In [1]:

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

In [2]:

```
df = pd.read_csv("demo1.csv")
df
```

Out[2]:

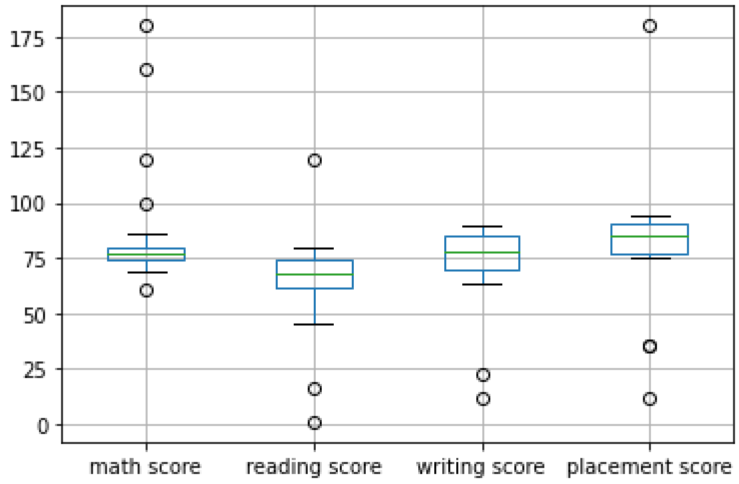
	math score	reading score	writing score	placement score	placement offer count	club join year
0	80	68	70	89	3	2019
1	71	61	85	91	3	2019
2	79	16	87	77	2	2018
3	61	77	74	76	2	2020
4	78	71	67	90	3	2019
5	73	68	90	80	2	2019
6	77	62	70	35	2	2020
7	74	45	80	12	1	2019
8	76	60	79	77	2	2020
9	75	65	85	87	3	2018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

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

Out[3]:

```
<AxesSubplot:>
```



In [4]:

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

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

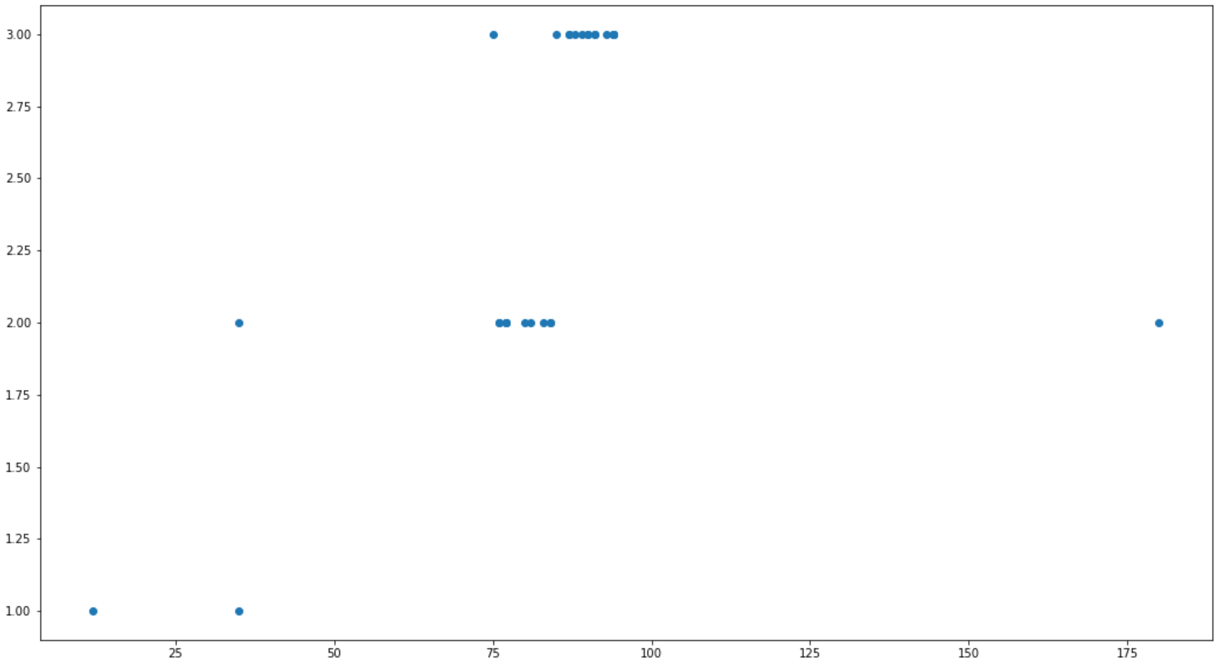
In [5]:

```
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 [6]:

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



In [7]:

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

Out[7]:

```
Text(3.200000000000017, 0.5, '(Full-value property-tax rate)/($10,000)')
```

In [8]:

```
print(np.where((df['placement score']<50) & (df['placement offer count']>1)))
print(np.where((df['placement score']>85) & (df['placement offer count']<3)))
```

```
(array([6], dtype=int64),)
(array([11], dtype=int64),)
```

Detecting outliers using Z-Score

In [9]:

```
from scipy import stats
```

In [10]:

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

In [11]:

```
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 [12]:

```
threshold = 0.18
```

In [13]:

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

Out[13]:

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

Detecting outliers using Inter Quantile Range(IQR):

In [14]:

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

In [15]:

```
sorted_rscore
```

Out[15]:

```
[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 [16]:

```
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
```

```
62.0 74.0
```

In [17]:

```
IQR = q3-q1
```

In [18]:

```
lwr_bound = q1-(1.5*IQR)
upr_bound = q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

```
44.0 92.0
```

In [19]:

```
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 [20]:

```
new_df=df
for i in sample_outliers:
    new_df.drop(i,inplace=True)
new_df
```

Out[20]:

	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

• Quantile based flooring and capping:

In [21]:

```
df_stud=df
```

In [22]:

```
ninetieth_percentile = np.percentile(df_stud['math score'], 90)
```

In [23]:

```
b = np.where(df_stud['math score']>ninetieth_percentile,
ninetieth_percentile, df_stud['math score'])
print("New Array : ", b)
```

```
New Array :  [ 71.  79.  61.  78.  73.  77.  74.  76.  75. 114.  79.  78.  75.  78.
  75.  69. 100.  72.  74.  75. 114.  72.  71. 114.]
```

In [24]:

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

Out[24]:

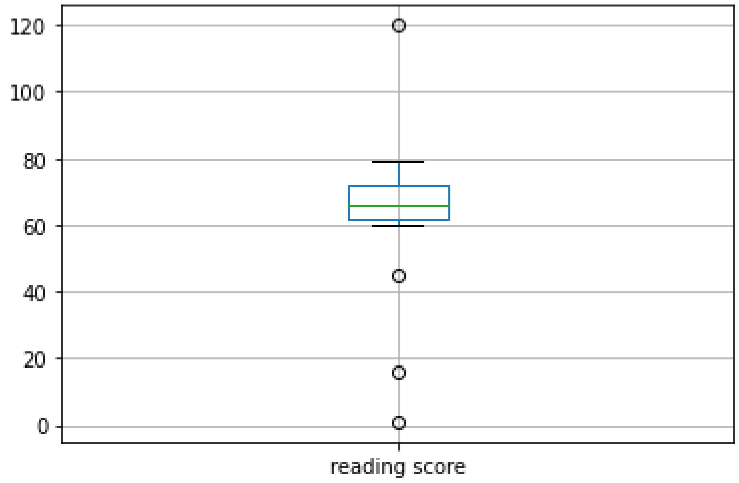
	math score	m score	reading score	writing score	placement score	placement offer count	club join year
1	71	71.0	61	85	91	3	2019
2	79	79.0	16	87	77	2	2018
3	61	61.0	77	74	76	2	2020
4	78	78.0	71	67	90	3	2019
5	73	73.0	68	90	80	2	2019
6	77	77.0	62	70	35	2	2020
7	74	74.0	45	80	12	1	2019
8	76	76.0	60	79	77	2	2020
9	75	75.0	65	85	87	3	2018
10	160	114.0	67	12	83	2	2020
11	79	79.0	72	88	180	2	2019
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
18	75	75.0	62	86	87	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	114.0	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	114.0	73	71	94	3	2019

In [25]:

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

Out[25]:

```
<AxesSubplot:>
```



In [26]:
```
median=np.median(sorted_rscore)
median
```

Out[26]:
```
68.0
```

In [27]:
```
refined_df=df
refined_df['reading score'] = np.where(refined_df['reading score'] >upr_bound, media
```

In [28]:
```
refined_df
```

Out[28]:

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
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	114.0	67.0	12	83	2	2020
11	79	79.0	72.0	88	180	2	2019
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
18	75	75.0	62.0	86	87	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

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
23	74	74.0	65.0	71	84	2	2019
24	75	75.0	77.0	83	77	2	2020
25	180	114.0	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	114.0	73.0	71	94	3	2019

In [29]:

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

In [30]:

```
refined_df
```

Out[30]:

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
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	114.0	67.0	12	83	2	2020
11	79	79.0	72.0	88	180	2	2019
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
18	75	75.0	62.0	86	87	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	114.0	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

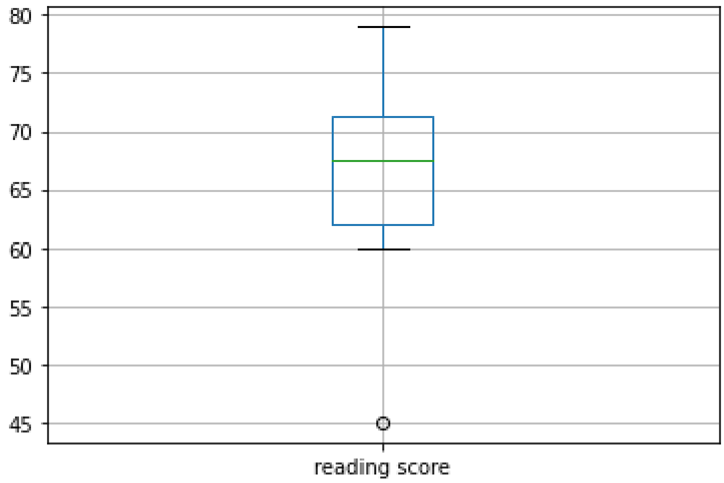
	math score	m score	reading score	writing score	placement score	placement offer count	club join year
28	120	114.0	73.0	71	94	3	2019

In [31]:

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

Out[31]:

```
<AxesSubplot:>
```



In [32]:

```
df
```

Out[32]:

	math score	m score	reading score	writing score	placement score	placement offer count	club join year
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	114.0	67.0	12	83	2	2020
11	79	79.0	72.0	88	180	2	2019
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
18	75	75.0	62.0	86	87	3	2019
20	69	69.0	65.0	84	35	1	2018
21	100	100.0	77.0	70	91	3	2018

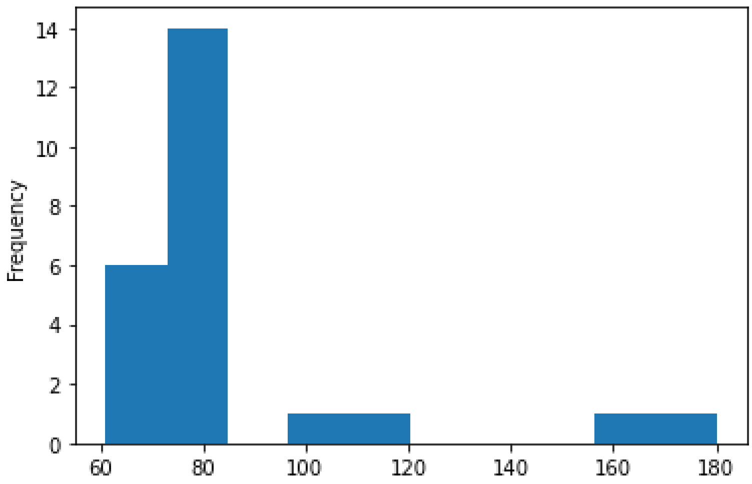
	math score	m score	reading score	writing score	placement score	placement offer count	club join year
22	72	72.0	60.0	78	94	3	2019
23	74	74.0	65.0	71	84	2	2019
24	75	75.0	77.0	83	77	2	2020
25	180	114.0	67.0	63	75	3	2021
26	72	72.0	68.0	70	84	2	2021
27	71	71.0	79.0	88	85	3	2021
28	120	114.0	73.0	71	94	3	2019

In [33]:

```
import matplotlib.pyplot as plt
new_df['math score'].plot(kind = 'hist')
```

Out[33]:

```
<AxesSubplot:ylabel='Frequency'>
```



In [34]:

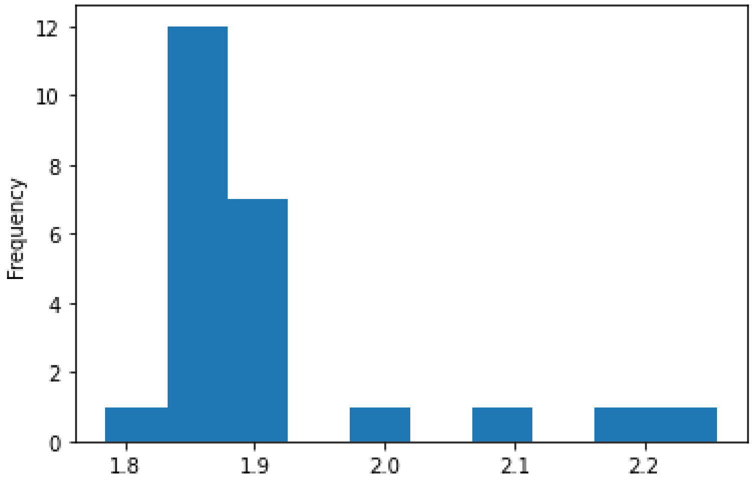
```
df['log_math'] = np.log10(df['math score'])
```

In [35]:

```
df['log_math'].plot(kind = 'hist')
```

Out[35]:

```
<AxesSubplot:ylabel='Frequency'>
```



In [ ]: