Lead Data Scientist Interview Questions And Answers Guide.

Global Guideline.

Lead Data Scientist Job Interview Preparation Guide.

Question # 1

Can you write the syntax to set the path for a current working directory in R environment?

Answer:-

Setwd("dir_path")
Different syntax can be asked in R Data science interview questions.

Read More Answers.

Question # 2

Do you know how to Merge the files into a single dataframe?

Answer:-

At last, we have to iterate the list of files in the current working directory. Also, we need to put them together to form a data frame. Moreover, when the script encounters the first file in the file_list, then it creates the main data frame to merge everything into. This is done using the !exists conditional:
If their dataset exists, then a temp_dataset called temporary data frame will be created and added to the dataset. Moreover, we have to delete temporary data frame. That is been removed when we're done with it using the rm(temp_dataset) command.
If dataset doesn't exist (!exists is true), then we have to create it.

Read More Answers.

Question # 3

Tell me how do you define big data?

Answer:-

It's likely that you'll be interviewed by an HR rep, an end business user, and an IT pro. Each person will probably ask you to explain what big data is, and how the data analysis discipline works with big data to produce insights.
You can start your answer with something fundamental, such as "big data analysis involves the collection and organization of data, and the ability to discover correlations between the data that provide revelations or insights that are actionable." You must be able to explain this in terms that resonate with each interviewer; the best way to do this is to illustrate the definition with an example.

Read More Answers.

Question # 4

Tell me what is bias, variance trade off?

Answer:-

Bias:
"Bias is error introduced in your model due to over simplification of machine learning algorithm." It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.
Low bias machine learning algorithms - Decision Trees, k-NN and SVM
Hight bias machine learning algorithms - Liear Regression, Logistic Regression
Variance:
"Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training dataset and performs bad on test dataset." It can lead high sensitivity and overfitting.

Read More Answers.

Question # 5

Tell me what is Random Forest? How does it work?

Answer:-

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimentionality reduction, treats missing values, outlier values. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.
In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes(Over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

Read More Answers.

Question # 6

Do you know what regularization is and why it is useful?

Answer:-

Regularization is the process of adding tunning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

Read More Answers.

Question # 7

Explain me what are feature vectors?

Answer:-

A feature vector is an n-dimensional vector of numerical features that represent some object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way.

Read More Answers.

Question # 8

Please explain me cross-validation?

Answer:-

It is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice. The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and gain insight on how the model will generalize to an independent data set.

Read More Answers.

Question # 9

Tell us what are confounding variables?

Answer:-

These are extraneous variables in a statistical model that correlate directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.

Read More Answers.

Question # 10

Can you please explain selective bias?

Answer:-

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.

Read More Answers.

Question # 11

Tell us what tools or devices help you succeed in your role as a data scientist?

Answer:-

This question's purpose is to learn the programming languages and applications the candidate knows and has experience using. The answer will show the candidate's need for additional training of basic programming languages and platforms or any transferable skills. This is vital to understand as it can cost more time and money to train if the candidate is not knowledgeable in all of the languages and applications required for the position. Answers to look for include:
* Experience in SAS and R programming
* Understanding of Python, PHP or Java programming languages
* Experience using data visualization tools

Read More Answers.

Question # 12

Tell us how would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression?

Answer:-

Proposed methods for model validation:
* If the values predicted by the model are far outside of the response variable range, this would immediately indicate poor estimation or model inaccuracy.
* If the values seem to be reasonable, examine the parameters; any of the following would indicate poor estimation or multi-collinearity: opposite signs of expectations, unusually large or small values, or observed inconsistency when the model is fed new data.
* Use the model for prediction by feeding it new data, and use the coefficient of determination (R squared) as a model validity measure.
* Use data splitting to form a separate dataset for estimating model parameters, and another for validating predictions.
* Use jackknife resampling if the dataset contains a small number of instances, and measure validity with R squared and mean squared error (MSE).

Read More Answers.

Question # 13

Do you know what is the difference between rnorm and runif functions?

Answer:-

rnorm function-
Basically, it generates "n" normal random numbers. That is totally based on the mean and standard deviation arguments passed to the function.

Syntax of rnorm function -
rnorm(n, mean = , sd =)
runif function-
Basically, it generates "n" unform random numbers in the interval. That is of minimum and maximum values passed to the function.
Syntax of runif function -
runif(n, min = , max =)

Read More Answers.

Question # 14

Explain me for loop control statement in R?

Answer:-

A loop is a sequence of instructions that is repeated until a certain condition is been reached. for, while and repeat, with the additional clauses break and next are used to construct loops.
For Example-
It is executed a known number of times for a block is been contained within curly braces.
x = c(1,2,3,4,5)
for(i in 1:5){
print(x[i])
}
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5

Read More Answers.

Question # 15

Tell me you develop a big data model, but your end user has difficulty understanding how the model works and the insights it can reveal. How do you communicate with the user to get your points across?

Answer:-

Many big data analysts come from statistics, engineering, and computer science disciplines; they're brilliant analysts, but their people and communications skills lag. Businesses understand that to obtain results, you need both strong execution and strong communication. You can expect your HR, end business, and IT interviewers to focus on your communications skills, and to try to test them with a hypothetical situation.

Read More Answers.

Question # 16

Tell me what cross-validation technique would you use on a time series dataset?

Answer:-

Instead of using k-fold cross-validation, you should be aware to the fact that a time series is not randomly distributed data - It is inherently ordered by chronological order.
In case of time series data, you should use techniques like forward chaining - Where you will be model on past data then look at forward-facing data.
fold 1: training[1], test[2]
fold 1: training[1 2], test[3]
fold 1: training[1 2 3], test[4]
fold 1: training[1 2 3 4], test[5]

Read More Answers.

Question # 17

Tell us why is resampling done?

Answer:-

Resampling is done in any of these cases:
* Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
* Substituting labels on data points when performing significance tests
* Validating models by using random subsets (bootstrapping, cross validation)

Read More Answers.

Question # 18

Tell me what methods do you use to identify outliers within a data set?

Answer:-

Data scientists must be able to go beyond classroom theoretical applications to real-world applications. Your candidate's answer to this question will show how they allocate their time to finding the best way to detect outliers. This information is important to know because it demonstrates the candidate's analytical skills. Look for answers that include:
* Raw data analysis
* Models
* Approaches

Read More Answers.

Question # 19

Can you describe strsplit() in R string manipulation?

Answer:-

Keywords
Character
Usage
strsplit(x, split, fixed = FALSE, perl = FALSE, useBytes = FALSE)
Arguments
a. x
It is a character vector, each element of which is to be split.
b. split
Basically, it is a character vector containing regular expression(s). That is used for splitting.
c. fixed
Since it is TRUE then it will match split exactly.
d. perl
Should Perl-compatible regexps be used?
e. useBytes
It is TRUE then the matching will do byte-by-byte rather than character-by-character, and inputs with marked encodings are not converted.

Read More Answers.

Question # 20

Explain me nchar() in R string manipulation?

Answer:-

To find out if elements of a character vector are non-empty strings or not then nzchar is the fastest way.
Keywords
character
Usage
nchar(x, type = "chars", allowNA = FALSE, keepNA = NA)
nzchar(x, keepNA = FALSE)
Arguments
a. x
Basically, a character vector or a vector will be restricted to a character vector. Giving a factor is an error.
b. type
character string: partial matching to one of c("bytes", "chars", "width").
c. allowNA
Should NA will return for invalid multibyte strings or "bytes"-encoded strings
d. keepNA
The default for nchar(), NA, means to use keepNA = TRUE unless type is "width". Used to be hardcoded to FALSE in R versions <= 3.2.0.

Read More Answers.

Question # 21

Tell me how do clean up and organize big data sets?

Answer:-

Data scientists frequently have to combine large amounts of information from various devices in several formats, such as data from a smartwatch or cellphone.
Answers to this question will demonstrate how your candidate's methods for organizing large data. This information is important to know because data scientists need
clean data to analyze information accurately to offer recommendations that solve business problems. Possible answers may include:
* Automation tools
* Value correction methods
* Comprehension of data sets

Read More Answers.

Question # 22

Do you know more functions in brief in R?

Answer:-

read.spss Function - read.spss
What it does - Reads spss data file
For Example- spss("myfile")
read.xport Function - read.xport
What it does - Reads SAS export file
For Example- export("myfile")
read.dta Function - read.dta
What it does - Reads stata binary file
For Example - read.dta("myfile")

Read More Answers.

Question # 23

Tell me what is exploding gradients?

Answer:-

"Exploding gradients are a problem where large error gradients accumulate and result in very large updates to neural network model weights during training." At an
extreme, the values of weights can become so large as to overflow and result in NaN values.
This has the effect of your model being unstable and unable to learn from your training data. Now let's understand what is the gradient.
Gradient:
Gradient is the direction and magnitude calculated during training of a neural network that is used to update the network weights in the right direction and by the right
amount.

Question # 24

Can you explain me what is logistic regression? Or State an example when you have used logistic regression recently?

Answer:-

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.

Question # 25

What are the steps in making a decision tree?

Answer:-

* Take the entire data set as input.
* Look for a split that maximizes the separation of the classes. A split is any test that divides the data into two sets.
* Apply the split to the input data (divide step).
* Re-apply steps 1 to 2 to the divided data.
* Stop when you meet some stopping criteria.
* This step is called pruning. Clean up the tree if you went too far doing splits.

Question # 26

Explain me what is the difference between machine learning and deep learning?

Answer:-

* Machine learning:
Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorized in following three categories.
* Supervised machine learning,
* Unsupervised machine learning,
* Reinforcement learning
* Deep learning:
Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

Question # 27

Tell me how do you overcome challenges to your findings?

Answer:-

The reason for asking this question is to discover how well the candidate approaches solving conflicts in a team environment. Their answer shows the candidate's problem-solving and interpersonal skills in stressful situations. Understanding these skills is significant because group dynamics and business conditions change. Consider answers that:
* Encourage discussion
* Demonstrate leadership
* Acknowledges recognizing and respecting different opinions

Question # 28

Tell me how to create a function in arguments using apply() in R?

Answer:-

What if we want to be able to find how many data points (n) are in each column of m?
We are using columns, MARGIN = 2, thus, we can use length function to do this:
apply(my.matrx, 2, length)
There isn't a function in R to find n-1 for each column. So if we want to, we have to create our own Function. Since the function is simple, you can create it right inside the arguments for applying. In the arguments, I created a function that returns length - 1.
apply(my.matrx, 2, function (x) length(x)-1)
The function returned a vector of n-1 for each column.

Question # 29

Explain me what is TF/IDF vectorization?

Answer:-

tf-idf is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Question # 30

Tell me what are the drawbacks of the linear model?

Answer:-

Some drawbacks of the linear model are:
* The assumption of linearity of the errors.
* It can't be used for count outcomes or binary outcomes
* There are overfitting problems that it can't solve

Read More Answers.

Question # 31

Tell us how do you work towards a random forest?

Answer:-

The underlying principle of this technique is that several weak learners combined to provide a strong learner. The steps involved are
* Build several decision trees on bootstrapped training samples of data
* On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
* Rule of thumb: At each split m=pYm=p
* Predictions: At the majority rule

Read More Answers.

Question # 32

Tell me how has your prior experience prepared you for a role in data science?

Answer:-

This question helps determine the candidate's experience from a holistic perspective and reveals experience in demonstrating interpersonal, communication and technical skills. It is important to understand this because data scientists must be able to communicate their findings, work in a team environment and have the skills to perform the task. Here are some possible answers to look for:
* Project management skills
* Examples of working in a team environment
* Ability to identify errors

Read More Answers.

Question # 33

Tell me which data object in R is used to store and process categorical data?

Answer:-

It seems like the Factor data objects in R are used to store and process categorical data in R

Read More Answers.

Question # 34

Tell me why do you want to work at this company as a data scientist?

Answer:-

The purpose of this question is to determine the motivation behind the candidate's choice of applying and interviewing for the position. Their answer should reveal their inspiration for working for the company and their drive for being a data scientist. It should show the candidate is pursuing the position because they are passionate about data and believe in the company, two elements that can determine the candidate's performance. Answers to look for include:
* Interest in data mining
* Respect for the company's innovative practices
* Desire to apply analytical skills to solve real-world issues with data

Read More Answers.

Question # 35

Tell me how can you produce co-relations and covariances?

Answer:-

Since Co-relations is produced by cor() and covariances are produced by cov() function, we need to use them.

Read More Answers.

Question # 36

Tell me what is selection bias?

Answer:-

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed. It is sometimes referred to as the selection effect. The phrase "selection bias" most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

Read More Answers.

Question # 37

Do you know what are the types of biases that can occur during sampling?

Answer:-

* Selection bias

* Under coverage bias
* Survivorship bias

Question # 38

Please explain what does not 'R' language do?

Answer:-

• Since R is open source language but still it does not consist of any graphical user interface.
• Also, it easily connects to Excel/Microsoft Office easily. Although, it does not provide any spreadsheet view of data.

Question # 39

Explain me a big data project you worked on?

Answer:-

Companies understand that they have to train and orient you to their business and technical environments, but they also expect you to bring skills, experience, and fresh ideas to the job.
The end business user and the IT interviewer will be especially interested in your project work. For the IT person, be sure to go into the data quality, analysis, publication, and actionalization processes, covering both the end business and the technical enablement details. For the end business person, review the project from a business results perspective, but avoid using technical jargon unless asked.

Question # 40

Tell me what is the difference between supervised and unsupervised machine learning?

Answer:-

Supervised Machine learning:
Supervised machine learning requires training labeled data.
Unsupervised Machine learning:
Unsupervised machine learning doesn't required labeled data.

Question # 41

Do you know how many data structures does R language have?

Answer:-

It has two data structures namely:
Homogeneous data structures-
It contains the same type of objects - Vector, Matrix, and Array.
Heterogeneous data structures-
It contains a different type of objects - Data frames and lists.

Question # 42

Tell me what is the difference between Regression and classification ML techniques?

Answer:-

Both Regression and classification machine learning techniques come under Supervised machine learning algorithms. In Supervised machine learning algorithm, we have to train the model using labeled dataset, While training we have to explicitly provide the correct labels and algorithm tries to learn the pattern from input to output. If our labels are discrete values then it will a classification problem, e.g A,B etc. but if our labels are continuous values then it will be a regression problem, e.g 1.23, 1.333 etc.

Question # 43

Tell me what is logistic regression?

Answer:-

Logistic Regression is also known as the logit model. It is a technique to forecast the binary outcome from a linear combination of predictor variables.

Question # 44

Explain me what is the goal of A/B Testing?

Answer:-

This is a statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.

Question # 45

Tell me how do you identify a barrier to performance?

Answer:-

This question will determine how the candidate approaches solving real-world issues they will face in their role as a data scientist. It will also determine how they approach problem-solving from an analytical standpoint. This information is vital to understand because data scientists must have strong analytical and problem-solving skills. Look for answers that reveal:
* Examples of problem-solving methods
* Steps to take to identify the barriers to performance
* Benchmarks for assessing performance

Read More Answers.

Question # 46

Tell me how regularly must an algorithm be updated?

Answer:-

You will want to update an algorithm when:
* You want the model to evolve as data streams through infrastructure
* The underlying data source is changing
* There is a case of non-stationarity

Read More Answers.

Question # 47

What are the different kernels functions in SVM?

Answer:-

There are four types of kernels in SVM.
* Linear Kernel
* Polynomial kernel
* Radial basis kernel
* Sigmoid kernel

Read More Answers.

Question # 48

Tell me what challenges have you encountered while working with big data?

Answer:-

Big data doesn't always work as advertised, which is why your IT interviewer will likely probe you about big data setbacks or limits that you've encountered, and ask how you worked through them. Be prepared to answer this question in a straightforward, factual manner, and cap your answer with a discussion of what you gained from the experience and how it benefits you now.

Read More Answers.

Question # 49

Do you know what is pruning in Decision Tree?

Answer:-

When we remove sub-nodes of a decision node, this procsss is called pruning or opposite process of splitting.

Read More Answers.

Question # 50

Do you know what is root cause analysis?

Answer:-

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from reoccurring.

Read More Answers.

Question # 51

Do you know what is the Law of Large Numbers?

Answer:-

It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample mean, the sample variance and the sample standard deviation converge to what they are trying to estimate.

Read More Answers.

Question # 52

Can you please explain survivorship bias?

Answer:-

It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

Read More Answers.

Question # 53

Tell me do gradient descent methods at all times converge to a similar point?

Answer:-

No, they do not because in some cases they reach a local minima or a local optima point. You would not reach the global optima point. This is governed by the data and the starting conditions.

Question # 54

Do you know what is selection Bias?

Answer:-

Selection bias occurs when sample obtained is not represantative of the population intended to be analyzed.

Question # 55

Tell us what are Eigenvalue and Eigenvector?

Answer:-

Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

Question # 56

Please explain what is Collaborative Filtering?

Answer:-

The process of filtering used by most recommender systems to find patterns and information by collaborating perspectives, numerous data sources, and several agents.

Question # 57

Explain me what are your technical competencies?

Answer:-

Before the interview, do your homework on the analytics environment that the interviewing company uses. During the IT interview, you will be asked to review your technical competencies and skillsets. How well the company feels your technical skills fit with the data analytics approaches and tools they use in their environment can have a make-or-break effect on whether you get the job.

Question # 58

Tell us what are Recommender Systems?

Answer:-

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.

Question # 59

Do you know what are Recommender Systems?

Answer:-

Recommender systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product.

Question # 60

Please explain star schema?

Answer:-

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

Data Warehouse Most Popular Interview Topics.

1 : [Data Warehousing Frequently Asked Interview Questions and Answers Guide.](#)

2 : [Ab Initio Frequently Asked Interview Questions and Answers Guide.](#)

3 : [Informatica Data Warehousing Frequently Asked Interview Questions and Answers Guide.](#)

4 : [ColdFusion Frequently Asked Interview Questions and Answers Guide.](#)

5 : [Cognos Data Warehouse Frequently Asked Interview Questions and Answers Guide.](#)

6 : [Business Objects Frequently Asked Interview Questions and Answers Guide.](#)

7 : [Micro Strategy Frequently Asked Interview Questions and Answers Guide.](#)

8 : [Warehouse Assistant Frequently Asked Interview Questions and Answers Guide.](#)

9 : [Extract, transform, load (ETL) Frequently Asked Interview Questions and Answers Guide.](#)

10 : [BO Designer Frequently Asked Interview Questions and Answers Guide.](#)

About Global Guideline.

Global Guideline is a platform to develop your own skills with thousands of job interview questions and web tutorials for fresher's and experienced candidates. These interview questions and web tutorials will help you strengthen your technical skills, prepare for the interviews and quickly revise the concepts. Global Guideline invite you to unlock your potentials with thousands of **Interview Questions with Answers** and much more. Learn the most common technologies at Global Guideline. We will help you to explore the resources of the World Wide Web and develop your own skills from the basics to the advanced. Here you will learn anything quite easily and you will really enjoy while learning. Global Guideline will help you to become a professional and Expert, well prepared for the future.

* This PDF was generated from https://GlobalGuideline.com at **November 29th, 2023**

* If any answer or question is incorrect or inappropriate or you have correct answer or you found any problem in this document then don't hesitate feel free and e-mail us we will fix it.

You can follow us on FaceBook for latest Jobs, Updates and other interviews material.
 www.facebook.com/InterviewQuestionsAnswers

Follow us on Twitter for latest Jobs and interview preparation guides
https://twitter.com/InterviewGuide

Best Of Luck.

Global Guideline Team
https://GlobalGuideline.com
Info@globalguideline.com