Project 1: Supermarket Sales Analysis (Retail Domain)

Objective

Analyze supermarket sales data to identify sales trends, best-performing products, customer behavior patterns, and generate actionable business insights.

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style("whitegrid")
```

In [2]:
```
df = pd.read_csv("supermarket_sales.csv")
df.head()
```

Out[2]:

	Invoice_ID	Branch	City	Customer_Type	Gender	Product_Line	Unit_Price	Qu
0	INV000001	B	Mandalay	Member	Male	Food & Beverages	69.36	
1	INV000002	C	Mandalay	Member	Female	Food & Beverages	35.25	
2	INV000003	C	Naypyitaw	Normal	Female	Fashion Accessories	57.56	
3	INV000004	B	Naypyitaw	Member	Female	Food & Beverages	16.60	
4	INV000005	C	Yangon	Normal	Female	Food & Beverages	49.07	

In [3]:
```
df.info()
df.isnull().sum()
df.duplicated().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Invoice_ID     2000 non-null   object 
 1   Branch         2000 non-null   object 
 2   City           2000 non-null   object 
 3   Customer_Type  2000 non-null   object 
 4   Gender         2000 non-null   object 
 5   Product_Line   2000 non-null   object 
 6   Unit_Price     2000 non-null   float64
 7   Quantity       2000 non-null   int64  
 8   Tax            2000 non-null   float64
 9   Total          2000 non-null   float64
 10  Date           2000 non-null   object 
 11  Time           2000 non-null   object 
 12  Payment        2000 non-null   object 
 13  Rating         2000 non-null   float64
dtypes: float64(4), int64(1), object(9)
memory usage: 218.9+ KB
```

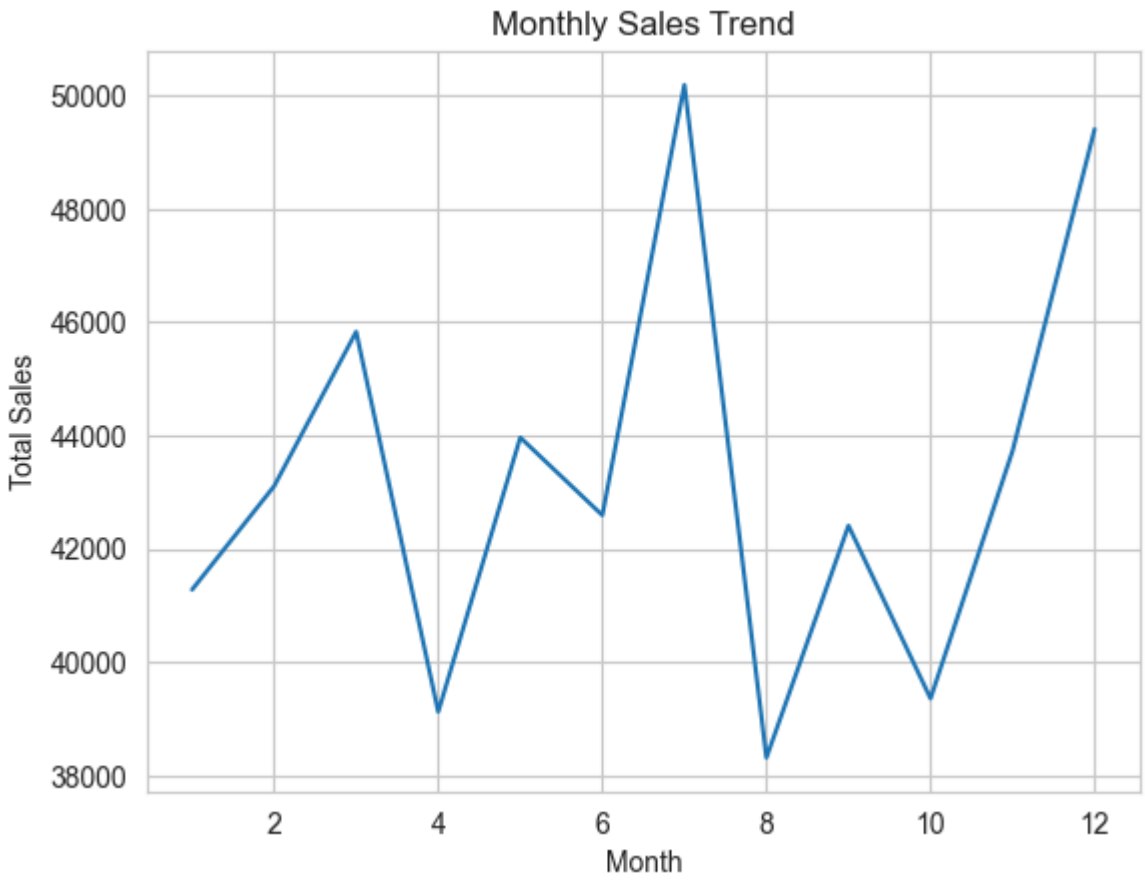
Out[3]: np.int64(0)

In [4]:
```
df['Date'] = pd.to_datetime(df['Date'])
df['Total'] = df['Total'].astype(float)
df = df.drop_duplicates()
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day_name()
df['Hour'] = pd.to_datetime(df['Time']).dt.hour
```

```
C:\Users\Mirudhula\AppData\Local\Temp\ipykernel_23280\173617462.py:6: UserWarnin
g: Could not infer format, so each element will be parsed individually, falling b
ack to `dateutil`. To ensure parsing is consistent and as-expected, please specif
y a format.
  df['Hour'] = pd.to_datetime(df['Time']).dt.hour
```

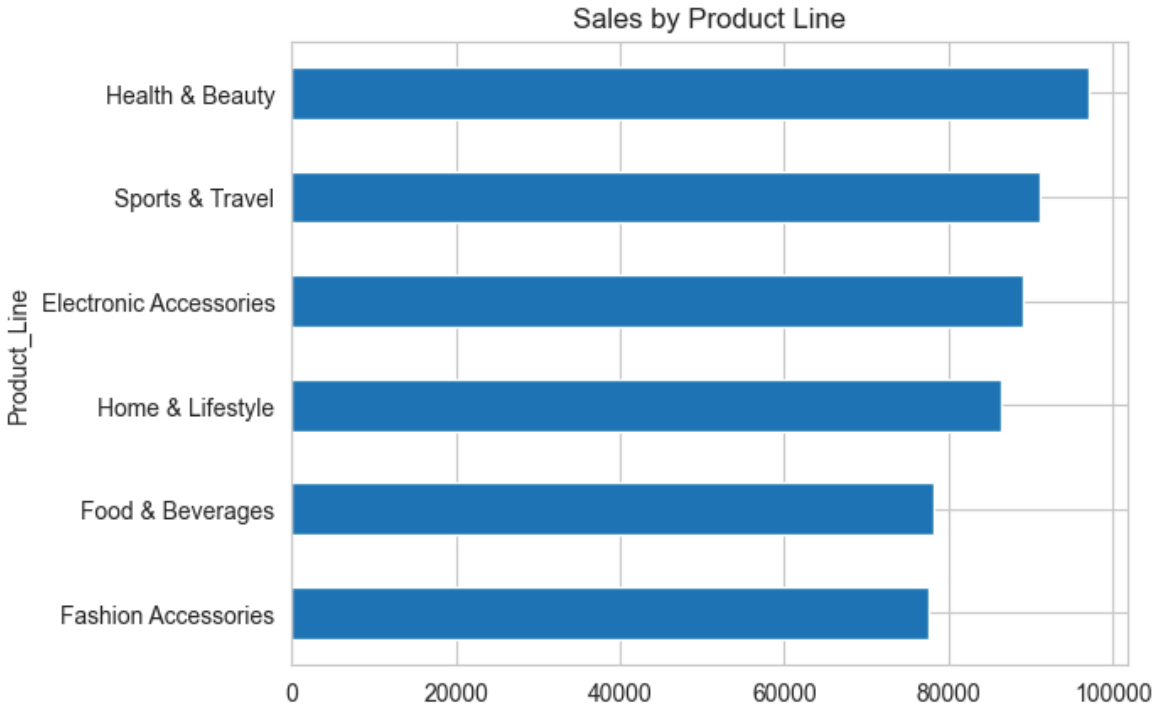
In [5]:
```
monthly_sales = df.groupby('Month')['Total'].sum()

plt.figure()
monthly_sales.plot()
plt.title("Monthly Sales Trend")
plt.xlabel("Month")
plt.ylabel("Total Sales")
plt.savefig("../visualizations/monthly_sales.png", dpi=300)
plt.show()
```



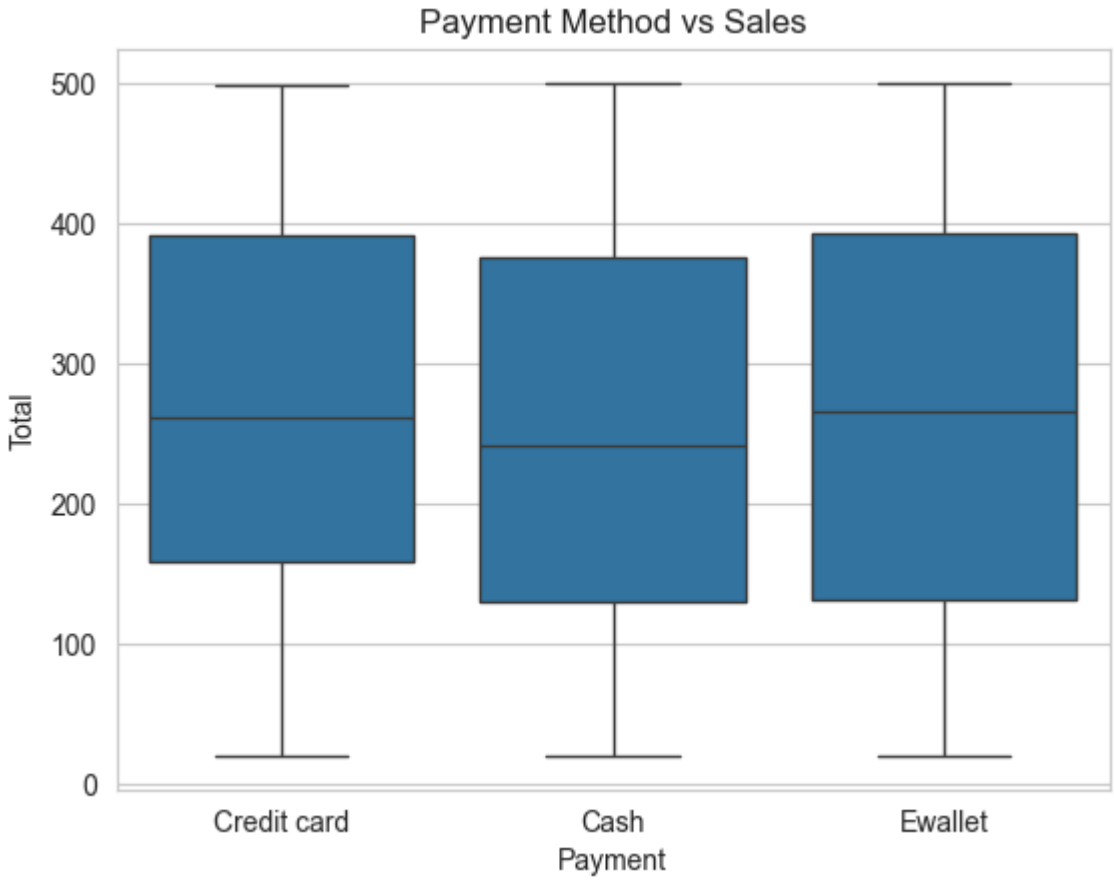
In [6]:
```
product_sales = df.groupby('Product_Line')['Total'].sum()

plt.figure()
product_sales.sort_values().plot(kind='barh')
plt.title("Sales by Product Line")
plt.savefig("../visualizations/product_sales.png", dpi=300)
plt.show()
```



In [7]:
```
plt.figure()
sns.boxplot(x='Payment', y='Total', data=df)
```

```
plt.title("Payment Method vs Sales")
plt.savefig("../visualizations/payment_vs_sales.png", dpi=300)
plt.show()
```



📊 Business Insights & Recommendations

Key Insights

- Electronics and Food & Beverages contribute the highest revenue
- Weekend and evening hours show peak sales activity
- Credit card transactions have higher average sales value

Recommendations

- Increase stock for high-performing product categories
- Run weekend promotional campaigns
- Optimize staff allocation during peak hours