Linear Regression

```
# Importing Libraries

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

```
# Reading the dataset
df=pd.read_csv('https://raw.githubusercontent.com/Deepsphere-AI/DataAnalyticsTraining/main/PredictiveAnalytics/housing.csv')
```

```
#Exploring dataset
df.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population
0	-122.23	37.88	41	880	129.0	322
1	-122.22	37.86	21	7099	1106.0	2401
2	-122.24	37.85	52	1467	190.0	496
3	-122.25	37.85	52	1274	235.0	558
4	-122.25	37.85	52	1627	280.0	565

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  int64  
 3   total_rooms         20640 non-null  int64  
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  int64  
 6   households          20640 non-null  int64  
 7   median_income       20640 non-null  float64
 8   ocean_proximity     20640 non-null  object 
 9   median_house_value  20640 non-null  int64  
dtypes: float64(4), int64(5), object(1)
memory usage: 1.6+ MB
```

```
print(df.size)
print(df.shape)
print(df.columns)
```

```
206400
(20640, 10)
Index(['longitude', 'latitude', 'housing_median_age', 'total_rooms',
       'total_bedrooms', 'population', 'households', 'median_income',
       'ocean_proximity', 'median_house_value'],
      dtype='object')
```

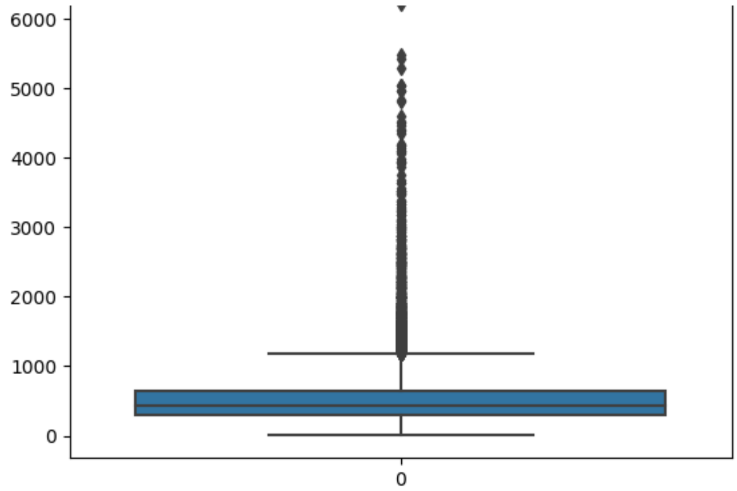
```
df.isna().sum()
```

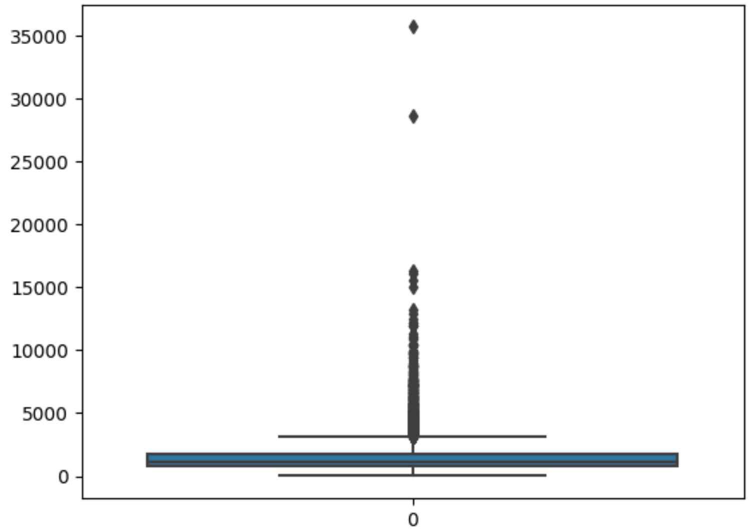
```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms        207
population              0
households              0
median_income           0
ocean_proximity         0
median_house_value      0
dtype: int64
```

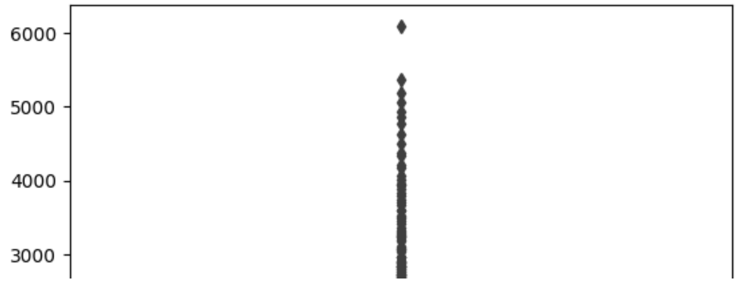
```
df.total_bedrooms.fillna(0)
```

```
0          129.0
1         1106.0
2          190.0
3          235.0
4          280.0
           ...
20635      374.0
20636      150.0
20637      485.0
20638      409.0
20639      616.0
Name: total_bedrooms, Length: 20640, dtype: float64
```

```
## Checking outliers
plt.show(sns.boxplot(df['latitude']))
plt.show(sns.boxplot(df['longitude']))
plt.show(sns.boxplot(df['housing_median_age']))
plt.show(sns.boxplot(df['median_house_value']))
plt.show(sns.boxplot(df['median_income']))
plt.show(sns.boxplot(df['total_bedrooms']))
plt.show(sns.boxplot(df['population']))
plt.show(sns.boxplot(df['households']))
plt.show(sns.boxplot(df['total_rooms']))
```


6000
5000
4000
3000
2000
1000
0
0


35000
30000
25000
20000
15000
10000
5000
0
0


6000
5000
4000
3000

```
df.drop(df[df['median_house_value']>500000].index,inplace=True)
df.drop(df[df['median_income']>8].index,inplace=True)
df.drop(df[df['total_bedrooms']>10000].index,inplace=True)
df.drop(df[df['total_rooms']>5000].index,inplace=True)
df.drop(df[df['households']>1000].index,inplace=True)
```

```
x=df.iloc[:,:1]
print(x.head())
y=df.iloc[:,1:]
print(y.head())
```

```
   longitude
2    -122.24
3    -122.25
4    -122.25
5    -122.25
6    -122.25
   latitude  housing_median_age  total_rooms  total_bedrooms  population  \
2     37.85                  52         1467           190.0         496
3     37.85                  52         1274           235.0         558
4     37.85                  52         1627           280.0         565
5     37.85                  52          919           213.0         413
6     37.84                  52         2535           489.0        1094

   households  median_income ocean_proximity  median_house_value
2         177         7.2574        NEAR BAY              352100
3         219         5.6431        NEAR BAY              341300
4         259         3.8462        NEAR BAY              342200
5         193         4.0368        NEAR BAY              269700
6         514         3.6591        NEAR BAY              299200
```

```
from sklearn import preprocessing

df['ocean_proximity']=preprocessing.LabelEncoder().fit_transform(df['ocean_proximity'])
df.ocean_proximity.unique()
```

```
array([3, 0, 1, 4, 2])
```

```
# plt.show(sns.boxplot(df['ocean_proximity']))
```