```
import pandas as pd

import numpy as np

df =
pd.read_csv(r'https://raw.githubusercontent.com/YBI-Foundation/Dataset
/main/Big%20Sales%20Data.csv')

df.head()

  Item_Identifier  Item_Weight Item_Fat_Content  Item_Visibility  \
0           FDT36         12.3          Low Fat         0.111448
1           FDT36         12.3          Low Fat         0.111904
2           FDT36         12.3               LF         0.111728
3           FDT36         12.3          Low Fat         0.000000
4           FDP12          9.8          Regular         0.045523

      Item_Type  Item_MRP Outlet_Identifier  Outlet_Establishment_Year
\
0  Baking Goods   33.4874            OUT049                       1999

1  Baking Goods   33.9874            OUT017                       2007

2  Baking Goods   33.9874            OUT018                       2009

3  Baking Goods   34.3874            OUT019                       1985

4  Baking Goods   35.0874            OUT017                       2007


  Outlet_Size Outlet_Location_Type        Outlet_Type
Item_Outlet_Sales
0      Medium               Tier 1  Supermarket Type1
436.608721
1      Medium               Tier 2  Supermarket Type1
443.127721
2      Medium               Tier 3  Supermarket Type2
564.598400
3       Small               Tier 1      Grocery Store
1719.370000
4      Medium               Tier 2  Supermarket Type1
352.874000

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14204 entries, 0 to 14203
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Item_Identifier            14204 non-null  object
```

```
 1   Item_Weight                11815 non-null  float64
 2   Item_Fat_Content           14204 non-null  object
 3   Item_Visibility            14204 non-null  float64
 4   Item_Type                  14204 non-null  object
 5   Item_MRP                   14204 non-null  float64
 6   Outlet_Identifier          14204 non-null  object
 7   Outlet_Establishment_Year  14204 non-null  int64
 8   Outlet_Size                14204 non-null  object
 9   Outlet_Location_Type       14204 non-null  object
 10  Outlet_Type                14204 non-null  object
 11  Item_Outlet_Sales          14204 non-null  float64
dtypes: float64(4), int64(1), object(7)
memory usage: 1.3+ MB
```

```
df.columns
```

```
Index(['Item_Identifier', 'Item_Weight', 'Item_Fat_Content',
'Item_Visibility',
       'Item_Type', 'Item_MRP', 'Outlet_Identifier',
       'Outlet_Establishment_Year', 'Outlet_Size',
'Outlet_Location_Type',
       'Outlet_Type', 'Item_Outlet_Sales'],
      dtype='object')
```

```
df.describe()
```

```
        Item_Weight  Item_Visibility      Item_MRP
Outlet_Establishment_Year  \
count  11815.000000     14204.000000  14204.000000
14204.000000
mean      12.788355         0.065953    141.004977
1997.830681
std        4.654126         0.051459     62.086938
8.371664
min        4.555000         0.000000     31.290000
1985.000000
25%        8.710000         0.027036     94.012000
1987.000000
50%       12.500000         0.054021    142.247000
1999.000000
75%       16.750000         0.094037    185.855600
2004.000000
max       30.000000         0.328391    266.888400
2009.000000

       Item_Outlet_Sales
count       14204.000000
mean         2185.836320
std          1827.479550
min            33.290000
25%           922.135101
```

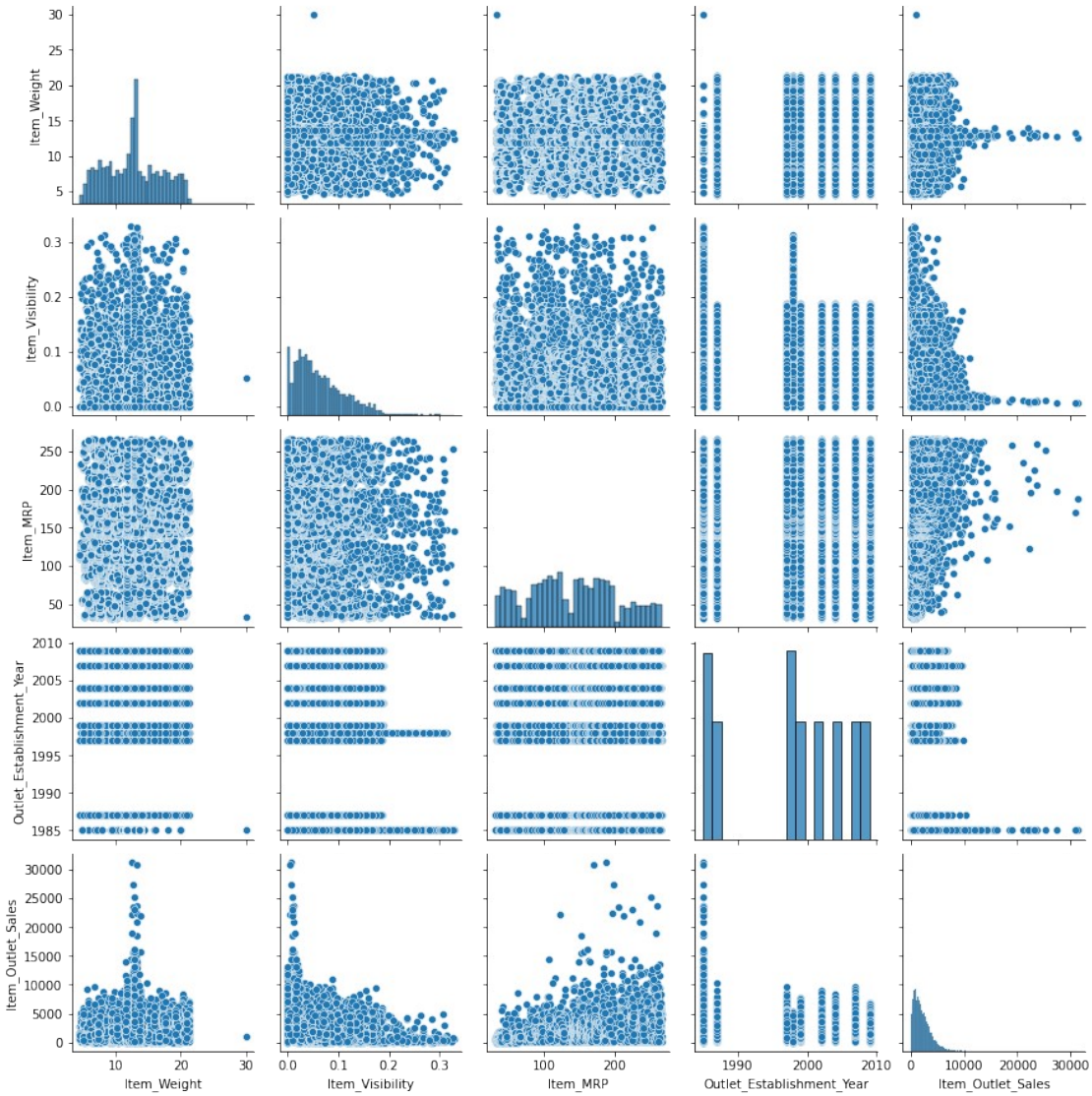
```
50%           1768.287680
75%           2988.110400
max          31224.726950
```

```
df['Item_Weight'].fillna(df.groupby(['Item_Type'])
['Item_Weight'].transform('mean'),inplace=True)
```

```
import seaborn as sns
```

```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f9329ed1e50>
```




```
df['Item_Fat_Content'].value_counts()
```

```
Low Fat    8485
Regular    4824
LF          522
```

```
reg         195
low fat     178
Name: Item_Fat_Content, dtype: int64
```

```
df.replace({'Item_Fat_Content' : {  'LF':'Low Fat', 'reg' :
'Regular' , 'low fat' : 'Low Fat' }},inplace= True)
```

```
df['Item_Fat_Content'].value_counts()
```

```
Low Fat    9185
Regular    5019
Name: Item_Fat_Content, dtype: int64
```

```
df.replace({'Item_Fat_Content' :{ 'Low Fat': 0 , 'Regular' :
1 }},inplace=True)
```

```
df['Item_Type'].value_counts()
```

```
Fruits and Vegetables    2013
Snack Foods              1989
Household                1548
Frozen Foods             1426
Dairy                    1136
Baking Goods             1086
Canned                   1084
Health and Hygiene        858
Meat                      736
Soft Drinks               726
Breads                    416
Hard Drinks               362
Others                    280
Starchy Foods             269
Breakfast                 186
Seafood                    89
Name: Item_Type, dtype: int64
```

```
df.replace({'Item_Type' : { 'Fruits and Vegetables'  : 0,
'Snack Foods'              :0,
'Household'                :1,
'Frozen Foods'             :0,
'Dairy'                    :0,
'Baking Goods'             :0,
'Canned'                   :0,
'Health and Hygiene'       :1,
'Meat'                     :0,
'Soft Drinks'              :0,
'Breads'                   :0,
'Hard Drinks'              :0,
'Others'                   :2,
'Starchy Foods'            :0,
'Breakfast'                :0,
'Seafood'                  :0      } },inplace=True)
```

```
df['Item_Type'].value_counts()
```

```
0    11518
1     2406
2      280
Name: Item_Type, dtype: int64
```

```
df[['Outlet_Identifier']].value_counts()
```

```
Outlet_Identifier
OUT027               1559
OUT013               1553
OUT035               1550
OUT046               1550
OUT049               1550
OUT045               1548
OUT018               1546
OUT017               1543
OUT010                925
OUT019                880
dtype: int64
```

```
df.replace({'Outlet_Identifier' : {
'OUT027':0,
'OUT013':1,
'OUT035':2,
'OUT046':3,
'OUT049':4,
'OUT045':5,
'OUT018':6,
'OUT017':7,
'OUT010':8,
'OUT019':9
}},inplace=True)
```

```
df['Outlet_Identifier'].value_counts()
```

```
0    1559
1    1553
4    1550
3    1550
2    1550
5    1548
6    1546
7    1543
8     925
9     880
Name: Outlet_Identifier, dtype: int64
```

```
df['Outlet_Size'].value_counts()
```

```
Medium    7122
Small     5529
High      1553
Name: Outlet_Size, dtype: int64
```

```
df.replace({'Outlet_Size' : {
'Medium'  :  1,
'Small'   :  0,
'High'    :  2
}},inplace=True)
```

```
df['Outlet_Size'].value_counts()
```

```
1    7122
0    5529
2    1553
Name: Outlet_Size, dtype: int64
```

```
df['Outlet_Location_Type'].value_counts()
```

```
Tier 3    5583
Tier 2    4641
Tier 1    3980
Name: Outlet_Location_Type, dtype: int64
```

```
df.replace({'Outlet_Location_Type' : {
            'Tier 3'  : 0,
            'Tier 2'  :  1,
            'Tier 1' :   2
}},inplace=True)
```

```
df['Outlet_Location_Type'].value_counts()
```

```
0    5583
1    4641
2    3980
Name: Outlet_Location_Type, dtype: int64
```

```
df['Outlet_Type'].value_counts()
```

```
Supermarket Type1    9294
Grocery Store        1805
Supermarket Type3    1559
Supermarket Type2    1546
Name: Outlet_Type, dtype: int64
```

```
df.replace({'Outlet_Type' : {
'Grocery Store'     : 0,
'Supermarket Type1' : 1,
'Supermarket Type3' : 2,
'Supermarket Type2' : 3
}},inplace=True)
```

```
df['Outlet_Type'].value_counts()
```

```
1    9294
0    1805
2    1559
3    1546
Name: Outlet_Type, dtype: int64
```

```
df.head()
```

```
  Item_Identifier  Item_Weight  Item_Fat_Content  Item_Visibility
Item_Type  \
0           FDT36         12.3                 0         0.111448
0
1           FDT36         12.3                 0         0.111904
0
2           FDT36         12.3                 0         0.111728
0
3           FDT36         12.3                 0         0.000000
0
4           FDP12          9.8                 1         0.045523
0

   Item_MRP  Outlet_Identifier  Outlet_Establishment_Year  Outlet_Size
\
0   33.4874                  4                       1999            1

1   33.9874                  7                       2007            1

2   33.9874                  6                       2009            1

3   34.3874                  9                       1985            0

4   35.0874                  7                       2007            1


   Outlet_Location_Type  Outlet_Type  Item_Outlet_Sales
0                     2            1         436.608721
1                     1            1         443.127721
2                     0            3         564.598400
3                     2            0        1719.370000
4                     1            1         352.874000
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14204 entries, 0 to 14203
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Item_Identifier            14204 non-null  object
```

```
 1   Item_Weight                14204 non-null  float64
 2   Item_Fat_Content           14204 non-null  int64
 3   Item_Visibility            14204 non-null  float64
 4   Item_Type                  14204 non-null  int64
 5   Item_MRP                   14204 non-null  float64
 6   Outlet_Identifier          14204 non-null  int64
 7   Outlet_Establishment_Year  14204 non-null  int64
 8   Outlet_Size                14204 non-null  int64
 9   Outlet_Location_Type       14204 non-null  int64
 10  Outlet_Type                14204 non-null  int64
 11  Item_Outlet_Sales          14204 non-null  float64
dtypes: float64(4), int64(7), object(1)
memory usage: 1.3+ MB
```

```
df.shape
```

```
(14204, 12)
```

```
y =df['Item_Outlet_Sales']
```

```
y.shape
```

```
(14204,)
```

```
X = df.drop(['Item_Outlet_Sales','Item_Identifier'],axis=1)
```

```
X
```

```
       Item_Weight  Item_Fat_Content  Item_Visibility  Item_Type
Item_MRP  \
0        12.300000                 0         0.111448          0
33.4874
1        12.300000                 0         0.111904          0
33.9874
2        12.300000                 0         0.111728          0
33.9874
3        12.300000                 0         0.000000          0
34.3874
4         9.800000                 1         0.045523          0
35.0874
...            ...               ...              ...        ...
...
14199    12.800000                 0         0.069606          0
261.9252
14200    12.800000                 0         0.070013          0
262.8252
14201    12.800000                 0         0.069561          0
263.0252
14202    13.659758                 0         0.069282          0
263.5252
14203    12.800000                 0         0.069727          0
263.6252
```

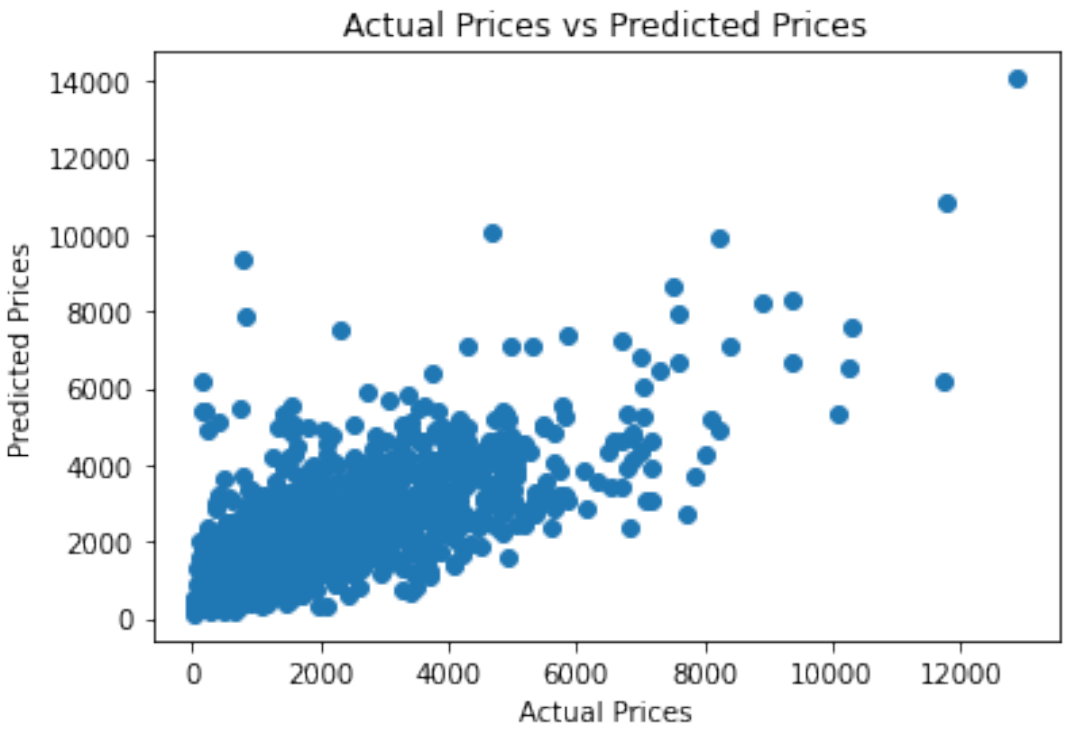
```
       Outlet_Identifier  Outlet_Establishment_Year  Outlet_Size  \
0                      4                       1999            1
1                      7                       2007            1
2                      6                       2009            1
3                      9                       1985            0
4                      7                       2007            1
...                  ...                        ...          ...
14199                  2                       2004            0
14200                  7                       2007            1
14201                  1                       1987            2
14202                  0                       1985            1
14203                  4                       1999            1

       Outlet_Location_Type  Outlet_Type
0                         2            1
1                         1            1
2                         0            3
3                         2            0
4                         1            1
...                     ...          ...
14199                     1            1
14200                     1            1
14201                     0            1
14202                     0            2
14203                     2            1

[14204 rows x 10 columns]
```

```
from sklearn.preprocessing import StandardScaler

sc=StandardScaler()

X_std= df[['Item_Weight',  'Item_Visibility'     ,'Item_MRP',
      'Outlet_Establishment_Year']]

X_std = sc.fit_transform(X_std)

X[['Item_Weight',     'Item_Visibility'     ,'Item_MRP',    'Outlet_Es
tablishment_Year']] =pd.DataFrame(X_std , columns = [['Item_Weight',
      'Item_Visibility'     ,'Item_MRP',    'Outlet_Establishment_Year'
]] )

X
```

```
       Item_Weight  Item_Fat_Content  Item_Visibility  Item_Type
Item_MRP  \
0        -0.115417                 0         0.884136          0 -
1.731787
1        -0.115417                 0         0.893006          0 -
1.723734
```

```
2        -0.115417                0         0.889583          0 -
1.723734
3        -0.115417                0        -1.281712          0 -
1.717291
4        -0.703509                1        -0.397031          0 -
1.706016
...            ...              ...              ...        ...
...
14199     0.002201                0         0.070990          0
1.947664
14200     0.002201                0         0.078898          0
1.962160
14201     0.002201                0         0.070120          0
1.965381
14202     0.204448                0         0.064694          0
1.973435
14203     0.002201                0         0.073349          0
1.975046

       Outlet_Identifier  Outlet_Establishment_Year  Outlet_Size  \
0                      4                   0.139681            1
1                      7                   1.095319            1
2                      6                   1.334228            1
3                      9                  -1.532686            0
4                      7                   1.095319            1
...                  ...                        ...          ...
14199                  2                   0.736955            0
14200                  7                   1.095319            1
14201                  1                  -1.293777            2
14202                  0                  -1.532686            1
14203                  4                   0.139681            1

       Outlet_Location_Type  Outlet_Type
0                         2            1
1                         1            1
2                         0            3
3                         2            0
4                         1            1
...                     ...          ...
14199                     1            1
14200                     1            1
14201                     0            1
14202                     0            2
14203                     2            1

[14204 rows x 10 columns]

from sklearn.model_selection import train_test_split
```

```
X_train,X_test,y_train, y_test =
train_test_split(X,y,test_size=0.1,random_state=4568)

X_train.shape,X_test.shape,y_train.shape,y_test.shape

((12783, 10), (1421, 10), (12783,), (1421,))

from sklearn.ensemble import RandomForestRegressor

rfr=RandomForestRegressor(random_state= 4576)

rfr.fit(X_train,y_train)

RandomForestRegressor(random_state=4576)

y_pred = rfr.predict(X_test)

y_pred.shape

(1421,)

y_pred

array([1915.30097647, 3341.56404881,  454.23156733, ...,
876.11269707,
        460.72865795, 1825.57694872])

from sklearn.metrics import mean_squared_error,
mean_absolute_error ,r2_score

mean_squared_error(y_test,y_pred)

1406674.1051420982

mean_absolute_error(y_test,y_pred)

801.3386658377389

r2_score(y_test,y_pred)

0.5062349783454325

import matplotlib.pyplot as plt

plt.scatter(y_test,y_pred)
plt.xlabel("Actual Prices")
plt.ylabel("Predicted Prices")
plt.title("Actual Prices vs Predicted Prices")
plt.show()
```


Actual Prices vs Predicted Prices
Predicted Prices
Actual Prices
0
2000
4000
6000
8000
10000
12000
14000