In [4]:

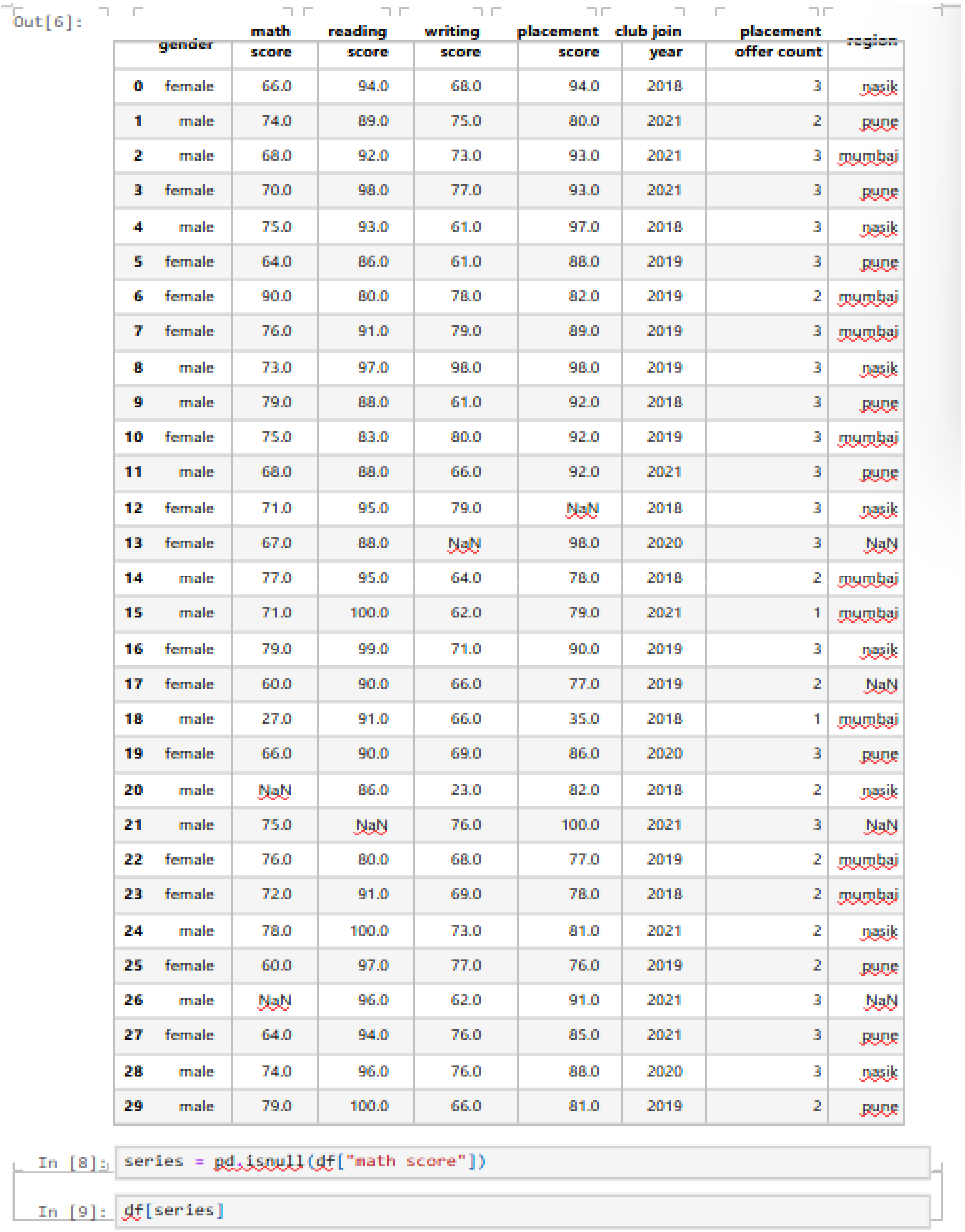
```
import pandas as pd
import numpy as np
```

In [5]:

```
df=pd.read_csv("/Users/Lenovo/Desktop/studentacademicperf.csv")
```

In [6]:

```
df
```



Out[6]:

	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	NaN	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	NaN	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 [8]:

```
series = pd.isnull(df["math score"])
```

In [9]:

```
df[series]
```

```
series1 = pd.notnull(df["math score"])
```

In [12]:

```
df[series1]
```

Out[12]:

	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

```
from sklearn.preprocessing import LabelEncoder
```

```
missing_values = ["Na", "na"]
```

```
df['math score'] = df['math score'].fillna(df['math score'].mean())
```

```
df['math score'] = df['math score'].fillna(df['math score'].median())
```

```
df['math score'] = df['math score'].fillna(df['math score'].std())
```

```
df['math score'] = df['math score'].fillna(df['math score'].min())
```

```
df['math score'] = df['math score'].fillna(df['math score'].max())
```

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

Out[23]:

	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 [25]:

```
ndf.dropna()
```

Out[25]:

	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
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
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
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
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 [27]:

```
ndf.dropna(axis = 1)
```

Out[27]:

	gender	math score	club join year	placement offer count
0	female	66.0	2018	3
1	male	74.0	2021	2
2	male	68.0	2021	3
3	female	70.0	2021	3
4	male	75.0	2018	3
5	female	64.0	2019	3
6	female	90.0	2019	2
7	female	76.0	2019	3
8	male	73.0	2019	3
9	male	79.0	2018	3
10	female	75.0	2019	3
11	male	68.0	2021	3
12	female	71.0	2018	3
13	female	67.0	2020	3
14	male	77.0	2018	2
15	male	71.0	2021	1
16	female	79.0	2019	3
17	female	60.0	2019	2
18	male	27.0	2018	1
19	female	66.0	2020	3
20	male	70.5	2018	2
21	male	75.0	2021	3
22	female	76.0	2019	2
23	female	72.0	2018	2
24	male	78.0	2021	2
25	female	60.0	2019	2
26	male	70.5	2021	3
27	female	64.0	2021	3
28	male	74.0	2020	3
29	male	79.0	2019	2

In [28]:

```
new_data = ndf.dropna(axis = 0, how ='any')
```

In [30]:

```
from scipy import stats
```

In [31]:

```
import matplotlib.pyplot as plt
```

In [32]:

```
df=pd.read_csv("/Users/Lenovo/Desktop/demo1.csv")
```

In [33]:

```
df
```

Out[33]:

	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
5	73	68	90	80	2	2019
6	77	62	70	35	2	2020
7	74	45	80	12	1	2019
8	76	60	79	77	2	2020
9	75	65	85	87	3	2018
10	160	67	12	83	2	2020
11	79	72	88	180	2	2019
12	80	80	78	94	3	2021
13	78	69	71	90	3	2019
14	75	1	71	81	2	2019
15	78	62	79	93	3	2021
16	86	78	80	88	3	2019
17	80	74	23	76	2	2021
18	75	62	86	87	3	2019
19	82	70	87	94	3	2019
20	69	65	84	35	1	2018
21	100	77	70	91	3	2018
22	72	60	78	94	3	2019
23	74	65	71	84	2	2019
24	75	77	83	77	2	2020
25	180	67	63	75	3	2021
26	72	120	70	84	2	2021
27	71	79	88	85	3	2021
28	120	73	71	94	3	2019

In [34]:

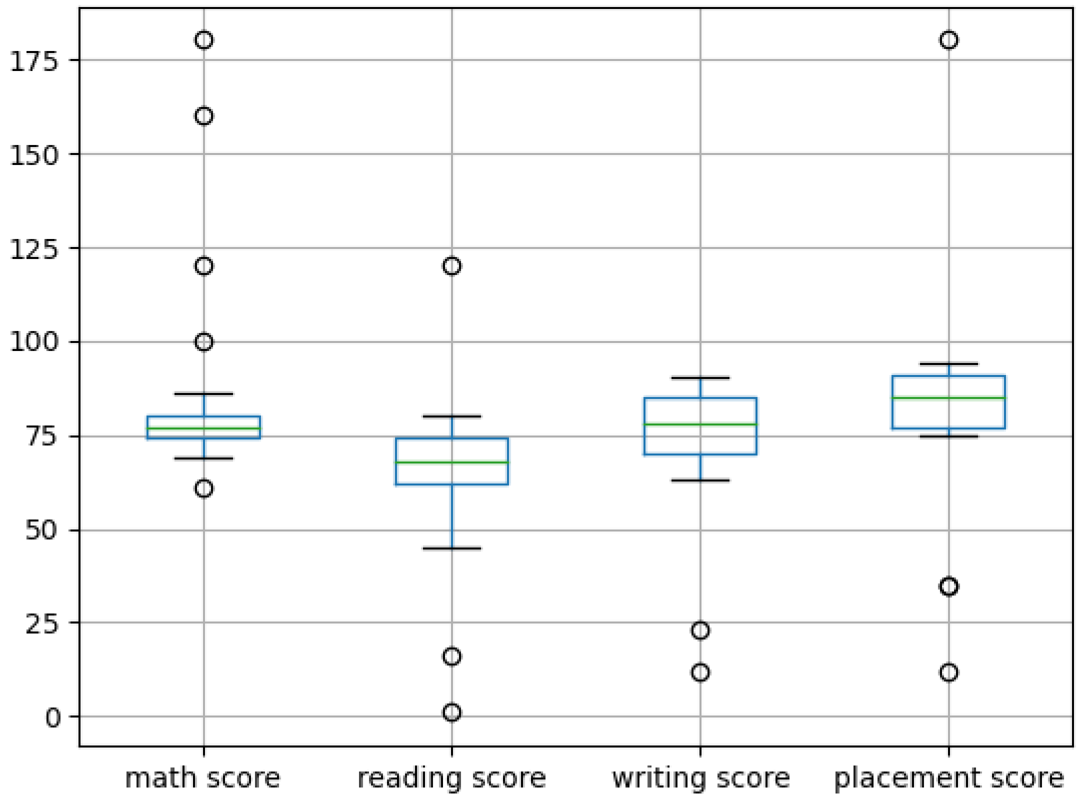
```
col = ['math score', 'reading score' , 'writing score', 'placement score']
```

In [35]:

```
df.boxplot(col)
```

Out[35]:

```
<AxesSubplot:>
```



```
In [36]: col = ['math score', 'reading score' , 'writing score','placement
         score'] df.boxplot(col)

Out[36]: <AxesSubplot:>
```

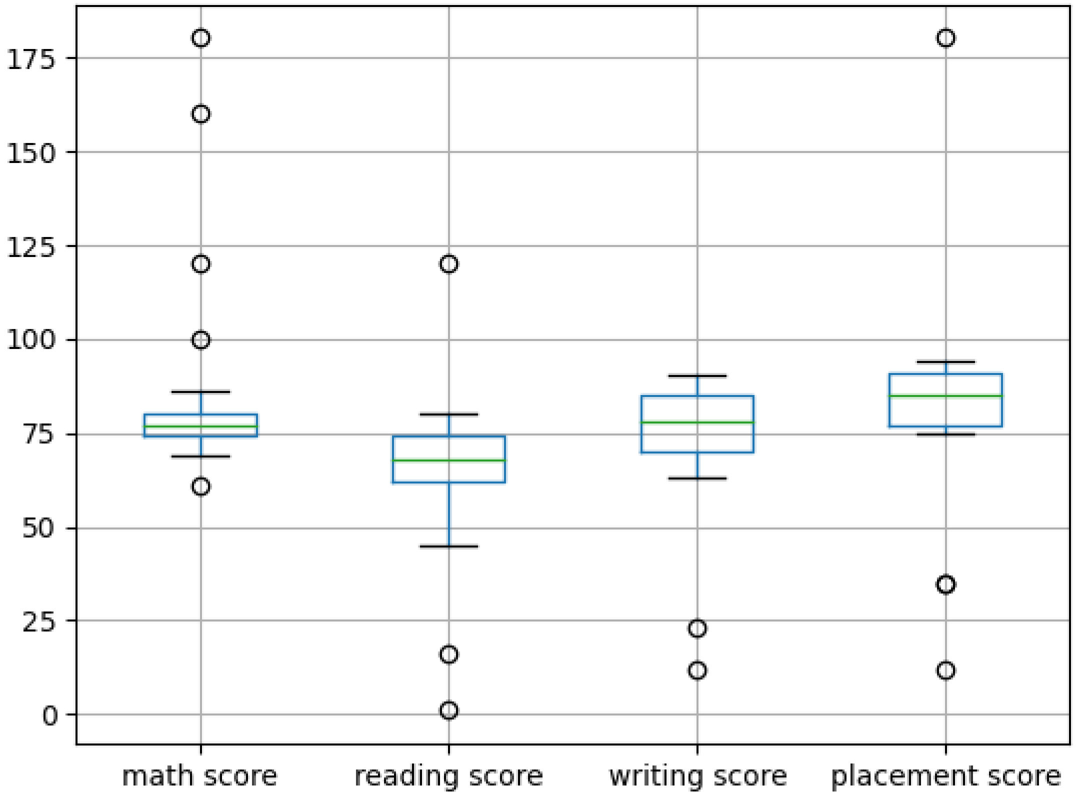
```
In [37]: print(np.where(df['math  score']>90))

         (array([10, 21, 25, 28], dtype=int64),)
```

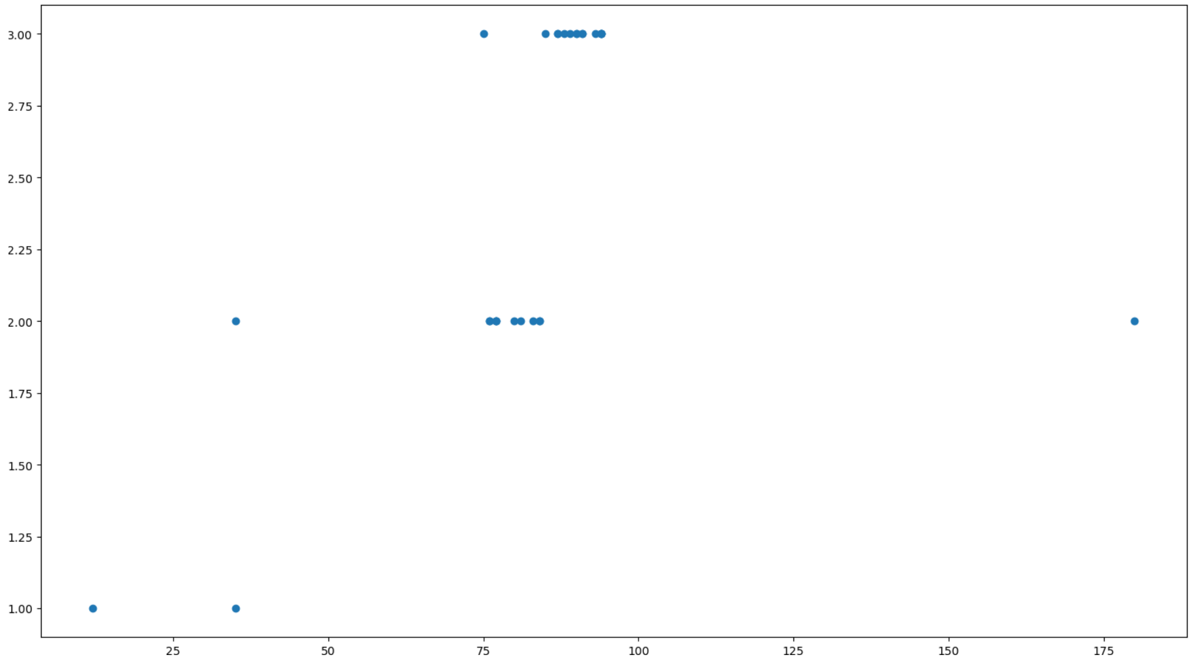
```
In [38]: print(np.where(df['reading  score']<25))
         print(np.where(df['writing  score']<30))

         (array([ 2, 14], dtype=int64),)
         (array([10,  17],  dtype=int64),)
```

```
In [39]: import matplotlib.pyplot as plt
```

```
In [40]: df=pd.read_csv("/Users/Lenovo/Desktop/demo1.csv")
```

```
In [41]: fig, ax = plt.subplots(figsize = (18,10))
         ax.scatter(df['placement  score'], df['placement  offer  count'])
         plt.show()
```





In [42]:
```
ax.set_xlabel('(Proportion non-retail business acres)/(town)')
ax.set_ylabel('(Full-value property-tax rate)/($10,000)')
```

Out[42]:
```
Text(4.444444444444452, 0.5, '(Full-value property-tax rate)/($10,000)')
```

In [43]:
```
print(np.where((df['placement score']<50) & (df['placement offer count']>1)))
```

```
(array([6], dtype=int64),)
```

In [44]:
```
print(np.where((df['placement score']>85) & (df['placement offer count']<3)))
```

```
(array([11], dtype=int64),)
```

In [45]:
```
from scipy import stats
```

In [46]:
```
z = np.abs(stats.zscore(df['math score']))
```

In [47]:

```
print(z)
```

```
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 [48]:

```
threshold = 0.18
```

In [49]:

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

Out[49]:

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

In
[50]:

```
sorted_rscore= sorted(df['reading score'])
```

In [51]:

```
sorted_rscore
```

Out[51]:
```
[1,
 16,
 45,
 60,
 60,
 61,
 62,
 62,
 62,
 65,
 65,
 65,
 67,
 67,
 68,
 68,
 69,
 70,
 71,
 72,
 73,
 74,
 77,
 77,
 77,
 78,
 79,
 80,
 120]
```

In [52]:
```
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
```

```
62.0 74.0
```

In [53]:
```
IQR = q3-q1
```

In [54]:
```
lwr_bound = q1-
(1.5*IQR) upr_bound =
q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

```
44.0 92.0
```

In [57]:
```
r_outliers = []
for i in sorted_rscore:
        if (i<lwr_bound or
                i>upr_bound):
                r_outliers.append(i)
                print(r_outliers)
```

```
[1]
[1, 16]
[1, 16, 120]
```

In [60]:
```
new_df=df
```

In [62]:
```
df=pd.read_csv("/Users/Lenovo/Desktop/demo1.csv")
df_stud=df
ninetieth_percentile = np.percentile(df_stud['math score'], 90)
b = np.where(df_stud['math score']>ninetieth_percentile, ninetieth_percentile, df_
print("New array:",b)
```

```
New array: [ 80.   71.   79.   61.   78.   73.   77.   74.   76.   75. 104.   79.   80.   78.
  75.   78.   86.   80.   75.   82.   69. 100.   72.   74.   75. 104.   72.   71.
 104.]
```

In [63]:

```
df_stud.insert(1,"m  score",b,True)
df_stud
```

Out[63]:

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
0	80	80.0	68	70	89	3	2019
1	71	71.0	61	85	91	3	2019
2	79	79.0	16	87	77	2	2018
3	61	61.0	77	74	76	2	2020
4	78	78.0	71	67	90	3	2019
5	73	73.0	68	90	80	2	2019
6	77	77.0	62	70	35	2	2020
7	74	74.0	45	80	12	1	2019
8	76	76.0	60	79	77	2	2020
9	75	75.0	65	85	87	3	2018
10	160	104.0	67	12	83	2	2020
11	79	79.0	72	88	180	2	2019
12	80	80.0	80	78	94	3	2021
13	78	78.0	69	71	90	3	2019
14	75	75.0	1	71	81	2	2019
15	78	78.0	62	79	93	3	2021
16	86	86.0	78	80	88	3	2019
17	80	80.0	74	23	76	2	2021
18	75	75.0	62	86	87	3	2019
19	82	82.0	70	87	94	3	2019
20	69	69.0	65	84	35	1	2018
21	100	100.0	77	70	91	3	2018
22	72	72.0	60	78	94	3	2019
23	74	74.0	65	71	84	2	2019
24	75	75.0	77	83	77	2	2020
25	180	104.0	67	63	75	3	2021
26	72	72.0	120	70	84	2	2021
27	71	71.0	79	88	85	3	2021
28	120	104.0	73	71	94	3	2019

In [64]:

```
col = ['reading
score']
```

Out[64]:

```
<AxesSubplot:>
```

In [65]:

```
median=np.median(sorted_rscore)
median
```

Out[65]:

```
68.0
```

In [69]:

```
refined_df=df
```

In [70]:

```
refined_df['reading score'] = np.where(refined_df['reading score'] >upr_bound, med
```

In [72]:

```
print(refined_df)
```

```
    math score  m score  reading score  writing score  placement score
\ 0         80     80.0           68.0             70               89
1           71     71.0           61.0             85               91
2           79     79.0           16.0             87               77
3           61     61.0           77.0             74               76
4           78     78.0           71.0             67               90
5           73     73.0           68.0             90               80
6           77     77.0           62.0             70               35
7           74     74.0           45.0             80               12
8           76     76.0           60.0             79               77
9           75     75.0           65.0             85               87
10         160    104.0           67.0             12               83
11          79     79.0           72.0             88              180
12          80     80.0           80.0             78               94
13          78     78.0           69.0             71               90
14          75     75.0            1.0             71               81
15          78     78.0           62.0             79               93
16          86     86.0           78.0             80               88
17          80     80.0           74.0             23               76
18          75     75.0           62.0             86               87
19          82     82.0           70.0             87               94
20          69     69.0           65.0             84               35
21         100    100.0           77.0             70               91
22          72     72.0           60.0             78               94
23          74     74.0           65.0             71               84
24          75     75.0           77.0             83               77
25         180    104.0           67.0             63               75
26          72     72.0           68.0             70               84
27          71     71.0           79.0             88               85
28         120    104.0           73.0             71               94

    placement offer count  club join
year 0                  3            2019
1                       3            2019
2                       2            2018
3                       2            2020
4                       3            2019
5                       2            2019
6                       2            2020
7                       1            2019
8                       2            2020
9                       3            2018
10                      2            2020
11                      2            2019
12                      3            2021
13                      3            2019
14                      2            2019
15                      3            2021
16                      3            2019
17                      2            2021
18                      3            2019
19                      3            2019
20                      1            2018
21                      3            2018
22                      3            2019
23                      2            2019
24                      2            2020
25                      3            2021
26                      2            2021
27                      3            2021
28                      3            2019
```

```
In [73]: refined_df['reading score'] = np.where(refined_df['reading score'] <lwr_bound, med
```

In [75]:

```
col = ['readingscore']
refined_df.boxplot(col)
```

Out[75]:

```
<AxesSubplot:>
```