1. **Define Data Science and Artificial Intelligence. How are they different?**

* Data Science is an interdisciplinary subject that involves collecting, cleaning, and analyzing data to get meaningful insights.
* Artificial Intelligence is the branch of computer science that aims to develop a system or machine that could do the tasks that human does.
* Data Science focuses on extracting knowledge and insights from data to generate reports, visualizations, and predictions. On the other hand, Artificial Intelligence (AI) focuses on building systems that can mimic human behavior, such as making decisions. AI uses predictive models and automated decision-making based on trained datasets. The insights and patterns discovered by data scientists are often used by AI developers to create and train intelligent models.

1. **What is the intersection between AI and Data Science?**

* The intersection between the AI and Data Science is mentioned below:

1. Machine Learning: Data Science use it to build a predictive models, whereas AI use it to give machines ability to learn from data.
2. Model Training and Evaluation: Data scientists collect, clean and analyze data to build a model that AI engineers deploy in their projects.
3. Big data and analytics: Both data scientists and AI engineers requires tools like python, Tensor Flow, dashboards and so on for handling the large datasets.
4. **List three real-world use cases where both are applied together.**
5. **Health Diagnosis and Analysis:**

Data Science analyzes the patient data from the past medical history.

Artificial Intelligence uses the trained models to predict diagnoses and treatments.

1. **Fraud Detection in Banking:**

Data Scientists analyzes the transaction data, past fraud cases, and user behaviors.

Artificial Intelligence learns the pattern of the transaction and flag unusual activity.

1. **E-Commerce Recommendation:**

Data Scientists gathers and analyze the purchase history of the customer and their behaviors.

Artificial Intelligence helps in filtering and recommending the right products that customer is looking or paying attention.

1. **List and explain three types of file formats commonly used to store datasets.**
2. **CSV (Comma-Separated Values):**

CSV is a simple text-based format where data is stored in rows and columns, separated by commas. It is easy to create, read, and use in tools like Excel, Python, R, etc. It is mostly used for basic tabular data.

1. **JSON (JavaScript Object Notation):**

JSON is a lightweight data format used to store and share structured data as key-value pairs. It is widely used in web apps, APIs, and is readable by both humans and machines. It is used for storing complex or nested data from web servers.

1. **XLS/XLSX (Excel Files):**

These files formats are used by Microsoft Excel to store data in a spreadsheet. They can contain multiple sheets, formulas, and charts, making them powerful for data analysis. This is mostly used for business and medium calculations.

1. **Define:**

**Population:** The information of a complete group.

**Sample:** It is a subset of the population which means suppose you want 400 student results out of thousand student.

**Variable:** Variable is the container to store the data or values.

1. **Explain four levels of measurement (nominal, ordinal, interval, and ratio) with examples.**
2. **Gaussian Distribution:**

It is a continuous probability distribution. It has a symmetric, bell-shaped curve centered around the mean. It is defined by two parameters: mean (μ) and standard deviation (σ). Most values cluster around the mean, and the probability decreases as they move further away.

Example: Student Test Scores

•In a large class, the distribution of student marks often follows a normal distribution.

•Most students score near the average, while few score extremely high or low.

1. **Binomial Distribution:**

It is a discrete distribution. It represents number of successes in a fixed number of independent trials, with only two possible outcomes: success or failure. It is defined by the number of trials (n) and probability of success (p).

Example: Quality Control in Manufacturing

•Suppose a factory produces light bulbs, and each bulb has a 2% chance of being defective.

•If 100 bulbs are tested, the binomial distribution can model the probability of finding exactly 5 defective bulbs.

1. **Define and differentiate between:**

**Mean:** It is the average of set of numbers. It shows the central value by averaging all data points.

**Median:** The middle value when the data is arranged in order is called the median. It is the middle number in an ordered list.

**Mode:** The number that appears frequently is called the mode.

1. **Describe any two probability distributions (e.g., Gaussian, Binomial, and Poisson) with real-world use cases.**
   * + 1. **Gaussian Distribution:**

It is a continuous probability distribution. It has a symmetric, bell-shaped curve centered around the mean. It is defined by two parameters: mean **(μ)** and standard deviation **(σ)**. Most values cluster around the mean, and the probability decreases as they move further away.

**Example:** **Student Test Scores**

* In a large class, the distribution of student marks often follows a normal distribution.
* Most students score near the average, while few score extremely high or low.
  + - 1. **Binomial Distribution:**

It is a discrete distribution. It represents number of successes in a fixed number of independent trials, with only two possible outcomes: success or failure. It is defined by the number of trials (n) and probability of success (p).

**Example:** **Quality Control in Manufacturing**

* Suppose a factory produces light bulbs, and each bulb has a 2% chance of being defective.
* If 100 bulbs are tested, the binomial distribution can model the probability of finding exactly 5 defective bulbs.

1. **Define and distinguish:**

**Type I Error (False Positive):** A Type I Error occurs when the null hypothesis (H0) is true, but we incorrectly reject it.

In simpler words, we detect an effect or difference that does not actually exist.

**Type II Error (False Negative):** A Type II Error occurs when the null hypothesis (H0) is false, but we fail to reject it.

We miss detecting a real effect or problem.

Type I Errors Basis Type II Error

Detecting something that isn’t there. Meaning Not detecting something that actually exists.

True, but rejected Null Hypothesis False, but accepted

Test says you’re sick, but you’re healthy Example Test says you’re healthy, but you are actually sick.

1. **What is a confusion matrix? Explain with an example.**

* A confusion matrix is the table used to calculate the performance of classification model. It has the following structure:

|  | **Predicted: Positive** | **Predicted: Negative** |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |
| **Actual: Positive** | True Positive (TP) | False Negative (FN) |

|  |  |  |
| --- | --- | --- |
| **Actual: Negative** | False Positive (FP) | True Negative (TN) |

**Suppose a medical test predicts whether the individual has a disease or not:**

 True **Positive (TP)**: Test correctly predicts disease → 1

 False **Negative (FN)**: Test misses disease → 1

 False **Positive (FP)**: Test wrongly says disease → 1

 True **Negative (TN)**: Test correctly says no disease → 1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | Predicted Positive | Predicted Negative | | Actual: Positive | 1 TP | 1(FN) | | Actual: Negative | 1 (FP) | 1 (TN) | |

1. **Explain the concepts of:**
2. **Null Hypothesis:** The **null hypothesis** is a statement that there is **no effect, no difference**, or **no relationship** between variables. It represents the **default or status quo assumption** that any observed outcome is due to **random chance.**

* **Purpose:** It is what you **try to disprove or reject** in hypothesis testing.
* **Example:**
  + H₀: The new medicine has **no effect** on blood pressure.
  + H₀: There is **no difference** in exam scores between two teaching methods.

1. **Alternative Hypothesis:** The **alternative hypothesis** is the statement that **there is an effect, a difference**, or **a relationship** between variables. It is what you **hope to support** with evidence from your data.

* **Purpose:** It is accepted if the null hypothesis is rejected.
* **Example:**
  + H₁: The new medicine **does affect** blood pressure.
  + H₁: There **is a difference** in exam scores between the two teaching methods.

1. **What is the purpose of a chi-square test? When is it used?**

* The **chi-square (χ²) test** is a **statistical test** used to determine whether there is a **significant association** between two categorical variables.

 **Chi-Square Test of Independence**

* **Used when**: You want to see if two categorical variables are related.
* **Example**: Is there a relationship between gender (male/female) and voting preference (party A/party B)?

 **Chi-Square Goodness of Fit Test**

* **Used when**: You want to see if a sample matches an expected distribution.
* **Example**: Do the colors in a bag of candies follow the company's stated distribution?

1. **Describe a real-world example where hypothesis testing would be necessary in business or healthcare.**

Hypothesis Testing is equally necessary in both business and health care sectors.

* 1. **Health Care Sectors:** In health sectors, a company develops a new drug that claims to lower blood pressure more effectively than the current drug. In this scenario, A Hypothesis Testing is done to determine whether the new drug is better than the current one.

In this case, the Null Hypothesis be like there is no difference in blood pressure reduction between the new and current drug. Like this all the data collection and testing are performed to determine the effectiveness of drug.

It helps in ensuring evidence-based medicine and prevents release of harmful drugs.

* 1. **Business:**

An e-commerce company is launching a new style of promotional email and wants to know if it increases customer click-through rates compared to the old version.

The null hypothesis in this scenario is that the new email campaign has the same click-through rate as the old one. And the alternative hypothesis be there is higher click-through rate than the old one.

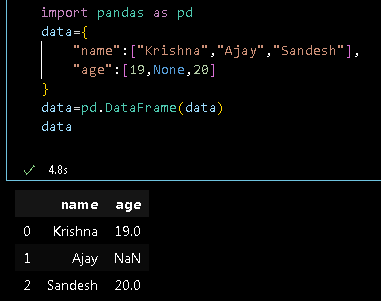
It helps with improving marketing decisions, increases conversion and revenues and helps in A/B testing strategies.

1. **What is the difference between Numpy arrays and Pandas DataFrames?**

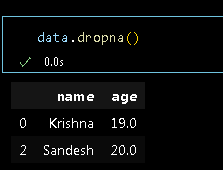
| **Feature** | **NumPy Array** | **Pandas DataFrame** |
| --- | --- | --- |
| **Structure** | Homogeneous (all elements must be of the same type) | Heterogeneous (different columns can have different types) |
| **Dimensions** | 1D, 2D, or higher-dimensional arrays | 2D tabular data (rows and columns) |
| **Indexing** | Uses integer-based indexing only | Supports labeled indexing with row and column labels |
| **Functionality** | Fast mathematical and matrix operations | Built-in support for data cleaning, aggregation, filtering |
| **Missing Data Handling** | Not handled well | Built-in support using NaN and functions like fillna() |
| **Best For** | Numerical computation and linear algebra | Data analysis and manipulation |
| **Library Origin** | Core package for scientific computing in Python | Built on top of NumPy, specialized for data manipulation |

1. **What is a missing value and how can you handle it in Pandas?**

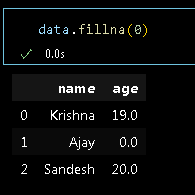
* The null, blank or NaN in the dataset which donot provide complete information of certain row or subject is called missing value. In pandas, it can be handled by below way:

****

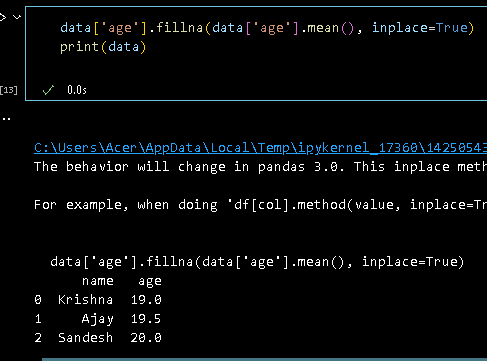
1. **Drop rows**

****

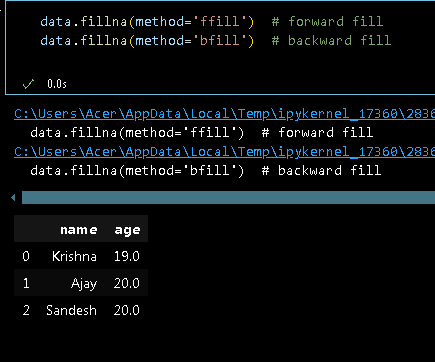
1. **Fill missing value**

****

1. **With mean/median/mode**

****

1. **Forward or backward fill**

****

1. **Use this sample dataset and apply the following (Practical):**

data = {

'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Eva'],

'Age': [25, 30, np.nan, 22, 28],

'Salary': [50000, 54000, 58000, 52000, np.nan],

'Department': ['HR', 'Finance', 'HR', 'IT', 'Finance']

}

**A computer screen shot of a black background with colorful text

AI-generated content may be incorrect.**

* 1. **Fill missing values with:**
     1. **Mean for Age**
     2. **Median for Salary**

**A computer screen shot of white text

AI-generated content may be incorrect.**

* 1. **Filter employees with Salary > 52000**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

* 1. **Group by Department and show average Salary**

**A screen shot of a computer

AI-generated content may be incorrect.**

* 1. **Add a new column Tax = Salary × 0.1 (10% tax)**

****

**A screenshot of a computer

AI-generated content may be incorrect.**

* 1. **Find the employee with the maximum salary**

**A screen shot of a computer

AI-generated content may be incorrect.**

1. Fill missing values with:

Mean for Age

Median for Salary

Filter employees with Salary > 52000

Group by Department and show average Salary

Add a new column Tax = Salary × 0.1 (10% tax)

1. **Find the employee with the maximum salary**
2. **What is SQL? Why is it important in data science?**

Structured Query Language is a programming language used to store, manage and retrieve data from relational databases.

Data Scientists often work with large datasets that are stored in databases. SQL is crucial because it allows them to access, extract, and manipulate this data before performing analysis or building models.

The importance of SQL in Data science includes:

* + 1. SQL is used for data extraction from databases using queries.
    2. SQL helps to filter, sort, and remove duplicates data.
    3. SQL helps in operations such as SUM, AVG, COUNT, GROUP, etc.
    4. SQL efficiently handles millions of rows in relational databases.
    5. SQL integrates easily with Python, R, Excel and many data tools.

1. **Define and differentiate between:**

**INNER JOIN vs. LEFT JOIN**

|  |  |  |
| --- | --- | --- |
| **Inner Join** | **Basis** | **Left Join** |
| Inner join only matches rows in both the tables | **Returns** | All rows from the left table with matched data are returned in left join. |
| Unmatched rows are excluded in the output. | **Unmatched rows** | Unmatched rows are included in the output. |
| It is used for retrieving the common data from both tables,  Like Data Intersection | **Use Case** | It is used for retrieving the data from left tables with the common datas available in right table. |

1. **Create a table (Practical):**

**CREATE TABLE employees (**

**id INTEGER PRIMARY KEY,**

**name TEXT,**

**department TEXT,**

**salary INTEGER,**

**experience INTEGER**

**);**

**Insert at least 5 rows, then write SQL queries to:**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**Insert at least 5 rows, then write SQL queries to:**

**A screenshot of a computer code

AI-generated content may be incorrect.**

* 1. **Select all employees in the Finance department**

**A screenshot of a computer

AI-generated content may be incorrect.**

* 1. **Get employees with experience > 3 years**

**A screenshot of a computer

AI-generated content may be incorrect.**

* 1. **Calculate average salary by department**

**A screenshot of a computer

AI-generated content may be incorrect.**

* 1. **Find the employee(s) with a salary greater than the overall average**

**A screenshot of a computer

AI-generated content may be incorrect.**

* 1. **Join with a new table department (id, name, location) to fetch department details**

**Apply following operations in pandas DataFrame (Practical):**

**# Students Table**

**students = pd.DataFrame({**

**'student\_id': [1, 2, 3, 4, 5],**

**'name': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'],**

**'class': ['10A', '10B', '10A', '10C', '10B']**

**})**

**# Scores Table**

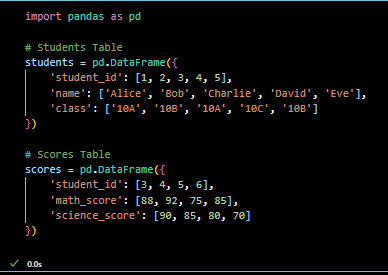
**scores = pd.DataFrame({**

**'student\_id': [3, 4, 5, 6],**

**'math\_score': [88, 92, 75, 85],**

**'science\_score': [90, 85, 80, 70]**

**})**

****

**Apply inner join, left join, right join, outer join**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Using the left join fill missing values in math\_score and science\_score with 0 and create a new column total\_score(math\_score+science\_score)**

**A computer screen shot of text

AI-generated content may be incorrect.**

