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

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

Out[6]:

:	9	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]:			das <b>as</b> pd py <b>as</b> 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
			-	-	Т	Т	т	False
	8	True	True	True	True	True	True	raise
n [10]:	se			True  f["math sc		True	True	T atse
	se	ries=pd.				placement score	placement offer count	Region
	se	ries=pd. [series]	notnull(d math	f["math sc	ore"]) writing	placement	placement offer	
	se df	ries=pd. [series] gender	math score	f["math sc reading score	ore"]) writing score	placement score	placement offer count	Region
	se df	ries=pd.[series] gender female	math score	reading score	ore"]) writing score 74.0	placement score 78.0	placement offer count	<b>Region</b> Pune
	se df 0	ries=pd.[series]  gender  female  female	math score 72 69	reading score 72 90	ore"]) writing score 74.0 88.0	placement score 78.0 NaN	placement offer count  1 2	Region Pune na
	se df 0 1 2	ries=pd.  [series]  gender  female  female  female	math score  72 69 90	reading score  72  90  95	ore"]) writing score 74.0 88.0 93.0	placement score 78.0 NaN 74.0	placement offer count  1 2 2	Region Pune na Nashik
	0 1 2	ries=pd.  [series]  gender  female female female male	math score 72 69 90 47	reading score  72  90  95  57	ore"])  writing score  74.0  88.0  93.0  NaN	placement score 78.0 NaN 74.0 78.0	placement offer count  1 2 2 1	Region Pune na Nashik Na
	se df  0 1 2 3 4	ries=pd.  [series]  gender  female female female male male	math score 72 69 90 47 na	reading score  72  90  95  57  78	ore"]) writing score 74.0 88.0 93.0 NaN 75.0	placement score 78.0 NaN 74.0 78.0 81.0	placement offer count  1 2 2 1 3	Region  Pune  na  Nashik  Na  Pune
	se df  0 1 2 3 4 5	ries=pd.[series]  gender  female female male male female	math score  72 69 90 47 na 71	reading score  72  90  95  57  78  Na	ore"])  writing score  74.0  88.0  93.0  NaN  75.0  78.0	placement score 78.0 NaN 74.0 78.0 81.0 70.0	placement offer count  1 2 2 1 3 4	Region  Pune  na  Nashik  Na  Pune  na
n [10]: ut[10]: n [11]:	se df  0 1 2 3 4 5 6 8	ries=pd.  [series]  gender  female female male male female female	math score  72  69  90  47  na  71  12  5	reading score  72  90  95  57  78  Na  44	ore"])  writing score  74.0  88.0  93.0  NaN  75.0  78.0  52.0  89.0	placement score 78.0 NaN 74.0 78.0 81.0 70.0 12.0 55.0	placement offer count  1 2 2 1 3 4 2	Region  Pune  na  Nashik  Na  Pune  na  Nashik
ut[10]:	sedf 0 1 2 3 4 5 6 8	ries=pd.[series]  gender  female female male male female female om sklea	math score 72 69 90 47 na 71 12 5	reading score  72  90  95  57  78  Na  44  77	ore"])  writing score  74.0  88.0  93.0  NaN  75.0  78.0  52.0  89.0  port Label	placement score 78.0 NaN 74.0 78.0 81.0 70.0 12.0 55.0	placement offer count  1 2 2 1 3 4 2	Region  Pune  na  Nashik  Na  Pune  na  Nashik
ut[10]:	sedf 0 1 2 3 4 5 6 8	ries=pd.[series]  gender  female female male male female female om sklea	math score 72 69 90 47 na 71 12 5 rn.prepro	reading score  72  90  95  57  78  Na  44  77	ore"])  writing score  74.0  88.0  93.0  NaN  75.0  78.0  52.0  89.0  port Label	placement score 78.0 NaN 74.0 78.0 81.0 70.0 12.0 55.0	placement offer count  1 2 2 1 3 4 2	Region  Pune  na  Nashik  Na  Pune  na  Nashik

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['m	ath score']	.mean()				
In [22]:	df	['math	score'].fil	lna(value:	=m_v, inpl	ace <b>=True</b> )		
In [22]: In [23]:	df df		score'].fil	lna(value	=m_v, inpl	ace= <b>True</b> )		
			score'].fil math score	lna(value: reading score	m_v, inpl writing score	placement	placement offer count	Region
In [23]:				reading	writing	placement	•	<b>Region</b> Pune
In [23]:	df	gender	math score	reading score	writing score	placement score	count	
In [23]:	df 0	gender female female	math score 72.000000	reading score 72.0	writing score	placement score 78.0	count	Pune
In [23]:	o 1 2	gender female female	math score 72.000000 69.000000	reading score 72.0 90.0	writing score 74.0 88.0	placement score 78.0 NaN	2 2	Pune NaN
In [23]:	o 1 2	gender female female	math score 72.000000 69.000000 90.000000	reading score 72.0 90.0 95.0	writing score 74.0 88.0 93.0	placement score 78.0 NaN 74.0	2 2	Pune NaN Nashik
In [23]:	o 1 2 3	gender female female male	math score 72.000000 69.000000 90.000000 47.000000	reading score 72.0 90.0 95.0 57.0	writing score 74.0 88.0 93.0 NaN	placement score 78.0 NaN 74.0 78.0	2 2 1	Pune NaN Nashik NaN
In [23]:	0 1 2 3	gender female female male male	math score 72.000000 69.000000 90.000000 47.000000 52.285714	reading score 72.0 90.0 95.0 57.0 78.0	writing score 74.0 88.0 93.0 NaN 75.0	placement score 78.0 NaN 74.0 78.0 81.0	count  1  2  2  1  3	Pune NaN Nashik NaN Pune

5.000000

In [24]: ndf.replace(to\_replace=np.nan,value=-99)

8 female

77.0

89.0

55.0

0

NaN

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
[n [25]:	nd	f.dropn	a()					
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]:	nd	f.dropn	a(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
[n [27]:	nd	f.dropn	a(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]:

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
2	20	69	65	84	35	1	2018
,	21	100	77	70	91	3	2018
:	22	72	60	78	94	3	2019
2	23	74	65	71	84	2	2019
2	24	75	77	83	77	2	2020
2	25	180	67	63	75	3	2021
2	26	72	120	70	84	2	2021
:	27	71	79	88	85	3	2021
2	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:>
                  φ
                                                    φ
        175
        150
        125
         100
         75
          50
              math score
                         reading score
                                    writing score placement score
 In [9]: print(np.where(df['math score']>90))
         (array([10, 21, 25, 28]),)
In [10]: print(np.where(df['reading score']<25))</pre>
         (array([ 2, 14]),)
In [11]: print(np.where(df['writing score']<30))</pre>
         (array([10, 17]),)
In [12]: fig,ax=plt.subplots(figsize=(18,10))
          ax.scatter(df['placement score'],df['placement offer count'])
          plt.show()
        3.00
        2.75
        2.25
        2.00
        1.50
        1.00
In [13]: ax.set_ylabel('(Proportion non-retail busines acres)/(town)')
Out[13]: Text(4.44444444444452, 0.5, '(Proportion non-retail busines acres)/(town)')
```

```
In [14]: ax.set_ylabel('(Full-value property-tax rate)/($10,000)')
Out[14]: Text(4.44444444444452, 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)))</pre>
        (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
              0.254010
        5
              0.449923
              0.293193
        6
        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
              0.410740
        23
        24
              0.371558
        25
              3.742601
        26
              0.489105
        27
              0.528288
              1.391653
        28
        Name: math score, dtype: float64
In [20]: threshold=0.18
In [21]: sample_outliers=np.where(z<threshold)</pre>
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]: |wr_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]:	<pre>df_st ninet b = n ninet</pre>	ud=df ieth_perce p.where(df	entile = np. _stud['math entile, df_s	percentile(	ds/demol(1).cs df_stud['math netieth_percen core'])	score'], 90)	
	New ar 8.	ray: [ 80.	71. 79.	61. 78.	73. 77. 74.	76. 75. 104. 7	79. 80. 7

```
5 of 12 26/03/25, 15:34
```

75. 78. 86. 80. 75. 82. 69. 100. 72. 74. 75. 104. 72. 71.

104.]

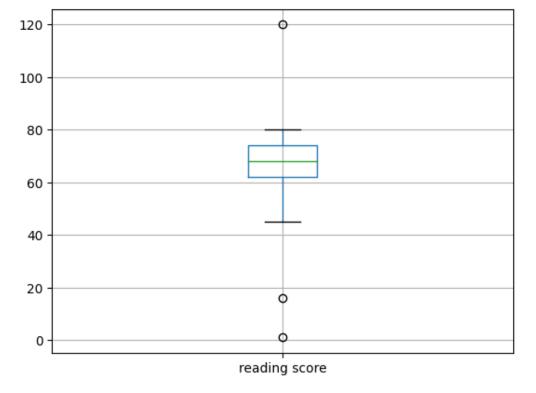
In [33]: df\_stud.insert(1,"m score",b,True)
 df\_stud

| Out[33]: | math<br>score | m<br>score | reading<br>score | writing<br>score | placement<br>score | placement offer<br>count | club join<br>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
```

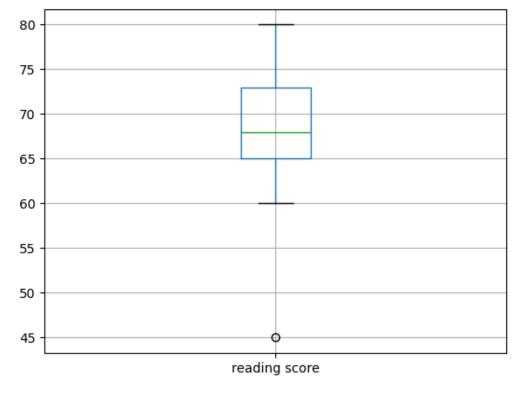
| Out[37]: | math<br>score | m<br>score | reading<br>score | writing<br>score | placement<br>score | placement offer<br>count | club join<br>year |
|----------|---------------|------------|------------------|------------------|--------------------|--------------------------|-------------------|
| 0        | 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</pre>

| Out[38]: | math<br>score | m<br>score | reading<br>score | writing<br>score | placement<br>score | placement offer<br>count | club join<br>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:>

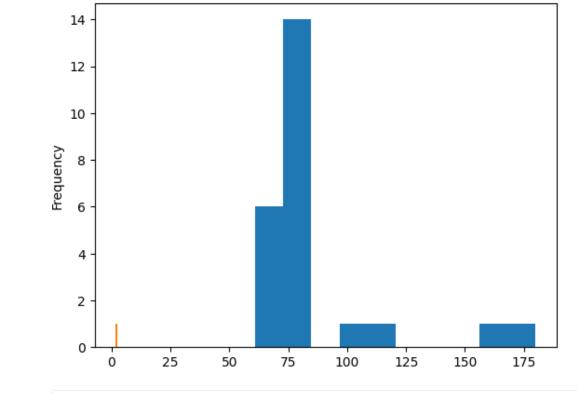


```
In [40]: import pandas as pd
import numpy as np
df=pd.read_csv("/home/jaihind/Downloads/demo1(1).csv")
In [41]: df
```

| Out[41]: | math<br>score | reading<br>score | writing<br>score | placement<br>score | placement offer<br>count | club join<br>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 [ ]: