Assignment 1.1 (Unit – 1) – Experiential Learning and Case Study

Name: Shrushti Sambhaji Shinde
Enrollment No: ADT23SOCB1090
Roll No: 48 – Batch B
Division: AIEC01
Course: Data Engineering

Answer 1: Research / Experiential Learning

In today's digital world, huge amounts of data are created every second. Hospitals, shops, apps, and even wearable devices continuously generate data. Organizations collect this data and analyze it to make better decisions. To do this effectively, we need to identify good **data sources** and use **modern tools** to process and visualize the information.

Examples of Real-World Data Sources

Healthcare Domain

- Hospital Management Systems → maintain patient records, billing information, and lab reports.
- Smart health devices (fitness bands, smartwatches) → track daily health activities such as heart rate and steps.
- Open datasets such as **WHO Global Health Data** or **Kaggle healthcare datasets →** provide disease statistics, hospital admissions, and patient outcomes.

Retail Domain

- Point of Sale (POS) systems → record transactions at shops.
- E-commerce platforms like Amazon or Flipkart → generate order and customer data.
- Customer feedback forms, loyalty programs, and CRM systems → give insights about buying behavior.
- ERP systems → store inventory and supply chain information.

Integration with Power BI and Modern Platforms

- **Excel** → useful for small datasets, calculations, pivot tables, and simple graphs.

- **Power BI** → connects with Excel, CSV, SQL, Google Analytics, AWS, and BigQuery. It is widely used for creating dashboards like patient admission trends or sales insights.
- **Python** → allows custom analysis with libraries like Pandas, Matplotlib, and Seaborn.
- **Modern Platforms**:
  - Snowflake and BigQuery → handle large-scale datasets.
  - Databricks → useful for big data processing and machine learning.
  - Apache Kafka → supports real-time streaming, such as live monitoring of hospital admissions.
  - ETL Tools (Talend, Apache Airflow) → automate data extraction, cleaning, and loading.

Key Learnings

1. Data always requires cleaning before analysis.
2. Power BI is excellent for business dashboards.
3. Python provides flexibility for deeper analysis and custom visualizations.
4. Large and real-time datasets need modern platforms like Snowflake or Databricks.
5. Even small sample datasets can provide meaningful insights if analyzed properly.

Conclusion

From this research, I understood that **data analytics is not just about tools but about asking the right questions**. With clean data and platforms like Power BI, Python, and modern databases, raw information can be transformed into valuable insights. For example, hospitals can predict which diseases need more attention, while retail shops can understand which products perform best.

Answer 2: Case Study (Mini Project in Healthcare Domain)

Introduction

Hospitals record a variety of data such as patient age, gender, admission date, discharge date, disease, treatment cost, and outcome. Without analysis, this information is just raw numbers. By applying the data lifecycle, hospitals can gain insights that help in resource allocation, budgeting, and planning for seasonal diseases.

In this mini project, I used a **healthcare dataset (1000 patient records)** to demonstrate the complete cycle: **Capture → Store → Clean → Analyze → Visualize**.

Problem Statement

The hospital management wants to analyze:

- Which diseases are most common among patients.
- Average treatment cost for each disease.
- Age-wise distribution of patients.
- Gender ratio in admissions.
- Monthly admission trends.

These insights help the hospital plan resources, manage costs, and improve patient care.

Steps in Data Lifecycle

1. Data Capture

- Dataset: `healthcare_dataset_1000.csv`
- Fields: PatientID, Age, Gender, Disease, AdmissionDate, DischargeDate, TreatmentCost, Outcome
- Represents real hospital records of patient admissions.

2. Data Storage

- The dataset is stored as a **CSV file**.

- In real hospitals, similar data is maintained in databases or EHR (Electronic Health Records).

3. Data Processing & Cleaning (Python)

- Removed duplicate records.
- Filled missing values:
  - Age → replaced with median age.
  - Gender → replaced with "Unknown."
  - Disease → replaced with "Not Specified."
  - Treatment Cost → replaced with average cost.
- Converted admission and discharge dates into proper datetime format.
- Added derived column **Age Group** (20–29, 30–39, etc.) for better segmentation.

4. Analysis and Insights

- **Most Common Diseases:** Flu and Diabetes had the highest number of admissions.
- **Average Treatment Cost:** Cardiac patients had the highest average treatment expenses.
- **Age Distribution:** Most patients were between 40–60 years.
- **Gender Distribution:** Slightly more male patients than female.
- **Monthly Trends:** Admissions peaked during February and June, showing seasonal variation.

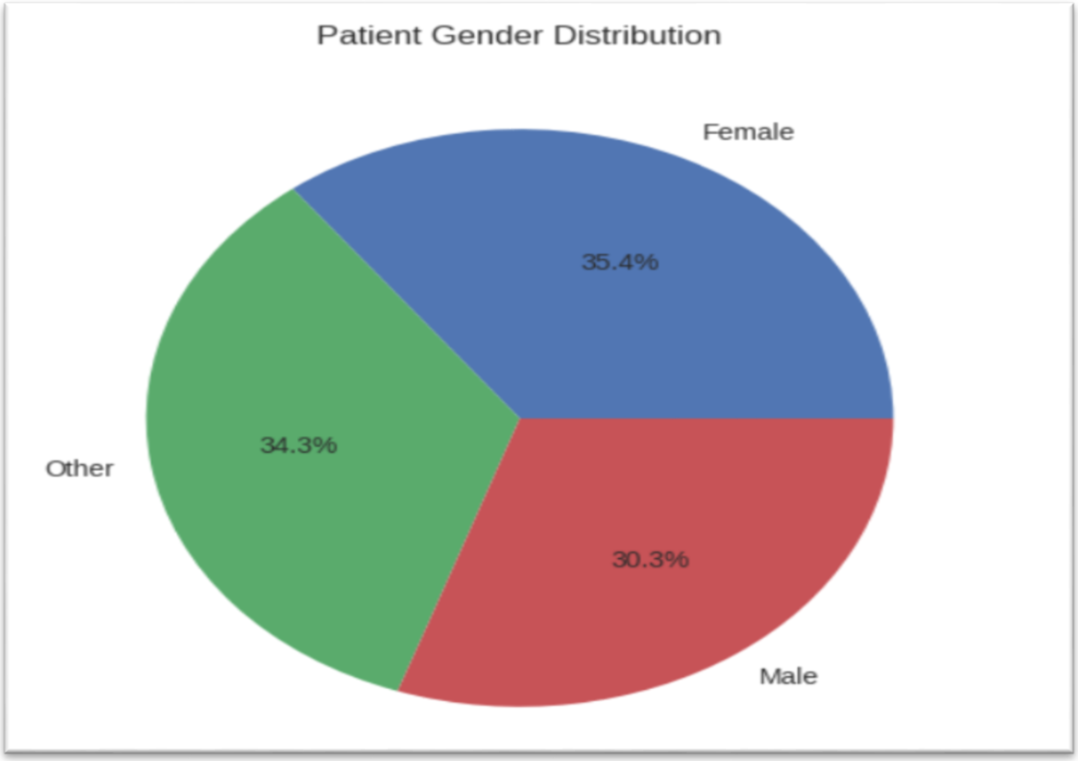
5. Visualization

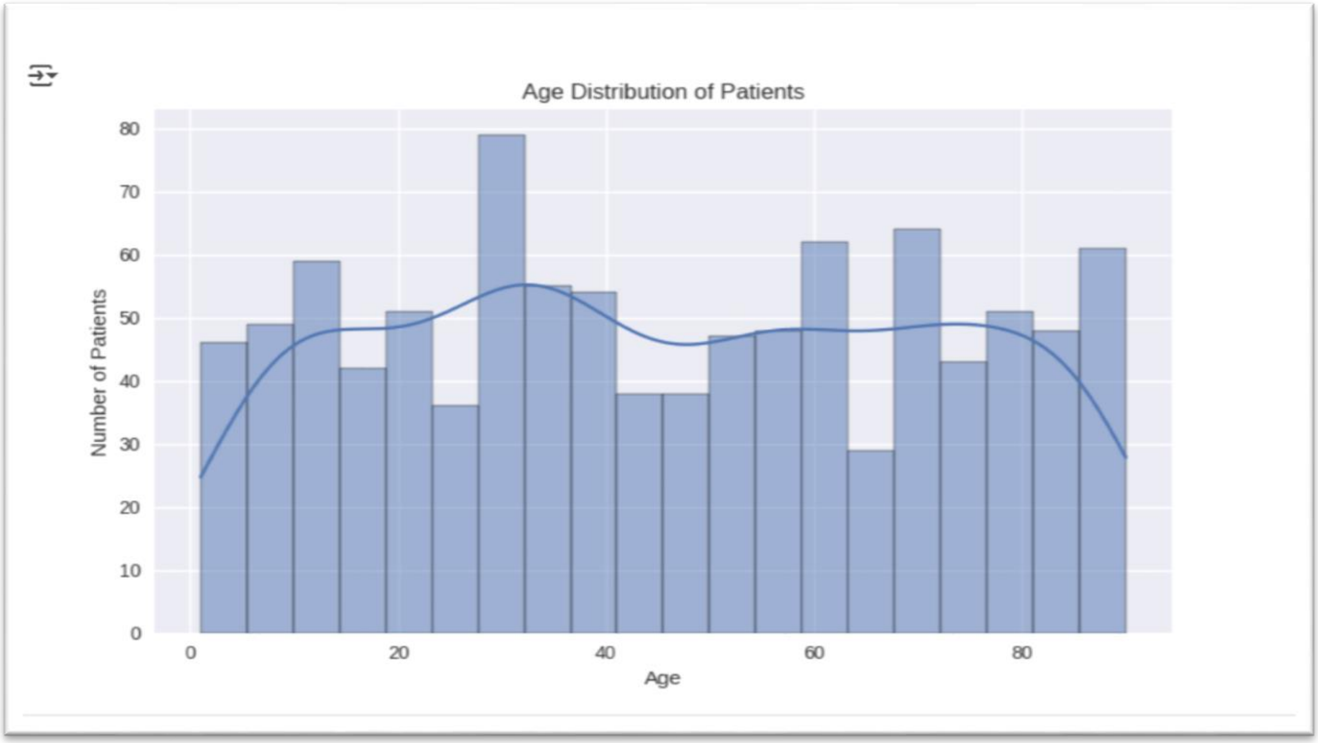
- **Python (Matplotlib/Seaborn):**
  - Bar chart → Disease frequency
  - Horizontal bar chart → Average treatment cost by disease
  - Pie chart → Gender distribution
  - Histogram → Age distribution
  - Line chart → Monthly admission trend
- **Power BI :**

The cleaned dataset was imported into **Power BI** to create an interactive dashboard. The following visuals were developed:

  - Bar Chart → showing the frequency of each disease.
  - Column Chart → displaying the average treatment cost by disease.
  - Pie Chart → representing the gender distribution of patients.
  - Histogram → for the age distribution of patients.
  - Line Chart → showing monthly admission trends over the dataset period.

These charts in Power BI make the data easier to understand and allow hospital management to quickly identify key patterns and insights.





Conclusion

This mini project shows how hospital data can be transformed into actionable insights. By analyzing the dataset, the hospital can:

- Allocate more resources for common diseases like Flu and Diabetes.
- Budget effectively for high-cost treatments such as Cardiac cases.
- Focus on care for middle-aged patients (40–60 years), who formed the majority.
- Prepare in advance for seasonal admission peaks.

The lifecycle **Capture → Store → Clean → Analyze → Visualize** proves to be effective in turning raw data into valuable information that supports better decision-making.

Mini Project GitHub Link

The mini project related to this assignment has been uploaded on GitHub.

Link: https://github.com/shrushti96-dot/hospital-patient-analytics