```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv("StudentsPerformance.csv")
```

```
df
```

| | Math_Score | Reading_Score | Writing_Score | Placement_Score | Club_Join_Date | Placement_Offer_Count | Region |
|---|---|---|---|---|---|---|---|
| **0** | 75 | 90 | 64 | 75 | 2019 | 2 | Pune |
| **1** | 74 | 86 | 79 | 99 | 2018 | 3 | NaN |
| **2** | 65 | 80 | 69 | 97 | 2021 | 1 | Nashik |
| **3** | 64 | 79 | 76 | 77 | 2020 | 2 | Pune |
| **4** | 95 | 76 | 63 | 97 | 2018 | 3 | Pune |
| **5** | 63 | 82 | 63 | 86 | 2018 | 3 | NaN |
| **6** | 74 | 92 | 80 | 98 | 2019 | 3 | Nashik |
| **7** | 79 | 86 | 62 | 96 | 2019 | 3 | Nashik |
| **8** | 63 | 92 | 70 | 84 | 2019 | 2 | Nashik |
| **9** | 68 | 77 | 64 | 78 | 2020 | 2 | Pune |
| **10** | 65 | 84 | 72 | 80 | 2018 | 2 | Nashik |
| **11** | 80 | 76 | 63 | 91 | 2021 | 1 | Pune |
| **12** | 72 | 80 | 75 | 95 | 2018 | 3 | Nashik |
| **13** | 78 | 85 | 78 | 83 | 2020 | 2 | Nashik |
| **14** | 77 | 84 | 72 | 95 | 2021 | 3 | NaN |
| **15** | 66 | 82 | 72 | 91 | 2019 | 3 | Nashik |
| **16** | 36 | 80 | 62 | 87 | 2020 | 3 | NaN |
| **17** | 66 | 85 | 66 | 100 | 2018 | 3 | Pune |
| **18** | 64 | 83 | 60 | 85 | 2018 | 3 | NaN |
| **19** | 73 | 92 | 75 | 96 | 2019 | 3 | Nashik |
| **20** | 77 | 90 | 68 | 87 | 2020 | 3 | NaN |
| **21** | 66 | 90 | 77 | 92 | 2021 | 3 | NaN |
| **22** | 64 | 95 | 77 | 76 | 2021 | 2 | Pune |
| **23** | 48 | 93 | 71 | 93 | 2020 | 3 | NaN |
| **24** | 77 | 85 | 75 | 83 | 2019 | 2 | NaN |
| **25** | 76 | 77 | 75 | 87 | 2020 | 3 | Pune |
| **26** | 60 | 89 | 80 | 89 | 2018 | 1 | Pune |
| **27** | 87 | 79 | 67 | 95 | 2019 | 3 | Nashik |
| **28** | 74 | 87 | 78 | 87 | 2019 | 3 | NaN |
| **29** | 75 | 83 | 61 | 75 | 2019 | 1 | NaN |

```
df.isnull()
```

| | Math_Score | Reading_Score | Writing_Score | Placement_Score | Club_Join_Date | Placement_Offer_Count | Region |
|---|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False | False |
| **1** | False | False | False | False | False | False | True |
| **2** | False | False | False | False | False | False | False |
| **3** | False | False | False | False | False | False | False |
| **4** | False | False | False | False | False | False | False |
| **5** | False | False | False | False | False | False | True |
| **6** | False | False | False | False | False | False | False |
| **7** | False | False | False | False | False | False | False |
| **8** | False | False | False | False | False | False | False |
| **9** | False | False | False | False | False | False | False |
| **10** | False | False | False | False | False | False | False |
| **11** | False | False | False | False | False | False | False |
| **12** | False | False | False | False | False | False | False |
| **13** | False | False | False | False | False | False | False |
| **14** | False | False | False | False | False | False | True |
| **15** | False | False | False | False | False | False | False |
| **16** | False | False | False | False | False | False | True |
| **17** | False | False | False | False | False | False | False |
| **18** | False | False | False | False | False | False | True |
| **19** | False | False | False | False | False | False | False |
| **20** | False | False | False | False | False | False | True |
| **21** | False | False | False | False | False | False | True |
| **22** | False | False | False | False | False | False | False |
| **23** | False | False | False | False | False | False | True |
| **24** | False | False | False | False | False | False | True |
| **25** | False | False | False | False | False | False | False |
| **26** | False | False | False | False | False | False | False |
| **27** | False | False | False | False | False | False | False |
| **28** | False | False | False | False | False | False | True |
| **29** | False | False | False | False | False | False | True |

```python
series = pd.isnull(df["Math_Score"])
series
```

```
0      False
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
Name: Math_Score, dtype: bool
```

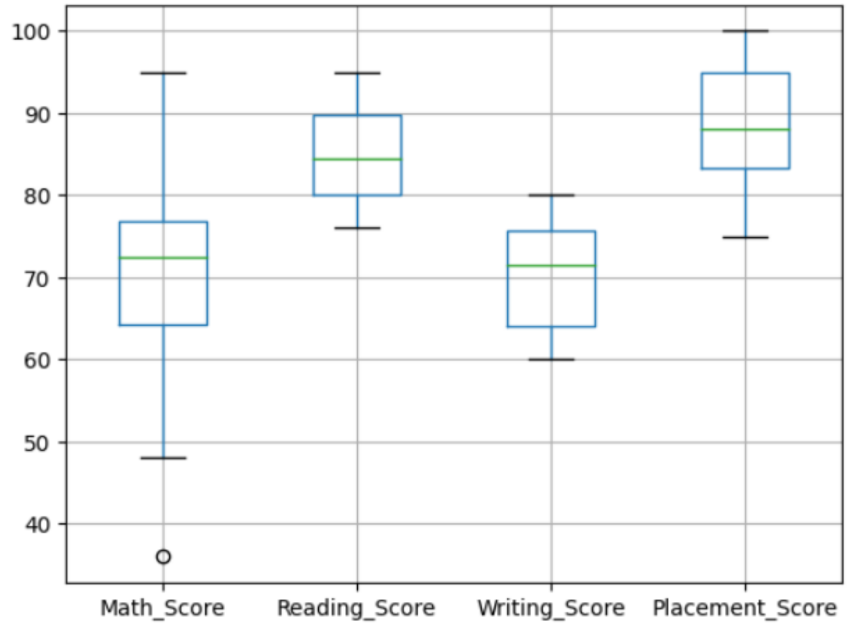
```
m_v=df['Math_Score'].mean()
df['Math_Score'].fillna(value=m_v, inplace=True)
df[series]
```

| | Math_Score | Reading_Score | Writing_Score | Placement_Score | Club_Join_Date | Placement_Offer_Count | Region |
|---|---|---|---|---|---|---|---|

```
ndf=df.dropna()
ndf
```

| | Math_Score | Reading_Score | Writing_Score | Placement_Score | Club_Join_Date | Placement_Offer_Count | Region |
|---|---|---|---|---|---|---|---|
| **0** | 75 | 90 | 64 | 75 | 2019 | 2 | Pune |
| **2** | 65 | 80 | 69 | 97 | 2021 | 1 | Nashik |
| **3** | 64 | 79 | 76 | 77 | 2020 | 2 | Pune |
| **4** | 95 | 76 | 63 | 97 | 2018 | 3 | Pune |
| **6** | 74 | 92 | 80 | 98 | 2019 | 3 | Nashik |
| **7** | 79 | 86 | 62 | 96 | 2019 | 3 | Nashik |
| **8** | 63 | 92 | 70 | 84 | 2019 | 2 | Nashik |
| **9** | 68 | 77 | 64 | 78 | 2020 | 2 | Pune |
| **10** | 65 | 84 | 72 | 80 | 2018 | 2 | Nashik |
| **11** | 80 | 76 | 63 | 91 | 2021 | 1 | Pune |
| **12** | 72 | 80 | 75 | 95 | 2018 | 3 | Nashik |
| **13** | 78 | 85 | 78 | 83 | 2020 | 2 | Nashik |
| **15** | 66 | 82 | 72 | 91 | 2019 | 3 | Nashik |
| **17** | 66 | 85 | 66 | 100 | 2018 | 3 | Pune |
| **19** | 73 | 92 | 75 | 96 | 2019 | 3 | Nashik |
| **22** | 64 | 95 | 77 | 76 | 2021 | 2 | Pune |
| **25** | 76 | 77 | 75 | 87 | 2020 | 3 | Pune |
| **26** | 60 | 89 | 80 | 89 | 2018 | 1 | Pune |

```
col = ['Math_Score', 'Reading_Score' , 'Writing_Score','Placement_Score']
df.boxplot(col)
plt.show()
```



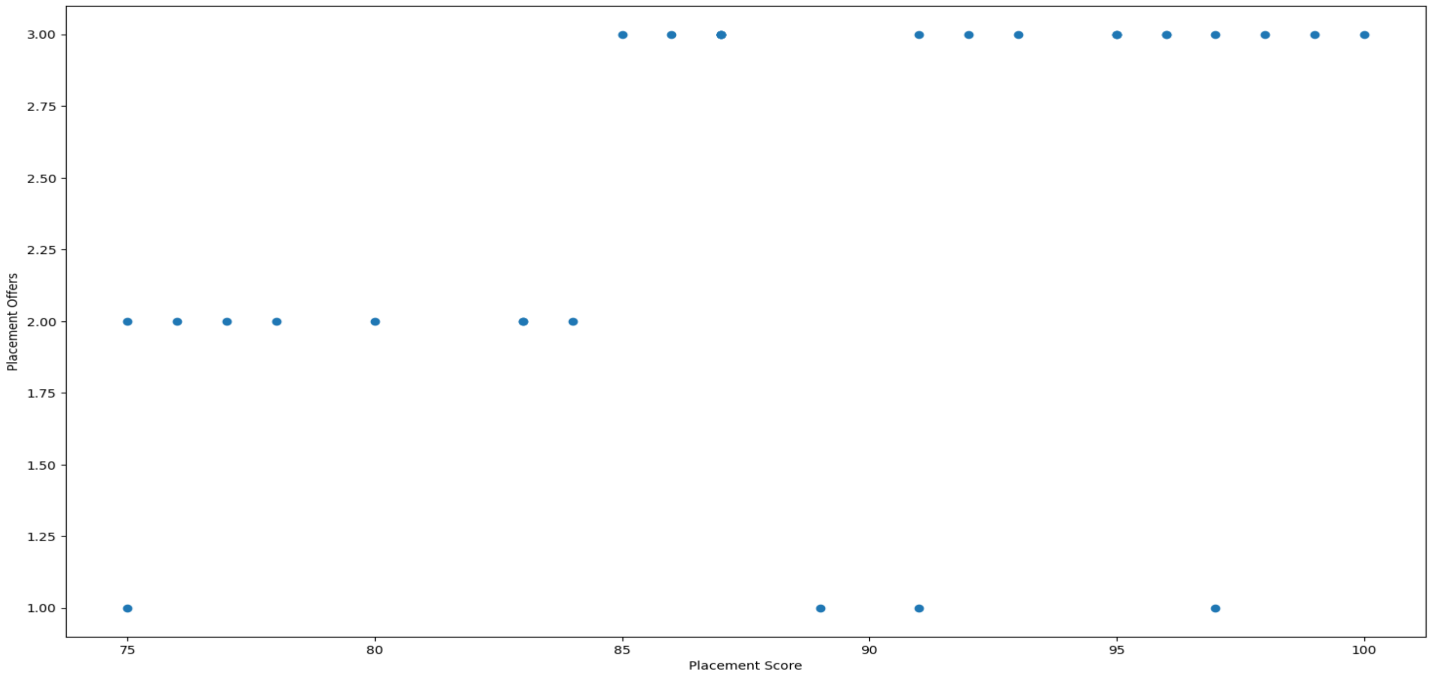
```
np.where(df['Math_Score']>80)
```

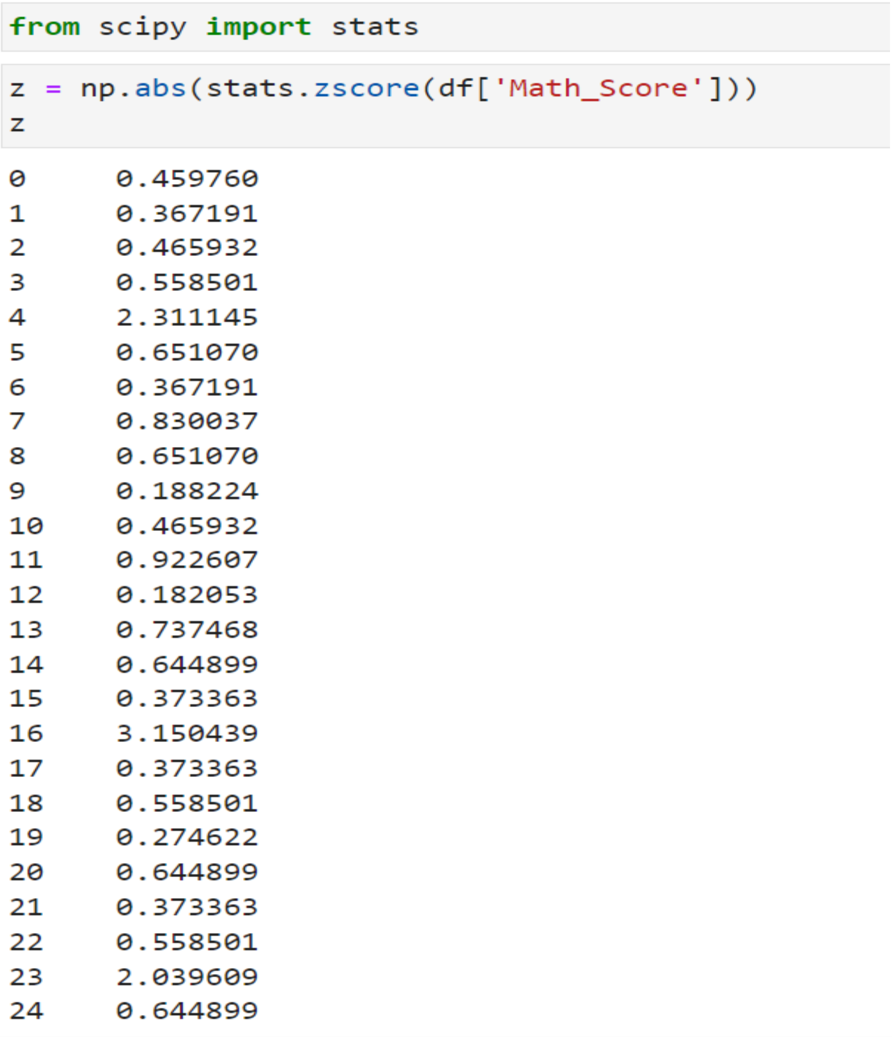
```
(array([ 4, 27], dtype=int64),)
```

```
np.where(df['Reading_Score']<80)
```

```
(array([ 3,  4,  9, 11, 25, 27], dtype=int64),)
```

```
fig, ax = plt.subplots(figsize = (18,10))
ax.scatter(df['Placement_Score'], df['Placement_Offer_Count'])
ax.set_ylabel('Placement Offers')
ax.set_xlabel('Placement Score')
plt.show()
```

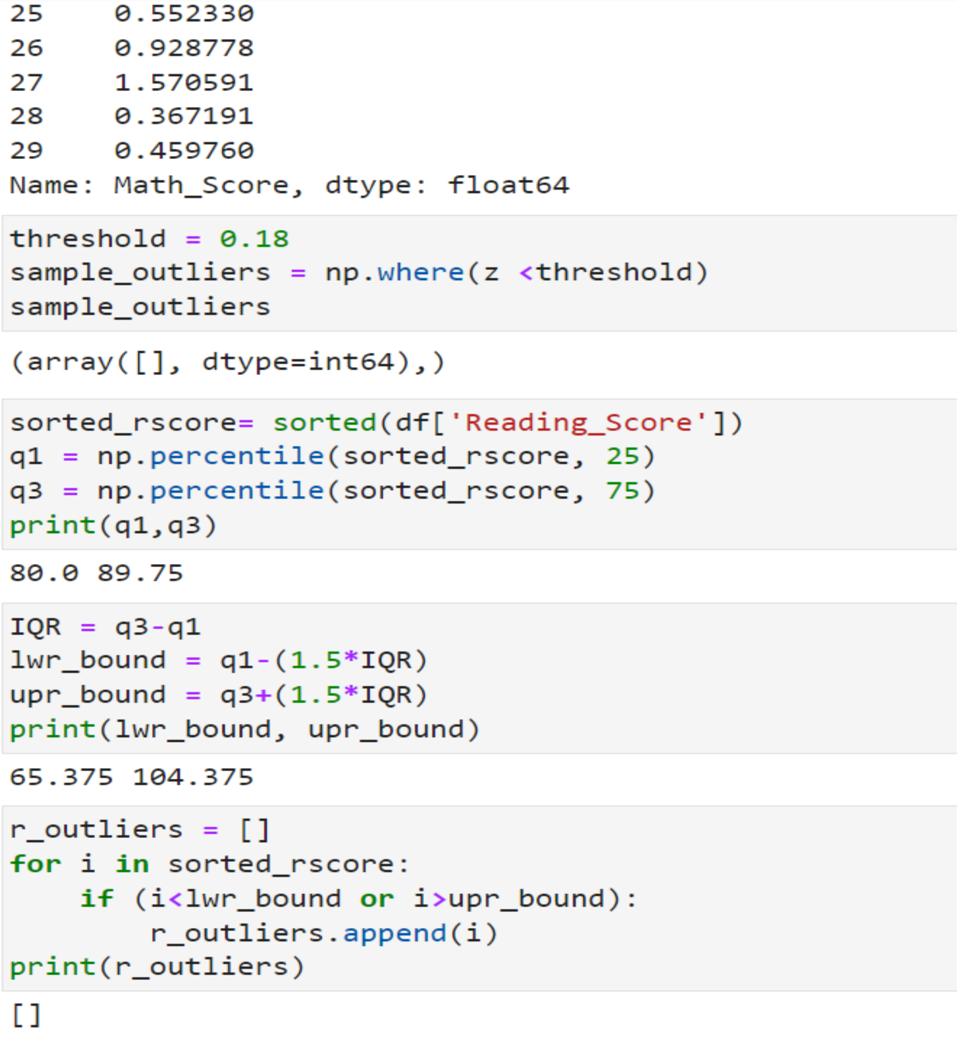




```
from scipy import stats
```

```
z = np.abs(stats.zscore(df['Math_Score']))
z
```

```
0     0.459760
1     0.367191
2     0.465932
3     0.558501
4     2.311145
5     0.651070
6     0.367191
7     0.830037
8     0.651070
9     0.188224
10    0.465932
11    0.922607
12    0.182053
13    0.737468
14    0.644899
15    0.373363
16    3.150439
17    0.373363
18    0.558501
19    0.274622
20    0.644899
21    0.373363
22    0.558501
23    2.039609
24    0.644899
```



```
25    0.552330
26    0.928778
27    1.570591
28    0.367191
29    0.459760
Name: Math_Score, dtype: float64
```

```
threshold = 0.18
sample_outliers = np.where(z <threshold)
sample_outliers
```

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

```
sorted_rscore= sorted(df['Reading_Score'])
q1 = np.percentile(sorted_rscore, 25)
q3 = np.percentile(sorted_rscore, 75)
print(q1,q3)
```

```
80.0 89.75
```

```
IQR = q3-q1
lwr_bound = q1-(1.5*IQR)
upr_bound = q3+(1.5*IQR)
print(lwr_bound, upr_bound)
```

```
65.375 104.375
```

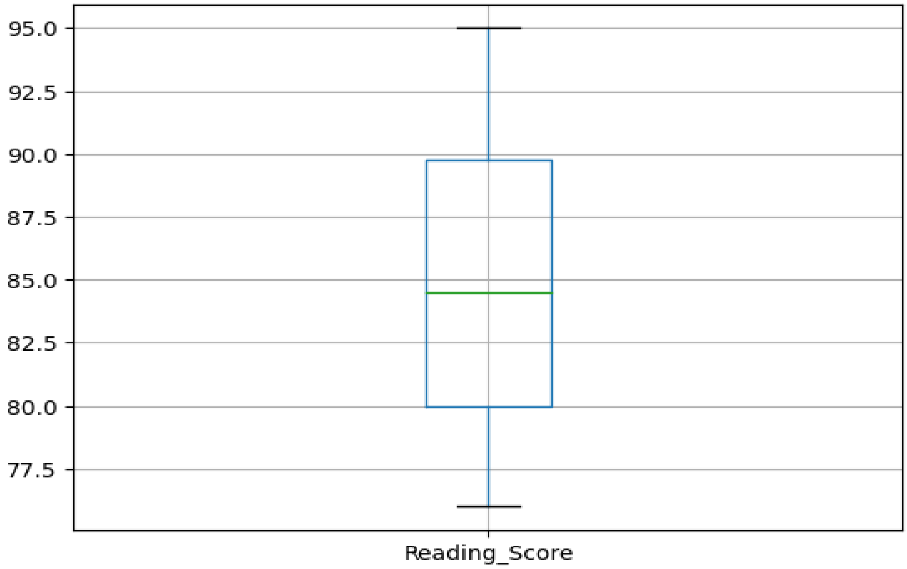
```
r_outliers = []
for i in sorted_rscore:
    if (i<lwr_bound or i>upr_bound):
        r_outliers.append(i)
print(r_outliers)
```

```
[]
```

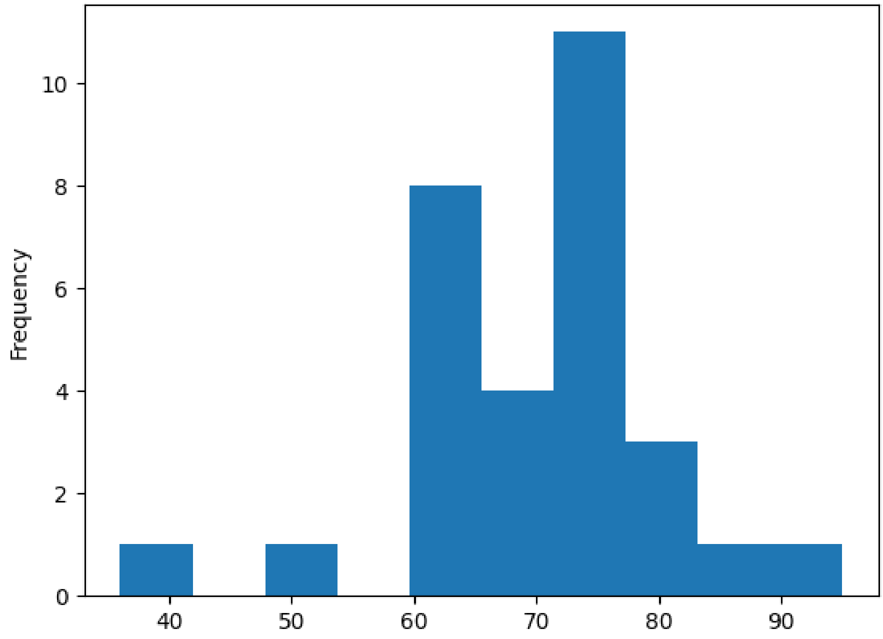
```
median=np.median(sorted_rscore)
refined_df=df
refined_df['Reading_Score'] = np.where(refined_df['Reading_Score'] >upr_bound, median,refined_df['Reading_Score'])
refined_df['Reading Score'] = np.where(refined_df['Reading_Score'] <lwr_bound, median,refined_df['Reading_Score'])
refined_df
```

| | Math_Score | Reading_Score | Writing_Score | Placement_Score | Club_Join_Date | Placement_Offer_Count | Region | Reading Score |
|---|---|---|---|---|---|---|---|---|
| **0** | 75 | 90.0 | 64 | 75 | 2019 | 2 | Pune | 90.0 |
| **1** | 74 | 86.0 | 79 | 99 | 2018 | 3 | NaN | 86.0 |
| **2** | 65 | 80.0 | 69 | 97 | 2021 | 1 | Nashik | 80.0 |
| **3** | 64 | 79.0 | 76 | 77 | 2020 | 2 | Pune | 79.0 |
| **4** | 95 | 76.0 | 63 | 97 | 2018 | 3 | Pune | 76.0 |
| **5** | 63 | 82.0 | 63 | 86 | 2018 | 3 | NaN | 82.0 |
| **6** | 74 | 92.0 | 80 | 98 | 2019 | 3 | Nashik | 92.0 |
| **7** | 79 | 86.0 | 62 | 96 | 2019 | 3 | Nashik | 86.0 |
| **8** | 63 | 92.0 | 70 | 84 | 2019 | 2 | Nashik | 92.0 |
| **9** | 68 | 77.0 | 64 | 78 | 2020 | 2 | Pune | 77.0 |
| **10** | 65 | 84.0 | 72 | 80 | 2018 | 2 | Nashik | 84.0 |
| **11** | 80 | 76.0 | 63 | 91 | 2021 | 1 | Pune | 76.0 |
| **12** | 72 | 80.0 | 75 | 95 | 2018 | 3 | Nashik | 80.0 |
| **13** | 78 | 85.0 | 78 | 83 | 2020 | 2 | Nashik | 85.0 |
| **14** | 77 | 84.0 | 72 | 95 | 2021 | 3 | NaN | 84.0 |
| **15** | 66 | 82.0 | 72 | 91 | 2019 | 3 | Nashik | 82.0 |
| **16** | 36 | 80.0 | 62 | 87 | 2020 | 3 | NaN | 80.0 |
| **17** | 66 | 85.0 | 66 | 100 | 2018 | 3 | Pune | 85.0 |
| **18** | 64 | 83.0 | 60 | 85 | 2018 | 3 | NaN | 83.0 |
| **19** | 73 | 92.0 | 75 | 96 | 2019 | 3 | Nashik | 92.0 |
| **20** | 77 | 90.0 | 68 | 87 | 2020 | 3 | NaN | 90.0 |
| **21** | 66 | 90.0 | 77 | 92 | 2021 | 3 | NaN | 90.0 |
| **22** | 64 | 95.0 | 77 | 76 | 2021 | 2 | Pune | 95.0 |
| **23** | 48 | 93.0 | 71 | 93 | 2020 | 3 | NaN | 93.0 |
| **24** | 77 | 85.0 | 75 | 83 | 2019 | 2 | NaN | 85.0 |
| **25** | 76 | 77.0 | 75 | 87 | 2020 | 3 | Pune | 77.0 |
| **26** | 60 | 89.0 | 80 | 89 | 2018 | 1 | Pune | 89.0 |
| **27** | 87 | 79.0 | 67 | 95 | 2019 | 3 | Nashik | 79.0 |
| **28** | 74 | 87.0 | 78 | 87 | 2019 | 3 | NaN | 87.0 |
| **29** | 75 | 83.0 | 61 | 75 | 2019 | 1 | NaN | 83.0 |

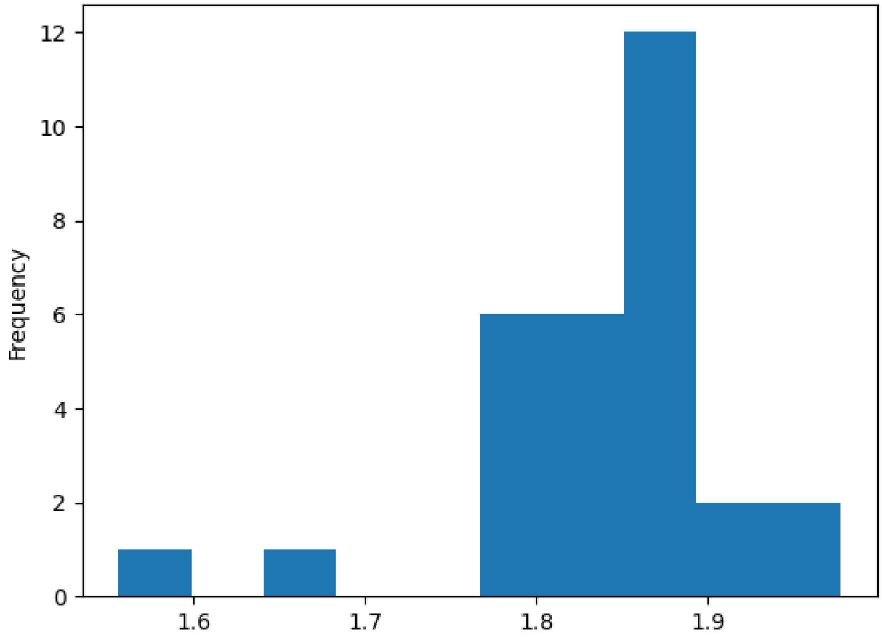
```
col = ['Reading_Score']
refined_df.boxplot(col)
plt.show()
```



```
refined_df['Math_Score'].plot(kind = 'hist')
plt.show()
```



```
df['log_math'] = np.log10(df['Math_Score'])
df['log_math'].plot(kind = 'hist')
plt.show()
```


Frequency