Task-1.R

Srikar R

2021-05-05

#OBJECTIVE: *To predict the percentage of a student based on the no. of study hours.*

#DATA DESCRIPTION: *The given dataset describes the number of hours a student studied and the mask scored by them. The dataset contains two #varibales i.e. Hours and Scores. It contains the data of 25 students.*

#Here the dependent variable is "Marks" (marks scored) and the independent variable is "Hours" (Number of hours studied)

```
library(readxl)

data1=read_excel("C:\\Users\\Srikar\\Desktop\\Study stuff\\Internship\\Grips Foundation\\Task 1\\dataset.xlsx")  #Importing dataset
head(data1)   # Viewing first 6 observations
```

```
## # 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
```

#SUMMARY STATISTICS:

```
summary(data1)
```

```
##      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
```

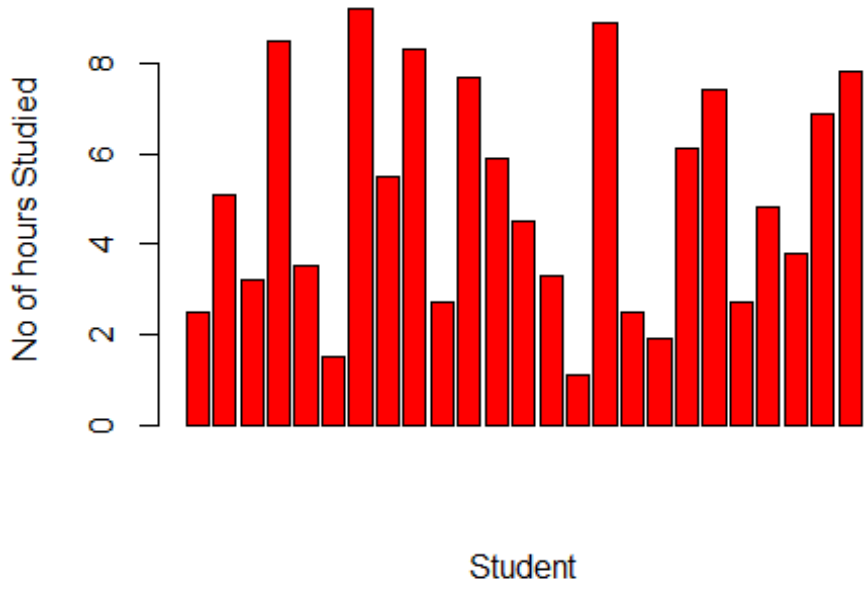
```
##  3rd Qu.:7.400   3rd Qu.:75.00
##  Max.   :9.200   Max.   :95.00
```

#The minimum hours any student student studied in the dataset is 1 hour and 10 minutes and the maximum hoursa student studied is for 9.2 hours.
#The average hours the entire population studied is apporximately 5 hours.

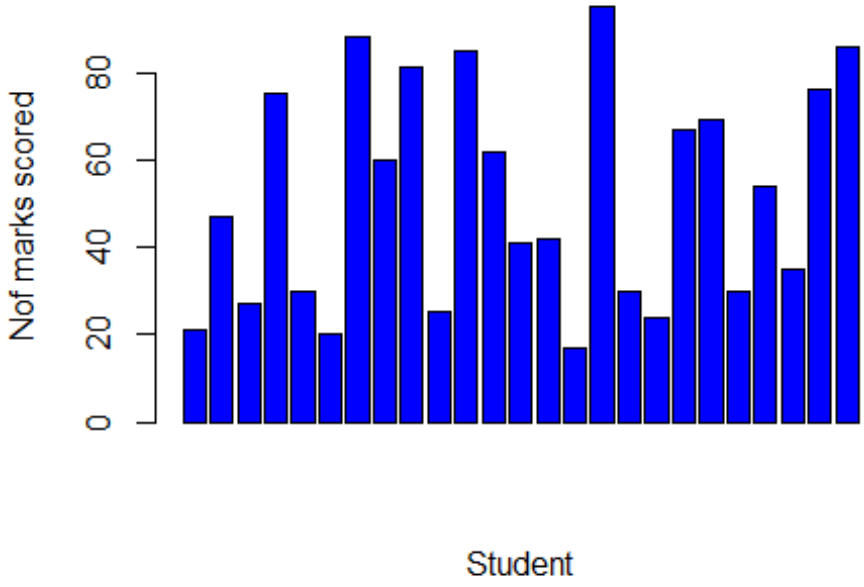
#The least marks scored by a student is 17 and the maximum marks scored by a student is 95.The average score of the class is 51.48.

#DATA VISUALIZATION:

```
barplot(data1$Hours,col = 'red',xlab = 'Student',ylab='No of hours Studied')
```




```
barplot(data1$Scores,col= 'blue',xlab='Student',ylab='Nof marks scored')
```



#The bargraphs show the variablity in data. The number of studied and the marks scored by students differ from each student.

#RELEVANT TERMS:

#I) Linear Regression- A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their
#interactions (often called x or explanatory variables).

#II)p-value-A p-value indicates whether or not you can reject or accept a hypothesis

#II)R-Sqaured- It is the coefficient of determination or R^2. This measure is defined by the proportion of the
#total variability explained by the regression model.

#R^2= (Explained Variation of the model)/ (Total variation of the model)

#ANALYSIS:

```
#1) Plotting the distribution of scores through scatter plot

plot(data1,col='purple')

#2) Constructing the regression model

Reg=lm(data1$Scores~data1$Hours,data=data1)
Reg

##
## Call:
## lm(formula = data1$Scores ~ data1$Hours, data = data1)
##
## Coefficients:
## (Intercept)  data1$Hours
##       2.484        9.776

summary(Reg)

##
## Call:
## lm(formula = data1$Scores ~ data1$Hours, data = data1)
##
## 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
## data1$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 'score' is
9.7758. It is observed that the
```

 # p-value of the varibale 'scores' is below 0.05 which means that it's an excellent addition to the model.

#The multiple R-sqaured is 0.9529 and adjusted R-sqaured is 0.9509 which are very close to 1. This means that it is explains 95% of
#the variability.

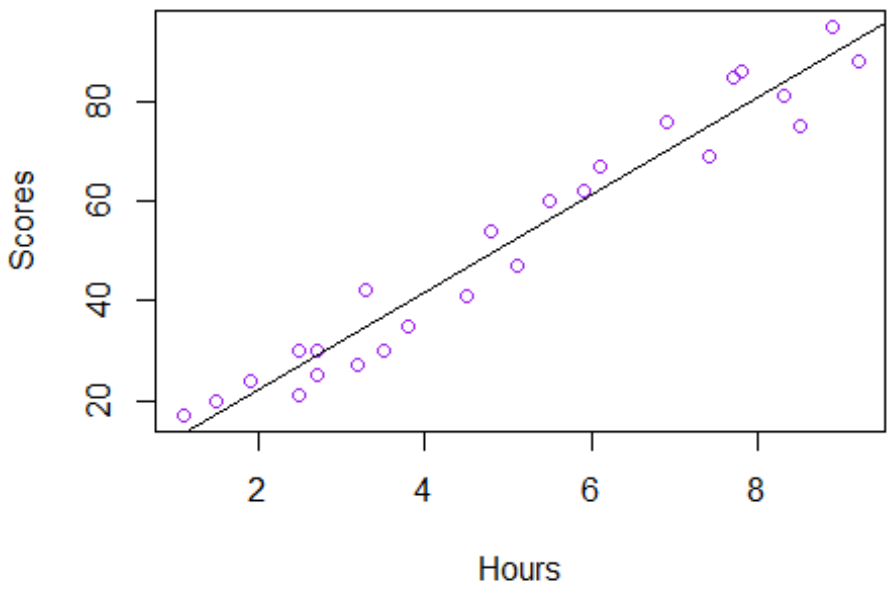
#The eqaution of the model is

*# Marks = 2.4837+ 9.7758 * X*

#Here X is the value of the variable, "Hours"

#3)Plotting the regression line

abline(Reg)



#We see that the data plot are very close to the regression line. This tells us that the residuals of the plot are very close to the line
 # which will be easy to predict the value as error is low.

#4)Testing the model with past 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))
Test.data=predict(Reg,newdata = hours)

comparison=data.frame(Test.data,data1$Scores)
comparison
```

```
##    Test.data data1.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
```

#We observe that values predicted by the model and the actual data are very close.
Using the same model we can predict the amount of
#marks a student will score if they studied for 9 hours and 25 minutes.

#5) Predicting score if student studied for 9.25 hours

```
P=2.4837+ 9.7758 * 9.25
P
```

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

#CONCLUSION:

*#I) The regression model of this data is 2.4837+ 9.7758 * X where X is the number of hours studied.*

#II) If a student studies for 9.25 hours then the predicted score would be 92.90 according to the regression model.