ZARA

Maximising Gross Margin through Data-Driven Retail Strategy

Bhushan Borse

NEXTLEAP | TOP FELLOW

✆ **9822168481**

✉ **bhushansb.career@gmail.com**

Table of Contents

Python Code File

Tableau Dashboard

1 Introduction:

Zara is one of the world's largest and most popular fashion brands. With stores across many countries, Zara is known for fast fashion, trendy designs, and efficient supply chains. As the company continues to grow, it now wants to shift focus from just increasing sales to improving profitability.

Zara's leadership has identified **Gross Margin** as a key business metric.

Gross Margin is the **difference** between how much **revenue** Zara earns and how much it **spends** to produce the products. Hence, by improving Gross Margin, Zara can keep its prices affordable while still earning healthy profits.
This study aims to analyse Zara's global retail dataset and leverage data-driven insights that will help Zara maximise its Gross Margin across different markets, product lines, and customer segments.

2 Objective:

The goal is to conduct a comprehensive analysis of Zara's global retail dataset, identifying key factors impacting Gross Margin for Zara and providing actionable insights to maximize the Gross Margin, enabling Zara to keep increasing sales with affordable prices while also earning healthy profits and improving the profitability.

This objective serves as the foundation for data exploration, hypothesis validation, and strategy formulation to enhance operations and drive business success for Zara.

3 Business Impact:

The business impact of this analysis can be highly significant, enabling improvements in pricing, product categories and customer engagement which in turns can **enhance** the **Gross Margin**.

This analysis can help Zara in –

3.1 Optimizing Pricing Strategy for Increased Profitability

By analysing how different product categories, customer segments, and discount strategies affect Gross Margin, Zara can modify its pricing strategy to maximize profitability while maintaining its competitive edge.

3.2 Improving Product Type/Categories

Identifying which product types and categories generate the highest margins will allow Zara to focus on high-value items, streamline production, and allocate resources more effectively, ensuring maximum returns on investment.

3.3 Targeted Marketing and Customer Engagement

Understanding which customer segments contribute most to Gross Margin will enable Zara to tailor its marketing efforts, personalize promotions, and enhance customer retention strategies, ultimately driving higher sales and profitability.

3.4 Improving Supply Chain Management

By analysing trends across time, regions, and stores, Zara can develop more accurate supply chain efficiency, allowing better production planning, reduced waste, and optimized inventory levels all of which will contribute to a healthier Gross Margin.

3.5 Financial Growth

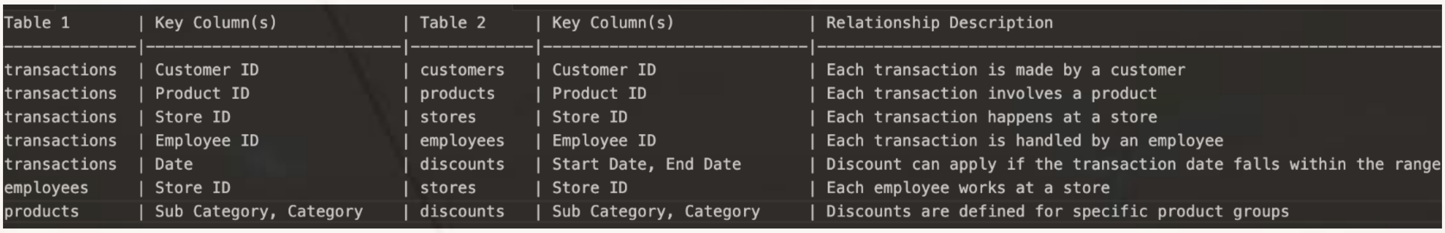
Improving Gross Margin will enhance the profitability making Zara to financially grow at a faster rate enabling reinvestment in innovation and expansion into new markets.

4 Dataset Overview:

- **Dataset Name**: ZARA Global Retail Dataset
- **Number of Rows**: 6414328
- **Number of Columns**: 27
- **Description**: This dataset captures key metrics related to retail transactions and store operations for Zara. Each row represents details of a transaction with relevant attributes.

The dataset is organised in 6 separate tables, one each for customers, discounts, employees, products, stores and transactions.

Each table is related to other as given below:



```
Table 1       | Key Column(s)             | Table 2     | Key Column(s)             | Relationship Description
--------------|---------------------------|-------------|---------------------------|------------------------------------------------------------------
transactions  | Customer ID               | customers   | Customer ID               | Each transaction is made by a customer
transactions  | Product ID                | products    | Product ID                | Each transaction involves a product
transactions  | Store ID                  | stores      | Store ID                  | Each transaction happens at a store
transactions  | Employee ID               | employees   | Employee ID               | Each transaction is handled by an employee
transactions  | Date                      | discounts   | Start Date, End Date      | Discount can apply if the transaction date falls within the range
employees     | Store ID                  | stores      | Store ID                  | Each employee works at a store
products      | Sub Category, Category    | discounts   | Sub Category, Category    | Discounts are defined for specific product groups
```

5 Column Definitions:

5.1 Customers Dataset

1. **CustomerID:** Unique number assigned to each customer.
2. **Name**: Full name of the customer. May include titles like "Mr." or job-related suffixes.
3. **Email**: Anonymized email address using fake domains (like fake_gmail.com).
4. **Telephone:** Customer's phone number.
5. **City**: City where the customer is located.
6. **Country**: Country where the customer resides.
7. **Gender**: Customer's gender. Values can be F (Female), M (Male), or D (Diverse).
8. **DateOfBirth**: Customer's date of birth in YYYY-MM-DD format.
9. **JobTitle**: Customer's occupation.

5.2 Discounts Dataset

1. **Start Date:** The date when the discount campaign begins, in YYYY-MM-DD format.
2. **End Date:** The date when the discount campaign ends, in YYYY-MM-DD format.
3. **Discount**: Decimal value representing the discount rate. For example, 0.20 means a 20% discount.
4. **Description:** A short text describing the purpose or theme of the discount campaign.
5. **Discount Percentage:** Another column showing the same discount rate as a decimal.
6. **Category**: The main product category to which the discount applies.
7. **Sub Category:** A more specific product sub-category under the main category where the discount applies.

5.3 Employee Dataset

1. **Employee ID:** A unique number assigned to each employee.

2. **Store ID**: Refers to the store where employee works. This links Store ID column in the stores dataset.
3. **Name**: Full name of the employee in the format: First Name followed by Last Name.
4. **Position:** The role of the employee in the store.

5.4 Products Dataset

1. **Product ID:** Unique number assigned to each product.
2. **Category:** Main classification of the product.
3. **Sub Category**: More specific product category within the main category.
4. **Description PT:** Product description written in Portuguese.
5. **Description DE:** Product description written in German.
6. **Description FR**: Product description written in French.
7. **Description ES:** Product description written in Spanish.
8. **Description EN:** Product description written in English.
9. **Description ZH:** Product description written in Chinese.
10. **Color:** The color of the product.
11. **Sizes:** Available sizes for the product, separated by pipes (|).
12. **Production Cost:** Cost in USD to produce one unit of the product.

5.5 Stores Dataset

1. **Store ID:** Unique number assigned to each store location.
2. **Country:** Country where the store is located.
3. **City:** City where the store operates.
4. **Store Name:** Name of the store, typically shown as "Store [City]".
5. **Number of Employees:** Total number of employees working at the store.
6. **ZIP Code:** Postal or ZIP code of the store's location.
7. **Latitude:** Geographical latitude of the store's position.
8. **Longitude:** Geographical longitude of the store's position.

5.6 Transactions Dataset

1. **Invoice ID:** A unique code for each transaction. It shows if it's a sale (INV) or return (RET), along with the country code, store ID, and a serial number. All line items in the same invoice share this ID.
2. **Line:** The position of the product in the invoice. One invoice may have multiple line items.
3. **Customer ID:** ID of the customer who made the purchase. Refers to the customers dataset.
4. **Product ID:** ID of the product that was purchased. Refers to the products dataset.
5. **Size:** Size of the product purchased (like S, M, L, XL). May be blank if not applicable.
6. **Color:** Color of the product purchased. May be blank if not applicable.
7. **Unit Price:** Price for a single unit of the product before applying any discount.
8. **Quantity:** Number of units purchased for this product line.
9. **Date:** Date and time of the transaction in YYYY-MM-DD HH:MM:SS format.
10. **Discount:** Discount applied to this line item, shown as a decimal (e.g., 0.30 means 30% off).
11. **Line Total:** Final cost for this line item after discount.
    Calculated as: Unit Price × Quantity × (1 - Discount).
12. **Store ID:** ID of the store where the purchase happened. Refers to the stores dataset.

13. **Employee ID:** ID of the employee who processed the sale. Refers to the employees dataset.
14. **Currency:** Three-letter currency code used in the transaction (like USD, EUR, GBP, CNY).
15. **Currency Symbol:** Symbol for the currency used in the transaction.
16. **SKU:** Stock Keeping Unit, a code made by combining Product ID, Size, and Color.
17. **Transaction Type:** Indicates whether the entry is a Sale or a Return.
18. **Payment Method:** How the customer paid — options may include Credit Card, Cash, etc.
19. **Invoice Total:** Total value of the full invoice. This value is repeated for all line items under the same Invoice ID.

6 Data Cleaning and Preparation:

Data cleaning and preparation being the crucial part of carrying out an accurate analysis was performed for the above-described 6 datasets of ZARA Global Retail data.

Each column of the dataset was examined to verify the kind of values and data types they hold, duplicates, missing/null and abnormal values were checked and corrected/imputed, outliers for numerical columns were identified and treated using IQR and few new features were derived from existing data values. This process ensured that the data was reliable and ready for further analysis.

6.1 - Step 1: Importing Data and required python libraries

▪ **Data Import:** Datasets from the provided google drive link were imported into Jupyter notebook for analysis. Also, all the necessary libraries such as NumPy, Pandas, Matplotlib, Seaborn were installed and imported to clean, handle and prepare the data for analysis and to further create visualizations.

6.2 - Step 2: Dataset Overview & Exploration

▪ **Shape**: As discussed earlier the dataset contains **6414328** rows and **27** columns.

▪ **Data Type**: The dataset contains values of various datatype like integer, float and object (text/string). Each column was separately examined to check for data type and the same was corrected wherever was required.

▪ **Count of non-null and null values**: Each attribute of the dataset was checked to count the number of non-null and null values. Most of provided datasets did not have null values. Obtained null values were imputed using existing proportion as per the observed scenarios to maintain the proportion of value distribution.

▪ **Duplicates:** Checked using duplicated().sum() function. There were few duplicates found in the transactions dataset which were removed from the dataset.

▪ **No. of Unique values**: Distinct values and their count for categorical columns were computed to understand distribution of values across dataset.

6.3 - Step 3: Handling Missing Values

Provided dataset has nearly zero missing values. Only a few null values were found in categorical columns of Discounts dataset.

▪ **Numerical Columns** – No missing values were found in any of the numerical columns of given dataset.

▪ **Categorical Columns** – Two categorical columns of Discounts dataset namely Category and Sub Category had null values which were imputed using existing proportion of values in these columns to maintain the distribution of data values in the dataset.

6.4 - Step 4: Modifying Data Values

It was observed that few values in the dataset needed modification to make them more interpretable and align with the required analysis hence such values were modified.

- **Country and City:** Few values in these columns were in present in Chinese. Such values were translated into English using Translator module from translatepy package to make these values interpretable.
- **Currency**: Values in Unit Price column were present in 4 different currencies, EUR, USD, CNY and GBP. To make all values uniform, values that belonged to EUR, CNY and GBP currencies were converted to USD using mapping.

6.5 - Step 5: Dealing with Outliers

▪ **IQR Analysis:** IQR (Inter Quartile Range) method was used to determine the outliers in numerical column of the transactions dataset like the Unit Price column. 25th percentile, 75th percentile and IQR were computed for this column.
Outlier count and outlier percentage for Unit Price value was computed by grouping on Product ID. Box plot and Scatter plot for Unit Price of top 12 Product ID on basis of outlier percentage were created.

▪ **Winsorization**: Upper bound and lower bound were calculated with ± 1.5*IQR with 75th percentile and 25th percentile respectively and the values that were greater than upper bound were set to upper bound and the values lesser than lower bound were set to lower bound. This limits extreme values and reduces their impact on analysis.

6.6 - Step 6: Datatype Correction

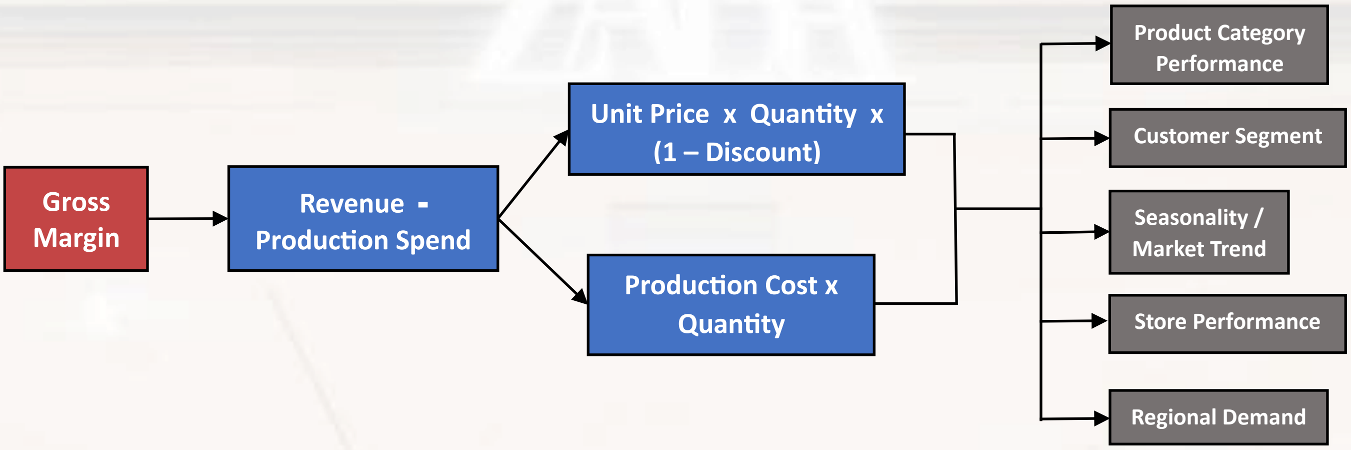
▪ **Verify Data Types:** Data type of each attribute was reviewed and converted to suitable data type after outlier and null value handling.
▪ **Converting Data Types:** All the identifier columns were assigned string datatype, all the date related columns were assigned datetime datatype, numerical columns were assigned integer datatype and categorical were converted to string type.

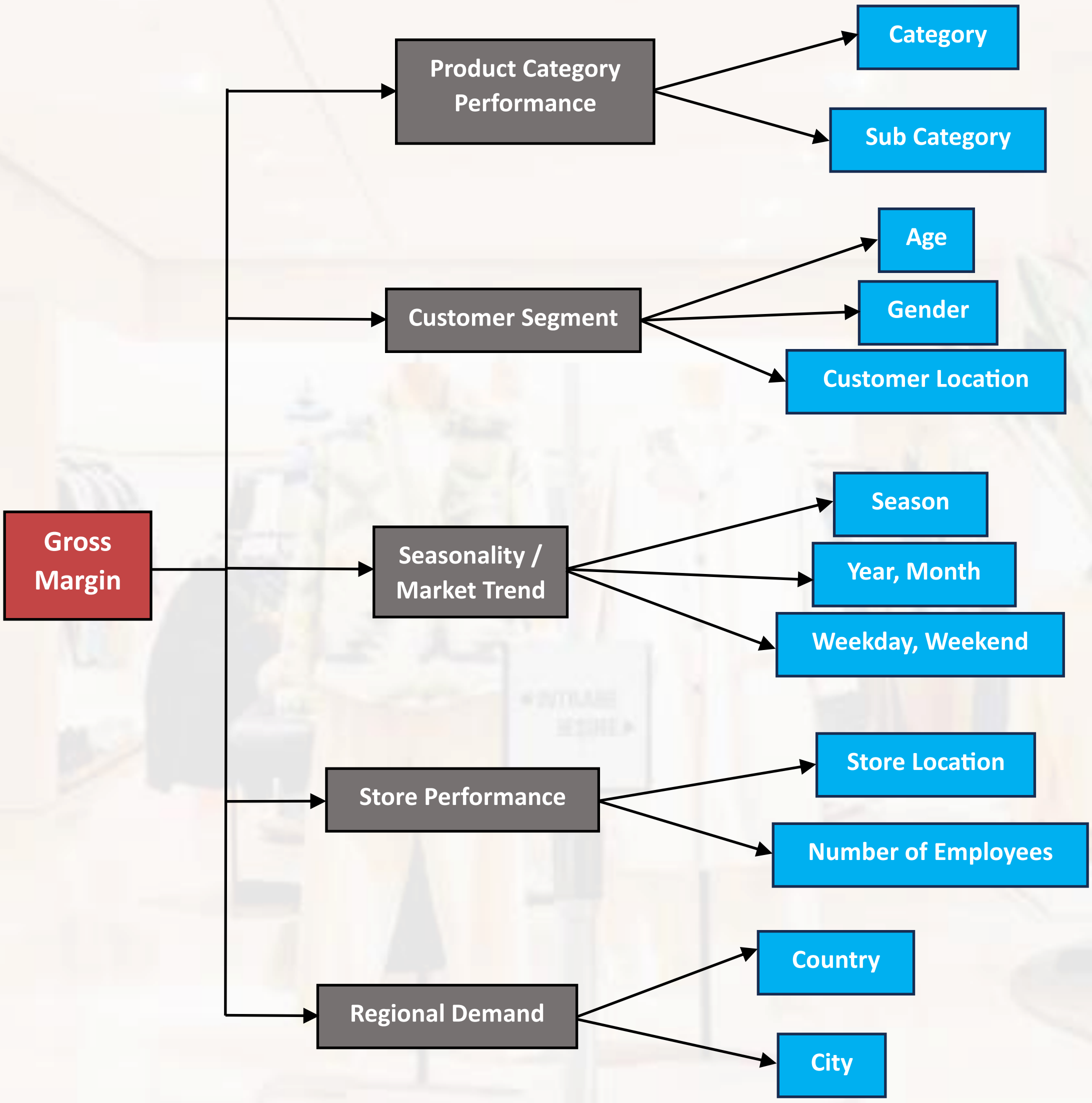
6.7 - Step 7: Feature Engineering

New variables were created using existing data value from the dataset for deeper insights.
▪ **Age:** Created new column Age using values from Date of Birth column of Customers dataset.
▪ **Year of Purchase:** Derived new column Year using values from Date column of transactions dataset.
▪ **Month of Purchase:** Derived new column Month using values from Date column of transactions dataset.
▪ **Weekday of Purchase:** Derived Weekday using values from Date column of transactions dataset.
▪ **Season:** Derived Season using derived Month values.

7 Metric Tree:

If the Gross Margin needs to be **improved** then the **Revenue** must be **increased** and the **Production Cost** must be **reduced**. Hence, the **factors** that are **impacting** Gross Margin need to be identified, **analysed** and worked upon to enhance revenue while managing to reduce the production cost leading to maximising Gross Margin.

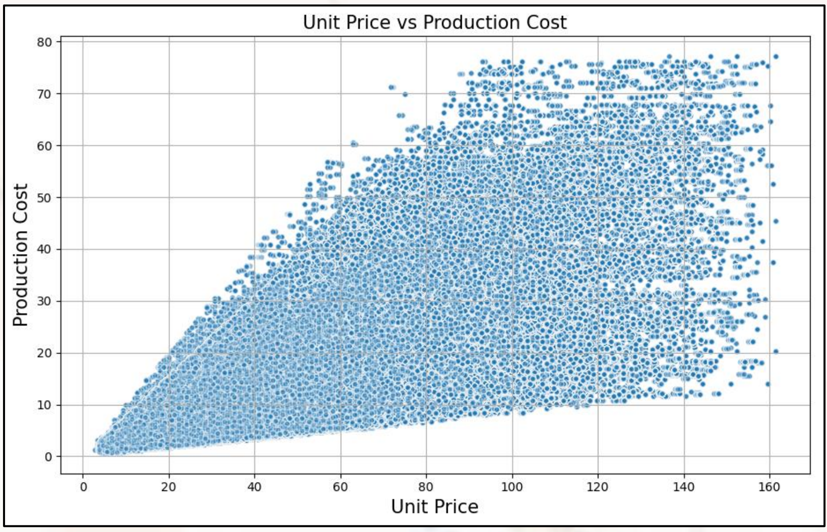

Gross Margin
Revenue - Production Spend
Unit Price x Quantity x (1 – Discount)
Production Cost x Quantity
Product Category Performance
Customer Segment
Seasonality / Market Trend
Store Performance
Regional Demand

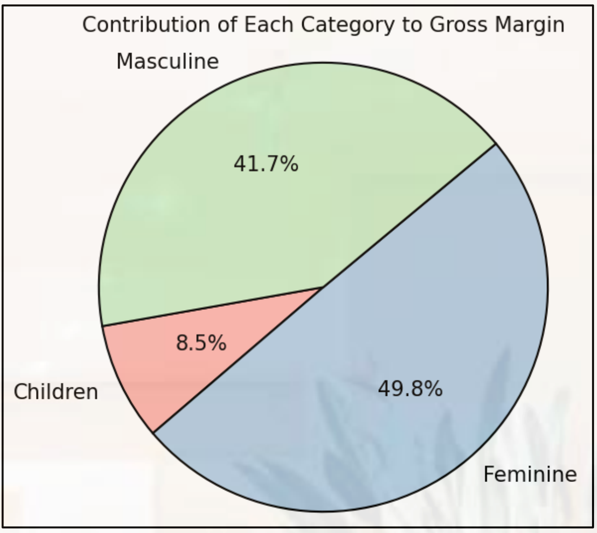


8 Exploratory Data Analysis (EDA):

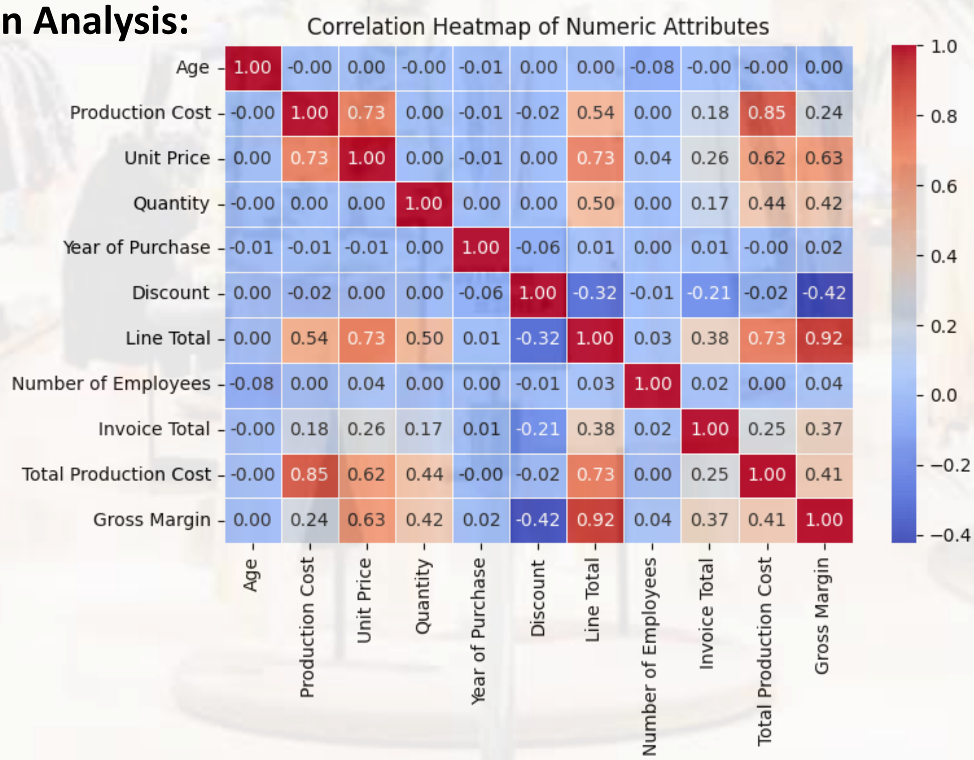
Key Summary Statistics:

Metric	Value
Total Revenue	308543527.8
Total Production Cost	115863150.4
Total Gross Margin	192680377.4
Average Discount	0.13
Average Unit Price	52.81
Total Quantity Sold	6684767
Number of Unique Products	17940
Number of Unique Categories	3
Number of Unique Sub Categories	21
Number of Unique Stores	35
Total Number of Transactions	4337270
Average Gross Margin per Transaction	31.71
Average Revenue per Transaction	50.78
Average Production Cost per Transaction	19.07





9 Correlation Analysis:



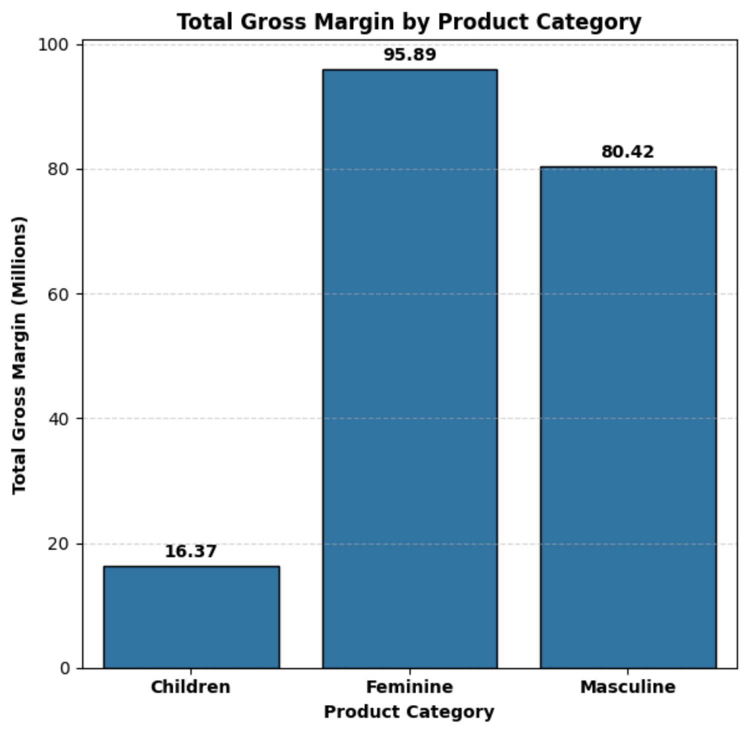
Below are the observations from above Key Summary Statistics and the Correlation Matrix heatmap:

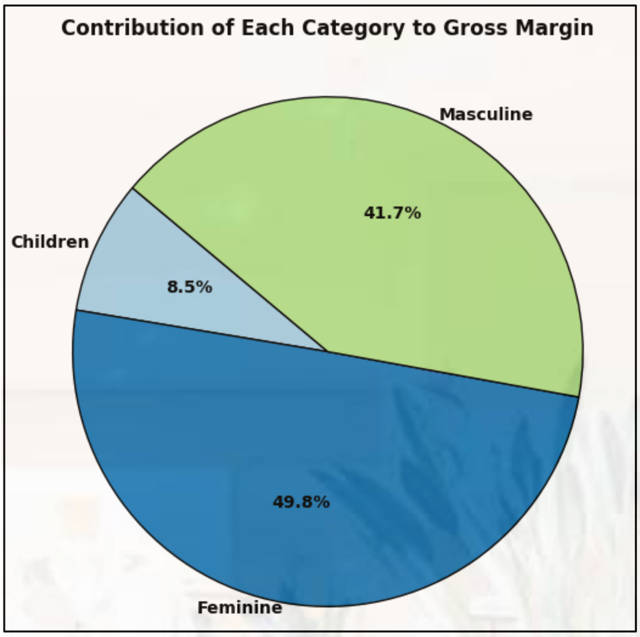
- Feminine and Masculine Product Categories are the major contributors in Gross Margin.
- A strong positive linear relationship can be observed between unit price - production cost and unit price - gross margin.
- Average discount is relatively low, suggesting Zara doesn't rely heavily on discounts for sales
- Zara has diverse set of products with 17940 unique products but has only 3 main categories hence there might be potential for expansion.

10 Hypothesis Formulation and Testing:

10.1 Product Category Performance –

Hypothesis 1.1) Some product categories contribute significantly more to Gross Margin than others.





Observation - The Children's category has the lowest Gross Margin with $16.37M while the Feminine category is the highest contributor with $95.89M. This proves the Hypothesis to be true.

Hence, there is a huge scope for Zara to expand the Children’s category which with proper planning could resulting in increased Gross Margin.

Recommendations –

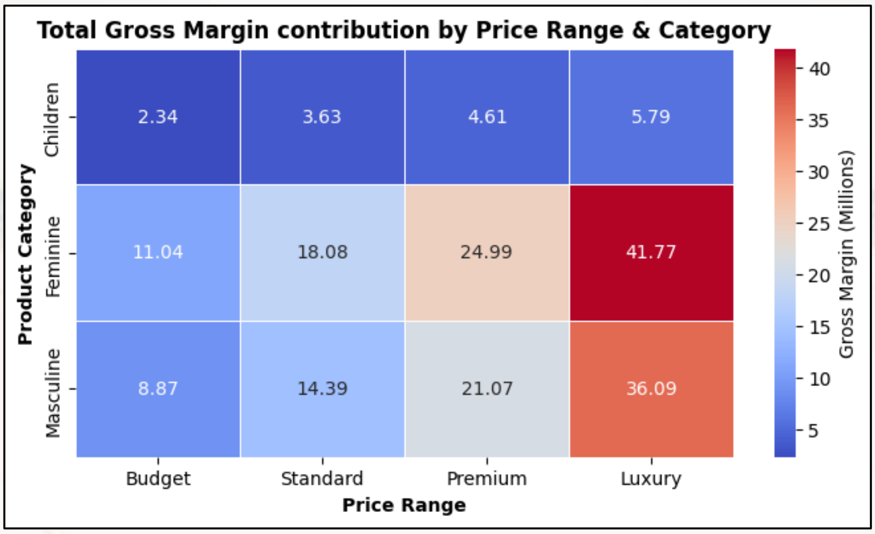
1) Introduce more premium children's apparel with higher margins which would help enhance overall margin.

2) Target parents through seasonal promotions and loyalty programs.

3) For Feminine category, introduce premium collections to further maximize margins.

4) For Masculine category, examine the production cost and try to reduce it to boost Gross Margin without raising prices.

Hypothesis 1.2) Higher priced products within each category contribute significantly more to Gross Margin compared to lower-priced products.



Observation - Luxury Products dominate the Gross Margin as this segment has the highest Gross Margin across Feminine ($41.77M) and Masculine ($36.09M) categories. But Children's products have significantly lower Gross Margins, even in the Premium segment. Also, the Standard and Premium price ranges show decent margins for all the product categories, so there is scope of improving margin in this groups.
Hence, the hypothesis turns out to be true for Feminine, Masculine and Children category.

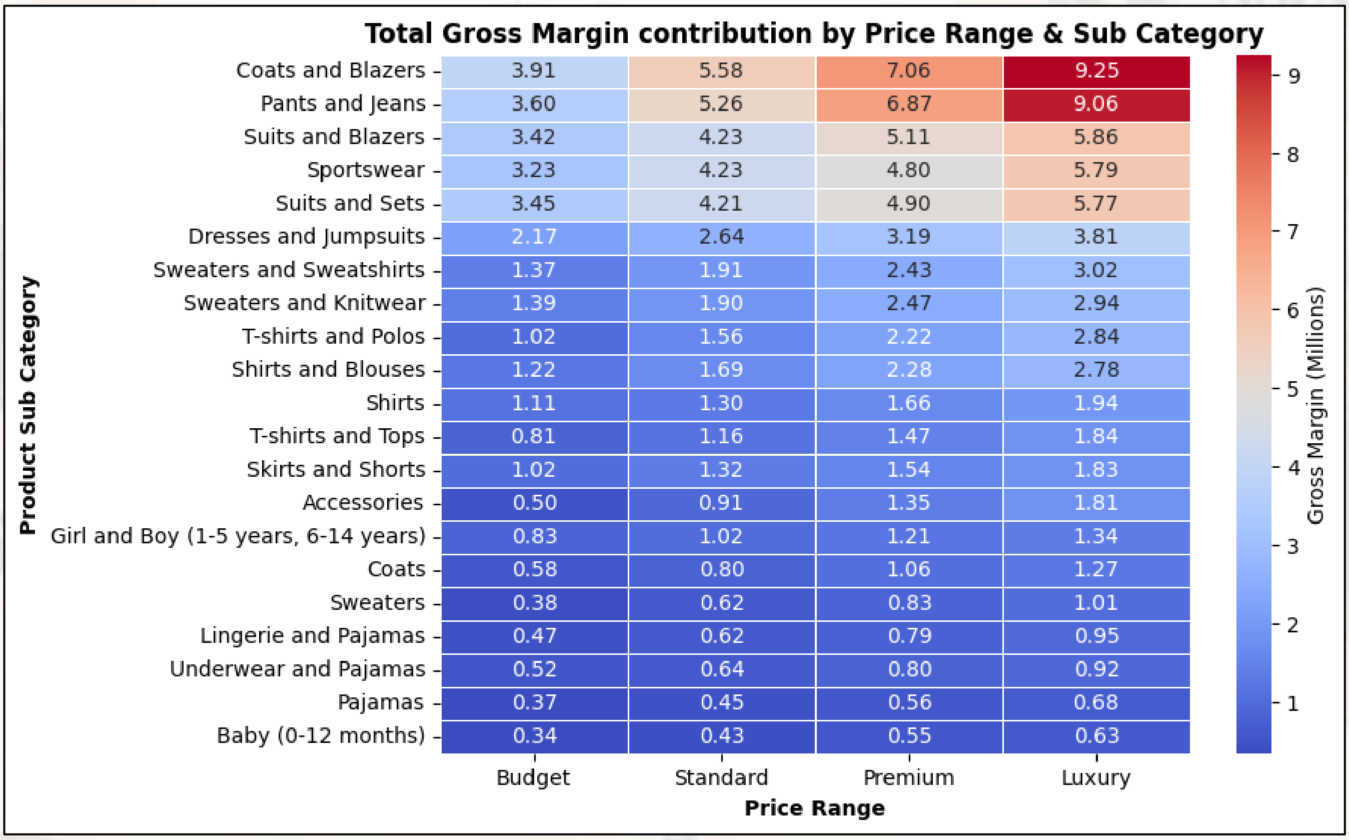
Recommendations –

1) Zara can invest more in premium collections, exclusive designs, and luxury product lines.

2) High-end children's products margins to be improved through quality and premium marketing.

3) Target the customers who frequently buy Standard/Premium items and offer tailored discounts.

Hypothesis 1.3) Higher priced products within each sub category contribute significantly more to Gross Margin compared to lower-priced products.



Observation -

All the Luxury Products belonging to all the Sub Categories dominate the Gross Margin with Coats & Blazers, Pants & Jeans leading the Luxury Segment. This indicates that the demand of high-end products is strong for Zara likely due to premium branding or material quality.

Sub Categories like Pajamas, Baby Clothing, Accessories show lower overall margins. This suggests less spending on premium versions of everyday items, with customers preferring budget/standard tiers.

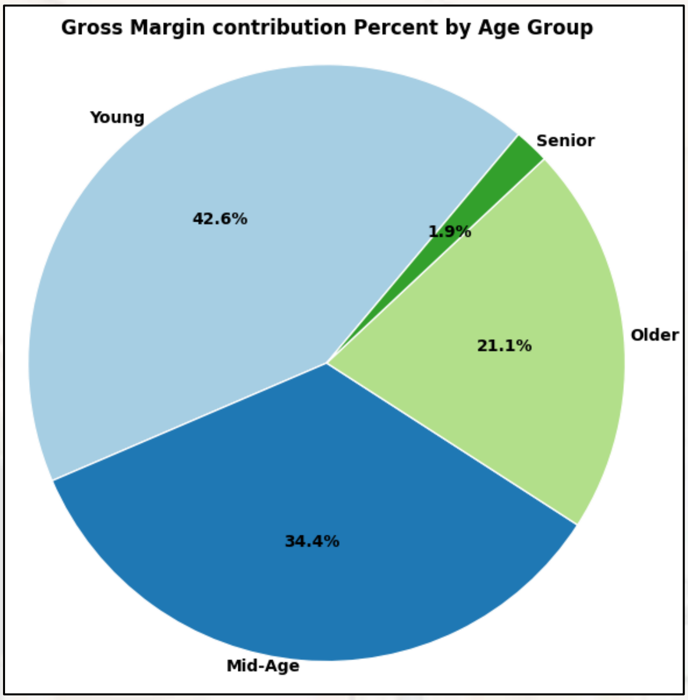
Hence, the hypothesis that the Higher priced products within each sub category contribute significantly more to Gross Margin turns out be true for all the sub categories.

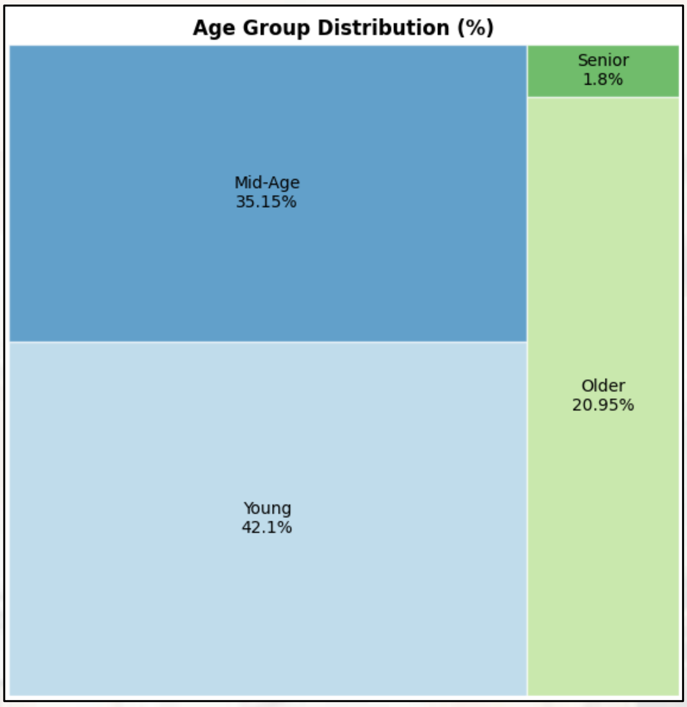
Recommendations –

1) Enhance visibility for Coats & Blazers, Pants & Jeans with limited editions & designer collaborations.
2) Target the customers who frequently buy Standard/Premium items and offer tailored discounts.
3) Optimize pricing strategy for middle-tier categories. Use strategic discounts for frequent and premium shoppers who might shift toward higher-end versions.
4) Improve the sales of low-margin segment using bundling strategies to increase sales of these items.

10.2 Customer Segment –

Hypothesis 2.1) Younger customers generate higher Gross Margin compared to older customers as a greater number of Younger customers tend to buy from Zara.





Observation - The Young customer segment contributes the highest share to Gross Margin with 42.10% contribution and $82.01M in Total Gross Margin. On the other hand, the Older Age Group has significantly lower contribution compared to younger groups with 20.95% contribution, $40.69M in Total Gross Margin. And, the Senior Age Group has Minimal impact on Gross Margin with just 1.80% contribution and $3.65M in Total Gross Margin. This proves the Hypothesis to be correct.

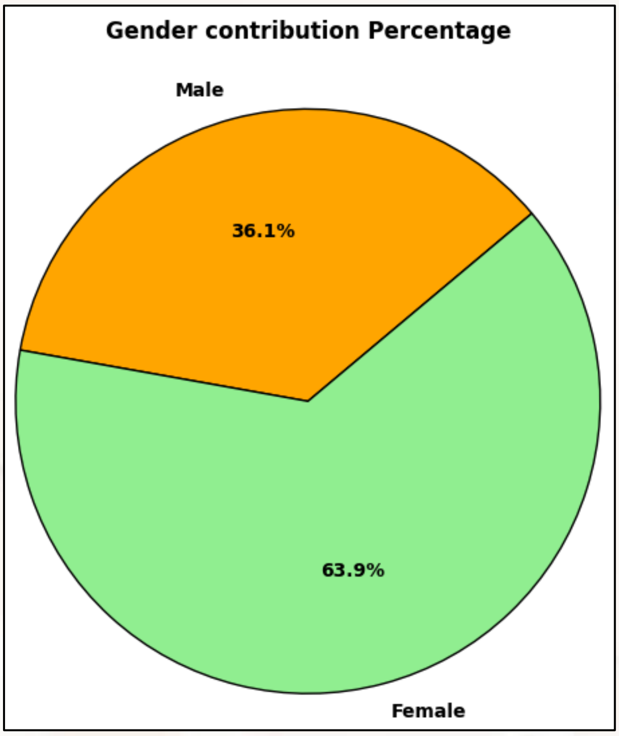
Hence, there is a huge scope to improving the products for older customer segments and targeting the older and senior customers.

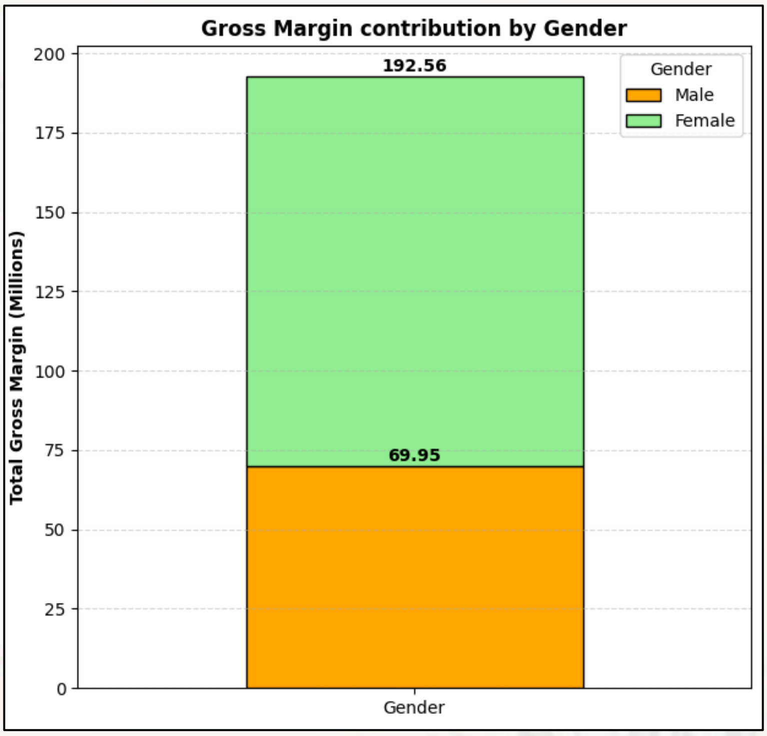
Recommendations -

1) Strengthen the Youth-Oriented Marketing. Focus on digital campaigns, influencer marketing, and social media promotions to maximize engagement.

2) Reevaluate the Senior customer segment strategy. Since demand is lower, consider targeted product expansion such as easy-wear & sustainable collections.

Hypothesis 2.2) Female customers contribute significantly more to Gross Margin than Male customers as more women buy from Zara than men.





Observation - Female customers dominate in Gross Margin contribution with 63.87% of Gross Margin, totalling $122.61M, significantly outperforming male consumers. This indicates strong demand for feminine product categories. While the males contribute 36.06% of Gross Margin totalling $69.95M the gap indicates potential opportunities for growth in the male customer segment.

Hence, this proves the hypothesis that Female customers contribute significantly more to Gross Margin than Male customers as more women buy from Zara than men to be true.

Recommendations -

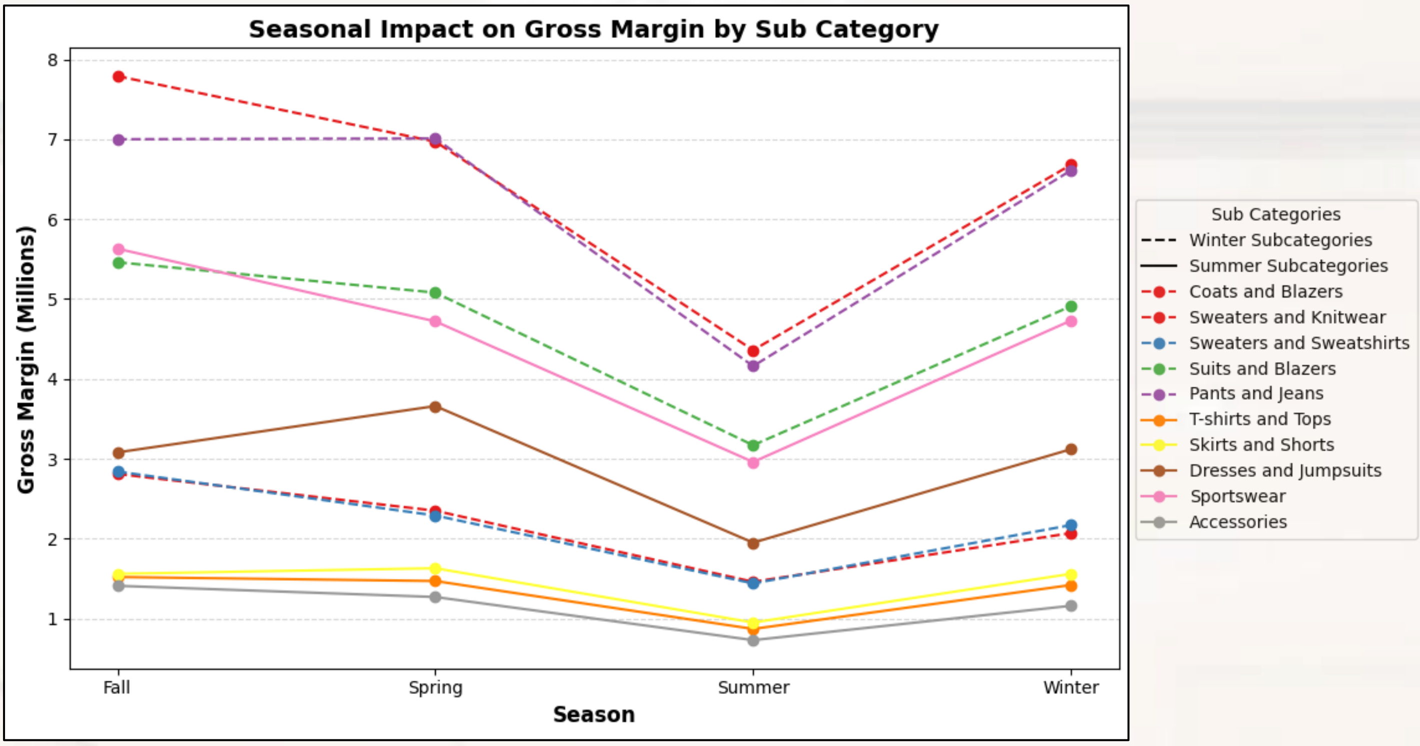
1) Improve the marketing efforts such as influencer promotions or personalized recommendations, targeting male shoppers.

2) Create exclusive memberships or loyalty programs tailored to men, encouraging repeat purchases.

3) Strengthen female centric category by introducing high-margin and premium collections to further maximize Gross Margin.

10.3 Seasonality / Market Trend –

Hypothesis 3.1) Product sub categories experience significant variations in Gross Margin based on seasons where the winter cloths peak in colder months and summer clothes dominate warmer months.



Observations – Winter clothings shows expected high performance in colder months:

- Coats and Blazers ($6.68M in Winter vs. $4.36M in Summer)
- Sweaters and Knitwear ($2.07M in Winter vs. $1.46M in Summer)
- Suits and Blazers ($4.91M in Winter vs. $3.17M in Summer)
- Pants and Jeans ($6.61M in Winter vs. $4.16M in Summer)

This demonstrates a clear seasonality effect with peak demand in colder months, confirming the hypothesis, aligning with temperature based seasonal trend.

However, Summer the clothings don't show extreme seasonality and actually perform slightly better in non-summer seasons.

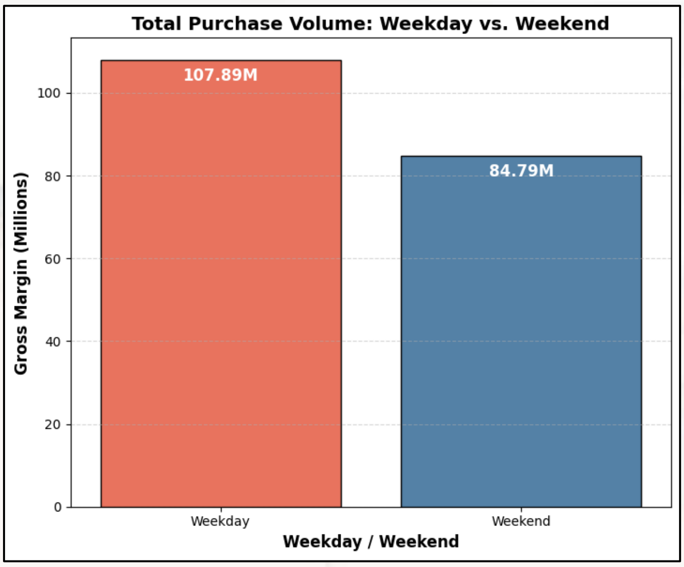
Recommendations -

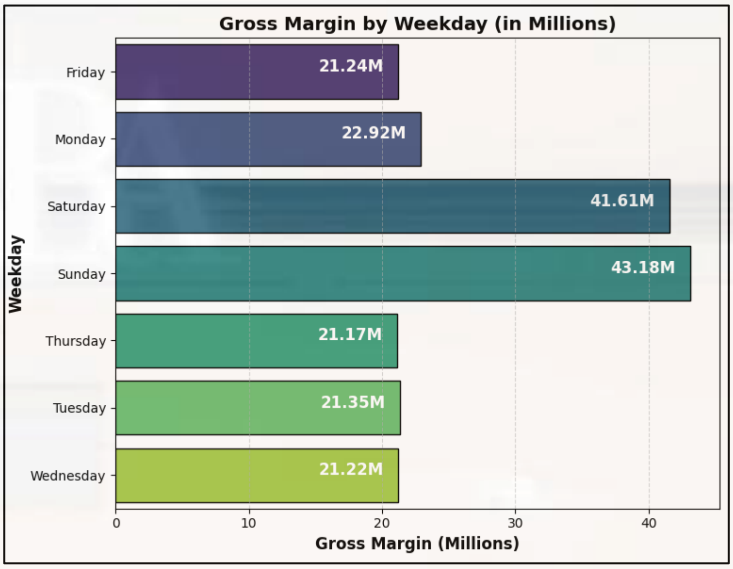
1) Enhance Winter products availability before peak demand. Stock up on Coats, Blazers, Sweaters, and Pants earlier to align inventory with seasonal spikes.

2) Introduce early winter discounts to capitalize on demand before full-season pricing begins.

3) Improve summer category positioning as T-shirts and Dresses aren't peaking as expected in summer hence increase visibility via targeted promotions.

Hypothesis 3.2) Higher purchase volumes and larger transaction values occur on weekends compared to weekdays.





Observations - Weekends contribute significantly to the overall gross margin but remain just slightly lower than weekday totals considering weekday total is for 5 days and weekend for 2 days, Saturday and Sunday. Weekday Total: $107.89M while Weekend Total: $84.79M.

This proves the hypothesis to be correct, hence higher purchase volumes and larger transaction values occur on weekends compared to weekdays.

Recommendations -

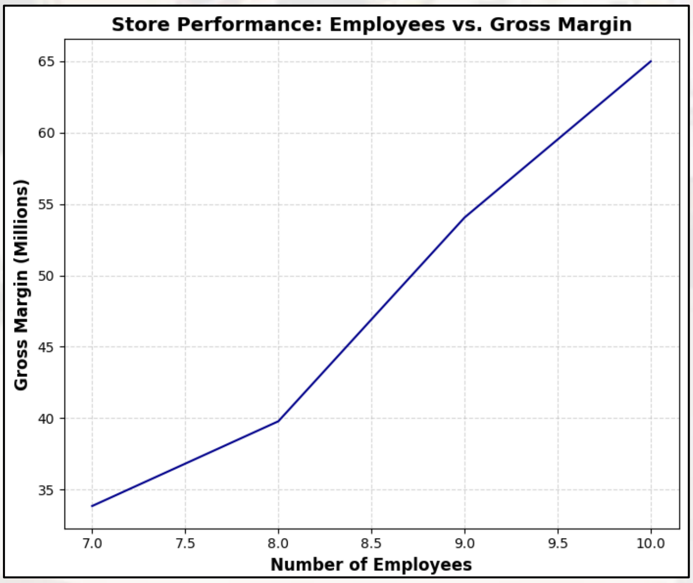
1) Enhance weekday promotions for routine shoppers, leverage mid-week flash discounts to boost engagement.

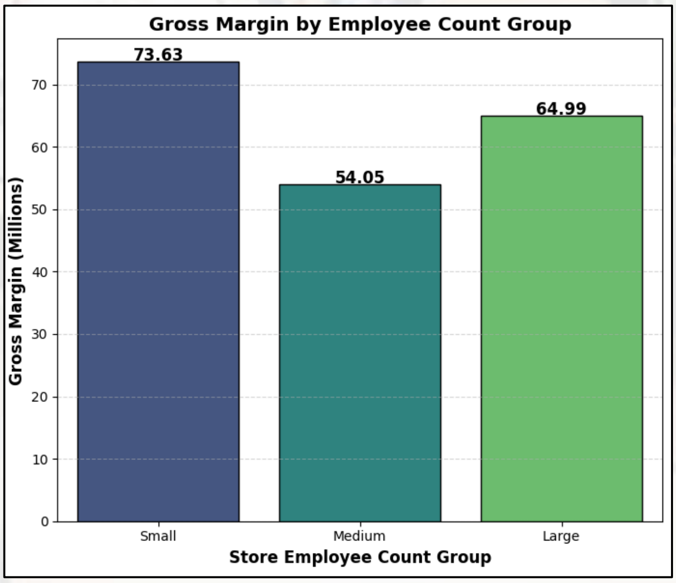
2) Increase weekend-specific promotions (e.g., "Saturday Specials") to drive impulse purchases.

3) Create targeted loyalty incentives for weekend shoppers.

10.4 Store Performance –

Hypothesis 4.1) Stores with more employees generate higher sales, leading to better customer service and improved purchase conversion rates.





Observation - Positive Correlation between employee count and Gross Margin can be seen as the line chart shows a clear upward trend, indicating that as the number of employees increases, the gross margin also grows. This suggests that larger teams may contribute to stronger sales performance, possibly due to better customer service, faster operations, and improved store management.

The bar chart shows that stores with a small number of employees have the highest gross margin ($73.63M), followed by large stores ($64.99M) and then medium stores ($54.05M). This contradicts that more employees always lead to better performance which might indicate efficiency in smaller teams or differences in store types and customer engagement.

Hence, considering the correlation and increasing gross margin with increase in employee count we can conclude the hypothesis to be correct.

Recommendations –

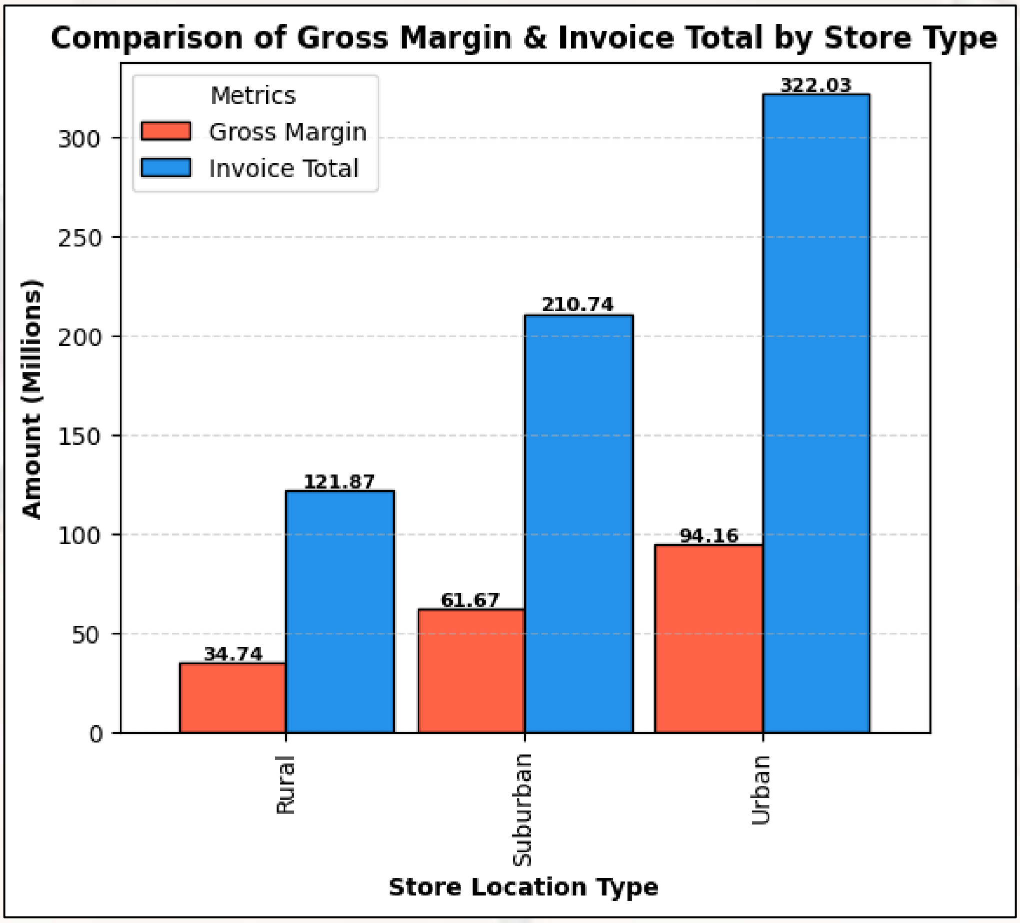
1) As small stores outperform larger stores, analyse best practices from small stores and apply these strategies to larger stores.

2) Optimize the employee count for efficiency by understanding why small teams outperform medium ones.

3) Investigate medium-sized stores, consider adjusting staff allocation or enhancing sales strategies to boost the gross margin of these stores.

10.5 Regional Demand –

Hypothesis 5.1) Stores in high-density urban areas generate higher sales volume and Gross Margin compared to stores in suburban or rural regions.



Observations - The grouped bar chart shows that urban locations generate the highest Gross Margin, confirming the hypothesis that high-density areas drive stronger revenue. This suggests that metropolitan areas contribute to greater profitability.

Suburban locations have lower Gross Margin than urban stores but still perform better than rural stores.

Rural locations consistently show lowest sales volume and Gross Margin confirming that lower population density struggle with fewer customers, leading to reduced foot traffic and purchasing power.

Recommendations –

1) Leverage marketing and promotions in urban locations to increase high-value purchases maximizing urban store revenue potential.

2) Focus on premium product offerings, as urban customers tend to have higher spending capacity.

3) Introduce localized promotions tailored to suburban areas to target suburban population.

4) Improve growth strategies for rural stores. Explore low-cost, high-demand essential products as per rural spending behaviour.

11 Strategic Insights and Actions:

1) **Product Category Performance:** Feminine category dominates revenue, while children category products display an untapped potential. Luxury products provide the highest Gross Margin.
Expand children's premium product line and boost marketing efforts in this segment.

2) **Customer Segment Trends:** Younger customers and female shoppers contribute significantly more to Gross Margin, confirming their huge role in Zara's success.
Strengthen youth-focused & female-targeted marketing strategies to maximize engagement.

3) **Seasonality & Purchase Behaviour:** Winter clothing peaks during colder months, and purchase volumes are higher on weekends.
Optimize pricing, inventory management for seasonal trends, ensuring winter wear is stocked as demand.

4) **Store Efficiency:** Higher employee count correlates with increased Gross Margin, but small stores show the highest efficiency.
Apply successful small-store strategies to larger outlets for improved efficiency.

5) **Regional Demand:** Urban stores outperform suburban and rural locations in sales and profitability.
Enhance urban store promotions and suburban and rural stores marketing to maximize sales potential.

12 Tools/ Libraries Used:

1) Pandas for data cleaning and manipulation.

2) NumPy for numerical computations.

3) Matplotlib and Seaborn for data visualization.

Environment: Google Colab for data processing and analysis.

13 Limitations of Analysis:

1) Data Scope: The dataset is limited to just a period of three years, which might not capture long term learning behaviours.

2) Outliers: Although outliers were identified and adjusted but the modification might not maintain data authenticity.