Stann-Omar Jones
Advanced Statistics for Data Science
Prof. Chong Liu

1. Inspect the iris data in R. (5 marks)

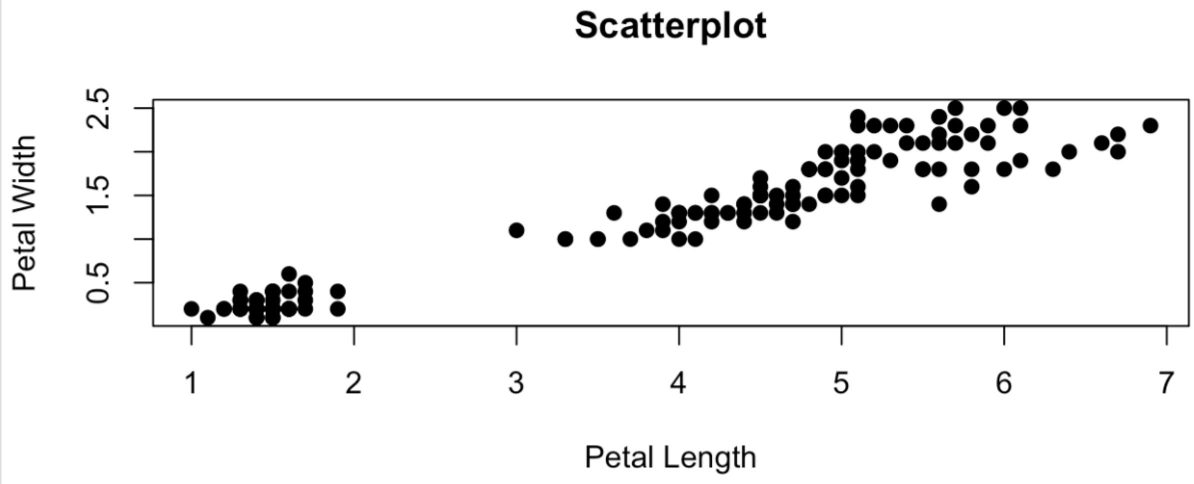
```
> dim(iris) #see number of rows and columns
[1] 150   5
> View(iris)
> str(iris)
'data.frame':	150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

We 150 observations across 5 attributes in this data sample.

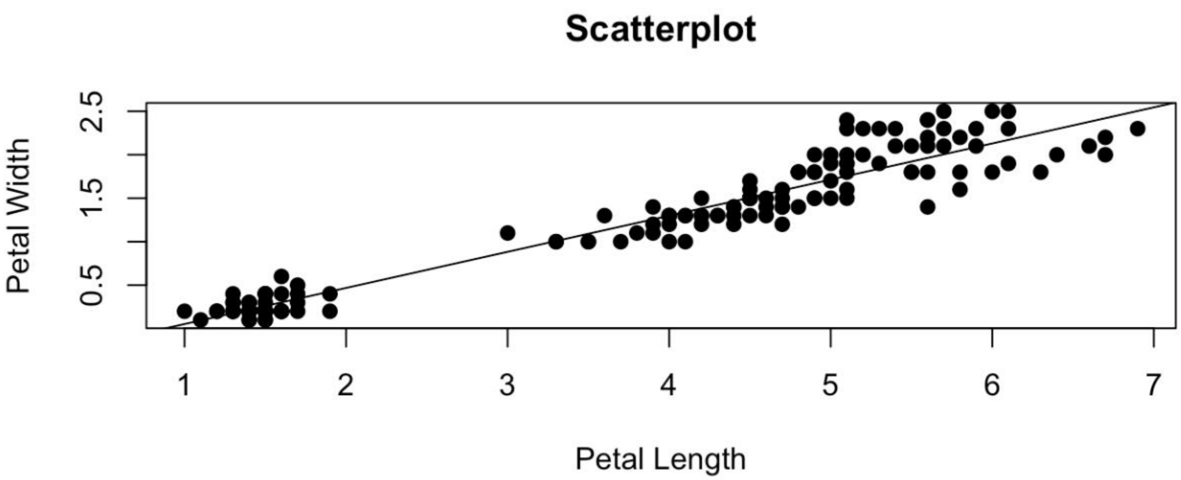
2. Use the summary code in R to perform descriptive analysis. Paste summary statistics into your report. (5 marks)

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```

3. Draw a scatter plot for petal length vs petal width. (5 marks)

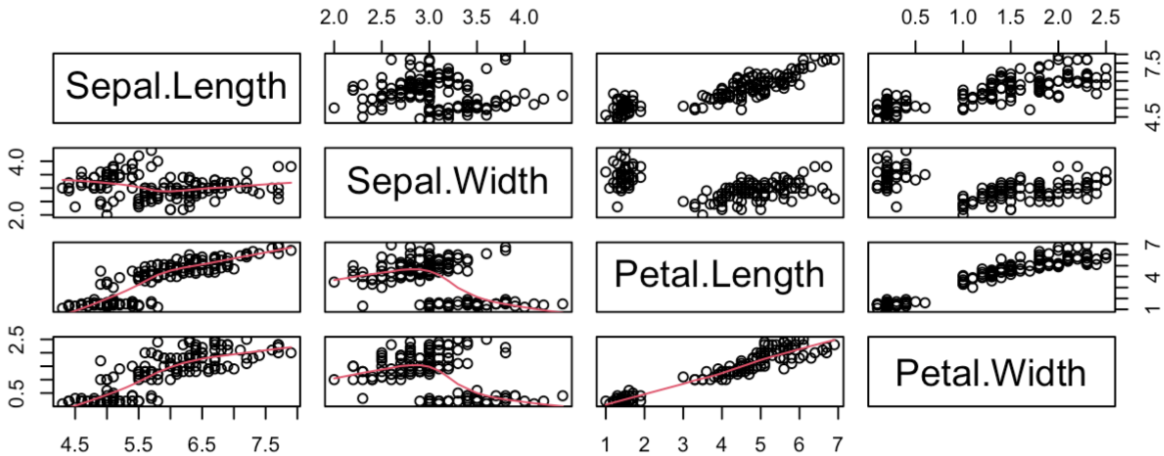


With abline:



We can see that there is a positive trend (reinforced by the abline) between petal length and petal width. Intuitively, the longer a petal is, the wider the petal will be also.

4. `Use pairs command for creating pairwise scatter plot for all variables in the data set. (5 marks)`



A pairwise scatter plot allows us to see the relationship between any two variables of the concerned data-set as a matrix.

```
5. Find all possible correlation between quantitative
   variables. (5 marks)
```

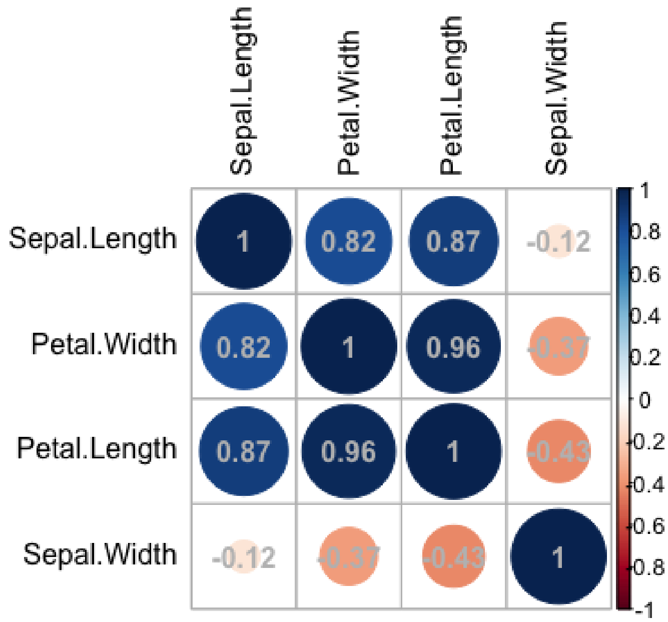
```
> cor(cor_matrix)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

From the pairwise scatterplot, we can confirm the direction of the relationships with the actual correlation measures between the variables.

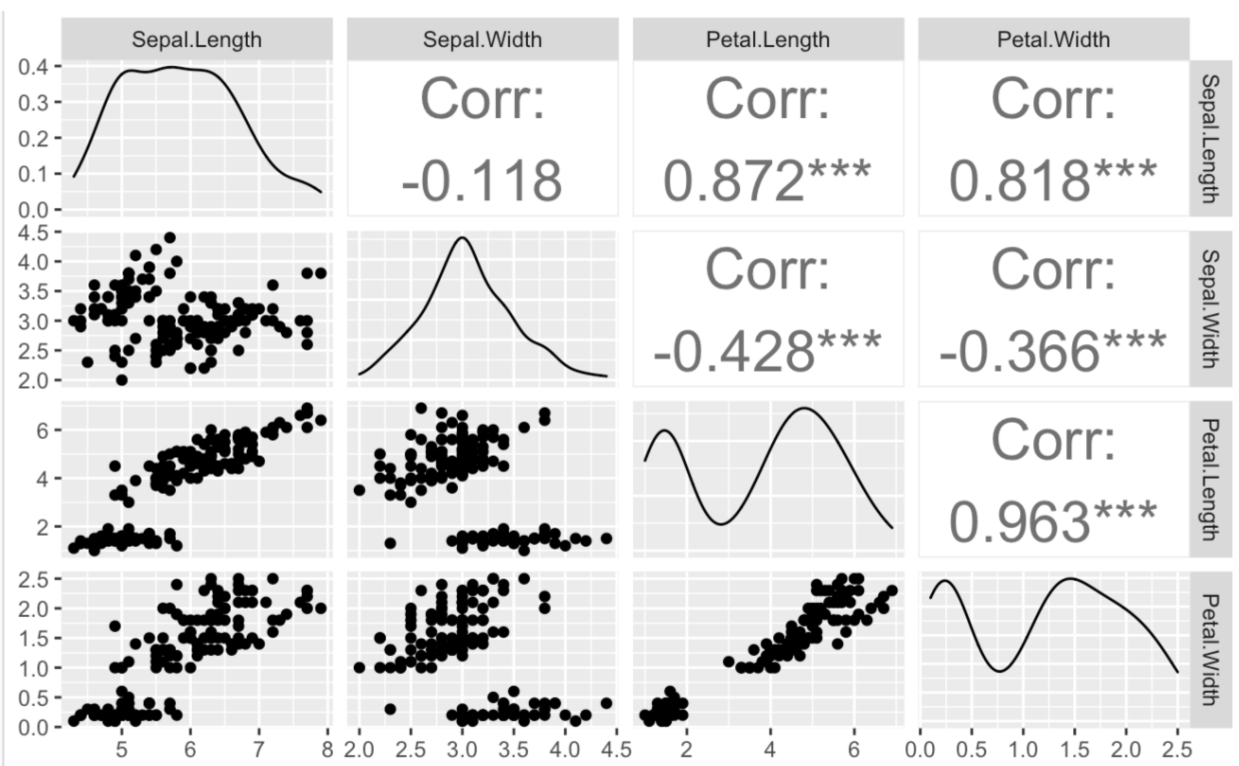
A correlation matrix displays values between -1 and 1 where:

- -1 indicates a perfectly negative linear correlation between two variables (when one variable increases as the other decreases)
- 0 indicates no linear correlation between two variables
- 1 indicates a perfectly positive linear correlation between two variables (when one variable increases as the other increases too)

It can difficult to detect whether a bivariate relationship is positive or negative sometimes e.g. between sepal length and sepal width. The correlation matrix shows us that these variables have a slightly negative relationship so as sepal length increases sepal width decreases.



This is another visual representation of the relationship between the quantitative variables in the iris dataset.



This is another visual representation of the relationship between the quantitative variables (using scatterplots and density plots) in the iris dataset.

6. Use isfit command for two highly correlated variables. (5 marks)

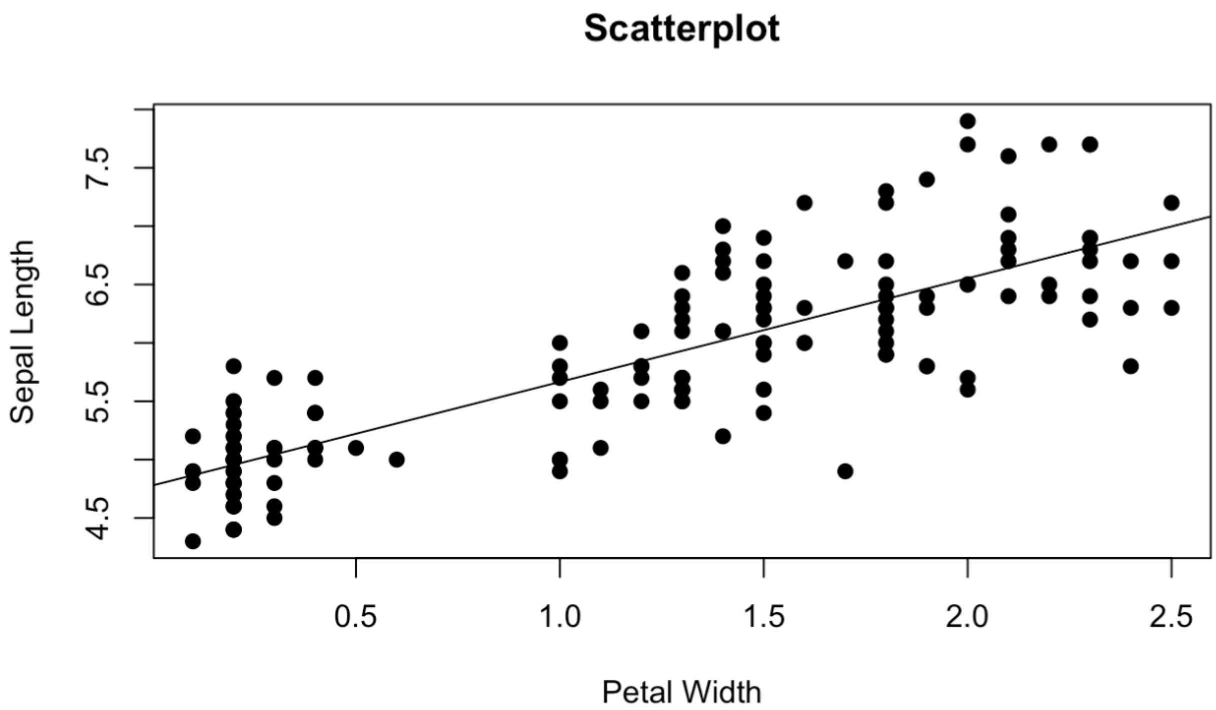
```
> lsfit(iris$Sepal.Length, iris$Petal.Width)
$coefficients
 Intercept          X
-3.2002150  0.7529176

$residuals
  [1] -0.439664606 -0.289081092 -0.138497578 -0.063205820 -0.364372849 -0.465539877
  [7]  0.036794180 -0.364372849  0.087377694 -0.389081092 -0.665539877 -0.213789335
 [13] -0.313789335  0.062669451 -0.966706905 -0.691415148 -0.465539877 -0.339664606
 [19] -0.791415148 -0.339664606 -0.665539877 -0.239664606 -0.063205820 -0.139664606
 [25] -0.213789335 -0.364372849 -0.164372849 -0.514956363 -0.514956363 -0.138497578
 [31] -0.213789335 -0.465539877 -0.614956363 -0.740831634 -0.289081092 -0.364372849
 [37] -0.740831634 -0.389081092  0.087377694 -0.439664606 -0.264372849  0.112085937
 [43]  0.087377694  0.035627151 -0.239664606 -0.113789335 -0.439664606 -0.063205820
 [49] -0.590248120 -0.364372849 -0.670207990 -0.118457448 -0.494916233  0.359168366
 [55] -0.193749205  0.208584852  0.056834309  0.510918908 -0.469040962  0.685043637
 [61]  0.435627151  0.258001338 -0.317290419  0.007417824  0.283876609 -0.444332719
 [67]  0.483876609 -0.166706905  0.032126066  0.083876609  0.558001338 -0.092582176
 [73] -0.043165691 -0.192582176 -0.318457448 -0.369040962 -0.519624476 -0.144332719
 [79]  0.182709581 -0.091415148  0.159168366  0.059168366  0.033293095  0.282709581
 [85]  0.634460123  0.282709581 -0.344332719 -0.243165691  0.283876609  0.359168366
 [91]  0.259168366  0.007417824  0.033293095  0.435627151  0.283876609  0.108584852
 [97]  0.208584852 -0.167873934  0.460335394  0.208584852  0.956834309  0.733293095
[103] -0.045499747  0.256834309  0.506250795 -0.421958532  1.210918908 -0.496083261
[109] -0.044332719  0.279208496  0.306250795  0.281542552  0.180375524  0.908584852
[115]  1.233293095  0.681542552  0.106250795 -0.397250290 -0.297250290  0.182709581
[121]  0.305083767  0.983876609 -0.597250290  0.256834309  0.255667281 -0.420791504
[127]  0.332126066  0.407417824  0.481542552 -0.620791504 -0.471375018 -0.747833804
[133]  0.581542552 -0.043165691  0.007417824 -0.297250290  0.856834309  0.181542552
[139]  0.482709581  0.105083767  0.555667281  0.305083767  0.733293095  0.380375524
[145]  0.655667281  0.455667281  0.356834309  0.306250795  0.832126066  0.558001338

$intercept
```

This simple linear regression tells us that when there is a single unit increase in sepal length (predictor variable), there is a 0.752 unit increase in petal width (response variable).

7. Plot a line of fit using abline command. (5 marks)



The line of fit shows that there is a positive relationship between sepal length and petal width.

8. Use function lm for developing a regression model and paste the summary of the regression model in your report-- Petal.Width ~ Petal.Length and for Sepal.Length ~ Sepal.Width (10 marks)

```
> summary(lm(formula = Petal.Width ~ Petal.Length, data = iris))

Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.56515 -0.12358 -0.01898  0.13288  0.64272 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,	Adjusted R-squared:  0.9266 
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

This simple linear regression output tells us that when there is a single unit increase in petal length (predictor variable), there is a 0.416 unit increase in petal width (response variable).
The significance level of an event (such as a statistical test) is the probability that the event could have occurred by chance. If the level is quite low, that is, the probability of occurring by chance is quite small, we say the event is *significant.*

As we can see, the petal length variable is significant in our output so the chance of these results being produced by chance are quite low.

```
> summary(lm(formula = Sepal.Length ~ Sepal.Width, data =iris))

Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width  -0.2234     0.1551   -1.44    0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,	Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

This simple linear regression output tells us that when there is a single unit increase in sepal length (predictor variable), there is a -0.2234 unit decrease in sepal width (response variable).

As we can see, the sepal width variable is not significant in our output.

Part 2 – A

A university investigation was conducted to determine if, on average, women and men complete medical school in significantly different amounts of time. Two independent random samples were selected and the following summary information concerning times to completion of medical school computed:

	Women	Men
Sample size	90	100
Sample mean	8.4 years	8.5 years
Sample standard deviation	0.6 years	0.5 years

Refer to Medical School Completion Narrative. Perform the appropriate test of the hypothesis to determine whether there is a significant difference in time regarding the completion of medical school between women and men. Test using . (10 marks)

Method One

```
> t.test2 <- function(m1,m2,s1,s2,n1,n2,m0=0,equal.variance=FALSE)
+ {
+   if( equal.variance==FALSE )
+   {
+     se <- sqrt( (s1^2/n1) + (s2^2/n2) )
+     # welch-satterthwaite df
+     df <- ( (s1^2/n1 + s2^2/n2)^2 )/( (s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1) )
+   } else
+   {
+     # pooled standard deviation, scaled by the sample sizes
+     se <- sqrt( (1/n1 + 1/n2) * ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2) )
+     df <- n1+n2-2
+   }
+   t <- (m1-m2-m0)/se
+   dat <- c(m1-m2, se, t, 2*pt(-abs(t),df))
+   names(dat) <- c("Difference of means", "Std Error", "t", "p-value")
+   return(dat)
+ }
> set.seed(0)
> x1 <- rnorm(90, mean = 8.4, sd = 0.6)
> x2 <- rnorm(100, mean = 8.5, sd = 0.5)
> (tt2 <- t.test2(mean(x1), mean(x2), sd(x1), sd(x2), length(x1), length(x2)))
Difference of means           Std Error                   t             p-value
        -0.08179856          0.07352345         -1.11255063          0.26739264
```

Method Two

```
> (tt <- t.test(x1, x2))

        Welch Two Sample t-test

data:  x1 and x2
t = -1.1126, df = 179.09, p-value = 0.2674
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.22688230  0.06328518
sample estimates:
mean of x mean of y
 8.397320  8.479118
```

The p-value is 0.267. This is greater than 0.05 so we cannot conclude that there is a significant difference in mean time to completion of medical school between women and men.

Part 2 – B

Suppose pulse rates of adult females have a normal curve distribution with mean and standard deviation . What is the probability that a randomly selected female has a pulse rate greater than 85?

Use **R** to calculate the probability, and paste screen shots for code and result. (10 marks)

```
> #z=pnorm(x,mean,std dev)
> (z=pnorm(85,75,8)) #P(Z<85)
[1] 0.8943502
> (x=1-z) #P(Z>85)=1-P(Z<85)
[1] 0.1056498
```

There is a 10.5% probability that a randomly selected woman will have a pulse over 85.

Part 3

Ten sampled students of 18-21 years of age received special training. They are given an IQ test that is N (100, 102) in the general population. Let μ be the mean IQ of these students who received special training. The observed IQ scores: 121, 98, 95, 94,102, 106, 112, 120, 108, 109. Test if the special training improves the IQ score using significance level α = 0.05.

a. What is the rejection region?

```
> qnorm(0.05,lower.tail=F)
[1] 1.644854
```

The critical value that defines the rejection region is 1.644854.

b. Calculate the p-value and state your conclusion.

```
> z.test(IQ,mu=100,stdev=sqrt(102),alt="g") #To test H0:mu=100 ag. H1:mu>100,

        One Sample z-test

data:  IQ
z = 2.0352, n = 10.0000, Std. Dev. = 10.0995, Std. Dev. of the sample mean
= 3.1937, p-value = 0.02091
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 101.2468      Inf
sample estimates:
mean of IQ
     106.5
```

Since p-value < 0.05, we can reject H0 and conclude that the IQ of the students have improved significantly after the special training.

c. What if the variance is unknown?

```
> t.test(IQ,mu=100,alt="g") # To test H0:mu=100 ag. H1:mu>100,

        One Sample t-test

data:  IQ
t = 2.1633, df = 9, p-value = 0.02937
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 100.9922      Inf
sample estimates:
mean of x
    106.5
```

To test the hypothesis when the variance is unknown, we use a one-sample t-test. Even though the p-value is different, it is still < 0.05 so we can reject H0 and conclude that the IQ of the students have improved significantly after the special training.

Use R studio to solve this problem. What codes you have to put in?