DATA SCIENCE & ANALYTICS ESSENTIALS (2022) INTERVIEW QUESTIONS

Search...

1. How do you create a PivotTable in Excel?

To create a PivotTable, select the data range, go to the Insert tab, click on PivotTable, choose the location for the PivotTable, and drag fields into the Rows and Values areas to summarize data.

2. How can you use conditional formatting in Excel?

Conditional formatting allows you to format cells based on specified conditions. Select the range, go to the Home tab, and choose Conditional Formatting options, such as highlighting cells with specific values or using color scales.

3. What is DAX, and how is it used in Power BI?

DAX (Data Analysis Expressions) is a formula language used in Power BI for creating custom calculations and aggregations. It is used to create calculated columns and measures.

4. How do you filter data in a Pandas DataFrame?

You can filter data in a Pandas DataFrame using conditional statements or boolean indexing. For example, df[df['Column'] > 50] filters rows where the value in the 'Column' is greater than 50.

5. What is linear regression, and how does it work?

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It works by finding the best-fit line that minimizes the sum of squared differences between observed and predicted values.

6. Explain the concept of random forest.

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

7. How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.

There are two useful methods in Pandas:

- IsNull() and dropna() will help to find the columns/rows with missing data and drop them
- Fillna() will replace the wrong values with a placeholder value

8. What is Semi-supervised Machine Learning?

Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data.

In the case of semi-supervised learning, the training data contains a small amount of labeled data and a large amount of unlabeled data.

9. Compare K-means and KNN Algorithms.

K-means	KNN
• K-Means (https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm) is unsupervised • K-Means is a clustering algorithm • The points in each cluster are similar to each other, and each cluster is different from its neighboring clusters	• KNN (https://www.simplilearn.com/tutorials/machine-learning-tutorial/knn-in-python) is supervised in nature • KNN is a classification algorithm • It classifies an unlabeled observation based on its K (can be any number) surrounding neighbors

10. What is a Random Forest?

A ‘ (https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm)Random Forest’ is a supervised machine learning algorithm that is generally used for classification problems. It operates by constructing multiple decision trees during the training phase. The random forest chooses the decision of the majority of the trees as the final decision.