assignment4.1_Chattapadhyay_Kausik.R

kausik

2022-09-22

```
# Assignment: ASSIGNMENT 4.1
# Name: Chattapadhyay, Kausik
# Date: 2022-09-21

## Load the ggplot2 package
library(ggplot2)
library(dplyr)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory
setwd("/Users/kausik/desktop/MS Data Science/DSC 520/dsc520-stats-r-assignments")

## Load the `data/scores.csv` to
scores_df <- read.csv("data/scores.csv")
head(scores_df)
```

```
##   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
```

```
## 1. What are the observational units in this study?
# course grades and total points earned in the course.

## 2. Identify the variables mentioned in the narrative paragraph and determine
## which are categorical and quantitative?
str(scores_df)
```

```
## '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" ...
```

```
# Categorical: Sports and regular
# Quantitative: Score and total points

## 3. Create one variable to hold a subset of your data set that contains only
```

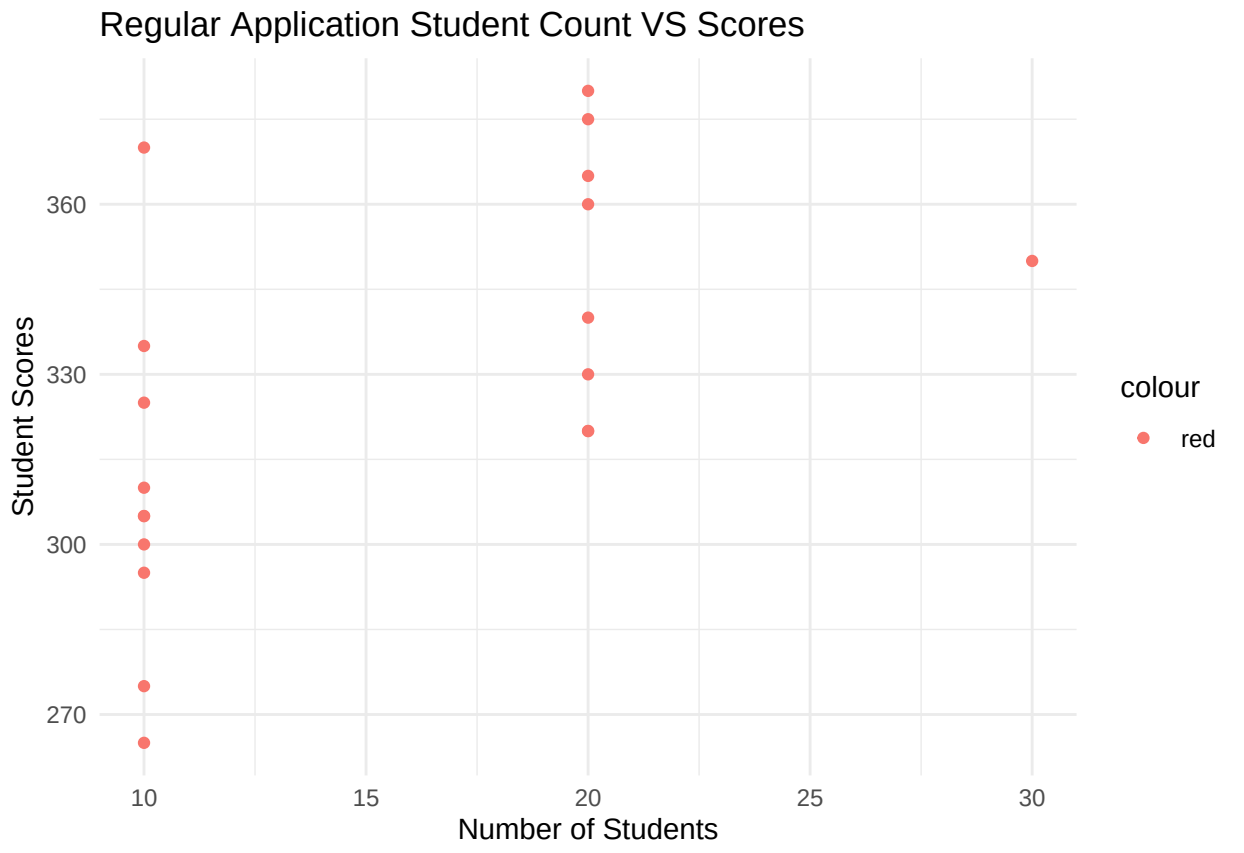
```
## the Regular Section and one variable for the Sports Section.
regular_scores <- scores_df %>%
                    filter(scores_df$Section == "Regular")
regular_scores
```

```
##    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
## 7     20   320 Regular
## 8     10   305 Regular
## 9     20   320 Regular
## 10    10   325 Regular
## 11    20   330 Regular
## 12    10   335 Regular
## 13    20   340 Regular
## 14    30   350 Regular
## 15    20   360 Regular
## 16    20   365 Regular
## 17    10   370 Regular
## 18    20   375 Regular
## 19    20   380 Regular
```

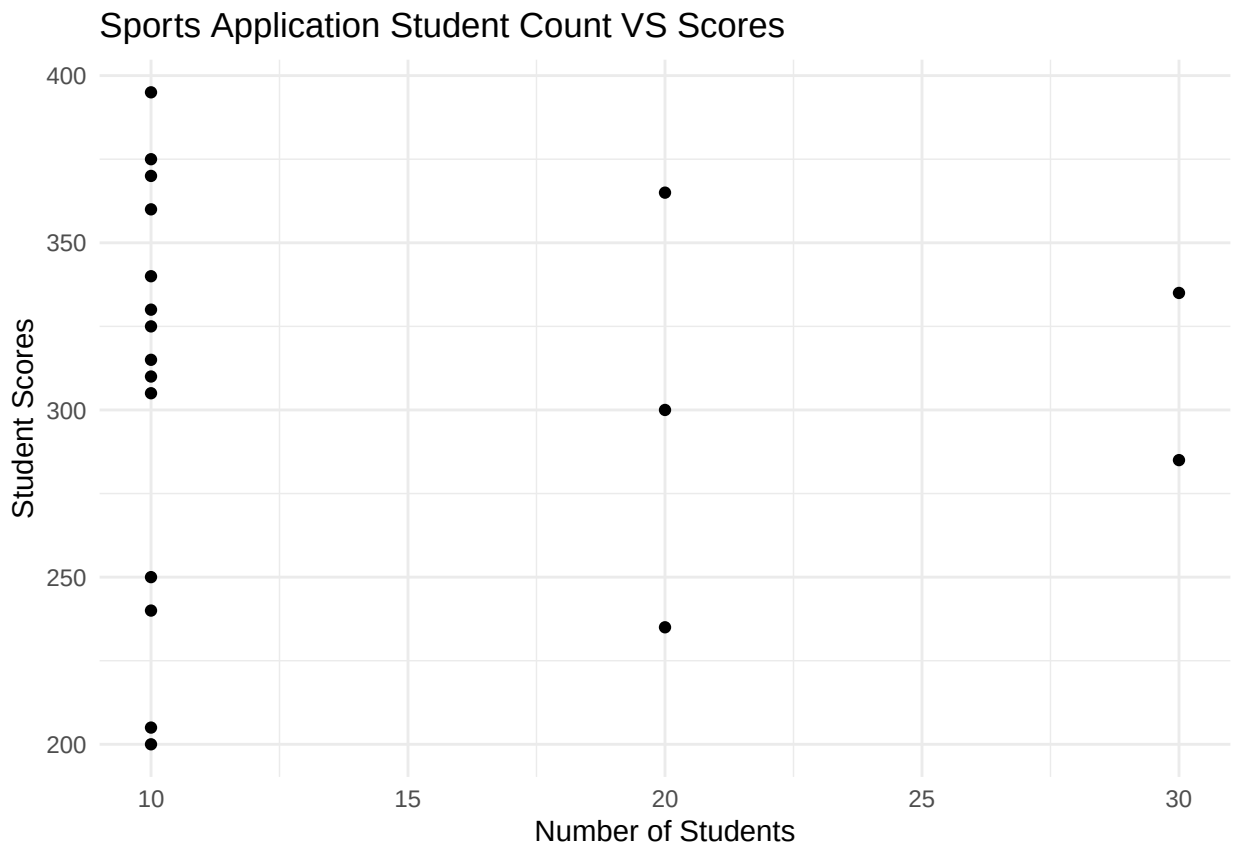
```
sports_scores <- scores_df %>%
    filter(scores_df$Section == "Sports")
sports_scores
```

```
##    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
## 7     20   300  Sports
## 8     10   305  Sports
## 9     10   310  Sports
## 10    10   315  Sports
## 11    10   325  Sports
## 12    10   330  Sports
## 13    30   335  Sports
## 14    10   340  Sports
## 15    10   360  Sports
## 16    20   365  Sports
## 17    10   370  Sports
## 18    10   375  Sports
## 19    10   395  Sports
```

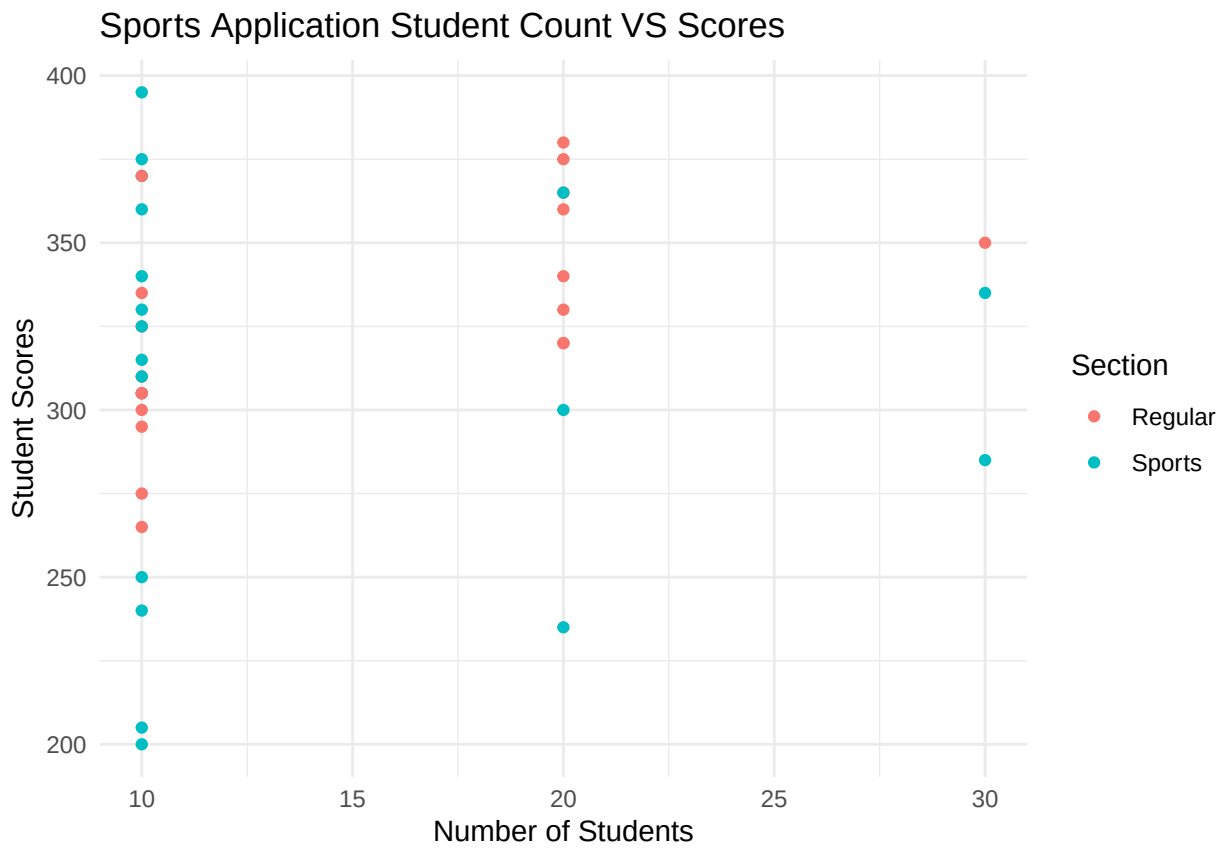
```
## 4. Use the Plot function to plot each Sections scores and the number of
## students achieving that score. Use additional Plot Arguments to label the
## graph and give each axis an appropriate label. Once you have produced your Plots
## answer the following questions:
ggplot(regular_scores, aes(x=Count, y=Score, color="red")) + labs(x = "Number of Students",
        y= "Student Scores", title = "Regular Application Student Count VS Scores") +
    geom_point()
```




```
ggplot(sports_scores, aes(x=Count, y=Score)) + labs(x = "Number of Students",
    y= "Student Scores",
    title = "Sports Application Student Count VS Scores") +
    geom_point()
```



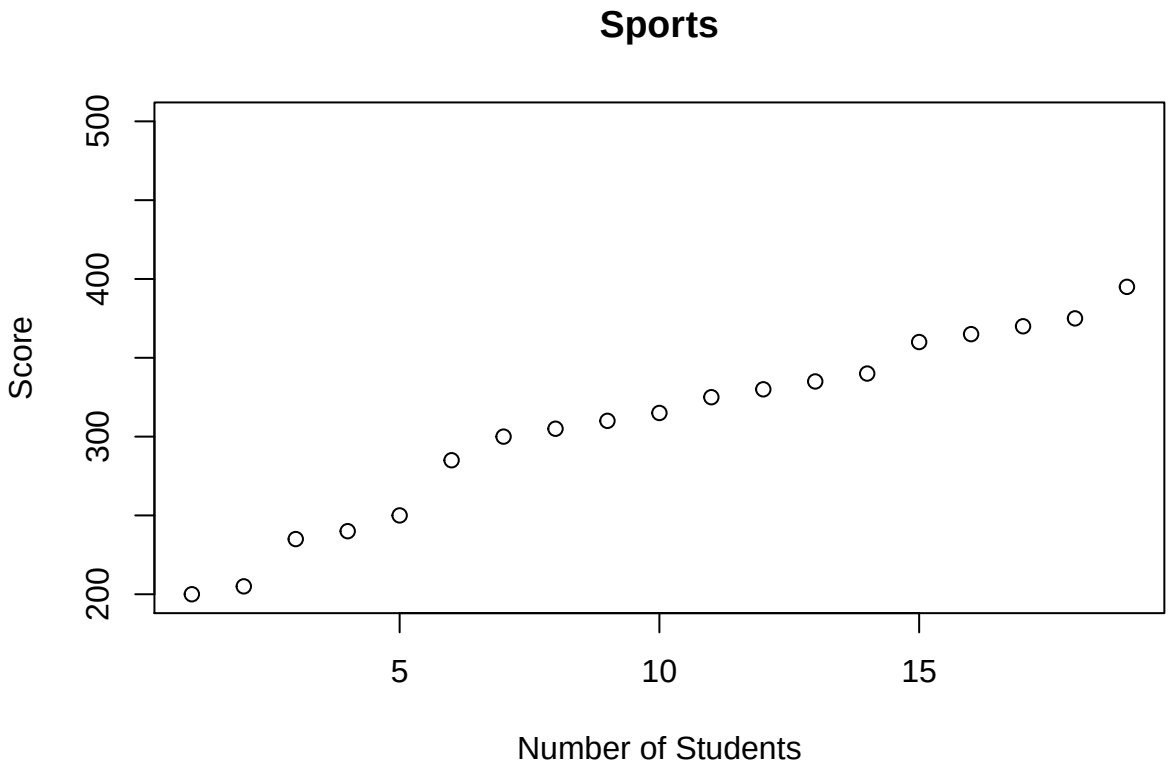

```
ggplot(scores_df, aes(x=Count, y=Score, color=Section)) + labs(x = "Number of Students",
                                                               y= "Student Scores",
                         title = "Sports Application Student Count VS Scores") +
    geom_point()
```



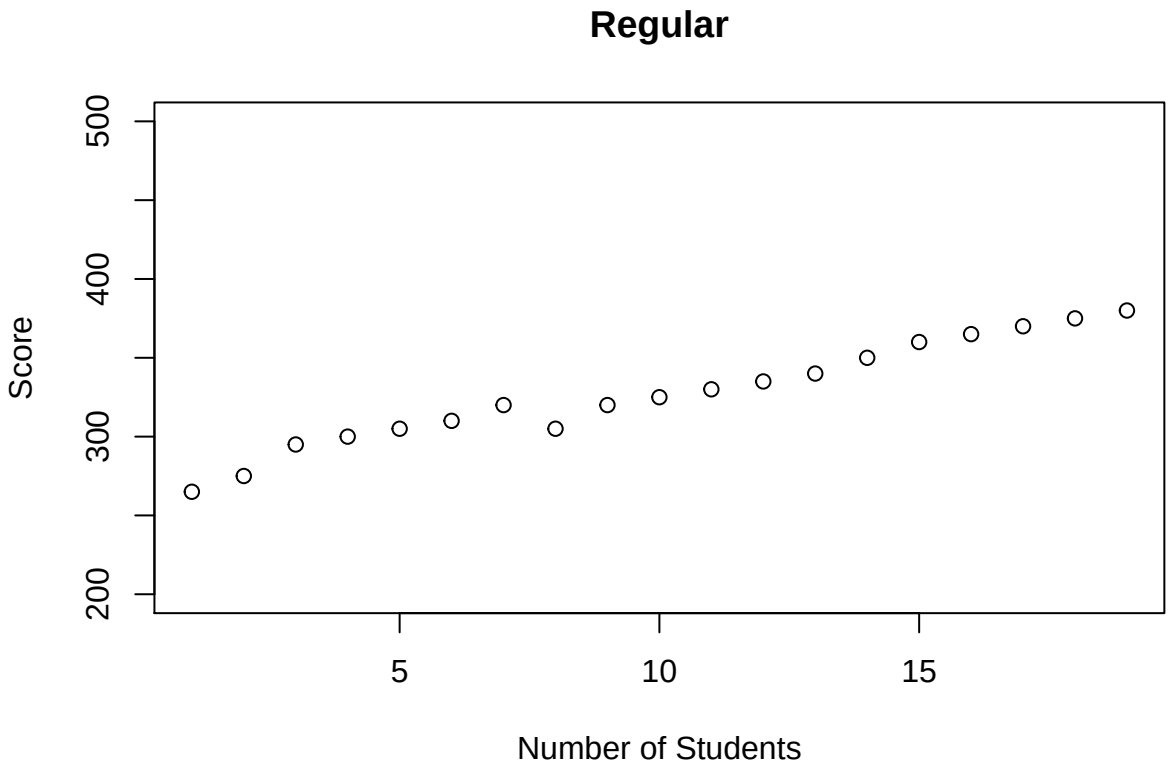

```
Score1=sports_scores[,2]
Score2=regular_scores[,2]

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

plot(Score1, xlab="Number of Students", ylab="Score", main="Sports", ylim=c(200, 500))
```




```
plot(Score2, xlab="Number of Students", ylab="Score", main="Regular", ylim=c(200, 500))
```




```
##    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.

#        The sports section tended to score more points than the regular section.
#        From looking at the plot, we can see that despite the begining of the
#        plot showing more points being given for "regular", by the time we get
#        towards the end of the plot, we see more points being given for "sports"

##    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.

#        Students in the sports section seem to have gotten more points than the
#        students in the regular section.

##    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?

#        In the narrative, it vaguely states that the professor "recently taught
#        two sections of the same class." It does not state if the two sections
#        were taught during the same semester/quarter. If semester/quarter (timeframe)
#        differs, that could influence student performance.
```