Data Science Interview Questions and Answers in R Programming

1) How can you merge two data frames in R language?

Data frames in R language can be merged manually using cbind () functions or by using the merge () function on common rows or columns.

2) Explain about data import in R language

R Commander is used to import data in R language. To start the R commander GUI, the user must type in the command Rcmdr into the console. There are 3 different ways in which data can be imported in R language-

- Users can select the data set in the dialog box or enter the name of the data set (if they know).

- Data can also be entered directly using the editor of R Commander via Data->New Data Set. However, this works well when the data set is not too large.

- Data can also be imported from a URL or from a plain text file (ASCII), from any other statistical package or from the clipboard.

3) Two vectors X and Y are defined as follows – X <- c(3, 2, 4) and Y <- c(1, 2). What will be output of vector Z that is defined as Z <- X*Y.

In R language when the vectors have different lengths, the multiplication begins with the smaller vector and continues till all the elements in the larger vector have been multiplied.

The output of the above code will be –

Z <- (3, 4, 4)

4) How missing values and impossible values are represented in R language?

NaN (Not a Number) is used to represent impossible values whereas NA (Not Available) is used to represent missing values. The best way to answer this question would be to mention that deleting missing values is not a good idea because the probable cause for missing value could be some problem with data collection or programming or the query. It is good to find the root cause of the missing values and then take necessary steps handle them.

5) R language has several packages for solving a particular problem. How do you make a decision on which one is the best to use?

CRAN package ecosystem has more than 6000 packages. The best way for beginners to answer this question is to mention that they would look for a package that follows good software development principles. The next thing would be to look for user reviews and find out if other data scientists or analysts have been able to solve a similar problem.

6) Which function in R language is used to find out whether the means of 2 groups are equal to each other or not?

t.tests ()

7) What is the best way to communicate the results of data analysis using R language?

The best possible way to do this is combine the data, code and analysis results in a single document using knitr for reproducible research. This helps others to verify the findings, add to them and engage in discussions. Reproducible research makes it easy to redo the experiments by inserting new data and applying it to a different problem.

8) How many data structures does R language have?

R language has Homogeneous and Heterogeneous data structures. Homogeneous data structures have same type of objects – Vector, Matrix ad Array. Heterogeneous data structures have different type of objects – Data frames and lists.

9) What is the value of f (2) for the following R code?

```
b <- 4
f <- function (a)
{
b <- 3
b^3 + g (a)
}
g <- function (a)
{
a*b
}
```

The answer to the above code snippet is 35. The value of “a” passed to the function is 2 and the value for “b” defined in the function f (a) is 3. So the output would be 3^3 + g (2). The function g is defined in the global environment and it takes the value of b as 4(due to lexical scoping in R) not 3 returning a value 2*4= 8 to the function f. The result will be 3^3+8= 35.

10) What is the process to create a table in R language without using external files?

MyTable= data.frame ()

edit (MyTable)

The above code will open an Excel Spreadsheet for entering data into MyTable.

Learn Data Science in R Programming to land a top gig as an Enterprise Data Scientist!

11) Explain about the significance of transpose in R language

Transpose t () is the easiest method for reshaping the data before analysis.

12) What are with () and BY () functions used for?

With () function is used to apply an expression for a given dataset and BY () function is used for applying a function each level of factors.

13) dplyr package is used to speed up data frame management code. Which package can be integrated with dplyr for large fast tables?

data.table

14) In base graphics system, which function is used to add elements to a plot?

boxplot () or text ()

15) What are the different type of sorting algorithms available in R language?

Bucket Sort

Selection Sort

Quick Sort

Bubble Sort

Merge Sort

15) What is the command used to store R objects in a file?

save (x, file="x.Rdata")

16) What is the best way to use Hadoop and R together for analysis?

HDFS can be used for storing the data for long-term. MapReduce jobs submitted from either Oozie, Pig or Hive can be used to encode, improve and sample the data sets from HDFS into R. This helps to leverage complex analysis tasks on the subset of data prepared in R.

17) What will be the output of log (-5.8) when executed on R console?

Executing the above on R console will display a warning sign that NaN (Not a Number) will be produced because it is not possible to take the log of negative number.

18) How is a Data object represented internally in R language?

unclass (as.Date ("2016-10-05″))

19) What will be the output of the below code -

```
printmessage <- function (a) {

                if (is.na (a))

                        print ("a is a missing value!")

                  else if (a < 0)

                        print ("a is less than zero")

                  else

                       print ("a is greater than or equal to zero")

                 invisible (a)

 }

printmessage (NA)
```

The output for the above R programming code will be “a is a missing value.” The function is.na () is used to check if the input passed is a missing value.

20) Which package in R supports the exploratory analysis of genomic data?

adegenet

21) What is the difference between data frame and a matrix in R?

Data frame can contain heterogeneous inputs while a matrix cannot. In matrix only similar data types can be stored whereas in a data frame there can be different data types like characters, integers or other data frames.

23) How do you split a continuous variable into different groups/ranks in R?

24) What are factor variable in R language?

Factor variables are categorical variables that hold either string or numeric values. Factor variables are used in various types of graphics and particularly for statistical modelling where the correct number of degrees of freedom is assigned to them.

25) What is the memory limit in R?

8TB is the memory limit for 64-bit system memory and 3GB is the limit for 32-bit system memory.

26) What are the data types in R on which binary operators can be applied?

Scalars, Matrices ad Vectors.

27) How do you create log linear models in R language?

Using the loglm () function

28) What will be the class of the resulting vector if you concatenate a number and NA?

number

29) What is meant by K-nearest neighbour?

K-Nearest Neighbour is one of the simplest machine learning classification algorithms that is a subset of supervised learning based on lazy learning. In this algorithm the function is approximated locally and any computations are deferred until classification.

30) What will be the class of the resulting vector if you concatenate a number and a character?

character

31) Write code to build an R function powered by C?

32) If you want to know all the values in c (1, 3, 5, 7, 10) that are not in c (1, 5, 10, 12, 14). Which in-built function in R can be used to do this? Also, how this can be achieved without using the in-built function.

Using in-built function - setdiff(c (1, 3, 5, 7, 10), c (1, 5, 10, 11, 13))

Without using in-built function - c (1, 3, 5, 7, 10) [! c (1, 3, 5, 7, 10) %in% c (1, 5, 10, 11, 13).

33) How can you debug and test R programming code?

R code can be tested using Hadley's testthat package.

34) What will be the class of the resulting vector if you concatenate a number and a logical?

number

35) Write a function in R language to replace the missing value in a vector with the mean of that vector.

mean impute <- function(x) {x [is.na(x)] <- mean(x, na.rm = TRUE); x}

36) What happens if the application object is not able to handle an event?

The event is dispatched to the delegate for processing.

37) Differentiate between lapply and sapply.

If the programmers want the output to be a data frame or a vector, then sapply function is used whereas if a programmer wants the output to be a list then lapply is used. There one more function known as vapply which is preferred over sapply as vapply allows the programmer to specific the output type. The disadvantage of using vapply is that it is difficult to be implemented and more verbose.

38) Differentiate between seq (6) and seq_along (6)

Seq_along(6) will produce a vector with length 6 whereas seq(6) will produce a sequential vector from 1 to 6 c((1,2,3,4,5,6)).

39) How will you read a .csv file in R language?

read.csv () function is used to read a .csv file in R language. Below is a simple example –

filcontent <-read.csv (sample.csv)

print (filecontent)

40) How do you write R commands?

The line of code in R language should begin with a hash symbol (#).

41) How can you verify if a given object "X" is a matric data object?

If the function call is.matrix(X) returns TRUE then X can be termed as a matrix data object.

42) What do you understand by element recycling in R?

If two vectors with different lengths perform an operation –the elements of the shorter vector will be re-used to complete the operation. This is referred to as element recycling.

Example – Vector A <-c(1,2,0,4) and Vector B<-(3,6) then the result of A*B will be (3,12,0,24). Here 3 and 6 of vector B are repeated when computing the result.

43) How can you verify if a given object "X" is a matrix data object?

If the function call is.matrix(X) returns true then X can be considered as a matrix data object otheriwse not.

44) How will you measure the probability of a binary response variable in R language?

Logistic regression can be used for this and the function glm () in R language provides this functionality.

45) What is the use of sample and subset functions in R programming language?

Sample () function can be used to select a random sample of size 'n' from a huge dataset.

Subset () function is used to select variables and observations from a given dataset.

46) There is a function fn(a, b, c, d, e) a + b * c - d / e. Write the code to call fn on the vector c(1,2,3,4,5) such that the output is same as fn(1,2,3,4,5).

do.call (fn, as.list(c (1, 2, 3, 4, 5)))

47) How can you resample statistical tests in R language?

Coin package in R provides various options for re-randomization and permutations based on statistical tests. When test assumptions cannot be met then this package serves as the best alternative to classical methods as it does not assume random sampling from well-defined populations.

48) What is the purpose of using Next statement in R language?

If a developer wants to skip the current iteration of a loop in the code without terminating it then they can use the next statement. Whenever the R parser comes across the next statement in the code, it skips evaluation of the loop further and jumps to the next iteration of the loop.

49) How will you create scatterplot matrices in R language?

A matrix of scatterplots can be produced using pairs. Pairs function takes various parameters like formula, data, subset, labels, etc.

The two key parameters required to build a scatterplot matrix are –

- formula- A formula basically like ~a+b+c . Each term gives a separate variable in the pairs plots where the terms should be numerical vectors. It basically represents the series of variables used in pairs.
- data- It basically represents the dataset from which the variables have to be taken for building a scatterplot.

50) How will you check if an element 25 is present in a vector?

There are various ways to do this-

i. It can be done using the match () function- match () function returns the first appearance of a particular element.
ii. The other is to use %in% which returns a Boolean value either true or false.
iii. Is.element () function also returns a Boolean value either true or false based on whether it is present in a vector or not.

51) What is the difference between library() and require() functions in R language?

There is no real difference between the two if the packages are not being loaded inside the function. require () function is usually used inside function and throws a warning

whenever a particular package is not found. On the flip side, library () function gives an error message if the desired package cannot be loaded.

52) What are the rules to define a variable name in R programming language?

A variable name in R programming language can contain numeric and alphabets along with special characters like dot (.) and underline (-). Variable names in R language can begin with an alphabet or the dot symbol. However, if the variable name begins with a dot symbol it should not be a followed by a numeric digit.

53) What do you understand by a workspace in R programming language?

The current R working environment of a user that has user defined objects like lists, vectors, etc. is referred to as Workspace in R language.

54) Which function helps you perform sorting in R language?

Order ()

55) How will you list all the data sets available in all R packages?

Using the below line of code-

data(package = .packages(all.available = TRUE))

56) Which function is used to create a histogram visualisation in R programming language?

Hist()

57) Write the syntax to set the path for current working directory in R environment.

Setwd("dir_path")

58) How will you drop variables using indices in a data frame?

Let's take a dataframe df<-data.frame(v1=c(1:5),v2=c(2:6),v3=c(3:7),v4=c(4:8))

```
df

##   v1 v2 v3 v4

## 1  1  2  3  4

## 2  2  3  4  5

## 3  3  4  5  6

## 4  4  5  6  7

## 5  5  6  7  8
```

Suppose we want to drop variables v2 & v3 , the variables v2 and v3 can be dropped using negative indicies as follows-

df1<-df[-c(2,3)]

```
df1

##   v1 v4

## 1  1  4

## 2  2  5

## 3  3  6

## 4  4  7

## 5  5  8
```

59) What will be the output of runif (7)?

It will generate 7 randowm numbers between 0 and 1.

60) What is the difference between rnorm and runif functions?

rnorm function generates "n" normal random numbers based on the mean and standard deviation arguments passed to the function.

Syntax of rnorm function -

rnorm(n, mean = , sd =)

runif function generates "n" unform random numbers in the interval of minimum and maximum values passed to the function.

Syntax of runif function -

runif(n, min = , max =)

R Interview Questions for Data Science

1) What is the need of factorizing variables in R?

2) List some of your favorite functions in R programming language along with their usage.

3) Explain the differences between Python and R.

4) What is multi-threading and how can you implement it in R programming language?

5) Implement string operations in R language.

6) dplyr <- "ggplot2" library(dplyr). Which package will be loaded on executing the command and why?

7) Why you should use R language for statistical work?

8) What according to you are disadvantages of R Programming over Python?

9) Which R objects have you most frequently worked with?

10) Build a binary search tree in R language.

11) How can you produce co-relations and covariances in R lanaguge?

12) How can you develop a package in R language and do version control?