TSF Task 1

Hrithik Agarwal

08/01/2021

Prediction using supervised ML (Linear Regression)

Importing the required libraries

```
library(ggplot2)
library(GGally)
```

Importing and Reading Data

```
d <- read.csv('https://raw.githubusercontent.com/AdiPersonalWorks/Random/master/student_scores%20-%20stu
```

```
dim(d)
```

```
## [1] 25  2
```

```
head(d)
```

```
##   Hours Scores
## 1   2.5     21
## 2   5.1     47
## 3   3.2     27
## 4   8.5     75
## 5   3.5     30
## 6   1.5     20
```

Cheacking for missing data

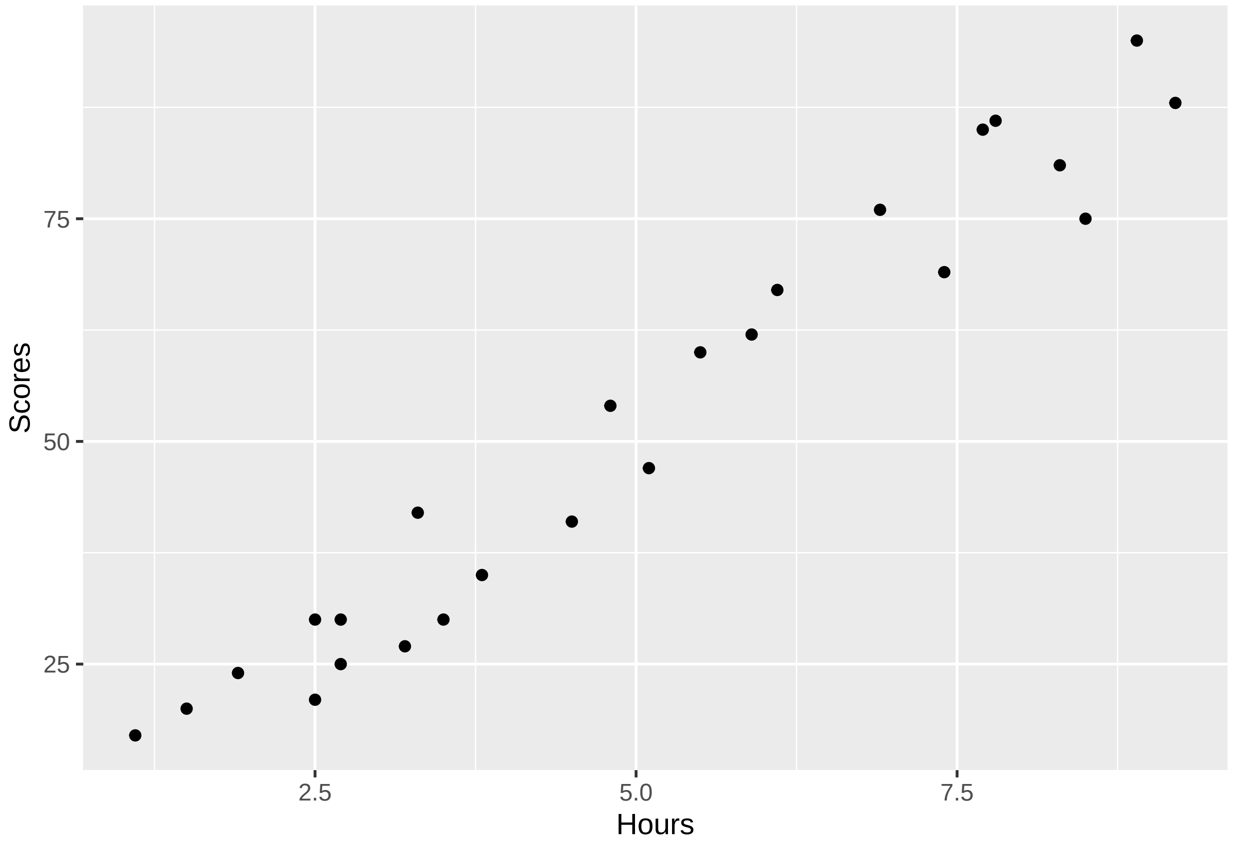
```
sum(is.na(d))
```

```
## [1] 0
```

Thus there is no Missing data.

Plotting the Data

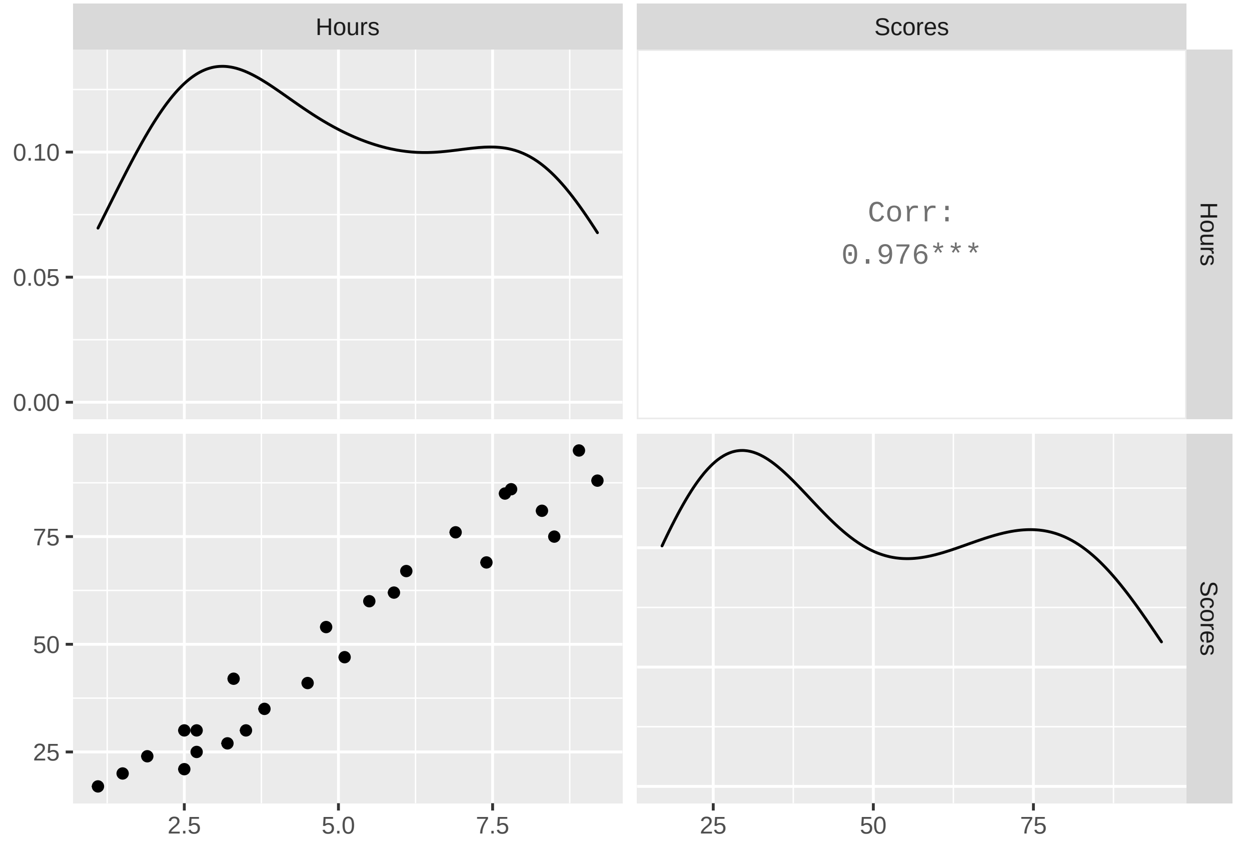
```
ggplot(data = d,aes(y = Scores , x = Hours)) + geom_point()
```



From the graph we can observe that there is no future engineering reuired in the data.

Correlation between the variables

```
ggpairs(data = d , columns = 1:2)
```



The Correlation is **0.976** which is very significant.

Training the model

```
model <- lm(Scores ~ Hours, data = d)
```

Summary of Model

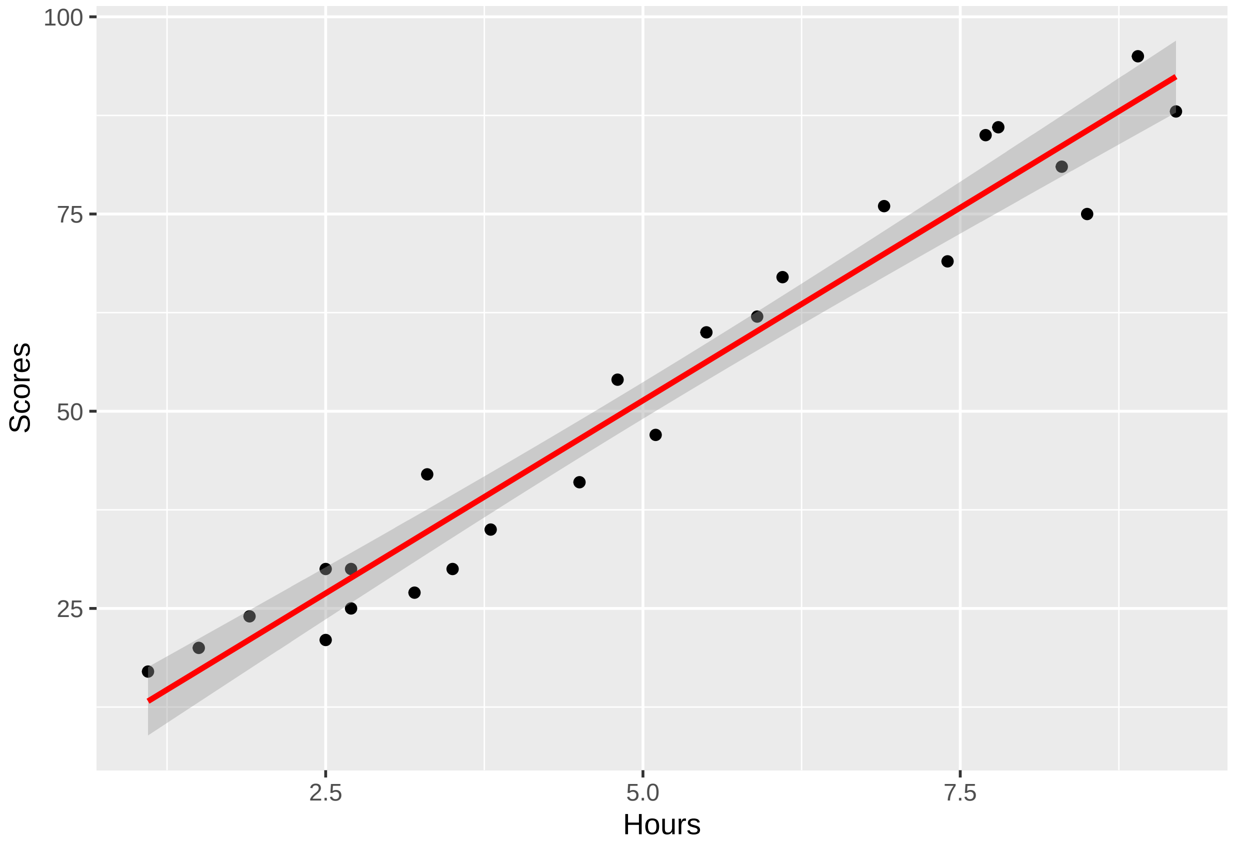
```
summary(model)
```

```
##
## Call:
## lm(formula = Scores ~ Hours, data = d)
##
## 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
## 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
```

Since **Pr(>|t|)** « 1, it indicates that model is highly significant.

Plotting the regression Line

```r
ggplot(data = d,aes(y = Scores , x = Hours)) + geom_point()+geom_smooth(method = ’lm’, color = ’red’)
```



Prediction on Our Dataset

```r
d2 <- data.frame(d$Hours)
names(d2)[1]<- ’Hours’
pred <- predict(model , )
pred <- as.numeric(pred)
pred
```

```
##  [1] 26.92318 52.34027 33.76624 85.57800 36.69899 17.14738 92.42106 56.25059
##  [9] 83.62284 28.87834 77.75736 60.16091 46.47479 34.74382 13.23706 89.48832
## [17] 26.92318 21.05770 62.11607 74.82462 28.87834 49.40753 39.63173 69.93672
## [25] 78.73494
```

Comparing the Actual and Predicted Scores.

```
actual <- as.numeric(d$Scores)
data.frame(actual,pred)
```

```
##    actual     pred
## 1      21 26.92318
## 2      47 52.34027
## 3      27 33.76624
## 4      75 85.57800
## 5      30 36.69899
## 6      20 17.14738
## 7      88 92.42106
## 8      60 56.25059
## 9      81 83.62284
## 10     25 28.87834
## 11     85 77.75736
## 12     62 60.16091
## 13     41 46.47479
## 14     42 34.74382
## 15     17 13.23706
## 16     95 89.48832
## 17     30 26.92318
## 18     24 21.05770
## 19     67 62.11607
## 20     69 74.82462
## 21     30 28.87834
## 22     54 49.40753
## 23     35 39.63173
## 24     76 69.93672
## 25     86 78.73494
```

Predicting for Hours = 9.25

```
predict(model, data.frame(Hours = 9.25))
```

```
##        1
## 92.90985
```

Hence our model Predicts Score of 92.90985