



MID-TERM PROJECT REPORT

Introduction To Data Science

Faculty

TOHEDUL ISLAM

Assistant Professor, Computer Science Dept.

MD. ABDULLAH AL NASIR

ID-19-41052-2

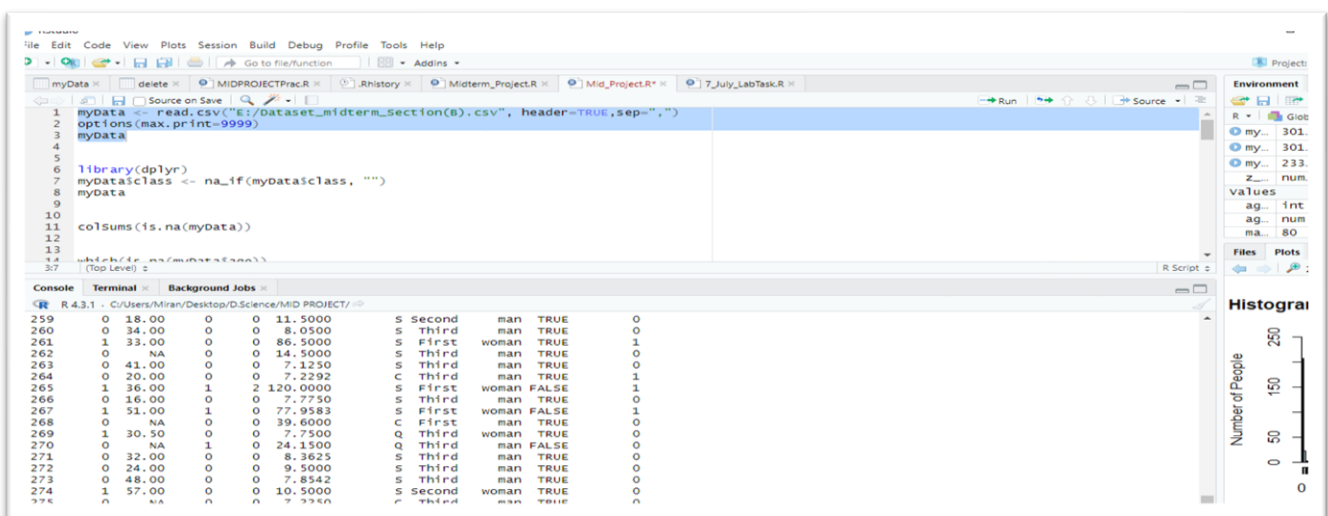
Section-Introduction To Data Science [B]

American International University – Bangladesh (AIUB)

About Dataset: Here, the dataset named "Titanic dataset" has 301 instances and 10 attributes. Attributes are "Gender", "age", "sibsp", "parch", "fare", "embarked", "class", "who", "alone" and "survived".

Here, six attributes are numerical. Those are: Gender, age, sibsp, parch, fare, embarked, and survived. The attribute named "class" is a categorical attribute. The target attribute of this dataset is "survived".

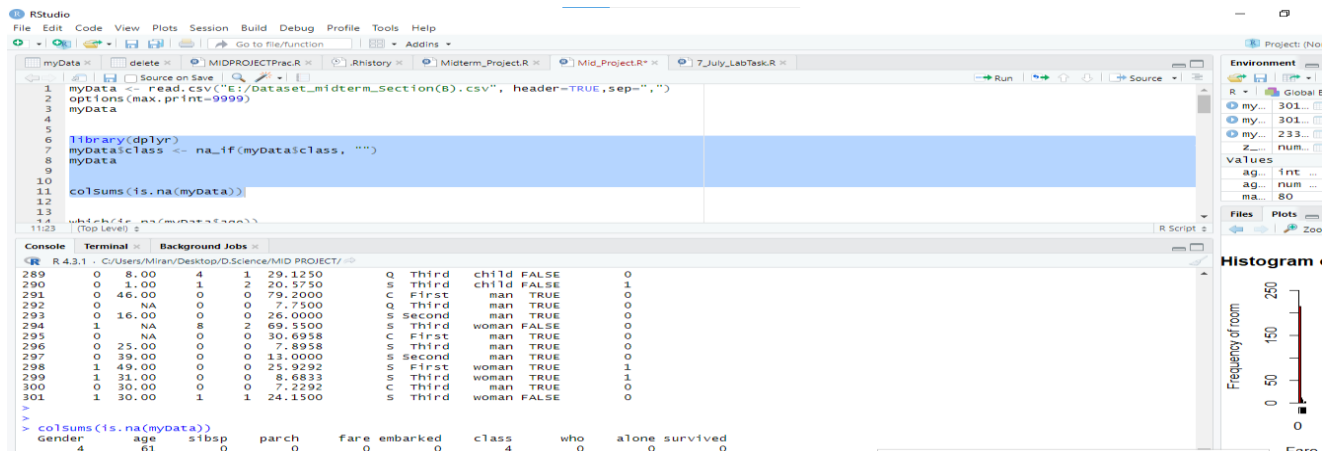
Step- 01: In the very first step, the dataset has been imported from the drive in a data frame named myData using the read function. Using the options () function, we have imported all the data values with the help of max.print=9999.



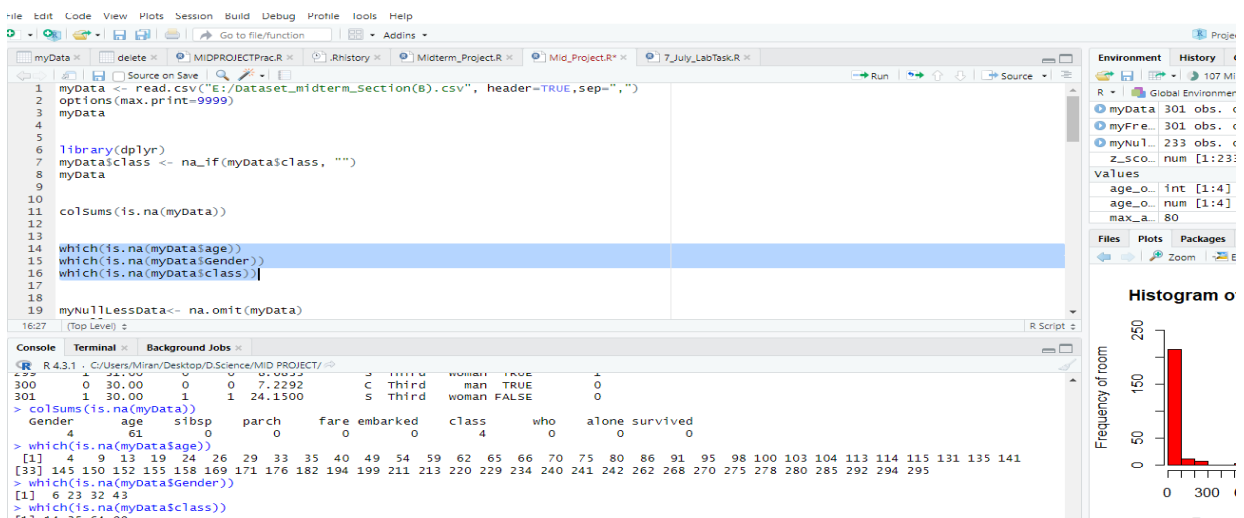
Step- 02: The library named "dplyr" has been loaded.

The na_if() function finds out if the cell is empty or not; if the cell is empty, then it replaces the value with NA. In the class attribute of the dataset, the na_if() function is called to replace all the missing values with NA.

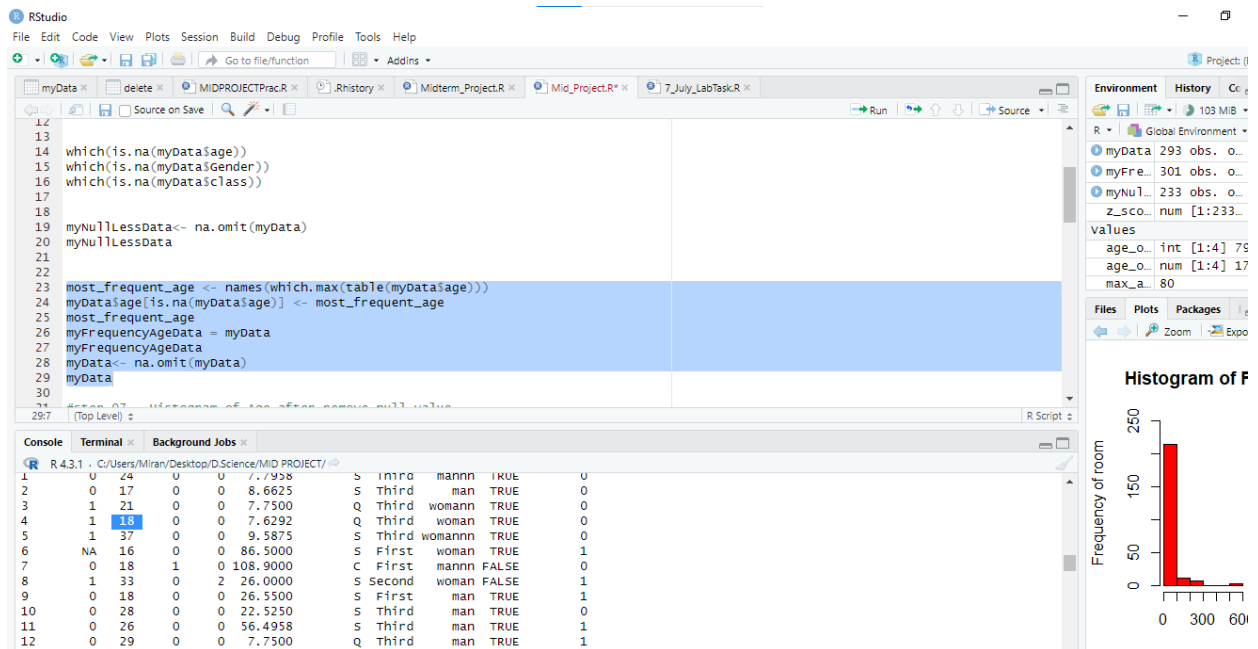
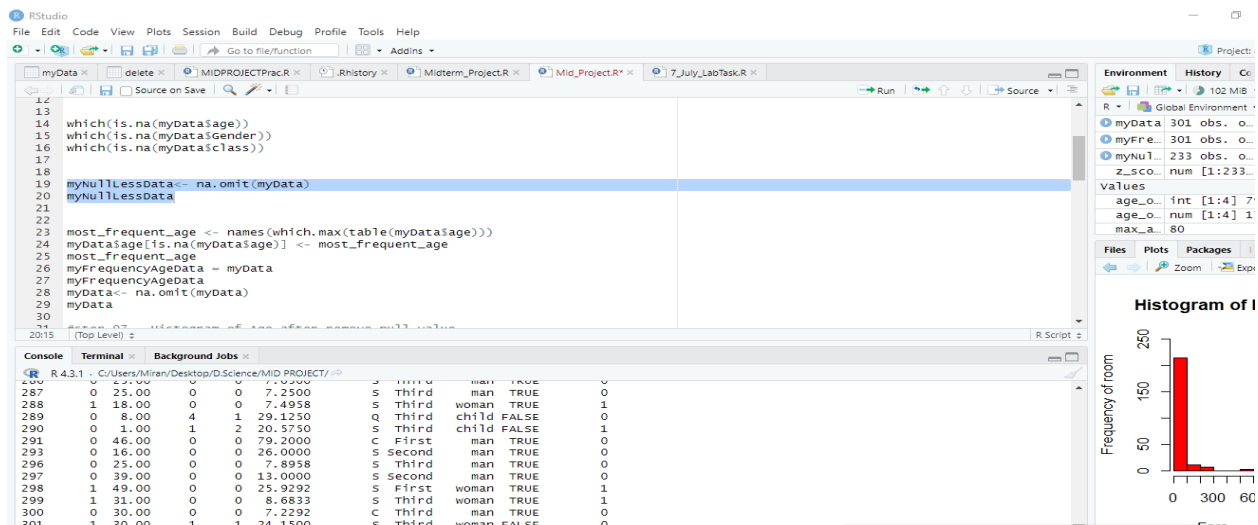
The function colSums () is used to calculate the column sums of a matrix or data frame. The function is.na() is used to identify null values (NA) in a vector, matrix, or data frame. To calculate the number of missing values in each column, the function used here is colSums(is.na()), keeping myData as a parameter. Here we got that there are 4 null values in Gender, 61 in age and 4 in class.



Step- 03: The function `which()` is used with `is.na()` to find out the specific row where the null value is located. Here, rows containing null values have been figured out for the attributes named gender, age and class.

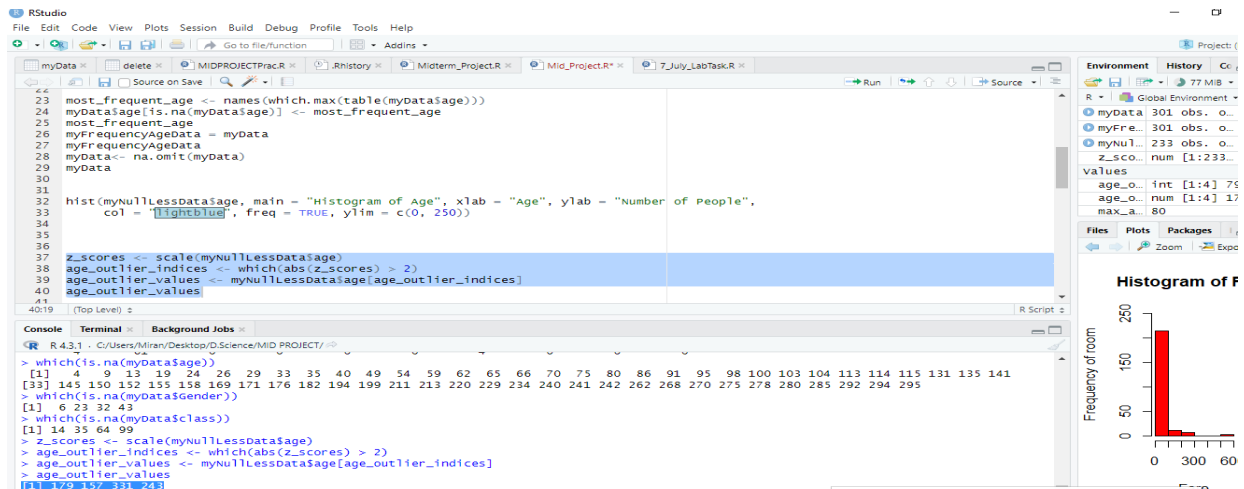


Step- 04: Discarding an instance is the simplest strategy to get rid of missing or null values in an instance. Therefore, a new data frame named 'myNullLessData' was created. Using the function `na.omit()`, myData has been modified and kept in a new data frame. Then myNullLessData printed in the console.

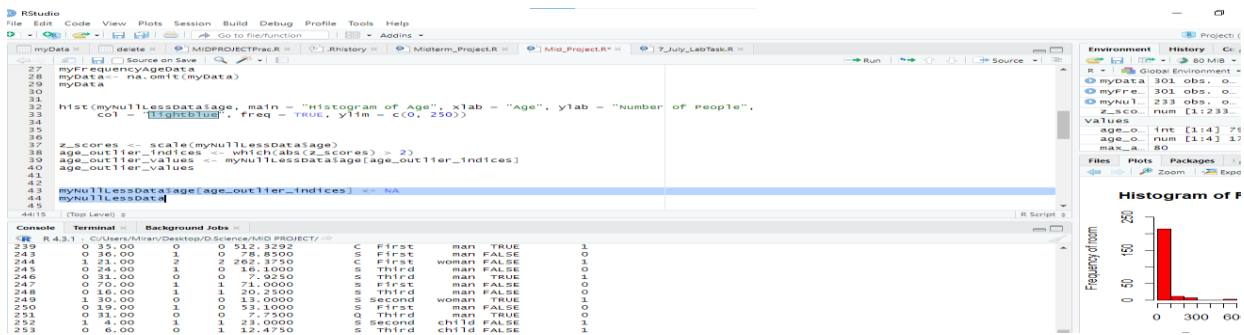


Step- 07:

Updating myNullLessData. Finding the outliers using scale(myNullLessData). The scale() function is used to identify outliers by examining the standardized values of age, abs() returns the absolute value of a number. Outliers we found here are 179,157,243,331.

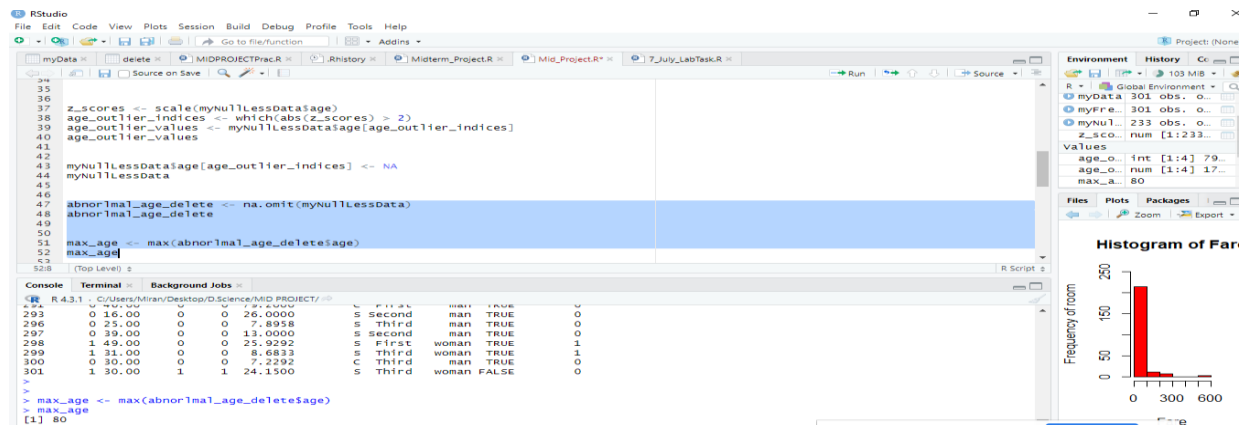


Step- 08: Now the outliers in the array named [age_outlier_indicates] are replaced with NA.

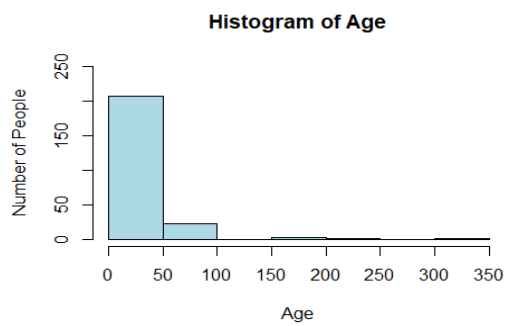


Step- 09: Now I'll delete abnormal age values that were previously replaced with NA. For doing so, na.omit() values containing NA are deleted and then printed in the console.

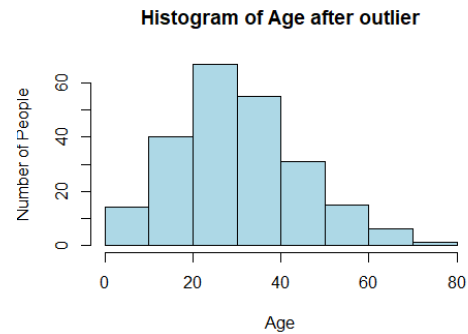
Then, Using the max () function, I figured out the maximum age to make sure that all the outliers were deleted. Here, the maximum age I found is 80, which seems relevant.



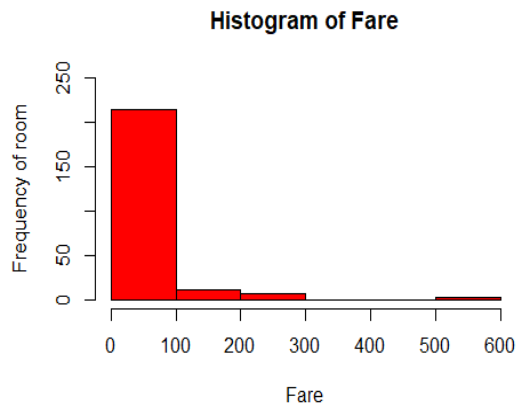
Step- 10 & Step- 11:



Histogram of Age before outliers detection and deletion.



Histogram of Age before outliers detection and deletion.



Histogram of Fare keeping frequency of room on Y-axis and Range of fare on X-axis.

```


63
64
65 sd(abnormal_age_delete$age)
66 sd(myNullLessData$fare)
67
68
69 myNullLessData$who <- gsub("^mann$", "
70 myNullLessData$who <- gsub("^woman$",
71
66:24 (Top Level)

```

Console

Terminal x

Background Jobs x

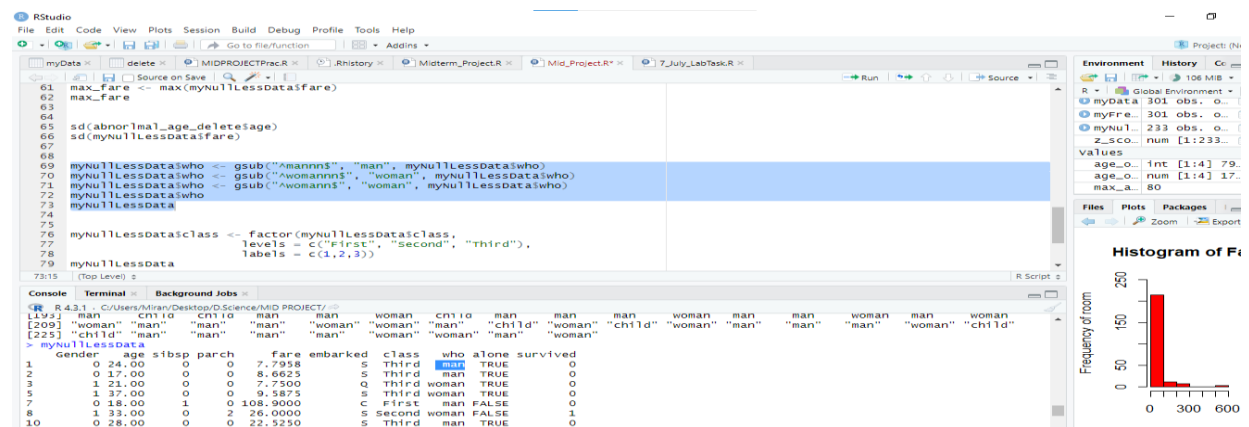
 R 4.3.1 · C:/Users/Miran/Desktop/D.Science/MID PROJECT/

```

[1] 80
> hist(myNullLessData$age,main = "Histogram of
1e",
+       col = "lightblue", freq = TRUE)
> hist(myNullLessData$fare, main = "Histogram
+       col = "red", freq = TRUE, ylim = c(0,
> sd(abnormal_age_delete$age)
[1] 14.30402
> sd(myNullLessData$fare)
[1] 61.38566
> |

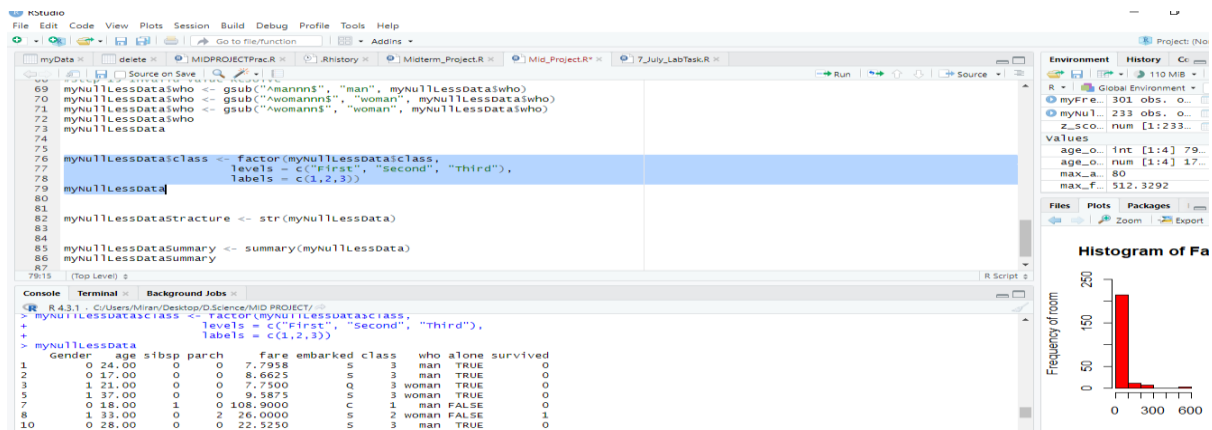
```

The syntax of the `gsub()` function is as follows: `gsub(pattern, replacement, vector where to perform substitution)`. `Mann` are replaced with `man` and `womannnn`, `womann` are replaced with `woman` in `who` column. Lastly showed in the console.

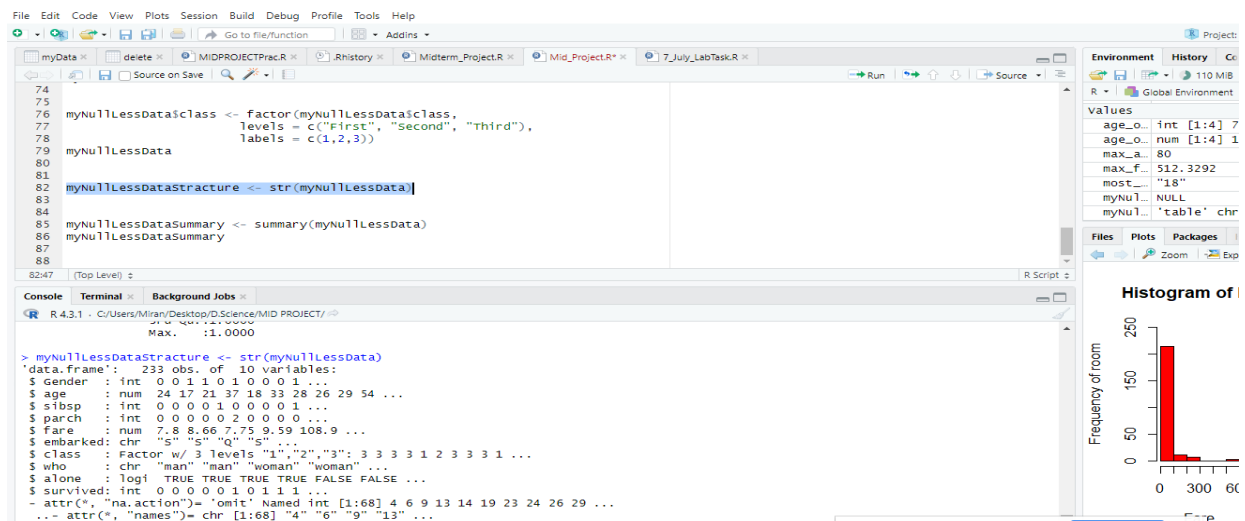


Step- 14: Annotations are used to represent categorical values. The attribute named “class” has categorical values. The function used here is factor() and the syntax is as follows:
factor(mynullLessData, levels, labels).

So, First, Second, Third has annotated as 1,2,3 respectively.



Step- 15: The str() function is used to display the structure of an R object. It provides a summary of the structure of an object, such as its type, dimensions, and content, especially useful for exploring data frames. Using str(), the summary of the structure of 'myNullLessData' has been shown.



Step- 16: The function `summary()` provides a summary of key statistics like minimum, median, mean, 3rd quartile, and maximum for numeric variables. `summary()` depends on the class of the object being summarized. Different classes provide different types of summaries.

