SUB : INTRO TO DATA SCIENCE LAB EXCERSICE 4

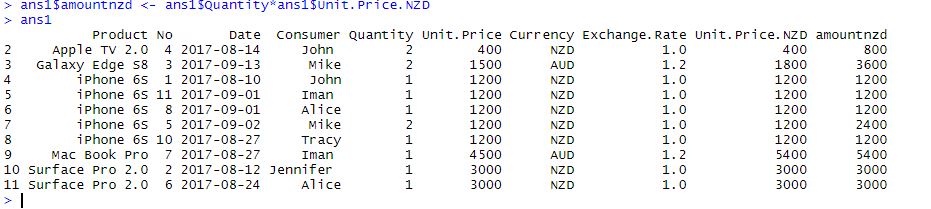
Name : Malhar Badheka (1522906)

# Question 1 [3.8 Marks]

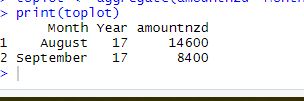
Develop a R Script file that can perform the following tasks on the data set given in “Data Set 1a.csv”:

1. Omitting all the transactions made before 1 Jan 2010 **[1.4 Marks]**
2. Calculating monthly total sale **[2 Marks]**
3. Visualizing the results (monthly total sale) by using appropriate charts/plots. **[0.4 Mark]**

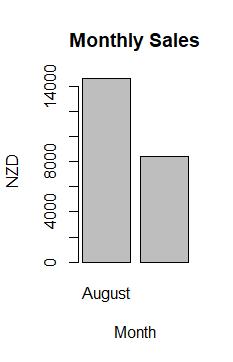
**ANS 1 :**

****

**ANS 2:**

****

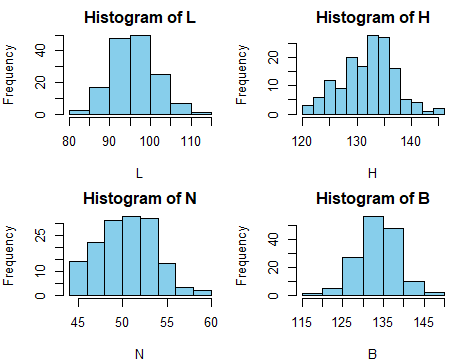
**ANS 3:**

****

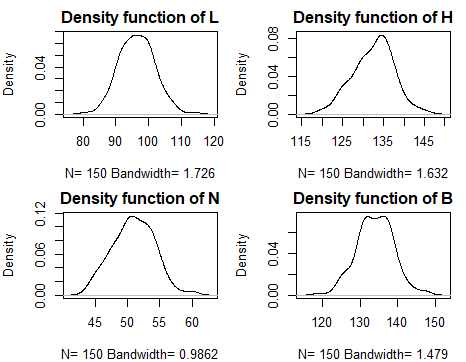
# Question 2 [4 Marks]

Develop a R Script file that can perform the following tasks on the Skulls data set given in “Data Set 3.csv”.

1. Plot the histograms of L, H, N, B in one page. The output should be similar to the following figure. **[1 Mark]**



1. Plot the density functions of L, H, N, B in one page. The output should be similar to the following figure. **[1 Mark]**

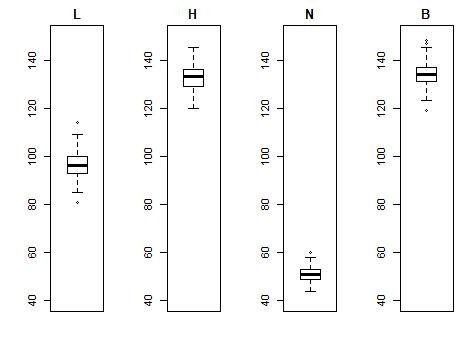


1. Compare the density functions against a normal density function. Comments on the symmetry and sharpness of the density functions. **[0.4 Marks]**

• *Hint: You may need to calculate shape measures and complete the following table.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **L** | **H** | **N** | **B** |
| Skewness | [1] 0.1394813 | - 0.1760332 | 0.08255358 | -0.02837297 |
| Kurtosis | 3.185601 | 2.889839 | 2.830727 | 3.25955 |

1. Create the boxplots of L, H, N and B using a similar scale. The output should be similar to the following figure. **[1 Mark]**



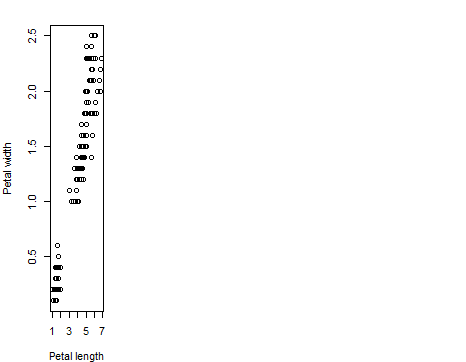
5 Calculate the mean, variance and standard deviation of L, H, N and B and complete the following table. **[0.6 Marks]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **L** | **H** | **N** | **B** |
| Mean | 96.46 | 132.5467 | 50.93333 | 133.9733 |
| Variance | 28.92121 | 24.39714 | 10.29083 | 23.91875 |
| Standard deviation | 5.377844 | 4.939346 | 3.207932 | 4.89068 |

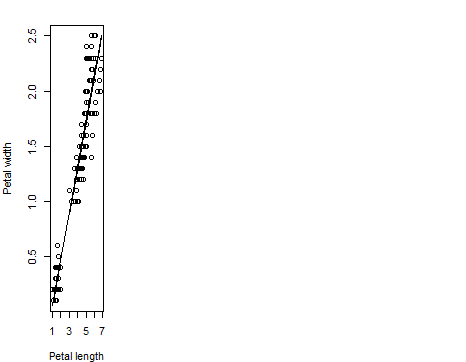
# Question 3 [4 Marks]

Develop a R Script file that can perform the following tasks on the Iris data set given in “Data Set 4.csv”.

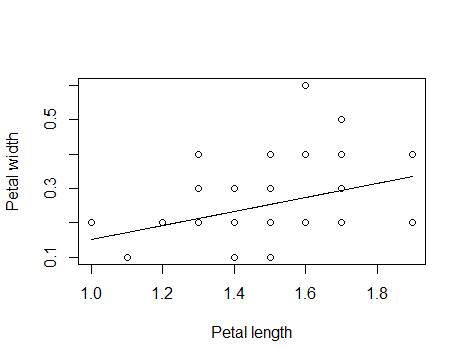
1. Create a scatter plot for petal length and width variables. **[0.1 Marks]**

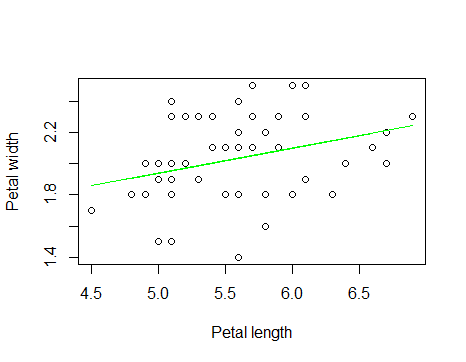
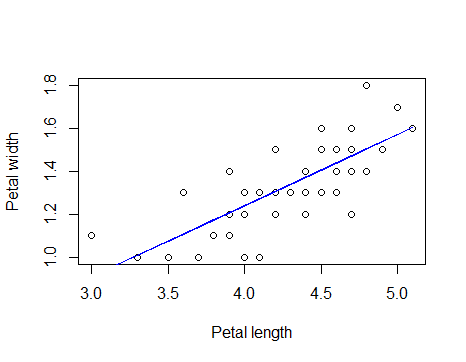


1. Calculate a liner model between petal length and width and show it in the scatter plot. **[0.2 Marks]**

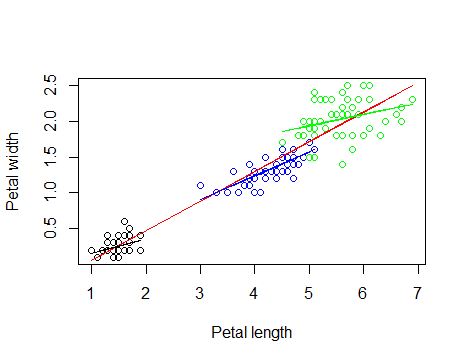


3. Based on the Species data, subdivide the iris dataset into three separate subsets (for each species).

4. Repeat steps 1 and 2 for each subsets. **[1.5 Marks]**



1. Plot all the results including scatter plots and linear models in one plot. The outcome should be similar to the following plot



3.6

> MSE

[1] 0.04206731

>

> error\_1<-model1$fitted.value-t1$Petal.Width

> MSE\_1<-mean(error\_1^2)

> MSE\_1

[1] 1.487158

>

> error\_2<-model2$fitted.value-t1$Petal.Width

> MSE\_2<-mean(error\_2^2)

> MSE\_2

[1] 0.6041109

>

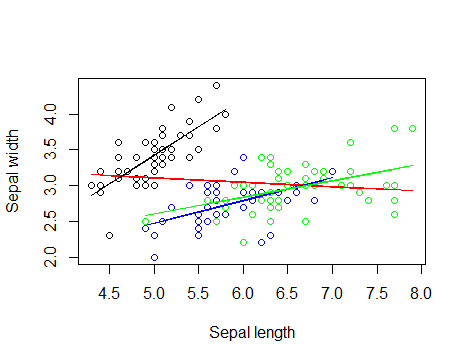
> error\_3<-model3$fitted.value-t1$Petal.Width

> MSE\_3<-mean(error\_3^2)

> MSE\_3

[1] 1.264536

**ALL SEPAL GRAPH**



> #3.6 for sepal

> x<-t1\_Sepal$Sepal.Length

> y<-t1\_Sepal$Sepal.Width

>

> mm<-lm(y ~ x)

> error<-model$fitted.values-y

> MSE\_SEPAL<-mean(error^2)

> MSE\_SEPAL

[1] 0.1861044

>

> error\_1<-model1$fitted.value-t1\_Sepal$Sepal.Width

> MSE\_SEPAL1<-mean(error\_1^2)

> MSE\_SEPAL1

[1] 0.3411616

>

> error\_2<-model2$fitted.value-t1\_Sepal$Sepal.Width

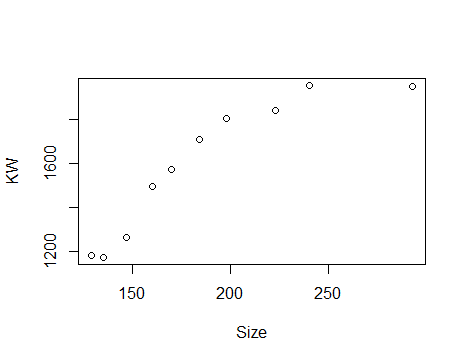
> MSE\_SEPAL2<-mean(error\_2^2)

> MSE\_SEPAL2

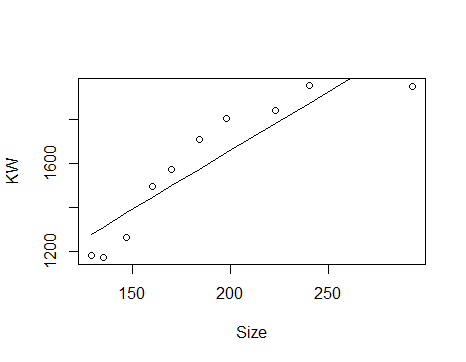
[1] 0.2861961

QUE 4:

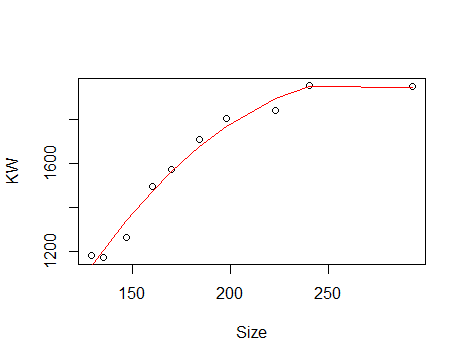
1. Create a scatter plot for the two variables. **[0.4 Marks]**

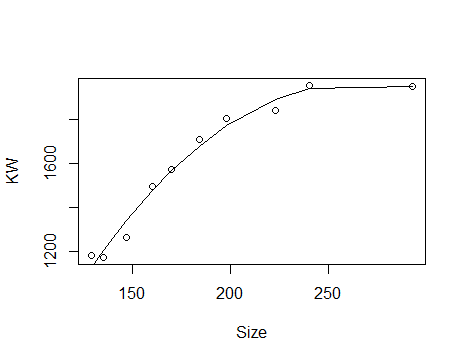
****

2. Calculate a linear regression model. **[0.2 Marks]**

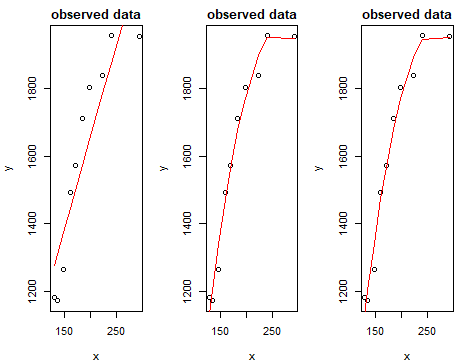


3. Calculate polynomial regression models of order 2 and 3. **[0.4 Marks]**





4. Plot the regression models (the result should be similar to the following figure) **[1 Mark]**



QUE 5 :

> z0 = (mean.machine1 -mean.machine2)/ (SP.Machine\*(sqrt((1/10) +( 1/10))))

> z0

[1] 0.7989355

Since t0 is 0.79835 which doesn't lie in the critical region, we accept the Null Hypothesis. X

QUE : 6

>

> L = 110 - (2.58 \* (10 /sqrt(30)))

> U = 110 + (2.58 \* (10 /sqrt(30)))

> L

[1] 105.2896

> U

[1] 114.7104

Based on value of z is around 5.05 and alfa value is less thaen given value so based on that u1 is not equal to u2 so we are rejecting the

hypotheseis