

ASSIGNMENT NO:2

Aim:

1. Creation of Dataset using Microsoft Excel.
3. Identification **and** Handling of Outliers
4. Data Transformation **for** the purpose of :
 - a. To change the scale **for** better understanding
 - b. To decrease the skewness **and** convert distribution into normal distribution

```
import pandas as pd  
import seaborn as sns  
import numpy as py  
df= pd.read_csv("C:/Users/Welcome/Music/Book1.csv")  
df
```

join year \	math score	reading score	writing score	placement score	club
0	60.0	63.0	76.0	95.0	
2021	75.0	70.0	64.0	85.0	
1	74.0	50.0	55.0	91.0	
2020	68.0	76.0	78.0	97.0	
3	NaN	67.0	71.0	93.0	
2020	70.0	64.0	80.0	98.0	
4	61.0	78.0	92.0	94.0	
2020	61.0	74.0	78.0	NaN	
5	64.0	76.0	79.0	76.0	
2018	65.0	95.0	75.0	90.0	
6	66.0	76.0	NaN	100.0	
2021	84.0	67.0	71.0	92.0	
9	69.0	NaN	70.0	86.0	
2020	74.0	65.0	65.0	80.0	
10	74.0	63.0	72.0	96.0	
2020					
12					
2021					
13					
2021					
14					

2018				
15	76.0	64.0	80.0	96.0
2020	60.0	64.0	54.0	91.0
16	77.0	70.0	72.0	99.0
2021	67.0	95.0	NaN	87.0
17	71.0	53.0	78.0	75.0
2020	58.0	65.0	56.0	NaN
18	68.0	63.0	62.0	94.0
2018	77.0	63.0	68.0	97.0
19	80.0	NaN	86.0	85.0
2018	84.0	63.0	67.0	83.0
20	68.0	67.0	73.0	88.0
2019	76.0	64.0	68.0	96.0
21	92.0	96.0	61.0	83.0
2021	60.0	68.0	59.0	93.0
22				
2021				
23				
2018				
24				
2018				
25				
2019				
26				
2021				
27				
2018				
28				
2020				

	placement offer count gender
0	3 female
1	3 male
2	3 male
3	3 female
4	3 male
5	3 female
6	3 male
7	2 male
8	2 male
9	3 female
10	1 male
11	3 male
12	3 female
13	2 male
14	3 male
15	3 male
16	3 female
17	3 male

```

18 19 20 21 22          3 female
23 24 25 26          2 female
27 28          3 female
df.isnull()
3          male
3 female
3 female
1          male
3 female
3 female
2          male
3          male

```

join year \	math score	reading score	writing score	placement score	club
0	False	False	False	False	False
False	False	False	False	False	False
1	False	False	False	False	False
False	False	False	False	False	False
2	False	False	False	False	False
False	True	False	False	False	False
3	False	False	False	False	False
False	False	False	False	False	False
4	False	False	False	False	False
False	False	False	False	False	True
5	False	False	False	False	False
False	False	False	False	False	False
6	False	False	False	False	False
False	False	False	False	True	False
7	False	False	False	False	False
False	False	False	True	False	False
8	False	False	False	False	False
False	False	False	False	False	False
9	False	False	False	False	False
False	False	False	False	False	False
10	False	False	False	False	False
False	False	False	False	False	False
11	False	False	False	False	False
False	False	False	False	False	False
12					
False					
13					
False					
14					
False					
15					
False					
16					
False					

17	False	False	False	False
False	False	False	True	False
18	False	False	False	False
False	False	False	False	True
19	False	False	False	False
False	False	False	False	False
20	False	True	False	False
False	False	False	False	False
21	False	False	False	False
False	False	False	False	False
22	False	False	False	False
False	False	False	False	False
23				
False				
24				
False				
25				
False				
26				
False				
27				
False				
28				
False				

	placement	offer	count	gender
0		False		False
1		False		False
2		False		False
3		False		False
4		False		False
5		False		False
6		False		False
7		False		False
8		False		False
9		False		False
10		False		False
11		False		False
12		False		False
13		False		False
14		False		False
15		False		False
16		False		False
17		False		False
18		False		False
19		False		False
20		False		False
21		False		False
22		False		False

```

23           False  False
24           False  False
25           False  False
26           False  False
27           False  False
28           False  False

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

      math score reading score      writing score placementt score club
join year \
4          NaN            67.0             71.0            93.0
2020

      placement offer count gender
4                  3     male
df.notnull()

      math score reading score      writing score placementt score club
join year \
0           True        True       True        True        True
True         True        True       True        True        True
1           True        True       True        True        True
True         True        True       True        True        True
2           True        True       True        True        True
True         True        True       True        True        True
3           False       True       True        True        True
True         True        True       True        True        True
4           True        True       True        True        True
True         True        True       True        True        False
5           True        True       True        True        True
True         True        True       True        True        True
6           True        True       True        True        True
True         True        True       True        True        True
7           True        True       True        False        True
True         True        True       True        True        True
8           True        True       True        True        True
True         True        False      True        True        True
9           True        True       True        True        True
True         True        True       True        True        True
10          True        True       True        True        True
True         True        True       True        True        True
11          True        True       True        True        True
True         True        True       True        True        True
12          True        True       True        True        True
True         True        True       True        True        True
13          True        True       True        True        True
True         True        True       True        True        True

```

14	True	True	True	True
True	True	True	True	True
15	True	True	True	True
True	True	True	True	True
16	True	True	False	True
True	True	True	True	True
17	True	True	True	False
True	True	True	True	True
18	True	True	True	True
True	True	False	True	True
19	True	True	True	True
True	True	True	True	True
20	True	True	True	True
True	True	True	True	True
21	True	True	True	True
True				
22				
True				
23				
True				
24				
True				
25				
True				
26				
True				
27				
True				
28				
True				

placement offer count gender

0	True	True
1	True	True
2	True	True
3	True	True
4	True	True
5	True	True
6	True	True
7	True	True
8	True	True
9	True	True
10	True	True
11	True	True
12	True	True
13	True	True
14	True	True
15	True	True
16	True	True

```

17          True   True
18          True   True
19          True   True
20          True   True
21          True   True
22          True   True
23          True   True
24          True   True
25          True   True
26          True   True
27          True   True
28          True   True

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

      math score  reading score    writing score  placementmt score   club
join year \
0            60.0           63.0         76.0             95.0
2021
1            75.0           70.0         64.0             85.0
2020
2            74.0           50.0         55.0             91.0
2020
3            68.0           76.0         78.0             97.0
2020
5            70.0           64.0         80.0             98.0
2018
6            61.0           78.0         92.0             94.0
2018
7            61.0           74.0         78.0             NaN
2021
8            64.0           76.0         79.0             76.0
2021
9            65.0           95.0         75.0             90.0
2019
8            66.0           76.0         NaN             100.0
2020
9            84.0           67.0         71.0             92.0
2020
10           69.0           NaN          70.0             86.0
2019
11           74.0           65.0         65.0             80.0
2020
12           74.0           63.0         72.0             96.0
2021
13           76.0           64.0         80.0             96.0
2021
14           60.0           64.0         54.0             91.0
2021
15
2020
16
2021

```

17	77.0	70.0	72.0	99.0
2020	67.0	95.0	NaN	87.0
18	71.0	53.0	78.0	75.0
2018	58.0	65.0	56.0	NaN
19	68.0	63.0	62.0	94.0
2018	77.0	63.0	68.0	97.0
20	80.0	NaN	86.0	85.0
2019	84.0	63.0	67.0	83.0
21	68.0	67.0	73.0	88.0
2021	76.0	64.0	68.0	96.0
22	92.0	96.0	61.0	83.0
2021	60.0	68.0	59.0	93.0
23				
2018				
24				
2018				
25				
2019				
26				
2021				
27				
2018				
28				
2020				

	placement	offer count	gender
0		3	female
1		3	male
2		3	male
3		3	female
5		3	female
6		3	male
7		2	male
8		2	male
9		3	female
10		1	male
11		3	male
12		3	female
13		2	male
14		3	male
15		3	male
16		3	female
17		3	male
18		3	female
19		2	female
20		3	female
21		3	male
22		3	female
23		3	female

```

24           1     male
25           3   female
26           3   female
27           2     male
28           3     male

from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['gender'] = le.fit_transform(df['gender'])
newdf = df
df

      math score reading score      writing score placementt score  club
join year \
0            60.0        63.0          76.0            95.0
2021         75.0        70.0          64.0            85.0
1            74.0        50.0          55.0            91.0
2020         74.0        76.0          78.0            97.0
2            68.0        67.0          71.0            93.0
2020         NaN         67.0          80.0            98.0
4            70.0        64.0          92.0            94.0
2020         61.0        78.0          78.0            NaN
5            61.0        74.0          79.0            76.0
2018         64.0        76.0          75.0            90.0
6            65.0        95.0          71.0            100.0
2021         66.0        76.0          NaN            92.0
8            84.0        67.0          70.0            86.0
2020         69.0        NaN           65.0            80.0
10           74.0        63.0          72.0            96.0
11           74.0        64.0          80.0            96.0
2021         60.0        64.0          54.0            91.0
13           77.0        70.0          72.0            99.0
14
2018
15
2020
16
2021
17
2020

```

18	67.0	95.0	NaN	87.0
2018	71.0	53.0	78.0	75.0
19	58.0	65.0	56.0	NaN
2018	68.0	63.0	62.0	94.0
20	77.0	63.0	68.0	97.0
2019	80.0	NaN	86.0	85.0
21	84.0	63.0	67.0	83.0
2021	68.0	67.0	73.0	88.0
22	76.0	64.0	68.0	96.0
2021	92.0	96.0	61.0	83.0
23	60.0	68.0	59.0	93.0

2018

24

2018

25

2019

26

2021

27

2018

28

2020

	placement	offer	count	gender
0		3	0	
1		3	1	
2		3	1	
3		3	0	
4		3	1	
5		3	0	
6		3	1	
7		2	1	
8		2	1	
9		3	0	
10		1	1	
11		3	1	
12		3	0	
13		2	1	
14		3	1	
15		3	1	
16		3	0	
17		3	1	
18		3	0	
19		2	0	
20		3	0	
21		3	1	
22		3	0	
23		3	0	
24		1	1	

25	3	0
26	3	0
27	2	11
28	3	

```
missing_values = ["Na","na"]
df = pd.read_csv("C:/Users/Welcome/Music/Book1.csv", na_values =
missing_values)
df
```

join year \	math score	reading score	writing score	placement score	club
0	60.0	63.0	76.0	95.0	
2021	75.0	70.0	64.0	85.0	
1	74.0	50.0	55.0	91.0	
2020	68.0	76.0	78.0	97.0	
3	Nan	67.0	71.0	93.0	
2020	70.0	64.0	80.0	98.0	
4	61.0	78.0	92.0	94.0	
2020	61.0	74.0	78.0	Nan	
5	64.0	76.0	79.0	76.0	
2018	65.0	95.0	75.0	90.0	
6	66.0	76.0	Nan	100.0	
2021	84.0	67.0	71.0	92.0	
8	69.0	Nan	70.0	86.0	
2019	74.0	65.0	65.0	80.0	
9	74.0	63.0	72.0	96.0	
2020	76.0	64.0	80.0	96.0	
10	60.0	64.0	54.0	91.0	
2021	77.0	70.0	72.0	99.0	
14	67.0	95.0	Nan	87.0	
2018					
15					
2020					
16					
2021					
17					
2020					
18					
2018					

19	71.0	53.0	78.0	75.0
2018	58.0	65.0	56.0	NaN
20	68.0	63.0	62.0	94.0
2019	77.0	63.0	68.0	97.0
21	80.0	NaN	86.0	85.0

2021 84.0 63.0 67.0 83.0

22 68.0 67.0 73.0 88.0

2021 76.0 64.0 68.0 96.0

23 92.0 96.0 61.0 83.0

2018 60.0 68.0 59.0 93.0

24

2018

25

2019

26

2021

27

2018

28

2020

	placement	offer count	gender
0		3	female
1		3	male
2		3	male
3		3	female
4		3	male
5		3	female
6		3	male
7		2	male
8		2	male
9		3	female
10		1	male
11		3	male
12		3	female
13		2	male
14		3	male
15		3	male
16		3	female
17		3	male
18		3	female
19		2	female
20		3	female
21		3	male
22		3	female
23		3	female
24		1	male
25		3	female
26		3	female

```
27 28
```

```
2 3 male  
male
```

```
ndf = df
```

```
ndf.fillna(0)
```

join year\	math score	reading score	writing score	placement score	club
0	60.0	63.0	76.0	95.0	
2021	75.0	70.0	64.0	85.0	
1	74.0	50.0	55.0	91.0	
2020	68.0	76.0	78.0	97.0	
2	0.0	67.0	71.0	93.0	
2020	70.0	64.0	80.0	98.0	
3	61.0	78.0	92.0	94.0	
2020	61.0	74.0	78.0	0.0	
4	64.0	76.0	79.0	76.0	
2021	65.0	95.0	75.0	90.0	
5	66.0	76.0	0.0	100.0	
2019	84.0	67.0	71.0	92.0	
6	69.0	0.0	70.0	86.0	
2020	74.0	65.0	65.0	80.0	
7	74.0	63.0	72.0	96.0	
2020	76.0	64.0	80.0	96.0	
8	60.0	64.0	54.0	91.0	
2021	77.0	70.0	72.0	99.0	
9	67.0	95.0	0.0	87.0	
2018	71.0	53.0	78.0	75.0	
10	58.0	65.0	56.0	0.0	
2020					
11					
2021					
12					
2020					
13					
2021					
14					
2018					
15					
2020					
16					
2021					
17					
2020					
18					
2018					
19					
2018					
20					
2019					

21	68.0	63.0	62.0	94.0
2021	77.0	63.0	68.0	97.0
22	80.0	0.0	86.0	85.0
2021	84.0	63.0	67.0	83.0
23	68.0	67.0	73.0	88.0

2018	76.0	64.0	68.0	96.0
24	92.0	96.0	61.0	83.0
2018	60.0	68.0	59.0	93.0

25

2019

26

2021

27

2018

28

2020

	placement	offer	count	gender
0			3	female
1			3	male
2			3	male
3			3	female
4			3	male
5			3	female
6			3	male
7			2	male
8			2	male
9			3	female
10			1	male
11			3	male
12			3	female
13			2	male
14			3	male
15			3	male
16			3	female
17			3	male
18			3	female
19			2	female
20			3	female
21			3	male
22			3	female
23			3	female
24			1	male
25			3	female
26			3	female
27			2	male
28			3	male

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

join year \	math score	reading score	writing score	placement score	club
0	60.000000	63.0	76.0	95.0	
2021	75.000000	70.0	64.0	85.0	
1	74.000000	50.0	55.0	91.0	
2020	68.000000	76.0	78.0	97.0	
2	70.678571	67.0	71.0	93.0	
2020	70.000000	64.0	80.0	98.0	
3	61.000000	78.0	92.0	94.0	
2018	61.000000	74.0	78.0	Nan	
6	64.000000	76.0	79.0	76.0	
2021	65.000000	95.0	75.0	90.0	
7	66.000000	76.0	Nan	100.0	
2019	84.000000	67.0	71.0	92.0	
9	69.000000	Nan	70.0	86.0	
2020	74.000000	65.0	65.0	80.0	
10	74.000000	63.0	72.0	96.0	
2019	76.000000	64.0	80.0	96.0	
11	60.000000	64.0	54.0	91.0	
2021	77.000000	70.0	72.0	99.0	
13	67.000000	95.0	Nan	87.0	
2018	71.000000	53.0	78.0	75.0	
15	58.000000	65.0	56.0	Nan	
2020	68.000000	63.0	62.0	94.0	
17					
2020					
18					
2018					
19					
2018					
20					
2019					
21					
2021					

22	77.000000	63.0	68.0	97.0
2021	80.000000	NaN	86.0	85.0
23	84.000000	63.0	67.0	83.0
2018	68.000000	67.0	73.0	88.0
24	76.000000	64.0	68.0	96.0
2018	92.000000	96.0	61.0	83.0
25	60.000000	68.0	59.0	93.0
2019				
26				
2021				
27				
2018				
28				
2020				

	placement offer count gender
0	3 female
1	3 male
2	3 male
3	3 female
4	3 male
5	3 female
6	3 male
7	2 male
8	2 male
9	3 female
10	1 male
11	3 male
12	3 female
13	2 male
14	3 male
15	3 male
16	3 female
17	3 male
18	3 female
19	2 female
20	3 female
21	male
22	3 female
23	3 female
24	male
25	1 female
26	3 female
27	male
28	2 male
	3

ndf.replace(to_replace = py.nan, value = -99)

math score	reading score	writing score	placement score	club
join year \				

0	60.000000	63.0	76.0	95.0
2021	75.000000	70.0	64.0	85.0
1	74.000000	50.0	55.0	91.0
2020	68.000000	76.0	78.0	97.0
2	70.678571	67.0	71.0	93.0
2020	70.000000	64.0	80.0	98.0
3	61.000000	78.0	92.0	94.0
2020	61.000000	74.0	78.0	-99.0
4	64.000000	76.0	79.0	76.0
2020	65.000000	95.0	75.0	90.0
5	66.000000	76.0	-99.0	100.0
2018	84.000000	67.0	71.0	92.0
6	69.000000	-99.0	70.0	86.0
2021	74.000000	65.0	65.0	80.0
7	74.000000	63.0	72.0	96.0
2021	76.000000	64.0	80.0	96.0
8	60.000000	64.0	54.0	91.0
2019	77.000000	70.0	72.0	99.0
9	67.000000	95.0	-99.0	87.0
2020	71.000000	53.0	78.0	75.0
10	58.000000	65.0	56.0	-99.0
2019	68.000000	63.0	62.0	94.0
11	77.000000	63.0	68.0	97.0
2020	80.000000	-99.0	86.0	85.0
12	84.000000	63.0	67.0	83.0
2021				
13				
2021				
14				
2018				
15				
2020				
16				
2021				
17				
2020				
18				
2018				
19				
2018				
20				
2019				
21				
2021				
22				
2021				
23				
2018				
24				

2018				
25	68.000000	67.0	73.0	88.0
2019	76.000000	64.0	68.0	96.0
26	92.000000	96.0	61.0	83.0
2021	60.000000	68.0	59.0	93.0
27				
2018				
28				
2020				
placement offer count gender				
0		3 female		
1		3 male		
2		3 male		
3		3 female		
4		3 male		
5		3 female		
6		3 male		
7		2 male		
8		2 male		
9		3 female		
10		1 male		
11		3 male		
12		3 female		
13		2 male		
14		3 male		
15		3 male		
16		3 female		
17		3 male		
18		3 female		
19		2 female		
20		3 female		
21		3 male		
22		3 female		
23		3 female		
24		1 male		
25		3 female		
26		3 female		
27		2 male		
28		3 male		
ndf.dropna()				
math score reading score writing score placement score club				
join year \				
0	60.000000	63.0	76.0	95.0
2021	75.000000	70.0	64.0	85.0
1	74.000000	50.0	55.0	91.0
2				

2020				
3	68.000000	76.0	78.0	97.0
2020	70.678571	67.0	71.0	93.0
4	70.000000	64.0	80.0	98.0
2020	61.000000	78.0	92.0	94.0
5	64.000000	76.0	79.0	76.0
2018	65.000000	95.0	75.0	90.0
6	84.000000	67.0	71.0	92.0
2021	74.000000	65.0	65.0	80.0
8	74.000000	63.0	72.0	96.0
2019	76.000000	64.0	80.0	96.0
9	60.000000	64.0	54.0	91.0
2020	77.000000	70.0	72.0	99.0
11	71.000000	53.0	78.0	75.0
2020	68.000000	63.0	62.0	94.0
13	77.000000	63.0	68.0	97.0
2021	84.000000	63.0	67.0	83.0
14	68.000000	67.0	73.0	88.0
2018	76.000000	64.0	68.0	96.0
15	92.000000	96.0	61.0	83.0
2020	60.000000	68.0	59.0	93.0

16

2021

17

2020

19

2018

21

2021

22

2021

24

2018

25

2019

26

2021

27

2018

28

2020

placement offer count gender

0		3	female
1		3	male
2		3	male
3		3	female
4		3	male
5		3	female

```

6 8 9 11 13 14 15 16 17 19 21 22 3      male
24 25 26 27 28          2      male
ndf.dropna(how = 'all')    3 female
                           3      male
                           2      male
                           3      male
                           3      male
                           3      male
                           3 female
                           3      male
                           2 female
                           3      male
                           3 female
                           1      male
                           3 female
                           3 female
                           2      male
                           3      male

```

join year\	math score	reading score	writing score	placement score	club
0	60.000000	63.0	76.0	95.0	
2021	75.000000	70.0	64.0	85.0	
1		50.0	55.0	91.0	
2020	74.000000				
2	68.000000	76.0	78.0	97.0	
2020	70.678571	67.0	71.0	93.0	
3			80.0	98.0	
2020	70.000000	64.0			
4	61.000000	78.0	92.0	94.0	
2020			78.0	NaN	
5	61.000000	74.0			
2018	64.000000	76.0	79.0	76.0	
6			75.0	90.0	
2021	65.000000	95.0			
7	66.000000	76.0	NaN	100.0	
2021	84.000000	67.0	71.0	92.0	
8			70.0	86.0	
2019	69.000000	NaN			
9	74.000000	65.0	65.0	80.0	
2020					
10					
2019					
11					
2020					
12					
2021					
13					
2021					

14	74.000000	63.0	72.0	96.0
2018	76.000000	64.0	80.0	96.0
15	60.000000	64.0	54.0	91.0
2020	77.000000	70.0	72.0	99.0
16	67.000000	95.0	NaN	87.0
2021	71.000000	53.0	78.0	75.0
17	58.000000	65.0	56.0	NaN
2020	68.000000	63.0	62.0	94.0
18	77.000000	63.0	68.0	97.0
2018	80.000000	NaN	86.0	85.0
19	84.000000	63.0	67.0	83.0
2018	68.000000	67.0	73.0	88.0
20	76.000000	64.0	68.0	96.0
2019	92.000000	96.0	61.0	83.0
21	60.000000	68.0	59.0	93.0
2021				
22				
2021				
23				
2018				
24				
2018				
25				
2019				
26				
2021				
27				
2018				
28				
2020				

	placement offer count gender
0	3 female
1	3 male
2	3 male
3	3 female
4	3 male
5	3 female
6	3 male
7	2 male
8	2 male
9	3 female
10	1 male
11	3 male
12	3 female
13	2 male
14	3 male
15	3 male
16	3 female

```

17 18 19 20 21 22 23 24      3     male
25 26 27 28                  3 female
ndf.dropna(axis = 1)
                                2 female
                                3 female
                                3 female
                                3     male
                                3 female
                                3 female
                                1     male
                                3 female
                                3 female
                                2     male
                                3     male

```

	math	score	club	join year	placement	offer count	gender
0	60.000000			2021		3	female
1	75.000000			2020		3	male
2	74.000000			2020		3	male
3	68.000000			2020		3	female
4	70.678571			2020		3	male
5	70.000000			2018		3	female
6	61.000000			2021		3	male
7	61.000000			2021		2	male
8	64.000000			2019		2	male
9	65.000000			2020		3	female
10	66.000000			2019		1	male
11	84.000000			2020		3	male
12	69.000000			2021		3	female
13	74.000000			2021		2	male
14	74.000000			2018		3	male
15	76.000000			2020		3	male
16	60.000000			2021		3	female
17	77.000000			2020		3	male
18	67.000000			2018		3	female
19	71.000000			2018		2	female
20	58.000000			2019		3	female
21	68.000000			2021		3	male
22	77.000000			2021		3	female
23	80.000000			2018		3	female
24	84.000000			2018		1	male
25	68.000000			2019		3	female
26	76.000000			2021		3	female
27	92.000000			2018		2	male
28	60.000000			2020		3	male

```

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

```

join year \	math score	reading score	writing score	placement score	club
0	60.000000	63.0	64.0	85.0	95.0
2021	75.000000	70.0	55.0	91.0	
1	74.000000	50.0	78.0	97.0	
2020	68.000000	76.0	71.0	93.0	
2	70.678571	67.0	80.0	98.0	
2020	70.000000	64.0	92.0	94.0	
3	61.000000	78.0	79.0	76.0	
2020	64.000000	76.0	75.0	90.0	
4	65.000000	95.0	71.0	92.0	
2020	84.000000	67.0	65.0	80.0	
5	74.000000	65.0	72.0	96.0	
2018	74.000000	63.0	80.0	96.0	
6	76.000000	64.0	54.0	91.0	
2021	60.000000	64.0	72.0	99.0	
8	77.000000	70.0	78.0	75.0	
2019	71.000000	53.0	62.0	94.0	
9	68.000000	63.0	68.0	97.0	
2020	77.000000	63.0	67.0	83.0	
11	84.000000	63.0	73.0	88.0	
2020	68.000000	67.0	68.0	96.0	
13	76.000000	64.0	61.0	83.0	
2021	92.000000	96.0	59.0	93.0	
14	60.000000	68.0			
2018					
15					
2020					
16					
2021					
17					
2020					
19					
2018					
21					
2021					
22					
2021					
24					
2018					
25					
2019					
26					
2021					
27					
2018					
28					
2020					
placement offer count gender					

```

0 1          3 female
2 3          3 male
4 5          3 male
6 8          3 female
9 11         3 male
13           3 female
14           3 male
15           2 male
16           3 female
17           3 male
19           2 male
21           3 male
22           3 male
24           3 female
25           3 male
26           2 female
27           3 male
28           3 female
             1 male
             3 female
             3 female
             2 male
             3 male

```

```

import matplotlib.pyplot as plt

df1= pd.read_csv("C:/Users/Welcome/Music/Book3.csv")
df1

      math score reading score      writing score placementmt score  club
join year\
0            60            63            76                  95
2021
1            75            70            64                  85
2020
2            74            50            55                  91
2020
3            68            76            78                  97
2020
4            94            67            71                  93
2020
5            70            64            80                  98
2020
6            61            78            92                  94
2018
7            61            74            78                  80
2021
8            64            76            79                  76
2021
9            65            95            75                  90
2019

```

	placement	offer count	gender	
10	66	76	67	100
2019	84	67	71	92
11	69	66	70	56
2020	74	65	65	80
12	74	63	72	96
2021	76	64	80	96
13	60	64	54	91
2021	77	70	72	99
14	67	95	64	87
2018	71	65	78	75
15	58	65	96	92
2020	68	63	62	94
16	77	63	68	97
2021	80	64	86	85
17	84	63	67	83
2020	68	67	73	88
18	76	64	68	96
2018	92	96	61	83
19	60	68	59	93

2018

20

2019

21

2021

22

2021

23

2018

24

2018

25

2019

26

2021

27

2018

28

2020

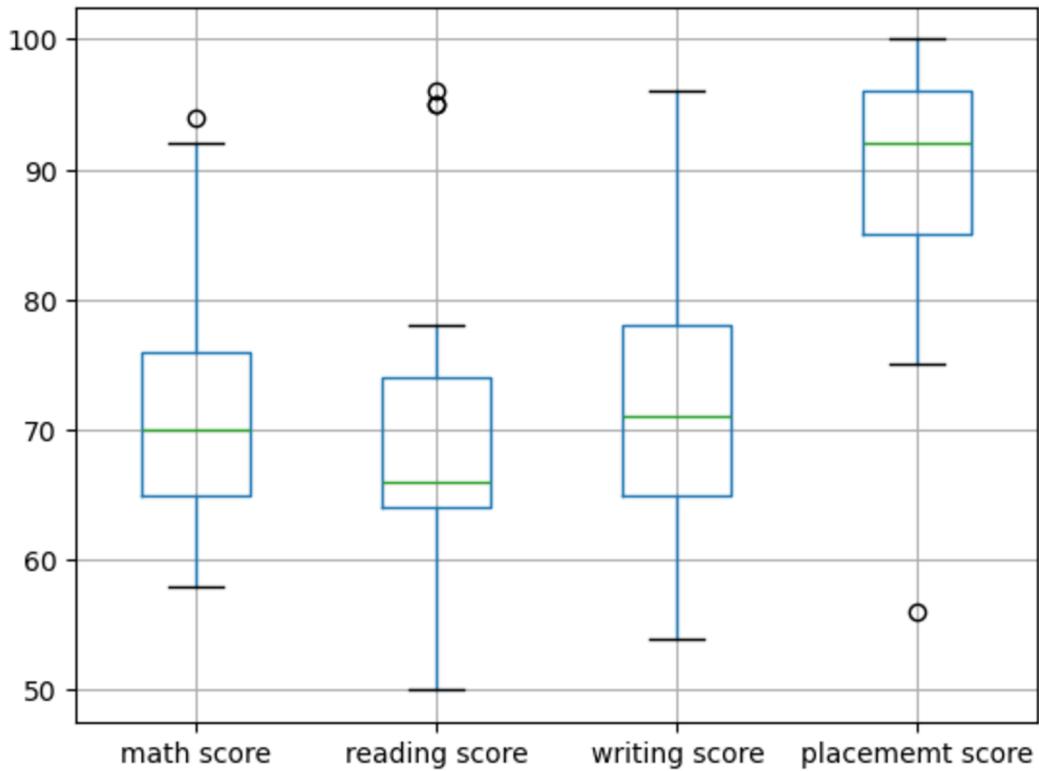
placement offer count gender

0	3	female
1	3	male
2	3	male
3	3	female
4	3	male
5	3	female
6	3	male
7	2	male

```
8 9          2     male
10           3 female
11           1     male
12           3     male
13           3 female
14           2     male
15           3     male
16           3     male
17           3 female
18           3     male
19           3 female
20           2 female
21           3 female
22           3     male
23           3 female
24           3 female
25           1     male
26           3 female
27           3 female
28           2     male
            3     male
```

```
col = ['math score', 'reading score', 'writing score', 'placement  
score']
df1.boxplot(col)
```

```
<Axes: >
```



```

print(py.where(df['math score']>90))
print(py.where(df['reading score']<25))
print(py.where(df['writing score']<30))

(array([ 4, 27], dtype=int64,))
(array([], dtype=int64,))
(array([], dtype=int64,))

```

pip install matplotlib

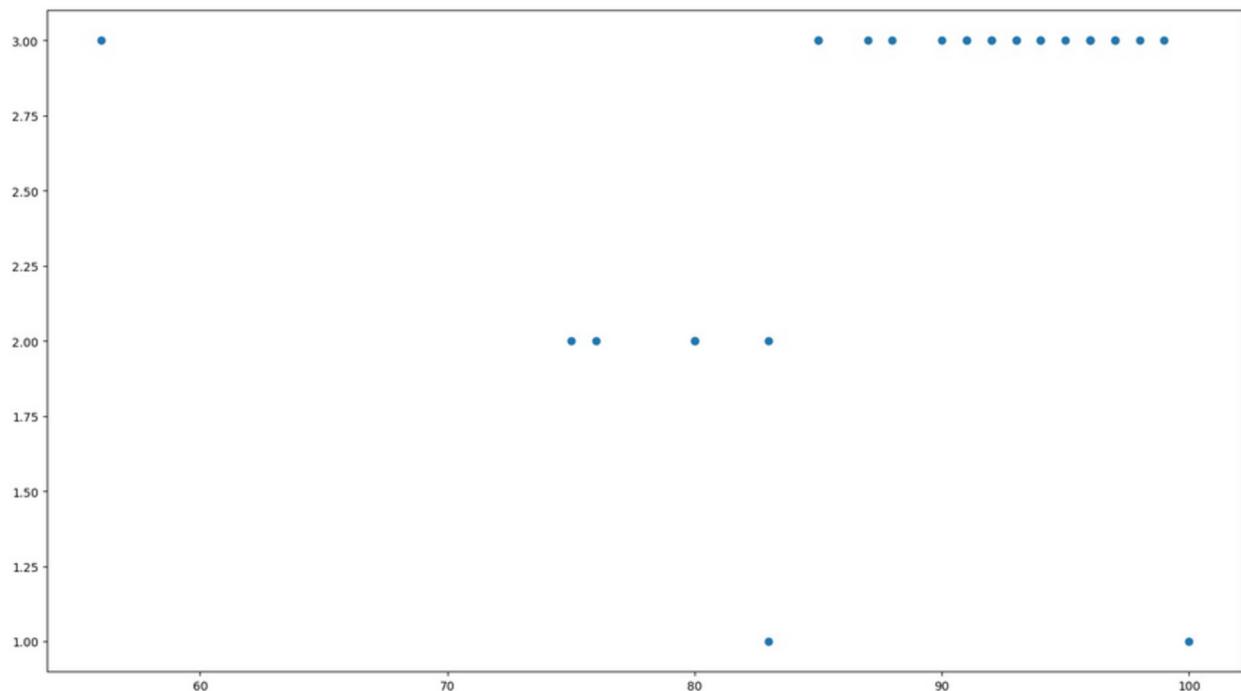
```

Requirement already satisfied: matplotlib in c:\users\welcome\anaconda3\lib\site-packages (3.8.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: numpy<2,>=1.21 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (1.26.4)
Requirement already satisfied: packaging>=20.0 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (23.1)

```

```
Requirement already satisfied: pillow>=6.2.0 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\welcome\anaconda3\lib\site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\welcome\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

```
fig, ax= plt.subplots(figsize = (18, 10))
ax.scatter(dfl['placement score'], dfl['placement offer count'])
plt.show()
ax.set_xlabel('((Proportion non-retail business acres)/(town))')
ax.set_ylabel('(Full-value property-tax rate)/($10,000)')
```



```
Text(4.444444444444452, 0.5, '(Full-value property-tax rate)/($10,000)')
print(py.where((dfl['placement score']<50) & (dfl['placement offer count']>1)))
print(py.where((dfl['placement score']>85) & (dfl['placement offer count']<3)))
(array([], dtype=int64),)
(array([10], dtype=int64),)
from scipy import stats
```

```
z = py.abs(stats.zscore(df1['math score']))

print(z)

0    1.252553
1    0.383665
2    0.274584
3    0.379903
4    2.456207
5    0.161741
6    1.143471
7    1.143471
8    0.816228
9    0.707147
10   0.598066
11   1.365395
12   0.270822
13   0.274584
14   0.274584
15   0.492746
16   1.252553
17   0.601827
18   0.488984
19   0.052660
20   1.470715
21   0.379903
22   0.601827
23   0.929071
24   1.365395
25   0.379903
26   0.492746
27   2.238044
28   1.252553
```

Name: math score, dtype: float64

threshold = 0.18

```
sample_outliers = py.where(z<threshold)
sample_outliers
(array([ 5, 19], dtype=int64,)

sorted_rscore = sorted(df1['reading score'])

sorted_rscore

[50,
 63,
 63,
 63,
 63,
```

```
63, 64, 64, 64, 64, 64, 65, 65, 65, 66, 67,  
67, 67, 68, 70, 70, 74, 76, 76, 76, 78, 95,  
95, 96]
```

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

```
64.0 74.0
```

```
IQR = q3-q1
```

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

```
49.0 89.0
```

```
r_outliers = []  
for i in sorted_rscore:
```

```
    if(i<lwr_bound or i>upr_bound):  
        r_outliers.append(i)  
print(r_outliers)
```

```
[95, 95, 96]
```

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

join year \	math score	reading score	writing score	placement score	club
0	60	63	64	76	95
2021	75	70	55	85	91
1	74	50	78	97	
2020	68	76	71	93	
2	94	67	92	94	
2020	61	78	78	80	
3	61	74	79	76	
2020	64	76	75	90	
4	65	95	67	100	
2020	66	76	71	92	
6	84	67	70	56	
2021	69	66	65	80	
7	74	65	72	96	
2021	74	63	80	96	
8	76	64	54	91	
2019	60	64	72	99	
9	77	70	64	87	
2020	67	95	96	92	
10	58	65	62	94	
2019	68	63	68	97	
11	77	63	86	85	
2020	80	64	67	83	
12	84	63	73	88	
2021	68	67			
13					
2021					
14					
2018					
15					
2020					
16					
2021					
17					
2020					
18					
2018					
20					
2019					
21					
2021					
22					
2021					
23					
2018					
24					
2018					
25					
2019					

26	76	64	68	96
2021	92	96	61	83
27	60	68	59	93
2018				
28				
2020				

```

placement offer count gender
0                      3 female
1                      3 male
2                      3 male
3                      3 female
4                      3 male
6                      3 male
7                      2 male
8                      2 male
9                      3 female
10                     1 male
11                     3 male
12                     3 female
13                     2 male
14                     3 male
15                     3 male
16                     3 female
17                     3 male
18                     3 female
20                     3 female
21                     3 male
22                     3 female
23                     3 female
24                     1 male
25                     3 female
26                     3 female
27                     2 male
28                     3 male
df_stud = df1

```

```

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: [60. 75. 74. 68. 84. 61. 61. 64. 65. 66. 84. 69. 74. 74.
76. 60. 77. 67.
58. 68. 77. 80. 84. 68. 76. 84. 60.]

```

```

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

```

	math score	m score	reading score	writing score	placement score
0	60	60.0	63	76	95
1	75	75.0	70	64	85
2	74	74.0	50	55	91
3	68	68.0	76	78	97
4	94	84.0	67	71	93
6	61	61.0	78	92	94
7	61	61.0	74	78	80
8	64	64.0	76	79	76
9	65	65.0	95	75	90
10	66	66.0	76	67	100
11	84	84.0	67	71	92
12	69	69.0	66	70	56
13	74	74.0	65	65	80
14	74	74.0	63	72	96
15	76	76.0	64	80	96
16	60	60.0	64	54	91
17	77	77.0	70	72	99
18	67	67.0	95	64	87
20	58	58.0	65	96	92
21	68	68.0	63	62	94
22	77	77.0	63	68	97
23	80	80.0	64	86	85
24	84	84.0	63	67	83
25	68	68.0	67	73	88
26	76	76.0	64	68	96

27	92	84.0	96	61	83
28	60	60.0	68	59	93

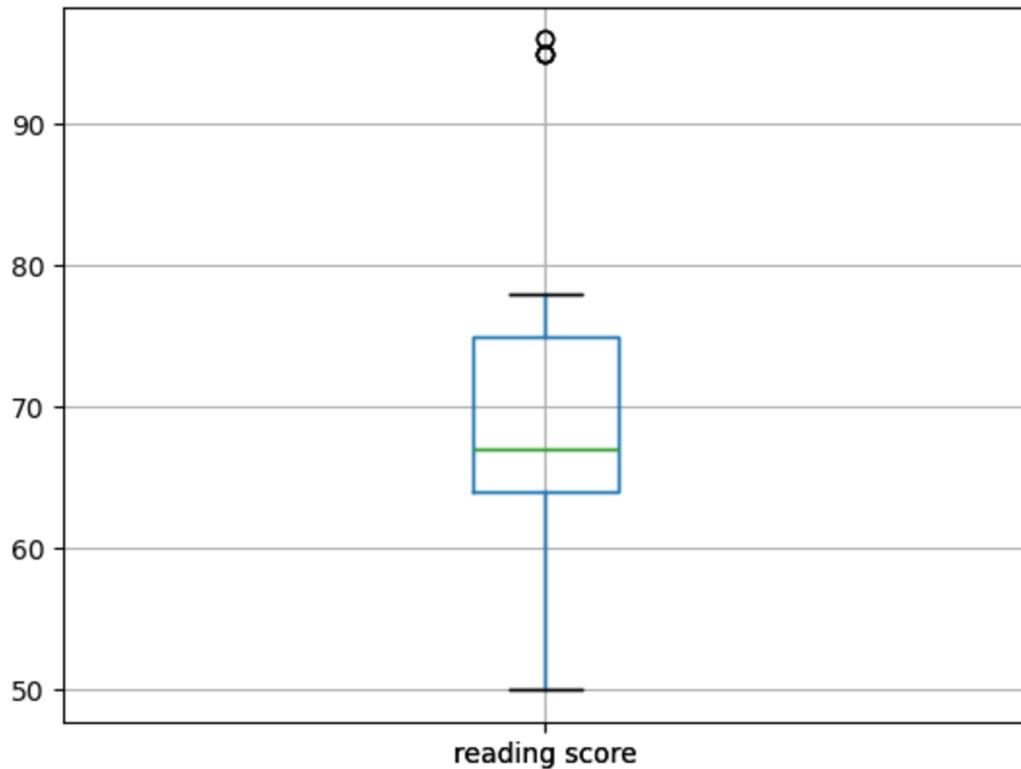
	club join year placement offer count	gender
0	2021	3 female
1	2020	3 male
2	2020	3 male
3	2020	3 female
4	2020	3 male
6	2021	3 male
7	2021	2 male
8	2019	2 male
9	2020	3 female
10	2019	1 male
11	2020	3 male
12	2021	3 female
13	2021	2 male
14	2018	3 male
15	2020	3 male
16	2021	3 female
17	2020	3 male
18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

```
col1 = ['reading score']
```

```
df1.boxplot(col1)
```

```
<Axes: >
```

```
plt.show()
```



```

median = np.median(sorted_rscores)
median
66.0
refined_df = df[
    refined_df['reading score'] > upr_bound, median, refined_df['reading score'])]
refined_df

```

	math score	m score	reading score	writing score	placement score
0	60	60.0	63.0	76	95
1	75	75.0	70.0	64	85
2	74	74.0	50.0	55	91
3	68	68.0	76.0	78	97
4	94	84.0	67.0	71	93
6	61	61.0	78.0	92	94
7	61	61.0	74.0	78	80

8	9	64	64.0	76.0	79	76
10		65	65.0	66.0	75	90
11		66	66.0	76.0	67	100
12		84	84.0	67.0	71	92
13		69	69.0	66.0	70	56
14		74	74.0	65.0	65	80
15		74	74.0	63.0	72	96
16		76	76.0	64.0	80	96
17		60	60.0	64.0	54	91
18		77	77.0	70.0	72	99
20		67	67.0	66.0	64	87
21		58	58.0	65.0	96	92
22		68	68.0	63.0	62	94
23		77	77.0	63.0	68	97
24		80	80.0	64.0	86	85
25		84	84.0	63.0	67	83
26		68	68.0	67.0	73	88
27		76	76.0	64.0	68	96
28		92	84.0	66.0	61	83
		60	60.0	68.0	59	93

	club	join year	placement offer count	gender
0		2021		3 female
1		2020		3 male
2		2020		3 male
3		2020		3 female
4		2020		3 male
6		2021		3 male
7		2021		2 male
8		2019		2 male
9		2020		3 female

10	2019	1	male
11	2020	3	male
12	2021	3	female
13	2021	2	male
14	2018	3	male
15	2020	3	male
16	2021	3	female
17	2020	3	male
18	2018	3	female
20	2019	3	female
21	2021	3	male
22	2021	3	female
23	2018	3	female
24	2018	1	male
25	2019	3	female
26	2021	3	female
27	2018	2	male
28	2020	3	male

```
refined_df['reading score'] = py.where(refined_df['reading score']<lwr_bound, median, refined_df['reading score'])
refined_df
```

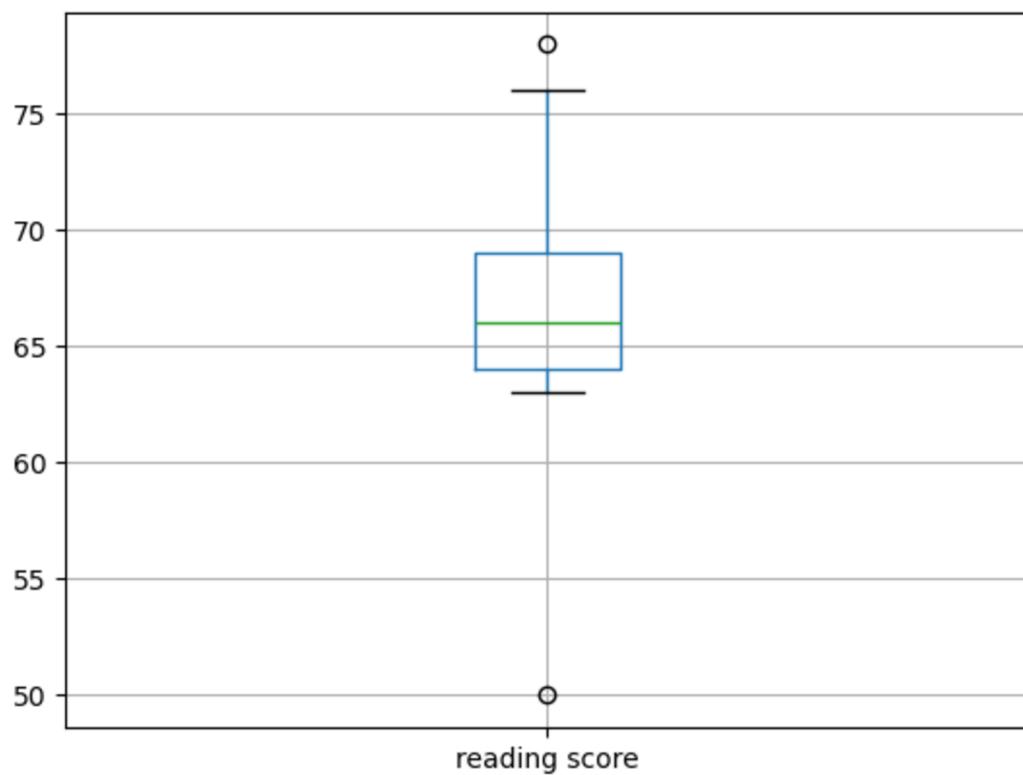
	math score	m score	reading score	writing score	placement score
0	60	60.0	63.0	76	95
1	75	75.0	70.0	64	85
2	74	74.0	50.0	55	91
3	68	68.0	76.0	78	97
4	94	84.0	67.0	71	93
6	61	61.0	78.0	92	94
7	61	61.0	74.0	78	80
8	64	64.0	76.0	79	76
9	65	65.0	66.0	75	90
10	66	66.0	76.0	67	100
11	84	84.0	67.0	71	92
12	69	69.0	66.0	70	56
13	74	74.0	65.0	65	80

14	74	74.0	63.0	72	96
15	76	76.0	64.0	80	96
16	60	60.0	64.0	54	91
17	77	77.0	70.0	72	99
18	67	67.0	66.0	64	87
20	58	58.0	65.0	96	92
21	68	68.0	63.0	62	94
22	77	77.0	63.0	68	97
23	80	80.0	64.0	86	85
24	84	84.0	63.0	67	83
25	68	68.0	67.0	73	88
26	76	76.0	64.0	68	96
27	92	84.0	66.0	61	83
28	60	60.0	68.0	59	93

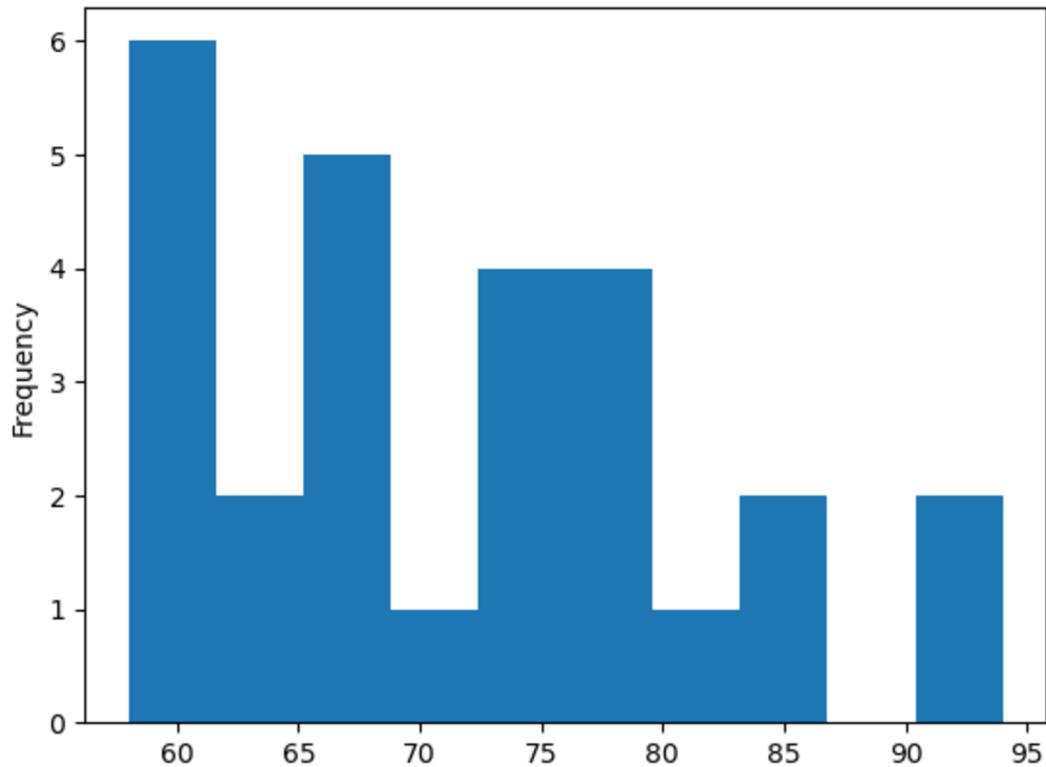
	club	join year	placement offer count	gender
0		2021	3	female
1		2020	3	male
2		2020	3	male
3		2020	3	female
4		2020	3	male
6		2021	3	male
7		2021	2	male
8		2019	2	male
9		2020	3	female
10		2019	1	male
11		2020	3	male
12		2021	3	female
13		2021	2	male
14		2018	3	male
15		2020	3	male
16		2021	3	female
17		2020	3	male
18		2018	3	female
20		2019	3	female
21		2021	3	male
22		2021	3	female

```
23          2018          3  female
24          2018          1   male
25          2019          3  female
26          2021          3  female
27          2018          2   male
28          2020          3   male
```

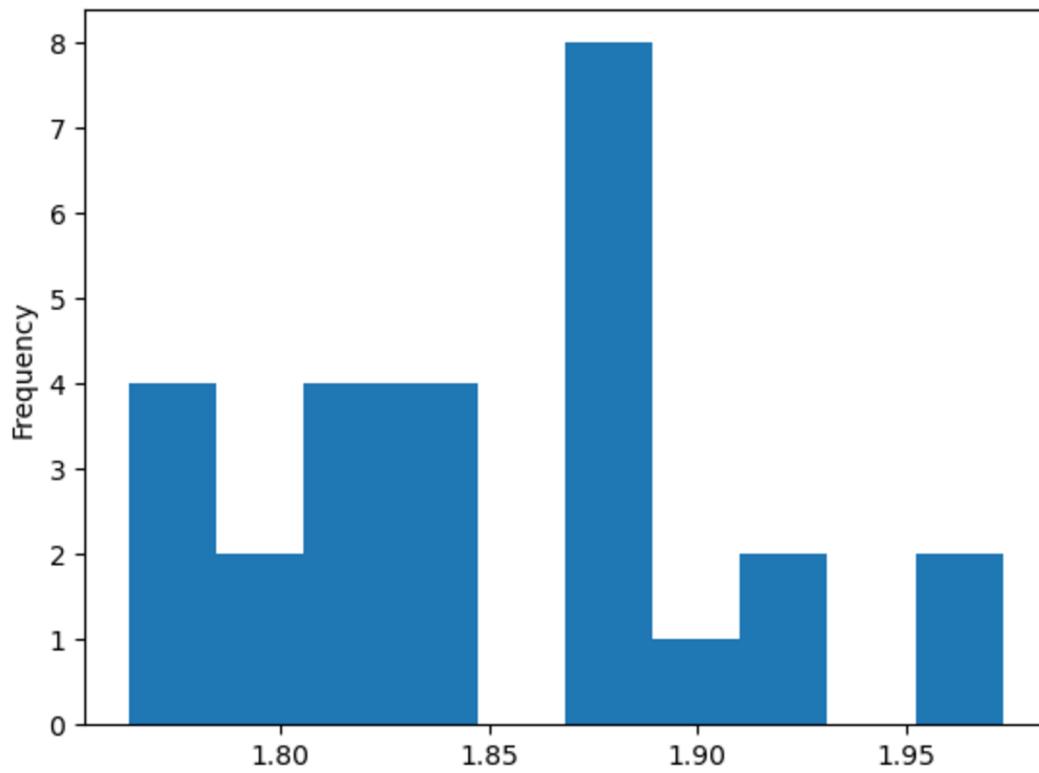
```
col2 = ['reading score']
refined_df.boxplot(col2)
<Axes: >
plt.show()
```



```
new_df['math score'].plot(kind = 'hist')
<Axes: ylabel='Frequency'>
plt.show()
```



```
df1['log_math'] = np.log10(df1['math score'])  
df1['log_math'].plot(kind = 'hist')  
<Axes: ylabel='Frequency'>  
plt.show()
```



Name:Patil Diksha Sudhakar

Rollno:13261 B3