Sales_EDA

October 15, 2024

0.1 Setting Up

[4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[5]:
```
df = pd.read_csv('superstore.csv')
df.head()
```

[5]:

	Category	City	Country	Customer.ID	Customer.Name	…	Sales
0	Office Supplies	Los Angeles	United States	LS-172304	Lycoris Saunders	…	19
1	Office Supplies	Los Angeles	United States	MV-174854	Mark Van Huff	…	19
2	Office Supplies	Los Angeles	United States	CS-121304	Chad Sievert	…	21
3	Office Supplies	Los Angeles	United States	CS-121304	Chad Sievert	…	111
4	Office Supplies	Los Angeles	United States	AP-109154	Arthur Prichep	…	6

	Discount	Market		Order.Date	Order.ID	…
0	0.0	US	1	2011-01-07 00:00:00.000	CA-2011-130813	…
1	0.0	US	1	2011-01-21 00:00:00.000	CA-2011-148614	…
2	0.0	US	1	2011-08-05 00:00:00.000	CA-2011-118962	…
3	0.0	US	1	2011-08-05 00:00:00.000	CA-2011-118962	…
4	0.0	US	1	2011-09-29 00:00:00.000	CA-2011-146969	…

	Segment	Ship.Date	Ship.Mode	Shipping.Cost
0	Consumer	2011-01-09 00:00:00.000	Second Class	4.37
1	Consumer	2011-01-26 00:00:00.000	Standard Class	0.94
2	Consumer	2011-08-09 00:00:00.000	Standard Class	1.81
3	Consumer	2011-08-09 00:00:00.000	Standard Class	4.59
4	Consumer	2011-10-03 00:00:00.000	Standard Class	1.32

	State	Sub.Category	Year	Market2	weeknum
0	California	Paper	2011	North America	2
1	California	Paper	2011	North America	4
2	California	Paper	2011	North America	32
3	California	Paper	2011	North America	32
4	California	Paper	2011	North America	40

```
[5 rows x 27 columns]
```

[6]:
```
df.columns
```

[6]:
```
Index(['Category', 'City', 'Country', 'Customer.ID', 'Customer.Name',
       'Discount', 'Market', '  ', 'Order.Date', 'Order.ID', 'Order.Priority',
       'Product.ID', 'Product.Name', 'Profit', 'Quantity', 'Region', 'Row.ID',
       'Sales', 'Segment', 'Ship.Date', 'Ship.Mode', 'Shipping.Cost', 'State',
       'Sub.Category', 'Year', 'Market2', 'weeknum'],
      dtype='object')
```

[7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 27 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Category        51290 non-null  object
 1   City            51290 non-null  object
 2   Country         51290 non-null  object
 3   Customer.ID     51290 non-null  object
 4   Customer.Name   51290 non-null  object
 5   Discount        51290 non-null  float64
 6   Market          51290 non-null  object
 7                   51290 non-null  int64
 8   Order.Date      51290 non-null  object
 9   Order.ID        51290 non-null  object
 10  Order.Priority  51290 non-null  object
 11  Product.ID      51290 non-null  object
 12  Product.Name    51290 non-null  object
 13  Profit          51290 non-null  float64
 14  Quantity        51290 non-null  int64
 15  Region          51290 non-null  object
 16  Row.ID          51290 non-null  int64
 17  Sales           51290 non-null  int64
 18  Segment         51290 non-null  object
 19  Ship.Date       51290 non-null  object
 20  Ship.Mode       51290 non-null  object
 21  Shipping.Cost   51290 non-null  float64
 22  State           51290 non-null  object
 23  Sub.Category    51290 non-null  object
 24  Year            51290 non-null  int64
 25  Market2         51290 non-null  object
 26  weeknum         51290 non-null  int64
dtypes: float64(3), int64(6), object(18)
memory usage: 10.6+ MB
```

[8]:
```
df.describe()
```

[8]:

	Discount		Profit	Quantity	Row.ID
count	51290.000000	51290.0	51290.000000	51290.000000	51290.00000
mean	0.142908	1.0	28.610982	3.476545	25645.50000
std	0.212280	0.0	174.340972	2.278766	14806.29199
min	0.000000	1.0	-6599.978000	1.000000	1.00000
25%	0.000000	1.0	0.000000	2.000000	12823.25000
50%	0.000000	1.0	9.240000	3.000000	25645.50000
75%	0.200000	1.0	36.810000	5.000000	38467.75000
max	0.850000	1.0	8399.976000	14.000000	51290.00000

	Sales	Shipping.Cost	Year	weeknum
count	51290.000000	51290.000000	51290.000000	51290.000000
mean	246.498440	26.375818	2012.777208	31.287112
std	487.567175	57.296810	1.098931	14.429795
min	0.000000	0.002000	2011.000000	1.000000
25%	31.000000	2.610000	2012.000000	20.000000
50%	85.000000	7.790000	2013.000000	33.000000
75%	251.000000	24.450000	2014.000000	44.000000
max	22638.000000	933.570000	2014.000000	53.000000

0.2 EDA starts here

[9]:
```
df.isnull().sum()
```

[9]:
```
Category          0
City              0
Country           0
Customer.ID       0
Customer.Name     0
Discount          0
Market            0
                  0
Order.Date        0
Order.ID          0
Order.Priority    0
Product.ID        0
Product.Name      0
Profit            0
Quantity          0
Region            0
Row.ID            0
Sales             0
Segment           0
Ship.Date         0
Ship.Mode         0
```

```
Shipping.Cost     0
State             0
Sub.Category      0
Year              0
Market2           0
weeknum           0
dtype: int64
```

[12]:
```
[features for features in df.columns if df[features].isnull().sum() > 0]
```

[12]:
```
[]
```

[14]:
```
df.dtypes
```

[14]:
```
Category           object
City               object
Country            object
Customer.ID        object
Customer.Name      object
Discount          float64
Market             object
                   int64
Order.Date         object
Order.ID           object
Order.Priority     object
Product.ID         object
Product.Name       object
Profit            float64
Quantity            int64
Region             object
Row.ID              int64
Sales               int64
Segment            object
Ship.Date          object
Ship.Mode          object
Shipping.Cost     float64
State              object
Sub.Category       object
Year                int64
Market2            object
weeknum             int64
dtype: object
```

[23]:
```
df=df.drop(df.columns[7], axis=1)
```

[24]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Category        51290 non-null  object
 1   City            51290 non-null  object
 2   Country         51290 non-null  object
 3   Customer.ID     51290 non-null  object
 4   Customer.Name   51290 non-null  object
 5   Discount        51290 non-null  float64
 6   Market          51290 non-null  object
 7   Order.Date      51290 non-null  object
 8   Order.ID        51290 non-null  object
 9   Order.Priority  51290 non-null  object
 10  Product.ID      51290 non-null  object
 11  Product.Name    51290 non-null  object
 12  Profit          51290 non-null  float64
 13  Quantity        51290 non-null  int64
 14  Region          51290 non-null  object
 15  Row.ID          51290 non-null  int64
 16  Sales           51290 non-null  int64
 17  Segment         51290 non-null  object
 18  Ship.Date       51290 non-null  object
 19  Ship.Mode       51290 non-null  object
 20  Shipping.Cost   51290 non-null  float64
 21  State           51290 non-null  object
 22  Sub.Category    51290 non-null  object
 23  Year            51290 non-null  int64
 24  Market2         51290 non-null  object
 25  weeknum         51290 non-null  int64
dtypes: float64(3), int64(5), object(18)
memory usage: 10.2+ MB
```

[17]:
```
df.Country.value_counts().head(10)
```

[17]:
```
Country
United States     9994
Australia         2837
France            2827
Mexico            2644
Germany           2065
China             1880
United Kingdom    1633
Brazil            1599
India             1555
Indonesia         1390
Name: count, dtype: int64
```

[20]:
```
df.City.value_counts().head(10)
```

[20]:
```
City
New York City    915
Los Angeles      747
Philadelphia     537
San Francisco    510
Santo Domingo    443
Manila           432
Seattle          428
Houston          377
Tegucigalpa      362
Jakarta          337
Name: count, dtype: int64
```

[21]:
```
df.Market.value_counts().head(10)
```

[21]:
```
Market
APAC      11002
LATAM     10294
EU        10000
US         9994
EMEA       5029
Africa     4587
Canada      384
Name: count, dtype: int64
```

[26]:
```
 df = df.rename(columns={'Customer.ID': 'Customer_ID', 'Customer.Name':␣
 ↪'Customer_Name', 'Order.Date': 'Order_Date', 'Order.ID': 'Order_ID',
                         'Order.Priority': 'Order_Priority', 'Product.ID':␣
 ↪'Product_ID', 'Product.Name': 'Product_Name', 'Row.ID': 'Row_ID',
                         'Ship.Date': 'Ship_Date', 'Ship.Mode': 'Ship_Mode',␣
 ↪'Shipping.Cost': 'Shipping_Cost', 'Sub.Category': 'Sub_Category' })
```

[27]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Category        51290 non-null  object
 1   City            51290 non-null  object
 2   Country         51290 non-null  object
 3   Customer_ID     51290 non-null  object
 4   Customer_Name   51290 non-null  object
 5   Discount        51290 non-null  float64
 6   Market          51290 non-null  object
```

```
 7   Order_Date      51290 non-null  object
 8   Order_ID        51290 non-null  object
 9   Order_Priority  51290 non-null  object
 10  Product_ID      51290 non-null  object
 11  Product_Name    51290 non-null  object
 12  Profit          51290 non-null  float64
 13  Quantity        51290 non-null  int64
 14  Region          51290 non-null  object
 15  Row_ID          51290 non-null  int64
 16  Sales           51290 non-null  int64
 17  Segment         51290 non-null  object
 18  Ship_Date       51290 non-null  object
 19  Ship_Mode       51290 non-null  object
 20  Shipping_Cost   51290 non-null  float64
 21  State           51290 non-null  object
 22  Sub_Category    51290 non-null  object
 23  Year            51290 non-null  int64
 24  Market2         51290 non-null  object
 25  weeknum         51290 non-null  int64
dtypes: float64(3), int64(5), object(18)
memory usage: 10.2+ MB
```

[28]:
```
df.Order_Priority.value_counts().head(10)
```

[28]:
```
Order_Priority
Medium      29433
High        15501
Critical     3932
Low          2424
Name: count, dtype: int64
```

[30]:
```
df.Product_Name.value_counts().head(20)
```

[30]:
```
Product_Name
Staples                                           227
Cardinal Index Tab, Clear                          92
Eldon File Cart, Single Width                      90
Rogers File Cart, Single Width                     84
Ibico Index Tab, Clear                             83
Sanford Pencil Sharpener, Water Color              80
Smead File Cart, Single Width                      77
Stanley Pencil Sharpener, Water Color              75
Acco Index Tab, Clear                              75
Avery Index Tab, Clear                             74
Tenex File Cart, Single Width                      70
Stockwell Paper Clips, Assorted Sizes              65
Boston Pencil Sharpener, Water Color               59
Binney & Smith Pencil Sharpener, Water Color       55
```

```
Stockwell Thumb Tacks, 12 Pack                  53
Binney & Smith Sketch Pad, Blue                 52
Cardinal Binding Machine, Economy               52
Wilson Jones 3-Hole Punch, Durable              52
Avery Binder Covers, Recycled                   52
Apple Smart Phone, Full Size                    51
Name: count, dtype: int64
```

[31]:
```
df.Category.value_counts().head(10)
```

[31]:
```
Category
Office Supplies    31273
Technology         10141
Furniture           9876
Name: count, dtype: int64
```

[32]:
```
df.Quantity.value_counts().head(10)
```

[32]:
```
Quantity
2     12748
3      9682
1      8963
4      6385
5      4882
6      3020
7      2385
8      1361
9       987
10      276
Name: count, dtype: int64
```

[33]:
```
df.Region.value_counts().head(10)
```

[33]:
```
Region
Central           11117
South              6645
EMEA               5029
North              4785
Africa             4587
Oceania            3487
West               3203
Southeast Asia     3129
East               2848
North Asia         2338
Name: count, dtype: int64
```

[34]:
```
df.Segment.value_counts().head(10)
```

[34]:
```
Segment
Consumer       26518
Corporate      15429
Home Office     9343
Name: count, dtype: int64
```

[35]:
```
df.Ship_Mode.value_counts().head(10)
```

[35]:
```
Ship_Mode
Standard Class    30775
Second Class      10309
First Class        7505
Same Day           2701
Name: count, dtype: int64
```

[36]:
```
df.Market2.value_counts().head(10)
```

[36]:
```
Market2
APAC             11002
North America    10378
LATAM            10294
EU               10000
EMEA              5029
Africa            4587
Name: count, dtype: int64
```

[37]:
```
df.Year.value_counts().head(10)
```

[37]:
```
Year
2014    17531
2013    13799
2012    10962
2011     8998
Name: count, dtype: int64
```

[41]:
```
df.weeknum.value_counts().head(10)
```

[41]:
```
weeknum
47    1527
46    1524
45    1508
52    1461
38    1453
48    1441
49    1440
39    1426
51    1381
50    1378
```

```
Name: count, dtype: int64
```

[42]:
```
df.Customer_ID.value_counts().head(10)
```

[42]:
```
Customer_ID
JG-158051    40
BC-111252    37
WB-218504    37
AF-108701    36
JG-158052    35
CS-121751    35
NH-186101    35
BW-111101    35
TZ-214453    34
MR-175452    34
Name: count, dtype: int64
```

[43]:
```
df.Customer_Name.value_counts().head(10)
```

[43]:
```
Customer_Name
Muhammed Yedwab    108
Steven Ward        106
Gary Hwang         102
Patrick O'Brill    102
Bill Eplett        102
Harry Greene       101
Eric Murdock       100
Art Ferguson        98
Brosina Hoffman     97
Bart Watters        96
Name: count, dtype: int64
```

0.3 Analyzing and Vizualising Data

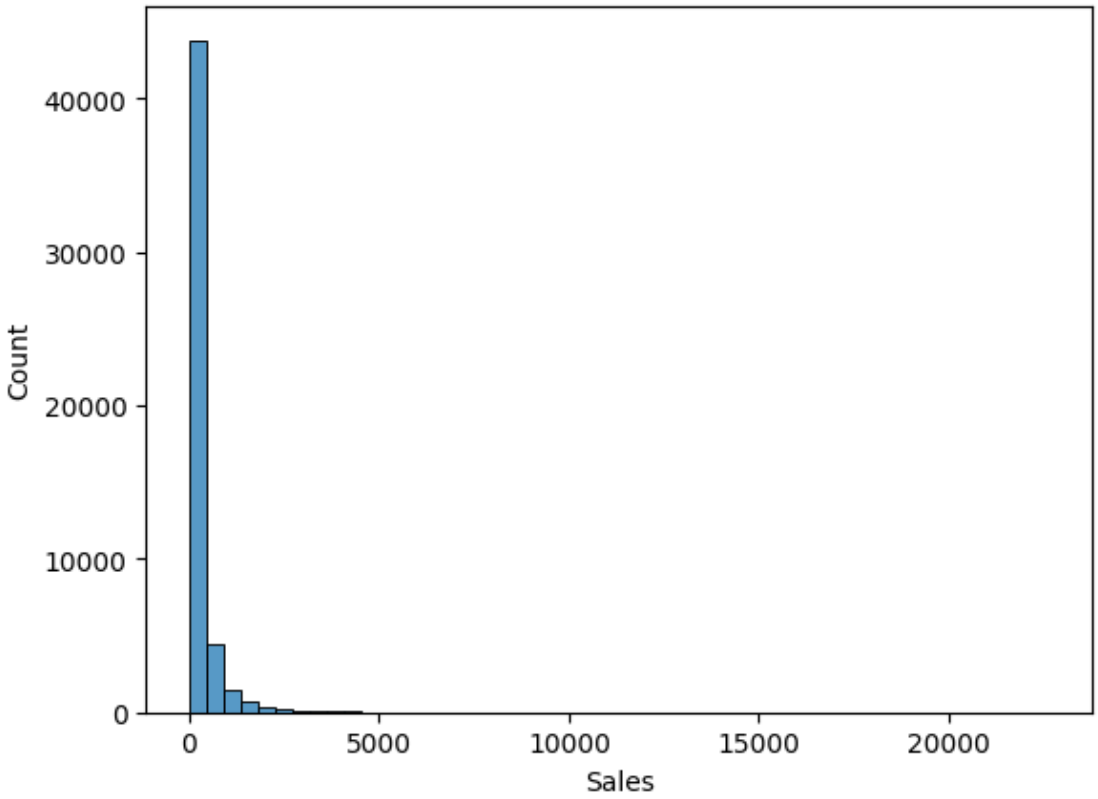
[46]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Category        51290 non-null  object
 1   City            51290 non-null  object
 2   Country         51290 non-null  object
 3   Customer_ID     51290 non-null  object
 4   Customer_Name   51290 non-null  object
 5   Discount        51290 non-null  float64
 6   Market          51290 non-null  object
```

```
 7   Order_Date      51290 non-null  object
 8   Order_ID        51290 non-null  object
 9   Order_Priority  51290 non-null  object
 10  Product_ID      51290 non-null  object
 11  Product_Name    51290 non-null  object
 12  Profit          51290 non-null  float64
 13  Quantity        51290 non-null  int64
 14  Region          51290 non-null  object
 15  Row_ID          51290 non-null  int64
 16  Sales           51290 non-null  int64
 17  Segment         51290 non-null  object
 18  Ship_Date       51290 non-null  object
 19  Ship_Mode       51290 non-null  object
 20  Shipping_Cost   51290 non-null  float64
 21  State           51290 non-null  object
 22  Sub_Category    51290 non-null  object
 23  Year            51290 non-null  int64
 24  Market2         51290 non-null  object
 25  weeknum         51290 non-null  int64
dtypes: float64(3), int64(5), object(18)
memory usage: 10.2+ MB
```

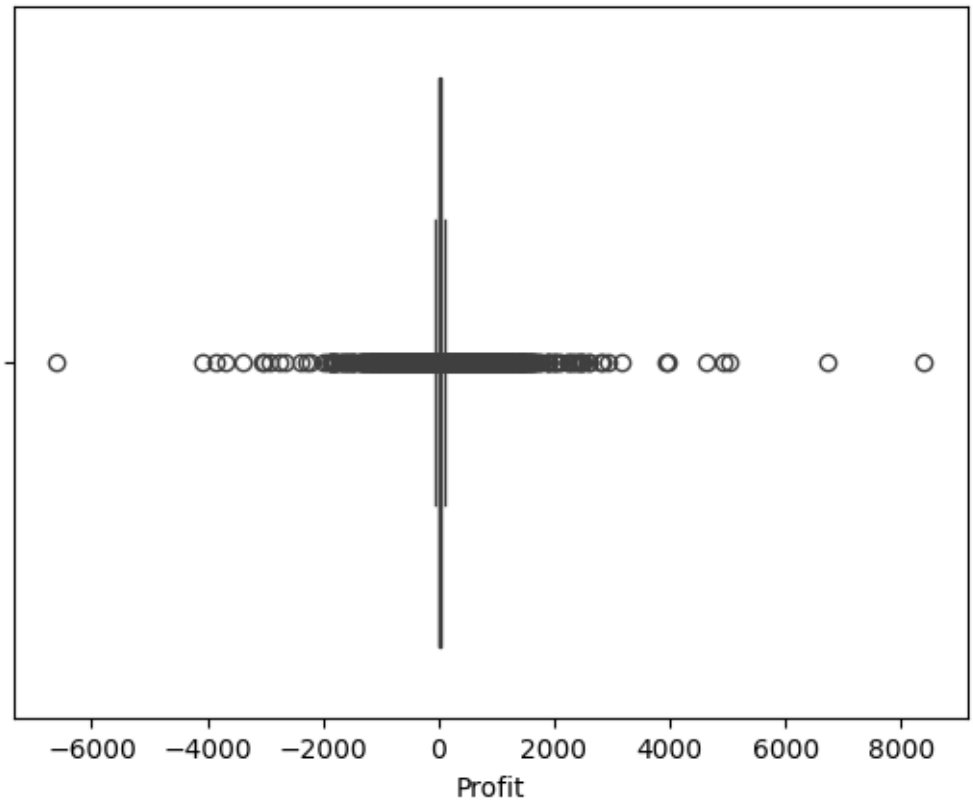
[47]:
```
sns.histplot(df['Sales'], bins=50)
```

[47]:
```
<Axes: xlabel='Sales', ylabel='Count'>
```



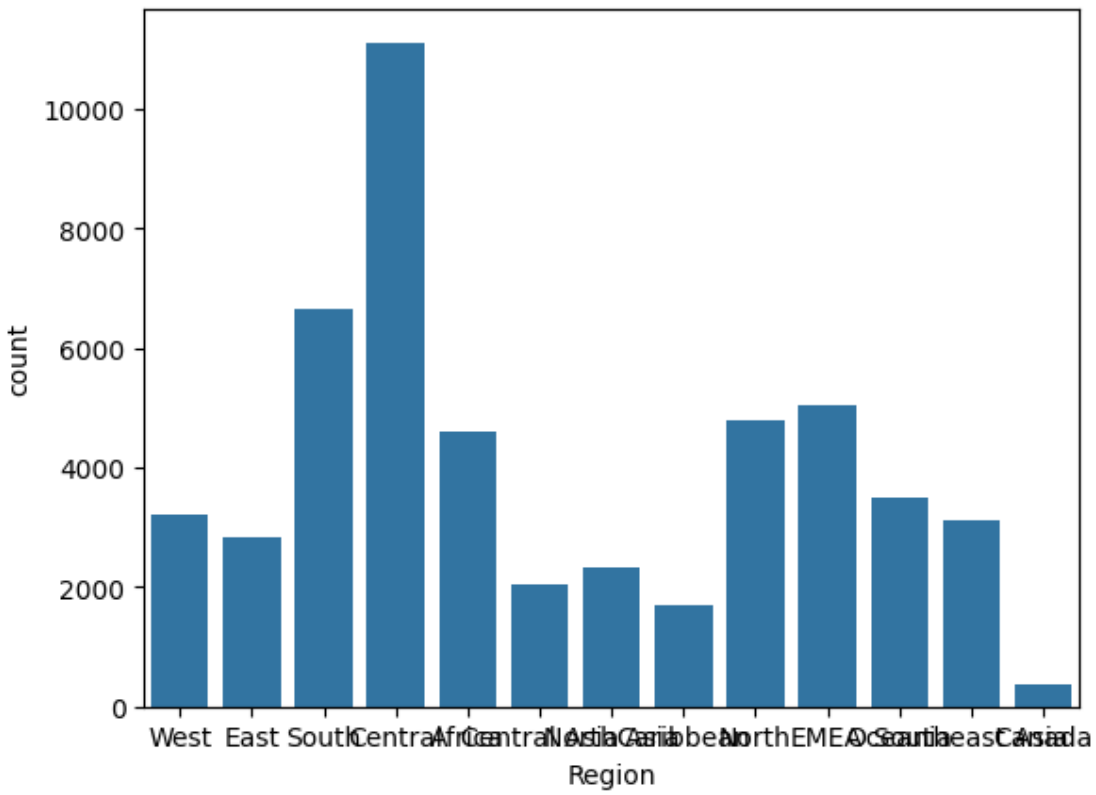
[53]:
```
sns.boxplot(x=df['Profit'])
```

[53]:
```
<Axes: xlabel='Profit'>
```



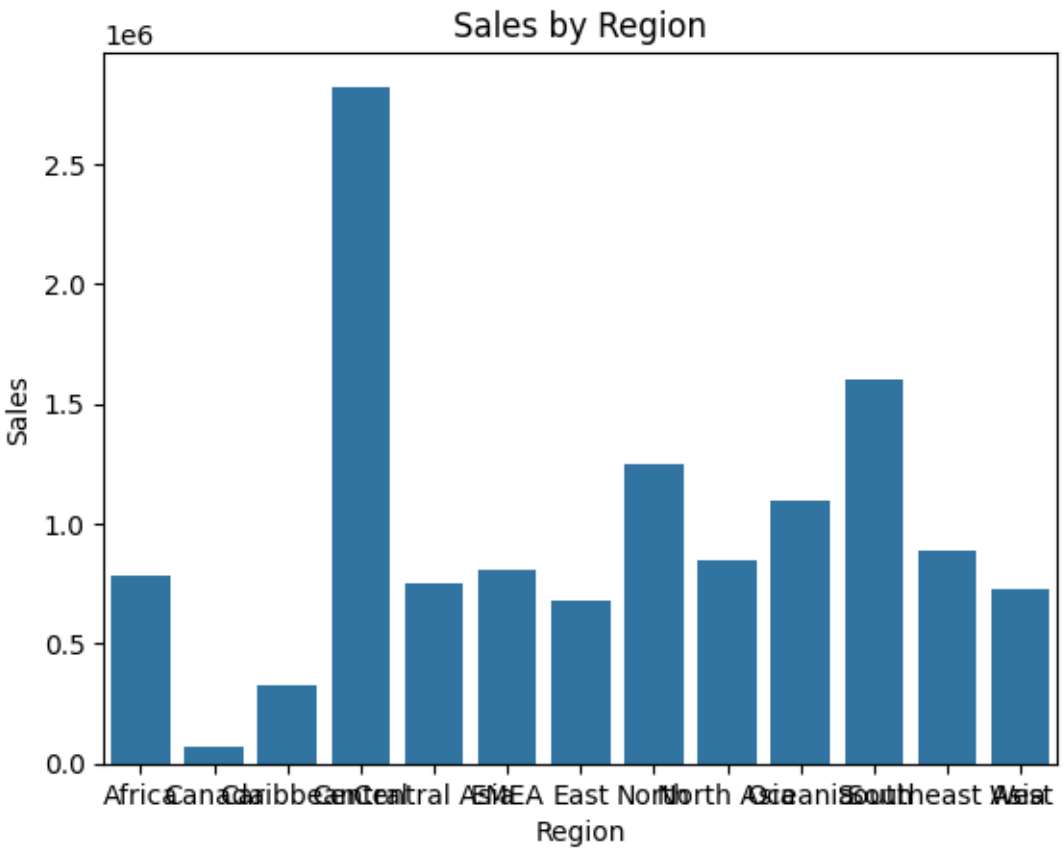
[54]:
```
sns.countplot(x='Region', data=df)
```

[54]:
```
<Axes: xlabel='Region', ylabel='count'>
```



[66]:
```
sales_by_region = df.groupby('Region')['Sales'].sum().reset_index()
sns.barplot(x='Region', y='Sales', data=sales_by_region)
plt.title('Sales by Region')
```

[66]: Text(0.5, 1.0, 'Sales by Region')


Sales by Region
1e6
2.5
2.0
1.5
1.0
0.5
0.0
Sales
Region

[ ]: