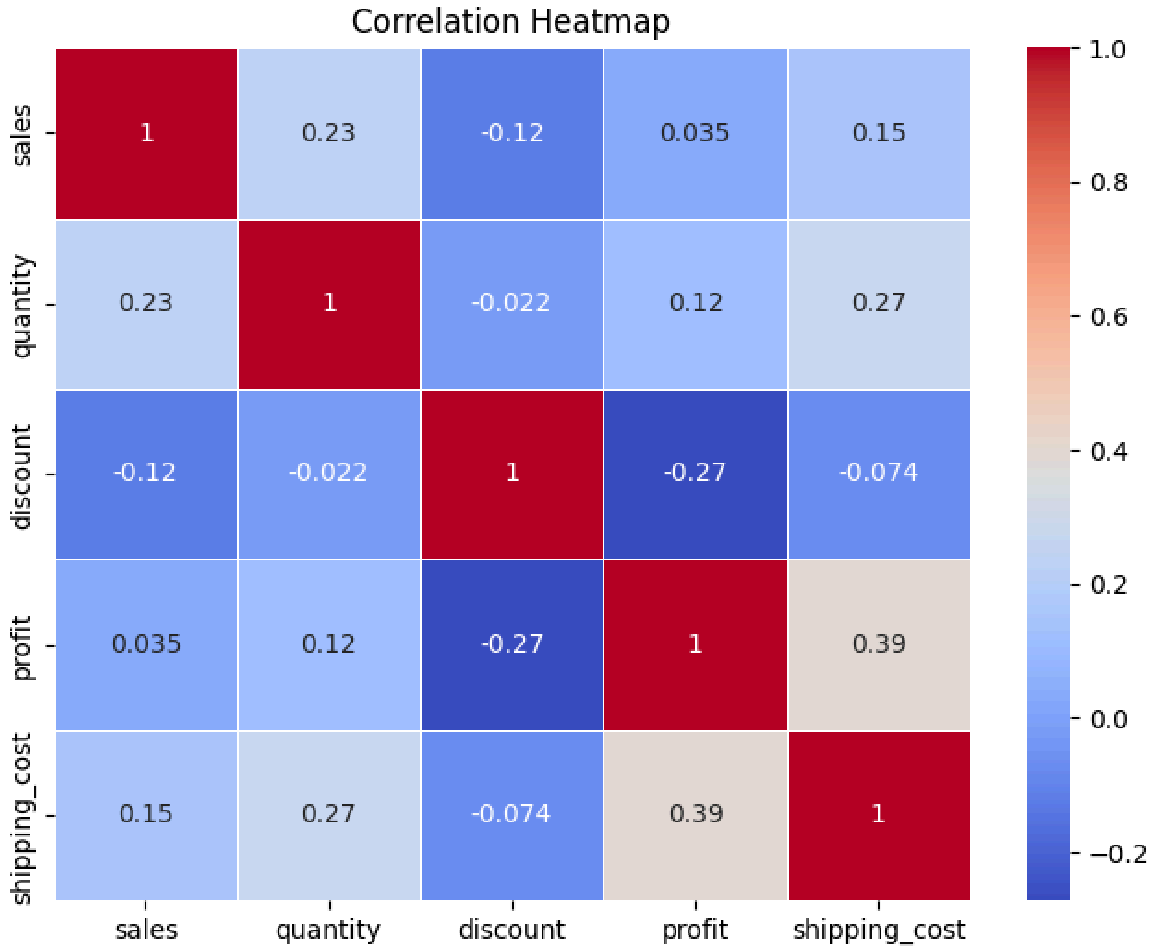
In [1]:

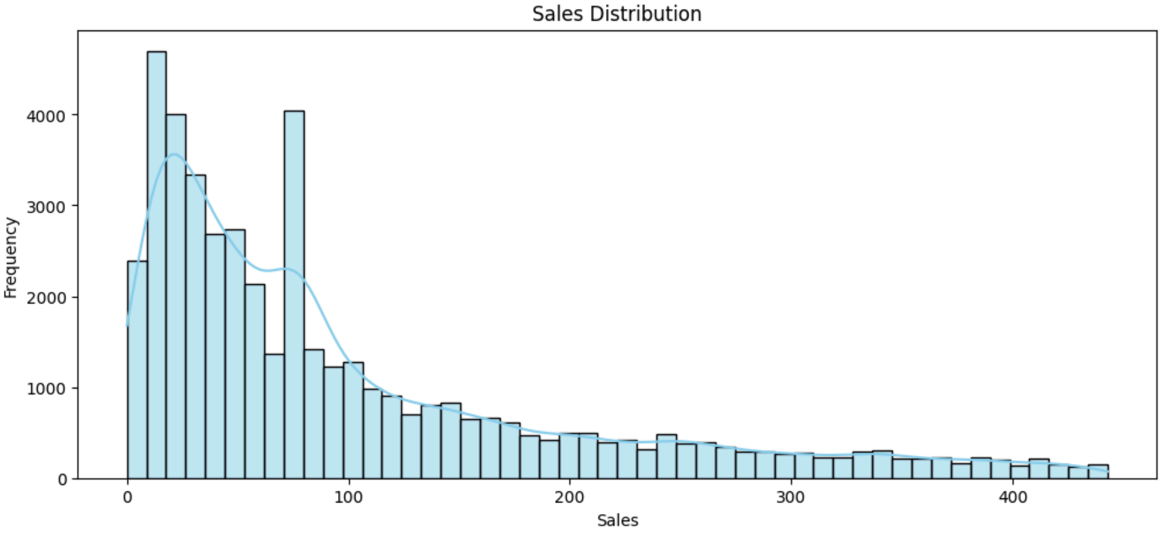
```
# 📌 Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error

# 📌 Load dataset (Update the file path if necessary)
file_path = r"D:\Internship\Main Flow\Task 2\sales data.csv"
df = pd.read_csv(file_path)

# ✅ Convert sales to numeric
df['sales'] = pd.to_numeric(df['sales'], errors='coerce')

# ✅ Convert date columns to datetime
df['order_date'] = pd.to_datetime(df['order_date'], errors='coerce')
df['ship_date'] = pd.to_datetime(df['ship_date'], errors='coerce')

# ✅ Fill missing values in sales with median
df['sales'] = df['sales'].fillna(df['sales'].median())

# ✅ Remove duplicates
df.drop_duplicates(inplace=True)

# ✅ Outlier Detection & Handling (Using IQR Method)
Q1 = df['sales'].quantile(0.25)
Q3 = df['sales'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
df = df[(df['sales'] >= lower_bound) & (df['sales'] <= upper_bound)]

# ✅ 📊 Statistical Analysis
print("🔍 Statistical Summary of the Dataset:")
print(df[['sales', 'profit', 'discount']].describe())

# ✅ 📊 Correlation Heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(df[['sales', 'quantity', 'discount', 'profit', 'shipping_cost']].cor
plt.title("Correlation Heatmap")
plt.show()

# ✅ 📊 Sales Distribution Histogram
plt.figure(figsize=(12, 5))
sns.histplot(df['sales'], bins=50, kde=True, color='skyblue')
plt.title("Sales Distribution")
plt.xlabel("Sales")
plt.ylabel("Frequency")
plt.show()

# ✅ 📊 Boxplot for Outliers in Sales
plt.figure(figsize=(10, 5))
sns.boxplot(x=df['sales'], color='lightcoral')
plt.title("Sales Outliers")
plt.show()
```

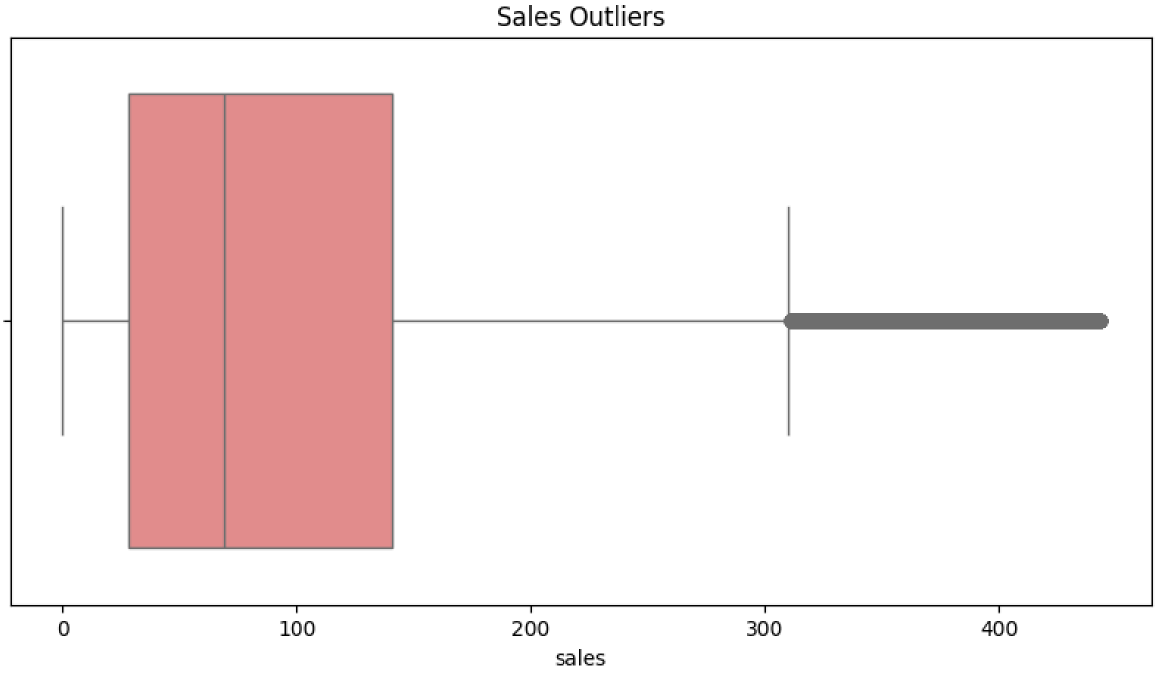
```
# ✅ 📊 Sales by Region (Bar Chart)
region_sales = df.groupby('region')['sales'].sum().sort_values(ascending=False)
plt.figure(figsize=(10, 5))
sns.barplot(x=region_sales.index, y=region_sales.values, hue=region_sales.index,
plt.title("Total Sales by Region")
plt.xlabel("Region")
plt.ylabel("Sales")
plt.xticks(rotation=45)
plt.show()

# ✅ 📊 Sales by Region (Pie Chart)
plt.figure(figsize=(8, 8))
plt.pie(region_sales, labels=region_sales.index, autopct='%1.1f%%', colors=sns.c
plt.title("Sales Distribution by Region")
plt.show()

# ✅ 📊 Sales by Category (Bar Chart)
category_sales = df.groupby('category')['sales'].sum().sort_values(ascending=Fal
plt.figure(figsize=(8, 5))
sns.barplot(x=category_sales.index, y=category_sales.values, hue=category_sales.
plt.title("Total Sales by Category")
plt.xlabel("Category")
plt.ylabel("Sales")
plt.show()

# ✅ 📊 Time Series Analysis (Sales over Time)
plt.figure(figsize=(12, 5))
df.groupby('order_date')['sales'].sum().plot()
plt.title("Sales Trend Over Time")
plt.xlabel("Date")
plt.ylabel("Total Sales")
plt.show()

# ✅ 📊 Scatter Plot: Profit vs Discount
plt.figure(figsize=(10, 5))
sns.scatterplot(x=df['discount'], y=df['profit'], alpha=0.5)
plt.title("Profit vs. Discount")
plt.xlabel("Discount")
plt.ylabel("Profit")
plt.show()

# ✅ 💡 Predictive Model: Linear Regression to Predict Sales
df_model = df[['profit', 'discount', 'sales']].dropna()

# Features and Target
X = df_model[['profit', 'discount']]
y = df_model['sales']

# Split Data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Train Model
model = LinearRegression()
model.fit(X_train, y_train)

# Predictions
y_pred = model.predict(X_test)

# Model Evaluation
r2 = r2_score(y_test, y_pred)
```

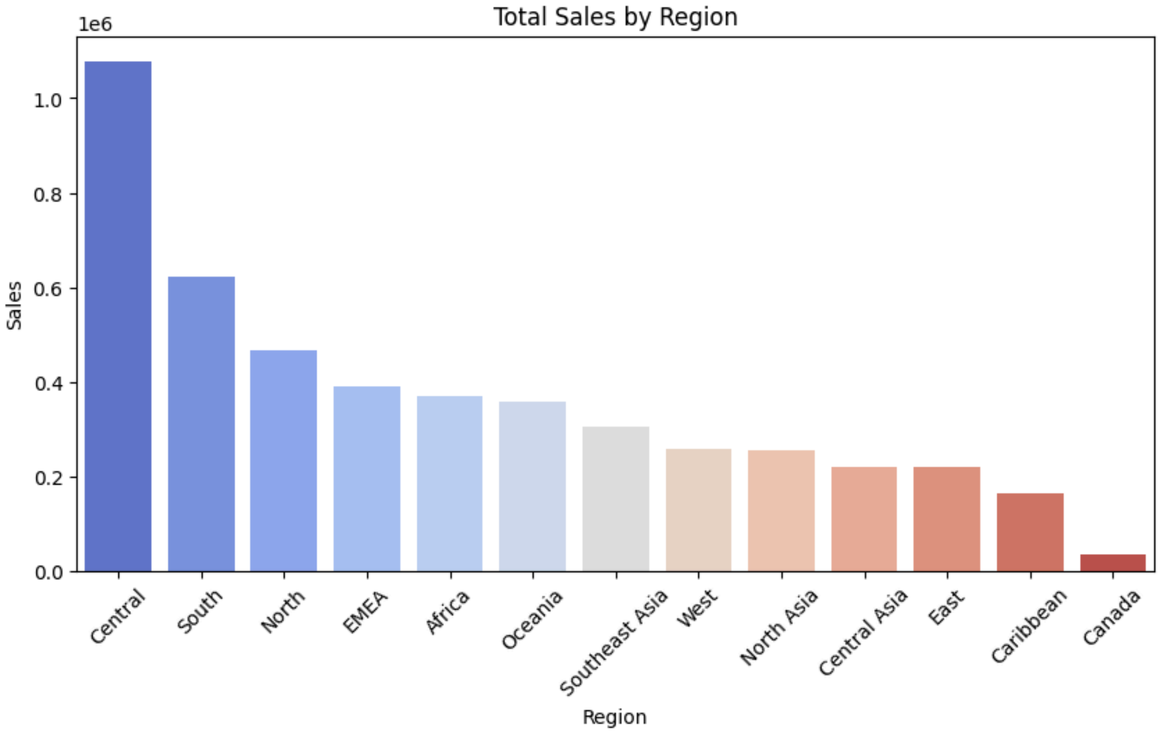
```python
mse = mean_squared_error(y_test, y_pred)

print(f"🔍 Linear Regression Model Performance:")
print(f"📈 R² Score: {r2:.4f}")
print(f"📉 Mean Squared Error: {mse:.2f}")

# ✅ 📊 Scatter Plot: Actual vs Predicted Sales
plt.figure(figsize=(8, 5))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.5, color="blue")
plt.plot(y_test, y_test, color="red", linestyle="--")  # 45-degree reference lin
plt.title("Actual vs Predicted Sales")
plt.xlabel("Actual Sales")
plt.ylabel("Predicted Sales")
plt.show()

# ✅ 🔥 Summary Insights & Recommendations
most_profitable_region = region_sales.idxmax()
least_profitable_region = region_sales.idxmin()
top_category = category_sales.idxmax()

insights = f"""
Key Insights:
1. Sales show a strong correlation with profit but a weak correlation with disco
2. The highest sales are observed in the '{most_profitable_region}' region, whil
3. The '{top_category}' category contributes the most to total sales.
4. Discounts do not always lead to higher profits, as observed in the scatter pl

Recommendations:
- Focus marketing efforts on top-performing regions like '{most_profitable_regio
- Optimize discount strategies based on the observed impact on profits.
- Continue tracking sales trends to refine business strategies.
"""
print(insights)
```

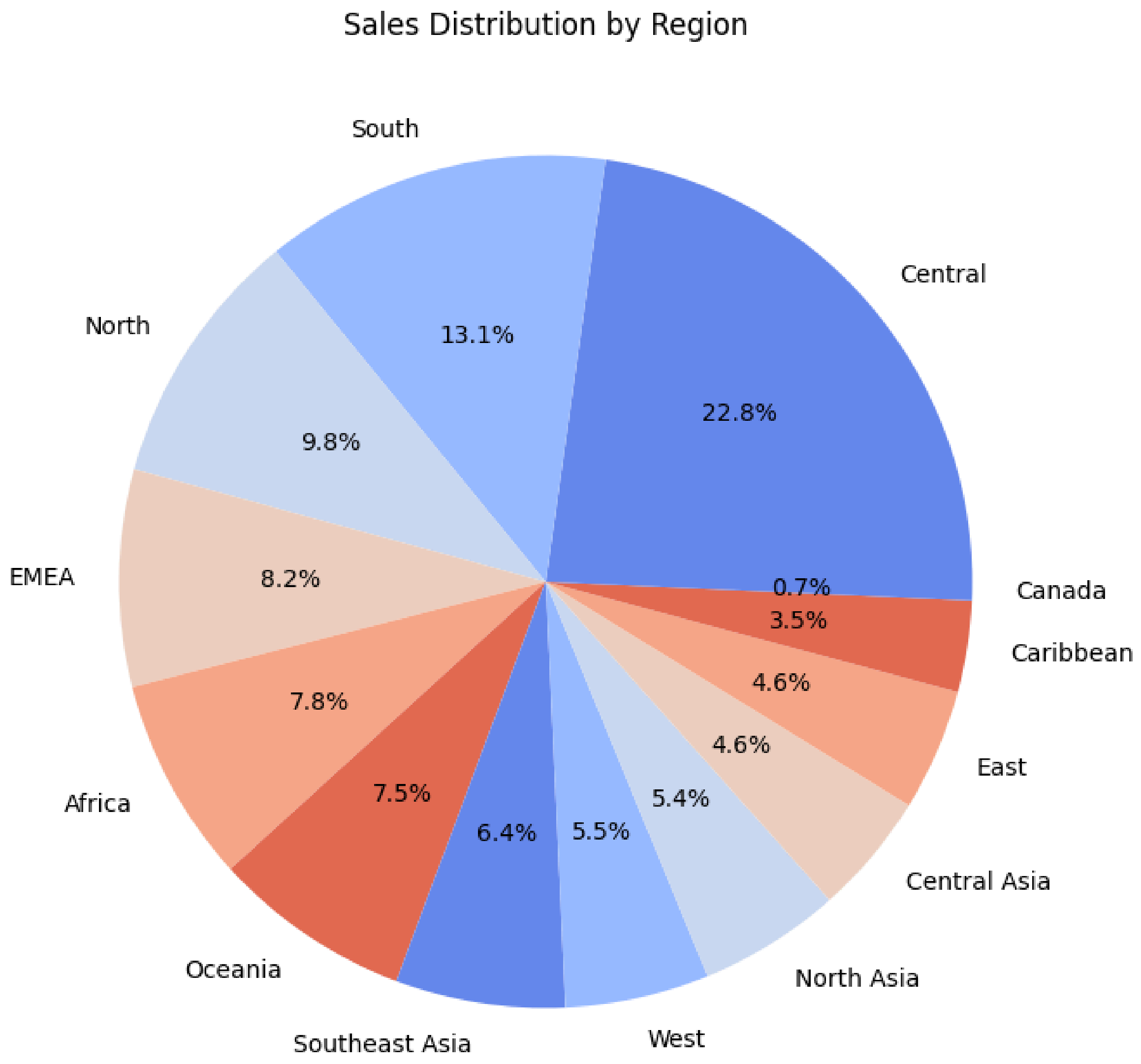
```
🔍 Statistical Summary of the Dataset:
              sales         profit      discount
count  46260.000000   46260.000000  46260.000000
mean     102.400432      24.194584      0.146222
std      101.429146     169.868055      0.216810
min        0.000000   -6599.978000      0.000000
25%       28.000000       0.000000      0.000000
50%       69.000000       8.080000      0.000000
75%      141.000000      28.500000      0.200000
max      443.000000    8399.976000      0.850000
```

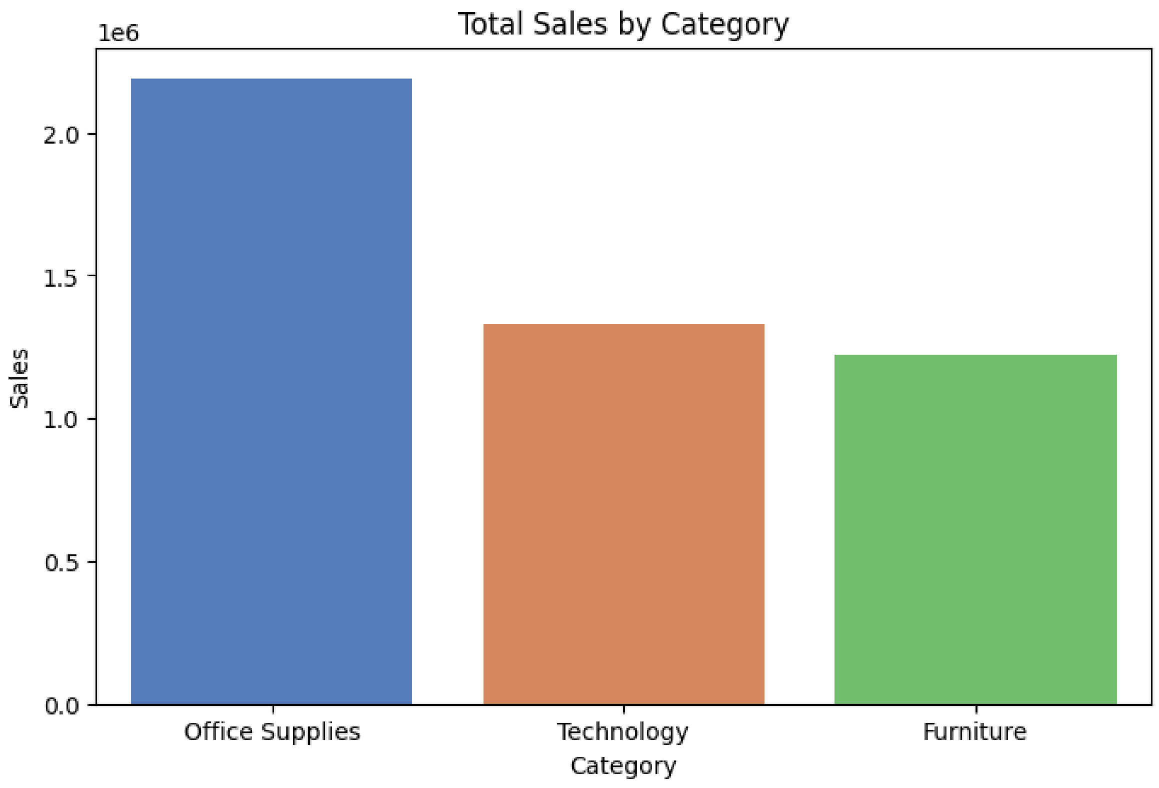

Correlation Heatmap
sales
quantity
discount
profit
shipping_cost
1
0.23
-0.12
0.035
0.15
0.23
1
-0.022
0.12
0.27
-0.12
-0.022
1
-0.27
-0.074
0.035
0.12
-0.27
1
0.39
0.15
0.27
-0.074
0.39
1
sales
quantity
discount
profit
shipping_cost
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

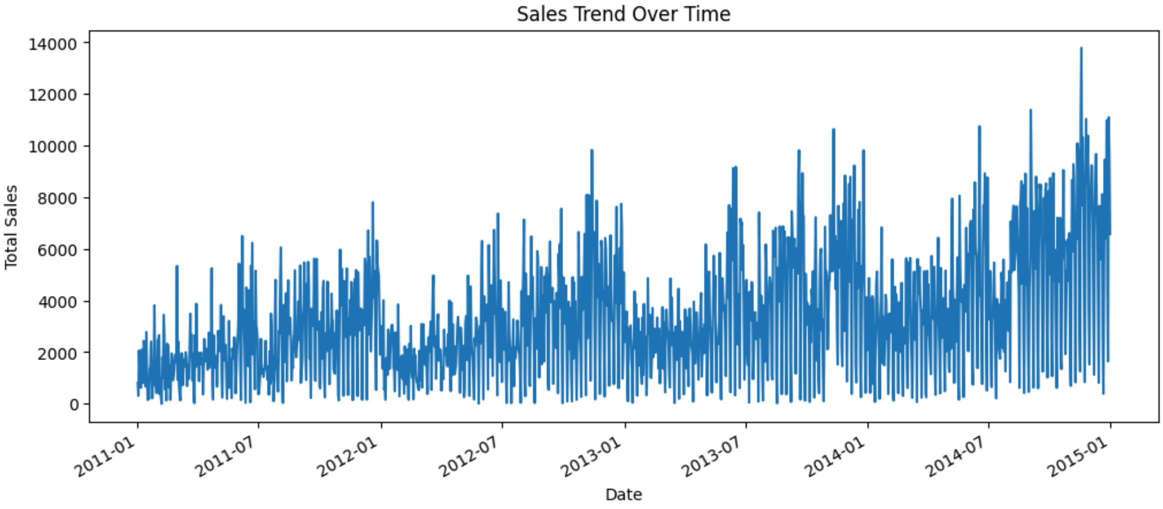

Sales Distribution
Frequency
4000
3000
2000
1000
0
0
100
200
300
400
Sales

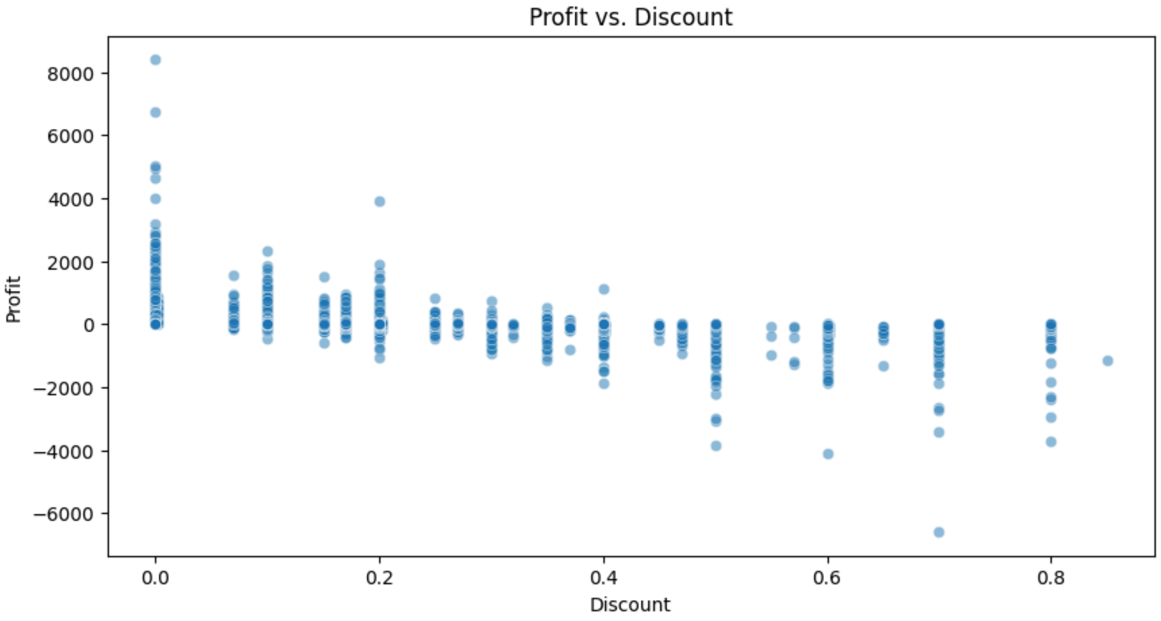

Sales Outliers
0
100
200
300
400
sales


Total Sales by Region
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Sales
Central
South
North
EMEA
Africa
Oceania
Southeast Asia
West
North Asia
Central Asia
East
Caribbean
Canada
Region

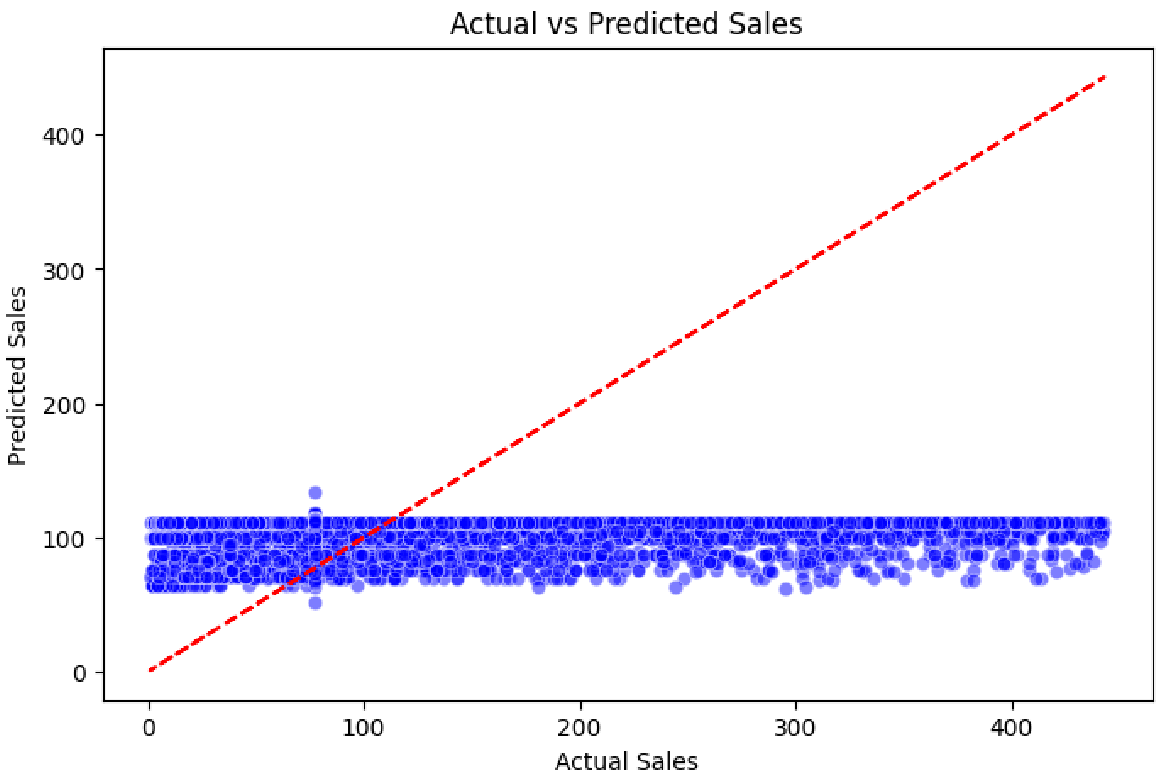

Sales Distribution by Region
South
13.1%
Central
22.8%
North
9.8%
EMEA
8.2%
Canada
0.7%
Caribbean
3.5%
East
4.6%
Central Asia
4.6%
North Asia
5.4%
West
5.5%
Southeast Asia
6.4%
Oceania
7.5%
Africa
7.8%


Total Sales by Category
1e6
2.0
1.5
1.0
0.5
0.0
Sales
Office Supplies
Technology
Furniture
Category


Sales Trend Over Time


Profit vs. Discount

```
🔍 Linear Regression Model Performance:
📈 R² Score: 0.0149
📉 Mean Squared Error: 9967.89
```



```
Key Insights:
1. Sales show a strong correlation with profit but a weak correlation with discou
nt.
2. The highest sales are observed in the 'Central' region, while the lowest sales
are in 'Canada'.
3. The 'Office Supplies' category contributes the most to total sales.
4. Discounts do not always lead to higher profits, as observed in the scatter plo
t.

Recommendations:
- Focus marketing efforts on top-performing regions like 'Central' and categories
like 'Office Supplies'.
- Optimize discount strategies based on the observed impact on profits.
- Continue tracking sales trends to refine business strategies.
```

```
In [2]: print("Amit Kumar Jha")
```

```
Amit Kumar Jha
```

```
In [ ]:
```