Live 2025-10-28

October 29, 2025

1 Regression on USA Housing data

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[2]:
```
housing = pd.read_csv("../LectureNotes/Data/USA_Housing.csv")
```

[6]:
```
featureCols = ['Avg. Area Income',
        'Avg. Area House Age',
        'Avg. Area Number of Rooms',
        'Avg. Area Number of Bedrooms',
        'Area Population',
         #'Price',
         #'Address'
        ]
targetCol = 'Price'

X = housing[featureCols]
y = housing[targetCol]
```

[7]:
```
X.describe()
```

[7]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039
std	10657.991214	0.991456	1.005833	1.234137	9925.650114
min	17796.631190	2.644304	3.236194		
25%	61480.562388	5.322283	6.299250		
50%	68804.286404	5.970429	7.002902		
75%	75783.338666	6.650808	7.665871		
max	107701.748378	9.519088	10.759588		

```
min                        2.000000       172.610686
25%                        3.140000     29403.928702
50%                        4.050000     36199.406689
75%                        4.490000     42861.290769
max                        6.500000     69621.713378
```

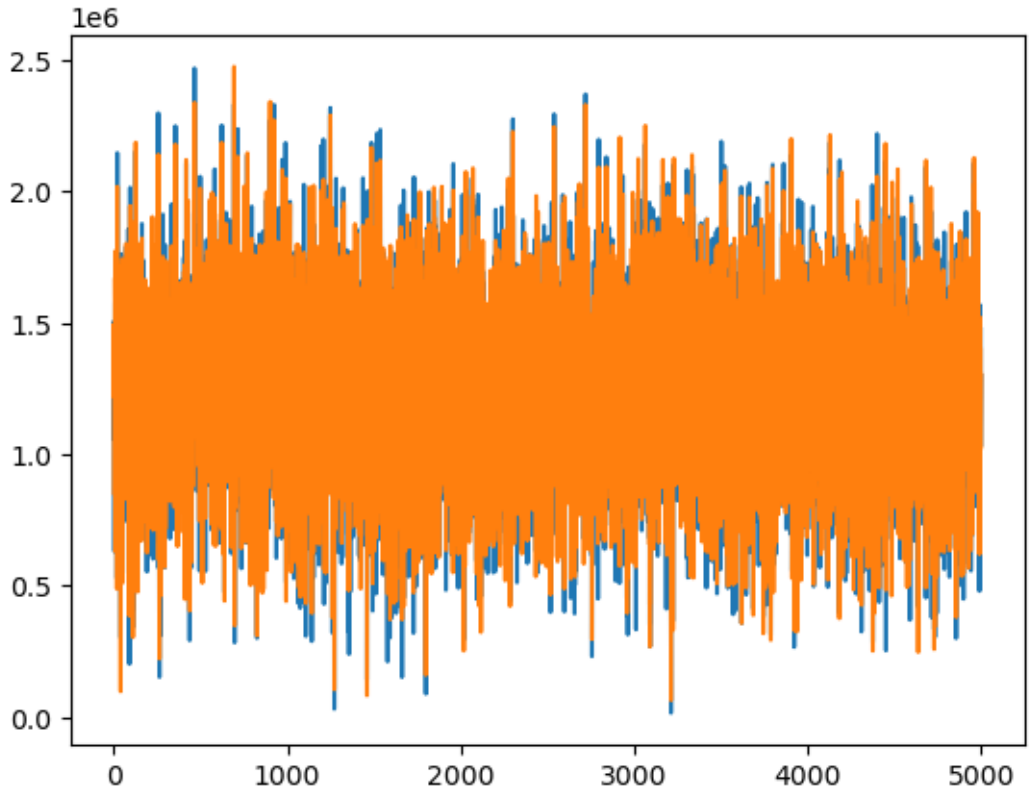
[8]:
```
y.describe()
```

[8]:
```
count    5.000000e+03
mean     1.232073e+06
std      3.531176e+05
min      1.593866e+04
25%      9.975771e+05
50%      1.232669e+06
75%      1.471210e+06
max      2.469066e+06
Name: Price, dtype: float64
```

[16]:
```
from sklearn.linear_model import Ridge

reg = Ridge(random_state=42)

reg.fit(X,y)

y_pred = pd.Series(reg.predict(X))
```

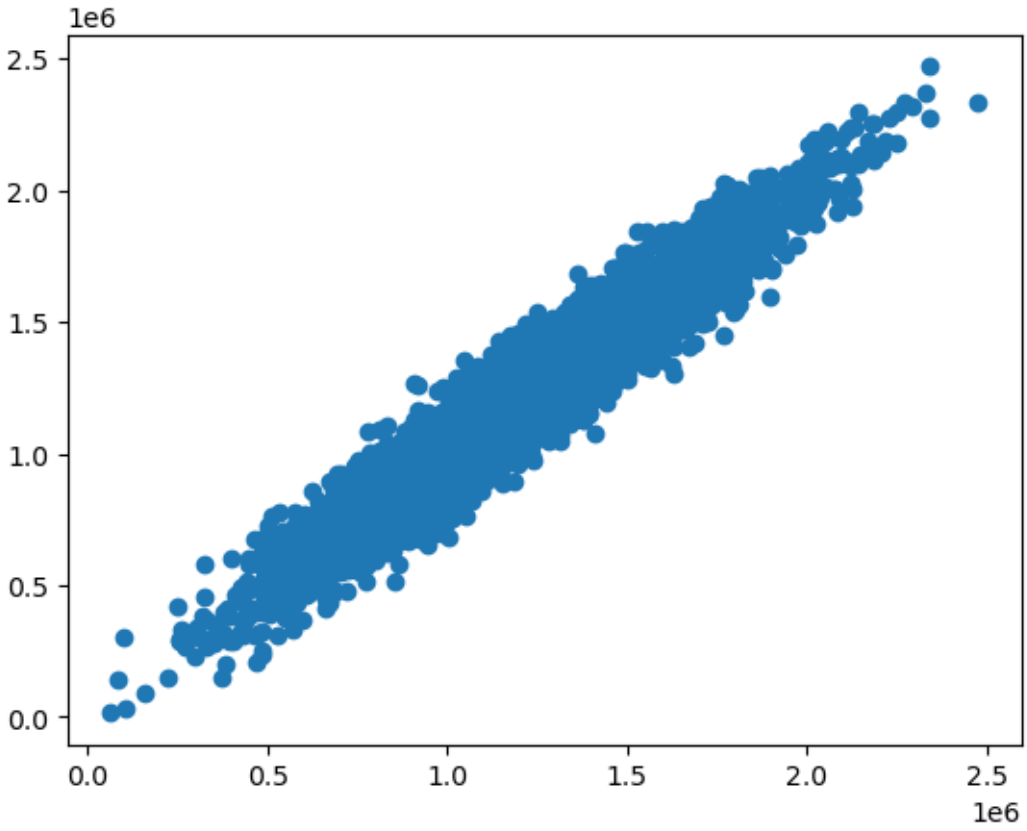
[18]:
```
plt.plot(y)
plt.plot(y_pred)
```

[18]:
```
[<matplotlib.lines.Line2D at 0x7ef885e65990>]
```



[20]:
```
plt.scatter(y_pred,y)
```

[20]:
```
<matplotlib.collections.PathCollection at 0x7ef88ea07fa0>
```


1e6
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
1e6

[ ]: