ANALYZING CUSTOMER SHOPPING TRENDS USING STATISTICAL METHODS

A term project report submitted for

MA 541 - Statistical Methods

by

Hemanth Dadi

E Navaneet Kumar

Jayanth Kumar Yanamandala

Under the guidance of

Prof. HONG DO

Stevens Institute of Technology

TABLE OF CONTENTS

1. Introduction

In the contemporary landscape of data-driven decision-making, this study undertakes a meticulous analysis of customer shopping trends within the retail sector, employing statistical methods to unveil key insights from supermarket sales records. The dataset under scrutiny encompasses a diverse array of factors, ranging from unit price and quantity to tax, branch location, and customer-related metrics.

As we delve into this exploration, the overarching objective is to decipher the intricate patterns governing sales dynamics, customer behavior, and the factors shaping overall revenue in supermarkets. Two fundamental questions guide our inquiry: firstly, how do unit price and quantity interact to influence total sales? Secondly, what is the relationship between tax and total sales, especially for higher-priced items? Through these inquiries, we aim to distill actionable insights that can inform strategic decision-making within the complex retail landscape.

In the subsequent sections of this report, we will navigate through the substantive context of our study, providing a summary of the dataset and the critical factors influencing customer shopping behavior. The "big questions" addressed by our data analyses will be expounded upon, offering concise summaries of the conclusions drawn from these inquiries. Furthermore, a brief outline of the remainder of the paper will be presented, providing readers with a roadmap for the ensuing sections that delve deeper into our analytical methods, findings, and implications for the retail sector. This introduction sets the stage for a comprehensive exploration of customer shopping trends through the lens of statistical analysis, inviting readers to join us on a journey to uncover the intricacies of retail dynamics.

1.1 Summary of the Study

This data analysis report aims to explore and analyze customer shopping trends based on a dataset containing transaction information. The dataset includes invoice ID, branch location, customer type, gender, product line, pricing information, taxes, and more. The focus is on extracting meaningful insights to understand customer behavior and preferences.

2. Data Overview

The dataset comprises transactions with various attributes that provide insights into the retail environment. It includes both quantitative and categorical variables, offering a comprehensive view of customer interactions with the business.

2.1 Data

2.1.1 About the Dataset

The dataset encapsulates a comprehensive record of supermarket transactions, each distinguished by a unique Invoice ID. Here's a concise breakdown of the key attributes within the dataset:

1. Invoice ID: A distinctive identifier for each transaction.

2. Branch: Denotes the specific branch or location where the transaction took place.

3. City: Specifies the city in which the branch is situated.

4. Customer Type: Indicates whether the customer is classified as a regular or new customer.

5. Gender: Specifies the gender of the customer involved in the transaction.

6. Product Line: Categorizes the type of product purchased, offering insights into the diversity of items available.

7. Unit Price: Represents the cost of a single unit of the purchased product.

8. Quantity: Specifies the number of units of the product acquired in the transaction.

9. Tax 5%: Reflects the amount of tax applied to the transaction, calculated as 5% of the total cost.

10. Total: Signifies the overall cost of the transaction, inclusive of tax.

11. Date: Captures the date on which the transaction occurred, contributing to temporal insights.

12. Time: Records the time of day when the transaction took place, adding a temporal dimension to the dataset.

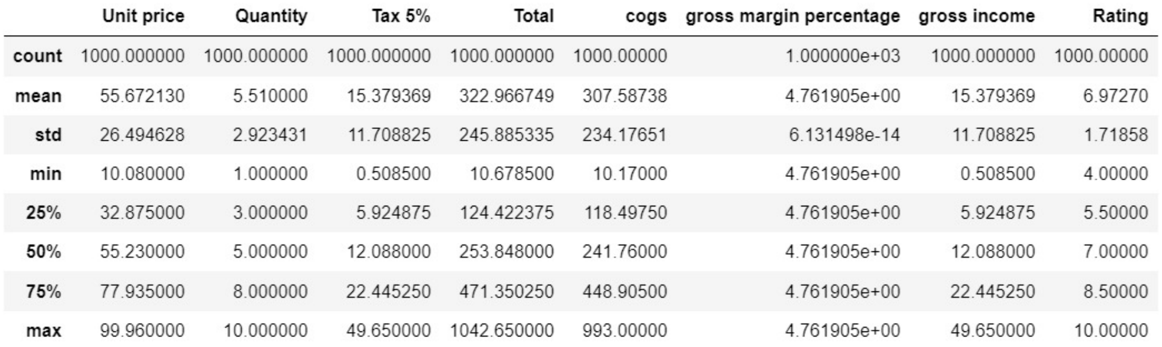
13. Payment: Identifies the payment method used for the transaction, ranging from credit cards to cash.

14. COGS (Cost of Goods Sold): Quantifies the direct costs associated with producing or purchasing the sold products.

15. Gross Margin Percentage: Specifies the profit margin percentage for each transaction, offering financial insights.

16. Gross Income: Represents the total profit earned from the transaction, contributing to financial analytics.

17. Rating: Captures customer satisfaction through a rating or feedback mechanism.



	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1.000000e+03	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905e+00	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	6.131498e-14	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905e+00	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905e+00	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905e+00	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905e+00	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905e+00	49.650000	10.00000

Summary Statistics:

The summary statistics provide a comprehensive overview of the numerical attributes within the dataset. Insights into the central tendency, spread, and distribution of variables such as Unit Price, Quantity, Tax 5%, Total, COGS, Gross Margin Percentage, Gross Income, and Rating are pivotal for understanding the overall transactional landscape.

Unique Values of Categorical Variables:

Categorical variables such as Invoice ID, Branch, City, Customer Type, Gender, Product Line, Date, Time, and Payment exhibit varying degrees of diversity, contributing to the rich complexity of the dataset. These unique values pave the way for nuanced analyses, providing a foundation for uncovering patterns and trends within supermarket transactions.

This foundational understanding of the dataset sets the stage for more in-depth exploration, enabling us to unravel the intricacies of customer shopping behavior and the factors influencing transactional dynamics.

2.2 Data Cleaning and Preprocessing

Before analysis, the dataset underwent cleaning and preprocessing steps to ensure data quality. This included handling missing values, checking for outliers, and converting data types where necessary. Additionally, any irrelevant or redundant information was removed to streamline the dataset for analysis.

2.3 Data Exploration Techniques

Data exploration involved visualizing trends and patterns in the dataset. Techniques included histograms, scatter plots, and pie charts to gain insights into the distribution of variables and potential relationships.

3. Visualizations

Visualizations serve as a powerful tool to convey patterns and trends within the dataset. Techniques such as histograms, scatter plots, and pie charts will be employed to visually represent the distribution of variables, relationships between attributes, and categorical proportions. These visualizations aim to enhance the interpretation of the dataset and facilitate a clearer understanding of customer shopping trends.

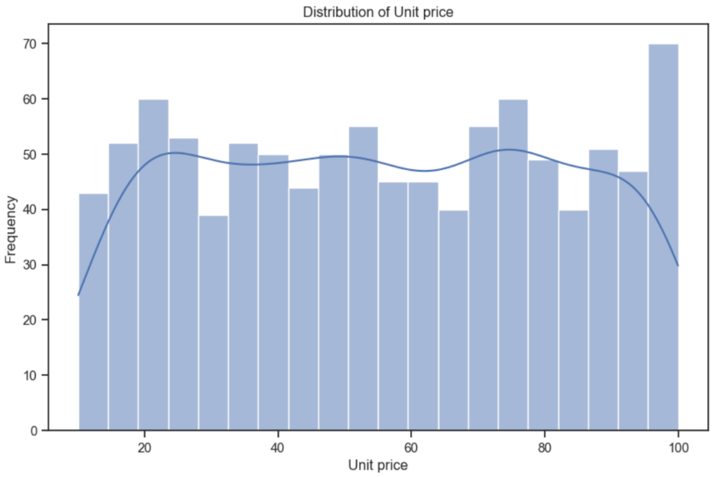
Visualizations play a crucial role in understanding the underlying patterns and trends in the dataset. The following visualization techniques have been employed to enhance the interpretation of the data:

3.1 Histogram

The histograms offer a visual representation of various business metrics. Here's an expanded analysis of each distribution based on the visual cues from the histograms:

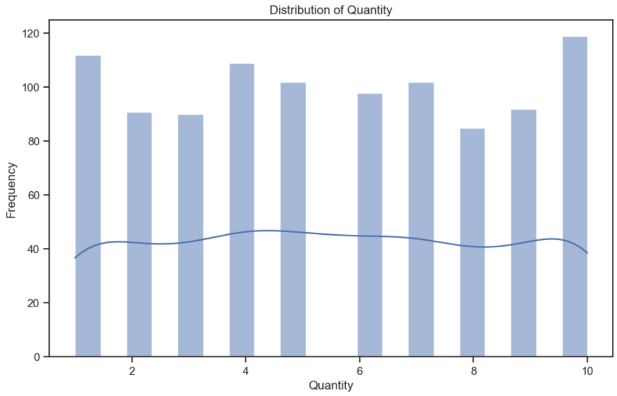
1. Distribution of Unit Price:

The multimodal nature of the unit price distribution suggests that certain price points are more common than others. This could reflect standard pricing strategies, psychological pricing points, or groupings of items at specific price ranges within the inventory. The presence of multiple peaks might indicate that items are grouped into distinct categories, each with its own common pricing level.



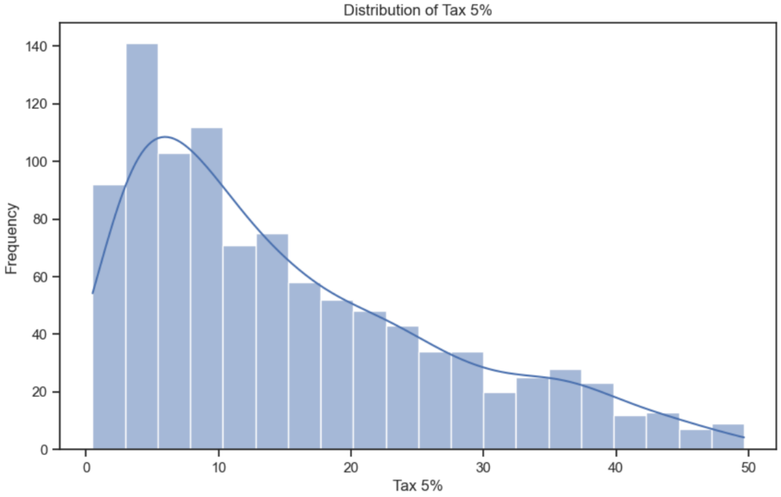
2. Distribution of Quantity:

The uniform distribution of quantities across various order sizes suggests that there isn't a particular quantity that is consistently preferred by customers. This could imply a diverse customer base with varied needs or that the business doesn’t incentivize purchasing in specific quantities. It may also reflect that the inventory consists of items that are equally likely to be bought in any quantity.



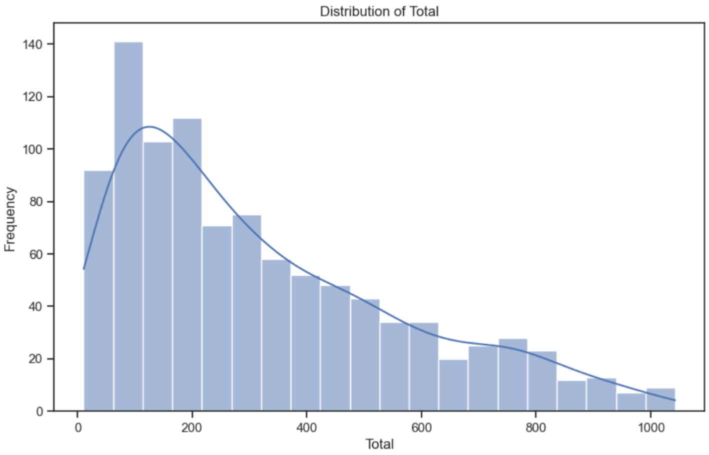
3. Distribution of Tax 5%:

The right-skewed distribution of tax values indicates that the majority of transactions are on the lower end of the tax scale, which in turn suggests that most purchases are of lower value. High-tax transactions are fewer, possibly indicating that expensive items or large purchases are less common.



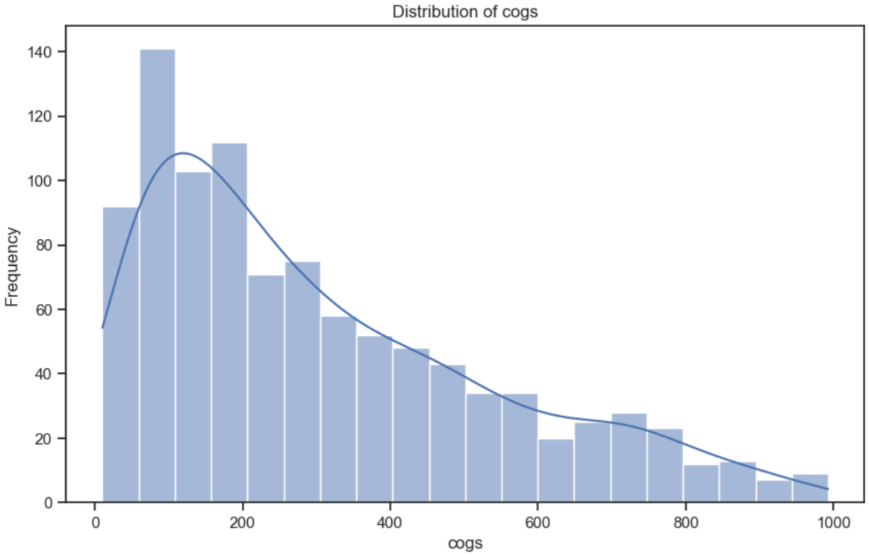
4. Distribution of Total:

The total value of transactions also shows a right-skewed distribution, hinting that while small to medium transactions occur frequently, large transactions are rare. This pattern often reflects typical consumer behavior in many retail environments where high-ticket items are less frequently purchased.



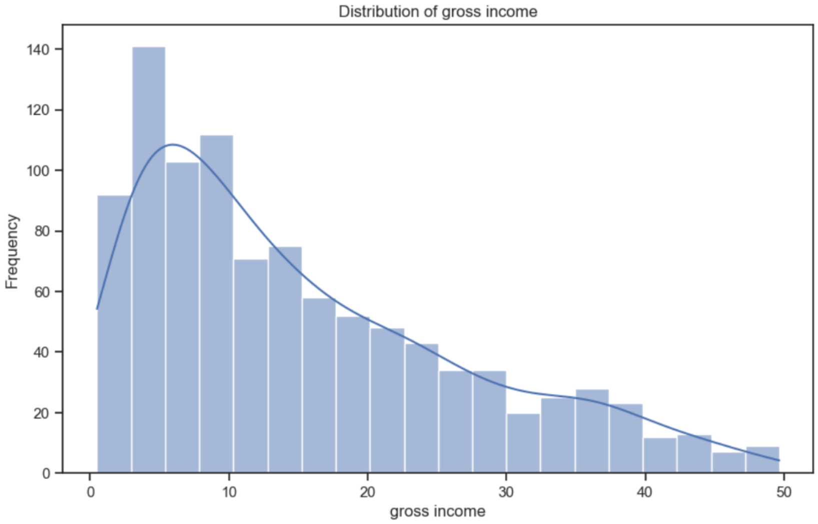
5. Distribution of COGS (Cost of Goods Sold):

The COGS distribution closely follows the total sales distribution, which is expected as they are directly related. The right skewness shows that transactions with high costs of goods sold are less frequent, aligning with the pattern that most sales transactions involve items with lower associated costs.



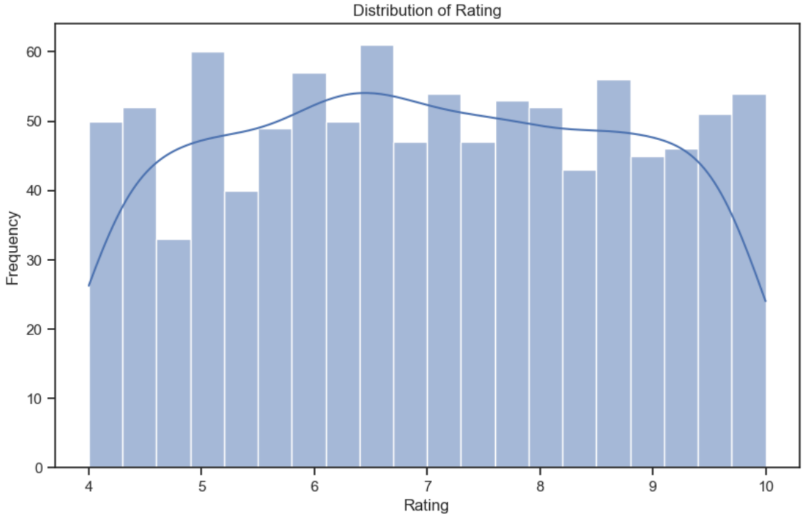
6. Distribution of Gross Income:

The gross income's right-skewed distribution suggests that the business typically earns lower gross income per transaction. This could be indicative of a volume-based business model where many small transactions contribute to the overall profit, rather than a few large transactions.



7. Distribution of Rating:

The concentration of customer ratings between 6 and 10 shows a trend toward positive feedback. The absence of lower ratings could indicate high customer satisfaction or possibly that less satisfied customers may be less inclined to leave a rating.



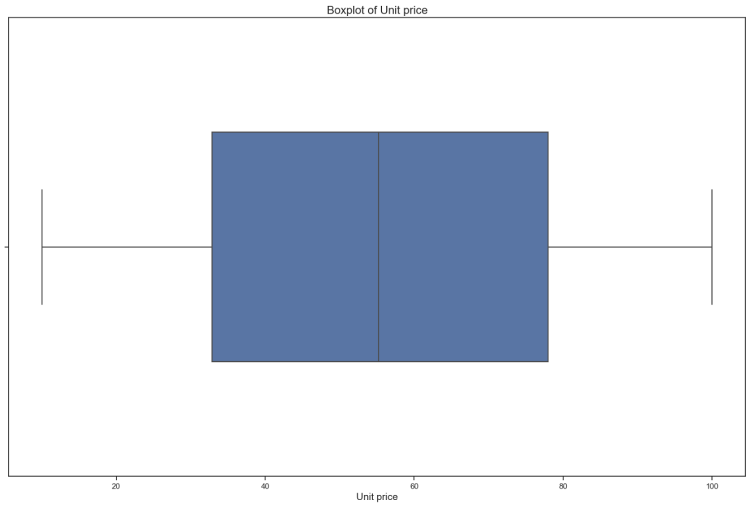
These distributions provide a quantitative look at the sales data, reflecting patterns in consumer behavior, pricing strategies, and the financial outcomes of transactions. Understanding these patterns can help the business make informed decisions about pricing, inventory management, and customer relations strategies.

3.2 Boxplots:

The boxplots provide another perspective on the distribution of the data across various business metrics. They summarize the central tendency, spread, and outliers within the datasets. Here's an expanded analysis based on the visual cues from the boxplots:

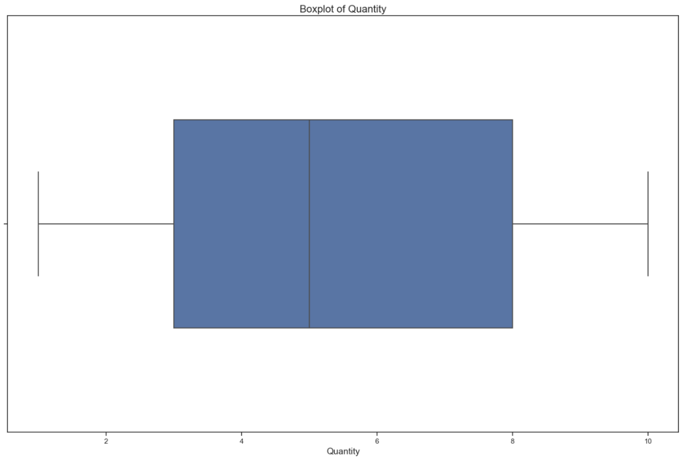
1. Boxplot of Unit Price:

The symmetric placement of the median indicates that unit prices are evenly distributed around a central value. This symmetry suggests that for every low-priced item, there's a correspondingly high-priced item. The absence of outliers implies price consistency and possibly a strategic range of pricing within the business model, with not abnormally low or high unit prices.



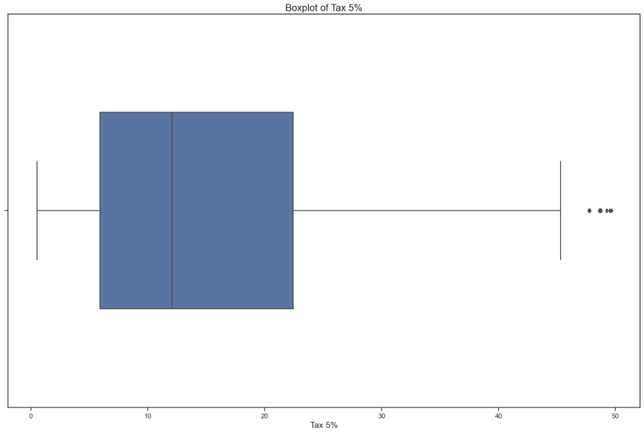
2. Boxplot of Quantity:

A median centered in the interquartile range (IQR) suggests an even spread of order quantities, which could indicate that the business does not cater to any specific order size but rather a wide range. The lack of outliers signifies uniformity in purchasing behavior, with all orders falling within a predictable range of quantities.



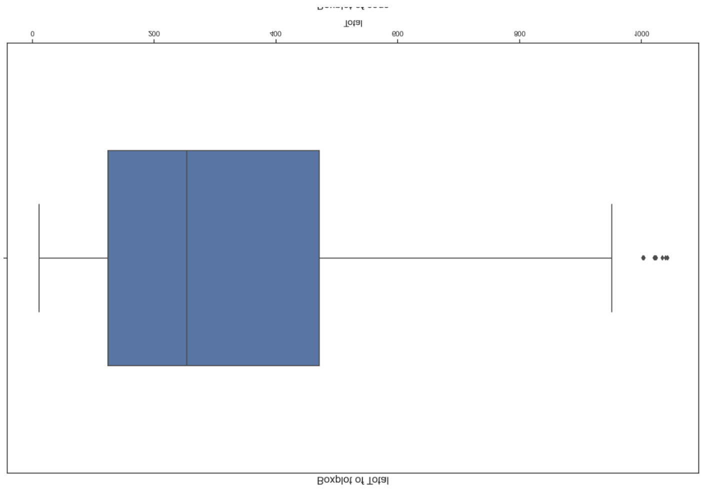
3. Boxplot of Tax 5%:

The lower-skewed median indicates that the majority of transactions involve lower tax amounts, which aligns with smaller transaction sizes being more common. Outliers on the higher end could represent unusually large transactions compared to the typical sales, perhaps due to bulk purchases or high-value items.



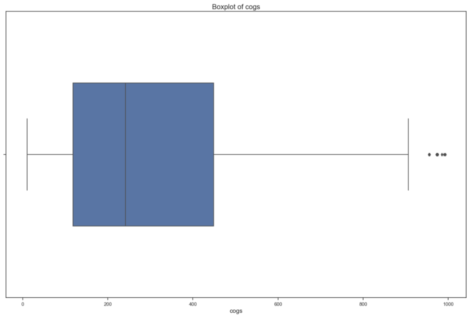
4. Boxplot of Total:

The right-skewed total amounts, with a median below the mean, reflect that while most transactions are smaller, there's a significant number of transactions that are considerably larger, evidenced by the outliers. This pattern might suggest that the business occasionally handles bulk orders or sells high-value items alongside more commonly sold lower-value items.



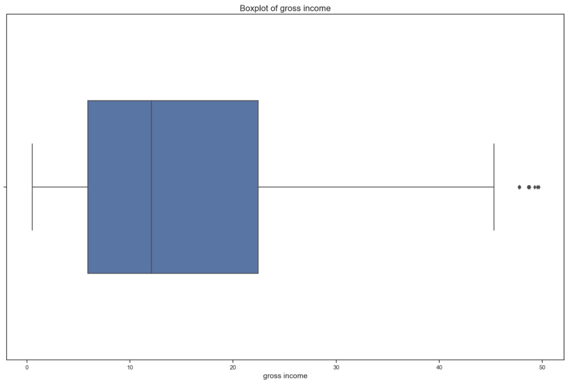
5. Boxplot of COGS (Cost of Goods Sold):

Mirroring the total sales, the COGS's distribution and outliers suggest that while most goods sold incur a standard cost, there are occasional sales that are much more costly for the business. This could also reflect that the business deals with a range of products with varying costs, from low-cost standard items to high-cost premium items.



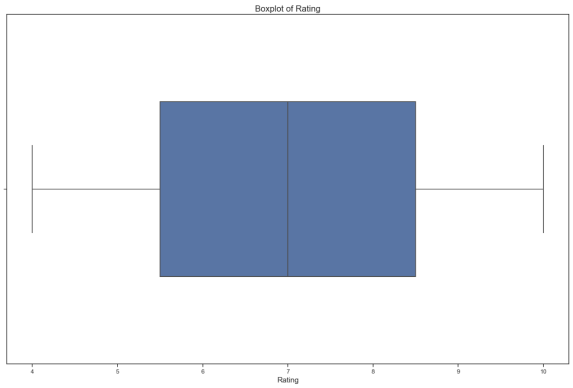
6. Boxplot of Gross Income:

The right-skewed distribution, with a median indicating moderate income per transaction, suggests that while the business consistently earns a certain level of income on most transactions, there are a few exceptionally profitable transactions. The outliers on the higher end could be due to sales with unusually high margins or larger-than-average sales volumes.



7. Boxplot of Rating:

The ratings show a concentration towards the higher end of the scale, with the median closer to the top of the box, suggesting that customers are generally satisfied with their experience. The relatively narrow IQR indicates that most ratings are clustered around a high median, which could be indicative of consistent quality of service or products provided by the business.



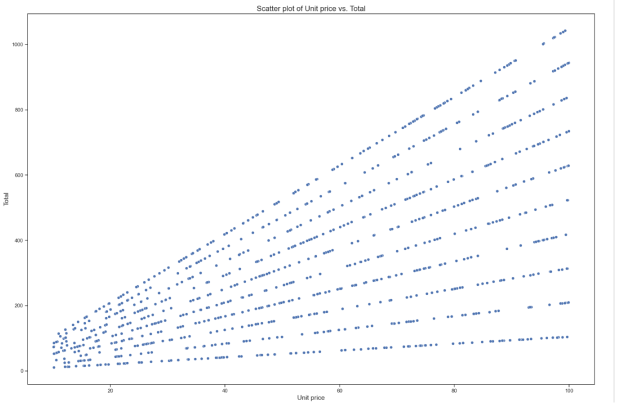
The boxplots provide a clear picture of variability and outliers, which are crucial for identifying any abnormal behavior or exceptional cases in the data. They can be particularly useful for decision-making processes related to inventory management, pricing strategies, and customer satisfaction initiatives.

3.3 Scatter Plots

The scatter plots are useful for understanding the relationships between different variables in a dataset. Here's an expanded analysis of each scatter plot:

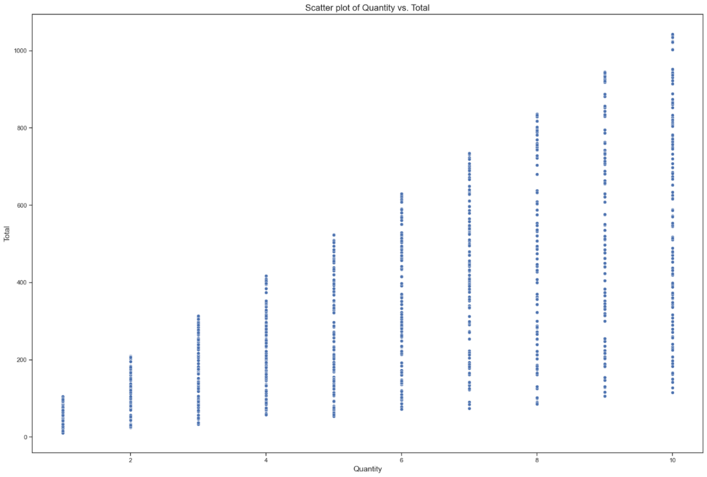
1. Scatter Plot of Unit Price vs. Total:

This plot shows a banding pattern, which suggests that for a given unit price, there are multiple total sales amounts. This could be due to the varying quantities of items sold at that price. The bands become denser as the unit price increases, indicating that higher-priced items may be sold in more varied quantities.



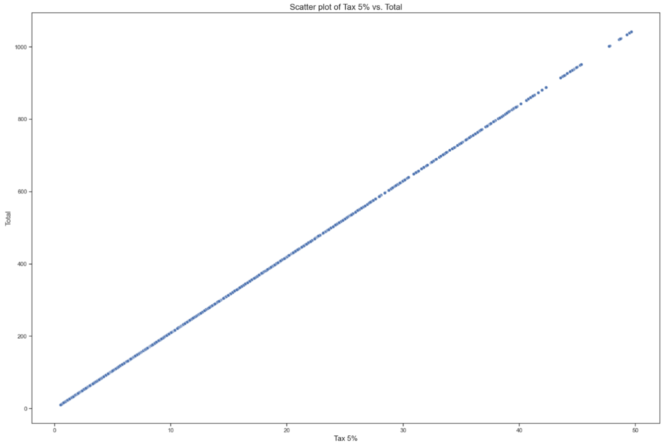
2. Scatter Plot of Quantity vs. Total:

The discrete horizontal lines represent distinct quantity levels. The linear increase in total sales with each step-in quantity indicates a direct, proportional relationship between quantity and total sales.



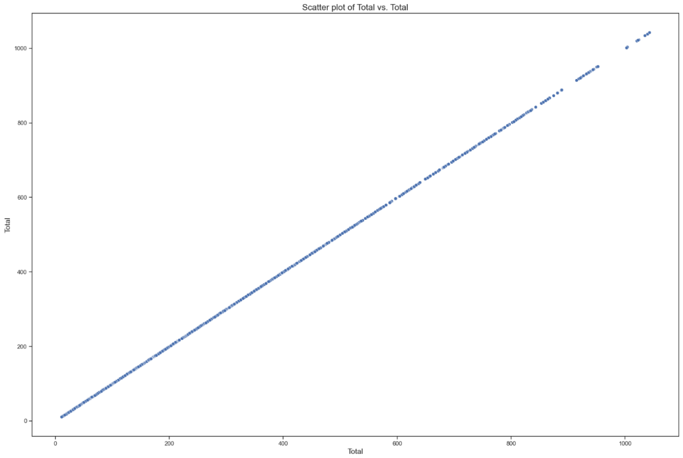
3. Scatter Plot of Tax 5% vs. Total:

The clear linear pattern shows the direct relationship between the sales total and the tax applied, which is expected given that tax is a percentage of the total sales. The uniformity of the points along the line indicates consistent tax application across all sales.



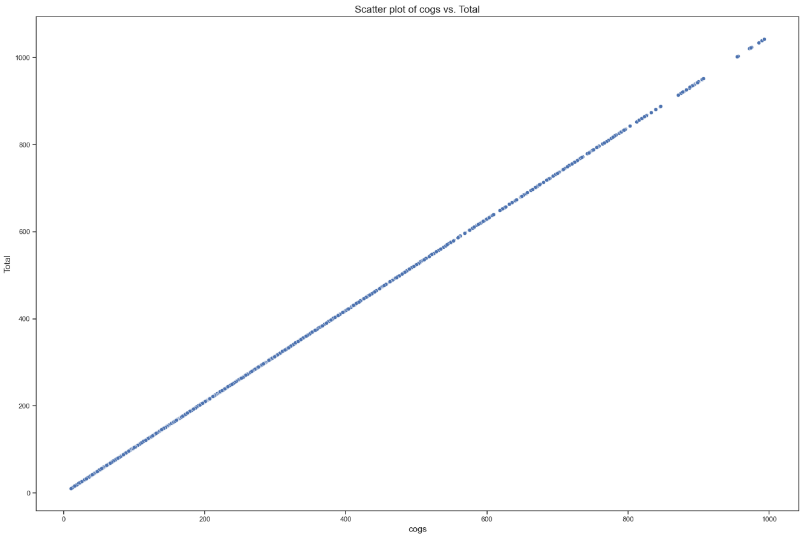
4. Scatter Plot of Total vs. Total:

This plot, likely a comparison of the total sales over different periods (like daily totals), shows a perfect linear relationship, indicating an error in the plot as it is comparing the same variable against itself.



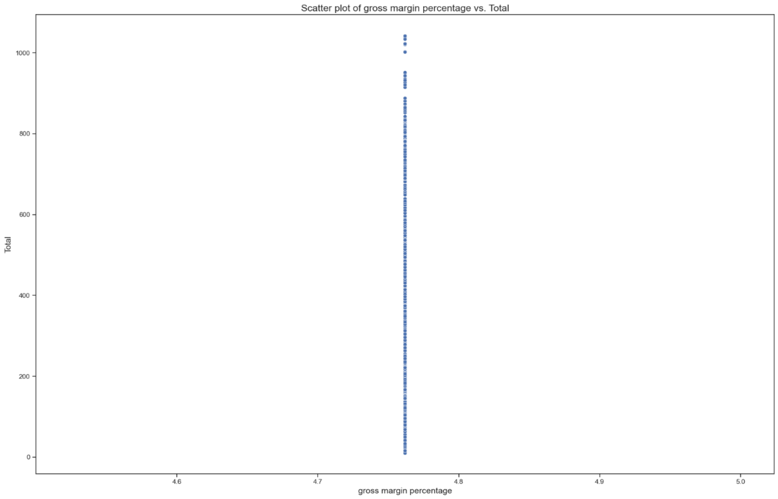
5. Scatter Plot of COGS vs. Total:

The linear relationship between COGS and total sales indicates that as the cost of goods sold increases, the total sales also increase proportionally. This suggests that markup or profit margin is relatively consistent across different sales values.



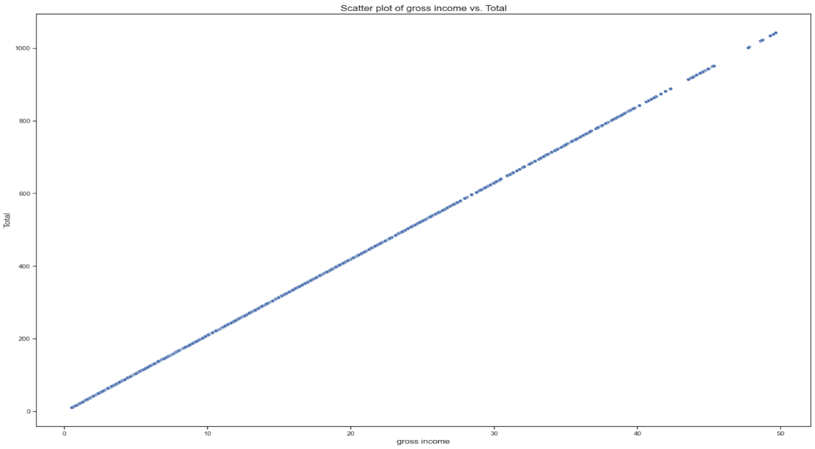
6. Scatter Plot of Gross Margin Percentage vs. Total:

The concentration of points at specific gross margin percentage values suggests that there is a standard margin applied across most sales. The lack of spread along the x-axis implies that the gross margin percentage does not vary much with changes in the total sales value.



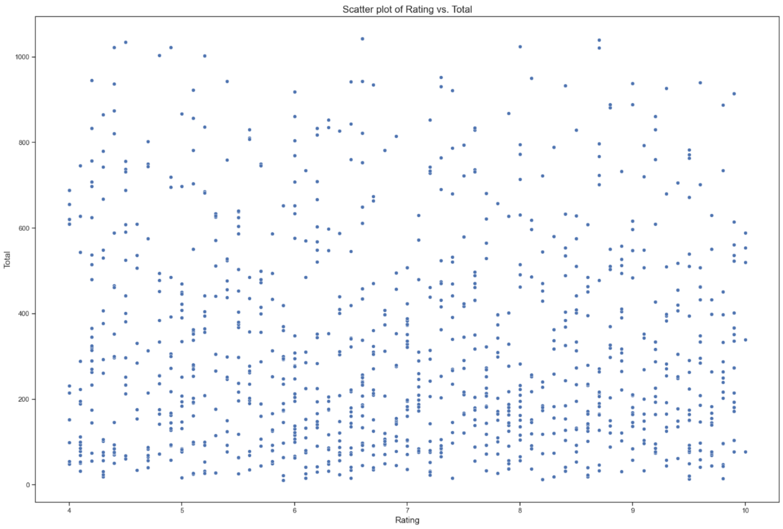
7. Scatter Plot of Gross Income vs. Total:

The linear pattern indicates a direct correlation between gross income and total sales. As total sales increase, the gross income increases proportionately, suggesting a consistent profit margin across different transaction sizes.



8. Scatter Plot of Rating vs. Total:

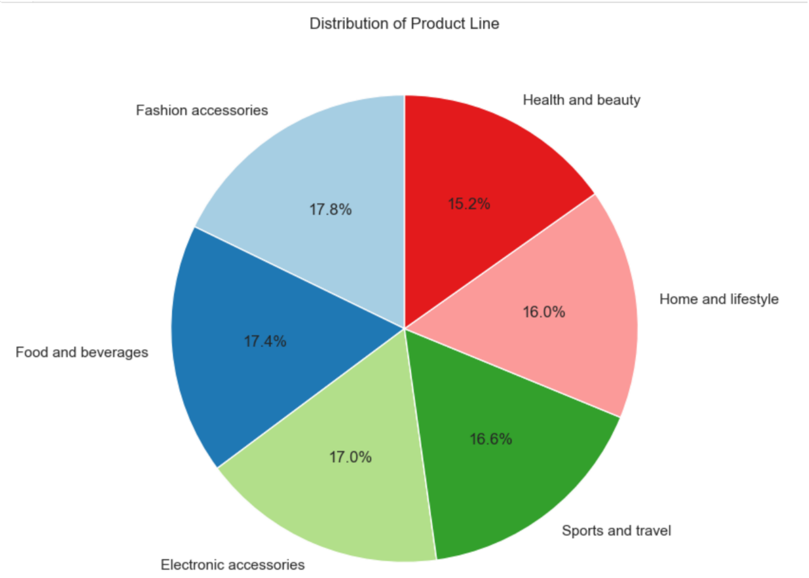
This plot shows a wide dispersion of points with no apparent trend, indicating that there is no clear relationship between the customer rating and the total sale amount. This suggests that customer satisfaction (as measured by rating) is not directly related to the amount spent.



These scatter plots can help identify trends, outliers, and the strength of relationships between variables. They are valuable for data analysis, allowing businesses to understand patterns and correlations in their sales data, which can inform strategic decisions in pricing, sales forecasting, and customer relationship management.

3.4 Pie Chart:

The pie chart shows the distribution of sales by product line for a business. Here’s an expanded analysis based on the visual information:



1. Fashion Accessories:

At 17.8%, fashion accessories take the largest share of sales distribution, suggesting a strong market presence or customer preference for these items within the business's product range.

2. Health and Beauty:

Health and beauty products account for 15.2% of sales, representing a significant portion but slightly less than fashion accessories. This indicates a healthy demand in this category, though not the leading one.

3. Home and Lifestyle:

With 16.0% of sales, home, and lifestyle items are in the middle range in terms of sales distribution. This category holds a substantial market share and is competitive with the leading categories.

4. Food and Beverages:

Food and beverages represent 17.4% of sales, nearly matching the leading category of fashion accessories. This suggests that consumable goods are a strong driver of sales for the business.

5. Electronic Accessories:

Electronic accessories account for 17.0% of sales, indicating strong consumer interest and market performance like that of food and beverages, and fashion accessories.

6. Sports and Travel:

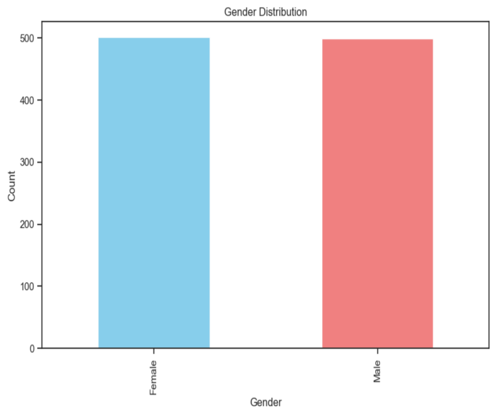
At 16.6%, sports and travel products have a significant but slightly smaller share than the leading categories, which may indicate a niche but solid customer base for these products.

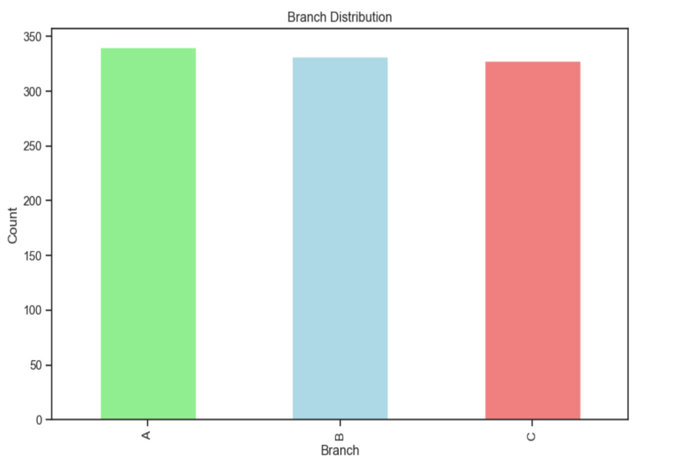
The fairly even distribution across product lines suggests a diversified business model with no single product line dominating the sales. It may also reflect a strategic balance in inventory or a wide customer base with varied interests. The business might benefit from maintaining this diversification, ensuring that changes in demand for one category don't disproportionately affect overall revenue.

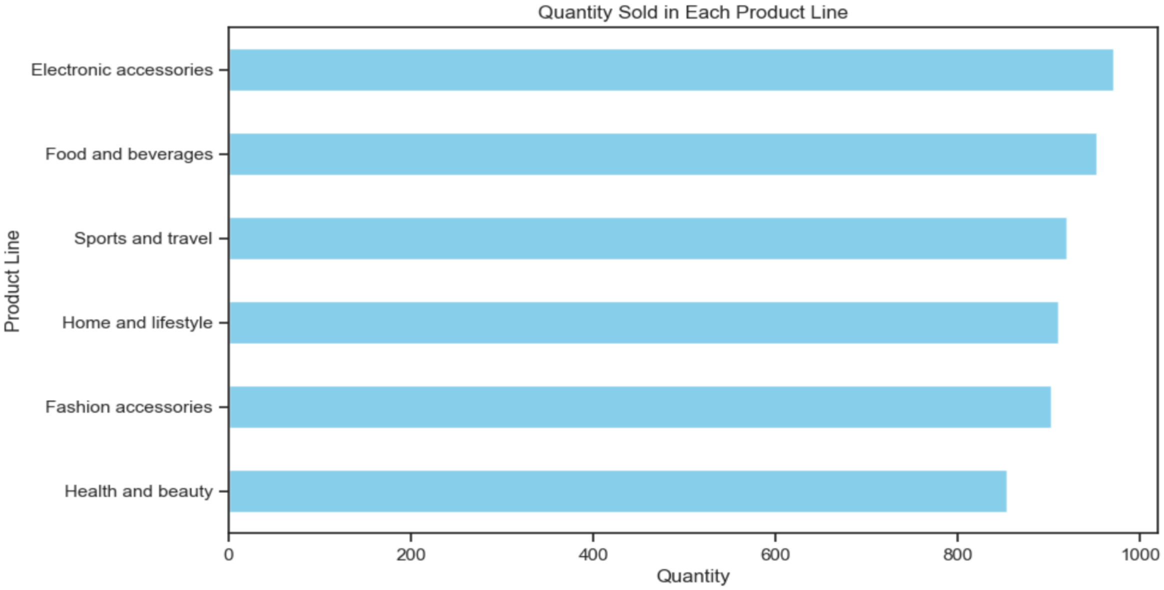
Understanding this distribution can help the business in decision-making related to inventory stocking, marketing strategies, and potential areas for expansion or promotion. It can also inform the business about consumer trends and preferences, guiding product development and customer engagement initiatives.

3.5 Bar Charts:

The bar charts you've provided show the distribution of counts across different categories, specifically gender, branch, and product line sales.







1. Gender Distribution:

The chart shows a close count between females and males, with females slightly higher. This suggests that the customer base is fairly balanced in terms of gender, with a minor skew towards female customers. This information can be useful for tailoring marketing strategies and product offerings to better serve the slight majority while still catering to all customers effectively.

2. Branch Distribution:

The distribution among branches A, B, and C indicates that Branch A has the lowest count, while Branch C has the highest. This could reflect various factors such as location, size, product availability, or local marketing effectiveness. The business might investigate why Branch C performs better and apply those strategies to other branches, or conversely, it could examine the challenges facing Branch A to improve its performance.

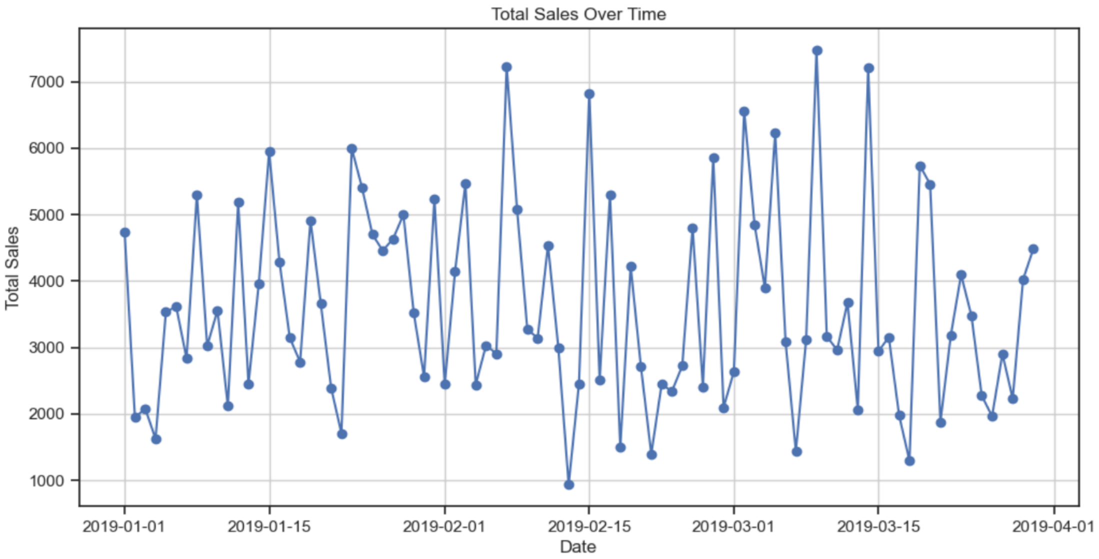
3. Quantity Sold in Each Product Line:

The horizontal bar chart illustrates that the quantity sold is highest for electronic accessories, followed closely by food and beverages, then sports and travel. Home and lifestyle, fashion accessories, and health and beauty have lower sales quantities in comparison. This hierarchy in product line popularity can guide inventory management, and promotional efforts, and potentially indicate which lines may benefit from further development or expansion.

For a business, understanding these distributions is critical for making informed decisions. The gender distribution can help in creating targeted marketing campaigns. The branch distribution data might lead to more focused management strategies to boost sales in underperforming branches. Lastly, the quantity sold by product line provides insights into consumer preferences, which can influence stock purchasing decisions and help in identifying potential areas for new product development or existing product improvement.

3.6 Line Chart:

The line chart provided illustrates the total sales over time for a given period. Here's an expanded analysis based on the visual information:



- The chart displays sales data from January to April 2019.

- There is considerable variability in total sales from day to day, as indicated by the significant fluctuations in the data points.

- The sales peaks suggest that there are days with particularly high sales volumes, which could correspond to promotional events, weekends, or other specific days when shopping activity is higher.

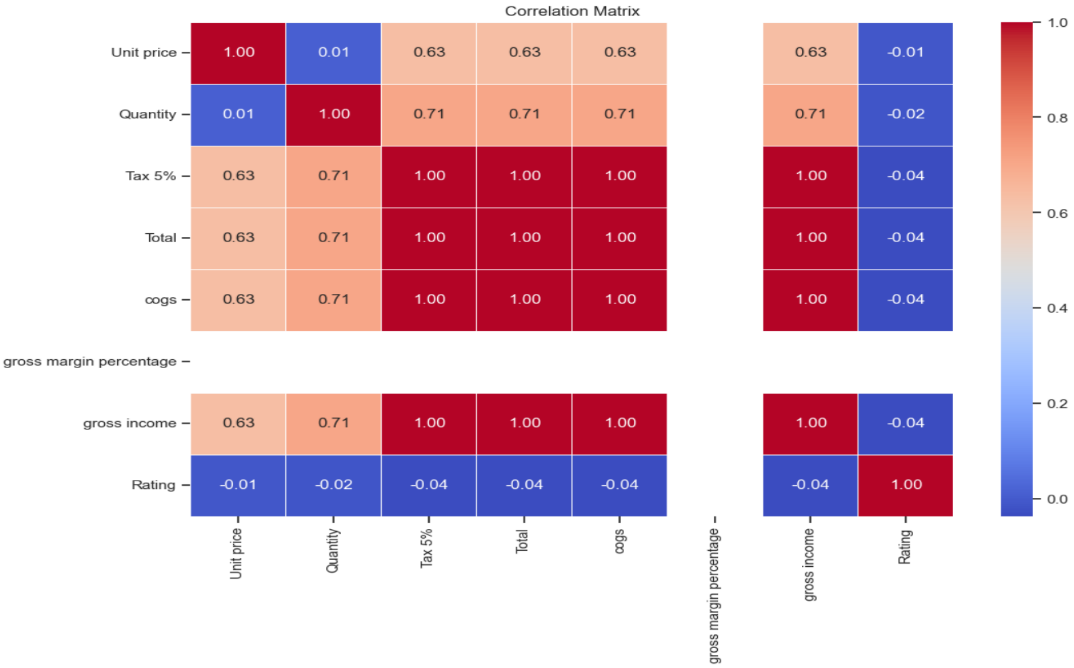
- Conversely, the troughs indicate days with lower sales volumes, which might align with weekdays or off-peak periods.

- There doesn't appear to be a clear upward or downward trend over the period shown, which suggests that sales are relatively stable month-to-month, without a significant increase or decrease in overall sales.

Some specific points or periods show unusually high or low sales, marked by sharp peaks or drops. These outliers could be due to special events, stock issues, market changes, or other factors affecting sales.

Understanding these patterns can help in inventory planning, staffing, and marketing campaigns. For example, preparing for peak sales periods may involve increasing stock levels or staffing, while trough periods could be used for training or maintenance activities. Additionally, identifying the causes of outliers could provide insights into successful sales strategies or potential issues that need to be addressed.

3.7 Correlation Matrix:



The correlation matrix shows correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The value is in the range of -1 to 1. If two variables have a high correlation (close to 1 or -1), it means that when one variable increases, the other variable tends to also increase (positive correlation) or decrease (negative correlation). If the correlation is close to 0, it indicates no linear relationship between the variables.

Unit Price and Quantity: There is virtually no correlation between unit price and quantity, suggesting that the price of an item does not significantly affect how many units are sold.

Unit Price and Tax 5%: There is a moderate positive correlation, indicating that higher-priced items tend to have higher tax values, which is expected since tax is usually a percentage of the price.

Unit Price and Total Sales: There is also a moderate positive correlation, suggesting that items with higher unit prices contribute to higher total sales, which could be due to a higher revenue per item sold.

Quantity and Total Sales, Tax 5%, COGS (Cost of Goods Sold), and Gross Income: All these have high positive correlations with quantity, indicating that as more items are sold, these metrics increase proportionally.

Tax 5%, Total Sales, COGS, and Gross Income: They all have perfect correlations with each other (1.00), which suggests they are directly linked. This is expected since they are all derived from the sales of goods.

Gross Margin Percentage: It shows a very slight negative correlation with most variables, indicating that as sales, COGS, and taxes increase, the gross margin percentage tends to decrease slightly, although the relationship is very weak.

Rating: It has a negligible correlation with all the financial metrics, suggesting that customer ratings do not show a linear relationship with sales, tax, COGS, gross income, or unit price in this dataset.

This correlation matrix is useful for identifying relationships between variables, which can help in predictive modeling, risk management, and strategic planning. For instance, if certain products have a strong positive correlation with total sales, the business might focus on promoting these products more heavily. Additionally, the lack of correlation between ratings and financial metrics could suggest that customer satisfaction does not directly translate to higher sales in this case.

These visualizations collectively contribute to a comprehensive understanding of customer shopping trends within the dataset. The diverse range of visual representations enhances interpretability and aids in drawing meaningful insights.

4 Methods

4.1 Statistical Methods Used

In the conducted analysis, five hypothesis tests were employed to explore various relationships and differences within the dataset. Firstly, the Chi-Square test was utilized to investigate associations between categorical variables, specifically examining relationships between branches and cities, customer types and genders, product lines and payment methods, branches, and customer types, as well as cities and payment methods. Secondly, the Kruskal-Wallis test was applied to assess differences in numerical variables across different categories, focusing on total amounts among branches, quantities among product lines, totals among customer types, and quantities among genders. The third method employed was the Paired t-test, utilized to examine paired observations, such as the differences between total amounts and cost of goods sold, unit prices and totals, quantities, and totals, as well as gross income and totals. Additionally, the Independent t-test was conducted to assess differences in means between two independent groups, including total amounts between regular and new customers, total amounts between female and male customers, and total amounts between branches A and B. Finally, an ANOVA test was performed to investigate differences in means across multiple groups, analyzing total amounts among different branches and various product lines. Each test was executed with a significance level of 0.05 to determine the statistical significance of the observed results.

4.2 Hypothesis Testing

Hypothesis testing involves the application of statistical tests to assess relationships and draw inferences about the population based on sample data. The specific tests outlined in the earlier section, including Chi-Square, Kruskal-Wallis, Paired t-test, independent t-test, and ANOVA, will be performed to address predefined hypotheses related to categorical and quantitative variables. Results will be interpreted to determine the significance of observed patterns and relationships.

For Categorical Data

Chi-Square Test

1) Branch and City Association

In our investigation into the association between Branch and City, we formulated the following hypotheses:

- Null Hypothesis (H0): There is no association between Branch and City.

- Alternative Hypothesis (Ha): There is an association between Branch and City.

We applied the Chi-Square test with a significance level of 0.05, examining the columns 'Branch' and 'City'. The test resulted in a p-value of 0.0. Given that this p-value is less than

0.05, we reject the null hypothesis. Consequently, we conclude that there is a statistically significant association between the Branch and City.

2) Customer Type and Gender Association

For the association between Customer Type and Gender, the formulated hypotheses were as follows:

- Null Hypothesis (H0): There is no association between Customer Type and Gender.

- Alternative Hypothesis (Ha): There is an association between Customer Type and Gender.

Applying the Chi-Square test with a significance level of 0.05 to the columns 'Customer Type' and 'Gender', the resulting p-value was 0.2295. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, we find no statistically significant association between Customer Type and Gender.

3) Product Line and Payment Association

Investigating the association between Product Line and Payment, we established the following hypotheses:

- Null Hypothesis (H0): There is no association between Product Line and Payment.

- Alternative Hypothesis (Ha): There is an association between Product Line and Payment.

Conducting the Chi-Square test with a significance level of 0.05 for the columns 'Product Line' and 'Payment', the resulting p-value was 0.5587. As this p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, there is no statistically significant association between Product Line and Payment.

4) Branch and Customer Type Association

For the association between Branch and Customer Type, the formulated hypotheses were as follows:

-Null Hypothesis (H0): There is no association between Branch and Customer Type.

- Alternative Hypothesis (Ha): There is an association between Branch and Customer Type.

Utilizing the Chi-Square test with a significance level of 0.05 and examining the columns 'Branch' and 'Customer Type', the resulting p-value was 0.8111. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Thus, there is no statistically significant association between Branch and Customer Type.

5) City and Payment Association

In our exploration of the association between City and Payment, we formulated the following hypotheses:

- Null Hypothesis (H0): There is no association between City and Payment.

- Alternative Hypothesis (Ha): There is an association between City and Payment.

Applying the Chi-Square test with a significance level of 0.05 to the columns 'City' and 'Payment', the resulting p-value was 0.509. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Consequently, we find no statistically significant association between City and Payment.

Kruskal-Wallis Test

1) Branch and Total:

In our examination of the differences in Total among Branches, the formulated hypotheses were as follows:

- Null Hypothesis (H0): There is no significant difference in Total among Branches.

- Alternative Hypothesis (Ha): There is a significant difference in Total among Branches.

Applying the Kruskal-Wallis test with a significance level of 0.05 to the columns 'Branch' and 'Total', the resulting p-value was 0.8119. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Consequently, we find no statistically significant difference in Total among Branches.

2) Product Line and Quantity:

For the differences in Quantity among Product Lines, the formulated hypotheses were as follows:

- Null Hypothesis (H0): There is no significant difference in Quantity among Product Lines.

- Alternative Hypothesis (Ha): There is a significant difference in Quantity among Product Lines.

Conducting the Kruskal-Wallis test with a significance level of 0.05 for the columns 'Product Line' and 'Quantity', the resulting p-value was 0.327. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, there is no statistically significant difference in Quantity among Product Lines.

3) Customer Type and Total:

In our exploration of the differences in Total among Customer Types, we established the following hypotheses:

- Null Hypothesis (H0): There is no significant difference in Total among Customer Types.

- Alternative Hypothesis (Ha): There is a significant difference in Total among Customer Types.

Utilizing the Kruskal-Wallis test with a significance level of 0.05 for the columns 'Customer Type' and 'Total', the resulting p-value was 0.591. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Thus, there is no statistically significant difference in Total among Customer Types.

4) Gender and Quantity:

For the differences in Quantity among Genders, the formulated hypotheses were as follows:

- Null Hypothesis (H0): There is no significant difference in Quantity among Genders.

- Alternative Hypothesis (Ha): There is a significant difference in Quantity among Genders.

Applying the Kruskal-Wallis test with a significance level of 0.05 to the columns 'Gender' and 'Quantity', the resulting p-value was 0.0195. Since this p-value is less than 0.05, we reject the null hypothesis. Therefore, there is a statistically significant difference in Quantity among Genders.

For Quantitative Data

Paired t-test

1) **Total and COGS:**
   For the investigation into the differences between Total and cogs, the following hypotheses were formulated:
   - Null Hypothesis (H0): There is no significant difference between Total and cogs.
   - Alternative Hypothesis (Ha): There is a significant difference between Total and cogs.
   With a significance level of 0.05, the paired t-test resulted in a p-value of 7.56e-220. Since this p-value is much less than 0.05, we reject the null hypothesis. Therefore, there is a statistically significant difference between Total and cogs.
2) **Unit Price and Total:**
   In examining the differences between Unit price and Total, the hypotheses were as follows:
   - Null Hypothesis (H0): There is no significant difference between Unit price and Total.
   - Alternative Hypothesis (Ha): There is a significant difference between Unit price and Total.
   The paired t-test, with a significance level of 0.05, yielded a p-value of 9.94e-188. Since this p-value is much less than 0.05, we reject the null hypothesis. Consequently, there is a statistically significant difference between Unit price and Total.
3) **Quantity and Total:**
   For the evaluation of differences between Quantity and Total, the formulated hypotheses were:
   - Null Hypothesis (H0): There is no significant difference between Quantity and Total.

- Alternative Hypothesis (Ha): There is a significant difference between Quantity and Total.

Applying the paired t-test with a significance level of 0.05, the resulting p-value was 1.98e-217. Since this p-value is much less than 0.05, we reject the null hypothesis. Therefore, there is a statistically significant difference between Quantity and Total.

4) **Gross Income and Total:**

In the exploration of differences between gross income and Total, the hypotheses were:

- Null Hypothesis (H0): There is no significant difference between gross income and Total.

- Alternative Hypothesis (Ha): There is a significant difference between gross income and Total.

The paired t-test, conducted with a significance level of 0.05, produced a p-value of 7.56e-220. Since this p-value is much less than 0.05, we reject the null hypothesis. Therefore, there is a statistically significant difference between gross income and Total.

Independent t-test

1) Total for Regular Customers and Total for New Customers:

To investigate the potential difference in Total between Regular and New Customers, the following hypotheses were tested:

- Null Hypothesis (H0): There is no significant difference in Total between Regular and New Customers.

- Alternative Hypothesis (Ha): There is a significant difference in Total between Regular and New Customers.

Conducting the independent t-test with a significance level of 0.05 resulted in a p-value of 0.534. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Thus, there is no statistically significant difference in Total between Regular and New Customer.

2) Total for 'Female' and 'Total' for 'Male':

Examining the potential difference in Total between Female and Male customers led to the formulation of the following hypotheses:

- Null Hypothesis (H0): There is no significant difference in Total between Female and Male customers.

- Alternative Hypothesis (Ha): There is a significant difference in Total between Female and Male customers.

The independent t-test, with a significance level of 0.05, yielded a p-value of 0.118. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, there is no statistically significant difference in Total between Female and Male customers.

4) Total for 'Branch A' and 'Total' for 'Branch B':

For the exploration of differences in Total between Branch A and Branch B, the following hypotheses were considered:

- Null Hypothesis (H0): There is no significant difference in Total between Branch A and Branch B.

- Alternative Hypothesis (Ha): There is a significant difference in Total between Branch A and Branch B.

Upon conducting the independent t-test with a significance level of 0.05, the resulting p-value was 0.681. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, there is no statistically significant difference in Total between Branch A and Branch B.

ANOVA Test

1) Total for Branch A, B, C:

To examine if there is a significant difference in Total among Branches (Branch A, Branch B, Branch C), the following hypotheses were formulated:

- Null Hypothesis (H0): There is no significant difference in Total among Branches.

- Alternative Hypothesis (Ha): There is a significant difference in Total among Branches.

The ANOVA test, conducted at a significance level of 0.05, produced a p-value of 0.4132. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Consequently, there is no statistically significant difference in Total among Branches based on the ANOVA test.

2) Total for 'Health and beauty', 'Electronic accessories', 'Home and lifestyle', 'Sports and travel', 'Food and beverages', and 'Fashion accessories':

To investigate if there is a significant difference in Total among Product Lines (Health and beauty, Electronic Accessories, Home and lifestyle, Sports and travel, Food and beverages, and Fashion accessories), the following hypotheses were considered:

- Null Hypothesis (H0): There is no significant difference in Total among Product Lines.

- Alternative Hypothesis (Ha): There is a significant difference in Total among Product Lines.

The ANOVA test, with a significance level of 0.05, resulted in a p-value of 0.8900. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, there is no statistically significant difference in Total among Product Lines based on the ANOVA test.

4.3 Regression Analysis

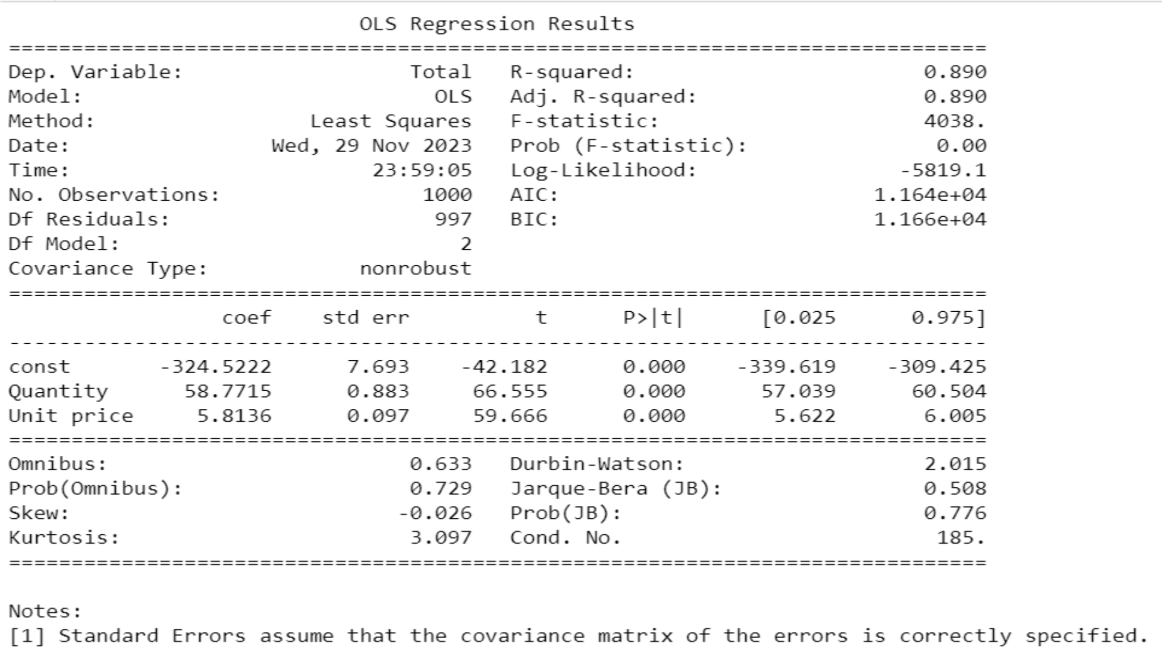
Regression analysis will be conducted to explore the relationships between dependent and independent variables. The focus will be on identifying factors influencing customer behavior and predicting outcomes. The chosen regression models will depend on the nature of the variables under consideration. The results will be analyzed to gain insights into how changes in independent variables relate to changes in the dependent variable.

2.3.4.1 Linear Regression: Total vs. Quantity and Unit Price

Model Summary:

The regression analysis of the model, where the dependent variable is "Total." This implies that the model aims to predict the total sales based on certain independent variables. The high R-squared and adjusted R-squared values of 0.890 suggest a strong fit, with 89% of the variability in Total being explained by the model. This is further reinforced by a very high F-statistic and an associated p-value near zero, indicating that the model's predictive capability is statistically significant.

Both "Quantity" and "Unit Price" are significant predictors with positive coefficients, meaning increases in these variables are associated with increases in total sales. The significance of these predictors is underscored by their p-values of 0.000, further solidifying their importance in the model. The robustness of the model is also indicated by



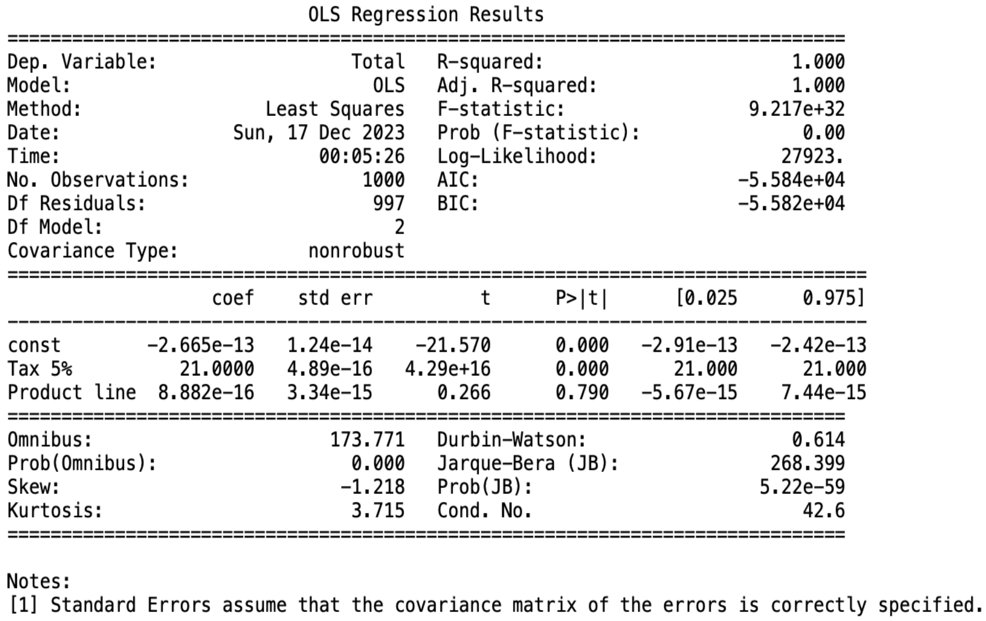
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Total   R-squared:                       0.890
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     4038.
Date:                Wed, 29 Nov 2023   Prob (F-statistic):               0.00
Time:                        23:59:05   Log-Likelihood:                -5819.1
No. Observations:                1000   AIC:                         1.164e+04
Df Residuals:                     997   BIC:                         1.166e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -324.5222      7.693    -42.182      0.000    -339.619    -309.425
Quantity      58.7715      0.883     66.555      0.000      57.039      60.504
Unit price     5.8136      0.097     59.666      0.000       5.622       6.005
==============================================================================
Omnibus:                        0.633   Durbin-Watson:                   2.015
Prob(Omnibus):                  0.729   Jarque-Bera (JB):                0.508
Skew:                          -0.026   Prob(JB):                        0.776
Kurtosis:                       3.097   Cond. No.                         185.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

the no robust covariance type, which suggests that the standard errors have been calculated assuming that the covariance matrix of the errors is correctly specified.

2.3.4.2 Linear Regression: Total vs. Tax 5% and Product Line

Model Summary



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Total   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 9.217e+32
Date:                Sun, 17 Dec 2023   Prob (F-statistic):               0.00
Time:                        00:05:26   Log-Likelihood:                 27923.
No. Observations:                1000   AIC:                        -5.584e+04
Df Residuals:                     997   BIC:                        -5.582e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -2.665e-13   1.24e-14    -21.570      0.000   -2.91e-13   -2.42e-13
Tax 5%          21.0000   4.89e-16   4.29e+16      0.000      21.000      21.000
Product line  8.882e-16   3.34e-15      0.266      0.790   -5.67e-15    7.44e-15
==============================================================================
Omnibus:                      173.771   Durbin-Watson:                   0.614
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              268.399
Skew:                          -1.218   Prob(JB):                     5.22e-59
Kurtosis:                       3.715   Cond. No.                         42.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The regression output centers on the dependent variable "Total," and it shows a model with R-squared and adjusted R-squared values at the maximum of 1.000. While this would typically indicate a perfect fit of the model to the data, such perfect values are exceedingly rare in practice and often signal potential issues such as overfitting, data leakage, or some form of error in the data or the model itself.

The F-statistic is exceptionally large, with a probability of 0.00, reinforcing the model's statistical significance. In terms of individual predictors, the variable "Tax 5%" has a positive and significant coefficient, which points to a strong positive influence on total sales when the tax rate is at 5%. On the contrary, "Product Line" has a p-value of 0.790, making it an insignificant predictor within this model, which suggests it does not have a meaningful association with total sales. This could mean that "Product Line" does not affect the "Total" variable, or it might not have been captured effectively in this particular model.

Lasso Regression

Experiment 1 (Total vs. Quantity and Unit Price)

In our regression analysis, the Mean Squared Error (MSE) was found to be 6228.25, which serves as a metric for the average of the squares of the errors. Specifically, it represents the average squared difference between the observed values and those predicted by the model. The relatively low MSE in our model indicates a close fit between our model's predictions and the actual data.

The R-squared (R2) value achieved in this model is 0.904, which explains approximately 90.4% of the variance within the dependent variable. This high R2 value suggests that the model has substantial explanatory power and is capable of accounting for a large proportion of the observed variability in the dataset.

The analysis has designated 'Quantity' and 'Unit price' as the selected features, underscoring their importance as significant predictors in the model. This selection was made after the application of a Lasso regression, which effectively penalizes the inclusion of less significant variables, thereby streamlining the model to focus on those factors that offer the most predictive power.

Moreover, the model has determined an optimal alpha value of 0.01. Alpha is a parameter in Lasso regression that regulates the strength of the penalty applied to the model's complexity. An optimal alpha of 0.01 indicates that this level of regularization strikes the best balance between the model's complexity and its performance, ensuring that the model remains robust without overfitting the data.

Experiment 2 (Total vs. Tax 5% and Product Line)

The regression model has yielded an exceedingly low Mean Squared Error (MSE) of approximately 0.000110549, suggesting that the predicted values are virtually identical to the actual data points. While such a precise MSE might typically be cause for celebration, it warrants a cautious approach in this context. The presence of such a minuscule error rate raises the possibility of overfitting, especially in light of the near-perfect R-squared (R2) value of approximately 1.0000000.

An R2 value that is essentially 1 denotes that the model explains virtually all the variance in the dependent variable, leaving almost no unexplained variance. This might imply an impeccable fit of the model to the data at hand; however, it is an unusual circumstance in real-world data analysis and can indicate that the model may be excessively tailored to this specific dataset, potentially at the expense of its ability to generalize to other datasets.

The model has identified 'Tax 5%' as the sole significant feature, which implies that this predictor alone accounts for almost all the variability in the outcome, aligning with the extraordinarily high R2 value. Although this may seem beneficial, such a result should be interpreted with a degree of skepticism, considering the potential for overfitting.

Furthermore, the model has consistently identified an optimal alpha value of 0.01 across different runs, which suggests that this level of regularization is suitable for this problem when employing Lasso regression. This stability in the alpha value might indicate that it is a robust parameter choice for the current modeling scenario, providing a balance between simplicity and fit.

Linear Regression vs Lasso Regression (Total vs. Quantity and Unit Price**)**

In our analytical endeavors, Lasso regression has proven to be a valuable method, yielding a slight increase in the R-squared value, which suggests an enhanced fit to our dataset. Lasso's innate ability to perform feature selection is one of its most salient characteristics. It accomplishes this by reducing certain coefficients to zero, thereby simplifying the model and eliminating less significant variables. This aspect of Lasso regression ensures that only the most relevant predictors are retained, enhancing the model's interpretability and predictive power.

The process of hyperparameter tuning is crucial within the framework of Lasso regression. The selection of the optimal alpha is particularly critical, as it dictates the level of regularization imposed on the model. It is this regularization that helps to prevent overfitting, ensuring that the model remains generalizable to new data. An optimal alpha of 0.01 has been identified as particularly effective, striking a balance that promotes the model's predictive performance while concurrently facilitating feature selection.

On the other hand, ordinary least squares (OLS) regression is preferable when the primary concern is the interpretability of coefficients without the influence of regularization. OLS provides a more direct interpretation of variable relationships, which can be invaluable in scenarios where understanding the impact of each predictor is essential.

When deciding between Lasso and OLS regression, it is imperative to weigh the specific objectives of the analysis against the trade-offs between predictive accuracy and interpretability. If the goal is to achieve a model that not only predicts well but also identifies the most impactful features without overcomplicating the model, Lasso regression, particularly with an optimally tuned alpha, is the superior choice.

Linear Regression vs Lasso Regression (Total vs. Tax 5% and Product Line**)**

The analysis utilizing both linear and Lasso regression models reveals an exceptional fit to the dataset, as evidenced by the high R-squared values obtained. The decision to employ either model should be informed by the analytical context and the specific objectives of the research.

Lasso regression, with its optimal alpha set at 0.01, inherently conducts feature selection and, in this instance, has ascribed significance solely to the 'Tax 5%' variable by allocating it a non-zero coefficient. This effectively positions 'Tax 5%' as the predominant factor influencing the dependent variable. It underscores the utility of Lasso when the goal is to distill the predictive elements to the most impactful variables.

In contrast, linear regression presents what appears to be a perfect model fit, with 'Tax 5%' and 'Product line' yielding coefficients approaching zero. This suggests that while these features do have a role, their influence is minimal, providing a more traditional understanding of how each variable contributes to the outcome.

Both models' high R-squared and low Mean Squared Error (MSE) values support their strong alignment with the observed data. The feature selection capability of Lasso becomes particularly pertinent when pinpointing key predictors is essential, whereas linear regression offers a clear, conventional elucidation of the coefficients.

Nevertheless, given the unusually high fit of the models, further scrutiny is advisable to ascertain the behavior of the 'Product line' variable and its true effect on the dependent variable. Additionally, the implications of such a high level of fit should be critically evaluated to rule out overfitting or anomalies within the dataset itself. This due diligence is crucial in ensuring the reliability and applicability of the model to broader contexts.

5. Conclusions:

The conclusions of our analysis reveal several key insights into the dataset and the effectiveness of various regression models. Notably, Lasso regression demonstrates a marginally better fit compared to ordinary least squares (OLS) regression, as evidenced by the improved R-squared value. This superior fit is indicative of Lasso regression's robustness, particularly in handling high-dimensional data. The method's inherent feature selection ability is advantageous for simplifying complex models, making it an essential tool for enhancing predictive accuracy in datasets with numerous variables.

Our study also sheds light on the significant role of pricing and quantity in influencing sales performance. The positive coefficients for unit price and quantity within the regression models suggest that as these variables increase, so too does the total sales volume. This finding underscores the importance of strategic pricing and the potential benefits of quantity-based promotions or discounts in driving sales growth.

Furthermore, the analysis highlights the notable influence of tax on sales. There is a strong positive relationship between the amount of tax applied and the total sales, suggesting that items with higher tax, which are typically higher-priced, contribute significantly to the overall sales figures. This correlation can be invaluable for understanding consumer purchasing patterns to taxed versus untaxed goods. Retailers and policymakers can leverage this information to refine their sales strategies and tax policies, potentially influencing consumer behavior to optimize sales outcomes.

In conclusion, the application of Lasso regression has provided a nuanced understanding of the factors contributing to sales. The insights gained from this analysis can inform strategic decisions in pricing, promotions, and taxation that align with consumer behavior and market dynamics, ultimately aiding in the pursuit of enhanced sales performance.