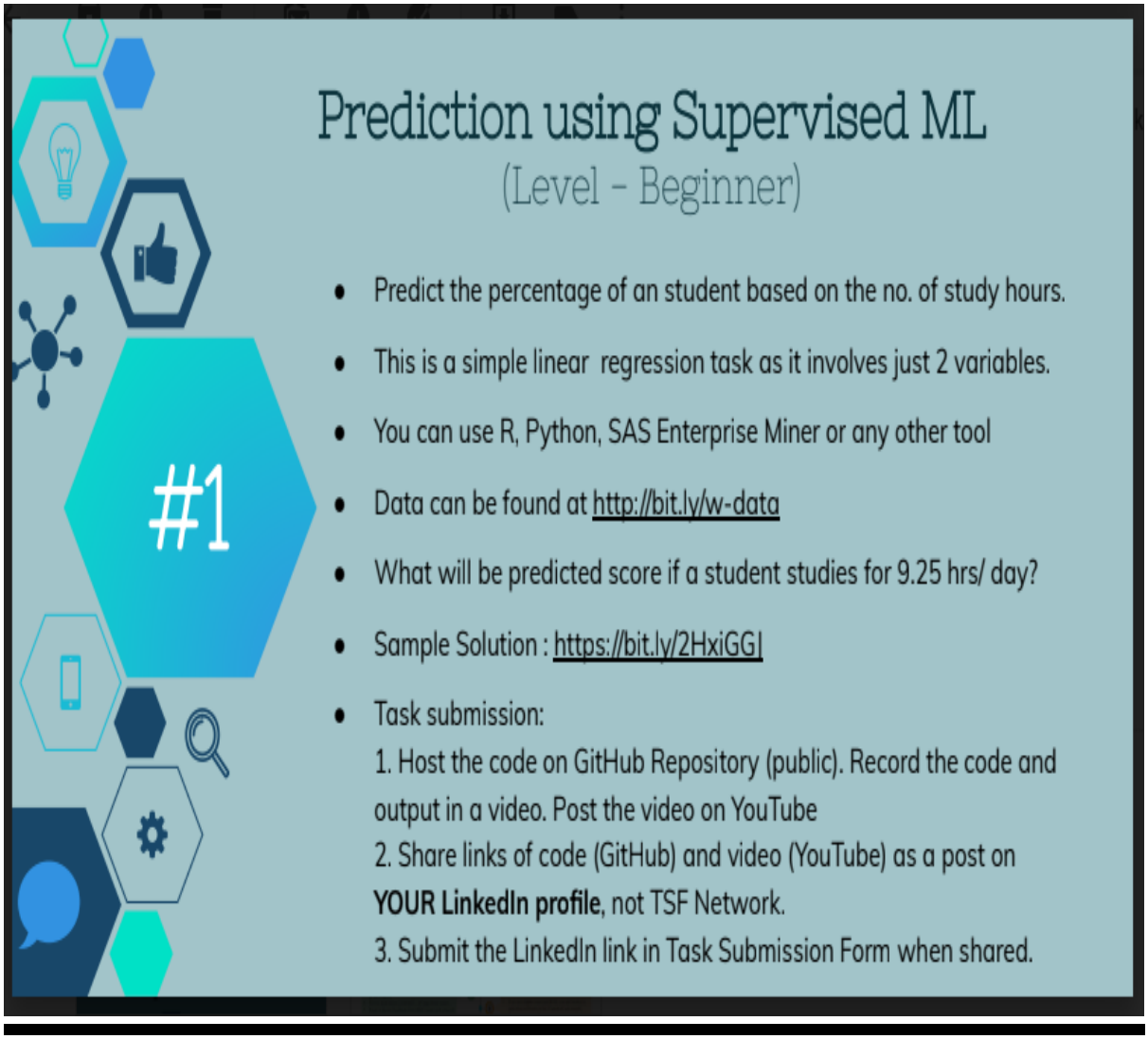
Graduate Rotational Internship Program

Author:Kartikey Gupta

The Sparks Foundation

TASK- 1


#1
Prediction using Supervised ML
(Level - Beginner)
• Predict the percentage of an student based on the no. of study hours.
• This is a simple linear regression task as it involves just 2 variables.
• You can use R, Python, SAS Enterprise Miner or any other tool
• Data can be found at http://bit.ly/w-data
• What will be predicted score if a student studies for 9.25 hrs/ day?
• Sample Solution : https://bit.ly/2HxiGGJ
• Task submission:
1. Host the code on GitHub Repository (public). Record the code and output in a video. Post the video on YouTube
2. Share links of code (GitHub) and video (YouTube) as a post on YOUR LinkedIn profile, not TSF Network.
3. Submit the LinkedIn link in Task Submission Form when shared.

PROVIDED DATA

Hours	Scores
2.5	21
5.1	47
3.2	27
8.5	75
3.5	30
1.5	20
9.2	88
5.5	60
8.3	81
2.7	25
7.7	85
5.9	62
4.5	41
3.3	42
1.1	17
8.9	95
2.5	30
1.9	24
6.1	67
7.4	69
2.7	30
4.8	54
3.8	35
6.9	76
7.8	86

R Programming code

######Data Science and Business Analytics######

###Task-1:"Prediction Using Supervised Machine Learning"###

######Author: Kartikey Gupta######

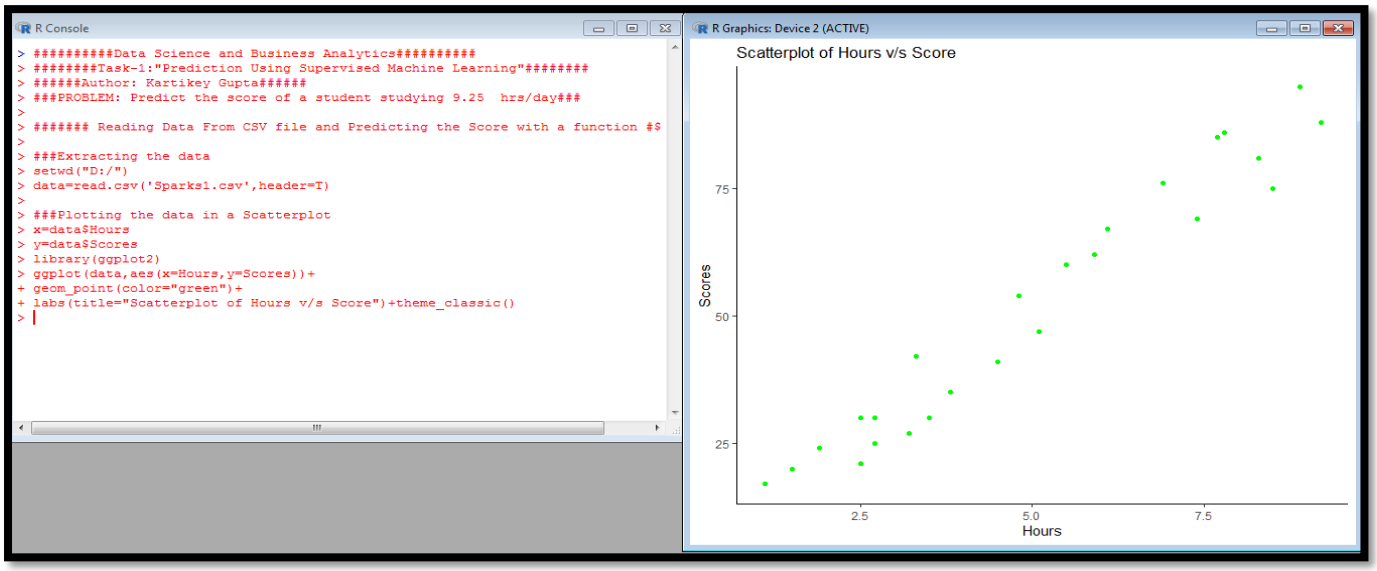
###PROBLEM: Predict the score of a student studying 9.25 hrs/day###

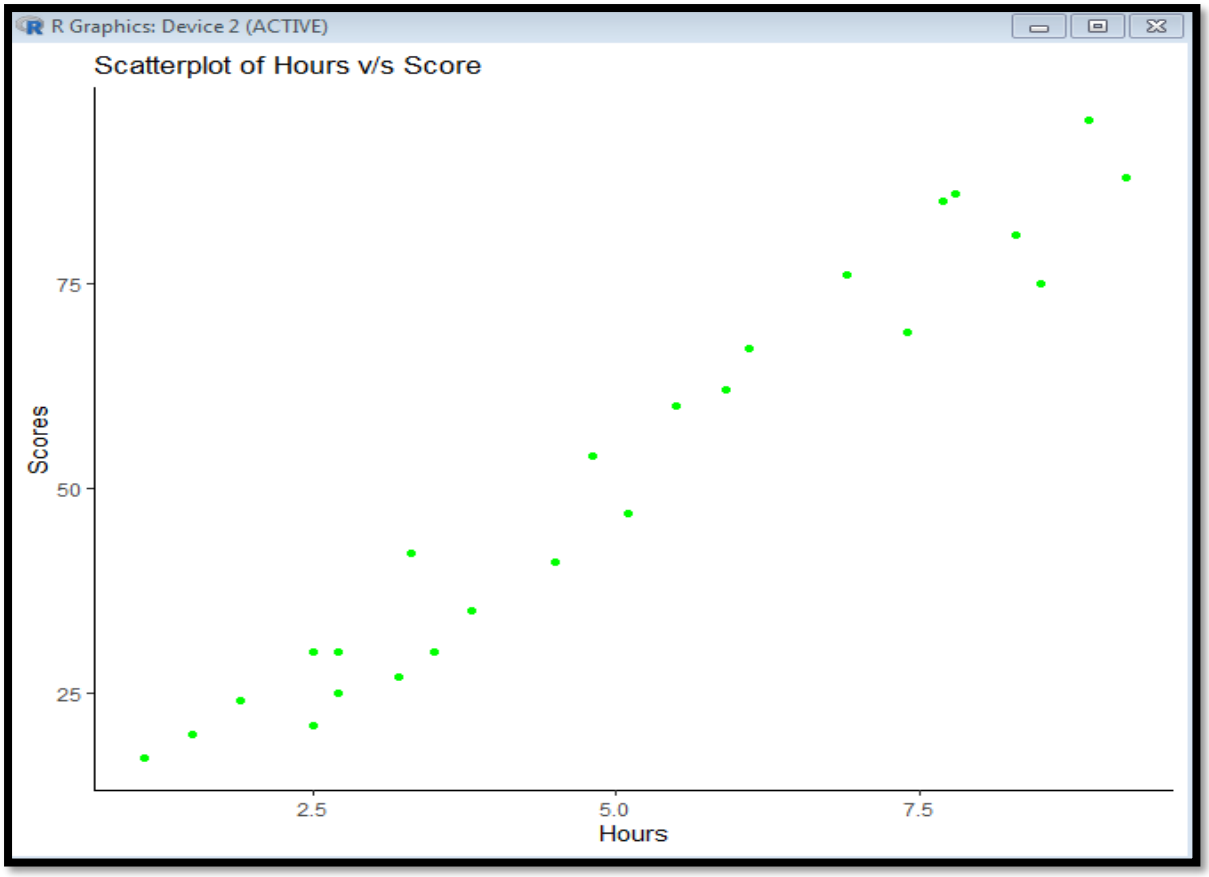
######## Reading Data from CSV file and Predicting the Score with a function ########

```
###Extracting the data###
setwd("D:/")
data=read.csv('Sparks1.csv',header=T)

###Plotting the data in a Scatterplot###
x=data$Hours
y=data$Scores
library(ggplot2)
ggplot(data,aes(x=Hours,y=Scores))+
geom_point(color="green")+
```

labs(title="Scatterplot of Hours v/s Score")+theme_classic()





###Fitting a linear regression model with an intercept term###

fit1=lm(y~x)

summary(fit1)

```
> ###Fitting a linear regression model with an itercept term
> fit1=lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

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
x             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
```

###Fitting a linear regression model with no intercept

fit2=lm(y~x-1)

summary(fit2)

```
> ###Fitting a linear regression model with no intercept
> fit2=lm(y~x-1)
> summary(fit2)

Call:
lm(formula = y ~ x - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-11.481  -4.784   2.530   4.937   8.425

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x  10.1743     0.2003   50.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.599 on 24 degrees of freedom
Multiple R-squared:  0.9908,    Adjusted R-squared:  0.9904
F-statistic:  2579 on 1 and 24 DF,  p-value: < 2.2e-16
```

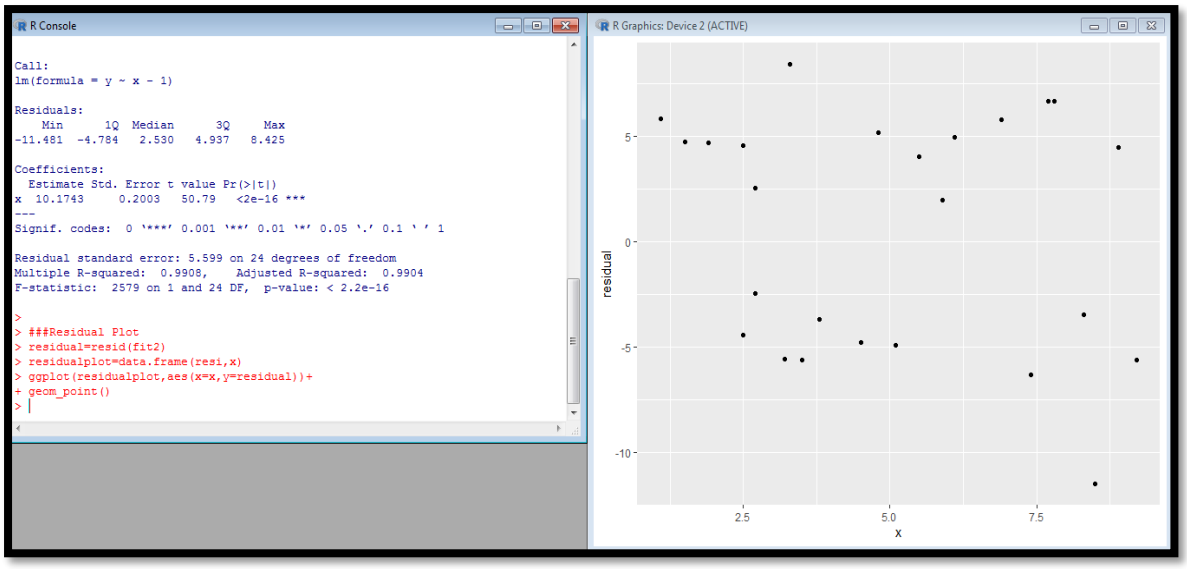
###Residual Plot###

residual=resid(fit2)

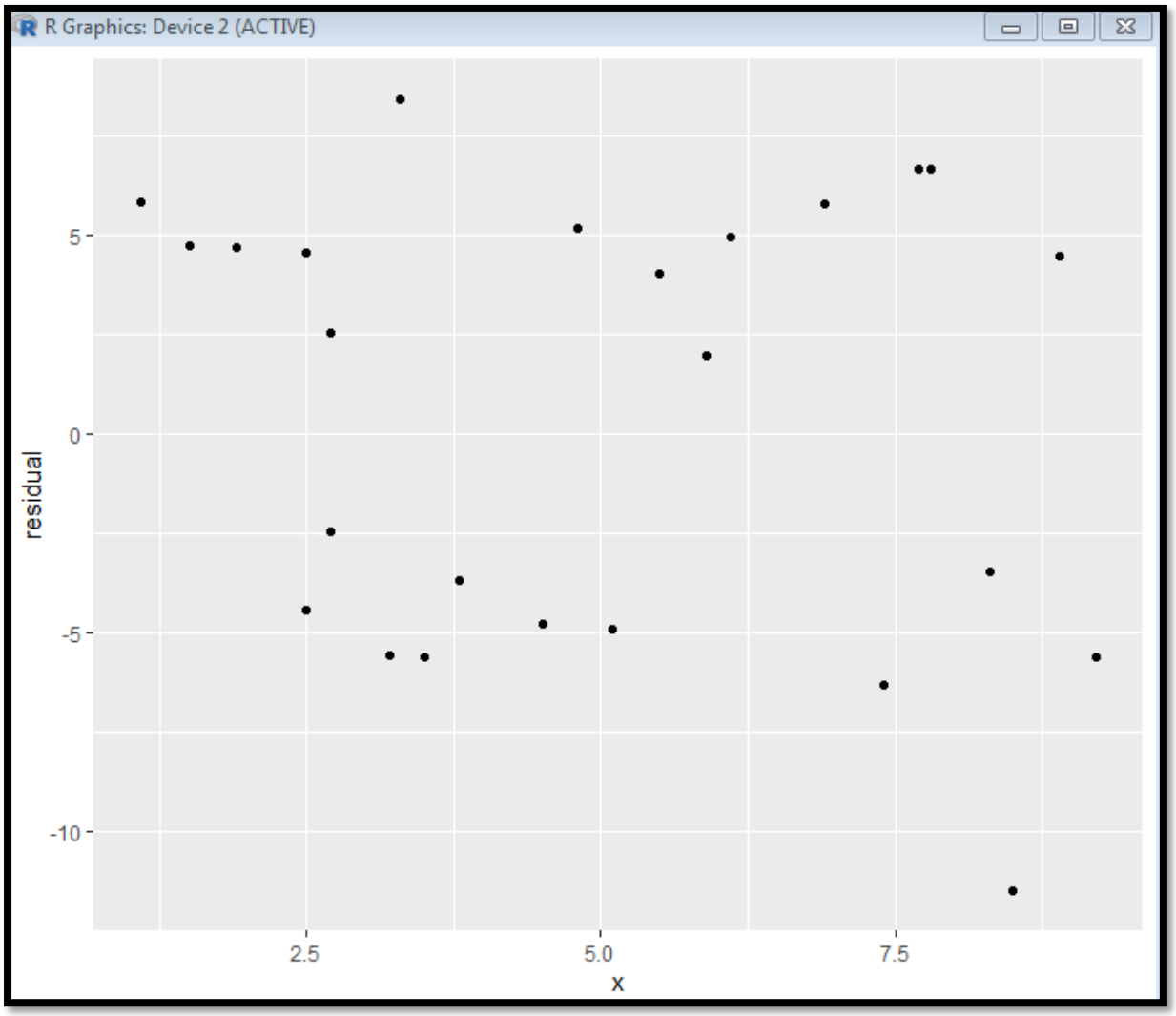
residualplot=data.frame(resid(fit2),x)

ggplot(residualplot,aes(x=x,y=residual))+

geom_point()


R Console
Call:
lm(formula = y ~ x - 1)
Residuals:
Min 1Q Median 3Q Max
-11.481 -4.784 2.530 4.937 8.425
Coefficients:
Estimate Std. Error t value Pr(>|t|)
x 10.1743 0.2003 50.79 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.599 on 24 degrees of freedom
Multiple R-squared: 0.9908, Adjusted R-squared: 0.9904
F-statistic: 2579 on 1 and 24 DF, p-value: < 2.2e-16
> ###Residual Plot
> residual=resid(fit2)
> residualplot=data.frame(resi,x)
> ggplot(residualplot,aes(x=x,y=residual))+
+ geom_point()
R Graphics: Device 2 (ACTIVE)
residual
x
5
0
-5
-10
2.5
5.0
7.5


R Graphics: Device 2 (ACTIVE)
residual
x
5
0
-5
-10
2.5
5.0
7.5

```
####Fitted Line PLot and Prediction####

PredictScr=function(tm)
{

  sdata=read.csv('Sparks1.csv',header=T)
  x=sdata$Hours
  y=sdata$Scores
  b=coefficients(lm(y~x))

  pred=b[1]+(b[2]*tm)

  plot(x,y,main="Regression Line Plot
",xlab="Hours",ylab="Scores","p",col="green")
  abline(lm(y~x),col='red')
cat('The Predicted Score for hour of study : ',tm,' is : ')

  return(as.numeric(pred))
}

PredictScr(9.25)
```

```
> ####Fitted Line PLot and Prediction
>
> PredictScr=function(tm)
+ {
+
+   sdata=read.csv('Sparks1.csv',header=T)
+   x=sdata$Hours
+   y=sdata$Scores
+   b=coefficients(lm(y~x))
+
+   pred=b[1]+(b[2]*tm)
+
+   plot(x,y,main="Regression Line Plot ",xlab="Hours",ylab="Scores","p",col="green")
+   abline(lm(y~x),col='red')
+
+   cat('The Predicted Score for hour of study : ',tm,' is : ')
+
+   return(as.numeric(pred))
+ }
>
> PredictScr(9.25)
The Predicted Score for hour of study :  9.25  is : [1] 92.90985
```