Exploratory data analysis (EDA)

it helps in understanding the dataset through various technique like

- Visualisation
- summary statics
- feature relationship

Seaborn

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

mydata={ 'name':['ram','sam','joe','asha'],
        'age':[23,22,26,47],
        'sal':[12000,4000,12000,34000],
        'exp':[2,1,3,10]
       }
df=pd.DataFrame(mydata)
df
```

	name	age	sal	exp
0	ram	23	12000	2
1	sam	22	4000	1
2	joe	26	12000	3
3	asha	47	34000	10

Histogram

```
plt.figure(figsize=(6,5))
sns.histplot(df['sal'],kde = True ,bins =2)
plt.title('Distribution of salary')
plt.show()
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

1. positive skew,large salary values
2. no outliers detected
3. Average salary is around 10000
4. majority salary values are between 5000 and 18000

```
mydata1={ 'name':['ram','sam','joe','asha'],
        'age':[23,22,26,47],
        'sal':[25000,500,6000,5000],
        'exp':[2,1,3,10]
       }
df1=pd.DataFrame(mydata1)
df1
```

	name	age	sal	exp
0	ram	23	25000	2
1	sam	22	500	1
2	joe	26	6000	3
3	asha	47	5000	10

```
plt.figure(figsize=(5,4))
sns.histplot(df1['sal'],kde = True ,bins =2)
plt.title('Distribution of salary')
plt.show()
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

1. positive skew,large salary values
2. no outliers detected
3. Average salary is around 50000
4. majority salary values are between 12000 and 25000

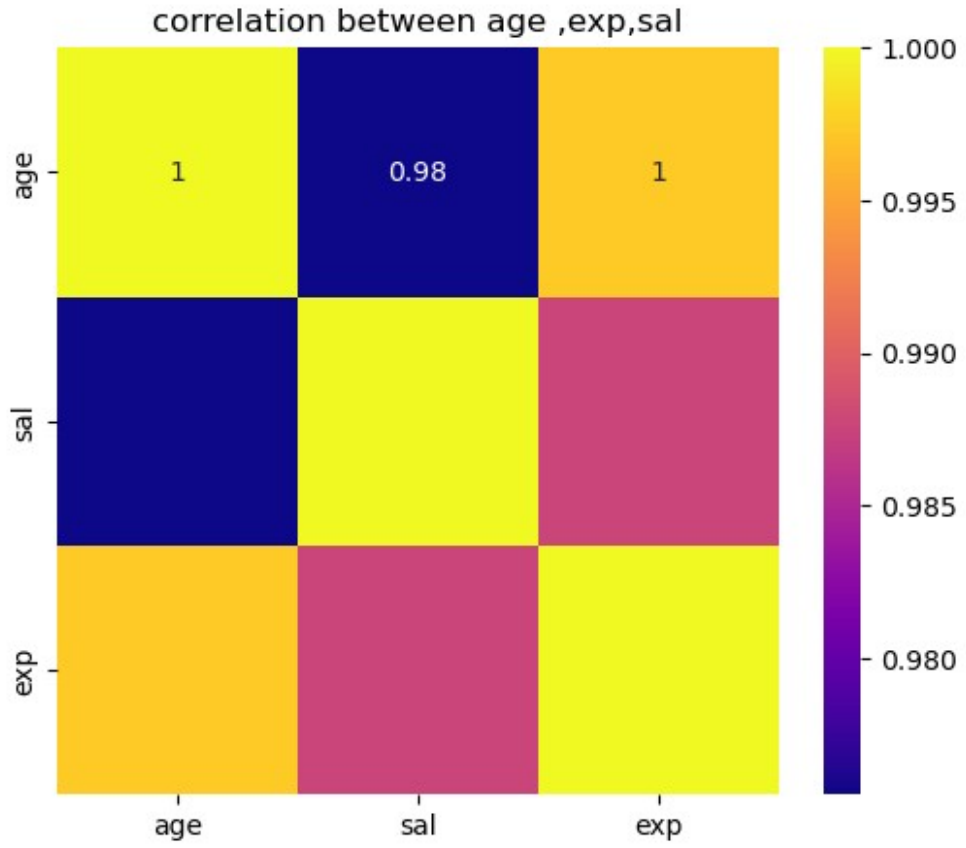
Correllaton matrix(heatmap)

```
#step 1: filter the numercial data
ndf=df.select_dtypes(include=['number'])
ndf.head()
```

	age	sal	exp
0	23	12000	2
1	22	4000	1
2	26	12000	3
3	47	34000	10

```
#step 2: heat map
plt.figure(figsize=(6,5))#rows and column
```

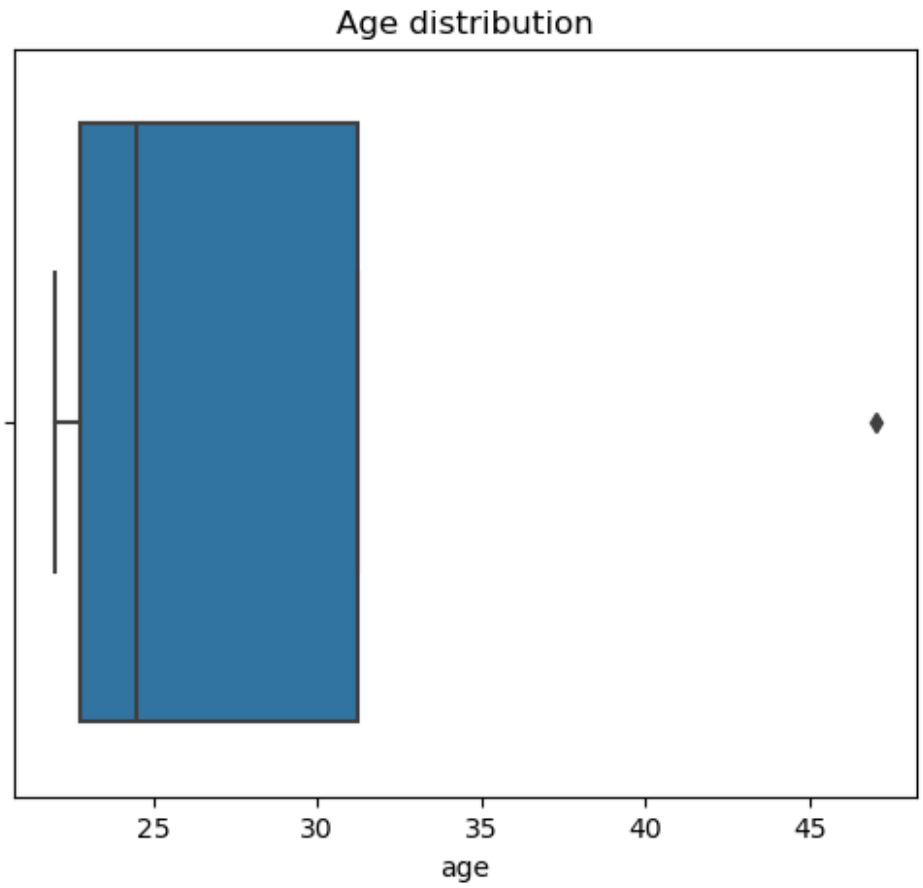
```
sns.heatmap(ndf.corr(),cmap='plasma',annot=True)#color --
plasma,coolwarm
plt.title("correlation between age ,exp,sal")
plt.show()
```



good correlation between age and experiance

poor correlation between the age and salary

```
plt.figure(figsize=(6,5))
sns.boxplot(x=df['age'])
plt.title('Age distribution')
plt.show()
```



- average age value is 25
- large value found towards right side
- abnormal/ outlier is around 45

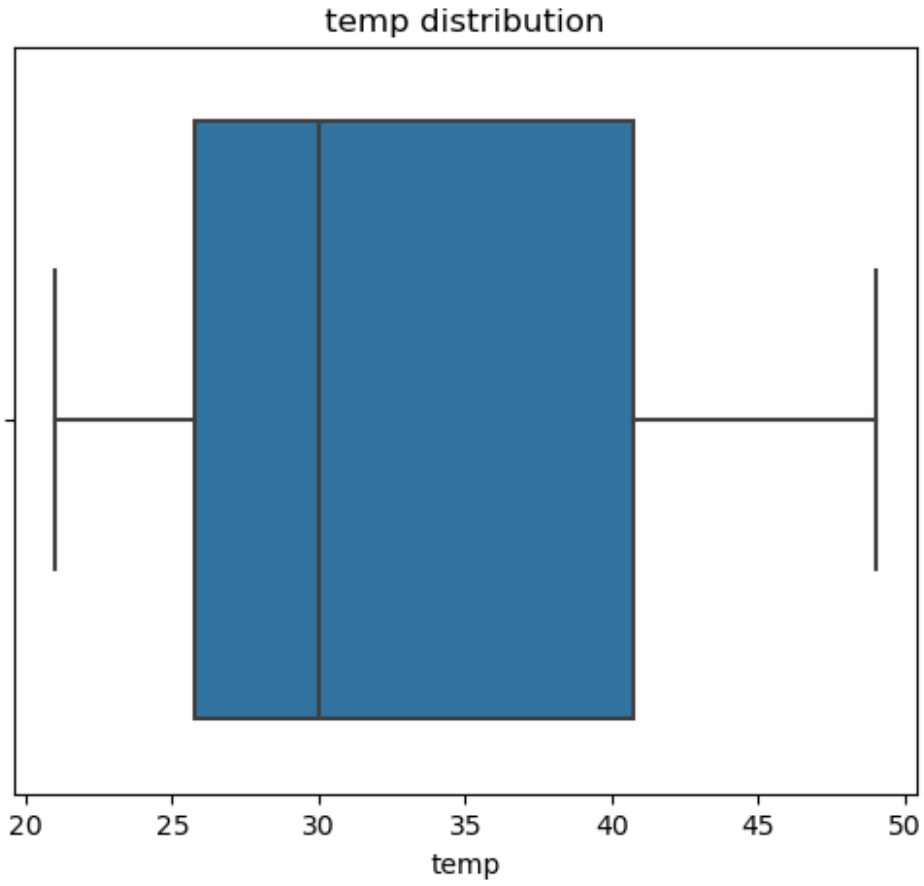
find the outliers in the following data : temp=[21,47,39,22,31,33,29,26,27,25,49,46]

```
mydata={
        'temp':[21,47,39,22,31,33,29,26,27,25,49,46]
       }
df2=pd.DataFrame(mydata)
df2
```

```
    temp
0     21
1     47
2     39
3     22
4     31
5     33
6     29
7     26
8     27
```

```
9     25
10    49
11    46

plt.figure(figsize=(6,5))
sns.boxplot(x=df2['temp'])
plt.title('temp distribution')
plt.show()
```



- average age value is 30
- large value found towards right side
- no abnormal/ outlier

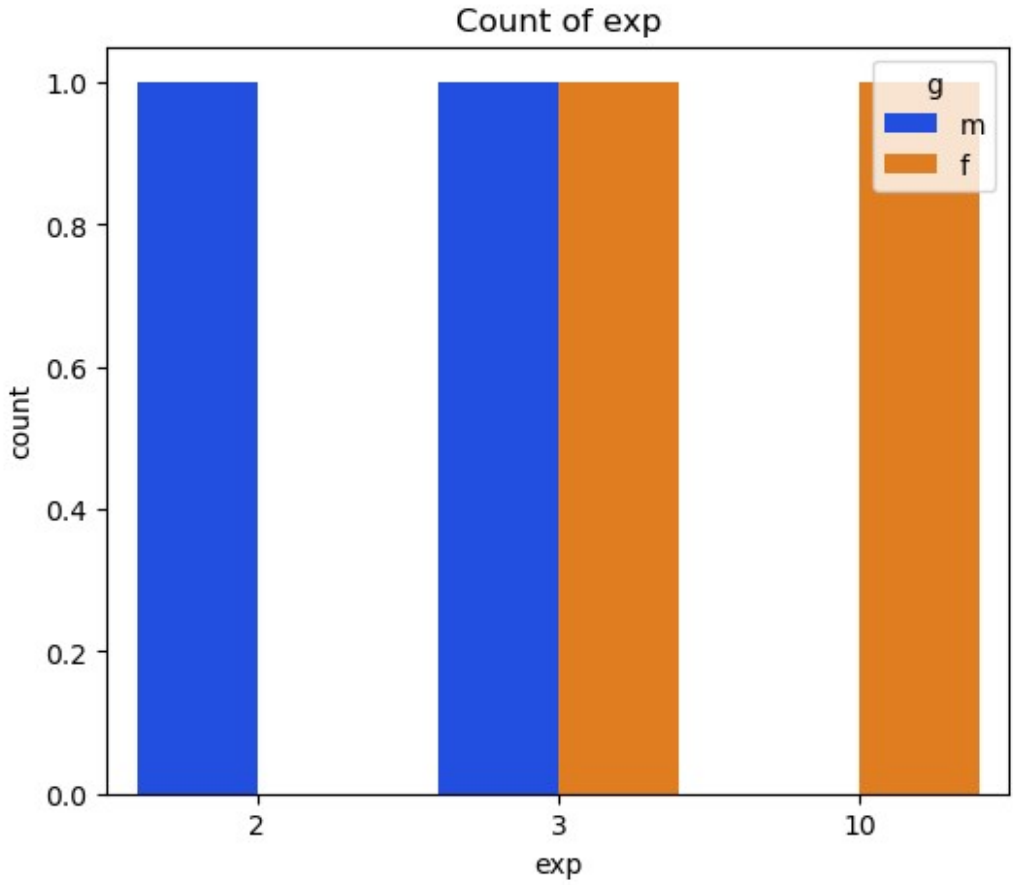
countplot

```
mydata2={ 'name':['ram','sam','joe','asha'],
        'age':[23,22,26,47],
        'sal':[12000,4000,12000,34000],
        'exp':[2,3,3,10],
        'g':['m','f','m','f']
       }
```

```
df=pd.DataFrame(mydata2)
df
```

	name	age	sal	exp	g
0	ram	23	12000	2	m
1	sam	22	4000	3	f
2	joe	26	12000	3	m
3	asha	47	34000	10	f

```
plt.figure(figsize=(6,5))#bright,pastel are the color we can use
sns.countplot(x=df['exp'],palette='bright',hue=df['g'])
plt.title('Count of exp')
plt.show()
```



- more experiance done by female

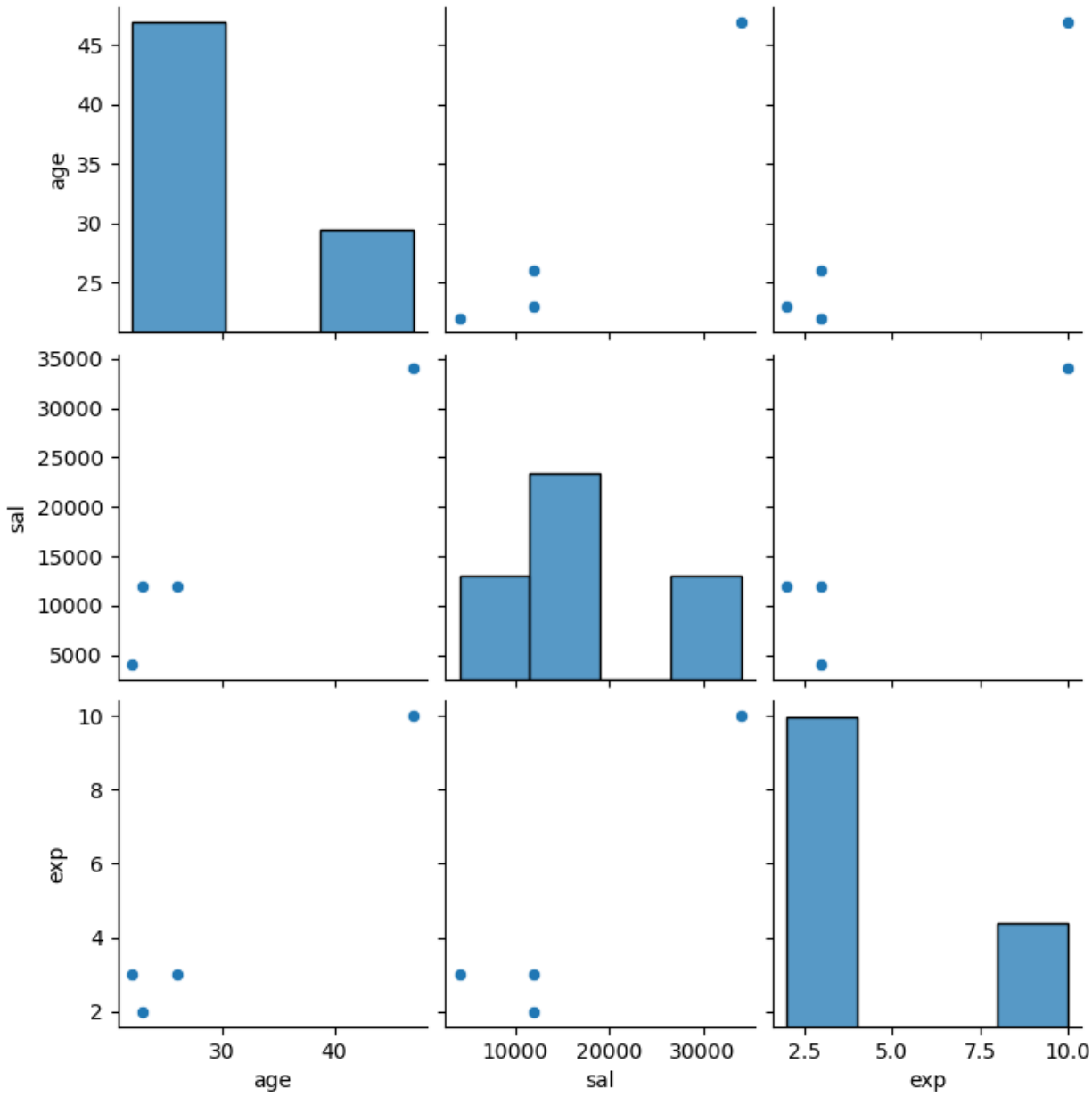
pair plot

```
sns.pairplot(df)
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
```

```
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):

<seaborn.axisgrid.PairGrid at 0x1a005254210>
```




```
* age ,exp, &sal have different level
```

```
  Cell In[66], line 1
    * age ,exp, &sal have different level
                ^
SyntaxError: invalid syntax
```

```
sns.pairplot(df,hue='g')
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
```

```
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):

<seaborn.axisgrid.PairGrid at 0x1a002b6f610>
```