SHRESTHA AGRAWAL

THE SPARKS FOUNDATION

TASK1_GRIP_MAY21

TASK *1_GRIP_MAY21*

Prediction using Supervised ML

****

***AIM**: To predict the percentage of students based on the number of study hours.*

****

***CONCEPT**:*

***Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.*
#Simple linear regression is used to estimate the relationship between two quantitative variables.

You can use simple linear regression when you want to know:
1) How strong the relationship is between two variables.
2) The value of the dependent variable at a certain value of the independent

variable

Linear regression *attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.*

*>> **p-value**: A p-value indicates whether or not you can reject or accept a hypothesis*

*>> **R^2:** It is the coefficient of determination or R^2. Defined by the proportion of the total variability explained by the regression model.*

*A **linear regression line** has an equation of the form **Y = a + bX,***

Where,

X is the explanatory variable and

Y is the dependent variable.

The slope of the line is b, and a is the intercept (the value of y when x = 0).

****

DATA DESCRIPTION:

..The dataset gives the description of the relationship between the number of hours student studies in a day and the marks scored by them.

The dataset contains two variables i.e. Number of hours student studies and marks scored by them.

..The dataset has the information regarding 25 students.

Dependent variable = "Marks"

Independent variable ="Hours"

****

R-CODE AND ANALYSIS

To load the R code and import the dataset

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.2
```

```
Task1 <- read_excel("C:/Users/HP/Desktop/internship/Task1.xlsx")
View(Task1)
```

To get top 6 observations

```
head(Task1)
```

```
## # A tibble: 6 x 2
##   Hours Scores
##   <dbl>  <dbl>
## 1   2.5     21
## 2   5.1     47
## 3   3.2     27
## 4   8.5     75
## 5   3.5     30
## 6   1.5     20
```

TO get the summary of the dataset:

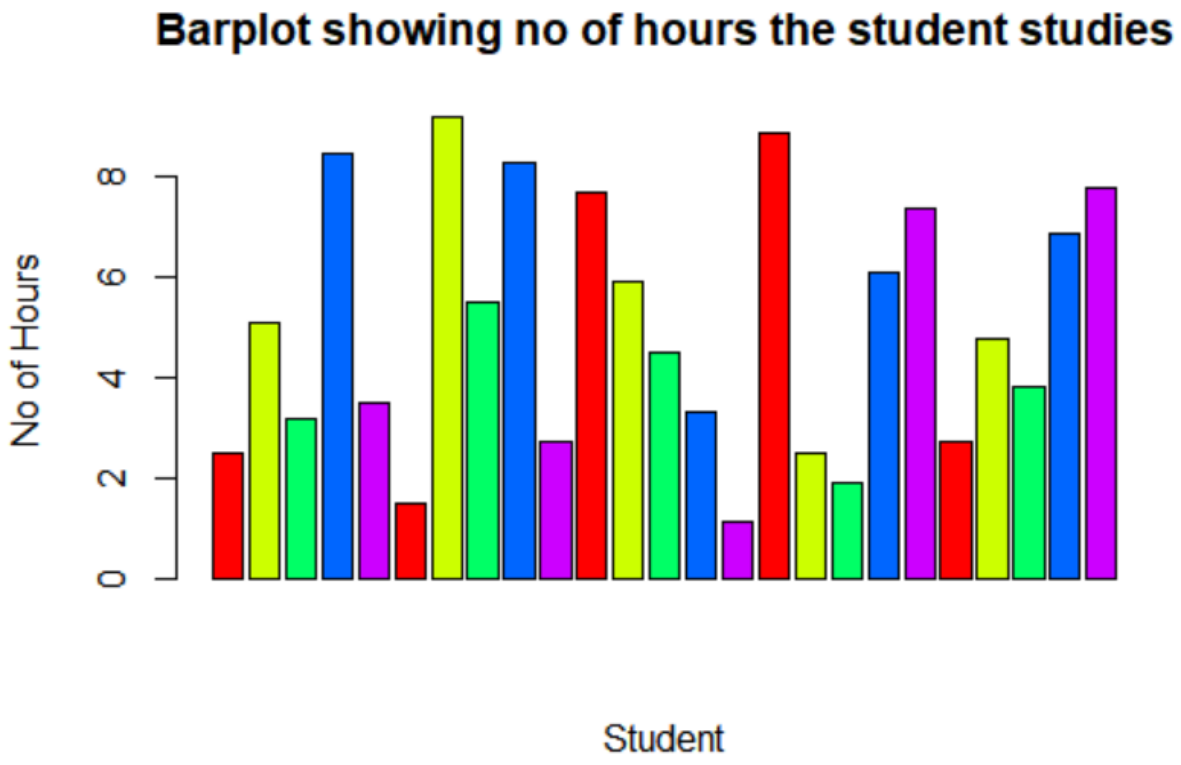
```
summary(Task1)
```

```
##      Hours           Scores     
##  Min.   :1.100   Min.   :17.00  
##  1st Qu.:2.700   1st Qu.:30.00  
##  Median :4.800   Median :47.00  
##  Mean   :5.012   Mean   :51.48  
##  3rd Qu.:7.400   3rd Qu.:75.00  
##  Max.   :9.200   Max.   :95.00
```

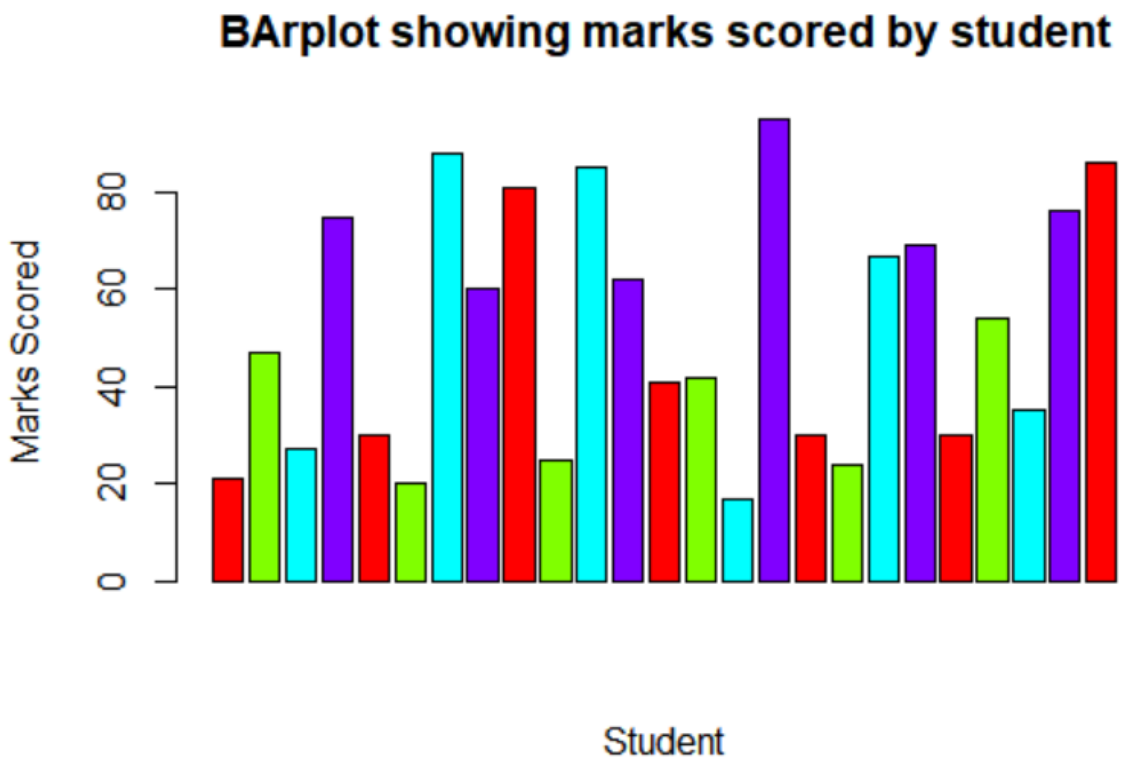
*# We get the average **number of hours** the student studies is **5.012 hrs**. The interval of hours within which the student studies is **[1.10, 9.20]***

*# We get the average **score** the student gets is **51.48**. The interval of score which the student scores is **[17.00, 95.00]***

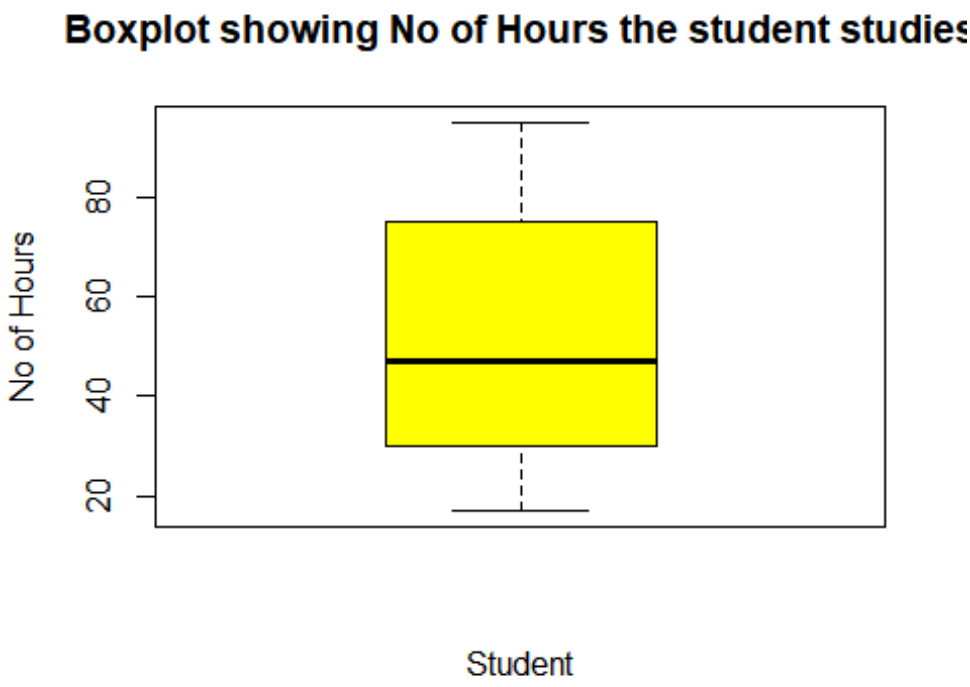
```r
#TO visualise the dataset
barplot(Task1$Hours, ylab="No of Hours", xlab="Student",main="Histogram
showing no of hours the student studies",col=rainbow(5.2))
```



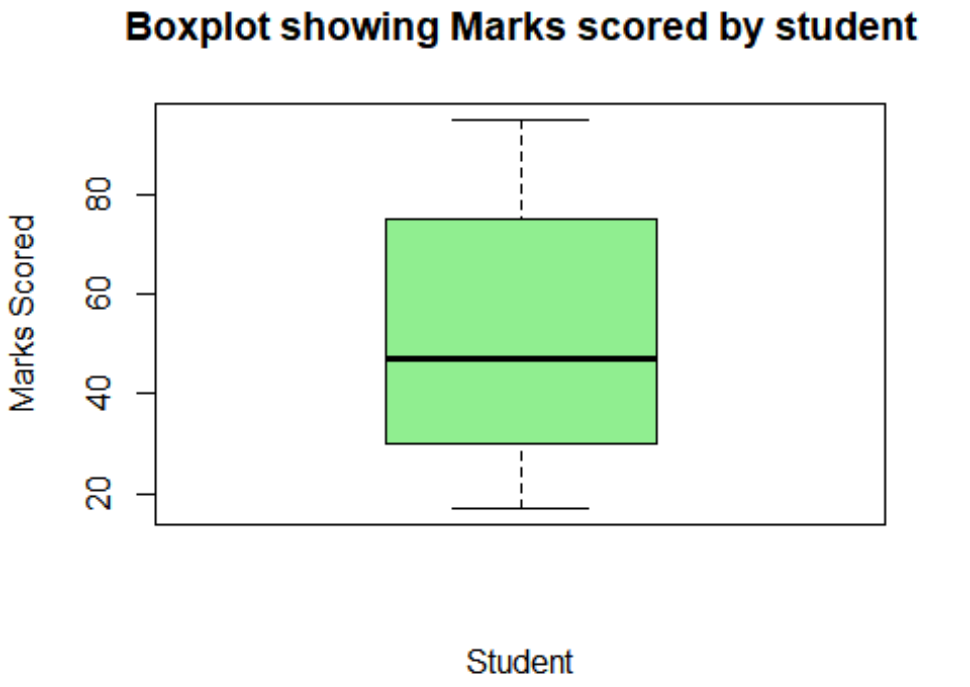

```r
barplot(Task1$Scores, ylab="Marks Scored",xlab="Student",main="Histogram
showing marks scored by student",col=rainbow(4.2))
```




```
boxplot(Task1$Scores, ylab="No of Hours",xlab="Student",main="Boxplot showing
No of Hours the student studies",col= "yellow")
```




```
boxplot(Task1$Scores, ylab="Marks Scored", xlab="Student",main="Boxplot
showing Marks scored by student",col="light green")
```



```
# We get the bargraph and the boxplot for the hours that the student devotes
and the marks scored by the students.
# The boxplot in itself gives the summary of the dataset.
```

Now we will check the regression analysis

```
 To plot the marks scored with scatter plot
plot(Task1,col='Green')
```

*To get the **regression model***

```
reg_model=lm(Task1$Scores~Task1$Hours,data=Task1)
reg_model
```

```
##
## Call:
## lm(formula = Task1$Scores ~ Task1$Hours, data = Task1)
##
## Coefficients:
## (Intercept)  Task1$Hours
##       2.484        9.776
```

```
summary(reg_model)
```

```
##
## Call:
## lm(formula = Task1$Scores ~ Task1$Hours, data = Task1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.578  -5.340   1.839   4.593   7.265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4837     2.5317   0.981    0.337
## Task1$Hours   9.7758     0.4529  21.583   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.603 on 23 degrees of freedom
## Multiple R-squared:  0.9529, Adjusted R-squared:  0.9509
## F-statistic: 465.8 on 1 and 23 DF,  p-value: < 2.2e-16
```

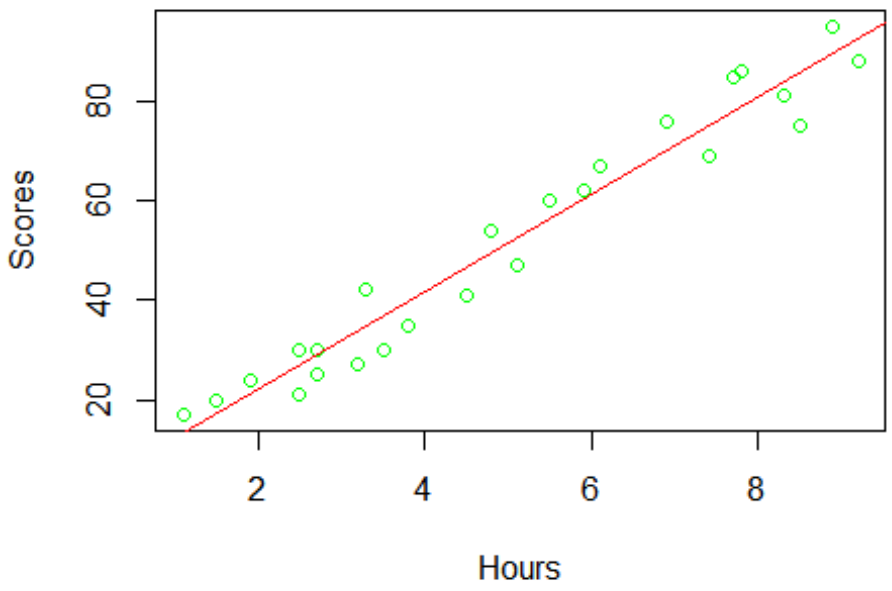
```
# The intercept of the model is 2.4837 and the coefficent for the variable
'Hours' is 9.7758.
# It is observed that the p-value of the varibale 'Hours' is below 0.05 which
```

```
means that it is benefitial to the model.
# The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

#The multiple R-sqaured is 0.9529 and adjusted R-sqaured is 0.9509~ 1.
# Multiple R-squared measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model. Adjusted R-squared adjusts the statistic based on the number of independent variables in the model
# R-squared is how well the regression model fits the observed data.
# Generally, a higher r-squared indicates a better fit for the model, here its 0.95 so it shows that its the best fit..

#The equation for the model is
#Marks_Scored = 2.4837+ 9.7758 * Hour_Studied

#To plot the regression line
abline(reg_model, col="red")
```




```
#From the graph we can see that the points lie very close and in sync with
the regression line. It shows that there
# exists a direct relation between the marks scored and the number of hours a
student studies.

#Checking the efficacy of our model with the given data.
Hours=data.frame(c(2.5,5.1,3.2,8.5,3.5,1.5,9.2,5.5,8.3,2.7,7.7,5.9,4.5,3.3,1.
1,8.9,2.5,1.9,6.1,7.4,2.7,4.8,3.8,6.9,7.8))
Check_Data=predict(reg_model,newdata = Hours)
comparison=data.frame(Check_Data,Task1$Scores)
comparison
```

```
##    Check_Data Task1.Scores
## 1    26.92318           21
## 2    52.34027           47
## 3    33.76624           27
## 4    85.57800           75
## 5    36.69899           30
## 6    17.14738           20
## 7    92.42106           88
## 8    56.25059           60
```

```
## 9    83.62284           81
## 10   28.87834           25
## 11   77.75736           85
## 12   60.16091           62
## 13   46.47479           41
## 14   34.74382           42
## 15   13.23706           17
## 16   89.48832           95
## 17   26.92318           30
## 18   21.05770           24
## 19   62.11607           67
## 20   74.82462           69
## 21   28.87834           30
## 22   49.40753           54
## 23   39.63173           35
## 24   69.93672           76
## 25   78.73494           86
```

#The scores that we obtain are pretty close to what was given to us. So we can say that our model is efficient.

#What will be predicted score if a student studies for 9.25 hrs/ day?

```
Hrs=9.25
Predicted_Score=2.4837+ 9.7758 * Hrs
Predicted_Score
```

```
## [1] 92.90985
```

#We get the score as 92.90985 when the student studies for 9.25hrs/day.

#

*#**

#INTERPRETATION:

*# **1)** We get the average number of hours the student studies is 5.012 hrs. The interval of hours within which the student studies is [1.10, 9.20]*

*# **2)** We get the average score the student gets is 51.48. The interval of score which the student scores is [17.00, 95.00]*

*# **3)** The regression model of this data is*

Marks_Scored = 2.4837+ 9.7758 * Hour_Studied

*# **4)** A student will get **92.90985 marks** if he/she studies for 9.25 hrs/day.*