EXPERIMENT – 2

OUTLIER DETECTION AND REMOVAL

AIM

To read the given data and perform outlier detection and removal of data to a file.

ALGORITHM

STEP 1

Read the given Data

STEP 2

Get the information about the data

STEP 3

Remove the null values from the data

STEP 4

Get that data and do the following operations to detect outliers

boxplot()

IQR(interquartile range)

STEP 5:

For IQR

Q1=df2.quantile(0.25)

Q2=df2.quantile(0.75)

IQR=Q2-Q1

In [2]:
```
import pandas as pd
```

In [3]:
```
df=pd.read_csv("C:\\Users\\banga\\gitremoterepo\\Ex-02_DS_Outlier\\weight.csv")
```

In [4]:
```
df
```

Out[4]:

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

10000 rows × 3 columns

In [5]:
```
df.drop("Gender",axis=1,inplace=True)
```

In [6]:
```
df
```

Out[6]:

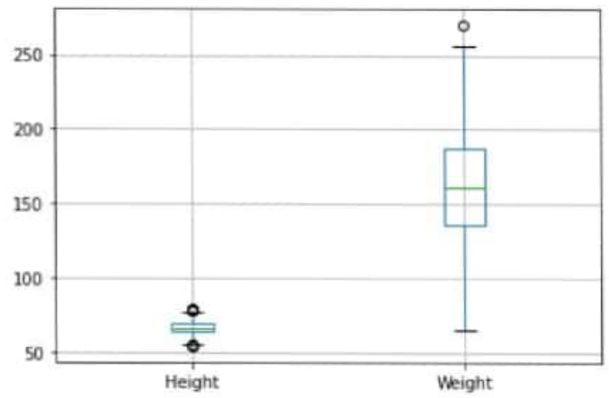
	Height	Weight
0	73.847017	241.893563
1	68.781904	162.310473
2	74.110105	212.740856
3	71.730978	220.042470
4	69.881796	206.349801
...	...	...
9995	66.172652	136.777454
9996	67.067155	170.867906
9997	63.867992	128.475319
9998	69.034243	163.852461
9999	61.944246	113.649103

10000 rows × 2 columns

In [7]:
```
# df=df.drop("Gender",axis=1,inplace=True)
```

In [8]:
```
df.boxplot()
```

Out[8]: <AxesSubplot:>



In [9]:
```
from scipy import stats
```

In [10]:
```
import numpy as np
```

In [11]:
```
z=np.abs(stats.zscore(df))
```

In [12]:
```
z
```

Out[12]:
```
array([[1.94406149, 2.50579697],
       [0.62753668, 0.02710064],
       [2.01244346, 1.59780623],
       ...,
       [0.64968792, 1.02672965],
       [0.69312469, 0.07512745],
       [1.14970831, 1.48850724]])
```

In [13]:
```
df
```

Out[13]:

	Height	Weight
0	73.847017	241.893563
1	68.781904	162.310473
2	74.110105	212.740856
3	71.730978	220.042470
4	69.881796	206.349801
...	...	...
9995	66.172652	136.777454
9996	67.067155	170.867906
9997	63.867992	128.475319
9998	69.034243	163.852461
9999	61.944246	113.649103

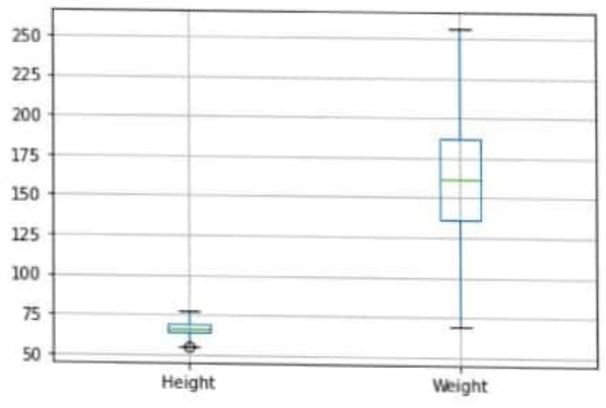
10000 rows × 2 columns

In [14]:
```
df1=df.copy()
```

In [15]:
```
df1=df1[(z<3).all(axis=1)]
```

In [16]:
```
df1.boxplot()
```

Out[16]: <AxesSubplot:>



In [17]:
```
df1
```

Out[17]:

	Height	Weight
0	73.847017	241.893563
1	68.781904	162.310473
2	74.110105	212.740856
3	71.730978	220.042470
4	69.881796	206.349801
...	...	...
9995	66.172652	136.777454
9996	67.067155	170.867906
9997	63.867992	128.475319
9998	69.034243	163.852461
9999	61.944246	113.649103

9993 rows × 2 columns

In [18]:
```
#interquartile method
df2=df.copy()
```

In [19]:
```
q1=df2.quantile(0.25)
```

In [20]:
```
q3=df2.quantile(0.75)
```

In [21]:
```
IQR=q3-q1
IQR
```

Out[21]:
```
Height     5.668641
Weight    51.351474
dtype: float64
```

In [22]:
```
IQR.Height
```

Out[22]:
```
5.668641245615746
```

In [23]:
```
df2_new=df2[((df2>=q1-1.5*IQR)&(df2<=q3+1.5*IQR)).all(axis=1)]
```

In [24]:
```
df2
```

Out[24]:

	Height	Weight
0	73.847017	241.893563
1	68.781904	162.310473
2	74.110105	212.740856
3	71.730978	220.042470
4	69.881796	206.349801
...	...	...
9995	66.172652	136.777454
9996	67.067155	170.867906
9997	63.867992	128.475319
9998	69.034243	163.852461
9999	61.944246	113.649103

10000 rows × 2 columns

In [ ]:
```

```