A real estate agent want help to predict the house price for regions in Usa.he gave us the dataset to work on to use linear Regression model.Create a model that helps him to estimate

Data Collection

In [5]:
```
#import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [22]:
```
#import the dataset
data=pd.read_csv(r"C:\Users\user\Desktop\Vicky\10_USA_Housing.csv")
```

In [23]:
```
#to display top 10 rows
data.head()
```

Out[23]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Addre
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry A 674\nLaurabury, N 3701
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Viev Suite 079\nLa Kathleen, CA
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabe Stravenue\nDanieltow WI 06482
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO A 448
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFF AE 093

In [24]:
```
#to display null values
data.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 [25]:
```
#to display summary of statistics
data.describe()
```

Out[25]:

	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
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

In [26]:
```
#to display columns name
data.columns
```

Out[26]:
```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Room
s',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Addres
s'],
      dtype='object')
```

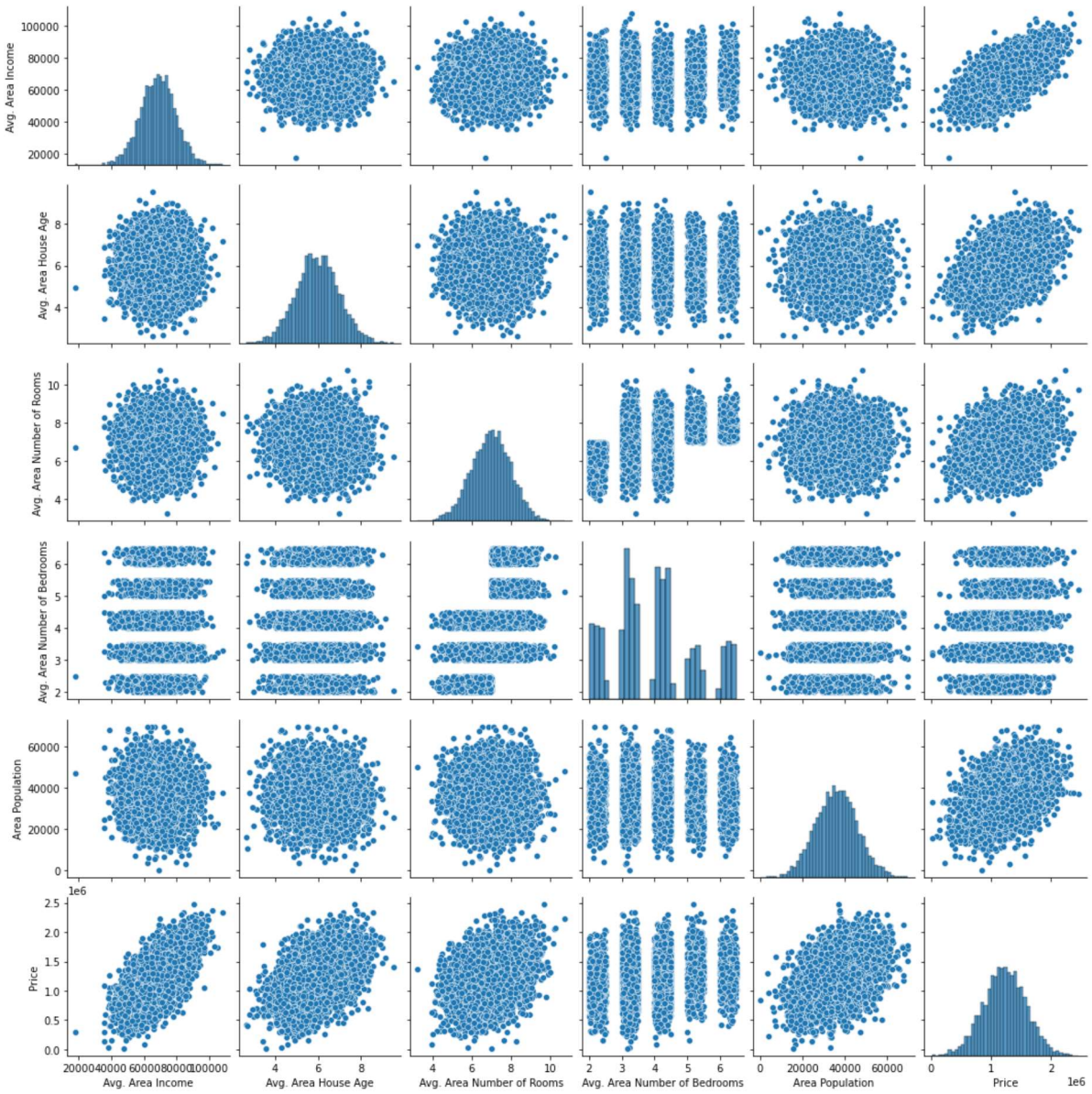
In [ ]:
```

```

EDA and Visualization

In [27]:
```
sns.pairplot(data)
```

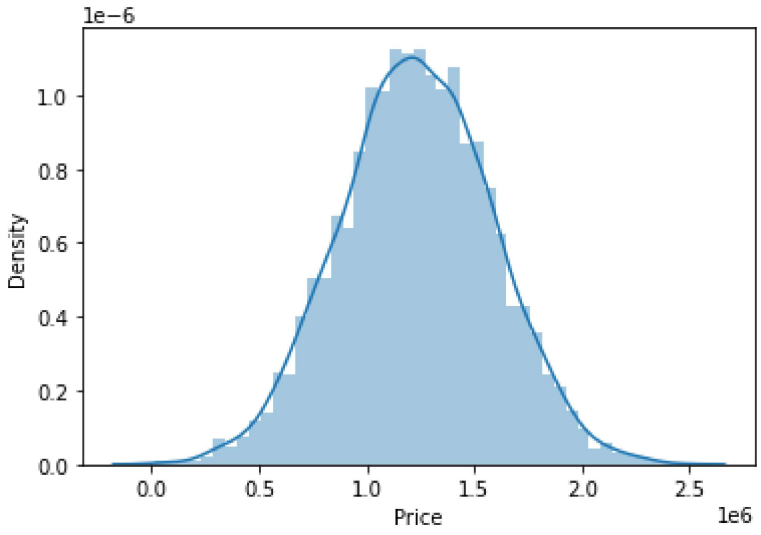
Out[27]:
```
<seaborn.axisgrid.PairGrid at 0x23ced545040>
```



In [28]:
```
sns.distplot(data['Price'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: Fut
ureWarning: `distplot` is a deprecated function and will be removed in a futu
re version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for hi
stograms).
  warnings.warn(msg, FutureWarning)
```

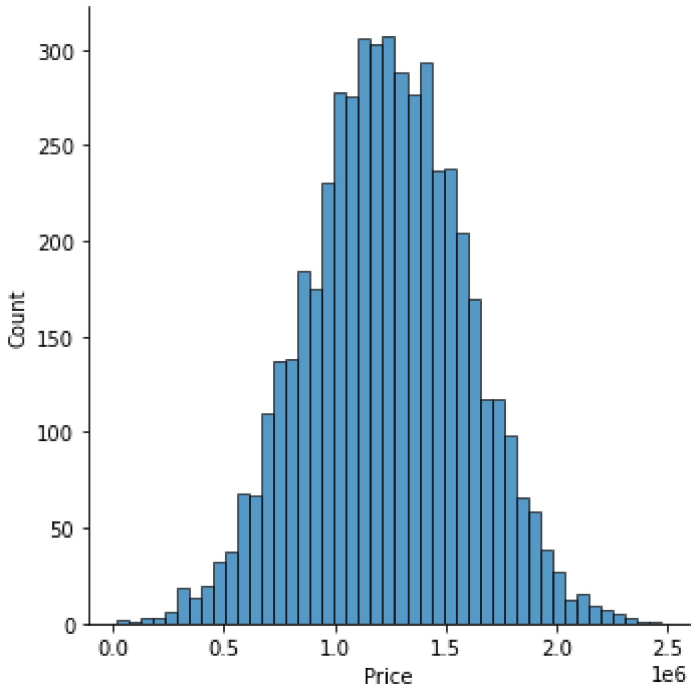
Out[28]: <AxesSubplot:xlabel='Price', ylabel='Density'>



In [29]:
```
data1=data[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Ro
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price']]
```

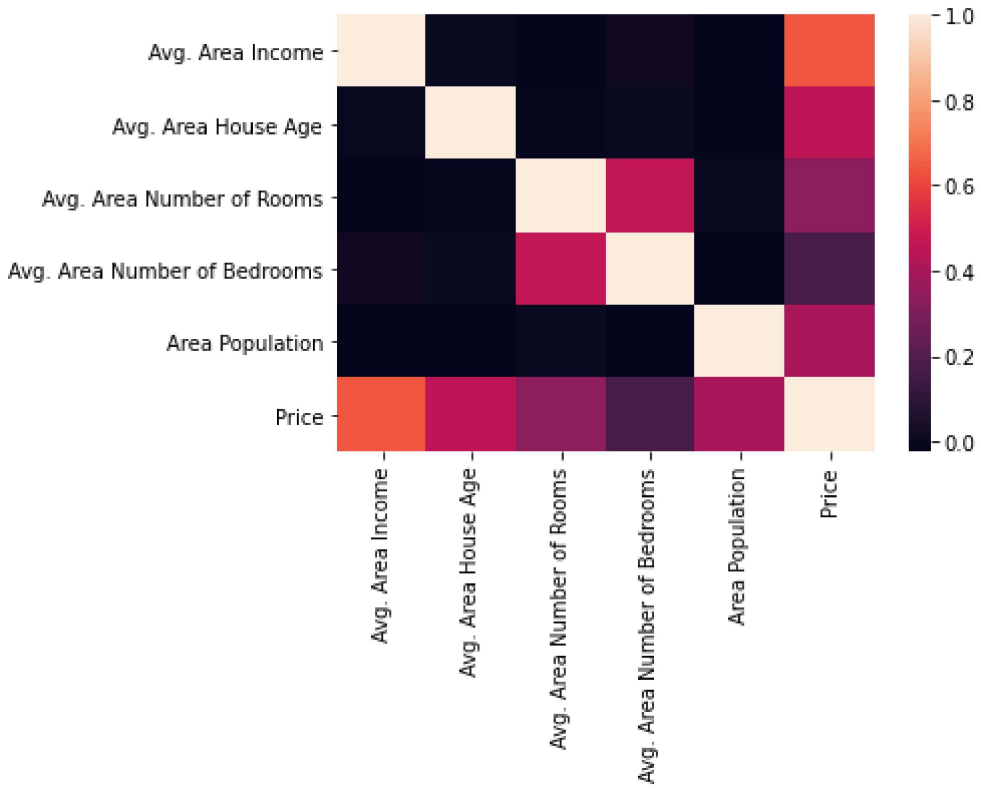
In [30]:
```
sns.displot(data['Price'])
```

Out[30]: <seaborn.axisgrid.FacetGrid at 0x23cf3482700>



In [31]:
```
sns.heatmap(data1.corr())
```

Out[31]: <AxesSubplot:>



To train the model

we are going to train the linear regression model ;We need to split the two variable x and y where x in independent variable (input) and y is dependent of x(output) so we could ignore address columns as it is not requires for our model

In [32]:
```
x=data1[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms
       'Avg. Area Number of Bedrooms', 'Area Population']]
y=data1['Price']
```

In [33]:
```


#To split test and train data
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.6)
```

In [34]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[34]: LinearRegression()

In [35]:
```
lr.intercept_
```

Out[35]:
```
-2647298.9378510206
```

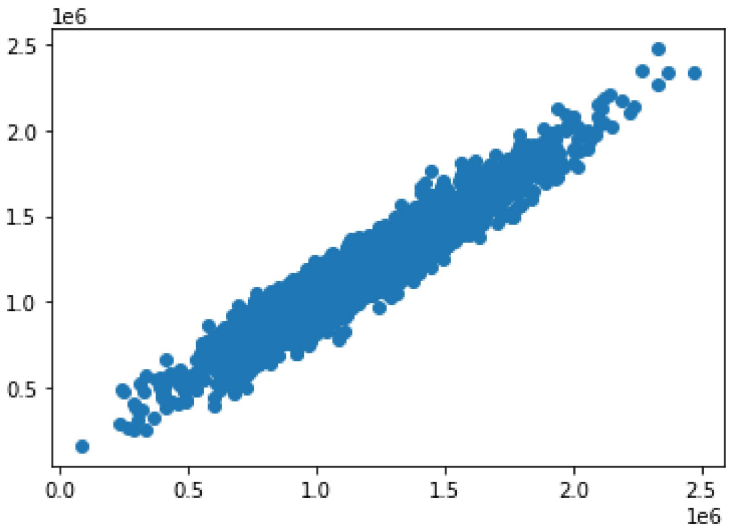
In [36]:
```
coeff = pd.DataFrame(lr.coef_,x.columns,columns=["Co-efficient"])
coeff
```

Out[36]:

	Co-efficient
Avg. Area Income	21.647983
Avg. Area House Age	165920.731183
Avg. Area Number of Rooms	120157.047189
Avg. Area Number of Bedrooms	3484.414564
Area Population	15.208294

In [37]:
```
prediction = lr.predict(x_train)
plt.scatter(y_train,prediction)
```

Out[37]:
```
<matplotlib.collections.PathCollection at 0x23cf3b20850>
```



In [38]:
```
lr.score(x_test,y_test)
```

Out[38]:
```
0.9190238850262678
```

In [ ]:

In [ ]: