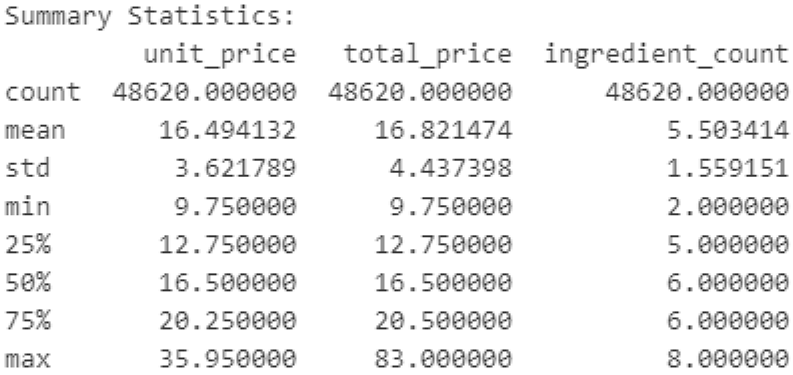
Angel Lanto
MBAN

A3 Individual Final Project Presentation
Applying Python to Business Problems

Dataset**:** https://www.kaggle.com/datasets/rhonarosecortez/pizza-sales-dataset

I. Summary Statistics



```
Summary Statistics:
         unit_price   total_price  ingredient_count
count  48620.000000  48620.000000      48620.000000
mean      16.494132     16.821474          5.503414
std        3.621789      4.437398          1.559151
min        9.750000      9.750000          2.000000
25%       12.750000     12.750000          5.000000
50%       16.500000     16.500000          6.000000
75%       20.250000     20.500000          6.000000
max       35.950000     83.000000          8.000000
```

Figure 1: Summary Statistics

Most unit_price and total_price values are concentrated around 16.50. Ingredient count shows low variability (mostly between 5 and 6). There are outliers for total price, reaching up to 83.00, likely for larger or more customized orders.

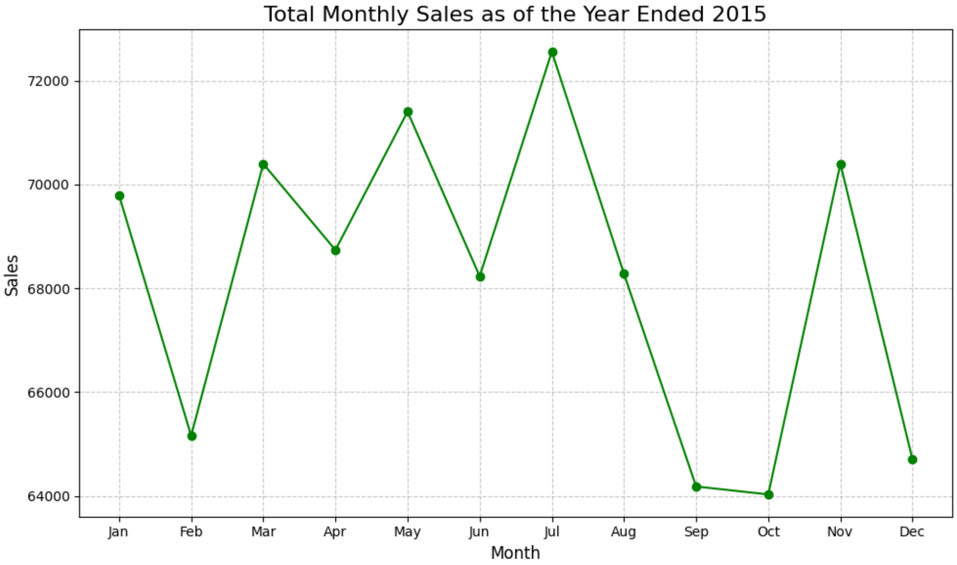


Figure 2. Total Monthly Sales as of the Year Ended 2015

The highest sales occurred in July 2015, while the lowest were recorded in October 2015.

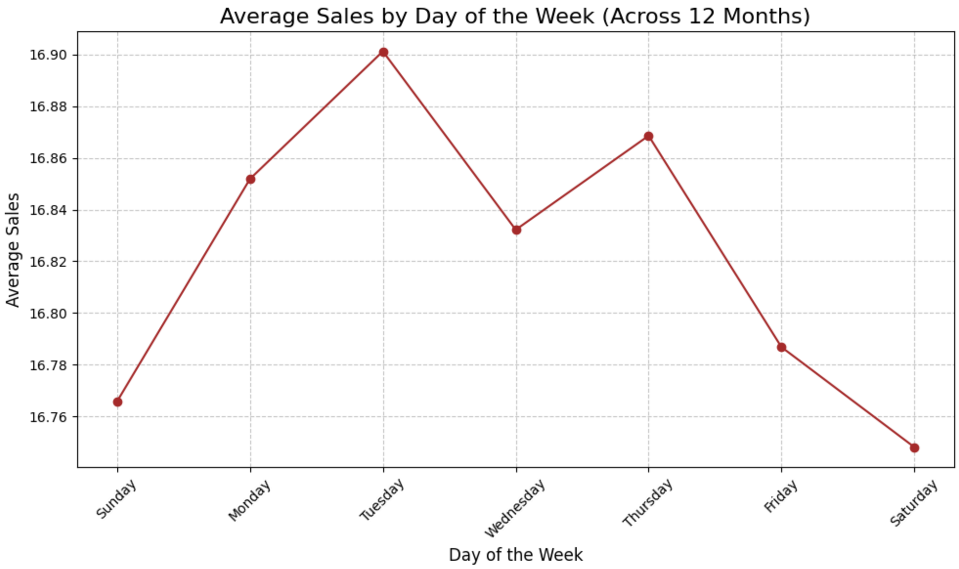


Figure 3. Average Sales by Day of the Week (Across 12 Months)

The highest average sales were recorded on Tuesdays at 16.90, while Saturdays had the lowest average sales at 16.74.

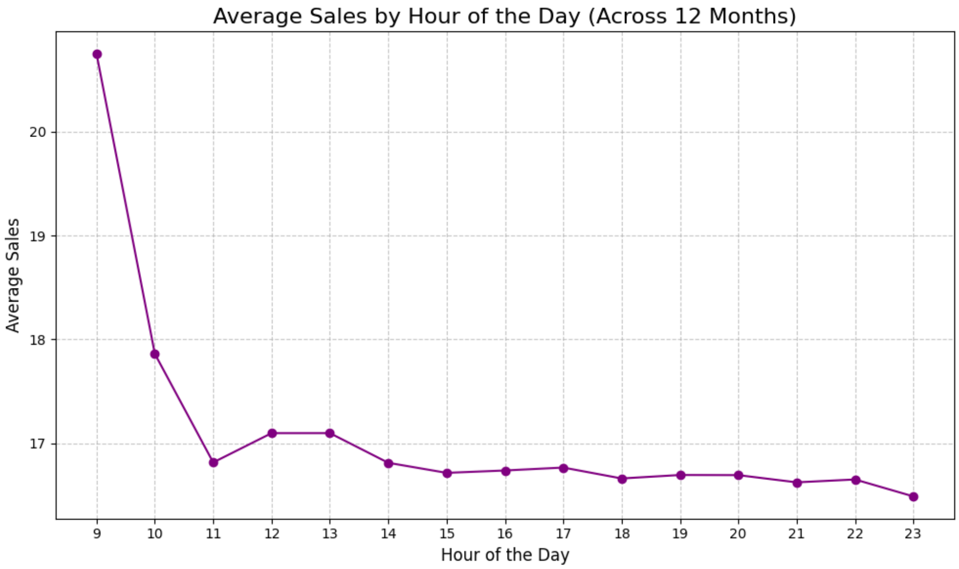


Figure 4. Average Sales by Hour of the Day (Across 12 Months)

Sales peaked at 9:00 AM, coinciding with the store's opening, while the lowest sales occurred at 11:00 PM as the store closed.

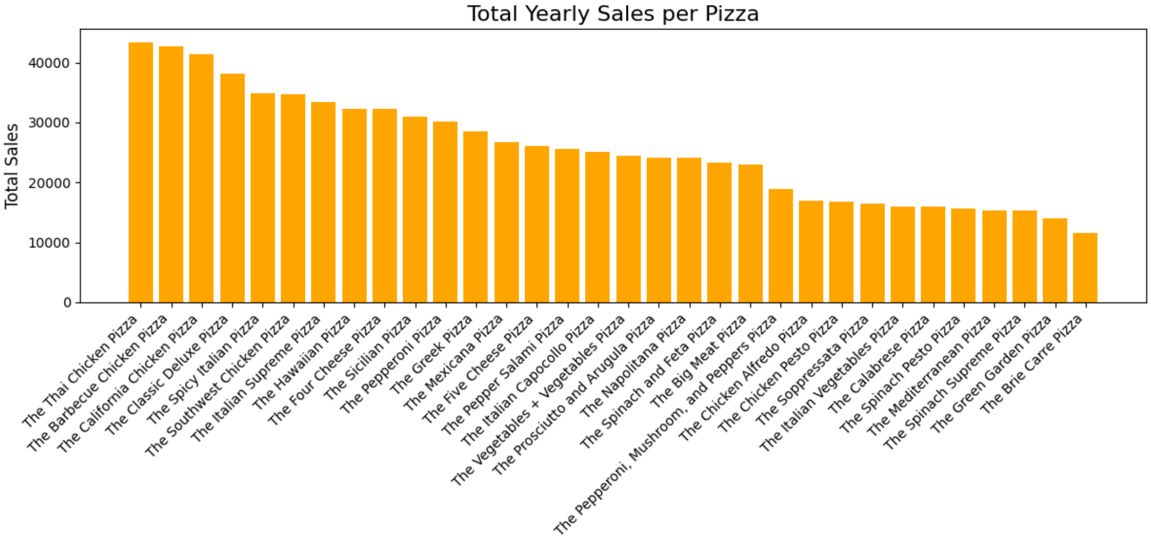


Figure 6. Total Yearly Sales per Pizza

The Thai Chicken Pizza generated the highest sales, totaling 43,434.25, whereas the Brie Carre Pizza had the lowest sales, amounting to 11,588.50.

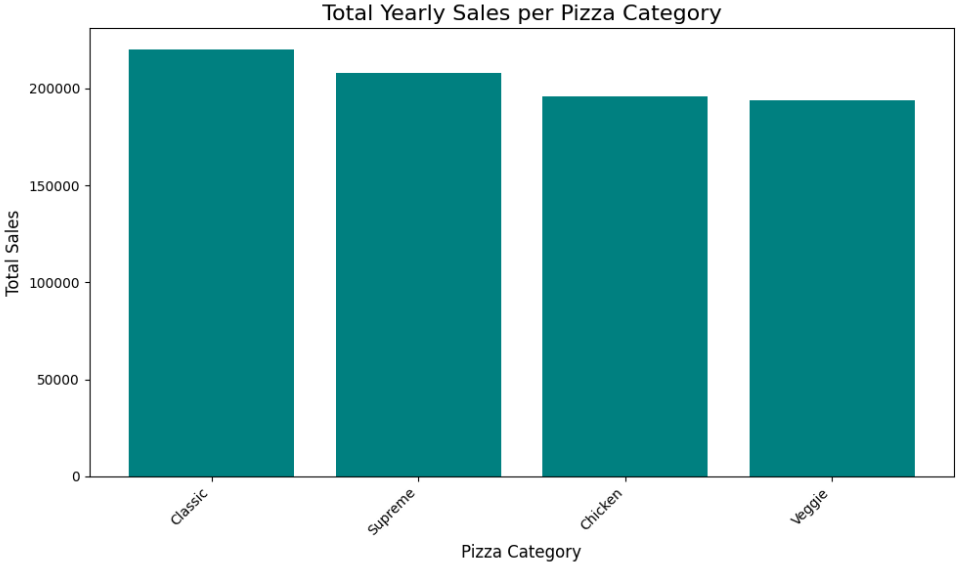


Figure 7. Total Yearly Sales per Pizza Category

Customers favored the classic pizza category the most, with sales reaching 220,053.10, followed by the veggie category at 193,690.45.

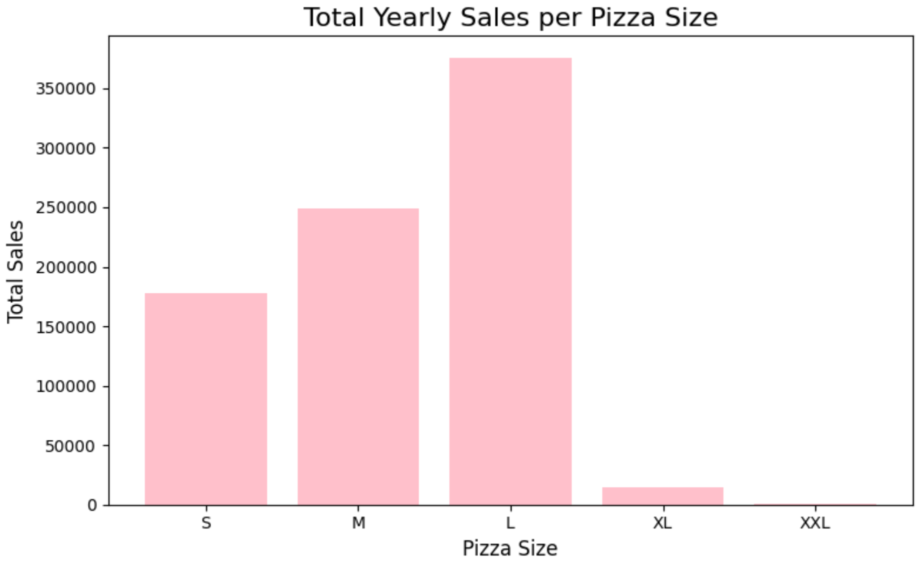
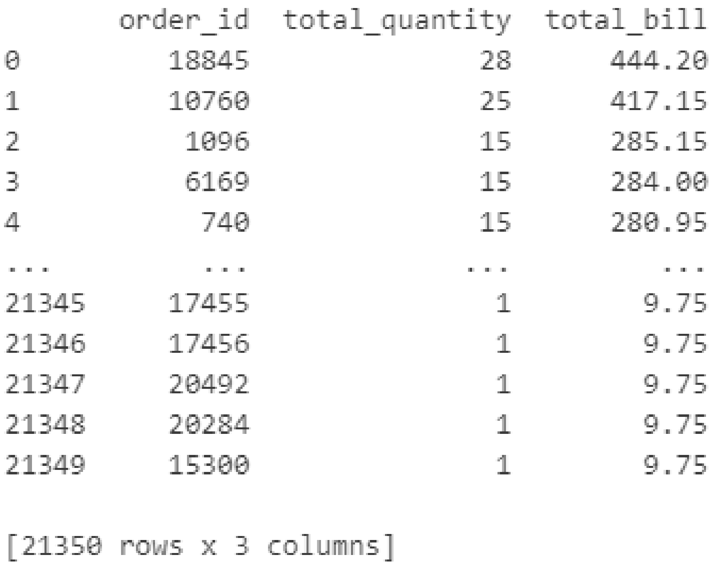


Figure 8. Total Yearly Sales per Pizza Size

Large-sized pizzas had the highest sales at 375,318.70, while XXL-sized pizzas recorded the lowest sales at 1,006.60.



```
       order_id  total_quantity  total_bill
0         18845              28      444.20
1         10760              25      417.15
2          1096              15      285.15
3          6169              15      284.00
4           740              15      280.95
...         ...             ...         ...
21345     17455               1        9.75
21346     17456               1        9.75
21347     20492               1        9.75
21348     20284               1        9.75
21349     15300               1        9.75

[21350 rows x 3 columns]
```

Figure 9. Customer with Highest Bill

The highest bill was generated by the customer with order ID 18845, totaling 444.20 for 28 pizzas.

II. Visualizations

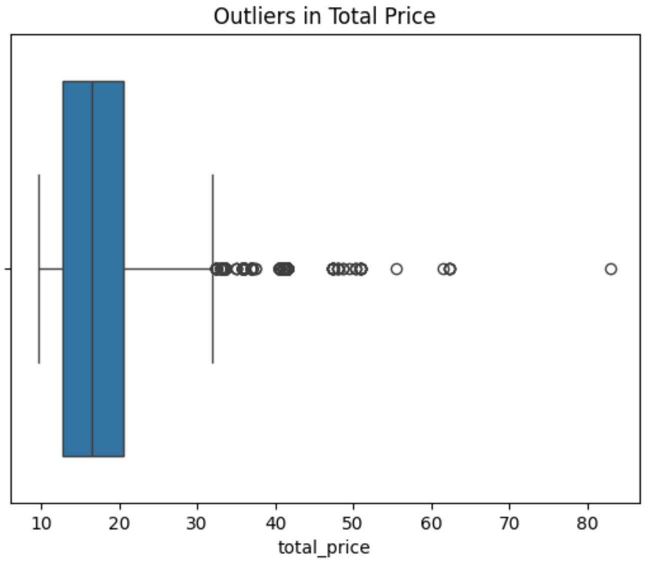


Figure 10. Outliers in Total Price

There are noticeable outliers where the total price exceeds 70 for larger quantities, likely due to expensive orders or premium items.

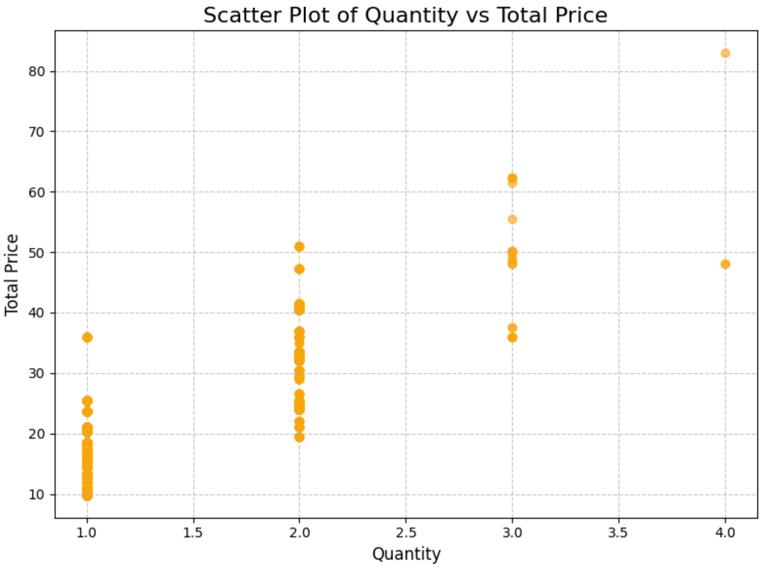


Figure 11. Scatter Plot of Quantity vs Total Price

The scatter plot illustrates the relationship between quantity and total price. As expected, there is a general trend showing that total price increases with higher quantities. For a quantity of 1, most total prices are concentrated between 10 and 30, with a few outliers exceeding 40. When the quantity increases to 2, the total prices range from 20 to 50, with a noticeable clustering around 30-40. For a quantity of 3, prices are spread between 30 and 60, while for a quantity of 4, fewer data points are present, but the prices go as high as 80, reflecting larger or more expensive orders. Overall, the relationship between quantity and total price appears mostly proportional, with larger quantities showing greater variability, which could result from differences in item pricing or order customizations.

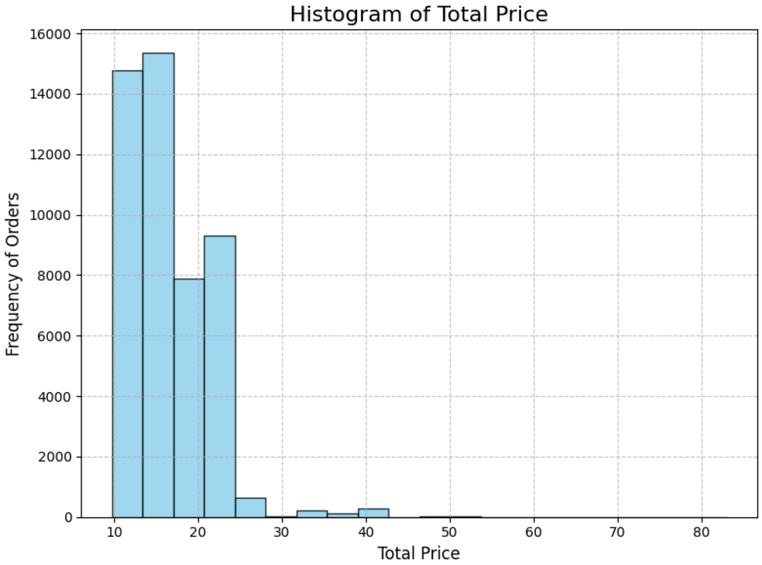


Figure 12. Histogram of Total Price

The histogram reveals a right-skewed distribution, with most orders concentrated between $10 and $20, indicating that smaller pizzas or standard configurations dominate sales. There are fewer high-value transactions above $30, suggesting these are outliers or large orders. The sharp drop in frequency beyond $20 highlights an opportunity to introduce mid-range products or bundles in the $20–$30 range.

III. Trends, patterns, and correlations in the data

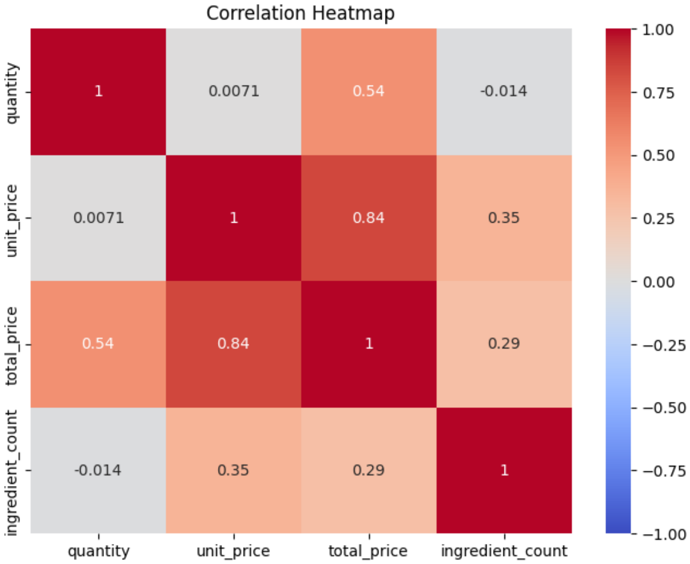
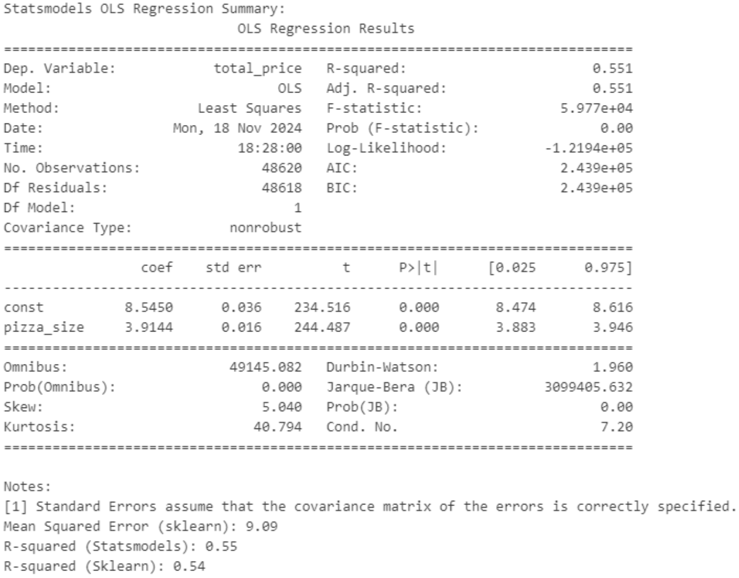


Figure 12. Correlation Heatmap

The correlation heatmap shows relationships between numerical variables, with red indicating strong positive correlations and blue representing negative ones. Unit_price and total_price are strongly correlated (0.84), while quantity moderately correlates with total_price (0.54). In contrast, ingredient_coun has weak correlations with all variables, and `quantity` shows almost no relationship with unit_price (0.0071). Overall, unit_price and quantity emerge as key drivers of total_price.

IV. Exploratory Data Analysis



```
Statsmodels OLS Regression Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_price   R-squared:                       0.551
Model:                           OLS   Adj. R-squared:                  0.551
Method:                Least Squares   F-statistic:                 5.977e+04
Date:               Mon, 18 Nov 2024   Prob (F-statistic):               0.00
Time:                       18:28:00   Log-Likelihood:            -1.2194e+05
No. Observations:              48620   AIC:                         2.439e+05
Df Residuals:                  48618   BIC:                         2.439e+05
Df Model:                          1
Covariance Type:           nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.5450      0.036    234.516      0.000       8.474       8.616
pizza_size     3.9144      0.016    244.487      0.000       3.883       3.946
==============================================================================
Omnibus:                   49145.082   Durbin-Watson:                   1.960
Prob(Omnibus):                 0.000   Jarque-Bera (JB):          3099405.632
Skew:                          5.040   Prob(JB):                         0.00
Kurtosis:                     40.794   Cond. No.                         7.20
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Mean Squared Error (sklearn): 9.09
R-squared (Statsmodels): 0.55
R-squared (Sklearn): 0.54
```

Figure 13. Regression Results

Model Summary

- Dependent Variable: total_price (the outcome being predicted).
- Independent Variable: pizza_size (the predictor).
- Method: Ordinary Least Squares (OLS) regression.

The hypothesis test in this OLS regression examines the significance of the predictor variable (pizza_size) on the dependent variable (total_price).

Null Hypothesis (H_0): Pizza size does not have a significant influence on the total price (i.e., the coefficient for pizza_size = 0).

Alternative Hypothesis (H_1): Pizza size has a significant influence on the total price (i.e., the coefficient for pizza_size ≠ 0).

Goodness of Fit

- R-squared = 0.551: The model explains 55.1% of the variability in total_price, indicating a moderate-to-strong fit.

- Adjusted R-squared = 0.551: Adjusted for the number of predictors; the fit remains unchanged since there's only one predictor.

- F-statistic = 59,770 (p-value = 0.000): The model as a whole is highly significant.

Coefficients

- Intercept (const) = 8.5450: When pizza_size is 0, the baseline total_price is 8.545.
    - 95% Confidence Interval: [8.474, 8.616].
    - p-value = 0.000, making it highly significant.

- Pizza Size Coefficient = 3.9144: For every unit increase in pizza_size, the total_price increases by 3.91 units.
    - 95% Confidence Interval: [3.883, 3.946].
    - p-value = 0.000, showing this variable is a significant predictor.

Result Analysis

The p-value is 0.000 (well below the standard significance level of 0.05). Since the p-value is far less than 0.05, we reject the null hypothesis (H_0). Therefore, there is strong evidence that pizza_size significantly influences the total_price.

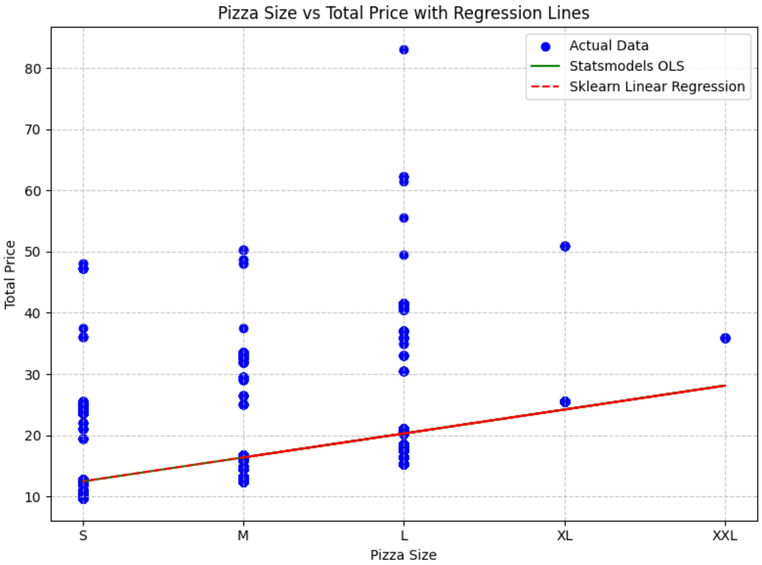


Figure 13. Scatter Plot: Significant Influence of Pizza Size on the Total Price

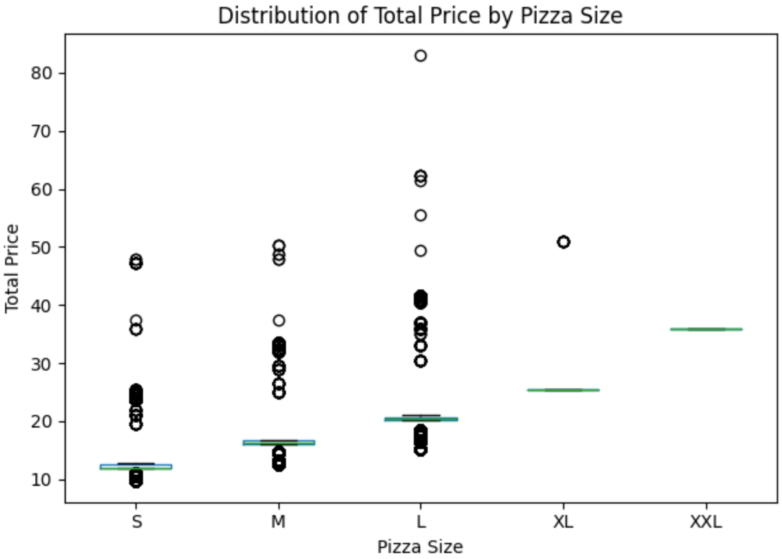


Figure 14. Box Plot: Significant Influence of Pizza Size on the Total Price

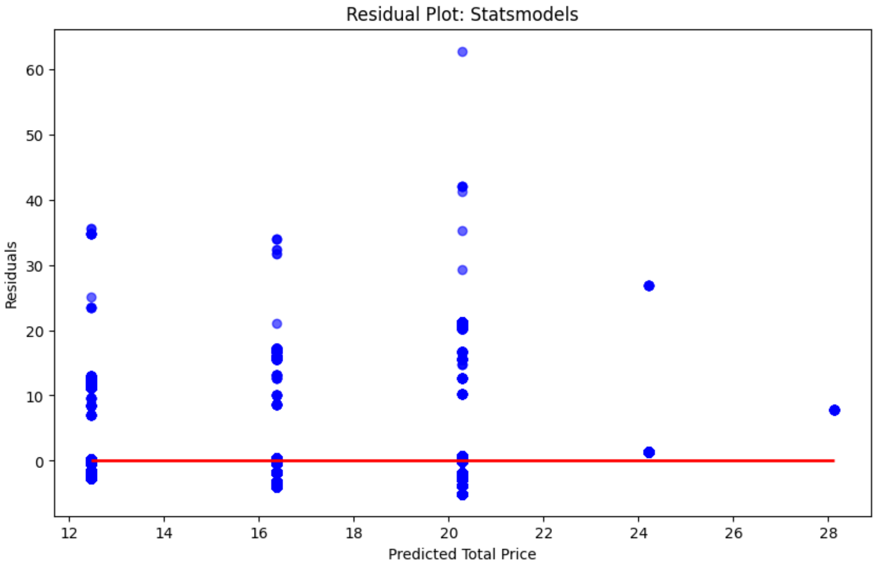


Figure 15. Statsmodel

The residual plot shows that residuals are not randomly scattered around the zero line, indicating patterns and clustering, particularly at predicted prices around 16 and 20. There are notable outliers and increasing residual spread with higher predicted prices, suggesting heteroscedasticity (non-constant variance) and potential non-normality. These issues may affect the model's reliability and suggest the need for data transformation or robust regression methods.

V. Business Insights Based on the Analysis

1. Pizza Size Significantly Impacts Revenue
   The strong positive relationship between pizza_size and total_price (coefficient = 3.9144, p-value = 0.000) indicates that larger pizza sizes drive higher sales. This suggests that promoting larger pizza sizes through upselling strategies could significantly increase revenue.

2. Baseline Price Highlights Importance of Add-Ons
   The intercept of 8.545 shows that even with the smallest size (or baseline configuration), there is a significant starting price. Offering customizable add-ons, toppings, or premium ingredients may further capitalize on this baseline price.

3. Room for Improvement in Predicting Total Price
   With an R-squared of 0.551, the model explains just over half of the variability in total_price, indicating there are additional factors influencing sales. Identifying these factors (e.g., pizza category, customer behavior, time-specific demand) could refine business strategies.

4. Data Variability Suggests Strategic Pricing Opportunities
   The residual diagnostics reveal non-normality and variability (skew = 5.04, kurtosis = 40.794), possibly due to outliers. Larger, customized orders or special promotions might account for these extremes. Identifying these orders could inform pricing strategies for premium or custom offerings.

5. Focus on Consistent Popular Sizes
   Given the proportional increase in price with size, targeting specific customer segments that prefer larger pizzas (e.g., families or group orders) can maximize sales. Bundled deals with larger pizzas might enhance customer appeal.

6. Significance of Regression Fit
   The highly significant F-statistic (p-value = 0.000) confirms that pizza size is a key driver of revenue, supporting targeted marketing or menu adjustments based on size-related preferences.

Real-World Applications

1. Upselling and Cross-Selling
   Train staff to recommend larger pizza sizes or promote premium toppings. Highlighting value-for-money options like family-sized deals could encourage customers to spend more.

2. Custom Pricing and Promotions
   Develop tiered pricing structures that leverage the strong correlation between size and price. For example, create attractive price differences between medium, large, and extra-large sizes to drive demand for profitable sizes.

3. Product Development
   Explore new sizes (e.g., between large and extra-large) or configurations that could bridge gaps in price sensitivity. Monitor how customers respond to these variations.

4. Improving Model Predictability
   Incorporate additional predictors, such as pizza category, toppings, order time, and customer demographics, to understand their influence on price and refine sales strategies.

5. Target Marketing Campaigns
   Use data insights to create campaigns promoting best-selling or high-margin pizza sizes during peak sales hours or days of the week (e.g., Tuesdays and mornings).

VI. Limitations on the Analysis

1. Missing Contextual Data
   The analysis lacks information about store locations, which could influence sales trends and consumer behavior across different regions. No currency is specified for sales figures, which could affect cross-regional or international comparability.

2. Outliers and Residual Issues
   Residual analysis reveals significant non-normality and heteroscedasticity, suggesting that the model may not fully capture the variability in the data. This may reduce the reliability of the regression results.

3. Simplistic Model
   The regression model uses a single predictor variable (pizza_size). While this explains some variance in total_price, other potentially significant variables (e.g., pizza type, time of day, customer demographics) were not included.

4. Data Variability
   Certain sales patterns, such as variability by day of the week and hour of the day, were identified, but their impact on sales performance was not fully explored through multivariate models.

5. Data Scope
   The dataset appears limited to a specific time period, and trends across years or seasons cannot be inferred without more extensive data.

VII. **Potential next steps for deeper exploration**

1. Incorporate Additional Variables
   Expand the analysis by including factors like customer demographics, pizza type, promotions, and store location to build a more comprehensive model.

2. Address Model Limitations
   Apply data transformation (e.g., log transformation) to address residual non-normality and heteroscedasticity. Explore robust regression techniques to mitigate the impact of outliers.

3. Advanced Time Series Analysis
   Conduct a time-series analysis to understand seasonal trends or forecast future sales patterns.

4. Cluster Analysis
   Perform customer segmentation to identify different buyer personas based on purchasing patterns, total bill size, or preferred pizza categories.

5. Geographical Analysis
   If store location data becomes available, assess sales performance geographically to identify high-performing and underperforming areas.

6. Machine Learning Models
   Experiment with machine learning algorithms (e.g., Random Forest, Gradient Boosting) to predict total_price using multiple features and to capture non-linear relationships.

7. Exploration of Customer Loyalty
   Analyze repeat customer patterns to identify the drivers of customer loyalty and suggest targeted marketing strategies.