**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

**PART B**

(PART B: TO BE COMPLETED BY STUDENTS)

| Roll No: C035 | Name: Krisha Goti |
|---|---|
| Class: B | Batch: EB2 |
| Date of Experiment: 23/07/2022 | Date of Submission |
| Grade | |

## B.1 Work done by student

*(Paste your gather information and the comparison table)*

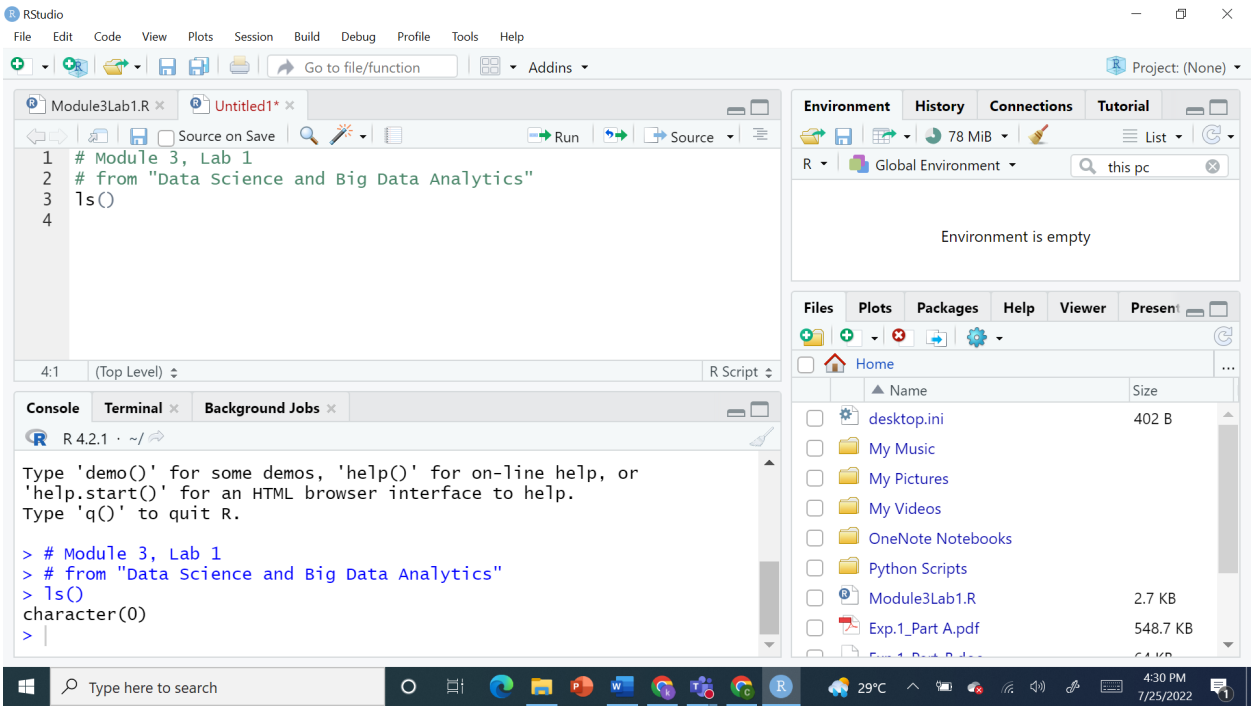
1. **Examine the Workspace:**

Type the following command into the R command panel, and hit [ENTER]
**ls()**
You should see the following:
**character(0)**

**Note:** R is telling you that you have nothing in your workspace.

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |




## 2. Getting Familiar with R

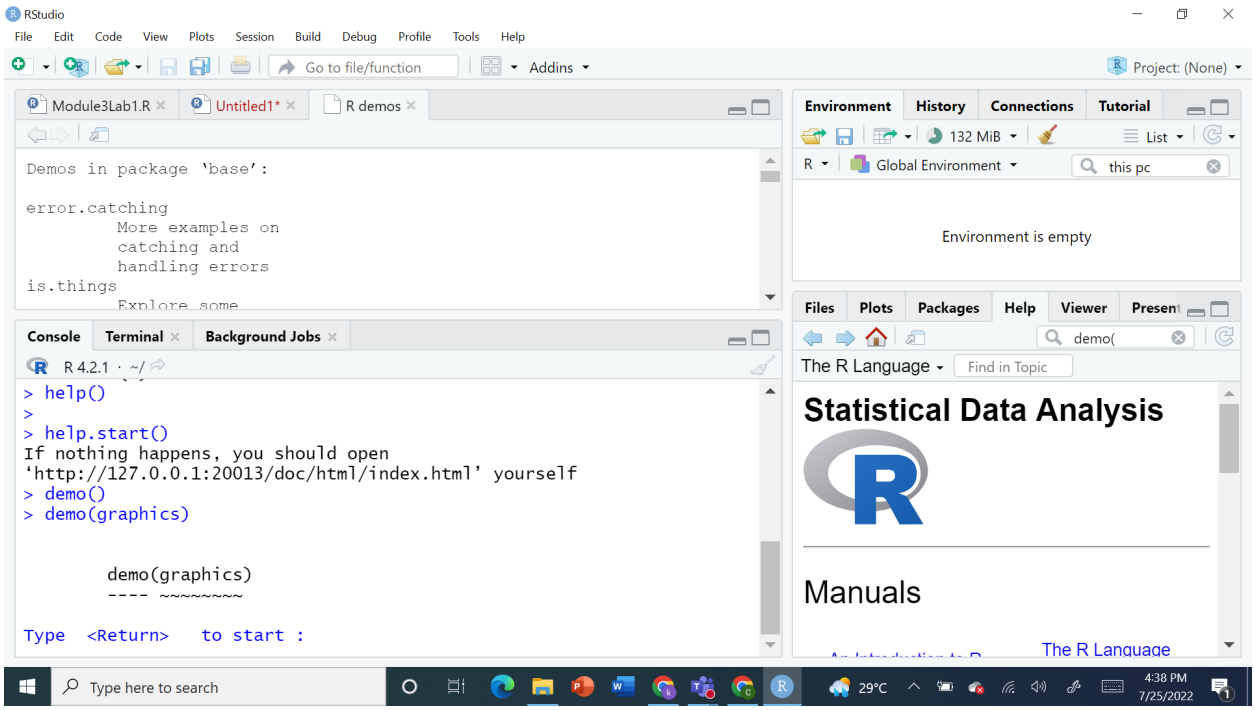
1. Click each tab in each panel. What happens?
2. Type the following commands into the R command panel

**help()**
**help.start()**
**demo()**
**demo(graphics)**

**Note:** Hit esc to exit out of the demo

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |




## 3. Working with R:

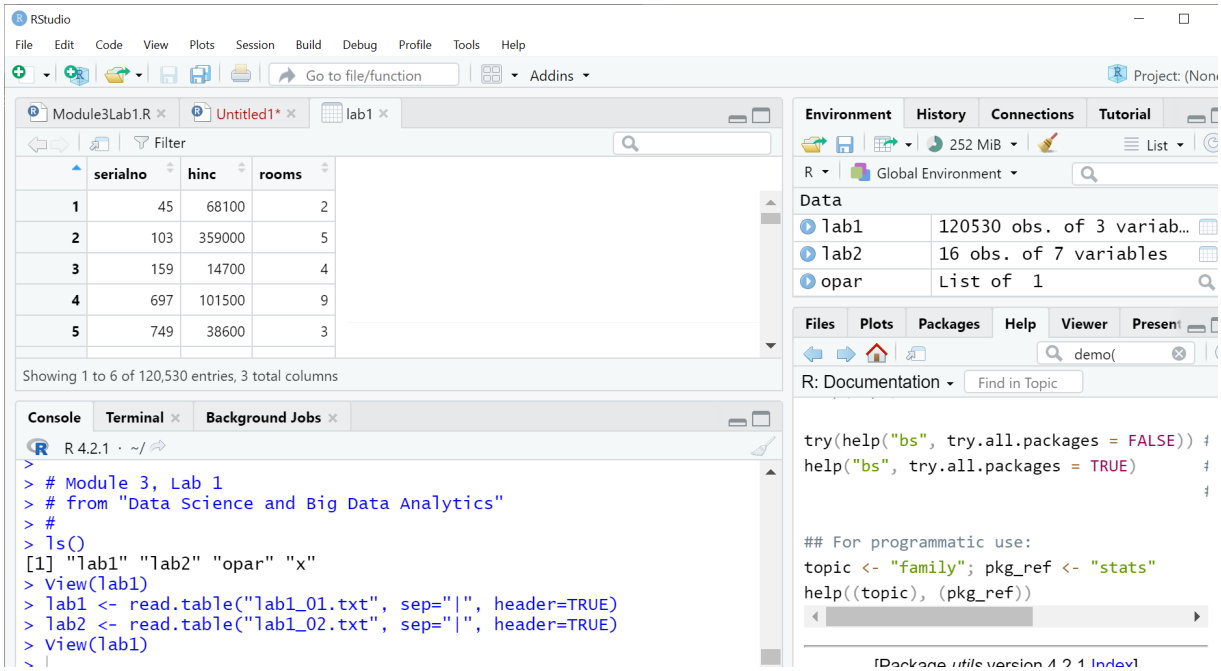
Load the .txt files you created in the first lab. Load the first file, **lab1_01.txt**

1. Set the working directory to LAB01 where we have stored the data. On the console window type:
   **setwd("~/LAB01")**

2. Select the line and press <ctl>**Enter**:
   **lab1 <- read.table("lab1_01.txt", sep = "|", header=TRUE)**

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

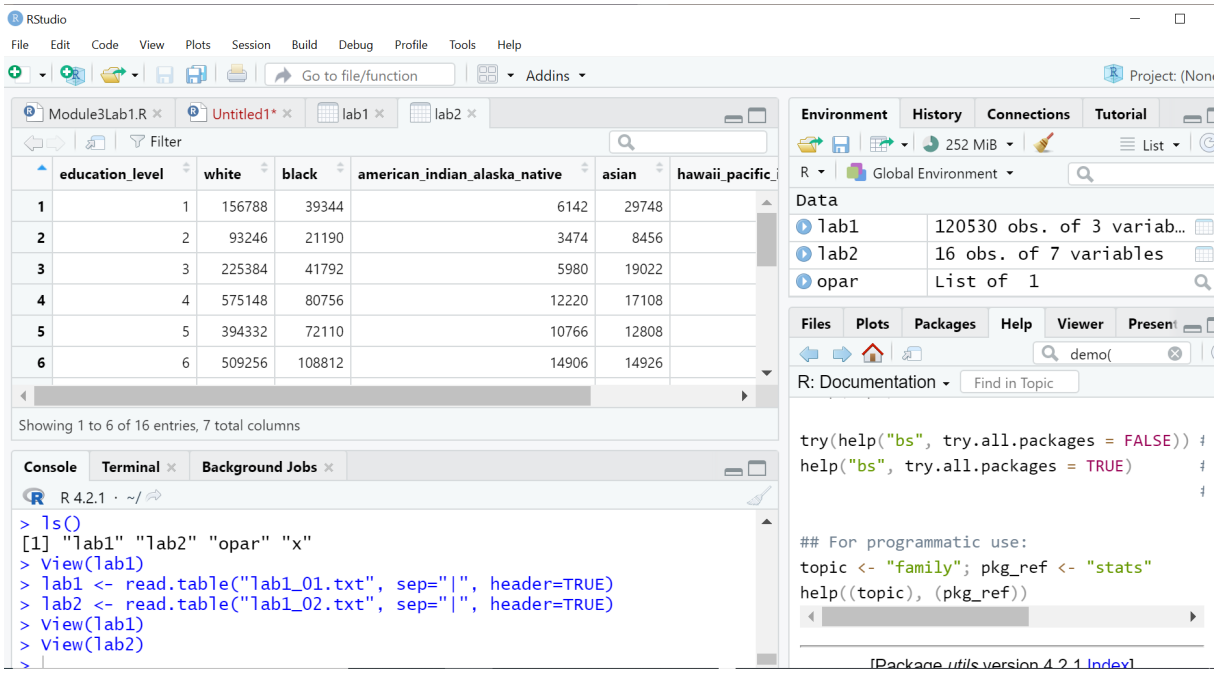



- If correct, R will simply return you to the command prompt ("> ").

3. Now load the second .txt file, **lab1_02.txt**, by modifying the command (using the line of code in the RStudio command panel) you just entered.
   (Use the up/down, left/right arrow buttons to move from and within lines; change each occurrence of "lab1" to "lab2".)
   The command should read:
   **lab2 <- read.table("lab1_02.txt", sep = "|", header=TRUE)**

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |




4. When you have completed the edits, make sure that your cursor is within the line, press **Enter**.
   **Note:** R supports copy and paste, as well as up and down arrows for moving to previous commands, left and right arrows to move within/between lines and home/end to move to the beginning or end of a line.

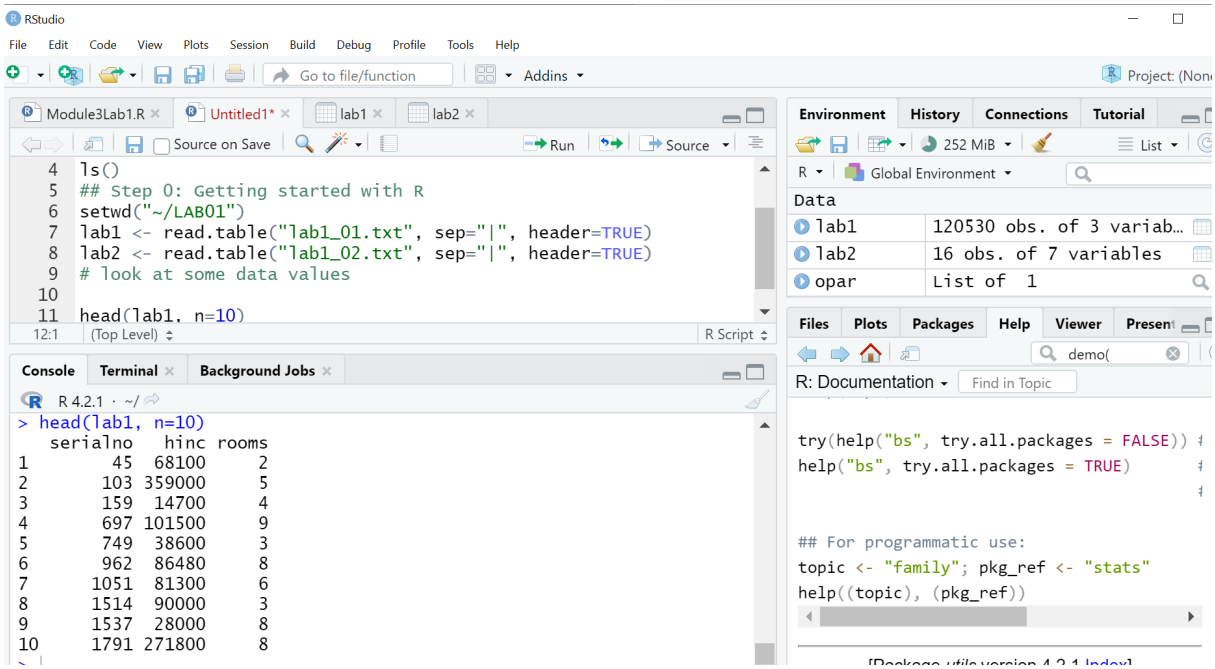
## 4. Verify the Contents of the Tables:

It is always a good idea to look at the data to make sure that everything works. You can use the **head()** command to print out the first 6 lines of a table or the, **tail()** command to print out the last 6 lines of the table.

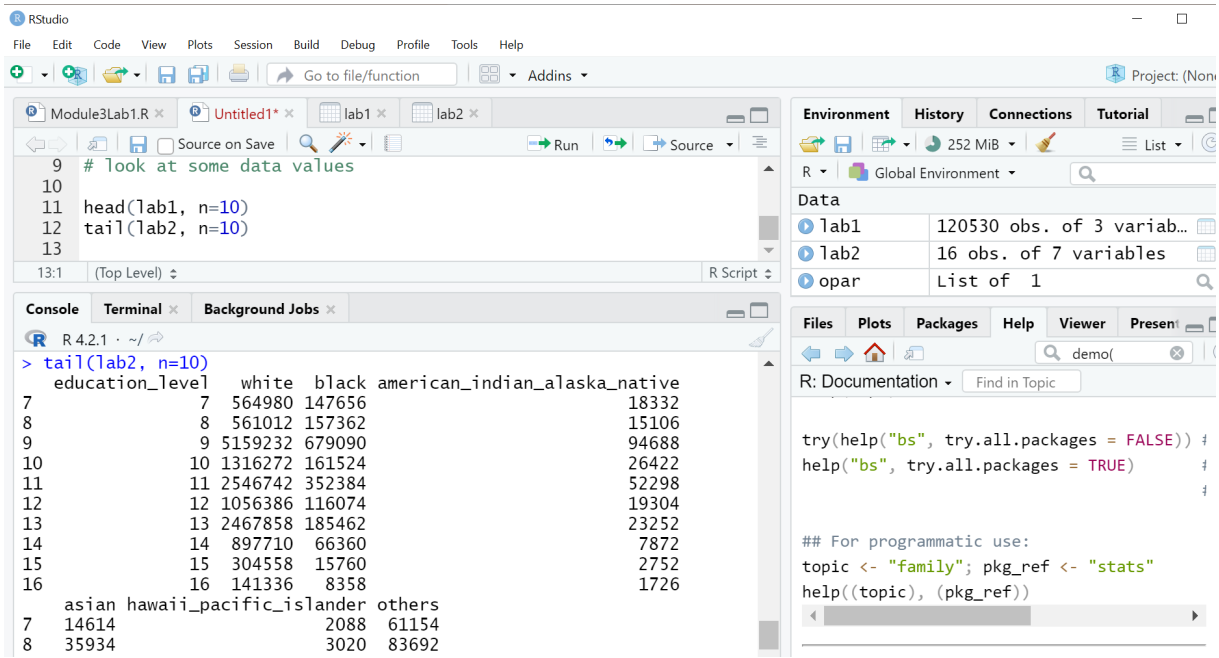
1. Select and run the command:
   **head(lab1,n=10)**
   Record the value of the 10th line here:

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |




2. Now do the same for the lab2 table, but use the **tail(lab2, n=10)** command instead.




3. Record the value of the 1st line here: 1 45 68100 2

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

## 5. Manipulating Data Tables (data frames) in R:

Examine the contents of the table in more detail.

1. Execute the following command:
   **summary(lab1)**
   Ignore the values for the *hinc* and *rooms* columns for now. The *serialnoid* field represents a unique identifier (it's the household identifier) from the Postgres database. You no longer need it and it will interfere with some of the procedures you want to run against this data set, so create a copy of the lab1 table without that column.

```
> summary(lab1)
    serialno            hinc             rooms
 Min.   :     45   Min.   :      4   Min.   :1.000
 1st Qu.:2489537   1st Qu.:  26000   1st Qu.:4.000
 Median :4992025   Median :  50300   Median :6.000
 Mean   :4996363   Mean   :  67152   Mean   :5.627
 3rd Qu.:7500553   3rd Qu.:  84200   3rd Qu.:7.000
 Max.   :9999998   Max.   :1620560   Max.   :9.000
```

2. Select and run:
   **nlab1 <- lab1[,2:3]**
   This uses a feature of R that allows us to refer to rows and columns in a dataframe as if they were entries in a matrix. A blank entry in a row or column position means "use all available." This statement says: use all the rows in the table, but only use columns 2 and 3

   You could have used the following for the same effect (Note that the following code is not part of the script you can see in the source file *Module3lab1.R)*:
   **hinc <- lab1$hinc**
   **rooms <- lab1$rooms**
   **nlab11 <- data.frame(hinc, rooms)**

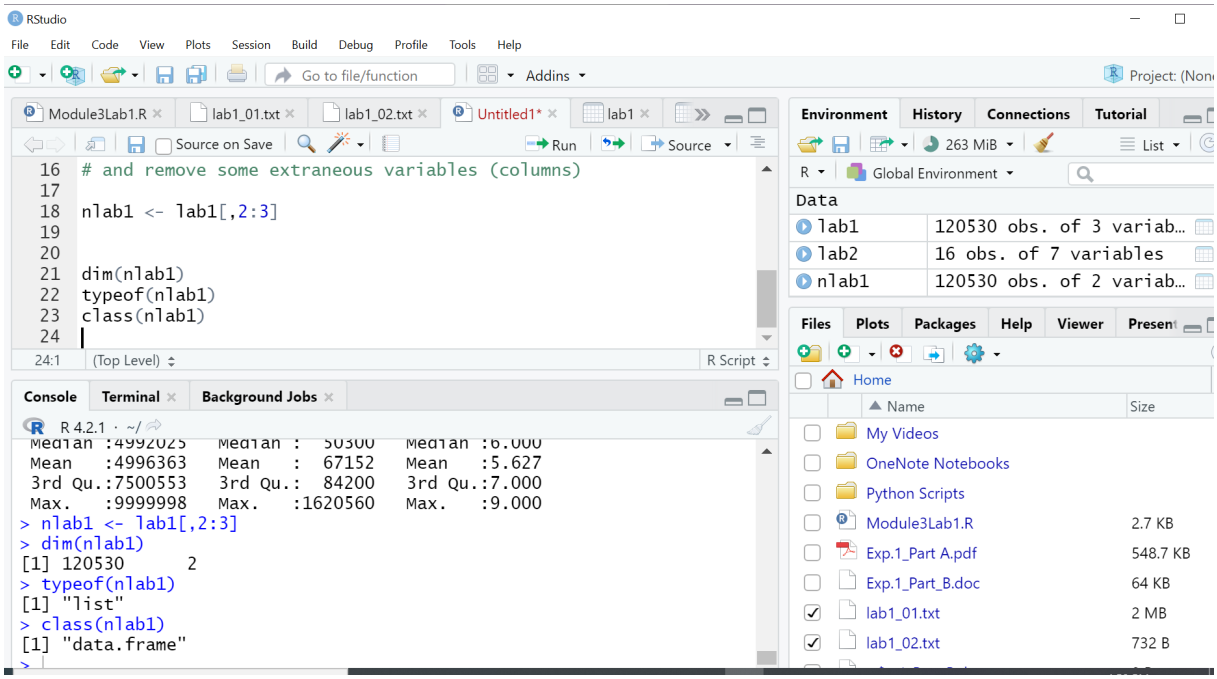
   You're taking advantage of R behavior that names the columns after the name of the variable. You could have used the following for the same effect:
   **nlab11 <- data.frame(lab1$hinc, lab1$rooms)**
   **names(nlab11) = c("hinc", "rooms")**

3. The dim(<table>) has the nice property of telling us how many rows exist in the table. Execute the following commands:

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

**dim(nlab1)**
**typeof(nlab1)**
**class(nlab1)**

Each of these commands tells us something about this particular object. You may not use these often, but they can be useful when R complains that it doesn't like something about the object that you just used.

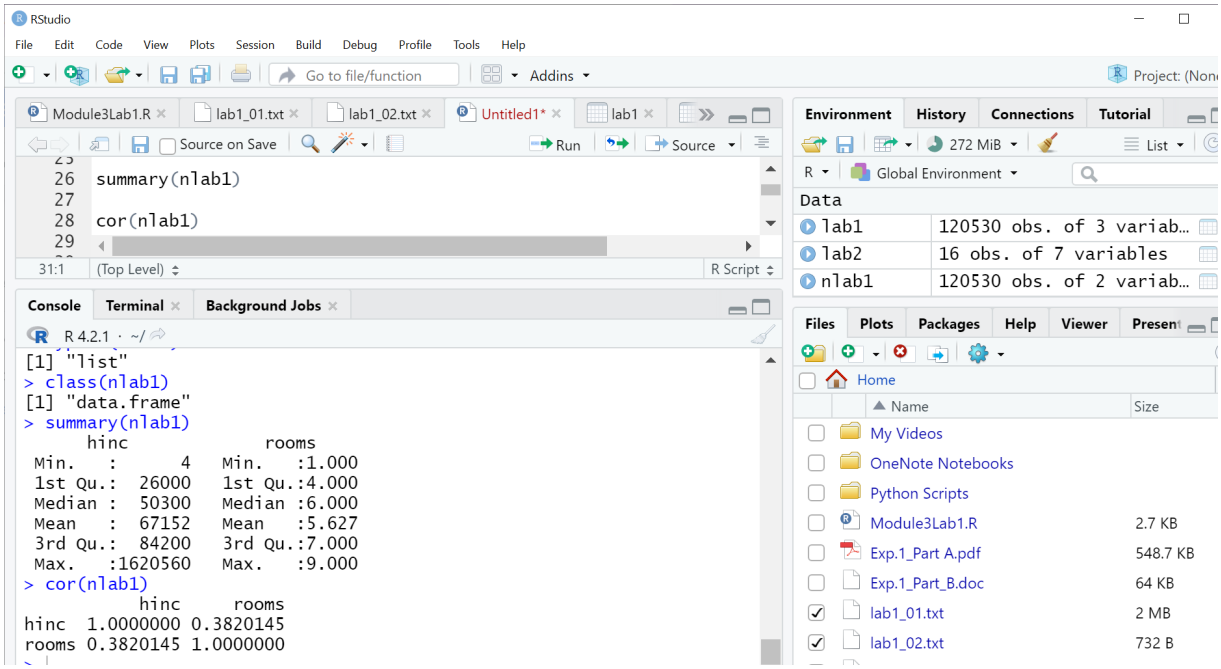



## 6. Continue to Investigate Your Data:

1. Select and execute the following commands:
   summary(nlab1)
   cor(nlab1)
   The summary function for data frames prints out summary statistics.

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |




2. Compare the median and the mean. What does it mean if the mean is less than the median? then data is skewed to the left
3. How about the mean greater than the median? the distribution is positively skewed
4. Does the min and max value for the quartiles make sense to you?

   Yes, the minimum and maximum values tells us about the most extreme values in the data set which then helps us to find and analyze the data easily.

Here again you have a chance to do further cleaning of your data sets, but postpone this until you've finished the next few lessons.

5. How do the values returned by the cor() function differ from the results obtained in lab 1?
   The values returned by the correlation function is 0.38 which shows that it is less correlated with the results obtained in lab1.

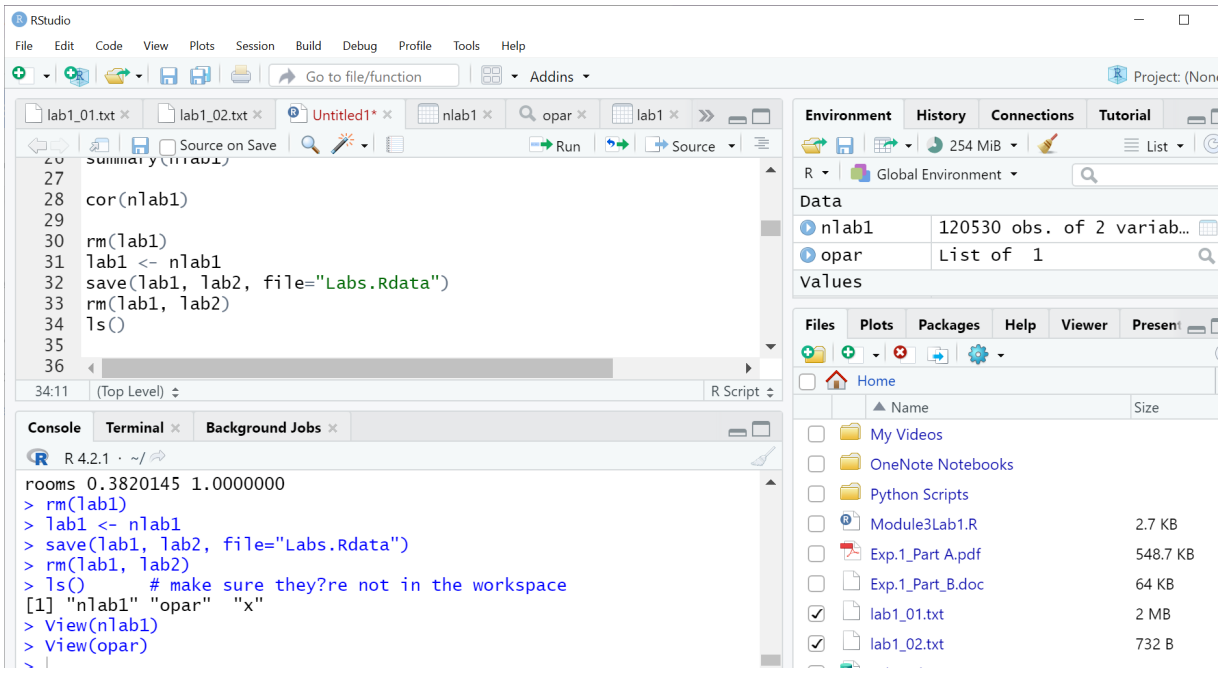
## 7. Save the Data Sets:

1. Execute the following commands:
   **rm(lab1)**
   **lab1 <- nlab1**
   **save(lab1, lab2, file = "Labs.Rdata")**
   **rm(lab1, lab2)**

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

**ls()** # make sure they"re not in the workspace




## 8. Examine Your Data:

1. Experiment with some of the examples used in the lecture portion of this lesson. Using the same selection techniques that you used earlier, run each line in the file.
    - Some commands don't print their results. If this is the case, type in the value of the variable you created in the command window. If the variable was named "x", you can type "x". You can also type "print(x)" which will do the same thing.
2. Experiment with R functions that identify the class and data type of a particular variable, type: **typeof(x), class(x), attributes(x), names(x), dim(x)**

3. Which ones work on which kind of data types?

Typeof: works with all kinds of data types as it is used to show the type of data given variable consist of on low-level.

Class: works with all kinds of data types as it is used to show the type of data given variable consist of on high-level.

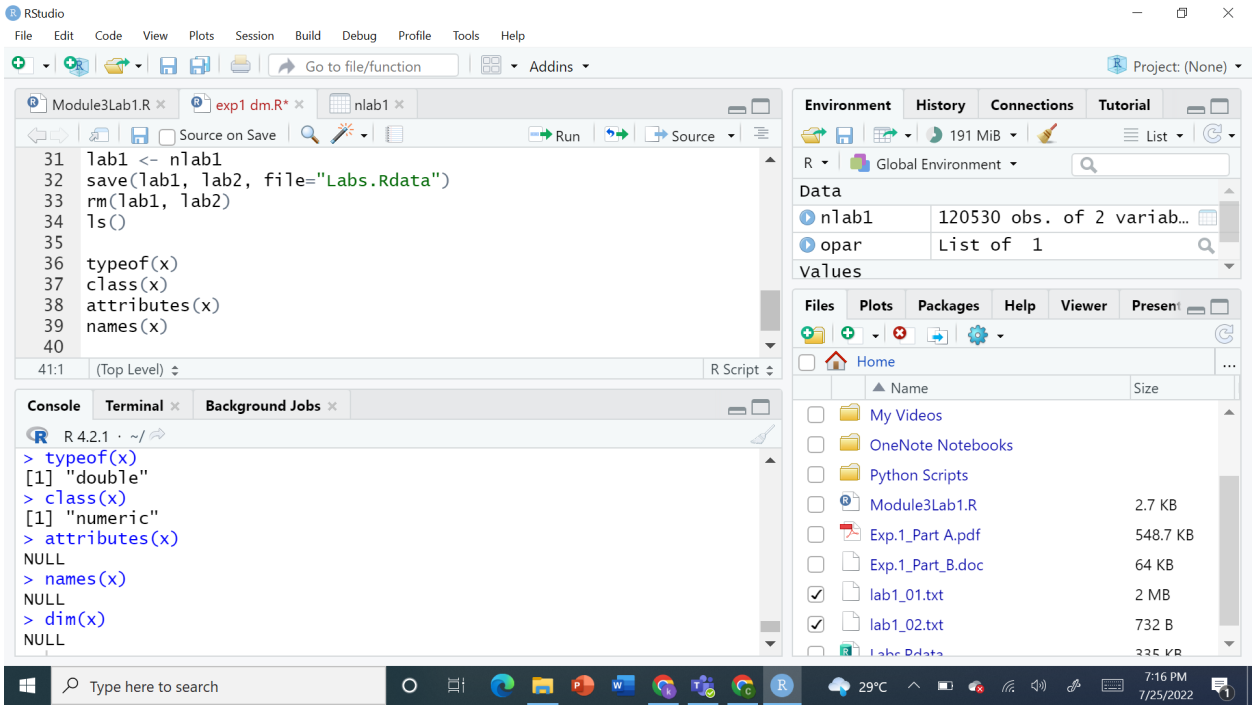
| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

Attributes: vector, matrix, data frame.

Names: vector, matrix, data frame.

Dim: vector, matrix, data frame.

4. Type these values into the RStudio command panel.




5. Typing all these commands for each variable is tedious. Alternatively, we will write a function *tellme* that takes a variable as an argument and performs typeof, class, names and str on that variable.

   Select and run the lines beginning with **"tellme <- function(x)**{extending through the right curly brace.

6. Now execute the following command

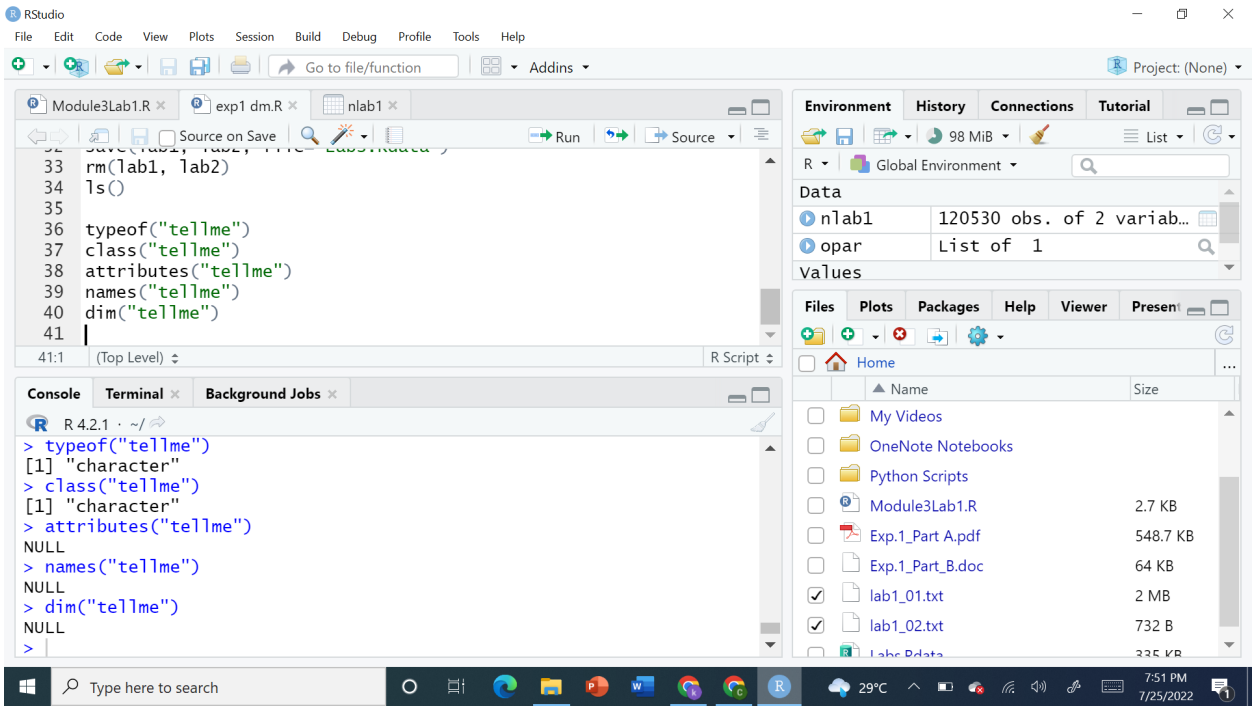
   **tellme**

   You should see the definition of the function that you just entered! This is because R doesn't interpret a plain **tellme** as a function, but rather as an object to be printed out. The

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |

default print function for a function is to print its definition. You can try this with any other R function. Type **mean** and inspect the results.

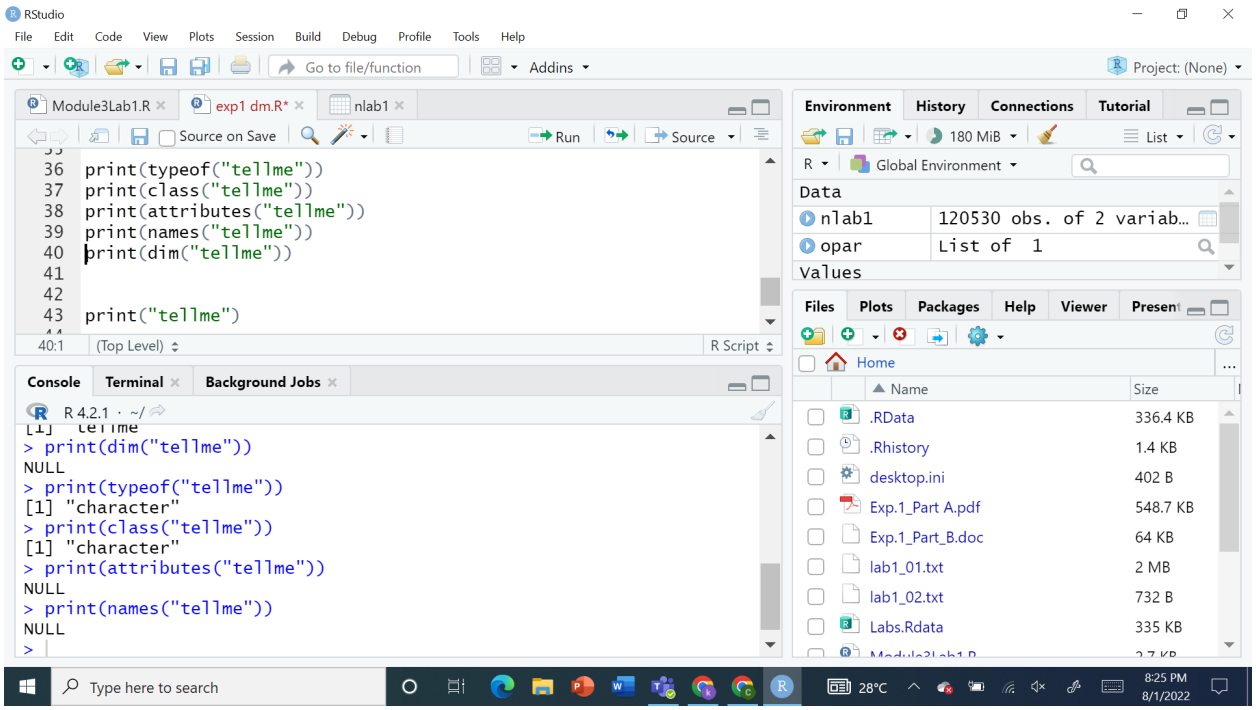
7. Try tellme**()** with a series of variables.




8. Which commands actually list something? print()
9. How might you get the other commands to list their return value?
   [Hint: try print()]

**SVKM'S NMIMS Deemed-to-be-University**
**Mukesh Patel School of Technology Management & Engineering**
**Department of Computer Engineering**

| Program | BTech Intg. | Branch | Computers |
|---|---|---|---|
| Semester | IX | Year | V |
| Name of the Faculty | Artika Singh | Class | Div B and Div C |
| Course Title | Data Mining | Academic year | 2022-23 |




## B.2 Conclusion

After completing this experiment, I have understood the basics of R studio, learned how to apply appropriate analytic techniques and tools to analyze big data, create statistical models and identify insights leading to actionable results.