```
import pandas as pd
import numpy as np

a = pd.read_csv("/content/new_dataset.csv")

a.describe();

Start coding or generate with AI.
```

a

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	NaN	WT	Pass	23.0
1	Ishani	12.0	DS	Pass	67.0
2	Akshada	13.0	CN	Pass	34.0
3	Sanika	14.0	DM	Pass	23.0
4	Titiksha	15.0	DS	Pass	56.0
5	Sneha	16.0	TOC	Pass	34.0
6	Arindita	90.0	Lp	NaN	NaN
7	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	NaN

```
a.head()
```

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	NaN	WT	Pass	23.0
1	Ishani	12.0	DS	Pass	67.0
2	Akshada	13.0	CN	Pass	34.0
3	Sanika	14.0	DM	Pass	23.0
4	Titiksha	15.0	DS	Pass	56.0

```
a.tail()
```

	Name	Roll_no	Sub	Result	Attendance
5	Sneha	16	TOC	Pass	34.0
6	Arindita	90	Lp	NaN	NaN
7	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	NaN

```
a.shape
```

```
(10, 5)
```

```
a.dtypes
```

```
Name          object
Roll_no       object
Sub           object
Result        object
Attendance   float64
dtype: object
```

```
a.columns
```

```
Index(['Name', 'Roll_no', 'Sub', 'Result', 'Attendance'], dtype='object')
```

a[0:3]

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	#NAME?	WT	Pass	23.0
1	Ishani	12	DS	Pass	67.0
2	Akshada	13	CN	Pass	34.0

a.loc[0:3]

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	#NAME?	WT	Pass	23.0
1	Ishani	12	DS	Pass	67.0
2	Akshada	13	CN	Pass	34.0
3	Sanika	14	DM	Pass	23.0

a.iloc[0:3]

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	#NAME?	WT	Pass	23.0
1	Ishani	12	DS	Pass	67.0
2	Akshada	13	CN	Pass	34.0

a

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	NaN	WT	Pass	23.000000
1	Ishani	12.0	DS	Pass	67.000000
2	Akshada	13.0	CN	Pass	34.000000
3	Sanika	14.0	DM	Pass	23.000000
4	Titiksha	15.0	DS	Pass	56.000000
5	Sneha	16.0	TOC	Pass	34.000000
6	Arindita	90.0	Lp	NaN	26.666667
7	NaN	NaN	NaN	NaN	26.666667
8	NaN	NaN	NaN	NaN	26.666667
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	26.666667

Handling Null values

a.isnull()

	Name	Roll_no	Sub	Result	Attendance
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	True	True
7	True	True	True	True	True
8	True	True	True	True	True
9	True	True	True	False	True

```
a.isna()
```

	Name	Roll_no	Sub	Result	Attendance
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	True	True
7	True	True	True	True	True
8	True	True	True	True	True
9	True	True	True	False	True

```
a.isnull().any()
```

```
Name          True
Roll_no       True
Sub           True
Result        True
Attendance    True
dtype: bool
```

```
a.isnull().sum()
```

```
Name          3
Roll_no       3
Sub           3
Result        3
Attendance    4
dtype: int64
```

```
a.Attendance.isnull().sum()
```

```
4
```

Use Replace method

```
#a['Attendance']=a['Attendance'].fillna(a['Roll_no'].mean())
```

```
a1 = pd.read_csv("/content/new_dataset.csv")
```

```
a1['Attendnce'] = a1['Attendance'].fillna(a['Roll_no'].min())
```

```
a1
```

	Name	Roll_no	Sub	Result	Attendance	Attendnce
0	Sakshi	NaN	WT	Pass	23.0	23.0
1	Ishani	12.0	DS	Pass	67.0	67.0
2	Akshada	13.0	CN	Pass	34.0	34.0
3	Sanika	14.0	DM	Pass	23.0	23.0
4	Titiksha	15.0	DS	Pass	56.0	56.0
5	Sneha	16.0	TOC	Pass	34.0	34.0
6	Arindita	90.0	Lp	NaN	NaN	12.0
7	NaN	NaN	NaN	NaN	NaN	12.0
8	NaN	NaN	NaN	NaN	NaN	12.0
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	NaN	12.0

```
a
```

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	NaN	WT	Pass	23.0
1	Ishani	12.0	DS	Pass	67.0
2	Akshada	13.0	CN	Pass	34.0
3	Sanika	14.0	DM	Pass	23.0
4	Titiksha	15.0	DS	Pass	56.0
5	Sneha	16.0	TOC	Pass	34.0
6	Arindita	90.0	Lp	NaN	90.0
7	NaN	NaN	NaN	NaN	90.0
8	NaN	NaN	NaN	NaN	90.0
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	90.0

```
a['Attendance'] = a['Attendance'].fillna(a['Roll_no'].median())
```

```
a
```

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	NaN	WT	Pass	23.000000
1	Ishani	12.0	DS	Pass	67.000000
2	Akshada	13.0	CN	Pass	34.000000
3	Sanika	14.0	DM	Pass	23.000000
4	Titiksha	15.0	DS	Pass	56.000000
5	Sneha	16.0	TOC	Pass	34.000000
6	Arindita	90.0	Lp	NaN	26.666667
7	NaN	NaN	NaN	NaN	26.666667
8	NaN	NaN	NaN	NaN	26.666667
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	26.666667

```
a['Attendance'] = a['Attendance'].fillna(a['Roll_no'].max())
```

```
a
```

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	NaN	WT	Pass	23.0
1	Ishani	12.0	DS	Pass	67.0
2	Akshada	13.0	CN	Pass	34.0
3	Sanika	14.0	DM	Pass	23.0
4	Titiksha	15.0	DS	Pass	56.0
5	Sneha	16.0	TOC	Pass	34.0
6	Arindita	90.0	Lp	NaN	90.0
7	NaN	NaN	NaN	NaN	90.0
8	NaN	NaN	NaN	NaN	90.0
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	90.0

```
a.replace(np.nan,value=0)
```

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	#NAME?	WT	Pass	23.0
1	Ishani	12	DS	Pass	67.0
2	Akshada	13	CN	Pass	34.0
3	Sanika	14	DM	Pass	23.0
4	Titiksha	15	DS	Pass	56.0
5	Sneha	16	TOC	Pass	34.0
6	Arindita	90	Lp	0	0.0
7	0	0	0	0	0.0
8	0	0	0	0	0.0
9	0	0	0	: RANDBETWEEN(80, 90)	0.0

Use Fill method

```
a.fillna(1);
```

```
a
```

	Name	Roll_no	Sub	Result	Attendance
0	Sakshi	#NAME?	WT	Pass	23.0
1	Ishani	12	DS	Pass	67.0
2	Akshada	13	CN	Pass	34.0
3	Sanika	14	DM	Pass	23.0
4	Titiksha	15	DS	Pass	56.0
5	Sneha	16	TOC	Pass	34.0
6	Arindita	90	Lp	NaN	NaN
7	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	NaN

Filling missing values using mean, median,max, min and standard deviation of that column

```
a
```

	Name	Roll_no	Sub	Result	Attendance	Attenadance
0	Sakshi	#NAME?	WT	Pass	23.0	23.0
1	Ishani	12	DS	Pass	67.0	67.0
2	Akshada	13	CN	Pass	34.0	34.0
3	Sanika	14	DM	Pass	23.0	23.0
4	Titiksha	15	DS	Pass	56.0	56.0
5	Sneha	16	TOC	Pass	34.0	34.0
6	Arindita	90	Lp	NaN	NaN	39.5
7	NaN	NaN	NaN	NaN	NaN	39.5
8	NaN	NaN	NaN	NaN	NaN	39.5
9	NaN	NaN	NaN	: RANDBETWEEN(80, 90)	NaN	39.5

```
a['Attendance'] = a['Attendance'].fillna(a['Attendance'].median)
```

```
a
```

	Name	Roll_no	Sub	Result	Attendance	Attenad
0	Sakshi	#NAME?	WT	Pass	23.0	
1	Ishani	12	DS	Pass	67.0	
2	Akshada	13	CN	Pass	34.0	
3	Sanika	14	DM	Pass	23.0	
4	Titiksha	15	DS	Pass	56.0	
5	Sneha	16	TOC	Pass	34.0	
6	Arindita	90	Lp	NaN	<bound method NDFrame._add_numeric_operations....	
7	NaN	NaN	NaN	NaN	<bound method NDFrame._add_numeric_operations....	

```
a.dropna()
```

	Name	Roll_no	Sub	Result	Attendance
1	Ishani	12.0	DS	Pass	67.0
2	Akshada	13.0	CN	Pass	34.0
3	Sanika	14.0	DM	Pass	23.0
4	Titiksha	15.0	DS	Pass	56.0
5	Sneha	16.0	TOC	Pass	34.0

```
a.dropna(axis=1)
```

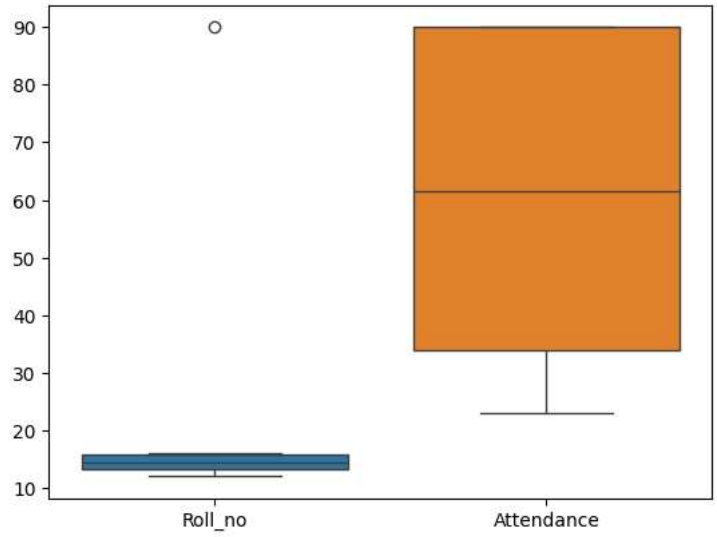
	Attendance
0	23.0
1	67.0
2	34.0
3	23.0
4	56.0
5	34.0
6	90.0
7	90.0
8	90.0
9	90.0

Outlier

```
#Import libraries
import seaborn as sns
import matplotlib.pyplot as plt
```

```
sns.boxplot(a)
```

```
<Axes: >
```

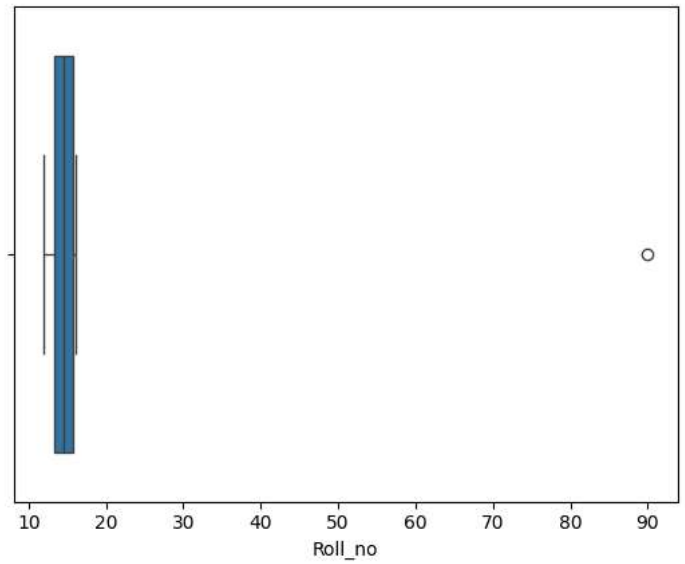



```
a.boxplot()
```

```
<Axes: >
```

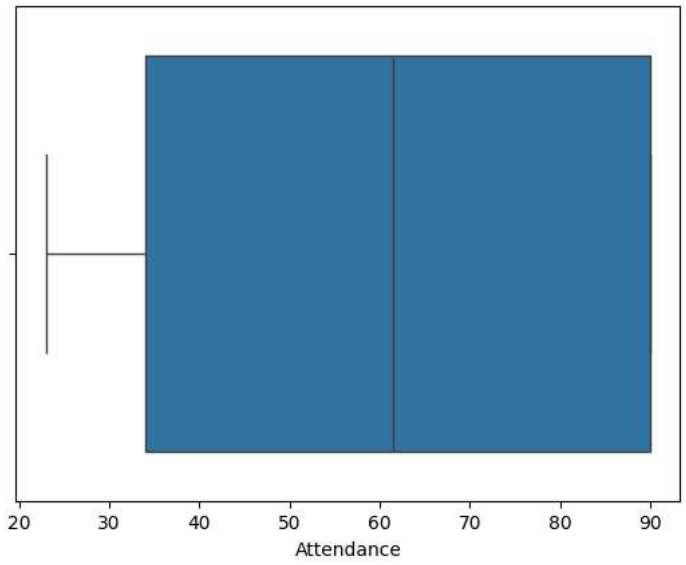

```
sns.boxplot(x=a.Roll_no)
```

```
<Axes: xlabel='Roll_no'>
```




```
sns.boxplot(x=a.Attendance)
```

```
<Axes: xlabel='Attendance'>
```



Removing Outlier i) Data cleaning process ii)Outliers are the observations in a dataset that deviate significantly from the rest of the data iii) Outliers can sometimes indicate errors or anomalies in the data. iv) The majority of the values in this feature range between 4.5−6.5 feet, but there is one value with 10 feet. This value would be considered an outlier v)Outliers can be drawn by boxplot method

```
Q1 = a['Roll_no'].quantile(0.25)
Q3 = a['Roll_no'].quantile(0.75)
IQR = Q3 - Q1
```