

**CSE 240 Data Science with R**

**STUDENT WORK BOOK**

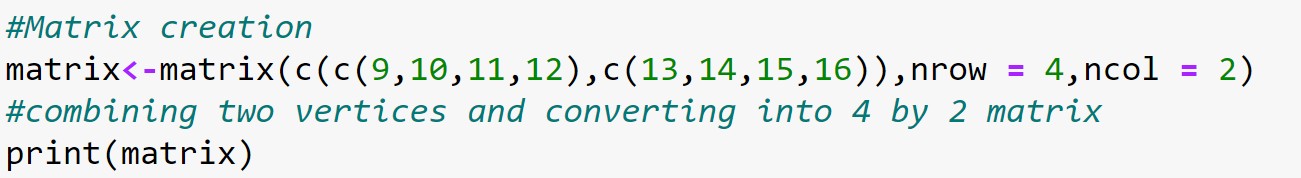
|  |  |  |
| --- | --- | --- |
| **Name** | **:** | M.Sathishkumar |
| **Unique ID** | **:** | E0119052 |
| **Year** | **:** | II |
| **Quarter** | **:** | Q6 |
| **Department** | **:** | B.Tech CSE (AI &ML) |
| **Faculty Name** | **:** | Prof.B.Nirmala |
| **Academic Year** | **:** | 2020-2021 |

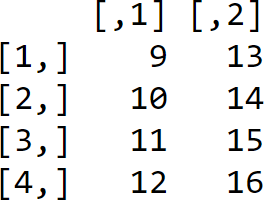
**Date:02/11/2020**

**Q.NO 1: Question**

Consider 2 vectors c(9,10,11,12) and c(13,14,15,16).

Create a 4 by 2 matrix from these two vectors

**Program:**

 **Output:**

**Explanation:**

* Matrix is a two-dimensional data structure in R programming.
* Matrix can be created using the matrix () function
* Dimension of the matrix can be defined by passing appropriate value guments nrow and ncol

**Date: 02/11/2020**

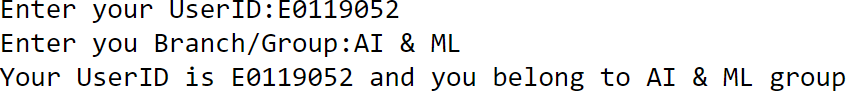
**Q.NO 2: Question**

**Program:**

Write an R program to take input from the user (userID and Group/Branch) and display the values



**Output:**



**Explanation:**

* readline reads a line from the terminal (in interactive use).
* readline() lets the user enter a one-line string at the terminal.
* The prompt argument is printed in front of the user input. It usually ends on ": ".

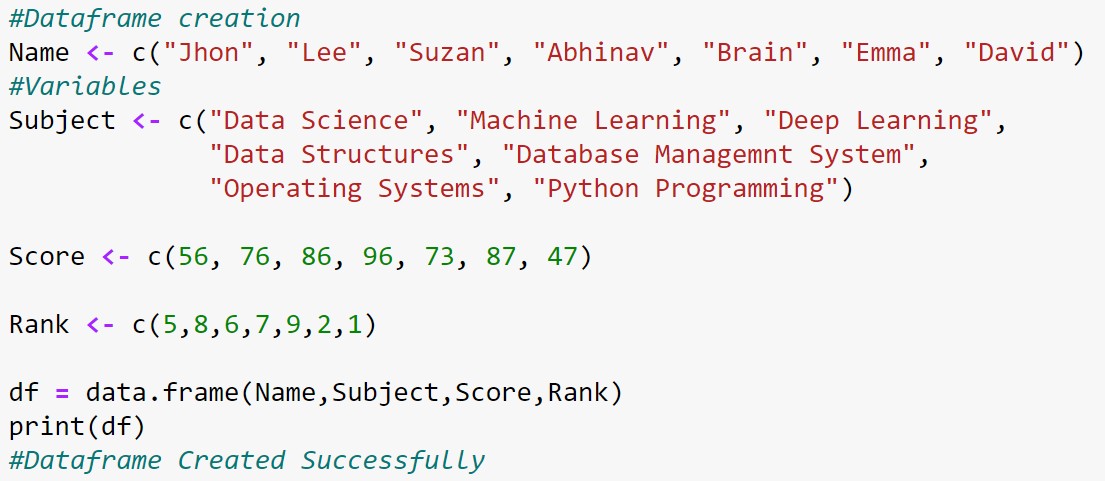
**Date: 02/11/2020**

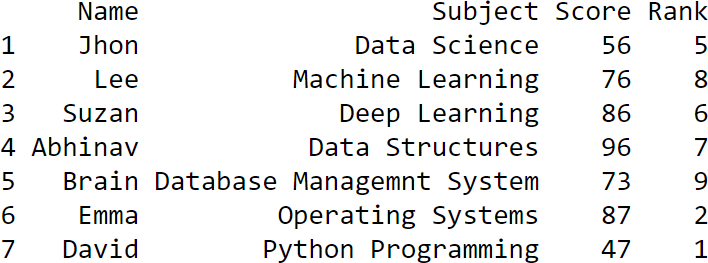
**Q.NO 3: Question**

Create a data frame Write a R program to create a data frame from four given vectors.

a name b. Subject C. Score d. Rank

**Program:**



**Output:**

**Explanation:**

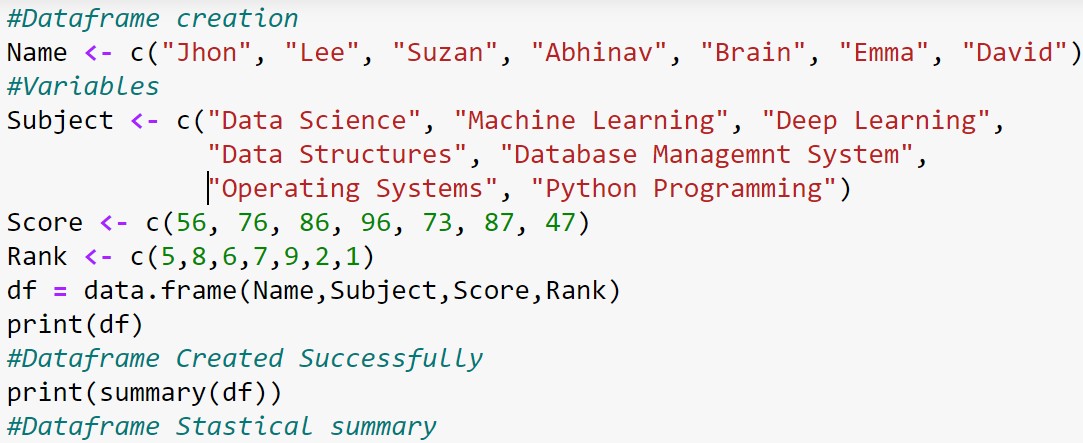
* A **data frame** is used for storing data tables. It is a list of vectors of equal length.
* The top line of the table, called the **header**, contains the column names.
* Each horizontal line afterward denotes a **data row**, which begins with header, and then followed by the actual data.
* Each data member of a row is called a **cell**.

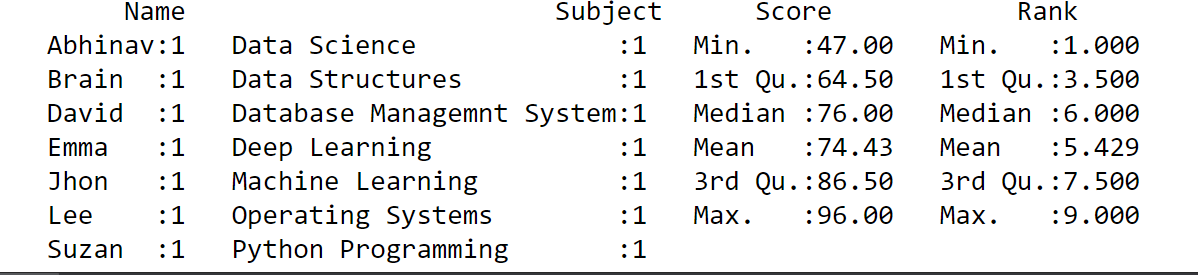
**Date: 02/11/2020**

**Q.NO 4: Question**

Write a R program to get the statistical summary and nature of the data of a given data frame. ( use 3rd Question dataframe

**Program:**



**Output:**

**Explanation:**

Summary is a generic function used to produce result summaries of the results of various model fitting functions.

The function invokes particular methods which depend on the class of the first argument

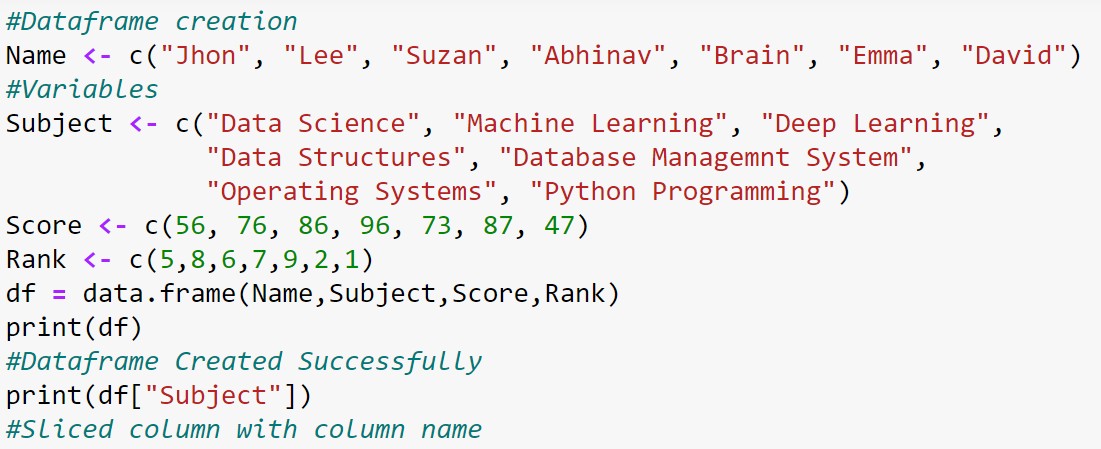
The form of the value returned by summary depends on the class of its argument

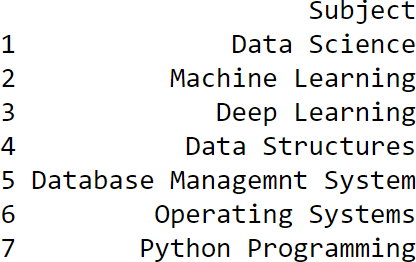
**Date: 02/11/2020**

**Q.NO 5: Question**

Write a R program to extract specific column from a data frame using column name

**Program:**



**Output:**

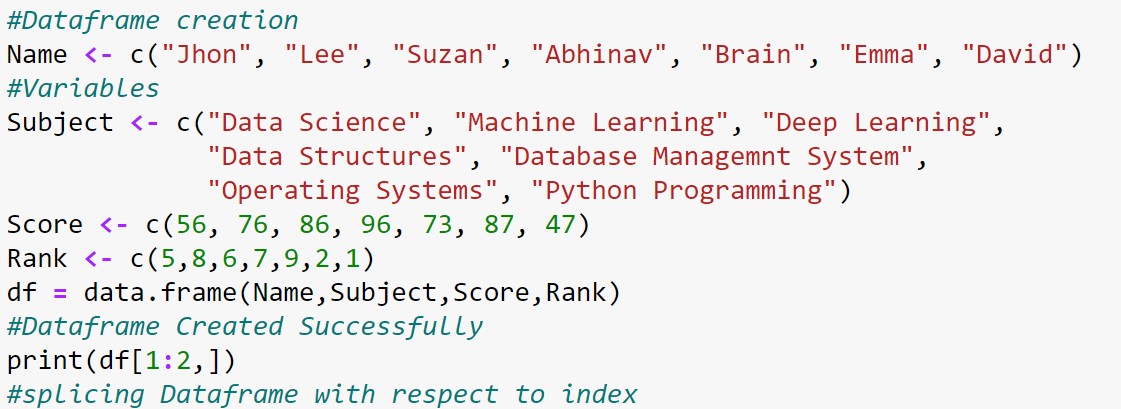
**Explanation:**

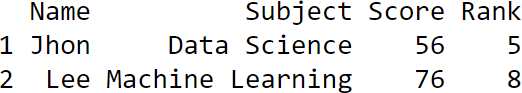
* We retrieve a data frame column slice with the single square bracket "[]" operator.
* Here We can retrieve the same column slice by its name.
* And we can pack the row names in an index vector in order to retrieve multiple rows

**Date: 02/11/2020**

**Q.NO 6: Question**

.Write a R program to extract first two rows from a given data frame

**Program:**

**Output:**

**Explanation:**

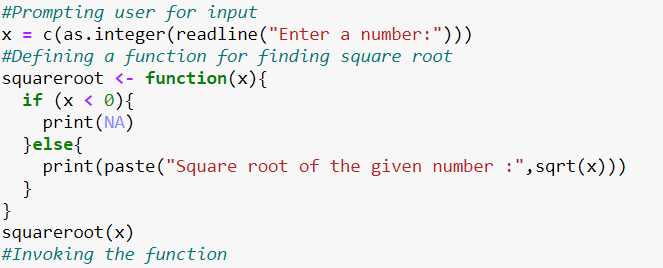
* We retrieve rows from a data frame with the single square bracket operator, just like what we did with columns.
* However, in additional to an index vector of row positions, we append an extra comma character.
* This is important, as the extra comma signals a wildcard match for the second coordinate for column positions.

**Date: 03/11/2020**

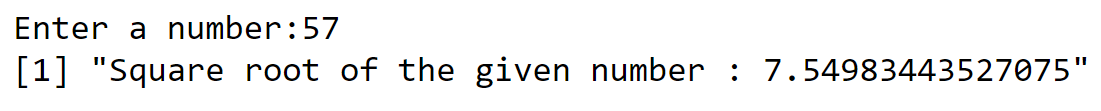
**Q.NO 7: Question**

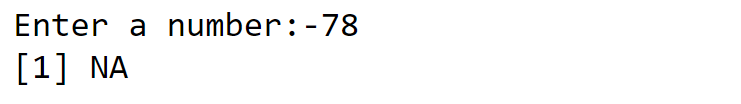
Create an R script that calculates the square root of a given integer vector x of length one, if the value contained in x is negative it should return NA.

**Program:**

****

**Output:**

****

****

**Explanation:**

* Missing values are represented by the symbol **NA** (not available)
* If the numeric\_Expression is a positive value, the sqrt function returns the square root of a given value.
* If the numeric\_Expression is a negative value, the sqrt function return ***NaN***.
* Numeric\_Expression is not a number (**NaN**), or Negative Infinity, then sqrt in R returns ***NaN***.

**Date: 03/11/2020**

**Q.NO 8: Question**

Demonstrate and examine the output of letter and LETTER

**Program:**

****

**Output:**



**Explanation:**

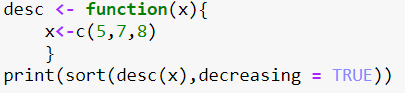
* This is useful because you can always view your data: just print it.
* You needn’t write special printing logic, even for complicated data structures.
* The print function has a significant limitation, it prints only one object at a time.
* Trying to print multiple items gives this mind-numbing error message

**Date: 03/11/2020**

**Q.NO 9: Question**

Create an R script that, given a numeric vector x with length 3, will print the elements by order from high to low.

**Program:**

****

**Output:**

****

**Explanation:**

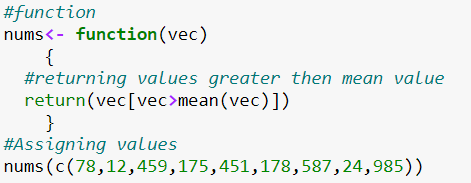
* Sort (or order) a vector or factor (partially) into ascending or descending order. For ordering along more than one variable, e.g., for sorting data frame
* To sort a data frame in R, use the **sort( )**function. By default, sorting is ASCENDING.
* Prepend the sorting variable by a minus sign to indicate DESCENDING order

**Date: 03/11/2020**

**Q.NO 10: Question**

Create an R script that returns the amount of values that are larger than the mean of a vector. You are allowed to use mean(). ( Use function)

**Program:**

****

**Output:**

****

**Explanation:**

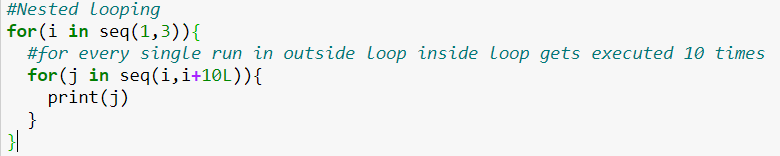
* It is calculated by taking the sum of the values and dividing with the number of values in a data series.
* The function **mean()** is used to calculate and decision operator “>” is used to compare the giving condition and take decision

**Date: 03/11/2020**

**Q.NO 11: Question**

Write a double for loop which prints 30 numbers (1:10, 2:11, 3:12). Those are three clusters of ten numbers each. The first loop determines the number of clusters (3) via its length; the second loop the numbers to be printed (1 to 10 at the beginning). Each cluster starts one number higher than the previous one

**Program:**

****

**Output:**

****

**Explanation:**

* In Nested for Loop, it makes use of the control structures to manage the execution of the expression, one such control structure is Nested for Loop a similar to basic ‘for’ loop executes.
* It can be defined as placing one ‘for’ loop inside the first ‘for’ loop is called as nesting or loop of loops in some terms, which takes the responsibility of two loops such that the outer loop controls the number of repetition of the whole inner detailed information until it is false,
* in other words, the inner loop executes n-times of every execution of the outer for loop and also, it’s a great tool to work with R Programming Language.

**Date: 03/11/2020**

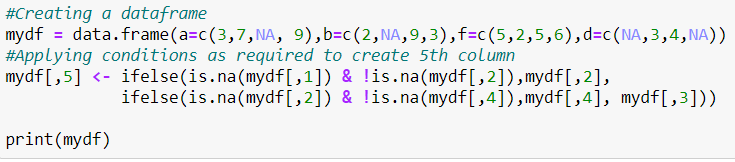
**Q.NO 12: Question**

a. You have the data.frame ‘mydf’ with four columns like below

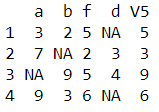
a = c(3,7,NA, 9) b = c(2,NA,9,3) f = c(5,2,5,6) d = c(NA,3,4,NA)

You want to add another column ‘5’: the 5th column contains the value of col 2 if col 1 is NA; the 5th column contains the value of col 4 if col 2 is NA; the 5th column contains the value of col 3 in all other cases.

**Program:**

****

**Output:**

****

**Explanation:**

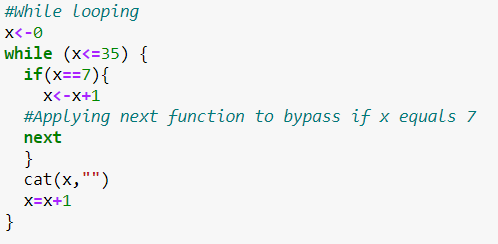
* An if statement can be followed by an optional else statement which executes when the Boolean expression is false.
* An if statement can be followed by an optional else if...else statement, which is very useful to test various conditions using single if...else if statement

**Date: 03/11/2020**

**Q.NO 13: Question**

Write a while loop starting with x = 0. The loop prints all numbers up to 35 but it skips number 7. Condition: If x== 7 next

**Program:**

****

**Output:**

****

**Explanation:**

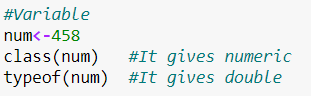
* The **next** statement in is useful when we want to skip the current iteration of a loop without terminating it.
* On encountering next, the R parser skips further evaluation and starts next iteration of the loop.in other words, the inner loop executes n-times of every execution of the outer for loop.

**Date: 03/11/2020**

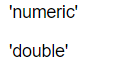
**Q.NO 14: Question**

Examine the difference between typeof and class () method using R program

**Program:**

****

**Output:**

****

**Explanation:**

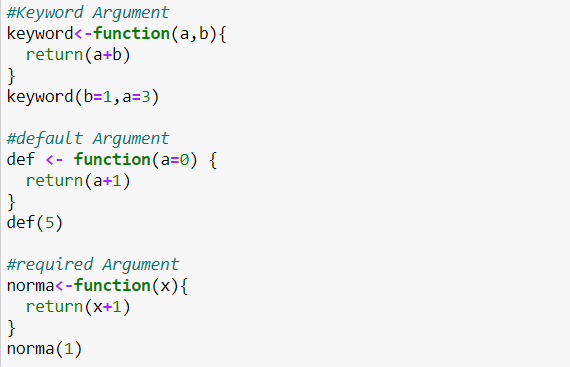
* 'class' is a property assigned to an object that determines how generic functions operate with it. It is not a mutually exclusive classification. If an object has no specific class assigned to it, such as a simple numeric vector, its class is usually the same as its mode, by convention.
* Typeof determines the (R internal) type or storage mode of any object

**Date: 03/11/2020**

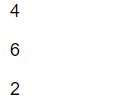
**Q.NO 15: Question**

Create a function and demonstrate their features like required, keyword, default.

**Program:**

****

**Output:**

****

**Explanation:**

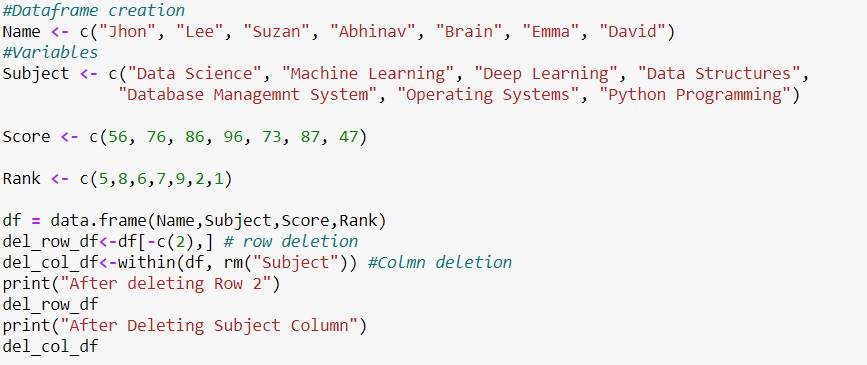
* Arguments are always named when you define a function. When you call a function, you do not have to specify the name of the argument.
* Arguments are optional; you do not have to specify a value for them. They can have a default value, which is used if you do not specify a value for that argument yourself.
* You can use as many arguments as you like, there is no limit to the number of arguments. An argument list comprises of comma-separated values that contain the various formal arguments.

**Date: 03/11/2020**

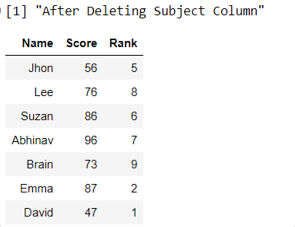
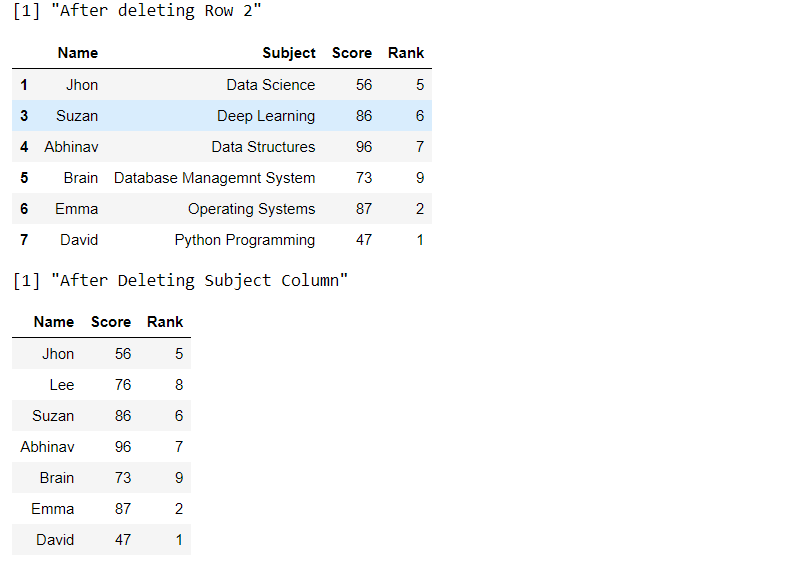
**Q.NO 16: Question**

Create a dataframe and delete the row and column. ( Use the own data values to create frame)

**Program:**

****

**Output:**

****

**Explanation:**

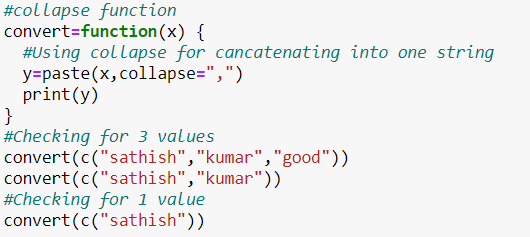
* -c(rowNum) is used to return a slice of a dataframe without the mentioned row i.e. row deletion
* The within function returns a subset of a dataframe with the second argument ad a function to apply to all the column rm function removes the “Subject” column out of the data frameTypeof determines the (R internal) type or storage mode of any object.

**Date: 04/11/2020**

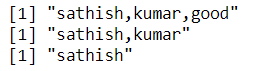
**Q.NO 17: Question**

Write a function that turns (e.g.) a vector c("a", "b", "c") into the string "a, b, and c". Think carefully about what it should do if given a vector of length 0, 1, or 2.

**Program:**

****

**Output:**

****

**Explanation:**

* **'collapse'** is a property Collapses a character vector of any length into a length 1 vector.
* Syntax - collapse(x, sep = "", width = Inf, last = "")
* **'paste'** Concatenate vectors after converting to character.

**Date: 04/11/2020**

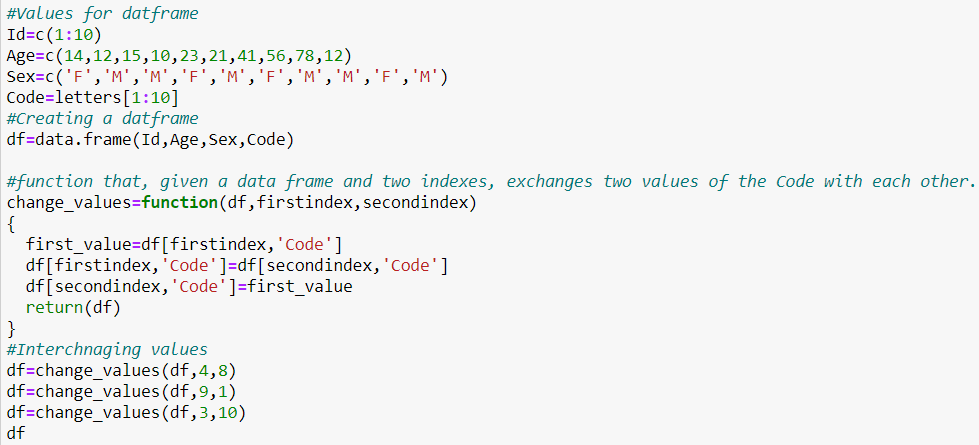
**Q.NO 18: Question**

Consider a data frame

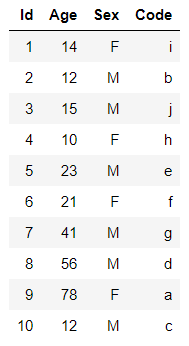
Create a function that, given a data frame and two indexes, exchanges two values ​​of the Code variable with each other.

For example, if the index is 1 and 3, you assign: df[1,'Code']=df[3,'Code'] df[3,'Code']=df[1,'Code']

**Program:**

****

**Output:**

****

**Explanation:**

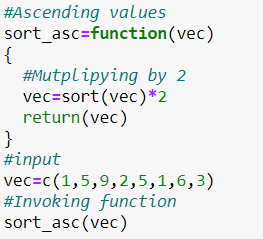
* A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

**Date: 03/11/2020**

**Q.NO 19: Question**

Create a function that given a numeric vector, sort this in ascending order and duplicate it by two.

**Program:**

****

**Output:**

****

**Explanation:**

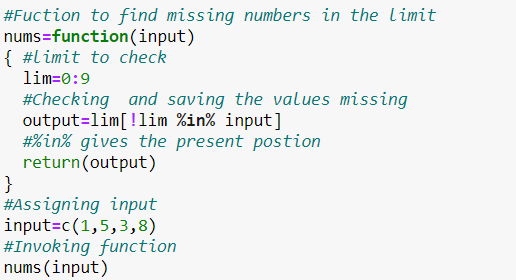
* Sort (or order) a vector or factor (partially) into ascending or descending order. For ordering along more than one variable
* By default, sorting is ASCENDING. Prepend the sorting variable by a minus sign to indicate DESCENDING order

**Date: 04/11/2020**

**Q.NO 20: Question**

Create a function that given a numeric vector X returns the digits 0 to 9 that are not in X. If X=0 2 4 8 the function return 1 3 5 6 7 9

**Program:**

****

**Output:**

****

**Explanation:**

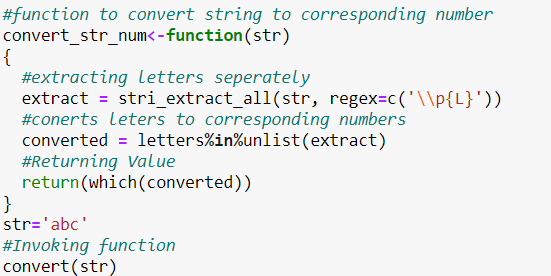
* %in% operator is used to identify if an element belongs to a vector or Dataframe.
* We can
  + select column of a dataframe in R using %in% operator.
  + create new variable of a column using %in% operator
  + drop column of a dataframe in R using %in% operator.

**Date: 04/11/2020**

**Q.NO 21: Question**

Create a function that given one word, return the position of word’s letters on letters vector. For example, if the word is ‘abc’, the function will return 1 2 3.

**Program:**

****

**Output:**

****

**Explanation:**

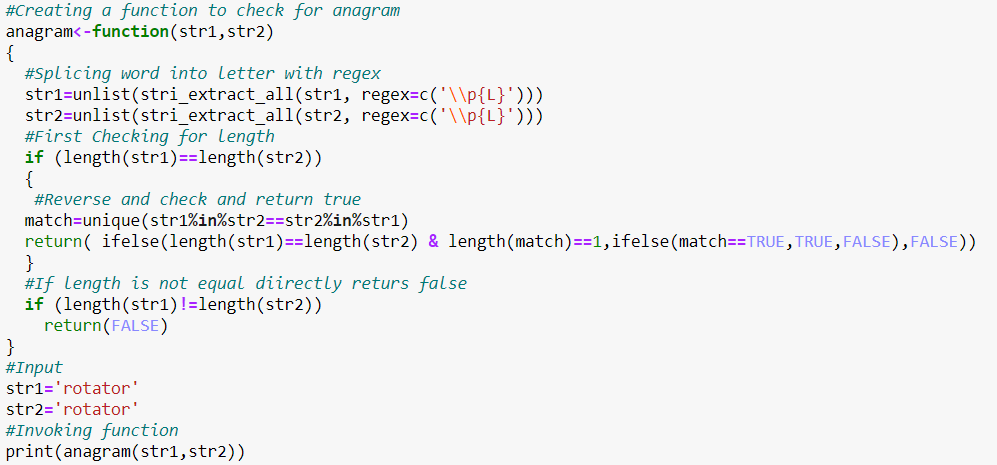
* Regular expressions are the default pattern engine in stringr. That means when you use a pattern matching function with a bare string, it’s equivalent to wrapping it in a call to REGEX()
* Regular expressions are a concise and flexible tool for describing patterns in strings.

**Date: 04/11/2020**

**Q.NO 22: Question**

Write a code to check the given string is anagram or not

**Program:**

****

**Output:**

****

**Explanation:**

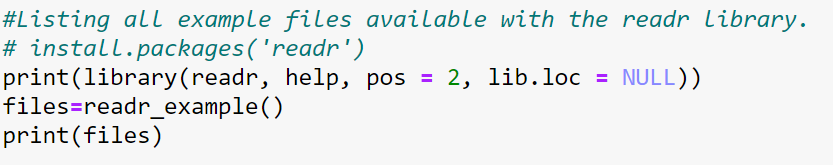
* Regular expressions are the default pattern engine in stringr. That means when you use a pattern matching function with a bare string, it’s equivalent to wrapping it in a call to REGEX()
* Regular expressions are a concise and flexible tool for describing patterns in strings.
* Basically, we’re splitting the word up into letters. Then using the unlist function to convert this into a vector. Then sorting the vector into alphabetical order.

**Date: 05/11/2020**

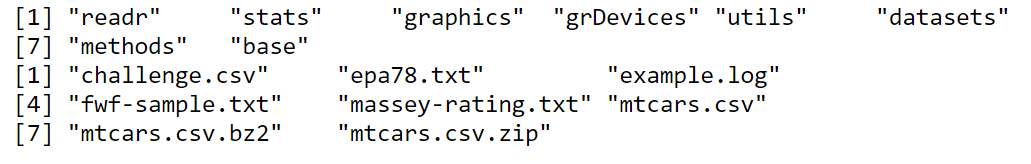
**Q.NO 23: Question**

List all example files available with the readr library.

**Program:**

****

**Output:**

****

**Explanation:**

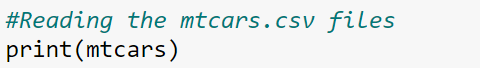
* The goal of readr is to provide a fast and friendly way to read rectangular data (like csv, tsv, and fwf).
* It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes.
* If you are new to readr, the best place to start is the data import chapter in R for data science.

**Date: 05/11/2020**

**Q.NO 24: Question**

Read the mtcars.csv file.

**Program:**

****

**Output:**

****

**Explanation:**

* We used the **mtcars** data set that is built-in to the **R** distribution .
* **mtcars** data comes from the 1974 Motor Trend magazine.
* The data includes fuel consumption data, and ten aspects of car design for then-current car models

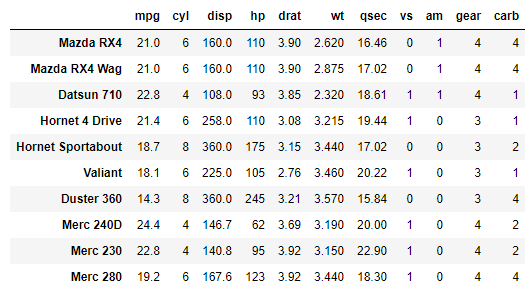
**Date: 05/11/2020**

**Q.NO 25: Question**

Read the first 10 lines from the mtcars.csv file.

**Program:**

****

**Output: **

**Explanation:**

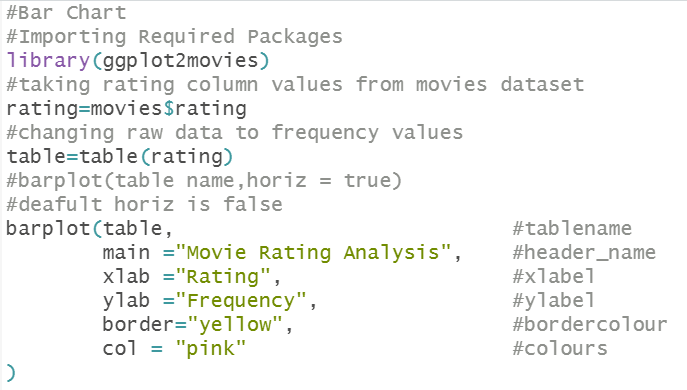
* We used the **mtcars** data set that is built-in to the **R** distribution .
* **mtcars** data comes from the 1974 Motor Trend magazine.
* The data includes fuel consumption data, and ten aspects of car design for then-current car models.
* Head function with specified parameter gives top rows of specified datax

**Date: 06/11/2020**

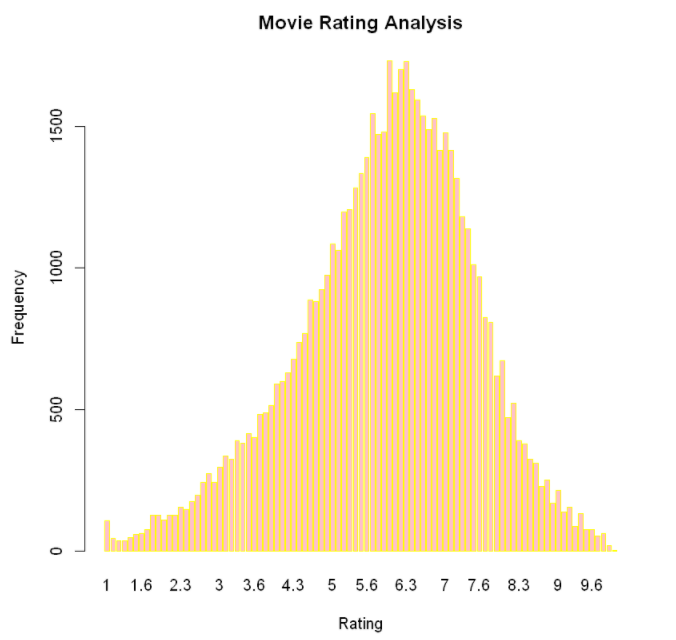
**Q.NO 26: Question**

**Program:**

Bar plot – Movies Dataset

****

**Output:**

****

**Explanation:**

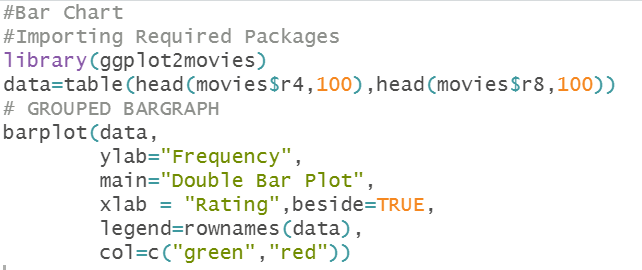
* A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function **barplot()** to create bar charts.

**Date: 06/11/2020**

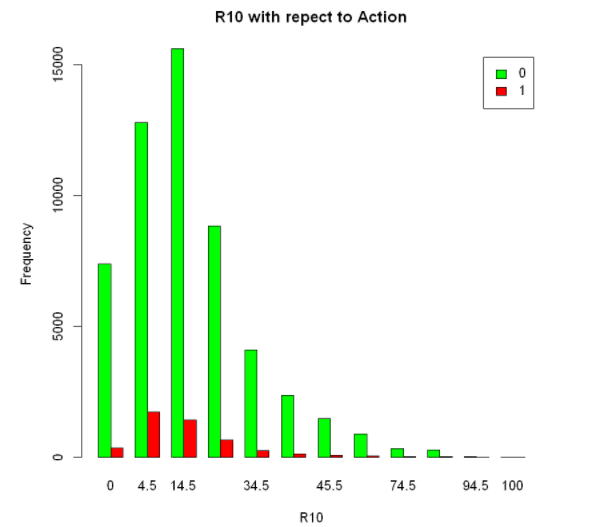
**Q.NO 27: Question**

**Program:**

Group BarGraph – Movies Dataset

****

**Output:**

****

**Explanation:**

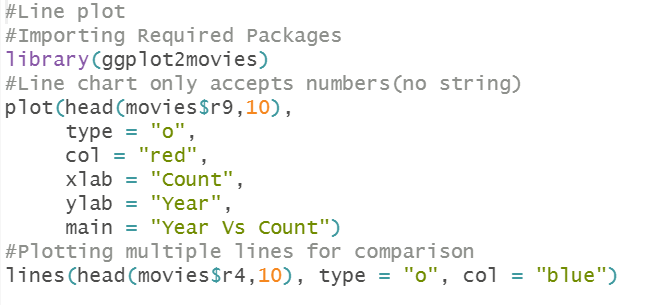
* A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function **barplot()** to create bar charts.

**Date: 06/11/2020**

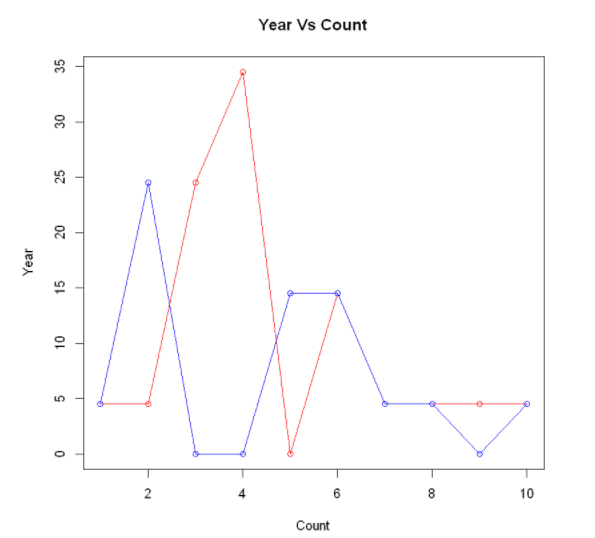
**Q.NO 28: Question**

**Program:**

Line Chart – Movies Dataset

****

**Output:**

****

**Explanation**

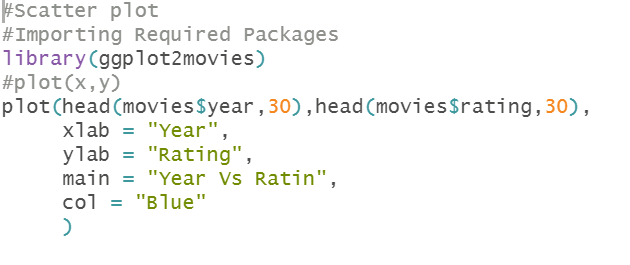
* A line chart is a graph that connects a series of points by drawing line segments between them. These points are ordered in one of their coordinate (usually the x-coordinate) ".

**Date: 06/11/2020**

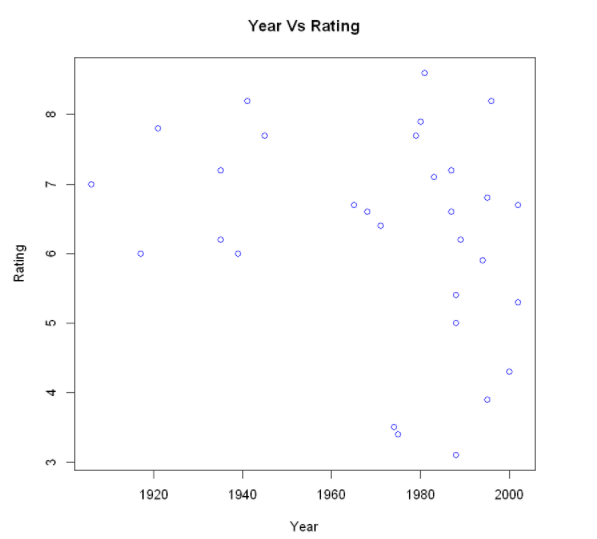
**Q.NO 29: Question**

**Program:**

Scatter plot – Movies Dataset

****

**Output:**

****

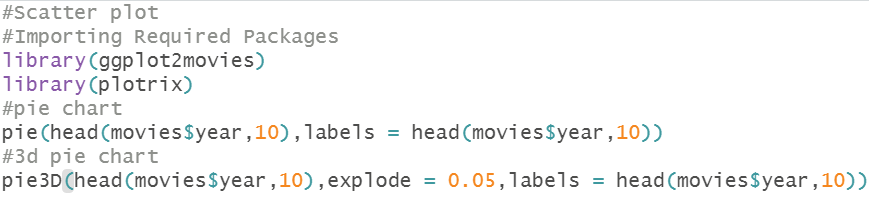
**Explanation:**

* Scatterplots show many points plotted in the Cartesian plane. Each point represents the values of two variables. One variable is chosen in the horizontal axis and another in the vertical axis.

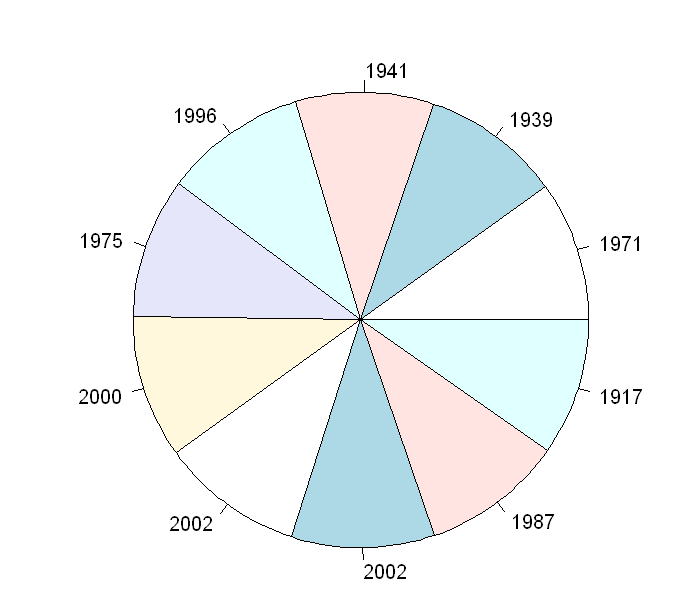
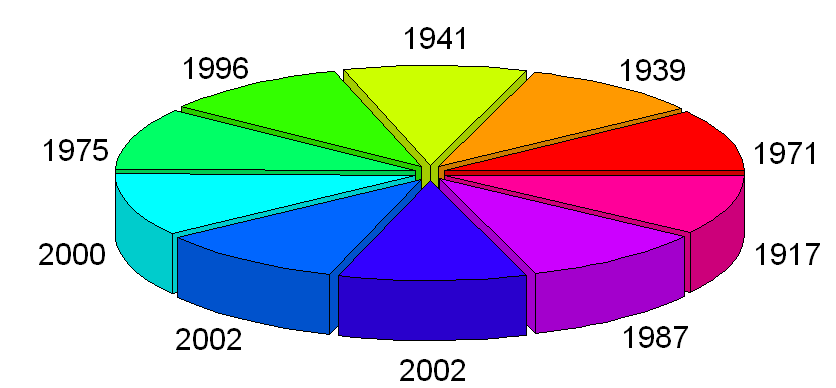
**Date: 06/11/2020**

**Q.NO 29: Question**

**Program:**

Pie Chart – Movies Dataset 

**Output:**

**** ****

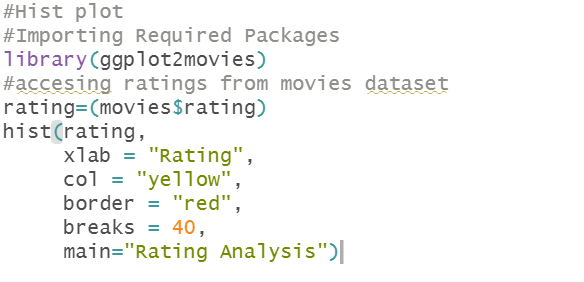
**Explanation:**

* pie-chart is a representation of values as slices of a circle with different colors. The slices are labeled and the numbers corresponding to each slice is also represented in the chart.

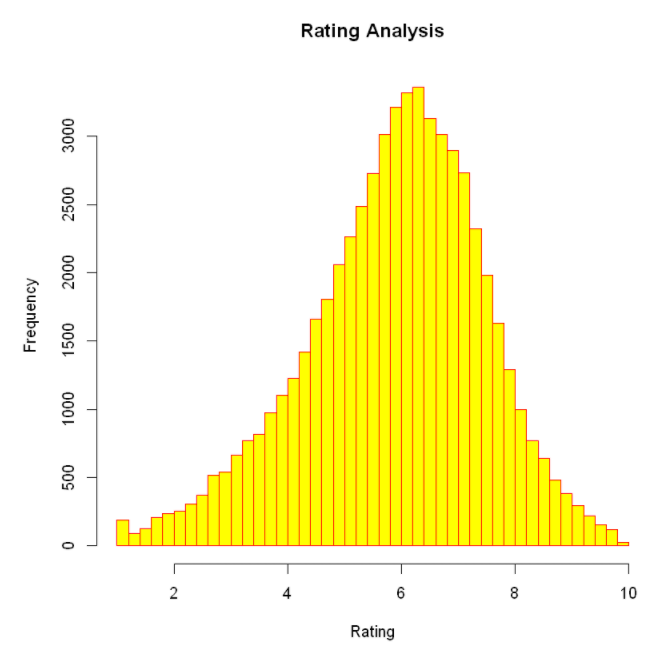
**Date: 06/11/2020**

**Q.NO 30: Question**

**Program:**

Hist plot – Movies Dataset ****

**Output:**

****

**Explanation:**

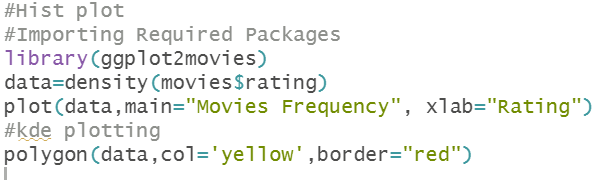
* A histogram represents the frequencies of values of a variable bucketed into ranges. Histogram is similar to bar chat but the difference is it groups the values into continuous ranges.

**Date: 06/11/2020**

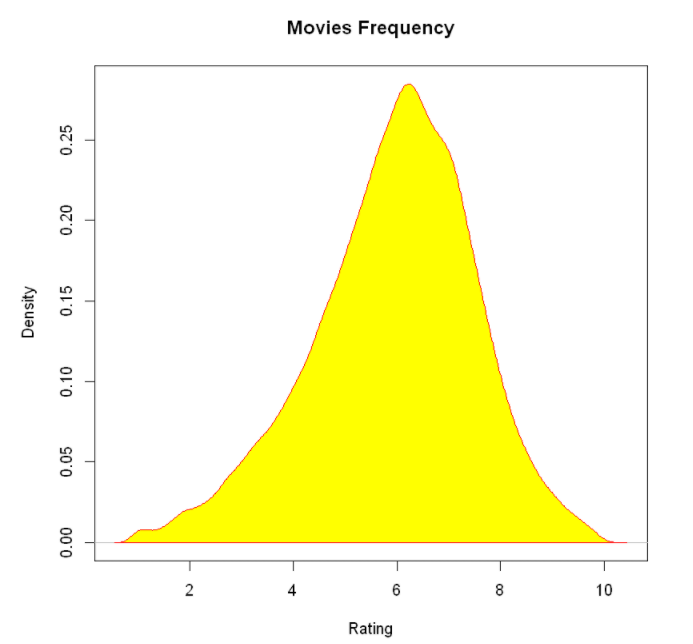
**Q.NO 31: Question**

**Program:**

difference plot – Movies dataset

****

**Output:**

****

**Explanation:**

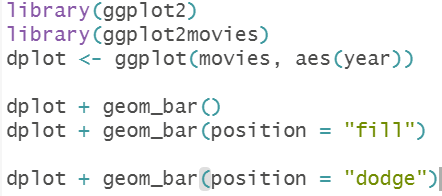
* polygon draws the polygons whose vertices are given in x and y.

**Date: 06/11/2020**

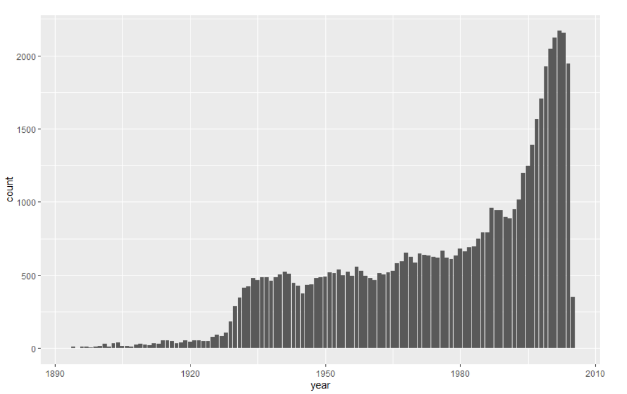
**Q.NO 32: Question**

**Program:**

Demonstrate the ggplot2 layer

****

**Output:**



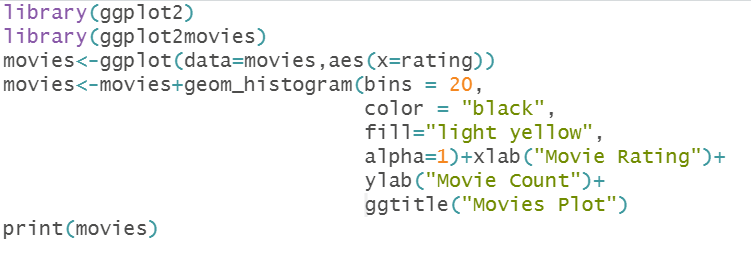
**Explanation:**

* ggplot() initializes a ggplot object. It can be used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.

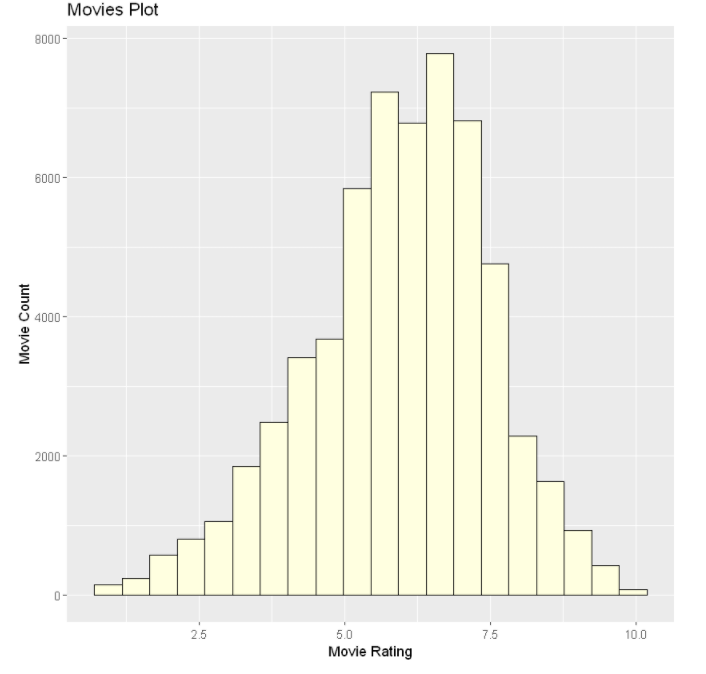
**Date: 06/11/2020**

**Q.NO 33: Question**

**Program:**

Ggplot – histogram plotting****

**Output:**

****

**Explanation:**

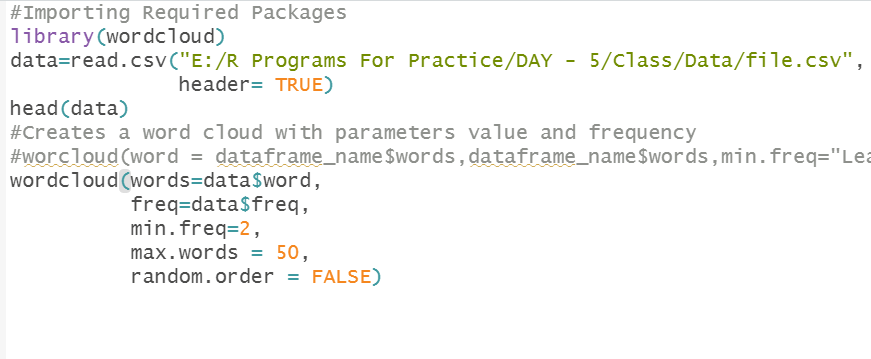
* ggplot() initializes a ggplot object. It can be used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.

**Date: 06/11/2020**

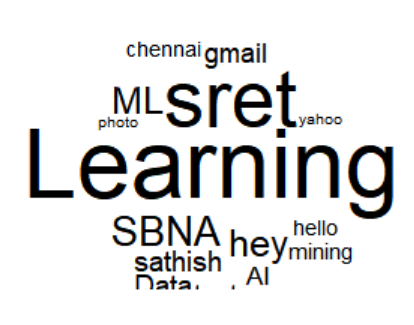
**Q.NO 34: Question**

**Program:**

Word Cloud

****

**Output:**

****

**Explanation:**

* wordcount() counts words. Currently a "word" is a clustering of characters separated from another clustering of charactersby at least 1 space. That is the law.

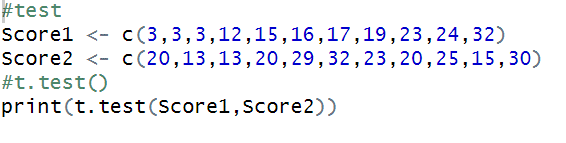
**Date: 07/11/2020**

**Q.NO 35: Question**

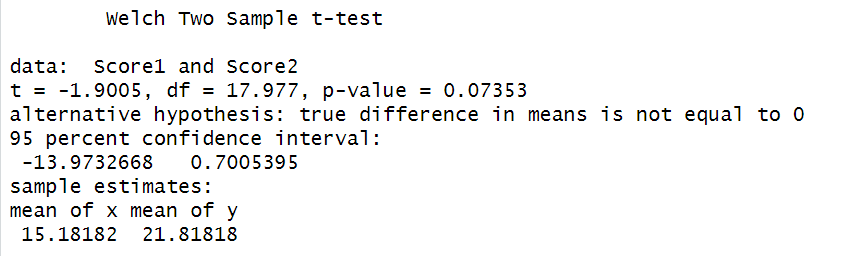
Perform the following operation:

1. T-Test (score 1 and score 2)

**Program:**

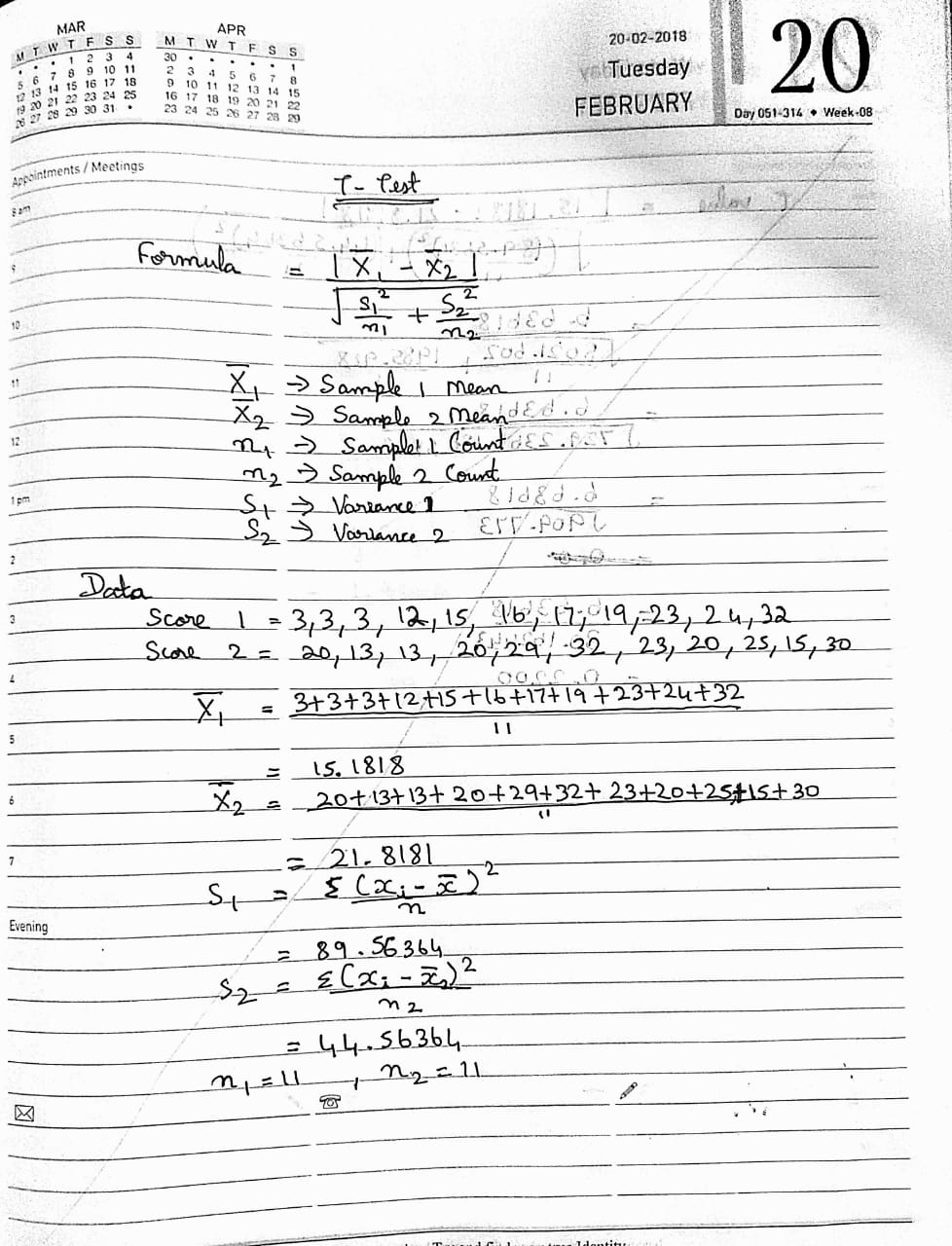


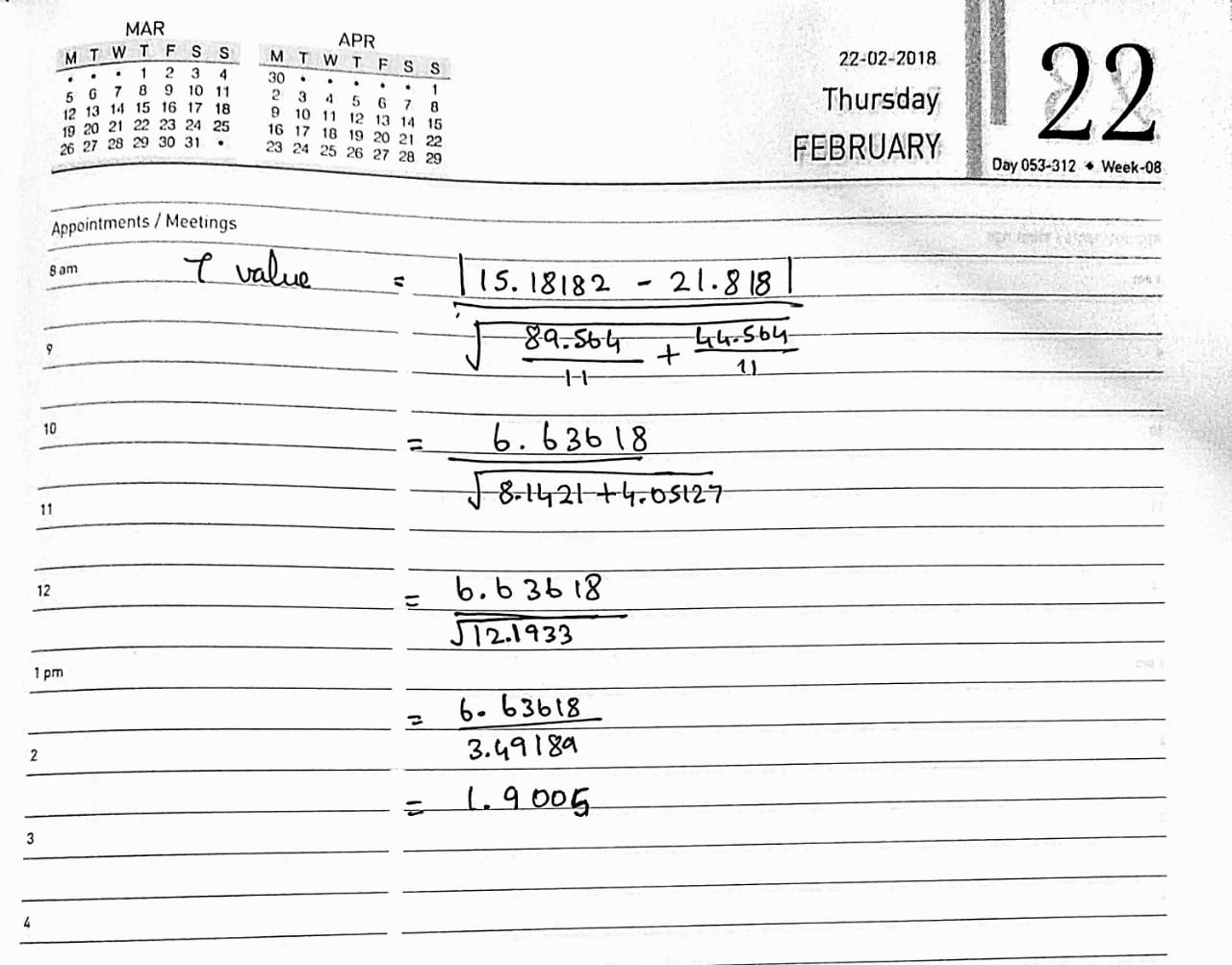
**Output:**

****

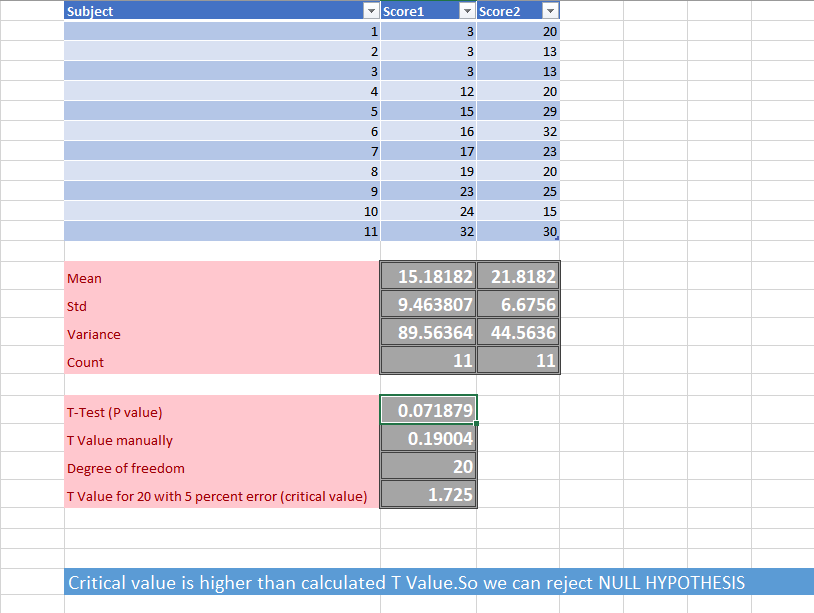
**Explanation:**

* A t-test allows us to compare the average values of the two data sets and determine if they came from the same population.
* In the above examples, if we were to take a sample of students from class A and another sample of students from class B, we would not expect them to have exactly the same mean and standard deviation.

Manual calculation for the T-value



Excel sheet R- Program

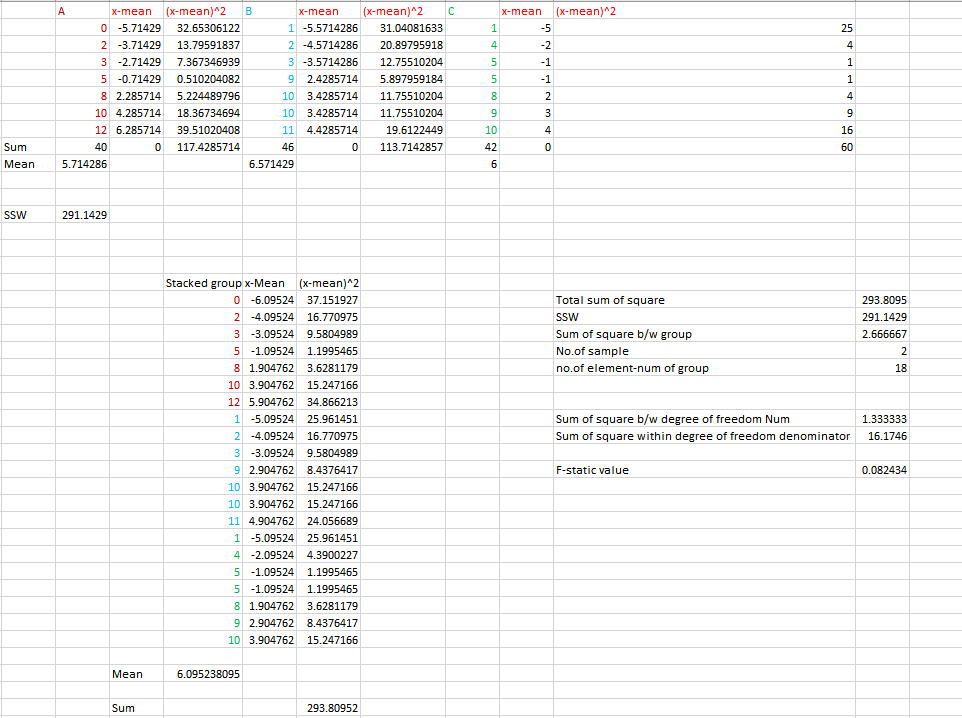
****

**Date: 07/11/2020**

**Q.NO 35: Question**

**Program:**

ANOVA Excel sheet

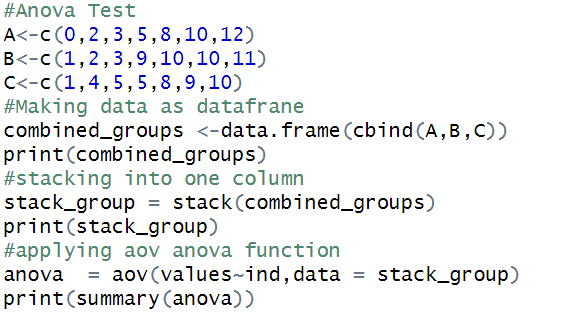
****

**Explanation:**

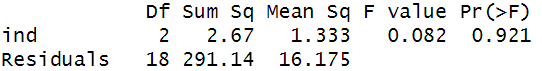
**Date: 06/11/2020**

R Program

**Code:**



**Output:**

****

**Explanation:**

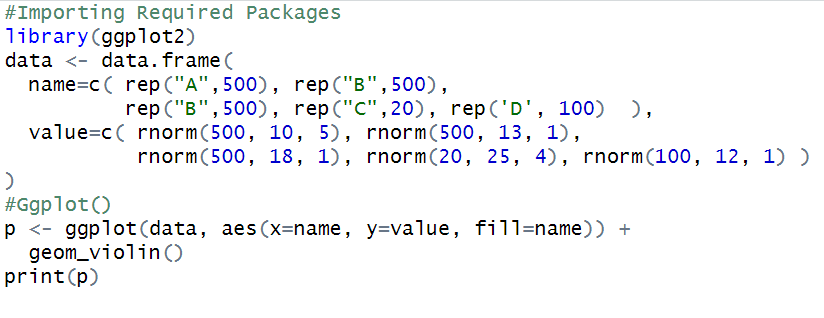
* An **ANOVA test** is a way to find out if survey or experiment results are significant.
* In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis.
* Basically, you're **testing** groups to see if there's a difference between them.

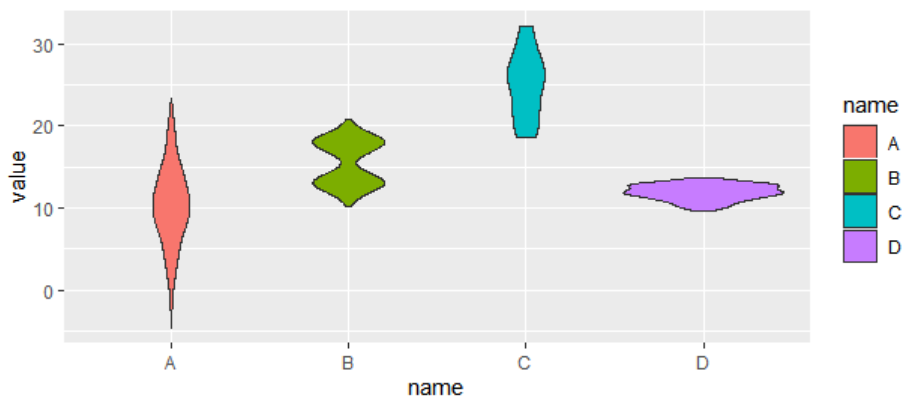
**Date: 06/11/2020**

**Q.NO 36: Question**

**Program:**

Draw the Violin plot



**Output:**

**Explanation:**

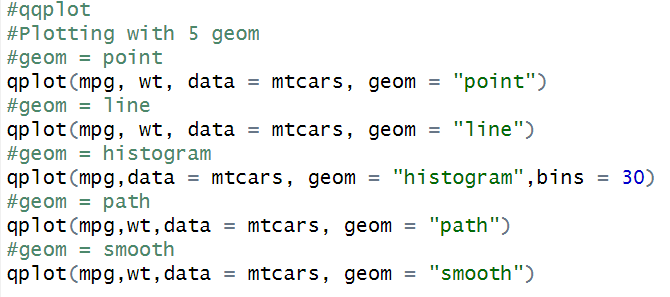
* A **violin plot** is a method of **plotting** numeric data. It is similar to a box **plot**, with the addition of a rotated kernel density **plot** on each side.
* While a box **plot** only shows summary statistics such as mean/median and interquartile ranges, the **violin plot** shows the full distribution of the data.

**Date: 07/11/2020**

**Q.NO 37: Question**

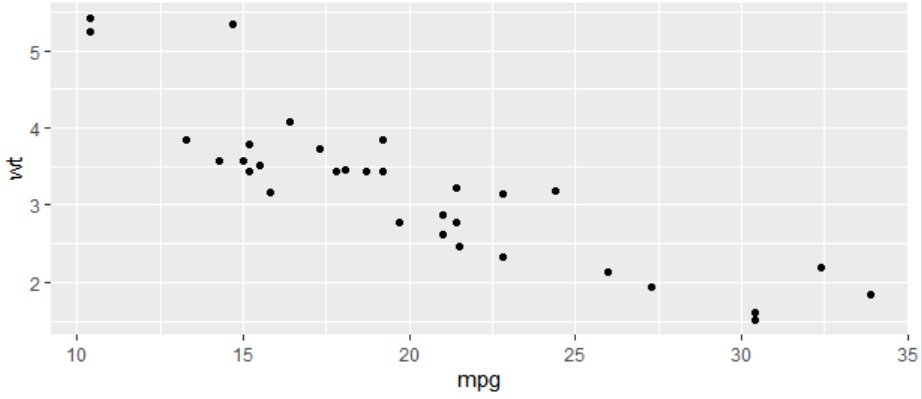
**Program:**

Draw the Qplot for 5 Types

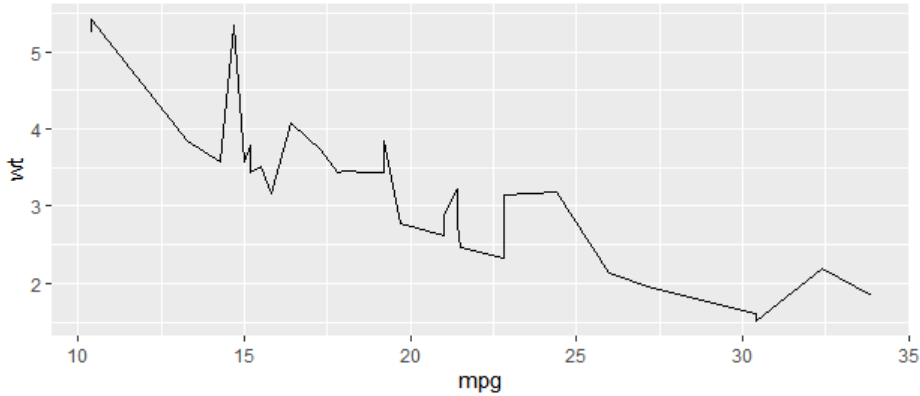
****

**Output:**

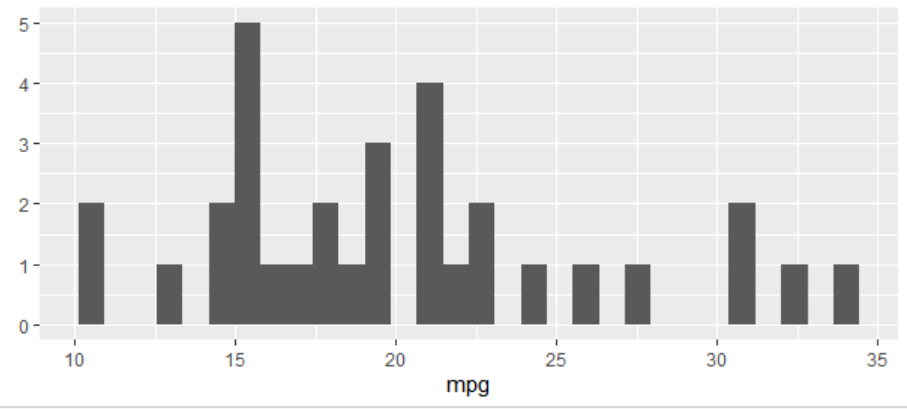
**Point plot**

****

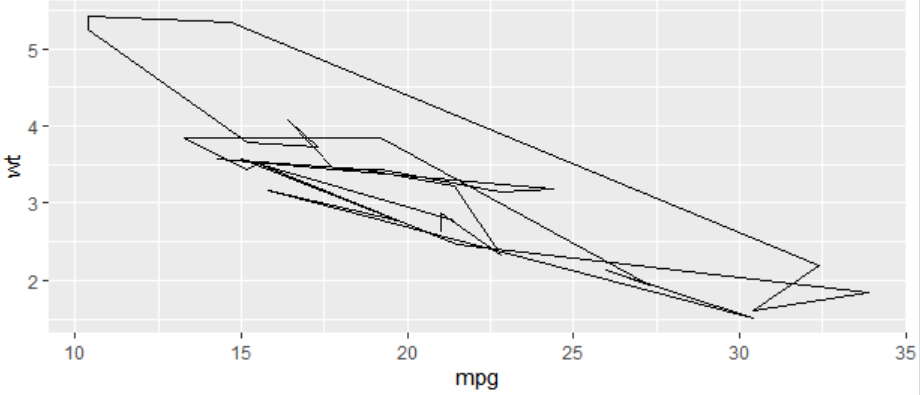
**Line plot**

****

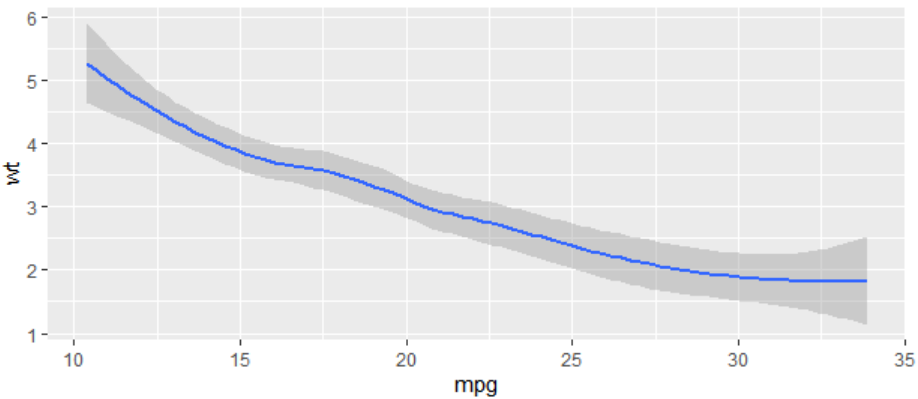
**Histogram**

****

**Path**



**Smooth**

****

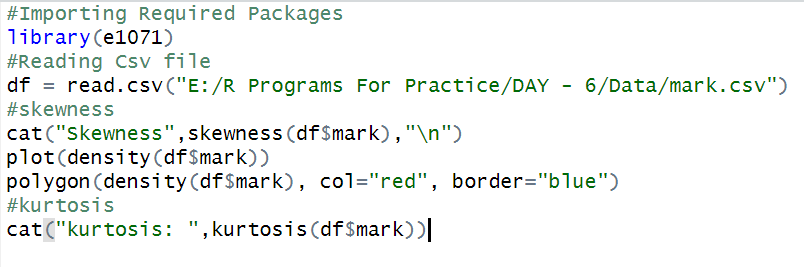
**Explanation:**

* **qplot**() is a shortcut designed to be familiar if you're used to base plot() .
* It's a convenient wrapper for creating a number of different types of plots using a consistent calling scheme.

**Date: 07/11/2020**

**Q.NO 38: Question**

**Program:**

Demonstrate the skewness and kurtosis using built-in dataset or your own dataset to get the data summary. 

**Output:**

****

**Explanation:**

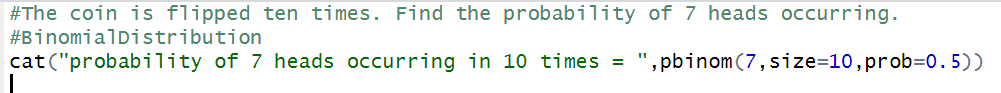
* **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry.
* **Kurtosis** is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

**Date: 10/11/2020**

**Q.NO 39: Question**

**Program:**

The coin is flipped ten times. Find the probability of 7 heads occurring.



**Output:**

****

**Explanation:**

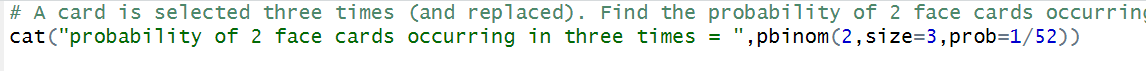
* The **binomial distribution**is a discrete probability distribution. It describes the outcome of *n*independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure.
* This function gives the cumulative probability of an event. It is a single value representing the probability.

**Date: 10/11/2020**

**Q.NO 40: Question**

**Program:**

A card is selected three times (and replaced). Find the probability of 2 face cards occurring.



**Output:**

****

**Explanation:**

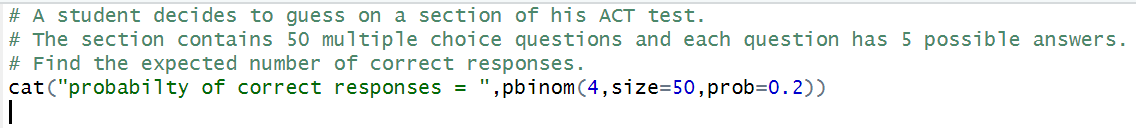
* The **binomial distribution**is a discrete probability distribution. It describes the outcome of *n*independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure.
* This function gives the cumulative probability of an event. It is a single value representing the probability.

**Date: 10/11/2020**

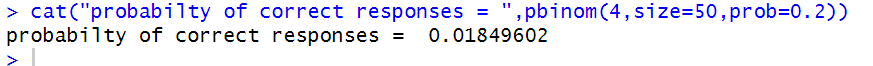
**Q.NO 41: Question**

**Program:**

A student decides to guess on a section of his ACT test. The section contains 50 multiple choice questions and each question has 5 possible answers. Find the expected number of correct responses.

****

**Output:**

****

**Explanation:**

* The **binomial distribution**is a discrete probability distribution. It describes the outcome of *n*independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure.
* This function gives the cumulative probability of an event. It is a single value representing the probability.

**Date: 10/11/2020**

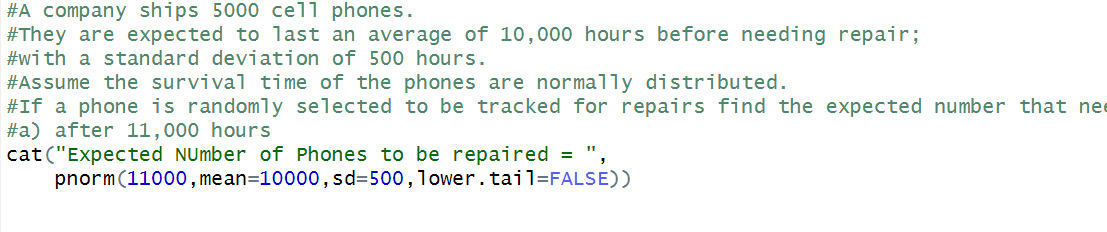
**Q.NO 42: Question**

**Program:**

A company ships 5000 cell phones. They are expected to last an average of 10,000 hours before needing repair; with a standard deviation of 500 hours.

Assume the survival time of the phones are normally distributed. If a phone is randomly selected to be tracked for repairs find the expected number that needs repair,

a) after 11,000 hours

****

**Output:**

****

**Explanation:**

* **pnorm** is the **R** function that calculates the c. d. f. where X is normal. Optional arguments described on the on-line documentation specify the parameters of the particular normal distribution.