Junior Data Scientist Coding question

Scenario

You are hired as a **Junior Data Scientist** at a retail company.
Your manager has provided a dataset named **`sales_data.csv`** that contains sales transactions for various stores.
You are required to **analyze, clean, visualize, and build a simple predictive model** to generate insights from the data.

Dataset Structure

Column Name	Description
`Order_ID`	Unique ID for each transaction
`Date`	Date of order (YYYY-MM-DD)
`Store_Location`	Location of the store
`Product_Category`	Category of the product sold
`Units_Sold`	Number of units sold
`Unit_Price`	Price per unit
`Total_Sales`	Total sales amount (Units_Sold × Unit_Price)
`Discount`	Discount applied (%)
`Profit`	Profit from each sale

Note: Some rows may contain missing or inconsistent values.

Your Task

Write a **single Python program** that performs the following tasks step by step.

---

1 Import & Explore the Data

- Load the dataset using **Pandas**.
- Display the **first 5 rows** and **data types**.
- Print the **number of missing values per column**.
- Print the **number of unique stores** and **product categories**.

---

2 Data Cleaning

- Handle missing values:
    - Replace missing numerical values (`Units_Sold`, `Unit_Price`, `Profit`) with the **column mean**.
    - Replace missing categorical values (`Store_Location`, `Product_Category`) with the **mode**.
- Convert the `Date` column to **datetime format**.
- Create a new column called `Month` extracted from `Date`.

---

3 Data Transformation

- Calculate a new column `Revenue` = `Units_Sold × Unit_Price`.
- Create a column `Discounted_Sales` = `Revenue - (Revenue × Discount / 100)`.
- Normalize the `Profit` column (min-max normalization).

---

4 Exploratory Data Analysis (EDA)

Perform some **basic analytics and visualizations** using **Matplotlib / Seaborn**:

- Find and print:
    - Average sales per month
    - Total profit per store
    - Top 3 product categories by revenue
- Create and display:
    - A **bar chart** of total sales per store
    - A **line chart** of monthly average sales
    - A **boxplot** of profit by product category

---

5 Statistical Summary

- Calculate and print:
    - Mean, median, mode, and standard deviation for `Profit`.
    - Correlation matrix between numerical columns (`Units_Sold`, `Unit_Price`, `Discount`, `Profit`, `Revenue`).
    - Visualize the correlation matrix using **Seaborn heatmap**.

---

6 Simple Predictive Model

- Use **Linear Regression** (from `scikit-learn`) to predict **Profit** based on the following features:
    1. `Units_Sold`, `Unit_Price`, and `Discount`
- Steps:
    1. Split data into **train/test** (80/20).

2. Train a **LinearRegression** model.
3. Predict on the test set.
4. Print model performance metrics:
    - **R^2 Score**
    - **Mean Absolute Error (MAE)**

- Visualize **actual vs predicted Profit** using a scatter plot.