Data Scientist Job Interview Questions And Answers

Interview Questions Answers

https://interviewquestionsanswers.org/

About Interview Questions Answers

Interview Questions Answers . ORG is an interview preparation guide of thousands of Job Interview Questions And Answers, Job Interviews are always stressful even for job seekers who have gone on countless interviews. The best way to reduce the stress is to be prepared for your job interview. Take the time to review the standard interview questions you will most likely be asked. These interview questions and answers on Data Scientist will help you strengthen your technical skills, prepare for the interviews and quickly revise the concepts.

If you find any **question or answer** is incorrect or incomplete then you can **submit your question or answer** directly with out any registration or login at our website. You just need to visit Data Scientist Interview Questions And Answers to add your answer click on the *Submit Your Answer* links on the website; with each question to post your answer, if you want to ask any question then you will have a link *Submit Your Question*; that's will add your question in Data Scientist category. To ensure quality, each submission is checked by our team, before it becomes live. This Data Scientist Interview preparation PDF was generated at **Wednesday 29th November, 2023**

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

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

If you need any further assistance or have queries regarding this document or its material or any of other inquiry, please do not hesitate to contact us.

Best Of Luck.

Interview Questions Answers.ORG Team
https://InterviewQuestionsAnswers.ORG/
Support@InterviewQuestionsAnswers.ORG

Data Scientist Interview Questions And Answers Guide.

Question - 1:

What is survivorship bias?

Ans:

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.

View All Answers

Question - 2:

Tell us what is Collaborative Filtering?

Ans:

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

View All Answers

Question - 3:

Explain me what is Interpolation and Extrapolation?

Ans:

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.

View All Answers

Question - 4:

Tell me what are Recommender Systems?

Ans:

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.

View All Answers

Question - 5:

Do you know what are confounding variables?

Ans:

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.

View All Answers

Question - 6:

Please explain what are Recommender Systems?

Ans:

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.

View All Answers

Question - 7:

Tell me what are Eigenvalue and Eigenvector?

Ans:

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.

View All Answers

Question - 8:

Tell me what is Collaborative filtering?

Ans:

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.

View All Answers

Question - 9:

Tell me Python or R - Which one would you prefer for text analytics?

Ans:

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.

View All Answers

Question - 10:

Tell me what is the Law of Large Numbers?

Ans:

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.

View All Answers

Question - 11:

Do you know what are feature vectors?

Ans:

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.

View All Answers

Question - 12:

Tell me what are the types of biases that can occur during sampling?

Ans:

* Selection bias
* Under coverage bias
* Survivorship bias

View All Answers

Question - 13:

Tell me what is Linear Regression?

Ans:

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.

View All Answers

Question - 14:

Tell me do gradient descent methods always converge to same point?

Ans:

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

View All Answers

Question - 15:

What is selective bias?

Ans:

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

View All Answers

Question - 16:

Explain me what makes CNNs translation invariant?

Ans:

As explained above, each convolution kernel acts as it's own filter/feature detector. So let's say you're doing object detection, it doesn't matter where in the image the object is since we're going to apply the convolution in a sliding window fashion across the entire image anyways.

View All Answers

Question - 17:

Please explain how do you overcome challenges to your findings?

Ans:

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

View All Answers

Question - 18:

Tell me which technique is used to predict categorical responses?

Ans:

Classification technique is used widely in mining for classifying data sets.

View All Answers

Question - 19:

Tell me how is True Positive Rate and Recall related?

Ans:

True Positive Rate = Recall. Yes, they are equal having the formula (TP/TP + FN).

View All Answers

Question - 20:

Tell me how do you know which Machine Learning model you should use?

Ans:

While one should always keep the "no free lunch theorem" in mind, there are some general guidelines.

View All Answers

Question - 21:

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

Ans:

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

View All Answers

Question - 22:

Tell us are expected value and mean value different?

Ans:

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.

View All Answers

Question - 23:

Tell me what is power analysis?

Ans:

An experimental design technique for determining the effect of a given sample size.

View All Answers

Question - 24:

Explain me when is Ridge regression favorable over Lasso regression?

Ans:

You can quote ISLR's authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.
Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.

View All Answers

Question - 25:

Tell me how do you work towards a random forest?

Ans:

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

View All Answers

Question - 26:

Tell us what are the drawbacks of the linear model?

Ans:

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

View All Answers

Question - 27:

Tell us what is root cause analysis?

Ans:

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.

View All Answers

Question - 28:

Tell us what is the significance of Residual Networks?

Ans:

The main thing that residual connections did was allow for direct feature access from previous layers. This makes information propagation throughout the network much easier. One very interesting paper about this shows how using local skip connections gives the network a type of ensemble multi-path structure, giving features multiple paths to propagate throughout the network.

View All Answers

Question - 29:

What is cross-validation?

Ans:

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.

View All Answers

Question - 30:

Do you know the steps in making a decision tree?

Ans:

* 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.

View All Answers

Question - 31:

Tell me why do segmentation CNNs typically have an encoder-decoder style / structure?

Ans:

The encoder CNN can basically be thought of as a feature extraction network, while the decoder uses that information to predict the image segments by "decoding" the features and upscaling to the original image size.

View All Answers

Question - 32:

Tell us how would you go about doing an Exploratory Data Analysis (EDA)?

Ans:

The goal of an EDA is to gather some insights from the data before applying your predictive model i.e gain some information. Basically, you want to do your EDA in a coarse to fine manner.
We start by gaining some high-level global insights. Check out some imbalanced classes. Look at mean and variance of each class. Check out the first few rows to see what it's all about. Run a pandas df.info() to see which features are continuous, categorical, their type (int, float, string).
Next, drop unnecessary columns that won't be useful in analysis and prediction. These can simply be columns that look useless, one's where many rows have the same value (i.e it doesn't give us much information), or it's missing a lot of values. We can also fill in missing values with the most common value in that column, or the median. Now we can start making some basic visualizations. Start with high-level stuff. Do some bar plots for features that are categorical and have a small number of groups. Bar plots of the final classes. Look at the most "general features".
Create some visualizations about these individual features to try and gain some basic insights. Now we can start to get more specific.
Create visualizations between features, two or three at a time. How are features related to each other? You can also do a PCA to see which features contain the most information. Group some features together as well to see their relationships. For example, what happens to the classes when A = 0 and B = 0? How about A = 1 and B = 0? Compare different features. For example, if feature A can be either "Female" or "Male" then we can plot feature A against which cabin they stayed in to see if Males and Females stay in different cabins.
Beyond bar, scatter, and other basic plots, we can do a PDF/CDF, overlayed plots, etc. Look at some statistics like distribution, p-value, etc. Finally it's time to build the ML model. Start with easier stuff like Naive Bayes and Linear Regression. If you see that those suck or the data is highly non-linear, go with polynomial regression, decision trees, or SVMs. The features can be selected based on their importance from the EDA. If you have lots of data you can use a Neural Network. Check ROC curve. Precision, Recall

View All Answers

Question - 33:

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

Ans:

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

View All Answers

Question - 34:

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

Ans:

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.

View All Answers

Question - 35:

Tell me is rotation necessary in PCA? If yes, Why? What will happen if you don't rotate the components?

Ans:

Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that's the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn't change, it only changes the actual coordinates of the points.
If we don't rotate the components, the effect of PCA will diminish and we'll have to select more number of components to explain variance in the data set.

View All Answers

Question - 36:

What is star schema?

Ans:

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.

View All Answers

Question - 37:

Explain me what tools or devices help you succeed in your role as a data scientist?

Ans:

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
"I believe I can excel in this position with my R, Python, and SQL programming skill set. I enjoy working on the FUSE and Tableau platforms to mine data and draw inferences."

View All Answers

Question - 38:

What is dimensionality reduction, where it's used, and it's benefits?

Ans:

Dimensionality reduction is the process of reducing the number of feature variables under consideration by obtaining a set of principal variables which are basically the important features. Importance of a feature depends on how much the feature variable contributes to the information representation of the data and depends on which technique you decide to use. Deciding which technique to use comes down to trial-and-error and preference. It's common to start with a linear technique and move to non-linear techniques when results suggest inadequate fit.
Benefits of dimensionality reduction for a data set may be:
(1) Reduce the storage space needed
(2) Speed up computation (for example in machine learning algorithms), less dimensions mean less computing, also less dimensions can allow usage of algorithms unfit for a large number of dimensions
(3) Remove redundant features, for example no point in storing a terrain's size in both sq meters and sq miles (maybe data gathering was flawed)
(4) Reducing a data's dimension to 2D or 3D may allow us to plot and visualize it, maybe observe patterns, give us insights
(5) Too many features or too complex a model can lead to overfitting.

View All Answers

Question - 39:

Do you know what is the goal of A/B Testing?

Ans:

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.

View All Answers

Question - 40:

Explain me why data cleaning plays a vital role in analysis?

Ans:

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.

View All Answers

Question - 41:

Do you know why is naive Bayes so 'naive' ?

Ans:

naive Bayes is so 'naive' because it assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.

View All Answers

Question - 42:

Tell us how regularly must an algorithm be updated?

Ans:

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

View All Answers

Question - 43:

Do you know what is logistic regression?

Ans:

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.

View All Answers

Question - 44:

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

Ans:

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

View All Answers

Question - 45:

Tell us why do we use convolutions for images rather than just FC layers?

Ans:

This one was pretty interesting since it's not something companies usually ask. As you would expect, I got this question from a company focused on Computer Vision. This answer has 2 parts to it. Firstly, convolutions preserve, encode, and actually use the spatial information from the image. If we used only FC layers we would have no relative spatial information. Secondly, Convolutional Neural Networks (CNNs) have a partially built-in translation in-variance, since each convolution kernel acts as it's own filter/feature detector.

View All Answers

Question - 46:

Explain me what is data normalization and why do we need it?

Ans:

I felt this one would be important to highlight. Data normalization is very important preprocessing step, used to rescale values to fit in a specific range to assure better convergence during backpropagation. In general, it boils down to subtracting the mean of each data point and dividing by its standard deviation. If we don't do this then some of the features (those with high magnitude) will be weighted more in the cost function (if a higher-magnitude feature changes by 1%, then that change is pretty big, but for smaller features it's quite insignificant). The data normalization makes all features weighted equally.

View All Answers

Question - 47:

Tell me how can outlier values be treated?

Ans:

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 -
* 1) To change the value and bring in within a range
* 2) To just remove the value.

View All Answers

Question - 48:

Can you differentiate between univariate, bivariate and multivariate analysis?

Ans:

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, analysing 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.

View All Answers

Question - 49:

Tell me how is kNN different from kmeans clustering?

Ans:

Don't get mislead by 'k' in their names. You should know that the fundamental difference between both these algorithms is, kmeans is unsupervised in nature and kNN is supervised in nature. kmeans is a clustering algorithm. kNN is a classification (or regression) algorithm.
kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

View All Answers

Question - 50:

Tell me why is resampling done?

Ans:

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)

View All Answers

Question - 51:

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

Ans:

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.

View All Answers

Question - 52:

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

Ans:

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

View All Answers

Question - 53:

Tell us how do you identify a barrier to performance?

Ans:

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
"My approach to determining performance bottlenecks is to conduct a performance test. I then evaluate the performance based on criteria set by the lead data scientist or company and discuss my findings with my team lead and group."

View All Answers

Question - 54:

Tell us why do we have max-pooling in classification CNNs?

Ans:

Again as you would expect this is for a role in Computer Vision. Max-pooling in a CNN allows you to reduce computation since your feature maps are smaller after the pooling. You don't lose too much semantic information since you're taking the maximum activation. There's also a theory that max-pooling contributes a bit to giving CNNs more translation in-variance.

View All Answers

Question - 55:

Tell me how do you handle missing or corrupted data in a dataset?

Ans:

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.
In Pandas, there are two very useful methods:
isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

View All Answers

Data Warehouse Most Popular & Related Interview Guides

1 : **Warehouse Assistant Interview Questions and Answers.**

2 : Ab Initio Interview Questions and Answers.

3 : Data Stage Interview Questions and Answers.

4 : ColdFusion Interview Questions and Answers.

5 : Data Warehouse Supervisor Interview Questions and Answers.

6 : Informatica Data Warehousing Interview Questions and Answers.

7 : Data Warehousing Interview Questions and Answers.

8 : Data Warehouse BI Interview Questions and Answers.

9 : ETL (Extract, transform, load) Interview Questions and Answers.

10 : Micro Strategy Interview Questions and Answers.

Follow us on FaceBook
www.facebook.com/InterviewQuestionsAnswers.Org

Follow us on Twitter
https://twitter.com/InterviewQA

For any inquiry please do not hesitate to contact us.

Interview Questions Answers.ORG Team
https://InterviewQuestionsAnswers.ORG/
support@InterviewQuestionsAnswers.ORG