# FUNDAMENTALS OF MACHINE LEARNING

## LAB ASSIGNMENT - 2

**NAME** - Kaparotu Venkata Surya Tharani
**USN** - 22BTRAD018
**BRANCH** - AI & DE

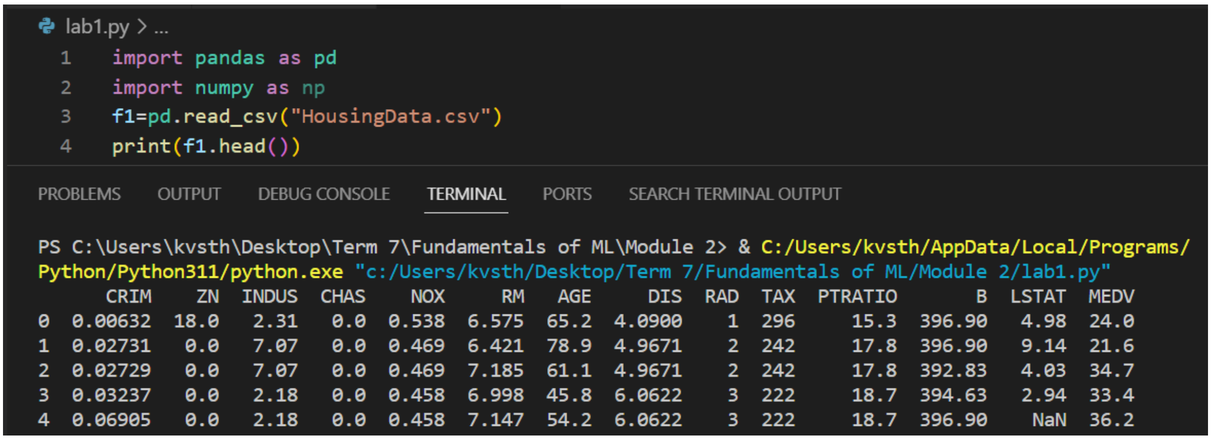
### Questions -

### 1. Load a dataset with outliers values (Boston Housing Dataset).

**CODE :**

```
import numpy as np
import pandas as pd
f1=pd.read_csv(“HousingData.csv”)
print(f1.head())
```

**CODE & OUTPUT :**



```
import pandas as pd
import numpy as np
f1=pd.read_csv("HousingData.csv")
print(f1.head())
```

```
PS C:\Users\kvsth\Desktop\Term 7\Fundamentals of ML\Module 2> & C:/Users/kvsth/AppData/Local/Programs/
Python/Python311/python.exe "c:/Users/kvsth/Desktop/Term 7/Fundamentals of ML/Module 2/lab1.py"
      CRIM    ZN  INDUS  CHAS    NOX     RM   AGE     DIS  RAD  TAX  PTRATIO       B  LSTAT  MEDV
0  0.00632  18.0   2.31   0.0  0.538  6.575  65.2  4.0900    1  296     15.3  396.90   4.98  24.0
1  0.02731   0.0   7.07   0.0  0.469  6.421  78.9  4.9671    2  242     17.8  396.90   9.14  21.6
2  0.02729   0.0   7.07   0.0  0.469  7.185  61.1  4.9671    2  242     17.8  392.83   4.03  34.7
3  0.03237   0.0   2.18   0.0  0.458  6.998  45.8  6.0622    3  222     18.7  394.63   2.94  33.4
4  0.06905   0.0   2.18   0.0  0.458  7.147  54.2  6.0622    3  222     18.7  396.90    NaN  36.2
```

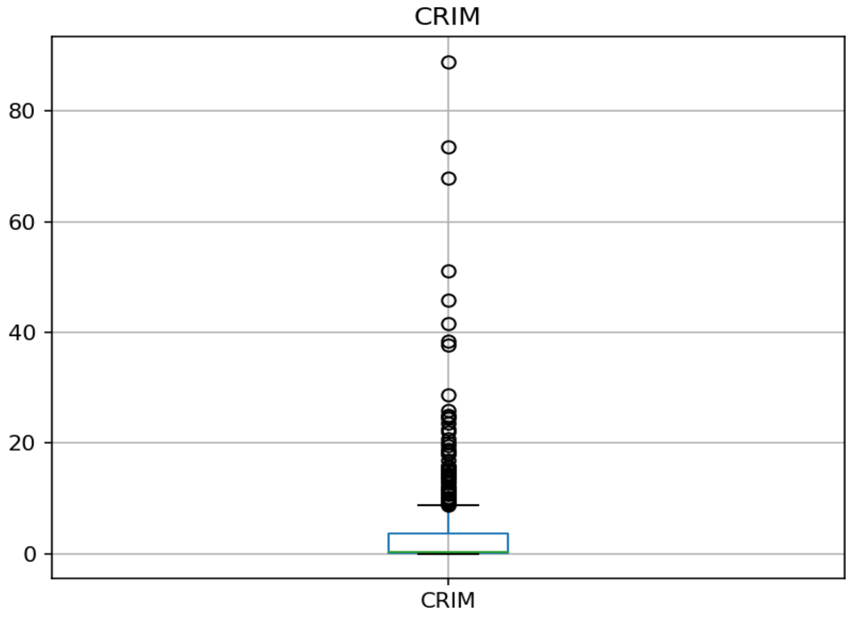
### 2. Use visualization or statistical methods to detect outliers.

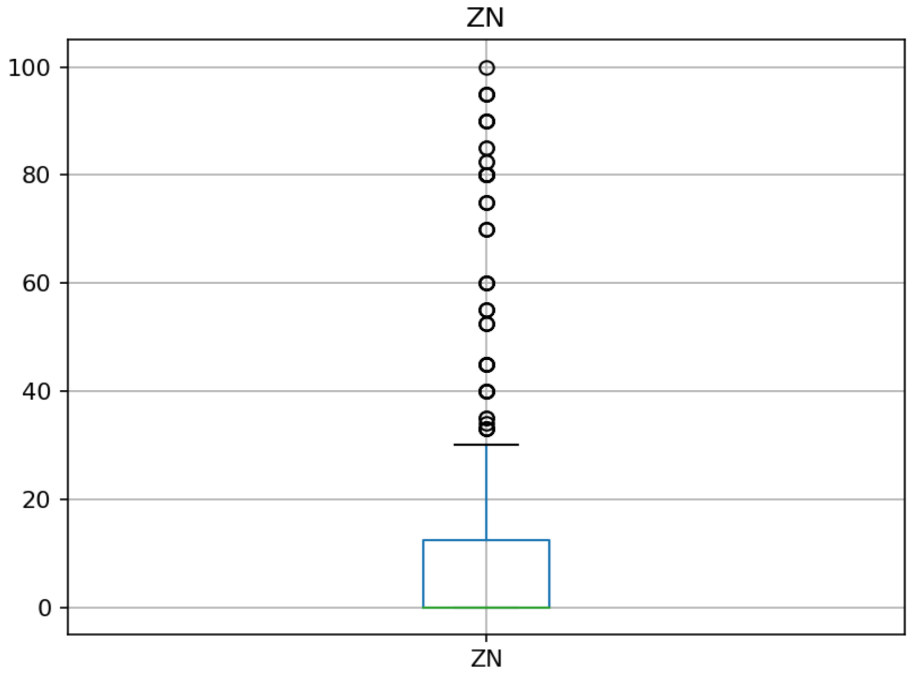
**Code :**

```
import matplotlib.pyplot as plt
for attr in f1.columns:
        f1.boxplot(column=attr)
        plt.title(attr)
        plt.show()
```

**Output :**

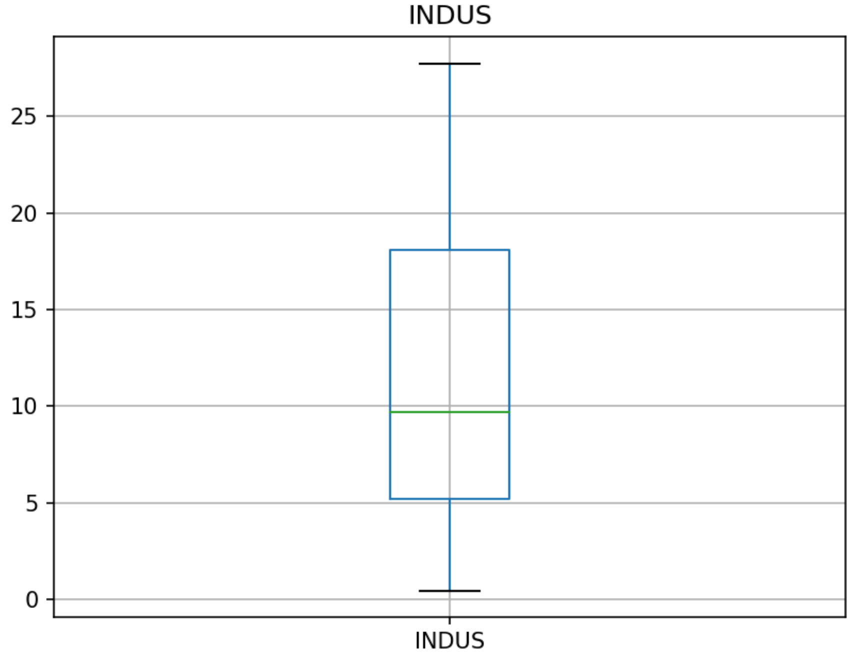
The outliers are present above 10.

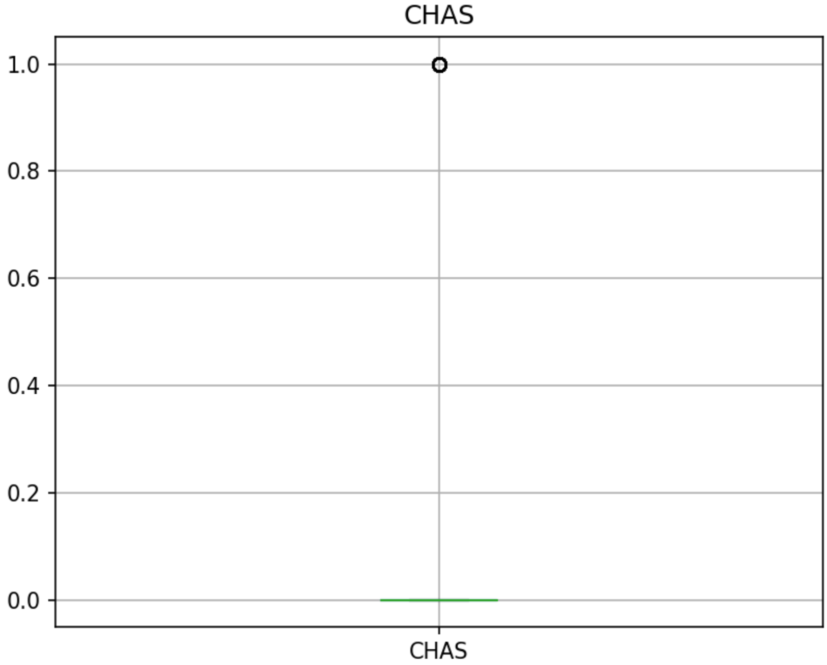




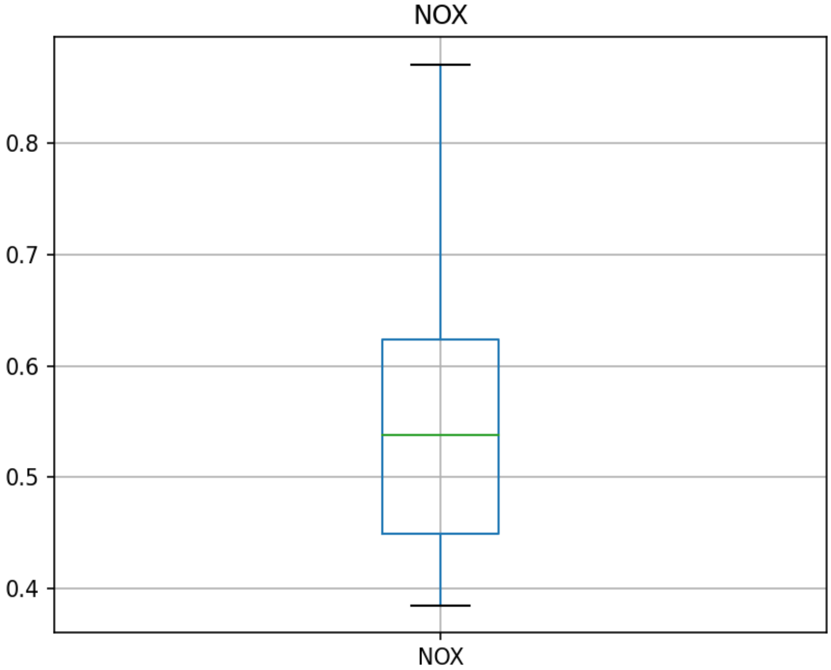
In this the outliers are present above 30.

There are no outliers present.

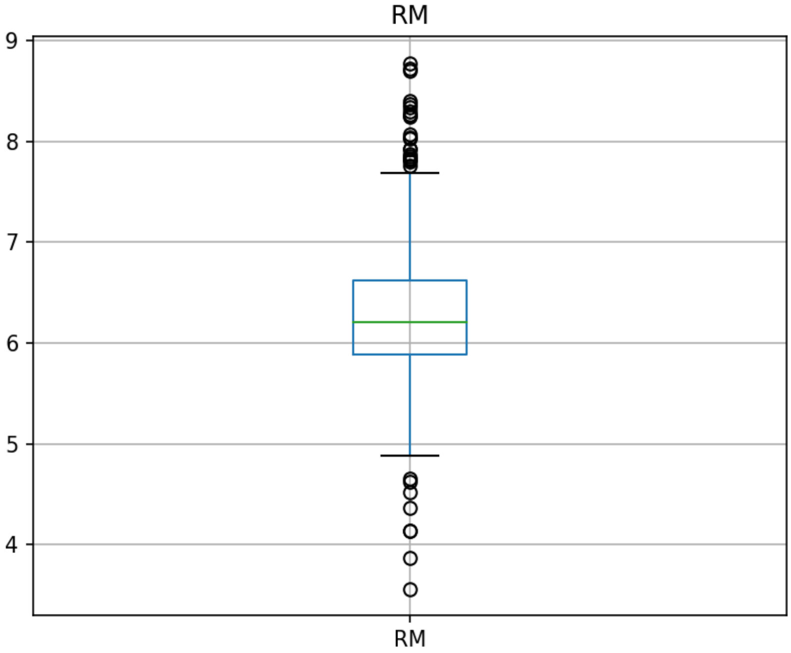




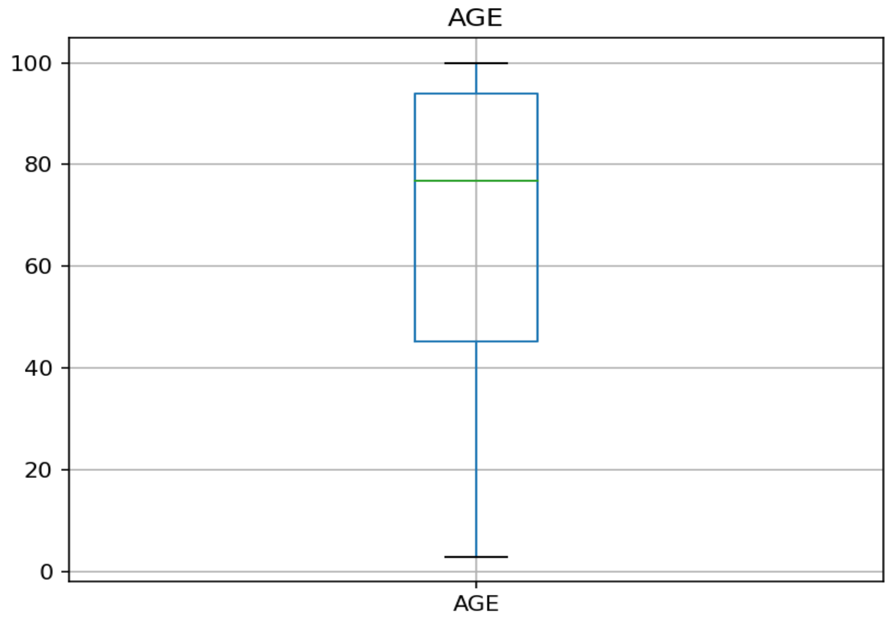
The outliers are present at 1.



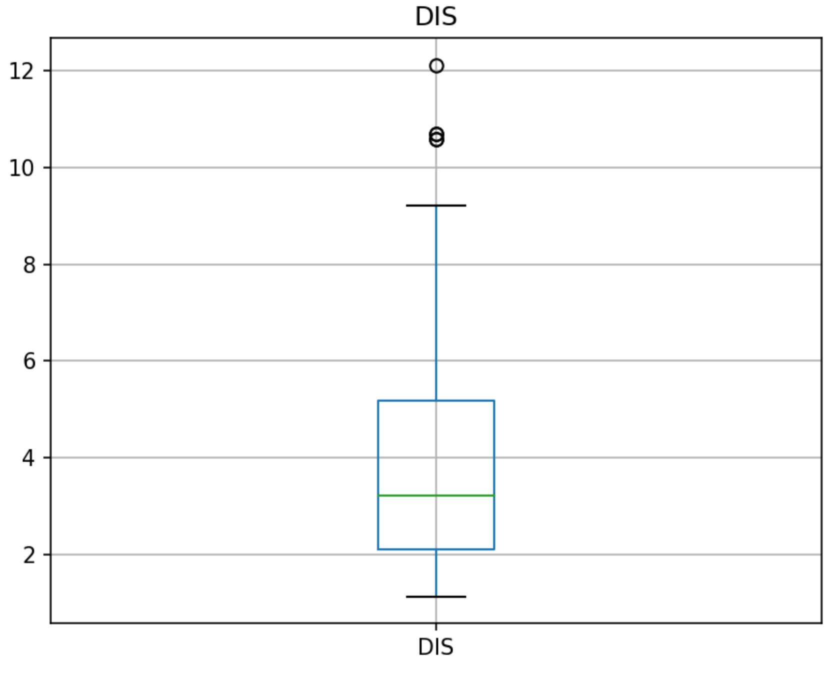
There are no outliers.



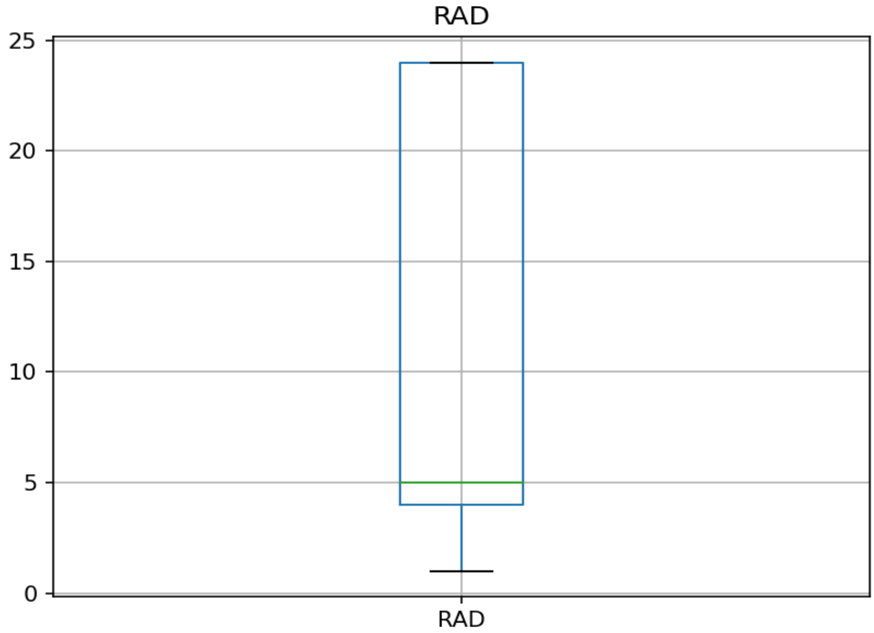
The outliers are present above 7.5 and below 5.7.



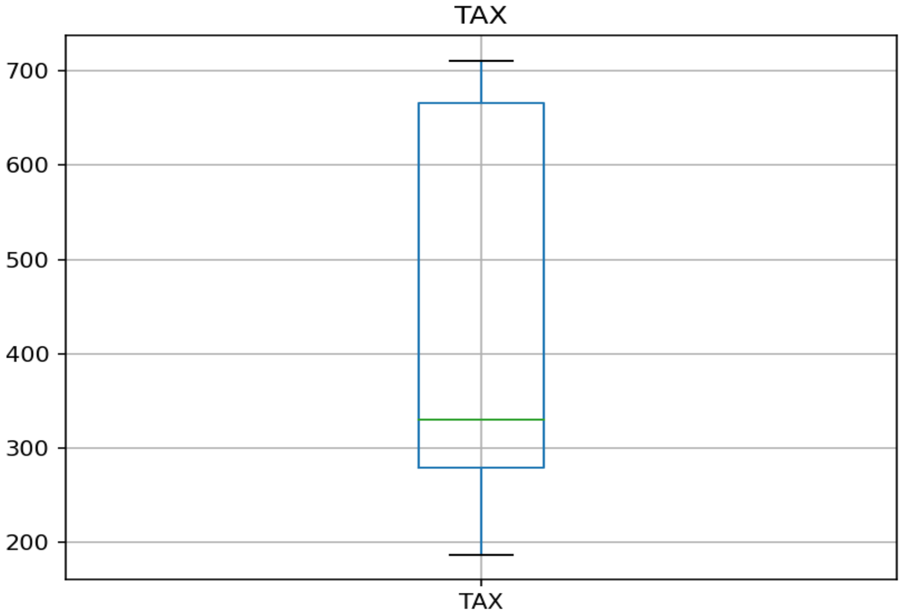
There are no outliers present.



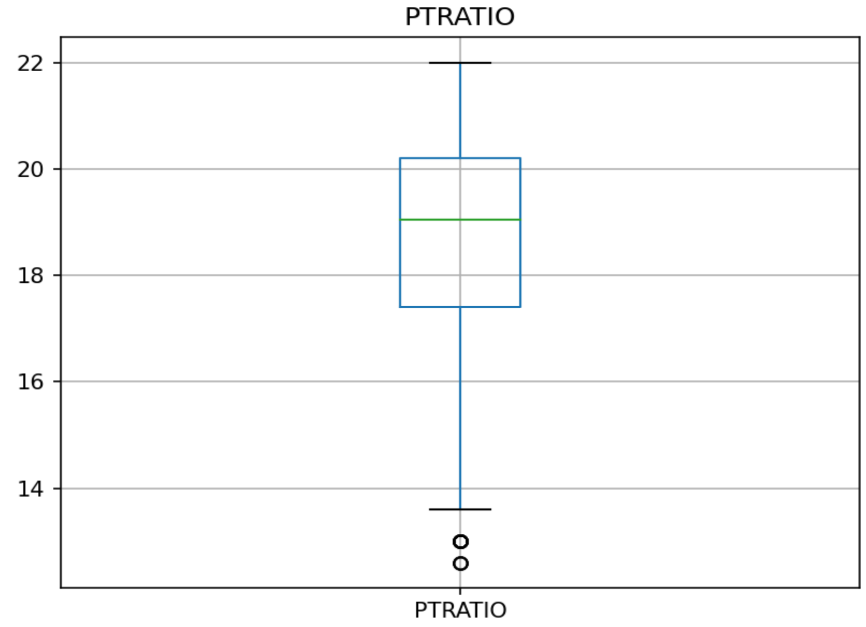
The outliers are present above 11.



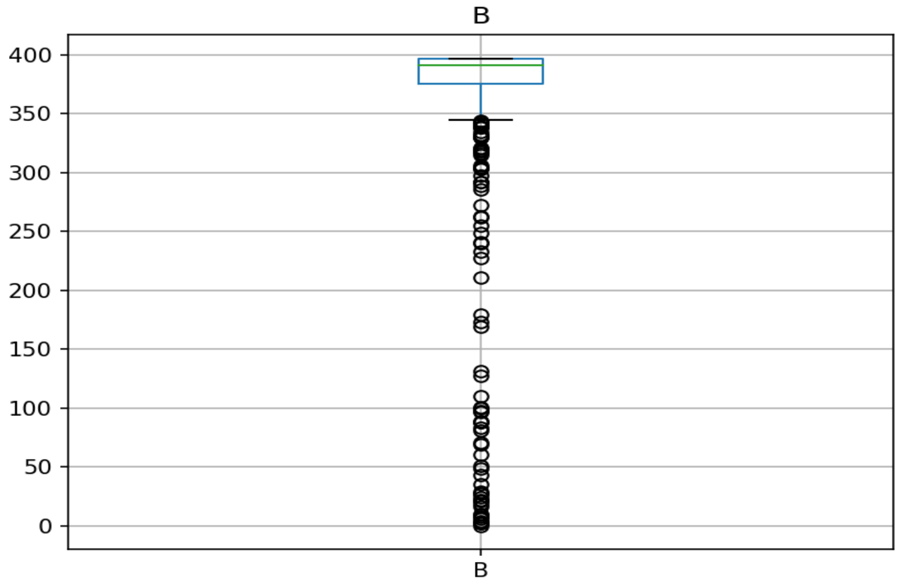
There are no outliers present.



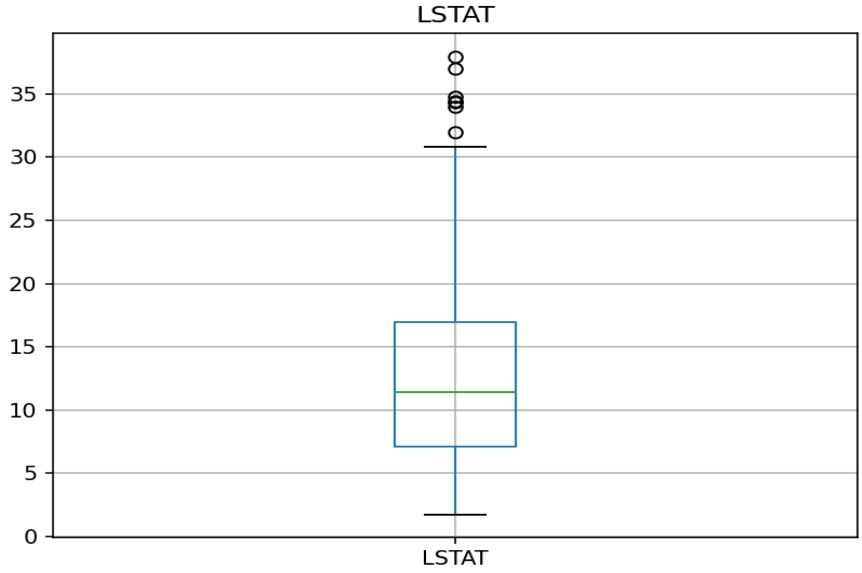
There are no outliers present.



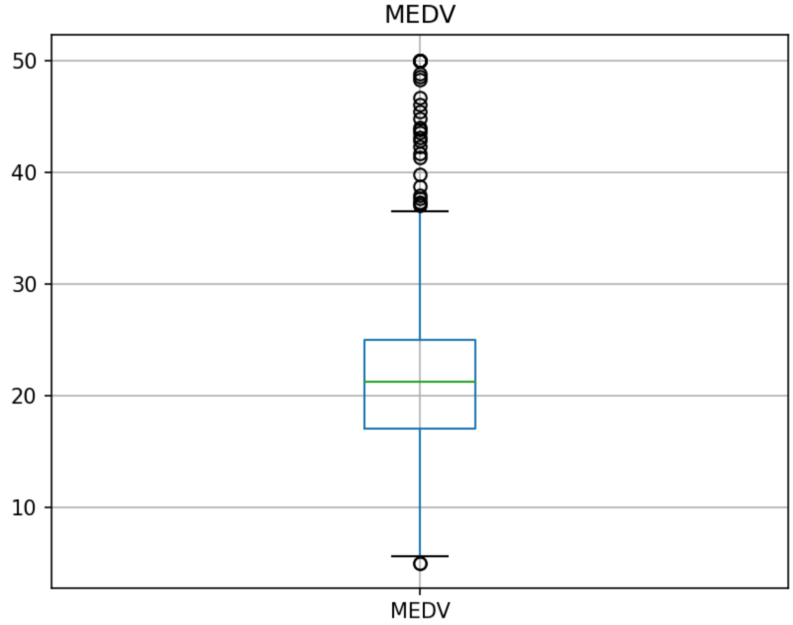
The outliers are present below 14.



The outliers are present below 350.



The outliers are present above 30.



The outliers are present above 35.

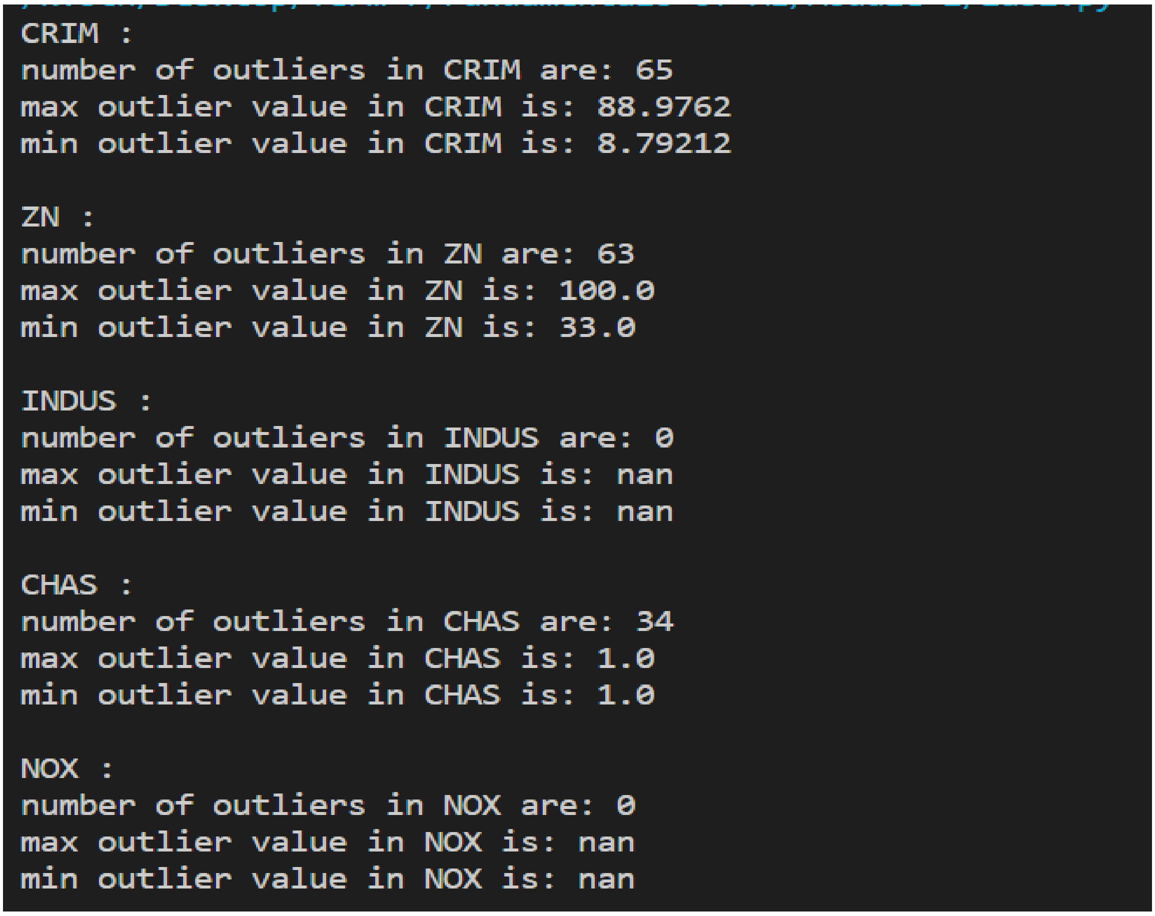
**Statistical methods:**

In statistical methods, we will compute the outlier data points using the statistical technique known as interquartile range (IQR).

**Code:**

```
def find_outliers_IQR(df):
        q1=df.quantile(0.25)
        q3=df.quantile(0.75)
        IQR=q3-q1
        outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]
        return outliers
for attr in f1.columns:
        outliers = find_outliers_IQR(f1[attr]) print(attr,':')
        print('number of outliers in',attr,'are: '+ str(len(outliers)))
        print('max outlier value in',attr,'is: '+ str(outliers.max()))
        print('min outlier value in',attr,'is: '+ str(outliers.min()),'\n')
```

**Output :**



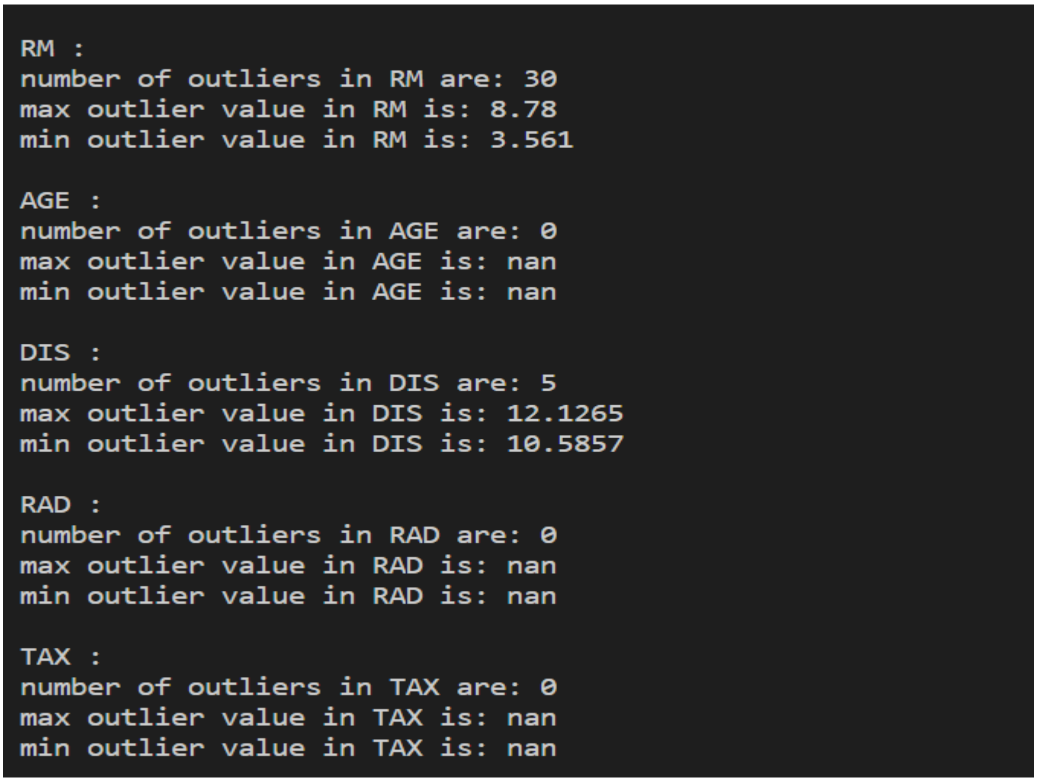
```
CRIM :
number of outliers in CRIM are: 65
max outlier value in CRIM is: 88.9762
min outlier value in CRIM is: 8.79212

ZN :
number of outliers in ZN are: 63
max outlier value in ZN is: 100.0
min outlier value in ZN is: 33.0

INDUS :
number of outliers in INDUS are: 0
max outlier value in INDUS is: nan
min outlier value in INDUS is: nan

CHAS :
number of outliers in CHAS are: 34
max outlier value in CHAS is: 1.0
min outlier value in CHAS is: 1.0

NOX :
number of outliers in NOX are: 0
max outlier value in NOX is: nan
min outlier value in NOX is: nan
```



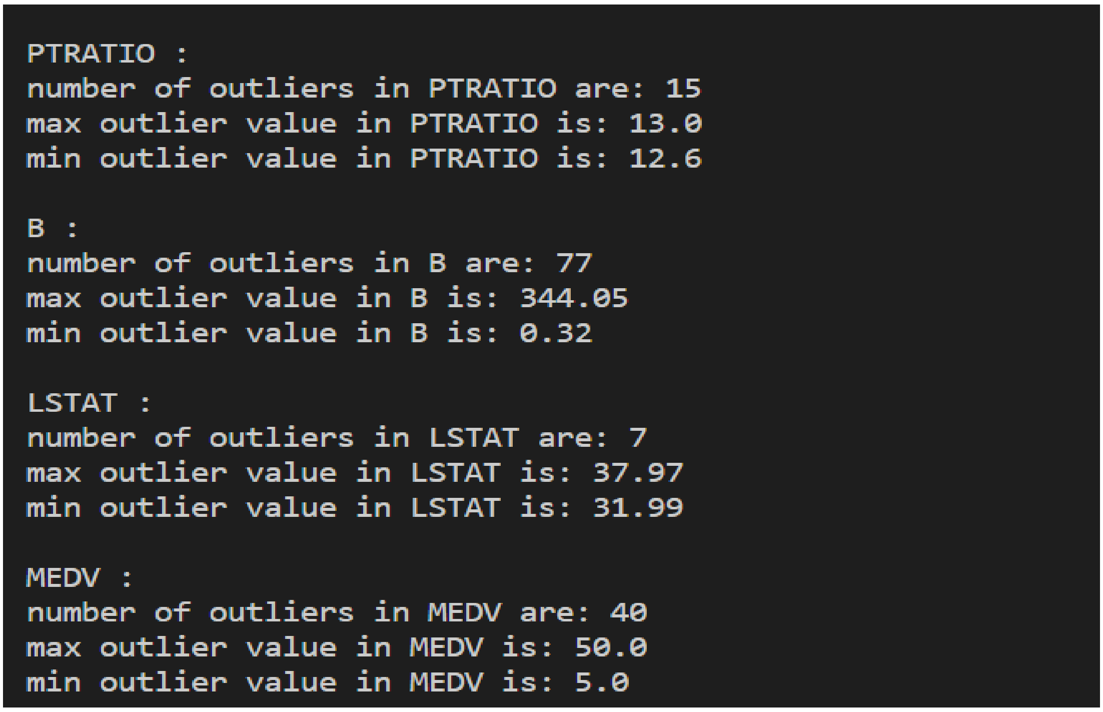
```
RM :
number of outliers in RM are: 30
max outlier value in RM is: 8.78
min outlier value in RM is: 3.561

AGE :
number of outliers in AGE are: 0
max outlier value in AGE is: nan
min outlier value in AGE is: nan

DIS :
number of outliers in DIS are: 5
max outlier value in DIS is: 12.1265
min outlier value in DIS is: 10.5857

RAD :
number of outliers in RAD are: 0
max outlier value in RAD is: nan
min outlier value in RAD is: nan

TAX :
number of outliers in TAX are: 0
max outlier value in TAX is: nan
min outlier value in TAX is: nan
```



```
PTRATIO :
number of outliers in PTRATIO are: 15
max outlier value in PTRATIO is: 13.0
min outlier value in PTRATIO is: 12.6

B :
number of outliers in B are: 77
max outlier value in B is: 344.05
min outlier value in B is: 0.32

LSTAT :
number of outliers in LSTAT are: 7
max outlier value in LSTAT is: 37.97
min outlier value in LSTAT is: 31.99

MEDV :
number of outliers in MEDV are: 40
max outlier value in MEDV is: 50.0
min outlier value in MEDV is: 5.0
```

## 3. Implement a strategy to handle outliers (e.g., removal and transformation).

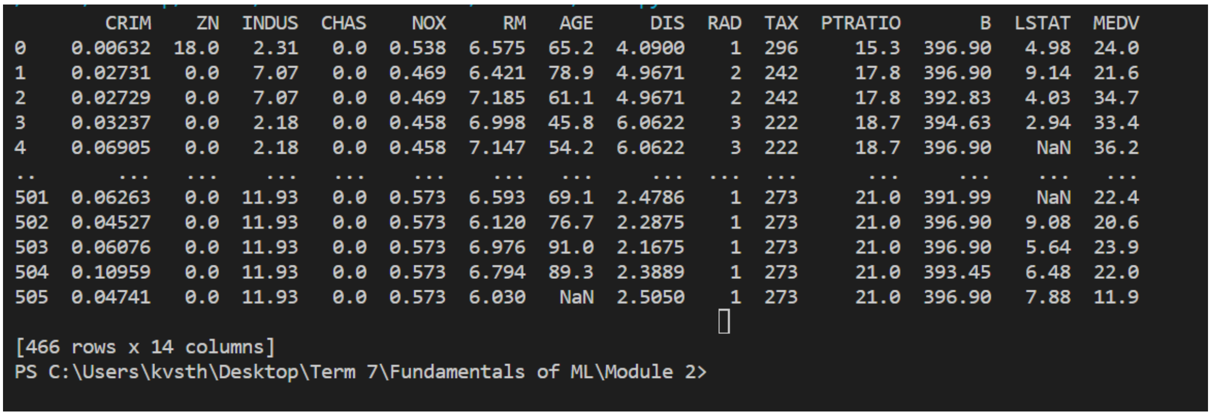
### ❖ Strategy to handle outliers using removal -

**Code:**

```
 import pandas as pd
 # Calculate the IQR for each feature
for attr in df.columns:
            q1 = f1[attr].quantile(0.25)
            q3 = f1[attr].quantile(0.75)
            iqr = q3 - q1

 # Remove outliers
df = f1[df[attr] >= q1 - 1.5 * iqr]
df = f1[df[attr] <= q3 + 1.5 * iqr]
print(df)
```

**Output :**



```
        CRIM    ZN  INDUS  CHAS    NOX     RM   AGE     DIS  RAD  TAX  PTRATIO       B  LSTAT  MEDV
0    0.00632  18.0   2.31   0.0  0.538  6.575  65.2  4.0900    1  296     15.3  396.90   4.98  24.0
1    0.02731   0.0   7.07   0.0  0.469  6.421  78.9  4.9671    2  242     17.8  396.90   9.14  21.6
2    0.02729   0.0   7.07   0.0  0.469  7.185  61.1  4.9671    2  242     17.8  392.83   4.03  34.7
3    0.03237   0.0   2.18   0.0  0.458  6.998  45.8  6.0622    3  222     18.7  394.63   2.94  33.4
4    0.06905   0.0   2.18   0.0  0.458  7.147  54.2  6.0622    3  222     18.7  396.90    NaN  36.2
..       ...   ...    ...   ...    ...    ...   ...     ...  ...  ...      ...     ...    ...   ...
501  0.06263   0.0  11.93   0.0  0.573  6.593  69.1  2.4786    1  273     21.0  391.99    NaN  22.4
502  0.04527   0.0  11.93   0.0  0.573  6.120  76.7  2.2875    1  273     21.0  396.90   9.08  20.6
503  0.06076   0.0  11.93   0.0  0.573  6.976  91.0  2.1675    1  273     21.0  396.90   5.64  23.9
504  0.10959   0.0  11.93   0.0  0.573  6.794  89.3  2.3889    1  273     21.0  393.45   6.48  22.0
505  0.04741   0.0  11.93   0.0  0.573  6.030   NaN  2.5050    1  273     21.0  396.90   7.88  11.9

[466 rows x 14 columns]
PS C:\Users\kvsth\Desktop\Term 7\Fundamentals of ML\Module 2>
```

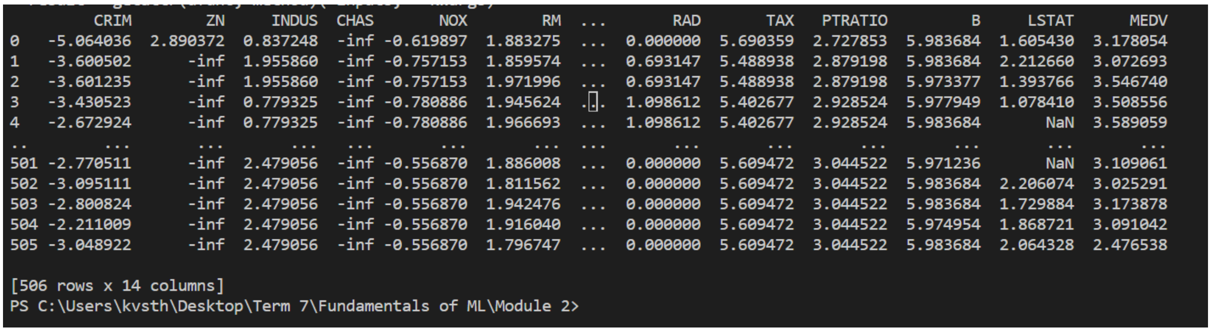
### ❖ Strategy to handle outliers using transformation -

**Code:**

```
 import pandas as pd
 import numpy as np

# Transform outliers using log transformation
for attr in f1.columns:
          if not pd.api.types.is_categorical_dtype(f1[attr]):
                    f1[attr] = np.log(f1[attr])
print(f1)
```

## Output :



```
          CRIM        ZN     INDUS  CHAS       NOX        RM  ...       RAD       TAX   PTRATIO         B     LSTAT      MEDV
0   -5.064036  2.890372  0.837248  -inf -0.619897  1.883275  ...  0.000000  5.690359  2.727853  5.983684  1.605430  3.178054
1   -3.600502      -inf  1.955860  -inf -0.757153  1.859574  ...  0.693147  5.488938  2.879198  5.983684  2.212660  3.072693
2   -3.601235      -inf  1.955860  -inf -0.757153  1.971996  ...  0.693147  5.488938  2.879198  5.973377  1.393766  3.546740
3   -3.430523      -inf  0.779325  -inf -0.780886  1.945624  ...  1.098612  5.402677  2.928524  5.977949  1.078410  3.508556
4   -2.672924      -inf  0.779325  -inf -0.780886  1.966693  ...  1.098612  5.402677  2.928524  5.983684       NaN  3.589059
..        ...       ...       ...   ...       ...       ...  ...       ...       ...       ...       ...       ...       ...
501 -2.770511      -inf  2.479056  -inf -0.556870  1.886008  ...  0.000000  5.609472  3.044522  5.971236       NaN  3.109061
502 -3.095111      -inf  2.479056  -inf -0.556870  1.811562  ...  0.000000  5.609472  3.044522  5.983684  2.206074  3.025291
503 -2.800824      -inf  2.479056  -inf -0.556870  1.942476  ...  0.000000  5.609472  3.044522  5.983684  1.729884  3.173878
504 -2.211009      -inf  2.479056  -inf -0.556870  1.916040  ...  0.000000  5.609472  3.044522  5.974954  1.868721  3.091042
505 -3.048922      -inf  2.479056  -inf -0.556870  1.796747  ...  0.000000  5.609472  3.044522  5.983684  2.064328  2.476538

[506 rows x 14 columns]
PS C:\Users\kvsth\Desktop\Term 7\Fundamentals of ML\Module 2>
```