

```
In [1]: import pandas as pd
```

```
In [2]: import numpy as np
```

```
In [3]: dict1={"math_score": [10,45,32,np.nan,62,300,57,84],  
             "reading_score": [75,80,np.nan,73,59,78,88,93],  
             "writing_score": [60,np.nan,63,np.nan,83,65,77,90],  
             "placement_score": [np.nan,np.nan,58,94,87,23,69,86],  
             "join_date": [2022,2023,2019,2021,2024,np.nan,2023,2021]}
```

```
In [4]: dict1
```

```
Out[4]: {'math_score': [10, 45, 32, nan, 62, 300, 57, 84],  
         'reading_score': [75, 80, nan, 73, 59, 78, 88, 93],  
         'writing_score': [60, nan, 63, nan, 83, 65, 77, 90],  
         'placement_score': [nan, nan, 58, 94, 87, 23, 69, 86],  
         'join_date': [2022, 2023, 2019, 2021, 2024, nan, 2023, 2021]}
```

```
In [5]: df=pd.DataFrame(dict1)
```

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

```
Out[6]:
```

	math_score	reading_score	writing_score	placement_score	join_date
0	10.0	75.0	60.0	NaN	2022.0
1	45.0	80.0	NaN	NaN	2023.0
2	32.0	NaN	63.0	58.0	2019.0
3	NaN	73.0	NaN	94.0	2021.0
4	62.0	59.0	83.0	87.0	2024.0
5	300.0	78.0	65.0	23.0	NaN
6	57.0	88.0	77.0	69.0	2023.0
7	84.0	93.0	90.0	86.0	2021.0

```
In [7]: df.isnull()
```

```
Out[7]:
```

	math_score	reading_score	writing_score	placement_score	join_date
0	False	False	False	True	False
1	False	False	True	True	False
2	False	True	False	False	False
3	True	False	True	False	False
4	False	False	False	False	False
5	False	False	False	False	True
6	False	False	False	False	False
7	False	False	False	False	False

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

```
Out[8]:
```

	math_score	reading_score	writing_score	placement_score	join_date
0	True	True	True	False	True
1	True	True	False	False	True
2	True	False	True	True	True
3	False	True	False	True	True
4	True	True	True	True	True
5	True	True	True	True	False
6	True	True	True	True	True
7	True	True	True	True	True

```
In [10]: df.index
```

```
Out[10]: RangeIndex(start=0, stop=8, step=1)
```

```
In [11]: dict2={"math_score": [10,45,32,np.nan,62,300,57,84],  
               "reading_score": [75,80,np.nan,73,59,78,88,93],  
               "writing_score": [60,np.nan,63,np.nan,83,65,77,90],  
               "placement_score": [np.nan,np.nan,58,94,87,23,69,86],  
               "join_date": [2022,2023,2019,2021,2024,np.nan,2023,2021],  
               "Gender": ['M','F','M','F','M','F','M','F'],  
               "Region": ['Pune','Mumbai','Satara','Pune','Maharashtra','Saswad','BHo
```

```
In [12]: df=pd.DataFrame(dict2)
```

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

```
Out[13]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.0	75.0	60.0	NaN	2022.0	M	Pune
1	45.0	80.0	NaN	NaN	2023.0	F	Mumbai
2	32.0	NaN	63.0	58.0	2019.0	M	Satara
3	NaN	73.0	NaN	94.0	2021.0	F	Pune
4	62.0	59.0	83.0	87.0	2024.0	M	Maharashtra
5	300.0	78.0	65.0	23.0	NaN	F	Saswad
6	57.0	88.0	77.0	69.0	2023.0	M	BHor
7	84.0	93.0	90.0	86.0	2021.0	F	Pune

```
In [14]: df.notnull()
```

```
Out[14]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	True	True	True	False	True	True	True
1	True	True	False	False	True	True	True
2	True	False	True	True	True	True	True
3	False	True	False	True	True	True	True
4	True	True	True	True	True	True	True
5	True	True	True	True	False	True	True
6	True	True	True	True	True	True	True
7	True	True	True	True	True	True	True

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

```
Out[15]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	False	False	False	True	False	False	False
1	False	False	True	True	False	False	False
2	False	True	False	False	False	False	False
3	True	False	True	False	False	False	False
4	False	False	False	False	False	False	False
5	False	False	False	False	True	False	False
6	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False

```
In [16]: df.columns
```

```
Out[16]: Index(['math_score', 'reading_score', 'writing_score', 'placement_score',  
              'join_date', 'Gender', 'Region'],  
              dtype='object')
```

```
In [17]: df.dtypes
```

```
Out[17]: math_score      float64  
reading_score    float64  
writing_score     float64  
placement_score   float64  
join_date        float64  
Gender           object  
Region           object  
dtype: object
```

```
In [18]: df.isnull().sum().sum()
```

```
Out[18]: 7
```

```
In [19]: df["math_score"]=df["math_score"].fillna(df["math_score"].mean())
```

```
In [20]: df
```

```
Out[20]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	NaN	NaN	2023.0	F	Mumbai
2	32.000000	NaN	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	NaN	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	NaN	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [24]: df.fillna(0)
```

```
Out[24]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	0.0	2022.0	M	Pune
1	45.000000	80.0	0.0	0.0	2023.0	F	Mumbai
2	32.000000	0.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	0.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	0.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [25]: df.fillna(70)
```

```
Out[25]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	70.0	2022.0	M	Pune
1	45.000000	80.0	70.0	70.0	2023.0	F	Mumbai
2	32.000000	70.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	70.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	70.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [26]: df["math_score"]=df["math_score"].fillna(df["math_score"].mean())
```

In [27]:

```
df
```

Out[27]:

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	NaN	NaN	2023.0	F	Mumbai
2	32.000000	NaN	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	NaN	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	NaN	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

In [31]: `df["writing_score"]=df["writing_score"].fillna(df["writing_score"].mode()[0])`

In [32]:

```
df
```

Out[32]:

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	NaN	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	NaN	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

In [33]: `df["reading_score"]=df["reading_score"].fillna(df["reading_score"].mode()[0])`

In [34]:

```
df
```

Out[34]:

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	NaN	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [35]: df["placement_score"]=df["placement_score"].fillna(df["placement_sco
```

```
In [36]: df
```

```
Out[36]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	NaN	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [37]: df["join_date"]=df["join_date"].fillna(df["join_date"].min())
```

```
In [38]: df
```

```
Out[38]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [39]: df.dropna()
```

```
Out[39]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [40]: df.dropna(how='all')
```

```
Out[40]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

```
In [41]: df.dropna(axis=1)
```

```
Out[41]:
```

	math_score	reading_score	writing_score	join_date	Gender	Region
0	10.000000	75.0	60.0	2022.0	M	Pune
1	45.000000	80.0	63.0	2023.0	F	Mumbai
2	32.000000	78.0	63.0	2019.0	M	Satara
3	84.285714	73.0	77.0	2021.0	F	Pune
4	62.000000	59.0	83.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	2023.0	M	BHor
7	84.000000	93.0	90.0	2021.0	F	Pune

```
In [44]: df.dropna(subset=('Region', 'Gender'))
```

```
Out[44]:
```

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

In [43]: df

Out[43]:

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

In [45]: df

Out[45]:

	math_score	reading_score	writing_score	placement_score	join_date	Gender	Region
0	10.000000	75.0	60.0	NaN	2022.0	M	Pune
1	45.000000	80.0	63.0	NaN	2023.0	F	Mumbai
2	32.000000	78.0	63.0	58.0	2019.0	M	Satara
3	84.285714	73.0	77.0	94.0	2021.0	F	Pune
4	62.000000	59.0	83.0	87.0	2024.0	M	Maharastra
5	300.000000	78.0	65.0	23.0	2019.0	F	Saswad
6	57.000000	88.0	77.0	69.0	2023.0	M	BHor
7	84.000000	93.0	90.0	86.0	2021.0	F	Pune

In [48]: df['Gender'].replace(to_replace=np.nan,value='F')

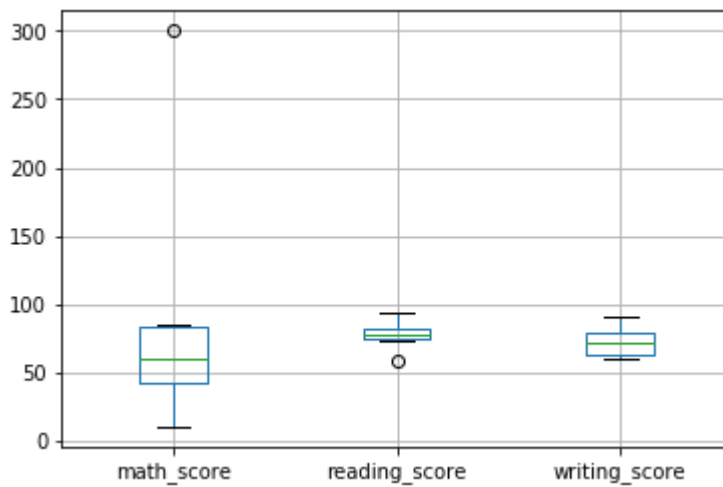
Out[48]:

0	M
1	F
2	M
3	F
4	M
5	F
6	M
7	F

Name: Gender, dtype: object


```
In [50]: coll=['math_score','reading_score','writing_score']
df.boxplot(coll)
```

Out[50]: <matplotlib.axes._subplots.AxesSubplot at 0x7f898cac6390>



```
In [51]: rscore=df['reading_score']
q1=np.percentile(rscore,25)
q3=np.percentile(rscore,75)
print(q1,q3)
```

74.5 82.0

```
In [52]: iqr=q3-q1
print(iqr)
```

7.5

```
In [53]: l_bound=q1-1.5*iqr
u_bound=q3+1.5*iqr
print(l_bound,u_bound)
```

63.25 93.25

```
In [55]: r_outlier=[]
for i in rscore:
    if(i<l_bound or i>u_bound):
        r_outlier.append(i)
print(r_outlier)
```

[59.0]

```
In [56]: median=np.median(rscore)
median
```

Out[56]: 78.0

```
In [59]: df['reading_score']=np.where(df['reading_score']>u_bound,median,df['
```

```
In [60]: df['reading_score']
```

```
Out[60]: 0    75.0
         1    80.0
         2    78.0
         3    73.0
         4    59.0
         5    78.0
         6    88.0
         7    93.0
         Name: reading_score, dtype: float64
```

```
In [63]: x=df.drop(axis=1,columns=["Region","Gender","join_date"])
```

```
In [64]: x
```

```
Out[64]:
```

	math_score	reading_score	writing_score	placement_score
0	10.000000	75.0	60.0	NaN
1	45.000000	80.0	63.0	NaN
2	32.000000	78.0	63.0	58.0
3	84.285714	73.0	77.0	94.0
4	62.000000	59.0	83.0	87.0
5	300.000000	78.0	65.0	23.0
6	57.000000	88.0	77.0	69.0
7	84.000000	93.0	90.0	86.0

```
In [66]: from sklearn import preprocessing
min_max_scaler=preprocessing.MinMaxScaler()
x_scaled = min_max_scaler.fit_transform(x)
df_normalized=pd.DataFrame(x_scaled)
```

```
In [67]: df_normalized
```

```
Out[67]:
```

	0	1	2	3
0	0.000000	0.470588	0.000000	NaN
1	0.120690	0.617647	0.100000	NaN
2	0.075862	0.558824	0.100000	0.492958
3	0.256158	0.411765	0.566667	1.000000
4	0.179310	0.000000	0.766667	0.901408
5	1.000000	0.558824	0.166667	0.000000
6	0.162069	0.852941	0.566667	0.647887
7	0.255172	1.000000	1.000000	0.887324

```
In [68]: df_normalized.rename(columns={0:'math_score',
1:'reading_score',2:'writing_score',3:'placement_score'},inplace=True)
```

In [69]: df_normalized

Out[69]:

	math_score	reading_score	writing_score	placement_score
0	0.000000	0.470588	0.000000	NaN
1	0.120690	0.617647	0.100000	NaN
2	0.075862	0.558824	0.100000	0.492958
3	0.256158	0.411765	0.566667	1.000000
4	0.179310	0.000000	0.766667	0.901408
5	1.000000	0.558824	0.166667	0.000000
6	0.162069	0.852941	0.566667	0.647887
7	0.255172	1.000000	1.000000	0.887324

In []: