Analyzing and Identifying Attributes of Apparel Products and their Effect on Sales

A Mid-Term report for the BDM capstone Project

Submitted by

Name: Tanmay Garg
Roll number: 22f2001630

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai Tamil Nadu, India, 600036

Contents

1 Executive Summary

V Mart, founded in 2002, is a rapidly growing value retail chain in India, catering to middle-class and lower-middle-class customers across Tier II and Tier III cities. The company specializes in affordable apparel and home products, operating 400+ stores nationwide. As the retail landscape becomes increasingly competitive, understanding the factors that drive sales performance is critical to sustaining growth and optimizing inventory management.

This study aims to identify the key product attributes—such as Apparel Group, Quantity, Promotions, and Discount—that significantly influence apparel sales. Currently, V Mart lacks a data-driven approach to assess how these attributes impact consumer preferences, leading to inventory inefficiencies and missed sales opportunities. External challenges like shifting fashion trends and increasing competition further complicate this issue. By leveraging statistical analysis, this research will uncover actionable insights to improve sales strategies, inventory planning, and customer satisfaction.

To address these challenges, this report follows a structured approach: Section 2 establishes proof of originality, ensuring that the dataset and analysis are unique and derived from credible sources. Section 3 outlines the metadata, providing a detailed breakdown of the dataset, including its source, structure, and attributes. Each attribute, such as Apparel Group, Quantity, Promotions, and Discount, plays a crucial role in determining sales performance. The section also highlights data types, missing values, and any preprocessing steps applied to standardize the dataset. Given that the data is not in an Excel format, a CSV of the dataset is provided to help reviewers link the metadata to the actual dataset, with a screen shot of meta data .

Section 4 presents descriptive statistics, including key metrics such as mean, median, standard deviation, and distribution analysis of variables like revenue, quantity sold, and pricing factors. This section provides an overview of sales trends and identifies potential outliers or patterns in the dataset. Section 5 delves into data collection, cleaning, and preprocessing techniques, ensuring that the dataset is free from inconsistencies or biases. It also explains the statistical and machine learning methods used for quantitative analysis, justifying why these approaches were chosen for evaluating attribute importance in driving sales.

Finally, Section 6 presents key findings based on the implemented methods, summarizing insights that can guide strategic decision-making at V Mart. This includes data-driven recommendations on inventory optimization, pricing strategies, and customer targeting. The section focuses on visual representation through charts and graphs to clearly communicate insights. By following this structured approach, the study ensures alignment with the BDM project expectations while providing actionable recommendations to enhance V Mart's sales performance and competitive positioning.

2 Proof of Originality

- Data Set Link:
  - parquet files as provided by organization: Link
  - CSV link for easy view: Link
- Letter from the organization: Link (Authorization Letter Link)
- Images:

- Video: Link

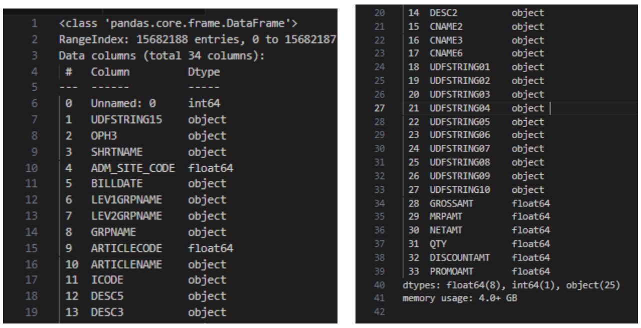
3 Metadata

- **CSV link for easy view**: Link
- **Data Source:** Billing Counters, generate data for every item a customer buys, billing counters generate a row of data for every bill that is generated. For online order a row data is generated after the user places its order.
- **Types of Attributes**: Region (Location), Customer Type, SKU attributes, Sales
- **Data file type:** parquet
- **Number of columns:** 32, Number of Rows: 15,682,188
- **Time period:** 3 months (May, Jue, July)
- **Tools Used for analysis:** Python, Google Sheets
- **Mentor:** Mr Arun Kumar Gaur, Mr. Akash Sharma
- **Attributes have been bucketed into 5 major categories as:**
    - **Region Attributes and Their Explanation:**
        - UDFSTRING15 (Data Type: String/Text) : Represents the region in reference to sales, including the digital region for catering to online sales presence.
        - OPH3 (Data Type: String/Text) : Indicates the sub-region within the broader region classification.
        - SHRTNAME (Data Type: String/Text) : Specifies the exact store location.
        - ADM_SITE_CODE (Data Type: Integer) : Unique code assigned to each store for administrative identification.
    - Customer Type Attributes and Their Explanation:

    - LEV1GRPNAME (Data Type: String/Text) : Categorizes customers based on gender and age group.
    - LEV2GRPNAME (Data Type: String/Text) : Subdivides the LEV1GRPNAME category into more specific classifications.

  - SKU Attributes and Their Explanation:
    - GRPNAME (Data Type: String/Text) : Represents the primary grouping of items.
    - ARTICLECODE (Data Type: Integer) : Unique identifier for each article.
    - ARTICLENAME (Data Type: String/Text) : The name assigned to each article.
    - ICODE (Data Type: Integer) : Unique code assigned to each SKU (Stock Keeping Unit).
    - CNAME2 (Data Type: String/Text) : Indicates the vendor or brand associated with the product.
    - CNAME3 (Data Type: String/Text) : Describes the style or pattern of the product.
    - CNAME6 (Data Type: String/Text) : Represents the color of the product.
    - DESC3 (Data Type: String/Text) : Details the material used in the product.
    - UDFSTRING01 (Data Type: String/Text): Specifies the neck style of the garment.
    - UDFSTRING02 (Data Type: String/Text) : Describes the fit of the garment.
    - UDFSTRING03 (Data Type: String/Text) : Indicates the sleeve length.
    - UDFSTRING04 (Data Type: String/Text) : Details the sleeve styling.
    - UDFSTRING05 (Data Type: String/Text) : Defines the type of garment.
    - UDFSTRING06 (Data Type: String/Text) : Highlights the trend or story associated with the garment.
    - UDFSTRING07 (Data Type: String/Text) : Explains the pattern coverage.
    - UDFSTRING08 (Data Type: String/Text) : Specifies the garment's length.
    - UDFSTRING09 (Data Type: String/Text) : Describes the shape of the garment.
    - UDFSTRING10 (Data Type: String/Text) : Indicates the wash type applied to the garment.

  - **Sales Attributes:**
    - GROSSAMT (Data Type: Float) : The gross amount of sales recorded.
    - MRPAMT (Data Type: Float) : Maximum Retail Price of the sold items.
    - NETAMT (Data Type: Float) : Net sales value after applicable deductions.
    - QTY (Data Type: Integer) : Quantity of items sold.
    - DISCOUNTAMT (Data Type: Float) : Total discount amount applied during the sale.
    - PROMOAMT (Data Type: Float) : Promotional amount or discounts offered.
  - Other Attributes and Their Explanation:
    - BILLDATE (Data Type: Date Type) : Date of bill generation for each transaction.
- **Image of the same**



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15682188 entries, 0 to 15682187
Data columns (total 34 columns):
 #   Column         Dtype
---  ------         -----
 0   Unnamed: 0     int64
 1   UDFSTRING15    object
 2   OPH3           object
 3   SHRTNAME       object
 4   ADM_SITE_CODE  float64
 5   BILLDATE       object
 6   LEV1GRPNAME    object
 7   LEV2GRPNAME    object
 8   GRPNAME        object
 9   ARTICLECODE    float64
 10  ARTICLENAME    object
 11  ICODE          object
 12  DESC5          object
 13  DESC3          object
 14  DESC2          object
 15  CNAME2         object
 16  CNAME3         object
 17  CNAME6         object
 18  UDFSTRING01    object
 19  UDFSTRING02    object
 20  UDFSTRING03    object
 21  UDFSTRING04    object
 22  UDFSTRING05    object
 23  UDFSTRING06    object
 24  UDFSTRING07    object
 25  UDFSTRING08    object
 26  UDFSTRING09    object
 27  UDFSTRING10    object
 28  GROSSAMT       float64
 29  MRPAMT         float64
 30  NETAMT         float64
 31  QTY            float64
 32  DISCOUNTAMT    float64
 33  PROMOAMT       float64
dtypes: float64(8), int64(1), object(25)
memory usage: 4.0+ GB
```

Jupyter output of dataFrame.info(), showing metadata

4 Descriptive Statistics

4.1 Summary Statistics

The dataset consists of 15,682,188 transactions with key financial attributes such as Gross Amount (GROSSAMT), Net Amount (NETAMT), Quantity (QTY), and Discount Amount (DISCOUNTAMT). The following table provides a summary of these attributes:

Statistic	**GROSSAMT**	**NETAMT**	**QTY**	**DISCOUNTAMT**
Count	15,682,188	15,682,188	15,682,188	15,682,188
Mean	₹ 328.22	₹ 318.81	0.96	₹ 9.41
Std Dev	₹ 321.62	₹ 296.43	0.34	₹ 81.64
Min	₹ -6,999.50	₹ -6,999.50	-9.00	₹ -2,161.57
25%	169.00	169.00	1.00	0.00
50%	249.00	249.00	1.00	0.00
75%	449.00	449.00	1.00	0.00
Max	₹ 17,999.00	₹ 13,000.00	25.00	₹ 8,500.00

Descriptive Statistics of Sales Transactions

Sales performance, including Gross Amount (GROSSAMT), Net Amount (NETAMT), Quantity Sold (QTY), and Discount Amount (DISCOUNTAMT). With 15,682,188 transactions, the data reveals the following trends:

- Average Sale: The mean gross amount per transaction is ₹328.22, while the net amount averages ₹318.81, indicating a slight reduction due to discounts.
- Discount Trends: The mean discount per transaction is ₹9.41, though the high standard deviation (₹81.64) suggests large variations in applied discounts.
- Quantity Sold: The average quantity per transaction is 0.96, with most transactions involving a single unit (median of 1).
- Data Distribution: The wide range in values (e.g., Gross Amount from ₹-6,999.50 to ₹17,999.00) highlights diverse transaction sizes and potential returns or refunds. The negative values suggest transactions recording returns made at the store.
- Skewness in Discounts: With a 0 median and high max value (₹8,500.00), most transactions receive no discount, but some have significant reductions.

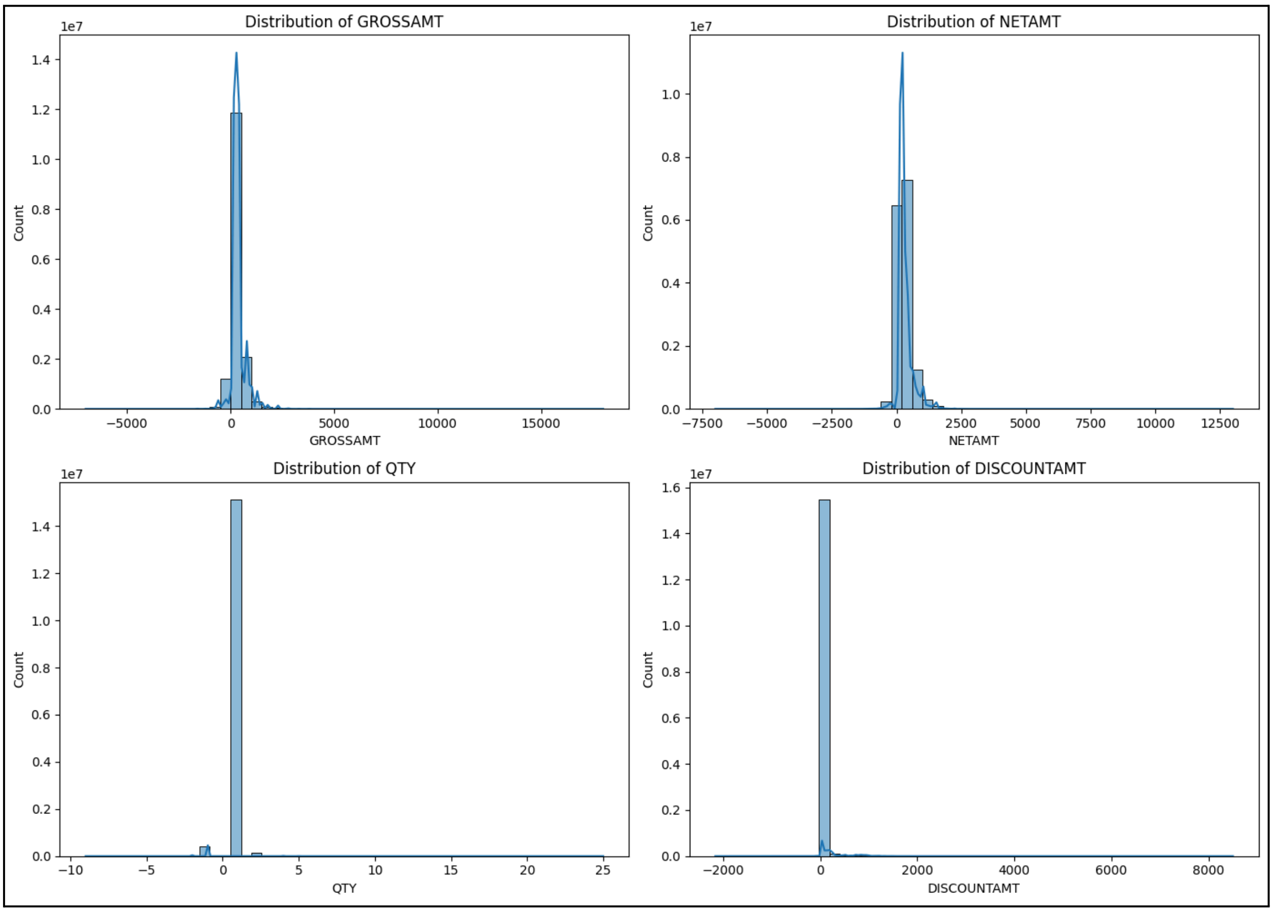
4.2 Distribution Analysis

This section displays the distributions of Gross Amount (GROSSAMT), Net Amount (NETAMT), Quantity Sold (QTY), and Discount Amount (DISCOUNTAMT) across transactions. Key observations:

- **GROSSAMT & NETAMT Distributions**
  - Both distributions are highly skewed, with most transactions clustered around lower values.

    - A few extreme values (both positive and negative) indicate potential refunds or high-value sales.
- **QTY Distribution**
    - Majority of transactions involve a quantity of 1, with very few cases exceeding this.
    - Some negative values suggest product returns.
- **DISCOUNTAMT Distribution**
    - Most transactions receive little to no discount, but some have substantial discounts.
    - The presence of negative discount values suggests possible adjustments or promotional corrections.
- **Skewness**
    - GROSSAMT (2.08) and NETAMT (1.50) indicate right-skewed distributions, suggesting a presence of extreme high values.
    - QTY (-4.40) indicates left-skewness, meaning there are a few extreme low values (negative quantities may indicate returns).
- **Kurtosis**
    - GROSSAMT (18.23) and NETAMT (14.26) suggest heavy-tailed distributions with significant outliers.
    - QTY (34.26) has an extremely high kurtosis, indicating a distribution with sharp peaks and heavy tails.

Overall, the distributions indicate a strong concentration of sales around lower amounts and quantities, with occasional high-value transactions and adjustments.



Distribution of Sales-Related Metrics

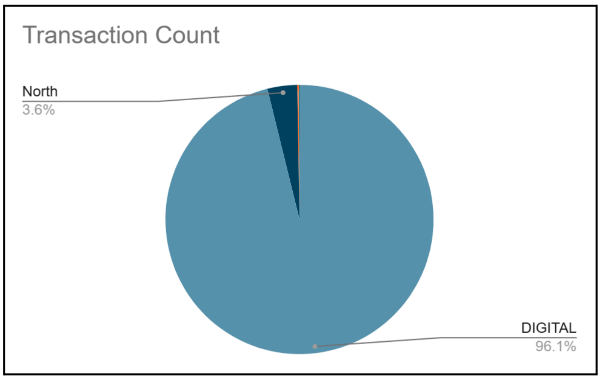
4.3 Regional Distribution

The dataset includes four main regions:

Region	**Transaction Count**
DIGITAL	13,176,621
North	491,404
Bihar/JH	22,546
South	14,433

Transaction Count by Region

The dataset consists of transaction counts from four main regions: DIGITAL, North, Bihar/JH, and South. The DIGITAL region dominates with 96.1% of total transactions, significantly surpassing physical locations. The North region follows with 3.6%, while Bihar/JH and South contribute a minimal share.



Transaction Count Distribution by Region

The pie chart visually emphasizes this dominance, with the DIGITAL segment overwhelmingly larger, reinforcing the importance of digital sales channels. The table provides exact transaction counts, highlighting the stark difference in distribution.

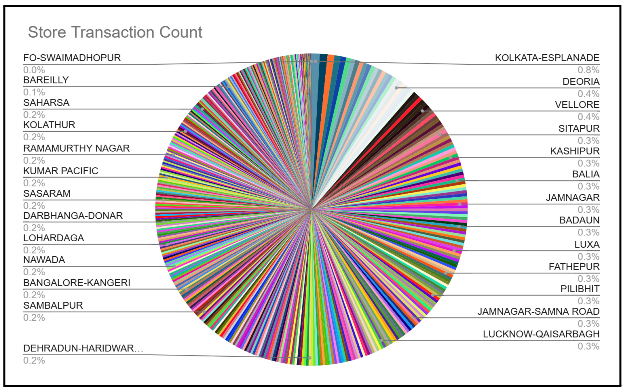
4.4 Store-wise Distribution

The dataset contains transactions from 515 unique stores. The top 5 stores with the highest transaction counts are:The dataset includes transactions from 515 unique stores. The top 5 stores with the highest transaction counts are Kolkata-Esplanade, Srinagar, Valsaravakkam Big MM, Birsa Chowk-Ranchi, and Bhubaneswar-Patia, with Kolkata-Esplanade leading at 132,920 transactions.

Store	**Transaction Count**
KOLKATA-ESPLANADE	132,920
SRINAGAR	125,975
VALASARAWAKKAM_BIG MM	112,144
BIRSA CHOWK-RANCHI	99,874
BHUBANESWAR-PATIA	92,151

Top 5 Stores by Transaction Count

The pie chart illustrates the distribution of transactions across all stores, showing the concentration of sales among a few high-performing stores while the rest are more evenly distributed. Smaller franchise outlets, such as FO-Sivamadehopur, recorded the lowest transactions at just 70.



Transaction Count Distribution by Region

4.5 Product Category Analysis

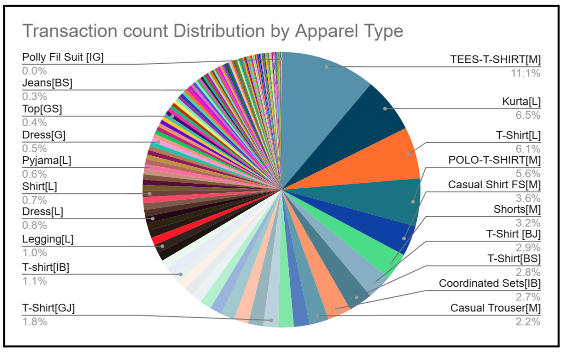
The dataset includes **255 unique product groups (GRPNAME)**. The top 5 product groups with the highest sales transactions are:

Product Group	Count
TEES-T-SHIRT[M]	1,740,826
Kurta[L]	1,020,221
T-Shirt[L]	949,919
POLO-T-SHIRT[M]	879,268
Casual Shirt FS[M]	562,783

Top 5 Product Groups by Transaction Count

T-Shirts (Multiple Variants): T-shirts are the leading category, with TEES-T-SHIRT[M] and T-Shirt[L] together accounting for nearly 2.7 million transactions. This indicates their popularity across different sizes and demographics. Kurta [L]: Traditional wear also holds a significant share, with 1,020,221 transactions, suggesting strong demand for ethnic and fusion wear. Casual Shirts & Polo T-Shirts: These categories rank high in transactions, likely appealing to both casual and semi-formal consumers.

The pie chart visualizes the distribution of transactions by apparel type, highlighting the dominance of T-shirts, Kurtas, Polo T-shirts, and Casual Shirts in the dataset. Other product categories contribute to a more fragmented portion of the sales, indicating a diverse product mix.



Transaction Count Distribution by Apparel Type

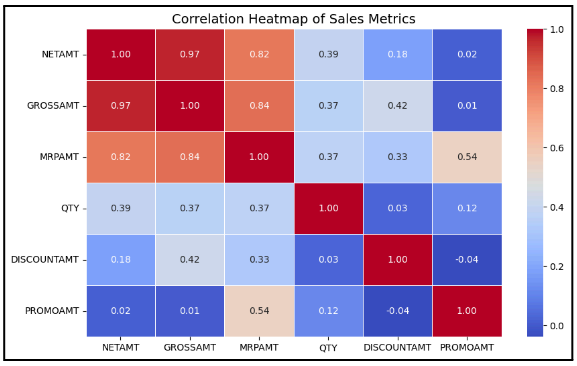
4.6 Missing Values

There are no missing value, notably. Over all categorical columns, have missing values but those in significant in analysis when using numerical columns such as Gross Amount (GROSSAMT), Net Amount (NETAMT), Quantity Sold (QTY), and Discount Amount (DISCOUNTAMT) for analysis of sales, which have no missing values. Categorical columns containing missing values:

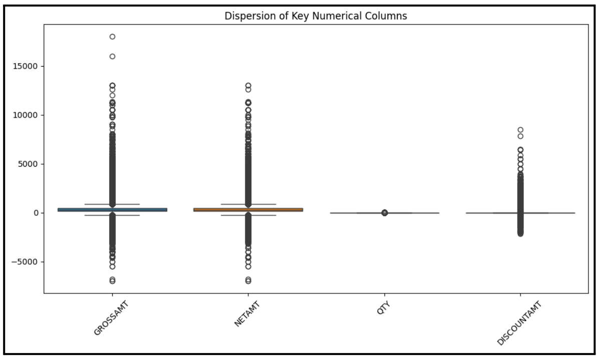
- UDFSTRING15 (Region) has 1,977,184 missing values.
- UDFSTRING01 to UDFSTRING10 have significant missing values, ranging from 3.25M to 11.18M.
- DESC3 (Product Description) has 59,712 missing values.

4.7 Observations & Insights

- The dataset has significant variability in GROSSAMT and NETAMT, with extreme high values suggesting the presence of high-value transactions.
- Negative values in QTY, GROSSAMT, and NETAMT indicate refunds.
- The majority of transactions (84%) are from DIGITAL channels, while other regions contribute relatively less.
- Gross Amount and Net Amount are key revenue metrics, with a large spread, refer fig., and strong correlation, refer fig. . High-value transactions significantly impact revenue, and net revenue calculations closely follow gross revenue trends.
- Discounting boosts sales, but excessive discounts can reduce profitability. Moderate correlation between DISCOUNTAMT and GROSSAMT , refer fig., suggests higher discounts boost sales, but outliers indicate high discounts, refer fig. . Customer segmentation can balance growth and profit margins.
- Promotional spending significantly influences market price variations, with a moderate correlation, refer fig., between PROMOAMT and MRPAMT. However, not all promotions lead to direct sales increases.



Correlation Heatmap of Sales Metrics



Dispersion of Key Numerical Columns (Boxplot)

5 Detailed Explanation of Analysis Process/Method

Vmart multi brand outlet, needs to evaluate it's business, on daily, weekly, monthly basis, based on certain parameters Which, should include time based comparison, of merchandise selling across various sub groups, price points, net sale value, gross sale value, promotions and offers.

- Retails companies Challenge and Opportunities:
  - ForeCasting Peak sale weeks/months.
  - Optimize Inventory Vs Sales.
  - Analysis of Fast moving/Slow Moving Sub Categories.
  - Identification of top revenue categories.
  - Identification of top margin contributor categories.
  - Planning promotion and consumer offer.
- Addressing the above:
  - Need a clear view through data of Peak sale weeks/months.
  - Analysis should be able to project the merchandise and time related challenges and opportunities.
- Help provided:
  - Pre alert to the product manufacturing and sourcing team.
  - Capacity planning for logistics and time lines.
  - Creating agile warehousing and prioritising movement of goods as per critical timelines.

5.1 Data Collection & Preprocessing

- The dataset was sourced from billing counters of the stores.
- The raw data underwent preprocessing steps, including:
  - Data Loading & Merging:
    - Multiple Parquet files containing transaction data were read and merged into a single DataFrame.
    - Data was then split into monthly datasets (May, June, and July) for focused analysis.
  - Datetime Conversion:
    - The "BILLDATE" column was converted to datetime format to facilitate time-based analysis.
    - This allowed accurate filtering and aggregation of sales data by month.
  - Handling Missing & Invalid Data:
    - Identified and checked for missing values in key columns like GROSSAMT, QTY, NETAMT, PROMOAMT, and DISCOUNTAMT.
    - Detected and flagged negative values in financial and quantity-related columns for further investigation.
  - Grouping & Aggregation:
    - Sales data was grouped by category levels (GRPNAME, LEV1GRPNAME) to analyze performance trends.
    - Aggregated sales metrics such as total quantity sold, revenue, and discounts for each category.
  - Exporting Clean Data:
    - Processed data was saved into separate CSV files for May, June, and July to support detailed month-wise analysis.

5.2 Analysis Methods Used

This enables business owners to optimize the cash flow, inventory planning and profitability of the company. To look in detail these parameters, we various tools like:

To understand the sales trends and their influencing factors, the following analytical techniques were applied:

- Month-on-Month (MoM) Sales Comparison
  - Evaluated sales trends across different months to identify fluctuations.
  - Analyzed the impact of promotions and discounts on monthly sales.

- Same-Week MoM Analysis (Seasonality Check)
  - Compared sales data for the same week across multiple months.
  - Identified recurring seasonal patterns affecting merchandise performance.

- Merchandise Sub-Category Sales Analysis
  - Assessed the contribution of each sub-category to total sales.
  - Tracked month-by-month variations (May, June, July) in sub-category sales to observe demand shifts.

6 Results and Findings

Data set given is at SKU attribute level, with their sales numbers(Qty, MRP, Discount, Promo). The following are analysis processes used:

Process:

1) **MOM Sales: May, June, July (overall SKUs). Along with Promotion Discount.**

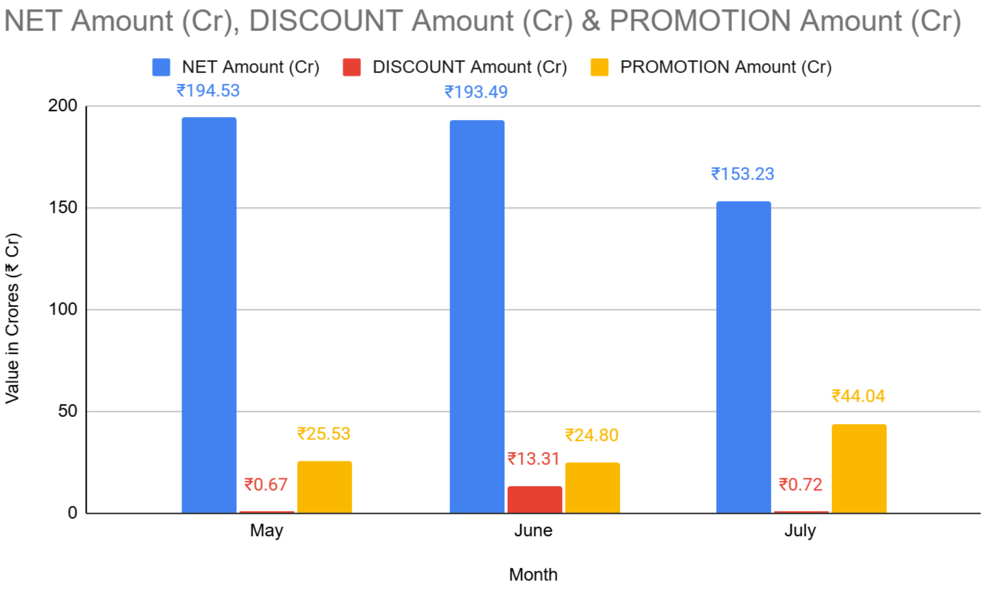


Chart shows us June is the month where heavy customer discounting was done to boost the sales. This is compromising the margins.

2) **Same Week MOM Sales: Each Month for comparison of NET Sales, Discount, Promotions**

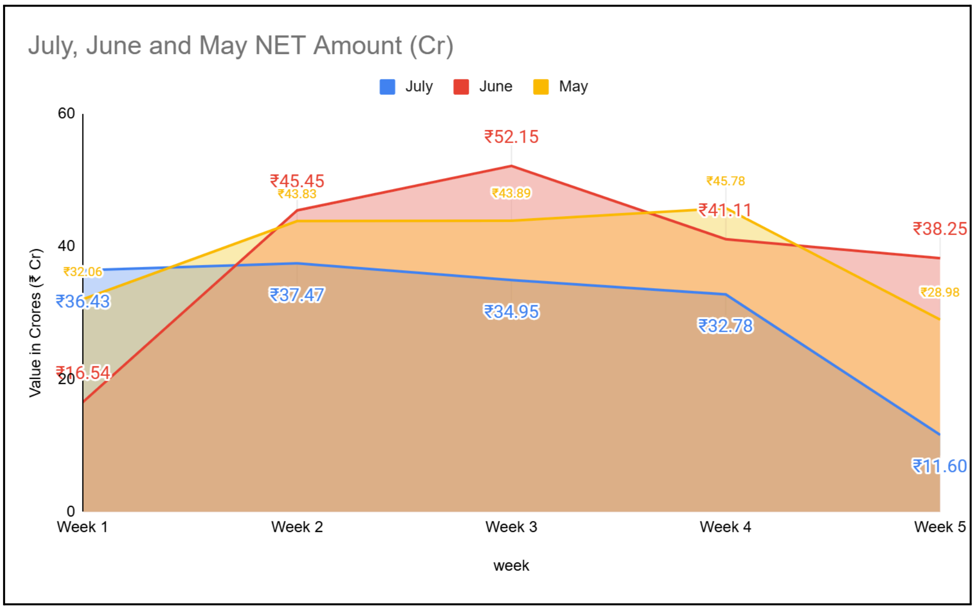
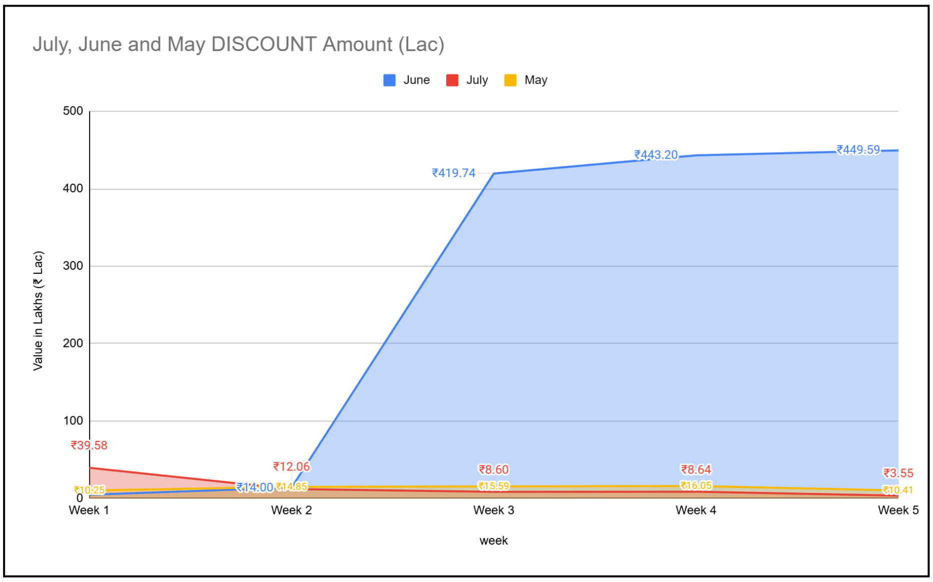
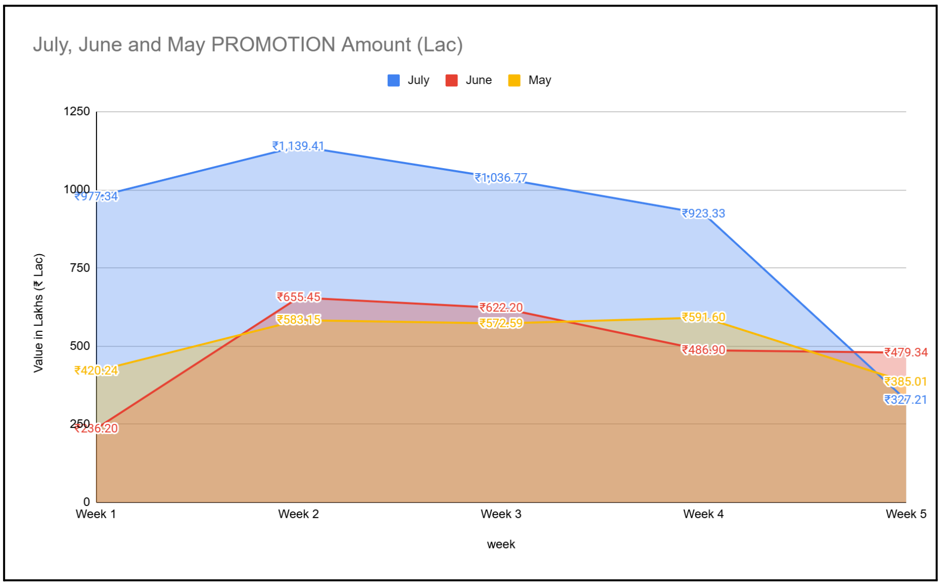


Chart shows July sales starting good plateauing towards the end of month week June sales Start slow from first week and picks up as discounting increases. May sales are normal month sales.

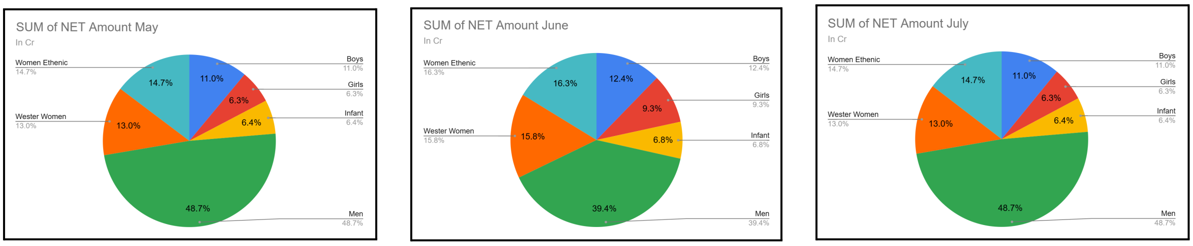


May and July Month shows less discounting were as June month from second week onwards till end heavy discounting go on.



Good Values of customer promotion running in the month of July. May and June are normal customer promotional months.

3) Merchandise Sub Category Sales: Percentage:



(a) (b) (c)

Comparison of May(a), June(b), July(c) sales mix sub category level, shows following trends. Mens sales percent dip in June, but picks up heavily in July on support discounts and promotion. Women Ethnic is constantly having no impact on promotion and discounts. Maybe not running. Same can be seen in Women Western.

7 Analysis sheet/code:

- **Jupyter Notebooks:** Link (all files in ipynb format)
- **Google Sheets:** Link1, Link2