Sales Analytics & Prediction System

Project Report
Prepared by: Dhanwin Sangishetty.
Date: August 20, 2025

---

1. Project Overview

This project demonstrates a comprehensive Sales Analytics & Prediction System built to analyze sales data and predict profitability. The system leverages SQL for data querying, Pandas for data processing, Seaborn/Matplotlib for visualizations, scikit-learn for machine learning, and Streamlit for a user-friendly web interface. Key components include:

- **SQL-based Reporting**: Aggregated sales, profit, and order data by region, category, month, quarter, and year.
- **Data Visualizations**: Bar and line charts to highlight sales trends and patterns.
- **Machine Learning Model**: A Logistic Regression model (94% accuracy) to predict whether a sale will be profitable or result in a loss.
- **Web Application**: A Streamlit app allowing users to input sales data and receive real-time profit/loss predictions.

This project aligns with the Implementation Analyst role's requirements, showcasing SQL proficiency, data analysis, visualization, and client-facing reporting skills.

---

2. Dataset Description

The dataset used is the Superstore Sales Dataset, sourced from Kaggle. It contains 21 columns, including:

- **Order Date**: Date of the sale (MM/DD/YYYY format, e.g., 11/8/2016).
- **Sales**: Revenue from the sale (e.g., 261.96).
- **Profit**: Profit or loss from the sale (e.g., 41.9136).
- **Quantity**: Number of items sold.
- **Discount**: Discount applied (0 to 1).
- **Region, Category, Customer Name**: Contextual fields for analysis.

The dataset was loaded into a SQLite database (sales.db) for efficient querying and analysis.

---

3. SQL Queries for Reporting

The project includes SQL queries to generate actionable insights, saved as Excel files for client delivery. Below are the key queries:

- **Sales by Region**:

sql

```
SELECT Region, SUM(Sales) AS Total_Sales
FROM sales
GROUP BY Region;
```

Output: Excel file (sales_by_region.xlsx) showing total sales per region.

- **Top 10 Customers**:

sql

```
SELECT "Customer Name", SUM(Sales) AS Revenue
FROM sales
GROUP BY "Customer Name"
ORDER BY Revenue DESC
LIMIT 10;
```

Output: Excel file (top_customers.xlsx) listing the top 10 customers by revenue.

- **Average Profit by Category**:

sql

```
SELECT Category, AVG(Profit) AS Avg_Profit
FROM sales
GROUP BY Category;
```

Output: Excel file (avg_profit_category.xlsx) showing average profit per product category.

- **Monthly Sales**:

sql

```
SELECT strftime('%Y-%m', "Order Date") AS Month,
    SUM(Sales) AS Total_Sales,
    SUM(Profit) AS Total_Profit,
    COUNT(*) AS Orders
FROM sales
GROUP BY strftime('%Y-%m', "Order Date")
ORDER BY Month;
```

Output: Excel file (monthly_sales.xlsx) with monthly sales, profit, and order counts.

- **Quarterly Sales**:

sql

```
SELECT
    strftime('%Y', "Order Date") || '-Q' ||
    CAST(((CAST(strftime('%m', "Order Date") AS INTEGER) - 1) / 3 + 1) AS TEXT) AS
Quarter,
    SUM(Sales) AS Total_Sales,
    SUM(Profit) AS Total_Profit,
    COUNT(*) AS Orders
FROM sales
GROUP BY strftime('%Y', "Order Date"),
        ((CAST(strftime('%m', "Order Date") AS INTEGER) - 1) / 3 + 1)
ORDER BY strftime('%Y', "Order Date"),
        ((CAST(strftime('%m', "Order Date") AS INTEGER) - 1) / 3 + 1);
```

Output: Excel file (quarterly_sales.xlsx) with quarterly sales, profit, and orders, formatted as YYYY-QN (e.g., 2016-Q1).

- **Yearly Sales**:

sql

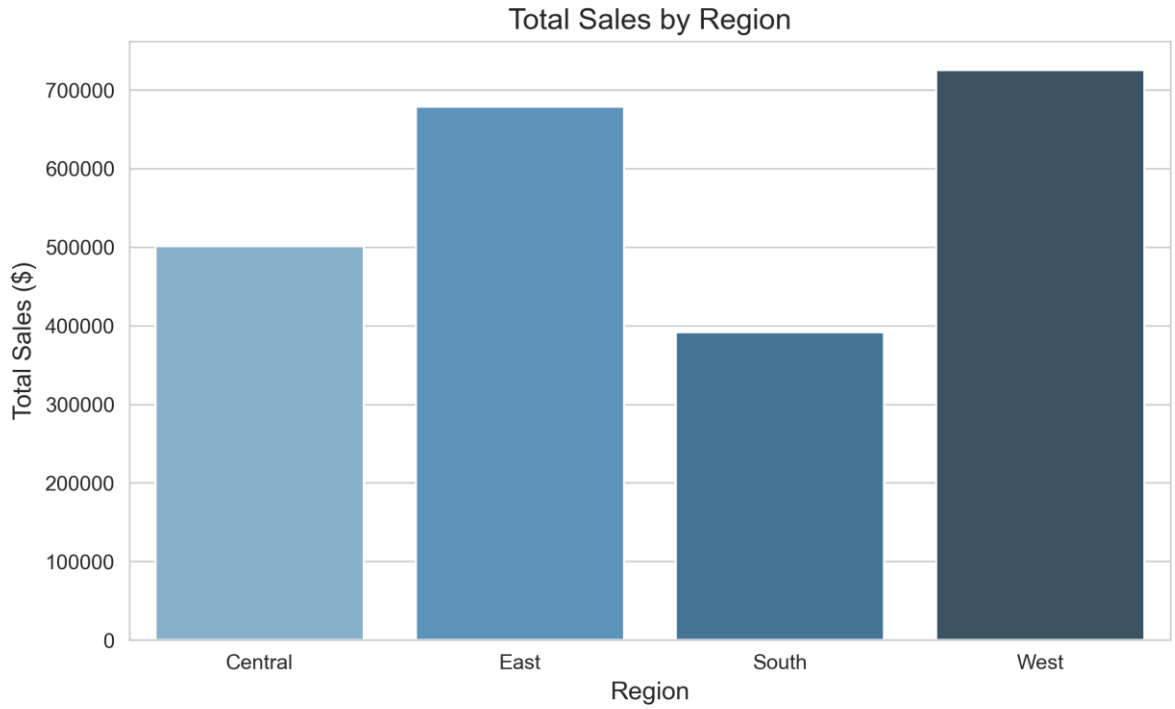
```
SELECT strftime('%Y', "Order Date") AS Year,
      SUM(Sales) AS Total_Sales,
      SUM(Profit) AS Total_Profit,
      COUNT(*) AS Orders
FROM sales
GROUP BY strftime('%Y', "Order Date")
ORDER BY Year;
```

Output: Excel file (yearly_sales.xlsx) with annual sales, profit, and orders.

4. Visualizations

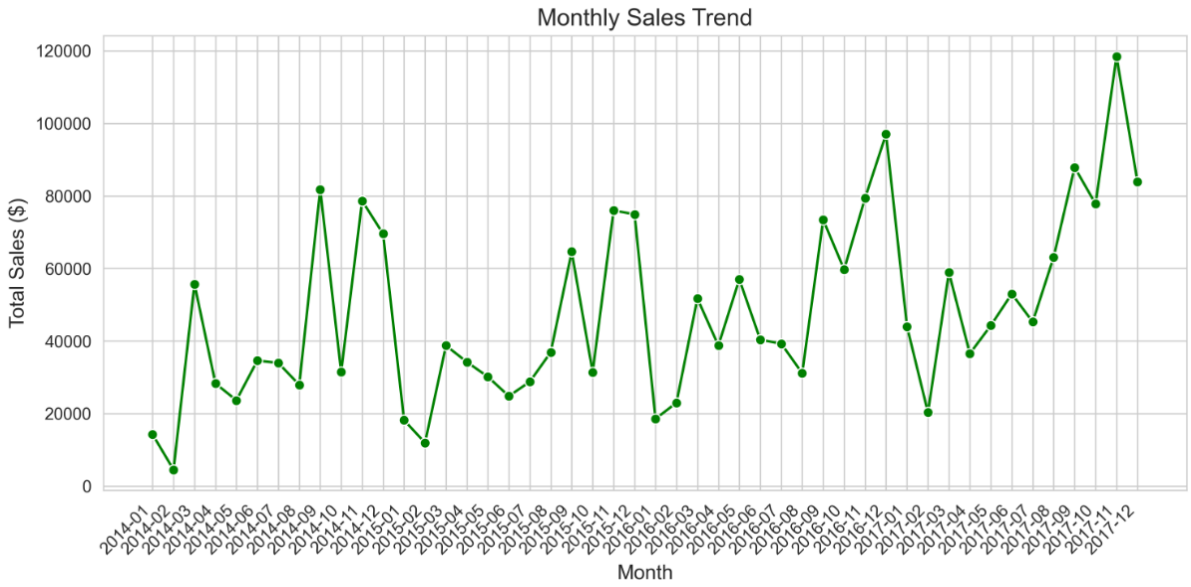
The project includes four visualizations generated using Python (Seaborn and Matplotlib) to provide intuitive insights into sales performance:

- **Sales by Region**: A bar chart showing total sales across regions, highlighting top-performing regions.



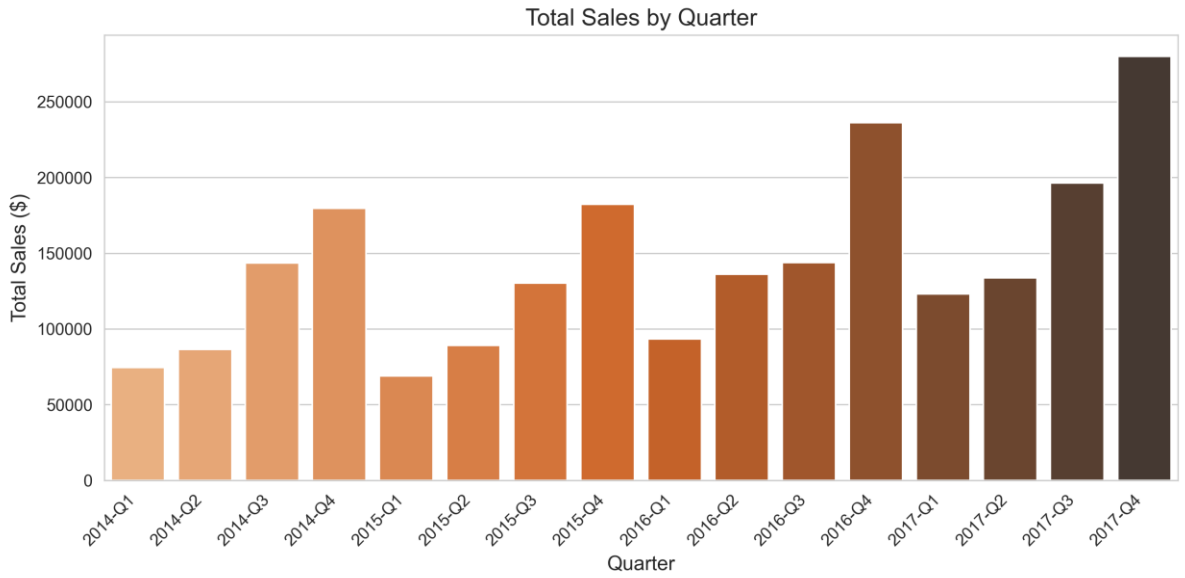
Description: Blue bars represent total sales per region, with clear labels for easy interpretation.

- **Monthly Sales Trend**: A line chart displaying total sales over time, with data points for each month



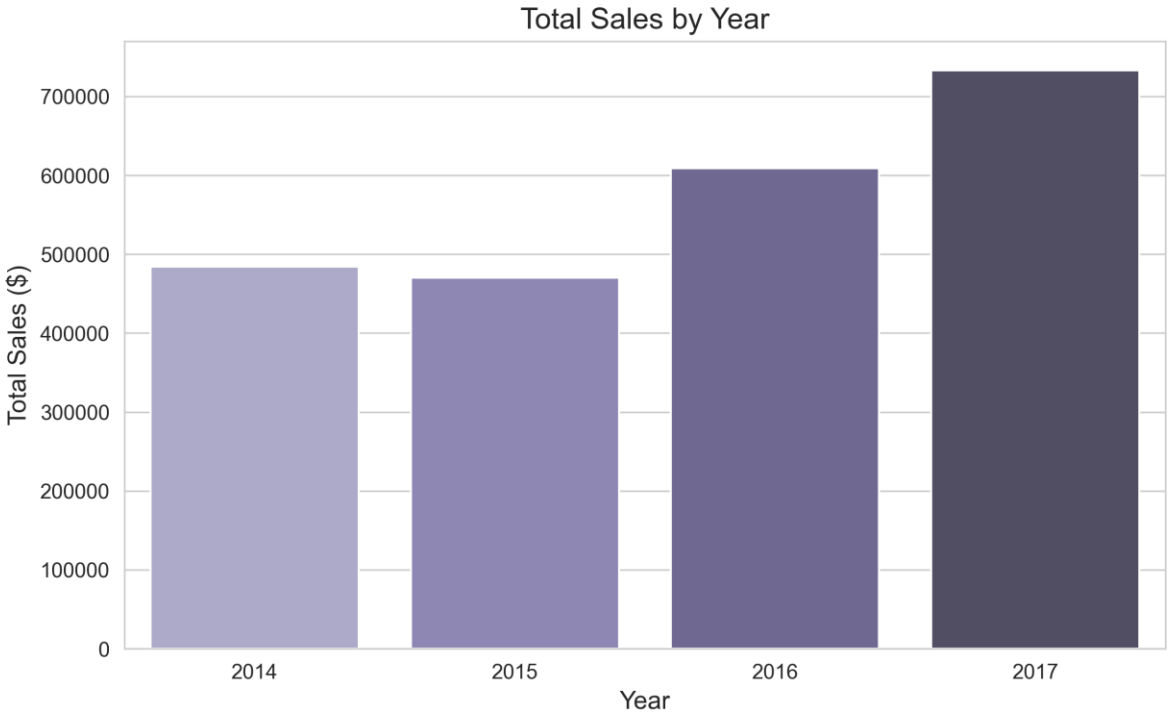
. *Description*: Green line with markers shows sales trends, with rotated x-axis labels for readability.

- **Quarterly Sales**: A bar chart illustrating total sales by quarter (e.g., 2016-Q1, 2016-Q2).



Description: Orange bars show quarterly sales, formatted for seasonal analysis.

- **Yearly Sales**: A bar chart summarizing total sales by year (e.g., 2016, 2017).



Description: Purple bars highlight annual performance trends.

5. Machine Learning Model

A Logistic Regression model was developed to predict whether a sale will be profitable (Profit > 0) or result in a loss (Profit ≤ 0). Key details:

- **Features**: Sales, Quantity, Discount.
- **Target**: Binary classification (1 = Profit, 0 = Loss).
- **Training**: Used scikit-learn with an 80/20 train-test split.
- **Performance**: Achieved 94% accuracy on the test set.
- **Output**: Predicts profit/loss for new inputs (e.g., Sales=$200, Quantity=3, Discount=0.1 → Profit).

---

6. Streamlit Web Application

A user-friendly web application was built using Streamlit to allow real-time profit/loss predictions:

- **Functionality**: Users input Sales, Quantity, and Discount via a browser interface and receive an instant prediction (Profit or Loss).
- **Implementation**: The app loads the pre-trained Logistic Regression model and displays predictions with a clean UI.
- **Sample Output**: For inputs Sales=$200, Quantity=3, Discount=0.1, the app predicts "Profit ✅" with 94% model accuracy.

Description: The interface shows input fields, a "Predict" button, and a clear Profit/Loss output.

---

7. Conclusion

This project demonstrates end-to-end skills in data analysis, visualization, and predictive modeling, tailored to the Implementation Analyst role. Key achievements include:

- **SQL Proficiency**: Developed queries for region, customer, category, monthly, quarterly, and yearly analyses.
- **Reporting**: Generated client-ready Excel reports.
- **Visualization**: Created professional bar and line charts for actionable insights.
- **AI/ML**: Built a 94% accurate Logistic Regression model for profit prediction.
- **User Interface**: Deployed a Streamlit app for interactive predictions.

The source code is available at

[GitHub Repository: https://github.com/DhanwinSangishetty/sales-analytics-prediction-system].