dsbda-assignment2-new-1

April 15, 2024

[ ]:
```
import pandas as pd
import numpy as np
```

[ ]:
```
from google.colab import files
files.upload()
```

```
<IPython.core.display.HTML object>
```

```
Saving dataset.csv to dataset.csv
```

[ ]:
```
{'dataset.csv':
b'Roll.no,Name,Sem,Attendance,Sub1,Sub2,Sub3,Sub4,Sub5,Percentage,Result,Total M
arks\r\n1,NaN,5,33,3,94,92,50,1,60,Pass,240\r\n2,NaN,6,36,56,74,99,75,3,76.75,Pa
ss,307\r\n3,NaN,5,92,74,46,10,52,3,46.25,Pass,185\r\n4,NaN,6,5,9,81,39,0,32,40.2
5,Pass,161\r\n5,NaN,5,23,98,34,93,3,32,65,Pass,260\r\n6,NaN,6,26,56,70,12,NaN,65
,50.75,Pass,203\r\n7,NaN,5,16,92,NaN,18,45,3,39.5,Pass,158\r\n8,NaN,6,39,71,85,-
21,65,48,62,Pass,248\r\n9,NaN,6,24,31,48,6,84,74,60.75,Pass,243\r\n10,NaN,5,60,5
3,6,16,81,180,59,Pass,336\r\n11,NaN,6,NaN,51,8,67,2,14,35.5,fail,142\r\n12,NaN,6
,21,68,68,28,73,26,65.75,Pass,263\r\n13,NaN,6,57,32,99,NaN,36,93,65,Pass,260\r\n
14,NaN,5,100,19,71,40,30,13,43.25,Pass,173\r\n15,NaN,5,24,47,-34,43,NaN,70,31.5,
fail,126\r\n16,NaN,6,18,38,17,57,69,48,57.25,Pass,229\r\n17,NaN,6,63,72,13,6,68,
6,41.25,Pass,165\r\n18,NaN,5,55,NaN,7,12,96,42,39.25,Pass,157\r\n19,NaN,5,92,60,
35,12,0,45,38,Pass,152\r\n20,NaN,6,66,28,10,196,94,55,70.75,Pass,383\r\n21,NaN,6
,69,86,97,73,38,37,82.75,Pass,331\r\n22,NaN,5,90,78,43,14,NaN,91,56.5,Pass,226\r
\n23,NaN,5,52,74,94,48,53,16,71.25,Pass,285\r\n24,NaN,6,NaN,30,40,41,21,74,51.5,
Pass,206\r\n25,NaN,6,NaN,76,39,63,23,19,55,Pass,220\r\n26,NaN,5,70,83,34,90,42,4
9,74.5,Pass,298\r\n27,NaN,6,40,14,13,42,43,20,33,fail,132\r\n28,NaN,6,71,45,18,2
5,31,27,36.5,Pass,146\r\n29,NaN,5,79,17,60,18,85,0,45,Pass,180\r\n30,NaN,6,35,21
,48,47,75,9,50,Pass,200\r\n'}
```

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

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
0	1	NaN	5	33.0	3.0	94.0	92.0	50.0	1	60.00	
1	2	NaN	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	NaN	5	92.0	74.0	46.0	10.0	52.0	3	46.25	
3	4	NaN	6	5.0	9.0	81.0	39.0	0.0	32	40.25	

4	5	NaN	5	23.0	98.0	34.0	93.0	3.0	32	65.00
5	6	NaN	6	26.0	56.0	70.0	12.0	NaN	65	50.75
6	7	NaN	5	16.0	92.0	NaN	18.0	45.0	3	39.50
7	8	NaN	6	39.0	71.0	85.0	-21.0	65.0	48	62.00
8	9	NaN	6	24.0	31.0	48.0	6.0	84.0	74	60.75
9	10	NaN	5	60.0	53.0	6.0	16.0	81.0	180	59.00
10	11	NaN	6	NaN	51.0	8.0	67.0	2.0	14	35.50
11	12	NaN	6	21.0	68.0	68.0	28.0	73.0	26	65.75
12	13	NaN	6	57.0	32.0	99.0	NaN	36.0	93	65.00
13	14	NaN	5	100.0	19.0	71.0	40.0	30.0	13	43.25
14	15	NaN	5	24.0	47.0	-34.0	43.0	NaN	70	31.50
15	16	NaN	6	18.0	38.0	17.0	57.0	69.0	48	57.25
16	17	NaN	6	63.0	72.0	13.0	6.0	68.0	6	41.25
17	18	NaN	5	55.0	NaN	7.0	12.0	96.0	42	39.25
18	19	NaN	5	92.0	60.0	35.0	12.0	0.0	45	38.00
19	20	NaN	6	66.0	28.0	10.0	196.0	94.0	55	70.75
20	21	NaN	6	69.0	86.0	97.0	73.0	38.0	37	82.75
21	22	NaN	5	90.0	78.0	43.0	14.0	NaN	91	56.50
22	23	NaN	5	52.0	74.0	94.0	48.0	53.0	16	71.25
23	24	NaN	6	NaN	30.0	40.0	41.0	21.0	74	51.50
24	25	NaN	6	NaN	76.0	39.0	63.0	23.0	19	55.00
25	26	NaN	5	70.0	83.0	34.0	90.0	42.0	49	74.50
26	27	NaN	6	40.0	14.0	13.0	42.0	43.0	20	33.00
27	28	NaN	6	71.0	45.0	18.0	25.0	31.0	27	36.50
28	29	NaN	5	79.0	17.0	60.0	18.0	85.0	0	45.00
29	30	NaN	6	35.0	21.0	48.0	47.0	75.0	9	50.00

	Result	Total Marks
0	Pass	240
1	Pass	307
2	Pass	185
3	Pass	161
4	Pass	260
5	Pass	203
6	Pass	158
7	Pass	248
8	Pass	243
9	Pass	336
10	fail	142
11	Pass	263
12	Pass	260
13	Pass	173
14	fail	126
15	Pass	229
16	Pass	165
17	Pass	157
18	Pass	152

```
19   Pass          383
20   Pass          331
21   Pass          226
22   Pass          285
23   Pass          206
24   Pass          220
25   Pass          298
26   fail          132
27   Pass          146
28   Pass          180
29   Pass          200
```

[ ]:
```
df.head()
```

[ ]:
```
   Roll.no  Name  Sem  Attendance  Sub1  Sub2  Sub3  Sub4  Sub5  Percentage  \
0        1   NaN    5        33.0   3.0  94.0  92.0  50.0     1       60.00
1        2   NaN    6        36.0  56.0  74.0  99.0  75.0     3       76.75
2        3   NaN    5        92.0  74.0  46.0  10.0  52.0     3       46.25
3        4   NaN    6         5.0   9.0  81.0  39.0   0.0    32       40.25
4        5   NaN    5        23.0  98.0  34.0  93.0   3.0    32       65.00

  Result  Total Marks
0   Pass          240
1   Pass          307
2   Pass          185
3   Pass          161
4   Pass          260
```

[ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Roll.no      30 non-null     int64
 1   Name         0 non-null      float64
 2   Sem          30 non-null     int64
 3   Attendance   27 non-null     float64
 4   Sub1         29 non-null     float64
 5   Sub2         29 non-null     float64
 6   Sub3         29 non-null     float64
 7   Sub4         27 non-null     float64
 8   Sub5         30 non-null     int64
 9   Percentage   30 non-null     float64
 10  Result       30 non-null     object
 11  Total Marks  30 non-null     int64
```



```
dtypes: float64(7), int64(4), object(1)
memory usage: 2.9+ KB
```

[ ]:
```
df.describe(include = "all")
```

[ ]:
```
          Roll.no  Name        Sem  Attendance       Sub1       Sub2  \
count   30.000000   0.0  30.000000   27.000000  29.000000  29.000000
unique        NaN   NaN        NaN         NaN        NaN        NaN
top           NaN   NaN        NaN         NaN        NaN        NaN
freq          NaN   NaN        NaN         NaN        NaN        NaN
mean    15.500000   NaN   5.566667   50.222222  51.103448  45.448276
std      8.803408   NaN   0.504007   26.966266  26.915668  33.803620
min      1.000000   NaN   5.000000    5.000000   3.000000 -34.000000
25%      8.250000   NaN   5.000000   25.000000  30.000000  17.000000
50%     15.500000   NaN   6.000000   52.000000  53.000000  43.000000
75%     22.750000   NaN   6.000000   69.500000  74.000000  71.000000
max     30.000000   NaN   6.000000  100.000000  98.000000  99.000000

              Sub3       Sub4       Sub5  Percentage Result  Total Marks
count    29.000000  27.000000  30.000000   30.000000     30    30.000000
unique         NaN        NaN        NaN         NaN      2          NaN
top            NaN        NaN        NaN         NaN   Pass          NaN
freq           NaN        NaN        NaN         NaN     27          NaN
mean     44.344828  49.407407  39.833333   53.458333    NaN   220.500000
std      42.224126  29.357556  38.046327   14.169857    NaN    66.978509
min     -21.000000   0.000000   0.000000   31.500000    NaN   126.000000
25%      14.000000  30.500000  13.250000   40.500000    NaN   162.000000
50%      40.000000  50.000000  32.000000   53.250000    NaN   213.000000
75%      63.000000  74.000000  53.500000   64.250000    NaN   260.000000
max     196.000000  96.000000 180.000000   82.750000    NaN   383.000000
```

[ ]:
```
df.shape
```

[ ]:
```
(30, 12)
```

[ ]:
```
df.dtypes
```

[ ]:
```
Roll.no          int64
Name           float64
Sem              int64
Attendance     float64
Sub1           float64
Sub2           float64
Sub3           float64
Sub4           float64
Sub5             int64
Percentage     float64
```

```
Result          object
Total Marks      int64
dtype: object
```

[ ]:
```
df.columns
```

[ ]:
```
Index(['Roll.no', 'Name', 'Sem', 'Attendance', 'Sub1', 'Sub2', 'Sub3', 'Sub4',
       'Sub5', 'Percentage', 'Result', 'Total Marks'],
      dtype='object')
```

[ ]:
```
df[0:5]
```

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
0	1	NaN	5	33.0	3.0	94.0	92.0	50.0	1	60.00	
1	2	NaN	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	NaN	5	92.0	74.0	46.0	10.0	52.0	3	46.25	
3	4	NaN	6	5.0	9.0	81.0	39.0	0.0	32	40.25	
4	5	NaN	5	23.0	98.0	34.0	93.0	3.0	32	65.00	

	Result	Total Marks
0	Pass	240
1	Pass	307
2	Pass	185
3	Pass	161
4	Pass	260

[ ]:
```
df.loc[0:2]
```

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
0	1	NaN	5	33.0	3.0	94.0	92.0	50.0	1	60.00	
1	2	NaN	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	NaN	5	92.0	74.0	46.0	10.0	52.0	3	46.25	

	Result	Total Marks
0	Pass	240
1	Pass	307
2	Pass	185

[ ]:
```
df.loc[0:2,'Sub1':'Sub5']
```

[ ]:

	Sub1	Sub2	Sub3	Sub4	Sub5
0	3.0	94.0	92.0	50.0	1
1	56.0	74.0	99.0	75.0	3
2	74.0	46.0	10.0	52.0	3

[ ]:
```
df.iloc[1:3]
```


[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
1	2	NaN	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	NaN	5	92.0	74.0	46.0	10.0	52.0	3	46.25	

	Result	Total Marks
1	Pass	307
2	Pass	185

[ ]:
```
df.iloc[1:5,1:5]
```

[ ]:

	Name	Sem	Attendance	Sub1
1	NaN	6	36.0	56.0
2	NaN	5	92.0	74.0
3	NaN	6	5.0	9.0
4	NaN	5	23.0	98.0

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

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	\
0	False	True	False	False	False	False	False	False	False	
1	False	True	False	False	False	False	False	False	False	
2	False	True	False	False	False	False	False	False	False	
3	False	True	False	False	False	False	False	False	False	
4	False	True	False	False	False	False	False	False	False	
5	False	True	False	False	False	False	False	True	False	
6	False	True	False	False	False	True	False	False	False	
7	False	True	False	False	False	False	False	False	False	
8	False	True	False	False	False	False	False	False	False	
9	False	True	False	False	False	False	False	False	False	
10	False	True	False	True	False	False	False	False	False	
11	False	True	False	False	False	False	False	False	False	
12	False	True	False	False	False	False	True	False	False	
13	False	True	False	False	False	False	False	False	False	
14	False	True	False	False	False	False	False	True	False	
15	False	True	False	False	False	False	False	False	False	
16	False	True	False	False	False	False	False	False	False	
17	False	True	False	False	True	False	False	False	False	
18	False	True	False	False	False	False	False	False	False	
19	False	True	False	False	False	False	False	False	False	
20	False	True	False	False	False	False	False	False	False	
21	False	True	False	False	False	False	False	True	False	
22	False	True	False	False	False	False	False	False	False	
23	False	True	False	True	False	False	False	False	False	
24	False	True	False	True	False	False	False	False	False	
25	False	True	False	False	False	False	False	False	False	
26	False	True	False	False	False	False	False	False	False	
27	False	True	False	False	False	False	False	False	False	

```
28    False  True  False       False  False  False  False  False  False
29    False  True  False       False  False  False  False  False  False

    Percentage  Result  Total Marks
0        False   False         False
1        False   False         False
2        False   False         False
3        False   False         False
4        False   False         False
5        False   False         False
6        False   False         False
7        False   False         False
8        False   False         False
9        False   False         False
10       False   False         False
11       False   False         False
12       False   False         False
13       False   False         False
14       False   False         False
15       False   False         False
16       False   False         False
17       False   False         False
18       False   False         False
19       False   False         False
20       False   False         False
21       False   False         False
22       False   False         False
23       False   False         False
24       False   False         False
25       False   False         False
26       False   False         False
27       False   False         False
28       False   False         False
29       False   False         False
```

[ ]:
```
df.isna()
```

[ ]:
```
    Roll.no  Name    Sem  Attendance   Sub1   Sub2   Sub3   Sub4   Sub5  \
0     False  True  False       False  False  False  False  False  False
1     False  True  False       False  False  False  False  False  False
2     False  True  False       False  False  False  False  False  False
3     False  True  False       False  False  False  False  False  False
4     False  True  False       False  False  False  False  False  False
5     False  True  False       False  False  False  False   True  False
6     False  True  False       False  False   True  False  False  False
7     False  True  False       False  False  False  False  False  False
8     False  True  False       False  False  False  False  False  False
```



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

    Percentage  Result  Total Marks
0        False   False         False
1        False   False         False
2        False   False         False
3        False   False         False
4        False   False         False
5        False   False         False
6        False   False         False
7        False   False         False
8        False   False         False
9        False   False         False
10       False   False         False
11       False   False         False
12       False   False         False
13       False   False         False
14       False   False         False
15       False   False         False
16       False   False         False
17       False   False         False
18       False   False         False
19       False   False         False
20       False   False         False
21       False   False         False
22       False   False         False
23       False   False         False
```

```
24      False   False       False
25      False   False       False
26      False   False       False
27      False   False       False
28      False   False       False
29      False   False       False
```

[ ]:
```
df.isnull().any()
```

[ ]:
```
Roll.no        False
Name            True
Sem            False
Attendance      True
Sub1            True
Sub2            True
Sub3            True
Sub4            True
Sub5           False
Percentage     False
Result         False
Total Marks    False
dtype: bool
```

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

[ ]:
```
Roll.no         0
Name           30
Sem             0
Attendance      3
Sub1            1
Sub2            1
Sub3            1
Sub4            3
Sub5            0
Percentage      0
Result          0
Total Marks     0
dtype: int64
```

[ ]:
```
df.Attendance.isnull().sum()
```

[ ]:
```
3
```

[ ]:
```
cols_with_na = []
for col in df.columns:
  if df[col].isna().any():
    cols_with_na.append(col)
```



```
cols_with_na
```

[ ]:
```
['Name', 'Attendance', 'Sub1', 'Sub2', 'Sub3', 'Sub4']
```

Handling missing values using dropna(),fillna(),replace()

[ ]:
```
df.replace(np.nan,value=0)
```

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
0	1	0.0	5	33.0	3.0	94.0	92.0	50.0	1	60.00	
1	2	0.0	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	0.0	5	92.0	74.0	46.0	10.0	52.0	3	46.25	
3	4	0.0	6	5.0	9.0	81.0	39.0	0.0	32	40.25	
4	5	0.0	5	23.0	98.0	34.0	93.0	3.0	32	65.00	
5	6	0.0	6	26.0	56.0	70.0	12.0	0.0	65	50.75	
6	7	0.0	5	16.0	92.0	0.0	18.0	45.0	3	39.50	
7	8	0.0	6	39.0	71.0	85.0	-21.0	65.0	48	62.00	
8	9	0.0	6	24.0	31.0	48.0	6.0	84.0	74	60.75	
9	10	0.0	5	60.0	53.0	6.0	16.0	81.0	180	59.00	
10	11	0.0	6	0.0	51.0	8.0	67.0	2.0	14	35.50	
11	12	0.0	6	21.0	68.0	68.0	28.0	73.0	26	65.75	
12	13	0.0	6	57.0	32.0	99.0	0.0	36.0	93	65.00	
13	14	0.0	5	100.0	19.0	71.0	40.0	30.0	13	43.25	
14	15	0.0	5	24.0	47.0	-34.0	43.0	0.0	70	31.50	
15	16	0.0	6	18.0	38.0	17.0	57.0	69.0	48	57.25	
16	17	0.0	6	63.0	72.0	13.0	6.0	68.0	6	41.25	
17	18	0.0	5	55.0	0.0	7.0	12.0	96.0	42	39.25	
18	19	0.0	5	92.0	60.0	35.0	12.0	0.0	45	38.00	
19	20	0.0	6	66.0	28.0	10.0	196.0	94.0	55	70.75	
20	21	0.0	6	69.0	86.0	97.0	73.0	38.0	37	82.75	
21	22	0.0	5	90.0	78.0	43.0	14.0	0.0	91	56.50	
22	23	0.0	5	52.0	74.0	94.0	48.0	53.0	16	71.25	
23	24	0.0	6	0.0	30.0	40.0	41.0	21.0	74	51.50	
24	25	0.0	6	0.0	76.0	39.0	63.0	23.0	19	55.00	
25	26	0.0	5	70.0	83.0	34.0	90.0	42.0	49	74.50	
26	27	0.0	6	40.0	14.0	13.0	42.0	43.0	20	33.00	
27	28	0.0	6	71.0	45.0	18.0	25.0	31.0	27	36.50	
28	29	0.0	5	79.0	17.0	60.0	18.0	85.0	0	45.00	
29	30	0.0	6	35.0	21.0	48.0	47.0	75.0	9	50.00	

	Result	Total Marks
0	Pass	240
1	Pass	307
2	Pass	185
3	Pass	161
4	Pass	260
5	Pass	203

6	Pass	158
7	Pass	248
8	Pass	243
9	Pass	336
10	fail	142
11	Pass	263
12	Pass	260
13	Pass	173
14	fail	126
15	Pass	229
16	Pass	165
17	Pass	157
18	Pass	152
19	Pass	383
20	Pass	331
21	Pass	226
22	Pass	285
23	Pass	206
24	Pass	220
25	Pass	298
26	fail	132
27	Pass	146
28	Pass	180
29	Pass	200

[ ]:
```
df.fillna(1)
```

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
0	1	1.0	5	33.0	3.0	94.0	92.0	50.0	1	60.00	
1	2	1.0	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	1.0	5	92.0	74.0	46.0	10.0	52.0	3	46.25	
3	4	1.0	6	5.0	9.0	81.0	39.0	0.0	32	40.25	
4	5	1.0	5	23.0	98.0	34.0	93.0	3.0	32	65.00	
5	6	1.0	6	26.0	56.0	70.0	12.0	1.0	65	50.75	
6	7	1.0	5	16.0	92.0	1.0	18.0	45.0	3	39.50	
7	8	1.0	6	39.0	71.0	85.0	-21.0	65.0	48	62.00	
8	9	1.0	6	24.0	31.0	48.0	6.0	84.0	74	60.75	
9	10	1.0	5	60.0	53.0	6.0	16.0	81.0	180	59.00	
10	11	1.0	6	1.0	51.0	8.0	67.0	2.0	14	35.50	
11	12	1.0	6	21.0	68.0	68.0	28.0	73.0	26	65.75	
12	13	1.0	6	57.0	32.0	99.0	1.0	36.0	93	65.00	
13	14	1.0	5	100.0	19.0	71.0	40.0	30.0	13	43.25	
14	15	1.0	5	24.0	47.0	-34.0	43.0	1.0	70	31.50	
15	16	1.0	6	18.0	38.0	17.0	57.0	69.0	48	57.25	
16	17	1.0	6	63.0	72.0	13.0	6.0	68.0	6	41.25	
17	18	1.0	5	55.0	1.0	7.0	12.0	96.0	42	39.25	
18	19	1.0	5	92.0	60.0	35.0	12.0	0.0	45	38.00	

19	20	1.0	6	66.0	28.0	10.0	196.0	94.0	55	70.75
20	21	1.0	6	69.0	86.0	97.0	73.0	38.0	37	82.75
21	22	1.0	5	90.0	78.0	43.0	14.0	1.0	91	56.50
22	23	1.0	5	52.0	74.0	94.0	48.0	53.0	16	71.25
23	24	1.0	6	1.0	30.0	40.0	41.0	21.0	74	51.50
24	25	1.0	6	1.0	76.0	39.0	63.0	23.0	19	55.00
25	26	1.0	5	70.0	83.0	34.0	90.0	42.0	49	74.50
26	27	1.0	6	40.0	14.0	13.0	42.0	43.0	20	33.00
27	28	1.0	6	71.0	45.0	18.0	25.0	31.0	27	36.50
28	29	1.0	5	79.0	17.0	60.0	18.0	85.0	0	45.00
29	30	1.0	6	35.0	21.0	48.0	47.0	75.0	9	50.00

	Result	Total Marks
0	Pass	240
1	Pass	307
2	Pass	185
3	Pass	161
4	Pass	260
5	Pass	203
6	Pass	158
7	Pass	248
8	Pass	243
9	Pass	336
10	fail	142
11	Pass	263
12	Pass	260
13	Pass	173
14	fail	126
15	Pass	229
16	Pass	165
17	Pass	157
18	Pass	152
19	Pass	383
20	Pass	331
21	Pass	226
22	Pass	285
23	Pass	206
24	Pass	220
25	Pass	298
26	fail	132
27	Pass	146
28	Pass	180
29	Pass	200

[ ]:
```
df['Sub1']=df['Sub1'].fillna(df['Sub1'].mean())
```

[ ]:
```
df.head(10)
```

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	Percentage	\
0	1	NaN	5	33.0	3.0	94.0	92.0	50.0	1	60.00	
1	2	NaN	6	36.0	56.0	74.0	99.0	75.0	3	76.75	
2	3	NaN	5	92.0	74.0	46.0	10.0	52.0	3	46.25	
3	4	NaN	6	5.0	9.0	81.0	39.0	0.0	32	40.25	
4	5	NaN	5	23.0	98.0	34.0	93.0	3.0	32	65.00	
5	6	NaN	6	26.0	56.0	70.0	12.0	NaN	65	50.75	
6	7	NaN	5	16.0	92.0	NaN	18.0	45.0	3	39.50	
7	8	NaN	6	39.0	71.0	85.0	-21.0	65.0	48	62.00	
8	9	NaN	6	24.0	31.0	48.0	6.0	84.0	74	60.75	
9	10	NaN	5	60.0	53.0	6.0	16.0	81.0	180	59.00	

	Result	Total Marks
0	Pass	240
1	Pass	307
2	Pass	185
3	Pass	161
4	Pass	260
5	Pass	203
6	Pass	158
7	Pass	248
8	Pass	243
9	Pass	336

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

[ ]:
```
Empty DataFrame
Columns: [Roll.no, Name, Sem, Attendance, Sub1, Sub2, Sub3, Sub4, Sub5,
Percentage, Result, Total Marks]
Index: []
```

[ ]:
```
df.dropna(how="all")
```

[ ]:

	Roll.no	Name	Sem	Attendance	Sub1	Sub2	Sub3	Sub4	Sub5	\
0	1	NaN	5	33.0	3.000000	94.0	92.0	50.0	1	
1	2	NaN	6	36.0	56.000000	74.0	99.0	75.0	3	
2	3	NaN	5	92.0	74.000000	46.0	10.0	52.0	3	
3	4	NaN	6	5.0	9.000000	81.0	39.0	0.0	32	
4	5	NaN	5	23.0	98.000000	34.0	93.0	3.0	32	
5	6	NaN	6	26.0	56.000000	70.0	12.0	NaN	65	
6	7	NaN	5	16.0	92.000000	NaN	18.0	45.0	3	
7	8	NaN	6	39.0	71.000000	85.0	-21.0	65.0	48	
8	9	NaN	6	24.0	31.000000	48.0	6.0	84.0	74	
9	10	NaN	5	60.0	53.000000	6.0	16.0	81.0	180	
10	11	NaN	6	NaN	51.000000	8.0	67.0	2.0	14	
11	12	NaN	6	21.0	68.000000	68.0	28.0	73.0	26	
12	13	NaN	6	57.0	32.000000	99.0	NaN	36.0	93	

13	14	NaN	5	100.0	19.000000	71.0	40.0	30.0	13
14	15	NaN	5	24.0	47.000000	-34.0	43.0	NaN	70
15	16	NaN	6	18.0	38.000000	17.0	57.0	69.0	48
16	17	NaN	6	63.0	72.000000	13.0	6.0	68.0	6
17	18	NaN	5	55.0	51.103448	7.0	12.0	96.0	42
18	19	NaN	5	92.0	60.000000	35.0	12.0	0.0	45
19	20	NaN	6	66.0	28.000000	10.0	196.0	94.0	55
20	21	NaN	6	69.0	86.000000	97.0	73.0	38.0	37
21	22	NaN	5	90.0	78.000000	43.0	14.0	NaN	91
22	23	NaN	5	52.0	74.000000	94.0	48.0	53.0	16
23	24	NaN	6	NaN	30.000000	40.0	41.0	21.0	74
24	25	NaN	6	NaN	76.000000	39.0	63.0	23.0	19
25	26	NaN	5	70.0	83.000000	34.0	90.0	42.0	49
26	27	NaN	6	40.0	14.000000	13.0	42.0	43.0	20
27	28	NaN	6	71.0	45.000000	18.0	25.0	31.0	27
28	29	NaN	5	79.0	17.000000	60.0	18.0	85.0	0
29	30	NaN	6	35.0	21.000000	48.0	47.0	75.0	9

	Percentage	Result	Total Marks
0	60.00	Pass	240
1	76.75	Pass	307
2	46.25	Pass	185
3	40.25	Pass	161
4	65.00	Pass	260
5	50.75	Pass	203
6	39.50	Pass	158
7	62.00	Pass	248
8	60.75	Pass	243
9	59.00	Pass	336
10	35.50	fail	142
11	65.75	Pass	263
12	65.00	Pass	260
13	43.25	Pass	173
14	31.50	fail	126
15	57.25	Pass	229
16	41.25	Pass	165
17	39.25	Pass	157
18	38.00	Pass	152
19	70.75	Pass	383
20	82.75	Pass	331
21	56.50	Pass	226
22	71.25	Pass	285
23	51.50	Pass	206
24	55.00	Pass	220
25	74.50	Pass	298
26	33.00	fail	132
27	36.50	Pass	146

28	45.00	Pass		180
29	50.00	Pass		200

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

[ ]:

	Roll.no	Sem	Sub1	Sub5	Percentage	Result	Total Marks
0	1	5	3.000000	1	60.00	Pass	240
1	2	6	56.000000	3	76.75	Pass	307
2	3	5	74.000000	3	46.25	Pass	185
3	4	6	9.000000	32	40.25	Pass	161
4	5	5	98.000000	32	65.00	Pass	260
5	6	6	56.000000	65	50.75	Pass	203
6	7	5	92.000000	3	39.50	Pass	158
7	8	6	71.000000	48	62.00	Pass	248
8	9	6	31.000000	74	60.75	Pass	243
9	10	5	53.000000	180	59.00	Pass	336
10	11	6	51.000000	14	35.50	fail	142
11	12	6	68.000000	26	65.75	Pass	263
12	13	6	32.000000	93	65.00	Pass	260
13	14	5	19.000000	13	43.25	Pass	173
14	15	5	47.000000	70	31.50	fail	126
15	16	6	38.000000	48	57.25	Pass	229
16	17	6	72.000000	6	41.25	Pass	165
17	18	5	51.103448	42	39.25	Pass	157
18	19	5	60.000000	45	38.00	Pass	152
19	20	6	28.000000	55	70.75	Pass	383
20	21	6	86.000000	37	82.75	Pass	331
21	22	5	78.000000	91	56.50	Pass	226
22	23	5	74.000000	16	71.25	Pass	285
23	24	6	30.000000	74	51.50	Pass	206
24	25	6	76.000000	19	55.00	Pass	220
25	26	5	83.000000	49	74.50	Pass	298
26	27	6	14.000000	20	33.00	fail	132
27	28	6	45.000000	27	36.50	Pass	146
28	29	5	17.000000	0	45.00	Pass	180
29	30	6	21.000000	9	50.00	Pass	200

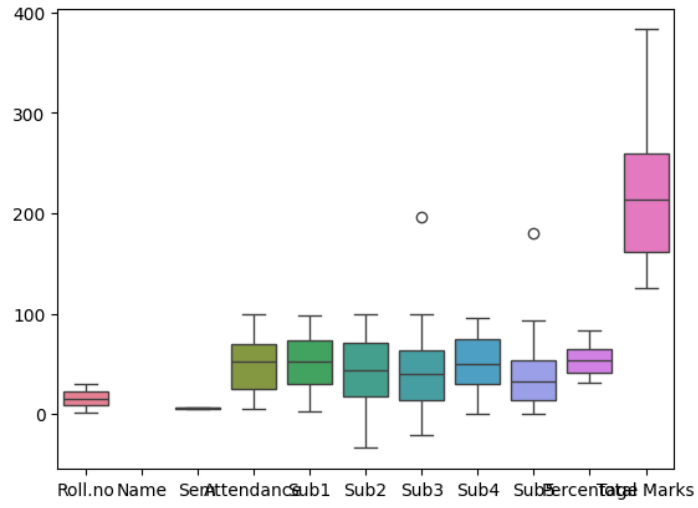
##Outlier Handling

[ ]:
```
import seaborn as sns
```

[ ]:
```
import matplotlib.pyplot as plt
```

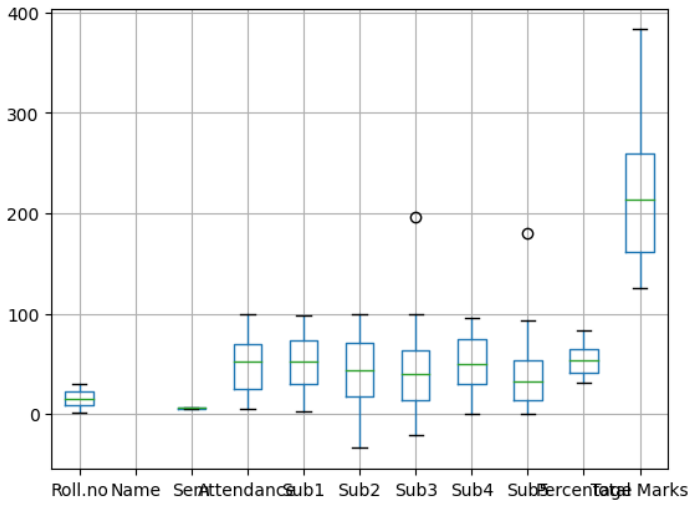
[ ]:
```
sns.boxplot(df)
```

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

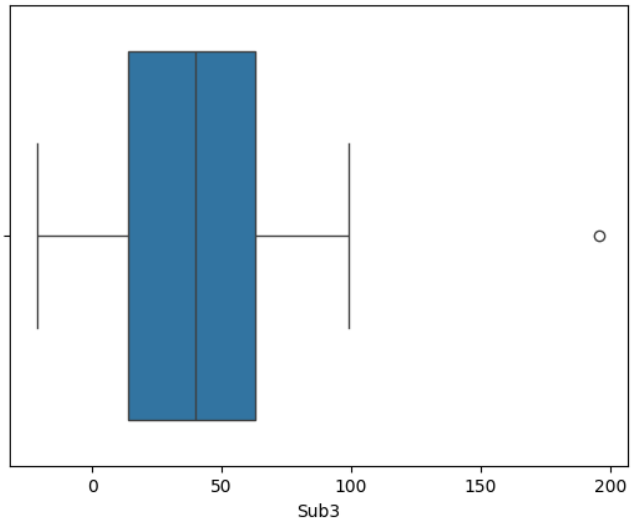
[ ]:
```
df.boxplot()
```

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



[ ]:
```
sns.boxplot(x=df.Sub3)
```

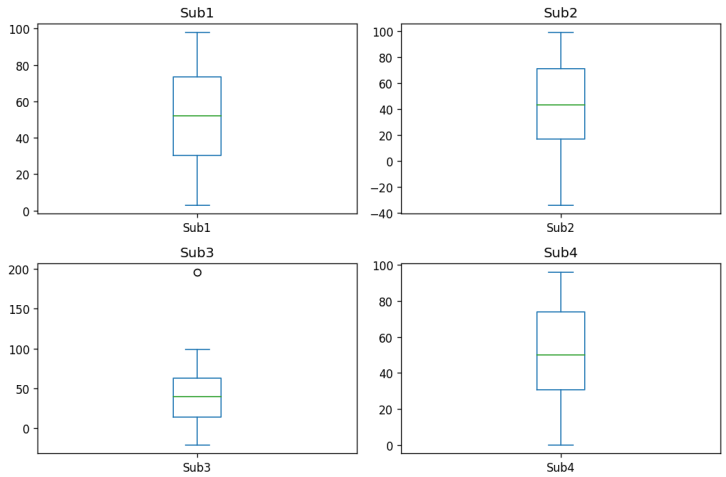
[ ]: <Axes: xlabel='Sub3'>



[ ]:
```
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = (9, 6)
df_list = ['Roll.no','Sub1','Sub2','Sub3','Sub4']
fig, axes = plt.subplots(2, 2)
fig.set_dpi(120)

count = 1  # Start from 1 as the first element in df_list is 'Roll.no'
for r in range(2):
    for c in range(2):
        df[df_list[count]].plot(kind='box', ax=axes[r, c])
        axes[r, c].set_title(df_list[count])
        count += 1

plt.tight_layout()
plt.show()
```



#Quantile Range

```
[ ]: Q1 = df['Sub5'].quantile(0.25)
     Q3 = df['Sub5'].quantile(0.75)
     IQR = Q3 - Q1
     Lower_limit = Q1 - 1.5 * IQR
     Upper_limit = Q3 + 1.5 * IQR
     print(f'Q1 = {Q1}, Q3 = {Q3}, IQR = {IQR}, Lower_limit = {Lower_limit},␣
      ↪Upper_limit = {Upper_limit}')
```

```
Q1 = 13.25, Q3 = 53.5, IQR = 40.25, Lower_limit = -47.125, Upper_limit = 113.875
```

```
[ ]: df[(df['Sub5'] < Lower_limit) | (df['Sub5'] > Upper_limit)]
```

```
[ ]:    Roll.no  Name  Sem  Attendance  Sub1  Sub2  Sub3  Sub4  Sub5  Percentage  \
     9       10   NaN    5        60.0  53.0   6.0  16.0  81.0   180        59.0

       Result  Total Marks
     9   Pass          336
```

#Handling Outliers

```
[ ]: outliers=[]
     for i in df.Sub5:
       if i<Lower_limit or i>Upper_limit:
         outliers.append(i)
     print("outliers are",outliers)
```

```
outliers are [180]
```

```
[ ]: Upper_limit
```

```
[ ]: 113.875
```

```
[ ]: Lower_limit
```

```
[ ]: -47.125
```

```
[ ]: df[df.Sub5<Lower_limit].index
```

```
[ ]: Int64Index([], dtype='int64')
```