STAT-S 520 Data Analysis Project

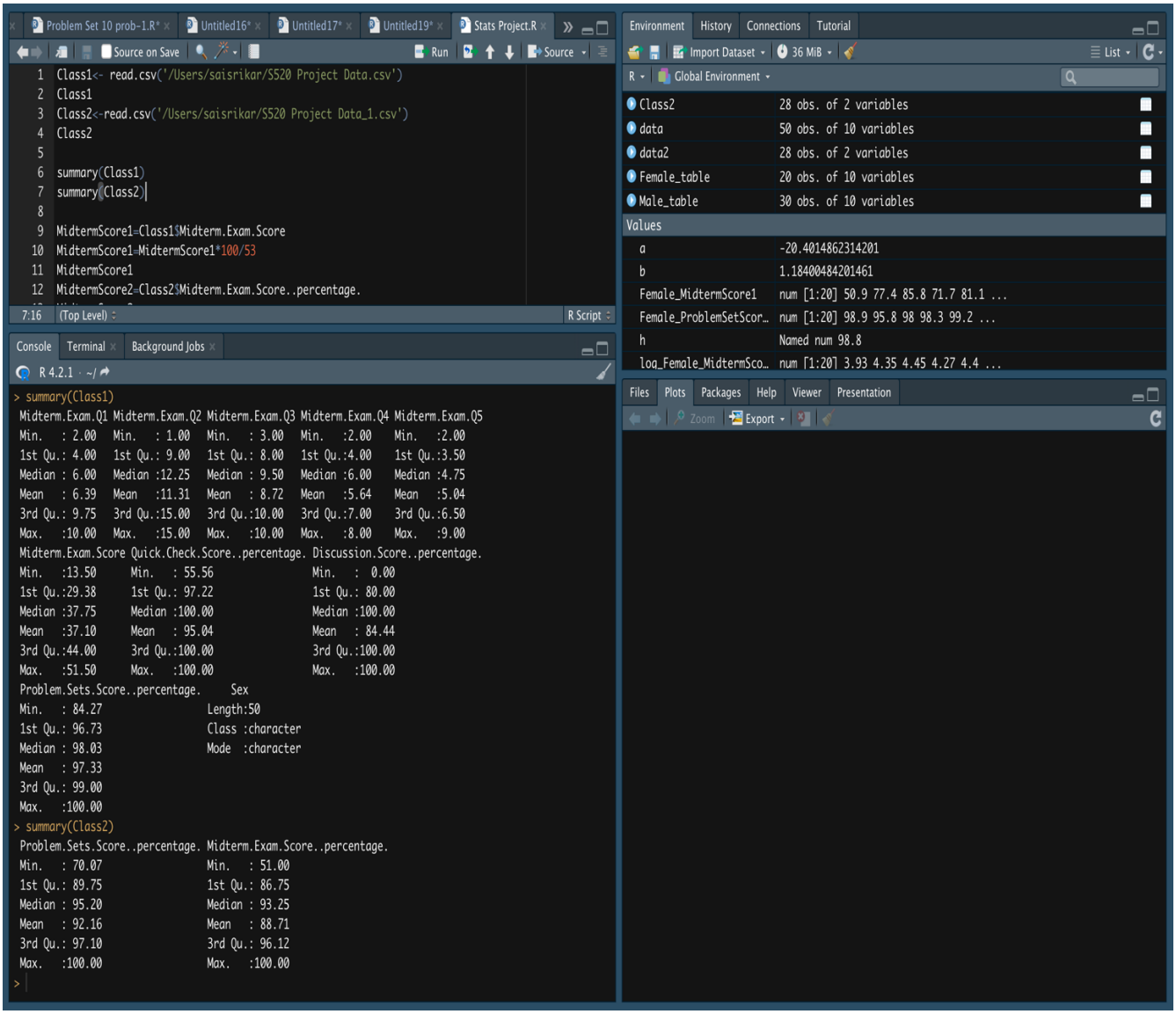
SAI SRIKAR GANDHE[1], MANASA GUDISE[1]

1. **Research question**: Compare these two classes' performance in the first half of Fall semester.

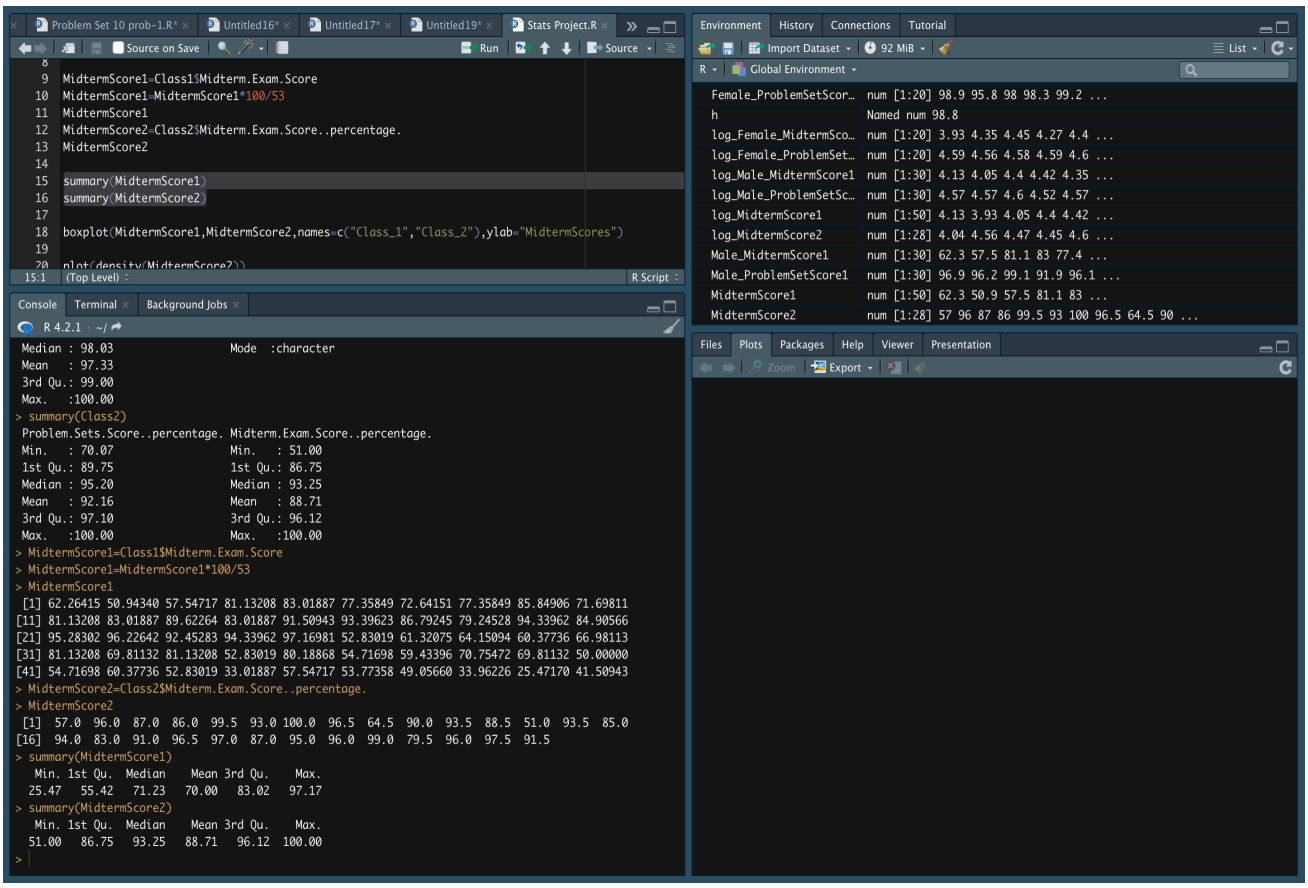
   Topic: Descriptive Statistics

   Purpose of the Topic: To compare two classes performance and see which class's performance is better which may help the instructor to decide which one of the classes students need more attention and help.

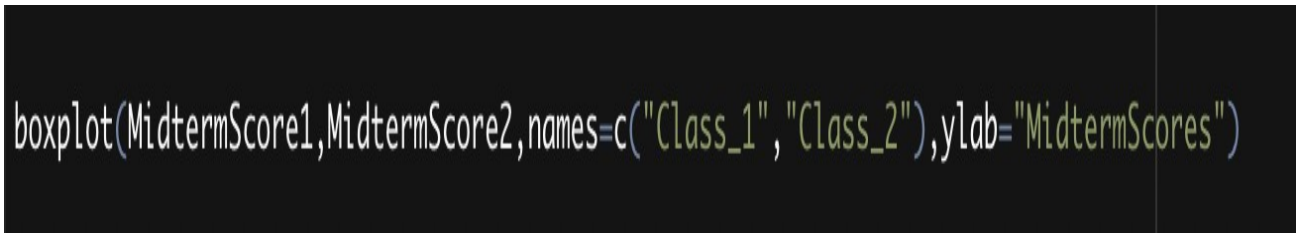
   Let's consider the midterm scores of both classes to see which class performance is better. Let's see the summary of both classes

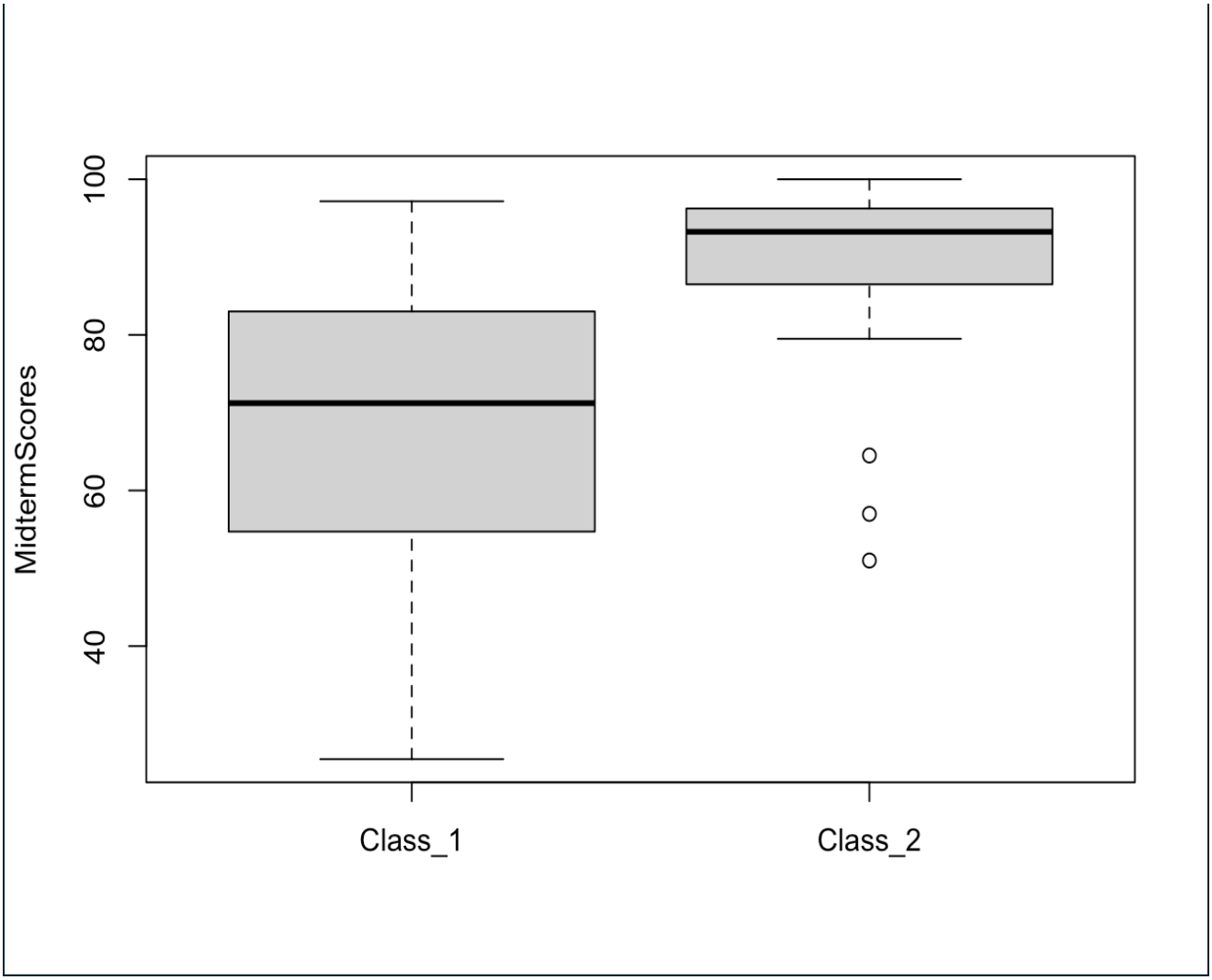


As we are considering midterm scores of both classes, let 'MidtermScore1' be the midterm score percentage of class1 and 'MidtermScore2' be the midterm score percentage of class2. Now, let's look at the numerical summaries of midterm scores of class1 and class2



From above summary statistics we can see that the mean and median of class2 midterm scores is higher than the class1 midterm scores. But we cannot say that for sure by just looking at numerical summary statistics. So, let's further investigate. Let's draw some boxplots for comparisons of midterm scores of class1 and class2. Below is a box plots of midterm scores of both classes

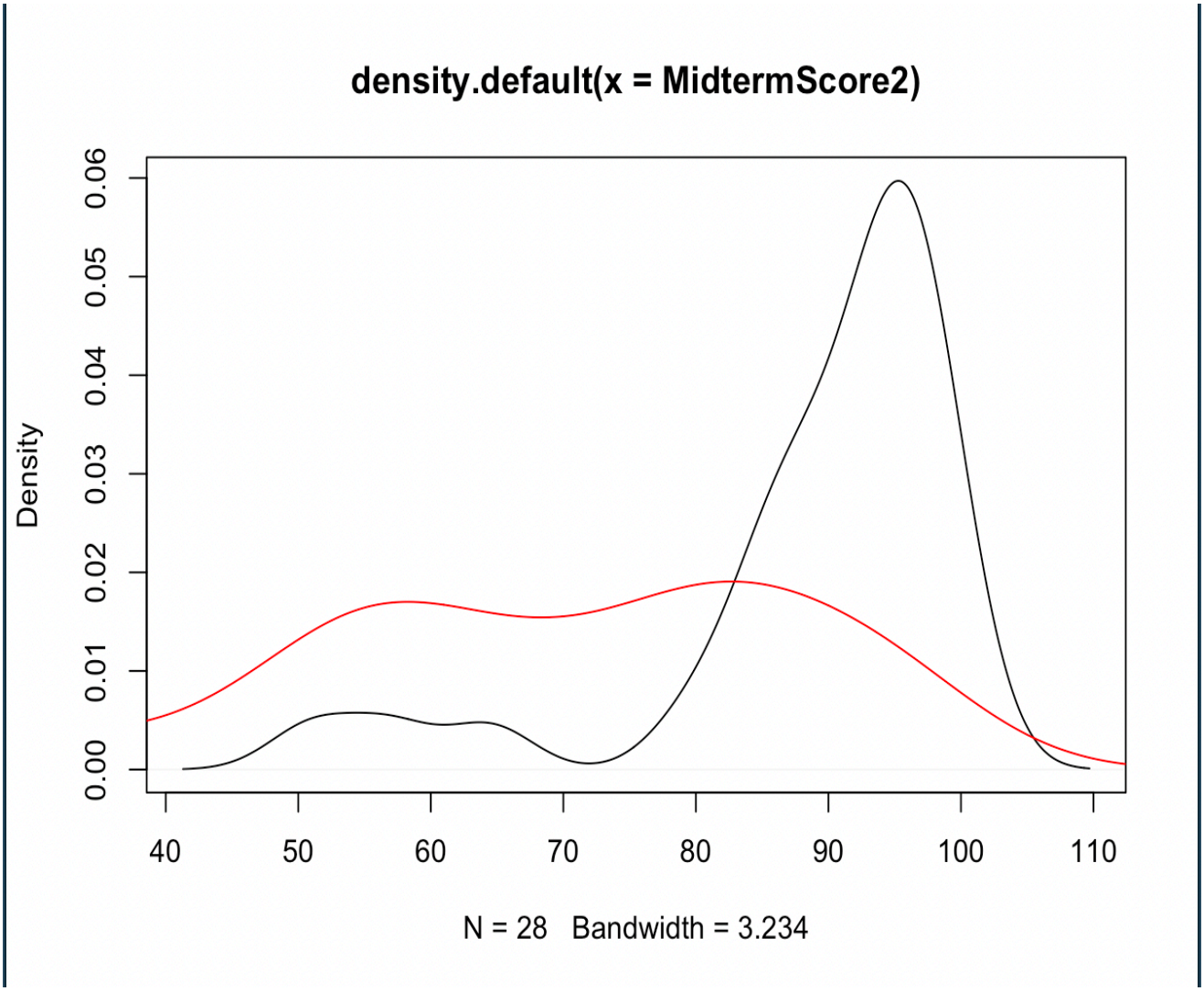




From the above boxplot, we can see that the median of the midterm scores of class1 lie around 71 and the median of the midterm scores of class2 lie around 93. We can also see that there are only three students in class2 whose midterm scores are less than 70%. Most of the students' scores in class2 lie between 86% and 100% whereas most of the students score in class1 lie between 55% and 97%.

Let's draw density plot of midterm scores of both classes

```
plot(density(MidtermScore2))
lines(density(MidtermScore1), col="red")
```



We can see that the distribution of midterm scores of class1 (red) is shifted to the left compared to midterm scores of class 2(black).

To find out more we can do a hypothesis test.

Hypothesis test:

Here we are considering midterm scores of both classes so it's a two-sample location problem and the parameter of interest is 'Δ' where $\Delta = \mu_2 - \mu_1$. Let μ_1 be population mean of the midterm scores of class1 and μ_2 be the population mean of midterm scores of class2. So, we want to see if the population mean of the class2 is greater than mean of class2. So, this becomes our alternative hypothesis i.e, $\mu_2 - \mu_1 > 0$ which says it is a right tailed test. Null hypothesis is exactly opposite to alternative hypothesis.

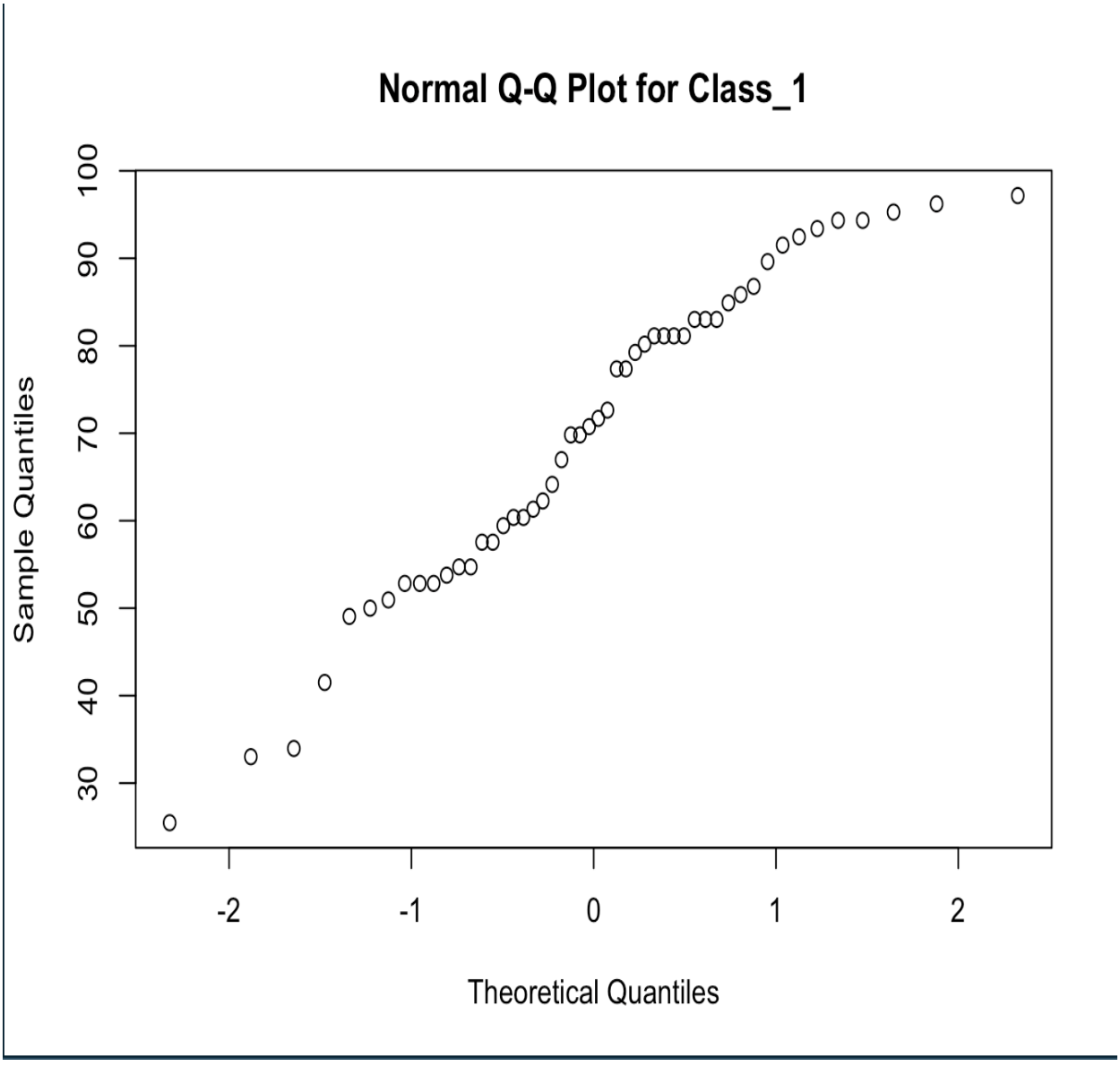
$$H_0: \Delta \leq 0$$

$$H_1: \Delta > 0$$

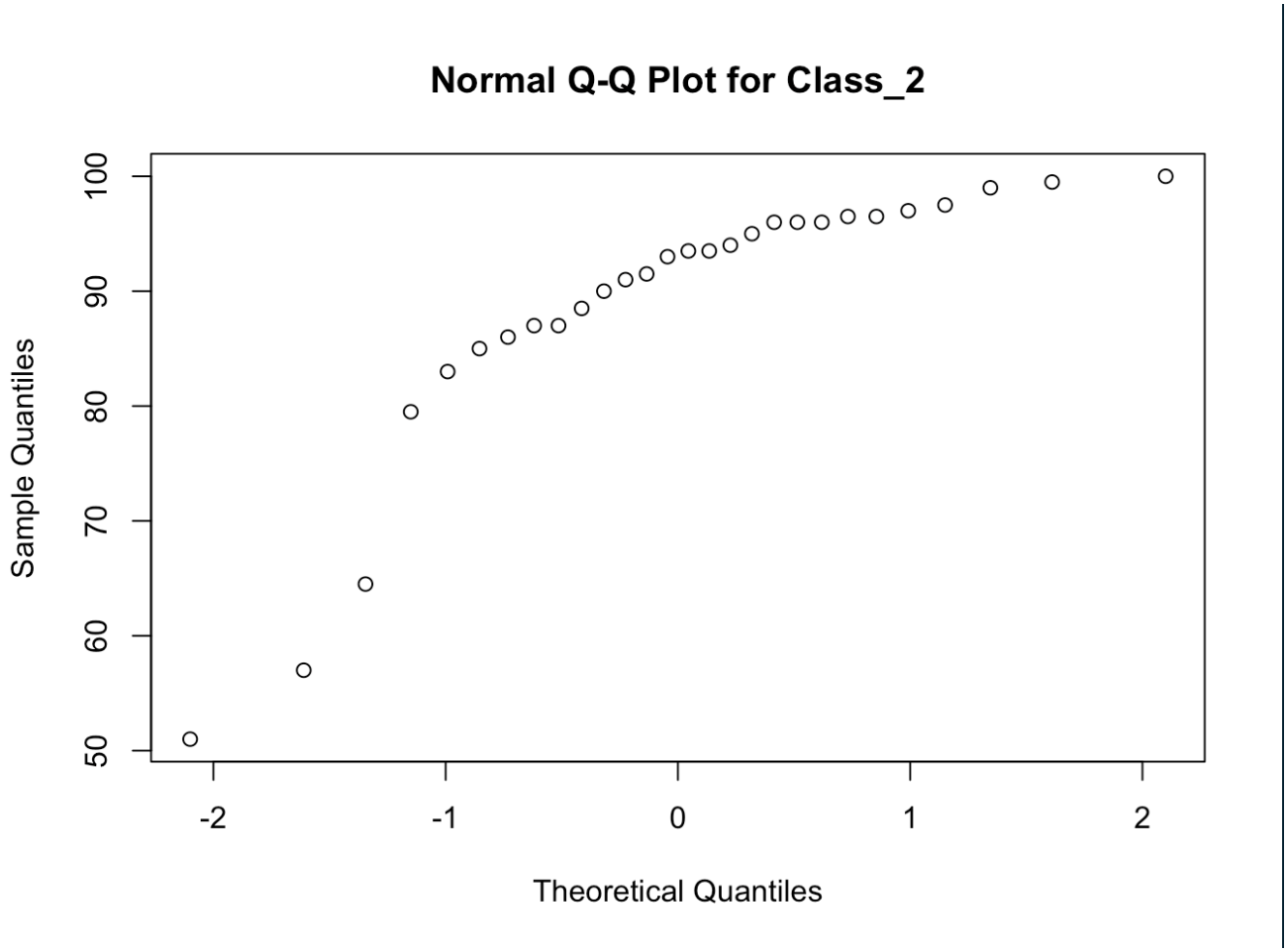
Now, let's draw QQ plots of midterm scores of both classes to see if the distribution is normal.

QQ plot of class1 midterm scores:

```
qqnorm(MidtermScore1,main='Normal Q-Q Plot for Class_1')
qqnorm(MidtermScore2,main='Normal Q-Q Plot for Class_2')
```



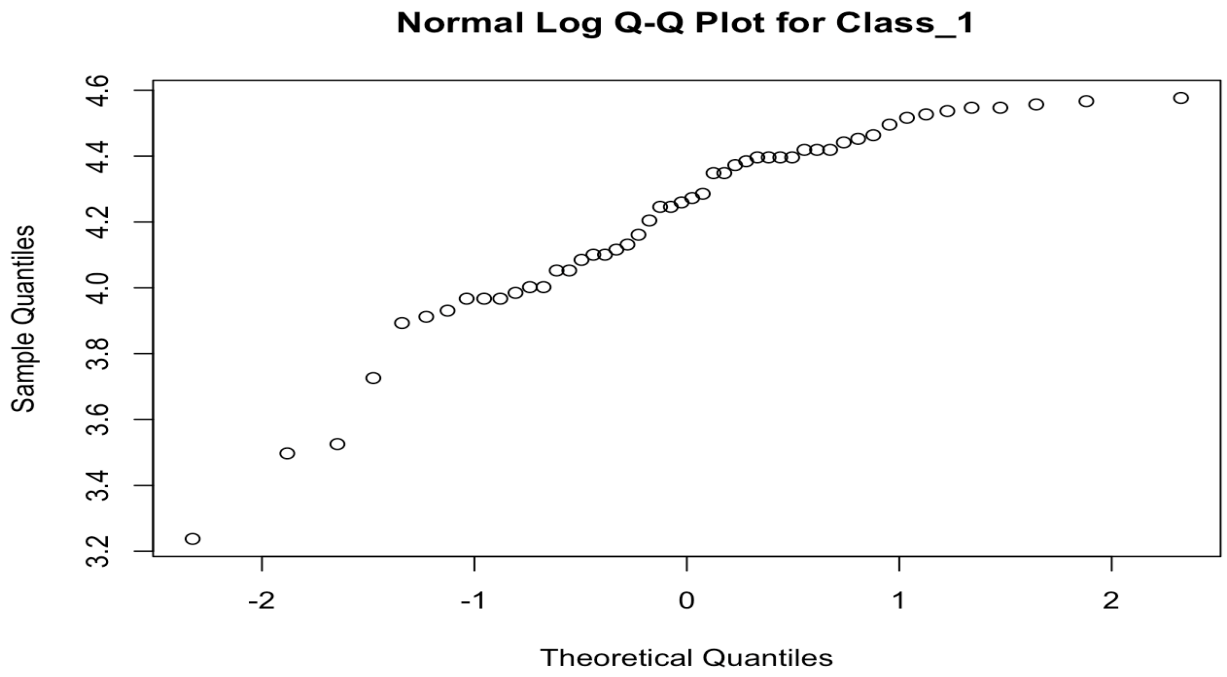
QQ plot of class2 midterm scores:



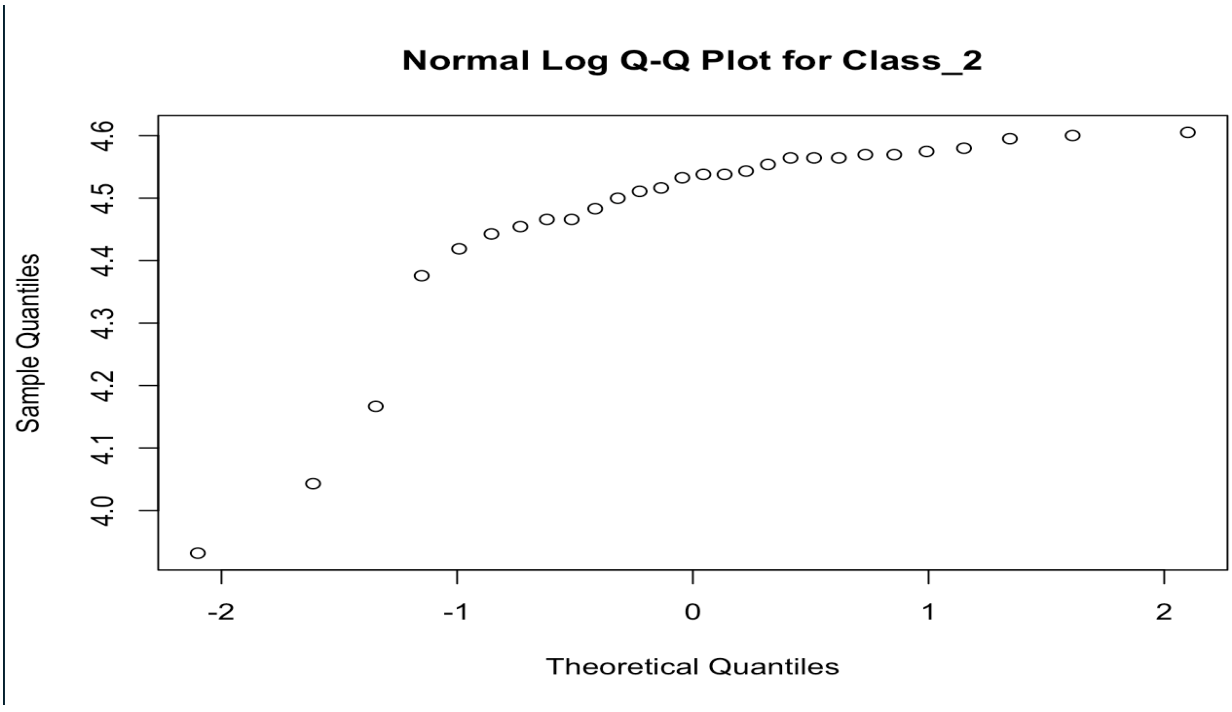
From above plots we can see that the QQ plot for class1 midterm scores looks almost like a straight line and the QQ plot for class2 midterm scores is more like curved line than a straight line. Now, let us transform the data and plot the graphs. Below is a QQ plot for log transformed midterm scores of class1.

RCode:

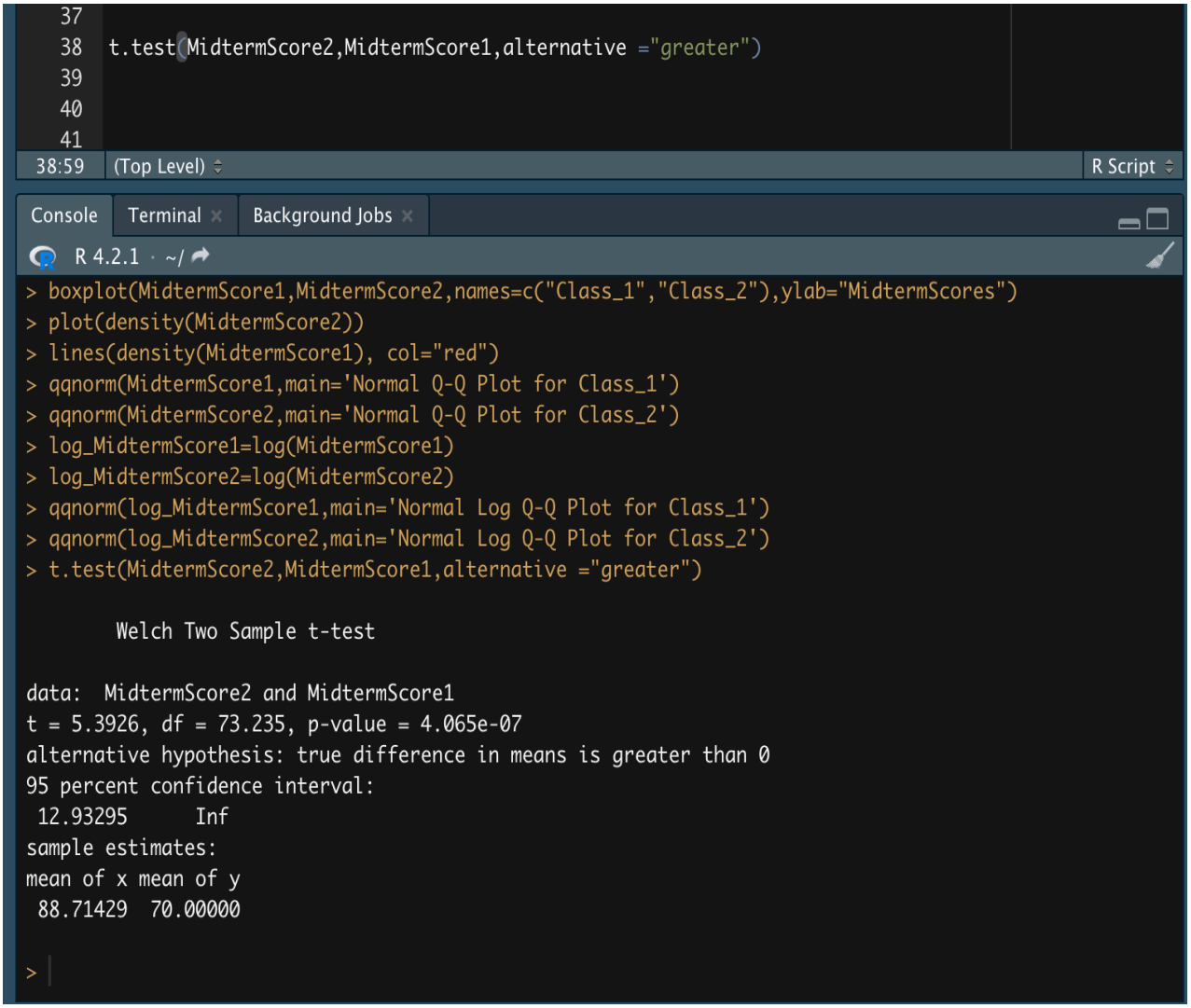
```
# using log
log_MidtermScore1=log(MidtermScore1)
log_MidtermScore2=log(MidtermScore2)

qqnorm(log_MidtermScore1,main='Normal Log Q-Q Plot for Class_1')
qqnorm(log_MidtermScore2,main='Normal Log Q-Q Plot for Class_2')
```



QQ plot of Log transformed Midterm Scores of class2:



The log transformation made the QQ plot of class1 more curved so let's stick with the untransformed data. Assuming that samples are IID samples from two independent normal populations, we can do Welch's two- sample t-test.



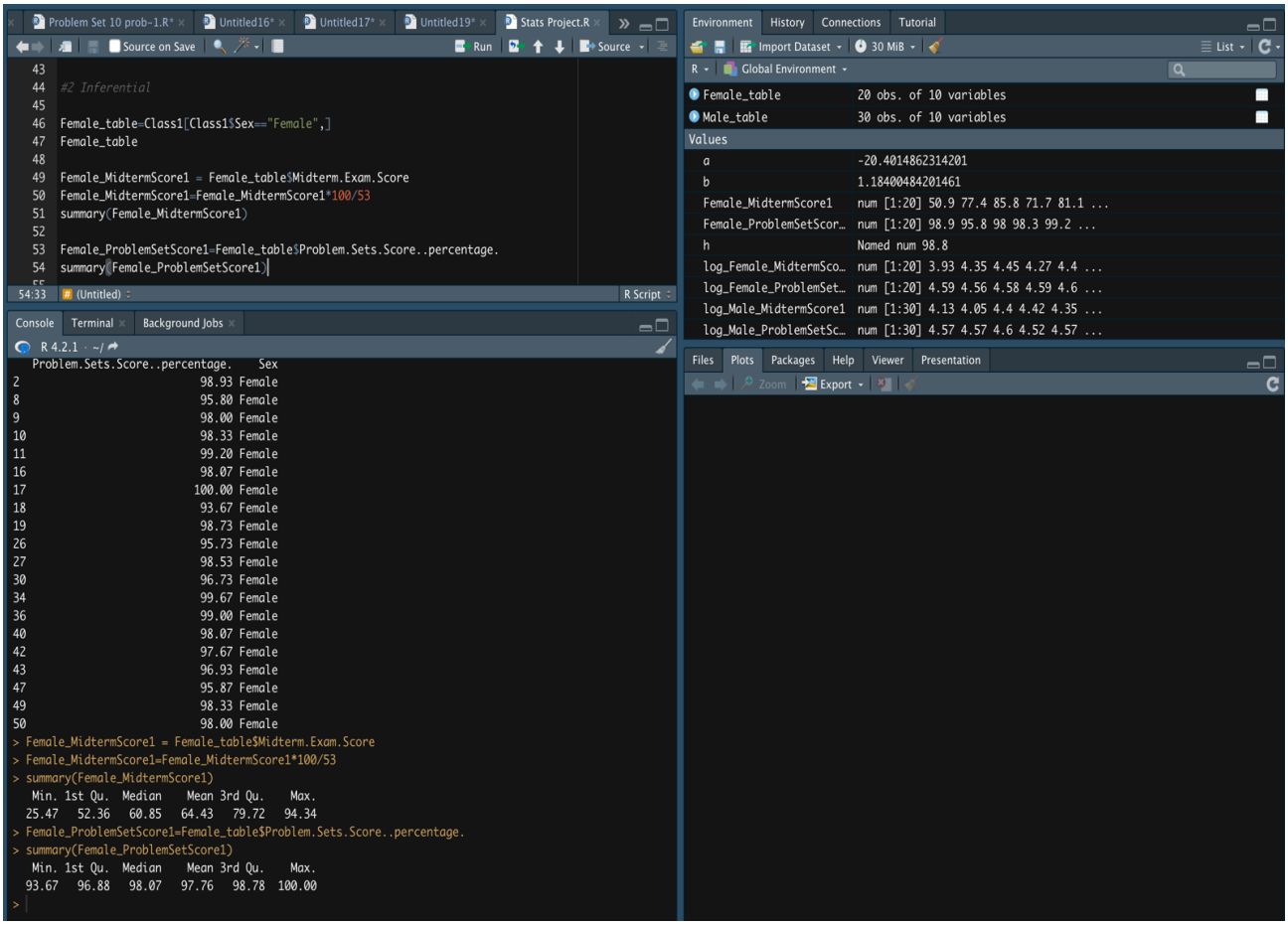
From above t-test we can see that the p-value is 4.065e-07 which is very small, so it means we have evidence against the null hypothesis. So, we need to reject null hypothesis. So, our alternative hypothesis becomes true. That is class2 performance is better than class1 performance when taken midterm scores into account. From above, we can see that the 95% confidence interval is (12.93295,∞).

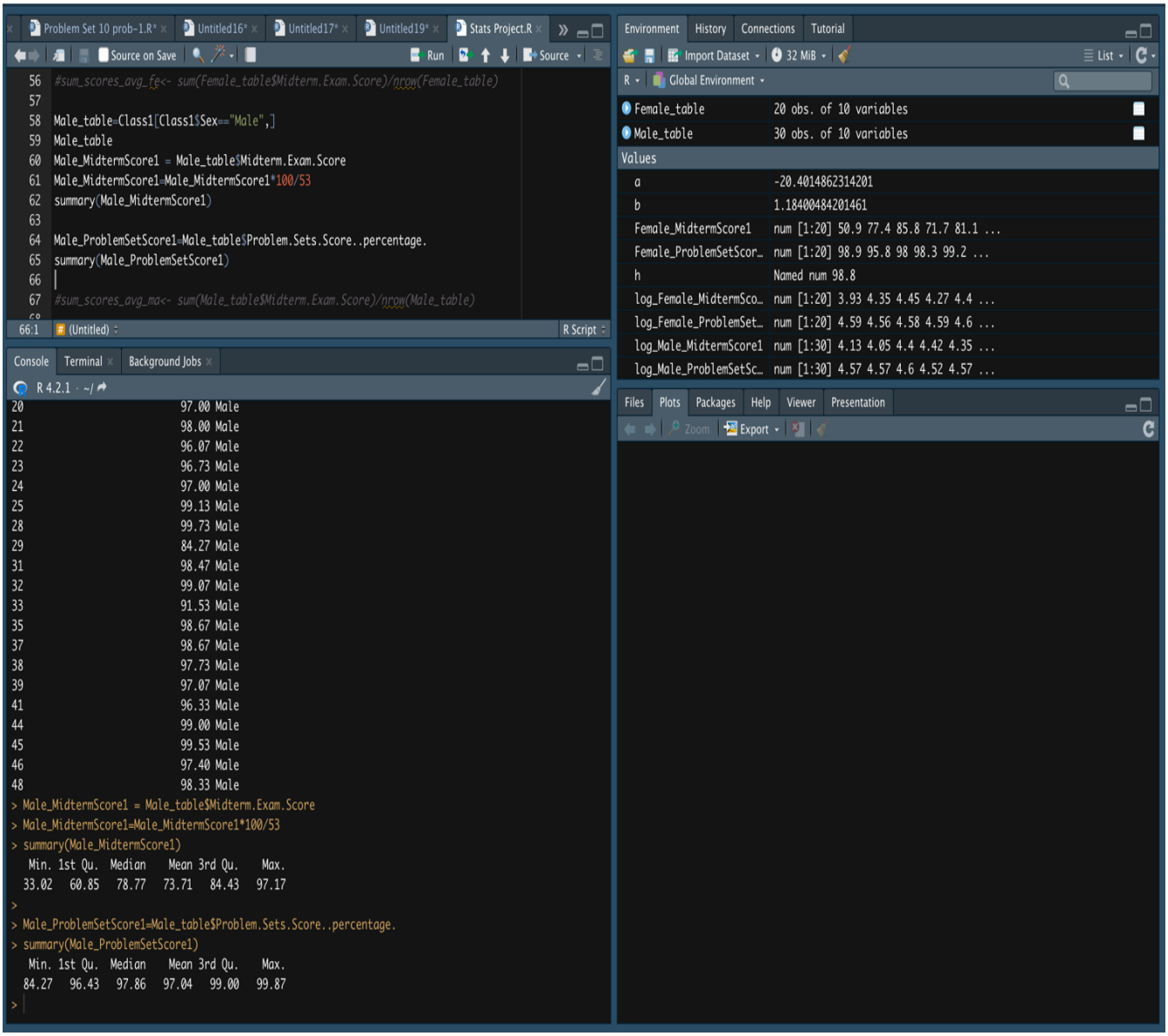
2. **Research question:** Does gender make a difference in student's performance on assignments and midterm exam in S520?

Topic: Inferential Statistics

Purpose of the Topic: To see if male students or female students perform better in assignments and midterm exam.

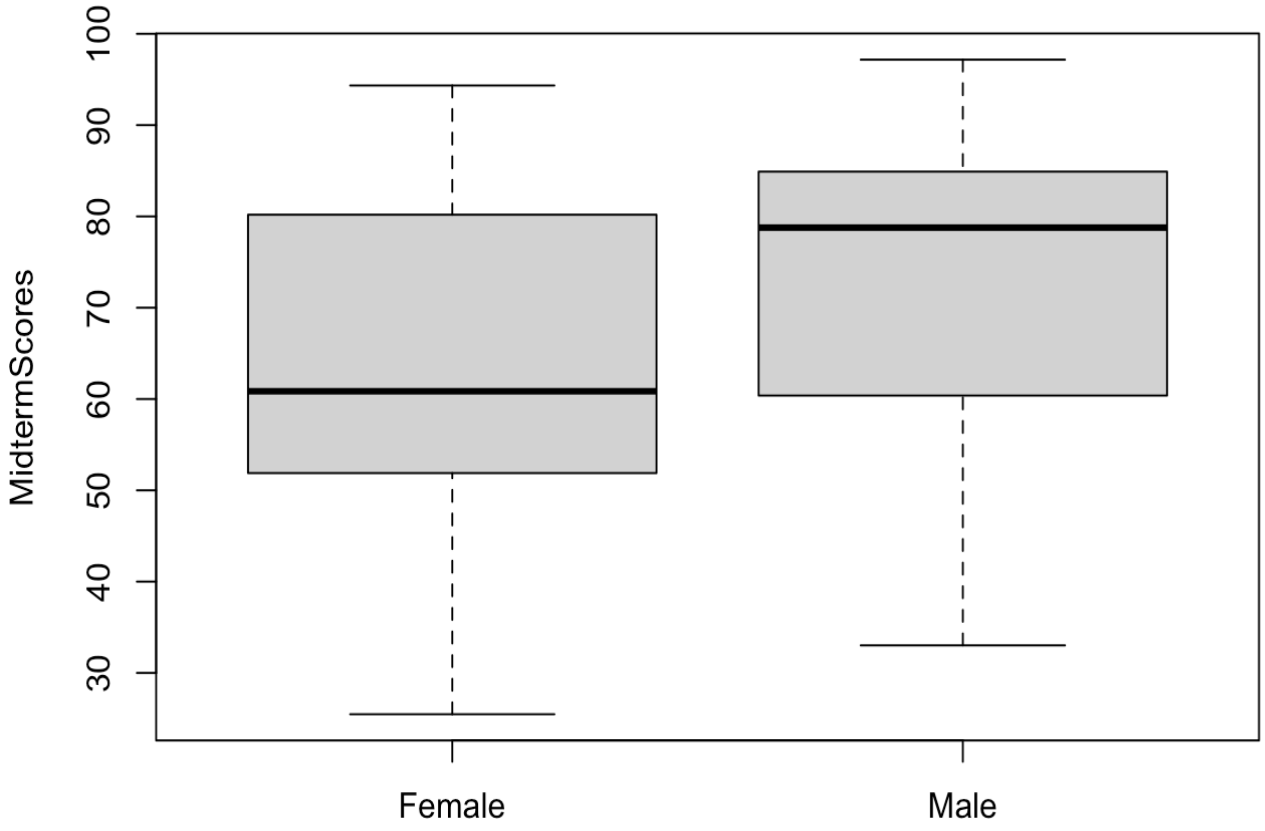
Firstly, let's see summary of midterm scores and problem set scores of female and male students





Let’s draw some boxplots for comparisons of midterm scores of female and male students of class1. Below is a box plots of midterm scores of both classes

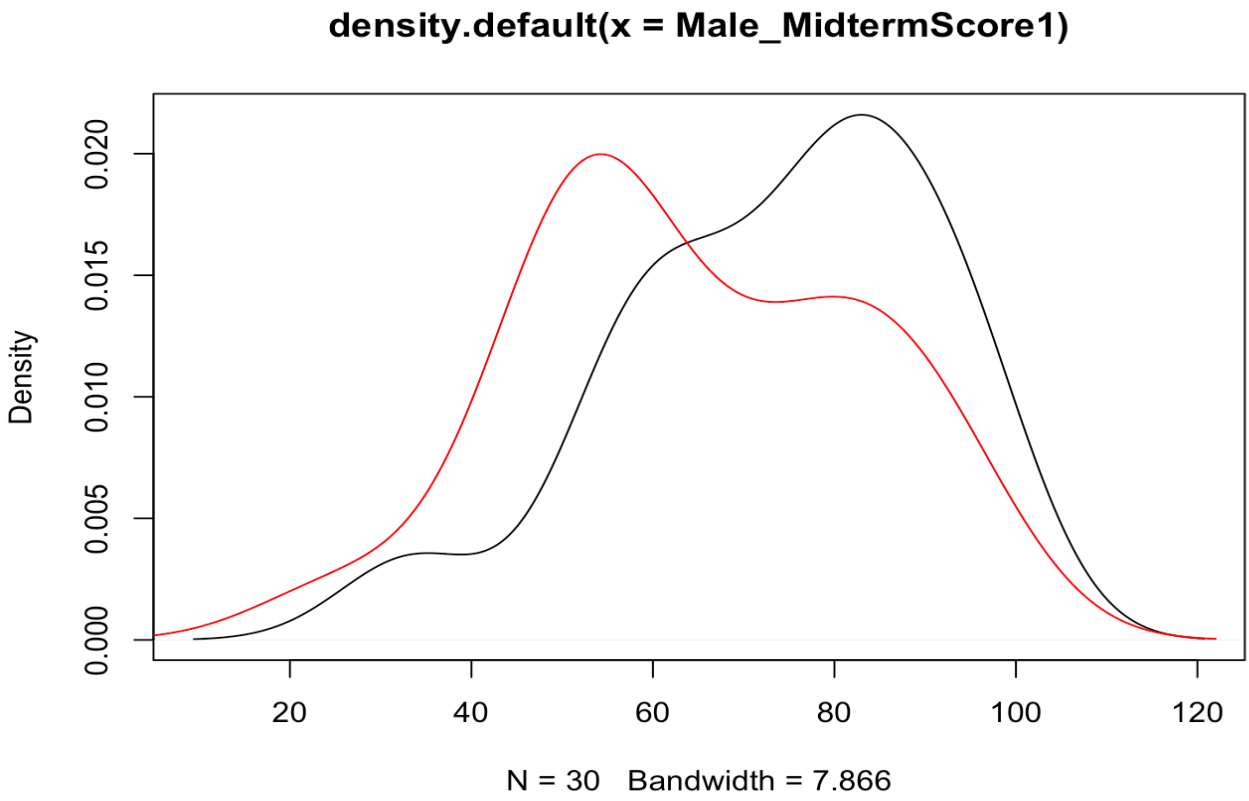
```
boxplot(Female_MidtermScore1,Male_MidtermScore1,names=c("Female","Male"),ylab="MidtermScores")
```



From the above boxplot, we can see that the median of the female midterm scores lie around 60 and the median of the midterm scores of male students lie around 78. 50% of the female students scores lie between 52 and 80 whereas male student scores lie between 60 and 84.

Let's draw density plots of midterm scores of male and female students.

```
plot(density(Male_MidtermScore1))
lines(density(Female_MidtermScore1), col="red")
```



We can see that the distribution of female midterm scores of class1 (red) is shifted to the left compared to male midterm scores of class1(black).

Hypothesis test:

Here we are considering midterm scores of both male and female students. so it's a two-sample location problem and the parameter of interest is 'Δ' where $\Delta = \mu_2 - \mu_1$. Let μ_1 be population mean of the midterm scores of male students and μ_2 be the population mean of midterm scores of female students. So, we want to see if the population mean of the female students is equal to mean of male students. So, this becomes our alternative hypothesis i.e, $\mu_2 - \mu_1 = 0$ which says it is a two-tailed test. Null hypothesis is exactly opposite to alternative hypothesis. Assuming the significance level (α) is 0.05.

$H_0: \Delta \neq 0$

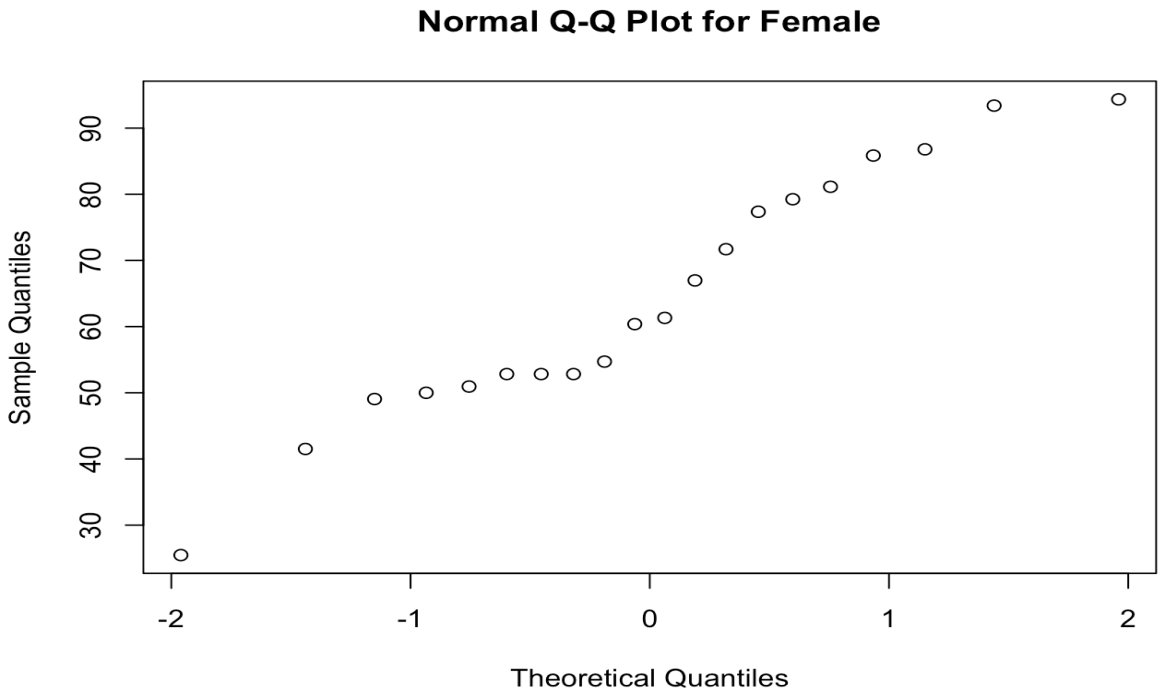
$H_1: \Delta = 0$

Now, let's draw QQ plots of midterm scores of both genders to see if the distribution is normal.

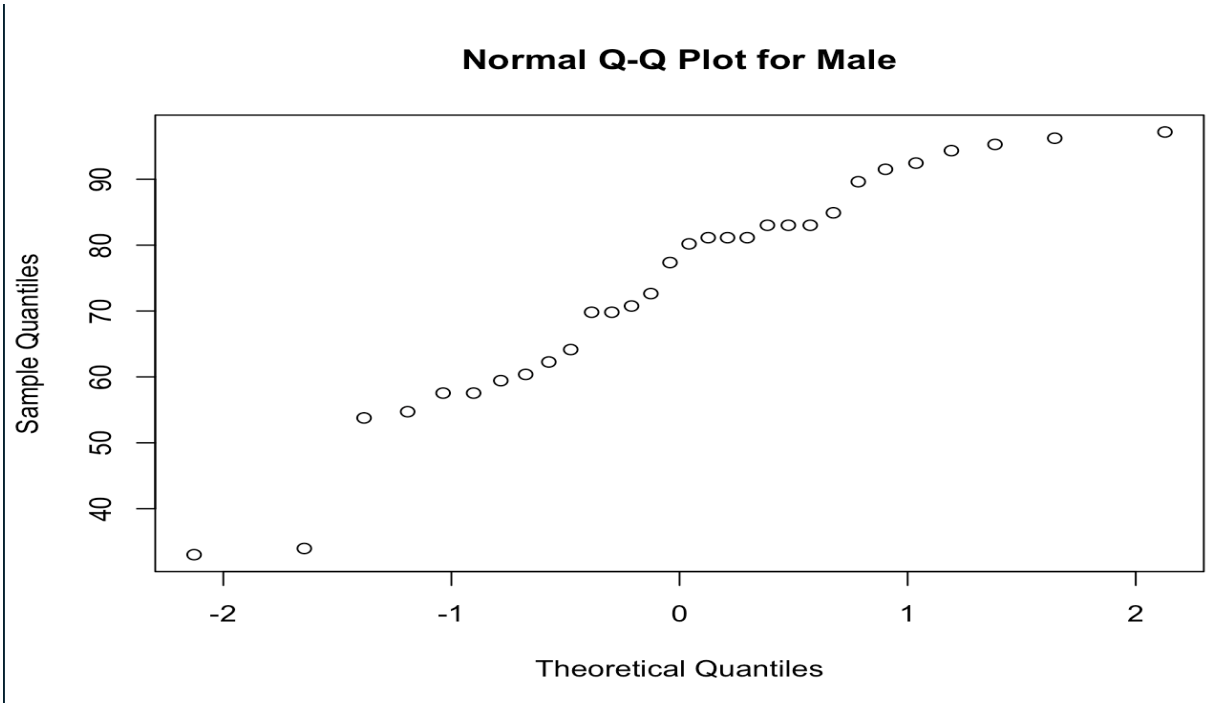
QQ plot of female and male student's midterm scores:

```
qqnorm(Female_MidtermScore1,main='Normal Q-Q Plot for Female')
qqnorm(Male_MidtermScore1,main='Normal Q-Q Plot for Male')
```

QQ plot of female student's midterm score:

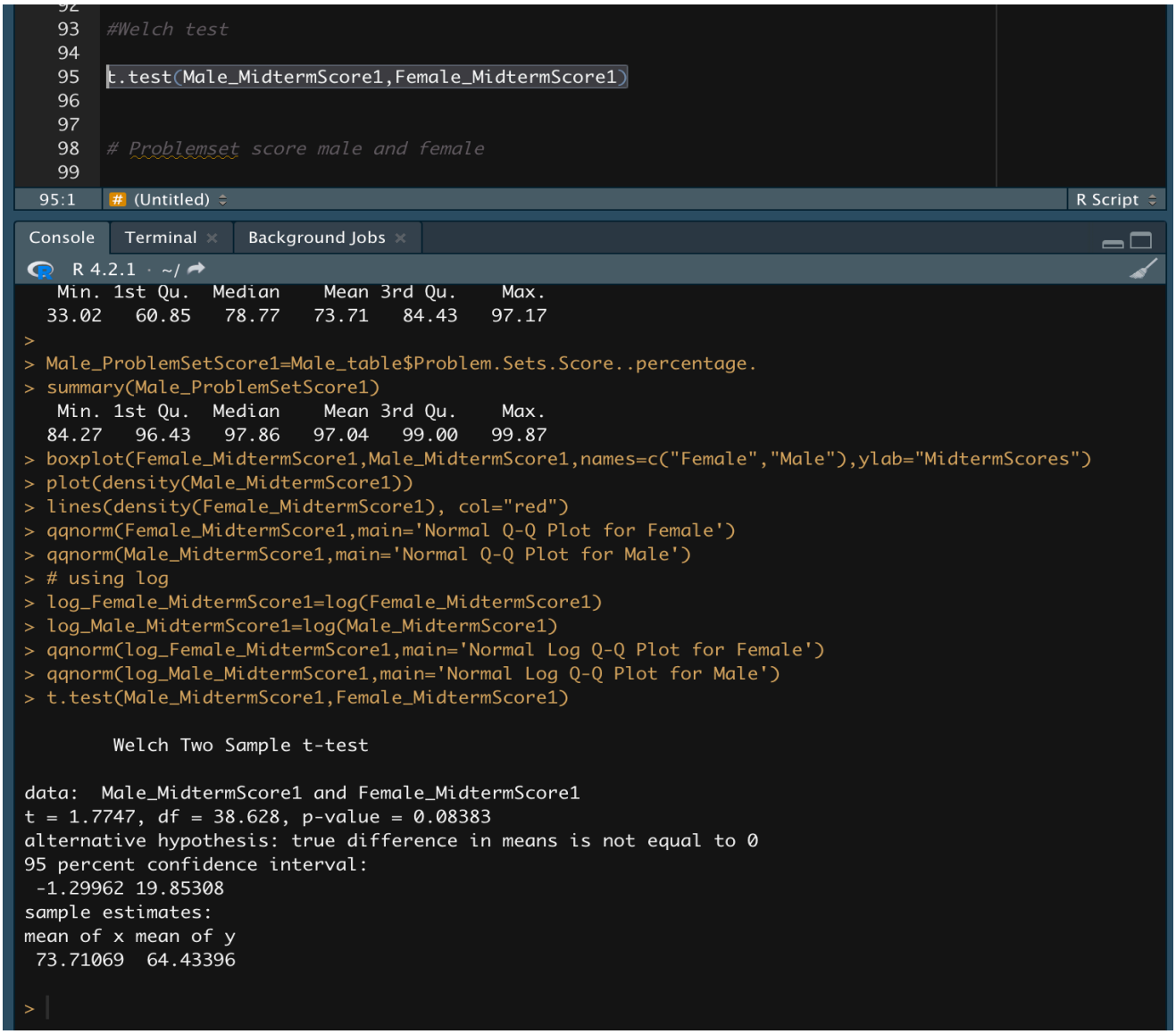


QQ plot of male student's midterm score:



The QQ plot of male students Midterm score looks like a straight line except one outlier which is at 33. The QQ plot of female student scores also looks nearly straight except it is bent around 60.

Assuming that samples are IID samples from two independent normal populations, we can do Welch's two- sample t-test.

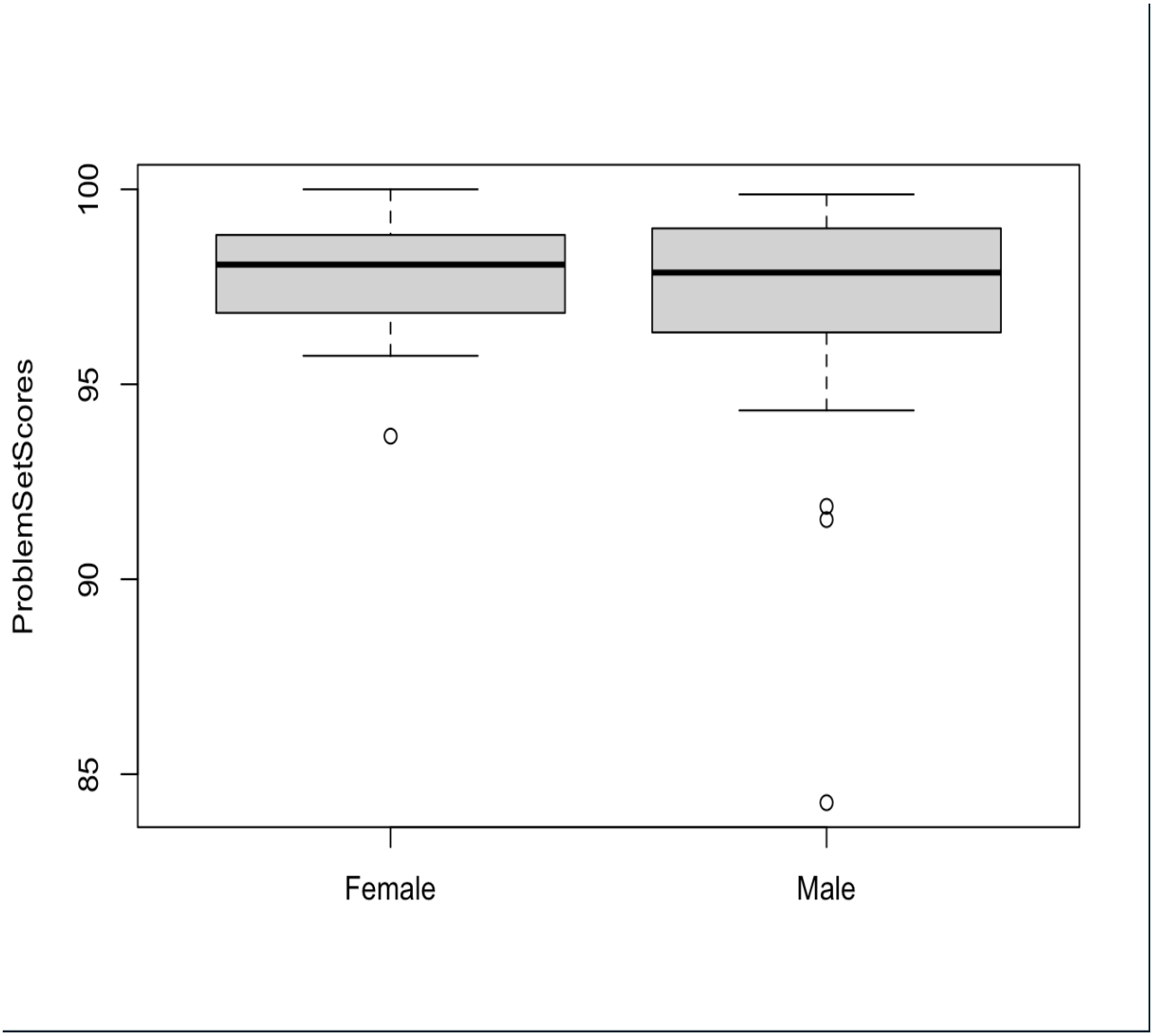


From above t-test we can see that the p-value is 0.08383 which is greater than significance level (0.05) small, so it means we do not have any evidence against the null hypothesis. So, we cannot reject null hypothesis. So, our null hypothesis becomes true. That is gender does make difference in performance when taken midterm scores into account. From above, we can see that the 95% confidence interval is (-1.29962,19.85308)

Now, Let's compare male and female students based on problem set scores.

Let’s draw some boxplots for comparisons of problem set scores of female and male students of class1. Below is a box plots of problem set scores of both classes.

```
boxplot(Female_ProblemSetScore1,Male_ProblemSetScore1,names=c("Female","Male"),ylab="ProblemSetScores")
```

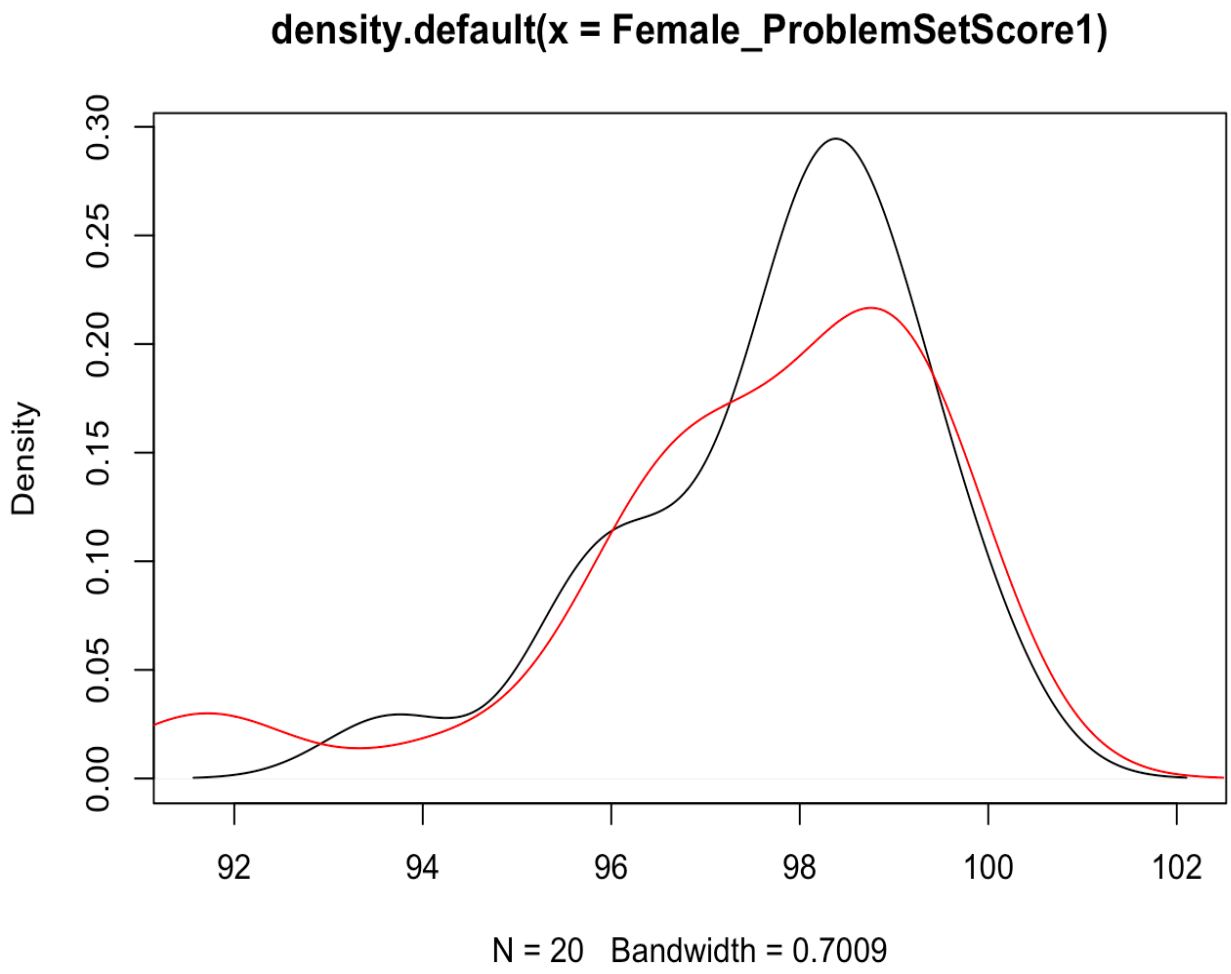


From above boxplot, we can see that the medians of problem set scores of both female and male students are nearly equal. We can also see that there is a outlier in female students scores which is around 93 .There are three outliers in male students scores in which two outliers lie between 90 and 92 and one outlier right below 85.

Let’s draw density plot of problem set scores of male and female students.

R Code:

```
plot(density(Female_ProblemSetScore1))
lines(density(Male_ProblemSetScore1), col="red")
```



We can see that the distribution of female problem set scores of class1 (black) is nearly same to the male problem set scores of class1(red).

Hypothesis test:

Here we are considering problem set scores of both male and female students. so it's a two-sample location problem and the parameter of interest is 'Δ' where $\Delta = \mu_2 - \mu_1$. Let μ_1 be population mean of the problem set scores of male students and μ_2 be the population mean of problem set scores of female students. So, we want to see if the population mean of the problem set scores of female students is equal to mean of problem set scores of male

students. So, this becomes our alternative hypothesis i.e, $\mu_2 - \mu_1 = 0$ which says it is a two-tailed test. Null hypothesis is exactly opposite to alternative hypothesis.

$$H_0: \Delta \neq 0$$

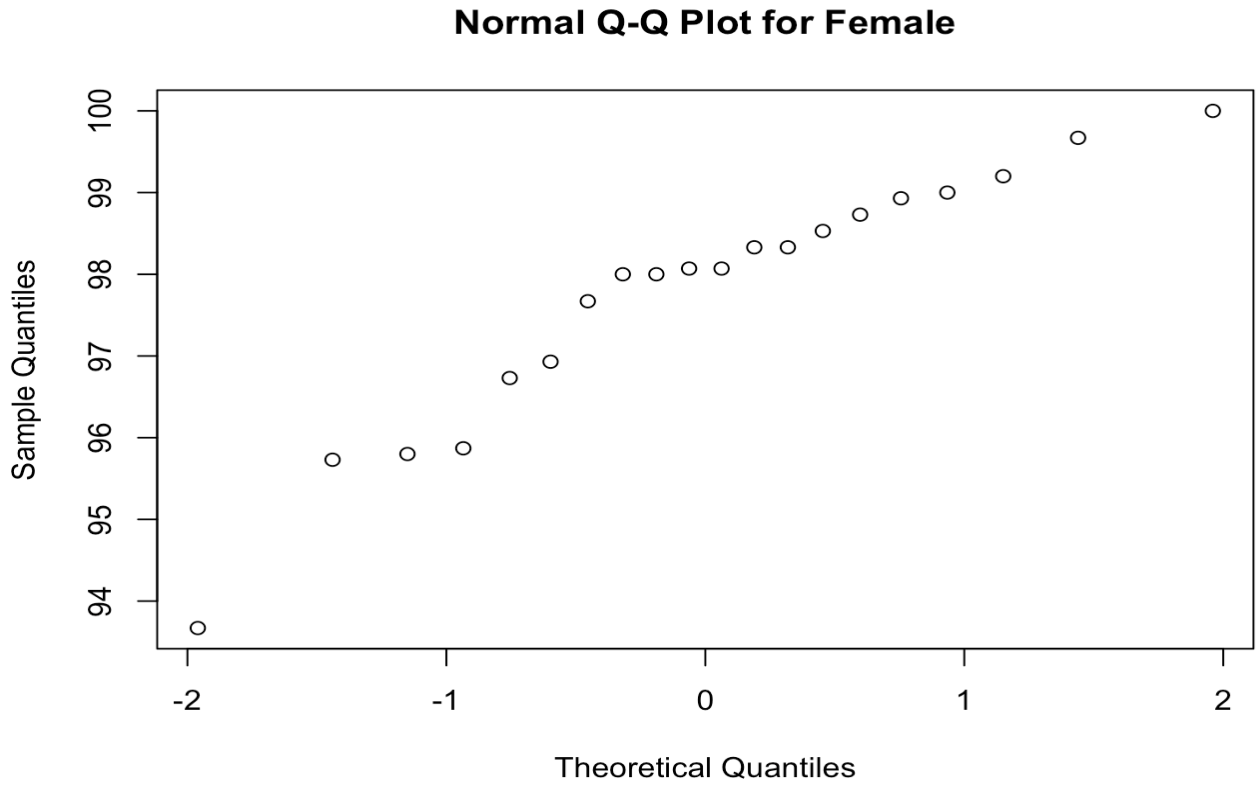
$$H_1: \Delta = 0$$

Now, let's draw QQ plots of problem set scores of both genders to see if the distribution is normal.

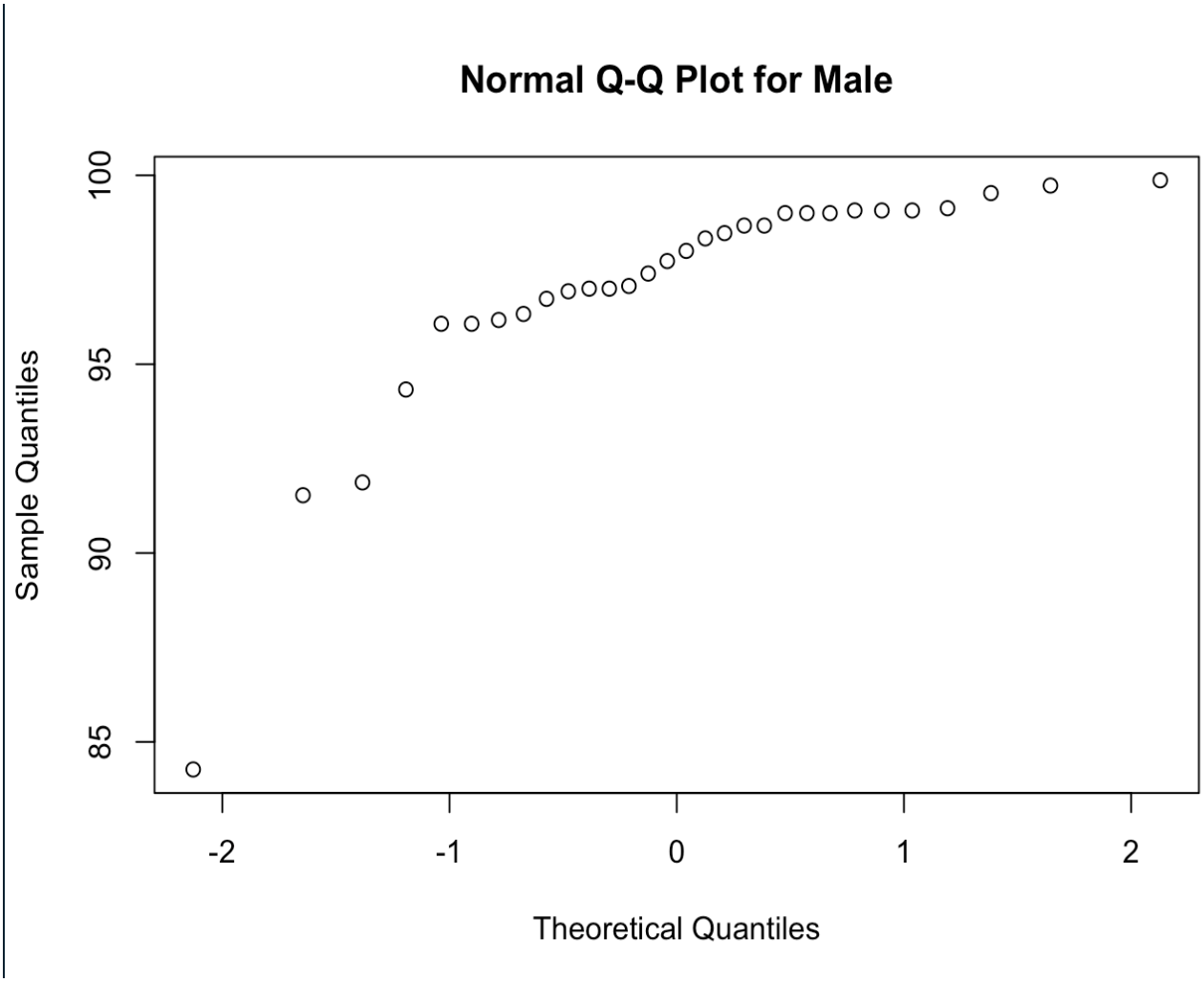
QQ plot of female and male student's problem set scores:

```
qqnorm(Female_ProblemSetScore1,main='Normal Q-Q Plot for Female')
qqnorm(Male_ProblemSetScore1,main='Normal Q-Q Plot for Male')
```

QQ plot of female student's problem set score:

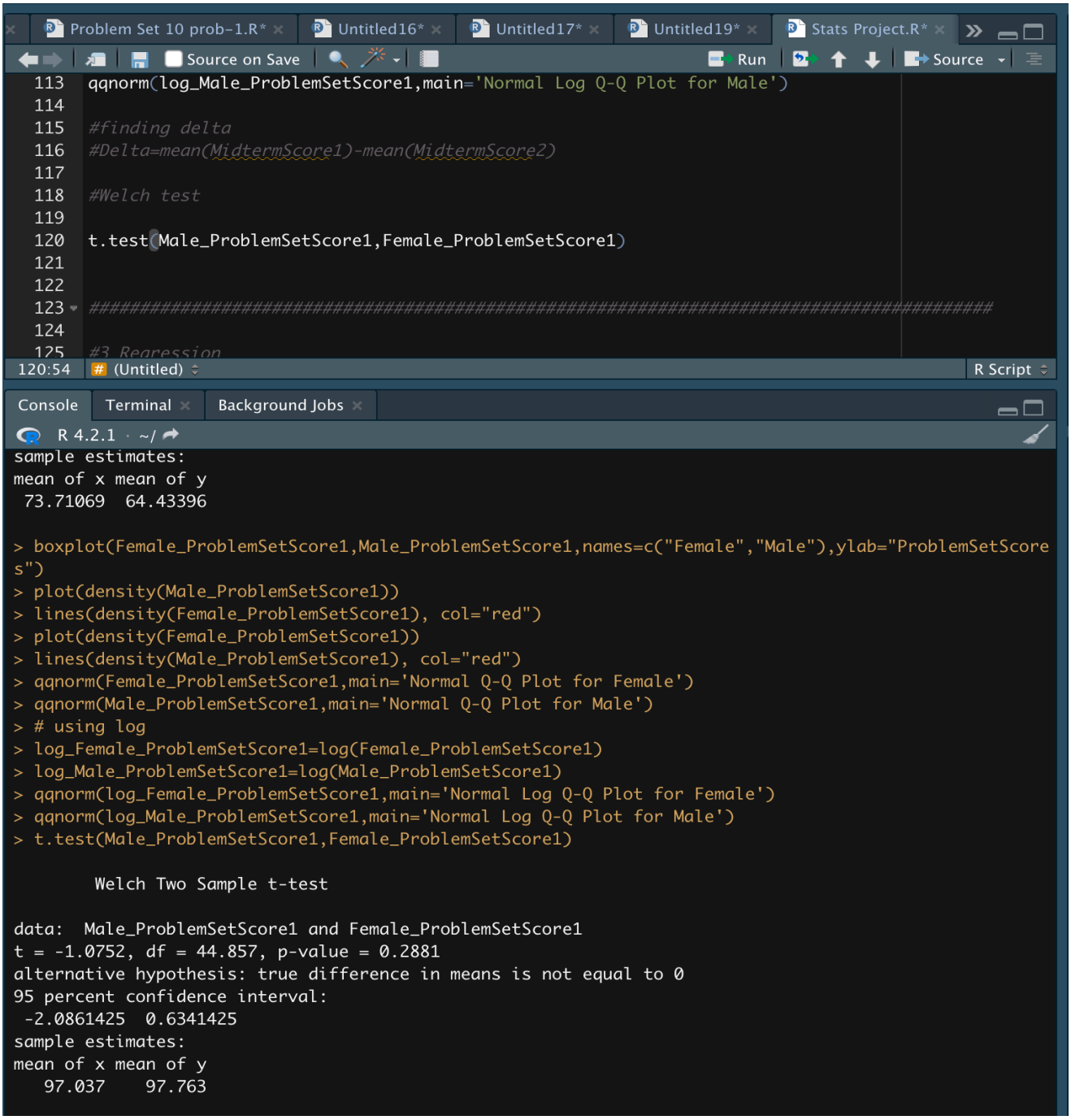


QQ plot of male student's problem set score:



The QQ plot of female student's problem set score looks like a straight line but the QQ plot of male student's problem set score does not look like a straight line it looks like a curved line with outlier right below 85.

Assuming that samples are IID samples from two independent normal populations, we can do Welch's two- sample t-test.



From above t-test we can see that the p-value is 0.2881 which is NOT small, so it means we do not have any evidence against the null hypothesis. So, we cannot reject null hypothesis. So, our null hypothesis becomes true. That is gender does make difference in performance when taken problem set scores into account. From above, we can see that the 95% confidence interval is (-2.0861425,0.6341425)

To conclude, gender does make difference in students' performance in assignments and midterm exam in S520.

3. **Research question:** Should we use students' assignment grades to predict their midterm exam scores? Is the conclusion the same for different programs (i.e., online vs residential)?

Topic: Regression and Prediction

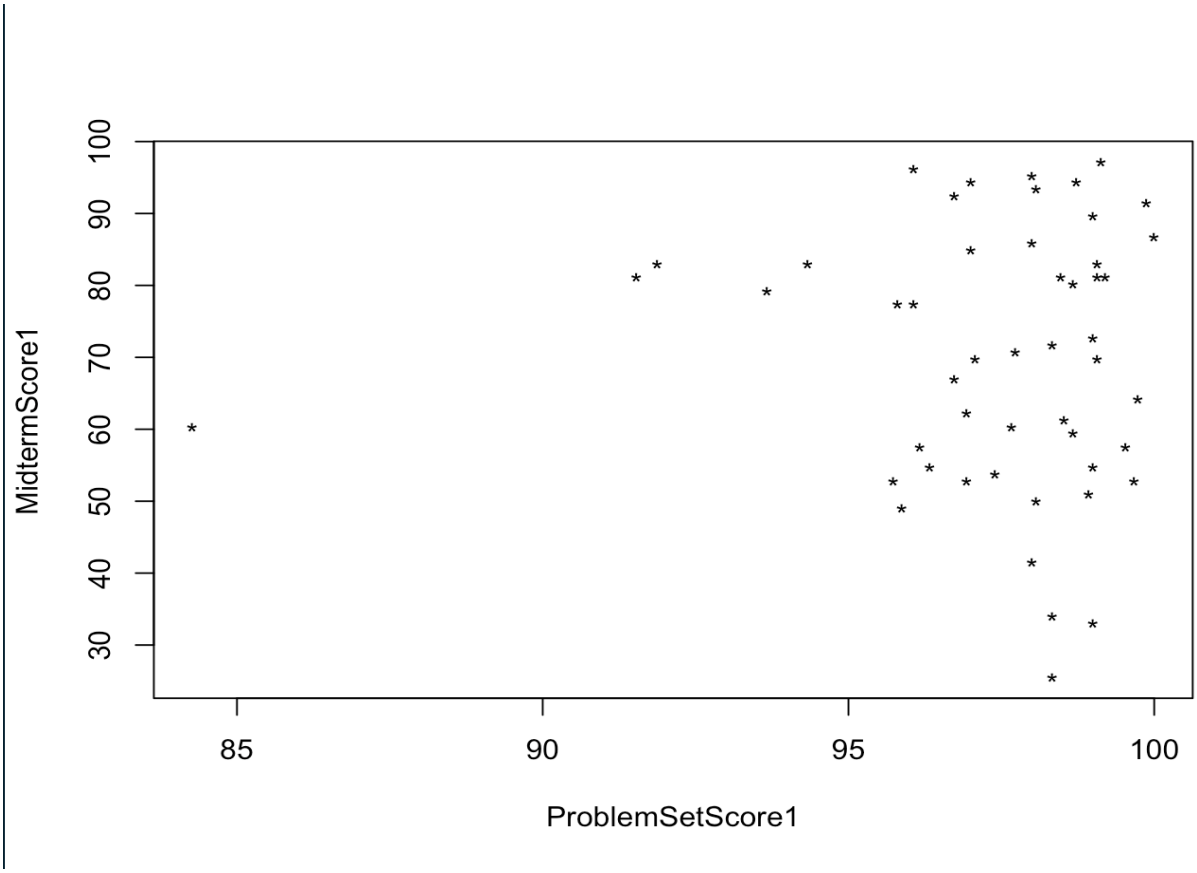
Purpose of the Topic: To check whether we can be able to predict student midterm exam score by using student's assignment grades for both online and residential.

Firstly, we have checked for Class 1.

For Class 1:

R Code:

```
plot(ProblemSetScore1, MidtermScore1, pch="*")
```



From the plot we can be able to observe that the data is less, and we cannot be able to say whether it is correlation high or low.

Next, we have found correlation between problem set score1 and mid-termscore1.

```
124
125 sd(ProblemSetScore1)
126 sd(MidtermScore1)
127 cor(ProblemSetScore1,MidtermScore1)
128
```

128:1 (Untitled) R Script

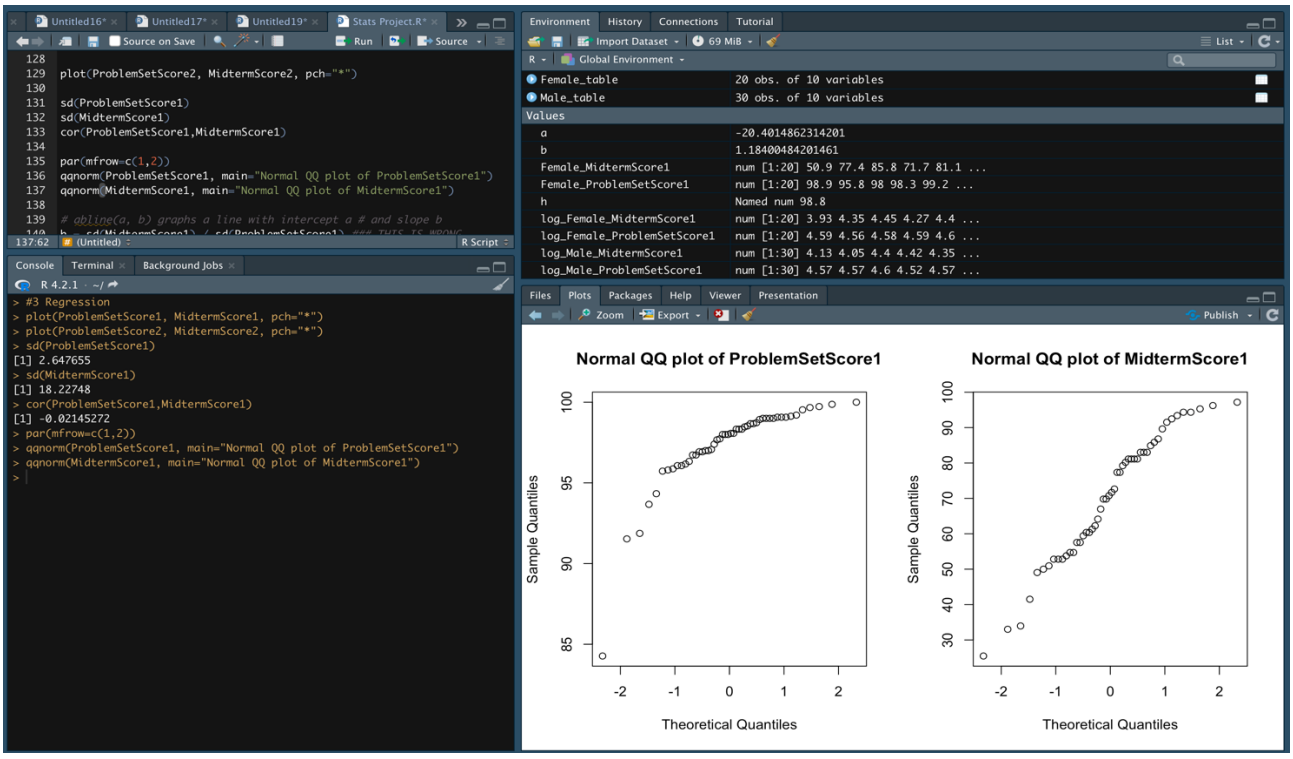
Console Terminal Background Jobs

R 4.2.1 · ~/

```
> sd(ProblemSetScore1)
[1] 2.647655
> sd(MidtermScore1)
[1] 18.22748
> cor(ProblemSetScore1,MidtermScore1)
[1] -0.02145272
>
```

As we can see there is -0.02145272 correlation between them, and we can say that if one value increases then other value decreases as it is a negative correlation.

For checking that the data Is bivariate normal, we have drawn QQ plots:



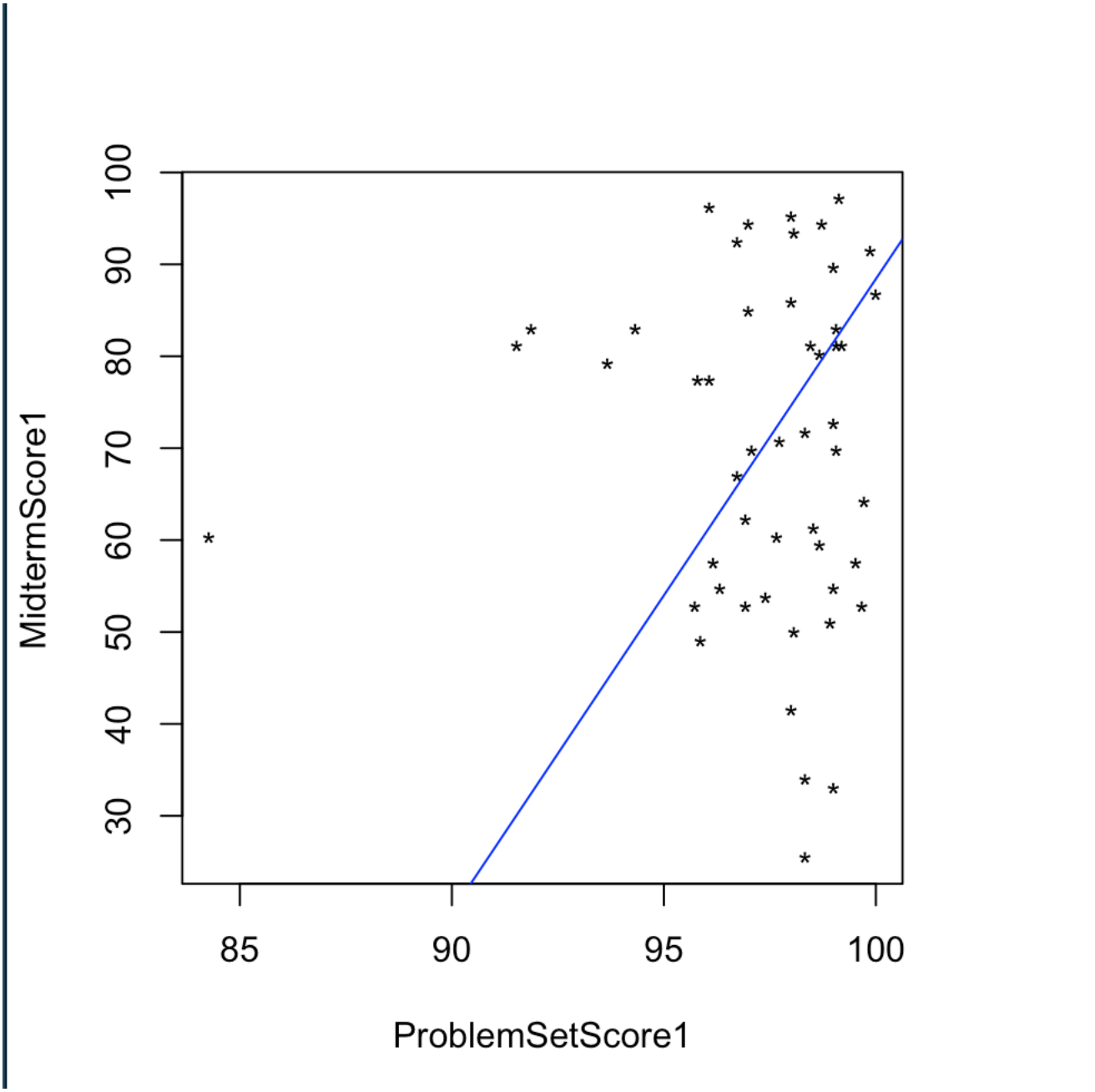
From above plots we can see that the QQ plot for class1 midterm scores looks almost like a straight line and the QQ plot for class1 problem set scores is more like curved line than a straight line. I cannot conclude that the data is very close to bivariate normal.

To find the prediction first we need to find slope so for that first we have done

SD-Line:

R Code:

```r
# abline(a, b) graphs a line with intercept a # and slope b
b = sd(MidtermScore1) / sd(ProblemSetScore1) ### THIS IS WRONG
a = mean(MidtermScore1) - b * mean(ProblemSetScore1)
plot(ProblemSetScore1, MidtermScore1, pch="*")
abline(a, b, col="blue")
```

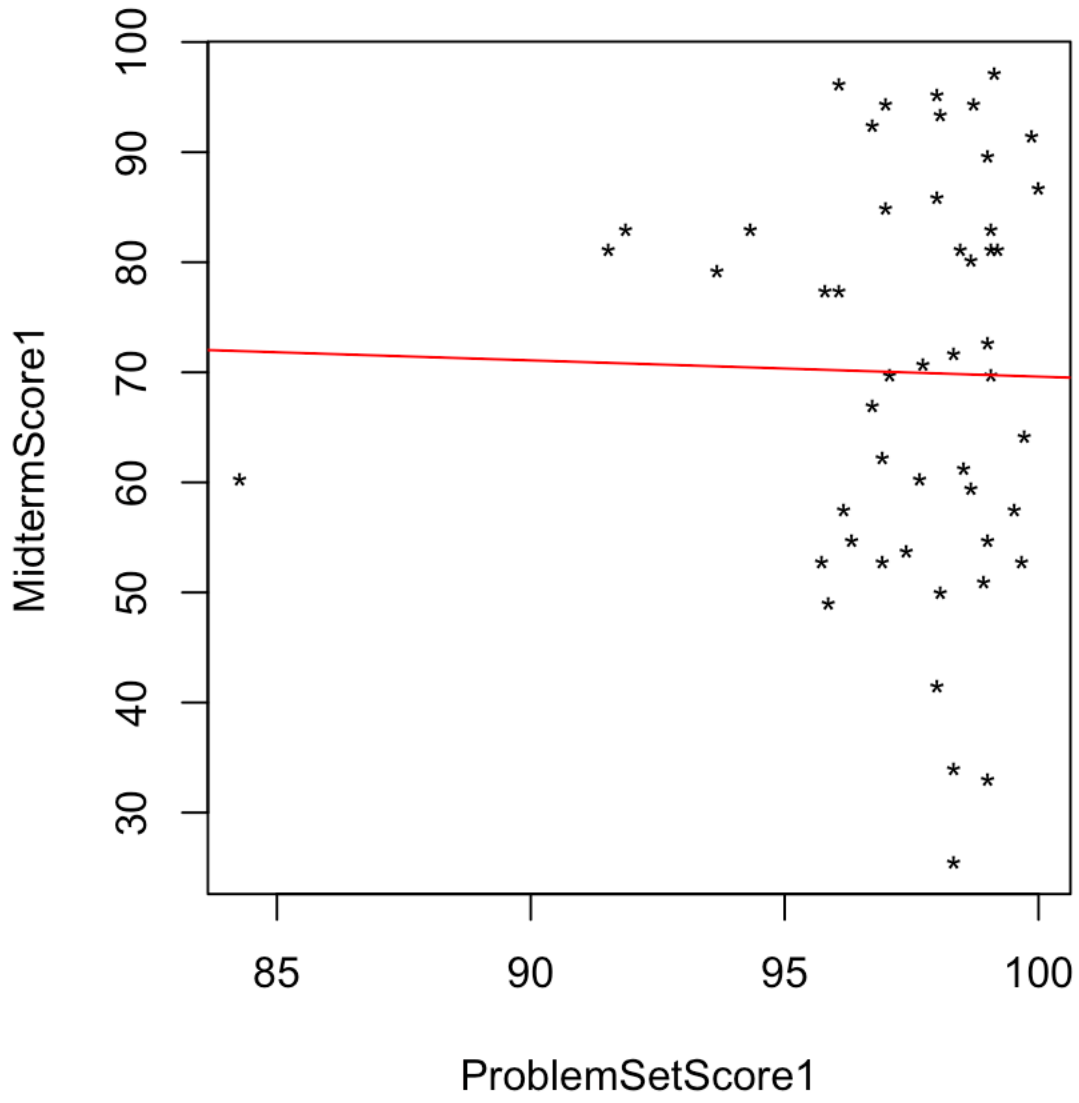


The blue line is called the SD line. It describes the data well but is terrible at prediction. The correct answer is not at all obvious until you do some calculus or linear algebra.

Correlation * SD of weight / SD of height:

R Code:

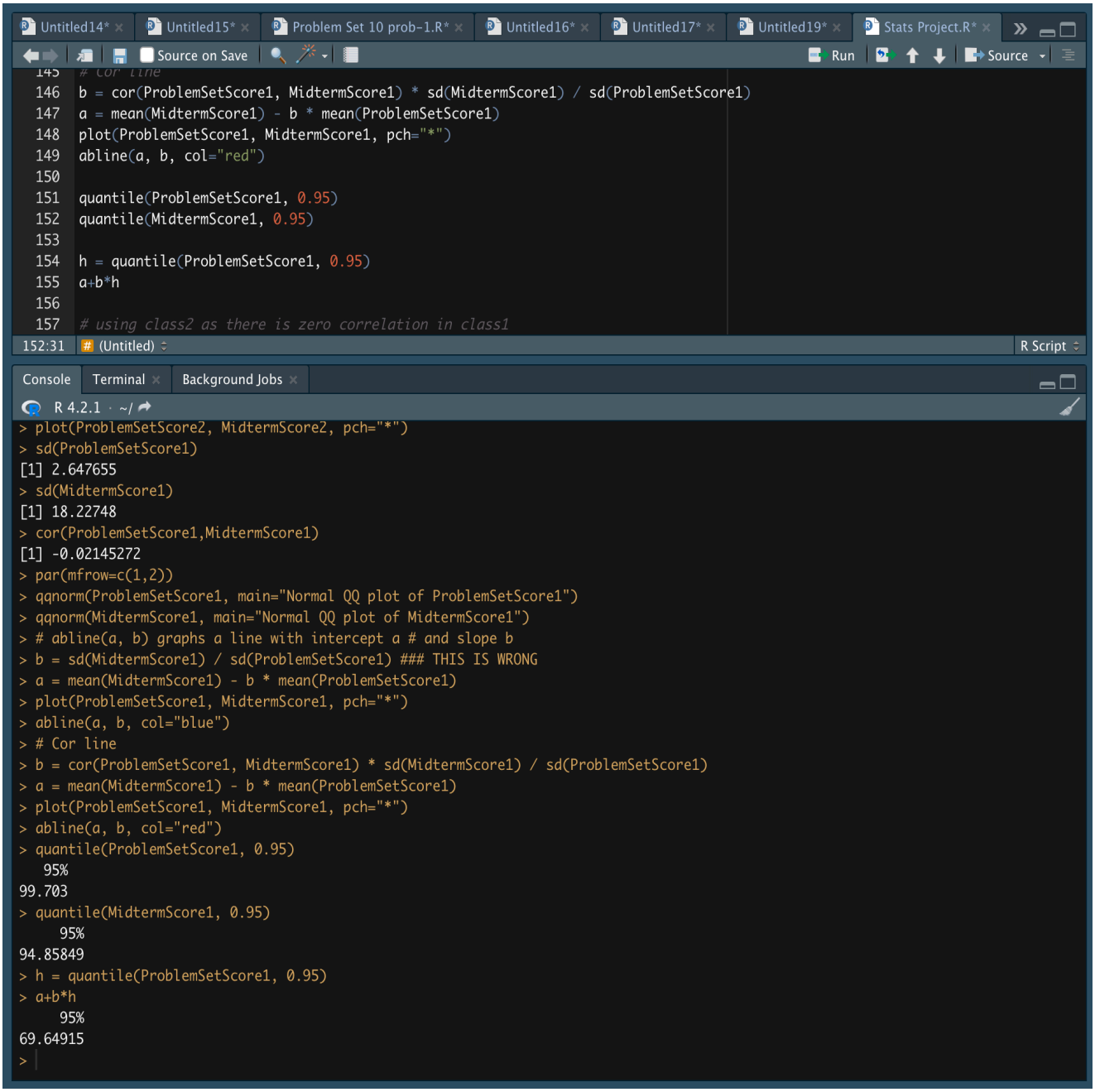
```r
# Cor line
b = cor(ProblemSetScore1, MidtermScore1) * sd(MidtermScore1) / sd(ProblemSetScore1)
a = mean(MidtermScore1) - b * mean(ProblemSetScore1)
plot(ProblemSetScore1, MidtermScore1, pch="*")
abline(a, b, col="red")
```



From finding slope by using correlation is the best way to predict rather than SD Line. It seems that's there is no such linear relationship between problem set score 1 and midterm score1.

Next, we have found Quantile:

R Code:

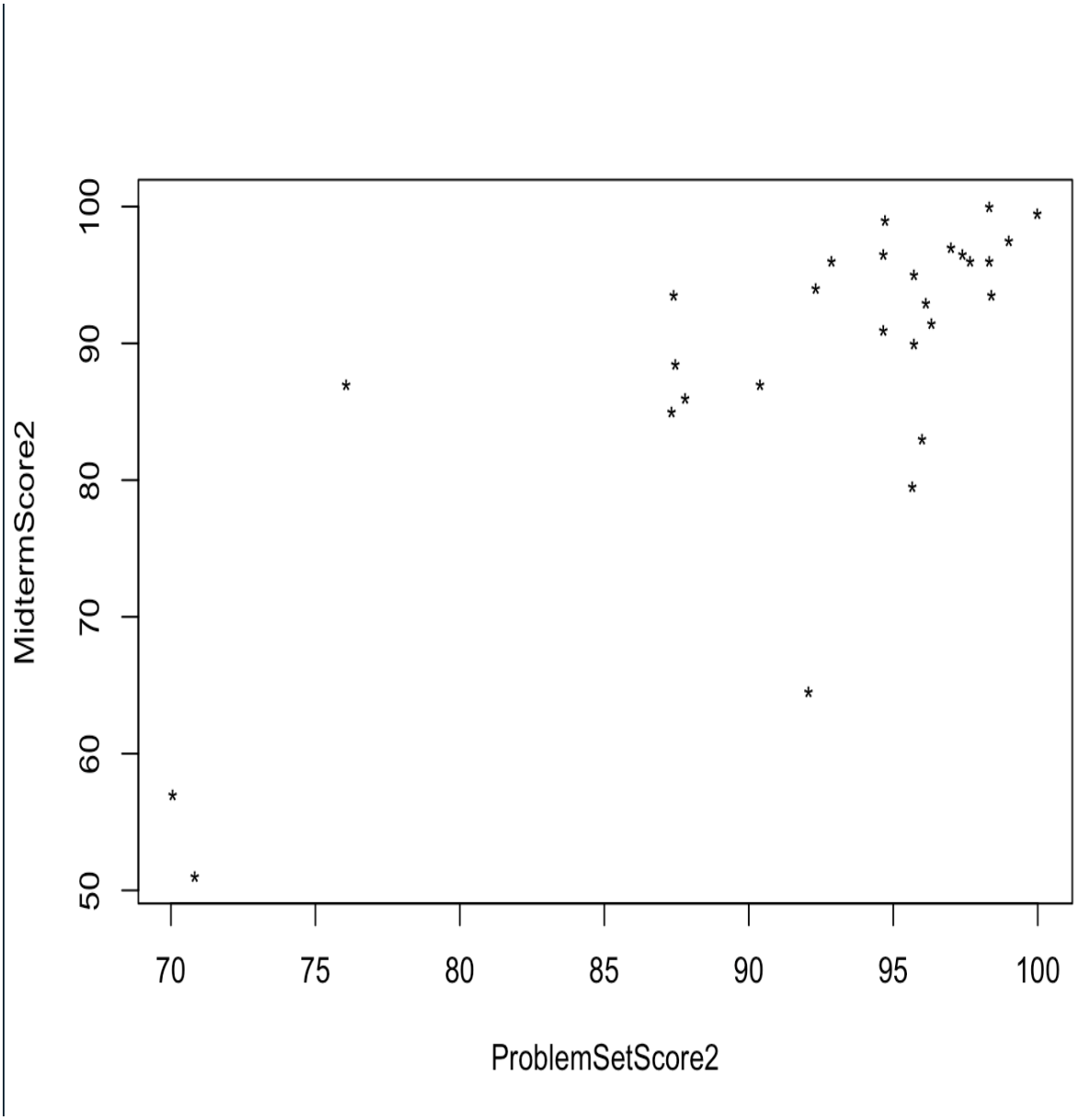


This is much less than the 95th percentile of MidtermScore1! So, we're saying that we do not predict midterm scores from problem set grades for Class1.

For Class 2:

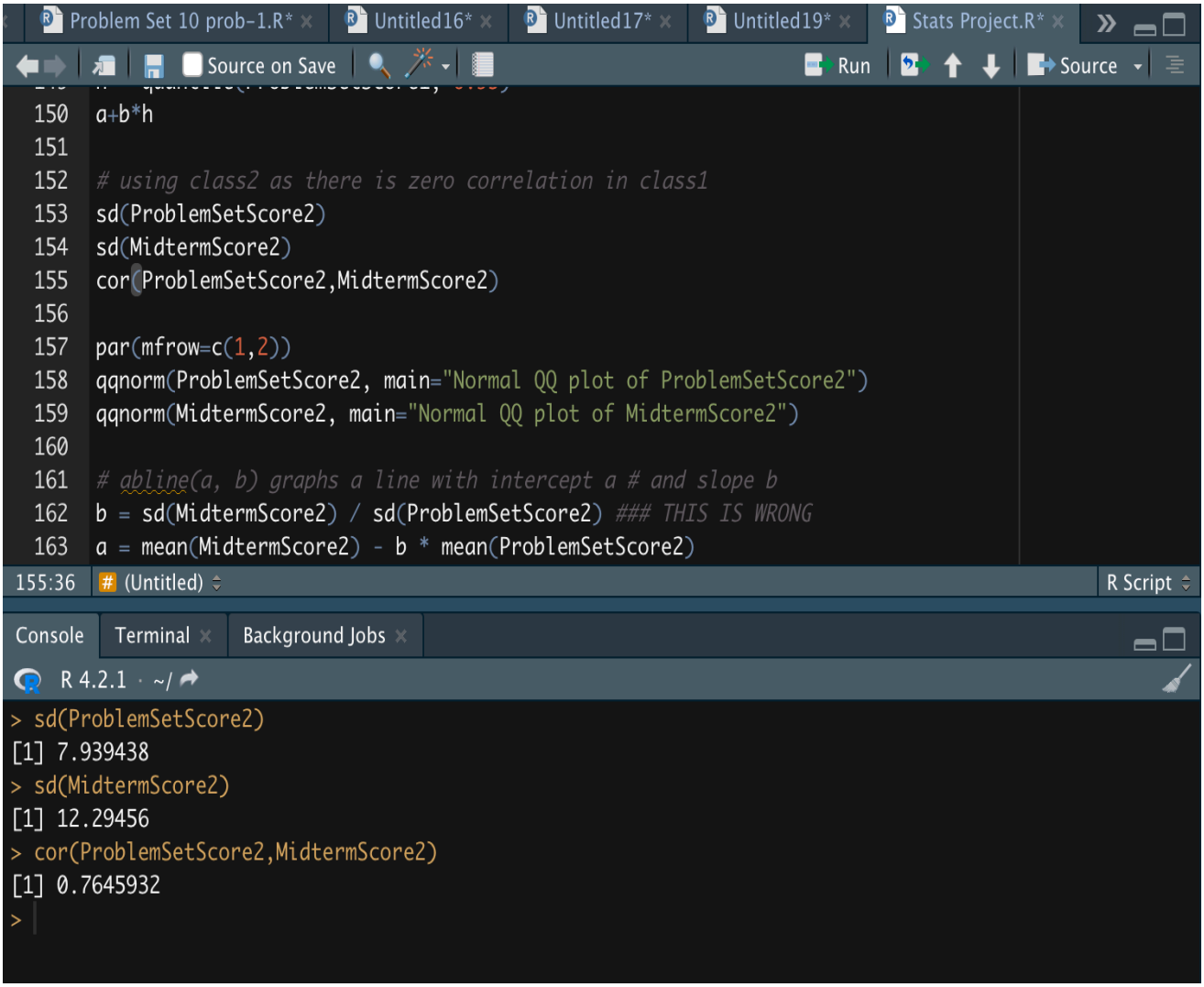
R Code:

```
plot(ProblemSetScore2, MidtermScore2, pch="*")
```



From the plot we can be able to observe that the data is less, and we cannot be able to say whether it is correlation high or low.

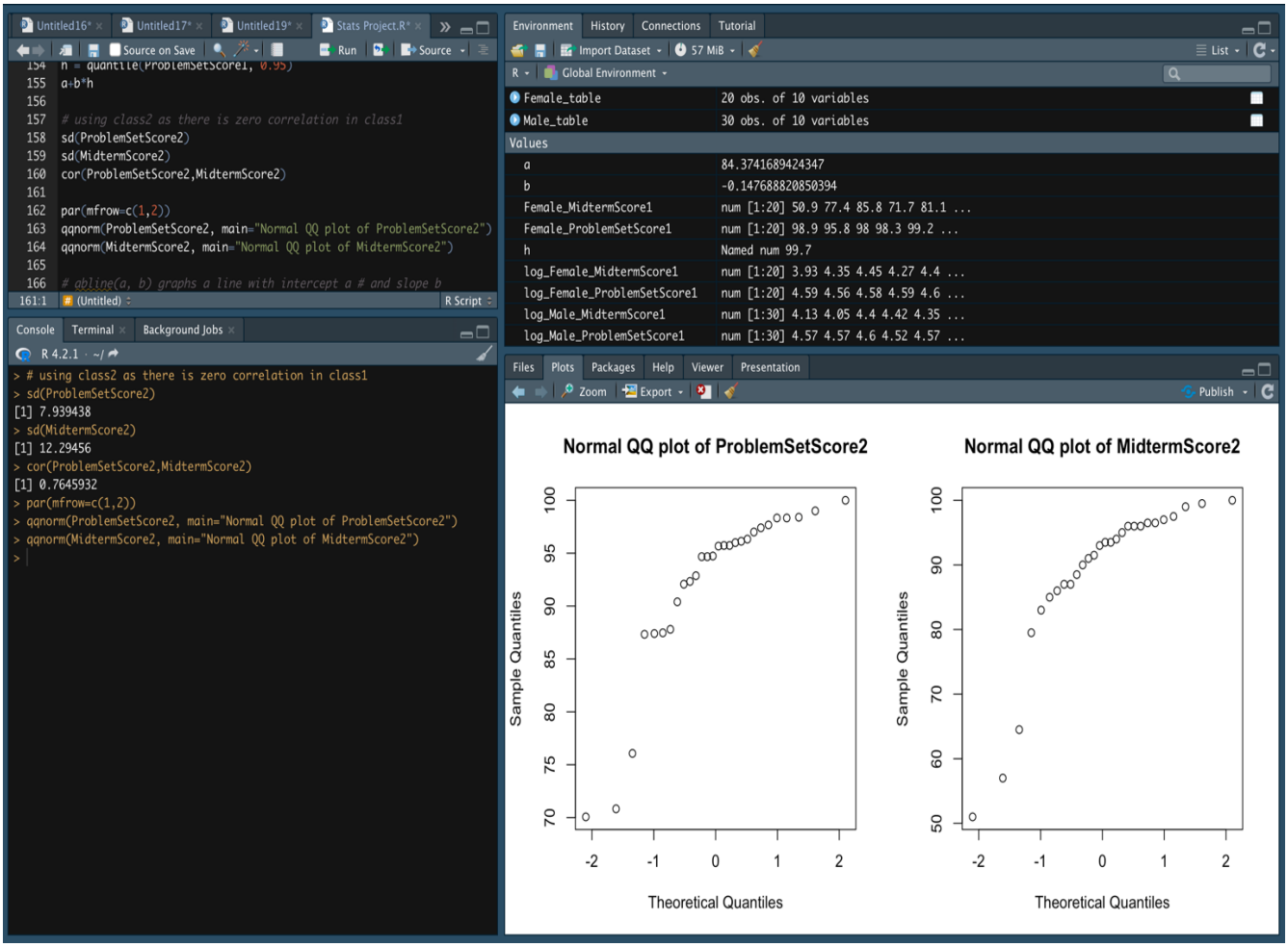
Next, we have found correlation between problem set score2 and mid-termscore2.



As we can see that there is 0.7645932 correlation between Problem set score 2 and midterm score 2, and we can say that it has good correlation between them.

For checking that the data Is bivariate normal, we have drawn QQ plots:

```
par(mfrow=c(1,2))
qqnorm(ProblemSetScore2, main="Normal QQ plot of ProblemSetScore2")
qqnorm(MidtermScore2, main="Normal QQ plot of MidtermScore2")
```



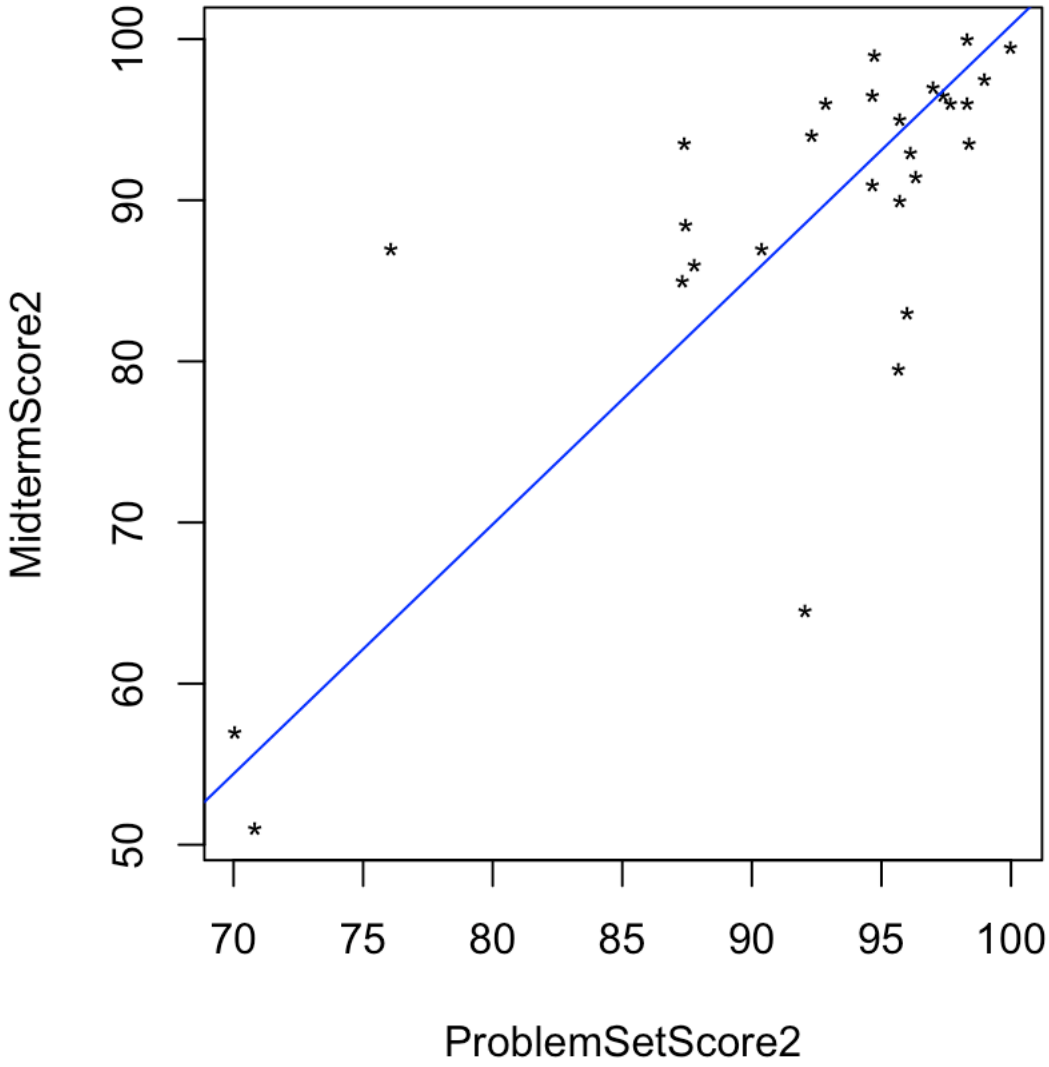
From above plots we can see that the QQ plots for class2 midterm scores and class2 problem set scores are more like curved line than a straight line. We cannot conclude that the data is very close to bivariate normal.

To find the prediction first we need to find slope so for that first we have done

SD-Line:

R Code:

```
# abline(a, b) graphs a line with intercept a # and slope b
b = sd(MidtermScore2) / sd(ProblemSetScore2) ### THIS IS WRONG
a = mean(MidtermScore2) - b * mean(ProblemSetScore2)
plot(ProblemSetScore2, MidtermScore2, pch="*")
abline(a, b, col="blue")
```

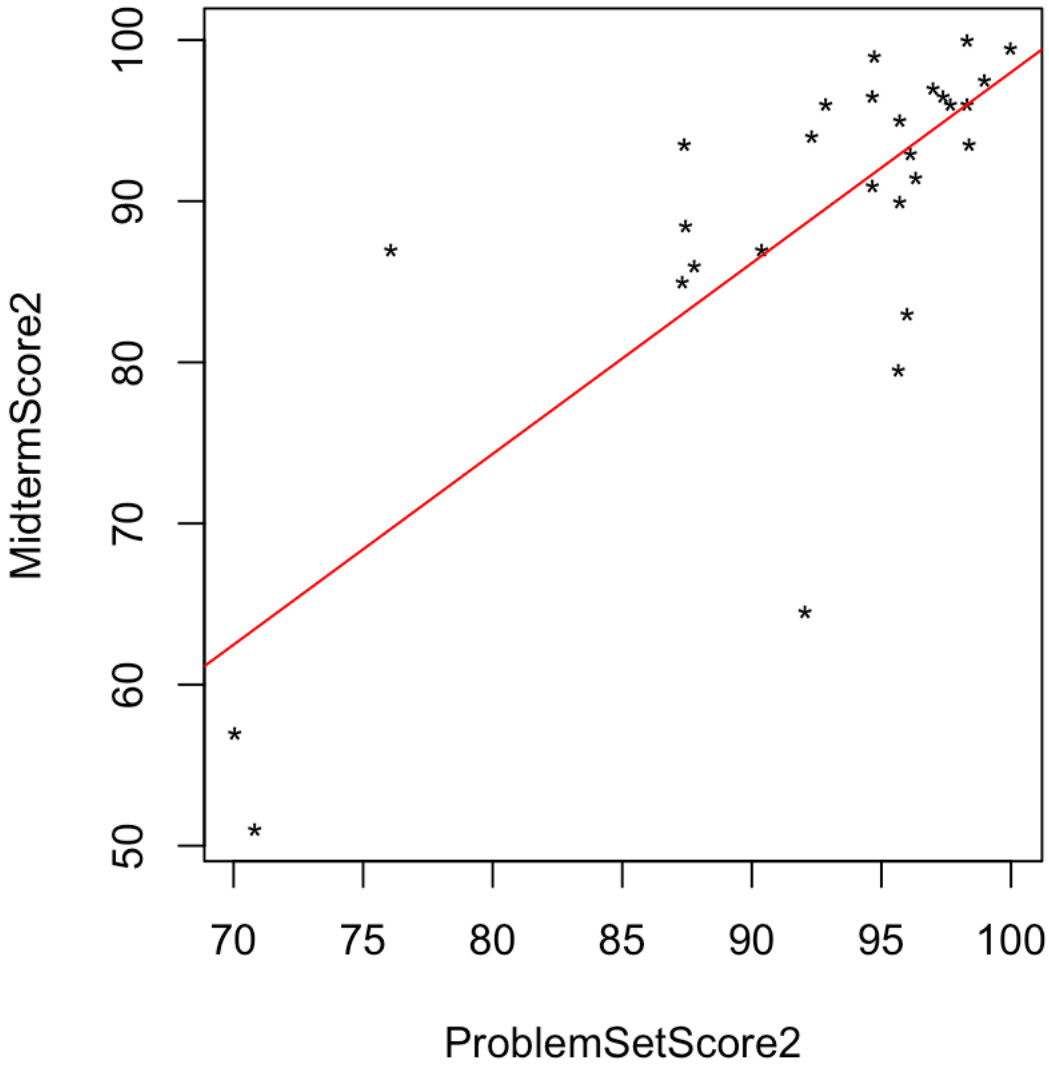


The blue line is called the SD line. It describes the data well but is terrible at prediction. The correct answer is not at all obvious until you do some calculus or linear algebra.

Correlation * SD of weight / SD of height:

R Code:

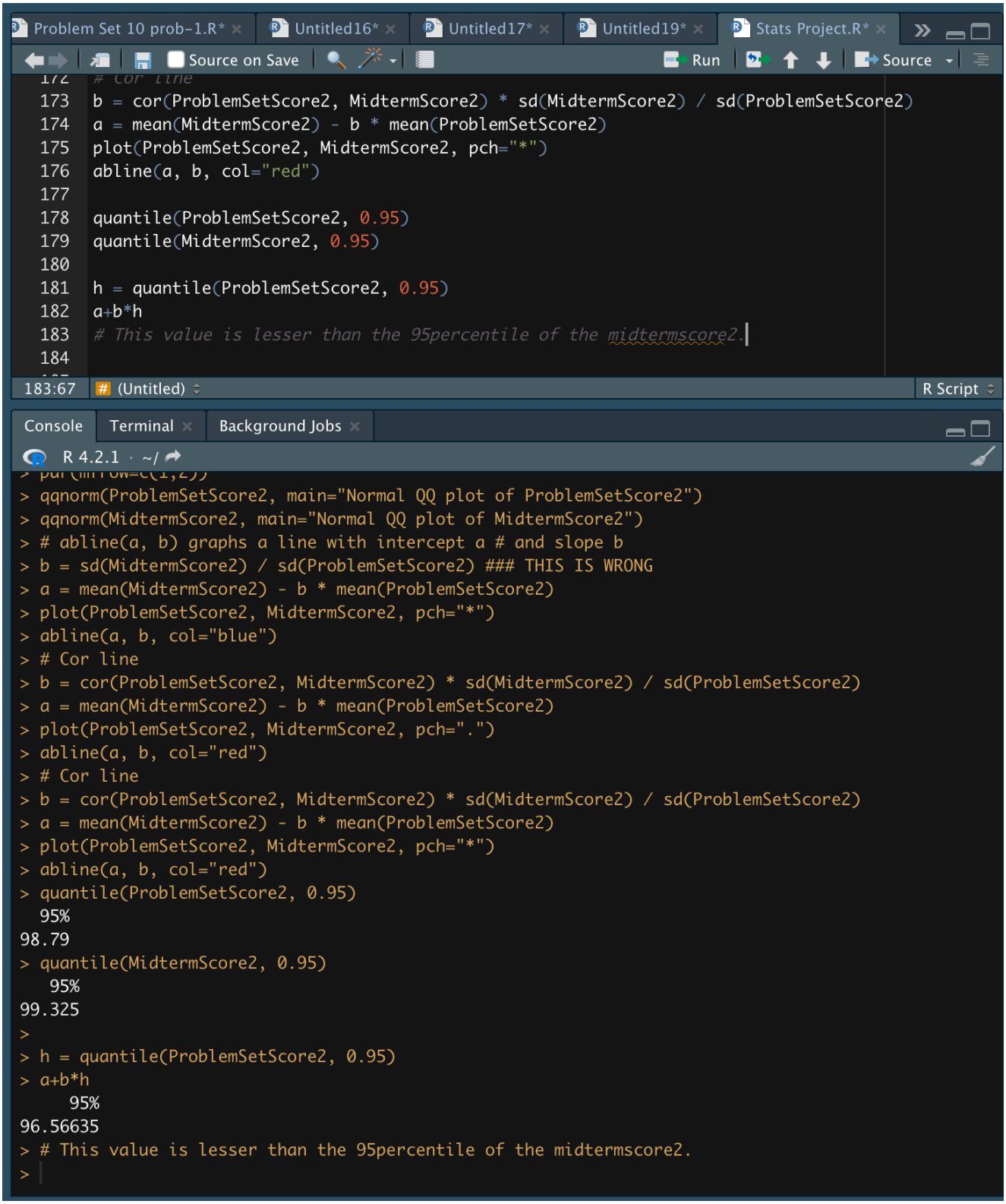
```r
# Cor line
b = cor(ProblemSetScore2, MidtermScore2) * sd(MidtermScore2) / sd(ProblemSetScore2)
a = mean(MidtermScore2) - b * mean(ProblemSetScore2)
plot(ProblemSetScore2, MidtermScore2, pch="*")
abline(a, b, col="red")
```



From finding slope by using correlation is the best way to predict rather than SD Line.

Next, we have found Quantile:

R Code:



We can predict the midterm scores by using problem set scores for Class 2 but I would help more if we have more data.

Total R-Code for 1 , 2 , 3:

```
Class1<- read.csv('/Users/saisrikar/S520 Project Data.csv')

Class1

Class2<-read.csv('/Users/saisrikar/S520 Project Data_1.csv')

Class2

summary(Class1)

summary(Class2)

MidtermScore1=Class1$Midterm.Exam.Score

MidtermScore1=MidtermScore1*100/53

MidtermScore1

MidtermScore2=Class2$Midterm.Exam.Score..percentage.

MidtermScore2

summary(MidtermScore1)

summary(MidtermScore2)

boxplot(MidtermScore1,MidtermScore2,names=c("Class_1","Class_2"),ylab="MidtermS
cores")

plot(density(MidtermScore2))

lines(density(MidtermScore1), col="red")

qqnorm(MidtermScore1,main='Normal Q-Q Plot for Class_1')

qqnorm(MidtermScore2,main='Normal Q-Q Plot for Class_2')

# using log

log_MidtermScore1=log(MidtermScore1)

log_MidtermScore2=log(MidtermScore2)
```

```
qqnorm(log_MidtermScore1,main='Normal Log Q-Q Plot for Class_1')

qqnorm(log_MidtermScore2,main='Normal Log Q-Q Plot for Class_2')

#Welch test

t.test(MidtermScore2,MidtermScore1,alternative ="greater")

##############################################################################
#############

#2 Inferential

Female_table=Class1[Class1$Sex=="Female",]

Female_table

Female_MidtermScore1 = Female_table$Midterm.Exam.Score

Female_MidtermScore1=Female_MidtermScore1*100/53

summary(Female_MidtermScore1)

Female_ProblemSetScore1=Female_table$Problem.Sets.Score..percentage.

summary(Female_ProblemSetScore1)

Male_table=Class1[Class1$Sex=="Male",]

Male_table

Male_MidtermScore1 = Male_table$Midterm.Exam.Score

Male_MidtermScore1=Male_MidtermScore1*100/53

summary(Male_MidtermScore1)

Male_ProblemSetScore1=Male_table$Problem.Sets.Score..percentage.

summary(Male_ProblemSetScore1)

boxplot(Female_MidtermScore1,Male_MidtermScore1,names=c("Female","Male"),ylab
="MidtermScores")

plot(density(Male_MidtermScore1))

lines(density(Female_MidtermScore1), col="red")
```

```
qqnorm(Female_MidtermScore1,main='Normal Q-Q Plot for Female')

qqnorm(Male_MidtermScore1,main='Normal Q-Q Plot for Male')

# using log

log_Female_MidtermScore1=log(Female_MidtermScore1)

log_Male_MidtermScore1=log(Male_MidtermScore1)

qqnorm(log_Female_MidtermScore1,main='Normal Log Q-Q Plot for Female')

qqnorm(log_Male_MidtermScore1,main='Normal Log Q-Q Plot for Male')

#Welch test

t.test(Male_MidtermScore1,Female_MidtermScore1)

# Problemset score male and female

boxplot(Female_ProblemSetScore1,Male_ProblemSetScore1,names=c("Female","Male")
,ylab="ProblemSetScores")

plot(density(Female_ProblemSetScore1))

lines(density(Male_ProblemSetScore1), col="red")

qqnorm(Female_ProblemSetScore1,main='Normal Q-Q Plot for Female')

qqnorm(Male_ProblemSetScore1,main='Normal Q-Q Plot for Male')

# using log

log_Female_ProblemSetScore1=log(Female_ProblemSetScore1)

log_Male_ProblemSetScore1=log(Male_ProblemSetScore1)

qqnorm(log_Female_ProblemSetScore1,main='Normal Log Q-Q Plot for Female')

qqnorm(log_Male_ProblemSetScore1,main='Normal Log Q-Q Plot for Male')

#Welch test

t.test(Male_ProblemSetScore1,Female_ProblemSetScore1)

##########################################################################
#################
```

```
#3 Regression
plot(ProblemSetScore1, MidtermScore1, pch="*")
plot(ProblemSetScore2, MidtermScore2, pch="*")
sd(ProblemSetScore1)
sd(MidtermScore1)
cor(ProblemSetScore1,MidtermScore1)
par(mfrow=c(1,2))
qqnorm(ProblemSetScore1, main="Normal QQ plot of ProblemSetScore1")
qqnorm(MidtermScore1, main="Normal QQ plot of MidtermScore1")
# abline(a, b) graphs a line with intercept a # and slope b
b = sd(MidtermScore1) / sd(ProblemSetScore1) ### THIS IS WRONG
a = mean(MidtermScore1) - b * mean(ProblemSetScore1)
plot(ProblemSetScore1, MidtermScore1, pch="*")
abline(a, b, col="blue")
# Cor line
b = cor(ProblemSetScore1, MidtermScore1) * sd(MidtermScore1) / sd(ProblemSetScore1)
a = mean(MidtermScore1) - b * mean(ProblemSetScore1)
plot(ProblemSetScore1, MidtermScore1, pch="*")
abline(a, b, col="red")
quantile(ProblemSetScore1, 0.95)
quantile(MidtermScore1, 0.95)
h = quantile(ProblemSetScore1, 0.95)
a+b*h
```

```
# using class2 as there is zero correlation in class1
sd(ProblemSetScore2)
sd(MidtermScore2)
cor(ProblemSetScore2,MidtermScore2)
par(mfrow=c(1,2))
qqnorm(ProblemSetScore2, main="Normal QQ plot of ProblemSetScore2")
qqnorm(MidtermScore2, main="Normal QQ plot of MidtermScore2")
# abline(a, b) graphs a line with intercept a # and slope b
b = sd(MidtermScore2) / sd(ProblemSetScore2) ### THIS IS WRONG
a = mean(MidtermScore2) - b * mean(ProblemSetScore2)
plot(ProblemSetScore2, MidtermScore2, pch="*")
abline(a, b, col="blue")
# Cor line
b = cor(ProblemSetScore2, MidtermScore2) * sd(MidtermScore2) / sd(ProblemSetScore2)
a = mean(MidtermScore2) - b * mean(ProblemSetScore2)
plot(ProblemSetScore2, MidtermScore2, pch="*")
abline(a, b, col="red")
quantile(ProblemSetScore2, 0.95)
quantile(MidtermScore2, 0.95)
h = quantile(ProblemSetScore2, 0.95)
a+b*h
# This value is lesser than the 95percentile of the midtermscore2.
```