Week4_Assignment_TestScores

Henry Pham

2023-04-08

```
# Set the working directory to the root of your DSC 520 directory

getwd()
```

```
## [1] "C:/R/DSC520"
```

```
dir("C://R//DSC520//data")
```

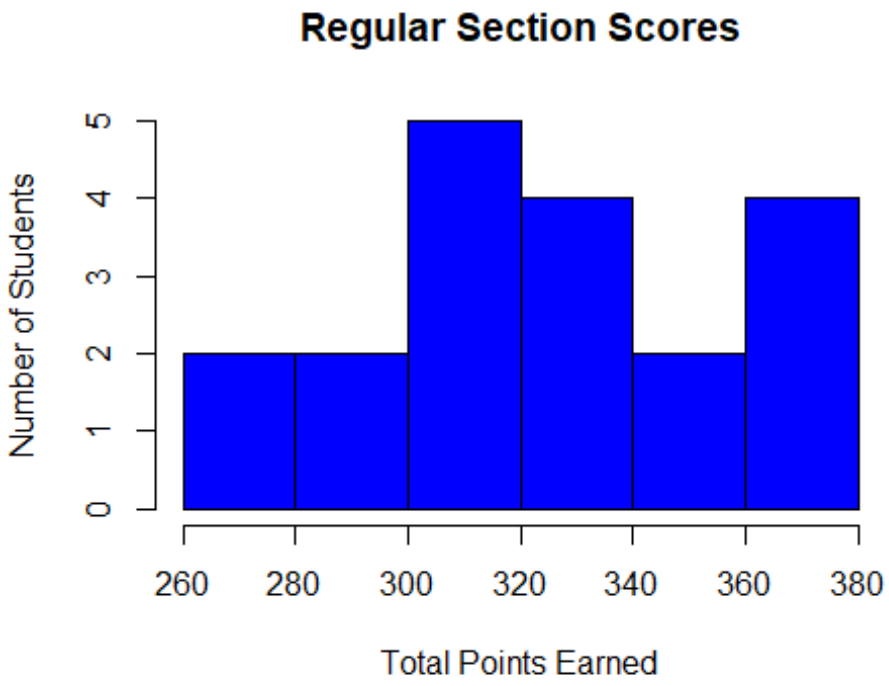
```
## [1] "acs-14-1yr-s0201.csv"           "G04ResultsDetail2004-11-02.xls"
## [3] "r4ds"                           "scores.csv"
## [5] "tidynomicon"
```

```
if (!file.exists("data"))
{
  # set data working directory
  setwd("C://R//DSC520//data")
}

# Load the data
scores_data <- read.csv("scores.csv")
head(scores_data)
```

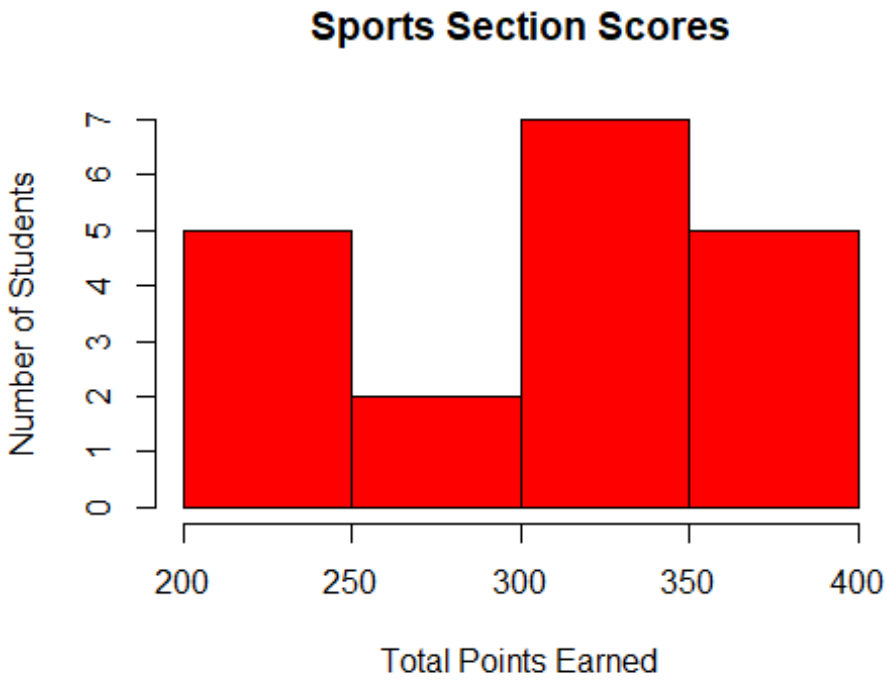
```
##   Count Score Section
## 1    10   200  Sports
## 2    10   205  Sports
## 3    20   235  Sports
## 4    10   240  Sports
## 5    10   250  Sports
## 6    10   265 Regular
```

```
# Q1.   What are the observational units in this study?
# The observational units in this study are students.

#  Q2.  Identify the variables mentioned in the narrative paragraph and
determine
# which are categorical and quantitative?

# The variables mentioned in the narrative paragraph are:
# Section(categorical): Regular or Sports
# Score(quantitative): final scores received in the course
# Count(quantitative): total number of scores earned in the course
```

```r
# Q3.   Create one variable to hold a subset of your data set that contains only the
# Regular Section and one variable for the Sports Section.

regular_section <- scores_data[scores_data$Section == "Regular", ]
sports_section <- scores_data[scores_data$Section == "Sports", ]


# Q4.   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.


library(ggplot2)

# Plot for Regular Section
hist(regular_section$Score, main = "Regular Section Scores", xlab = "Total Points Earned", ylab = "Number of Students", col = "blue")
```




```r
# Plot for Sports Section
hist(sports_section$Score, main = "Sports Section Scores", xlab = "Total Points Earned", ylab = "Number of Students", col = "red")
```




```
# Once you have produced your Plots answer the following questions:
#   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.

# To compare the point distributions between the two sections, we can
visually inspect
# the histograms. If one section tended to score more points than the other,
# we would expect to see a higher peak or more values in the upper end of the
distribution for that section.
# However, based on the histograms, it's difficult to say whether one section
tended to score more points than the other.
# Both distributions appear roughly similar in shape and spread.

# 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.

# It is unlikely that every student in one section scored more points than
every
# student in the other section. Instead, a statistical tendency means that,
on average,
# students in one section may have scored more points than students in the
other section.
```

```r
# This can be assessed using measures of central tendency, such as the mean
or median.
# regular_section has a right-skewed dist while sports_section has a left-
skewed dist.

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.2     ✓ tibble    3.2.0
## ✓ purrr     1.0.1     ✓ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors

regular_section %>% summarise(mean(Score), median(Score))

##   mean(Score) median(Score)
## 1    327.6316           325

sports_section %>% summarise(mean(Score), median(Score))

##   mean(Score) median(Score)
## 1    307.3684           315

# 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?


# One additional variable that could be influencing the point distributions
between
# the two sections is student motivation. It's possible that students who
were interested
# in sports were more motivated to succeed in the sports-themed section,
# which could have influenced their scores.
```