In [57]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [58]:
```
df = pd.read_csv(r"C:\Users\CSP\Downloads\Order.csv\Order.csv")

df.head(5)
```

Out[58]:

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	
0	1	CA-2016-152156	11/8/2016	1101102016	Second Class	CG-12520	Claire Gute	Consumer	States,
1	2	CA-2016-152156	11/8/2016	1101102016	Second Class	CG-12520	Claire Gute	Consumer	States,
2	3	CA-2016-138688	6/12/2016	601602016	Second Class	DV-13045	Darrin Van Huff	Corporate	Unitec
3	4	US-2015-108966	10/11/2015	1001802015	Standard Class	SO-20335	Sean O'Donnell	Consumer	
4	5	US-2015-108966	10/11/2015	1001802015	Standard Class	SO-20335	Sean O'Donnell	Consumer	

In [32]:
```
print("Shape:", df.shape)
print("\nColumns:\n", df.columns)
df.info()
```

```
Shape: (9994, 20)

Columns:
 Index(['Row ID', 'Order ID', 'Order Date', 'Ship Date', 'Ship Mode',
       'Customer ID', 'Customer Name', 'Segment', 'Location', 'State',
       'Postal Code', 'Region', 'Product ID', 'Category', 'Sub-Category',
       'Product Name', 'Sales', 'Quantity', 'Discount', 'Profit'],
      dtype='object')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Row ID         9994 non-null   int64
 1   Order ID       9994 non-null   object
 2   Order Date     9994 non-null   object
 3   Ship Date      9994 non-null   int64
 4   Ship Mode      9994 non-null   object
 5   Customer ID    9994 non-null   object
 6   Customer Name  9994 non-null   object
 7   Segment        9994 non-null   object
 8   Location       9994 non-null   object
 9   State          9994 non-null   object
 10  Postal Code    9994 non-null   int64
 11  Region         9994 non-null   object
 12  Product ID     9994 non-null   object
 13  Category       9994 non-null   object
 14  Sub-Category   9994 non-null   object
 15  Product Name   9994 non-null   object
 16  Sales          9994 non-null   float64
 17  Quantity       9994 non-null   int64
 18  Discount       9994 non-null   float64
 19  Profit         9994 non-null   float64
dtypes: float64(3), int64(4), object(13)
memory usage: 1.5+ MB
```

Data cleaning

In [34]:
```
df = df.drop_duplicates()
```

In [35]:
```
print(df.isnull().sum())
```

```
Row ID          0
Order ID        0
Order Date      0
Ship Date       0
Ship Mode       0
Customer ID     0
Customer Name   0
Segment         0
Location        0
State           0
Postal Code     0
Region          0
Product ID      0
Category        0
Sub-Category    0
Product Name    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
```

In [78]:
```
df = df.dropna()
```

In [82]:
```
df['Sales'] = df['Sales'].fillna(df['Sales'].mean())
```

fix data types

In [87]:
```
df[['Sales','Profit','Quantity','Discount']].dtypes
```

Out[87]:
```
Sales       float64
Profit      float64
Quantity      int64
Discount    float64
dtype: object
```

In [89]:
```
df['Sales'] = df['Sales'].astype(float)
```

convert order date

In [91]:
```
df['Order Date'] = pd.to_datetime(df['Order Date'])
```

create new columns

In [93]:
```
df['Month'] = df['Order Date'].dt.to_period('M')
```

In [95]:
```
monthly_summary = df.groupby('Month').agg({
    'Sales': 'sum',
    'Profit': 'sum'
```

```
}).reset_index()

monthly_summary.head()
```

Out[95]:

	Month	Sales	Profit
0	2014-01	14236.895	2450.1907
1	2014-02	4519.892	862.3084
2	2014-03	55691.009	498.7299
3	2014-04	28295.345	3488.8352
4	2014-05	23648.287	2738.7096

Top 5 Products by Sales

In [97]:
```
top_products = df.groupby('Product Name')['Sales'].sum().sort_values(ascending=Fals

top_products
```

Out[97]:
```
Product Name
Canon imageCLASS 2200 Advanced Copier                                615
99.824
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind    274
53.384
Cisco TelePresence System EX90 Videoconferencing Unit                226
38.480
HON 5400 Series Task Chairs for Big and Tall                         218
70.576
GBC DocuBind TL300 Electric Binding System                           198
23.479
Name: Sales, dtype: float64
```

Sales by Region

In [99]:
```
region_sales = df.groupby('Region')['Sales'].sum()

region_sales
```

Out[99]:
```
Region
Central    501239.8908
East       678781.2400
South      391721.9050
West       725457.8245
Name: Sales, dtype: float64
```

Outlier Detection(iqr method)

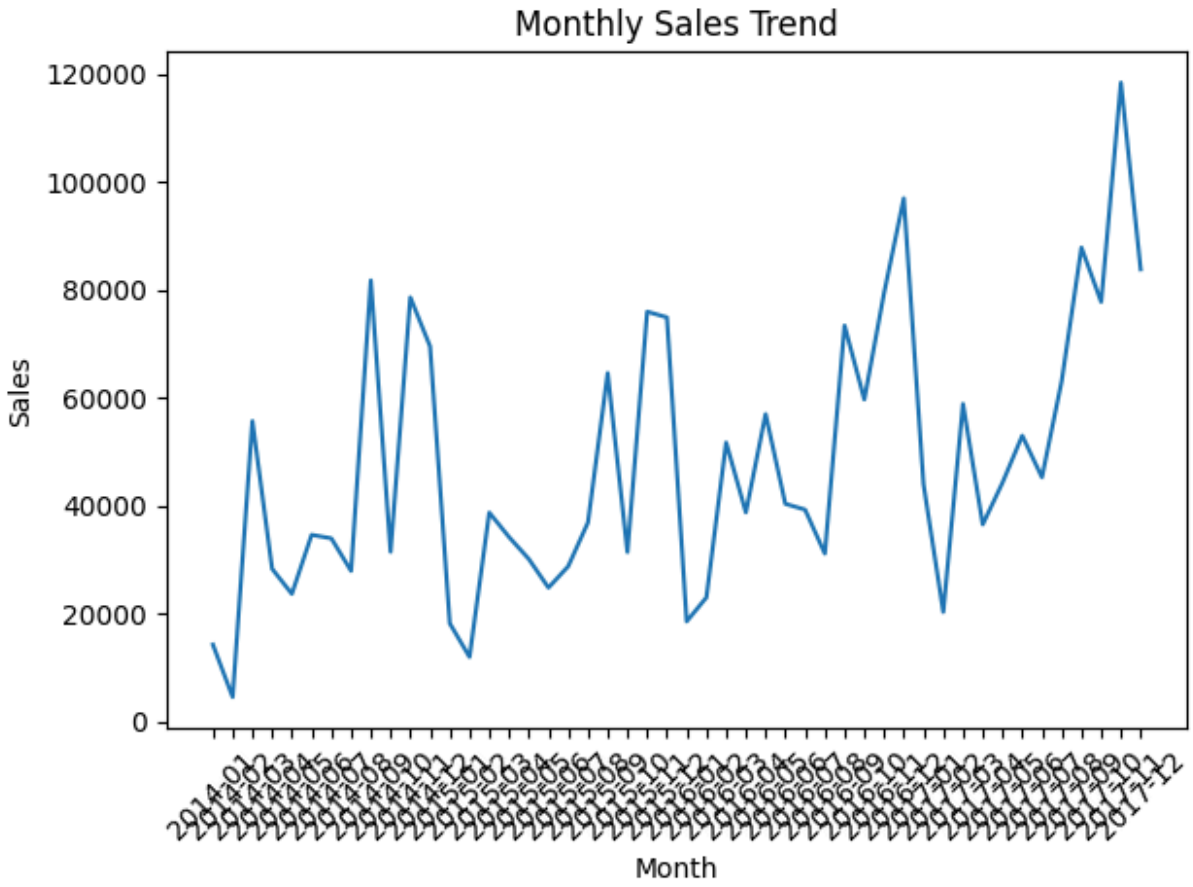
In [101…
```
Q1 = df['Sales'].quantile(0.25)
Q3 = df['Sales'].quantile(0.75)
IQR = Q3 - Q1
```

In [103…
```
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

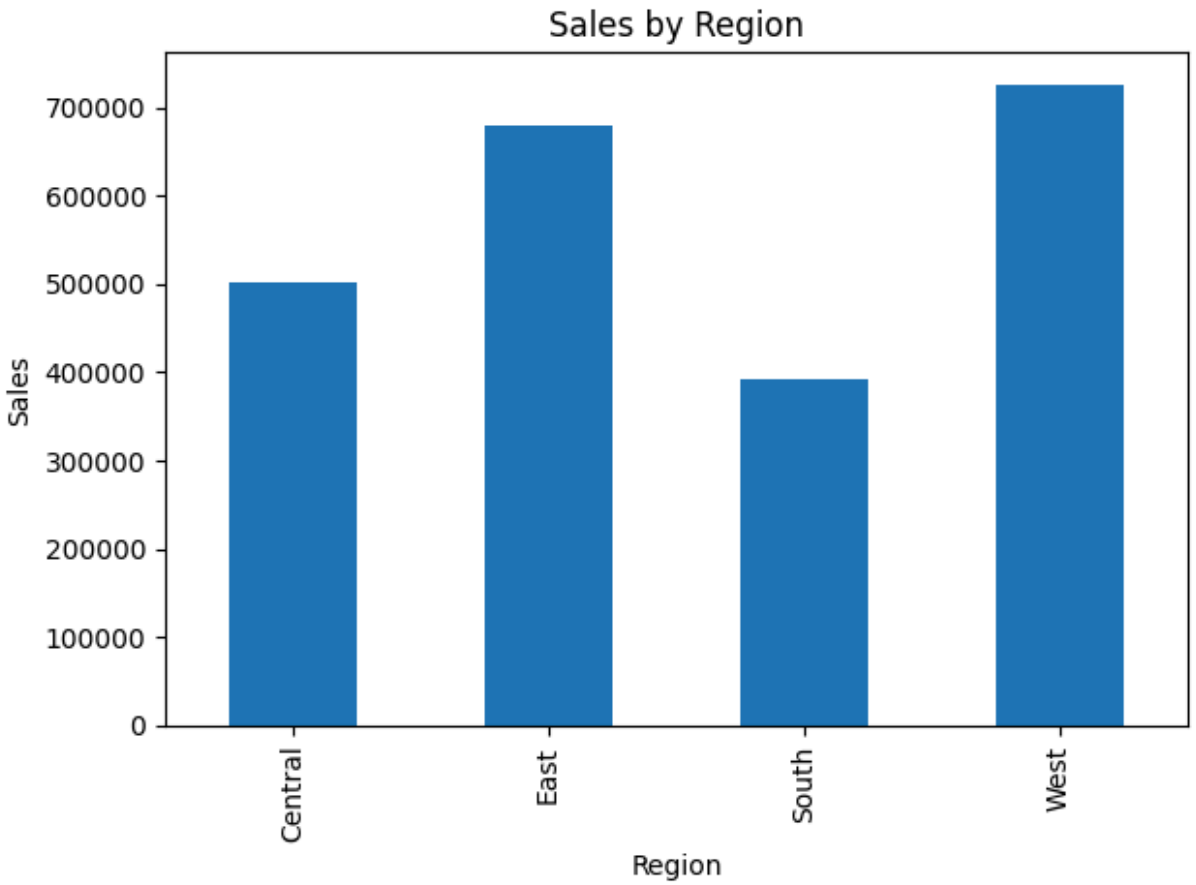
In [105…
```
df_clean = df[(df['Sales'] >= lower_bound) &
              (df['Sales'] <= upper_bound)]

print("Original rows:", df.shape[0])
print("After removing outliers:", df_clean.shape[0])
```

```
Original rows: 9994
After removing outliers: 8827
```

In [50]:
```
plt.figure()
plt.plot(monthly_summary['Month'].astype(str), monthly_summary['Sales'])
plt.xticks(rotation=45)
plt.title("Monthly Sales Trend")
plt.xlabel("Month")
plt.ylabel("Sales")
plt.tight_layout()
plt.savefig("monthly_sales_trend.png")
plt.show()
```



In [107…
```
plt.figure()
region_sales.plot(kind='bar')
plt.title("Sales by Region")
plt.xlabel("Region")
plt.ylabel("Sales")
plt.tight_layout()
plt.savefig("sales_by_region.png")
plt.show()
```



In [53]:
```
import os
print(os.getcwd())
```

```
C:\Users\CSP
```

In [111…
```
df_clean.to_csv(r"C:/Users/CSP/Downloads/cleaned_sale_data.csv", index=False)
```

In [ ]:

In [ ]: