1) What is the difference between Data Science, Machine Learning and Artificial Intelligence?
Ans:

	Data Science	**Machine Learning**	**Artificial Intelligence**
Definition	Data Science is not exactly a subset of machine learning, but it uses machine learning to analyze and make future predictions.	A subset of AI that focuses on narrow range of activities.	A wide term that focuses on applications ranging from Robotics to Text Analysis.
Role	It can take on a business role.	It is a purely technical role.	It is a combination of both business and technical aspects.
Scope	Data Science is a broad term for diverse disciplines and is not merely about developing and training models.	Machine learning fits within the data science spectrum.	AI is a sub-field of computer science.
AI	Loosely integrated	Machine learning is a sub field of AI and is tightly integrated.	A sub- field of computer science consisting of various task like planning, moving around in the world, recognizing objects and sounds, speaking, translating, performing social or business transactions, creative work..

2) Python or R – Which is the best programming language for Text Analytics?

Ans: I prefer Python because it has Pandas library that provides easy to use data structures and high-performance data analysis tools. Python also provides packages such NLTK and SCAPY that helps us performing text analytics with ease.

3) Which technique is used to predict categorical responses?

Ans: Classification technique is used widely in mining for classifying data sets.

4) **What is logistic regression? State an example when you have used logistic regression recently?**

Ans: Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a candidate, the amount of time spent in campaigning, etc.

5) **What are Recommender Systems?**

Ans: A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

6) **Why data cleaning plays a vital role in analysis?**

Ans: Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

7) **Differentiate between univariate, bivariate and multivariate analysis.**

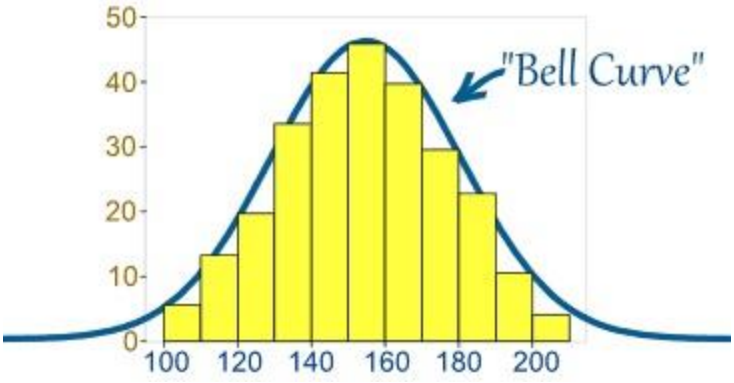
Ans: These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as **univariate analysis**.

If the analysis attempts to understand the difference between 2 variables at time as in a scatterplot, then it is referred to as bivariate analysis. For example, analyzing the volume of sale and a spending can be considered as an example of **bivariate analysis**.

Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as **multivariate analysis**.

8) **What do you understand by the term Normal Distribution?**

Ans: Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve. The random variables are distributed in the form of a symmetrical bell-shaped curve.



Bell Curve

9) What is Linear Regression?

Ans: Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable

10) What is Interpolation and Extrapolation?

Ans: Estimating a value from 2 known values from a list of values is Interpolation. Extrapolation is approximating a value by extending a known set of values or facts.

11) What is power analysis?

Ans: An experimental design technique for determining the effect of a given sample size.

12) What is K-means? How can you select K for K-means?

Ans: K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K

There is a popular method known as **elbow** method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm

13) What does p-value signify about the statistical data?

Ans: p-value is used determine the significance of results after a hypothesis test in statistics. **p-value** helps us draw conclusions and **p-value** is always between 0 and 1

- **p-value > 0.05** represents weak evidence against the null hypothesis i.e. the null hypothesis **can't be rejected**
- **p-value <= 0.05** represents strong evidence against the null hypothesis i.e. the null hypothesis **can be rejected**
- **p-value = 0.05** this is marginal value that indicates that it will go either way

14) What are categorical variables?

Ans: Categorical variables are also called **qualitative** variables. Any variable / feature that can be classified into names / labels. Example: color of a pen – red, blue, green etc.

15) What is the difference between Supervised Learning an Unsupervised Learning?

Ans: If an algorithm learns something from the training data so that the knowledge can be applied to the test data, then it is referred to as Supervised Learning. Classification is an example for Supervised Learning. If the algorithm does not learn anything beforehand because there is no response variable or any training data, then it is referred to as unsupervised learning. Clustering is an example for unsupervised learning.

Supervised learning is simply a process of learning algorithm from the training dataset. Supervised learning is where you have input variables and an output variable, and you use an algorithm to learn the mapping function from the input to the output. The aim is to approximate the mapping function so that when we have new input data we can predict the output variables for that data.

Unsupervised learning is modeling the underlying or hidden structure or distribution in the data in order to learn more about the data. Unsupervised learning is where you only have input data and no corresponding output variables.

16) What is an Eigenvalue and Eigenvector?

Ans: Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the **eigenvectors** for a correlation or covariance matrix. **Eigenvectors** are the directions along which a linear transformation acts by flipping, compressing or stretching. **Eigenvalue** can be referred to as the strength of the transformation in the direction of **eigenvector** or the factor by which the compression occurs.

Eigenvectors and **eigenvalues** are used to reduce noise in data. They can help us improve efficiency in computationally intensive tasks. They also eliminate features that have a strong correlation between them and help in reducing over-fitting

17) What is Gradient Descent?

Ans: Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). Gradient descent is best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm

Gradient descent is by far the most popular optimization strategy, used in machine learning and deep learning. It is used while training your model, can be combined with every algorithm and is easy to understand and implement. Therefore, everyone who works with Machine Learning should understand its concept.

18) What are outliers? How Outliers can be treated?

Ans: Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values:

- To change the value and bring in within a range
- To just remove the value.

19) How do we deal with missing values?

Ans: We must identify the patterns with regards to the missing data. Here are the data missing patterns:

- **Missing at Random (MAR):** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data

- **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables

- **Missing not at Random (MNAR):** Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So, we have to be really careful before removing observations. Note that imputation does not necessarily give better results

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. There are various factors to be considered when answering this question-

Understand the problem statement, understand the data and then give the answer. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.
Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

20) What is multicollinearity and how you can overcome it?

Ans: Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

In regression, "multicollinearity" refers to predictors that are correlated with other predictors. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other

We can take any of the below mentioned approaches in dealing with **Multicollinearity**

- ⇨ **Remove highly correlated predictors from the model**
- ⇨ **Use Partial Least Squares Regression (PLS)** or **Principal Component Analysis (PCA)** methods to cut down the number of predictors to a smaller set of uncorrelated components

21) How do you decide whether your linear regression model fits the data?

Ans: We have several parameters using which we can evaluate the fitness of the Regression model. Here is the list of common metrics / parameters used:

- ⇨ R-Squared or coefficient of determination
- ⇨ Adjusted R-Squared
- ⇨ Standard Error
- ⇨ F Statistics
- ⇨ t statistics

22) What is the difference between skewed and uniform distribution?

Ans: When the observations in a dataset are spread equally across the range of distribution, then it is referred to as uniform distribution. There are no clear perks in a uniform distribution. Distributions that have more observations on one side of the graph than the other are referred to as skewed distribution. Distributions with fewer observations on the left (towards lower values) are said to be skewed left and distributions with fewer observation on the right (towards higher values) are said to be skewed right.

23) **What are Linear Regression Assumptions?**

Ans: LINE is the easy way to recall the assumptions

Linear relation exists between X and Y
Independent of each other
Normally distributed
Equal variance among variables

24) **What is the difference between overfitting and underfitting?**

Ans: In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, to be able to make reliable predictions on general untrained data.

In **overfitting**, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

25) How does data cleaning play a vital role in analysis?

Ans:
Data cleaning can help in analysis because:

- ⇨ Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- ⇨ Data Cleaning helps to increase the accuracy of the model in machine learning. It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
- ⇨ It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.