Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A

EXPERIMENT-11

EXPT11 : Implement Linear regression using R tool

Theory: Linear regression is a regression model that uses a straight line to describe the relationship between variables. It finds the line of best fit through given data by searching for the value of the regression coefficient(s) that minimizes the total error of the model. There are two main types of linear regression:

- Simple linear regression uses only one independent variable
- Multiple linear regression uses two or more independent variables

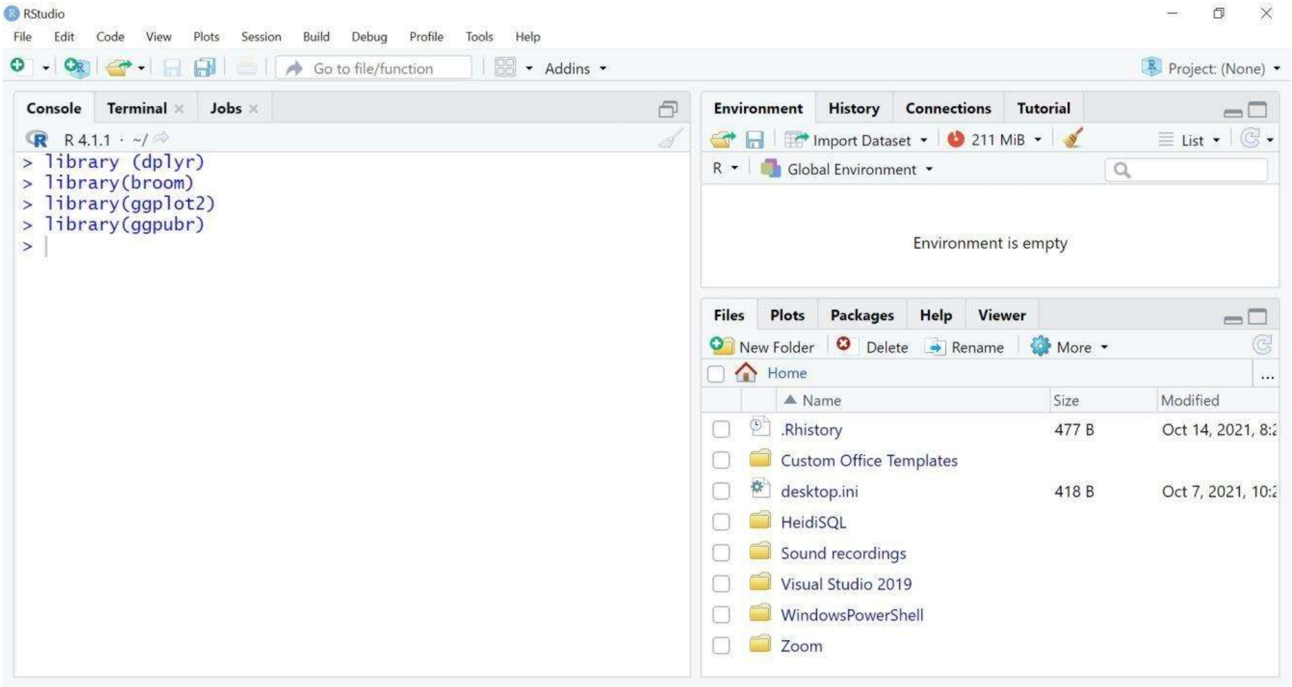
Simple Linear Regression: The first dataset contains observations about income (in a range of $15k to $75k) and happiness (rated on a scale of 1 to 10) in an imaginary sample of 500 people. The income values are divided by 10,000 to make the income data match the scale of the happiness scores (so a value of $2 represents $20,000, $3 is $30,000, etc.)

Multiple Linear Regression: The second dataset contains observations on the percentage of people biking to work each day, the percentage of people smoking, and the percentage of people with heart disease in an imaginary sample of 500 towns. Steps to implement Linear Regression in R:

Step 1: Open RStudio and click on File > New File > R Script.
Then install the packages needed for the analysis, using following code (only need to do this once):

```
install.packages("ggplot
2")
install.packages("dplyr"
)
install.packages("broom
")
install.packages("ggpub
r") library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)
```

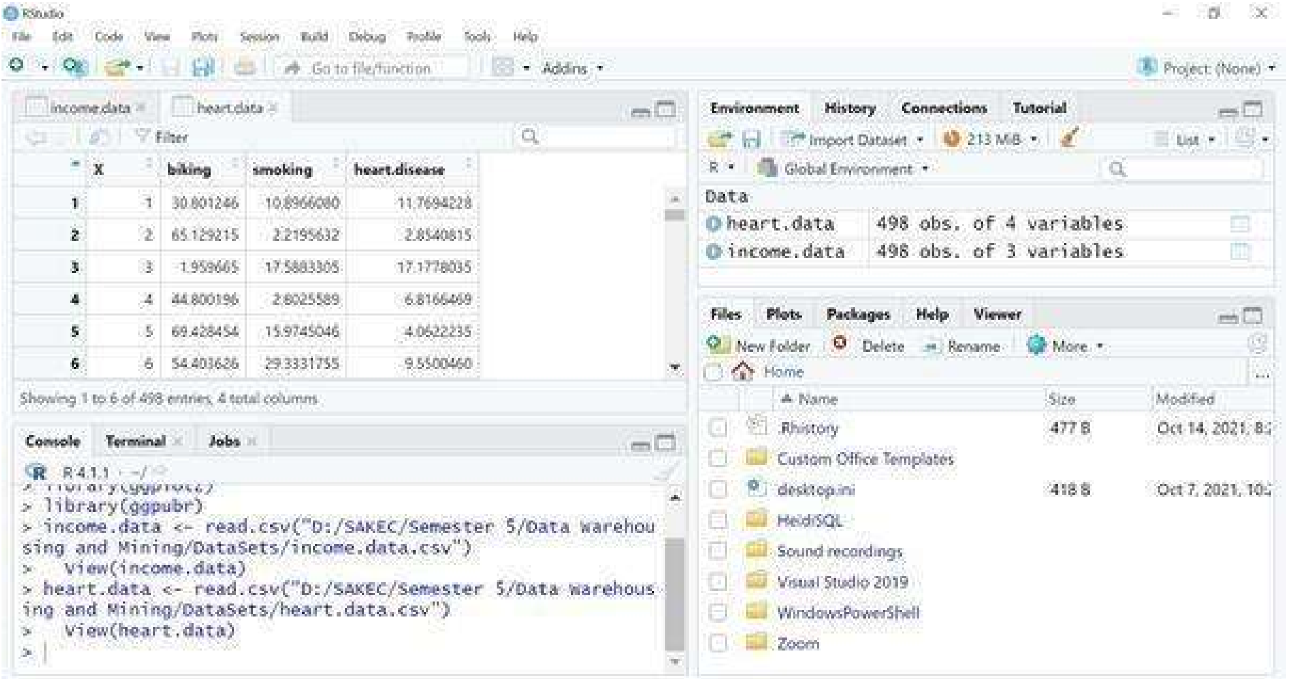
Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A



```
R 4.1.1 · ~/
> library (dplyr)
> library(broom)
> library(ggplot2)
> library(ggpubr)
> |
```

Environment is empty

Name	Size	Modified
.Rhistory	477 B	Oct 14, 2021, 8::
Custom Office Templates		
desktop.ini	418 B	Oct 7, 2021, 10::
HeidiSQL		
Sound recordings		
Visual Studio 2019		
WindowsPowerShell		
Zoom		



	X	biking	smoking	heart.disease
1	1	30.801246	10.8966080	11.7694228
2	2	65.129215	2.2195632	2.8540815
3	3	1.959665	17.5883305	17.1778035
4	4	44.800196	2.8025589	6.8166469
5	5	69.428454	15.9745046	4.0622235
6	6	54.403626	29.3331755	9.5500460

Showing 1 to 6 of 498 entries, 4 total columns

```
R 4.1.1 · ~/
> library(ggplot2)
> library(ggpubr)
> income.data <- read.csv("D:/SAKEC/Semester 5/Data warehou
sing and Mining/DataSets/income.data.csv")
>   View(income.data)
> heart.data <- read.csv("D:/SAKEC/Semester 5/Data warehous
ing and Mining/DataSets/heart.data.csv")
>   View(heart.data)
> |
```

Data

```
heart.data    498 obs. of 4 variables
income.data   498 obs. of 3 variables
```

Name	Size	Modified
.Rhistory	477 B	Oct 14, 2021, 8::
Custom Office Templates		
desktop.ini	418 B	Oct 7, 2021, 10::
HeidiSQL		
Sound recordings		
Visual Studio 2019		
WindowsPowerShell		
Zoom		

Step 2: Load the data into R

Follow these four steps for each dataset:

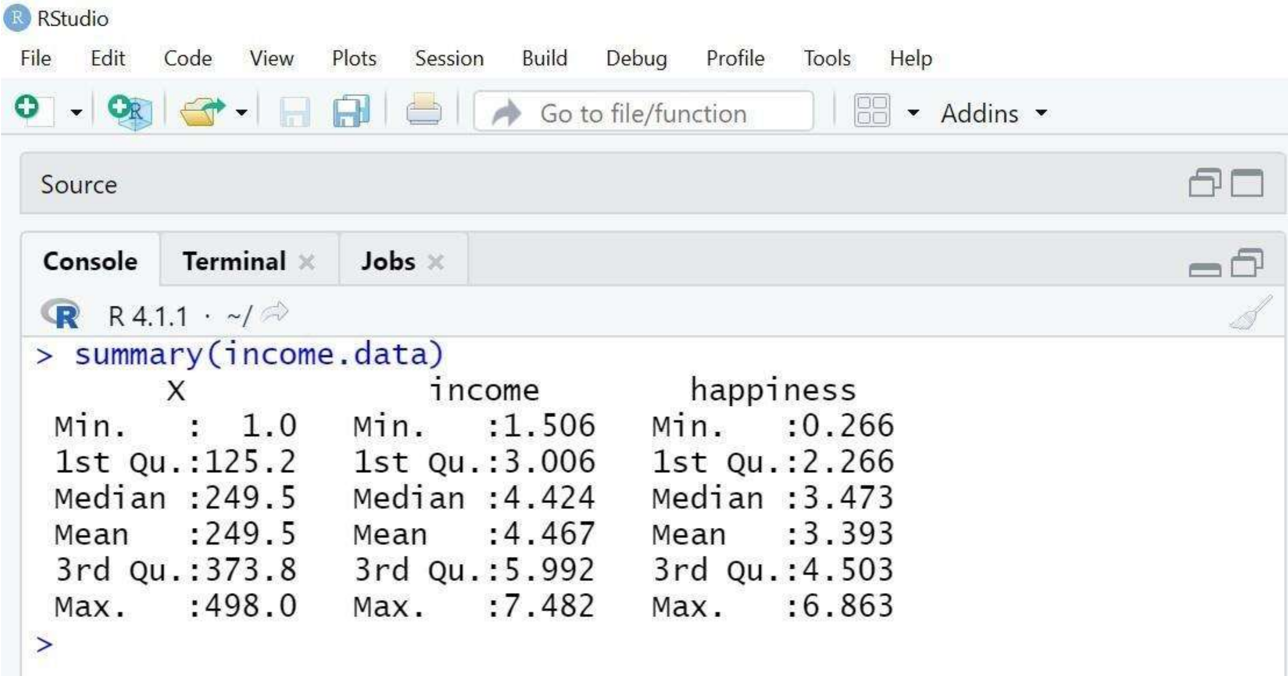
- In RStudio, go to File > Import dataset > From Text (base).

Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A

- Choose the data file you have downloaded (income.data or heart. Data), and an Import Dataset window pops up.
- In the Data Frame window, X (index) column and columns listing the data for each of the variables (income and happiness or biking, smoking, and heart.disease) is there.
- Click on the Import button and the file should appear in R Environment tab on the upper right side of the RStudio screen.

Once the data is loaded, check that it has been read in correctly using summary().

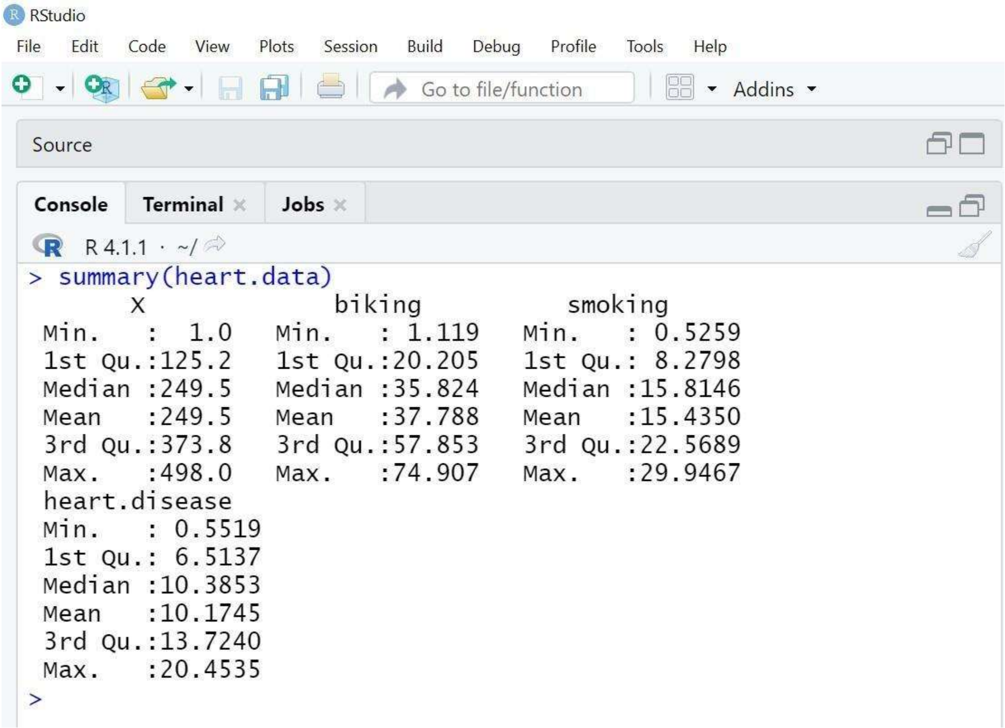
1. Simple regression: Because both the variables are quantitative, when this function executed it shows output as a table with a numeric summary of the data. This tells that the minimum, median, mean, and maximum values of the independent variable (income) and dependent variable (happiness):



```
> summary(income.data)
       X             income        happiness    
 Min.   :  1.0   Min.   :1.506   Min.   :0.266  
 1st Qu.:125.2   1st Qu.:3.006   1st Qu.:2.266  
 Median :249.5   Median :4.424   Median :3.473  
 Mean   :249.5   Mean   :4.467   Mean   :3.393  
 3rd Qu.:373.8   3rd Qu.:5.992   3rd Qu.:4.503  
 Max.   :498.0   Max.   :7.482   Max.   :6.863  
>
```

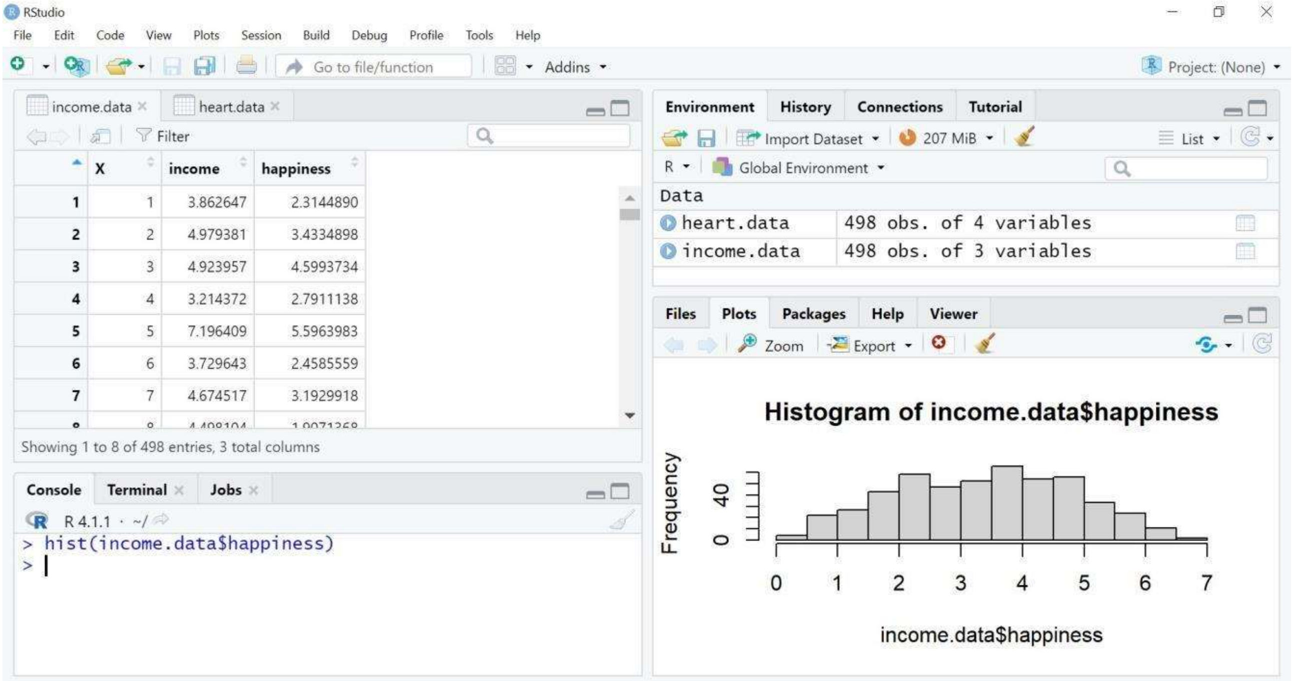
2. Multiple regression: Again, because the variables are quantitative, running the code produces a numeric summary of the data for the independent variables (smoking and biking) and the dependent variable (heart disease):

Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A



```
> summary(heart.data)
       X             biking          smoking       
 Min.   :  1.0   Min.   : 1.119   Min.   : 0.5259  
 1st Qu.:125.2   1st Qu.:20.205   1st Qu.: 8.2798  
 Median :249.5   Median :35.824   Median :15.8146  
 Mean   :249.5   Mean   :37.788   Mean   :15.4350  
 3rd Qu.:373.8   3rd Qu.:57.853   3rd Qu.:22.5689  
 Max.   :498.0   Max.   :74.907   Max.   :29.9467  
 heart.disease    
 Min.   : 0.5519  
 1st Qu.: 6.5137  
 Median :10.3853  
 Mean   :10.1745  
 3rd Qu.:13.7240  
 Max.   :20.4535  
>
```

Step 3: To check whether the dependent variable follows a normal distribution, use the hist() function



	X	income	happiness
1	1	3.862647	2.3144890
2	2	4.979381	3.4334898
3	3	4.923957	4.5993734
4	4	3.214372	2.7911138
5	5	7.196409	5.5963983
6	6	3.729643	2.4585559
7	7	4.674517	3.1929918
8	8	4.498104	1.9071368

Showing 1 to 8 of 498 entries, 3 total columns

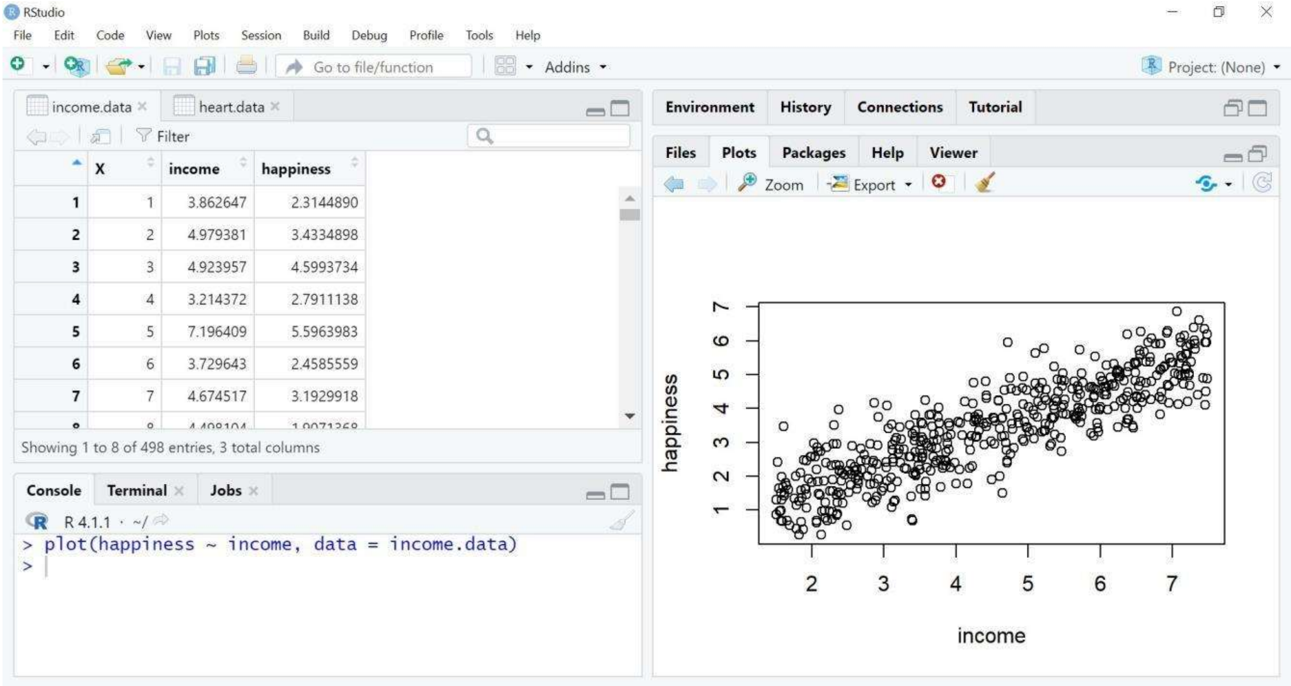
Data

heart.data	498 obs. of 4 variables
income.data	498 obs. of 3 variables

```
> hist(income.data$happiness)
> 
```

Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A

Step 4: The relationship between the independent and dependent variable must be linear. To test this visually with a scatter plot to see if the distribution of data points could be described with a straight line or not.



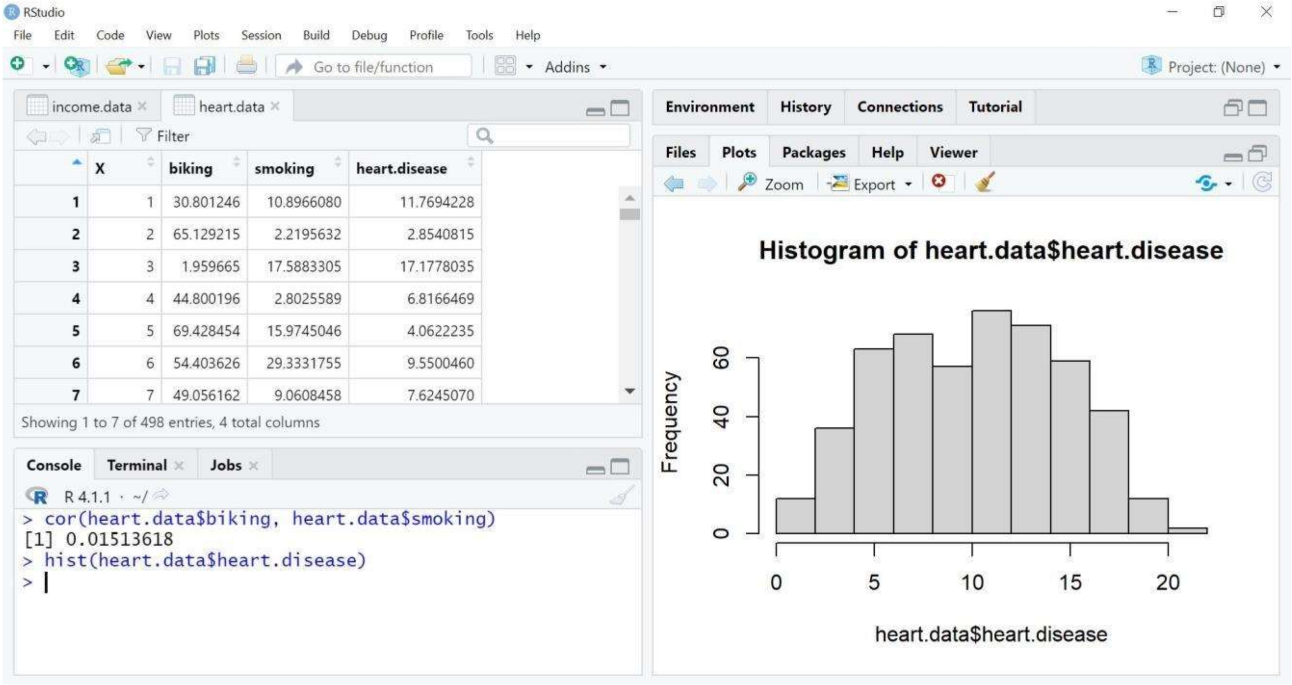
	X	income	happiness
1	1	3.862647	2.3144890
2	2	4.979381	3.4334898
3	3	4.923957	4.5993734
4	4	3.214372	2.7911138
5	5	7.196409	5.5963983
6	6	3.729643	2.4585559
7	7	4.674517	3.1929918
8	8	4.498104	1.9071368

Showing 1 to 8 of 498 entries, 3 total columns

```
> plot(happiness ~ income, data = income.data)
>
```

Step 5: Use the cor() function to test the relationship between independent variables and make sure they aren't too highly correlated.
When this code is executed, the output is 0.015. The correlation between biking and smoking is small (0.015 is only a 1.5% correlation), so that include both parameters in our model. Step 6:
Use the hist() function to test whether your dependent variable follows a normal distribution.

Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A

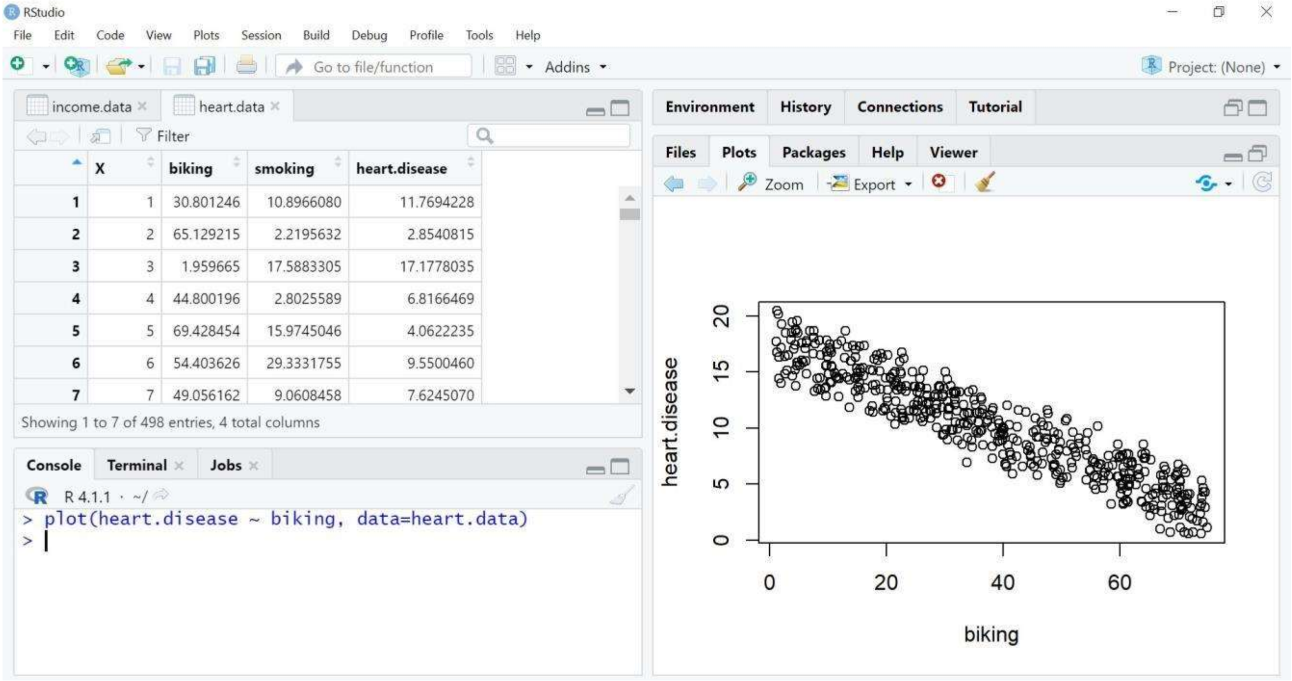

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
income.data
heart.data
Filter
Console Terminal Jobs
R 4.1.1 · ~/
Environment History Connections Tutorial
Files Plots Packages Help Viewer
Zoom Export
Histogram of heart.data$heart.disease
Frequency
0
20
40
60
0
5
10
15
20
heart.data$heart.disease

	X	biking	smoking	heart.disease
1	1	30.801246	10.8966080	11.7694228
2	2	65.129215	2.2195632	2.8540815
3	3	1.959665	17.5883305	17.1778035
4	4	44.800196	2.8025589	6.8166469
5	5	69.428454	15.9745046	4.0622235
6	6	54.403626	29.3331755	9.5500460
7	7	49.056162	9.0608458	7.6245070

Showing 1 to 7 of 498 entries, 4 total columns

```
> cor(heart.data$biking, heart.data$smoking)
[1] 0.01513618
> hist(heart.data$heart.disease)
> |
```

Step 7: Linearity property is checked using two scatterplots: one for biking and heart disease, and one for smoking and heart disease.

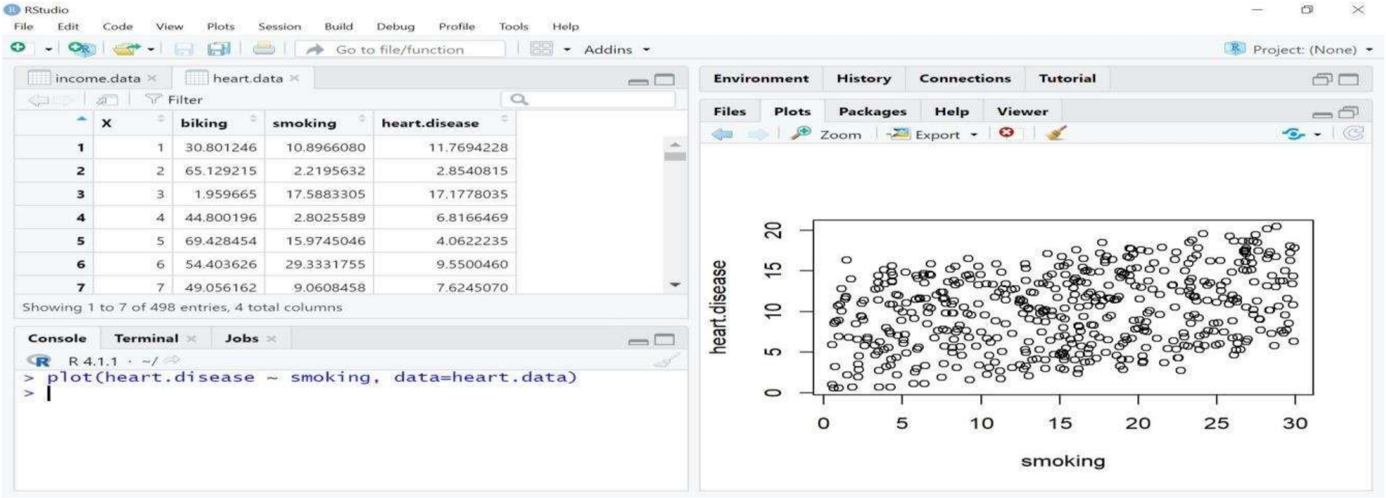


	X	biking	smoking	heart.disease
1	1	30.801246	10.8966080	11.7694228
2	2	65.129215	2.2195632	2.8540815
3	3	1.959665	17.5883305	17.1778035
4	4	44.800196	2.8025589	6.8166469
5	5	69.428454	15.9745046	4.0622235
6	6	54.403626	29.3331755	9.5500460
7	7	49.056162	9.0608458	7.6245070

Showing 1 to 7 of 498 entries, 4 total columns

```
> plot(heart.disease ~ biking, data=heart.data)
> |
```

Name:Chavan Dhananjay Rajendra
Div:TE3 Roll No:6
Batch:A



	X	biking	smoking	heart.disease
1	1	30.801246	10.8966080	11.7694228
2	2	65.129215	2.2195632	2.8540815
3	3	1.959665	17.5883305	17.1778035
4	4	44.800196	2.8025589	6.8166469
5	5	69.428454	15.9745046	4.0622235
6	6	54.403626	29.3331755	9.5500460
7	7	49.056162	9.0608458	7.6245070

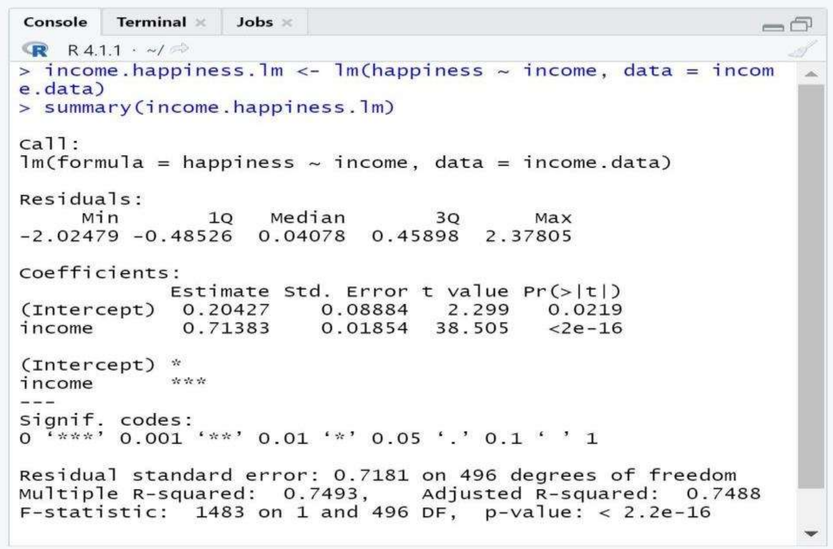
Showing 1 to 7 of 498 entries, 4 total columns

```
> plot(heart.disease ~ smoking, data=heart.data)
> |
```

Step 8: Perform the linear regression analysis

When the data meet the assumptions, perform a linear regression analysis to evaluate the relationship between the independent and dependent variables.

A. Simple regression: income and happiness



```
> income.happiness.lm <- lm(happiness ~ income, data = incom
e.data)
> summary(income.happiness.lm)

Call:
lm(formula = happiness ~ income, data = income.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.02479 -0.48526  0.04078  0.45898  2.37805 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20427    0.08884   2.299   0.0219
income       0.71383    0.01854  38.505   <2e-16
               
(Intercept) *  
income      ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7181 on 496 degrees of freedom
Multiple R-squared:  0.7493,	Adjusted R-squared:  0.7488 
F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

Conclusion:

The above result shows that there is a significant positive relationship between income and happiness (*p*-value < 0.001), with a 0.713-unit (+/- 0.01) increase in happiness for every unit increase in income.