SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

CSE 240 R Data Science with R

Project Report

Name : Bhuvann V

Unique ID : E0420018

Year : II

Quarter : Q2

Department : B. Tech MED

Faculty Name : Prof. Ramya M

Academic Year : 2020-2021

CONTENTS

Students Performance in Exam

PROBLEM STATEMENT

This dataset contains the students details and the marks of a particular subject based on their performance. We need to predict the mark of the student based on the marks which was awarded to them based on their skills

OBJECTIVE :

The main objective of this project:

To outline contains the marks of the student after completing an exam and estimated marks based on their skill.

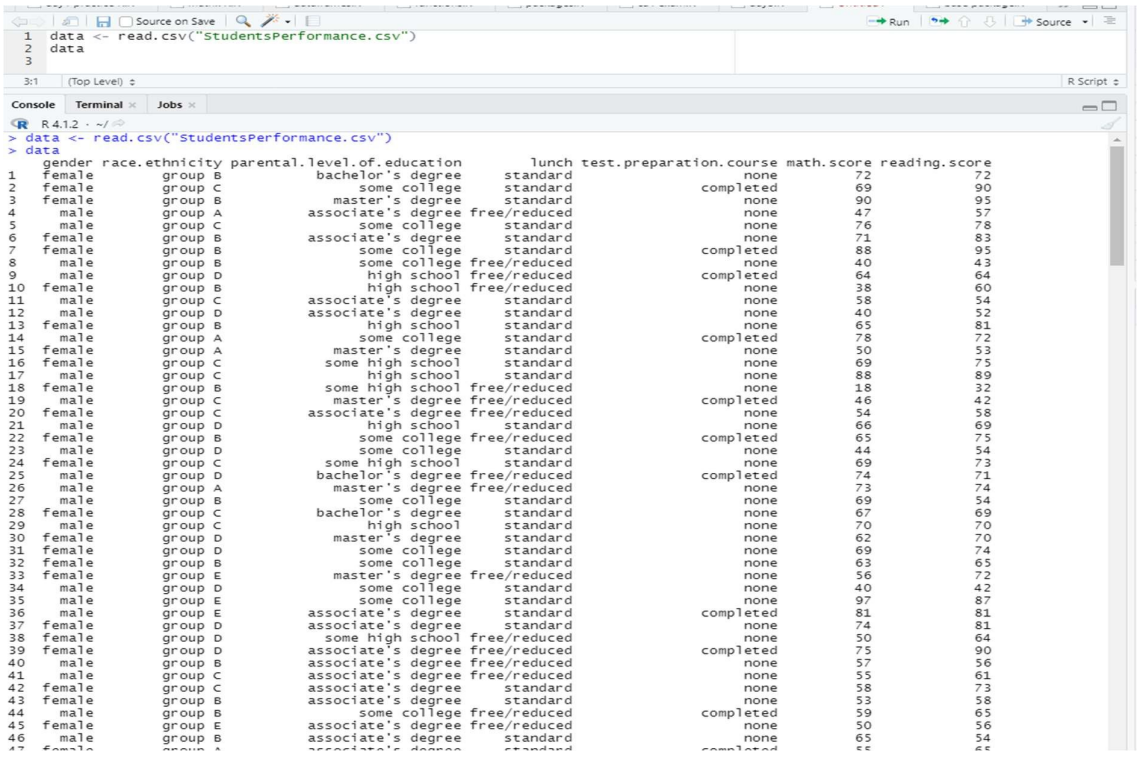
To select appropriate prediction model algorithm that combines

several features to estimate the mark of the student.

To track the mark of the student based on correct model.

To forecast and visualize the mark obtained based on the dataset

DATA LOADING



```
data <- read.csv("StudentsPerformance.csv")
data
```

```
> data <- read.csv("StudentsPerformance.csv")
> data
```

	gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	math.score	reading.score
1	female	group B	bachelor's degree	standard	none	72	72
2	female	group C	some college	standard	completed	69	90
3	female	group B	master's degree	standard	none	90	95
4	male	group A	associate's degree	free/reduced	none	47	57
5	male	group C	some college	standard	none	76	78
6	female	group B	associate's degree	standard	none	71	83
7	female	group B	some college	standard	completed	88	95
8	male	group B	some college	free/reduced	none	40	43
9	male	group D	high school	free/reduced	completed	64	64
10	female	group B	high school	free/reduced	none	38	60
11	male	group C	associate's degree	standard	none	58	54
12	male	group D	associate's degree	standard	none	40	52
13	female	group B	high school	standard	none	65	81
14	male	group A	some college	standard	completed	78	72
15	female	group A	master's degree	standard	none	50	53
16	female	group C	some high school	standard	none	69	75
17	male	group C	high school	standard	none	88	89
18	female	group B	some high school	free/reduced	none	18	32
19	male	group C	master's degree	free/reduced	completed	46	42
20	female	group C	associate's degree	free/reduced	none	54	58
21	male	group D	high school	standard	none	66	69
22	female	group B	some college	free/reduced	completed	65	75
23	male	group D	some college	standard	none	44	54
24	female	group C	some high school	standard	none	69	73
25	male	group D	bachelor's degree	free/reduced	completed	74	71
26	male	group A	master's degree	free/reduced	none	73	74
27	male	group B	some college	standard	none	69	54
28	female	group C	bachelor's degree	standard	none	67	69
29	male	group C	high school	standard	none	70	70
30	female	group D	master's degree	standard	none	62	70
31	female	group D	some college	standard	none	69	74
32	female	group B	some college	standard	none	63	65
33	female	group E	master's degree	free/reduced	none	56	72
34	male	group D	some college	standard	none	40	42
35	male	group E	some college	standard	none	97	87
36	male	group E	associate's degree	standard	completed	81	81
37	female	group D	associate's degree	standard	none	74	81
38	female	group D	some high school	free/reduced	none	50	64
39	female	group D	associate's degree	free/reduced	completed	75	90
40	male	group B	associate's degree	free/reduced	none	57	56
41	male	group C	associate's degree	free/reduced	none	55	61
42	female	group C	associate's degree	standard	none	58	73
43	female	group B	associate's degree	standard	none	53	58
44	male	group B	some college	free/reduced	completed	59	65
45	female	group E	associate's degree	free/reduced	none	50	56
46	male	group B	associate's degree	standard	none	65	54

```
R 4.1.2 · ~/
> data <- read.csv("StudentsPerformance.csv")
> head(data)
  gender race.ethnicity parental.level.of.education       lunch test.preparation.course math.score reading.score
1 female        group B           bachelor's degree    standard                    none         72            72
2 female        group C                some college    standard               completed         69            90
3 female        group B             master's degree    standard                    none         90            95
4   male        group A          associate's degree free/reduced                    none         47            57
5   male        group C                some college    standard                    none         76            78
6 female        group B          associate's degree    standard                    none         71            83
  writing.score
1            74
2            88
3            93
4            44
5            75
6            78
>
```

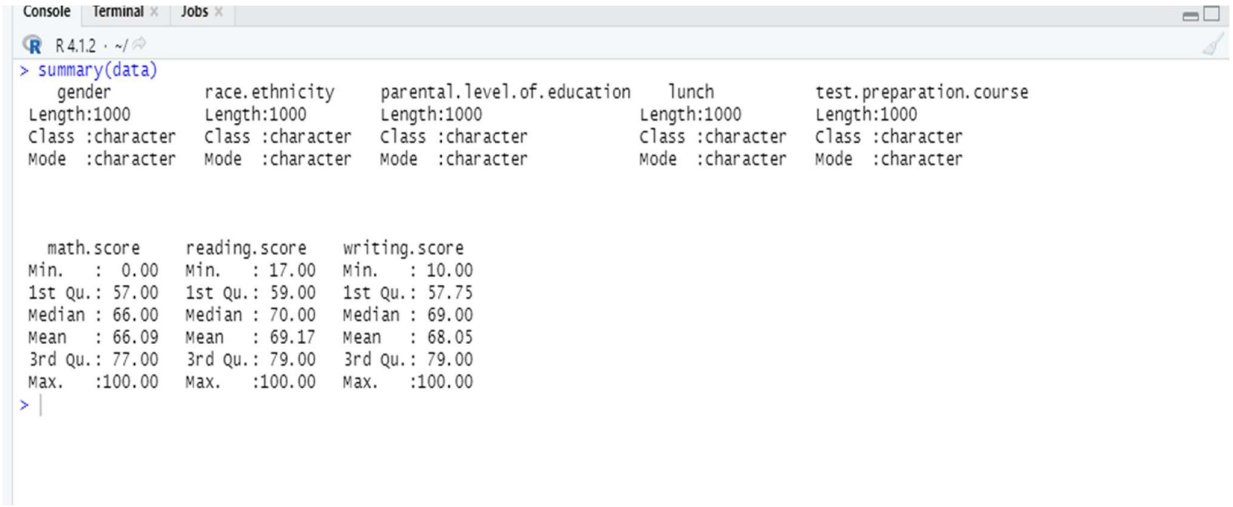
THIS DATASET IS STUDENT PERFORMANCE DATASET, WHICH IS SAVED AS A CSV FILE, LOADED AND VIEWED IN THE WORKING ENVIRONMENT

DATA EXPLORATION

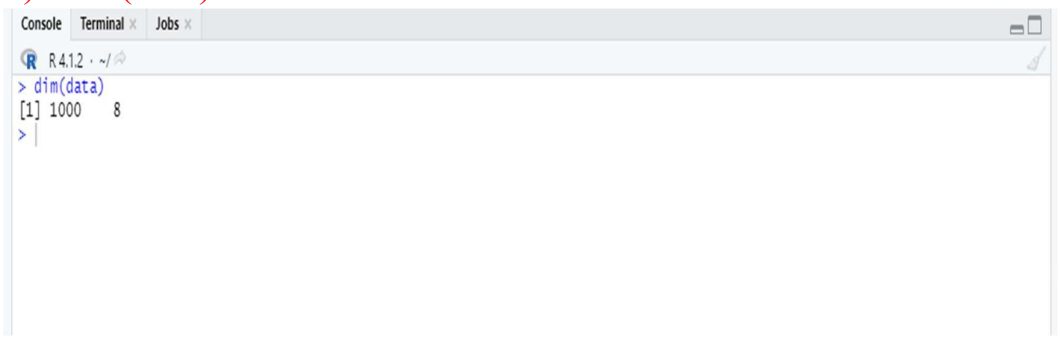
1) Str(data)

```
Console  Terminal  Jobs
R 4.1.2 · ~/
> str(data)
'data.frame':   1000 obs. of  8 variables:
 $ gender                    : chr  "female" "female" "female" "male" ...
 $ race.ethnicity            : chr  "group B" "group C" "group B" "group A" ...
 $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree" "associate's degree" ...
 $ lunch                     : chr  "standard" "standard" "standard" "free/reduced" ...
 $ test.preparation.course   : chr  "none" "completed" "none" "none" ...
 $ math.score                : int  72 69 90 47 76 71 88 40 64 38 ...
 $ reading.score             : int  72 90 95 57 78 83 95 43 64 60 ...
 $ writing.score             : int  74 88 93 44 75 78 92 39 67 50 ...
>
```

2) summary(data)

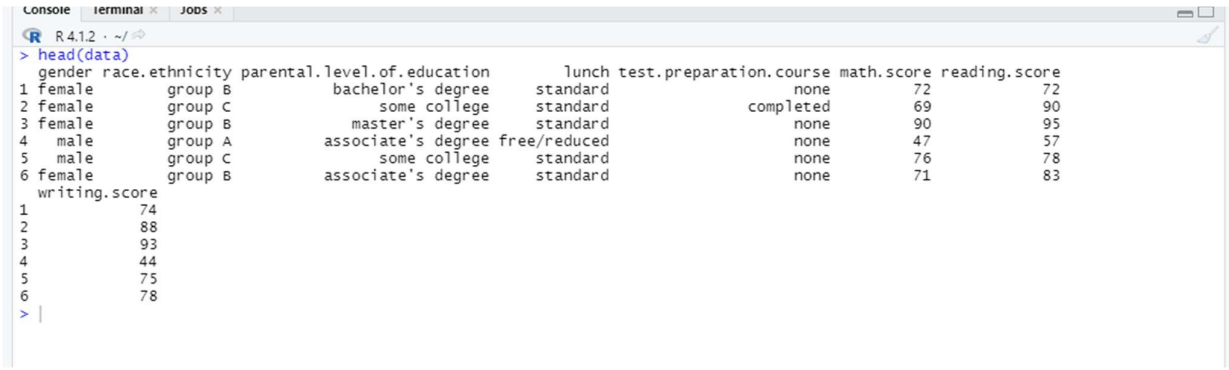


3) dim(data)



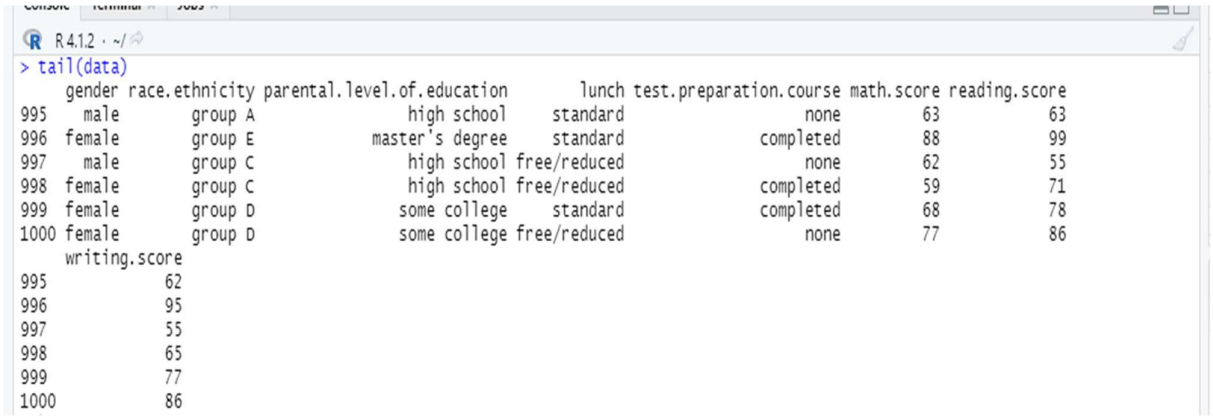
DIMENTION FUNCTION RETURNING THE COUNT OF TOTAL NUMBER OF ROWS AND COLUMNS IN THE DATA SET AND HEAD FUNCTION DISPLAYING THE FIRST FIVE ROWS

4)head(data)



```
> head(data)
  gender race.ethnicity parental.level.of.education       lunch test.preparation.course math.score reading.score
1 female        group B           bachelor's degree    standard                    none         72            72
2 female        group C                some college    standard               completed         69            90
3 female        group B             master's degree    standard                    none         90            95
4   male        group A          associate's degree free/reduced                    none         47            57
5   male        group C                some college    standard                    none         76            78
6 female        group B          associate's degree    standard                    none         71            83
  writing.score
1            74
2            88
3            93
4            44
5            75
6            78
>
```

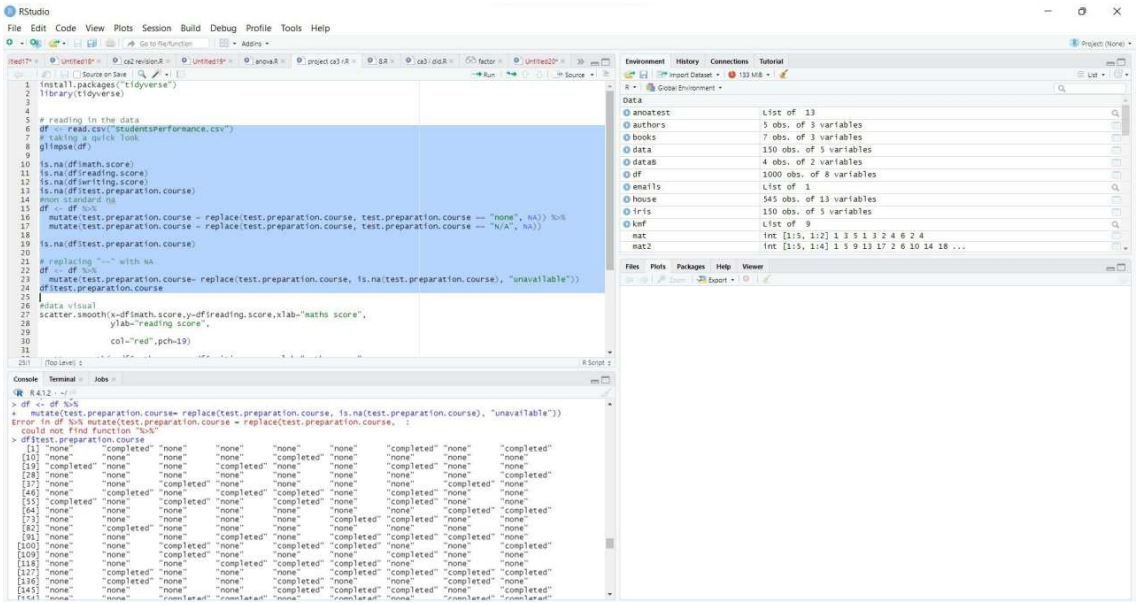
5) tail(data)



```
> tail(data)
     gender race.ethnicity parental.level.of.education       lunch test.preparation.course math.score reading.score
995    male        group A                high school    standard                    none         63            63
996  female        group E            master's degree    standard               completed         88            99
997    male        group C                high school free/reduced                    none         62            55
998  female        group C                high school free/reduced               completed         59            71
999  female        group D               some college    standard               completed         68            78
1000 female        group D               some college free/reduced                    none         77            86
     writing.score
995             62
996             95
997             55
998             65
999             77
1000            86
```

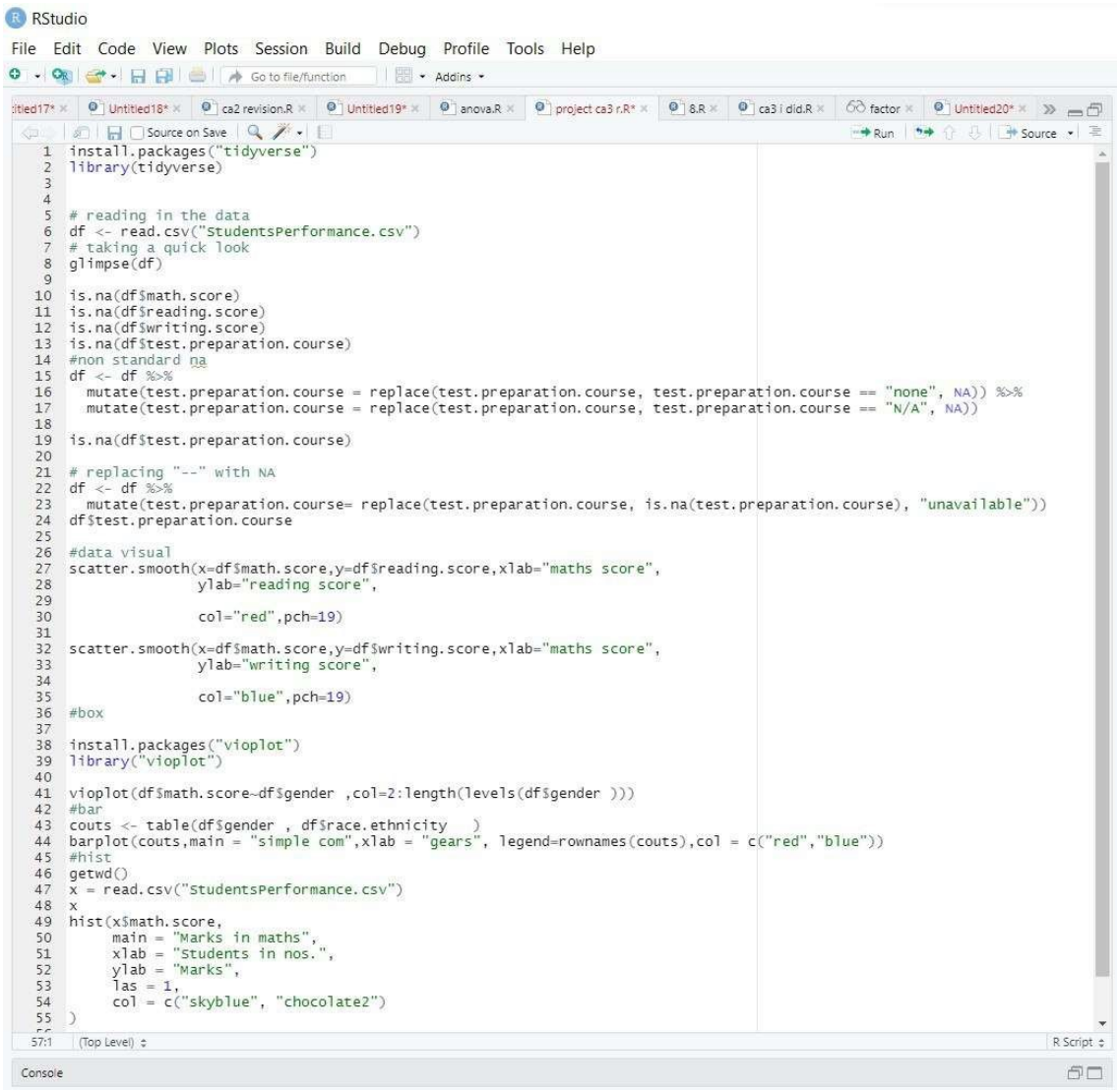
```
Console  Terminal  Jobs
R 4.1.2 · ~/
> colnames(data)<- c("gender","race_ethnicity","parential_level_of_education","lunch","test_preparation_course","math_score","reading_score","writing score")
> head(data)
  gender race_ethnicity parential_level_of_education       lunch test_preparation_course math_score reading_score
1 female        group B           bachelor's degree    standard                    none         72            72
2 female        group C                some college    standard               completed         69            90
3 female        group B            master's degree    standard                    none         90            95
4   male        group A         associate's degree free/reduced                    none         47            57
5   male        group C                some college    standard                    none         76            78
6 female        group B         associate's degree    standard                    none         71            83
  writing score
1            74
2            88
3            93
4            44
5            75
6            78
>
```

Data Cleaning:



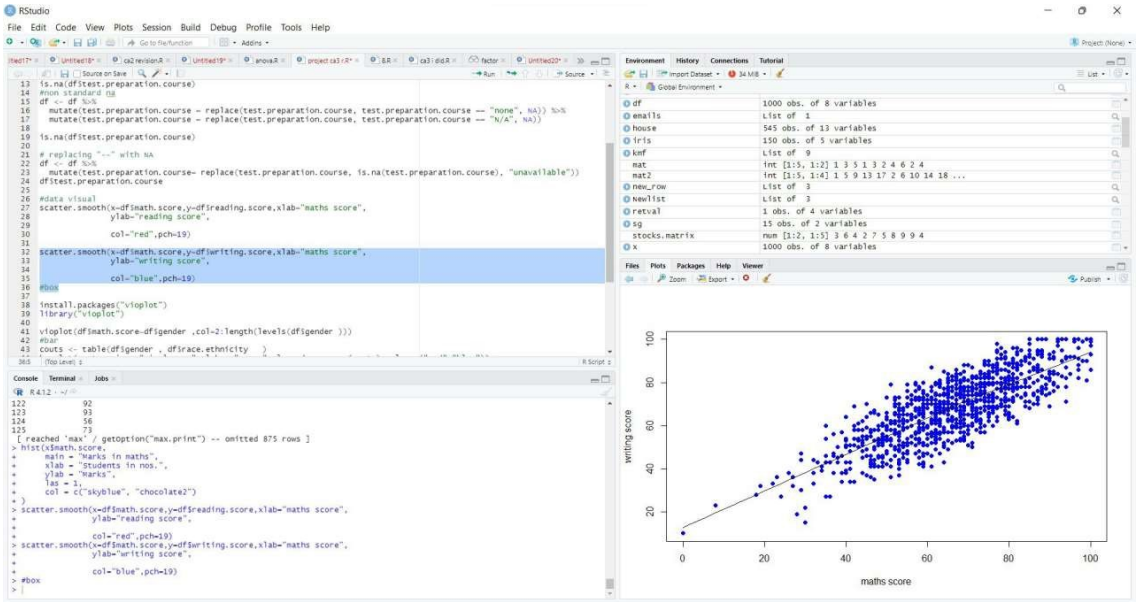
```
  [1] "none"      "completed" "none"      "none"      "none"      "none"      "completed" "none"      "completed"
 [10] "none"      "none"      "none"      "none"      "completed" "none"      "none"      "none"      "none"
 [19] "completed" "none"      "none"      "completed" "none"      "none"      "completed" "none"      "none"
 [28] "none"      "none"      "none"      "none"      "none"      "none"      "none"      "none"      "completed"
 [37] "none"      "none"      "completed" "none"      "none"      "none"      "none"      "completed" "none"
 [46] "none"      "completed" "none"      "completed" "completed" "none"      "completed" "none"      "none"
 [55] "completed" "none"      "completed" "completed" "none"      "none"      "completed" "none"      "none"
 [64] "none"      "none"      "none"      "none"      "none"      "none"      "none"      "completed" "completed"
 [73] "none"      "none"      "none"      "none"      "none"      "completed" "completed" "none"      "none"
 [82] "none"      "none"      "completed" "none"      "none"      "none"      "none"      "none"      "none"
 [91] "none"      "none"      "none"      "completed" "none"      "completed" "completed" "completed" "none"
[100] "none"      "none"      "completed" "none"      "completed" "completed" "none"      "none"      "completed"
[109] "none"      "none"      "completed" "none"      "none"      "none"      "completed" "none"      "none"
[118] "none"      "none"      "none"      "completed" "completed" "none"      "none"      "none"      "none"
[127] "none"      "completed" "none"      "none"      "none"      "completed" "completed" "completed" "completed"
[136] "none"      "completed" "none"      "none"      "completed" "none"      "none"      "completed" "none"
[145] "none"      "none"      "none"      "none"      "completed" "completed" "completed" "none"      "completed"
```

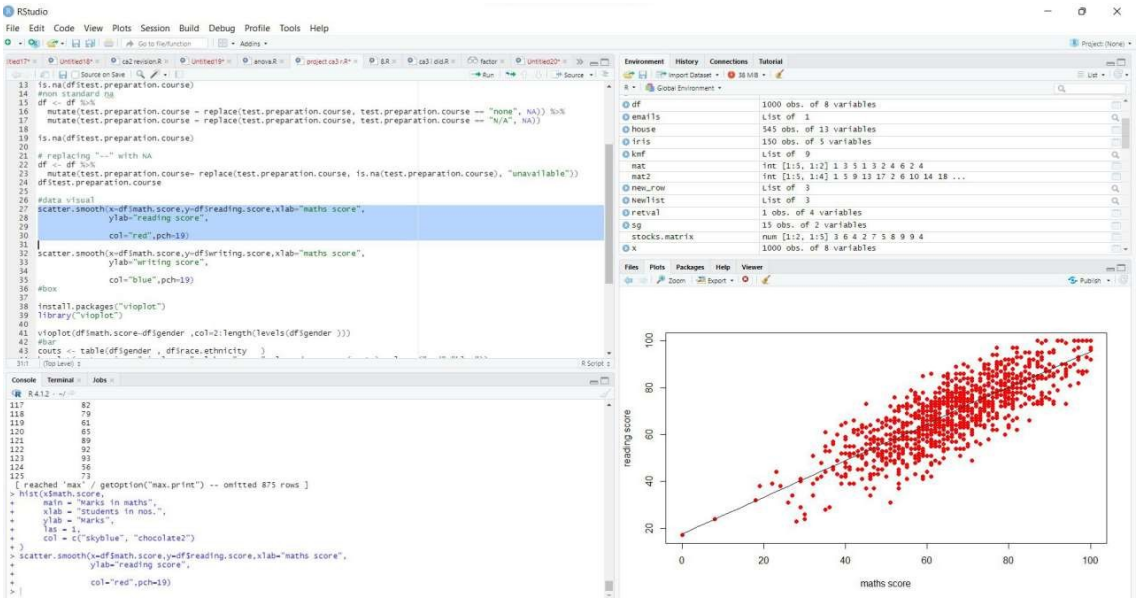
Code:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled18*
ca2 revision.R
Untitled19*
anova.R
project ca3 r.R*
8.R
ca3 i did.R
factor
Untitled20*
Source on Save
Run
Source
install.packages("tidyverse")
library(tidyverse)
# reading in the data
df <- read.csv("StudentsPerformance.csv")
# taking a quick look
glimpse(df)
is.na(df$math.score)
is.na(df$reading.score)
is.na(df$writing.score)
is.na(df$test.preparation.course)
#non standard na
df <- df %>%
mutate(test.preparation.course = replace(test.preparation.course, test.preparation.course == "none", NA)) %>%
mutate(test.preparation.course = replace(test.preparation.course, test.preparation.course == "N/A", NA))
is.na(df$test.preparation.course)
# replacing "--" with NA
df <- df %>%
mutate(test.preparation.course= replace(test.preparation.course, is.na(test.preparation.course), "unavailable"))
df$test.preparation.course
#data visual
scatter.smooth(x=df$math.score,y=df$reading.score,xlab="maths score",
ylab="reading score",
col="red",pch=19)
scatter.smooth(x=df$math.score,y=df$writing.score,xlab="maths score",
ylab="writing score",
col="blue",pch=19)
#box
install.packages("vioplot")
library("vioplot")
vioplot(df$math.score~df$gender ,col=2:length(levels(df$gender)))
#bar
couts <- table(df$gender , df$race.ethnicity)
barplot(couts,main = "simple com",xlab = "gears", legend=rownames(couts),col = c("red","blue"))
#hist
getwd()
x = read.csv("StudentsPerformance.csv")
x
hist(x$math.score,
main = "Marks in maths",
xlab = "Students in nos.",
ylab = "Marks",
las = 1,
col = c("skyblue", "chocolate2")
)
57:1
(Top Level)
R Script
Console

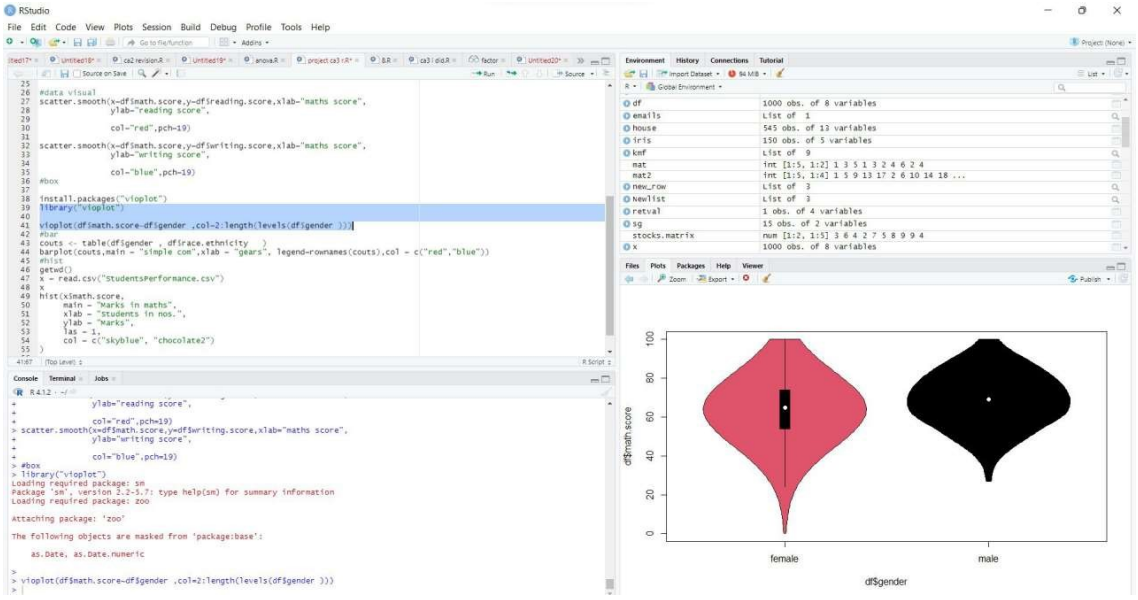
Data Visualization:

Scatter Plot:

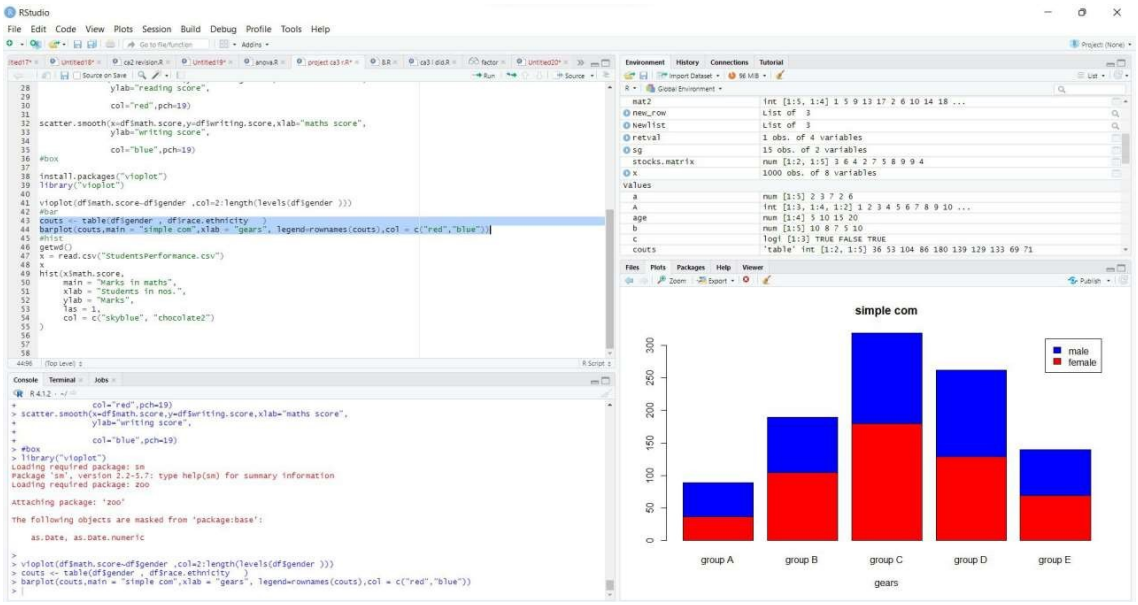



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
reading score
maths score
0
20
40
60
80
100

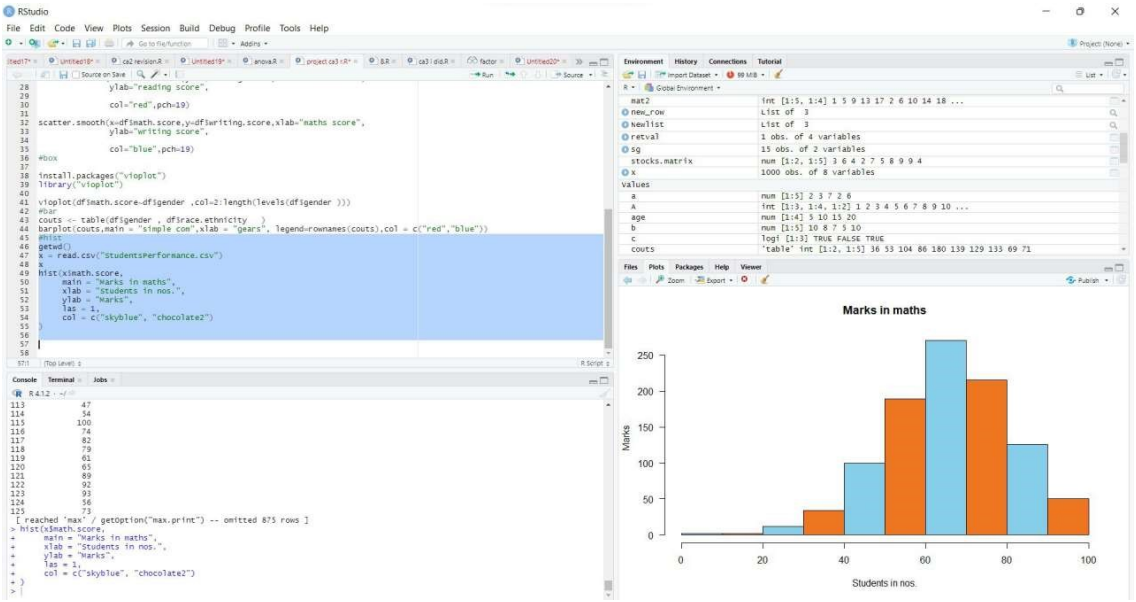
Violin Plot:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
df$math.score
female
male
df$gender
0
20
40
60
80
100

Bar Chart:



Histogram:

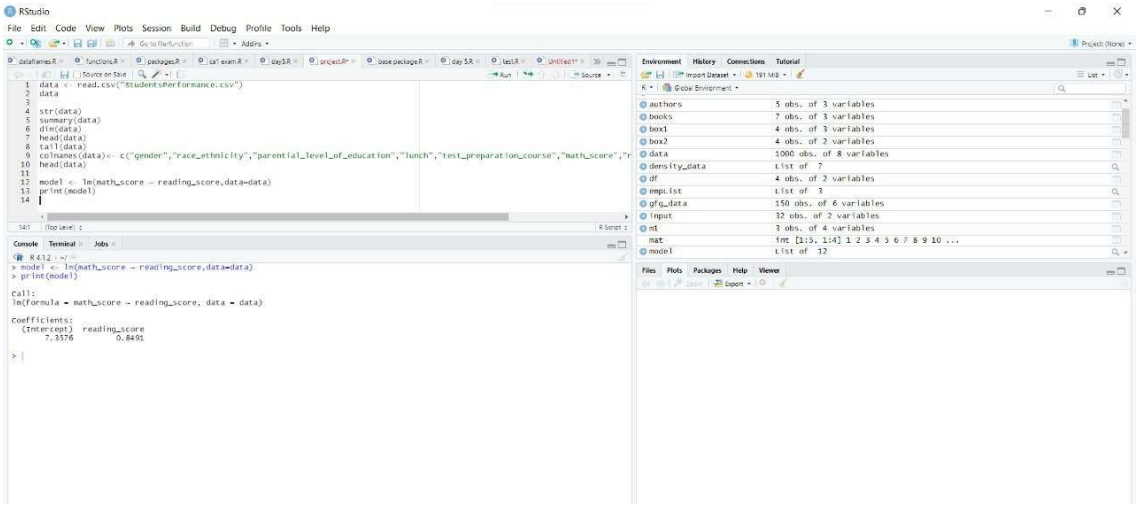


ML Algorithm model:

SINGLE LINEAR REGRESSION

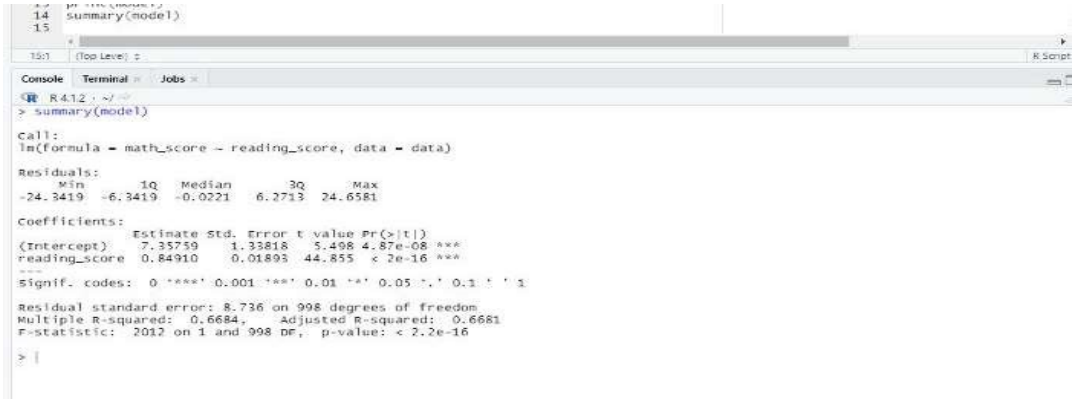
#BASIC LINEAR REGRESSION

#PREDICTING MATH SCORE WITH READING SCORE



#THE MODEL PREDICTS THE FORMULA BELOW:

#MATH_SCORE = 7.3576 + 0.8491 * READING_SCORE




```
# Accuracy
sqrt(mean((data$reading_score-predicted_value)^2))
```

20.6478715883146

```
model2 <- lm(math_score ~ writing_score, data=data)
print(model2)
```

```
Call:
lm(formula = math_score ~ writing_score, data = data)

Coefficients:
  (Intercept)  writing_score
      11.5831         0.8009
```

```
summary(model2)
```

```
Call:
lm(formula = math_score ~ writing_score, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-24.8467  -6.4600   0.1464   6.4356  25.5515

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.58310    1.31369   8.817   <2e-16 ***
writing_score  0.80092    0.01884  42.511   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.049 on 998 degrees of freedom
Multiple R-squared:  0.6442,    Adjusted R-squared:  0.6439
F-statistic:  1807 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
b<-data.frame(writing_score=80)
predicted_value_two=predict(model2,b)
predicted_value_two
```

1: 75.6568060570767

```
# Accuracy
sqrt(mean((data$writing_score-predicted_value_two)^2))
```

16.9846914585318

MULTIPLE LINEAR REGRESSION

```
##MULTIPLE LINEAR REGRESSION

#Predicting math scores for reading and writing score
model3 = lm(math.score~reading.score + writing.score, data =df)
print(model3)
```

```
Call:
lm(formula = math.score ~ reading.score + writing.score, data = df)

Coefficients:
  (Intercept)  reading.score  writing.score
       7.5241         0.6013         0.2494
```

```
#Review the results
summary(model3)
```

```
Call:
lm(formula = math.score ~ reading.score + writing.score, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-23.8779  -6.1750   0.2693   6.0184  24.8727

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.52409    1.32823   5.665 1.93e-08 ***
reading.score  0.60129    0.06304   9.538  < 2e-16 ***
writing.score  0.24942    0.06057   4.118 4.14e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.667 on 997 degrees of freedom
Multiple R-squared:  0.674,     Adjusted R-squared:  0.6733
F-statistic:  1031 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
b <-data.frame(reading.score=90,writing.score=88)
predicted_value <-predict(model3,b)
```

```
predicted_value
```

1: 83.5894767553443

```
#Redidual Standard Error
RSE=sigma(model3)/mean(df$math.score)
RSE
```

0.131133840006564

```
#Accuracy
sqrt(mean((predicted_value)^2))
```

83.5894767553443

Conclusion:

- From the given dataset we can able to predict the mark obtained by each student using the linear regression model.

- Here we use both Single and Multiple Linear Regression, and we use Multiple Linear Regression since the given dataset contain more than 2 columns and it has more accuracy compared to Single Linear Regression.

- So we are able to predict the mark of the student using the Multiple Linear regression model.