Q1)

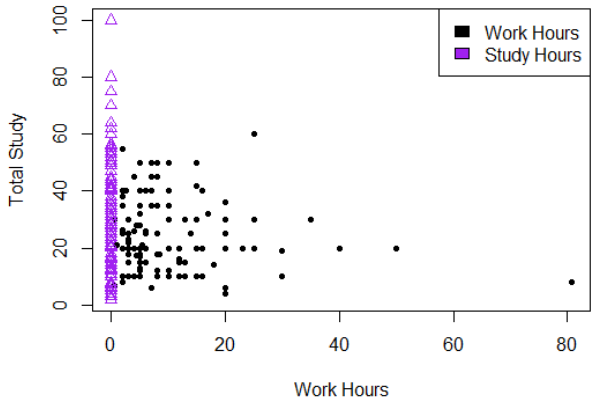
a) The variables that were associated with outliers were Sleep, totalstudy, onestudy and classhour
b) Sleep was the most outstanding one because the survey asked how long a person slept a day and for sleep one was 63 hours which is obviously nonsense. The student most likely believed they were being asked about the whole week not just one day.
c) We could simply remove the ones over 24 hours, but the data is too valuable, so instead we can divide the number of hours by 7 days and we can get the average sleep per day. Since the answers are estimates by the respondents, so we can estimate too without altering the data.
d) I used summary(labdata) which returned to me all the variables with a lot of information, in this case I was looking for the max in each variable. I noticed the max number of hours under sleep was 63 and then I realized that it has happened multiple times.

Q2)

a) There are 5 variables that have NA's and they are: workhours, numcourses, classhours, totalstudy, and onestudy.
b) I will decide to go with the workhours one because if the value is missing then it could mean that the respondent doesn't have a job not necessarily a miss entry.
c) Again, I used summary(labdata) to find the NA's associated with variables.

Q3)

```
> plot(labdata$workhours, labdata$totalstudy,xlab = "Work Hours", ylab= "Tota
l Study", col=ifelse(labdata$workhours>0,'black','purple'), pch=ifelse(labdat
a$workhours>0, 20, 2))
```



I have decided to use scatter plot to plot this graph. Using the two variables, totalstudy and workhours since we are comparing the two. I used col=ifelse() and pch=ifelse() to have two different colours and shapes in my graph so we can identify which one is which. I also added a legend and for that I used the following code:

```
> legend('topright', legend=c("Work Hours","Study Hours"), fill=c('black','pu
rple'))
```

4)

a) `> labdata$freetime <- 168 - labdata$workhours - labdata$classhours - labdata$sleep - labdata$totalstudy`

b) to calculate the mean and standard deviation I first removed the NA's from the freetime variable using newdata<-na.omit(labdata$freetime)

and to calculate the mean I did: mean(newdata) and I got 113.7435

and to calculate the standard deviation I did: sd(newdata) and I got 17.52923