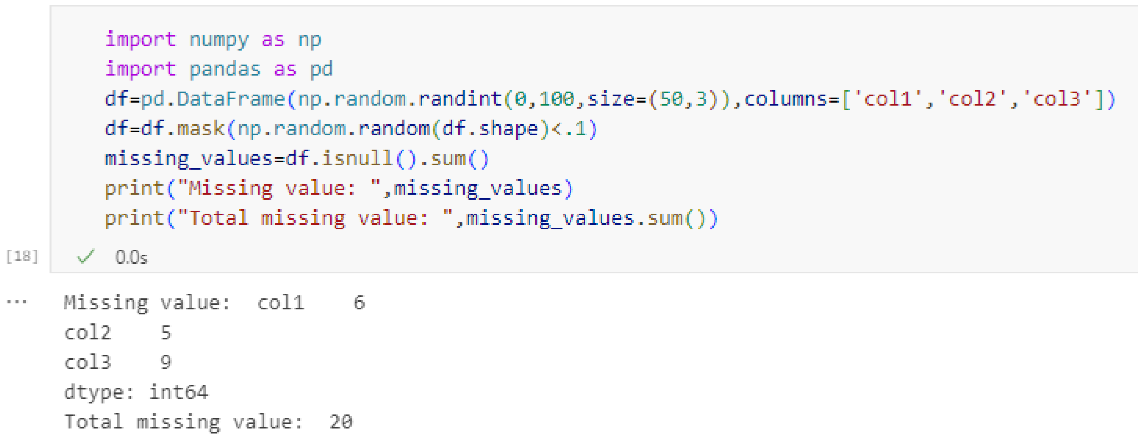
NAME : GARIMA ROHILLA
ROLL NO : CSC/21/37

# PRACTICAL - 03

Create a dataframe having at least 3 columns and 50 rows to store numeric data generated using a random function. Replace 10% of the values by null values whose index positions are generated using random function.

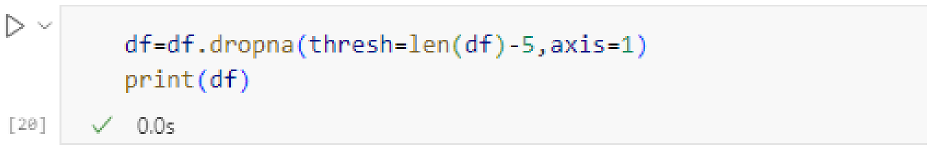
a. Identify and count missing values in a dataframe.



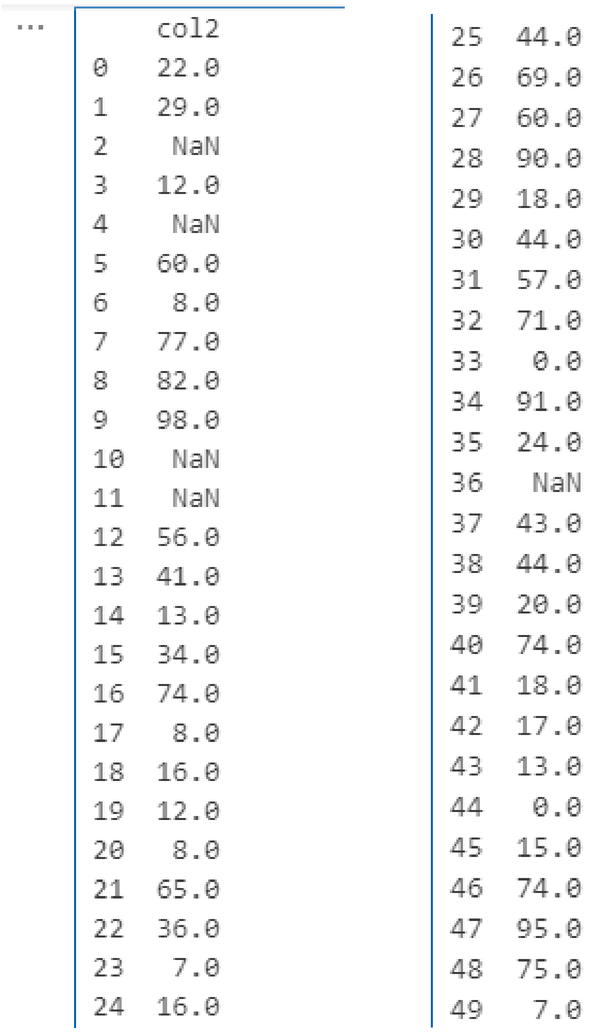
```python
import numpy as np
import pandas as pd
df=pd.DataFrame(np.random.randint(0,100,size=(50,3)),columns=['col1','col2','col3'])
df=df.mask(np.random.random(df.shape)<.1)
missing_values=df.isnull().sum()
print("Missing value: ",missing_values)
print("Total missing value: ",missing_values.sum())
```

```
Missing value:  col1    6
col2    5
col3    9
dtype: int64
Total missing value:  20
```

b. Drop the column having more than 5 null values.

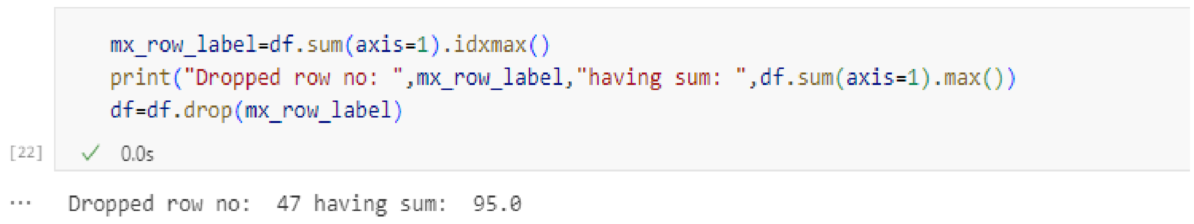


```python
df=df.dropna(thresh=len(df)-5,axis=1)
print(df)
```



```
     col2
0    22.0
1    29.0
2     NaN
3    12.0
4     NaN
5    60.0
6     8.0
7    77.0
8    82.0
9    98.0
10    NaN
11    NaN
12   56.0
13   41.0
14   13.0
15   34.0
16   74.0
17    8.0
18   16.0
19   12.0
20    8.0
21   65.0
22   36.0
23    7.0
24   16.0
25   44.0
26   69.0
27   60.0
28   90.0
29   18.0
30   44.0
31   57.0
32   71.0
33    0.0
34   91.0
35   24.0
36    NaN
37   43.0
38   44.0
39   20.0
40   74.0
41   18.0
42   17.0
43   13.0
44    0.0
45   15.0
46   74.0
47   95.0
48   75.0
49    7.0
```

c. Identify the row label having maximum of the sum of all values in a row and drop that row.



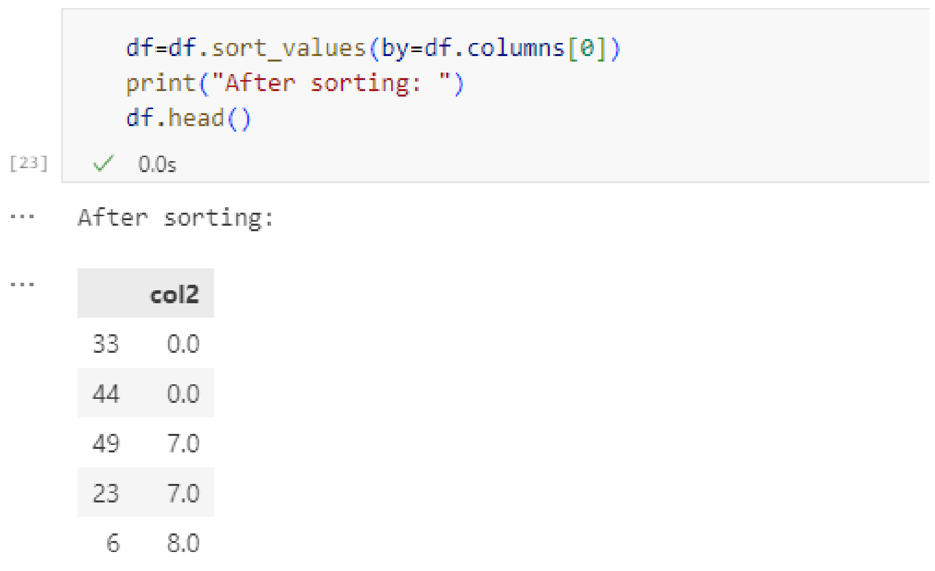
```
mx_row_label=df.sum(axis=1).idxmax()
print("Dropped row no: ",mx_row_label,"having sum: ",df.sum(axis=1).max())
df=df.drop(mx_row_label)
```

```
[22]  ✓ 0.0s
```

```
Dropped row no:  47 having sum:  95.0
```

d. Sort the dataframe on the basis of the first column.

NAME : GARIMA ROHILLA
ROLL NO : CSC/21/37



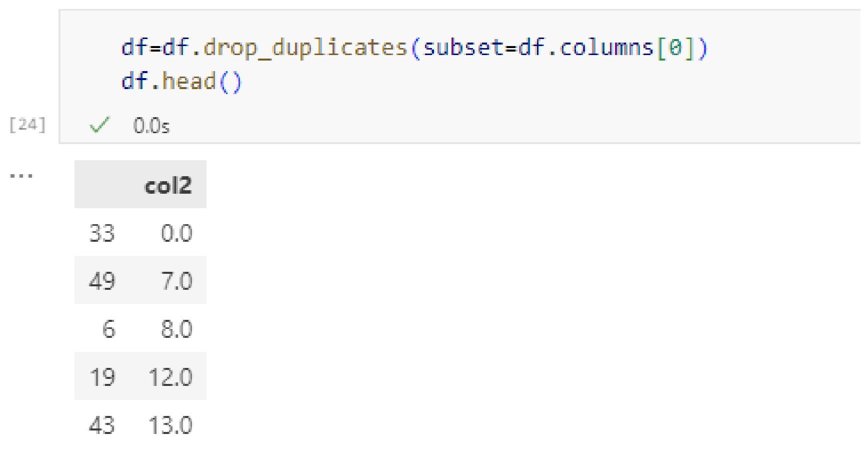
```
df=df.sort_values(by=df.columns[0])
print("After sorting: ")
df.head()
```

```
[23]  ✓  0.0s
```

```
After sorting:

      col2
33     0.0
44     0.0
49     7.0
23     7.0
 6     8.0
```

e. Remove all duplicates from the first column.



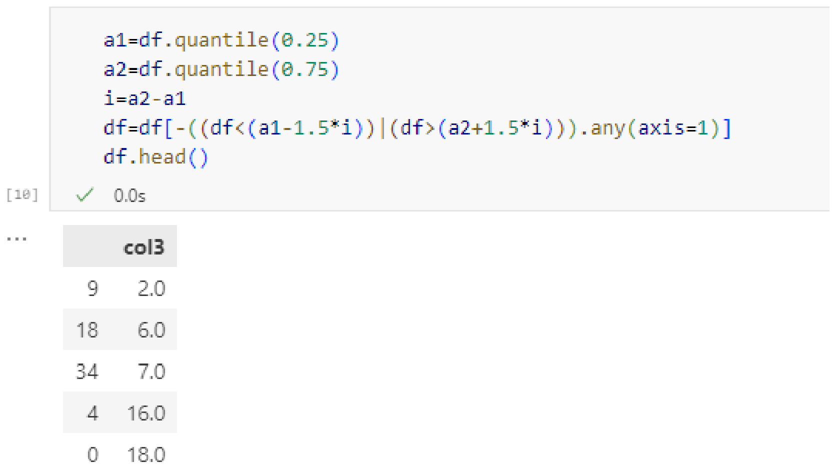
```
df=df.drop_duplicates(subset=df.columns[0])
df.head()
```

```
[24]  ✓  0.0s
```

```
      col2
33     0.0
49     7.0
 6     8.0
19    12.0
43    13.0
```

f. Detect the outliers and remove the rows having outliers.

NAME : GARIMA ROHILLA
ROLL NO : CSC/21/37

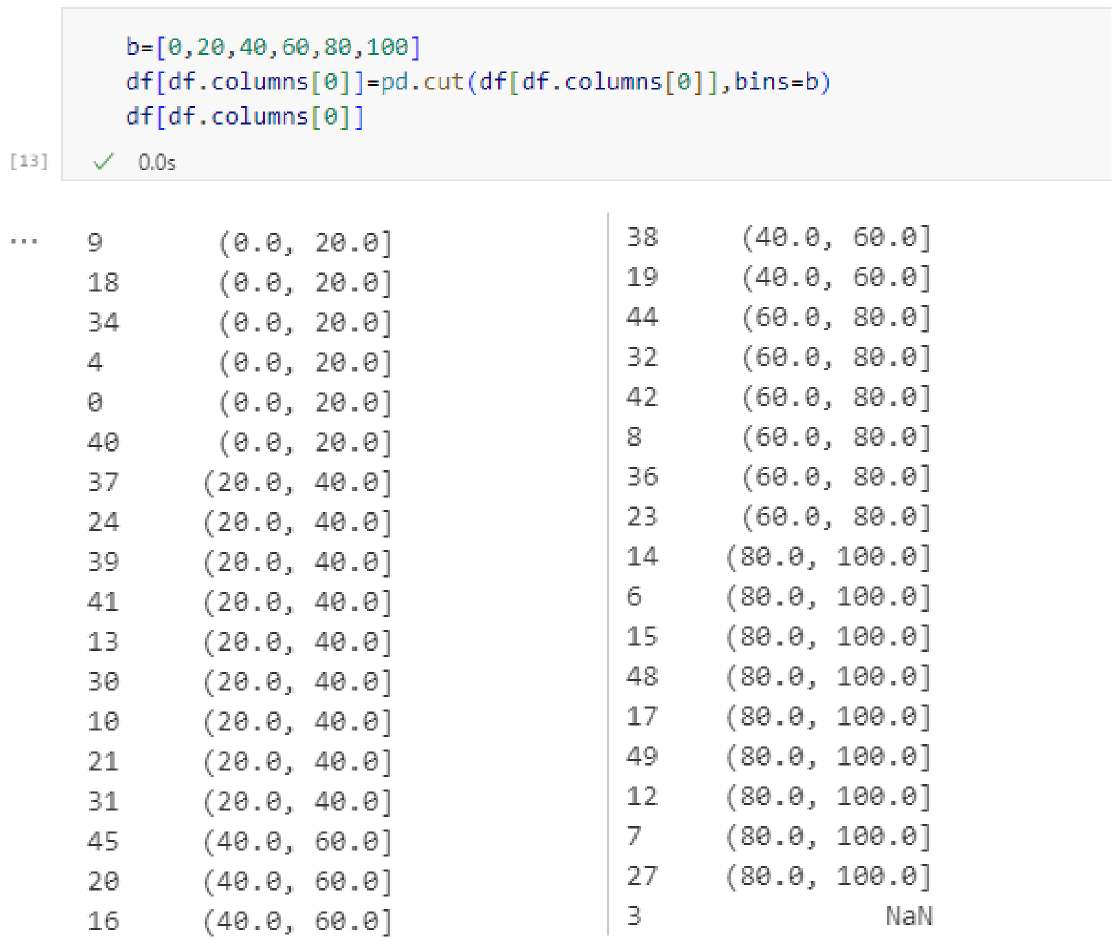


```python
a1=df.quantile(0.25)
a2=df.quantile(0.75)
i=a2-a1
df=df[-((df<(a1-1.5*i))|(df>(a2+1.5*i))).any(axis=1)]
df.head()
```

[10] ✓ 0.0s

|    | col3 |
|----|------|
| 9  | 2.0  |
| 18 | 6.0  |
| 34 | 7.0  |
| 4  | 16.0 |
| 0  | 18.0 |

g. Discretize second column and create 5 bins.



```python
b=[0,20,40,60,80,100]
df[df.columns[0]]=pd.cut(df[df.columns[0]],bins=b)
df[df.columns[0]]
```

[13] ✓ 0.0s

```
9        (0.0, 20.0]
18       (0.0, 20.0]
34       (0.0, 20.0]
4        (0.0, 20.0]
0        (0.0, 20.0]
40       (0.0, 20.0]
37      (20.0, 40.0]
24      (20.0, 40.0]
39      (20.0, 40.0]
41      (20.0, 40.0]
13      (20.0, 40.0]
30      (20.0, 40.0]
10      (20.0, 40.0]
21      (20.0, 40.0]
31      (20.0, 40.0]
45      (40.0, 60.0]
20      (40.0, 60.0]
16      (40.0, 60.0]
38      (40.0, 60.0]
19      (40.0, 60.0]
44      (60.0, 80.0]
32      (60.0, 80.0]
42      (60.0, 80.0]
8       (60.0, 80.0]
36      (60.0, 80.0]
23      (60.0, 80.0]
14     (80.0, 100.0]
6      (80.0, 100.0]
15     (80.0, 100.0]
48     (80.0, 100.0]
17     (80.0, 100.0]
49     (80.0, 100.0]
12     (80.0, 100.0]
7      (80.0, 100.0]
27     (80.0, 100.0]
3                NaN
```