101 Data Science Interview Questions shared by Ravit Jain

By this Data Science Interview Questions and answers, many students are got placed in many reputed companies with high package salaries. So utilize our Data Science Interview Questions and answers to grow in your career.

Q1) What are the types of machine learning?

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement Learning**

Q2) What is the Supervised learning in machine learning?

þÿSupervised learning When you know your target variable for the probl
Supervised learning. This can be applied to perform regression and classification.

Example: Linear Regression and Logistic Regression.

Q3) What is the Unsupervised learning in machine learning?

þÿUnsupervised learning When you do not know your target variable
becomes Unsupervised learning. This is widely used to perform Clustering.

Example: K-Means and Hierarchical clustering.

Q4) What are the commonly used python packages?

- **Numpy**
- **Pandas**
- **SCI-KIT Learn**
- **Matplot library**

Q5) What are the commonly used R packages?

- **Caret**
- **Data.Table**

- Reshape
- Reshape2
- E1071
- DMwR
- Dplyr
- Lubridate

Q6) Name the commonly used algorithms.

- Linear regression

Type

- Logistic regression
- Random Forest
- KNN

Q7) What is precision?

The ration of predicted positive against the actual positive.

It is the most commonly used error metric is n classification mechanism.

The range is from 0 to 1, where 1 represents 100%.

Q8) What is recall?

The ratio of the true positive rate against the actual positive rate.

The range is again from 0 to 1

Q9) Which metric acts like accuracy in classification problem statement?

þÿF1 Score 2 * (Precision*Recall)/Precision + Recall

Q10) What is a normal distribution?

When the data distribution is equally distributed as such the mean, median and mode are equal.

Q11) What is overfitting?

Any prediction rate which has high inconsistency between the training error and the test error leads ta a high business problem, if the error rate in training set is low and the error rate ithe n test set is high, then we can conclude it as overfitting model.

Q12) What is underfitting?

Any prediction rate which has provides low prediction in the training error and the test error leads to a high business problem, if the error rate in training set is high and the error rate inthe test set is also high, then we can conclude it as overfitting model.

Q13) What is a univariate analysis?

An Analysis that can be applied to one attribute at a time is called as a univariate analysis.

Boxplot is one of the widely used univariate model.

þÿScatter plot and cook s distance are other methods used for bivariate a

Q14) Name few methods for Missing Value Treatments.

þÿCentral Imputation This method acts more like central tendencies. Al
with mean and median mode respective to numerical and categorical datatypes.

þÿKNN K Nearest Neighbour imputation.

þÿDistance between two or multiple attributes are calculated using Eucli
be used to treat the missing values. Mean and mode will agaibe n used as in CI.

Q15) What is the Pearson correlation?

Correlation between predicted and actual data can be examined and understood using this method.

The range is from -1 to +1.

-1 refers to negative 100% whereas +1 refers to positive 100%.

The formula is Sd(x)*m/Sd.(y)

Q16) How and by what methods data visualizations can be effectively used?

In addition to giving insights in a very effective and efficient manner, data visualization can also be used in such a way that it is not only restricted to bar, line or some stereotypic graphs. Data can be represented in a much more visually pleasing manner.

One thing have to be taken care of is to convey the intended insight or finding correctly to the audience. Once the baseline is set. Innovative and creative part can help you come up with better looking and functional dashboards. There is a fine line between the simple insightful dashboard and awesome looking 0 fruitful insight dashboards.

Q17) How to understand the problems faced during data analysis?

Most of the problem faced during hands on analysis or data science is because of poor understanding of the problem in hand and concentrating more on tools, end results and other aspects of the project.

Breaking the problem down to a granular level and understanding takes a lot of time and practice to master. Coming back to square one in data science projects can be seen in lot of companies and even in your own project or kaggle problems.

Q18) Advantages of Tableau Prep?

Tableau Prep will reduce a lot of time like how its parent software (Tableau) does when creating impressive visualizations. The tool has a lot of potentials in taking professionals from data cleaning, merging step to creating final usable data that can be linked to Tableau desktop for getting visualization and business insights. A lot of manual tasks will be reduced and the time can be used to make better findings and insights.

Q19) What is the common perception about visualization?

People think visualization as just charts and summary information. But they are beyond that and drive business with a lot of underlying principles. Learning design principles can help anyone build effective and efficient visualizations and this Tableau prep tool can drastically increase our time on focusing more important part. The only issue with Tableau is, it is paid and companies need to pay for leveraging that

awesome tool.

Q20) What are the time series algorithms?

Time series algorithms like ARIMA, ARIMAX, SARIMA, Holts winters are very interesting to learn and use as well to solve a lot of complex problems for businesses. Data preparation for time series analysis plays a vital role. The stationarity, seasonality, cycles and noises need time and attention. Take as much time as you would like to make the data right. Then you can run any model on top of it.

Q21) How to choose the right chart in case of creating a viz?

Using the right chart to represent data is one of the key aspects of data visualization and design principle. You will always have options to choose from when deciding on a chart. But fixing to the right chart comes only by experience, practice and deep understanding of end-user needs. That dictates everything in the dashboard.

Q22) Where to seek help in case of discrepancies in Tableau?

When you face any issue regarding Tableau, try searching in the Tableau community forum. It is one of the best places to get your queries answered. You can always write your question and get the query answered with an hour or a day. You can always post on LinkedIn and follow people.

Q23) Now companies are heavily investing their money and time to make the dashboards. Why?

To make stakeholders more aware about the business through data. Working on visualization projects helps you develop one of the key skills every data scientist should possess i.e. Thinking from the shoes of the end user.

þÿIf you re learning any visualization tool, download a dataset from kagg
the dashboard should be the last step. Research more about the domain and think about the KPIs you
þÿwould like to see in the dashboard if you re going to be the end user.
piece by piece.

Q24) How can I achieve accuracy in the first model that I built?

þÿBuilding machine learning models involves a lot of interesting steps.

in the very first attempt. You have done a lot of better feature selection techniques to get that point, which means it involves a lot of trial and error. The process will help you learn new concepts in statistics, math and probability.

Q25) What is the basic responsibility of a Data Scientist?

As a data scientist, we have the responsibility to make complex things simple enough that anyone without context should understand, what we are trying to convey.

The moment, we start explaining even the simple things the mission of making the complex simple goes away. This happens a lot when we are doing data visualization.

Less is more. Rather than pushing too much information on to readers brain, we need to figure out how easily we can help them consume a dashboard or a chart.

The process is simple to say but difficult to implement. You must bring the complex business value out þÿof a self-explanatory chart. It s a skill every data scientist should striv
arsenal.

Q26) How do I enhance a SAS analyst?

Step 1: Earn a College Degree. Businesses prefer SAS programmers who have completed a

þÿstatistics or computer science bachelor s degree program.

Step 2: Acquire SAS Certification.

Step 3: Consider Getting an Advanced Degree.

Step 4: Gain SAS Program Coding Work Background.

Q27) What does SAS stand out to be the best over other data analytics tools?

Ease to understand: The provisions included in SAS are remarkably easy to learn. Further, it offers the most suitable option for those who already are aware of the SQL. On the other hand, R comes with a steep training cover which is supposed to be a low-level programming style.

Data Handling Capacities: it is at par the most leading tool which also includes the R& Python.

If it advances before handling the huge data, it is the best platform to engage Graphical Capacities: it comes with functional graphical capacities and has a limited knowledge field.

It is useful to customize the plots Better tool management: It benefits in a release the updates with regards to the controlled conditions.

This is the main reason why it is well tested. Whereas if you considered R&Python, it has open contribution also the risk of errors in the current development is also high.

Q28) What is RUN-Group processing?

To practice RUN-group processing, you start the system and then submit many RUNgroups.

A RUN-group is a group of records that contain at least one product group including ends with a RUN statement. It can contain different SAS statements such as AXIS, BY, GOPTIONS,

LEGEND, Power, or WHERE.

Q29) Definitions of is BY-Group processing?

Definitions for BY-Group Processing. is a method of preparing observations from one or numerous SAS data sets that are arranged or ordered by importance of individual or more

shared variables. All data sets that are being connected must include one or more BY variables.

Q30) What is the right way to validate the SAS program?

The OPTIONS OBS=0 through the commencement of the code needs to be written but if yourself require to perform the same then their mind be any log which gets recognized by the

colors that get highlighted.

Q31) Do you know any SAS functions and Call Routines?

Can be a mutable type, uniform, or any SAS expression, including different use. This product also a

letter from contentions that SAS allows are called by special purposes. Multiple arguments are separated with a comma.

Q32) What is means by precision and Recall?

Recall:

It is known as a true real rate. The number of positives that your model has claimed related to the original defined number of positives available during this data.

Precision:

It is also known as a positive predicted value. This is more based on the prediction. That indicates a time like a number of accurate positives that the model needs when compared to the number of positives it actually claims.

Q33) What is deep learning?

Deep learning is a process where it is considered to be a subset of machine learning process.

Q34) What is the F1 score?

þÿThe F1 score is defined as a measure of a model s performance.

35) How is F1 score is used?

The average of Precision and Recall of a model is nothing but F1 score measure. Based on the results, the F1 score is 1 then it is classified as best and 0 being the worst

Q36) What is the difference between Machine learning Vs Data Mining?

Data mining is about working on unlimited data and then extract it to a level anywhere the unusual and unknown patterns are identified.

Machine learning is any method about a study whether it closely relates to design, development concerning the algorithms that provide an ability to certain computers to capacity to learn.

Q37) What are confounding variables?

These are obvious variables in a scientific model that correlates directly or inversely with both the subject and the objective variable. The study fails to account for the confounding factor.

Q38) How can you randomize the items of a list in place in Python?

Consider the example shown below:

```
from random import shuffle

þÿx = [Data, Class, Blue, Flag, Red, Slow]

shuffle(x)

print(x)
```

The output of the following code is as below.

```
þÿ[Red, Data, Blue, Slow, Class, Flag]
```

Q39) How to get indices of N maximum values in a NumPy array?

We can get the indices of N maximum values in a NumPy array using the below code:

```
import numpy as np

arr = np.array([1, 3, 2, 4, 5])

print(arr.argsort()[-3:][::-1])
```

Output

```
[ 4 3 1 ]
```

Q40) How make you 3D plots/visualizations using NumPy/SciPy?

Like 2D plotting, 3D graphics is beyond the scope of NumPy and SciPy, but just as in this 2D example,

packages exist that integrate with NumPy. Matplotlib provides primary 3D plotting in

the mplot3d subpackage, whereas Mayavi produces a wide range of high-quality 3D visualization features, utilizing the powerful VTK engine.

Q41) What are the types of biases that can occur during sampling?

Some simple models of selection bias are described below. Undercoverage occurs when some members þÿof the population live badly represented inside the sample. & The surv of telephone directories and car registration lists.

- **Selection bias**
- **Under coverage bias**
- **Survivorship bias**

Q42) Which Python library is used for data visualization?

Plotly. The fifth tool is Plotly, also called as Plot.ly because of its main platform online. It is an interactive online visualization tool that is being used for data analytics, scientific graphs, and

other visualization. This contains some great API including one for Python

Q43) Write code to sort a DataFrame in Python in descending order.

```
þÿDataFrame.sort_values(by, axis=0, ascending=True, inplace=False, kind

þÿna_position= last )[source]
```

Sort by the values along either axis

Parameters:

by: str or list of str

Name or list of names to sort by.

þÿif an axis is 0 or index then by may contain index levels and/or colum

þÿif the axis is 1 or columns then by may contain column levels and/or i

Changed in version 0.23.0: Allow specifying index or column level names.

þÿaxis : {0 or index, 1 or columns}, default 0

Axis to be sorted

ascending: bool or list of bool, default True

Sort ascending vs. descending. Specify list for multiple sort orders. If this is a list of bools, must match the length of the by.

in place: bool, default False

if True, perform operation in-place

þÿkind: {quicksort, mergesort, heapsort}, default quicksort

Choice of sorting algorithm. See also array.np.sort for more information. mergesort is the only stable algorithm. For DataFrames, this option is only applied when sorting on a single column or label.

þÿna_position : {first, last}, default last

first puts NaNs at the beginning, last put NaNs at the end

Returns:

sorted_obj: DataFrame

Q44) Why you should use NumPy arrays instead of nested Python lists?

þÿAns : lets say you have a list a of numbers, and you want to add 1 to e

In regular python, you would do:

a = [6, 2, 1, 4, 3]

b = [e + 1 fore in a]

Whereas with numpy, you simply have to do:

import numpy as np

a = np.array([6, 2, 1, 4, 3])

b = a + 1

It also works for every numpy mathematics function: you can take the exponential of every

element of a list using np.exp for example.

Q45) Why is an import statement required in Python?

Ans : To be able to use any functionality, the respective code logic needs to be accessible for the Python interpreter. With the help of the import statement, we can use specific scripts. However, there are thousands of such scripts available and every script available cannot be used at once. Hence we import statement to use only the scripts that we want to use

- **import pandas as pd**
- **import numpy as np**

Q46) What is alias in import statement? Why is it used?

Ans : Aliases are used in import statements for ease of usage. If the imported module has a large name, for example import multiprocessing . Everytime we want to access any scrript present in multiprocessing module, we need to use the word multiprocessing.

However if an alias is used, import

multiprocessing as mp, we can simply replace the words multiprocessing with mp

Q47) Are the aliases used for a module fixed/static ?

Ans : No, the aliases are not pre-fixed. The alias can be named as per your convenience. However, the documentation of a respective module sometimes specifies the alias to be used for ease of understanding.

Q48) How to access a specific script inside a module?

Ans : If the whole module needs to be imported, we simply can use from pandas import *

Q49) What is a nonparametric test used for?

Ans : Non parametric tests do not assume that the data follows a specific distribution. They can be used whenever the data do not meet the assumptions of parametric test.

Q50) What are the pros and cons of Decision Trees algorithm?

Ans :

þÿ Pros Easy to interpret. Will ignore irrelevant independant variables s
minimal. Can handle missing data. Fast modelling.
þÿ Cons Many combinations are possible to create a
tree. There are chances that it might not find the best tree possible.

Q51) Name some Classification Algorithms.

Ans : Linear Classifiers: Logistic Regression, Naive Bayes Classifier, Decision Trees, Random Forest, Neural Networks, K Nearest Neighbor.

Q52) What are pros and cons of Naive Bayes algorithm?

Ans :

Big sized data is handled easily

Multiclass perfomance is good and accurate

It is not process
intensive

Cons: Assumea independence of predictor variables.

Q53) What are the types of Skewness?

Ans : A dataset that is skewed right or left are the two types.

Q54) What is skewed data?

Ans : A data distribution that is has skewed data towards the right or left.

Q55) What is the skewness of this data? 27 ; 28 ; 30 ; 32 ; 34 ; 38 ; 41 ; 42 ; 43 ; 44 ; 46 ; 53 ; 56 ; 62

Ans : The data set is skewed left

Q56) What is an outlier?

An outlier is a value that is very much away from the rest of the values in the data set.

Q57) Mention the characteristics of symmetric data distribution?

Ans : The mean is equal to the median and the tails of the distribution are balanced.

Q58) What are the applications of data science?

Ans : Optical character recognition, recommendation engines, fitering algorithms, personal assistants, advertising, surveillance, autonomous driving, facial recognition and more.

Q59) Define EDA?

Ans : EDA [exploratory data analysis] is an apporach to analysing data to summarise their main characteriscs, often with visual methods.

Q60) What are the steps in exploratory data analysis?

Ans :

- **Make summary of observations**
- **describe central tendencies or core part of dataset**
- **desribe shape of data**
- **identify potential associations**

- develop insight into errors, misssing values and major deviations

Q61) What are the types of data available in Enterprises?

Ans :

- Structured data
- unstructured data
- big data from social media, surveys, pictures, audio, video, drawings, maps.
- Machine generated data from instruments
- real time data feeds

Q62) What are the various types of analysis on type of data?

Ans :

þÿ Univariate 1 variable

þÿ bivariate 2 variables

þÿ multivariate more than 2 variables

Q63) What is difference between primary data and secondary data?

Ans :

Data collected by the interested/self is primary data. This data is collected afresh and first time.

Someone else has collected the data and being used by you is secondary data.

Q64) What is the difference between qualitative & quantitative ?

Ans : Quantitative method analyses the data based on numbers. Qualitative method analyses the data by attributes.

Q65) What is histogram?

Ans : Histogram is the accurate representation of numerical data based on their occurrences/frequencies.

Q66) What are the common measures of central tendancies?

Ans :

- **Mean**
- **Median**
- **Mode**

Q67) What are quartiles?

Ans : Quartiles are three points in the data, that divide the data into four groups. Each group consisting of a quarter of data.

Q68) What are the commonly used error metrics in regression tasks?

Ans :

þÿ MSE Mean squared error Average of square of errors

þÿ RMSE Root mean square error root of MSE

þÿ MAPE Mean absolute percentage error

Q69) What are the commonly used error metrics for classification tasks?

Ans :

- **F1 score**
- **Accuracy**
- **Sensitivity**
- **Specificity**
- **Recall**
- **Precision**

Q70) What is it called when there are more than 1 explanatory variables in the regression task?

Ans : Multiple linear regression

Q71) What are residuals in a regression task?

Ans : The difference between the predicted value and the actual value is called the residual.

Q72) What are the main classifications in Machine learning?

Ans :

- **Supervised learning**
- **Unsupervised learning**
- **Reinforcement learning**

Q73) What are the main types of supervised learning tasks?

Ans :

Classification task [categorical in nature] Regression task [continuous in nature]

Q74) Can Random forest be used for classification and regression?

Ans : Yes, it can be used

Give a simple representation for Linear Equation.

Ans : Y = mx + c ; where y is

the dependant variable; c is the independant variable;m is slope

Q75) What is R square value?

Ans : R squared values tells us how close the regression line is fit to the actual values.

Q76) What are some common ways of imputation?

Ans : Mean imputation, median imputation, KNN imputation, Stochastic regression, substitution

Q77) What is the difference between series and list

Ans :

list is size and data mutable

series is data mutable but not size mutable

Q78) Which function is used to get descriptive statistics of a dataframe?

Ans : describe()

Q79) What parameter is used to update the data without explicitly assigning data to a variable.

Ans : Inplace is used to assign result of function to itself. If inplace = True , there is no need to explicitly assign to a variable

Q80) What is the difference between a dictionary and a set?

Ans :

Dictionary has key value pair

set does not have key value pairs

set has only unique elements

Q81) How to create a series with letters as index?

þÿAns : Series({a:1, b:2}) will create a and b as index. 1 and 2 as their r

Q82) Which function can be used to filter a DataFrame?

Ans : The query function can be used to filter a dataframe.

Q83) What is the function to create test train split?

Ans : From sklearn.metrics import test_train_split . This function is used to create test train split from the data.

Q84) What is pickling?

Ans : Pickling is the process of saving a data structure into the physical drive or hard disk.

Q85) What is unpickling?

Ans : Unpickling is used to read a pickled file from hard disk or physical storage drive.

Q86) What are the most common web frameworks of Python?

Ans : Django and Flask.

Q87) How to convert n number of series to a dataframe?

þÿAns : DataFrame(data = {col1 :series1, col2 :series2}).

Q88) How to select a section of a dataframe?

Ans : Using iloc and loc functions the rows and columns can be selected.

Q89) How are exceptions handled in Python?

Ans : Exceptions can be handled using the try except statements.

Q90) Is multiprocessing possible in python?

Ans : Yes it is possible using the multiprocessing module.

Q91) Can the values be replaced in tuple?

Ans : No values cannot be replaced in tuple as tuple is data immutable

Q92) What are lambda function in Python and how it is different from def (defining functions) in Python?

Ans : Lambda function in Python is used for evaluating an expression and then return a value. Where as def needs a function name, and the program logic is broken into smaller chunks. Lambda is an inline function consisting of only a single expression, It can take any number of arguments.

Q93) Difference between supervised and unsupervised machine learning?

Ans : Supervised learning is a method where it needs training specified data. When it gets to þÿUnsupervised learning it doesn t need data labeling.

Q94) How to differentiate from KNN and K-means clustering?

Ans : KNN is standing for the K- Nearest Neighbours, it remains classified because a supervised algorithm.K-means is an unsupervised cluster algorithm.

Q95) What is your opinion on our current data process &nbs p;?

Ans : This type of questions signifies asked and the individuals must to carefully listen to their value case and at this same time, the return should be in a constructive also insightful manner. Based on your responses, the interviewer mind has a future to review and know whether you imply a vague reply to their team or not.

Q96) Difference between Machine learning and Data Mining?

Ans :

- Data mining is about going about unstructured data and when extracting this to a level anywhere that interesting also unknown patterns remain identified.
- Machine learning is any process or a concept whether it closely relates designing, development of the algorithms that give an experience within these machines on the capacity to learn.

Q97) Explain about from capture of the correlation between continuous and categorical variable?

Ans : It is possible to that using ANCOVA technique. It exists for Analysis of Covariance. It is used to calculate this association between continuous including categorical variables.

Q98) Difference between an Array and a Linked list?

Ans : An array is an established method of collection objects. A linked program is a group of objects that are prepared into sequential order.

þÿQ99) Difference between long and wide format data?

þÿAns : In the wide form, each subject s happened responses will remain answer is into a separate column. In the long format, each data is a one-time time by subject. You can understand data in wide form by that fact that columns usually design groups.

Q100) What do you know by the term Normal Distribution?

Ans :

Data is usually distributed under many ways including a bias on the port or over the benefit or it can all

- be jumbled up.
- However, there continue indications that data is distributed on a central position without bias to the left or right more gives natural order in some form of a bell-shaped curve.

Q101) Differences between overfitting and underfitting

Ans :

- In statistics and machine learning, individual of that most basic tasks is to fit one model on a collection of training data, so doing to be ready to provide reliable predictions of general untrained data.
- Underfitting happens at a statistical design or machine learning algorithm cannot get this underlying trend of the data.