E-COMMERCE SALES PERFORMANCE AND REVENUE ANALYSIS

Project Objective:

The primary objective of this analysis is to transform raw transactional and master data into meaningful business insights that support strategic decision-making. By leveraging customer demographics, product cost and pricing, discount structures, revenue, total sales amount, and profit or loss, the study aims to assess overall business performance and identify opportunities for growth and optimization. The following analysis covers the following core areas.

Customer Analysis

- Customer loyalty levels and their impact on revenue generation.
- Identification of high-value customers based on purchasing behavior and revenue contribution.

Product Performance Analysis

- Performance analysis by product category and Inventory status .
- Regional performance analysis to identify geographical sales trends and opportunities.

Sales and Revenue Analysis

- Trend analysis based on order date and quantity sold.
- Measurement of revenue, and profit or loss by product category
- Impact assessment of discounts on revenue and profitability.
- Evaluation of payment type preferences

Dataset Description:

The dataset represents an e-commerce sales environment and consists of four primary tables: Customer, Product, Store, and Sales. These tables collectively capture customer demographics, product details, store information, and transactional sales records. The structured design of the dataset enables comprehensive analysis of customer behavior, product performance, store operations, and overall sales and revenue trends.

Customer Table - Column Description

Column Name	Description
Customer_ID	Unique identifier for each customer

Name	Full name of the customer
Age	Age of the customer
Gender	Gender of the customer
City	Customer’s city
State	Customer’s state
Loyalty_Level	Loyalty program tier (e.g., Gold, Silver, Bronze)

Product Table – Column Description

Column Name	Description
Product_ID	Unique identifier for each product
Product_Name	Name of the product
Category	Main product category
Sub_Category	Detailed product category
Brand	Brand name of the product
Cost	Cost price of the product
Stock	Available quantity in inventory

Store Table – Column Description

Column Name	Description
Store_ID	Unique identifier for each store
Store_Name	Name of the store
Region	Geographic region of the store
City	City of the store
Store_Type	Type of store (e.g., flagship, outlet)

Sales Table – Column Description

Column Name	Description
Sales_ID	Unique identifier for each sales transaction
Order_Date	Date when the order was placed
Customer_ID	ID of the customer making the purchase
Product_ID	ID of the product sold
Store_ID	ID of the store where sale occurred
Region	Geographical location of the sale, used for region-wise sales and performance analysis
Quantity	Number of units sold in the transaction
Product Cost	Cost price of the product per unit, used to calculate profit or loss when compared with selling price.
Unit_Price	Price per unit of the product
Discount	Discount applied on the transaction
Revenue	Total income generated from product sales and used to measure sales performance
Total_Amount	Final amount after applying discount
Payment_Type	Payment method used (Cash, Card, Online)

Data cleaning & transformation:

Data Import Process

The dataset was imported into the analysis environment using Microsoft Excel's data import feature. The following steps were followed to load the data:

1. Navigate to the **Data** tab in Excel.
2. Select **Get Data** and choose **From File**.
3. Click on **Excel Workbook** and browse to the *E-Commerce Sales Dataset* file.
4. Select the required worksheets from the workbook.
5. Click **Load** to import the data into Excel for further processing and analysis.

1. Customer Data Cleaning and Transformation

- **Customer_ID Standardization:**

The Customer_ID column contained inconsistencies in formatting (for example, use of hyphens and varying prefixes). These inconsistencies were corrected using the **Find and Replace** method to standardize the Customer_ID format across the dataset.

Inconsistent Customer_ID	Standardized Customer_ID
C-1416	CUST1416
C-1486	CUST1486
C-1491	CUST1491
C-1499	CUST1499

This standardization ensures uniformity and improves data integrity for accurate customer-level analysis.

- **Duplicate Validation in Customer_ID:**

A duplicate check was performed on the Customer_ID column using the **Remove Duplicates** feature. As no duplicates were detected and **CLEAN** and **TRIM** functions were applied to remove hidden characters and extra spaces, ensuring consistency in the Customer_ID values.
Formula used =**CLEAN(TRIM(A2))**

- **Name Column Standardization:**

The Name column contained honorific titles and academic degree suffixes such as **Mr, Mrs, Ms, Dr, PhD, MD, and MBA**, which caused inconsistency in customer names. These were removed to maintain uniform formatting and improve data quality.
The cleaning process was performed using:

- **Data → Data Tools → Text to Columns**
- Delimiters: **Space and Dot**
- Followed by re-combining name components

After standardization, the following formula was applied to format names correctly: **=CLEAN(TRIM(PROPER(CONCAT([@Name]," ",[@Column4]," ",[@Column3]))))**

This ensured proper casing, removed unnecessary spaces, and delivered consistent name formatting across the dataset.

- **Age Group Classification:**

A new column, Age Group, was created to categorize customers into defined age segments (Young, Adult, Senior). This classification supports demographic analysis and enables better understanding of customer purchasing behaviour across different life stages.

Formula used

=IFS([@Age]>=50,"Senior",[@Age]>=30,"Adult",[@Age]>=18,"Young")

- **Loyalty_Level Data Imputation**

Missing values in the Loyalty_Level column were handled through imputation based on customer purchase activity. Loyalty levels were assigned according to the number of sales transactions contributed by each customer.

Formula=IF(ISBLANK([@[Loyalty_Level]]),IF(COUNTIF(Sales_Fact[Customer_ID],[@[Customer_ID]])>5,"Platinum",IF(COUNTIF(Sales_Fact[Customer_ID],[@[Customer_ID]])>=3,"Gold",IF(COUNTIF(Sales_Fact[Customer_ID],[@[Customer_ID]])>=2,"Silver",""))),[@[Loyalty_Level]])

The Customer City attribute was removed from the dataset, as all records belong to a single country and the field did not provide additional analytical value.

Additionally, all columns were formatted in accordance with their respective data types, including text, numeric to ensure consistency and enable accurate analytical reporting.

2. Product Table Data Cleaning and Transformation

- **Product_ID Standardization**

The Product_ID column contained inconsistencies in formatting (for example, use of hyphens). These inconsistencies were corrected using the **Find and Replace** method to standardize the Product_ID format across the dataset.

Inconsistent Product_ID	Standardized Product_ID
P-3	PROD3
P-16	PROD16

- **Duplicate Validation in Product_ID**

  A duplicate check was performed on the Product_ID column using the **Remove Duplicates** feature. As no duplicates were detected and **CLEAN** and **TRIM** functions were applied to remove hidden characters and extra spaces, ensuring consistency in the Product_ID values.

Formula Used**=CLEAN(TRIM([@[Product_ID]]))**

- **Product Name Standardization**

  The Product_Name data was reviewed and found to be largely consistent. As part of data standardization, CLEAN and TRIM functions were applied to eliminate hidden characters and extra spaces, ensuring uniformity across Product_Name values. The cleaned results were then converted to values and used for subsequent processing.

Formula Used=**CLEAN(TRIM([@[Product_Name]]))**

- **Product Category Standardization**

  The Category data was reviewed and found to be largely consistent. As part of data standardization, CLEAN and TRIM functions were applied to eliminate hidden characters and extra spaces, ensuring uniformity across Category values. The cleaned results were then converted to values and used for subsequent processing.

Formula Used **=CLEAN(TRIM([@Category]))**

- **Product Sub_Category Standardization**

  The Sub_Category data was reviewed and found to be largely consistent. As part of data standardization, CLEAN and TRIM functions were applied to eliminate hidden characters and extra spaces, ensuring uniformity across Sub_Category values. The cleaned results were then converted to values and used for subsequent processing.

Formula Used =**CLEAN(TRIM([@[Sub_Category]]))**

- **Brand Name Standardization**

  Brand names were standardized to ensure consistent formatting across the dataset. Data cleansing activities included the removal of extra spaces and hidden characters, correction of capitalization inconsistencies, and standardization. Legal

entity suffixes (such as Inc, LLC, Ltd, and PLC) were retained and formatted consistently where available. No assumptions were made to append missing suffixes, and brand names without legal identifiers were preserved as provided to maintain data accuracy and integrity.
Formula Used =**CLEAN(TRIM(PROPER([@Brand])))**

- **Product Cost Standardization**

Missing Cost values were addressed through category-level imputation, whereby the average Cost of products within the same Sub_Category was applied. This method was used solely for blank records and did not override any existing values, ensuring data integrity and analytical consistency.
Formula
Used=**IF(ISBLANK([@Cost]),AVERAGEIF([Sub_Category],[@[Sub_Category]],[Cost]),[@Cost])**

- **Stock Standardization**

Missing Stock values were handled by replacing blank entries with the average product cost within the same sub-category. This imputation method was applied only to missing records and did not overwrite any existing values, thereby preserving data integrity and ensuring consistency for analytical purposes.
Formula used
=IF(ISBLANK([@Stock]),AVERAGEIF([Category],[@Category],$G$2:$G$101),[@Stock])

Additionally, all columns were formatted in accordance with their respective data types, including text, numeric, and currency formats, to ensure consistency and enable accurate analytical reporting.

3. Store Dimension Table Data Cleaning and Transformation

- **Store ID Standardization**

The Store_ID field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces, ensuring the Store_ID values are fully standardized and ready for downstream analysis.
Formula used=**CLEAN(TRIM([@[Store_ID]]))**

- **Store Name Standardization**

During the review of the Store_Name field, all values were found to be valid and consistent, except for minor capitalization inconsistencies, such as “Mcdonald Inc,” which should be standardized to “McDonald Inc” using **Find & Replace** to ensure uniform Title Case formatting. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces.
Formula Used=**CLEAN(TRIM(PROPER([@[Store_Name]])))**

- **Region Standardization**

The Region field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces, ensuring the Region values are fully standardized and ready for downstream analysis.
Formula Used=**CLEAN(TRIM([@Region]))**

- **City Standardization**

The City field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces, ensuring the City values are fully standardized and ready for downstream analysis.
Formula used=**CLEAN(TRIM([@City]))**

- **Store Type Standardization**

The Store_Type field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces, ensuring the Store_Type values are fully standardized and ready for downstream analysis.
Formula Used=**CLEAN(TRIM([@[Store_Type]]))**

Additionally, all columns were formatted in accordance with their respective data types, including text, numeric, and currency formats, to ensure consistency and enable accurate analytical reporting.

4. Sales Fact Table Data Cleaning and Transformation

- **Sales Id Standardization**

The Sales_ID field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces, ensuring the Sales_ID values are fully standardized and ready for downstream analysis.
Formula used**=CLEAN(TRIM([@[Sales_ID]]))**

- **Order date Standardization**

The Order_Date field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, the CLEAN and TRIM functions were applied to remove any hidden characters and extra spaces, ensuring that the date values were fully standardized and suitable for analysis. Additionally, the date format was converted from Year-Month-Day (YYYY-MM-DD) to Day-Month-Year (DD-MM-YYYY) to maintain consistency with reporting standards and improve readability for analytical and visualization purposes.
Formula used**=CLEAN(TRIM([@[Order_Date]]))**

- **Quantity Standardization**

During the data cleaning process, it was found that some records in the Quantity field were missing. To maintain consistency and ensure accurate analysis, these missing values were filled with the **average quantity** calculated from the available data. This approach ensures that all records have valid quantity values, allowing reliable calculation of revenue, profit, and other key metrics.
Formula used
=IF(ISBLANK([@Quantity]),AVERAGE([Quantity]),[@Quantity])

- **Product_cost Standardization**

To calculate **Revenue** and **Profit**, the Product_cost was included in the sales dataset by using the **VLOOKUP** function. This allowed us to match each Product_ID in the sales data with its corresponding Product_cost from the product table. By having the Product_cost available, we were able to compute **Revenue** as the total amount received from sales and **Profit** as the difference between Revenue and Product_cost, ensuring accurate financial analysis.
Formula Used**=VLOOKUP([@[Product_ID]],Product_Dim,6,FALSE)**

- **Unit Price Standardization**

During the data preparation process, it was identified that the Unit_Price field contained missing values. To ensure data completeness and support accurate downstream calculations, these missing unit prices were imputed using the **average unit price** derived from the available data. This approach ensures consistency in pricing data and enables reliable analysis in subsequent processes such as revenue and profit calculation
Formula Used=**IF(ISBLANK([@[Unit_Price]]),AVERAGE($H$2:$H$20),[@[Unit_Price]])**

- **Discount Standardization**

During data cleaning, missing values in the Discount field were identified and replaced with the **average discount** calculated from the existing records. This ensured consistency in discount values and supported accurate calculation of total amount, revenue, and profit in subsequent analysis.
Formula used=**IF(OR(ISBLANK([@Discount]),[@Discount]=0),AVERAGE($I$2:$I$27),[@Discount])**

- **Revenue Standardization**

A new column was added to the dataset to calculate **Revenue**, which was derived by multiplying the Quantity by the Unit_Price. This ensures a clear and consistent representation of sales value for each transaction and supports further analysis such as profit calculation and performance evaluation.
Total Revenue: =[@Quantity]*[@[Unit_Price]]

- **Total_Amount Standardization**

Missing values in the Total_Amount column were identified during data validation. These values were calculated by first applying the discount to the revenue, computed as Revenue × Discount, and then subtracting the discount amount from the revenue to derive the final total amount. This ensured consistency and accuracy in the total sales values used for further analysis.
Formula Used=J2-([@Revenue]*[@Discount])

- **Profit Standardization**

A Profit column was created and utilized to evaluate whether each transaction resulted in a profit or a loss. This measure enables clear identification of profitable and non-profitable sales and supports overall profitability analysis across products, customers, and regions.
Formula used=[@[Total Amount]]-([@[Product_cost]]*[@Quantity])

- **Sore region Standardization**

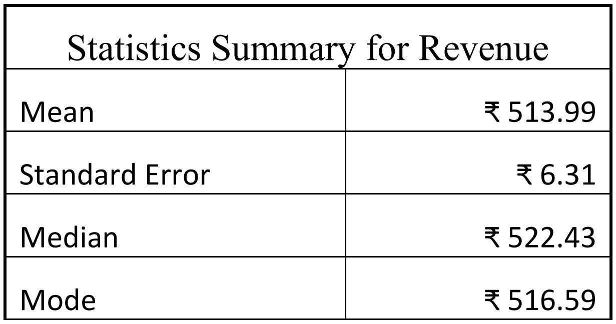
Store and region attributes were included in the analysis to evaluate sales performance at both store and regional levels, enabling store-wise and region-wise comparison of sales trends and revenue contribution. These attributes were retrieved from the Store table using the VLOOKUP function and integrated into the Sales table for consolidated analysis.
Formula Used=VLOOKUP([@[Store_ID]],Store_Dim[[Store_ID]:[Region]],3)

- **Payment Type Standardization**

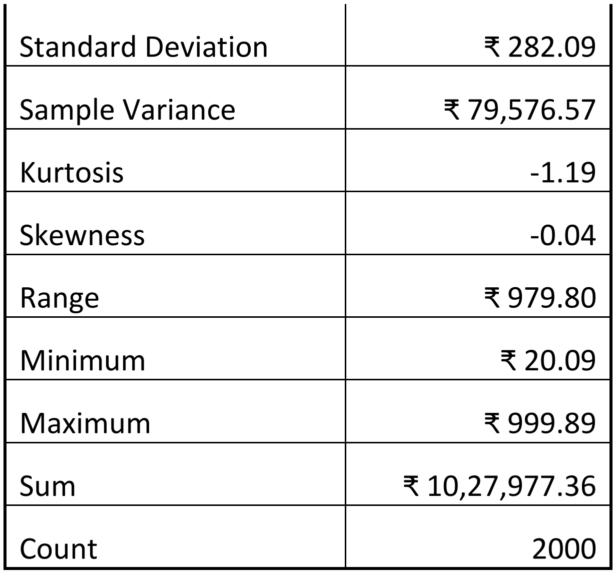
The Payment Type field was reviewed and found to be free of null values and inconsistencies. As a standard data quality measure, CLEAN and TRIM functions were applied to remove any hidden characters or extra spaces, ensuring the Sales_ID values are fully standardized and ready for downstream analysis.
Formula used=**CLEAN(TRIM([@[Payment_Type]]))**

Total sold qty calculated by using =SUM(G2:G2001)
Total Revenue Calculated by using =SUM(K2:K2001)
Profit Calculated by Using =SUMIF(M2:M2001, ">0")
Loss calculated by using=SUMIF(M2:M2001, "<0")

Statistics Description:



Statistics Summary for Revenue	
Mean	₹ 513.99
Standard Error	₹ 6.31
Median	₹ 522.43
Mode	₹ 516.59



Standard Deviation	₹ 282.09
Sample Variance	₹ 79,576.57
Kurtosis	-1.19
Skewness	-0.04
Range	₹ 979.80
Minimum	₹ 20.09
Maximum	₹ 999.89
Sum	₹ 10,27,977.36
Count	2000

The revenue Statistic consists of 2,000 observations with a total revenue of ₹10,27,977.36. The mean revenue value is ₹513.99, indicating the average revenue generated per transaction.

The median revenue of ₹522.43 is close to the mean, suggesting that the data is symmetrically distributed around the central value.

The mode of ₹516.59 further supports the consistency of frequently occurring revenue values.

The standard deviation of ₹282.09 indicates a moderate level of variability in revenue, showing that individual transaction values deviate considerably from the average.

The sample variance of ₹79,576.57 reinforces the presence of dispersion within the dataset.

The standard error of ₹6.31 reflects a reliable estimate of the population mean, given the large sample size.

Skewness is recorded at -0.04, which indicates an approximately symmetrical distribution with no significant bias toward higher or lower revenue values.

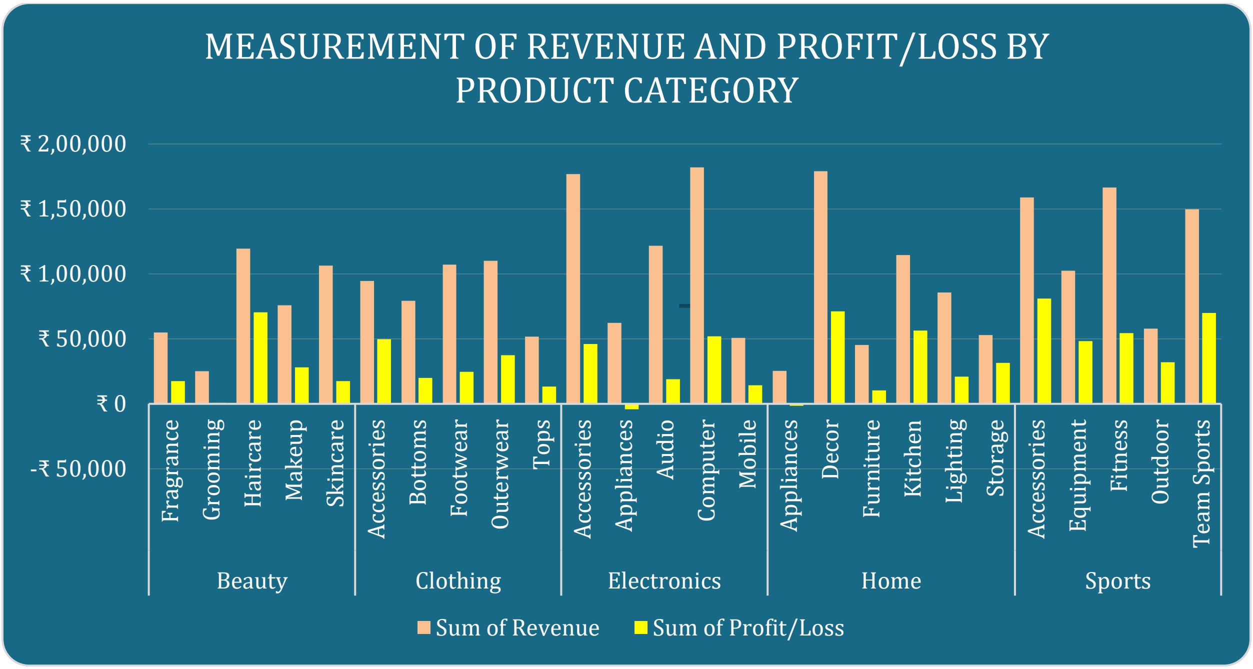
The kurtosis value of -1.19 suggests a flatter distribution compared to a normal distribution, implying fewer extreme outliers and a more evenly spread set of revenue values.

The revenue range of ₹979.80, with a minimum of ₹20.09 and a maximum of ₹999.89, shows that transaction values span a wide interval. This demonstrates that the dataset includes both low-value and high-value transactions, contributing to overall revenue diversity.

The revenue distribution exhibits balanced central tendency, moderate dispersion, and minimal skewness, indicating stable and consistent revenue behavior across transactions. These statistical measures provide a reliable foundation for further trend analysis, forecasting, and performance evaluation.

Visualization and Insights

The Dataset provides an overview of sales performance, customer characteristics, and product performance. Key metrics such as total sales, total revenue, total profit, and number of orders are calculated. Sales trends are analyzed over time to identify patterns. Product-wise and category-wise summaries are used to identify best-selling products. The Appliances category shows low profit or loss even though stock levels are high. This means that a large quantity of appliances is stored but not sold effectively. Store-wise and region-wise performance is also summarized. This indicates poor performance of this category and possible overstocking issues.



Pricing & Profitability

- Many product categories are facing losses because the Unit Price is lower than the Product Cost.
- To earn profit, the Unit Price must be higher than the Product Cost, even after applying discounts.
- The Appliance category has the highest losses.
- All stores should ensure products are sold above manufacturing or purchase cost to maintain profit margins.

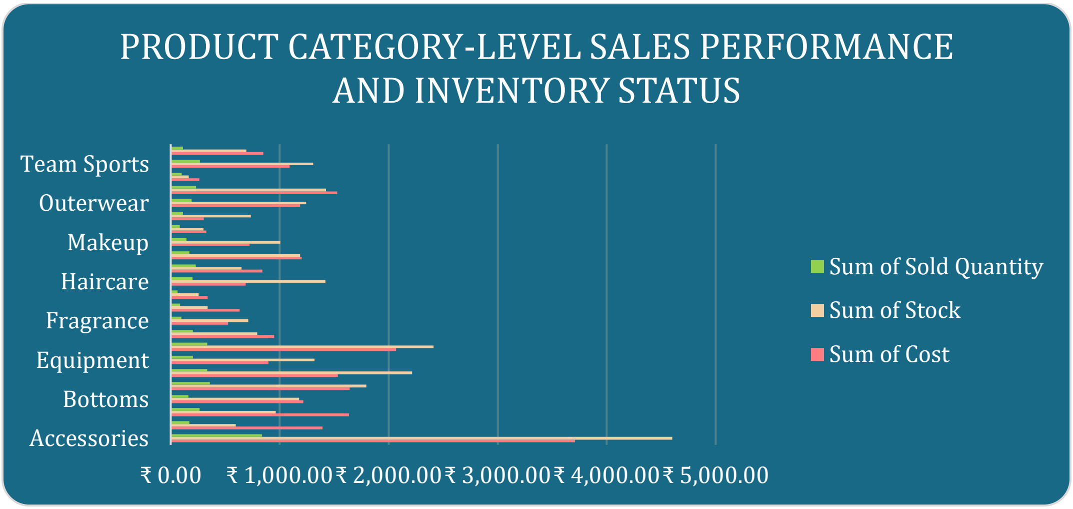
Stock & Inventory Analysis

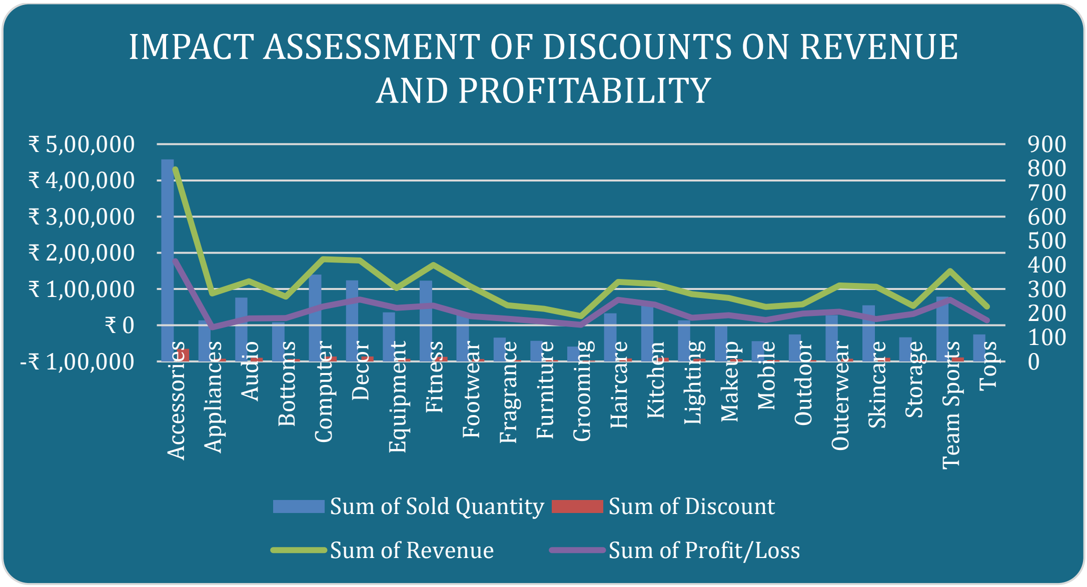
- All products maintain a minimum stock level, which is good for operations.
- The Accessories category has very high stock but low sales, showing:

    - Overstocking
    - Low demand
- Sales strategies are needed to:
    - Increase Accessories sales
    - Reduce excess inventory

Sales & Regional Performance

- Computers are the highest-selling products.
- North region has the highest sales volume.
- East region has the lowest sales volume.
- More focus is needed to:
    - Improve sales in the East region
    - Maintain performance in other regions





Category Performance (Future)

- Computers and Electronics will continue to give high revenue due to strong demand.

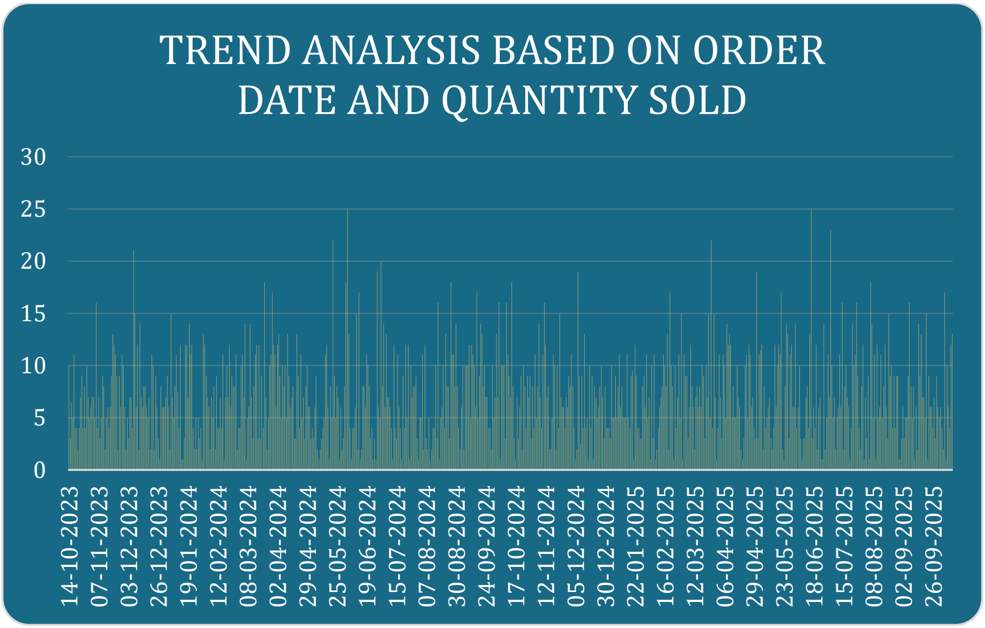
- Appliance category may continue to show losses if prices are not corrected.
- Appliance prices should be higher than product cost to avoid losses.

Inventory (Stock)

- Accessories have high stock but low sales.
- If this continues:
  - Storage cost will increase
  - Profit will reduce
- Promotions and better stock planning can:
  - Increase sales
  - Reduce overstock

Regional Sales

- North region will likely remain the top sales region.
- East region will likely stay the lowest performing region unless:
  - Better marketing is done
  - Stores are improved
- South and West regions are expected to show stable growth



Pricing & Discounts

- Prices and discounts should match:
  - Product cost
  - Customer demand
- Revenue changes each month, so:
  - Pricing should be more consistent
  - Planning should be based on demand

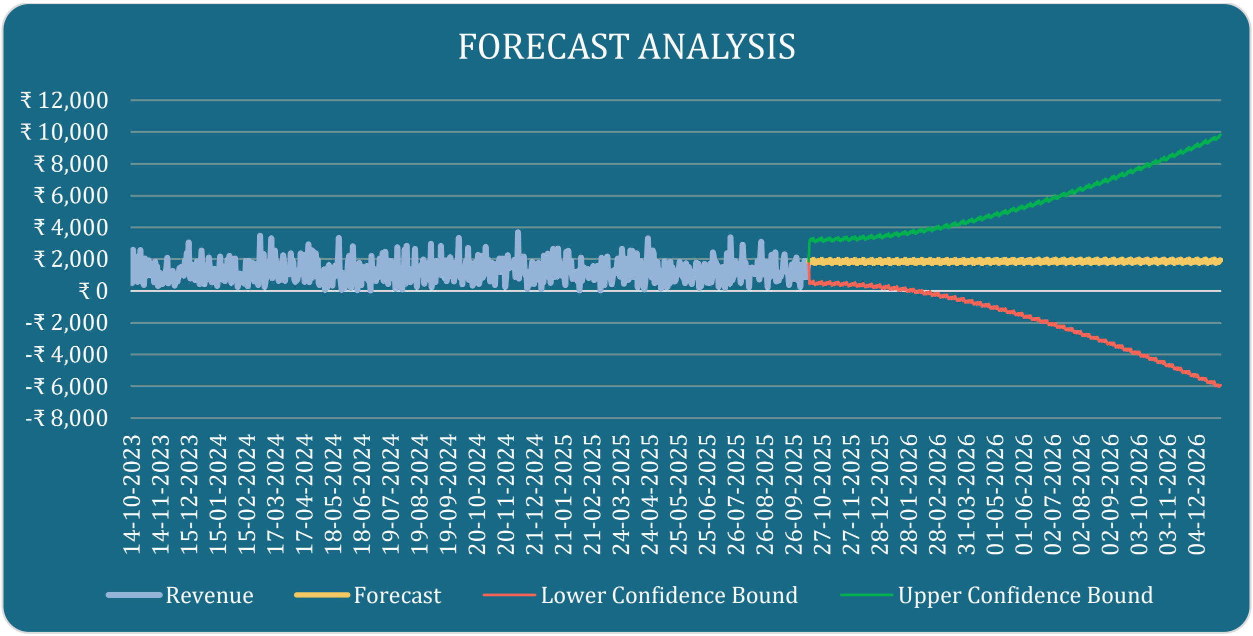
Promotion Strategy

- Focus promotions on categories that:
  - Give high revenue
  - Perform well with small discounts
- Review or reduce discounts for categories that:
  - Lose profit even with high discounts

Discount Optimization

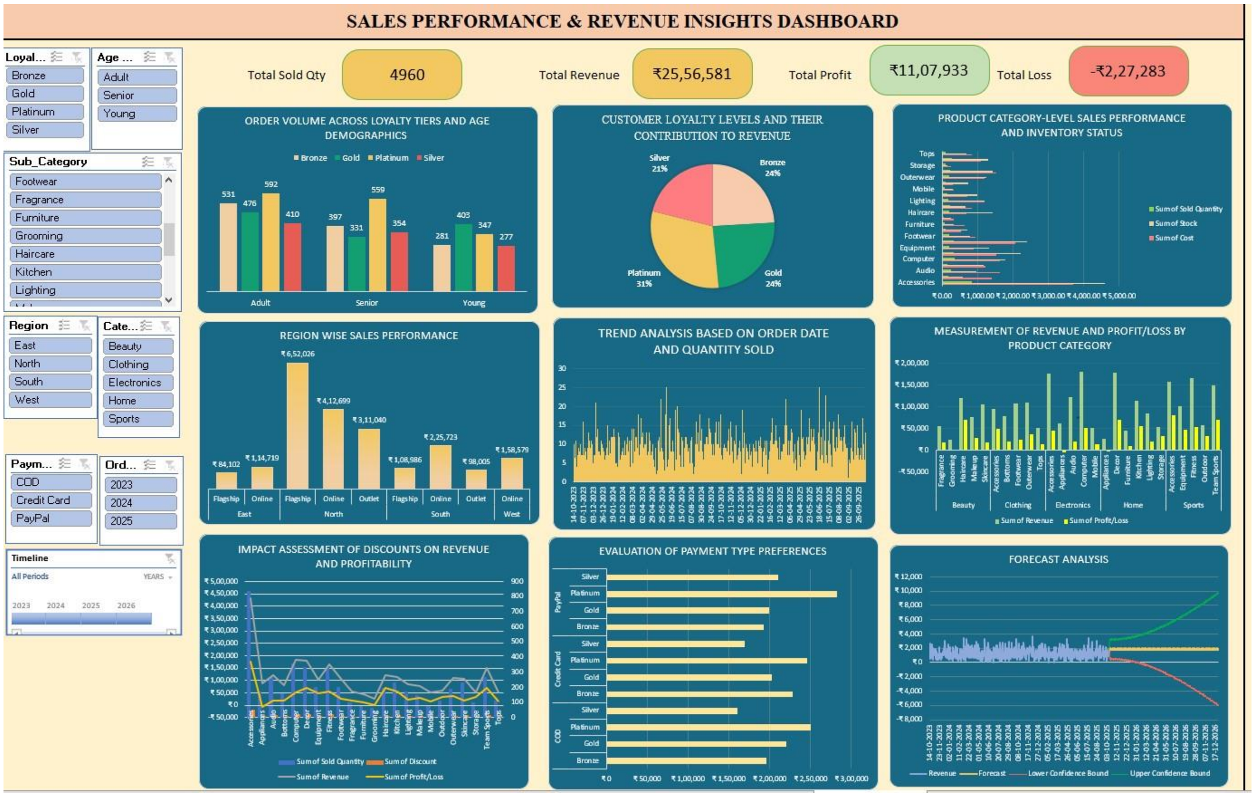
- Do not give the same discount to all products.
- Use different discount strategies for different products:
  - High-demand products → low discount
  - Low-performing products → short-term offers
- Avoid heavy discounts on low-profit products.

Revenue Forecast



Revenue Forecast Summary

- A time-series model was used to predict future revenue.
- Past revenue shows a mostly stable pattern with small fluctuations.
- Revenue is expected to increase gradually over time.
- Upper bound shows possible higher growth if market conditions are good.
- Lower bound shows possible decline if business conditions are poor.
- Uncertainty increases for long-term predictions.
- The forecast helps in planning, budgeting, and setting sales targets.



Conclusion:

- Sales are strong, but both profit and loss exist, showing the need for better pricing and cost control.
- Some product categories generate high revenue, while others make loss due to low selling price compared to cost.
- High-loyalty customers contribute a large share of revenue, proving the importance of customer retention.
- Different age groups need targeted marketing strategies.
- Certain products have high stock but low sales, indicating poor inventory management.
- North region has the highest sales, while the East region has the lowest and needs improvement.
- Revenue shows fluctuations over time, so demand forecasting and stable pricing are required.
- Excessive discounts reduce profit, especially for low-margin products.
- Customers prefer different payment methods based on loyalty levels.
- Business performance can improve by optimizing pricing, inventory, regional strategies, and promotions.