Homework 2

Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. Only **one person** in the group should submit to Canvas.
Due: check on Canvas.

1. The following questions use the data collected from the anonymous survey collected at the beginning of the course and from previous semesters. The dataset can be downloaded from the Datasets module on Canvas. Perform the following steps.

   a. Load the survey data into a variable called "survey". Use built-in function `read.csv()`
      `survey <- read.csv("C:/Users/username/Downloads/surveydataFall2023.csv")`
      Or you can set the working directory in RStudio by `Session->Set Working Directory->Choose Directory` and then call
      `survey <- read.csv("surveydataFall2023.csv")`

```
> survey <- read.csv("Downloads/surveydataFall2023.csv")
```

   b. How many rows are there in the data? (code, answer)

```
> nrow(survey)
[1] 561
```

   c. Are there any NAs in the dataset? Hint: use complete.cases()

```
> any(is.na(survey))
[1] TRUE
```

   d. How many rows have at least one NA?

```
> nrow(survey) - sum(complete.cases(survey))
[1] 17
```

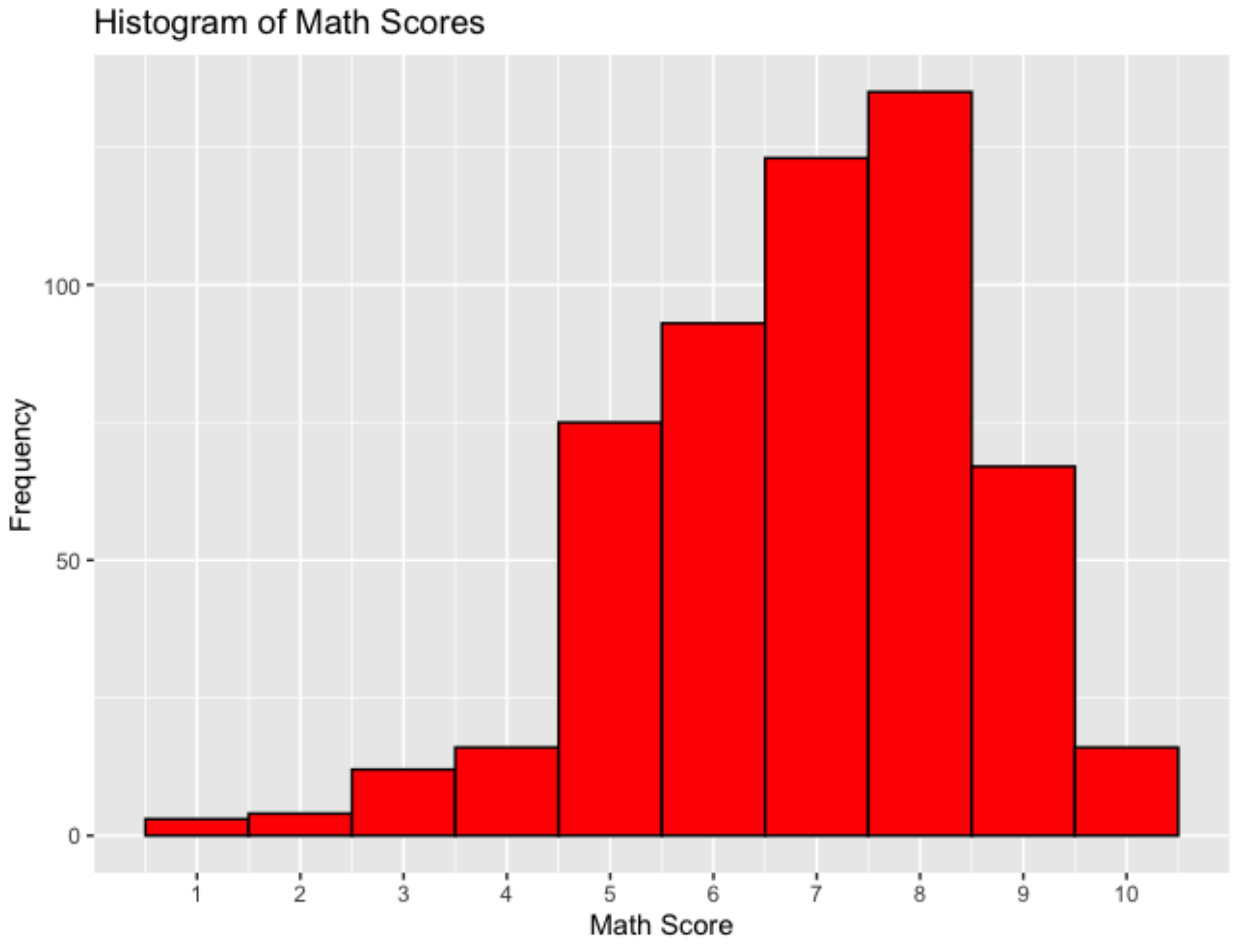
   e. Write R code to remove all rows which contain an NA value. Give your R code. How many rows does your data contain?

```
> survey_clean <- survey[complete.cases(survey), ]
> nrow(survey_clean)
[1] 544
```

2. Use the same data from the previous question (after removing NAs) to generate the following graphs. All plots should use **ggplot**. Include **both** the R code and paste the plot as an image

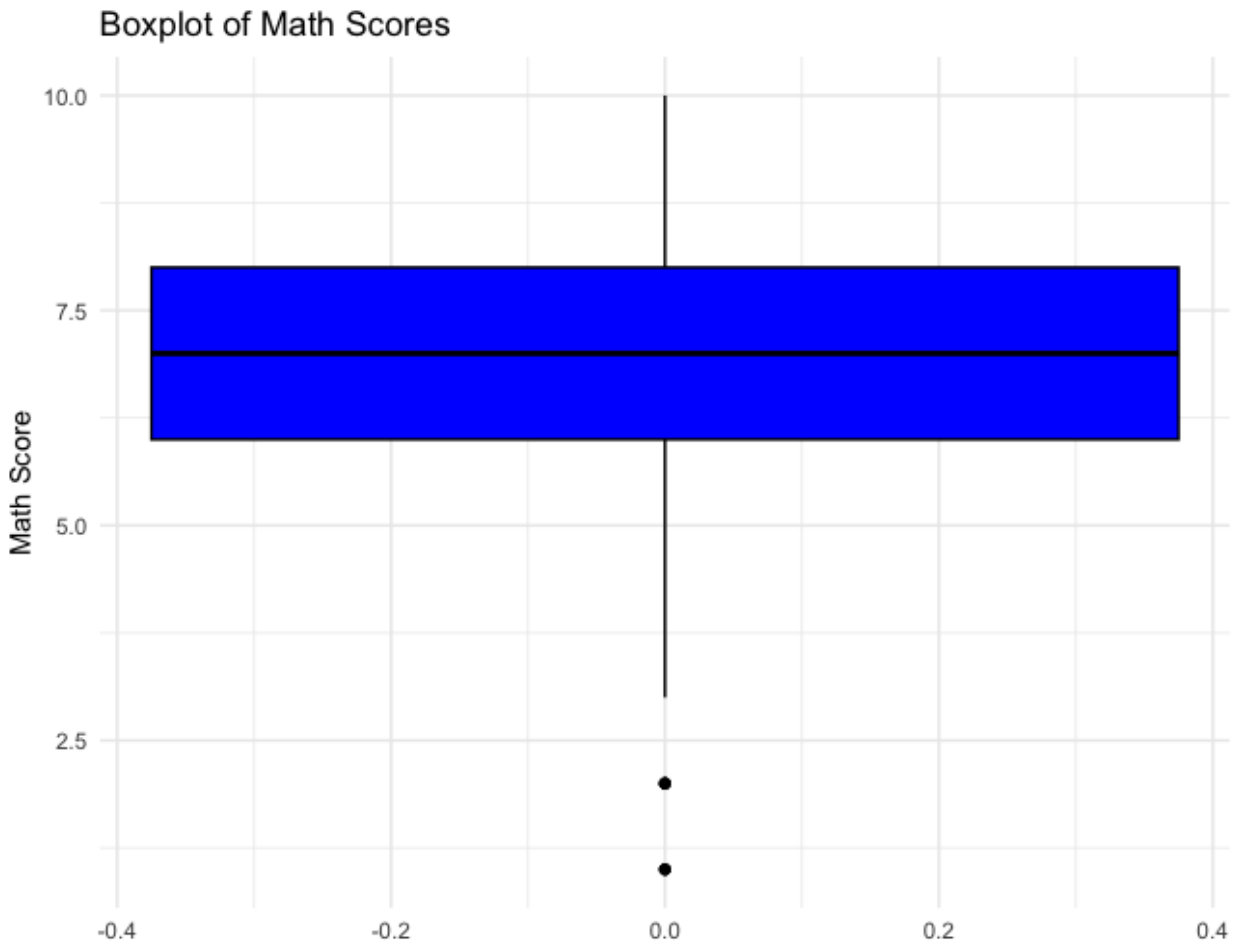
a. Plot a histogram of variable *Math*
b. The plot above likely has x-axis "breaks" not aligned with the bars. Provide your own breaks to match the bars. Hint: use the breaks argument in scale_x_continuous().

```
> ggplot(survey_clean, aes(x = Math)) + geom_histogram(bins = 10,
fill = "red", color = "black") + labs(title = "Histogram of Math
Scores", x = "Math Score", y = "Frequency") +
scale_x_continuous(breaks = seq(1, 10, by = 1))
```



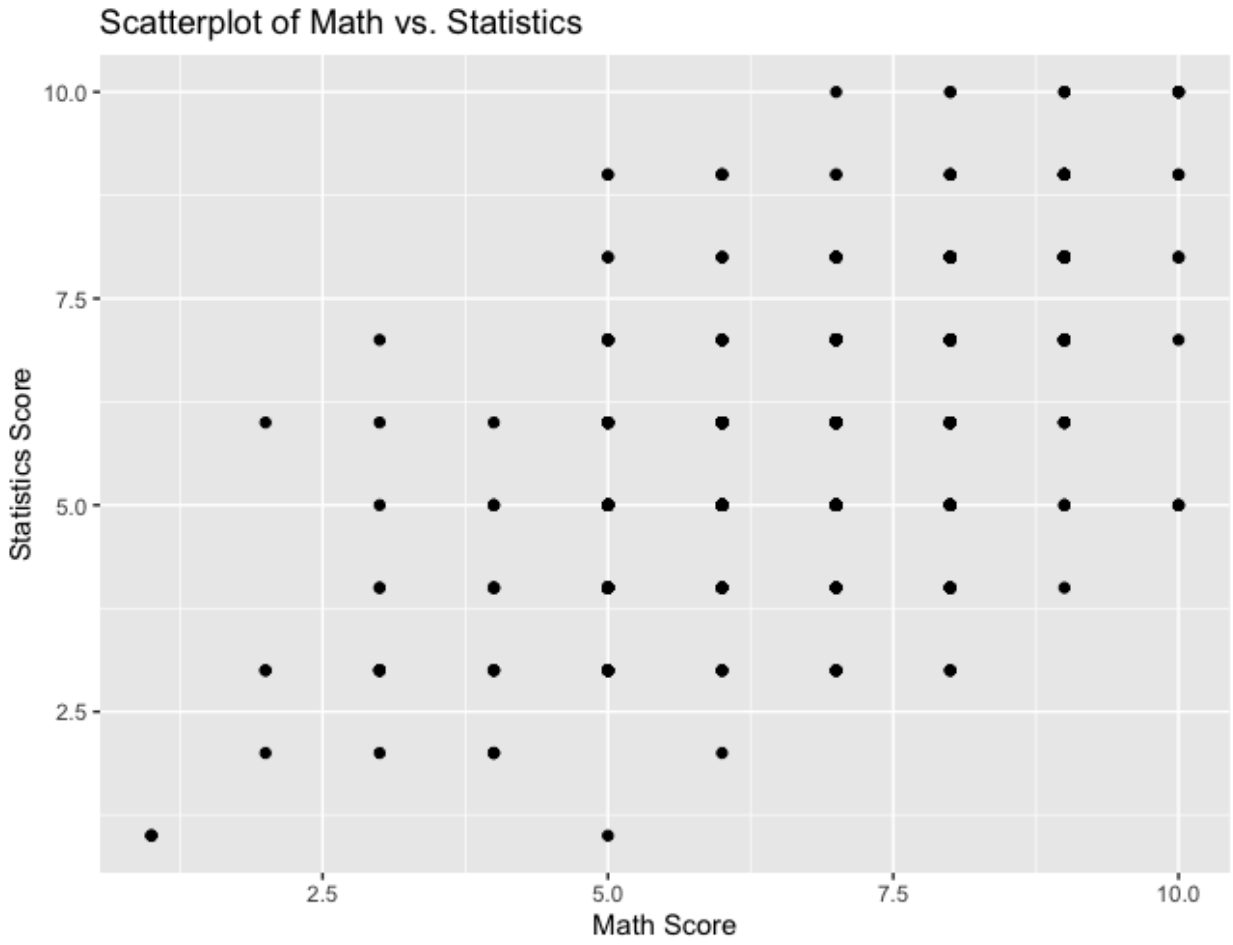
c. Plot a boxplot of variable *Math*

```
> ggplot(survey_clean, aes(y = Math)) + geom_boxplot(fill = "blue",
color = "black") + labs(title = "Boxplot of Math Scores", y = "Math
Score") + theme_minimal()
```



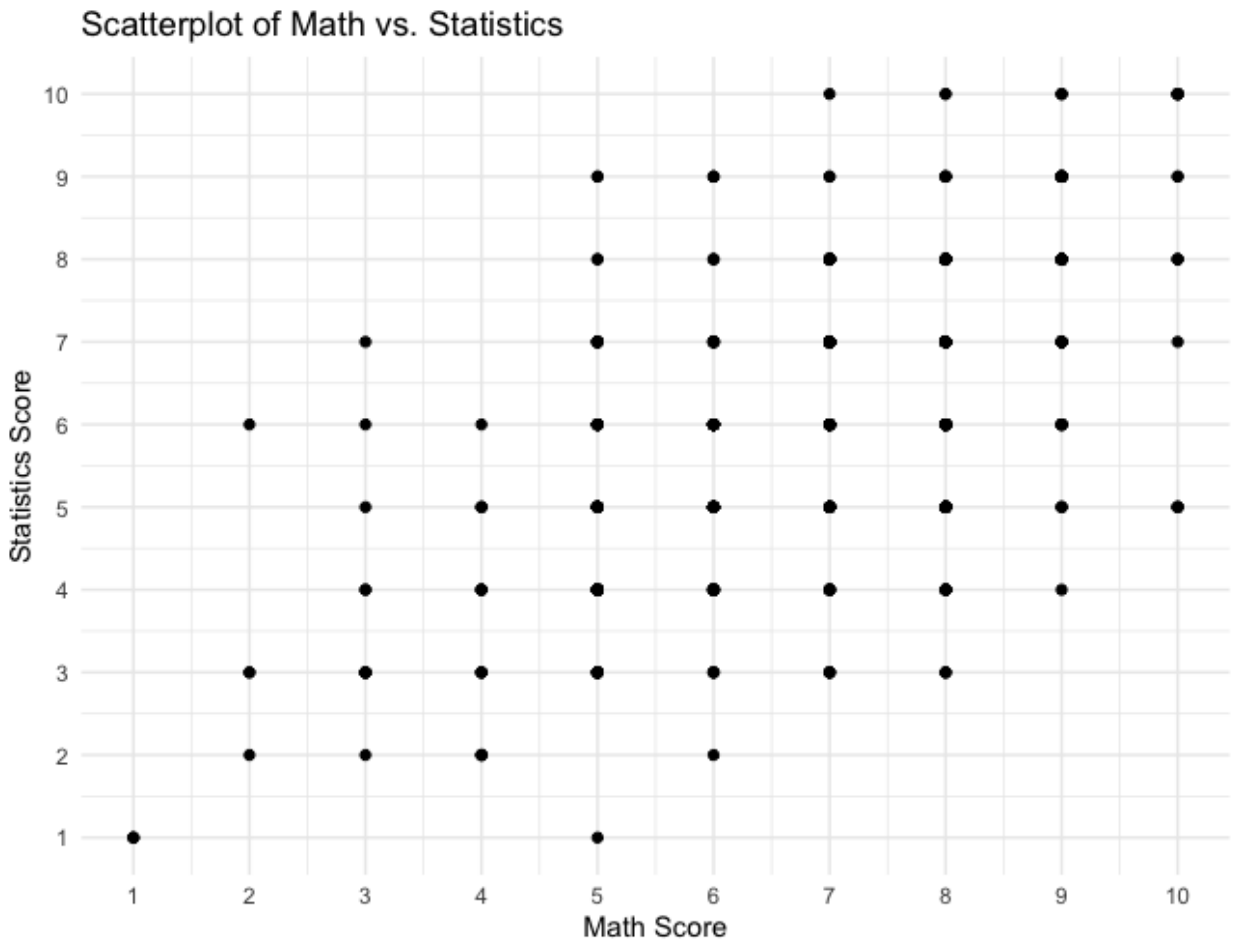
d. Plot a scatterplot of variables *Math* and *Statistics*.

```
> ggplot(survey_clean, aes(x = Math, y = Statistics)) + geom_point()
+ labs(title = "Scatterplot of Math vs. Statistics", x = "Math
Score", y = "Statistics Score")
```



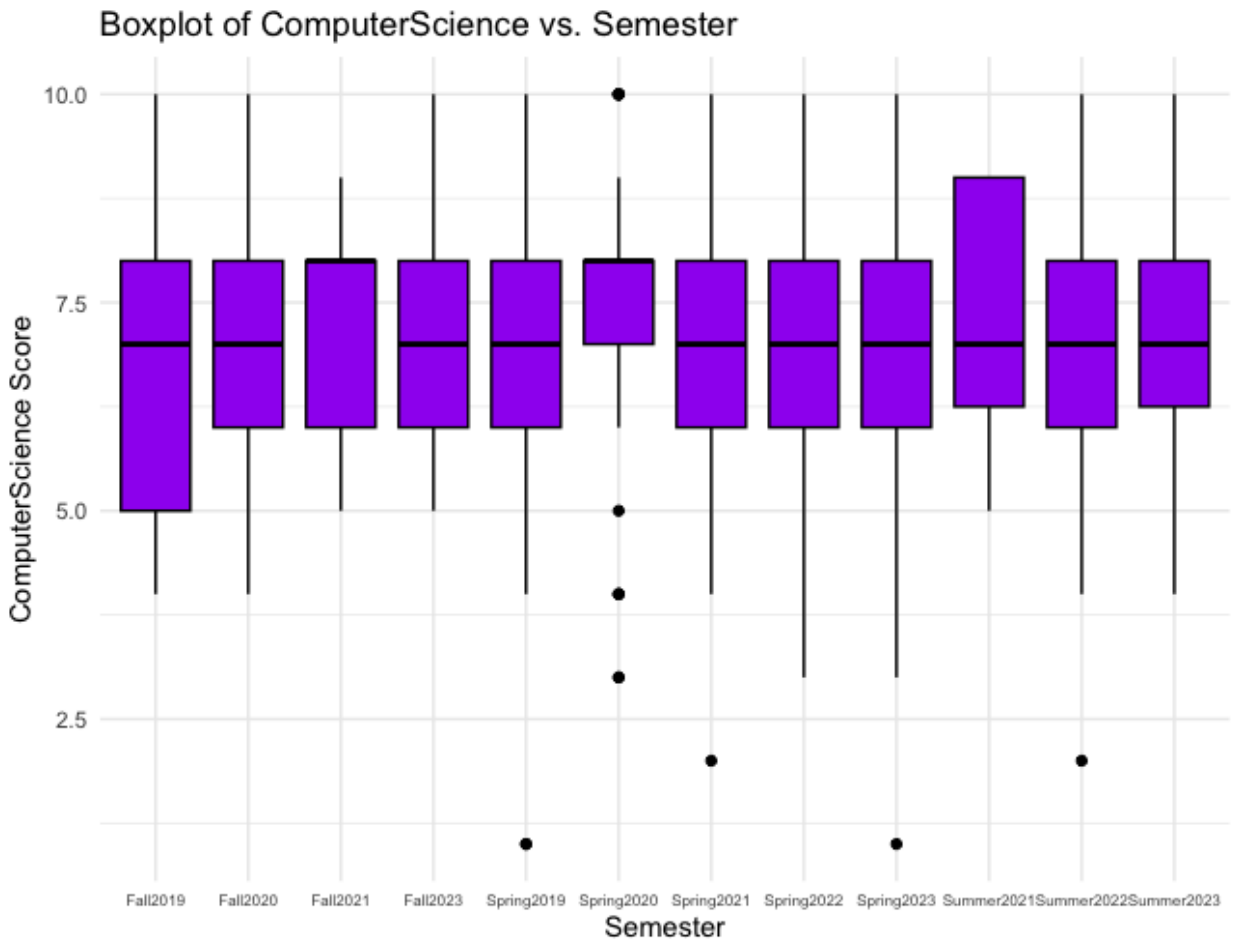
e. Redraw the previous scatterplot but also:
   i. Add more descriptive x and y-axis labels, add a title that should be the names of all group members, set both x-axis and y-axis limits to (1,10), and make sure x-axis and y-axis breaks are aligned with the points.

```
> ggplot(survey_clean, aes(x = Math, y = Statistics)) + geom_point()
+ labs(title = "Scatterplot of Math vs. Statistics", x = "Math
Score", y = "Statistics Score") + xlim(1, 10) + ylim(1, 10) +
scale_x_continuous(breaks = seq(1, 10, by = 1)) +
scale_y_continuous(breaks = seq(1, 10, by = 1)) + theme_minimal()
```



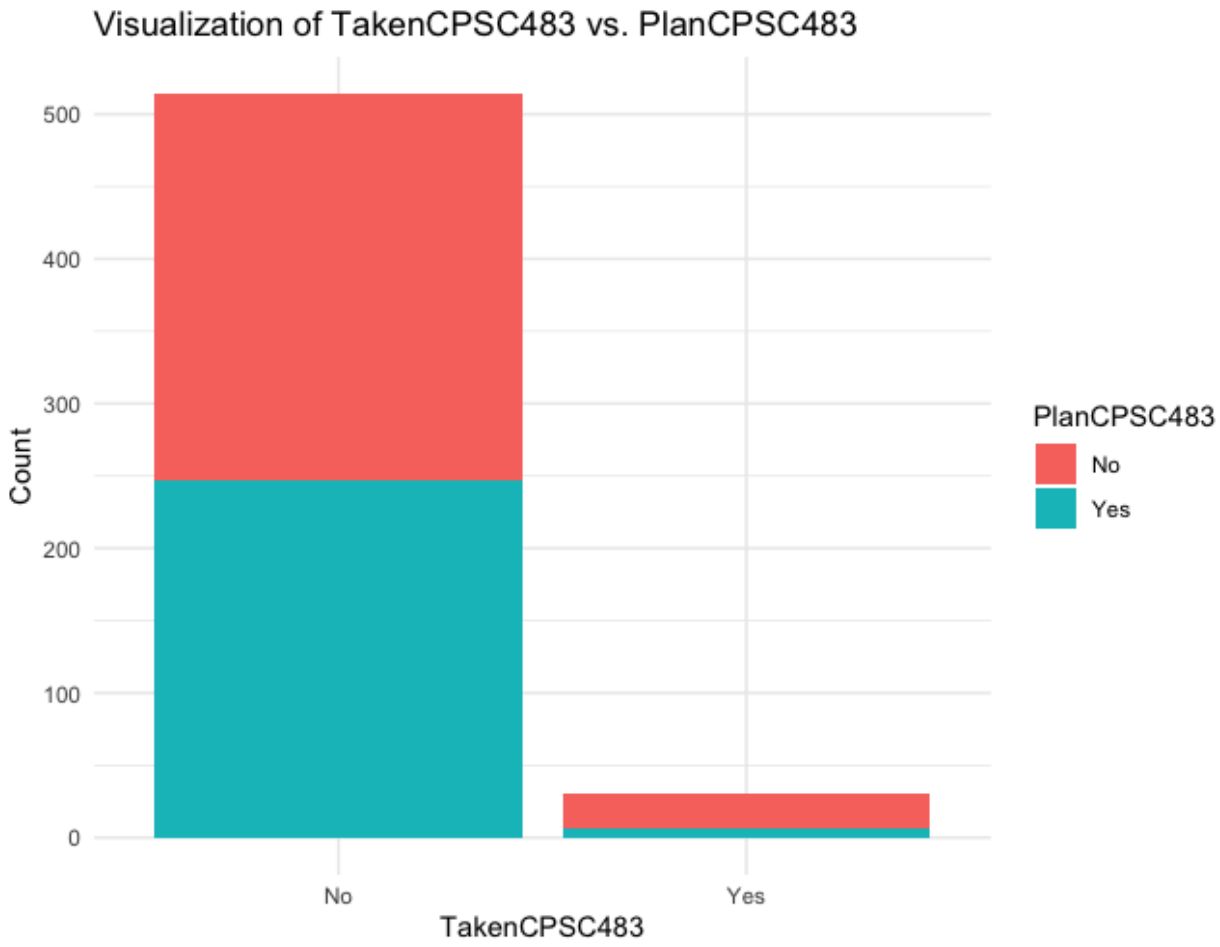
f. Plot a boxplot of variable *ComputerScience* vs. *Semester* (Note: you can plot either horizontally or vertically)

```
> ggplot(survey_clean, aes(x = as.factor(Semester), y =
ComputerScience)) + geom_boxplot(fill = "purple", color = "black") +
labs(title = "Boxplot of ComputerScience vs. Semester", x =
"Semester", y = "ComputerScience Score") + theme_minimal() +
theme(axis.text.x = element_text(size = 6))
```



g. Visualize the two categorical variables *TakenCPSC483* and *PlanCPSC483*.

```
> ggplot(survey_clean, aes(x = TakenCPSC483, fill = PlanCPSC483)) +
geom_bar() + labs(title = "Visualization of TakenCPSC483 vs.
PlanCPSC483", x = "TakenCPSC483", y = "Count") + theme_minimal()
```


Visualization of TakenCPSC483 vs. PlanCPSC483
Count
500
400
300
200
100
0
No
Yes
TakenCPSC483
PlanCPSC483
No
Yes