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

In [8]:
```
df = pd.read_excel('StudentsPerformanceTest1.xlsx')
```

In [9]:
```
df
```

Out[9]:

	gender	math score	reading score	writing score	Placement Score	placement offer count	Region
0	female	72.0	72.0	74.0	78.0	1	Pune
1	female	69.0	90.0	88.0	NaN	2	NaN
2	female	90.0	95.0	93.0	74.0	2	Nashik
3	male	47.0	57.0	NaN	78.0	1	NaN
4	male	NaN	78.0	75.0	81.0	3	Pune
5	female	71.0	NaN	78.0	70.0	4	NaN
6	male	12.0	44.0	52.0	12.0	2	Nashik
7	male	NaN	65.0	67.0	49.0	1	Pune
8	male	5.0	77.0	89.0	55.0	0	NaN

In [12]:
```
df.isna()
```

Out[12]:

	gender	math score	reading score	writing score	Placement Score	placement offer count	Region
0	False	False	False	False	False	False	False
1	False	False	False	False	True	False	True
2	False	False	False	False	False	False	False
3	False	False	False	True	False	False	True
4	False	True	False	False	False	False	False
5	False	False	True	False	False	False	True
6	False	False	False	False	False	False	False
7	False	True	False	False	False	False	False
8	False	False	False	False	False	False	True

In [13]:
```
df.isnull()
```

Out[13]:

	gender	math score	reading score	writing score	Placement Score	placement offer count	Region
0	False	False	False	False	False	False	False
1	False	False	False	False	True	False	True
2	False	False	False	False	False	False	False
3	False	False	False	True	False	False	True
4	False	True	False	False	False	False	False
5	False	False	True	False	False	False	True
6	False	False	False	False	False	False	False
7	False	True	False	False	False	False	False
8	False	False	False	False	False	False	True

In [14]:
```
df.isna().sum()
```

Out[14]:
```
gender                   0
math score               2
reading score            1
writing score            1
Placement Score          1
placement offer count    0
Region                   4
dtype: int64
```

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
gender                   0
math score               2
reading score            1
writing score            1
Placement Score          1
placement offer count    0
Region                   4
dtype: int64
```

In [16]:
```
# implacing missing values with the mean
mean_math_score = df['math score'].mean()
df['math score'].fillna(mean_math_score, inplace=True)
```

In [17]:
```
df.isnull().sum()
```

Out[17]:
```
gender                  0
math score              0
reading score           1
writing score           1
Placement Score         1
placement offer count   0
Region                  4
dtype: int64
```

In [18]:
```
mean_reading_score = df['reading score'].mean()
df['reading score'].fillna(mean_reading_score, inplace=True)
```

In [20]:
```
mean_writing_score = df['writing score'].mean()
df['writing score'].fillna(mean_writing_score, inplace=True)
```

In [30]:
```
mean_placement_score = df['Placement Score'].mean()
df['Placement Score'].fillna(mean_placement_score, inplace=True)
```

In [24]:
```
df.isnull().sum()
```

Out[24]:
```
gender                  0
math score              0
reading score           0
writing score           0
Placement Score         0
placement offer count   0
Region                  4
dtype: int64
```

In [28]:
```
df.describe()
```

Out[28]:

	math score	reading score	writing score	Placement Score	placement offer count
count	9.000000	9.000000	9.0000	9.000000	9.000000
mean	52.285714	72.250000	77.0000	62.125000	1.777778
std	28.123452	15.698328	12.5499	21.791268	1.201850
min	5.000000	44.000000	52.0000	12.000000	0.000000
25%	47.000000	65.000000	74.0000	55.000000	1.000000
50%	52.285714	72.250000	77.0000	70.000000	2.000000
75%	71.000000	78.000000	88.0000	78.000000	2.000000
max	90.000000	95.000000	93.0000	81.000000	4.000000

In [32]:
```
# dealing with missing string values by dropping
df.drop('Region',axis=1,inplace=True)
```

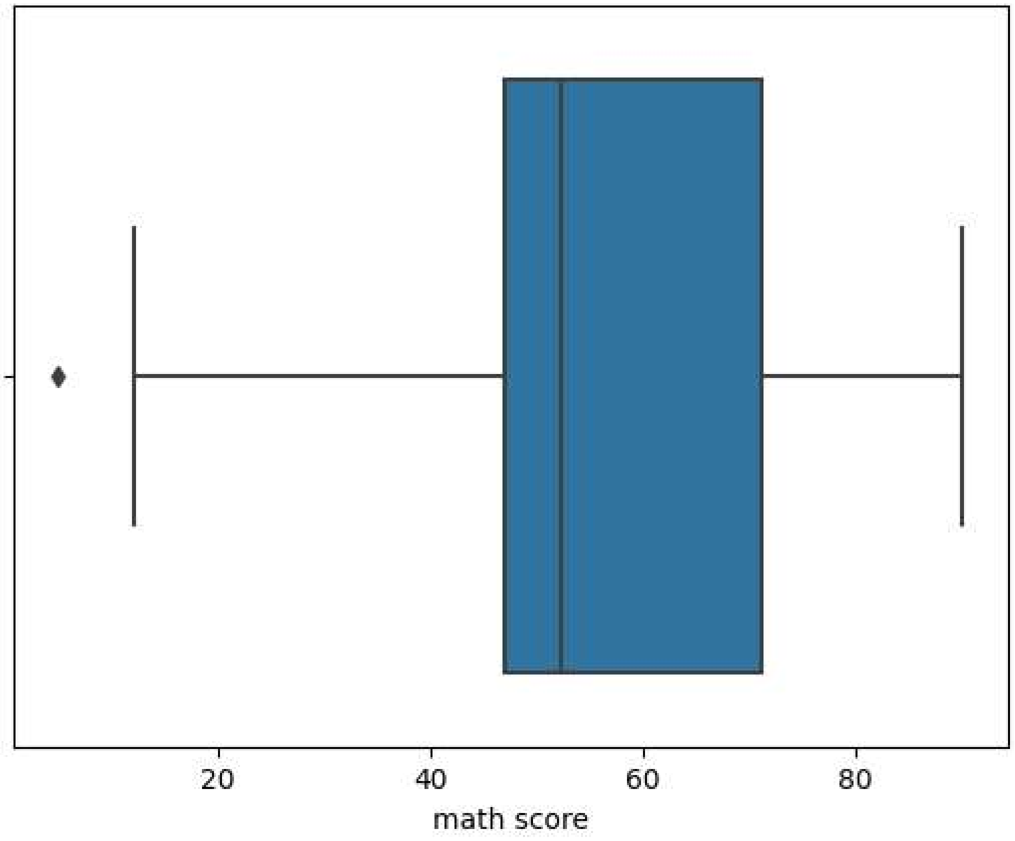
In [33]:
```
df
```

Out[33]:

	gender	math score	reading score	writing score	Placement Score	placement offer count
0	female	72.000000	72.00	74.0	78.000	1
1	female	69.000000	90.00	88.0	62.125	2
2	female	90.000000	95.00	93.0	74.000	2
3	male	47.000000	57.00	77.0	78.000	1
4	male	52.285714	78.00	75.0	81.000	3
5	female	71.000000	72.25	78.0	70.000	4
6	male	12.000000	44.00	52.0	12.000	2
7	male	52.285714	65.00	67.0	49.000	1
8	male	5.000000	77.00	89.0	55.000	0

In [34]:
```
# dealing with outliers
sns.boxplot(x=df['math score'])
```

Out[34]:
```
<AxesSubplot: xlabel='math score'>
```



In [48]:

```
# removing outliers that are beyond 1.5 times the interquartile range
Q1 = df['math score'].quantile(0.25)
Q3 = df['math score'].quantile(0.75)
IQR = Q3 - Q1
threshold = 1.5 * IQR

upper = Q3 + threshold
lower = Q1 - threshold

upper_array = np.array(df['math score'] >= upper)
lower_array = np.array(df['math score'] >= lower)

outliers = df[(df['math score'] > lower) & (df['math score'] < upper)]
```

In [49]:

```
outliers
```

Out[49]:

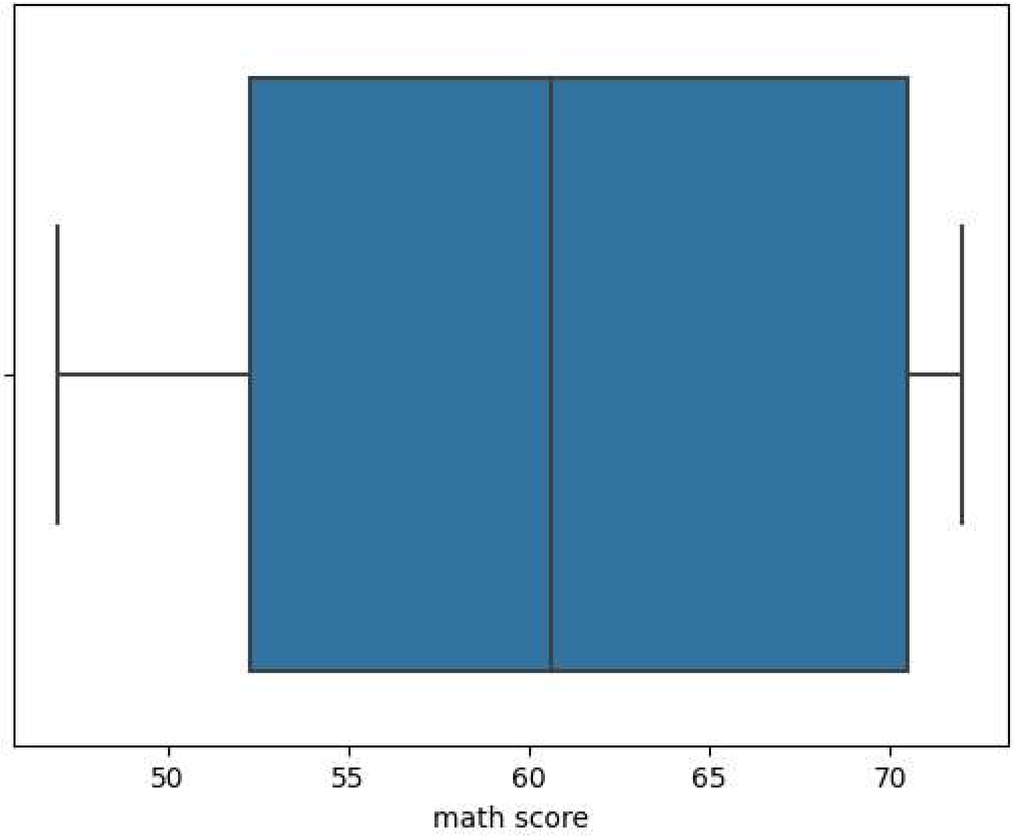
	gender	math score	reading score	writing score	Placement Score	placement offer count
0	female	72.000000	72.00	74.0	78.000	1
1	female	69.000000	90.00	88.0	62.125	2
3	male	47.000000	57.00	77.0	78.000	1
4	male	52.285714	78.00	75.0	81.000	3
5	female	71.000000	72.25	78.0	70.000	4
7	male	52.285714	65.00	67.0	49.000	1

In [50]:

```
df = df[(df['math score'] <= upper) & (df['math score'] >= lower)]
```

In [52]:
```
sns.boxplot(x = df['math score'])
```

Out[52]:
```
<AxesSubplot: xlabel='math score'>
```



In [ ]: