Assignment_2

March 6, 2023

[22]:
```
import pandas as pd
```

[23]:
```
df = pd.read_csv("data.csv")
```

[24]:
```
df
```

[24]:

	Math_Score	Reading	Writing_Score	Placement_Score	Club_Join_date	\
0	64	82	77	88	2020	
1	66	78	75	80	2019	
2	80	77	62	90	2021	
3	72	80	61	91	2021	
4	150	90	62	91	2018	
5	74	76	64	76	2019	
6	63	75	61	83	2020	
7	67	76	63	90	2020	
8	64	84	78	86	2020	
9	63	91	70	99	2018	
10	64	85	71	93	2019	
11	68	93	77	92	2020	
12	75	88	62	89	2020	
13	68	79	76	89	2020	
14	71	94	75	92	2019	
15	69	88	67	91	2018	
16	60	80	74	75	2021	
17	70	100	63	93	2020	
18	70	79	71	92	2021	
19	79	90	89	78	2020	
20	75	86	68	96	2020	
21	62	79	78	82	2019	
22	79	78	75	93	2018	
23	61	88	70	99	2020	
24	66	85	60	93	2020	
25	63	88	64	80	2020	
26	79	91	75	87	2020	
27	60	77	71	92	2018	
28	74	90	72	94	2021	
29	70	90	77	92	2018	

	Placement_Offer_Count
0	NaN
1	2.0
2	3.0
3	3.0
4	3.0
5	2.0
6	2.0
7	3.0
8	3.0
9	3.0
10	3.0
11	3.0
12	3.0
13	3.0
14	3.0
15	3.0
16	2.0
17	3.0
18	3.0
19	2.0
20	3.0
21	2.0
22	3.0
23	3.0
24	1.0
25	2.0
26	3.0
27	3.0
28	3.0
29	3.0

[25]:
```
DF = pd.DataFrame(df)
DF
```

[25]:

	Math_Score	Reading	Writing_Score	Placement_Score	Club_Join_date	\
0	64	82	77	88	2020	
1	66	78	75	80	2019	
2	80	77	62	90	2021	
3	72	80	61	91	2021	
4	150	90	62	91	2018	
5	74	76	64	76	2019	
6	63	75	61	83	2020	
7	67	76	63	90	2020	
8	64	84	78	86	2020	
9	63	91	70	99	2018	

10	64	85	71	93	2019
11	68	93	77	92	2020
12	75	88	62	89	2020
13	68	79	76	89	2020
14	71	94	75	92	2019
15	69	88	67	91	2018
16	60	80	74	75	2021
17	70	100	63	93	2020
18	70	79	71	92	2021
19	79	90	89	78	2020
20	75	86	68	96	2020
21	62	79	78	82	2019
22	79	78	75	93	2018
23	61	88	70	99	2020
24	66	85	60	93	2020
25	63	88	64	80	2020
26	79	91	75	87	2020
27	60	77	71	92	2018
28	74	90	72	94	2021
29	70	90	77	92	2018

	Placement_Offer_Count
0	NaN
1	2.0
2	3.0
3	3.0
4	3.0
5	2.0
6	2.0
7	3.0
8	3.0
9	3.0
10	3.0
11	3.0
12	3.0
13	3.0
14	3.0
15	3.0
16	2.0
17	3.0
18	3.0
19	2.0
20	3.0
21	2.0
22	3.0
23	3.0
24	1.0

```
25                    2.0
26                    3.0
27                    3.0
28                    3.0
29                    3.0
```

[26]:
```
DF.isnull()
```

[26]:
```
    Math_Score  Reading  Writing_Score  Placement_Score  Club_Join_date  \
0        False    False          False            False           False
1        False    False          False            False           False
2        False    False          False            False           False
3        False    False          False            False           False
4        False    False          False            False           False
5        False    False          False            False           False
6        False    False          False            False           False
7        False    False          False            False           False
8        False    False          False            False           False
9        False    False          False            False           False
10       False    False          False            False           False
11       False    False          False            False           False
12       False    False          False            False           False
13       False    False          False            False           False
14       False    False          False            False           False
15       False    False          False            False           False
16       False    False          False            False           False
17       False    False          False            False           False
18       False    False          False            False           False
19       False    False          False            False           False
20       False    False          False            False           False
21       False    False          False            False           False
22       False    False          False            False           False
23       False    False          False            False           False
24       False    False          False            False           False
25       False    False          False            False           False
26       False    False          False            False           False
27       False    False          False            False           False
28       False    False          False            False           False
29       False    False          False            False           False

    Placement_Offer_Count
0                    True
1                   False
2                   False
3                   False
4                   False
5                   False
```

```
6                   False
7                   False
8                   False
9                   False
10                  False
11                  False
12                  False
13                  False
14                  False
15                  False
16                  False
17                  False
18                  False
19                  False
20                  False
21                  False
22                  False
23                  False
24                  False
25                  False
26                  False
27                  False
28                  False
29                  False
```

[27]:
```
DF.isna().sum()
```

[27]:
```
Math_Score              0
Reading                 0
Writing_Score           0
Placement_Score         0
Club_Join_date          0
Placement_Offer_Count   1
dtype: int64
```

[28]:
```
DF_r = DF.fillna(0.0)
```

[29]:
```
DF_r
```

[29]:

	Math_Score	Reading	Writing_Score	Placement_Score	Club_Join_date	\
0	64	82	77	88	2020	
1	66	78	75	80	2019	
2	80	77	62	90	2021	
3	72	80	61	91	2021	
4	150	90	62	91	2018	
5	74	76	64	76	2019	
6	63	75	61	83	2020	

7	67	76	63	90	2020
8	64	84	78	86	2020
9	63	91	70	99	2018
10	64	85	71	93	2019
11	68	93	77	92	2020
12	75	88	62	89	2020
13	68	79	76	89	2020
14	71	94	75	92	2019
15	69	88	67	91	2018
16	60	80	74	75	2021
17	70	100	63	93	2020
18	70	79	71	92	2021
19	79	90	89	78	2020
20	75	86	68	96	2020
21	62	79	78	82	2019
22	79	78	75	93	2018
23	61	88	70	99	2020
24	66	85	60	93	2020
25	63	88	64	80	2020
26	79	91	75	87	2020
27	60	77	71	92	2018
28	74	90	72	94	2021
29	70	90	77	92	2018

	Placement_Offer_Count
0	0.0
1	2.0
2	3.0
3	3.0
4	3.0
5	2.0
6	2.0
7	3.0
8	3.0
9	3.0
10	3.0
11	3.0
12	3.0
13	3.0
14	3.0
15	3.0
16	2.0
17	3.0
18	3.0
19	2.0
20	3.0
21	2.0

```
22                    3.0
23                    3.0
24                    1.0
25                    2.0
26                    3.0
27                    3.0
28                    3.0
29                    3.0
```

[30]:
```
DF_r.isnull().sum()
```

[30]:
```
Math_Score              0
Reading                 0
Writing_Score           0
Placement_Score         0
Club_Join_date          0
Placement_Offer_Count   0
dtype: int64
```

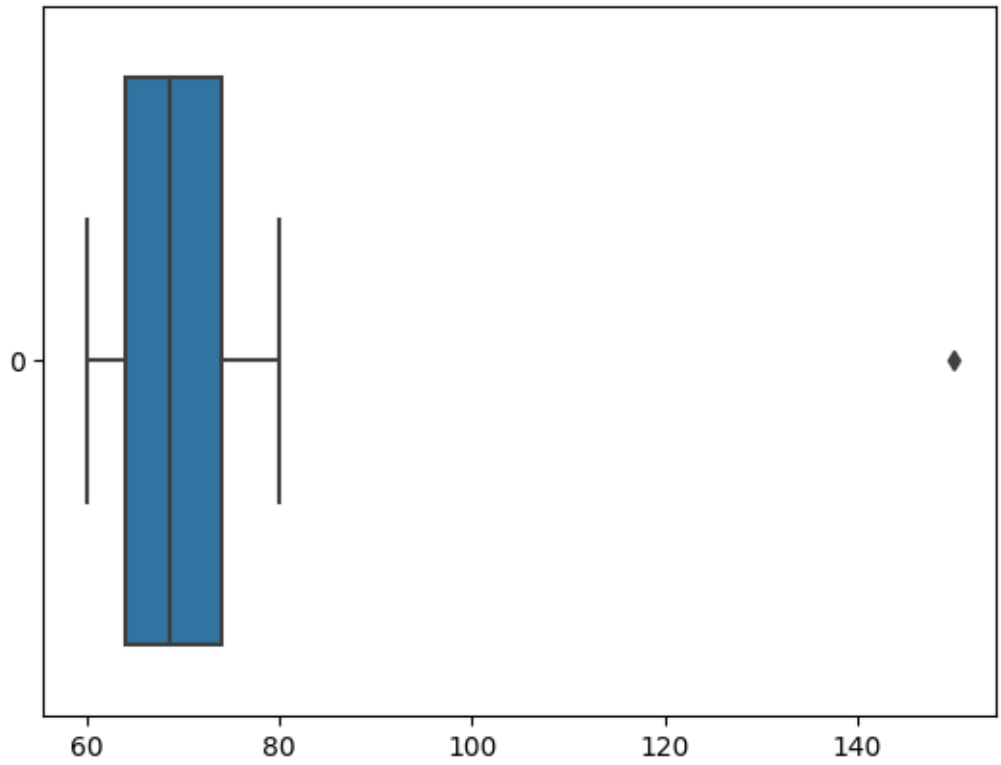
[31]:
```
DF_r.describe()
```

[31]:
```
       Math_Score     Reading  Writing_Score  Placement_Score  Club_Join_date  \
count   30.000000   30.000000      30.000000        30.000000       30.000000
mean    71.533333   84.566667      70.266667        88.866667     2019.600000
std     15.984331    6.510910       7.021805         6.279322        1.003442
min     60.000000   75.000000      60.000000        75.000000     2018.000000
25%     64.000000   79.000000      63.250000        86.250000     2019.000000
50%     68.500000   85.000000      71.000000        91.000000     2020.000000
75%     74.000000   90.000000      75.000000        92.750000     2020.000000
max    150.000000  100.000000      89.000000        99.000000     2021.000000

       Placement_Offer_Count
count              30.000000
mean                2.600000
std                 0.723974
min                 0.000000
25%                 2.000000
50%                 3.000000
75%                 3.000000
max                 3.000000
```

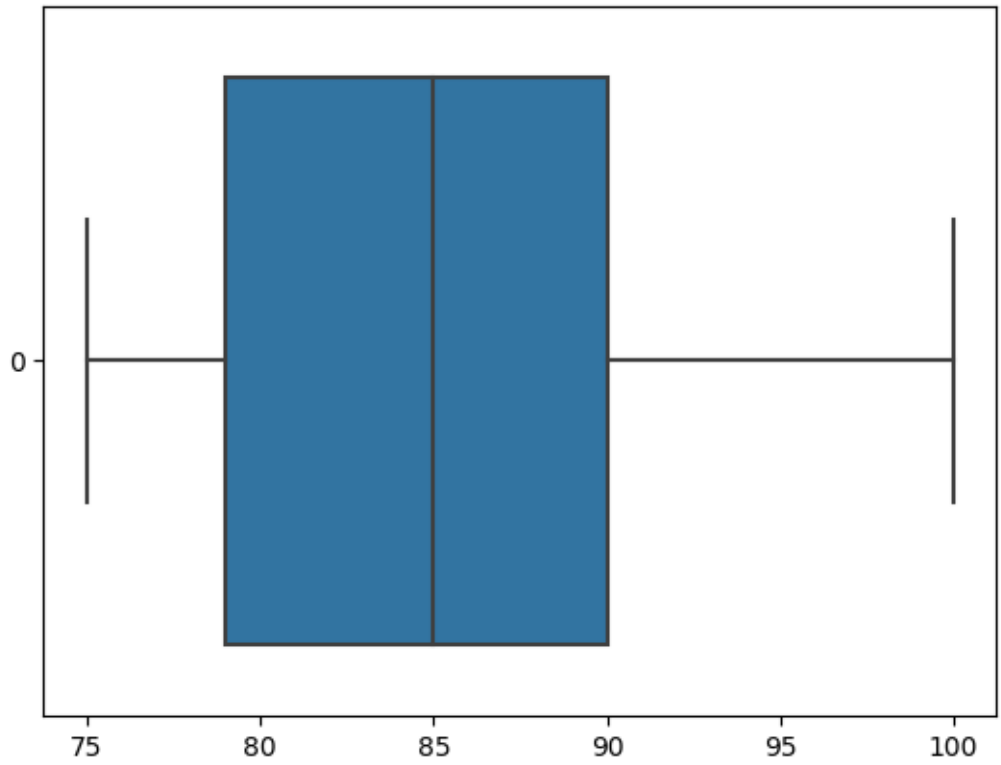
[50]:
```
import seaborn as sns
sns.boxplot(DF_r['Math_Score'],orient='h')
```

[50]:
```
<Axes: >
```



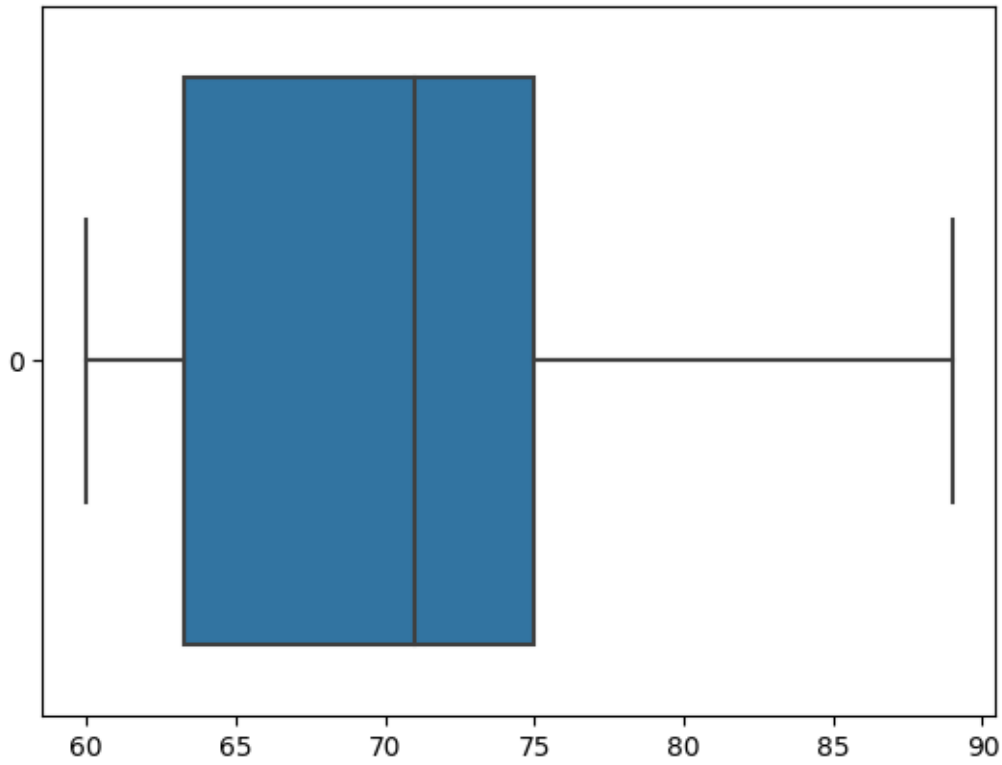
[51]:
```
sns.boxplot(DF_r['Reading'],orient='h')
```

[51]: <Axes: >



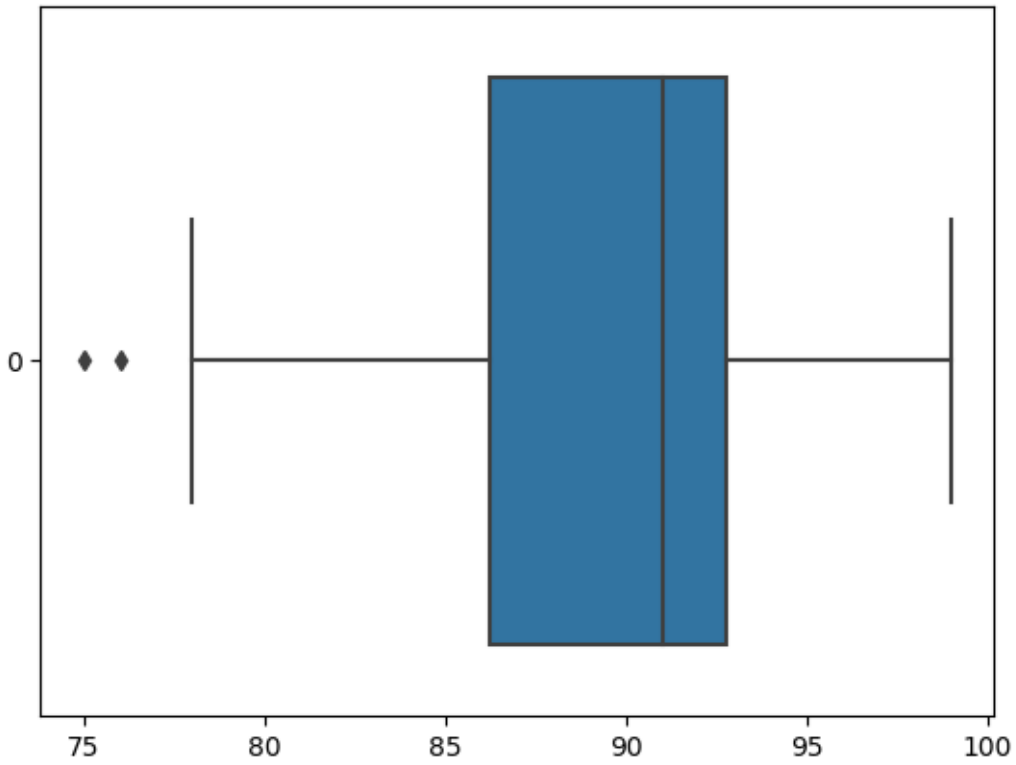
[52]:
```
sns.boxplot(DF_r['Writing_Score'],orient='h')
```

[52]: <Axes: >



[53]:
```
sns.boxplot(DF_r['Placement_Score'],orient='h')
```

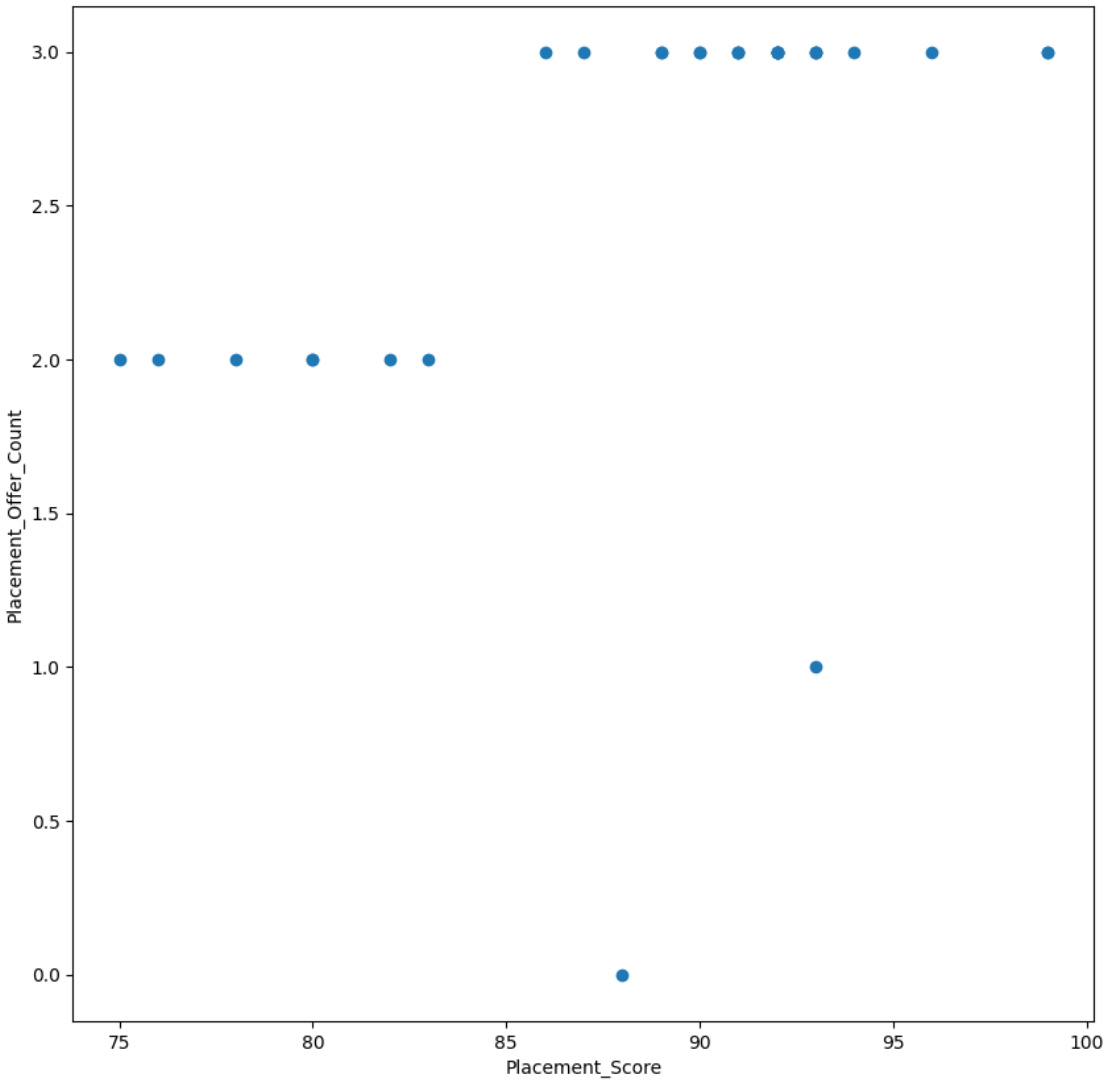
[53]: <Axes: >



[57]:
```
import  matplotlib.pyplot as plt
# Scatter plot
fig, ax = plt.subplots(figsize = (10,10))
ax.scatter(DF_r['Placement_Score'], DF_r['Placement_Offer_Count'])

# x-axis label
ax.set_xlabel('Placement_Score')

# y-axis label
ax.set_ylabel('Placement_Offer_Count')
plt.show()
```



[64]:
```
import numpy as np
print(np.where(DF_r['Math_Score']>80))
print(np.where(DF_r['Math_Score']<60))
```

```
(array([4]),)
(array([], dtype=int64),)
```

[66]:
```
import numpy as np
print(np.where(DF_r['Reading']>95))
print(np.where(DF_r['Reading']<75))
```

```
(array([17]),)
(array([], dtype=int64),)
```

[67]:
```
import numpy as np
print(np.where(DF_r['Writing_Score']>80))
print(np.where(DF_r['Writing_Score']<60))
```

```
(array([19]),)
(array([], dtype=int64),)
```

[68]:
```
import numpy as np
print(np.where(DF_r['Placement_Score']>100))
print(np.where(DF_r['Placement_Score']<75))
```

```
(array([], dtype=int64),)
(array([], dtype=int64),)
```

[83]:
```
from scipy import stats
x = np.abs(stats.zscore(DF_r['Math_Score']))
print(x)

threshold = 0.99
outliers_mscore = np.where(x>threshold)
```

```
0     0.479352
1     0.352090
2     0.538740
3     0.029694
4     4.992894
5     0.156956
6     0.542983
7     0.288459
8     0.479352
9     0.542983
10    0.479352
11    0.224829
12    0.220587
13    0.224829
14    0.033936
15    0.161198
16    0.733875
17    0.097567
18    0.097567
19    0.475110
20    0.220587
21    0.606613
22    0.475110
23    0.670244
24    0.352090
25    0.542983
26    0.475110
```

```
27    0.733875
28    0.156956
29    0.097567
Name: Math_Score, dtype: float64
```

[82]:
```
from scipy import stats
x = np.abs(stats.zscore(DF_r['Reading']))
print(x)

threshold = 1.99
outliers_rscore = np.where(x>threshold)
```

```
0     0.400949
1     1.025805
2     1.182019
3     0.713377
4     0.848763
5     1.338233
6     1.494447
7     1.338233
8     0.088521
9     1.004977
10    0.067693
11    1.317405
12    0.536335
13    0.869591
14    1.473619
15    0.536335
16    0.713377
17    2.410902
18    0.869591
19    0.848763
20    0.223907
21    0.869591
22    1.025805
23    0.536335
24    0.067693
25    0.536335
26    1.004977
27    1.182019
28    0.848763
29    0.848763
Name: Reading, dtype: float64
```

[81]:
```
from scipy import stats
x = np.abs(stats.zscore(DF_r['Writing_Score']))
print(x)
```

```
threshold = 2.0
outliers_wscore = np.where(x>threshold)
```

```
0     0.975311
1     0.685614
2     1.197411
3     1.342259
4     1.197411
5     0.907715
6     1.342259
7     1.052563
8     1.120159
9     0.038626
10    0.106222
11    0.975311
12    1.197411
13    0.830463
14    0.685614
15    0.473171
16    0.540766
17    1.052563
18    0.106222
19    2.713488
20    0.328322
21    1.120159
22    0.685614
23    0.038626
24    1.487107
25    0.907715
26    0.685614
27    0.106222
28    0.251070
29    0.975311
Name: Writing_Score, dtype: float64
```

[74]:
```
from scipy import stats
x = np.abs(stats.zscore(DF_r['Placement_Score']))
print(x)
```

```
0     0.140379
1     1.436181
2     0.183572
3     0.345547
4     0.345547
5     2.084083
6     0.950255
7     0.183572
```

```
8     0.464329
9     1.641350
10    0.669498
11    0.507523
12    0.021597
13    0.021597
14    0.507523
15    0.345547
16    2.246058
17    0.669498
18    0.507523
19    1.760132
20    1.155424
21    1.112231
22    0.669498
23    1.641350
24    0.669498
25    1.436181
26    0.302354
27    0.507523
28    0.831473
29    0.507523
Name: Placement_Score, dtype: float64
```

[94]:
```
print("--outliers--")
print(outliers_mscore + outliers_rscore + outliers_wscore)
```

```
--outliers--
(array([4]), array([17]), array([19]))
```

[100]:
```
newDF = DF_r

for i in outliers_mscore:
    newDF.drop(i,inplace=True)

for i in outliers_rscore:
    newDF.drop(i,inplace=True)

for i in outliers_wscore:
    newDF.drop(i,inplace=True)

newDF
```

[100]:

	Math_Score	Reading	Writing_Score	Placement_Score	Club_Join_date	\
0	64	82	77	88	2020	
1	66	78	75	80	2019	
2	80	77	62	90	2021	

3	72	80	61	91	2021
5	74	76	64	76	2019
6	63	75	61	83	2020
7	67	76	63	90	2020
8	64	84	78	86	2020
9	63	91	70	99	2018
10	64	85	71	93	2019
11	68	93	77	92	2020
12	75	88	62	89	2020
13	68	79	76	89	2020
14	71	94	75	92	2019
15	69	88	67	91	2018
16	60	80	74	75	2021
18	70	79	71	92	2021
20	75	86	68	96	2020
21	62	79	78	82	2019
22	79	78	75	93	2018
23	61	88	70	99	2020
24	66	85	60	93	2020
25	63	88	64	80	2020
26	79	91	75	87	2020
27	60	77	71	92	2018
28	74	90	72	94	2021
29	70	90	77	92	2018

	Placement_Offer_Count
0	0.0
1	2.0
2	3.0
3	3.0
5	2.0
6	2.0
7	3.0
8	3.0
9	3.0
10	3.0
11	3.0
12	3.0
13	3.0
14	3.0
15	3.0
16	2.0
18	3.0
20	3.0
21	2.0
22	3.0
23	3.0

```
24                    1.0
25                    2.0
26                    3.0
27                    3.0
28                    3.0
29                    3.0
```

[111]:
```
newDF['Placement_Offer_Count'] = newDF['Placement_Offer_Count'].astype(int)
```

[112]:
```
newDF
```

[112]:

	Math_Score	Reading	Writing_Score	Placement_Score	Club_Join_date	\
0	64	82	77	88	2020	
1	66	78	75	80	2019	
2	80	77	62	90	2021	
3	72	80	61	91	2021	
5	74	76	64	76	2019	
6	63	75	61	83	2020	
7	67	76	63	90	2020	
8	64	84	78	86	2020	
9	63	91	70	99	2018	
10	64	85	71	93	2019	
11	68	93	77	92	2020	
12	75	88	62	89	2020	
13	68	79	76	89	2020	
14	71	94	75	92	2019	
15	69	88	67	91	2018	
16	60	80	74	75	2021	
18	70	79	71	92	2021	
20	75	86	68	96	2020	
21	62	79	78	82	2019	
22	79	78	75	93	2018	
23	61	88	70	99	2020	
24	66	85	60	93	2020	
25	63	88	64	80	2020	
26	79	91	75	87	2020	
27	60	77	71	92	2018	
28	74	90	72	94	2021	
29	70	90	77	92	2018	

	Placement_Offer_Count
0	0
1	2
2	3
3	3
5	2
6	2

7	3
8	3
9	3
10	3
11	3
12	3
13	3
14	3
15	3
16	2
18	3
20	3
21	2
22	3
23	3
24	1
25	2
26	3
27	3
28	3
29	3

[113]:
```
newDF['Duration'] = newDF.apply(lambda row: 2023 - row.Club_Join_date, axis = 1)
newDF
```

[113]:

	Math_Score	Reading	Writing_Score	Placement_Score	Club_Join_date	\
0	64	82	77	88	2020	
1	66	78	75	80	2019	
2	80	77	62	90	2021	
3	72	80	61	91	2021	
5	74	76	64	76	2019	
6	63	75	61	83	2020	
7	67	76	63	90	2020	
8	64	84	78	86	2020	
9	63	91	70	99	2018	
10	64	85	71	93	2019	
11	68	93	77	92	2020	
12	75	88	62	89	2020	
13	68	79	76	89	2020	
14	71	94	75	92	2019	
15	69	88	67	91	2018	
16	60	80	74	75	2021	
18	70	79	71	92	2021	
20	75	86	68	96	2020	
21	62	79	78	82	2019	
22	79	78	75	93	2018	
23	61	88	70	99	2020	

24	66	85	60	93	2020
25	63	88	64	80	2020
26	79	91	75	87	2020
27	60	77	71	92	2018
28	74	90	72	94	2021
29	70	90	77	92	2018

	Placement_Offer_Count	Duration
0	0	3
1	2	4
2	3	2
3	3	2
5	2	4
6	2	3
7	3	3
8	3	3
9	3	5
10	3	4
11	3	3
12	3	3
13	3	3
14	3	4
15	3	5
16	2	2
18	3	2
20	3	3
21	2	4
22	3	5
23	3	3
24	1	3
25	2	3
26	3	3
27	3	5
28	3	2
29	3	5