200 Interview Questions on DS

Python Programming

- What is the difference between list and tuple?
- What is the difference between append and extend?
- What is List Comprehension? Take a list of numbers from users, remove all odd numbers from the list.
- What is the lambda function and write one example?
- Take a string and find the first non-repeating character in it and return its index.
- What do you understand by oops? What is the purpose of using oops concepts?
- What is the function of self?
- What are the four main features of oops?
- What is Inheritance? Name the main types of python inheritance.

Python for Data Science

- What are the most common NumPy data types?
- How to change the data type of a NumPy array?
- What are the advantages of numpy over a regular python list?
- What is the easiest way to calculate percentiles when using Python?
- What are the different ways of creating DataFrame in pandas?
- What do you mean by data aggregation?
- What do you mean by pandas indexing?
- What is the best way to generate histograms in matplotlib?
- Can you provide me with an example of when a scatter graph would be more appropriate than a line chart?

Exploratory Data Analysis

- What is the significance of Exploratory data analysis?
- What is the difference between dependent and independent variables?
- What is the difference between Univariate, Bivariate and Multivariate analysis?
- How to perform Bivariate analysis Numerical-numerical, Categorical-Categorical, and Numerical-Categorical variables?
- During the data preprocessing step, how should one treat missing/null values? How will you deal with them?

- **What is an outlier and how to identify them?**

- **How is overfitting different from underfitting?**
- **What is multicollinearity? How can you solve it?**

Data Analysis using SQL

What are DDL and DML languages? Give an example.
Explain the effect of delete, drop and truncate statements?
What is the difference between Inner, Left, Right, Full outer join?
What is the difference between join and union?
Can multiple primary keys exist on a single table?
Explain the cases to use where versus having?
What is the difference between clustered and non-clustered indexes?
How would you extract the last four characters from a string?

Statistics

What is the difference between descriptive and inferential statistics?
What is the probability of throwing two fair dice when the sum is 5 and 8?
What are some of the properties of a Normal Distribution?
What is Skewness and Kurtosis?
What is the relationship between the confidence level and the significance level in statistics?
What is the Central Limit Theorem? Explain it. Why is it important?
How to screen outliers in a data set using z-score method and Interquartile Range(IQR).
What is Poisson Distribution and also explain the meaning of KPI in statistics?

Hypothesis Testing

1. **What is null hypothesis and alternate hypothesis?**
2. **How is the statistical significance of an insight assessed?**
3. **What is Type-I error and Type-II error?**
4. **What does a 95% confidence interval mean?**
5. **What is A/B testing?**
6. **How to decide if a hypothesis test is a One-Tailed Test or a Two Tailed Test?**
7. **What is Chi-Square Test for normality?**
8. **What is ANNOVA and how it is used?**

Linear Regression

What are the assumptions of Linear Regression?
What is the difference between R square and adjusted R square?
What are the disadvantages of linear model?
What is the difference between Ridge & Lasso Regression?
What will be the impact of multicollinearity on linear regression model?
How to find the multicollinearity?
What is MSE and RMSE? Why we calculate both of these?
If You have one independent variable. How many coefficient will you required to estimate in the simple linear regression model?
Why do we use optimization? Explain.

Logistic Regression

What is the difference between linear regression and logistic regression?
How will you deal with Categorical Independent Variable?
What are the key matrices used to check the performance of logistic regression?
What is multicollinearity?
How many kinds of techniques to fill the null values? Explain all of them.
What is the formula of the sigmoid function and why do we use it?
What is the issue of high dimensionality?
Why is logistic regression termed as regression and not classification?
What is a confusion matrix? Explain AUC and ROC curve.

Decision Trees

What type of node is consider pure?
What are the advantages of using Decision Tree?
What is the difference between gini impurity and information gain? Explain Entropy also.
What do you need to prune the decision tree?
What is the difference between post-pruning and pre-pruning?
Explain how ID3 produces classification trees?
What is the difference between ID3 and C4.5 algorithms?
How would you deal with an overfitted Decision tree?

Classification Techniques - CART

Explain how the CART algorithm performs the pruning?
Explain the measure of goodness used by CART.
Explain the difference between the CART and ID3 algorithm.
What are the limitations of CART?
What are the advantages of CART?

Support Vector Machines(SVMs)

11.1. What is a support vector machine (SVM) and how does it differ from other classification algorithms?
11.2. Can you provide an example of how SVM has been used in a real-world application?
11.3. How do you handle imbalanced datasets when using SVM?
11.4. Can you explain the concept of the kernel trick and how it is used in SVM?
11.5. What are some common challenges you have encountered when implementing SVM?
11.6. How do you choose the appropriate kernel function for an SVM model?
11.7. Can you describe your experience with non-linear SVM and how it differs from linear SVM?
11.8. How do you handle multi-class classification using SVM?
11.9. Can you explain the concept of support vectors and how they are used in SVM?
11.10. How do you tune hyperparameters in SVM?

Clustering

12.1. What is clustering and how does it differ from classification?
12.2. Can you provide an example of how clustering has been used in a real-world application?
12.3. What are some common clustering algorithms and how do they differ from each other?
12.4. How do you evaluate the effectiveness of a clustering algorithm?
12.5. Can you explain the concept of centroids and how they are used in clustering?
12.6. What are some common challenges you have encountered when implementing clustering algorithms?

12.7. How do you handle multi-dimensional data when performing clustering?
12.8. Can you describe your experience with hierarchical clustering and how it differs from k-means clustering?
12.9. How do you choose the appropriate number of clusters for a given dataset?
12.10. Can you explain the concept of cluster validity and how it is used to evaluate clustering algorithms?

Principal Component Analysis(PCA)

13.1. Can you explain what PCA is and how it works?
13.2. How is PCA related to eigenvectors and eigenvalues?
13.3. What is the goal of PCA and why is it used?
13.4. How does PCA reduce the dimensionality of a dataset?
13.5. Can you give an example of when PCA might be used in a real-world application?
13.6. How do you decide how many principal components to keep when using PCA?
13.7. How does PCA handle missing data or outliers in the dataset?
13.8. Can PCA be used with categorical data, or is it only applicable to numerical data?
13.9. How does PCA compare to other dimensionality reduction techniques, such as singular value decomposition (SVD) or independent component analysis (ICA)?
13.10. Can you discuss any potential limitations or drawbacks of using PCA?

Ensemble Modelling

14.1. Can you explain what ensemble modeling is and how it works?
14.2. How do ensemble models improve the performance of a model compared to using a single model?
14.3. Can you describe the different types of ensemble models and provide examples of when each might be used?
14.4. How do you decide which base models to include in an ensemble?
14.5. Can you discuss the trade-offs between using a more complex ensemble model versus a single model?
14.6. How do you evaluate the performance of an ensemble model?
14.7. Can you discuss any potential limitations or drawbacks of using ensemble modeling?

14.8. Can you give an example of when ensemble modeling has been successful in a real-world application?
14.9. How does ensemble modeling compare to other techniques for improving model performance, such as boosting or bagging?
14.10. Can you describe the process for training and tuning an ensemble model?

Time Series Analysis

What are the four main components of Time Series Analysis?
What is moving average?
What is Auto Regression?
Why does a time series have to be stationary?
What is an additive and multiplicative time series?
What is the difference between ARMA & ARIMA?
Can you explain RNN and LSTM, and when you use each for TSA?

BI Tools: Tableau

What is data visualization in Tableau?
What are filters? How many types of filters are there in tableau?
What is aggregation and disaggregation of data?
What is the benefit of the Tableau extract file over the live connection?
What is the difference between .twb and .twbx file extensions. Please explain.
What is the maximum number of tables that can be joined in Tableau?
What are sets and groups in Tableau?
What is the difference between joining and blending in Tableau?
What are parameters in Tableau?How do they work?
What is the difference between INDEX and RANK in Tableau?
What are dashboards in Tableau and why they are used?

Natural Language Processing and Speech

17.1. What is natural language processing (NLP) and how does it differ from traditional data processing?
17.2. Can you provide an example of how NLP has been used in a real-world application?
17.3. How do you handle missing or incomplete data when performing NLP tasks?
17.4. How do you pre-process text data before applying NLP techniques?
17.5. Can you explain the concept of tokenization in NLP and how it is used?
17.6. What are some common techniques for stemming and lemmatization in NLP?

17.7. How do you perform Part-of-Speech (POS) tagging and named entity recognition (NER)?
17.8. Can you explain the concept of machine translation and how it is used in NLP?
17.9. What are some common challenges you have encountered when performing NLP tasks on text data?
17.10. Can you describe your experience with speech recognition and how it is used in NLP?

Text Mining and Sentiment Analysis

18.1. What is text mining and how does it differ from traditional data mining?
18.2. Can you provide an example of how text mining has been used in a real-world application?
18.3. How do you handle missing or incomplete data when performing text mining?
18.4. How do you pre-process text data before applying text mining techniques?
18.5. Can you explain the concept of sentiment analysis in text mining and how it is used?
18.6. What are some common challenges you have encountered when performing sentiment analysis?
18.7. How do you measure the effectiveness of a text mining model?
18.8. Can you explain the concept of word embeddings and how they are used in text mining?
18.9. What are some common techniques for identifying and extracting features from text data?
18.10. How do you handle large amounts of text data when performing text mining?

Reinforcement Learning

19.1. What is reinforcement learning and how does it differ from supervised and unsupervised learning?
19.2. Can you provide an example of how reinforcement learning has been used in a real-world application?
19.3. How do you define the reward function in reinforcement learning?
19.4. Can you explain the concept of the action-value function and how it is used in reinforcement learning?
19.5. What are some common challenges you have encountered when implementing reinforcement learning algorithms?
19.6. How do you handle exploration-exploitation trade-offs in reinforcement learning?
19.7. Can you explain the concept of temporal difference learning and how it is used in reinforcement learning?

19.8. How do you handle off-policy learning in reinforcement learning?
19.9. Can you describe your experience with model-based reinforcement learning and how it differs from model-free reinforcement learning?
19.10. How do you handle delayed rewards in reinforcement learning?

Introduction to AI & Deep Learning

What is a perceptron?
How is Deep Learning better than machine learning?
What are activation functions?
What is the use of loss functions?
What is backpropagation and forward propagation?
What are hyperparameters in Deep Learning?
Why do we use dropout in deep learning?
What is the difference between ANN and CNN?

Advanced Deep Learning and Computer Vision

What is the vanishing gradient?
What is early stopping?
What is transfer learning and what are the applications of transfer learning?
What is the meaning of bagging and boosting in deep learning?
Can you define “Digital Image”?
What is the purpose of grayscaling?
Can you explain a scenario where you might use the anchor box?
Can you explain what the mach band effect is?

BI Tools: Google Data Studio

What is Google Data Studio?
How does google data studio work?
Why do you think Google Data Studio is better than other similar solutions like Tableau, PowerBI or Looker?
What is the difference between a table chart and a pivot table in Google Data Studio?
What are the differences between filters and control widgets in Google Data Studio?
What are some best practices for developing reports in Google Data Studio?

BI Tools: Power BI

What is PowerBI and why should we use it?

Difference between PowerBI and Tableau.
Difference between Power Query and Power Pivot.
What is PowerBI Desktop?
What is Power Query?
What is DAX?
What are the variety of filters available in Power BI?
Name the different connectivity modes available in Power BI?
What is a Dashboard in Power BI?

Getting started with R

What is R? Name the different data structures in R.
How to install a package in R?
Compare R & Python?
Can we store mixed data types in a R Vector?
What are the R packages used in Data Imputation?
Which function is used for adding datasets in R?
What are some advantages and disadvantages of using R?
What is t-tests() in R?