Unit 1: Overview of Data Science

Short Answer Questions

1. Define Data Science.

Data Science is the art of using scientific methods, programming skills, and knowledge of statistics to find meaningful stories and insights from data.

Real-Life Example: Think of a detective solving a case. The detective (a data scientist) collects clues (data) from a crime scene, analyzes them for patterns, and uses them to figure out what happened (gain insights). Just like the detective, data science helps us understand what the data is telling us.

2. Differentiate between Data Science and Machine Learning.

Data Science is the entire process of collecting, cleaning, analyzing, and getting insights from data. Machine Learning is a tool used within data science that helps computers learn from data to make predictions automatically.

Real-Life Example: Imagine building a custom car.

Data Science is the whole project: designing the car, sourcing the parts, assembling it, and painting it.

Machine Learning is like the powerful engine you choose to put in the car. It's a critical component, but not the whole car itself.

3. What are the three types of data: structured, unstructured, and semi-structured? Give examples.

Structured Data: Highly organized data that fits neatly into rows and columns, like a spreadsheet.

Example: An Excel sheet of student information with columns for Roll Number, Name, and Marks.

Unstructured Data: Data that has no specific format or organization.

Example: The text of a social media post, a photo, a video, or an audio recording.

Semi-structured Data: A mix of both. It isn't in a strict table format but contains tags or markers to separate elements.

Example: An email. It has some structure (To, From, Subject fields) but also unstructured content (the body of the email).

4. List any two key benefits of Data Science in business.

Better Decision Making: Businesses can stop guessing and start making decisions based on facts and patterns found in their data.

Example: A retail store can analyze past sales data to decide which products to stock up on for an upcoming festival.

Personalized Customer Experience: Companies can understand what individual customers like and offer them tailored recommendations.

Example: Netflix analyzes your viewing history to suggest movies and TV shows you are likely to enjoy.

5. Who first used the term "Data Science" and in which year?

Peter Naur, a Danish computer science pioneer, first used the term "Data Science" in 1974.

Long Answer Questions (5 Marks Each)

6. Explain the evolution of Data Science from the 1960s to the present AI era.

The journey of Data Science has been a story of growing with technology. It evolved over several decades from simple data analysis to the complex, AI-driven field it is today.

1960s-1970s (Early Beginnings): Before "Data Science" was a common term, this era was about basic statistics and the birth of relational databases (like simple digital filing cabinets). The focus was on storing and retrieving data efficiently.

1990s (The Rise of Data Mining): With more data being collected, "Data Mining" became popular. The goal was to dig through large datasets to find hidden patterns. Think of it as panning for gold in a river of information.

2000s (The Big Data Explosion): The internet boom created a massive amount of data, known as "Big Data." Traditional tools couldn't handle this volume. This is when Data Science became a formal field, using new technologies like Hadoop and advanced machine learning to make sense of huge datasets.

2010s-Present (The AI and Deep Learning Era): Today, Data Science is powered by Artificial Intelligence (AI) and Deep Learning. We now have complex models that can understand images, text, and speech. This has led to innovations like self-driving cars, virtual assistants, and powerful recommendation engines. The focus has shifted from just analyzing the past to predicting the future and automating complex tasks.

7. Compare Data Mining, Data Analysis, and Data Analytics.

While these terms are often used together, they refer to different stages of working with data. Let's use a detective analogy to understand the difference.

Term	What It Is	Goal	Detective Analogy
Data Mining	The process of automatically discovering hidden patterns and relationships in large datasets using algorithms.	To find previously unknown and useful information.	The detective uses special tools (like fingerprint analysis) to find a hidden clue that no one saw at first glance.
Data Analysis	The process of inspecting, cleaning, transforming, and modeling data. It's a more hands-on process.	To discover useful information and support decision-making.	The detective takes all the clues, organizes them on a board, and starts connecting the dots to form a theory.

Term	What It Is	Goal	Detective Analogy
Data Analytics	A broader term that encompasses the entire process of managing and analyzing data, from collection to generating reports. It focuses on using specialized systems to draw conclusions.	To make informed business decisions by examining data.	This is the detective's whole investigation, from securing the crime scene to presenting the final conclusions in court.

Export to Sheets
In short, Data Mining finds the hidden clues, Data Analysis connects the clues, and Data Analytics is the entire investigation from start to finish.

8. Describe the components of Data Science (Data, Big Data, Machine Learning, Statistics, Programming).

Data Science is like a complex recipe with five essential ingredients. Each component plays a vital role in creating the final result: valuable insights.

Data: This is the foundation. It can be anything from customer details in a spreadsheet (structured) to photos on Instagram (unstructured). Without data, there is no science to be done.

Big Data: This refers to datasets that are so large and complex that traditional data-processing tools can't handle them. It's characterized by its high volume, velocity (speed), and variety. Think of the massive amount of data generated by social media every second.

Machine Learning: This is the "brain" of data science. It involves algorithms that learn patterns from data without being explicitly programmed. It's used to make predictions, such as forecasting sales or identifying spam emails.

Statistics & Probability: This is the scientific backbone. Statistics helps us summarize data (like finding the average) and make inferences, while probability helps us understand uncertainty and predict future events. It ensures our conclusions are mathematically sound.

Programming: This is the "hands" that do the work. Programming languages like Python and R are used to clean, process, analyze, and visualize data. They are the tools that bring all the other components together.

9. Explain the Data Science Process with steps.

The Data Science Process is a step-by-step lifecycle for turning raw data into actionable insights. It’s like following a recipe to bake a cake.

Frame the Problem (Understand the Goal): This is the most important step. Before you start, you must understand the business problem you are trying to solve. What question are we trying to answer? For example, "How can we reduce the number of customers leaving our service?"

Collect Raw Data: Once you know the problem, you gather all the relevant data you need. This data can come from databases, web scraping, surveys, etc.

Process and Clean the Data: Raw data is often messy. It might have missing values, duplicates, or errors. In this step, you clean and format the data to make it ready for analysis. This is like washing and chopping vegetables before cooking.

Explore the Data (Exploratory Data Analysis): Here, you dive deep into the data to find initial patterns and relationships. You use charts and graphs (data visualization) to understand the data's story. It's like a first look at your ingredients to see what you're working with.

Perform In-depth Analysis (Model Building): In this step, you use machine learning algorithms to build predictive models. For our example, you might build a model that predicts which customers are most likely to leave.

Communicate Results: Finally, you present your findings to stakeholders (like managers) in a clear and understandable way, often using reports and visualizations. The goal is to turn your insights into actions that solve the original business problem.

10. Discuss real-world applications of Data Science in healthcare, e-commerce, and logistics.

Data Science is transforming industries by turning data into powerful actions. Here's how it's used in three key sectors:

Healthcare:

- **Medical Image Analysis:** Data Science helps in analyzing medical images like X-rays and MRIs to detect diseases like cancer at a very early stage, often more accurately than the human eye.
- **Drug Discovery:** By analyzing biological data, data science can speed up the process of developing new drugs, making it faster and cheaper to find cures for diseases.
- **Predictive Analytics:** Hospitals can predict patient admissions and disease outbreaks, helping them manage resources like staff and beds more effectively.

E-commerce:

- **Recommendation Engines:** Websites like Amazon and Flipkart use data science to analyze your past purchases and browsing history. This allows them to create a personalized "Recommended for you" section, which significantly boosts sales.
- **Price Optimization:** Companies can use data science to set dynamic prices for products based on demand, competitor prices, and customer behavior to maximize profit.

Logistics:

- **Route Optimization:** Companies like FedEx and DHL use data science to find the fastest and most fuel-efficient routes for their delivery vehicles. This saves time and money and also considers factors like traffic and weather.
- **Demand Forecasting:** Logistics companies can predict when and where demand for shipments will be high, allowing them to allocate their trucks and staff efficiently, especially during peak seasons.

11. How does Data Science help in driverless cars to make driving decisions?

A driverless car uses Data Science to see, think, and act like a human driver, but much faster.

Seeing (Data Collection): The car is equipped with sensors, cameras, and LiDAR that constantly collect huge amounts of data about its surroundings—other cars, pedestrians, road signs, and lane markings.

Thinking (Data Processing & Modeling): This is where data science comes in. Complex machine learning models, trained on millions of miles of driving data, process this real-time information. They perform tasks like:

- **Object Recognition:** Identifying a pedestrian from a bicycle.
- **Path Prediction:** Predicting where that pedestrian is likely to move next.
- **Decision Making:** Based on these predictions, the model decides the safest action—whether to slow down, stop, change lanes, or continue.

Acting (Execution): The model's decision is instantly sent to the car's controls to perform the action (e.g., apply the brakes).

12. Explain the role of Data Science in stock market predictions.

Data Science helps traders move from gut feelings to data-driven strategies for stock market predictions.

- **Data Collection:** It starts by gathering vast amounts of historical data, including stock prices, trading volumes, company financial reports, and even news articles.
- **Pattern Recognition:** Machine learning algorithms are used to analyze this data to find complex patterns that are impossible for humans to see. For example, how a particular piece of news affects a company's stock price.
- **Sentiment Analysis:** Data science can analyze news articles and social media posts to gauge public sentiment (positive or negative) about a stock, which can be a powerful predictor of price movements.
- **Predictive Modeling:** Based on all this data, a predictive model is built to forecast whether a stock's price is likely to go up or down in the future, helping investors make more informed decisions.

13. How is Data Science used for image recognition in social media platforms like Facebook?

When you upload a photo to Facebook, its ability to suggest tagging your friends is a direct application of Data Science and Deep Learning.

- **Training the Model:** Facebook has trained its image recognition model on billions of photos that users have already tagged. The model learns to associate specific facial features with specific people.
- **Face Detection:** When you upload a new photo, the system first detects all the faces in the image.
- **Feature Analysis:** For each face, it creates a unique digital signature by analyzing key features like the distance between the eyes, the shape of the nose, and the jawline.
- **Matching and Suggestion:** It then compares this signature with the signatures of your friends in its database. If it finds a close match, it suggests a tag. The more you tag, the smarter the model gets.

14. Suggest how an e-commerce website can use Data Science to improve customer experience.

An e-commerce website can use Data Science to make shopping feel personal, smart, and easy for every customer.

- **Personalized Recommendations:** Analyze a user's browsing history, past purchases, and even items they've added to their cart to create a personalized homepage and product recommendations. For example, "Because you watched The Avengers, you might like..."

- **Smarter Search Results:** Instead of just matching keywords, use Data Science to understand the user's intent. If someone searches for "summer dress," show them popular, highly-rated summer dresses, not just any item with "dress" in the name.
- **Customer Segmentation:** Group customers into different segments based on their buying habits (e.g., "frequent buyers," "deal seekers"). This allows the website to send targeted marketing emails and offers that are more likely to be relevant.
- **Sentiment Analysis on Reviews:** Automatically analyze thousands of product reviews to understand what customers love or hate about a product. This feedback can be used to improve products and provide better descriptions on the site.

Unit 2: Mathematics and Statistics in Data Science

Short Answer Questions

1. Define Linear Algebra and explain its role in Data Science.

Linear Algebra is a branch of math that deals with vectors, matrices (grids of numbers), and the rules for transforming them.

Role in Data Science: In data science, all data (like images, text, or tables) is converted into numbers. Linear algebra provides the tools to work with these numbers efficiently. It's the engine that powers many machine learning algorithms.

2. What is a vector? Give an example.

A vector is a one-dimensional list or array of numbers. You can think of it as a single row or column of data that represents a point in space.

Real-Life Example: Imagine you are describing a house. A vector could represent its key features: `[number of bedrooms, number of bathrooms, square footage]`. For a specific house, this might look like: `[3, 2, 1500]`.

3. Differentiate between matrix addition and matrix multiplication.

Feature	Matrix Addition	Matrix Multiplication
How it Works	You add the corresponding elements in each position. The matrices must be the same size.	It's a "row-by-column" operation. The number of columns in the first matrix must equal the number of rows in the second.
Analogy	Like adding two shopping lists together item by item.	Like calculating the total cost of items from different stores, where prices and quantities are in separate grids.

4. Define determinant. Why is it important in linear algebra?

A determinant is a special number that can be calculated from a square matrix.

Importance: It tells us important properties about the matrix. For example, if the determinant is zero, it means the transformations described by the matrix are "squashing" data into a smaller dimension, and the system of equations it represents might not have a unique solution.

5. What is meant by a vector space? List its properties.

A vector space is a collection of all possible vectors of a certain type, where you can perform two basic operations: adding vectors together and scaling them (multiplying by a number).

Properties: It must satisfy rules like commutativity (A + B = B + A), associativity, having a zero vector, and having additive inverses.

6. Differentiate between descriptive and inferential statistics.

- **Descriptive Statistics:** This is about summarizing and organizing data so it's easy to understand. It describes what the data shows.
  - **Example:** Calculating the average score (mean) and the most common score (mode) for a class of students. You are simply describing the class's performance.
- **Inferential Statistics:** This is about using data from a small sample to make educated guesses (inferences) or predictions about a larger population.
  - **Example:** Surveying 100 people in a city to predict who will win a city-wide election. You are using a small group to infer about the whole city.

7. Define mean, median, and mode with examples.

- **Mean:** The average of all the numbers. You add them all up and divide by how many there are.
  - **Example:** For scores `[10, 20, 30]`, the mean is `(10+20+30)/3 = 20`.
- **Median:** The middle value when the numbers are sorted.
  - **Example:** For `[1, 3, 5, 8, 9]`, the median is `5`. For `[1, 3, 5, 8]`, the median is the average of the two middle numbers: `(3+5)/2 = 4`.
- **Mode:** The number that appears most often.
  - **Example:** For `[1, 5, 5, 8, 9]`, the mode is `5`.

8. What is variance? How is it related to standard deviation?

Variance measures how spread out the data points are from the average (mean). A high variance means the data is widely scattered.

Relation to Standard Deviation: The **Standard Deviation** is simply the **square root of the variance**. It's often easier to interpret because it's in the same unit as the original data.

Analogy: If variance is the area of a square, standard deviation is the length of one of its sides.

9. Define probability with an example.

Probability is a measure of how likely an event is to happen. It's a number between 0 (impossible) and 1 (certain).

Formula: Probability = (Number of favorable outcomes) / (Total number of possible outcomes)

Example: The probability of rolling a 4 on a single six-sided die is 1/6, because there is only one "4" and there are six possible sides.

10. What is conditional probability?

Conditional probability is the likelihood of an event happening, given that another event has already occurred.

Real-Life Example: The probability of the ground being wet is low on its own. But the *conditional probability* of the ground being wet, *given that it has just rained*, is very high.

Long Answer Questions (5 Marks Each)

11. Explain different matrix operations with examples. (Corrected)

Matrices are fundamental in data science, and we can perform several key operations on them. Let's use two simple matrices, A and B, for our examples:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$

1. **Addition and Subtraction:** You can add or subtract matrices only if they have the same dimensions. The operation is done by adding or subtracting the elements in the corresponding positions.
   - **Example (Addition):**

     $$A + B = \begin{pmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

   - **Example (Subtraction):**

     $$A - B = \begin{pmatrix} 1-5 & 2-6 \\ 3-7 & 4-8 \end{pmatrix} = \begin{pmatrix} -4 & -4 \\ -4 & -4 \end{pmatrix}$$

2. **Scalar Multiplication:** This means multiplying every single element of a matrix by one number (a scalar).
   - **Example:** Let's multiply matrix A by the scalar 3:

     $$3A = 3 \times \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix}$$

$$\begin{pmatrix} 3\times1 & 3\times2 \\ 3\times3 & 3\times4 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 9 & 12 \end{pmatrix}$$

3. **Matrix Multiplication:** This is a "row-by-column" operation. To get the element in the first row and first column of the result, you multiply the elements of the first row of matrix A by the corresponding elements of the first column of matrix B and sum them up.
   - **Example:**

$$A * B = \begin{pmatrix} (1\times5 + 2\times7) & (1\times6 + 2\times8) \\ (3\times5 + 4\times7) & (3\times6 + 4\times8) \end{pmatrix}$$

$$= \begin{pmatrix} (5 + 14) & (6 + 16) \\ (15 + 28) & (18 + 32) \end{pmatrix}$$

$$= \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

4. **Transpose:** The transpose of a matrix is found by swapping its rows and columns. The first row becomes the first column, and so on. It is denoted by a superscript T.
   - **Example:**

$$A^T = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

12. Discuss applications of Linear Algebra in Machine Learning and Neural Networks.

Linear Algebra is the mathematical foundation on which modern Machine Learning (ML) and Neural Networks are built.

- **Data Representation:** In ML, all data, whether it's an image, text, or a table of user information, is represented as matrices or vectors. An image is a matrix of pixel values, and a sentence can be converted

into a vector of numbers.

- **Machine Learning Algorithms:** Many core ML algorithms are expressed in terms of linear algebra.
  - **Linear Regression:** This algorithm finds the best-fit line through data points, and the entire process is a problem of solving a system of linear equations.
  - **Support Vector Machines (SVMs):** These use dot products between vectors to find the best boundary to separate different classes of data.
- **Neural Networks:** A neural network is essentially a series of interconnected layers, and the transformation of data from one layer to the next is a matrix multiplication. The "weights" of the network are stored in a matrix, and the learning process involves continuously updating this matrix to improve predictions. Data flows through the network as a series of vector and matrix operations.

13. Explain measures of central tendency (mean, median, mode) with examples.

Measures of central tendency are single values that attempt to describe the "center" or a typical entry of a dataset. They give us a quick summary of the data.

1. **Mean (The Average):**
   - **What it is:** The sum of all values divided by the number of values. It is sensitive to outliers (extremely high or low values).
   - **When to use it:** Best for data that is symmetrically distributed without extreme outliers, like the average height of students in a class.
   - **Example:** For the dataset `[10, 20, 30, 40, 100]`, the mean is `(10+20+30+40+100)/5 = 40`. The outlier `100` significantly pulls the average up.
2. **Median (The Middle Value):**
   - **What it is:** The value that separates the higher half from the lower half of a dataset when it is sorted. It is not affected by outliers.
   - **When to use it:** Best for skewed data or data with outliers, like house prices in a city (where a few mansions shouldn't distort the "typical" price).
   - **Example:** For the dataset `[10, 20, 30, 40, 100]`, the median is `30`. It gives a better sense of the center than the mean in this case.
3. **Mode (The Most Frequent Value):**
   - **What it is:** The value that appears most often in the dataset. A dataset can have one mode, more than one mode, or no mode.
   - **When to use it:** Best for categorical data (non-numerical data like colors or brands).
   - **Example:** In a survey of favorite T-shirt colors `[Red, Blue, Blue, Green, Red, Blue]`, the mode is `Blue`.

14. What are measures of variability? Explain with examples.

Measures of variability (or dispersion) describe the spread or scatter of a dataset. They tell us how much the data points differ from each other and from the center.

1. **Range:**
   - **What it is:** The simplest measure, calculated as the difference between the highest and lowest values.
   - **Example:** For the scores `[60, 65, 70, 90, 95]`, the range is `95 - 60 = 35`. It's sensitive to outliers.
2. **Variance:**
   - **What it is:** The average of the squared differences from the Mean. A larger variance means the data is more spread out.
   - **Example:** For the dataset `[1, 2, 3, 4, 5]`, the mean is 3. The variance would be calculated based on the squared differences: $(1-3)^2$, $(2-3)^2$, $(3-3)^2$, $(4-3)^2$, $(5-3)^2$.

3. **Standard Deviation:**
    - **What it is:** The square root of the variance. It's the most common measure of spread and is in the same units as the data, making it easy to interpret.
    - **Real-Life Example:** If the average height of students is 5'5" with a standard deviation of 2 inches, it means most students' heights are clustered between 5'3" and 5'7". A smaller standard deviation means the heights are very similar to each other.

15. Explain probability with suitable examples of coin toss or dice roll.

Probability is the measure of the likelihood that an event will occur. It is the foundation for making predictions from data.

Basic Formula:
Probability(Event) = Number of Ways the Event Can Happen / Total Number of Possible Outcomes

- **Example 1: Tossing a Single Coin**
    - **Possible Outcomes:** {Heads, Tails} (Total = 2)
    - **Probability of getting Heads:** There is only 1 way to get heads. So, P(Heads) = 1 / 2 = 0.5 or 50%.
    - **Probability of getting Tails:** There is only 1 way to get tails. So, P(Tails) = 1 / 2 = 0.5 or 50%.
- **Example 2: Rolling a Single Die**
    - **Possible Outcomes:** {1, 2, 3, 4, 5, 6} (Total = 6)
    - **Probability of rolling a 3:** There is only one '3' on a die. So, P(3) = 1 / 6.
    - **Probability of rolling an even number:** The even numbers are {2, 4, 6}. There are 3 ways this can happen. So, P(Even) = 3 / 6 = 1 / 2 or 50%.
- **Example 3: Tossing Two Coins**
    - **Possible Outcomes:** {HH, HT, TH, TT} (Total = 4)
    - **Probability of getting exactly two heads (HH):** There is only 1 way for this to happen. So, P(HH) = 1 / 4 = 25%.
    - **Probability of getting at least one head:** The outcomes are {HH, HT, TH}. There are 3 ways. So, P(At least one Head) = 3 / 4 = 75%.

Application/Case-based Questions

16. Solve: Calculate the mean, median, and mode of the dataset: [5, 7, 2, 3, 1, 4, 8].

1. **Sort the data:** `[1, 2, 3, 4, 5, 7, 8]`
2. **Mean (Average):**
    - Sum = 1 + 2 + 3 + 4 + 5 + 7 + 8 = 30
    - Number of values = 7
    - Mean = 30 / 7 ≈ 4.29
3. **Median (Middle Value):**
    - In the sorted list `[1, 2, 3, 4, 5, 7, 8]`, the middle value is `4`.
4. **Mode (Most Frequent):**
    - Each number appears only once, so there is **no mode**.

17. Solve: If two coins are tossed, what is the probability of getting at least one tail?

1. **List all possible outcomes:** {HH, HT, TH, TT}. (Total = 4 outcomes)
2. **Identify favorable outcomes (at least one tail):** {HT, TH, TT}. (There are 3 favorable outcomes)
3. **Calculate the probability:**
    - Probability = (Favorable Outcomes) / (Total Outcomes) = 3 / 4 = 0.75 or 75%.

18. Using Python, write code to find variance and standard deviation of student marks.

You can use the powerful **NumPy** library for this.

```
import numpy as np

# Sample student marks
marks = np.array([60, 70, 75, 80, 85, 90, 95])

# Calculate variance
variance_marks = np.var(marks)

# Calculate standard deviation
std_dev_marks = np.std(marks)

print(f"Student Marks: {marks}")
print(f"Variance of marks: {variance marks:.2f}")
print(f"Standard Deviation of marks: {std_dev_marks:.2f}")
```

19. Explain how linear algebra is used in image representation in computer vision.

In computer vision, an image is not seen as a picture but as a grid of numbers (a matrix).

- **Grayscale Image:** A simple black and white image can be represented as a single 2D matrix. Each element in the matrix is a number representing the pixel intensity, usually from 0 (black) to 255 (white).
- **Color Image:** A color image is typically represented as three separate 2D matrices, one for each color channel: **R**ed, **G**reen, and **B**lue. These three matrices are stacked together to form a 3D matrix (or a tensor).
- **Image Manipulation:** Operations like brightening an image, rotating it, or applying a filter are all mathematical operations performed on these matrices using the principles of linear algebra. For example, brightening an image involves adding a constant value to every element in its matrix.

20. A bank wants to use probability to predict loan default. Explain how.

A bank can use probability to build a risk model that predicts the likelihood of a customer failing to repay a loan (defaulting).

1. **Data Collection:** The bank collects historical data on thousands of past loan applicants. This data includes features like `Credit Score`, `Income`, `Loan Amount`, `Age`, and most importantly, whether they `Defaulted` or `Paid` the loan.
2. **Calculating Probabilities:** Using this data, the bank calculates conditional probabilities. For example:
    - What is the probability of default *given* a person has a low credit score? `P(Default | Low Credit Score)`
    - What is the probability of default *given* a person has a low income? `P(Default | Low Income)`
3. **Building a Predictive Model:** The bank uses a statistical model (like Logistic Regression) that combines these probabilities. When a new customer applies for a loan, the model takes their information (credit score, income, etc.) and calculates a **probability score** of them defaulting.
4. **Decision Making:** If the calculated probability of default is above a certain threshold (e.g., 20%), the bank might reject the loan application or offer it at a higher interest rate to cover the risk.

Unit 3: Data Pre-processing

Short Answer Questions

15. Define Data Preprocessing and explain why it is necessary.

Data Preprocessing is the process of cleaning and preparing raw data to make it suitable for analysis and building machine learning models.

Why it's necessary: Raw data from the real world is often "dirty"—it can be incomplete, inconsistent, and contain errors. If you use dirty data for analysis, your results will be wrong. It's like cooking: you must wash and prepare your ingredients before you can make a good meal.

Analogy: You can't build a strong house with broken, mismatched bricks. Preprocessing is like cleaning the bricks, making sure they are all the right shape and size before you start building.

16. What are the common methods of data collection? (List any four).

1. **Surveys:** Asking people questions directly through questionnaires or interviews.
    - **Example:** A company sending out a customer satisfaction survey.
2. **Web Scraping:** Using automated tools to extract large amounts of data from websites.
    - **Example:** Collecting product prices from different e-commerce websites to compare them.
3. **Sensors:** Collecting data automatically from devices like GPS, temperature sensors, or smartwatches.
    - **Example:** A fitness app collecting step count data from your phone's sensor.
4. **Observational Studies:** Watching and recording behavior or events without interfering.
    - **Example:** A traffic engineer counting the number of cars passing an intersection at different times of the day.

17. What is the difference between measurement error and data entry error?

- **Measurement Error:** This error happens when the tool used to measure something is faulty. The data is collected incorrectly from the source.
    - **Example:** A weighing scale is not calibrated correctly and adds 1 kg to every measurement. Everyone who uses it will have their weight recorded incorrectly.
- **Data Entry Error:** This is a human error that happens when data is being manually typed into a system.
    - **Example:** A person is typing a customer's age as "32" but accidentally types "23" or "322". The original information was correct, but it was recorded incorrectly.

18. Define missing data and explain the types (MCAR, MAR, NMAR).

Missing data refers to the absence of values for certain variables in a dataset.

1. **MCAR (Missing Completely at Random):** The missingness has no reason and is purely random. It doesn't depend on any other data, seen or unseen.
    - **Example:** A participant in a survey accidentally skips a question.
2. **MAR (Missing at Random):** The missingness can be explained by other variables in the dataset, but not by the missing data itself.
    - **Example:** Men might be less likely to answer a survey question about depression. Here, the missingness of the "depression" answer depends on the "gender" variable.

3. **NMAR (Not Missing at Random):** The missingness is related to the value of the missing data itself. This is the most difficult type to handle.
    - **Example:** People with very high incomes are less likely to reveal their income in a survey. The missingness of the "income" data is directly related to how high the income is.

19. What is meant by consistent data? Give an example.

Consistent data means that the same piece of information is represented in the same way throughout the entire dataset. There are no contradictions.

- **Example of Inconsistent Data:** In a "State" column of a customer address table, you might find entries like "New York," "NY," and "N.Y." for the same state.
- **Example of Consistent Data:** After cleaning, all of these entries would be standardized to a single format, such as "NY," making the data consistent.

20. Define feature engineering.

Feature engineering is the creative process of using your domain knowledge to create new input variables (features) from your existing data. The goal is to make your machine learning models more effective.

Example: Instead of using a customer's `date of birth` as a feature, you could engineer a new, more useful feature called `age`. Or, from a `start_date` and `end_date`, you could create a `duration` feature.

Long Answer Questions (5 Marks Each)

21. Discuss different types of data errors with examples.

Real-world data is rarely perfect and can contain various types of errors that need to be fixed before analysis.

1. **Missing Data:** Values are absent in the dataset.
    - **Example:** A survey respondent leaves the "Age" field blank.
2. **Duplicate Data:** The same record appears more than once.
    - **Example:** Due to a system glitch, a customer's order is recorded twice in the sales database.
3. **Inconsistent Data:** The same entity is represented in different ways.
    - **Example:** In a company database, the same employee might be listed as "John Smith" in one table and "J. Smith" in another. Or a state is listed as both "California" and "CA".
4. **Incorrect or Invalid Data (Outliers):** Data points that are clearly wrong or fall far outside the expected range.
    - **Example:** A person's age is entered as "150" years, or a product price is listed as "-$10". These are likely data entry errors and are considered outliers.
5. **Formatting Errors:** Data is not in a standard format.
    - **Example:** A "Date" column contains dates in multiple formats like "10/01/2023," "Jan 10, 2023," and "2023-01-10". For analysis, they all need to be converted to a single format.

22. Explain the major tasks in data preprocessing.

Data preprocessing is a multi-step process to convert raw, dirty data into a clean, high-quality dataset. The major tasks are:

1. **Data Cleaning:** This is the first step and involves fixing the "dirt" in the data. This includes:
    - **Handling Missing Values:** Filling in or removing empty fields.

    - **Smoothing Noisy Data:** Removing errors and outliers.
    - **Resolving Inconsistencies:** Standardizing data formats and values.
2. **Data Integration:** This involves combining data from multiple different sources into a single, unified dataset.
    - **Example:** A company might combine customer data from its website, mobile app, and in-store purchases to get a complete 360-degree view of the customer.
3. **Data Reduction:** This task aims to reduce the size of the dataset without losing important information, making analysis faster and more efficient. This can be done by:
    - **Dimensionality Reduction:** Reducing the number of variables (columns).
    - **Numerosity Reduction:** Reducing the number of records (rows), for instance, by sampling.
4. **Data Transformation:** This involves converting the data into a more suitable format for modeling. Common techniques include:
    - **Normalization/Standardization:** Scaling all numeric data to a common range (e.g., 0 to 1) so that no single feature dominates the others.
    - **Aggregation:** Summarizing data, such as calculating monthly sales from daily sales data.

23. Describe different methods of handling missing data.

When faced with missing values in a dataset, a data scientist has several strategies to choose from, depending on the type and amount of missing data.

1. **Remove the Data (Deletion):**
    - **Listwise Deletion:** Delete the entire row if it contains any missing value. This is simple but can lead to significant data loss if many rows have missing values.
    - **Pairwise Deletion/Dropping a Variable:** If a particular column has too many missing values and is not very important, you can delete the entire column.
2. **Impute the Missing Values (Filling them in):** This is often a better approach than deletion.
    - **Mean/Median/Mode Imputation:** For numerical data, you can fill the missing values with the mean or median of the column. For categorical data, you can use the mode (most frequent value). This is simple but can distort the data's variance.
    - **Constant Value Imputation:** Fill the missing value with a constant like "0" or "Unknown." This can be useful, but the model might treat this constant as a new category.
    - **Advanced Imputation (e.g., Regression):** Use other variables to predict the missing value. For example, you could use a regression model to predict a missing `age` based on a person's `education level` and `job title`. This is more accurate but computationally more expensive.

24. Explain Data Cleaning as a process.

Data Cleaning is the process of detecting and correcting corrupt, inaccurate, or irrelevant records from a dataset. It is arguably the most important step in data preprocessing because it directly impacts the quality of any analysis or model. The process can be broken down into these steps:

1. **Identify the "Dirty" Data:** The first step is to systematically find the problems. This involves checking for:
    - **Missing Values:** Finding columns with empty cells.
    - **Outliers:** Identifying values that are far from the norm (e.g., an age of 200).
    - **Inconsistencies:** Looking for formatting issues (e.g., "NY" vs. "New York") or contradictions.
    - **Duplicates:** Finding identical rows.
2. **Handle Missing Data:** Choose a strategy like removing the rows or imputing the missing values with the mean, median, or a predicted value.

3. **Correct Invalid Data and Outliers:** Fix typos and decide what to do with outliers—you might remove them, cap them (e.g., set any age above 100 to 95), or treat them as missing values.
4. **Standardize and Ensure Consistency:** Convert all data in a column to a standard format. This includes making sure date formats are the same, units of measurement are consistent, and categorical values like state names are standardized.
5. **Remove Duplicates:** Delete any duplicate records from the dataset.
6. **Verify:** After cleaning, it's important to review the data to ensure that the cleaning process didn't introduce new errors and that the data is now ready for analysis.

25. Write short notes on:

a) Data Integration
Data Integration is the process of combining data from different sources to create a single, unified view. In today's world, data is often stored in many places—a company might have a sales database, a marketing database, and a customer support system. Data integration brings all this data together. Challenges in this process include **schema integration** (matching columns like `Cust-ID` and `Customer-No` that mean the same thing) and **entity resolution** (identifying that "Bill Clinton" and "William Clinton" are the same person).

b) Data Reduction
Data Reduction involves techniques to reduce the volume of a dataset while preserving its analytical value. The goal is to make storage and analysis more efficient. Two main strategies are:

- **Dimensionality Reduction:** Reducing the number of features (columns). For example, if `length` and `width` are highly correlated with `area`, you might just keep the `area` feature and drop the other two.
- **Numerosity Reduction:** Reducing the number of records (rows). This can be done by clustering data points and storing only the cluster representatives, or by simply taking a random sample of the data.

c) Data Transformation and Normalization
Data Transformation is the process of changing the format, structure, or values of data. A key part of this is **Normalization**, which is the technique of scaling numerical data from different columns to a common range, typically between 0 and 1 (Min-Max Normalization) or with a mean of 0 and standard deviation of 1 (Z-score Standardization). This is crucial because machine learning algorithms can be biased towards features with larger values. For example, if you have an `age` feature (e.g., 20-60) and an `income` feature (e.g., 50,000-200,000), the income feature would dominate the model simply because its numbers are bigger. Normalization puts all features on a level playing field.

Application/Case-based Questions

26. Suppose you are working with a dataset containing missing customer ages. Suggest different methods to handle missing values.

If I had a dataset with missing customer ages, I would consider the following methods, from simplest to most complex:

1. **Remove the Rows:** If only a very small percentage of rows (e.g., < 2%) have a missing age, the easiest solution is to simply delete those customer records. This is quick but not ideal if the missingness is not random or if you lose too much data.
2. **Impute with Mean/Median:** I could calculate the mean or median age of all the other customers and fill in the missing spots with that value. I would prefer the **median** because age data can be skewed (e.g., by a few very old customers), and the median is less sensitive to outliers.

3. **Impute with a Constant:** I could fill the missing ages with a value like "-1" or "Unknown." This makes it clear that the data was originally missing, and the model might even learn a pattern from this "Unknown" category.
4. **Predict the Missing Age (Regression Imputation):** This is the most sophisticated method. I could use other features in the dataset, like `years_of_education`, `job_title`, or `purchase_history`, to build a regression model that predicts the age. I would then use this model to fill in the missing age values. This is likely to be the most accurate approach.

27. A company collects data from multiple sources (web scraping, surveys, sensors). Discuss the challenges of data integration.

Integrating data from such diverse sources presents several challenges:

- **Schema Mismatches:** The data from each source will have a different structure. The survey data might have a column named `customer_name`, while the web data has `user_name`. A key challenge is mapping these different schemas so that the data can be merged correctly.
- **Data Format Inconsistencies:** The sensor data might record timestamps in UTC, while the survey data records them in a local time zone. Dates, numbers, and text will likely be in different formats that need to be standardized.
- **Entity Resolution Problem:** The same customer might be represented differently in each source. For example, a user might be "John D." in the web data, have a specific user ID in the sensor data, and be "John Doe" in the survey. The challenge is to correctly identify that these all refer to the same person.
- **Data Redundancy and Contradictions:** When you combine sources, you will likely get redundant data (the same information recorded multiple times) and conflicting data (e.g., a customer's address is different in the survey and web data). You need to establish rules to decide which source is the "single source of truth."

28. Consider a dataset containing "New York" and "NY" as state names. How will you preprocess this data for consistency?

To handle this inconsistency, I would perform a standardization process. The steps would be:

1. **Explore the Data:** First, I would get a list of all unique values in the "state" column to identify all the variations (e.g., "New York," "NY," "N.Y.", "new york").
2. **Define a Standard Mapping:** I would create a dictionary or a mapping rule that defines the standard format for each state. For this example, the rule would be to convert all variations to the two-letter code.
    - `{"New York": "NY", "N.Y.": "NY", "new york": "NY"}`
3. **Apply the Transformation:** I would then apply this mapping to the entire "state" column. This can be easily done using a function like `replace()` in Python's pandas library.
4. **Verify:** After the transformation, I would check the unique values again to ensure that only the standardized formats (e.g., "NY") remain.

29. Explain with Python code how to detect and handle outliers using the IQR method.

The Interquartile Range (IQR) method defines outliers as any data points that fall below the first quartile (Q1) by 1.5 times the IQR or above the third quartile (Q3) by 1.5 times the IQR.

Here's how you can do it in Python using the pandas library:

```
import pandas as pd
import numpy as np
```

```
# 1. Create a sample dataset with outliers
data = {'score': [25, 28, 29, 30, 32, 33, 34, 35, 38, 40, 75, 100]}
df = pd.DataFrame(data)

print("Original DataFrame:")
print(df)

# 2. Calculate Q1, Q3, and IQR
Q1 = df['score'].quantile(0.25)
Q3 = df['score'].quantile(0.75)
IQR = Q3 - Q1

# 3. Define the outlier boundaries
lower bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

print(f"\nQ1: {Q1}, Q3: {Q3}, IQR: {IQR}")
print(f"Lower Bound for Outliers: {lower bound}")
print(f"Upper Bound for Outliers: {upper_bound}")

# 4. Detect the outliers
outliers = df[(df['score'] < lower_bound) | (df['score'] > upper_bound)]

print("\nDetected Outliers:")
print(outliers)

# 5. Handle the outliers by removing them
df no outliers = df[~((df['score'] < lower_bound) | (df['score'] >
upper_bound))]

print("\nDataFrame after removing outliers:")
print(df_no_outliers)
```

Unit 4: Data Visualization

Short Answer Questions

21. Define a graph or chart in data science.

A graph or chart in data science is a visual representation of numerical data. It helps in understanding complex data, identifying patterns, trends, and outliers, and communicating insights more effectively than raw numbers.

Analogy: Think of a chart as a map for your data. A long list of addresses (data) is hard to understand, but a map (chart) instantly shows you where everything is located and how places relate to each other.

22. What is visual encoding? List its key aspects.

Visual encoding is the process of translating data into visual elements on a chart. It's how we map data values to visual properties like position, color, and size.

Key Aspects:

- **Position:** Where points are on the X and Y axes (most effective).
- **Length:** The length of bars in a bar chart.
- **Angle:** The size of slices in a pie chart.
- **Color:** To differentiate categories or show intensity.
- **Size:** The size of bubbles in a bubble chart to show magnitude.
- **Shape:** Different shapes for different categories in a scatter plot.

23. What is the use of Matplotlib in data visualization?

Matplotlib is a fundamental and widely-used Python library for creating static, animated, and interactive visualizations. It provides a huge amount of control over every aspect of a plot, making it the foundation for many other visualization libraries.

Analogy: Matplotlib is like a full set of artist's tools. You have every brush, color, and canvas you need to create any kind of painting (visualization) you can imagine, from a simple sketch to a detailed masterpiece.

24. Differentiate between line plot and scatter plot.

- **Line Plot:** Connects a series of data points with a continuous line. It is primarily used to show trends or changes over a continuous interval or time.
    - **Example:** Tracking the change in a company's stock price over a month.
- **Scatter Plot:** Shows individual data points as dots without connecting them. It is used to show the relationship or correlation between two numerical variables.
    - **Example:** Plotting a person's height versus their weight to see if there is a relationship.

25. What is a histogram and when do we use it?

A histogram is a type of bar chart that shows the frequency distribution of a set of continuous numerical data. It groups numbers into ranges (called "bins") and the height of the bar shows how many data points fall into that range.

When to use it: Use a histogram when you want to understand the underlying distribution of a single variable, like its shape, center, and spread.

Example: To see the distribution of student exam scores, a histogram can show how many students scored between 90-100, 80-89, 70-79, and so on.

26. Explain the purpose of annotations and legends in plots.

- **Annotations:** These are notes or text added directly onto a plot to point out specific data points or provide extra information. Their purpose is to draw attention to something important that the viewer might otherwise miss.
    - **Example:** Adding an arrow and text to a sales chart that says "New marketing campaign launched" to explain a sudden spike in sales.
- **Legends:** A legend acts as a key for the plot. It explains what the different colors, shapes, or line styles represent, especially when multiple datasets are plotted on the same graph.
    - **Example:** In a line chart showing sales for two different products, the legend would show that the blue line represents "Product A" and the red line represents "Product B".

27. Why is choosing the right graph important?

Choosing the right graph is crucial because the type of visualization directly affects how well the data's story is told and understood. An incorrect chart can confuse the audience or even lead to wrong conclusions.

Example: If you want to show the percentage breakdown of a company's budget, a pie chart is perfect. But using a line chart for the same data would be meaningless and confusing, as it would imply a trend over time where none exists.

28. What is the role of colors and markers in visualizations?

- **Colors:** Color is a powerful tool used to differentiate between categories (e.g., blue bars for male, pink for female), show intensity (e.g., light red for low temperature, dark red for high temperature), or highlight specific data points.
- **Markers:** Markers are symbols (like circles, squares, or triangles) used to represent individual data points, typically in scatter plots or line plots. They help distinguish between different groups of data on the same plot, for example, using circles for one dataset and triangles for another.

Long Answer Questions (5 Marks Each)

29. Explain different types of plots (line, bar, histogram, scatter, pie) with examples.

1. **Line Plot:** Shows trends over time. Data points are connected by a line, making it easy to see how a value changes.
    - **Example:** Tracking website traffic (number of visitors) each day for a month.
2. **Bar Chart:** Used to compare quantities across different categories. Each category has a bar, and its length represents the quantity.
    - **Example:** Comparing the total sales figures for different products (e.g., Laptops, Phones, Tablets) in a quarter.
3. **Histogram:** Shows the distribution of a single continuous variable. It groups data into bins and shows the frequency of data points in each bin.
    - **Example:** Visualizing the distribution of heights of people in a large group to see if it follows a normal (bell-shaped) curve.

4. **Scatter Plot:** Used to visualize the relationship between two continuous variables. Each point represents one observation.
    - **Example:** Plotting hours spent studying against exam scores to see if more study time correlates with higher scores.
5. **Pie Chart:** Shows parts of a whole, representing proportions or percentages. The entire pie represents 100%, and each slice represents a category's share.
    - **Example:** Showing the market share of different smartphone brands (e.g., Apple, Samsung, Google) in a country.

30. Discuss the importance of visual encoding in effective data visualization.

Visual encoding is the foundation of effective data visualization. It's the science and art of mapping data values to visual properties (like position, size, and color) in a way that our brains can easily decode. Its importance lies in several key areas:

1. **Clarity and Accuracy:** Good visual encoding ensures that the message is clear and not misleading. The most important data should be encoded with the most effective visual cues. For instance, our brains are very good at judging position and length, which is why bar charts are often more accurate for comparisons than pie charts (where we are worse at judging angles).
2. **Efficiency:** It allows viewers to grasp information quickly. A well-encoded chart can convey a complex pattern in seconds, whereas the same information in a table might take minutes to understand. For example, using color to distinguish categories allows for instant recognition.
3. **Highlighting Insights:** It can be used to draw attention to the most important parts of the data. For instance, using a bright, contrasting color for an outlier in a scatter plot immediately draws the viewer's eye to it.
4. **Revealing Relationships:** By encoding different variables to different visual properties, we can reveal complex relationships. A scatter plot can encode one variable on the X-axis, another on the Y-axis, a third to the size of the points, and a fourth to the color, all in one chart.

Ultimately, a visualization fails or succeeds based on its encoding. Poor choices lead to confusion and misinterpretation, while thoughtful choices lead to insight and understanding.

31. Explain how Matplotlib properties (figure, axes, plot, legend, color) improve graphs.

Matplotlib properties allow a data scientist to customize every detail of a graph, transforming it from a basic, uninformative chart into a clear, professional, and insightful visualization.

- **Figure:** This is the outermost container for the entire visualization. Properties like `figsize` allow you to control the overall dimensions (width and height) of the plot, ensuring it fits well in a report or presentation. `facecolor` lets you set a background color for the entire figure.
- **Axes:** The axes represent the actual plotting area (the individual graph). You can control the `title` of the plot, and set the `xlabel` and `ylabel` to tell the viewer what the axes represent. This is fundamental for making a graph understandable.
- **Plot:** This refers to the actual data being plotted (the lines, bars, etc.). You can control the `color` of the plot to distinguish it, the `linestyle` (e.g., solid, dashed) to show different types of data, and the `marker` (e.g., circle, square) to highlight specific data points.
- **Legend:** The legend is crucial when you have multiple datasets on one graph. The `legend()` function uses the `label` property from each plot to create a key that explains what each line or bar represents, preventing confusion.
- **Color:** Color is used throughout Matplotlib to improve clarity. You can set the color of lines, bars, markers, text, and backgrounds. Using color effectively helps to group related items, highlight important data, and make the graph visually appealing.

32. Compare bar charts, histograms, and pie charts.

Aspect	Bar Chart	Histogram	Pie Chart
What it Shows	Comparison of quantities between discrete, separate categories.	Frequency distribution of a single continuous variable.	Proportions or percentages of a whole.
Data Type	Categorical data (e.g., products, countries).	Continuous numerical data (e.g., age, height, temperature).	Categorical data where parts add up to a whole.
X-Axis	Represents the distinct categories. The bars have gaps between them.	Represents continuous numerical ranges (bins). The bars touch each other.	Not applicable (categories are represented by slices).
Best For	Ranking or comparing values, like sales per city.	Understanding the shape and spread of data, like the distribution of exam scores.	Showing market share or budget allocation.
Limitation	Can become cluttered with too many categories.	The choice of bin size can drastically change the look of the histogram.	Hard to compare slices accurately, especially if they are similar in size. Not good for more than a few categories.

33. Explain the role of data visualization in decision-making.

Data visualization plays a critical role in modern decision-making by bridging the gap between raw data and human understanding. It translates complex numerical data into an intuitive visual format, which helps stakeholders at all levels make faster, more informed decisions.

1. **Identifying Trends and Patterns Quickly:** A line chart can instantly show whether sales are trending up or down, an insight that might be missed by looking at a large table of numbers. This allows decision-makers to react quickly to opportunities or threats.
2. **Simplifying Complexity:** Visualizations can summarize millions of data points into a single, understandable graphic. A dashboard with a few key charts can give a CEO a high-level overview of the company's health in minutes, which is far more effective than reading pages of reports.
3. **Communicating Insights Effectively:** A well-designed chart is a powerful communication tool. It can tell a compelling story and persuade an audience. When presenting to a board of directors, showing a bar chart that clearly illustrates how one product is outperforming all others is far more impactful than just stating the numbers.
4. **Spotting Outliers and Anomalies:** Visualizations make it easy to spot data points that don't fit the pattern. For example, a scatter plot might reveal a cluster of fraudulent transactions that would be nearly impossible to find by manually scanning through data. This helps in risk management and process improvement.

In essence, data visualization empowers decision-makers by making data accessible, understandable, and actionable.

Application/Case-based Questions

34. Write a Python code to create a line plot using student scores.

Here is a simple Python code using Matplotlib to create a line plot showing a student's scores across five different tests.

```
import matplotlib.pyplot as plt

# Data: Test numbers and the scores achieved
tests = [1, 2, 3, 4, 5]
scores = [75, 80, 78, 85, 90]

# Create the plot
plt.plot(tests, scores, marker='o', linestyle='-', color='b')

# Add titles and labels for clarity
plt.title("Student's Test Scores Over Time")
plt.xlabel("Test Number")
plt.ylabel("Score")

# Add a grid for easier reading
plt.grid(True)

# Display the plot
plt.show()
```

35. Using Titanic dataset, draw a bar chart showing number of passengers by class.

This code uses pandas to load the Titanic dataset and Matplotlib to create a bar chart showing the count of passengers in each class (1st, 2nd, and 3rd).

```
import pandas as pd
import matplotlib.pyplot as plt

# Assuming 'titanic.csv' is in the same directory
# You can download it from many sources online, like Kaggle
df = pd.read_csv('titanic.csv')

# Count the number of passengers in each class
passenger_class_counts = df['Pclass'].value_counts().sort_index()

# Create the bar chart
plt.figure(figsize=(8, 6)) # Set the figure size
passenger class_counts.plot(kind='bar', color=['skyblue', 'lightgreen',
'salmon'])

# Add titles and labels
plt.title('Number of Passengers by Class on the Titanic')
plt.xlabel('Passenger Class')
plt.ylabel('Number of Passengers')
plt.xticks(rotation=0) # Keep the x-axis labels horizontal

# Display the plot
plt.show()
```

36. Write a Python program to draw a histogram showing age distribution of passengers.

This code will create a histogram to visualize the age distribution of the Titanic passengers. It's important to drop missing age values (`.dropna()`) before plotting.

```
 import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('titanic.csv')

# Create the histogram
plt.figure(figsize=(10, 6))
# We drop missing 'Age' values to avoid errors
plt.hist(df['Age'].dropna(), bins=20, color='teal', edgecolor='black')

# Add titles and labels
plt.title('Age Distribution of Titanic Passengers')
plt.xlabel('Age')
plt.ylabel('Number of Passengers (Frequency)')

# Display the plot
plt.show()
```

37. Create a scatter plot for Age vs Fare using the Titanic dataset.

A scatter plot is excellent for seeing if there's a relationship between two numerical variables like age and the fare passengers paid.

```
 import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('titanic.csv')

# Create the scatter plot
plt.figure(figsize=(10, 6))
plt.scatter(df['Age'], df['Fare'], alpha=0.5, color='purple')

# Add titles and labels
plt.title('Scatter Plot of Age vs. Fare on the Titanic')
plt.xlabel('Age')
plt.ylabel('Fare Paid')

# Display the plot
plt.show()
```

38. Suppose you want to show the percentage of students passing vs failing. Which visualization would you choose and why?

For showing the percentage of students passing versus failing, the best choice would be a **Pie Chart**.

Why a Pie Chart?

1. **Shows Parts of a Whole:** A pie chart is specifically designed to show how different parts make up a whole. In this case, the "whole" is the total number of students, and the "parts" are the two distinct categories: "Passing" and "Failing".
2. **Easy to Understand Percentages:** The visual size of the slices in a pie chart makes it instantly clear what proportion of the whole each category represents. It’s very intuitive for an audience to see the percentage breakdown.
3. **Simple and Clean:** Since there are only two categories, a pie chart will be very simple, clean, and not cluttered. It directly answers the question "What percentage of students passed?" without any extra distractions.

A bar chart could also work, but it is better for comparing the absolute number of students, whereas a pie chart is superior for highlighting the percentage relationship between the two groups.