Project: Predicting House Prices Using Machine Learning

Phase 3: Development Part 1 – Loading and Preprocessing the Dataset

Importing Dependencies :

[8]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import r2_score, mean_absolute_error,mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR

%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

[84]:
```
dataset = pd.read_csv('C:/Users/VS/Desktop/Sakthivel/USA_Housing.csv       ')
```

[91]:
```
dataset
```

[91]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms
0	79545.458574	5.682861	7.009188
1	79248.642455	6.002900	6.730821
2	61287.067179	5.865890	8.512727
3	63345.240046	7.188236	5.586729
4	59982.197226	5.040555	7.839388
…	…	…	…
4995	60567.944140	7.830362	6.137356
4996	78491.275435	6.999135	6.576763
4997	63390.686886	7.250591	4.805081
4998	68001.331235	5.534388	7.130144
4999	65510.581804	5.992305	6.792336

```
Avg. Area Number of Bedrooms Area
Population Price
0                          4.09    23086.800503
                                   1.059034e+06
1                          3.09    40173.072174
                                   1.505891e+06
2                          5.13    36882.159400
                                   1.058988e+06
3                          3.26    34310.242831
                                   1.260617e+06
4                          4.23    26354.109472
                                   6.309435e+05
…                         …            …          …
4995                       3.46    22837.361035
                                   1.060194e+06
4996                       4.02    25616.115489
                                   1.482618e+06
4997                       2.13    33266.145490
                                   1.030730e+06
4998                       5.44    42625.620156
                                   1.198657e+06
4999                       4.07    46501.283803
                                   1.298950e+06
                                             Addr
ess 0 208 Michael Ferry Apt. 674\nLaurabury, NE
3701…
1    188 Johnson Views Suite 079\nLake Kathleen, CA…
2    9127 Elizabeth Stravenue\nDanieltown, WI 06482…
3    USS Barnett\nFPO AP 44820 4 USNS Raymond\nFPO AE 09386
…                                                 …
4995  USNS Williams\nFPO AP 30153-7653 4996 PSC
9258, Box 8489\nAPO AA 42991-3352
4997                       4215     Tracy     Garden     Suite
                           076\nJoshualand, VA 01…
4998                       USS Wallace\nFPO AE 73316
4999                       37778  George  Ridges  Apt.  509\nEast
                           Holly, NV 2…

[5000 rows x 7 columns]
```

[13]:
```
dataset.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
3Avg. Area Number of Bedrooms 5000 non-nullfloat64
4   Area Population           5000 non-null float64
5    Price                    5000 non-null float64
 6    Address    5000   non-null   dtypes:   object
float64(6),   object(1)   memory   usage:
273.6+ KB
```

[14]:
```
dataset.describe()
```

[14]:
```
  Avg. Area Income Avg. Area House Age Avg. Area Number of Rooms \
count     5000.000000       5000.000000            5000.000000
mean     68583.108984          5.977222               6.987792
std      10657.991214          0.991456               1.005833
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

       Avg. Area Number of Bedrooms Area Population   Price
count                   5000.000000    5000.000000 5.000000e+03
mean                       3.981330   36163.516039 1.232073e+06
std                        1.234137    9925.650114 3.531176e+05
min                        2.000000     172.610686 1.593866e+04
25%                        3.140000   29403.928702 9.975771e+05
50%                        4.050000   36199.406689 1.232669e+06
75%                        4.490000   42861.290769 1.471210e+06
max                        6.500000   69621.713378 2.469066e+06
```

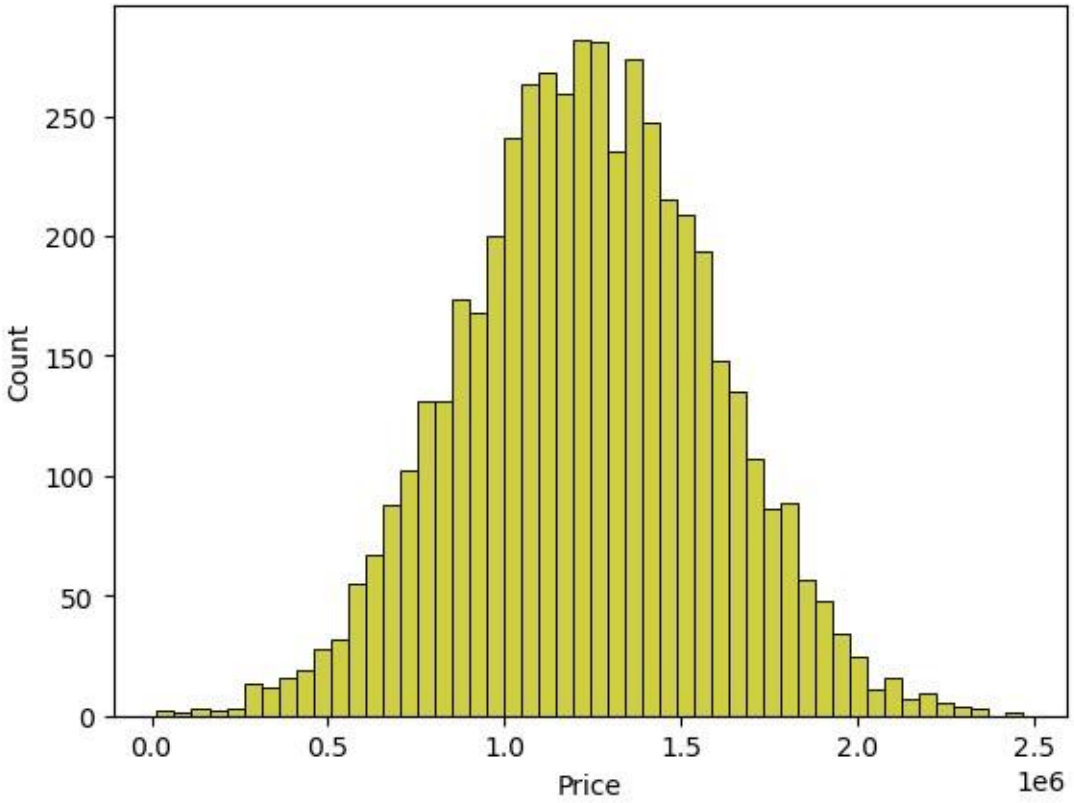
[17]:
```
dataset.columns
```

[17]:
```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number
       of Rooms', 'Avg. Area Number of Bedrooms', 'Area Population',
       'Price', 'Address'], dtype='object')
```

Visualisation and Pre-Processing of Data:

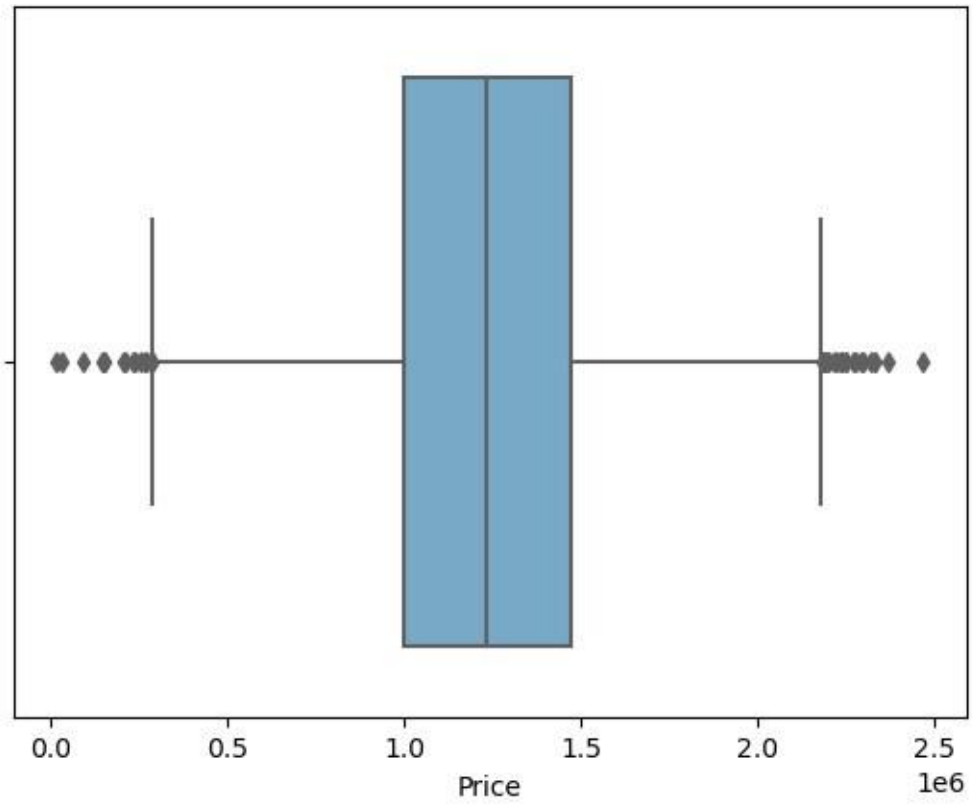
[18]:
```
sns.histplot(dataset, x='Price', bins=50, color='y')
```

[18]: <Axes: xlabel='Price', ylabel='Count'>



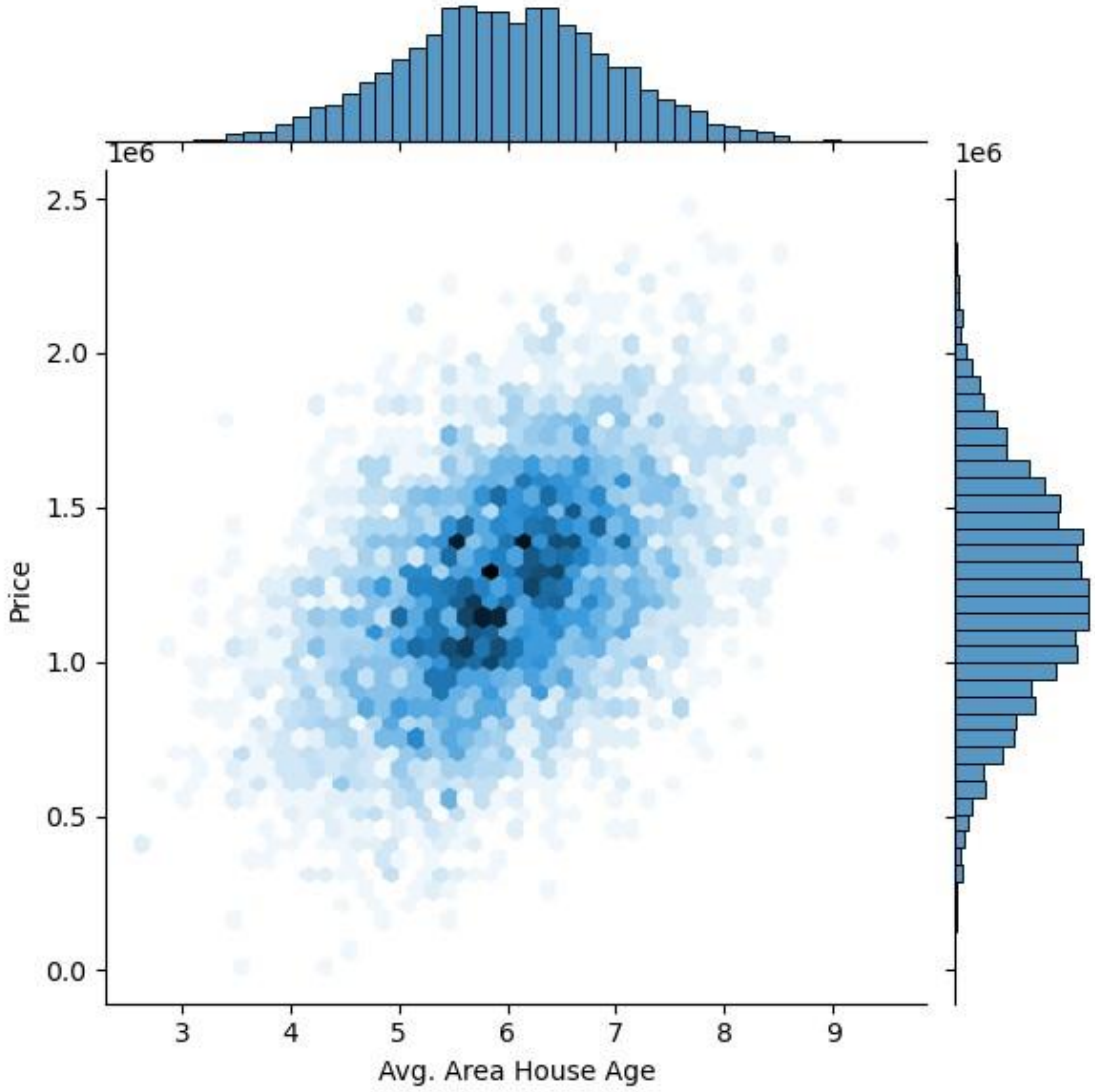
[20]:
```
sns.boxplot(dataset, x='Price', palette='Blues')
```

[20]: <Axes: xlabel='Price'>



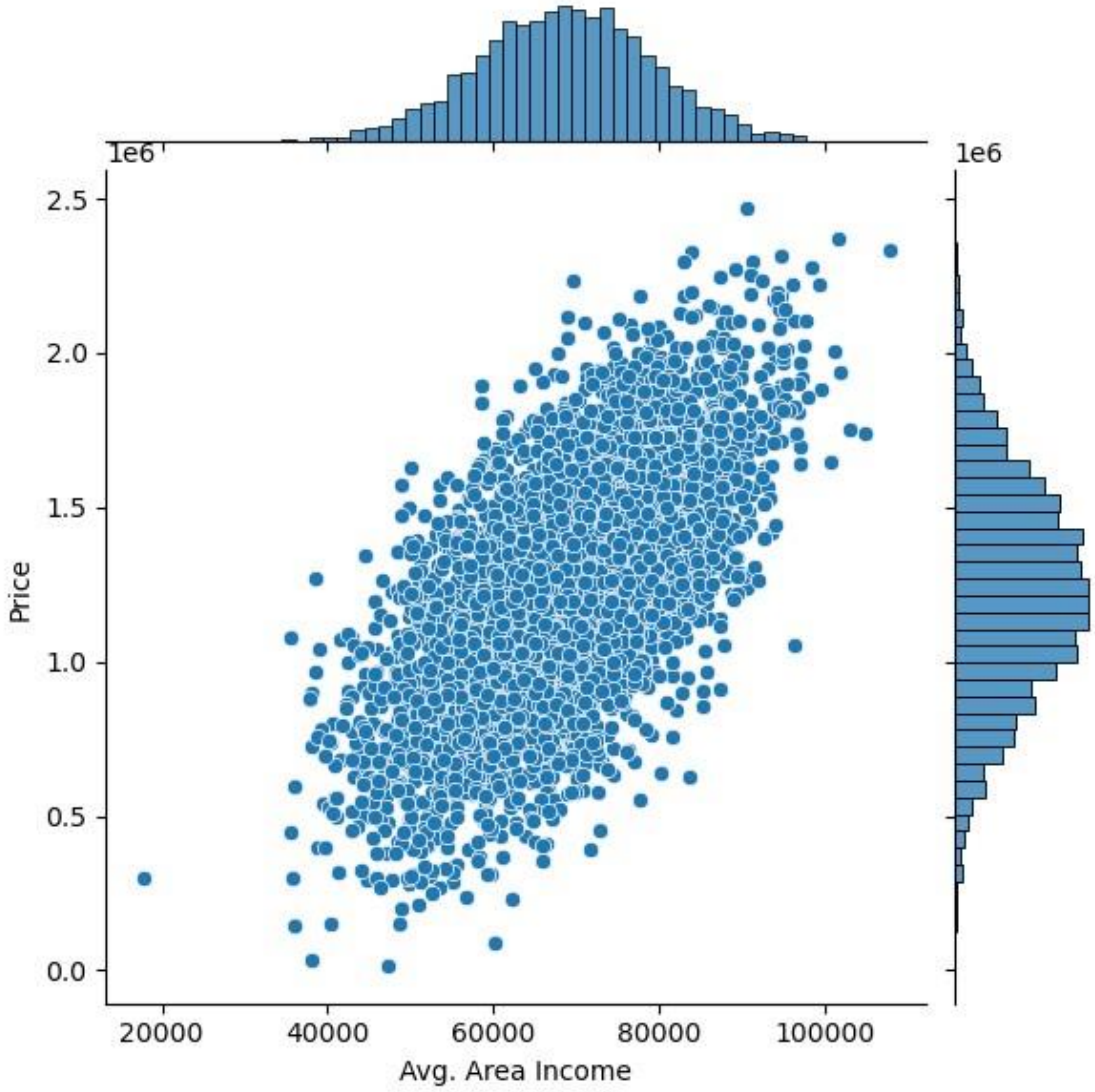
[21]:
```
sns.jointplot(dataset, x='Avg. Area House Age', y='Price',
kind='   h ex')
```

[21]: <seaborn.axisgrid.JointGrid at 0x1570cc77690>



[22]:
```
sns.jointplot(dataset, x='Avg. Area Income', y='Price')
```

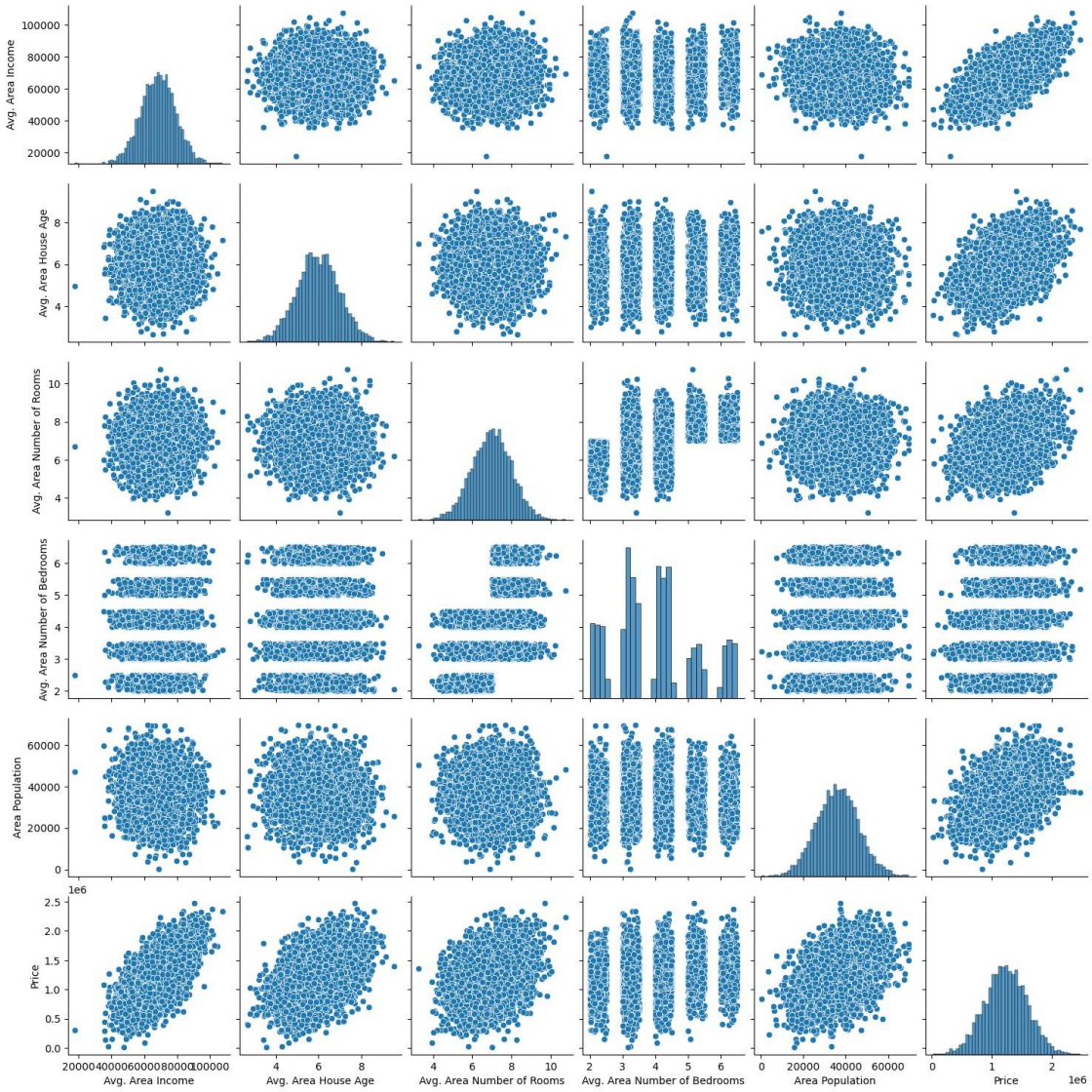
[22]:
```
<seaborn.axisgrid.JointGrid at 0x1570dfa73d0>
```



[32]:
```
plt.figure(figsize=(12,8))
sns.pairplot(dataset)
```

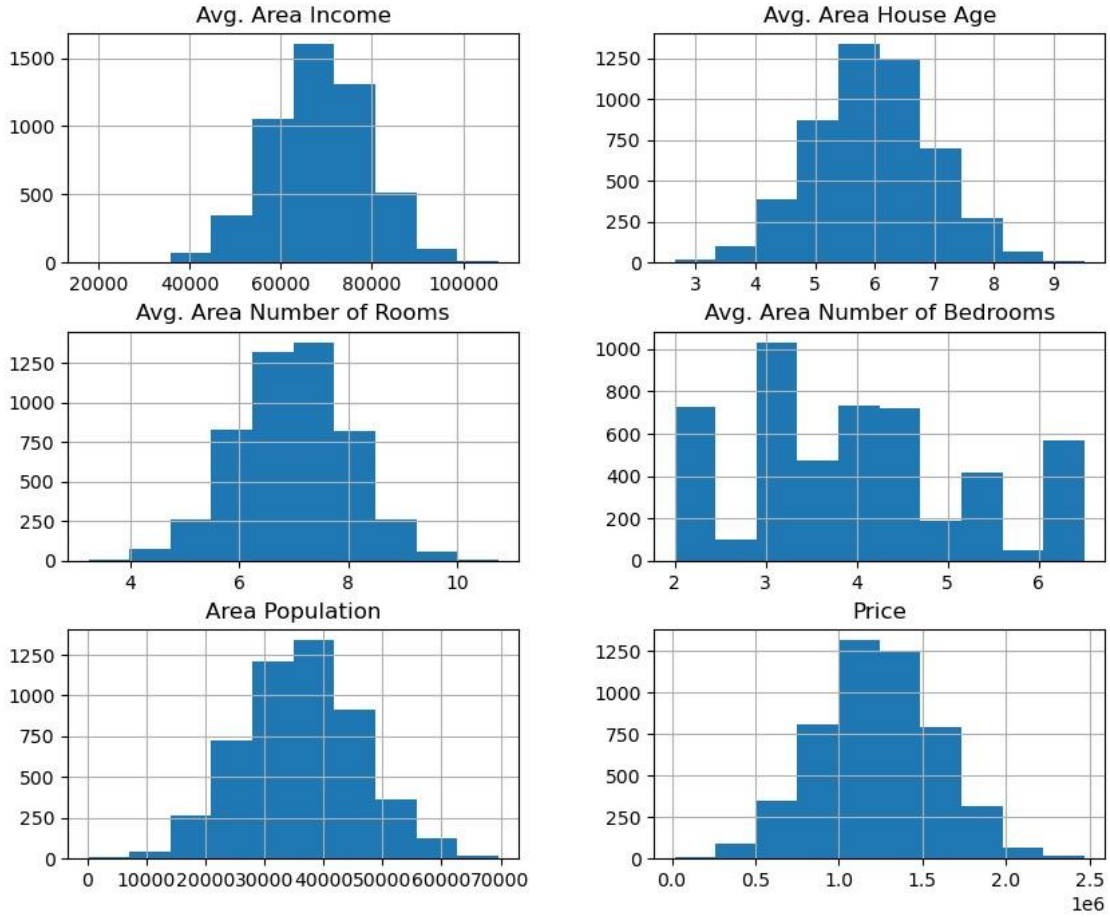
[32]: <seaborn.axisgrid.PairGrid at 0x15723570250>

<Figure size 1200x800 with 0 Axes>



[33]:
```
dataset.hist(figsize=(10,8))
```

[33]:
```
array([[<Axes: title={'center': 'Avg. Area Income'}>,
        <Axes: title={'center': 'Avg. Area House Age'}>],
       [<Axes: title={'center': 'Avg. Area Number of Rooms'}>,
<Axes: title={'center': 'Avg. Area Number of Bedrooms'}>],
[<Axes: title={'center': 'Area Population'}>,
        <Axes: title={'center': 'Price'}>]], dtype=object)
```


Avg. Area Income
Avg. Area House Age
Avg. Area Number of Rooms
Avg. Area Number of Bedrooms
Area Population
Price
1e6

Visualising Correlation:

```
[34]: dataset.corr(numeric_only=True)
```

```
[34]:                          Avg. Area Income Avg. Area House Age \
      Avg. Area Income                 1.000000          -
                                                          0.002007
      Avg. Area House Age             -0.002007           1.000000
      Avg. Area Number of Rooms       -0.011032          -
                                                          0.009428
      Avg.   Area   Number   of        0.019788           0.006149
      Bedrooms
      Area Population                 -0.016234          -
                                                          0.018743
      Price                            0.639734           0.452543
                                Avg. Area Number of Rooms \
     Avg. Area Income                          -0.011032
      Avg. Area House Age   -0.009428 Avg. Area Number
      of Rooms   1.000000
    Avg. Area Number of Bedrooms                0.462695
     Area Population                            0.002040
      Price                                     0.335664
Avg. Area Number of Bedrooms Area Population \
      Avg. Area Income                          0.019788      -
                                                              0.016234
      Avg. Area House Age                       0.006149      -
                                                              0.018743
      Avg. Area Number of Rooms                 0.462695       0.002040

      Avg.   Area   Number   of                 1.000000      -
      Bedrooms                                                0.022168
      Area Population                          -0.022168       1.000000

      Price                                     0.171071       0.408556

                                  Price
      Avg. Area Income         0.639734

      Avg. Area House Age      0.452543

      Avg. Area Number of Rooms 0.335664
      Avg. Area Number of Bedrooms 0.171071
      Area Population          0.408556
      Price                    1.000000
```

[31]:
```
plt.figure(figsize=(10,5)

sns.heatmap(dataset.corr(numeric_only=True),annot=True)
```

[35]:<Axes:>