Customer Segmentation Analysis and Sales Dashboard

Group Number: 3

November 27, 2024

Contents

1 Overview of the Project

This project uses Python and its data analysis libraries to create a customer segmentation dashboard for marketing and sales teams. By segmenting customers based on purchasing behavior, we aim to provide actionable insights for improving retention, optimizing inventory, and increasing sales. By analyzing these segments, marketing, and sales campaigns can be tailored for maximum efficiency and effectiveness.

Key challenges include handling an extensive dataset with missing values, ensuring data consistency for clustering, and providing meaningful interpretations of customer behavior. The methods employed address these challenges while aligning with the project's strategic objectives.

Objective

This project employs clustering techniques to segment customers based on purchasing behavior, frequency, and spending, providing valuable insights into distinct customer profiles for strategic retention and growth. By examining sales trends, revenue generation, and category performance alongside daily and seasonal sales patterns, we aim to provide improved inventory and marketing and sales strategies. An interactive sales dashboard will offer marketers data-driven visualizations to optimize product promotion, resource allocation, and campaign effectiveness.

Our Team

- **Suranjana Sarkar:** Data reprocessing, revenue contribution, and product segmentation analysis and visualization. Preparing the report.
- **Romaisa Ali Bhutta:** Customer purchase pattern and daily sales trend analysis and visualization. Creating dashboard.
- **Suraj Kanwar:** Order Behavior analysis and marketing campaign efficiency analysis and cluster visualization. Preparing the report.

2 Report and Analysis

2.1 Dataset Source

The dataset used in this analysis is the Online Retail Dataset, which contains information on products sold, quantities, prices, and transaction dates. Link: UCI Repository

Variables Table

Variable Name	Role	Type	Description
InvoiceNo	ID	Categorical	A 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'C', it indicates a cancellation.
StockCode	ID	Categorical	A 5-digit integral number uniquely assigned to each distinct product.
Description	Feature	Categorical	Product name.
Quantity	Feature	Integer	The quantities of each product (item) per transaction.
InvoiceDate	Feature	Date	The day and time when each transaction was generated.
UnitPrice	Feature	Continuous	Product price per unit.
CustomerID	Feature	Categorical	A 5-digit integral number uniquely assigned to each customer.
Country	Feature	Categorical	The name of the country where each customer resides.

Table 1: Key fields of variables in the dataset.

2.2 Methods Used on Dataset

Preprocessing

The preprocessing of the dataset is critical for ensuring the quality and reliability of subsequent analyses. Below, we detail and justify each step undertaken in the code to clean, normalize, and handle missing values within the dataset.

- **Checking Missing Values:** Initially, the dataset is examined for missing values using the `isnull().sum()` function. This step identifies the extent of missing data in each column, enabling informed decisions on how to handle these gaps.
- **Handling Missing Descriptions:** Rows where the `Description` column is null are removed. This decision is justified as the `Description` column contains critical information about the products. Retaining rows without this information could lead to inaccuracies in product-level analysis.
- **Filling Missing Customer IDs:** Missing values in the `CustomerID` column are replaced with `0`. This approach ensures that the dataset remains

complete for analyses requiring this field. While this may group unidentified customers under a common placeholder, it avoids omitting potentially valuable transaction data.

- **Removing Duplicates:** Duplicate rows are dropped from the dataset using the `drop_duplicates()` function. Duplicates can lead to inflated metrics and misrepresentations of customer behavior, and their removal ensures data integrity.

- **Filtering Invalid Data:** Rows with negative values in the `Quantity` or `UnitPrice` columns are removed. Negative quantities or prices may indicate returns or data entry errors. These transactions, unless specifically relevant, are excluded to focus on valid sales data.

- **Date Formatting:** The `InvoiceDate` column is converted to datetime format using `pd.to_datetime()`. This conversion ensures consistency in date representation and facilitates time-series analysis.

Visualization

Visualization is a critical aspect of data analysis, enabling the intuitive understanding of patterns, distributions, and relationships. The code utilized the powerful `matplotlib` and `seaborn` libraries to create meaningful visual representations through bar charts, histograms, and scatter plots.

1. A bar chart is used to compare discrete categories, such as product descriptions, against a numerical metric like order count. This makes it particularly useful for identifying the most popular products or categories. Easy to interpret and visually intuitive. Provides a clear comparison between categories. Effective for identifying trends or outliers in categorical data.

2. A histogram were used to visualize the frequency distribution of a numerical variable, such as the number of orders per product. It helps to identify the shape of the distribution (e.g., normal, skewed) and detect outliers or irregular patterns. Displays the spread and variability of data. Highlights central tendencies like mean and median. Useful for understanding the overall behavior of a dataset.

3. Scatter plots, combined with Principal Component Analysis (PCA) from the `sklearn` library, visualized customer clusters in various dimensions. This approach simplifies high-dimensional data, making patterns and clusters easier to interpret. Simplifies complex, multi-dimensional data for better visual understanding. Reveals groupings or separations among data points. Useful for presenting clustering results

Clustering

Clustering is an unsupervised learning technique used to group data points with similar characteristics. Applied K-means clustering to group products and customers based on revenue, sales frequency, and other metrics.

- **Purpose:** K-Means clustering partitions data into a pre-defined number of clusters based on feature similarity, minimizing the variance within clusters while maximizing the variance between clusters.
- **Advantages:**
  - Simple and efficient for large datasets.
  - Produces well-defined, non-overlapping clusters.
  - Useful for customer segmentation, identifying distinct groups within a population.
- **Considerations:**
  - The number of clusters (k) must be pre-determined, typically using techniques like the Elbow Method to balance the trade-off between model complexity and fit.
  - Sensitive to outliers, which can distort cluster centroids and reduce clustering accuracy.

Evaluation Metrics

Silhouette Score and Elbow Method to determine optimal cluster count.

2.3 Analysis Performed

Six key analyses were conducted on the dataset:

1. **Revenue Contribution Analysis:** Identifying products or categories contributing the most to overall revenue.

2. **Product Segmentation:** Grouping products based on sales volume, revenue, and daily sales trends.

3. **Customer Purchase Patterns:** Segmenting customers based on order frequency and total spending.

4. **Daily Sales Trends:** Analyzing daily sales data to identify peak periods and consistent performers.

5. **Order Behavior Analysis:** Examining the distribution and frequency of total orders per product.

6. **Customer segmentation clustering:** Utilizing cluster insights in multiple customer segmentation.

7. **Dashboard:** Creating a sales Dashboard offering an interactive and insightful overview of the sales data depicted by the Online Retail dataset

2.4 Results and Findings

2.4.1 Top 20 Products contributing in total revenue

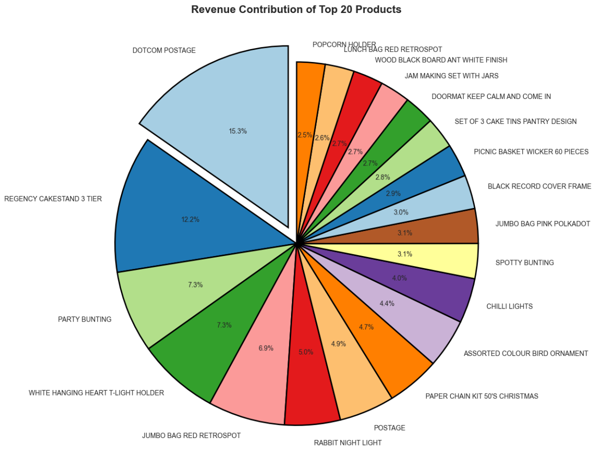


Figure 1: This pie chart represents the top 20 items which bring the most revenue and also shows the contribution in percentage. The top product is **DOTCOM POSTAGE** with a revenue of 15.3% and the 20th item is **Popcorn holder** with a revenue of 2.5%. Also, for better understanding, there is a bar graph.

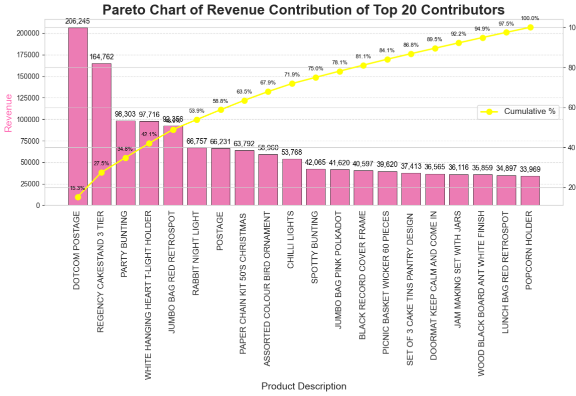


Figure 2: Top 20 Products of Revenue Contribution

Purpose and Insights:

- Here, the x-axis represents the Revenue, and the y-axis represents the product descriptions. Also here is a second y-axis showing the revenue percentage growth of the top 20 items.
- This graph helps identify which products are the top revenue contributors and which are under-performing. This can guide business decisions, such as focusing on high-revenue products or improving the performance of low-revenue ones.

2.4.2 Top 20 Products by Order Count

The following chart displays the top 20 products based on their order count. This analysis identifies high-demand items, which are critical for inventory management and marketing strategies.

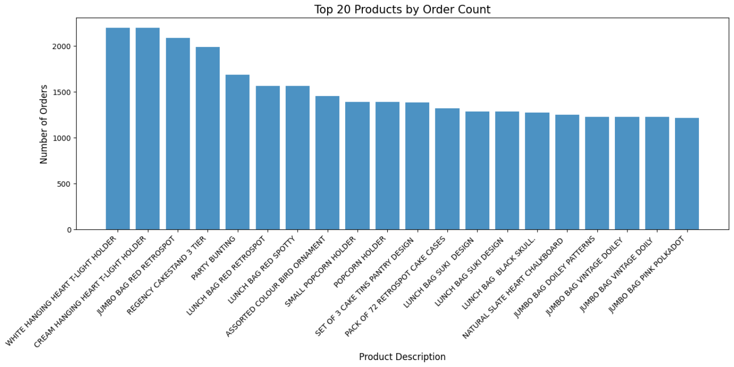


Figure 3: Top 20 Products by Order Count

Insights from the Chart:

- The top products include items such as *White Hanging Heart T-Light Holder* and *Cream Hanging Heart T-Light Holder*, which have the highest order counts, exceeding 2,000 orders.
- Many of the top-performing products belong to similar categories, suggesting a trend in customer preferences for decorative or utility items.
- These high-demand products are key revenue drivers and should be prioritized for stock replenishment and promotional activities.

Recommendations: Focus marketing campaigns around top-selling products to boost overall sales. We can also monitor inventory levels of these high-demand items to prevent stockouts and finally, investigate potential cross-selling opportunities with other products in similar categories.

2.4.3 Product Segmentation

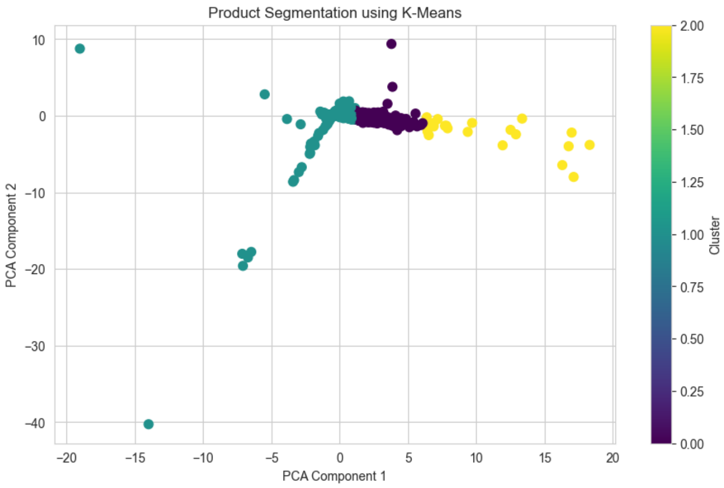


Figure 4: The above graph shows the product segmentation based on 3 features: Avg Sales Volume, Avg Revenue, and Avg Daily Sales.

Clustering: Here we can understand that:

- **Cluster 1: Avg Sales Volume** Products with high sales volume but average revenue and daily sales. These products could have low prices but are bought very frequently. Focus on promotions and discounts to maintain high sales volume.

- **Cluster 2: Avg Revenue** High revenue with low sales volume. These could be high-value products with a high selling price but a lower frequency of purchase. Consider increasing visibility or bundling with other products to boost sales volume.

- **Cluster 3: Avg Daily Sales** Products with moderate sales volume and revenue, while having high daily sales trends. This could be products that, over time, are gaining popularity. Monitor trends and potentially introduce new products in this category to capitalize on the growing popularity.

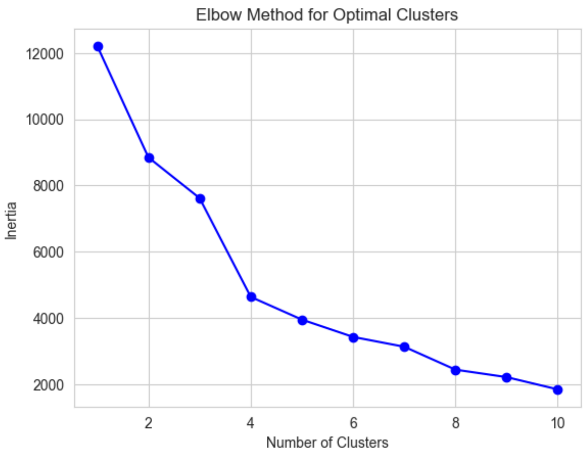


Figure 5: Elbow Plot to Determine Optimal Number of Clusters

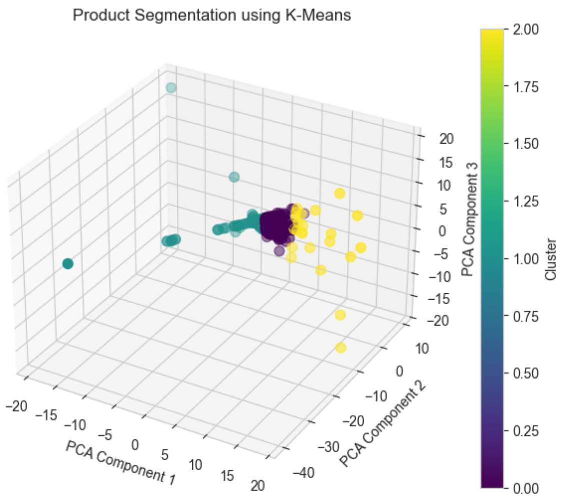


Figure 6: 3-D Cluster Model Visualization

Cluster	Avg Sales Volume	Avg Revenue	Avg Daily Sales
0	7256.64	13294.07	45.21
1	642.64	1185.29	2.00
2	25716.17	57314.59	102.77

Table 2: Product Segmentation Results based on K-Means Clustering

Product Segmentation Findings

- Based on the K-Means clustering analysis, the products have been segmented into three clusters. Below is the summary of the clusters, showing the average sales volume, average revenue, and average daily sales for each cluster:

- **Cluster 1**: Products with moderate sales (7,257 units) and revenue (13,294). These items have steady daily sales (45), indicating consistent demand. Stable yet there is potential for increased exposure.
- **Cluster 2:** Products with low sales (643 units) and revenue (1,185), and very low daily sales (2). These items are likely niche and need targeted promotion. Marketing focus required for niche products.
- **Cluster 3:** Products with high sales (25,716 units) and revenue (57,315), along with high daily sales (103). These are high-demand items that perform well consistently. High performers require strategies to maintain momentum.

2.4.4 Customer purchase pattern

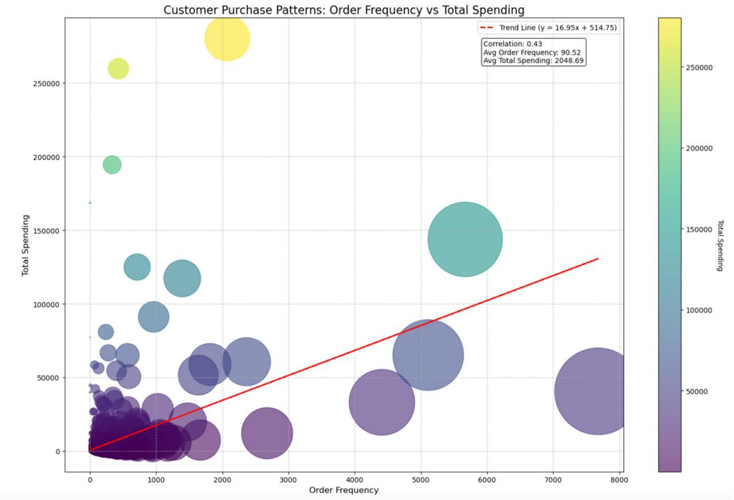


Figure 7: The scatter plot displays customer purchase patterns showing order frequency in comparison to total spending. Each point represents a customer, where the size of the point indicates order frequency and color reflects total spending. A trend line indicates the overall relationship between these two variables, suggesting how frequently customers purchase correlates with how much they spend.

Insights from the Chart:

- The majority of customers fall in the low-frequency, low-spending buyer category, while a small portion of high-frequency, high-spending customers generates most of the revenue. The occasional big spenders which are low frequency, and high spending represent unique opportunities.
- There is a moderate positive relationship between order frequency and total spending, showing that with every additional order, an average customer spent $16.95 more.

- Customers in the top-right region have both high order frequency and high spending, making them the most valuable contributors to revenue. These customers form a smaller, distinct group in the plot above.
- The top-left outliers represent customers with low order frequency but exceptionally high spending. These could be bulk buyers or customers who purchase high-ticket items. The middle-right outliers show customers with high order frequency but moderate spending, possibly suggesting loyalty-driven behavior with smaller purchases.

Recommendations

- Find customers with high order frequency but lower spending for potential upselling or cross-selling strategies.
- Design re-engagement strategies for low-frequency customers, offering them bundle deals or limited-time promotions to convert them into repeat buyers.
- Reward frequent customers with points, discounts, or free shipping. Moreover, use gamification to incentivize higher spending or more orders.

2.4.5 Daily Sales Trend

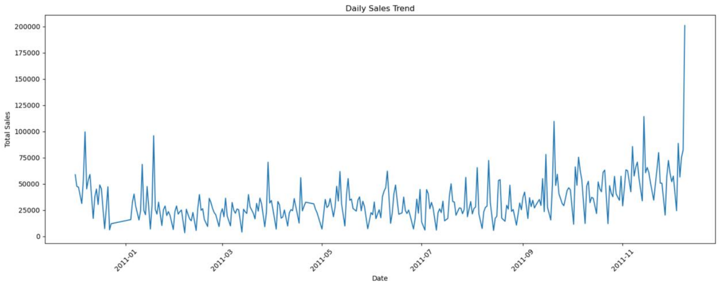


Figure 8: The graph shows the total sales amount on the y-axis and dates on the x-axis. Each point on the line represents the total sales for a specific day, and the line connects these points to show the trend over time.

Insights from the Chart:

- **Consistent Base Level:** While the sale generally hovers between $\$25,000$ and $\$50,000$ on most days, there are frequent spikes and occasional dips, indicating an active sales pattern with variations.
- **Trends:** Several days have a sharp rise in sales, with the largest peak nearing $\$200,000$. These peaks probably reflect special events, holidays, or promotional campaigns. There are dips below $\$25,000$ which are scattered throughout, with small clusters at certain periods. This decline could indicate off-peak seasons, operational downtime, or missed opportunities for engagement.

- **Year-End Growth:** A clear upward trend can be seen at the end of the year because of holiday shopping, such as Black Friday and Christmas, and year-end sales. This growth reflects the impact of seasonal demand and may indicate a strategic period for maximizing revenue.

Recommendations:

- Identify the reasons behind the steep peaks in sales, for instance, promotions, holidays, or product launches, and reproduce successful strategies. Plan extensive marketing campaigns and product launches on days that have traditionally seen high sales.
- Introduce loyalty or subscription models to encourage customers to continue buying through the week and not rely on peak days. Introduce targeted discounts, flash sales, or email campaigns to drive traffic during traditionally low-performing days.

2.4.6 Customer Segmentation using PCA

The segmentation of customers was performed using the K-means clustering algorithm. To better understand and visualize the clusters, PCA (Principal Component Analysis) was applied to reduce the dimensionality of the dataset to two components. This chart shows the distribution of customers across distinct clusters based on their purchasing behavior.

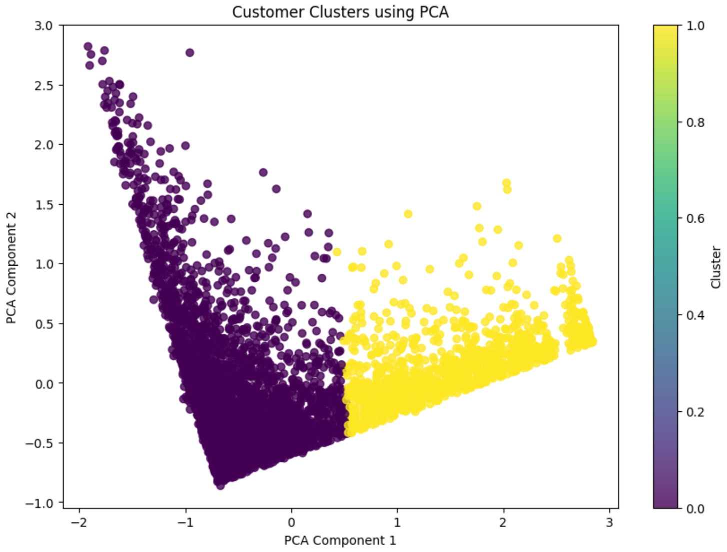


Figure 9: Product Clusters Visualized in 2D Space

Insights from the Chart: The distinct separation between clusters indicates the success of segmentation. A majority of customers fall into Cluster 1 (Purple), suggesting a need for re-engagement strategies. However, Cluster 2 (Yellow) highlights a smaller, high-value customer segment critical for revenue.

- **Cluster 1 (Purple):** Represents customers with lower overall purchasing activity, characterized by:

  - Low recency (they haven't made recent purchases).
  - Low frequency (infrequent buyers).
  - Moderate to low monetary value.
  - Potentially one-time buyers or customers who have churned.

  Marketing Recommendation: Design re-engagement campaigns or loyalty programs to bring back these customers.

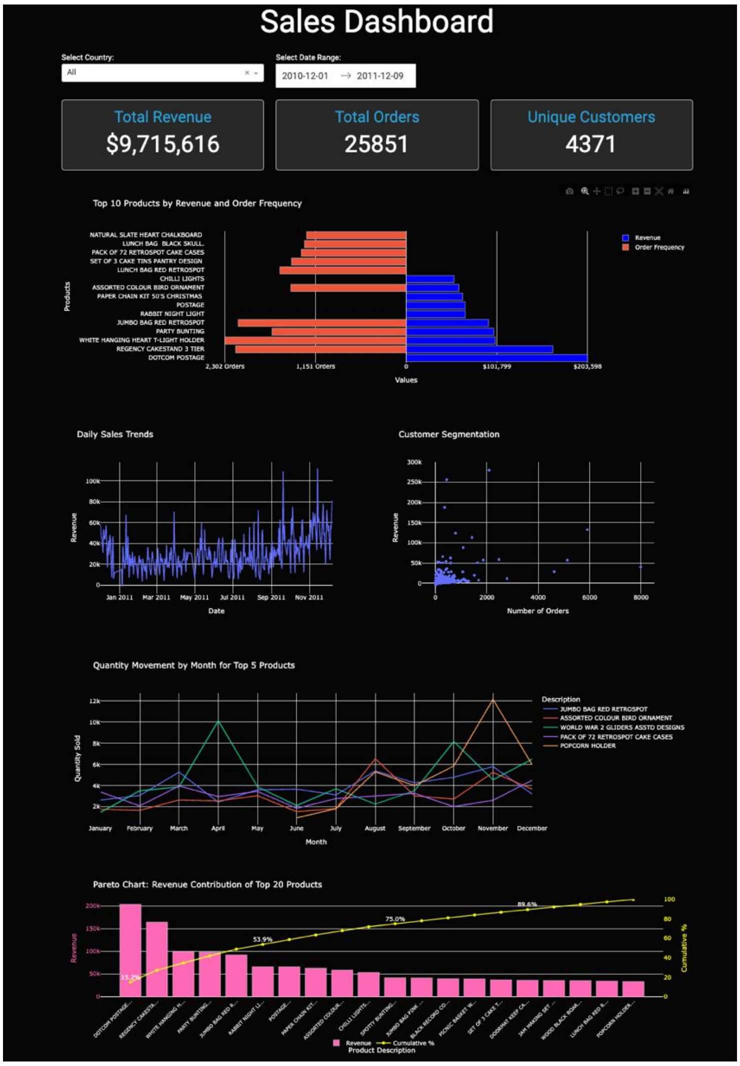
- **Cluster 2 (Yellow):** Represents high-value customers with:
  - High recency (recent purchases).
  - High frequency (frequent buyers).
  - High monetary value (significant contributors to revenue).

  Marketing Recommendation: Focus retention efforts on these customers, offer personalized promotions, and target them with exclusive deals to maintain loyalty.

The insights derived from the analysis have significant potential to enhance product performance and revenue growth.

2.5 Dashboard

- **Overview** This sales dashboard interacts with the Online Retail dataset to highlight sales patterns, product performance, and customer behavior in intuitive charts and KPIs. Additionally, country and date range filtering is supported for focused analysis.



Sales Dashboard

Select Country: All

Select Date Range: 2010-12-01 → 2011-12-09

Total Revenue $9,715,616

Total Orders 25851

Unique Customers 4371

Figure 10: Sales analysis Dashboard

- **KPI Metrics**

- Total Revenue: This shows the total revenue generated for the selected filters.

- Total Orders: A count of unique orders placed, considered as invoices.

- Unique Customers: The number of customers placing orders is considered unique.
- These KPI's are added to provide an overall snapshot of the performance, enabling the organization to gauge business health.
- **Date Country Filters** The country filter is added to allow the business to analyze performance across different regions, making it easy to identify strong or weak markets. Similarly, the date filter lets them zoom in on specific timeframes, such as peak holiday periods or promotional campaigns, to track their impact. This flexibility makes the dashboard adaptable to diverse scenarios.
- Charts and Visualizations
- The dashboard was designed to ensure usability and visual clarity. The layout is structured to flow logically, starting with high-level KPIs at the top, followed by detailed charts that support deeper analysis.

Sales Dashboard Insights

Visualizations

1. **Pyramid Plot:**
   This shows the top 10 products by revenue vs. order frequency. It helps to identify the products with high revenue as compared to products that are ordered frequently, identifying both high-value and high-demand items.
2. **Daily Sales Trend-Line Chart:**
   It captures the daily revenue over time, highlighting sales peaks and trends for a selected time-period. This makes the periodic trends easily detectable.
3. **Customer Segmentation-Scatter Plot:**
   Maps customers based on the number of orders against their revenue contribution. This helps in identifying high-value customers and highlights potential opportunities to retain such customers.
4. **Quantity Movement-Line Chart:**
   This displays top 5 products by quantity sold across months, showing the overall sales trend. It is helpful to show seasonality or monthly demand, aiding inventory and marketing planning.
5. **Pareto Chart:**
   Combines revenue bars with a cumulative percentage line for the top 20 products. This helps focus on products contributing to the majority of the revenue.

Applications of the Sales Dashboard

This dashboard assists retail businesses in optimizing strategies for better growth through actionable insights:

- **Product Optimization:** Targeted promotions for high-revenue items and highlighting frequently ordered low-revenue products for bundling or upselling opportunities.
- **Customer Segmentation Insights:** Enable nurturing of high-value customers with loyalty programs and engagement of occasional big spenders with personalized offers to increase their purchase frequency.
- **Seasonal Trends:** The Quantity Movement and Daily Sales Trend charts support demand forecasting and inventory alignment to prepare for peak sales periods and avoid overstocking or stockouts.
- **Sales Campaign Planning:** Analyzing performance by day to replicate successful strategies and mitigate low-sale periods with flash sales or discounts.
- **Regional and Temporal Adaptation:** Filters enable businesses to tailor strategies to specific markets and timeframes, making the dashboard crucial for data-driven decision-making.
- The dashboard is a powerful tool that drives informed decisions across multiple aspects in an online retail business. It enables the observation of seasonal trends through the sales data for specific date ranges so that businesses can optimize inventory and marketing strategies during the peak time, such as holidays or promotional events. Moreover, businesses can uncover regional preferences, tailoring their campaigns and product offerings to specific markets. Visualizations like the Pyramid Plot and Pareto Chart provide insights for product strategy, allowing a focus on high-revenue items while addressing frequently ordered but low-margin products.
- In conclusion, this dashboard seamlessly integrates interactive features, user-friendly design, and purpose-driven visualizations to support strategic decision-making for the online retail business.
- **Future Potential of the Dashboard**
- **Real-time updates:** The dashboard can be integrated with a live data feed to provide real-time numbers enabling hourly performance, allowing to act timely on declining performance.
- **Scalability:** With additional data integrations, such as supplier performance, customer demographics, and cost structures, the dashboard could further enhance predictive analytics, demand forecasting, and resource allocation, benefiting all teams.
- **• Enhanced Visualization:**Introduce advanced visualizations like tree maps, or funnel charts to represent complex data. These charts can reveal bottlenecks or drop-off points in customer engagement or purchase processes.

3 Conclusion

This project successfully demonstrates the use of clustering techniques and data visualization to derive actionable marketing insights. The dashboard allows marketers to:

- Focus on high-performing product categories.
- Develop targeted campaigns for low-engagement items.
- Optimize resource allocation based on product and customer segments.

Attachments

The following files are included in the email:

- **Report:** This document (PDF).
- **Dataset:** The Online retail dataset in XLSX.
- **Code:** Python scripts for data pre-processing, clustering, analysis, and dashboard creation.

Email Subject

Subject: Group 3 - Customer Segmentation Dashboard Project