# Problem statement

A real estate agent help to predict the price of the house for region in USA.He gave us the

# DATA COLLECTION

In [1]:

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

In [2]:

```
#to import dataset
data=pd.read_csv(r"C:\Users\user\Downloads\10_USA_Housing.csv")
data
```

Out[2]:

| | Avg. Area Income | Avg. Area House Age | Avg. Area Number of Rooms | Avg. Area Number of Bedrooms | Area Population | Price | Ac |
|---|---|---|---|---|---|---|---|
| **0** | 79545.458574 | 5.682861 | 7.009188 | 4.09 | 23086.800503 | 1.059034e+06 | 208 Michael Fer 674\nLaurabu |
| **1** | 79248.642455 | 6.002900 | 6.730821 | 3.09 | 40173.072174 | 1.505891e+06 | 188 Johnson Suite 079 Kathleer |
| **2** | 61287.067179 | 5.865890 | 8.512727 | 5.13 | 36882.159400 | 1.058988e+06 | 9127 Eli Stravenue\nDanie WI 0 |
| **3** | 63345.240046 | 7.188236 | 5.586729 | 3.26 | 34310.242831 | 1.260617e+06 | USS Barnett\nF |
| **4** | 59982.197226 | 5.040555 | 7.839388 | 4.23 | 26354.109472 | 6.309435e+05 | USNS Raymond AE |
| **...** | ... | ... | ... | ... | ... | ... | |
| **4995** | 60567.944140 | 7.830362 | 6.137356 | 3.46 | 22837.361035 | 1.060194e+06 | USNS Williams AP 3015 |
| **4996** | 78491.275435 | 6.999135 | 6.576763 | 4.02 | 25616.115489 | 1.482618e+06 | PSC 925 8489\nAPO AA |
| **4997** | 63390.686886 | 7.250591 | 4.805081 | 2.13 | 33266.145490 | 1.030730e+06 | 4215 Tracy Suite 076\nJoshu V |
| **4998** | 68001.331235 | 5.534388 | 7.130144 | 5.44 | 42625.620156 | 1.198657e+06 | USS Wallace\nF |
| **4999** | 65510.581804 | 5.992305 | 6.792336 | 4.07 | 46501.283803 | 1.298950e+06 | 37778 George Apt. 509\nEas |

5000 rows × 7 columns

In [3]:

```
#to display top 5 rows
data.head()
```

Out[3]:

|  | 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<br>674\nLaurabury,<br>370 |
| **1** | 79248.642455 | 6.002900 | 6.730821 | 3.09 | 40173.072174 | 1.505891e+06 | 188 Johnson Vie<br>Suite 079\nLa<br>Kathleen, C |
| **2** | 61287.067179 | 5.865890 | 8.512727 | 5.13 | 36882.159400 | 1.058988e+06 | 9127 Elizab<br>Stravenue\nDanieltov<br>WI 0648 |
| **3** | 63345.240046 | 7.188236 | 5.586729 | 3.26 | 34310.242831 | 1.260617e+06 | USS Barnett\nFPO<br>448 |
| **4** | 59982.197226 | 5.040555 | 7.839388 | 4.23 | 26354.109472 | 6.309435e+05 | USNS Raymond\nF<br>AE 093 |

In [4]:

```
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 [6]:

```
#to display summary of statistics

data.describe()
```

Out[6]:

| | 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 [11]:

```
#to display columns heading

data.columns
```

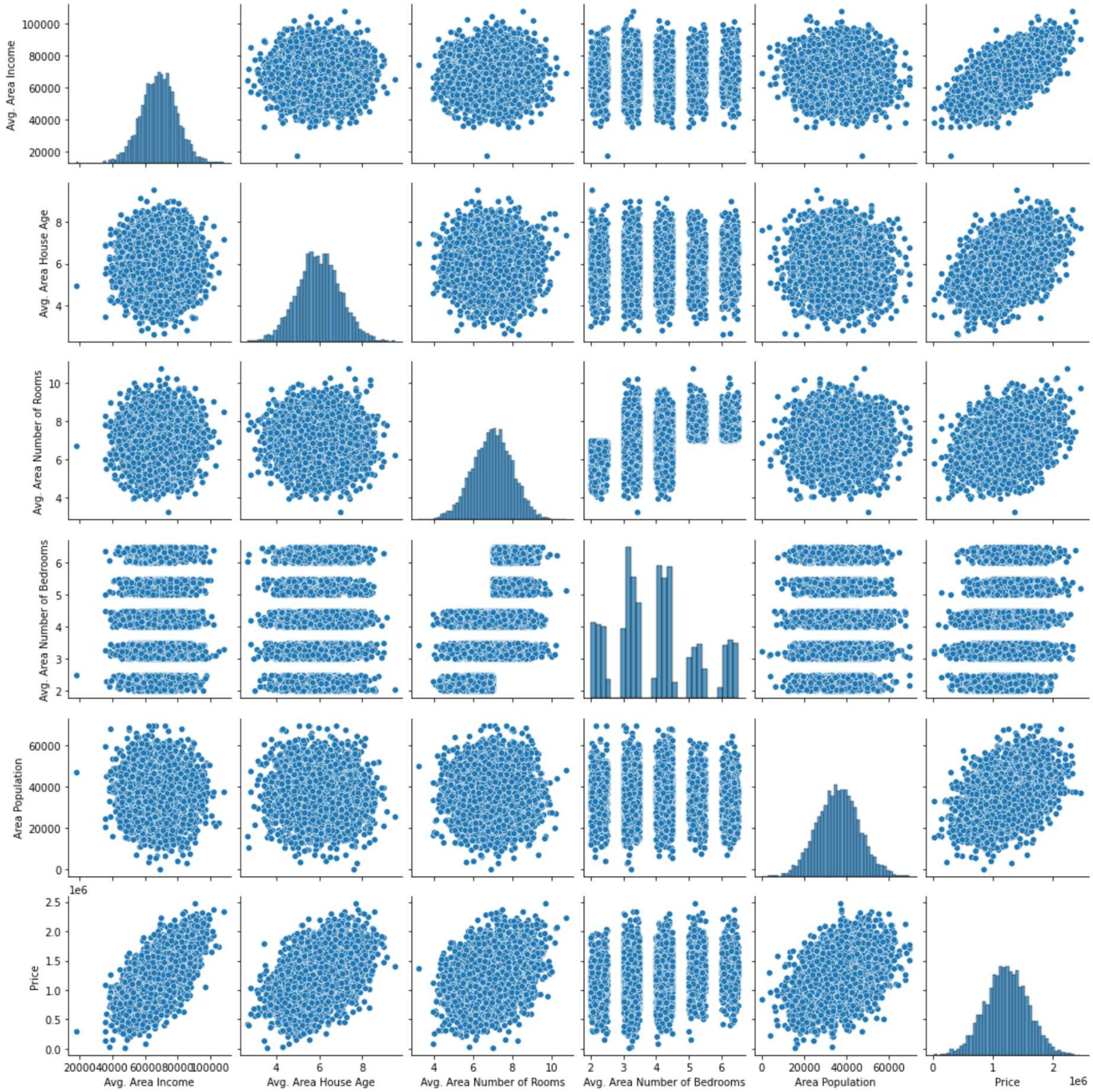
Out[11]:

```
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')
```

# EDA and VISUALIZATION

```
In [12]: sns.pairplot(data)
```

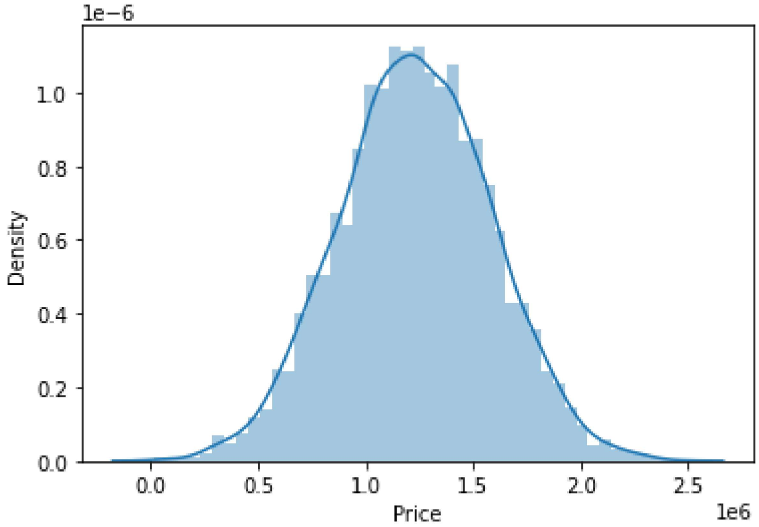
```
Out[12]: <seaborn.axisgrid.PairGrid at 0x2a8c629f3d0>
```



In [15]:
```
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[15]:
```
<AxesSubplot:xlabel='Price', ylabel='Density'>
```



In [16]:
```
da=data[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price']]
da
```

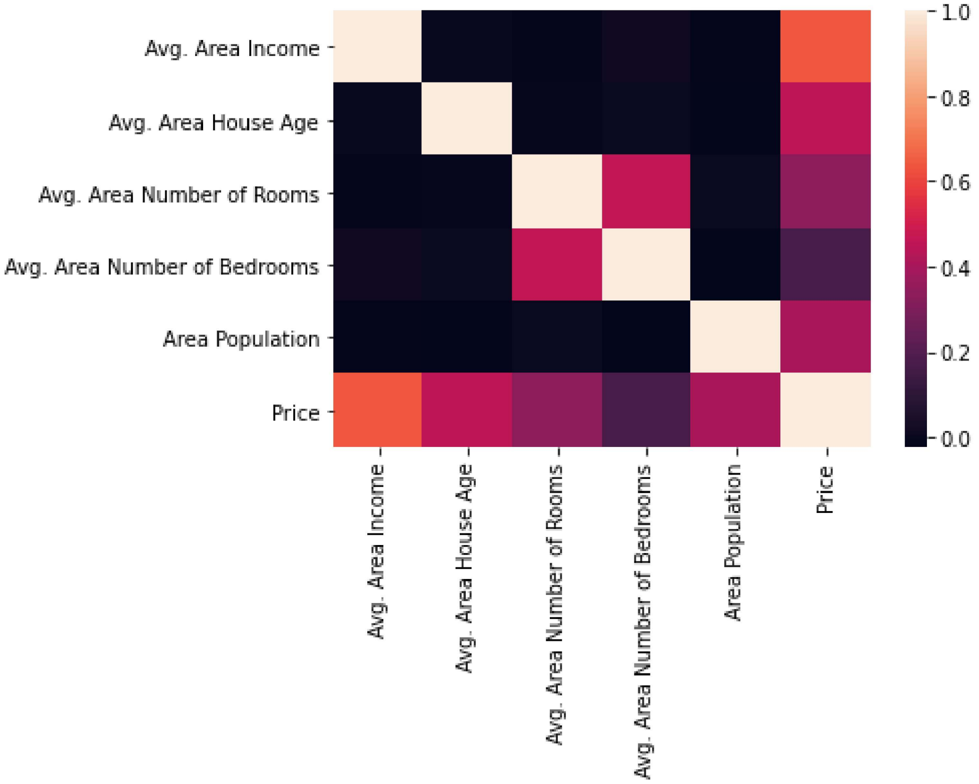
Out[16]:

| | 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 |
| **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 |
| **...** | ... | ... | ... | ... | ... | ... |
| **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 |
| **4998** | 68001.331235 | 5.534388 | 7.130144 | 5.44 | 42625.620156 | 1.198657e+06 |
| **4999** | 65510.581804 | 5.992305 | 6.792336 | 4.07 | 46501.283803 | 1.298950e+06 |

5000 rows × 6 columns

In [17]:
```
sns.heatmap(da.corr())
```

Out[17]:
```
<AxesSubplot:>
```



# To train the model-model building

We are gng to train linear regression model;we need to split out the date into two variables x and y where x is independent variable(input) and y is dependent on x(output)we could ignore addrress as it is not required for our model

In [29]:
```
x=da[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population']]
y=da['Price']
```

In [30]:
```
# to split my dataset into training and test data

from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
```

In [31]:
```
from sklearn.linear_model import LinearRegression

lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[31]:
```
LinearRegression()
```

In [32]:

```
print(lr.intercept_)
```

```
-2635801.6119889
```

In [33]:

```
coeff = pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[33]:

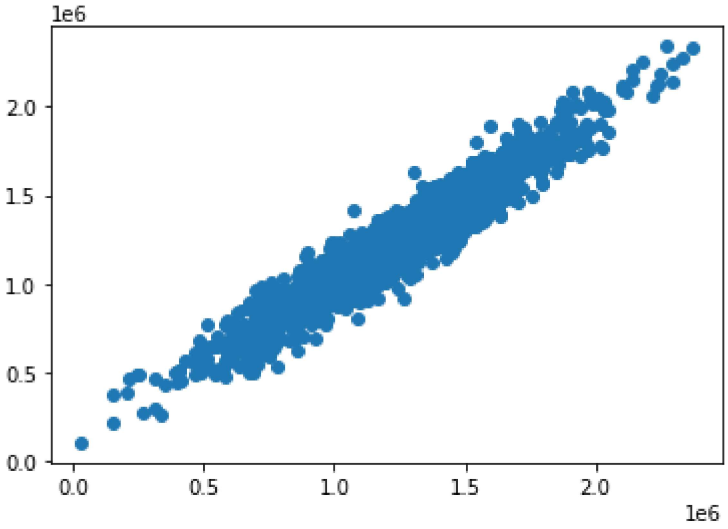
| | Co-efficient |
|---|---|
| **Avg. Area Income** | 21.547838 |
| **Avg. Area House Age** | 165399.074090 |
| **Avg. Area Number of Rooms** | 120409.007529 |
| **Avg. Area Number of Bedrooms** | 2292.767419 |
| **Area Population** | 15.259347 |

In [27]:

```
prediction=lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[27]:

```
<matplotlib.collections.PathCollection at 0x2a8cd5423a0>
```



In [28]:

```
print(lr.score(x_test,y_test))
```

```
0.9182950803369223
```

In [ ]:

```
 
```