1. **What do you mean by data science? Explain its scope and limitation. What are the common misconceptions about data science.**

Data science is the study that deals with high volumes of data using modern tools and techniques to find unseen patterns and meaningful information to make business decisions. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, computer engineering and domain specific expert to analyze large amounts of data.

Scope of Data Science

Data driven technology is rapidly advancing and data is becoming more valuable. So data science future scope looks promising. Some of the career scope of the data science are:

1) Data Engineer
2) Data Architect
3) Data Manager
4) Data Scientist
5) Data Analyst
6) ML Engineer
7) AI Engineer

Limitations of Data Science

1) Data Quality
   Data is collected from different sources, but the collected data might not be reliable.
2) Data Quantity
   Data science requires large amount of data to
3) Domain Expertise
   Understanding the context of data within specific industries or fields is crucial. Not everyone can be an expert in multiple domains.

Common Misconceptions about Data Science

1) Data Science is Easy and Fun.
2) Data Science needs a large volume of data.
3) Data Science is the same as ML/AI Engineer.
4) Data Science requires just coding skills.
5) Data Science is a one man show.
6) Data Science is used only in big companies.

2. Who is data scientist? What are their roles and responsibilities?

A data scientist is a professional who knows how to collect and analyze large amounts of data using analytical, statistical, and programming skills and extract meaningful information and interpret them.

Roles and responsibilities of Data Scientist

1) Data Collection and Analysis:
   Data scientists fetch information from various sources, including databases, APIs, and raw files then analyze it to gain a clear understanding of how an organization performs.

2) Pattern Recognition and Insights:
   Data scientists identify patterns and trends within the data then extract valuable insights that can drive business decisions.

3) Domain Knowledge:
   Data scientists often specialize in specific domains (e.g., healthcare, finance, marketing). Understanding the context and nuances of the industry they work in is vital.

4) Continuous Learning:
   The field of data science is evolving rapidly. Data scientists need to stay updated with the latest techniques, tools, and trends.

3. **How does the roles and responsibilities of data scientists differ from Data Engineers and Data Analysts? Explain.**

	Data Engineer	Data Analyst	Data Science
Tools	Hadoop, Spark, Kafka	Tableau, Power BI, Google Analytics	Python, R
Objective	Data Collection, Data pipeline development, Data transformation	EDA, Visualization and Reporting	Discover trend and patterns for decision making
Responsibilities	ETL process, data storage, optimization	Data collection, cleaning, and visualization	Modeling, analyses, ML deployment
Goal	Create Data pipeline for continuous reliable data.	Analyze the insight	Understand the business problem and solve using data driven approach
Working Environment	Big data platform	Business Intelligence.	Should familiar with all the platform.

4. **Explain CRISP-DM lifecycle for Agile implementation in any data science project with any suitable example of your own.**

CRISP-DM (CRoss Industry Standard Process for Data Mining) is a widely used methodology for conducting data science projects. When combined with Agile principles, it provides a structured framework for iterative and adaptive development in data science projects It has six sequential phases:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

CRISP-DM (Cross-Industry Standard Process for Data Mining) lifecycle implemented for a spam email detection project with an Agile framework.

1) **Business Understanding**
   Conduct meetings to gather requirements and define user stories related to spam detection. Identify key metrics.
   Example: In our spam email detection project, the business objective is to reduce the number of spam emails
2) **Data Understanding**
   Explore available email datasets and assess data quality issues (e.g., missing values, noise). Continuously, refine data requirements based on feedback and insights from initial analysis.
   Example: Collects email datasets containing labeled examples of spam and non-spam emails. Then analyze the data to identify common features like keywords, sender information, and email structure.
3) **Data Preparation**
   Break down data preprocessing tasks into smaller user stories and address them iteratively. Prioritize tasks such as text cleaning, tokenization, and feature engineering based on their impact on spam detection.
   Example: Agile focuses on tasks like removing HTML tags from emails, sentimental analysis, tokenizing text into words, and extracting features like word frequencies or presence of specific keywords relevant to spam.
4) **Modeling**
   Experiment with different modeling approaches (e.g., Naive Bayes, Logistic Regression, Neural Networks) within Agile iterations. Continuously evaluate model performance and refine approaches based on feedback and validation results.
   Example: Agile involves building and testing spam detection models using supervised learning algorithms. Models are trained on labeled email data and iteratively improved based on validation metrics like accuracy and false positive rate.

5) **Evaluation**
   Conduct regular evaluations using test datasets and validation techniques (e.g., cross-validation). Gather feedback from users and stakeholders to assess model effectiveness and identify areas for improvement.
   Example: Evaluating model performance metrics such as accuracy, precision, recall and F1-score. Stakeholders provide feedback on the model's ability to accurately classify spam and non-spam emails.

6) **Deployment**
   Plan deployment tasks in Agile release cycles, ensuring seamless integration with email servers and user interfaces. Monitor model performance post-deployment and iterate on deployment processes as needed based on user feedback.
   Example: Focus on deploying the trained spam detection model into the email server infrastructure. The team collaborates with IT operations to integrate the model and monitors its performance to ensure accurate spam filtering.

5. **Perform a case study on TDSP Lifecycle for data science.**

The TDSP lifecycle is a structured framework for managing data science projects, particularly those that involve building and deploying machine learning models. It emphasizes collaboration, communication, and an iterative approach to ensure successful project outcomes

Case Study: Fraud Detection in Commercial Bank Transactions using TDSP Lifecycle
A commercial bank is experiencing an increase in fraudulent transactions. So, the aim is to build a data science model to detect fraudulent activities in real-time and prevent financial losses.

The TDSP lifecycle can be a valuable guide for this project:

1) **Business Understanding**
   Understand the objectives to reduce fraudulent transactions by 80% within 6 months. Data scientist, fraud investigation team, security team.
2) **Data Acquisition and Understanding**
   Securely collect transaction data, customer data, and external fraud indicators (e.g., blacklisted accounts). Analyze transaction data to identify patterns and characteristics typical of fraudulent activities. Address missing information, outliers, and potential inconsistencies in the data. Perform feature engineering to create new features based on transaction behavior (e.g., unusual purchase locations, sudden spikes in spending).
3) **Modeling**
   Choose appropriate machine learning models for anomaly detection and train the models on historical transaction data labeled as fraudulent or legitimate. Optimize the models to minimize false positives (flagging legitimate transactions as fraud) and false negatives (missing actual fraud). Create Confusion matrix and evaluate the accuracy, precision, recall of the model.
4) **Deployment**
   Choose the model with the best balance of accuracy and low false positive rate. Integrate the model into the bank's transaction processing system for real-time fraud detection. Establish an alert system to notify fraud investigators of suspicious transactions flagged by the model.