STA141 Midterm Exam 1

Richard McCormick

2023-10-05

The following code block if for the packages you want to use. I've included those necessary for data.

INSTRUCTIONS: You may use any materials available, including your notes, textbook, and online information. If any information is used outside of the textbooked and in-class notes, details must be given on what was used and how it works. **You may NOT share solutions among classmates.** The exam has no time limit besides the due date of Tuesday (10/10/2023) before 11:00 AM.

Please prepare your solutions using the RMD file provided. You may change options to suit your style, but be sure to keep the document organized. *Justify all free response answers. Type solutions after each prompt and try to keep your PDF organized.* The points for each question are given.

Please submit your exam as an organized PDF document directly from RMD. The solutions should be presented in the order the questions were asked. If there are problems with your PDF you will be asked to resubmit. Organization, clarity and correct preparation of solutions will be worth **5 points**.

This assignment is worth a total of 100 points, which is 20 percent of you overall grade for this course.

Due Tuesday October 10th, 2023 before 11:00 AM.

Exam Questions

Question 1 (10 points)

The President of NAU was told by a Department Chair that exams were too difficult. The Chair surveyed students in the data science major to see if they agreed by asking 'Do you believe that your exams are too easy?'. Would the data collected be:

(a) Representative?

The data is not representative. The Chair is only polling Data Science students, which means that the final data can only represent Data Science students, not all students at NAU.

(b) Sufficient?

The data collected is not sufficient to answer the question. The question, "Do you believe that your exams are too easy" is not enough to answer whether or not exams are too hard. For example, many students could believe that exams are not too easy, but also that they are not too hard.

(c) Unbiased?

The data is collected in an unbiased way. The collector of the data, the Chair, has no obvious interest in swaying the results one way or another.

Explain all of your answers!

Question 2 (10 points)

The President of NAU was told by a Department Chair that exams were too difficult. The Chair sampled the GPA of all students to see if there was evidence of many students struggling academically. Would the data collected be:

(a) Representative?

The data would be representative, because all students are being included in the survey.

(b) Sufficient?

The data is sufficient. GPA is a direct measurement of academic performance, so using this as a measure of whether or not students are academically struggling seems appropriate. **However**, in terms of the President's original question of whether exams are too difficult or not, this would not be sufficient. For example, students may do fine on exams but do poorly on homework, lowering their GPA.

(c) Unbiased?

The data would be unbiased. The Chair has no reason to want to influence the results of the data in either direction (assuming the Chair made their conclusions after the data was collected and not before).

Explain all of your answers!

Question 3 (10 points)

If we wanted to observe the relationship between the value of houses in Flagstaff and their floor area, number of bedrooms, number of bathrooms, distance from downtown, and the year it was built.

(a) What would be our explanatory variable(s)?

The explanatory variables would be floor area, number of bedrooms, number of bathrooms, distance from downtown, and the year the house was built.

(b) What would be our response variable(s)?

The response variable is the value of a house in Flagstaff.

(c) Would the floor area be a numerical or categorical variable?

Floor area would be numerical, as there are not a standardized list of floor areas and each house would likely have a different floor area measured in square feet.

(d) Would the number of bedrooms be a numerical or categorical variable?

The number of bedrooms would be a categorical variable. Most homes in Flagstaff have roughly the same number of bedrooms (within a certain range), so it would make sense to use a categorical variable.

(e) Would the distance to downtown be a numerical or categorical variable?

The distance to downtown would be a numerical variable. The distance would be slightly different for each house, so it makes most sense to have a numerical variable instead of many different levels.

Explain your answers!

Question 4 (10 points)

Run the following code block to save a sample of data called `my.q4`.

```
set.seed(141)
q4.data <- data.frame(sport = sample(c("basketball",
                                       "hockey",
                                       "football",
                                       "volleyball",
                                       "golf"),
                                     1000,
                                     replace=TRUE,
                                     prob=c(0.25, 0.2, 0.4, 0.05, 0.1)))
```

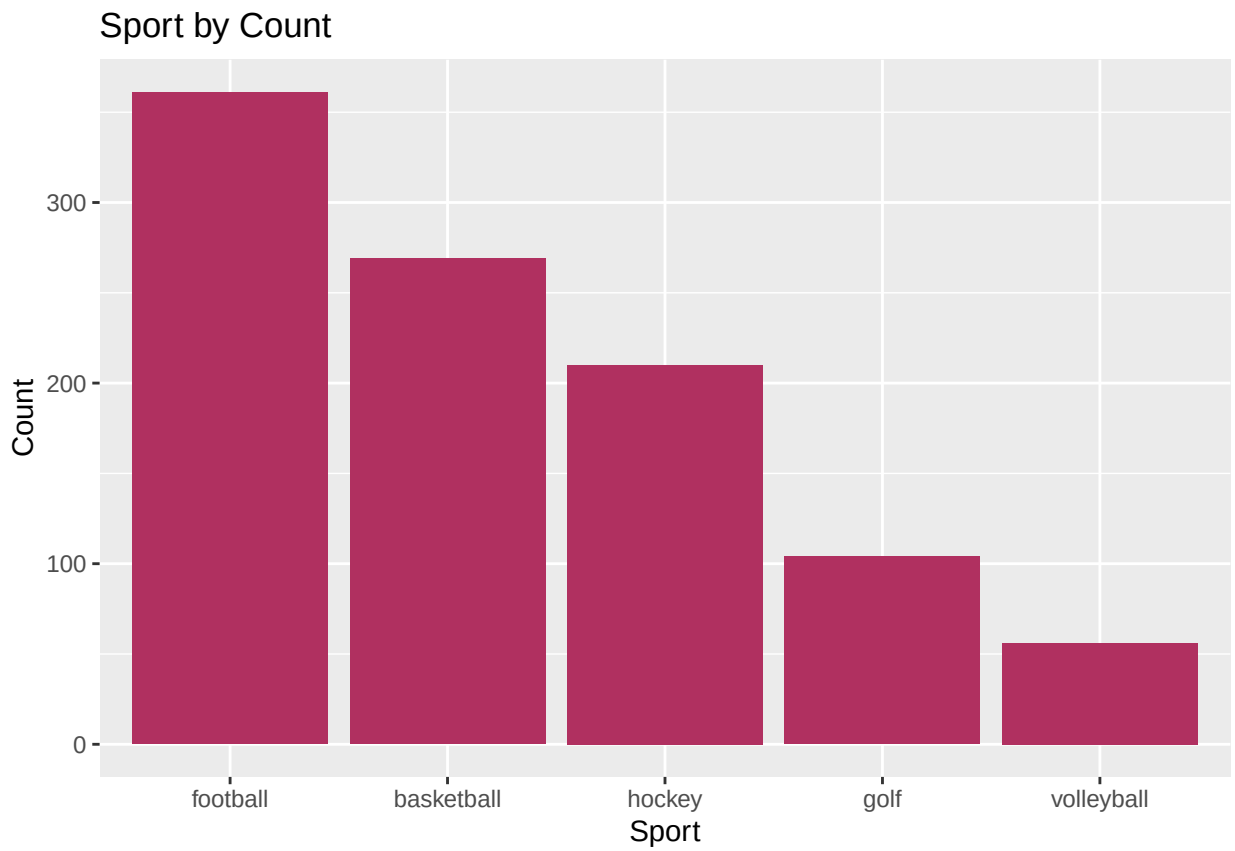
(a) Create a frequency distribution table of the `sport` variable.

```
table( q4.data )
```

```
## q4.data
## basketball   football       golf     hockey volleyball
##        269        361        104        210         56
```

(b) Plot a bar chart of the count of the `sport` variable. Order the bars in descending order of counts, and color the bars maroon.

```
q4.data <- q4.data %>% group_by( sport ) %>% mutate(count_name_occurr = n() )

sport_bar_chart <- ggplot( data=q4.data ) +
  geom_bar( aes( x=reorder( sport, -count_name_occurr ),
                 y=after_stat( count ) ), fill='maroon' ) +
  labs( title="Sport by Count", x="Sport", y="Count" )

sport_bar_chart
```



(c) Why is there a gap between bars when plotting a bar chart? *Hint: you might want to compare this to a histogram.*

There is a gap between the bars because the variable being plotted is categorical. In a histogram, where the variable is numeric, there are no gaps.

Question 5 (15 points)

Use the following code block to save a copy of the `Alfalfa` dataset from the `Stat2Data` package.

```
data("Alfalfa")
my.alfalfa <- Alfalfa
```

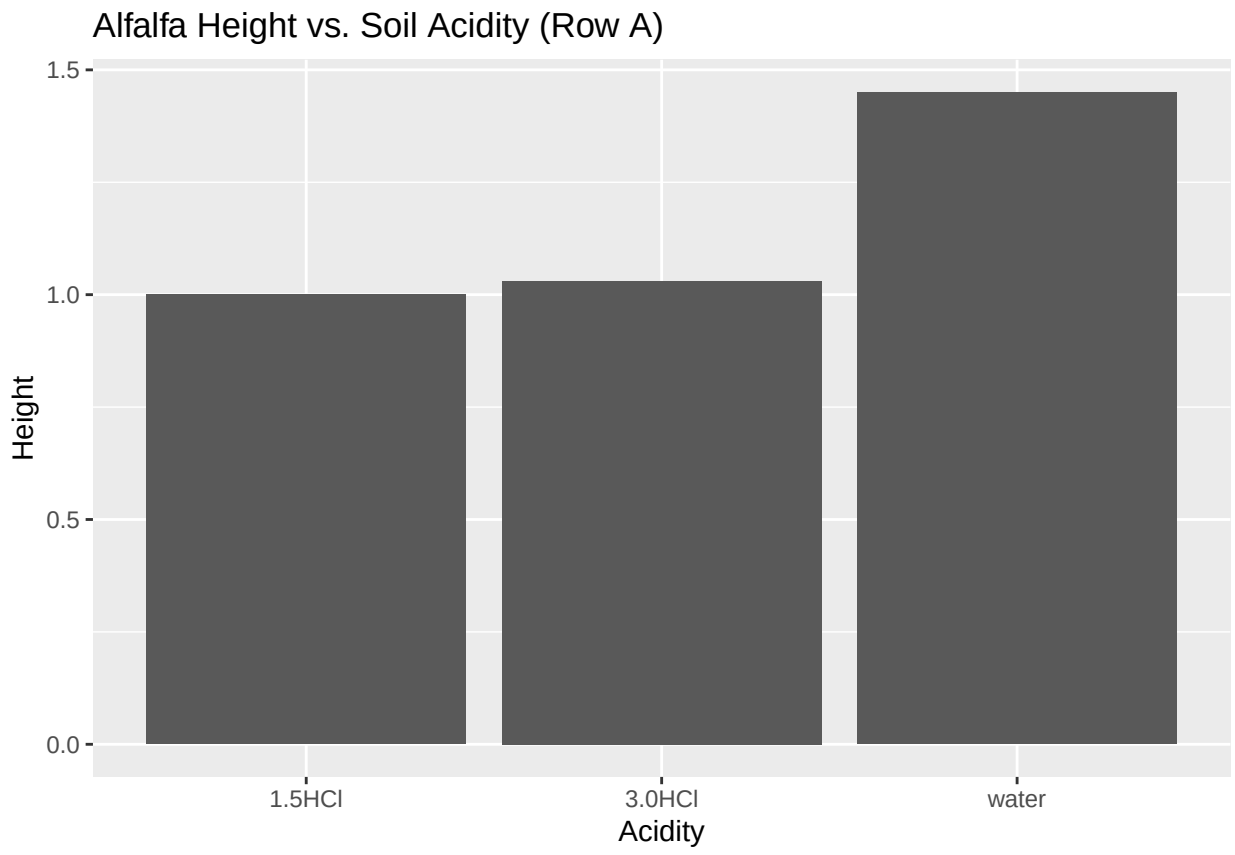
The data describes experiments run by students to see how an acidic environment might affect the growth of alfalfa seeds. The `Ht4` variable describes the height of the alfalfa after four days of growing; the `Acid` variable describes whether the seed was treated with water only, a moderate amount of acid, or a high quantity of acid; the `Row` variable describes how close to the window the experiment was located: with "a" being the closest and "e" being the farthest. Using your dataset answer the following:

(a) How many experiments were there in total? Were they evenly divided between water, moderate acid, and high acid?

There are 15 observations in total. They appear to be evenly divided between water, moderate acid, and high acid (in terms of number of data points). Water appears to have significantly higher growth than either of the acids.

(b) Create a bivariate bar chart of the height of the alfalfa against the acid for row `a` only. Do these experiments suggest that acid increases or decreases growth?

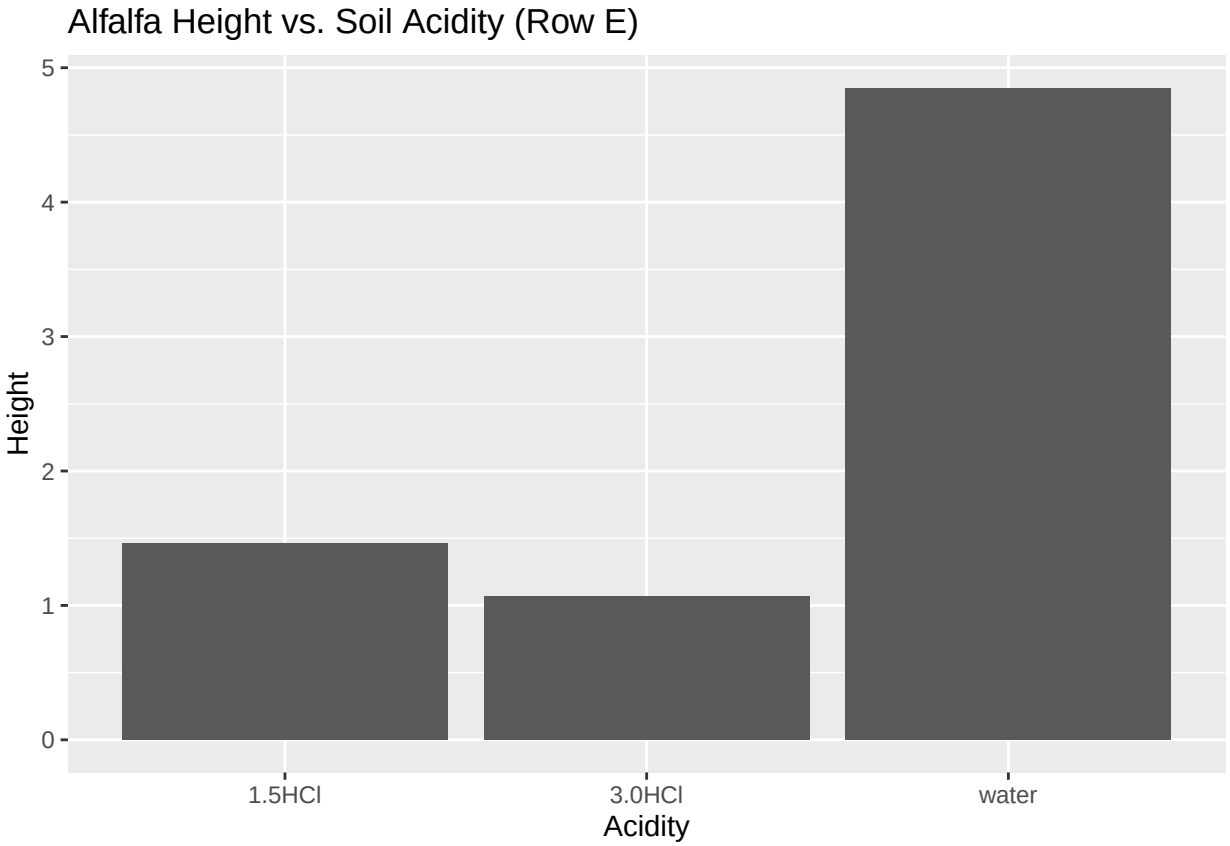
```
alfalfa_chart <- ggplot( data=my.alfalfa[my.alfalfa$Row == 'a', ] ) +
  geom_bar( stat='identity', aes( x=Acid, y=Ht4 ) ) +
  labs( title="Alfalfa Height vs. Soil Acidity (Row A)", x='Acidity', y='Height' )

alfalfa_chart
```



This chart suggests that acidity reduces the height of alfalfa growth.

(c) Create a bivariate bar chart of the height of the alfalfa against the acid for row `e` only. Do these experiments suggest that being near the window increases or decreases growth? That is, compare this plot to that from part `(b)`.

```
alfalfa_chart <- ggplot( data=my.alfalfa[my.alfalfa$Row == 'e', ] ) +
  geom_bar( stat='identity', aes( x=Acid, y=Ht4 ) ) +
  labs( title="Alfalfa Height vs. Soil Acidity (Row E)", x='Acidity', y='Height' )

alfalfa_chart
```



Comparing the two charts does seem to indicate that being closer to the window helped the growth of the alfalfa. The plants with high acidity remained about the same height, but the plant with only water grew significantly higher than its counterpart in Row A, away from the window.

Question 6 (15 points)

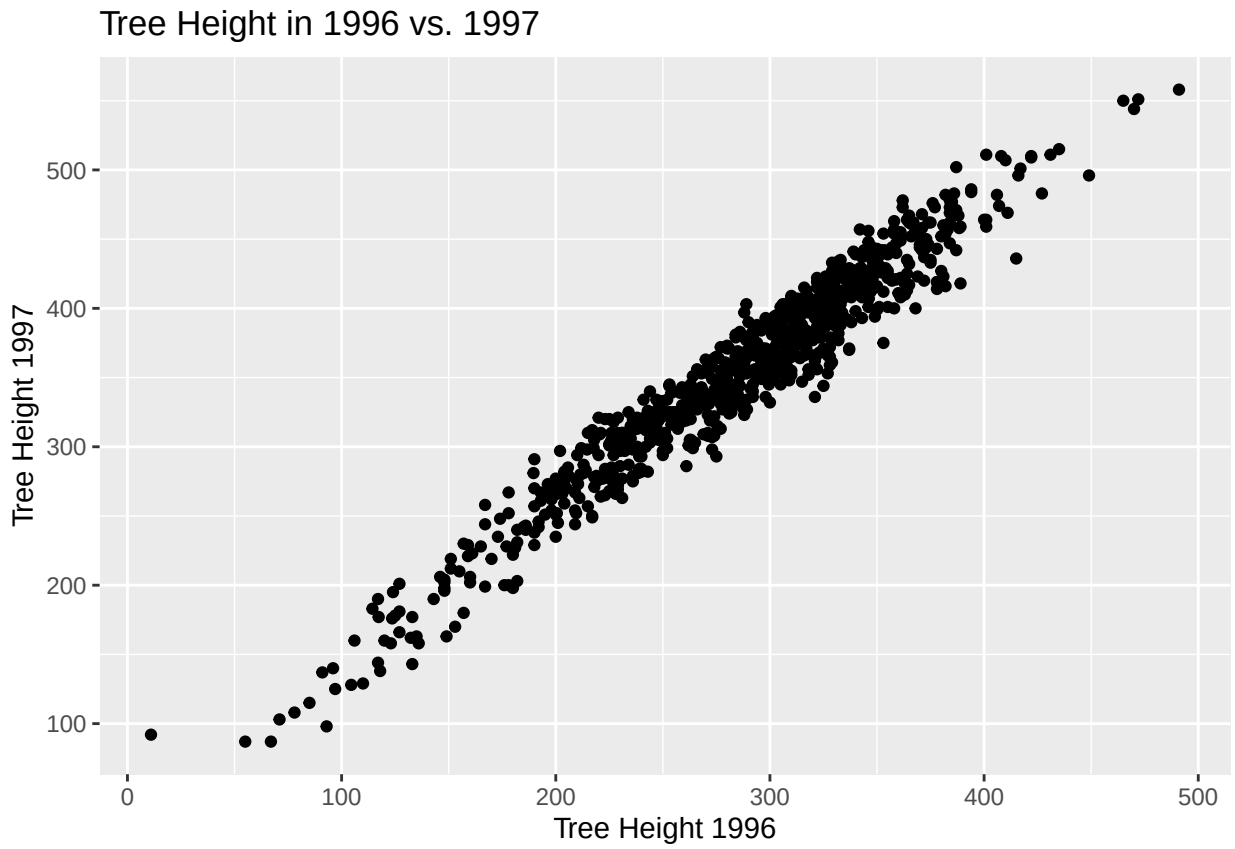
Use the following code block to save a copy of the `Pines` dataset as `my.pines` from the `Stat2Data` package.

```
data("Pines")
my.pines <- Pines
my.pines <- my.pines %>% filter((!is.na(Hgt90)) & (!is.na(Hgt96)) & (!is.na(Hgt97))) # this will just r
```

The data describes the growth of pine seedlings planted in 1990. The data contains 15 measurements on 1000 different pine trees. Using your dataset answer the following:

(a) Produce a scatterplot of tree height in 1997 against tree height in 1996.

```
pine_plot <- ggplot( data=my.pines ) +
  geom_point( aes( x=Hgt96, y=Hgt97 ) ) +
  labs( title="Tree Height in 1996 vs. 1997", x="Tree Height 1996",
        y="Tree Height 1997" )

pine_plot
```

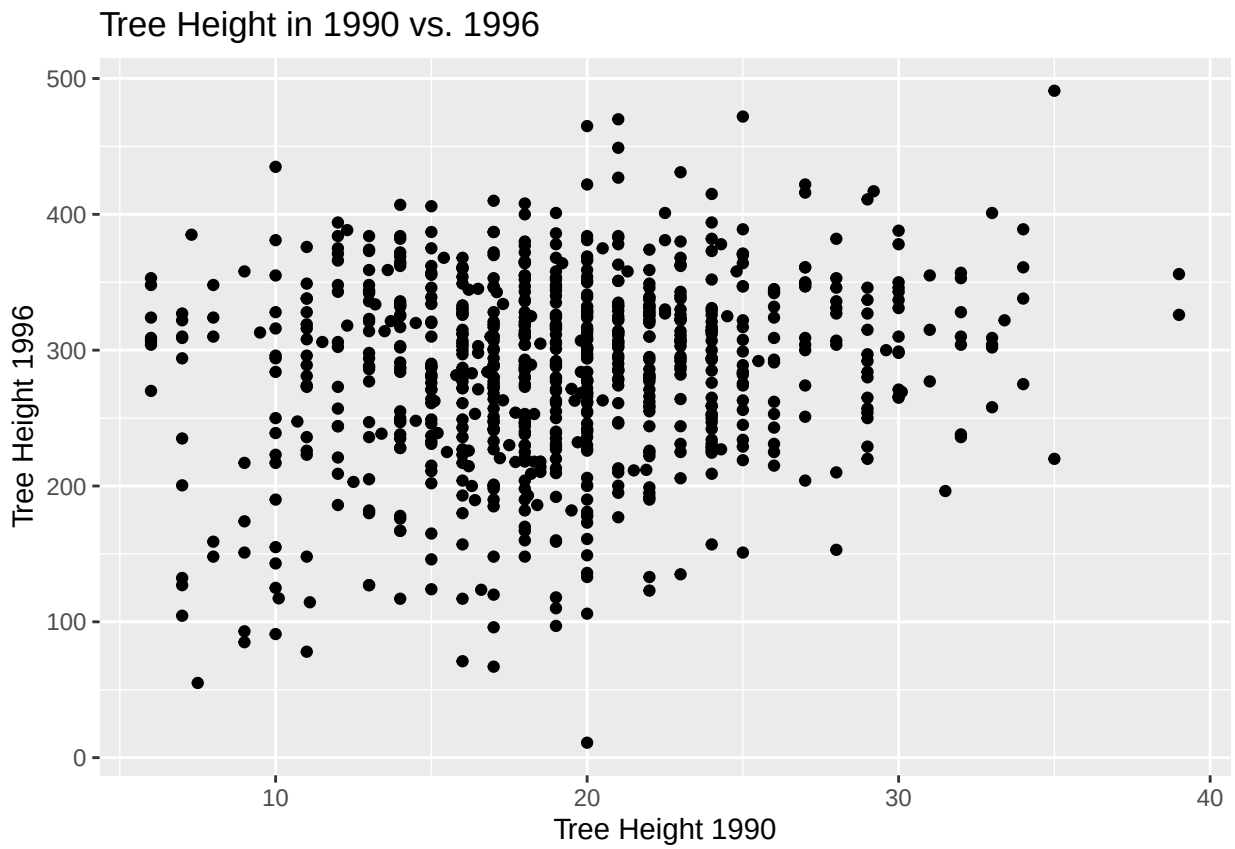


(b) If a tree was taller than average in 1996 did that result in it being taller than average in 1997?

(c) Now produce a scatterplot of tree height in 1996 against tree height in 1990.

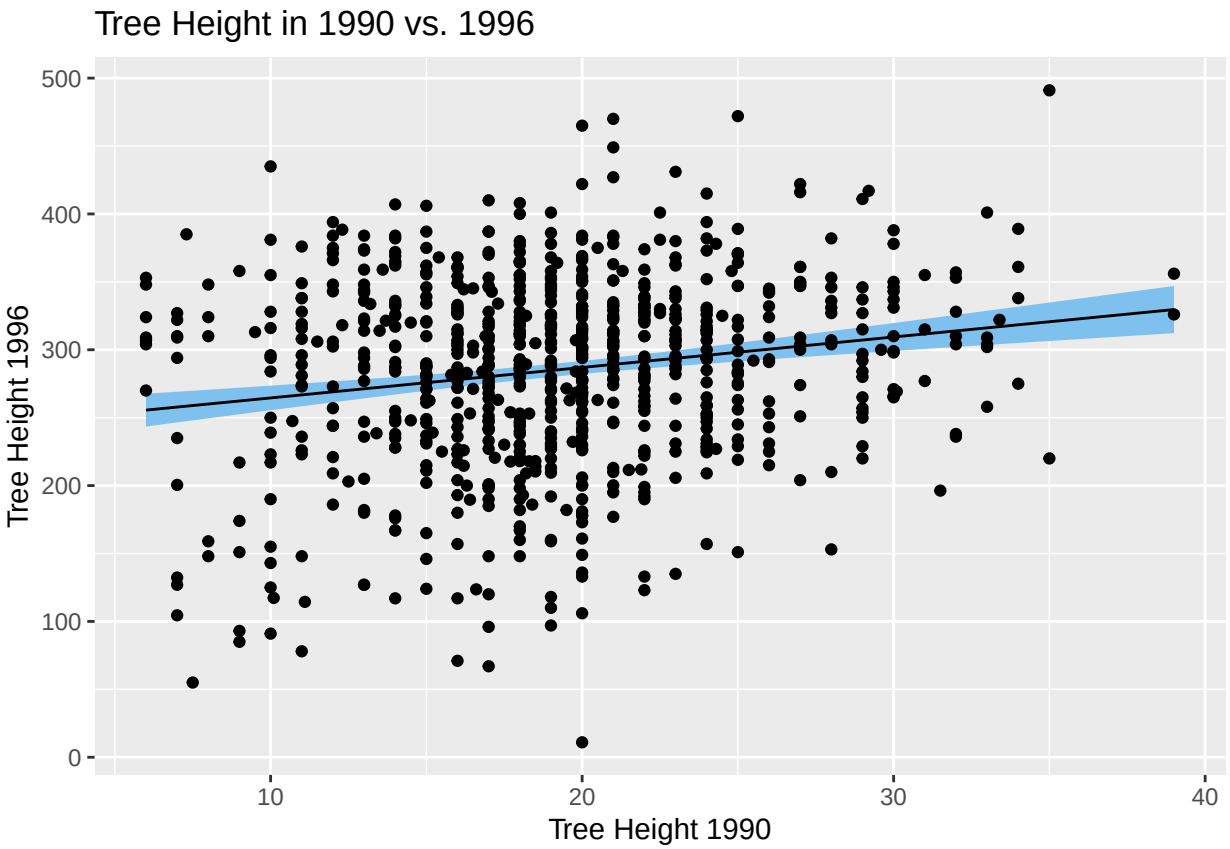
```
pine_plot <- ggplot( data=my.pines ) +
  geom_point( aes( x=Hgt90, y=Hgt96 ) ) +
```

```
  labs( title="Tree Height in 1990 vs. 1996", x="Tree Height 1990",
        y="Tree Height 1996" )

pine_plot
```



(d) If a tree was taller than average when planted did that result in it being taller than average in 1996? Use a line of best fit to support your answer.

```
SLR <- lm( my.pines$Hgt96 ~ my.pines$Hgt90 )
my.pines <- my.pines %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(SLR, newdata=., interval='confidence') )

pine_plot <- ggplot( data=my.pines ) +
  geom_ribbon( aes( x=Hgt90, ymin=lwr, ymax=upr ), fill='skyblue2' ) +
  geom_line( aes( x=Hgt90, y=fit ) ) +
  geom_point( aes( x=Hgt90, y=Hgt96 ) ) +
  labs( title="Tree Height in 1990 vs. 1996", x="Tree Height 1990",
        y="Tree Height 1996" )

pine_plot
```



From the data and the line of best fit, it does appear that the height of a tree when it is planted slightly correlates to height in 1996. Plants that were taller in 1990 tended to be slightly taller in 1996.

Question 7 (15 points)

Use the following code block to save a copy of the `BlueJays` dataset from the `Stat2Data` package.

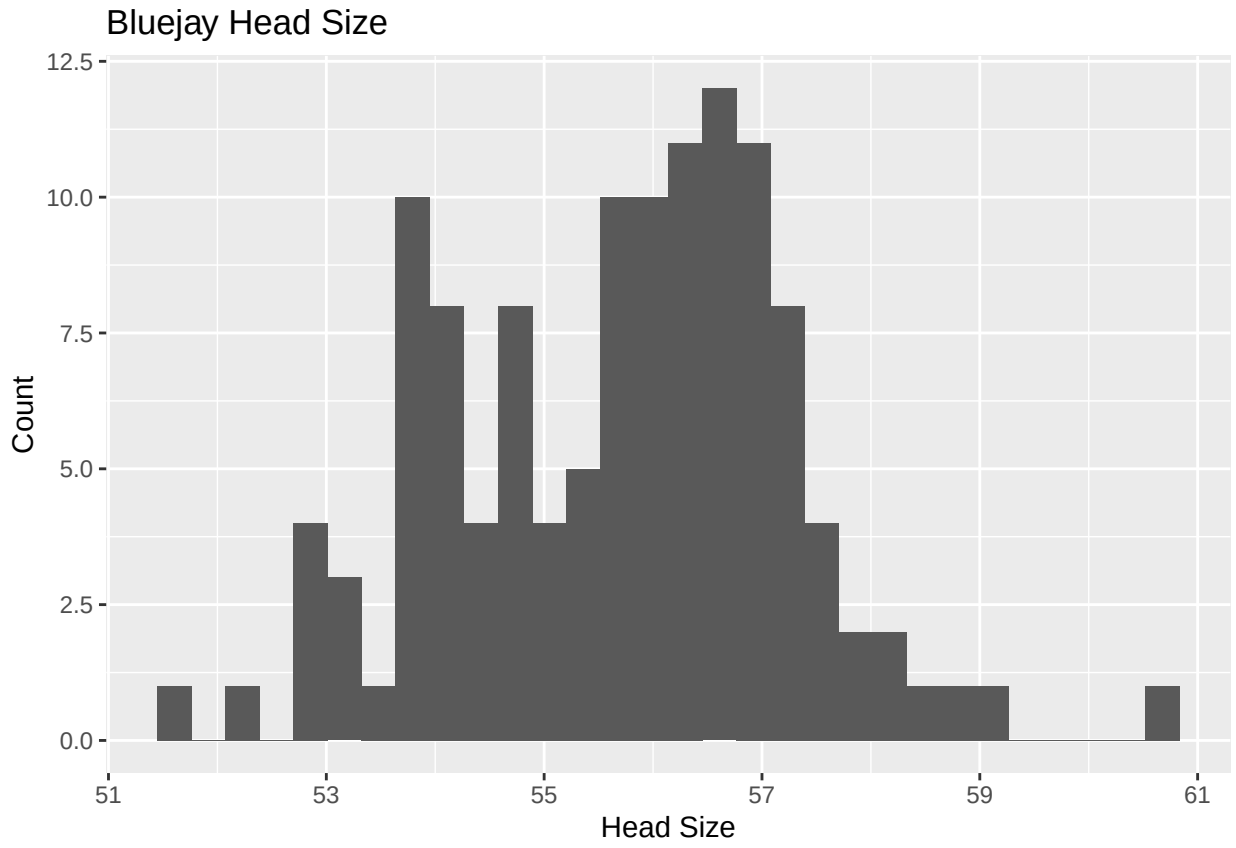
```
data("BlueJays")
my.bluejays <- BlueJays
```

The data describes body measurements of 123 blue jays (the birds, not the baseball team). Use your dataset to answer the following questions:

(a) Create a histogram for the `Head` variable for male blue jays.

```
bird_plot <- ggplot( data=my.bluejays ) +
  geom_histogram( aes( x=Head ) ) +
  labs( title="Bluejay Head Size", x="Head Size", y="Count" )

bird_plot
```

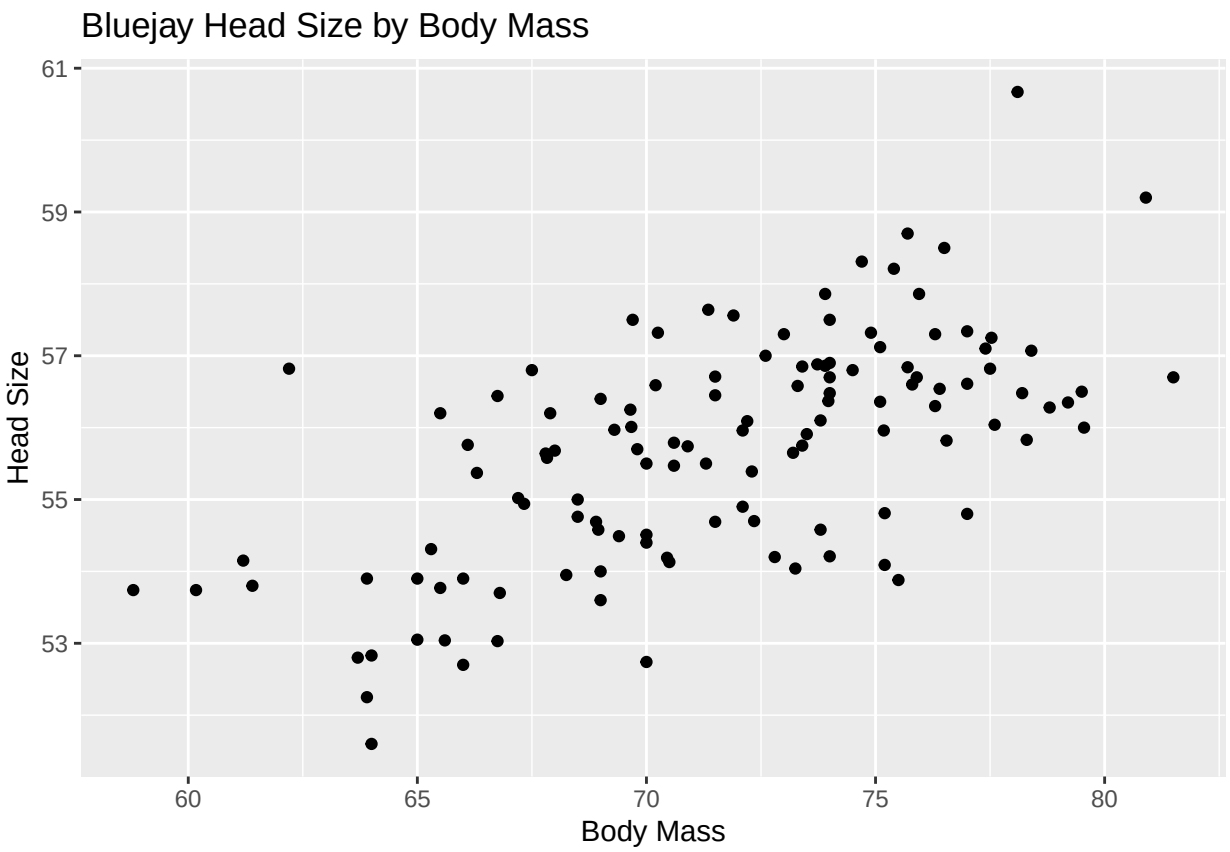
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The above is a histogram for Bluejay Head size by count.

(b) Do blue jays with higher body mass have longer heads? Use a scatterplot to support your answer.

```
bird_chart <- ggplot( data=my.bluejays ) +
  geom_point( aes( x=Mass, y=Head ) ) +
  labs( title="Bluejay Head Size by Body Mass", x="Body Mass", y="Head Size" )

bird_chart
```

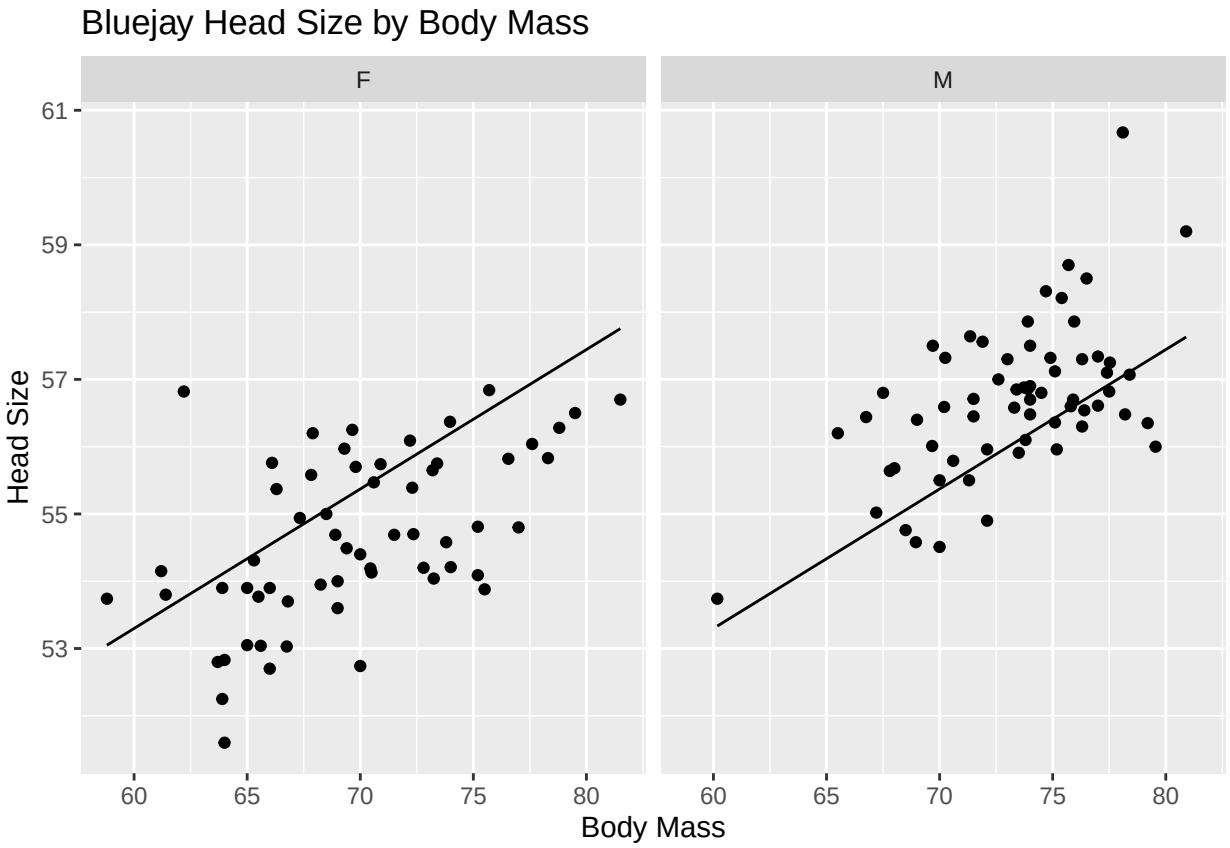


From the data and scatterplot, it appears that bluejays with a larger body mass tend to have larger heads.

(c) If I measure a new bird to have a Mass of 70g, how long would you expect its head to be if it were a male? Is the answer different if it were a female? *Hint: use the line of best fit because it's our best estimate!*

```
SLR.birds <- lm( my.bluejays$Head ~ my.bluejays$Mass )
my.bluejays <- my.bluejays %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(SLR.birds, newdata=., interval='confidence') )

male.birds <- my.bluejays[ my.bluejays[ 'KnownSex' ] == 'M', ]
m.Mass <- male.birds$Mass
SLR.birds.male <- lm( male.birds$Head ~ m.Mass )
male.birds <- male.birds %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(SLR.birds.male, newdata=., interval='confidence') )

female.birds <- my.bluejays[ my.bluejays[ 'KnownSex' ] == 'F', ]
f.Mass <- female.birds$Mass
SLR.birds.female <- lm( female.birds$Head ~ f.Mass )
```

```
female.birds <- female.birds %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(SLR.birds.female, newdata=., interval='confidence') )

bird_chart <- ggplot( data=my.bluejays ) +
  geom_point( aes( x=Mass, y=Head ) ) +
  geom_line( aes( x=Mass, y=fit ) ) +
  facet_grid( ~KnownSex) +
  labs( title="Bluejay Head Size by Body Mass", x="Body Mass", y="Head Size" )

bird_chart
```




```
print( paste("Predicted head size of a Female bird of 70g:",
             predict(SLR.birds.female, newdata=data.frame( f.Mass=70 ) ) ) )
```

```
## [1] "Predicted head size of a Female bird of 70g: 54.6716675470589"
```

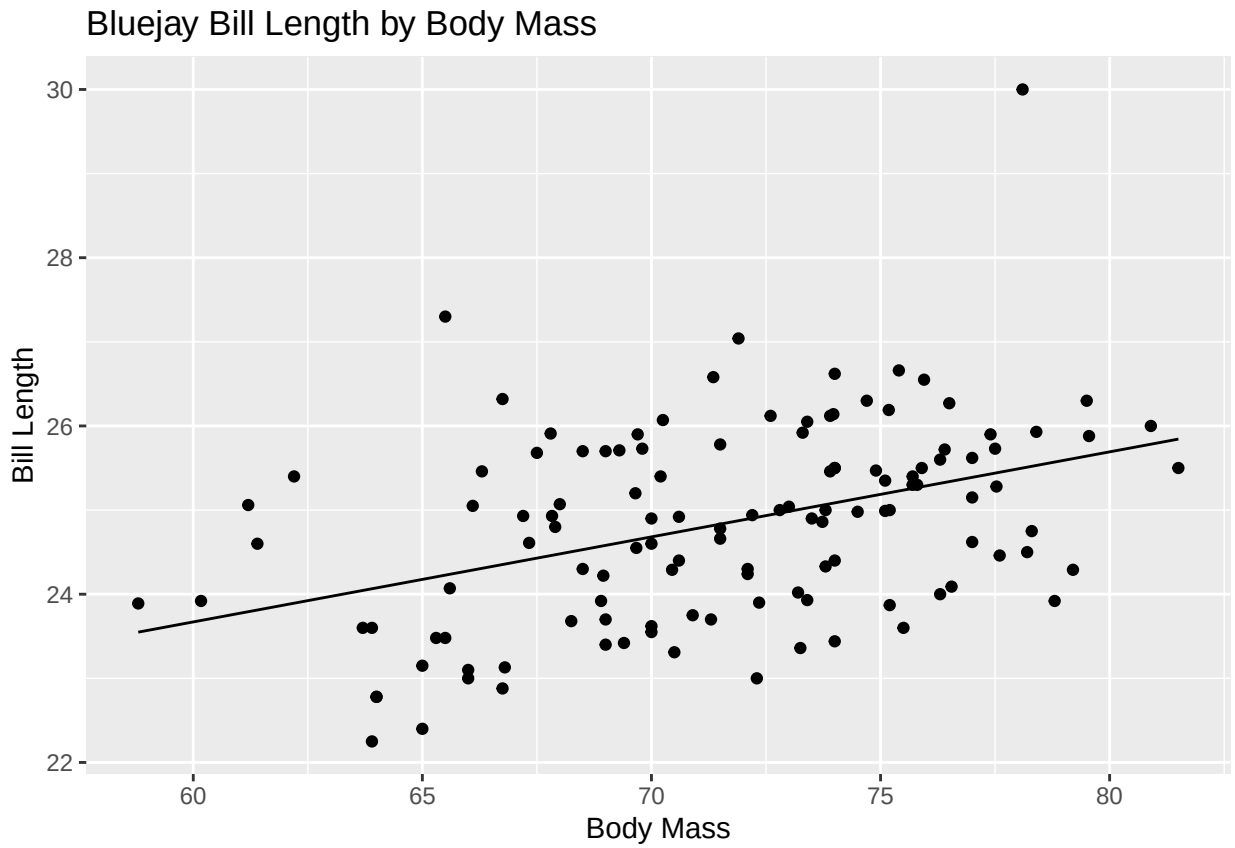
```
print( paste("Predicted head size of a Male bird of 70g:",
             predict(SLR.birds.male, newdata=data.frame( m.Mass=70 ))) )
```

```
## [1] "Predicted head size of a Male bird of 70g: 56.1444581028735"
```

We can see from the data that on average, the male birds have a higher than average head size, and the female birds have a lower than average head size. For a bird of 70 grams in mass, we could expect it to have a head size of 54.67mm if it is female, and 56.14mm if it is male.

(d) Do heavier blue jays have longer bills? Produce a scatterplot of bill length against body mass to support your answer. Include a line of best fit *if* it helps support your answer.

```
SLR.birds <- lm( my.bluejays$BillLength ~ my.bluejays$Mass )
my.bluejays <- my.bluejays %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(SLR.birds, newdata=., interval='confidence') )

bird_chart <- ggplot( data=my.bluejays ) +
  geom_point( aes( x=Mass, y=BillLength ) ) +
  geom_line( aes( x=Mass, y=fit ) ) +
  labs( title="Bluejay Bill Length by Body Mass", x="Body Mass", y="Bill Length" )

bird_chart
```



From the data, it does appear that heavier bluejays do tend to have longer bills.

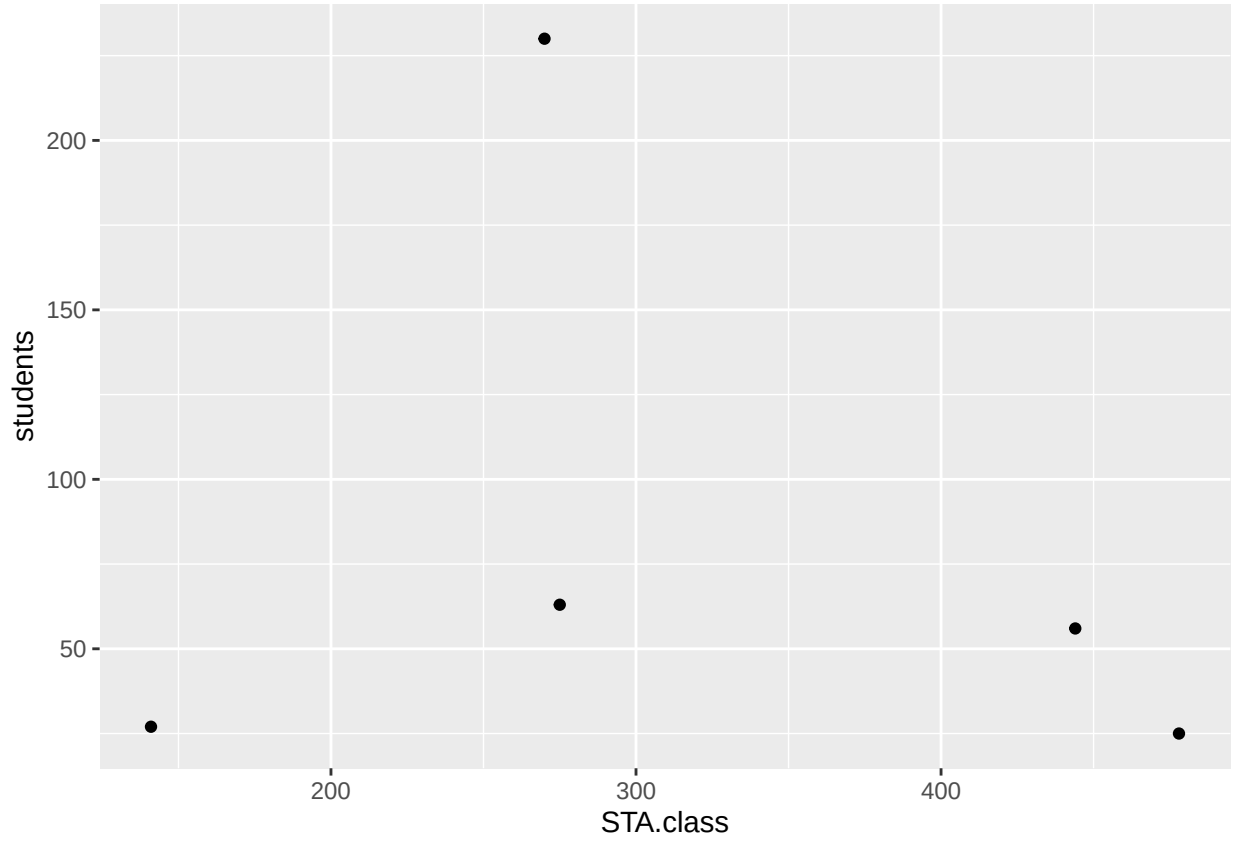
Question 8 (15 points)

The following code block will create a dataset that shows the number of students enrolled in certain NAU STA classes.

```
enrollment <- data.frame(STA.class=c(141,270,275,444,478),
                         students=c(27, 230, 63, 56, 25))
```

A student was asked to produce a plot showing the number of students in different Math classes. The following code block shows their attempt:

```
ggplot(data=enrollment,
       mapping=aes(x=STA.class, y=students))+
  geom_point()
```



(a) Another student suggests that a scatterplot was not a good choice. Is this student correct and why/why not?

The student suggesting that a scatterplot was not the best choice is correct. For univariate data like this, it is best to employ a univariate bar chart.

(b) Create a plot that you think best display the number of students enrolled in each of these 5 classes. Explain why yours is better.

```
class_plot <- ggplot( data=enrollment ) +
  geom_bar( stat='identity', aes( x=as.factor( STA.class ), y=students ) ) +
  labs( title="Students Enrolled in STA Classes",
```

```
        x="STA Class Number", y="Students Enrolled")

class_plot
```

This chart is better than the scatterplot because it displays the information in a way which is easy to read and understand. The X-axis is made into a categorical rather than a numerical variable, which eliminates much of the white space and makes more logical sense. The bars clearly show how many students are in each class, and it is easier to compare enrollment between classes.