Sales Data ETL, Analysis & Profit Prediction – Project Report

---

1. Overview

This project focuses on understanding sales performance, analysing profit trends, and predicting profit using machine learning models.
It includes all major steps — **ETL (Extract, Transform, Load)**, **SQL Analysis**, **Visualisation**, and **Prediction**.

---

2. Tools & Technologies

- **Languages:** Python, SQL
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, SQLAlchemy
- **Database:** MySQL
- **IDE:** Jupyter Notebook
- **Visualisation:** Matplotlib, Seaborn

---

3. Step 1: Data Engineering (ETL)

- Cleaned the raw dataset and removed ASCII characters and null values.
- Converted `Order Date` to proper datetime format.
- Created new columns like `Month`, `Year`, and `Discounted_Sales`.
- Loaded the cleaned data into **MySQL** using SQLAlchemy.

✅ **Output:** Clean and structured data ready for analysis.

---

4. Step 2: SQL Data Analysis

Performed analysis using SQL queries to get key business insights:

- Average Profit by Category
- Sales by Region and Sub-Category
- Top 10 Selling Products
- Monthly and Yearly Sales Trends
- Relation between Discount and Profit

✅ **Output:** Understood sales patterns and profit behaviour across different categories.

---

5. Step 3: Visualisation & Insights

- Created bar charts, line graphs, and heatmaps using Python.
- Visualised monthly sales, profit vs discount, and regional performance.
- Found that discounts impact profit negatively and some categories perform better in specific regions.

✅ **Output:** Clear understanding of trends and relationships in data.

---

6. Step 4: Predictive Modelling

Built machine learning models to **predict Profit** based on sales and discount data.
Models used:

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor

Compared models using **MSE, MAE, and R^2 Score**.
✅ **Best Model: Random Forest Regressor** (lowest error and better accuracy).

---

7. Step 5: Final Insights & Recommendations

- Focus on high-profit product categories.
- Avoid giving high discounts on low-margin items.
- Concentrate marketing in top-performing regions.
- Use the Random Forest model for future profit predictions.

---

8. Results Summary

Step	Task	Output
1	Data Cleaning & Transformation	Clean, formatted dataset
2	SQL Analysis	Key insights on sales and profit
3	Visualization	Graphs and trend analysis
4	ML Modeling	Profit prediction with Random Forest
5	Final Insights	Actionable business recommendations

---

9. Conclusion

This project successfully demonstrates a complete **data analytics workflow**, from raw data to predictive insights.
It shows how data cleaning, SQL analysis, and machine learning can work together to help businesses make better decisions.