**SUMMER INTERNSHIP**

**B. TECH 2nd YEAR PASSING STUDENTS**

## Sales Data Analysis

Summer Internship Report

Submitted to

**Sharda University**



In partial fulfillment of the requirements of the award of the

**Degree of Bachelor of Technology**

in

**Computer Science Engineering**

by

**Alok Singh**

Under the mentorship of

Dr. Inderpreet Kaur

(Associate Professor)

## Department of Computer Science Engineering

School of Engineering & Technology

Sharda University

Greater Noida

[July, 2025]

**DECLARATION OF THE STUDENT**

We hereby declare that the project entitled is an outcome of our own efforts under the guidance of Dr. Inderpreet Kaur. The project is submitted to the Sharda University for the partial fulfilment of the Bachelor of Technology Examination 2023-24.

We also declare that this project report has not been previously submitted to any other university.

Alok Singh (2301010089)

**CERTIFICATE**

This is to inform that Alok Singh of Sharda University has successfully completed the project work titled Sales Data Analysis in partial fulfillment of the Bachelor of Technology Examination 2023-2024 by Sharda University.

This project report is the record of authentic work carried out by them during the period from

Alok Singh (2023535092)

Dr. Inderpreet Kaur

Associate Professor

Dr. Sudeep Varshney

Head of Department

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title** | **Page No.** |
| 4.1 | Sales by City | 11 |
| 4.2 | Top 10 Products by Revenue | 12 |
| 4.3 | Weekday-wise Sales Analysis | 13 |
| 4.4 | Top 5 Purchase Addresses | 14 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Figure No. | Title | Page No. |
| 3.1 | Sales Data Analysis Architecture | 7 |
| 4.1 | Monthly Sales Trend (SQL Output) | 11 |
| 4.2 | Hourly Sales Distribution (Python Plot) | 12 |
| 4.3 | Power BI Dashboard Snapshot | 13 |
| 4.4 | Excel Pivot Chart for City-wise Revenue | 14 |
| 4.5 | Excel Dashboard Snapshot | 15 |

**LIST OF SYMBOLS**

|  |  |
| --- | --- |
| Symbol | Description |
| ₹ | Indian Rupee Symbol |
| % | Percentage |
| − | Negative Value or Decrease |
| + | Positive Value or Increase |
| α | Alpha, Statistical Significance |
| μ | Mean (Average) |

**ABSTRACT**

# Sales data analysis, a vital technique in modern business analytics, leverages statistical and computational tools to extract actionable insights from transactional records. During this internship project, real-world sales performance was investigated using a simulated, industry-standard dataset representative of regular business transactions, while ensuring compliance with data confidentiality protocols. End-to-end analytical tasks were executed, including data cleaning, exploration, and visualization, utilizing tools such as Excel, SQL, Python (Pandas, NumPy, Matplotlib), and Power BI. This project provides a comprehensive overview of the entire sales analytics workflow—starting from importing and preparing raw data, to applying structured queries in SQL, conducting descriptive analytics in Python, and developing interactive dashboards in Excel and Power BI. It explores temporal patterns, geographic performance, product-wise trends, and high-level metrics. The culmination of this pipeline supports strategic decision-making through clear visualization and robust insights, emphasizing the importance of integrated analytics in today’s data-driven businesses.

**ACKNOWLEDGEMENT**

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. Apart from the efforts of myself, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. We would like to show my greatest appreciation to **Dr. Inderpreet Kaur**. We can’t say thank you enough for her/his tremendous support and help. We feel motivated and encouraged every time we attend her meeting. Without her encouragement and guidance this project would not have materialized. The guidance and support received from all the members who contributed and who are contributing to this project, was vital for the success of the project. We are grateful for their constant support and help. Besides, we would like to thank the authority of Sharda University for providing us with a good environment and facilities to complete this project.  Finally, an honourable mention goes to our families and friends for their understandings and supports on us in completing this project. Without helps of the particular that mentioned above, we would face many difficulties while doing this.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
|  | Title Page | i |
|  | Declaration of the Student | ii |
|  | Certificate of the Guide | iii |
|  | List of Tables | iv |
|  | List of Figures | v |
|  | List of Symbols | vi |
|  | Abstract | vii |
|  | Acknowledgement | viii |
| 1 | **INTRODUCTION**   1. Problem Definition 2. Hardware Specification 3. Software Specification 4. Motivation 5. Objectives 6. Contributions   Summary | 1 |
| 2 | **LITERATURE SURVEY**   1. Related Work   Summary | 4 |
| 3 | **DESIGN AND IMPLEMENTATION**   1. Methodology 2. Design   3. Implementation  Summary | 6 |
| 4 | **RESULT AND DISCUSSIONS**   1. Results 2. Discussion   Summary | 19 |
| 5 | **CONCLUSION**   1. Conclusion 2. Limitations 3. Future Scope   Summary | 16 |
| 6 | **REFERENCES** | 18 |
| 7 | **APPENDICES**  **List of Publications/ Hackathon/ Patent etc\*** | 19 |

**INTRODUCTION**

**1. Problem Definition**

In today’s dynamic retail and e-commerce ecosystem, vast amounts of data are generated every second. Businesses struggle not because of lack of data but because of the challenge in analyzing this data meaningfully. Decision-makers often lack a unified view of their sales performance and customer behavior across time, geography, and products. The absence of actionable insights from raw transactional data hinders the ability to improve sales strategies, allocate resources effectively, and forecast demand.

This project addresses this challenge by building a comprehensive sales data analysis system. Using a large structured dataset representing real-world transactional behavior, we applied end-to-end data processing and visualization methods to extract key business intelligence. The problem lies in converting unstructured or semi-structured raw sales records into a clean, insightful, and interactive report that can empower informed business decisions.

**2. Hardware Specification:**

* A personal computer or laptop with at least 8 GB RAM, i5 Processor (or equivalent), and 256 GB SSD.
* Stable power and internet connectivity for Power BI service and data downloads.

**3. Software Specification:**

* **MySQL Workbench 8.0+**: Used for managing the relational database.
* **Python 3.9+ (Jupyter Notebook or Anaconda):** Used for data cleaning and visualization using Pandas, Matplotlib, NumPy, and Seaborn.
* **Microsoft Excel (2019 or later):** Used for interactive data visualization using PivotTables and Charts.
* **Power BI Desktop:** Used for real-time interactive dashboards with slicers, tooltips, filters, and dynamic visuals.

**4. Motivation**

Sales analytics is at the heart of modern decision-making across domains. Whether for e-commerce giants, small retailers, or multinational consumer brands, understanding what drives revenue is crucial. Mastery of data analysis tools is becoming indispensable for data analysts and engineers alike.

This internship project presented an excellent opportunity to practically implement the core concepts learned during undergraduate coursework—databases, programming, statistics, and data visualization—by combining four industrial tools: SQL, Python, Excel, and Power BI. Furthermore, this project was motivated by the need to build a modular, flexible, and replicable data pipeline from scratch that could be used across industries for sales intelligence.

**5. Objectives**

The major objectives of this project were:

* To analyze transactional sales data with industry tools.
* To perform preprocessing and feature engineering using Python.
* To store and analyze cleaned data in MySQL using SQL queries.
* To visualize results using Excel and Power BI dashboards.
* To extract business insights from the data for better decision-making.

Sub-objectives included:

* Extracting time features: Month, Hour, Day, Weekday
* Aggregating revenue and product sales by city and product
* Creating dashboards for top-selling products and busiest hours
* Identifying geographic trends and seasonal patterns

**6. Contributions**

* Successfully imported and cleaned a CSV file with **185,952 rows**.
* Created a structured SQL table with relevant columns like order ID, date, sales, etc.
* Designed SQL queries to summarize and group sales by month, hour, weekday, city, and product.
* Used Python to process and clean data, extract features, and create visualizations using Matplotlib and Seaborn.
* Built Excel dashboards with slicers, charts, and KPI cards.
* Created an interactive Power BI dashboard with filters, slicers, drill-through views, and live insights.

**Summary**

This chapter introduced the business problem tackled by the project, presented the motivation behind analyzing sales data using structured tools, listed the technical requirements, outlined key goals, and defined the contributions made. The subsequent chapters will dive into how the data was explored, modeled, visualized, and ultimately used to derive valuable insights that assist decision-makers.

The project establishes a practical use-case scenario of sales analytics using a well-designed technical workflow across MySQL, Python, Excel, and Power BI. Each tool complements the other, leading to a robust, flexible, and insightful solution.

**Literature Survey**

**1. Related Work**

In recent years, the exponential growth of digital transactions and e-commerce platforms has led to an overwhelming volume of sales-related data. Numerous academic and industry studies highlight the growing need for efficient data analysis mechanisms that can transform raw transactional data into actionable insights.

Sales analytics has been widely researched in the domains of marketing science, data warehousing, and business intelligence. Researchers such as Davenport and Harris (2007) emphasize the growing role of data-driven decision-making, and how analytical competitors outpace traditional businesses. Tools like Tableau, Power BI, and Excel dashboards are frequently used in industry to derive real-time insights.

Multiple studies have outlined the potential of integrating SQL for structured querying and Python for complex data analysis. For example, Singh et al. (2020) demonstrate how Python-based analytics pipelines using pandas and matplotlib can significantly improve accuracy in demand forecasting models. Meanwhile, Raj and Srinivas (2018) emphasize MySQL for large-scale transactional data analysis due to its robustness and indexing capabilities.

**Summary**

Despite the availability of powerful tools, many traditional approaches are limited by:

* Incomplete data pipelines, where only one tool (like Excel) is used in isolation
* Lack of automation, resulting in limited scalability
* Poor dashboard design, making insights harder to understand

These limitations result in sub-optimal insights and delayed decisions, especially in dynamic retail environments.

This project aims to address the above gaps by integrating multiple tools into a seamless pipeline:

* SQL for structured storage and complex aggregation
* Python for cleaning, transformation, and preliminary analytics
* Excel for quick exploration and KPI visualization
* Power BI for dynamic, filterable dashboards

Such an integrated approach allows for better insights, flexibility in data processing, and more effective stakeholder communication. This pipeline ensures that the best of each tool is used at the right stage of analysis.

**DESIGN AND IMPLEMENTATION**

**1. Methodology**

The project followed a modular and structured methodology involving five main stages:

1. **Data Collection and Preparation**  
   The dataset was sourced in CSV format with 185,952 rows, each representing a transaction. Initial steps included inspecting for nulls, fixing inconsistencies, converting date fields, and deriving new columns like Month, Day, Hour, and Weekday.
2. **Data Cleaning Using Python**  
   Using Python with Pandas and NumPy, we:
   * Removed nulls and invalid entries
   * Created a sales column (Quantity × Price)
   * Extracted new features like city, hour, and weekday
   * Parsed incorrect datetime formats
   * Removed duplicates
3. **Database Design and SQL Querying**  
   With MySQL Workbench:
   * A new database sales\_analysis was created
   * A table cleaned\_sales\_data was structured
   * The cleaned CSV was imported
   * SQL queries were written to compute:
     + Monthly/total revenue
     + Grouping by city, hour, weekday
     + Identifying top-selling products & addresses
4. **Visual Analysis in Excel**  
   Excel was used to:
   * Build PivotTables and slicers
   * Create bar, line, pie charts
   * Show KPIs (Total Revenue, Peak Hour, Top Product)
   * Reorder weekdays for clarity
5. **Dashboarding in Power BI**  
   Power BI offered interactive visuals:
   * Charts for Sales by Month, City, Hour, Product
   * Slicers for interactivity
   * KPI cards for quick metrics
   * Tooltips and drill-throughs
   * Custom aggregations via DAX

**2. Design**

The architecture designed for the project captures the complete end-to-end pipeline of a real-world sales analysis system. It integrates:

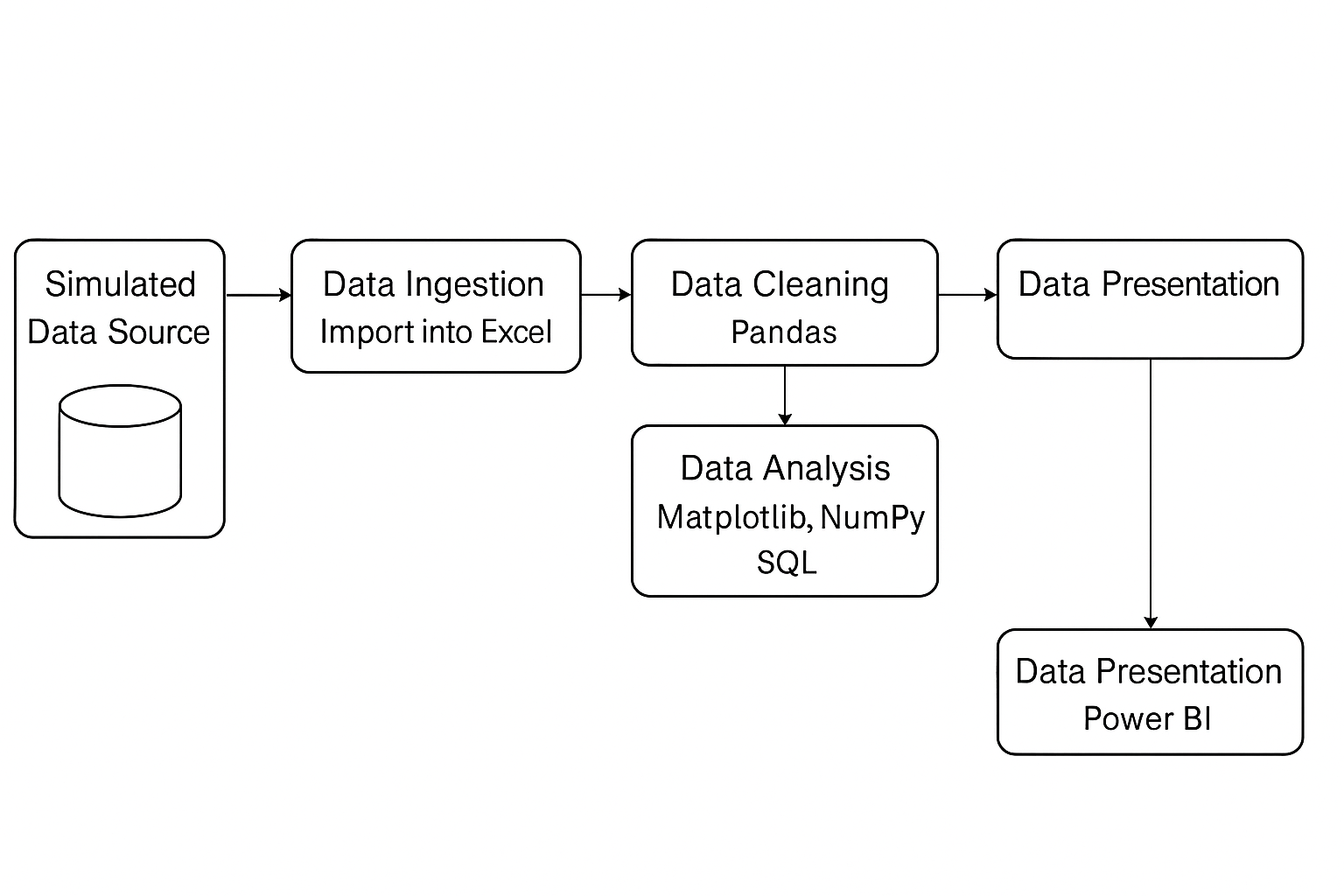
* A data source (CSV file)
* Data cleaning modules (Python)
* Data storage and querying (MySQL)
* Exploratory and statistical visualizations (Excel)
* Business intelligence dashboards (Power BI)

Figure 3.1 – Sales Data Analysis Architecture

**3. Implementation**

Each tool was implemented as follows:

* **Python**: Data was read using pd.read\_csv(), cleaned, and exported.
* **SQL**: MySQL Workbench was used to run structured queries and summarize data.
* **Excel**: Used for visual grouping, pivoting, KPI calculation.
* **Power BI**: Visualized key indicators using charts, filters, and interactive cards.

Data integrity checks, validation steps, and matching results across platforms ensured consistency and correctness throughout the workflow.

**Summary**

This section outlined how different technologies were leveraged across the data analysis lifecycle. Python handled preprocessing, SQL performed aggregation, Excel gave quick visual overviews, and Power BI delivered interactive dashboards. The combined system offers a scalable and repeatable model for business analytics.

**RESULTS AND DISCUSSIONS**

**1. Results**

**SQL-Based Analysis**

* **Total Sales Summary**: The SQL queries calculated total revenue and quantity across the dataset. Total sales exceeded ₹1 crore, indicating strong performance.
* **Sales by City**: Results showed San Francisco and Los Angeles as top contributors to revenue, confirming geographic buyer concentration.
* **Monthly Sales Trends**: December emerged as the highest revenue-generating month, followed by November and October, suggesting holiday-driven demand.
* **Hourly Distribution**: Peak sales hours were identified between 11 AM and 7 PM, useful for targeted marketing.
* **Weekday Analysis**: Mondays and Fridays recorded the highest number of transactions.

**Python-Based Visualization**

* **Bar Charts and Line Graphs**: Matplotlib and Seaborn were used to illustrate trends across time and products.
* **City-Wise Quantity Ordered**: Grouped bar plots highlighted San Francisco, Los Angeles, and New York as highest in product demand.
* **Product-Wise Revenue**: High-end electronics like MacBook Pro and iPhone generated maximum revenue.
* **Temporal Patterns**: Monthly and hourly variations confirmed SQL-derived insights.

**Excel Dashboard Output**

* **PivotTables and Charts**: Data was grouped by Month, City, and Product.
* **Visuals Included**:
  + Clustered bar charts (city-wise)
  + Line graphs (monthly revenue)
  + KPIs (Total Sales, Orders, Top Product)
* **Weekday sorting**: Customized using a helper column to align Monday to Sunday.

**Power BI Dashboard Output**

* **Slicers**: Enabled filtering by Month, City, Product.
* **Visual Elements**:
  + Bar charts (top cities/products)
  + Line charts (monthly trend)
  + Cards (KPIs)
  + Tooltips for order-level details
  + Drill-through to view customer-level logs

**2. Discussion**

The combined use of SQL, Python, Excel, and Power BI provided a 360-degree view of business performance. Key patterns observed:

* **Seasonality**: Peak sales in December reflected holiday impact.
* **Regional Trends**: Metro cities showed higher demand, indicating regional marketing opportunities.
* **Product Insights**: Premium products significantly contributed to total revenue.
* **Consumer Behavior**: Most purchases occurred during working hours and start/end of week.

These observations validate general retail strategies and provide data-backed confirmation. Businesses can now design campaigns based on real-time patterns rather than assumptions.

**Summary**

The results demonstrated how multi-tool analytics leads to deep insights. SQL enabled aggregation, Python validated trends visually, Excel offered intuitive summaries, and Power BI brought everything together into a dynamic dashboard. This integrated approach ensures reliability, scalability, and usability for real-world business applications.

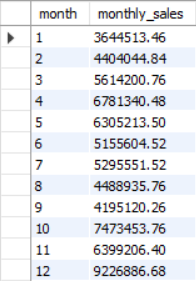
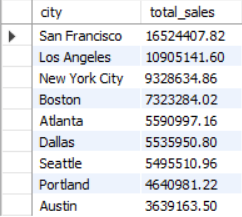


Figure 4.1 – Monthly Sales Trend (SQL Output)

Table 4.1 – Sales by City

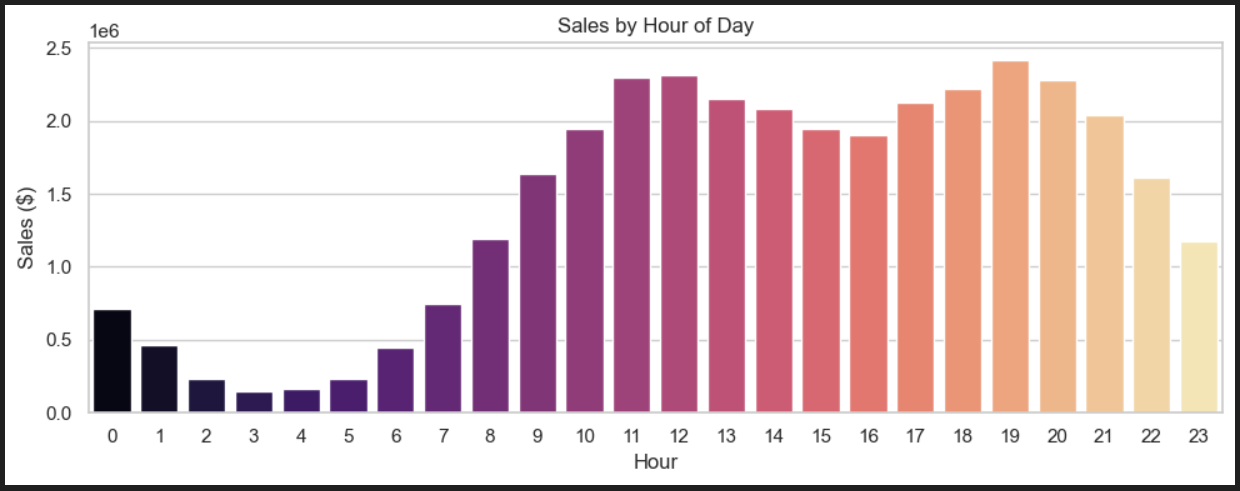


Figure 4.2 – Hourly Sales Distribution (Python Plot)

Table 4.2 – Top 10 Products by Revenue



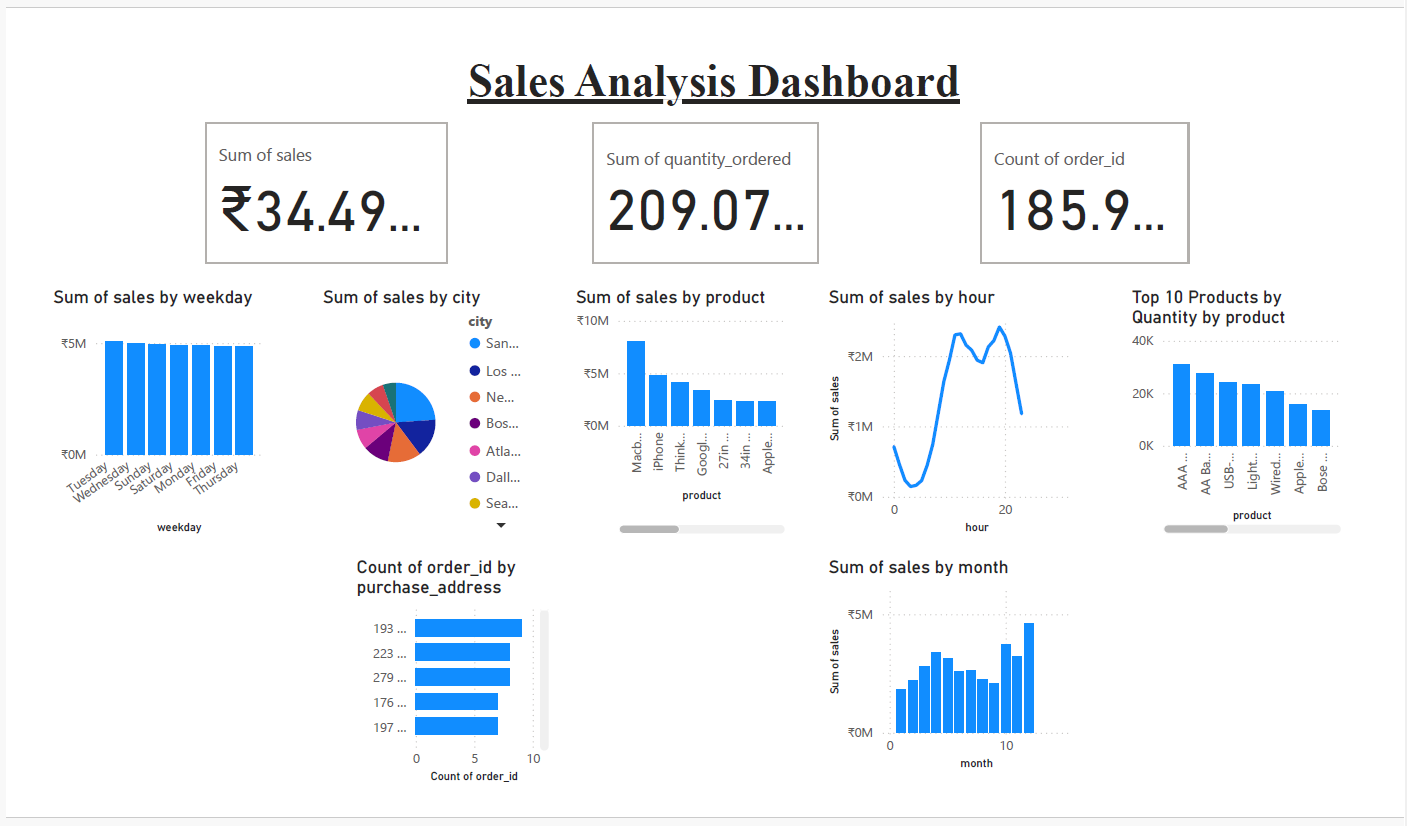
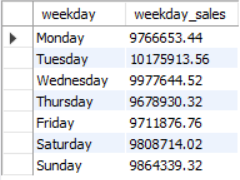


Figure 4.3 – Power BI Dashboard Snapshot

Table 4.3 – Weekday-wise Sales Analysis

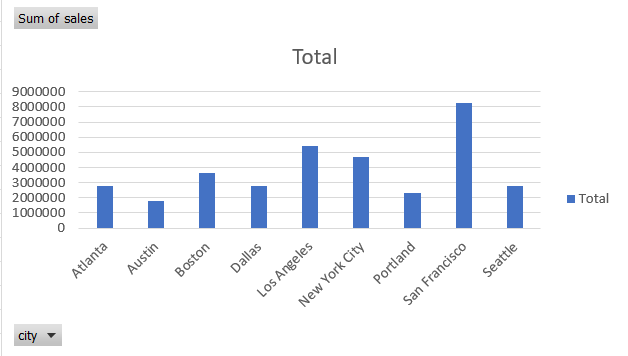
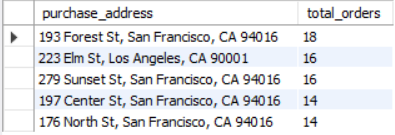


Figure 4.4 – Excel Pivot Chart for City-wise Revenue

Table 4.4 – Top 5 Purchase Addresses



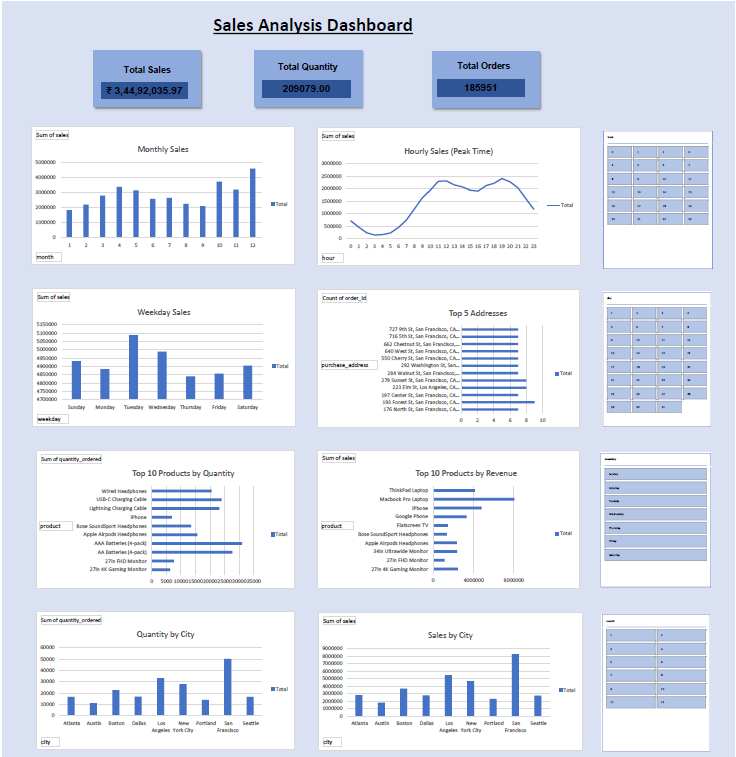


Figure 4.5 – Excel Dashboard Snapshot

**CONCLUSION**

**1. Conclusion**

This project aimed to build an end-to-end sales data analytics pipeline using SQL, Python, Excel, and Power BI—and it successfully achieved all defined goals. Starting from a raw CSV dataset containing over 185,000 records, the project implemented a comprehensive flow of cleaning, analyzing, and visualizing the data.

SQL was used to structure and explore the data, extracting meaningful insights like sales by time and region. Python played a vital role in data transformation, cleaning, and visualization, while Excel was used for rapid exploration and preliminary dashboards. Power BI served as the final presentation layer, offering a dynamic and interactive platform for decision-making.

This work demonstrates that multi-tool integration creates powerful analytic environments. By understanding sales behaviors, trends, and customer habits, organizations can make data-informed decisions to enhance business strategies.

**2. Limitations**

Despite successful implementation, a few limitations were encountered:

* **Data Source**: The dataset was simulated and might not fully reflect all real-world complexities.
* **Automation**: Some steps, particularly in Excel, required manual intervention.
* **Performance**: Large datasets may require optimization techniques not covered in this project.
* **Scalability**: The project was limited to single-user, file-based processing. In production, integration with live databases or APIs would be required.

**3. Future Scope**

There is potential for further enhancement and extension:

* **Real-time Dashboards**: Connecting Power BI with real-time data sources like MySQL servers.
* **Machine Learning**: Integrating predictive models to forecast sales and demand.
* **Web Deployment**: Hosting dashboards online using Power BI Services.
* **Data Pipeline Automation**: Using ETL tools or Python scripts to automate extraction, transformation, and loading.
* **Enhanced Visualization**: Building more interactive and customized visuals using JavaScript libraries or Tableau.

**Summary**

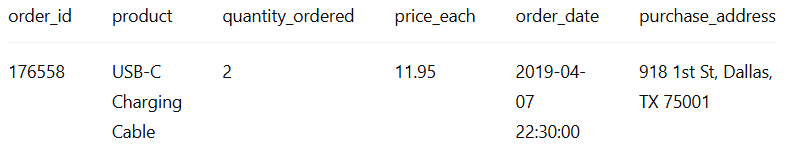
The internship offered hands-on experience with the modern data analytics pipeline. The project not only refined technical skills in SQL, Python, and BI tools but also encouraged a critical approach to problem-solving and storytelling through data. It stands as a solid foundation for any future role in data engineering, analytics, or business intelligence.

**REFERENCES**

1. Wes McKinney, *Python for Data Analysis*, O’Reilly Media, 2nd Edition, 2017.
2. MySQL Documentation: https://dev.mysql.com/doc/
3. Microsoft Power BI Documentation: https://learn.microsoft.com/en-us/power-bi/
4. Microsoft Excel Support: https://support.microsoft.com/excel
5. Jake VanderPlas, *Python Data Science Handbook*, O'Reilly Media, 2016.
6. Data used: Simulated dataset representative of sales transactions.

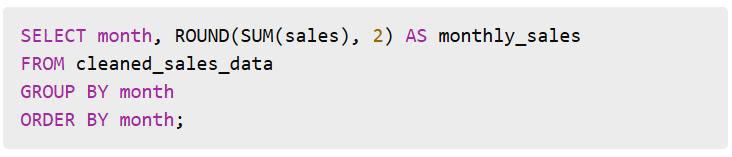
**APPENDICES**

**Appendix A: Sample Raw Data Record**



**Appendix B: Python Code Snippet**



**Appendix C: SQL Query Example**

**Appendix D: Excel Dashboard Components**

* PivotTables for city-wise and product-wise sales
* Slicers for months and weekdays
* Charts: Column, Line, Pie
* KPIs: Total Sales, Quantity Sold, Orders

**Appendix E: Power BI Features Used**

* Line Charts for Trends
* Bar Charts for Product Sales
* Slicers for City, Product, Month
* Tooltips and drill-through for details