USER PC

[Email address]

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

[Document title]

[Document subtitle]

Title: Statistical Programming Assessment Report

Module Code: LDS7001M

Student ID:

Submission Date:

Word Count:

Table of Contents

[1.0 Introduction 3](#_Toc182331054)

[2.0 Data Preparation and Cleaning 4](#_Toc182331055)

[3.Exploratory Data Analysis 5](#_Toc182331056)

[3.1 Monthly Fluctuation in Revenue and Transactions. 5](#_Toc182331057)

[3.1.1Monthly Transaction Trend 5](#_Toc182331058)

[Key Insight from the Monthly Transaction Data 6](#_Toc182331059)

[3.1.2Monthly Revenue Trend 7](#_Toc182331060)

[Key Insights from the Monthly Revenue Trend: 8](#_Toc182331061)

[3.2 Product Category Revenue Analysis 9](#_Toc182331062)

[3.2.1 Products performing well over time 9](#_Toc182331063)

[3.2.2 Testing the Sustainability trend 10](#_Toc182331064)

[Key Insights from the trend chart 10](#_Toc182331065)

[3.3 Seasonal Sales Variation 12](#_Toc182331066)

[3.4 Customer Purchasing behaviour 14](#_Toc182331067)

[4.0. Summary of Insights and Recommendations 21](#_Toc182331068)

[5.0 Limitations and Caveats 21](#_Toc182331069)

# 1.0 Introduction

The purpose of this report is to perform an exploratory data analysis (EDA) on historical transaction data for an e-commerce company. The company seeks to optimize its sales and marketing strategies by gaining deeper insights into customer purchasing behaviour, product performance, and revenue trends.

The dataset contains key variables such as customer ID, transaction date, product ID, product category, quantity purchased, and total price. By analyzing this data, we aim to uncover patterns and trends that can guide the company in making data-driven decisions.

To achieve this, Python was used as the primary tool, leveraging powerful libraries such as Pandas for data manipulation and Matplotlib/Seaborn for visualizations. The analysis focuses on four key objectives:

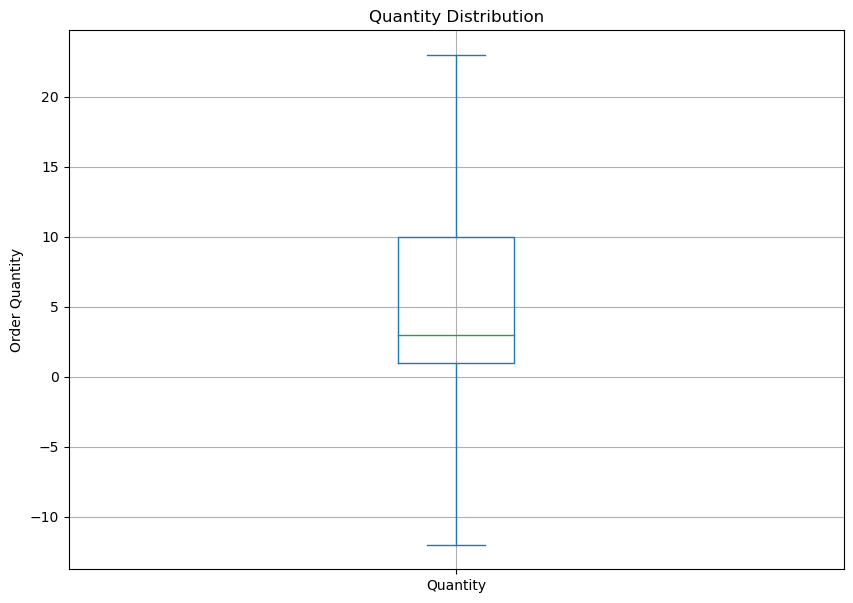
1. Analysing monthly fluctuations in total revenue and transaction count.
2. Identifying product categories with the highest revenue and consistent growth.
3. Exploring seasonal variations in product sales.
4. Examining shifts in customer purchasing behaviour across multiple transactions.

This report will present actionable insights supported by visualizations and tables to help the company enhance its sales strategy effectively.

# **2.0 Data Preparation and Cleaning**

The dataset, containing fields such as Customer ID, Transaction Date, Product ID, Product Category, Quantity Purchased, and Total Price, was loaded and examined for quality. Key steps included:

* **Handling Missing Values:** There was no missing value in the dataset.
* **Data Transformation:** InvoiceDate column was initially an object data type but was converted to DateTime data type.
* **Data Structuring:** Aggregation and categorization to facilitate monthly, seasonal, and categorical analyses.
* **Data Filtering:** It was discovered that some input data in the Quantity column were negative due to canceled orders, so they were removed completely from the dataset, for analysis sake.
* **Anomaly Treatment:** It was discovered that the data for the month of December, 2011 was incomplete, the data recorded was from December 1 to 10. So this month was removed to avoid been misrepresented.



# 3.Exploratory Data Analysis

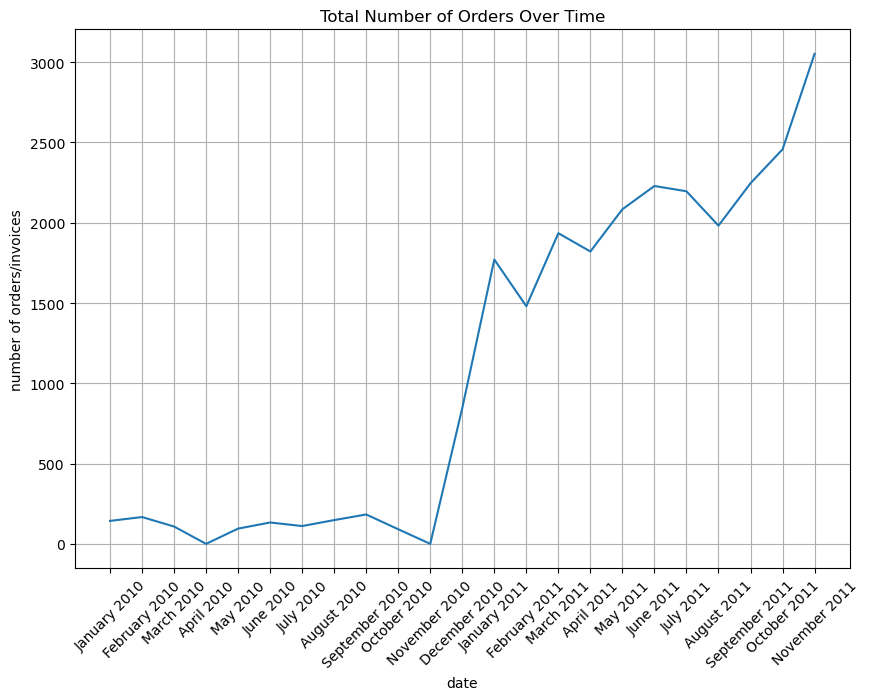
## 3.1 Monthly Fluctuation in Revenue and Transactions.

### 3.1.1Monthly Transaction Trend

The first step in analysing monthly fluctuations in transaction was to calculate total numbers of orders for each month. The dataset was then aggregated on a monthly basis using Python’s Pandas library to the total number of order for each month.

|  |  |  |
| --- | --- | --- |
| index | Month | Number of Orders |
| 1. | January, 2010 | 143 |
| 2. | February, 2010 | 167 |
| 3. | March, 2010 | 108 |
| 4. | April, 2010 | 0 |
| 5. | May, 2010 | 95 |
| 6. | June, 2010 | 133 |
| 7. | July, 2010 | 111 |
| 8. | August, 2010 | 148 |
| 9. | September, 2010 | 183 |
| 10. | October, 2010 | 91 |
| 11. | November, 2010 | 0 |
| 12. | December, 2010 | 846 |
| 13. | January, 2011 | 1771 |
| 14. | February, 2011 | 1480 |
| 15. | March, 2011 | 1935 |
| 16. | April, 2011 | 1821 |
| 17. | May, 2011 | 2084 |
| 18. | June, 2011 | 2229 |
| 19. | July, 2011 | 2196 |
| 20. | August, 2011 | 1982 |
| 21. | September, 2011 | 2246 |
| 22. | October, 2011 | 2457 |
| 23 | November, 2011 | 3057 |
| 24. | December, 2011 | -------- |

Note: December, 2011 will be misrepresented with an incomplete data so it was disregarded.



### Key Insight from the Monthly Transaction Data

1. There was a dramatic increase in orders starting from October-November 2010, showing a clear inflection point in the business. Before this point, orders were relatively flat and low (around 200 orders per month).

2. After the sharp increase in late 2010, the business maintained a generally upward trend through 2011, with order volumes consistently above 1,500 units.

3. The most recent period (late 2011) shows particularly strong growth, reaching approximately 3,000 orders - the highest point in the entire timeline.

4. There are some seasonal fluctuations visible in 2011, with minor dips and peaks, but the overall trend remains positive.

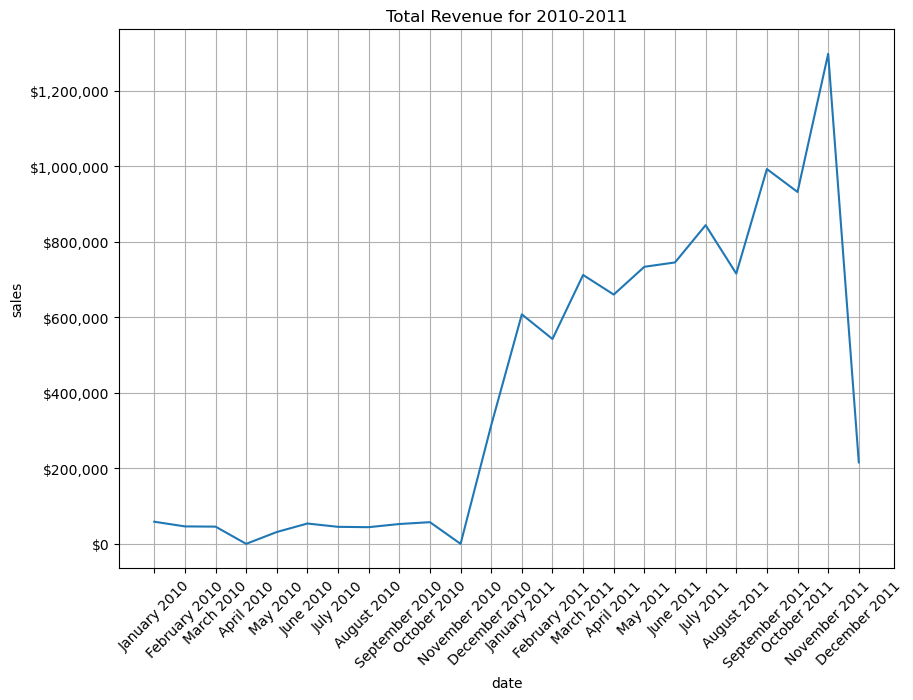
5. The jump from around 200 orders to over 1,500 orders represents roughly a 750% increase in business volume, suggesting a major business event occurred (possibly a new marketing strategy, product launch, or market expansion).

The most significant insight is the transformational change that occurred in late 2010, which marked a permanent shift from a small-volume operation to a much larger-scale business, followed by sustained growth

### **3.1.2Monthly Revenue Trend**

The first step in analysing monthly fluctuations in revenue was to calculate total sales for each transaction. This was achieved by introducing a new column, Sales, computed as the product of Quantity and Unit Price. The dataset was then aggregated on a monthly basis using Python’s Pandas library to sum the revenue for each month.

|  |  |  |
| --- | --- | --- |
| index | Month | Revenue($) |
| 1. | January, 2010 | 58635.560 |
| 2. | February, 2010 | 46207.280 |
| 3. | March, 2010 | 45620.460 |
| 4. | April, 2010 | 0.000 |
| 5. | May, 2010 | 31383.950 |
| 6. | June, 2010 | 53860.180 |
| 7. | July, 2010 | 45059.050 |
| 8. | August, 2010 | 44189.840 |
| 9. | September, 2010 | 52532.130 |
| 10. | October, 2010 | 57404.910 |
| 11. | November, 2010 | 0.00 |
| 12. | December, 2010 | 314063.660 |
| 13. | January, 2011 | 607748.110 |
| 14. | February, 2011 | 542511.540 |
| 15. | March, 2011 | 712059.510 |
| 16. | April, 2011 | 660149.571 |
| 17. | May, 2011 | 733769.320 |
| 18. | June, 2011 | 745252.340 |
| 19. | July, 2011 | 843891.231 |
| 20. | August, 2011 | 715632.720 |
| 21. | September, 2011 | 992555.452 |
| 22. | October, 2011 | 931852.160 |
| 23 | November, 2011 | 1297606.420 |
| 24. | December, 2011 | 215762.540 |



### Key Insights from the Monthly Revenue Trend:

1. **Steady Growth in Revenue**:
   * There is a noticeable upward trend in revenue, particularly from mid-2010 to late 2011. This indicates a period of business growth and increased customer activity.
2. **Seasonal Peaks**:
   * Revenue spikes are observed around November and December, suggesting that the business may benefit from seasonal events such as holiday shopping or end-of-year sales.
3. **Sudden Drop in December 2011**:
   * Despite the usual increase in revenue during the last quarter, December 2011 shows a sharp decline. This could indicate a data anomaly, an operational issue, or external factors such as economic downturns. According analysis this is due to the incomplete data recorded. Data spans from December 01 – 10, 2010
4. **Low Revenue Periods**:
   * The early months in the dataset (January to June 2010) show relatively low revenue, possibly indicating that the business was either in a nascent stage or undergoing challenges during that period.
5. **Opportunities for Analysis**:
   * The sharp increase from August 2010 could be linked to successful marketing campaigns, product launches, or expansion. Understanding what drove this growth can help replicate the success in the future.

## 3.2 Product Category Revenue Analysis

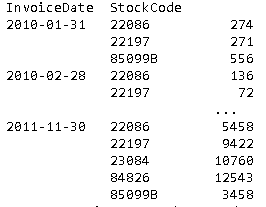
### 3.2.1 Products performing well over time

The table below show the top-performing based on total revenue over the time

|  |  |  |
| --- | --- | --- |
| Index | StockCode | Quanity |
| 1 | **84826** | **12543** |
| 2 | **23084** | **10760** |
| 3 | **22197** | **9422** |
| 4 | **85123A** | **6842** |
| 5 | **22086** | **5458** |
| 6 | **84879** | **5129** |
| 7 | **84077** | **4699** |
| 8 | **21108** | **4541** |
| 9 | **22578** | **3677** |
| 10 | **22577** | **3640** |

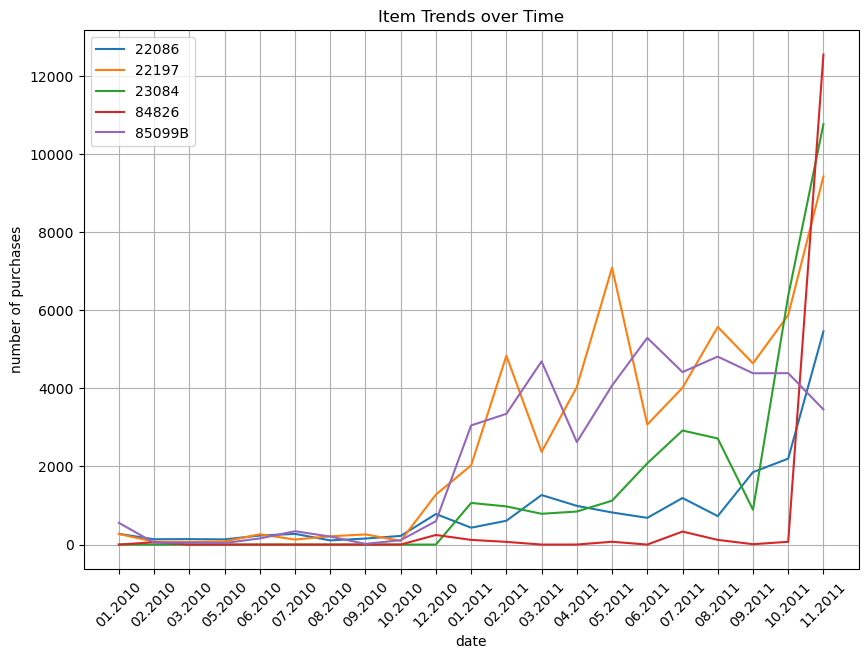
As we can see from the result, the products with codes 84826, 23084, 22797, 85123A and 22086 were the top best sellers in the month November 2011.

Now that we know what the top 5 sellers were in November 20, lets aggregate the monthly sales data for these five products.



### 3.2.2 Testing the Sustainability trend

The figure below shows the trend.



### Key Insights from the trend chart

1. Major Surge in Late 2011:

- Multiple items (particularly 84826 and 23084) show a dramatic spike in purchases in November 2011

- This surge is far above previous levels, reaching around 12,000 units for item 84826

2. Different Growth Patterns:

- Item 22197 (orange line) shows more volatility throughout 2011 with several peaks and troughs

- Item 85099B (purple line) shows moderate but consistent performance through most of 2011

- Item 22086 (blue line) maintains the most stable pattern with gradual growth

- Item 84826 (red line) remains relatively flat until the dramatic increase at the end

3. Timing of Changes:

- Similar to the previous graph, there's a notable change in behaviour starting around late 2010/early 2011

- Before this point, all items had relatively low and stable purchase numbers

- After this point, there's increased volatility and generally higher volumes

4. Performance Comparison:

- Items showed different levels of success:

- Some maintained steady growth

- Others showed high volatility

- A few remained relatively flat until the final surge

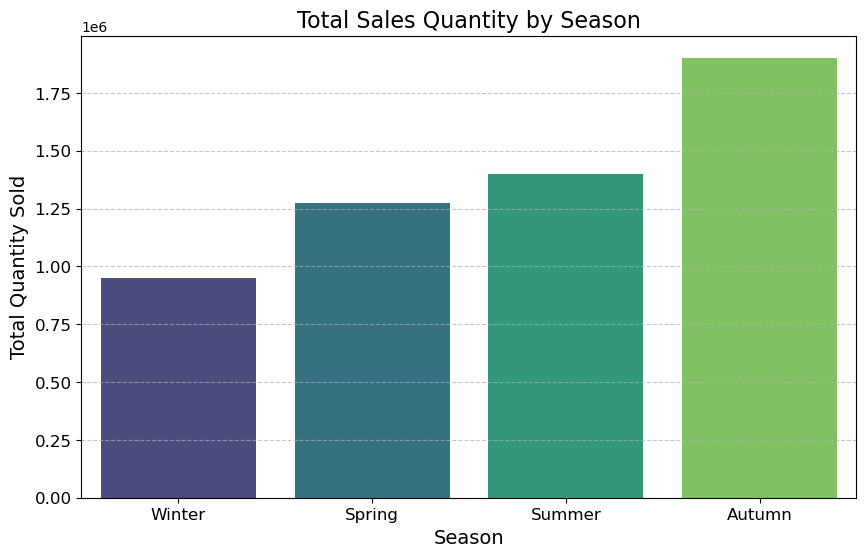
5. The most striking insight is the synchronized sharp uptick in multiple items at the end of the period, suggesting a possible seasonal effect or company-wide event affecting multiple product lines simultaneously.

## 3.3 Seasonal Sales Variation

The table below shows the product category variation across different seasons analysing the top performing products during each seasons ranging from summer, autumn, spring and winter.

|  |  |  |
| --- | --- | --- |
| Season | Description | Quantity Sold |
| Autumn | Paper craft, little birdie | 80995 |
| Winter | Medium ceramic top storage jar | 74327 |
| Autumn | Popcorn holder | 18715 |
| Spring | World war 2 Glider assisted designs | 18452 |
| Autumn | Rabbit Night Light | 18048 |
| Autumn | World war 2 gliders assisted design | 16053 |
| Summer | Jumbo bag red retrospot | 15554 |
| Autumn | Jumbo bag red retrospot | 12990 |
| Autumn | Assisted design 3D paper stickers | 12624 |
| Spring | Assorted Colours Silk Fan | 12311 |

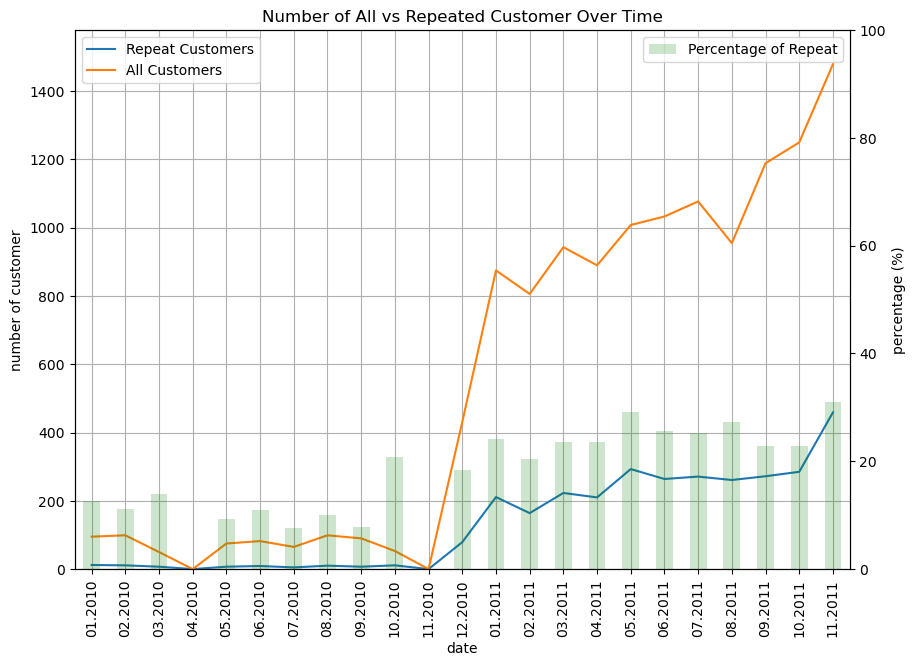
The total Sales Quantity by Season

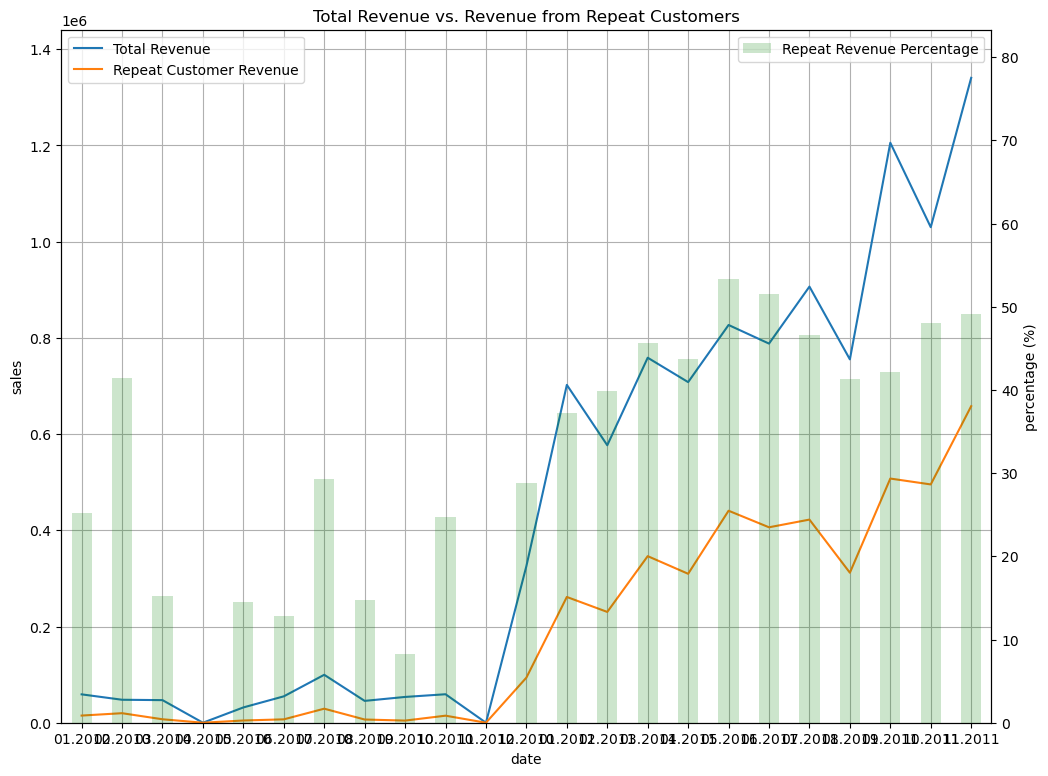
Generally, we have more sales during Autumn.

## 3.4 Customer Purchasing behaviour

**Objective:** To analyse shifts in purchasing patterns to find recurring trends.

* **Transaction Frequency and Spend Patterns:** Analysis of average transaction frequency and spending per customer revealed recurring behaviour patterns.
* **Customer Segmentation:** By grouping customers based on purchasing behaviour (e.g., high-frequency, high-spend customers), insights for targeted marketing campaigns were developed.

Fig 1. 



focusing on the 2010 to 2011 time period in the image:

Valuable Insights:

1. The total revenue increased significantly from 2010 to 2011, going from around 0.8 to over 1.4.
2. The revenue from repeat customers also grew rapidly during this period, going from around 0.2 to over 0.4.
3. The repeat customer revenue percentage increased from around 25% in 2010 to over 60% in 2011.

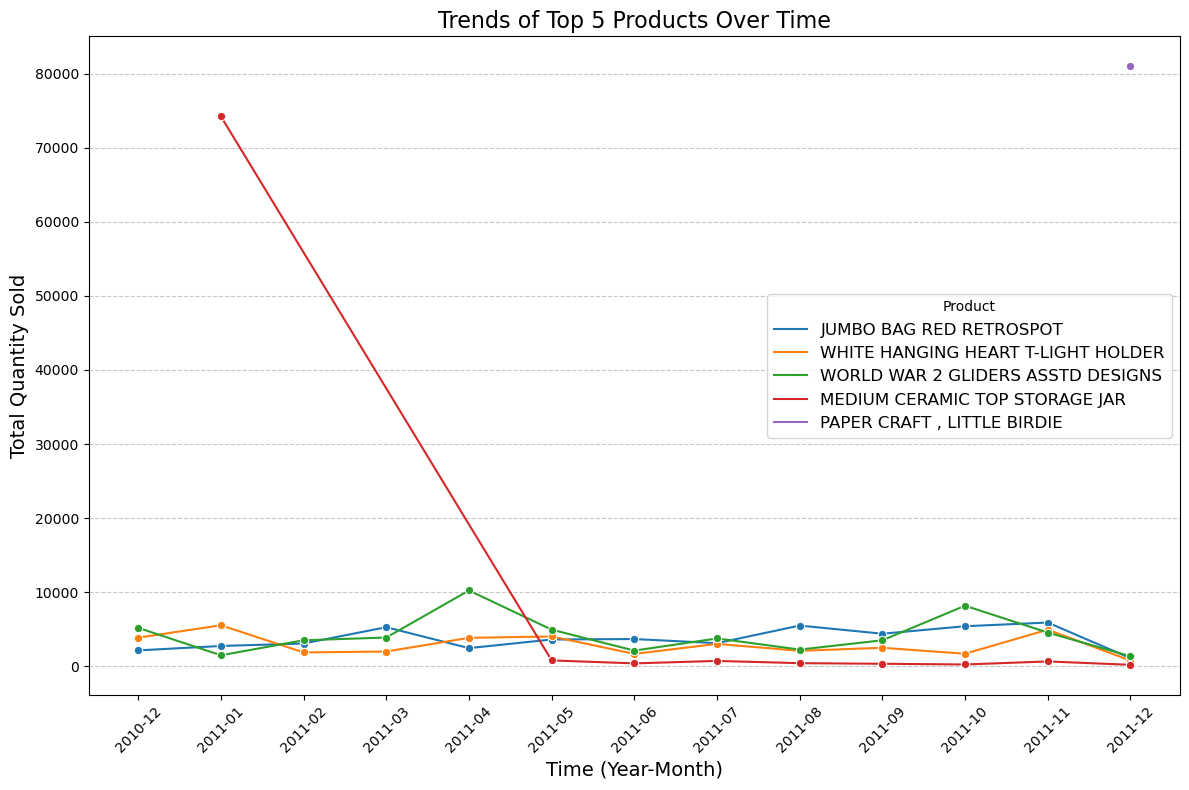
Action Insights:

1. The company should continue focusing on retaining and growing its repeat customer base, as this segment is driving a larger portion of total revenue.
2. Analyse what strategies or initiatives led to the sharp increase in repeat customer revenue from 2010 to 2011, and look for ways to further capitalize on these successful approaches.
3. Consider investing more resources into customer loyalty programs, repeat purchase incentives, or other initiatives that can help maintain and build upon the strong repeat customer growth seen in this time period.
4. Evaluate the overall customer acquisition and retention strategies to understand what is driving the total revenue growth, and see if there are opportunities to further optimize these efforts.

We went further to discover the top customers as well as the top 5 product overtime

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 13089.0  index | CustomerID | AvgTransaction Value($) | Avg Items Per Purchase | Unique Product Per Transaction | Purchase frequecy |
| 1. | 12748.0 | 160.570143 | 122.609524 | 20.400000 | 210 |
| 2. | 14911.0 | 715.547562 | 400.572139 | 28.218905 | 201 |
| 3. | 17841.0 | 330.577177 | 186.056452 | 60.282258 | 124 |
| 4. | 13089.0 | 606.451856 | 320.309278 | 18.680412 | 97 |
| 5. | 14606.0 | 130.716667 | 66.924731 | 28.602151 | 93 |
| 6. | 15311.0 | 667.779121 | 419.714286 | 25.901099 | 91 |
| 7. | 12971.0 | 130.115233 | 108.011628 | 3.546512 | 86 |
| 8. | 14646.0 | 3786.567838 | 2668.797297 | 28.108108 | 74 |
| 9. | 16029.0 | 1286.108571 | 638.222222 | 3.809524 | 63 |
| 10. | 13408.0 | 453.500645 | 261.806452 | 7.709677 | 62 |

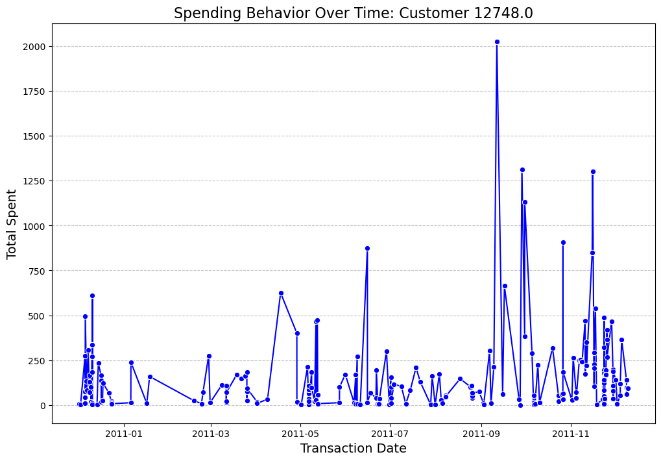
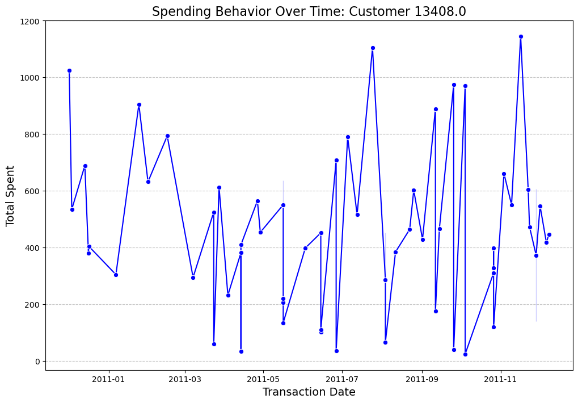
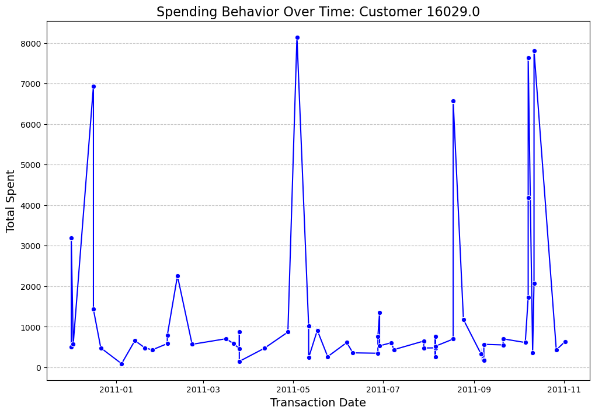
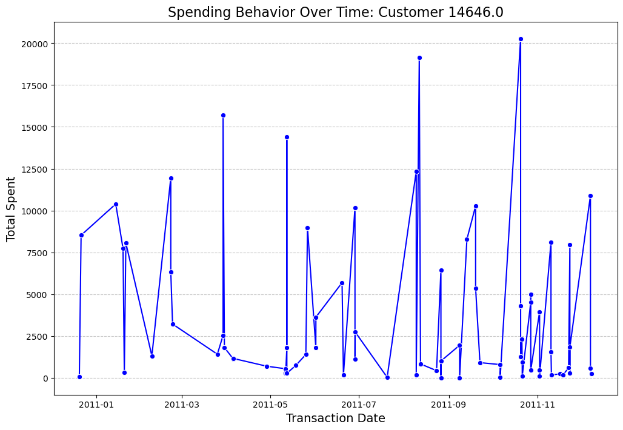
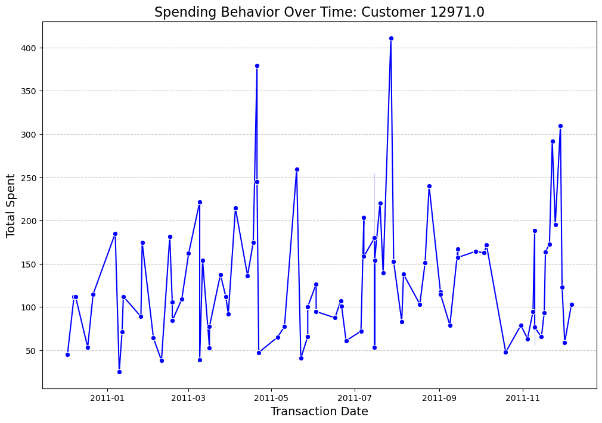
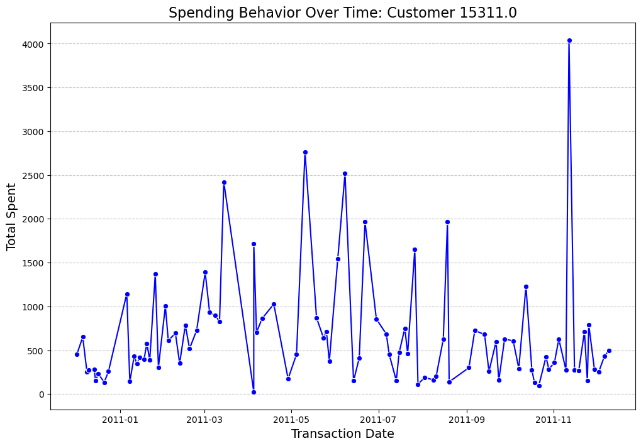
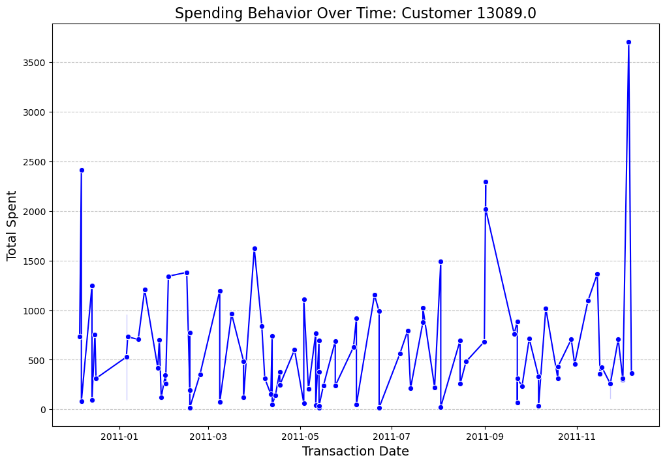
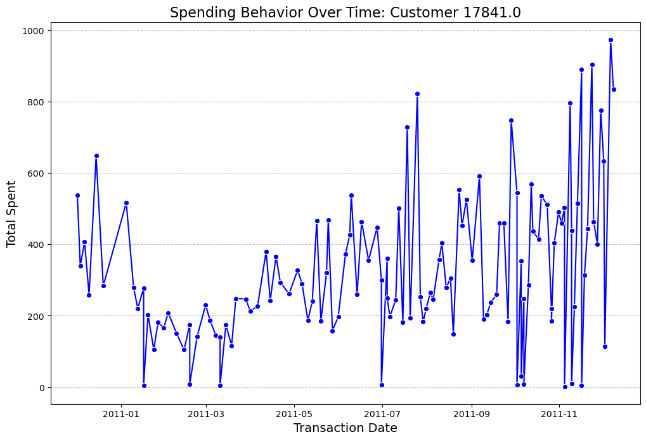
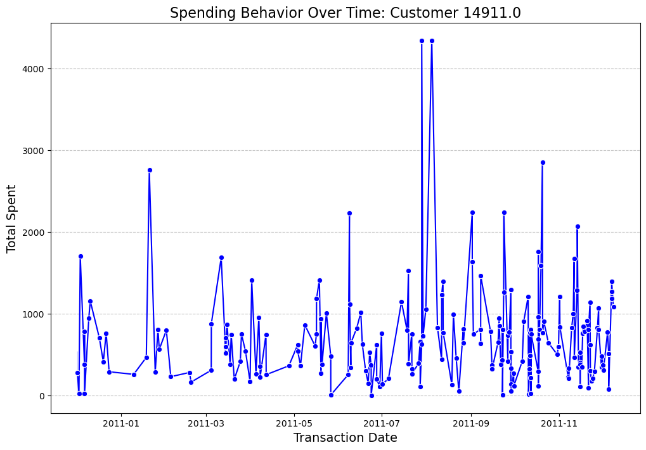
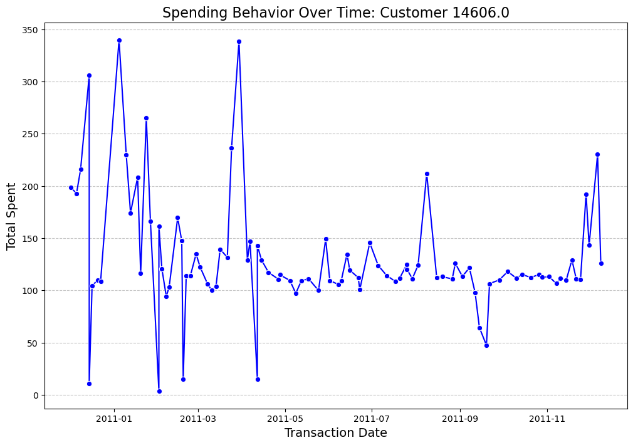
.



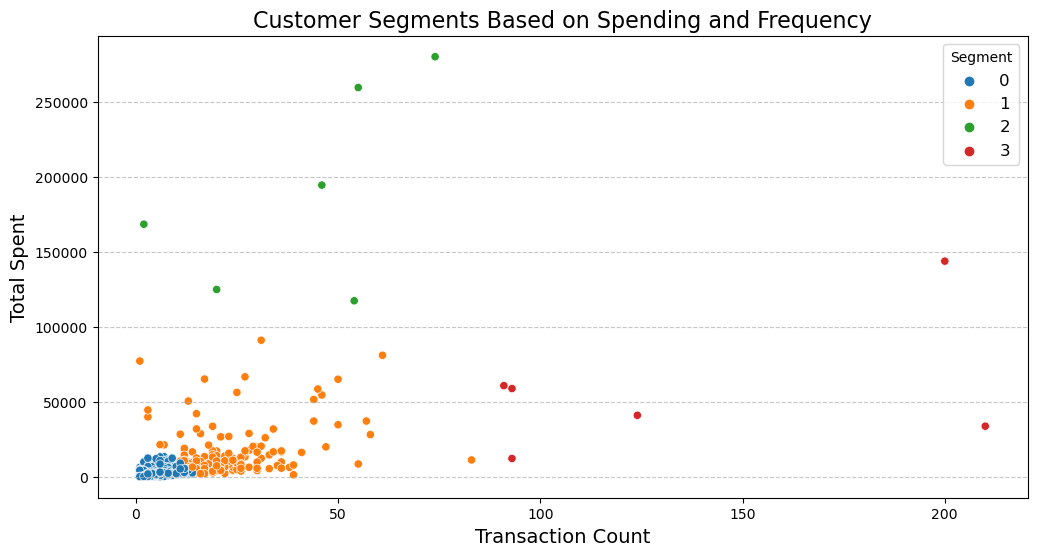
Analysing the spending pattern of each top customer

The plots below were used to study the spending pattern of top spending customer and some insight were gotten

|  |  |
| --- | --- |
| CustomerID | Peak spending periods |
| 14606 | First Quarter |
| 14911 | Third quarter |
| 17641 | Third Quarter, Fourth Quarter |
| 13089 | Fourth Quarter |
| 15311 | Second Quarter, Fourth Quarter |
| 12971 | Second Quarter, Third Quarter |
| 14646 | Second Quarter, Third Quarter |
| 16029 | Third Quarter, Third Quarter, Fourth Quarter |
| 13408 | First Quarter, Third Quarter, Fourth Quarter |
| 12748 | Third Quarter, partially Fourth Quarter |



Customer Segmentation



Legend interpretation

0 - Blue - These customers have a low transaction count and low total spend.

1- Orange -  These customers have a low transaction count but high total spend.

2 - Green - These customers have a high transaction count but moderate total spend.

3 - Red - These customers have a high transaction count but moderate total spend.

Here are the key insights and recommendations based on the customer segment analysis:

Insights:

1. Segment 0 (low spend, low frequency): These are likely occasional or new customers. Opportunities to increase their engagement and spend through targeted promotions, onboarding campaigns, and incentives to drive more frequent purchases.
2. Segment 1 (high spend, low frequency): These are high-value customers, but their infrequent purchases suggest potential churn risk. Focus on building loyalty through personalized offers, exclusive benefits, and proactive outreach to retain them.
3. Segment 2 (moderate spend, high frequency): These are the core, engaged customers. Maintain strong relationships through consistent positive experiences, loyalty programs, and cross-sell/upsell opportunities.
4. Segment 3 (high spend, high frequency): These are the most valuable, loyal customers. Prioritize providing exceptional service, personalized experiences, and exclusive perks to solidify their loyalty and advocacy.

Recommendations:

1. Develop targeted marketing strategies for each segment to optimize engagement and lifetime value:
   1. Segment 0: Acquisition campaigns, onboarding, and incentives to increase spend and frequency
   2. Segment 1: Retention initiatives, loyalty programs, and proactive outreach to prevent churn
   3. Segment 2: Loyalty programs, cross-sell/upsell, and continuous service improvements
   4. Segment 3: Personalized experiences, exclusive benefits, and opportunities for advocacy
2. Leverage customer insights to personalize the user experience, product offerings, and communication channels for each segment.
3. Continuously monitor and refine the segmentation model to adapt to changing customer behaviors and market conditions.
4. Identify opportunities to migrate customers between segments (e.g., from Segment 0 to Segment 2) through targeted initiatives that drive increased engagement and spending.
5. Analyze the drivers of customer behavior within each segment to uncover opportunities for optimization and growth.
6. The key is to use these customer insights to develop tailored strategies that maximize the lifetime value of each segment, while also fostering long-term loyalty and advocacy.

# 4.0. Summary of Insights and Recommendations

* Revenue Peaks and Promotions: Align marketing campaigns with identified seasonal peaks to capitalize on natural customer behaviour.
* Product Category Focus: Invest in categories with consistent revenue growth for a stable income source.
* Seasonal Inventory Adjustments: Stock levels for seasonal products should be managed in anticipation of high-demand periods.
* Customer Segmentation for Targeted Marketing: Tailor marketing efforts based on customer segmentation to increase engagement and retention.

# 5.0 Limitations and Caveats

1. Data Quality:

* Limited to 2010-2011 timeframe
* December 2011 data incomplete (only first 10 days)
* Cancelled orders removed from analysis

1. Analysis Constraints:

* Geographic variations not deeply analysed
* Price sensitivity not fully explored
* Limited customer demographic information