STAT S-520 DATA ANALYSIS PROJECT

Group No. 3
Sumeet Suvarna
Ameya Parab
Anuja Merwade

Introduction:

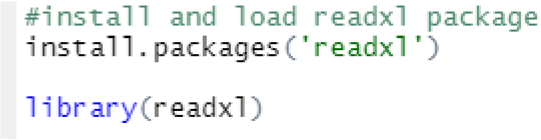
In this data analysis project, we have analyzed the data set provided consisting of STAT S520 Class 1(residential students) and Class 2(online students) with their midterm exam scores. Some of the analysis includes comparison between the performance of both the classes, gender-wise performance, etc. We have incorporated some of the statistical concepts learnt throughout the semester such as hypothesis testing of two-sample problems, analysis of variance (ANOVA), regression analysis, etc. and derived some conclusions which helped us observe the trend and understand the data better.

Analysis of given data and pre-processing:

The "Class 1" sheet in the given Excel file consists of granular data (up to scores of students in each question) of the midterm scores of Class 1 having 50 students as compared to "Class 2" sheet that has data of 28 students. We perform some preprocessing on Class 1 data to have the same variable to perform analysis against in the Inferential Statistics section. Also, the count of Class 1 students is higher than that of Class 2 students but overall, we have a quite small dataset.

Installing excel package in R

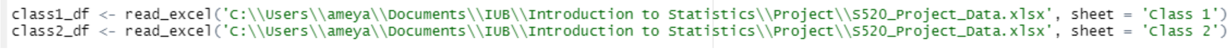
(https://www.statology.org/import-excel-into-r/#:~:text=The%20easiest%20way%20to%20import,function%20from%20the%20readxl%20package.&text=where%3A,sheet%3A%20The%20sheet%20to%20read)



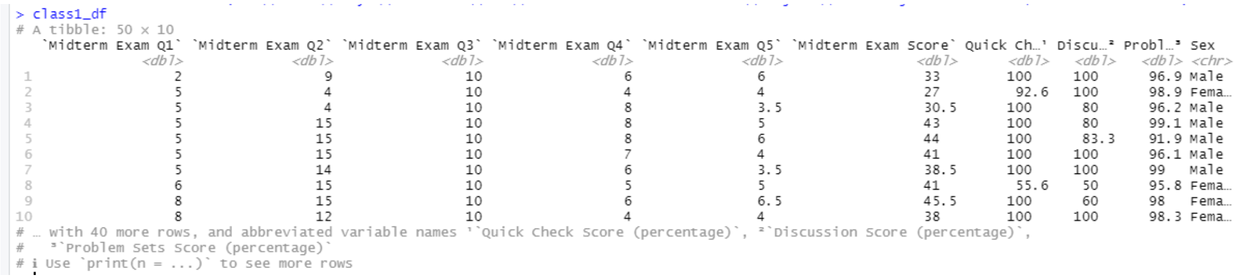
```
#install and load readxl package
install.packages('readxl')

library(readxl)
```

Creating dataframes

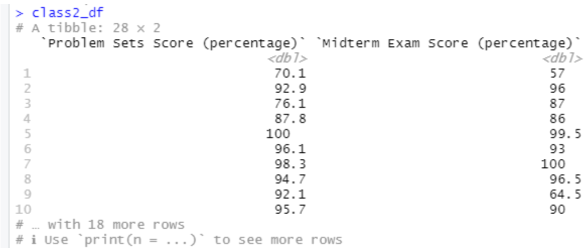
(https://www.statology.org/r-add-a-column-to-dataframe/)



```
class1_df <- read_excel('C:\\Users\\ameya\\Documents\\IUB\\Introduction to Statistics\\Project\\S520_Project_Data.xlsx', sheet = 'Class 1')
class2_df <- read_excel('C:\\Users\\ameya\\Documents\\IUB\\Introduction to Statistics\\Project\\S520_Project_Data.xlsx', sheet = 'Class 2')
```

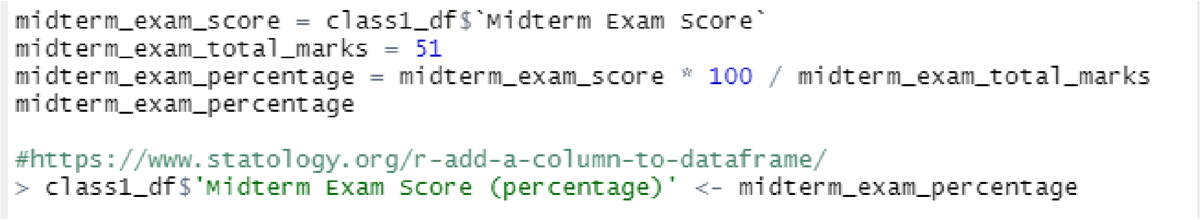


```
> class1_df
# A tibble: 50 × 10
   `Midterm Exam Q1` `Midterm Exam Q2` `Midterm Exam Q3` `Midterm Exam Q4` `Midterm Exam Q5` `Midterm Exam Score` Quick Ch…¹ Discu…² Probl…³ Sex  
               <dbl>             <dbl>             <dbl>             <dbl>             <dbl>                <dbl>      <dbl>   <dbl>   <dbl> <chr>
 1                 2                 9                10                 6               6                    33       100     100      96.9 Male 
 2                 5                 4                10                 4               4                    27        92.6   100      98.9 Fema…
 3                 5                 4                10                 8               3.5                  30.5     100      80      96.2 Male 
 4                 5                15                10                 8               5                    43       100      80      99.1 Male 
 5                 5                15                10                 8               6                    44       100      83.3    91.9 Male 
 6                 5                15                10                 7               4                    41       100     100      96.1 Male 
 7                 5                14                10                 6               3.5                  38.5     100     100      99   Male 
 8                 6                15                10                 5               5                    41        55.6    50      95.8 Fema…
 9                 8                15                10                 6               6.5                  45.5     100      60      98   Fema…
10                 8                12                10                 4               4                    38       100     100      98.3 Fema…
# … with 40 more rows, and abbreviated variable names ¹`Quick Check Score (percentage)`, ²`Discussion Score (percentage)`,
#   ³`Problem Sets Score (percentage)`
# ℹ Use `print(n = ...)` to see more rows
```



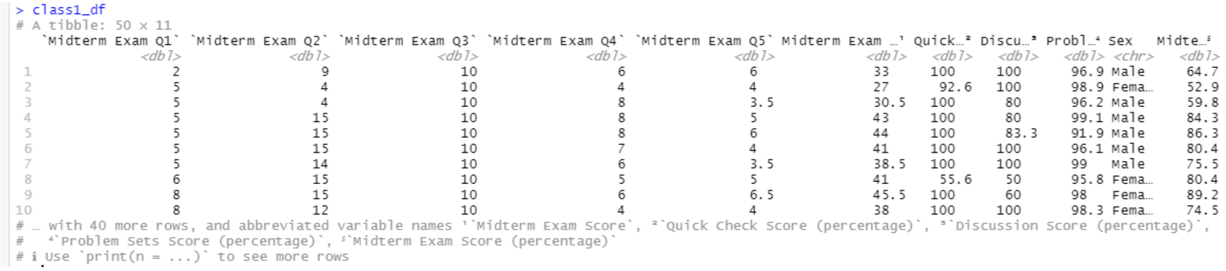
```
> class2_df
# A tibble: 28 × 2
   `Problem Sets Score (percentage)` `Midterm Exam Score (percentage)`
                               <dbl>                              <dbl>
 1                              70.1                               57  
 2                              92.9                               96  
 3                              76.1                               87  
 4                              87.8                               86  
 5                             100                                 99.5
 6                              96.1                               93  
 7                              98.3                              100  
 8                              94.7                               96.5
 9                              92.1                               64.5
10                              95.7                               90  
# … with 18 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Conversion of midterm exam scores to percentage



```
midterm_exam_score = class1_df$`Midterm Exam Score`
midterm_exam_total_marks = 51
midterm_exam_percentage = midterm_exam_score * 100 / midterm_exam_total_marks
midterm_exam_percentage

#https://www.statology.org/r-add-a-column-to-dataframe/
> class1_df$'Midterm Exam Score (percentage)' <- midterm_exam_percentage
```

Viewing Data-frame after pre-processing

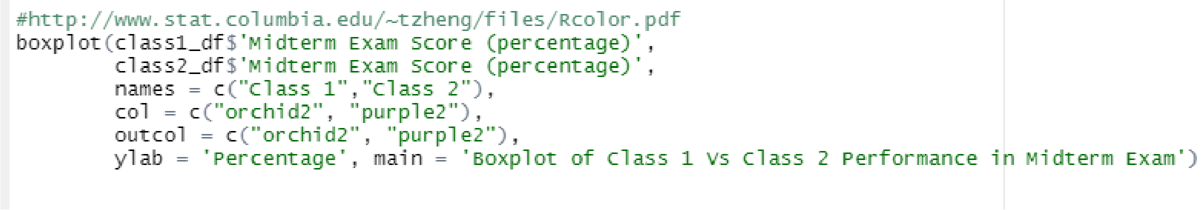


```
> class1_df
# A tibble: 50 × 11
   `Midterm Exam Q1` `Midterm Exam Q2` `Midterm Exam Q3` `Midterm Exam Q4` `Midterm Exam Q5` Midterm Exam …¹ Quick…² Discu…³ Probl…⁴ Sex   Midte…⁵
               <dbl>             <dbl>             <dbl>             <dbl>             <dbl>           <dbl>   <dbl>   <dbl>   <dbl> <chr>   <dbl>
 1                 2                 9                10                 6               6              33     100     100      96.9 Male     64.7
 2                 5                 4                10                 4               4              27      92.6   100      98.9 Fema…    52.9
 3                 5                 4                10                 8               3.5            30.5   100      80      96.2 Male     59.8
 4                 5                15                10                 8               5              43     100      80      99.1 Male     84.3
 5                 5                15                10                 8               6              44     100      83.3    91.9 Male     86.3
 6                 5                15                10                 7               4              41     100     100      96.1 Male     80.4
 7                 5                14                10                 6               3.5            38.5   100     100      99   Male     75.5
 8                 6                15                10                 5               5              41      55.6    50      95.8 Fema…    80.4
 9                 8                15                10                 6               6.5            45.5   100      60      98   Fema…    89.2
10                 8                12                10                 4               4              38     100     100      98.3 Fema…    74.5
# … with 40 more rows, and abbreviated variable names ¹`Midterm Exam Score`, ²`Quick Check Score (percentage)`, ³`Discussion Score (percentage)`,
#   ⁴`Problem Sets Score (percentage)`, ⁵`Midterm Exam Score (percentage)`
# ℹ Use `print(n = ...)` to see more rows
```

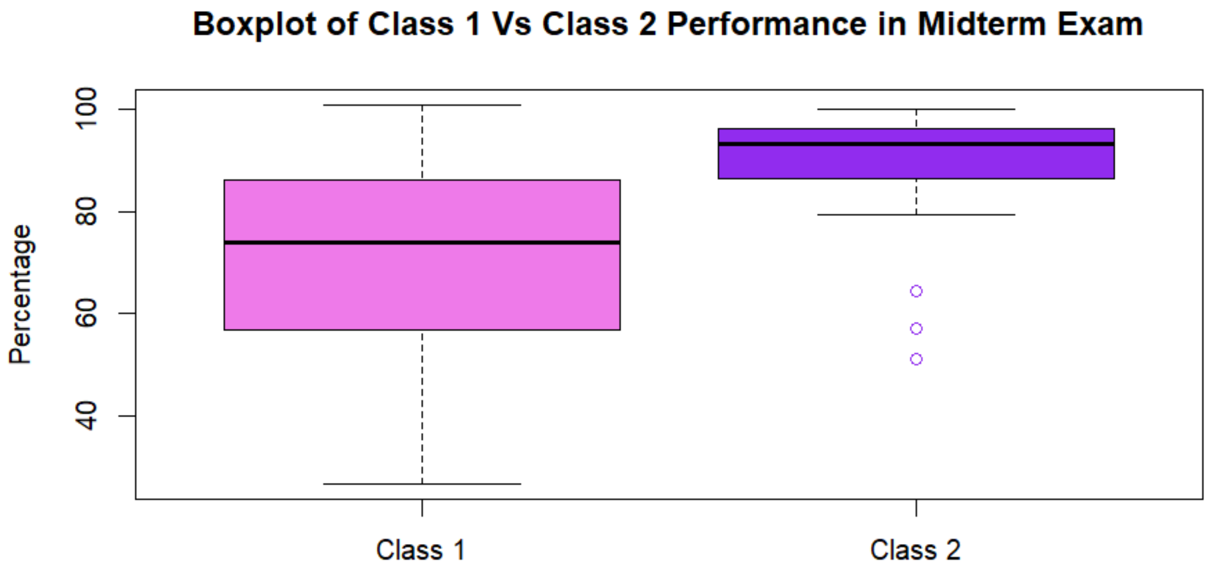
Descriptive analysis:

Plotting a Boxplot in R

(http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf)



```
#http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf
boxplot(class1_df$'Midterm Exam Score (percentage)',
        class2_df$'Midterm Exam Score (percentage)',
        names = c("Class 1","Class 2"),
        col = c("orchid2", "purple2"),
        outcol = c("orchid2", "purple2"),
        ylab = 'Percentage', main = 'Boxplot of Class 1 Vs Class 2 Performance in Midterm Exam')
```



Here, we have compared the performance of Class 1 and Class 2 in the midterm exams based on their percentage. As observed in the boxplot, the entire range in which Class 1 scored is between 58% and 85% and Class 2 scored between 85% and 95%. The medians of midterm percentages of Class 1 and Class 2 are approximately 78% and 92% respectively. Overall, we can observe that Class 2 has performed better than Class 1 in the midterm exam.

Inferential Statistics:

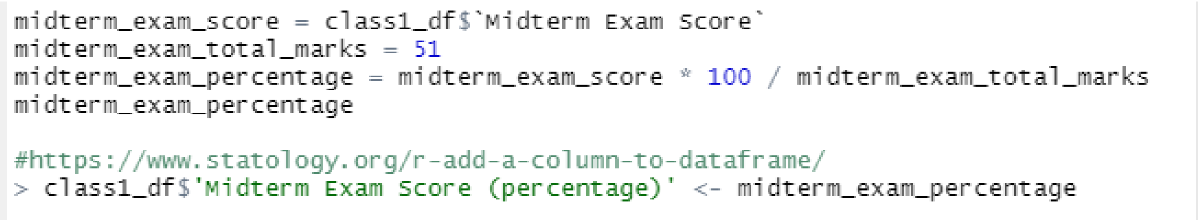
Below we have answered a question that can help us reach to some inferences.

Q. Do students of Class 2 get higher exam grades than students of Class 1 in S520?

The experimental unit is students, and they were drawn from two populations of students namely Class 1 and Class 2. Hence, this problem is a two-sample problem. The measurement taken on students is their midterm exam scores.

As the data values for midterm exam score of Class 1 students are in marks and Class 2 students are in percentage, we need to convert the Class 1 students marks in percentage which would bring the comparison parameter of midterm exam score column in the same format.

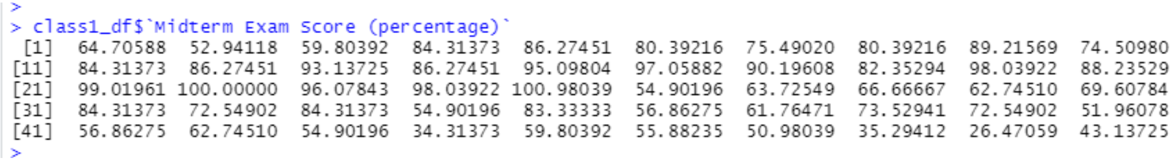
Midterm score preprocessing



```
midterm_exam_score = class1_df$`Midterm Exam Score`
midterm_exam_total_marks = 51
midterm_exam_percentage = midterm_exam_score * 100 / midterm_exam_total_marks
midterm_exam_percentage

#https://www.statology.org/r-add-a-column-to-dataframe/
> class1_df$'Midterm Exam Score (percentage)' <- midterm_exam_percentage
```

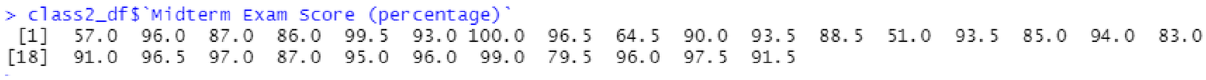
midterm exam score in percentage for Class 1 and Class 2 students are as follows:

> Class 1_df$'midterm exam score (percentage)'



```
>
> class1_df$`Midterm Exam Score (percentage)`
 [1]  64.70588  52.94118  59.80392  84.31373  86.27451  80.39216  75.49020  80.39216  89.21569  74.50980
[11]  84.31373  86.27451  93.13725  86.27451  95.09804  97.05882  90.19608  82.35294  98.03922  88.23529
[21]  99.01961 100.00000  96.07843  98.03922 100.98039  54.90196  63.72549  66.66667  62.74510  69.60784
[31]  84.31373  72.54902  84.31373  54.90196  83.33333  56.86275  61.76471  73.52941  72.54902  51.96078
[41]  56.86275  62.74510  54.90196  34.31373  59.80392  55.88235  50.98039  35.29412  26.47059  43.13725
>
```

> Class2_df$'midterm exam score (percentage)'



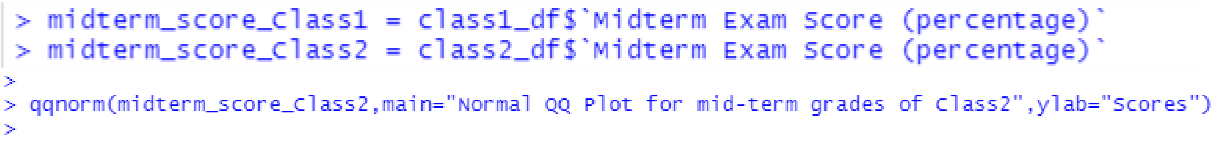
```
> class2_df$`Midterm Exam Score (percentage)`
 [1]  57.0  96.0  87.0  86.0  99.5  93.0 100.0  96.5  64.5  90.0  93.5  88.5  51.0  93.5  85.0  94.0  83.0
[18]  91.0  96.5  97.0  87.0  95.0  96.0  99.0  79.5  96.0  97.5  91.5
```

Let X_i denote the midterm exam score(percentage) of student i in the Class 1 sample and let X_j denote the midterm exam score(percentage) of student j in the Class 2 sample.

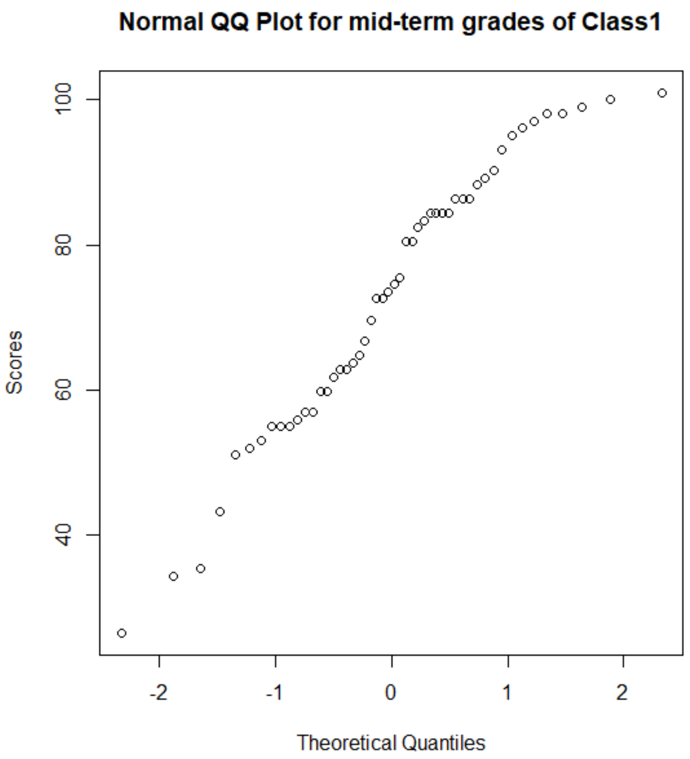
For this 2-sample problem, the parameter of interest is $\Delta = \mu_1 - \mu_2$ where $\mu_1 = EX_i$ and $\mu_2 = EX_j$.

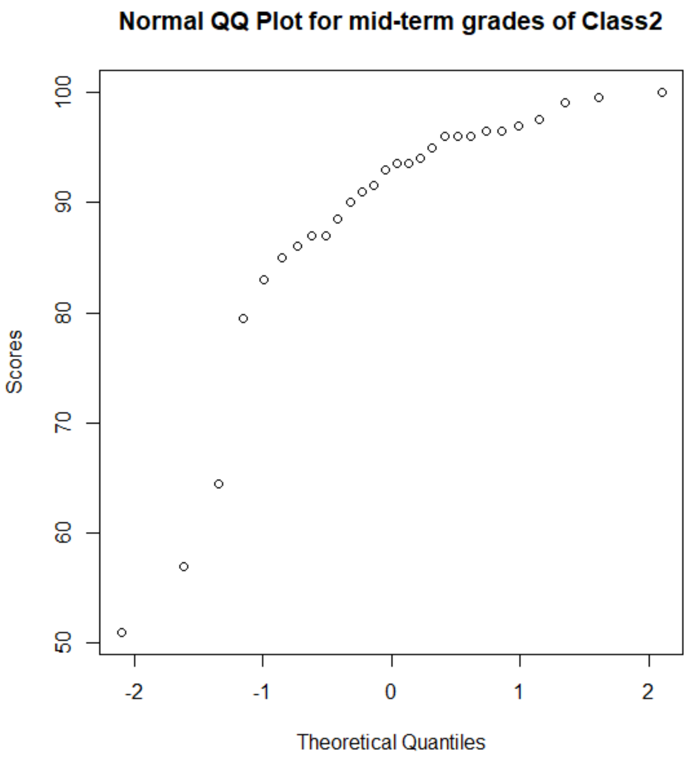
Now, let's check normality assumption for these samples.

Plotting Q-Q plot for midterm grades of Class 1 and Class 2 students.



```
> midterm_score_Class1 = class1_df$`Midterm Exam Score (percentage)`
> midterm_score_Class2 = class2_df$`Midterm Exam Score (percentage)`
>
> qqnorm(midterm_score_Class2,main="Normal QQ Plot for mid-term grades of Class2",ylab="Scores")
>
```

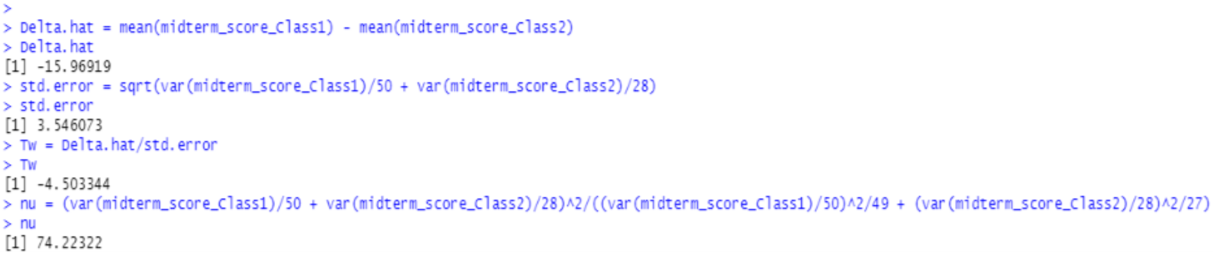

Normal QQ Plot for mid-term grades of Class1
Scores
100
80
60
40
-2
-1
0
1
2
Theoretical Quantiles


Normal QQ Plot for mid-term grades of Class2
Scores
100
90
80
70
60
50
-2
-1
0
1
2
Theoretical Quantiles

The Q-Q plot of midterm grades of Class 1 students seem normal as it is close to straight line. But this is not the case with the Q-Q plot of midterm grades of Class 2 students where Q-Q plot is slightly bent upwards with few outliers. For a small sample size, the normal population might give a straight Q-Q plot or curved Q-Q plot while same can be true for the skewed population, so the Q-Q plot won't give us the definite answer as the sample size of the students is too small to be confident about the normality assumption.
So, let's assume that both the population sample is normal for defining the hypothesis.
Let's consider that students of Class 1 perform well in midterm exam as compared to Class 2 students.
As, now we are considering that the Class 1 students would perform better and score higher grades in midterm exam as compared to Class 2 students, we will need compelling evidence to prove the theory that Class 2 students score higher grades in exam.
Hence, we formulate the alternative hypothesis that Class 2 students score higher grades in exam than the Class 1 students.
Therefore, the theory is $\mu_1 < \mu_2$.

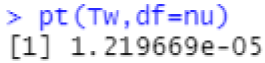
Let Δ be the population mean of Class 1 student's midterm exam grades minus the population mean of Class 2 student's midterm exam grades.
$\Delta = \mu_1 - \mu_2$
Accordingly, the null hypothesis will be formulated as follows:
$H_0 : \Delta \geq 0$
And alternate hypothesis would be: $H_1 : \Delta < 0$

We test the hypothesis using Welch's approximate t-test.



```
>
> Delta.hat = mean(midterm_score_Class1) - mean(midterm_score_Class2)
> Delta.hat
[1] -15.96919
> std.error = sqrt(var(midterm_score_Class1)/50 + var(midterm_score_Class2)/28)
> std.error
[1] 3.546073
> Tw = Delta.hat/std.error
> Tw
[1] -4.503344
> nu = (var(midterm_score_Class1)/50 + var(midterm_score_Class2)/28)^2/((var(midterm_score_Class1)/50)^2/49 + (var(midterm_score_Class2)/28)^2/27)
> nu
[1] 74.22322
```

To find the p-value of left tailed test, we use below formula:



```
> pt(Tw,df=nu)
[1] 1.219669e-05
```

As, the P-value = $1.2196 * 10^{-5}$ is less than $\propto = 0.05$ so we reject the null hypothesis. And we can conclude that the Class 1 students doesn't have higher midterm grades than the Class 2 students which in turn means that performance of students in Class 2 was better than the performance of students in Class 1 in the midterm exam.

An approximate 95% confidence interval for the difference in the population mean of midterm exam grades between Class 1 and Class 2 students can be calculated using below

$$\hat{\Delta} \pm q_t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

= -15.96919 ± 1.992444 * 3.546073

= (-8.903838 , -23.03454)

Conclusion: The performance of students in Class 2 was better than the performance of students in Class 1 in the midterm exam by approximately 8% to 23%.

Q. Does gender make a difference in student's performance on assignments?

The percentage distribution of marks obtained by male and female students in the problem sets assignments for Class 1 and Class 2 were measured. We want to analyze whether the two classes have similar average performance in the problem set assignments.

Since it is a two-sample problem, the experimental units are the students with two samples belonging to Class 1 and Class 2. The samples have unequal sample size with the measurement in consideration being the problem set score.

Let the population mean percentage of problem set score for Class 1 be $\mu_1$and that for Class 2 be μ_2.

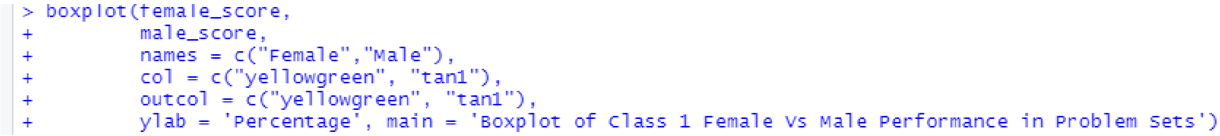
As per the research question the null hypothesis states that the two classes have same mean score

$H_0: \mu_1 = \mu_2$

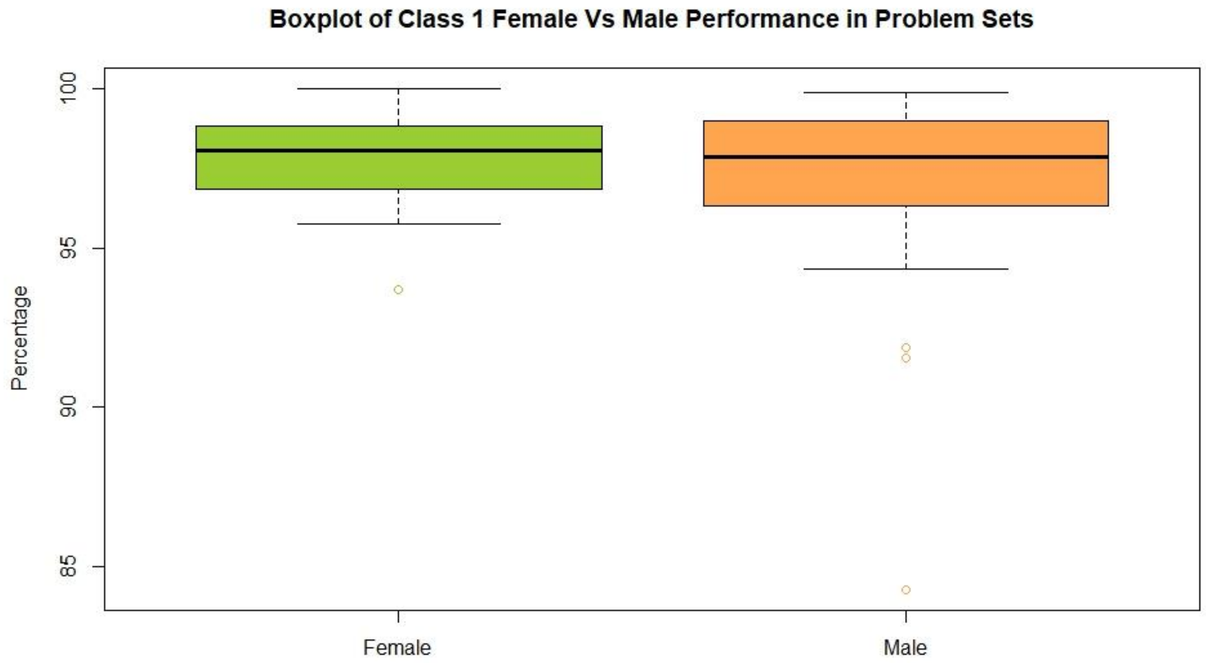
And the alternative hypothesis states that the means differ

$H_1: \mu_1 \neq \mu_2$

Plotting Boxplot of the two samples using R:

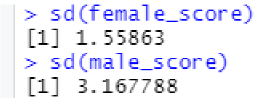


```
> boxplot(female_score,
+         male_score,
+         names = c("Female","Male"),
+         col = c("yellowgreen", "tan1"),
+         outcol = c("yellowgreen", "tan1"),
+         ylab = 'Percentage', main = 'Boxplot of Class 1 Female Vs Male Performance in Problem Sets')
```



From the boxplots, we can observe that the mean of both the population seem to be equal.

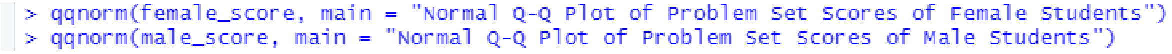
Checking the standard deviation of both the populations:



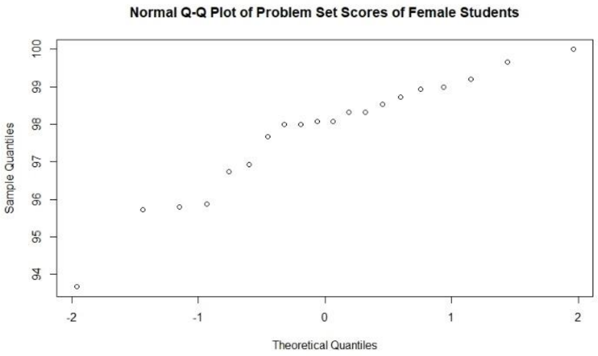
```
> sd(female_score)
[1] 1.55863
> sd(male_score)
[1] 3.167788
```

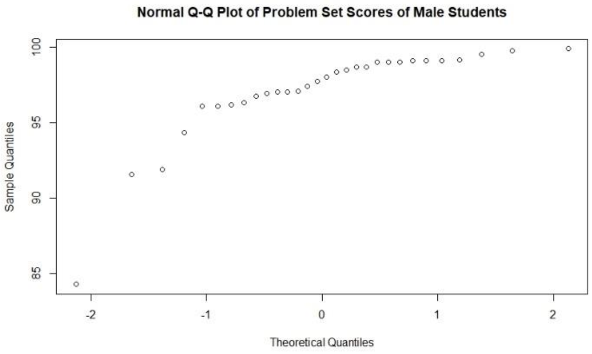
There is a major difference in the standard deviation of male and female scores. Hence, the assumption of Homoscedasticity of data doesn't seem true.

Checking Normality of the data:



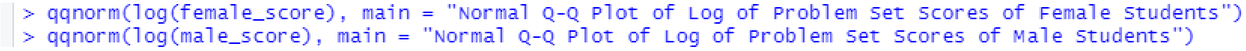
```
> qqnorm(female_score, main = "Normal Q-Q Plot of Problem Set Scores of Female Students")
> qqnorm(male_score, main = "Normal Q-Q Plot of Problem Set Scores of Male Students")
```



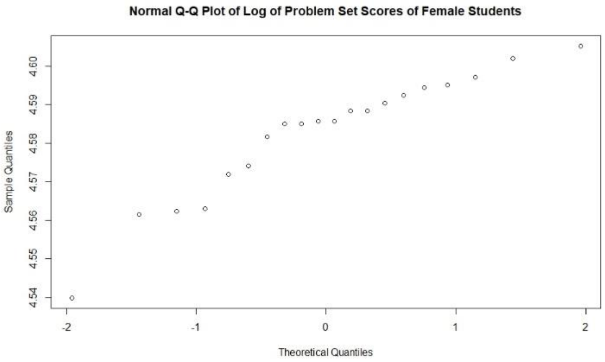


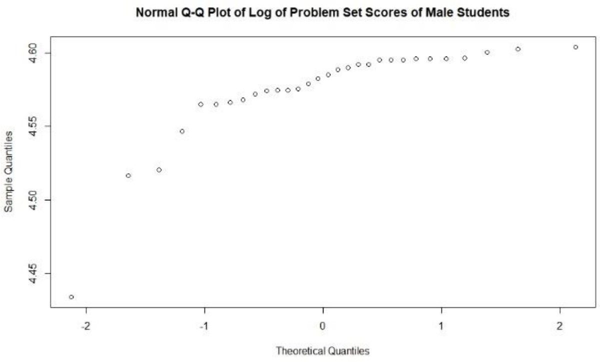
From the Q-Q plots, it is clear that the data for both the samples data is skewed towards left and does not form a straight line. Hence, the data is not normal.

Transforming the data by taking log of the scores:

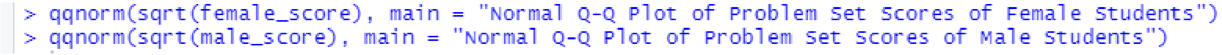


```
> qqnorm(log(female_score), main = "Normal Q-Q Plot of Log of Problem Set Scores of Female Students")
> qqnorm(log(male_score), main = "Normal Q-Q Plot of Log of Problem Set Scores of Male Students")
```

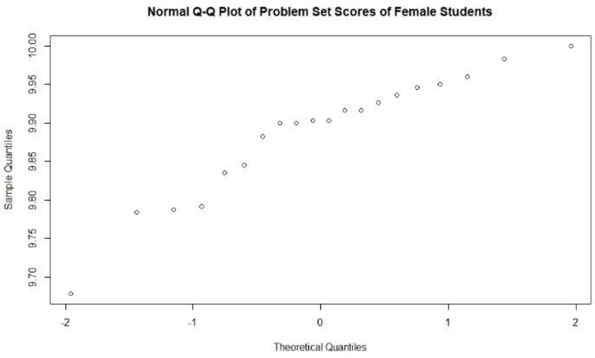


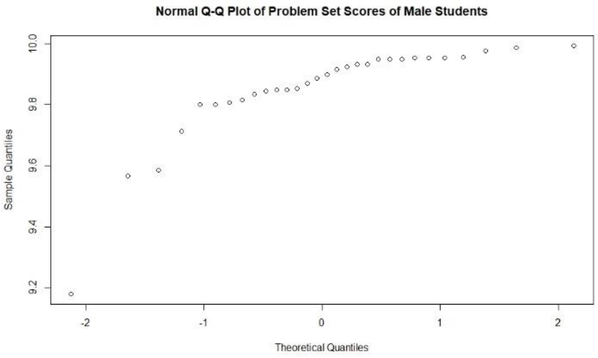


Transforming the data by taking square root of the scores:



```
> qqnorm(sqrt(female_score), main = "Normal Q-Q Plot of Problem Set Scores of Female Students")
> qqnorm(sqrt(male_score), main = "Normal Q-Q Plot of Problem Set Scores of Male Students")
```



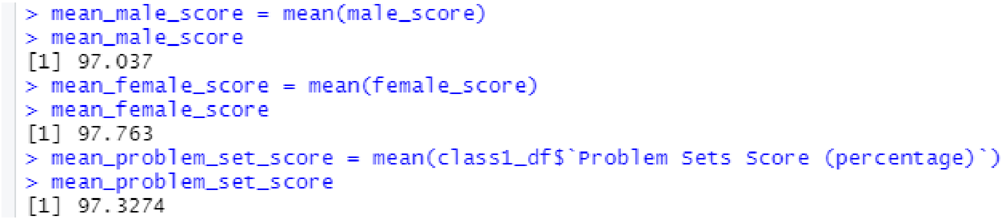


Even after transforming the data by taking log and square root, the Q-Q normal plots are similar with data being skewed on the left side.

For a small sample size, the normal population might give a straight Q-Q plot or curved Q-Q plot while same can be true for the skewed population, so the Q-Q plot won’t give us the definite answer as the sample size of the students’ data is too small to be confident about the normality assumption.

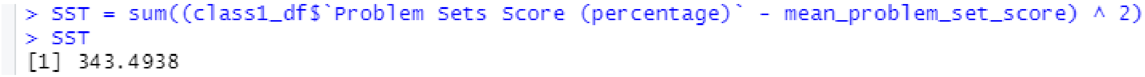
Creating ANOVA table by finding each value manually:

Calculating mean of each sample:



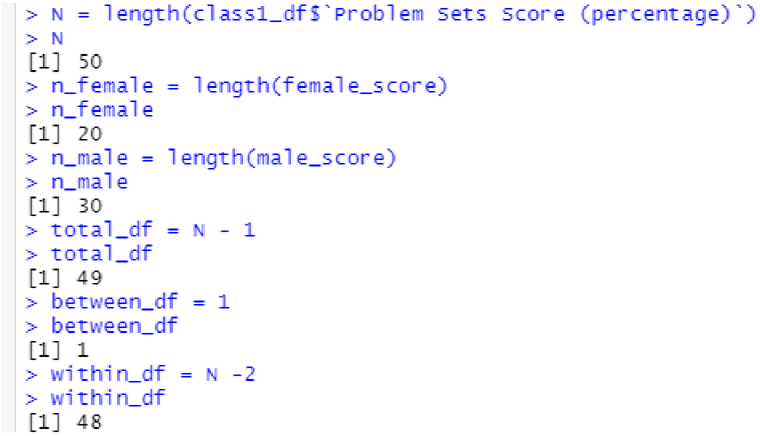
```
> mean_male_score = mean(male_score)
> mean_male_score
[1] 97.037
> mean_female_score = mean(female_score)
> mean_female_score
[1] 97.763
> mean_problem_set_score = mean(class1_df$`Problem Sets Score (percentage)`)
> mean_problem_set_score
[1] 97.3274
```

Sum of squares:



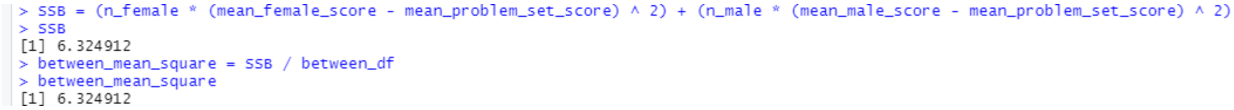
```
> SST = sum((class1_df$`Problem Sets Score (percentage)` - mean_problem_set_score) ^ 2)
> SST
[1] 343.4938
```

Sample size and degree of freedom:



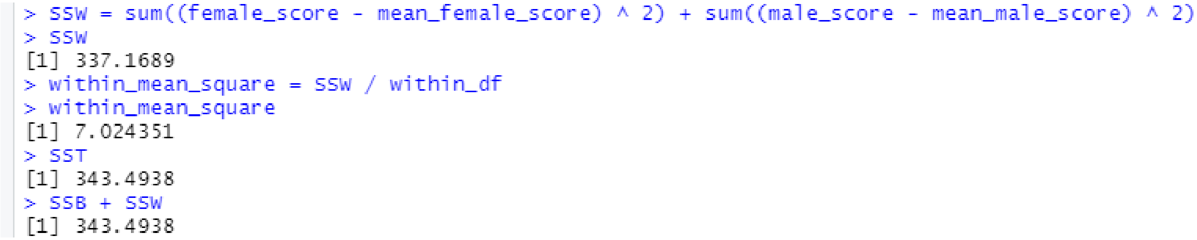
```
> N = length(class1_df$`Problem Sets Score (percentage)`)
> N
[1] 50
> n_female = length(female_score)
> n_female
[1] 20
> n_male = length(male_score)
> n_male
[1] 30
> total_df = N - 1
> total_df
[1] 49
> between_df = 1
> between_df
[1] 1
> within_df = N -2
> within_df
[1] 48
```

Between sum of squares:



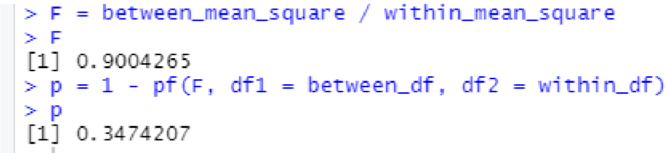
```
> SSB = (n_female * (mean_female_score - mean_problem_set_score) ^ 2) + (n_male * (mean_male_score - mean_problem_set_score) ^ 2)
> SSB
[1] 6.324912
> between_mean_square = SSB / between_df
> between_mean_square
[1] 6.324912
```

Within sum of squares:



```
> SSW = sum((female_score - mean_female_score) ^ 2) + sum((male_score - mean_male_score) ^ 2)
> SSW
[1] 337.1689
> within_mean_square = SSW / within_df
> within_mean_square
[1] 7.024351
> SST
[1] 343.4938
> SSB + SSW
[1] 343.4938
```

Calculate test statistic and p-value:

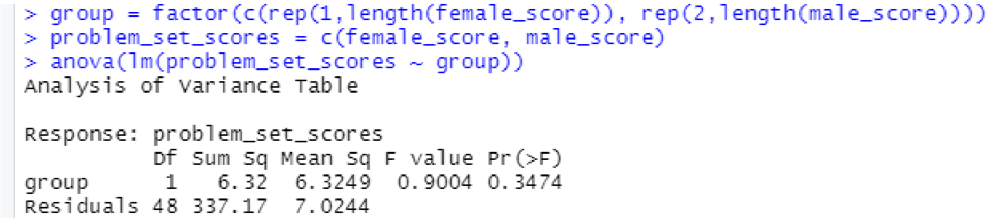


```
> F = between_mean_square / within_mean_square
> F
[1] 0.9004265
> p = 1 - pf(F, df1 = between_df, df2 = within_df)
> p
[1] 0.3474207
```

ANOVA table:

Source	SS	df	MS	F	p
Between	6.324912	1	6.324912	0.9004265	0.3474207
Within	337.1689	48	7.024351		
Total	343.4938	49			

Creating ANOVA table using R function:



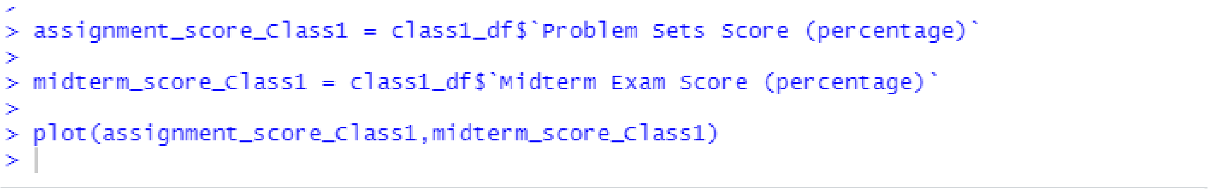
```
> group = factor(c(rep(1,length(female_score)), rep(2,length(male_score))))
> problem_set_scores = c(female_score, male_score)
> anova(lm(problem_set_scores ~ group))
Analysis of Variance Table

Response: problem_set_scores
          Df Sum Sq Mean Sq F value Pr(>F)
group      1   6.32  6.3249  0.9004 0.3474
Residuals 48 337.17  7.0244
```

Conclusion: We cannot reject H_0, which means that both the genders perform equally in the problem set assignments.

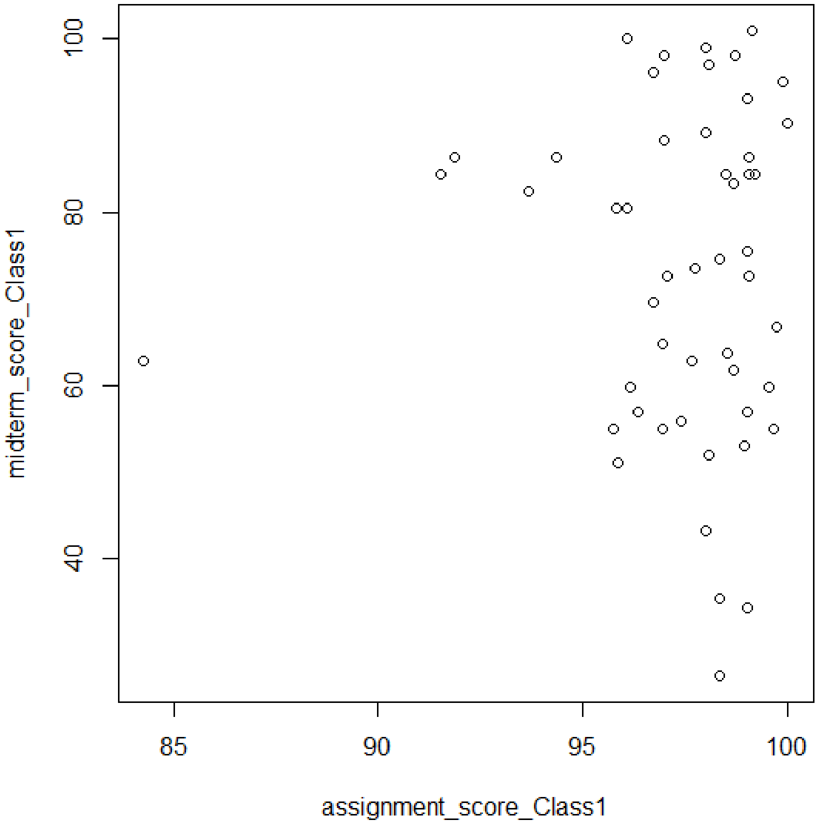
Regression and Prediction:

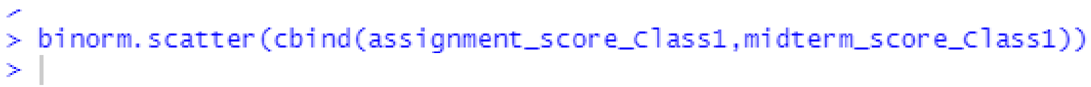
Q. Should we use students' assignment grades to predict their midterm exam scores? Is the conclusion the same for different programs (i.e., online vs residential)?

Analyzing Class 1 data:

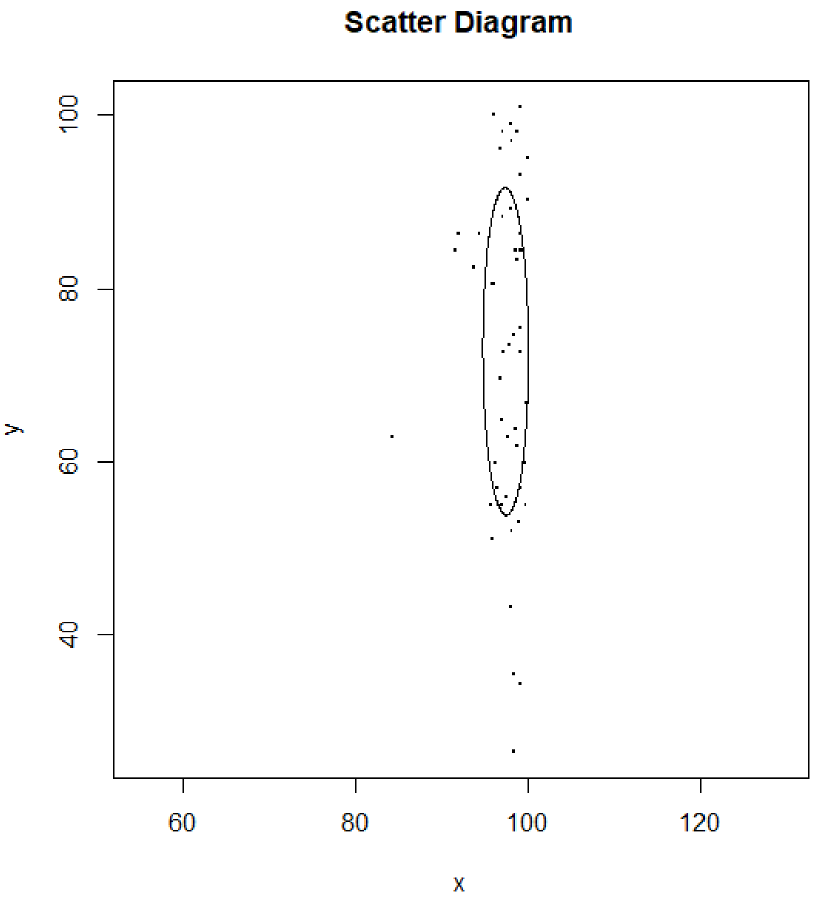


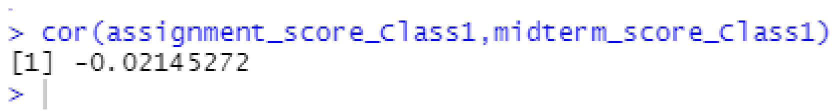
```
> assignment_score_Class1 = class1_df$`Problem Sets Score (percentage)`
>
> midterm_score_Class1 = class1_df$`Midterm Exam Score (percentage)`
>
> plot(assignment_score_Class1,midterm_score_Class1)
>
```





```
> binorm.scatter(cbind(assignment_score_Class1,midterm_score_Class1))
>
```



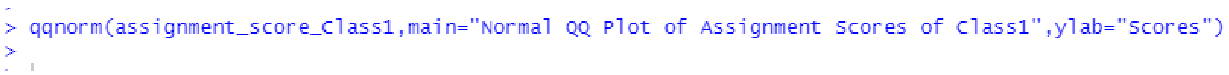


```
> cor(assignment_score_Class1,midterm_score_Class1)
[1] -0.02145272
> |
```

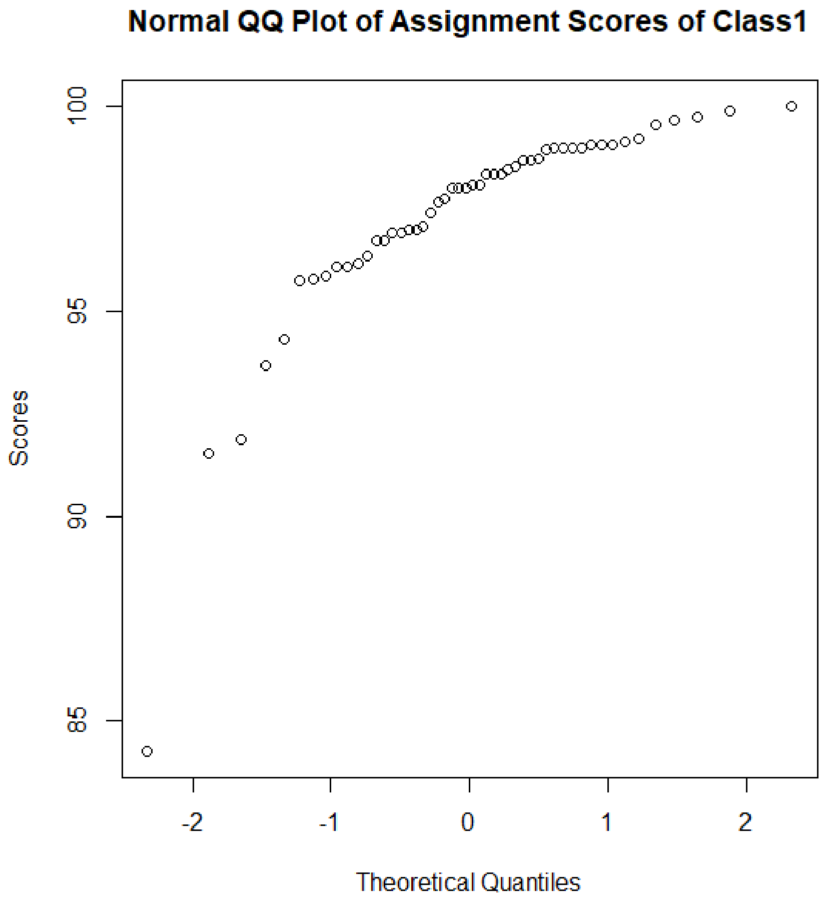
Cor(x,y) = -002145272

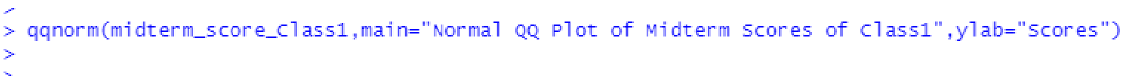
The association is harder to see here, as it is not clear whether we have positive or negative association, and it is visible that the relationship is not linear because few students who have higher assignment score also have less score in midterm. And based on the correlation coefficient value we can say that the association between two variables is too weak.

To check whether two variables are well approximated by a bivariate normal distribution, we will check normal Q-Q plots:

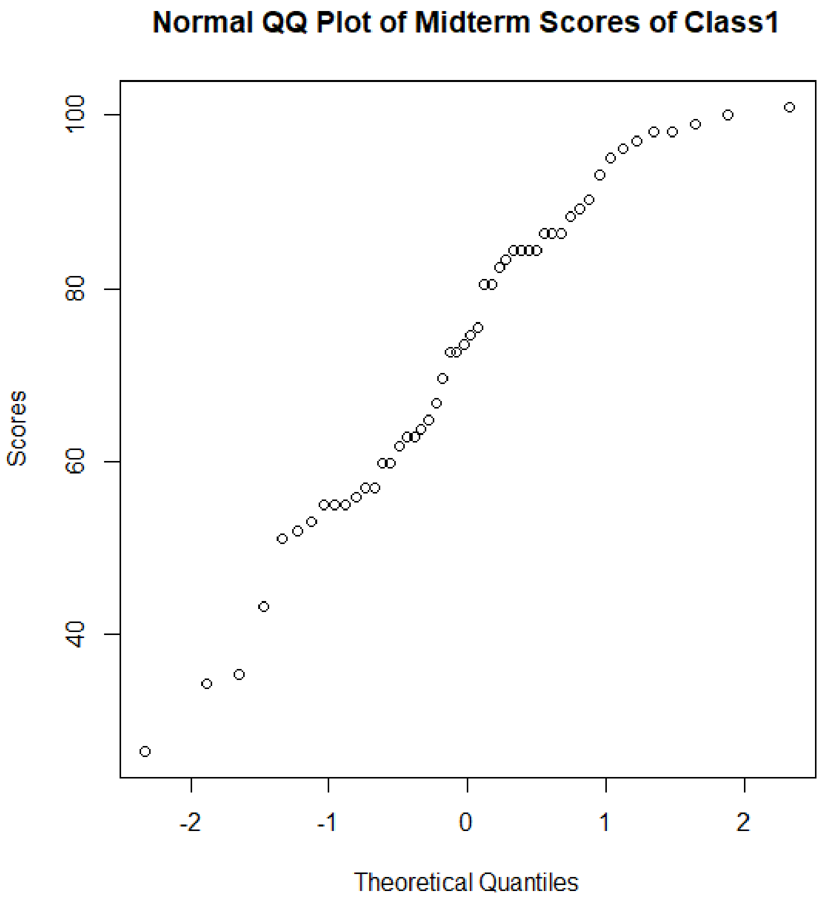


```
> qqnorm(assignment_score_Class1,main="Normal QQ Plot of Assignment Scores of Class1",ylab="Scores")
>
```





```
> qqnorm(midterm_score_Class1,main="Normal QQ Plot of Midterm Scores of Class1",ylab="Scores")
>
```



As with small sample sizes, we can be definite whether the data is really close to normal or not. In case of class 2 students at least we don’t see extreme skewness or outliers that would contradict approximate normality.

Now let’s do hypothesis testing:

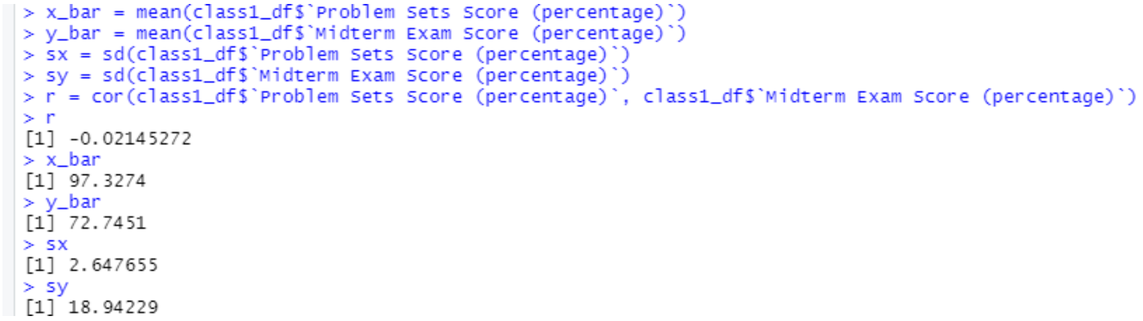
Null Hypothesis (H_0) states that there isn’t enough evidence that knowing problem set marks helps in prediction of the midterm scores for Class 1.

$H_0: \beta_1 = 0$

And Alternate Hypothesis (H_1) will be that there is enough evidence that knowing problem set marks can help in prediction of the midterm scores for Class 1.

$H_1: \beta_1 \neq 0$

Calculating sample statistics for the data using R:



```
> x_bar = mean(class1_df$`Problem Sets Score (percentage)`)
> y_bar = mean(class1_df$`Midterm Exam Score (percentage)`)
> sx = sd(class1_df$`Problem Sets Score (percentage)`)
> sy = sd(class1_df$`Midterm Exam Score (percentage)`)
> r = cor(class1_df$`Problem Sets Score (percentage)`, class1_df$`Midterm Exam Score (percentage)`)
> r
[1] -0.02145272
> x_bar
[1] 97.3274
> y_bar
[1] 72.7451
> sx
[1] 2.647655
> sy
[1] 18.94229
```

r = -0.02145272

$\bar{x}$ = 97.3274

$\bar{y}$ = 72.7451

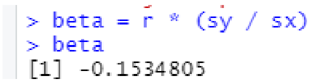
$S_x = 2.647655$

$S_y = 18.94229$

Using t-test to find the p-value

$$\hat{\beta}_1 = r\frac{S_y}{S_x}$$

Using R:

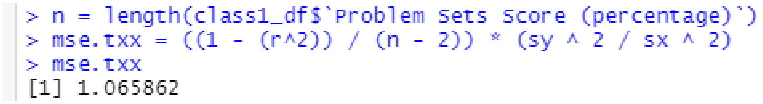


```
> beta = r * (sy / sx)
> beta
[1] -0.1534805
```

$\hat{\beta}_1 = -0.1534805$

$$MS_E\,/t_{xx} = \frac{1 - r^2}{n - 2}\left(\frac{{S_y}^2}{{S_x}^2}\right)$$

Using R:

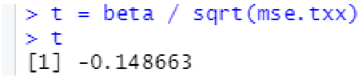


```
> n = length(class1_df$`Problem Sets Score (percentage)`)
> mse.txx = ((1 - (r^2)) / (n - 2)) * (sy ^ 2 / sx ^ 2)
> mse.txx
[1] 1.065862
```

$MS_E\,/t_{xx} = 1.0065862$

t = $\frac{\hat{\beta}_1}{\sqrt{MS_E\,/t_{xx}}}$

Using R:

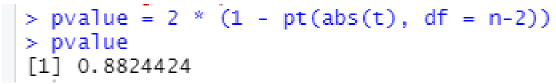


```
> t = beta / sqrt(mse.txx)
> t
[1] -0.148663
```

Since it’s a two-tailed test:

p-value = 2 * (1 - pt(abs(t), df = n-2))

Using R:

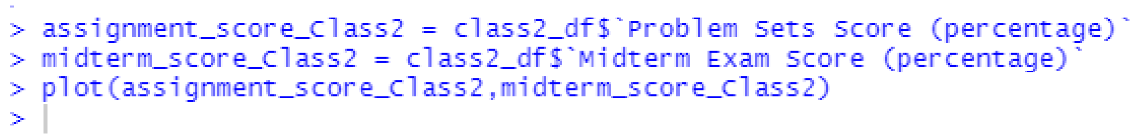


```
> pvalue = 2 * (1 - pt(abs(t), df = n-2))
> pvalue
[1] 0.8824424
```

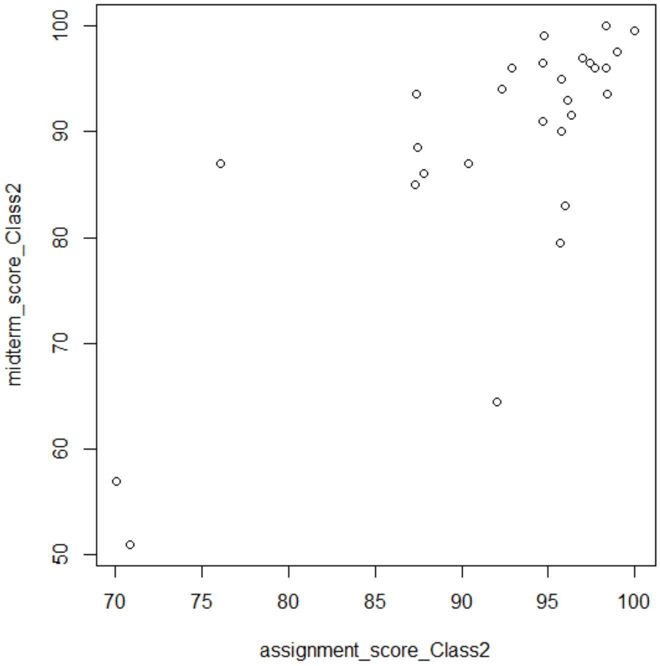
Since p-value = 0.8824424 > $\alpha = 0.05$, we do not reject $H_{0.}$

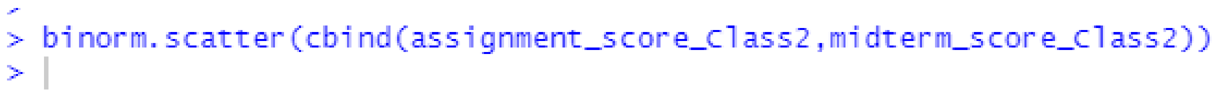
Conclusion: There is not enough evidence that evidence that knowing problem set marks helps in prediction of the midterm scores for Class 1.

<u>Analyzing Class 2 data:</u>

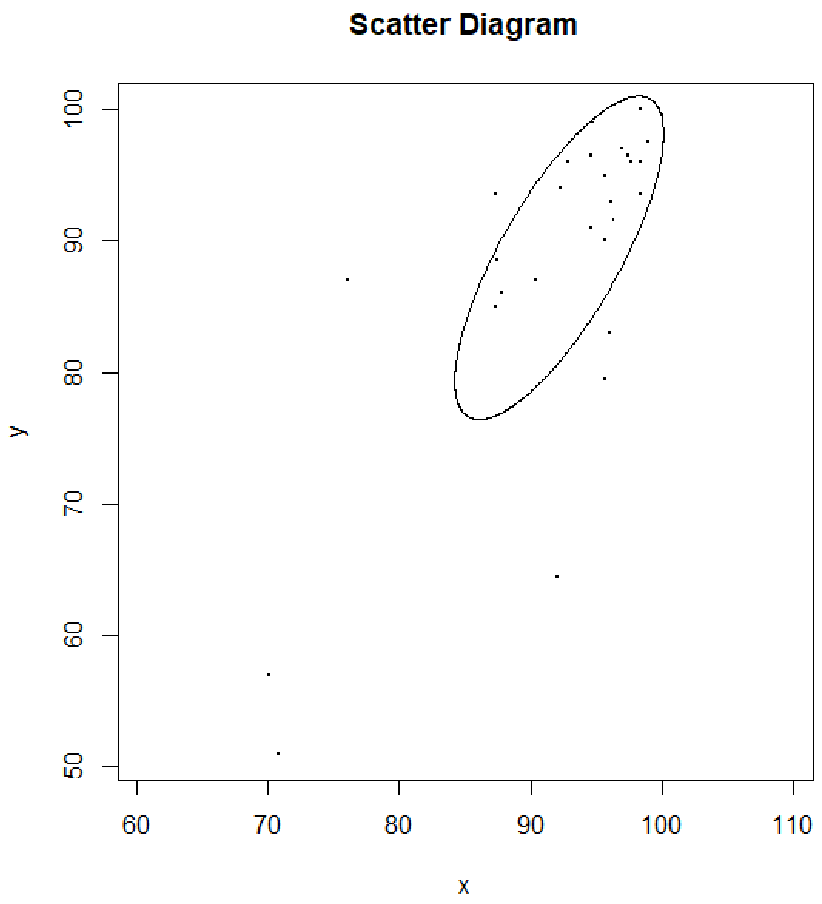


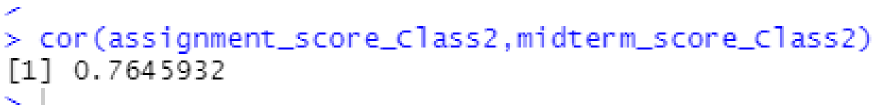
```
> assignment_score_Class2 = class2_df$`Problem Sets Score (percentage)`
> midterm_score_Class2 = class2_df$`Midterm Exam Score (percentage)`
> plot(assignment_score_Class2,midterm_score_Class2)
> 
```





```
> binorm.scatter(cbind(assignment_score_Class2,midterm_score_Class2))
> 
```



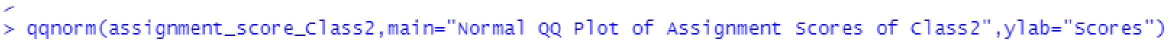


```
> cor(assignment_score_Class2,midterm_score_Class2)
[1] 0.7645932
> 
```

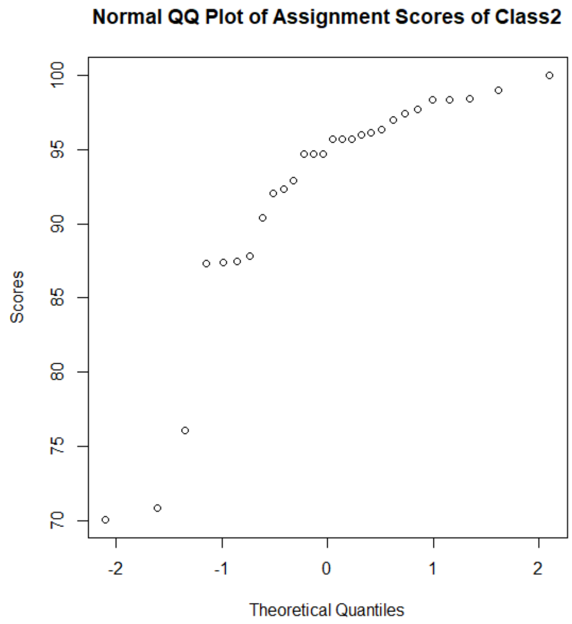
Cor = 0.7645932

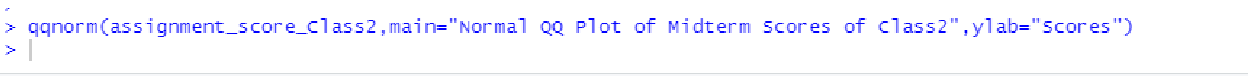
As ellipse indicates the shape of the data cloud, it shows that there is a relationship between tow variables which is the assignment scores of Class 2 students and midterm scores of Class 1 students. The relationship seems to be of the form y = x which is a linear association. As the x goes up i.e. the assignment scores goes up, y tends to go up i.e their midterm tends to go up as well. So, the association is quite linear and strong with ellipse pointing upwards which shows that the association is positive.

To check whether two variables are well approximated by a bivariate normal distribution, we will check normal Q-Q plots:

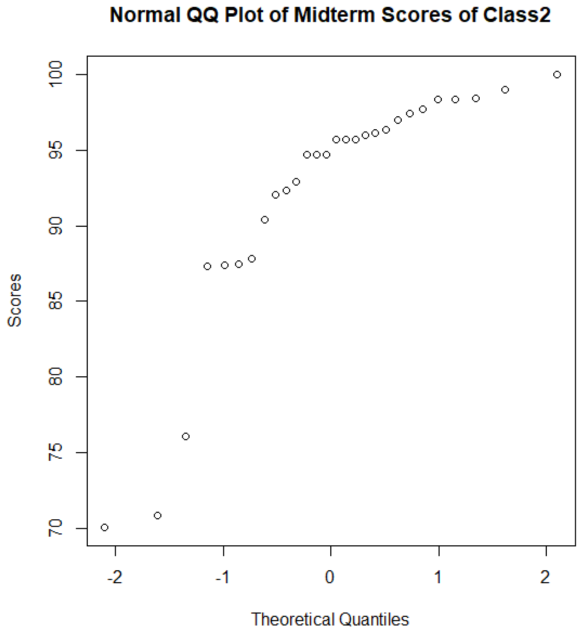


```
> qqnorm(assignment_score_Class2,main="Normal QQ Plot of Assignment Scores of Class2",ylab="Scores")
```





```
> qqnorm(assignment_score_Class2,main="Normal QQ Plot of Midterm Scores of Class2",ylab="Scores")
> 
```



As with small sample sizes, we can be definite whether the data is really close to normal or not. In case of class 2 students at least we don't see extreme skewness or outliers that would contradict approximate normality.

Now let's do hypothesis testing:

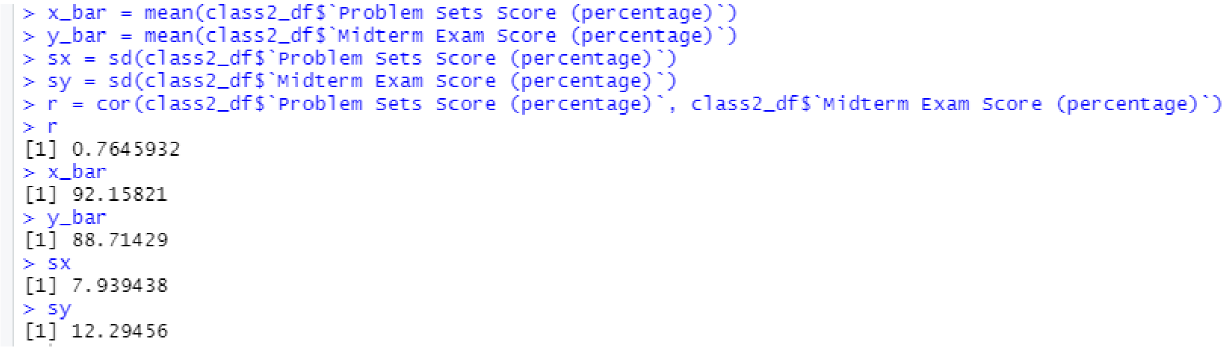
Null Hypothesis (H_0) states that there isn't enough evidence that knowing problem set marks helps in prediction of the midterm scores for Class 2.

$H_0: \beta_1 = 0$

And Alternate Hypothesis (H_1) will be that there is enough evidence that knowing problem set marks can help in prediction of the midterm scores for Class 2.

$H_1: \beta_1 \neq 0$

Calculating sample statistics for the data using R:



```
> x_bar = mean(class2_df$`Problem Sets Score (percentage)`)
> y_bar = mean(class2_df$`Midterm Exam Score (percentage)`)
> sx = sd(class2_df$`Problem Sets Score (percentage)`)
> sy = sd(class2_df$`Midterm Exam Score (percentage)`)
> r = cor(class2_df$`Problem Sets Score (percentage)`, class2_df$`Midterm Exam Score (percentage)`)
> r
[1] 0.7645932
> x_bar
[1] 92.15821
> y_bar
[1] 88.71429
> sx
[1] 7.939438
> sy
[1] 12.29456
```

r = 0.7645932
$\bar{x}$ = 92.15821

$\bar{y}$ = 88.71429

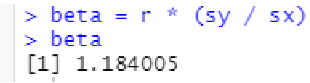
$S_x = 7.939438$

$S_y = 12.29456$

Using t-test to find the p-value

$$\hat{\beta}_1 = r\frac{S_y}{S_x}$$

Using R:

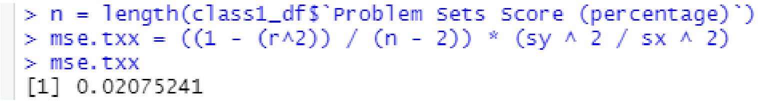


```
> beta = r * (sy / sx)
> beta
[1] 1.184005
```

$\hat{\beta}_1 = 1.184005$

$$MS_E\ /t_{xx} = \frac{1\ -\ r^2}{n\ -\ 2}(\frac{{S_y}^2}{{S_x}^2})$$

Using R:

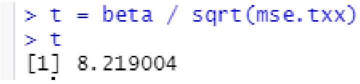


```
> n = length(class1_df$`Problem Sets Score (percentage)`)
> mse.txx = ((1 - (r^2)) / (n - 2)) * (sy ^ 2 / sx ^ 2)
> mse.txx
[1] 0.02075241
```

$MS_E\ /t_{xx} =$ 0.02075241

t = $\frac{\widehat{\beta}_1}{\sqrt{MS_E\ /t_{xx}}}$

Using R:

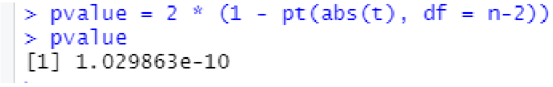


```
> t = beta / sqrt(mse.txx)
> t
[1] 8.219004
```

Since it’s a two-tailed test:

p-value = 2 * (1 - pt(abs(t), df = n-2))

Using R:



```
> pvalue = 2 * (1 - pt(abs(t), df = n-2))
> pvalue
[1] 1.029863e-10
```

Since p-value = 1.029863e-10 < $\alpha = 0.05$, we reject $H_{0.}$

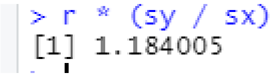
Hence, using problem set marks we can predict the midterm scores for Class 2.

Let’s try to predict a Midterm Score using a random Problem Set Score.

The slope of the regression line is given by:

b = $\frac{r\frac{S_y}{}}{S_x}$

Using R:



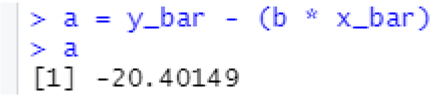
```
> r * (sy / sx)
[1] 1.184005
```

b = 1.184005

The intercept of the regression line is:

a = $\bar{y} - b\ \bar{x}$

Using R:

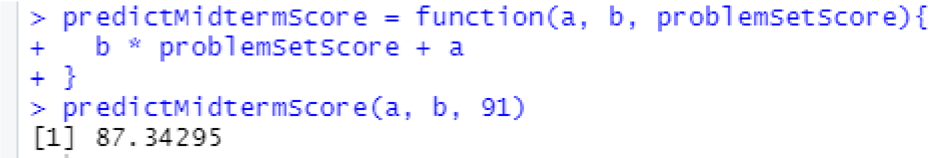


```
> a = y_bar - (b * x_bar)
> a
[1] -20.40149
```

The predicted value can be given by equation:

Midterm Exam Predicted Score (Percentage) $= b\ *\ (Problem\ Set\ Score)\ +\ a$

We have created a function to predict midterm exam scores using a problem set score in R:



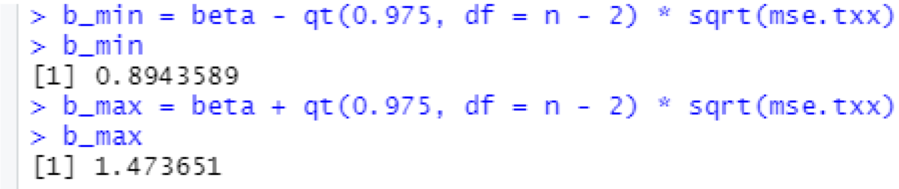
```
> predictMidtermScore = function(a, b, problemSetScore){
+   b * problemSetScore + a
+ }
> predictMidtermScore(a, b, 91)
[1] 87.34295
```

The confidence interval is calculated as:

$$\hat{\beta}_1 \pm q_t \sqrt{\frac{MS_E}{t_{xx}}}$$

Consider 95% level of confidence

Using R:

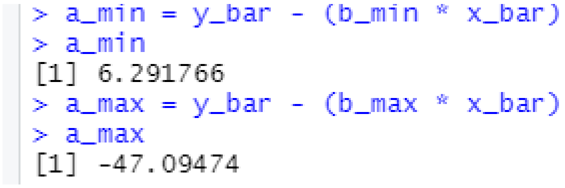


```
> b_min = beta - qt(0.975, df = n - 2) * sqrt(mse.txx)
> b_min
[1] 0.8943589
> b_max = beta + qt(0.975, df = n - 2) * sqrt(mse.txx)
> b_max
[1] 1.473651
```

Confidence interval = (0.8943591, 1.473651)

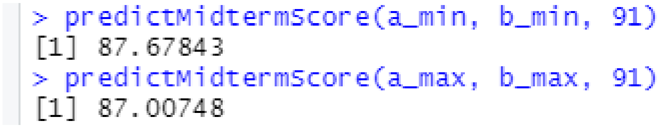
The confidence interval of the slope of line lies between (0.8943591, 1.473651)

The intercept of the regression line will be in interval:



```
> a_min = y_bar - (b_min * x_bar)
> a_min
[1] 6.291766
> a_max = y_bar - (b_max * x_bar)
> a_max
[1] -47.09474
```

This means the predicted value could also be in a range.



```
> predictMidtermScore(a_min, b_min, 91)
[1] 87.67843
> predictMidtermScore(a_max, b_max, 91)
[1] 87.00748
```

Thus, the predicted value can lie in the range 87% to 87.67%

Conclusion: From this observation, we can predict that as compared to the problem set assignments, the percentage score of students of Class 2 would be somewhat lesser in the midterm exam.

Final Conclusion:

By observing the data and using Welch's t-test for analysis, we inferred that Class 2 performed better than Class 1 in midterm exams. Consequently, we examined the performance of male and female students in assignments of Class 1 and derived that both genders had performed equally well implying that gender did not impact the grades. Furthermore, we checked if we can do predictive analysis on the student assignment scores for both the classes. Interestingly, we could only implement predictive analysis using Simple Linear Regression on the scores of Class 2 students as there was considerable linear association between the parameters of interest.

References:

- How to Import Excel Files into R (Step-by-Step)
  https://www.statology.org/import-excel-into-r/#:~:text=The%20easiest%20way%20to%20import,function%20from%20the%20readxl%20package.&text=where%3A,sheet%3A%20The%20sheet%20to%20read
- How to Add a Column to a Data Frame in R (With Examples)
  https://www.statology.org/r-add-a-column-to-dataframe/
- Colors in R by Dr. Ying Wei
  http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf