Project Title:
In-Depth Analysis of Supermarket Sales Across Three Branches

Project overview

Description:
This project involves an in-depth analysis of sales data from three supermarket branches over a three-month period (January to March). The goal is to explore sales trends, customer behavior, and branch performance to uncover valuable insights that can drive business improvements. The analysis covers various aspects, including sales by product line, customer demographics, peak shopping times, and customer satisfaction.

The dataset, which is available on Kaggle, contains detailed information about sales transactions, including customer types, product categories, pricing, payment methods, and customer ratings. This analysis highlights key insights on gross income, sales performance, and customer feedback, helping businesses optimize their strategies.

Key Analyses:

1. **Customer Behavior Analysis:**
    1. Product line popularity by customer type and gender.
    2. Identification of peak shopping times by day and hour.
    3. Deeper insights by combining customer type, gender, and time.
2. **Branch Performance Analysis:**
    1. Total sales performance by branch.
    2. Gross income comparison.
    3. Customer satisfaction analysis across branches.
3. **Product Insights:**
    1. Sales volume and gross income contribution by product line.
    2. Average gross income per sale and the relationship between sales volume and gross income.
    3. Comparison of product line performance by customer type.

1. Data reading

First, i imported the required libraries for the analysis

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

After downloading the dataset from Kaggle, I used the Pandas read_csv function to load it into a DataFrame:

```
df = pd.read_csv("supermarket_sales.csv")
```

To get a quick understanding of the dataset, I used the head() function to see the first few rows:

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3

Next, I examined the structure and size of the dataset:

```
df.shape
```

output:

(1000, 17)

The dataset contains **1000 rows** and **17 columns**.

I used describe() to get some basic statistics about the dataset:

```
df.describe()
```

output:

	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1.000000e+03	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905e+00	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	6.131498e-14	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905e+00	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905e+00	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905e+00	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905e+00	22.445250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	993.00000	4.761905e+00	49.650000	10.00000

Finally, to understand the type of data in each column, I used:

```
df.info()
```

output:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Invoice ID               1000 non-null   object
 1   Branch                   1000 non-null   object
 2   City                     1000 non-null   object
 3   Customer type            1000 non-null   object
 4   Gender                   1000 non-null   object
 5   Product line             1000 non-null   object
 6   Unit price               1000 non-null   float64
 7   Quantity                 1000 non-null   int64
 8   Tax 5%                   1000 non-null   float64
 9   Total                    1000 non-null   float64
 10  Date                     1000 non-null   object
 11  Time                     1000 non-null   object
 12  Payment                  1000 non-null   object
 13  cogs                     1000 non-null   float64
 14  gross margin percentage  1000 non-null   float64
 15  gross income             1000 non-null   float64
 16  Rating                   1000 non-null   float64
dtypes: float64(7), int64(1), object(9)
memory usage: 132.9+ KB
```

2. Data QA:

The data is clean and ready for analysis, there is no null values, no mismatched character every thing looks good, one thing is that i chose to first merge Date & Time to simplify the dataset.

Merging Date & Time to one column and setting this new column to datetime type:

```
df["Datetime"] = pd.to_datetime(df["Date"] + " " + df["Time"])
df.drop(["Date", "Time"], axis = 1, inplace = True)
```

3. Exploratory data analysis (EDA)

3.1 Customer behavior analysis

This section covers customer behavior, by analyzing product line popularity by customer type and gender, identifying peak shopping times by day and hour, and combining customer type, gender, and time for deeper insights. This analysis will help understand which products are most popular among different customer demographics and when peak shopping times occur.

3.1.1 Analyzing product line popularity by customer type

To understand which product lines are favored by different customer types, i performed a grouped analysis based on the `Customer type` and `Product line`, summarizing the total quantity purchased for each combination.

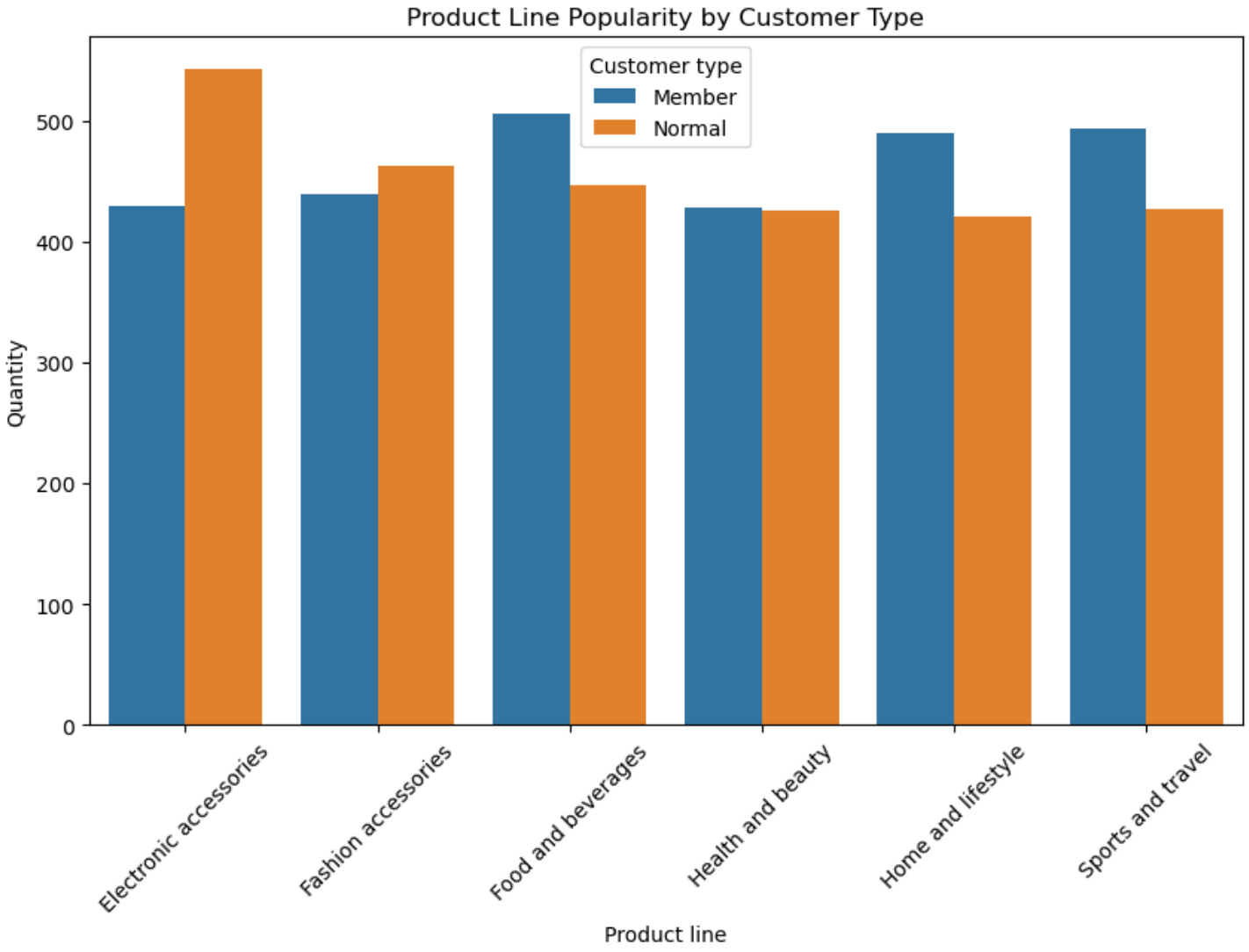
```
customer_type_product_popularity = df.groupby(['Customer type', 'Product
line'])['Quantity'].sum().reset_index()

customer_type_product_popularity.pivot_table(
    index = "Product line",
    columns = "Customer type",
    values = "Quantity",
).style.background_gradient(cmap = "RdYlGn", axis = None)
```

output:

Here, a table display product line popularity by customer type with a color gradient representing quantities.

Customer type	Member	Normal
Product line		
Electronic accessories	429	542
Fashion accessories	439	463
Food and beverages	506	446
Health and beauty	428	426
Home and lifestyle	490	421
Sports and travel	493	427

```
plt.figure(figsize=(10,6))
sns.barplot(x='Product line', y='Quantity', hue='Customer type',
data=customer_type_product_popularity)
plt.title('Product Line Popularity by Customer Type')
plt.xticks(rotation=45)
plt.show()
```



Interpretation of Results:

- For **Normal** customers, the most popular product line is **Electronic Accessories**, followed by **Fashion Accessories** and **Food and Beverages**. Interestingly, **Health and Beauty**, **Home and Lifestyle**, and **Sports and Travel** are equally favored.
- For **Member** customers, the top-selling product line is **Food and Beverages**, followed by **Home and Lifestyle** and **Sports and Travel**. It is notable that **Electronic Accessories** and **Fashion Accessories** share similar sales, highlighting different priorities compared to Normal customers.

This suggests that Normal customers prioritize tech-related products and fashion, while Members tend to spend more on food and household items, indicating potential differences in customer loyalty and spending patterns.

3.1.2 Analyzing product line popularity by gender

Next, i explored product line popularity based on customer gender, which helps in tailoring product recommendations and marketing campaigns.

```
gender_product_popularity = df.groupby(['Gender', 'Product
line'])['Quantity'].sum().reset_index()
```

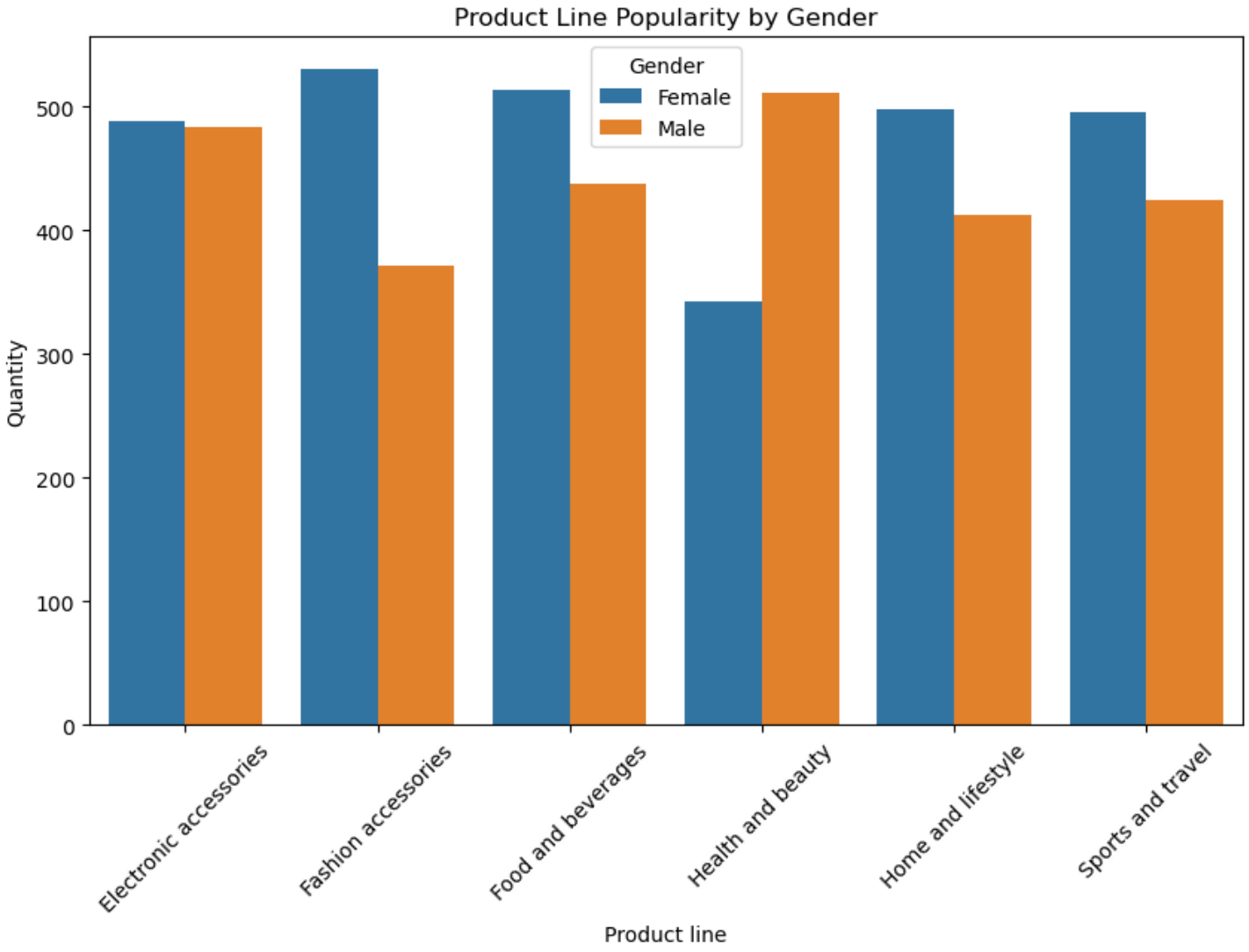
```
gender_product_popularity.pivot_table(
    index = "Product line",
    columns = "Gender",
    values = "Quantity",
).astype(int).style.background_gradient(cmap = "RdYlGn", axis = None)
```

output:

A table with gender-based product line popularity is displayed with quantities visually represented by a gradient.

Gender	Female	Male
Product line		
Electronic accessories	488	483
Fashion accessories	530	372
Food and beverages	514	438
Health and beauty	343	511
Home and lifestyle	498	413
Sports and travel	496	424

```
plt.figure(figsize=(10,6))
sns.barplot(x='Product line', y='Quantity', hue='Gender', data=gender_product_popularity)
plt.title('Product Line Popularity by Gender')
plt.xticks(rotation=45)
plt.show()
```



Interpretation of Results:

- For **Females**, **Fashion Accessories** dominate sales, followed by a relatively even distribution across the other categories, with the exception of **Health and Beauty**, which shows significantly lower sales. This suggests that fashion is a key driver for female shoppers, while health products are less appealing.
- For **Males**, **Health and Beauty** leads, followed by **Electronic Accessories**. The rest of the categories, including **Food and Beverages**, **Home and Lifestyle**, and **Sports and Travel**, exhibit similar popularity, with **Fashion Accessories** being the least popular among male customers.

These insights show clear gender preferences, with fashion being a female-driven category and health products resonating more with males, allowing for gender-specific promotions.

3.1.3 Identifying peak shopping times by day and hour

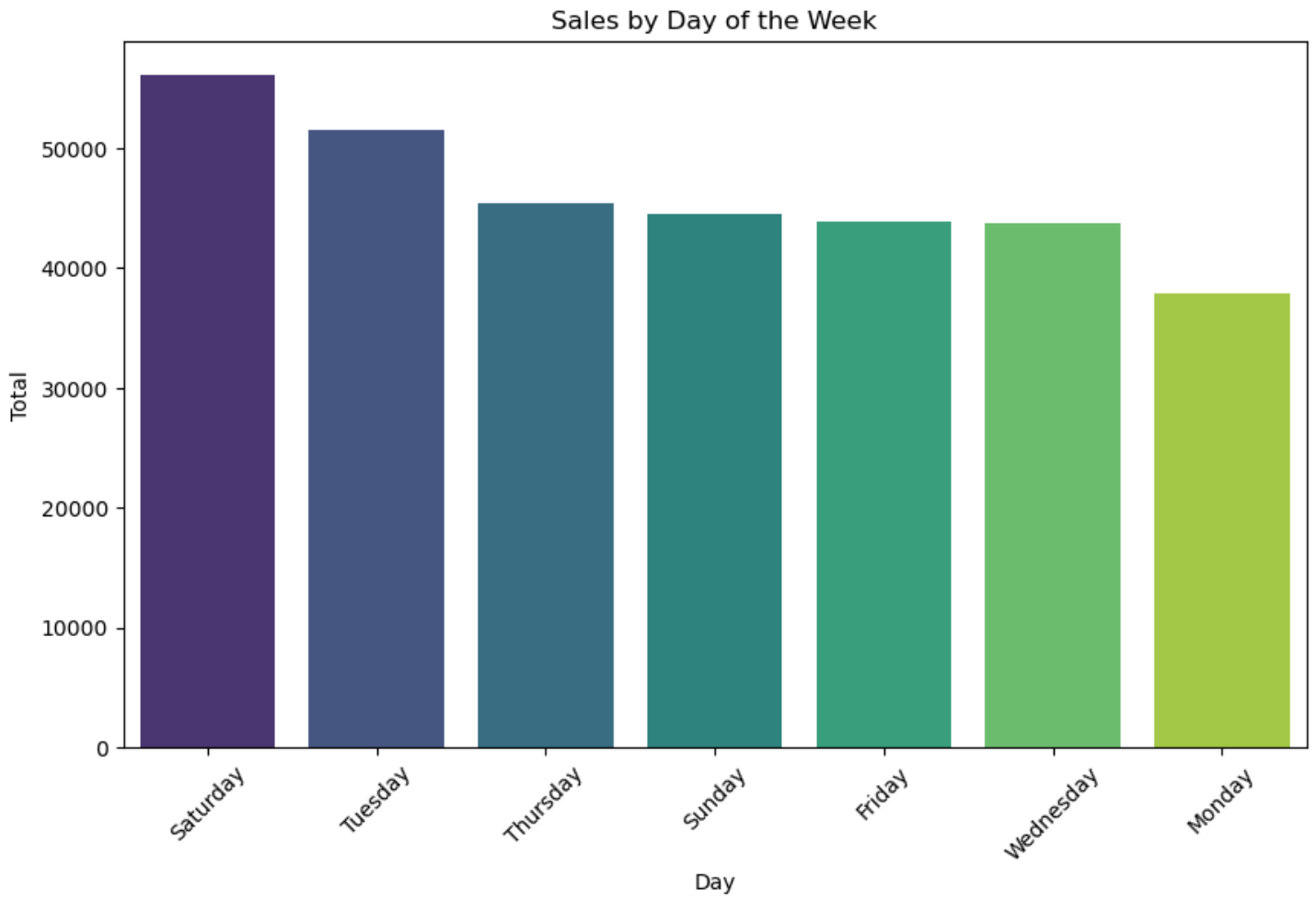
I further analyzed peak shopping times to determine which days and hours have the highest sales. This can help in optimizing staffing, marketing campaigns, and promotions.

First, i assigned two new columns to represent `Day` and `Hour`, and analyzed sales by day of the week:

```
df = df.assign(
```

```
    Day = df['Datetime'].dt.day_name(),
    Hour = df['Datetime'].dt.hour,
)

sales_by_day = df.groupby('Day')['Total'].sum().reset_index().sort_values(by='Total',
ascending=False)

plt.figure(figsize=(10,6))
sns.barplot(x='Day', y='Total', data=sales_by_day)
plt.title('Sales by Day of the Week')
plt.xticks(rotation=45)
plt.show()
```

output:



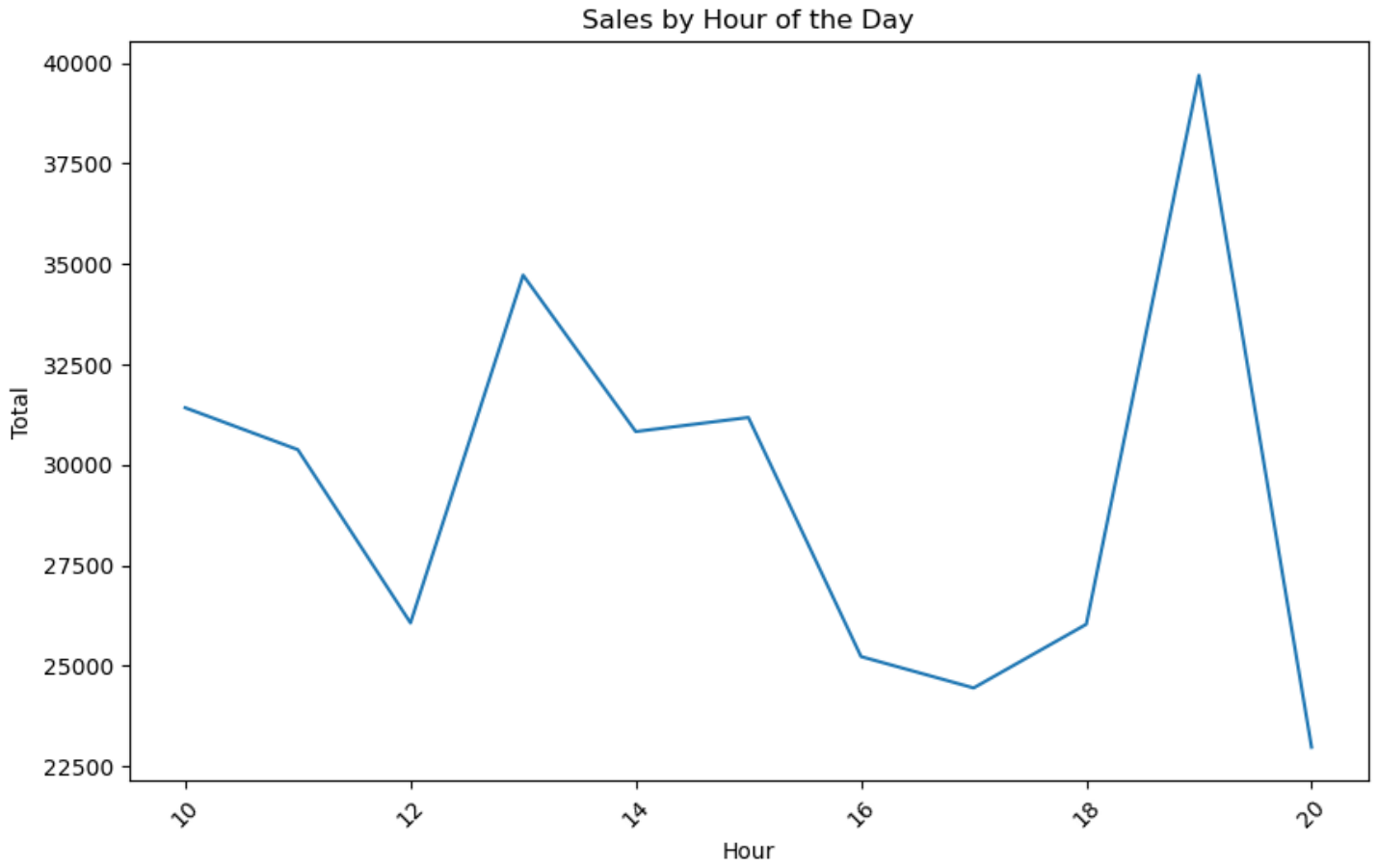
Interpretation of Results:

- **Saturday** and **Tuesday** are the most active shopping days, suggesting weekends and early-week promotions could be effective.
- **Monday** is the least active, potentially an opportunity for discounts or special deals to boost sales.
- The remaining days exhibit relatively consistent sales, indicating stable weekday activity.

Next, i examined sales by hour of the day:

```
sales_by_hour = df.groupby('Hour')['Total'].sum().reset_index().sort_values(by='Hour',
ascending=False)

plt.figure(figsize=(10,6))
sns.lineplot(x='Hour', y='Total', data=sales_by_hour)
plt.title('Sales by Hour of the Day')
plt.xticks(rotation=45)
plt.show()
```

output:



Interpretation of Results:

- Sales show distinct peaks at the **start of the day** and at **1 PM**, which likely corresponds to the lunchtime shopping rush.
- There is a steady decline after 2 PM until a final peak at **7 PM**, suggesting post-work or evening shoppers.
- Sales drop off after 8 PM, indicating that late-night shopping is minimal.

These insights can help in scheduling staff during peak hours and planning promotions to align with high-traffic times.

3.1.4 Combining customer type, gender, and time for deeper insights

To dig deeper, i analyzed customer behavior by combining customer type, gender, and time (day and hour) to see how these factors influence sales:

```
customer_behavior_analysis = (df.groupby(['Customer type', 'Gender', 'Day', 'Hour'])['Total']
                              .sum()
                              .reset_index()
                              .sort_values(by='Total', ascending=False))
```

output:

	Customer type	Gender	Day	Hour	Total
179	Normal	Female	Saturday	19	4171.0200
58	Member	Female	Tuesday	15	3963.4350
62	Member	Female	Tuesday	19	3295.9605
191	Normal	Female	Thursday	10	3188.0730
144	Member	Male	Wednesday	15	3076.4055
...	...	...	...	...	...
168	Normal	Female	Monday	18	49.7700
294	Normal	Male	Wednesday	16	44.5935
154	Normal	Female	Friday	15	41.3910
290	Normal	Male	Wednesday	12	32.1405
75	Member	Male	Friday	11	20.6850

Interpretation of Results:

- **Female** customers dominate the top spending slots, with the highest sales occurring on **Saturday at 7 PM**, followed by **Tuesday at 3 PM** and **7 PM**, and **Thursday at 10 AM**.
- **Male** customers appear only in **Wednesday at 3 PM**, indicating that men tend to shop less frequently but still make notable purchases mid-week.

These insights suggest that marketing campaigns targeting female customers on weekends and late afternoons could be highly effective. For male customers, mid-week promotions may yield better results.

3.2 Branch performance analysis

In this section, i analyzed branch performance through sales, customer satisfaction (ratings), and gross income. Additionally, i explored the performance of each branch based on product lines and customer types. This gives us a comprehensive view of how well each branch is performing in terms of sales volume, customer feedback, and profitability.

3.2.1 Sales performance by branch

I started by analyzing the total sales for each branch:

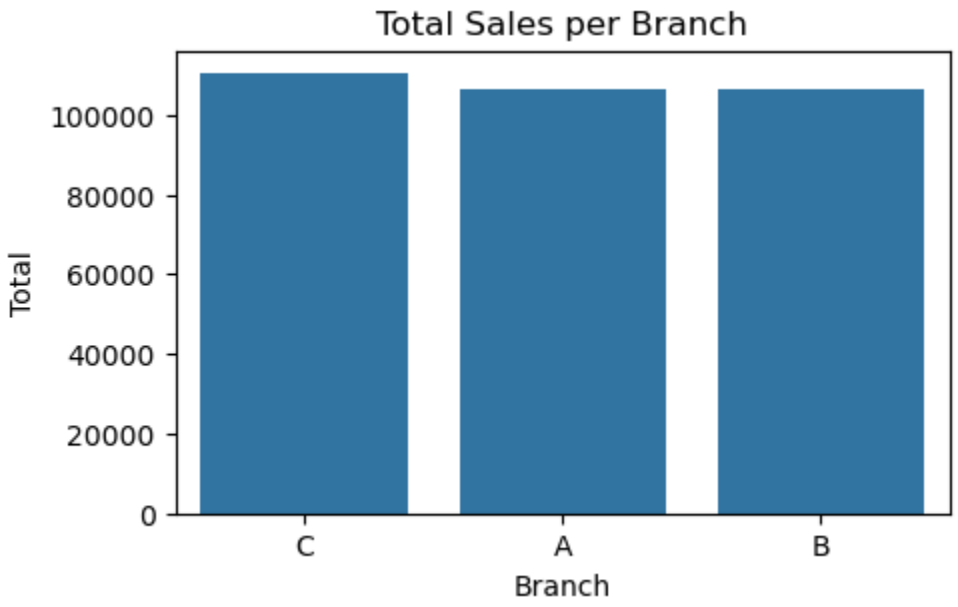
```
sales_per_branch = df.groupby('Branch')['Total'].sum().reset_index().sort_values(by='Total',
ascending=False)

sales_per_branch
```

output:

	Branch	Total
2	C	110568.7065
0	A	106200.3705
1	B	106197.6720

```
plt.figure(figsize=(5, 3))
sns.barplot(x='Branch', y='Total', data=sales_per_branch)
plt.title('Total Sales per Branch')
plt.show()
```

output:



Results interpretation:

Results indicate that **Branch C** marginally outperforms **Branches A** and **B**. Specifically:

- **Branch C** leads with 4.05% higher total sales than **Branch A**.
- **Branch A** outperforms **Branch B** by just **0.01%**, showing that their sales performances are the same.

Despite the slight differences, **Branch C** remains the top performer in sales, which suggests stronger customer traffic or higher-priced transactions.

3.2.2 Customer satisfaction (Ratings) by branch

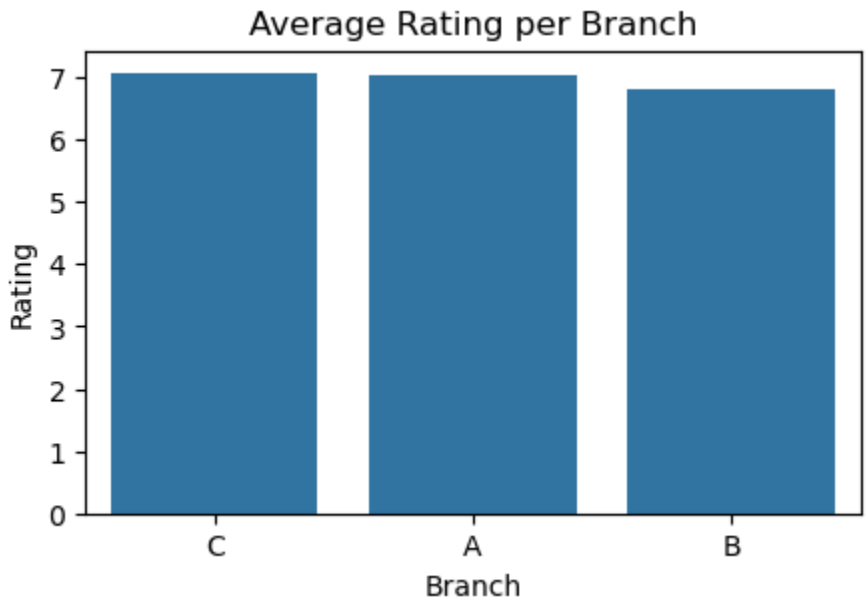
Next, i analyzed the average customer ratings for each branch to get insights into customer satisfaction:

```
ratings_per_branch =
df.groupby('Branch')['Rating'].mean().reset_index().sort_values(by='Rating', ascending=False)

ratings_per_branch
```

output:

	Branch	Rating
2	C	7.072866
0	A	7.027059
1	B	6.818072

```
plt.figure(figsize=(5, 3))
sns.barplot(x='Branch', y='Rating', data=ratings_per_branch)
plt.title('Average Rating per Branch')
plt.show()
```



Results interpretation:

- Branch C has the highest customer satisfaction, outperforming Branch A by a slight margin of 0.65%.
- Branch A follows, outperforming Branch B by 3.8%.

While the difference between Branches C and A is minor, Branch B trails behind significantly in customer satisfaction. This could point to service quality or product availability issues in Branch B that require attention.

3.2.3 Gross income comparison by branch

I also examined the gross income for each branch to assess profitability:

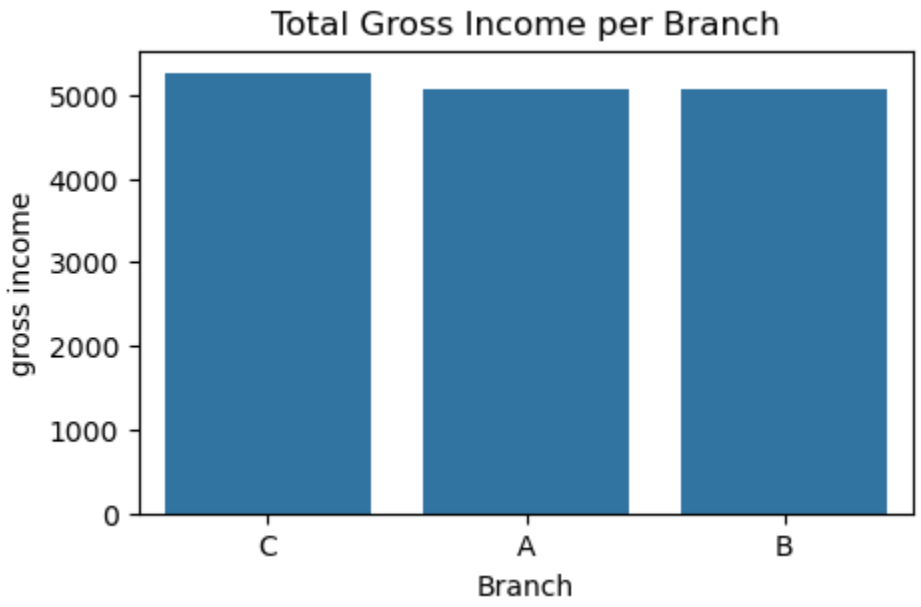
```
gross_income_per_branch = (df.groupby('Branch')['gross income']
                           .sum()
                           .reset_index()
                           .sort_values(by='gross income', ascending=False))

gross_income_per_branch
```

output:

	Branch	gross income
2	C	5265.1765
0	A	5057.1605
1	B	5057.0320

```
plt.figure(figsize=(5, 3))
sns.barplot(x='Branch', y='gross income', data=gross_income_per_branch)
plt.title('Total Gross Income per Branch')
plt.show()
```

output:



Results interpretation:

Results show a similar pattern as sales:

- Branch C leads by a 4.5% margin over Branch A.
- Branch A barely edges out Branch B with a 0.01% difference.

Gross income trends are in line with total sales performance, confirming that Branch C is the most profitable, while Branch B lags slightly behind. These numbers align with the higher sales performance of Branch C and A.

3.2.4 Combining sales, ratings, and gross income for a complete view

By combining total sales, average ratings, and gross income, i got a full picture of each branch's performance:

```
branch_performance = pd.merge(sales_per_branch, ratings_per_branch, on='Branch')
branch_performance = pd.merge(branch_performance, gross_income_per_branch, on='Branch')

branch_performance.columns = ['Branch', 'Total Sales', 'Average Rating', 'Total Gross
Income']
```

```
branch_performance = branch_performance.sort_values(by='Total Sales', ascending=False)

branch_performance
```

output:

	Branch	Total Sales	Average Rating	Total Gross Income
0	C	110568.7065	7.072866	5265.1765
1	A	106200.3705	7.027059	5057.1605
2	B	106197.6720	6.818072	5057.0320

The combined analysis highlights the dominance of Branch C in all categories (sales, customer satisfaction, and gross income). Branch A consistently performs better than Branch B, but only marginally. This suggests that Branch B may require strategic interventions to improve both customer satisfaction and profitability.

3.2.5 Sales by product line for each branch

To get deeper insights, i analyzed how each product line performs across branches:

```
product_sales_per_branch = (df.groupby(['Branch', 'Product line'])['Total']
                            .sum()
                            .reset_index()
                            .sort_values(by=['Branch', 'Total'], ascending=False))

product_sales_per_branch.pivot_table(
    index="Product line",
    columns="Branch",
    values="Total"
).style.format("{:.2f}").background_gradient(cmap="RdYlGn", axis=None)
```

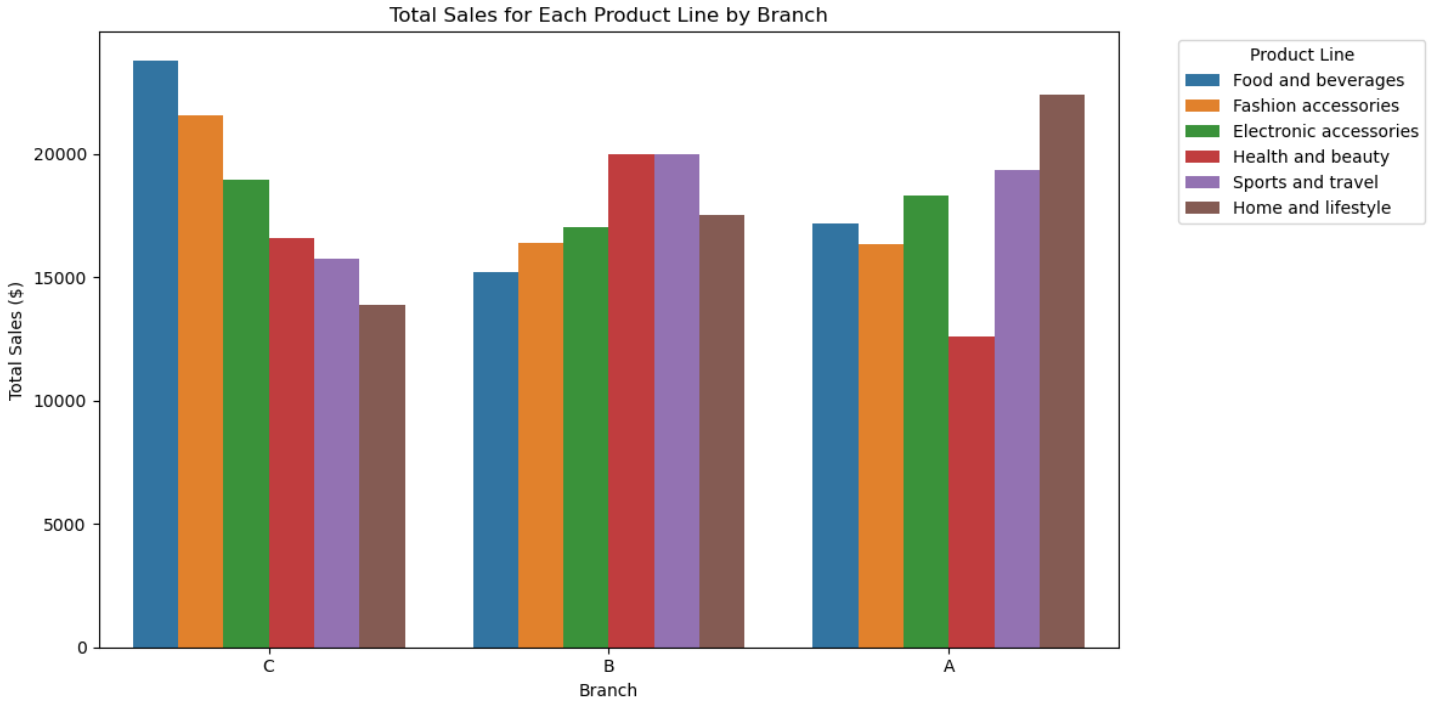
output:

Branch	A	B	C
Product line			
Electronic accessories	18317.11	17051.44	18968.97
Fashion accessories	16332.51	16413.32	21560.07
Food and beverages	17163.10	15214.89	23766.85
Health and beauty	12597.75	19980.66	16615.33
Home and lifestyle	22417.20	17549.16	13895.55
Sports and travel	19372.70	19988.20	15761.93

```
plt.figure(figsize=(12,6))
sns.barplot(x='Branch', y='Total', hue='Product line', data=product_sales_per_branch)
plt.title('Total Sales for Each Product Line by Branch')
plt.ylabel('Total Sales ($)')
```

```
plt.xlabel('Branch')
plt.legend(title='Product Line', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

output:



Results interpretation:

The heatmap and bar plots reveal the following patterns:

- Electronic Accessories: Sales are well-distributed across branches, but Branch C has the highest sales, followed by Branch A. Branch B shows a slightly lower performance.
- Fashion Accessories: Branch C clearly stands out with the highest sales. Branches A and B have similar but lower sales, indicating room for growth in these branches.
- Food and Beverages: Branch C dominates in this category with significantly higher sales. Branch A shows moderate sales, while Branch B falls behind, suggesting the need to improve this product line.
- Health and Beauty: Branch B leads with the highest sales, followed by Branch C. Branch A significantly lags in this category, indicating a need for more targeted marketing or product range improvements.
- Home and Lifestyle: Branch A outperforms the other branches, indicating a strong demand for this category. Branch C has the lowest sales here, suggesting an opportunity to boost marketing or stock availability in this segment.
- Sports and Travel: Branch B excels in this category, with Branch A showing good performance. Branch C lags behind, potentially pointing to a missed opportunity for Branch C in promoting or stocking these products.

Key Insights:

- Branch C excels in Fashion Accessories and Food and Beverages, indicating a strong customer preference for these categories.
- Branch A dominates in Home and Lifestyle and performs well in Sports and Travel, but struggles with Health and Beauty.

- Branch B leads in Health and Beauty and Sports and Travel, but needs to address lower performance in Food and Beverages and Electronic Accessories.

This analysis suggests potential areas for improvement. For example, Branch A could focus on boosting Health and Beauty sales, while Branch B should aim to improve performance in Food and Beverages and Electronic Accessories.

3.2.6 Comparing performance by customer type

Lastly, i compared the performance of each branch by customer type:

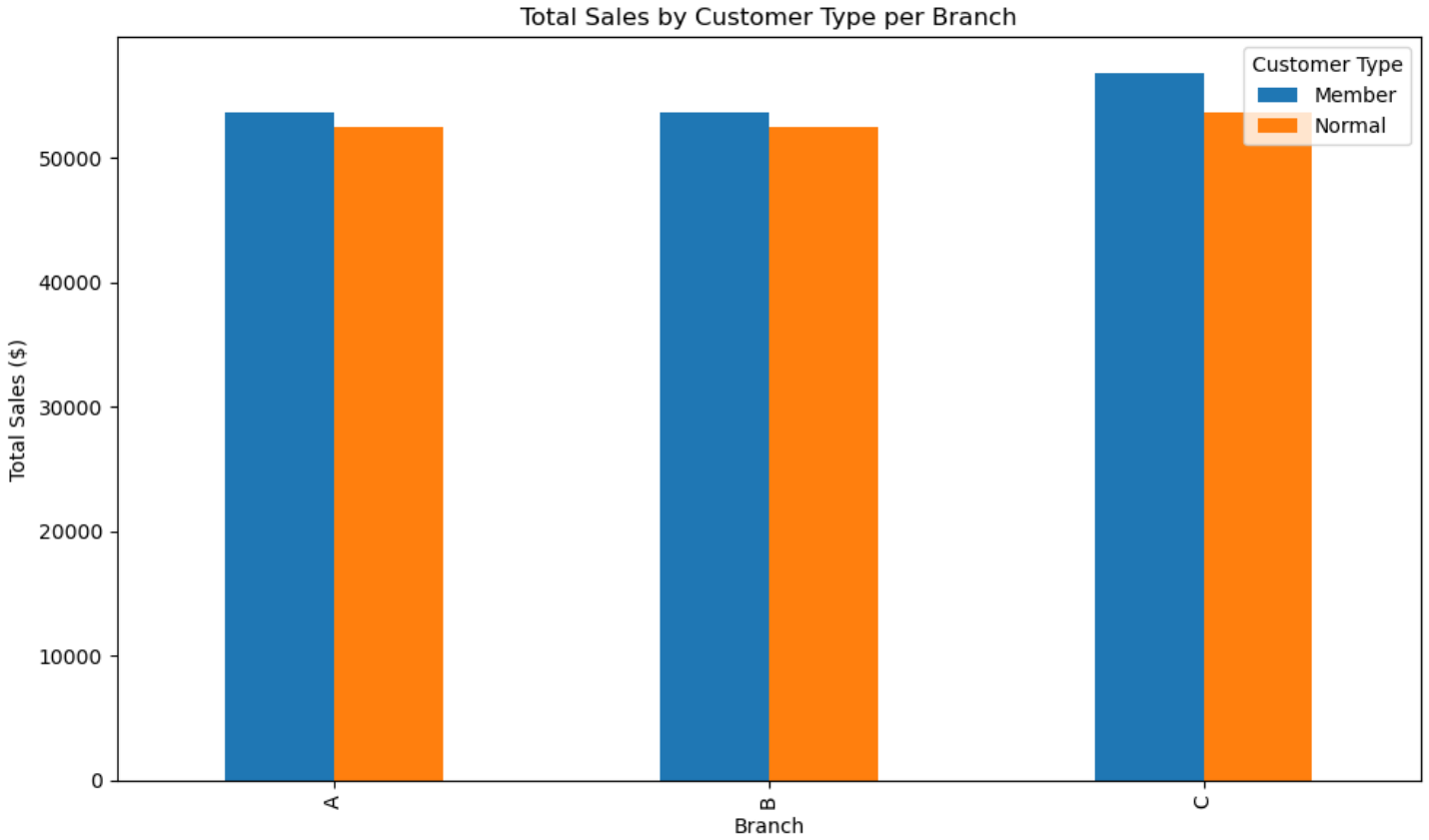
```
customer_type_sales_per_branch = (df.groupby(['Branch', 'Customer type'])['Total']
                                  .sum()
                                  .reset_index()
                                  .sort_values(by=['Branch', 'Total'], ascending=False))
customer_type_sales_per_branch.pivot_table(
    index = "Customer type",
    columns = "Branch",
    values = "Total",
).style.format("{:.2f}").background_gradient(cmap = "RdYlGn", axis = None)
```

output:

Branch	A	B	C
Customer type			
Member	53637.48	53704.69	56881.28
Normal	52562.89	52492.99	53687.42

```
customer_type_pivot = customer_type_sales_per_branch.pivot(
    index='Branch',
    columns='Customer type',
    values='Total')

customer_type_pivot.plot(kind='bar', figsize=(10,6))

plt.title('Total Sales by Customer Type per Branch')
plt.ylabel('Total Sales ($)')
plt.xlabel('Branch')
plt.legend(title='Customer Type')
plt.tight_layout()
plt.show()
```

output:



Results interpretation:

- Branch C's strong performance among members indicates successful customer retention strategies, possibly through loyalty programs or personalized experiences.
- The similar buying tendencies between Branches A and B suggest uniform performance across customer segments, though both branches could focus on increasing member engagement for higher spending.

Overall, Branch C continues to dominate in both customer types, while Branches A and B show areas for growth, particularly in boosting member engagement and improving product-specific sales.

3.3 Customer satisfaction analysis

In this section, i analyzed customer satisfaction based on the ratings provided. I first categorized customer ratings into three levels: Low, Medium, and High. Then, i explored how these satisfaction levels relate to product lines and payment methods.

3.3.1 Categorizing customer ratings

i began by categorizing the ratings into three levels:

- **Low** (ratings ≤ 4),
- **Medium** (ratings between 5 and 7),
- **High** (ratings ≥ 8).

```
df = df.assign(
    Satisfaction_Level = np.where(df["Rating"] <= 4, "Low",
```

```
                                    np.where((df["Rating"] >= 5) & (df["Rating"] <= 7),
"Medium", "High"))
).rename(columns = {"Satisfaction_Level": "Satisfaction Level"})

df[['Rating', 'Satisfaction Level']]
```

3.3.2 Product line vs. Satisfaction level

Next, i explored how customer satisfaction is distributed across different product lines. This allows us to identify which product lines lead to higher customer satisfaction.

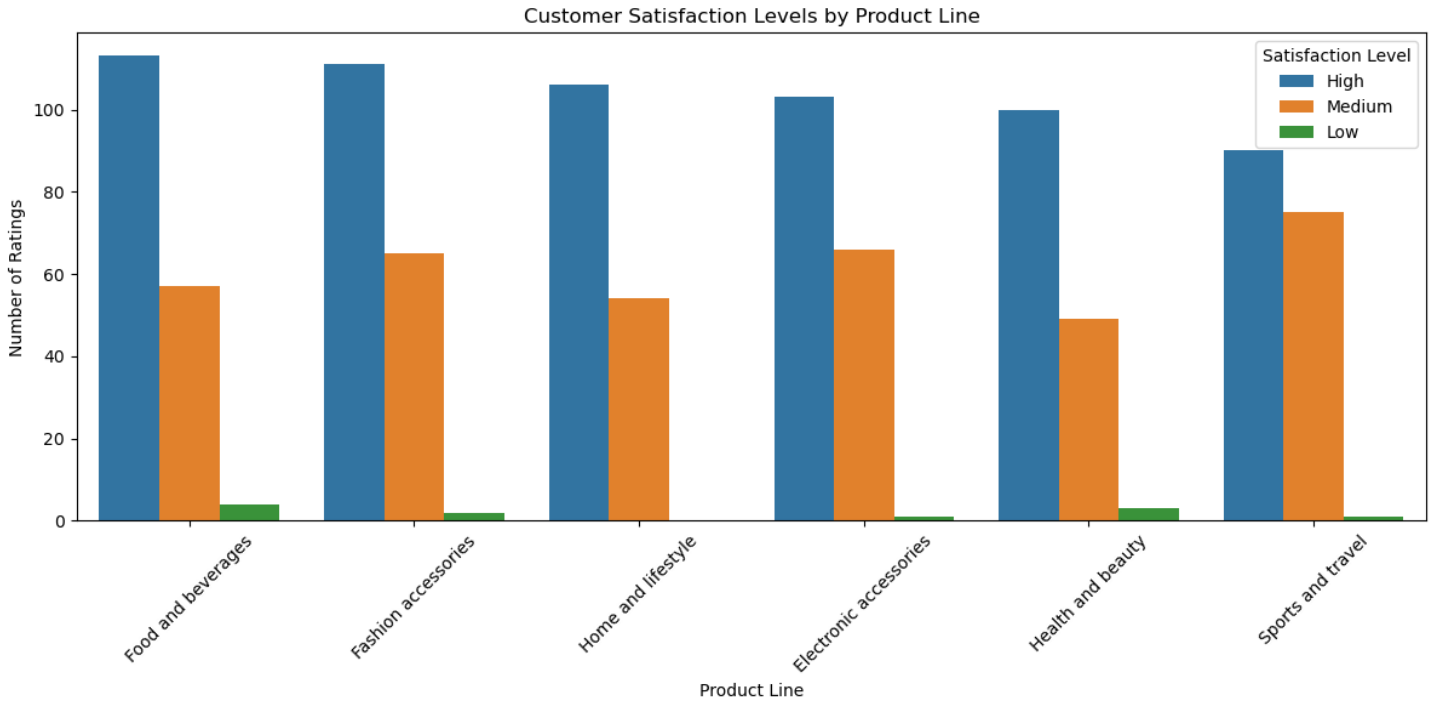
```
product_line_satisfaction = df.groupby(['Product line', 'Satisfaction Level'])['Invoice
ID'].count().reset_index()

product_line_satisfaction.rename(columns={'Invoice ID': 'Count'}, inplace=True)

product_line_satisfaction = product_line_satisfaction.sort_values(by='Count',
ascending=False)

product_line_satisfaction.pivot_table(
    index = "Product line",
    columns = "Satisfaction Level",
    values = "Count",
).replace(np.NAN, 0).style.format("{:.0f}").background_gradient(cmap="RdYlGn", axis=None)
```

output:

Satisfaction Level	High	Low	Medium
Product line			
Electronic accessories	103	1	66
Fashion accessories	111	2	65
Food and beverages	113	4	57
Health and beauty	100	3	49
Home and lifestyle	106	0	54
Sports and travel	90	1	75

```
plt.figure(figsize=(12,6))
sns.barplot(x='Product line', y='Count', hue='Satisfaction Level',
data=product_line_satisfaction)
plt.title('Customer Satisfaction Levels by Product Line')
plt.ylabel('Number of Ratings')
plt.xlabel('Product Line')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

output:



Results interpretation:

- Food and Beverages and Fashion Accessories lead in customer satisfaction, showing strong overall performance.
- Home and Lifestyle stands out for having no low satisfaction ratings, indicating consistently high performance.
- Electronic Accessories and Health and Beauty have a larger portion of medium ratings, which may point to opportunities for enhancing the customer experience.
- Sports and Travel has the most medium ratings, suggesting that while customers are not dissatisfied, their experience could be improved to elevate satisfaction levels.

3.3.3 Payment method vs. Satisfaction level

To gain further insights, i analyzed how satisfaction levels differ across various payment methods. This can help businesses understand if certain payment methods are associated with higher or lower customer satisfaction.

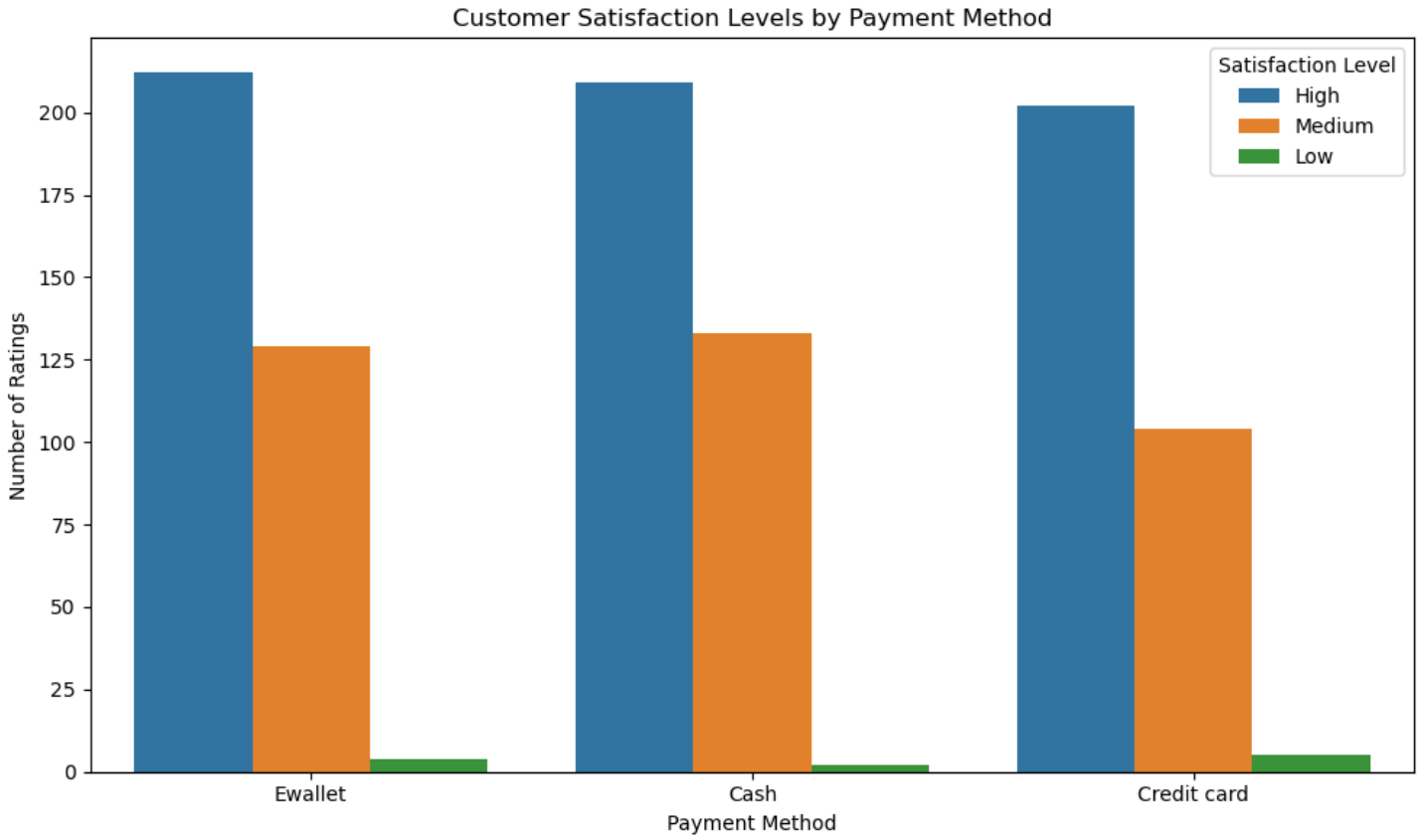
```
payment_satisfaction = df.groupby(['Payment', 'Satisfaction Level'])['Invoice
ID'].count().reset_index()

payment_satisfaction.rename(columns={'Invoice ID': 'Count'}, inplace=True)
payment_satisfaction = payment_satisfaction.sort_values(by='Count', ascending=False)

payment_satisfaction.pivot_table(
    index = "Payment",
    columns = "Satisfaction Level",
    values = "Count",
).style.format("{:.0f}").background_gradient(cmap="RdYlGn", axis=None)
```

output:

Satisfaction Level	High	Low	Medium
Payment			
Cash	209	2	133
Credit card	202	5	104
Ewallet	212	4	129

```
plt.figure(figsize=(10,6))
sns.barplot(x='Payment', y='Count', hue='Satisfaction Level', data=payment_satisfaction)
plt.title('Customer Satisfaction Levels by Payment Method')
plt.ylabel('Number of Ratings')
plt.xlabel('Payment Method')
plt.tight_layout()
plt.show()
```

output:



Results interpretation:

- E-Wallet payments lead in high satisfaction ratings, making them the most favored payment method overall. The low dissatisfaction levels further solidify e-wallets as a top-performing option.
- Cash payments show high satisfaction as well but have the most medium satisfaction ratings. This suggests that while customers are generally satisfied, there is room for improvement to convert these neutral experiences into highly positive ones.
- Credit Card payments also perform well but show a slightly higher rate of dissatisfaction compared to cash and e-wallets. This could be an area to address, possibly by examining factors such as transaction speed or fees associated with credit card payments.

3.4 Product insights

In this section, i explored various aspects of product performance, including sales volume, gross income, and customer-type comparisons. These insights help us identify which product lines generate the most sales and income and how different customer types interact with these product lines.

3.4.1 Sales volume analysis by product line

I started by analyzing the total sales volume for each product line. This helps us understand which products are sold the most.

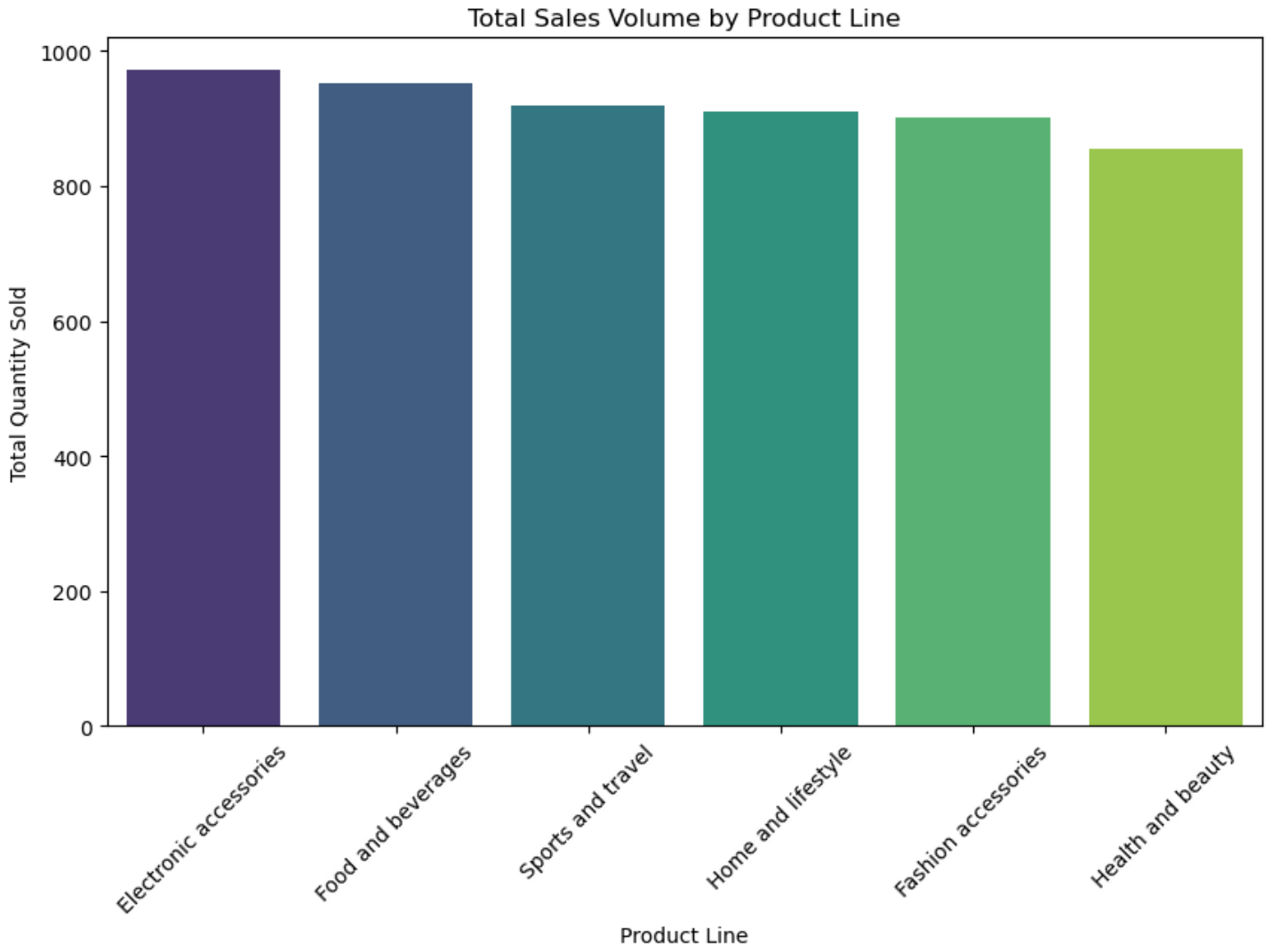
```
product_sales_volume = df.groupby('Product
line')['Quantity'].sum().sort_values(ascending=False).reset_index()

product_sales_volume
```

output:

	Product line	Quantity
0	Electronic accessories	971
1	Food and beverages	952
2	Sports and travel	920
3	Home and lifestyle	911
4	Fashion accessories	902
5	Health and beauty	854

```
plt.figure(figsize=(10,6))
sns.barplot(data = product_sales_volume, x="Product line", y="Quantity", palette='viridis')
plt.title('Total Sales Volume by Product Line')
plt.xlabel('Product Line')
plt.ylabel('Total Quantity Sold')
plt.xticks(rotation=45)
plt.show()
```

output:



Results interpretation:

- Electronic Accessories and Food and Beverages lead in total sales volume, indicating strong demand for these products across the supermarket branches.
- Sports and Travel, Home and Lifestyle, and Fashion Accessories have similar sales volumes, suggesting steady but moderate interest.
- Health and Beauty rank the lowest in sales volume, potentially indicating weaker demand or room for promotional efforts to increase visibility and sales.

3.4.2 Gross income contribution by product line

Next, i analyzed which product lines contribute the most to gross income. This highlights which products are driving revenue.

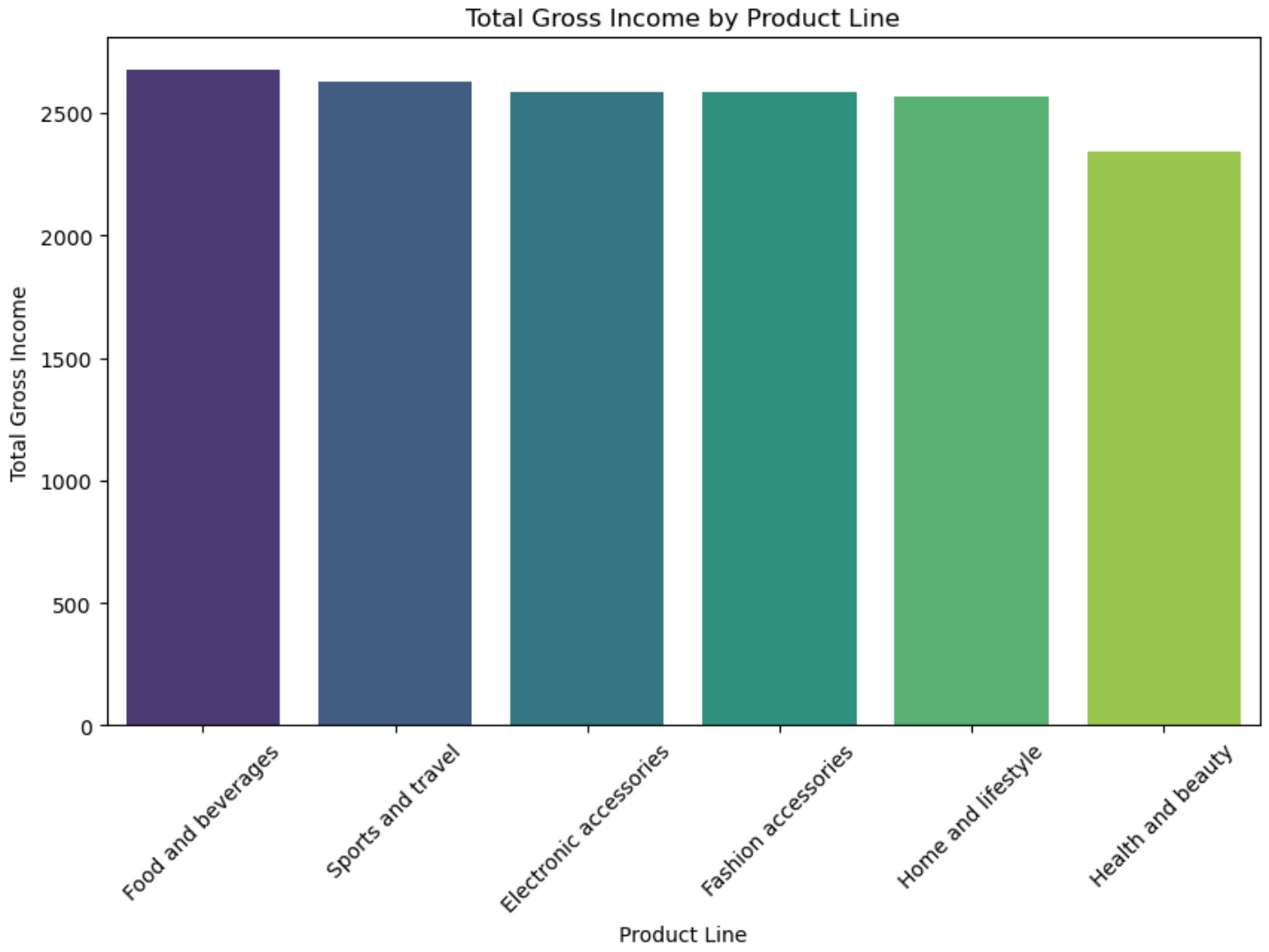
```
product_gross_income = df.groupby('Product line')['gross
income'].sum().sort_values(ascending=False).reset_index()

product_gross_income
```

output:

	Product line	gross income
0	Food and beverages	2673.5640
1	Sports and travel	2624.8965
2	Electronic accessories	2587.5015
3	Fashion accessories	2585.9950
4	Home and lifestyle	2564.8530
5	Health and beauty	2342.5590

```
plt.figure(figsize=(10,6))
sns.barplot(data = product_gross_income, x="Product line", y="gross income",
palette='virdis')
plt.title('Total Gross Income by Product Line')
plt.xlabel('Product Line')
plt.ylabel('Total Gross Income')
plt.xticks(rotation=45)
plt.show()
```

output:



Results interpretation:

- Food and Beverages and Sports and Travel lead in gross income, indicating that these product lines not only sell well but also generate significant revenue, likely due to higher profit margins.
- Electronic Accessories, Fashion Accessories, and Home and Lifestyle have similar gross income, indicating steady profitability but not at the top of the chart.
- Health and Beauty have the lowest gross income, mirroring its low sales volume. This may indicate either a lower profit margin or less consumer interest, making it a category that requires more attention.

3.4.3 Average gross income per sale by product line

Here, i analyzed the average gross income per sale, which can indicate the profitability of each product line on a per-transaction basis.

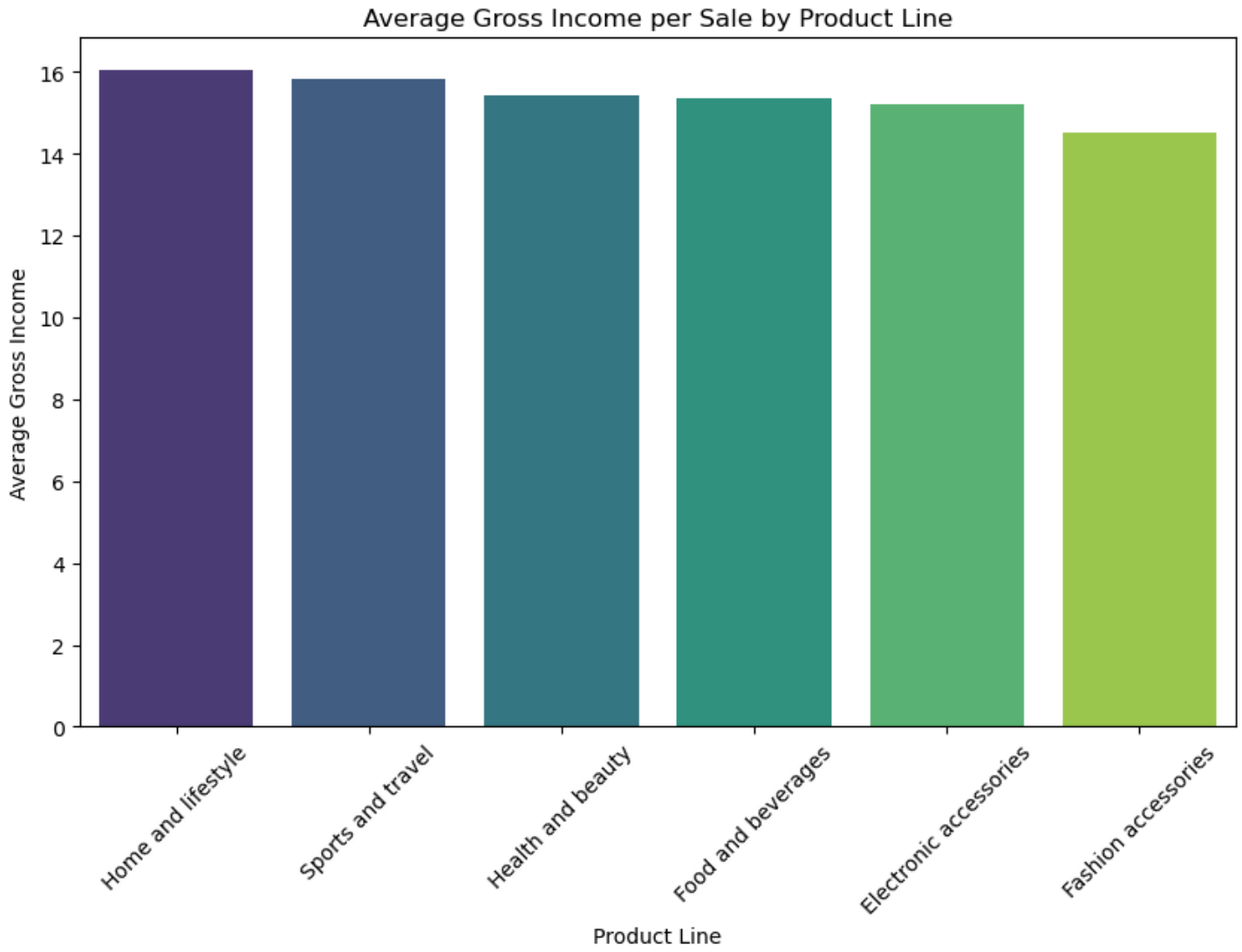
```
avg_gross_income_per_sale = df.groupby('Product line')['gross
income'].mean().sort_values(ascending=False).reset_index()

avg_gross_income_per_sale
```

output:

	Product line	gross income
0	Home and lifestyle	16.030331
1	Sports and travel	15.812630
2	Health and beauty	15.411572
3	Food and beverages	15.365310
4	Electronic accessories	15.220597
5	Fashion accessories	14.528062

```
plt.figure(figsize=(10,6))
sns.barplot(data = avg_gross_income_per_sale, x="Product line", y="gross income",
palette='viridis')
plt.title('Average Gross Income per Sale by Product Line')
plt.xlabel('Product Line')
plt.ylabel('Average Gross Income')
plt.xticks(rotation=45)
plt.show()
```

output:



Results interpretation:

- Home and Lifestyle leads in average gross income per sale, indicating high profitability per transaction, which could be attributed to higher-priced items.
- Sports and Travel follows closely, suggesting that although it doesn't lead in total volume, it offers good margins on each sale.
- Health and Beauty, Food and Beverages, and Electronic Accessories exhibit similar average gross income, showing that while these categories have steady sales, their per-sale profitability is moderate.
- Fashion Accessories has the lowest average gross income, indicating lower profitability per transaction, possibly due to lower-priced items or higher competition.

3.4.4 Correlation between sales volume and gross income

In this section, i explored whether there is a relationship between the sales volume and gross income for each product line.

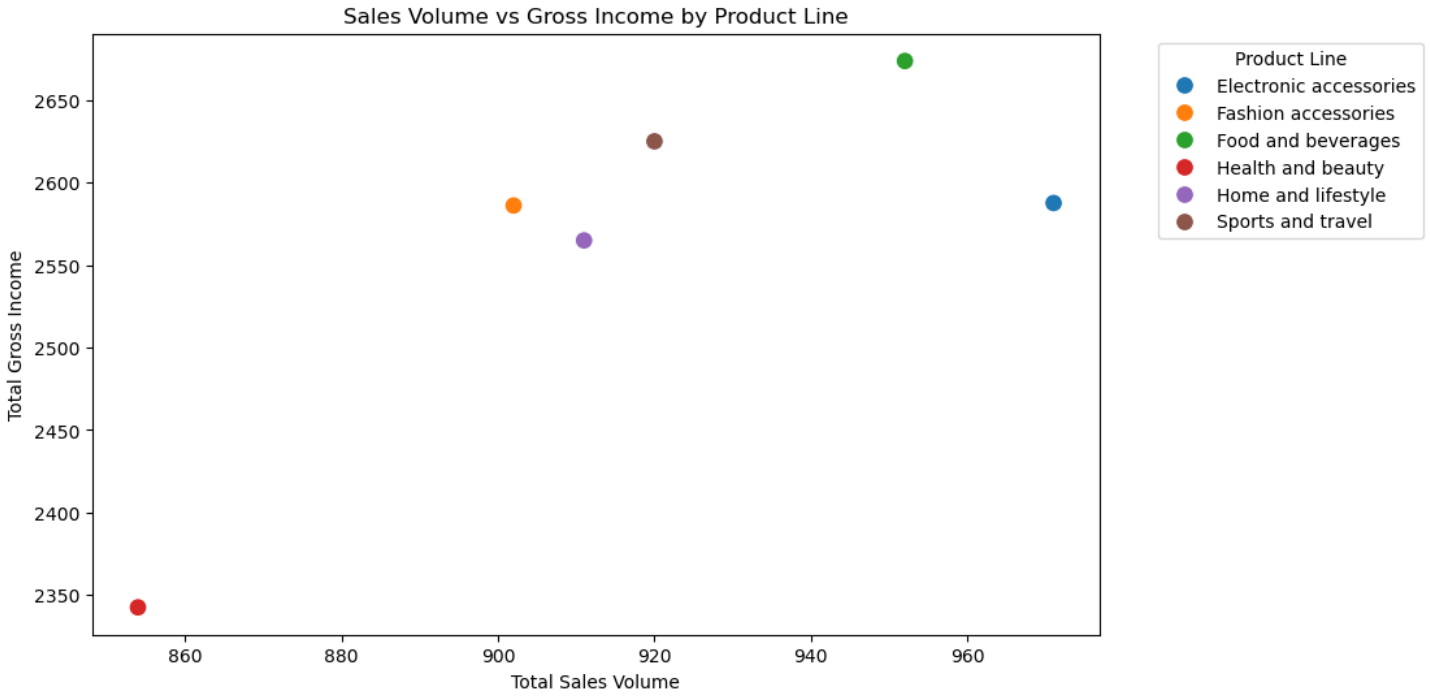
```
product_analysis = df.groupby('Product line').agg({'Quantity': 'sum', 'gross income':
'sum'}).reset_index()

product_analysis
```

output:

	Product line	Quantity	gross income
0	Electronic accessories	971	2587.5015
1	Fashion accessories	902	2585.9950
2	Food and beverages	952	2673.5640
3	Health and beauty	854	2342.5590
4	Home and lifestyle	911	2564.8530
5	Sports and travel	920	2624.8965

```
plt.figure(figsize=(10,6))
sns.scatterplot(data=product_analysis, x='Quantity', y='gross income', hue="Product line", s
= 100)
plt.title('Sales Volume vs Gross Income by Product Line')
plt.xlabel('Total Sales Volume')
plt.ylabel('Total Gross Income')
plt.legend(title='Product Line', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

output:



Results interpretation:

- Food and Beverages stands out as the top performer in both sales volume and gross income, with 952 units sold and a gross income of 2673.56, highlighting it as a major contributor to profitability.
- Electronic Accessories sold the highest number of units (971) and generated a gross income of 2587.50, but its gross income is slightly lower than Food and Beverages, indicating that the profit margin per unit is lower.
- Health and Beauty has the lowest gross income (2342.56) and the smallest sales volume (854). This product line might need special attention to either increase sales or improve its profitability.
- Sports and Travel generated a relatively high gross income (2624.90) with a sales volume of 920, showing good profitability despite selling fewer units than Electronic Accessories and Food and Beverages.

- Home and Lifestyle shows balanced performance, with 911 units sold and a gross income of 2564.85, indicating consistent performance across branches but not the highest in either category.
- Fashion Accessories sold 902 units, generating a gross income of 2586.00. It performs similarly to Home and Lifestyle, suggesting stable sales and profitability.

Correlation Insights:

- There is a positive correlation between sales volume and gross income, as expected. Products that sell more units tend to generate higher gross income.
- However, the relationship is not strictly linear. For example, Electronic Accessories sold the most units but doesn't have the highest gross income, which suggests that some product lines have higher margins (e.g., Food and Beverages and Sports and Travel).
- Health and Beauty is an outlier with relatively low sales and gross income, signaling a potential area for improvement.

3.4.5 Product line comparison by customer type

Finally, i compared product performance across customer types to understand which product lines are preferred by different customer segments.

```
product_customer_analysis = df.groupby(['Product line', 'Customer type']).agg({'Quantity': 'sum', 'gross income': 'sum'}).unstack()

product_customer_analysis
```

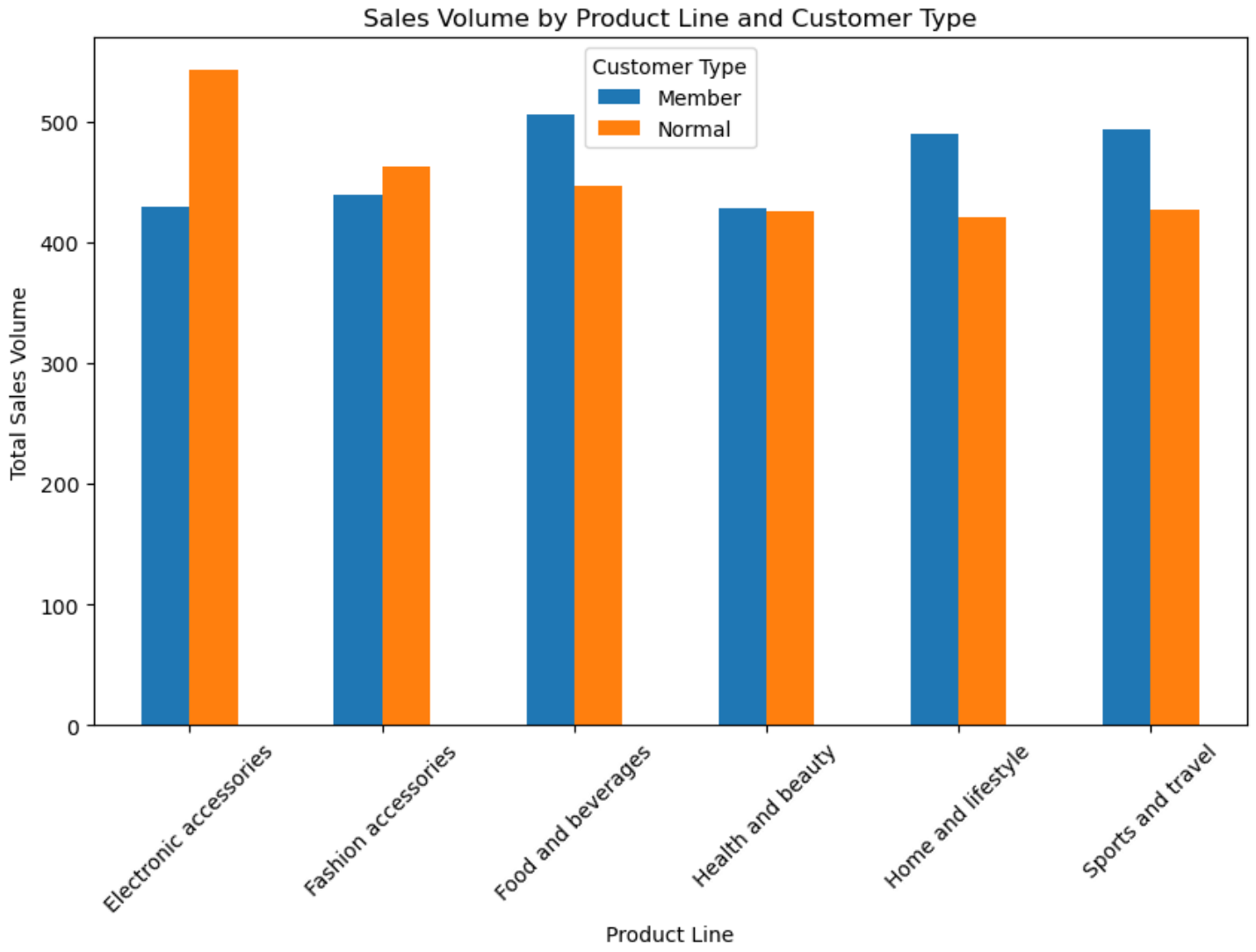
output:

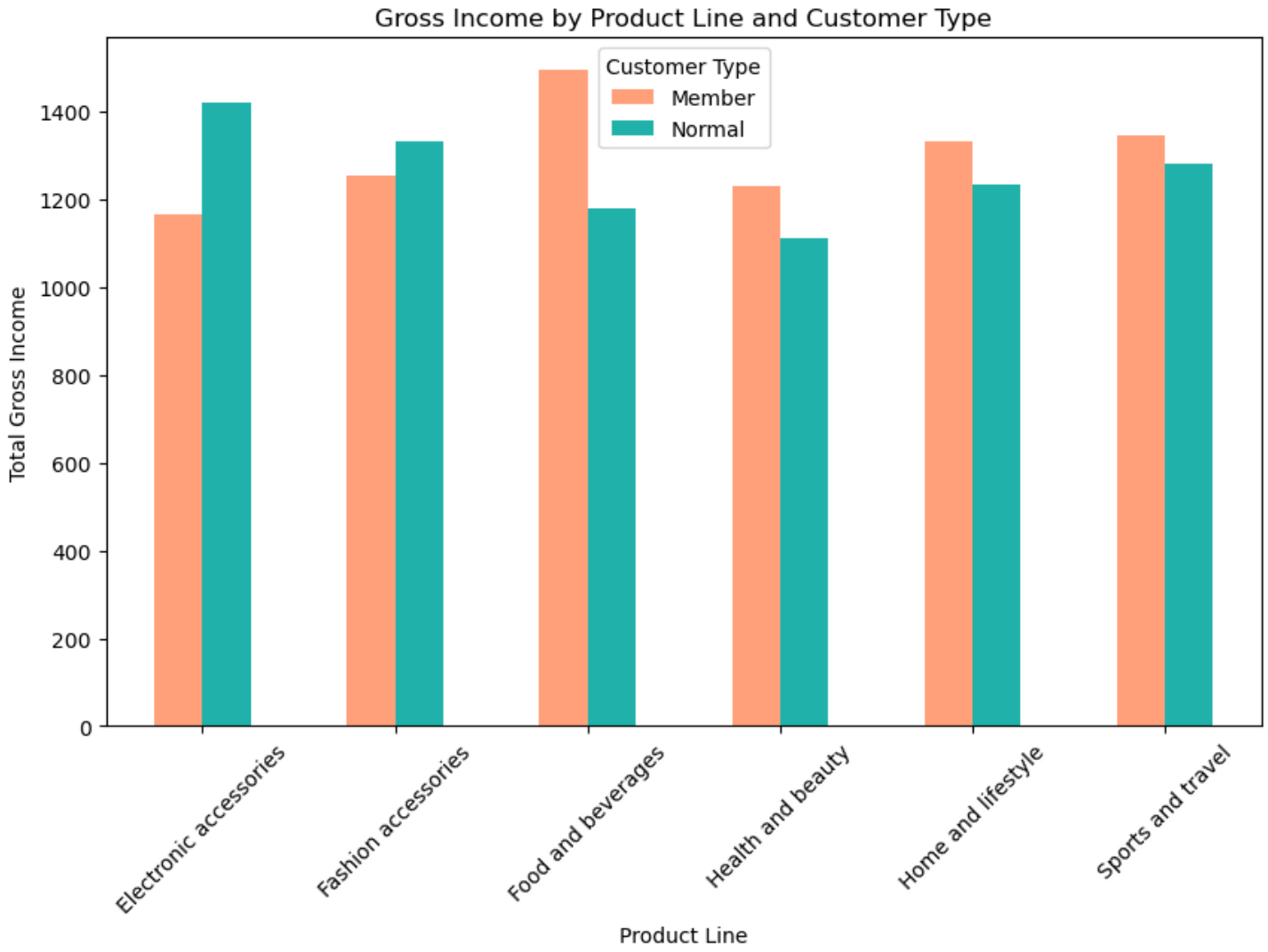
	Quantity		gross income	
Customer type	**Member**	**Normal**	**Member**	**Normal**
Product line				
Electronic accessories	429	542	1166.5950	1420.9065
Fashion accessories	439	463	1253.5220	1332.4730
Food and beverages	506	446	1493.2200	1180.3440
Health and beauty	428	426	1230.0495	1112.5095
Home and lifestyle	490	421	1332.2870	1232.5660
Sports and travel	493	427	1344.4905	1280.4060

```
product_customer_analysis['Quantity'].plot(kind='bar', figsize=(10,6))
plt.title('Sales Volume by Product Line and Customer Type')
plt.xlabel('Product Line')
plt.ylabel('Total Sales Volume')
plt.xticks(rotation=45)
plt.legend(title='Customer Type')
plt.show()

product_customer_analysis['gross income'].plot(kind='bar', figsize=(10,6), color=['#FFA07A', '#20B2AA'])
```

```
plt.title('Gross Income by Product Line and Customer Type')
plt.xlabel('Product Line')
plt.ylabel('Total Gross Income')
plt.xticks(rotation=45)
plt.legend(title='Customer Type')
plt.show()
```

output:





Results interpretation:

- Electronic Accessories: Normal customers purchase more (542 units) than members (429 units), contributing significantly to the total sales volume. Similarly, gross income is higher from normal customers ($1420.91) than members ($1166.60), suggesting that electronic accessories attract a higher non-member customer base.
- Fashion Accessories: Sales are fairly balanced between members (439 units) and normal customers (463 units). Gross income is similar for both groups, with normal customers generating $1332.47 and members $1253.52. This indicates that both customer types show comparable interest in fashion accessories, with slightly better sales among normal customers.
- Food and Beverages: Members contribute more in terms of sales volume (506 units) and gross income ($1493.22), outperforming normal customers who generate lower sales volume (446 units) and gross income ($1180.34). This suggests that members have a stronger preference for food and beverages, which could be due to loyalty programs or membership benefits.
- Health and Beauty: Sales volume is nearly identical for members (428 units) and normal customers (426 units). However, gross income from members ($1230.05) is higher than from normal customers ($1112.51), indicating members tend to spend more per purchase in this category.
- Home and Lifestyle: Members generate more sales volume (490 units) and gross income ($1332.29) compared to normal customers (421 units and $1232.57). This suggests a stronger preference for home and lifestyle products among members, potentially due to repeat purchases or higher spending capacity.
- Sports and Travel: Similar to other categories, members purchase slightly more (493 units) than normal customers (427 units), and the gross income reflects this pattern ($1344.49 vs. $1280.41). Although the difference is not

drastic, members contribute more to both sales and revenue in this category, indicating that they are key consumers in sports and travel.

Key Insights:

- Members tend to contribute more gross income across most product lines, even when the sales volume is similar or lower than that of normal customers. This could be due to higher spending per transaction.
- Normal customers are strong contributors in Electronic Accessories and Fashion Accessories, suggesting targeted marketing efforts could boost sales from this group in other product lines.
- The preference for Food and Beverages is much stronger among members, making this a potentially lucrative category for driving membership growth.
- Home and Lifestyle and Sports and Travel also perform well among members, reinforcing the idea that membership brings higher engagement in these categories.