Assignment 2

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:

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

In [3]:

```
df
```

Out[3]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
...	...	...	...	...	...	...	...	...
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

1000 rows × 8 columns

In [4]:

```
df.head()
```

Out[4]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

In [5]:
```
df.tail()
```

Out[5]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

In [6]:
```
df.describe()
```

Out[6]:

	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

In [7]:

```
df.isnull()
```

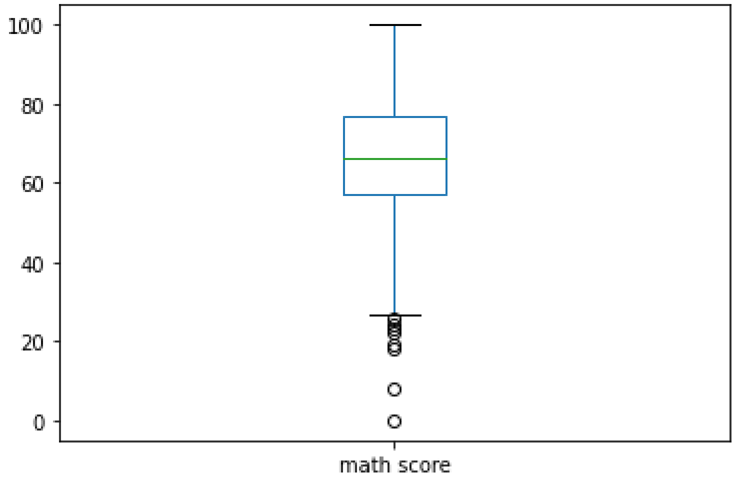
Out[7]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
995	False	False	False	False	False	False	False	False
996	False	False	False	False	False	False	False	False
997	False	False	False	False	False	False	False	False
998	False	False	False	False	False	False	False	False
999	False	False	False	False	False	False	False	False

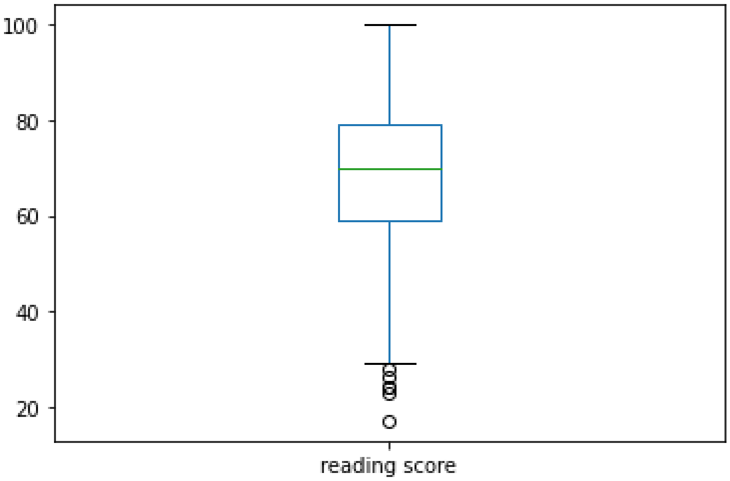
1000 rows × 8 columns

In [8]:

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



In [9]:
```
def plot_boxplot(df,ft):
    df.boxplot(column=[ft])
    plt.grid(False)
plt.show()
plot_boxplot(df,'reading score')
```



In [10]:
```
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
```

In [11]:
```
index_list=[]
for features in ['math score','reading score']:
    index_list.extend(outliers(df,features))
```

In [12]:
```
index_list
```

Out[12]:
```
[17, 59, 145, 338, 466, 787, 842, 980, 59, 76, 211, 327, 596, 980]
```

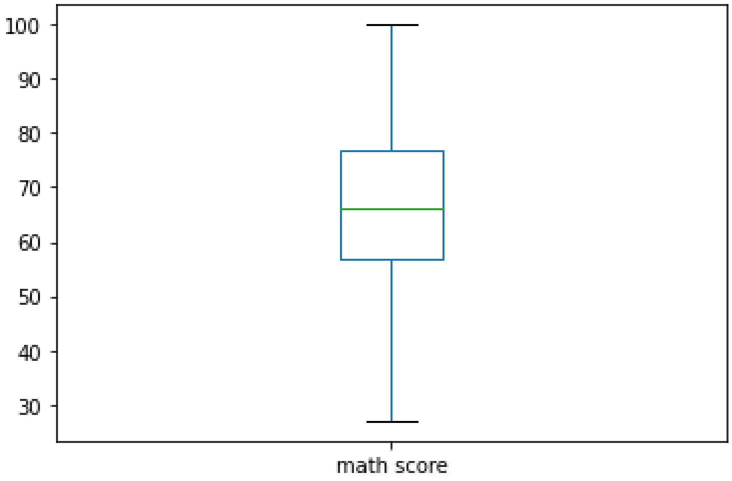
In [13]:
```
def remove(df,ls):
    ls=sorted(set(ls))
    df=df.drop(ls)
    return df
```

In [14]:
```
df_cleaned=remove(df,index_list)
```

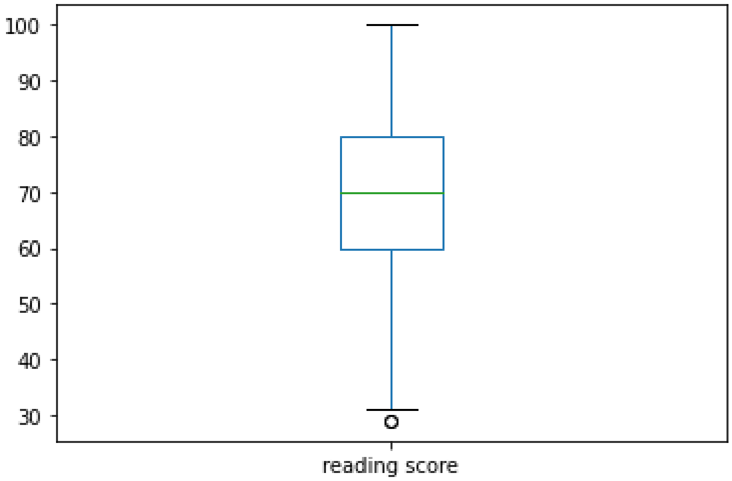
In [15]:
```
df_cleaned.shape
```

Out[15]:
```
(988, 8)
```

In [16]:
```
plot_boxplot(df_cleaned,'math score')
```



In [17]:
```
plot_boxplot(df_cleaned,'reading score')
```


reading score