ANALYZING CUSTOMER SHOPPING TRENDS USING STATISTICAL METHODS

By

Hemanth Dadi

E Navaneet Kumar

Jayanth Kumar Yanamandala

INTRODUCTION AND OVERVIEW

- In the era of data-driven decision-making, analyzing customer shopping trends using statistical methods is paramount for businesses.
- The presentation aims to provide a comprehensive overview of sales data from a retail perspective.
- The dataset, derived from supermarket sales records, encompasses various factors, including unit price, quantity, tax, branch location, customer type, and customer-related metrics more.
- Understanding the intricacies of this dataset allows for valuable insights into sales patterns, customer behavior, and the factors influencing overall revenue.
- **Key Objectives :**
  - 1. Pricing and Quantity Impact: How do unit price and quantity interact to influence total sales?
  - 2. Tax Influence on Sales: What is the relationship between tax and total sales, particularly for higher-priced items?
- Through an in-depth exploration of these objectives, we aim to uncover key insights that can inform strategic decision-making in the retail landscape.

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.

.

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
233-67-5758	C	Naypyitaw	Normal	Male	Health and beauty	40.35	1	2.0175	42.3675	1/29/2019	13:46	Ewallet	40.35	4.761905	2.0175	6.2
303-96-2227	B	Mandalay	Normal	Female	Home and lifestyle	97.38	10	48.6900	1022.4900	3/2/2019	17:16	Ewallet	973.80	4.761905	48.6900	4.4
727-02-1313	A	Yangon	Member	Male	Food and beverages	31.84	1	1.5920	33.4320	2/9/2019	13:22	Cash	31.84	4.761905	1.5920	7.7

Dataset Description

The dataset consists of transactions recorded with various attributes. Key attributes include: - Invoice ID: Unique identifier for each transaction.

- Branch: Location of the transaction.
- City: City where the branch is situated.
- Customer Type: Indicates if the customer is a regular or new customer. - Gender: Gender of the customer.
- Product Line: Category or type of product purchased Unit Price: Price of a single unit of the product.
- Quantity: Number of units of the product purchased.
- Tax 5%: Tax applied to the transaction (5% of the total cost). - Total: Total cost of the transaction, including tax.
- Date: Date of the transaction.
- Time: Time of day when the transaction occurred.
- Payment: Payment method used (e.g., credit card, cash).
- COGS (Cost of Goods Sold): Direct costs associated with producing or purchasing the products sold.
- Gross Margin Percentage: Profit margin percentage for the transaction.
- Gross Income: Total profit earned from the transaction.
- Rating: Customer satisfaction rating or feedback.

Statistical Methods Used

The analysis employed various statistical methods, including:

- **Descriptive Statistics**: To summarize and describe the main features of the dataset.
- **Hypothesis Testing**: To assess relationships and make inferences about the population based on sample data.
- **Regression Analysis**: To explore the relationships between dependent and independent variables.

Descriptive Statistics

Summary Statistics:

	Unit price	Quantity	Tax 5%	Total	cogs
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738
std	26.494628	2.923431	11.708825	245.885335	234.17651
min	10.080000	1.000000	0.508500	10.678500	10.17000
25%	32.875000	3.000000	5.924875	124.422375	118.49750
50%	55.230000	5.000000	12.088000	253.848000	241.76000
75%	77.935000	8.000000	22.445250	471.350250	448.90500
max	99.960000	10.000000	49.650000	1042.650000	993.00000

	gross margin percentage	gross income	Rating
count	1.000000e+03	1000.000000	1000.00000
mean	4.761905e+00	15.379369	6.97270
std	6.131498e-14	11.708825	1.71858
min	4.761905e+00	0.508500	4.00000
25%	4.761905e+00	5.924875	5.50000
50%	4.761905e+00	12.088000	7.00000
75%	4.761905e+00	22.445250	8.50000
max	4.761905e+00	49.650000	10.00000

- Descriptive statistics provide a summary of key metrics to better understand the dataset. This includes measures such as mean, median, standard deviation, minimum, and maximum values for relevant variables.
- The analysis will focus on extracting insights into the central tendencies and variability in the data, contributing to a comprehensive overview.

Visualizations

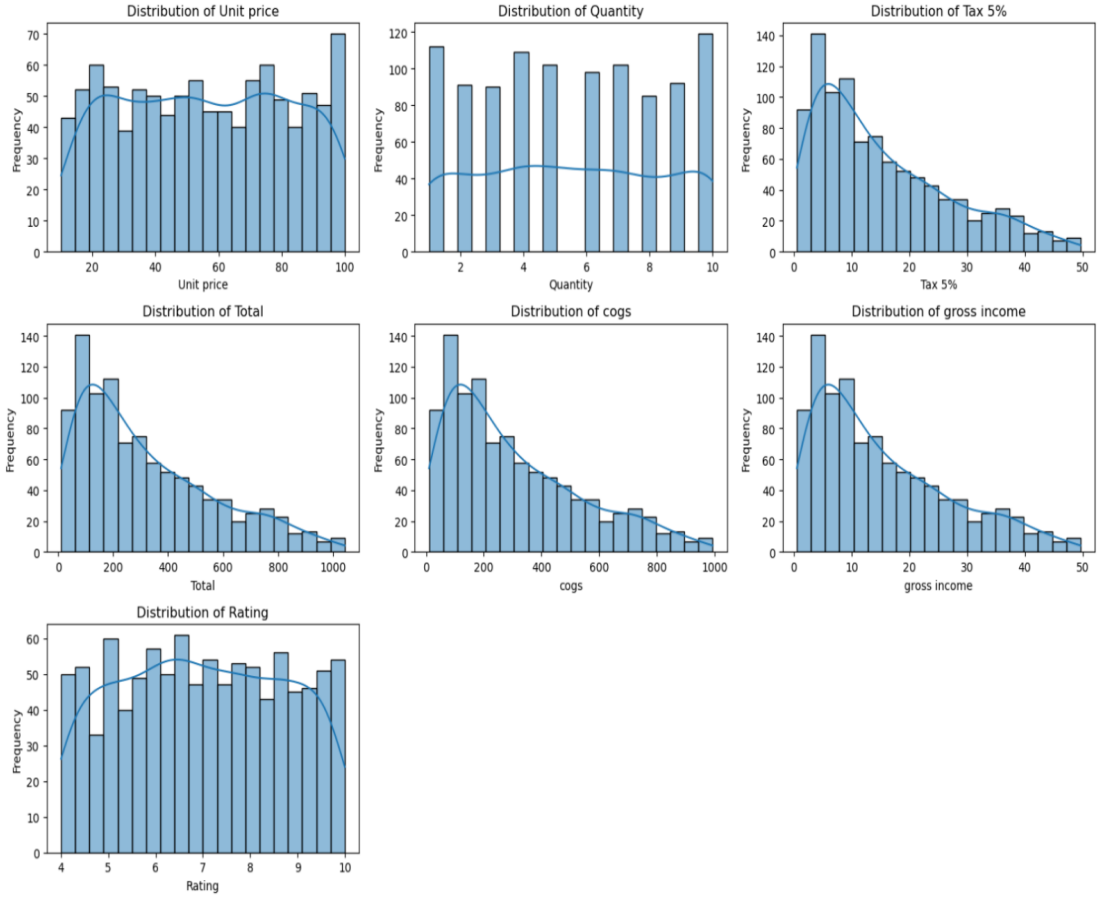
Histograms : To illustrate the distribution of various quantitative variables.

- **Distribution of Unit Price:**
  - The distribution appears multimodal, with several peaks, which could indicate common price points at which items are sold.
- **Distribution of Quantity:**
  - The quantity shows a uniform distribution across the bins, suggesting no specific quantity preference in orders.
- **Distribution of Tax 5%:**
  - The right-skewed distribution indicates that most transactions have a lower tax value, with fewer transactions having a higher tax.

Distribution of Total:

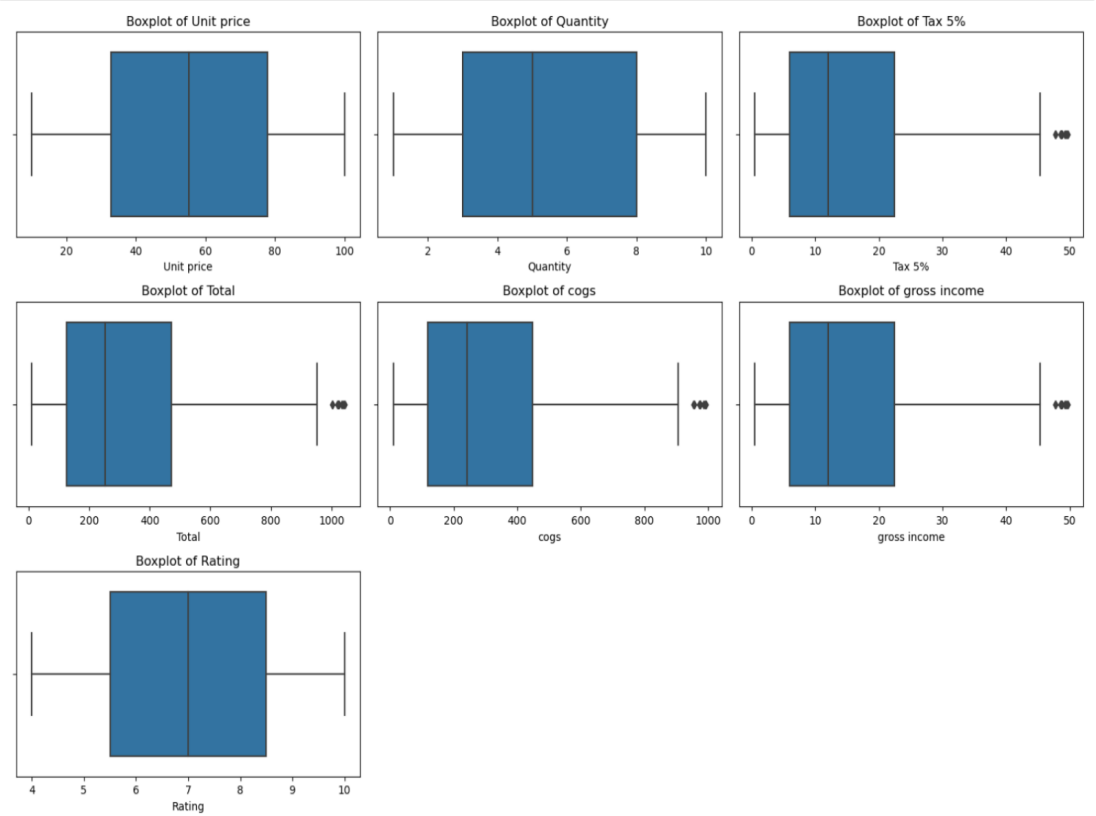
- A right-skewed distribution for the total value of transactions is observed, indicating that higher total sales are less frequent.

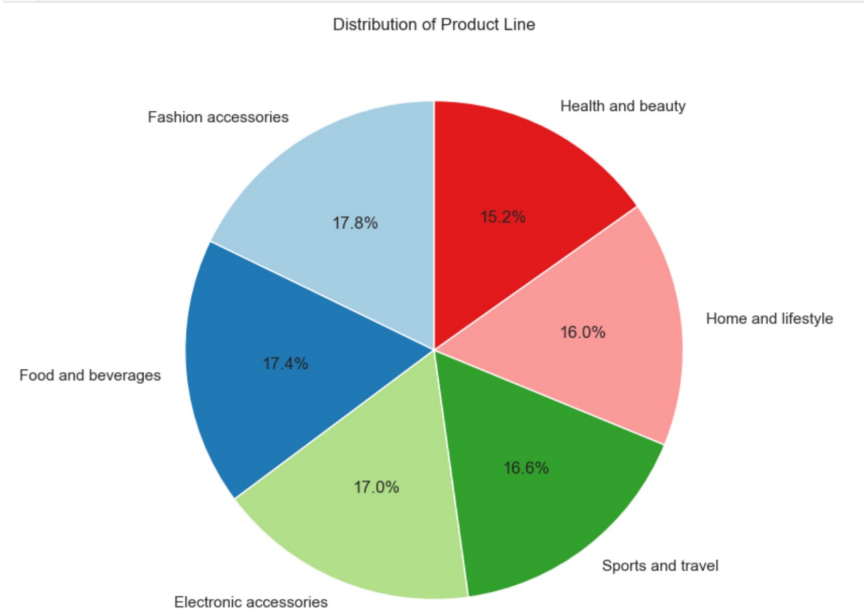
- **Distribution of COGS (Cost of Goods Sold):**
  - The distribution mirrors that of the total sales, as expected, since COGS and total sales are typically correlated.
- **Distribution of Gross Income:**
  - Gross income also shows a right-skewed distribution, with most transactions resulting in a lower gross income.
- **Distribution of Rating:**
  - Customer ratings are concentrated between 6 and 10, suggesting generally positive feedback.



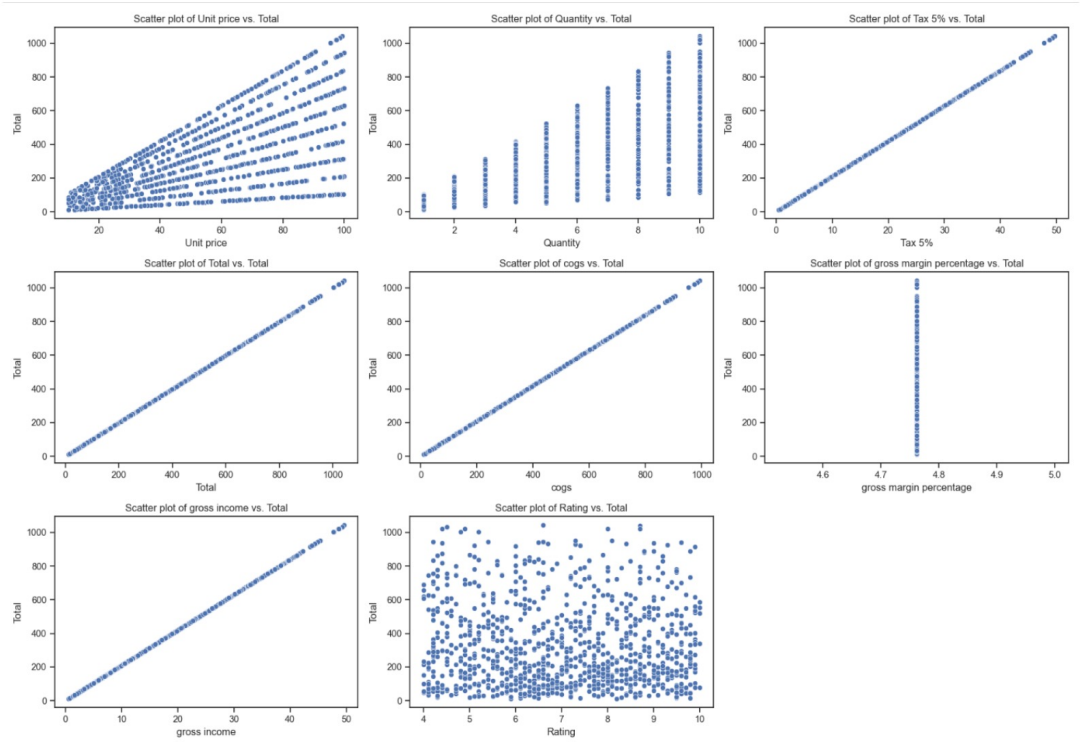
Boxplots : To provide a concise summary of the distribution, highlighting key statistical measures and identifying potential outliers.

- **Boxplot of Unit Price:**
  - The median unit price is centrally located, indicating a symmetrical distribution of prices.
  - There are no outliers, suggesting that all unit prices fall within a typical range without extreme values.
- **Boxplot of Quantity:**
  - The median quantity is around the middle of the range, indicating a balanced distribution of purchase quantities.
  - There are no visible outliers, which implies consistent ordering behavior among customers.
- **Boxplot of Tax 5%:**
  - The tax amount has a median that skews towards the lower end, with a few outliers indicating some transactions with significantly higher tax values.
- **Boxplot of Total:**
  - The total transaction amounts are right-skewed, with the median below the average, and there are a few high-value outliers, suggesting occasional larger purchases.
- **Boxplot of COGS (Cost of Goods Sold):**
  - The distribution of COGS closely mirrors that of the total sales, indicating a direct relationship between the cost to sell goods and the sales amount, with a few outliers on the higher end.
- **Boxplot of Gross Income:**
  - Gross income also displays a right-skewed distribution, with a median that indicates most transactions result in a moderate amount of income, and outliers suggest some transactions are particularly profitable.
- **Boxplot of Rating:**
  - The ratings are relatively evenly distributed across the range, with the median close to the upper quartile, suggesting customers tend to give higher ratings.



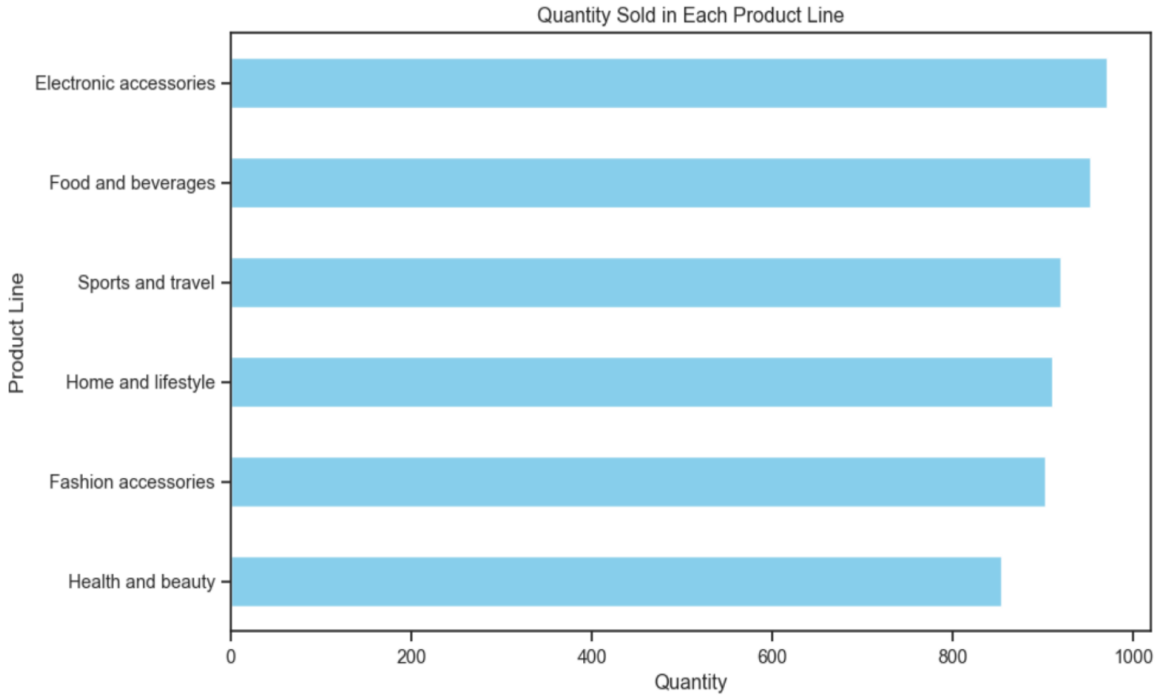


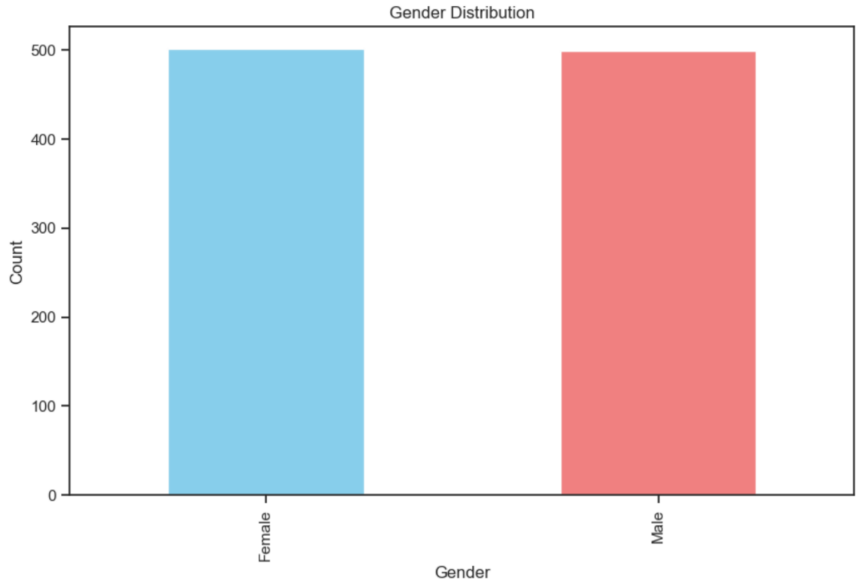
Pie Chart : To visualize the distribution of product lines.

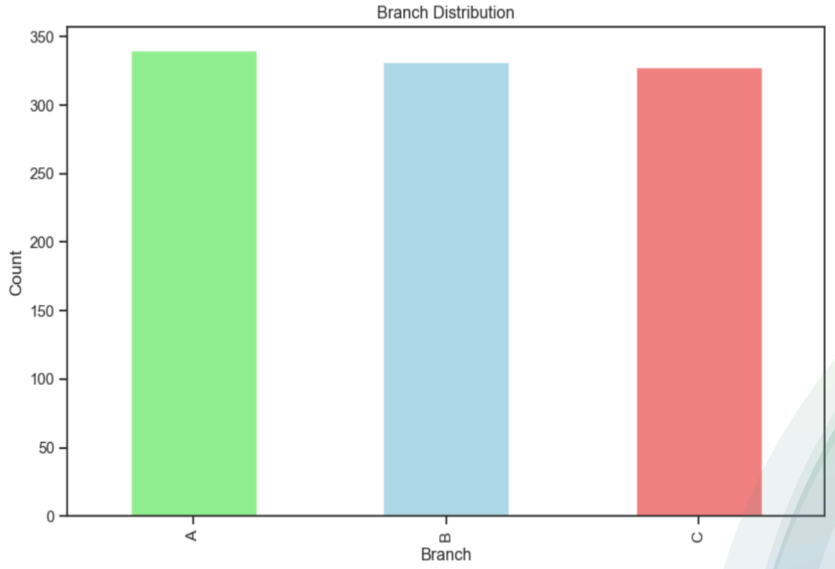


Scatter plots : To visualize relationships. between two quantitative variables.

Bar Charts







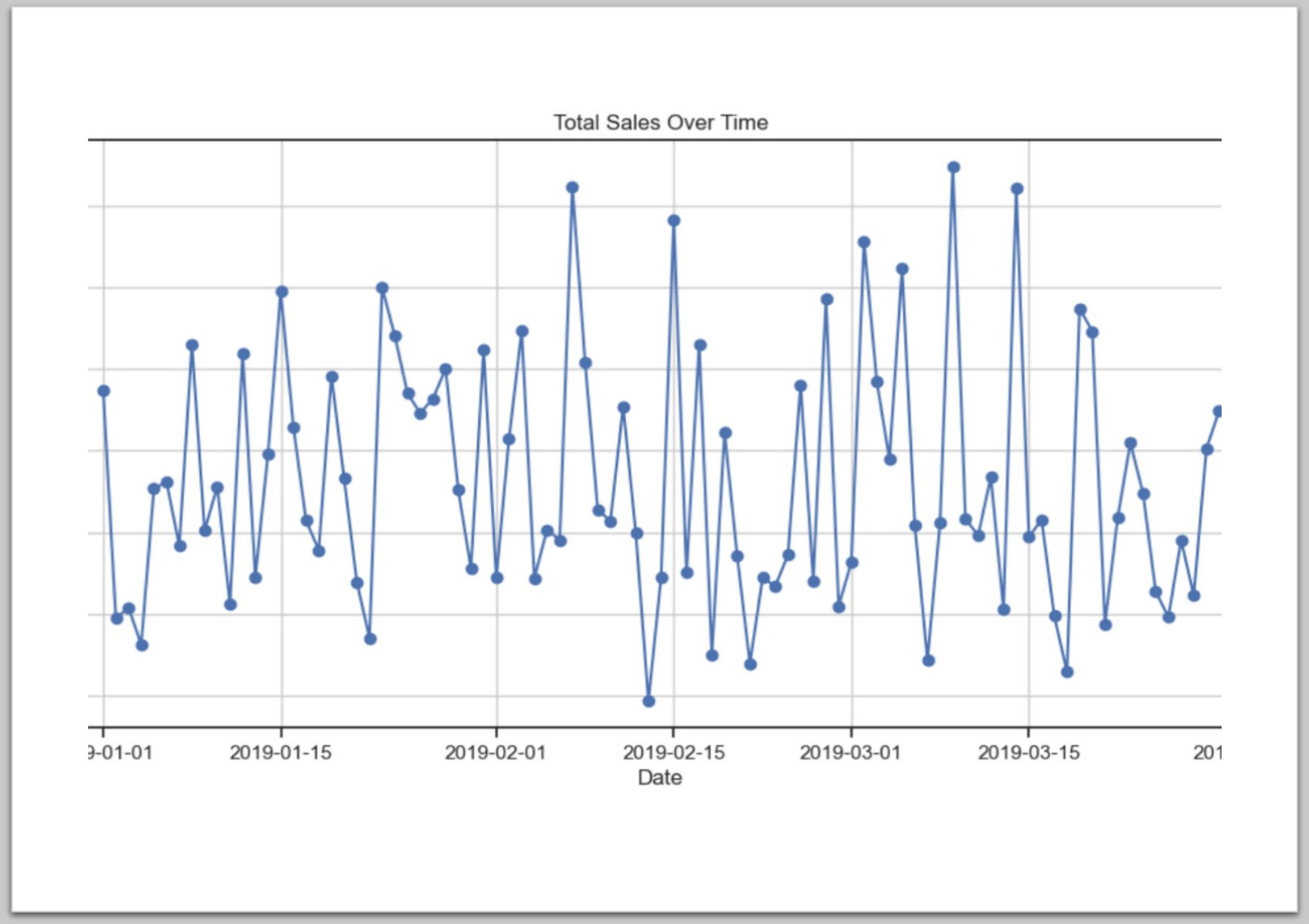
Bar Charts :

To visualize the distribution of categorical variables.

Line Chart

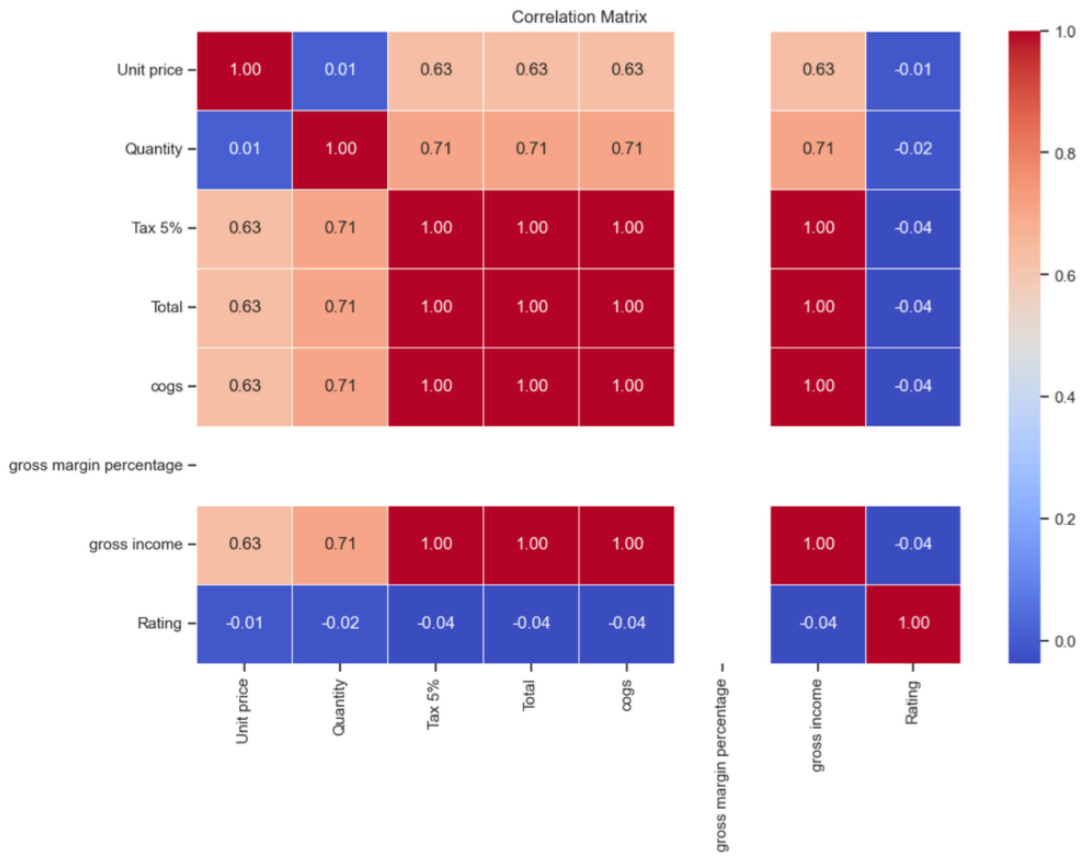
A line chart is utilized to visualize sales trends over time.

Total Sales Over Time- Captures the trend in sales over the specified time period.



A correlation matrix is employed to visualize the relationships between quantitative variables.

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.



Correlation Matrix

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

- **Branch and City:**
  - Objective: Explore the relationship between the branch location and the city.
    - Chi-Square Value: 2000.0
    - P-value: 0.0
- **Customer Type and Gender:**
  - Objective: Investigate if there is an association between customer type and gender. Test Results
    - Chi-Square Value: 1.4437
    - P-value: 0.2295

Product Line and Payment:

- Objective: Examine the relationship between the product line and the payment method.
  - Chi-Square Value: 8.7214
  - P-value: 0.5587
- **Branch and Customer Type:**
  - Objective: Assess if the branch location is associated with the customer type.
    - Chi-Square Value: 0.4188
    - P-value: 0.8111
- **City and Payment:**
  - Objective: Determine if there is a significant relationship between the city and the payment method.
    - Chi-Square Value: 3.2997
    - P-value: 0.5090

Kruskal-Wallis Test

- Branch and Total:
  - Objective: Evaluate if there are significant differences in transaction totals among different branches.
    - Kruskal-Wallis Value: 0.4168
    - P-value: 0.8119
- **Product Line and Quantity:**
  - Objective: Investigate if product lines impact the quantity of items purchased.
    - Kruskal-Wallis Value: 5.7923
    - P-value: 0.3270
- **Customer Type and Total:**
  - Objective: Assess if there are differences in transaction totals between customer types.
    - Kruskal-Wallis Value: 0.2887
    - P-value: 0.5911
- **Gender and Quantity:**
  - Objective: Examine if there are variations in the quantity of items purchased based on gender.
    - Kruskal-Wallis Value: 5.4544
    - P-value: 0.0195

ANOVA Test

- **Total for Branch A, B, C:**
  - Objective: Explore if there are significant differences in transaction totals among different branches.
    - F-statistic: 0.8846
    - P-value: 0.4132
- **Total for 'Health and beauty', 'Electronic accessories', 'Home and lifestyle', 'Sports and travel', 'Food and beverages', 'Fashion accessories':**
  - Objective: Assess if there are significant variations in transaction totals among different product lines.
    - F-statistic: 0.3380
    - P-value: 0.8900

For Quantitative Data - Paired t-test

- **Total and COGS:**
  - Objective: Evaluate if there is a significant difference between total cost and cost of goods sold.
    - T-statistic: 41.5360
    - P-value: 7.5619e-220
- **Unit Price and Total:**
  - Objective: Assess if there is a significant relationship between unit price and total transaction cost.
    - T-statistic: 41.5360
    - P-value: 7.5619e-220
- **Quantity and Total:**
  - Objective: Investigate if there is a significant correlation between quantity and total cost.
    - T-statistic: -41.1713
    - P-value: 1.9755e-217
- **Gross Income and Total:**
  - Objective: Examine if there is a significant association between gross income and total transaction cost.
    - T-statistic: -41.5360
    - P-value: 7.5619e-220

Independent t-test

- **Total for Regular Customers and Total for New Customers:**
  - Objective: Determine if there is a significant difference in total transaction amounts between regular and new customers.
    - T-statistic: 0.6215
    - P-value: 0.5344
- **Total for 'Member' and 'Total' for 'Normal':**
  - Objective: Investigate if there is a significant difference in total transaction amounts between member and non-member customers.
    - T-statistic: 0.6215
    - P-value: 0.5344
- **Total for 'Female' and 'Total' for 'Male':**
  - Objective: Assess if there is a significant difference in total transaction amounts between female and male customers.
    - T-statistic: 1.5641
    - P-value: 0.1181
- **Total for 'Branch A' and 'Total' for 'Branch B':**
  - Objective: Examine if there is a significant difference in total transaction amounts between different branches.
    - T-statistic: -0.4111
    - P-value: 0.6811

Regression Analysis

- In our project, we employed both linear regression and lasso regression techniques in two separate experiments. Linear regression allowed us to model the relationship between variables, while lasso regression provided a means of feature selection, aiding in building a more parsimonious and interpretable model.
  - **Experiment 1**
    - Independent variables(features): **“Quantity, Unit price”**
    - Target variable: “**Total**”
  - **Experiment 2**
    - Independent variables(features): **“Tax 5%, Product line”**
    - Target variable: “**Total**”

Linear Regression Experiment 1

- The dependent variable (the one being predicted) is "Total."
- The model has an R-squared and adjusted R-squared of 0.890, indicating that approximately 89% of the variability in the total sales can be explained by the model's inputs.
- The F-statistic is very high, leading to a probability close to 0, which indicates the model is statistically significant.
- The coefficients for "Quantity" and "Unit Price" are both positive, suggesting that as these increase, the total sales also increase.
- The p-values for both "Quantity" and "Unit Price" are 0.000, indicating these variables are statistically significant predictors of the total sales.

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.

Linear Regression Experiment 2

- The dependent variable is "Total."
- This model has perfect R-squared and adjusted R-squared values of 1.000, which would usually indicate that the model perfectly fits the data. However, such a result is highly unusual and might suggest overfitting, data leakage, or an error in the model.
- The F-statistic is again extremely high with a probability of 0.00, indicating a statistically significant model.
- The coefficient for "Tax 5%" is positive and significant, implying a strong relationship with total sales. However, the coefficient for "Product Line" is not statistically significant (p-value: 0.790), suggesting it does not have a predictive relationship with the total sales in this model.

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.

Lasso Regression

- **Experiment 1**
  - Mean Squared Error (MSE): 6228.245960239633
  - R-squared (R2): 0.9042677361570542
  - Selected Features: Index(['Quantity', 'Unit price'], dtype='object’)
  - Optimal Alpha: 0.01
- **Experiment 2**
  - Mean Squared Error (MSE): 0.00011054938912303335
  - R-squared (R2): 0.9999999983007827
  - Selected Features: Index(['Tax 5%'], dtype='object’)
  - Optimal Alpha: 0.01

Linear Regression vs Lasso Regression (Expt 1)

- Lasso regression has been instrumental in our analysis, demonstrating a slight improvement in the R-squared value, indicating a better fit to the data.
- One of the distinctive features of Lasso regression is its automatic feature selection capability. By strategically shrinking coefficients, Lasso effectively sets some coefficients to zero, streamlining the model.
- Hyperparameter tuning plays a pivotal role in Lasso regression. The optimal alpha, carefully selected, serves as a control for the amount of regularization applied to the model.
- For those aiming at enhanced predictive performance and automatic feature selection, Lasso regression with an optimal alpha of 0.01 emerges as a favorable choice.
- Conversely, linear regression (OLS) becomes advantageous when the focus is on interpreting coefficients without regularization. This method offers a clear and straightforward interpretation of the relationships between variables.
- In the decision between the two models, consider your specific goals and the trade-off between predictive accuracy and interpretability. If feature selection and regularization are priorities, Lasso regression stands out as a more suitable option.

Linear Regression vs Lasso Regression (Expt 2)

- Both models exhibit exceptional fit to the data, as indicated by the high R-squared values in linear and Lasso regression. The choice between the two models depends on the context of your analysis and specific goals.
- With an optimal alpha of 0.01, Lasso automatically performs feature selection, assigning a non-zero coefficient only to 'Tax 5%.' This identifies 'Tax 5%' as the primary contributor to the dependent variable.
- Linear Regression suggests a perfect fit with coefficients close to zero for both 'Tax 5%' and 'Product line.' While both features contribute to the dependent variable, their coefficients are minimal.
- The high R-squared values and low Mean Squared Error (MSE) in both models affirm a strong fit to the data. Lasso's feature selection is advantageous if crucial, while linear regression provides a traditional interpretation of coefficients.
- However, further investigation is recommended to understand the nature of the 'Product line' feature and its potential impact on the dependent variable. Consider the practical implications of such high model fit, raising questions about overfitting or potential dataset issues.

Conclusions

- **Superior Fit with Lasso**: Lasso regression outperforms linear regression (OLS) with a slight R-squared improvement, showcasing its superior fit to the data.
- **Efficient Handling of High-Dimensional Data:** Lasso's automatic feature selection streamlines the model for efficient management of high-dimensional datasets, enhancing predictive performance.
- **Pricing and Quantity Impact** : The analysis indicates that both unit price and quantity have a significant positive impact on sales performance. As the unit price and quantity sold increase, there is a corresponding increase in total sales, suggesting that careful pricing strategies coupled with quantity discounts or promotions can effectively drive sales.
- **Tax Influence on Sales** : The strong positive correlation between tax and total sales suggests that higher-priced items, which incur more tax, contribute substantially to total sales. This relationship may be leveraged to understand consumer behavior towards taxed and untaxed products and adjust strategies accordingly.


THANK YOU AND
HAPPY HOLIDAYS