HOUSE PRICE PREDICTION

Phase 1:Problem Definiton and Design Thinking

Problem Definiton:

The housing market is a pivotal sector that profoundly impacts individuals and families by representing one of the most substantial investments in their lifetimes. Accurate house price prediction is paramount to empower both buyers and sellers with the information they need to make well-informed decisions. This project aims to leverage machine learning techniques to predict house prices based on a comprehensive set of features, including but not limited to location, square footage, number of bedrooms and bathrooms, and other pertinent factors.

Design Thinking:

Collection Data and Integration:

We will collect and consolidate a diverse dataset comprising historical housing information, encompassing details such as property characteristics, transaction history, and location attributes. This dataset will serve as the foundation for our predictive model.

Data Preprocessing and Feature Engineering:

Prior to modeling, we will preprocess the data, addressing issues such as missing values, outliers, and feature scaling. Additionally, we will engineer new features to extract valuable insights and relationships from the raw data.

Model Development:

Employing state-of-the-art machine learning algorithms, including but not limited to linear regression, decision trees, and gradient boosting, we will develop predictive models capable of estimating house prices accurately.

Hyperparameter Tuning:

To optimize model performance, we will conduct hyperparameter tuning using techniques such as grid search or random search, ensuring that our models are fine-tuned to the specific characteristics of the data.

Model Evaluation:

We will rigorously evaluate the models using appropriate evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to assess their predictive accuracy.

Validation and Testing:

To measure the model's generalization capabilities, we will validate its performance on a holdout testing dataset, assessing its ability to make accurate predictions on previously unseen data.

Interpretability and Explainability:

In addition to predictive accuracy, we will aim to make our models interpretable and explainable, enabling stakeholders to understand the rationale behind the price predictions.

Ethical Considerations:

We will proactively address potential biases in the data to ensure fairness in predictions, avoiding discrimination related to factors such as race, gender, or socioeconomic status

Privacy and Security:

The handling of sensitive data, including personal information, will strictly adhere to privacy regulations and security best practices.

Deployment and Integration:

Once a satisfactory model is achieved, we will deploy it in a production environment where it can provide real-time house price predictions, potentially integrated into a user-friendly platform.

Continuous Monitoring and Maintenance:

Regular model updates, data refreshes, and performance monitoring will be conducted to ensure the model remains accurate and relevant over time.

Dataset:

The dataset used in the project is USA_Hosing. which is downloaded from kaagle. Kaagle is a subsidiary of Google, it is an online community of data scientists and machine learning engineers. Kaggle allows users to find datasets they want to use in building AI models, publish datasets, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. The link for the dataset is given below.

Dataset Link: **https://www.kaggle.com/datasets/vedavyasv/usa-housing**

Data Preprocessing:

Data preprocessing includes methods like data cleaning, data transformation, data reduction, handling imbalanced data, data visualization etc..

Model Training:

Feed the training data into the model and use an optimization algorithm (e.g., gradient descent) to update the model's parameters to minimize a loss function. This process involves iteratively adjusting the model's weights to make better predictions.

Coding:

#**Predicting House Prices using Machine Learning**

#Import Modules

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn import metrics
from sklearn.model_selection import train_test_split
%matplotlib inline
from sklearn.metrics import r2_score
import warnings
```

#Load the housing dataset

```
data=pd.read_csv('/content/USA_Housing.csv')

data
```

```
      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   

                                                Address  
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]
```

#Preprocess the data

```
data.head()

   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   

   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   

                                             Address  
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
```

```
data.shape
```

```
(5000, 7)
```

```
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')
```

```
data.isnull().sum
```

```
<bound method NDFrame._add_numeric_operations.<locals>.sum of
Avg. Area Income  Avg. Area House Age  Avg. Area Number of Rooms  \
0                 False                False                      False

1                 False                False                      False

2                 False                False                      False

3                 False                False                      False

4                 False                False                      False

...                 ...                  ...                        ...

4995              False                False                      False

4996              False                False                      False

4997              False                False                      False

4998              False                False                      False

4999              False                False                      False


      Avg. Area Number of Bedrooms  Area Population  Price  Address
0                            False            False  False    False
1                            False            False  False    False
2                            False            False  False    False
3                            False            False  False    False
4                            False            False  False    False
...                            ...              ...    ...      ...
4995                         False            False  False    False
4996                         False            False  False    False
4997                         False            False  False    False
```

```
4998                              False            False  False    False
4999                              False            False  False    False

[5000 rows x 7 columns]>

data.info

<bound method DataFrame.info of       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

                                                Address
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]>
```

```
data.describe
```

```
<bound method NDFrame.describe of       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

                                                Address
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]>
```

#Model Training

```
x=data.iloc[:, :-1]
x

      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

[5000 rows x 6 columns]
```

```
y=data.iloc[:, -1]
y
```

```
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...
Name: Address, Length: 5000, dtype: object
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,rando
m_state=4)
```

```
x_train
```

```
      Avg. Area Income  Avg. Area House Age  Avg. Area Number of Rooms
\
476       84439.855749             4.313978                   7.698765

2298      70689.364339             5.865246                   6.462900

3813      71068.996114             4.746896                   9.387913

4538      74497.673077             6.166026                   8.142658

1068      79575.641539             4.970709                   5.850243

...                ...                  ...                        ...

3671      67097.092120             6.086754                   7.211963

709       62357.030953             6.725271                   7.126592

2487      79687.761870             6.010368                   7.337394

174       83347.669697             5.468158                   5.475253
```

```
1146      65846.171039             6.385374                   6.804131


      Avg. Area Number of Bedrooms  Area Population         Price
476                           4.48     19835.247317  1.242422e+06
2298                          3.29     21350.099746  9.730686e+05
3813                          6.20     35724.018492  1.355557e+06
4538                          4.01     28160.457535  1.204753e+06
1068                          4.04     31050.102814  1.141917e+06
...                            ...              ...           ...
3671                          3.05     27191.506877  1.027428e+06
709                           5.00     23382.539386  9.724178e+05
2487                          6.09     20867.669885  1.360101e+06
174                           3.14     48226.718928  1.453382e+06
1146                          3.18     28214.363551  9.289500e+05

[4000 rows x 6 columns]
```

x_test

```
      Avg. Area Income  Avg. Area House Age  Avg. Area Number of Rooms
\
2175      66083.165901             4.213323                   8.908381

3156      56180.591431             6.201921                   7.180671

337       77733.731186             5.624500                   5.967832

444       47065.053303             5.767575                   7.266028

2334      71028.175896             3.895831                   6.623776

...                ...                  ...                        ...

1862      60288.475915             6.170239                   7.014315

1028      61839.767863             7.740113                   5.937847

4430      78529.527679             7.060888                   7.634762

3025      56505.827795             5.300534                   7.795375

1807      61264.201530             4.944536                   7.322068


      Avg. Area Number of Bedrooms  Area Population         Price
2175                          4.17     39185.995034  1.051173e+06
3156                          4.27     41036.284152  1.139073e+06
337                           3.23     32074.575986  1.236932e+06
444                           5.49     24125.875810  5.668962e+05
2334                          2.50     43922.630172  9.346104e+05
```

```
...                               ...          ...          ...
1862                             3.28     34651.072317  1.144938e+06
1028                             3.24     26064.820316  9.615391e+05
4430                             5.20     23897.116272  1.578087e+06
3025                             3.43     30995.488167  9.764817e+05
1807                             3.32     43208.356563  7.647561e+05

[1000 rows x 6 columns]
```

```
y_train
```

```
476          125 Jesse Spring\nNew Benjaminberg, NY 16741
2298                     PSC 7179, Box 6714\nAPO AA 57159
3813                       USNV Wright\nFPO AA 70734-4928
4538    03161 Lori Meadows Suite 563\nAndersonfurt, MT...
1068    4827 Kelsey Glen Suite 220\nMichaeltown, MD 34529
                              ...
3671    052 Thomas Square Apt. 034\nWrightmouth, OR 04272
709                      PSC 9682, Box 5865\nAPO AA 11465
2487    932 Schwartz Park Suite 892\nSouth Brian, CT 5...
174          0647 Ramirez Hill\nNew Crystalport, AZ 33060
1146    12315 Johnson Corners Suite 788\nWest Tyler, W...
Name: Address, Length: 4000, dtype: object
```

```
y_test
```

```
2175    54042 Proctor Corner Apt. 796\nNew Staceyville...
3156     121 Morris Rue Apt. 772\nWillisborough, NM 03840
337                672 Larson Ramp\nRobertside, NC 16903
444     006 Miller Orchard Suite 211\nPort Louis, WY 0...
2334            3757 Price Rue\nEast Colin, MD 62622-8672
                              ...
1862    0163 Samantha Coves Apt. 848\nPort Heidiville,...
1028                     PSC 9596, Box 0250\nAPO AE 81289
4430         2631 Ellis Walk\nSamanthatown, VT 51809-6834
3025               770 Cole Rest\nLunafurt, FL 70678-5139
1807        0995 Olivia Land Apt. 728\nAlexport, CA 92200
Name: Address, Length: 1000, dtype: object
```

```
x_train = x_train.iloc[:, 1:]
x_test = x_test.iloc[:, 1:]
```

```
x_train["Avg. Area Number of Rooms"].value_counts()
```

```
7.698765    1
7.253766    1
8.787825    1
7.768934    1
9.098980    1
           ..
4.242191    1
```

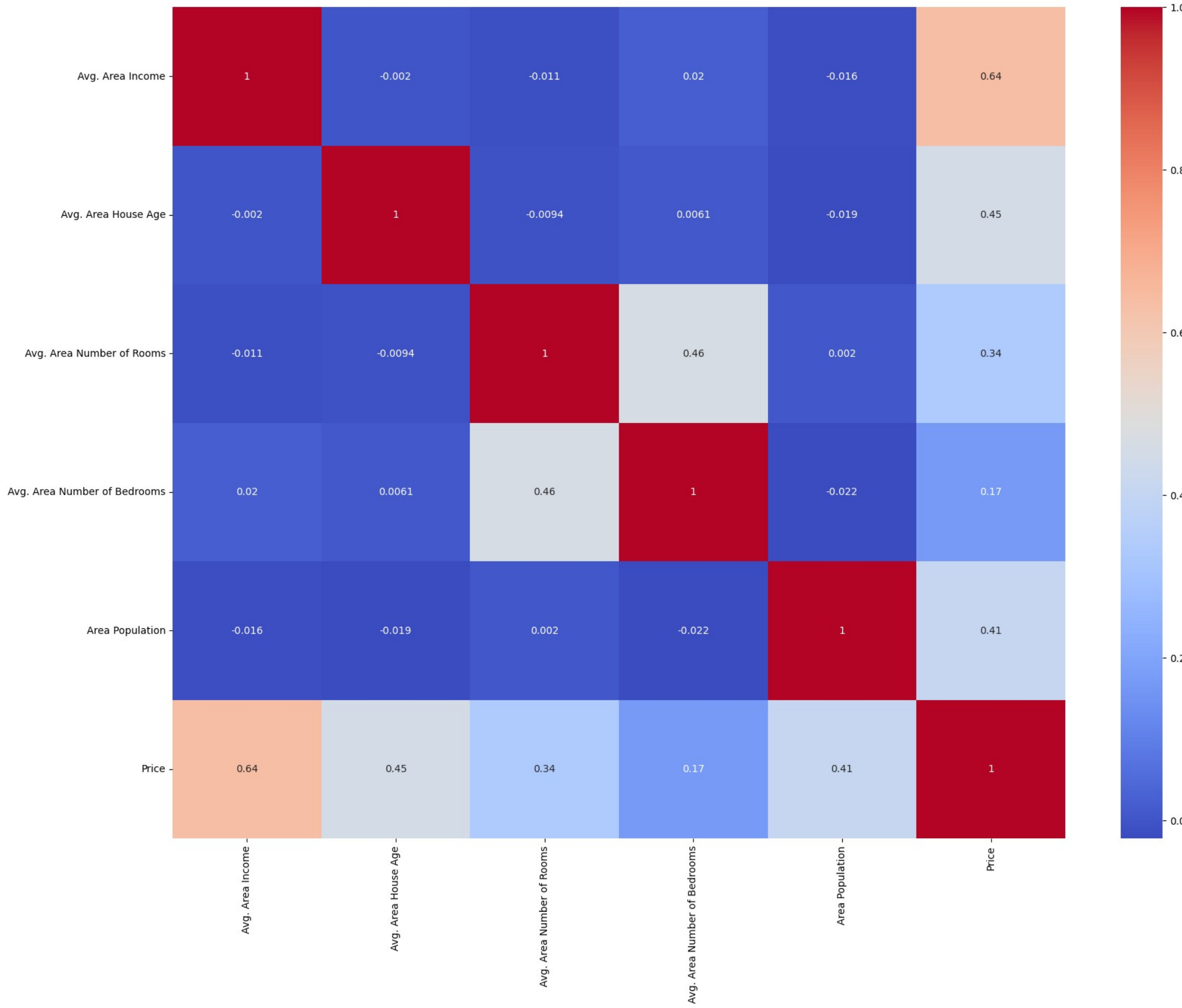
```
6.478152    1
8.025554    1
6.769326    1
6.804131    1
Name: Avg. Area Number of Rooms, Length: 4000, dtype: int64
```

```
make_train = x_train["Avg. Area Number of Rooms"].str.split(" ",
expand=True)
make_test = x_test["Avg. Area Number of Rooms"].str.split(" ",
expand=True)
```

#Visualize

```
plt.figure(figsize=(20, 15))
correlations = data.corr()
sns.heatmap(correlations, cmap="coolwarm", annot=True)
plt.show()
```

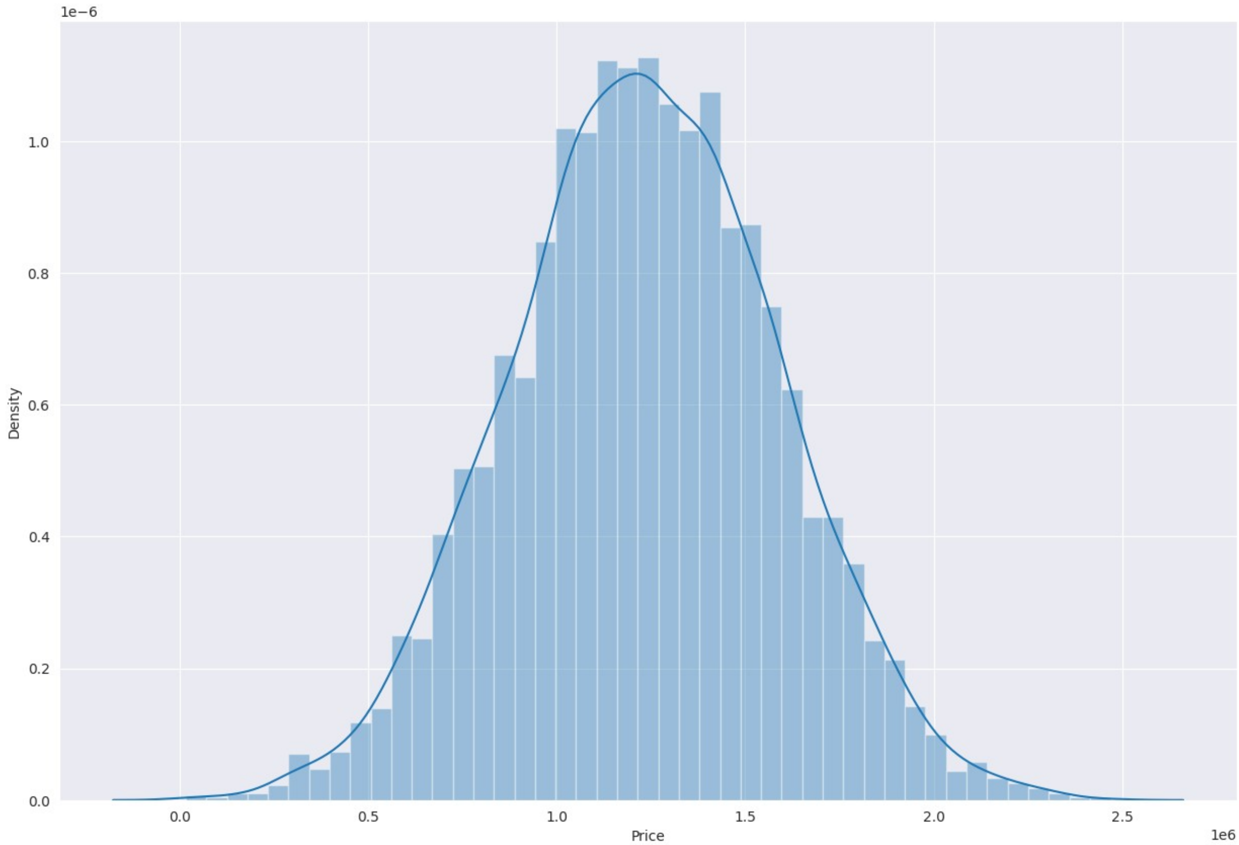
```
<ipython-input-23-555d4168b84a>:2: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  correlations = data.corr()
```



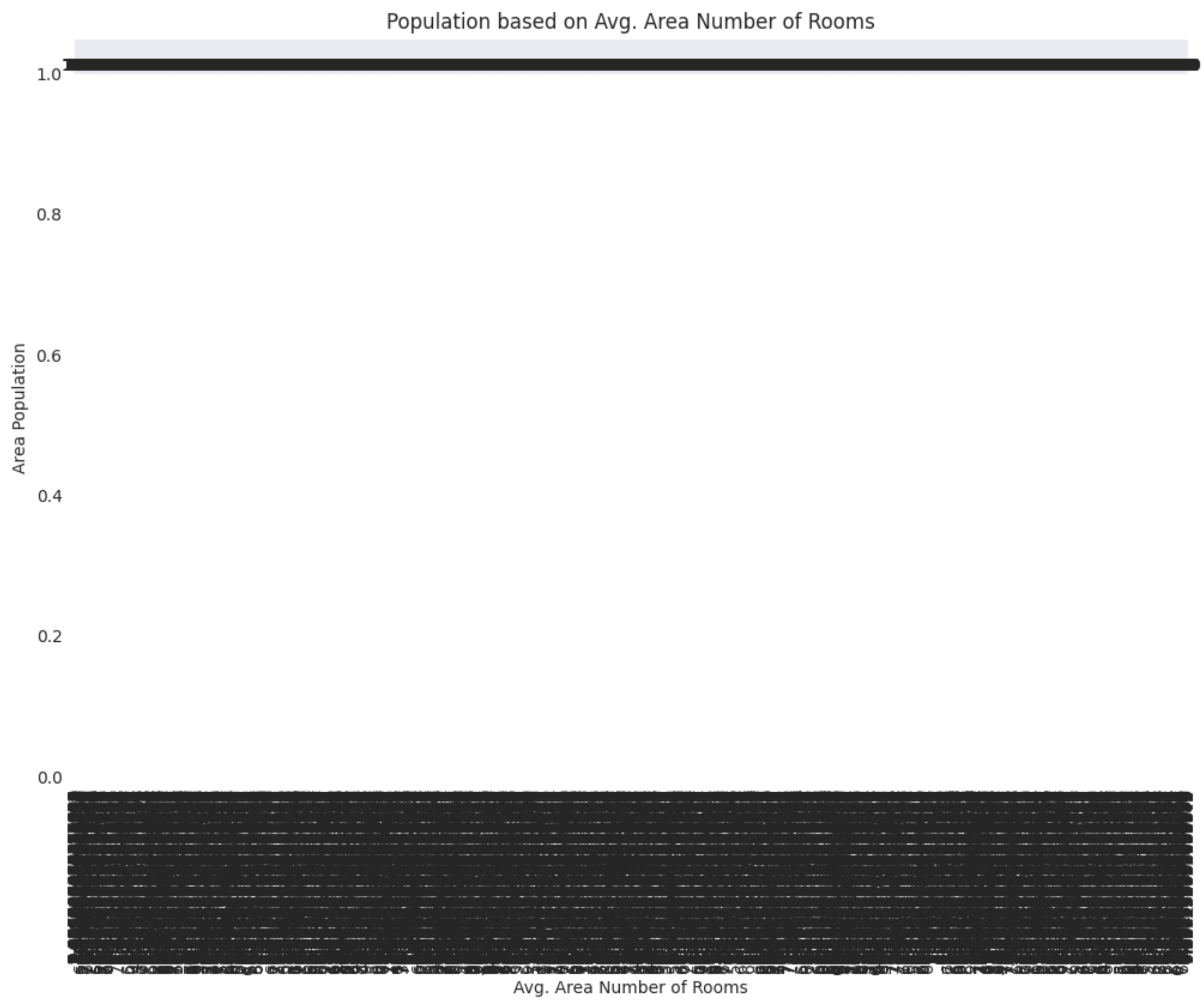

```
sns.set_style("darkgrid")
plt.figure(figsize=(15, 10))
sns.distplot(data.Price)
plt.show()
```

```
<ipython-input-30-a70e66c40847>:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn
v0.14.0.

Please adapt your code to use either `displot` (a figure-level
function with
similar flexibility) or `histplot` (an axes-level function for
histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751
```

```
sns.distplot(data.Price)
```



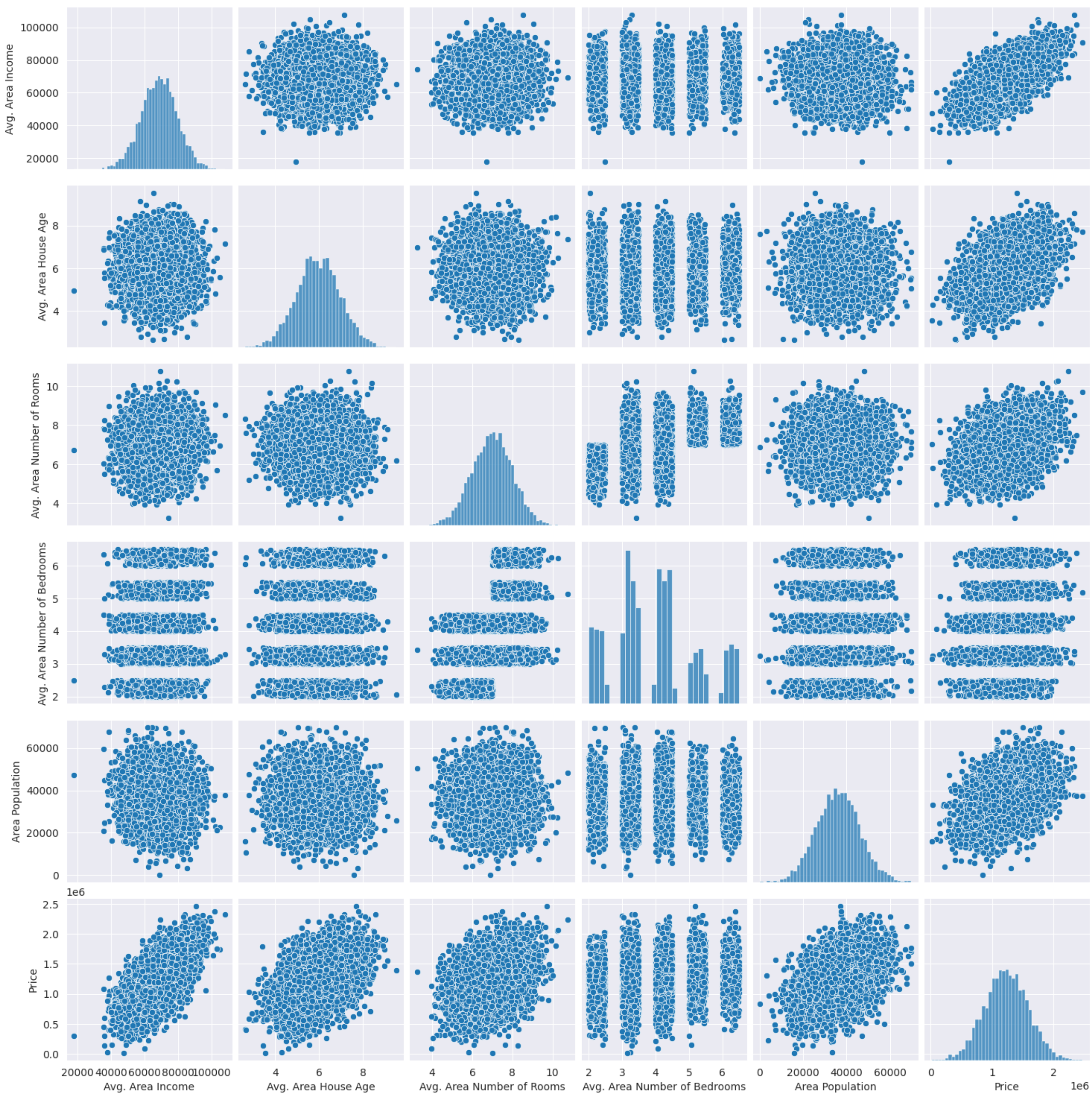

```
plt.figure(figsize=(12,8))
plot=sns.countplot(x='Avg. Area Number of Rooms',data=x_train)
plt.xticks(rotation=90)
for p in plot.patches:
    plot.annotate(p.get_height(),
                        (p.get_x() + p.get_width() / 2.0,
                         p.get_height()),
                        ha = 'center',
                        va = 'center',
                        xytext = (0, 5),
                        textcoords = 'offset points')

plt.title("Population based on Avg. Area Number of Rooms")
plt.xlabel("Avg. Area Number of Rooms")
plt.ylabel("Area Population")

Text(0, 0.5, 'Area Population')
```




```
missing_cols = set(x_train.columns) - set(x_test.columns)
for col in missing_cols:
    x_test[col] = 0
x_test = x_test[x_train.columns]
```

#Exploratory Data Analysis

```
sns.pairplot(data)

<seaborn.axisgrid.PairGrid at 0x7a999352eec0>
```

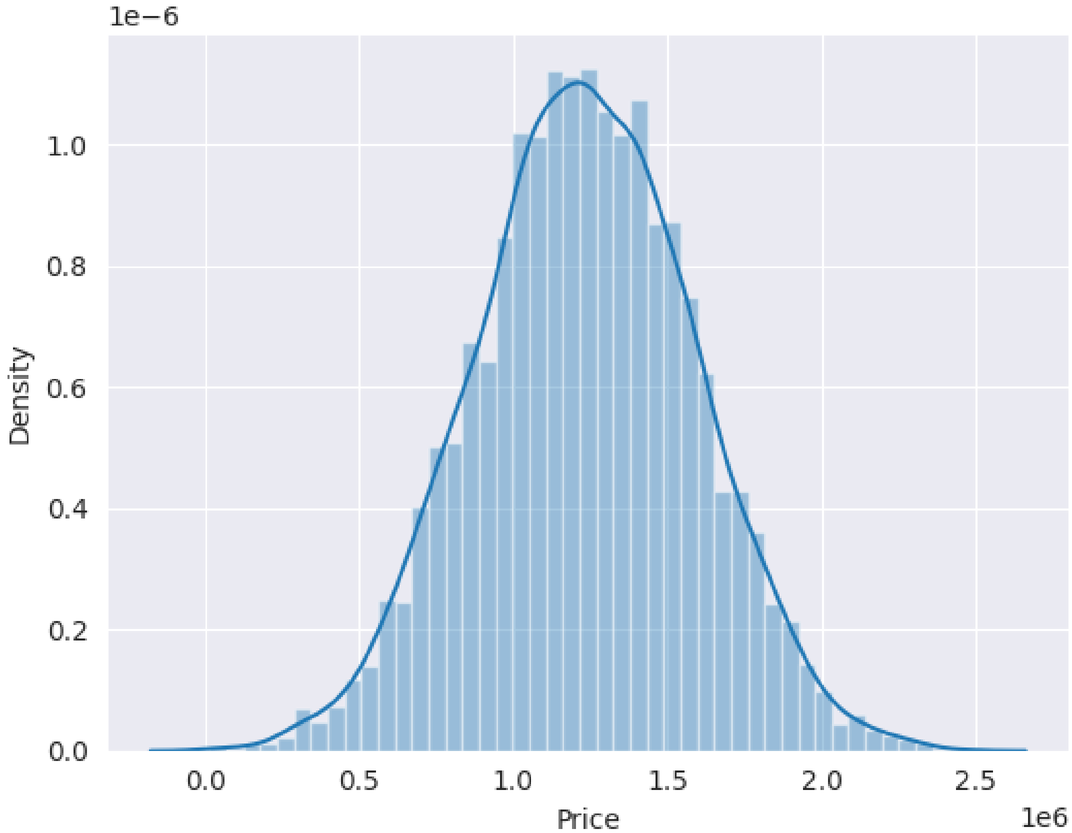



```
sns.distplot(data['Price'])
```

```
<ipython-input-32-049e7ab17fe1>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn
v0.14.0.

Please adapt your code to use either `displot` (a figure-level
function with
similar flexibility) or `histplot` (an axes-level function for
histograms).

For a guide to updating your code to use the new functions, please see
```

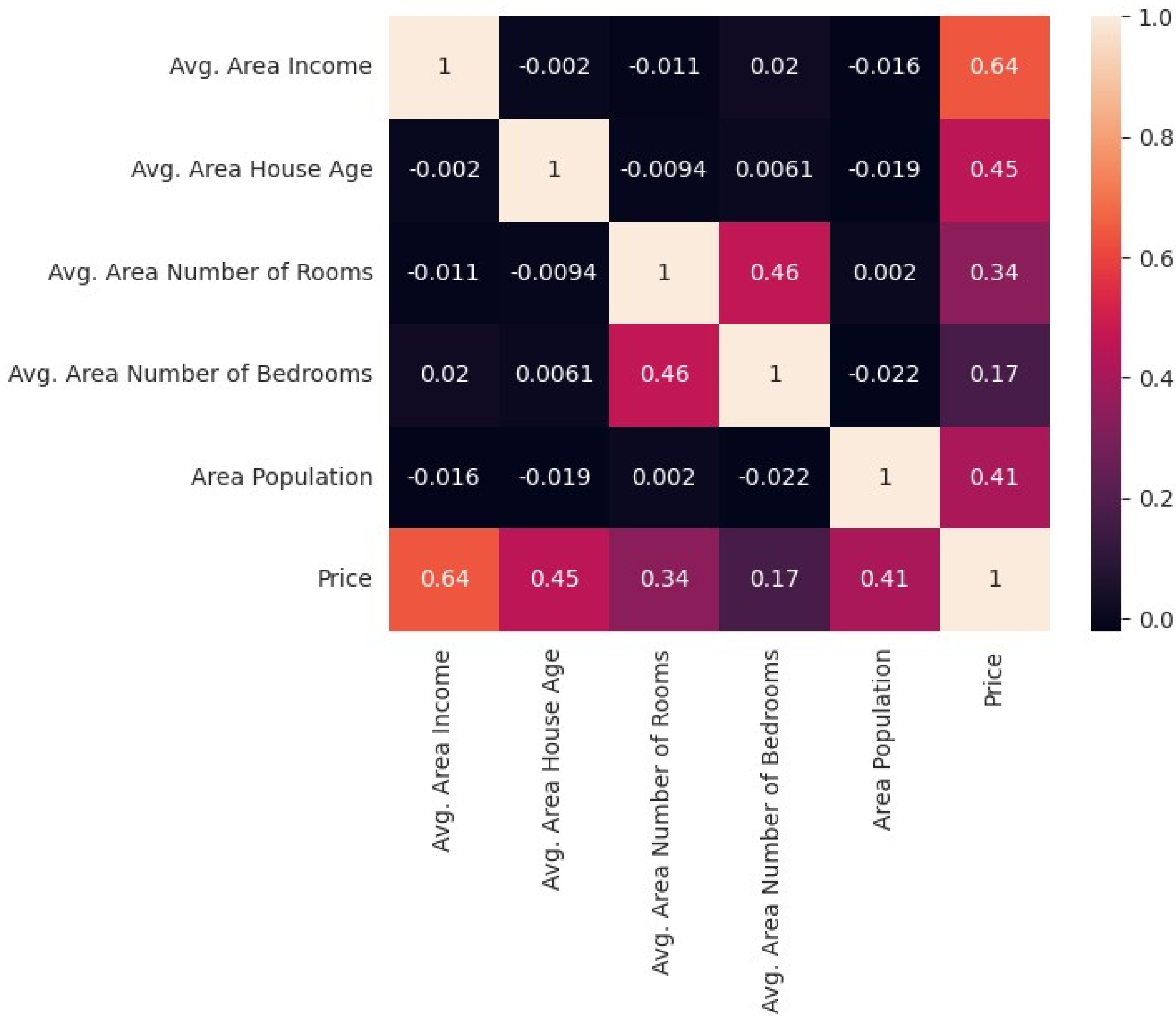
```
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data['Price'])
```

```
<Axes: xlabel='Price', ylabel='Density'>
```




```
sns.heatmap(data.corr(), annot=True)
```

```
<ipython-input-34-b699050ce883>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  sns.heatmap(data.corr(), annot=True)

<Axes: >
```



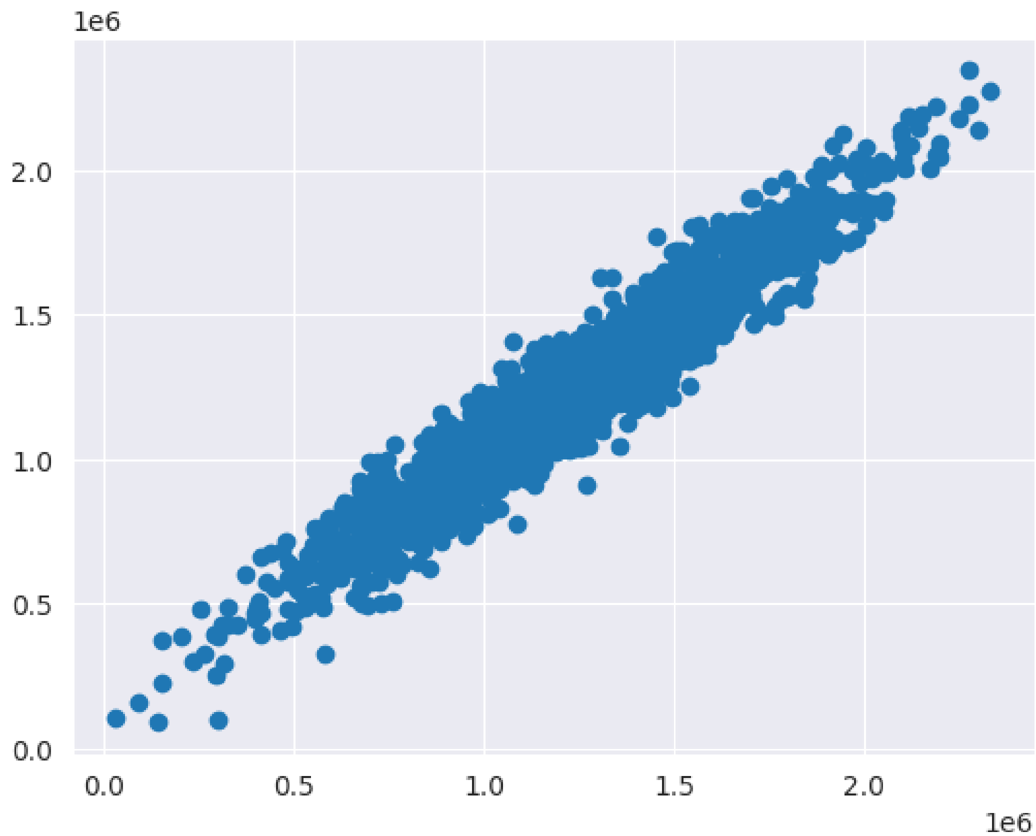
#Linear Regression

```
X = data[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number
of Rooms',
               'Avg. Area Number of Bedrooms', 'Area Population']]

y = data['Price']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.4, random_state=101)

lm = LinearRegression()

lm.fit(X_train,y_train)

LinearRegression()

print(lm.intercept_)
```

```
-2640159.7968526953

coeff_df = pd.DataFrame(lm.coef_,X.columns,columns=['Coefficient'])
coeff_df

                                 Coefficient
Avg. Area Income                   21.528276
Avg. Area House Age            164883.282027
Avg. Area Number of Rooms      122368.678027
Avg. Area Number of Bedrooms     2233.801864
Area Population                    15.150420
```

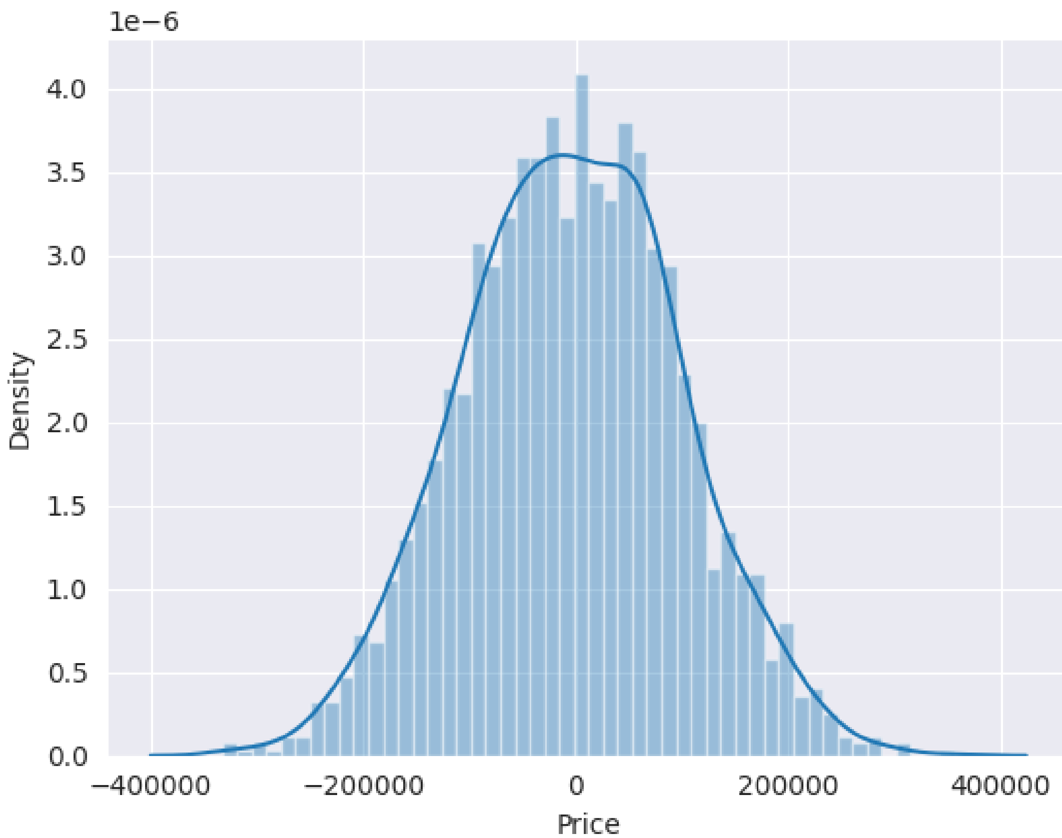
#Prediction from Linear Regression

```
predictions = lm.predict(X_test)

plt.scatter(y_test,predictions)

<matplotlib.collections.PathCollection at 0x7a9986443970>
```




```
sns.distplot((y_test-predictions),bins=50);

<ipython-input-49-5f2bc21c0ef7>:1: UserWarning:
```

```
`distplot` is a deprecated function and will be removed in seaborn
v0.14.0.

Please adapt your code to use either `displot` (a figure-level
function with
similar flexibility) or `histplot` (an axes-level function for
histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot((y_test-predictions),bins=50);
```



#Regression Evaluation Metrics

```python
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test,
predictions)))
```

```
MAE: 82288.22251914942
MSE: 10460958907.208977
RMSE: 102278.82922290897
```

By the representation of above predicted models,the data are accurately predicted