Data Analytics I Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset (https://www.kaggle.com/c/boston-housing (https://www.kaggle.com/c/boston-housing)). The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset. The objective is to predict the value of prices of the house using the given features.

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
#Step 2: Import the Boston Housing dataset
from sklearn.datasets import load_boston
boston = load_boston()
```

In [10]:
```
data = pd.DataFrame(boston.data)
```

In [28]:
```
data.shape
```

Out[28]:
```
(506, 14)
```

In [11]:
```
data.columns = boston.feature_names
data.head()
```

Out[11]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33

In [12]:
```
data['PRICE'] = boston.target
```

In [13]:
```
data.isnull().sum()
```

Out[13]:
```
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
PRICE      0
dtype: int64
```

In [14]:
```
x = data.drop(['PRICE'], axis = 1)
y = data['PRICE']
```

In [16]:
```
from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest =train_test_split(x, y, test_size =0.2,random_state = 0)
```

In [17]:
```
import sklearn
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
model=lm.fit(xtrain, ytrain)
```

In [25]:
```
lm.intercept_
```

Out[25]:
```
38.138692713393205
```

In [26]:
```
lm.coef_
```

Out[26]:
```
array([-1.18410318e-01,  4.47550643e-02,  5.85674689e-03,  2.34230117e+00,
       -1.61634024e+01,  3.70135143e+00, -3.04553661e-03, -1.38664542e+00,
        2.43784171e-01, -1.09856157e-02, -1.04699133e+00,  8.22014729e-03,
       -4.93642452e-01])
```

In [18]:
```
ytrain_pred = lm.predict(xtrain)
ytest_pred = lm.predict(xtest)
```

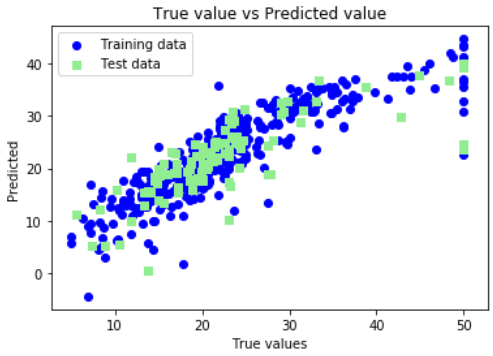
In [19]:
```
df=pd.DataFrame(ytrain_pred,ytrain)
df=pd.DataFrame(ytest_pred,ytest)
```

In [20]:
```
from sklearn.metrics import mean_squared_error, r2_score
mse = mean_squared_error(ytest, ytest_pred)
print(mse)
mse = mean_squared_error(ytrain_pred,ytrain)
print(mse)
```

```
33.4507089676911 85
19.330019357349375
```

In [21]:
```
mse = mean_squared_error(ytest, ytest_pred)
```

In [23]:
```
plt.scatter(ytrain ,ytrain_pred,c='blue',marker='o',label='Training data')
plt.scatter(ytest,ytest_pred ,c='lightgreen',marker='s',label='Test data')
plt.xlabel('True values')
plt.ylabel('Predicted')
plt.title("True value vs Predicted value")
plt.legend(loc= 'upper left')
#plt.hlines(y=0,xmin=0,xmax=50)
plt.plot()
plt.show()
```



In [ ]:
```
.
```