Assignment 2:Data Wrangling II Perform the following operations using Python on any open source dataset (eg. data.csv)

1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.
3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution. Reason and document your approach properly.

In [ ]:

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [26]:
```
pwd
```

Out[26]:
```
'C:\\Users\\Admin'
```

In [27]:
```
df=pd.read_csv("C:\\Users\\Admin\\Desktop\\StudentPerformance.csv")
```

In [28]:
```
print(df)
```

```
    Maths_Score  Reading_Score  Writing_Score  Placement_Score  \
0          70.0           93.0           61.0             82.0
1          77.0           84.0           65.0             88.0
2          69.0           84.0           68.0             93.0
3          72.0           81.0           73.0             91.0
4          78.0           95.0           73.0             96.0
5           NaN           94.0            NaN             80.0
6          69.0           86.0           79.0             91.0
7          76.0           92.0           61.0             79.0
8          79.0           81.0           77.0             80.0
9          65.0           85.0           78.0             76.0
10         66.0           78.0           69.0             94.0
11         75.0            NaN            NaN             90.0
12         75.0           81.0           74.0             88.0
13         94.0           75.0           80.0             83.0
14         69.0           79.0           79.0              NaN
15         60.0           88.0           61.0             81.0
16         79.0           84.0           75.0             76.0
17         68.0           80.0           66.0             89.0
18         65.0           85.0           68.0             92.0
19         63.0           75.0           75.0             84.0
20         71.0           78.0           67.0             83.0
21         67.0           89.0           95.0             78.0
22         74.0           77.0           72.0             81.0
23         64.0           76.0           67.0             82.0
24         61.0           87.0           63.0             98.0
25         76.0           91.0           60.0             88.0
26         67.0           93.0           76.0             90.0
27         93.0           88.0           99.0             91.0
28         62.0           79.0           67.0             86.0

    Club_Join_Date  Placement offer count
0             2020                      2
1             2019                      3
2             2020                      3
3             2019                      3
4             2020                      3
5             2020                      2
6             2018                      3
7             2019                      2
8             2018                      2
9             2020                      2
10            2018                      3
11            2018                      3
12            2019                      3
13            2020                      2
14            2020                      2
15            2018                      2
16            2018                      2
17            2020                      3
18            2020                      3
19            2018                      2
20            2018                      2
21            2019                      2
22            2020                      2
23            2018                      2
24            2020                      3
25            2019                      3
26            2019                      3
27            2019                      3
28            2020                      3
```

In [29]:
```
df
```

Out[29]:

	Maths_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement offer count
0	70.0	93.0	61.0	82.0	2020	2
1	77.0	84.0	65.0	88.0	2019	3
2	69.0	84.0	68.0	93.0	2020	3
3	72.0	81.0	73.0	91.0	2019	3
4	78.0	95.0	73.0	96.0	2020	3
5	NaN	94.0	NaN	80.0	2020	2
6	69.0	86.0	79.0	91.0	2018	3
7	76.0	92.0	61.0	79.0	2019	2
8	79.0	81.0	77.0	80.0	2018	2
9	65.0	85.0	78.0	76.0	2020	2
10	66.0	78.0	69.0	94.0	2018	3
11	75.0	NaN	NaN	90.0	2018	3
12	75.0	81.0	74.0	88.0	2019	3
13	94.0	75.0	80.0	83.0	2020	2
14	69.0	79.0	79.0	NaN	2020	2
15	60.0	88.0	61.0	81.0	2018	2
16	79.0	84.0	75.0	76.0	2018	2
17	68.0	80.0	66.0	89.0	2020	3
18	65.0	85.0	68.0	92.0	2020	3
19	63.0	75.0	75.0	84.0	2018	2
20	71.0	78.0	67.0	83.0	2018	2
21	67.0	89.0	95.0	78.0	2019	2
22	74.0	77.0	72.0	81.0	2020	2
23	64.0	76.0	67.0	82.0	2018	2
24	61.0	87.0	63.0	98.0	2020	3
25	76.0	91.0	60.0	88.0	2019	3
26	67.0	93.0	76.0	90.0	2019	3
27	93.0	88.0	99.0	91.0	2019	3
28	62.0	79.0	67.0	86.0	2020	3

In [30]:
```
df.isnull()
```

Out[30]:

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

In [31]:
```
df.isnull().sum()
```

Out[31]:
```
Maths_Score             1
Reading_Score           1
Writing_Score           2
Placement_Score         1
Club_Join_Date          0
Placement offer count   0
dtype: int64
```

In [32]:
```
df.notnull()
```

Out[32]:

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

In [33]:
```
series1=pd.notnull(df["Maths_Score"])
```

In [34]:
```
df[series1]
```

Out[34]:

	Maths_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement offer count
0	70.0	93.0	61.0	82.0	2020	2
1	77.0	84.0	65.0	88.0	2019	3
2	69.0	84.0	68.0	93.0	2020	3
3	72.0	81.0	73.0	91.0	2019	3
4	78.0	95.0	73.0	96.0	2020	3
6	69.0	86.0	79.0	91.0	2018	3
7	76.0	92.0	61.0	79.0	2019	2
8	79.0	81.0	77.0	80.0	2018	2
9	65.0	85.0	78.0	76.0	2020	2
10	66.0	78.0	69.0	94.0	2018	3
11	75.0	NaN	NaN	90.0	2018	3
12	75.0	81.0	74.0	88.0	2019	3
13	94.0	75.0	80.0	83.0	2020	2
14	69.0	79.0	79.0	NaN	2020	2
15	60.0	88.0	61.0	81.0	2018	2
16	79.0	84.0	75.0	76.0	2018	2
17	68.0	80.0	66.0	89.0	2020	3
18	65.0	85.0	68.0	92.0	2020	3
19	63.0	75.0	75.0	84.0	2018	2
20	71.0	78.0	67.0	83.0	2018	2
21	67.0	89.0	95.0	78.0	2019	2
22	74.0	77.0	72.0	81.0	2020	2
23	64.0	76.0	67.0	82.0	2018	2
24	61.0	87.0	63.0	98.0	2020	3
25	76.0	91.0	60.0	88.0	2019	3
26	67.0	93.0	76.0	90.0	2019	3
27	93.0	88.0	99.0	91.0	2019	3
28	62.0	79.0	67.0	86.0	2020	3

In [35]:
```
ndf=df
ndf.fillna(0)
```

Out[35]:

	Maths_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement offer count
0	70.0	93.0	61.0	82.0	2020	2
1	77.0	84.0	65.0	88.0	2019	3
2	69.0	84.0	68.0	93.0	2020	3
3	72.0	81.0	73.0	91.0	2019	3
4	78.0	95.0	73.0	96.0	2020	3
5	0.0	94.0	0.0	80.0	2020	2
6	69.0	86.0	79.0	91.0	2018	3
7	76.0	92.0	61.0	79.0	2019	2
8	79.0	81.0	77.0	80.0	2018	2
9	65.0	85.0	78.0	76.0	2020	2
10	66.0	78.0	69.0	94.0	2018	3
11	75.0	0.0	0.0	90.0	2018	3
12	75.0	81.0	74.0	88.0	2019	3
13	94.0	75.0	80.0	83.0	2020	2
14	69.0	79.0	79.0	0.0	2020	2
15	60.0	88.0	61.0	81.0	2018	2
16	79.0	84.0	75.0	76.0	2018	2
17	68.0	80.0	66.0	89.0	2020	3
18	65.0	85.0	68.0	92.0	2020	3
19	63.0	75.0	75.0	84.0	2018	2
20	71.0	78.0	67.0	83.0	2018	2
21	67.0	89.0	95.0	78.0	2019	2
22	74.0	77.0	72.0	81.0	2020	2
23	64.0	76.0	67.0	82.0	2018	2
24	61.0	87.0	63.0	98.0	2020	3
25	76.0	91.0	60.0	88.0	2019	3
26	67.0	93.0	76.0	90.0	2019	3
27	93.0	88.0	99.0	91.0	2019	3
28	62.0	79.0	67.0	86.0	2020	3

In [36]:
```
df['Maths_Score']=df['Maths_Score'].fillna(df['Maths_Score'].mean())
```

In [37]:
```
df
```

Out[37]:

	Maths_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement offer count
0	70.000000	93.0	61.0	82.0	2020	2
1	77.000000	84.0	65.0	88.0	2019	3
2	69.000000	84.0	68.0	93.0	2020	3
3	72.000000	81.0	73.0	91.0	2019	3
4	78.000000	95.0	73.0	96.0	2020	3
5	71.571429	94.0	NaN	80.0	2020	2
6	69.000000	86.0	79.0	91.0	2018	3
7	76.000000	92.0	61.0	79.0	2019	2
8	79.000000	81.0	77.0	80.0	2018	2
9	65.000000	85.0	78.0	76.0	2020	2
10	66.000000	78.0	69.0	94.0	2018	3
11	75.000000	NaN	NaN	90.0	2018	3
12	75.000000	81.0	74.0	88.0	2019	3
13	94.000000	75.0	80.0	83.0	2020	2
14	69.000000	79.0	79.0	NaN	2020	2
15	60.000000	88.0	61.0	81.0	2018	2
16	79.000000	84.0	75.0	76.0	2018	2
17	68.000000	80.0	66.0	89.0	2020	3
18	65.000000	85.0	68.0	92.0	2020	3
19	63.000000	75.0	75.0	84.0	2018	2
20	71.000000	78.0	67.0	83.0	2018	2
21	67.000000	89.0	95.0	78.0	2019	2
22	74.000000	77.0	72.0	81.0	2020	2
23	64.000000	76.0	67.0	82.0	2018	2
24	61.000000	87.0	63.0	98.0	2020	3
25	76.000000	91.0	60.0	88.0	2019	3
26	67.000000	93.0	76.0	90.0	2019	3
27	93.000000	88.0	99.0	91.0	2019	3
28	62.000000	79.0	67.0	86.0	2020	3

In [38]:
```
ndf.replace(to_replace=np.nan,value=-99)
```

Out[38]:

	Maths_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement offer count
0	70.000000	93.0	61.0	82.0	2020	2
1	77.000000	84.0	65.0	88.0	2019	3
2	69.000000	84.0	68.0	93.0	2020	3
3	72.000000	81.0	73.0	91.0	2019	3
4	78.000000	95.0	73.0	96.0	2020	3
5	71.571429	94.0	-99.0	80.0	2020	2
6	69.000000	86.0	79.0	91.0	2018	3
7	76.000000	92.0	61.0	79.0	2019	2
8	79.000000	81.0	77.0	80.0	2018	2
9	65.000000	85.0	78.0	76.0	2020	2
10	66.000000	78.0	69.0	94.0	2018	3
11	75.000000	-99.0	-99.0	90.0	2018	3
12	75.000000	81.0	74.0	88.0	2019	3
13	94.000000	75.0	80.0	83.0	2020	2
14	69.000000	79.0	79.0	-99.0	2020	2
15	60.000000	88.0	61.0	81.0	2018	2
16	79.000000	84.0	75.0	76.0	2018	2
17	68.000000	80.0	66.0	89.0	2020	3
18	65.000000	85.0	68.0	92.0	2020	3
19	63.000000	75.0	75.0	84.0	2018	2
20	71.000000	78.0	67.0	83.0	2018	2
21	67.000000	89.0	95.0	78.0	2019	2
22	74.000000	77.0	72.0	81.0	2020	2
23	64.000000	76.0	67.0	82.0	2018	2
24	61.000000	87.0	63.0	98.0	2020	3
25	76.000000	91.0	60.0	88.0	2019	3
26	67.000000	93.0	76.0	90.0	2019	3
27	93.000000	88.0	99.0	91.0	2019	3
28	62.000000	79.0	67.0	86.0	2020	3

In [39]:
```
ndf.dropna()
```

Out[39]:

	Maths_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join_Date	Placement offer count
0	70.0	93.0	61.0	82.0	2020	2
1	77.0	84.0	65.0	88.0	2019	3
2	69.0	84.0	68.0	93.0	2020	3
3	72.0	81.0	73.0	91.0	2019	3
4	78.0	95.0	73.0	96.0	2020	3
6	69.0	86.0	79.0	91.0	2018	3
7	76.0	92.0	61.0	79.0	2019	2
8	79.0	81.0	77.0	80.0	2018	2
9	65.0	85.0	78.0	76.0	2020	2
10	66.0	78.0	69.0	94.0	2018	3
12	75.0	81.0	74.0	88.0	2019	3
13	94.0	75.0	80.0	83.0	2020	2
15	60.0	88.0	61.0	81.0	2018	2
16	79.0	84.0	75.0	76.0	2018	2
17	68.0	80.0	66.0	89.0	2020	3
18	65.0	85.0	68.0	92.0	2020	3
19	63.0	75.0	75.0	84.0	2018	2
20	71.0	78.0	67.0	83.0	2018	2
21	67.0	89.0	95.0	78.0	2019	2
22	74.0	77.0	72.0	81.0	2020	2
23	64.0	76.0	67.0	82.0	2018	2
24	61.0	87.0	63.0	98.0	2020	3
25	76.0	91.0	60.0	88.0	2019	3
26	67.0	93.0	76.0	90.0	2019	3
27	93.0	88.0	99.0	91.0	2019	3
28	62.0	79.0	67.0	86.0	2020	3

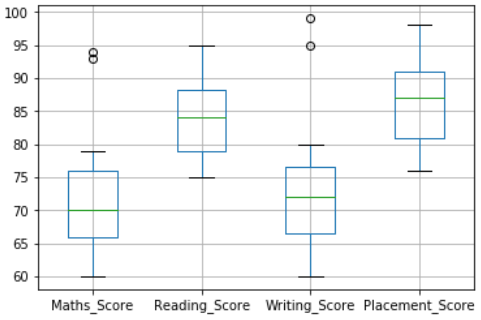
Module 2:Detection of Outlier 1.we can plot the outlier by using Boxplot, Scatterplot

1. Techniques of Detecting outlier a. Z-score b.IQR

Boxplot:

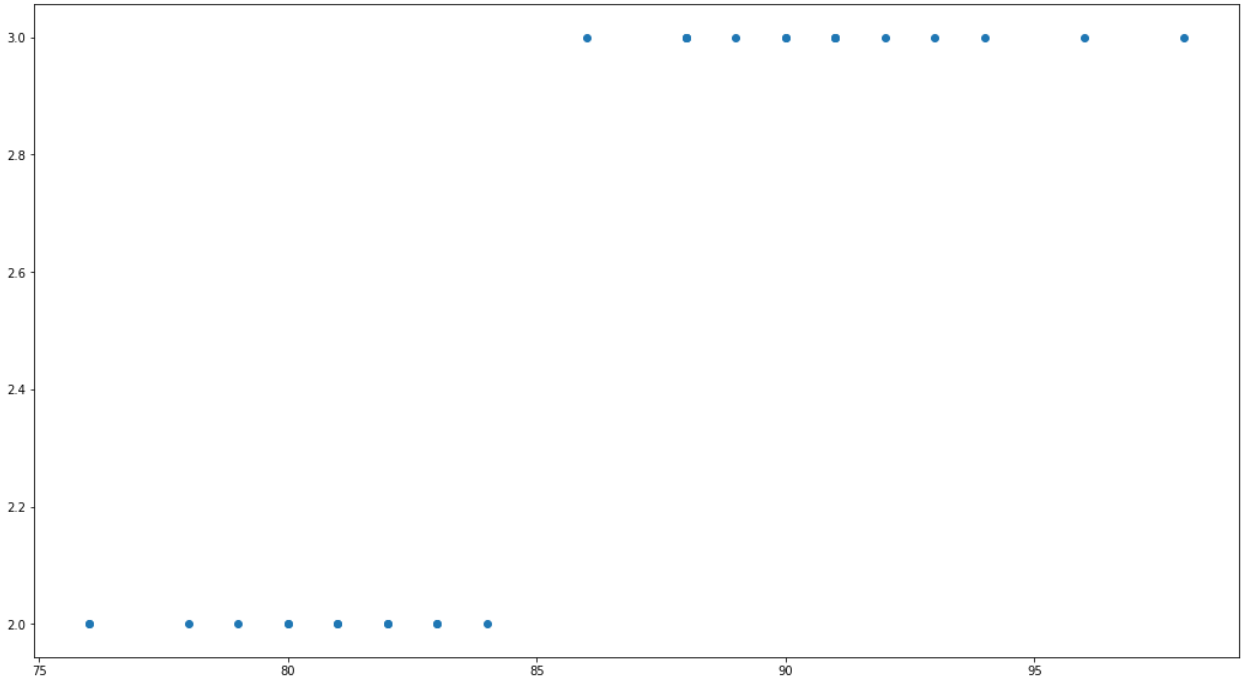
In [40]:
```
# Boxplot--> Summaries sample data using 25th, 50th,75th value
col=['Maths_Score','Reading_Score','Writing_Score','Placement_Score']
df.boxplot(col)
```

Out[40]: `<matplotlib.axes._subplots.AxesSubplot at 0x11240b0>`



Scatterplot:It is used when you have paired numerical data, or when your dependent variable has multiple values for each reading independent variable, or when trying to determine the relationship between the two variables. In the process of utilizing the scatter plot, one can also use it for outlier detection.

In [41]:
```
fig,ax=plt.subplots(figsize=(18,10))
ax.scatter(df['Placement_Score'],df['Placement offer count'])
plt.show()
```



In [ ]: