Homework 2 - Descriptive Statistics

NAME: Your Name

NETID: Your NetID

DUE DATE: September 7, 2016 by 1:00pm

Homework 2 Instructions

1. In this homework we will explore the `StudentSurvey` data described below. For each problem:
    a) Answer all questions
    b) Insert code chunks directly under any problems that require you to use R and type in code as needed. In particular, make sure code chunks are included for any requested plots.
    c) Answer any questions related to the problem in the .Rmd document directly under the question
    d) Note: Occasionally when you insert a code chunk it may not go where you intend it to. If this happens, you can cut and paste it into the correct spot. Make sure the code chunk is aligned to the left margin of this document. Often it may be easier to just store a code chunk on your clip board and paste it in when you need one.
    e) You may need to knit your document occasionally to answer questions related to R output.
2. Submit two documents: a R Markdown file and a pdf. These files should be named "*LastF*-HW2.Rmd" and "*LastF*-HW2.pdf".

SurveyData

An in-class survey was given to 362 introductory statistics students over several years. The StudentSurvey data contains 17 variables recorded for each student. They are as follows:

Year: Year in school: FirstYear, Sophomore, Junior, or Senior

Gender: M or F

Smoke: "No" or "Yes"

Award: Preferred award: Academy, Nobel or Olympic

HigherSAT: Which SAT score was higher: Math or Verbal

Exercise: Hours of exercise per week

TV: Hours of TV viewing per week

Height: Height in inches

Weight: Weight in pounds

Siblings: Number of Siblings

BirthOrder: Birth order: 1=oldest, 2=second oldest, etc.

VerbalSAT: Verbal SAT score

MathSAT: Math SAT score

SAT: Combined Verbal + Math SAT

GPA: College GPA

Pulse: Pulse Rate (beats per minute)

Piercings: Number of body piercings

The **StudentSurvey** data can be downloaded from the folder for homework 2 on Blackboard. Put this data set in your folder for homework 2.

To read these data into your R Console:

i. In the menu for RStudio above, select *Tools->Import Dataset->From Text File....*

ii. Navigate to the correct file in your folder for homework 2.

iii. Click on the StudentSurvey file and choose *Open*.

iv. A window will pop up where you can preview the data set and possibly choose different options for downloading this data. For this data set, the defaults are appropriate. Click once on *Import* to read the data into the R Console.

You now should see this data listed in the "Environment" window in the upper right corner of RStudio.

Problem 1

To read the data into this R Markdown document, we will use the `read.csv()` function in R. Fortunately, this function was just used in the R Console.

a) Create a code chunk here. Copy the code in the R console below that starts with `StudentSurvey<-read.csv` and paste it into this code chunk.

```
StudentSurvey <- read.csv("~/BTRY6010HW2/StudentSurvey.txt")
colnames(StudentSurvey)
```

```
##  [1] "Year"       "Gender"     "Smoke"      "Award"      "HigherSAT"
##  [6] "Exercise"   "TV"         "Height"     "Weight"     "Siblings"
## [11] "BirthOrder" "VerbalSAT"  "MathSAT"    "SAT"        "GPA"
## [16] "Pulse"      "Piercings"
```

```
dim(StudentSurvey)
```

```
## [1] 362  17
```

b) In the code chunk above also include code to do the following:

i. list the variable names
ii. get the dimension of the data

c) Suppose the population of interest is all college students. What would you call the sampling method used for this study? How does this affect the interpretation of any analysis performed on these data?

Convenience sampling since the procedure is subjective that college students not enrolling in introductory statistics have zero probability to be included in the sample. Since the sample is non-representative for all college students, we can only derive limited inference and can not project analysis results to the population.

d) List the variable types of the following (be as specific as possible!):

1. `TV: quantitative`

2. `Award: nominal categorical`

3. `Birth Order: ordinal categorical`

4. `Pulse: discrete quantitative`

5. `GPA: discrete quantitative`

6. `Piercings: discrete quantitative`

Problem 2

One of the questions asked on the `StudentSurvey` was, "Which award would you prefer to win: an Academy Award, a Nobel Prize, or an Olympic gold medal?"

a) Which award was most popular amongst students? Create a table of counts for `Award` with R's `table()` function.

```
table(StudentSurvey$Award)
```

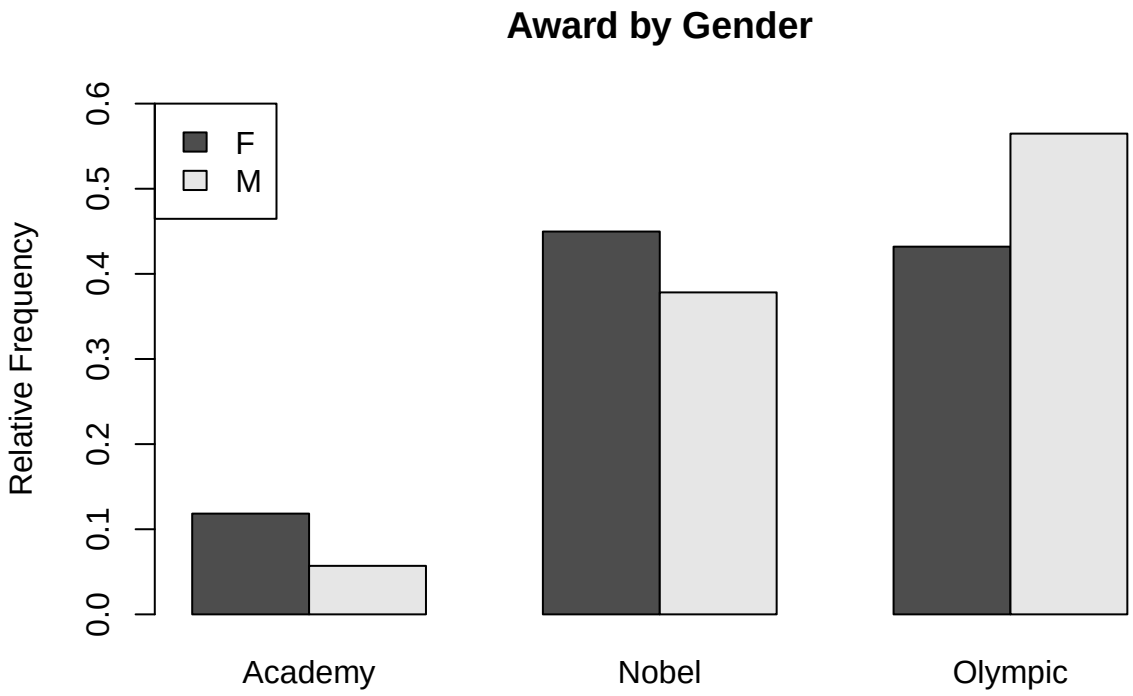
```
##
## Academy   Nobel Olympic
##      31     149     182
```

Olympic is the most popular one.

b) Was the proportion of students preferring each award different for women and men? Explain. Complete the following steps to answer this question.

i. Create a relative frequency bar chart for `Award` by `Gender`. The proportions of the preferred awards for each gender should sum to 1. You may get the necessary counts using the `table()` function in R, but it may take more than 1 step. Do all calculations necessary in the code chunk.
ii. Title this chart "Award by Gender"
iii. The bars for Males and Females should be side by side
iv. Include a legend for gender using "F" and "M" as the labels
v. Make the bars vertical
vi. Set `ylim=c(0, 0.6)`
vii. Include the option, `args.legend = list(x="topleft")`
viii. Don't forget to answer the question!

```r
awardbygen <- table(StudentSurvey$Gender, StudentSurvey$Award)
awardbygen_prop <- awardbygen/rowSums(awardbygen)
barplot(awardbygen_prop, names.arg = c('Academy', 'Nobel', 'Olympic'),
        legend.text = c('F', 'M'), beside = TRUE, horiz = FALSE,
        ylim = c(0, 0.6), main = 'Award by Gender', ylab = "Relative Frequency",
        args.legend = list(x="topleft"))
```



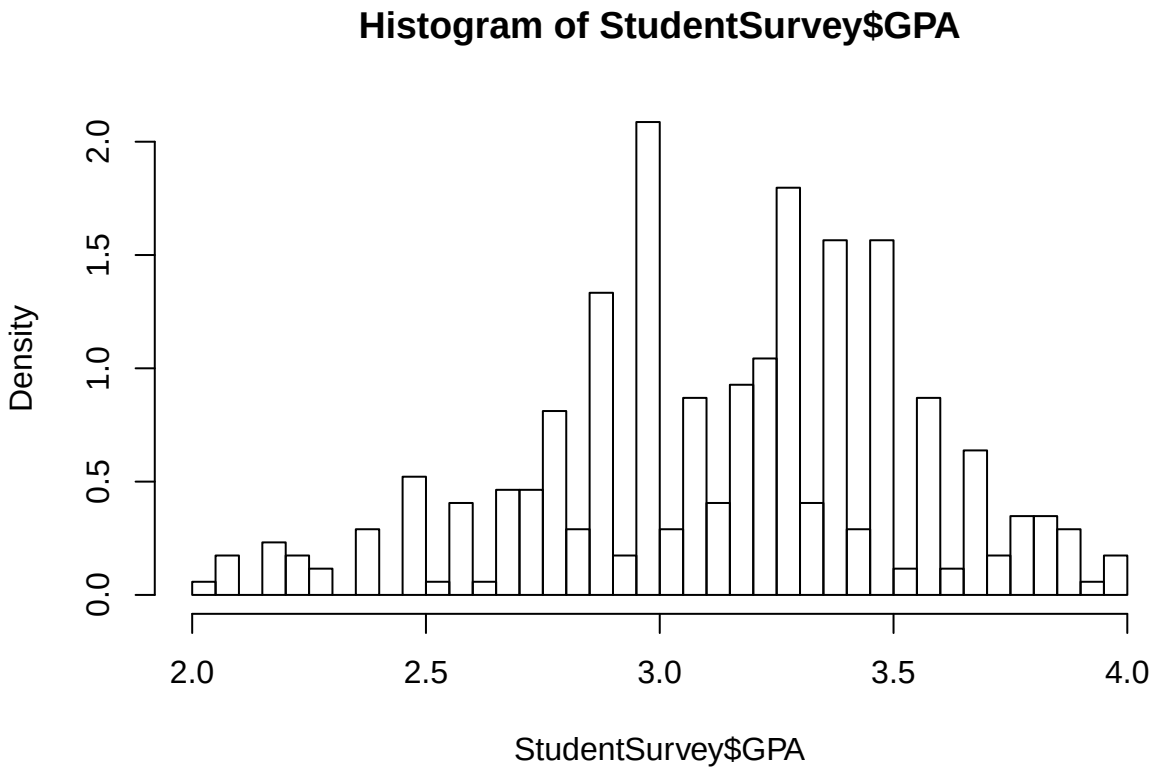
The overall distribution of preference towards the awards is similar for both males and females: Olympic is the most popular award, followed by Nobel. However, the discrepancy between genders within each award is obvious: females have a slightly larger proportion preferring Academy or Nobel and a slightly smaller proportion preferring Olympic than males.

Problem 3

Another variable recorded for the `StudentSurvey` is college GPA. Here we will look at the relationship between college GPA and `Award`.

a) First, create a probability histogram of `GPA`, set `breaks=50`.

```r
hist(StudentSurvey$GPA, breaks = 50, freq = FALSE)
```



i. How would you describe the distribution of GPA?

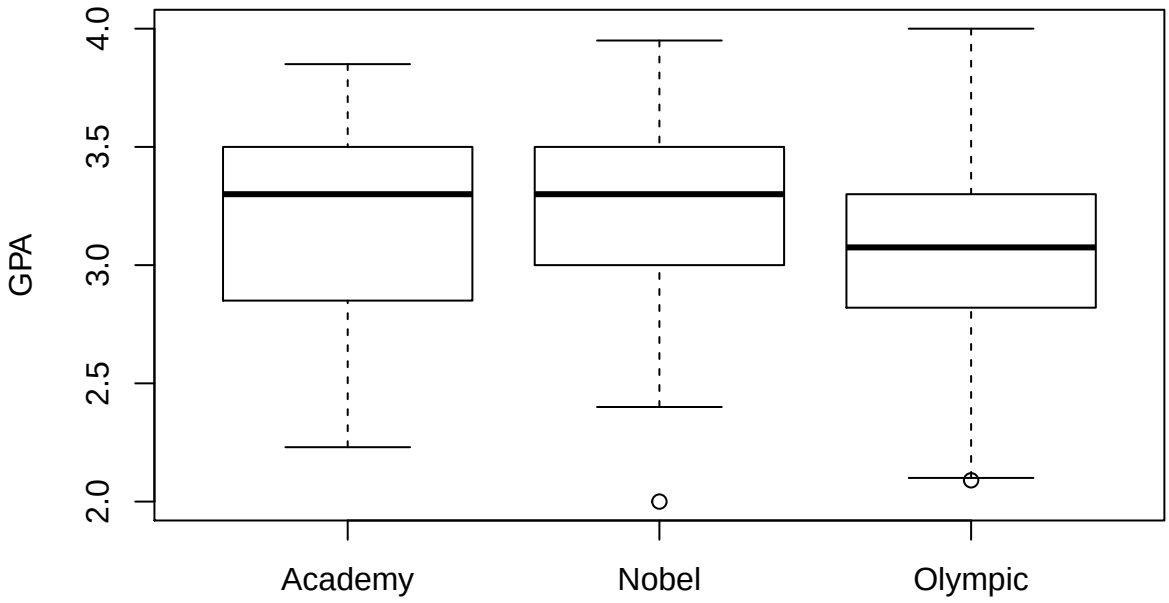
The distribution of GPA is unimodal and slightly skewed to the left.

ii. Based on the histogram alone, estimate the range of the most common GPA values.

2.8-3.5

b) Create boxplots for `GPA` separated by `Award`.

```
boxplot(StudentSurvey$GPA ~ StudentSurvey$Award, ylab = "GPA")
```



i. Do there appear to be any differences between the mean GPAs of the three groups? Support your ansu

Our initial goal was to compare the three medians, since boxplots don't directly show the means. Based on the boxplots, the GPA median of the Olympic group is slightly smaller than that of the other two groups. However, since Academy GPAs are skewed, its mean may not be the same as its median. Therefore, it is hard to compare the mean GPAs of the three groups.

ii. One group has a student with a very optimisitic outlook on life if he/she plans to get his/her p

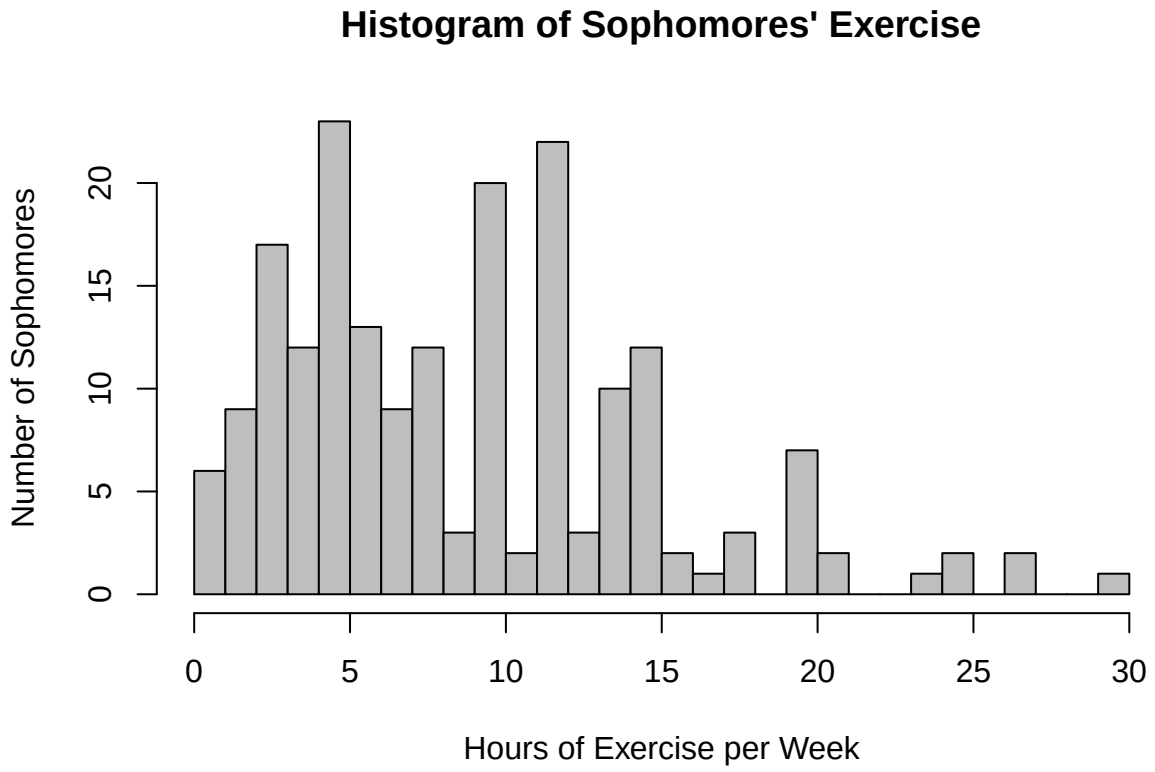
One student in the Nobel prize group has the lowest GPA across all students. This GPA is an "outlier" in the Nobel group. Of course Einstein apparently had poor grades in school, so this student might just be following this less traditional path.

Problem 4

Yet another variable collected by the student survey was `Exercise`. This variable recorded the number of hours each student exercised per week. Here we will look at the relationship between `Exercise` and `Award`.

a) Use the function `hist` to create a histogram of the number of hours of exercise per week for Sophomores. Be sure to customize the plot so that it is clear what it is conveying (e.g., label the axes to convey what is being shown) and perhaps adjust `breaks` manually (recall that `?hist` will give you information about the arguments). How would you describe the distribution of `Exercise` for the Sophomore students?

```
sophomores <- subset(StudentSurvey, Year == "Sophomore")
hist(sophomores$Exercise, xlab = "Hours of Exercise per Week", ylab = "Number of Sophomores",
     main = "Histogram of Sophomores' Exercise", col = "gray", breaks = 35)
```



The majority of the Sophomore students exercise less than 15 hours per week. The distribution is skewed to the right with a few Sophomore students exercise more than 25 hours per week.

b) Use the `summary` function to get summary statistics for `Exercise`. What was the range of `Exercise`? If a student exercised more hours per week than half of the sample, what is the least amount of exercise he/she was getting per week?

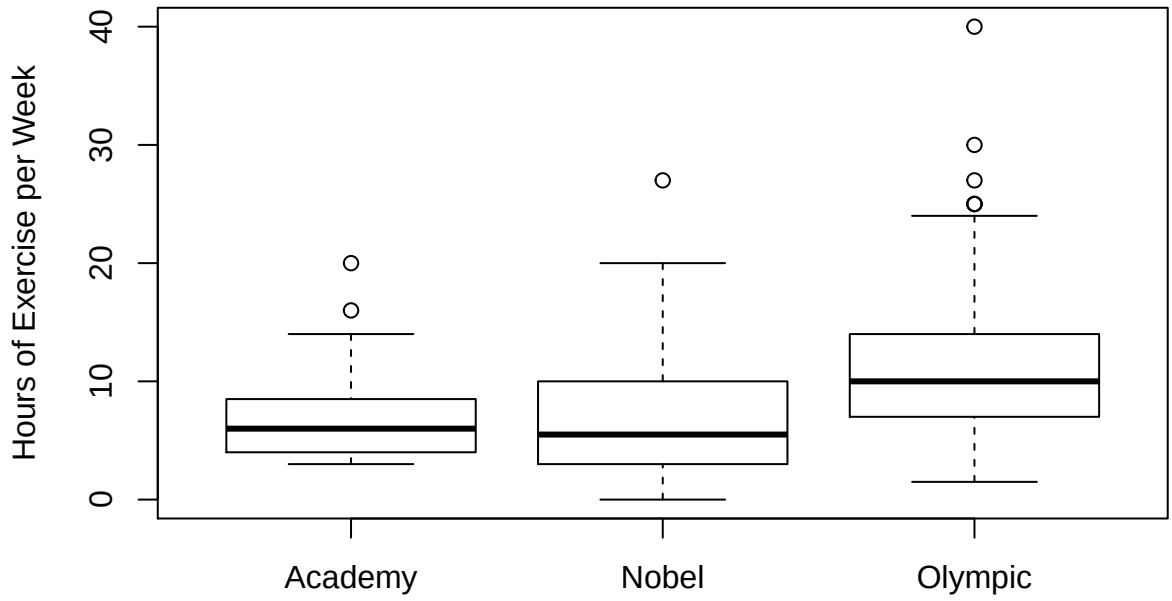
```
summary(StudentSurvey$Exercise)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   5.000   8.000   9.054  12.000  40.000       1
```

The range of `Exercise` is 40. The student exercised at least 8 hours per week (median value), if he/she exercised more hours than half of the sample.

c) Create boxplots of `Exercise` by `Award`. What can be said about the distribution of `Exercise` for the students who preferred to win an Olympic gold medal in comparison to the distribution of `Exercise` for those who chose an Academy Award or a Nobel Prize?

```
boxplot(StudentSurvey$Exercise ~ StudentSurvey$Award, ylab = "Hours of Exercise per Week")
```



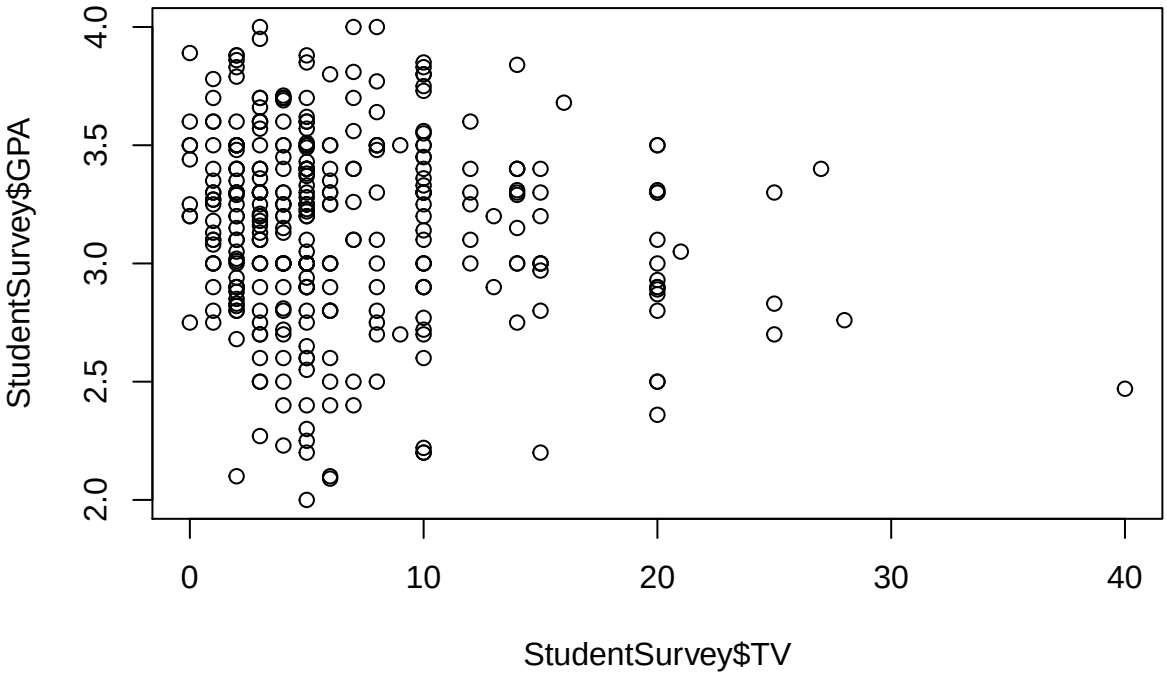
The distribution of `Exercise` for students who preferred to win an Olympic gold medal has larger median value (and mean value) than that for students who chose an Academy Award or a Nobel Prize. There are more outliers with much larger values in `Exercise` for students who chose an Olympic gold medal. This ourliers demonstrate there are some students truly loving sports and they devote more effort in training.

Problem 5

Is there a relationship between the number of hours of TV you watch and your GPA?

a) Create a scatterplot of `GPA` by `TV` using the code below.

```
plot(StudentSurvey$TV, StudentSurvey$GPA)
```

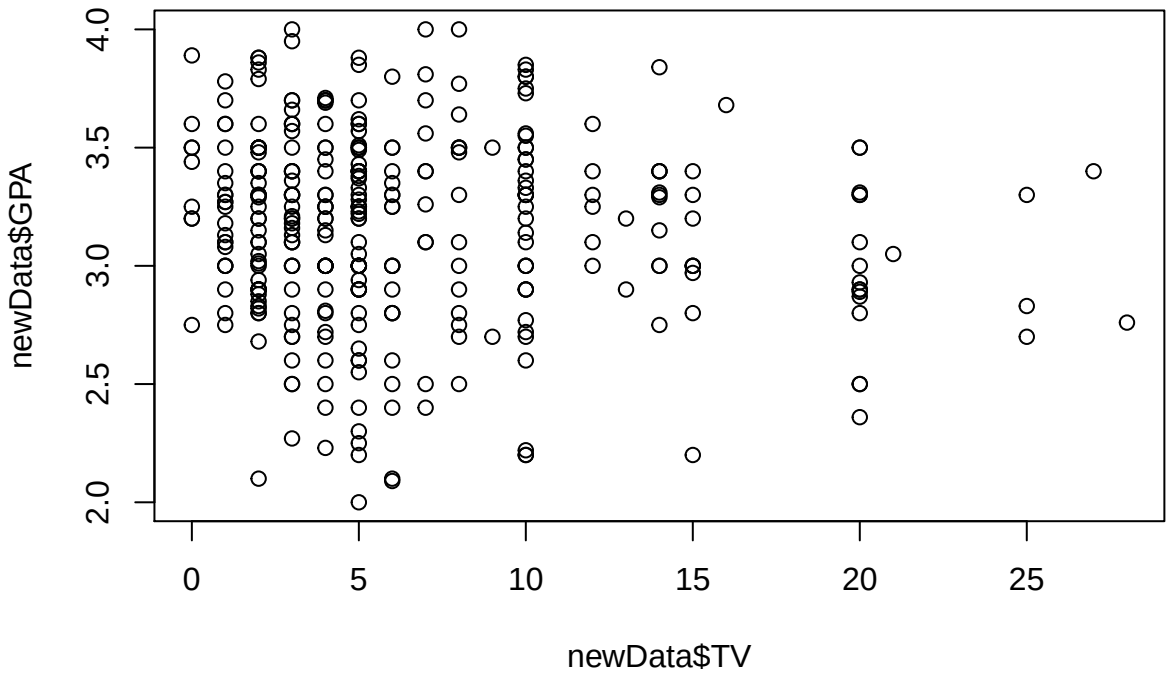


b) Does there appear to be a relationship between the number of hours watching TV and college GPA?

With the student who watches 40 hours per week, it looks like there might be a slight downward trend or perhaps no trend.

c) Let's take a look at this relationship again after excluding students who watches 40 hours of TV a week. Do this in two lines. First, create a new data frame (using the `subset` function). Second, use `plot`.

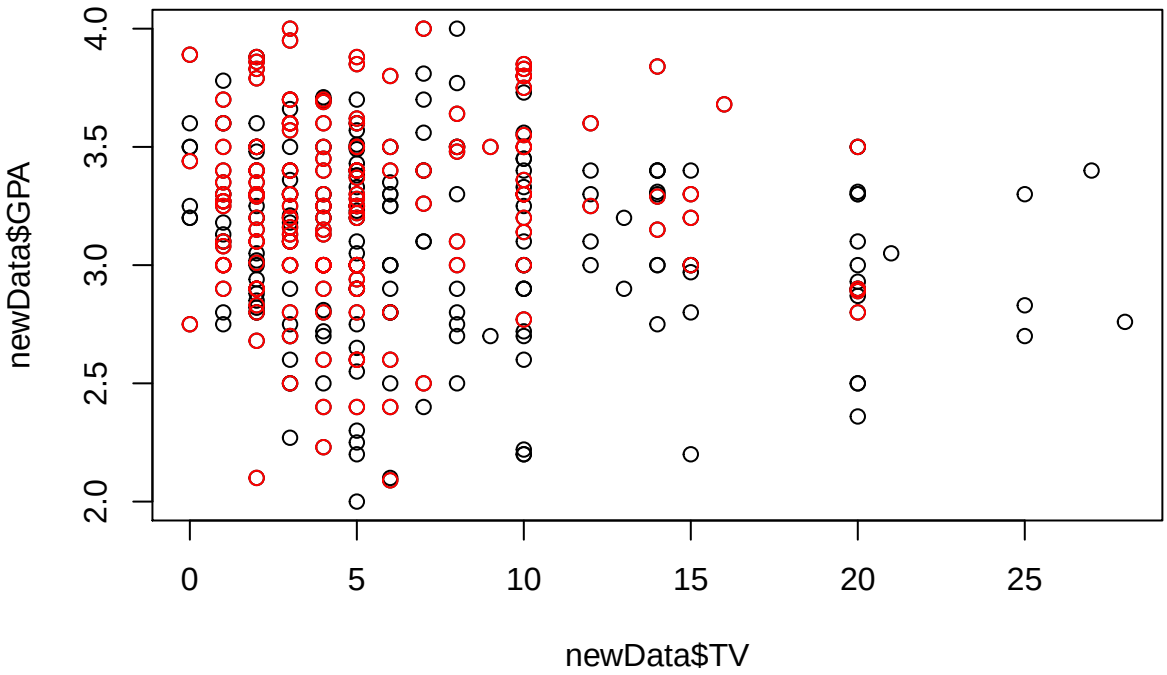
```
# First line: create new data frame that keeps only students with TV under 40
newData <- subset(StudentSurvey, TV != 40)
# Second line: call plot function with this newly created data frame
plot(newData$TV, newData$GPA)
```



Without that student it is difficult to see by eye any sort of relationship.

d) We can look at the difference in this relationship between males and females by coloring the female observation red. Is this relationship any different for females compared to males? You will do this in three lines of code. In the first line, simply repeat the previous call to `plot` that you wrote in part c; in the second line, create a data frame called `females` that only has the rows corresponding to women; the third line is written for you.

```
# First line: Call to plot function from second line of part c
plot(newData$TV, newData$GPA)
# Second line: create data frame called females
females <- subset(newData, Gender == "F")
points(females$TV, females$GPA, col='red') # this is third line
```



The relationship between GPA and TV doesn't seem to be different for females compared to males.

e) What was the effect of the `points()` function above?

We used `points()` to overlay female observations with red circles.

Problem 6

For this problem, we will examine the variable, `Piercings`.

a) In R, output from using the `class()` function on a variable tells you what class R has given that variable.

```
class(StudentSurvey$Year)
```

```
## [1] "factor"
```

```
class(StudentSurvey$Piercings)
```

```
## [1] "integer"
```

```
class(StudentSurvey$GPA)
```

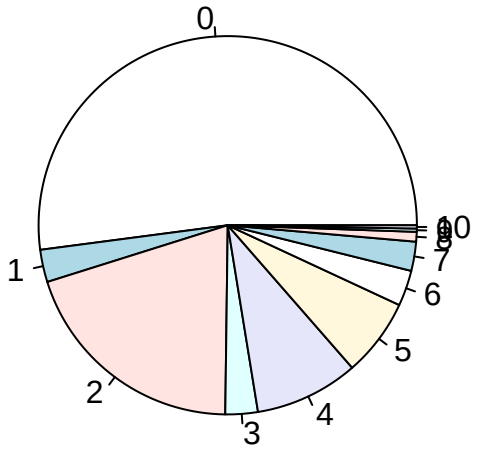
```
## [1] "numeric"
```

b) `Piercings` is of class "integer". For this problem, we will consider `Piercings` as a "factor" (categorical variable). Run the following to change the class of `Piercings`.

```
Piercings <- as.factor(StudentSurvey$Piercings)
```

c) Create a pie chart of `Piercings` using the following code. Is this a good graphical summary of these data? Explain.

```
pie(table(Piercings))
```

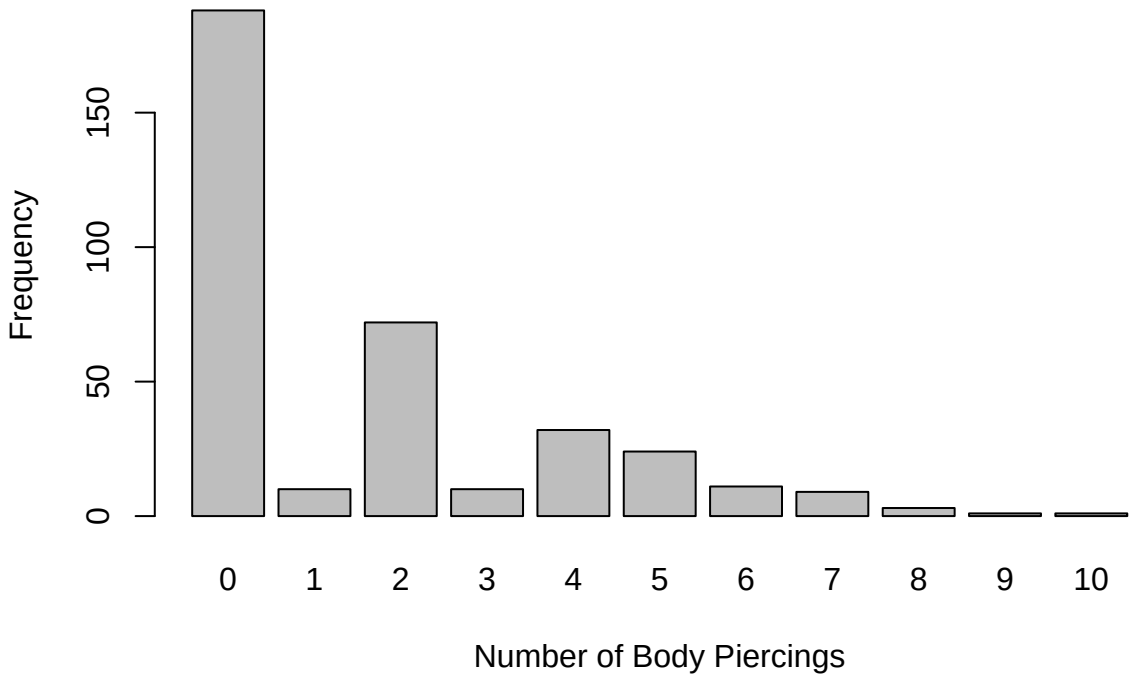


No. Many of the large-valued levels overlap and it's hard to tell the difference among them. Also, pie charts have been shown to be an unreliable way of displaying data. It is difficult to distinguish the differences between similar sized wedges and also rotating a pie chart affects a human's perception of the size of wedges.

d) What might be a better way to graphically describe the distribution of `Piercings`?

A bar chart would be a better way visualize the distribution of `Piercings`.

```
barplot(table(StudentSurvey$Piercings), xlab="Number of Body Piercings", ylab = "Frequency")
```



e) Suppose we want to reduce the number of levels for `Piercings` from 11 to 8. What might be the best way to re-group these data so that the pie chart is a better representation of the distribution of `Piercings`?

We can combine the three levels of `Piercings` with values 8, 9 and 10 into the level `above 7`. These three levels overlap in the pie chart and all have low frequency based on the bar chart in (d).