TITLE: Predicting House Prices using Machine Learning

Phase 3: Development Part 1

Start building the house price prediction model by loading and preprocessing the dataset.

Importing Dependencies:

linkcode

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import r2_score,
mean_absolute_error,mean_squared_error

from sklearn.linear_model import LinearRegression

from sklearn.ensemble import RandomForestRegressor

import xgboost as xg%matplotlib inline

import warnings

warnings.filterwarnings("ignore")
```

```
/opt/conda/lib/python3.10/site-packages/scipy/__init__.py:146: UserWarning: A NumPy version
>=1.16.5 and <1.23.0 is required for this version of SciPy (detected version 1.23.5

  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

Loading Dataset:

Dataset = https://www.kaggle.com/datasets/vedavyasv/usa.housing

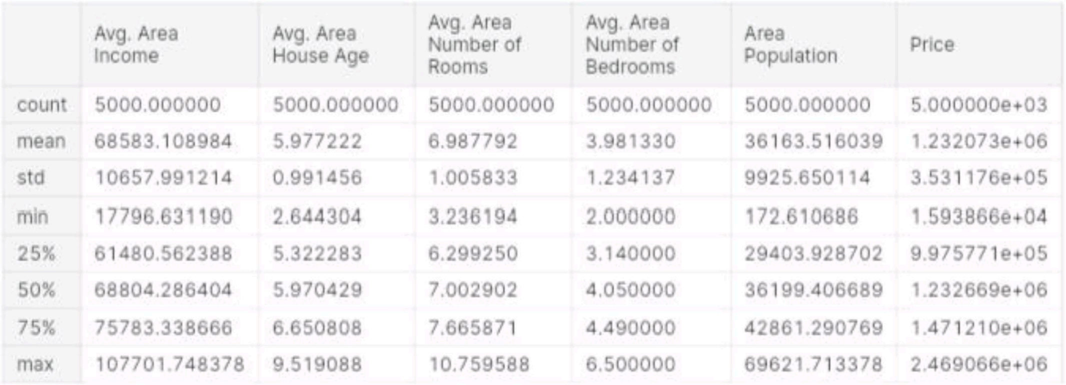
Data Exploration:

```
dataset.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
```

dataset.describe():



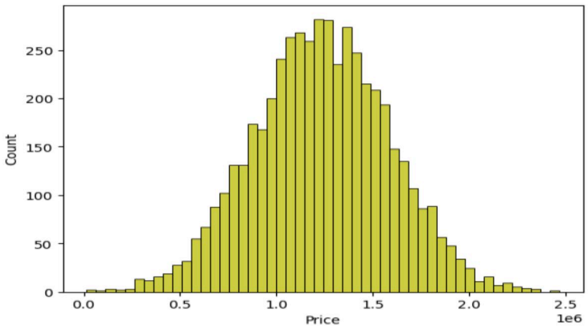
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	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

dataset.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')
```

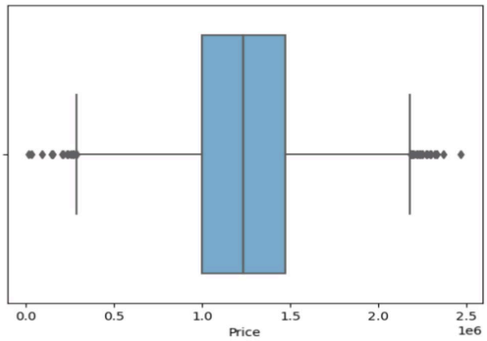
Visualisation and Pre-Processing of Data:

```
Sns.histplot(dataset, x=’Price’, bins=50, color=’y’)
<Axes: xlabel=’Price’, ylabel=’Count’>
```



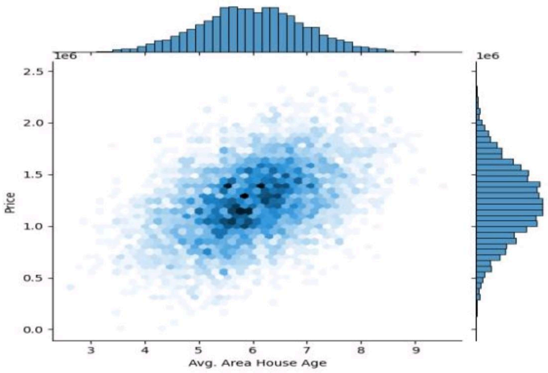
```
Sns.boxplot(dataset, x=’Price’,  palette=’Blues’)
```

```
<Axes: xlabel=’Price’>
```



```
sns.jointplot(dataset, x=’Avg . Area House Age’, y=’Price’, kind=’hex’)
```

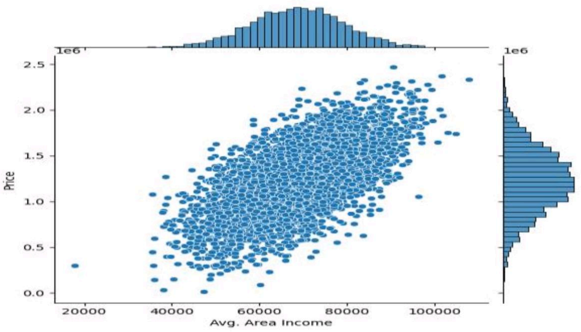
```
<seaborn.axisgrid.JointGrid at 0x7caf1d571810>
```



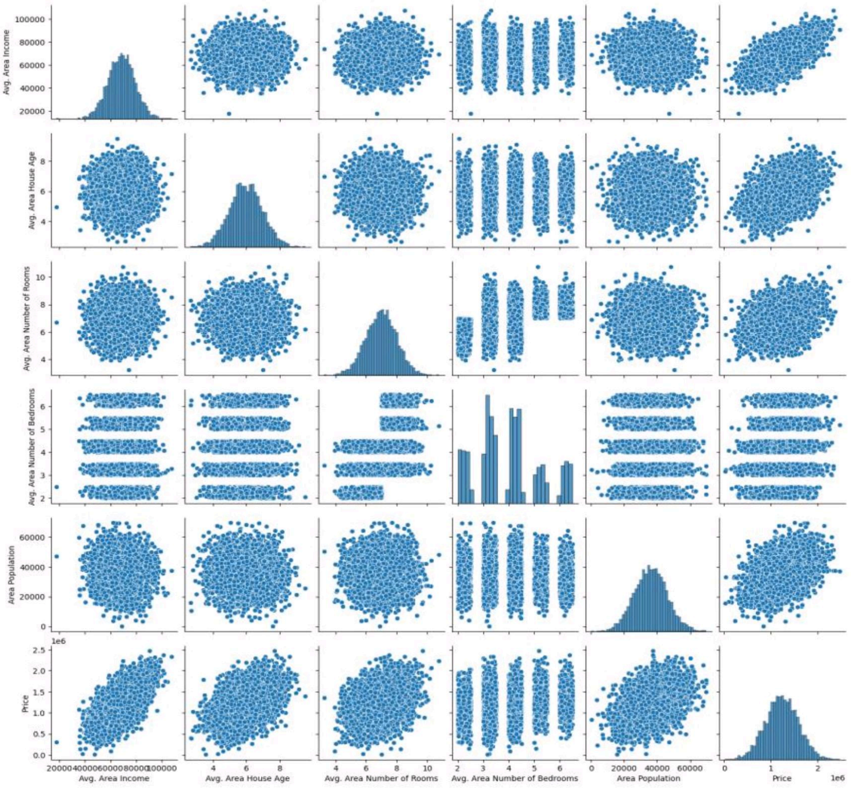
```
sns.jointplot(dataset, x=’Avg. Area Income’, y=’Price’)
```

```
<seaborn.axisgrid.JointGrid at 0x7caf1d8bf7f0>
```

```
plt.figure(figsize=(12,8))

sns.pairplot(dataset)
```

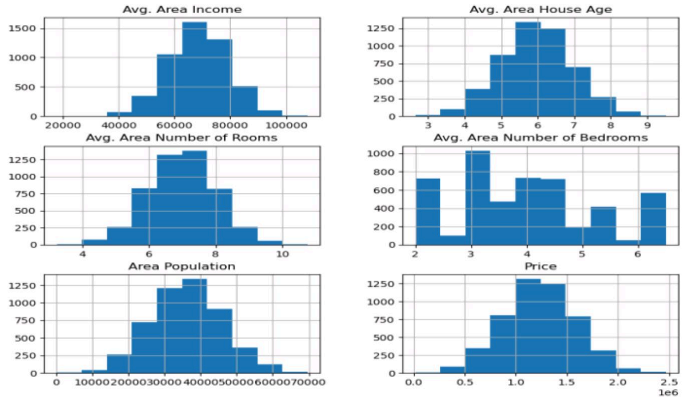


```
<seaborn.axisgrid.PairGrid at 0x7caf0c2ac550>

<Figure size 1200x800 with 0 Axes>

dataset.hist(figsize=(10,8))
```



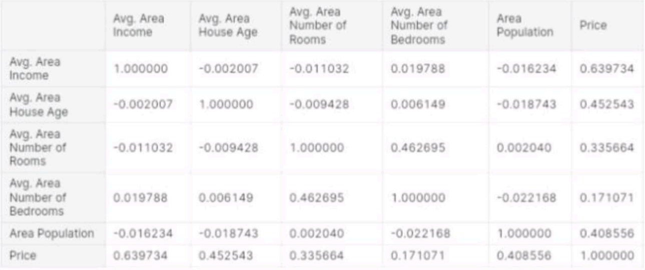
```
array([[<Axes: title={‘center’: ‘Avg. Area Income’}>,

        <Axes: title={‘center’: ‘Avg. Area House Age’}>],

       [<Axes: title={‘center’: ‘Avg. Area Number of Rooms’}>,

        <Axes: title={‘center’: ‘Avg. Area Number of Bedrooms’}>],

       [<Axes: title={‘center’: ‘Area Population’}>
```

<Axes: title={'center': 'Price'}>]], dtype=object)



Visualising Correlation:

Dataset.corr(numeric_only=True)

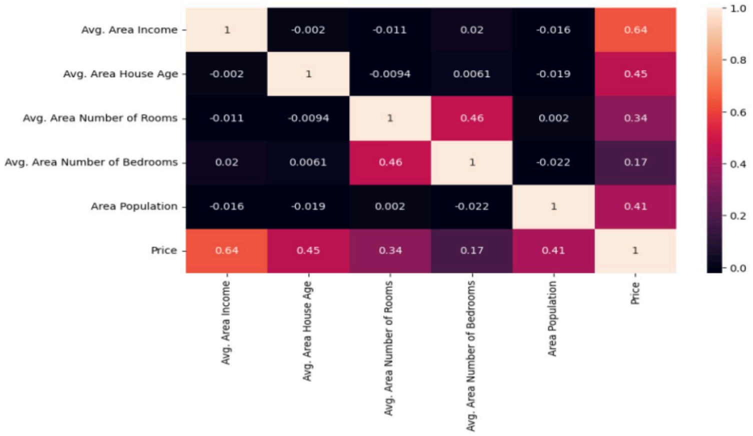


	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1.000000	-0.002007	-0.011032	0.019788	-0.016234	0.639734
Avg. Area House Age	-0.002007	1.000000	-0.009428	0.006149	-0.018743	0.452543
Avg. Area Number of Rooms	-0.011032	-0.009428	1.000000	0.462695	0.002040	0.335664
Avg. Area Number of Bedrooms	0.019788	0.006149	0.462695	1.000000	-0.022168	0.171071
Area Population	-0.016234	-0.018743	0.002040	-0.022168	1.000000	0.408556
Price	0.639734	0.452543	0.335664	0.171071	0.408556	1.000000

Plt.figure(figsize=(10,5))

Sns.heatmap(dataset.corr(numeric_only = True), annot=True

<Axes: >



	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1	-0.002	-0.011	0.02	-0.016	0.64
Avg. Area House Age	-0.002	1	-0.0094	0.0061	-0.019	0.45
Avg. Area Number of Rooms	-0.011	-0.0094	1	0.46	0.002	0.34
Avg. Area Number of Bedrooms	0.02	0.0061	0.46	1	-0.022	0.17
Area Population	-0.016	-0.019	0.002	-0.022	1	0.41
Price	0.64	0.45	0.34	0.17	0.41	1