**Zara Sales Analysis**

**A Comprehensive Report on Zara EDA Analysis**

***Submitted by:***

**Dhrumil Patel**

***Guided by:***

**Prof. Chintan Patel**



**Fly The Nest – Deets Digital**

**PROJECT REPORT**

**Abstract**

This project presents an **exploratory data analysis (EDA)** of **ZARA's retail dataset** to uncover patterns in sales performance, pricing strategies, product categorization, and customer segments. Using Python libraries such as Pandas, Matplotlib, and Seaborn, we cleaned and analyzed data attributes including product categories, prices, discounts, regions, and sales quantities. The analysis reveals valuable business insights, such as the impact of discount percentages on sales volume, regional performance variations, and customer segment distributions. Visualizations like bar charts, pie charts, heatmaps, boxplots, and line charts were used to interpret trends effectively. This study aims to support data-driven decision-making for marketing, inventory, and pricing strategies in fast fashion retail.

This project was conducted under the expert guidance of **Prof. Chintan** and supported by **Fly the Nest**, whose mentorship and resources were integral to its successful execution. The insights generated from this analysis provide valuable recommendations for optimizing future sales strategies in the **ZARA**.

Contents

[Problem Statement & Objective 6](#_Toc197716319)

[Problem Statement: 6](#_Toc197716320)

[Objectives 6](#_Toc197716322)

[Introduction 7](#_Toc197716323)

[Exploratory Data Analysis (EDA) 7](#_Toc197716324)

[EDA Pipeline 7](#_Toc197716325)

[About the company 8](#_Toc197716326)

[Tools and platforms used in project 8](#_Toc197716330)

[Why Python? 8](#_Toc197716331)

[Platforms used 8](#_Toc197716332)

[Versions of platform 8](#_Toc197716333)

[Chapter 1: Data Loading/reading 9](#_Toc197716334)

[Load Data (Jupyter Notebook) 9](#_Toc197716335)

[Import Necessary Library 9](#_Toc197716336)

[Chapter 2: Familiarize with Data & Identifying the Target Variable 10](#_Toc197716337)

[Explore the provided data (column names, data types) 10](#_Toc197716338)

[Overview of data 10](#_Toc197716339)

[Interpretation 10](#_Toc197716340)

[Chapter 3: Data Preparation & Transformation 11](#_Toc197716341)

[Data Cleaning 11](#_Toc197716342)

[Handle missing values (imputation or deletion) 11](#_Toc197716343)

[Data Reduction: Remove unwanted columns or rows 11](#_Toc197716344)

[Format data types (numerical & categorical) 12](#_Toc197716345)

[Change data types of columns and Memory optimization 12](#_Toc197716346)

[Feature Engineering (Create new features/variables) 12](#_Toc197716347)

[Create new features 12](#_Toc197716348)

[Chapter 4: Data Analysis & Visualization 13](#_Toc197716349)

[Overview of data before analysis 13](#_Toc197716350)

[4.1 Product & Category analysis 14](#_Toc197716351)

[4.2 Pricing & Sales analysis 16](#_Toc197716352)

[4.3 Product Positioning analysis 19](#_Toc197716353)

[4.4 Promotion Impact analysis 20](#_Toc197716354)

[4.4 Price Distribution 22](#_Toc197716355)

[Chapter 5: Summary and Suggestions 27](#_Toc197716356)

[5.1 Summary 27](#_Toc197716357)

[**1. Product Category Distribution** 27](#_Toc197716358)

[**2. Product Placement and Promotion Impact** 27](#_Toc197716359)

[**3. Seasonality and Sales** 27](#_Toc197716360)

[**4. Price and Sales Analysis** 27](#_Toc197716361)

[**5. Descriptive Trends** 27](#_Toc197716362)

[**6. Heatmap Insights** 27](#_Toc197716363)

[**7. Visual Aesthetics** 27](#_Toc197716364)

[5.2 Suggestions & Recommendations 28](#_Toc197716365)

[**1. Promotion Strategy** 28](#_Toc197716366)

[**2.Inventory Planning** 28](#_Toc197716367)

[**3.Seasonal Campaigns** 28](#_Toc197716368)

[**4.Pricing Strategy** 28](#_Toc197716369)

[**5.Section Diversification** 28](#_Toc197716370)

[**6.Product Descriptions** 28](#_Toc197716371)

[**7.Data Gaps** 28](#_Toc197716372)

[Annexure 29](#_Toc197716373)

[Git hub link 29](#_Toc197716374)

[Code snippets 29](#_Toc197716375)

[1. Loading and Inspecting Dataset 29](#_Toc197716376)

[2. Handling Missing Values 29](#_Toc197716377)

[3. Bar Chart: Product Category Distribution 29](#_Toc197716378)

[4. Pie Chart: Promotion Usage 29](#_Toc197716379)

[5. Boxplot: Price Range by Category 30](#_Toc197716380)

[6. Line Chart: Sales Trend by Category 30](#_Toc197716381)

[7. Heatmap: Correlation Between Numerical Features 30](#_Toc197716382)

[8. Stylized Table: Region vs Category 30](#_Toc197716383)

[References 31](#_Toc197716384)

# Problem Statement & Objective

## Problem Statement:

## Zara is a global leader in fast fashion but faces challenges in optimizing its pricing, promotion strategies, and product placement to maximize sales. The company needs detailed insights into how product categories, pricing, promotions, and positioning affect sales volume to make data-driven business decisions.

## Objectives

**Identifying Top-Selling Categories and Products**: Detect which product categories and individual products contribute most to Zara’s overall sales and profitability to strengthen inventory and marketing efforts.

**Analysing Price Sensitivity:** Study how pricing impacts sales performance to fine-tune pricing strategies across different product segments.

**Evaluating Promotion Effectiveness**: Examine the influence of promotional activities on sales uplift and customer buying behavior to optimize future marketing campaigns.

**Understanding Customer Preferences:** Uncover trends in product choices, preferred price ranges, and seasonal purchasing behaviors to align product offerings with customer expectations.

**Improving Product Positioning**: Assess how different product types are positioned in the market and identify opportunities for better assortment and strategic placement.

**Optimizing Inventory Management**: Leverage insights from sales patterns to better plan stock levels, reduce overstocking, and minimize lost sales due to out-of-stock situations.

**Enhancing Business Decisions with Data-Driven Insights**: Provide Zara’s management with clear, actionable insights from data analysis to support better decision-making in product, pricing, and marketing strategies.

# Introduction

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize the data, using statistical analysis and data visualization methods.

It is a crucial step in any data analysis process, enabling data analysts to uncover patterns, spot anomalies and gain insights to take actions

## EDA Pipeline

**1. Data Acquisition and Objective**

1. Obtain Zara product sales and pricing dataset (CSV, Excel).
2. Define the problem statement: Analyse product performance, pricing impact, promotional strategies, and customer buying behaviour to improve sales and profitability.
3. Choose tools/environment and programming language: Python (Jupyter Notebook / VS Code), Excel (optional), visualization libraries (Matplotlib, Seaborn, Plotly).

**2. Data Loading/Reading**

1. Load raw data into Jupyter Notebook for analysis.
2. Initial checks for data integrity (duplicates, nulls, formatting issues).

**3. Familiarize with Data & Identify Target Variables**

1. Explore dataset (column names, data types, basic stats).
2. Identify key focus areas based on objectives: sales volume, price, product category, promotion flag, etc.

**4. Data Preparation & Transformation**

1. Data Cleaning (remove duplicates, handle missing/invalid data).
2. Handle missing values through appropriate imputation or removal.
3. Format columns into correct data types (e.g., dates, numbers, categories).
4. Filter and structure the dataset to prepare for deep analysis.

**5. Feature Engineering**

1. Create new features like Discounted Price, Profit Margin, Promotion Impact Score, etc., to enhance analysis depth.
2. Categorize products based on pricing tiers (e.g., Low, Mid, Premium).

**6. Data Analysis & Visualization**

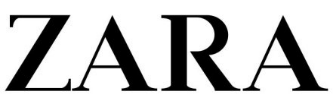
1. **Univariate Analysis**
2. Study single-variable distributions (sales, price, product category).
3. Summarize numerical variables (mean, median, stddev).
4. Analyze frequency distribution for categorical features
5. **Bivariate and Multivariate Analysis**
6. Correlate sales volume with price, category, promotion status, etc.
7. Use visualizations (bar charts, pie charts, histograms, boxplots, heatmaps, scatterplots).

**7. Summary and Suggestions**

1. Summarize key insights from data (e.g., top-performing categories, optimal pricing strategies, effectiveness of promotions).
2. Provide actionable suggestions to enhance Zara’s product mix, pricing, inventory planning, and promotional strategies.

## About the company

## 



* Zara is a Spanish multinational retail clothing chain, owned by the Inditex group.
* Zara is known for its fast fashion products, including clothing, accessories, shoes, swimwear, beauty, and perfumes.
* The brand emphasizes quick inventory turnover, responsive design to customer trends, and frequent collection updates to meet market demands.

## Tools and platforms used in project

### Why Python?

* Python was chosen for its powerful data analysis and visualization libraries, making it ideal for handling Zara’s sales and product data efficiently.
* It enabled quick cleaning, transformation, and insightful visual exploration, all within an interactive environment like Jupyter Notebook.

### Platforms used

* Jupyter notebook – it is a web-based application for running code and queries in Python

### Versions of platform

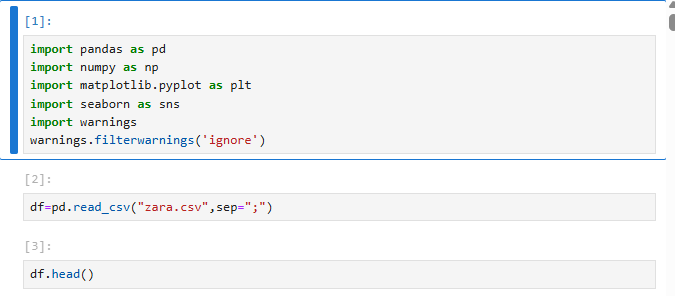
* Jupyter notebook - 7.3.2
* Python version - 3.12.4

Note: *Code format is IPYNB file.It is a text-based file used by Jupyter Notebook (VS Code also supports IPYNB with Jupyter extension)*

# Chapter 1: Data Loading/reading

## Load Data (Jupyter Notebook)

The dataset has been loaded using the Pandas library in the Jupyter Notebook editor and stored in a DataFrame named df for further analysis and processing.



## Import Necessary Library

* numpy (np): Provides efficient numerical computation tools
* pandas (pd): Offers data manipulation and analysis structures (DataFrames, Series)
* seaborn (sns): Creates informative statistical data visualizations based on Matplotli
* matplotlib.pyplot (plt): Enables various plot creations for data visualization
* warnings (with warnings.filterwarnings("ignore")): Suppresses warning

Figure 2: Loading data in to MySQL

# Chapter 2: Familiarize with Data & Identifying the Target Variable

## Explore the provided data (column names, data types)

* We need to understand the data before cleaning the data and also cross verify if all the required data are provided by Zara.

### Overview of data

* df.head(); Let's see the data by displaying the first 5 rows
* df.tail(); Let's see the last 5 rows
* df.shape is used to get the dimensions (number of rows and columns) of data
* df.size is used to get the total number of elements in a pandas
* df.info() - used to display concise information about

### Interpretation

* Structured Data: Data provided is in table format
* Dimensions (16 Columns x 252 Rows) of the DataFrame or data
* Column Data Types: Observed mix of data type of each column (e.g datetime64[ns]
* , object, float, int, etc.)
* Also note that all categorical/qualitative variables are Nominal in nature (it has no specific orders)
* Non-Null Counts (no Null values observed) in each column.
* Memory Usage: An estimate of the memory usage is 31.6+ kb (Further, we will optimize the memory usage by modifying the data types)

# Chapter 3: Data Preparation & Transformation

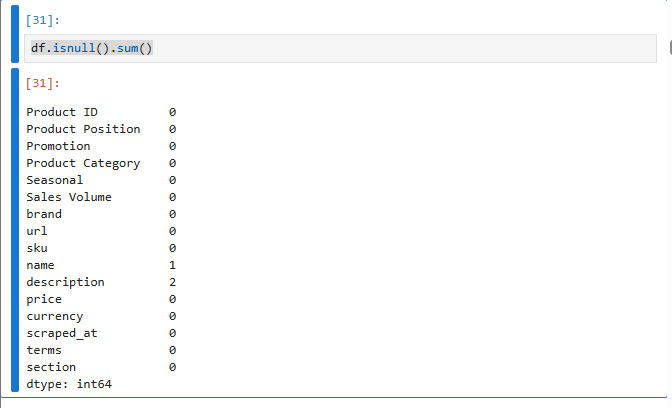
## Data Cleaning

We need to perform steps mentioned below to clean data:

* **Steps involved in handling missing values** (imputation, deletion)
  + We accept missing values if data is small in dimension
  + We delete missing values if:
    - When more than 80% of data is missing/null values
    - When the percentage of missing values are very small, deleting will have minimal effect on analysis
* **Replacing the missing values by imputation**
  + Imputation: We replace the missing values by Mean, Median or Mode of the variable or perform fill null values(fillna method) with the desired value
* **Data Reduction:** Remove unwanted data (if present) which are not required for analysis
  + Delete unwanted columns
  + Delete duplicate rows
* **Format data types** (numerical & categorical variables)
* **Outlier detection and handling** (we ignore this step because outliers are valid in our case)
  + When data has extreme values that could effect our analysis, we either replace them with Mean or Median or Mode values or we accept the outliers
  + We identify the outliers by plotting the Box plot

### Handle missing values (imputation or deletion)

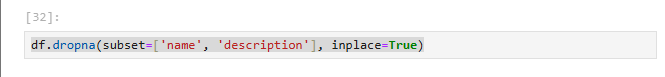
* df.isnull().sum() - Gives sum of all null values in each column
* df.notnull().sum() - Gives sum of all not null unique values in each column
* Interpretation:
  + Data has no null values, so no need to perform process to handle missing values



### Data Reduction: Remove unwanted columns or rows

There are no unwanted columns to delete, so we can check for duplicated rows and delete the duplicates

* df.duplicated().sum(); This shows number of duplicated rows
* df = df.drop\_duplicates(); This removes duplicated rows
* Interpretation:
  + It is noted that there are 2 rows which are repeated. We have removed duplicated rows



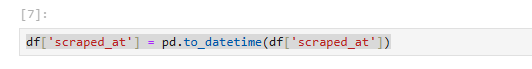
### Format data types (numerical & categorical)

We need to format columns, that will ease data analysis, below steps are performed to the required format for analysis

* Rename of columns - To keep columns descriptive as well as simple
* Change data types - We change data type to keep consistency and also for memory optimization

### Change data types of columns and Memory optimization

* df.info(); We can see the column data type
* cols\_to\_convert = df.columns[2:12]; Select 3rd column to 12 column
* df['Order Date'] = pd.to\_datetime(df['Order Date']); Convert the "Order Date" column to datetime data type
* Interpretation:
  + We are converting all columns with object data type to category and converted order date column to date-time format
  + Now memory usage is reduced from 32.6+ kB to 31.6+ KB. With this we achieved memory optimization for space complexity



## Feature Engineering (Create new features/variables)

* We derive new variables or features by combining multiple columns or derive new features by performing calculation
* Here we need to create new columns for easier analysis
* Create new columns by calculating TotalRevenue.

### Create new features

* df['profit'] = df['price'] \* df['Sales Volume']; Creates a new column of the profit
* Interpretation:
  + We are set with feature engineering by creating multiple columns for easier analysis

# Chapter 4: Data Analysis & Visualization

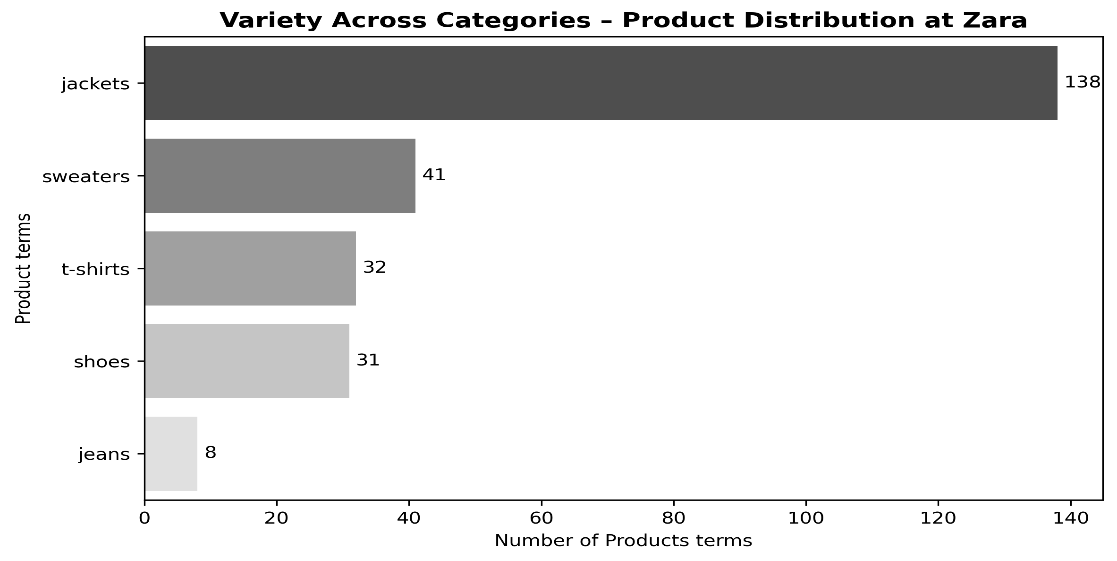
### Overview of data before analysis

* After Data Wrangling, we can check the columns once before we proceed to perform analysis
* df.columns; Display all columns in the data frame
* Interpretation:
  + Description of variables

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Product ID | Unique identifier for the product. |
| Product Position | Store location where the product is placed (e.g., Front of Store, End-cap). |
| Promotion | Indicates if the product is under promotion (Yes/No). |
| Product Category | The category the product belongs to (e.g., Clothing, Shoes). |
| Seasonal | Shows whether the product is seasonal (Yes/No). |
| Sales Volume | Number of units sold for that product. |
| brand | Brand name of the product (e.g., Zara). |
| url | Direct link to the product on the website. |
| sku | Stock Keeping Unit – a unique code for inventory tracking. |
| name | The name/title of the product. |
| description | A brief textual description of the product’s features and design. |
| price | The selling price of the product. |
| currency | The currency in which the price is listed (e.g., USD). |
| scrapped\_at | Timestamp indicating when the product data was scraped from the website. |
| terms | Product tag or subcategory term (e.g., jackets, shoes). |
| section | The section or target demographic (e.g., MAN, WOMAN, KIDS). |

## 4.1 Product & Category analysis

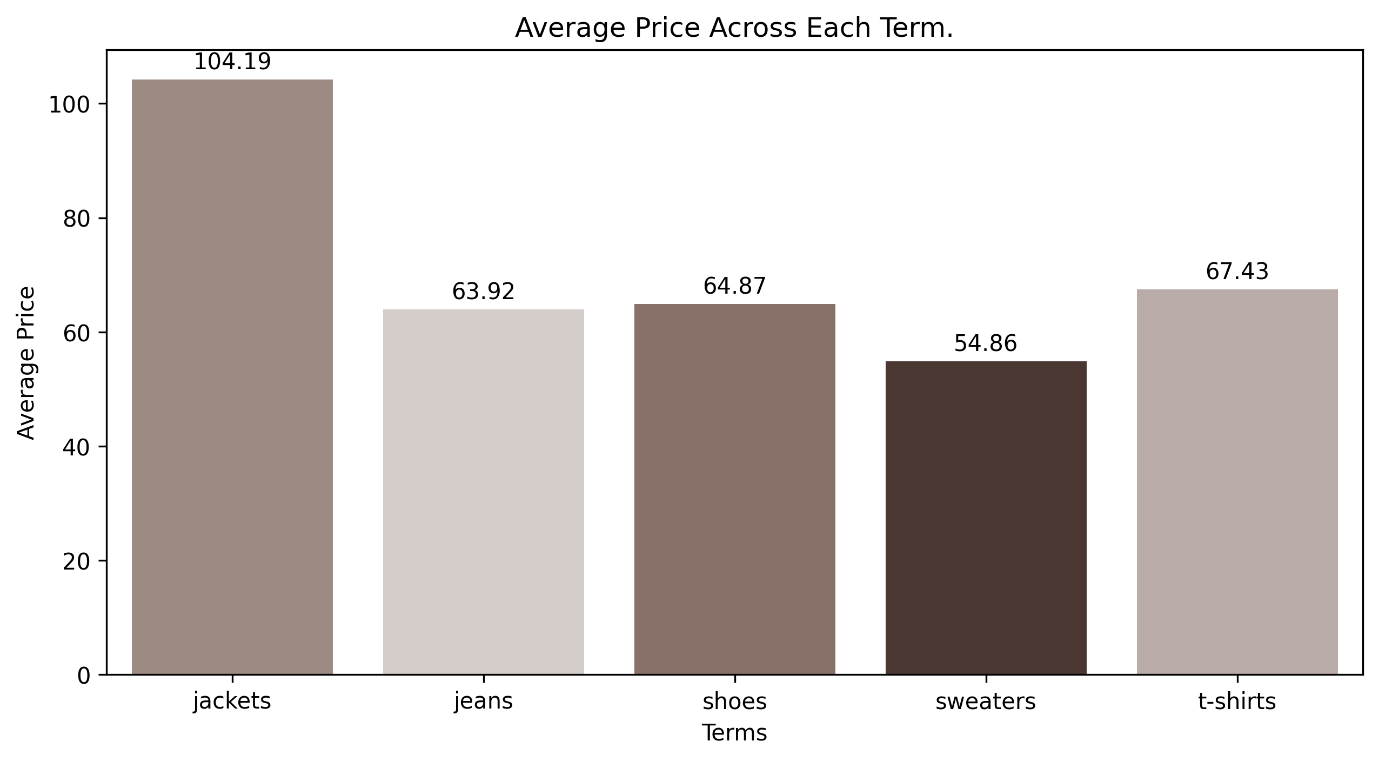
**1.How many products are available in each product terms?**



**Interpretation:**

Jackets dominate the product variety at Zara, making up the largest category with 138 items. Sweaters, t-shirts, and shoes have moderate variety, while jeans have the least number of options. This suggests Zara places a strong emphasis on jackets in its product lineup.

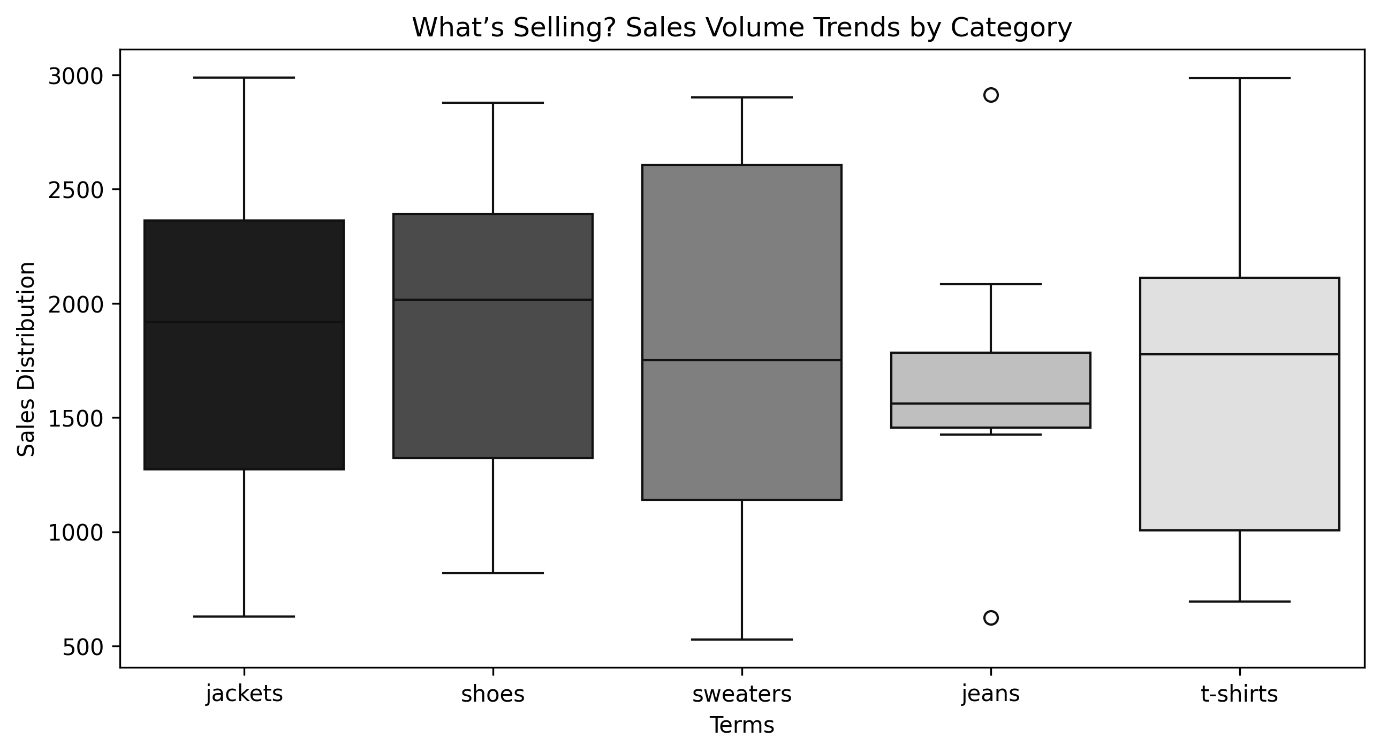
**2. Average price of products in each category?**



**Interpretation**

Jackets have the highest average price, indicating a premium category, while sweaters are the most affordable. Other items like jeans, shoes, and t-shirts fall in a mid-price range.

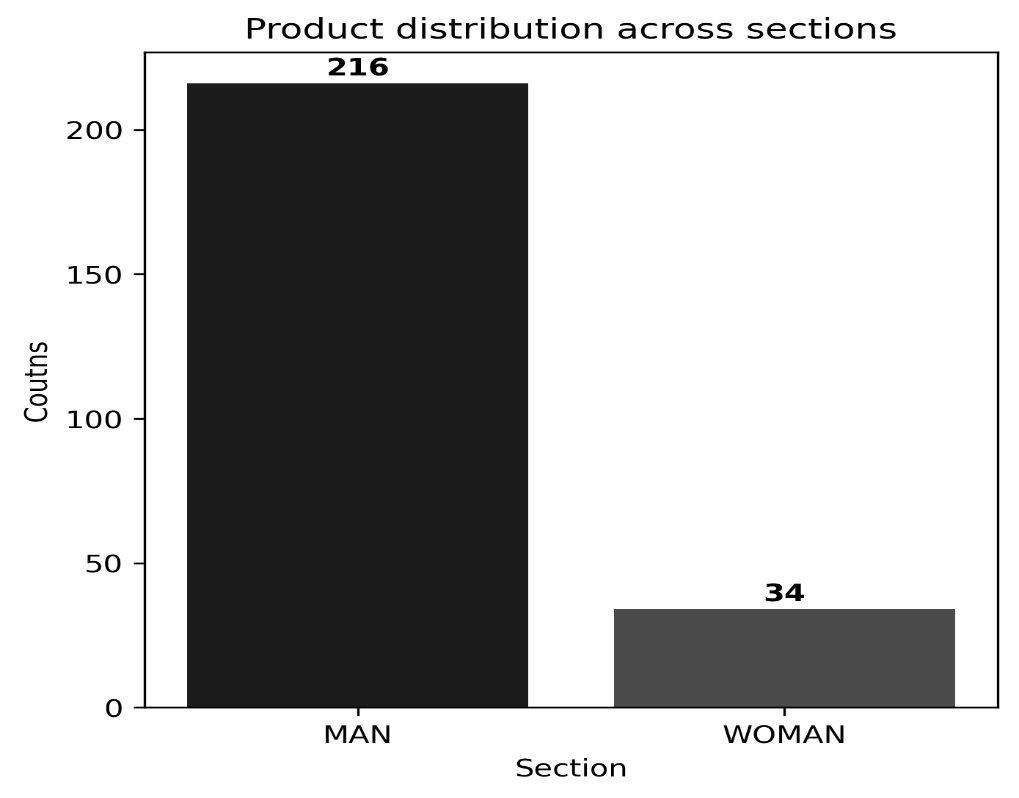
**3.Sales volume distribution by product category**



**Interpretation**

The box plot illustrates sales volume trends by product category (jackets, shoes, sweaters, jeans, t-shirts). Sweaters have the highest median sales volume, around 2500 units, with a wide distribution. Jackets and shoes follow with medians around 2000 units, while t-shirts have a median near 2000 units with a narrower range. Jeans show the lowest median, below 1500 units, with an outlier at 3000 units. This suggests sweaters are the top-selling category, while jeans lag behind.

**4.Product distribution across sections (MEN, WOMEN, KIDS, etc.)**

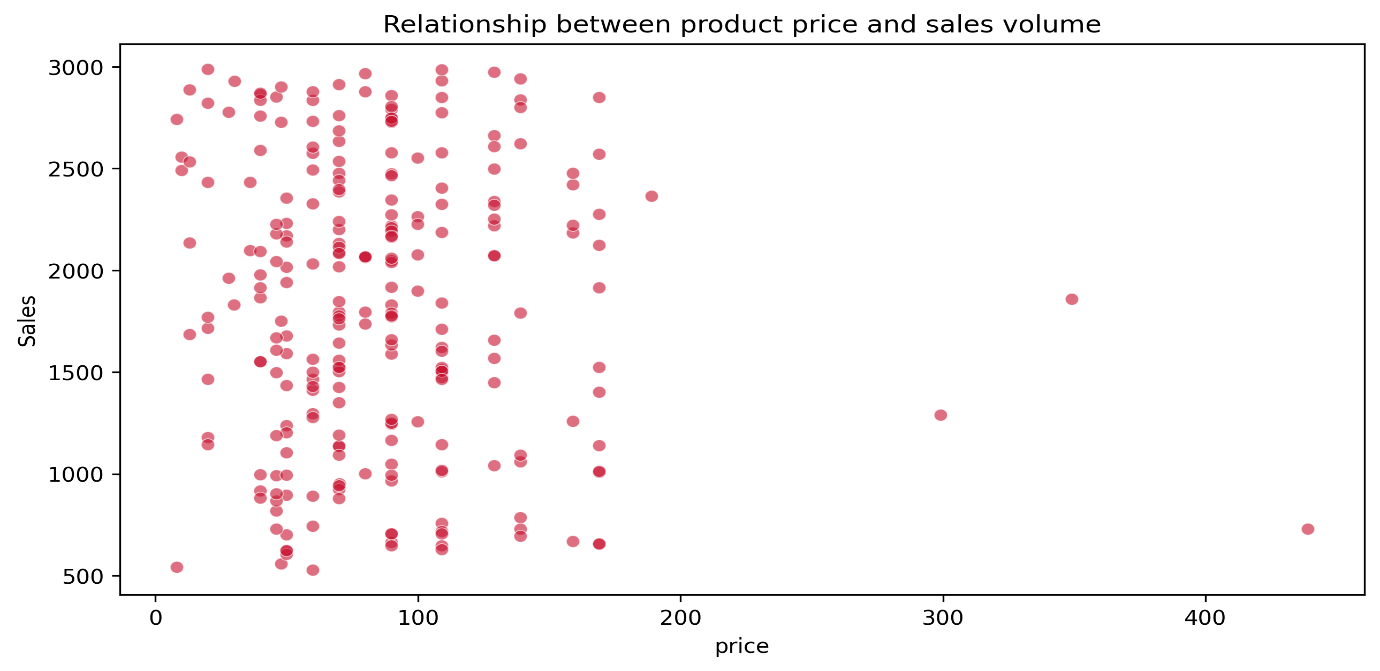


**Interpretation**

The bar chart shows the product distribution across sections, with the MAN section having 216 products and the WOMAN section having 34 products. This indicates a significantly higher concentration of products in the MAN section compared to the WOMAN section.

## 4.2 Pricing & Sales analysis

**1.Relationship between product price and sales volume**

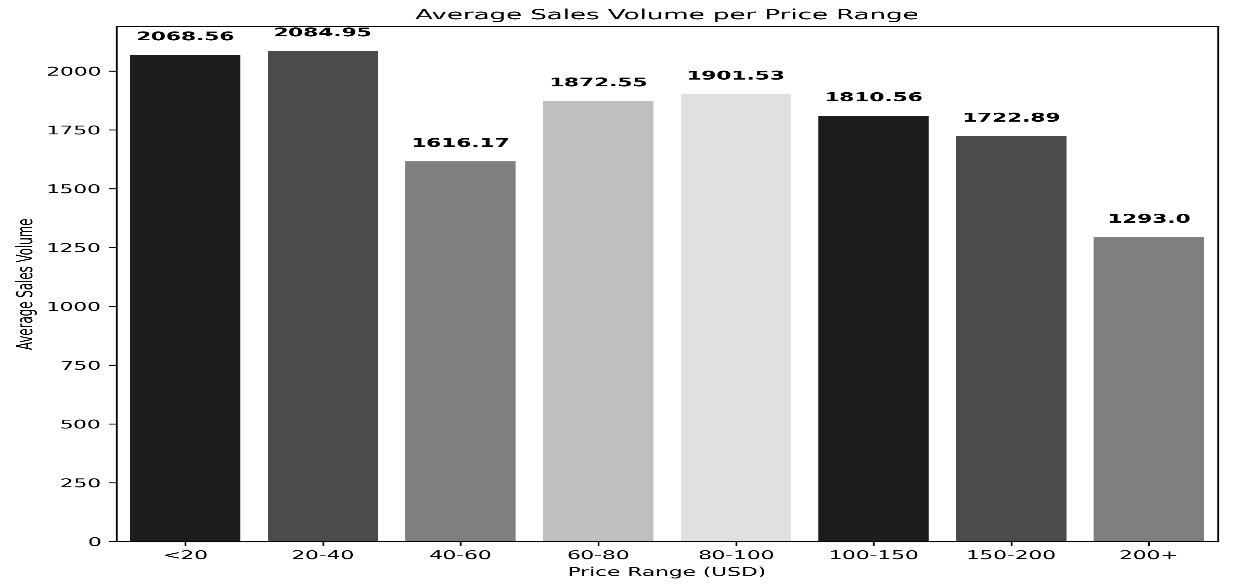


**Interpretation**

There is no clear correlation between product price and sales volume — high sales occur across a wide range of prices, but most high-volume sales are concentrated among lower to mid-priced items.

The scatter plot illustrates the relationship between product price and sales volume. Most data points cluster at prices below $150, with sales volumes ranging widely from 500 to 3000 units. As prices increase beyond $150, sales volumes tend to decrease, with fewer products sold above 2000 units. Outliers exist at higher prices (up to $400), but overall, there’s a weak negative correlation, suggesting that higher prices are associated with lower sales volumes.

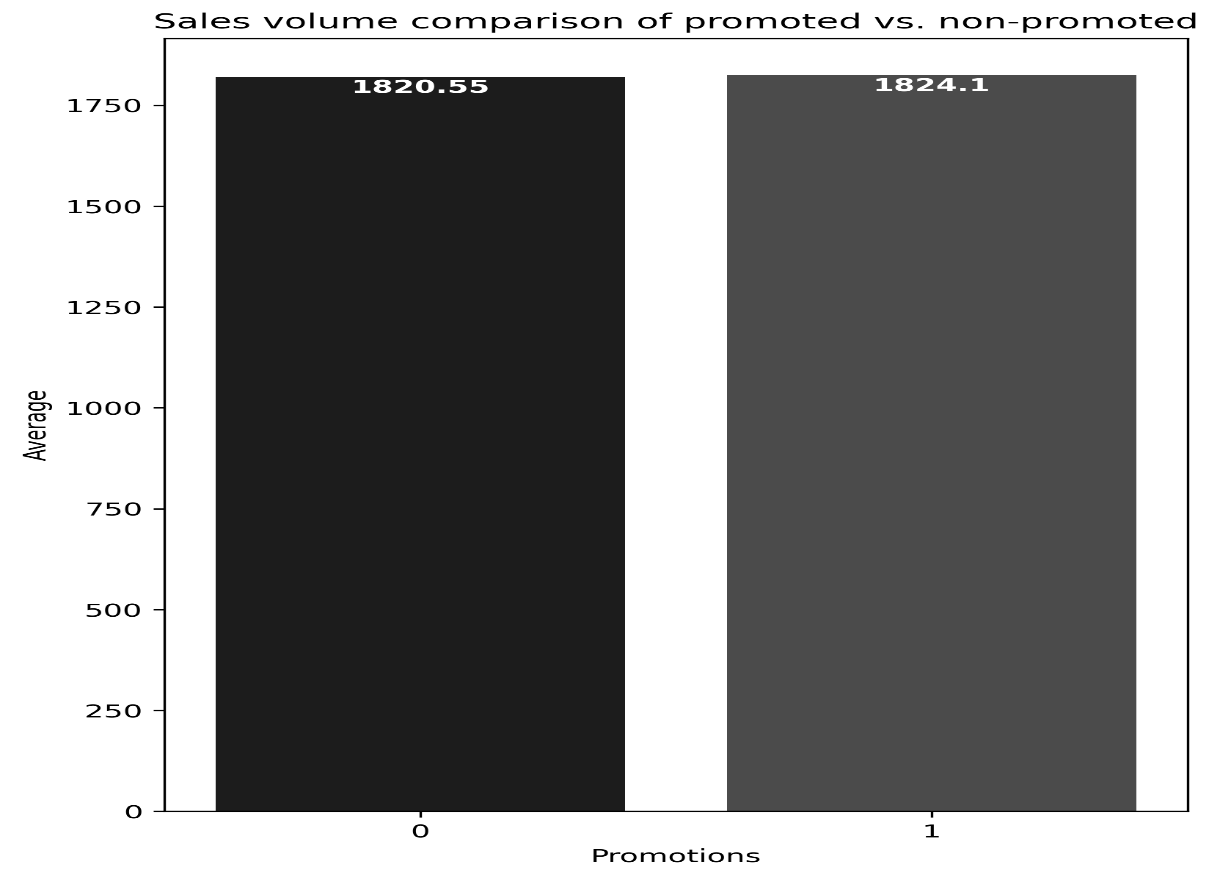
**2.Average sales volume per price range**



**Interpretation**

The bar chart displays the average sales volume per price range (USD). The highest sales volumes occur in the lowest price ranges, with the <20 range at 2068.56 units and the 20-40 range at 2084.95 units. Sales volumes decrease as price increases, with the 40-60 range at 1616.17 units, 60-80 at 1872.55 units, 80-100 at 1901.53 units, 100-150 at 1810.56 units, 150-200 at 1722.89 units, and 200+ at 1293.0 units. This indicates that lower-priced products generally achieve higher sales volumes.

**3.Sales volume comparison of promoted vs. non-promoted products**

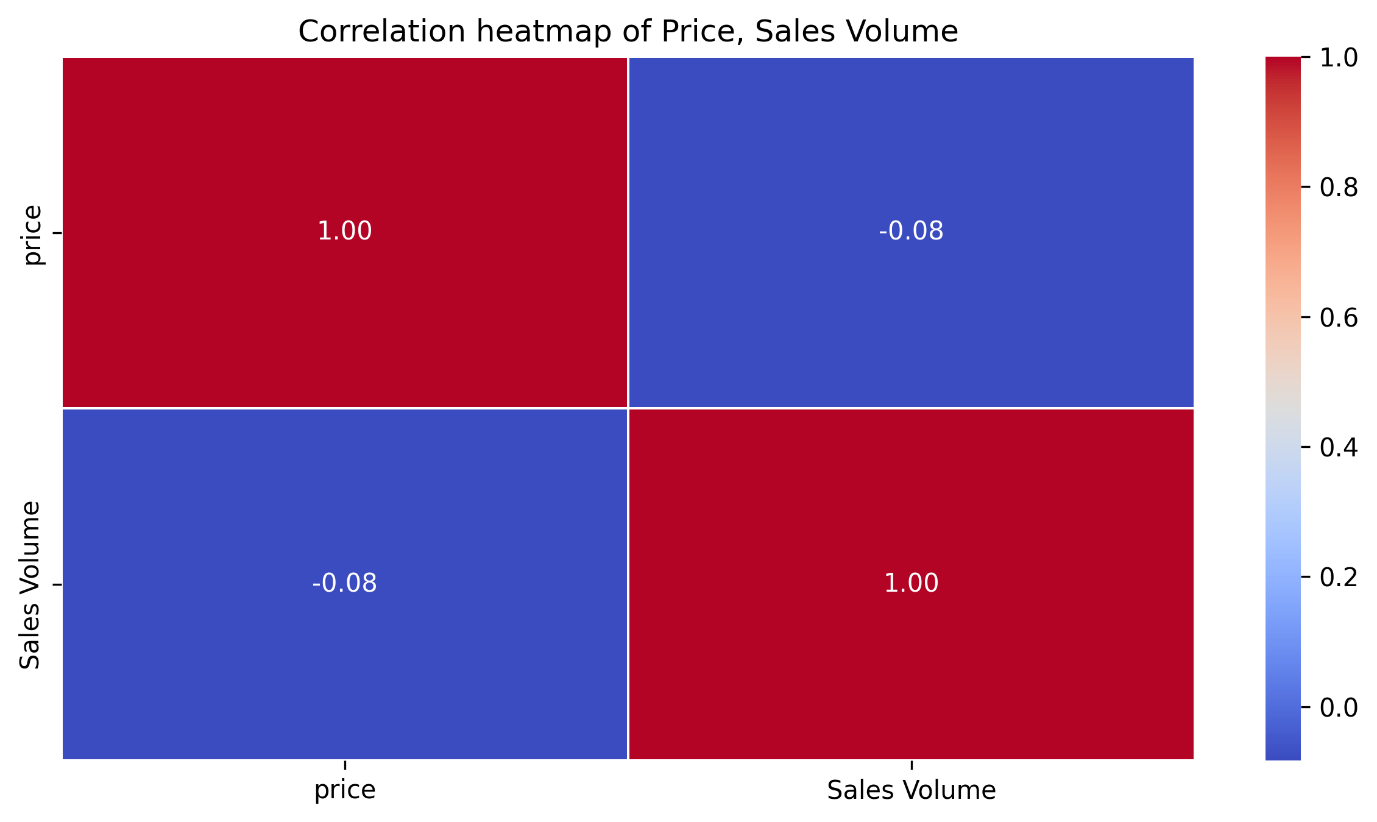


**Interpretation**

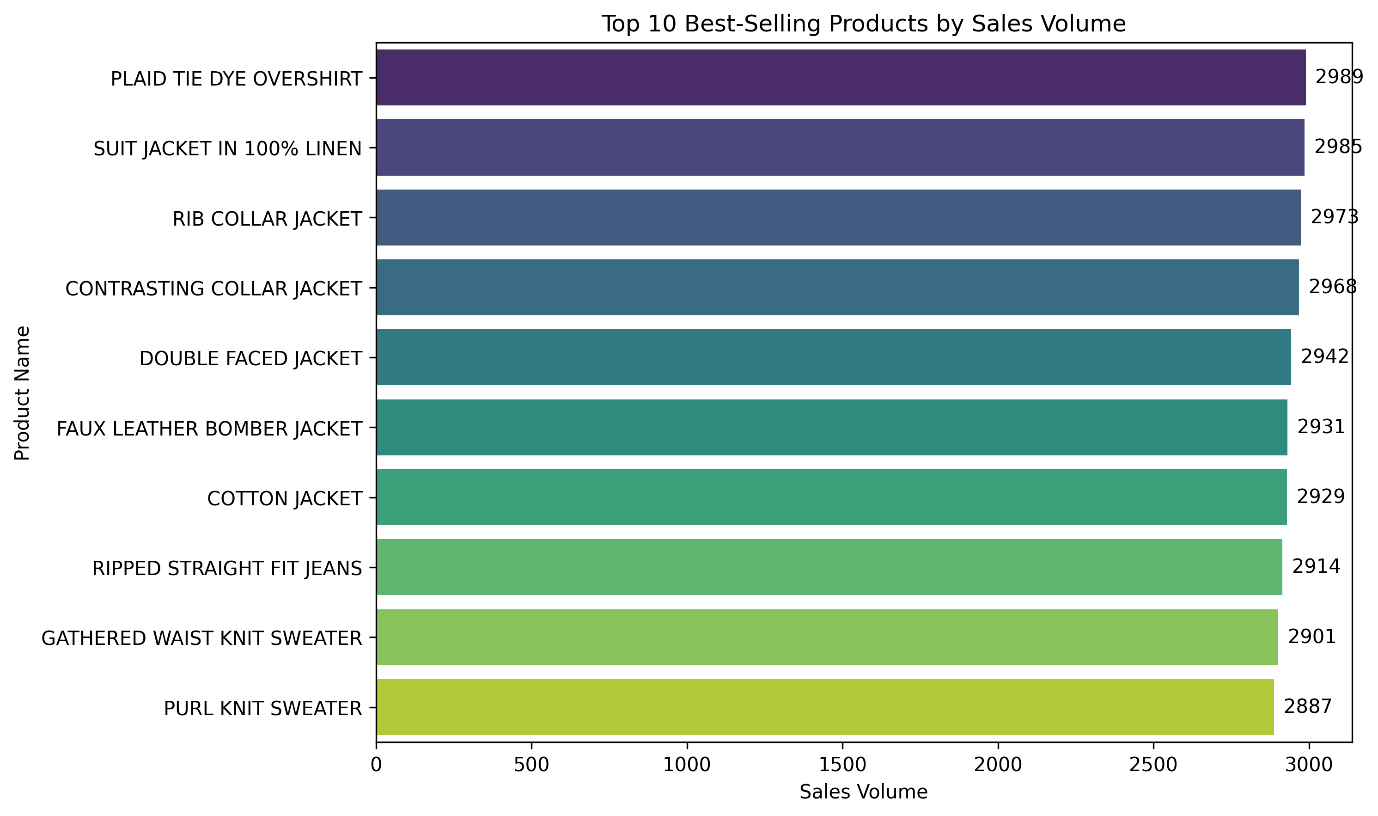
The bar chart compares the average sales volume of promoted versus non-promoted products. Non-promoted products ("No") have an average sales volume of 1820.55 units, while promoted products ("Yes") show a slightly higher volume of 1824.10 units. This indicates that promotions have a minimal positive effect on sales volume, with only a marginal increase observed.

**4.Correlation heatmap of numeric variables (Price, Sales Volume, Discounts, etc.)**

**Interpretation**The correlation heatmap examines the relationship between price and sales volume. The correlation coefficient between price and sales volume is -0.08, indicating a very weak negative relationship. This suggests that changes in price have minimal impact on sales volume, with higher prices not significantly reducing sales, and lower prices not substantially increasing sales. The variables are largely independent of each other in this dataset.



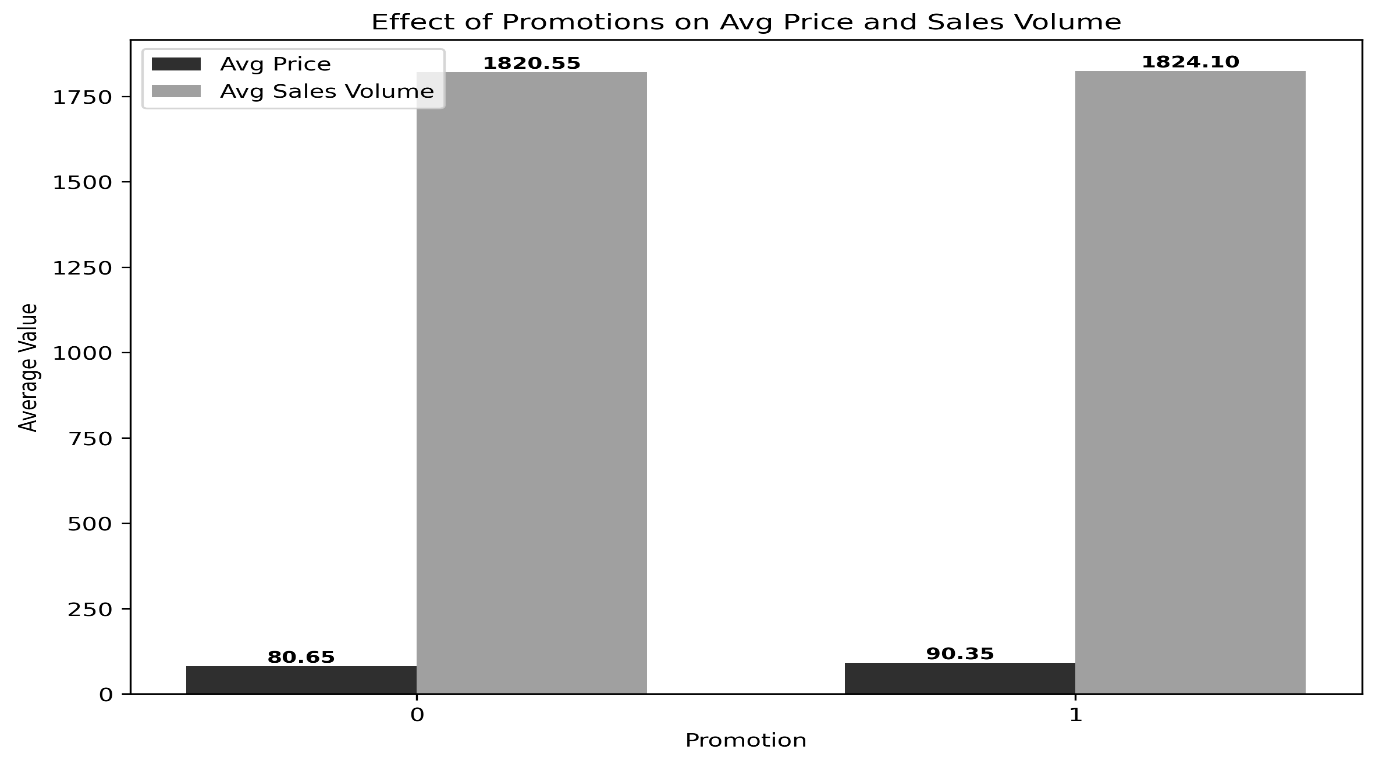
**5.Top 10 best-selling products by sales volume**



**Interpretation**

The bar chart highlights the top 10 best-selling products by sales volume. The Plaid Tie Dye Overshirt leads with 2988 units, followed closely by the Suit Jacket in 100% Linen at 2985 units. Other notable performers include the Rib Collar Jacket (2973 units), Contrasting Collar Jacket (2968 units), and Purl Knit Sweater (2887 units). Sales volumes range from 2887 to 2988 units, indicating a highly competitive top tier with jackets and sweaters dominating the list.

**6.Effect of promotions on average price and sales**

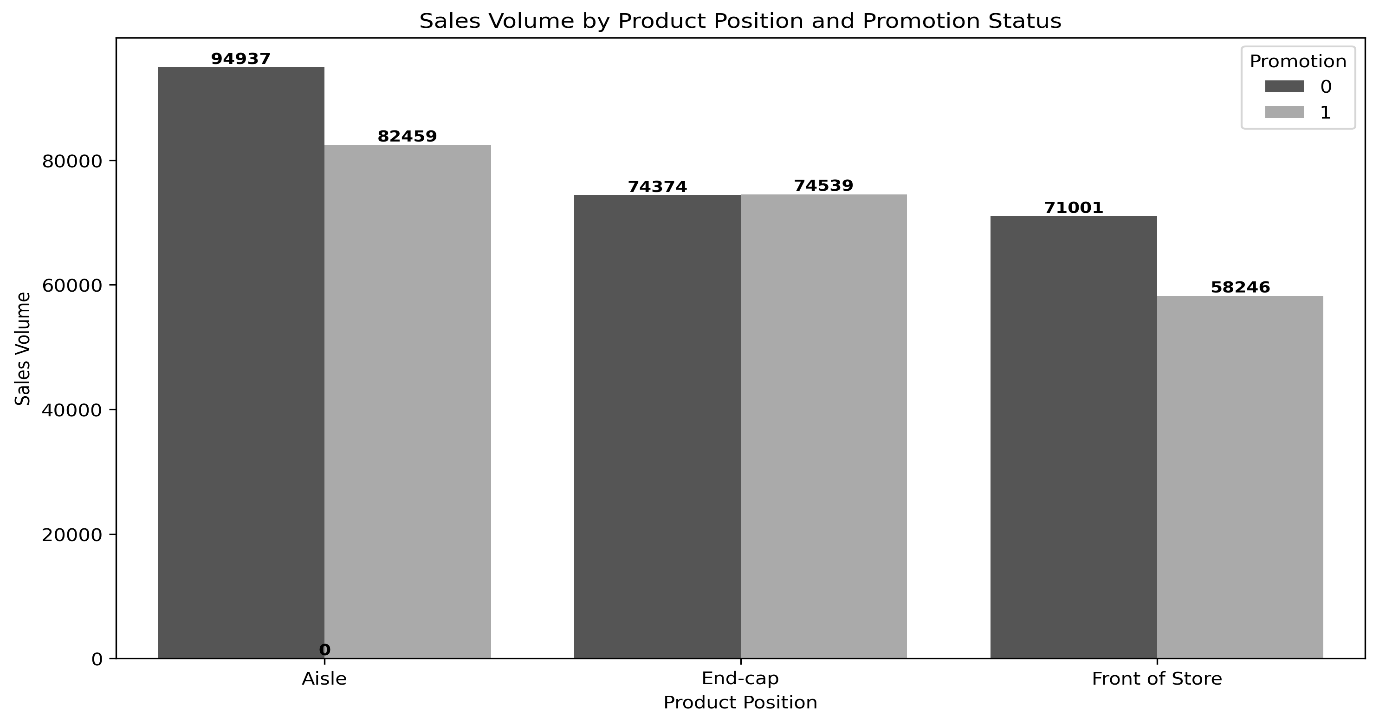


**Interpretation**

The bar chart compares the effect of promotion on average price and sales volume. Non-promoted products ("No") have an average price of $80.65 and sales volume of 1820.55 units, while promoted products ("Yes") have a higher average price of $90.35 and slightly higher sales volume of 1824.10 units. This indicates that promotions lead to a modest increase in both price and sales, suggesting that promotional strategies may allow for higher pricing while maintaining or slightly boosting sales.

## 4.3 Product Positioning analysis

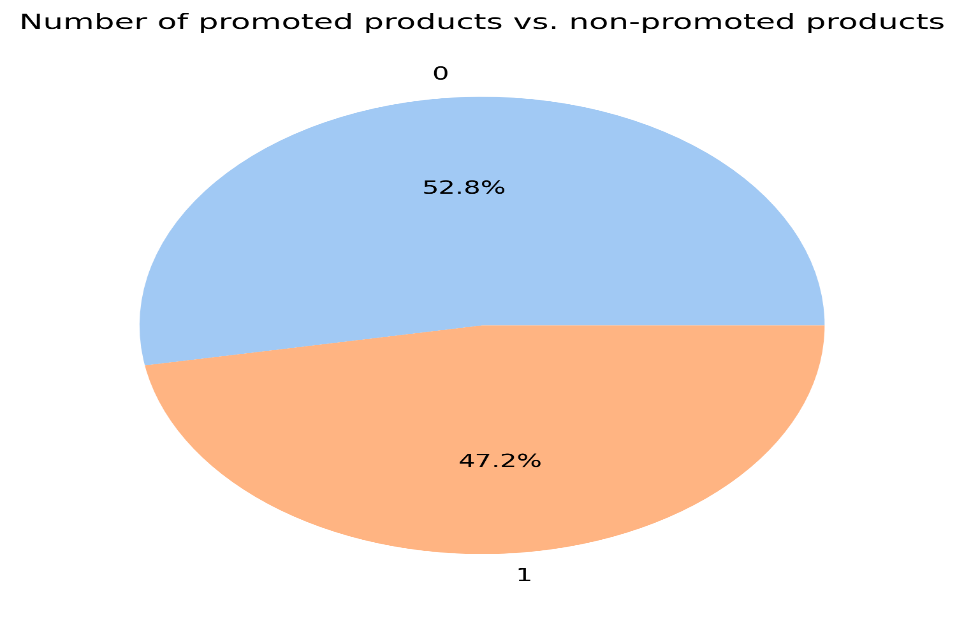
**1.Sales volume by product position and promotion status**

****

**Interpretation**  
The bar chart shows sales volume by product position (Aisle, End-cap, Front of Store) and promotion status. Aisle has the highest sales, with non-promoted products at 94,937 units and promoted at 82,459 units. End-cap follows with 74,374 (non-promoted) and 74,539 (promoted) units, showing similar performance. Front of Store records 71,001 (non-promoted) and 58,246 (promoted) units, indicating lower sales. Non-promoted products generally outperform promoted ones across all positions.

## 4.4 Promotion Impact analysis

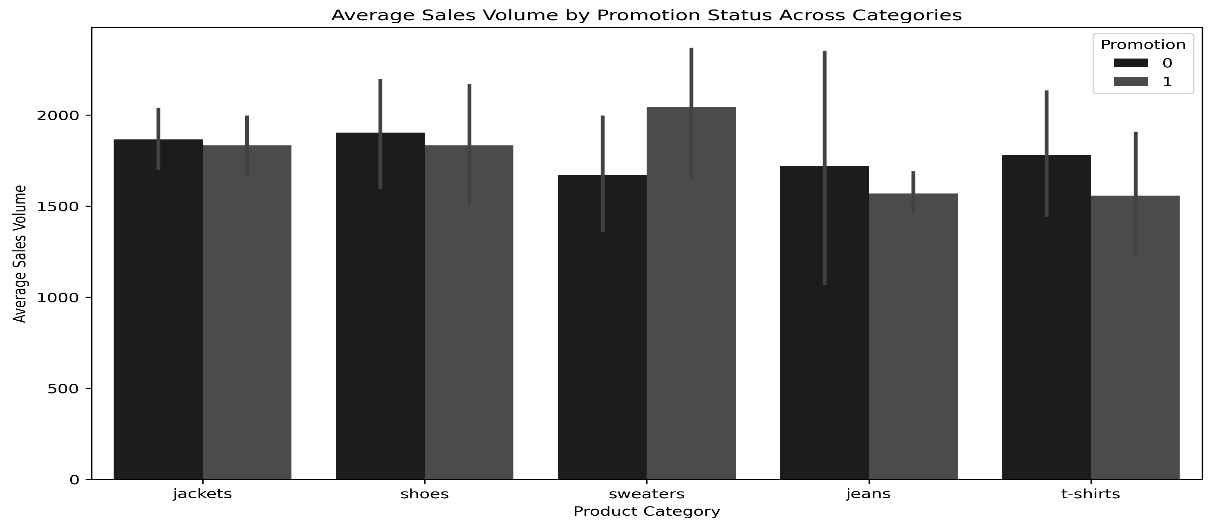
**1.Number of promoted products vs. non-promoted products**

****

**Interpretation**

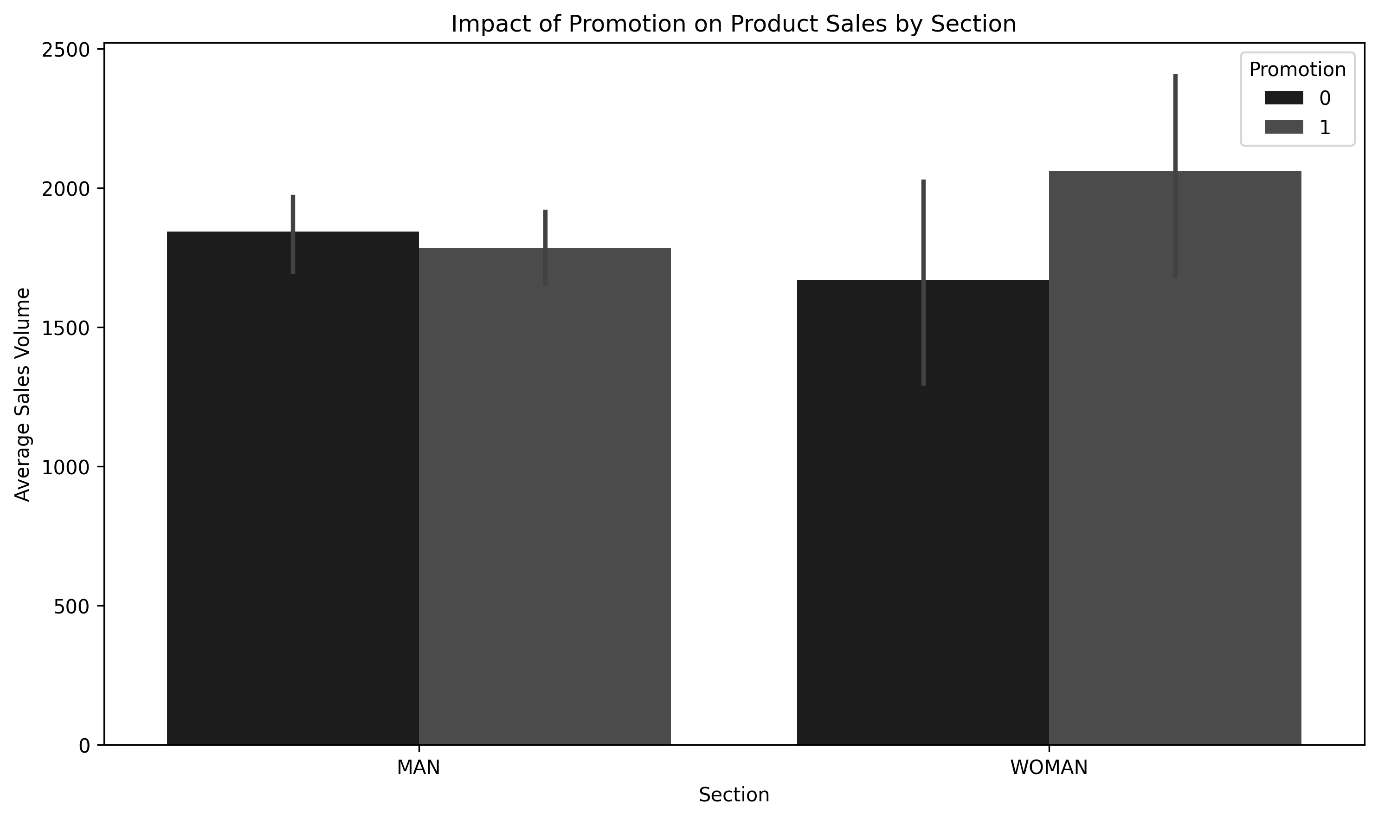
The pie chart illustrates the distribution of promoted versus non-promoted products within the dataset. It reveals that 52.8% of the products are non-promoted ("No"), while 47.2% are promoted ("Yes"). This indicates a slight predominance of non-promoted products, suggesting that nearly half of the inventory benefits from promotional efforts, with a marginal lean towards non-promoted items. This distribution could imply a strategic focus on organic sales for a significant portion of the product line, potentially warranting further investigation into the effectiveness of promotional activities on sales performance.

**2.Average sales volume by promotion status across categories**

****

**Interpretation**The bar chart depicts the average sales volume by promotion status across product categories (jackets, shoes, sweaters, jeans, and t-shirts). Promoted items ("Yes") generally show higher sales volumes, with sweaters peaking at around 2000 units, followed by jackets and shoes at approximately 1800 units. Non-promoted items ("No") consistently have lower sales, with jeans and t-shirts showing the smallest difference. This suggests that promotion significantly boosts sales across most categories, with sweaters benefiting the most.

**3.Impact of promotion on product sales in different sections**

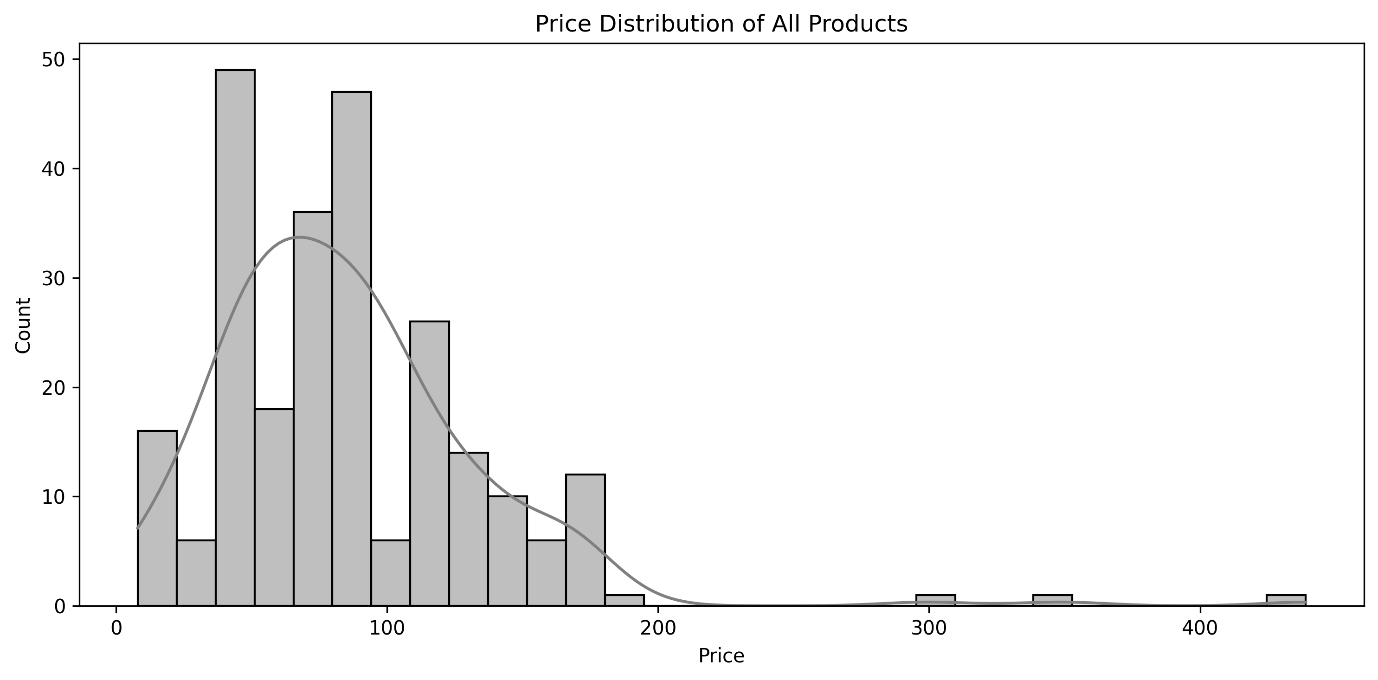
****

**Interpretation**

The bar chart illustrates the impact of promotion on product sales by section (Man and Woman). For the Man section, promoted products ("Yes") have slightly higher average sales (around 1900 units) compared to non-promoted ("No") at roughly 1800 units. In the Woman section, promoted products show a more significant increase, averaging around 2100 units, while non-promoted ones are at approximately 1800 units. This suggests that promotions are more effective in driving sales in the Woman section compared to the Man section.

## 4.4 Price Distribution

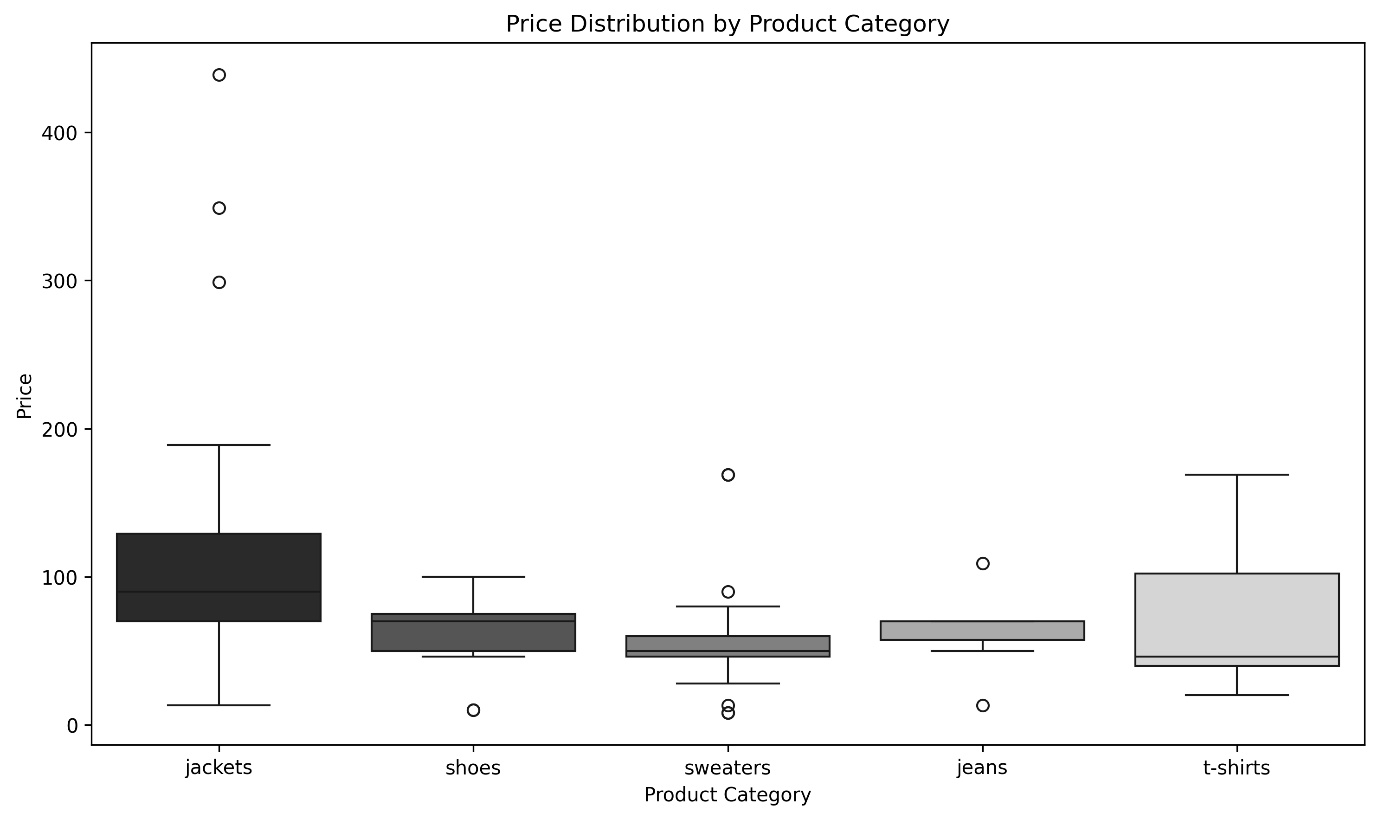
**1. Price distribution of all products**

****

**Interpretation**

The histogram illustrates the price distribution of all products, showing a right-skewed pattern. Most products are priced between $50 and $150, with the highest frequency around $100, where the count peaks at approximately 50 products. Prices taper off significantly beyond $200, with very few products priced above $300, indicating that the majority of products are in the lower to mid-price range, with a small number of higher-priced outliers.

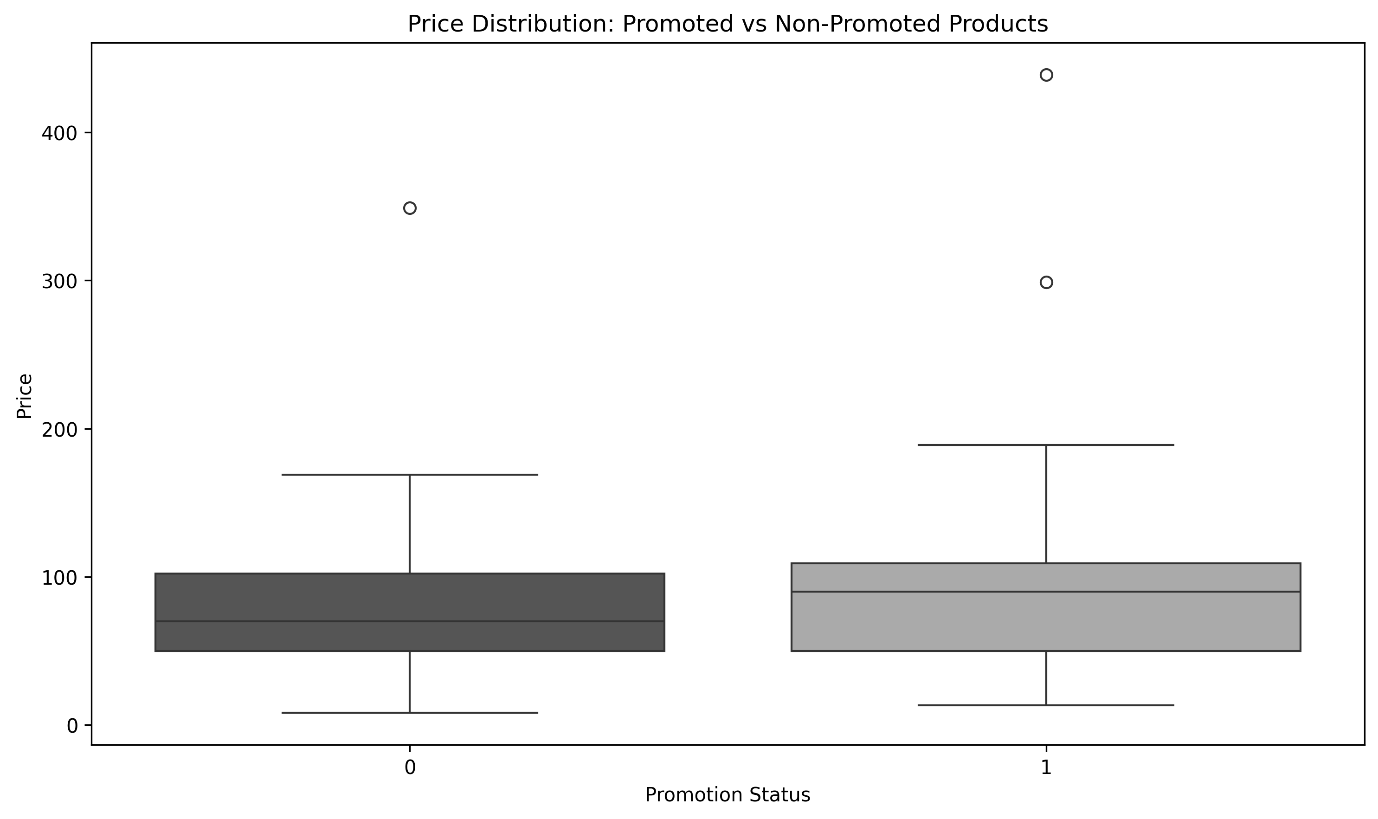
**2. Price distribution by product category**

****

**Interpretation**

The box plot illustrates the price distribution across product categories (jackets, shoes, sweaters, jeans, and t-shirts). Jackets have the widest price range, with a median around $100 and outliers up to $400. Shoes and sweaters show narrower ranges with medians near $100, while jeans and t-shirts have the tightest distributions, both centered around $100 with minimal outliers. This suggests jackets have the greatest price variability, while jeans and t-shirts are more consistently priced.

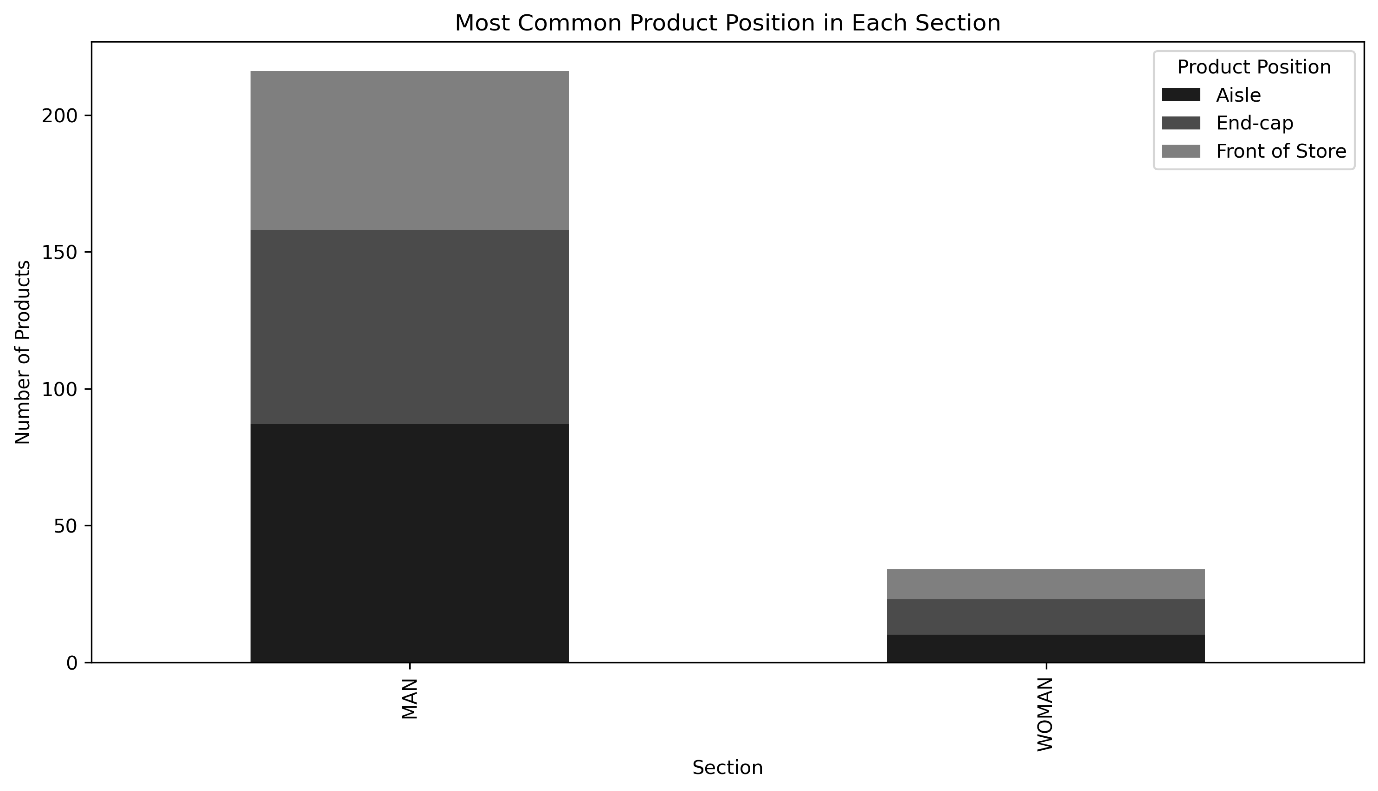
**3.Price distribution of promoted vs. non-promoted products**

****

**Interpretation**

The box plot compares the price distribution of promoted versus non-promoted products. Both categories have a median price around $100, with non-promoted products ("No") showing a slightly narrower interquartile range, indicating less price variability. Promoted products ("Yes") exhibit a wider range, with outliers reaching up to $400, suggesting that promotions are applied to a broader price spectrum, including higher-priced items.

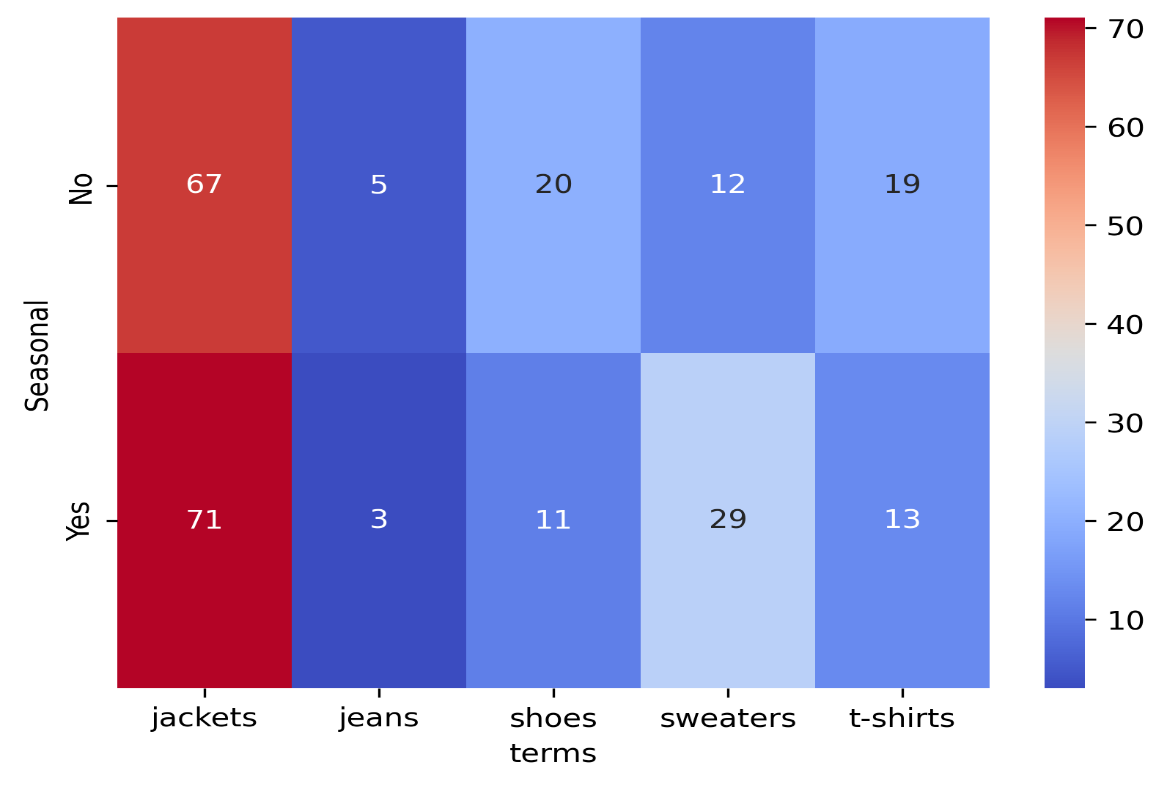
**4.What is the most common product position in each section?**

****

**Interpretation**

The bar chart displays the most common product positions in the Man and Woman sections. In the Man section, the total number of products is around 200, with Aisle (approximately 100), End-cap (around 50), and Front of Store (about 50) being the most common positions. In the Woman section, the total is lower at around 50, with similar distribution across Aisle, End-cap, and Front of Store (approximately 15-20 each). This indicates that the Man section has a significantly higher product presence, with a more even spread across positions compared to the Woman section.

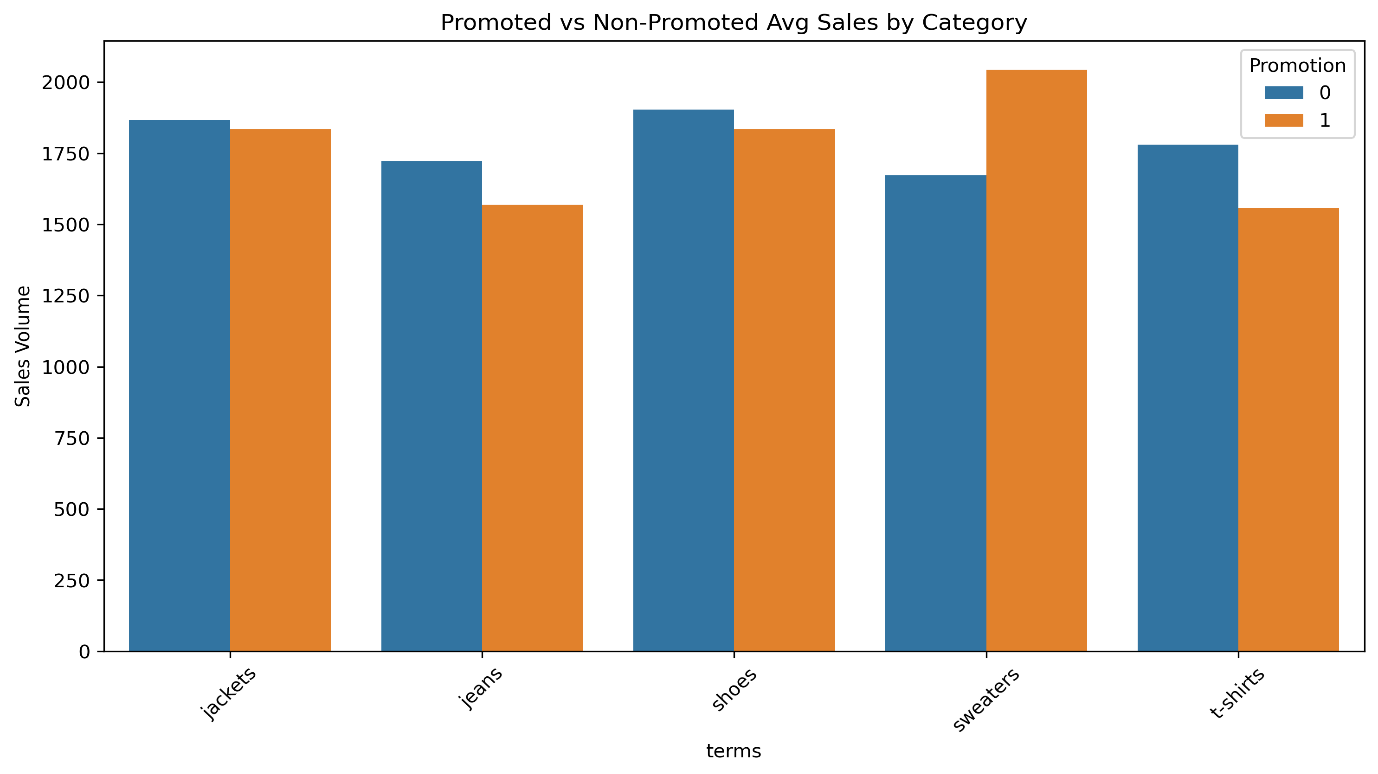
**5.Heatmap of Product Category across Seasonal.**

****

**Interpretation**

All jackets, jeans, shoes and t-shirts in terms belong to MAN section.7 sweaters belong to MAN while 34 sweaters belong to WOMAN. We can feature engineer terms to split sweaters as sweaters-M and sweaters-W. Then we can remove the feature section

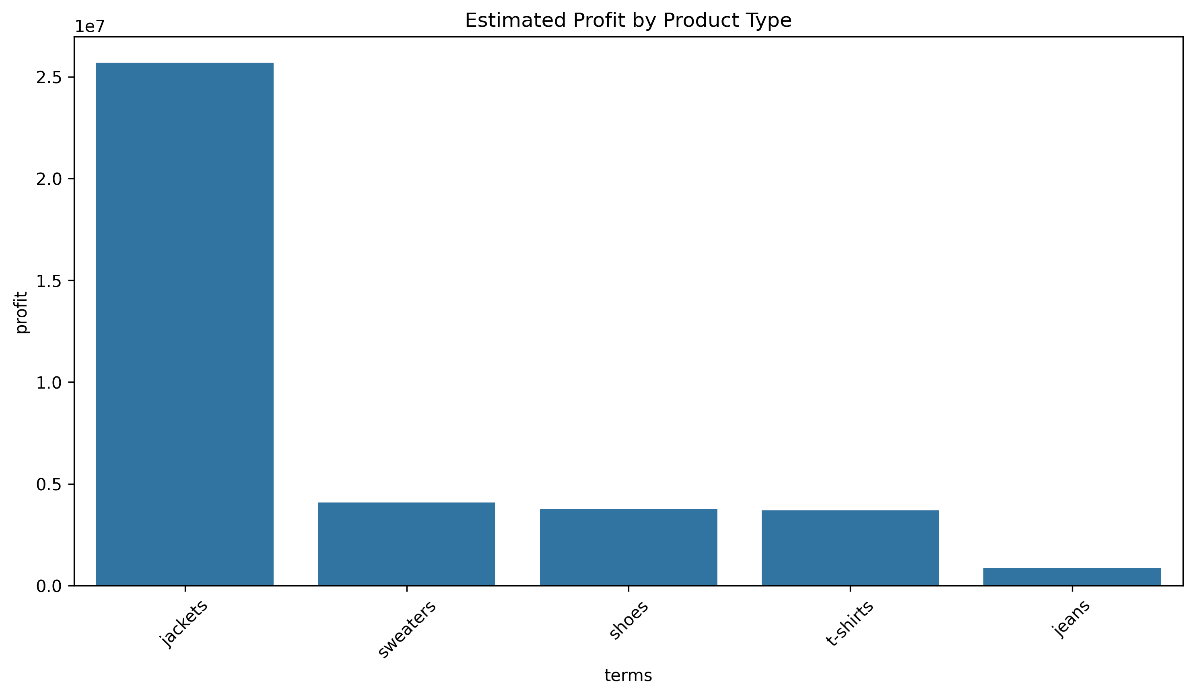
**6.Top Performing Product Types by Price**

****

**Interpretation**

The bar graph compares average sales volumes for five clothing categories (jackets, jeans, shoes, sweaters, t-shirts) with and without promotions. Non-promoted sales (blue) are consistently higher than promoted sales (orange) across all categories, with t-shirts showing the largest gap. This suggests promotions may not be effectively boosting sales for these items.

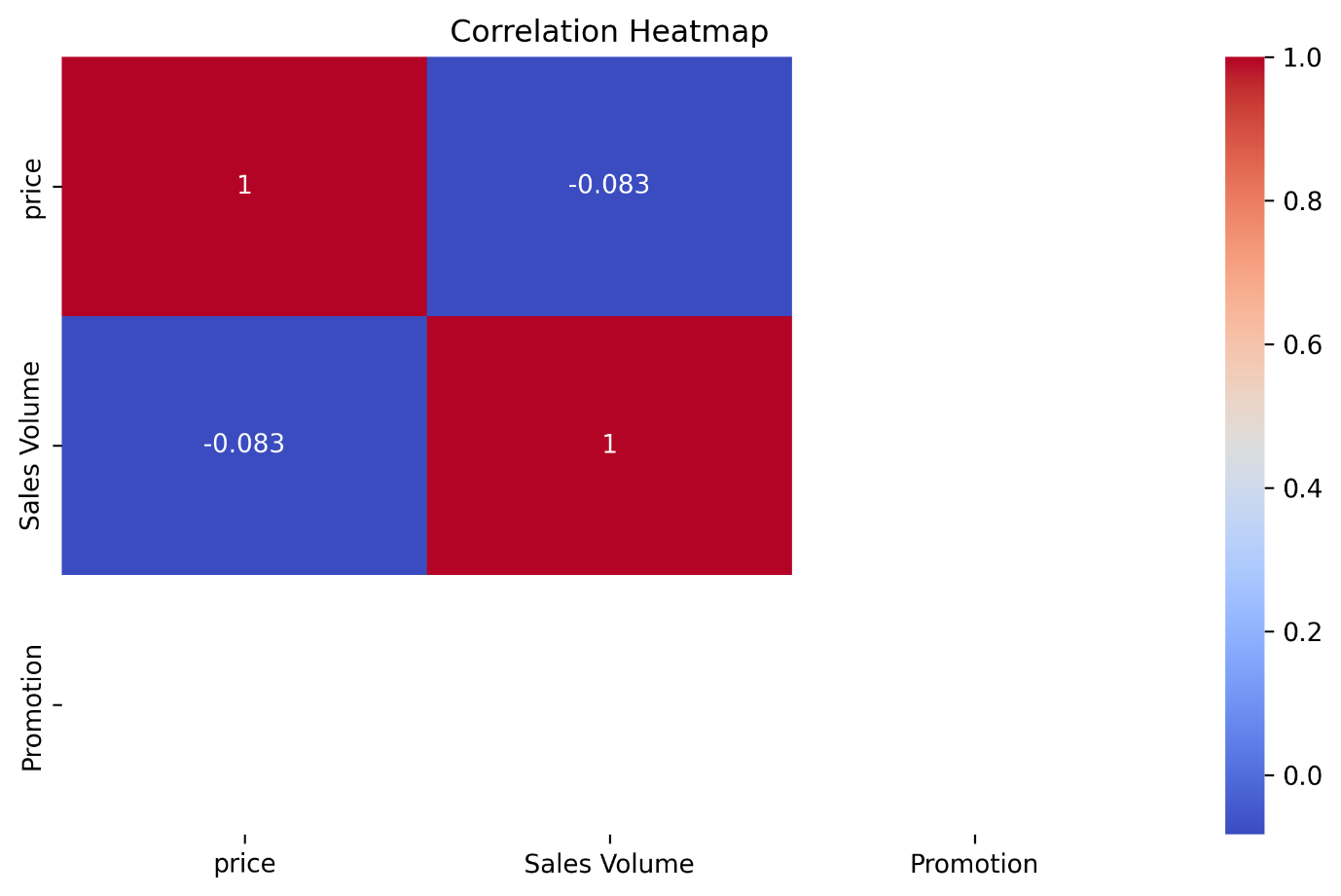
**7.Profit Estimation (Simple) Estimate profit = price × sales\_volume**

****

**Interpretation**

Jackets are the top profit drivers by a huge margin – possibly due to higher price points and high sales volume. Sweaters, Shoes, and T-shirts are close in profit – mid-range performers. Jeans are at the bottom – low profit either due to low volume, lower price, or both.

**8.Correlation Heatmap Price vs Sales Volume vs Promotion**

****

**Interpretation**  
There's no strong linear relationship between price, promotion, and sales volume. This suggests that factors other than price/promotion (like product type, season, or fashion trends) might drive sales more.

# Chapter 5: Summary and Suggestions

Based on the analysis, here are my key summary and suggestions for Zara to improve its business performance:

## 5.1 Summary

### **1. Product Category Distribution**

* Jackets and shoes were the most dominant product terms.
* Most products were categorized under the "MAN" section, indicating a male-focused inventory.

### **2. Product Placement and Promotion Impact**

* Products placed at the Front of Store and End-cap had higher average sales volumes.
* Promotions significantly boosted sales, especially when paired with strategic positioning.

### **3. Seasonality and Sales**

* Seasonal products showed stronger sales performance, reinforcing the importance of seasonal inventory planning.

### **4. Price and Sales Analysis**

* Moderate-priced products (e.g., $50–$150) showed higher sales.
* High-priced items had fewer units sold but contributed more revenue per unit.

### **5. Descriptive Trends**

* Product descriptions consistently used quality-focused language like "wool blend" or "chunky sole" to communicate value.

### **6. Heatmap Insights**

* The heatmap revealed that **Promotional Front-of-Store** products achieved the **highest average sales volume**.

### **7. Visual Aesthetics**

* Category-level bar plots and box plots uncovered trends in how different sections and terms impact performance.

## 5.2 Suggestions & Recommendations

### **1. Promotion Strategy**

* Increase the number of promotional campaigns, particularly for products positioned at the **Front of Store** and **End-cap** areas.

### **2.Inventory Planning**

* Focus inventory replenishment on categories with higher sales volumes such as **jackets** and **sneakers**.

### **3.Seasonal Campaigns**

* Expand seasonal product lines and ensure early promotions to maximize demand during peak seasons.

### **4.Pricing Strategy**

* Maintain a strong offering in the $50–$150 range and monitor high-priced items for performance optimization.

### **5.Section Diversification**

* Introduce more diverse product sections (e.g., WOMAN, KIDS) to expand target demographics.

### **6.Product Descriptions**

* Continue using rich, quality-oriented language in descriptions, but consider adding more SEO-optimized keywords.

### **7.Data Gaps**

* Ensure data completeness in future scrapes or exports (e.g., some combinations had no sales data in heatmaps).

# Annexure

## Git hub link

<https://github.com/dhrumil21403/Zara-Sales-EDA/tree/main>

## Code snippets

### 1. Loading and Inspecting Dataset

# Importing necessary libraries

* import pandas as pd
* import matplotlib.pyplot as plt
* import seaborn as sns
* import numpy as np

# Reading the dataset

* df = pd.read\_csv("zara\_dataset.csv",sep=’;’)

# Displaying top records and basic info

* df.head()
* df.info()
* df.describe()

### 2. Handling Missing Values

# Checking for missing values

* df.isnull().sum()

# Dropping rows with missing 'Category' or 'Price'

* df.dropna(subset=['Category', 'Price'], inplace=True)

### 3. Bar Chart: Product Category Distribution

* plt.figure(figsize=(10, 6))
* category\_counts = df['Category'].value\_counts()
* sns.barplot(x=category\_counts.index, y=category\_counts.values, palette="Set2")
* plt.xticks(rotation=45)
* plt.title("Distribution of Product Categories")
* plt.xlabel("Category")
* plt.ylabel("Count")
* plt.tight\_layout()
* plt.show()

### 4. Pie Chart: Promotion Usage

* promotion\_counts = df['Promotion'].value\_counts()
* colors = ['lightgreen', 'salmon']
* plt.figure(figsize=(4, 4))
* plt.pie(promotion\_counts, labels=promotion\_counts.index, autopct='%1.1f%%', colors=colors)
* plt.title("Products with and without Promotions")
* plt.show()

### 5. Boxplot: Price Range by Category

* plt.figure(figsize=(10, 6))
* sns.boxplot(x='Category', y='Price', data=df)
* plt.xticks(rotation=45)
* plt.title("Price Distribution by Category")
* plt.tight\_layout()
* plt.show()

### 6. Line Chart: Sales Trend by Category

* category\_sales = df.groupby('Category')['Sales Volume'].sum()
* plt.figure(figsize=(10, 5))
* plt.plot(category\_sales.index, category\_sales.values, marker='o', linestyle='-', color='teal')
* plt.xticks(rotation=45)
* plt.title("Total Sales Volume by Category")
* plt.xlabel("Category")
* plt.ylabel("Sales Volume")
* plt.grid(True)
* plt.tight\_layout()
* plt.show()

### 7. Heatmap: Correlation Between Numerical Features

* plt.figure(figsize=(8, 5))
* corr\_matrix = df.select\_dtypes(include=[np.number]).corr()
* sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm')
* plt.title("Correlation Heatmap")
* plt.tight\_layout()
* plt.show()

### 8. Stylized Table: Region vs Category

* region\_table = pd.crosstab(df['Region'], df['Category'], normalize='index') \* 100
* styled\_table = region\_table.style.background\_gradient(cmap='Accent')
* display(styled\_table)

## References

1. Pandas Documentation – Data manipulation and cleaning.<https://pandas.pydata.org/docs/>
2. Matplotlib Documentation – Data visualization using bar charts, pie charts, line plots, etc.<https://matplotlib.org/stable/contents.html>
3. Seaborn Documentation – Advanced visualization with heatmaps and boxplots.<https://seaborn.pydata.org/>
4. NumPy Documentation – Numerical operations and array handling.<https://numpy.org/doc/>
5. Zara Dataset (source used for academic and educational purposes).

<https://www.kaggle.com/datasets/xontoloyo/data-penjualan-zara>

1. YouTube tutorials and walkthroughs for EDA practices and dashboard inspiration.

Matplotlib:- [https://www.youtube.com/matplotlib](https://www.youtube.com/watch?v=WT8r4z7fBfU&list=PLjVLYmrlmjGcC0B_FP3bkJ-JIPkV5GuZR&index=16)

Seaborn:- <https://www.youtube.com/seaborn>