

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
```

```
%matplotlib inline
```

```
from google.colab import files
uploaded=files.upload()
```

Choose Files

No file chosen

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving USA Housing.csv to USA_Housing.csv

```
df = pd.read_csv(io.StringIO(uploaded['USA_Housing.csv'].decode('utf-8')))
df.head(5)
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael 674\nLau
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 John Suite Kath
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Stravenue\nΓ

```
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
```

```
df.describe()
```

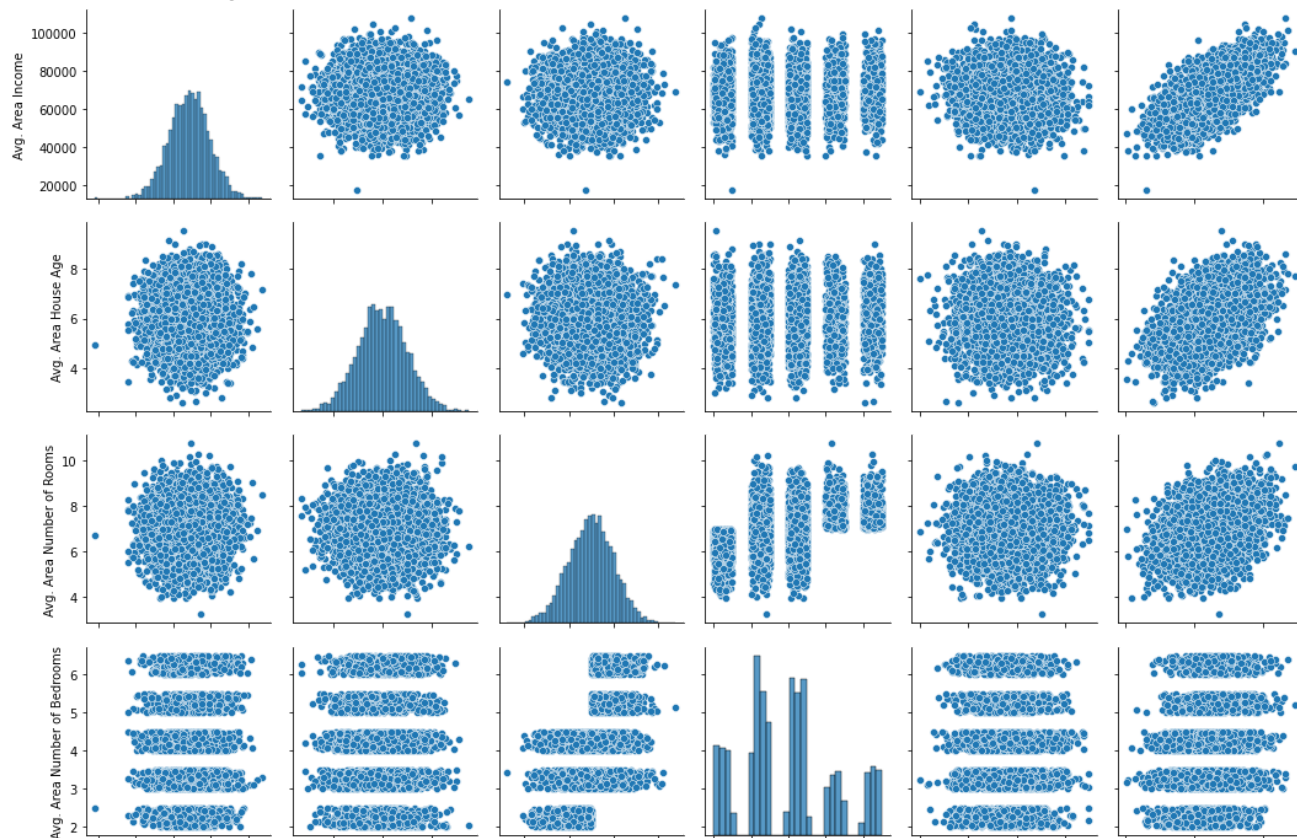
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	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%	68804.286404	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

```
df.columns
```

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

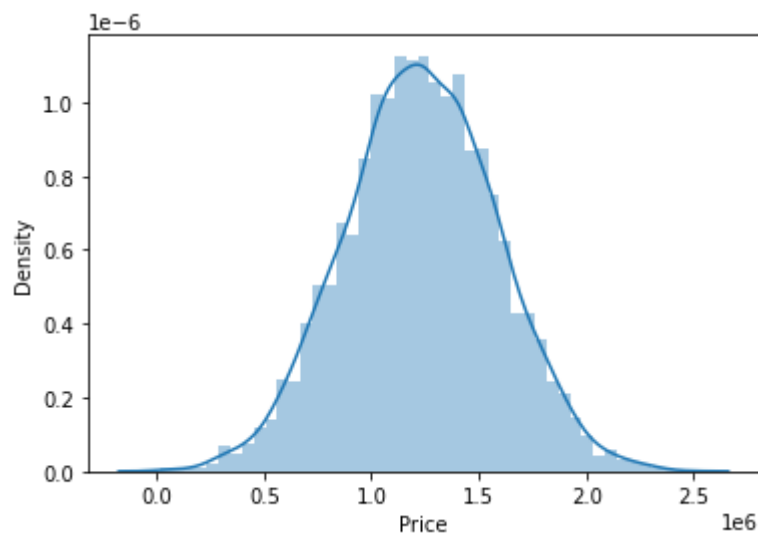
```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f4972cfb550>
```



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

```
/usr/local/lib/python3.6/dist-packages/seaborn/distributions.py:2557: FutureWarning:
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f4968bcf860>
```



```
sns.heatmap(df.corr(),annot=True)
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f4966cfb2b0>



```
df.columns
```

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

```
X=df[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
      'Avg. Area Number of Bedrooms', 'Area Population']]
```

```
y=df['Price']
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)
```

```
X_train
```

	Avg. Area	Avg. Area	Avg. Area Number of	Avg. Area Number	Area
X_test					
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population
1718	66774.995817	5.717143	7.795215	4.32	36788.980327
2511	62184.539375	4.925758	7.427689	6.22	26008.309124
345	73643.057298	6.766853	8.337085	3.34	43152.139577
2521	61909.041438	6.228343	6.593138	4.29	28953.925377
54	72942.705059	4.786222	7.319886	6.41	24377.909049
...
1776	65173.050438	7.679469	6.602618	4.23	44125.540782
4269	42969.659393	6.295501	7.885507	4.38	29594.089863
1661	48735.924512	5.543730	6.091906	2.43	19682.347295
2410	65081.584048	5.433570	9.212518	5.14	37594.493458
2302	65969.707036	7.325976	8.020966	4.09	61772.756810

```
from sklearn.linear_model import LinearRegression
lm=LinearRegression()

lm.fit(X_train,y_train)

LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

print(lm.intercept_)

-2640159.796853739

print(lm.coef_)

[2.15282755e+01 1.64883282e+05 1.22368678e+05 2.23380186e+03
 1.51504200e+01]

cdf = pd.DataFrame(lm.coef_, X.columns, columns= ['Coefficient'])
cdf
```

	Coefficient
Avg. Area Income	21.528276
Avg. Area House Age	164883.282027
Avg. Area Number of Bedrooms	2233.801864

▼ Prediction

```
prediction = lm.predict(X_test)
prediction
```

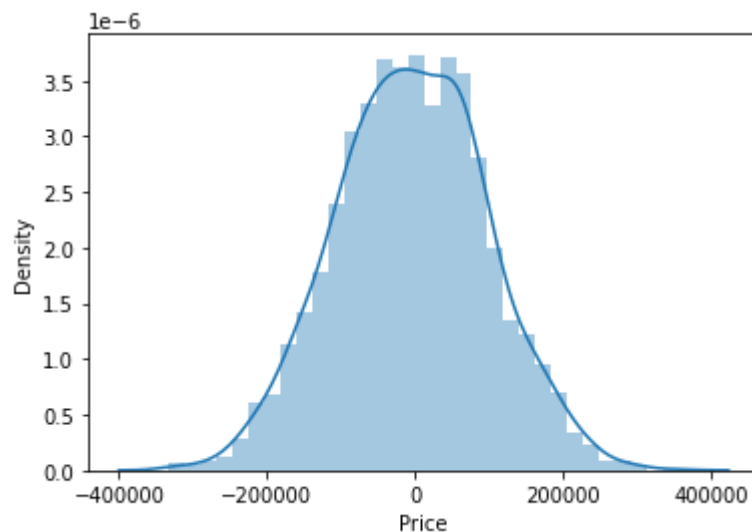
```
array([1260960.70567629,  827588.75560301, 1742421.24254363, ...,
        372191.40626868, 1365217.15140901, 1914519.54178955])
```

```
y_test
```

```
1718    1.251689e+06
2511    8.730483e+05
345     1.696978e+06
2521    1.063964e+06
54      9.487883e+05
...
1776    1.489520e+06
4269    7.777336e+05
1661    1.515271e+05
2410    1.343824e+06
2302    1.906025e+06
Name: Price, Length: 2000, dtype: float64
```

```
sns.distplot((y_test-prediction))
```

```
/usr/local/lib/python3.6/dist-packages/seaborn/distributions.py:2557: FutureWarning:
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f49637c21d0>
```



```
from sklearn import metrics  
metrics.mean_absolute_error(y_test,prediction)
```

```
82288.22251914928
```

```
metrics.mean_squared_error(y_test,prediction)
```

```
10460958907.208244
```

```
np.sqrt(metrics.mean_squared_error(y_test,prediction))
```

```
102278.82922290538
```