4_2

2022-07-02

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# libraries
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(pastecs)
```

```
## Warning: package 'pastecs' was built under R version 4.2.1
```

```
##
## Attaching package: 'pastecs'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
# Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/darge/OneDrive/Documents/1. Data Science/DSC 520 - Statistics for Data Science/dsc520")

## Load test_scores.csv
test_scores <- read.csv("data/scores.csv")
```

1. What are the observational units in this study? The observational units are the different recordings of each class. The professor recorded 38 observations

2. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative? The variables are count, score, and section. Count is quantitative, score is quantitative, and section is categorical

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.

```r
sports <- filter(test_scores, Section == "Sports")
print(sports)
```

```
##    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
```

```r
regular <- filter(test_scores, Section == "Regular")
print(regular)
```

```
##    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
```

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:

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 number of students in the regular section scored higher and were more consistent.
The regular section has a higher sum than the the sports section. The data isn't specific to the number of students that scored or the class size in the Count Column. The regular section has a higher mean and median which we can use to determine that the regular section has higher test scores

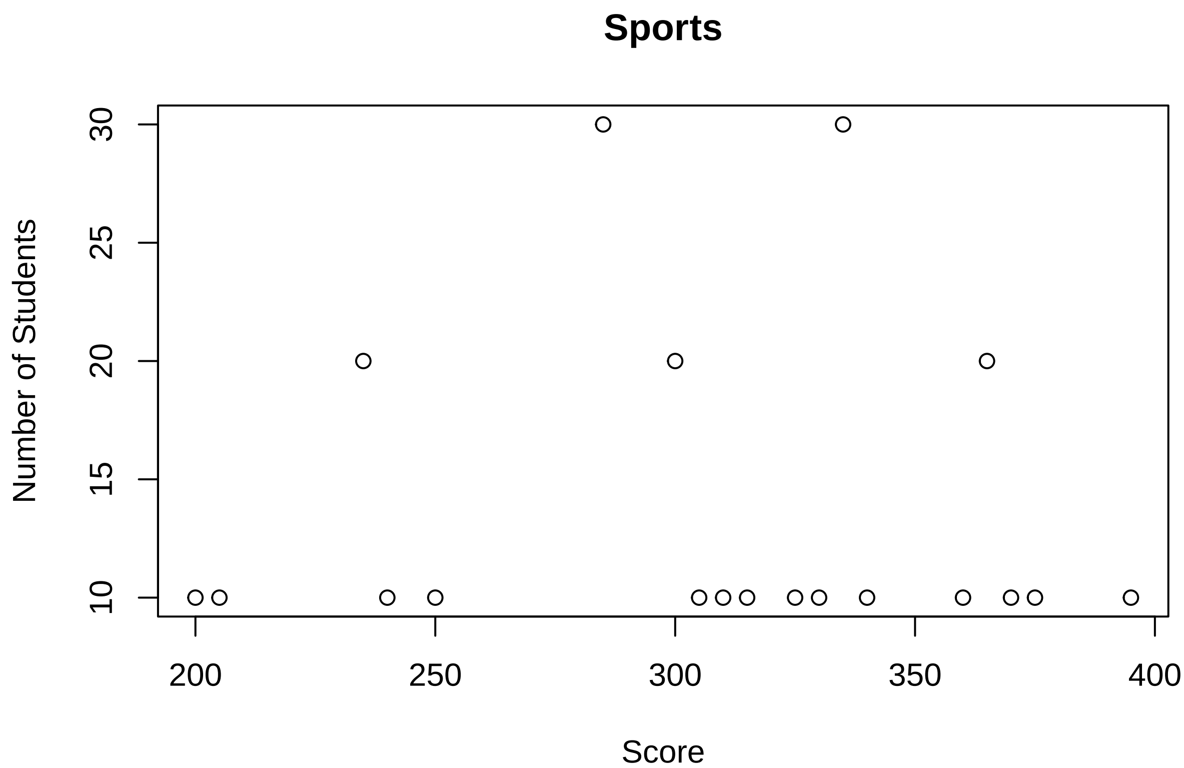
Did every student in one section score more points than every student in the other section? If not, explain what a statistical tendency means in this context.

No, the regular section had higher scores

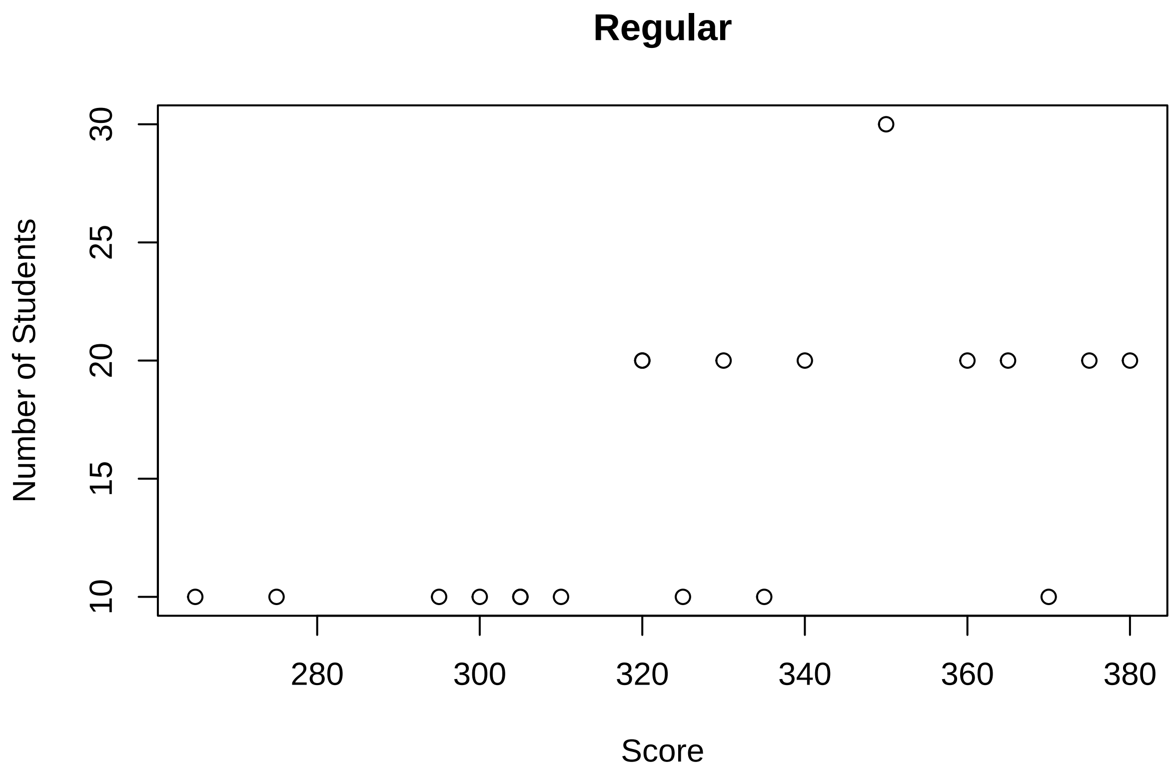
What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

An additional variable that was not mentioned in the narrative that could be influencing the point distributions between the the two sections is class size.

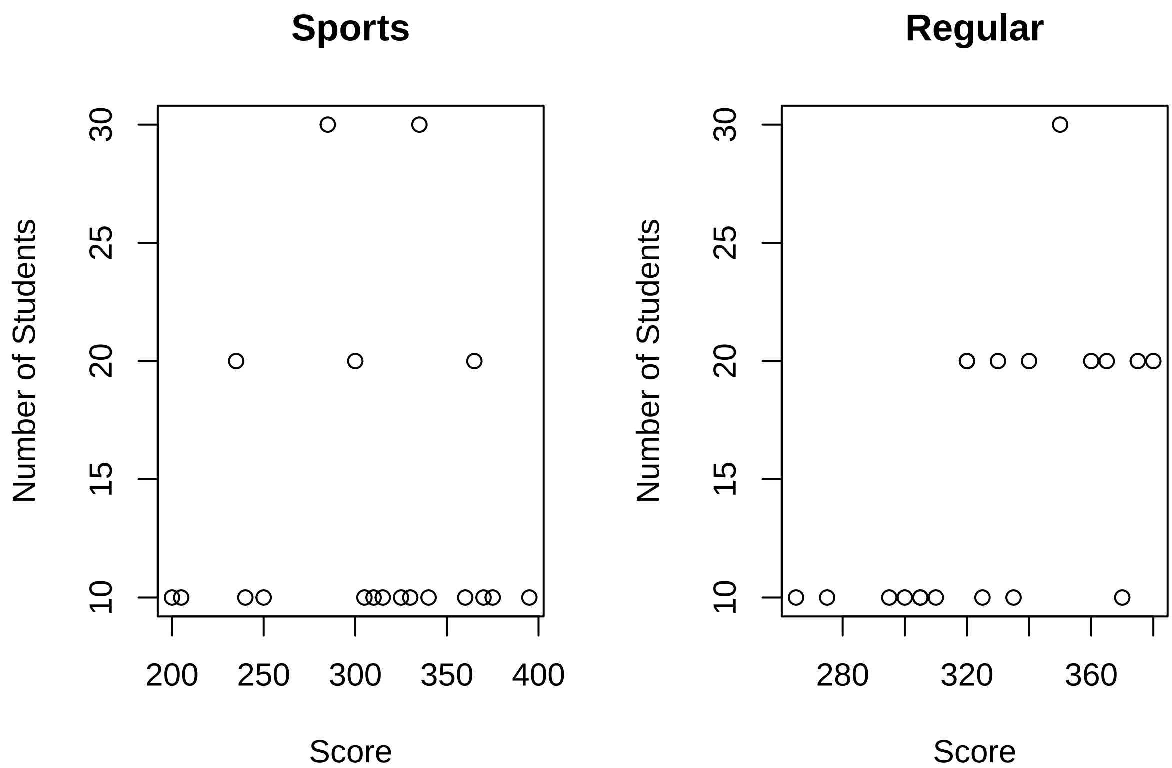
```
sports_chart <- plot(x = sports$Score, y = sports$Count, main = "Sports", xlab = "Score", ylab = "Numbe
```




```
regular_chart <- plot(x = regular$Score, y = regular$Count, main = "Regular", xlab = "Score", ylab = "Nu
```




```
par(mfcol = c(1,2))
plot(x = sports$Score, y = sports$Count, main = "Sports", xlab = "Score", ylab = "Number of Students")
plot(x = regular$Score, y = regular$Count, main = "Regular", xlab = "Score", ylab = "Number of Students
```




```
stat.desc(sports)
```

```
##                    Count         Score Section
## nbr.val       19.0000000   19.0000000       NA
## nbr.null       0.0000000    0.0000000       NA
## nbr.na         0.0000000    0.0000000       NA
## min           10.0000000  200.0000000       NA
## max           30.0000000  395.0000000       NA
## range         20.0000000  195.0000000       NA
## sum          260.0000000 5840.0000000       NA
## median        10.0000000  315.0000000       NA
## mean          13.6842105  307.3684211       NA
## SE.mean        1.5691705   13.3134085       NA
## CI.mean.0.95   3.2967049   27.9704333       NA
## var           46.7836257 3367.6900585       NA
## std.dev        6.8398557   58.0318021       NA
## coef.var       0.4998356    0.1888021       NA
```

```
stat.desc(regular)
```

```
##                    Count         Score Section
## nbr.val       19.0000000   19.0000000       NA
## nbr.null       0.0000000    0.0000000       NA
## nbr.na         0.0000000    0.0000000       NA
```

```
## min           10.0000000  265.0000000      NA
## max           30.0000000  380.0000000      NA
## range         20.0000000  115.0000000      NA
## sum          290.0000000 6225.0000000      NA
## median        10.0000000  325.0000000      NA
## mean          15.2631579  327.6315789      NA
## SE.mean        1.4035088    7.6315789      NA
## CI.mean.0.95   2.9486625   16.0333524      NA
## var           37.4269006 1106.5789474      NA
## std.dev        6.1177529   33.2652814      NA
## coef.var       0.4008183    0.1015326      NA
```

```
mean(sports$Score)
```

```
## [1] 307.3684
```

```
median(sports$Score)
```

```
## [1] 315
```

```
mean(regular$Score)
```

```
## [1] 327.6316
```

```
median(regular$Score)
```

```
## [1] 325
```