In [1]:
```
import pandas as pd
import numpy as np
```

In [4]:
```
df=pd.read_csv(r"C:\Users\akash\Downloads\StudentPerformance.csv")
df
```

Out[4]:

	math score	reading score	writing score	placement score	club join year	placement offer count	gender
0	79.0	70.0	79.0	100.0	2018	3	male
1	77.0	64.0	68.0	NaN	2018	1	female
2	21.0	NaN	NaN	87.0	2021	3	female
3	75.0	60.0	60.0	99.0	2018	3	female
4	72.0	NaN	NaN	94.0	2018	3	male
5	69.0	61.0	69.0	97.0	2020	3	female
6	60.0	64.0	72.0	86.0	2020	3	male
7	20.0	60.0	72.0	NaN	2018	1	female
8	NaN	80.0	63.0	77.0	2020	2	female
9	72.0	78.0	78.0	75.0	2020	2	female
10	72.0	NaN	NaN	NaN	2020	1	male
11	70.0	72.0	74.0	95.0	2018	3	female
12	61.0	71.0	67.0	81.0	2020	2	female
13	52.0	71.0	77.0	81.0	2021	2	male
14	73.0	NaN	68.0	97.0	2019	3	female
15	NaN	61.0	77.0	100.0	2020	3	female
16	11.0	75.0	70.0	NaN	2018	1	male
17	66.0	62.0	NaN	78.0	2021	2	male
18	61.0	NaN	62.0	87.0	2021	3	female
19	66.0	66.0	70.0	80.0	2020	2	female
20	NaN	69.0	65.0	86.0	2021	3	female
21	24.0	72.0	75.0	NaN	2018	1	male
22	69.0	76.0	62.0	76.0	2018	2	female
23	71.0	78.0	NaN	77.0	2021	2	male
24	66.0	63.0	71.0	76.0	2019	2	female
25	62.0	NaN	80.0	88.0	2019	3	female
26	72.0	71.0	77.0	80.0	2019	2	male
27	75.0	67.0	NaN	NaN	2021	1	female
28	70.0	70.0	75.0	97.0	2018	3	male
29	71.0	64.0	67.0	87.0	2021	3	male

In [5]:

```
df.isnull()
```

Out[5]:

	math score	reading score	writing score	placement score	club join year	placement offer count	gender
0	False	False	False	False	False	False	False
1	False	False	False	True	False	False	False
2	False	True	True	False	False	False	False
3	False	False	False	False	False	False	False
4	False	True	True	False	False	False	False
5	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False
7	False	False	False	True	False	False	False
8	True	False	False	False	False	False	False
9	False	False	False	False	False	False	False
10	False	True	True	True	False	False	False
11	False	False	False	False	False	False	False
12	False	False	False	False	False	False	False
13	False	False	False	False	False	False	False
14	False	True	False	False	False	False	False
15	True	False	False	False	False	False	False
16	False	False	False	True	False	False	False
17	False	False	True	False	False	False	False
18	False	True	False	False	False	False	False
19	False	False	False	False	False	False	False
20	True	False	False	False	False	False	False
21	False	False	False	True	False	False	False
22	False	False	False	False	False	False	False
23	False	False	True	False	False	False	False
24	False	False	False	False	False	False	False
25	False	True	False	False	False	False	False
26	False	False	False	False	False	False	False
27	False	False	True	True	False	False	False
28	False	False	False	False	False	False	False
29	False	False	False	False	False	False	False

In [6]:

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

Out[6]:

	math score	reading score	writing score	placement score	club join year	placement offer count	gender
8	NaN	80.0	63.0	77.0	2020	2	female
15	NaN	61.0	77.0	100.0	2020	3	female
20	NaN	69.0	65.0	86.0	2021	3	female

In [7]:

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

Out[7]:

	math score	reading score	writing score	placement score	club join year	placement offer count	gender
0	79.0	70.0	79.0	100.0	2018	3	male
1	77.0	64.0	68.0	NaN	2018	1	female
2	21.0	NaN	NaN	87.0	2021	3	female
3	75.0	60.0	60.0	99.0	2018	3	female
4	72.0	NaN	NaN	94.0	2018	3	male
5	69.0	61.0	69.0	97.0	2020	3	female
6	60.0	64.0	72.0	86.0	2020	3	male
7	20.0	60.0	72.0	NaN	2018	1	female
9	72.0	78.0	78.0	75.0	2020	2	female
10	72.0	NaN	NaN	NaN	2020	1	male
11	70.0	72.0	74.0	95.0	2018	3	female
12	61.0	71.0	67.0	81.0	2020	2	female
13	52.0	71.0	77.0	81.0	2021	2	male
14	73.0	NaN	68.0	97.0	2019	3	female
16	11.0	75.0	70.0	NaN	2018	1	male
17	66.0	62.0	NaN	78.0	2021	2	male
18	61.0	NaN	62.0	87.0	2021	3	female
19	66.0	66.0	70.0	80.0	2020	2	female
21	24.0	72.0	75.0	NaN	2018	1	male
22	69.0	76.0	62.0	76.0	2018	2	female
23	71.0	78.0	NaN	77.0	2021	2	male
24	66.0	63.0	71.0	76.0	2019	2	female
25	62.0	NaN	80.0	88.0	2019	3	female
26	72.0	71.0	77.0	80.0	2019	2	male
27	75.0	67.0	NaN	NaN	2021	1	female
28	70.0	70.0	75.0	97.0	2018	3	male
29	71.0	64.0	67.0	87.0	2021	3	male

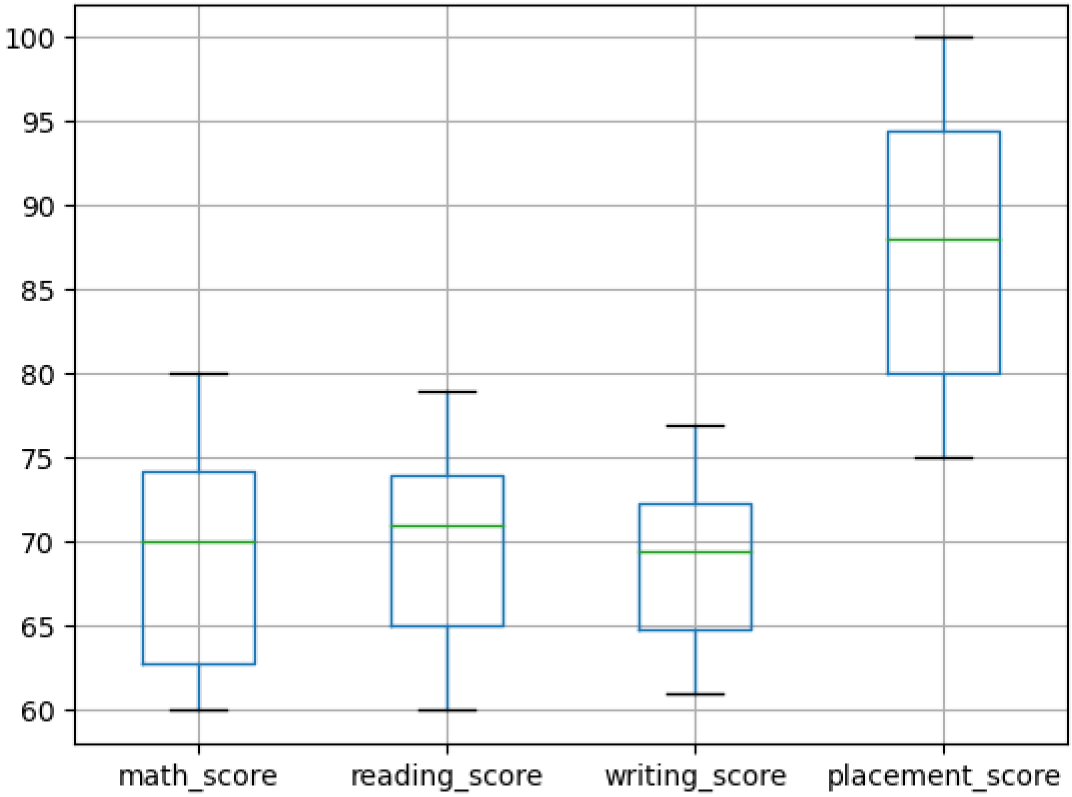
In [8]:

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

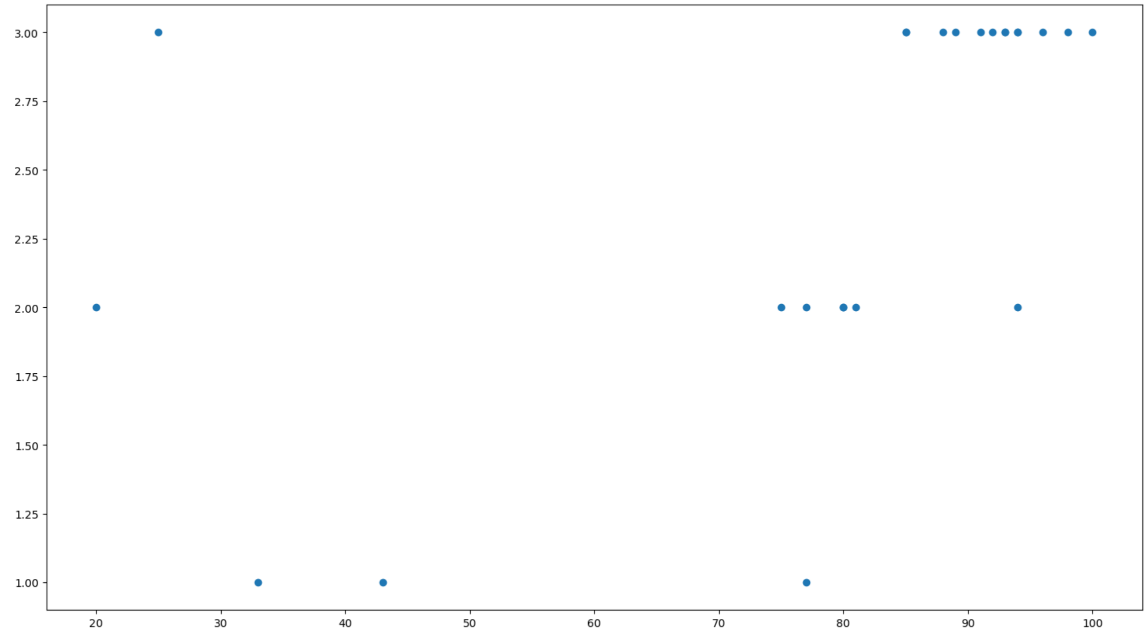
Out[8]:

	math score	reading score	writing score	placement score	club join year	placement offer count	gender
0	79.0	70.0	79.0	100.0	2018	3	1
1	77.0	64.0	68.0	NaN	2018	1	0
2	21.0	NaN	NaN	87.0	2021	3	0
3	75.0	60.0	60.0	99.0	2018	3	0
4	72.0	NaN	NaN	94.0	2018	3	1
5	69.0	61.0	69.0	97.0	2020	3	0
6	60.0	64.0	72.0	86.0	2020	3	1
7	20.0	60.0	72.0	NaN	2018	1	0
8	NaN	80.0	63.0	77.0	2020	2	0
9	72.0	78.0	78.0	75.0	2020	2	0
10	72.0	NaN	NaN	NaN	2020	1	1
11	70.0	72.0	74.0	95.0	2018	3	0
12	61.0	71.0	67.0	81.0	2020	2	0
13	52.0	71.0	77.0	81.0	2021	2	1
14	73.0	NaN	68.0	97.0	2019	3	0
15	NaN	61.0	77.0	100.0	2020	3	0
16	11.0	75.0	70.0	NaN	2018	1	1
17	66.0	62.0	NaN	78.0	2021	2	1
18	61.0	NaN	62.0	87.0	2021	3	0
19	66.0	66.0	70.0	80.0	2020	2	0
20	NaN	69.0	65.0	86.0	2021	3	0
21	24.0	72.0	75.0	NaN	2018	1	1
22	69.0	76.0	62.0	76.0	2018	2	0
23	71.0	78.0	NaN	77.0	2021	2	1
24	66.0	63.0	71.0	76.0	2019	2	0
25	62.0	NaN	80.0	88.0	2019	3	0
26	72.0	71.0	77.0	80.0	2019	2	1
27	75.0	67.0	NaN	NaN	2021	1	0
28	70.0	70.0	75.0	97.0	2018	3	1
29	71.0	64.0	67.0	87.0	2021	3	1

```
In [54]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         df=pd.read_csv(r"C:\Users\akash\Downloads\StudentPerformance.csv")
         col = ['math_score', 'reading_score' , 'writing_score','placement_score']
         df.boxplot(col)
         plt.show()
```



In [61]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv(r"C:\Users\akash\Downloads\StudentPerformance.csv")
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['placement_score'], df['placement_offer_count'])
plt.show()
```

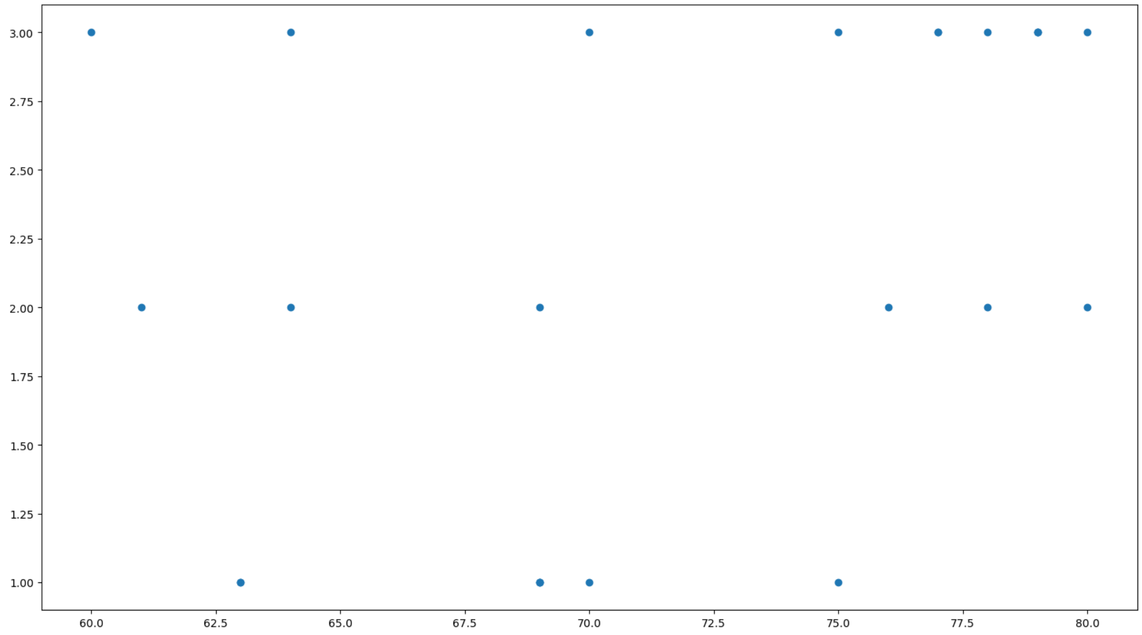


In [62]:
```
print(np.where((df['placement_score']<50) & (df['placement_offer_count']>1)
print(np.where((df['placement_score']>85) & (df['placement_offer_count']<3)
```

```
(array([1, 6], dtype=int64),)
(array([11], dtype=int64),)
```

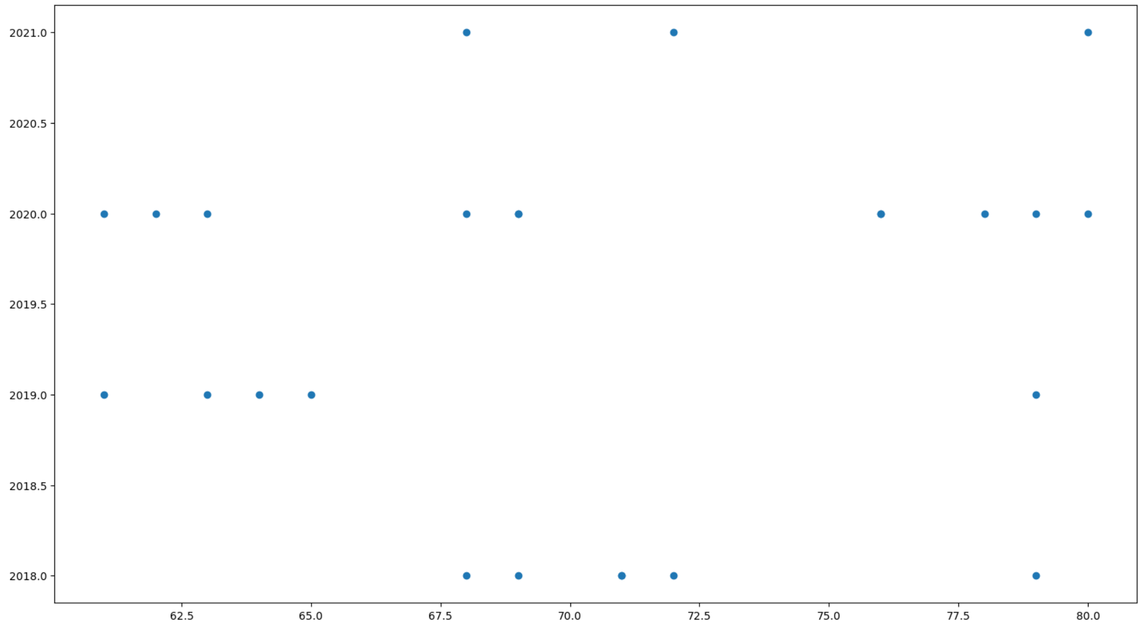
In [63]:

```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['math_score'], df['placement_offer_count'])
plt.show()
```



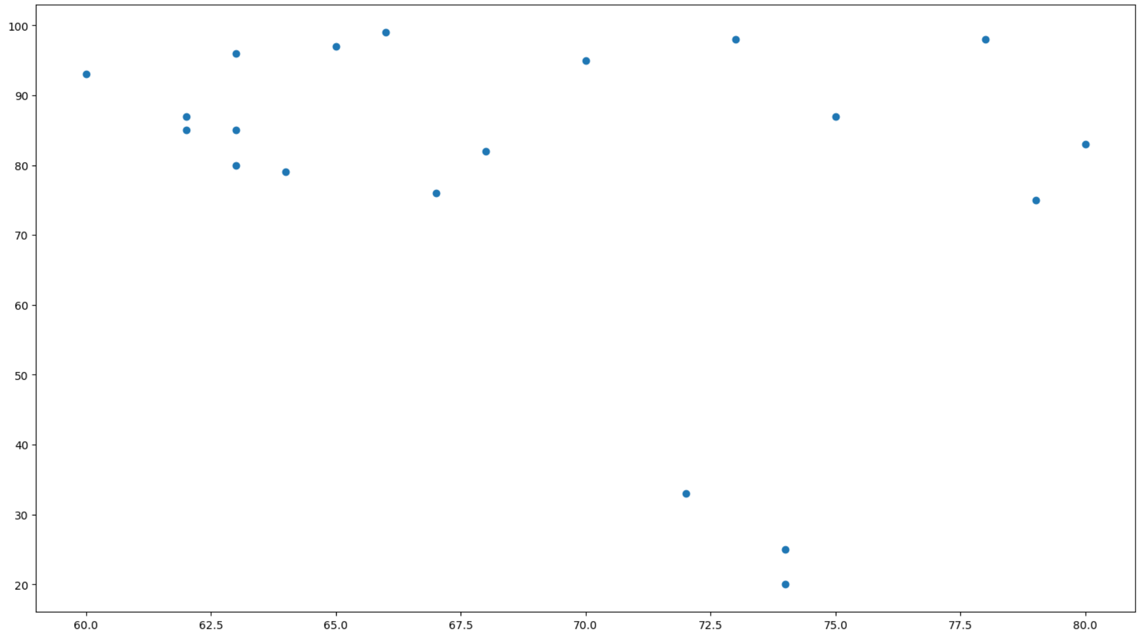
In [64]:

```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['reading_score'], df['club_join_year'])
plt.show()
```



In [77]:

```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['reading_score'], df['placement_score'])
plt.show()
```



In [67]:

```
# Detecting outliers using Z-Score
from scipy import stats
df=pd.read_csv(r"C:\Users\akash\Downloads\StudentPerformance.csv")
z = np.abs(stats.zscore(df['math_score']))
print(z)
```

```
0     0.452445
1     0.150815
2     0.271467
3     2.865487
4     0.392119
5     0.693749
6     0.211141
7     0.392119
8     2.684509
9     0.090489
10    0.633423
11    0.331793
12    0.331793
13    2.020922
14    0.211141
15    0.452445
16    0.271467
17    0.573097
18    1.960596
19    0.331793
20    0.392119
21    0.452445
22    0.392119
23    0.331793
24    1.658966
25    0.512771
26    0.090489
27    0.271467
28    0.090489
29    0.633423
Name: math_score, dtype: float64
```

In [68]:

```
threshold = 0.18
sample_outliers = np.where(z <threshold)
sample_outliers
```

Out[68]:

```
(array([ 1,  9, 26, 28], dtype=int64),)
```

In [69]:

```
# Detecting outliers using Inter Quantile Range(IQR)
sorted_rscore= sorted(df['math_score'])
sorted_rscore
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
```

```
63.0 75.0
```

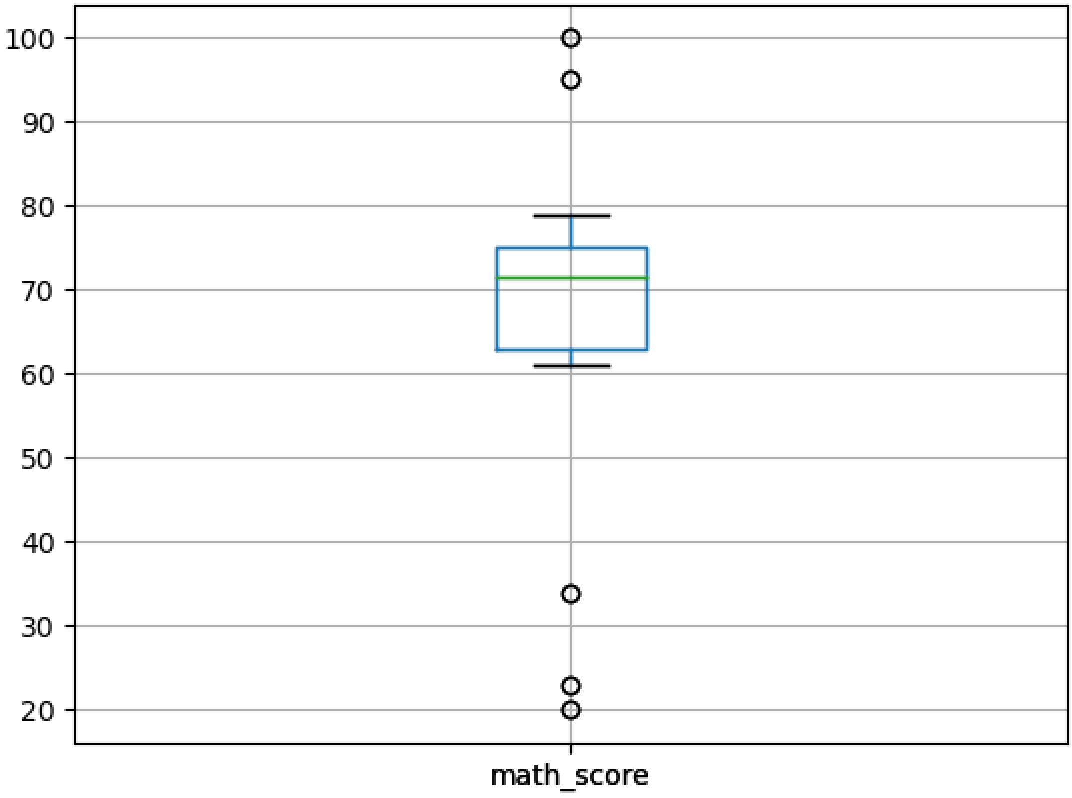
```
In [70]: IQR = q3-q1
         lwr_bound = q1-(1.5*IQR)
         upr_bound = q3+(1.5*IQR)
         print(lwr_bound, upr_bound)

         45.0 93.0
```

```
In [73]: r_outliers = []
         for i in sorted_rscore:
             if (i<lwr_bound or i>upr_bound):
                 r_outliers.append(i)
         print(r_outliers)

         [20, 23, 34, 95, 100]
```

```
In [75]: col = ['math_score']
         df.boxplot(col)
         plt.show()
```



```
In [76]: median=np.median(sorted_rscore)
         median

Out[76]: 71.5
```

# Ayush_Apte_13111_TE_A1