PROBLEM STATEMENT ¶

A realestate agent want to help tp predict the house price for regions in USA .

DATA PROCESSING

In [4]:
```
#to import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [9]:
```
#to import dataset
data=pd.read_csv(r"C:\Users\user\Downloads\10_USA_Housing - 10_USA_Housing.csv")
data
```

Out[9]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Addres
0	79545.45857	5.682861	7.009188	4.09	23086.80050	1.059034e+06	208 Michael Ferry Ap 674\nLaurabury, N 3701.
1	79248.64245	6.002900	6.730821	3.09	40173.07217	1.505891e+06	188 Johnson View Suite 079\nLak Kathleen, CA.
2	61287.06718	5.865890	8.512727	5.13	36882.15940	1.058988e+06	9127 Elizabet Stravenue\nDanieltowr WI 06482.
3	63345.24005	7.188236	5.586729	3.26	34310.24283	1.260617e+06	USS Barnett\nFPO A 4482
4	59982.19723	5.040555	7.839388	4.23	26354.10947	6.309435e+05	USNS Raymond\nFP(AE 0938
...	...	...	...	...	...	...	.
4995	60567.94414	7.830362	6.137356	3.46	22837.36103	1.060194e+06	USNS Williams\nFP(AP 30153-765
4996	78491.27543	6.999135	6.576763	4.02	25616.11549	1.482618e+06	PSC 9258, Bo 8489\nAPO AA 42991 335
4997	63390.68689	7.250591	4.805081	2.13	33266.14549	1.030730e+06	4215 Tracy Garde Suite 076\nJoshualanc VA 01.
4998	68001.33124	5.534388	7.130144	5.44	42625.62016	1.198657e+06	USS Wallace\nFPO A 7331
4999	65510.58180	5.992305	6.792336	4.07	46501.28380	1.298950e+06	37778 George Ridge Apt. 509\nEast Holl NV 2.

5000 rows × 7 columns

In [11]:
```
#to display top 5 rows
data.head()
```

Out[11]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.45857	5.682861	7.009188	4.09	23086.80050	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.64245	6.002900	6.730821	3.09	40173.07217	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.06718	5.865890	8.512727	5.13	36882.15940	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.24005	7.188236	5.586729	3.26	34310.24283	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.19723	5.040555	7.839388	4.23	26354.10947	6.309435e+05	USNS Raymond\nFPO AE 09386

DATA CLEANING AND PREPROCESSING

In [12]:
```
#
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 [13]:
```
#to display summary of statistics(here to know min max value)
data.describe()
```

Out[13]:

	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%	61480.562390	5.322283	6.299250	3.140000	29403.928700	9.975771e+05
50%	68804.286405	5.970429	7.002902	4.050000	36199.406690	1.232669e+06
75%	75783.338665	6.650808	7.665871	4.490000	42861.290770	1.471210e+06
max	107701.748400	9.519088	10.759588	6.500000	69621.713380	2.469066e+06

In [15]:
```
#to display the column heading
data.columns
```

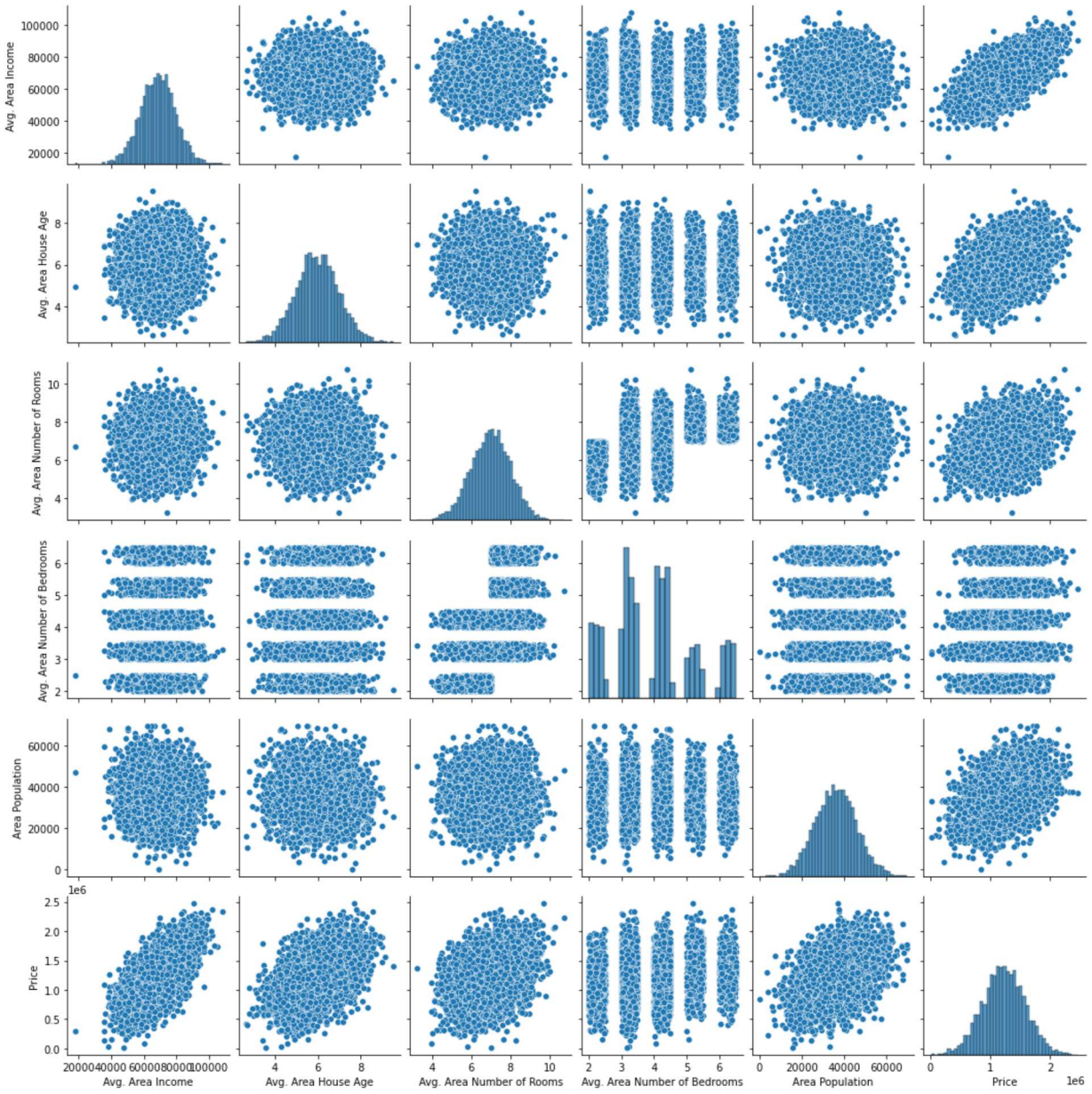
Out[15]:
```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Address'],
      dtype='object')
```

In [16]:
```
#here there is no missing values (identified through info() 5000 data are descrit
```

EDA and DATA VISUALIZATION

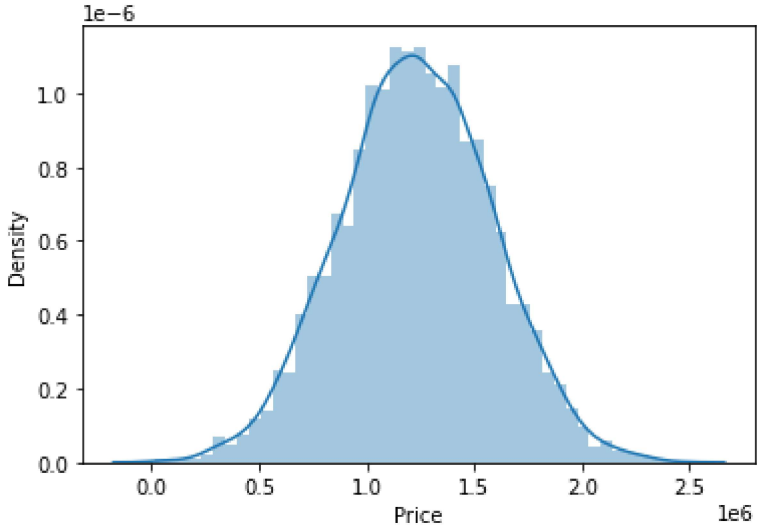
In [21]:
```
sns.pairplot(data)
```

Out[21]: <seaborn.axisgrid.PairGrid at 0x20c59dca490>



In [23]:
```
sns.distplot(data["Price"])
```

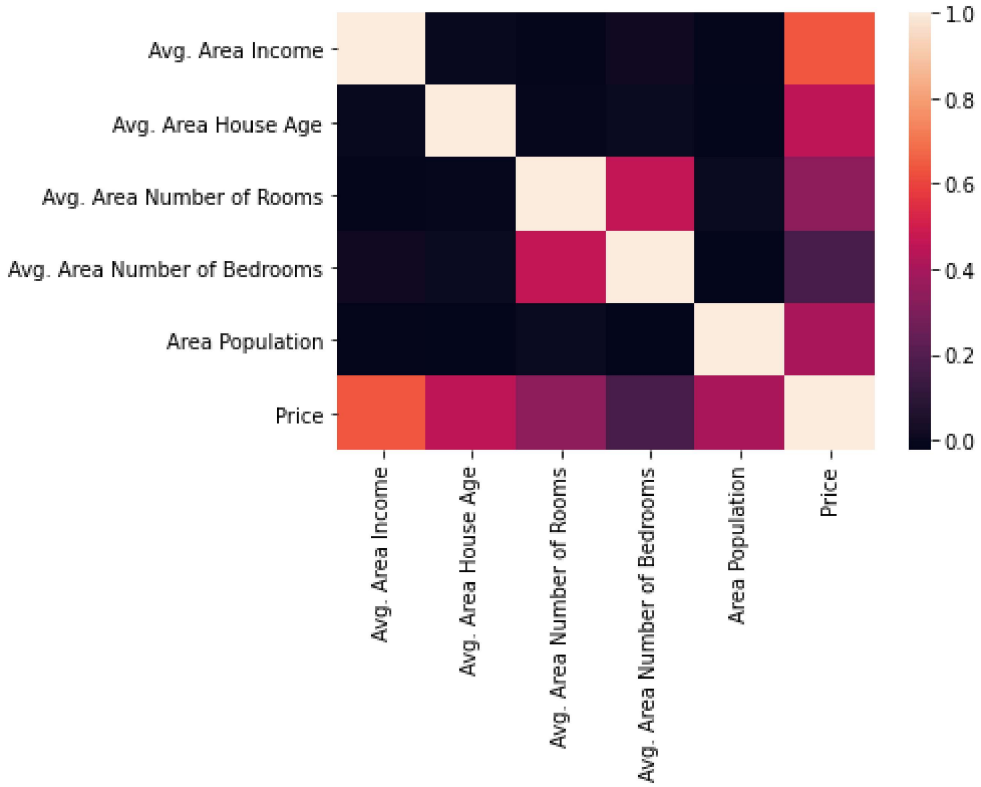
Out[23]: <AxesSubplot:xlabel='Price', ylabel='Density'>



In [24]:
```
df=data[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price']]
```

In [25]:
```
sns.heatmap(df.corr())
```

Out[25]: <AxesSubplot:>



TO TRAIN MODEL

MODEL BUILDING We are going to train linear regression model; we need to split out the data into two variables x and y where x is independent variables (input) and y is dependent on x(output) we could ignore address column as it is not required for our model

In [42]:
```
x=df[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population']]
y=df['Price']
```

In [43]:
```
#to split my dataset into trainning and test

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 [44]:
```
from sklearn.linear_model import LinearRegression

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

Out[44]:
```
LinearRegression()
```

In [45]:
```
#to find intercept
print(lr.intercept_)
```

```
-2648428.1571136247
```

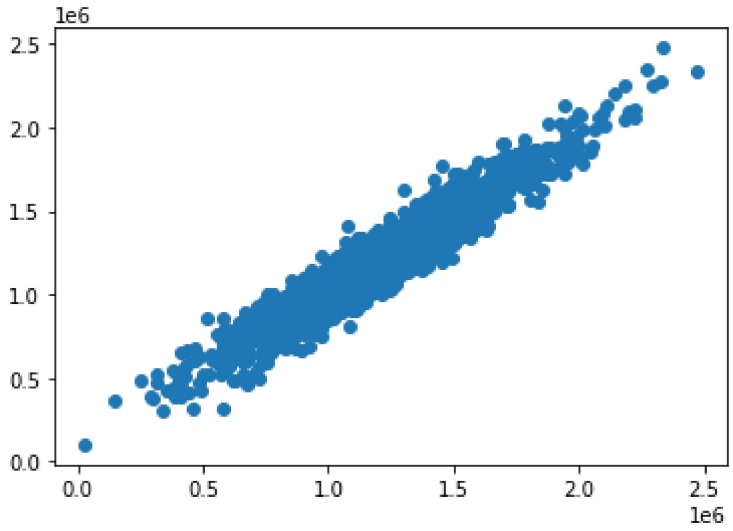
In [46]:
```
coeff = pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[46]:

	Co-efficient
Avg. Area Income	21.582037
Avg. Area House Age	166159.637150
Avg. Area Number of Rooms	121418.629535
Avg. Area Number of Bedrooms	1315.966575
Area Population	15.318730

In [48]:
```
prediction = lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[48]: <matplotlib.collections.PathCollection at 0x20c5dd28580>



In [50]:
```
print(lr.score(x_test,y_test))
```

```
0.9142656851538081
```

In [ ]: