In [23]:
```
import pandas as pd
data=pd.read_csv("C:\\Users\\swati\\OneDrive\\D
print(type(data))
```

```
<class 'pandas.core.frame.DataFrame'>
```

In [24]:
```
data.info
```

Out[24]:
```
<bound method DataFrame.info of    SL.NO     N
AME    PLACE   AGE   SALARY   INVEST
0      1  Antarik      Goa  25.0  40000.0  100
00.0
1      2    Suvam   Odisha  29.0  50000.0  150
00.0
2      3    Swati   Odisha  20.0  60000.0   50
00.0
3      4   Soumya      Goa  23.0  70000.0
NaN
4      5  Antarik      Goa  25.0  40000.0   20
00.0
5      6    Suvam   Odisha   NaN  20000.0    5
00.0
6      7   Suresh    Assam  22.0  10000.0
NaN
7      8   Sweety  Manipur  19.0  50000.0   30
00.0
8      9   Suresh      NaN  27.0      NaN
NaN
9     10     Rosy      NaN  19.0  70000.0   80
00.0>
```

In [13]:
```
data.describe()
```

Out[13]:

	SL.NO	AGE	SALARY	INV
count	10.00000	9.000000	9.000000	7.000
mean	5.50000	23.222222	45555.555556	6214.285
std	3.02765	3.562926	20682.789410	5114.172
min	1.00000	19.000000	10000.000000	500.000
25%	3.25000	20.000000	40000.000000	2500.000
50%	5.50000	23.000000	50000.000000	5000.000
75%	7.75000	25.000000	60000.000000	9000.000
max	10.00000	29.000000	70000.000000	15000.000

In [24]:
```
data=data.drop_duplicates()
data
```

Out[24]:

	SL.NO	NAME	PLACE	AGE	SALARY	INVEST
0	1	Antarik	Goa	25.0	40000.0	10000.0
1	2	Suvam	Odisha	29.0	50000.0	15000.0
2	3	Swati	Odisha	20.0	60000.0	5000.0
3	4	Soumya	Goa	23.0	70000.0	NaN
4	5	Antarik	Goa	25.0	40000.0	2000.0
5	6	Suvam	Odisha	NaN	20000.0	500.0
6	7	Suresh	Assam	22.0	10000.0	NaN
7	8	Sweety	Manipur	19.0	50000.0	3000.0
8	9	Suresh	NaN	27.0	NaN	NaN
9	10	Rosy	NaN	19.0	70000.0	8000.0

In [26]:
```
data.isnull()
```

Out[26]:

	SL.NO	NAME	PLACE	AGE	SALARY	INVEST
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	True
4	False	False	False	False	False	False
5	False	False	False	True	False	False
6	False	False	False	False	False	True
7	False	False	False	False	False	False
8	False	False	True	False	True	True
9	False	False	True	False	False	False

In [27]:
```
data.isnull().sum()
```

Out[27]:
```
SL.NO     0
NAME      0
PLACE     2
AGE       1
SALARY    1
INVEST    3
dtype: int64
```

In [28]:
```
data.notnull()
```

Out[28]:

	SL.NO	NAME	PLACE	AGE	SALARY	INVEST
0	True	True	True	True	True	True
1	True	True	True	True	True	True
2	True	True	True	True	True	True
3	True	True	True	True	True	False
4	True	True	True	True	True	True
5	True	True	True	False	True	True
6	True	True	True	True	True	False
7	True	True	True	True	True	True
8	True	True	False	True	False	False
9	True	True	False	True	True	True

In [29]:
```
data.isnull().sum().sum()
```

Out[29]: 7

In [47]:
```
data2=data.fillna(value=0)
data2
```

Out[47]:

	SL.NO	NAME	PLACE	AGE	SALARY	INVEST
0	1	Antarik	Goa	25.0	40000.0	10000.0
1	2	Suvam	Odisha	29.0	50000.0	15000.0
2	3	Swati	Odisha	20.0	60000.0	5000.0
3	4	Soumya	Goa	23.0	70000.0	0.0
4	5	Antarik	Goa	25.0	40000.0	2000.0
5	6	Suvam	Odisha	0.0	20000.0	500.0
6	7	Suresh	Assam	22.0	10000.0	0.0
7	8	Sweety	Manipur	19.0	50000.0	3000.0
8	9	Suresh	0	27.0	0.0	0.0
9	10	Rosy	0	19.0	70000.0	8000.0

In [31]:
```
data3=data.fillna(method='pad')
data3
```

Out[31]:

	SL.NO	NAME	PLACE	AGE	SALARY	INVEST
0	1	Antarik	Goa	25.0	40000.0	10000.0
1	2	Suvam	Odisha	29.0	50000.0	15000.0
2	3	Swati	Odisha	20.0	60000.0	5000.0
3	4	Soumya	Goa	23.0	70000.0	5000.0
4	5	Antarik	Goa	25.0	40000.0	2000.0
5	6	Suvam	Odisha	25.0	20000.0	500.0
6	7	Suresh	Assam	22.0	10000.0	500.0
7	8	Sweety	Manipur	19.0	50000.0	3000.0
8	9	Suresh	Manipur	27.0	50000.0	3000.0
9	10	Rosy	Manipur	19.0	70000.0	8000.0

In [32]:

```
 # filling the null value with the next value
data4=data.fillna(method='bfill')
data4
```

Out[32]:

	SL.NO	NAME	PLACE	AGE	SALARY	INVEST
0	1	Antarik	Goa	25.0	40000.0	10000.0
1	2	Suvam	Odisha	29.0	50000.0	15000.0
2	3	Swati	Odisha	20.0	60000.0	5000.0
3	4	Soumya	Goa	23.0	70000.0	2000.0
4	5	Antarik	Goa	25.0	40000.0	2000.0
5	6	Suvam	Odisha	22.0	20000.0	500.0
6	7	Suresh	Assam	22.0	10000.0	3000.0
7	8	Sweety	Manipur	19.0	50000.0	3000.0
8	9	Suresh	NaN	27.0	70000.0	8000.0
9	10	Rosy	NaN	19.0	70000.0	8000.0

In [25]:
```
import numpy as np
from scipy import stats
```

In [28]:
```
#detect the outliers using IQR
data2.columns
```

Out[28]:
```
Index(['SL.NO', 'NAME', 'PLACE', 'AGE', 'SALAR
Y', 'INVEST'], dtype='object')
```

In [48]:
```
data2.drop(['NAME','PLACE'],axis=1,inplace=True
data2
```

Out[48]:

	SL.NO	AGE	SALARY	INVEST
0	1	25.0	40000.0	10000.0
1	2	29.0	50000.0	15000.0
2	3	20.0	60000.0	5000.0
3	4	23.0	70000.0	0.0
4	5	25.0	40000.0	2000.0
5	6	0.0	20000.0	500.0
6	7	22.0	10000.0	0.0
7	8	19.0	50000.0	3000.0
8	9	27.0	0.0	0.0
9	10	19.0	70000.0	8000.0

In [51]:
```
Q1=data2.quantile(0.25)
Q3=data2.quantile(0.75)
IQR=Q3-R1
print(IQR)
```

```
SL.NO         4.50
AGE           5.75
SALARY    32500.00
INVEST     7125.00
dtype: float64
```

In [52]:
```
data2=data2[~((data2<(Q1-1.5*IQR))|(data2>(Q3+1
data2
```

Out[52]:

	SL.NO	AGE	SALARY	INVEST
0	1	25.0	40000.0	10000.0
1	2	29.0	50000.0	15000.0
2	3	20.0	60000.0	5000.0
3	4	23.0	70000.0	0.0
4	5	25.0	40000.0	2000.0
6	7	22.0	10000.0	0.0
7	8	19.0	50000.0	3000.0
8	9	27.0	0.0	0.0
9	10	19.0	70000.0	8000.0

In [42]:
```
data2.describe()
```

Out[42]:

	SL.NO	AGE	SALARY	INVE
count	10.00000	10.00000	10.000000	10.0000
mean	5.50000	20.90000	41000.000000	4350.0000
std	3.02765	8.07534	24244.128728	5142.6862
min	1.00000	0.00000	0.000000	0.0000
25%	3.25000	19.25000	25000.000000	125.0000
50%	5.50000	22.50000	45000.000000	2500.0000
75%	7.75000	25.00000	57500.000000	7250.0000
max	10.00000	29.00000	70000.000000	15000.0000

In [ ]: