4.2.1 Exercise

Felipe Rodriguez

2022-01-07

Import Libraries

```
library(dplyr)
```

```
##
## 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)
```

Read File

```
# Read in Scores file
setwd('/Users/feliperodriguez/OneDrive - Bellevue University/Github/dsc520/')
scores <- read.csv(file='data/scores.csv')
```

What are the observational units in this study? Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?

```
# Display a sample of the data
head(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    10   265 Regular
```

The observational units of this study are score and count. The variables will be analyzed to compare the differences between the sections.

The quantitative variables are count and score since they are represented by an amount. The categorical variable is section which represents the group the other variables are in.

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

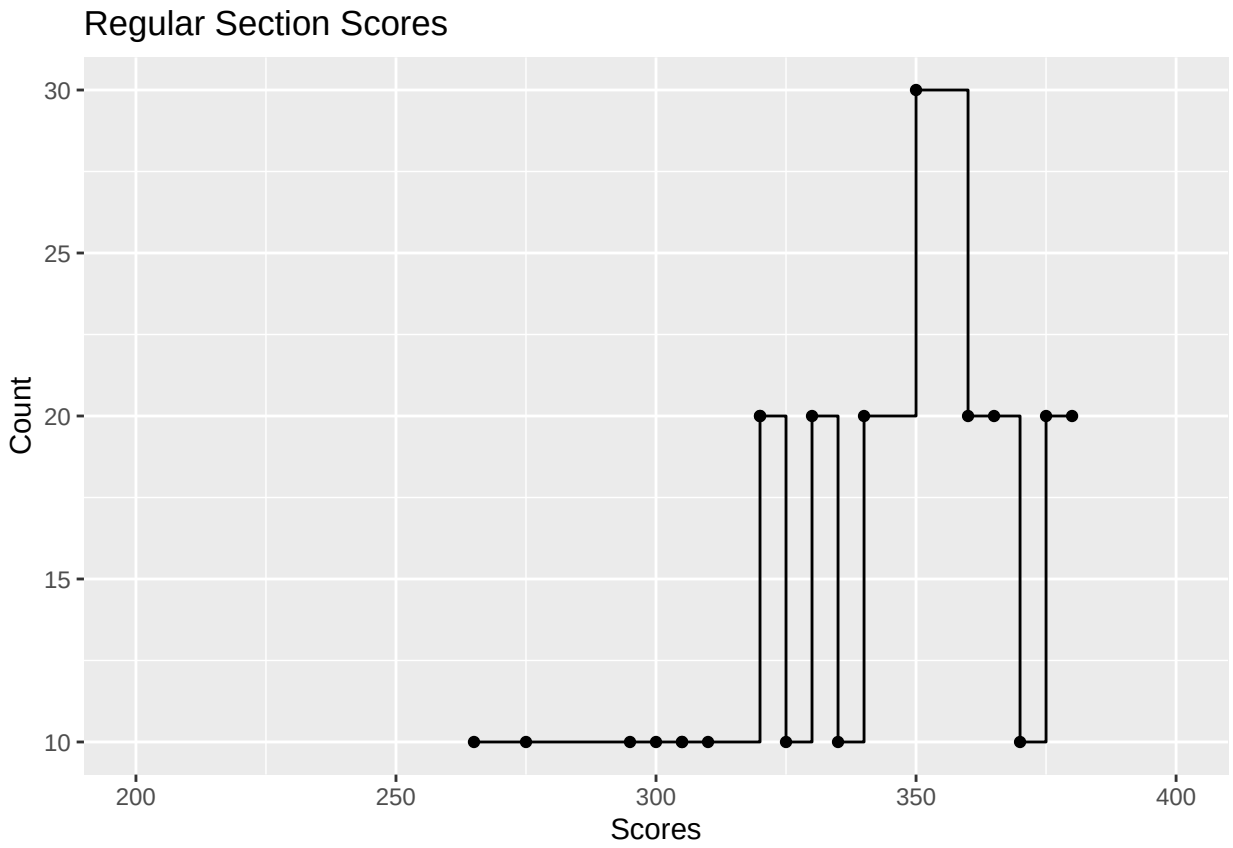
```
# New df that contains Regular Section
regular_section <- filter(scores, Section=='Regular')
# Displays Regular Section
head(regular_section)
```

```
##   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
```

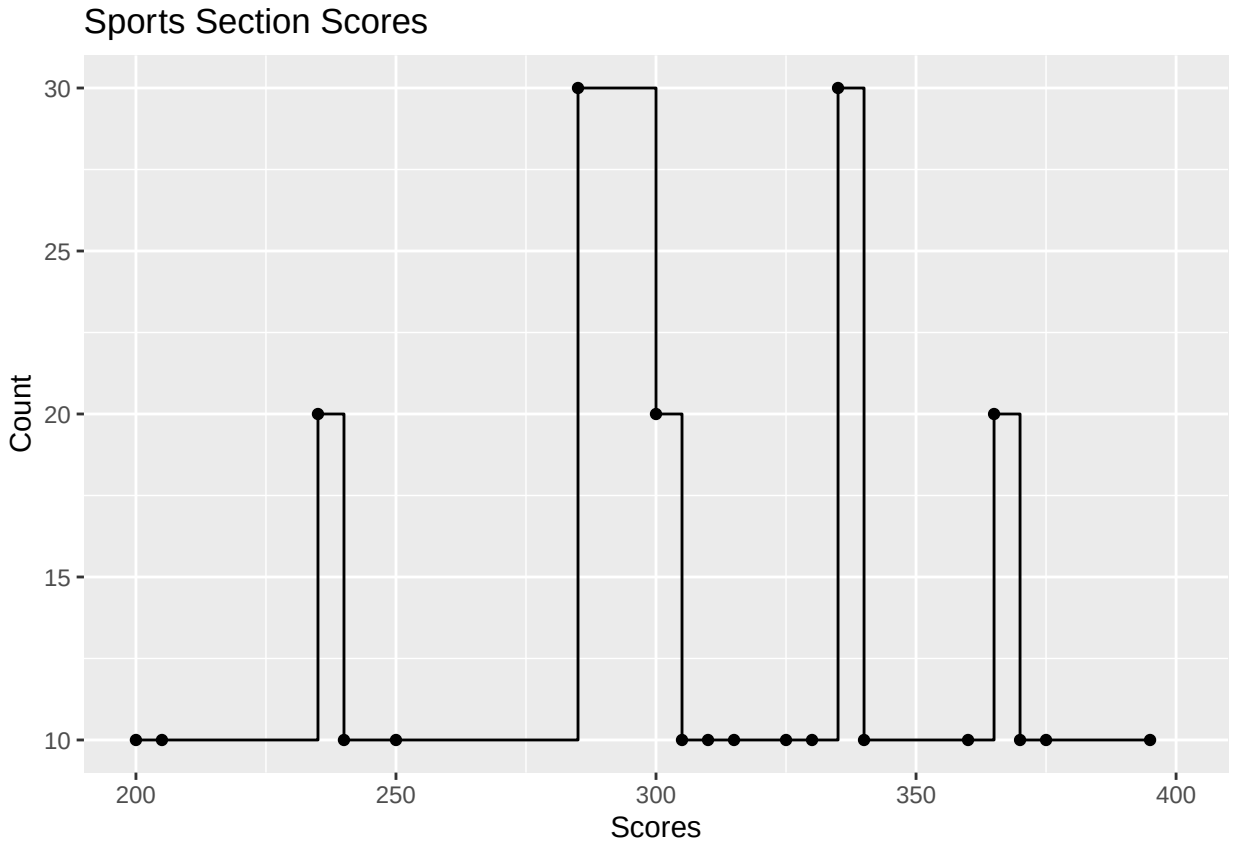
```
# New df that contains Sports Section
sports_section <- filter(scores, Section=='Sports')
head(sports_section)
```

```
##   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
```

```
regular_plot <- ggplot(regular_section, aes(x=Score, y=Count)) + ggtitle('Regular Section Scores') + xla
regular_plot
```




```
sports_plot <- ggplot(sports_section, aes(x=Score, y=Count)) + ggtitle('Sports Section Scores') + xlab(
sports_plot
```



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.

By looking at the visuals, the Regular Section has higher counts for scores above 300. This gives an indication that the Regular Section score more points than the other.

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.

Yes, the regular sections scored higher scores. The lowest score in the Regular Section was 260 while the lowest score in the Sports Section was 200. Look at the graphs, the majority of students scored above 330 in the Regular Section, while the majority of students score around 300 in the Sports Section.

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

Another variable that can be influencing this point distribution is the type of score. For example the scores are not defined by whether they are homework scores, test scores, or participation. This additional factors can change the distribution by further partitioning the data.