```
#import the libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#load  the dataset
df = pd.read_csv('sales_data.csv', encoding='latin-1')


#display the first few rows
print("first 5 rows of the dataset:")
display(df.head())

#basic information about the dataset
print("\nDataset information:")
df.info()

#statistical summary of numerical columns
print("\nStatistical summary:")
display(df.describe())


#check for duplicates
duplicates = df.duplicated().sum()
print(f"number of duplicate rows: {duplicates}")

#remove duplicates
df = df.drop_duplicates()

#handle missing values
print(f"missing values before cleaning:\n{df.isnull().sum()}")

# Select only numeric columns for filling with mean
numeric_cols = df.select_dtypes(include=['number']).columns
df[numeric_cols] = df[numeric_cols].fillna(df[numeric_cols].mean())

print(f"missing values after cleaning:\n{df.isnull().sum()}")

#convert 'date' column to datetime format
print(f"missing values after cleaning:\n{df.isnull().sum()}")

#convert 'Order Date' column to datetime format
df['Order Date'] = pd.to_datetime(df['Order Date'], format='%d-%m-%Y')

# verify the changes
print("\ndata after cleaning:")
display(df.head())

#plot sales trends over time
plt.figure(figsize=(10,6))
df.groupby('Order Date')['Sales'].sum().plot(kind='line',color='blue')
plt.title('Sales Trends Over Time')
plt.xlabel('Order Date')
plt.ylabel('Total Sales')
plt.show()

#scatter plot : profit vs discount
plt.figure(figsize=(8,6))
sns.scatterplot(x='Discount',y='Profit',data=df,color='violet')
plt.title('Profit vs Discount')
plt.xlabel('Discount')
plt.ylabel('Profit')
plt.show()

#sales distribution by region
plt.figure(figsize=(8,6))
region_sales = df.groupby('Region')['Sales'].sum()
region_sales.plot(kind='bar',color='green')
plt.title('sales by region')
plt.ylabel('total sales')
plt.show()

#heatmap for correlation
plt.figure(figsize=(8,6))
# Calculate correlation only for numeric columns
numeric_df = df.select_dtypes(include=['number'])
sns.heatmap(numeric_df.corr(),annot=True,cmap='coolwarm')
plt.title('correlation matrix')
plt.show()
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

#select features and target
x = df[['Profit','Discount']]
y = df['Sales']

#split the dataset into training and test sets
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2,random_state=42)

#train the linear regression model
model = LinearRegression()
model.fit(x_train,y_train)

#make predictions on the test set
y_pred = model.predict(x_test)

#evaluate the model
print(f"mean squared error: {mean_squared_error(y_test,y_pred):.2f}")
print(f"R-squared: {r2_score(y_test,y_pred):.2f}")
```

```
first 5 rows of the dataset:
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	...	Postal Code	Region	Product ID	Category	Ca
0	1	CA-2016-152156	08-11-2016	11-11-2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420	South	FUR-BO-10001798	Furniture	Boo
1	2	CA-2016-152156	08-11-2016	11-11-2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420	South	FUR-CH-10000454	Furniture	
2	3	CA-2016-138688	12-06-2016	16-06-2016	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...	90036	West	OFF-LA-10000240	Office Supplies	
3	4	US-2015-108966	11-10-2015	18-10-2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311	South	FUR-TA-10000577	Furniture	
4	5	US-2015-108966	11-10-2015	18-10-2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311	South	OFF-ST-10000760	Office Supplies	

5 rows × 21 columns

```
Dataset information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 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   object 
 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   Country        9994 non-null   object 
 9   City           9994 non-null   object 
 10  State          9994 non-null   object 
 11  Postal Code    9994 non-null   int64  
 12  Region         9994 non-null   object 
 13  Product ID     9994 non-null   object 
 14  Category       9994 non-null   object 
 15  Sub-Category   9994 non-null   object 
 16  Product Name   9994 non-null   object 
 17  Sales          9994 non-null   float64
 18  Quantity       9994 non-null   int64  
 19  Discount       9994 non-null   float64
 20  Profit         9994 non-null   float64
dtypes: float64(3), int64(3), object(15)
memory usage: 1.6+ MB

Statistical summary:
```

	Row ID	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	4997.500000	55190.379428	229.858001	3.789574	0.156203	28.656896
std	2885.163629	32063.693350	623.245101	2.225110	0.206452	234.260108
min	1.000000	1040.000000	0.444000	1.000000	0.000000	-6599.978000
25%	2499.250000	23223.000000	17.280000	2.000000	0.000000	1.728750
50%	4997.500000	56430.500000	54.490000	3.000000	0.200000	8.666500
75%	7495.750000	90008.000000	209.940000	5.000000	0.200000	29.364000
max	9994.000000	99301.000000	22638.480000	14.000000	0.800000	8399.976000

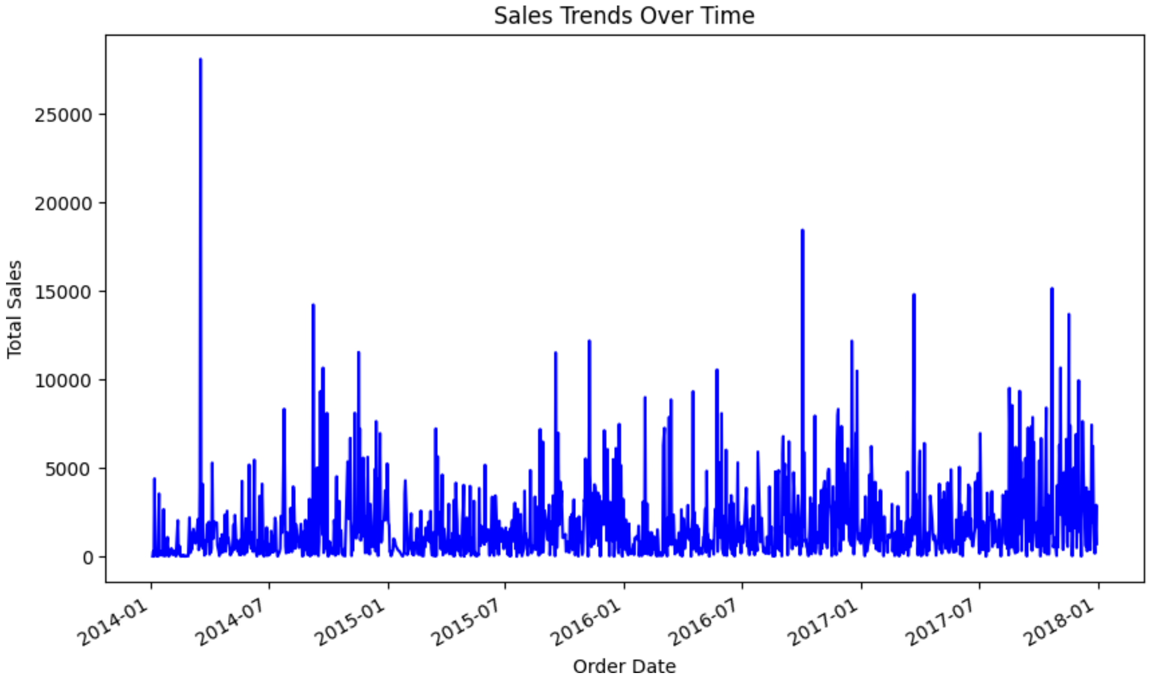
```
number of duplicate rows: 0
missing values before cleaning:
Row ID           0
Order ID         0
Order Date       0
Ship Date        0
Ship Mode        0
Customer ID      0
Customer Name    0
```

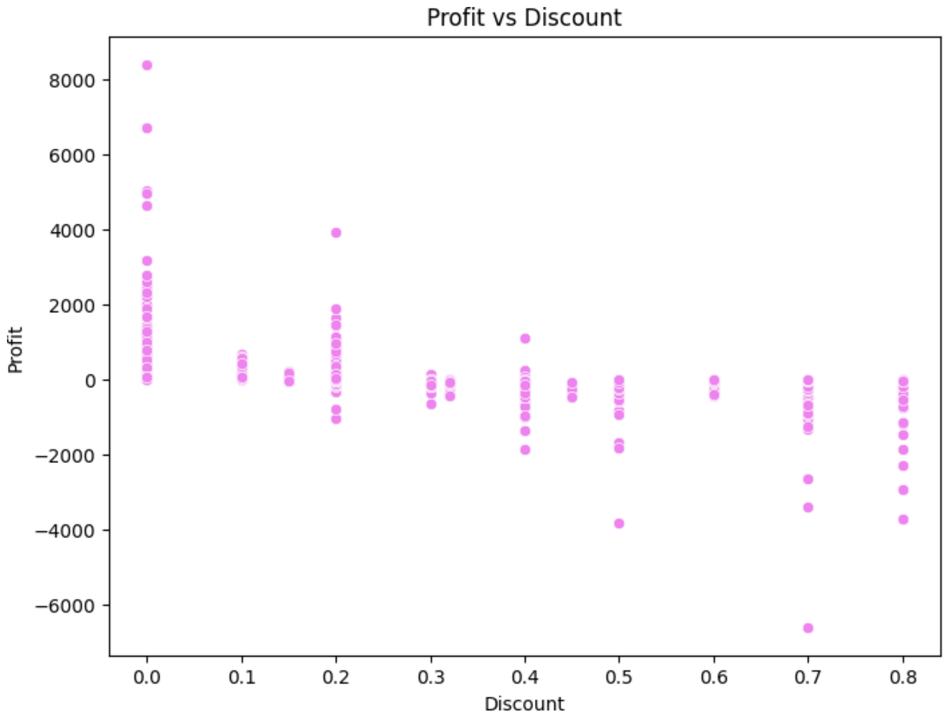
```
Segment          0
Country          0
City             0
State            0
Postal Code      0
Region           0
Product ID       0
Category         0
Sub-Category     0
Product Name     0
Sales            0
Quantity         0
Discount         0
Profit           0
dtype: int64
missing values after cleaning:
Row ID           0
Order ID         0
Order Date       0
Ship Date        0
Ship Mode        0
Customer ID      0
Customer Name    0
Segment          0
Country          0
City             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
missing values after cleaning:
Row ID           0
Order ID         0
Order Date       0
Ship Date        0
Ship Mode        0
Customer ID      0
Customer Name    0
Segment          0
Country          0
City             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

data after cleaning:
```

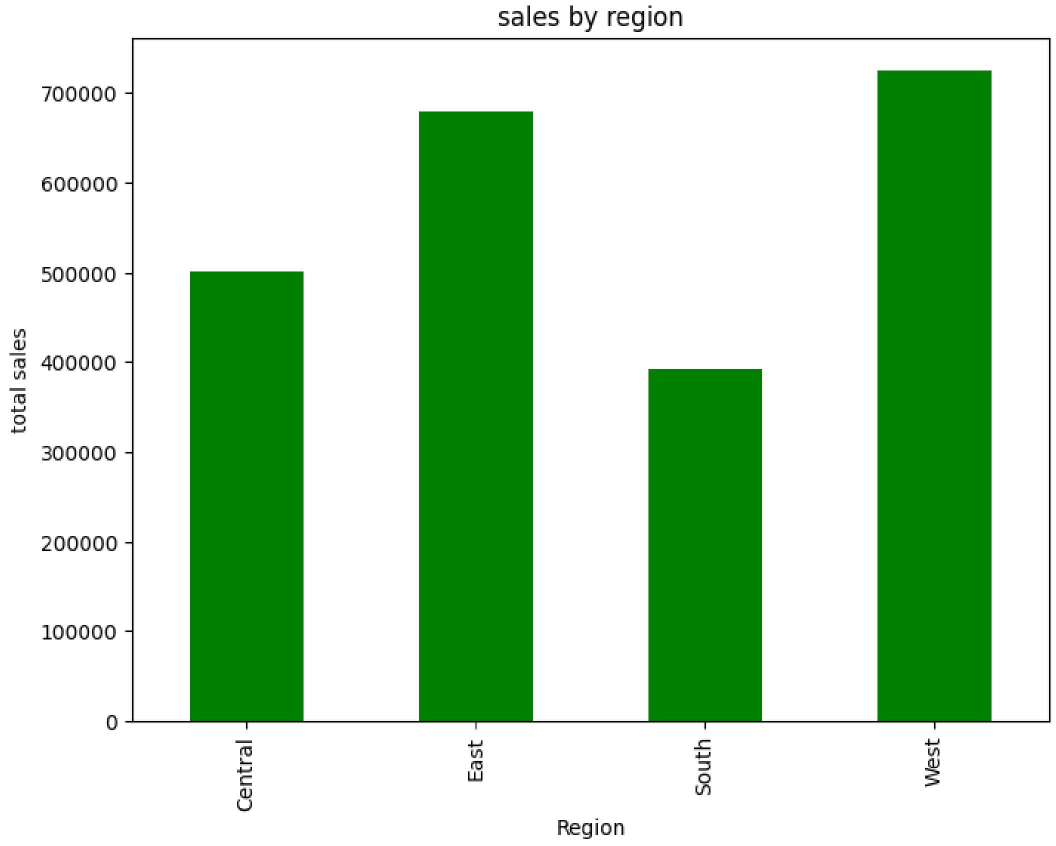
	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	...	Postal Code	Region	Product ID	Category	Ca
0	1	CA-2016-152156	2016-11-08	11-11-2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420	South	FUR-BO-10001798	Furniture	Boo
1	2	CA-2016-152156	2016-11-08	11-11-2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420	South	FUR-CH-10000454	Furniture	
2	3	CA-2016-138688	2016-06-12	16-06-2016	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...	90036	West	OFF-LA-10000240	Office Supplies	
3	4	US-2015-108966	2015-10-11	18-10-2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311	South	FUR-TA-10000577	Furniture	
		US-	2015-	18-	Standard	SO-	Sean		United	Fort				OFF-ST-	Office	

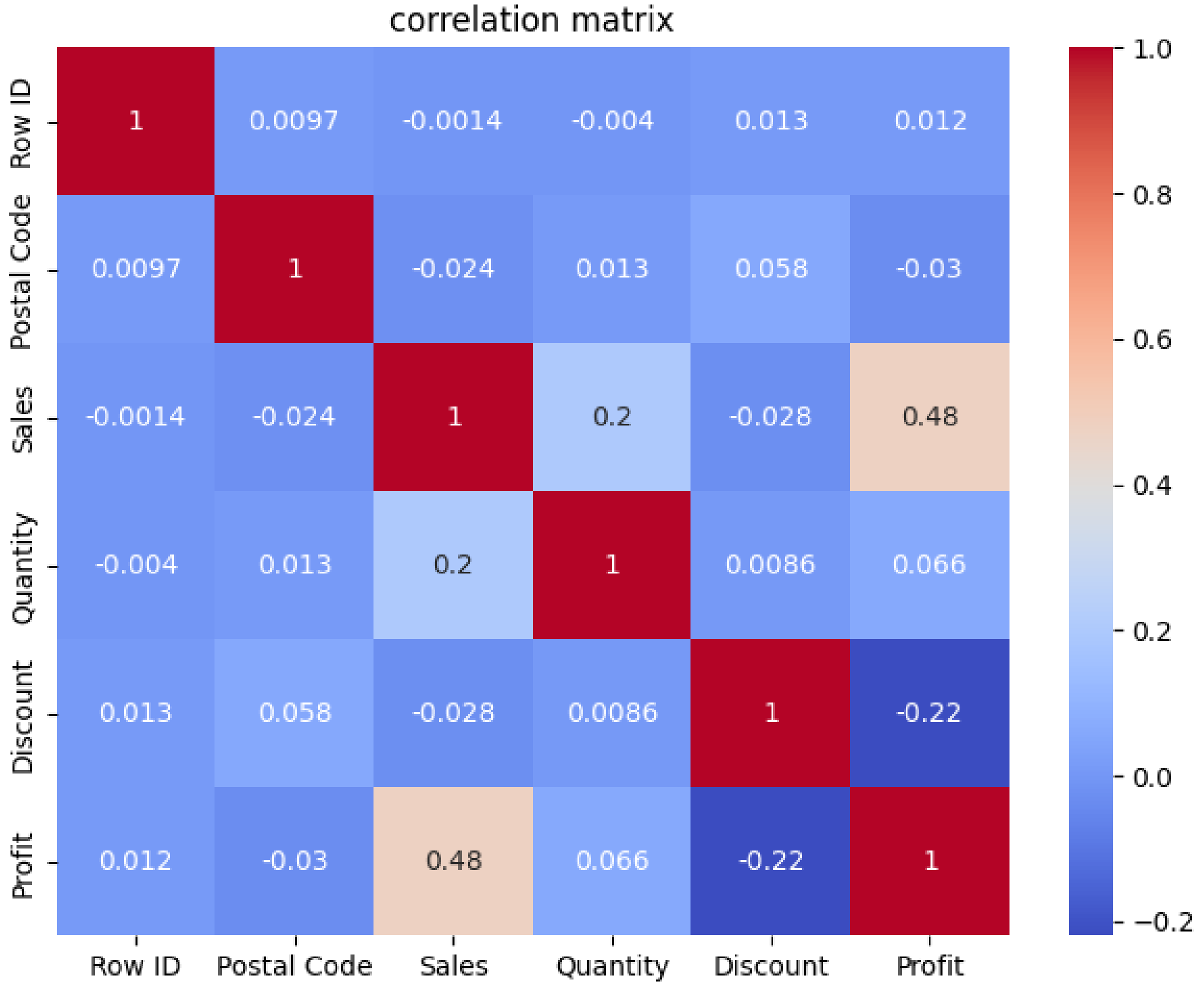
4	5	2015- 108966	10-11	10- 2015	Class	20335	O'Donnell	Consumer	States	Lauderdale	...	33311	South	10000760	Supplies

5 rows × 21 columns









	Row ID	Postal Code	Sales	Quantity	Discount	Profit
Row ID	1	0.0097	-0.0014	-0.004	0.013	0.012
Postal Code	0.0097	1	-0.024	0.013	0.058	-0.03
Sales	-0.0014	-0.024	1	0.2	-0.028	0.48
Quantity	-0.004	0.013	0.2	1	0.0086	0.066
Discount	0.013	0.058	-0.028	0.0086	1	-0.22
Profit	0.012	-0.03	0.48	0.066	-0.22	1

```
mean squared error: 700271.89
R-squared: -0.19
```