```
!pip install matplotlib and seaborn
import warnings
```

```
Requirement already satisfied: matplotlib in c:\users\deepika singh\
anaconda3\lib\site-packages (3.5.1)
Collecting and
  Using cached and-66.0.3.tar.gz (1.2 kB)
Requirement already satisfied: seaborn in c:\users\deepika singh\
anaconda3\lib\site-packages (0.11.2)
Requirement already satisfied: cycler>=0.10 in c:\users\deepika singh\
anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\deepika
singh\anaconda3\lib\site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: pillow>=6.2.0 in c:\users\deepika
singh\anaconda3\lib\site-packages (from matplotlib) (9.0.1)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\
deepika singh\anaconda3\lib\site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: numpy>=1.17 in c:\users\deepika singh\
anaconda3\lib\site-packages (from matplotlib) (1.22.4)
Requirement already satisfied: packaging>=20.0 in c:\users\deepika
singh\anaconda3\lib\site-packages (from matplotlib) (21.3)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\deepika
singh\anaconda3\lib\site-packages (from matplotlib) (3.0.4)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\deepika
singh\anaconda3\lib\site-packages (from matplotlib) (1.3.2)
Requirement already satisfied: pandas>=0.23 in c:\users\deepika singh\
anaconda3\lib\site-packages (from seaborn) (1.4.2)
Requirement already satisfied: scipy>=1.0 in c:\users\deepika singh\
anaconda3\lib\site-packages (from seaborn) (1.13.1)
Requirement already satisfied: pytz>=2020.1 in c:\users\deepika singh\
anaconda3\lib\site-packages (from pandas>=0.23->seaborn) (2025.1)
Requirement already satisfied: six>=1.5 in c:\users\deepika singh\
anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib)
(1.16.0)
Building wheels for collected packages: and
  Building wheel for and (setup.py): started
  Building wheel for and (setup.py): finished with status 'error'
  Running setup.py clean for and
Failed to build and
Installing collected packages: and
    Running setup.py install for and: started
    Running setup.py install for and: finished with status 'error'

  ERROR: Command errored out with exit status 1:
   command: 'C:\Users\Deepika Singh\anaconda3\python.exe' -u -c
'import io, os, sys, setuptools, tokenize; sys.argv[0] = '"'"'C:\\
Users\\Deepika Singh\\AppData\\Local\\Temp\\pip-install-lt89irlu\\
and_0270bcf340af4f098467f5bec7b36e2d\\setup.py'"'"';
__file__='"'"'C:\\Users\\Deepika Singh\\AppData\\Local\\Temp\\pip-
install-lt89irlu\\and_0270bcf340af4f098467f5bec7b36e2d\\
```

```
setup.py'"'"';f = getattr(tokenize, '"'"'open'"'"', open)(__file__) if
os.path.exists(__file__) else io.StringIO('"'"'from setuptools import
setup; setup()'"'"');code = f.read().replace('"'"'\r\n'"'"', '"'"'\
n'"'"');f.close();exec(compile(code, __file__, '"'"'exec'"'"'))'
bdist_wheel -d 'C:\Users\Deepika Singh\AppData\Local\Temp\pip-wheel-
qypyw182'
       cwd: C:\Users\Deepika Singh\AppData\Local\Temp\pip-install-
lt89irlu\and_0270bcf340af4f098467f5bec7b36e2d\
  Complete output (35 lines):
  C:\Users\Deepika Singh\anaconda3\lib\site-packages\setuptools\
_distutils\dist.py:275: UserWarning: Unknown distribution option:
'readme'
    warnings.warn(msg)
  running bdist_wheel
  running build
  C:\Users\Deepika Singh\anaconda3\lib\site-packages\setuptools\
command\install.py:34: SetuptoolsDeprecationWarning: setup.py install
is deprecated. Use build and pip and other standards-based tools.
    warnings.warn(
  installing to build\bdist.win-amd64\wheel
  running install
  Traceback (most recent call last):
    File "<string>", line 1, in <module>
    File "C:\Users\Deepika Singh\AppData\Local\Temp\pip-install-
lt89irlu\and_0270bcf340af4f098467f5bec7b36e2d\setup.py", line 10, in
<module>
      setup(
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\__init__.py", line 87, in setup
      return distutils.core.setup(**attrs)
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\core.py", line 148, in setup
      return run_commands(dist)
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\core.py", line 163, in run_commands
      dist.run_commands()
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\dist.py", line 967, in run_commands
      self.run_command(cmd)
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\dist.py", line 1214, in run_command
      super().run_command(command)
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\dist.py", line 986, in run_command
      cmd_obj.run()
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\wheel\
bdist_wheel.py", line 335, in run
      self.run_command('install')
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
```

```
setuptools\_distutils\cmd.py", line 313, in run_command
      self.distribution.run_command(command)
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\dist.py", line 1214, in run_command
      super().run_command(command)
    File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\dist.py", line 986, in run_command
      cmd_obj.run()
    File "C:\Users\Deepika Singh\AppData\Local\Temp\pip-install-
lt89irlu\and_0270bcf340af4f098467f5bec7b36e2d\setup.py", line 7, in
run
      raise RuntimeError("You are trying to install a stub package
and. Maybe you are using the wrong pypi?")
  RuntimeError: You are trying to install a stub package and. Maybe
you are using the wrong pypi?
  ----------------------------------------
  ERROR: Failed building wheel for and
    ERROR: Command errored out with exit status 1:
     command: 'C:\Users\Deepika Singh\anaconda3\python.exe' -u -c
'import io, os, sys, setuptools, tokenize; sys.argv[0] = '"'"'C:\\
Users\\Deepika Singh\\AppData\\Local\\Temp\\pip-install-lt89irlu\\
and_0270bcf340af4f098467f5bec7b36e2d\\setup.py'"'"';
__file__='"'"'C:\\Users\\Deepika Singh\\AppData\\Local\\Temp\\pip-
install-lt89irlu\\and_0270bcf340af4f098467f5bec7b36e2d\\
setup.py'"'"';f = getattr(tokenize, '"'"'open'"'"', open)(__file__) if
os.path.exists(__file__) else io.StringIO('"'"'from setuptools import
setup; setup()'"'"');code = f.read().replace('"'"'\r\n'"'"', '"'"'\
n'"'"');f.close();exec(compile(code, __file__, '"'"'exec'"'"'))'
install --record 'C:\Users\Deepika Singh\AppData\Local\Temp\pip-
record-vqo4pxlr\install-record.txt' --single-version-externally-
managed --compile --install-headers 'C:\Users\Deepika Singh\anaconda3\
Include\and'
         cwd: C:\Users\Deepika Singh\AppData\Local\Temp\pip-install-
lt89irlu\and_0270bcf340af4f098467f5bec7b36e2d\
    Complete output (24 lines):
    C:\Users\Deepika Singh\anaconda3\lib\site-packages\setuptools\
_distutils\dist.py:275: UserWarning: Unknown distribution option:
'readme'
      warnings.warn(msg)
    running install
    C:\Users\Deepika Singh\anaconda3\lib\site-packages\setuptools\
command\install.py:34: SetuptoolsDeprecationWarning: setup.py install
is deprecated. Use build and pip and other standards-based tools.
      warnings.warn(
    Traceback (most recent call last):
      File "<string>", line 1, in <module>
      File "C:\Users\Deepika Singh\AppData\Local\Temp\pip-install-
lt89irlu\and_0270bcf340af4f098467f5bec7b36e2d\setup.py", line 10, in
<module>
```

```
        setup(
      File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\__init__.py", line 87, in setup
        return distutils.core.setup(**attrs)
      File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\core.py", line 148, in setup
        return run_commands(dist)
      File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\core.py", line 163, in run_commands
        dist.run_commands()
      File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\dist.py", line 967, in run_commands
        self.run_command(cmd)
      File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\dist.py", line 1214, in run_command
        super().run_command(command)
      File "C:\Users\Deepika Singh\anaconda3\lib\site-packages\
setuptools\_distutils\dist.py", line 986, in run_command
        cmd_obj.run()
      File "C:\Users\Deepika Singh\AppData\Local\Temp\pip-install-
lt89irlu\and_0270bcf340af4f098467f5bec7b36e2d\setup.py", line 7, in
run
        raise RuntimeError("You are trying to install a stub package
and. Maybe you are using the wrong pypi?")
    RuntimeError: You are trying to install a stub package and. Maybe
you are using the wrong pypi?
    ----------------------------------------
ERROR: Command errored out with exit status 1: 'C:\Users\Deepika
Singh\anaconda3\python.exe' -u -c 'import io, os, sys, setuptools,
tokenize; sys.argv[0] = '"'"'C:\\Users\\Deepika Singh\\AppData\\
Local\\Temp\\pip-install-lt89irlu\\
and_0270bcf340af4f098467f5bec7b36e2d\\setup.py'"'"';
__file__='"'"'C:\\Users\\Deepika Singh\\AppData\\Local\\Temp\\pip-
install-lt89irlu\\and_0270bcf340af4f098467f5bec7b36e2d\\
setup.py'"'"';f = getattr(tokenize, '"'"'open'"'"', open)(__file__) if
os.path.exists(__file__) else io.StringIO('"'"'from setuptools import
setup; setup()'"'"');code = f.read().replace('"'"'\r\n'"'"', '"'"'\
n'"'"');f.close();exec(compile(code, __file__, '"'"'exec'"'"'))'
install --record 'C:\Users\Deepika Singh\AppData\Local\Temp\pip-
record-vqo4pxlr\install-record.txt' --single-version-externally-
managed --compile --install-headers 'C:\Users\Deepika Singh\anaconda3\
Include\and' Check the logs for full command output.
```

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

data = {
    'emp_id':[101, 102, 103, 104, 105, 106, 107, 108, 109, 110],
```

```
    'emp_name':['Maria','Kumar','bhwani',
None,'Viji','Savitha','Sidharth','Soumya','usha','Ramar'],
    'Age': [25, 30, None, 40, 35, 29, 50, 25, -5, 120],
    'Salary': [50000, 60000, 70000, None, 45000, 55000, None, 50000,
48000, 75000],
    'City': ['Chennai','Madurai','Rameshwaram','Covai','new
delhi','Hyderabad',None,'Kolkata','Trivandrum', 'varanasi'],
}

df =  pd.DataFrame(data)
df.head()
```

	emp_id	emp_name	Age	Salary	City
0	101	Maria	25.0	50000.0	Chennai
1	102	Kumar	30.0	60000.0	Madurai
2	103	bhwani	NaN	70000.0	Rameshwaram
3	104	None	40.0	NaN	Covai
4	105	Viji	35.0	45000.0	new delhi

```
df['emp_name'].str.strip()
```

```
0       Maria
1       Kumar
2      bhwani
3        None
4        Viji
5     Savitha
6    Sidharth
7      Soumya
8        usha
9       Ramar
Name: emp_name, dtype: object
```

```
df['emp_name'].str.upper()
```

```
0       MARIA
1       KUMAR
2      BHWANI
3        None
4        VIJI
5     SAVITHA
6    SIDHARTH
7      SOUMYA
8        USHA
9       RAMAR
Name: emp_name, dtype: object
```

```
df['emp_name'].str.rstrip()
```

```
0       Maria
1       Kumar
```

```
2      bhwani
3        None
4        Viji
5     Savitha
6    Sidharth
7      Soumya
8        usha
9       Ramar
Name: emp_name, dtype: object
```

```
df['emp_name'].replace('Viji','Vijender',inplace=True)
df
```

	emp_id	emp_name	Age	Salary	City
0	101	Maria	25.0	50000.0	Chennai
1	102	Kumar	30.0	60000.0	Madurai
2	103	bhwani	NaN	70000.0	Rameshwaram
3	104	None	40.0	NaN	Covai
4	105	Vijender	35.0	45000.0	new delhi
5	106	Savitha	29.0	55000.0	Hyderabad
6	107	Sidharth	50.0	NaN	None
7	108	Soumya	25.0	50000.0	Kolkata
8	109	usha	-5.0	48000.0	Trivandrum
9	110	Ramar	120.0	75000.0	varanasi

```
df.isnull().sum()
```

```
emp_id      0
emp_name    1
Age         1
Salary      2
City        1
dtype: int64
```

```
df.isnull()
```

	emp_id	emp_name	Age	Salary	City
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	True	False	False
3	False	True	False	True	False
4	False	False	False	False	False
5	False	False	False	False	False
6	False	False	False	True	True
7	False	False	False	False	False
8	False	False	False	False	False
9	False	False	False	False	False

```
mean=df['Salary'].mean()
```

```
df['Salary'].fillna(mean)
```

```
0    50000.0
1    60000.0
2    70000.0
3    56625.0
4    45000.0
5    55000.0
6    56625.0
7    50000.0
8    48000.0
9    75000.0
Name: Salary, dtype: float64
```

```
age=df['Age'].median()
```

```
df['Age'].fillna(age,inplace=True)
df
```

	emp_id	emp_name	Age	Salary	City
0	101	Maria	25.0	50000.0	Chennai
1	102	Kumar	30.0	60000.0	Madurai
2	103	bhwani	30.0	70000.0	Rameshwaram
3	104	None	40.0	NaN	Covai
4	105	Vijender	35.0	45000.0	new delhi
5	106	Savitha	29.0	55000.0	Hyderabad
6	107	Sidharth	50.0	NaN	None
7	108	Soumya	25.0	50000.0	Kolkata
8	109	usha	-5.0	48000.0	Trivandrum
9	110	Ramar	120.0	75000.0	varanasi

```
df.duplicated().sum()
```

```
0
```

```
df['emp_name'].duplicated().sum()
```

```
0
```

```
df['emp_name'].drop_duplicates()
```

```
0       Maria
1       Kumar
2      bhwani
3        None
4    Vijender
5     Savitha
6    Sidharth
7      Soumya
8        usha
9       Ramar
Name: emp_name, dtype: object
```

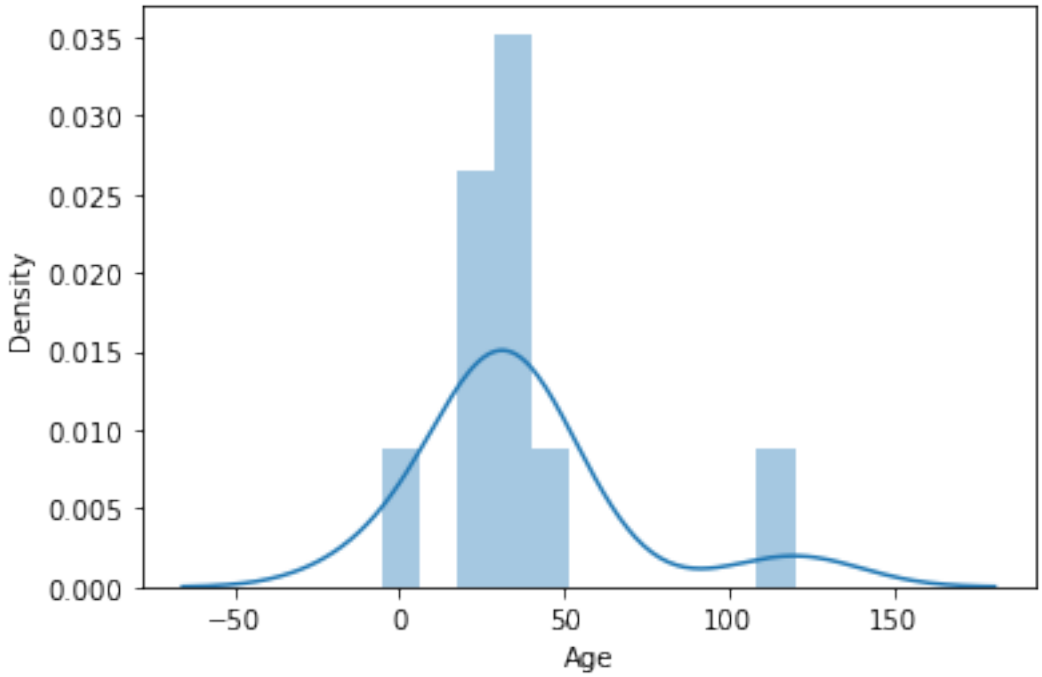
```
print(df['Salary'].std())
print(df['Salary'].mean())
print(df['Salary'].median())
print(df['Salary'].mode())
print(df['Salary'].var())
print(df['Salary'].sum())
print(df['Salary'].max())
print(df['Salary'].min())
print(df['Salary'].count())
```

```
10875.102626773558
56625.0
52500.0
0    50000.0
Name: Salary, dtype: float64
118267857.14285715
453000.0
75000.0
45000.0
8
```

```
sns.distplot(df['Age'])
```

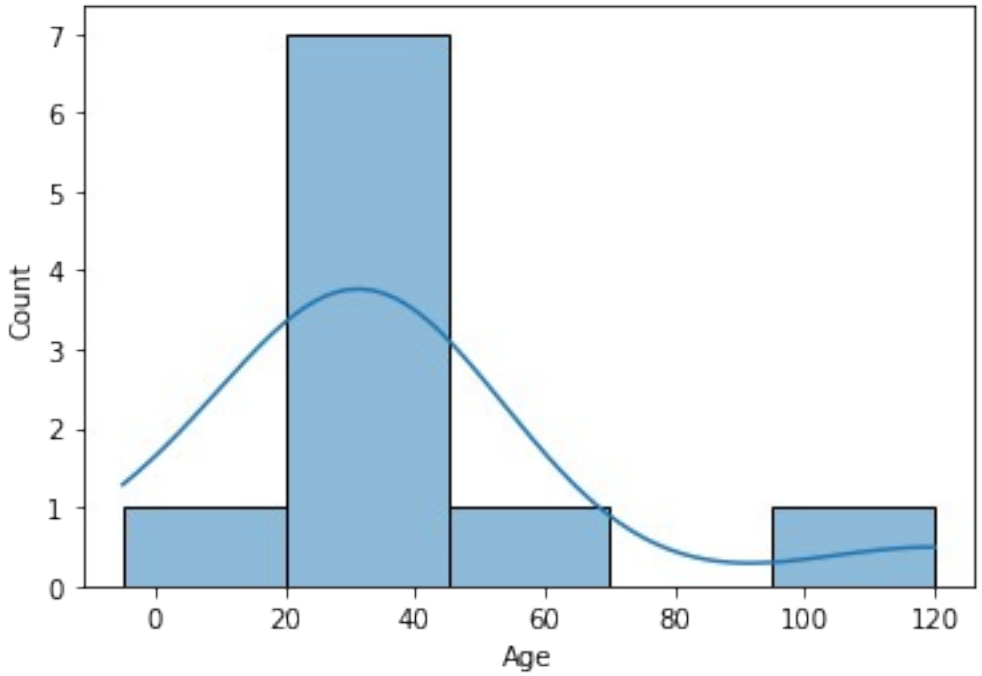
```
C:\Users\Deepika Singh\anaconda3\lib\site-packages\seaborn\
distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



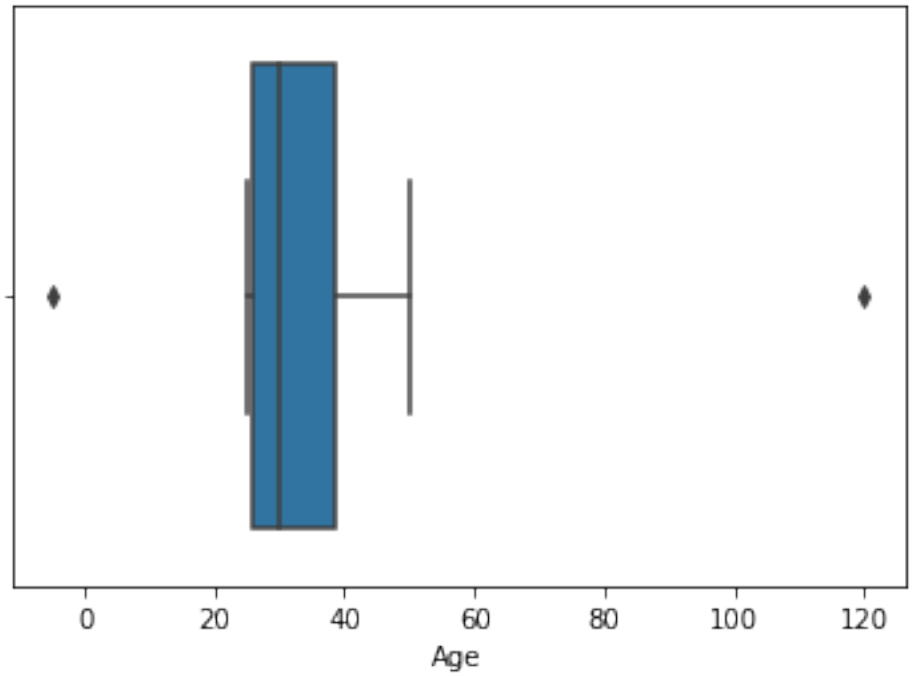

```
sns.histplot(df['Age'],bins=5,kde=True)
```

```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```



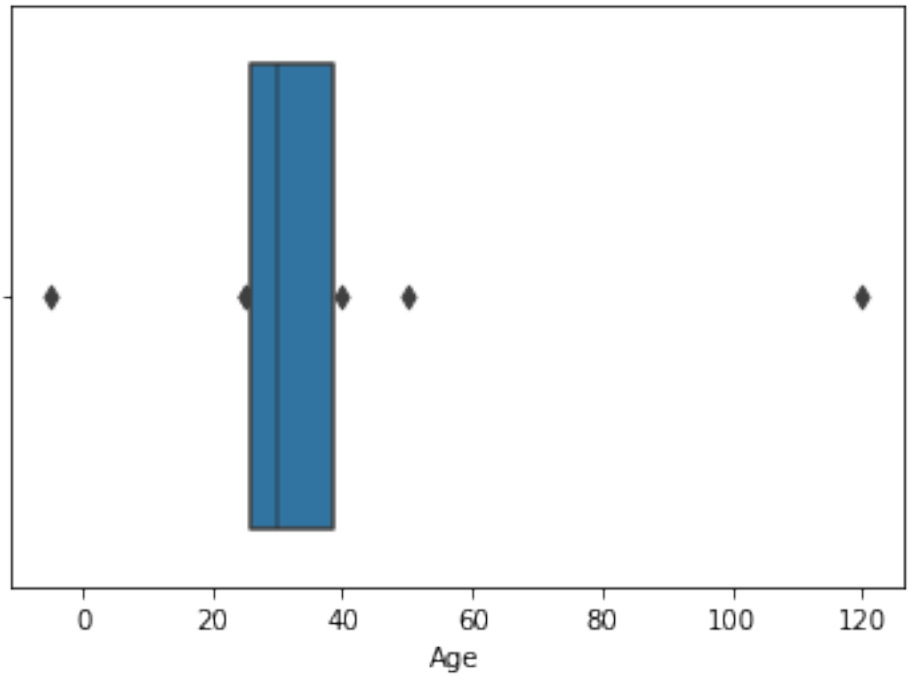

```
sns.boxplot(df['Age'])
plt.show()
```

```
C:\Users\Deepika Singh\anaconda3\lib\site-packages\seaborn\
_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit
keyword will result in an error or misinterpretation.
  warnings.warn(
```




```
sns.boxenplot(df['Age'])
```

```
C:\Users\Deepika Singh\anaconda3\lib\site-packages\seaborn\
_decorators.py:36: FutureWarning: Pass the following variable as a
keyword arg: x. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit
keyword will result in an error or misinterpretation.
  warnings.warn(

<AxesSubplot:xlabel='Age'>
```




```
# load the data
df=pd.read_csv('SampleSuperstore.csv')
#Display the first few rows
df.head()
```

```
        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

   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

   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
```

```
#Checking for missing values
df.isna().sum()
```

```
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
```

```
#Remove rows with missing values
cleaned=df.dropna()
cleaned.head()
```

```
        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

   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

   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
```

```
#Filling missing value with the column mean
df['Sales']=df['Sales'].fillna(df['Sales'].mean())
df['Sales']
```

```
0       261.9600
1       731.9400
2        14.6200
3       957.5775
4        22.3680
         ...
9989     25.2480
9990     91.9600
9991    258.5760
9992     29.6000
9993    243.1600
Name: Sales, Length: 9994, dtype: float64
```

```
#Filling missing value with the column median
df['Discount']=df['Discount'].fillna(df['Sales'].median())
df['Discount']
```

```
0       0.00
1       0.00
2       0.00
3       0.45
4       0.20
       ...
9989    0.20
9990    0.00
9991    0.20
9992    0.00
```

```
9993    0.00
Name: Discount, Length: 9994, dtype: float64
```

```
#Filling missing value with the most frequent value(mode)
df['Category']=df['Category'].fillna(df['Category'].mode()[0])
df['Category']
```

```
0             Furniture
1             Furniture
2       Office Supplies
3             Furniture
4       Office Supplies
              ...
9989          Furniture
9990          Furniture
9991         Technology
9992    Office Supplies
9993    Office Supplies
Name: Category, Length: 9994, dtype: object
```

```
#Filling missing value with a specific value
df['State']=df['State'].fillna('Unknown')
df['State']
```

```
0         Kentucky
1         Kentucky
2       California
3          Florida
4          Florida
           ...
9989       Florida
9990    California
9991    California
9992    California
9993    California
Name: State, Length: 9994, dtype: object
```

```
#Forward fill (propagate previous value)
df.ffill(inplace=True)
df
```

```
           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]
```

```
#Backward fill (propagate next value)
df.bfill(inplace=True)
df
```

```
           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]
```

```
#checking the basic info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Ship Mode     9994 non-null   object 
 1   Segment       9994 non-null   object 
 2   Country       9994 non-null   object 
 3   City          9994 non-null   object 
 4   State         9994 non-null   object 
 5   Postal Code   9994 non-null   int64  
 6   Region        9994 non-null   object 
 7   Category      9994 non-null   object 
 8   Sub-Category  9994 non-null   object 
 9   Sales         9994 non-null   float64
 10  Quantity      9994 non-null   int64  
 11  Discount      9994 non-null   float64
 12  Profit        9994 non-null   float64
```

```
dtypes: float64(3), int64(2), object(8)
memory usage: 1015.1+ KB
```

```
#Check for duplicate rows
df.duplicated()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
9989    False
9990    False
9991    False
9992    False
9993    False
Length: 9994, dtype: bool
```

```
#Display the number of duplicated rows
df.duplicated().sum()
```

```
17
```

```
#Remove duplicate rows
df.drop_duplicates(inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9977 entries, 0 to 9993
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Ship Mode     9977 non-null   object 
 1   Segment       9977 non-null   object 
 2   Country       9977 non-null   object 
 3   City          9977 non-null   object 
 4   State         9977 non-null   object 
 5   Postal Code   9977 non-null   int64  
 6   Region        9977 non-null   object 
 7   Category      9977 non-null   object 
 8   Sub-Category  9977 non-null   object 
 9   Sales         9977 non-null   float64
 10  Quantity      9977 non-null   int64  
 11  Discount      9977 non-null   float64
 12  Profit        9977 non-null   float64
dtypes: float64(3), int64(2), object(8)
memory usage: 1.1+ MB
```

```
#Standardizing category column value
#Convert to lowercase
```

```
df['Category']=df['Category'].str.lower()
df['Category']
```

```
0             furniture
1             furniture
2       office supplies
3             furniture
4       office supplies
             ...
9989          furniture
9990          furniture
9991         technology
9992    office supplies
9993    office supplies
Name: Category, Length: 9977, dtype: object
```

```
#Standardizing category column value
#Convert to uppercase
df['Category']=df['Category'].str.upper()
df['Category']
```

```
0             FURNITURE
1             FURNITURE
2       OFFICE SUPPLIES
3             FURNITURE
4       OFFICE SUPPLIES
             ...
9989          FURNITURE
9990          FURNITURE
9991         TECHNOLOGY
9992    OFFICE SUPPLIES
9993    OFFICE SUPPLIES
Name: Category, Length: 9977, dtype: object
```

```
#Checking the unique values after standardization
df['Category'].unique()
```

```
array(['FURNITURE', 'OFFICE SUPPLIES', 'TECHNOLOGY'], dtype=object)
```

```
#To check datatype
df.dtypes
```

```
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
```

```
#Convert sales column to numeric (if it is incorrectly formatted)
df['Sales']=pd.to_numeric(df['Sales'],errors='coerce')
df.dtypes
```

```
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
```

```
#Replace currency with currency symbol
df['Sales']=df['Sales'].replace({'€': 'EUR','$':'USD'})
df['Sales']
```

```
0       261.9600
1       731.9400
2        14.6200
3       957.5775
4        22.3680
          ...
9989     25.2480
9990     91.9600
9991    258.5760
9992     29.6000
9993    243.1600
Name: Sales, Length: 9977, dtype: float64
```

```
#ensures that the 'Profit' values are positive
df['Profit']=df['Profit'].abs()
df['Profit']
```

```
0        41.9136
1       219.5820
2         6.8714
3       383.0310
4         2.5164
```

```
          ...
9989      4.1028
9990     15.6332
9991     19.3932
9992     13.3200
9993     72.9480
Name: Profit, Length: 9977, dtype: float64
```

```
df[['Sales','Profit']].head()
```

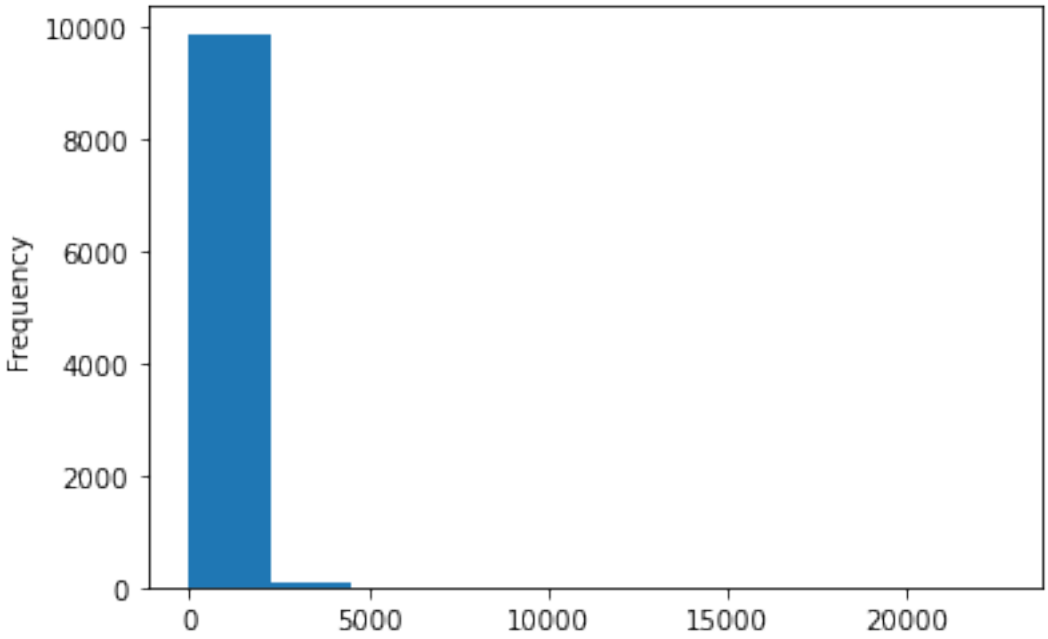
```
      Sales    Profit
0  261.9600   41.9136
1  731.9400  219.5820
2   14.6200    6.8714
3  957.5775  383.0310
4   22.3680    2.5164
```

```
#analyse sales column
df['Sales'].describe()
```

```
count     9977.000000
mean       230.148902
std        623.721409
min          0.444000
25%         17.300000
50%         54.816000
75%        209.970000
max      22638.480000
Name: Sales, dtype: float64
```

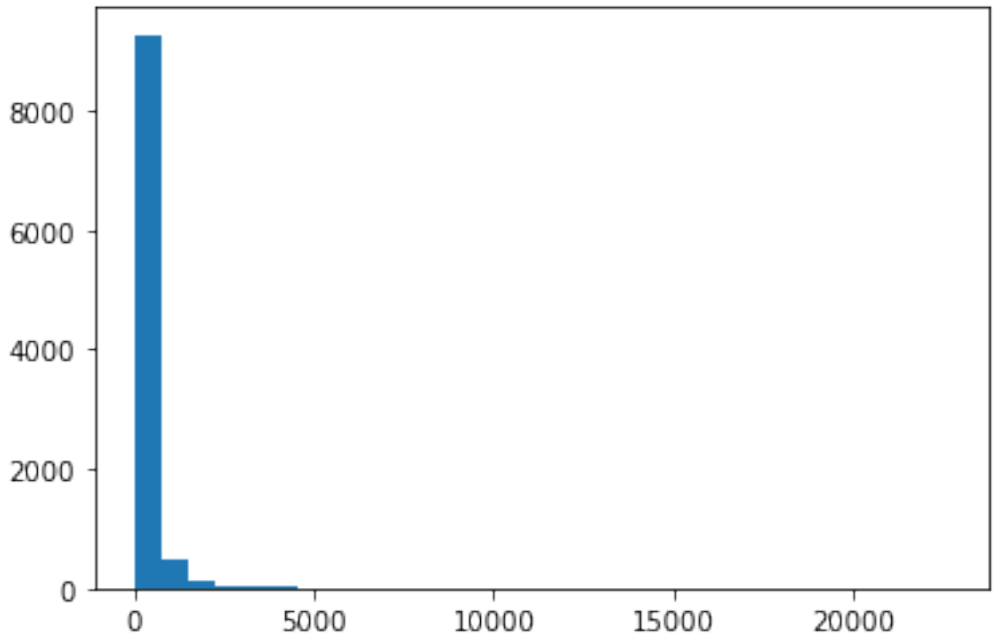
```
df['Sales'].plot(kind='hist')
```

```
<AxesSubplot:ylabel='Frequency'>
```

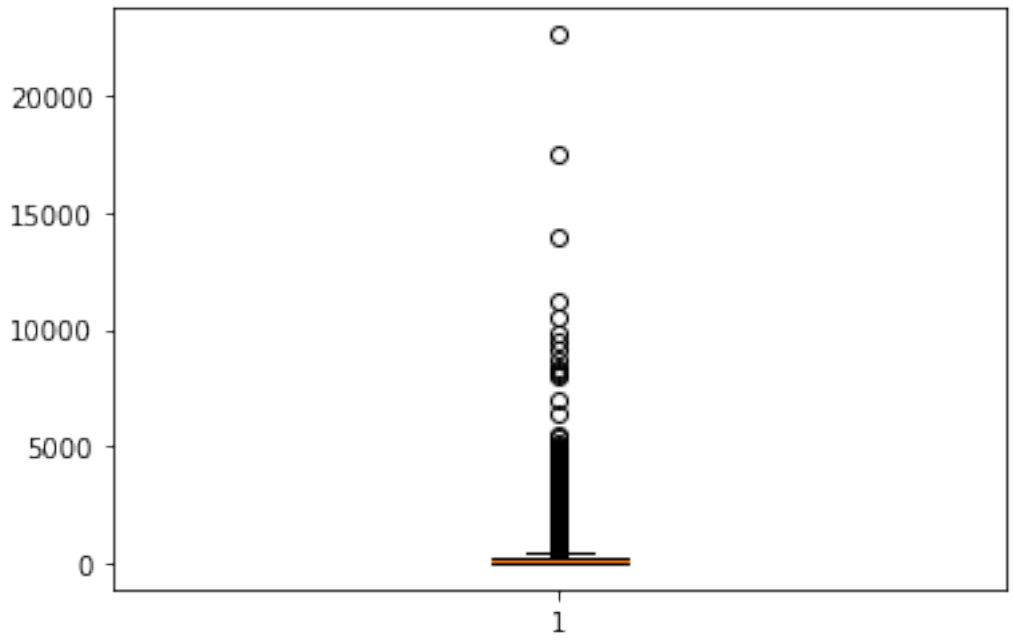



```
plt.hist(df['Sales'],bins=30)
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

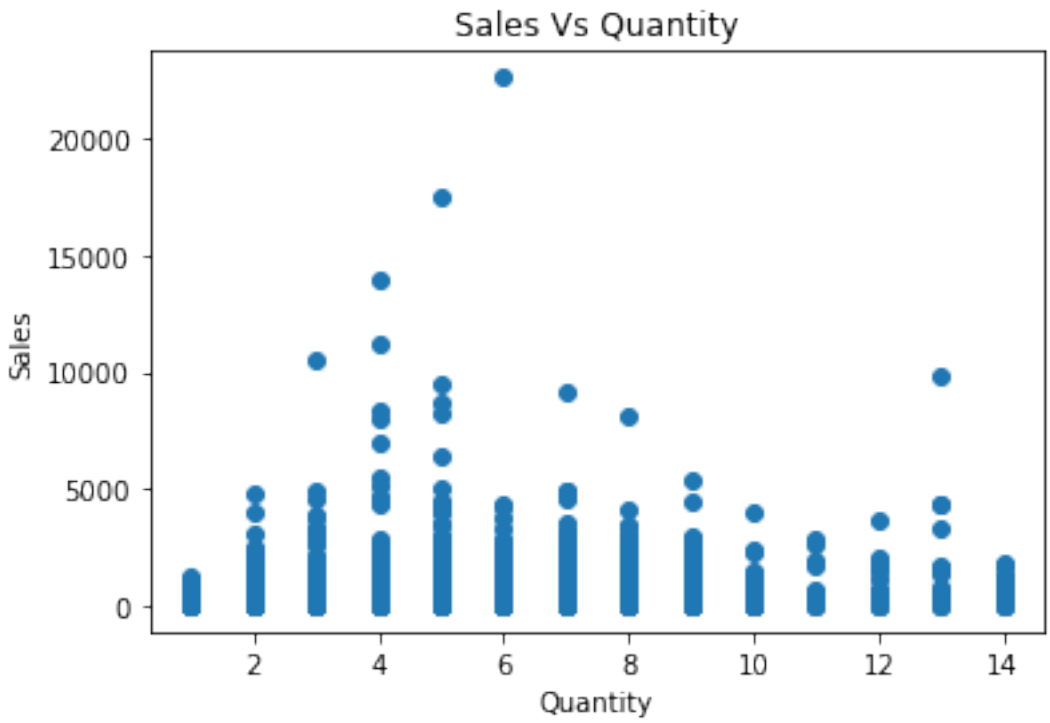



```
plt.boxplot(df['Sales'])
plt.show()
```



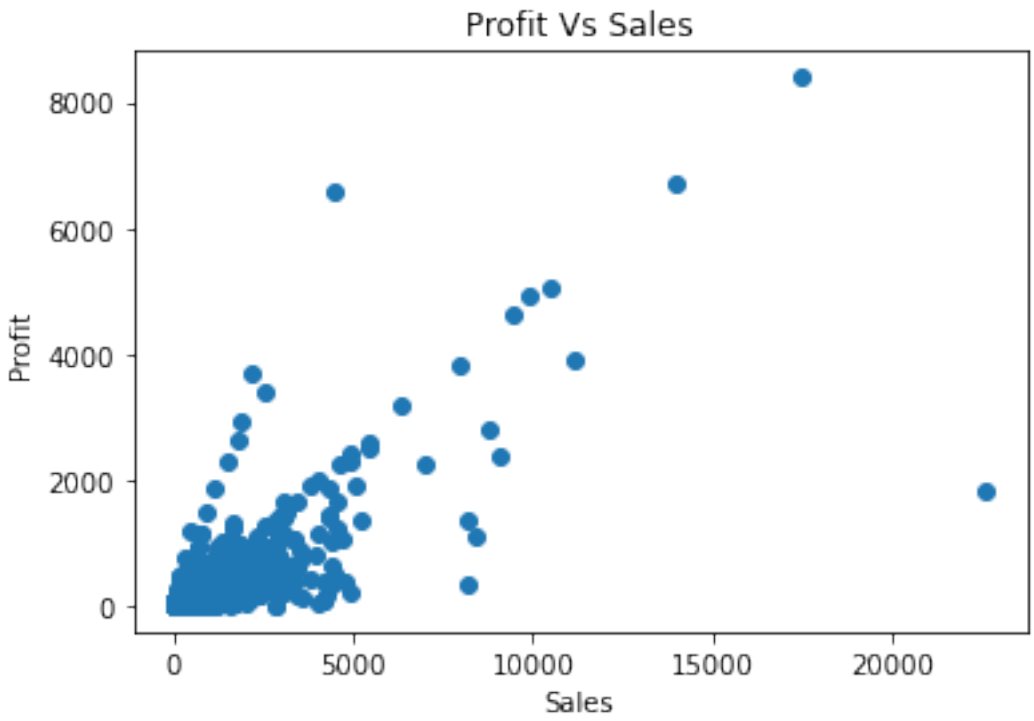

```
#Bivariate Visualisation
# Lets plot a scatter plot to see the relation between sales and
quantity
plt.scatter(df['Quantity'],df['Sales'])
plt.ylabel('Sales')
plt.xlabel('Quantity')
plt.title('Sales Vs Quantity')
```

```
Text(0.5, 1.0, 'Sales Vs Quantity')
```




```
#Sales Vs Profit
plt.scatter(df['Sales'],df['Profit'])
plt.ylabel('Profit')
plt.xlabel('Sales')
plt.title('Profit Vs Sales')
```

```
Text(0.5, 1.0, 'Profit Vs Sales')
```



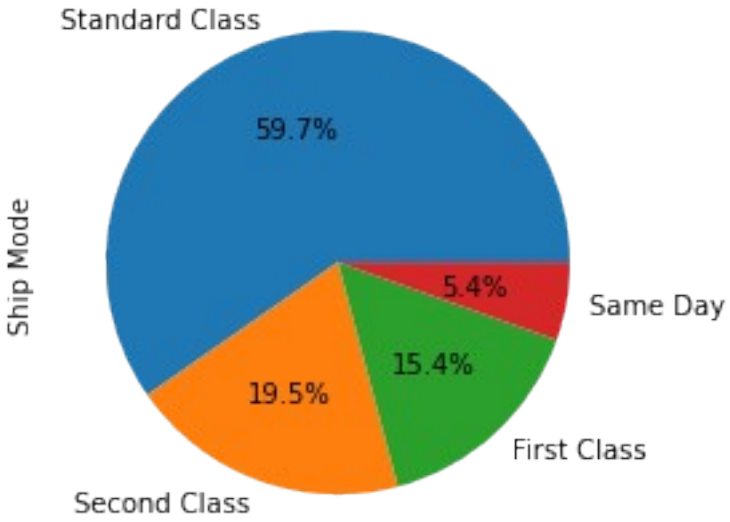

```
df
```

```
           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

[9977 rows x 13 columns]
```

```
df['Ship Mode'].value_counts().plot(kind='pie',autopct='%1.1f%%')
```

```
<AxesSubplot:ylabel='Ship Mode'>
```

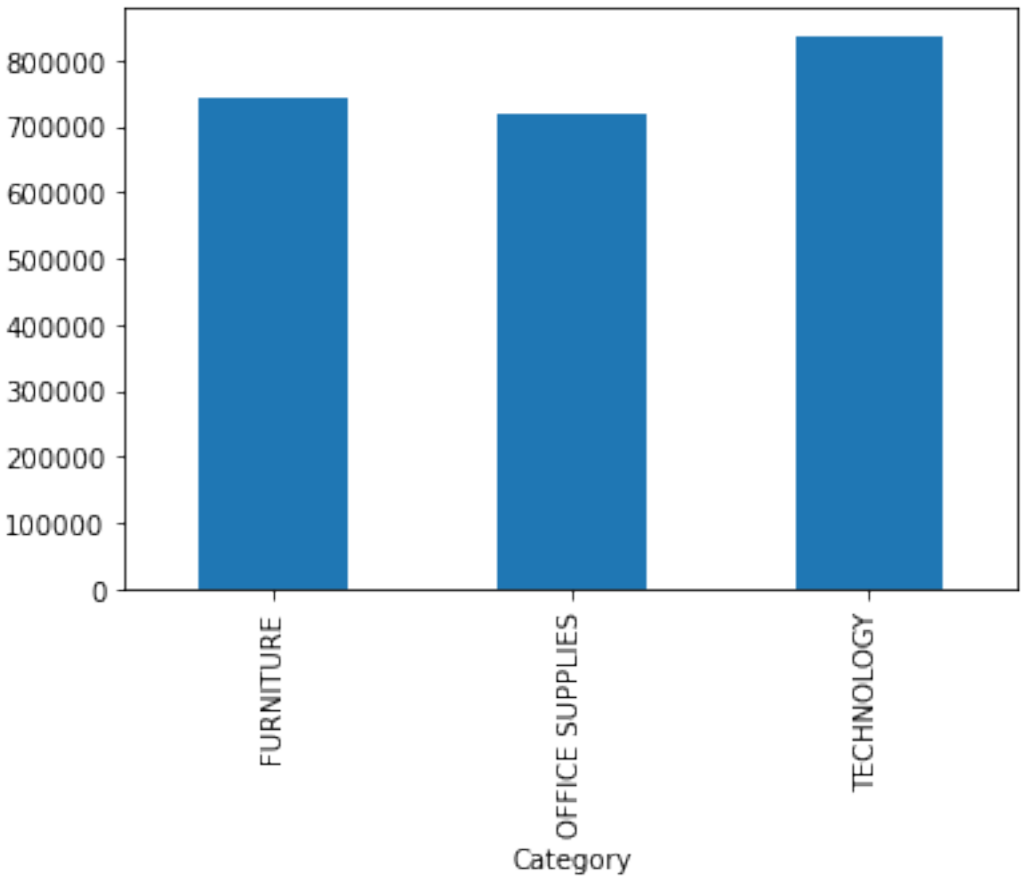



```
#lets say i want to see sales category wise
df['Category'].unique()
```

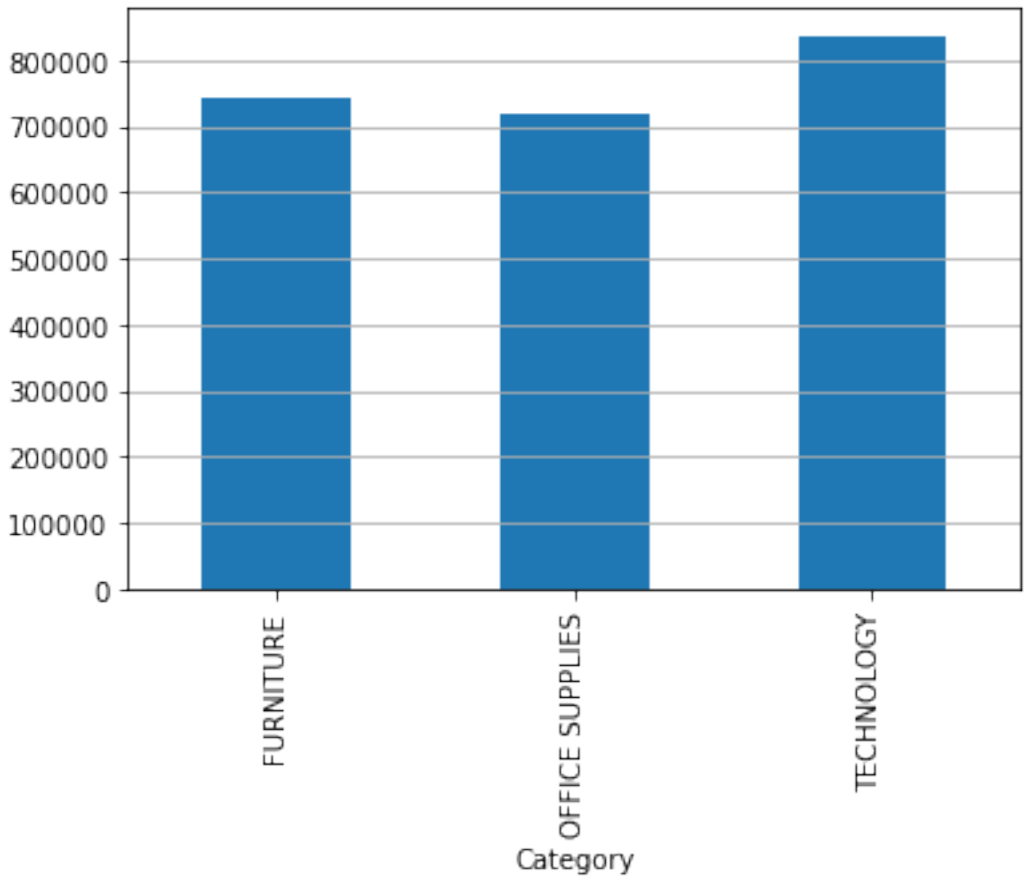
```
array(['FURNITURE', 'OFFICE SUPPLIES', 'TECHNOLOGY'], dtype=object)
```

```
df.groupby('Category')['Sales'].sum().plot(kind='bar')
```

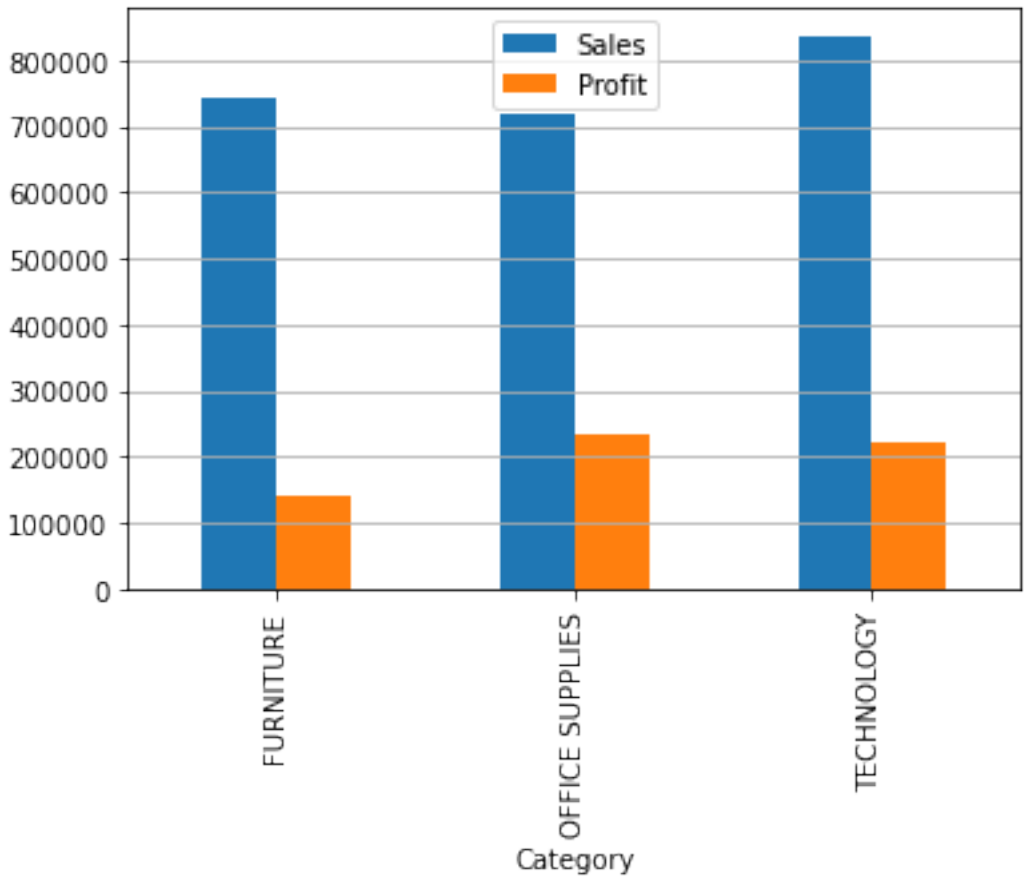
```
<AxesSubplot:xlabel='Category'>
```




```
df.groupby('Category')['Sales'].sum().plot(kind='bar')
plt.grid(True, axis='y')
```



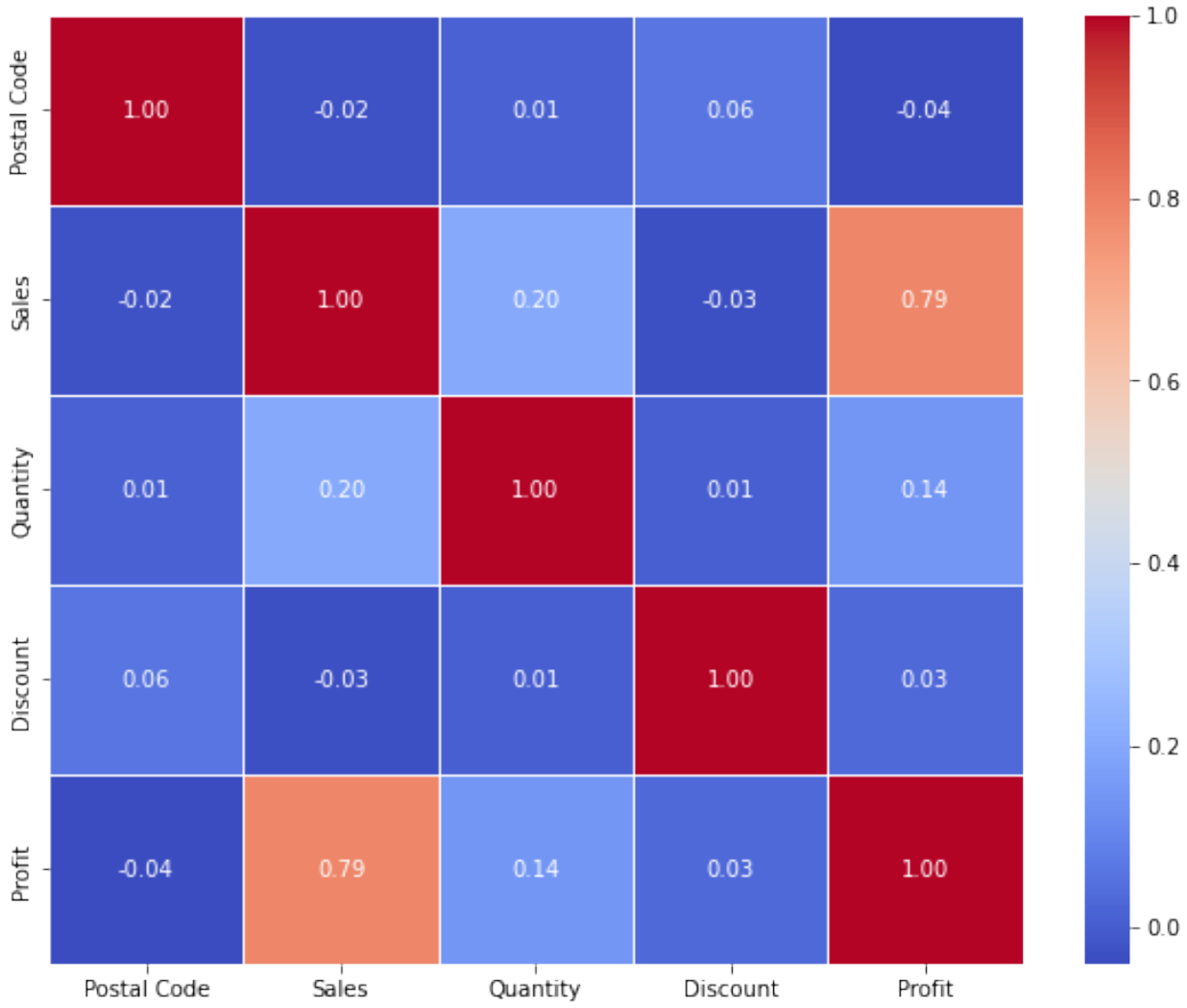

```
df.groupby('Category')[['Sales','Profit']].sum().plot(kind='bar')
plt.grid(True, axis='y')
```




```
df['Sales'].corr(df['Profit'])
```

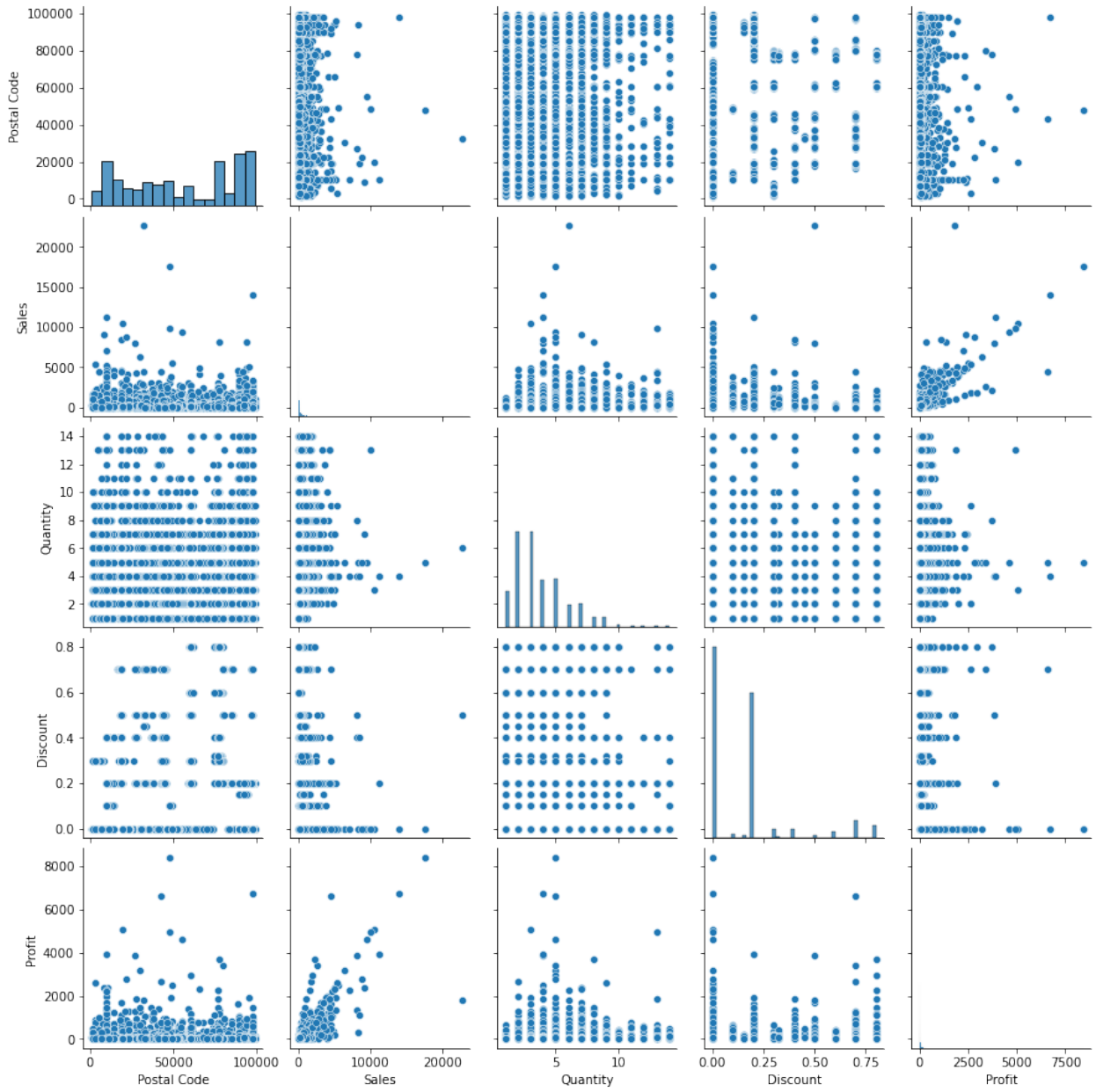
```
0.7874220530442166
```

```
numeric_columns =
df.select_dtypes(include=['number']).columns.tolist()

# Compute correlation matrix
corr = df[numeric_columns].corr()

# Plot heatmap
plt.figure(figsize=(10, 8))  # Optional: set figure size
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt='.2f',
linewidths=0.5)
plt.show()
```




```
#To plot all possible plots
sns.pairplot(df[numeric_columns])
```

```
<seaborn.axisgrid.PairGrid at 0x2370fa8da30>
```




```
numeric_columns
```

```
['Postal Code', 'Sales', 'Quantity', 'Discount', 'Profit']
```