Exploratory Data Analysis

- StudentID: 22100748

- Name: Yunyoung Choi

- 1st Major: Life Science

- 2nd Major: AI

This EDA report plays a crucial role in suggesting two main strategies: (1) tailoring marketing strategies by customer categorization and (2) utilizing the most effective distribution channels. To come up with these strategies, two datasets were used: `S_data.csv` *containing sales transaction data and* `HR_data.csv` *providing information about employees. Based on the univariate and multivariate analysis conducted in this EDA report, implementing these two strategies could enable the company to enhance overall profitability.*

1. Data overview

1. `S_data.csv`

- **Description**: This dataset contains sales transaction data for various items across different countries, regions, sales channels, and order priorities.
- **sample size**: 5000000 X 14
- **Key variables**(data type):
    - `Item Type`(object): The type of item sold
      (e.g. Office Supplies, Beverages, Cereal, ... , Cosmetics)
    - `Order Priority`(object): Priority of the order (e.g. H(High), L(Low), M(Medium), C(Critical))
    - `Units Sold`(int64): The number of units sold for each item in the transaction
    - `Unit Price`(float64): The price per unit of the item
    - `Unit Cost`(float64): The cost per unit of the item
    - `Total Revenue`(float64): The total revenue generated from the sale (`Units Sold` * `Unit Price`)
    - `Total Cost`(float64): The total cost incurred for the sale (`Units Sold` * `Unit Cost`)
    - `Total Profit`(float64): The total profit earned from the sale (`Total Revenue` - `Total Cost`)

2. `HR_data.csv`

- **Description**: This dataset contains information about employees, including their personal details, employment history, and location in the United States.
- **Sample Size**: 5000000 X 37
- **Key Variables**(data type):
  - `Age in Yrs.`(float64): Age of the employee in years (min: 21, max: 60)
  - `State`(object): State where the employee is located (OH, DC, CA, TX, ..., LA)

2. Univariate analysis

2.1 Average Profit Margin by Item Types in United States

To determine which `item type` generates the highest `profit margin` in the United States, the data was grouped by `Item Type` and analyzed. (In the dataset, there were data for many countries. However, to link it with the `HR_data`, only the data for the United States were utilized.)

The variables relevant for profit analysis include `Cost`, `Revenue`, and `Profit`. These variables are crucial for analyzing profitability, as they provide insights into the financial aspects of the transactions.

To calculate a more precise measure of **profitability**, new indexes were created:

1. `Net Profit Margin`

- This index represents **the net income or profit** generated as a **percentage of revenue**.
- It provides a clearer picture of the company's profitability by considering all expenses, not just the direct costs associated with producing or acquiring the items.
- The formula for calculating the `Net Profit Margin` is as follows:

$$NetProfitMargin = \frac{TotalProfit}{TotalRevenue} * 100$$

Following the calculation of the `Net Profit Margin`, the average `Net Profit Margin` was computed for each `Item Type` by taking the mean of the `Net Profit Margin` values corresponding to each item type in the dataset.

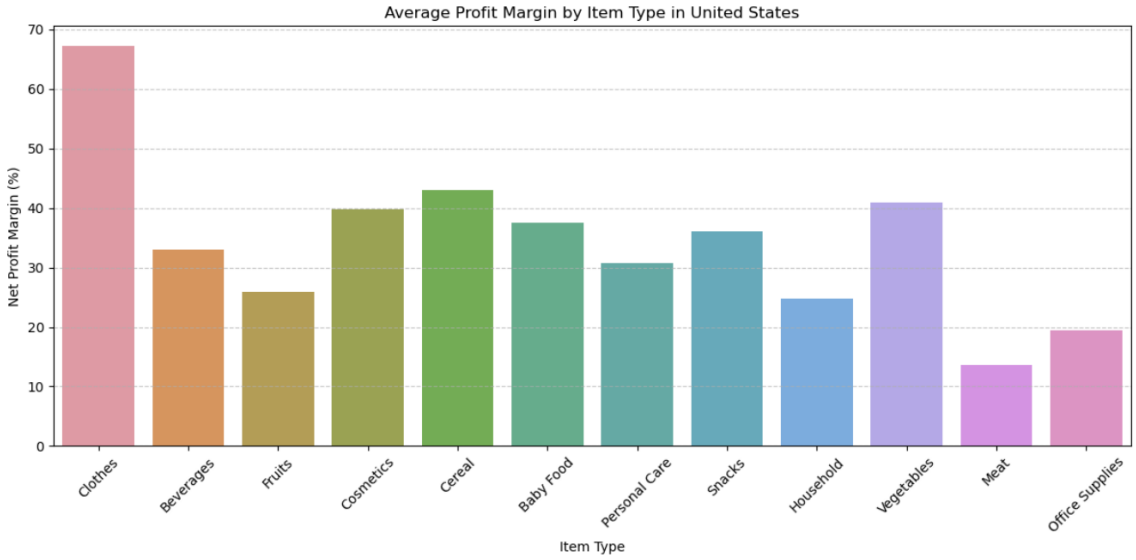

Figure 1. Average Profit Margin by Item Type in United States

2. `Profit per unit`

- To further compare with the `Net Profit Margin` results, a new index called the `Profit per unit` was also utilized.
- This index is obtained by calculating the difference between the `unit price` and the `unit cost` of the items sold.
- It represents the **profit earned per unit sold**.
- The formula for calculating the `Profit per unit` is as follows:

  `Profit per unit` = `Unit Price` - `Unit Cost`

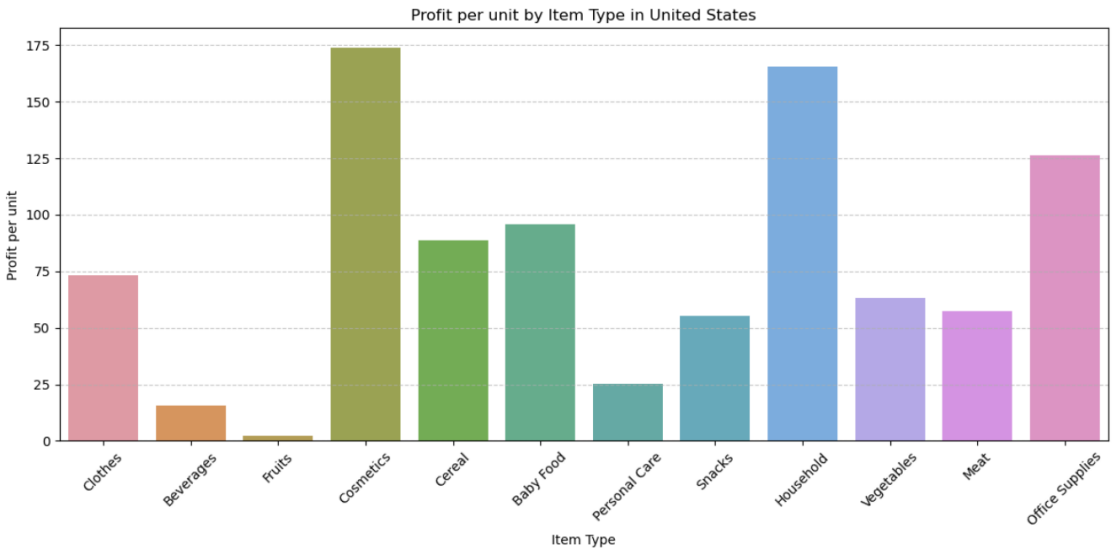

Figure 2. Profit per unit by Item Type in United States

In Figure 1 and Figure 2, we can observe two types of profits based on the item types: `Net Profit Margin` and `Profit per unit`. It appears that the analysis of the `Net Profit Margin` graph shows that `Clothes` have the highest margin, indicating that this item type

generates the highest proportion of profit relative to revenue. On the other hand, when examining the Profit per unit graph, `Cosmetics` and `Household` appear to have higher profit margins on a per-unit basis. **So, what should the company sell to maximize profits?**

This discrepancy underscores the importance of considering multiple factors when making business decisions. While `Cosmetics` and `Household` may exhibit higher profits per unit, `Clothes` boast a superior overall margin, indicating potentially greater long-term profitability when considering various realistic factors such as labor costs, taxes, and overhead expenses. Thus, focusing on the sale of `Clothes` could lead to sustainable profitability.

2.2 Units Sold by Order Date

To analyze when the most profitable item(`Clothes`) sells best in the `S_data`, the number of `units sold` was recorded based on the `order date`. Additionally, considering that the `sales channel` may also impact the number of `units sold`, the data was segmented into *Online* and *Offline* sales channels. This segmentation aims to understand whether there are any differences in sales patterns between online and offline channels for the `Clothes` item.

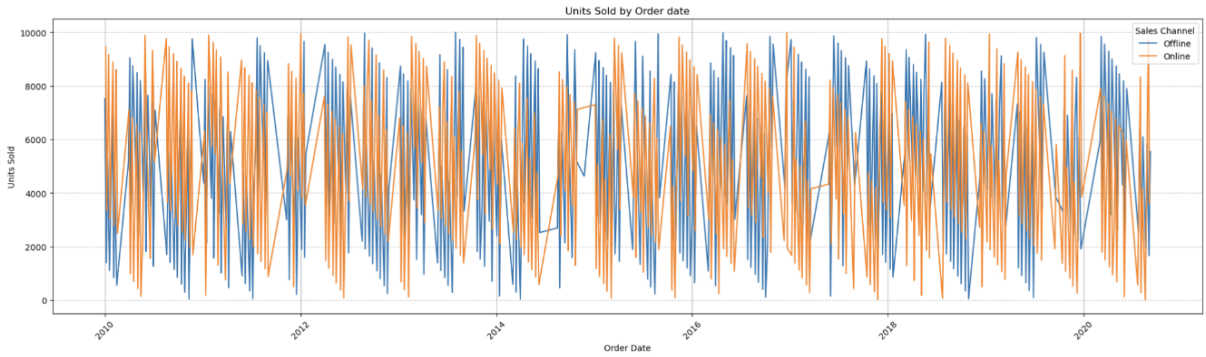

Figure 3. Units Sold by order date

The number of `units sold` for each item by `order date` shows a consistent pattern. For *Online* sales, the highest sales volumes are observed around January, April, July, and October, while for *Offline* sales, they peak around March, June, September, and December.

These patterns indicate that the `Sales Channel` may have varying impacts depending on the time of year. While this data alone does not provide a precise understanding of the influencing factors, selecting the appropriate sales channel according to the season could lead to higher profits. Further analysis incorporating additional variables such as marketing campaigns, consumer trends, and economic indicators may help uncover the specific factors influencing sales channel effectiveness.

2.3 Number of People grouped by age and states

To obtain insights into the population distribution across different age groups in each state based on the information provided in the `HR_data`:

1. confirmed that the age range is between 21 and 60 years.
2. categorized individual data into three age groups(reference):

- Young Adults: 18-35 years old
- Middle-aged Adults: 36-55 years old
- Older Adults: older than 55 years old

3. Grouped `States` and `Age Groups`

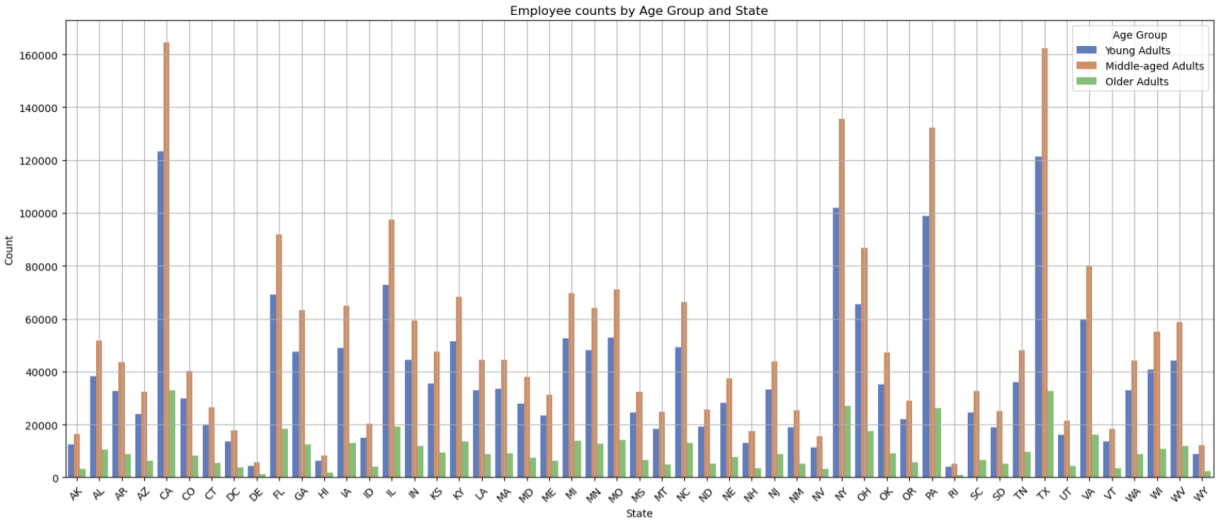

Figure 4. Employee Counts by Age Group and States

- On the X-axis, each state in the United States is represented.
- On the Y-axis, the number of individuals is represented based on their age group.

The graph clearly indicates significantly higher values for CA, NY, PA, and TX. When considering this alongside the earlier analyses, it underscores the potential effectiveness of concentrating sales efforts on `Clothes` items targeting middle-aged adults in these four key states. By strategically timing sales initiatives and tailoring strategies to specific age groups, it presents an opportunity to drive successsful sales and maximize profitability.

However, it's important to note that these insights are derived from specific organization HR data and may not fully represent the population and age distribution across the entire United States. This limitation underscores the need for additional data sources and thorough analysis when making strategic decisions.

3. Multivariate analysis

3.1 Correlation of cost and profit in `S_data`

Correlation analysis examines the strength and direction of the relationship between two numerical variables, measured by correlation coefficients. The correlation coefficient ranges

from -1 to 1. A coefficient closer to 1 indicates a strong positive correlation, meaning that as one variable increases, the other also tends to increase. A coefficient close to 0 suggests no linear relationship between the variables.

I wanted to investigate whether there is a relationship between `Price` and `Cost`, as well as between `Revenue` and `Profit`. For this purpose, I excluded categorical variables and included only meaningful numerical variables.

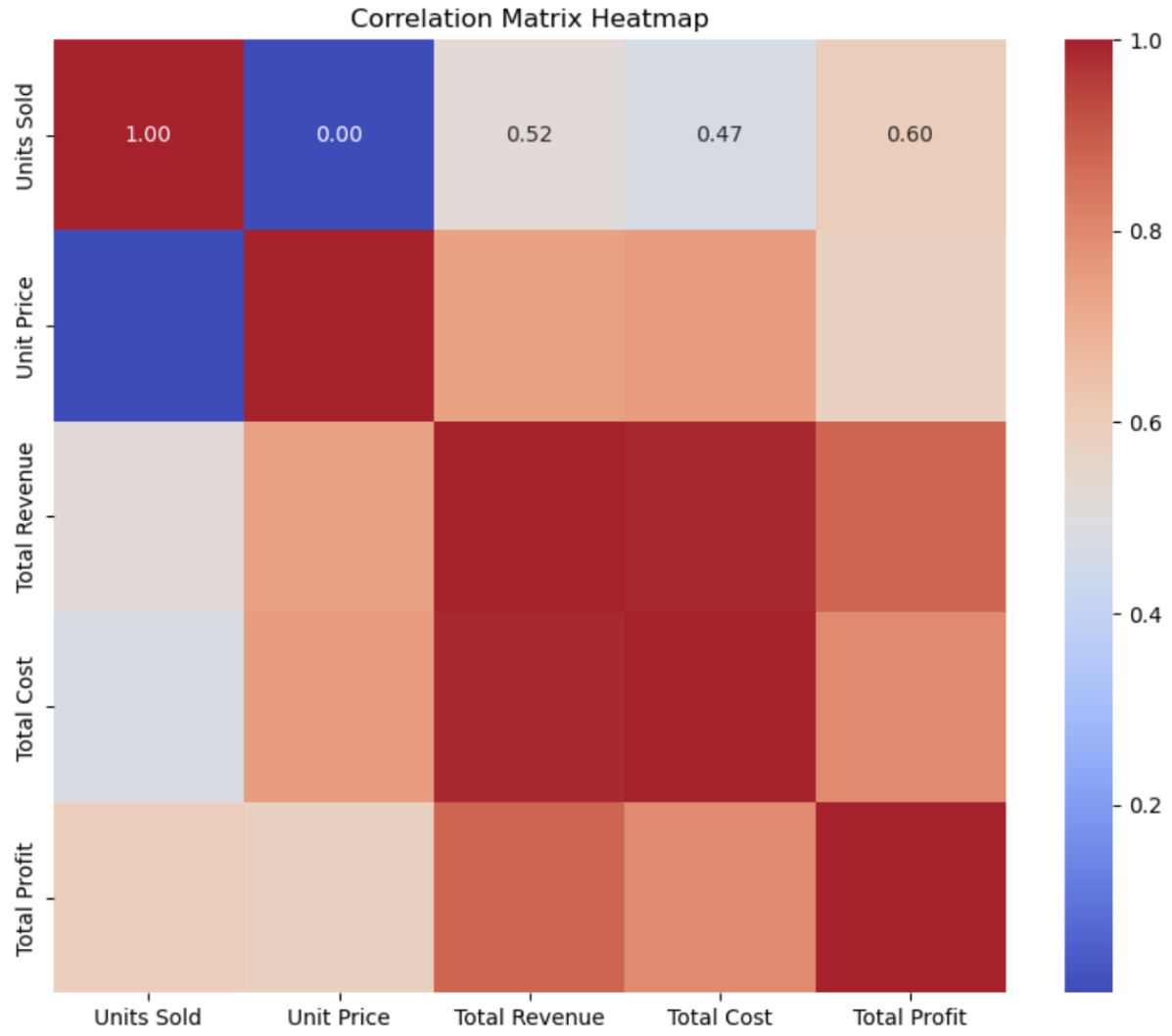

Figure 5. Correlation Matrix for USA data

The high correlations among `Unit Price` - `Total Cost`, `Total Revenue` - `Total Cost`, `Total Revenue` - `Total Profit`, and `Total Cost` - `Total Profit` indicate their interrelatedness. The positive correlation values suggest that when one variable's value is high, the other variable's value is likely to be high as well.

Particularly noteworthy is the strong relationship between `Total Revenue` and `Total Cost`. It can be attributed to the fact that, in the `S_data` dataset, each `item type` has only one associated cost.

```
USA_data.groupby('Item Type').nunique()
```

Item Type	Region	Country	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit	Profit_per_unit	Profit Margin
Baby Food	1	1	2	4	945	945	794	945	1	1	945	945	945	1	4
Beverages	1	1	2	4	946	946	807	946	1	1	946	946	946	1	2
Cereal	1	1	2	4	945	945	812	945	1	1	945	945	945	1	3
Clothes	1	1	2	4	945	945	831	945	1	1	945	945	945	1	3
Cosmetics	1	1	2	4	944	944	808	944	1	1	944	944	944	1	2
Fruits	1	1	2	4	944	944	801	944	1	1	944	944	944	1	3
Household	1	1	2	4	944	944	826	944	1	1	944	944	944	1	3
Meat	1	1	2	4	945	945	825	945	1	1	945	945	945	1	3
Office Supplies	1	1	2	4	944	944	808	944	1	1	944	944	944	1	3
Personal Care	1	1	2	4	944	944	832	944	1	1	944	944	944	1	3
Snacks	1	1	2	4	945	945	825	945	1	1	945	945	945	1	3
Vegetables	1	1	2	4	945	945	796	945	1	1	945	945	945	1	2

Table 1. The unique variable numbers in USA_data

As seen in Table 2, a subset of data for the `Clothes` item type reveals identical values for `Unit Price`, `Unit Cost`, and `Profit per unit`. Therefore, while correlations provide insights, they alone cannot precisely determine cause and effect. Additional variables such as delivery location, workforce, etc., are necessary to understand what influences changes in `Units Sold`.

3.2 ANOVA analysis of Units Sold by Order Priority

ANOVA (Analysis of Variance) analysis is a statistical method used to compare the means of three or more groups to determine if there are statistically significant differences between them. It assesses whether the variability between group means is greater than the variability within groups, considering both the differences between group means and the variation within each group.

$H0 : \mu1 = \mu2 = \cdots = \mu i$
$H1$: at least two means differ

As discussed in the previous section *3.1 Correlation of cost and profit in `S_data`*, an analysis was conducted to explore the factors influencing `Units Sold`, particularly examining its relationship with `Order Priority`. The null hypothesis (H0) posits that the means of `Units Sold` across different `Order Priority` categories are equal (μ1 = μ2 = · · · = μl), while the alternative hypothesis (H1) suggests that at least two of the means differ. This hypothesis testing enables to assess whether `Order Priority` significantly impacts the mean `Units Sold`.

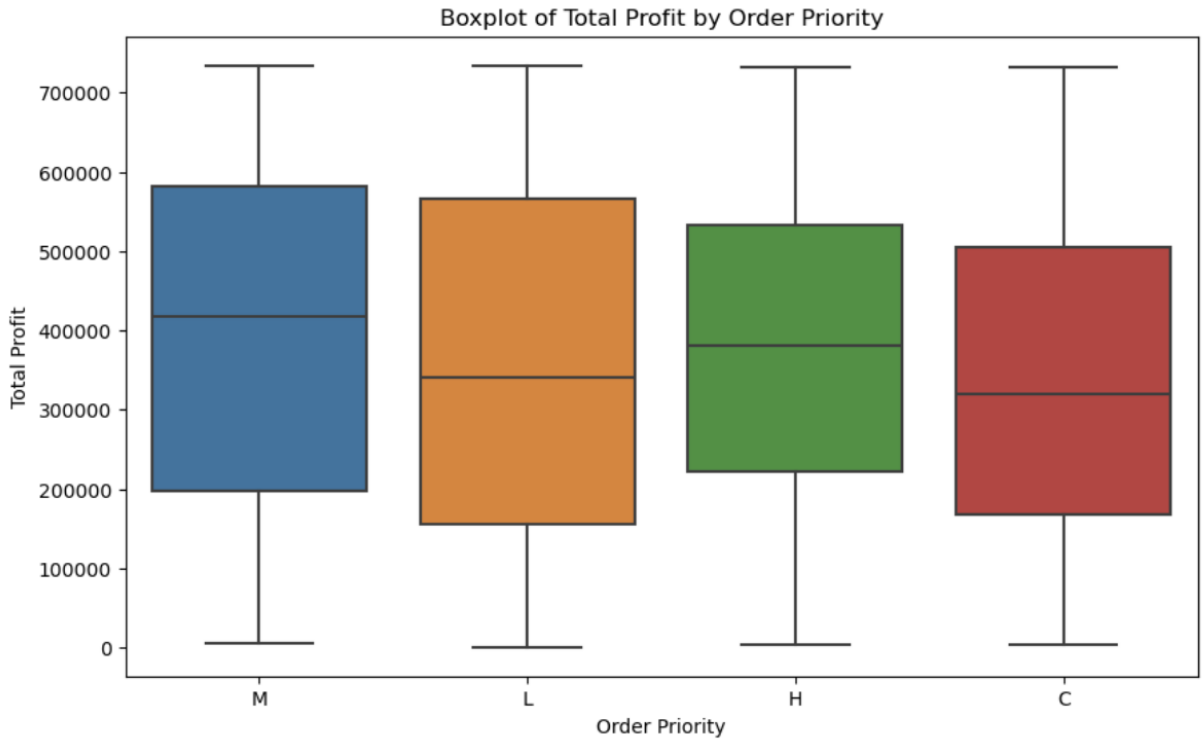

Figure 6. Boxplot of Total Profit by order priority

- ANOVA table

	sum_sq	df	F	PR(>F)
C(OrderPriority)	7.473710e+11	3.0	5.602042	0.000797
Residual	9.939079e+13	2235.0	NaN	NaN

Table 2. ANOVA analysis of Units Sold by Order Priority

Table 2 provides information about the analysis of variance for the relationship between `OrderPriority` and `TotalProfit` in the dataset.

Since the **p-value** (PR(>F)) associated with the **F-statistic** is 0.000797, which is less than the significance level (typically 0.05), it is possible to reject the null hypothesis.

Therefore, we have sufficient evidence to conclude that there is a significant difference in the mean `TotalProfit` based on `OrderPriority`.

4. Suggestion

Based on the results obtained from both *univariate* and *multivariate* analysis, two strategies can be suggested:

1. **Optimize Sales Product and Customers**

   By prioritizing items that generate significant profits, such as `Clothes`(Figure 1), the sales product can be optimized. This optimization means **focusing on high-profit items** and adjusting the product mix based on key sales regions and channels.

   In addition, utilizing customer categorization can be helpful to identify specific customer groups and **develop targeted marketing** strategies for each group. For example, targeting regions with high purchasing power among `middle-aged` consumers(Figure 4) can be a good strategy based on this EDA results.

2. **Optimize Sales Channel**

   Sales Channel Optimization aims to **improve sales performance** by identifying and utilizing **the most effective distribution channels**. This involves analyzing sales data to understand channel effectiveness, implementing strategies to enhance customer engagement, exploring new channels, and optimizing existing ones.

   For example, online sales are better in certain months, like January and April, while offline sales do well in different months, like March and June(Figure 3). So, the company can adjust things like **how much they stock, their promotions, and their ads to make the most of each sales channel** at the right times. They could also improve their online shopping experience or open more physical stores where sales are good. By optimizing sales channels based on the EDA insights, the company can enhance customer satisfaction, increase sales revenue, and improve overall profitability.

If you want to see the detailed code including all the figures and tables, please refer to the following link:
https://github.com/foryourjoy/EDA-BigDataDesign