TEAM DAXXX

Team 2025113

CONTENTS

1. OVERVIEW:

1.1) Business Context

ElectroMart is a Canada-based e-commerce company specializing in electronic products. Over the past year, the company has made significant marketing investments, including digital advertisements, traditional media campaigns, and promotional sales events. The leadership team now seeks data-driven insights to optimize marketing spending for the following year.

1.2) Problem Statement

The objective is to analyze past marketing investments, evaluate their impact on revenue, and allocate future budgets effectively to maximize returns. This involves:

a. Identifying Performance Drivers – Determining which KPIs significantly impact revenue
b. Measuring Marketing ROI – Quantifying the impact of different marketing channels
c. Optimizing Future Marketing Budget Allocation – Reallocating resources to the most impactful marketing strategies while minimizing unproductive spending
d. Visualization – Creating a dashboard that can display company statistics in an intuitive manner

1.3) Approach

We used data visualization tools to generate insights from Daily Consumer, Media Investment, Weather, Sales, and Holiday datasets. Next, we conducted univariate and multivariate analysis of the variables. Based on these results, we engineered features, Key Performance Indicators (KPIs), Key Risk Indicators (KRIs) and identified key impact areas. To build a robust model pipeline, we categorized the data by product type, applying Marketing Mix Modeling to each segment's monthly media investment and gross merchandise value (GMV) data. Finally, we integrated the outputs of these models to determine the optimal budget allocation for the upcoming year.

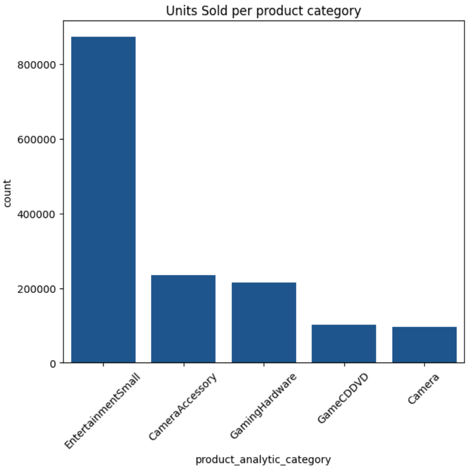
1.4) Technology Stack

To conduct a thorough data analysis and build an interactive solution, the following technology stack was used:

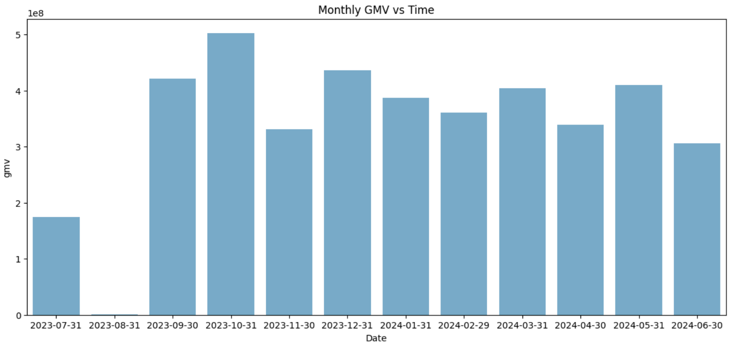
1. Data Processing & Analysis: Python (Pandas, NumPy, Scipy, Statsmodels) - Python provides efficient data processing and analysis techniques.
2. Frontend Development: React.js - Used to create a responsive & dynamic frontend.
3. Backend Development: Node.js - For robust & scalable backend development
4. Version Control: GitHub - To ensure version control & collaboration

2. Exploratory Data Analysis:

2.1) Data Summary



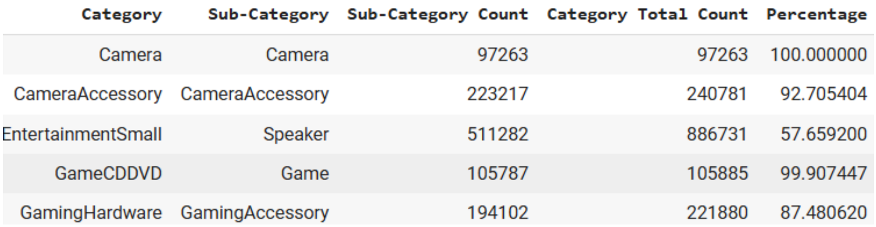
The customer order dataset has 20611 unique products divided into 74 verticals, 14 sub-categories, and 5 categories. Over the course of one year, 123360 unique customers have used the website to buy something. The dataset is imbalanced both in terms of Total Investment and in terms of product categories.



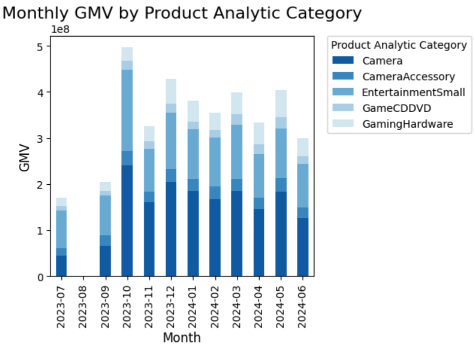
The bar chart illustrates the monthly Gross Merchandise Value over time, showing fluctuations across different months. October 2023 recorded the highest GMV, whereas August 2023 had a smaller bar due to fewer data points.

2.2) Univariate Analysis

First, we analyze product information. The table illustrates the most sold sub-categories within each category. The "EntertainmentSmall" category has the highest overall sales, with "Speaker" being the most sold sub-category, contributing 57.65% of the category's sales. The data reveals a significant class imbalance, not only within sub-categories but also across product categories. The "EntertainmentSmall" category dominates sales, while categories like "GameCDDVD" and "Camera" have relatively lower contributions.

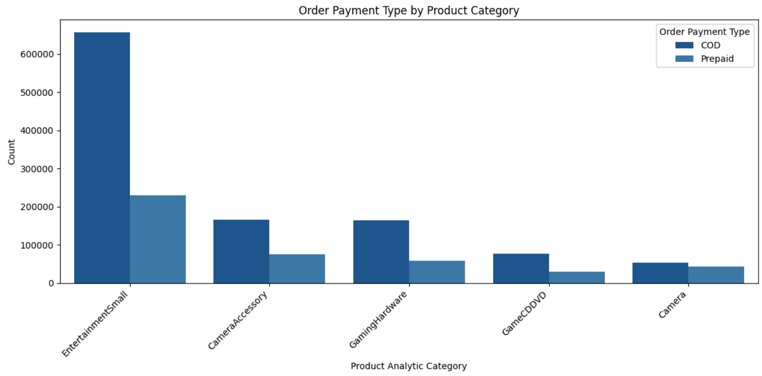


Category	Sub-Category	Sub-Category Count	Category Total Count	Percentage
Camera	Camera	97263	97263	100.000000
CameraAccessory	CameraAccessory	223217	240781	92.705404
EntertainmentSmall	Speaker	511282	886731	57.659200
GameCDDVD	Game	105787	105885	99.907447
GamingHardware	GamingAccessory	194102	221880	87.480620



The stacked bar chart illustrates the Monthly GMV (Gross Merchandise Value) by Product Category, highlighting sales trends across different product categories over time. Camera and EntertainmentSmall categories consistently contribute the largest share of GMV, demonstrating their strong market presence.

The chart below illustrates the distribution of order payment types, Cash on Delivery (COD) and Prepaid, across different product categories. COD is the dominant payment method for all the product categories. This trend indicates a strong customer preference for COD, particularly in high-selling categories.

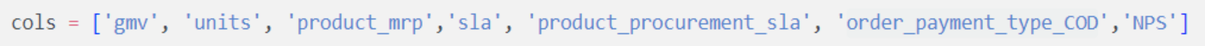


2.3) Distribution Analysis

To analyze the dataset effectively, we first need to determine the distribution of key variables. Identifying whether they follow a normal distribution is essential for selecting appropriate statistical techniques and ensuring valid inferences. So, we chose the Kolmogorov-Smirnov test because it is effective for both small and large sample sizes and does not assume prior knowledge of parameters.

Kolmogorov-Smirnov (K-S) test is a non-parametric test that compares the empirical distribution function of the sample data with the expected cumulative distribution function of a normal distribution.

The test was applied to the following variables:



```
cols = ['gmv', 'units', 'product_mrp','sla', 'product_procurement_sla', 'order_payment_type_COD','NPS']
```

The K-S test results indicate that all selected variables follow a normal distribution. Each variable's P-value was greater than 0.05, confirming that the data follows a normal distribution.

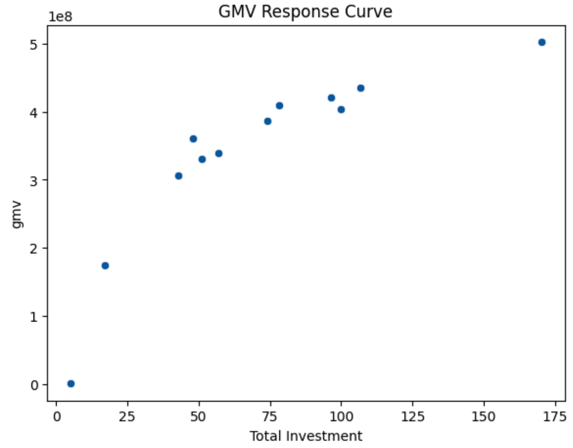
The normal distribution of these variables facilitates analysis in the following ways:

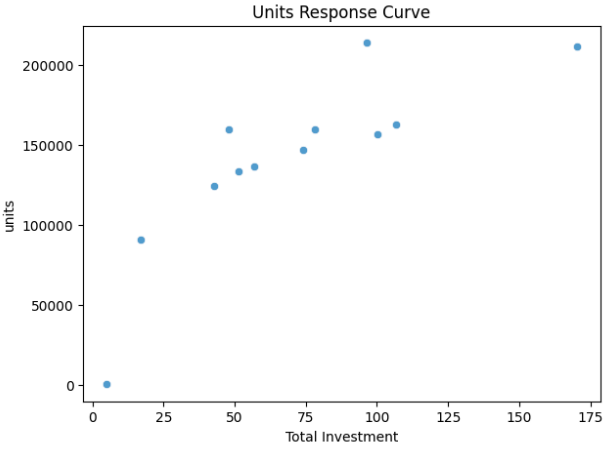
1. Improved Model Performance - Having target variables with better distributional properties can improve optimization.

2. Better Statistical Properties in Small Samples - While large samples are robust to non-normality due to the Central Limit Theorem, in smaller samples, the normality of residuals becomes more critical for valid hypothesis testing.

2.4) Bivariate Analysis

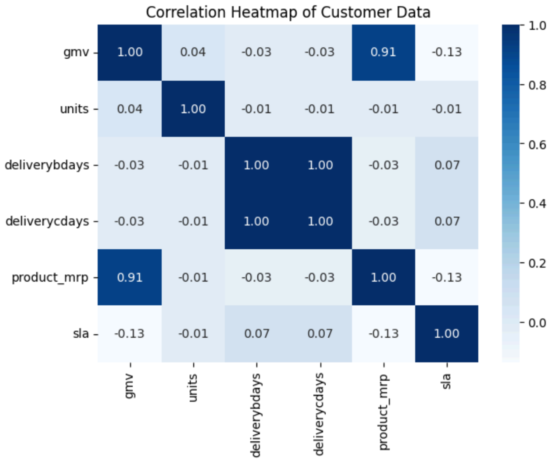
Below is the scatter plot for GMV and Units with Total Investment. Initially, there was a rise in GMV and units sold with increased investment, but beyond a certain point, the incremental gains did not match the additional spending.



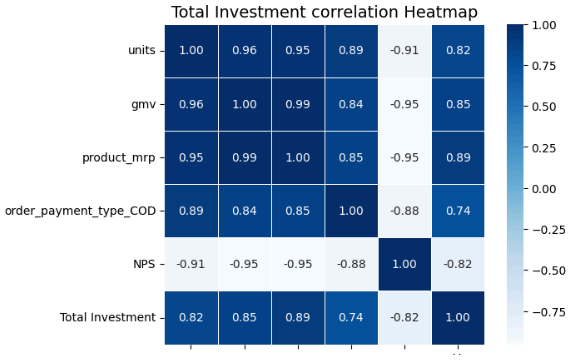


2.5) Pearson Correlation

The Pearson correlation test is a statistical method used to measure the linear relationship between two continuous variables. It provides the Pearson correlation coefficient (r), which quantifies the strength and direction of this relationship on a scale from -1 to 1. A score of 1 signifies a high positive correlation, while -1 is a high negative correlation.



The heatmap displays the correlation between various customer data attributes. Strong positive correlations are observed between GMV (Gross Merchandise Value) and product_mrp (0.91), as well as ‘deliverycdays’ and ‘deliverybdays’ (1.00) which are in accordance with their definitions.



The correlation heatmap of Total Investment with other monthly variables is displayed. Strong correlations are found between GMV and units (0.96). On the other hand, NPS was found to have highly negative correlations with all other variables

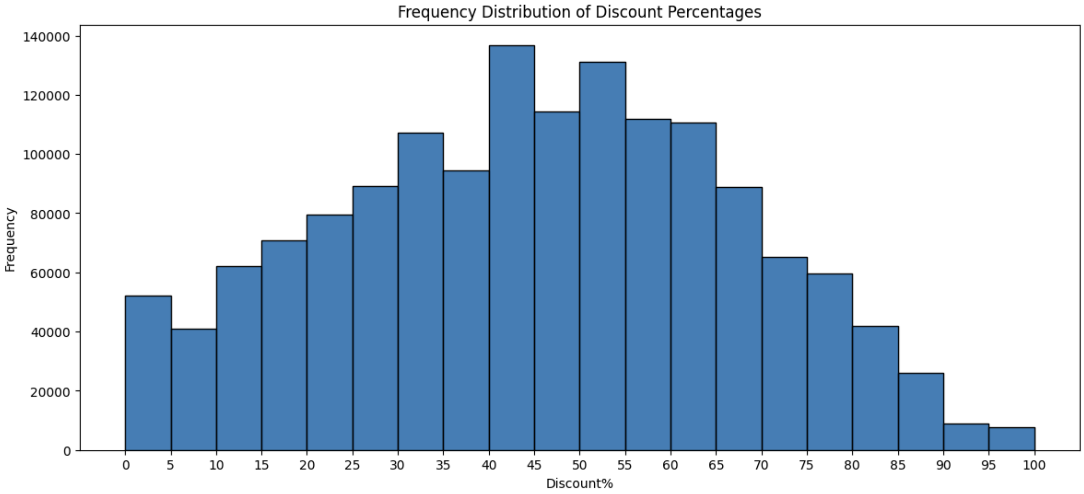
3. Feature Engineering:

3.1) Sale Price

The sale price represents the Gross Merchandise Value (GMV) per unit, which is the actual selling price of a product after applying discounts or promotional adjustments. Unlike the MRP (Maximum Retail Price), the sale price reflects customers' effective transactional value for each unit. This metric is essential for analyzing pricing strategies, assessing product performance, and optimizing revenue streams in competitive markets.

3.2) Discount Percentage

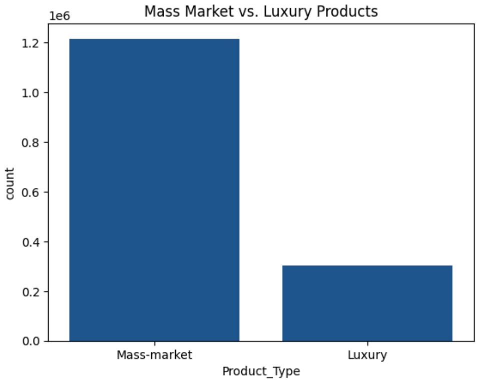
In marketing, the discount percentage refers to the reduction in price offered on a product or service, usually expressed as a percentage of the original price. It is a common pricing strategy used to attract customers, increase sales volume, and create a sense of urgency. By offering discounts, businesses can clear inventory, boost customer loyalty, and enhance their competitive edge.



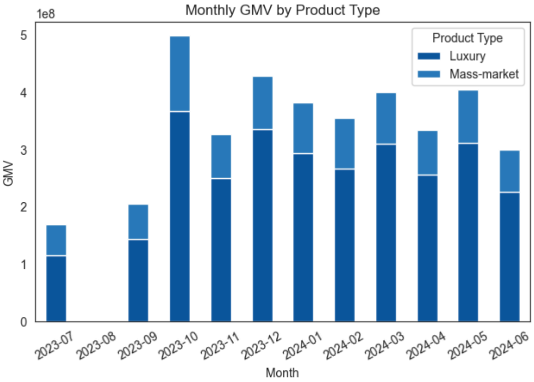
3.3) Product Type

The Product Type feature categorizes products into two segments based on their Gross Merchandise Value (GMV):

A. Luxury: Products in the top 20% of GMV.
B. Mass-market: Products in the remaining 80% of GMV.



There is a significant disparity between the sales of mass-market and luxury products. Mass-market products overwhelmingly dominate units sold, suggesting that consumer demand is significantly higher for affordable, widely accessible products than for premium or high-end alternatives.



The Monthly GMV by Product Type chart reveals that while the Luxury segment had significantly lower sales volume than Mass-market products, it contributed a substantial share to the overall GMV. This indicates that luxury products have a much higher price per unit, making them a key revenue driver despite their lower sales figures.

3.4) Payday Week

The Payday Week feature identifies whether an order was placed in a week, including a typical payday. Paydays were assumed to be on the 1st and 15th of each month, and weeks containing these dates were labeled as Payday Weeks (1), while others were labeled as Non-Payday Weeks (0).

3.5) Order Delay

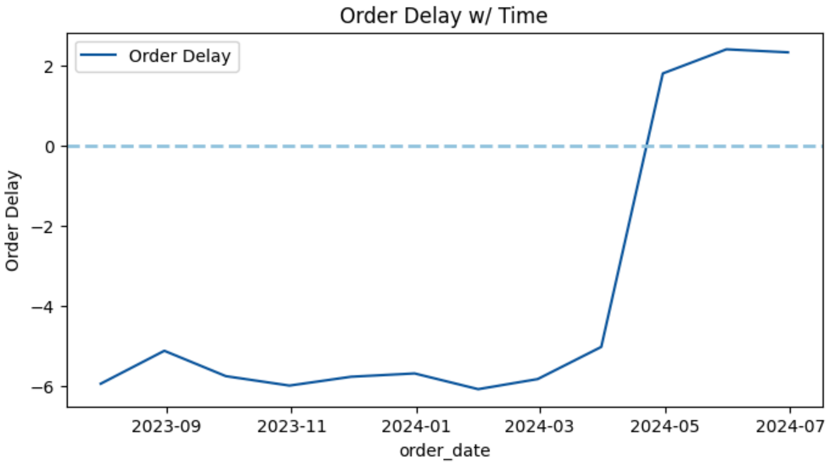
The Order Delays feature represents the total time taken for an order to be delivered, incorporating multiple components of the fulfillment process:

A. ‘Deliverybdays’: The time required to retrieve the item from the warehouse and prepare it for shipping
B. ‘Deliverycdays’: The time taken to transport and deliver the item to the customer.
C. ‘SLA’: The standard number of days it typically takes to deliver the product as per the expected timeline

The total order delay is calculated as follows:

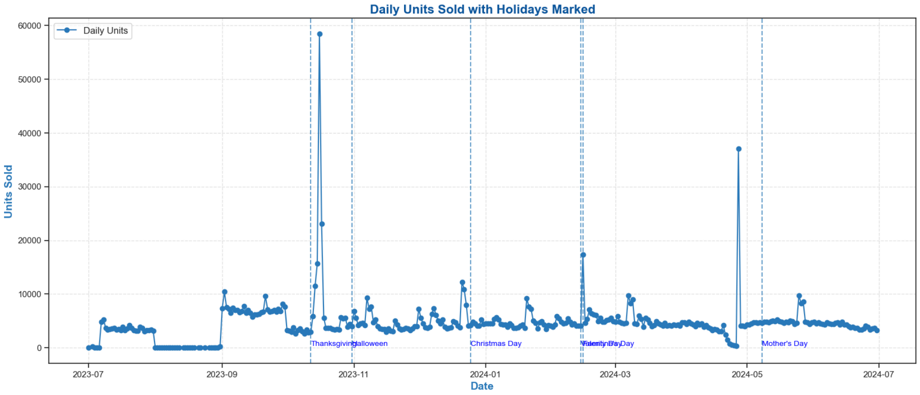
$$Order\ Delays = Deliverybdays + Deliverycdays - SLA$$

A drastic delay in order delivery was observed from March 24 onwards, represented as follows:



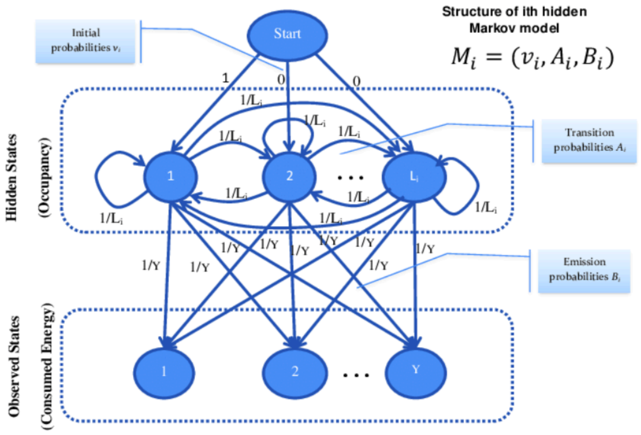
3.6) Anomalous Data

We plotted the total units sold per day vs time along with holidays and observed the increase in sales near special occasions such as holidays, paydays, and sale days. The observed spikes in sales in October, May, and February appear to correlate with them, indicating that these periods increased consumer activity. To model the behavior of this activity, we used the (HMM) since it incorporates state-specific thresholds and sequential patterns.



Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a probabilistic model used to analyze time series data where the system transitions between hidden states over time while generating observable outputs. In our time series analysis, we used HMM to model the effects of Special Occasions The key problems in HMM—evaluation, decoding, and learning—are solved using the Forward Algorithm, Viterbi Algorithm, and Baum-Welch Algorithm, respectively.

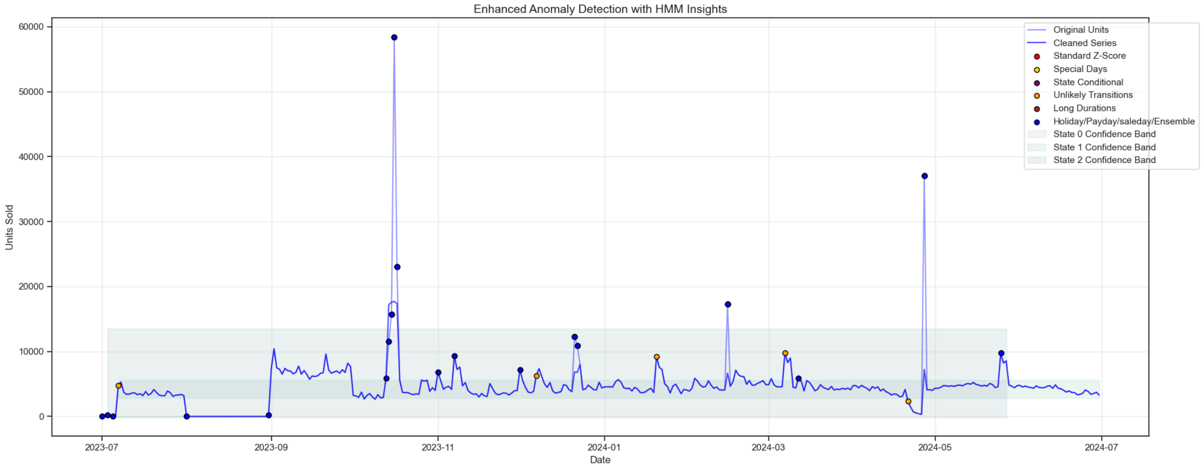

Structure of ith hidden Markov model
$M_i = (v_i, A_i, B_i)$
Initial probabilities v_i
Transition probabilities A_i
Emission probabilities B_i
Start
1
0
0
$1/L_i$
$1/Y$
1
2
...
L_i
Y
Hidden States
(Occupancy)
Observed States
(Consumed Energy)

Why HMM?

HMMs are highly effective in detecting anomalies in time-series data because they model sequential patterns and latent dependencies, enabling a robust understanding of temporal dynamics.

HMMs provide high anomaly detection accuracy with a low false alarm rate, making them particularly useful for distinguishing between normal and anomalous behavior in noisy or uncertain environments.

Result: Most Anomalous points occur during some special occasion



4. Performance Indicators:

Financial KPIs

Financial KPIs provide critical insights into revenue generation, profitability, and overall economic health of the business. These metrics help track performance and guide strategic financial decisions.

1) Monthly Sales Growth (MSG)

Monthly Sales Growth measures the percentage increase or decrease in sales revenue from one month to the next, serving as a fundamental indicator of business momentum. Calculated as:

$$MSG = \frac{(Current\ Month\ GMV - Previous\ Month\ GMV\ ) \times 100}{Previous\ Month\ GMV}$$

This metric reveals whether a business is expanding, contracting, or maintaining steady revenue streams. Positive growth indicates successful sales strategies and increasing market demand, while negative growth may signal competitive pressures, seasonal fluctuations, or underlying business challenges.

2) Sales by Payment Type

Sales by Payment Type tracks the distribution between payment options like Cash on Delivery (COD) and Prepaid transactions. Understanding this breakdown helps businesses optimize cash flow management, as COD orders typically delay revenue recognition while potentially increasing return rates. While providing immediate revenue, prepaid transactions may present conversion challenges for some customer segments. Calculated as:

$$SPT = \frac{Payment\ Type\ Value}{Total\ GMV\ (per\ month)} \times 100$$

3) Marketing Return on Investment

Marketing Return on Investment (ROI) measures the financial return generated from marketing expenditures, serving as the ultimate accountability metric for marketing effectiveness. Calculated as:

$$MROI \ = \ \frac{Sales\ Growth - Total\ Investment}{Total\ Investment} \times 100$$

This KPI evaluates how efficiently marketing budgets translate into revenue growth. A positive ROI indicates successful campaigns generating more revenue than cost, while negative values signal ineffective spending requiring strategic adjustment.

4) Special Day Performance Index (SDPI)

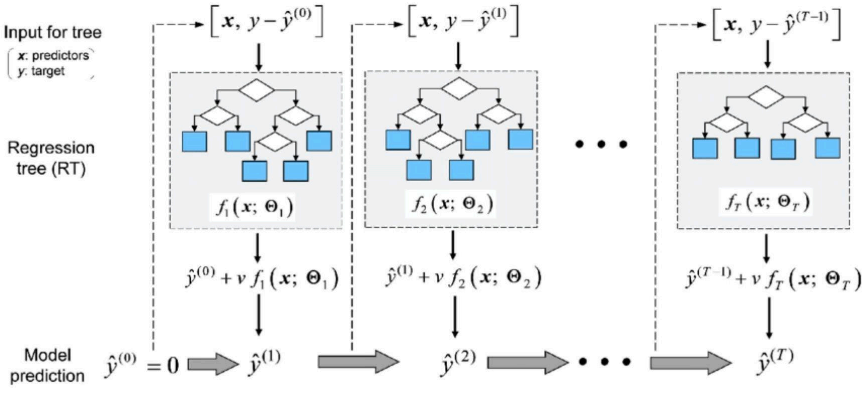
The Sales Deviation Per Holiday Index (SDPI) is a metric designed to quantify how sales on special days perform relative to forecasts, effectively isolating the incremental or underperforming revenue impacts of special days (Holidays, Paydays, Saledays), calculated as

$$SDPI \ = \ \frac{Current\ sales - Expected\ sales}{No.\ of\ holidays\ (per\ month) + Sale\ days(per\ month) + Paydays} \ ,$$

To calculate SDPI, an XGBoost model is trained on non-anomalous daily sales data, incorporating autoregressive features, pricing metrics such as mean product MRP, and revenue indicators such as Gross Merchandise Value (GMV).

XGBoost

XGBoost Regressor is a machine-learning algorithm designed for regression tasks. It builds on the principles of gradient boosting, creating an ensemble of decision trees that work together to make accurate predictions. XGBoost incorporates several innovative features, including regularization to prevent overfitting, efficient handling of sparse data, and optimized tree-building techniques. These enhancements make XGBoost highly effective in capturing complex relationships in data while maintaining computational efficiency.



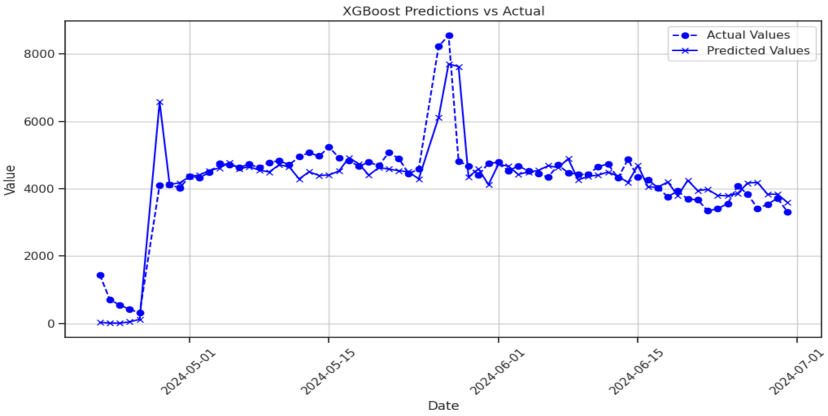
Why Xgboost?

1. Autoregressive features leverage past sales data, enabling the model to learn from temporal patterns like trend, seasonality, and cyclic behaviors, which are crucial for sales prediction.
2. Xgboost is suitable for large datasets with multiple autoregressive features. It supports parallel processing and optimization.

Results

The XGBoost model gave us an R2 score of 0.75.

This model then predicts expected sales for the whole year.



A positive SDPI indicates stronger-than-expected performance, while a negative value signals underperformance. This metric enables strategic decision-making by identifying high-potential holidays for promotions or inventory boosts.

Operational KPIs

Operational KPIs are measurable metrics that track the efficiency and effectiveness of business processes across areas like manufacturing, logistics, and customer service.

Delivery Performance Index (DPI)

Delivery Performance measures how effectively a business fulfills its order delivery commitments, directly impacting customer satisfaction and retention. Calculated as:

$$DPI = \frac{orders\ delivered\ within\ SLA}{Total\ Orders} \times 100$$

It provides visibility into fulfillment reliability. Strong delivery performance builds customer trust and drives repeat purchases, while poor performance leads to increased customer service costs and reduced lifetime value.

Key Result Areas (KRA)

KRAs define the critical focus areas that directly impact business performance and success. These areas represent measurable objectives that help track progress, optimize operations, and achieve strategic goals.

1) GMV

GMV has a marginal relationship with Total Investment, meaning investment impacts it to a limited extent. As a key financial KPI, GMV closely reflects revenue and market performance.

2) Units

Units sold also show a marginal link to Total Investment, with other factors like demand playing a role. It serves as a vital metric for tracking product performance and sales trends.

Key Risk Indicators (KRI)

Key Risk Indicators (KRIs) monitor potential risks threatening an organization's objectives. Unlike KPIs, which track performance, KRIs are forward-looking and identify vulnerabilities, such as financial instability or operational disruptions. KRIs enable proactive risk management and enhance organizational resilience.

1) Cost Per Acquisition (CPA)

Cost Per Acquisition measures the average expense required to acquire a new customer, serving as a critical efficiency metric for marketing and sales operations. Calculated as:

$$CPA = \frac{Total\ Investment}{no.\ new\ customers}$$

A rising CPA signals competition or declining effectiveness, while a lower CPA suggests improved targeting and optimization. Regular monitoring helps refine strategies and maximize efficiency.

2) Order Delay

By analyzing Order Delays, we can identify inefficiencies in warehousing, dispatch, and transportation, helping to optimize logistics, enhance customer experience, and improve overall supply chain performance.

3) Net Promoter Score (NPS)

Net Promoter Score (NPS) measures customer loyalty by subtracting the percentage of detractors (0-6 ratings) from promoters (9-10 ratings), excluding passives (7-8). Ranging from -100 to 100, NPS helps businesses assess customer sentiment and identify areas for improvement. A high NPS correlates with better retention, repeat purchases, and long-term growth, making it a key metric for brand health and customer loyalty.

5. Dashboard:

The dashboard is built using React.js for a responsive and dynamic user interface and Node.js for a scalable backend infrastructure for data processing.

Features include:

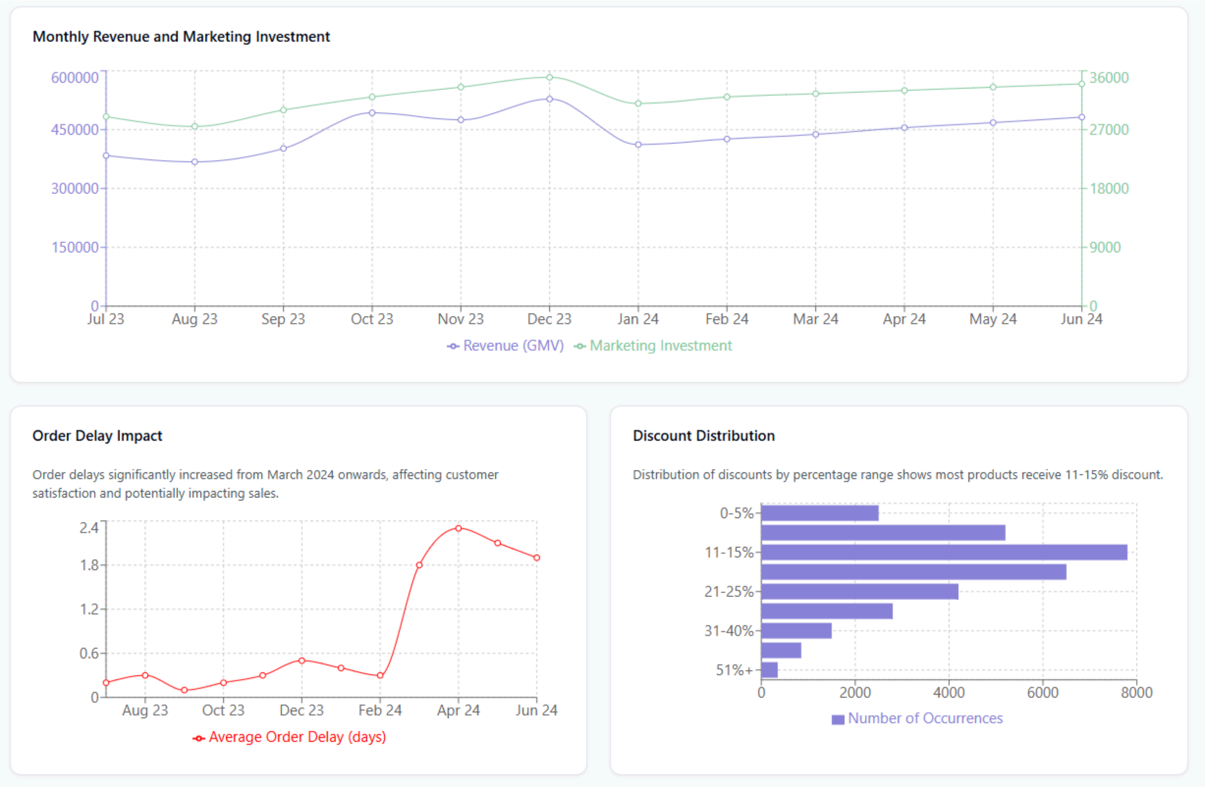
1. Performance Overview: Tracks monthly revenue, marketing investment, order delays, and discount distribution.

2. Marketing ROI Analysis: Visualizes returns on investment with response curves, channel-specific ROI, and marketing channel weights. It also allows investment details to be filtered by channel.

3. Budget Allocation: Provides an optimal budget distribution breakdown and shows the impact of different budget percentages.

4. Product Categories: Analyzes sales trends across product categories through interactive charts and delivers product insights.

5. KPI & KRI Monitoring: Visualizes key performance and risk indicators while offering detailed insights.



Monthly Revenue and Marketing Investment

Order Delay Impact

Order delays significantly increased from March 2024 onwards, affecting customer satisfaction and potentially impacting sales.

Discount Distribution

Distribution of discounts by percentage range shows most products receive 11-15% discount.

6. Regression Model and Model Pipeline:

Now that we have identified our KRAs and KPIs, we use them to find the optimal budget. Clearly, the budget that maximizes our KRAs will maximize our revenue. To this end, we have the task of finding the relation between GMV, Sales, and the various marketing channels and then maximizing it

Marketing Mix Modeling

It is a statistical analysis technique used to measure the impact of different marketing activities on business outcomes like sales, revenue, or customer engagement. It helps optimize budget allocation by identifying which marketing channels contribute the most to performance.

Why Marketing Mix Modeling?

1. It uses historical data to analyze trends and helps forecast and decide future marketing strategies.
2. It effectively identifies marketing channels that drive product sales and help forecast and plan future marketing strategies.

Key Components of Marketing Mix Modeling:

1) Dependent Variable - GMV, Units
2) Independent Variables - Marketing expenses
3) Mathematical Approach - Uses regression modeling to estimate relationships

Regression

Since the relationship between GMV and Total Investment has a diminishing nature, we are using the Michaelis-Menten model to express their relationship. The equation is given by:

$$y \ = \sum \frac{\alpha x}{1+x}$$

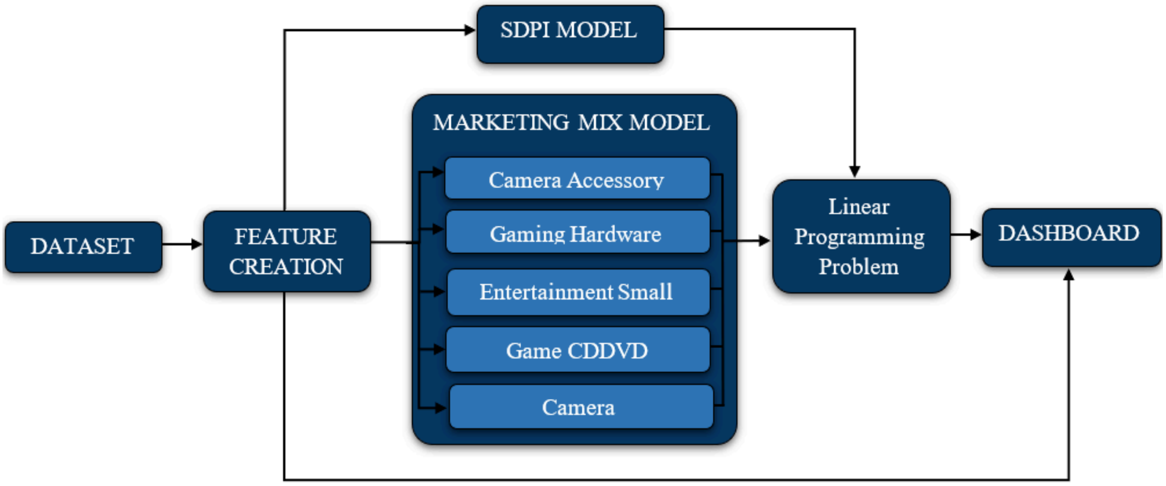
Where y is the dependent variable, x is the independent variable (input), and a (alpha) represents the maximum saturation level.

Model Pipeline

The dataset is used to create features, KPIs, and KRIs after which the dataset is then categorized into five groups for each product category.

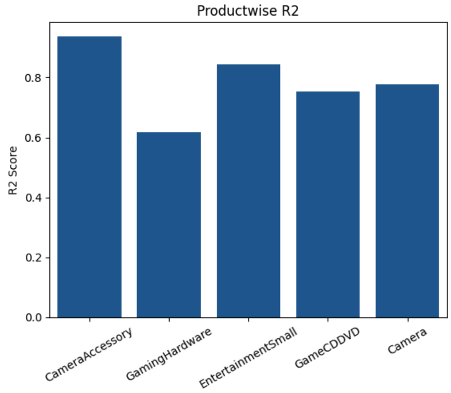
Model Overview – The extracted features are passed into five Michaelis-Menten models, each corresponding to a specific category. The computed coefficients from these models are then fed into a Linear Programming Problem (LPP) solver with their product weights and SDPI to get the optimized model.

All the features of importance created during this process are then converted to JSON objects and fed to the Dashboard website to be presented as shown below



Results

After training the marketing mix model we get R2 scores of each product category with the highest being 0.93 of Camera Accessories for GMV and 0.87 for Entertainment Small on Sales



7. Conclusion:

We conducted extensive Exploratory Data Analysis (EDA), including univariate and multivariate analysis which examined distributions, and performed statistical tests and hypothesis testing to identify key performance indicators (KPIs), key risk indicators (KRIs), and key result areas (KRAs). Following this, we applied the Michaelis-Menten model to estimate the impact of individual investments—such as TV, Sponsorships, and affiliates—on Gross Merchandise Value (GMV) and Units. Using this model, we optimized the marketing budget allocation to maximize returns. To make these insights accessible and actionable, we built an interactive dashboard enabling stakeholders to explore the results and make data-driven decisions dynamically.

Annexure

8.1) Additional Helpful Information

For each marketing channel more information on how each of them contribute exactly to the KPIs would have been helpful such as:

1. TV Marketing – Gross Rating Points (GRPs): The basic measurement unit that quantifies ad impressions as a percentage of a target market. Studies track the rate of delivery of a television commercial's sales effectiveness per GRP.
2. Digital Marketing – Engagement Metrics: Click-through rates (CTR), bounce rates, and exit rates that indicate content relevance.
3. Sponsorship Marketing – Audience Reach Data: Raw numbers of people exposed to a brand through sponsored events across TV broadcasts, social media, and live attendance.
4. Content Marketing – Site Interaction Data: How users engage with different content types, including time spent on page and scroll depth.
5. Online Marketing (Cross-Channel) – Budget Analytics: Cost per visitor, revenue per visitor, and other financial metrics across online channels.
6. Affiliate Marketing – Raw Clicks: The total number of times an affiliate link is clicked, providing a fundamental count of user interactions.

8.2) Alternative approach

Instead of using correlations in the data to optimize the budget we can forecast all the relevant columns to create synthetic data and use reinforcement learning to adjust budgets dynamically.

1) Forecasting: Use Deep Learning methods such as LSTM Neural Networks to forecast relevant features which are driven by marketing levers. Conduct a Hypothesis to validate the impact of each marketing channel. Consider Instrumental Variables to filter out noise from confounding factors like seasonality, holidays, and weather.
2) Reinforcement Learning for Dynamic Budget Allocation: Using AI-driven budget optimization with Deep Q-Networks, adjusting spending in the forecasted data on what works best. The goal? Maximize ROI while staying within budget and avoiding overspending on less effective channels

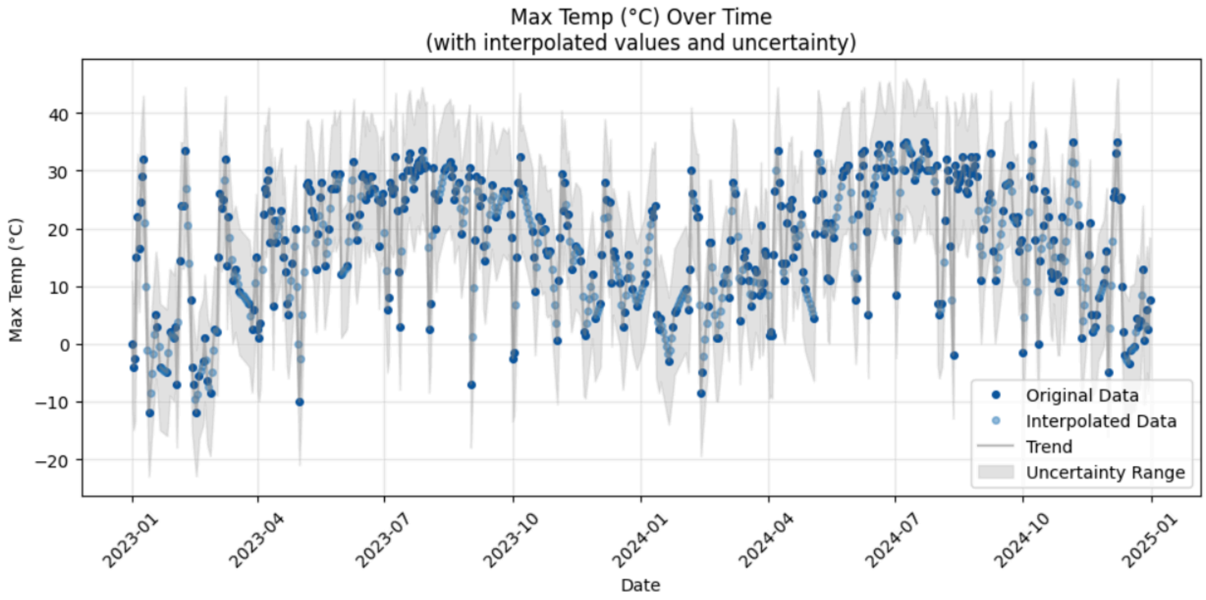
8.3) Interpolation of the Weather Data

From the combined weather dataset, an analysis of missing values revealed a significant distinction between rows with NaN (missing) values and those with complete data. By performing a count of NaN rows, it was observed that specific weather attributes had missing entries, potentially due to sensor failures or incomplete data collection. The number of fully populated (non-NaN) rows significantly outweighed the missing ones, indicating a generally well-maintained dataset with minor gaps.

The provided data, which consisted of weather, had missing values and missing datetime rows, so we must compute and impute the missing values. We used interpolation with a linear method for small gaps (limit = 3 days), and for longer days, we used time interpolation and also added some random noise to avoid perfect linearity.

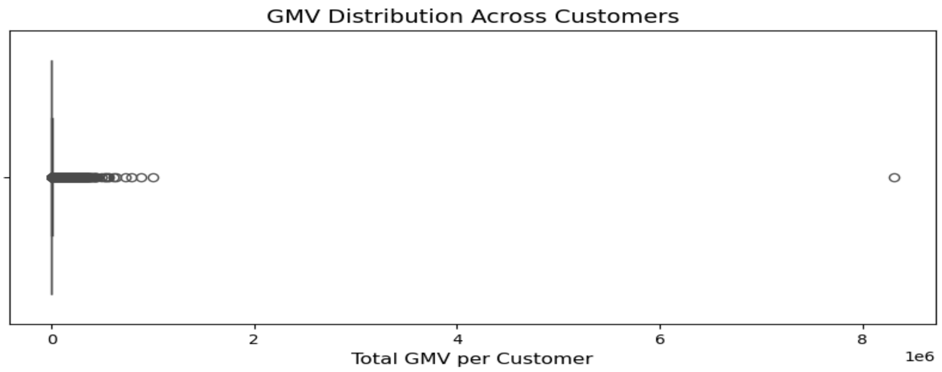
In the case of precipitation analysis, linear interpolation is applied over a span of 2 days. For a longer timeframe, we used a rolling median for the same week for different years.

Heat Degree Days (HDD) and Cool Degree Days (CDD) are calculated based on the difference between the base temperature (18°C) and the mean temperature. HDD is the difference when the mean temperature is below the base temperature, indicating heating needs, and is set to 0 if the mean temperature is higher. CDD is the difference when the mean temperature is above the base temperature, indicating cooling needs, and is set to 0 if the mean temperature is lower. These calculations are applied to rows with missing HDD or CDD values by subtracting the mean temperature from the base temperature or vice versa, depending on the condition.



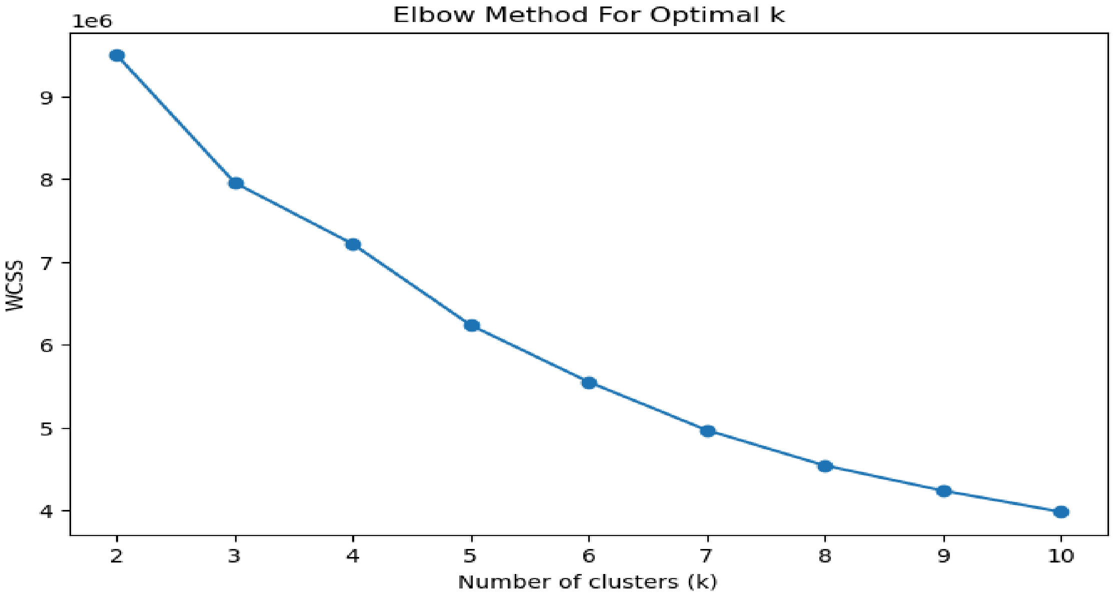
8.4) Clustering

We applied a clustering algorithm on the daily dataset to determine whether we can divide the customers into clusters to find customer categories, such as customers more dominated by gaming products, cameras, etc. For this, there is an anomaly in the data, which we identified from the following graph:



From the above boxplot, we can infer that most of the GMV (Gross Merchandise Value) per customer is concentrated on the lower end, while extreme values on the higher end lead to a long right tail. The plot shows multiple individual points far from the main distribution, indicating that some customers have significantly higher GMV than the rest. One extreme outlier appears around 8 million GMV. For this reason, after keeping this particular customer aside, we plotted the graph between WCSS (Within-Cluster Sum of Squares).

We used the Elbow Method plot to determine the optimal number of clusters (K) in K-Means clustering by analyzing the Within-Cluster Sum of Squares (WCSS). The plot shows how WCSS decreases as K increases.



The graph does not show a clear elbow point, as the WCSS decreases gradually without a sharp drop at any specific k value. This suggests that the data may not have a well-defined clustering structure. Since the optimal number of clusters is ambiguous, selecting k confidently is difficult. The results indicate that K-Means may not be the best approach, and other techniques like the silhouette score or hierarchical clustering should be explored for better insights