In [16]:
```
#Name: Atharv Santosh Danave
#Roll No: 11
#Practical no: 02
#Academic year: 2024-25
```

In [2]:
```
import pandas as pd
```

In [3]:
```
import numpy as np
```

In [4]:
```
df=pd.read_csv("/home/jaihind/Desktop/StudentPerformance.csv")
```

In [5]:
```
df
```

Out[5]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72	72	74.0	78.0	1	Pune
1	female	69	90	88.0	NaN	2	na
2	female	90	95	93.0	74.0	2	Nashik
3	male	47	57	NaN	78.0	1	Na
4	male	na	78	75.0	81.0	3	Pune
5	female	71	Na	78.0	70.0	4	na
6	male	12	44	52.0	12.0	2	Nashik
7	male	NaN	65	67.0	49.0	1	Pune
8	female	5	77	89.0	55.0	0	NaN

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

Out[6]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	False	False	False	False	False	False	False
1	False	False	False	False	True	False	False
2	False	False	False	False	False	False	False
3	False	False	False	True	False	False	False
4	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False
7	False	True	False	False	False	False	False
8	False	False	False	False	False	False	True

In [7]:
```
series=pd.isnull(df["math score"])
df[series]
```

Out[7]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
7	male	NaN	65	67.0	49.0	1	Pune

In [8]:
```
import pandas as pd
import numpy as np
```

In [9]:
```
df.notnull()
```

Out[9]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	True	True	True	True	True	True	True
1	True	True	True	True	False	True	True
2	True	True	True	True	True	True	True
3	True	True	True	False	True	True	True
4	True	True	True	True	True	True	True
5	True	True	True	True	True	True	True
6	True	True	True	True	True	True	True
7	True	False	True	True	True	True	True
8	True	True	True	True	True	True	False

In [10]:
```
series=pd.notnull(df["math score"])
df[series]
```

Out[10]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72	72	74.0	78.0	1	Pune
1	female	69	90	88.0	NaN	2	na
2	female	90	95	93.0	74.0	2	Nashik
3	male	47	57	NaN	78.0	1	Na
4	male	na	78	75.0	81.0	3	Pune
5	female	71	Na	78.0	70.0	4	na
6	male	12	44	52.0	12.0	2	Nashik
8	female	5	77	89.0	55.0	0	NaN

In [11]:
```
from sklearn.preprocessing import LabelEncoder
```

In [12]:
```
from sklearn.preprocessing import LabelEncoder
```

In [13]:
```
le=LabelEncoder()
```

In [14]:
```
df['gender']=le.fit_transform(df['gender'])
```

In [15]:
```
newdf=df
df
```

Out[15]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	0	72	72	74.0	78.0	1	Pune
1	0	69	90	88.0	NaN	2	na
2	0	90	95	93.0	74.0	2	Nashik
3	1	47	57	NaN	78.0	1	Na
4	1	na	78	75.0	81.0	3	Pune
5	0	71	Na	78.0	70.0	4	na
6	1	12	44	52.0	12.0	2	Nashik
7	1	NaN	65	67.0	49.0	1	Pune
8	0	5	77	89.0	55.0	0	NaN

In [16]:
```
missing_values=["Na","na"]
```

In [17]:
```
df=pd.read_csv("/home/jaihind/Desktop/StudentPerformance.csv",na_values=mi
```

In [18]:
```
df
```

Out[18]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.0	72.0	74.0	78.0	1	Pune
1	female	69.0	90.0	88.0	NaN	2	NaN
2	female	90.0	95.0	93.0	74.0	2	Nashik
3	male	47.0	57.0	NaN	78.0	1	NaN
4	male	NaN	78.0	75.0	81.0	3	Pune
5	female	71.0	NaN	78.0	70.0	4	NaN
6	male	12.0	44.0	52.0	12.0	2	Nashik
7	male	NaN	65.0	67.0	49.0	1	Pune
8	female	5.0	77.0	89.0	55.0	0	NaN

In [19]:
```
ndf=df
```

In [20]:
```
ndf.fillna(0)
```

Out[20]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.0	72.0	74.0	78.0	1	Pune
1	female	69.0	90.0	88.0	0.0	2	0
2	female	90.0	95.0	93.0	74.0	2	Nashik
3	male	47.0	57.0	0.0	78.0	1	0
4	male	0.0	78.0	75.0	81.0	3	Pune
5	female	71.0	0.0	78.0	70.0	4	0
6	male	12.0	44.0	52.0	12.0	2	Nashik
7	male	0.0	65.0	67.0	49.0	1	Pune
8	female	5.0	77.0	89.0	55.0	0	0

In [21]:
```
m_v=df['math score'].mean()
```

In [22]:
```
df['math score'].fillna(value=m_v, inplace=True)
```

In [23]:
```
df
```

Out[23]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.000000	72.0	74.0	78.0	1	Pune
1	female	69.000000	90.0	88.0	NaN	2	NaN
2	female	90.000000	95.0	93.0	74.0	2	Nashik
3	male	47.000000	57.0	NaN	78.0	1	NaN
4	male	52.285714	78.0	75.0	81.0	3	Pune
5	female	71.000000	NaN	78.0	70.0	4	NaN
6	male	12.000000	44.0	52.0	12.0	2	Nashik
7	male	52.285714	65.0	67.0	49.0	1	Pune
8	female	5.000000	77.0	89.0	55.0	0	NaN

In [24]:
```
ndf.replace(to_replace=np.nan,value=-99)
```

Out[24]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.000000	72.0	74.0	78.0	1	Pune
1	female	69.000000	90.0	88.0	-99.0	2	-99
2	female	90.000000	95.0	93.0	74.0	2	Nashik
3	male	47.000000	57.0	-99.0	78.0	1	-99
4	male	52.285714	78.0	75.0	81.0	3	Pune
5	female	71.000000	-99.0	78.0	70.0	4	-99
6	male	12.000000	44.0	52.0	12.0	2	Nashik
7	male	52.285714	65.0	67.0	49.0	1	Pune
8	female	5.000000	77.0	89.0	55.0	0	-99

In [25]:
```
ndf.dropna()
```

Out[25]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.000000	72.0	74.0	78.0	1	Pune
2	female	90.000000	95.0	93.0	74.0	2	Nashik
4	male	52.285714	78.0	75.0	81.0	3	Pune
6	male	12.000000	44.0	52.0	12.0	2	Nashik
7	male	52.285714	65.0	67.0	49.0	1	Pune

In [26]:
```
ndf.dropna(how='all')
```

Out[26]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.000000	72.0	74.0	78.0	1	Pune
1	female	69.000000	90.0	88.0	NaN	2	NaN
2	female	90.000000	95.0	93.0	74.0	2	Nashik
3	male	47.000000	57.0	NaN	78.0	1	NaN
4	male	52.285714	78.0	75.0	81.0	3	Pune
5	female	71.000000	NaN	78.0	70.0	4	NaN
6	male	12.000000	44.0	52.0	12.0	2	Nashik
7	male	52.285714	65.0	67.0	49.0	1	Pune
8	female	5.000000	77.0	89.0	55.0	0	NaN

In [27]:
```
ndf.dropna(axis=1)
```

Out[27]:

	gender	math score	placement offer count
0	female	72.000000	1
1	female	69.000000	2
2	female	90.000000	2
3	male	47.000000	1
4	male	52.285714	3
5	female	71.000000	4
6	male	12.000000	2
7	male	52.285714	1
8	female	5.000000	0

In [28]:
```
new_data=ndf.dropna(axis=0, how='any')
```

In [29]:
```
new_data
```

Out[29]:

	gender	math score	reading score	writing score	placement score	placement offer count	Region
0	female	72.000000	72.0	74.0	78.0	1	Pune
2	female	90.000000	95.0	93.0	74.0	2	Nashik
4	male	52.285714	78.0	75.0	81.0	3	Pune
6	male	12.000000	44.0	52.0	12.0	2	Nashik
7	male	52.285714	65.0	67.0	49.0	1	Pune

In [ ]:

In [2]:
```
import pandas as pd
```

In [3]:
```
import numpy as np
```

In [4]:
```
df=pd.read_csv("/home/jaihind/Downloads/demo1(1).csv")
```

In [5]:
```
df
```

Out[5]:

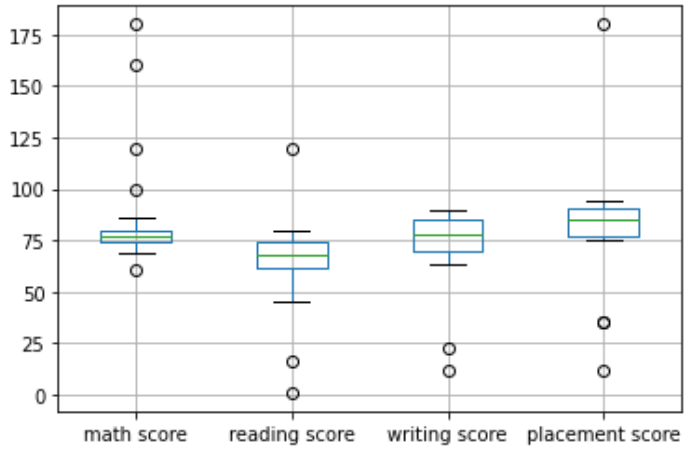
	math score	reading score	writing score	placement score	placement offer count	club join year
0	80	68	70	89	3	2019
1	71	61	85	91	3	2019
2	79	16	87	77	2	2018
3	61	77	74	76	2	2020
4	78	71	67	90	3	2019
5	73	68	90	80	2	2019
6	77	62	70	35	2	2020
7	74	45	80	12	1	2019
8	76	60	79	77	2	2020
9	75	65	85	87	3	2018
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 [6]:
```
import matplotlib.pyplot as plt
```

In [7]:
```
col=['math score','reading score','writing score','placement score']
```

In [8]:
```
df.boxplot(col)
```

Out[8]: <AxesSubplot:>



In [9]:
```
print(np.where(df['math score']>90))
```

```
(array([10, 21, 25, 28]),)
```

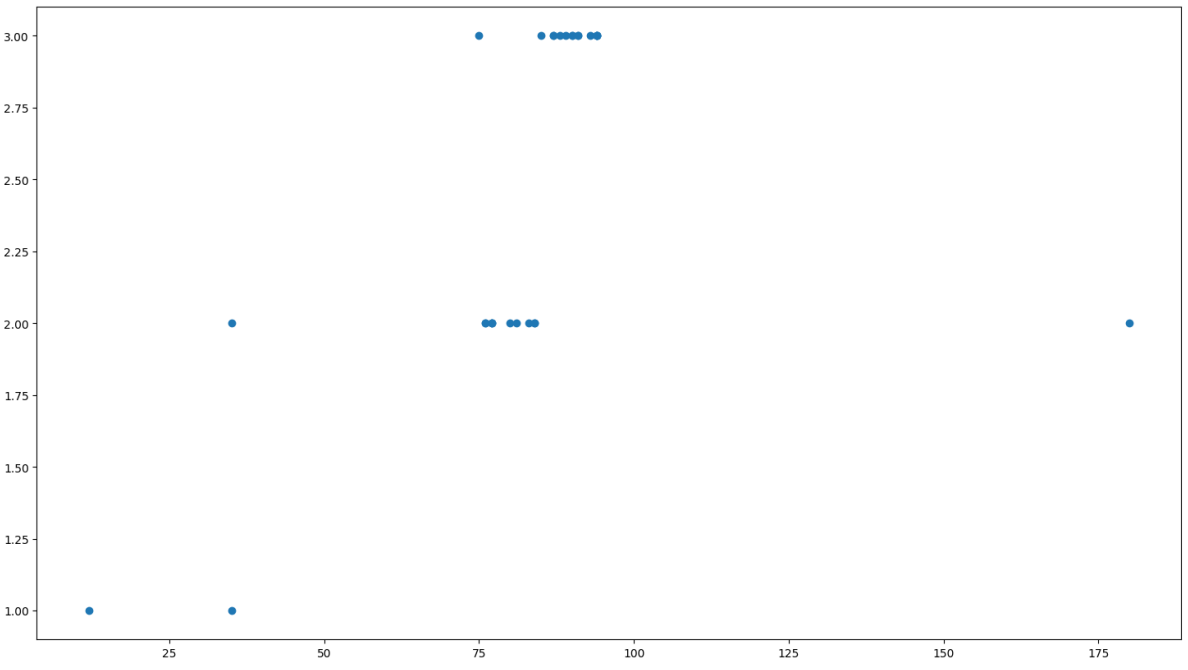
In [10]:
```
print(np.where(df['reading score']<25))
```

```
(array([ 2, 14]),)
```

In [11]:
```
print(np.where(df['writing score']<30))
```

```
(array([10, 17]),)
```

In [12]:
```
fig,ax=plt.subplots(figsize=(18,10))
ax.scatter(df['placement score'],df['placement offer count'])
plt.show()
```



In [13]:
```
ax.set_ylabel('(Proportion non-retail busines acres)/(town)')
```

Out[13]: Text(4.444444444444452, 0.5, '(Proportion non-retail busines acres)/(town)')

In [14]:
```
ax.set_ylabel('(Full-value property-tax rate)/($10,000)')
```

Out[14]:
```
Text(4.444444444444452, 0.5, '(Full-value property-tax rate)/($10,000)')
```

In [15]:
```
print(np.where((df['placement score']<50) & (df['placement offer count']>1)))
```

```
(array([6]),)
```

In [16]:
```
print(np.where((df['placement score']>85) & (df['placement offer count']<3)))
```

```
(array([11]),)
```

In [17]:
```
from scipy import stats
```

In [18]:
```
z=np.abs(stats.zscore(df['math score']))
```

In [19]:
```
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 [20]:
```
threshold=0.18
```

In [21]:
```
sample_outliers=np.where(z<threshold)
```

In [22]:
```
sample_outliers
```

Out[22]:
```
(array([ 0, 12, 16, 17, 19]),)
```

In [23]:
```
sorted_rscore=sorted(df['reading score'])
```

In [24]:
```
sorted_rscore
```

Out[24]:
```
[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 [25]:
```
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
```

```
62.0 74.0
```

In [26]:
```
IQR = q3-q1
```

In [27]:
```
lwr_bound = q1-(1.5*IQR)
upr_bound = q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

```
44.0 92.0
```

In [28]:
```
r_outliers = []
for i in sorted_rscore:
    if (i<lwr_bound or i>upr_bound):
        r_outliers.append(i)
print(r_outliers)
```

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

In [29]:
```
new_df=df
```

In [30]:
```
for i in sample_outliers:
    new_df.drop(i,inplace=True)
new_df
```

Out[30]:

	math score	reading score	writing score	placement score	placement offer count	club join year
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
13	78	69	71	90	3	2019
14	75	1	71	81	2	2019
15	78	62	79	93	3	2021
18	75	62	86	87	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 [31]:
```
df=pd.read_csv("/home/jaihind/Downloads/demo1(1).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_stud['math score'])
print("New array:",b)
```

```
New array: [ 80.  71.  79.  61.  78.  73.  77.  74.  76.  75. 104.  79.  80.  7
8.
  75.  78.  86.  80.  75.  82.  69. 100.  72.  74.  75. 104.  72.  71.
 104.]
```

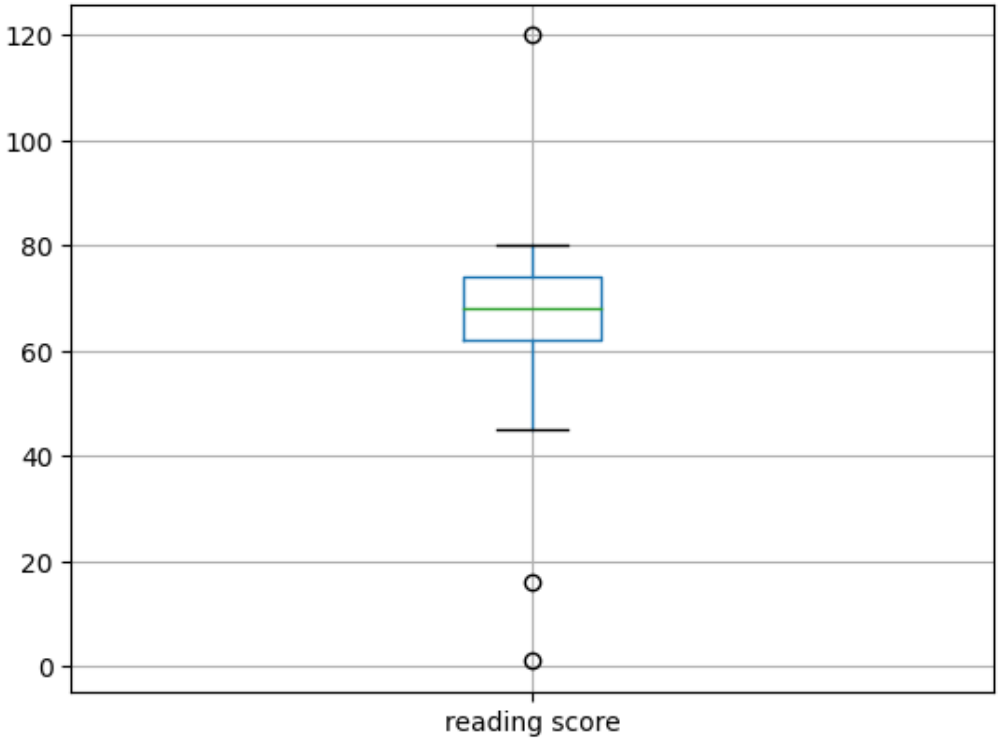
In [33]:
```
df_stud.insert(1,"m score",b,True)
df_stud
```

Out[33]:

	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 [34]:
```
col=['reading score']
df.boxplot(col)
```

Out[34]: <AxesSubplot:>



In [35]:
```
median=np.median(sorted_rscore)
median
```

Out[35]: 68.0

In [37]:
```
refined_df=df
refined_df['reading score']=np.where(refined_df['reading score']>upr_bound,media
refined_df
```

Out[37]:

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
0	80	80.0	68.0	70	89	3	2019
1	71	71.0	61.0	85	91	3	2019
2	79	79.0	16.0	87	77	2	2018
3	61	61.0	77.0	74	76	2	2020
4	78	78.0	71.0	67	90	3	2019
5	73	73.0	68.0	90	80	2	2019
6	77	77.0	62.0	70	35	2	2020
7	74	74.0	45.0	80	12	1	2019
8	76	76.0	60.0	79	77	2	2020
9	75	75.0	65.0	85	87	3	2018
10	160	104.0	67.0	12	83	2	2020
11	79	79.0	72.0	88	180	2	2019
12	80	80.0	80.0	78	94	3	2021
13	78	78.0	69.0	71	90	3	2019
14	75	75.0	1.0	71	81	2	2019
15	78	78.0	62.0	79	93	3	2021
16	86	86.0	78.0	80	88	3	2019
17	80	80.0	74.0	23	76	2	2021
18	75	75.0	62.0	86	87	3	2019
19	82	82.0	70.0	87	94	3	2019
20	69	69.0	65.0	84	35	1	2018
21	100	100.0	77.0	70	91	3	2018
22	72	72.0	60.0	78	94	3	2019
23	74	74.0	65.0	71	84	2	2019
24	75	75.0	77.0	83	77	2	2020
25	180	104.0	67.0	63	75	3	2021
26	72	72.0	68.0	70	84	2	2021
27	71	71.0	79.0	88	85	3	2021
28	120	104.0	73.0	71	94	3	2019

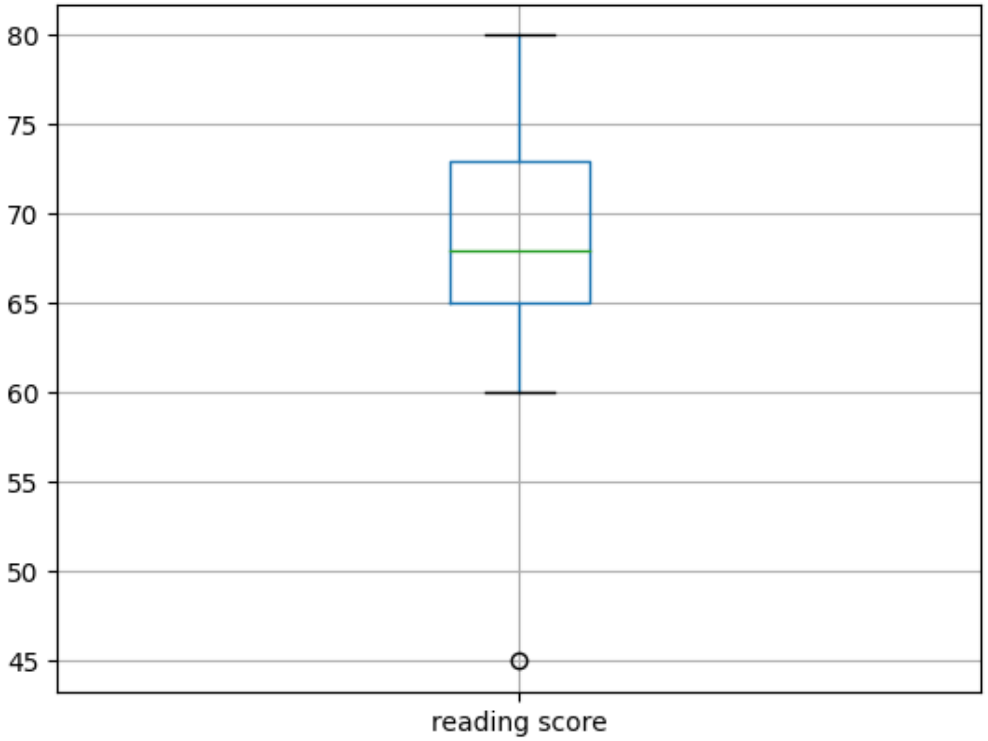
In [38]:
```
refined_df['reading score']=np.where(refined_df['reading score']<lwr_bound,media
refined_df
```

Out[38]:

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
0	80	80.0	68.0	70	89	3	2019
1	71	71.0	61.0	85	91	3	2019
2	79	79.0	68.0	87	77	2	2018
3	61	61.0	77.0	74	76	2	2020
4	78	78.0	71.0	67	90	3	2019
5	73	73.0	68.0	90	80	2	2019
6	77	77.0	62.0	70	35	2	2020
7	74	74.0	45.0	80	12	1	2019
8	76	76.0	60.0	79	77	2	2020
9	75	75.0	65.0	85	87	3	2018
10	160	104.0	67.0	12	83	2	2020
11	79	79.0	72.0	88	180	2	2019
12	80	80.0	80.0	78	94	3	2021
13	78	78.0	69.0	71	90	3	2019
14	75	75.0	68.0	71	81	2	2019
15	78	78.0	62.0	79	93	3	2021
16	86	86.0	78.0	80	88	3	2019
17	80	80.0	74.0	23	76	2	2021
18	75	75.0	62.0	86	87	3	2019
19	82	82.0	70.0	87	94	3	2019
20	69	69.0	65.0	84	35	1	2018
21	100	100.0	77.0	70	91	3	2018
22	72	72.0	60.0	78	94	3	2019
23	74	74.0	65.0	71	84	2	2019
24	75	75.0	77.0	83	77	2	2020
25	180	104.0	67.0	63	75	3	2021
26	72	72.0	68.0	70	84	2	2021
27	71	71.0	79.0	88	85	3	2021
28	120	104.0	73.0	71	94	3	2019

In [39]:
```
col=['reading score']
refined_df.boxplot(col)
```

Out[39]: <AxesSubplot:>



In [40]:
```
import pandas as pd
import numpy as np
df=pd.read_csv("/home/jaihind/Downloads/demo1(1).csv")
```

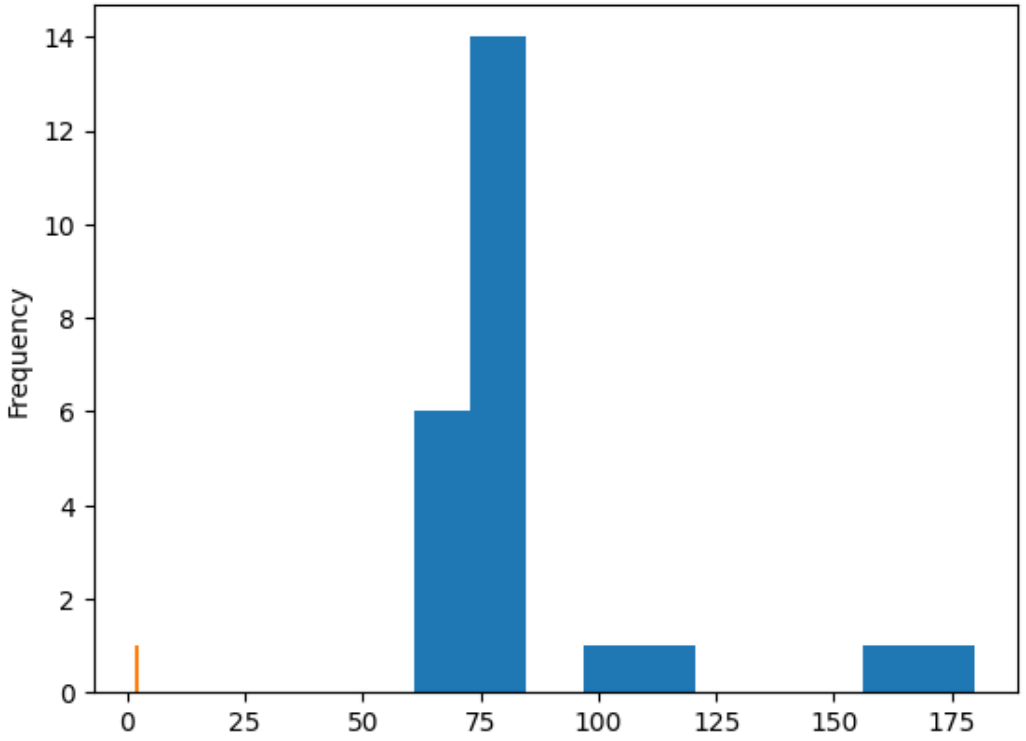
In [41]:
```
df
```

Out[41]:

	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 [43]:
```
import matplotlib.pyplot as plt
new_df['math score'].plot(kind='hist')
df['log_math']=np.log10(df['math score'])
df['log_math'].plot(kind='hist')
```

Out[43]: <AxesSubplot:ylabel='Frequency'>



In [ ]: