# CAPSTONE PROJECT

# (MAHFUJ RAZA)

# Myntra Online Retail Customer Segmentation

# Unsupervised ML

## Abstract

This project analyzes the online retail operations of Myntra Gifts Ltd., a UK-based division of Myntra specializing in all-occasion giftware. The dataset spans from December 1, 2009, to December 9, 2011, and includes transaction records from its non-store online platform. The objective of this study is to derive actionable insights to optimize business strategies. By examining purchasing trends, product performance, customer behavior, pricing strategies, and inventory management, we aim to enhance decision-making for sales growth and operational efficiency. Exploratory Data Analysis (EDA) techniques are employed to uncover patterns and relationships within the data, aiding in strategic recommendations.

## Introduction

Myntra is a leading Indian fashion e-commerce company, and Myntra Gifts Ltd. is its UK-based division focusing on giftware retail through an online platform. With the increasing competition in the e-commerce space, businesses must leverage data-driven insights to optimize their strategies. This project delves into the sales data of Myntra Gifts Ltd. to identify key trends and patterns that can improve product selection, pricing strategies, and customer engagement. The study involves data exploration, trend identification, and clustering techniques to enhance the company's market positioning.

## Problem Statement

Myntra Gifts Ltd. operates in the online retail sector, where understanding consumer behavior, optimizing product offerings, and managing inventory efficiently are crucial for profitability. The company seeks insights into purchasing trends, product demand, and pricing impacts to improve marketing efforts, streamline inventory planning, and enhance customer segmentation. The key challenge is to analyze the dataset effectively to extract meaningful patterns that can inform business decisions and drive sales growth.

## Data Description

Before starting any analysis, it's important to familiarize ourselves with the dataset. After loading it, we reviewed the first and last few rows and confirmed the dataset contains 541909 rows and 8 feature columns.

The dataset includes the following features:

- **InvoiceNo**: Unique identifier for each transaction.
- **StockCode**: Product code.
- **Description**: Name of the product.
- **Quantity**: Number of units purchased.
- **InvoiceDate**: Date and time of purchase.
- **UnitPrice**: Price per unit of the product.
- **CustomerID**: Unique identifier for each customer.
- **Country**: Country where the customer is located.

This dataset provides a comprehensive view of sales transactions, allowing for detailed analysis of purchasing behavior and product performance.

# Exploratory Data Analysis (EDA)

## Data Cleaning and Preprocessing:

- Handling missing values in 'CustomerID' and 'Description' fields.
- Converting product descriptions to lowercase and standardizing the 'InvoiceDate' format in the data cleaning and preprocessing section.

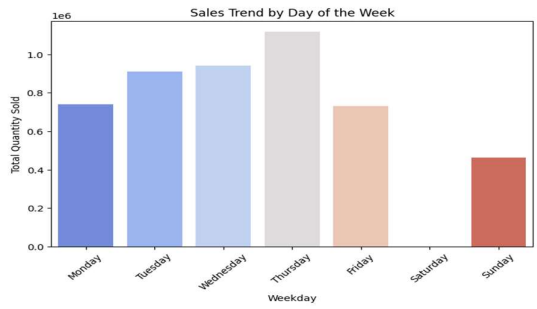
# Data Visualization:

## Identifying Purchasing Trends:

### Analyze Monthly Sales Trends

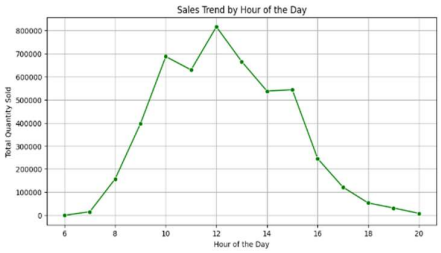

The peak sales volume occurs in November 2011

### Identify Weekly & Daily Trends



Sales reach their highest point on Thursday, indicating a strong demand on this day.

**Identify Peak Purchase Hours**




The peak hours are 11am to 1pm and at 12pm the highest sales reach

# Evaluating Product Performance:

**Identify Best-Selling Products**

Here is the list of top 10 best selling product:-

| | StockCode | Description | Quantity |
|---|---|---|---|
| 3028 | 84077 | world war 2 gliders asstd designs | 53215 |
| 3460 | 85099B | jumbo bag red retrospot | 45066 |
| 3288 | 84879 | assorted colour bird ornament | 35314 |
| 3475 | 85123A | white hanging heart t-light holder | 34147 |
| 434 | 21212 | pack of 72 retrospot cake cases | 33409 |
| 1112 | 22197 | popcorn holder | 30504 |
| 2010 | 23084 | rabbit night light | 27094 |
| 1387 | 22492 | mini paint set vintage | 25880 |
| 1509 | 22616 | of 12 london tissues | 25321 |
| 930 | 21977 | pack of 60 pink paisley cake cases | 24163 |

The best selling product is world war 2 gliders asstd designs.

**Identify Least-Selling Products**

Here is the list of least 10 selling product:-

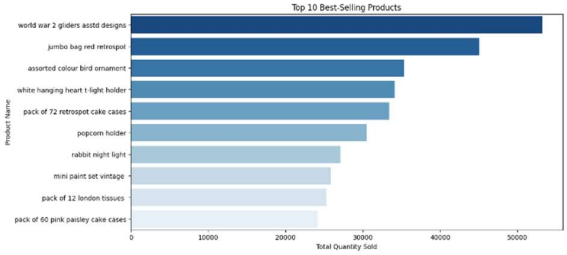
| | StockCode | Description | Quantity |
|---|---|---|---|
| 3062 | 84347 | rotating silver angels t-light hldr | -1460 |
| 3911 | D | discount | -1194 |

| | | | |
|---|---|---|---|
| 702 | 21645 | assorted tutti frutti round box | -24 |
| 3910 | CRUK | cruk commission | -16 |
| 387 | 21144 | pink poodle hanging decoration | -12 |
| 957 | 22034 | robin christmas card | -9 |
| 2970 | 79323W | white cherry lights | -8 |
| 2618 | 35400 | wooden box advent calendar | -6 |
| 161 | 20703 | blue padded soft mobile | -6 |
| 567 | 21412 | vintage gold tinsel reel | -6 |

Negative quantities indicating returns

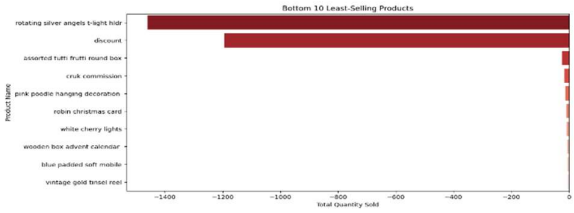
rotating silver angels t-light hldr is least-selling or highest returing product

## Visualizing Product Performance




The best selling product is world bar 2 gliders asstd design.

## Bottom 10 Least-Selling Products




The least selling product is rotating silver angles t-light hldr.

## Analyzing Product Revenue

- The highest-selling product is "Regency Cakestand 3 Tier" with a revenue of 132,870.40, followed by:
- White Hanging Heart T-Light Holder (93,823.85)
- Jumbo Bag Red Retrospot (83,236.76)
- Party Bunting (67,687.53)

# Understanding Customer Behavior:

## Analyze Customer Purchase Frequency

| | CustomerID | TotalPurchases |
|---|---|---|
| 1895 | 14911.0 | 248 |
| 330 | 12748.0 | 224 |
| 4042 | 17841.0 | 169 |
| 1674 | 14606.0 | 128 |
| 568 | 13089.0 | 118 |
| 2192 | 15311.0 | 118 |
| 487 | 12971.0 | 89 |
| 1615 | 14527.0 | 86 |
| 803 | 13408.0 | 81 |
| 1703 | 14646.0 | 77 |

Customer 14911 has made the highest number of purchases (248 transactions), indicating high engagement and loyalty.

## Analyze Average Purchase Value Per Customer

| | CustomerID | TotalSpending |
|---|---|---|
| 1703 | 14646.0 | 279489.02 |
| 4233 | 18102.0 | 256438.49 |
| 3758 | 17450.0 | 187482.17 |
| 1895 | 14911.0 | 132572.62 |
| 55 | 12415.0 | 123725.45 |
| 1345 | 14156.0 | 113384.14 |
| 3801 | 17511.0 | 88125.38 |
| 3202 | 16684.0 | 65892.08 |
| 1005 | 13694.0 | 62653.10 |
| 2192 | 15311.0 | 59419.34 |

Customer 14646 is the highest spender, having spent 279,489.02.

## Analyze Customer Retention (Repeat Buyers vs One-Time Buyers)

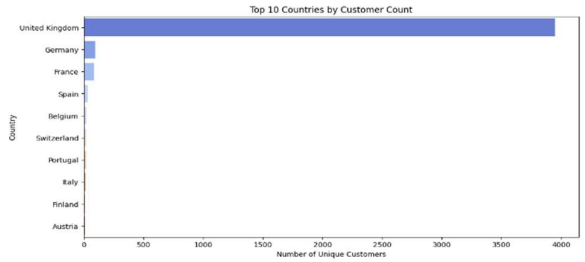
- Repeat Customers: 3059
- One-Time Customers: 1313

## Geographic Distribution of Customers

| | Country | UniqueCustomers |
|---|---|---|
| 35 | United Kingdom | 3950 |
| 14 | Germany | 95 |
| 13 | France | 87 |
| 30 | Spain | 31 |
| 3 | Belgium | 25 |
| 32 | Switzerland | 21 |
| 26 | Portugal | 19 |
| 18 | Italy | 15 |
| 12 | Finland | 12 |
| 1 | Austria | 11 |

The highest unique customers are from United Kingdom. There are 3,950 unique customers from the United Kingdom.

### Visualizing Customer Distribution



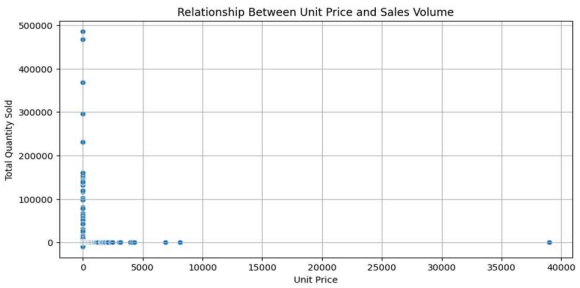

## Customer Segmentation Based on Spending Behavior

| | SpendingCategory | CustomerID |
|---|---|---|
| 0 | Low Spenders | 4257 |
| 1 | Moderate Spenders | 1173 |
| 2 | Good Spenders | 108 |
| 3 | High Spenders | 54 |
| 4 | VIP Customers | 6 |

There are 4,257 customers who fall into the "Low Spenders" category and there are 54 customers who fall into the "High Spenders" category.

## Optimizing Pricing Strategies:

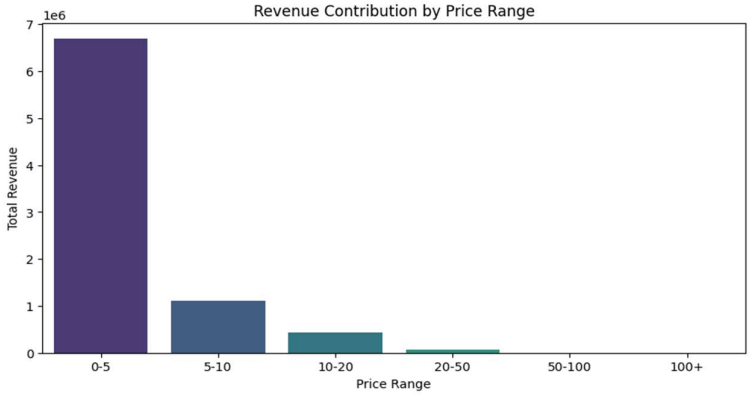
### Understand the Relationship Between Price and Sales Volume




Lower-priced products tend to have higher sales volumes, as they attract a larger customer base due to affordability.

### Identify the Price Range with Highest Sales

| | PriceRange | Quantity |
|---|---|---|
| 0 | 0-5 | 4704441 |
| 1 | 5-10 | 152532 |
| 2 | 10-20 | 33932 |
| 3 | 20-50 | 1860 |
| 4 | 50-100 | 53 |
| 5 | 100+ | 186 |




Most of the quantity comes under the price range of 5-10.

## Correlation Between Price and Revenue

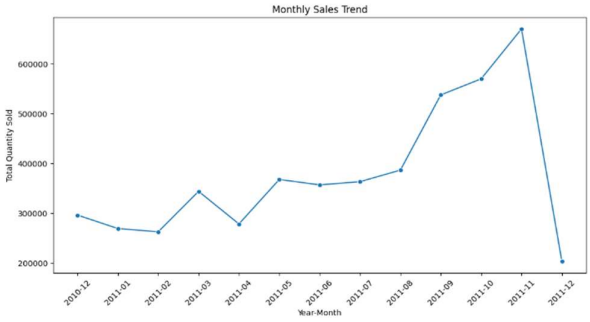
| | UnitPrice | Revenue |
|---|---|---|
| UnitPrice | 1.000000 | -0.129296 |
| Revenue | -0.129296 | 1.000000 |

## Streamlining Inventory Management:

### Identify Fast-Moving and Slow-Moving Products

The analysis highlights fast-moving products like World War 2 Gliders and Jumbo Bag Red Retrospot, which enjoy high demand and frequent sales. In contrast, slow-moving items, such as Rotating Silver Angels T-Light Holder, show low or negative sales, indicating weak demand. These insights help optimize inventory and sales strategies**.**

### Forecast Future Demand




The sales trend indicates strong seasonal demand, with peaks in the last quarter of the year, particularly in November. Future forecasting should account for these seasonal spikes, ensuring sufficient inventory and marketing efforts during high-demand periods. Additionally, strategies to stabilize sales in slower months, such as targeted promotions or new product launches, can help maintain consistent revenue. A predictive model using historical data can further optimize inventory planning and sales strategies.

## Detect Overstock Risks

The analysis identifies overstocked items, such as the Flower Shop Design Mug and Ruby Glass Cluster Necklace, with a Stock-to-Sales Ratio of 2.0, indicating excess inventory compared to sales. Overstocking can lead to increased storage costs, potential product obsolescence, and cash flow constraints. To mitigate this risk, businesses should implement demand forecasting, promotional discounts, or bundling strategies to clear excess stock efficiently.

## Identify Stockout Risks

The analysis highlights stockout risks for high-demand items such as the White Hanging Heart T-Light Holder and Jumbo Bag Red Retrospot, which have low inventory levels despite strong sales. A low Stock-to-Sales Ratio indicates that these products may soon be unavailable, leading to missed revenue opportunities and dissatisfied customers. To prevent stockouts, businesses should enhance demand forecasting, optimize restocking schedules, and consider safety stock strategies to ensure product availability during peak demand periods.

# Modelling

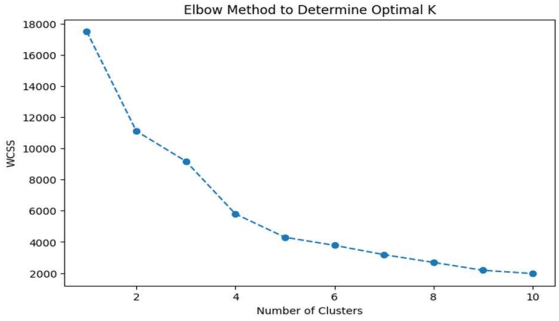
# Feature Engineering:

Feature engineering enhances the dataset by deriving valuable insights for customer analysis. The Revenue column calculates total spending, while grouping by CustomerID helps track key metrics like total revenue, purchase frequency, total quantity purchased, and recency (days since last purchase). These features are essential for customer segmentation, lifetime value prediction, and personalized marketing strategies, enabling businesses to optimize sales and customer retention efforts.

## Data Normalization

Data normalization ensures that all features contribute equally to clustering by standardizing them to a common scale. Using StandardScaler, the Revenue, Purchase Frequency, Quantity, and Recency columns are transformed to have a mean of 0 and a standard deviation of 1. This prevents dominant influence from larger values and improves the performance of clustering algorithms, leading to more accurate customer segmentation.

## Apply K-Means Clustering

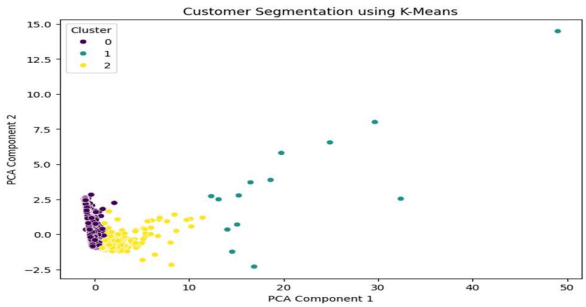



- The WCSS decreases as the number of clusters increases, which is expected because more clusters mean smaller distances between data points and their respective centroids.
- The "elbow point" is the value of K where the decrease in WCSS slows down significantly.
- In this graph, the elbow appears around K = 3 or K = 4, suggesting that these values are optimal for clustering.

## Apply K-Means with Optimal K

- Cluster 0: 3,969 customers (or data points) were assigned to this cluster.
- Cluster 2: 389 customers were grouped into this cluster.
- Cluster 1: Only 14 customers were assigned to this cluster.

## Cluster Interpretation

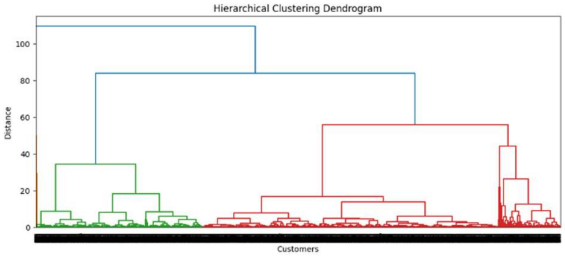



- Cluster 0 (Purple): Most data points are tightly packed on the left side. Likely represents regular customers with lower spending or low transaction frequency.
- Cluster 2 (Yellow): Spread out but still close to Cluster 0. Could represent moderate spenders or customers with higher transaction frequency.
- Cluster 1 (Green): Sparsely distributed across the right side. A few points are far apart (outliers), suggesting high-value customers who spend significantly more or have unusual purchasing behavior.

| Cluster | Revenue | Purchase_Frequency | Quantity | Recency |
|---|---|---|---|---|
| 0 | 921.520394 | 3.226757 | 561.580499 | 98.940791 |
| 1 | 110488.785000 | 96.071429 | 63508.928571 | 3.785714 |
| 2 | 7958.119229 | 20.663239 | 4598.586118 | 13.647815 |

- For Cluster 0 (Low-Spending Customers): Implement discount strategies to encourage more frequent shopping. Use email marketing to re-engage them (since their recency is high).
- For Cluster 1 (VIP Customers): Offer exclusive loyalty programs or personalized promotions. Provide priority customer support to increase retention.
- For Cluster 2 (Moderate Buyers): Encourage them to shop more frequently with limited-time deals. Use targeted ads and push notifications to increase conversion.

# Hierarchical Clustering




- Clusters are merged from bottom to top: Initially, each data point is its own cluster. As we move up, clusters are combined based on similarity.
- The height of the vertical lines (distance): Short lines → More similar data points, merged early. Tall lines → Distant clusters, merged later. The biggest vertical gap (elbow effect) suggests the ideal number of clusters.

# Conclusions:

**Successfully segmented customers into meaningful groups:** Customers were grouped based on their purchasing patterns using clustering techniques (e.g., K-Means and Hierarchical Clustering). This segmentation allows the business to tailor its marketing and sales strategies to different customer groups.

**Identified high-value customers to increase retention:** High-spending customers (VIPs) were recognized as an important segment.Strategies like exclusive promotions, loyalty programs, and priority customer support can be used to retain these customers and encourage repeat purchases.

**Found lost customers to reactivate them**: Some customers showed infrequent purchases or had stopped engaging with the platform.By identifying them, businesses can use email campaigns, discounts, or personalized offers to re-engage and bring them back.

**Created insights to optimize marketing and inventory strategies:** Understanding customer behavior helps in targeted marketing, ensuring that promotions reach the right audience.It also aids in inventory management, ensuring that popular products are well-stocked while reducing surplus inventory for low-demand items.

Overall, these insights help the company improve customer satisfaction, boost revenue, and streamline operations.

# How It’s Useful to Stakeholders:

**Marketing Team:** Can run targeted campaigns based on customer segments.With customer segmentation, the marketing team can design personalized campaigns for different groups.

For example, high-value customers can receive exclusive deals, while inactive customers can be re-engaged with special discounts or reminders.

**Inventory Management:** Can stock high-demand products based on top customer clusters.By analyzing customer purchasing behavior, businesses can predict which products will be in high demand.

This helps optimize stock levels, reducing overstocking of low-demand items and ensuring availability of popular products.

**Finance Team:** Can analyze high-value customer behaviors for revenue predictions.Understanding spending patterns of top customers helps in forecasting revenue and financial planning.

The finance team can use these insights to set realistic revenue targets and optimize budget allocation.

**Customer Retention Team:** Can reduce churn by identifying and re-engaging lost customers.The segmentation helps identify customers who haven't made recent purchases or whose activity has dropped.

The retention team can implement strategies like personalized offers, loyalty rewards, or direct outreach to bring these customers back.

**Overall Benefit:** Each team uses customer segmentation insights to make data-driven decisions, leading to higher sales, better resource allocation, and improved customer satisfaction.