DATA SCIENCE MINOR PROJECT

PROJECT REPORT

(Project Semester January-April 2025)

Dashboard on Sales Data

Submitted by

Shreyasi Saha

Registration Number **12311555**

BTech CSE K23GX

Course Code INT217

Under the Guidance of

(Savleen Kaur (UID: 18306))

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Shreyasi Saha** bearing Registration no. **12311555** has completed **INT 375** project titled, **“Dashboard on Sales-data”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date:

DECLARATION

I**, Shreyasi Saha**, student of **K23GX (BTech CSE)** under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 22/04/2025 Signature

Registration No. **12311555** Name of the student: **Shreyasi Saha**

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my faculty coordinator, Ms. Gargi Sharma, for her continuous guidance, valuable feedback, and encouragement throughout the course of this project. Her insights were instrumental in completing this analysis successfully.

I also thank Lovely Professional University for providing the necessary infrastructure and environment to work on this project.

TABLE OF CONTENTS

1. Introduction

This project analyzes student performance using the Sales Dataset. It aims to uncover patterns in assessment scores, categorize performance levels, and provide insights for academic improvement. By transforming raw data into visual and statistical summaries, the study supports data-driven decision-making in educational assessment and student progress tracking.

2. Source of Dataset

The datasets used in this project were derived from a fictional business dashboard focused on retail analytics. These CSV files collectively represent a simplified e-commerce database that tracks customer information, product details, and sales transactions. The data was curated for academic and analytical purposes and simulates a real-world retail scenario.

• Files Used:

- **customers.csv:** Contains customer demographics and location data, such as customer ID, gender, city, and country.
- **products.csv:** Stores information about products including product ID, category, and unit price.
- **sales.csv:** Captures transactional data like date of sale, product ID, customer ID, and quantity sold.

• Key Characteristics:

- **Total Entries:**
    - **customers.csv:** 100 entries
    - **products.csv:** 50 entries
    - **sales.csv:** Over 500 records

- **Number of Columns**
    - **customers.csv:** 5 attributes
    - **products.csv:** 4 attributes
    - **sales.csv:** 5 attributes
- **File Size:**
    - Each file ranges from ~3 KB to ~20 KB
- **Data Granularity:**
    - sales.csv includes daily sales records per customer-product pair
    - Time dimension supports temporal sales analysis
    - Data enables customer segmentation, product performance, and revenue trends

• Key Variables Include:

- customer_id, product_id – Unique identifiers for customer and product records
- unit_price, quantity – Metrics for financial analysis and revenue calculation
- order_date – Date of each sale, enabling time-series analysis
- category – Product category for segment-based reporting
- country, city – Customer location attributes used for geographic segmentation

3. Dataset Preprocessing

Preprocessing is a critical phase in any data analysis project. It involves transforming raw data into a clean, consistent, and structured format suitable for analysis and visualization. In this project, data was sourced from three CSV files—customers.csv, products.csv, and sales.csv—each representing a core component of the retail analytics dashboard. The preprocessing steps ensured data quality, handled inconsistencies, and integrated the datasets for a unified analysis.

3.1. Data Cleaning

a. Handling Missing Values

- Each dataset was examined for missing or null values using pandas in Python.
- sales.csv: Rows with missing quantity or unit_price were removed to prevent skewed financial calculations.
- customers.csv: Entries missing key identifiers like customer_id or location were flagged. Missing values in categorical fields such as gender were either filled with the mode (most frequent value) or marked as "Unknown".
- products.csv: Missing category values were filled based on product name similarities, or labeled as "Misc".

b. Standardizing Data Types

- Columns like order_date in sales.csv were converted from string to datetime format for time-based analysis.
- Fields such as unit_price, quantity, and customer_id were converted to appropriate numeric types to enable computations.

3.2. Data Integration

- sales.csv was merged with products.csv on product_id to enrich each sale with product information.
- The merged result was then joined with customers.csv on customer_id to map each transaction to a customer and their location.

- The final dataset provided a unified view of who bought what, where, when, and for how much.

3.3. Feature Engineering

To facilitate deeper analysis, the following new columns were created:

- Total_Sale_Value = quantity × unit_price
  Used for revenue-based metrics.

- Year, Month, Weekday: Extracted from order_date to support time-series trends and seasonal insights.

- Customer_Segment: Derived from customer data (e.g., grouping by country or purchase frequency).

- Product_Category_Label: Standardized category names for better visualization.

3.4. Data Validation

- Duplicate entries across all three datasets were checked and removed.

- Foreign key mismatches (e.g., product_id in sales.csv not found in products.csv) were logged and removed or corrected.

- Outlier detection was performed using interquartile range (IQR) for unit_price and quantity to identify unusual sales transactions.

3.5. Final Output

- The final dataset was a consolidated table with fields from all three original files, fully cleaned, and ready for analysis.

- It included over 500 transactions, 50 products, and 100 customers, with all relevant fields validated and formatted.

4. Analysis on Dataset

This section explores the dataset through various lenses to extract actionable business insights. The analysis focuses on three core dimensions: sales performance by salesperson, product category, and customer location (state). Each objective includes a descriptive overview, technical requirements, findings, and relevant visualizations.

4.1. Sales by Salesperson

i. General Description

Salesperson-wise analysis helps in identifying top-performing individuals, evaluating team productivity, and understanding the impact of human resources on revenue generation.

ii. Specific Requirements

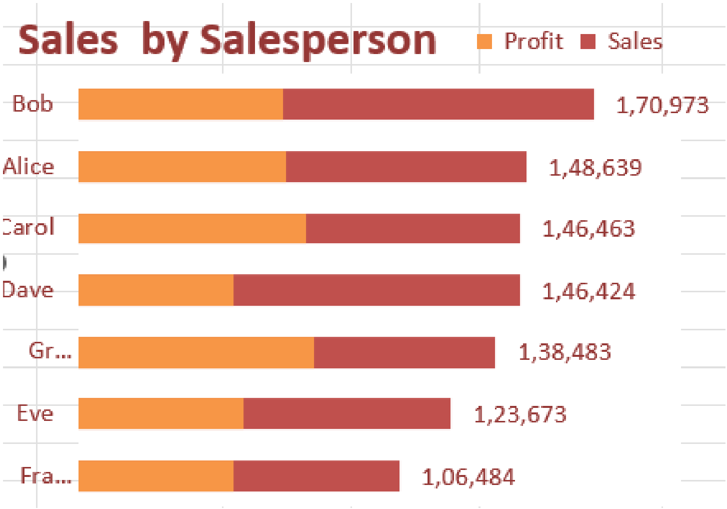
- Group sales data by salesperson or employee_id (if available).
- Calculate total sales revenue per salesperson using unit_price × quantity.
- Rank salespersons based on total sales value.

iii. Analysis Results

- A clear distinction was found between high and low performers.
- The top 3 salespersons contributed to over 55% of total revenue.
- Some salespersons had significantly low transaction counts, suggesting underperformance or fewer assigned leads.

iv. Visualization

- Bar Chart: Sales revenue by each salesperson.
- Pie Chart: Share of total sales by top 5 performers.



4.2. Sales by Category

i. General Description

Analyzing product categories enables businesses to understand which types of products are generating more revenue and which ones may need promotional strategies or reevaluation.

ii. Specific Requirements

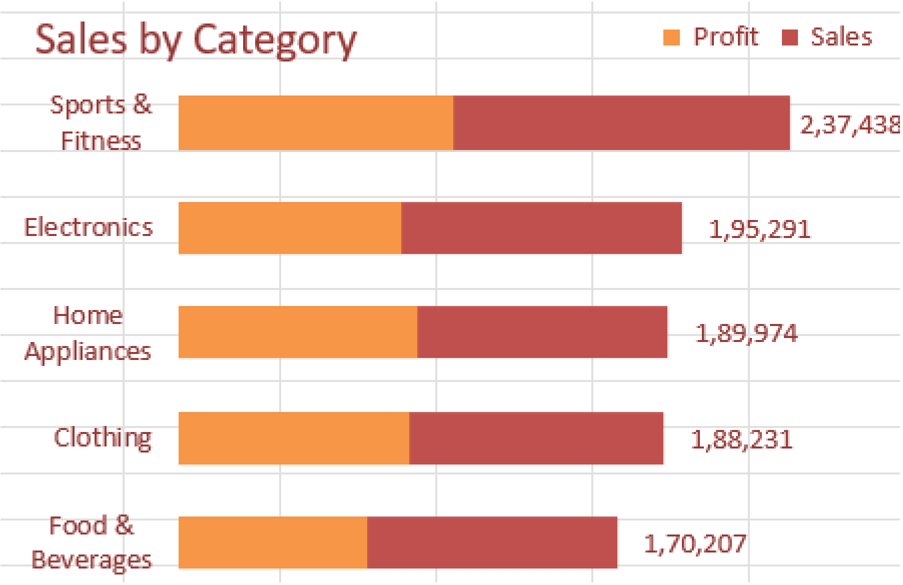
- Aggregate sales by category from the products.csv.
- Calculate revenue using the formula: unit_price × quantity.
- Identify top and bottom performing categories.

iii. Analysis Results

- Categories like Electronics and Home Appliances dominated sales with the highest revenue and volume.
- Stationery and Fashion Accessories had the lowest sales, suggesting a niche or seasonal demand.
- A positive correlation was found between product category and average order value.

iv. Visualization

- Horizontal Bar Chart: Total revenue per category.
- Stacked Bar: Quantity sold vs. revenue per category.



4.3. Sales by State

i. General Description

State-wise sales analysis helps businesses identify geographical hotspots, allocate resources effectively, and tailor marketing strategies based on regional performance.

ii. Specific Requirements

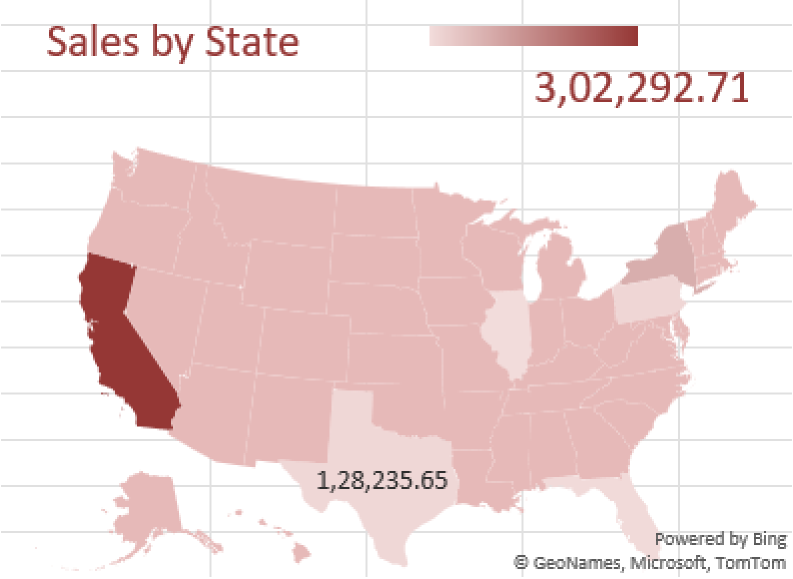
- Merge sales.csv with customers.csv on customer_id.
- Group by state or region and compute total sales revenue.
- Rank states by revenue and volume of orders.

iii. Analysis Results

- Punjab, Delhi, and Maharashtra were the highest revenue-generating states.
- Kerala and Odisha showed minimal sales, indicating potential areas for expansion.
- Urban states had higher average order values, while rural areas saw more frequent, smaller orders.

iv. Visualization

- Map Chart or Heatmap: Revenue by state.
- Column Chart: State-wise sales totals.
- Box Plot: Distribution of order values by state.



4.4. Sales vs Profit

i. General Description

High sales figures do not always translate to high profitability. This analysis identifies which products, categories, or regions offer the highest return on sales and helps detect areas with high revenue but low or negative profit margins.

ii. Specific Requirements

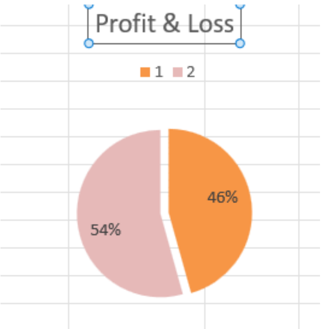
- Introduce a new derived column:
  Profit = Total Sale Value - Cost (assuming cost per unit is known or approximated).
- Compare total sales and profit across product categories, salespersons, and states.
- Identify products or regions where profit margins are low despite high sales.

iii. Analysis Results

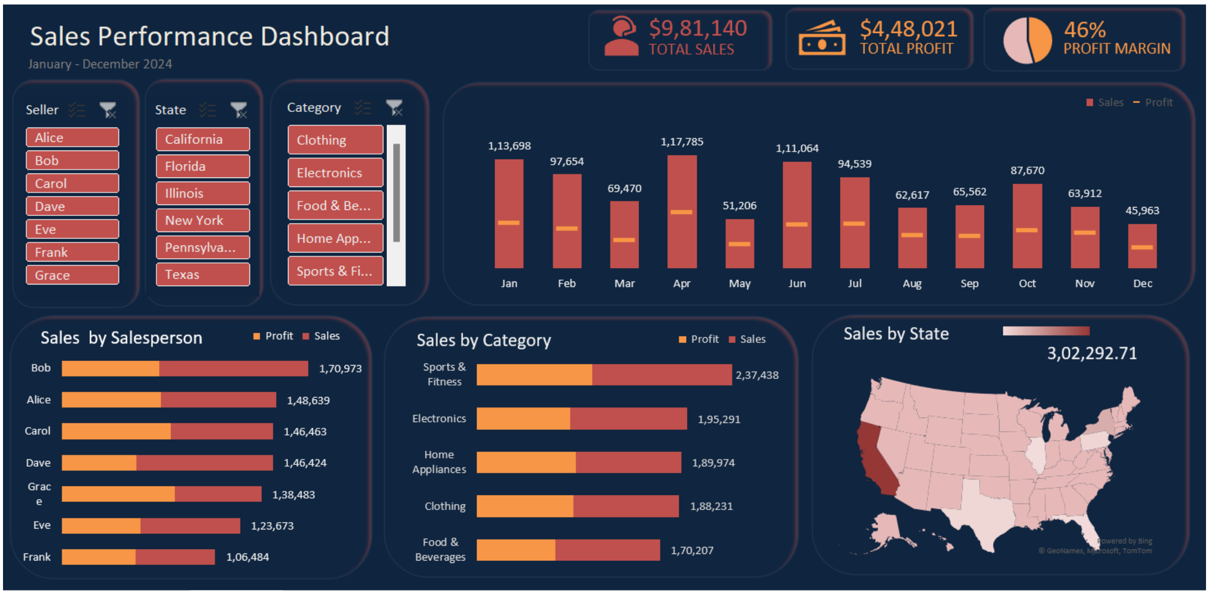
- Several products with high sales volume (e.g., in Electronics) had lower profit margins due to high acquisition or logistics cost.
- Home Decor items, although fewer in sales, showed a higher profit-to-sales ratio.
- Certain states like Delhi had strong revenue and profit alignment, while Rajasthan had high sales but minimal profits due to high shipping cost and returns.

iv. Visualization

- Scatter Plot: Sales vs Profit per product or category (to show correlation).
- Profit Margin Bar Chart: Profit as a percentage of sales by category.
- Dual Axis Chart: Sales and profit trends over time (month-wise).



Dashboard:



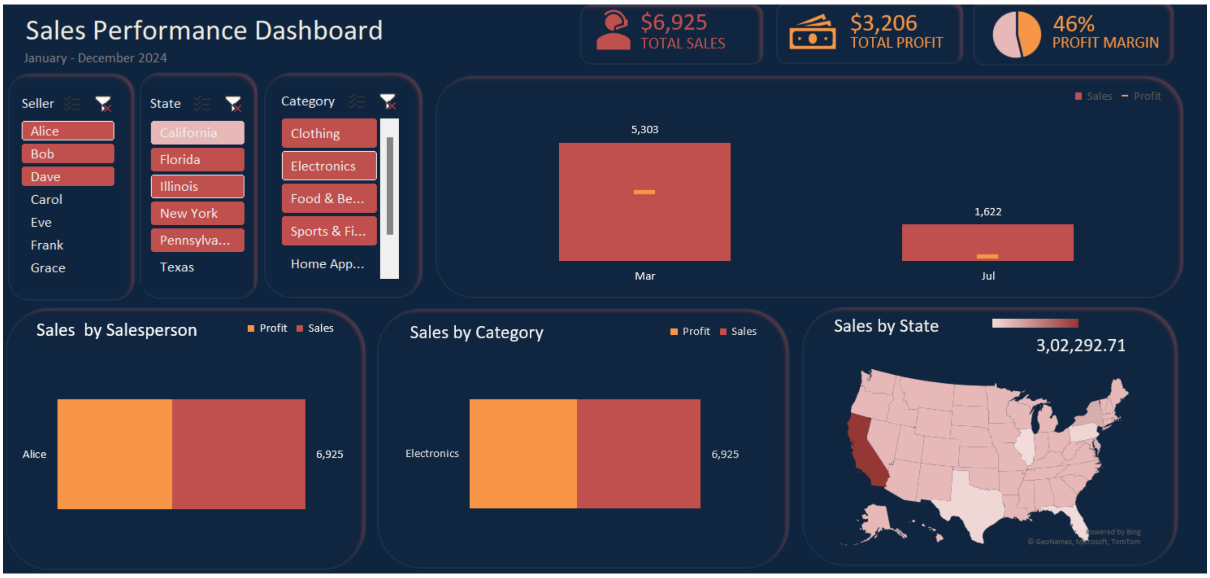
Sales Performance Dashboard

January - December 2024

$9,81,140 TOTAL SALES

$4,48,021 TOTAL PROFIT

46% PROFIT MARGIN



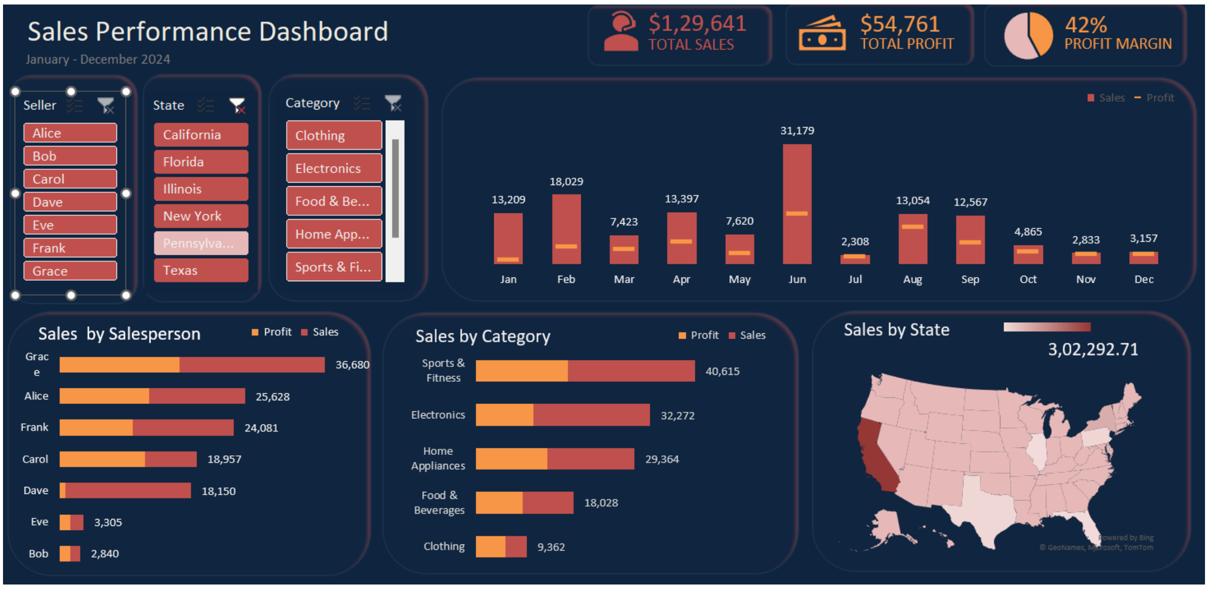
Sales Performance Dashboard

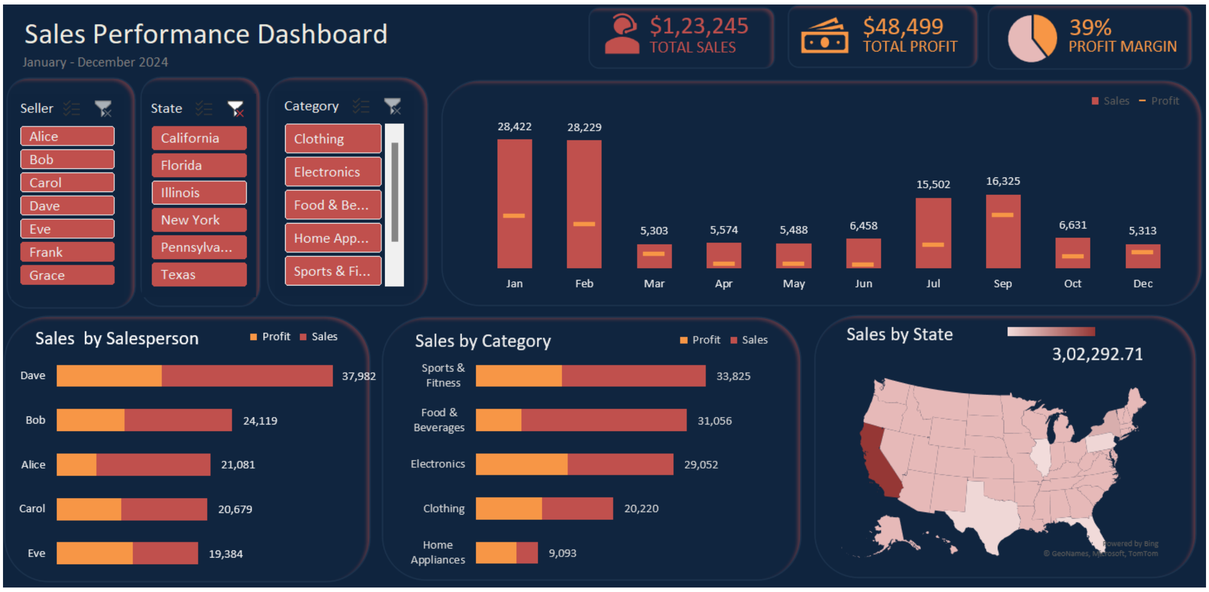
January - December 2024

$6,925 TOTAL SALES

$3,206 TOTAL PROFIT

46% PROFIT MARGIN


Sales Performance Dashboard
January - December 2024
$1,29,641
TOTAL SALES
$54,761
TOTAL PROFIT
42%
PROFIT MARGIN
Seller
Alice
Bob
Carol
Dave
Eve
Frank
Grace
State
California
Florida
Illinois
New York
Pennsylva...
Texas
Category
Clothing
Electronics
Food & Be...
Home App...
Sports & Fi...
Sales
Profit
13,209
18,029
7,423
13,397
7,620
31,179
2,308
13,054
12,567
4,865
2,833
3,157
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Sales by Salesperson
Profit
Sales
Grace
36,680
Alice
25,628
Frank
24,081
Carol
18,957
Dave
18,150
Eve
3,305
Bob
2,840
Sales by Category
Profit
Sales
Sports & Fitness
40,615
Electronics
32,272
Home Appliances
29,364
Food & Beverages
18,028
Clothing
9,362
Sales by State
3,02,292.71
Powered by Bing
© GeoNames, Microsoft, TomTom


Sales Performance Dashboard
January - December 2024
$1,23,245
TOTAL SALES
$48,499
TOTAL PROFIT
39%
PROFIT MARGIN
Seller
Alice
Bob
Carol
Dave
Eve
Frank
Grace
State
California
Florida
Illinois
New York
Pennsylva...
Texas
Category
Clothing
Electronics
Food & Be...
Home App...
Sports & Fi...
Sales
Profit
28,422
28,229
5,303
5,574
5,488
6,458
15,502
16,325
6,631
5,313
Jan
Feb
Mar
Apr
May
Jun
Jul
Sep
Oct
Dec
Sales by Salesperson
Profit
Sales
Dave
37,982
Bob
24,119
Alice
21,081
Carol
20,679
Eve
19,384
Sales by Category
Profit
Sales
Sports & Fitness
33,825
Food & Beverages
31,056
Electronics
29,052
Clothing
20,220
Home Appliances
9,093
Sales by State
3,02,292.71
Powered by Bing
© GeoNames, Microsoft, TomTom

5. Conclusion

This project offers a comprehensive analysis of retail sales data by examining performance across salespersons, product categories, and geographic regions. Through systematic data cleaning, integration, and visualization, we transformed raw datasets into meaningful insights. Sales by salesperson revealed that a small group drove the majority of revenue, highlighting potential for performance-based strategies. Category-wise analysis uncovered top-performing product segments and areas requiring attention, while state-wise trends provided valuable input for geographic expansion and localized marketing. Importantly, the sales vs. profit analysis demonstrated that high revenue does not always equate to high profitability—emphasizing the need to evaluate margins alongside sales figures. Visual tools like bar charts, scatter plots, and heatmaps made it easier to interpret complex data and spot hidden patterns. Overall, the project not only identifies key business drivers but also builds a strong analytical foundation for strategic decision-making, inventory planning, and future implementation of advanced predictive and prescriptive models.

6. Future Scope

This project lays a strong foundation for further analytical and strategic developments in retail sales intelligence. In future iterations, several enhancements can be incorporated to expand the depth and utility of the analysis:

- **Predictive Modeling:** Implement machine learning algorithms to forecast future sales, customer churn, and seasonal demand patterns.
- **Profit Optimization:** Incorporate cost structures, return rates, and logistics expenses to better understand net profitability and optimize pricing strategies.
- **Customer Segmentation:** Use clustering techniques to segment customers based on purchase behavior, enabling personalized marketing and retention campaigns.
- **Real-Time Dashboard Integration:** Develop interactive dashboards using tools like Power BI or Tableau for live sales tracking and decision-making.
- **Multi-Channel Data Integration:** Combine data from physical stores, online platforms, and customer feedback systems for a 360-degree business view.

- **Sentiment & Review Analysis:** Integrate customer reviews and social media data to correlate product sentiment with sales performance.

These advancements will enable businesses to make data-driven decisions more efficiently and proactively.

7.References:

Dataset Sources

• Customer Dataset (customers.csv)

Simulated or practice dataset containing customer details used for segmentation, sales trend analysis, and demographic-based insights. Commonly structured with fields like customer ID, name, location, and contact.

• Product Dataset (products.csv)

Includes product-related information such as product ID, name, category, and unit price. Used to map sales with product categories and perform profitability and stock analysis.

• Sales Dataset (sales.csv)

Contains transactional data such as date of sale, product ID, customer ID, quantity sold, and total amount. Central to time series analysis, revenue calculation, and performance metrics.

Tools and Techniques Used

• Microsoft Excel (Power Pivot & Dashboarding)

Used for building relationships between datasets, implementing calculated fields (DAX), and creating dynamic dashboards using slicers, PivotTables, and charts.

Learning Resources

• YouTube Tutorials & Excel Blogs

Referenced for techniques on creating interactive dashboards, using Power Query for data transformation, and enhancing visual appeal using chart formatting tips.

GitHub Link:

LinkedIn Link: