**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.

Overall, understanding the distribution of sales across different countries provides valuable insights for customer lifetime value. By recognizing the countries with the highest sales percentages, companies can prioritize resources and investments to capitalize on their strongest markets while also identifying potential opportunities for growth in regions with lower sales contributions.

**Data Statistics:**

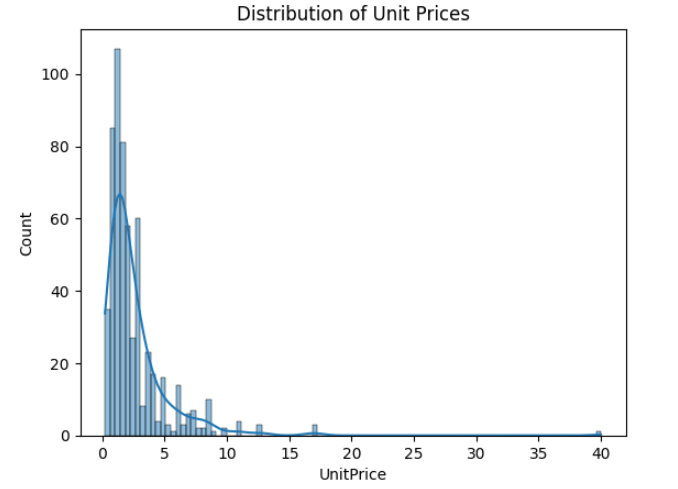
The mean quantity purchased is approximately 11.90, indicating the average quantity of items bought in each transaction.

The median quantity, at 12 units, signifies the middle value of the quantities purchased, with half of the transactions involving 12 units or fewer.

The mode quantity is also 12, suggesting that 12 units are the most frequently purchased quantity.

Percentile statistics further elucidate the distribution of quantities, with the 25th percentile at 6 units and the 75th percentile also at 12 units. This indicates that a significant portion of transactions involve quantities ranging from 6 to 12 units.

**Analysis of Unit Price Distribution:**

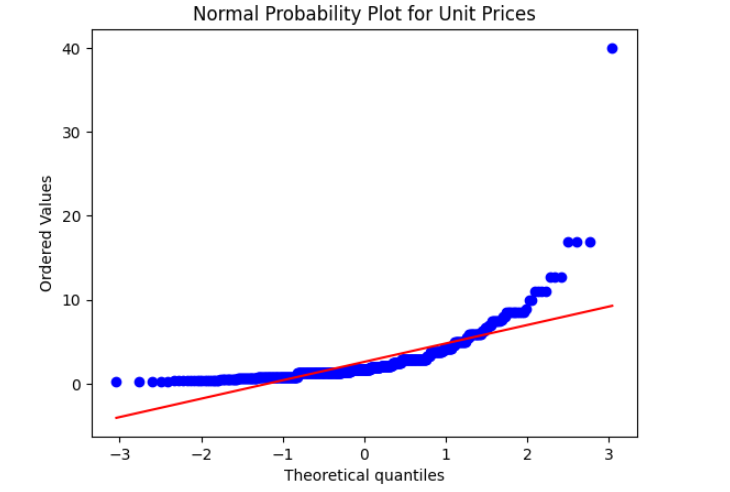


The distribution of unit prices in the dataset shows a clear trend towards lower-priced products. The majority of unit prices fall within the range of 0.21 to 16.95 unit price, with a peak occurrence around 1.25 unit price. As the unit price increases beyond 17, there is a significant decline in frequency 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.

**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.

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.

**Normality Test for Unit Prices:**

Shapiro-Wilk Test for Normality: The Shapiro-Wilk test is used to assess the normality of the distribution of unit prices. With a very low p-value of 2.517e-34, the test suggests that the distribution significantly deviates from a normal distribution. This indicates that the assumption of normality may not hold for the unit price data. This is what we concluded from normal probability plot where the blue dots deviate from the red line slightly, particularly at the higher end.

**Estimating Parameters**

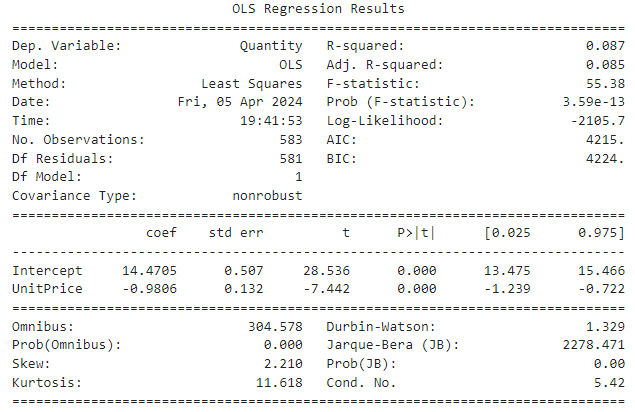
The sample mean of unit prices is 2.62 units, indicating the average price of items in the dataset.

The sample standard deviation of unit prices is 2.82 units, reflecting the variability or dispersion of prices around the mean.

**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.

**OLS Regression Analysis:**



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.

**Correlation Analysis:**

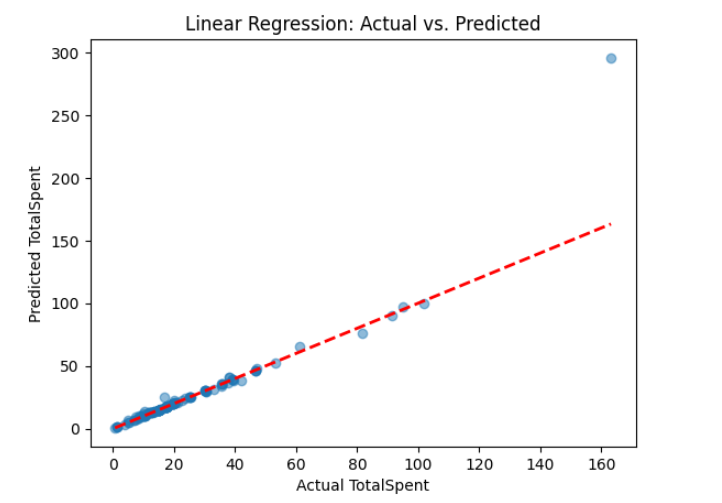
The correlation matrix shows a negative correlation of approximately -0.295 between unit prices and quantity purchased. This indicates that as unit prices increase, the quantity purchased tends to decrease, although the correlation is not very strong.

Overall, the analysis indicates that while the unit price distribution deviates from normality, there is a significant relationship between unit prices and the quantity purchased. Higher prices are associated with lower quantities purchased, suggesting that pricing strategies may influence customer purchasing behaviour.

**Machine Learning Models**

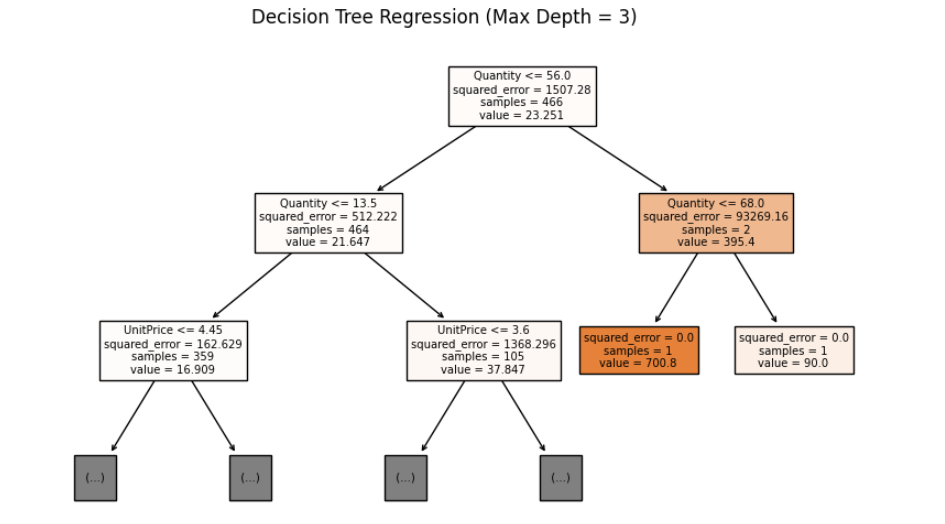
We have conducted predictive analysis using historical transaction data to pinpoint customers with the highest likelihood of making substantial purchases. Through feature engineering, we calculated the total amount spent per transaction and categorizes customers into different spending tiers. Subsequently, 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 utilizes 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.

Linear Regression:



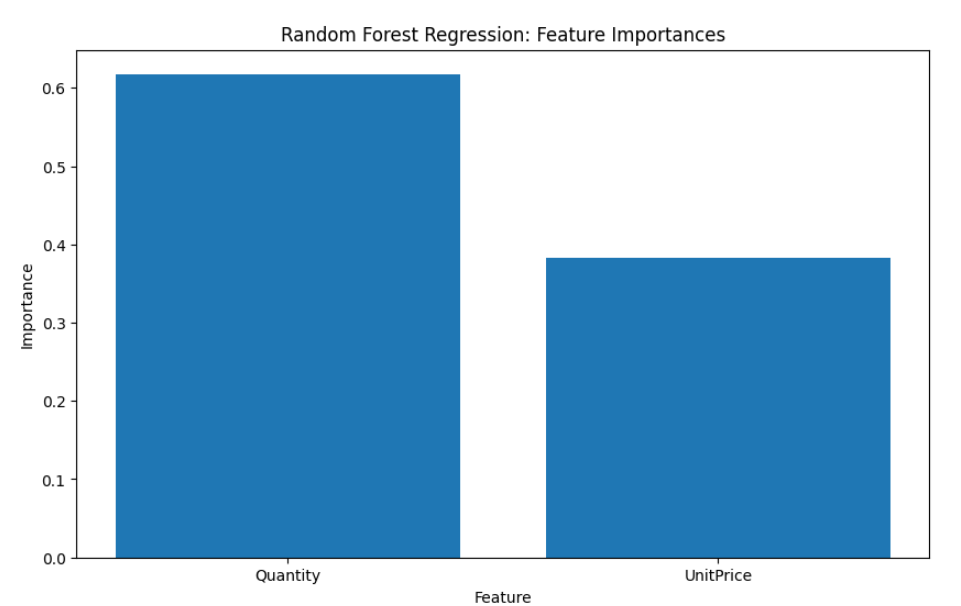
The linear regression model achieves a moderate performance with an MSE of 203.63 and an R-squared value of 0.58. The MSE measures the average squared difference between the actual and predicted values, indicating how well the model fits the data. An R-squared value of 0.58 suggests that approximately 58% of the variance in the dependent variable (Total Spent) can be explained by the independent variables (Quantity and Unit Price) in the model. The predictions made by the linear regression model are relatively close to the actual values.

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.

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.