Steps:

Import all the relavant libraries and modules

- **EDA**
  - Basic EDA
  - Missing Values
  - Duplicate Values
  - Data Visualization
  - Outliers/Anomalies Detection
  - Feature Encoding
  - Feature Selection

- **Model Building**
  - Separate your Independent and Dependent data
  - Split your data into train and test
  - Model Selection
  - Model Training
  - Model Prediction
  - Model Evaluation

In [ ]:

1

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

import sklearn
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor

from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

import warnings
warnings.filterwarnings('ignore')
```

In [ ]:

Load the dataset

In [3]:

```
df = pd.read_csv('USA_Housing.csv')
df.head()
```

Out[3]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

In [ ]:

```
df.shape
```

Out[4]: (5000, 7)

In [ ]:

Basic EDA

In [5]:

```
df.shape
```

Out[5]: (5000, 7)

In [6]:

```
df.columns
```

Out[6]:
```
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 [8]:

```
df.head(3)
```

Out[8]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...

In [ ]:

```
7       df.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 [ ]:

In [9]:
```
df.describe()
```

Out[9]:

	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 [ ]:

Missing Values

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
Avg. Area Income              0
Avg. Area House Age           0
Avg. Area Number of Rooms     0
Avg. Area Number of Bedrooms  0
Area Population               0
Price                         0
Address                       0
dtype: int64
```

In [11]:
```
df.isnull().mean()
```

Out[11]:
```
Avg. Area Income              0.0
Avg. Area House Age           0.0
Avg. Area Number of Rooms     0.0
Avg. Area Number of Bedrooms  0.0
Area Population               0.0
Price                         0.0
Address                       0.0
dtype: float64
```

In [ ]:

Duplicate Values

In [13]:
```
df.duplicated().sum()
```

Out[13]:
```
0
```

In [ ]:

In [14]:
```
df[df.duplicated()]
```

Out[14]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address

In [15]:
```
df.drop_duplicates(keep='first', inplace=True)
```

In [ ]:

Outliers/Anomalies Detection

Using Boxplot

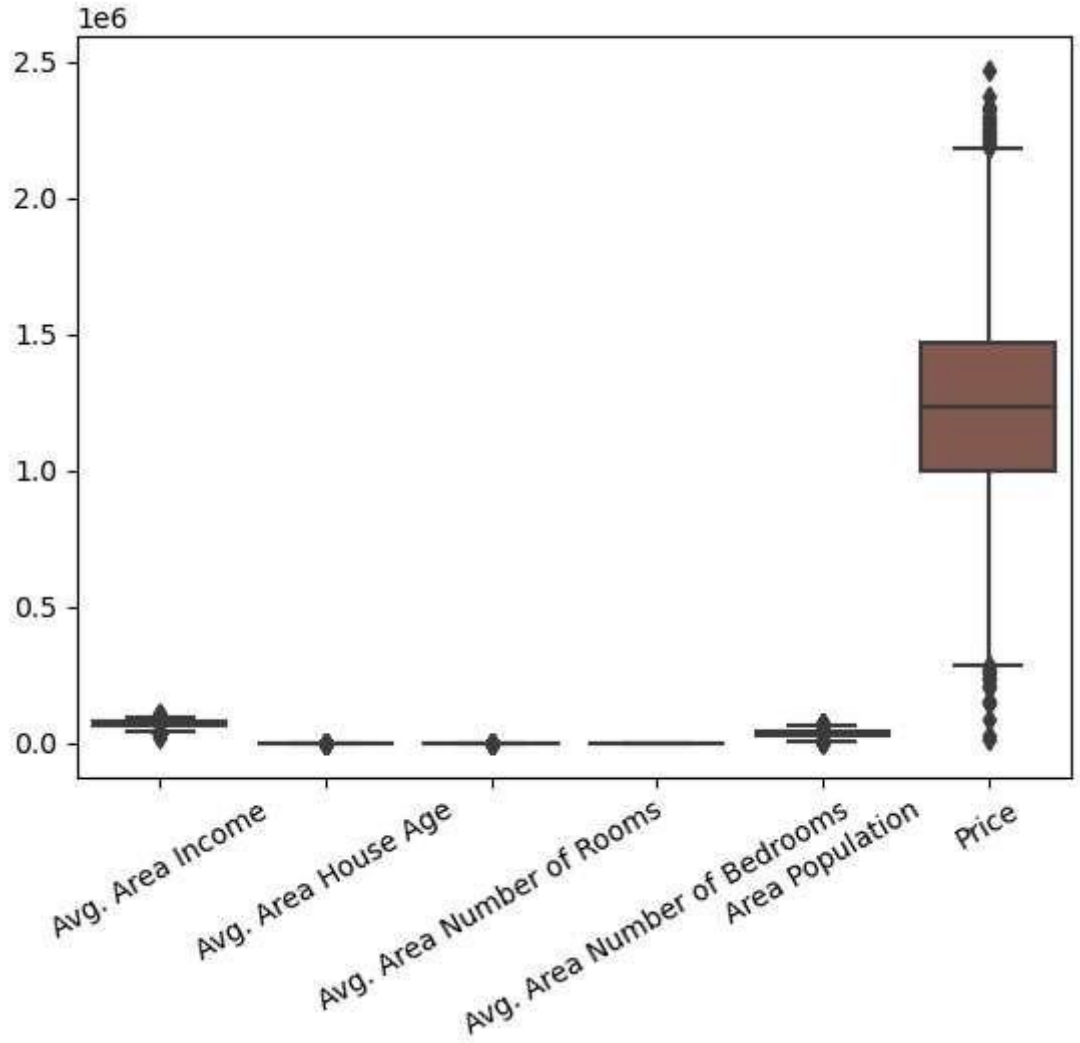
In [17]:
```
df.columns
```

Out[17]:
```
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 [ ]:

20

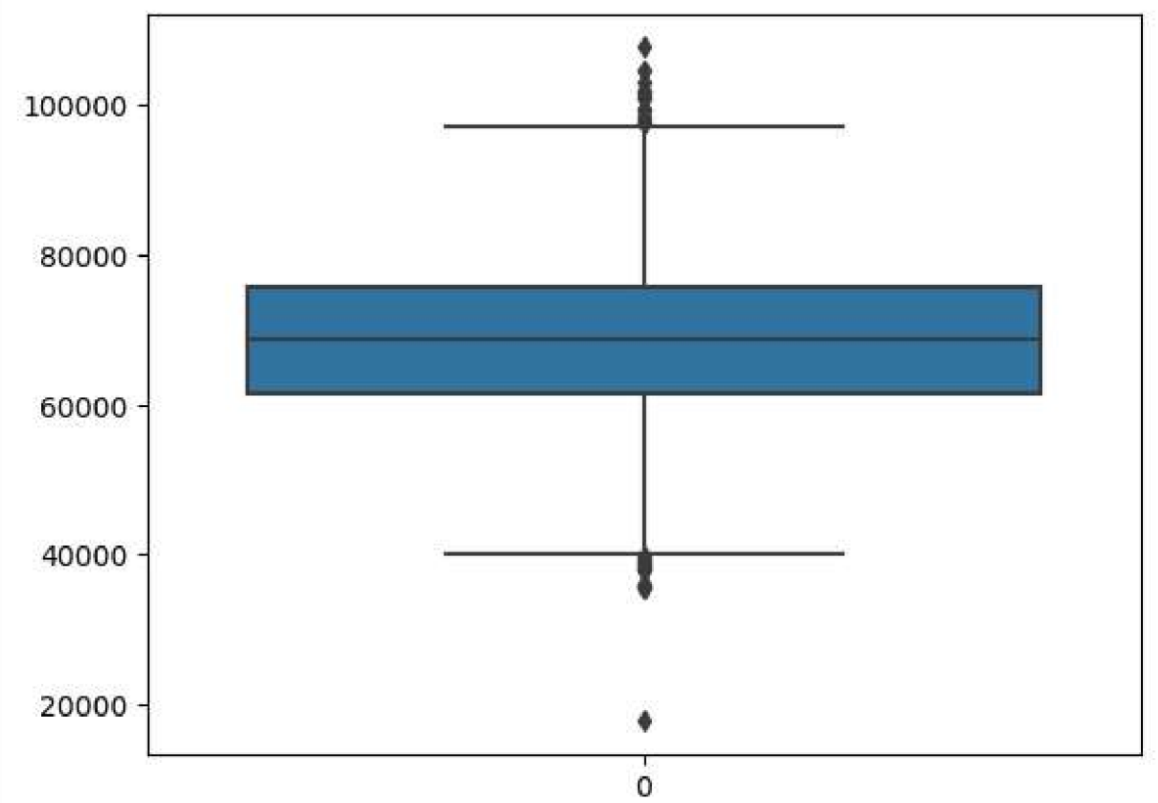
```
sns.boxplot(df)
plt.xticks(rotation=30)
plt.show()
```



In [ ]:

In [ ]:

22
```
sns.boxplot(df['Avg. Area Income'])
plt.show()
```



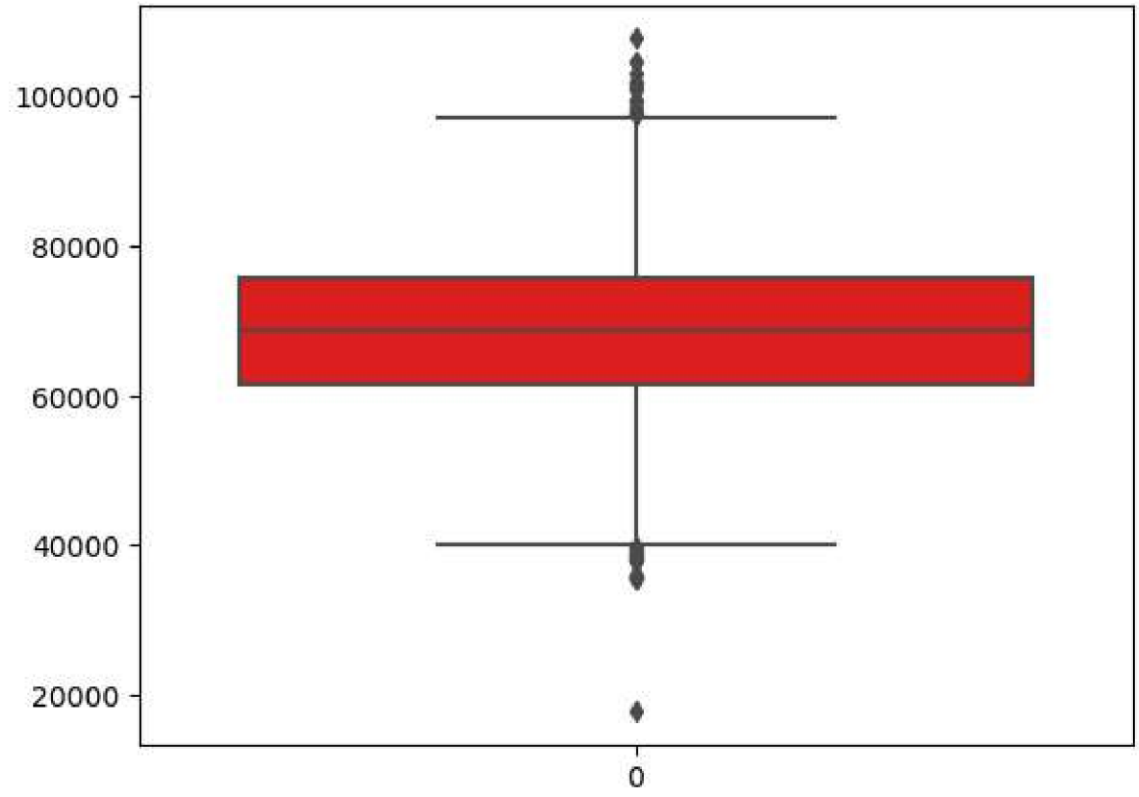
In [23]:
```
df.columns
```

Out[23]:
```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Address'],
      dtype='object')
```

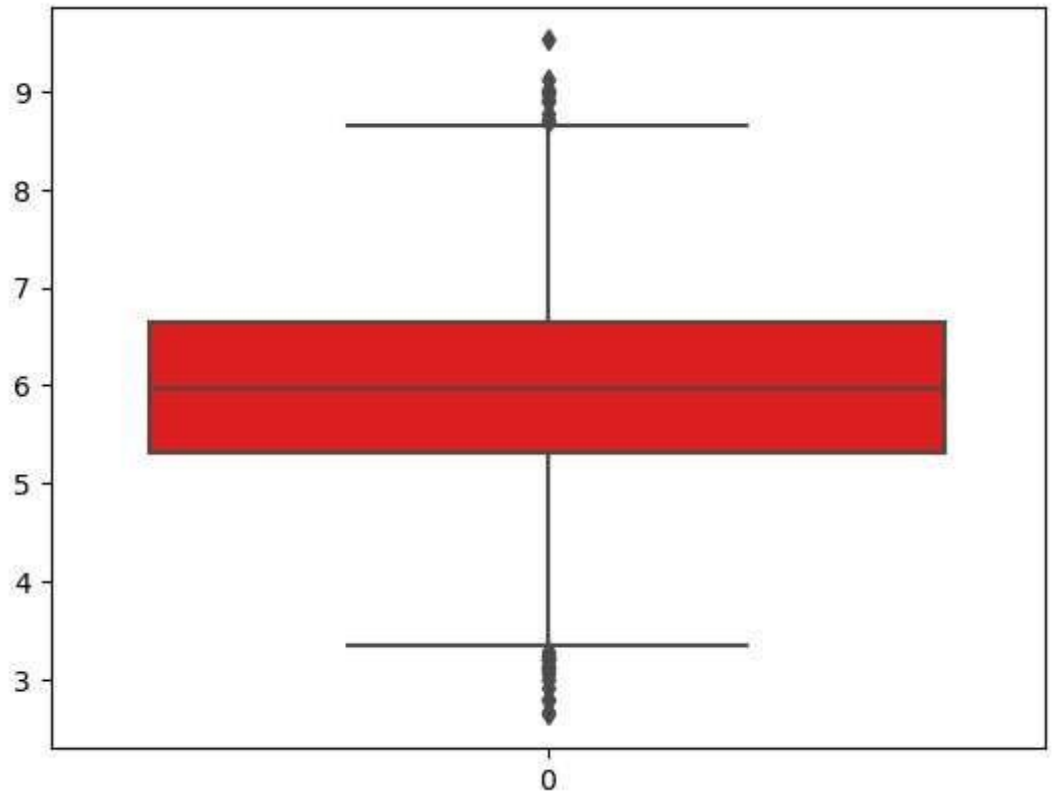
In [31]:

```
for i in df.columns:
    if i in ['Price', 'Address']:
        pass
    else:
        print(f'------------------------{i}-------------------------')
        sns.boxplot(df[i], color='r')
        plt.show()
```

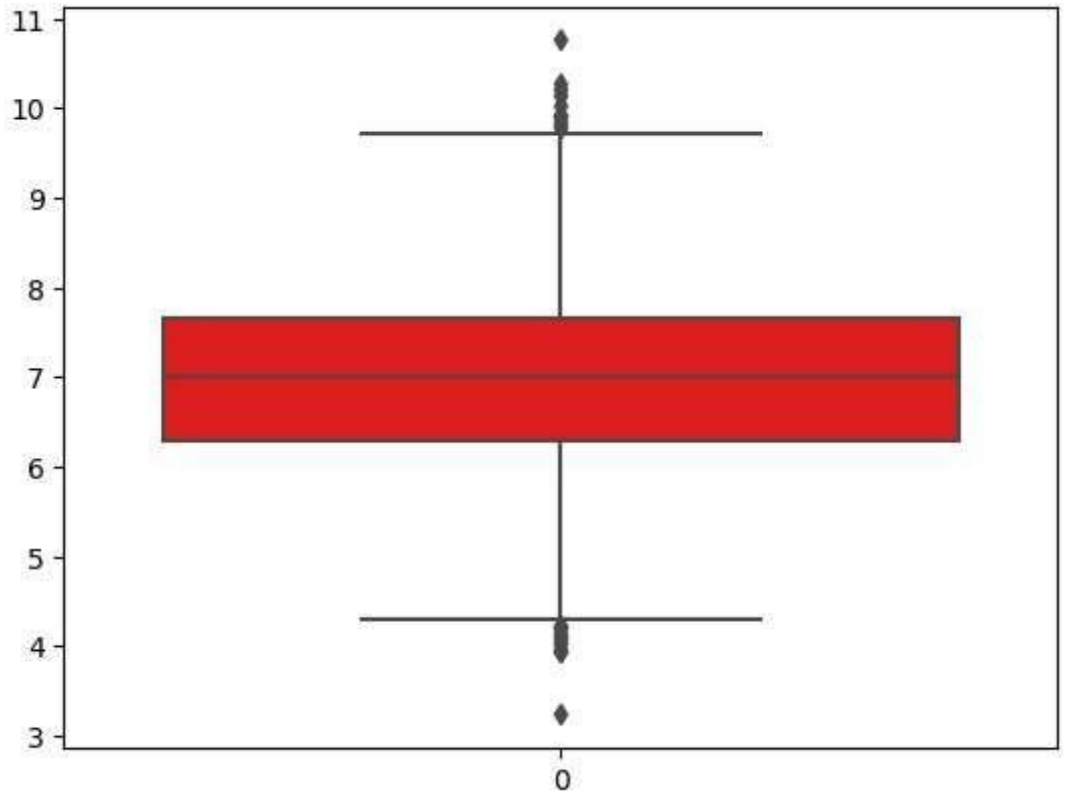
```
------------------------Avg. Area Income-------------------------
```



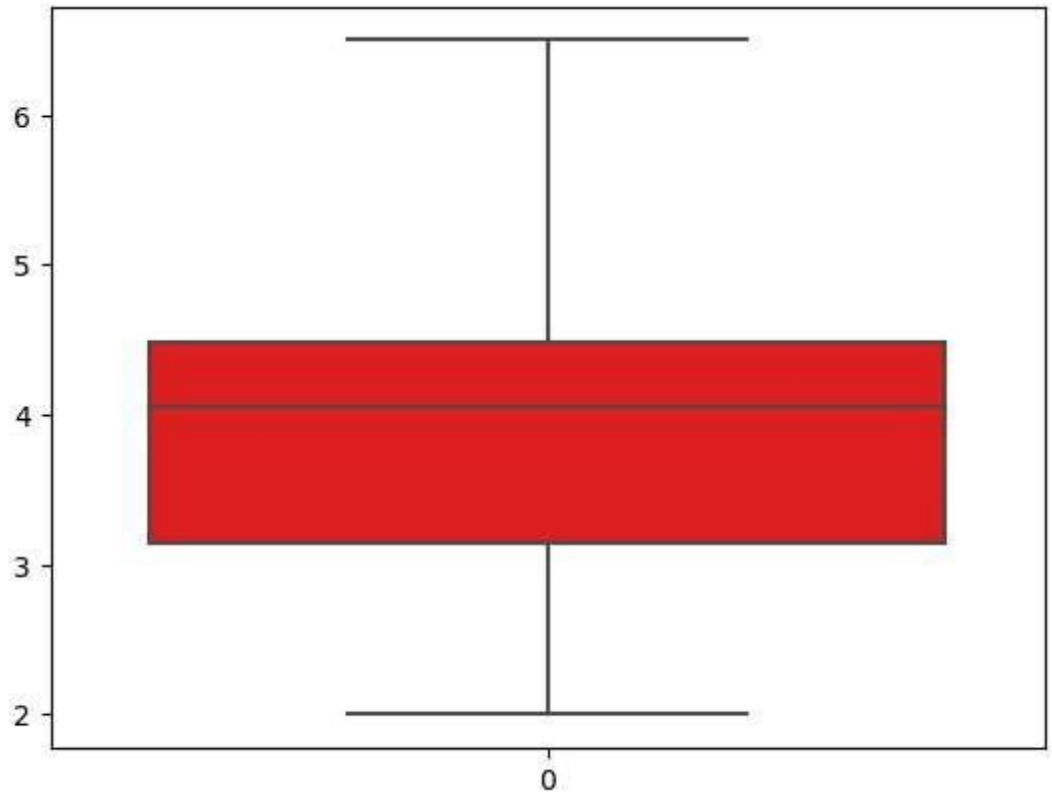

```
------------------------Avg. Area House Age-------------------------
```



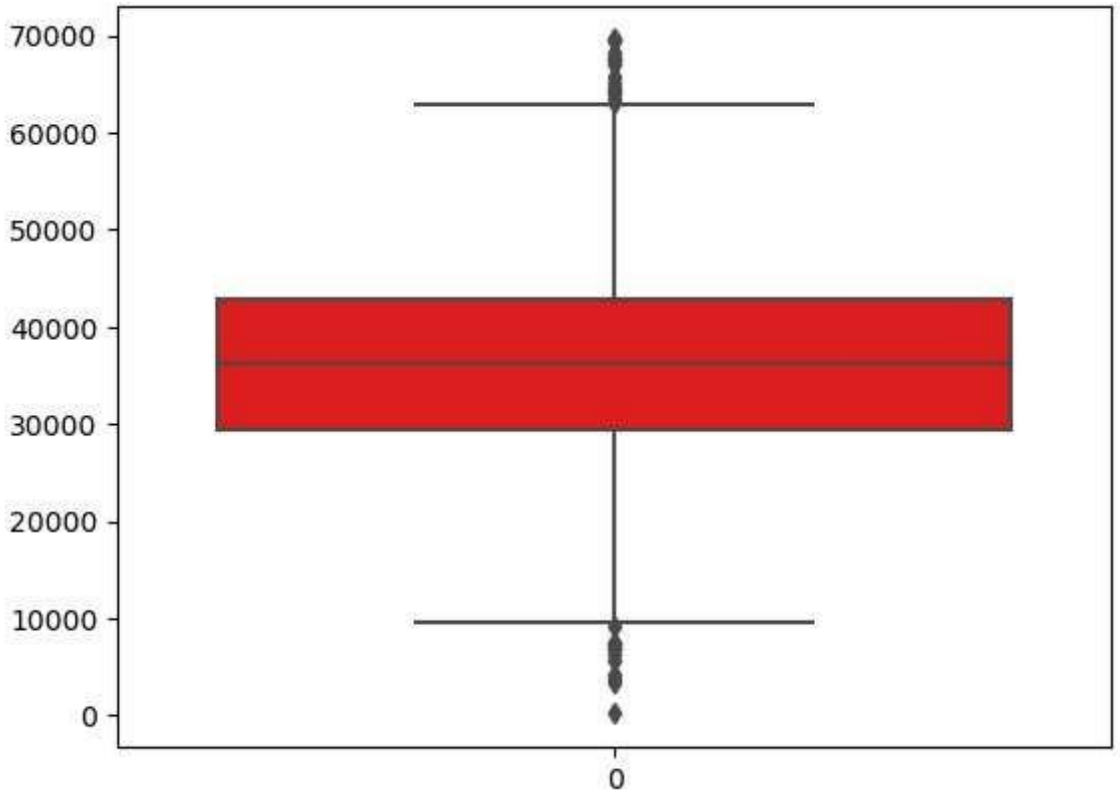
----------Avg. Area Number of Rooms----------



----------Avg. Area Number of Bedrooms----------



----------Area Population----------



In [ ]:

Using IQR

In [ ]:

In [ ]:

In [ ]:

```
Data Visualization
Feature Encoding
Feature Selection
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: