Basic in R

PETER NDUNGU

2024-10-21

##BASIC IN R  
  
# Input data  
age <- c(2, 2.5, 3, 4, 4.5, 4.5, 5, 3, 6, 6.5)  
mileage <- c(22, 34, 33, 37, 40, 45, 49, 30, 58, 58)

Reading the data.

# Create a data frame  
mydata <- data.frame(Age = age, Mileage = mileage)

creating a data frame.

# Display the data frame  
print(mydata)

## Age Mileage  
## 1 2.0 22  
## 2 2.5 34  
## 3 3.0 33  
## 4 4.0 37  
## 5 4.5 40  
## 6 4.5 45  
## 7 5.0 49  
## 8 3.0 30  
## 9 6.0 58  
## 10 6.5 58

the data frame has two columns.

# Save the data frame as a CSV file  
write.csv(mydata, file = "mydata.csv", row.names = FALSE)

saving my data into a specific file (csv file)

# Save the data frame to a specific folder  
write.csv(mydata, file = "C:/Users/PC/Desktop/3.1 N/statistical programming practicals/mydata.csv", row.names = FALSE)

saving my file into a specific location using path

# Load the dataset into the R environment and name it Loan  
Loan <- read.csv("C:/Users/PC/Desktop/3.1 N/statistical programming practicals/mydata.csv")

Loadind the csv file for use

# Display the first 5 rows of the Loan dataset  
head(Loan, 5)

## Age Mileage  
## 1 2.0 22  
## 2 2.5 34  
## 3 3.0 33  
## 4 4.0 37  
## 5 4.5 40

Limitted to only five rows to be displayed.

# Display the last 5 rows of the Loan dataset  
tail(Loan, 5)

## Age Mileage  
## 6 4.5 45  
## 7 5.0 49  
## 8 3.0 30  
## 9 6.0 58  
## 10 6.5 58

limited to only 5 last rows to be displayed.

# Generate summary statistics of the data variables  
summary(Loan)

## Age Mileage   
## Min. :2.000 Min. :22.00   
## 1st Qu.:3.000 1st Qu.:33.25   
## Median :4.250 Median :38.50   
## Mean :4.100 Mean :40.60   
## 3rd Qu.:4.875 3rd Qu.:48.00   
## Max. :6.500 Max. :58.00

# v. Output the dimension of the dataset  
dim(Loan)

## [1] 10 2

The dimension is displayed.

# Input the data into the R environment  
# Creating a data frame for the Student Performance Data  
student\_data <- data.frame(  
 Gender = c("Female", "Female", "Female", "Male", "Male", "Female"),  
 Race = c("Group B", "Group C", "Group B", "Group A", "Group C", "Group B"),  
 Lunch = c("Standard", "Standard", "Standard", "Free", "Standard", "Standard"),  
 Prep\_Course = c("None", "Completed", "None", "None", "None", "None"),  
 Statistics\_Score = c(72, 69, 90, 47, 76, 71),  
 Reading\_Score = c(72, 90, 95, 57, 78, 83),  
 Writing\_Score = c(74, 88, 93, 44, 75, 78)  
)

Creating a data frame for the Student Performance Data

#Output the data frame  
print(student\_data)

## Gender Race Lunch Prep\_Course Statistics\_Score Reading\_Score  
## 1 Female Group B Standard None 72 72  
## 2 Female Group C Standard Completed 69 90  
## 3 Female Group B Standard None 90 95  
## 4 Male Group A Free None 47 57  
## 5 Male Group C Standard None 76 78  
## 6 Female Group B Standard None 71 83  
## Writing\_Score  
## 1 74  
## 2 88  
## 3 93  
## 4 44  
## 5 75  
## 6 78

The table has 5 columns.

#Output the first 5 rows of the data  
head(student\_data, 5)

## Gender Race Lunch Prep\_Course Statistics\_Score Reading\_Score  
## 1 Female Group B Standard None 72 72  
## 2 Female Group C Standard Completed 69 90  
## 3 Female Group B Standard None 90 95  
## 4 Male Group A Free None 47 57  
## 5 Male Group C Standard None 76 78  
## Writing\_Score  
## 1 74  
## 2 88  
## 3 93  
## 4 44  
## 5 75

first 5 rows are displayed.

#Filter observations for female students  
female\_students <- subset(student\_data, Gender == "Female")  
print(female\_students)

## Gender Race Lunch Prep\_Course Statistics\_Score Reading\_Score  
## 1 Female Group B Standard None 72 72  
## 2 Female Group C Standard Completed 69 90  
## 3 Female Group B Standard None 90 95  
## 6 Female Group B Standard None 71 83  
## Writing\_Score  
## 1 74  
## 2 88  
## 3 93  
## 6 78

Only female students displayed.

#sort the observation using the statistic score in ascending order and reading score in descending order.   
sort<-student\_data[order(student\_data$Statistics\_Score,student\_data$Reading\_Score),]   
sort

## Gender Race Lunch Prep\_Course Statistics\_Score Reading\_Score  
## 4 Male Group A Free None 47 57  
## 2 Female Group C Standard Completed 69 90  
## 6 Female Group B Standard None 71 83  
## 1 Female Group B Standard None 72 72  
## 5 Male Group C Standard None 76 78  
## 3 Female Group B Standard None 90 95  
## Writing\_Score  
## 4 44  
## 2 88  
## 6 78  
## 1 74  
## 5 75  
## 3 93

Sorting the observations by Statistics Scores in ascending order

#Create a new column titled "Total\_Score" that calculates the total score for each student  
student\_data$Total\_Score <- student\_data$Statistics\_Score + student\_data$Reading\_Score + student\_data$Writing\_Score  
print(student\_data)

## Gender Race Lunch Prep\_Course Statistics\_Score Reading\_Score  
## 1 Female Group B Standard None 72 72  
## 2 Female Group C Standard Completed 69 90  
## 3 Female Group B Standard None 90 95  
## 4 Male Group A Free None 47 57  
## 5 Male Group C Standard None 76 78  
## 6 Female Group B Standard None 71 83  
## Writing\_Score Total\_Score  
## 1 74 218  
## 2 88 247  
## 3 93 278  
## 4 44 148  
## 5 75 229  
## 6 78 232

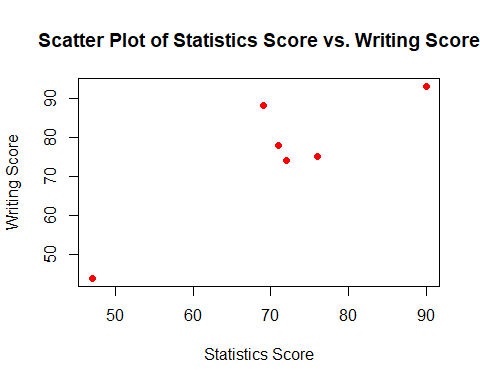
the column of a total score is displayed

#Perform correlation between Statistics Score and Writing Score  
correlation <- cor(student\_data$Statistics\_Score, student\_data$Writing\_Score)  
print(paste("Correlation between Statistics Score and Writing Score:", correlation))

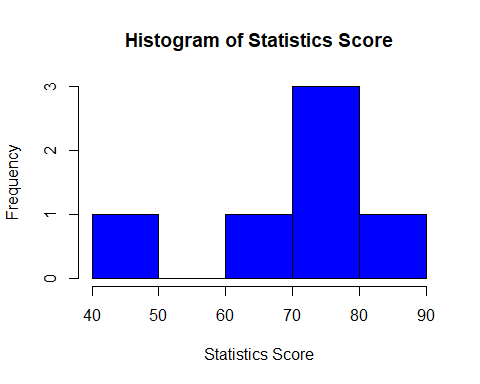
## [1] "Correlation between Statistics Score and Writing Score: 0.890605141184602"

There is an evidence of a high correlation between Statistics Score and Writing Score

#Plot a scatter plot of Statistics Score vs. Writing Score  
plot(student\_data$Statistics\_Score, student\_data$Writing\_Score,  
 main = "Scatter Plot of Statistics Score vs. Writing Score",  
 xlab = "Statistics Score",  
 ylab = "Writing Score",  
 pch = 19, col = "red")

 Visualisation of a scatter plot of Statistics Score vs. Writing Score.There is a higher writing score above a statistic score of 70.

#Plot a histogram of the Statistics Score  
hist(student\_data$Statistics\_Score,  
 main = "Histogram of Statistics Score",  
 xlab = "Statistics Score",  
 col = "blue",  
 border = "black")



#subtopic .dealing with matrices   
Mat1<-matrix(c(5,0,6,1,3,5,9,5,7,1,5,3),3,4)   
Mat1

## [,1] [,2] [,3] [,4]  
## [1,] 5 1 9 1  
## [2,] 0 3 5 5  
## [3,] 6 5 7 3

Visualisation of statistics score in form of a histogram, whith the highest score been between 70-80 mark.

library(MASS)   
Mat1\_pseudoinv<- ginv(Mat1)#solving the inverse Mat1\_pseudoinv   
Mat1\_pseudoinv

## [,1] [,2] [,3]  
## [1,] -0.01788491 -0.141135303 0.142690513  
## [2,] -0.14580093 -0.002008813 0.166861068  
## [3,] 0.14346812 0.063310005 -0.097524624  
## [4,] -0.05598756 0.137895283 -0.002592017

Finding the inverse of a matrix.

#marices   
mat2<-matrix(c(3,3,2,1,1,4,1,1,0,0,1,5,1,2,2),ncol = 3)   
mat2

## [,1] [,2] [,3]  
## [1,] 3 4 1  
## [2,] 3 1 5  
## [3,] 2 1 1  
## [4,] 1 0 2  
## [5,] 1 0 2

Forming a matrix.

#creating rowand column labels   
colnames(mat2)<-c("gold","silver","bronze")  
rownames(mat2)<-c("United States","Great Britain","Canada","Russia","Switzerland")  
mat2

## gold silver bronze  
## United States 3 4 1  
## Great Britain 3 1 5  
## Canada 2 1 1  
## Russia 1 0 2  
## Switzerland 1 0 2

Namming columns and rows.

#dimension of the matrix  
dim(mat2)

## [1] 5 3

Finding the dimensuon of a matrix

#using sequences to create matrices  
b<-1:9   
b

## [1] 1 2 3 4 5 6 7 8 9

The Cbind matrix.

cbind(b)

## b  
## [1,] 1  
## [2,] 2  
## [3,] 3  
## [4,] 4  
## [5,] 5  
## [6,] 6  
## [7,] 7  
## [8,] 8  
## [9,] 9

The Rbind matrix.

rbind(b)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]  
## b 1 2 3 4 5 6 7 8 9