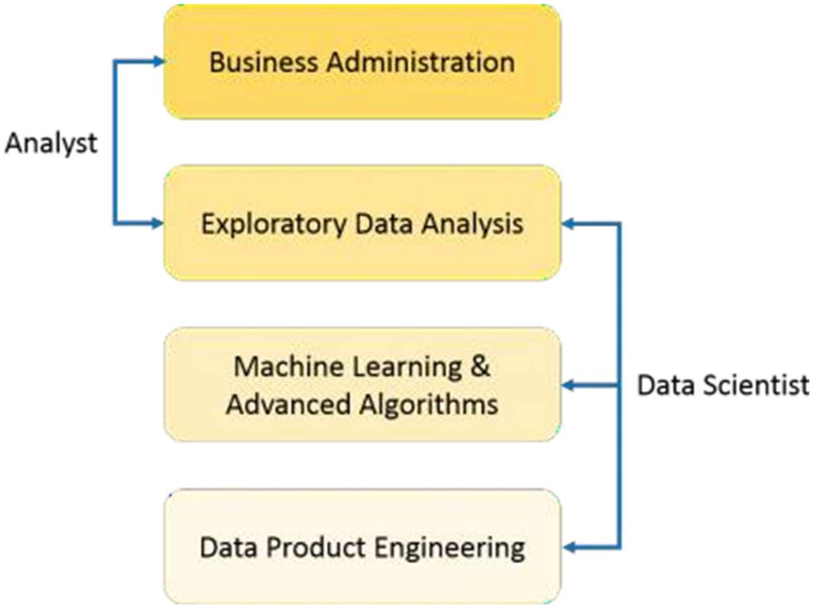
Data Science :

Q1. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years? The answer lies in the difference between explaining and predicting.



The differences between supervised and unsupervised learning are as follows;

Supervised Learning	Unsupervised Learning
Input data is labelled.	Input data is unlabelled.
Uses a training data set.	Uses the input data set.
Used for prediction.	Used for analysis.
Enables classification and regression.	Enables Classification, Density Estimation, & Dimension Reduction

Q2. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.
The types of selection bias include:

1. **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2. **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
3. **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.

Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

3. If you are having 4GB RAM in your machine and you want to train your model on 10GB data set. How would you go about this problem? Have you ever faced this kind of problem in your machine learning/data science experience so far?

First of all, you have to ask which ML model you want to train.

For Neural networks: Batch size with Numpy array will work.

Steps:

1. Load the whole data in the Numpy array. Numpy array has a property to create a mapping of the complete data set, it doesn't load complete data set in memory.
2. You can pass an index to Numpy array to get required data.
3. Use this data to pass to the Neural network.

4. Have a small batch size.

For SVM: Partial fit will work

Steps:

1. Divide one big data set in small size data sets.
2. Use a partial fit method of SVM, it requires a subset of the complete data set.
3. Repeat step 2 for other subsets.

However, you could actually face such an issue in reality. So, you could check out the ***best laptop for Machine Learning*** to prevent that

4. **Explain machine learning to me like a 5-year-old.**

It's simple. It's just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bend position. The next time they fall down, they feel pain. They cry. But, they learn 'not to stand like that again'. In order to avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.

This is how a machine works & develops intuition from its environment.

5. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn't. How ?

Using one hot encoding, the dimensionality (with perspective of features) in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let's say we have a variable 'color'. The variable has 3 levels namely Red, Blue and Green. One hot encoding 'color' variable will generate three new variables as `Color.Red`, `Color.Blue` and `Color.Green` containing 0 and 1 value.

In label encoding, the levels of a categorical variables gets encoded as 0 and 1, so no new variable is created. Label encoding is majorly used for binary variables.

6. What Is the Cost Function?

Also referred to as “loss” or “error,” cost function is a measure to evaluate how good your model’s performance is. It’s used to compute the error of the output layer during backpropagation. We push that error backwards through the neural network and use that during the different training functions.

7. What Are Hyperparameters?

With Machine Learning, you’re usually working with ***hyperparameters*** once the data is formatted correctly. A hyperparameter is a parameter whose value is set before the learning process begins. It determines how a model is trained and the structure of the network.

8. Name a few ***Machine Learning libraries*** for various purposes.

Purpose	Libraries
Scientific Computation	Numpy
Tabular Data	Pandas
Data Modelling & Preprocessing	Scikit Learn
Time-Series Analysis	Statsmodels
Text processing	Regular Expressions, NLTK
Deep Learning	Tensorflow, Pytorch

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

RMSE and MSE are two of the most common measures of accuracy for a linear regression model.

RMSE indicates the Root Mean Square Error.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

MSE indicates the Mean Square Error.

$$MSE=\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}$$

10. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that's easy to analyze.

11.What are recommender systems?

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

12. What is selection bias?

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

13.What are the types of biases that can occur during sampling?

1. Selection bias
2. Undercoverage bias
3. Survivorship bias

14. What is Skewness in data (Right and Left) ?

It measures the degree of symmetry
Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of **data**. If the curve is shifted to the **left** or to the **right**, it is said to be skewed. **Skewness** can be quantified as a representation of the extent to which a given distribution varies from a normal distribution

15. What are types of statistics Descriptive and Inferential Statistics ?

Descriptive statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question

16. What is Multi-Collinearity ?

When the independent variables depend on each other

17.What is Hetroskedacity and Homskedacity ?

Homskedacity - if all its random variables have the same finite variance.
Hetroskedacity - the variance of the residual term, or error term, in a regression model varies widely

18. Role Of Interaction ?

Interaction effects occur when the effect of one variable depends on the value of another variable.

19. What is Generalization and Regularization?

Regularization is used to control over fitting
Generalization is used to control under fitting

20. What is Curse of Dimensionality?

Curse of Dimensionality refers to non-intuitive properties of data observed when working in high-dimensional space.

21. What is training error and testing error ?

Training error is the error that you get when you run the trained model back on the training data. Remember that this data has already been used to train the model and this necessarily doesn't mean that the model once trained will accurately perform when applied back on the training data itself.