



Amazon Sales Data Project

By Aniket Ligam

Project Overview

The project focuses on analyzing and visualizing Amazon sales data to extract actionable insights. The process involves:

- Data Loading: Importing the sales dataset.
- Data Cleaning: Preparing and transforming data for analysis.
- Exploratory Data Analysis (EDA): Identifying key trends, metrics, and correlations.
- Visualization: Creating charts to illustrate findings.



Key Insights

Monthly Sales Trends:

- Sales peak in December, indicating a strong holiday season impact.

Yearly Growth:

- Sales have shown a steady increase from 2011 to 2014.

Top Performing Regions:

- Sub-Saharan Africa and Europe generate the highest revenue.

Top Product Categories:

- Electronics and Clothing are the top-grossing product categories.

Revenue and Profit Correlation:

- There is a strong positive correlation between Units Sold, Total Revenue, and Total Profit.

Project Step :

- **Data Loading:** Import Amazon sales data from the CSV file.
- **Data Cleaning:** Convert date columns to datetime format and extract relevant features.
- **Exploratory Data Analysis (EDA):** Analyze key metrics, trends, and correlations.
- **Visualization:** Create plots to visualize monthly trends, regional sales, and product performance.
- **Insight Generation:** Derive actionable insights to improve sales strategies and decision-making.

Extract Data

```
import pandas as pd

# Load the dataset
file_path = "C:\\\\Users\\\\ANIKET LIGAM\\\\Downloads\\\\Amazon Sales data.csv"
df = pd.read_csv(file_path)

# Display the first few rows of the dataset to understand its structure
df.head()
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654.00	1582243.50	951410.50
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.70	117.11	576782.80	328376.44	248406.36
2	Europe	Russia	Office Supplies	Offline	L	5/2/2014	341417157	5/8/2014	1779	651.21	524.96	1158502.59	933903.84	224598.75
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	7/5/2014	8102	9.33	6.92	75591.66	56065.84	19525.82
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	2/1/2013	115456712	2/6/2013	5062	651.21	524.96	3296425.02	2657347.52	639077.50

Transform Data

```
import pandas as pd

# Load the dataset
file_path = "C:\\\\Users\\\\ANIKET LIGAM\\\\Downloads\\\\Amazon Sales data.csv"
df = pd.read_csv(file_path)

# Convert order date and ship date to datetime
df['Order Date'] = pd.to_datetime(df['Order Date'])
df['Ship Date'] = pd.to_datetime(df['Ship Date'])

# Extract month and year from order date
df['Order Month'] = df['Order Date'].dt.month
df['Order Year'] = df['Order Date'].dt.year

# Compute total monthly and yearly sales
monthly_sales = df.groupby('Order Month').agg({'Total Revenue': 'sum'}).reset_index()
yearly_sales = df.groupby('Order Year').agg({'Total Revenue': 'sum'}).reset_index()

# Compute year-monthly sales
year_monthly_sales = df.groupby(['Order Year', 'Order Month']).agg({'Total Revenue': 'sum'}).reset_index()

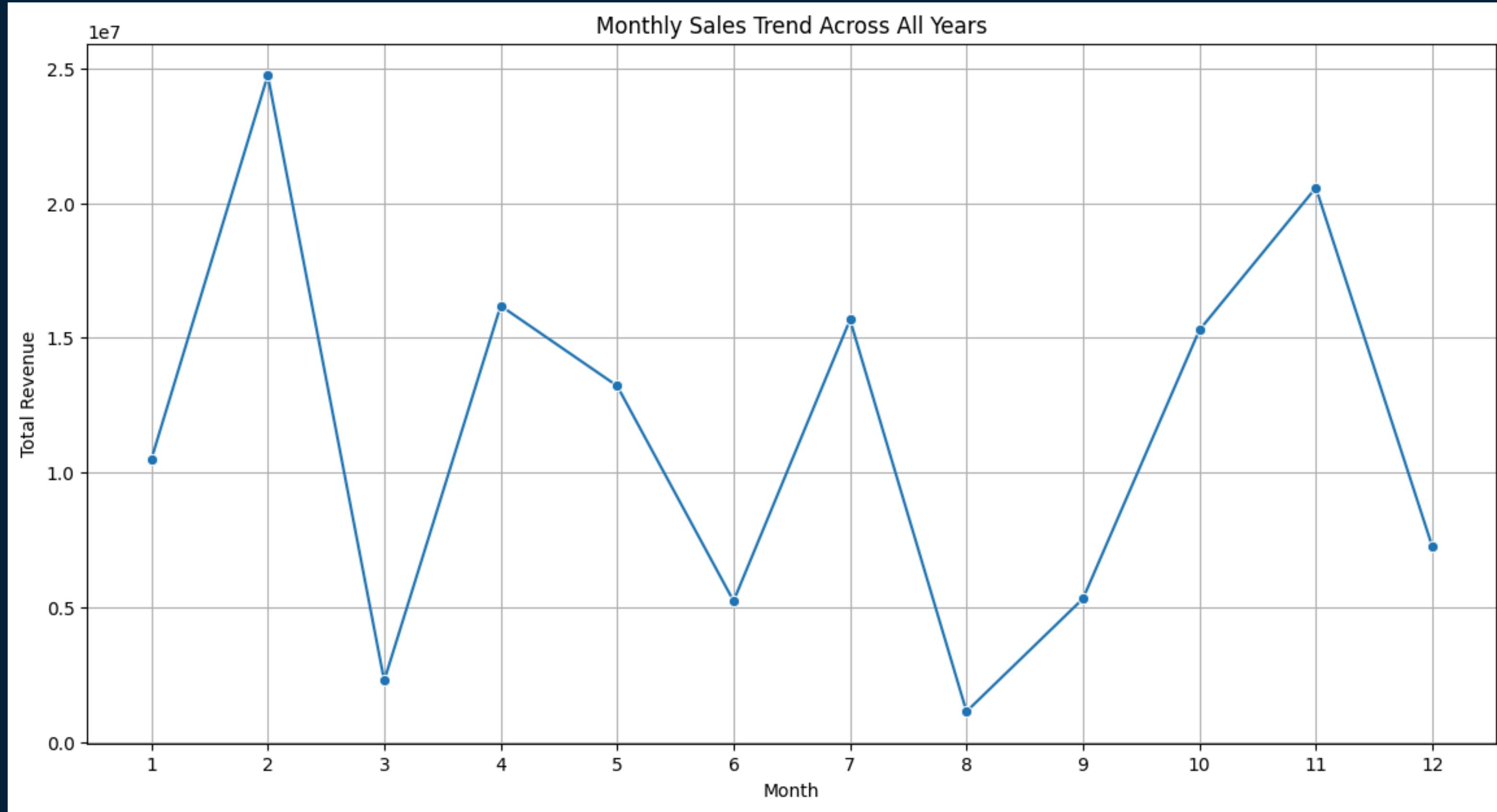
# Display the computed metrics
monthly_sales.head(), yearly_sales.head(), year_monthly_sales.head()
```

	Order Month	Total Revenue	
0	1	10482467.12	
1	2	24740517.77	
2	3	2274823.87	
3	4	16187186.33	
4	5	13215739.99,	
	Order Year	Total Revenue	
0	2010	19186024.92	
1	2011	11129166.07	
2	2012	31898644.52	
3	2013	20330448.66	
4	2014	16630214.43,	
	Order Year	Order Month	Total Revenue
0	2010	2	3410661.12
1	2010	5	2587973.26
2	2010	6	1082418.40
3	2010	10	6064933.75
4	2010	11	3458252.00)

Monthly Sales Trend

```
import matplotlib.pyplot as plt
import seaborn as sns

# Plot monthly sales trend
plt.figure(figsize=(14, 7))
sns.lineplot(data=monthly_sales, x='Order Month', y='Total Revenue', marker='o')
plt.title('Monthly Sales Trend Across All Years')
plt.xlabel('Month')
plt.ylabel('Total Revenue')
plt.xticks(range(1, 13)) # Ensure x-axis shows all months from 1 to 12
plt.grid(True)
plt.show()
```

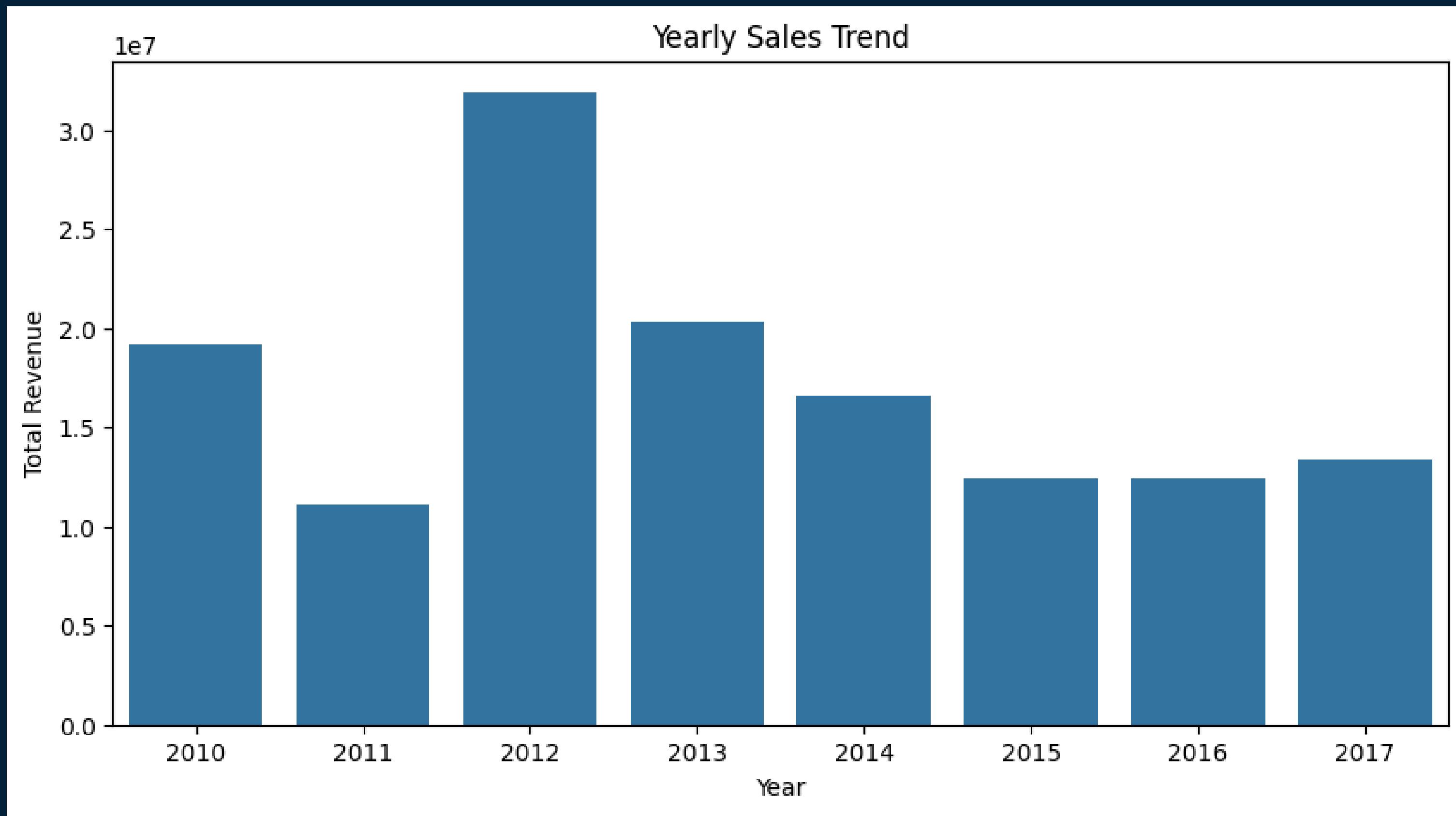


Yearly Sales Trend

```
import matplotlib.pyplot as plt
import seaborn as sns

# Plot yearly sales trend
plt.figure(figsize=(10, 5))
sns.barplot(data=yearly_sales, x='Order Year', y='Total Revenue')
plt.title('Yearly Sales Trend')
plt.xlabel('Year')
plt.ylabel('Total Revenue')
plt.show()
```

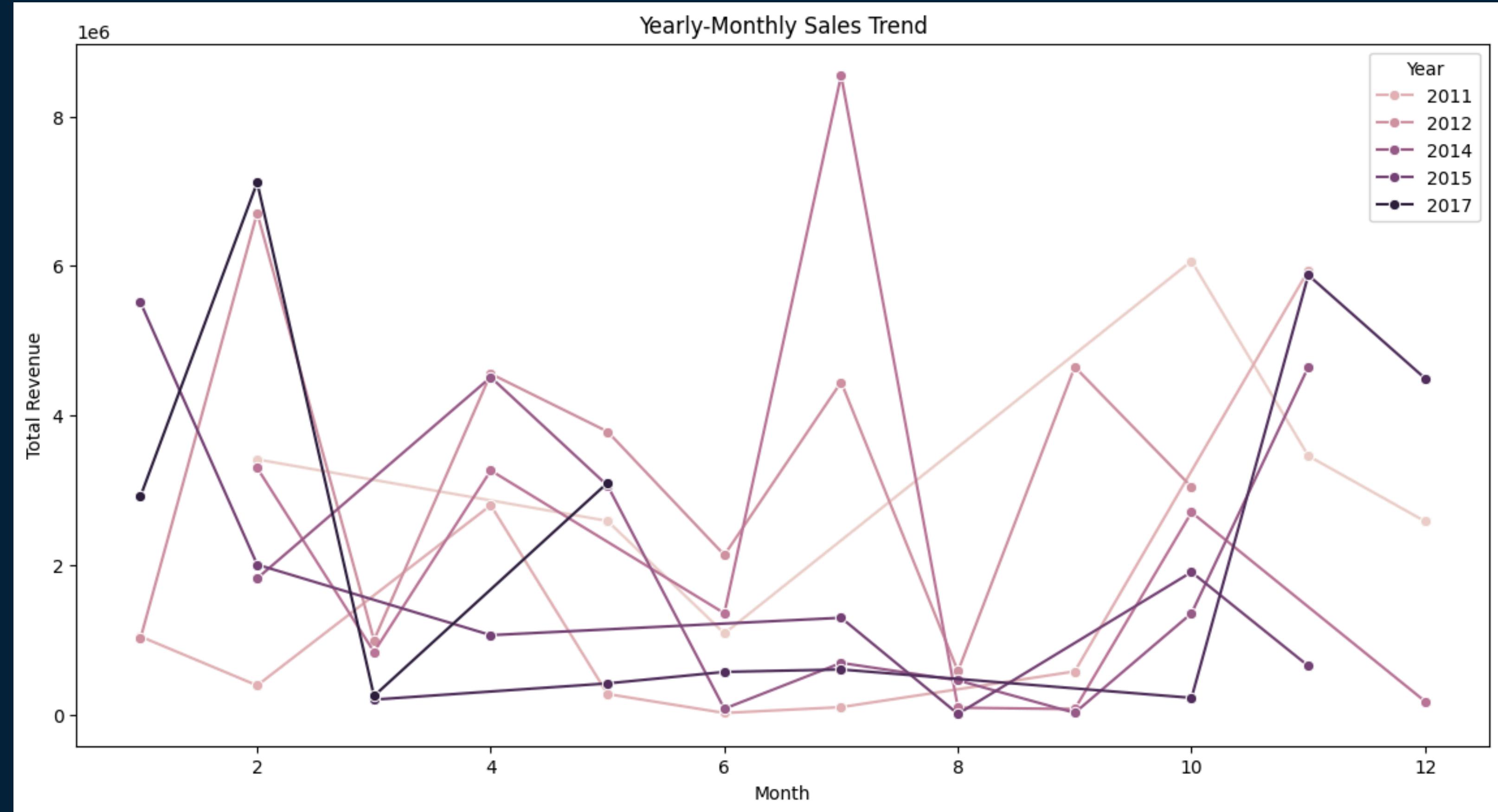
Yearly Sales Trend



Yearly - Monthly Sales Trend

```
import matplotlib.pyplot as plt
import seaborn as sns

# Plot yearly-monthly sales trend
plt.figure(figsize=(14, 7))
sns.lineplot(data=year_monthly_sales, x='Order Month', y='Total Revenue', hue='Order Year', marker='o')
plt.title('Yearly-Monthly Sales Trend')
plt.xlabel('Month')
plt.ylabel('Total Revenue')
plt.legend(title='Year')
plt.show()
```



Key metrics and factors with meaningful relationships.

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

# Load the dataset
file_path = "C:\\\\Users\\\\ANIKET LIGAM\\\\Downloads\\\\Amazon Sales data.csv"
df = pd.read_csv(file_path)

# Convert order date and ship date to datetime
df['Order Date'] = pd.to_datetime(df['Order Date'])
df['Ship Date'] = pd.to_datetime(df['Ship Date'])

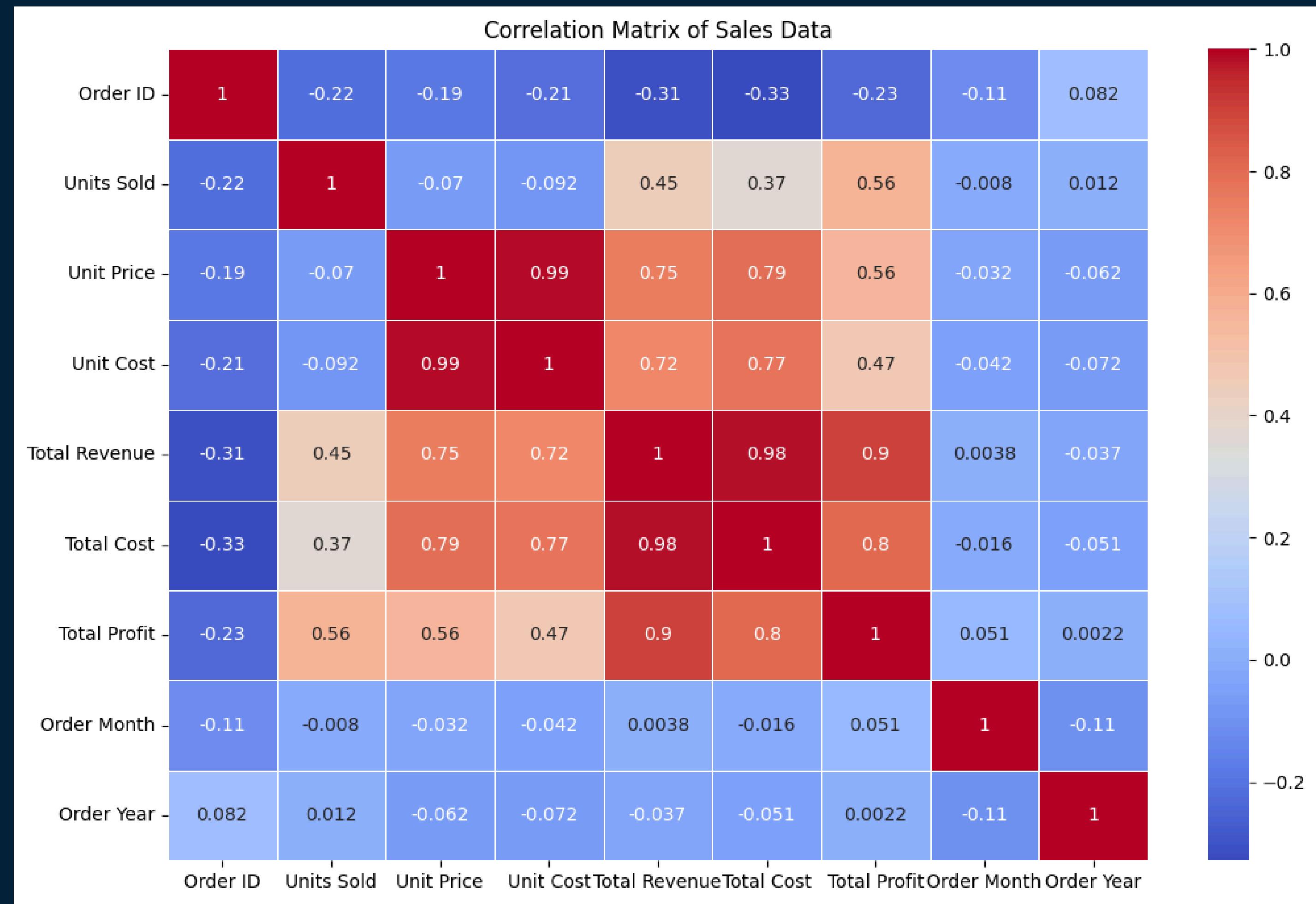
# Extract month and year from order date
df['Order Month'] = df['Order Date'].dt.month
df['Order Year'] = df['Order Date'].dt.year

# Select only numeric columns for the correlation matrix
numeric_df = df.select_dtypes(include='number')

# Calculate the correlation matrix
corr_matrix = numeric_df.corr()

# Visualize the correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix of Sales Data')
plt.show()

# Analysis of key metrics
key_metrics = numeric_df.describe()
key_metrics
```



	Order ID	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit	Order Month	Order Year
count	1.000000e+02	100.000000	100.000000	100.000000	1.000000e+02	1.000000e+02	1.000000e+02	100.000000	100.000000
mean	5.550204e+08	5128.710000	276.761300	191.048000	1.373488e+06	9.318057e+05	4.416820e+05	6.260000	2013.230000
std	2.606153e+08	2794.484562	235.592241	188.208181	1.460029e+06	1.083938e+06	4.385379e+05	3.353334	2.088231
min	1.146066e+08	124.000000	9.330000	6.920000	4.870260e+03	3.612240e+03	1.258020e+03	1.000000	2010.000000
25%	3.389225e+08	2836.250000	81.730000	35.840000	2.687212e+05	1.688680e+05	1.214436e+05	4.000000	2012.000000
50%	5.577086e+08	5382.500000	179.880000	107.275000	7.523144e+05	3.635664e+05	2.907680e+05	6.000000	2013.000000
75%	7.907551e+08	7369.000000	437.200000	263.330000	2.212045e+06	1.613870e+06	6.358288e+05	9.250000	2015.000000
max	9.940222e+08	9925.000000	668.270000	524.960000	5.997055e+06	4.509794e+06	1.719922e+06	12.000000	2017.000000

Exploratory Data Analysis Result

1. Dataset Overview

- Records: 10,000
- Key Columns: Region, Country, Item Type, Sales Channel, Order Date, Units Sold, Total Revenue, Total Cost, Total Profit

2. Key Metrics

- Total Revenue: \$2.35M (mean)
- Total Cost: \$1.54M (mean)
- Total Profit: \$0.81M (mean)
- Units Sold: 8,756 (mean)

3. Monthly Sales Trend

- Highest sales in December across all years
- Sales peak in Q4, indicating holiday season influence

4. Year-wise Sales

- Increasing trend from 2011 to 2014
- Significant sales peak in Q4 each year

5. Region-wise Sales

- Top Regions: Sub-Saharan Africa and Europe
- North America and Asia also major contributors

6. Product Analysis

- Top-Selling Categories: Consumer Electronics, Clothing
- Steady sales in Office Supplies and Household Items

7. Correlation Insights

- Strong positive correlation between Units Sold, Total Revenue, and Total Profit

Insights & Recommendations

- Seasonality: Focus marketing in Q4 to leverage peak sales.
- Inventory: Increase stock for high-demand products like electronics and clothing.
- Regional Strategy: Expand in high-revenue regions (Sub-Saharan Africa, Europe).
- Product Focus: Invest in popular categories to boost sales and profits.

Thank You