importing libraries

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

loding and viewing data

```
df=pd.read_csv(r"C:\Users\Bhoomika.G\OneDrive\Documents\
Salary_EDA.csv")
df.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary
0   90000.0
1   65000.0
2  150000.0
3   60000.0
4   60000.0
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 375 entries, 0 to 374
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Age                  373 non-null    float64
 1   Gender               371 non-null    object
 2   Education Level      372 non-null    object
 3   Job Title            370 non-null    object
 4   Years of Experience  373 non-null    float64
 5   Salary               372 non-null    float64
dtypes: float64(3), object(3)
memory usage: 17.7+ KB
```

```
conclusion:
    -Age ,salary, year of exprience have in float datatype
    -gender , job titles, have the object data types
    -null values exit became no same non null values
    -they are 6-feature and 375 columns

df.isnull().sum()
```

```
Age                    2
Gender                 4
Education Level        3
Job Title              5
Years of Experience    2
Salary                 3
dtype: int64
```

```
df.dropna(inplace=True)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 366 entries, 0 to 374
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Age                  366 non-null    float64
 1   Gender               366 non-null    object 
 2   Education Level      366 non-null    object 
 3   Job Title            366 non-null    object 
 4   Years of Experience  366 non-null    float64
 5   Salary               366 non-null    float64
dtypes: float64(3), object(3)
memory usage: 20.0+ KB
```

```
df.isnull().sum()
```

```
Age                    0
Gender                 0
Education Level        0
Job Title              0
Years of Experience    0
Salary                 0
dtype: int64
```

```
conclusion: All null values are droped ,now the features are null

df.describe(include='all')
```

```
                Age Gender Education Level             Job Title  \
count   366.000000    366             366                   366   
unique         NaN      2               3                   169   
```

top	NaN	Male	Bachelor's	Director of Marketing
freq	NaN	189	220	12
mean	37.459016	NaN	NaN	NaN
std	6.962303	NaN	NaN	NaN
min	23.000000	NaN	NaN	NaN
25%	32.000000	NaN	NaN	NaN
50%	36.000000	NaN	NaN	NaN
75%	44.000000	NaN	NaN	NaN
max	53.000000	NaN	NaN	NaN

	Years of Experience	Salary
count	366.000000	366.000000
unique	NaN	NaN
top	NaN	NaN
freq	NaN	NaN
mean	10.045082	100492.759563
std	6.517102	48013.732434
min	0.000000	350.000000
25%	4.000000	56250.000000
50%	9.000000	95000.000000
75%	15.000000	140000.000000
max	25.000000	250000.000000

Conclusion:

1.Age:

- -the average age is 37.459016
- -the majority age is the falls become 32 and 44
- -minimum age=23.00
- -maximum age=53.0

2.Gender:

- -the unique values are male and female
- -among the 366,189 entites are male ,177 entites are female,so we can say that male are slightly dominting

3.Education level:

- -most of the data concentrate on the bachelor's(dominating)

4.job Titile:

- - among 366,12 times are directed of markting of markiting is repeated Others are repeated less than 12 times,which means no jobing titles is dominating the dataset

5.years of exprience:

- -minimum exprience is 0,Maxmium exprience is 25,the average expeience is 10 tears
- -majority of exprience is b/w thw 4 and 5

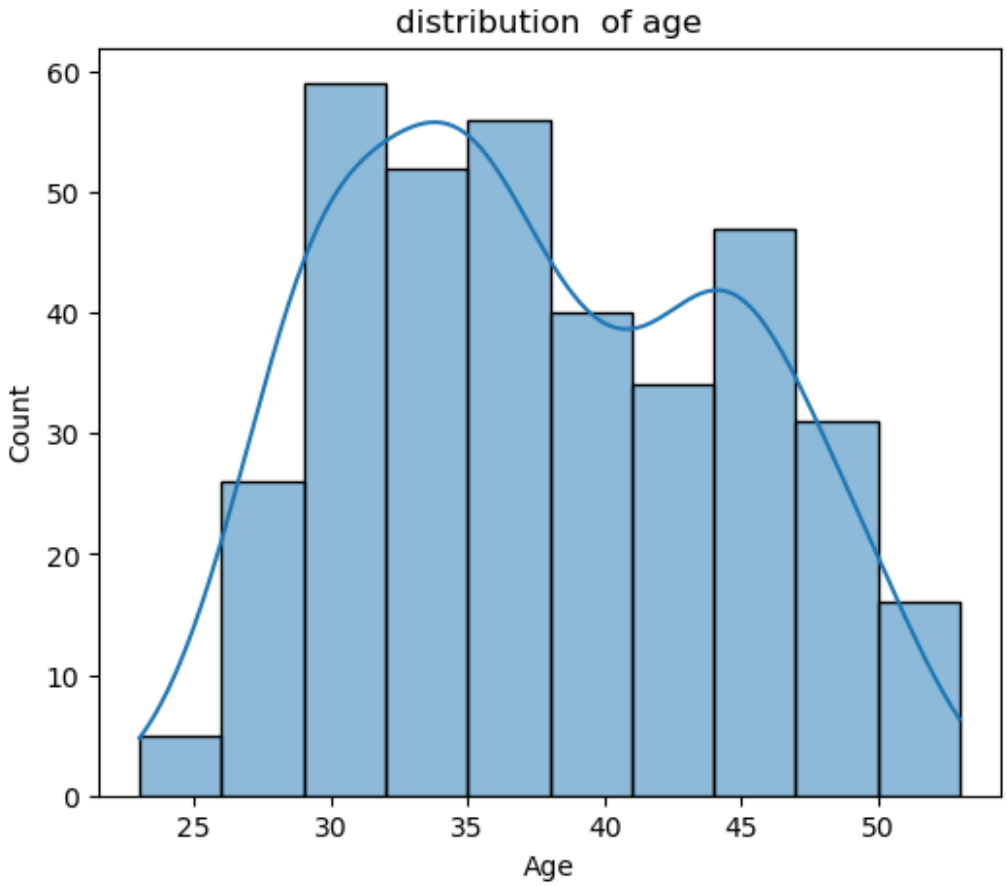
6.Salary:

- -minimum salry is 350,maximum salary is 250000,average salary is

```
1lakh
    -the majority of the salary is b/w 56250 to 140000
    -ther might be outliners,min=350,avg=1lakh,there are lot of
time(error part time)
```

1.Analysis age distribution

```
plt.figure(figsize =(6,5))
sns.histplot(df['Age'],kde=True,bins=10)
plt.title('distribution  of age')
plt.show()
```



Conclusions: -On outline -the majority is 30 to 40 -few peoples are less in 50 and less in 25 -skew on the sightly positive side -the average age 34

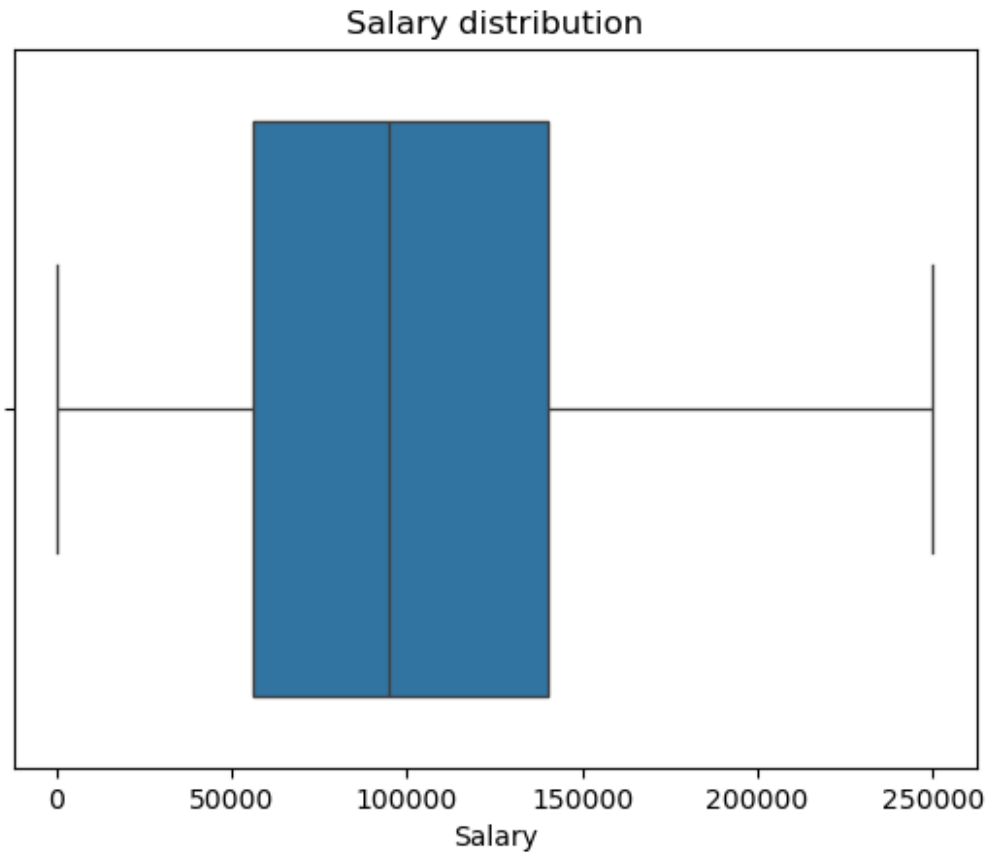
Analyse the distribution of salary

```
plt.figure(figsize =(6,5))
sns.histplot(df['Salary'],kde=True,bins=10)
plt.title('distribution  of salary')
plt.show()
```


```
Conclusions:
-it is outline
-the majority is 56250.0 to 140000
-few peoples are less in 50000 and less in
-skew  on the sightly positive side
-the average   50000
```

Analyse of salary distribution

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



Conclusions: -the majority salary is the 50000 to 150000 -the average is 90000 -the box plot is normal -on outline

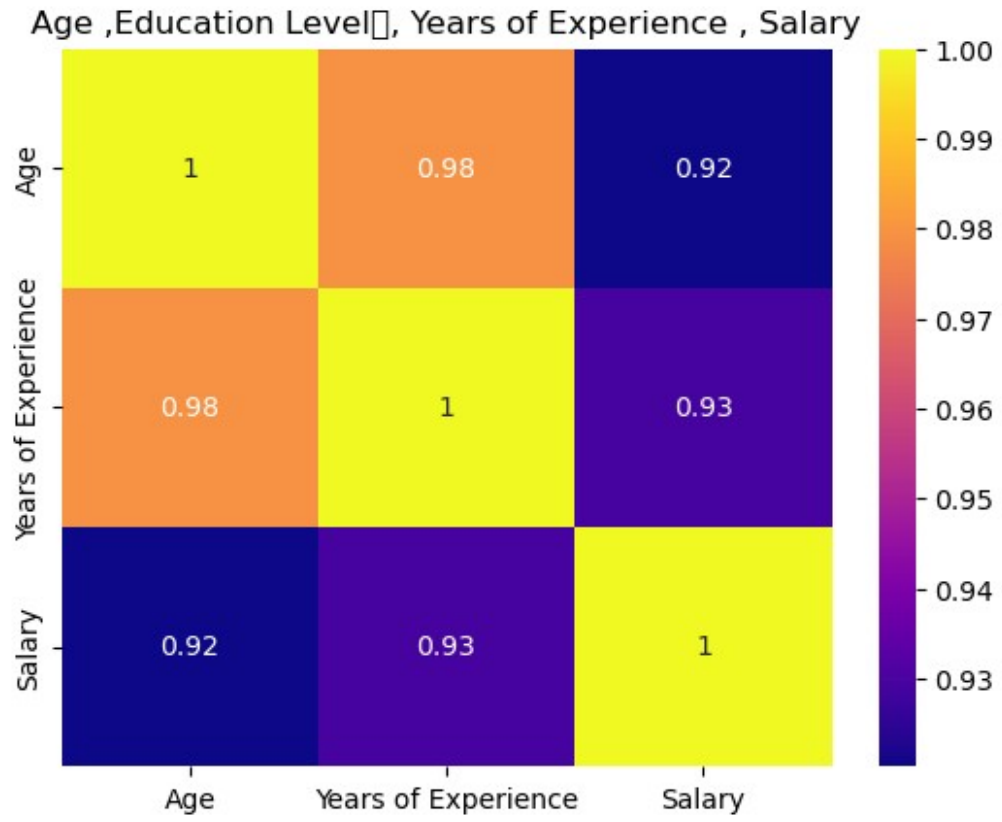
find the corelation matrix

```
ndf=df.select_dtypes(include=['number'])
ndf.head()
```

```
    Age  Years of Experience    Salary
0  32.0                  5.0   90000.0
1  28.0                  3.0   65000.0
2  45.0                 15.0  150000.0
3  36.0                  7.0   60000.0
4  36.0                  7.0   60000.0
```

```
plt.figure(figure=(6,5))
sns.heatmap(ndf.corr(),cmap='plasma',annot=True)
plt.title('Age ,Education Level  , Years of Experience , Salary')
plt.show()
```

```
C:\ProgramData\anaconda3\Lib\site-packages\IPython\core\
pylabtools.py:170: UserWarning: Glyph 9 (   ) missing from font(s)
DejaVu Sans.
  fig.canvas.print_figure(bytes_io, **kw)
```

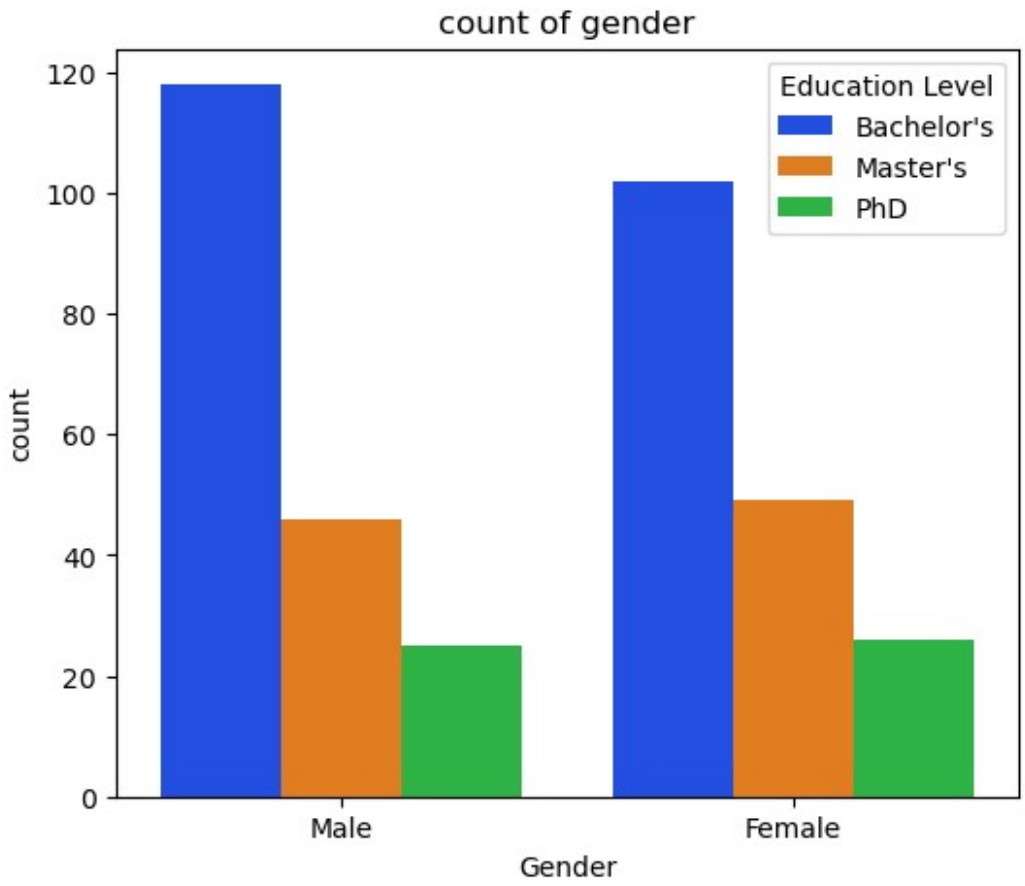



```
concusion:
-age  sal sal exp  are the corelations
-age and salary is the lowest corelation
```

plot the count plot for Gender based on the education

```
plt.figure(figsize=(6,5))
sns.countplot(x=df['Gender'],palette='bright',hue=df['Education
Level'])
plt.title('count of gender')
```

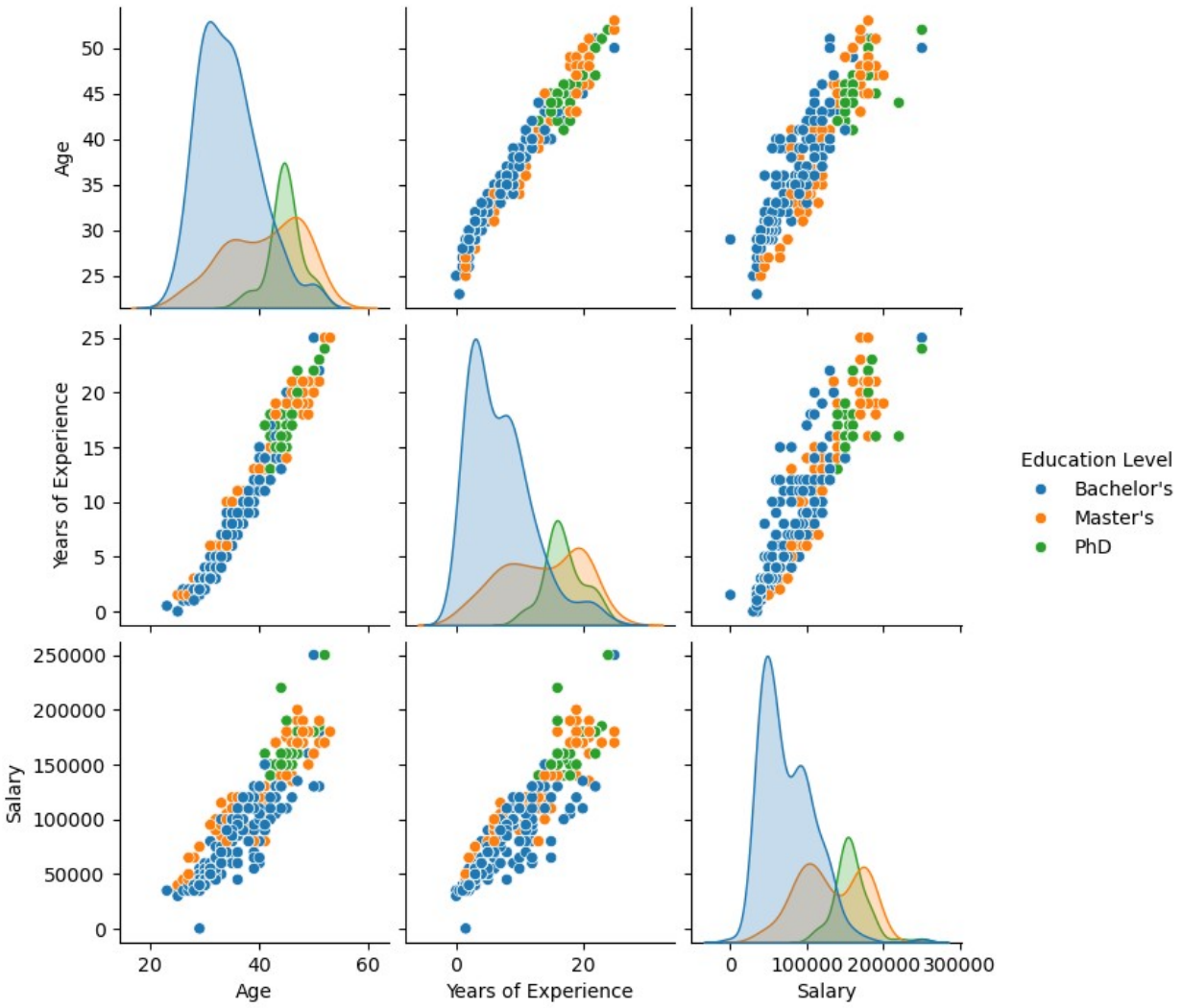
```
Text(0.5, 1.0, 'count of gender')
```



- construct the pair plot for the dataset of Education Level

```
sns.pairplot(df,hue='Education Level')
```

```
<seaborn.axisgrid.PairGrid at 0x272140cae40>
```




```
Conclusion:
-we absover that  when age
-the peak salary are given to bacholer degree people
-empoyes are in bachlores degree  are conistences in the job
-salary is also effect by the years of expriences
```

- group eduction level and find avg salary for every catagary
- filter data set it will exprince is more than 20 years and find the avg salary

```
g1=df.groupby('Education Level')['Salary'].mean()
g1

Education Level
Bachelor's     74683.409091
Master's      129473.684211
PhD           157843.137255
Name: Salary, dtype: float64
```

```
g1=df[(df['Gender']=='Female')&(df['Education Level']=="Master's")]
g1['Salary'].mean()
```

```
121020.40816326531
```

```
e=df[df['Years of Experience']>20]
e['Salary'].mean()
```

```
175892.85714285713
```

Aggregation

```
df.groupby("Education Level").agg({'Age':['count','mean']})
```

	Age	
	count	mean
Education Level		
Bachelor's	225	34.364444
Master's	96	40.718750
PhD	51	44.725490