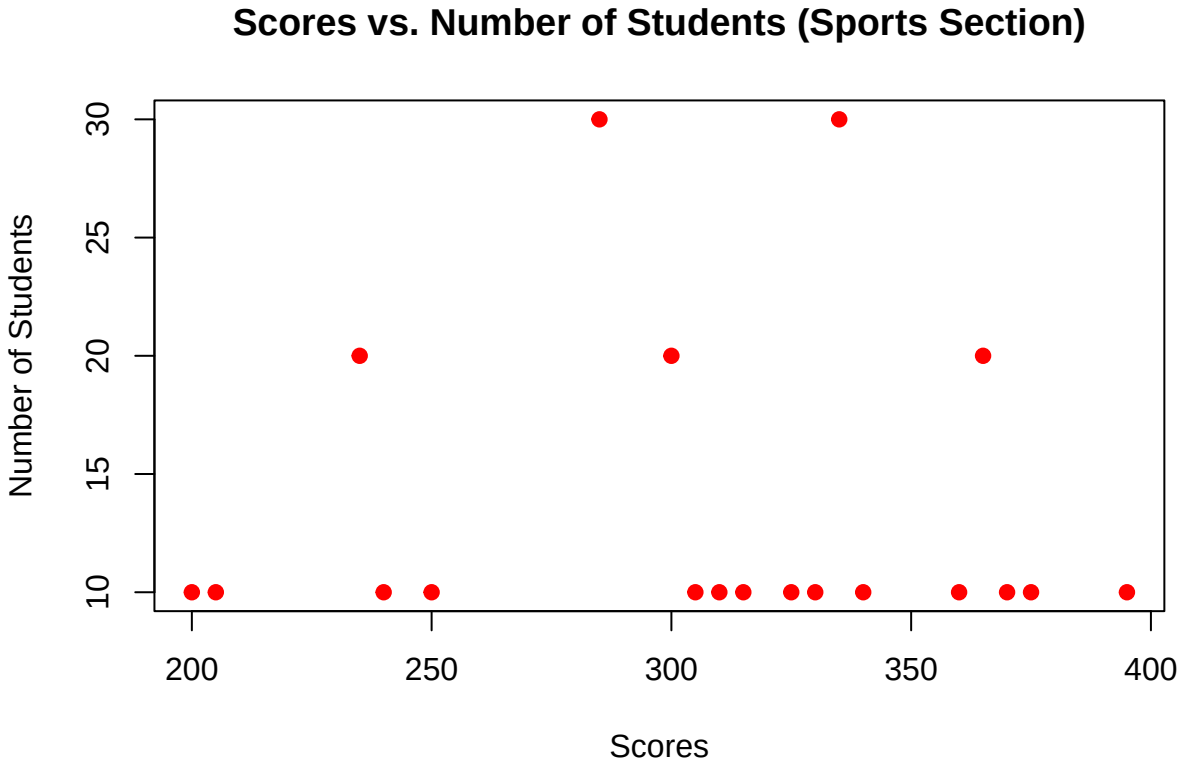
```r
# Creating a subset for Sports Section.
sports_section <- scores_data_file %>% filter(Section == "Sports")
head(sports_section)
```

```
##   Count Score Section
## 1    10   200  Sports
## 2    10   205  Sports
## 3    20   235  Sports
## 4    10   240  Sports
## 5    10   250  Sports
## 6    30   285  Sports
```

```r
# 4) 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:

# Answer:

# Creating a subset for Regular Section.
regular_section <- scores_data_file %>% filter(Section == "Regular")

# Creating a subset for Sports Section.
sports_section <- scores_data_file %>% filter(Section == "Sports")

# Defining a common x-axis ranges.
x_range <- range(c(regular_section$Score, sports_section$Score))

# Plot for the Regular Section
plot(regular_section$Score, regular_section$Count,
     main = "Scores vs. Number of Students (Regular Section)",
     xlab = "Scores", ylab = "Number of Students",
     col = "blue", pch = 19, xlim = x_range)
```




```r
# Plot for the Sports Section
plot(sports_section$Score, sports_section$Count,
     main = "Scores vs. Number of Students (Sports Section)",
     xlab = "Scores", ylab = "Number of Students",
     col = "red", pch = 19, xlim = x_range)
```




```
# a. 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.

# Answer: Based on the descriptive statistics we can see that Regular point
#         distribution has a mean of 327.63 and Sports point distribution has a
#         mean of 307.37. Since, Regular mean is higher than Sports mean. It
#         means that the Regular students scoring more points because its
#         tendency distribution is higher. Also, we can see that the standard
#         deviation of Regular is 33.27 which is lower than standard deviation
#         of 58.03 for Sports distribution. This means that the Sports
#         distribution points spread more from the mean. At the same time it
#         means that that the Regular distribution points spread less from the
#         mean and it is more consistent.

# b. 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.

# Answer: Not every student in one section scored more points than in another
#         section. We can easily and quickly see by taking the minimum and
#         maximum number for each distribution points. For Regular min and max
#         are 265 and 380 respectively. And for Sports min and max are 200 and
#         395 respectively. We can see clearly that some points on Sports
#         distribution have scores for students that are higher with 395 vs. 380
#         for Regular students. However, statistical tendency means that on
#         average students for Regular distribution score higher.

# c. What could be one additional variable that was not mentioned in the
```

```
#    narrative that could be influencing the point distributions between the two
#    sections?

# Answer: Extra participation in some sport curriculum could be the additional
#         variable that probably influencing students performance on scores. We
#         could say that students in regular section have more time to spend and
#         concentrate on their studies and homework than students from sport
#         section. This could be a good explanation why scoring is higher for
#         regular students on average.
```