```
#Assignment 2

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('StudentsPerformance.csv')

df

     gender race/ethnicity parental level of education
lunch  \
0    female        group B           bachelor's degree      standard

1    female        group C                some college      standard

2    female        group B             master's degree      standard

3      male        group A          associate's degree  free/reduced

4      male        group C                some college      standard

..      ...            ...                         ...           ...

995  female        group E             master's degree      standard

996    male        group C                 high school  free/reduced

997  female        group C                 high school  free/reduced

998  female        group D                some college      standard

999  female        group D                some college  free/reduced


    test preparation course  math score  reading score  writing score

0                      none          72             72             74

1                 completed          69             90             88

2                      none          90             95             93

3                      none          47             57             44

4                      none          76             78             75

..                      ...         ...            ...            ...

995               completed          88             99             95

996                    none          62             55             55
```

```
997                   completed          59              71             65

998                   completed          68              78             77

999                        none          77              86             86


[1000 rows x 8 columns]
```

```
df.head()
```

```
   gender race/ethnicity parental level of education         lunch  \
0  female        group B           bachelor's degree      standard
1  female        group C                some college      standard
2  female        group B             master's degree      standard
3    male        group A          associate's degree  free/reduced
4    male        group C                some college      standard

  test preparation course  math score  reading score  writing score
0                    none          72             72             74
1               completed          69             90             88
2                    none          90             95             93
3                    none          47             57             44
4                    none          76             78             75
```

```
df.tail()
```

```
     gender race/ethnicity parental level of education
lunch  \
995  female        group E             master's degree      standard

996    male        group C                 high school  free/reduced

997  female        group C                 high school  free/reduced

998  female        group D                some college      standard

999  female        group D                some college  free/reduced


    test preparation course  math score  reading score  writing score

995               completed          88             99             95

996                    none          62             55             55

997               completed          59             71             65

998               completed          68             78             77
```

```
999                    none          77             86             86
```

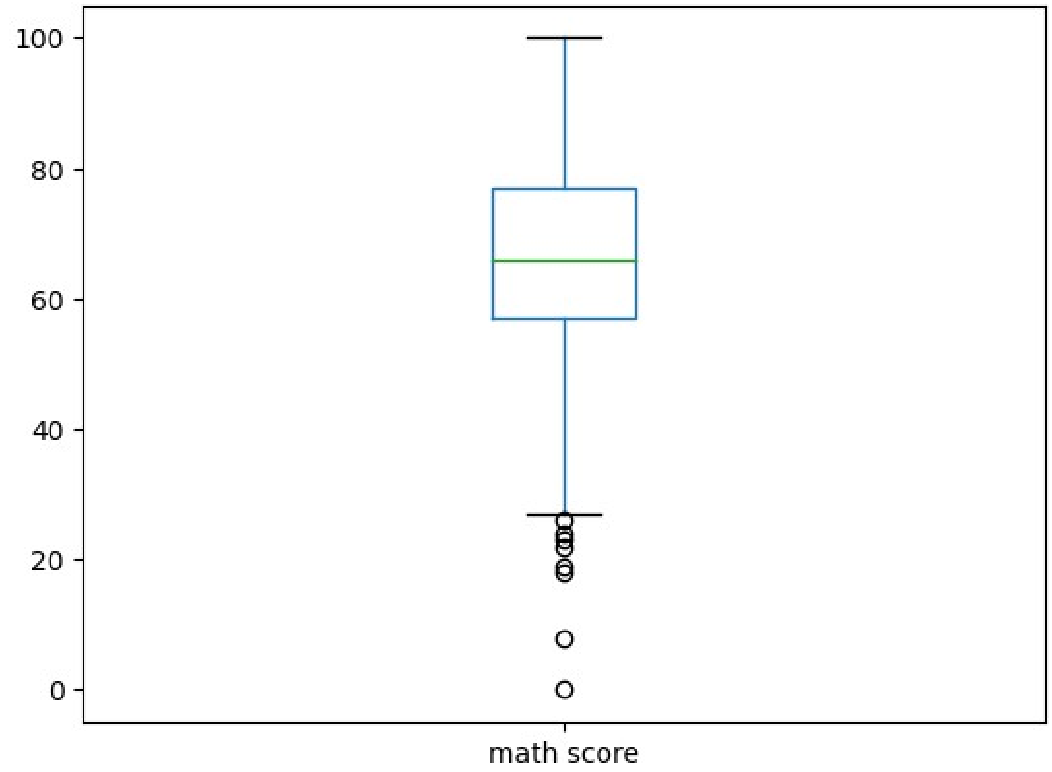
```
df.describe()
```

```
       math score  reading score  writing score
count  1000.00000    1000.000000    1000.000000
mean     66.08900      69.169000      68.054000
std      15.16308      14.600192      15.195657
min       0.00000      17.000000      10.000000
25%      57.00000      59.000000      57.750000
50%      66.00000      70.000000      69.000000
75%      77.00000      79.000000      79.000000
max     100.00000     100.000000     100.000000
```

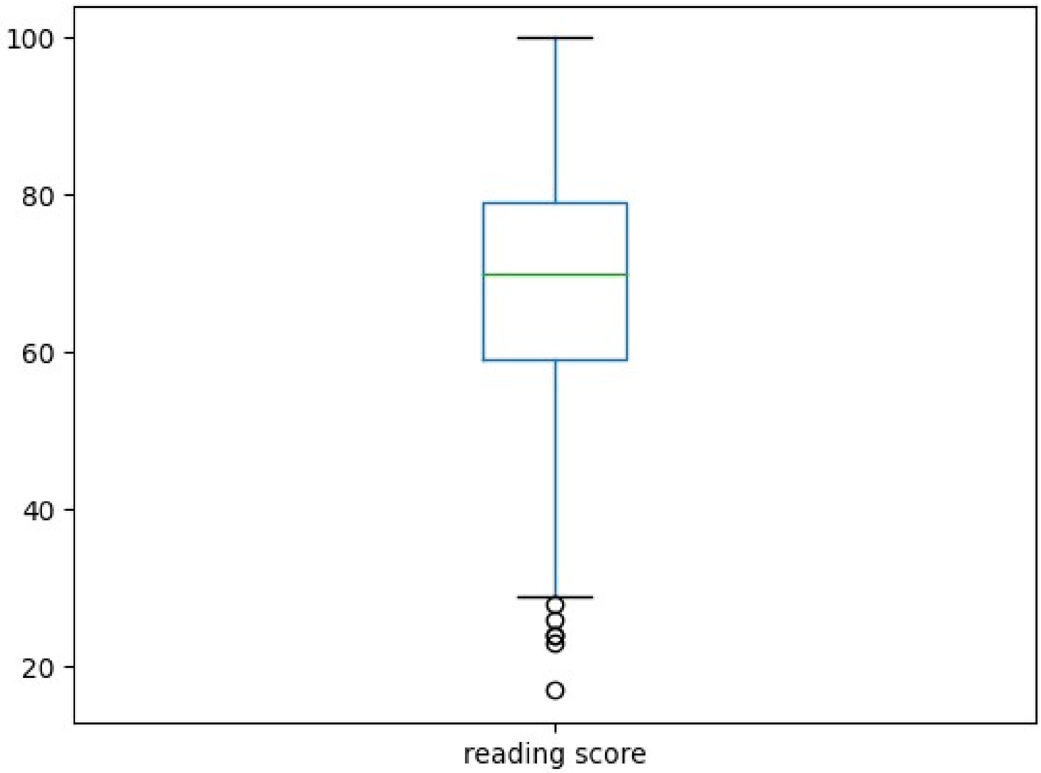
```
df.isnull()
```

```
     gender  race/ethnicity  parental level of education  lunch  \
0     False           False                        False  False   
1     False           False                        False  False   
2     False           False                        False  False   
3     False           False                        False  False   
4     False           False                        False  False   
..      ...             ...                          ...    ...   
995   False           False                        False  False   
996   False           False                        False  False   
997   False           False                        False  False   
998   False           False                        False  False   
999   False           False                        False  False   

     test preparation course  math score  reading score  writing score  

0                      False       False          False          False  

1                      False       False          False          False  

2                      False       False          False          False  

3                      False       False          False          False  

4                      False       False          False          False  

..                       ...         ...            ...            ...  

995                    False       False          False          False  

996                    False       False          False          False  

997                    False       False          False          False  

998                    False       False          False          False  
```

```
999                    False       False          False          False


[1000 rows x 8 columns]

def plot_boxplot(df,ft):
    df.boxplot(column=[ft])
    plt.grid(False)
plt.show()
plot_boxplot(df,'math score')
```



```
def plot_boxplot(df,ft):
    df.boxplot(column=[ft])
    plt.grid(False)
plt.show()
plot_boxplot(df,'reading score')
```



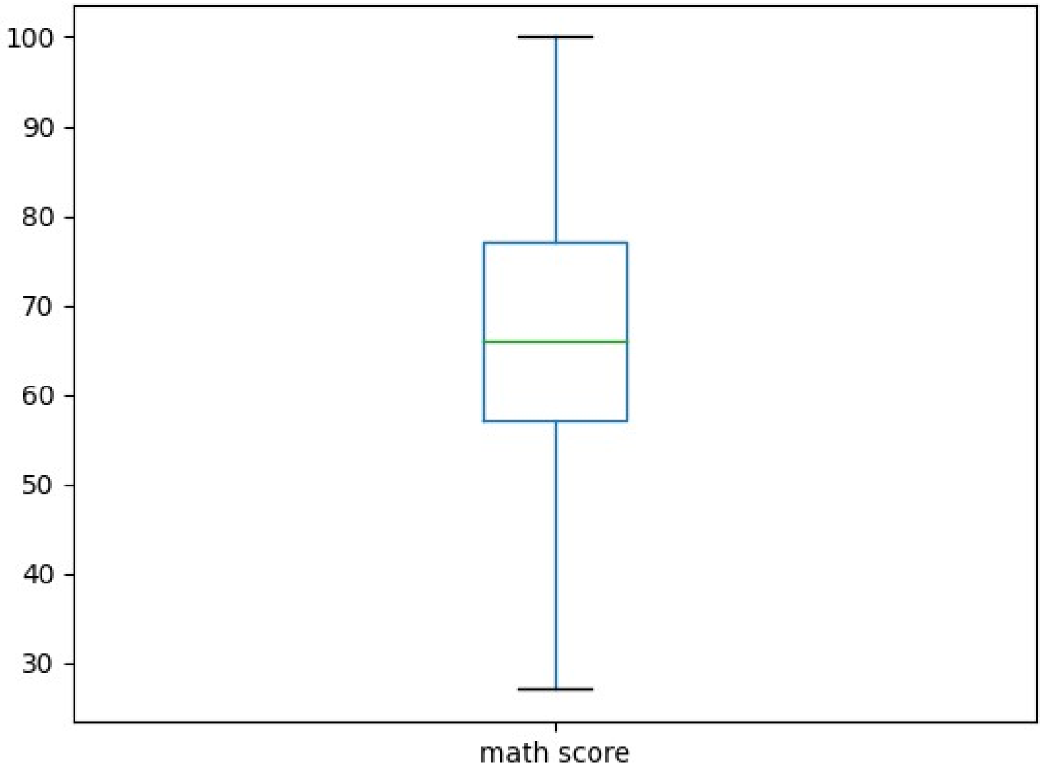
```
def outliers(df,ft):
   Q1=df[ft].quantile(0.25)
   Q3=df[ft].quantile(0.75)
   IQR=Q3-Q1
   lower_bound=Q1-1.5 *IQR
   upper_bound=Q3 +1.5 *IQR
   ls=df.index[(df[ft] < lower_bound) | (df[ft] > upper_bound)]
   return ls

index_list=[]
for features in ['math score','reading score']:
    index_list.extend(outliers(df,features))

index_list
```

```
[17, 59, 145, 338, 466, 787, 842, 980, 59, 76, 211, 327, 596, 980]
```

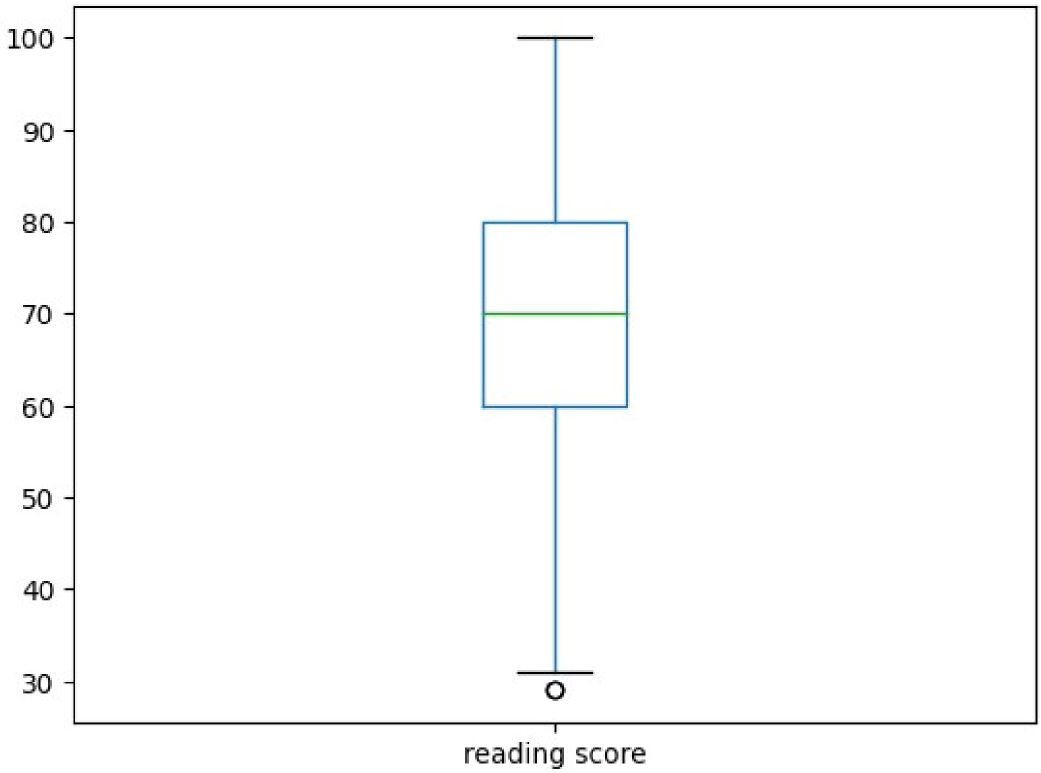
```
def remove(df,ls):
    ls=sorted(set(ls))
    df=df.drop(ls)
    return df

df_cleaned=remove(df,index_list)

df_cleaned.shape
```

```
(988, 8)
```

```
plot_boxplot(df_cleaned,'math score')
```



```
plot_boxplot(df_cleaned,'reading score')
```


100
90
80
70
60
50
40
30
reading score