```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
data=pd.read_csv("USA_Housing.csv")
data
```

Files ×

..

sample_data

USA_Housing.csv

	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	
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	S
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	U
...	...	...	...	...	...	...	
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06	
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06	
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06	S
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06	U

Next steps: View recommended plots

```
data.info(verbose=True)
```

```
<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
```

```
data.describe(percentiles=[0.1,0.25,0.5,0.75,0.9])
```

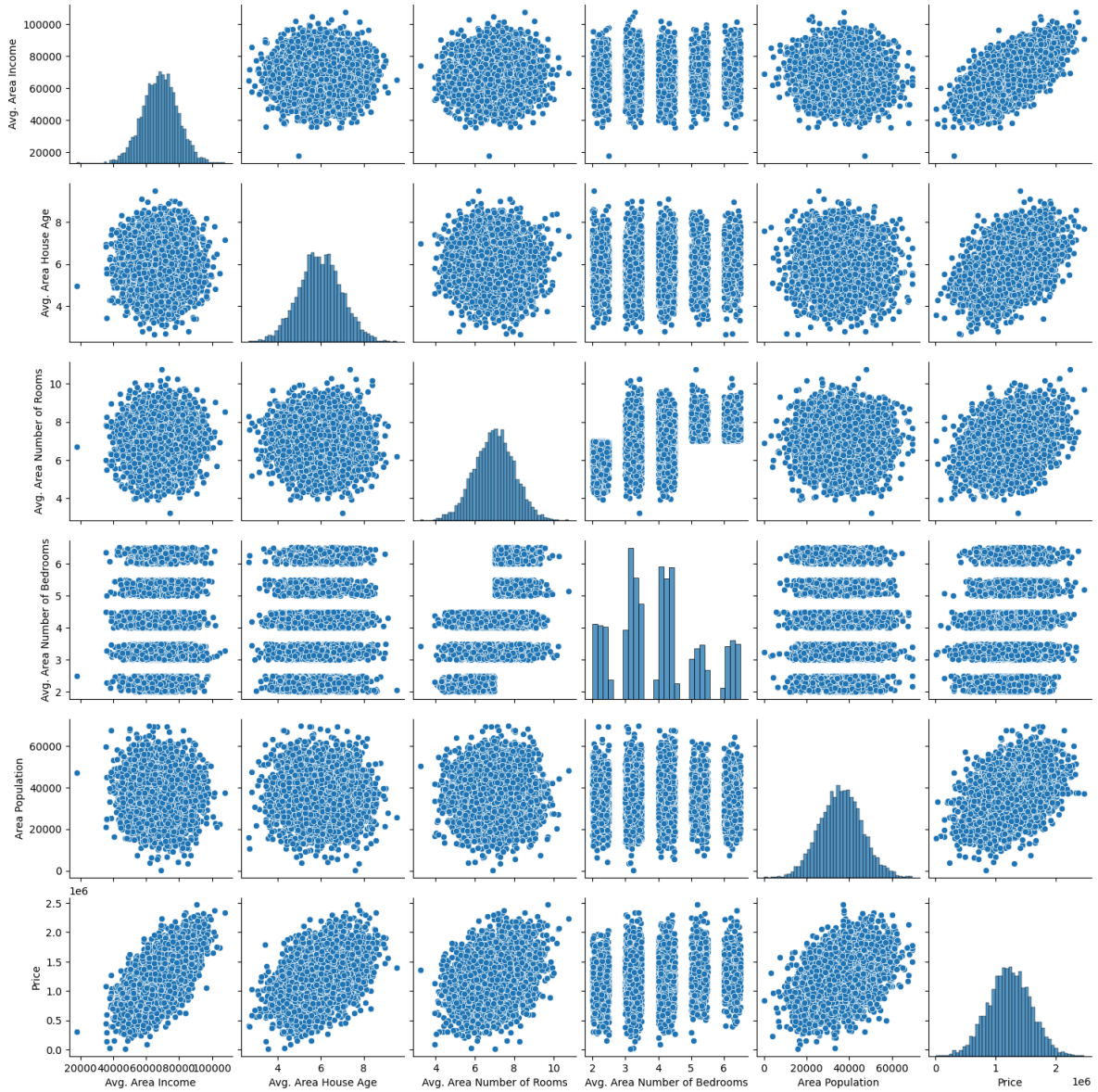
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.00(
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.23:
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.53
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.59:
10%	55047.633980	4.697755	5.681951	2.310000	23502.845262	7.72(
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.97:
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.23:
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.47
90%	82081.188283	7.243978	8.274222	6.100000	48813.618633	1.68

```
data.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')
```

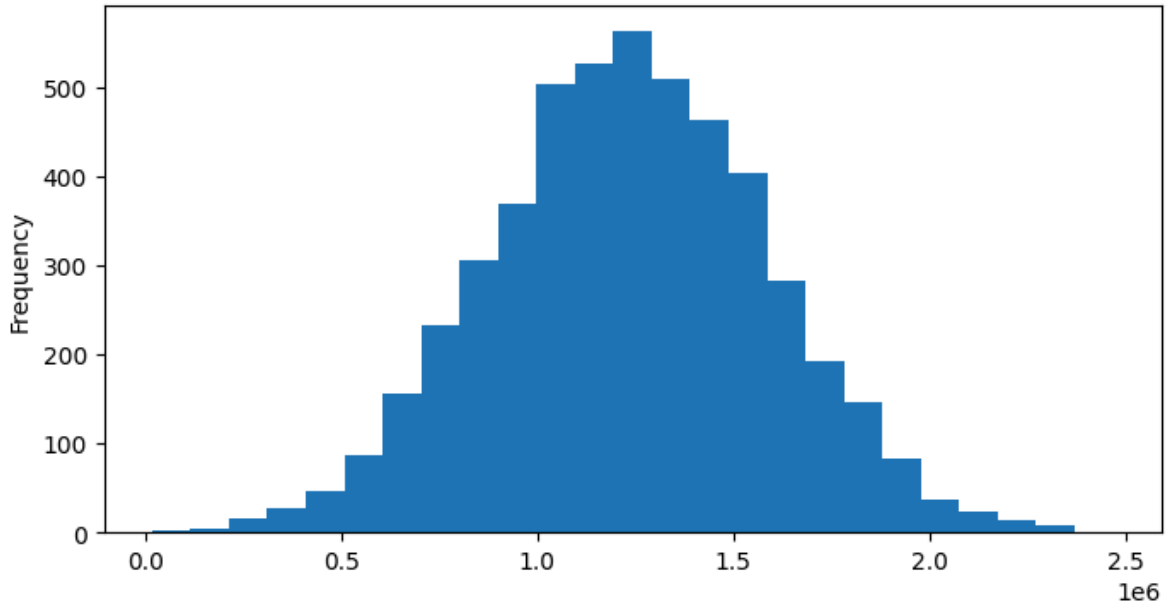
```
sns.pairplot(data)
```

```
<seaborn.axisgrid.PairGrid at 0x7b12fd742920>
```



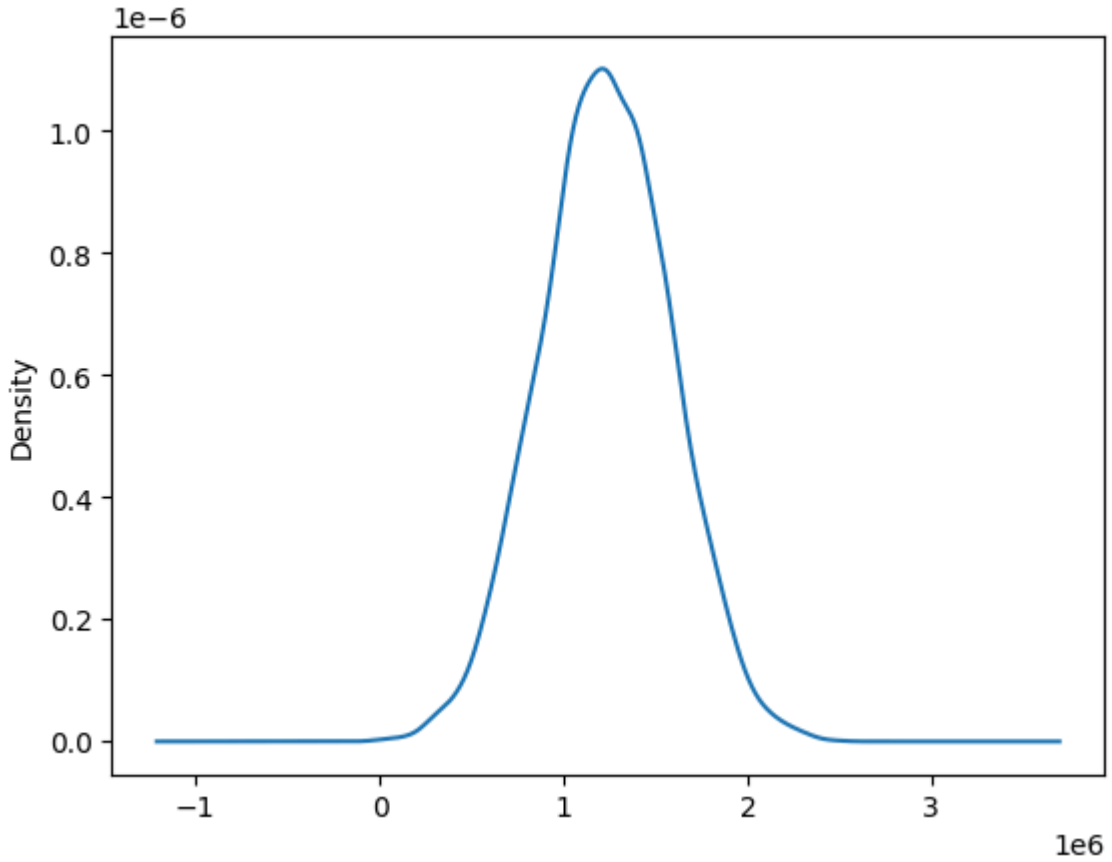
```
data['Price'].plot.hist(bins=25,figsize=(8,4))
```

```
<Axes: ylabel='Frequency'>
```




```
data['Price'].plot.density()
```

```
<Axes: ylabel='Density'>
```



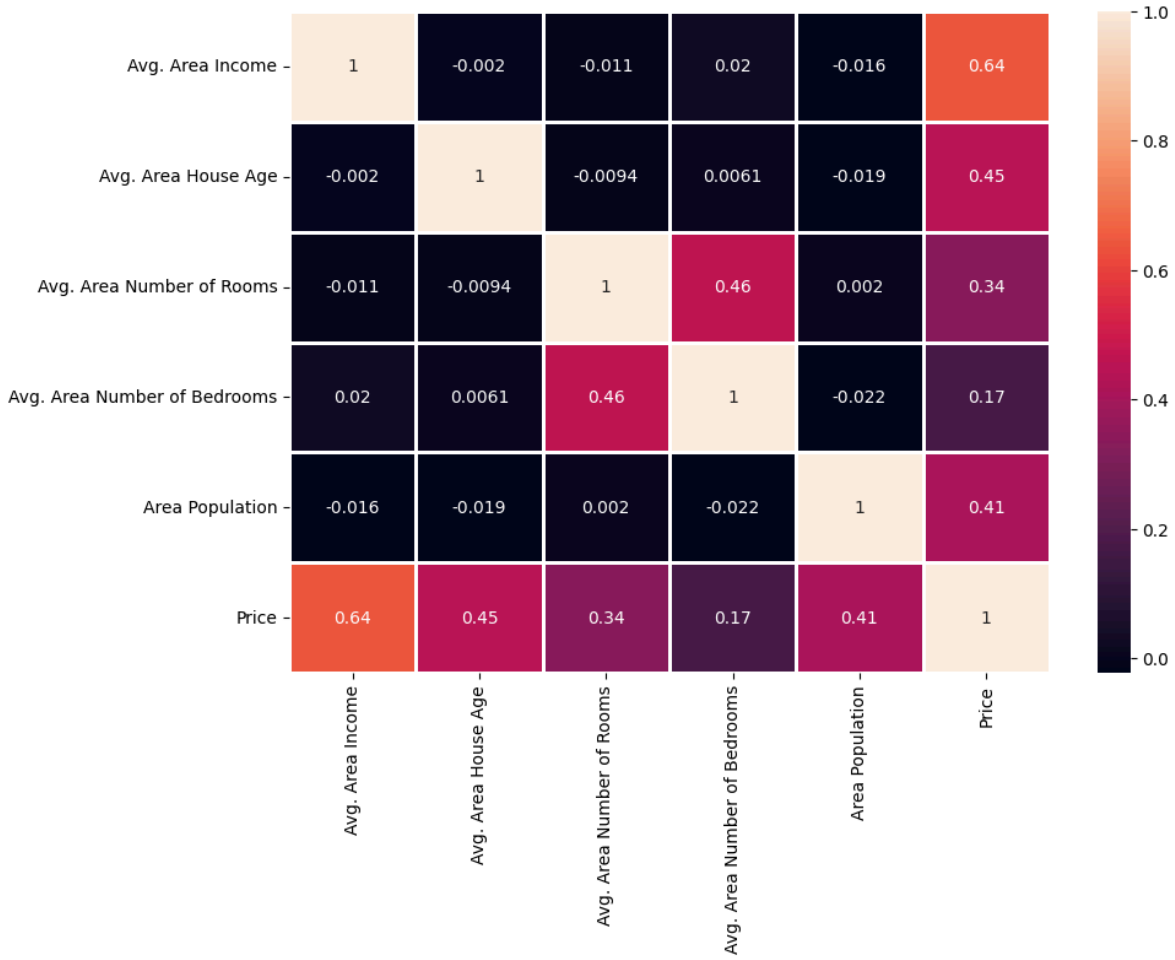

```
data.corr()
```

```
<ipython-input-19-c44ded798807>:1: FutureWarning: The default value of numeri
  data.corr()
```

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

```
plt.figure(figsize=(10,7))
sns.heatmap(data.corr(),annot=True,linewidths=2)
```

```
<ipython-input-20-6c85ed40dd2c>:2: FutureWarning: The default value of numeri
  sns.heatmap(data.corr(),annot=True,linewidths=2)
<Axes: >
```



```
l_column = list(data.columns) # Making a list out of column names
len_feature = len(l_column) # Length of column vector list
l_column
```

```
['Avg. Area Income',
 'Avg. Area House Age',
 'Avg. Area Number of Rooms',
 'Avg. Area Number of Bedrooms',
 'Area Population',
 'Price',
 'Address']
```

```
X = data[l_column[0:len_feature-2]]
y = data[l_column[len_feature-2]]
```

```
print("Feature set size:",X.shape)
print("Variable set size:",y.shape)
```

```
Feature set size: (5000, 5)
Variable set size: (5000,)
```

```
X.head()
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population
0	79545.458574	5.682861	7.009188	4.09	23086.800503
1	79248.642455	6.002900	6.730821	3.09	40173.072174
2	61287.067179	5.865890	8.512727	5.13	36882.159400
3	63345.240046	7.188236	5.586729	3.26	34310.242831
4	59982.197226	5.040555	7.839388	4.23	26354.109472

Next steps: View recommended plots

```
y.head()
```

```
0    1.059034e+06
1    1.505891e+06
2    1.058988e+06
3    1.260617e+06
4    6.309435e+05
Name: Price, dtype: float64
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                test_size=0.3, random_state=123)
print("Training feature set size:",X_train.shape)
print("Test feature set size:",X_test.shape)
print("Training variable set size:",y_train.shape)
print("Test variable set size:",y_test.shape)
```

```
Training feature set size: (3500, 5)
Test feature set size: (1500, 5)
Training variable set size: (3500,)
Test variable set size: (1500,)
```

```
from sklearn.linear_model import LinearRegression
from sklearn import metrics
lm = LinearRegression() # Creating a Linear Regression object 'lm'
lm.fit(X_train,y_train) # Fit the linear model on to the 'lm' object itself i.e.
print("The intercept term of the linear model:", lm.intercept_)
print("The coefficients of the linear model:", lm.coef_)
```

```
The intercept term of the linear model: -2631028.9017454907
The coefficients of the linear model: [2.15976020e+01 1.65201105e+05 1.190614
 1.52281212e+01]
```

```
#idict = {'Coefficients':lm.intercept_}
#idf = pd.DataFrame(data=idict,index=['Intercept'])
cdf = pd.DataFrame(data=lm.coef_, index=X_train.columns, columns=["Coefficients"]
#cdf=pd.concat([idf,cdf], axis=0)
cdf
n=X_train.shape[0]
k=X_train.shape[1]
dfN = n-k
train_pred=lm.predict(X_train)
train_error = np.square(train_pred - y_train)
sum_error=np.sum(train_error)
se=[0,0,0,0,0]
for i in range(k):
    r = (sum_error/dfN)
    r = r/np.sum(np.square(X_train[
        list(X_train.columns)[i]]-X_train[list(X_train.columns)[i]].mean()))
    se[i]=np.sqrt(r)
cdf['Standard Error']=se
cdf['t-statistic']=cdf['Coefficients']/cdf['Standard Error']
cdf
```

	Coefficients	Standard Error	t-statistic
Avg. Area Income	21.597602	0.160361	134.681505
Avg. Area House Age	165201.104954	1722.412068	95.912649
Avg. Area Number of Rooms	119061.463868	1696.546476	70.178722
Avg. Area Number of Bedrooms	3212.585606	1376.451759	2.333962

Next steps: View recommended plots

```
print("Therefore, features arranged in the order of importance for predicting the
l=list(cdf.sort_values('t-statistic',ascending=False).index)
print(' > \n'.join(l))
```

```
Therefore, features arranged in the order of importance for predicting the ho
--------------------------------------------------------------------------------
Avg. Area Income >
Avg. Area House Age >
Area Population >
Avg. Area Number of Rooms >
Avg. Area Number of Bedrooms
```

```
l=list(cdf.index)
from matplotlib import gridspec
fig = plt.figure(figsize=(18, 10))
gs = gridspec.GridSpec(2,3)
#f, ax = plt.subplots(nrows=1,ncols=len(l), sharey=True)
ax0 = plt.subplot(gs[0])
ax0.scatter(data[l[0]],data['Price'])
ax0.set_title(l[0]+" vs. Price", fontdict={'fontsize':20})

ax1 = plt.subplot(gs[1])
```

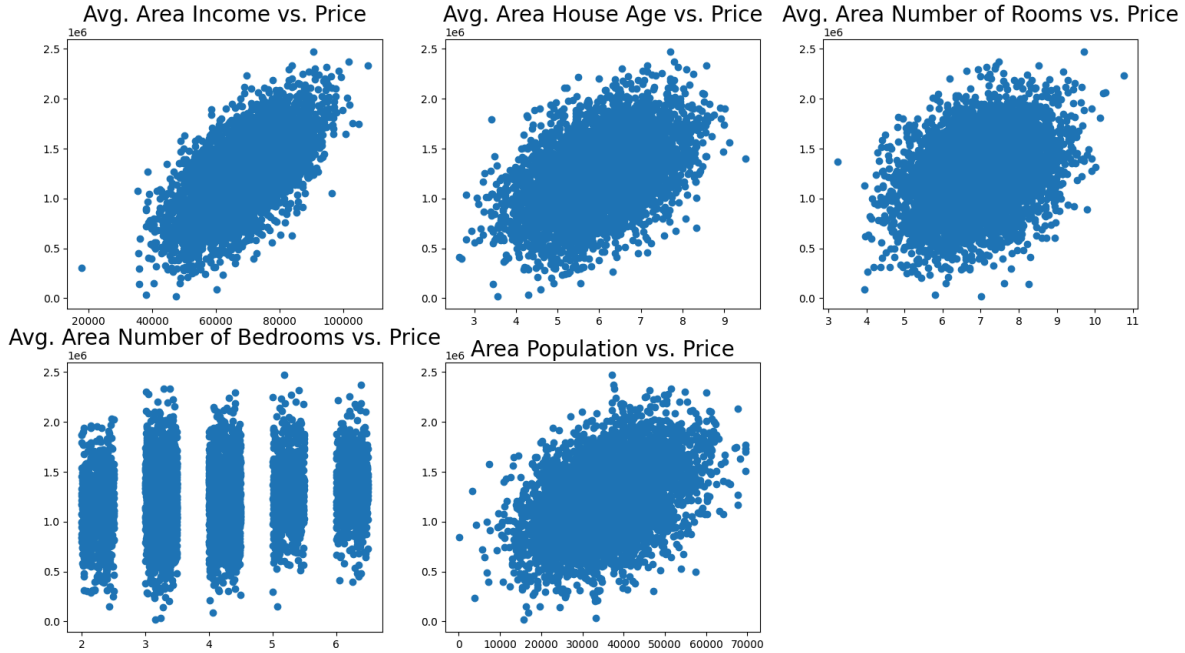
```
    ax1.scatter(data[l[1]],data['Price'])
    ax1.set_title(l[1]+" vs. Price",fontdict={'fontsize':20})

    ax2 = plt.subplot(gs[2])
    ax2.scatter(data[l[2]],data['Price'])
    ax2.set_title(l[2]+" vs. Price",fontdict={'fontsize':20})

    ax3 = plt.subplot(gs[3])
    ax3.scatter(data[l[3]],data['Price'])
    ax3.set_title(l[3]+" vs. Price",fontdict={'fontsize':20})

    ax4 = plt.subplot(gs[4])
    ax4.scatter(data[l[4]],data['Price'])
    ax4.set_title(l[4]+" vs. Price",fontdict={'fontsize':20})
```

```
Text(0.5, 1.0, 'Area Population vs. Price')
```




```
print("R-squared value of this fit:",round(metrics.r2_score(y_train,train_pred),3
```

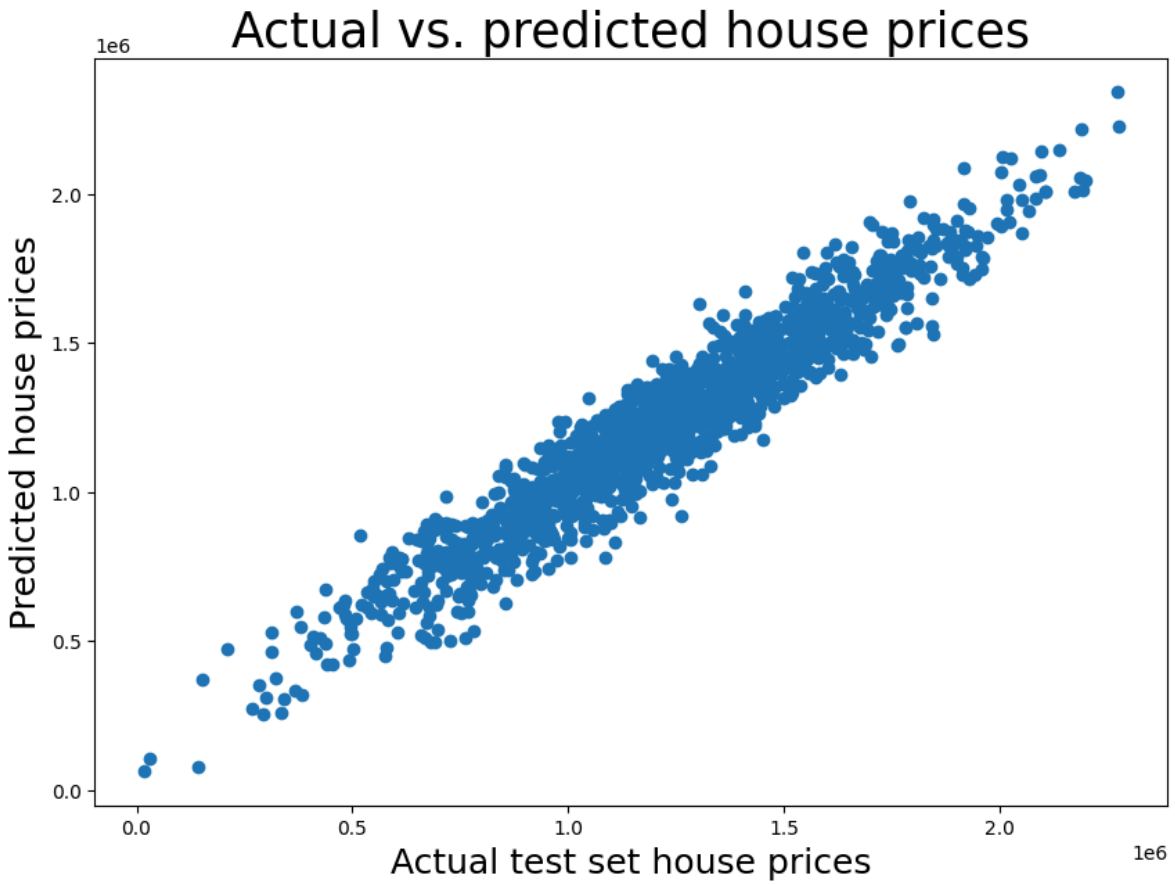
```
R-squared value of this fit: 0.917
```

```
predictions = lm.predict(X_test)
print ("Type of the predicted object:", type(predictions))
print ("Size of the predicted object:", predictions.shape)
```

```
Type of the predicted object: <class 'numpy.ndarray'>
Size of the predicted object: (1500,)
```

```
plt.figure(figsize=(10,7))
plt.title("Actual vs. predicted house prices",fontsize=25)
plt.xlabel("Actual test set house prices",fontsize=18)
plt.ylabel("Predicted house prices", fontsize=18)
plt.scatter(x=y_test,y=predictions)
```

```
<matplotlib.collections.PathCollection at 0x7b12f21e4f70>
```



```
plt.figure(figsize=(10,7))
plt.title("Histogram of residuals to check for normality",fontsize=25)
plt.xlabel("Residuals",fontsize=18)
plt.ylabel("Kernel density", fontsize=18)
sns.histplot([y_test-predictions])
```

```
<Axes: title={'center': 'Histogram of residuals to check for normality'},
xlabel='Residuals', ylabel='Kernel density'>
```