Module-4 Practical Assessment
Data Analysis with Python — Scenario Based Questions

Question-1 — Student Performance Analytics

Your college has recorded student marks and attendance for internal assessment. As a Data Analyst, you are asked to analyze this performance data so teachers can identify high-performing students.

Tasks

Create a Pandas DataFrame using the given dataset.
Handle missing values:

- Replace missing **Marks** with class average
- Replace missing **Attendance** with 0

Display list of students scoring **more than 80 marks**.
Sort students by attendance in **descending** order.
Group students by **Subject** and calculate average marks per subject to support subject-wise performance tracking.

Question-2 — Retail Store Price Management

You are working with a retail business to analyze product pricing. You receive price details for 6 products.

Tasks

Create a Pandas Series for product prices with product codes as Index.
Retrieve the price of product **P4** — customer has asked for it.
Identify premium products with price **greater than ₹300** for promotional strategy.
Increase prices by **10%** due to rising market costs.
Find the **highest** and **lowest** priced products after the update.

Question-3 — Company Sales Analysis

You are employed as a data analyst for a sales company. Management wants insights based on recent sales and category performance.

Tasks

Create a DataFrame using given sales dataset.

Filter orders where **Sales > ₹10,000** for premium billing review.

Sort data by Sales in Show descending order to analyze top-performing records.

Group data by **Category** to find:

- Total Sales per Category
- Total Quantity sold per Category

Region-wise Total Sales to find the best-performing region.

Question-4 — Monthly Sales Trend Report

A file sales_data.csv contains one year's sales of a retail chain. You must generate a performance report for management.

Tasks

Load and display first 10 rows.

Check dataset completeness using info() & describe().

Plot **monthly/quarterly sales trend** (line chart) to detect seasonality.

Compare **Sales by Product Category** using a bar graph.

Scatter plot for relationship between **Sales Amount vs Units Sold**.

Write **one insight** discovered from the above analyses.

Question-5 — Customer Spending Analysis

You are hired to study customers' buying behavior from customers.csv.

Tasks

Load and explore data structure (shape, head, data types).

Group by **Age Group/Gender** to compute average spending.

Bar chart — **Purchase count by Gender** for marketing focus.

Pie chart — **Customer distribution by Region**.

Provide **two business insights** (e.g., "North region customers spend more").

Question-6 — Sales & Customer Relationship Insights

A company combined its sales & customer feedback into sales_customer_combined.csv. You must find relational patterns.

Tasks

Load and show column statistics.

Generate **Correlation Heatmap** for:

- Sales Amount
- Discount
- Customer Rating
- Units Purchased

Scatter plot — **Discount vs Sales Amount**.

Line chart — **Customer purchase frequency trend** over the year.

Identify **two strong correlations** and give business meaning.

Statistics and Probability — Scenario Based

Question-7 — Academic Performance Study

You are analyzing whether students who score well in Math also perform well in Science.

Tasks

✓ Mean, Median, Mode of Math Marks

✓ Range, Variance, Standard Deviation

✓ Interpret performance and variation

✓ Covariance & Pearson Correlation (Math vs Science)

✓ Conclusion: Are Math & Science **positively correlated**?

Question-8 — Titanic Data Sampling & Insights

You work at a historical research institute analyzing Titanic passenger data.

Part A — Sampling

Load Titanic dataset

Perform **Stratified Sampling** by Passenger Class

Select **10% records** from each class

Compare class-wise counts before & after

Part B — Correlation

Find Covariance & Correlation between **Fare & Age**

Scatter plot visualization

Mention:

- Direction (positive/negative)
- Strength (weak/moderate/strong)

Question-9 — Fraud Detection Data Cleaning

You are working with a cybersecurity firm analyzing transaction fraud in fraud_transactions.csv.

Tasks

Load dataset and show summary stats

Fix missing values (mean/median/mode)

Remove duplicate records

Detect & remove **outliers** in transaction_amount using **IQR**

Ensure timestamps are correct datetime format

Export cleaned dataset as fraud_cleaned.csv

Question-10 — Fraud Detection ML Model

Using the cleaned dataset from Q-9, build a **Fraud Classification Model**.

Tasks

✓ Feature-Target split

✓ Encode categorical columns

✓ Scale numerical features

✓ Train-Test split (80:20)

✓ Train Logistic Regression Model

✓ Evaluate: Accuracy, Precision, Recall, F1

✓ Show Confusion Matrix

→ Provide baseline fraud detection performance

Power BI & Dashboarding — Scenario Based

Question-11 — E-Commerce Sales BI Dashboard

Your company wants an interactive dashboard using sales_data.xlsx.

Clean data → fix data types, missing category/sales

Visuals:

✓ Monthly Sales Trend (Line)
✓ Region-wise Sales (Geo Map)
✓ Sales by Category (Bar/Column)
✓ Top 10 customers table
Add slicers: Year & Region
Create KPI cards:
• Total Revenue • Total Orders • Average Order Value
Goal: Help management track **revenue growth & regional performance**

Question-12 — Workforce Productivity Dashboard

HR wants insights from attendance_productivity.csv.

Clean data → remove duplicates, correct date format
Visuals:
✓ Attendance Trend (Line)
✓ Productivity by Region (Map)
✓ Department-wise Bar Chart
✓ Work Hours Matrix
Create DAX Measures
✓ Total Working Hours
✓ Avg Productivity Score
✓ Attendance Compliance %
Dashboard should support workforce planning decisions

Retail Analytics & ML

Question-13 — Market Basket + Customer Retention

You analyze retail_transactions.csv for:
• Buying patterns
• Customer retention (WillBuyAgain)

🔹 **Part A — Association Rules**
→ Use Apriori to find "frequently bought together" products
→ Provide business recommendations (cross-selling ideas)

🔹 **Part B — Predicting WillBuyAgain**

→ ML models:

✓ RandomForest

✓ Logistic Regression

→ Compare Accuracy, Precision, Recall, F1, ROC-AUC

→ Decide best model for future purchase predictions

Customer Insights + Churn Prediction

Question-14 — Customer Segmentation & Churn Analysis

You are consulting a telecom company using customer_data.csv.

🔹 **Part A — Clustering**

→ Segment customers using K-Means (Income + Spending)

→ Visualize clusters → Explain segment behaviors

🔹 **Part B — Churn Classification**

→ Build + compare:

✓ RandomForest

✓ Logistic Regression

→ Explain why **Recall or ROC-AUC** is more important than accuracy
(e.g., missing actual churners is costly)