Homework 2 - Descriptive Statistics

NAME: Andres Castano Zuluaga

NETID: ac986

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.

```r
StudentSurvey<-read.csv("~/Dropbox/CORNELL/Fall 2016/BTRY6010/Homework/HW2/StudentSurvey.txt")
names(StudentSurvey)
```

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

```r
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?

Despite there is not enough information to determine what is the objective (or objectives) of study, if the population of interest is all students, then it is clear that sampling method used is judgment sampling. Maybe the researchers or statisticians involved in the sample's choosing method considered (in my point of view erroneously) that the students of introductory statistics are representative of all university. Given that this sampling method is not probabilistic, it is not possible to compute the sampling error and all the interpretation of any analysis can not be extrapolated to the population.

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

```
1. TV -- discrete numerical
```

```
2. Award -- ordinal categorical
```

```
3. Birth Order -- ordinal categorical
```

```
4. Pulse -- discrete numerical.
```

Beats per minute can take only a finite number of distinct values and have a theorical maximun of 220 (see http://www.sciencedirect.com/science/article/pii/S0735109700010548)

```
5. GPA -- continuos numerical
```

```
6. Piercings -- discrete numerical
```

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.

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

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

The most popular award was winning an Olympic Gold Medal.

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!

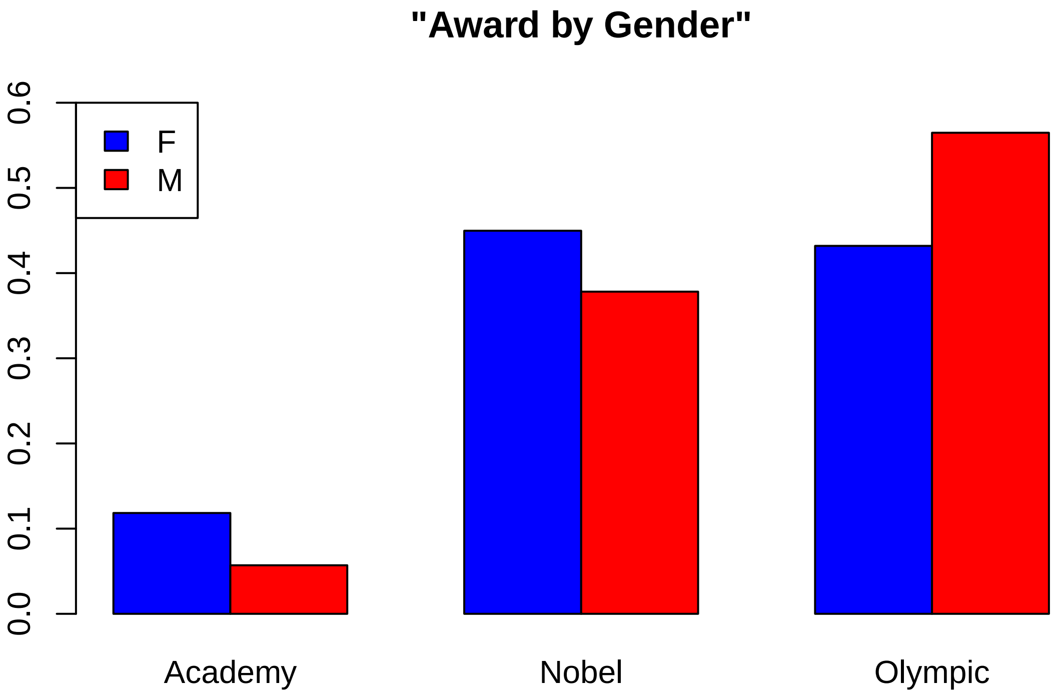
```
awardbygen <- table(StudentSurvey$Gender , StudentSurvey$Award)
awardbygen
```

```
##
##     Academy Nobel Olympic
##   F      20    76      73
##   M      11    73     109
```

```
prop_awardbygen <- prop.table(awardbygen,1)
prop_awardbygen
```

```
##
##        Academy      Nobel    Olympic
##   F 0.11834320 0.44970414 0.43195266
##   M 0.05699482 0.37823834 0.56476684
```

```
barplot(prop_awardbygen, ylim = c(0, 0.6), beside = TRUE, horiz = FALSE
        ,legend.text=(rownames(prop_awardbygen)), args.legend = list(x="topleft")
        , main = '"Award by Gender"', col = c("blue", "red"))
```



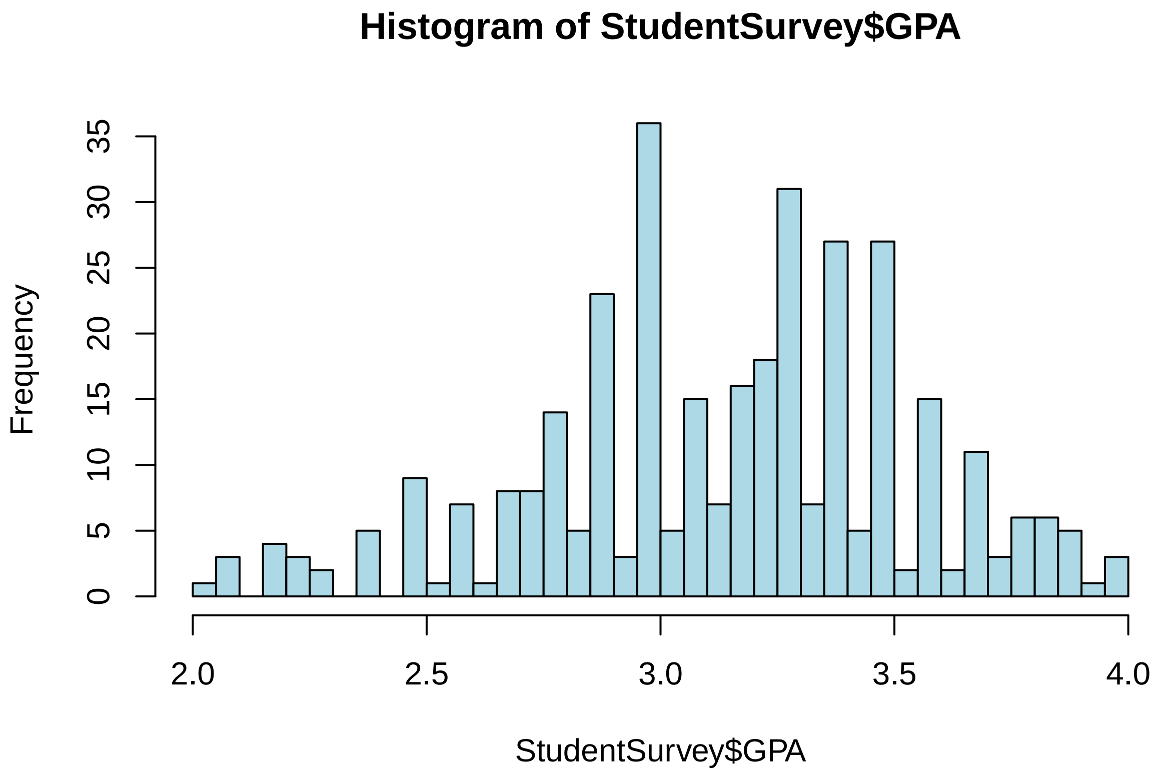
Graphically we can appreciate a difference regarding the preferred award by gender. The men have a stronger preference to win an Olympic Gold Medal than women (56% vs 43%); the women showed a stronger preference regarding the Nobel Prize Award than men (45% vs 38%); and finally, the women also prefer more to win an Academy Award than men (12% vs 5%). Despite this illustrative information is important to depict the award preferences by gender, it is not enough to reach a definitive conclusion about gender award preferences, we need to use some appropriate statistical test to be sure.

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`.

```
hist(StudentSurvey$GPA, breaks = 50, col='lightblue')
```



```
mean(StudentSurvey$GPA, na.rm=TRUE)
```

```
## [1] 3.157942
```

```
median(StudentSurvey$GPA, na.rm=TRUE)
```

```
## [1] 3.2
```

```
mode<-3.0
```

```
i. How would you describe the distribution of GPA?
```

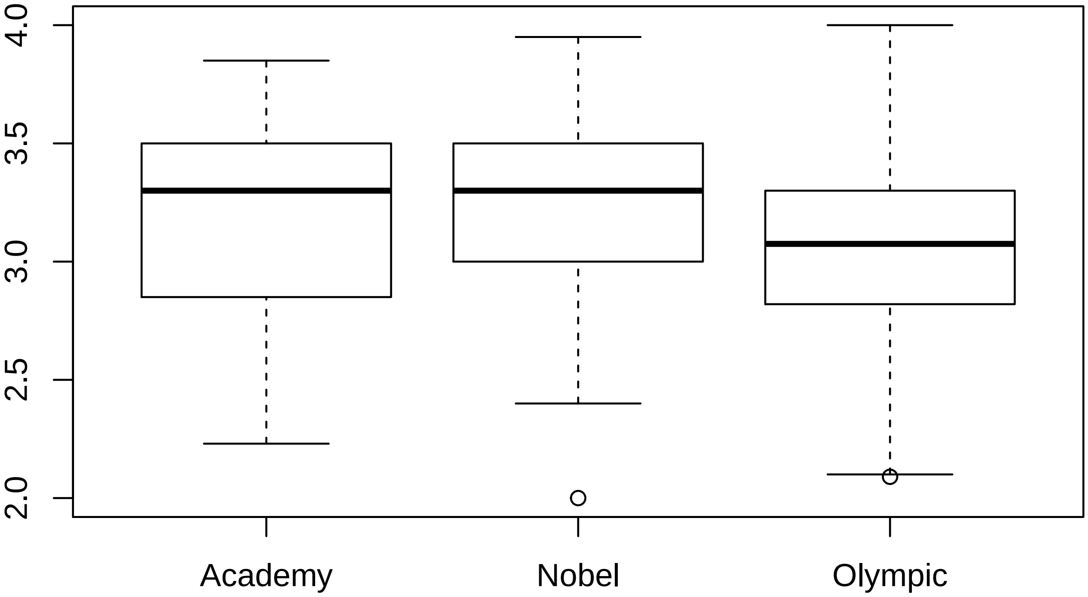
The distribution of GPA is slightly left-skewed, which means that an important portion of the students have high GPAs (for example, over 3.3). The distribution also seems to be unimodal. In this case mean=3.15, median=3.2 and mode=3.0.

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

Given that we have divided the histogram into 50 bins, the most common values according to the histogram should be between 2.951 and 3.0.

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

```r
boxplot(StudentSurvey$GPA ~ StudentSurvey$Award)
```



i. Do there appear to be any differences between the mean GPAs of the three groups? Support your answer using information from the plots.

A box plot does not give information about the mean unless our variable of study follows a normal distribution, in that case, the median=mean. In our case, we do not know if there are differences between the mean GPAs of the three groups and it will be not recommendable get conclusions about it based on a box plot.

ii. One group has a student with a very optimisitic outlook on life if he/she plans to get his/her preferred award. Which group does he/she belong to? What is the statistical term for this observation?

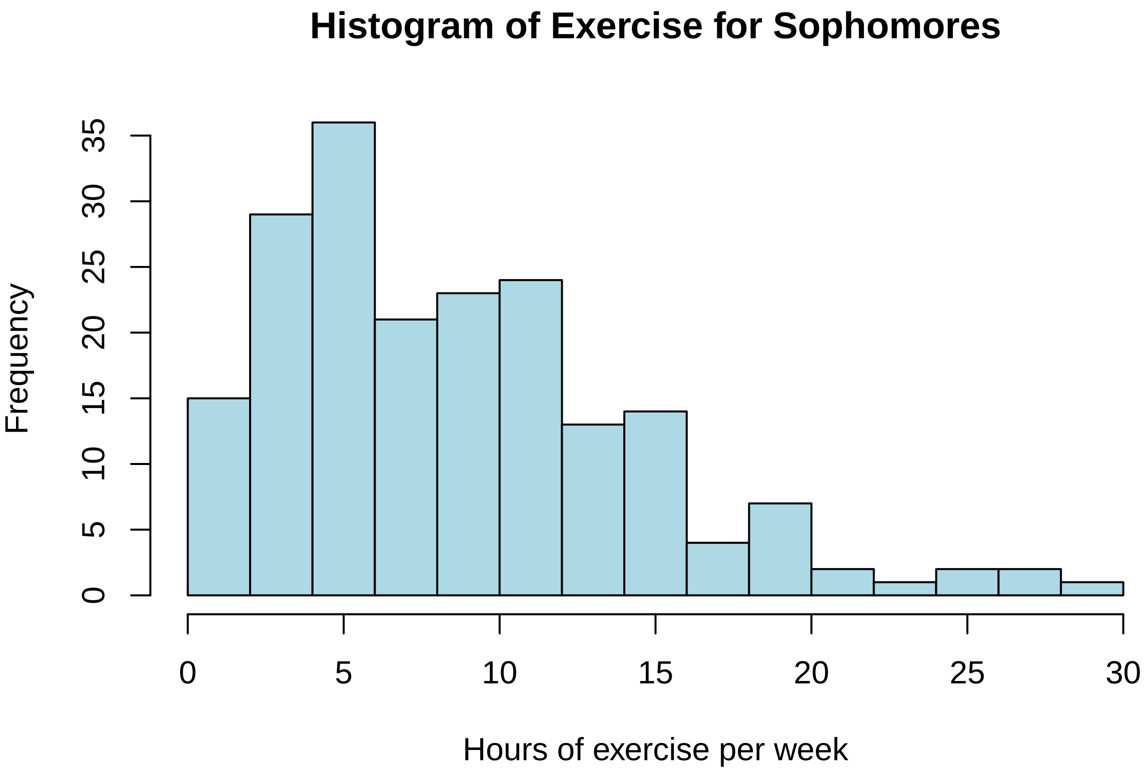
He belongs to the group that prefers a Nobel Award and this observation is typically called “outlier”.

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?

```r
sophomore <- subset(StudentSurvey, Year== "Sophomore")
hist(sophomore$Exercise, breaks = 20, col='lightblue', main = "Histogram of Exercise for Sophomores", xl
```



The distribution of the hours of exercise per week among sophomores is right skewed, which means that an important part of sophomore students exercise no many hours (for example, between 0 and 12 hours per week) and a small percentage exercise many hours (for example more than 15 hours or more).

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?

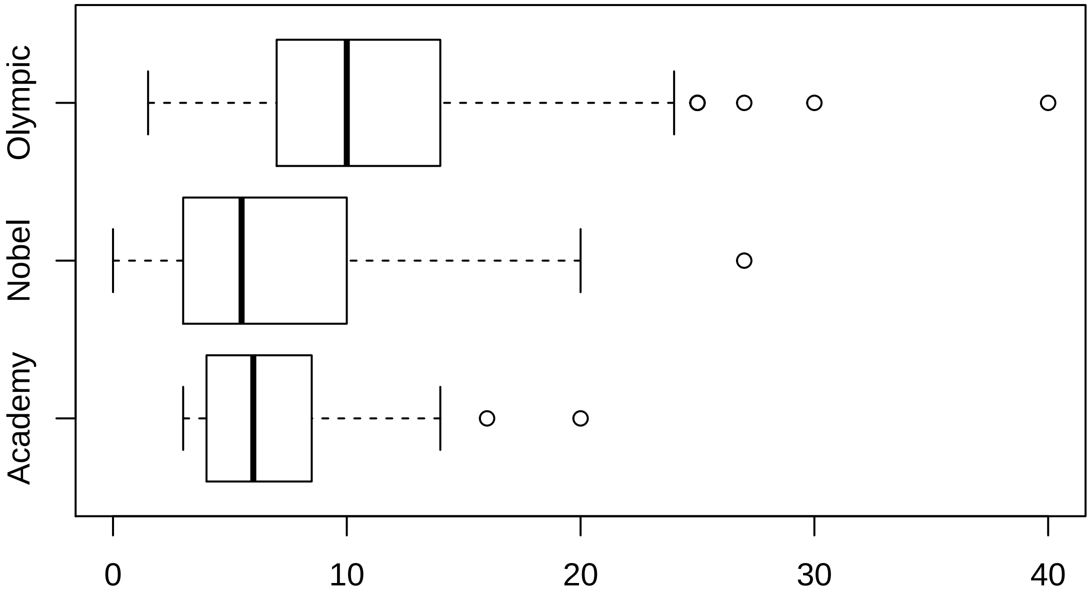
```r
summary(StudentSurvey$Exercise)
```

```
##    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 was 40. On the other hand, If a student exercised more hours per week than half of the sample, the least amount of exercise that he/she does should be greater than or equal to 8.0 hours of exercise per week. In our case, when we order the 362 values of the variable exercise from the lowest to greatest and found the medium point, we get (8+8)/(2)=8 (because we have an even number of observations). Every value to the right of this value must be a value that is greater than half of the sample. In our case, the next 11 values are also equal to 8, which means that even getting 8 hours of exercise per week you could be over the 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, horizontal=TRUE)
```



```
summary(StudentSurvey$Exercise, StudentSurvey$Award)
```

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

The distribution of exercise for the students who preferred to win a gold medal is more skewed to the right compared with those who chose to win an Academy Award, but less skewed to the right compared with those who chose a Nobel award. This means, that for the firs two groups there is an important percentage of people that exercise few hours, and few percentage that exercises many hours per week (for example, more than 20 hours).

On the other hand, base on the range, we can say that the distribution of exercised hours shows more variability in those who preferred an Olympic medal compared with the other groups.

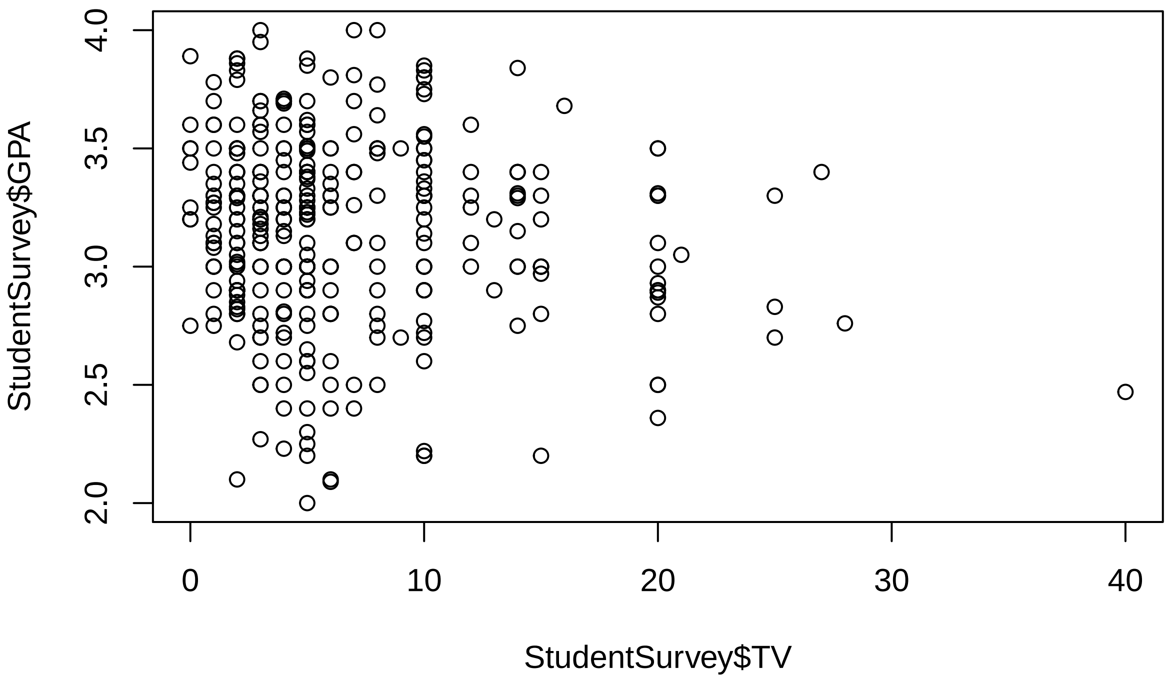
Finally, as we might have supposed, the median of exercised hours per week among those who preferred an Olympic Gold Medal is greater compared to the other groups (maybe they are passionate about sports and enjoy training). At the same time, it is not difficult to detect that the mean of hours per week must be also higher in the Olympic Gold Medal group compared to the other groups. This can be assumed for two reasons: first, the number of outliers (in the right part) and second, that 75% of the students who preferred an Olympic Gold Medal have 7 or more hours of exercise per week (for the other groups, this value is less than 5).

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.

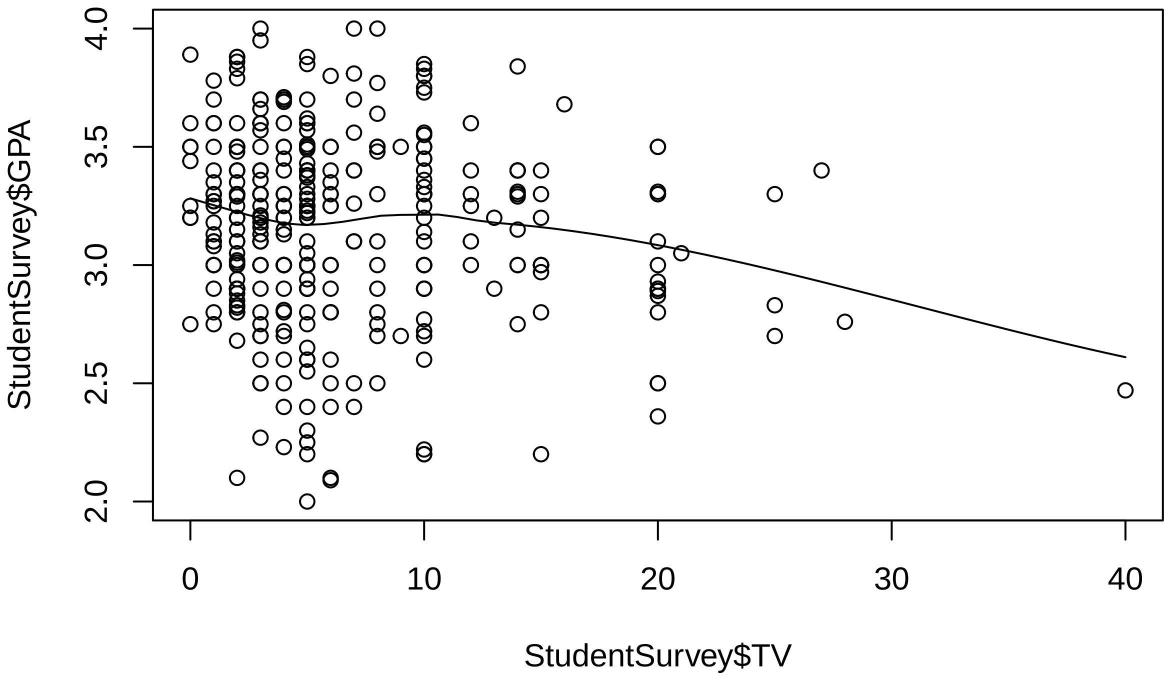
```
plot(StudentSurvey$TV, StudentSurvey$GPA)
```



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

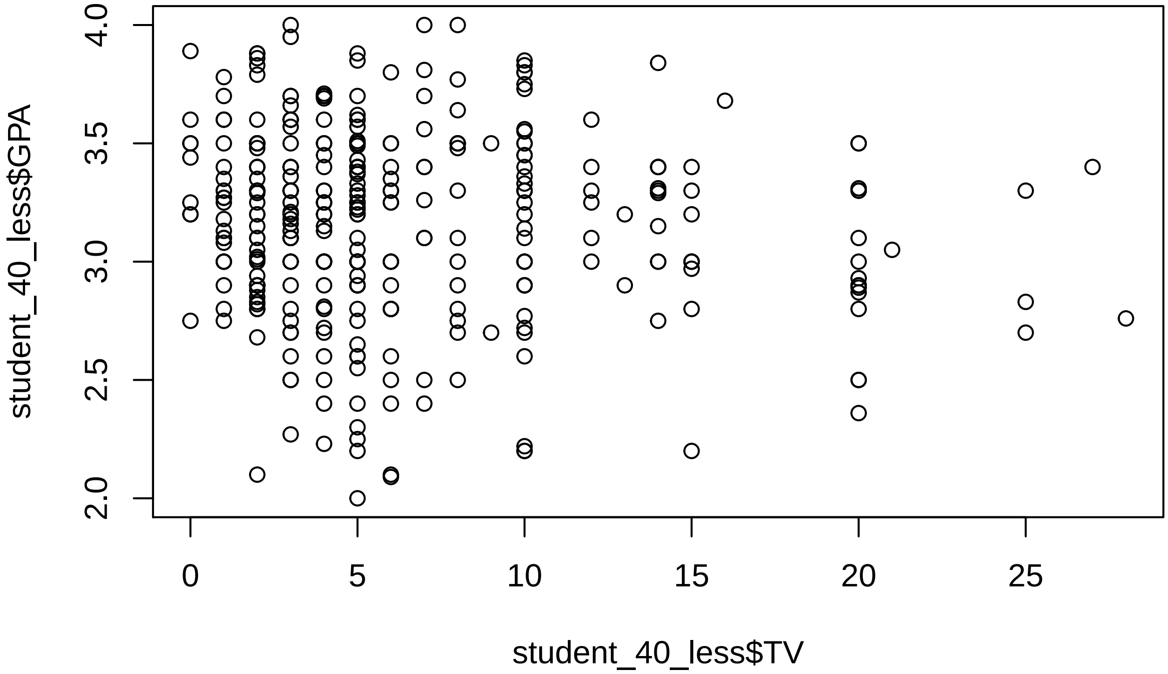
It seems that there is a negative relationship between watch tv and the GPA. Despite that this relationship is weak, it might suggest that the more time you spend watching television, less time you spend studying and maybe this could affect your GPA. I have done the same graph adding a smooth curve to verify this initial intuition.

```
scatter.smooth(StudentSurvey$TV, StudentSurvey$GPA)
```



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`.

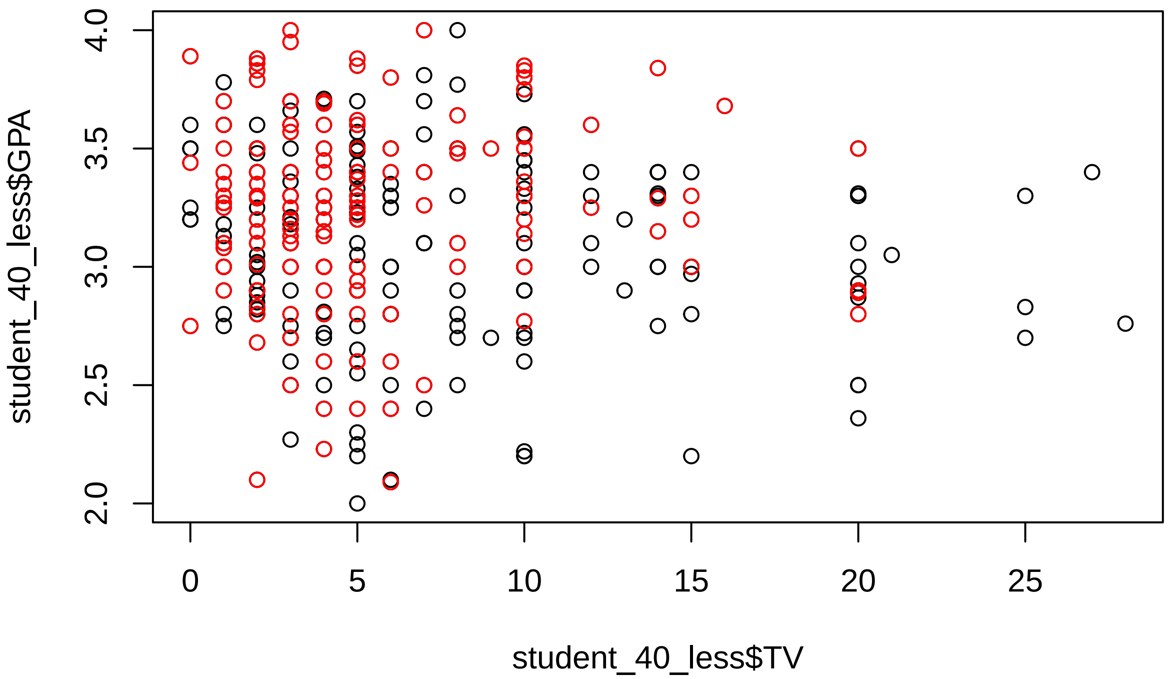
```
student_40_less <- subset(StudentSurvey, TV<40)
plot(student_40_less$TV, student_40_less$GPA)
```



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.

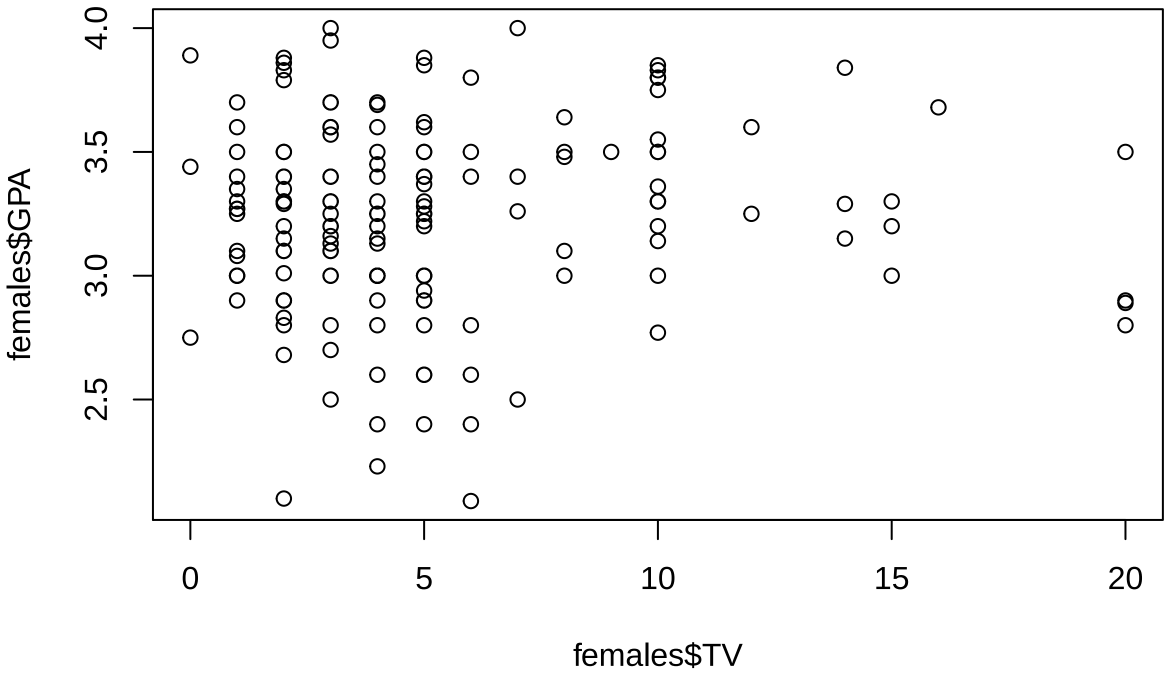
Using the same data of the point c:

```
plot(student_40_less$TV, student_40_less$GPA)
females <- subset(student_40_less, Gender == "F")
points(females$TV, females$GPA, col='red') # this is third line
```

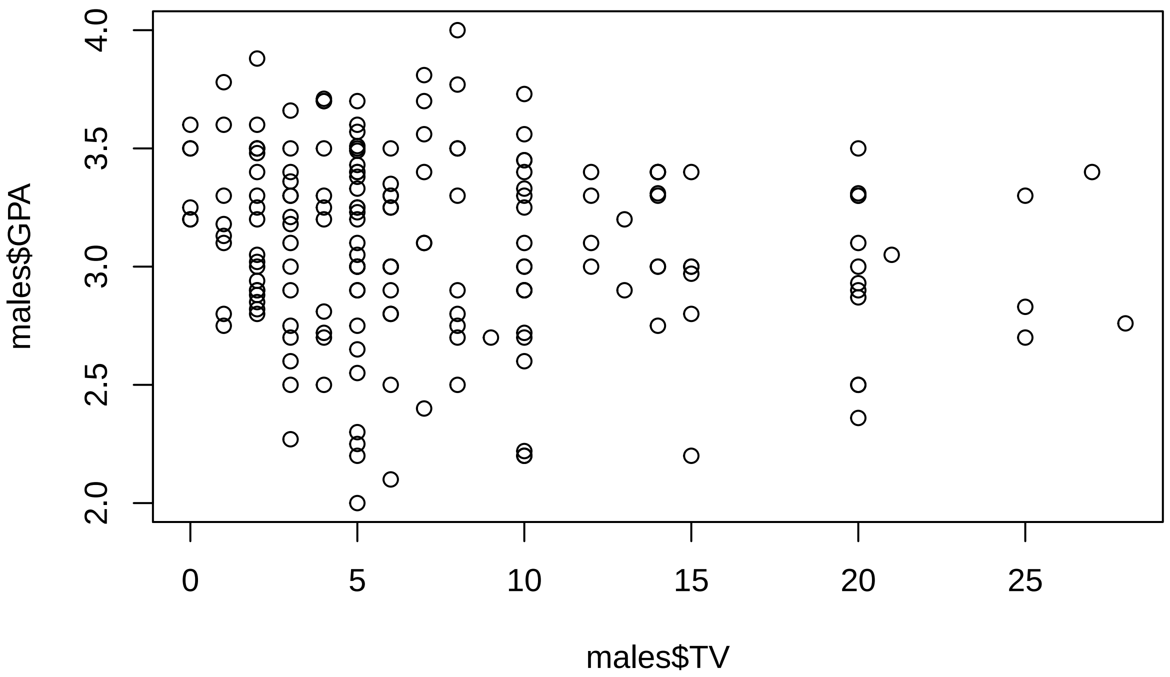


I think it is difficult to realize if the relationship is different between genders using the above framework. I have preferred to separate the graphs for women and men as follows:

```
males <- subset(student_40_less, Gender == "M")
plot(females$TV, females$GPA)
```



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



Now, it seems that there are differences between genders. Whereas in men graph the relationship between TV hours and GPA is negative (albeit weak), in the women's graph is more difficult to find a pattern.

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

The effect was that it marked with red in the scatterplot the points for the relation TV and GPA that are associated with the answers of women. Then we can see if might be different patterns in this relation between genders.

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)
```

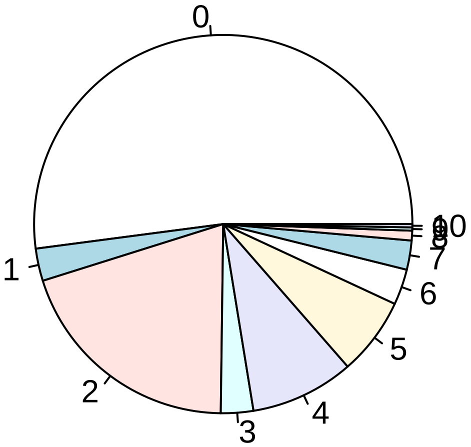
```
## [1] "numeric"
```

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))
```



This graph is not a good summary for at least three reasons:

1) The graphic does not have title and labels, thereby, no one except yourself can understannd what it represents.

2) Despite that you can visualize that almost a half of the sample do not have a piercing, it is more difficult compare, for example, persons that have 4 piercings with those who have 5, the same happens with the people that have 6 when you are trying to compared with those who have 7.

3. As we discuss in class the angle in which we are observing the graph affects our perceived size.

d) What might be a better way to graphically describe the distribution of `Piercings`? A better way to do it is with a bar chart as follows:

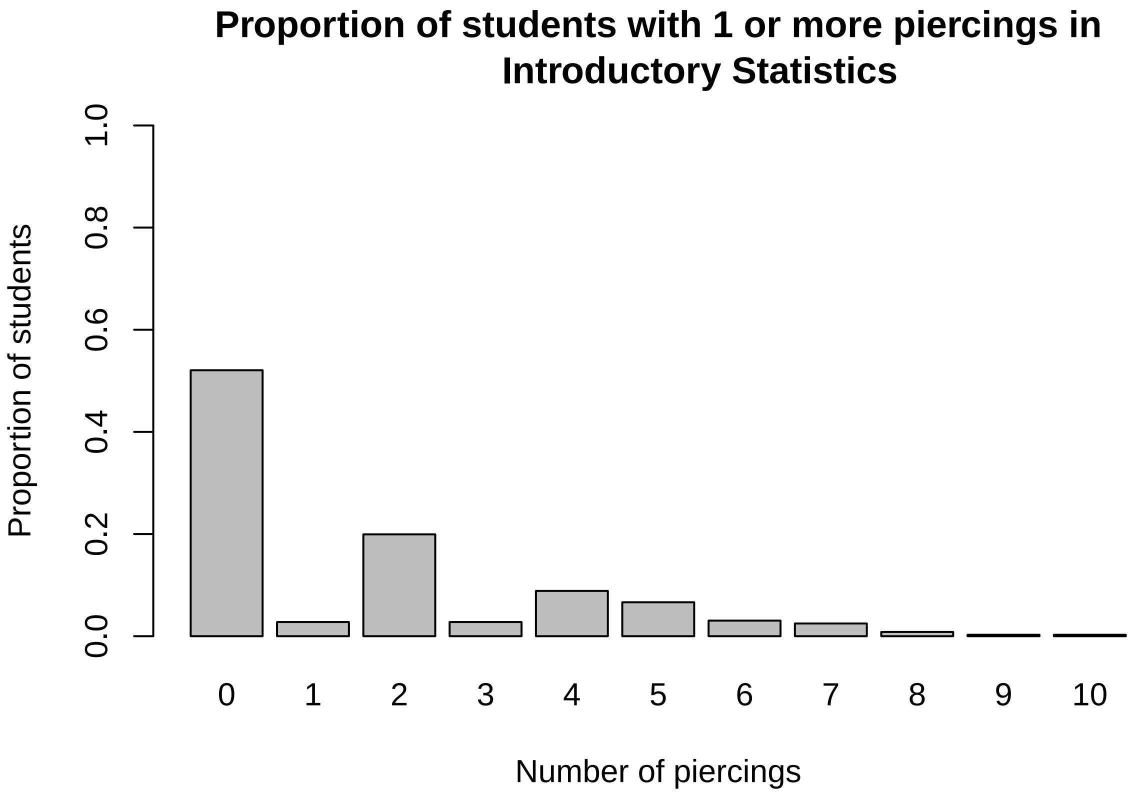
```
resul_piercings <-(table(Piercings))
resul_piercings
```

```
## Piercings
##   0   1   2   3   4   5   6   7   8   9  10 
## 188  10  72  10  32  24  11   9   3   1   1
```

```
prop_piercings <- prop.table(resul_piercings)
prop_piercings
```

```
## Piercings
##           0           1           2           3           4           5 
## 0.520775623 0.027700831 0.199445983 0.027700831 0.088642659 0.066481994 
##           6           7           8           9          10 
## 0.030470914 0.024930748 0.008310249 0.002770083 0.002770083
```

```
barplot(prop_piercings, main="Proportion of students with 1 or more piercings in
        Introductory Statistics",
        ylab="Proportion of students", xlab="Number of piercings", ylim=c(0,1))
```



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`?

One way to do it is by grouping the categories in order to get categories that represent percentages that are visually different, for example, agruping in one category 1 and 3, in another category 6 and 7 and in a final category 8 and 9.