Team 2025110

Table of Contents

Introduction

ElectroMart is an Ontario-based **e-commerce company** specializing in electronic products, including categories such as camera accessories, home audio, and gaming accessories. Over the past year, ElectroMarta llocated significant resources to various marketing initiatives such as commercials, online campaigns, and major promotional events similar to “Flipkart's Big Billion Day”. Despite the substantial investment, the Chief Financial Officer (CFO) has expressed concerns regarding the effectiveness and impact of these marketing efforts. The CFO suggests the necessity of reviewing the current marketing strategy, either by reducing the budget or by optimizing the allocation across various marketing channels to enhance revenue generation.

As the marketing team prepares to establish the marketing budget for the upcoming year, there is a critical need to analyze historical marketing performance comprehensively. Leveraging data analytics, ElectroMart aims to understand the effectiveness of past marketing spend, identify key performance drivers, quantify the impact of different marketing channels on revenue, and ultimately optimize the budget to maximize future returns.

Objectives

- **Performance Driver Analysis:**
  - Identify key performance indicators (KPIs) impacting revenue.
  - Analyze relationships among variables including ad spend, discounts, NPS scores, payment methods, holidays, and payday sales effects.
- **Marketing ROI Impact Analysis:**
  - Conduct a quantitative assessment of the impact each marketing channel (commercials, online campaigns, promotions) has on revenue.
  - Assess the ROI of promotional events versus routine marketing activities.
- **Optimization of Marketing Spending:**
  - Recommend optimal budget allocation across marketing channels.
  - Prioritize product subcategories (camera accessories, home audio, gaming accessories) for targeted marketing.
  - Identify and rationalize the most effective marketing channels to utilize for each targeted product subcategory.
- **Current Spending Strategy:**

The company’s current advertising strategy places a significant emphasis on Sponsorship, with ₹365.4 crore allocated, followed by Online Marketing at ₹193.6 crore. Search Engine Marketing (SEM) receives ₹91.2 crore, while Affiliates account for ₹61.4 crore. Traditional media remains part of the mix, with Television at ₹44.4 crore and Radio at ₹4.7 crore, along with Digital channels at ₹29.7 crore and Content Marketing at ₹8 crore. Additionally, the “Other” category amounts to ₹48 crore. Overall, this distribution underscores a robust investment in both brand-building and performance channels, ensuring a balanced yet strategically weighted approach across multiple advertising platforms.

Data Description

The files used for analysis are:

- **ONTARIO-2023 and ONTARIO-2024:** The Datasets that Captures climate conditions in Ontario, Canada which may influence purchasing behavior of Customers in the given Time Frame.
- **Net Promoter Score (NPS) and Stock Index Data:** The Dataset that Provides insights into customer satisfaction and company performance trends month wise for the given Time Frame.
- **Media Investment Data:** The Dataset that Includes monthly spending on various advertising channels for the given Time Frame.
- **Canada Holiday List:** List of National Holidays in Canada that may impact consumer purchasing patterns.
- **Product SKU Mapping Data:** It Provides product hierarchy details, including category and subcategory classifications.
- **Customer Orders Data:** It Contains transaction-level details, including order dates, product details, revenue (GMV), payment types, and customer IDs.
- **Media Investment Data:** It Includes monthly spending on various advertising channels.
- **Sales Calendar:** It Lists dates of promotional offers and special sales events in the given Time Frame.

Data Cleaning and Preprocessing

CUSTOMER DATA PREPROCESSING

- **Handling Erroneous Entries:**
  - 4,189 MRP values recorded as 0 (invalid price point) were removed from Customers_Orders_Data.csv to maintain accuracy.
  - 5,890 rows with missing GMV values were removed, as GMV is essential for sales analysis.
  - Rows with undecipherable pincode inputs were discarded to ensure data consistency.
- **Computing Pricing Metrics:**
  - List price was computed, and discount percentage was derived.
  - Discount values below 0 were excluded to maintain data integrity.
- **Handling Missing Data:**
  - Columns deliverybdays and deliverycdays had 13 lakh missing values out of 16 lakh entries and were removed to prevent inconsistencies.
- **Product Categorization:**
  - Products were categorized as luxury (top 20th percentile) and mass market (remaining 80%).
- **Sales Influencing Factors Considered:**
  - Payday effects, holidays, and special sale days were accounted for in the analysis to assess their impact on purchasing behavior and sales performance.

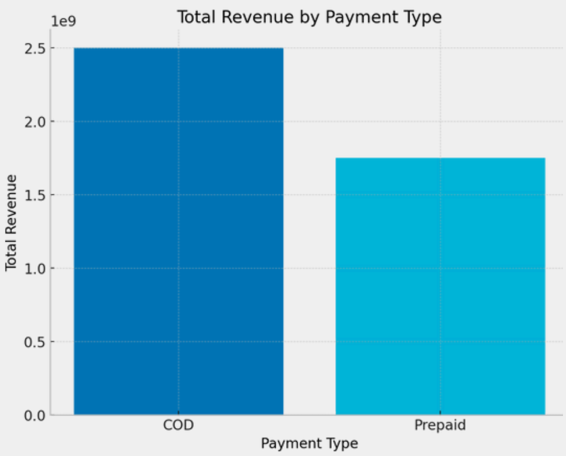
Data Cleaning and Preprocessing

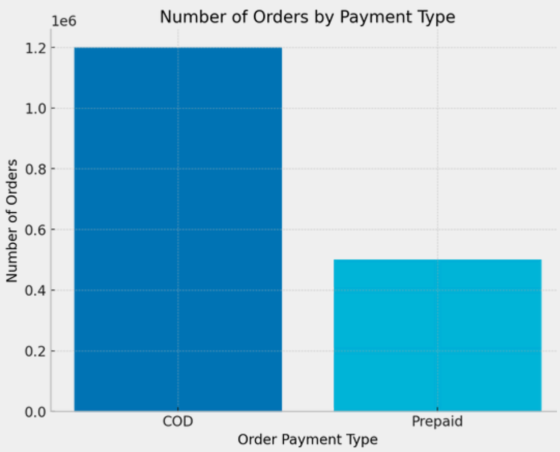
WEATHER DATA PREPROCESSING

- **Column Removal:**
  - Removed columns containing name ‘flags’('Max Temp Flag', 'Min Temp Flag', 'Mean Temp Flag', 'Heat Deg Days Flag', 'Cool Deg Days Flag', 'Total Rain Flag', 'Total Snow Flag', 'Total Precip Flag', 'Snow on Grnd Flag', 'Dir of Max Gust Flag', 'Spd of Max Gust Flag') at the end due to high percentage of empty value.
  - Dropped the last 2 columns(,i.e, ‘Dir of Max Gust (10s deg)’, ‘Spd of Max Gust (km/h)’) due to a high percentage of empty values.
- **Handling Null Values:**
  - After Column removals the Columns which left had null values.
  - Applied forward fill method to fill missing values in those left columns, ensuring data continuity**.**

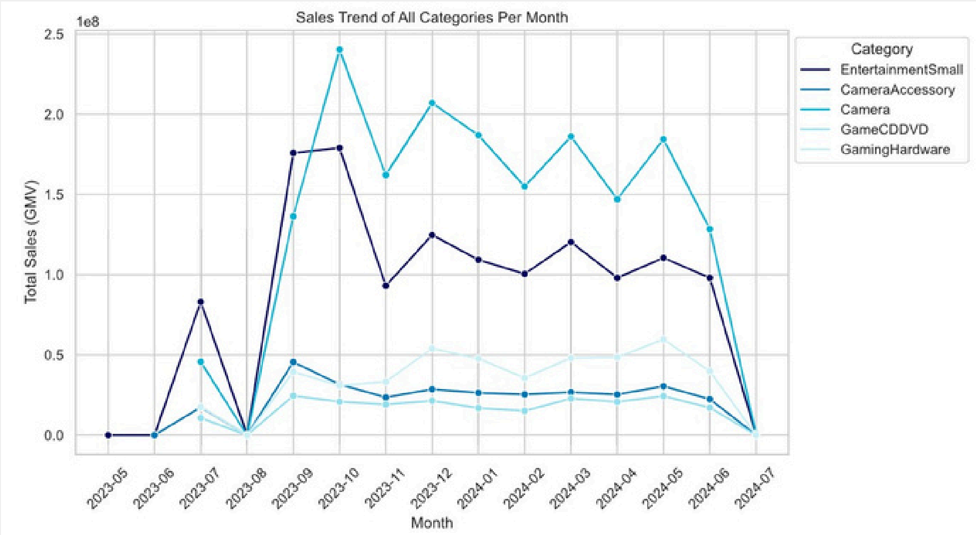
Exploratory Data Analysis

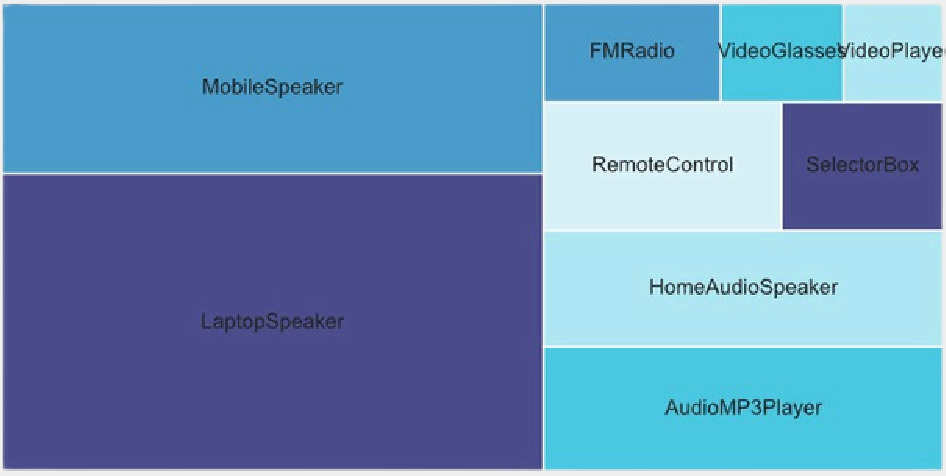
i) Relationship between revenue and Payment Types



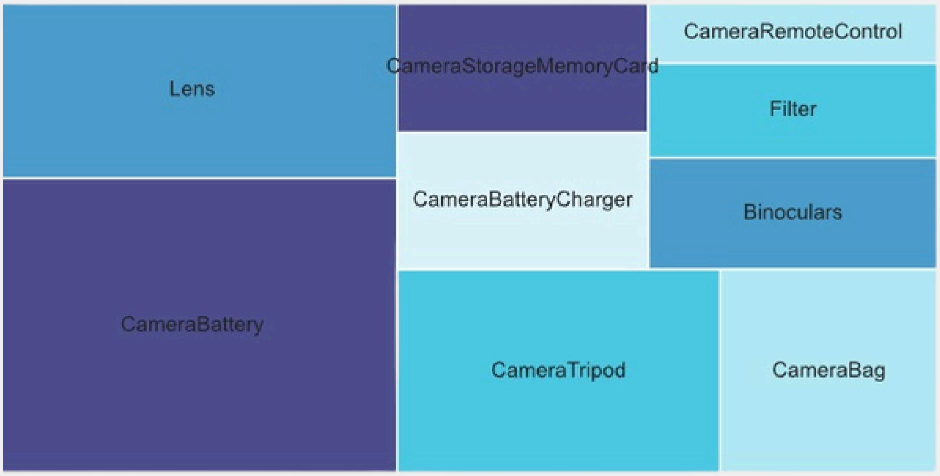


- COD contributes significantly more to total revenue and GMV than prepaid across most product sub-categories.
- The highest COD contribution is observed in the "Game" category
- COD is more common for products under ₹10,000.
- Prepaid share increases with price, overtaking COD as prices rise.
- Products above ₹1,50,000 are almost entirely prepaid, with minimal COD.





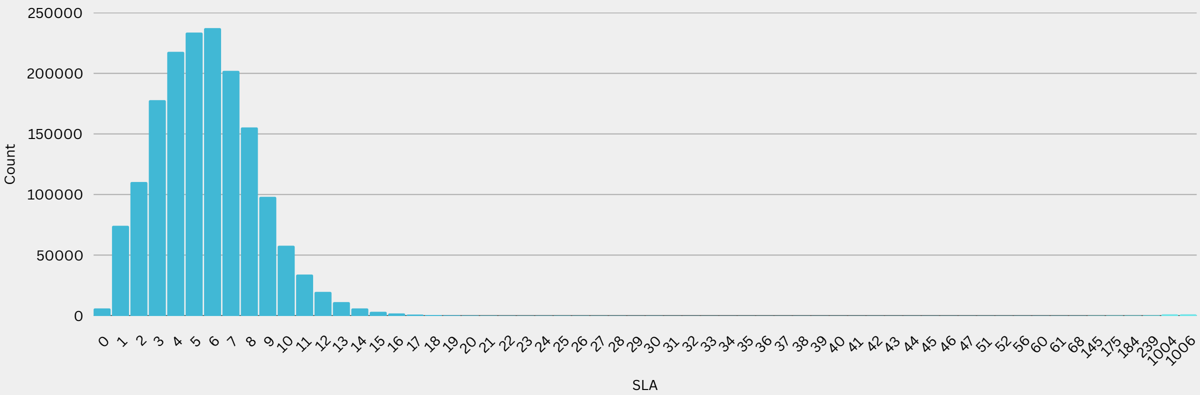
Units sold for EntertainmentSmall



Units sold for CameraAccesory

- The most selling category over the year is EntertainmentSmall and CameraAcessory
- Among EntertainmentSmall, Laptop speakers and Mobile speakers are in high demand.
- Among CameraSmall, Camera battery and lens are most selling product followed by camera tripod.

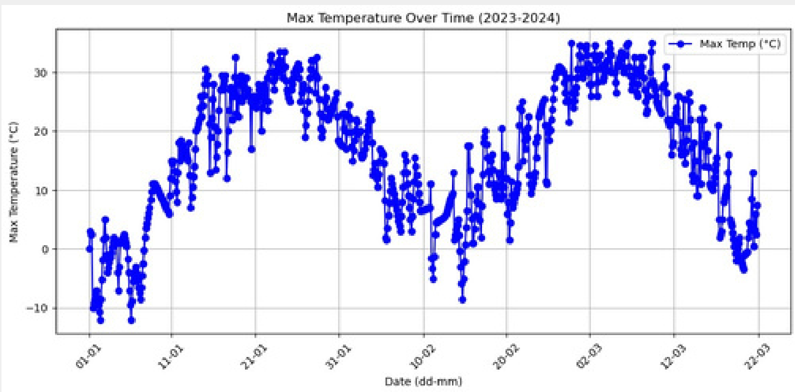
ii) SLA Frequency Distribution



The SLA distribution shows most orders fall between 3 to 8 days, peaking at 5 to 7 days. Beyond 10 days, occurrences drop sharply, with only a few orders having very high SLA values.

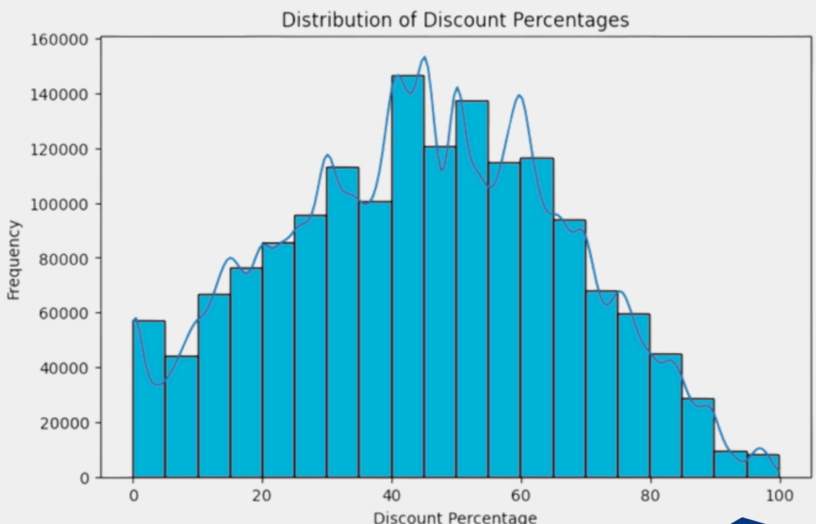
The data distribution shows most SLA values range between 3 to 10 days, with a sharp drop beyond this. While some outliers exist, a reasonable maximum SLA can be set at 15 to 20 days, as values beyond this are rare.

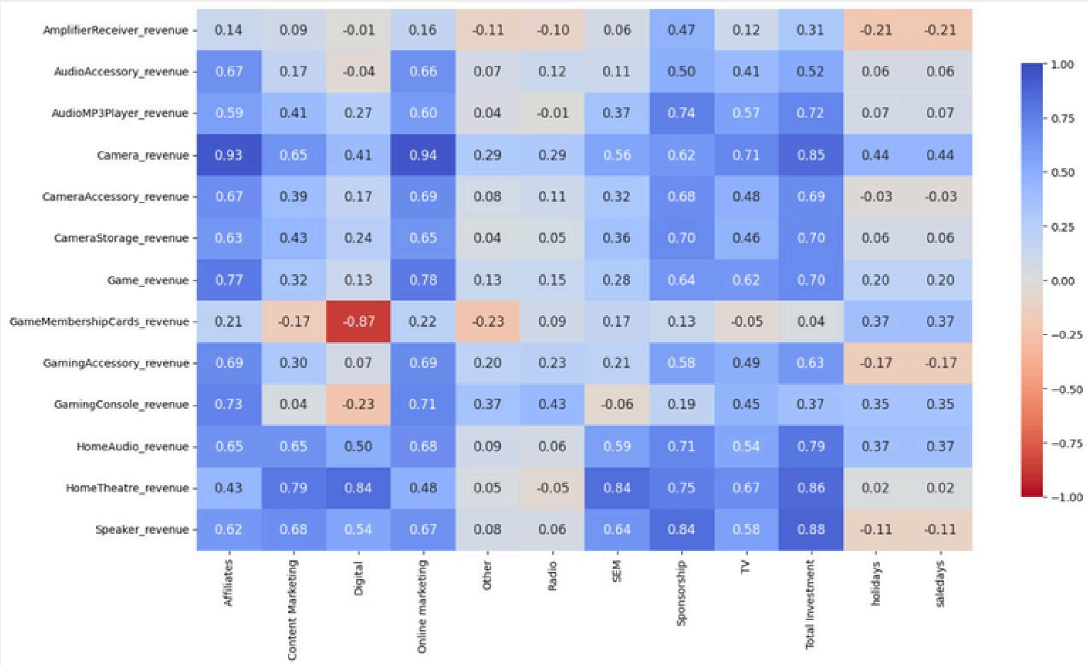
iii) Temperature Variation



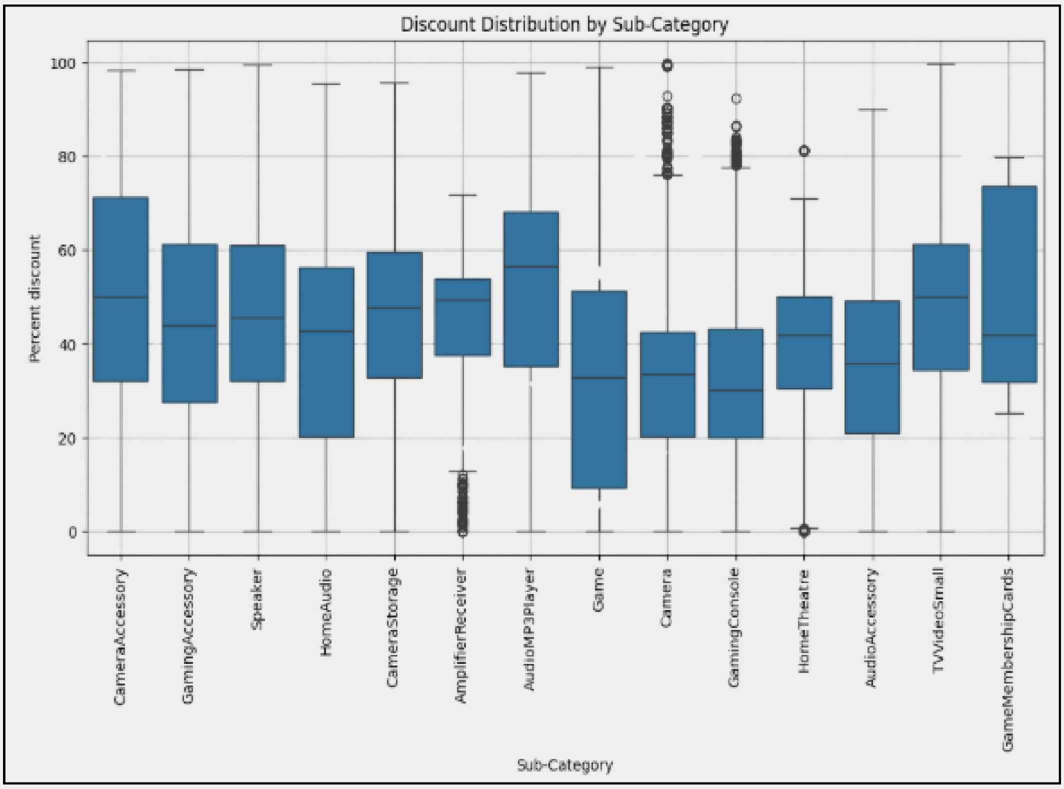
A steady rise in temperature is observed from early January to mid-February, indicating a transition from winter to early spring. However, fluctuations in March suggest periodic weather disturbances before a further rise in temperature.

iv) Discount and Revenue





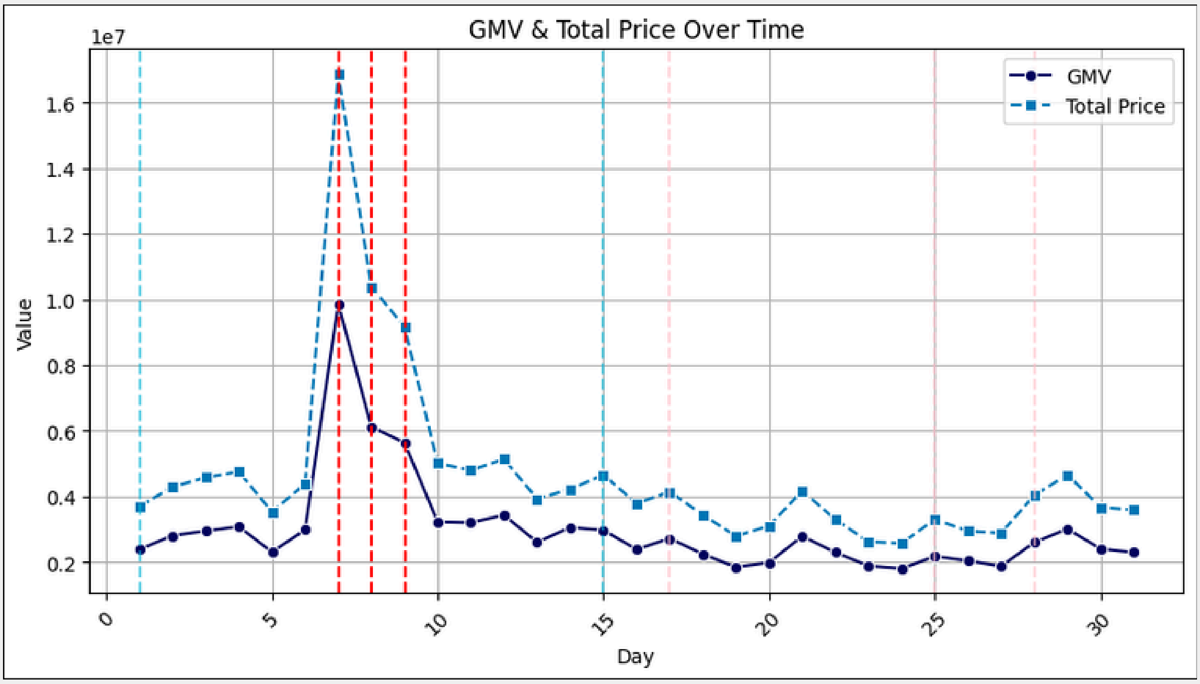
This correlation plot visually represents the impact of various marketing channels on revenue across different product categories. The intensity of blue and red shades indicates the strength and direction of these influences, with blue showing positive correlation and red indicating negative correlation.

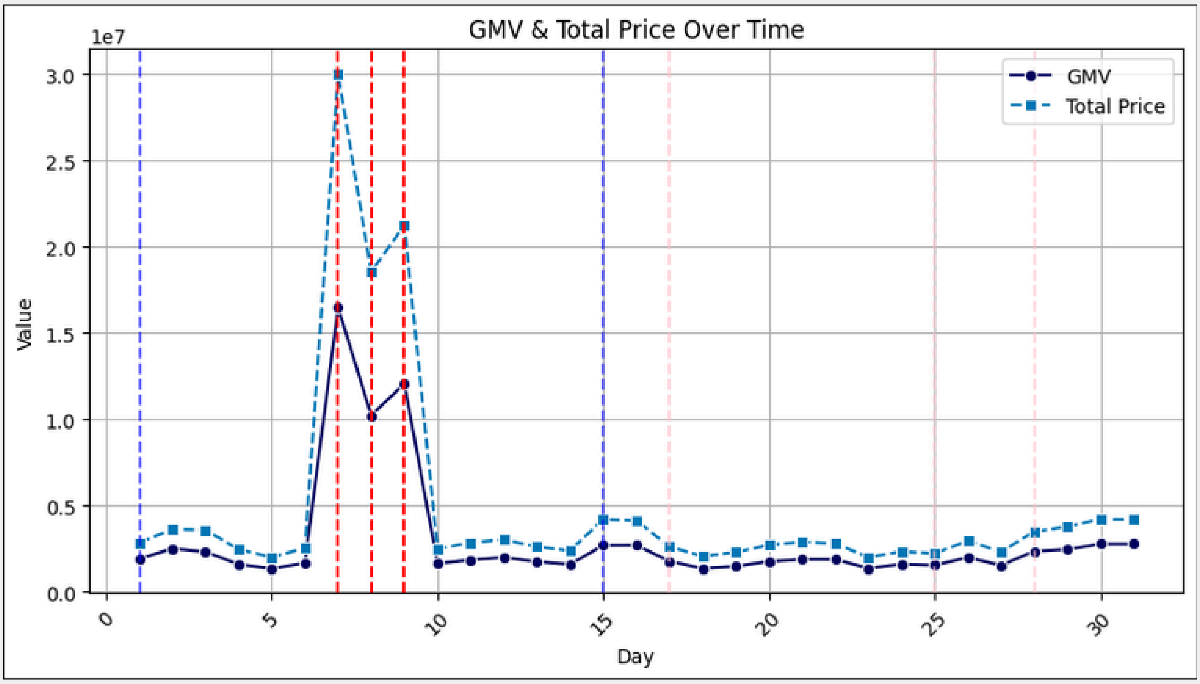


The box plot visualizes discount distribution across sub-categories, highlighting variations and outliers. Game Membership Cards, Audio MP3 Players, and Camera Accessories have the highest median discounts, while Cameras, Gaming Consoles, and Home Theatres show lower medians with wider ranges. Game and Camera Storage exhibit high variability with extreme outliers, whereas Amplifier/Receivers have the lowest discount range, indicating minimal price reductions.

v) Impact of holidays and paydays

Comparison of cumulative MRP and total revenue





The graph demonstrates that revenue on sale days is substantially higher than on regular days. It also indicates that discount rates are notably increased during sales events. The peak in revenue occurs a few days after the start of the month, suggesting a trend where consumers tend to purchase comfort items after allocating funds for essential expenses like bills.

Key Performance Indicators (KPIs)

Key Performance Indicators (KPIs) are measurable values that demonstrate how effectively an organization is achieving its key business objectives. We have also computed five main KPI's to quantitively measure the effectiveness of our project. These are as follows:

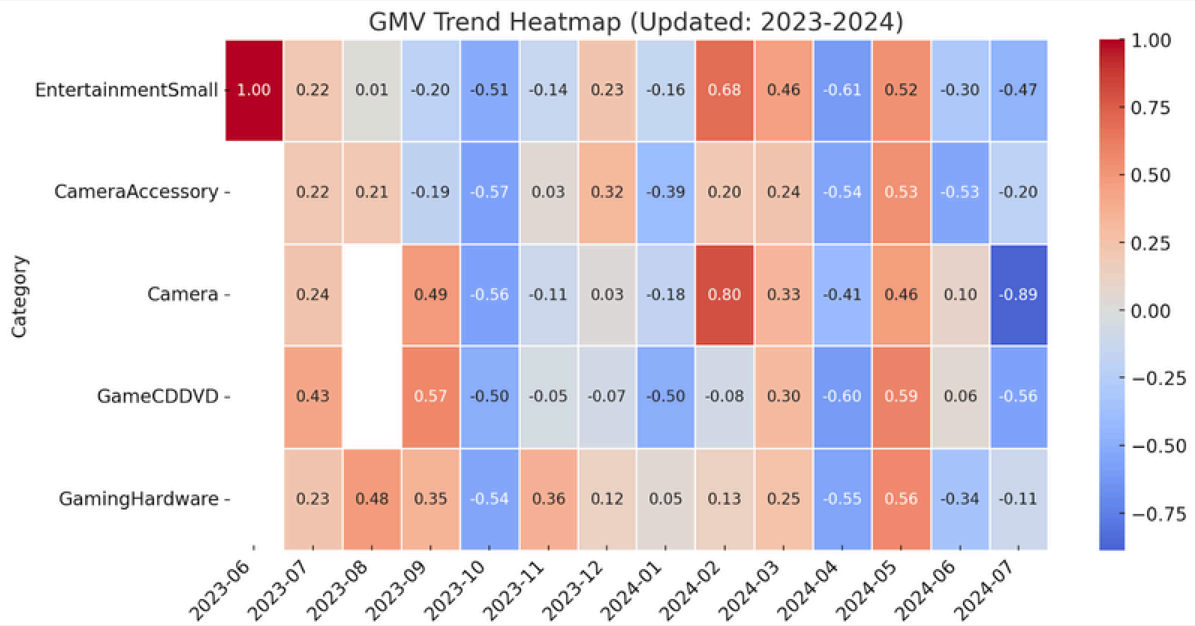
1. **Average GMV:** Represents the mean revenue per order or time period, calculated by dividing the total GMV by the number of orders or time units. This metric helps assess overall sales performance.
2. **Ad-stock for 3 Categories** (Camera, CameraAccessory, EntertainmentSmall, GameCDDVD, GamingHardware): Measures how advertising in these specific product categories influences future sales, capturing the carryover effect of marketing efforts across multiple sales periods.
3. **Moving Average of Last 3 Weeks** (GMV per Unit, Discount Percentage): Helps smooth out short-term fluctuations by calculating the rolling average of GMV per unit and discount percentage over the last three weeks, providing a clearer trend in sales performance.
4. **Lag Variables for 3 Weeks** (GMV per Unit, Discount Percentage): Tracks the delayed impact of GMV per unit and discounts from previous weeks on current sales, helping identify patterns in purchasing behavior and the long-term effect of pricing strategies.

Secondary Key Performance Indicators

1. **Average Discount**: Measures the average percentage of discounts applied across all sales. This helps assess the effectiveness of pricing strategies, promotions, and their impact on sales volume.
2. **Customer Repeat Rate:** Tracks the percentage of customers who make multiple purchases within a specific period. A higher repeat rate indicates strong customer loyalty and retention.
3. **Sale Day Impact:** Analyzes how major sales events (e.g., Black Friday, holiday discounts) influence revenue, customer traffic, and buying behavior, helping in future promotional planning.

Seasonal Trend Analysis

This analysis explores how temperature correlates with units sold across five product categories. The correlations indicate how strongly sales move in sync with temperature changes, with positive values suggesting higher sales in warmer months and negative values indicating lower sales in warm months (or vice versa).



Category	Best Months (High Correlation)	Worst Months (Low Correlation)	Seasonality Impact
EntertainmentSmall	Jun, Feb, May	Apr, Oct, Jul	High
CameraAccessory	May, Dec, Jul	Oct, Apr, Jun	Moderate
Camera	Feb, May, Sept	Jul, Oct, Apr	High
GameCDDVD	July Sept, May	Apr, Oct, Jul	High
GamingHardware	Aug, May, Nov	Apr, Oct, Jun	Moderate

Key Takeaways:

- Warmer months boost sales for EntertainmentSmall, Cameras, and GameCDDVD.
- GamingHardware & CameraAccessory show mixed trends, but still favor warm weather.
- Winter months (October–February) negatively impact most categories.
- April is consistently a low-sales month across multiple categories.

Model Selection

In the provided dataset, we initially had 16 lakh data points collected on a day-to-day basis. However, since the corresponding investment data was available only on a monthly basis, we aggregated the daily data into monthly figures, reducing the dataset to just 12 points to derive meaningful patterns. This aggregation reduced our dataset significantly, leaving us with only 12 data points.

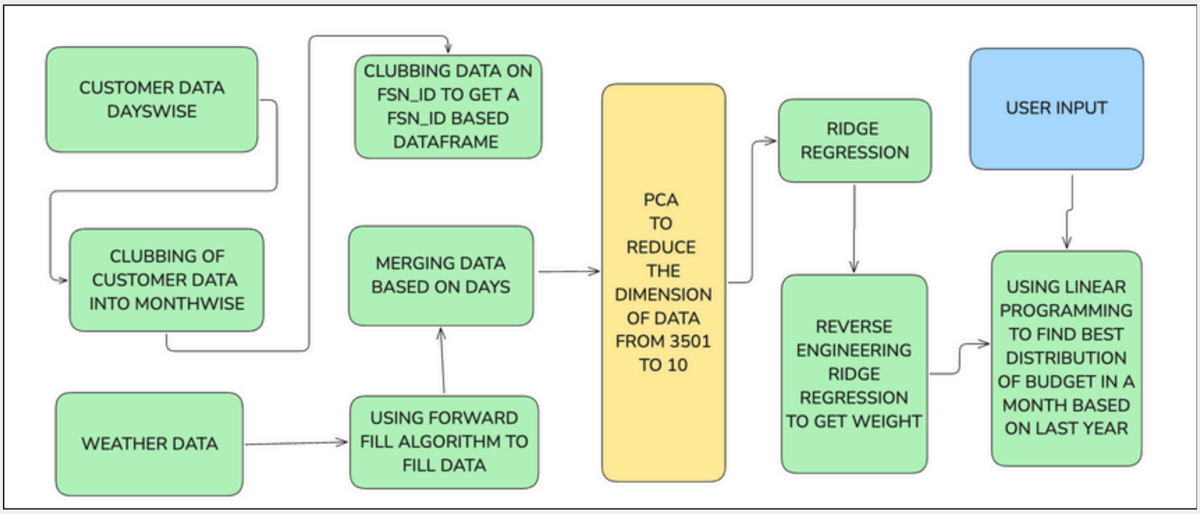
With only 12 monthly points and 60 analytical category points, we needed models optimized for small datasets that still deliver accurate insights. We evaluated the following six models:

1. **Linear Model:**
    - This model assumes a direct linear relationship between advertising spend and sales, effectively capturing straightforward correlations. However, it does not account for diminishing returns or saturation effects with increased advertising expenditure.
2. **Multiplicative Model**:
    - This model assumes advertising effects interact with factors like promotions, pricing, or seasonality. Unlike the linear model, it captures complex interactions where variables amplify each other, offering more nuanced insights.
3. **Koyck Model (Lagged Effect Model):**
    - The Koyck model is particularly valuable for capturing delayed or lagged effects of advertising. It acknowledges that advertising might not lead to immediate sales but could influence consumer behavior over subsequent periods. This model incorporates historical advertising data into current period forecasts, thus providing a realistic representation of delayed consumer responses.
4. **Lasso Regression (L1 Regularization):**
    - Lasso regression is an advanced modeling technique that performs feature selection by penalizing the absolute values of regression coefficients. This regularization helps identify and retain the most impactful predictors while eliminating or reducing less influential variables. It is highly beneficial for small datasets, preventing overfitting and improving model interpretability.
5. **Ridge Regression (L2 Regularization):**
    - Ridge regression is similar to Lasso but penalizes the squared magnitude of coefficients. This approach shrinks the coefficients towards zero without entirely removing any variables. Ridge regression is particularly useful when dealing with multicollinearity in small datasets, stabilizing the estimates and improving prediction accuracy.
6. **Polynomial Regression:**
    - Polynomial regression captures non-linear relationships by including polynomial terms (e.g., squared or cubic terms of predictors). This model can represent more complex, curved relationships between advertising spend and sales, potentially capturing diminishing returns or saturation effects better than linear models. However, caution is needed with small datasets, as high-order polynomials may cause overfitting.

Finally we decided to take the month based classification over the subclassification based division as it was easier for us to assume that the investments were equally distributed in the month based classification while we couldn't do the same in the latter.

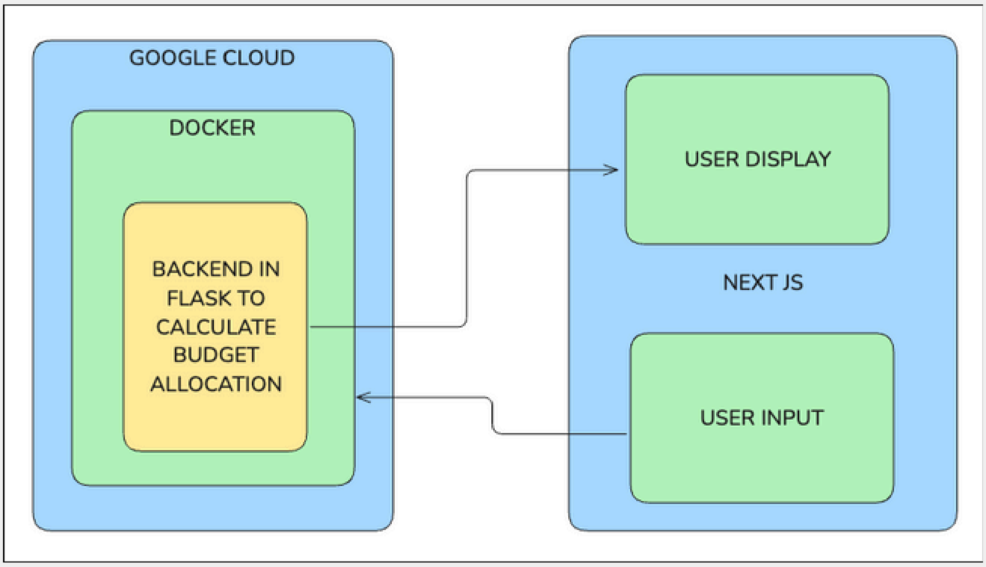
Workflow and Application Design

Model Pipeline



This pipeline starts with collecting and clustering customer data daywise, then aggregating it monthwise alongside weather data. Data is also grouped by FSN_ID, merged by days, and missing values are imputed using a forward fill algorithm. Principal Component Analysis (PCA) reduces dimensionality from 3501 features to 10, followed by Ridge Regression modeling. The model's weights are extracted (reverse-engineered), then linear programming is used to determine the optimal monthly budget distribution based on last year's data. Finally, the user can input parameters to refine or customize this process.

Application Design



A Docker container deployed on Google Cloud hosts a Flask backend that handles the core budget allocation logic. The frontend, built with Next.js, provides a user interface for inputting parameters and displaying results. When a user submits data, it's sent to the Flask backend for processing, which then returns the calculated budget allocations. Next.js updates the interface in real-time, ensuring the user can quickly view and adjust their budget scenarios as needed.

BackEnd: Flask (Python) , Frontend : Next.Js , Cloud Service : Google Cloud Service

Results

Model Name	R^2
Linear Model	-0.89
Multiplicative Model	-1.2
Koyck Model (Lagged Effect Model)	0.11
Lasso Regression (L1 Regularization)	0.05
Ridge Regression (L2 Regularization)	0.87
Polynomial Regression	0.1

This table compares how well each model fits the data based on the R^2 value (closer to 1 means better fit). Both the Linear and Multiplicative Models have negative R^2 scores, indicating they fit the data worse than a naive baseline. Koyck (which accounts for lagged effects), Lasso, and Polynomial Regression all achieve low positive R^2 values, showing limited predictive power. By contrast, Ridge Regression yields a strong R^2 of 0.87, making it the most accurate model for this particular budget allocation problem.

Dashboard Overview

We have created an interactive dashboard to visualise the company's status at a given point through charts and graphs

1. **Home tab (Monthly, Yearly):** This Provides an at-a-glance view of the company's overall performance, including total revenue, total sales, net profit, and other key metrics broken down by month and year.
2. **Categories tab:** Focuses on product categories and subcategories, highlighting sales figures, discount rates, and customer retention statistics. This section helps identify top-performing categories and areas needing improvement.
3. **Investment tab:** Gives managers a tool to select dates, forecast sales, and see projected profit or loss. By adjusting time ranges and sales parameters, managers can assess potential outcomes for different investment scenarios.
4. **Details tab:** Displays the entire dataset in an interactive format, allowing managers to filter, sort, and explore the data for deeper insights. This helps in identifying patterns or correlations that might not be immediately visible.
5. **Discount + Weather tab:** Shows how seasonal changes and weather patterns influence sales. It visualizes correlations between discounts, external factors, and customer demand, offering insights into how to optimize promotions throughout the year.

Conclusion

This analysis highlights the value of a comprehensive, data-driven approach for optimizing ElectroMart's marketing budget. By carefully cleaning and aggregating daily transaction data into monthly figures, we addressed the mismatch in data granularity, while also capturing important seasonal and weather-related patterns. Warmer months consistently showed higher sales for categories like EntertainmentSmall, Cameras, and GameCDDVD, whereas winter months (October–February) tended to reduce overall demand. GamingHardware and CameraAccessory displayed mixed but generally positive correlations with warmer temperatures, and April emerged as a consistently low-sales month across multiple categories.

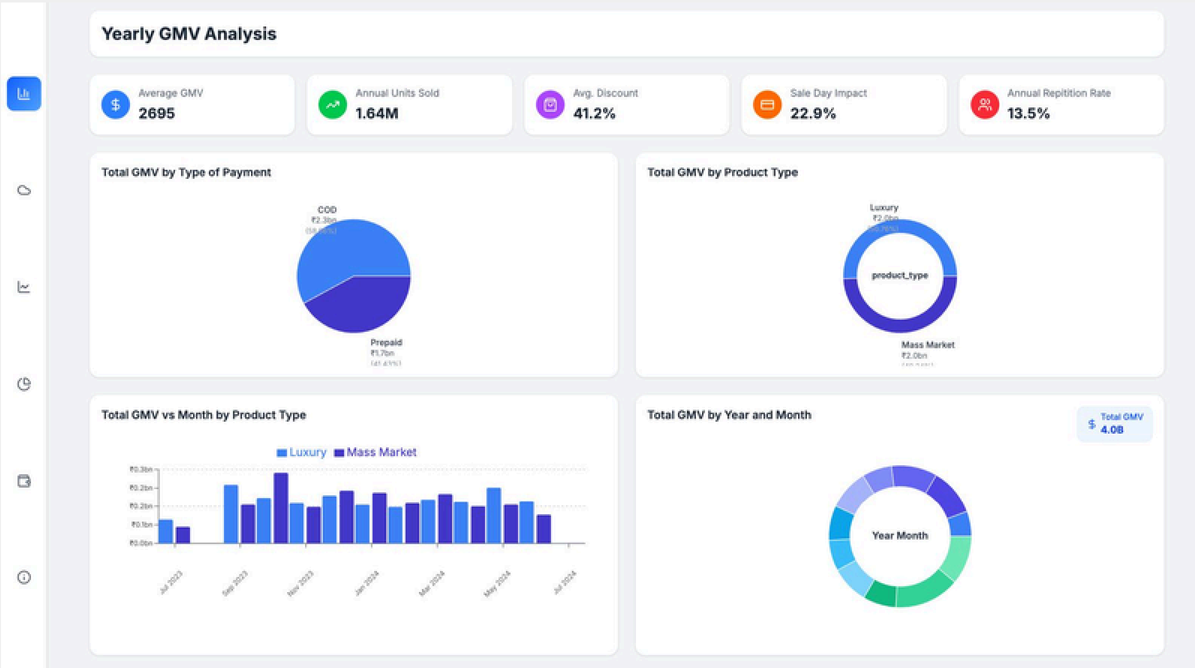
On the modeling front, Ridge Regression performed best, achieving an **R^2 of 0.87**, indicating a strong predictive capacity even with the limited monthly dataset. This model was then integrated into a budget optimization framework, enabling refined allocation of marketing resources based on historical spend and performance data.
The application's architecture ensures seamless data flow and user interaction. A Docker container on Google Cloud hosts the Flask backend, which handles core budget allocation logic, while a Next.js frontend provides an intuitive interface for parameter input and real-time visualization of results. This setup allows managers to explore different scenarios, monitor performance metrics, and adjust their marketing strategies in a dynamic, user-friendly environment.

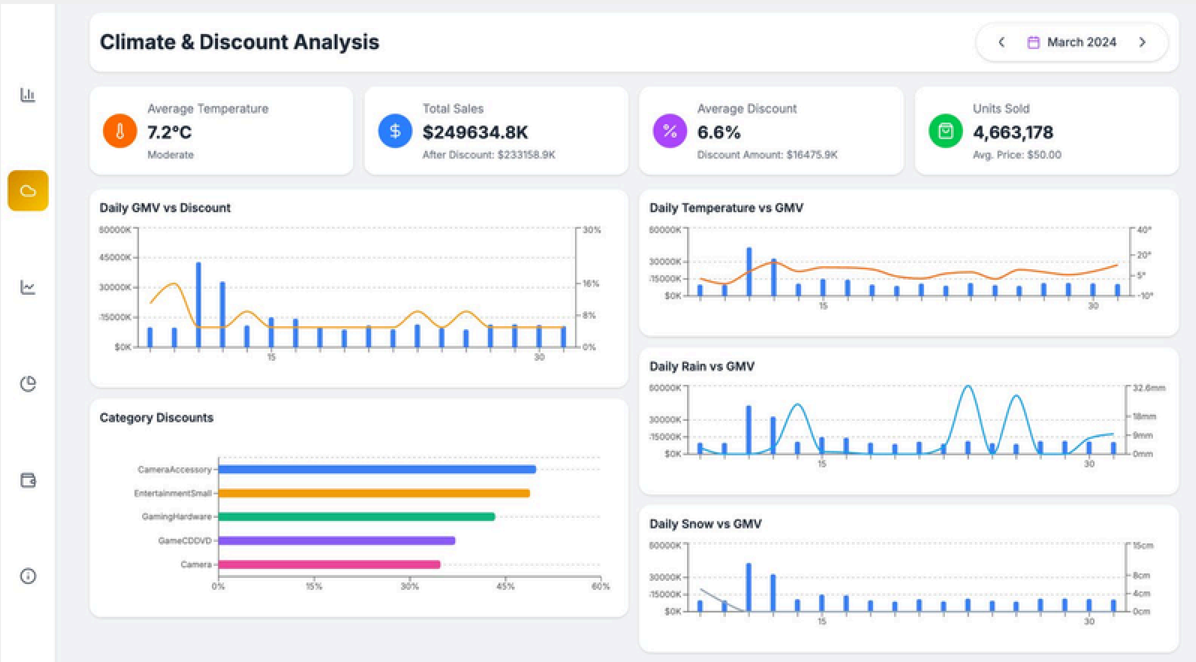
Overall, the analysis underscores the importance of incorporating seasonal trends, rigorous modeling, and robust application design. By continuously refining these components, ElectroMart can adapt to changing market conditions and optimize future marketing investments for maximum return.

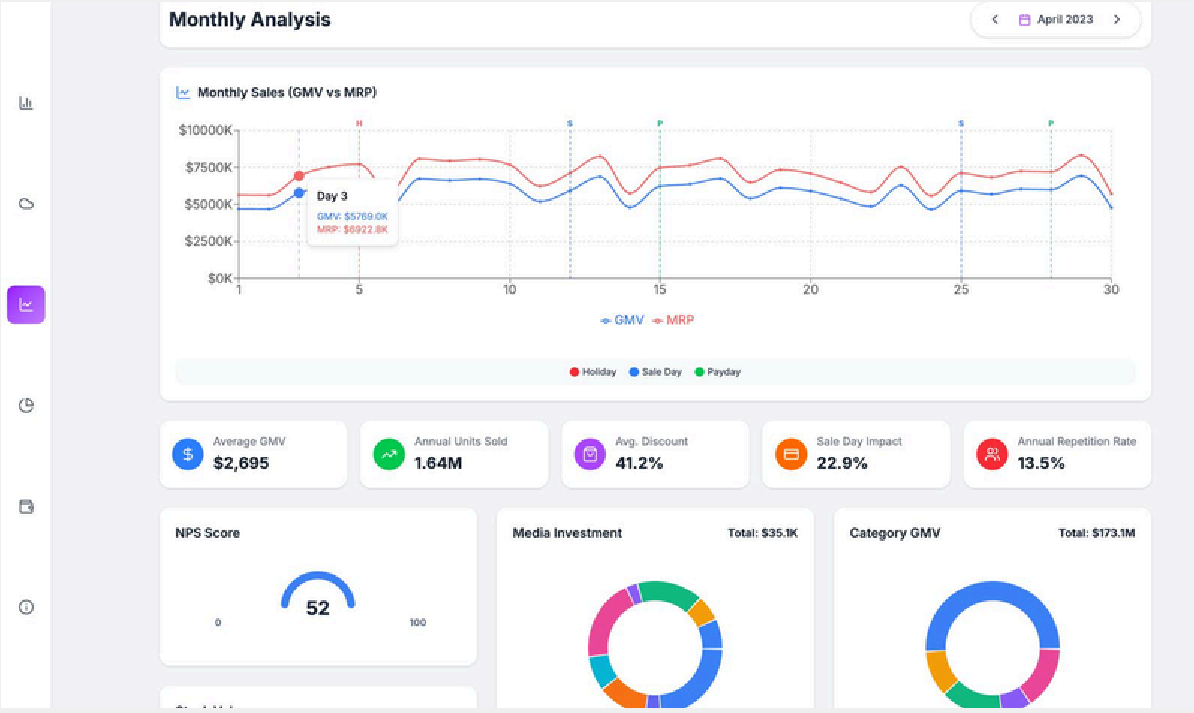
Website Link:

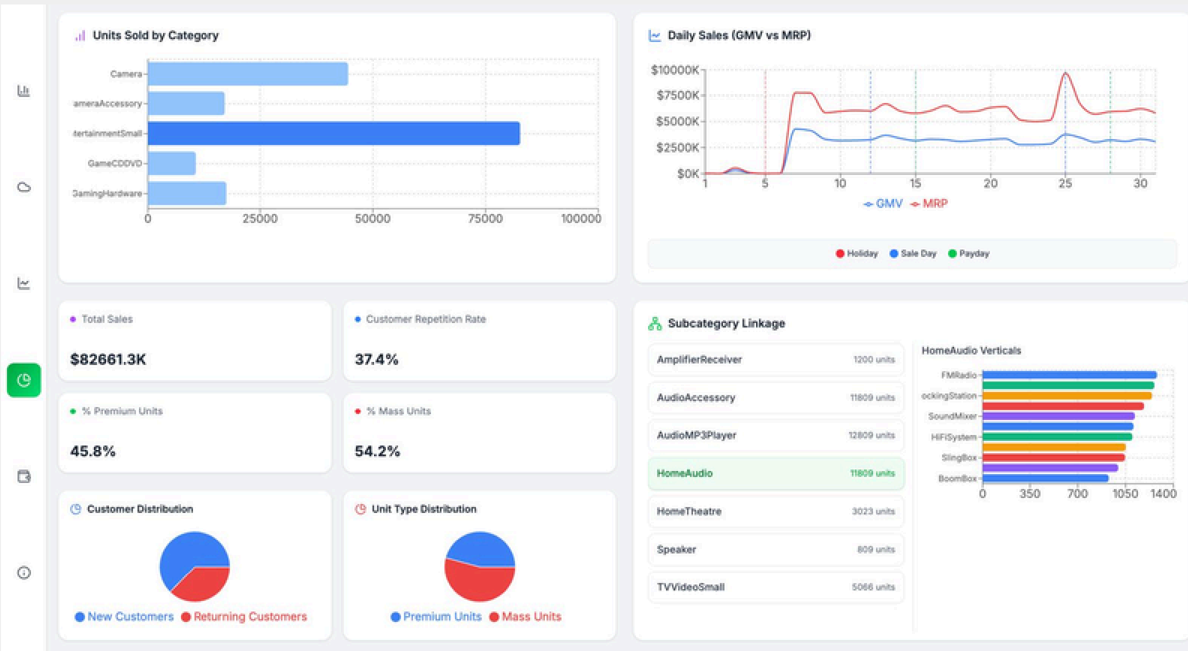
https://bit.ly/4iNI7Aq

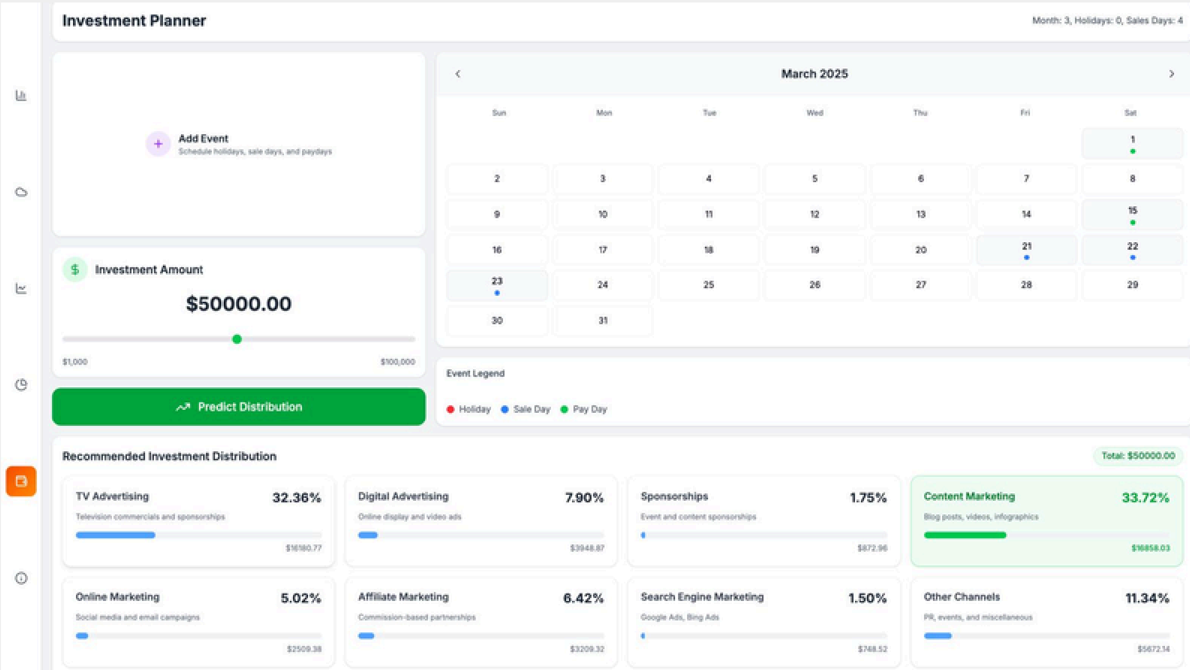
ANNEXURE


Yearly GMV Analysis
Average GMV
2695
Annual Units Sold
1.64M
Avg. Discount
41.2%
Sale Day Impact
22.9%
Annual Repitition Rate
13.5%
Total GMV by Type of Payment
COD
Prepaid
Total GMV by Product Type
Luxury
product_type
Mass Market
Total GMV vs Month by Product Type
Luxury
Mass Market
Total GMV by Year and Month
Total GMV
4.0B
Year Month

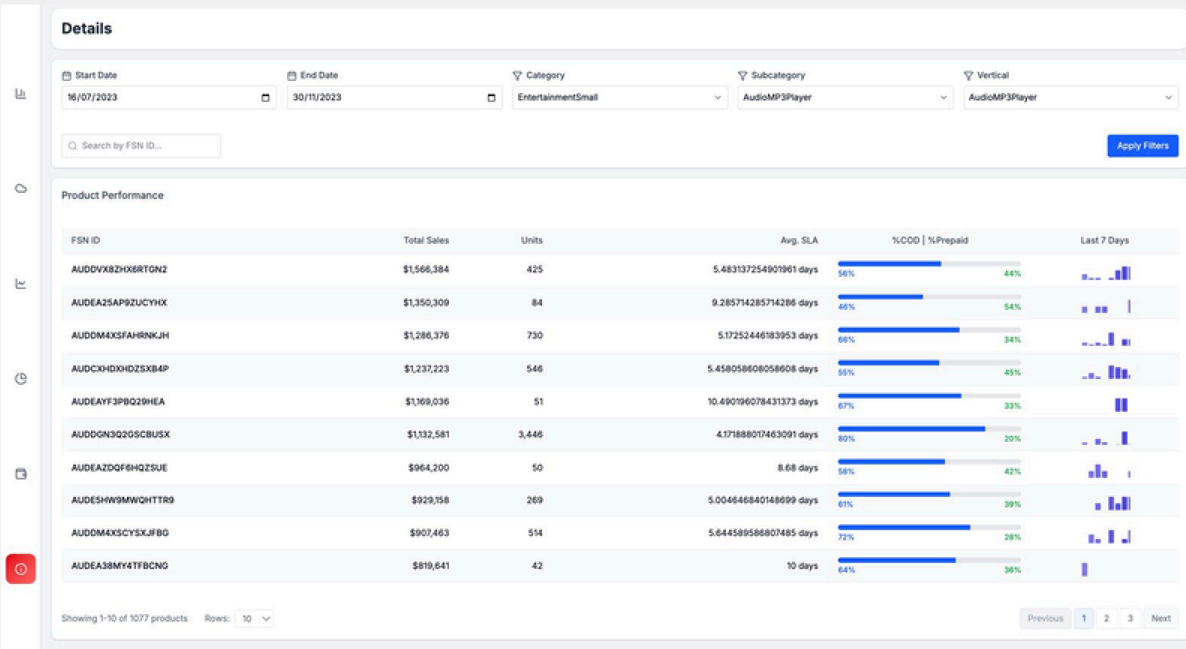

Climate & Discount Analysis
March 2024
Average Temperature
7.2°C
Moderate
Total Sales
$249634.8K
After Discount: $233158.9K
Average Discount
6.6%
Discount Amount: $16475.9K
Units Sold
4,663,178
Avg. Price: $50.00
Daily GMV vs Discount
Daily Temperature vs GMV
Category Discounts
CameraAccessory
EntertainmentSmall
GamingHardware
GameCDDVD
Camera
Daily Rain vs GMV
Daily Snow vs GMV


Monthly Analysis
April 2023
Monthly Sales (GMV vs MRP)
$10000K
$7500K
$5000K
$2500K
$0K
Day 3
GMV: $5769.0K
MRP: $6922.8K
GMV
MRP
Holiday
Sale Day
Payday
Average GMV
$2,695
Annual Units Sold
1.64M
Avg. Discount
41.2%
Sale Day Impact
22.9%
Annual Repetition Rate
13.5%
NPS Score
52
0
100
Media Investment
Total: $35.1K
Category GMV
Total: $173.1M


Units Sold by Category
Camera
GameCDDVD
0
25000
50000
75000
100000
Daily Sales (GMV vs MRP)
$10000K
$7500K
$5000K
$2500K
$0K
GMV
MRP
Holiday
Sale Day
Payday
Total Sales
$82661.3K
Customer Repetition Rate
37.4%
% Premium Units
45.8%
% Mass Units
54.2%
Subcategory Linkage
AmplifierReceiver
1200 units
AudioAccessory
11809 units
AudioMP3Player
12809 units
HomeAudio
11809 units
HomeTheatre
3023 units
Speaker
809 units
TVVideoSmall
5066 units
HomeAudio Verticals
FMRadio
SoundMixer
HiFiSystem
SlingBox
BoomBox
0
350
700
1050
1400
Customer Distribution
New Customers
Returning Customers
Unit Type Distribution
Premium Units
Mass Units



Investment Planner
Month: 3, Holidays: 0, Sales Days: 4
Add Event
Schedule holidays, sale days, and paydays
March 2025
Sun Mon Tue Wed Thu Fri Sat
Investment Amount
$50000.00
$1,000
$100,000
Predict Distribution
Event Legend
Holiday
Sale Day
Pay Day
Recommended Investment Distribution
Total: $50000.00
TV Advertising 32.36%
Television commercials and sponsorships
$16180.77
Digital Advertising 7.90%
Online display and video ads
$3948.87
Sponsorships 1.75%
Event and content sponsorships
$872.96
Content Marketing 33.72%
Blog posts, videos, infographics
$16858.03
Online Marketing 5.02%
Social media and email campaigns
$2509.38
Affiliate Marketing 6.42%
Commission-based partnerships
$3209.32
Search Engine Marketing 1.50%
Google Ads, Bing Ads
$748.52
Other Channels 11.34%
PR, events, and miscellaneous
$5672.14



Details
Start Date
16/07/2023
End Date
30/11/2023
Category
EntertainmentSmall
Subcategory
AudioMP3Player
Vertical
AudioMP3Player
Search by FSN ID...
Apply Filters
Product Performance
FSN ID Total Sales Units Avg. SLA %COD | %Prepaid Last 7 Days
AUDDVX8ZHX6RTGN2 $1,566,384 425 5.483137254901961 days 56% 44%
AUDEA25AP9ZUCYHX $1,350,309 84 9.285714285714286 days 46% 54%
AUDDM4XSFAHRNKJH $1,286,376 730 5.17252446183953 days 66% 34%
AUDCXHDXHDZSXB4P $1,237,223 546 5.458058608058608 days 55% 45%
AUDEAYF3PBQ29HEA $1,169,036 51 10.490196078431373 days 67% 33%
AUDDGN3Q2GSCBUSX $1,132,581 3,446 4.171888017463091 days 80% 20%
AUDEAZDQF6HQZSUE $964,200 50 8.68 days 58% 42%
AUDESHW9MWQHTTR9 $929,158 269 5.004646840148699 days 61% 39%
AUDDM4XSCYSXJFBG $907,463 514 5.644589586807485 days 72% 28%
AUDEA38MY4TFBCNG $819,641 42 10 days 64% 36%
Showing 1-10 of 1077 products
Rows: 10
Previous 1 2 3 Next