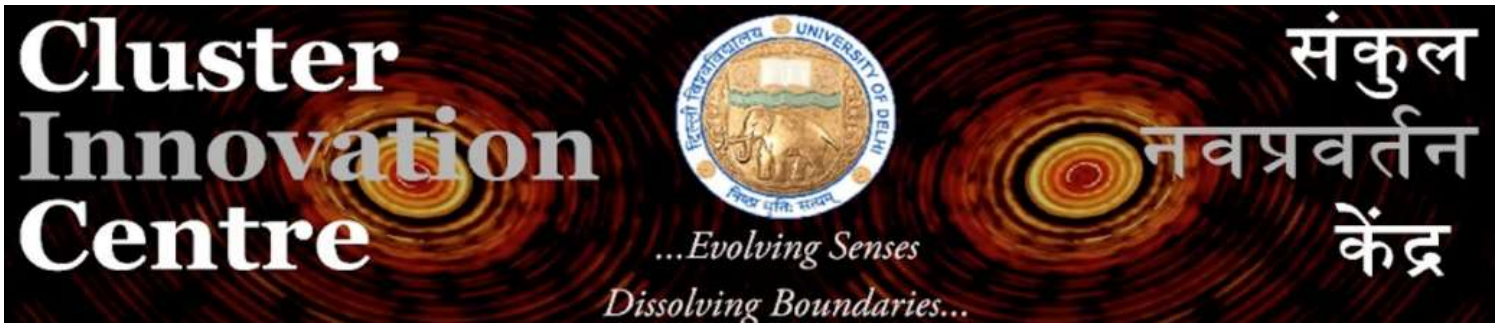


Customer Purchase Behaviour Analysis and Visualization

by

Tushitaa Narayan Ojha

Cluster Innovation Centre

University of Delhi

At

Institute of Informatics and Communication,

University of Delhi, South Campus

Under the mentorship and supervision of

Dr Amit Pundir

Associate Professor

Maharaja Agrasen College,

University of Delhi

Nitisha Aggarwal

Senior Research Fellow

Institute of Informatics and Communication,

University of Delhi, South Campus

CERTIFICATE OF ORIGINALITY

This is to certify that the report on the project titled “Customer Purchase Behaviour Analysis and Visualization” submitted by Tushitaa Narayan Ojha for the paper “Industrial Mini Project” is the result of my own independent and original work. All the work presented in this report has been conducted solely by me. I have made sure to properly acknowledge all sources from which ideas and content have been drawn.

I confirm that this project is free from any form of plagiarism and has not been submitted elsewhere for any academic or professional purpose. Every external reference used in the report has been appropriately cited and credited.

I understand the importance of academic honesty and that any breach, including plagiarism, will result in strict penalties and possible disciplinary action by the University.

Dr Amit Pundir

Mrs. Nitisha Aggarwal

Tushitaa Narayan Ojha

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the Institute of Informatics and Communication, University of Delhi, South Campus, for providing the resources and environment necessary for the successful completion of this project.

I am deeply indebted to Dr Amit Pundir and Dr Geetika Jain Saxena, Associate Professor at Maharaja Agrasen College, University of Delhi, for their mentorship and supervision throughout this endeavour. Their guidance and expertise were crucial in shaping the direction and outcome of this project.

My heartfelt thanks also go to Nitisha Aggarwal, Senior Research Fellow at the Institute of Informatics and Communication, University of Delhi, South Campus for her valuable insights and encouragement, which greatly contributed to the progress and completion of this work.

The contributions of these esteemed individuals have been invaluable, and their support has significantly enriched the quality of this work.

Tushitaa Narayan Ojha

ABSTRACT

The "Customer Purchase Behaviour Analysis and Visualization" project aims to empower businesses with actionable insights into customer behaviour by leveraging advanced data analytics and visualization techniques. This project involves a systematic approach, beginning with data preparation to ensure accuracy and reliability. Detailed analysis uncovers patterns and trends in customer purchases, segments customers based on their buying habits, and evaluates product performance. A key aspect of the project is calculating Customer Lifetime Value (CLV) to assess long-term customer potential and building predictive models to forecast future sales trends.

The analysis utilizes data from the UCI Machine Learning Repository, comprising comprehensive online retail transaction records. Techniques such as Exploratory Data Analysis (EDA) and clustering (e.g., K-means) are employed to uncover critical insights into purchase behaviours, customer segmentation, and product trends. The findings are presented in an interactive dashboard and a detailed report, offering clear recommendations to enhance customer engagement, optimize marketing strategies, and improve inventory management.

This project highlights the transformative potential of data analytics in driving business growth and fostering a deeper understanding of customers. It provides a framework for turning raw data into valuable insights, enabling businesses to remain competitive in a dynamic market.

Keywords: Customer behaviour analysis, data visualization, customer segmentation, CLV, predictive modelling, business strategy, data analytics, EDA, online retail, K-means clustering.

Content

1. INTRODUCTION

This project, titled *Customer Purchase Behaviour Analysis and Visualization*, focuses on transforming raw customer data into meaningful insights that businesses can use to better understand their customers. The key objective of this project is to analyse customer purchasing patterns and trends. With the ever-increasing amount of customer data, businesses often find it challenging to extract actionable insights. My goal was to bridge this gap using data analytics techniques, providing businesses with information that helps them make data-driven decisions, improve customer engagement, and refine their marketing strategies.

1.1 Background and Context

The "Customer Purchase Behaviour Analysis and Visualization" project was created to help businesses better understand how their customer's shop. In today's world, where data is a powerful tool, companies collect a lot of information about customer purchases. However, this data is often not fully utilized, and important insights are missed.

This project aims to change that by carefully analysing the data to reveal patterns in customer behaviour, such as what products people buy, how much they spend, and how often they shop. The data used for this analysis comes from the UCI Machine Learning Repository, which includes detailed records of online retail transactions.

A key part of the project is grouping customers based on their shopping habits, which allows businesses to target different types of customers more effectively. Additionally, the project includes predicting future sales trends, helping companies plan better and make smarter decisions about inventory and promotions.

Overall, this project is about turning raw data into useful information that can guide businesses in improving their strategies, understanding their customers better, and staying ahead in a competitive market.

1.2 Scope and Objective

The "Customer Purchase Behaviour Analysis and Visualization" project aims to provide businesses with a deeper understanding of their customers through detailed data analysis and visualization. The project involves several key activities, starting with the preparation of data from the UCI Machine Learning Repository, ensuring that

it is clean and ready for analysis. The analysis focuses on exploring patterns and trends in customer purchases, segmenting customers based on their buying behaviour, and assessing the performance of products and categories. Additionally, the project calculates Customer Lifetime Value (CLV) to evaluate the long-term value of customers and builds predictive models to forecast future sales. The ultimate goal is to present these findings through a comprehensive report and an interactive dashboard, offering clear insights and actionable recommendations to help businesses refine their strategies, improve customer engagement, and drive growth in a competitive market.

2. FORMULATION OF THE PROBLEM

2.1 Problem Statement

In today's fast-paced digital economy, many businesses struggle to fully understand their customers' purchase behaviour, which can lead to missed opportunities for growth and customer retention. Without clear insights into what drives customer purchases, companies often find themselves unable to tailor their strategies effectively, resulting in lost revenue and a lack of customer loyalty.

This project seeks to address this pressing issue by providing a thorough analysis of customer purchase behaviour for an online retail store. The project is structured in several phases, beginning with the crucial task of data cleaning, which includes addressing missing values, eliminating duplicates, and ensuring that data types are correctly formatted for analysis. This foundation allows for an accurate Exploratory Data Analysis (EDA), where key variables such as Unit_Price, Quantity, and ProductDescription will be examined. The project will also involve identifying outliers and visualizing purchase trends over time, utilizing auto EDA tools.

A significant aspect of the project is customer segmentation, which will use clustering techniques like K-means to group customers based on their purchasing habits. By understanding these segments, businesses can tailor their marketing strategies more effectively. The project will also include a detailed analysis of product performance, helping businesses identify top-selling products and trends over time.

Furthermore, the calculation of Customer Lifetime Value (CLV) will provide businesses with critical insights into the long-term value of their customers, enabling more informed decision-making. The project will also forecast future sales using historical data, ensuring

that businesses are prepared for future trends. Finally, the findings will be compiled into a comprehensive report and dashboard, making it easy for businesses to implement the recommendations and improve their overall performance.

2.2 Software / Tools Used

Python: Python is a programming language that lets you work more quickly and integrate your systems more effectively. Python is used successfully in thousands of real-world business applications around the world, including many large and mission critical systems.

Colab: Colab is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It supports multiple programming languages including Python, which is one of the most widely used programming languages in the data science and scientific computing communities. It is often used for data analysis, numerical simulation, machine learning, and more.

Pandas - Pandas is a software library written for the Python programming language for data manipulation and analysis.

Power BI – Power BI is a powerful business analytics tool developed by Microsoft that enables users to visualize and share insights from their data. It provides interactive data

visualization, robust reporting, and advanced data analysis capabilities, allowing businesses to transform raw data into meaningful insights.

With Power BI, users can connect to a wide variety of data sources, including databases, spreadsheets, cloud services, and more. The platform supports data cleaning, transformation, and modelling, making it easier to prepare data for analysis. One of Power BI's key features is its drag-and-drop interface, which allows users to create visually compelling reports and dashboards without needing extensive technical expertise. These visualizations can be customized and tailored to suit specific business needs, and they can include a range of charts, graphs, maps, and tables.

Power BI also offers real-time data updates, enabling users to monitor business performance as it happens. Reports and dashboards can be shared across an organization, ensuring that all stakeholders have access to the latest insights. Additionally, Power BI integrates seamlessly with other Microsoft products, such as Excel, Azure, and Office 365, enhancing its functionality and ease of use.

2.3 Methodology

- **Data Cleaning**

The first step in the data analysis process was to ensure that the dataset was clean and ready for further analysis. This involved several crucial tasks:

1. Removal of Duplicated Rows:

   The dataset was examined for any duplicated entries that could skew the analysis results. Duplicate rows are often caused by errors in data collection or data entry, leading to misleading conclusions if not addressed. In this case, all duplicated rows were identified and removed from the dataset to ensure the accuracy and reliability of the analysis.

2. Data Type Correction:

   It is essential for each column in the dataset to have the correct data type to avoid errors during data processing and analysis. For example, columns containing date and time information were converted to the appropriate datetime format, numerical columns were verified and converted to numeric types, and categorical variables were appropriately labelled. This step was vital in ensuring that the data was correctly interpreted by the analysis tools used in the subsequent steps.

3. Handling Missing Values:

Missing values can occur due to various reasons, such as errors in data collection or incomplete records. To address this issue, different strategies were considered based on the nature of the missing data:

- Removal of Records: In cases where missing values were few and scattered, the affected rows were removed, assuming that their absence would not significantly impact the analysis.
- Imputation: For columns where, missing values were more substantial, imputation techniques were employed. These techniques included:
    1. Mean/Median Imputation: For numerical data, missing values were replaced with the mean or median of the respective column, depending on the data distribution.
    2. Mode Imputation: For categorical data, missing values were filled in with the mode, or the most frequently occurring value in the column.
    3. Forward/Backward Fill: In time series data, missing values were filled using forward or backward filling techniques, ensuring continuity in the dataset.

This data cleaning process was critical in ensuring that the dataset was of high quality, thereby enhancing the accuracy and validity of the analysis. By carefully removing duplicates, correcting data types, and addressing missing values, the dataset was made ready for the subsequent stages of data exploration and modelling.

- **Curve Fitting**

1. Linear Regression

Curve fitting is a mathematical technique used to analyse the relationship between two variables by fitting a curve that best represents the observed data points. Unlike linear regression, which assumes a straight-line relationship, curve fitting allows for more complex relationships by fitting curves such as polynomials, exponentials, or other functional forms to the data. The goal is to capture the overall trend and patterns that may not be well-represented by a straight line.

The shape and equation of the fitted curve indicate the nature of the relationship between the variables. For example, a positive curve that rises indicates a direct relationship, where increases in one variable are associated with increases in the other,

while a downward-sloping or inverted curve suggests an inverse relationship. Curve fitting enables analysts to better represent nonlinear relationships, where changes in one variable may have varying effects on the other.

The quality of the curve's fit can be evaluated using statistical measures like the coefficient of determination (R-squared), which shows the proportion of variation in the dependent variable explained by the fitted curve. A higher R-squared value suggests that the curve closely follows the data points, indicating a strong relationship and effective model performance.

2. **Polynomial Regression**

Polynomial curve fitting is an extension of basic curve fitting that models the relationship between two variables by fitting a non-linear curve. Instead of a straight line, as in linear models, polynomial curve fitting uses higher-order polynomials to capture more intricate patterns in the data, making it well-suited for situations where the relationship between variables is not linear.

In customer purchase history analysis, polynomial curve fitting uncovers complex trends between features like Quantity and Total Price or purchase behaviours over time. It can identify cyclical buying patterns or diminishing returns on purchases as quantities increase, insights that simpler linear methods might overlook.

While higher-degree polynomials offer greater flexibility to fit the data, they also increase the risk of overfitting, where the curve fits the noise in the data rather than capturing meaningful trends. To prevent this, techniques like cross-validation are employed to strike a balance between model complexity and accuracy. Metrics like R-squared and RMSE are used to evaluate how well the fitted curve explains the variation in the data.

This approach allows for a more nuanced understanding of non-linear relationships, making it valuable for uncovering deeper insights in various domains, including customer behaviour analysis.

- **Heatmap**

Heatmaps can play a valuable role in our analysis of customer sales data. By utilizing heatmaps, we can visually represent and compare various key indicators such as purchase trends, product popularity, and customer behaviour. For example, we can

create heatmaps to display patterns in sales volume, product categories, geographical sales distribution, and customer segmentation.

These heatmaps allow us to easily identify areas of high or low sales activity, helping us highlight trends, popular products, or regions with more engaged customers. By visually encoding the data in this manner, heatmaps provide an intuitive way to spot correlations, trends, and outliers across different dimensions of sales performance.

Furthermore, heatmaps can aid in pinpointing specific areas for further analysis or strategic interventions, based on observed trends. Whether investigating product demand, price sensitivity, or customer demographics, heatmaps provide a powerful visual tool to synthesize complex sales data and support data-driven decision-making for better sales strategies.

- **Scatter Plot**

  Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. We know that correlation is a statistical measure of the relationship between the two variables relative movements. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the closer the points will touch the line.

  The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

1. Positive Correlation
2. Negative Correlation
3. No Correlation

Positive Correlation: When the points in the graph are rising, moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another.

Negative Correlation: When the points in the scatter graph fall while moving left to right, then it is called a negative correlation. It means the values of one variable are decreasing with respect to another.

No Correlation: When the points are scattered all over the graph and it is difficult to conclude whether the values are increasing or decreasing, then there is no correlation between the variables.

- **Quadrant Analysis**

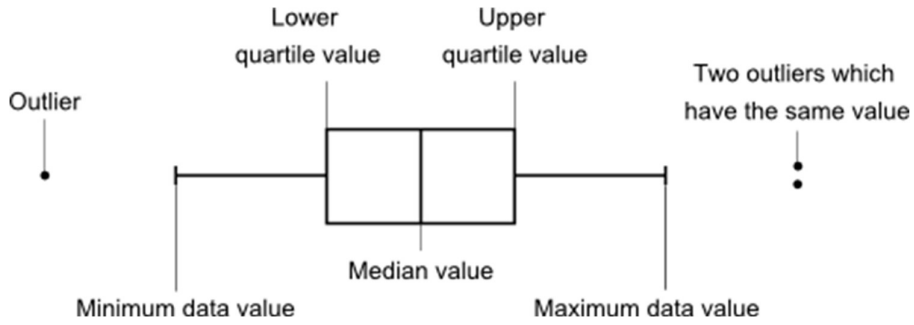
  Quadrant analysis is a visual technique used to categorize data points into four segments based on their position relative to the mean of two variables. In a scatter plot, the graph is divided into quadrants using vertical and horizontal lines at the average values of each axis. Each quadrant provides insights into different data patterns:

1. Top Right (High-High): Points in this quadrant represent cases with high values for both variables, indicating strong performance or high value.
2. Bottom Right (High-Low): This area includes instances where one variable is high while the other is low, often suggesting bulk or high-volume but low-cost outcomes.
3. Top Left (Low-High): Represents low-volume but high-value cases, such as luxury or specialized items.
4. Bottom Left (Low-Low): Shows cases where both variables are low, typically reflecting low-value, low-priority items.

   Quadrant analysis helps identify trends, correlations, and key focus areas, making it an effective tool for decision-making and strategy formulation across various fields.

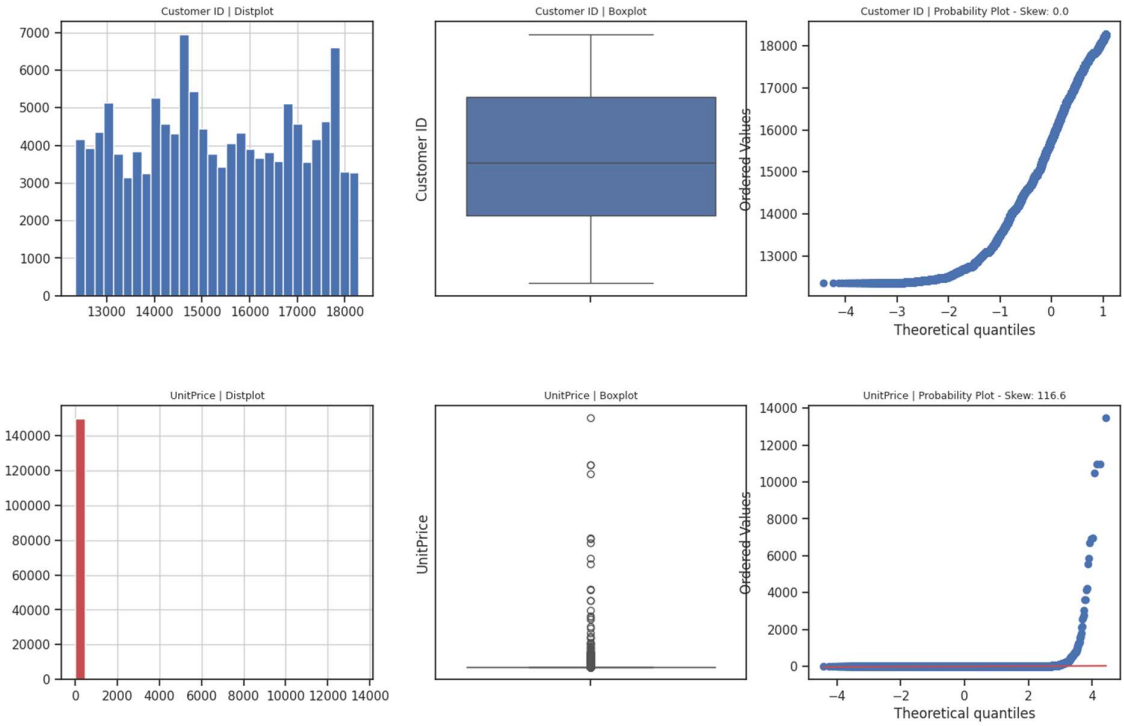
- **Box Plot**

  A box and whisker plot or diagram (otherwise known as a boxplot), is a graph summarising a set of data. The shape of the boxplot shows how the data is distributed and it also shows any outliers. It is a useful way to compare different sets of data as you can draw more than one boxplot per graph. The line splitting the box in two represents the median value. This shows that 50% of the data lies on the left-hand side of the median value and 50% lies on the right-hand side. The left edge of the box represents the lower quartile; it shows the value at which the first 25% of the data falls up to. The right edge of the box shows the upper quartile; it shows that 25% of the data lies to the right of the upper quartile value. The values at which the horizontal lines stop at are the values of the upper and lower values of the data. The single points on the diagram show the outliers.



3. RESULTS

Exploratory Data Analysis



1. Distribution of Customer IDs

The distribution of Customer IDs spans a broad range, indicating that no single group of customers is overrepresented. The even spread suggests that the dataset likely reflects a diverse customer base, free from any obvious biases toward specific customer groups. This balanced

distribution is ideal for generating unbiased insights into customer behaviour and allows for generalizable results.

2. **Distribution of Unit Prices**

The distribution of Unit Prices is highly skewed, with most values clustered near zero. This reinforces the observation that the majority of products sold are inexpensive, while a few high-cost items create a long tail on the right-hand side of the distribution. Such a pattern is typical in retail, where low-cost items dominate, and high-cost items are sold less frequently. The skewed nature of the data may require transformation or specialized handling to ensure accurate analysis, especially in predictive modelling tasks.

3. **Box Plots**

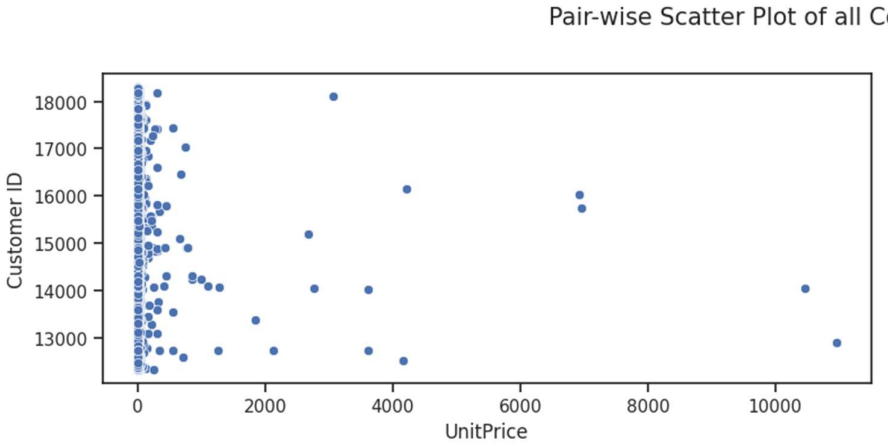
Customer ID Box Plot: The box plot for Customer ID shows a relatively even spread, with no significant outliers. This indicates that the dataset covers a wide range of customers fairly evenly, supporting the conclusion that the customer base is diverse.

Unit Price Box Plot: The box plot for Unit Price highlights the presence of outliers on the higher end, with most prices concentrated near the lower end. This further emphasizes the skewed nature of the Unit Price distribution and underscores the importance of addressing outliers in the analysis.

4. **Probability Plots**

Customer ID Probability Plot: The probability plot for Customer ID aligns well with a normal distribution, suggesting that the data is evenly distributed across the customer base. This symmetry implies that Customer ID is likely a uniformly assigned identifier without any inherent biases.

Unit Price Probability Plot: The probability plot for Unit Price diverges significantly from normality, reflecting the earlier observation of a heavily skewed distribution. This skewness is indicative of a retail environment where most products are relatively inexpensive, but a few high-priced items create a long tail.



Scatter Plot Analysis: Customer ID vs. Unit Price

The scatter plot of Customer ID versus Unit Price offers insights into the pricing and transaction distribution:

Concentration of Low-Priced Items

Observation: Most UnitPrice values cluster at the lower end, indicating a high volume of low-cost transactions.

Interpretation: The business model likely relies on selling inexpensive items in large quantities, typical of retail environments focusing on everyday goods.

Presence of Outliers

Observation: A few high UnitPrice values create a "long tail" on the right.

Interpretation: These outliers may represent premium products or potential data entry errors. Further investigation is needed to confirm their validity.

Impact on Analysis

Observation: The skewed distribution of UnitPrice affects statistical analyses assuming normality.

Interpretation: Skewness may lead to biased results in models assuming normal distribution, with the mean UnitPrice potentially inflated by outliers.

Business Implications

Observation: Sales are predominantly driven by low-cost items, with occasional high-value transactions.

Interpretation: The business likely uses a high-volume, low-margin strategy, with a few high-margin products enhancing profitability. Understanding this distribution is crucial for pricing, inventory, and marketing strategies.

Need for Transformation

Observation: The UnitPrice distribution's skewness may necessitate transformation for more accurate statistical modeling.

Interpretation: A log transformation can reduce skewness, normalize the data, and improve model accuracy by compressing extreme values and stabilizing variance.

Log Transformation can address skewed data issues

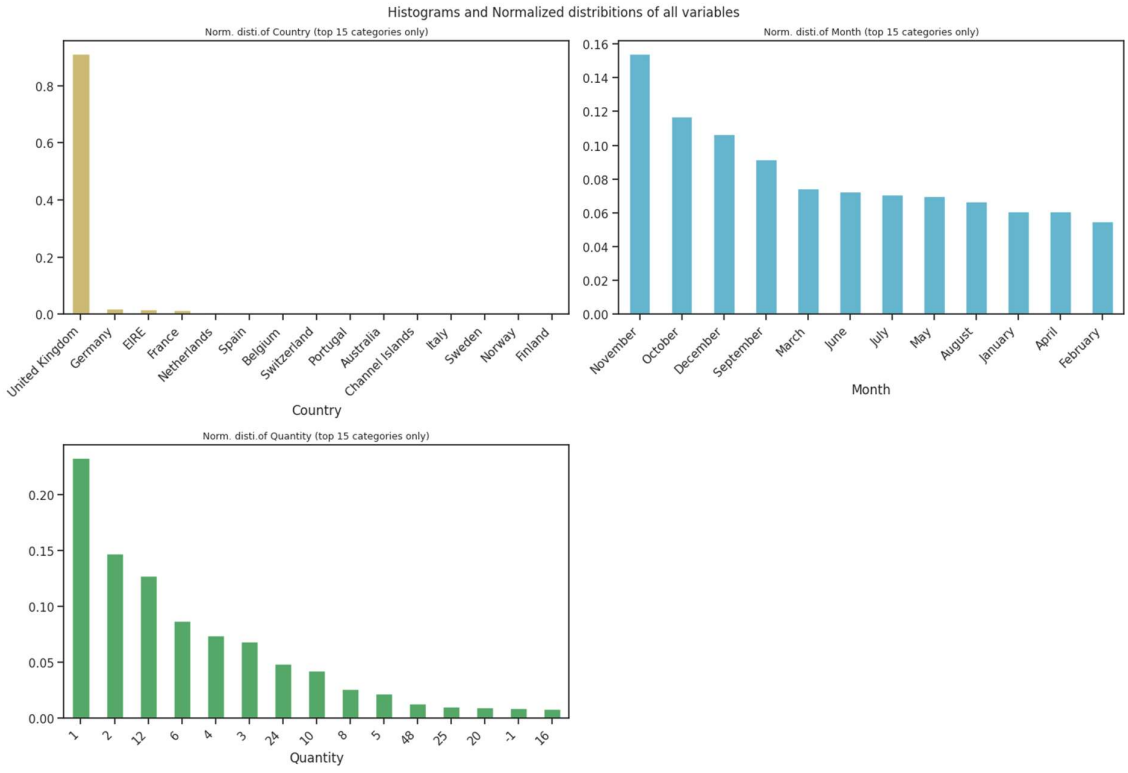
Reduces Skewness: Compresses larger values, leading to a more symmetrical distribution.

Stabilizes Variance: Mitigates heteroscedasticity by making variance more uniform across the dataset.

Compresses the Scale: Lessens the impact of extreme values on overall analysis.

Linearizes Exponential Growth: Converts exponential trends into linear relationships, simplifying analysis with linear models.

Country Distribution (Top 15 Categories)



Histogram Analysis

1**. Country Distribution (Top-Left)**

Observation: The United Kingdom overwhelmingly dominates the dataset, with the majority of transactions originating from this country. Other countries, such as Germany, EIRE, France, and the Netherlands, are represented but in significantly smaller proportions.

Interpretation: Insights and analyses derived from this dataset will likely be skewed toward the behaviour and trends prevalent in the United Kingdom. This dominance suggests that findings

may not generalize well to other countries unless further investigation into regional behaviours is conducted.

2. Month Distribution (Top-Right)

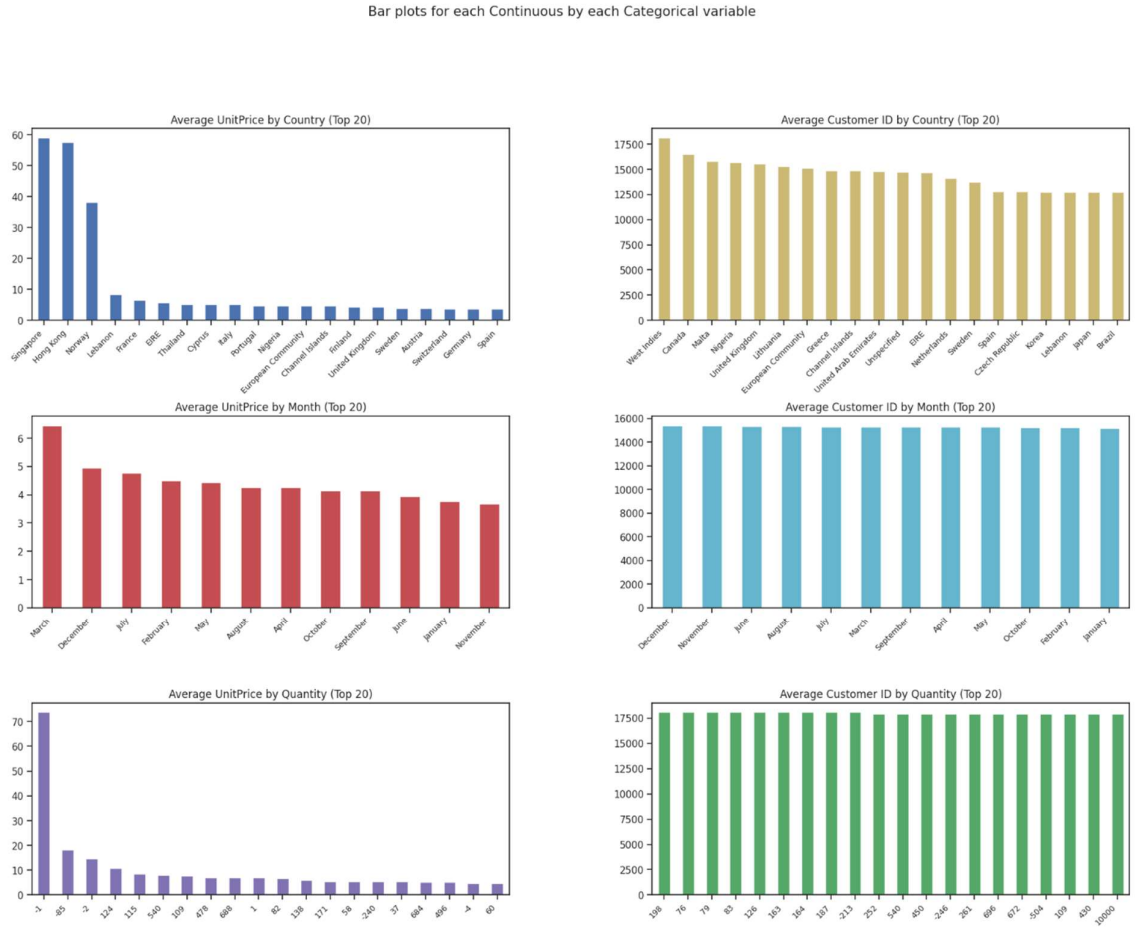
Observation: The month of November shows the highest number of transactions, followed by October and December. This indicates a spike in sales activity toward the end of the year, likely due to holiday shopping or other seasonal factors.

Interpretation: The data reflects typical retail trends, where sales volume increases during the holiday season. Conversely, January and February show the lowest activity, indicating a post-holiday decline in transactions. Businesses may leverage this seasonal pattern for promotional planning and inventory management during high-demand periods.

3. Quantity Distribution (Bottom-Left)

Observation: Most transactions involve the purchase of 1 or 2 units of a product, highlighting a trend of smaller purchases. Instances of larger quantity purchases are present but occur less frequently.

Interpretation: The data suggests that customers tend to make smaller, individual purchases rather than bulk buying. This insight is valuable for inventory planning and may influence promotional strategies aimed at encouraging bulk purchases or offering discounts on larger quantities.



Bar Plot Analysis

4. Average Unit Price by Country (Top-Left)

Observation: Countries like Singapore and Hong Kong exhibit the highest average unit prices, indicating that products sold in these regions are of higher value, or that these markets have higher price points.

Interpretation: This suggests regional differences in pricing strategies or the types of products offered in these countries. Markets like Lithuania, Israel, and the United Arab Emirates also demonstrate higher average prices, which may be driven by product demand or market positioning.

5. Average Customer ID by Country (Top-Right)

Observation: Canada, USA, and Hong Kong have higher average Customer IDs, which may indicate a more established customer base or that customer accounts in these regions were created earlier.

Interpretation: Despite the United Kingdom having the highest transaction volume, it shows a lower average Customer ID. This could imply that the UK has a broader, newer customer base, as more recent customers tend to have lower ID values, indicating ongoing customer acquisition.

6. Average Unit Price by Month (Middle-Left)

Observation: March and December show the highest average unit prices, which could be attributed to specific sales events or the launch of higher-value products during these months.

Interpretation: The lower average unit price in November, despite having the highest transaction volume, could point to discounting or promotional activities during this period, such as Black Friday sales, where high transaction volumes occur at lower price points.

7. Average Customer ID by Month (Middle-Right)

Observation: The average Customer ID remains consistent across all months, indicating that customer engagement and registration are stable throughout the year.

Interpretation: This consistency suggests that there are no significant seasonal fluctuations in new customer acquisition, and the customer base grows steadily over time.

8. Average Unit Price by Quantity (Bottom-Left)

Observation: Transactions involving larger quantities tend to have lower average unit prices, which aligns with typical bulk-buying behaviour who ere discounts are applied.

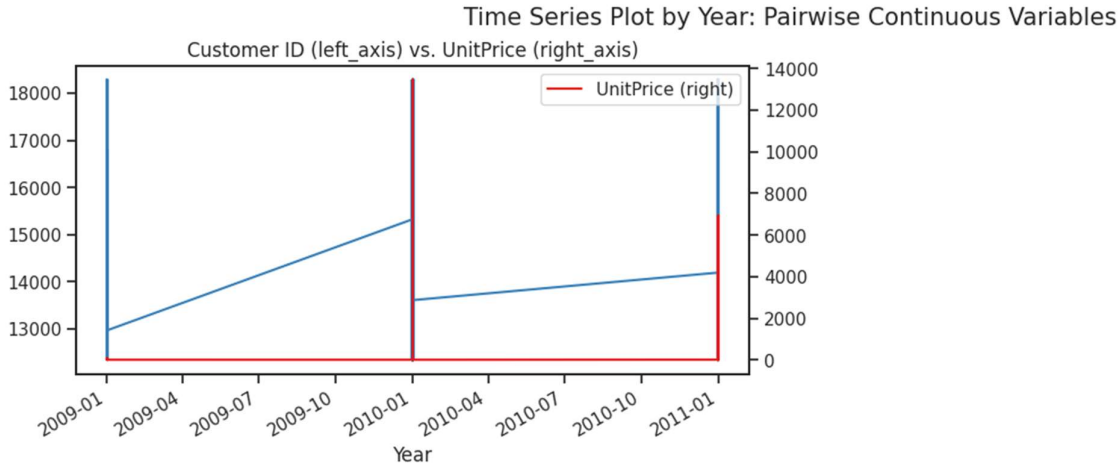
Interpretation: Smaller quantity purchases show higher average unit prices, suggesting that higher-margin products or single-item purchases dominate smaller transactions. This relationship is essential for understanding customer buying behaviour and could help businesses optimize pricing strategies for different purchase volumes.

9. Customer ID and Unit Price Distributions

Observation:

1. Customer ID Distribution: The distribution of Customer IDs is relatively uniform, indicating a balanced representation across customers without significant skew.
2. Unit Price Distribution: The Unit Price distribution is highly skewed to the right, with many low-priced transactions and a few at extremely high price points, indicating the presence of outliers.

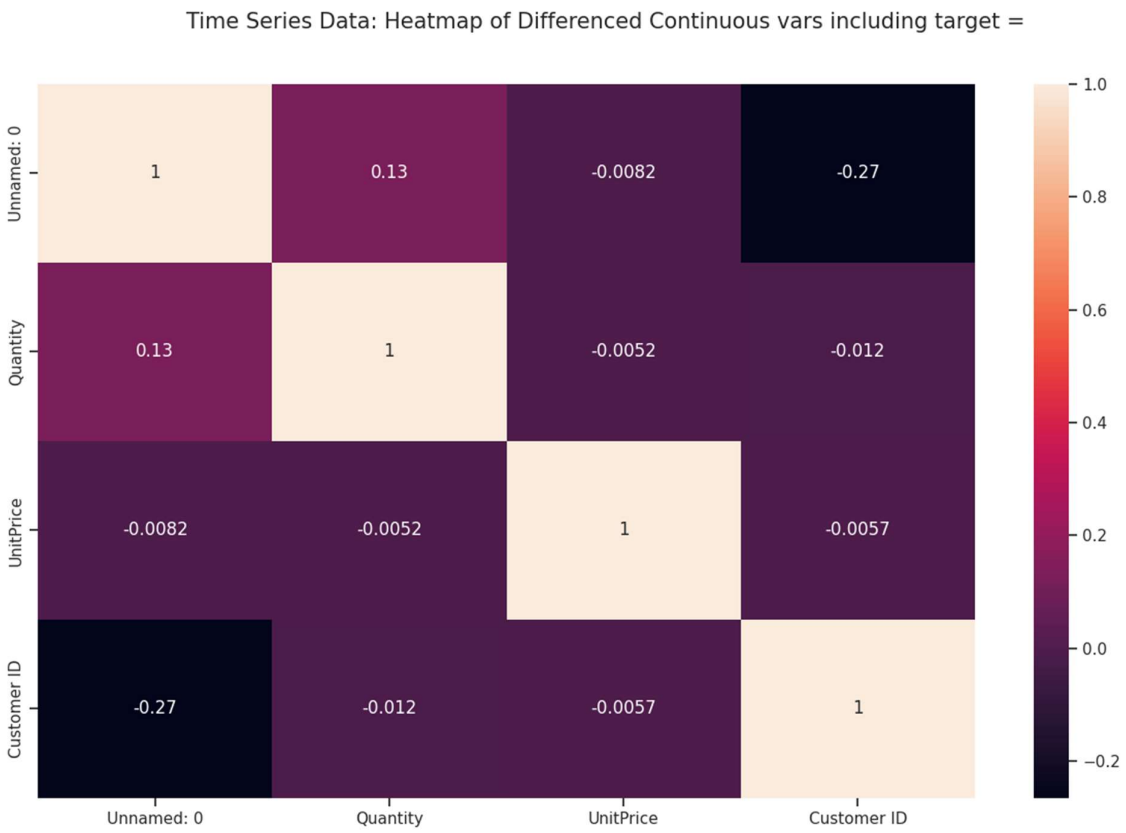
Interpretation: The uniformity in Customer IDs suggests no dominant customer group, while the skewness in Unit Price highlights the need for handling outliers in analysis, as the majority of transactions are at lower price points, but higher-priced items can disproportionately impact average values.



Analysis of the Time Series Plot by Year: Pairwise Continuous Variables

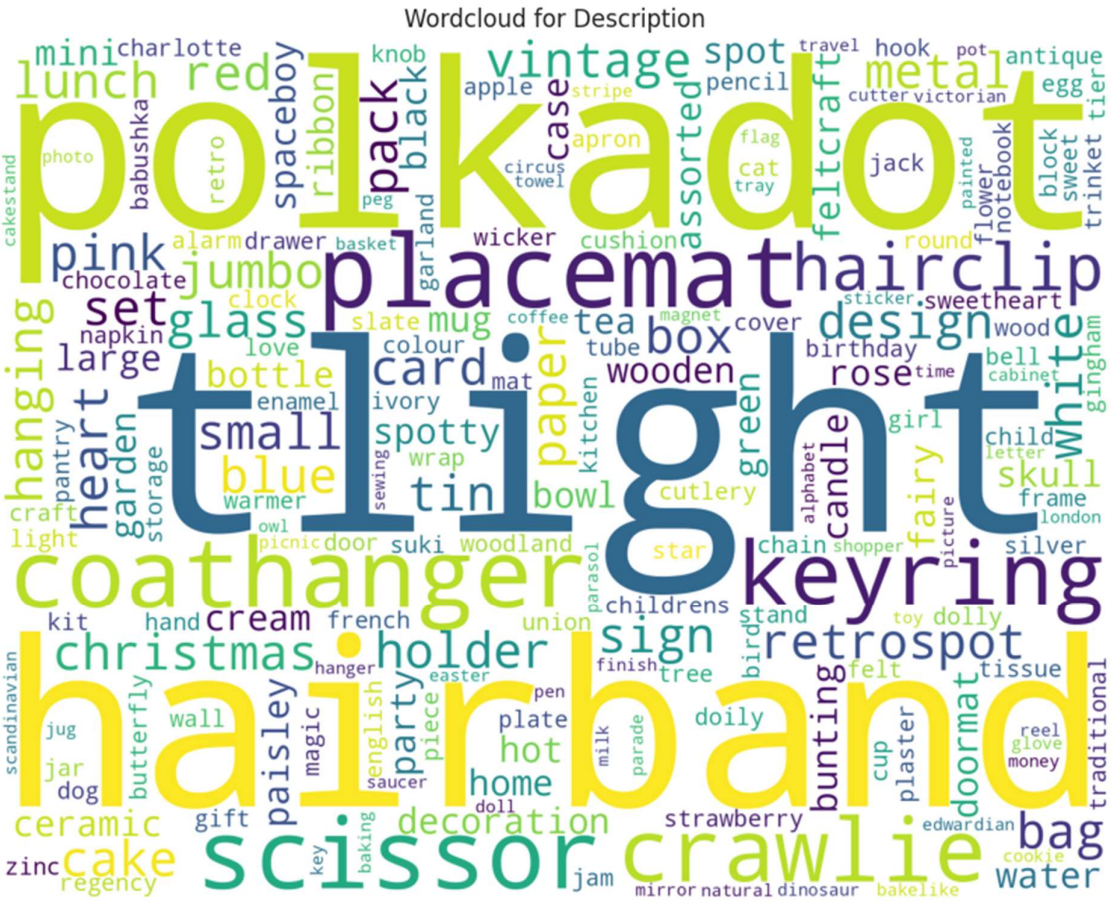
1. Time Series Plot by Year: Pairwise Continuous Variables

    a. Description: This plot tracks the relationship between Customer ID and UnitPrice across multiple years. The left axis measures Customer ID, and the right axis tracks UnitPrice. The lines on the plot indicate how these two variables evolve over time.
    b. Key Observations:
        i. The plot highlights fluctuations in customer interactions and unit prices over the years.
        ii. Some years display more pronounced shifts in Customer ID and UnitPrice, with potential seasonal spikes or abrupt changes that might signify business trends or anomalies.
    c. Implications:
        i. Stable vs. Fluctuating Pricing: The trends in UnitPrice could indicate periods of stable pricing or suggest that specific years saw significant price changes, which may be tied to economic conditions or promotional activities.
        ii. Customer Purchase Behaviour: The changes in Customer ID might reflect growth or decline in customer base or could suggest periods of high engagement (e.g., during seasonal sales) or low activity.



2. Heatmap of Differenced Continuous Variables:

- Description: The heatmap illustrates the correlation between various continuous variables in the dataset, including Quantity, UnitPrice, Customer ID, and one unnamed variable. Each cell in the heatmap represents the strength and direction of the correlation between two variables, with colour intensity showing the degree of correlation.
- Key Observations:
    1. Quantity and Customer ID: A weak positive correlation exists between Quantity and Customer ID, suggesting that there isn't a strong relationship between the number of items purchased and the specific customers making the purchases.
    2. Quantity and UnitPrice: Similarly, Quantity and UnitPrice have a weak positive correlation, indicating that the price of an item does not significantly influence the number of items purchased.



3. Word Cloud for Product Descriptions:

- Description: The word cloud showcases the most frequently occurring terms in the product descriptions. The size of each word represents its frequency in the dataset, with larger words indicating that they appear more often in product descriptions.
- Key Observations:
    1. Words like "polka dot," "hairband," "light," and "coat hanger" stand out as the most prominent, suggesting that products associated with these words are among the most popular or frequently purchased.
    2. Other common words such as "vintage," "set," and "heart" also appear frequently, hinting at broader product categories or features that attract customer interest.

Customer centric Analysis

Customer Lifetime Value (CLV) estimates the total revenue a business can expect from a customer throughout their relationship.

Definition:
CLV is calculated by multiplying the average purchase value, purchase frequency, and average customer lifespan, representing the total profit a business can expect from a customer over time.

CLV helps businesses:

- Identify valuable customers.
- Optimize marketing and retention strategies.
- Make informed decisions on pricing, product development, and customer service.

CLV Calculation Steps:

1. **Average Purchase Value:**

Average Purchase Value = Total Revenue\ Number of Purchases Total Revenue

2. **Purchase Frequency Rate:**

Purchase Frequency Rate= Number of Purchases\ Number of Unique Customers

3. **Customer Value:**

Customer Value = Average Purchase Value × Purchase Frequency Rate

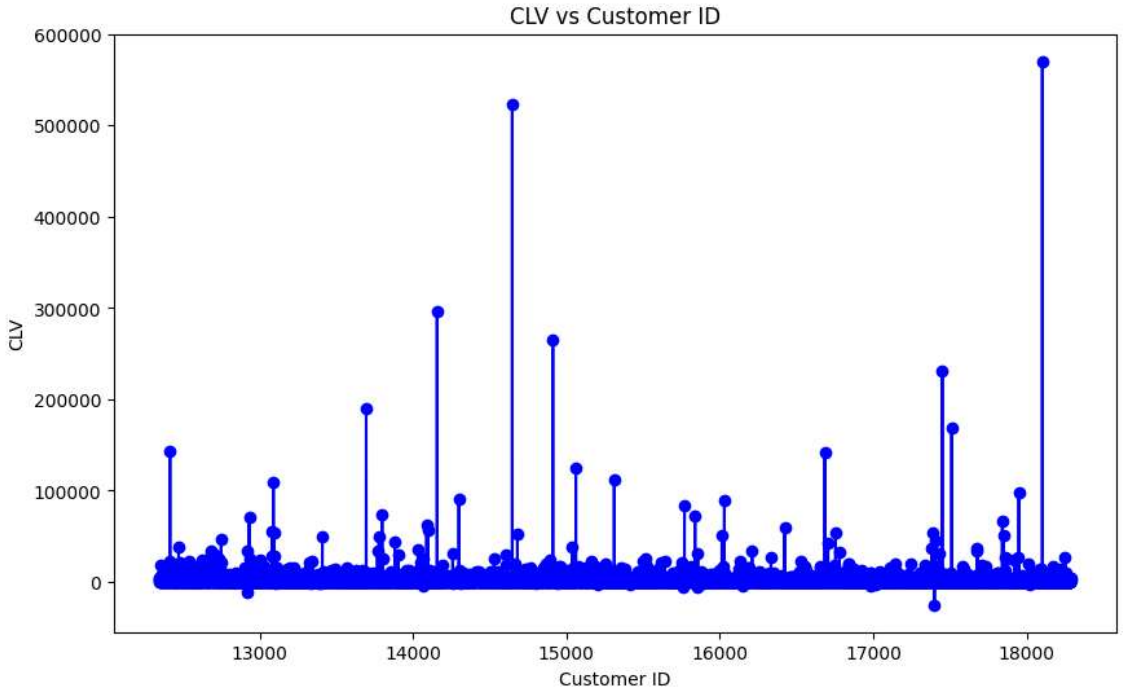
4. **Average Customer Lifespan:**

Average Customer Lifespan = Sum of Customer Lifespans/Number of Customers

5. **CLV:**

CLV = Customer Value × Average Customer Lifespan

.



This means each customer is expected to generate £1,500 in revenue over their lifetime with the business.

CLV helps businesses determine how much to spend on customer acquisition and retention, optimize marketing strategies, and focus on long-term profitability by identifying and nurturing high-value customers.

The graph titled "CLV vs Customer ID" illustrates the Customer Lifetime Value (CLV) for various customers, showing that a few customers contribute significantly more value to the company compared to others. These large spikes in the graph represent customers whose CLV is exceptionally high, indicating that they generate substantial revenue for the business.

From the provided table, we can identify the top 10 customers with the highest CLV, such as those with Customer IDs 18102, 14646, and 14156, who have CLVs of £570,380.61, £523,342.07, and £296,063.44, respectively. These customers not only bring in substantial total revenue but also have high average revenue per transaction, with Customer ID 18102 having an average of £543.22 per transaction.

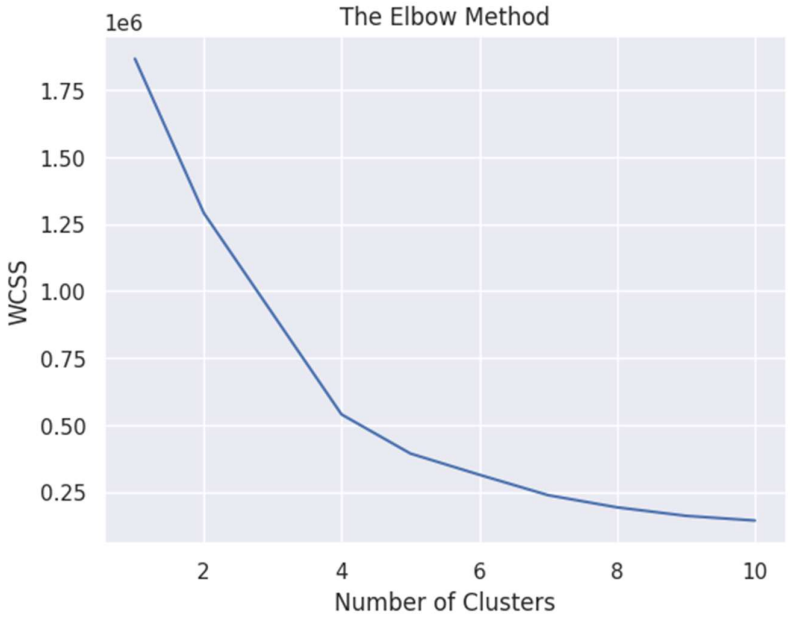
These insights suggest that these top customers are vital to the company's revenue stream. By focusing on retaining and nurturing relationships with these high-value customers, the company can maximize its profitability and ensure long-term success.

Top Customers with Highest CLV

```
Top 10 Customers by CLV:
      Customer ID            CLV
2300      14646.0  295979.003331
5756      18102.0  234909.284438
1810      14156.0  143166.879194
2565      14911.0  124355.302446
1348      13694.0  114870.920470
5104      17450.0  102786.414044
5165      17511.0   91065.659933
4338      16684.0   74804.723216
69        12415.0   74245.447687
2965      15311.0   59716.069112
```

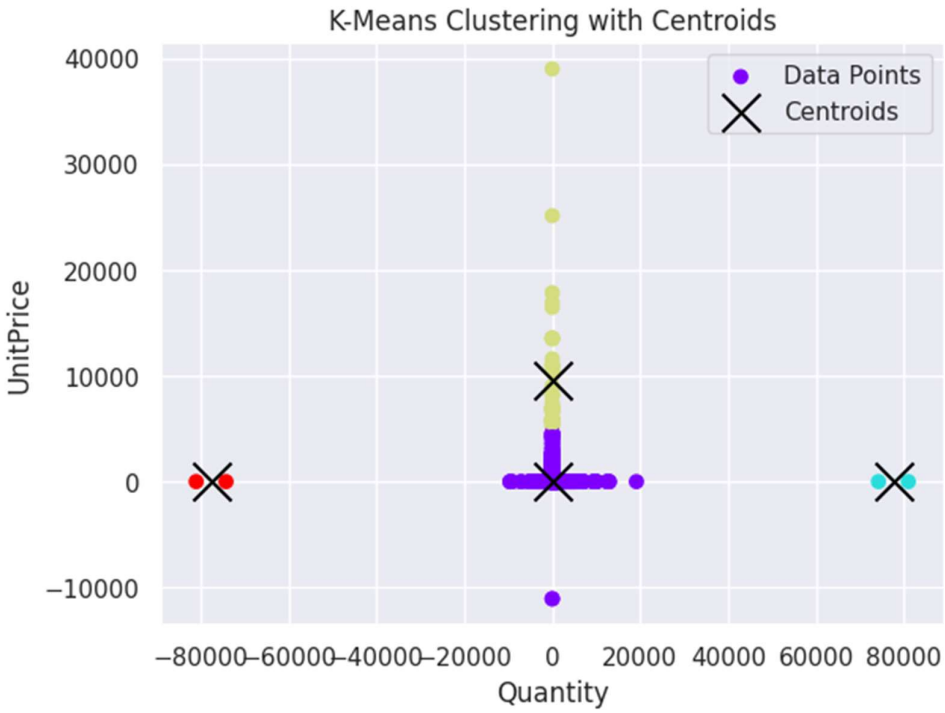
K-Means clustering

K-Means clustering is an unsupervised machine learning algorithm used to group data points into *K* clusters based on their similarity. The algorithm begins by initializing *K* random centroids, then assigns each data point to the nearest centroid. It iteratively updates the centroids by calculating the mean of the data points in each cluster and reassigns points to the closest centroids until the centroids stabilize. The aim is to minimize the distance between data points and their respective centroids, creating compact clusters. K-Means is commonly used in applications such as customer segmentation and pattern recognition, and the optimal number of clusters can be determined using techniques like the Elbow Method.



Elbow Point Graph:

The Elbow Point Graph indicates that the optimal number of clusters for this dataset is 4. This is identified by the point where the WCSS value significantly decreases and then begins to flatten out. This suggests that further increasing the number of clusters would not result in significant improvements in clustering performance.



Customer Segmentation:

This plot shows K-Means clustering with centroids, where data points are segmented based on Quantity and UnitPrice. The centroids, represented by large black crosses, signify the centre of each cluster, and the different colours indicate distinct customer segments.

Segment Analysis:

1. Blue Segment (Right):

   This group is characterized by high quantity and moderate unit price, suggesting that these customers are most likely wholesalers who purchase in bulk, leading to large quantities per transaction.

2. Purple Segment (Centre):

This cluster has a moderate quantity and low unit price, potentially representing regular customers or retail buyers. The relatively lower purchase amounts are indicative of typical retail transactions, where customers purchase in smaller quantities.

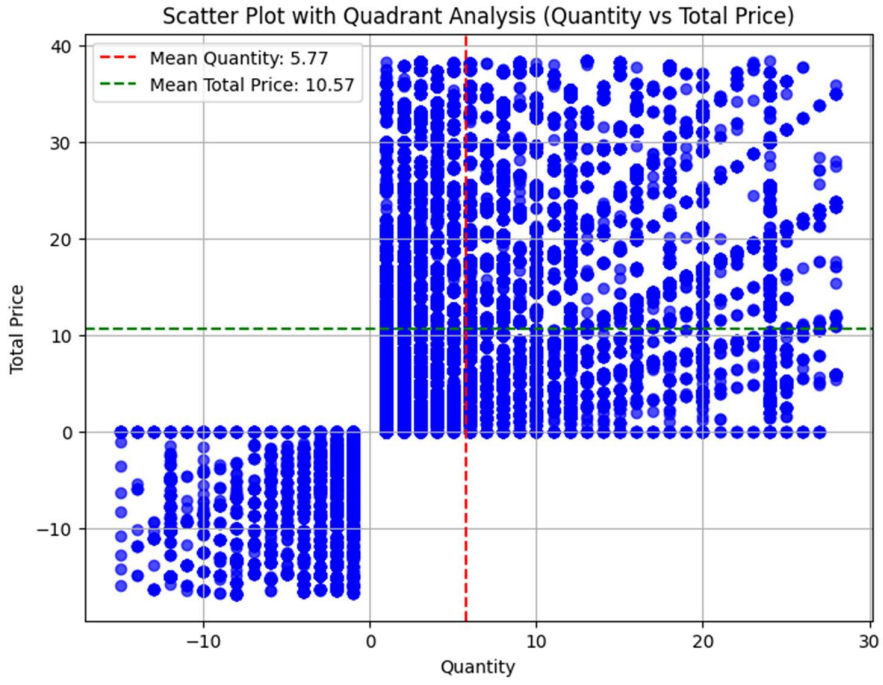
3. Red Segment (Left):

   Customers in this segment display negative or very low quantity and low unit price. The negative quantity could indicate returns or refunds, where items are being sent back or adjustments are made after purchase.

4. Yellow Segment (Top Centre):

   This segment shows moderate to high unit price with low quantity, likely indicating premium or specialized customers who purchase fewer items at higher prices, perhaps representing niche buyers or luxury item customers.

Quadrant Analysis



- Mean Quantity is represented by the vertical red dashed line at 5.77.
- Mean Total Price is represented by the horizontal green dashed line at 10.57.
- The plot is divided into four quadrants:

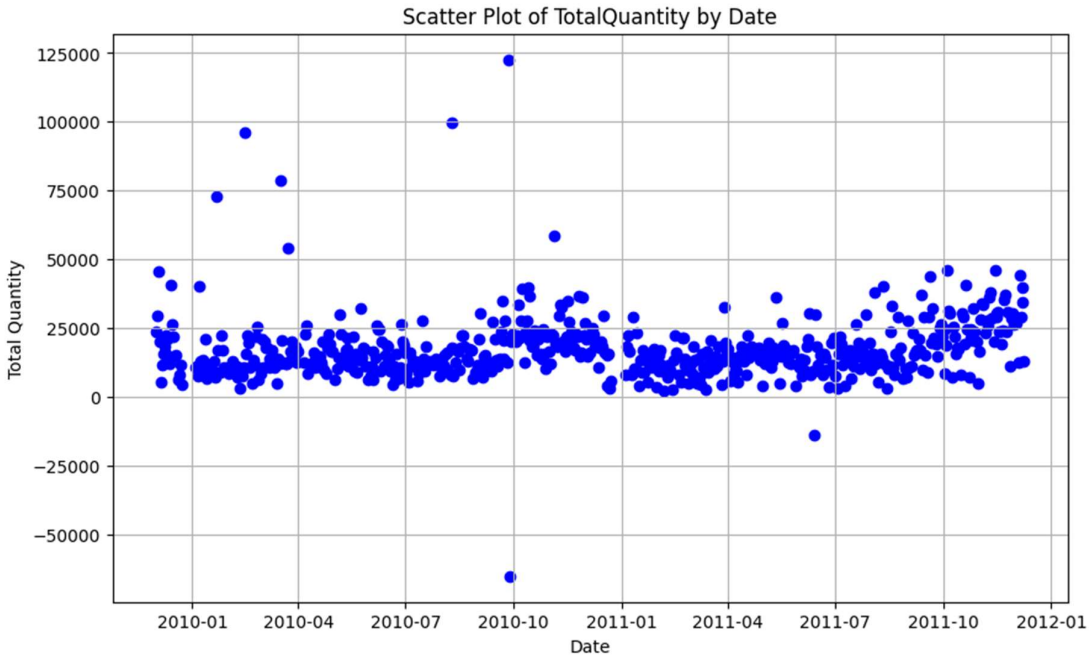
1. Top Right Quadrant (High Quantity, High Price): Points in this quadrant represent transactions with a high quantity of items and a high total price. These are typically high-value orders.
2. Bottom Right Quadrant (High Quantity, Low Price): Transactions with a high number of items but a relatively low total price. This could represent bulk purchases of low-cost items.
3. Top Left Quadrant (Low Quantity, High Price): These points correspond to orders where fewer items were purchased, but the total price is high. This might indicate purchases of expensive items in small quantities.
4. Bottom Left Quadrant (Low Quantity, Low Price): This quadrant includes transactions with both low quantities and low prices, possibly representing small, low-value orders.

Correlation:

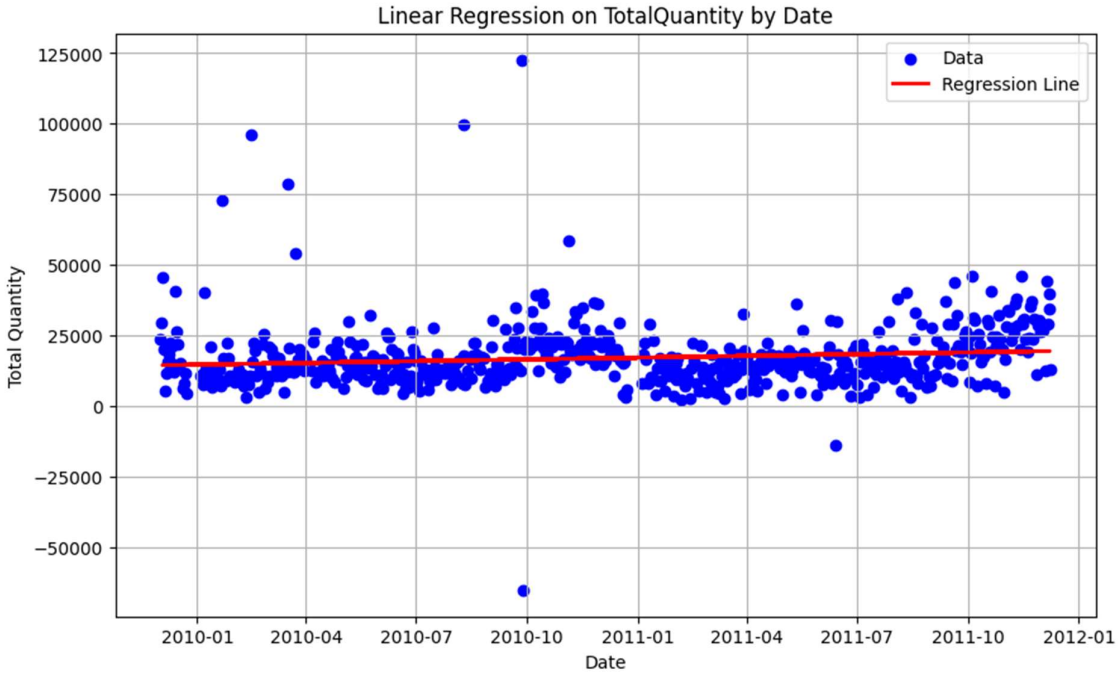
- The data shows a positive correlation between Quantity and Total Price. As the quantity of items increases, the total price also tends to increase. This is expected since total price is calculated as the product of quantity and unit price, leading to a linear trend in this relationship.
- The density of points in the top-right quadrant indicates that many high-value transactions involve large quantities of items, reflecting a direct relationship between the two variables.

Product Centric Approach

Scatter Plot of Total Quantity by Date for Website total Sales Analysis



Curve Fitting




```
Mean Squared Error (MSE): 114969410.28333817
R-squared (R2): -0.009813334466801216
```

Linear regression Curve Fitting Model Evaluation

Model Evaluation: Total Quantity Prediction (Before Removing Outliers)

Before removing outliers, the curve fitting model's performance for predicting Total Quantity based on Date was notably poor, as reflected in the following metrics:

- **Mean Squared Error (MSE):**
  The MSE was 114,969,410.28, showing a significant gap between the observed and predicted values for Total Quantity. This high error value indicates that the fitted curve poorly captured the relationship, with predictions deviating substantially from actual values.

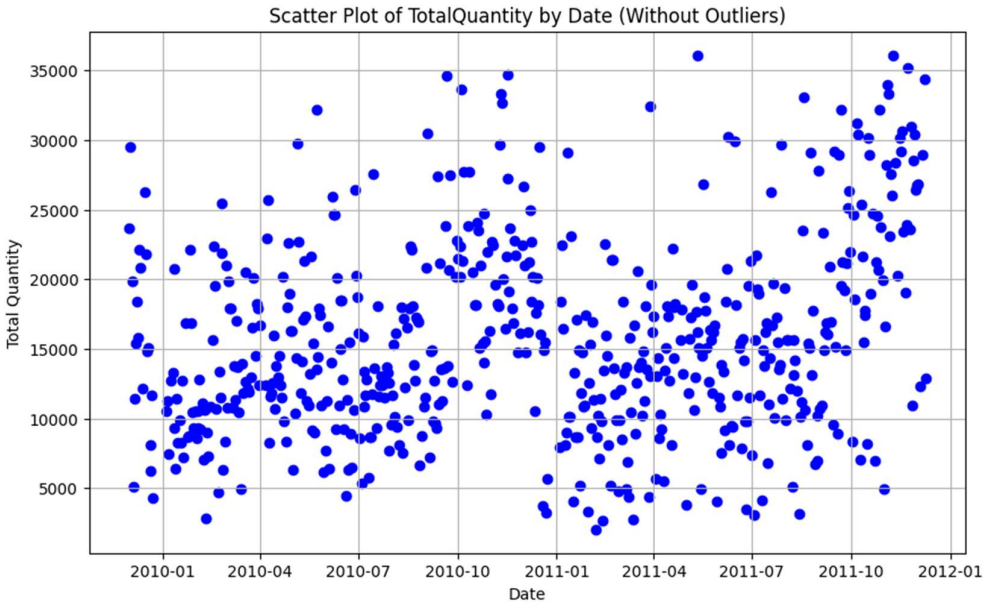
- **R-squared (R^2):**
  The R^2 value was -0.0098, suggesting that the fitted curve explained almost none of the variation in Total Quantity. The negative R^2 indicates that the model performed worse than simply predicting the mean for all observations, emphasizing that Date was not an effective predictor without outlier removal.

The combination of the large MSE and the negative R^2 underscores the poor quality of the initial curve fit, highlighting the importance of removing outliers to improve the model's accuracy and reliability.

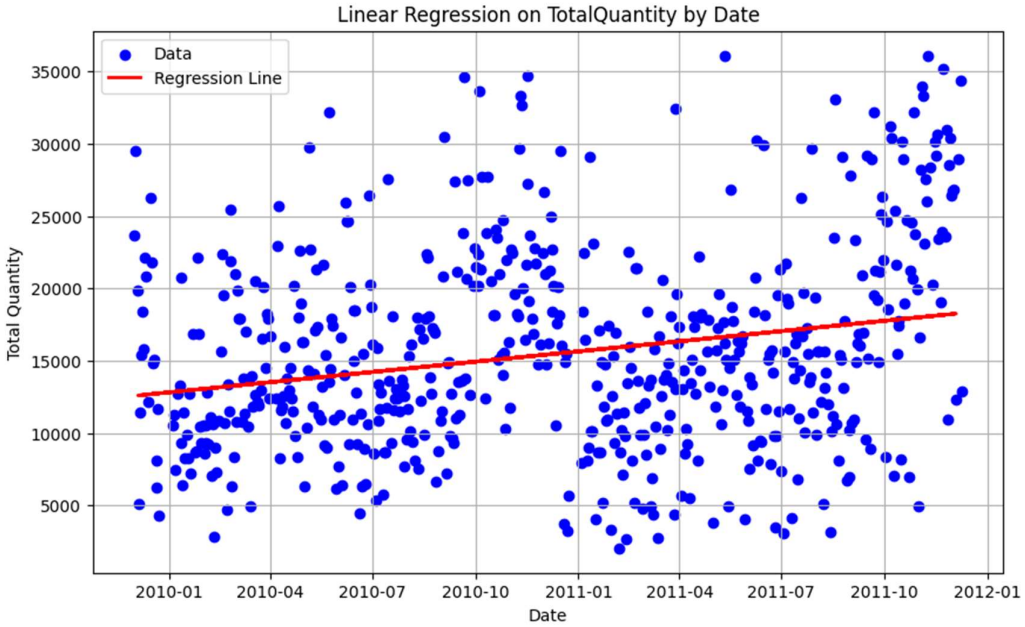
Scatter Plot of Total Quantity after Removing Outliers by Date for Website Total Sales Analysis (Using Interquartile Range (IQR) for Outlier Detection)

- **Interquartile Range (IQR) and Outlier Detection:**
  The IQR was 10,127, with the first quartile (Q1) at 10,844 and the third quartile (Q3) at 20,971. Using the IQR method for outlier detection, outliers were identified with a lower bound of -4,346.5 and an upper bound of 36,161.5. Data points falling outside these bounds were removed to improve the curve's fit and predictive capability.

By removing outliers based on these bounds, the model became more robust, resulting in a curve that better represents the true relationship between Date and Total Quantity, ultimately leading to enhanced accuracy and reliability in predictions.



Curve Fitting




```
Mean Squared Error (MSE): 40548981.186312675
R-squared (R2): 0.026807573478857982
```

Linear regression Curve Fitting Model Evaluation

Model Evaluation: Total Quantity Prediction (After Removing Outliers)

The performance of the curve fitting model for predicting Total Quantity based on Date was assessed using key metrics such as Mean Squared Error (MSE) and R-squared (R^2).

- **Mean Squared Error (MSE):**
  The MSE, calculated at 40,548,981.19, reflects the average squared difference between the actual and predicted Total Quantity values. Although lower than the MSE before outlier removal, this high value still suggests that the fitted curve deviates significantly from the actual data, pointing to poor predictive accuracy. A lower MSE would indicate a better fit of the curve to the data.

- **R-squared (R^2):**
  The R^2 value of 0.027 shows that the curve explains only 2.7% of the variance in Total Quantity. Since R^2 ranges from 0 (no explanation) to 1 (perfect explanation), this low value emphasizes that Date is a weak predictor for Total Quantity, and the curve fitting model fails to capture much of the data's variability.

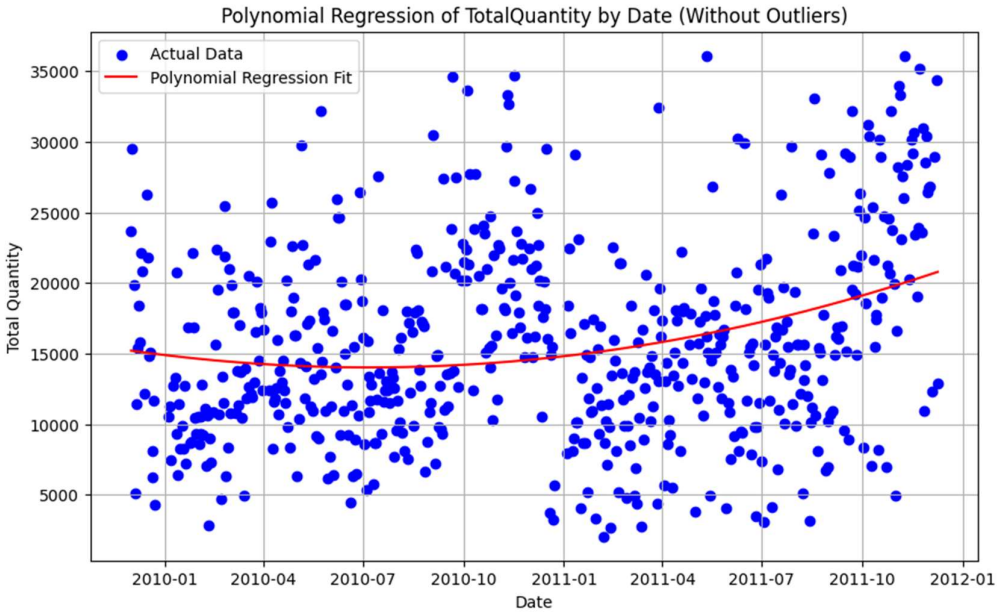
While outlier removal improved the model's performance, the high MSE and low R^2 still indicate that the curve is not an effective representation of the relationship between Date and Total Quantity, suggesting that other variables or more complex models may be needed to enhance prediction accuracy.

Result

Linear Regression		
Methods	**Before removing outliers**	**After removing outliers**
Mean Squared Error (MSE)	**114969410.28333817**	**40548981.186312675**
R-squared (R2)	**-0.009813334466801216**	**0.026807573478857982**

After removing outliers, the **MSE significantly decreased**, and **R^2 slightly improved**, indicating that removing outliers had a positive impact on model performance, though the improvement in R^2 suggests that Date still remains a weak predictor of Total Quantity.

Polynomial Regression Analysis:




```
Mean Squared Error (MSE): 45577576.87638631
R-squared (R2): 0.07368477714731281
```

Polynomial Curve Fitting Model Evaluation

After applying polynomial curve fitting to model the relationship between Total Quantity and Date, the following insights were derived:

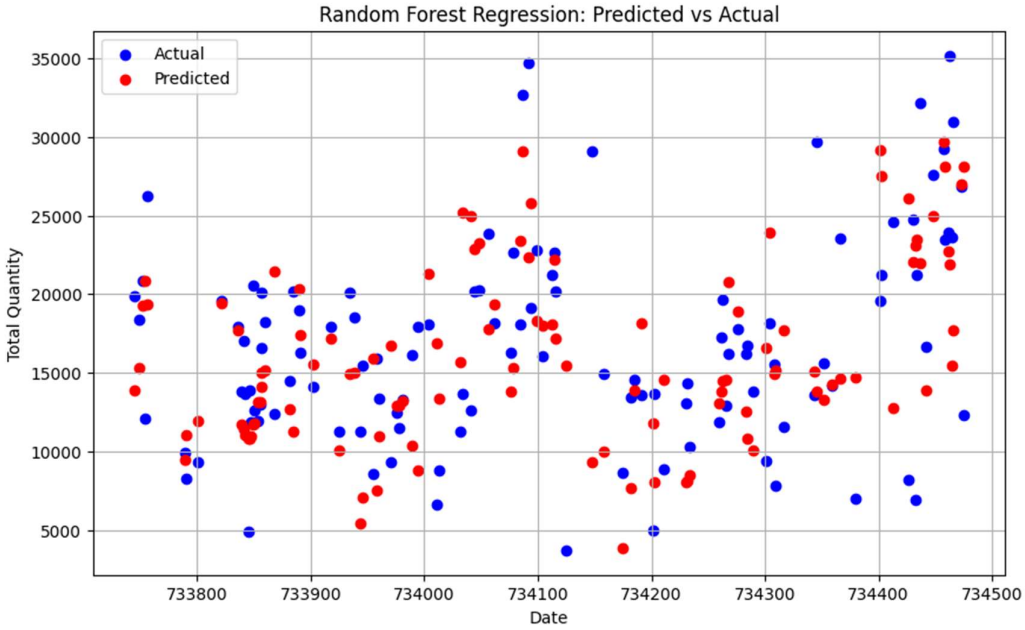
- **Model Performance:**

  The Mean Squared Error (MSE) for the polynomial curve fitting model was 45,577,576.88, indicating a significant gap between the actual and predicted values of Total Quantity. This high MSE suggests that the fitted curve struggles to effectively capture the underlying relationship between Total Quantity and Date. The R-squared (R^2) value of 0.074 indicates that the model explains only 7.4% of the variance in Total Quantity, highlighting a weak fit. The low R^2 value implies that the model does not account for much of the variability in the data.

- **Model Interpretation:**

  Despite the removal of outliers, the polynomial curve fitting model fails to adequately capture the complexities and patterns in the dataset. The low R^2 suggests that this particular polynomial curve may not be suitable for representing the relationship between Date and Total Quantity. It indicates that other influencing factors are likely impacting sales quantities over time. Further investigation and the consideration of

alternative modelling approaches are necessary to improve the model's accuracy and predictive performance.




```
Mean Squared Error (MSE): 43413815.46754568
R-squared (R²): -0.041949641725236964
```

Model Performance Evaluation

- **Mean Squared Error (MSE):**
  The MSE of the model was 43,413,815.47, indicating a high level of deviation between the predicted and actual values. This suggests substantial prediction errors and highlights potential issues with model fit or data quality, necessitating further investigation.

- **R-squared (R^2) Value:**
  The R^2 value was -0.042, which is notably negative. A negative R^2 signifies that the model performs worse than a baseline prediction using the mean of the dependent variable. Typically, R^2 values range from 0 to 1, with higher values indicating better model performance. The negative R^2 indicates that the model fails to explain the variability in the Quantity variable, potentially due to incorrect model specification, insufficient features, or complexities in the dataset that the model cannot capture.

- **Conclusion:**
  The high MSE and negative R^2 suggest that the current regression model is not effective

for predicting Quantity based on Invoice Date. These results point to significant issues with the model's structure and its ability to generalize, indicating the need for revaluation or alternative modelling approaches.

Comparative Analysis

Models	**Linear Regression**	**Polynomial Regression**	**Random Forest**
Mean Squared Error (MSE)	**40548981.186**	**45577576.876**	**43413815.467**
R-squared (R2)	**0.0268**	**0.736**	**-0.042**

1. **Linear Regression:** With an R^2 of 0.0268 and an MSE of 40,548,981.19, this model has a poor fit, explaining very little of the variability in Quantity and showing high prediction errors.
2. **Polynomial Regression:** Exhibited a much higher R^2 of 0.736 and an MSE of 45,577,576.88, indicating a better fit than the linear model. However, the high MSE suggests that there is still room for improvement.
3. **Random Forest:** Had an MSE of 43,413,815.47 and a negative R^2 of -0.042, indicating worse performance than a baseline model and potential issues such as overfitting or inadequate feature representation.

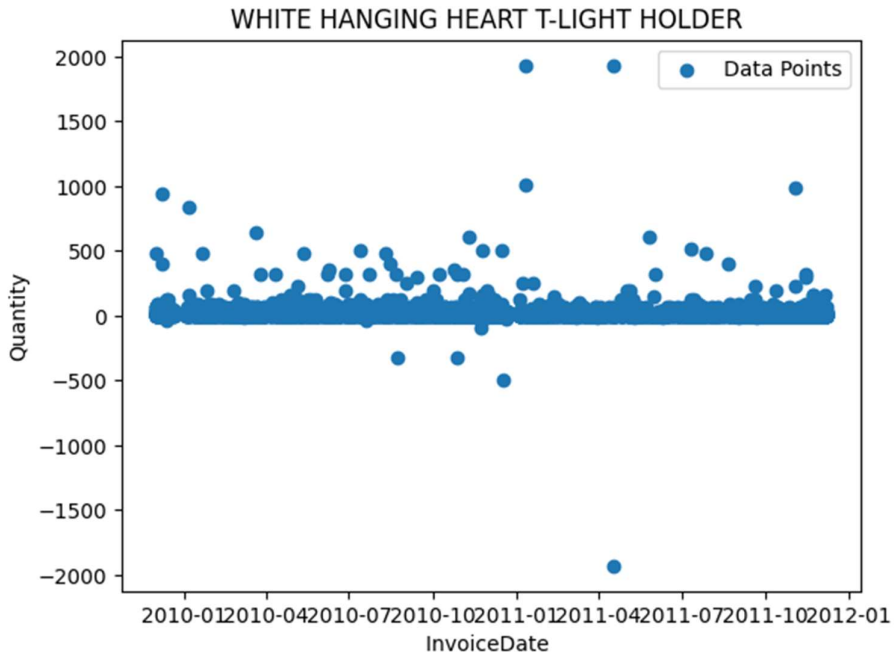
The polynomial regression model provides the best fit among the models tested, as evidenced by its higher R^2 value. However, all models show significant prediction errors, and further refinement or exploration of alternative modelling approaches may be necessary to improve predictive performance.

Top Selling Products

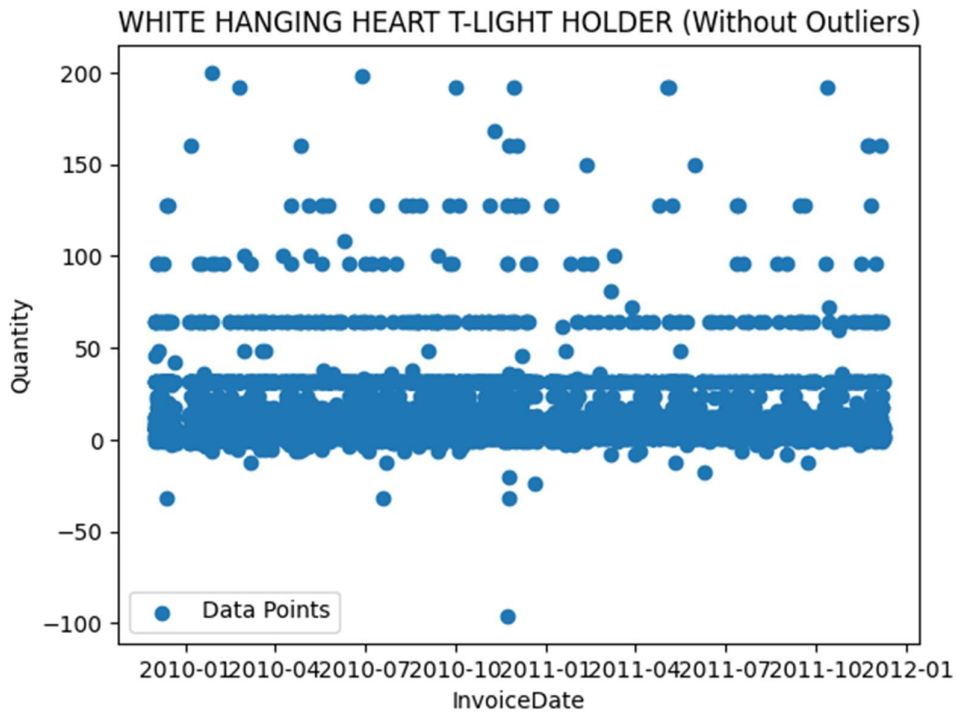
```
Top 10 Selling Products:
WORLD WAR 2 GLIDERS ASSTD DESIGNS
WHITE HANGING HEART T-LIGHT HOLDER
ASSORTED COLOUR BIRD ORNAMENT
JUMBO BAG RED RETROSPOT
BROCADE RING PURSE
PACK OF 60 PINK PAISLEY CAKE CASES
60 TEATIME FAIRY CAKE CASES
PACK OF 72 RETROSPOT CAKE CASES
SMALL POPCORN HOLDER
PACK OF 72 RETRO SPOT CAKE CASES
```

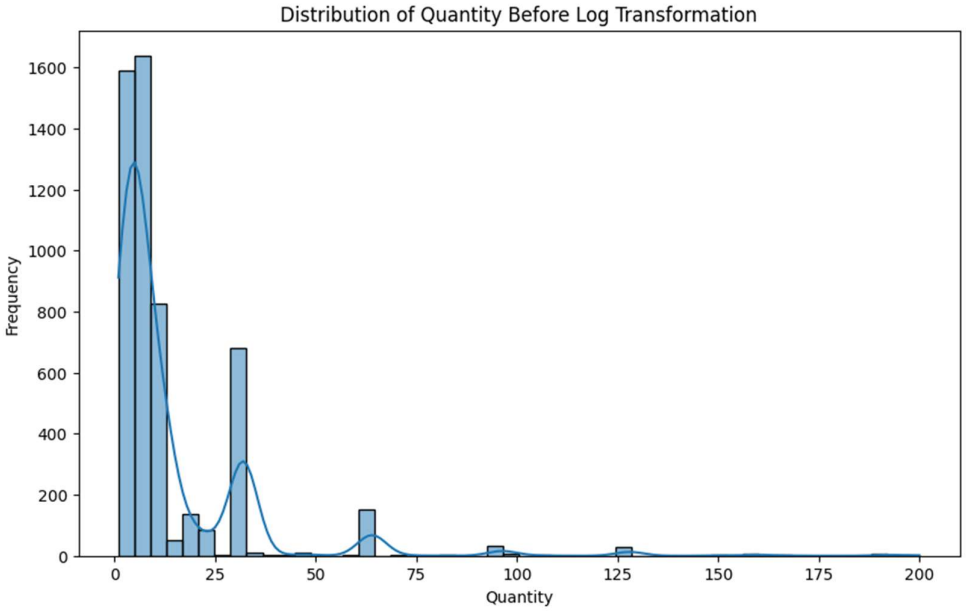
Visualization of Sales Trends Across Products

White hanging heart T- light holder

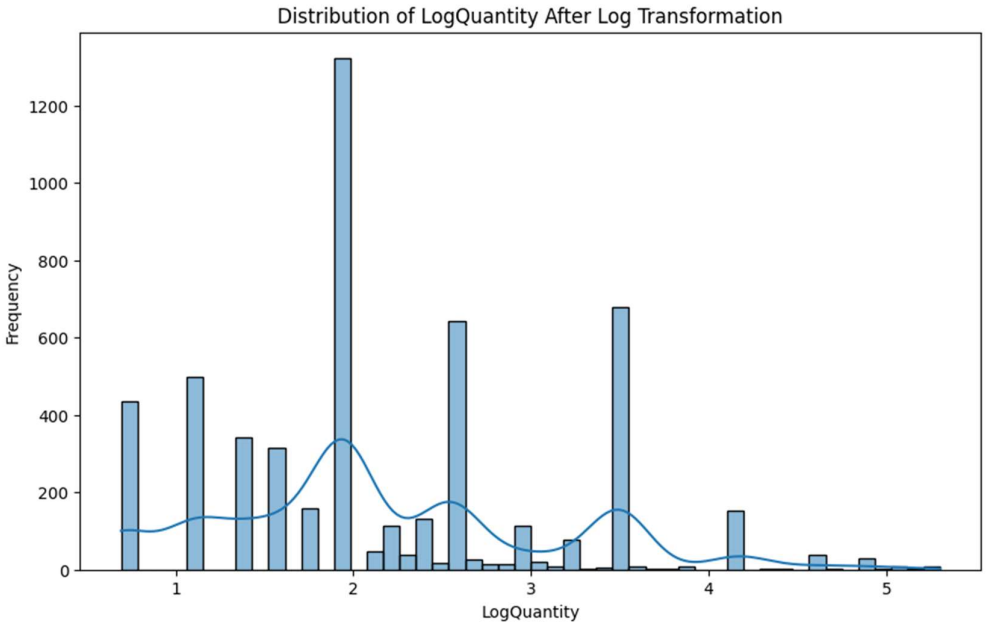


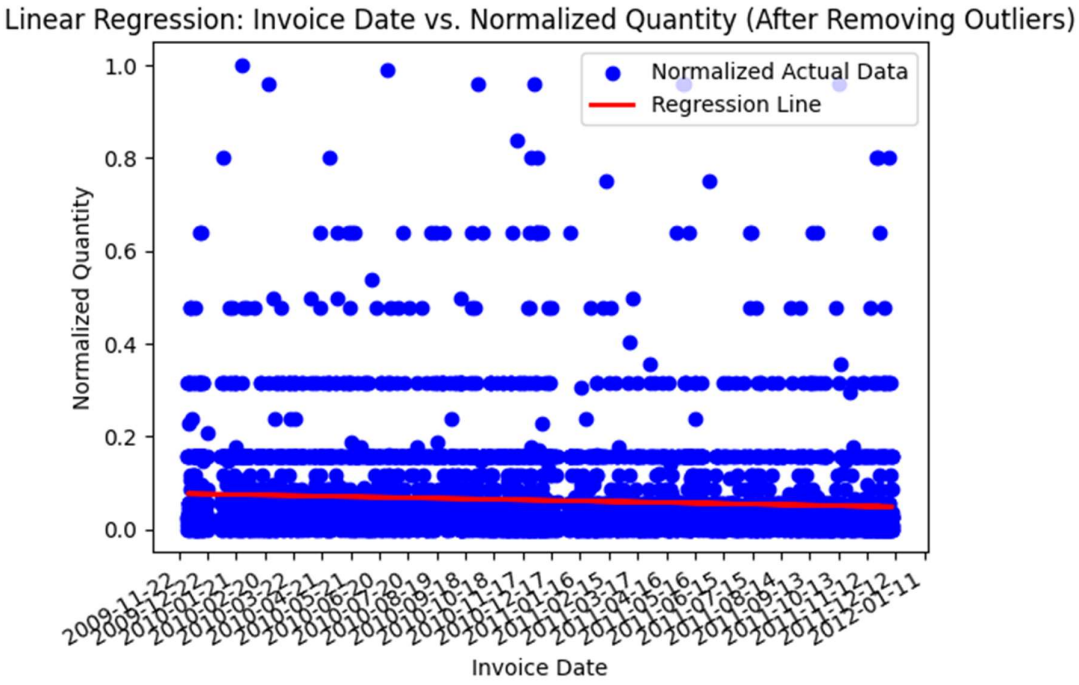
After removal of outliers





After Log transformation





Model Performance Metrics

- Mean Squared Error (MSE): 0.0094

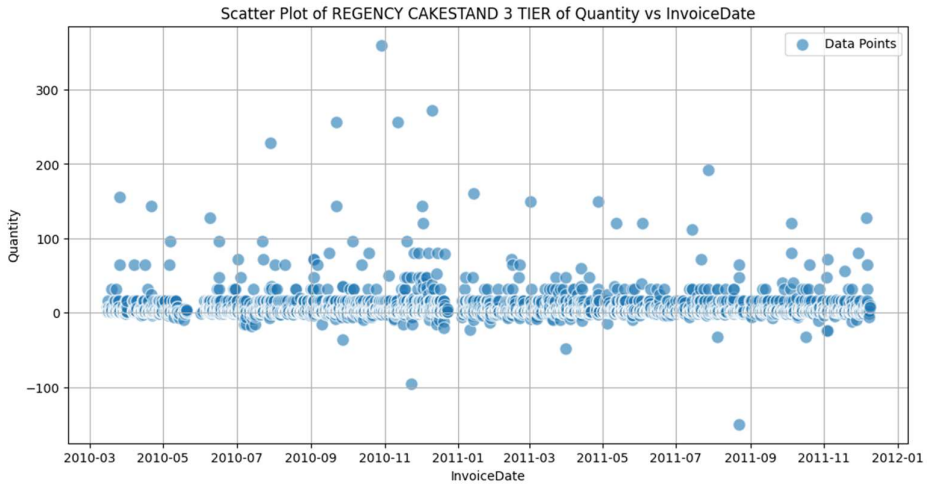
  The MSE indicates the average squared difference between actual and predicted values. A lower MSE generally signifies a better fit, though it does not fully capture the model's performance relative to the data's scale.

- Root Mean Squared Error (RMSE): 0.0969
  The RMSE, being in the same units as Quantity, shows that, on average, the model's predictions deviate by approximately 0.097 units from the actual values. Depending on the range of Quantity, this deviation may be considered small or large.
- Mean Absolute Error (MAE): 0.0610
  The MAE represents the average absolute difference between actual and predicted values. It indicates that the model's predictions are off by about 0.061 units on average, providing context for the model's error in relation to the actual data.
- R-squared (R^2): 0.0039
  The R^2 value indicates that only about 0.39% of the variance in Quantity is explained by the model. This low R^2 suggests a weak relationship between InvoiceDateTime_num and Quantity, highlighting that the model does not capture much of the underlying data patterns.

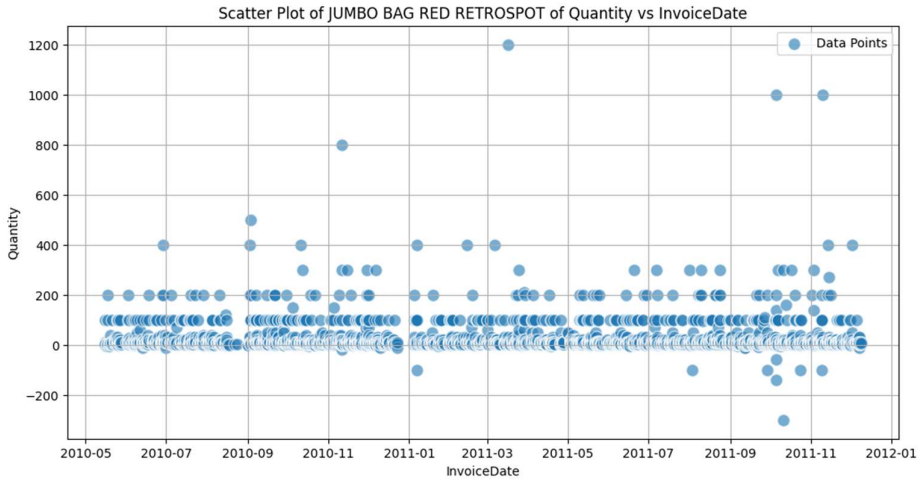
Conclusion:

The combination of a low R^2 and small error metrics indicates that while the model does not make large prediction errors, it fails to explain or predict the variability in the data effectively. This suggests that the model may be underfitting and missing important factors that influence Quantity. Further refinement or additional features may be needed to improve model performance.

Regency Cake Stand 3 Tier



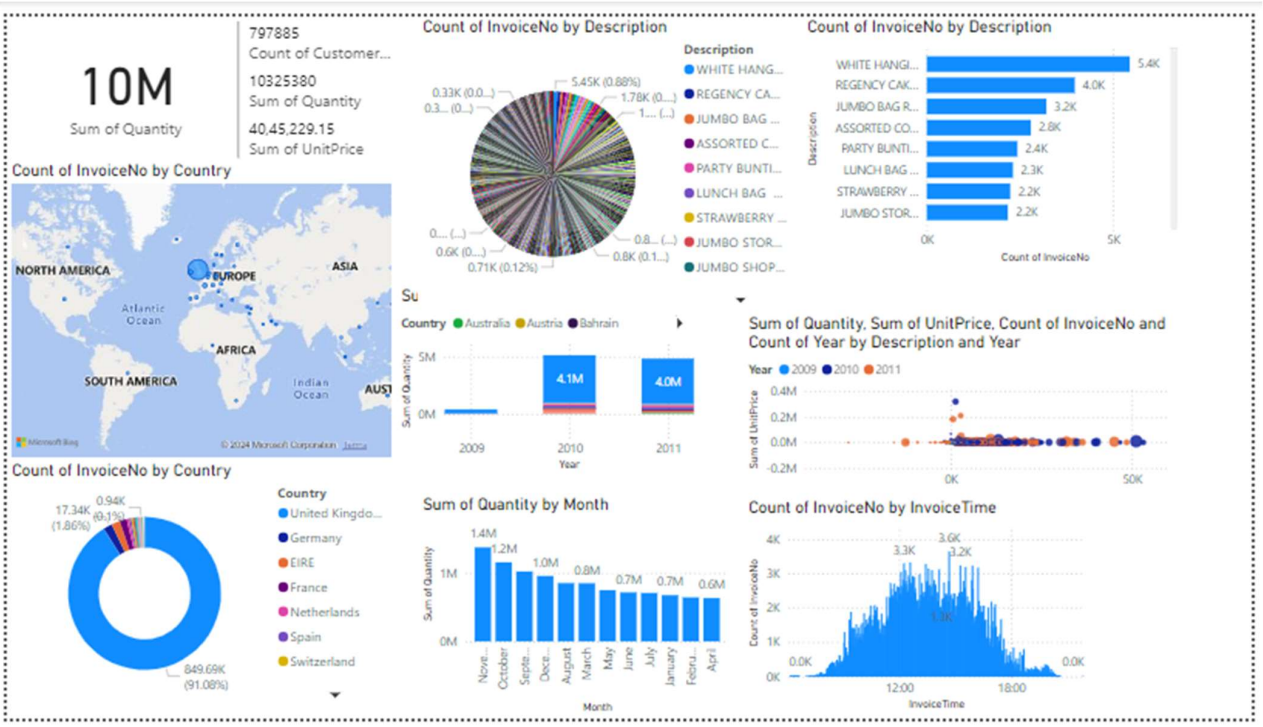
Jumbo Bag Retrospot

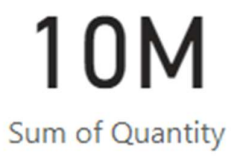


Model Performance Analysis Other Products

The analysis conducted for the "White Hanging Heart T-Light Holder" revealed that the model exhibited a low R^2 value and relatively small error metrics, indicating it failed to capture the variability in Quantity effectively. Similar results were observed for other products, including the "Regency Cake Stand 3 Tier" and the "Jumbo Bag Retrospot." These products, along with others in the dataset, demonstrated comparable performance characteristics, suggesting that the model's limitations are consistent across various product types. This underscores the need for further refinement or exploration of additional features to improve the model's ability to predict and explain the Quantity for a diverse range of products.

4. DASHBOARDING USING POWER BI



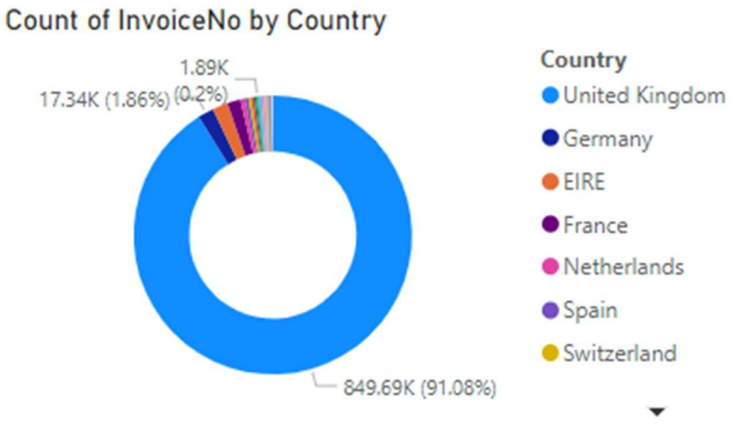


Total Quantity Purchased

An analysis of the dashboard reveals that the total quantity of products purchased from the website amounts to 10 million units. This figure highlights the volume of transactions and reflects the scale of customer activity on the platform.

Geographical Sales Distribution

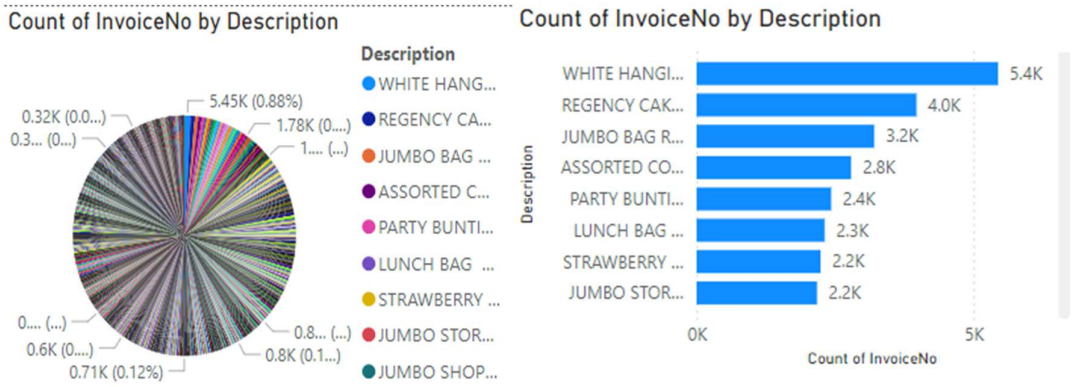
The analysis of daily invoice counts by country shows that the majority of sales occur in Europe. This conclusion is supported by visualizations, where the largest concentration of activity is represented by a significant data point cluster from European countries. As such, it can be concluded that Europe is the primary market for the online retail store.



Sales by Country

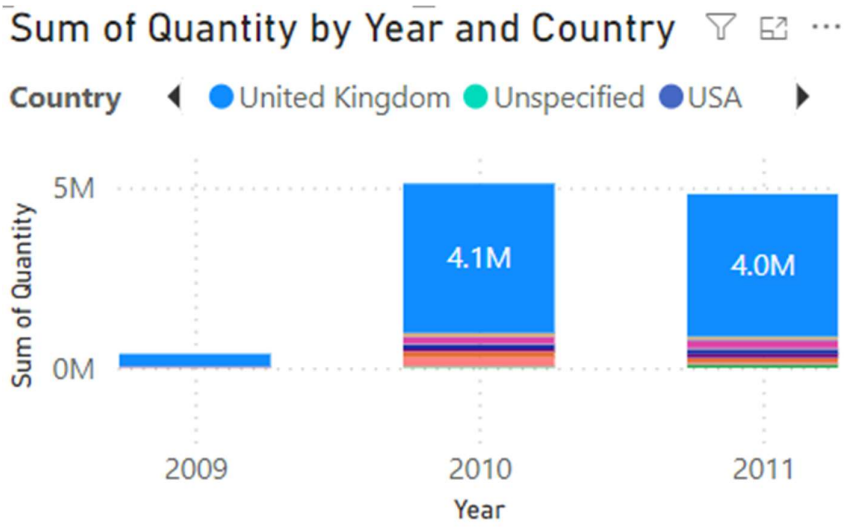
As evident from the doughnut chart, **91.08%** of the total sales i.e. **849.69K invoices**, originate from the **United Kingdom**, making it the dominant market. While other countries like Germany, France, and Switzerland etc. contribute to the overall sales, their share is

significantly smaller in comparison, further emphasizing the United Kingdom's leading role in driving sales for the online retail store.



Product Sales Analysis

The analysis of the pie chart and bar chart reveals that the product with the highest sales is the **White Hanging Heart T-Light Holder**, which achieved impressive sales of 5.4 thousand units. In contrast, the product with the lowest sales is the **Polka Dot Rain Hat**, with only 0.3 thousand units sold. This stark contrast in sales performance highlights the varying levels of consumer interest and demand for different products within the inventory.

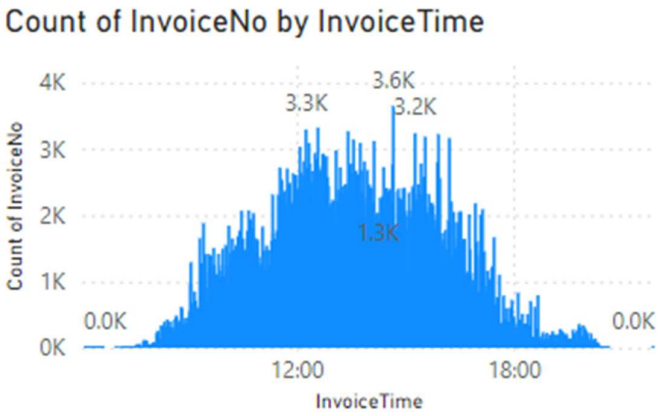


Trends in Sales Performance (2009-2011)

The sales analysis for the years 2009 to 2011 reveals notable fluctuations in product demand. In 2009, sales quantity remained relatively low, reflecting a period of limited demand influenced by various external factors. However, in 2010, the market experienced a remarkable turnaround, with sales increasing multiple fold to reach an impressive total of 4.1 million units. This surge in demand indicates a successful recovery and highlights the impact of potential improvements in marketing strategies or product offerings. Conversely, in 2011, sales experienced a slight decline, dropping to 4.0 million units. While this decrease was marginal, it prompts an investigation into the factors affecting consumer behaviour during this period to develop effective strategies for stabilizing sales and promoting growth in the future.

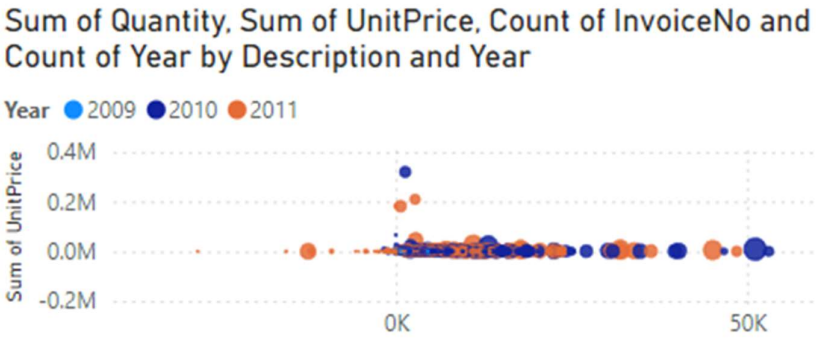
Monthly Sales Distribution Analysis

Upon analysing the month-wise distribution of sales, it is evident that November recorded the highest sales, with a total of 1.4 million products sold, followed closely by October with 1.2 million products sold. In contrast, the months of February and April exhibited the lowest sales figures, each reaching only 0.6 million products. This distribution highlights seasonal trends in consumer purchasing behaviour and underscores the importance of further investigation into the factors that drive sales during peak and low periods.



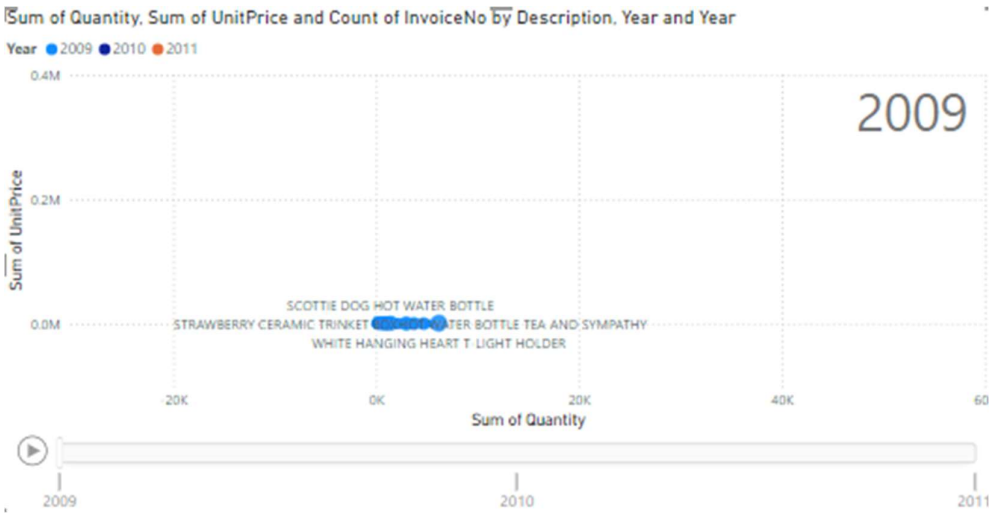
Invoice Timing Distribution

The analysis of invoice activity over the course of the day, as shown in the chart, reveals that **invoice times follow a normal distribution**. The peak transaction period occurs between **12:00 PM and 3:00 PM**, where the highest volume of invoices is recorded, with counts reaching as high as **3.6K** around midday. The activity gradually declines after this peak, indicating that most transactions are concentrated in the afternoon, with fewer sales occurring in the early morning and late evening.



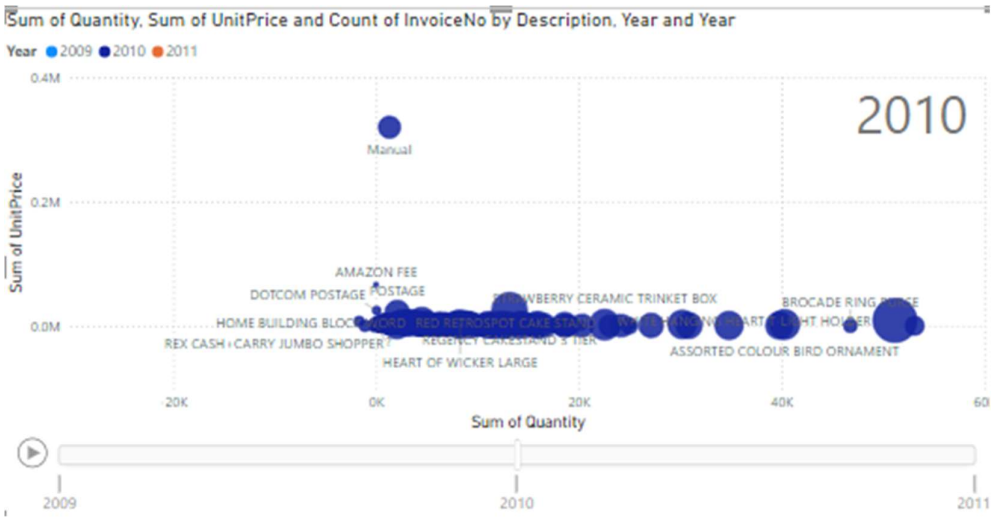
Yearly Comparison of Sales by Quantity and Unit Price

The scatter plot compares the sum of **Quantity**, **Unit Price**, and **Invoice Count** across different years (2009, 2010, and 2011). From the plot, it can be observed that the bulk of transactions in all three years cluster around a **low unit price range** with relatively moderate quantities. However, some **outliers** are present, indicating a few transactions with **high unit prices**, especially in **2010** and **2011**. These year's show more significant deviations in pricing, reflecting some large sales during this period. Overall, the majority of transactions exhibit consistent pricing and quantities over the three years, with a notable concentration of sales around lower price points.



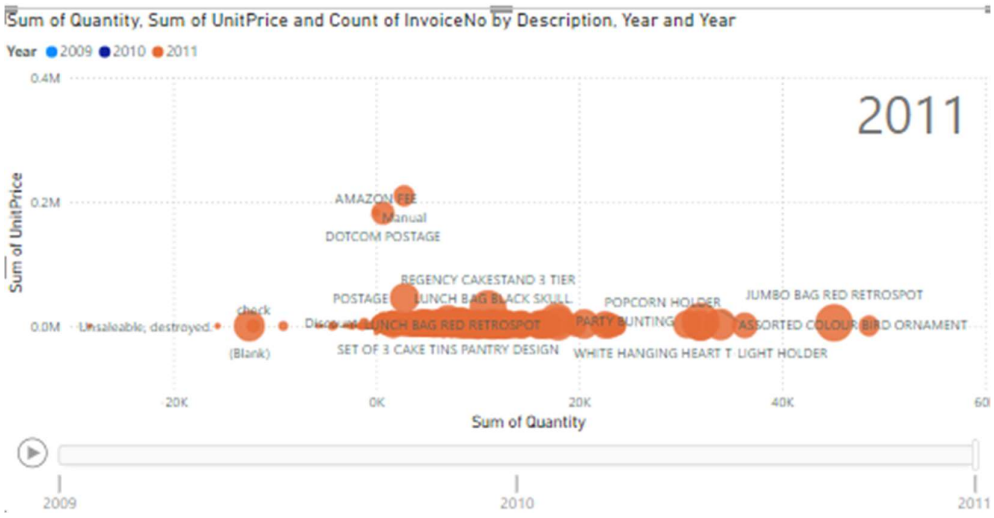
Sales Analysis for 2009

The analysis of sales data for the year 2009 indicates that the sales quantity remained relatively low compared to other years. This observation suggests a period of limited product demand, which could have been influenced by various external factors.



Sales Analysis for 2010

The sales data for the year 2010 reflects a significant rebound in product demand, with sales quantity increasing multiple fold to reach an impressive total of 4.1 million units. This substantial growth indicates a successful recovery from the low sales observed in 2009 and suggests that various factors, such as improved marketing strategies, product enhancements, or changes in consumer behaviour, may have played a crucial role in driving this positive trend.



Sales Analysis for 2011

In 2011, the sales data indicates a slight decline in product demand, with sales quantity dropping to 4.0 million units from the previous year's peak of 4.1 million units. These marginal decreases, while not drastic, raises questions about potential underlying factors that may have influenced consumer purchasing behaviour. Analysing the causes behind this decline will be essential for developing targeted strategies to stabilize sales and encourage growth in subsequent years.

5. CONCLUSION

The "Customer Purchase Behaviour Analysis and Visualization" project has successfully provided a detailed understanding of customer purchasing patterns and product performance within an online retail context. Through a comprehensive approach that included data cleaning, exploratory data analysis, and advanced modelling techniques, we have identified significant trends and behaviours that influence customer interactions.

Our analysis revealed a distinct trend where the majority of transactions are characterized by low-cost items, punctuated by occasional high-value purchases. This finding aligns with typical retail dynamics, where a strategy focused on high-volume, low-margin sales can be observed. However, the model evaluations indicated challenges in accurately capturing the complexities of customer purchase behaviour. Metrics such as high Mean Squared Error (MSE) and low R-squared (R^2) values suggest that while some models, particularly polynomial regression,

improved the fit and explained a portion of the variability in sales, there remains substantial room for enhancement in predictive accuracy.

The insights gained underscore the necessity for further refinement of the predictive models. Incorporating additional features and potentially exploring advanced modelling techniques could yield better results. Understanding Customer Lifetime Value (CLV) emerged as a critical factor, allowing us to identify and prioritize customers who contribute significantly to the bottom line. This understanding can enable businesses to tailor their marketing strategies more effectively and allocate resources where they will have the most impact.

Moreover, the visualization of sales trends and customer segments provided actionable insights that can guide strategic decision-making. By focusing on high-value segments and adapting strategies to meet the needs of these customers, businesses can enhance their overall profitability.

An important observation during the analysis was the presence of negative quantities in the dataset, indicating product returns. Investigating the reasons behind these returns could offer valuable insights for the company. Analysing return reasons can help identify areas for improvement in product quality, customer service, or marketing strategies, ultimately reducing return rates and enhancing customer satisfaction.

In summary, while this project has yielded valuable insights into customer purchase behaviour and product performance, it also highlights the need for ongoing refinement and exploration of innovative analytical approaches. By leveraging these insights and continuing to enhance our understanding of customer dynamics, businesses can make informed strategic decisions that drive growth and improve customer engagement in a competitive marketplace. The future scope of this analysis includes investigating return reasons, refining predictive models, and exploring additional variables that may influence customer behaviour, ensuring that the findings are not only actionable but also sustainable for long-term success.

6. Future Scope

The "Customer Purchase Behaviour Analysis and Visualization" project has successfully uncovered key insights into customer behaviour, product performance, and sales trends. However, several areas can be explored further to add even greater value to businesses:

1. **Return Analysis:** A noteworthy finding during the analysis was the presence of numerous negative quantities in the dataset, which indicate product returns. If data on the reasons for these returns were available, it could significantly enhance the analysis. Understanding why products are being returned—whether due to defects, dissatisfaction, or logistical issues—would allow the company to implement targeted strategies to reduce return rates. This could lead to improved product quality, more effective customer service, and higher levels of customer retention.

2. **Advanced Customer Segmentation:** While customer segmentation based on purchase behaviour has already been conducted using techniques like K-means clustering, further refinement is possible. Incorporating additional features such as customer demographics, browsing history, and interaction patterns could lead to even more precise segments. This would enable businesses to develop highly personalized marketing strategies that cater to specific customer needs and preferences.

3. **Predictive Analytics Enhancement:** The current predictive models focus on sales forecasting based on historical data. Future work could explore more advanced machine learning models, such as deep learning or ensemble methods, to improve prediction accuracy. Additionally, the integration of external factors, such as economic indicators, seasonality, and competitor actions, could enhance the robustness of the predictive models.

4. **Real-Time Data Processing**: The project relied on historical data for analysis and modelling. Implementing real-time data processing pipelines could enable businesses to monitor customer behaviour and sales performance dynamically. This would allow for immediate adjustments to marketing campaigns, inventory management, and customer engagement strategies based on real-time insights.

5. **Customer Experience Enhancement**: Insights from the current analysis can be expanded to focus on improving customer experience. By integrating feedback mechanisms, sentiment analysis from customer reviews, and customer support

interactions, businesses could better understand pain points in the customer journey and proactively address them.

6. **Product Bundling and Cross-Selling Opportunities:** Further analysis could be conducted to identify potential product bundling and cross-selling opportunities. By analysing purchase patterns, businesses could develop targeted recommendations to offer complementary products, increasing the average transaction value and boosting customer satisfaction.

By exploring these future directions, businesses can leverage data-driven insights to refine their strategies, improve customer engagement, and drive sustained growth in an increasingly competitive market.

7. Appendix

Link of the data set used

https://archive.ics.uci.edu/dataset/352/online+retail

https://archive.ics.uci.edu/dataset/502/online+retail+ii

Link to the colab files

https://colab.research.google.com/drive/1WpSv0dKN8JAZGhw7A7arPPr2F8Xv9Y1U?usp=sharing

https://colab.research.google.com/drive/1pUEvHkunZ0XQLYjEFXewGBhM3NVjbFVK?usp=sharing

https://colab.research.google.com/drive/1lNw9gyJsj7heMb4jg5sjVSp9Q2mPbVZY?usp=sharing

https://colab.research.google.com/drive/1LKrbluajUM3yQdPsDwaK0uqBk-tSCZmb?usp=sharing

https://colab.research.google.com/drive/1BjFnHp80EOUyjp5uM0nfvYWq4bwEGgoi?usp=sharing

Resources used

https://scikit-learn.org/stable/

https://www.kaggle.com/code/ekami66/detailed-exploratory-data-analysis-with-python

https://www.geeksforgeeks.org/regression-in-machine-learning/

https://www.geeksforgeeks.org/k-means-clustering-introduction/

https://learn.microsoft.com/en-us/power-bi/fundamentals/desktop-getting-started