```
import pandas as pd
import numpy as np
```

```
data = pd.read_csv("sales_data.csv")
print(data.head())
```

```
  Item_Identifier  Item_Weight  ...        Outlet_Type  Item_Outlet_Sales
0           FDA15         9.30  ...  Supermarket Type1          3735.1380
1           DRC01         5.92  ...  Supermarket Type2           443.4228
2           FDN15        17.50  ...  Supermarket Type1          2097.2700
3           FDX07        19.20  ...      Grocery Store           732.3800
4           NCD19         8.93  ...  Supermarket Type1           994.7052

[5 rows x 12 columns]
```

```
#drop irrelevant columns
data.drop(["Item_Identifier", "Outlet_Identifier"], axis=1, inplace=True)
```

```
#replace null values with median
data = data.fillna(data.median())
```

```
#Clean Item Fat Content Column
print(data['Item_Fat_Content'].unique())
data['Item_Fat_Content'] = data['Item_Fat_Content'].replace(['Low Fat', 'LF', 'low fat'], 0)
data['Item_Fat_Content'] = data['Item_Fat_Content'].replace(['Regular', 'reg'], 1)
print(data['Item_Fat_Content'].unique())
```

```
['Low Fat' 'Regular' 'low fat' 'LF' 'reg']
[0 1]
```

```
#Clean Item Type Column
print(data['Item_Type'].unique())
perishable = ["Breads", "Breakfast", "Dairy", "Fruits and Vegetables", "Meat", "Seafood"]
non_perishable = ["Baking Goods", "Canned", "Frozen Foods", "Hard Drinks", "Health an
              "Soft Drinks", "Snack Foods", "Starchy Foods", "Others"]
data["Item_Type"] = data["Item_Type"].replace(to_replace=perishable, value="perishable
data["Item_Type"] = data["Item_Type"].replace(to_replace=non_perishable, value="non_
data["Item_Type"] = data["Item_Type"].replace('perishable', 0)
data["Item_Type"] = data["Item_Type"].replace('non_perishable', 1)
print(data['Item_Type'].unique())
```

```
['Dairy' 'Soft Drinks' 'Meat' 'Fruits and Vegetables' 'Household'
 'Baking Goods' 'Snack Foods' 'Frozen Foods' 'Breakfast'
 'Health and Hygiene' 'Hard Drinks' 'Canned' 'Breads' 'Starchy Foods'
 'Others' 'Seafood']
[0 1]
```

```
#Clean Outlet Size Column
```

```
print(data['Outlet_Size'].unique())
data["Outlet_Size"] = data["Outlet_Size"].replace('High', 3)
data["Outlet_Size"] = data["Outlet_Size"].replace('Medium', 2)
data["Outlet_Size"] = data["Outlet_Size"].replace('Small', 1)
data["Outlet_Size"] = data["Outlet_Size"].replace(np.nan, 0)
print(data['Outlet_Size'].unique())
```

```
['Medium' nan 'High' 'Small']
[2. 0. 3. 1.]
```

```
#Clean Outlet Location Column
print(data['Outlet_Location_Type'].unique())
data["Outlet_Location_Type"] = data["Outlet_Location_Type"].replace("Tier 1", 1)
data["Outlet_Location_Type"] = data["Outlet_Location_Type"].replace("Tier 2", 2)
data["Outlet_Location_Type"] = data["Outlet_Location_Type"].replace("Tier 3", 3)
print(data['Outlet_Location_Type'].unique())
```

```
['Tier 1' 'Tier 3' 'Tier 2']
[1 3 2]
```

```
#Clean Outlet Type Column
print(data['Outlet_Type'].unique())
data["Outlet_Type"] = data["Outlet_Type"].replace("Grocery Store", 0)
data["Outlet_Type"] = data["Outlet_Type"].replace("Supermarket Type1", 1)
data["Outlet_Type"] = data["Outlet_Type"].replace("Supermarket Type2", 1)
data["Outlet_Type"] = data["Outlet_Type"].replace("Supermarket Type3", 3)
print(data['Outlet_Type'].unique())
```

```
['Supermarket Type1' 'Supermarket Type2' 'Grocery Store'
 'Supermarket Type3']
[1 0 3]
```

```
#print cleaned data
print(data.head(10))
```

```
   Item_Weight  Item_Fat_Content  ...  Outlet_Type  Item_Outlet_Sales
0        9.300                 0  ...            1          3735.1380
1        5.920                 1  ...            1           443.4228
2       17.500                 0  ...            1          2097.2700
3       19.200                 1  ...            0           732.3800
4        8.930                 0  ...            1           994.7052
5       10.395                 1  ...            1           556.6088
6       13.650                 1  ...            1           343.5528
7       12.600                 0  ...            3          4022.7636
8       16.200                 1  ...            1          1076.5986
9       19.200                 1  ...            1          4710.5350

[10 rows x 10 columns]
```

```
y  = data['Item_Outlet_Sales']
x = data.drop(['Item_Outlet_Sales'], axis=1)
```

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=42)
```

```
from sklearn.linear_model import LinearRegression
linearRegression = LinearRegression()
linearRegression.fit(x_train, y_train)
prediction = linearRegression.predict(x_test)
```

```
final = pd.DataFrame(list(zip(y_test, prediction)), columns =['Actual', 'Predicted'])
print("Actual Sales VS Predicted Sales")
print(final.head(10))
```

```
Actual Sales VS Predicted Sales
      Actual    Predicted
0  1743.0644   961.028092
1   356.8688   662.454505
2   377.5086   949.707263
3  5778.4782  4456.067478
4  2356.9320  3224.061636
5   865.5400   747.585254
6  4613.9940  4851.021897
7  2410.8618  2457.356477
8  1948.1308  1772.296631
9  1937.4780  3181.449336
```

✓ 0s completed at 14:42