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

Name: Samarth Sham Kale

Enrollment No: ADT23SOCB0955

Roll No: 46-Batch B

Division: AIEC01

Course: Data Engineering

Answer 1: Research / Experimental Learning

Research and identify real-world data sources and integration with tools like Power BI or modern data platforms.

In today's digital world, huge amounts of data are created every second. Businesses collect this data from different sources and use it for decision-making. Some common real-world examples are:

Retail Domain

- Point of Sale (POS) machines record daily transactions in shops.
- E-commerce websites like Amazon or Flipkart generate online order data.
- Customer data comes from loyalty programs, CRM systems, or feedback forms.
- Inventory and supply chain data is stored in ERP systems.

Healthcare Domain

- Hospital Management Systems keep patient records, billing, and lab reports.
- Smart devices and IoT like fitness bands or smartwatches track health data.
- Insurance companies maintain claim history and premium data.

Integration with Power BI and Modern Platforms

Different tools are used depending on the type of analysis:

1. **Excel** – Best for beginners to do basic calculations, pivot tables, and graphs.
2. **Power BI** – Useful for creating dashboards. It can connect to Excel, CSV, SQL, Google Analytics, AWS, Azure, or BigQuery. Companies use it for sales, profits, and customer behavior dashboards.

3. **Python (with PyCharm IDE)** – For deeper analysis and coding. Using libraries like Matplotlib and Seaborn, I created simple graphs and visualizations. PyCharm made it easy to run and test these programs.

4. **Modern Data Platforms:**
    - Snowflake and BigQuery → store and query large volumes of data.
    - Databricks → big data processing and machine learning.
    - Apache Kafka → real-time streaming (e.g., live sales updates).
    - ETL Tools (Talend, Apache Airflow) → used for extracting, cleaning, and loading data automatically.

Key Learnings

1. Data needs cleaning before analysis.
2. Power BI is excellent for creating business dashboards.
3. Python with Matplotlib helps in visualization beyond Excel.
4. Modern platforms (Snowflake, Kafka, Databricks) are necessary for large or real-time data.
5. Experimental practice on a small dataset builds confidence for handling bigger projects.

Conclusion

Through this research and experimental learning, I understood that **data analytics is not just about using tools but about asking the right questions and finding patterns**. Even a small dataset can give useful insights if analyzed properly. With tools like Excel, Power BI, Python, and modern platforms, it is possible to convert raw data into decisions that support business growth.

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

Introduction

Retail businesses generate large amounts of data every day through customer purchases, billing systems, and inventory management. For a shop owner, it is difficult to manually track which products are performing well, which customers contribute the most to revenue, or how sales vary over time. Data analysis provides a structured way to capture, clean, analyze, and visualize sales data so that better business decisions can be made.

This case study explains how a **sample retail dataset** was created, processed in Python, and visualized using charts. The goal is to help the shop owner identify sales patterns and make informed decisions.

Problem Statement

A shop owner wants to analyze:

- **Daily sales trends** → to know which days perform better.
- **Product performance** → to see which items sell the most.
- **Category-wise revenue** → to check which product categories bring maximum sales.
- **Customer contribution** → to identify loyal or high-spending customers.

By performing this analysis, the owner can decide which products to stock more, which categories need improvement, and which customers can be given loyalty benefits.

Steps in Data Lifecycle (Capture → Visualization)

1. Data Capture

A **sample dataset (retail_sales_1500.csv)** containing **1500 transactions** was prepared to simulate real-world retail sales data.
The dataset contains the following fields:

- **Date** → transaction date
- **Customer** → customer name
- **Product** → product purchased
- **Category** → Electronics, Clothing, Grocery, etc.
- **Price** → per unit cost of the product
- **Quantity** → number of items bought
- This represents the type of data generated by real **Point of Sale (POS) systems** or **ERP software** in shops.

2. Data Storage

The data was stored in a **CSV file**, which is a common format used in business for maintaining sales reports. In real-world retail, this could come from:

- POS systems
- ERP systems
- E-commerce platforms

For this project, CSV simulates the actual business database.

---

3. Data Processing & Cleaning (Python)

Using Python (in **PyCharm IDE**), the dataset was cleaned and prepared for analysis:

- **Removed duplicates** → avoided repeating transactions.
- **Handled missing values** → ensured consistency in data.
- **Created new column** → Total_Sales = Price × Quantity to calculate the sales value of each transaction.

This cleaned dataset became the base for further insights.

---

4. Analysis and Insights

Here are the **actual insights from your 1500-entry retail dataset**:

- **Total Revenue**: ₹32,82,55,375 (overall sales from all transactions)
- **Best Selling Product**: **Jacket**
- **Top Customer**: **Michael**
- **Revenue by Category**:
    - Electronics → ₹11,34,54,394
    - Clothing → ₹11,03,36,214
    - Grocery → ₹10,44,64,767
- **Daily Sales Trend (Peak Days)**:
    - 17th July 2025 → ₹87,70,031
    - 30th July 2025 → ₹85,73,545
    - 27th July 2025 → ₹80,27,285
    - 19th August 2025 → ₹79,10,398

These results directly show the shop owner **where business is strong and where improvements are needed**.

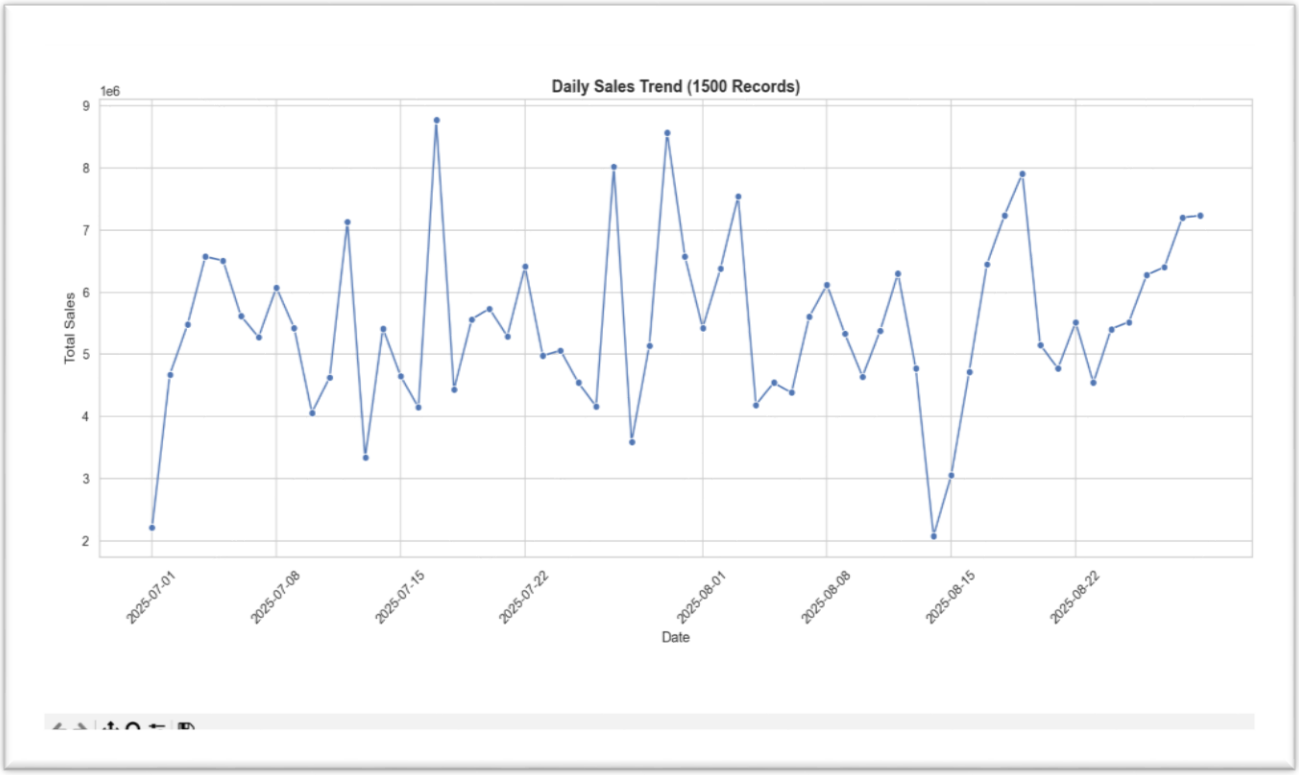
---

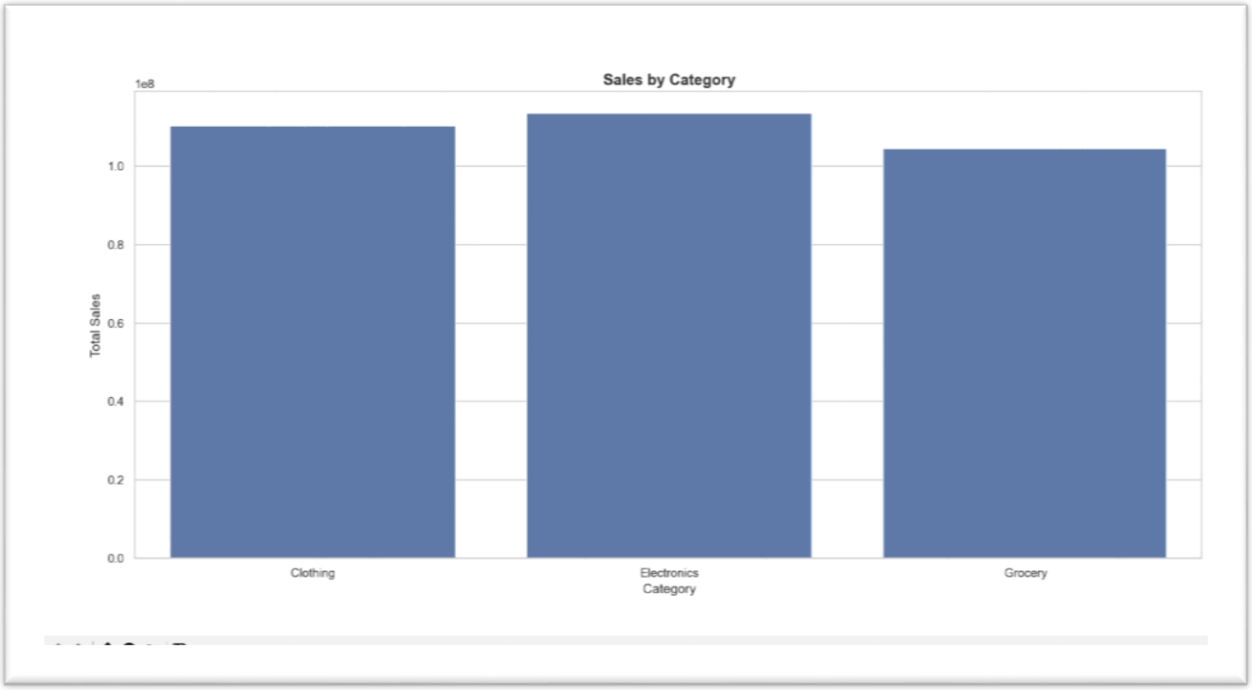
5. Visualization (Python – Matplotlib/Seaborn)

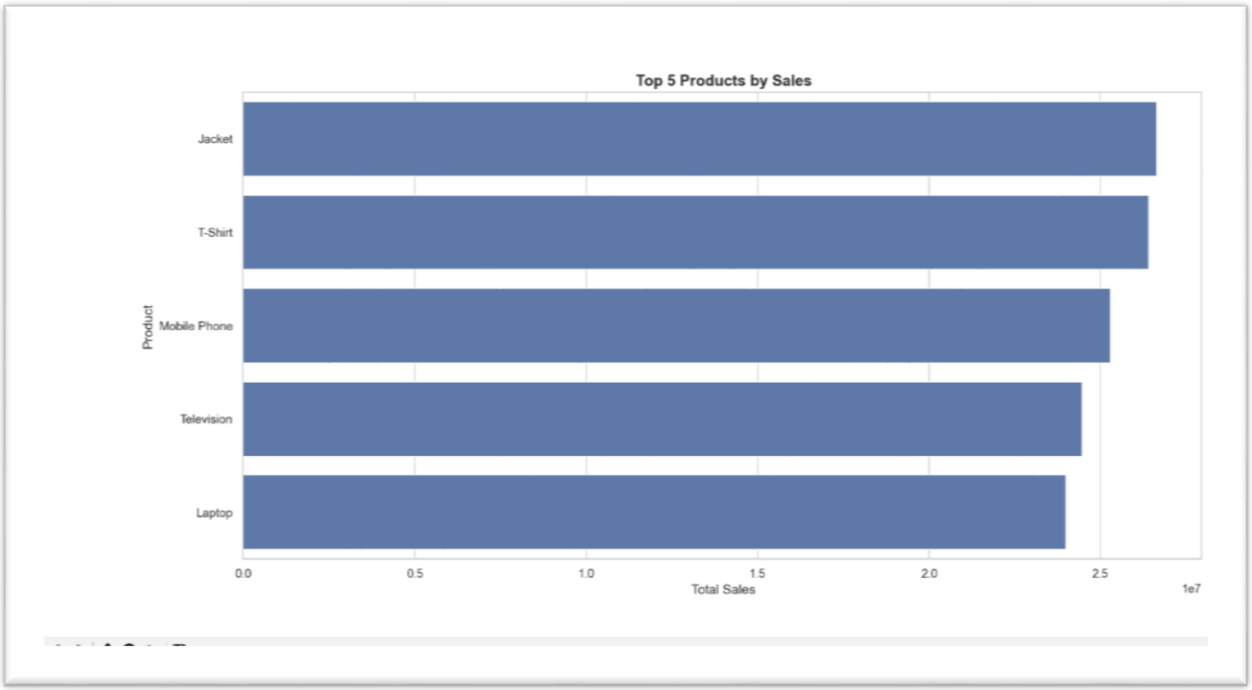
To make the insights easier to understand, visual charts were created:

1. **Daily Sales Trend → Line Chart**
   Shows how sales changed over different dates, highlighting peak sales days.
2. **Sales by Category → Bar Chart**
   Compares Electronics, Clothing, and Grocery sales, showing Electronics as the leader.
3. **Top 5 Products by Sales → Horizontal Bar Chart**
   Ranks the best products, showing clearly which items bring the most revenue.

OUTPUT (Screenshots):




Sales by Category
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Total Sales
Clothing
Electronics
Grocery
Category


Top 5 Products by Sales
Jacket
T-Shirt
Mobile Phone
Television
Laptop
Product
0.0
0.5
1.0
1.5
2.0
2.5
1e7
Total Sales

Conclusion

This mini project demonstrates how a simple retail dataset can be transformed into valuable insights using Python. The lifecycle followed was **Capture → Store → Process → Analyze → Visualize.**

From the results, the shop owner can:

- Focus on stocking more electronics, especially laptops.
- Reward loyal customers like Ramesh to retain them.
- Monitor low-performing categories like Grocery.
- Plan special offers on peak days to maximize revenue.

Mini Project GitHub Link

The mini project related to this assignment has been uploaded on GitHub.
Link: https://github.com/SamarthKale5556/data-engineering-mini-project