
DATA
SCIENCE
Interview Questions
and Answers

Table of Contents

GENERAL DATA SCIENCE INTERVIEW QUESTIONS AND ANSWERS 2

R PROGRAMMING DATA SCIENCE INTERVIEW QUESTIONS AND ANSWERS 19

PYTHON DATA SCIENCE INTERVIEW QUESTIONS AND ANSWERS 35

General Data Science Interview Questions and Answers

1. **How would you create a taxonomy to identify key customer trends in unstructured data?**
   The best way to approach this question is to mention that it is good to check with the business owner and understand their objectives before categorizing the data. Having done this, it is always good to follow an iterative approach by pulling new data samples and improving the model accordingly by validating it for accuracy by soliciting feedback from the stakeholders of the business. This helps ensure that your model is producing actionable results and improving over the time.

2. **Python or R – Which one would you prefer for text analytics?**
   The best possible answer for this would be Python because it has Pandas library that provides easy to use data structures and high performance data analysis tools.

3. **Which technique is used to predict categorical responses?**
   Classification technique is used widely in mining for classifying data sets.

4. **What is logistic regression? Or State an example when you have used logistic regression recently.**
   Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

5. **What are Recommender Systems?**
   A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

6. **Why data cleaning plays a vital role in analysis?**
   Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

7. **Differentiate between univariate, bivariate and multivariate analysis.**
   These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis.

If the analysis attempts to understand the difference between 2 variables at time as in a scatterplot, then it is referred to as bivariate analysis. For example, analyzing the volume of sale and a spending can be considered as an example of bivariate analysis.

Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

8. **What do you understand by the term Normal Distribution?**
   Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell shaped curve. The random variables are distributed in the form of an symmetrical bell shaped curve.

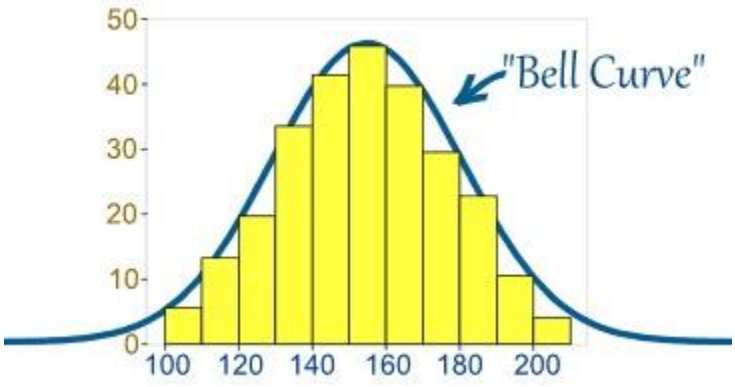


Image Credit : mathisfun.com

9. **What is Linear Regression?**
   Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.

10. **What is Interpolation and Extrapolation?**
    Estimating a value from 2 known values from a list of values is Interpolation. Extrapolation is approximating a value by extending a known set of values or facts.

11. **What is power analysis?**
    An experimental design technique for determining the effect of a given sample size.

12. **What is Collaborative filtering?**
    The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources and multiple agents.

13. What is the difference between Cluster and Systematic Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection, or cluster of elements. Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example for systematic sampling is equal probability method.

14. Are expected value and mean value different?

They are not different but the terms are used in different contexts. Mean is generally referred when talking about a probability distribution or sample population whereas expected value is generally referred in a random variable context.

For Sampling Data

Mean value is the only value that comes from the sampling data.

Expected Value is the mean of all the means i.e. the value that is built from multiple samples. Expected value is the population mean.

For Distributions

Mean value and Expected value are same irrespective of the distribution, under the condition that the distribution is in the same population.

15. What does P-value signify about the statistical data?

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

- P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
- P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
- P-value=0.05is the marginal value indicating it is possible to go either way.

16. Do gradient descent methods always converge to same point?

No, they do not because in some cases it reaches a local minima or a local optima point. You don't reach the global optima point. It depends on the data and starting conditions**.**

17. A test has a true positive rate of 100% and false positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?

Let's suppose you are being tested for a disease, if you have the illness the test will end up saying you have the illness. However, if you don't have the illness- 5% of the times the test will end up saying you have the illness and 95% of the times the test will give accurate result that you don't have the illness. Thus there is a 5% error in case you do not have the illness.

Out of 1000 people, 1 person who has the disease will get true positive result.

Out of the remaining 999 people, 5% will also get true positive result.

Close to 50 people will get a true positive result for the disease.

This means that out of 1000 people, 51 people will be tested positive for the disease even though only one person has the illness. There is only a 2% probability of you having the disease even if your reports say that you have the disease.

18. What is the difference between Supervised Learning an Unsupervised Learning?

If an algorithm learns something from the training data so that the knowledge can be applied to the test data, then it is referred to as Supervised Learning. Classification is an example for Supervised Learning. If the algorithm does not learn anything beforehand because there is no response variable or any training data, then it is referred to as unsupervised learning. Clustering is an example for unsupervised learning.

19. What is the goal of A/B Testing?

It is a statistical hypothesis testing for randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of an interest. An example for this could be identifying the click through rate for a banner ad.

20. What is an Eigenvalue and Eigenvector?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

21. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values –

- To change the value and bring in within a range
- To just remove the value.

22. How can you assess a good logistic model?

There are various methods to assess the results of a logistic regression analysis-

- Using Classification Matrix to look at the true negatives and false positives.
- Concordance that helps identify the ability of the logistic model to differentiate between the event happening and not happening.
- Lift helps assess the logistic model by comparing it with random selection.

23. What are various steps involved in an analytics project?

- Understand the business problem
- Explore the data and become familiar with it.
- Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.
- After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step till the best possible outcome is achieved.
- Validate the model using a new data set.
- Start implementing the model and track the result to analyze the performance of the model over the period of time.

24. How can you iterate over a list and also retrieve element indices at the same time?

This can be done using the enumerate function which takes every element in a sequence just like in a list and adds its location just before it.

25. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. There are various factors to be considered when answering this question:

- Understand the problem statement, understand the data and then give the answer. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

- If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.

- If you have a distribution of data coming, for normal distribution give the mean value.

- Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

26. Can you use machine learning for time series analysis?

Yes, it can be used but it depends on the applications.

27. Write a function that takes in two sorted lists and outputs a sorted list that is their union.

First solution which will come to your mind is to merge two lists and short them afterwards

Python code-

```
def return_union(list_a, list_b):
    return sorted(list_a + list_b)
```

R code-

```
return_union <- function(list_a, list_b)
{
list_c<-list(c(unlist(list_a),unlist(list_b)))
return(list(list_c[[1]][order(list_c[[1]])]))
}
```

Generally, the tricky part of the question is not to use any sorting or ordering function. In that case you will have to write your own logic to answer the question and impress your interviewer.

Python code-

```
def return_union(list_a, list_b):
    len1 = len(list_a)
    len2 = len(list_b)
    final_sorted_list = []
    j = 0
    k = 0

    for i in range(len1+len2):
        if k == len1:
            final_sorted_list.extend(list_b[j:])
            break
        elif j == len2:
            final_sorted_list.extend(list_a[k:])
            break
        elif list_a[k] < list_b[j]:
            final_sorted_list.append(list_a[k])
            k += 1
        else:
            final_sorted_list.append(list_b[j])
            j += 1
    return final_sorted_list
```

Similar function can be returned in R as well by following the similar steps.

```
return_union <- function(list_a,list_b)
{
#Initializing length variables
len_a <- length(list_a)
len_b <- length(list_b)
len <- len_a + len_b

#initializing counter variables

j=1
k=1

#Creating an empty list which has length equal to sum of both the lists

list_c <- list(rep(NA,len))
```

```
#Here goes our for loop

for(i in 1:len)
 {
  if(j>len_a)
    {
     list_c[i:len] <- list_b[k:len_b]
     break
    }
  else if(k>len_b)
    {
     list_c[i:len] <- list_a[j:len_a]
     break
    }
  else if(list_a[[j]] <= list_b[[k]])
    {
     list_c[[i]] <- list_a[[j]]
     j <- j+1
    }
  else if(list_a[[j]] > list_b[[k]])
  {
   list_c[[i]] <- list_b[[k]]
   k <- k+1
  }
 }
 return(list(unlist(list_c)))

 }
```

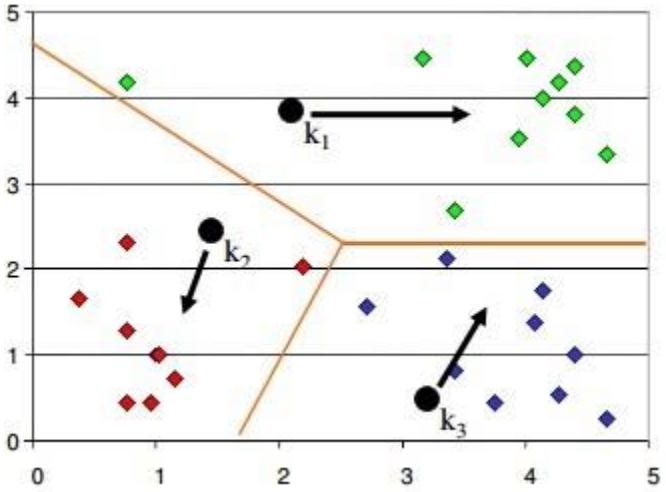
28. What is Machine Learning?

The simplest way to answer this question is – we give the data and equation to the machine. Ask the machine to look at the data and identify the coefficient values in an equation.

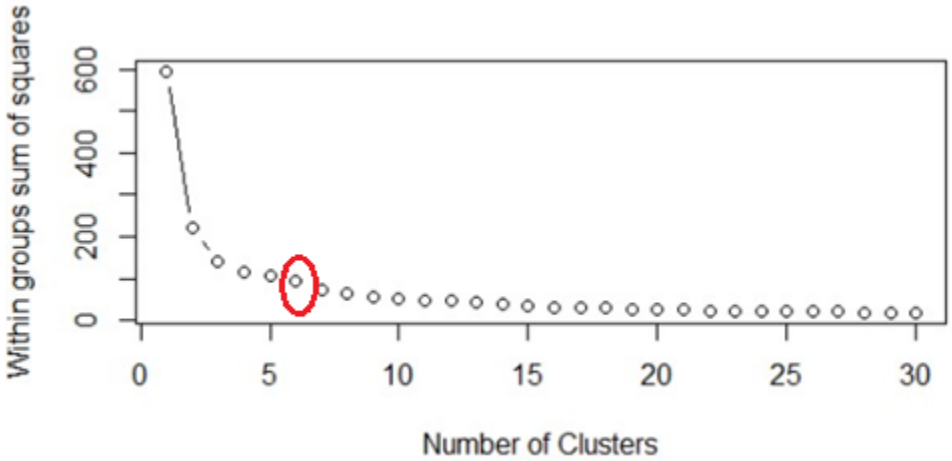
For example for the linear regression y=mx+c, we give the data for the variable x, y and the machine learns about the values of m and c from the data.

29. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question will mostly be asked in reference to K-Means clustering where "K" defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.
For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don't see any decrement in WSS. This point is known as bending point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendograms and identify the distinct groups from there.

30. Is it possible to perform logistic regression with Microsoft Excel?

It is possible to perform logistic regression with Microsoft Excel. There are two ways to do it using Excel.

- One is to use Add-ins provided by many websites which we can use.
- Second is to use fundamentals of logistic regression and use Excel's computational power to build a logistic regression

But when this question is being asked in an interview, interviewer is not looking for a name of Add-ins rather a method using the base excel functionalities.

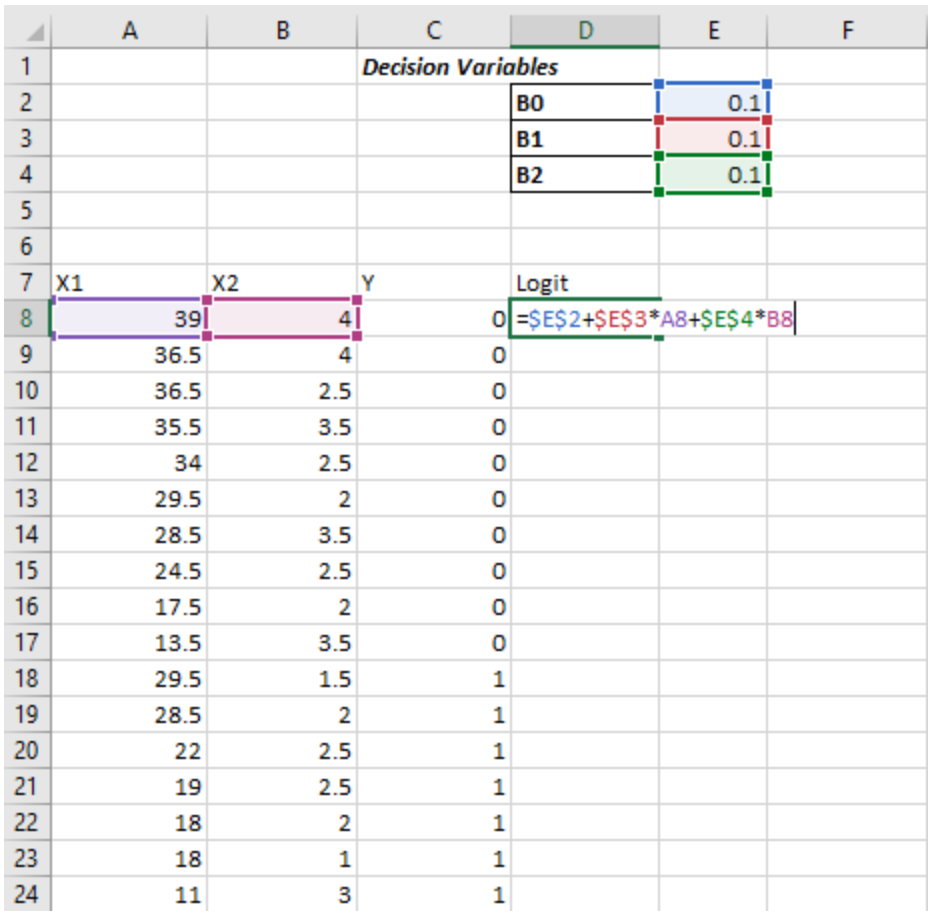
Let's use a sample data to learn about logistic regression using Excel. (Example assumes that you are familiar with basic concepts of logistic regression)

	A	B	C
6			
7	X1	X2	Y
8	39	4	0
9	36.5	4	0
10	36.5	2.5	0
11	35.5	3.5	0
12	34	2.5	0
13	29.5	2	0
14	28.5	3.5	0
15	24.5	2.5	0
16	17.5	2	0
17	13.5	3.5	0
18	29.5	1.5	1
19	28.5	2	1
20	22	2.5	1
21	19	2.5	1
22	18	2	1
23	18	1	1
24	11	3	1
25	11	2.5	1
26	7.5	2	1
27	5	3	1

Data shown above consists of three variables where X1 and X2 are independent variables and Y is a class variable. We have kept only 2 categories for our purpose of binary logistic regression classifier.

Next we have to create a logit function using independent variables, i.e.

Logit = L = $\beta_0 + \beta_1 * X_1 + \beta_2 * X_2$



	A	B	C	D	E	F
1			*Decision Variables*			
2				B0	0.1	
3				B1	0.1	
4				B2	0.1	
5						
6						
7	X1	X2	Y	Logit		
8	39	4	0	=$E$2+$E$3*A8+$E$4*B8		
9	36.5	4	0			
10	36.5	2.5	0			
11	35.5	3.5	0			
12	34	2.5	0			
13	29.5	2	0			
14	28.5	3.5	0			
15	24.5	2.5	0			
16	17.5	2	0			
17	13.5	3.5	0			
18	29.5	1.5	1			
19	28.5	2	1			
20	22	2.5	1			
21	19	2.5	1			
22	18	2	1			
23	18	1	1			
24	11	3	1			

31. You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?

Since the question asked, is about post model building exercise, we will assume that you have already tested for null hypothesis, multi collinearity and Standard error of coefficients.

Once you have built the model, you should check for following –

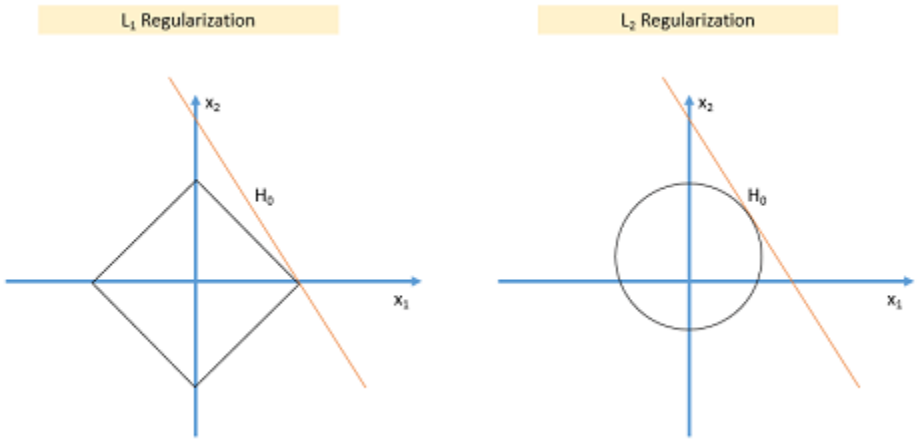
- Global F-test to see the significance of group of independent variables on dependent variable
- R^2
- Adjusted R^2
- RMSE, MAPE

In addition to above mentioned quantitative metrics you should also check for-

- Residual plot
- Assumptions of linear regression

32. Why L1 regularizations causes parameter sparsity whereas L2 regularization does not?

Regularizations in statistics or in the field of machine learning is used to include some extra information in order to solve a problem in a better way. L1 & L2 regularizations are generally used to add constraints to optimization problems.



In the example shown above H_0 is a hypothesis. If you observe, in L_1 there is a high likelihood to hit the corners as solutions while in L_2, it doesn't. So in L_1 variables are penalized more as compared to L2 which results into sparsity.

In other words, errors are squared in L_2, so model sees higher error and tries to minimize that squared error.

33. How can you deal with different types of seasonality in time series modelling?

Seasonality in time series occurs when time series shows a repeated pattern over time. E.g., stationary sales decreases during holiday season, air conditioner sales increases during the summers etc. are few examples of seasonality in a time series.

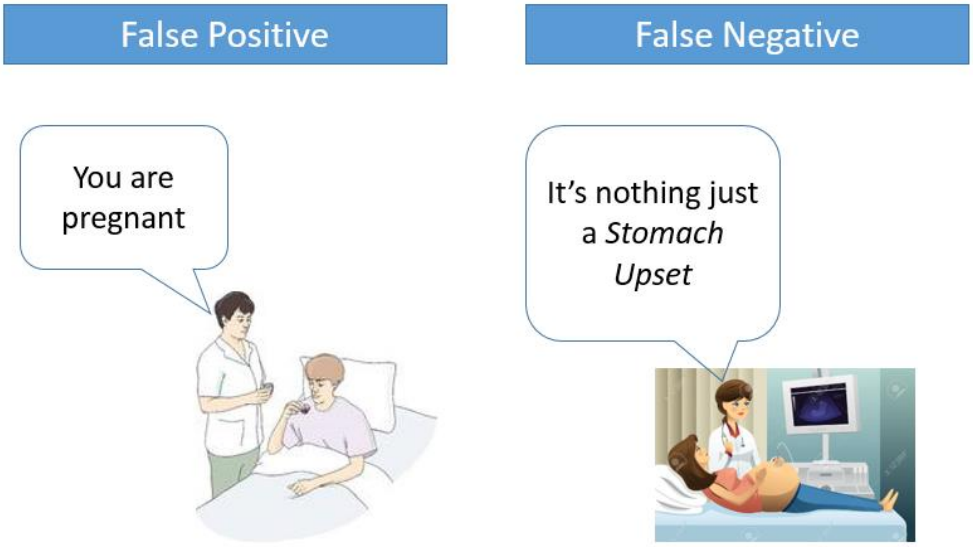
Seasonality makes your time series non-stationary because average value of the variables at different time periods. Differentiating a time series is generally known as the best method of removing seasonality from a time series. Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

34. Can you cite some examples where a false positive is important than a false negative?

Before we start, let us understand what false positives are and what false negatives are.

False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.

And, False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.



In medical field, assume you have to give chemo therapy to patients. Your lab tests patients for certain vital information and based on those results they decide to give radiation therapy to a patient.

Assume a patient comes to that hospital and he is tested positive for cancer (But he doesn't have cancer) based on lab prediction. What will happen to him? (Assuming Sensitivity is 1)

One more example might come from marketing. Let's say an ecommerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $5000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 5K.

Now what if they have sent it to false positive cases?

35. Can you cite some examples where a false negative important than a false positive?

Assume there is an airport 'A' which has received high security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to shortage of staff they decided to scan passenger being predicted as risk positives by their predictive model.

What will happen if a true threat customer is being flagged as non-threat by airport model?

Another example can be judicial system. What if Jury or judge decide to make a criminal go free?

What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after few years and realize that you had a false negative?

36. Can you cite some examples where both false positive and false negatives are equally important?

In the banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point of time they don't want to acquire bad customers. In this scenario both the false positives and false negatives be**come very important to measure.**

These days we hear many cases of players using steroids during sport competitions. Every player has to go through a steroid test before the game starts. A false positive can ruin the career of a Great sportsman and a false negative can make the game unfair.

37. Can you explain the difference between a Test Set and a Validation Set?

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid Overfitting of the model being built. On the other hand, test set is used for testing or evaluating the performance of a trained machine leaning model.

In simple terms, the differences can be summarized as-

- Training Set is to fit the parameters i.e. weights.
- Test Set is to assess the performance of the model i.e. evaluating the predictive power and generalization.
- Validation set is to tune the parameters.

38. What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF etc.). Sensitivity is nothing but "Predicted TRUE events/ Total events". True events here are the events which were true and model also predicted them as true.

Calculation of seasonality is pretty straight forward-

Seasonality = True Positives /Positives in Actual Dependent Variable

Where, True positives are Positive events which are correctly classified as Positives.

39. Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.

SVM and Random Forest are both used in classification problems.

a) If you are sure that your data is outlier free and clean then go for SVM. It is the opposite - if your data might contain outliers then Random forest would be the best choice.
b) Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for Random Forest machine learning algorithm.
c) Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose Random Forest machine learning algorithm.
d) Random Forest machine learning algorithms are preferred for multiclass problems.
e) SVM is preferred in multi-dimensional problem set - like text classification but as a good data scientist, you should experiment with both of them and test for accuracy or rather you can use ensemble of many Machine Learning techniques.

40. How do data management procedures like missing data handling make selection bias worse?

Missing value treatment is one of the primary tasks which a data scientist is supposed to do before starting data analysis. There are multiple methods for missing value treatment. If not done properly, it could potentially result into selection bias. Let see few missing value treatment examples and their impact on selection-

Complete Case Treatment: Complete case treatment is when you remove entire row in data even if one value is missing. You could achieve a selection bias if your values are

not missing at random and they have some pattern. Assume you are conducting a survey and few people didn't specify their gender. Would you remove all those people? Can't it tell a different story?

Available case analysis: Let say you are trying to calculate correlation matrix for data so you might remove the missing values from variables which are needed for that particular correlation coefficient. In this case your values will not be fully correct as they are coming from population sets.

Mean Substitution: In this method missing values are replaced with mean of other available values. This might make your distribution biased e.g., standard deviation, correlation and regression are mostly dependent on the mean value of variables.

Hence, various data management procedures might include selection bias in your data if not chosen correctly.

41. What are the basic assumptions to be made for linear regression?
Normality of error distribution, statistical independence of errors, linearity and additivity.

42. Can you write the formula to calculat R-square?

R-Square can be calculated using the below formula -

1 - (Residual Sum of Squares/ Total Sum of Squares)

43. What is the advantage of performing dimensionality reduction before fitting an SVM?
Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

44. How will you assess the statistical significance of an insight whether it is a real insight or just by chance?
Statistical importance of an insight can be accessed using Hypothesis Testing.

R Programming Data Science Interview questions and Answers

1. **Compare R and Python programming languages for Predictive Modelling.**

Python vs R for Predictive Modelling		
Feature	Python is Better	R Language is Better
Model Building	Both are Similar	Both are Similar
Model Interpretability	Not better than R.	R is better
Production	Python is Better	Not better than Python
Community Support	Not better than R.	R has good community support over Python.
Data Science Libraries	Both are similar.	Both are similar
Data Visualizations	Not better than R	R has good data visualizations libraries and tools.
Learning Curve	Learning Python is easier than learning R.	R has a steep learning curve.

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.

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

16. **What is the command used to store R objects in a file?**

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

17. **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.

18. **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.

19. **How is a Data object represented internally in R language?**

unclass (as.Date (“2016-10-05″))

20. **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.

21. **Which package in R supports the exploratory analysis of genomic data?**

adegenet

22. **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 can you add datasets in R?**

rbind () function can be used add datasets in R language provided the columns in the datasets should be same.

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. **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).

32. **How can you debug and test R programming code?**

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

33. **What will be the class of the resulting vector if you concatenate a number and a logical?**

number

34. **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}

35. **What happens if the application object is not able to handle an event?**

The event is dispatched to the delegate for processing.

36. **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.

37. **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)).

38. **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)
```

39. **How do you write R commands?**

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

40. **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.

41. **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.

42. **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.

43. **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.

44. **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.

45. **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)))

46. **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.

47. **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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. **Which function helps you perform sorting in R language?**

Order ()

54. **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))

55. **Which function is used to create a histogram visualisation in R programming language?**

Hist()

56. **Write the syntax to set the path for current working directory in R environment?**

Setwd("dir_path")

57. **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
```

58. **What will be the output of runif (7)?**

    It will generate 7 randowm numbers between 0 and 1.

59. **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 = )
    ```

60. **What will be the output on executing the following R programming code –**

    mat<-matrix(rep(c(TRUE,FALSE),8),nrow=4)

    sum(mat)

    8

61. **How will you combine multiple different string like “Data”, “Science”, “in” ,“R”, “Programming” as a single string “Data_Science_in_R_Programmming” ?**

    paste(“Data”, “Science”, “in” ,“R”, “Programming”,sep="_")

62. **Write a function to extract the first name from the string “Mr. Tom White”.**

    substr (“Mr. Tom White”,start=5, stop=7)

63. **Can you tell if the equation given below is linear or not?**

Emp_sal= 2000+2.5(emp_age)^2

Yes it is a linear equation as the coefficients are linear.

64. **What will be the output of the following R programming code?**

```
var2<- c("I","Love,"DeZyre")

var2
```

It will give an error.

65. **What will be the output of the following R programming code?**

```
x<-5

if(x%%2==0)

 print("X is an even number")

else

 print("X is an odd number")
```

Executing the above code will result in an error as shown below -

```
## Error: :4:1: unexpected 'else'

## 3:   print("X is an even number")

## 4: else

##    ^
```

R programming language does not know if the else related to the first 'if' or not as the first if() is a complete command on its own.

66. **I have a string "contact@dezyre.com". Which string function can be used to split the string into two different strings "contact@dezyre" and "com" ?**

This can be accomplished using the strsplit function which splits a string based on the identifier given in the function call. The output of strsplit() function is a list.

strsplit("contact@dezyre.com",split = ".")

Output of the strsplit function is -

[[1]]

[1] " contact@dezyre" "com"

67. What is R Base package?

R Base package is the package that is loaded by default whenever R programming environent is loaded .R base package provides basic fucntionalites in R environment like arithmetic calcualtions, input/output.

68. How will you merge two dataframes in R programming language?

Merge () function is used to combine two dataframes and it identifies common rows or columns between the 2 dataframes. Merge () function basically finds the intersection between two different sets of data.

Merge () function in R language takes a long list of arguments as follows –

Syntax for using Merge function in R language -

merge (x, y, by.x, by.y, all.x or all.y or all)

- X represents the first dataframe.
- Y represents the second dataframe.
- by.X- Variable name in dataframe X that is common in Y.
- by.Y- Variable name in dataframe Y that is common in X.
- all.x - It is a logical value that specifies the type of merge. all.X should be set to true, if we want all the observations from dataframe X . This results in Left Join.
- all.y - It is a logical value that specifies the type of merge. all.y should be set to true , if we want all the observations from dataframe Y . This results in Right Join.
- all – The default value for this is set to FALSE which means that only matching rows are returned resulting in Inner join. This should be set to true if you want all the observations from dataframe X and Y resulting in Outer join.

69. Write the R programming code for an array of words so that the output is displayed in decreasing frequency order.

R Programming Code to display output in decreasing frequency order -

```
tt <- sort(table(c("a", "b", "a", "a", "b", "c", "a1", "a1", "a1")),
dec=T)
depth <- 3
tt[1:depth]
```

Output -

```
1) a a1  b
2) 3  3  2
```

70. How to check the frequency distribution of a categorical variable?

The frequency distribution of a categorical variable can be checked using the table function in R language. Table () function calculates the count of each categories of a categorical variable.

gender=factor(c("M","F","M","F","F","F"))

table(sex)

Output of the above R Code –

Gender

F M

4 2

Programmers can also calculate the % of values for each categorical group by storing the output in a dataframe and applying the column percent function as shown below -

t = data.frame(table(gender))
t$percent= round(t$Freq / sum(t$Freq)*100,2)

Gender	Frequency	Percent
F	4	66.67

M	2	33.33

71. What is the procedure to check the cumulative frequency distribution of any categorical variable?

The cumulative frequency distribution of a categorical variable can be checked using the cumsum () function in R language.

Example –

```
gender = factor(c("f","m","m","f","m","f"))
y = table(gender)
cumsum(y)
```

Output of the above R code-

```
Cumsum(y)
f m
3 3
```

72. What will be the result of multiplying two vectors in R having different lengths?

The multiplication of the two vectors will be performed and the output will be displayed with a warning message like – "Longer object length is not a multiple of shorter object length." Suppose there is a vector a<-c (1, 2, 3) and vector b <- (2, 3) then the multiplication of the vectors a*b will give the resultant as 2 6 6 with the warning message. The multiplication is performed in a sequential manner but since the length is not same, the first element of the smaller vector b will be multiplied with the last element of the larger vector a.

73. R programming language has several packages for data science which are meant to solve a specific problem, how do you decide which one to use?

CRAN package repository in R has more than 6000 packages, so a data scientist needs to follow a well-defined process and criteria to select the right one for a specific task. When looking for a package in the CRAN repository a data scientist should list out all the requirements and issues so that an ideal R package can address all those needs and issues.

The best way to answer this question is to look for an R package that follows good software development principles and practices. For example, you might want to look at the quality documentation and unit tests. The next step is to check out how a particular R package is used and read the reviews posted by other users of the R package. It is important to know if other data scientists or data analysts have been able to solve a similar problem as that of yours. When you in doubt choosing a

particular R package, I would always ask for feedback from R community members or other colleagues to ensure that I am making the right choice.

74. **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.

75. **Explain the usage of which() function in R language.**

which() function determines the postion of elemnts in a logical vector that are TRUE. In the below example, we are finding the row number wherein the maximum value of variable v1 is recorded.

```
mydata=data.frame(v1 = c(2,4,12,3,6))
which(mydata$v1==max(mydata$v1))
```

It returns 3 as 12 is the maximum value and it is at 3rd row in the variable x=v1.

76. **How will you convert a factor variable to numeric in R language?**

A factor variable can be converted to numeric using the as.numeric() function in R language. However, the variable first needs to be converted to character before being converted to numberic because the as.numeric() function in R does not return original values but returns the vector of the levels of the factor variable.

```
X <- factor(c(4, 5, 6, 6, 4))
X1 = as.numeric(as.character(X))
```

Python Data Science Interview Questions and Answers

1. **Name a few libraries in Python used for Data Analysis and Scientific computations.**

   NumPy, SciPy, Pandas, SciKit, Matplotlib, Seaborn

2. **Which library would you prefer for plotting in Python language: Seaborn or Matplotlib?**

   Matplotlib is the python library used for plotting but it needs lot of fine-tuning to ensure that the plots look shiny. Seaborn helps data scientists create statistically and aesthetically appealing meaningful plots. The answer to this question varies based on the requirements for plotting data.

3. **Which method in pandas.tools.plotting is used to create scatter plot matrix?**

   Scatter_matrix

4. **How can you check if a data set or time series is Random?**

   To check whether a dataset is random or not use the lag plot. If the lag plot for the given dataset does not show any structure then it is random.

5. **What are the possible ways to load an array from a text data file in Python? How can the efficiency of the code to load data file be improved?**

   numpy.loadtxt ()

6. **Which is the standard data missing marker used in Pandas?**

   NaN

7. **Which Python library would you prefer to use for Data Munging?**

   Pandas

8. **Write the code to sort an array in NumPy by the nth column?**

Using argsort () function this can be achieved. If there is an array X and you would like to sort the nth column then code for this will be x[x [: n-1].argsort ()]

9. **Which python library is built on top of matplotlib and Pandas to ease data plotting?**

   Seaborn

10. **Which plot will you use to access the uncertainty of a statistic?**

    Bootstrap

11. **What is pylab?**

    A package that combines NumPy, SciPy and Matplotlib into a single namespace.

12. **Which python library is used for Machine Learning?**

    SciKit-Learn

13. **How can you copy objects in Python?**

    The functions used to copy objects in Python are-

    1) Copy.copy () for shallow copy

    2) Copy.deepcopy () for deep copy

    However, it is not possible to copy all objects in Python using these functions. For instance, dictionaries have a separate copy method whereas sequences in Python have to be copied by 'Slicing'.

14. **What is the difference between tuples and lists in Python?**

    Tuples can be used as keys for dictionaries i.e. they can be hashed. Lists are mutable whereas tuples are immutable - they cannot be changed. Tuples should be used when the order of elements in a sequence matters. For example, set of actions that need to be executed in sequence, geographic locations or list of points on a specific route.

15. **What is PEP8?**

    PEP8 consists of coding guidelines for Python language so that programmers can write readable code making it easy to use for any other person, later on.

16. **Is all the memory freed when Python exits?**

    No it is not, because the objects that are referenced from global namespaces of Python modules are not always de-allocated when Python exits.

17. **What does _init_.py do?**

    init_.py is an empty py file used for importing a module in a directory. _init_.py provides an easy way to organize the files. If there is a module maindir/subdir/module.py,_init_.py is placed in all the directories so that the module can be imported using the following command-

    import maindir.subdir.module

18. **What is the different between range () and xrange () functions in Python?**

    range () returns a list whereas xrange () returns an object that acts like an iterator for generating numbers on demand.

19. **How can you randomize the items of a list in place in Python?**

    Shuffle (lst) can be used for randomizing the items of a list in Python

20. **What is a pass in Python?**

    Pass in Python signifies a no operation statement indicating that nothing is to be done.

21. **If you are gives the first and last names of employees, which data type in Python will you use to store them?**

    You can use a list that has first name and last name included in an element or use Dictionary.

22. **What happens when you execute the statement mango=banana in Python?**

    A name error will occur when this statement is executed in Python.

23. **Optimize the below python code-**

    word = 'word'

    print word.__len_()

Answer: print 'word'._len_()

24. **What is monkey patching in Python?**

    Monkey patching is a technique that helps the programmer to modify or extend other code at runtime. Monkey patching comes handy in testing but it is not a good practice to use it in production environment as debugging the code could become difficult.

25. **Which tool in Python will you use to find bugs if any?**

    Pylint and Pychecker. Pylint verifies that a module satisfies all the coding standards or not. Pychecker is a static analysis tool that helps find out bugs in the course code.

26. **How are arguments passed in Python- by reference or by value?**

    The answer to this question is neither of these because passing semantics in Python are completely different. In all cases, Python passes arguments by value where all values are references to objects.

27. **You are given a list of N numbers. Create a single list comprehension in Python to create a new list that contains only those values which have even numbers from elements of the list at even indices. For instance if list[4] has an even value the it has be included in the new output list because it has an even index but if list[5] has an even value it should not be included in the list because it is not at an even index.**

    [x for x in list [1::2] if x%2 == 0]

    The above code will take all the numbers present at even indices and then discard the odd numbers.

28. **Explain the usage of decorators.**

    Decorators in Python are used to modify or inject code in functions or classes. Using decorators, you can wrap a class or function method call so that a piece of code can be executed before or after the execution of the original code. Decorators can be used to check for permissions, modify or track the arguments passed to a method, logging the calls to a specific method, etc.

29. **How can you check whether a pandas data frame is empty or not?**

    The attribute df.empty is used to check whether a data frame is empty or not.

30. **What will be the output of the below Python code –**

```
def multipliers ():

return [lambda x: i * x for i in range (4)]

print [m (2) for m in multipliers ()]
```

The output for the above code will be [6, 6,6,6]. The reason for this is that because of late binding the value of the variable i is looked up when any of the functions returned by multipliers are called.

31. **What do you mean by list comprehension?**

    The process of creating a list while performing some operation on the data so that it can be accessed using an iterator is referred to as List Comprehension.

Example:

```
[word (j) for j in string.ascii_uppercase]
```

[65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90]

32. **What will be the output of the below code**

    word = ‘aeioubcdfg'

    print word [:3] + word [3:]

    The output for the above code will be: ‘aeioubcdfg'.

    In string slicing when the indices of both the slices collide and a “+” operator is applied on the string it concatenates them.

33. **list= [‘a’,’e’,’i’,’o’,’u’]**

    print list [8:]

    The output for the above code will be an empty list []. Most of the people might confuse the answer with an index error because the code is attempting to access a member in the list whose index exceeds the total number of members in the list. The reason being the code is trying to access the slice of a list at a starting index which is greater than the number of members in the list.

34. **What will be the output of the below code :**

```
def foo (i= []):
    i.append (1)
    return i
>>> foo ()
>>> foo ()
```

The output for the above code will be-

[1]

[1, 1]

Argument to the function foo is evaluated only once when the function is defined. However, since it is a list, on every all the list is modified by appending a 1 to it.

35. **Can the lambda forms in Python contain statements?**

No, as their syntax is restricted to single expressions and they are used for creating function objects which are returned at runtime.