In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns
```

In [2]:

```
df = pd.read_csv("F:\\DSE\\3rd year engineering\\5th sem\\6th sem\\DSBDA\\dataset\\stude
```

In [3]:

```
df.head()
```

Out[3]:

	gender	math score	reading score	writing score	placement score	club join year	placement offer count	region
0	female	66.0	94.0	68.0	94.0	2018	3	nasik
1	male	74.0	89.0	75.0	80.0	2021	2	pune
2	male	68.0	92.0	73.0	93.0	2021	3	mumbai
3	female	70.0	98.0	77.0	93.0	2021	3	pune
4	male	75.0	93.0	61.0	97.0	2018	3	nasik

In [4]:

```
df.isnull()
```

Out[4]:

	gender	math score	reading score	writing score	placement score	club join year	placement offer count	region
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
5	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	False
10	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False
12	False	False	False	False	True	False	False	False
13	False	False	False	True	False	False	False	True
14	False	False	False	False	False	False	False	False
15	False	False	False	False	False	False	False	False
16	False	False	False	False	False	False	False	False
17	False	False	False	False	False	False	False	True
18	False	False	False	False	False	False	False	False
19	False	False	False	False	False	False	False	False
20	False	True	False	False	False	False	False	False
21	False	False	True	False	False	False	False	True
22	False	False	False	False	False	False	False	False
23	False	False	False	False	False	False	False	False
24	False	False	False	False	False	False	False	False
25	False	False	False	False	False	False	False	False
26	False	True	False	False	False	False	False	True
27	False	False	False	False	False	False	False	False
28	False	False	False	False	False	False	False	False
29	False	False	False	False	False	False	False	False

In [5]:

```
df.isnull().sum()
```

Out[5]:

```
gender                  0
math score              2
reading score           1
writing score           1
placement score         1
club join year          0
placement offer count   0
region                  4
dtype: int64
```

In [6]:

```
df.mean()
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_16104\3698961737.py:1: FutureWarn
ing: The default value of numeric_only in DataFrame.mean is deprecated. In
a future version, it will default to False. In addition, specifying 'numer
ic_only=None' is deprecated. Select only valid columns or specify the valu
e of numeric_only to silence this warning.
  df.mean()
```

Out[6]:

```
math score                 70.500000
reading score              91.965517
writing score              69.655172
placement score            85.241379
club join year           2019.433333
placement offer count       2.533333
dtype: float64
```

In [7]:

```
df['math score'].fillna(value=70.5, inplace=True)
```

In [8]:

```
df
```

Out[8]:

	gender	math score	reading score	writing score	placement score	club join year	placement offer count	region
0	female	66.0	94.0	68.0	94.0	2018	3	nasik
1	male	74.0	89.0	75.0	80.0	2021	2	pune
2	male	68.0	92.0	73.0	93.0	2021	3	mumbai
3	female	70.0	98.0	77.0	93.0	2021	3	pune
4	male	75.0	93.0	61.0	97.0	2018	3	nasik
5	female	64.0	86.0	61.0	88.0	2019	3	pune
6	female	90.0	80.0	78.0	82.0	2019	2	mumbai
7	female	76.0	91.0	79.0	89.0	2019	3	mumbai
8	male	73.0	97.0	98.0	98.0	2019	3	nasik
9	male	79.0	88.0	61.0	92.0	2018	3	pune
10	female	75.0	83.0	80.0	92.0	2019	3	mumbai
11	male	68.0	88.0	66.0	92.0	2021	3	pune
12	female	71.0	95.0	79.0	NaN	2018	3	nasik
13	female	67.0	88.0	NaN	98.0	2020	3	NaN
14	male	77.0	95.0	64.0	78.0	2018	2	mumbai
15	male	71.0	100.0	62.0	79.0	2021	1	mumbai
16	female	79.0	99.0	71.0	90.0	2019	3	nasik
17	female	60.0	90.0	66.0	77.0	2019	2	NaN
18	male	27.0	91.0	66.0	35.0	2018	1	mumbai
19	female	66.0	90.0	69.0	86.0	2020	3	pune
20	male	70.5	86.0	23.0	82.0	2018	2	nasik
21	male	75.0	NaN	76.0	100.0	2021	3	NaN
22	female	76.0	80.0	68.0	77.0	2019	2	mumbai
23	female	72.0	91.0	69.0	78.0	2018	2	mumbai
24	male	78.0	100.0	73.0	81.0	2021	2	nasik
25	female	60.0	97.0	77.0	76.0	2019	2	pune
26	male	70.5	96.0	62.0	91.0	2021	3	NaN
27	female	64.0	94.0	76.0	85.0	2021	3	pune
28	male	74.0	96.0	76.0	88.0	2020	3	nasik
29	male	79.0	100.0	66.0	81.0	2019	2	pune

In [9]:

```
df.median()
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_16104\530051474.py:1: FutureWarni
ng: The default value of numeric_only in DataFrame.median is deprecated. I
n a future version, it will default to False. In addition, specifying 'num
eric_only=None' is deprecated. Select only valid columns or specify the va
lue of numeric_only to silence this warning.
  df.median()
```

Out[9]:

```
math score                  71.5
reading score               92.0
writing score               69.0
placement score             88.0
club join year            2019.0
placement offer count        3.0
dtype: float64
```

In [10]:

```
df['placement score'].fillna(value=88.0,inplace= True)
```

In [11]:

```
df
```

Out[11]:

	gender	math score	reading score	writing score	placement score	club join year	placement offer count	region
0	female	66.0	94.0	68.0	94.0	2018	3	nasik
1	male	74.0	89.0	75.0	80.0	2021	2	pune
2	male	68.0	92.0	73.0	93.0	2021	3	mumbai
3	female	70.0	98.0	77.0	93.0	2021	3	pune
4	male	75.0	93.0	61.0	97.0	2018	3	nasik
5	female	64.0	86.0	61.0	88.0	2019	3	pune
6	female	90.0	80.0	78.0	82.0	2019	2	mumbai
7	female	76.0	91.0	79.0	89.0	2019	3	mumbai
8	male	73.0	97.0	98.0	98.0	2019	3	nasik
9	male	79.0	88.0	61.0	92.0	2018	3	pune
10	female	75.0	83.0	80.0	92.0	2019	3	mumbai
11	male	68.0	88.0	66.0	92.0	2021	3	pune
12	female	71.0	95.0	79.0	88.0	2018	3	nasik
13	female	67.0	88.0	NaN	98.0	2020	3	NaN
14	male	77.0	95.0	64.0	78.0	2018	2	mumbai
15	male	71.0	100.0	62.0	79.0	2021	1	mumbai
16	female	79.0	99.0	71.0	90.0	2019	3	nasik
17	female	60.0	90.0	66.0	77.0	2019	2	NaN
18	male	27.0	91.0	66.0	35.0	2018	1	mumbai
19	female	66.0	90.0	69.0	86.0	2020	3	pune
20	male	70.5	86.0	23.0	82.0	2018	2	nasik
21	male	75.0	NaN	76.0	100.0	2021	3	NaN
22	female	76.0	80.0	68.0	77.0	2019	2	mumbai
23	female	72.0	91.0	69.0	78.0	2018	2	mumbai
24	male	78.0	100.0	73.0	81.0	2021	2	nasik
25	female	60.0	97.0	77.0	76.0	2019	2	pune
26	male	70.5	96.0	62.0	91.0	2021	3	NaN
27	female	64.0	94.0	76.0	85.0	2021	3	pune
28	male	74.0	96.0	76.0	88.0	2020	3	nasik
29	male	79.0	100.0	66.0	81.0	2019	2	pune

In [12]:

```
df.replace(to_replace = np.nan, value = 'pune',inplace=True)
```

part B

In [13]:

```
df1 = pd.read_csv('F:\\DSE\\3rd year engineering\\5th sem\\6th sem\\DSBDA\\dataset\\demo
```

In [14]:

```
df1.head()
```

Out[14]:

	math score	reading score	writing score	placement score	placement offer count	club join year
0	80	68	70	89	3	2019
1	71	61	85	91	3	2019
2	79	16	87	77	2	2018
3	61	77	74	76	2	2020
4	78	71	67	90	3	2019

In [15]:

```
col = ['math score','reading score','writing score','placement score']
```

In [16]:

```
col
```

Out[16]:

```
['math score', 'reading score', 'writing score', 'placement score']
```

In [18]:

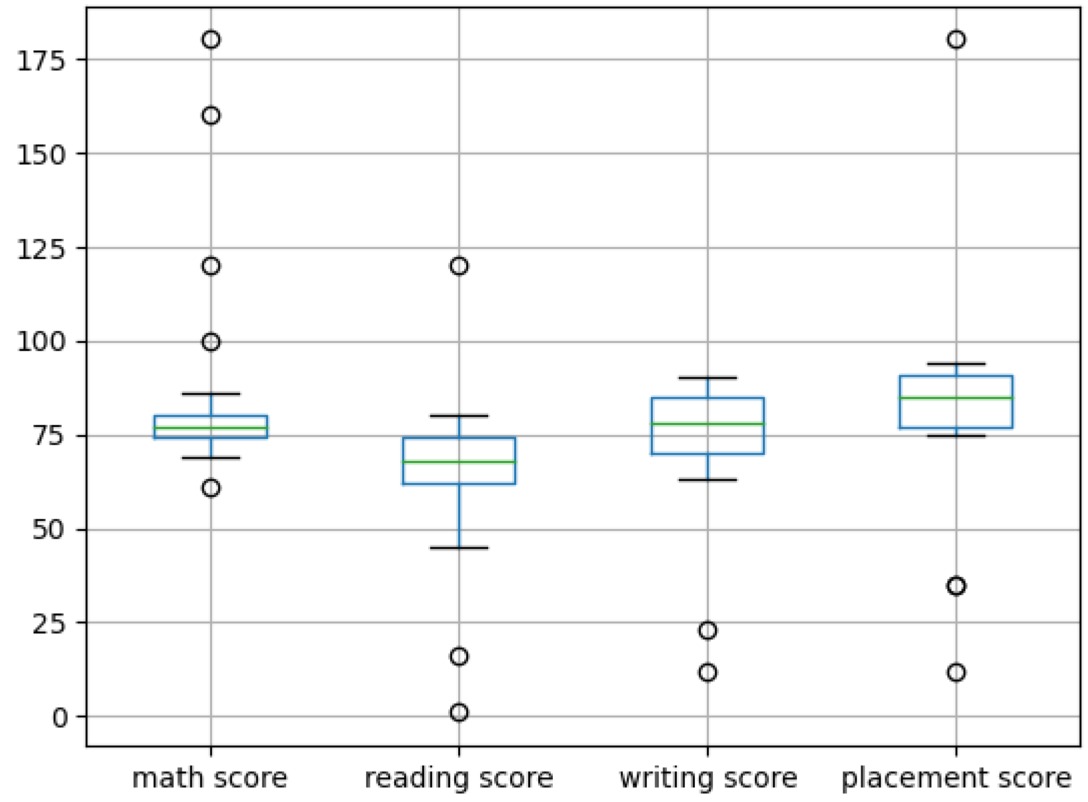
```
plt.show()
```

In [19]:

```
df1.boxplot(col)
```

Out[19]:

```
<Axes: >
```



In [20]:

```
print(np.where(df['math score']>90))
```

```
(array([], dtype=int64),)
```

In [21]:

```
z =np.abs(stats.zscore(df1['math score']))
```

In [22]:

```
z
```

Out[22]:

```
0     0.175646
1     0.528288
2     0.214828
3     0.920112
4     0.254010
5     0.449923
6     0.293193
7     0.410740
8     0.332375
9     0.371558
10    2.958952
11    0.214828
12    0.175646
13    0.254010
14    0.371558
15    0.254010
16    0.059449
17    0.175646
18    0.371558
19    0.097281
20    0.606653
21    0.608004
22    0.489105
23    0.410740
24    0.371558
25    3.742601
26    0.489105
27    0.528288
28    1.391653
Name: math score, dtype: float64
```

In [24]:

```
threshold = 0.18
```

In [25]:

```
sample_outliers=np.where(z<threshold)
```

In [26]:

```
sample_outliers
```

Out[26]:

```
(array([ 0, 12, 16, 17, 19], dtype=int64),)
```

In [ ]:

```
```