In [56]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
df = pd.read_csv("C:/Users/karra/Downloads/SampleSuperstore.csv",delimiter=',')
df.head(15)
```

Out[56]:

	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage
5	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Furniture	Furnishings
6	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Office Supplies	Art
7	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Technology	Phones
8	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Office Supplies	Binders
9	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Office Supplies	Appliances
10	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Furniture	Tables
11	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Technology	Phones
12	Standard Class	Consumer	United States	Concord	North Carolina	28027	South	Office Supplies	Paper
13	Standard Class	Consumer	United States	Seattle	Washington	98103	West	Office Supplies	Binders
14	Standard Class	Home Office	United States	Fort Worth	Texas	76106	Central	Office Supplies	Appliances

In [5]:
```
df.nunique()
```

Out[5]:
```
Ship Mode          4
Segment            3
Country            1
City             531
State             49
Postal Code      631
Region             4
Category           3
Sub-Category      17
Sales           5825
Quantity          14
Discount          12
Profit          7287
dtype: int64
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
Ship Mode       0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Category        0
Sub-Category    0
Sales           0
Quantity        0
Discount        0
Profit          0
dtype: int64
```

In [9]:
```
df.dtypes
```

Out[9]:
```
Ship Mode        object
Segment          object
Country          object
City             object
State            object
Postal Code       int64
Region           object
Category         object
Sub-Category     object
Sales           float64
Quantity          int64
Discount        float64
Profit          float64
dtype: object
```

In [10]:
```
df.shape
```

Out[10]:
```
(9994, 13)
```

In [54]:
```
df.info
```

Out[54]:
```
<bound method DataFrame.info of            Ship Mode    Segment        Country
City       State  \
0       Second Class   Consumer  United States        Henderson    Kentucky
1       Second Class   Consumer  United States        Henderson    Kentucky
2       Second Class  Corporate  United States      Los Angeles  California
3     Standard Class   Consumer  United States  Fort Lauderdale     Florida
4     Standard Class   Consumer  United States  Fort Lauderdale     Florida
...              ...        ...            ...              ...         ...
9989    Second Class   Consumer  United States            Miami     Florida
9990  Standard Class   Consumer  United States       Costa Mesa  California
9991  Standard Class   Consumer  United States       Costa Mesa  California
9992  Standard Class   Consumer  United States       Costa Mesa  California
9993    Second Class   Consumer  United States      Westminster  California

      Postal Code Region         Category Sub-Category     Sales  Quantity  \
0           42420  South        Furniture    Bookcases  261.9600         2
1           42420  South        Furniture       Chairs  731.9400         3
2           90036   West  Office Supplies       Labels   14.6200         2
3           33311  South        Furniture       Tables  957.5775         5
4           33311  South  Office Supplies      Storage   22.3680         2
...           ...    ...              ...          ...       ...       ...
9989        33180  South        Furniture  Furnishings   25.2480         3
9990        92627   West        Furniture  Furnishings   91.9600         2
9991        92627   West       Technology       Phones  258.5760         2
9992        92627   West  Office Supplies        Paper   29.6000         4
9993        92683   West  Office Supplies   Appliances  243.1600         2

      Discount    Profit
0         0.00   41.9136
1         0.00  219.5820
2         0.00    6.8714
3         0.45 -383.0310
4         0.20    2.5164
...        ...       ...
9989      0.20    4.1028
9990      0.00   15.6332
9991      0.20   19.3932
9992      0.00   13.3200
9993      0.00   72.9480

[9994 rows x 13 columns]>
```

In [14]:
```
df.corr()
```

Out[14]:

	Postal Code	Sales	Quantity	Discount	Profit
Postal Code	1.000000	-0.023854	0.012761	0.058443	-0.029961
Sales	-0.023854	1.000000	0.200795	-0.028190	0.479064
Quantity	0.012761	0.200795	1.000000	0.008623	0.066253
Discount	0.058443	-0.028190	0.008623	1.000000	-0.219487
Profit	-0.029961	0.479064	0.066253	-0.219487	1.000000

In [82]:
```
df.describe()
```

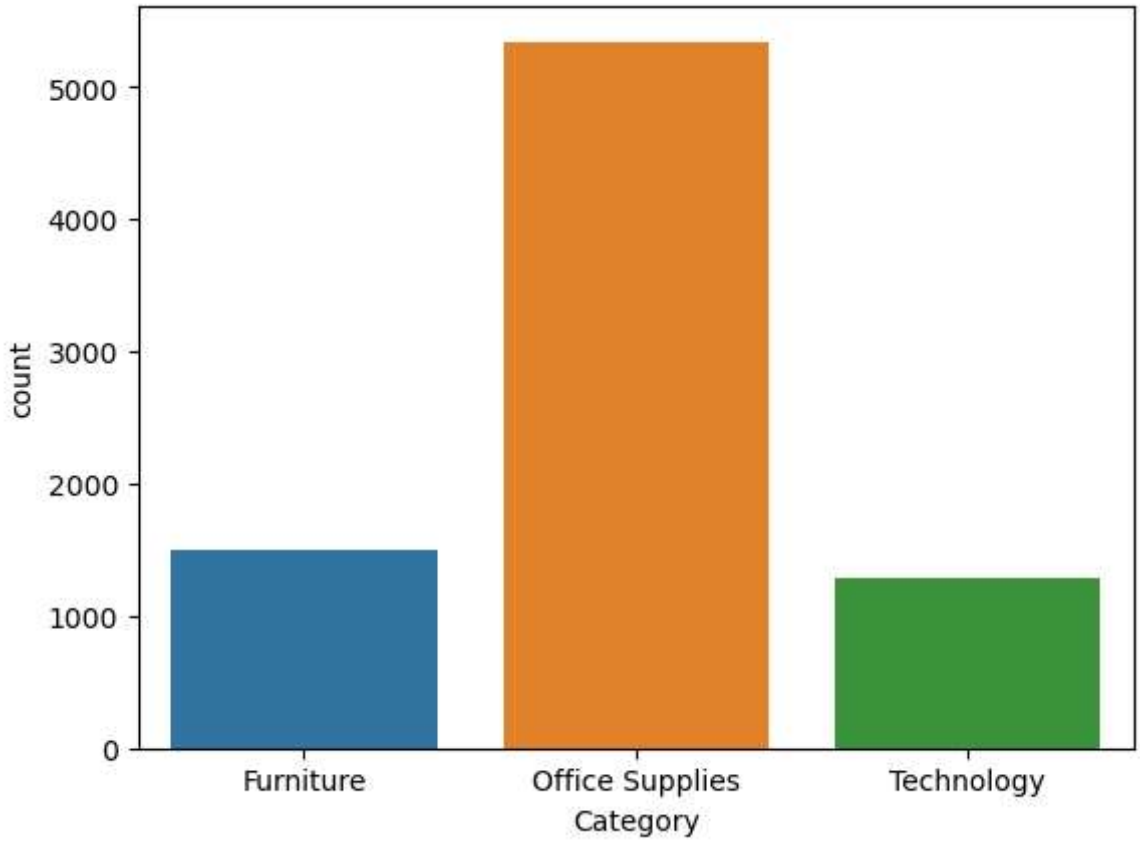
Out[82]:

	Postal Code	Sales	Quantity	Discount	Profit
count	8113.000000	8113.000000	8113.000000	8113.000000	8113.000000
mean	56065.988044	91.051345	3.550475	0.148920	11.604086
std	32077.418326	156.189857	2.076080	0.197695	18.641425
min	1040.000000	0.444000	1.000000	0.000000	-39.637000
25%	24153.000000	13.970000	2.000000	0.000000	2.049200
50%	60505.000000	35.440000	3.000000	0.200000	7.257600
75%	90032.000000	99.870000	5.000000	0.200000	19.034400
max	99301.000000	2803.920000	14.000000	0.800000	70.722000

In [43]:
```
df.count
```

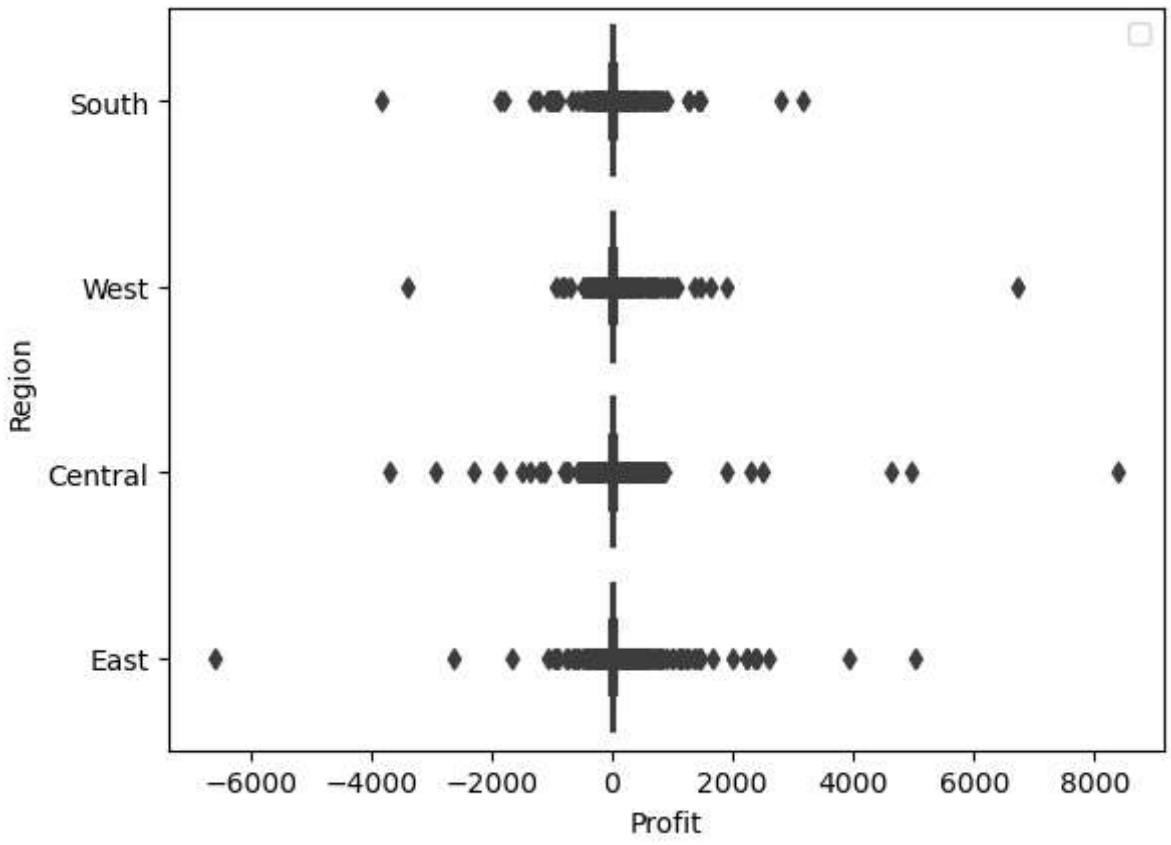
Out[43]:
```
<bound method DataFrame.count of             Country             City       State
Postal Code Region  \
0     United States        Henderson    Kentucky        42420  South
1     United States        Henderson    Kentucky        42420  South
2     United States      Los Angeles  California        90036   West
3     United States  Fort Lauderdale     Florida        33311  South
4     United States  Fort Lauderdale     Florida        33311  South
...             ...              ...         ...          ...    ...
9989  United States            Miami     Florida        33180  South
9990  United States       Costa Mesa  California        92627   West
9991  United States       Costa Mesa  California        92627   West
9992  United States       Costa Mesa  California        92627   West
9993  United States      Westminster  California        92683   West

             Category Sub-Category     Sales  Quantity  Discount    Profit
0           Furniture    Bookcases  261.9600         2      0.00   41.9136
1           Furniture      Chairs   731.9400         3      0.00  219.5820
2     Office Supplies      Labels    14.6200         2      0.00    6.8714
3           Furniture      Tables   957.5775         5      0.45 -383.0310
4     Office Supplies     Storage    22.3680         2      0.20    2.5164
...               ...         ...       ...       ...       ...       ...
9989        Furniture  Furnishings   25.2480         3      0.20    4.1028
9990        Furniture  Furnishings   91.9600         2      0.00   15.6332
9991       Technology       Phones  258.5760         2      0.20   19.3932
9992  Office Supplies        Paper   29.6000         4      0.00   13.3200
9993  Office Supplies   Appliances  243.1600         2      0.00   72.9480

[9994 rows x 11 columns]>
```

In [83]:
```
sns.countplot(x='Category', data=df, )
plt.show()
```



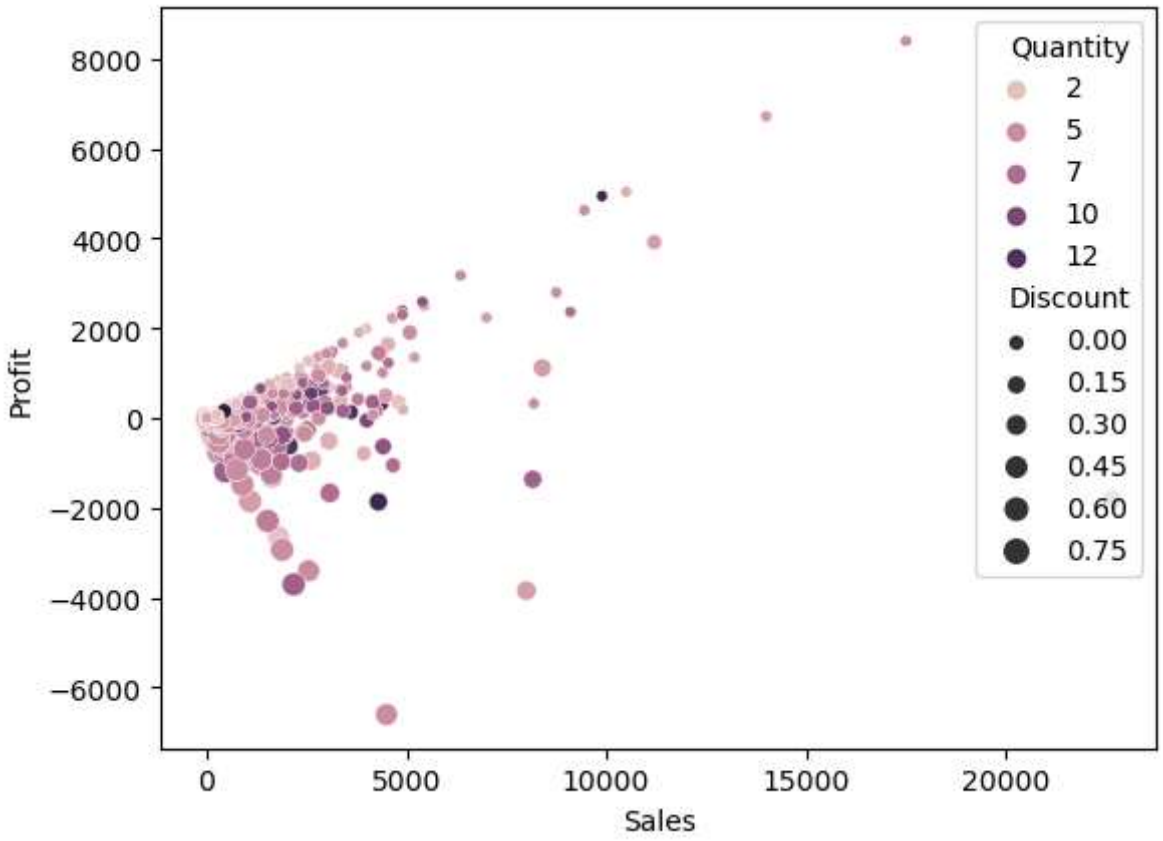
In [66]:
```
sns.boxplot(x='Profit',y='Region',data=df)
plt.legend()
plt.show()
```

```
No artists with labels found to put in legend.  Note that artists whose label star
t with an underscore are ignored when legend() is called with no argument.
```



In [68]:
```
sns.scatterplot( x="Sales",y='Profit', data=df,
                hue='Quantity', size='Discount')
```

Out[68]: <AxesSubplot:xlabel='Sales', ylabel='Profit'>



In [70]:
```
sns.pairplot(df, hue='Category', height=2)
```

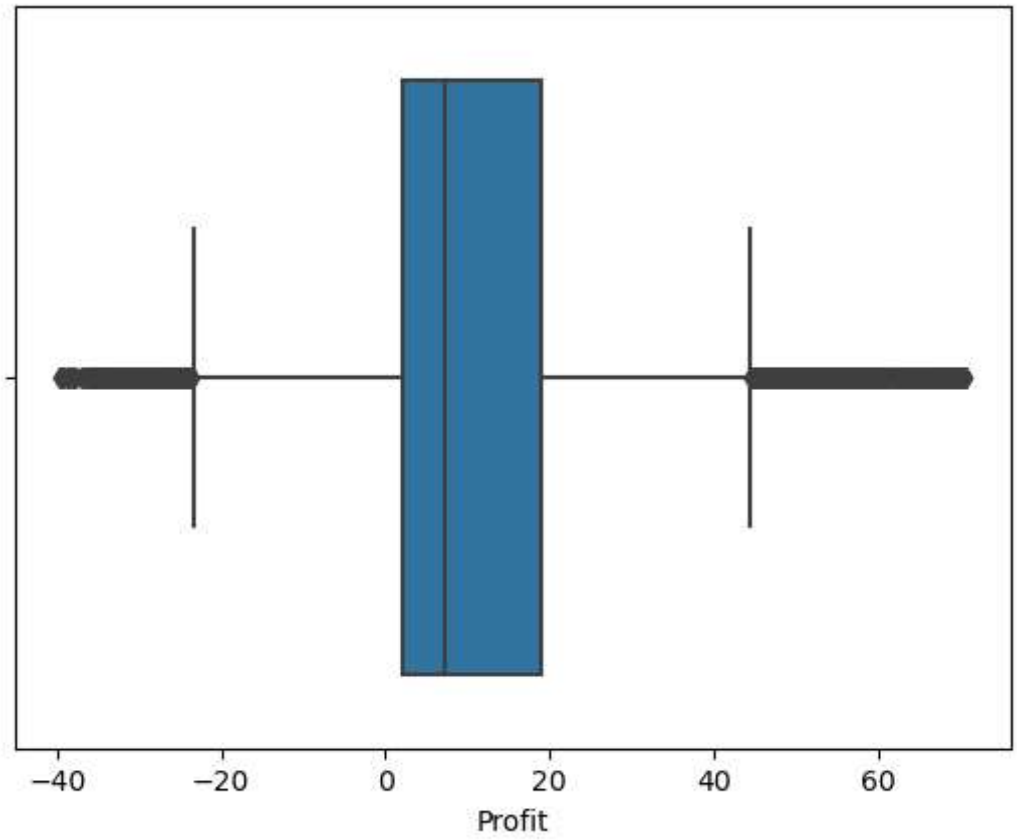
Out[70]: <seaborn.axisgrid.PairGrid at 0x29df9d6c040>

In [77]:
```
import sklearn
from sklearn.datasets import load_boston
Q1 = np.percentile(df['Profit'], 25,interpolation = 'midpoint')
Q3 = np.percentile(df['Profit'], 75,interpolation = 'midpoint')
IQR = Q3 - Q1
print("Old Shape: ", df.shape)
upper = np.where(df['Profit'] >= (Q3+1.5*IQR))
lower = np.where(df['Profit'] <= (Q1-1.5*IQR))
df.drop(upper[0], inplace = True)
df.drop(lower[0], inplace = True)
print("New Shape: ", df.shape)
sns.boxplot(x='Profit', data=df)
```

```
Old Shape:  (9994, 13)
New Shape:  (8113, 13)
```

```
C:\Users\karra\AppData\Local\Temp\ipykernel_2972\1231543111.py:4: DeprecationWarni
ng: the `interpolation=` argument to percentile was renamed to `method=`, which ha
s additional options.
Users of the modes 'nearest', 'lower', 'higher', or 'midpoint' are encouraged to r
eview the method they used. (Deprecated NumPy 1.22)
  Q1 = np.percentile(df['Profit'], 25,interpolation = 'midpoint')
C:\Users\karra\AppData\Local\Temp\ipykernel_2972\1231543111.py:5: DeprecationWarni
ng: the `interpolation=` argument to percentile was renamed to `method=`, which ha
s additional options.
Users of the modes 'nearest', 'lower', 'higher', or 'midpoint' are encouraged to r
eview the method they used. (Deprecated NumPy 1.22)
  Q3 = np.percentile(df['Profit'], 75,interpolation = 'midpoint')
```

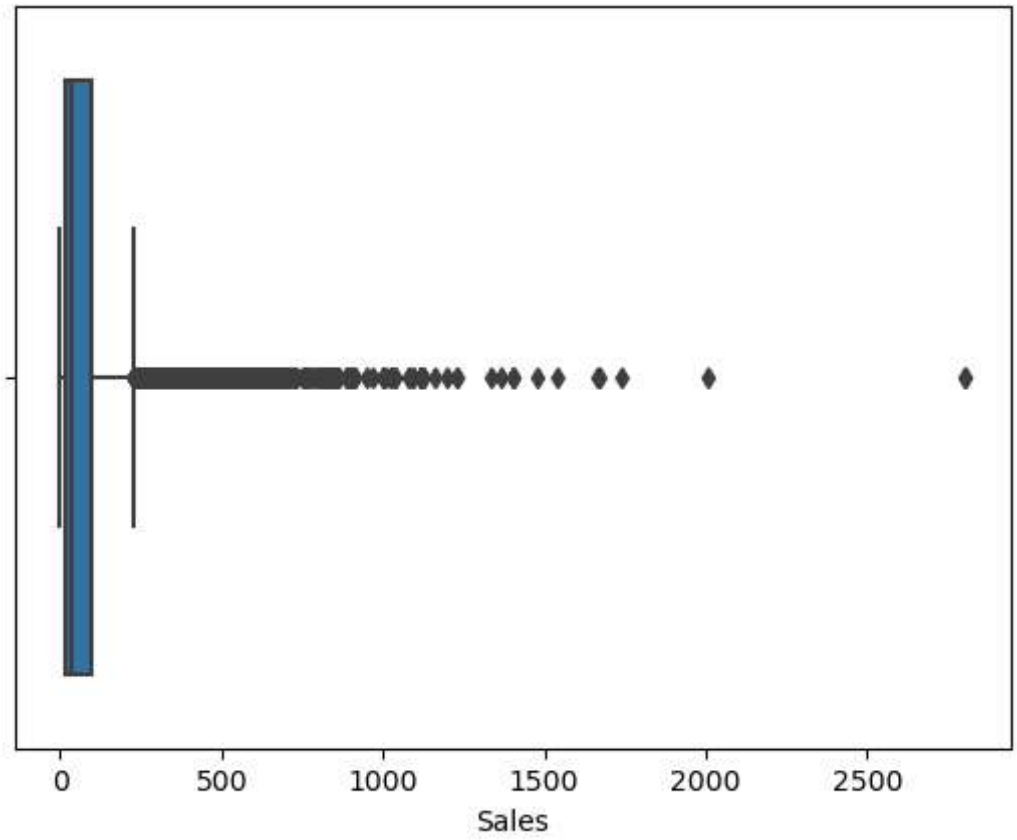
Out[77]:
```
<AxesSubplot:xlabel='Profit'>
```



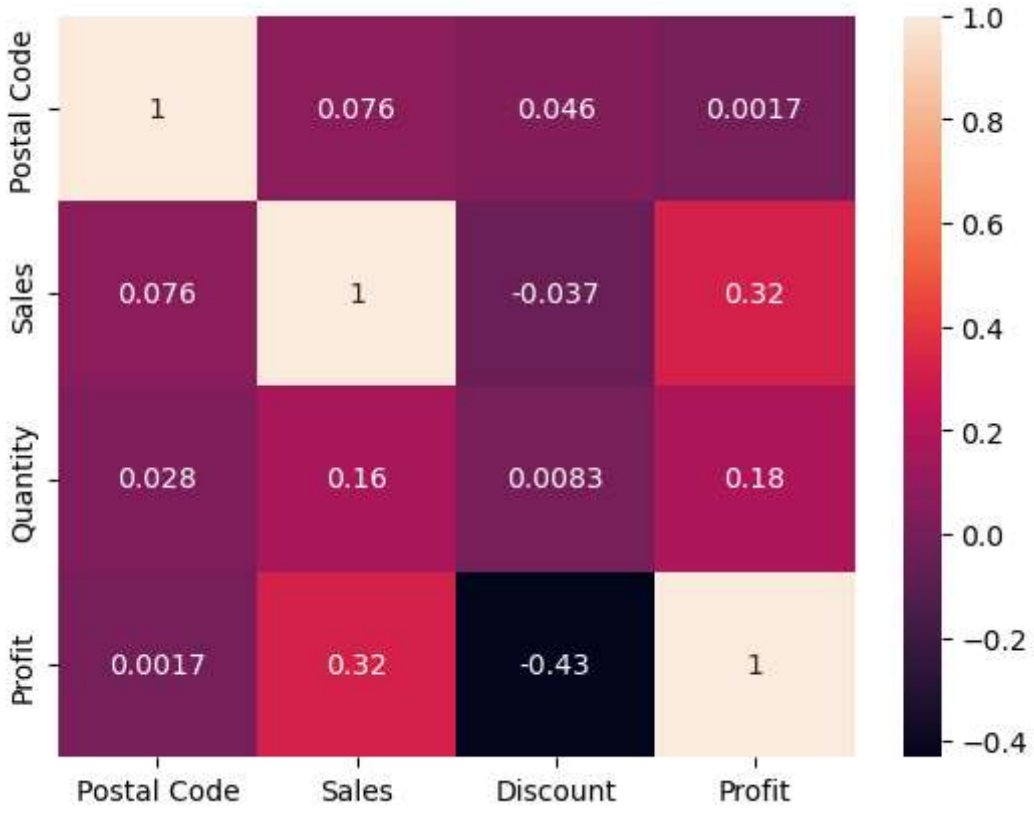
In [81]:
```
import sklearn
from sklearn.datasets import load_boston
Q1 = np.percentile(df['Sales'], 25,interpolation = 'midpoint')
Q3 = np.percentile(df['Sales'], 75,interpolation = 'midpoint')
IQR = Q3 - Q1
print("Old Shape: ", df.shape)
upper = np.where(df['Sales'] >= (Q3+1.5*IQR))
lower = np.where(df['Sales'] <= (Q1-1.5*IQR))
print("New Shape: ", df.shape)
sns.boxplot(x='Sales', data=df)
```

```
Old Shape:  (8113, 13)
New Shape:  (8113, 13)
C:\Users\karra\AppData\Local\Temp\ipykernel_2972\3739968152.py:3: DeprecationWarni
ng: the `interpolation=` argument to percentile was renamed to `method=`, which ha
s additional options.
Users of the modes 'nearest', 'lower', 'higher', or 'midpoint' are encouraged to r
eview the method they used. (Deprecated NumPy 1.22)
  Q1 = np.percentile(df['Sales'], 25,interpolation = 'midpoint')
C:\Users\karra\AppData\Local\Temp\ipykernel_2972\3739968152.py:4: DeprecationWarni
ng: the `interpolation=` argument to percentile was renamed to `method=`, which ha
s additional options.
Users of the modes 'nearest', 'lower', 'higher', or 'midpoint' are encouraged to r
eview the method they used. (Deprecated NumPy 1.22)
  Q3 = np.percentile(df['Sales'], 75,interpolation = 'midpoint')
```

Out[81]:
```
<AxesSubplot:xlabel='Sales'>
```



In [85]:
```
sns.heatmap(df.corr(method='pearson').drop(
  ['Quantity'], axis=1).drop(['Discount'], axis=0),
            annot = True);

plt.show()
```



In [ ]: