**KING KHALID UNIVERSITY**

**College: Computer Science**

**Department: Information Systems**

**Course Instructor: Ms. Wejdan Mansoor**

**Course code: 373 CIS-4**

**Course Name: Data Mining**

**Total Marks: 10 Marks**

# LAB PROJECT

| Students Names : | Student ID.No : |
|---|---|
| | |
| | |
| | |
| | |

***INSTRUCTIONS:***

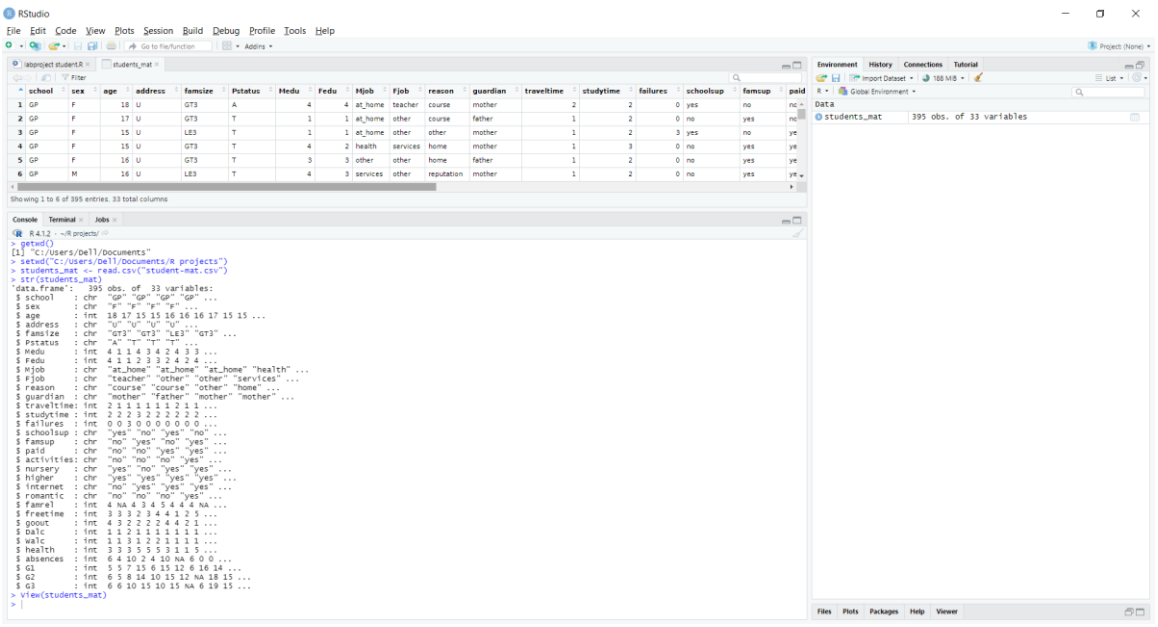
1) The deadline is on **10\2\2023, 19\7\1444.**
2) Groups members are from 3 to 4 members.
3) **Copied assignments will lead to marks deduction**.
4) **Make screens shot for all steps and write down all the related codes also arrange document properly.**

---

• **Open RStudio software and apply the following questions:**

*1.* Import the Students-mat to R studio, *check the attribute information click on the following link:* Math Students | Kaggle

```
students_mat <- read.csv("student-mat.csv")
str(students_mat)
View(students_mat)
```

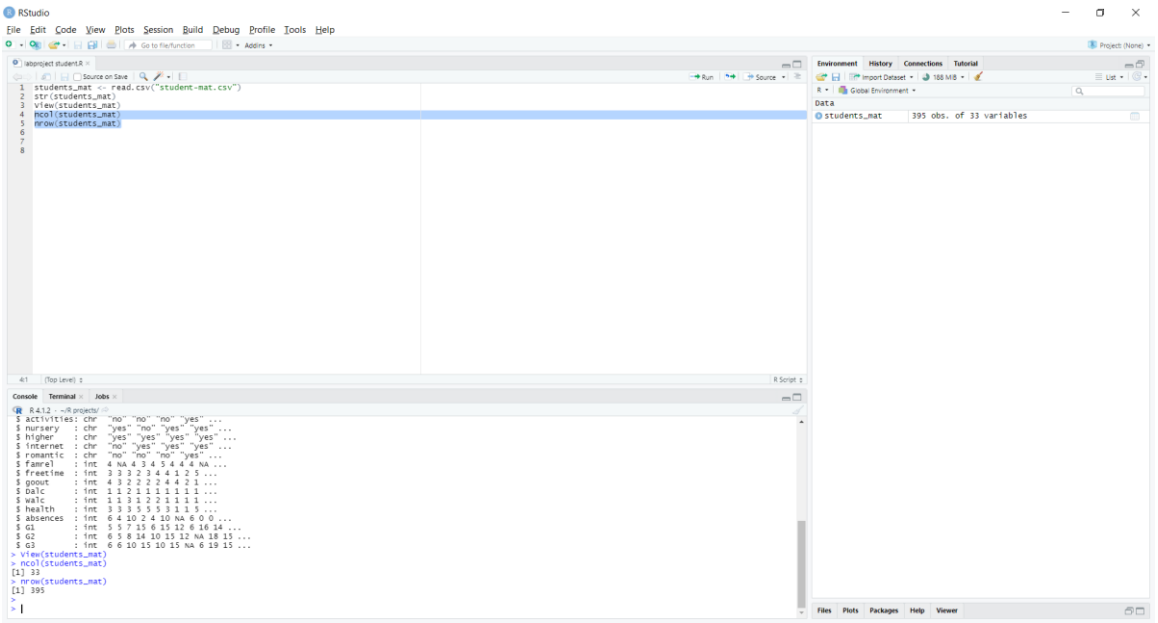


2. How many columns and observations in the dataset.

```
ncol(students_mat)
```

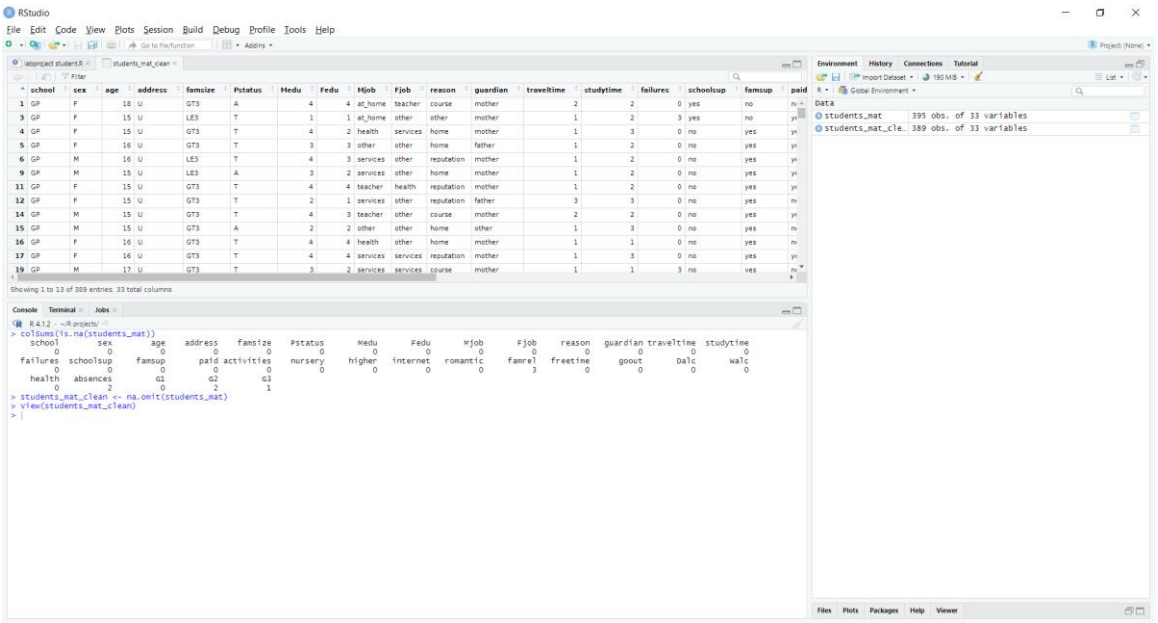
nrow(students_mat)

columns = 33
rows = 395




3. Specify the missing values in which columns then clean your dataset by drop the tuples the include the missing data.

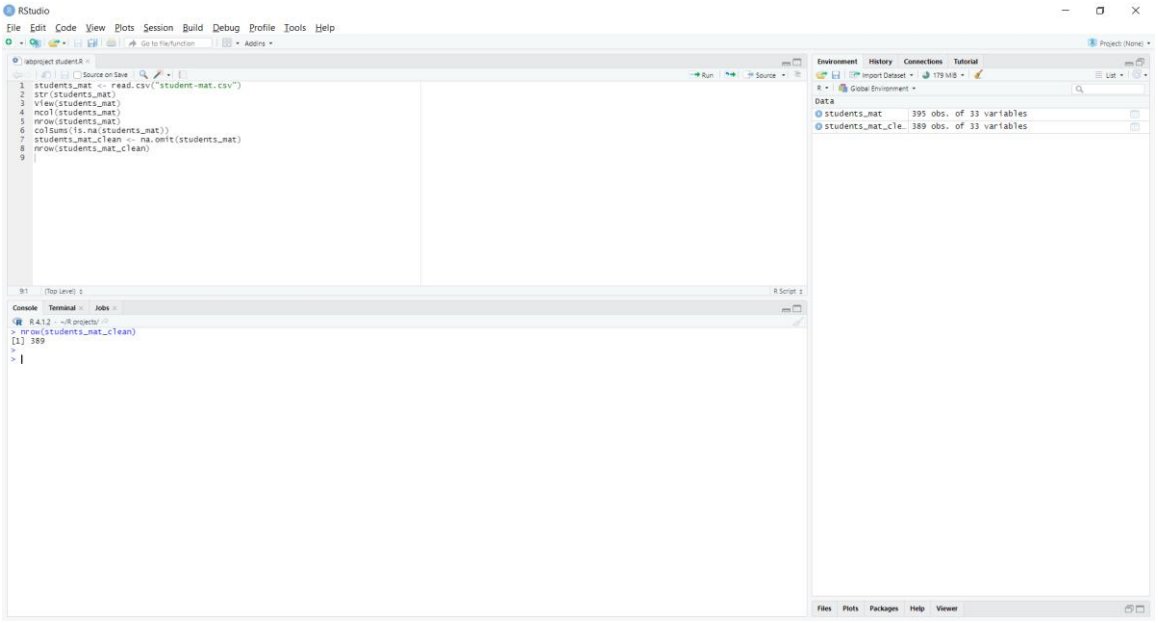
colSums(is.na(students_mat))
students_mat_clean <- na.omit(students_mat)



4. How many tuples do you have now after cleaning the data set?

nrow(students_mat_clean)

rows = 389

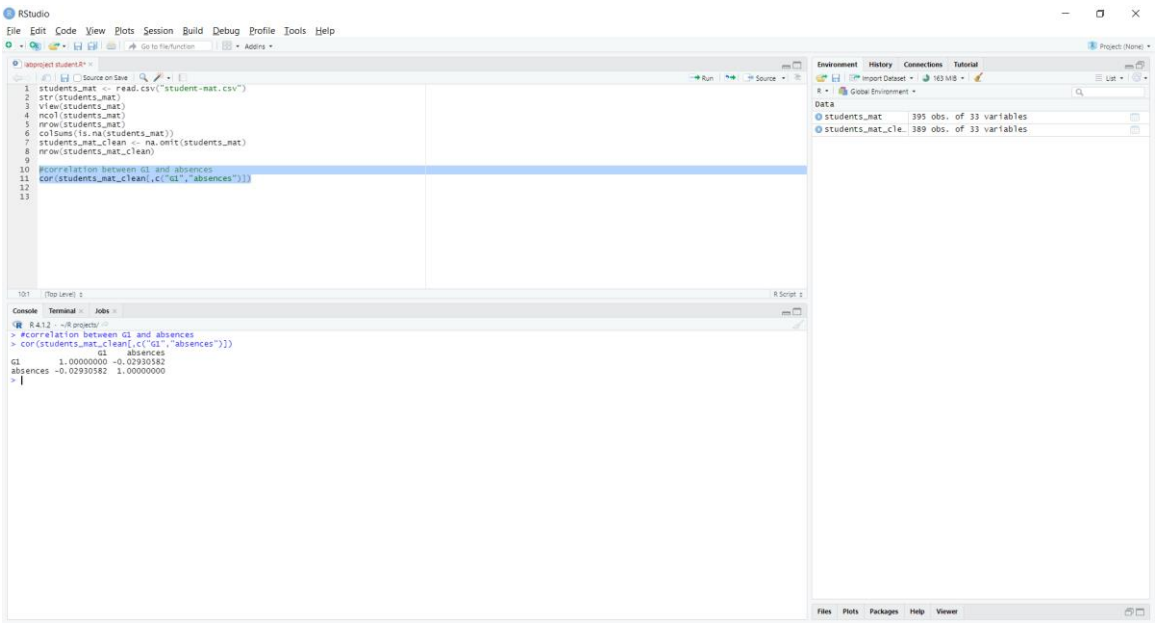



5. Find the correlation between **"G1"** attribute (Grade1) and the following attributes:

- " absences "
- " studytime "
- " G2 "
- " G3"

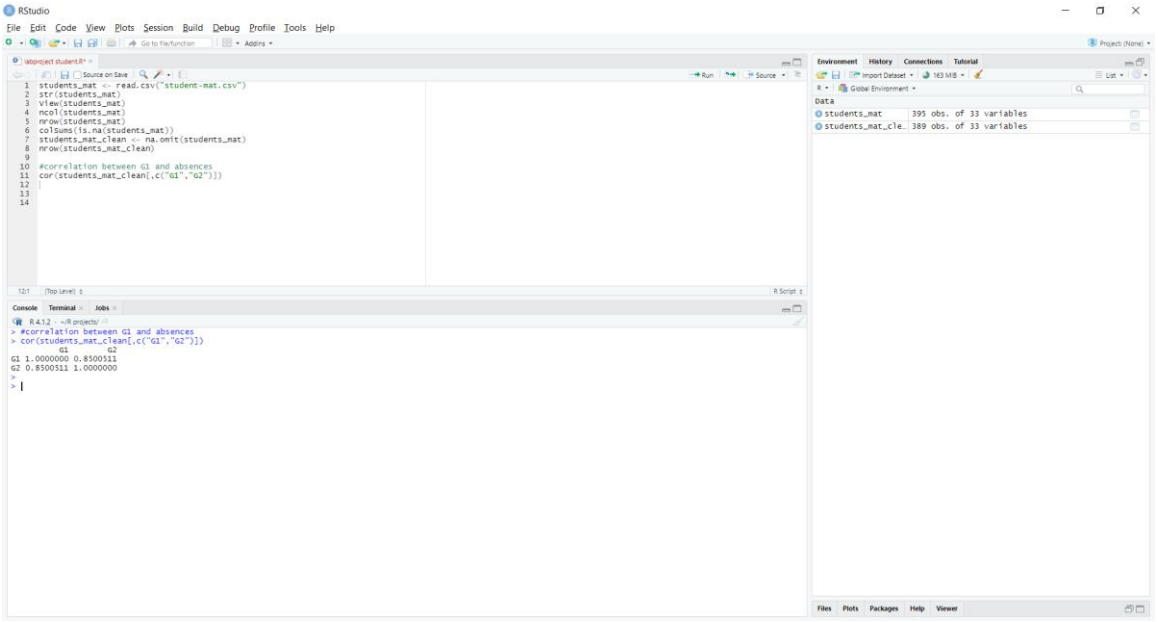
#correlation between G1 and absences

cor(students_mat_clean[,c("G1","absences")])



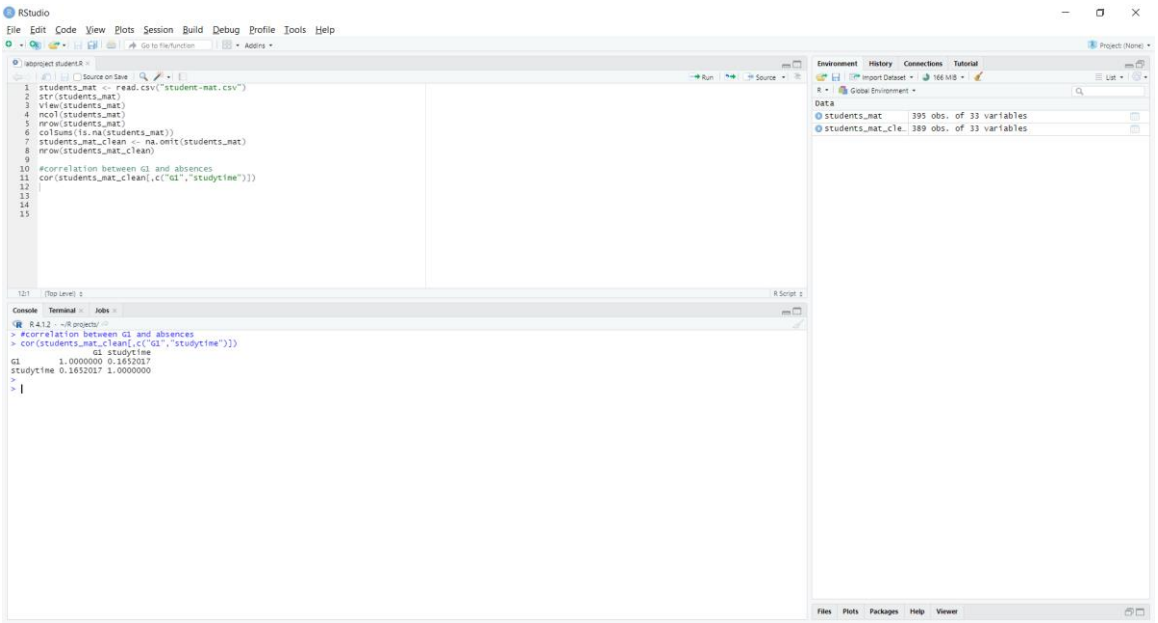

#correlation between G1 and absences

cor(students_mat_clean[,c("G1","G2")])



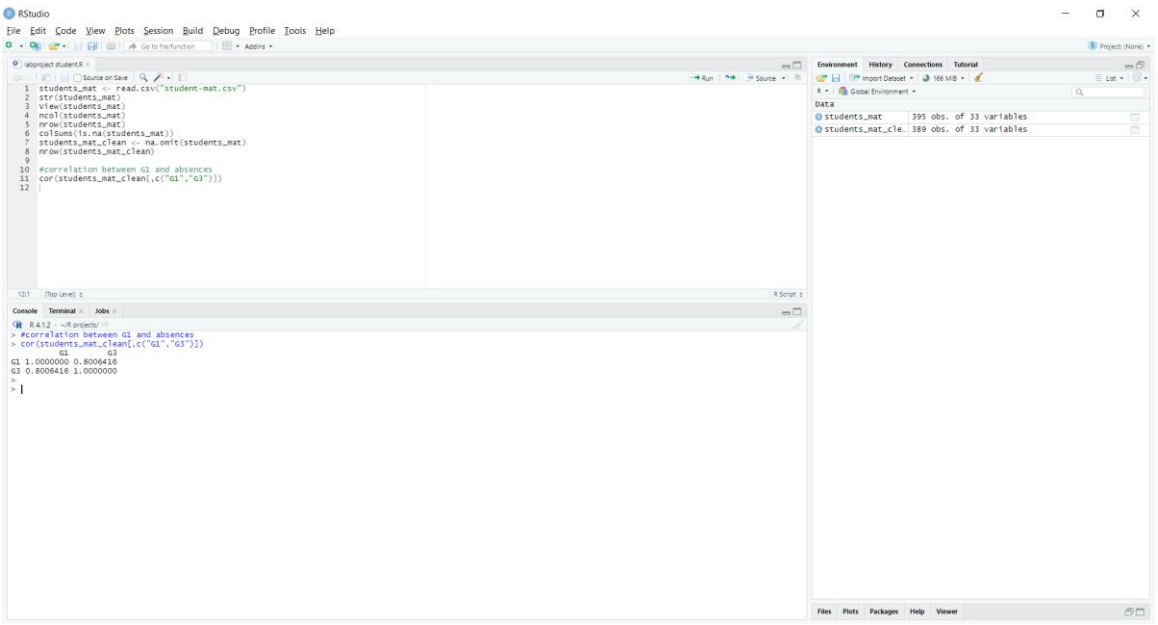

#correlation between G1 and absences

cor(students_mat_clean[,c("G1","studytime")])



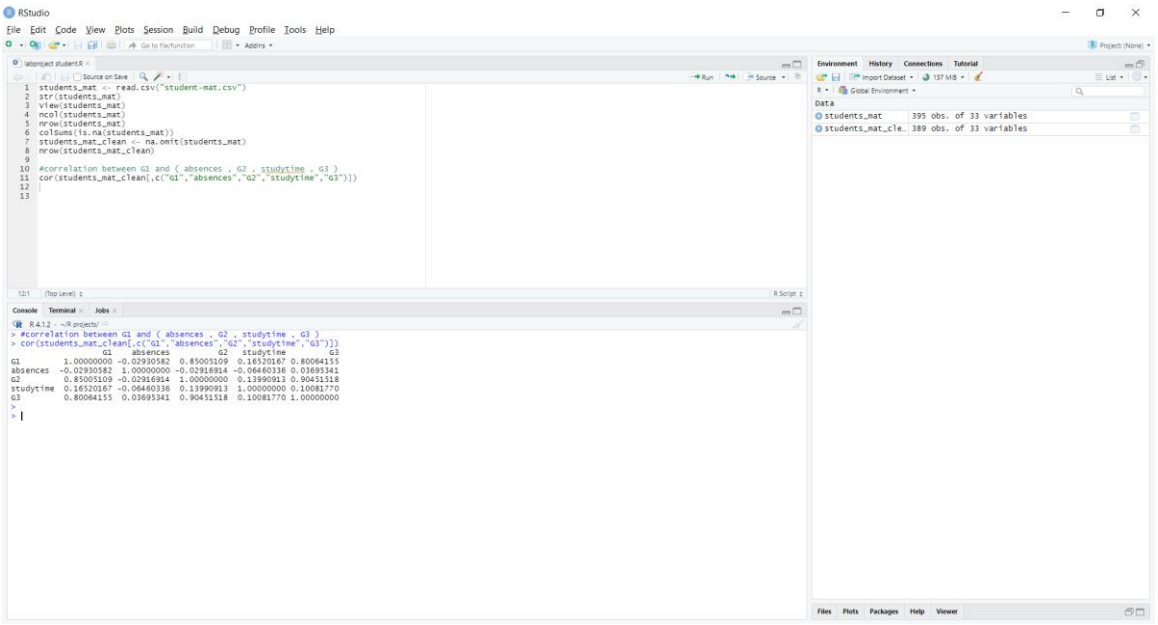

#correlation between G1 and absences

cor(students_mat_clean[,c("G1","G3")])




#correlation between G1 and ( absences , G2 , studytime , G3 )

cor(students_mat_clean[,c("G1","absences","G2","studytime","G3")])




**6.** Specify the correlation types in *point 2* and plot the relations using **scatter plot diagram** (Note: plot each relation and explain it).

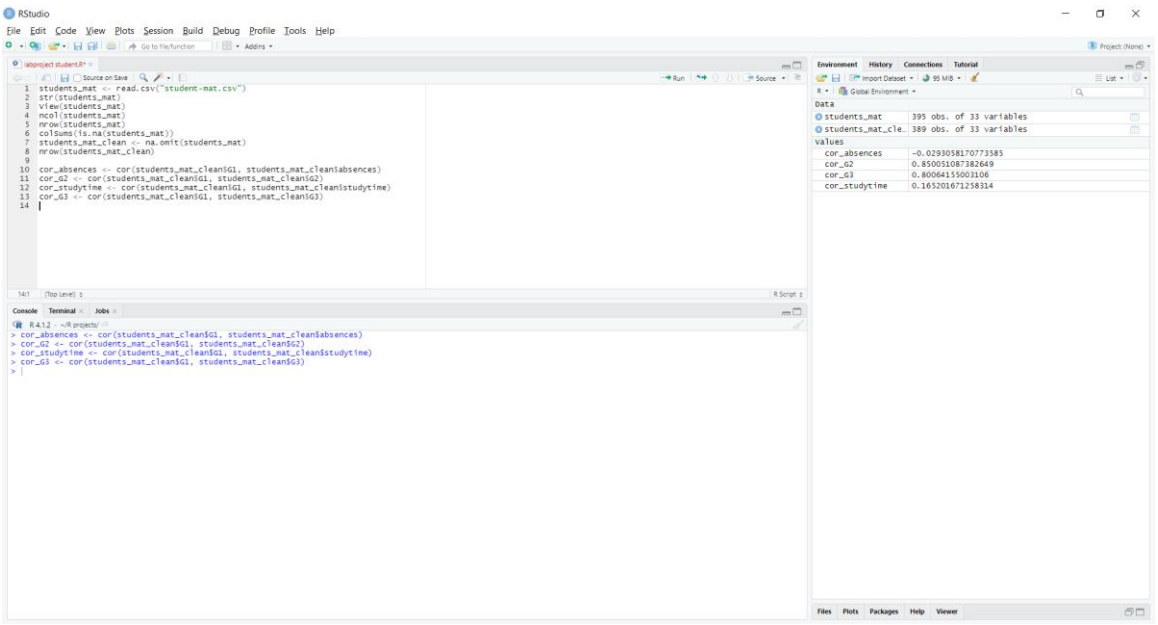
To find the correlation types between "G1" and "absences", "G2", "studytime", "G3", you can use the cor function in R:

cor_absences <- cor(students_mat_clean$G1, students_mat_clean$absences)

cor_G2 <- cor(students_mat_clean$G1, students_mat_clean$G2)

cor_studytime <- cor(students_mat_clean$G1, students_mat_clean$studytime)

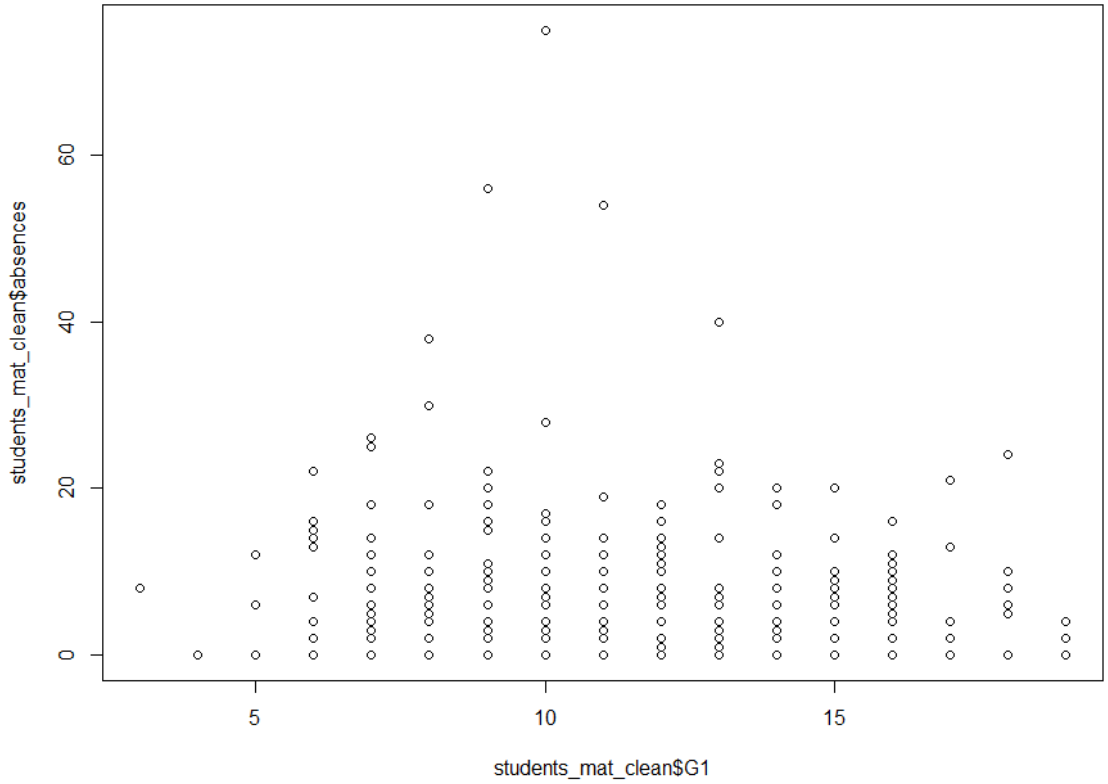
cor_G3 <- cor(students_mat_clean$G1, students_mat_clean$G3)

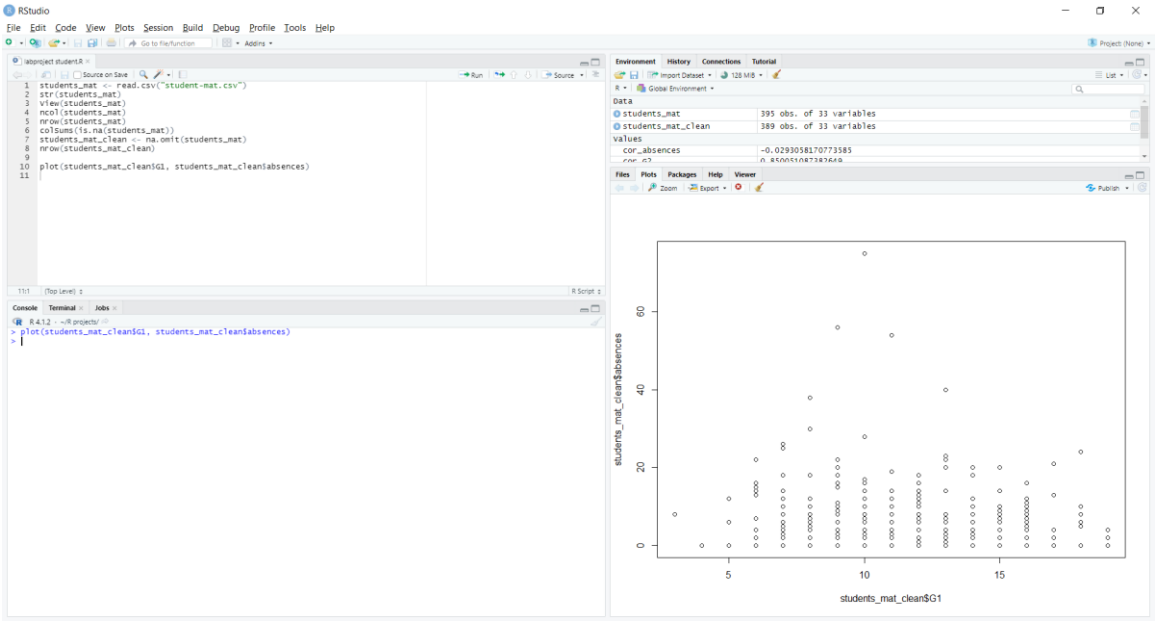



The cor function returns the Pearson's correlation coefficient, which measures the linear relationship between two variables. A positive correlation means that as one variable increases, the other variable increases as well. A negative correlation means that as one variable increases, the other variable decreases. A value of 1 indicates a perfect positive correlation, and a value of -1 indicates a perfect negative correlation. A value of 0 indicates no correlation.

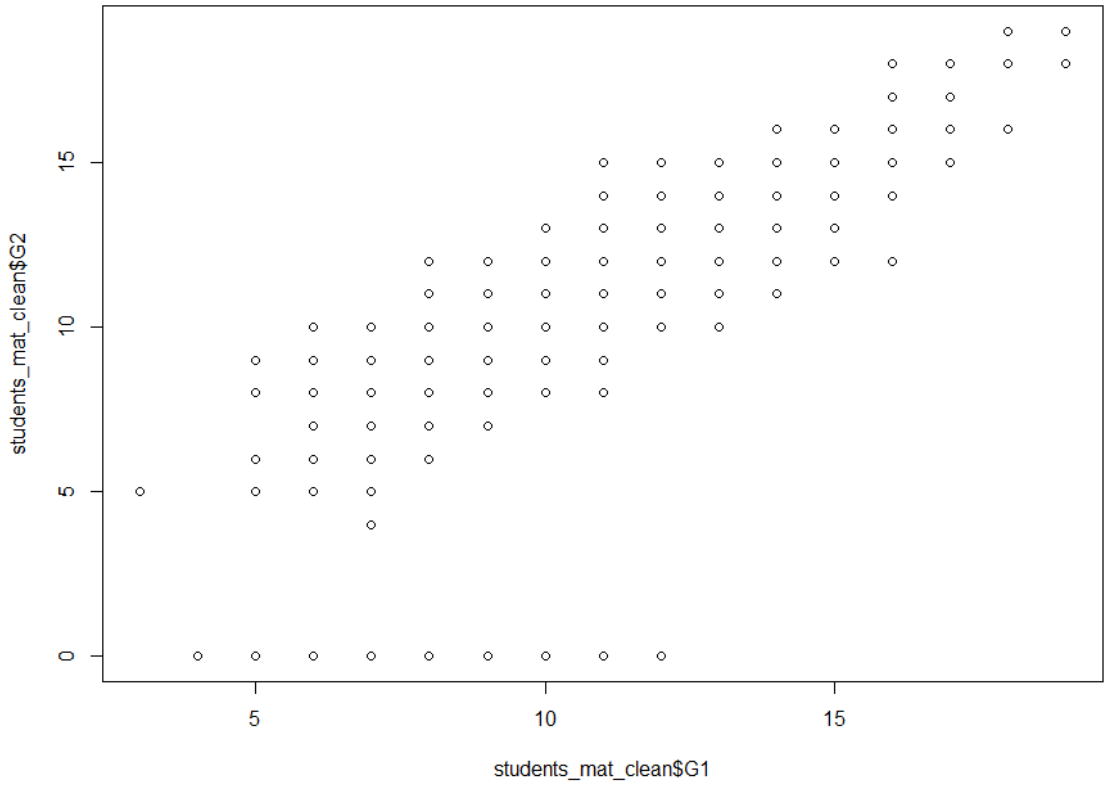
To plot the scatter plots, you can use the plot function in R:

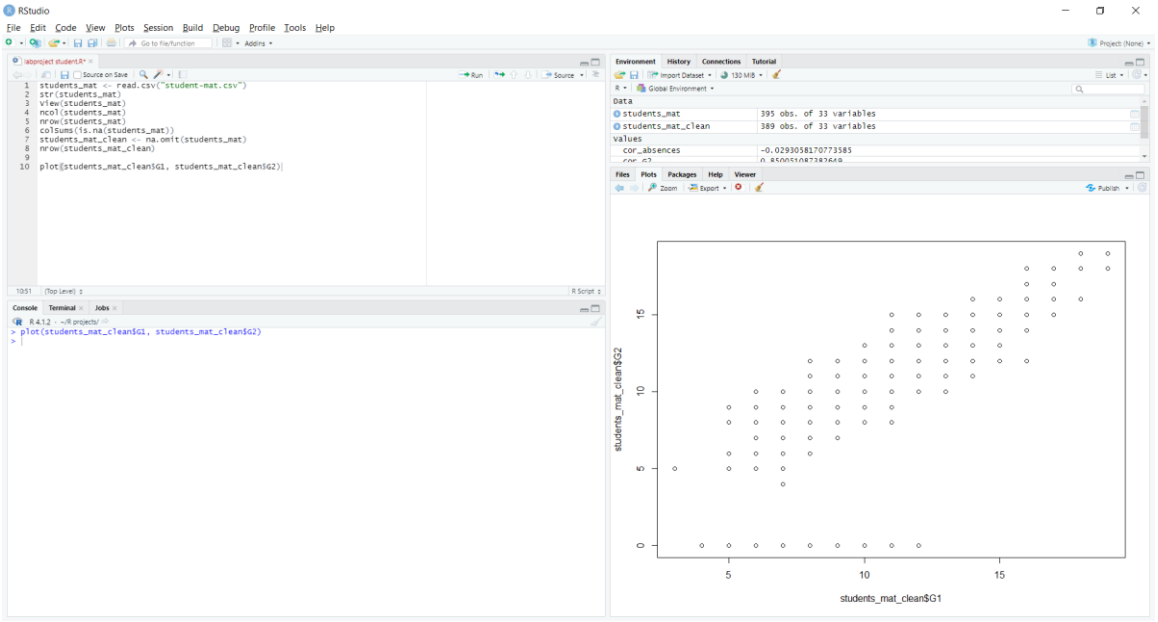
plot(students_mat_clean$G1, students_mat_clean$absences)



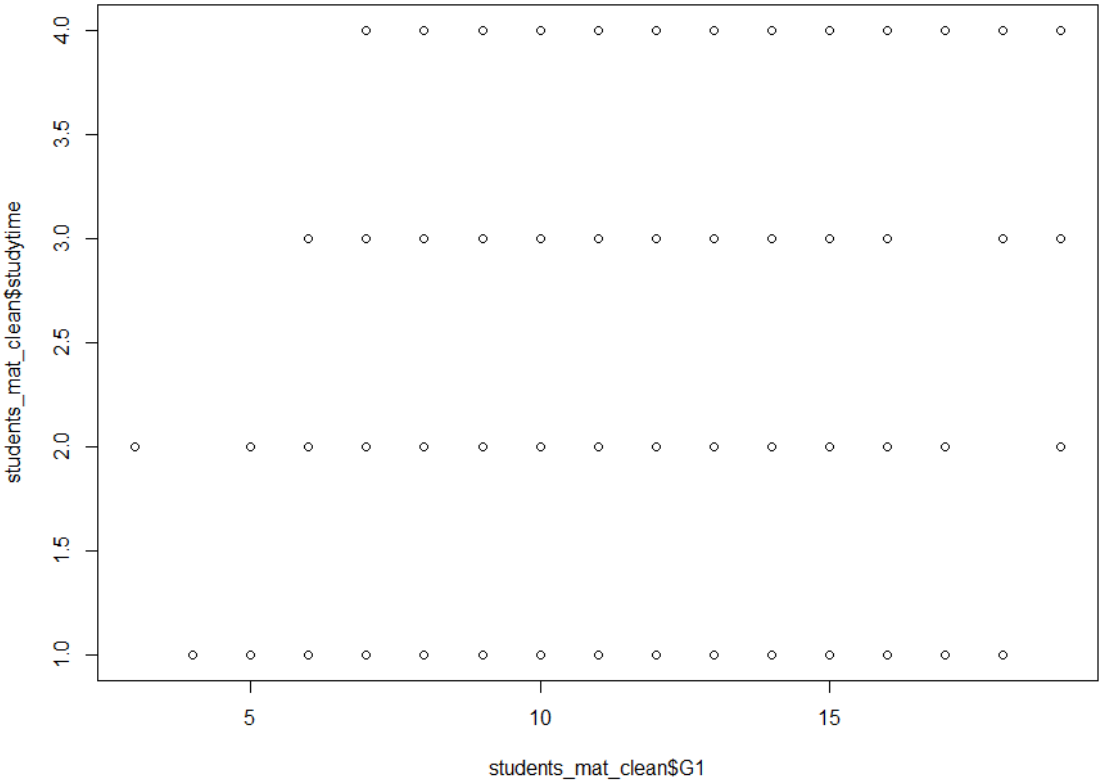


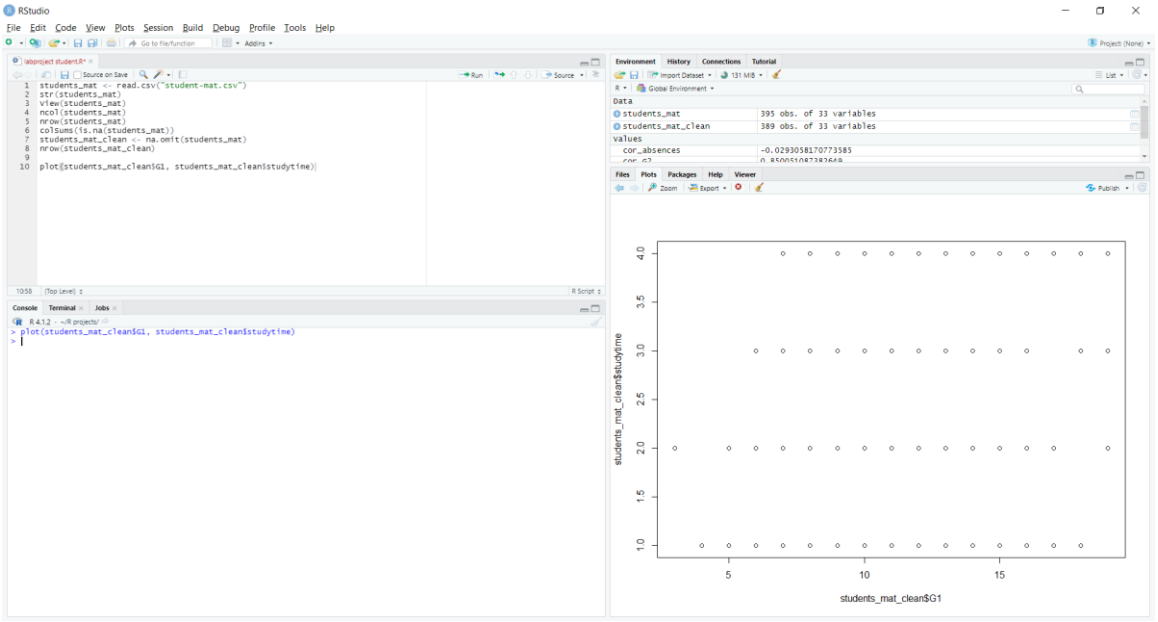

plot(students_mat_clean$G1, students_mat_clean$G2)



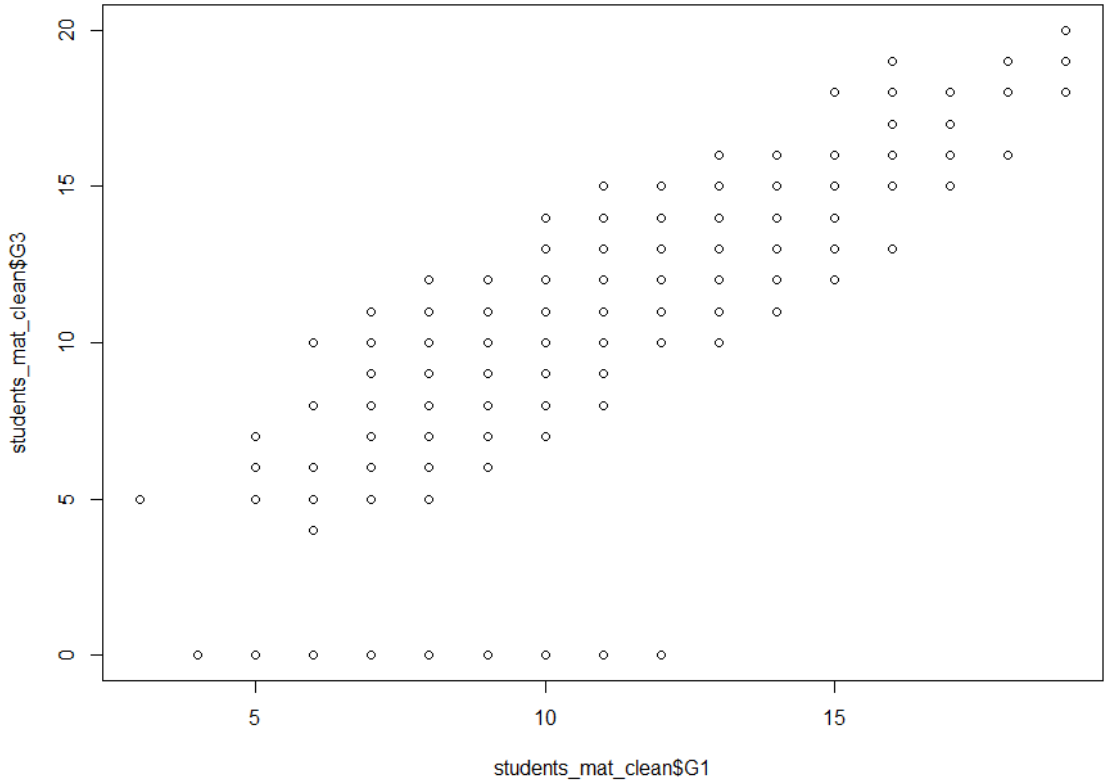


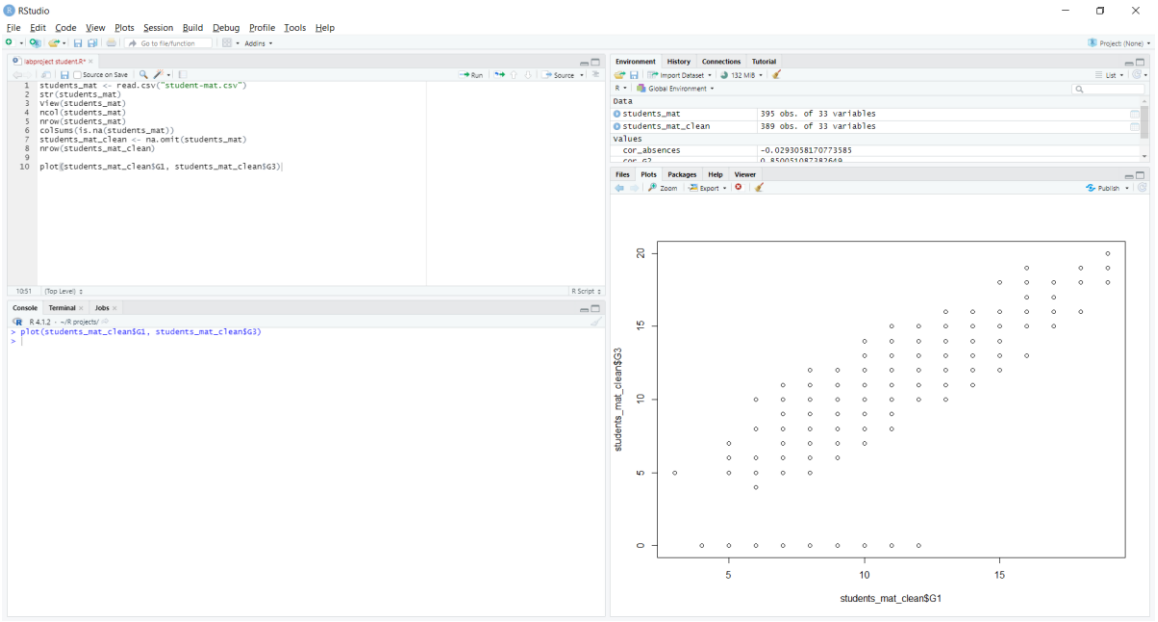

plot(students_mat_clean$G1, students_mat_clean$studytime)






plot(students_mat_clean$G1, students_mat_clean$G3)

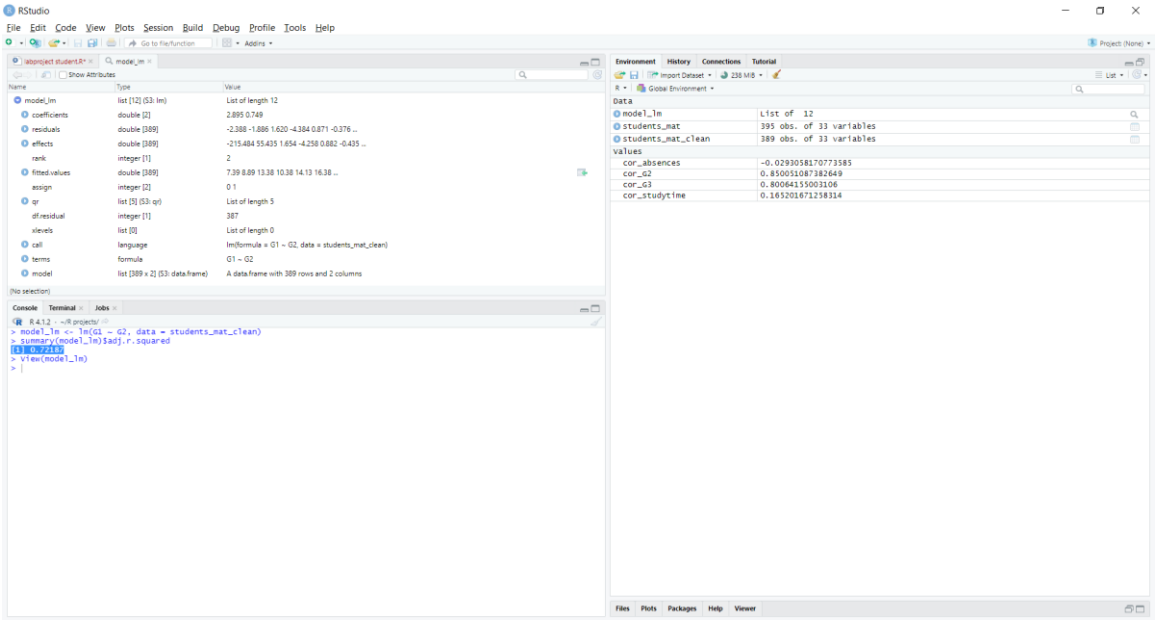





7. Apply linear regression algorithm on **"Grade1"** and **" Grade2"** attributes to calculate the Adjusted R Squared.

```
model_lm <- lm(G1 ~ G2, data = students_mat_clean)
summary(model_lm)$adj.r.squared
```

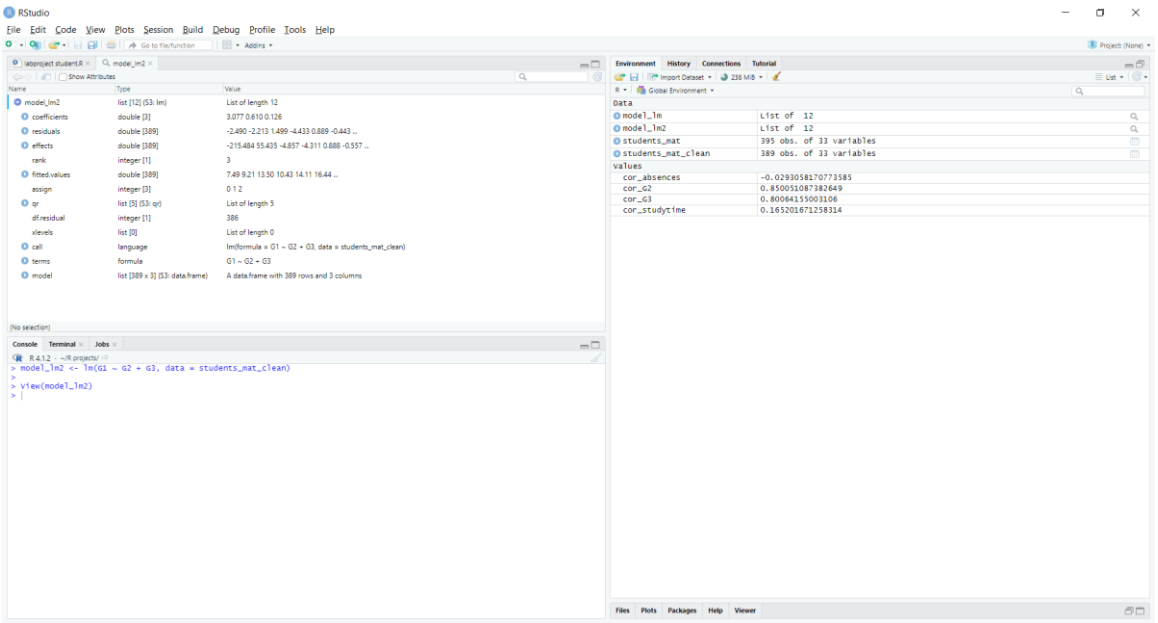
Adjusted R Squared = 0.72187



8. Apply the multiple regression algorithm to (G1,G2,G3) columns where the G1 as dependent variable and the two others as independent.

To apply multiple regression to predict "G1" based on "G2" and "G3", you can use the lm function:

model_lm2 <- lm(G1 ~ G2 + G3, data = students_mat_clean)

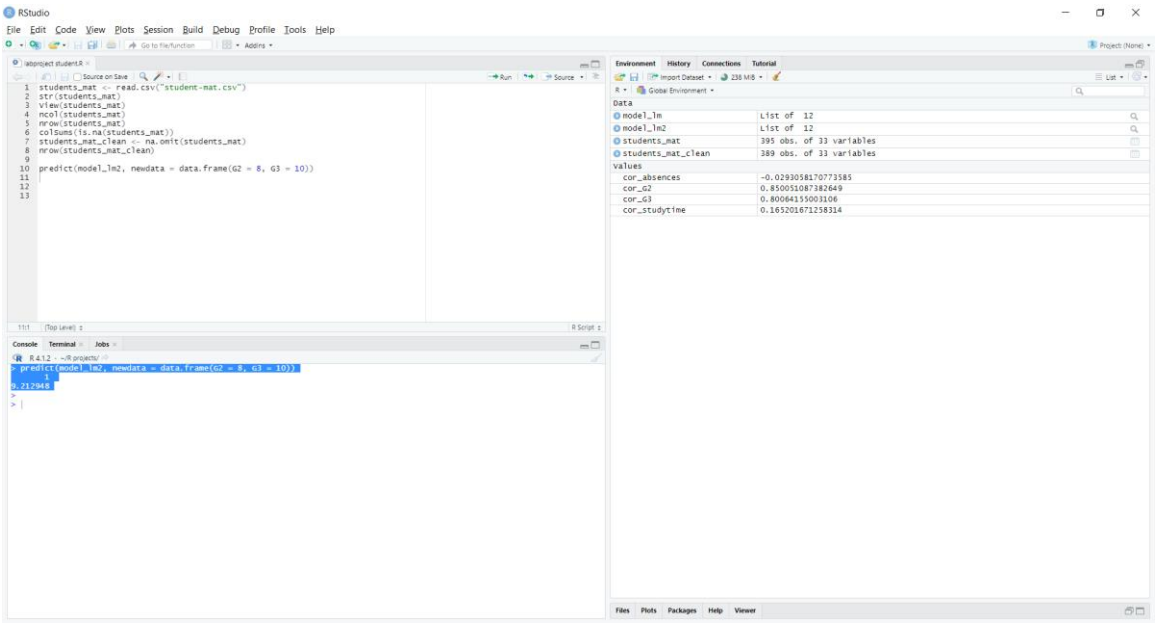



9. If we assume that the student got G2 as 8 and G3 as 10 what will be the value ofG1?

(Note: apply the equation for multiple regression model)

To make a prediction for "G1" based on the assumption that "G2" is 8 and "G3" is 10, you can use the predict function:

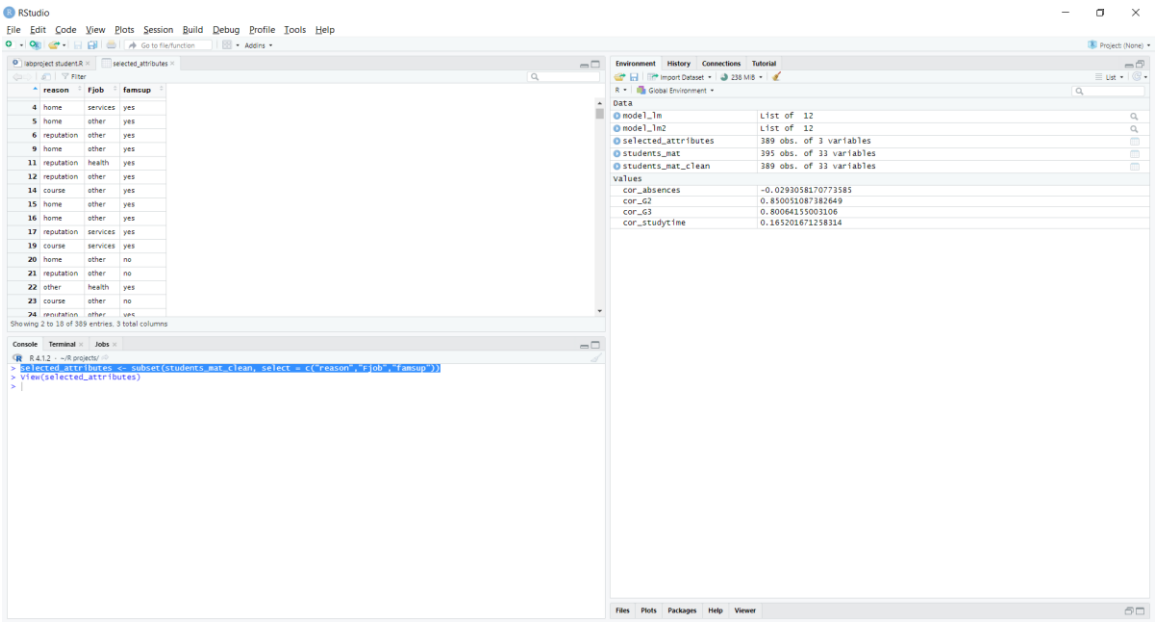
predict(model_lm2, newdata = data.frame(G2 = 8, G3 = 10))



10. Select the ("reason","Fjob" ,"famsup") attributes then Apply decision tree algorithm to the selected dataset by setting the training dataset as **80%**, the seed as **100**, to predict the "**reason**" **attribute. Then** Show the decision tree.

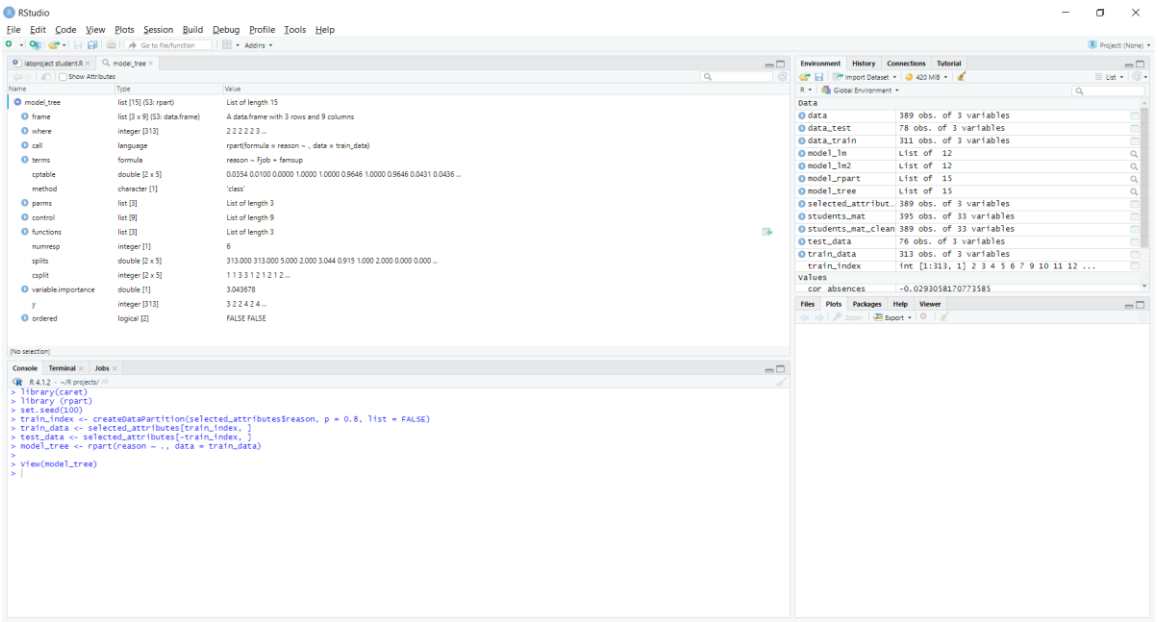
To select the "reason", "Fjob", and "famsup" attributes, you can use the subset function:

```
selected_attributes <- subset(students_mat_clean, select = c("reason","Fjob","famsup"))
```



To apply decision tree to predict the "reason" attribute, you can use the rpart function from the "rpart" package:

```
library(caret)
library (rpart)
set.seed(100)
train_index <- createDataPartition(selected_attributes$reason, p = 0.8, list = FALSE)
train_data <- selected_attributes[train_index, ]
test_data <- selected_attributes[-train_index, ]
model_tree <- rpart(reason ~ ., data = train_data)
```



To show the decision tree, you can use the plot function from the "rpart.plot" package:

```
library(rpart.plot)
prp(model_tree)
```