The Duke GPA Data Analysis: In Class

the class

2023-09-08

Task 1.

First, we will read the data from our csv file "gpa.csv" into a data.frame called `gpa.data`:

```
gpa.data<-read.csv("gpa.csv")
#gpa.data
```

If you want to see what your data.frame looks like, you can click on it in the **Global environment** in the upper right pane. The data.frame will get displayed in the upper left pane.

Task 2.

You interested in the types and names of the variables in your data.frame. What do you run?

```
ls.str(gpa.data)
## gender :  chr [1:55] "female" "female" "female" "male" "female" "male" "female" ...
## gpa :  num [1:55] 3.89 3.9 3.75 3.6 4 ...
## out :  num [1:55] 3 1 1 4 3 3 1 3 2 4 ...
## sleepnight :  num [1:55] 6 6 7 6 7 7 6 8 8 8 ...
## studyweek :  int [1:55] 50 15 15 10 25 20 15 10 12 2 ...
```

You see that the students/**cases** all have corresponding **rows**. They are labeled by the row indices. The **column** names stand for the variable names.

Then, you can do a bit of exploratory analysis.

Task 3.

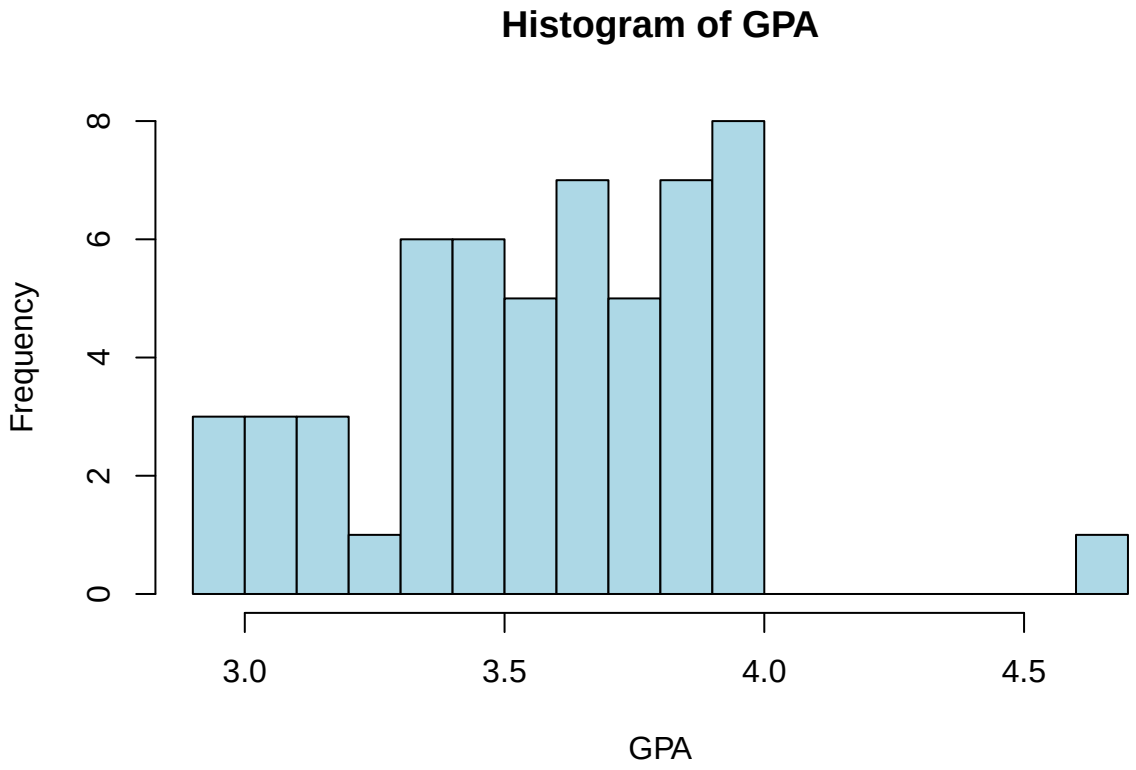
What are the minimum and maximum GPAs? What is the mean GPA? What is the median GPA?

```
gpa<-gpa.data$gpa
min(gpa)
## [1] 2.9
max(gpa)
## [1] 4.67
mean(gpa)
## [1] 3.600073
sd(gpa)
## [1] 0.3356183
summary(gpa)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.900   3.400   3.650   3.600   3.825   4.670
```

Task 4.

Plot the histogram of the GPAs. Make sure that your plot has the main title and that the axes are also labeled.

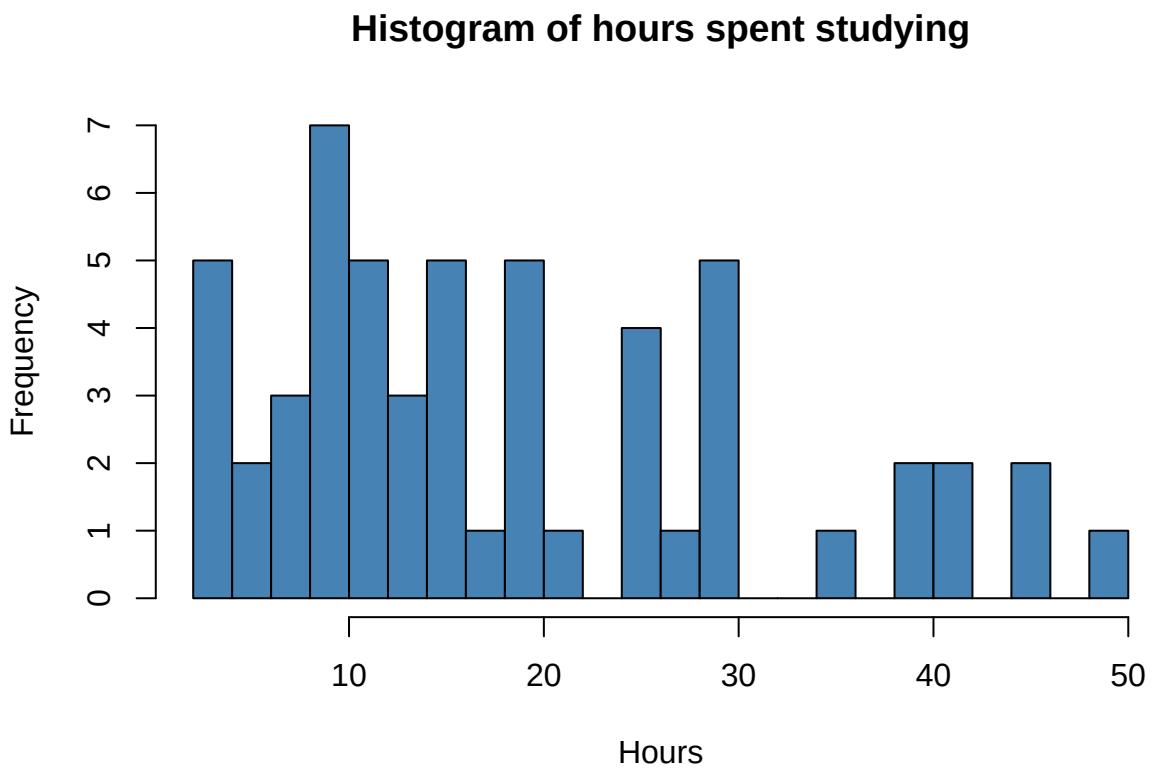
```
hist(gpa,
     breaks=15,
     main="Histogram of GPA",
     xlab="GPA",
     ylab="Frequency",
     col="lightblue")
```



Task 5.

Plot the histogram of the number of hours spent studying per week. Make sure that your plot has the main title and that the axes are also labeled.

```
hrs<-gpa.data$studyweek
hist(hrs,
     breaks=18,
     main="Histogram of hours spent studying",
     xlab="Hours",
     ylab="Frequency",
     col="steelblue")
```



Task 6.

Is the mean number of hours spent studying different for females than for males?

```
gpa.data$gender
##  [1] "female" "female" "female" "male"   "female" "male"   "female" "female"
##  [9] "female" "male"   "female" "female" "female" "female" "female" "male"
## [17] "male"   "female" "female" "female" "female" "female" "male"   "female"
## [25] "female" "female" "female" "female" "female" "female" "female" "female"
## [33] "female" "male"   "female" "female" "female" "female" "male"   "female"
## [41] "female" "female" "male"   "female" "female" "male"   "male"   "male"
## [49] "female" "female" "female" "female" "female" "female" "female"
gpa.data$gender=="female"
##  [1]  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [13]  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [37]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [49]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
hrs.f=hrs[gpa.data$gender=="female"]
hrs.f
##  [1] 50 15 15 25 15 10 12 10 30 30 21 10 12  4 45  6 10 13 35 10 40 14 30  8  8
## [26] 20 40 25 10 18 15 11 28  4 25 42 20  7  6 20 45 30 20
hrs.m=hrs[gpa.data$gender=="male"]
hrs.m
##  [1] 10 20  2 14 12 12 15 30  4  3 42 25
mean(hrs.f)-mean(hrs.m)
## [1] 4.343023
```

Task 7.

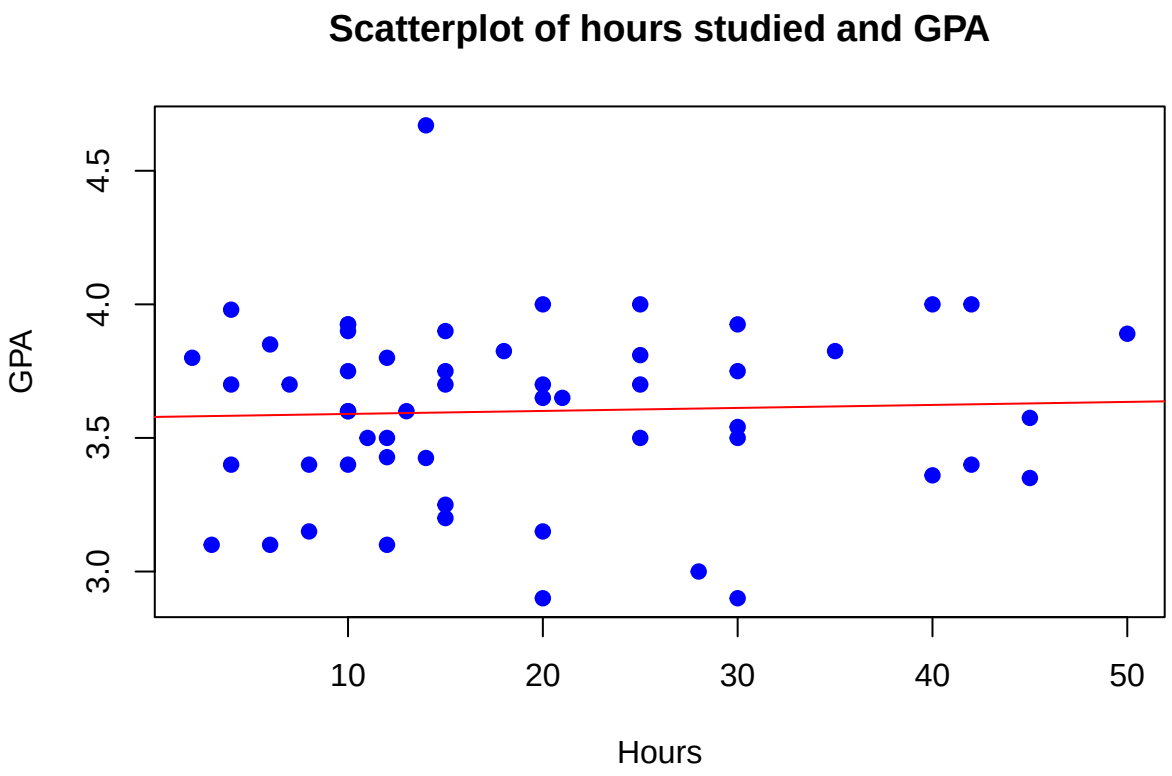
Any difference in the GPA?

```
gpa.f=gpa[gpa.data$gender=="female"]
gpa.m=gpa[gpa.data$gender=="male"]
mean(gpa.f)-mean(gpa.m)
## [1] 0.05125581
```

Task 8.

Is there a relationship between the hours studied and the GPA?

```
plot(hrs,gpa, pch=19, col="blue",
     main="Scatterplot of hours studied and GPA",
     xlab="Hours",
     ylab="GPA")
slr=lm(gpa ~ hrs)
slr
##
## Call:
## lm(formula = gpa ~ hrs)
##
## Coefficients:
## (Intercept)          hrs
##    3.578490     0.001127
summary(slr)
##
## Call:
## lm(formula = gpa ~ hrs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71231 -0.18864  0.04784  0.22274  1.07573
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.578490   0.084568  42.315   <2e-16 ***
## hrs         0.001127   0.003719   0.303    0.763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3385 on 53 degrees of freedom
## Multiple R-squared:  0.001731,   Adjusted R-squared:  -0.0171
## F-statistic: 0.0919 on 1 and 53 DF,  p-value: 0.763
abline(slr, col="red")
```



Task x.

What else could we ask?

```
plot(gpa.data$studyweek,gpa.data$gpa, pch=16,
     col=as.factor(gpa.data$gender))
legend(40, 4.75, legend=c("Male", "Female"),
       col=c("red", "black"), pch=16, cex=0.8)
```