In [1]:

```
pip install scikit-learn
```

```
Defaulting to user installation because normal site-packages is not wr
iteable
Requirement already satisfied: scikit-learn in ./.local/lib/python3.8/
site-packages (1.2.1)
Requirement already satisfied: scipy>=1.3.2 in ./.local/lib/python3.8/
site-packages (from scikit-learn) (1.10.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in ./.local/lib/py
thon3.8/site-packages (from scikit-learn) (3.1.0)
Requirement already satisfied: numpy>=1.17.3 in ./.local/lib/python3.
8/site-packages (from scikit-learn) (1.22.4)
Requirement already satisfied: joblib>=1.1.1 in ./.local/lib/python3.
8/site-packages (from scikit-learn) (1.2.0)
WARNING: You are using pip version 22.0.4; however, version 23.0.1 is
 available.
You should consider upgrading via the '/usr/bin/python3 -m pip install
--upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

In [57]:

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

In [3]:

```
dict={'gender':['f','m','f','f','m','f','m','f','m','m'],'maths_score':[78,12,67,30
      'reading_score':[70,72,12,76,89,90,np.nan,86,79,200],'writing_score':[70,72,7
      'placement_score':[20,81,82,83,84,90,300,98,np.nan,90],
      'college_joining_year':[2019,2020,2021,2020,2020,2019,2020,2021,202,2021],
      'placement_offer_count':[1,1,1,1,0,np.nan,np.nan,np.nan,3,4],
      'region':['Pune','Nashik','Mumbai','Pune','Pune','Pune','Nashik','Mumbai','Mu
```

In [58]:

```
df=pd.DataFrame(data=dict)
```

In [59]:

```
df
```

Out[59]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	70	20.0	2019
1	m	12	72.0	72	81.0	2020
2	f	67	12.0	74	82.0	2021
3	f	30	76.0	76	83.0	2020
4	m	80	89.0	89	84.0	2020
5	f	67	90.0	90	90.0	2019
6	m	1112	NaN	87	300.0	2020
7	f	72	86.0	86	98.0	2021
8	m	73	79.0	79	NaN	202
9	m	79	200.0	70	90.0	2021

In [60]:

```
df.isnull()
```

Out[60]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
5	False	False	False	False	False	False
6	False	False	True	False	False	False
7	False	False	False	False	False	False
8	False	False	False	False	True	False
9	False	False	False	False	False	False

In [61]:

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

Out[61]:

```
gender                  0
maths_score             0
reading_score           1
writing_score           0
placement_score         1
college_joining_year    0
placement_offer_count   3
region                  0
dtype: int64
```

In [8]:

```
df.notnull()
```

Out[8]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	True	True	True	True	True	True
1	True	True	True	True	True	True
2	True	True	True	True	True	True
3	True	True	True	True	True	True
4	True	True	True	True	True	True
5	True	True	True	True	True	True
6	True	True	False	True	True	True
7	True	True	True	True	True	True
8	True	True	True	True	False	True
9	True	True	True	True	True	True

In [9]:

```
s=pd.isnull(df['maths_score'])
```

In [10]:

```
df
```

Out[10]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	70	20.0	2019
1	m	12	72.0	72	81.0	2020
2	f	67	12.0	74	82.0	2021
3	f	30	76.0	76	83.0	2020
4	m	80	89.0	89	84.0	2020
5	f	67	90.0	90	90.0	2019
6	m	1112	NaN	87	300.0	2020
7	f	72	86.0	86	98.0	2021
8	m	73	79.0	79	NaN	202
9	m	79	200.0	70	90.0	2021

In [11]:

```
s=pd.notnull(df['maths_score'])
df
```

Out[11]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	70	20.0	2019
1	m	12	72.0	72	81.0	2020
2	f	67	12.0	74	82.0	2021
3	f	30	76.0	76	83.0	2020
4	m	80	89.0	89	84.0	2020
5	f	67	90.0	90	90.0	2019
6	m	1112	NaN	87	300.0	2020
7	f	72	86.0	86	98.0	2021
8	m	73	79.0	79	NaN	202
9	m	79	200.0	70	90.0	2021

In [12]:

```
df['maths_score']=df['maths_score'].fillna(df['maths_score']).mean()
df
```

Out[12]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	167.0	70.0	70	20.0	2019
1	m	167.0	72.0	72	81.0	2020
2	f	167.0	12.0	74	82.0	2021
3	f	167.0	76.0	76	83.0	2020
4	m	167.0	89.0	89	84.0	2020
5	f	167.0	90.0	90	90.0	2019
6	m	167.0	NaN	87	300.0	2020
7	f	167.0	86.0	86	98.0	2021
8	m	167.0	79.0	79	NaN	202
9	m	167.0	200.0	70	90.0	2021

In [13]:

```
df=pd.DataFrame(data=dict)
df
```

Out[13]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	70	20.0	2019
1	m	12	72.0	72	81.0	2020
2	f	67	12.0	74	82.0	2021
3	f	30	76.0	76	83.0	2020
4	m	80	89.0	89	84.0	2020
5	f	67	90.0	90	90.0	2019
6	m	1112	NaN	87	300.0	2020
7	f	72	86.0	86	98.0	2021
8	m	73	79.0	79	NaN	202
9	m	79	200.0	70	90.0	2021

In [14]:

```
df['maths_score']=df['maths_score'].fillna(df['maths_score']).median()
df
```

Out[14]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	72.5	70.0	70	20.0	2019
1	m	72.5	72.0	72	81.0	2020
2	f	72.5	12.0	74	82.0	2021
3	f	72.5	76.0	76	83.0	2020
4	m	72.5	89.0	89	84.0	2020
5	f	72.5	90.0	90	90.0	2019
6	m	72.5	NaN	87	300.0	2020
7	f	72.5	86.0	86	98.0	2021
8	m	72.5	79.0	79	NaN	202
9	m	72.5	200.0	70	90.0	2021

In [15]:

```
df['reading_score']=df['reading_score'].fillna(df['reading_score']).median()
df
```

Out[15]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	72.5	79.0	70	20.0	2019
1	m	72.5	79.0	72	81.0	2020
2	f	72.5	79.0	74	82.0	2021
3	f	72.5	79.0	76	83.0	2020
4	m	72.5	79.0	89	84.0	2020
5	f	72.5	79.0	90	90.0	2019
6	m	72.5	79.0	87	300.0	2020
7	f	72.5	79.0	86	98.0	2021
8	m	72.5	79.0	79	NaN	202
9	m	72.5	79.0	70	90.0	2021

In [16]:

```
df=pd.DataFrame(data=dict)
df
```

Out[16]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	70	20.0	2019
1	m	12	72.0	72	81.0	2020
2	f	67	12.0	74	82.0	2021
3	f	30	76.0	76	83.0	2020
4	m	80	89.0	89	84.0	2020
5	f	67	90.0	90	90.0	2019
6	m	1112	NaN	87	300.0	2020
7	f	72	86.0	86	98.0	2021
8	m	73	79.0	79	NaN	202
9	m	79	200.0	70	90.0	2021

In [17]:

```
df['writing_score']=df['writing_score'].fillna(df['writing_score']).std()
df
```

Out[17]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
5	f	67	90.0	8.01457	90.0	2019
6	m	1112	NaN	8.01457	300.0	2020
7	f	72	86.0	8.01457	98.0	2021
8	m	73	79.0	8.01457	NaN	202
9	m	79	200.0	8.01457	90.0	2021

In [18]:

```
df1=df.dropna()
```

In [19]:

```
df1
```

Out[19]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
9	m	79	200.0	8.01457	90.0	2021

In [20]:

```
df1=df.dropna(axis=1)
df1
```

Out[20]:

	gender	maths_score	writing_score	college_joining_year	region
0	f	78	8.01457	2019	Pune
1	m	12	8.01457	2020	Nashik
2	f	67	8.01457	2021	Mumbai
3	f	30	8.01457	2020	Pune
4	m	80	8.01457	2020	Pune
5	f	67	8.01457	2019	Pune
6	m	1112	8.01457	2020	Nashik
7	f	72	8.01457	2021	Mumbai
8	m	73	8.01457	202	Mumbai
9	m	79	8.01457	2021	Pune

In [21]:

```
df3=df.dropna(how='all')
df3
```

Out[21]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
5	f	67	90.0	8.01457	90.0	2019
6	m	1112	NaN	8.01457	300.0	2020
7	f	72	86.0	8.01457	98.0	2021
8	m	73	79.0	8.01457	NaN	202
9	m	79	200.0	8.01457	90.0	2021

In [22]:

```
df4=df.dropna(axis=0,how="any")
df4
```

Out[22]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
9	m	79	200.0	8.01457	90.0	2021

In [23]:

```
df4=df.dropna(axis=1,how="any")
df4
```

Out[23]:

	gender	maths_score	writing_score	college_joining_year	region
0	f	78	8.01457	2019	Pune
1	m	12	8.01457	2020	Nashik
2	f	67	8.01457	2021	Mumbai
3	f	30	8.01457	2020	Pune
4	m	80	8.01457	2020	Pune
5	f	67	8.01457	2019	Pune
6	m	1112	8.01457	2020	Nashik
7	f	72	8.01457	2021	Mumbai
8	m	73	8.01457	202	Mumbai
9	m	79	8.01457	2021	Pune

In [24]:

```
df4=df.dropna(how="any")
df4
```

Out[24]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
9	m	79	200.0	8.01457	90.0	2021

In [25]:

```
df6=df.replace(to_replace=np.nan,value=60)
df6
```

Out[25]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
5	f	67	90.0	8.01457	90.0	2019
6	m	1112	60.0	8.01457	300.0	2020
7	f	72	86.0	8.01457	98.0	2021
8	m	73	79.0	8.01457	60.0	202
9	m	79	200.0	8.01457	90.0	2021

In [26]:

```
df7=df['placement_offer_count'].replace(to_replace=np.nan,value=2)
df7
```

Out[26]:

```
0    1.0
1    1.0
2    1.0
3    1.0
4    0.0
5    2.0
6    2.0
7    2.0
8    3.0
9    4.0
Name: placement_offer_count, dtype: float64
```

In [27]:

```
df
```

Out[27]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
5	f	67	90.0	8.01457	90.0	2019
6	m	1112	NaN	8.01457	300.0	2020
7	f	72	86.0	8.01457	98.0	2021
8	m	73	79.0	8.01457	NaN	202
9	m	79	200.0	8.01457	90.0	2021

In [28]:

```
sorted_rscore=sorted(df6['reading_score'])
sorted_rscore
```

Out[28]:

```
[12.0, 60.0, 70.0, 72.0, 76.0, 79.0, 86.0, 89.0, 90.0, 200.0]
```

In [29]:

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

```
70.5 88.25
```

In [30]:

```
IQR=q3-q1
IQR
```

Out[30]:

```
17.75
```

In [31]:

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

```
43.875 114.875
```

In [32]:

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

```
[12.0]
[12.0, 200.0]
```

In [33]:

```
pip install matplotlib
```

```
Defaulting to user installation because normal site-packages is not wr
iteable
Requirement already satisfied: matplotlib in ./.local/lib/python3.8/si
te-packages (3.7.0)
Requirement already satisfied: python-dateutil>=2.7 in ./.local/lib/py
thon3.8/site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: cycler>=0.10 in ./.local/lib/python3.8/
site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in ./.local/lib/pytho
n3.8/site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: pyparsing>=2.3.1 in ./.local/lib/python
3.8/site-packages (from matplotlib) (3.0.8)
Requirement already satisfied: packaging>=20.0 in ./.local/lib/python
3.8/site-packages (from matplotlib) (21.3)
Requirement already satisfied: pillow>=6.2.0 in /usr/lib/python3/dist-
packages (from matplotlib) (7.0.0)
Requirement already satisfied: importlib-resources>=3.2.0 in ./.local/
lib/python3.8/site-packages (from matplotlib) (5.7.1)
Requirement already satisfied: numpy>=1.20 in ./.local/lib/python3.8/s
ite-packages (from matplotlib) (1.22.4)
Requirement already satisfied: contourpy>=1.0.1 in ./.local/lib/python
3.8/site-packages (from matplotlib) (1.0.7)
Requirement already satisfied: fonttools>=4.22.0 in ./.local/lib/pytho
n3.8/site-packages (from matplotlib) (4.38.0)
Requirement already satisfied: zipp>=3.1.0 in ./.local/lib/python3.8/s
ite-packages (from importlib-resources>=3.2.0->matplotlib) (3.8.0)
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packa
ges (from python-dateutil>=2.7->matplotlib) (1.14.0)
WARNING: You are using pip version 22.0.4; however, version 23.0.1 is
 available.
You should consider upgrading via the '/usr/bin/python3 -m pip install
--upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

In [34]:

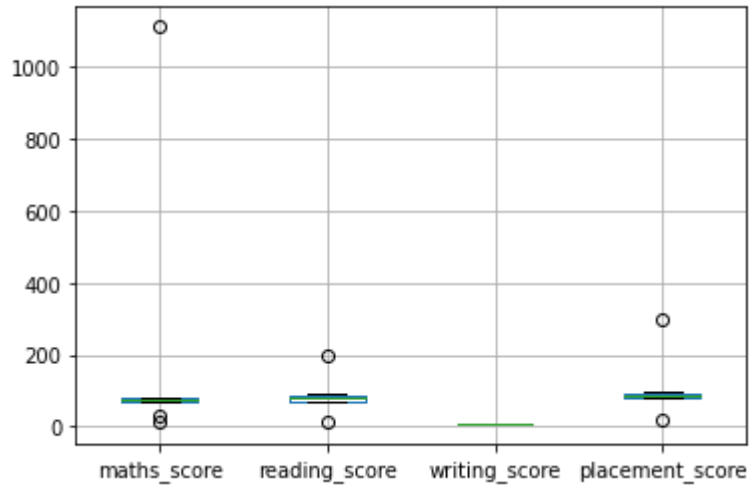
```
import matplotlib.pyplot as plt
```

In [35]:

```
col=['maths_score','reading_score','writing_score','placement_score']
df.boxplot(col)
```

Out[35]:

```
<Axes: >
```



In [36]:

```
sorted_rscore=sorted(df6['reading_score'])
sorted_rscore
```

Out[36]:

```
[12.0, 60.0, 70.0, 72.0, 76.0, 79.0, 86.0, 89.0, 90.0, 200.0]
```

In [37]:

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

```
70.5 88.25
```

In [38]:

```
IQR=q3-q1
IQR
```

Out[38]:

```
17.75
```

In [39]:

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

```
43.875 114.875
```

In [40]:

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

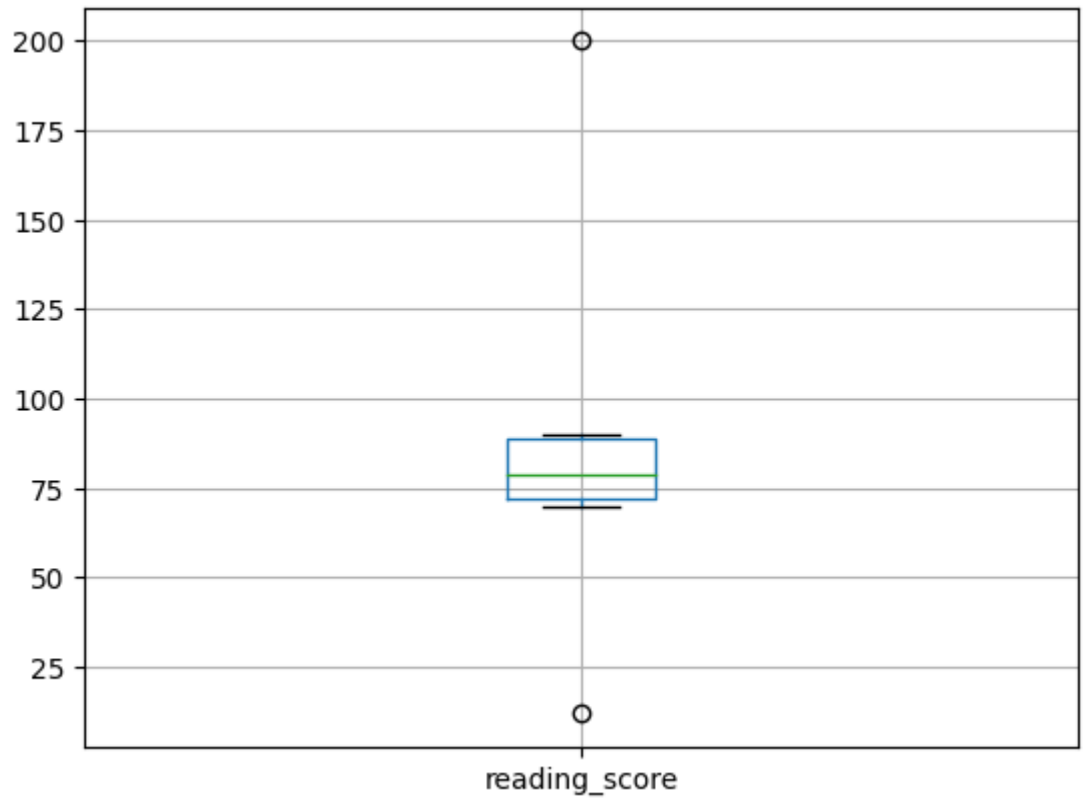
```
[12.0]
[12.0, 200.0]
```

In [41]:

```
col=['reading_score']
df.boxplot(col)
```

Out[41]:

```
<Axes: >
```



In [42]:

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

Out[42]:

77.5

In [43]:

```
refined_df=df
refined_df['reading_score']=np.where(refined_df['reading_score']>upr_bound,median,r
```

In [44]:

```
df
```

Out[44]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	12.0	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
5	f	67	90.0	8.01457	90.0	2019
6	m	1112	NaN	8.01457	300.0	2020
7	f	72	86.0	8.01457	98.0	2021
8	m	73	79.0	8.01457	NaN	202
9	m	79	77.5	8.01457	90.0	2021

In [45]:

```
refined_df=df
refined_df['reading_score']=np.where(refined_df['reading_score']<iwr_bound,median,r
df
```

Out[45]:

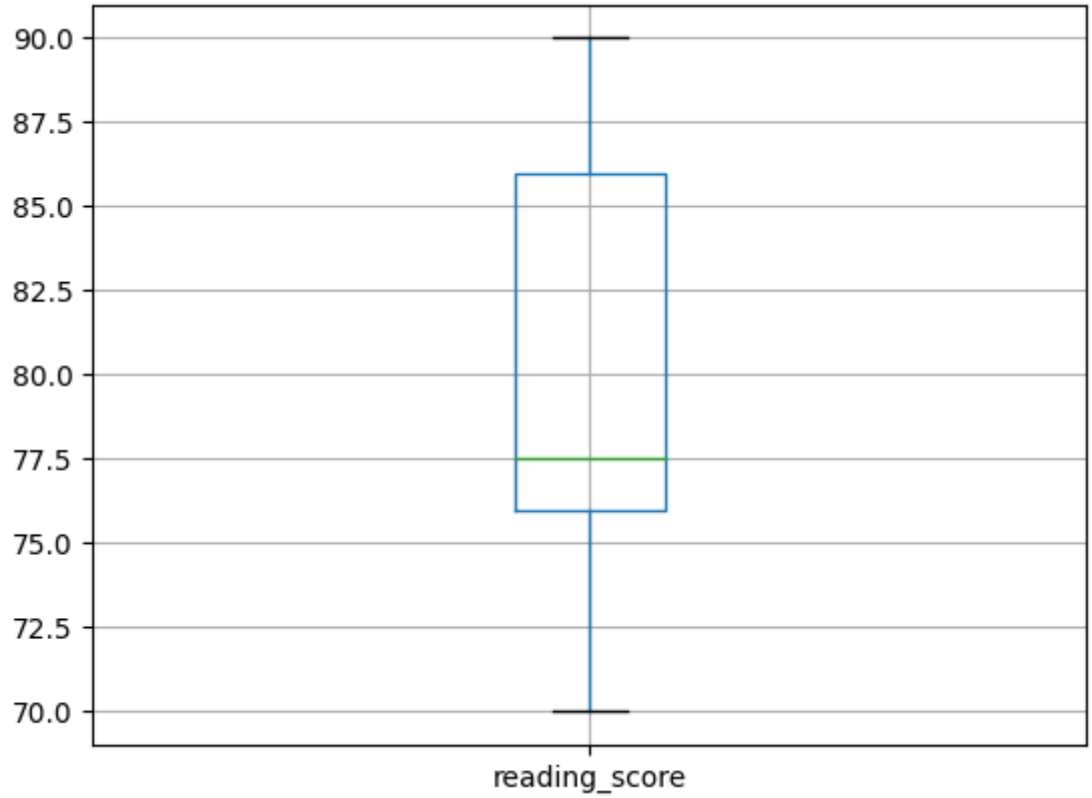
	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78	70.0	8.01457	20.0	2019
1	m	12	72.0	8.01457	81.0	2020
2	f	67	77.5	8.01457	82.0	2021
3	f	30	76.0	8.01457	83.0	2020
4	m	80	89.0	8.01457	84.0	2020
5	f	67	90.0	8.01457	90.0	2019
6	m	1112	NaN	8.01457	300.0	2020
7	f	72	86.0	8.01457	98.0	2021
8	m	73	79.0	8.01457	NaN	202
9	m	79	77.5	8.01457	90.0	2021

In [46]:

```
col=['reading_score']
df.boxplot(col)
```

Out[46]:

```
<Axes: >
```



In [47]:

```
sorted_rscore=sorted(df['maths_score'])
sorted_rscore
```

Out[47]:

```
[12, 30, 67, 67, 72, 73, 78, 79, 80, 1112]
```

In [48]:

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

```
67.0 78.75
```

In [49]:

```
IQR=q3-q1
IQR
```

Out[49]:

```
11.75
```

In [50]:

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

```
49.375 96.375
```

In [51]:

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

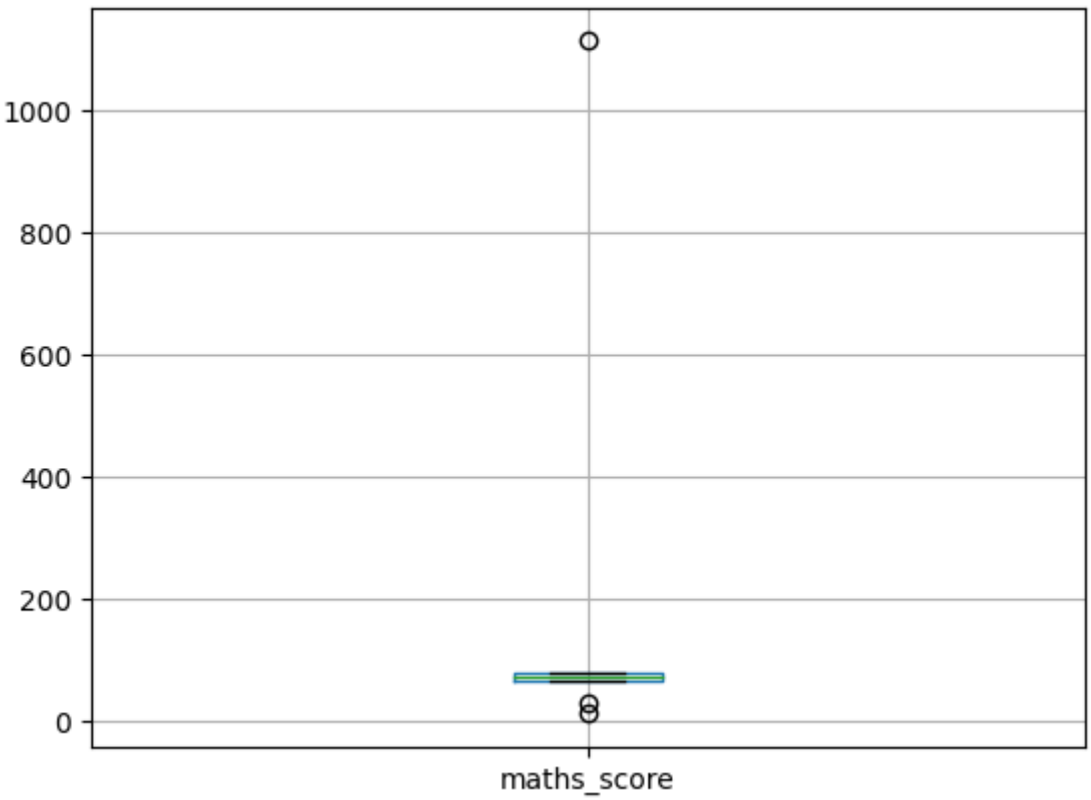
```
[12]
[12, 30]
[12, 30, 1112]
```

In [52]:

```
col=['maths_score']
df.boxplot(col)
```

Out[52]:

```
<Axes: >
```



In [53]:

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

Out[53]:

72.5

In [54]:

```
refined_df=df
refined_df['maths_score']=np.where(refined_df['maths_score']>upr_bound,median,refin
df
```

Out[54]:

	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78.0	70.0	8.01457	20.0	2019
1	m	12.0	72.0	8.01457	81.0	2020
2	f	67.0	77.5	8.01457	82.0	2021
3	f	30.0	76.0	8.01457	83.0	2020
4	m	80.0	89.0	8.01457	84.0	2020
5	f	67.0	90.0	8.01457	90.0	2019
6	m	72.5	NaN	8.01457	300.0	2020
7	f	72.0	86.0	8.01457	98.0	2021
8	m	73.0	79.0	8.01457	NaN	202
9	m	79.0	77.5	8.01457	90.0	2021

In [55]:

```
refined_df=df
refined_df['maths_score']=np.where(refined_df['maths_score']<iwr_bound,median,refin
df
```

Out[55]:

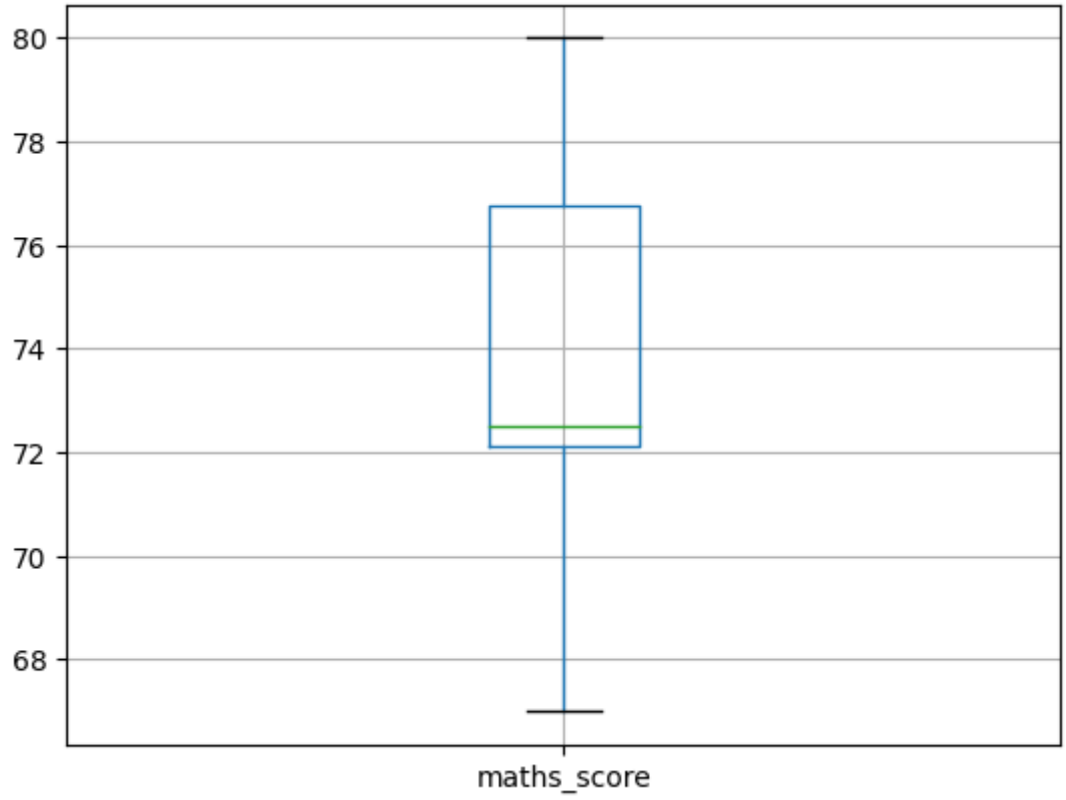
	gender	maths_score	reading_score	writing_score	placement_score	college_joining_year
0	f	78.0	70.0	8.01457	20.0	2019
1	m	72.5	72.0	8.01457	81.0	2020
2	f	67.0	77.5	8.01457	82.0	2021
3	f	72.5	76.0	8.01457	83.0	2020
4	m	80.0	89.0	8.01457	84.0	2020
5	f	67.0	90.0	8.01457	90.0	2019
6	m	72.5	NaN	8.01457	300.0	2020
7	f	72.0	86.0	8.01457	98.0	2021
8	m	73.0	79.0	8.01457	NaN	202
9	m	79.0	77.5	8.01457	90.0	2021

In [56]:

```
col=['maths_score']
df.boxplot(col)
```

Out[56]:

<Axes: >



In [62]:

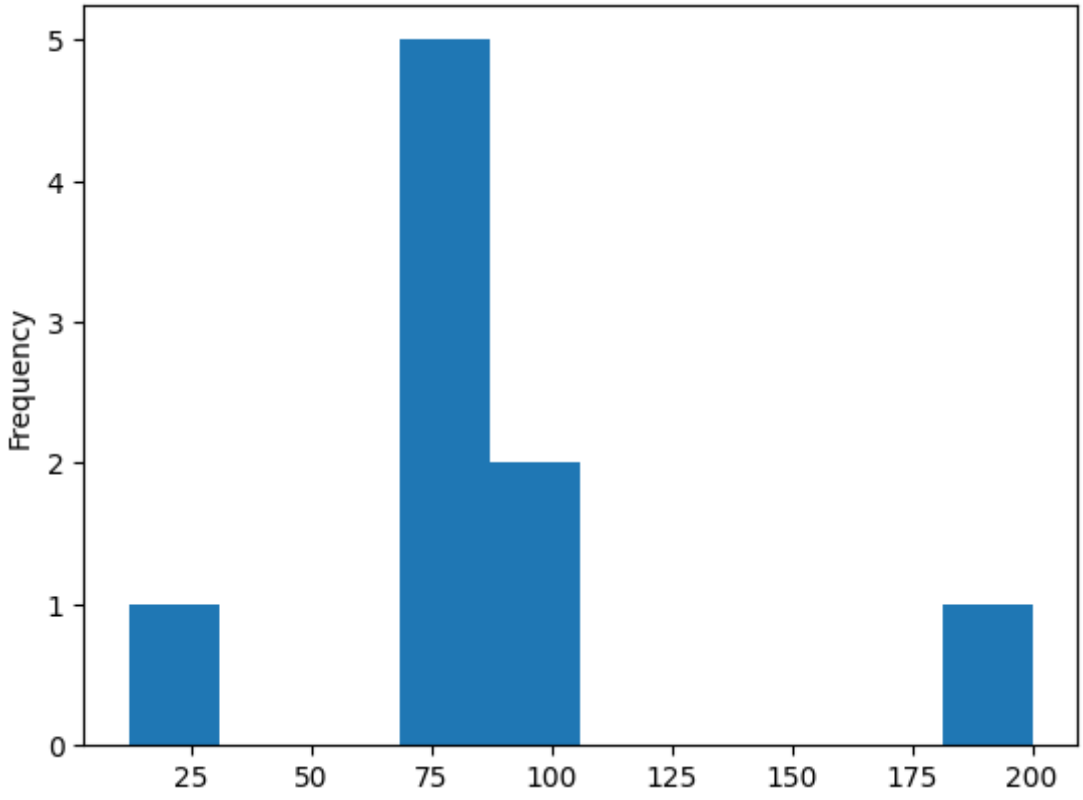
```
import matplotlib.pyplot as plt
```

In [63]:

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

Out[63]:

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

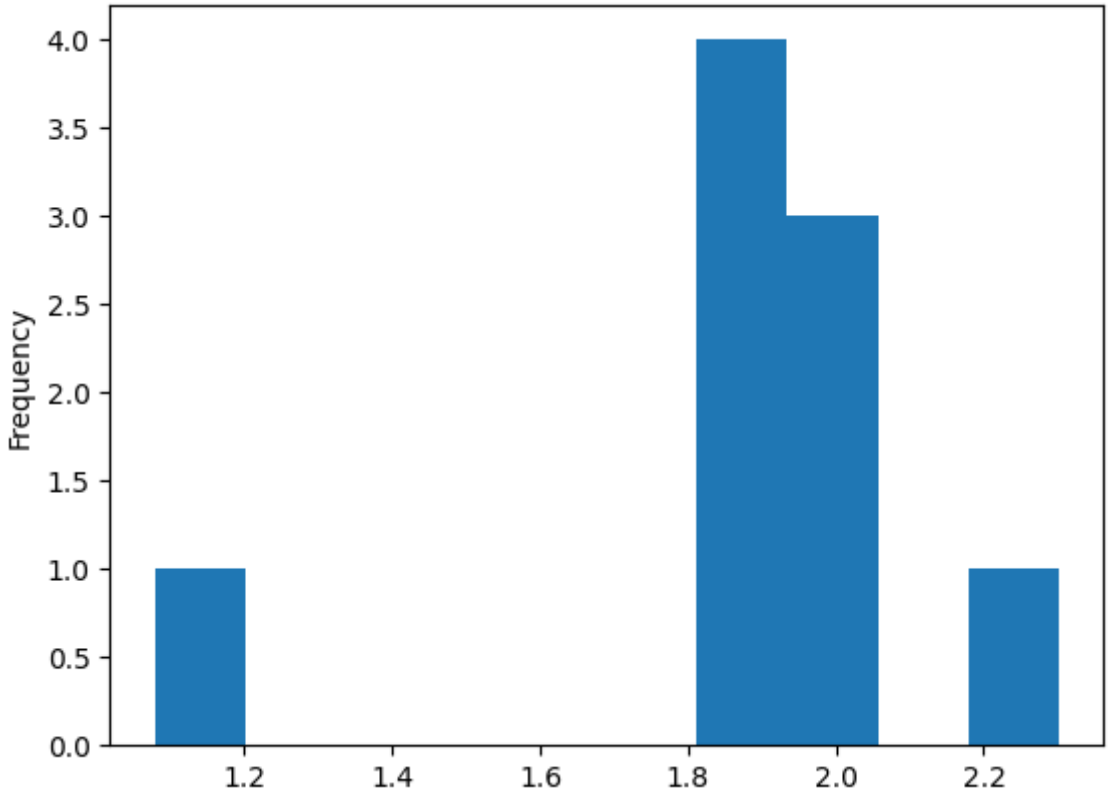


In [66]:

```
df['log_reading']=np.log10(df['reading_score'])
df['log_reading'].plot(kind='hist')
```

Out[66]:

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



In [67]:

In [ ]: