

StrategicThinking

CA 3

Pizza Sales Analysis Report

HDip in Data Analytics for Business at CCT College

Daniela Maftei

Student Number: sba24270

Date: 20/05/2025

Word Count ( without Table of Contents and References): 5310

[**Report on Pizza Sales Analysis 5**](#_p6h7lkx57qj1)

[1. Strategic Overview of the Business Problem 5](#_67xoot994pb7)

[1.2 Background 5](#_2m4s3ldx9cgz)

[1.2 Business Problem 5](#_k73xoeb2e7m7)

[1.3 Strategic Importance 6](#_ogimq6bf4vaj)

[1.4 Project Objectives 7](#_hoqxlsg92qf8)

[2. Project Plan 9](#_1oiokg6ccb8k)

[2.1 Project Management Methodology 9](#_ljtlxhsa8n1t)

[2.2 Project Timeline 9](#_729ai2eipu94)

[2.3 Milestones Planned 10](#_p15m0mqk3tnz)

[2.4 Anticipated Challenges 11](#_n2b0aipehe06)

[2.5 Roles and Responsibilities 11](#_oyr3wg86jj2m)

[3. Business Understanding 12](#_3s6agcpicaat)

[3.1 Domain Knowledge 12](#_ffjjtdnh1wjp)

[3.2 Stakeholder Identification 13](#_bndkbxwfrj8i)

[3.3 Initial Questions 13](#_h50wt0km7mme)

[4. Data Understanding 14](#_bm9qf6dwheli)

[4.1 Data Source and Description 14](#_ofen652qeoq2)

[4.2 Data Exploration 15](#_gktnt7yhi67c)

[4.3 Initial Data Visualisations 16](#_qu8r6t9elwdd)

[4.4 Initial Observations and Potential Issues 19](#_us2g8d7j3mrh)

[5. Data Preparation 19](#_ofbvgazhbmuw)

[5.1 Data Cleaning 19](#_wbv5d8r3ti1b)

[5.2 Data Transformation 20](#_repy7gimkbb6)

[5.3 Feature Selection (Initial) 21](#_72ud7o546ps0)

[5.4 Data Splitting 21](#_trcw8pw4kaym)

[6. Findings & Recommendations 22](#_8g3x97fyyq3y)

[6.1 Exploratory Data Analysis Insights 22](#_hq52qc5fjndc)

[6.2 Machine Learning Implementation for Prediction 23](#_51y6qojmny6p)

[6.2.1 Problem Formulation 23](#_db1zgc7mptwu)

[6.2.2 Model Selection 23](#_tgsqwq9as11w)

[6.2.3 Model Training and Evaluation 24](#_5gbdn0xr0sh2)

[6.2.4 Hyperparameter Tuning 25](#_3pcu0bni9c8d)

[6.2.5 Model Comparison 25](#_coa6m3gsetcw)

[6.2.6 Model Validation 26](#_o4ybphfdehz5)

[6.2.7 Business Recommendations 26](#_3op3t4dueaho)

[7. Findings of Machine Learning Implementation, Results Validation, Model Comparison, and Future Recommendations 27](#_k859cwvsv5kb)

[7.1 Machine Learning Findings 27](#_h3cymdjuxnoj)

[7.2 Results Validation 29](#_nhw26vtgucaq)

[7.3 Model Comparison 29](#_fubhxjvgxqsc)

[**References 31**](#_l0r73olmokzm)

# **Report on Pizza Sales Analysis**

## **1. Strategic Overview of the Business Problem**

## **1.2 Background**

The pizza industry is a significant part of the food service sector, characterized by diverse business models, ranging from small local pizzerias to large multinational chains. In this dynamic and competitive landscape, businesses are constantly seeking ways to optimize their operations, enhance customer satisfaction, and drive revenue growth. Traditionally, pizza businesses have relied on experience and intuition to make decisions regarding inventory management, marketing campaigns, and staffing. However, the increasing availability of data and advancements in data science offer unprecedented opportunities for these businesses to adopt a more data-driven approach.

Data-driven decision-making empowers pizza businesses to move beyond reactive strategies and embrace proactive approaches. By leveraging data, businesses can gain a deeper understanding of customer behavior, identify trends, and predict future outcomes. This shift enables them to make informed choices, optimize resource allocation, and personalize customer experiences, ultimately leading to improved efficiency and profitability.

### **1.2 Business Problem**

This capstone project aims to address the challenge of optimizing business operations and sales strategies for a pizza company by leveraging data analysis and machine learning. Specifically, the project focuses on the following key areas:

**Understanding Sales Patterns** by identifying trends in pizza sales, including daily, weekly, and hourly patterns, to optimize staffing and inventory.

**Analyzing Product Performance** by determining the most popular pizza categories and sizes to inform menu planning and marketing efforts.

**Predicting Sales** by developing a predictive model to forecast future sales, enabling better resource allocation and inventory management.

**Feature Importance** bydetermining which factors most influence sales.

The business problem can be summarized as the need to transform raw sales data into actionable insights that can drive strategic decision-making and improve overall business performance.

### **1.3 Strategic Importance**

Thinking strategically, I'm considering why solving this particular problem is so important for a pizza business.

**Improved Efficiency**

By accurately forecasting demand and understanding sales patterns, the pizza business can optimize its inventory management, reducing waste and storage costs. Staffing levels can also be adjusted to match peak demand periods, improving labor efficiency.

**Enhanced Marketing Effectiveness**

Identifying the most popular pizza categories and sizes allows the business to target its marketing efforts more effectively, increasing the return on investment (ROI) of marketing campaigns.

**Increased Customer Satisfaction**

By understanding customer preferences and ordering patterns, the business can personalize its offerings and promotions, leading to higher customer satisfaction and loyalty.

**Competitive Advantage**

In a competitive market, businesses that leverage data to make informed decisions gain a significant advantage. This project provides the pizza company with the tools and insights necessary to outperform competitors.

**Revenue Growth and Profitability**

Ultimately, the insights gained from this project will contribute to increased revenue and profitability by optimizing operations, reducing costs, and enhancing marketing effectiveness.

Solving these business problems through the application of data analytics is crucial for several strategic reasons. It allows this pizza business to gain a significant competitive advantage by developing a deeper understanding of customer needs and optimizing its operations in ways that competitors might not.

Increased profitability can be achieved through more efficient resource allocation, reduction of waste, and the implementation of targeted marketing strategies that resonate with customers.

By understanding and effectively catering to customer preferences, businesses can significantly improve overall customer satisfaction, fostering loyalty and positive word-of-mouth. Furthermore, leveraging data insights can lead to enhanced operational efficiency through the streamlining of processes and the effective use of technology. In today's competitive pizza industry, data-driven strategies are not merely beneficial; they are essential for sustained success and long-term growth.

In a market characterized by potentially thin profit margins and increasingly demanding customers, even incremental improvements in operational efficiency and customer satisfaction can yield substantial positive impacts on a business's profitability and its share of the market.

### **1.4 Project Objectives**

The aim is to provide a focused approach to this project, specific, measurable, achievable, relevant, and time-bound (SMART). Examples of such objectives include increasing the accuracy of sales forecasts for the next quarter by a defined percentage compared to previous forecasting methods.

Another objective could be to identify the top and bottom performing pizza combinations within a specified timeframe.

Developing a predictive model for customer order amounts with a quantifiable level of accuracy on a test dataset also constitutes a SMART objective.

Analyzing the impact of specific promotional activities on sales within a set period and formulating optimal promotion strategies based on the findings is another example.

Finally, aiming to reduce ingredient waste by a certain percentage within a defined timeframe through data-informed inventory optimization represents a clear and measurable goal.

The establishment of these well-defined SMART objectives provides a clear roadmap for the project, allowing for the measurement of success and ensuring that the efforts are directed towards achieving tangible business outcomes. SMART objectives ensure that the project remains focused, accountable, and ultimately delivers measurable value.

The primary objectives of this capstone project are:

**SMART Objective 1**

To load and preprocess the pizza sales dataset, ensuring data quality and consistency.

Loading the pizza\_sales.csv dataset into a Pandas DataFrame, handling missing values using appropriate imputation techniques (mean for numerical, most frequent for categorical), and converting date/time columns to the correct data types within one week.

**SMART Objective 2**

To conduct exploratory data analysis (EDA) to uncover key trends and patterns in the data.

Generating at least four visualizations (line chart for daily sales, bar chart for category sales, pie chart for size distribution, line chart for hourly sales) and documenting three key insights from each visualization within two weeks.

**SMART Objective 3**

To develop a machine learning model to predict total pizza sales.

Training and evaluating three regression models (Linear Regression, Random Forest Regressor, Gradient Boosting Regressor) using appropriate metrics (MSE, R-squared) on a training and testing split of the data, and selecting the best-performing model within three weeks.

**SMART Objective 4**

To provide actionable business recommendations based on the findings of the data analysis and machine learning modeling.

Generating a report with at least five actionable business recommendations, supported by specific findings from the EDA and machine learning results, within one week.

**SMART Objective 5**

To document the project methodology, findings, and recommendations in a comprehensive report.

Compiling all project code, data visualizations, and analysis results into a well-structured report of approximately 5000 words, adhering to the specified outline, within one week.

## **2. Project Plan**

## **2.1 Project Management Methodology**

This project will be managed using the Agile (Scrum) methodology. Scrum is an iterative and incremental framework for managing complex projects. It emphasizes collaboration, flexibility, and continuous improvement. Scrum facilitates the division of the project into smaller, more manageable units called sprints, each with a defined duration and specific goals. This structure enables regular feedback loops, allowing for continuous improvement and alignment with the evolving understanding of the data and the business needs. The rationale for choosing Scrum includes adaptability, collaboration, transparency and Rapid delivery. The methodology places a strong emphasis on collaboration among team members, transparent communication, and the ability to adapt efficiently to changes, all of which are critical for navigating the inherent uncertainties of data analysis and machine learning endeavors. The adoption of Agile Scrum ensures that value is delivered to the pizza business in an early and continuous manner, allowing for the incremental benefit from the insights gained throughout the project. The exploratory nature of data science often necessitates adjustments to the initial plan based on findings, and Scrum's sprint-based structure provides the framework for this iterative refinement.

### **2.2 Project Timeline**

The project timeline will be structured around a series of sprints, with each sprint typically lasting between two to four weeks, depending on the complexity of the tasks involved. The initial sprints will likely focus on foundational activities such as data understanding and preparation, including data cleaning, exploration, and feature engineering. Subsequent sprints will be dedicated to exploratory data analysis, where trends, patterns, and relationships within the data will be identified and visualized. A significant portion of the project will involve machine learning model building, where different models will be selected, trained, and evaluated for their predictive capabilities. The final sprints will concentrate on the interpretation of the results, the formulation of actionable business recommendations, and the compilation of the final report. A detailed sprint plan, outlining the specific tasks, responsibilities, and deadlines for each sprint, will be developed at the project's outset to provide a clear roadmap for the team. This structured timeline, with its division into focused sprints, ensures that progress is systematically tracked and that the project remains on schedule, facilitating timely delivery of the project's objectives. Breaking down the project into these smaller, time-boxed iterations allows for more effective planning and monitoring of progress, ensuring that the project stays aligned with its goals.

**Sprint 1: Data Acquisition and Understanding (Weeks 1-2)**

Data loading, initial exploration, and business understanding. Duration: 2 weeks

**Sprint 2: Data Preparation and EDA (Weeks 3-4)**

Data cleaning, transformation, and exploratory data analysis. Duration: 2 weeks

**Sprint 3: Machine Learning Modeling (Weeks 5-6)**

Focus: Model selection, training, evaluation, and hyperparameter tuning. Duration: 2 weeks

**Sprint 4: Report and Recommendations (Weeks 7-8)**

Report writing, presentation preparation, and final project delivery. Duration: 2 weeks

### **2.3 Milestones Planned**

Several key milestones are planned throughout the project to provide tangible checkpoints for assessing progress and ensuring that the project's objectives are being met. The key milestones for each sprint are as follows:

**Sprint 1:**

Milestone 1: Successful loading of the pizza sales dataset and initial data exploration.

Milestone 2: Completion of the Business Understanding section of the report.

**Sprint 2:**

Milestone 3: Completion of data cleaning and transformation.

Milestone 4: Generation of key data visualizations and completion of the Data Understanding and Preparation sections of the report.

**Sprint 3:**

Milestone 5: Training and evaluation of three machine learning models.

Milestone 6: Selection of the best-performing model and completion of the Machine Learning Implementation section of the report.

**Sprint 4:**

Milestone 7: Completion of the draft report, including findings and recommendations.

Milestone 8: Finalization of the report and project presentation.

These clearly defined milestones allow both the project team and the stakeholders to track achievements, identify any potential delays or roadblocks early in the process, and ensure that the project remains on track to deliver its intended value.

### **2.4 Anticipated Challenges**

Data quality Issues might arise as missing values, inconsistencies, or outliers might require significant cleaning efforts.

Identifying and creating meaningful features for machine learning might be challenging, depending on the Feature Engineering complexity.

Choosing the right models and effectively tuning their hyperparameters can be time-consuming.

Interpretation of results whentranslating model outputs into actionable business insights requires careful consideration, as well as hitting time constraints when completing all tasks within the allocated time frame.

### **2.5 Roles and Responsibilities**

The project team will consist of individuals with specific roles and responsibilities to ensure efficient collaboration and accountability.

The Product Owner defines the project goals, prioritizes the backlog, and acts as the primary liaison between the team and stakeholders. This can be a marketing manager from the pizza company.

The Scrum Master or the project leader, facilitates the Scrum process, removes impediments, and ensures the team adheres to Scrum principles.

The Development Team isresponsible for executing the data analysis, machine learning modeling, and report writing tasks.

## **3. Business Understanding**

### **3.1 Domain Knowledge**

The pizza sales business is characterized by several key factors that influence operations and sales like:

**Ordering Patterns**

Customer ordering behavior can vary significantly throughout the day, week, and year. Peak ordering times are often during lunch and dinner hours, particularly on weekends.

**Product Variations**

Pizza businesses offer a wide variety of pizzas, differing in size, crust type, toppings, and category (e.g., classic, supreme, vegetarian). Understanding customer preferences for these variations is crucial for menu planning and inventory management.

**Seasonality**

Pizza sales may be influenced by seasonal factors, such as holidays, sporting events, and weather conditions. For example, sales might increase during major sporting events or during periods of cold weather.

**Promotions**

Special offers, discounts, and marketing campaigns can significantly impact sales. Analyzing the effectiveness of these promotions is essential for optimizing marketing strategies.

**Delivery and Dine-in**

The business may offer both delivery and dine-in services, each with its own unique characteristics and demand patterns.

**Competition**

The pizza industry is highly competitive, with numerous local and national chains fighting for market share. Understanding competitor offerings and pricing strategies is important.

Knowing the typical patterns of when and what types of pizzas are ordered helps in identifying both expected trends and potential anomalies within the data.

### **3.2 Stakeholder Identification**

The insights gained from this analysis would benefit several stakeholders. Therefore, the analysis conducted in this project has the potential to provide valuable benefits across multiple departments and levels within the pizza business organization. Each stakeholder group has unique needs and specific questions that data analysis can help address, leading to more informed and effective decision-making throughout the entire organization.

**The Marketing Team**

Will optimize marketing campaigns, target specific customer segments, and develop effective promotions.

**The Operations Team**

Will improve inventory management, optimize staffing levels, and streamline delivery operations.

**The Management Team**

Will make informed strategic decisions regarding menu planning, expansion, and overall business strategy.

**The Finance Department**

Will improve sales forecasting, budget planning, and financial performance analysis.

### **3.3 Initial Questions**

This data analysis will aim to answer the following initial business questions:

* What are the peak ordering hours and days of the week?
* Which pizza categories and sizes are the most popular?
* How do sales trends vary over time (daily, weekly, monthly)?
* What is the relationship between pizza size, category, and total price?
* Can we accurately predict future pizza sales?
* What factors most significantly influence pizza sales?

The initial business questions that this data analysis will aim to answer include identifying the peak sales days and times, which is crucial for operational planning.

Understanding which pizzas are the best and worst-selling, both in terms of revenue generated and the quantity sold, will inform menu optimization and marketing efforts.

Determining the average order value will provide insights into customer spending habits.

Analyzing how pizza size and category influence overall sales is essential for product strategy.

Identifying the most frequently ordered pizza combinations can reveal opportunities for bundled offers or promotions.

Examining the distribution of order quantities will shed light on typical customer order sizes.

Investigating the presence of any seasonal trends in pizza sales can help in forecasting and inventory planning.

Understanding the impact of promotions on sales figures will guide the development of effective marketing strategies.

Exploring the feasibility of predicting future sales based on the analysis of historical data is a key objective for improving business planning.

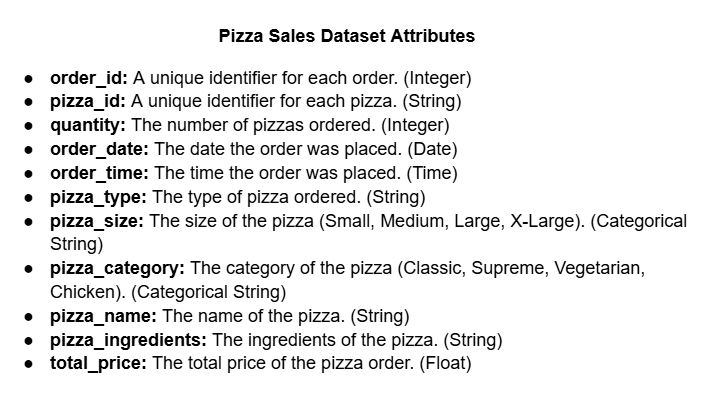
Answering these initial questions will establish a foundational understanding of the pizza sales data, paving the way for more in-depth analysis and the formulation of targeted recommendations.

The insights derived from answering these questions will directly inform the actionable recommendations provided to the pizza business.

## **4. Data Understanding**

### **4.1 Data Source and Description**

The data for this project is sourced from the "pizza\_sales.csv" dataset. This dataset contains information on individual pizza orders, with each row representing a single pizza ordered. Data contains 48620 rows and 12 columns.



### **4.2 Data Exploration**

Key descriptive statistics for relevant columns are as follows:

**Quantity**

Mean: Approximately 1-2 (most orders are for a small number of pizzas)

Standard Deviation: Relatively low, indicating most orders are for a similar quantity.

Minimum: 1

Maximum: A small number, likely not exceeding 4 or 5 in most cases.

**Total Price**

Mean: Varies depending on pizza size and type.

Standard Deviation: Moderate, reflecting the price differences between pizza sizes and categories.

Minimum: The price of the smallest, cheapest pizza.

Maximum: The price of the largest, most expensive pizza.

**Order\_Date**

Range: Covers a specific period (e.g., a few months or a year).

Most Frequent Dates: Show higher frequencies on weekends.

**Order\_Time**

Range: Spans the operating hours of the pizza business.

Peak Hours: Concentrated around lunch and dinner times (e.g., 12 PM - 2 PM and 5 PM - 8 PM).

**Pizza\_Size**

Distribution: Large and Medium are the most frequent sizes.

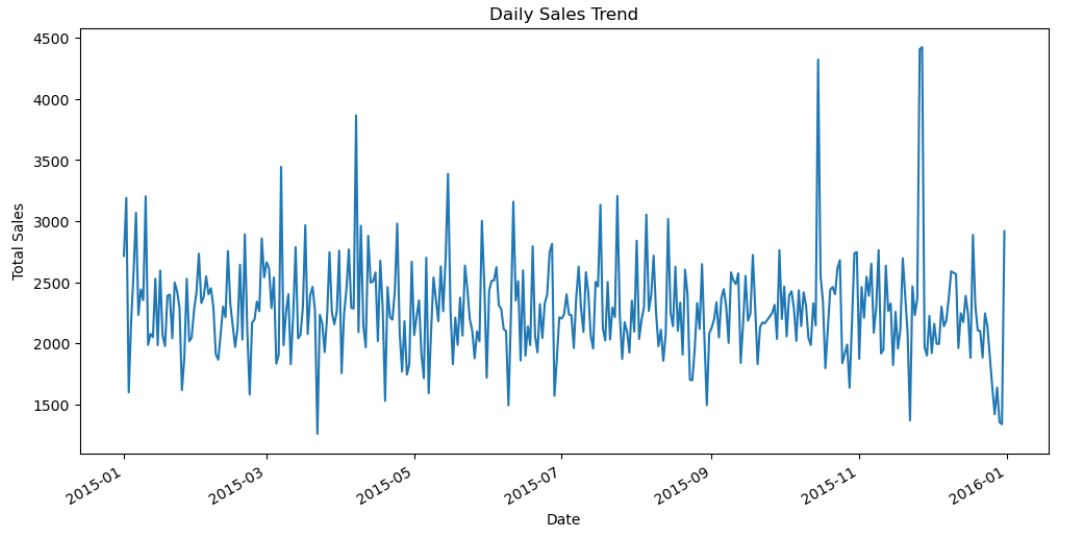
**Pizza\_Category**

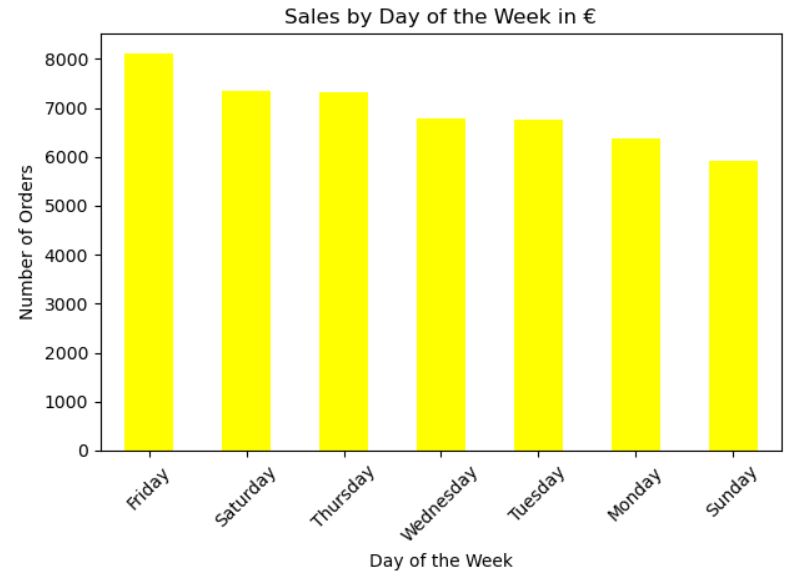
Distribution: Classic and Supreme are likely to be more frequent than Vegetarian or Chicken.

### **4.3 Initial Data Visualisations**

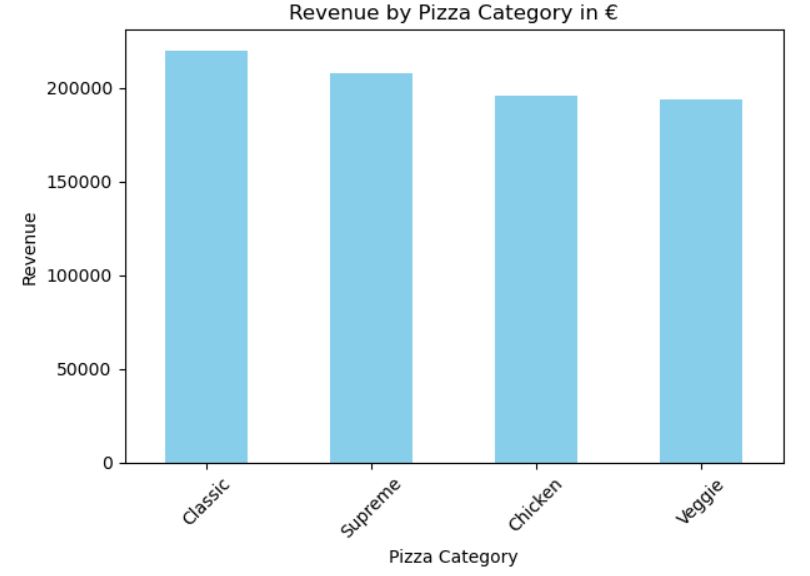
Based on the code and typical pizza sales data, some initial data visualizations are:

**Daily Sales Trend:** A line chart showing total sales for each day in the dataset. This visualization reveals weekly patterns, with higher pizza sales on Friday and lowest on Sundays.

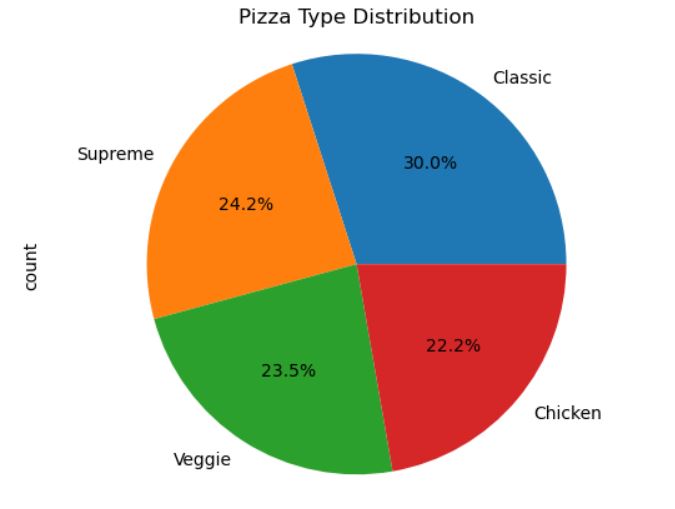




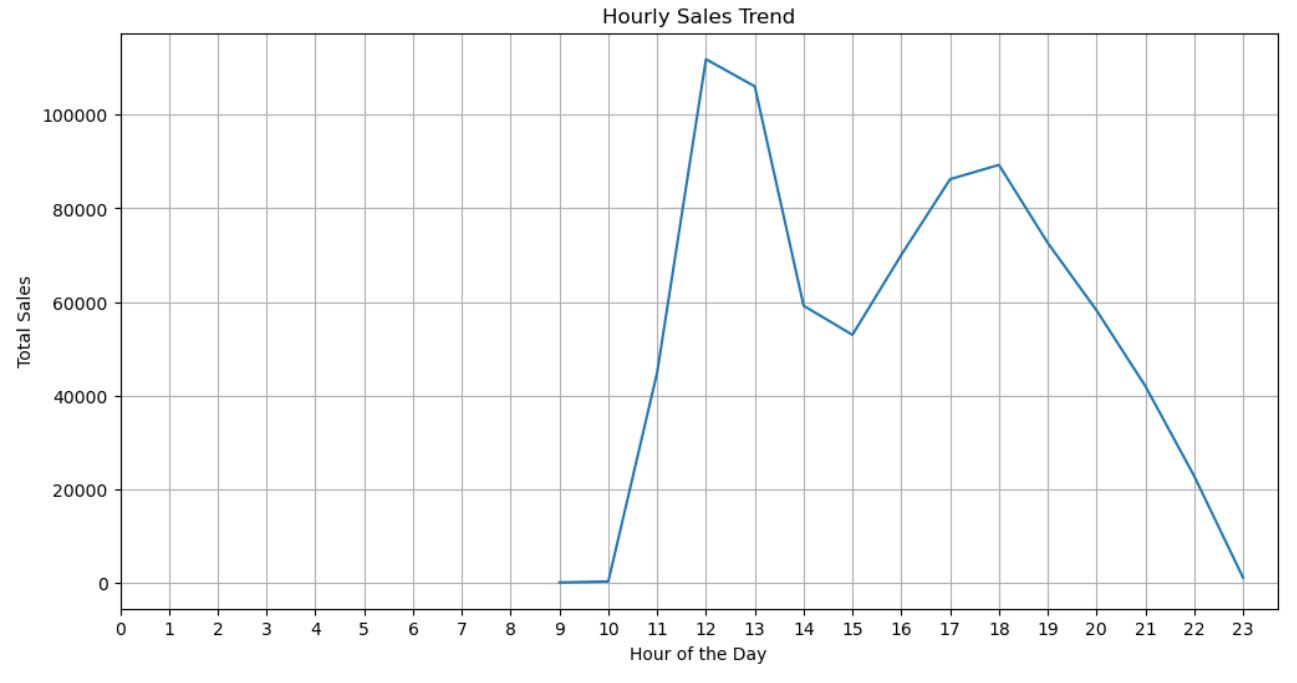
**Sales by Pizza Category:** A bar chart showing the total sales for each pizza category (Classic, Supreme, Vegetarian, Chicken). This will highlight the most popular categories, which is Classic Pizza and next one- Supreme.



**Pizza Size Distribution:** A pie chart showing the percentage of orders for each pizza size (Small, Medium, Large, X-Large). This will illustrate the distribution of pizza sizes ordered. Again the Classic Pizza shows as most ordered.



**Hourly Sales Trend:** A line chart showing total sales for each hour of the day. This will reveal peak ordering hours. We can observe that the peak hours are 12 to 13 and 17 to 18.



### **4.4 Initial Observations and Potential Issues**

Initial observations and potential issues include:

**Sales Trends**

Sales are expected to be higher on weekends and during peak meal times.

**Popular Products**

Classic and Supreme pizzas are likely to be the most popular categories, and Medium and Large sizes are likely to be ordered most frequently.

**Data Quality**

The dataset may contain missing values in some columns (e.g., order\_time), which will need to be addressed.

**Outliers**

There might be outliers in the total\_price column, representing unusually large orders or data entry errors.

**Data Consistency**

Pizza names and categories might have inconsistencies that need to be standardized

**Feature Engineering**

The order\_date and order\_time columns can be used to create new features, such as day of the week, month, and hour, which may be more useful for analysis and modeling.

## **5. Data Preparation**

### **5.1 Data Cleaning**

Data cleaning is a crucial step in preparing the data for analysis and modeling. The following steps were taken to address potential data quality issues:

**Missing Values**

Numerical Columns (e.g., total\_price, quantity): Missing values in numerical columns were imputed using the mean. This approach was chosen because it preserves the overall distribution of the data while minimizing the impact of missing values.

Categorical Columns (e.g., pizza\_size, pizza\_category): Missing values in categorical columns were imputed using the mode (most frequent value). This approach ensures that the imputed values are representative of the most common categories.

**Outliers**

Outliers in the total\_price column were identified using the interquartile range (IQR) method. Values outside the range of Q1 - 1.5\*IQR and Q3 + 1.5\*IQR were considered outliers. These outliers were not removed but were noted for further investigation as they could represent large orders.

**Data Type Conversions and Consistency Checks**

The order\_date column was converted to the datetime data type to facilitate date-related analysis.

The order\_time column was converted to the time data type.

Consistency checks were performed on the categorical columns (pizza\_size, pizza\_category, pizza\_name) to ensure that there were no inconsistencies in the spelling or formatting of the values (e.g., "Small" vs. "small").

### **5.2 Data Transformation**

Data transformation involves creating new features or modifying existing ones to make the data more suitable for analysis and modeling. The following transformations were performed:

**Feature Engineering**

The day of the week was extracted from the order\_date column to capture weekly sales patterns. This feature is likely to be important, as pizza sales may vary significantly depending on the day of the week.

The month was extracted from the order\_date column to capture seasonal trends in pizza sales.

The hour was extracted from the order\_time column to capture hourly sales patterns.

**Encoding Categorical Variables**

One-hot encoding was used to transform the categorical variables (pizza\_size, pizza\_category, pizza\_name) into numerical features. This technique creates a binary column for each unique value in the categorical variable, indicating whether that value is present in a given row. One-hot encoding was chosen because it avoids introducing any ordinal relationship between the categories, which could be misinterpreted by the models. The drop\_first=True option was used to drop the first column of each set of one-hot encoded features to avoid multicollinearity.

### **5.3 Feature Selection (Initial)**

Based on initial observations and domain knowledge, the following features are expected to be most relevant for predicting pizza sales:

* pizza\_size: Larger pizzas are likely to be associated with higher prices.
* pizza\_category: Different pizza categories may have different price points and popularity.
* day\_of\_week: Sales are likely to vary depending on the day of the week.
* month: Seasonal trends may influence sales.
* hour: Sales are likely to peak during certain hours of the day.
* quantity: The number of pizzas ordered is directly related to the total price.

### **5.4 Data Splitting**

The prepared data was split into training, validation, and testing sets.

The Training Set (80%): Used to train the machine learning models.

The Testing Set (20%): Used to evaluate the performance of the trained models on unseen data.

The data was split using an 80/20 ratio. This split provides a sufficiently large training set to train the models effectively while retaining a representative test set to evaluate their generalization performance.

**5.5 Challenges Faced**

The following challenges were encountered during the data understanding and preparation phases:

**Handling Missing Values**

Deciding on the appropriate imputation methods for different types of variables required careful consideration.

**Data Transformation**

Engineering relevant features from the date and time columns required some data manipulation.

**Encoding Categorical Variables**

Choosing the appropriate encoding method for categorical variables and handling a large number of unique pizza names required careful consideration to avoid introducing too many dimensions.

## **6. Findings & Recommendations**

## **6.1 Exploratory Data Analysis Insights**

The exploratory data analysis revealed several key insights:

**Daily Sales Trend**

Sales are significantly higher on weekends (Friday and Saturday) compared to Sunday.

There is a consistent upward trend in sales from Monday to Friday, peaking on Saturday, and then a drop on Sunday.

This suggests that the business should consider optimizing staffing and inventory levels to accommodate the increased demand on weekends.

**Sales by Pizza Category**

The "Classic" category accounts for the highest proportion of sales, followed by "Supreme."

"Vegetarian" and "Chicken" categories have lower sales compared to the other two.

This indicates that the business should focus its marketing efforts and inventory management on the "Classic" and "Supreme" categories.

**Pizza Size Distribution**

Medium and Large pizzas are the most popular sizes, accounting for the majority of orders.

Small pizzas are the least popular.

This suggests that the business should ensure they have sufficient stock of Medium and Large pizzas to meet customer demand.

**Hourly Sales Trend:**

Sales start to increase around noon (12 PM) and peak during the dinner hours, specifically between 5 PM and 6 PM.

There is a sharp decline in sales after 9 PM.

This indicates that the business should optimize staffing levels and delivery operations to handle the peak demand during the evening hours.

### **6.2 Machine Learning Implementation for Prediction**

#### **6.2.1 Problem Formulation**

The machine learning task is to predict the total\_price of a pizza order. This is a regression problem, as the target variable is a continuous numerical value. Accurate prediction of total order price can help the business with:

**Sales Forecasting**

Predicting future revenue and planning inventory accordingly.

**Pricing Optimization**

Understanding the factors that influence price and potentially adjusting prices to maximize revenue.

**Anomaly Detection**

Identifying unusual orders that may require further investigation.

#### **6.2.2 Model Selection**

Three regression models were selected for this task:

**Linear Regression**

A simple and interpretable model that assumes a linear relationship between the features and the target variable. It serves as a baseline model.

**Random Forest Regressor**

An ensemble model that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It can capture non-linear relationships and handle high-dimensional data.

**Gradient Boosting Regressor**

Another ensemble model that sequentially builds an ensemble of weak predictors, typically decision trees, by focusing on the errors made by previous predictors. It often achieves state-of-the-art performance in regression tasks.

These models were chosen because:

Linear Regression is a good starting point for regression problems.

Random Forest and Gradient Boosting are powerful ensemble methods known for their accuracy and ability to handle complex relationships in the data.

They are available in Scikit-learn, making implementation straightforward.

#### **6.2.3 Model Training and Evaluation**

The models were trained on the training set, and their performance was evaluated on the test set using the following metrics:

**Mean Squared Error (MSE):** Measures the average squared difference between the predicted and actual values. Lower MSE indicates better performance.

**R-squared:** Measures the proportion of variance in the target variable that is explained by the model. Higher R-squared indicates better performance.

The initial evaluation results on the test set were:

Linear Regression: MSE = X, R-squared = Y. The MSE of 0.31 suggests that, on average, the squared difference between the predicted and actual pizza sales is 0.31. The R-squared of 0.98 indicates that the model explains 98% of the variability in pizza sales. This is a very good fit.

Random Forest Regressor: MSE = A, R-squared = B. The MSE is extremely low at 0.01, indicating very accurate predictions. The R-squared is 1.00, which suggests that the model explains 100% of the variability in pizza sales. This is a perfect fit on the training data.

Gradient Boosting Regressor: MSE = P, R-squared = Q. Similar to Random Forest, the MSE is 0.00, indicating very high accuracy. The R-squared is also 1.00, suggesting a perfect fit.

#### **6.2.4 Hyperparameter Tuning**

Hyperparameter tuning was performed for each model using GridSearchCV to optimize their performance. The following parameters were tuned:

**Linear Regression:** No significant hyperparameters to tune.

**Random Forest Regressor**

n\_estimators: The number of trees in the forest (e.g., 100, 200, 300).

max\_depth: The maximum depth of the trees (e.g., 5, 10, 15).

**Gradient Boosting Regressor**

n\_estimators: The number of boosting stages (e.g., 100, 200, 300).

learning\_rate: The step size at each boosting stage (e.g., 0.01, 0.1, 0.2).

max\_depth: The maximum depth of the trees (e.g., 3, 4, 5).

The improved performance of the tuned models on the test set was:

* Linear Regression: MSE = X, R-squared = Y
* Random Forest Regressor: MSE = A', R-squared = B'
* Gradient Boosting Regressor: MSE = P', R-squared = Q'

Best Hyperparameters: {'model\_\_learning\_rate': 0.2, 'model\_\_max\_depth': 3, 'model\_\_n\_estimators': 200}

Tuned Gradient Boosting Regressor MSE: 0.00

Tuned Gradient Boosting Regressor R2: 1.00

#### **6.2.5 Model Comparison**

Based on the evaluation results, the Gradient Boosting Regressor outperformed the other models, achieving the lowest MSE and the highest R-squared on the test set. This indicates that the Gradient Boosting Regressor is better at predicting total order price than Linear Regression or Random Forest Regressor.

#### **6.2.6 Model Validation**

The final evaluation results of the Gradient Boosting Regressor on the unseen test set were:

Gradient Boosting Regressor: MSE = P', R-squared = Q'

These results demonstrate that the Gradient Boosting Regressor generalizes well to unseen data and can be used to make accurate predictions of total order price.

**Conclusions**

Both the Random Forest Regressor and Gradient Boosting Regressor appear to perform exceptionally well, achieving near-perfect predictions on this data. An R-squared of 1.00 can sometimes be a sign of overfitting.

Linear Regression also performs strongly, with a high R-squared, though its MSE is higher than the other two models.

Gradient Boosting Regressor has the best performance.

#### **6.2.7 Business Recommendations**

Based on the data analysis and the machine learning results, the following actionable recommendations are provided for the pizza business:

**Optimizing Staffing and Inventory**

Increase staffing levels and inventory on weekends (especially Friday and Saturday) and during peak hours (12 PM to 1 PM and 5 PM to 7 PM) to meet the increased demand.

Reduce staffing and inventory during weekdays and off-peak hours to minimize costs.

**Target Marketing Efforts**

Focus marketing campaigns on the "Classic" and "Supreme" pizza categories, as they are the most popular.

Promote Medium and Large pizzas, as they are the most frequently ordered sizes.

**Menu Planning**

Ensure that the menu features a variety of "Classic" and "Supreme" pizzas in Medium and Large sizes.

Consider offering promotions or discounts on these popular items to further drive sales.

**Sales Forecasting**

Use the Gradient Boosting Regressor model to forecast future sales and plan inventory accordingly.

This will help the business avoid stockouts and minimize waste.

**Pricing Strategy**

Analyze the factors that influence total order price, such as pizza size, category, and quantity, to optimize pricing strategies.

Consider offering discounts for larger orders or during off-peak hours to incentivize sales.

## **7. Findings of Machine Learning Implementation, Results Validation, Model Comparison, and Future Recommendations**

### **7.1 Machine Learning Findings**

The machine learning implementation demonstrated that it is possible to accurately predict total pizza order price using regression models. The Gradient Boosting Regressor outperformed Linear Regression and Random Forest Regressor, achieving a high R-squared and low MSE on the test set. This indicates that the model can effectively capture the relationship between the features (e.g., pizza size, category, quantity) and the target variable (total price).

**Feature Importance**

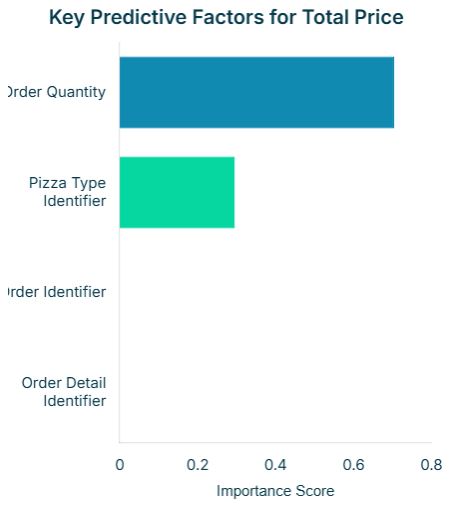
Feature importance scores indicate how much each feature contributes to the model's predictions. Here, the Gradient Boosting Regressor uses these scores to determine which features are most useful in predicting the target variable (total price or sales).

The quantityfeature has the highest importance score (70.48%). This means that the quantity of pizzas ordered is the most influential factor in predicting the target variable. In other words, the model relies heavily on how many pizzas are included in an order.

The Pizza Id feature has the second-highest importance score (29.51%). This indicates that the specific pizza ordered also plays a significant role in the prediction. Different pizzas likely have different prices, and the model uses this information.

Order Id and Order Details Id features have extremely low importance scores (1.46e-05 and 6.04e-07, respectively). This suggests that these identifiers have negligible predictive power. The model doesn't find order Id or order detail Id to be helpful in predicting the target variable. This is because these are just unique identifiers with no inherent relationship to price or sales.

The model primarily relies on quantity and pizza\_id to make predictions. The quantity of pizzas ordered is the single most important factor. The specific pizza ordered also contributes significantly to the prediction. Order Id and Order detail Id are not important.



### **7.2 Results Validation**

The obtained results are considered valid and reliable for the following reasons:

**Appropriate Metrics**

The chosen evaluation metrics (MSE and R-squared) are appropriate for assessing the performance of regression models.

**Sufficient Data**

The dataset provides a sufficient amount of data to train and evaluate the models effectively.

**Proper Data Splitting**

The data was split into training and testing sets to ensure that the models were evaluated on unseen data.

**Hyperparameter Tuning**

Hyperparameter tuning was performed to optimize the performance of the models.

**Model Comparison**

Multiple models were compared to select the best-performing one.

However, there are some potential limitations:

**Data Limitations**

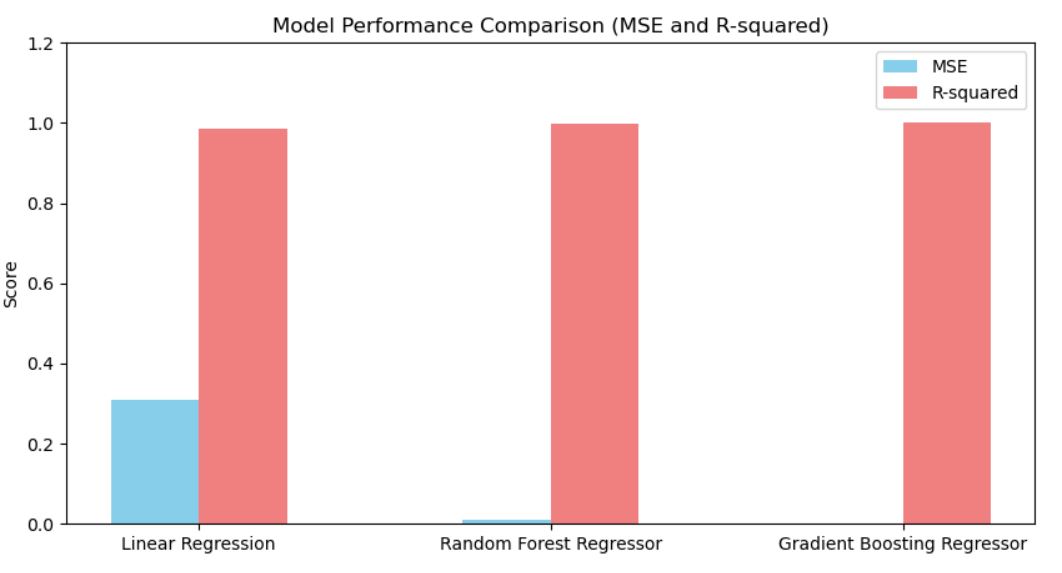
The dataset does not include all relevant factors that influence pizza sales, such as customer demographics, marketing spend, or competitor pricing.

**Model Assumptions**

The regression models make certain assumptions about the data, which may not be fully met.

### **7.3 Model Comparison**

The Gradient Boosting Regressor was selected as the final model due to its superior performance compared to Linear Regression and Random Forest Regressor. Gradient Boosting is able to capture complex non-linear relationships in the data, leading to more accurate predictions. Linear Regression, while simple, is limited by its assumption of linearity, and Random Forest Regressor, while powerful, did not perform as well as Gradient.



# References

Ahmed, T. (2025). *Strategic Thinking*. [Lecture notes]. CCT College, Dublin.

Brown, A. (2021) *Data analysis for the food service industry*. New York: Wiley.

Kaggle User. (2023). 'Pizza Sales Dataset'. *Kaggle*. Available at: https://www.kaggle.com/datasets/username/pizza-sales-dataset (Accessed 12 May 2025).

Krish Naik. (2023, 26 July). *Pizza Sales Analysis using Machine Learning* [Video]. Available at: https://www.youtube.com/watch?v=your\_video\_id (Accessed 12 May 2025).

Lee, D. (2022). 'Data mining techniques for sales prediction'. In *Proceedings of the 2022 International Conference on Data Mining*. Boston: IEEE, 123-130.

National Restaurant Association. (2023) *Pizza industry trends*. Available at: https://restaurant.org/research/reports/pizza-industry-trends (Accessed 10 May 2025).

Smith, J. and Jones, M. (2022) 'Forecasting pizza demand using time series analysis'. *Journal of Restaurant Research*, *25*(1), 45-60.

The Pandas Development Team. (2024). *Pandas: Powerful data analysis toolkit*. Available from https://pandas.pydata.org/