NARESH i technologies

Data Science Interview Questions :

1. What is the difference between Type I Error & Type II Error? Also, Explain the Power of the test?

When we perform hypothesis testing we consider two types of Error, Type I error and Type II error, sometimes we reject the null hypothesis when we should not or choose not to reject the null hypothesis when we should.

A Type I Error is committed when we reject the null hypothesis when the null hypothesis is actually true. On the other hand, a Type II error is made when we do not reject the null hypothesis and the null hypothesis is actually false.

The probability of a Type I error is denoted by α and the probability of Type II error is denoted by β.

For a given sample n, a decrease in α will increase β and vice versa. Both α and β decrease as n increases.

The table given below explains the situation around the Type I error and Type II error:

Decision	**Null Hypothesis is true**	**Null hypothesis is false**
Reject the Null Hypothesis	Type I error	Correct Decision
Fail to reject Null Hypothesis	Correct Decision	Type II error

Two correct decisions are possible: not rejecting the null hypothesis when the null hypothesis is true and rejecting the null hypothesis when the null hypothesis is false.

Conversely, two incorrect decisions are also possible: Rejecting the null hypothesis when the null hypothesis is true(Type I error), and not rejecting the null hypothesis when the null hypothesis is false (Type II error).

Type I error is false positive while Type II error is a false negative.

Power of Test: The Power of the test is defined as the probability of rejecting the null hypothesis when the null hypothesis is false. Since β is the probability of a Type II error, the power of the test is defined as 1- β. In advanced statistics, we compare various types of tests based on their size and power, where the size denotes the actual proportion of rejections when the null is true and the power denotes the actual proportion of rejections when the null is false.

2. What do you understand by Over-fitting and Under-fitting?

Overfitting is observed when there is a small amount of data and a large number of variables, If the model we finish with ends up modelling the noise as well, we call it "overfitting" and if we are not modelling all the information, we call it "underfitting". Most commonly underfitting is observed when a linear model is fitted to a non-linear data.

The hope is that the model that does the best on testing data manages to capture/model all the information but leave out all the noise. Overfitting can be avoided by using cross-validation techniques (like K Folds) and regularisation techniques (like Lasso regression).

3. When do you use the Classification Technique over the Regression Technique?

Classification problems are mainly used when the output is the categorical variable (Discrete) whereas Regression Techniques are used when the output variable is Continuous variable.

In the Regression algorithm, we attempt to estimate the mapping function (f) from input variables (x) to numerical (continuous) output variable (y).

For example, Linear regression, Support Vector Machine (SVM) and Regression trees.

In the Classification algorithm, we attempt to estimate the mapping function (f) from the input variable (x) to the discrete or categorical output variable (y).

For example, Logistic Regression, naïve Bayes, Decision Trees & K nearest neighbours.

Both Classifications, as well as Regression techniques, are Supervised Machine Learning Algorithms.

4. What is the importance of Data Cleansing?

Ans. As the name suggests, data cleansing is a process of removing or updating the information that is incorrect, incomplete, duplicated, irrelevant, or formatted improperly. It is very important to improve the quality of data and hence the accuracy and productivity of the processes and organisation as a whole.

Real-world data is often captured in formats which have hygiene issues. There are sometimes errors due to various reasons which make the data inconsistent and sometimes only some features of the data. Hence data cleansing is done to filter the usable data from the raw data, otherwise many systems consuming the data will produce erroneous results.

5. Which are the important steps of Data Cleaning?

Different types of data require different types of cleaning, the most important steps of Data Cleaning are:

1. Data Quality
2. Removing Duplicate Data (also irrelevant data)
3. Structural errors
4. Outliers
5. Treatment for Missing Data

Data Cleaning is an important step before analysing data, it helps to increase the accuracy of the model. This helps organisations to make an informed decision.

Data Scientists usually spends 80% of their time cleaning data.

6. How is k-NN different from k-means clustering?

Ans. K-nearest neighbours is a classification algorithm, which is a subset of supervised learning. K-means is a clustering algorithm, which is a subset of unsupervised learning.

And K-NN is a Classification or Regression Machine Learning Algorithm while K-means is a Clustering Machine Learning Algorithm.

K-NN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample, whereas K-means is the number of clusters the algorithm is trying to learn from the data.

7. What is p-value?

Ans. p-value helps you determine the strengths of your results when you perform a hypothesis test. It is a number between 0 and 1. The claim which is on trial is called the Null Hypothesis. Lower p-values, i.e. ≤ 0.05, means we can reject the Null Hypothesis. A high p-value, i.e. ≥ 0.05, means we can accept the Null Hypothesis. An exact p-value 0.05 indicates that the Hypothesis can go either way.

P-value is the measure of the probability of events other than suggested by the null hypothesis. It effectively means the probability of events rarer than the event being suggested by the null hypothesis.

8. How is Data Science different from Big Data and Data Analytics?

Ans. Data Science utilises algorithms and tools to draw meaningful and commercially useful insights from raw data. It involves tasks like data modelling, data cleansing, analysis, pre-processing etc.
Big Data is the enormous set of structured, semi-structured, and unstructured data in its raw form generated through various channels.
And finally, Data Analytics provides operational insights into complex business scenarios. It also helps in predicting upcoming opportunities and threats for an organisation to exploit.

Essentially, big data is the process of handling large volumes of data. It includes standard practices for data management and processing at a high speed maintaining the consistency of data. Data analytics is associated with gaining meaningful insights from the data through mathematical or non-mathematical processes. Data Science is the art of making intelligent systems so that they learn from data and then make decisions according to past experiences.

Statistics in Data Science Interview Questions

9. What is the use of Statistics in Data Science?

Ans. Statistics in Data Science provides tools and methods to identify patterns and structures in data to provide a deeper insight into it. Serves a great role in data acquisition, exploration, analysis, and validation. It plays a really powerful role in Data Science.

Data Science is a derived field which is formed from the overlap of statistics probability and computer science. Whenever one needs to do estimations, statistics is involved. Many algorithms in data science are built on top of statistical formulae and processes. Hence statistics is an important part of data science.

Also Read: *Practical Ways to Implement Data Science in Marketing*

10. What is the difference between Supervised Learning and Unsupervised Learning?

Ans. Supervised Machine Learning requires labelled data for training while Unsupervised Machine Learning does not require labelled data. It can be trained on unlabelled data.

To elaborate, supervised learning involves training of the model with a target value whereas unsupervised has no known results to learn and it has a state-based or adaptive mechanism to learn by itself. Supervised learning involves high computation costs whereas unsupervised learning has low training cost. Supervised learning finds applications in classification and regression tasks whereas unsupervised learning finds applications in clustering and association rule mining.

11. What is a Linear Regression?

Ans. The linear regression equation is a one-degree equation with the most basic form being Y = mX + C where m is the slope of the line and C is the standard error. It is used when the response variable is continuous in nature for example height, weight, and the number of hours. It can be a simple linear regression if it involves continuous dependent variable with one independent variable and a multiple linear regression if it has multiple independent variables.

Linear regression is a standard statistical practice to calculate the best fit line passing through the data points when plotted. The best fit line is chosen in such a way so that the distance of each data point is minimum from the line which reduces the overall error of the system. Linear regression assumes that the various features in the data are linearly related to the target. It is often used in predictive analytics for calculating estimates in the foreseeable future.

12. What is Logistic Regression?

Ans. Logistic regression is a technique in predictive analytics which is used when we are doing predictions on a variable which is dichotomous(binary) in nature. For example, yes/no or true/false etc. The equation for this method is of the form Y = eX + e – X . It is used for classification based tasks. It finds out probabilities for a data point to belong to a particular class for classification.

13. Explain Normal Distribution

Ans. Normal Distribution is also called the Gaussian Distribution. It is a type of probability distribution such that most of the values lie near the mean. It has the following characteristics:

- The mean, median, and mode of the distribution coincide
- The distribution has a bell-shaped curve
- The total area under the curve is 1
- Exactly half of the values are to the right of the centre, and the other half to the left of the centre

14. Mention some drawbacks of the Linear Model

Ans. Here a few drawbacks of the linear model:

- The assumption regarding the linearity of the errors
- It is not usable for binary outcomes or count outcome
- It can’t solve certain overfitting problems
- It also assumes that there is no multicollinearity in the data.

15. Which one would you choose for text analysis, R or Python?

Ans. Python would be a better choice for text analysis as it has the Pandas library to facilitate easy-to-use data structures and high-performance data analysis tools. However, depending on the complexity of data one could use either which suits best.

16. What steps do you follow while making a decision tree?

Ans. The steps involved in making a decision tree are:

1. Determine the Root of the Tree Step
2. Calculate Entropy for The Classes Step
3. Calculate Entropy After Split for Each Attribute
4. Calculate Information Gain for each split
5. Perform the Split
6. Perform Further Splits Step
7. Complete the Decision Tree

17. What is correlation and covariance in statistics?

Ans. Correlation is defined as the measure of the relationship between two variables. If two variables are directly proportional to each other, then its positive correlation. If the variables are indirectly proportional to each other, it is known as a negative correlation. Covariance is the measure of how much two random variables vary together.

18. What is 'Naive' in a Naive Bayes?

Ans. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Basically, it's "naive" because it makes assumptions that may or may not turn out to be correct.

19. How can you select k for k-means?

Ans. The two methods to calculate the optimal value of k in k-means are:

1. Elbow method
2. Silhouette score method

Silhouette score is the most prevalent while determining the optimal value of k.

20. What Native Data Structures Can You Name in Python? Of These, Which Are Mutable, and Which Are Immutable?

Ans. The native python data structures are:

- Lists
- Tuples
- Sets
- Dictionary

Tuples are immutable. Others are mutable.

21. What libraries do data scientists use to plot data in Python?

Ans. The libraries used for data plotting are:

- matplotlib
- seaborn
- ggplot.

Apart from these, there are many opensource tools, but the aforementioned are the most used in common practice.

22. How is Memory Managed in Python?

Ans. Memory management in Python involves a private heap containing all Python objects and data structures. The management of this private heap is ensured internally by the Python memory manager.

23. What is a recall?

Ans. Recall gives the rate of true positives with respect to the sum of true positives and false negatives. It is also known as true positive rate.

24. What are lambda functions?

Ans. A lambda function is a small anonymous function. A lambda function can take any number of arguments, but can only have one expression.

25. What is reinforcement learning?

Ans. Reinforcement learning is an unsupervised learning technique in machine learning. It is a state-based learning technique. The models have predefined rules for state change which enable the system to move from one state to another, while the training phase.

26. What is Entropy and Information Gain in decision tree algorithm?

Ans. Entropy is used to check the homogeneity of a sample. If the value of entropy is ‘0’ then the sample is completely homogenous. On the other hand, if entropy has a value ‘1’, the sample is equally divided. Entropy controls how a Decision Tree decides to split the data. It actually affects how a Decision Tree draws its boundaries.

The information gain depends on the decrease in entropy after the dataset is split on an attribute. Constructing a decision tree is always about finding the attributes that return highest information gain.

27. What is Cross-Validation?

Ans. It is a model validation technique to asses how the outcomes of a statistical analysis will infer to an independent data set. It is majorly used where prediction is the goal and one needs to estimate the performance accuracy of a predictive model in practice.
The goal here is to define a data-set for testing a model in its training phase and

limit overfitting and underfitting issues. The validation and the training set is to be drawn from the same distribution to avoid making things worse.

Also Read: *Why Data Science Jobs Are in Demand*

28. What is Bias-Variance tradeoff?

Ans. The error introduced in your model because of over-simplification of the algorithm is known as Bias. On the other hand, Variance is the error introduced to your model because of the complex nature of machine learning algorithm. In this case, the model also learns noise and perform poorly on the test dataset.

The bias-variance tradeoff is the optimum balance between bias and variance in a machine learning model. If you try to decrease bias, the variance will increase and vice-versa.

Total Error= Square of bias+variance+irreducible error. Bias variance tradeoff is the process of finding the exact number of features while model creation such that the error is kept minimum, but also taking effective care such that the model does not overfit or underfit.

29. Mention the types of biases that occur during sampling?

Ans. The three types of biases that occur during sampling are:
a. Self-Selection Bias
b. Under coverage bias
c. Survivorship Bias

Self selection is when the participants of the analysis select themselves. Undercoverage occurs when very few samples are selected from a segment of the population. Survivorship bias occurs when the observations recorded at the end of the investigation are a non-random set of those present at the beginning of the investigation.

30. What is the Confusion Matrix?

Ans. A confusion matrix is a 2X2 table that consists of four outputs provided by the binary classifier.

A binary classifier predicts all data instances of a test dataset as either positive or negative. This produces four outcomes-

1. True positive(TP) — Correct positive prediction
2. False-positive(FP) — Incorrect positive prediction
3. True negative(TN) — Correct negative prediction
4. False-negative(FN) — Incorrect negative prediction

It helps in calculating various measures including error rate (FP+FN)/(P+N), specificity(TN/N), accuracy(TP+TN)/(P+N), sensitivity (TP/P), and precision(TP/(TP+FP)).

A confusion matrix is essentially used to evaluate the performance of a machine learning model when the truth values of the experiments are already known and the target class has more than two categories of data. It helps in visualisation and evaluation of the results of the statistical process.

31. Explain selection bias

Ans. Selection bias occurs when the research does not have a random selection of participants. It is a distortion of statistical analysis resulting from the method of collecting the sample. Selection bias is also referred to as the selection effect. When professionals fail to take selection bias into account, their conclusions might be inaccurate.

Some of the different types of selection biases are:

- Sampling Bias – A systematic error that results due to a non-random sample
- Data – Occurs when specific data subsets are selected to support a conclusion or reject bad data
- Attrition – Refers to the bias caused due to tests that didn’t run to completion.

32. What are exploding gradients?

Ans. Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and result in NaN values. Hence the model becomes unstable and is unable to learn from the training data.

33. Explain the Law of Large Numbers

Ans. The ‘Law of Large Numbers’ states that if an experiment is repeated independently a large number of times, the average of the individual results is close to the expected value. It also states that the sample variance and standard deviation also converge towards the expected value.

34. What is the importance of A/B testing

Ans. The goal of A/B testing is to pick the best variant among two hypotheses, the use cases of this kind of testing could be a web page or application responsiveness, landing page redesign, banner testing, marketing campaign performance etc.
The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

35. Explain Eigenvectors and Eigenvalues

Ans. Eigenvectors depict the direction in which a linear transformation moves and acts by compressing, flipping, or stretching. They are used to understand linear transformations and are generally calculated for a correlation or covariance matrix.
The eigenvalue is the strength of the transformation in the direction of the eigenvector.

An eigenvector’s direction remains unchanged when a linear transformation is applied to it.

36. Why Is Re-sampling Done?

Ans. Resampling is done to:

- Estimate the accuracy of sample statistics with the subsets of accessible data at hand
- Substitute data point labels while performing significance tests
- Validate models by using random subsets

37. What is systematic sampling and cluster sampling

Ans. Systematic sampling is a type of probability sampling method. The sample members are selected from a larger population with a random starting point but a fixed periodic interval. This interval is known as the sampling interval. The sampling interval is calculated by dividing the population size by the desired sample size.

Cluster sampling involves dividing the sample population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. Analysis is conducted on data from the sampled clusters.

38.What are Autoencoders?

Ans. An autoencoder is a kind of artificial neural network. It is used to learn efficient data codings in an unsupervised manner. It is utilised for learning a representation (encoding) for a set of data, mostly for dimensionality reduction, by training the network to ignore signal “noise”. Autoencoder also tries to generate a representation as close as possible to its original input from the reduced encoding.

39. What are the steps to build a Random Forest Model?

A Random Forest is essentially a build up of a number of decision trees. The steps to build a random forest model include:

Step1: Select ‘k’ features from a total of ‘m’ features, randomly. Here k << m

Step2: Calculate node D using the best split point — along the ‘k’ features

Step 3: Split the node into daughter nodes using best splitStep 4: Repeat Steps 2 and 3 until the leaf nodes are finalised

Step5: Build a Random forest by repeating steps 1-4 for ‘n’ times to create ‘n’ number of trees.

40. How do you avoid the overfitting of your model?

Overfitting basically refers to a model that is set only for a small amount of data. It tends to ignore the bigger picture. Three important methods to avoid overfitting are:

- Keeping the model simple—using fewer variables and removing major amount of the noise in the training data
- Using cross-validation techniques. E.g.: k folds cross-validation
- Using regularisation techniques — like LASSO, to penalise model parameters that are more likely to cause overfitting.

41. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate data, as the name suggests, contains only one variable. The univariate analysis describes the data and finds patterns that exist within it.

Bivariate data contains two different variables. The bivariate analysis deals with causes, relationships and analysis between those two variables.

Multivariate data contains three or more variables. Multivariate analysis is similar to that of a bivariate, however, in a multivariate analysis, there exists more than one dependent variable.

42. How is random forest different from decision trees?

Ans. A Decision Tree is a single structure. Random forest is a collection of decision trees.

43. What is dimensionality reduction? What are its benefits?

Dimensionality reduction is defined as the process of converting a data set with vast dimensions into data with lesser dimensions — in order to convey similar information concisely.

This method is mainly beneficial in compressing data and reducing storage space. It is also useful in reducing computation time due to fewer dimensions. Finally, it helps remove redundant features — for instance, storing a value in two different units (meters and inches) is avoided.

In short, dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

44. For the given points, how will you calculate the Euclidean distance in Python? plot1 = [1,3] ; plot2 = [2,5]

Ans.

```
import math
```

```
# Example points in 2-dimensional space...
x = (1,3)
y = (2,5)
distance = math.sqrt(sum([(a - b) ** 2 for a, b in zip(x, y)]))
print("Euclidean distance from x to y: ",distance)
```

45. Mention feature selection methods used to select the right variables.

The methods for feature selection can be broadly classified into two types:

Filter Methods: These methods involve:

- Linear discrimination analysis
- ANOVA
- Chi-Square

Wrapper Methods: These methods involve

- Forward Selection: One feature at a time is tested and a good fit is obtained
- Backward Selection: All features are reviewed to see what works better
- Recursive Feature Elimination: Every different feature is looked at recursively and paired together accordingly.

Others are Forward Elimination, Backward Elimination for Regression, Cosine Similarity-Based Feature Selection for Clustering tasks, Correlation-based eliminations etc.

Machine Learning in Data Science Interview Questions

46. What are the different types of clustering algorithms?

Ans. Kmeans Clustering, KNN (K nearest neighbour), Hierarchial clustering, Fuzzy Clustering are some of the common examples of clustering algorithms.

47. How should you maintain a deployed model?

Ans. A deployed model needs to be retrained after a while so as to improve the performance of the model. Since deployment, a track should be kept of the predictions made by the model and the truth values. Later this can be used to retrain the model with the new data. Also, root cause analysis for wrong predictions should be done.

48. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables? K-

means clustering Linear regression K-NN (k-nearest neighbour) Decision trees

Ans. KNN and Kmeans

49. What is a ROC Curve? Explain how a ROC Curve works?

Ans. AUC – ROC curve is a performance measurement for the classification problem at various thresholds settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

50. How do you find RMSE and MSE in a linear regression model?

Ans. Mean square error is the squared sum of (actual value-predicted value) for all data points. It gives an estimate of the total square sum of errors. Root mean square is the square root of the squared sum of errors.

51. Can you cite some examples where a false negative holds more importance than a false positive?

Ans. In cases of predictions when we are doing disease prediction based on symptoms for diseases like cancer.

52. How can outlier values be treated?

Ans. Outlier treatment can be done by replacing the values with mean, mode, or a cap off value. The other method is to remove all rows with outliers if they make up a small proportion of the data. A data transformation can also be done on the outliers.

53. How can you calculate accuracy using a confusion matrix?

Ans. Accuracy score can be calculated by the formula: (TP+TN)/(TP+TN+FP+FN), where TP= True Positive, TN=True Negatives, FP=False positive, and FN=False Negative.

54. What is the difference between “long” and “wide” format data?

Ans. Wide-format is where we have a single row for every data point with multiple columns to hold the values of various attributes. The long format is where for each data point we have as many rows as the number of attributes and each row contains the value of a particular attribute for a given data point.

55. Explain the SVM machine learning algorithm in detail.

Ans. SVM is an ML algorithm which is used for classification and regression. For classification, it finds out a muti dimensional hyperplane to distinguish between classes. SVM uses kernels which are namely linear, polynomial, and rbf. There are few parameters which need to be passed to SVM in order to specify the points to consider while the calculation of the hyperplane.

56. What are the various steps involved in an analytics project?

Ans. The steps involved in a text analytics project are:

1. Data collection
2. Data cleansing
3. Data pre-processing
4. Creation of train test and validation sets
5. Model creation
6. Hyperparameter tuning
7. Model deployment

57. Explain Star Schema.

Ans. Star schema is a data warehousing concept in which all schema is connected to a central schema.

58. How Regularly Must an Algorithm be Updated?

Ans. It completely depends on the accuracy and precision being required at the point of delivery and also on how much new data we have to train on. For a model trained on 10 million rows its important to have new data with the same volume or close to the same volume. Training on 1 million new data points every alternate week, or fortnight won't add much value in terms of increasing the efficiency of the model.

59. What is Collaborative Filtering?

Ans. Collaborative filtering is a technique that can filter out items that a user might like on the basis of reactions by similar users. It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user.

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

Ans. By determining the Silhouette score and elbow method, we determine the number of clusters in the algorithm.

61. What is Ensemble Learning? Define types.

Ans. Ensemble learning is clubbing of multiple weak learners (ml classifiers) and then using aggregation for result prediction. It is observed that even if the classifiers perform poorly individually, they do better when their results are aggregated. An example of ensemble learning is random forest classifier.

62. What are the support vectors in SVM?

Ans. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximise the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

63. What is pruning in Decision Tree?

Ans. Pruning is the process of reducing the size of a decision tree. The reason for pruning is that the trees prepared by the base algorithm can be prone to overfitting as they become incredibly large and complex.

64. What are the various classification algorithms?

Ans. Different types of classification algorithms include logistic regression, SVM, Naive Bayes, decision trees, and random forest.

65. What are Recommender Systems?

Ans. A recommendation engine is a system, which on the basis of data analysis of the history of users and behaviour of similar users, suggests products, services, information to users. A recommendation can take user-user relationship, product-product relationships, product-user relationship etc. for recommendations.

Data Analysis Interview Questions

66. List out the libraries in Python used for Data Analysis and Scientific Computations.

Ans. The libraries NumPy, Scipy, Pandas, sklearn, Matplotlib which are most prevalent. For deep learning Pytorch, Tensorflow is great tools to learn.

67. State the difference between the expected value and the mean value.

Ans. Mathematical expectation, also known as the expected value, is the summation or integration of possible values from a random variable. Mean value is the average of all data points.

68. How are NumPy and SciPy related?

Ans. NumPy and SciPy are python libraries with support for arrays and mathematical functions. They are very handy tools for data science.

69. 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 ()]
```

Ans. Error

70. What do you mean by list comprehension?

Ans. List comprehension is an elegant way to define and create a list in Python. These lists often have the qualities of sets but are not in all cases sets. List comprehension is a complete substitute for the lambda function as well as the functions map(), filter(), and reduce().

71. What is __init__ in Python?

Ans. “__init__” is a reserved method in python classes. It is known as a constructor in object-oriented concepts. This method is called when an object is created from the class and it allows the class to initialise the attributes of the class.

72. What is the difference between append() and extend() methods?

Ans. append() is used to add items to list. extend() uses an iterator to iterate over its argument and adds each element in the argument to the list and extends it.

73. What is the output of the following? x = [‘ab’, ‘cd’] print(len(list(map(list, x))))

Ans. 2

74. Write a Python program to count the total number of lines in a text file.

Ans.

```
count=0

with open ('filename.txt','rb') as f:

   for line in f:

      count+=1
```

```
print count
```

75. How will you read a random line in a file?

Ans.

```
import random

def random_line(fname): lines = open(fname).read().splitlines()

   return random.choice(lines) print(random_line('test.txt'))
```

76. How would you effectively represent data with 5 dimensions?

Ans. It can be represented in a NumPy array of dimensions (n*n*n*n*5)

77. Whenever you exit Python, is all memory de-allocated?

Ans. Objects having circular references are not always free when python exits. Hence when we exit python all memory doesn’t necessarily get deallocated.

78. How would you create an empty NumPy array?

Ans.

```
"import numpy as np

np.empty([2, 2])"
```

79. Treating a categorical variable as a continuous variable would result in a better predictive model?

Ans. There is no substantial evidence for that, but in some cases, it might help. It’s totally a brute force approach. Also, it only works when the variables in question are ordinal in nature.

80. How and by what methods data visualisations can be effectively used?

Ans. Data visualisation is greatly helpful while creation of reports. There are quite a few reporting tools available such as tableau, Qlikview etc. which make use of plots, graphs etc for representing the overall idea and results for analysis. Data visualisations are also used in exploratory data analysis so that it gives us an overview of the data.

81. You are given a data set consisting of variables with more than 30 per cent missing values. How will you deal with them?

Ans. If 30 per cent data is missing from a single column then, in general, we remove the column. If the column is too important to be removed we may impute values. For imputation, several methods can be used and for each method of imputation, we need to evaluate the model. We should stick with one that model which gives us the best results and generalises well to unseen data.

82. What is skewed Distribution & uniform distribution?

Ans. The skewed distribution is a distribution in which the majority of the data points lie to the right or left of the centre. A uniform distribution is a probability distribution in which all outcomes are equally likely.

83. What can be used to see the count of different categories in a column in pandas?

Ans. value_counts will show the count of different categories.

84. What is the default missing value marker in pandas, and how can you detect all missing values in a DataFrame?

Ans. NaN is the missing values marker in pandas. All rows with missing values can be detected by is_null() function in pandas.

85. What is root cause analysis?

Ans. Root cause analysis is the process of tracing back of occurrence of an event and the factors which lead to it. It's generally done when a software malfunctions. In data science, root cause analysis helps businesses understand the semantics behind certain outcomes.

86. What is a Box-Cox Transformation?

Ans. A Box Cox transformation is a way to normalise variables. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

87. What if instead of finding the best split, we randomly select a few splits and just select the best from them. Will it work?

Ans. The decision tree is based on a greedy approach. It selects the best option for each branching. If we randomly select the best split from average splits, it would give us a locally best solution and not the best solution producing sub-par and sub-optimal results.

88. What is the result of the below lines of code?

```
def fast (items= []):
```

```
items.append (1)

return items



print fast ()

print fast ()
```

Ans. [1]

89. How would you produce a list with unique elements from a list with duplicate elements?

Ans.

```
l=[1,1,2,2]

l=list(set(l))

l
```

90. How will you create a series from dict in Pandas?

Ans.

```
import pandas as pd



# create a dictionary

dictionary = {'cat' : 10, 'Dog' : 20}



# create a series

series = pd.Series(dictionary)
```

```
print(series)
```

91. How will you create an empty DataFrame in Pandas?

Ans.

```
column_names = ["a", "b", "c"]



df = pd.DataFrame(columns = column_names)
```

92. How to get the items of series A not present in series B?

Ans. We can do so by using series.isin() in pandas.

93. How to get frequency counts of unique items of a series?

Ans. pandas.Series.value_counts gives the frequency of items in a series.

94. How to convert a numpy array to a dataframe of given shape?

Ans. If matrix is the numpy array in question: df = pd.DataFrame(matrix) will convert matrix into a dataframe.

95. What is Data Aggregation?

Ans. Data aggregation is a process in which aggregate functions are used to get the necessary outcomes after a groupby. Common aggregation functions are sum, count, avg, max, min.

96. What is Pandas Index?

Ans. An index is a unique number by which rows in a pandas dataframe are numbered.

97. Describe Data Operations in Pandas?

Ans. Common data operations in pandas are data cleaning, data preprocessing, data transformation, data standardisation, data normalisation, data aggregation.

98. Define GroupBy in Pandas?

Ans. groupby is a special function in pandas which is used to group rows together given certain specific columns which have information for categories used for grouping data together.

99. How to convert the index of a series into a column of a dataframe?

Ans. df = df.reset_index() will convert index to a column in a pandas dataframe.

Advanced Data Science Interview Questions

100. How to keep only the top 2 most frequent values as it is and replace everything else as ‘Other’?

Ans.

```
"s = pd.Series(np.random.randint(1, 5, [12]))

print(s.value_counts())

s[~s.isin(ser.value_counts().index[:2])] = 'Other'

s"
```

101. How to convert the first character of each element in a series to uppercase?

Ans. pd.Series([x.title() for x in s])

102. How to get the minimum, 25th percentile, median, 75th, and max of a numeric series?

Ans.

```
"randomness= np.random.RandomState(100)

s = pd.Series(randomness.normal(100, 55, 5))

np.percentile(ser, q=[0, 25, 50, 75, 100])"
```

103. What kind of data does Scatterplot matrices represent?

Ans. Scatterplot matrices are most commonly used to visualise multidimensional data. It is used in visualising bivariate relationships between a combination of variables.

104. What is the hyperbolic tree?

Ans. A hyperbolic tree or hypertree is an information visualisation and graph drawing method inspired by hyperbolic geometry.

105. What is scientific visualisation? How it is different from other visualisation techniques?

Ans. Scientific visualization is representing data graphically as a means of gaining insight from the data. It is also known as visual data analysis. This helps to understand the system that can be studied in ways previously impossible.

106. What are some of the downsides of Visualisation?

Ans. Few of the downsides of visualisation are: It gives estimation not accuracy, a different group of the audience may interpret it differently, Improper design can cause confusion.

107. What is the difference between a tree map and heat map?

Ans. A heat map is a type of visualisation tool that compares different categories with the help of colours and size. It can be used to compare two different measures. The 'tree map' is a chart type that illustrates hierarchical data or part-to-whole relationships.

108. What is disaggregation and aggregation of data?

Ans. Aggregation basically is combining multiple rows of data at a single place from low level to a higher level. Disaggregation, on the other hand, is the reverse process i.e breaking the aggregate data to a lower level.

109. What are some common data quality issues when dealing with Big Data?

Ans. Some of the major quality issues when dealing with big data are duplicate data, incomplete data, the inconsistent format of data, incorrect data, the volume of data(big data), no proper storage mechanism, etc.

110. What is clustering?

Ans. Clustering means dividing data points into a number of groups. The division is done in a way that all the data points in the same group are more similar to each other than the data points in other groups. A few types of clustering are Hierarchical clustering, K means clustering, Density-based clustering, Fuzzy clustering etc.

111. What are the data mining packages in R?

Ans. A few popular data mining packages in R are Dplyr- data manipulation, Ggplot2- data visualisation, purrr- data wrangling, Hmisc- data analysis, datapasta- data import etc.

112. What are techniques used for sampling? Advantage of sampling

There are various methods for drawing samples from data.

The two main Sampling techniques are

1. Probability sampling
2. Non-probability sampling

Probability sampling

Probability sampling means that each individual of the population has a possibility of being included in the sample. Probability sampling methods include –

- Simple random sampling

In simple random sampling, each individual of the population has an equivalent chance of being selected or included.

- Systematic sampling

Systematic sampling is very much similar to random sampling. The difference is just that instead of randomly generating numbers, in systematic sampling every individual of the population is assigned a number and are chosen at regular intervals.

- Stratified sampling

In stratified sampling, the population is split into sub-populations. It allows you to conclude more precise results by ensuring that every sub-population is represented in the sample.

- Cluster sampling

Cluster sampling also involves dividing the population into sub-populations, but each subpopulation should have analogous characteristics to that of the whole sample. Rather than sampling individuals from each subpopulation, you randomly select the entire subpopulation.

Non-probability sampling

In non-probability sampling, individuals are selected using non-random ways and not every individual has a possibility of being included in the sample.

- Convenience sampling

Convenience sampling is a method where data is collected from an easily accessible group.

- Voluntary Response sampling
- Voluntary Response sampling is similar to convenience sampling, but here instead of researchers choosing individuals and then contacting them, people or individuals volunteer themselves.
- Purposive sampling

Purposive sampling also known as judgmental sampling is where the researchers use their expertise to select a sample that is useful or relevant to the purpose of the research.

- Snowball sampling

Snowball sampling is used where the population is difficult to access. It can be used to recruit individuals via other individuals.

Advantages of Sampling

- Low cost advantage
- Easy to analyze by limited resources
- Less time than other techniques
- Scope is considered to be considerably high
- Sampled data is considered to be high
- Organizational convenience

113. What is imbalance data?

Imbalance data in simple words is a reference to different types of datasets where there is an uneven distribution of observations to the target class. Which means, one class label has higher observations than the other comparatively.

114. Define Lift, KPI, Robustness, Model fitting and DOE

Lift is used to understand the performance of a given targeting model in predicting performance, when compared against a randomly picked targeting model.

KPI or Key performance indicators is a yardstick used to measure the performance of an organization or an employee based on organizational objectives.

Robustness is a property that identifies the effectiveness of an algorithm when tested with a new independent dataset.

Model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained.

Design of Experiment (DOE) is a set of mathematical methods for process optimization and for quality by design (QbD).

115. Define Confounding Variables

A confounding variable is an external influence in an experiment. In simple words, these variables change the effect of a dependent and independent variable. A variable should satisfy below conditions to be a confounding variable :

- Variables should be correlated to the independent variable.
- Variables should be informally related to the dependent variable.

For example, if you are studying whether a lack of exercise has an effect on weight gain, then the lack of exercise is an independent variable and weight gain is a dependent variable. A confounder variable can be any other factor that has an effect on weight gain. Amount of food consumed, weather conditions etc. can be a confounding variable.

116. Why are time series problems different from other regression problems?

Time series is extrapolation whereas Regression is interpolation. Time-series refers to an organized chain of data. Time-series forecasts what comes next in the sequence. Time-series could be assisted with other series which can occur together.

Regression can be applied to Time-series problems as well as to non-ordered sequences which are termed as Features. While making a projection, new values of Features are presented and Regression calculates results for the target variable.

117. What is the difference between the Test set and validation set?

Test set : Test set is a set of examples used only to evaluate the performance of a fully specified classifier. In simple words, it is used to fit the parameters. It is used to test the data which is passed as input to your model.

Validation set : Validation set is a set of examples used to tune the parameters of a classifier. In simple words, it is used to tune the parameters. Validation set is used to validate the output which is produced by your model.

Kernel Trick

A Kernel Trick is a method where a linear classifier is used to solve non-linear problems. In other words, it is a method where a non-linear object is projected to a higher dimensional space to make it easier to categorize where the data would be divided linearly by a plane.

Let’s understand it better,

Let’s define a Kernel function K as xi and xj as just being the dot product.

$\mathbf{K(x_i,x_j) = x_i \, . \, x_j = x^T{}_i x_j}$

If every data point is mapped into the high-dimensional space via some transformation

$\mathbf{\Phi{:}x \text{ -> } \Phi(x)}$

The dot product becomes:

$\mathbf{K(x_i,x_j) = \Phi x^T{}_i \Phi x_j}$

Box Plot and Histograms

Box Plot and Histogram are types of charts that represent numerical data graphically. It is an easier way to visualize data. It makes it easier to compare characteristics of data between categories.