ASSIGNMENT 2 ¶

In [12]:

```
import pandas as pd
import numpy as np
df=pd.read_csv("StudentsPerformanceTest1.csv")
df.head()
```

Out[12]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.0	88.0	65.0	93.0	2019	3	Pune
1	Male	67.0	87.0	45.0	77.0	2019	2	Mumbai
2	Male	NaN	95.0	65.0	94.0	2020	3	Pune
3	Male	45.0	85.0	68.0	34.0	2018	1	NaN
4	Male	70.0	80.0	23.0	79.0	2019	2	Nashik

In [13]:

```
df.isnull().sum()
```

Out[13]:

```
Gender                   0
Math Score               3
Reading Score            1
Writing Score            1
Placement Score          1
Club Join Year           0
Placement Offer Count    0
Region                   6
dtype: int64
```

In [14]:

```
series = pd.isnull(df["Math Score"])
df[series]
```

Out[14]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
2	Male	NaN	95.0	65.0	94.0	2020	3	Pune
18	Male	NaN	92.0	72.0	80.0	2018	2	NaN
28	Male	NaN	76.0	67.0	91.0	2020	3	Nashik

In [15]:

```
df.notnull().sum()
```

Out[15]:

```
Gender                  30
Math Score              27
Reading Score           29
Writing Score           29
Placement Score         29
Club Join Year          30
Placement Offer Count   30
Region                  24
dtype: int64
```

In [29]:

```
series1 = pd.notnull(df["Math Score"])
df[series1].head()
```

Out[29]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
3	Male	45.000000	85.0	68.0	34.0	2018	1	NaN
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik

In [17]:

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

Out[17]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	0	72.0	88.0	65.0	93.0	2019	3	Pune
1	1	67.0	87.0	45.0	77.0	2019	2	Mumbai
2	1	NaN	95.0	65.0	94.0	2020	3	Pune
3	1	45.0	85.0	68.0	34.0	2018	1	NaN
4	1	70.0	80.0	23.0	79.0	2019	2	Nashik

In [18]:

```
missing_values = ["Na", "na"]
df=pd.read_csv("StudentsPerformanceTest1.csv", na_values = missing_values)
df.head()
```

Out[18]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.0	88.0	65.0	93.0	2019	3	Pune
1	Male	67.0	87.0	45.0	77.0	2019	2	Mumbai
2	Male	NaN	95.0	65.0	94.0	2020	3	Pune
3	Male	45.0	85.0	68.0	34.0	2018	1	NaN
4	Male	70.0	80.0	23.0	79.0	2019	2	Nashik

In [19]:

```
import pandas as pd
import numpy as np
df=pd.read_csv("StudentsPerformanceTest1.csv")
df.head()
```

Out[19]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.0	88.0	65.0	93.0	2019	3	Pune
1	Male	67.0	87.0	45.0	77.0	2019	2	Mumbai
2	Male	NaN	95.0	65.0	94.0	2020	3	Pune
3	Male	45.0	85.0	68.0	34.0	2018	1	NaN
4	Male	70.0	80.0	23.0	79.0	2019	2	Nashik

In [30]:

```
ndf=df
ndf.fillna(0).head()
```

Out[30]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
3	Male	45.000000	85.0	68.0	34.0	2018	1	0
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik

In [21]:

```
m_v=df["Math Score"].mean()
df['Math Score'].fillna(value=m_v,inplace=True)
df.head()
```

Out[21]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
3	Male	45.000000	85.0	68.0	34.0	2018	1	NaN
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik

In [22]:

```
ndf.replace(to_replace=np.nan,value=-99).head()
```

Out[22]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
3	Male	45.000000	85.0	68.0	34.0	2018	1	-99
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik

In [31]:

```
#delete null values usig dropna
ndf.dropna().head()
```

Out[31]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik
5	Female	75.000000	82.0	63.0	77.0	2020	2	Pune

In [32]:

```
#To Drop rows if all values in that row aremissing
ndf.dropna(how='all').head()
```

Out[32]:

	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
3	Male	45.000000	85.0	68.0	34.0	2018	1	NaN
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik

In [33]:

```
#To Drop columns with at least 1 null value.
ndf.dropna(axis=1).head()
```

Out[33]:

	Gender	Math Score	Club Join Year	Placement Offer Count
0	Female	72.000000	2019	3
1	Male	67.000000	2019	2
2	Male	68.777778	2020	3
3	Male	45.000000	2018	1
4	Male	70.000000	2019	2

In [34]:

```
#To drop rows with at least 1 null value inCSV file.making new data frame with dropped NA v
new_data=ndf.dropna(axis=0,how='any')
new_data.head()
```

Out[34]:

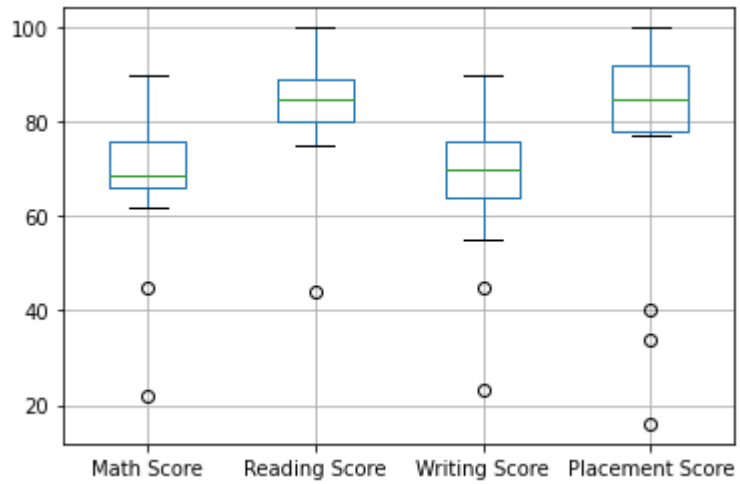
	Gender	Math Score	Reading Score	Writing Score	Placement Score	Club Join Year	Placement Offer Count	Region
0	Female	72.000000	88.0	65.0	93.0	2019	3	Pune
1	Male	67.000000	87.0	45.0	77.0	2019	2	Mumbai
2	Male	68.777778	95.0	65.0	94.0	2020	3	Pune
4	Male	70.000000	80.0	23.0	79.0	2019	2	Nashik
5	Female	75.000000	82.0	63.0	77.0	2020	2	Pune

In [27]:

```
#Select the columns for boxplot and draw theboxplot.
import matplotlib.pyplot as plt
df
col = ['Math Score','Reading Score','Writing Score','Placement Score']
df.boxplot(col)
```

Out[27]:

```
<AxesSubplot:>
```



In [28]:

```
import matplotlib.pyplot as plt
#Draw the scatter plot with placement scoreand placement offer count
fig,ax = plt.subplots(figsize=(10,7))
ax.scatter(df['Placement Score'],df['Placement Offer Count'])
plt.show()
```