

UNIVERSITY OF ECONOMICS AND LAW
FACULTY OF INFORMATION SYSTEMS



FINAL PROJECT STAGE 2 REPORT
ARTIFICIAL INTELLIGENCE IN BUSINESS ANALYTICS COURSE
TOPIC: SALES PERFORMANCE ANALYSIS

Lecturer: Le Hoanh Su, PhD.

Group Live Laugh Love:

- 1. K214160989 – Tran Thi Minh Hien**
- 2. K214162148 – Cao Nguyen Hai Nhu**
- 3. K214162140 – Le Quoc Dan An**
- 4. K214162143 – Tran Hoang Anh**

Ho Chi Minh City, December 30th 2023

Members





NO.	Full name	Student ID	Point / 10 (Individual Contribution)	Signature
1	Tran Thi Minh Hien	K214160989	10	
			- Theoretical Basis - BCG Matrix - Chat GPT-4 Approach	
2	Cao Nguyen Hai Nhu	K214162148	10	
			- Methodology - Product Segmentation (ABC Analysis, Demand Stability)	
3	Le Quoc Dan An	K214162140	10	
			- Customer Segmentation - Buyer Persona	
4	Tran Hoang Anh	K214162143	10	
			- Data Preparation (Data understanding, EDA)	

Table of Contents

Project Overview	6
1. Background	7
1.1.Theoretical Basis	7
1.2.Methodology	9
2. Data Preparation	13
2.1.Data Understanding	13
2.2.Data Exploration.....	17
3. Product Segmentation	47
3.1 Overview.....	47
3.2.ABC Analysis	48
3.3.Demand Stability.....	48
3.4. Visualization	51
3.5. BCG Matrix	53
4. Customer Segmentation	55
4.1.Data Processing	55
4.2.Segmentation Result.....	58
4.3.Buyer Persona.....	61
5. Chat GPT-4.....	66
6. Conclusion	72
References.....	73

List of Figure

Figure 2-1 Distribution of Net Price Category	17
Figure 2-2 Distribution of Sold Quantity Categories	18
Figure 2-3 Distribution of Price Group	19
Figure 2-4 Distribution of Product Group	20
Figure 2-5 Distribution of Size Group	21
Figure 2-6 Distribution of Age Group	22
Figure 2-7 Distribution of Activity Group	23
Figure 2-8 Distribution of Lifestyle Group	24
Figure 2-9 Distribution of Color Group	25
Figure 2-10 Profitability Distribution Chart.....	26
Figure 2-11 Top-selling products Chart.....	27
Figure 2-12 Sales Trend Over Time	28
Figure 2-13 Scatter Plot between Net price and Quantity.....	29
Figure 2-14 Total Sold Quantity and Revenue in Distribution Channels	30
Figure 2-15 Distribution of Sales Across Different Regions	31
Figure 2-16 Average Net Price by Urbanization Level	32
Figure 2-17 Sold Quantity and Revenue by Gender and Price Group	33
Figure 2-18 Sold Quantity and Revenue by Brand and Product Group.....	36
Figure 2-19 Sales Across Distribution Channels.....	39
Figure 2-20 Sold Quantity and Revenue by Gender and Age Group.....	41
Figure 2-21 Sold Quantity and Revenue by Gender and Lifestyle	44
Figure 3-1 Pareto Chart for Product Category and Subcategory	47
Figure 3-2 Demand Stability Result.....	49
Figure 3-3 Product Contribution by Class.....	52
Figure 3-4 Cumulative Contribution by Class	52
Figure 3-5 BCG Matrix Explanation.....	53
Figure 4-1 Distribution of RFM Metrics.....	56

Figure 4-2 RFM Scatter Plot Pre-transformation.....	56
Figure 4-3 RFM Scatter Plot Post-transformation	57
Figure 4-4 Elbow Result.....	57
Figure 4-5 The mean RFM value of clusters.....	58
Figure 4-6 A buyer persona of High-value customer cluster, N. J. Keria	61
Figure 4-7 A buyer persona of Lapsed customer cluster, NS Hillison	62
Figure 4-8 A buyer persona of Potential customer cluster, M. Adam	63
Figure 4-9 A buyer persona of Loyal customer cluster, A. Arnold	64
Figure 5-1 Result of Fragment 1	67
Figure 5-2 Result of Fragment 2	68
Figure 5-3 Result of Fragment 3	68
Figure 5-4 Result of Fragment 4	69
Figure 5-5 Result of Fragment 5	69
Figure 5-6 Result of 6 Fragments.....	70
Figure 5-7 Chat GPT-4 Analysis	70
Figure 5-8 Chat GPT-4 Limitation	71

Project Overview

Reasons

As a result of this cutting-edge era, businesses have gradually begun to utilize data to solve business cases in order to maximize their profits. Most retail businesses in Vietnam, however, continue to use manual and traditional model to propose business strategies. Not only does this take time and effort, but it also does not guarantee accuracy and dependability. The emergence of the digital transformation era is a significant step forward, allowing businesses to apply scientific and technological advances to their operations. This is where machine learning and artificial intelligence comes into play.

Objectives

The project aims to present sales performance analysis in the retail industry, providing insights for managerial decision-making, as well as ensuring the accuracy of the results in order to increase revenue for the business. The models presented in this project is suitable for retail businesses as well as those undergoing digital transformation. After analysis, suitable business strategies will be proposed to enhance business' sales performance.

Objects and scope

Object: Sales and Inventory data from a business

Time scope: 01/2022 – 07/2023

Space scope: a number of retail businesses in HCMC, Vietnam

Project structure

This project include ... figures and ... tables. The report is divided into ... parts, including: ...

1. Project Background

1.1. Theoretical Basis

1.1.1. ABC Analysis

ABC analysis is one of the most commonly employed inventory classification techniques. Conventional ABC classification was developed for use by General Electric during the 1950s. The classification scheme is based on the Pareto principle, or the 80/20 rule, that employs the following rule of thumb: “vital few and trivial many.” The process of ABC analysis classifies inventory items into A, B, or C categories based on so-called annual dollar usage. Annual dollar usage is calculated by multiplying the dollar value per unit by the annual usage rate (Cohen & Ernst, 1988; Partovi & Anandarajan, 2002). Inventory items are then arranged according to the descending order of their annual dollar usage. Class A items are relatively small in number, but account for the greatest amount of annual dollar usage. In contrast, class C items are relatively large in number, but make up a rather small amount of annual dollar usage. Items between classes A and C are categorized as class B.

1.1.2. BCG Matrix

The BCG matrix (BCG Product portfolio Analysis model, BCG product portfolio matrix or BCG Growth Share Matrix) is a product portfolio analysis model created by Henderson for the Boston Consulting Group to help enterprise make resource allocation decisions related to their SBUs/products portfolio strategies (Henderson, 1979). BCG Growth-Share Matrix is plotted on a two-dimensional four celled grid (2×2 matrix). Using the BCG grid, a company classifies all its SBUs/Products according to two dimensions: (1) on the horizontal axis is the Relative Market Share (RMS)—this serves as a measure of product strength (competitive advantage) in the market. The market share of the product in the market is defined as compared to Rule-Based BCG Matrix for Product Portfolio Analysis its competitors and overall product/category; (2) on the vertical axis is the Market Growth Rate (MGR)—this provides a measure of market attractiveness.

1.1.3. Elbow

The number of k clusters to be created is specified by the K-means algorithm. It determines how efficient the algorithm is (Hong-Dien, Phuc-Son, Hoang-Uyen, & Van-Hinh, 2019). In order to determine the optimal number of clusters in the analytical data set, our team used the Elbow method. The process includes plotting the explained variation as a function of cluster count and selecting the curve's elbow as the number of clusters to use.

The Sum of Squared Errors (SSE) value is calculated by adding the values of k one by one:

$$SSE = \sum_{K=1}^K \sum_{x_i \in S_K} \|X_i - C_k\|_2^2$$

SSE is the sum of each point's average Euclidean Distance from the centroid. The value of k is discovered when the value drops dramatically and forms a smaller angle. For example, start with $k = 2$ and gradually increase the SSE value, where $k_n = k + 1$, the largest $SSE_{k_n} - SSE_{k_{n-1}}$ is the point at which the optimal k value is found. When k is re-added, the new Cluster is similar to the previous Cluster or the number of errors does not change significantly, resulting in the value of k (Marutho, Handaka, Wijaya, & Muljono, 2018).

1.1.4. Kmeans

The K-means algorithm is applied in labeling data points based on the characteristics of clusters where each Cluster is represented by a point called the center, the center is the most representative and characteristic point for a cluster. Furthermore, is equal to the average of all observations in each Cluster (S.Khan & Ahmad, 2004). Initially, if the number of cluster centers is not given, the algorithm will randomly generate the number of cluster centers with the number of cluster centers not more significant than the number of data points. Then proceed to label the data points and redefine the number of cluster centers, the algorithm will stop when the number of centers is larger than the number of data points or all data points have been labeled. This process takes place in turn in 2 stages as follows (Dinh-Khanh, n.d.):

- (1) Initialize a random number of k cluster centers such that no more than the number of data points. In this algorithm, the centers are usually expressed $\mu_1, \mu_2, \mu_3, \dots, \mu_k$. Call X_1, X_2, \dots, X_N as the data points with $i \leq N$, j is the number of the Cluster such that $j \leq k$, C_i is the label where the data point is attached to $i \leq N$ and i is the sequence number of the data point.
- (2) Conduct clustering and repeat the process of determining the center of the Cluster:
 - a. Define labels for C_i data points based on the range of data points to the center of the Cluster using the algorithm:

$$c_i = \arg \min_j \|X_i - \mu_j\|_2^2$$

- b. Recalculate the center for each Cluster based on data from points located in the same Cluster, specifically:

$$\mu_j := \frac{\sum_{i=1}^n 1(c_i = j)X_i}{\sum_{i=1}^n 1(c_i = j)}$$

Typically, the number of k cluster centuries is determined through the Elbow method so that the number of clusters is the most optimal, and the K-means algorithm is used to find and label data points according to the characteristics based on that k cluster center. The labeling process will work according to the mechanism that the value of μ_j will be returned as 1 if C_i belongs to the cluster j and vice versa, at this time j is the number of the Cluster to which C_i belongs.

The K-means clustering algorithm is considered one of the most influential and popular data mining algorithms in the research community (Ahmed, Sera, & Islam, 2020).

1.2. Methodology

1.2.1. AI Approach

The AI approach, specifically employing the Chat GPT-4 model, initiates its process with the provisioning of datasets and an exploratory data analysis (EDA) file by the team. As a result, Chat GPT-4 greatly contributes to the methodical investigation of the offered material by

leveraging its powerful Data Analysis mode and taking into account the exact needs expressed within the provided prompts.

During the first phase, our team ensures the availability of comprehensive datasets containing data related to the given assignment. Our EDA file serves as a critical foundation for the further analysis performed by Chat GPT-4. This advanced model then incorporates its Data Analysis mode, a tool designed for in-depth investigation and comprehension of the presented data.

Furthermore, the effectiveness of the Chat GPT-4 technique is tightly connected to the details of the prompts supplied. These prompts serve as a framework for the analysis, explaining the precise requirements and expectations. By responding to these prompts, Chat GPT-4 refines its focus, ensuring that the succeeding phases in the process are tailored to the specific requirements of the particular work.

The interaction of the dataset, EDA file, and the model's Data Analysis mode is critical to gaining a comprehensive knowledge of the data. This thorough technique eliminates needless complexity while maintaining the analytical process's coherence and organization.

In short, the Chat GPT-4 methodology is a powerful and efficient approach that leverages datasets, expertly crafted EDA files, and advanced Data Analysis capabilities to conduct a systematic and coherent examination of provided data, all guided by the specific requirements outlined in the prompts.

1.2.2. Traditional Approach

- ***Data Collection***

The Python, or, traditional approach initiates with the collection of data relevant to the analysis, which is about sales and inventory of a retail business in HCMC.

- ***Data Preprocessing***

This approach takes over the data preprocessing phase. Python scripts clean and standardize various flat files, addressing issues such as missing values and outliers. This step assures that the datasets, regardless of differences in structure and content, are ready for unified analysis.

- ***Exploratory Data Analysis (EDA)***

This step visually explores and interprets the datasets' features using Python packages. Python's flexibility enables dynamic visualization of patterns, correlations, and anomalies while taking into account the specific characteristics of each department's data.

- ***Data Modeling***

Selecting ML models that correspond to the specific patterns observed during the EDA phase.

- ***Visualization***

Creating visual representations of data trends and patterns, aiding in a comprehensive understanding.

- ***Evaluation***

Python scripts are used to implement, train, and evaluate the selected ML models on standardized datasets. The models are fine-tuned to capture the unique characteristics of each set of data, and their performance is measured using appropriate criteria.

The AI approach makes use of Chat GPT-4's Data Analysis mode to conduct the analysis based on the dataset, the EDA data, and the provided prompts. The Traditional approach, on the other hand, follows a systematic sequence from data collection, preprocessing, and EDA to modeling, visualization, and evaluation. The combination of both methodologies enables a comprehensive sales performance analysis strategy, delivering a deeper insight and increased project maintainability.

When the AI approach is compared to Chat GPT-4 and the classic Python approach, it is clear that each methodology has advantages. While AI provides creative assistance, it may encounter difficulties in circumstances of complicated data and intricate interactions between

factors. ChatGPT may not deliver the intended results in such cases. When combined with processed files, however, it exhibits the capacity to build basic EDA graphs, supplementing the old approach and adding to a more complete sales performance analysis strategy. The interaction of the two techniques improves overall efficacy by offering a thorough and nuanced understanding of the data.

2. Data Preparation

2.1. Data Understanding

The data used by the group related to business data of products in the footwear industry includes 3 datasets, of which 2 main datasets are sales and inventory data and the remaining is master data. With the sales and inventory dataset, the dataset includes many files but the files all have the same number of columns and data types. Master data is a bit different. This dataset includes many files with different numbers of columns and data types, but most of them serve as master data for a specific category such as COGS (Cost of Goods Sold), Products, calendar, prices,... For a better understanding, details about the datasets will be presented below.

Sales data:

This dataset contains files related to recorded sales data, including month, week, site, branch, channel, distribution channel, sold price, net price, customer, and product. The specific meaning of these columns can be understood as follows:

- **month:** Indicates a specific month, formatted as a numerical code.
- **week:** Refers to a specific week, also formatted as a numerical code.
- **site:** Site code or identifier.
- **branch_id:** Identifies different branches, likely of a company or organization.
- **channel_id:** Describes the channel, for example, "Online" or "Offline(CHTT)".
- **distribution_channel:** A more descriptive form of the distribution channel, like "Online" or "Bán lẻ" (which translates to "Retail" in English).
- **distribution_channel_code:** A coded form of the distribution channel, like "ZF2" or "FP".
- **sold_quantity:** The number of items that have been sold.
- **cost_price:** The cost price of the items.
- **net_price:** The selling price of the items.

- **customer_id:** Unique identifiers for customers.
- **product_id:** Unique identifiers for products, including a combination of letters, numbers, and a product code.

Inventory data:

Similar to Sales data, Inventory data provides detailed information about inventory data in the warehouse. These data relate to product inventory time, quality, and product information. Detailed information about this dataset can be understood as below:

- **Unnamed:** Appears to be an auto-generated index column by Excel, typically used for internal tracking.
- **index:** Index column, possibly indicating the row number or a specific record identifier.
- **plant:** This column contains numerical codes, possibly identifying different plants or production facilities.
- **calendar_year:** The year for the data record, which in this case is 2022.
- **calendar_year_week:** A numerical code likely representing a specific week in the calendar year, formatted as YYYYMMDD.
- **Sloc:** This could stand for "Storage Location" or something similar, denoted by a numerical code.
- **quantity:** The number of items in stock.
- **total_amount:** The total value or amount of the stock, which appears to be 0 in the first few rows.
- **product_id:** Unique identifiers for products, consisting of a combination of letters, numbers, and a product code.

Master data:

Unlike the above two datasets, the master dataset will not contain information in a multidimensional way, but it will serve as a dataset that describes more detailed information for a subject such as price, product, distribution channels, ...

- **COGS (Cost of Goods Sold):** This table contains data about the purchase price of a product, including the columns 'index', the purchase price of the product 'amount', 'valid from' and 'valid to' meaning the validity period of the product's import price. product, 'product_id' represents the product's information.
- **Distribution Channel:** In general, the Distribution_channel table contains information about distribution channels, in which numerical data includes:
 - 'index' and 'site_store' can contain information about the index and information about the store code.
 - *Categorical variable:* 'b2b_b2c', 'channel_id', 'region', 'city_level', 'store_concept', 'trade_term', 'area_range', 'store_type', 'urbanization', 'branch_area', 'start_month', 'start_year', 'end_month', và 'end_year'.
 - *Variables with text data include:* 'address_2', 'address_3', 'note', and 'customer_name'.
 - *The variable can be a foreign key:* 'customer_id'
 - *Variables with time data:* 'start_month', 'start_year', 'end_month', 'end_year'.
- **Master Calendar:** The Master Calendar table contains time information including 8 columns with 260 rows. This table acts as a Dim Time table when building a data warehouse.
- **New Core Classification:** This table contains information about the business activities of products. Specifically:
 - 'index': Index of row in DataFrame.
 - 'launch_season': Launch season
 - 'lauch_season_num': Number representing the launch season (can be a number or an assigned number).
 - 'sales_season': Sales season.
 - 'sales_season_num': Number representing the sales season.
 - 'final_status': The final status of a product
 - 'b2c_assortment': Whether the retail (B2C) format of the product is available

- *'b2b_assortment'*: Whether the wholesale (B2B) version of the product is available
- *'total_assortment'*: Total number of product variants available (can be a binary variable or can have a value other than 1.0).
- *'product_syle_color'*: Number or code representing the product style and color.

In there:

- Column *'lauch_season'* is missing values (NaN).
- The *'final_status'* column contains the product's final status (CORE), which appears to be a categorical variable.
- The columns *'b2c_assortment'*, *'b2b_assortment'*, and *'total_assortment'* may be being used to represent different aspects of product classification.
- The *'product_syle_color'* column appears to be a numeric code or code representing the product style and color.
- **Product master:** This table comprises 94,867 rows and 26 columns, indicating its substantial size and wealth of information. The data exhibits diversity in data types, including integers (int64), floating-point numbers (float64), and character strings (object). Notably, there are missing values in several columns such as color, listing_price, size, color_group, and others, prompting the need for careful consideration of data imputation strategies. The listing_price column has an average value of 238,632.4 and a maximum value of 5,889,927, while the size column averages 35.43 with a maximum value of 47. Numerous categorical variables like color, color_group, gender, product_group, and launch_season are present, and examining unique values and their frequencies is essential for a deeper understanding of the data's diversity.
- **Retail price:** Similar to COGS, this table contains data about the retail price of the product and the validity of this price. The prices of the items range from free to 900

million, however, some products, although their prices are listed, are not for sale but are only for display.

2.2. Data Exploration

2.2.1. Univariate Analysis

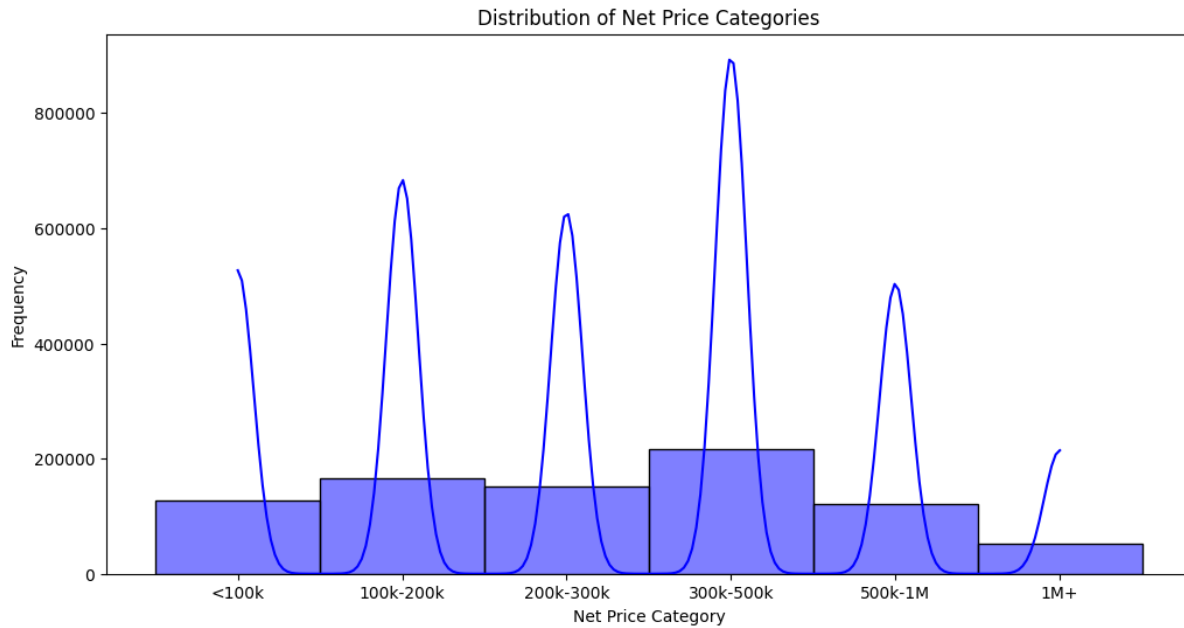


Figure 2-1 Distribution of Net Price Category

Notably, the preeminence of the 300k-500k and 100k-200k categories, as evidenced by their higher counts, underscores the substantial presence of fashion establishments operating within the mid-tier and lower price brackets. This prevailing trend suggests a market environment where affordability constitutes a pivotal determinant influencing consumer decision-making processes. Furthermore, the presence of fashion shops in the <100k category signifies an inclusive market ethos, appealing to budget-conscious consumers and contributing to the establishment of a more expansive customer base. Although the 1M+ category exhibits the lowest count, the presence of more than 50,000 fashion shops in this higher-end range indicates a discernible market niche for luxury or premium products. This, however, is accompanied by a comparatively smaller market share in contrast to the mid-tier and lower price categories.

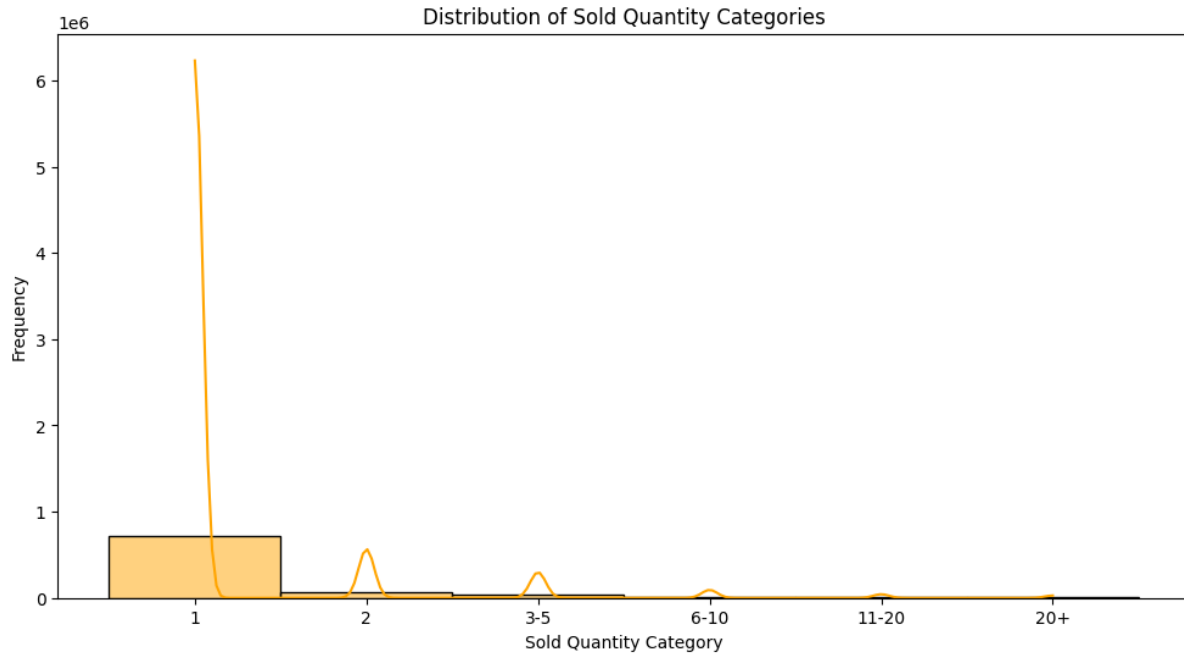


Figure 2-2 Distribution of Sold Quantity Categories

In terms of frequency, the top-performing category is '1', boasting a hefty count of around 714,000 instances, signaling a substantial prevalence of solo fashion item sales. Following closely, the '2' category comes into play with roughly 65,000 occurrences, hinting at a noteworthy presence of products bundled in pairs. The '3-5' category, clocking in at approximately 34,000 instances, points to transactions involving a moderate number of items, while the '6-10' category, with around 10,600 occurrences, suggests a somewhat less common but still notable frequency of sales involving larger quantities. Stepping up the volume ladder, the '11-20' category, boasting approximately 4,900 instances, signifies transactions with a higher volume of items, and the '20+' category, featuring roughly 3,200 instances, indicates a specialized market segment dealing with substantial quantities of items sold.

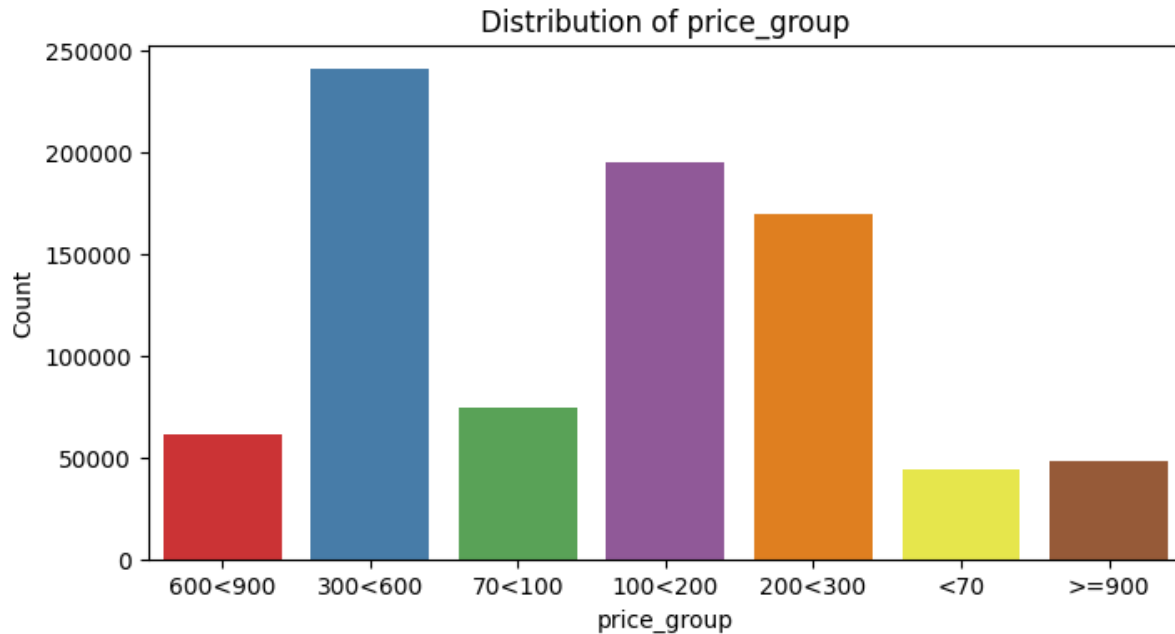


Figure 2-3 Distribution of Price Group

The '300<600' category stands out with a hefty count of around 240,000 occurrences, suggesting a strong presence of products priced between 300 and 600 units. Following closely, the '100<200' category clocks in at approximately 195,000 instances, indicating a significant share of reasonably priced items. The '200<300' category, with roughly 169,000 occurrences, points to a notable proportion of products falling within the 200 to 300 unit price range. Moving into the mid-range, the '70<100' category shows up with a frequency of around 74,000 instances, spotlighting products priced between 70 and 100 units.

In the higher price zones, both the '600<900' and '>=900' categories boast approximately 61,000 and 48,000 occurrences, respectively, highlighting a specific market segment for pricier items. Meanwhile, the '<70' category, featuring about 44,000 instances, represents products with a more wallet-friendly price point.

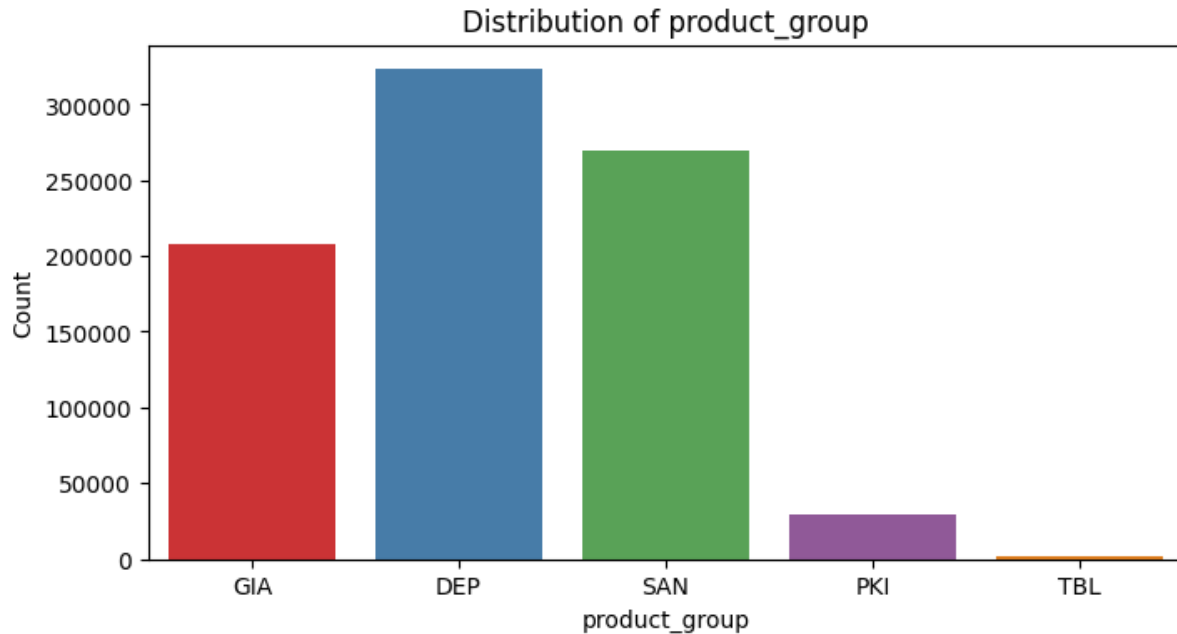


Figure 2-4 Distribution of Product Group

The 'DEP' category takes the lead with a hefty count of around 323,000 occurrences, pointing to a significant representation of products falling under this category. Coming in closely, the 'SAN' category registers approximately 269,000 instances, emphasizing a notable presence of products classified in this way. The 'GIA' category, with roughly 208,000 occurrences, signifies another substantial grouping within the dataset. Moving to more specific segments, the 'PKI' category, boasting around 29,000 instances, represents a distinct subset of products, while the 'TBL' category, featuring about 2,200 occurrences, suggests a more niche representation within the dataset.

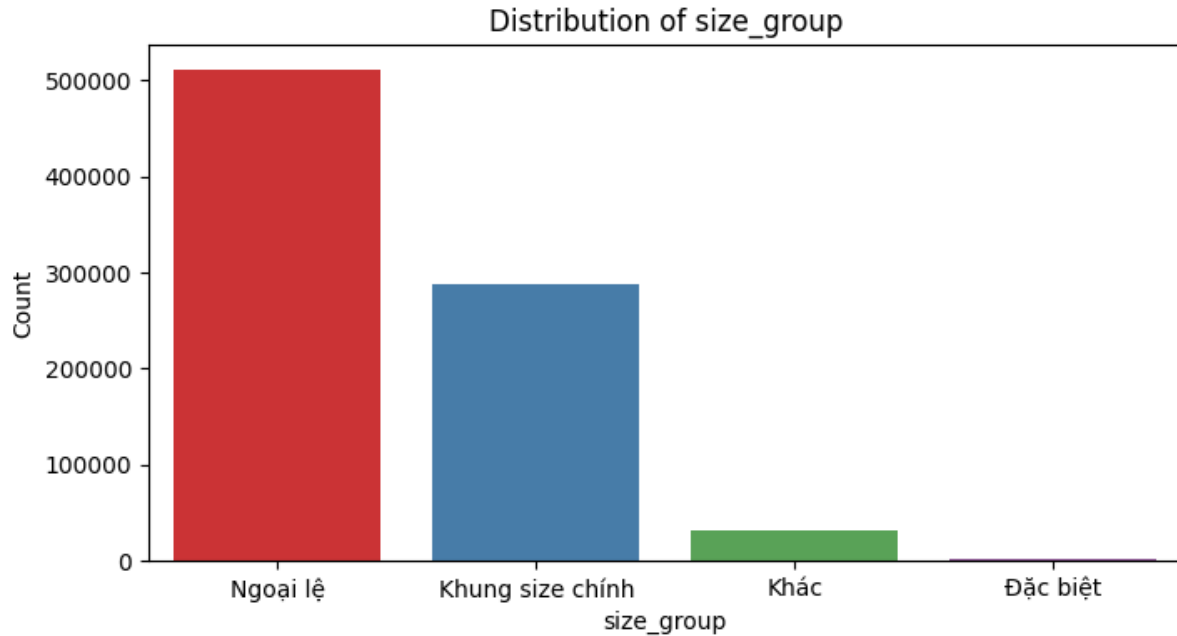


Figure 2-5 Distribution of Size Group

The 'Ngoại lệ' category stands out with a robust count of approximately 511,000 instances, indicating a substantial representation of products falling into this exceptional size grouping. Following closely, the 'Khung size chính' category logs in at around 288,000 instances, highlighting a significant presence of products adhering to the main size framework. The 'Khác' category, featuring roughly 31,000 occurrences, suggests a diverse range of products that don't neatly fit into the primary size categories. Lastly, the 'Đặc biệt' category, with approximately 1,600 instances, draws attention to a segment of products characterized as special or unique in terms of size.

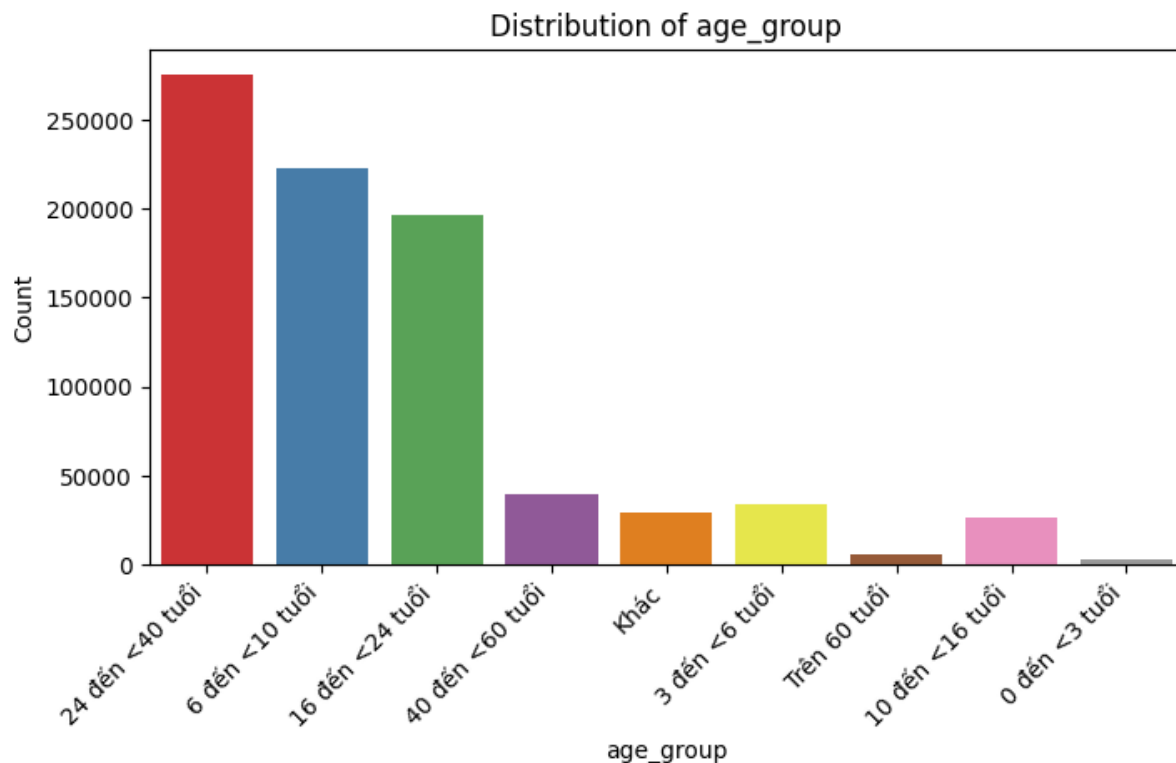


Figure 2-6 Distribution of Age Group

The dataset outlining age groupings in the fashion retail sector unveils valuable insights into the customer distribution across various age categories. The age group '24 đến <40 tuổi' (24 to less than 40 years old) takes the lead with a count of approximately 275,000, showcasing a significant representation of customers in this age bracket. Following closely, the '6 đến <10 tuổi' (6 to less than 10 years old) category logs around 223,000 instances, indicating a noteworthy presence of younger customers. The '16 đến <24 tuổi' (16 to less than 24 years old) category, featuring around 197,000 occurrences, represents another sizable segment of the customer base.

Moving into the mid-age range, the '40 đến <60 tuổi' (40 to less than 60 years old) and '3 đến <6 tuổi' (3 to less than 6 years old) categories, with approximately 39,000 and 34,000 instances respectively, suggest the presence of customers in middle age and early childhood. The 'Khác' (Other) category, featuring around 29,000 occurrences, adds diversity to the customer age groups. The '10 đến <16 tuổi' (10 to less than 16 years old) category, with around 27,000 instances, denotes a segment of teenagers.

In the older age brackets, 'Trên 60 tuổi' (Above 60 years old) and '0 đến <3 tuổi' (0 to less than 3 years old) categories, featuring about 5,800 and 2,900 instances respectively, represent the older and very young customer demographics.

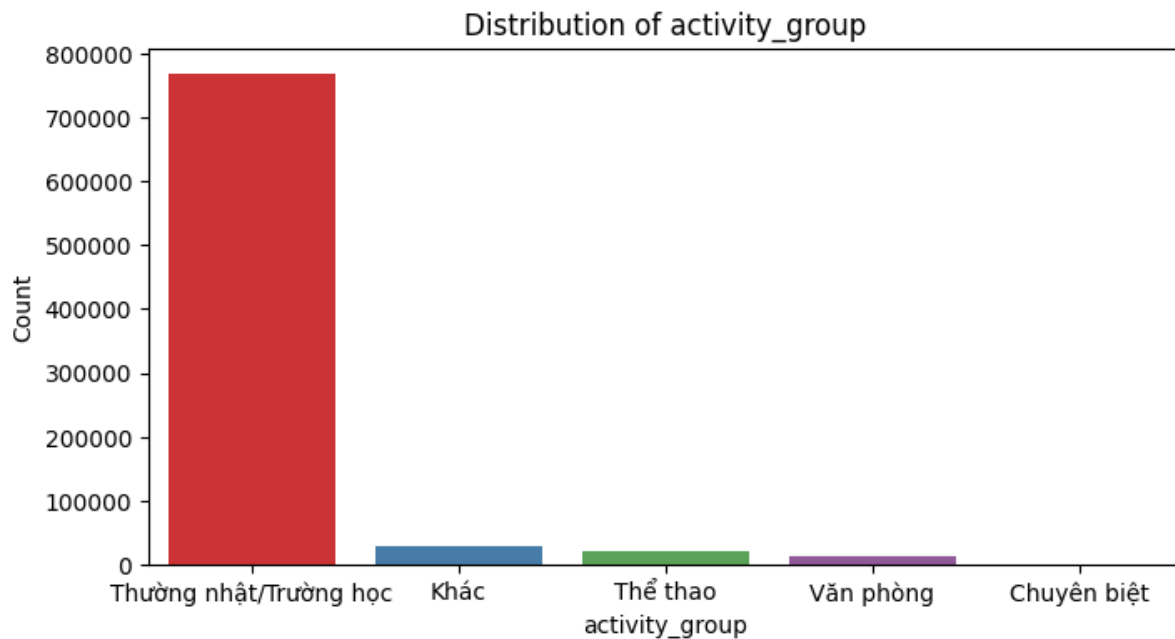


Figure 2-7 Distribution of Activity Group

The category 'Thường nhật/Trường học' (Everyday/School) takes the lead as the predominant activity group, boasting a count of approximately 768,000. This suggests a substantial representation of individuals involved in daily activities or school-related pursuits. The 'Khác' (Other) category, with around 29,000 instances, hints at a variety of miscellaneous activities among customers. 'Thể thao' (Sports) and 'Văn phòng' (Office) categories, featuring approximately 21,000 and 12,900 instances, respectively, indicate customers engaged in sports or office-related activities. The 'Chuyên biệt' (Specialized) category, with roughly 150 instances, points to a smaller but distinct segment of customers engaged in specialized activities.

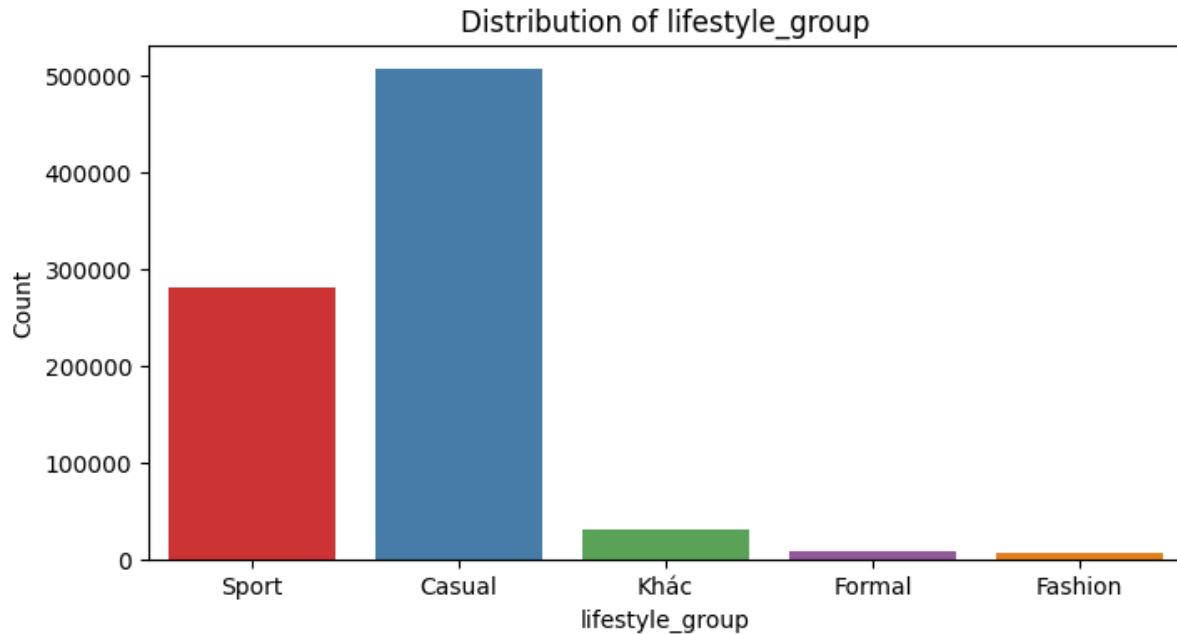


Figure 2-8 Distribution of Lifestyle Group

The 'Casual' lifestyle takes the lead as the predominant category, boasting a count of approximately 507,000, indicating a significant representation of customers who favor casual fashion choices. Following closely is the 'Sport' lifestyle, with around 280,000 instances, highlighting a substantial presence of individuals inclined towards sporty and active lifestyles. The 'Khác' (Other) category, featuring approximately 29,800 occurrences, suggests additional diversity in lifestyle preferences among customers. The 'Formal' lifestyle, with around 8,300 instances, represents a segment of customers with a preference for formal and professional attire, while the 'Fashion' lifestyle, with about 7,000 instances, denotes a group of fashion-forward individuals likely prioritizing trendy and stylish clothing choices.

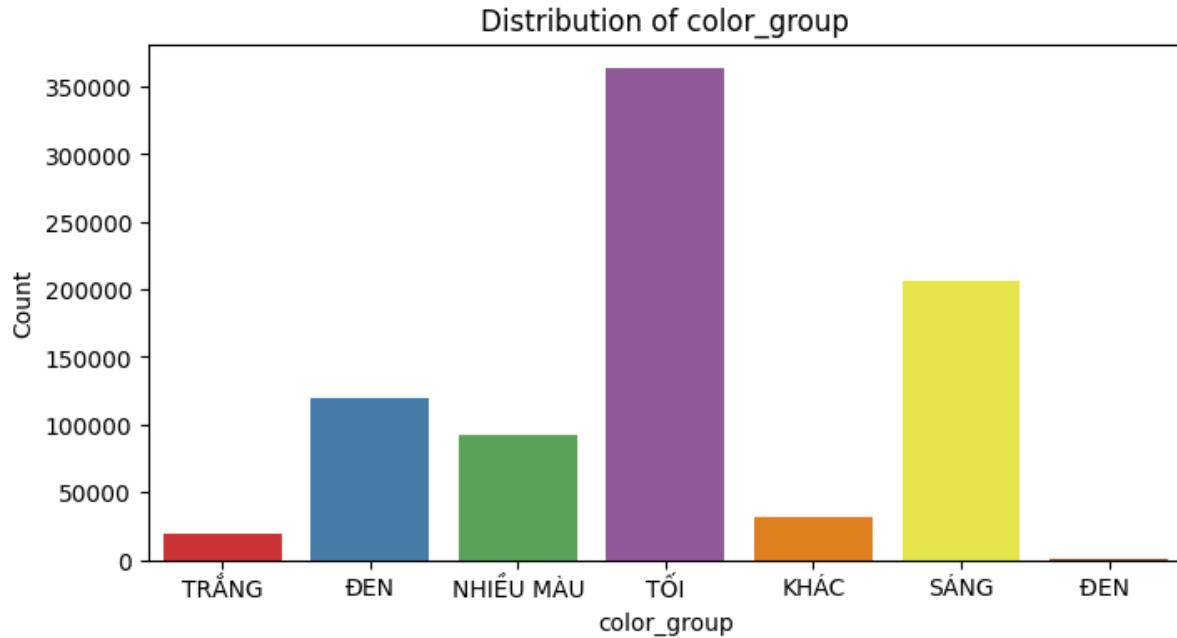


Figure 2-9 Distribution of Color Group

The 'TỐI' (Dark) color group takes the lead as the most prevalent, boasting a count of approximately 363,000, indicating a substantial representation of products in darker shades. Following closely is the 'SÁNG' (Bright) color group, with around 206,000 instances, suggesting a significant presence of products in lighter and brighter colors. The 'ĐEN' (Black) color group, listed separately with 119,000 occurrences, and an additional 'ĐEN' entry with 219 instances, highlights a specific preference for black-colored products. The 'NHIỀU MÀU' (Multi-color) category, with approximately 91,800 instances, indicates a demand for products featuring a variety of colors. The 'KHÁC' (Other) color group, featuring around 31,400 occurrences, suggests additional diversity in color preferences among customers. The 'TRẮNG' (White) color group, with about 19,600 instances, denotes a preference for products in white hues.

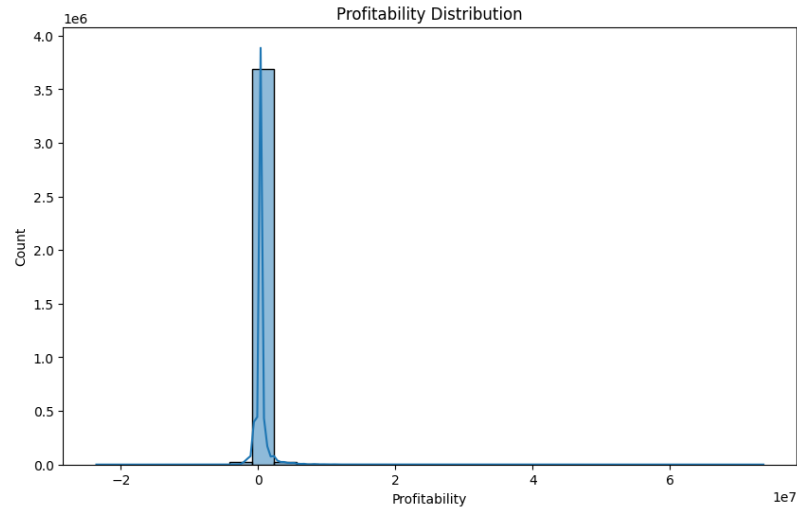


Figure 2-10 Profitability Distribution Chart

Although the total number of products sold is relatively high, this number is about 4 million products, however, the profit of each product is mostly only 100,000 VND (0.1×10^7). Even though some products are not profitable, they still have to be sold to maintain customers and create diversity in business products. Besides, there are still a small number of products with higher revenue than the average, but these products have extremely low sales.

Most products have affordable prices to suit most customers, so profits are not high. However, most profits come from this product group, so business strategies will focus on product prices to diversify products while still ensuring consumption and revenue

Top-selling product:

Perhaps because of security issues, this data set does not disclose specific product names, however, finding products with high purchases to focus on and boost revenue is a good way. For example, we can see that the product has the highest Total Sold Quantity about more than 3000 products.

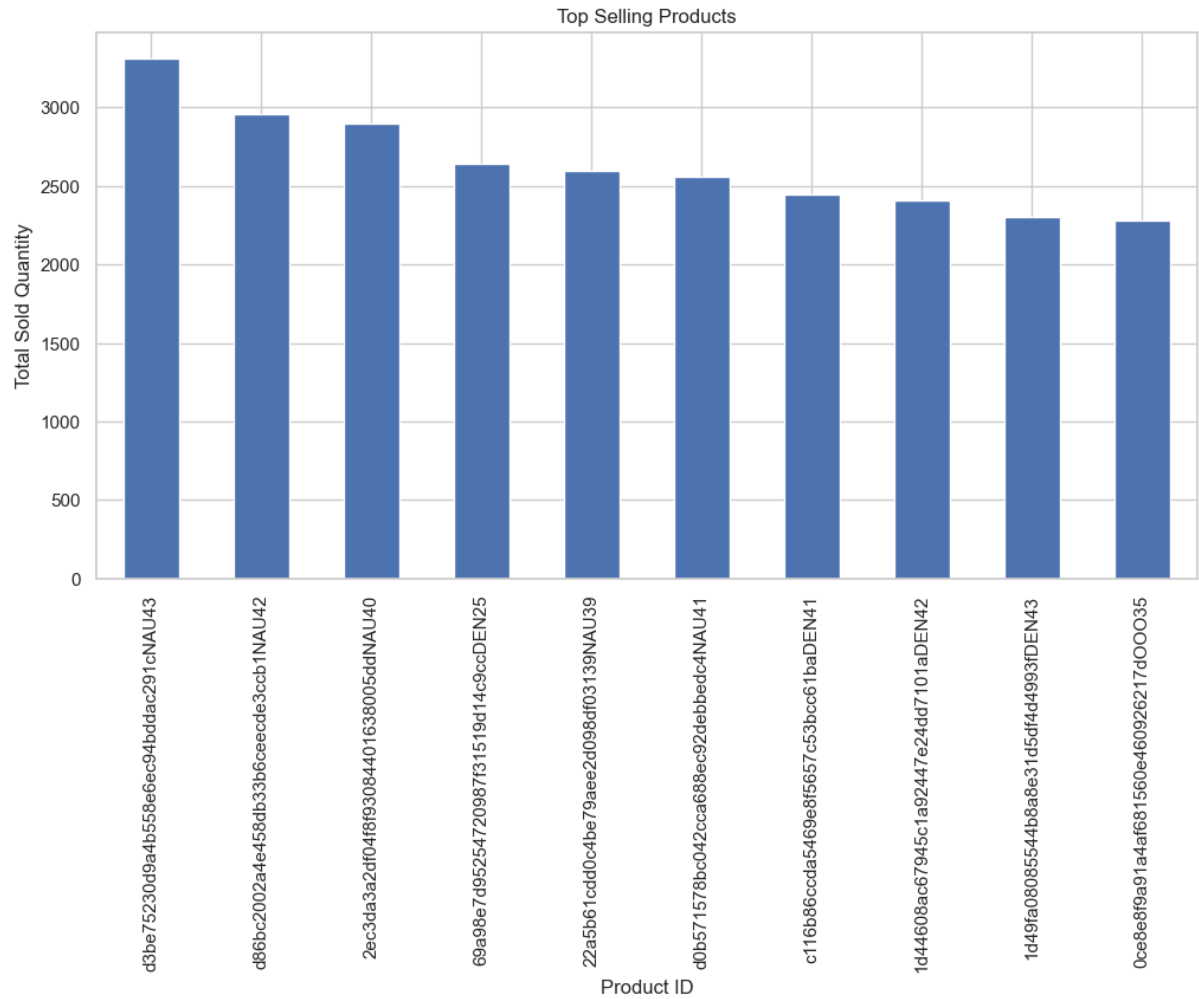


Figure 2-11 Top-selling products Chart

2.2.2. Multivariate Analysis

2.2.2.1. Basic Analysis

What is the trend of total sales over year?

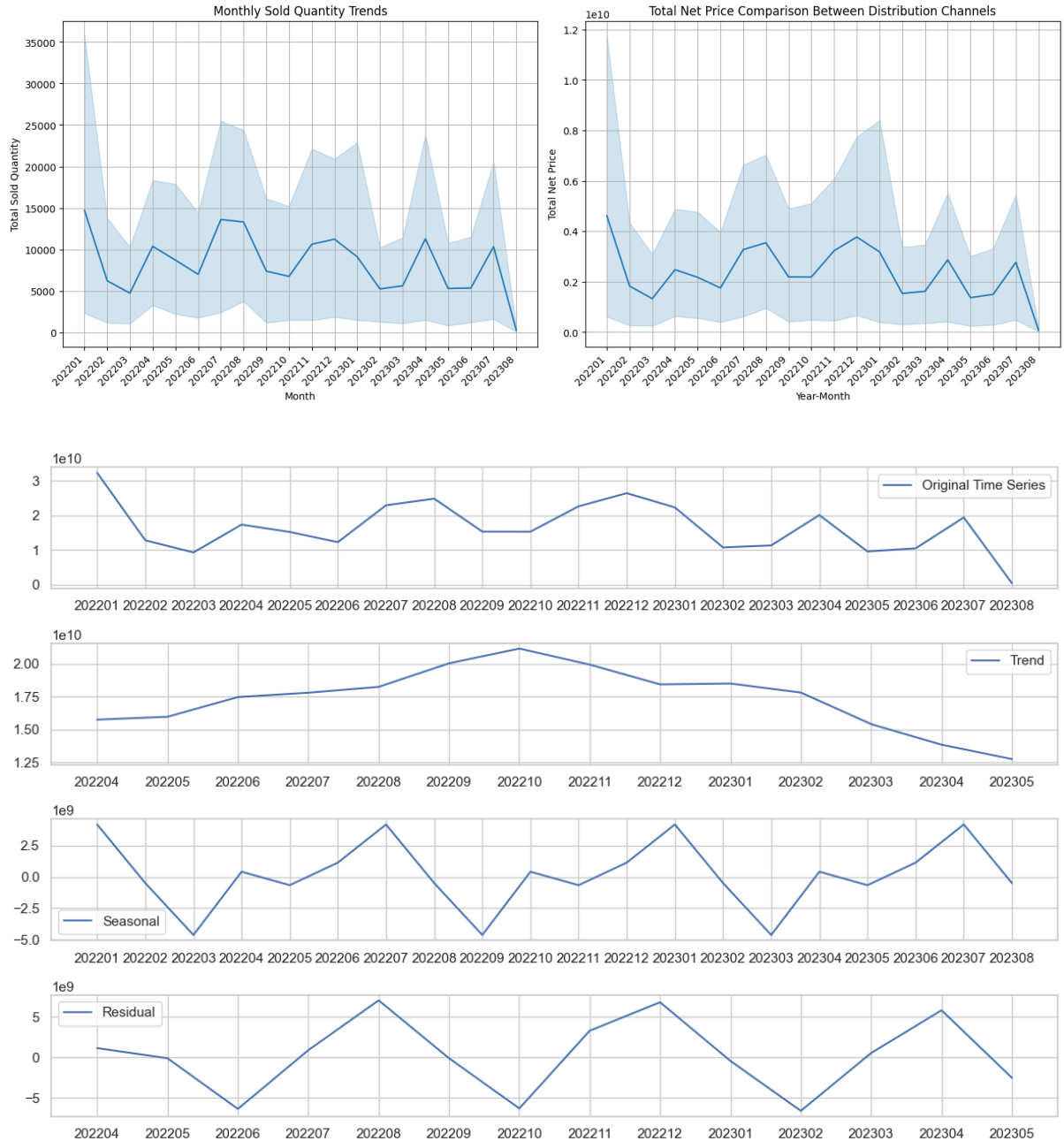
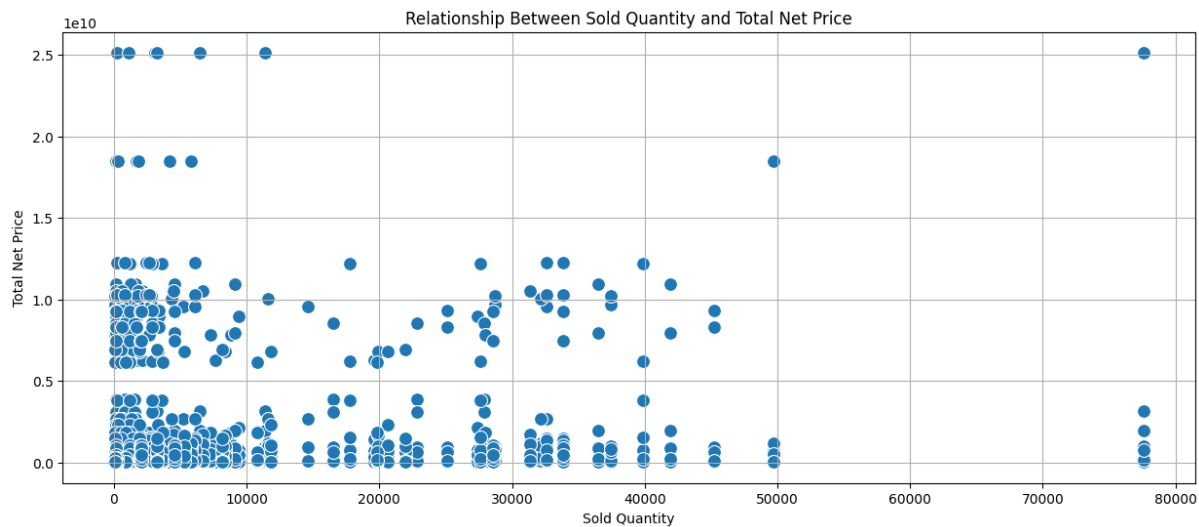


Figure 2-12 Sales Trend Over Time

The visual representation indicates a zenith in sales during October 2022, followed by a descending trajectory. This warrants attention, as sustained periods of dwindling sales may point to issues related to product appeal, market competition, or external factors affecting consumer interest. Observable 6-month cycles hint at seasonal patterns, with peaks in January and July and troughs in March and September. Understanding these seasonal undulations is pivotal for effective inventory management, marketing planning, and resource allocation.

To address potential market appeal decline, consider diversifying product offerings. Introduce new products or variations aligned with changing consumer preferences or seasonal demands. This can broaden the audience and mitigate the impact of declining sales for existing products. Reevaluate and enhance marketing strategies, particularly during peak months. Identify factors contributing to success in October 2022 and replicate those strategies. Implement targeted marketing campaigns during off-peak months to maintain consistent consumer engagement throughout the year. Given the seasonal nature, optimize inventory management to align with 6-month cycles. Ensure stock levels adjust appropriately based on anticipated demand during peak and off-peak periods to prevent overstocking or stockouts, enhancing overall operational efficiency.

Is there any correlation between net price and quantity?



```
correlation = merged_data_price_quantity['sold_quantity'].corr(merged_data_price_quantity['net_price'])
print("Correlation Coefficient:", correlation)
✓ 0.0s
Correlation Coefficient: 0.06275266160946817
```

Figure 2-13 Scatter Plot between Net price and Quantity

In the context of sold quantity and net price, a correlation coefficient of 0.06275 suggests that there is a slightly positive but very weak relationship between them. It implies that while there might be a tendency for the net price to increase as the sold quantity increases, the relationship is not strong, and other factors likely have a more significant impact on the net price.

How does the distribution channel impact sales?

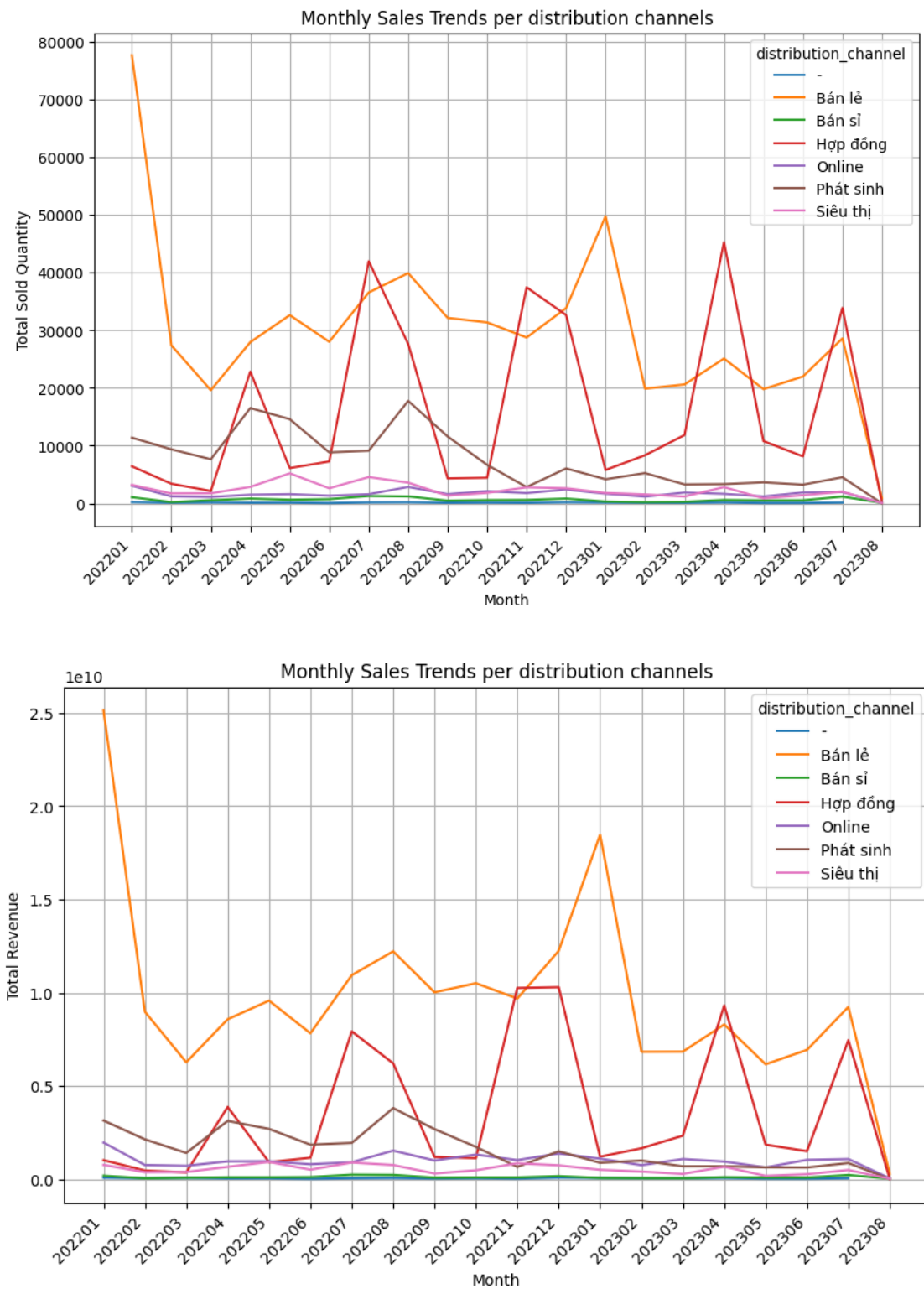


Figure 2-14 Total Sold Quantity and Revenue in Distribution Channels

Overall, retail and contractual channels significantly contribute to the highest revenue, outperforming other channels. While channels like supermarkets and online sales show improvement by the end of 2022, a notable decline is evident by 2023.

2.2.2.2. Customer Demographic Analysis

What is the distribution of sales across different regions?

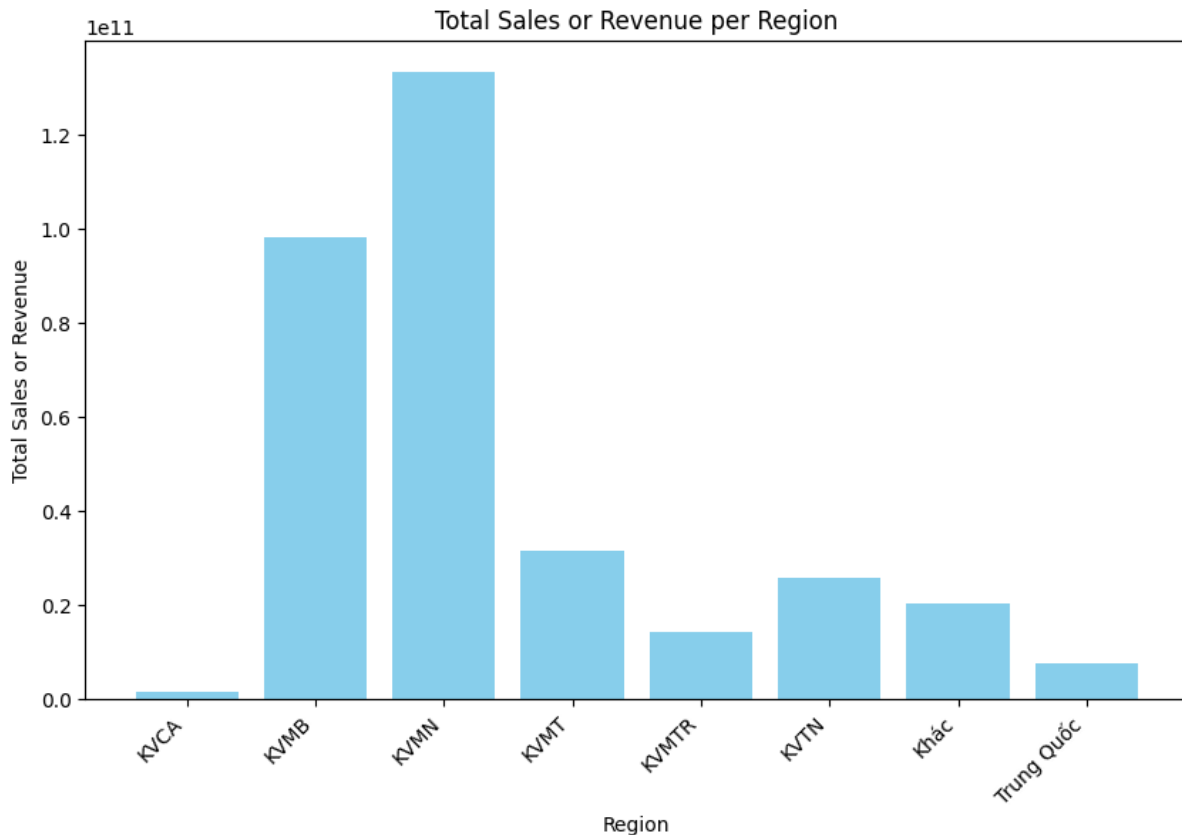


Figure 2-15 Distribution of Sales Across Different Regions

Notably, the southern region of Vietnam emerges as the preeminent contributor to the highest revenue, followed sequentially by the northern, central, and Central Highlands regions. A distinctive inclusion is a dataset subset corresponding to the geographic region of China, introducing an international dimension to the analysis. It is evident that pivotal economic regions in Vietnam exhibit robust and significantly higher sales revenues than other areas.

How does urbanization affect store performance?

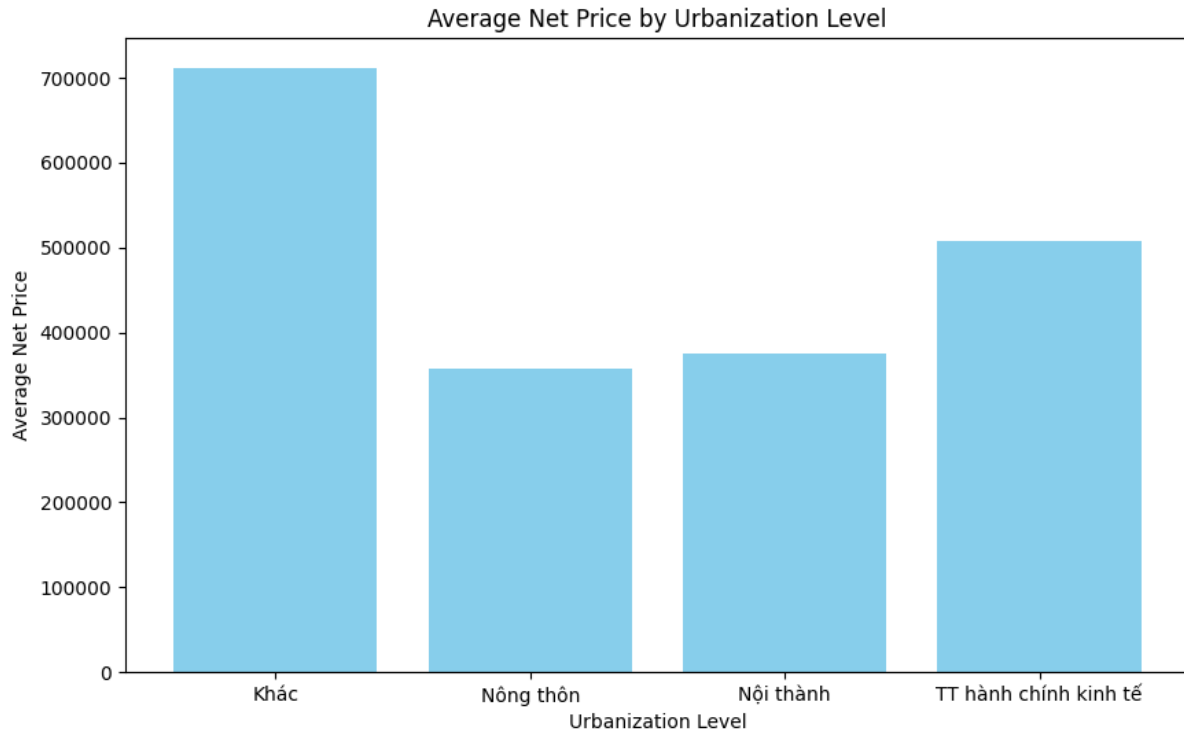


Figure 2-16 Average Net Price by Urbanization Level

The category denoted as "TT hành chính kinh tế" (Administrative and Economic Centers) exhibits the highest observed average net price, indicative of premium pricing within this specific classification. Concurrently, the "Khác" (Other) category also manifests a relatively elevated average net price, suggesting a comparable propensity for premium or specialized products within this diverse grouping. In contrast, the categories "Nội thành" (Urban) and "Nông thôn" (Rural) present lower average net prices, implying a discernible pricing differentiation. Urban and rural categories feature comparatively more accessible or cost-effective products. Disparities in average net prices across these categories contribute to a nuanced understanding of pricing dynamics.

2.2.2.3. Product Analysis

How do the sold quantity and revenue vary across genders and different price groups? Are there specific gender group combinations contributing significantly to both sold quantity and net revenue and are there any notable patterns or trends in the data?

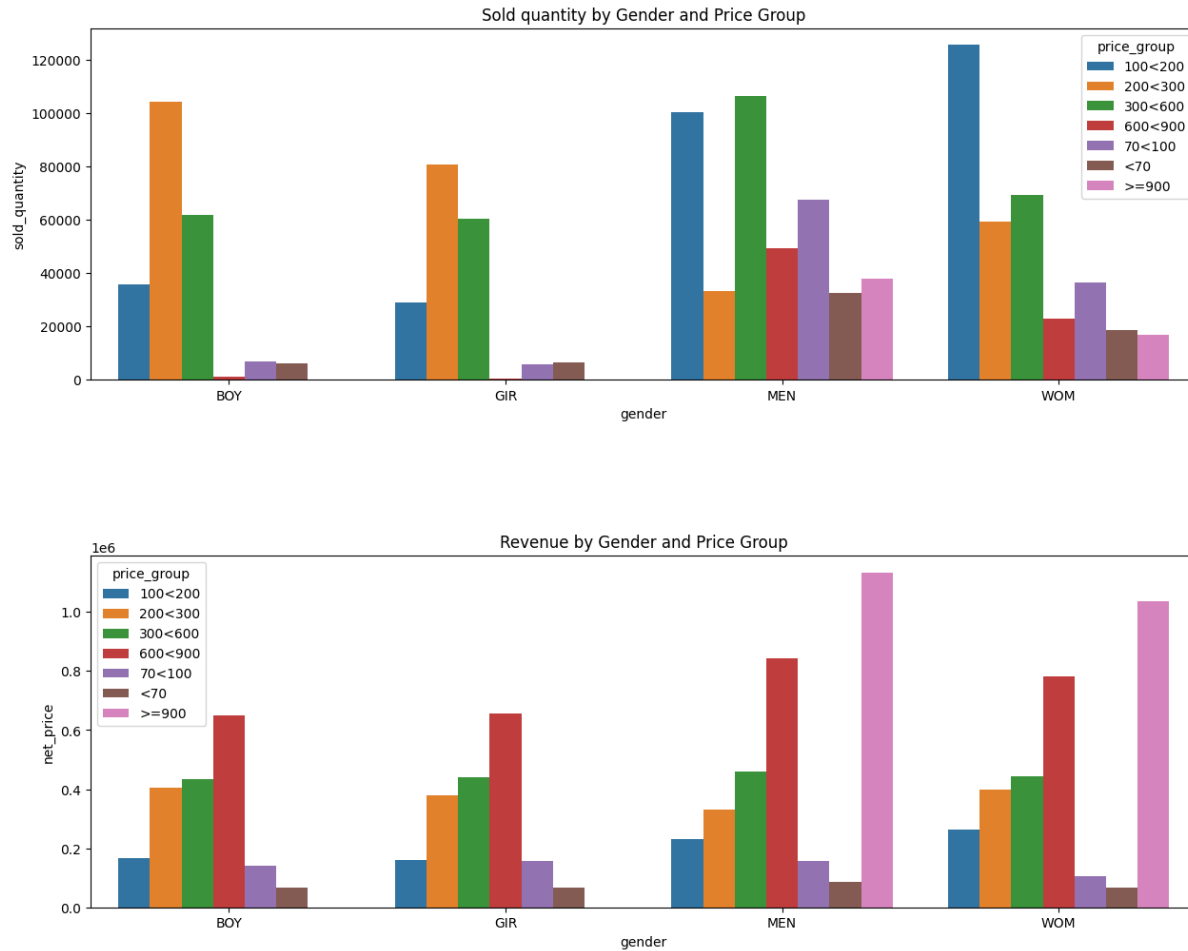


Figure 2-17 Sold Quantity and Revenue by Gender and Price Group

- Boys (BOY):
 - 0 to <3 years & <70 price group: The lower sold quantity but decent net price might correspond to the 0 to <3 years group, where fewer but possibly higher-priced items are sold.
 - 10 to <16 years & 200<300 price group: The significant quantity sold with a high net price in this range aligns with the observation that this age group is a key market. The data shows the highest quantity sold in the 200<300 range, confirming this as a crucial segment.

- 3 to <6 years & 6 to <10 years & 100<200 and 300<600 price groups: Good sales volume in these price ranges supports the notion of a solid market presence in middle childhood years.
- Khác (Others) & 600<900 price group: Lower sales and a higher net price might indicate miscellaneous or specialized items.
- Girls (GIR):
 - 0 to <3 years & <70 price group: Similar trends to boys with lower sales volume but higher net prices, indicating a demand for higher-priced items in the youngest age group.
 - 10 to <16 years & 200<300 and 300<600 price groups: High sold quantity and net price, especially in the 300<600 range, suggest a strong market presence and higher spending in this demographic.
 - 3 to <6 years & 6 to <10 years & 100<200 price group: Consistent with the observed good sales volume and strong net price, indicating a solid market in these age groups.
 - Khác (Others) & 600<900 price group: The lowest quantity and high net price could represent niche or premium items.
- Men (MEN):
 - 16 to <24 years & 24 to <40 years & 100<200 and 300<600 price groups: The highest sold quantities in these ranges align with the observed strong market presence among young and middle-aged men.
 - 40 to <60 years & 600<900 price group: Lower quantity with a decent net price corresponds to fewer purchases of possibly higher-value items.
 - Trên 60 tuổi (Over 60) & ≥ 900 price group: Very low quantity with a very high net price suggests a niche market for luxury items targeted at older men.

- Khác (Others) & 70<100 price group: Moderate quantity and lower net price likely represent miscellaneous items.
- Women (WOM):
 - 16 to <24 years & 24 to <40 years & 100<200 and 200<300 price groups: High sales volume and net price in these ranges indicate these are key demographic segments for women.
 - 40 to <60 years & 600<900 price group: Lower quantity but decent net price suggests a continued interest in higher-value items.
 - Trên 60 tuổi (Over 60) & >=900 price group: Lower quantity but high net price indicates a niche market for older women, possibly in luxury items.
 - Khác (Others) & 70<100 price group: Moderate quantity with a lower net price, likely representing miscellaneous items.
- General Observations & Strategic Considerations:
 - Target Age Groups: For boys and girls, products in the 200<300 and 300<600 price ranges are crucial, especially for the 10 to <16 age group. For men and women, the 100<200 and 300<600 ranges are key, with luxury items (>900) also significant for older age groups.
 - Premium Products: High net prices in the upper price ranges across all genders suggest a market for premium or specialized products. Brands might consider expanding offerings in these ranges, especially for older demographics.
 - Market Expansion: The 'Khác' category across all genders, particularly in higher price ranges, indicates a potential for niche market exploration or the introduction of new, innovative products.

- Price Sensitivity: The variation in sold quantities across price ranges indicates different levels of price sensitivity among consumers. Brands might consider price-tiered strategies to cater to a wider range of customers.

When considering sold quantity and revenue, how does the performance vary across different product groups within each brand? Are there specific brandproduct group combinations that stand out in terms of both sold quantity and net revenue, and are there any discernible patterns or trends in the sales data?

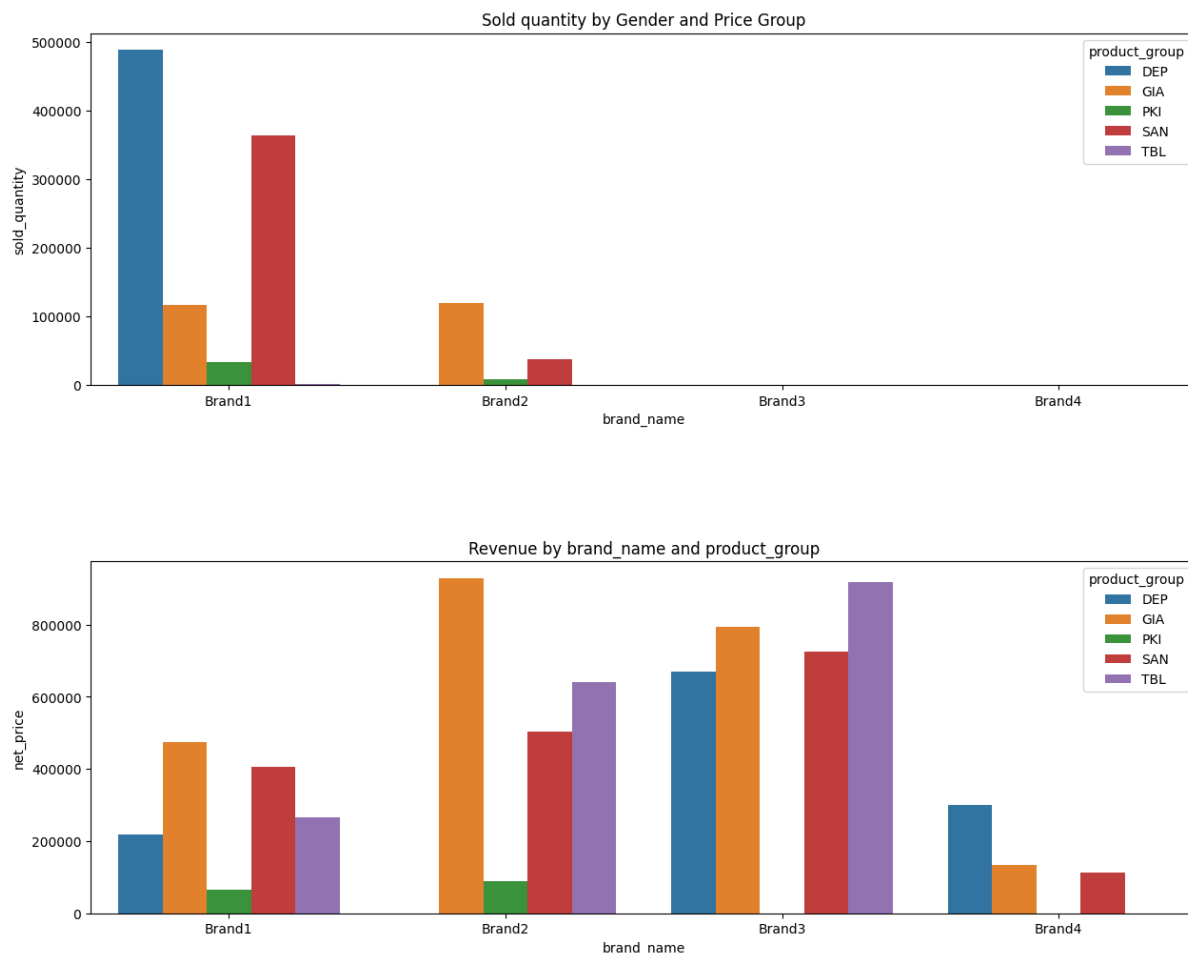


Figure 2-18 Sold Quantity and Revenue by Brand and Product Group

- Brand1:

- DEP: High sold quantity with a relatively lower net price, indicating a popular, possibly more affordable or essential product group.
- GIA: Moderate sold quantity with a higher net price, suggesting a premium or less frequently purchased product group.
- PKI: Lower sold quantity and net price, indicating a niche or less popular product group within this brand.
- SAN: High sold quantity with a moderate net price, suggesting a popular and wellpositioned product group.
- TBL: Very low sold quantity but a very high net price, indicating a premium, possibly luxury or specialized product group.
- Brand2:
 - GIA: High sold quantity with the highest net price among Brand2's products, indicating a successful, likely premium product group.
 - PKI: Low sold quantity with a lower net price, suggesting a niche market or less popular product group.
 - SAN: Moderate sold quantity and net price, indicating a reasonably popular product group.
 - TBL: Extremely low quantity but with a very high net price, suggesting an exclusive or luxury product group.
- Brand3:
 - DEP: Very low sold quantity but the highest net price, indicating an extremely premium or specialized product group.
 - GIA, SAN, TBL: Low quantities but very high net prices across all these groups, suggesting Brand3 specializes in premium, highcost products.

- Brand4:
 - DEP: Low sold quantity with a moderate net price, indicating a niche or moderately positioned product group.
 - GIA: Extremely low quantity and lower net price, suggesting this product group is not a strong performer.
 - SAN: Very low quantity with a low net price, further indicating a less popular or niche product group.
- General Observations & Strategic Considerations:
 - Brand Positioning: Brand1 shows a balanced portfolio with both high volume, lower priced products and low volume, higher-priced products. Brand2 and Brand3 seem to be positioned in the premium market, especially Brand3, with consistently high net prices. Brand4 appears to have a more niche market with generally lower quantities sold.
 - Product Group Performance:
 - DEP: Varies significantly across brands, from high volume and lower prices in Brand1 to very high prices and low volume in Brand3.
 - GIA: Generally suggests a premium positioning, especially in Brand2 and Brand3.
 - SAN: Appears popular across brands but with a wide range in net prices.
 - TBL: Consistently shows as a premium or luxury product group with high net prices and generally low quantities.
 - Market Dynamics: High net prices with low quantities suggest premium or luxury positioning, while higher quantities with lower prices suggest massmarket products. Brands and product groups need to be evaluated within the context of their market positioning and strategy.

How does the distribution channel impact sales?

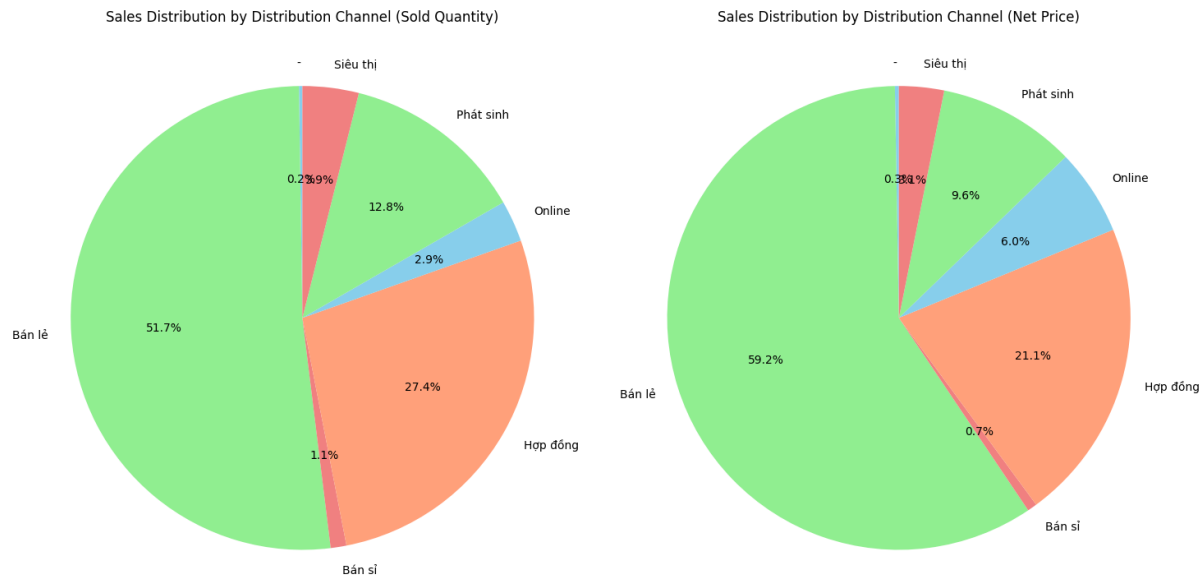


Figure 2-19 Sales Across Distribution Channels

- Evaluation:
 - Retail: the dominant channel in terms of both quantity and revenue, indicating a strong retail market presence. The substantial revenue suggests a successful retail strategy with a broad customer base.
 - Wholesale: A smaller quantity is sold, but considering the nature of wholesale (bulk selling), these figures could represent significant transactions.
 - Contractual: A substantial quantity sold, likely to contractual partners or large orders. The high revenue indicates significant contracts or partnerships that contribute greatly to the overall sales.
 - Online: The online channel shows a moderate sales volume but with a high revenue. This suggests a successful online strategy, possibly selling higherpriced items or having a strong market presence.

- Incidental: This could represent sporadic or occasional sales. The relatively high revenue and quantity suggest that while these sales are not the primary channel, they are significant.
- Supermarket: The supermarket channel has a decent sales volume and contributes a significant amount to the revenue. This indicates a strong presence in supermarkets, which can be an important retail strategy.
- General Observations & Strategy Considerations:
 - Retail Dominance: Retail is the dominant channel, suggesting a focus on enhancing retail operations, customer experience, and marketing.
 - Online Potential: The online channel, while not the highest in volume, generates substantial revenue. This indicates a growing or strong market segment worth investing in further.
 - Contractual Importance: The high revenue from contracts suggests valuable partnerships or largescale deals that are crucial to the business.
 - Diverse Portfolio: The presence of multiple effective channels indicates a diverse sales strategy, which can be beneficial for risk distribution and market coverage.

How does the sold quantity and revenue vary across different age groups within each gender category? Are there specific genderage group combinations that demonstrate notable differences in both sold quantity and average net price, and do any trends or patterns emerge from this analysis?

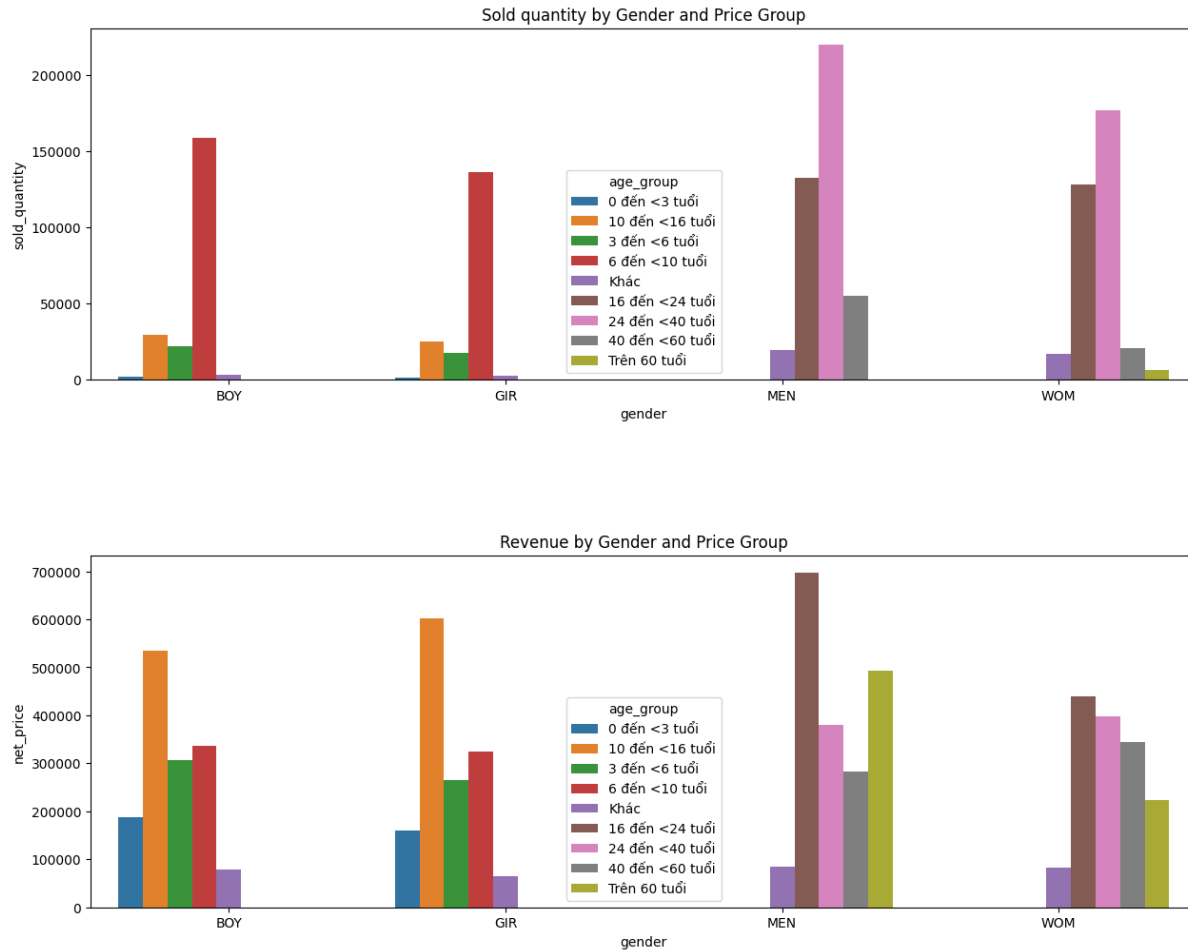


Figure 2-20 Sold Quantity and Revenue by Gender and Age Group

- Boys (BOY):
 - 0 to <3 years: Low quantity sold but with a decent average net price, indicating fewer but possibly higher-priced items.
 - 10 to <16 years: Significant quantity sold with a high net price, suggesting this age group is a key market for boys' products.
 - 3 to <6 years & 6 to <10 years: Both segments show good sales volume, with 6 to <10 being particularly strong. This suggests a solid market presence in the middle childhood years.
 - Khác (Others): Lower sales, which could be miscellaneous or uncategorized items.

- Girls (GIR):
 - 0 to <3 years: Similar to boys, lower sales volume but with a higher net price than the boys' equivalent age group.
 - 10 to <16 years: High sold quantity and net price, indicating strong market demand in this age group, even higher than the boys' equivalent.
 - 3 to <6 years & 6 to <10 years: Good sales volume, with the older age group (6 to <10) demonstrating particularly strong sales and net price.
 - Khác (Others): Lowest in quantity and net price, indicating a niche or miscellaneous category.
- Men (MEN):
 - 16 to <24 years & 24 to <40 years: These are the prime segments with the highest sold quantities and substantial net prices, indicating a strong market presence among young and middleaged men.
 - 40 to <60 years: Lower quantity but with a decent net price, suggesting fewer purchases but possibly of higher value items.
 - Trên 60 tuổi (Over 60): Very low quantity but with a very high net price, indicating niche, possibly luxury items targeted at older men.
 - Khác (Others): Moderate quantity and lower net price, likely miscellaneous items.
- Women (WOM):
 - 16 to <24 years & 24 to <40 years: High sales volume and net price, indicating these are key demographic segments for women, similar to men.
 - 40 to <60 years: Lower quantity but a decent net price, similar to the trend in men, suggesting continued interest but possibly in highervalue items.

- Trên 60 tuổi (Over 60): Lower quantity but a relatively high net price, indicating a niche market for older women.
- Khác (Others): Similar to men, a moderate quantity with a lower net price.
- General Observations & Strategic Considerations:
 - Key Demographics: For both boys and girls, the 6 to <16 age range shows high activity, indicating a robust market. For adults, the 16 to <40 age range is crucial.
 - Premium Pricing in Older Age Groups: Older demographics (40 years and above) tend to have a lower sold quantity but higher net prices, suggesting a market for premium or specialized products.
 - GenderSpecific Trends: Girls' items in the 10 to <16 age group have higher net prices and quantities compared to boys, suggesting potential for targeted marketing and product development.
 - Potential for Expansion: The 'Khác' category across all genders indicates miscellaneous or unclassified items. Understanding this category better could reveal new market opportunities or areas for data improvement.

When examining sales performance based on gender and lifestyle group, how do the sold quantity and revenue vary across different lifestyle groups within each gender category? Are there specific lifestyle group combinations that exhibit significant differences in both sold quantity and net revenue, and are there any notable trends or insights that can be derived from this analysis?

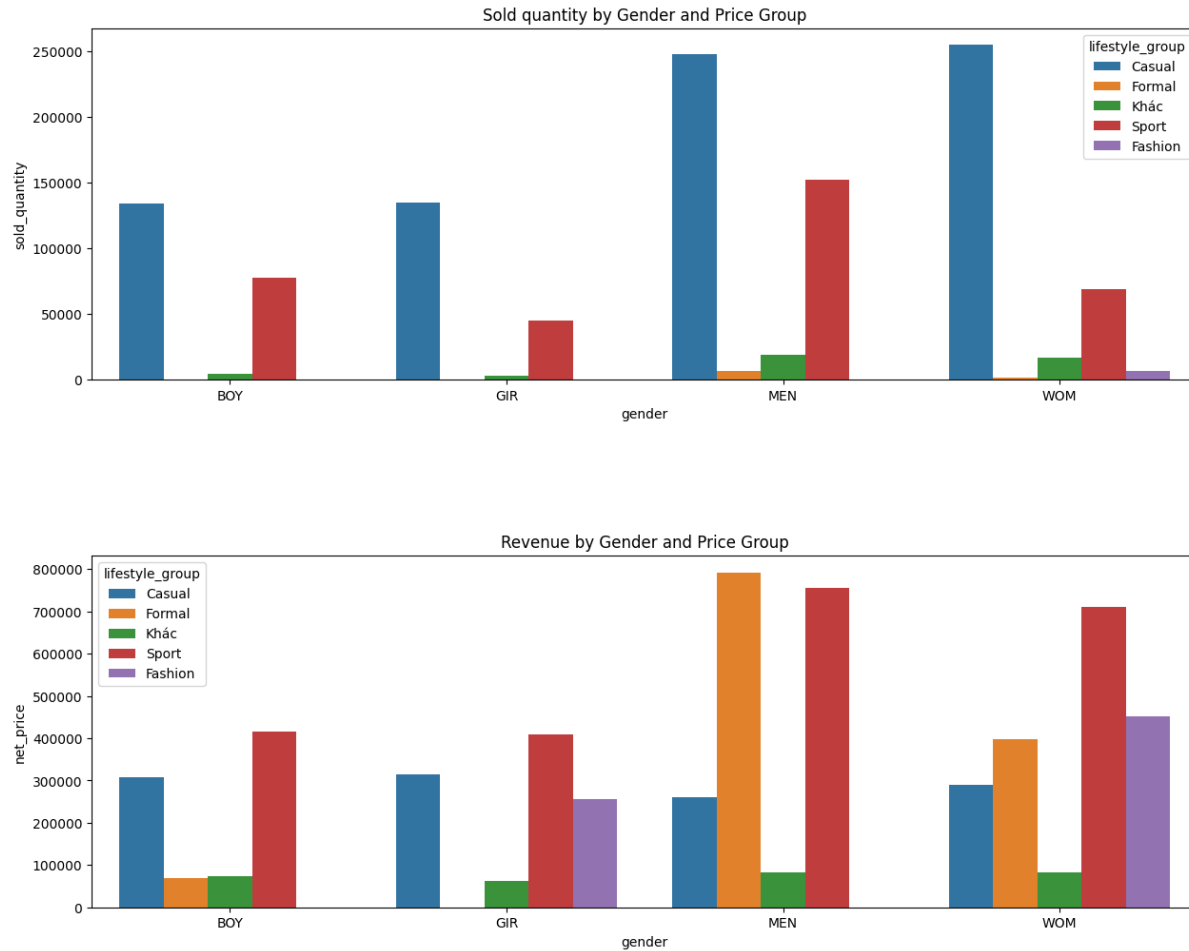


Figure 2-21 Sold Quantity and Revenue by Gender and Lifestyle

- Boys (BOY):
 - Casual: High sold quantity with a moderate net price, indicating a popular and potentially essential product group.
 - Formal: Extremely low sold quantity but a higher net price, suggesting a niche or less frequently purchased category.
 - Khác (Others): Low sold quantity with a lower net price, indicating a less popular or miscellaneous category.
 - Sport: Significant sold quantity with a high net price, suggesting a strong market presence and possibly premium offerings in sports attire.
- Girls (GIR):

- Casual: Similar to boys, high sold quantity with a moderate net price, indicating popularity and a primary choice for everyday wear.
 - Fashion: Very low sold quantity but with a very high net price, suggesting exclusive or luxury fashion items.
 - Khác (Others): Low quantity and net price, similar to boys, indicating less popular or miscellaneous items.
 - Sport: Moderate sold quantity with a high net price, indicating a solid market for girls' sportswear, possibly with premium offerings.
- Men (MEN):
 - Casual: Very high sold quantity with the lowest net price among MEN's categories, indicating a strong market presence and possibly a focus on value.
 - Formal: Moderate sold quantity with the highest net price, suggesting premium pricing and a strong market for men's formal wear.
 - Khác (Others): Moderate quantity with a lower net price, likely miscellaneous or less defined items.
 - Sport: High sold quantity with a very high net price, indicating a strong market and potentially premium offerings in men's sportswear.
 - Women (WOM):
 - Casual: Highest sold quantity among women's categories with a moderate net price, indicating a strong market presence and potentially a primary wardrobe choice.
 - Fashion: Low sold quantity but with a very high net price, suggesting luxury or highend fashion items.

- Formal: Low quantity with a high net price, indicating a niche market for premium women's formal wear.
- Khác (Others): Moderate quantity with a lower net price, similar to other demographics, indicating less popular or miscellaneous items.
- Sport: Moderate sold quantity with the highest net price among WOM's categories, suggesting a strong market for premium women's sportswear.

3. Product Segmentation

3.1. Overview

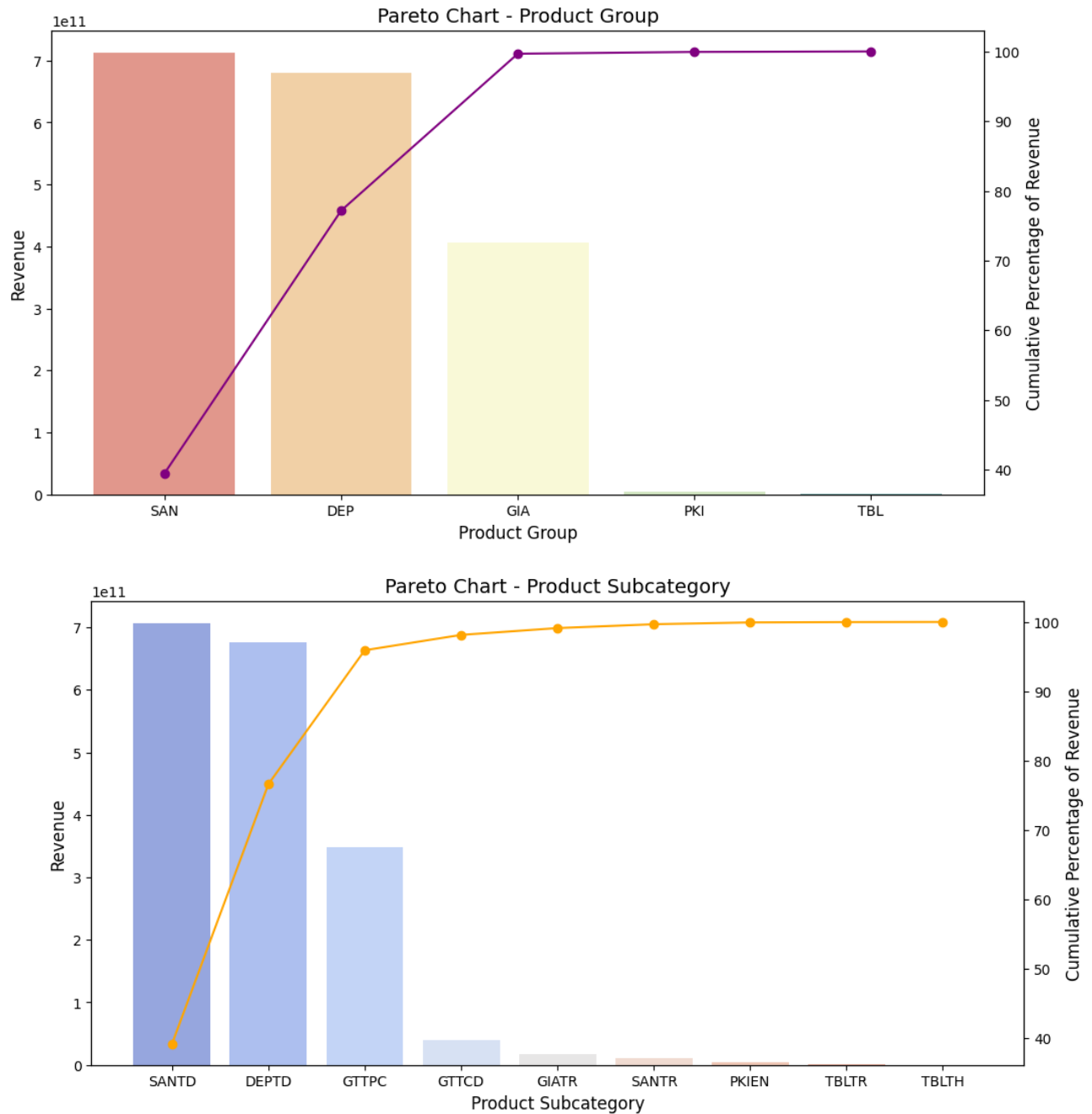


Figure 3-1 Pareto Chart for Product Category and Subcategory

Based on the two Pareto charts above, it can be observed that both product groups SAN and DEP collectively contribute to 80% of the revenue, with their respective subcategories being

SANTD and DEPTD. This implies a high concentration of revenue in these product groups and subcategories. These Pareto charts serve as a useful instrument for identifying and prioritizing the most important contributors to overall revenue, allowing for more targeted strategies and resource allocation.

3.2. ABC Analysis

What are the references that are driving most of the company's sales?

Product Rotation (using Pareto):

- Very Fast Movers: top 5% (Class A)
- The following 15% of fast movers (Class B)
- The remaining 80% of very slow movers (Class C)

In understanding the dynamics of product segmentation, our group decided to employ a comprehensive analysis, including the ABC Analysis. This categorization into Classes A, B, and C provides insights into the products driving the majority of the company's sales. Class A, constituting the top 5% of references, plays a pivotal role, contributing to around 80% of the total revenue. The identification of these critical references is imperative for strategic resource allocation and effective sales management for this footwear company.

Further delving into product rotation through the Pareto principle, the analysis distinguishes between Very Fast Movers (Class A), representing the top 5% of SKUs with the highest turnover rate, Fast Movers (Class B), encompassing the subsequent 15% of SKUs, and Slow Movers (Class C), comprising the remaining 80% of SKUs. This segmentation aids in understanding the speed at which products move through the inventory, providing valuable insights into the overall product portfolio.

3.3. Demand Stability

How stable is the company's customers' demand?

- Average Sales: μ

- Standard Deviation: σ
- Coefficient of Variation: $CV = \sigma/\mu$

A crucial aspect of product management involves evaluating the stability of customer demand. This is achieved through the assessment of average sales (μ), representing the mean quantity of products sold, and standard deviation (σ), which measures the extent of variation or dispersion from the average. The Coefficient of Variation ($CV = \sigma/\mu$) is then calculated, offering a ratio that indicates the relative variability of demand.

Interpreting the results, SKUs with high CV values suggest more erratic demand patterns. These high-variance products may present challenges such as workload peaks, increased forecasting complexity, and a higher risk of stock-outs. Recognizing and addressing SKUs with elevated CV values becomes crucial for optimizing inventory management processes, ensuring product availability, and mitigating potential disruptions in the supply chain. This comprehensive approach to product segmentation and demand stability enhances the company's ability to strategically manage its inventory and meet customer needs effectively.

	product_id	product_group	detail_product_group	sold_quantity	contribution	cumulative_contribution	class	SKU_ %
25064	d3be75230d9a4b558e6ec94bddac291cNAU43	DEP	DEPTD	3312	0.002820	0.002820	A	82.540258
25611	d86bc2002a4e458db33b6ceecde3ccb1NAU42	DEP	DEPTD	2961	0.002521	0.005341	A	84.341555
5477	2ec3da3a2df04f8f93084401638005ddNAU40	DEP	DEPTD	2899	0.002468	0.007809	A	18.039319
12441	69a98e7d95254720987f31519d14c9ccDEN25	PKI	PKIEN	2640	0.002248	0.010057	A	40.972108
4060	22a5b61cd0c4be79aee2d098df03139NAU39	DEP	DEPTD	2595	0.002209	0.012266	A	13.373069
...
15308	81573b191a02429dadda9428e9b24c4cNAU30	SAN	SANTD	-2	-0.000002	1.000013	C	50.413278
30358	fff116f43dd8455e8ee2c5423c5db50eHOG28	GIA	GTTPC	-3	-0.000003	1.000010	C	99.973656
7185	3d642155dae84692b29b69fe541ed052XNH33	GIA	GTTPC	-3	-0.000003	1.000008	C	23.663846
27220	e5f9126adced448a9f0f29e95c086de8XMN40	GIA	GTTPC	-4	-0.000003	1.000004	C	89.640070
21790	b730bd45622641e9a2c9b909d347beddXDG42	DEP	DEPTD	-5	-0.000004	1.000000	C	71.758817

30367 rows x 8 columns

Figure 3-2 Demand Stability Result

The above table provides information about various products, including their product_id, product_group, detail_product_group, sold_quantity, contribution, cumulative_contribution, class, and SKU_ %:

- product_id: A unique identifier for each product.
- product_group: The broader category or group to which the product belongs.

- `detail_product_group`: A more specific subcategory or detail within the broader product group.
- `sold_quantity`: The quantity of units sold for each product.
- `contribution`: The contribution of each product to the total turnover. It seems to be calculated as the ratio of the sold quantity of the product to the total turnover.
- `cumulative_contribution`: The cumulative contribution, which is the sum of contributions up to the current product. It represents the cumulative impact of products on total turnover.
- `class`: The ABC classification assigned to each product based on its cumulative contribution. It has values 'A', 'B', or 'C'.
- `SKU_%`: The percentage of SKUs (Stock Keeping Units) represented by each product, based on the total number of SKUs.

`SKU_%` represents the percentage of Stock Keeping Units (SKUs) that a particular product contributes to the total turnover. It indicates the proportion of total turnover attributed to a specific product relative to the entire product range.

Class A (High Contribution): Products in this class have a high contribution to total turnover. They typically represent a smaller percentage of SKUs but contribute significantly to overall revenue. In the given result, products classified as 'A' have high `SKU_%`, indicating a concentrated contribution from a smaller set of products.

Class B (Moderate Contribution): Products in this class have a moderate contribution to total turnover. They represent a moderate percentage of SKUs and contribute moderately to overall revenue.

Class C (Low Contribution): Products in this class have a low contribution to total turnover. They represent a larger percentage of SKUs but contribute less to overall revenue. In the given result, it seems that 'C' class products still have notable `SKU_%`, suggesting a larger number of products with lower individual contributions.

Some products with negative sold_quantity are classified as 'C', indicating a low contribution to turnover. Despite their low contribution, these products might have a relatively high SKU_%, implying that a significant portion of the total products falls into this category. The negative sold_quantity may indicate returns or adjustments, and the 'C' class designation reflects their lower impact on overall turnover.

Overall result insights:

The analysis classifies products into 'A', 'B', and 'C' based on their cumulative contribution to turnover. The distribution of products across these classes, along with their corresponding SKU_%, provides a strategic understanding of the product portfolio.

The presence of 'A' class products with high SKU_% highlights the concentration of revenue among a select few products. On the other hand, the 'C' class products, although numerous, individually contribute less to overall turnover.

Overall, this analysis helps identify key products driving revenue ('A' class) and those with moderate or lower impact ('B' and 'C' classes). It aids in focusing resources on high-impact products for strategic decision-making and inventory management.

3.4. Visualization

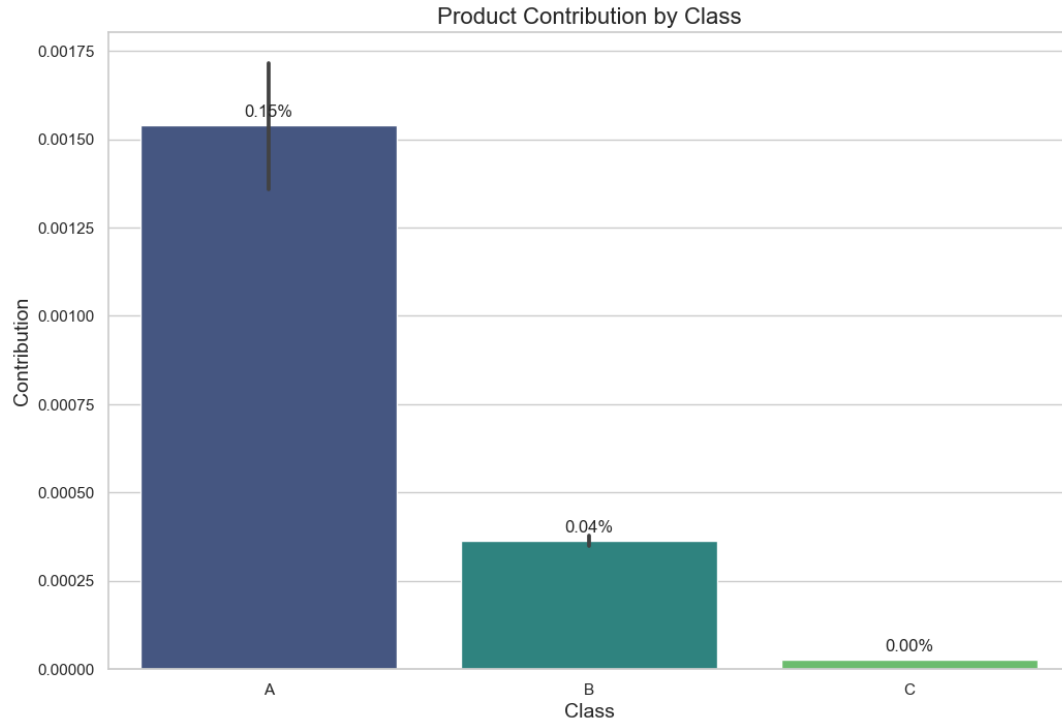


Figure 3-3 Product Contribution by Class



Figure 3-4 Cumulative Contribution by Class

The graph illustrates the unequal contributions of the three classes (A, B, and C) to the cumulative total. Specifically, A Class consistently stands out as the leading contributor, making the largest impact. For instance, at product index 10,000, A Class contributes approximately 0.6 to the cumulative total, surpassing both B Class, which contributes around 0.2, and C Class, which contributes approximately 0.1.

In summary, the overall pattern revealed by the graph emphasizes the significant discrepancy in contributions among the classes. A Class consistently holds the highest contribution, followed by B Class and then C Class, underscoring the unequal distribution of cumulative contributions across these three classes.

3.5. BCG Matrix

To build a strategy after product segmentation, our group will apply the BCG Matrix.



Figure 3-5 BCG Matrix Explanation

The BCG Matrix is a strategic management tool that categorizes product portfolio into four quadrants: Stars (high market share, high growth), Cash Cows (high market share, low growth), Question Marks (low market share, high growth), and Dogs (low market share, low growth). It helps businesses allocate resources, prioritize products, and develop appropriate strategies for each category, guiding decisions on investment, divestment, and growth. The integration between ABC Analysis and BCG Matrix can be briefly summarized as follows.

Stars (High Market Share, High Growth Rate):

Align high-value A category products (5%) with Stars in the BCG Matrix. These are products with high market share and growth potential. We should consider investing in innovative features, limited editions, or collaborations to maintain their high market share and growth rate.

Cash Cows (High Market Share, Low Growth Rate):

Some A category products, especially those with a high market share in mature markets, may be classified as Cash Cows. Continuously optimize these products to generate steady cash flow.

Question Marks (Low Market Share, High Growth Rate):

Evaluate B category products with growth potential (Question Marks). Consider strategic investments, marketing campaigns, or product development initiatives to turn them into Stars.

Dogs (Low Market Share, Low Growth Rate):

Assess products in the C category and those with low market share and growth potential. Consider phasing out or repositioning these products based on their strategic fit and contribution.

Overall

In formulating a comprehensive product strategy, the footwear company should focus on meticulous segmentation and innovation for Class A products, representing the 5% of very fast-moving items. Understanding consumer demographics and preferences within this category will guide tailored marketing and innovation efforts, ensuring a sustained competitive advantage. Simultaneously, strategic investments in brand-building activities will fortify the identity of Cash Cow products, encompassing both Class A and potentially Class B items. Leveraging the steady cash flow from these products is key to enhancing overall brand equity and fostering consumer loyalty.

To strategically elevate Question Marks, particularly Class B products with high growth potential, the company will embark on targeted marketing campaigns aimed at bolstering market share. Exploring collaborations or partnerships will further extend the customer base. Concurrently, for slow-moving products classified as Dogs in Class C, the focus will be on portfolio optimization. This involves a comprehensive evaluation of performance and the implementation of strategies such as inventory management, limited-time promotions, or potential repositioning within the market.

Adaptability is fundamental to this proposed strategy. Continuous monitoring of market trends, competitor activities, and consumer behavior is imperative. Establishing a robust feedback mechanism from sales channels and customers will facilitate agile decision-making based on real-time data and the ever-evolving dynamics of the footwear industry. The company will remain agile, ready to make informed adjustments to its strategy to stay ahead in the competitive market landscape.

4. Customer Segmentation

4.1. Data Processing

Data Transformation

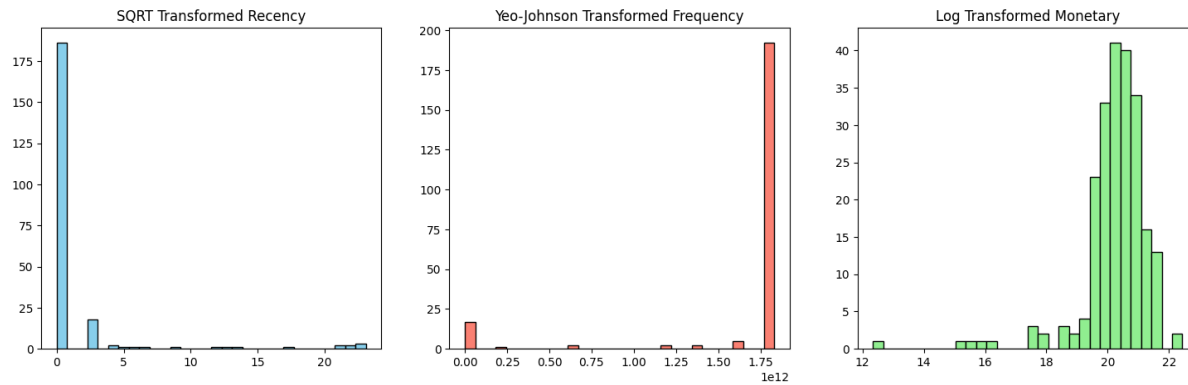


Figure 4-1 Distribution of RFM Metrics

This study focuses on data transformation and customer segmentation techniques to analyze customer behavior and revenue distribution in the retail sector. The dataset undergoes specific transformations to enhance its interpretability. Recency, frequency, and monetary variables are subjected to square root, Yeo-Johnson, and log transformations, respectively.

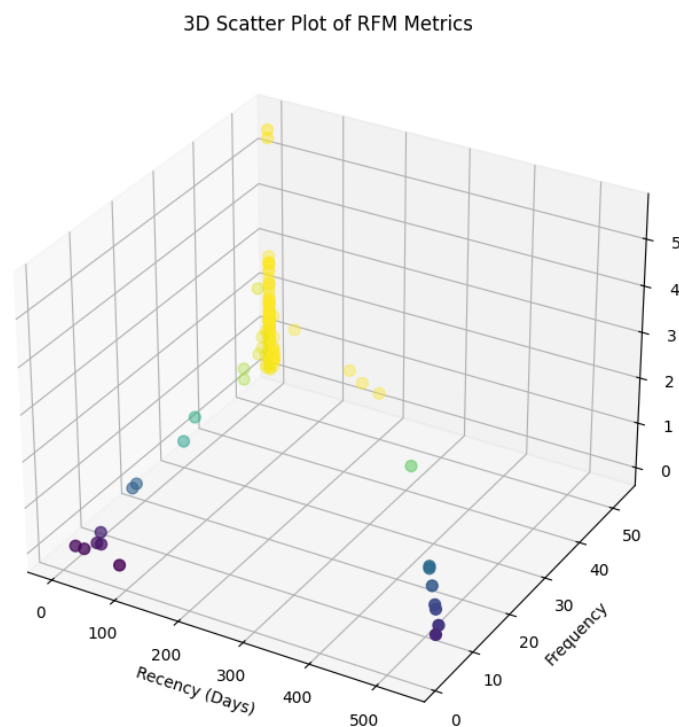


Figure 4-2 RFM Scatter Plot Pre-transformation

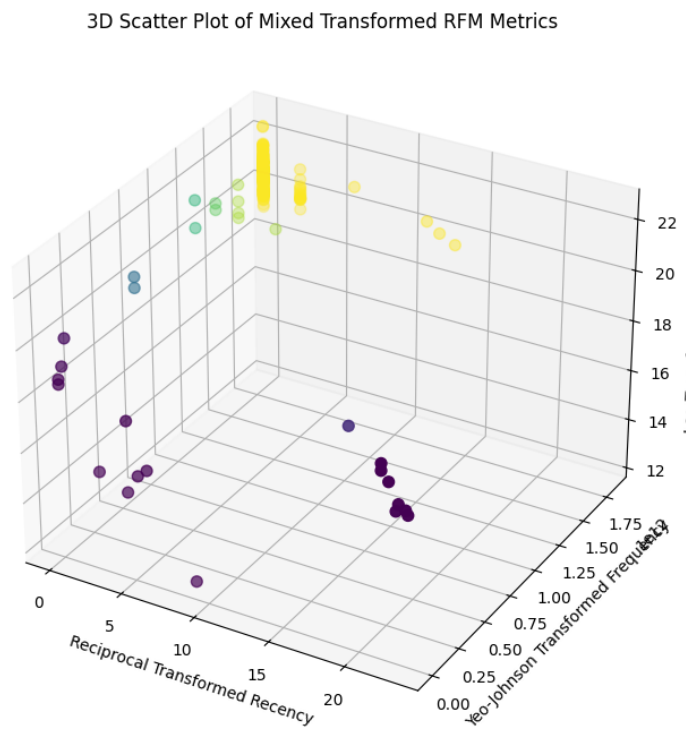


Figure 4-3 RFM Scatter Plot Post-transformation

Identify optimal cluster

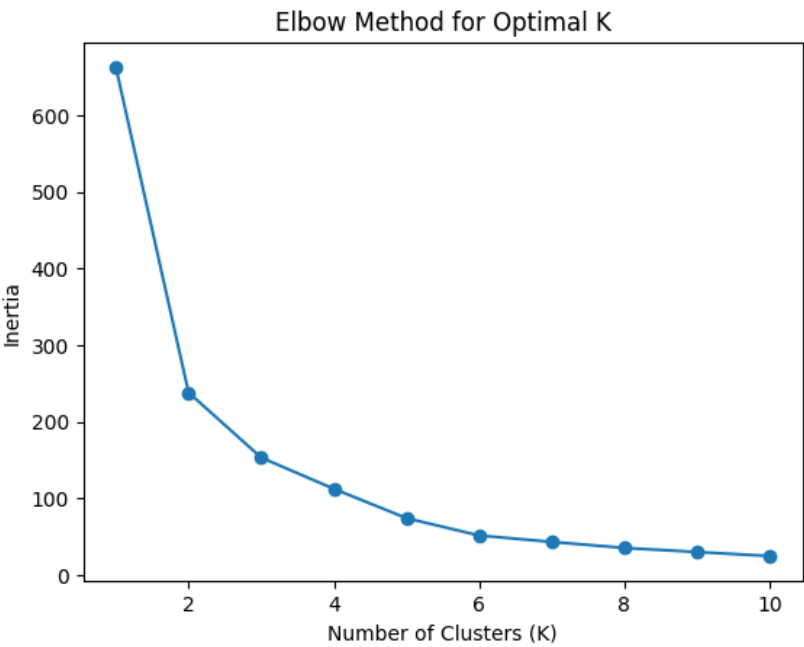


Figure 4-4 Elbow Result

Optimal cluster identification is crucial for understanding customer segments. The elbow, silhouette, and dendrogram methods are employed, suggesting that the optimal number of clusters is four. Subsequently, k-means clustering is applied to segment customers into distinct groups.

4.2. Segmentation Result

Based on the results of the RFM analysis and Kmeans clustering method, we have decided to divide our customers into three clusters, as shown below:

	recency	frequency	monetary
cluster			
0	5.519231	52.778846	5.071430e+08
1	469.000000	16.875000	1.076092e+08
2	19.600000	11.700000	1.038117e+08
3	0.565657	52.939394	1.433148e+09

Figure 4-5 The mean RFM value of clusters

Cluster 0: High-value customer

Based on the above data, we call this customer group the "High-value customer" group. This customer cluster has a relatively frequent purchasing frequency and a fairly regular purchasing frequency based on the R, F, and M indicators as follows:

- Recency: This group of customers' last purchase was relatively short, just over 5 days. The distance between purchases of this group of customers may be even and this shows that the return rate of customers is quite high.
- Frequency: This group has a relatively large purchasing frequency, on average they will purchase more than 52 orders per week. This partly brings great value to businesses.
- Monetary: The total amount of money spent by this group of customers is relatively large, with more than 500 million VND per month.

From the above indicators and based on purchasing characteristics, we decided to call this customer group the "High-value customer" group. The indicators also show that this group of customers has regular monthly purchases and a relatively high amount of money spent, so it can be predicted that this group of customers are small distributors specializing in buying and reselling. Paying attention to retaining and increasingly promoting business with this customer group will bring significant benefits in the long run.

Cluster 1: Lapsed customer

The indicators of the customer group in cluster 1 are relatively low. These figures show that the value that customers bring at the present time is not really too high, specifically:

- **Recency:** The recency index of this customer group is extremely low with the average last purchase being more than 1 year ago, specifically 15 months. This shows that the customer churn rate is extremely high.
- **Frequency:** The average weekly purchase frequency is 16 times. Although this number is higher than cluster 2, if you consider the recency index, this number shows that this group of customers only transacts with the business for a short time and after that, they almost stopped trading.
- **Monetary:** The total revenue that this customer group brings is at a stable level with a total of 100 million VND.

Indicators show that this group of customers tends to buy very little or only buy for a short period of time and then does not continue to buy, even though revenue is not too low. Therefore, we decided to call this group "Lapsed customers" based on the indicators we analyzed. At the same time, retention strategies to turn this customer group into high-value customers will be prioritized.

Cluster 2: Potential customer

The R, F, M indexes of this customer group are not too high, specifically:

- Recency: The recency index of this customer group is only at a good level, the average last purchase is about 19 days. This shows that the customer return rate is stable at about 1 time per month.
- Frequency: On average, this group of customers has a purchasing frequency of 11 times per week. Because this is a footwear product and customers can be regular consumers or small traders, this number can be considered at a good level.
- Monetary: The monetary index is low with an average of about 100 million VND, almost equal to the lapsed customer group.

From the above indicators and analysis, we have decided to call this group "Potential customer". The frequency of purchases and turnover rate are only stable, and the revenue is not high, showing that the customer group only purchases a few times per month. Therefore, strategies that focus on driving purchase intensity will be prioritized for this customer group.

Cluster 3: Loyal customer

This group of customers is labeled "Loyal customer" based on extremely high R, F, M indexes, specifically:

- Recency: The average last purchase of this customer group is extremely short at only 0.5 days. This clearly shows that the customer return rate is extremely high and will stick with the business for a long time.
- Frequency: On average, this group of customers has a number of purchases of more than 52 orders. Regardless of the total value of orders, a high-frequency index shows that customers are sympathetic to the business.
- Monetary: The monetary index of this customer group is quite high, up to 1.4 billion VND. Along with high purchasing frequency, this group of customers brings in large revenue with each order.

From the indicators we mentioned above, we decided to call this group "Loyal customer". Through the relatively high R, F, M indexes and this is also data related to the footwear

industry, we predict that these may be large distributors. If there are appropriate strategies or incentives for this customer group, it will bring significant benefits in terms of revenue.

4.3. Buyer Persona

Each customer group will have different purchasing characteristics and values, so depending on the customer group, there will be separate strategies.

Cluster 0:

As analyzed, this is the customer group that does not bring in the highest revenue but has the most stable indicators. Therefore, business strategies for this customer group are focused on driving revenue per order. To have more specific strategies for this customer group, we recommend based on the buyer personas of customers in this cluster.



Figure 4-6 A buyer persona of High-value customer cluster, N. J. Keria

The above information shows that this is a businessman specializing in selling fashion products. Business products are extremely diverse and footwear is one of those products. However, the products are in the near-high-end segment and the business scale is quite large, so the average value of each order is quite high. Therefore, to boost revenue from this customer, appropriate strategies are needed, specifically:

- Nurturing program and periodic communication: Communicate with customers to ask about service quality or product feedback so that customers feel cared for and increase their affinity for the business. At the same time, this also helps us understand customers to make appropriate service adjustments and boost revenue.
- Accumulate points and receive rewards: After each purchase, customers will be given a certain number of accumulated points based on the total order value. The higher the accumulated points, the greater the incentives from the business. This boosts revenue on each order and customers will tend to stick with the business for a long time to receive more incentives.

Cluster 1:

Customers from this group have an extremely low return rate despite the relatively stable purchase frequency, it is highly likely that these customers are looking for suppliers with better prices and do not continue to buy.



Figure 4-7 A buyer persona of Lapsed customer cluster, NS Hillson

This customer's information shows she and other entrepreneurs in the footwear industry are looking for a supplier with a good price, so the most suitable strategy is to retain this group of

customers to increase the rate of returning to the business. Our recommended strategies are as follows:

- Ask for customer feedback: Always ask customers about service quality after each purchase via email, this both helps increase goodwill with the business and helps the business overcome limitations.
- Give away short-term vouchers: Give away short-term vouchers of about 1 month or 3 months to encourage customers to come back and increase purchasing frequency.
- Send timely reminders: Through email, send notifications about new products or special events with promotions for customers to refer to. Maybe customers won't buy at all events, but it somehow increases brand recognition with customers.

Cluster 2:

Similar to the customer group in cluster 1, the "Potential customer" group will be focused on by businesses to increase purchase frequency. A typical buyer persona is as follows:



Figure 4-8 A buyer persona of Potential customer cluster, M. Adam

It can be seen that, because the nature of their work is mainly labor in the fields, most of these customers do not use a variety of products but only focus on highly durable products, and their purchasing frequency is also low. very few so strategies will focus on driving purchase frequency. Strategies include:

- Recommend personalized products: Send notifications about new and relevant products based on purchase history to increase personalization for customers. Because the purchasing frequency is not high, recommending products that suit each individual's preferences will attract customers' attention and have a higher tendency to purchase than products that are best-selling but not suitable.
- Send promotions via email: Purchasing demand is not high, so the interest in products of this customer group is very low. Sending voucher announcements or information about discounts to help information reach customers is one of the most optimal ways.

Cluster 3:

The customer group in the cluster 1 is considered the customer group that brings the main revenue to the business. Revenue, frequency, and return rate are all very high, so businesses must maximize revenue and retain them with the business.



Figure 4-9 A buyer persona of Loyal customer cluster, A. Arnold

Typically, the above information is the buyer persona of a merchant specializing in selling affordable and high-end fashion products. Because the business scale is large and diverse, the frequency of purchases and revenue from orders are extremely high. Strategies to retain and boost revenue will be the priority strategies to create strong relationships between customers and businesses. Specifically:

- Proactive Care and Excellent Customer Service: Focus on customer service and dedicated care. Ensure that every customer interaction is handled professionally and responded to quickly. Build a dedicated support system and trained staff to meet customer needs fully.
- Exclusive incentives: Build a high-end loyalty program with special incentives and benefits only for customer groups that bring in high revenue. Offer exclusive offers, rewards points, and valuable rewards based on their level of shopping activity.
- Give special offers and free product sampling: Let customers experience new products for free or have special offers to promote new product purchases. This also helps businesses promote new products through customers.

5. Chat GPT-4

We provide Chat GPT-4 the prompt as below:

“Use SalesPortion data. You are now tasked with analyzing an extensive dataset from a footwear company in Vietnam. The dataset includes a wide range of columns such as date, channel distribution, sold quantity, net price, customer demographics, and more. Some of the content includes Vietnamese words, and it's crucial to interpret and analyze these as well.

Here are some additional information. In the region, KVMN means Southern, KVMT means Middle, KVMB means Northern, KVCA means Asia, Trung Quốc means China, Khác means Other.

Your main goal is to conduct various analyses and provide detailed insights and recommendations for improving sales performance and customer demographics. In addition, you must create clear and coherent visualizations for each question to illustrate your findings effectively.

Remember, for each analysis, you need to provide comprehensive insights and recommendations. Ensure to address the following questions:

1. Analyze sales trends over time, define the seasonal pattern if needed, and propose possible reasons for the results.
2. Analyze sales revenue by distribution channels, regions, urbanization, city level, age group, and gender, providing insights and possible reasons.
3. Analyze sold quantity by gender, price group, and style group, and provide insights and possible reasons.
4. Segment customers by RFM into 4 clusters, analyze each cluster and provide strategies.
5. Segment products by ABC Analysis and BCG Matrix, and include strategies.

6. Analyze customer satisfaction and loyalty based on their purchase behavior and demographic information.
7. Analyze the correlation between customer satisfaction, loyalty, and other variables such as product group, store concept, and activity group.
8. Analyze the correlation between net price and sold quantity.

Your responses should be detailed, insightful, and supported by visualizations to enhance understanding. Remember to provide specific recommendations for improving sales performance and customer demographics in each analysis.”

Our dataset have nearly 700,000 rows, in order to assist Chat GPT in handling the dataset, we split the dataset into 20 files, each containing 35,000 rows. After uploading the files and send out the prompt, Chat GPT started to process each file, it started with question number 1 and file number 1. The result is as follows.

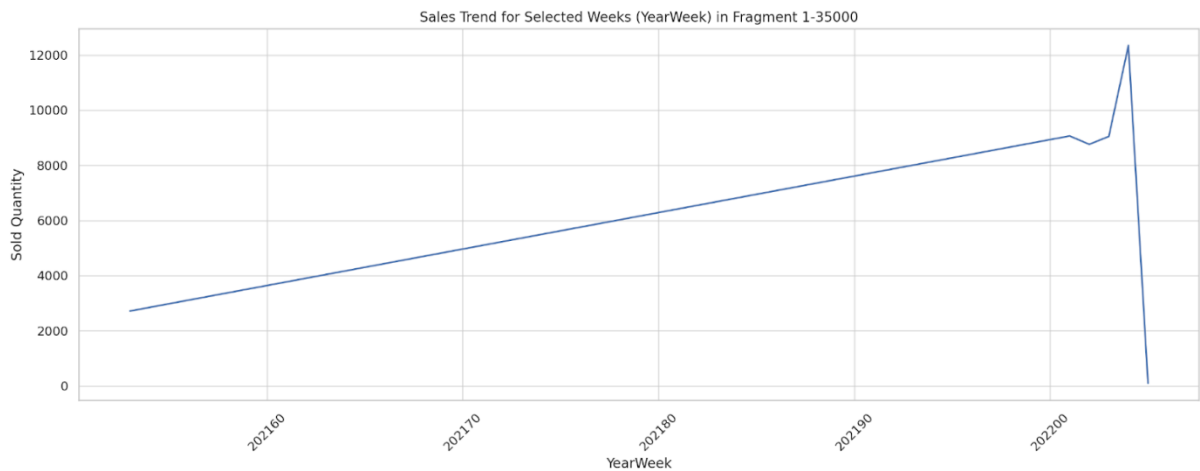


Figure 5-1 Result of Fragment 1

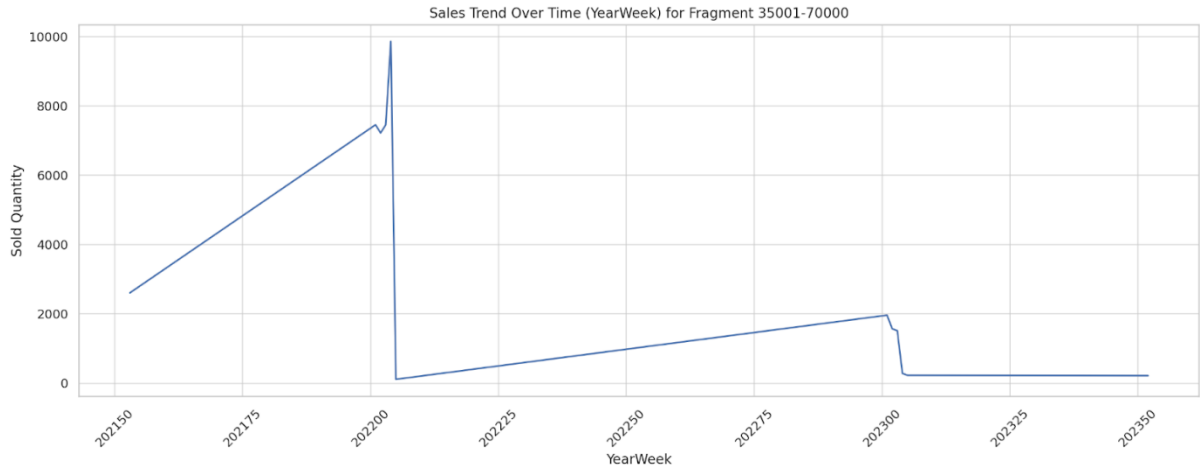


Figure 5-2 Result of Fragment 2

However, from the partition number 3 (row 70,000 to 105,000), the line disappeared.

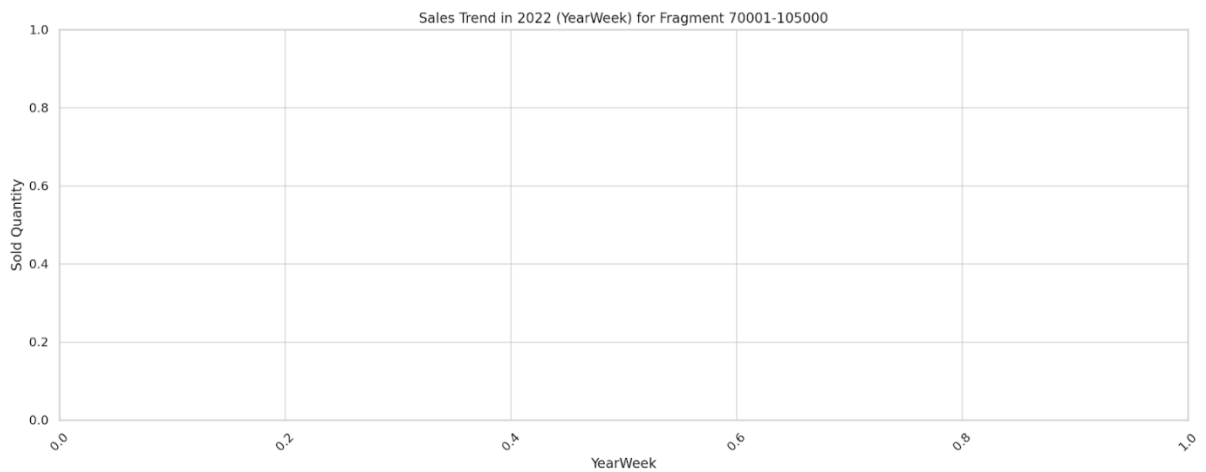


Figure 5-3 Result of Fragment 3

The next file (105,001 to 140,000) begins to look normal again.

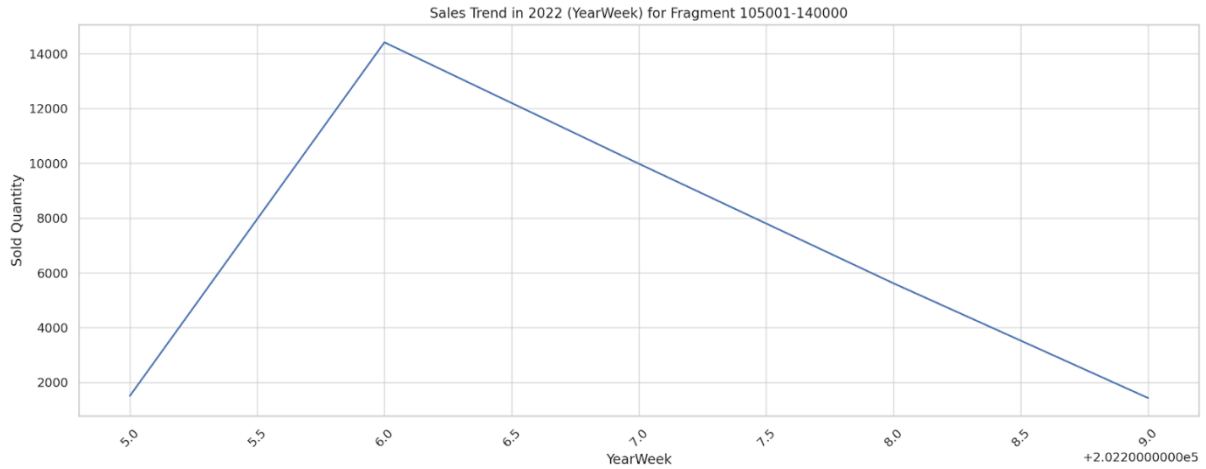


Figure 5-4 Result of Fragment 4

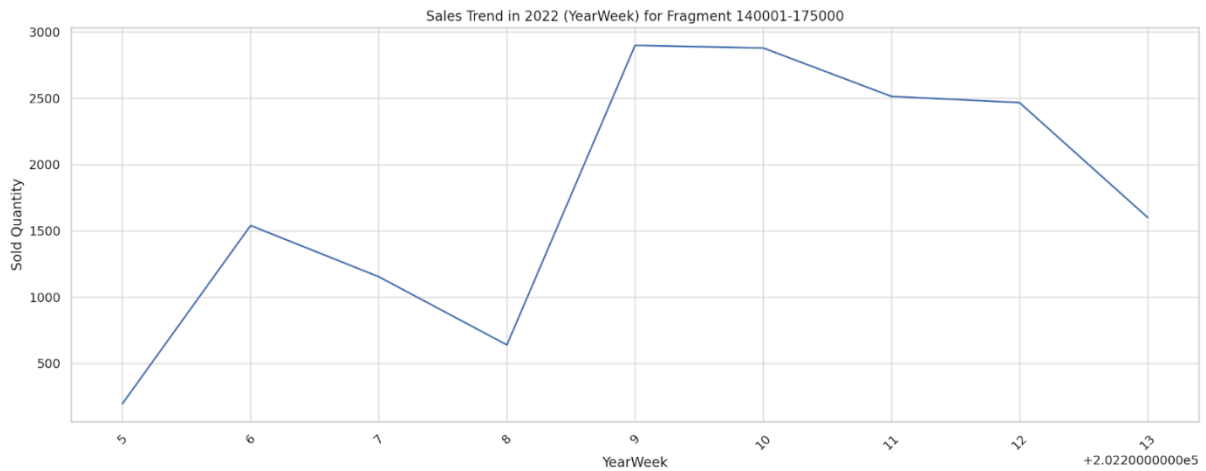


Figure 5-5 Result of Fragment 5

It stops at the seventh file due to large dataset, and then give the summarize results by combining previous fragments. The chart is not really look like a combination of above charts; however, it do look like the beginning of our traditional approach.

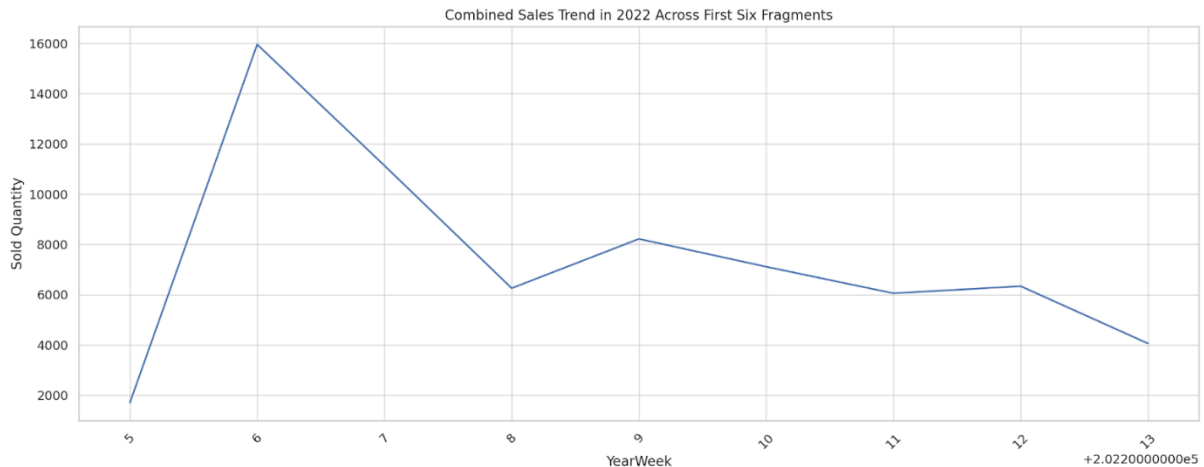


Figure 5-6 Result of 6 Fragments

It appears that we encountered an issue while attempting to visualize the sales trend for the seventh fragment. Given the challenges in processing and visualizing such a large dataset, I suggest focusing on summarizing the insights we've gathered so far from the analysis of the previous fragments.

From the analysis of the first six fragments, here are some overarching insights about the sales trends in 2022:

- Variability in Weekly Sales:** Across all fragments analyzed, we observed notable fluctuations in sales from week to week. This pattern suggests that sales are influenced by a variety of factors that vary over time.
- Potential Seasonal Influences:** While individual fragments provide limited insights into seasonal trends, some recurring peaks and troughs could hint at seasonal influences, such as holidays or weather-related shopping behaviors.
- Impact of Marketing and Promotions:** Peaks in sales may correlate with marketing campaigns, promotional events, or special discounts, indicating the effectiveness of these strategies in driving sales.
- Product and Market Dynamics:** Introduction of new products, changes in stock availability, and consumer preferences likely play a significant role in sales variations.
- External Factors:** External events, cultural occasions, and economic conditions can also impact consumer purchasing behavior, as evidenced by the fluctuations in sales.

Figure 5-7 Chat GPT-4 Analysis

Therefore, it can be concluded that Chat GPT-4 is not really effective for large dataset but there is a slight possibility that it provided a similar result to the traditional approach. Provided that we are not putting the process time into consideration.

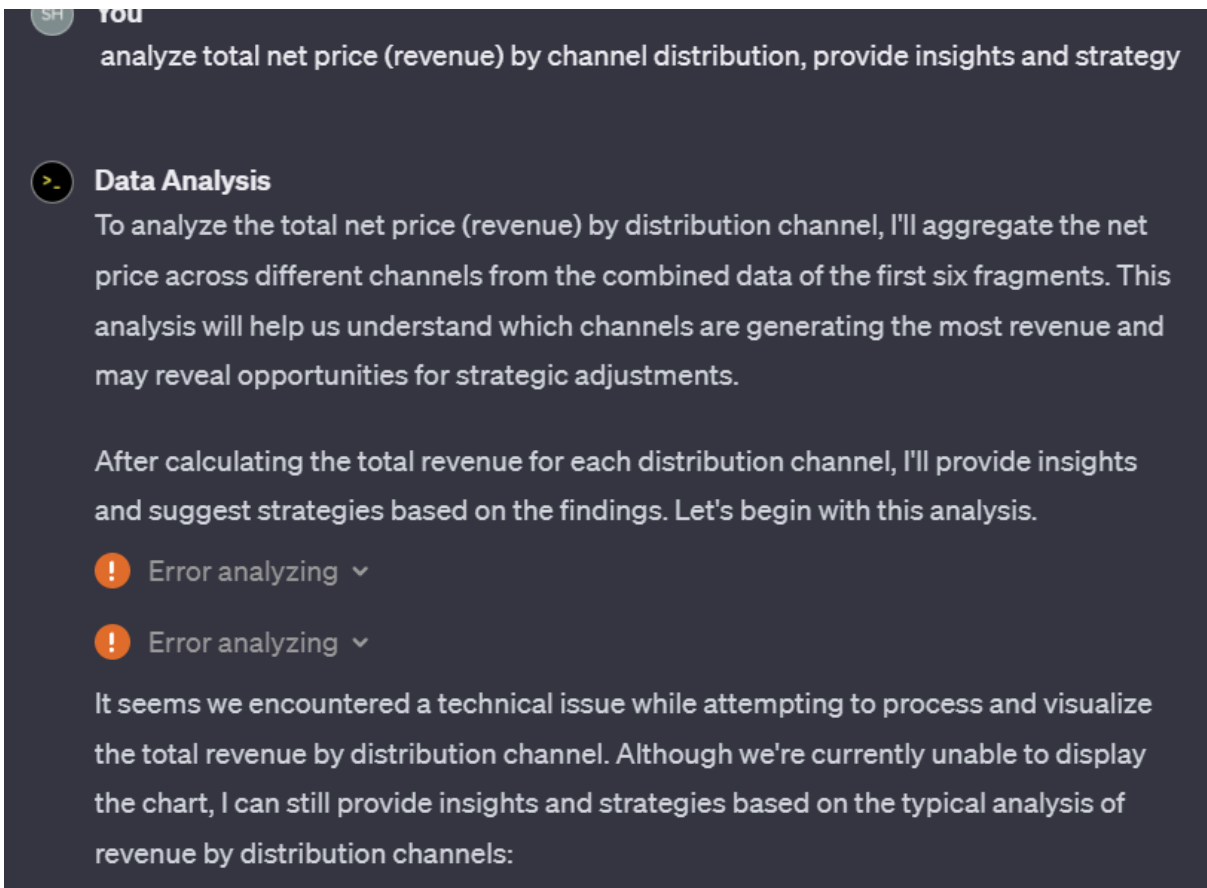


Figure 5-8 Chat GPT-4 Limitation

6. Conclusion

In conclusion, the segmentation analysis conducted in this report has provided valuable insights into both product and customer categories, laying the groundwork for informed decision-making and strategic planning. The diverse methodologies employed, including ABC analysis, demand stability assessment, visualization techniques, and the BCG Matrix, have collectively contributed to a comprehensive understanding of the company's product landscape.

The product segmentation has enabled the identification of key products based on their significance and contribution to the overall business. ABC analysis has facilitated the prioritization of products, allowing for a focused approach in resource allocation and inventory management. Additionally, the visualization techniques applied have enhanced the clarity and accessibility of complex data, aiding stakeholders in making more informed decisions.

On the customer segmentation front, the meticulous data processing and segmentation methodologies utilized have yielded a nuanced understanding of the customer base. The segmentation results provide actionable information that can be leveraged to tailor marketing strategies, enhance customer experiences, and optimize engagement. The development of buyer personas further refines this understanding, creating vivid profiles that guide personalized interactions with distinct customer segments.

References

- Ahmed, M., Sera, R., & Islam, S. M. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*. doi:10.3390/electronics9081295
- Cohen, M. A., & Ernst, R. (1988). Multi-item classification and generic inventory stock control policies. *Production and Inventory Management Journal*, 29(3), 6–8.
- Dinh-Khanh, P. (n.d.). Các bước của thuật toán k-Means Clustering. Retrieved November 9, 2022, from Deep AI KhanhBlog: https://phamdinhkhanh.github.io/deepai-book/ch_ml/KMeans.html
- Henderson, B.: *Corporate Strategy*. Abt Books Publisher, Cambridge (1979)
- Hong-Dien, L., Phuc-Son, N., Hoang-Uyen, P., & Van-Hinh, L. (2019). On a segmentation of Coopextra customers in Thu Duc district. *VNUHCM Journal of Economics, Business and Law*, 28-36.
- Marutho, D., Handaka, S. H., Wijaya, E., & Muljono. (2018). The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. 2018 International Seminar on Application for Technology of Information and Communication (pp. 533-538). Semarang: IEEE. doi:10.1109/ISEMANTIC.2018.8549751
- Partovi, F. Y., & Anandarajan, M. (2002). Classifying inventory using artificial neural network approach. *Computers and Industrial Engineering*, 41, 389–404.
- S.Khan, S., & Ahmad, A. (2004). Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters*, 1293-1302. doi:<https://doi.org/10.1016/j.patrec.2004.04.007>