

```
In [3]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

```
In [5]: House_DF = pd.read_csv("USA_Housing.csv")

In [9]: House_DF.tail(10)
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
4990	52723.876555	5.452237	8.124571	6.39	14802.088438	4.795006e+05	86727 Kelly Plaza\r\nLake Veronica, IL 04474
4991	74102.191890	5.657841	7.683993	3.13	24041.270592	1.263721e+06	2871 John Lodge\r\nAmychester, GU 61734-5597
4992	87499.125743	6.403473	4.836091	4.02	40815.199679	1.568701e+06	Unit 2096 Box 9559\r\nDPO AE 80983-8797
4993	69639.140896	5.007510	7.778375	6.05	54056.128430	1.381831e+06	5259 David Causeway Apt. 975\r\nSouth Alexstad...
4994	73060.846226	5.293682	6.312253	4.16	22695.695480	9.053549e+05	5224 Lamb Passage\r\nNancystad, GA 16579
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06	USNS Williams\r\nFPO AP 30153-7653
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06	PSC 9258, Box 8489\r\nAPO AA 42991-3352
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06	4215 Tracy Garden Suite 076\r\nJoshualand, VA ...
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06	USS Wallace\r\nFPO AE 73316
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	1.298950e+06	37778 George Ridges Apt. 509\r\nEast Holly, NV...

```
In [10]: House_DF.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Avg. Area Income     5000 non-null  float64
1   Avg. Area House Age  5000 non-null  float64
2   Avg. Area Number of Rooms  5000 non-null float64
3   Avg. Area Number of Bedrooms  5000 non-null float64
4   Area Population      5000 non-null float64
5   Price                5000 non-null float64
6   Address              5000 non-null object
dtypes: float64(6), object(1)
memory usage: 273.6+ KB
```

```
In [12]: House_DF.describe()
```

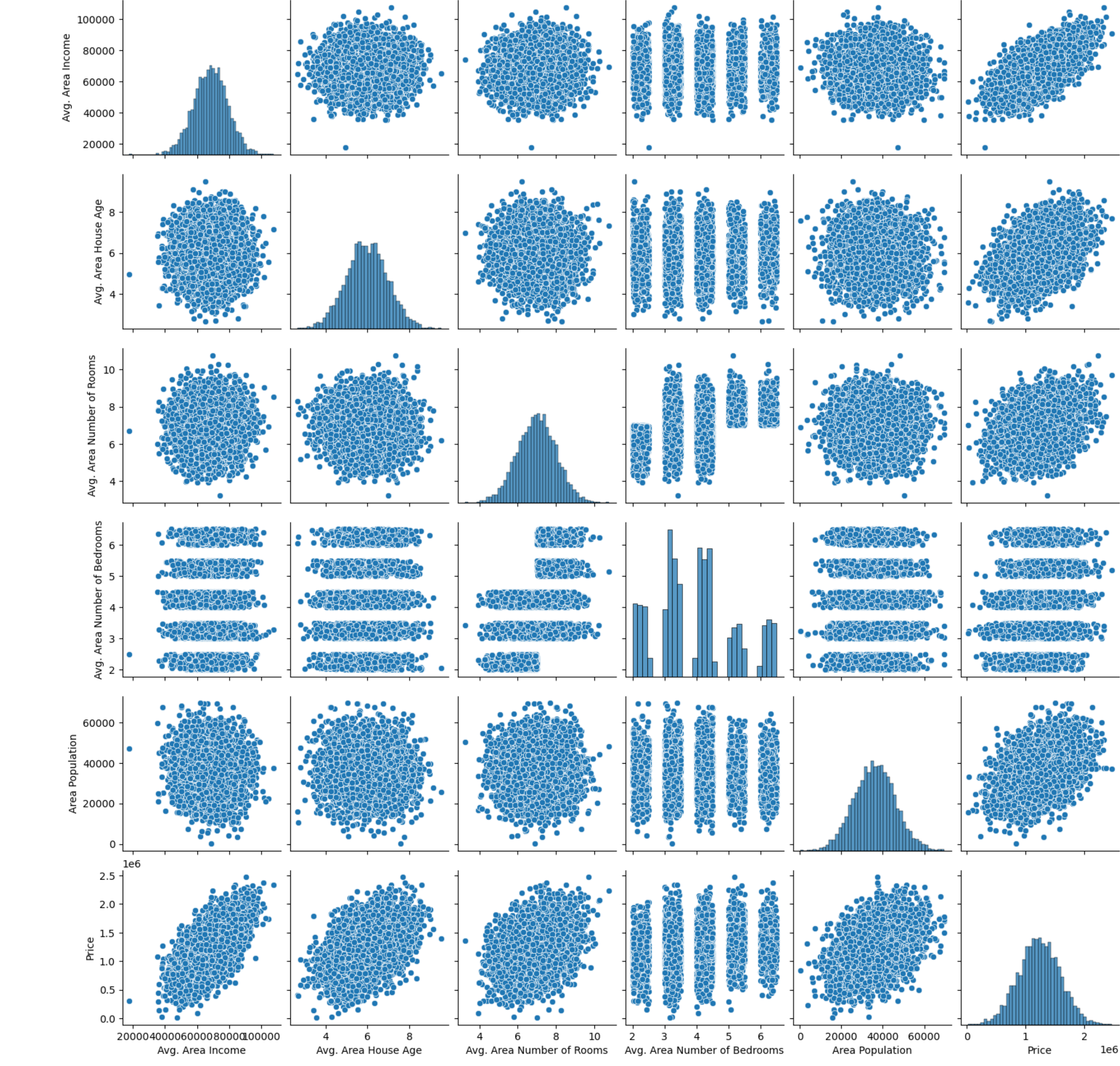
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68904.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

```
In [13]: House_DF.columns

Out[13]: Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
          'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Address'],
          dtype='object')
```

```
In [14]: sns.pairplot(House_DF)

Out[14]: <seaborn.axisgrid.PairGrid at 0x247728f6fa0>
```



```
In [17]: sns.heatmap(House_DF.corr(),annot=True)
```

```
Out[17]: <AxesSubplot:>
```



```
In [24]: X=House_DF[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
          'Avg. Area Number of Bedrooms', 'Area Population']]
Y=House_DF['Price']
```

```
In [25]: from sklearn.model_selection import train_test_split
```

```
>>> X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.50,random_state=101)
```

```
In [29]: X_train
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population
1762	82073.855417	6.712377	5.884170	4.38	34317.617986
4553	71348.726518	7.844341	6.766798	3.45	34405.947643
1049	66293.884402	6.839681	7.328415	6.13	48877.500969
224	71424.406007	5.598933	8.522538	3.49	37242.561593
590	70696.402449	6.515077	5.875832	3.33	43313.165120
...
4171	56610.642563	4.846832	7.558137	3.29	25494.740298
599	70596.850945	6.548274	6.539986	3.10	51614.830136
1361	55621.899104	3.735942	6.868291	2.30	63184.613147
1547	63044.460096	5.935261	5.913454	4.10	32725.279544
4959	75078.791516	7.644779	8.440726	4.33	56148.449322

2500 rows x 5 columns

```
In [30]: from sklearn.linear_model import LinearRegression
```

```
In [35]: LR = LinearRegression()
```

```
In [38]: LR.fit(X_train, Y_train)

Out[38]: LinearRegression()
```

```
In [43]: coeff_df = pd.DataFrame(LR.coef_,X.columns,columns = ['Coefficient'])
```

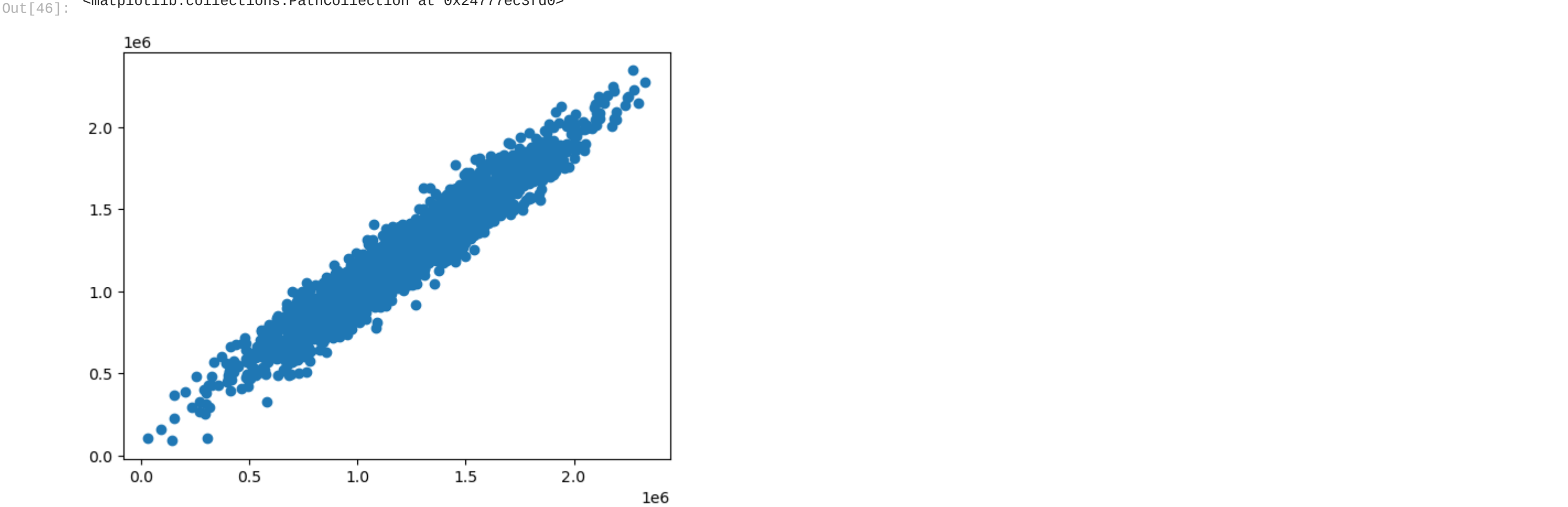
```
In [44]: coeff_df
```

	Coefficient
Avg. Area Income	21.527435
Avg. Area House Age	163813.349215
Avg. Area Number of Rooms	122376.422344
Avg. Area Number of Bedrooms	1327.089080
Area Population	15.292655

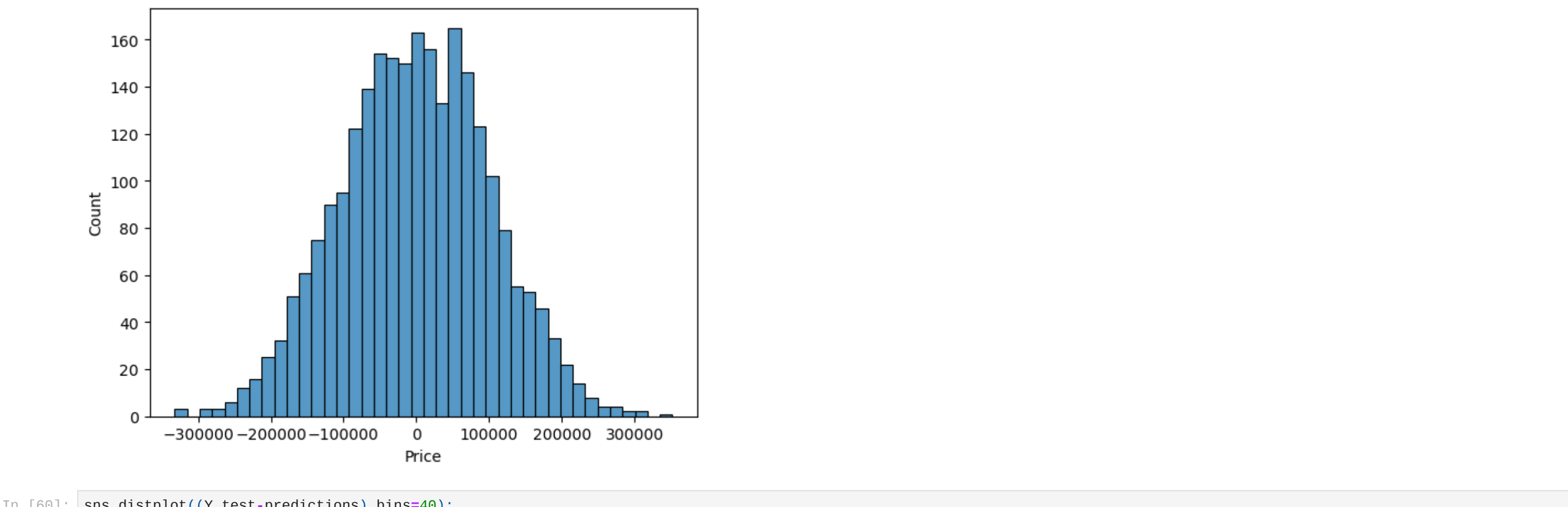
```
In [45]: predictions = LR.predict(X_test)
```

```
In [46]: plt.scatter(Y_test , predictions)
```

```
Out[46]: <matplotlib.collections.PathCollection at 0x24777ec3fd0>
```

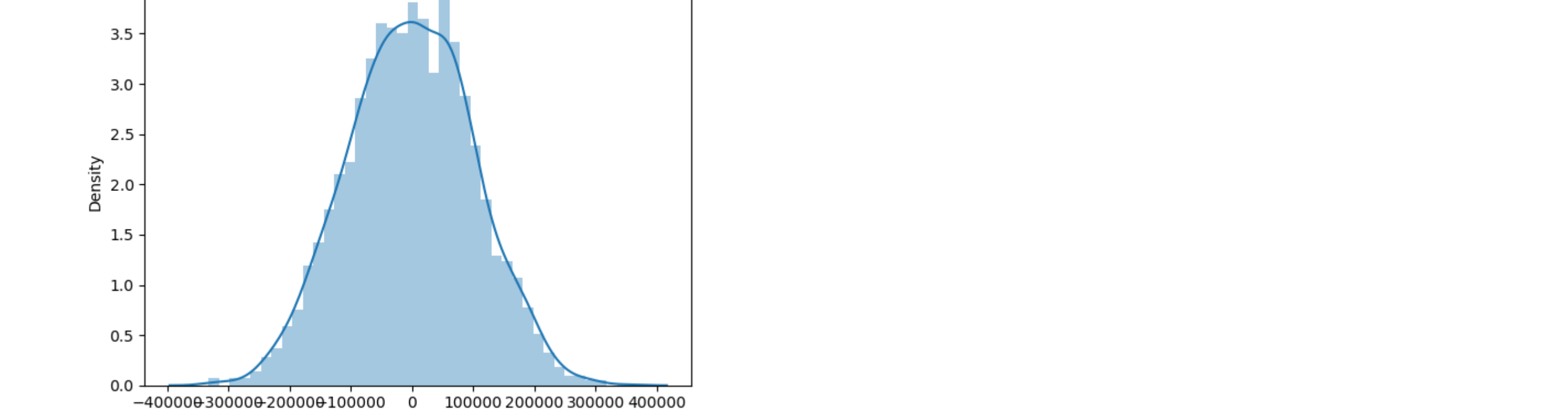


```
In [61]: sns.histplot((Y_test-predictions),bins=40);
```



```
In [60]: sns.distplot((Y_test-predictions),bins=40);
```

D:\Programfiles\Anaconda\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).



```
In [ ]:
```