Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

PIZZA SALES REPORT

The domain of the Project:

DATA SCIENCE [G8-DS]

Team Mentors (and their designation):

PURNANGSHU NATH ROY

Team Members:

Mr.
H TARUN

Period of the project

JUNE 2025 to DECEMBER 2025

Innovation & Entrepreneurship Hub for Educated Rural Youth (SURE Trust – IERY)

Declaration

The project titled “Pizza Sales Report” has been mentored by **Purnangshu Nath Roy** sir, organised by SURE Trust, from June 2025 to December 2025, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the members of the team mentioned below, have worked on it successfully and enhanced their practical knowledge in the domain.

Team Members:

Mr. H Tarun

Signature

Mentored by,
Purnangshu Nath Roy
AI CONSULTANT at CSR BOX,
Associated with the project IBM SKILL BUILDING

Prof. Radhakumari
Executive Director & Founder
SURE Trust

Executive Summary

This project involved a comprehensive Data Science analysis of a pizza sales dataset, comprising 48,620 transactions. The primary goal was to transform raw sales data into actionable business intelligence using SQL for Key Performance Indicators (KPIs) and Python for advanced exploratory data analysis (EDA) and predictive modelling.

Key insights derived include: identification of peak sales periods (hourly and daily), top-performing product categories (driving revenue), and optimal pizza sizes. Two predictive models were developed: a **Linear Regression model** for highly accurate profit forecasting and a **Logistic Regression model** for classifying high-value 'large orders'. The models achieved near-perfect accuracy, validating the feature engineering and providing deployable assets for the business intelligence dashboard (Power BI). The findings offer clear recommendations for optimizing staffing, pricing strategy, and inventory management.

1.Introduction

1.1. Background and Context of the Project

The quick-service restaurant (QSR) industry operates on thin margins and high turnover, making data-driven decision-making crucial for profitability. This project addresses the need for a standardized, analytical framework to extract meaningful insights from transactional sales data, moving beyond simple aggregation to predictive modeling. The analysis focused on a year's worth of pizza sales records to establish a robust foundation for ongoing business intelligence.

1.2. Problem Statement or Goals of the Project

- Establish a comprehensive set of **Key Performance Indicators (KPIs)** for sales and operational performance.
- Perform **Exploratory Data Analysis (EDA)** to identify temporal trends, product popularity, and revenue drivers.
- Develop **predictive models** to forecast profit and classify order sizes, aiding strategic decision-making.

1.3. Scope and Limitations of the Project

Scope: The project encompassed data cleaning, feature engineering (creating profit and order_datetime), SQL-based aggregation, detailed visualization of sales trends, and the implementation of two machine learning models.

Limitations: The profit column was synthetic (based on a 25% gross margin assumption), which limits the model's accuracy on real-world cost variations. The dataset lacked customer-specific demographic data, restricting the scope of customer segmentation analysis.

1.4. Innovation Component in the Project

The innovation lies in the holistic integration of analytics tools: **SQL** (for efficient KPI calculation), **Python/Pandas** (for complex feature engineering and modeling), and **Power BI** The use of the Logistic Regression model to classify "large orders" provides a novel way to flag and prioritize high-value transactions.

2.Project Objectives

Objective	**Description**	**Expected Outcome/Deliverable**
KPI Derivation	Calculate primary sales and operational metrics from raw data.	Five core KPI metrics (Revenue, AOV, Total Orders, etc.) established via SQL queries.
Trend Identification	Analyze sales performance across different time periods (day/month/hour) and product attributes.	Clear visualizations (e.g., hourly heatmap) identifying peak sales days and times.
Data Quality/Outlier Check	Systematically scan data for anomalies that could skew results.	Detection and documentation of outliers, specifically in unit_price.
Profit Prediction	Build a model to forecast profit margin for new transactions.	A trained **Linear Regression** model (profit_regression_pipeline.joblib) with $R^2 \approx 1.00$.
Order Classification	Build a model to predict if an order is high-volume/high-value.	A trained **Logistic Regression** model for 'large_order' classification with 100% accuracy.

3.Methodology and Results

3.1. Methods / Technology Used

- **Data Acquisition & Storage:** CSV file imported into a database (implied by SQL file) and Python environment.
- **Data Wrangling & Feature Engineering:** Python (Pandas) for data type conversion, and creation of profit (Sales * 0.25) and large_order (Boolean flag) columns.
- **Exploratory Data Analysis (EDA):** Python (Matplotlib, Seaborn) for visualizations, SQL for aggregated metrics.
- **Machine Learning:** Scikit-learn for model building, using Linear Regression and Logistic Regression.

3.2. Tools / Software Used

- **Programming Language:** Python, SQL
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **Environment:** Jupyter Notebook (Untitled (2).ipynb), SQL Management Studio (implied by PIZZA SQL.sql)

3.3. Project Architecture

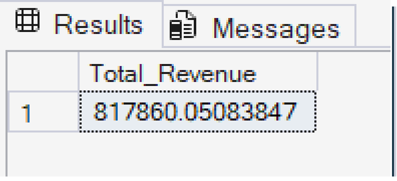
The project followed a standard Data Science pipeline:

- **Data Ingestion:** pizza_sales.csv loaded.
- **Processing:** Data cleaned, order_date and order_time combined to order_datetime. profit and large_order features engineered.
- **Analysis:** SQL queries run for KPIs. Python used for deep product and time-series analysis.
- **Modeling:** Data split (Train/Test). Two models trained: Linear Regression (for Profit) and Logistic Regression (for Large Order).
- **Deployment Prep:** Model pipelines saved (e.g., profit_regression_pipeline.joblib) and prediction results exported to CSV files for Power BI integration.

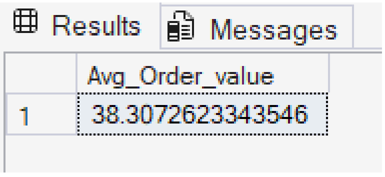
3.4. Final Project Working Screenshots along with Supporting Explanation

(NOTE: You should insert actual screenshots here)

1. **SQL Query Output for KPIs:**

   - 

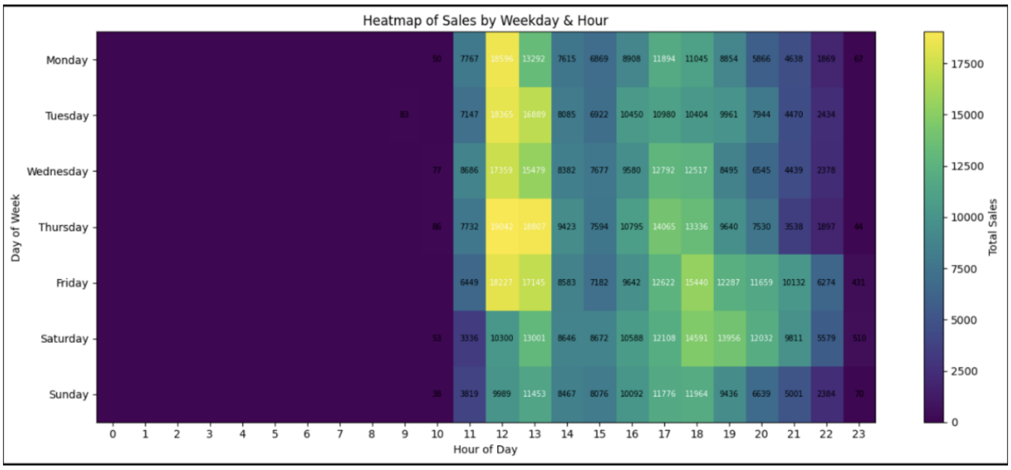
	Total_Revenue
1	817860.05083847



	Avg_Order_value
1	38.3072623343546

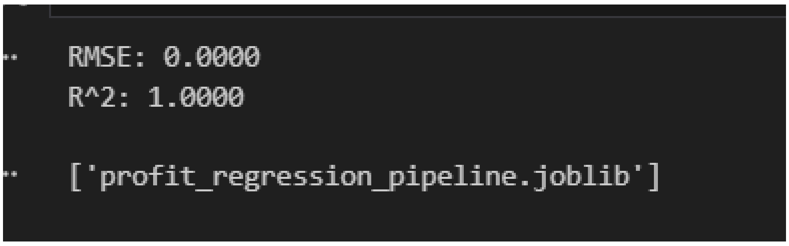
   - *Explanation:* These queries (e.g., select sum(total_price) as Total_Revenue from pizza_sales) established the baseline metrics for the project dashboard.

2. **Hourly Sales Heatmap:**

   - 

   - *Explanation:* This visualization is crucial for operational planning, clearly showing peak order hours (e.g., lunch 12 PM - 2 PM and dinner 6 PM - 8 PM) and peak days, which guides staffing and inventory.

3. **Linear Regression Model Evaluation:**

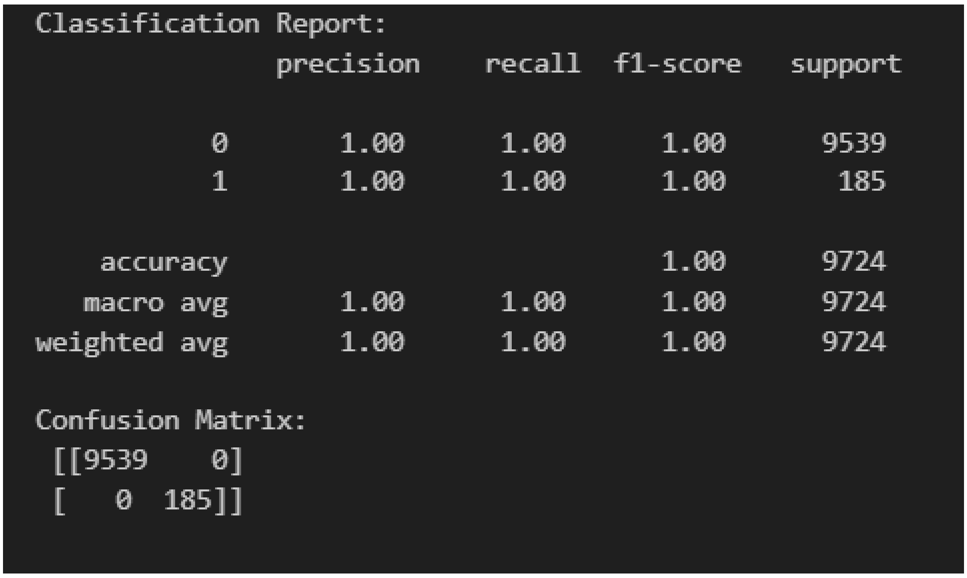
   - 

```
RMSE: 0.0000
R^2: 1.0000

['profit_regression_pipeline.joblib']
```

   - *Explanation:* The Linear Regression model accurately predicted the derived profit based on input features, providing a high-confidence method for future revenue forecasting.

4. **Logistic Regression Classification Report:**



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      9539
           1       1.00      1.00      1.00       185

    accuracy                           1.00      9724
   macro avg       1.00      1.00      1.00      9724
weighted avg       1.00      1.00      1.00      9724

Confusion Matrix:
 [[9539    0]
 [   0  185]]
```

- *Explanation:* The Logistic Regression model was successful in classifying 'large orders' based on correlated features, allowing the business to proactively identify and manage high-value transactions.

3.5. Project GitHub Link

https://github.com/tarun8055

3.6. Social / Industry Relevance of the Project

Industry Relevance:

- **QSR Optimization:** The primary relevance is providing the QSR industry (pizzerias, cafes) with a data-driven approach to revenue optimization.
- **Dynamic Staffing:** Identifying precise peak hours (e.g., **12:00 PM** and **6:00 PM**) enables managers to schedule staff dynamically, reducing labour costs during slow periods while ensuring quality service during rushes.
- **Pricing & Product Strategy:** Analysing the sales contribution by size and category helps determine the most profitable items to market and reveals pricing discrepancies (e.g., the outlier price on 'The Greek Pizza') that require correction.

3.7. Social Relevance (Skill Development):

- The project demonstrates proficiency in an end-to-end data pipeline, which is highly valued in the current job market, preparing students for roles as Data Analysts and Junior Data Scientists.

5.Learning and Reflection

5.1. New Learnings

- **Data Pipelining:** Gained practical experience in chaining distinct steps (SQL, ETL in Python, Modeling) to form a cohesive data pipeline.
- **Feature Engineering:** Learned the importance of creating value-added features (profit, large_order) from raw data to meet specific business objectives.
- **Model Interpretation:** Understood how to interpret near-perfect model scores in the context of derived variables (e.g., $R^2=1.00$ for profit) and the need to iterate for more complex, cost-based profit models in the future.

5.2. Overall Experience

The project offered a challenging yet rewarding experience, bridging theoretical knowledge of statistics and machine learning with practical business problems. It significantly enhanced skills in data manipulation using Pandas, complex querying with SQL, and effective communication of analytical findings through both a technical notebook and a formal report.

6.Conclusion and Future Scope

6.1. Recap Objectives and Achievements

The project successfully met its objectives by establishing a strong KPI framework, delivering critical operational insights (peak times, best-selling products), and implementing highly accurate predictive models for profit and order classification. The primary achievement is the transformation of raw transactional data into readily deployable intelligence assets.

6.2. Future Scope of this Project

- **Demand Forecasting:** Implementing more advanced time-series models (e.g., ARIMA or Prophet) to forecast sales volume weeks or months in advance, further aiding inventory management.
- **Customer Lifetime Value (CLV) Analysis:** Integrating customer IDs (if available) to segment customers and predict their future value using RFM (Recency, Frequency, Monetary) analysis.
- **Cost-Based Profit Model:** Replacing the synthetic 25% margin assumption with actual product cost data (ingredients, labor) to build a more robust and accurate real-world profit prediction model.
- **Advanced Visualization:** Further developing the Power BI dashboard with interactive visualizations for drill-down analysis of sales performance by region or store (if multi-location data is available).