1) Explain what is R?

This should be an easy one for data science job applicants. R is an open-source language and environment for statistical computing and analysis, or for our purposes, data science.

R is data analysis software which is used by analysts, quants, statisticians, data scientists and others.

10) What are the applications of R?

There are various applications available in real-time. These applications are as follows:

1. Facebook
2. Google
3. Twitter
4. HRDAG
5. NDAA

11) Explain RStudio.

RStudio is an integrated development environment which allows us to interact with R more readily. RStudio is similar to the standard RGui, but it is considered more user-friendly. This IDE has various drop-down menus, windows with multiple tabs, and so many customization processes. The first time when we open RStudio, we will see three Windows. The fourth Window will be hidden by default.

19) Differentiate b/w "%%" and "%/%".

The "%%" provides a reminder of the division of the first vector with the second, and the "%/%" gives the quotient of the division of the first vector with the second.

20) Why do we use apply() function in R?

This is used to apply the same function to each of the elements in an Array. For example, finding the mean of the rows in every row.

27) Explain the use of the table() function.

This function is used to create the frequency table in R.

R Interview Questions

A list of frequently asked **R Interview Questions and answers** are given below.

1) What is R?

R is an interpreted computer programming language which was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand". It is a software environment used to analyze statistical information, graphical representation, reporting, and data modeling. R is the implementation of the S programming language, which is combined with lexical scoping semantics.

---

2) Differentiate between vector, List, Matrix, and Data frame.

A **vector** is a series of data elements of the same basic type. The members in the vector are known as a component.

The R object that contains elements of different types such as numbers, strings, vectors, or another list inside it, is known as **List**.

A two-dimensional data structure used to bind the vectors from the same length, known as the **matrix**. The matrix contains the same types of elements.

A **Data** frame is a generic form of a matrix. It is a combination of lists and matrices. In the Data frame, different data columns contain different data types.

---

3) Give names of those packages which are used for data imputation.

There are the following packages which are used for data imputation

1. MICE
2. missFores
3. Mi
4. Hmisc
5. Amelia
6. imputeR

---

4) Explain initialize() function in R?

This function is used to initialize the private data members while declaring the object.

---

5) How can we find the mean of one column with respect to another?

In iris dataset, there are five columns, i.e., Sepal.Length, Sepal.Width, Petal.Length, Petal.Width and Species. We will calculate the mean of Sepal-Length across different species of iris flower using the mean() function from the mosaic package.

1. mean(iris$Sepal.Length~iris$Species)

6) What is a Random Walk model?

A random walk is the simplest example of a non-stationary process. A random walk has no specified mean or variance, strong dependence over time, and its changes or increments are white noise. Simulating random walk in R:

arima.sim(model=list(order=c(0,1,0)),n=40)->rw ts.plot(rw)

---

7) What is a White Noise model?

It is a basic time series model and a simple example of a stationary process. A white noise model has a fixed constant mean, a fixed constant variance, and no correlation over time. We can simulate a white noise model in the following way:

arima.sim(model=list(order=c(0,0,0)),n=50)->wn

---

8) Give any five features of R.

1. Simple and effective programming language.
2. It is a data analysis software.
3. It gives effective storage facility and data handling.
4. It gives high extensible graphical techniques.
5. It is an interpreted language.

---

9) Differentiate between R and Python in terms of functionality?

For data analysis, R has inbuilt functionality, but in Python, the data analysis functionalities are not inbuilt. They are available by packages like Pandas and Numpy.

---

10) What are the applications of R?

There are various applications available in real-time. These applications are as follows:

1. Facebook
2. Google
3. Twitter
4. HRDAG
5. NDAA

---

11) Explain RStudio.

RStudio is an integrated development environment which allows us to interact with R more readily. RStudio is similar to the standard RGui, but it is considered more user-friendly. This IDE has various drop-down menus, windows with multiple tabs, and so many customization processes. The first time when we open RStudio, we will see three Windows. The fourth Window will be hidden by default.

---

12) What are the advantages and disadvantages of R?

Advantages

1. Open Source
2. Data Wrangling

3. Array of Packages
4. Platform Independent
5. Machine Learning Operations

Disadvantages

1. Weak origin
2. Data Handling
3. Basic Security
4. Complicated Language
5. Lesser Speed

---

13) What is the purpose behind R and Hadoop integration?

1. For executing Hadoop to execute R code.
2. For using R to access the data stored in Hadoop.

---

14) Give the name of the Hadoop integration methods.

1. R Hadoop
2. Hadoop Streaming
3. RHIPE
4. ORCH

---

15) What will be the output of the expression all(NA==NA)?

[1] NA

16) What is the difference b/w sample() and subset() in R?

The sample() method is used to choose a random sample of size n from a dataset while the subset method is used to choose variables and observations.

---

17) Why do we use the command - install.packages(file.choose(), repos=NULL)?

This command is used to install an R package from the local directory by browsing and selecting the file.

---

18) Give the command to create a histogram and to remove a vector from the R workspace?

hist() and rm() function are used as a command to create a histogram and remove a vector from the R workspace.

---

19) Differentiate b/w "%%" and "%/%".

The "%%" provides a reminder of the division of the first vector with the second, and the "%/%" gives the quotient of the division of the first vector with the second.

---

20) Why do we use apply() function in R?

This is used to apply the same function to each of the elements in an Array. For example, finding the mean of the rows in every row.

---

21) Differentiate between library() and require() functions.

If the desired package cannot be loaded, then the library() function gives an error message and display while the required () function is used inside the function and throws a warning message whenever a particular package is not found.

---

22) What is the t-test() in R?

The t-test() function is used to determine that the mean of the two groups are equal or not.

---

23) What is the use of with() and by() functions in R?

The with() function applies an expression to a dataset, and the by() function applies a function to each level of factors.

---

24) Differentiate b/w lapply and sapply.

The lapply is used to show the output in the form of the list, whereas sapply is used to show the output in the form of a vector or data frame.

---

25) Explain aggregate() function.

The aggregate() function is used to aggregate data in R. There are two methods which are collapsing data by using one or more BY variable and other is an aggregate() function in which By variable should be in the list.

---

26) Explain the doBy package?

This package is used to define the desired table using function and model formula.

---

27) Explain the use of the table() function.

This function is used to create the frequency table in R.

---

28) Explain fitdistr() function?

This function is used to give the maximum likelihood fitting of univariate distribution and defined under the MASS package.

---

29) What are GGobi and iPlots?

The GGobi is an open-source program for visualization to exploring high dimensional typed data, and the iPlots is a package which provides bar plots, mosaic plots, box plots, parallel plots, histograms, and scatter plots.

---

30) Explain the lattice package.

The lattice package is meant to improve upon the base R graphics by giving better defaults and has the ability to display multivariate relationships easily.

---

31) Explain anova() function.

The anova() function is used for comparing the nested models.

---

32) Explain cv.lm() and stepAIC() function.

The cv.lm() function is defined under the DAAG package used for k-fold validation while the stepAIC() function is defined under the MASS package that performs stepwise model selection under exactAIC.

---

33) Explain leaps() function.

The leaps() function is used to perform the all-subsets regression and defined under the leaps package.

---

34) Explain relaimpo and robust package.

This package is used to measure the relative importance of every predictor in the model, and the robust package gives a library of robust methods, including regression.

---

35) Give full form of MANOVA and what is the use of it.

MANOVA stands for Multivariate Analysis of Variance, and it is used to test more than one dependent variable simultaneously.

---

36) Explain mashapiro.test() and barlett.test().

This function defines in the mvnormtest package and produces the Shapiro-wilk test to multivariate normality. The barlett.test() is used to provide a parametric k-sample test of the equality of variances.

---

37) Explain the use of the forecast package.

The forecast package gives the functions which are used to automatic selection of exponential and ARIMA models.

---

38) Differentiate between qda() and lda() function.

The qda() function prints a quadratic discriminant function while lda() function print the discriminant functions based on the centered variable.

---

39) Explain the auto.arima() and principal() function.

The auto.arima() function handle both the seasonal and non-seasonal ARIMA model and the principal() function used for rotating and extracting the principal components.

---

40) Explain FactoMineR.

The FactoMineR is a package that includes qualitative and quantitative variables. The observations and supplementary variables are also included in these packages.

---

41) What is the full form of SEM and CFA?

CFA stands for Confirmatory Factor Analysis, and SEM stands for Structural Equation Modeling.

---

42) Define cluster.stats() and pvclust() function().

The cluster.stats() function define in the fpc package that provides a method for comparing the similarity of two cluster solutions using different validation criteria, and the pvclust() function is defined in the pvclust package that provides p-values for hierarchical clustering.

---

43) Define MATLAB and party packages.

This package includes wrapper functions and variable which are used for replicating Matlab function calls.

---

44) Explain S3 and S4 systems.

In oops, the S3 is used to overload any function. So that we can call the functions with different names, and it depends on the type of input parameter or the number of parameters, and the S4 is the most important characteristic of oops. However, this is a limitation, as it is quite difficult to debug. There is an optional reference class for S4.

45) Give names of visualization packages.

There are the following packages of visualization in R:

1. Plotly
2. ggplot2
3. tidyquant
4. geofacet
5. googleVis
6. Shiny

Explain Pie chart in R.

R programming language has several libraries for creating charts and graphs. A pie-chart is a representation of values in the form of slices of a circle with different colors.

Explain Histogram.

A histogram is a type of bar chart which shows the frequency of the number of values which are compared with a set of values ranges. The histogram is used for the distribution, whereas a bar chart is used for comparing different entities. In the histogram, each bar represents the height of the number of values present in the given range.

1. Compare R & Python

Model Building is similar to Python	Model Building is similar to R.

Model Interpretability is good	Model Interpretability is not good
Production is not better than Python.	Production is good
R has good community support over Python.	Community Support is not better than R
Data Science Libraries are same as Python.	Data Science Libraries are same as R.
R has good data visualizations libraries and tools	Data visualization is not better than R
R has a steep learning curve.	Learning Curve in Python is easier than learning R.

Difference between library () and require () functions in R language.

Library () function gives an error message display, if the desired package cannot be loaded.	Require () function is used inside function and throws a warning messages whenever a particular package is not Found

It loads the packages whether it is already loaded or not,	It just checks that it is loaded, or loads it if it isn't (use in functions that rely on a certain package). The documentation explicitly states that neither function will reload an already loaded package.

4) In R how you can import Data?

You use R commander to import Data in R, and there are three ways through which you can enter data into it

- You can enter data directly via Data New Data Set
- Import data from a plain text (ASCII) or other files (SPSS, Minitab, etc.)
- Read a data set either by typing the name of the data set or selecting the data set in the dialog box

3)Mention what does not 'R' language do?

- **Though R programming can easily connects to DBMS is not a database**
- **R does not consist of any graphical user interface**
- **Though it connects to Excel/Microsoft Office easily, R language does not provide any spreadsheet view of data**

11) What are the data structures in R that is used to perform statistical analyses and create graphs?

R has data structures like

- **Vectors**
- **Matrices**
- **Arrays**
- **Data frames**

12) Explain general format of Matrices in R?

General format is

```
Mymatrix< - matrix (vector, nrow=r , ncol=c ,
byrow=FALSE,
dimnames = list ( char_vector_ rowname,
char_vector_colnames))
```

13) In R how missing values are represented ?

In R missing values are represented by NA (Not Available), why impossible values are represented by the symbol NaN (not a number).

14) Explain what is transpose?

For re-shaping data before, analysis R provides various method and transpose are the simplest method of reshaping a dataset. To transpose a matrix or a data frame t () function is used.

Can you write and explain some of the most common syntax in R?

Again, this is an easy—but crucial—one to nail. For the most part, this can be demonstrated through any other code you might write for other R interview questions, but sometimes this is asked as a standalone. Some of the basic syntax for R that's used most often might include:

 # — as in many other languages, # can be used to introduce a line of comments. This tells the compiler not to process the line, so it can be used to make code more readable by reminding future inspectors what blocks of code are intended to do.

"" — quotes operate as one might expect; they denote a string data type in R.

<- — one of the quirks of R, the assignment operator is <- rather than the relatively more familiar use of =. This is an essential thing for those using R to know, so it would be good to display your knowledge of it if the question comes up.

\ — the backslash, or reverse virgule, is the escape character in R. An escape character is used to "escape" (or ignore) the special meaning of certain characters in R and, instead, treat them literally.

13. What are the advantages of R?

- **The advantages are:-**
- **It is used for managing and manipulating of data.**
- **No license restrictions**
- **Free and open source software.**
- **Graphical capabilities of R are good.**
- **Runs on many Operating system and different hardware and also run on 32 & 64 bit processors etc.**

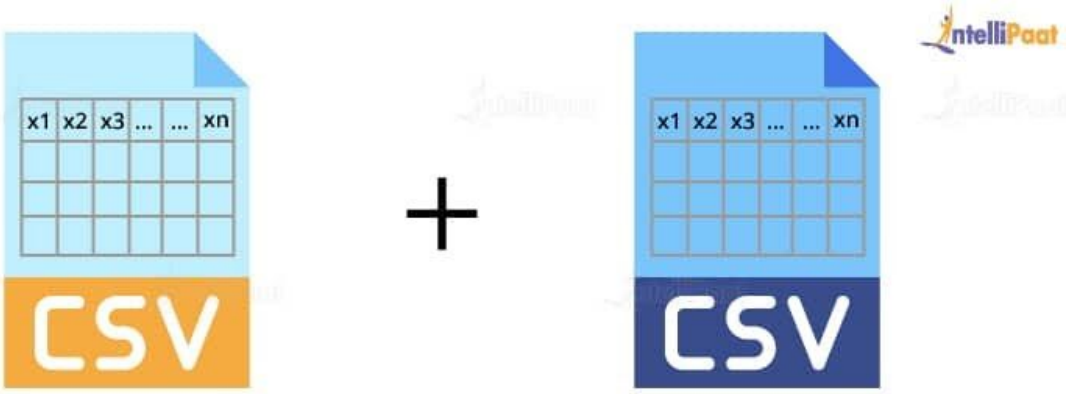
8. What are the disadvantages of R Programming?

The disadvantages are:-

- Lack of standard GUI
- Not good for big data.
- Does not provide spreadsheet view of data.

14. What is the function used for adding datasets in R?

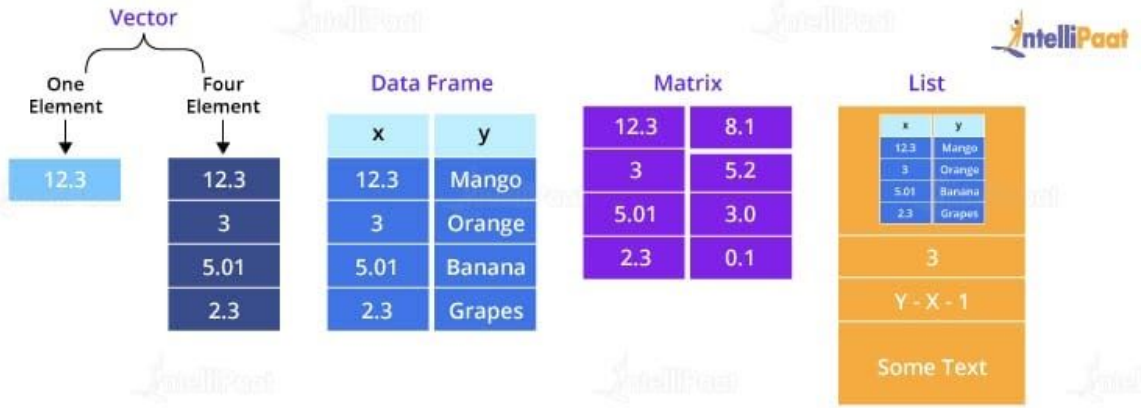
For adding two datasets rbind() function is used but the column of two datasets must be same.




```
Syntax: rbind(x1,x2……) where x1,x2: vector, matrix, data frames.
```

What is difference between matrix and dataframes?

Dataframe can contain different type of data but matrix can contain only similar type of data. Here are the different types of data structures in R:



What is difference between lapply and sapply?

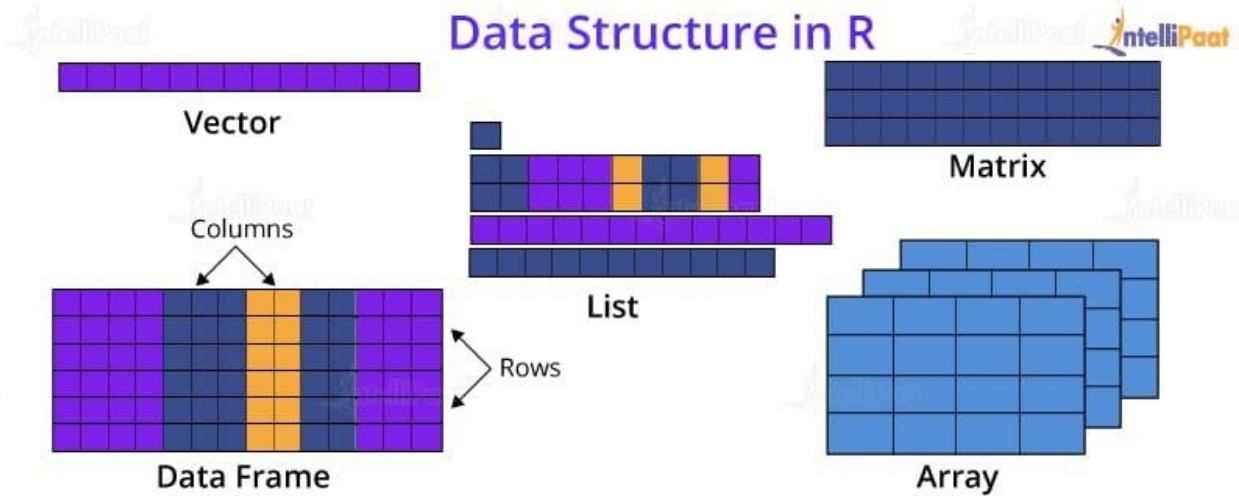
lapply is used to show the output in the form of list whereas sapply is used to show the output in the form of vector or data frame

What is the memory limit of R?

In 32 bit system memory limit is 3Gb but most versions limited to 2Gb and in 64 bit system memory limit is 8Tb.

How many data structures R has?

There are 5 data structure in R i.e. vector, matrix, array which are of homogenous type and other two are list and data frame which are heterogeneous.



When is it appropriate to use the “next” statement in R?

A data scientist will use next to skip an iteration in a loop. As an example:

```
x <- 1:20
for (val in x) {
      if (val == 15) {
      next
      }
print(val)
}
```

9. What is the use of With () and By () function in R?

with() function applies an expression to a dataset.

`#with(data,expression)`

By() function applies a function t each level of a factors.

`#by(data,factorlist,function)`

10. In R programming, how missing values are represented?

In R missing values are represented by NA which should be in capital letters.

What are the different data types/objects in R?

This is another good opportunity to show that you *know* R, and you're not winging it. Unlike other object-oriented languages such as C, R doesn't ask users to declare a data type when assigning a variable. Instead, everything in R correlates to an R data object. When you assign a variable in R, you assign it a data object and that object's data type determines the data type of the variable. The most commonly used data objects include:

- **Vectors**
- **Matrices**
- **Lists**
- **Arrays**

- Factors
- Data frames

Why use R?

This is a variant of the "advantages of R" question. Reasons to use R include its open-source nature and the fact that it's a versatile tool for statistical plotting, analysis, and portrayal. Don't be afraid to give some personal reasons as well. Maybe you simply love the assignment operator in R or feel that it's more elegant than other languages—but always remember to explicate. You should be answering follow-up questions before they're even asked.

Write a custom function in R

An example of a custom function

```
myFunction <- function(arg1, arg2, ... ){

statements

return(object)

}
```

Functions can be simple or complex, but they should make your code more extensible, readable, and efficient. This is a chance to show your ingenuity and experience.

How do you import data in R?

Let's use CSV as an example, as it's a very common data format. Simply make sure the file is saved in a CSV format, then use the read function to import the data.

```
yourRDateHere <- read.csv("Data.csv", header = TRUE)
```

Though not required, strictly speaking, the argument `header = TRUE` is used to ensure that labels are not parsed as data.

16. How do you install a package in R?

There are many ways to install a package in R. Some even include using the GUI. We're coders, so we're not going to give those attention.

Type the following into your console and hit enter:

```
install.packages("package_name")
```

Followed by:

```
library(package_name)
```

19. When is it appropriate to use mode()?

By default, `mode()` gets or sets the storage mode of an object. It's default usage is equivalent to storage.mode(). A sample usage:

```
x <- 1:25

mode(x)

[1] "numeric"

y <- "helloWorld"

mode(y)
```

```
[1] "character"
```

```
mode(state.name)
```

```
[1] "character"
```

21. When is it appropriate to use the which() function?

The `which()` function loops through a logical object until the condition returns TRUE and returns the index `(position)` of the element.

To get a sense of how this works, plug in the letters array and search for the index of a specific letter using `which()`.

```
letters
[1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r"
"s" "t" "u" "v" "w" "x" "y" "z"
which(letters == "a")
[1] 1
which(letters == "z")
[1] 26
which(letters == "m")
[1] 13
```

In my console, I've checked the letters array, which contains the English alphabet in lowercase. I've used `which()` to find the positions of `a`, `z`, and `m`, which returned the indexes `1`, `26`, and `13`, respectively, because these are the positions in the array, as they are typically the positions in the alphabet

How do you concatenate strings in R?

Concatenating strings in R is less than intuitive. You don't use a `.` operator, nor a `+` operator, and forget about the `&` operator. In fact, you don't use an operator at all. Concatenating strings in R requires the use of the `paste()` function. Here's an example:

```
hello <- "Hello, "

world <- "World."

paste(hello, world)

[1] "Hello, World."
```

I've stored `Hello,` and `World.` in variables aptly named `hello` and `world`. With paste(), I've simply plugged in the two variables, and it concatenates them such that it creates the single phrase "Hello, World."

How many sorting algorithms are available?

There are 5 types of sorting algorithms are used which are:-

- Bubble Sort
- Selection Sort
- Merge Sort
- Quick Sort
- Bucket Sort

24. How to create new variable in R programming?

```
For creating new variable assignment operator '<-' is used
For e.g. mydata$sum <- mydata$x1 + mydata$x2
```

25. What are R packages?

Packages are the collections of data, R functions and compiled code in a well-defined format and these packages are stored in library. One of the strengths of R is the user-written function in R language.

26. What is the workspace in R?

Workspace is the current R working environment which includes any user defined objects like vector, lists etc.

27. What is the function which is used for merging of data frames horizontally in R?

Merge()function is used to merge two data frames

```
Eg. Sum<-merge(data frame1,data frame 2,by='ID')
```

28. what is the function which is used for merging of data frames vertically in R?

rbind() function is used to merge two data frames vertically.

```
Eg. Sum <- rbind(data frame1,data frame 2)
```

How to create axes in the graph?

Using axes() function custom axes are created.

44. What is the use of abline() function?

abline() function is add the reference line to a graph.

```
Syntax: abline(h=yvalues, v=xvalues)
```

52. Why library() function is used?

This function is used to show the packages which are installed.

Give any five features of R.

1. **Simple and effective programming language.**
2. **It is a data analysis software.**
3. **It gives effective storage facility and data handling.**
4. **It gives high extensible graphical techniques.**
5. **It is an interpreted language.**