Assignment no 2

```
Aim:
1.Creation of dataset using microsoft excel.
2.Identification and handling of null values.
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
```

In [57]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

In [2]:
```
df=pd.read_csv("Student_performance.csv")
```

In [3]:
```
df
```

Out[3]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.0	90.0	80.0	79.0	2024.0	2
1	67.0	92.0	60.0	75.0	2023.0	2
2	62.0	75.0	61.0	75.0	2022.0	2
3	64.0	77.0	73.0	91.0	2022.0	3
4	76.0	85.0	79.0	75.0	2024.0	2
5	72.0	92.0	77.0	NaN	2025.0	3
6	77.0	NaN	76.0	95.0	2022.0	3
7	78.0	79.0	71.0	78.0	2025.0	2
8	62.0	81.0	80.0	83.0	2022.0	2
9	74.0	84.0	68.0	77.0	2026.0	2
10	76.0	75.0	200.0	80.0	2022.0	2
11	78.0	89.0	63.0	84.0	2022.0	2
12	66.0	76.0	68.0	86.0	2026.0	3
13	60.0	83.0	NaN	100.0	2022.0	3
14	76.0	92.0	78.0	94.0	2023.0	3
15	62.0	79.0	60.0	76.0	2025.0	2
16	71.0	81.0	75.0	82.0	2026.0	2
17	73.0	81.0	61.0	98.0	2025.0	3
18	60.0	77.0	67.0	95.0	2024.0	3
19	79.0	92.0	70.0	85.0	2022.0	3
20	75.0	79.0	75.0	83.0	2026.0	2
21	73.0	86.0	66.0	98.0	2025.0	3
22	78.0	90.0	70.0	97.0	2026.0	3
23	60.0	91.0	75.0	93.0	2026.0	3
24	63.0	91.0	63.0	99.0	2025.0	3
25	NaN	81.0	65.0	97.0	2023.0	3
26	76.0	82.0	77.0	81.0	2025.0	2
27	76.0	84.0	79.0	89.0	2024.0	3
28	71.0	NaN	NaN	NaN	NaN	Na

In [4]:
```
df.isnull()
```

Out[4]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	False	False	False	False	False	Fal
1	False	False	False	False	False	Fal
2	False	False	False	False	False	Fal
3	False	False	False	False	False	Fal
4	False	False	False	False	False	Fal
5	False	False	False	True	False	Fal
6	False	True	False	False	False	Fal
7	False	False	False	False	False	Fal
8	False	False	False	False	False	Fal
9	False	False	False	False	False	Fal
10	False	False	False	False	False	Fal
11	False	False	False	False	False	Fal
12	False	False	False	False	False	Fal
13	False	False	True	False	False	Fal
14	False	False	False	False	False	Fal
15	False	False	False	False	False	Fal
16	False	False	False	False	False	Fal
17	False	False	False	False	False	Fal
18	False	False	False	False	False	Fal
19	False	False	False	False	False	Fal
20	False	False	False	False	False	Fal
21	False	False	False	False	False	Fal
22	False	False	False	False	False	Fal
23	False	False	False	False	False	Fal
24	False	False	False	False	False	Fal
25	True	False	False	False	False	Fal
26	False	False	False	False	False	Fal
27	False	False	False	False	False	Fal
28	False	True	True	True	True	Tr

In [5]:
```
series=pd.isnull(df["Math_Score"])
df[series]
```

Out[5]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
25	NaN	81.0	65.0	97.0	2023.0	3

In [6]:
```
df.notnull()
```

Out[6]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	True	True	True	True	True	Tr
1	True	True	True	True	True	Tr
2	True	True	True	True	True	Tr
3	True	True	True	True	True	Tr
4	True	True	True	True	True	Tr
5	True	True	True	False	True	Tr
6	True	False	True	True	True	Tr
7	True	True	True	True	True	Tr
8	True	True	True	True	True	Tr
9	True	True	True	True	True	Tr
10	True	True	True	True	True	Tr
11	True	True	True	True	True	Tr
12	True	True	True	True	True	Tr
13	True	True	False	True	True	Tr
14	True	True	True	True	True	Tr
15	True	True	True	True	True	Tr
16	True	True	True	True	True	Tr
17	True	True	True	True	True	Tr
18	True	True	True	True	True	Tr
19	True	True	True	True	True	Tr
20	True	True	True	True	True	Tr
21	True	True	True	True	True	Tr
22	True	True	True	True	True	Tr
23	True	True	True	True	True	Tr
24	True	True	True	True	True	Tr
25	False	True	True	True	True	Tr
26	True	True	True	True	True	Tr
27	True	True	True	True	True	Tr
28	True	False	False	False	False	Fal

In [10]:
```
series1=pd.notnull(df["Math_Score"])
df[series1]
```

Out[10]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.0	90.0	80.0	79.0	2024.0	2
1	67.0	92.0	60.0	75.0	2023.0	2
2	62.0	75.0	61.0	75.0	2022.0	2
3	64.0	77.0	73.0	91.0	2022.0	3
4	76.0	85.0	79.0	75.0	2024.0	2
5	72.0	92.0	77.0	NaN	2025.0	3
6	77.0	NaN	76.0	95.0	2022.0	3
7	78.0	79.0	71.0	78.0	2025.0	2
8	62.0	81.0	80.0	83.0	2022.0	2
9	74.0	84.0	68.0	77.0	2026.0	2
10	76.0	75.0	200.0	80.0	2022.0	2
11	78.0	89.0	63.0	84.0	2022.0	2
12	66.0	76.0	68.0	86.0	2026.0	3
13	60.0	83.0	NaN	100.0	2022.0	3
14	76.0	92.0	78.0	94.0	2023.0	3
15	62.0	79.0	60.0	76.0	2025.0	2
16	71.0	81.0	75.0	82.0	2026.0	2
17	73.0	81.0	61.0	98.0	2025.0	3
18	60.0	77.0	67.0	95.0	2024.0	3
19	79.0	92.0	70.0	85.0	2022.0	3
20	75.0	79.0	75.0	83.0	2026.0	2
21	73.0	86.0	66.0	98.0	2025.0	3
22	78.0	90.0	70.0	97.0	2026.0	3
23	60.0	91.0	75.0	93.0	2026.0	3
24	63.0	91.0	63.0	99.0	2025.0	3
26	76.0	82.0	77.0	81.0	2025.0	2
27	76.0	84.0	79.0	89.0	2024.0	3
28	71.0	NaN	NaN	NaN	NaN	Na

In [16]:
```
ndf=df
ndf.fillna(0)
```

Out[16]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.0	90.0	80.0	79.0	2024.0	2
1	67.0	92.0	60.0	75.0	2023.0	2
2	62.0	75.0	61.0	75.0	2022.0	2
3	64.0	77.0	73.0	91.0	2022.0	3
4	76.0	85.0	79.0	75.0	2024.0	2
5	72.0	92.0	77.0	0.0	2025.0	3
6	77.0	0.0	76.0	95.0	2022.0	3
7	78.0	79.0	71.0	78.0	2025.0	2
8	62.0	81.0	80.0	83.0	2022.0	2
9	74.0	84.0	68.0	77.0	2026.0	2
10	76.0	75.0	200.0	80.0	2022.0	2
11	78.0	89.0	63.0	84.0	2022.0	2
12	66.0	76.0	68.0	86.0	2026.0	3
13	60.0	83.0	0.0	100.0	2022.0	3
14	76.0	92.0	78.0	94.0	2023.0	3
15	62.0	79.0	60.0	76.0	2025.0	2
16	71.0	81.0	75.0	82.0	2026.0	2
17	73.0	81.0	61.0	98.0	2025.0	3
18	60.0	77.0	67.0	95.0	2024.0	3
19	79.0	92.0	70.0	85.0	2022.0	3
20	75.0	79.0	75.0	83.0	2026.0	2
21	73.0	86.0	66.0	98.0	2025.0	3
22	78.0	90.0	70.0	97.0	2026.0	3
23	60.0	91.0	75.0	93.0	2026.0	3
24	63.0	91.0	63.0	99.0	2025.0	3
25	0.0	81.0	65.0	97.0	2023.0	3
26	76.0	82.0	77.0	81.0	2025.0	2
27	76.0	84.0	79.0	89.0	2024.0	3
28	71.0	0.0	0.0	0.0	0.0	0

In [17]:
```
m_v=df['Math_Score'].mean()
df['Math_Score'].fillna(value=m_v,inplace=True)
df
```

Out[17]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.000000	90.0	80.0	79.0	2024.0	2
1	67.000000	92.0	60.0	75.0	2023.0	2
2	62.000000	75.0	61.0	75.0	2022.0	2
3	64.000000	77.0	73.0	91.0	2022.0	3
4	76.000000	85.0	79.0	75.0	2024.0	2
5	72.000000	92.0	77.0	NaN	2025.0	3
6	77.000000	NaN	76.0	95.0	2022.0	3
7	78.000000	79.0	71.0	78.0	2025.0	2
8	62.000000	81.0	80.0	83.0	2022.0	2
9	74.000000	84.0	68.0	77.0	2026.0	2
10	76.000000	75.0	200.0	80.0	2022.0	2
11	78.000000	89.0	63.0	84.0	2022.0	2
12	66.000000	76.0	68.0	86.0	2026.0	3
13	60.000000	83.0	NaN	100.0	2022.0	3
14	76.000000	92.0	78.0	94.0	2023.0	3
15	62.000000	79.0	60.0	76.0	2025.0	2
16	71.000000	81.0	75.0	82.0	2026.0	2
17	73.000000	81.0	61.0	98.0	2025.0	3
18	60.000000	77.0	67.0	95.0	2024.0	3
19	79.000000	92.0	70.0	85.0	2022.0	3
20	75.000000	79.0	75.0	83.0	2026.0	2
21	73.000000	86.0	66.0	98.0	2025.0	3
22	78.000000	90.0	70.0	97.0	2026.0	3
23	60.000000	91.0	75.0	93.0	2026.0	3
24	63.000000	91.0	63.0	99.0	2025.0	3
25	70.714286	81.0	65.0	97.0	2023.0	3
26	76.000000	82.0	77.0	81.0	2025.0	2
27	76.000000	84.0	79.0	89.0	2024.0	3
28	71.000000	NaN	NaN	NaN	NaN	Na

In [18]:
```
ndf.replace(to_replace=np.nan,value=-99)
```

Out[18]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.000000	90.0	80.0	79.0	2024.0	2
1	67.000000	92.0	60.0	75.0	2023.0	2
2	62.000000	75.0	61.0	75.0	2022.0	2
3	64.000000	77.0	73.0	91.0	2022.0	3
4	76.000000	85.0	79.0	75.0	2024.0	2
5	72.000000	92.0	77.0	-99.0	2025.0	3
6	77.000000	-99.0	76.0	95.0	2022.0	3
7	78.000000	79.0	71.0	78.0	2025.0	2
8	62.000000	81.0	80.0	83.0	2022.0	2
9	74.000000	84.0	68.0	77.0	2026.0	2
10	76.000000	75.0	200.0	80.0	2022.0	2
11	78.000000	89.0	63.0	84.0	2022.0	2
12	66.000000	76.0	68.0	86.0	2026.0	3
13	60.000000	83.0	-99.0	100.0	2022.0	3
14	76.000000	92.0	78.0	94.0	2023.0	3
15	62.000000	79.0	60.0	76.0	2025.0	2
16	71.000000	81.0	75.0	82.0	2026.0	2
17	73.000000	81.0	61.0	98.0	2025.0	3
18	60.000000	77.0	67.0	95.0	2024.0	3
19	79.000000	92.0	70.0	85.0	2022.0	3
20	75.000000	79.0	75.0	83.0	2026.0	2
21	73.000000	86.0	66.0	98.0	2025.0	3
22	78.000000	90.0	70.0	97.0	2026.0	3
23	60.000000	91.0	75.0	93.0	2026.0	3
24	63.000000	91.0	63.0	99.0	2025.0	3
25	70.714286	81.0	65.0	97.0	2023.0	3
26	76.000000	82.0	77.0	81.0	2025.0	2
27	76.000000	84.0	79.0	89.0	2024.0	3
28	71.000000	-99.0	-99.0	-99.0	-99.0	-99

In [19]:
```
df=pd.read_csv("Student_performance.csv")
df
```

Out[19]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.0	90.0	80.0	79.0	2024.0	2
1	67.0	92.0	60.0	75.0	2023.0	2
2	62.0	75.0	61.0	75.0	2022.0	2
3	64.0	77.0	73.0	91.0	2022.0	3
4	76.0	85.0	79.0	75.0	2024.0	2
5	72.0	92.0	77.0	NaN	2025.0	3
6	77.0	NaN	76.0	95.0	2022.0	3
7	78.0	79.0	71.0	78.0	2025.0	2
8	62.0	81.0	80.0	83.0	2022.0	2
9	74.0	84.0	68.0	77.0	2026.0	2
10	76.0	75.0	200.0	80.0	2022.0	2
11	78.0	89.0	63.0	84.0	2022.0	2
12	66.0	76.0	68.0	86.0	2026.0	3
13	60.0	83.0	NaN	100.0	2022.0	3
14	76.0	92.0	78.0	94.0	2023.0	3
15	62.0	79.0	60.0	76.0	2025.0	2
16	71.0	81.0	75.0	82.0	2026.0	2
17	73.0	81.0	61.0	98.0	2025.0	3
18	60.0	77.0	67.0	95.0	2024.0	3
19	79.0	92.0	70.0	85.0	2022.0	3
20	75.0	79.0	75.0	83.0	2026.0	2
21	73.0	86.0	66.0	98.0	2025.0	3
22	78.0	90.0	70.0	97.0	2026.0	3
23	60.0	91.0	75.0	93.0	2026.0	3
24	63.0	91.0	63.0	99.0	2025.0	3
25	NaN	81.0	65.0	97.0	2023.0	3
26	76.0	82.0	77.0	81.0	2025.0	2
27	76.0	84.0	79.0	89.0	2024.0	3
28	71.0	NaN	NaN	NaN	NaN	Na

In [21]:
```
df.dropna(how='all')
```

Out[21]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.0	90.0	80.0	79.0	2024.0	2
1	67.0	92.0	60.0	75.0	2023.0	2
2	62.0	75.0	61.0	75.0	2022.0	2
3	64.0	77.0	73.0	91.0	2022.0	3
4	76.0	85.0	79.0	75.0	2024.0	2
5	72.0	92.0	77.0	NaN	2025.0	3
6	77.0	NaN	76.0	95.0	2022.0	3
7	78.0	79.0	71.0	78.0	2025.0	2
8	62.0	81.0	80.0	83.0	2022.0	2
9	74.0	84.0	68.0	77.0	2026.0	2
10	76.0	75.0	200.0	80.0	2022.0	2
11	78.0	89.0	63.0	84.0	2022.0	2
12	66.0	76.0	68.0	86.0	2026.0	3
13	60.0	83.0	NaN	100.0	2022.0	3
14	76.0	92.0	78.0	94.0	2023.0	3
15	62.0	79.0	60.0	76.0	2025.0	2
16	71.0	81.0	75.0	82.0	2026.0	2
17	73.0	81.0	61.0	98.0	2025.0	3
18	60.0	77.0	67.0	95.0	2024.0	3
19	79.0	92.0	70.0	85.0	2022.0	3
20	75.0	79.0	75.0	83.0	2026.0	2
21	73.0	86.0	66.0	98.0	2025.0	3
22	78.0	90.0	70.0	97.0	2026.0	3
23	60.0	91.0	75.0	93.0	2026.0	3
24	63.0	91.0	63.0	99.0	2025.0	3
25	NaN	81.0	65.0	97.0	2023.0	3
26	76.0	82.0	77.0	81.0	2025.0	2
27	76.0	84.0	79.0	89.0	2024.0	3
28	71.0	NaN	NaN	NaN	NaN	Na

In [22]:
```
df.dropna()
```

Out[22]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	placement_cou
0	75.0	90.0	80.0	79.0	2024.0	2
1	67.0	92.0	60.0	75.0	2023.0	2
2	62.0	75.0	61.0	75.0	2022.0	2
3	64.0	77.0	73.0	91.0	2022.0	3
4	76.0	85.0	79.0	75.0	2024.0	2
7	78.0	79.0	71.0	78.0	2025.0	2
8	62.0	81.0	80.0	83.0	2022.0	2
9	74.0	84.0	68.0	77.0	2026.0	2
10	76.0	75.0	200.0	80.0	2022.0	2
11	78.0	89.0	63.0	84.0	2022.0	2
12	66.0	76.0	68.0	86.0	2026.0	3
14	76.0	92.0	78.0	94.0	2023.0	3
15	62.0	79.0	60.0	76.0	2025.0	2
16	71.0	81.0	75.0	82.0	2026.0	2
17	73.0	81.0	61.0	98.0	2025.0	3
18	60.0	77.0	67.0	95.0	2024.0	3
19	79.0	92.0	70.0	85.0	2022.0	3
20	75.0	79.0	75.0	83.0	2026.0	2
21	73.0	86.0	66.0	98.0	2025.0	3
22	78.0	90.0	70.0	97.0	2026.0	3
23	60.0	91.0	75.0	93.0	2026.0	3
24	63.0	91.0	63.0	99.0	2025.0	3
26	76.0	82.0	77.0	81.0	2025.0	2
27	76.0	84.0	79.0	89.0	2024.0	3

In [23]:
```
df.dropna(axis=1)
```

Out[23]:

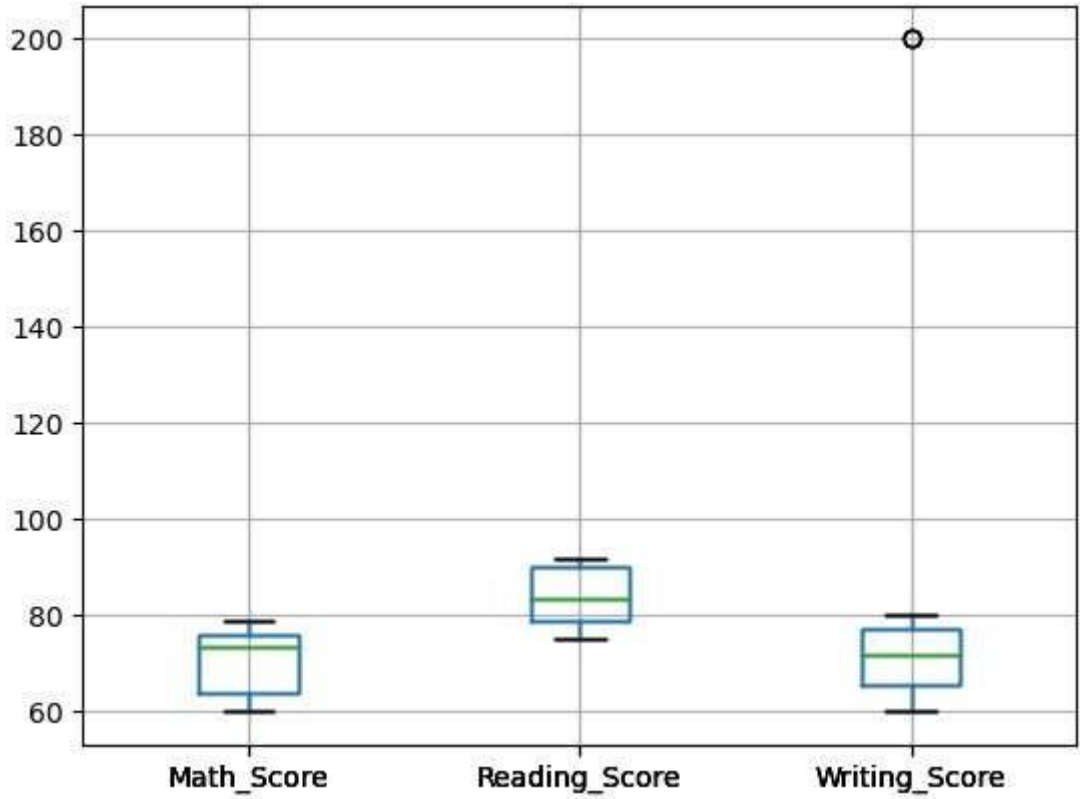
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

In [32]:
```
df=pd.read_csv('demods.csv')
df
```

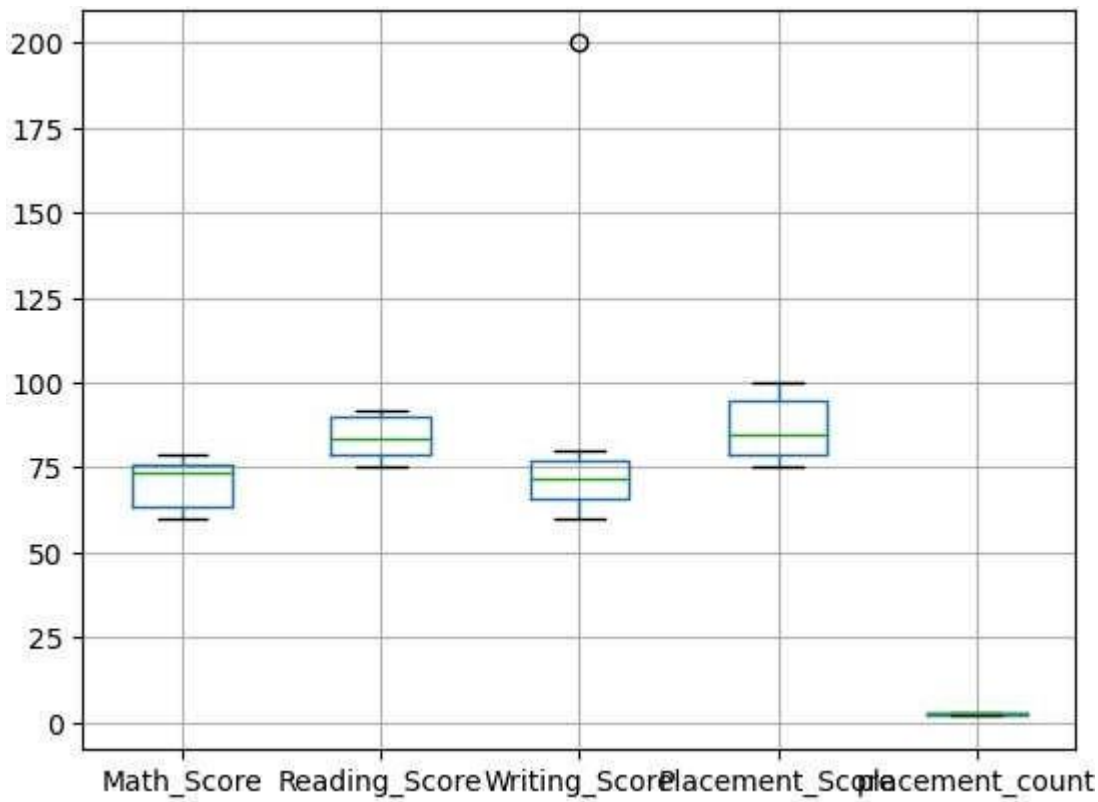
Out[32]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	placement_count
0	75	90	80	79	2
1	67	92	60	75	2
2	62	75	61	75	2
3	64	77	73	91	3
4	76	85	79	75	2
5	72	92	77	78	3
6	77	90	76	95	3
7	78	79	71	78	2
8	62	81	80	83	2
9	74	84	68	77	2
10	76	75	200	80	2
11	78	89	63	84	2
12	66	76	68	86	3
13	60	83	79	100	3
14	76	92	78	94	3
15	62	79	60	76	2
16	71	81	75	82	2
17	73	81	61	98	3
18	60	77	67	95	3
19	79	92	70	85	3
20	75	79	75	83	2
21	73	86	66	98	3
22	78	90	70	97	3
23	60	91	75	93	3
24	63	91	63	99	3
25	75	81	65	97	3
26	76	82	77	81	2
27	76	84	79	89	3

In [42]:
```
col=['Math_Score','Reading_Score','Writing_Score']
df.boxplot(col)
plt.show()
```



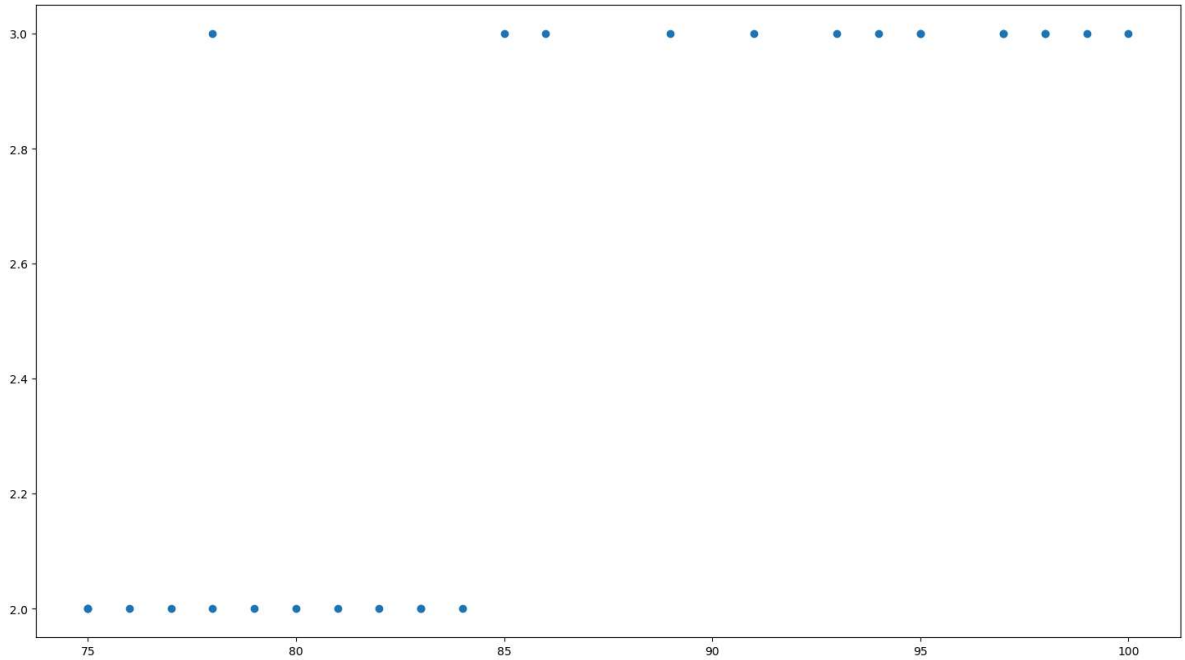
In [43]:
```
col=['Math_Score','Reading_Score','Writing_Score','Placement_Score','placement_
df.boxplot(col)
plt.show()
```



In [44]:
```
print(np.where(df['Math_Score']>70))
```

```
(array([ 0,  4,  5,  6,  7,  9, 10, 11, 14, 16, 17, 19, 20, 21, 22, 25, 26,
       27], dtype=int64),)
```

In [45]:
```
fig,ax=plt.subplots(figsize=(18,10))
ax.scatter(df['Placement_Score'],df['placement_count'])
plt.show()
```



In [48]:
```
z=np.abs(stats.zscore(df['Math_Score']))
print(z)
```

```
0     0.633511
1     0.589821
2     1.354403
3     1.048570
4     0.786427
5     0.174762
6     0.939344
7     1.092260
8     1.354403
9     0.480595
10    0.786427
11    1.092260
12    0.742737
13    1.660236
14    0.786427
15    1.354403
16    0.021845
17    0.327678
18    1.660236
19    1.245177
20    0.633511
21    0.327678
22    1.092260
23    1.660236
24    1.201486
25    0.633511
26    0.786427
27    0.786427
Name: Math_Score, dtype: float64
```

In [49]:
```
threshold=0.18
```

In [51]:
```
sample_outliers=np.where(z<threshold)
sample_outliers
```

Out[51]:
```
(array([ 5, 16], dtype=int64),)
```

In [52]:
```
sorted_rscore=sorted(df['Reading_Score'])
sorted_rscore
```

Out[52]:
```
[75,
 75,
 76,
 77,
 77,
 79,
 79,
 79,
 81,
 81,
 81,
 81,
 82,
 83,
 84,
 84,
 85,
 86,
 89,
 90,
 90,
 90,
 91,
 91,
 92,
 92,
 92,
 92]
```

In [53]:
```
q1=np.percentile(sorted_rscore,25)
q3=np.percentile(sorted_rscore,75)
print(q1,q3)
```

```
79.0 90.0
```

In [54]:
```
IQR=q3-q1
lwr_bound=q1-(1.5*IQR)
upr_bound=q3+(1.5*IQR)
print(lwr_bound,upr_bound)
```

```
62.5 106.5
```

Name : Devesh Y Mali
Roll No : 13228
Batch : B2