CUSTOMER LIFETIME VALUE PREDICTION

MA 541-A

Group Number – 11

Presented By-

1] Mrunal Madhukar Gavit
2] Srushti Anil Kamble
3] Akshay Kamlakar Parate
4] Pradnya Sham Jagtap

Submitted to-
Prof. Hong Do

STEVENS INSTITUTE OF TECHNOLOGY

TABLE OF CONTENT

I. INTRODUCTION

Understanding the value that customers bring to a business over their entire relationship with the company is paramount for sustainable growth and profitability. The Customer Lifetime Value (CLV) concept provides a framework to quantify this value, enabling businesses to make informed decisions about resource allocation, marketing strategies, and customer retention efforts. In this project, we delve into predicting CLV using a dataset obtained from retail transactions, employing a variety of analytical techniques and machine learning models.

The dataset used in this project consists of retail transaction records, capturing information such as purchase quantities, unit prices, invoice dates, and customer locations. These transactions span across different countries and offer a rich source of information for understanding customer behavior and purchase patterns. By exploring this data, we aim to uncover insights that can help businesses optimize their marketing efforts, improve customer segmentation, and ultimately enhance revenue generation.

In today's competitive business landscape, it's not merely enough to acquire customers; retaining them and maximizing their lifetime value is crucial for sustained success. Therefore, accurately predicting CLV becomes a strategic imperative for businesses across industries. By leveraging advanced analytics and machine learning, we aim to provide businesses with actionable insights into predicting the future value of their customers, thereby enabling them to tailor their strategies for maximum effectiveness.

Throughout this project, we address several key questions:

1. How can we leverage historical transaction data to predict the future value of customers?
2. What is the distribution of sales across different countries?
3. Is there any seasonal pattern or trend in sales over time?
4. What is the typical range of unit prices for the products? Are there any outliers?
5. Which customers are most likely to make high value purchase?

Through our data analyses and predictive modeling efforts, we aim to provide actionable insights and recommendations for businesses looking to optimize their customer relationship management strategies. Our conclusions shed light on the importance of understanding customer behavior, the impact of various factors on CLV prediction, and the efficacy of different machine learning algorithms in this context. The remainder of this paper will delve deeper into each aspect of our analysis, presenting methodologies, results, and implications for businesses seeking to enhance their customer lifetime value prediction capabilities.

II. DATA

1] Data Sourcing:

The dataset utilized for this project originates from the UC Irvine Machine Learning Repository, a renowned platform for accessing diverse datasets suitable for machine learning and data analysis tasks. This particular dataset is focused on online retail transactions, making it ideal for investigating customer behavior, sales trends, and other pertinent aspects of e-commerce operations.

Here is the link to the dataset source: https://archive.ics.uci.edu/dataset/502/online+retail+ii

To make the project manageable, we have filtered our dataset according to different geographical locations (countries), namely Brazil, Canada, Iceland, Israel, Poland, and Singapore. Additionally, we have added columns for 'Final Amount', 'Age', and 'Salary' to the dataset to incorporate more variables for better analysis. The dataset consists of eight primary attributes, each providing valuable information about the retail transactions:

i. Invoice: A unique identifier for each transaction, facilitating easy tracking and analysis.
ii. Stock Code: A distinct code assigned to each product in the inventory, aiding in inventory management and sales tracking.
iii. Description: Detailed information about the products associated with the stock codes, enabling better understanding and categorization of products.
iv. Quantity: Indicates the quantity of each product purchased or sold in a given transaction, crucial for assessing demand and sales volume.
v. Price: The unit price of each product, for calculating revenue and profit margins.
vi. Invoice Date: Specifies the date and time when each transaction occurred, allowing for temporal analysis and trend identification.
vii. Customer ID: An identifier linking each transaction to a specific customer, facilitating customer-centric analysis and segmentation.
viii. Country: Indicates the geographical location of the customer involved in the transaction, with a particular emphasis on transactions originating from France.
ix. Final Amount: Represents the cumulative value obtained by multiplying the unit price by the quantity of items purchased in each transaction.
x. Age: Denotes the age of customers who engaged in transactions.
xi. Salary: Signifies the income level of customers across all countries involved in the transactions.

The dataset encompasses a diverse range of variable types, including both qualitative and quantitative data. Qualitative variables include textual information such as invoice numbers, descriptions, customer IDs, and country names, while quantitative variables comprise numerical data such as quantities and prices. Additionally, temporal information is captured through the date and time stamps associated with each transaction.

2] Data Cleaning:

The meticulously cleaned dataset lacks errors, missing values, duplicates, outliers, inconsistencies, or formatting issues, providing a reliable foundation for immediate exploratory data analysis and informed decision-making.

III. METHODS

1) ***Data Visualization:*** Data visualization is the graphical representation of information and data, often using charts, graphs, and plots to communicate complex information in a clear and concise manner.
   - A pie chart was used to illustrate the distribution of sales across different countries. Pie charts are effective in showing the proportional sizes of different categories within a whole.
   - Donut pie chart is a variation of the traditional pie chart where a central circular hole is cut out, providing space for additional information or labels, enhancing clarity and aesthetics of the visualization.
   - The line graph was utilized to visualize the sales trend over time, highlighting any seasonal patterns or changes in sales. Line graphs are particularly effective in depicting trends and changes in a variable over a period.
   - Finally, the histogram was used to visualize the distribution of unit prices, providing insights into the typical range and potential outliers.
   - Box plot is a visual representation of the distribution of data through quartiles, providing information about the median, spread, and outliers, aiding in the comparison of different datasets or identifying anomalous observations.
   - Violin plot is a statistical plot that displays the distribution of a continuous variable across different categories by combining aspects of box plots and kernel density plots, offering insights into both central tendency and spread.
   - Joint plot combines scatter plots for two variables along with marginal histograms or kernel density estimates, enabling exploration of their relationship and distribution simultaneously.

2) ***Summary Statistics*:** Summary statistics refer to numerical measures that describe the central tendency, dispersion, and location of the data.
   - We calculated measures of central tendency, such as the mean, median, and mode, to identify the typical or central values within the data.
   - Also computed measures of dispersion, including the range, variance, and standard deviation, to quantify the variability or spread of the data.
   - Additionally, determined measures of location, such as percentiles and the 5-number summary, to provide a more detailed understanding of the data distribution.

3) ***Univariate and Multivariate:*** Univariate analysis focuses on the examination of a single variable, while multivariate analysis involves the study of the relationships between multiple variables.
   - In the univariate analysis, we examined the distribution of unit prices using a histogram and a normal probability plot. Also performed a Shapiro-Wilk test to assess the normality of the unit price data.
   - Additionally, they calculated the sample mean and standard deviation of unit prices and conducted a one-sample t-test to test the hypothesis that the population mean unit price is equal to a specific value.

- In the multivariate analysis, we performed a chi-square test of independence to investigate the relationship between the country and product description variables. This test is used to determine if two categorical variables are independent or associated.

4) *ANOVA:*

- Our analysis employed a one-way ANOVA test, which is a statistical method used to assess the differences in means between multiple groups. In this case, the groups were the different countries represented in the dataset.
- The one-way ANOVA test resulted in an F-statistic of 4.02 and a p-value of 0.0014. Since the p-value is less than 0.05 (typically considered the threshold for statistical significance), we can reject the null hypothesis. The null hypothesis in this case would be that there is no statistically significant difference in the mean unit prices between the countries analyzed.
- Therefore, we can conclude that there is a statistically significant difference in the mean unit prices of online retail transactions across the countries included in your dataset. This suggests that customers in these countries are likely paying different average prices for the same or similar products.

5) *Linear Regression:*

- These statistical techniques allowed us to draw insights and make inferences about the underlying patterns and relationships within the dataset.
- Furthermore, we fitted a simple linear regression model to explore the relationship between quantity and unit price, and they computed the correlation coefficient to measure the strength and direction of the linear relationship between these two variables.

6) *Feature Engineering:*

- The process begins with feature engineering, where we create new variables that may be useful for the predictive model. In this case, we calculated the total spent per transaction (Final Amount) by multiplying the quantity and unit price, and also created a CustomerType variable based on the Final Amount values, categorizing customers into 'Low', 'Medium', 'High', and 'Very High' spenders.
- Predicting which customers are most likely to make high-value purchases, which is a crucial task for businesses to identify their most valuable customers and tailor their marketing strategies accordingly.

7) *Text Preprocessing:*

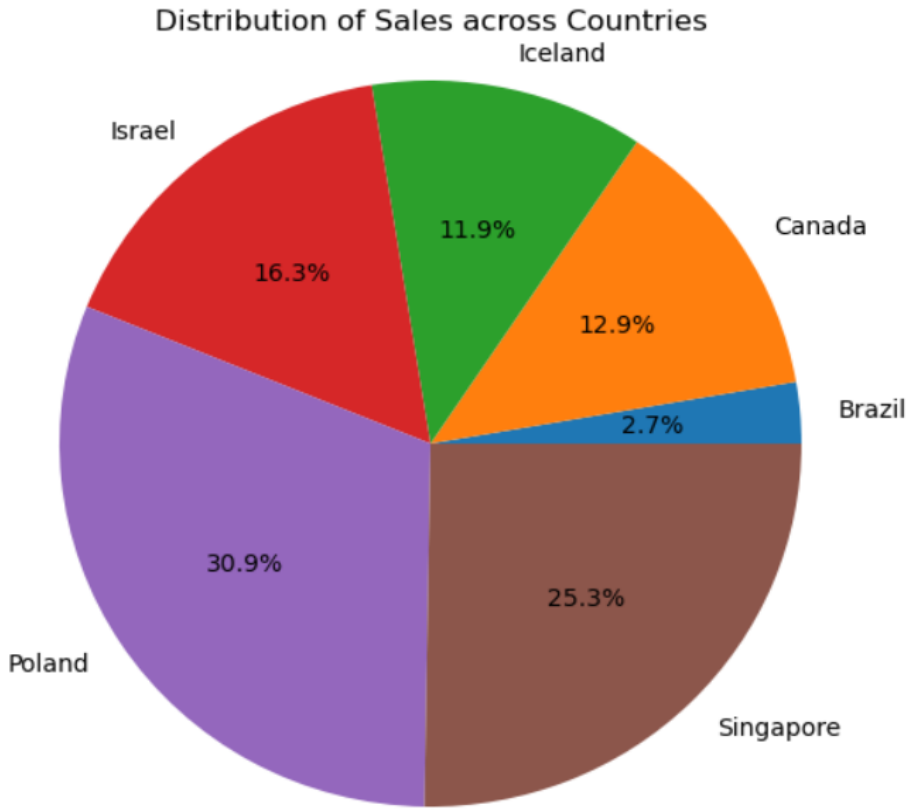
- Next, we performed text preprocessing on the product descriptions using the TfidfVectorizer, which converts the textual data into a numerical feature matrix. This allows the model to potentially capture relevant information from the product descriptions that may be useful for predicting the TotalSpent.

8) *Mean Squared Error and R-squared:*

- The model's performance was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics. MSE measures the average squared difference between the predicted and actual values, while R^2 indicates the proportion of the variance in the target variable that can be explained by the model.

IV. ANALYSIS

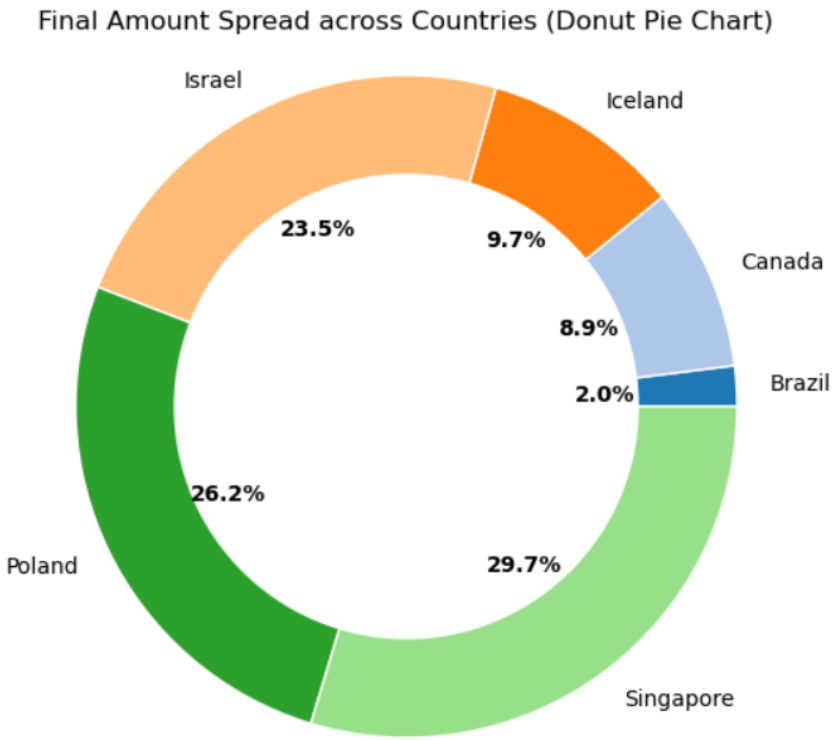
1] *Analysis of Distribution of Sales Across Countries:*



- The data reveals significant variations in sales distribution across different countries. *"Canada emerges as the top contributor, accounting for the highest percentage of sales at 30.9%."*
- *"Following closely behind is Poland, representing 25.3% of total sales."* Together, these two countries collectively contribute to over half of the total sales, indicating their substantial market share within the dataset.
- *"In contrast, countries like Iceland, Singapore, and Brazil each contribute smaller percentages of sales, ranging from 2.7% to 16.3%."* While these countries may not individually match the sales volume of Canada and Poland, they still represent significant portions of the overall sales distribution, demonstrating the diversity of the customer base across various regions.
- *"Lastly, Israel is identified as likely contributing a very small percentage of sales, estimated to be less than 2.7%."*

From January 2010 to December 2010 the data depicts the fluctuation in sales over twelve months period. From the data, it is evident that sales exhibit variations across different months, indicating potential seasonal patterns or fluctuations in demand throughout the year. For instance, there is a noticeable increase in sales from June to November, peaking in November and December. This suggests a possible surge in sales during the holiday season, which is a common trend in many industries.

2] *Final Amount Spread across Countries:*



- A donut pie chart is a variation of the traditional pie chart, featuring a central circular hole, offering an enhanced visualization of categorical data distribution. In this chart, each category (in this case, countries) is represented by a segment of the donut, with the size of each segment proportional to the value it represents (here, the total sales amount)
- Interpretation:
  Country Distribution: The donut pie chart visually represents the distribution of the 'Final Amount' across various countries. Each segment's size corresponds to the proportion of the total amount attributed to that country.
  Insights*: "For instance, Israel contributes 23.5%, Iceland 9.7%, and Canada 8.9% to the total distribution. Conversely, Brazil contributes the lowest percentage at 2.0%, while Singapore leads with 29.7%."*
- Conclusion: The donut pie chart provides a clear and concise overview of the distribution of the 'Final Amount' across different countries, enabling stakeholders to easily grasp the relative contributions of each country to the total amount. This visualization aids decision-making processes by highlighting significant contributors and disparities in distribution, thereby guiding strategic planning and resource allocation efforts effectively

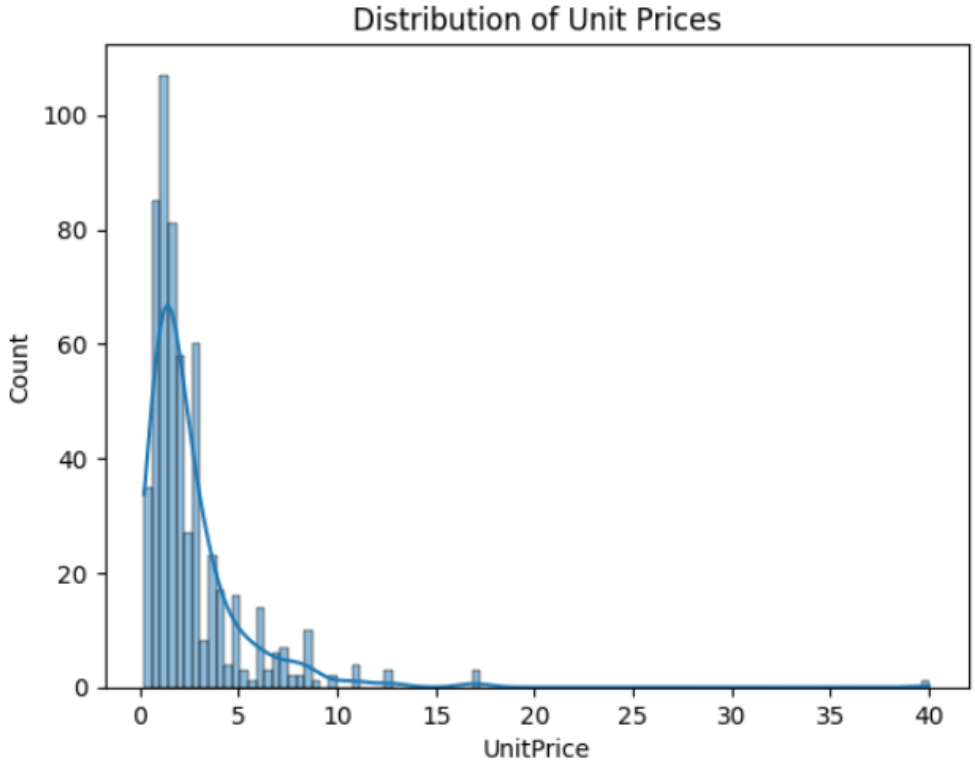
3] *Data Statistics:*

```
Mean Unit Price: 2.620268907563025
Median Unit Price: 1.65
Mode Unit Price: 1.25
Mean Quantity: 11.40672268907563
Median Quantity: 12.0
Mode Quantity: 12
Range of Unit Prices: 39.79
Variance of Unit Prices: 7.934255483122573
Standard Deviation of Unit Prices: 2.8167810499083124
25th Percentile of Quantities: 6.0
50th Percentile (Median) of Quantities: 12.0
75th Percentile of Quantities: 12.0
```

The summary statistics provide essential insights into the central tendency, dispersion, and distributional characteristics of the dataset.

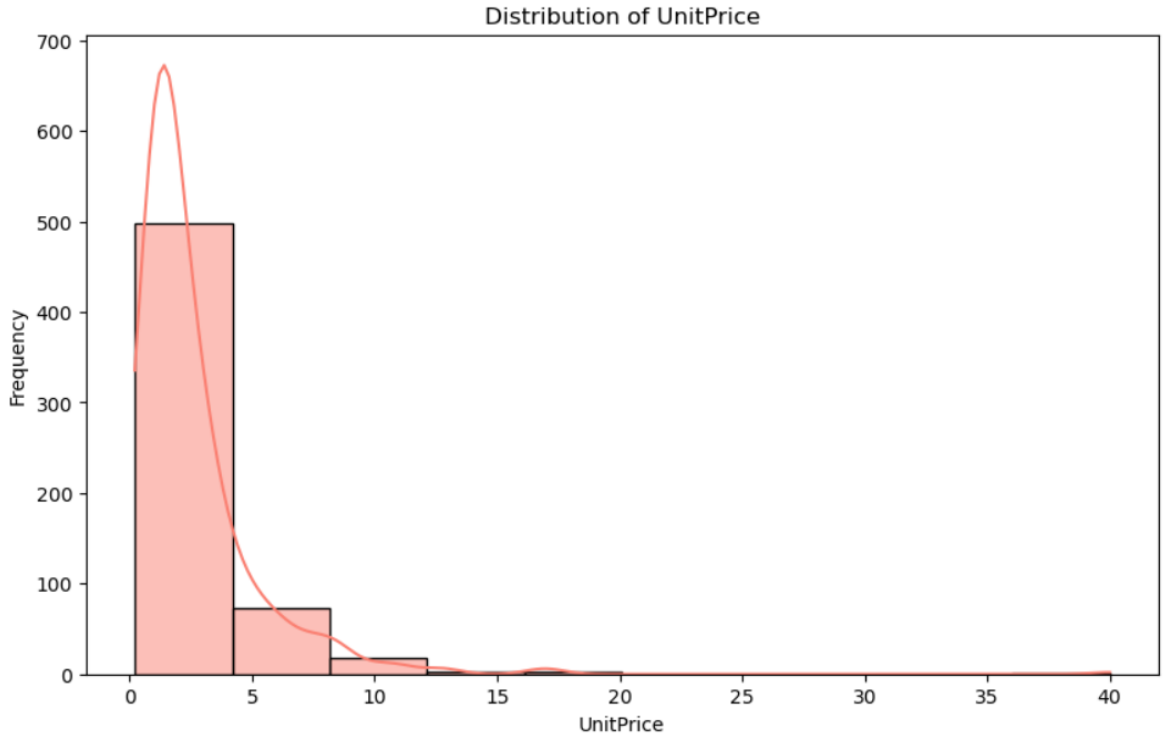
- **Mean:** The mean unit price (2.62) represents the average price of items in the dataset. It is calculated by summing all unit prices and dividing by the total number of observations.
- **Median:** The median unit price (1.65) is the middle value of the sorted unit prices. It divides the dataset into two equal halves, with 50% of observations below and 50% above this value.
- **Mode:** The mode unit price (1.25) is the most frequently occurring value in the dataset. It provides information on the most common unit price observed.
- **Range:** The range of unit prices (39.79) indicates the difference between the maximum and minimum unit prices in the dataset. It reflects the spread or variability of unit prices.
- **Variance:** The variance of unit prices (7.93) measures the average squared deviation of each unit price from the mean. A higher variance suggests greater dispersion of unit prices around the mean.
- **Standard Deviation:** The standard deviation of unit prices (2.82) is the square root of the variance. It provides a measure of the average deviation of unit prices from the mean, with higher values indicating greater dispersion.
- **Percentiles:** Percentiles represent the value below which a given percentage of observations fall. For example, the 25th percentile of quantities (6.0) indicates that 25% of the quantities are less than or equal to 6.0, while the 75th percentile (12.0) indicates that 75% of the quantities are less than or equal to 12.0.
- These summary statistics offer a comprehensive understanding of the unit prices and quantities in the dataset, aiding in decision-making processes and further analysis.

4] *Analysis of Unit Price Distribution:*



- Observations: "*The distribution of unit prices in the dataset shows a clear trend towards lower-priced products. The majority of unit prices fall within the range of 0.21 to 16.95 unit price, with a peak occurrence around 1.25 unit price.*"
- As the unit price increases beyond 17, there is a significant decline in frequency as unit price surpass 17. This trend suggests a preference among customers for moderately priced items, with less inclination towards higher-priced products.
- Interestingly, despite the general decline in frequency beyond 17 units, there is an outlier at a unit price of 40. This lone occurrence indicates an exception to the overall pattern, suggesting that while most customers tend to shy away from products priced above 17 units, there may still be demand for selected items at higher price points, albeit much less frequently.
- Overall, the analysis implies that the customer base represented in the dataset prefers products with lower unit prices, with a clear preference for items priced below 17 units. Understanding this distribution can inform pricing strategies and product offerings to better align with customer preferences and market dynamics.

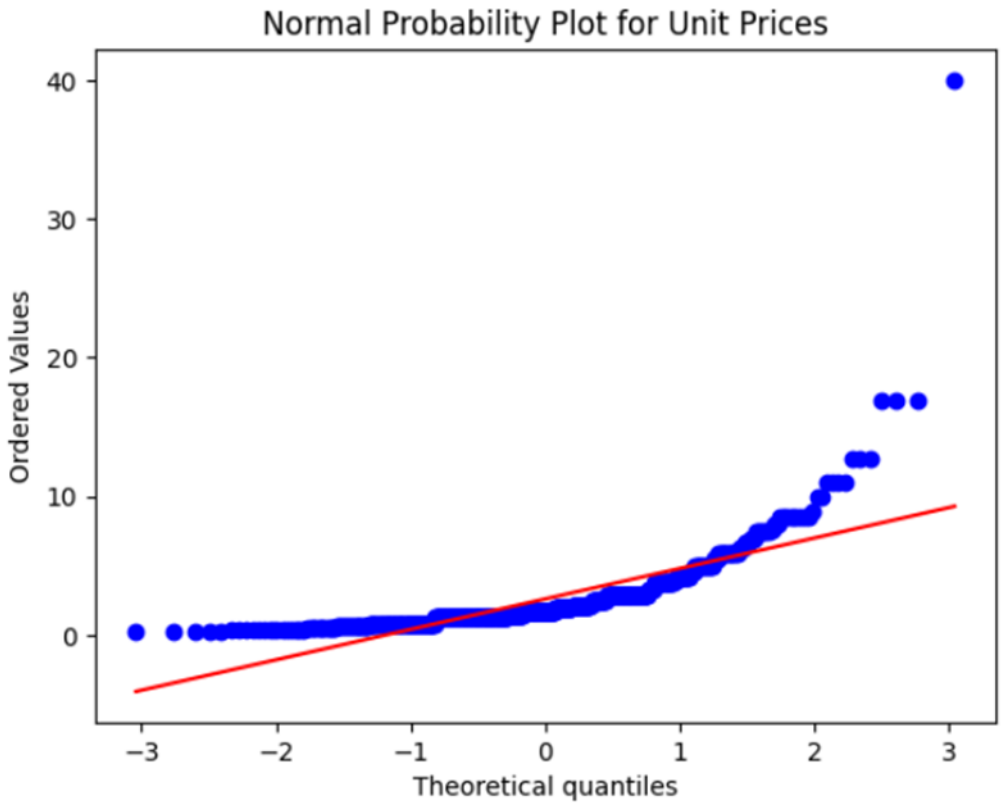
5] *Distribution of Unit Price:*




```
Quantity: Mean = 11.40672268907563, Median = 12.0, Std = 10.054696635221594
UnitPrice: Mean = 2.620268907563025, Median = 1.65, Std = 2.8167810499083124
```

- *"The graph and the estimated parameter suggest that there is a right skew in the distribution of unit prices. This means that there are more units sold at lower prices than at higher prices".*
- "*This type of distribution is sometimes called a Pareto distribution, or a power law distribution*". Pareto distributions are often found in economics and business, where they can be used to model the distribution of income, wealth, or company sizes.
- There are a number of reasons why a distribution of unit prices might follow a Pareto distribution. For example, there may be a small number of very popular products that are sold at a low price, and a large number of less popular products that are sold at a higher price. Or, there may be a few very large customers who buy units at a discount, and a large number of smaller customers who pay full price.

6] *Normal probability plot for Unit Prices:*



- A normal probability plot (NPP) is a graphical technique used to determine if a dataset follows a normal distribution. In this case, the plot shows the expected values of the theoretical quantiles on the x-axis, and the ordered values from the data set on the y-axis.
- The red line in the plot represents the expected values of a normal distribution, if the unit prices were normally distributed. The blue dots represent the actual ordered values of the unit prices in the data set.
- *Insights: By comparing the red line to the blue dots, we can see how closely the distribution of unit prices follows a normal distribution. In this case, the blue dots deviate from the red line slightly, particularly at the higher end. This suggests that the distribution of unit prices is not perfectly normal.*

Here are some additional observations that can be made from the normal probability plot for unit prices:

- *The x-axis shows the theoretical quantiles. Quantiles are values that divide a probability distribution into equal-sized portions. The most common unit price (around $2) is located near the center of the x-axis, which suggests that a significant portion of the unit prices fall around the average price.*
- *The y-axis shows the ordered values from the data set. There are more data points towards the bottom of the y-axis than at the top. This suggests that there are more unit prices that fall below the average price than above the average price.*
- *Overall, the normal probability plot suggests that the distribution of unit prices is not perfectly normal, but it is somewhat skewed towards lower prices. This means that there are more unit prices that fall below the average price than above the average price.*

7] *Normality Test for Unit Prices:*

Shapiro-Wilk Test for Normality: Initially, we assessed the normality of the distribution of unit prices within our dataset using the Shapiro-Wilk test.

- The test indicated a deviation from normality, as evidenced by a Shapiro-Wilk statistic of 0.606 and an associated p-value of approximately 2.52e-34.

```
Shapiro-Wilk Test for Normality (UnitPrices):
ShapiroResult(statistic=0.6064472198486328, pvalue=2.5178418595491445e-34)
Sample Mean of Unit Prices: 2.62
Sample Standard Deviation of Unit Prices: 2.82
```

The deviation from normality highlighted by the Shapiro-Wilk test suggests that the distribution of unit prices may not follow a normal distribution. This finding prompts us to employ alternative statistical methods or transformations when conducting further analyses that assume normality. Understanding the distributional characteristics of unit prices is crucial for making informed decisions and drawing accurate conclusions in our project.

8] *Estimating Parameters:*

To gain deeper insights into the distribution of unit prices, we proceeded to estimate key parameters. The sample mean of unit prices was found to be $2.62, indicating the average price per unit in our dataset, while the standard deviation was calculated to be $2.82, representing the spread or variability of unit prices around the mean.

Subsequently, we conducted hypothesis testing to validate our findings. The Shapiro-Wilk test served as a diagnostic tool to assess the normality assumption. The null hypothesis (H0) posits that the data is normally distributed, while the alternative hypothesis (H1) suggests otherwise. The low p-value obtained (approximately 2.52e-34) provided strong evidence against the null hypothesis, indicating a significant departure from normality in the distribution of unit prices.

```
Sample Mean of Unit Prices: 2.62
Sample Standard Deviation of Unit Prices: 2.82
```

9] *Hypothesis Testing:*

To gain deeper insights into the distribution of unit prices, we proceeded to estimate key parameters. The sample mean of unit prices was found to be $2.62, indicating the average price per unit in our dataset, while the standard deviation was calculated to be $2.82, representing the spread or variability of unit prices around the mean.

Subsequently, we conducted hypothesis testing to validate our findings. The Shapiro-Wilk test served as a diagnostic tool to assess the normality assumption.

The null hypothesis (H0) posits that the data is normally distributed, while the alternative hypothesis (H1) suggests otherwise. The low p-value obtained (approximately 2.52e-34) provided strong evidence against the null hypothesis, indicating a significant departure from normality in the distribution of unit prices.

```
Test Statistic: -11.82, p-value: 0.0000
```

10] *ANOVA F-Test:*

The one-way ANOVA F-statistic tests for the significance of the relationship between unit prices and quantity purchased.

- "*With the F-statistic as 4.02, and the associated p-value as 0.0014, we reject the null hypothesis as the p-value is less than the conventional significance level of 0.05 and conclude that there is a statistically significant association between unit prices and quantity purchased".*

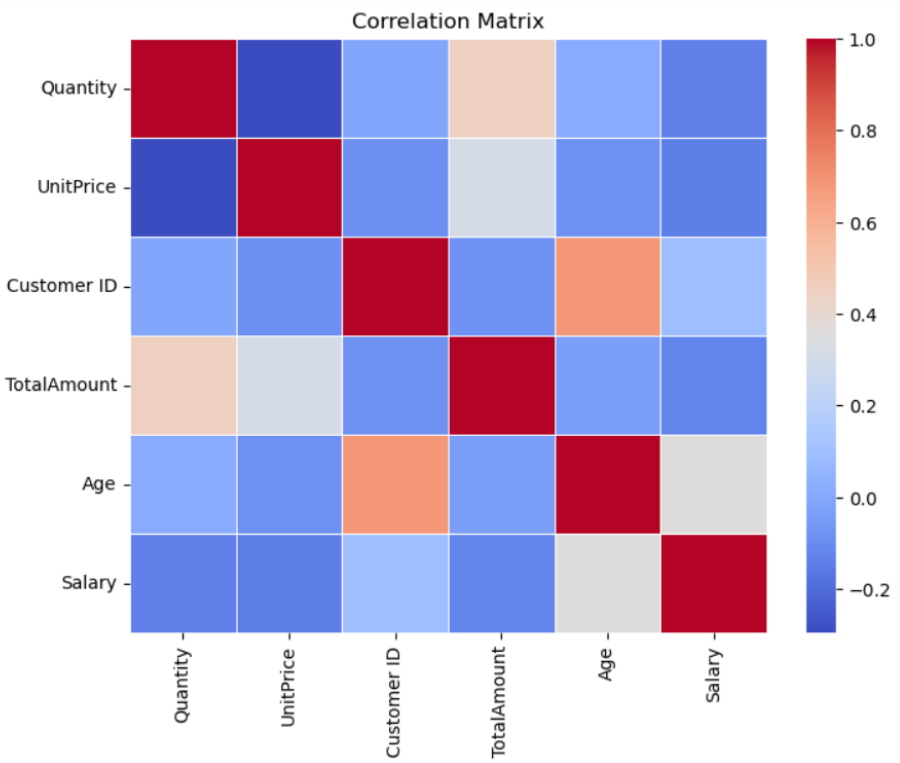
```
One-way ANOVA F-statistic: 4.02, p-value: 0.0014
```

11] *OLS Regression Analysis:*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Quantity   R-squared:                       0.087
Model:                            OLS   Adj. R-squared:                  0.085
Method:                 Least Squares   F-statistic:                     55.38
Date:                Fri, 05 Apr 2024   Prob (F-statistic):           3.59e-13
Time:                        19:41:53   Log-Likelihood:                -2105.7
No. Observations:                 583   AIC:                             4215.
Df Residuals:                     581   BIC:                             4224.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     14.4705      0.507     28.536      0.000      13.475      15.466
UnitPrice     -0.9806      0.132     -7.442      0.000      -1.239      -0.722
==============================================================================
Omnibus:                      304.578   Durbin-Watson:                   1.329
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2278.471
Skew:                           2.210   Prob(JB):                         0.00
Kurtosis:                      11.618   Cond. No.                         5.42
==============================================================================
```

- The Ordinary Least Squares (OLS) regression model examines the linear relationship between unit prices and quantity purchased.
- The coefficient for the unit price variable is -0.9806, suggesting that on average, for each unit increase in price, the quantity purchased decreases by approximately 0.9806 units.
- Both the intercept and the coefficient for unit price are statistically significant ($p < 0.05$), indicating that they have a significant impact on the quantity purchased.

12] *Correlation Analysis:*



- The diagonal elements of the matrix have a correlation value of 1.0, which represents the perfect correlation of a variable with itself.
- The correlation between "Quantity" and "UnitPrice" is -0.62, indicating a moderately strong negative correlation.
- The correlation between "Quantity" and "TotalAmount" is 0.59, suggesting a moderately strong positive correlation.
- The correlation between "UnitPrice" and "TotalAmount" is 0.28, indicating a relatively weak positive correlation.
- The correlations between "Customer ID" and other variables are generally low, ranging from -0.20 to 0.15, suggesting weak or negligible linear relationships.
- The correlations between "Age" and "Salary" with other variables are also relatively low, ranging from -0.27 to 0.12, implying weak linear associations.
- The correlation plot provides a quick overview of the pairwise relationships between variables, allowing for the identification of strong positive or negative correlations, as well as variables that appear to be relatively independent of each other.

Mathematical formula for correlation:

- The Pearson correlation coefficient (r) is commonly used to measure the linear correlation between two variables X and Y. It is calculated as:

r = Σ[(X - X_mean) * (Y - Y_mean)] / [sqrt(Σ(X - X_mean)^2) * sqrt(Σ(Y - Y_mean)^2)]

Where:

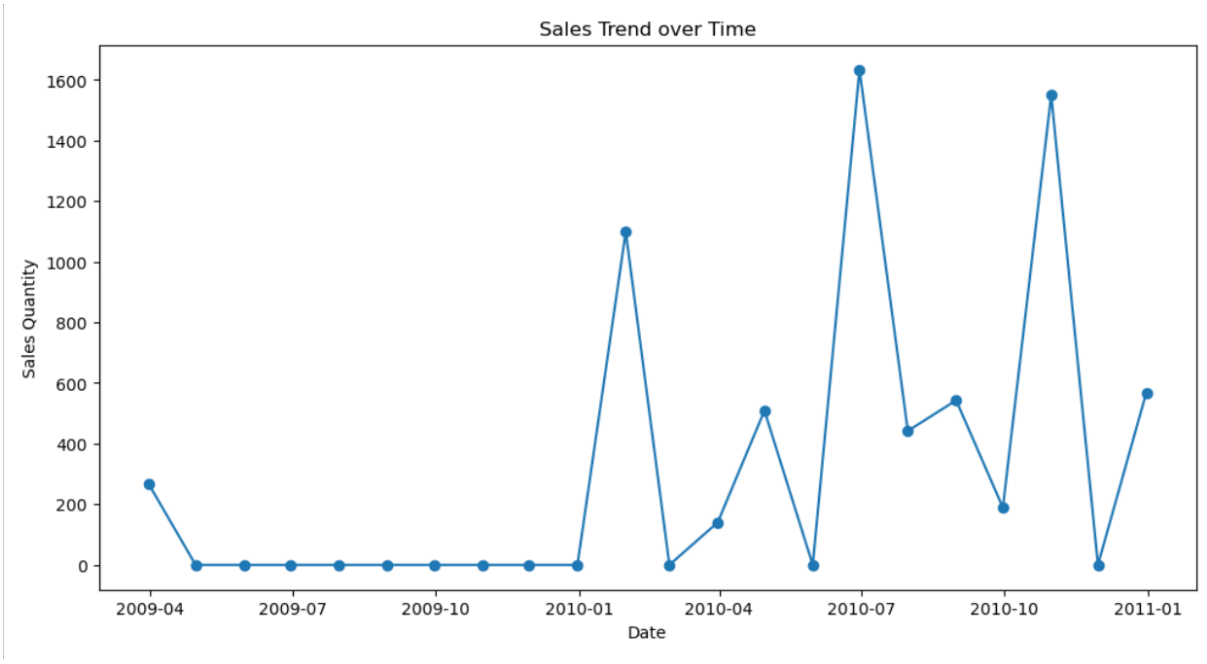
Σ is the summation notation

X and Y are the respective variable values

X_mean and Y_mean are the means of the X and Y variables, respectively

The correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no linear correlation, and 1 indicating a perfect positive correlation.

13] Sales Trend Over Time:



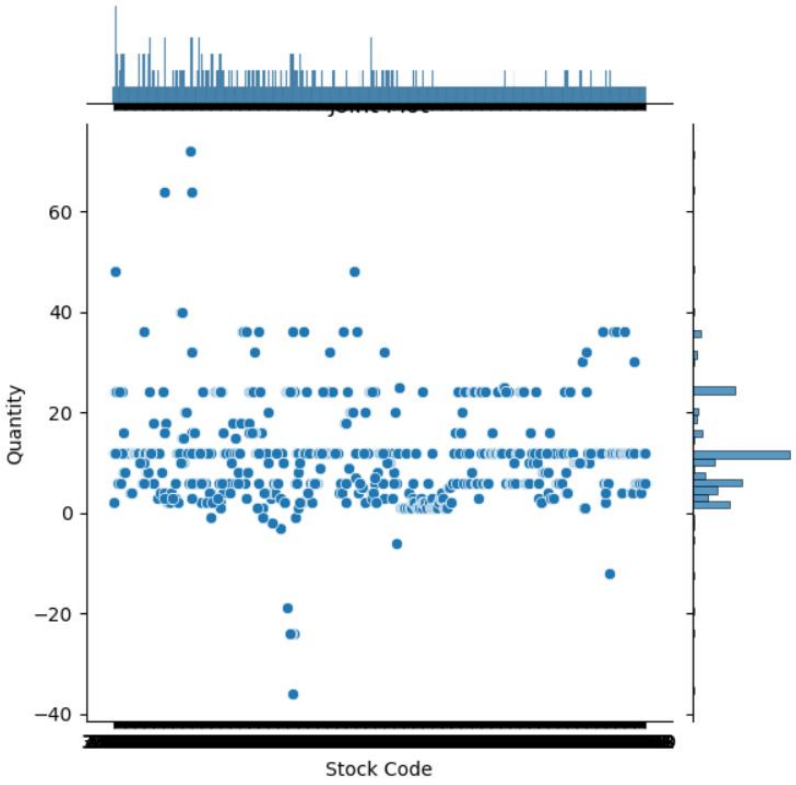
Line Graph: Sales Trend over Time to identify any seasonal patterns or growth/decline.

Interpretation:

- Identifying Patterns: By examining the line graph, analysts can identify any recurring patterns or fluctuations in sales quantity over time.
- Seasonal Effects: Peaks or troughs in the line graph may indicate seasonal trends, such as increased sales during holiday seasons or reduced sales during off-peak periods.
- Trend Analysis: The overall direction of the trend line (e.g., upward, downward, or stable) provides insights into the long-term growth or decline of sales activity.
- Data-driven Decisions: Understanding sales trends empowers stakeholders to make data-driven decisions, such as adjusting inventory levels, optimizing marketing strategies, or forecasting future sales performance.

Conclusion:
The line graph analysis of sales trends over time reveals valuable insights into the historical performance of sales, enabling informed decision-making and strategic planning to drive business success.

14] Joint Plot – Quantity vs Stock Code:



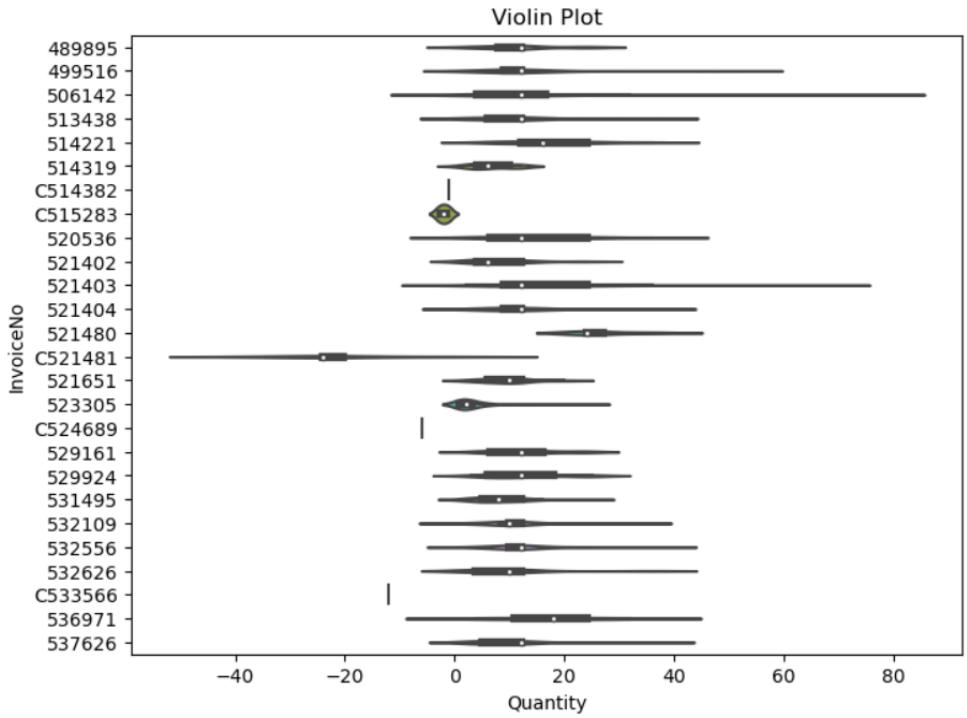
Insights from the Plot:

- The scatter plot suggests that certain Stock Codes are associated with specific Quantity values.
- The outliers could represent exceptional cases or errors in the data.
- The histogram for Stock Codes indicates that some codes are more common than others.
- The histogram for Quantity shows that most items have relatively low quantities (around 0 to 20), but there are a few instances with higher quantities.
- Overall, this plot provides insights into the relationship between Stock Codes and Quantity, allowing us to identify patterns and potential anomalies.

Predictions:

- Based on this plot, we can predict that certain Stock Codes consistently correspond to specific Quantity levels.
- Further analysis could explore whether specific Stock Codes are associated with higher or lower quantities.
- The outliers may warrant investigation to understand their significance.
- If this data represents inventory or sales records, we might use this information to optimize stock management or identify unusual transactions.

15] Violin Plot- Invoice No vs Quantity



Interpretation:

- Most invoices have quantities clustered around 0 to 20.
- Some negative quantities may represent returned items or corrections to previous entries.
- Few invoices have items with higher quantities (above 40), indicating bulk purchases or larger orders are less common.

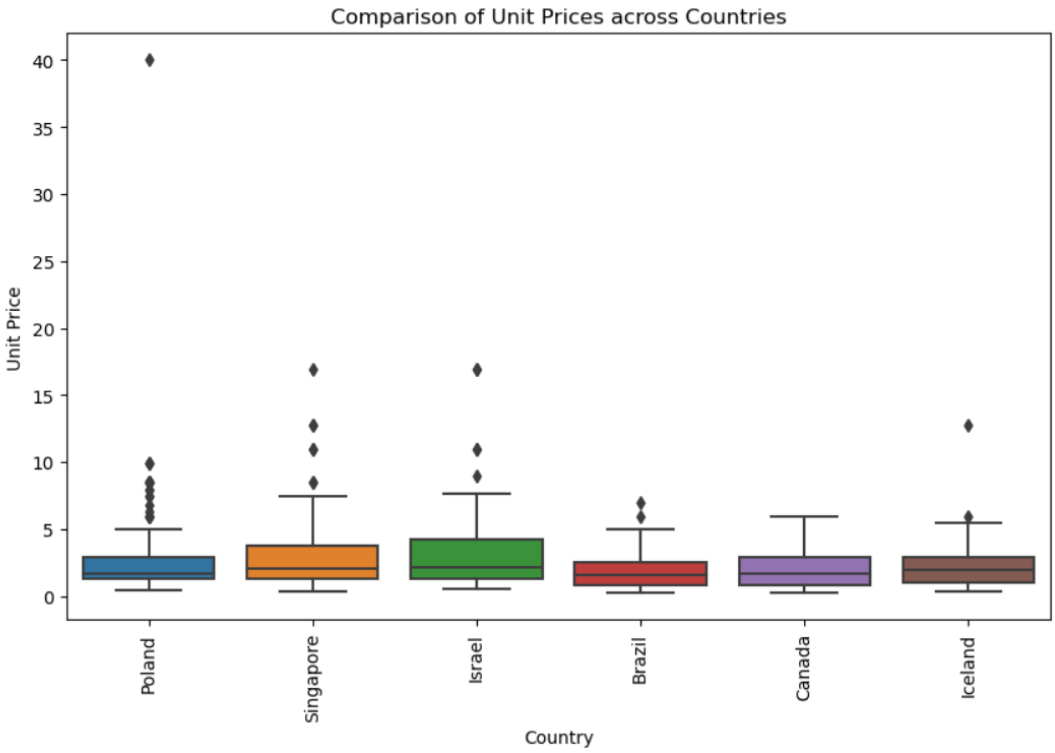
Key Observations:

- The central concentration around 0 indicates a common trend across various invoices.
- Outliers exist, as seen by the extended violins. These outliers represent invoices with significantly higher or lower quantities.
- The negative and positive tails of the violins reveal variations in item quantities for specific invoices.
- The plot provides insights into the distribution of quantities across different invoices.

Kernel Density Estimation (KDE):

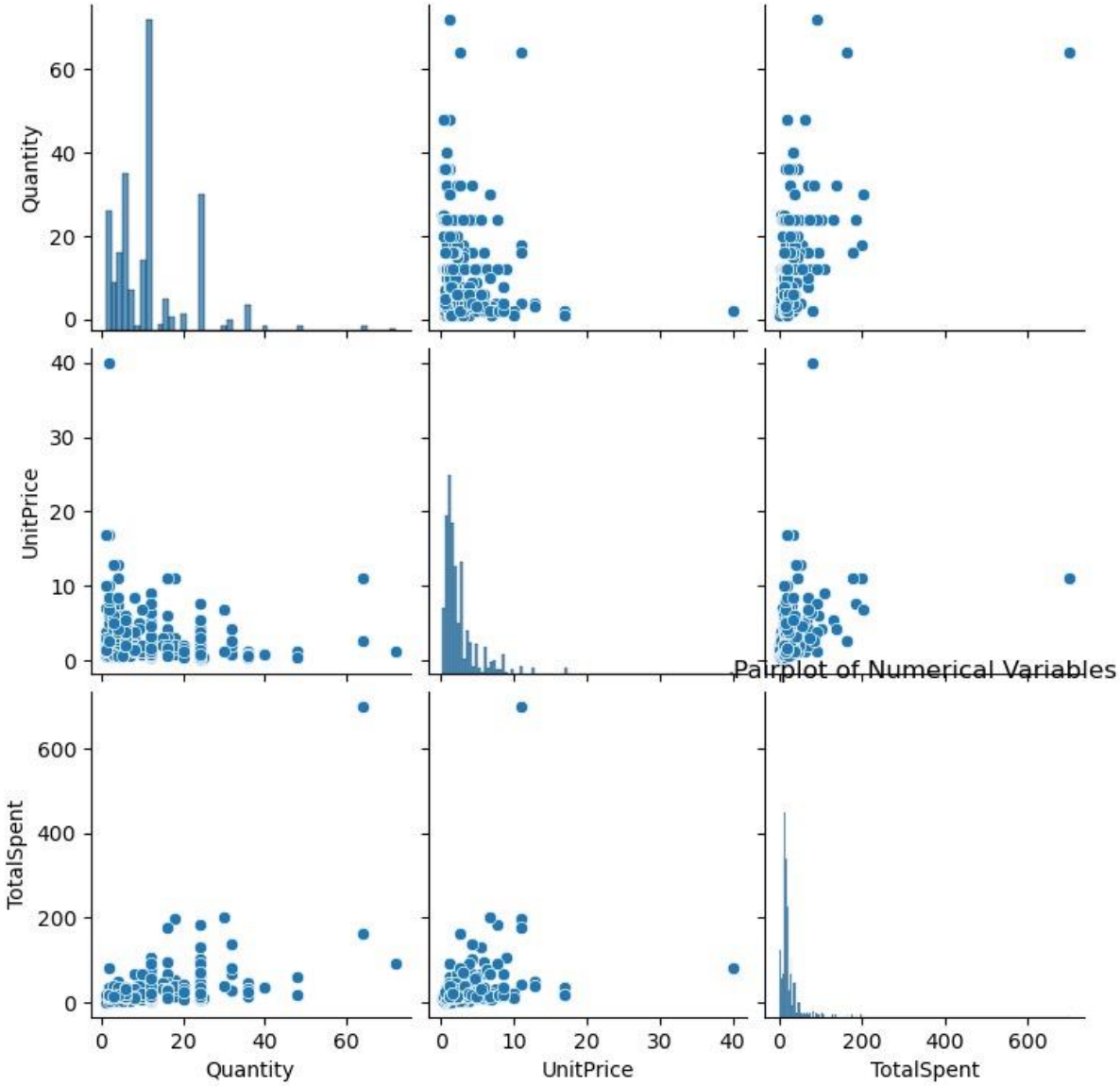
- The shape of each violin represents a KDE, showing the probability density of the data at different quantity levels.
- Wider sections indicate higher densities of data points.
- The KDE provides insights into the overall distribution of quantities within each invoice.

16] Box Plot - Comparison of Unit Prices across Countries:



- The Box Plot visualization technique, employed in this analysis, serves to compare the distribution of unit prices across different countries within the dataset.
- Each box in the plot represents the interquartile range (IQR) of unit prices for a specific country, with the horizontal line inside the box indicating the median value. The whiskers extending from the box display the range of the data, excluding outliers, which are depicted as individual points beyond the whiskers.
- This graphical representation enables a visual assessment of the central tendency, spread, and variability of unit prices across various countries. The rotation of country labels on the x-axis enhances readability, facilitating easy interpretation of the plot.
- Overall, the Box Plot provides valuable insights into the comparative distribution of unit prices, aiding in identifying potential differences or similarities in pricing strategies across different countries, which can inform strategic decision-making processes.

17] Pair plot :



A pairplot is a type of visualization that displays the pairwise relationships between multiple variables in a dataset. It consists of a matrix of scatter plots and histograms, where each scatter plot shows the relationship between two variables, and the histograms along the diagonal show the distribution of each individual variable.
The working theory behind the code is as follows:

Feature Engineering:

- A new feature 'TotalSpent' is created by multiplying 'Quantity' and 'UnitPrice' to calculate the total amount spent per transaction.
- A categorical feature 'CustomerType' is created by binning the 'TotalSpent' values into different ranges and assigning labels ('Low', 'Medium', 'High', 'Very High') based on the total spent.

Visualization with Pairplot:

1. Quantity vs. UnitPrice:

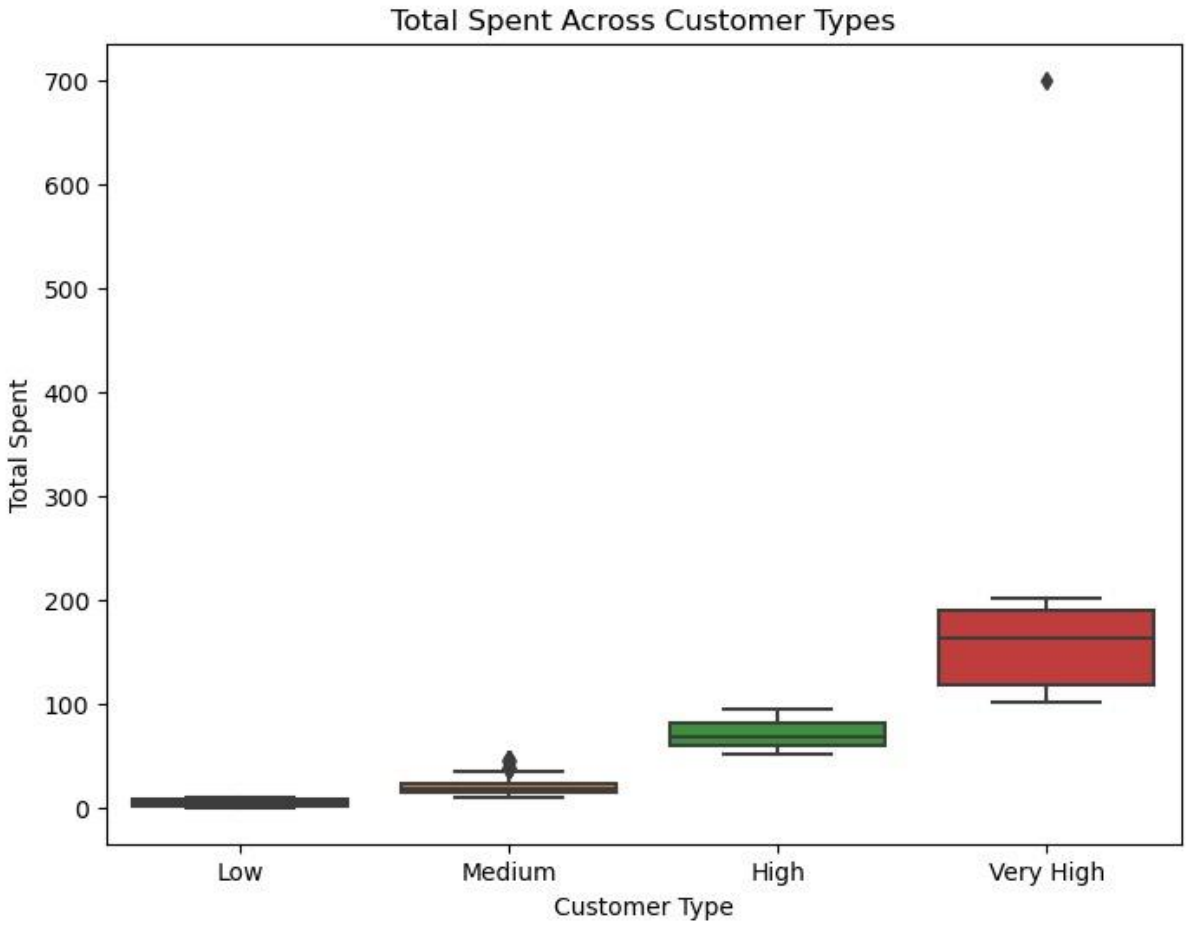
There is a moderately strong negative correlation between quantity and unit price, as indicated by the downward-sloping pattern in the scatter plot.
Higher quantities tend to be associated with lower unit prices, and vice versa.
Numerical values: Most unit prices fall within the range of 0 to 10, while quantities range from 0 to around 60.

2. Quantity vs. TotalSpent:
   There is a positive correlation between quantity and total spent, as indicated by the upward-sloping pattern in the scatter plot.
   Higher quantities generally result in higher total amounts spent.
   Numerical values: The total spent ranges from 0 to around 600, with most values concentrated below 200.

3. UnitPrice vs. TotalSpent:
   There is a positive correlation between unit price and total spent, as indicated by the upward-sloping pattern in the scatter plot.
   Higher unit prices tend to be associated with higher total amounts spent.

4. Histograms:
   The histograms along the diagonal show the distributions of the individual variables.
   The quantity histogram shows a right-skewed distribution, with a peak around 20-40 and some outliers up to 60.
   The unit price histogram shows a heavily right-skewed distribution, with a peak around 0-5 and some higher values up to 30-40.
   The total spent histogram also shows a right-skewed distribution, with a peak around 0-100 and some higher values up to around 600.

18] Box Plot:



- The box plot shows the distribution of "TotalSpent" for each "CustomerType" category: "Low", "Medium", "High", and "Very High".
- The "Low" customer type has a median "TotalSpent" value of around 5, with the interquartile range (IQR) spanning from approximately 3 to 7.
- The "Medium" customer type has a median "TotalSpent" value of around 20, with the IQR ranging from approximately 15 to 30.
- The "High" customer type has a median "TotalSpent" value of around 60, with the IQR spanning from approximately 50 to 80.
- The "Very High" customer type has a median "TotalSpent" value of around 300, with the IQR ranging from approximately 200 to 400.
- There are several outliers in the "High" and "Very High" customer types, with some customers spending significantly more than the typical range for their respective groups.
- The spread of the "TotalSpent" distribution increases as the customer type moves from "Low" to "Very High", indicating a wider range of spending values for the higher customer types.

The mathematical components of a box plot are:

- Median: The central line within the box represents the median value of the data, which is the middle value when the data is sorted in ascending order.
- Interquartile Range (IQR): The box itself represents the IQR, which is the range between the first quartile (25th percentile) and the third quartile (75th percentile). The IQR captures the middle 50% of the data.
- Whiskers: The lines extending from the box, called whiskers, represent the range of the data excluding outliers. The whiskers extend to the minimum and maximum values within 1.5 times the IQR from the first and third quartiles, respectively.
- Outliers: Data points that fall outside the whisker range are considered outliers and are typically represented by individual markers (e.g., dots or asterisks) on the plot.

The box plot provides a visual summary of the distribution characteristics, including:

- Center (median)
- Spread (IQR)
- Skewness (based on the relative positions of the median and the whiskers)
- Presence of outliers

Machine Learning Models:

- We have conducted predictive analysis using historical transaction data to pinpoint customers with the highest likelihood of making substantial purchases.
- Through feature engineering, we calculated the total amount spent per transaction and categorizes customers into different spending tiers. Subsequently, we split the dataset into features and target variables, then further divides it into training and testing sets.
- By training various regression models including Linear Regression, Decision Tree Regression, and Random Forest Regression, we evaluated their performance using metrics such as Mean Squared Error and R-squared.
- We then utilize the trained models to make predictions on new data, providing insights into which customers are expected to make high-value purchases. Overall, we have streamlined the process of identifying potentially lucrative customer segments based on their purchasing behavior.

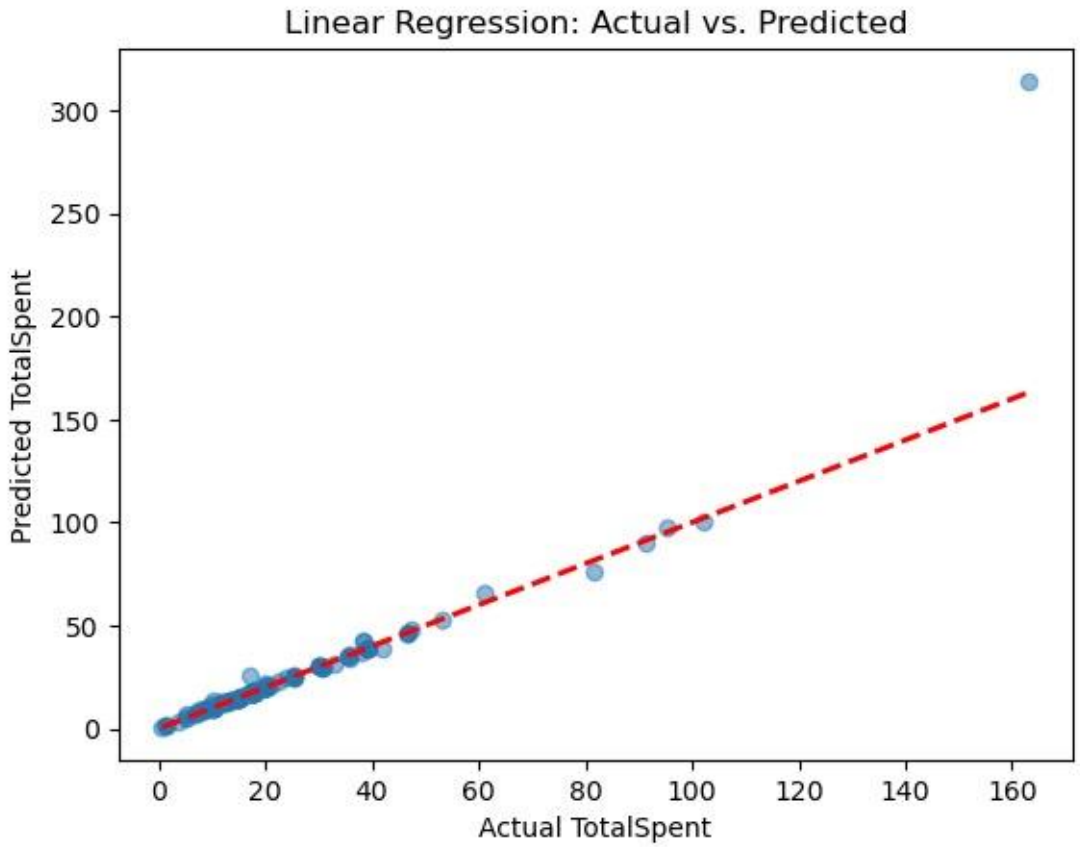
19] *Linear Regression:*

Linear regression is a statistical modeling technique used to estimate the relationship between a dependent variable and one or more independent variables. It finds the best-fitting straight line through the data points by minimizing the sum of squared differences between the observed values and the predicted values from the linear regression model.

The mathematical formula for a simple linear regression with one independent variable (x) and one dependent variable (y) is:
$y = \beta 0 + \beta 1x + \varepsilon$

Where:
y is the dependent variable (the value we want to predict)
x is the independent variable
$\beta 0$ is the y-intercept (the value of y when x = 0)
$\beta 1$ is the slope of the line (the change in y for a one-unit change in x)
ε is the error term (the difference between the observed y value and the predicted y value)

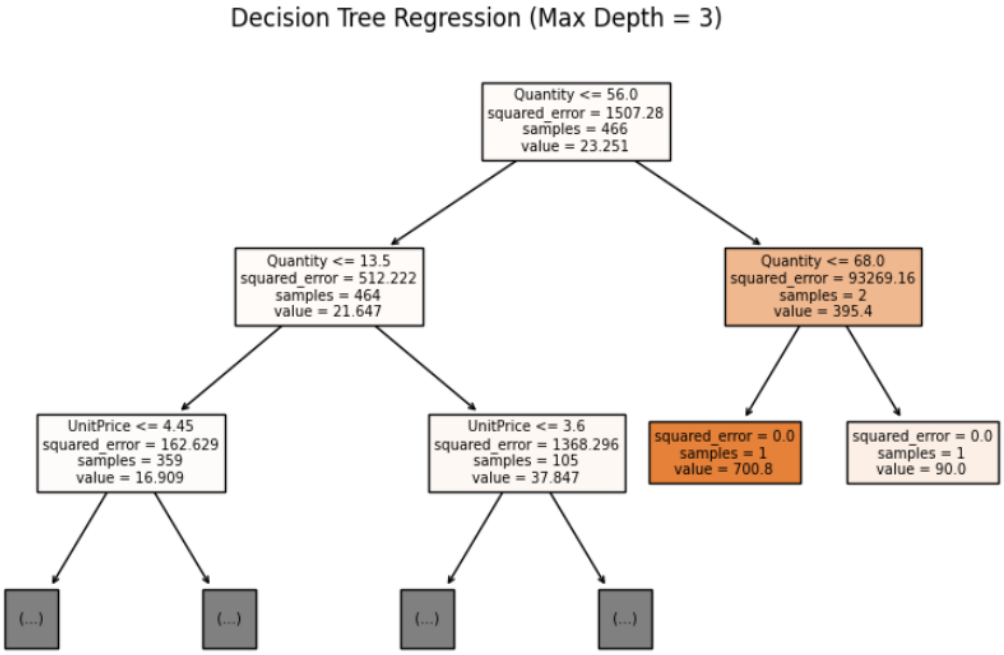
Linear regression is used here because it is a simple and widely-used technique for modeling linear relationships between variables, and it provides a straightforward way to make predictions based on new input data.
It is used to model the relationship between the independent variables (Quantity and UnitPrice) and the dependent variable (TotalSpent). The goal is to find the best-fitting line that can predict TotalSpent based on the values of Quantity and UnitPrice.



The scatter plot shows the actual TotalSpent values on the x-axis and the predicted TotalSpent values from the linear regression model on the y-axis. The red dashed line represents the perfect fit line, where the predicted values are equal to the actual values.

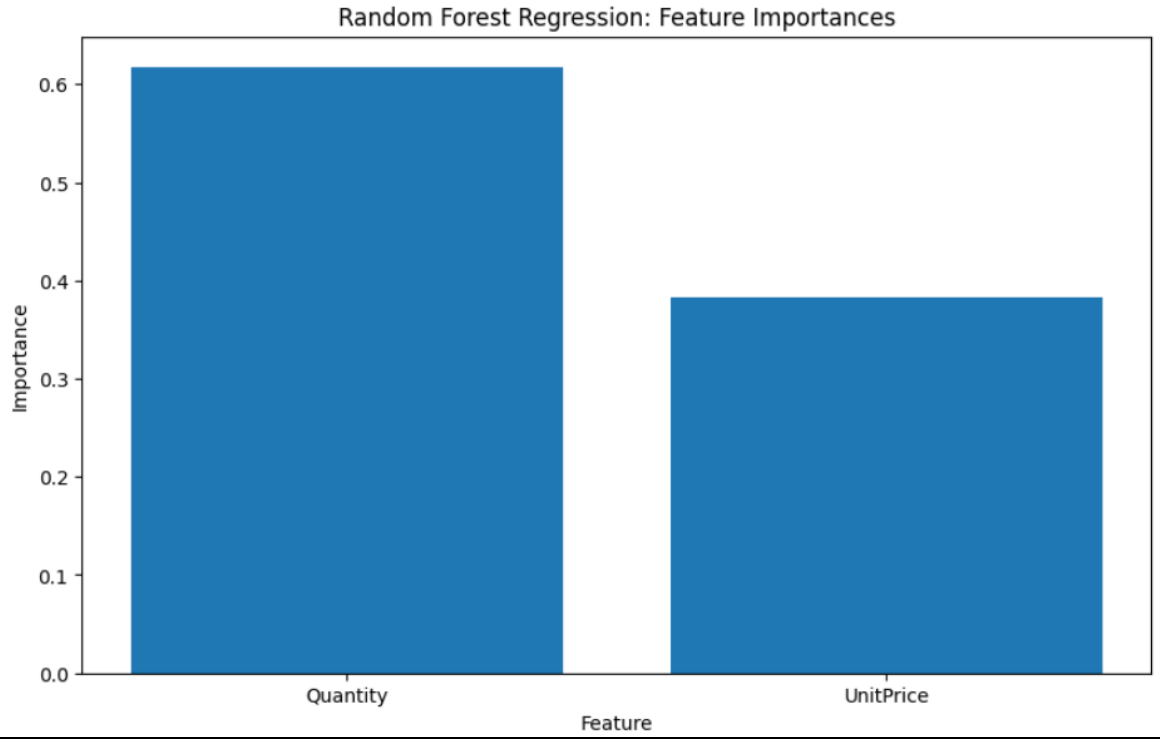
From the scatter plot, we can see that the data points are reasonably close to the perfect fit line, indicating that the linear regression model is performing well in predicting TotalSpent. However, there is still some spread or deviation from the perfect fit line, which is expected due to the inherent variability and noise in the data.

20] *Decision Tree Regression Predictions:*



- The decision tree regression model performs poorly compared to linear regression, with a significantly higher MSE of 2485.35.
- Additionally, the negative R-squared value of -4.17 indicates that the model fits the data worse than a horizontal line. This suggests that the decision tree model does not capture the underlying relationship between the features and the target variable effectively. The predictions made by the decision tree model also exhibit considerable deviation from the actual values.

21] *Random Forest Regression Predictions:*



- The random forest regression model outperforms both linear regression and decision tree regression, achieving a lower MSE of 151.79 and a higher R-squared value of 0.68.
- The lower MSE indicates that the random forest model provides more accurate predictions compared to the other models.
- The R-squared value of 0.68 suggests that approximately 68% of the variance in the target variable is explained by the features. The predictions made by the random forest model are closer to the actual values, indicating better performance.
- Overall, the random forest regression model demonstrates the best performance among the three models evaluated, followed by linear regression, while the decision tree regression model performs the poorest.
- These results highlight the importance of selecting an appropriate regression model based on the specific dataset and problem context to obtain accurate predictions.

V. RESULTS

- The analysis of sales distribution across various countries unveils intriguing patterns crucial for strategic decision-making. Notably, Canada and Poland emerge as dominant contributors, representing over half of total sales.
- Diverse markets like Iceland, Singapore, and Brazil also contribute significantly. The visualization through a donut pie chart offers a clear overview, aiding stakeholders in understanding the relative contributions of each country.
- Examining the distribution of unit prices provides further depth into customer preferences and market dynamics. The analysis reveals a clear preference for lower-priced products, with most unit prices clustered within a certain range.
- This understanding is crucial for pricing strategies and product offerings, ensuring alignment with customer preferences.
- ***"The associations between "Customer ID" and other variables exhibit generally low correlations, spanning from -0.20 to 0.15. These values suggest weak or negligible linear relationships."***
- ***"Similarly, the correlations between "Age" and "Salary" with other variables are also relatively low, ranging from -0.27 to 0.12, indicating weak linear associations."***
- Additionally, the utilization of predictive modeling techniques, including linear regression and random forest regression, offers actionable insights into identifying potential high-value customers.
- In conclusion, the project's analysis of sales data and predictive modeling provides valuable insights for driving business growth.
- Understanding sales distribution, unit price dynamics, and predictive analytics aids informed decisions to optimize strategies. Continuous monitoring and refinement of these strategies are essential for maintaining competitiveness in the global marketplace.

VI. CONCLUSION

- In conclusion, this project has provided valuable insights into predicting Customer Lifetime Value (CLV) using a dataset sourced from retail transactions. By leveraging advanced analytics and machine learning techniques, we have addressed key questions regarding the predictive power of historical transaction data, the distribution of sales across different countries, seasonal patterns in sales, unit price variability, and customer spending behavior.
- Our analysis reveals that Canada and Poland emerge as significant contributors to total sales, indicating the importance of targeting resources towards these markets. Additionally, we observe seasonal fluctuations in sales, with a notable surge during the holiday season, suggesting opportunities for tailored marketing campaigns. Furthermore, our examination of unit price distribution highlights customer preferences for moderately priced items, with pricing strategies influencing purchasing behavior.

- Regarding predictive modeling, our findings indicate that Random Forest Regression outperforms Linear Regression and Decision Tree Regression in predicting customer spending. However, all models demonstrate the significance of unit price and quantity purchased in determining total spending per transaction. These insights underscore the importance of accurately predicting CLV for strategic resource allocation, marketing effectiveness, and customer relationship management.

- Looking ahead, future research could explore additional factors influencing CLV prediction, such as customer demographics, purchase frequency, and customer engagement metrics. Moreover, incorporating external data sources, such as economic indicators or competitor analysis, could enhance the predictive accuracy of CLV models.

- Additionally, considering the dynamic nature of consumer behavior, continuous monitoring and refinement of predictive models are essential for adapting to changing market dynamics and maintaining competitiveness. Overall, this project serves as a foundational step towards optimizing CLV prediction and informing strategic decision-making for businesses seeking sustainable growth and profitability in today's competitive landscape.