Importing librabries

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

loading and veiwing

```
df=pd.read_csv(r"C:\Users\DELL\Downloads\my_python\Salary_EDA.csv")
df

      Age  Gender Education Level                      Job Title  \
0    32.0    Male      Bachelor's              Software Engineer   
1    28.0  Female        Master's                   Data Analyst   
2    45.0    Male             PhD                 Senior Manager   
3    36.0  Female      Bachelor's                Sales Associate   
4    36.0  Female      Bachelor's                Sales Associate   
..    ...     ...             ...                            ...   
370  35.0  Female      Bachelor's       Senior Marketing Analyst   
371  43.0    Male        Master's         Director of Operations   
372  29.0  Female      Bachelor's         Junior Project Manager   
373  34.0    Male      Bachelor's  Senior Operations Coordinator   
374  44.0  Female             PhD        Senior Business Analyst   

     Years of Experience    Salary  
0                    5.0   90000.0  
1                    3.0   65000.0  
2                   15.0  150000.0  
3                    7.0   60000.0  
4                    7.0   60000.0  
..                   ...       ...  
370                  8.0   85000.0  
371                 19.0  170000.0  
372                  2.0   40000.0  
373                  7.0   90000.0  
374                 15.0  150000.0  

[375 rows x 6 columns]

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
```

observation/conclusion : info()

1. age ,year of experiance ,& salary is in float datatype
2. gender , education ,job title is in object datatype
3. null values exist because no same non -null values
4. 6-features ,375 rows

```
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.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 dropped .now the features have no null values

summary statistics

```
df.describe()

              Age  Years of Experience         Salary
count  366.000000           366.000000     366.000000
mean    37.459016            10.045082  100492.759563
std      6.962303             6.517102   48013.732434
min     23.000000             0.000000     350.000000
25%     32.000000             4.000000   56250.000000
50%     36.000000             9.000000   95000.000000
75%     44.000000            15.000000  140000.000000
max     53.000000            25.000000  250000.000000
```

```
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
   - minimum age is 23,maximum age is 53,average age is 37.4
   - majority of age falls between 32-44
   - few entries from 50's
2. gender
   - there are 2 unique values male female
   - among 366 ,189-males,177-females.so we can say male is slightly dominating
3. educational level
   - most of the data concentrates on bachelor's(dominating)
4. job title
   - among 366 ,12 times directoe of marketing is repeated .Other are repeated less than 12 times .which means no job title is dominating in the dataset
5. years of experiance
   - minimum age is 0,maximum age is 25,average age is 10
   - majority of age falls between 4-15
6. salary
   - minimum age is 350,maximum age is 250000,average age is 1lakh

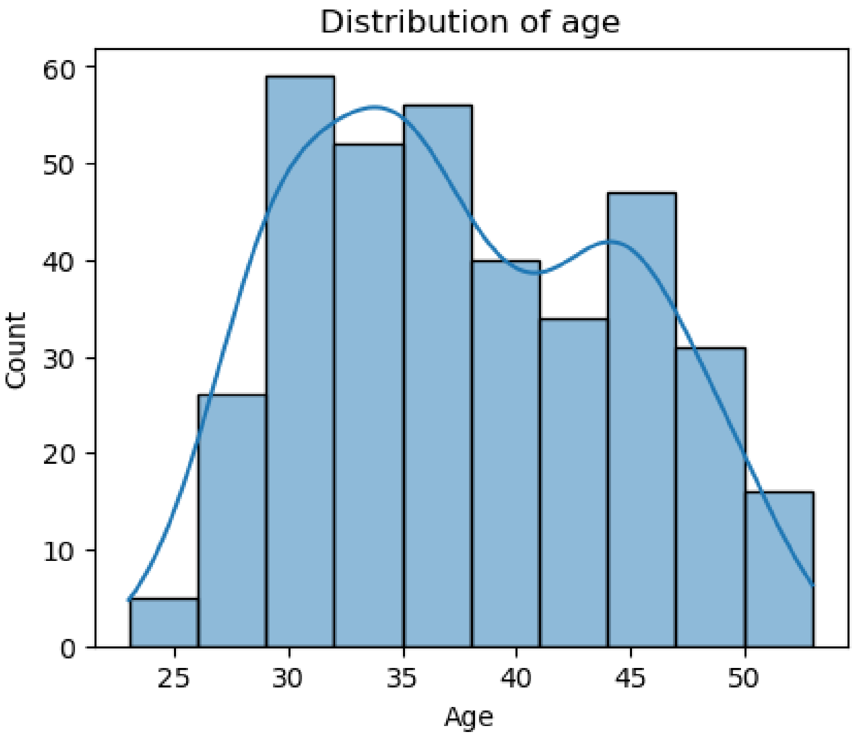
- majority of age falls between 56000-1lakh
- there might be outlier ,min-350,avg-1lakh,there is lot difference (error ,part-time)

visualization

1.analyze age distribution[histogram]

```
plt.figure(figsize=(5,4))
sns.histplot(df['Age'],kde = True ,bins =10)
plt.title('Distribution of age')
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):
```



majority of age is 30-40

average age are fall between is 30-35

minimum age are 25

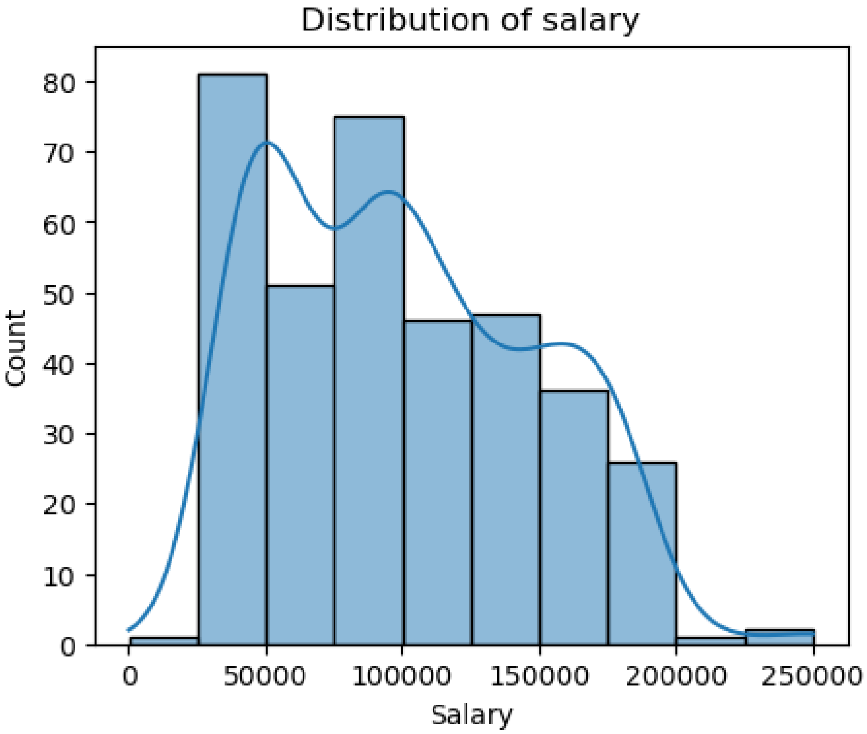
maximum age are 50

there is no outlier in the age data set

analyse the distribution of salary using histogram

```
plt.figure(figsize=(5,4))
sns.histplot(df['Salary'],kde = True ,bins =10)
plt.title('Distribution of salary')
plt.show()
```

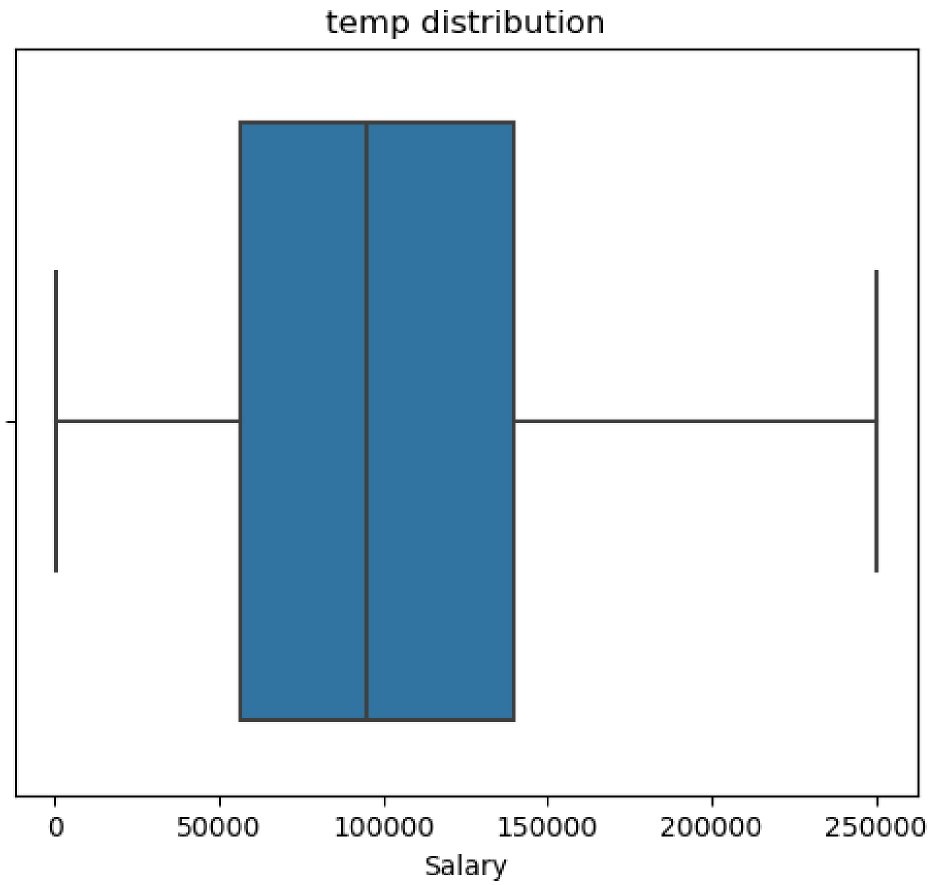
```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



minimum salary is 0-350 maximum salary is 250000, average salaryis 1lakh majority ofsalary falls between 56000-1lakh there might be outlier ,min-350,avg-1lakh,there is lot difference (error ,part-time)

analyse salary distribution using boxplot

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



- majority of salary fall between 100000-250000

the large values are towards right

there is no outlier

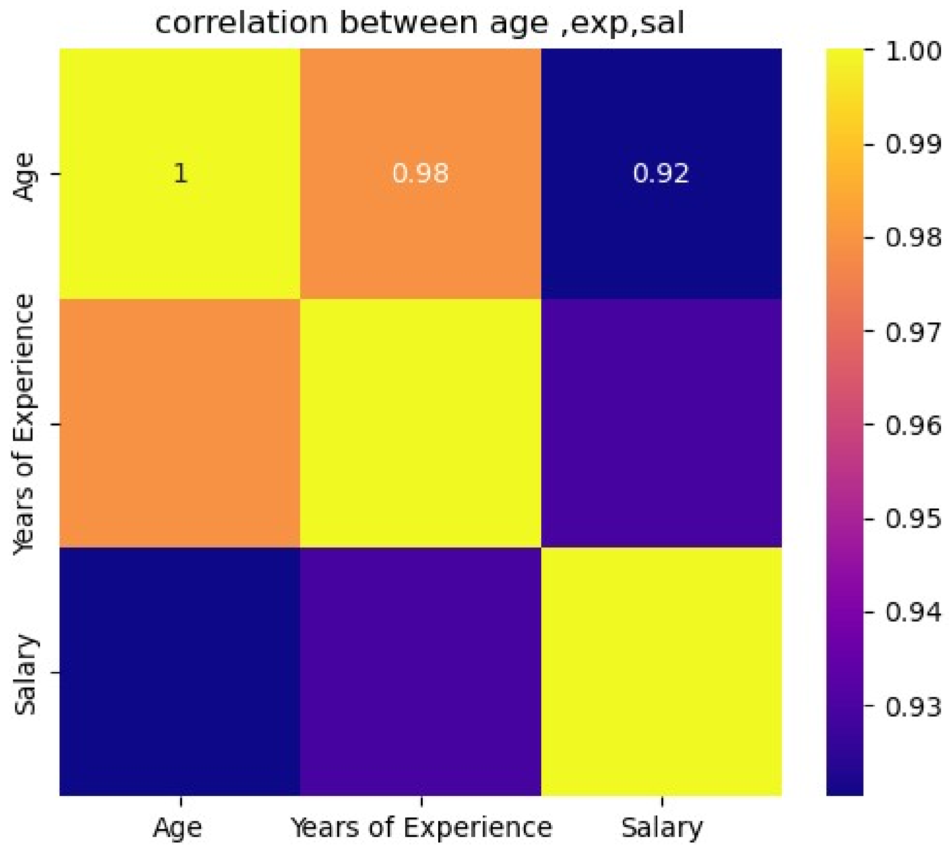
# find the correlation matrix

```
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
```

```
#step 2: heat map
plt.figure(figsize=(6,5))#rows and column
sns.heatmap(ndf.corr(),cmap='plasma',annot=True)#color --
plasma,coolwarm
```

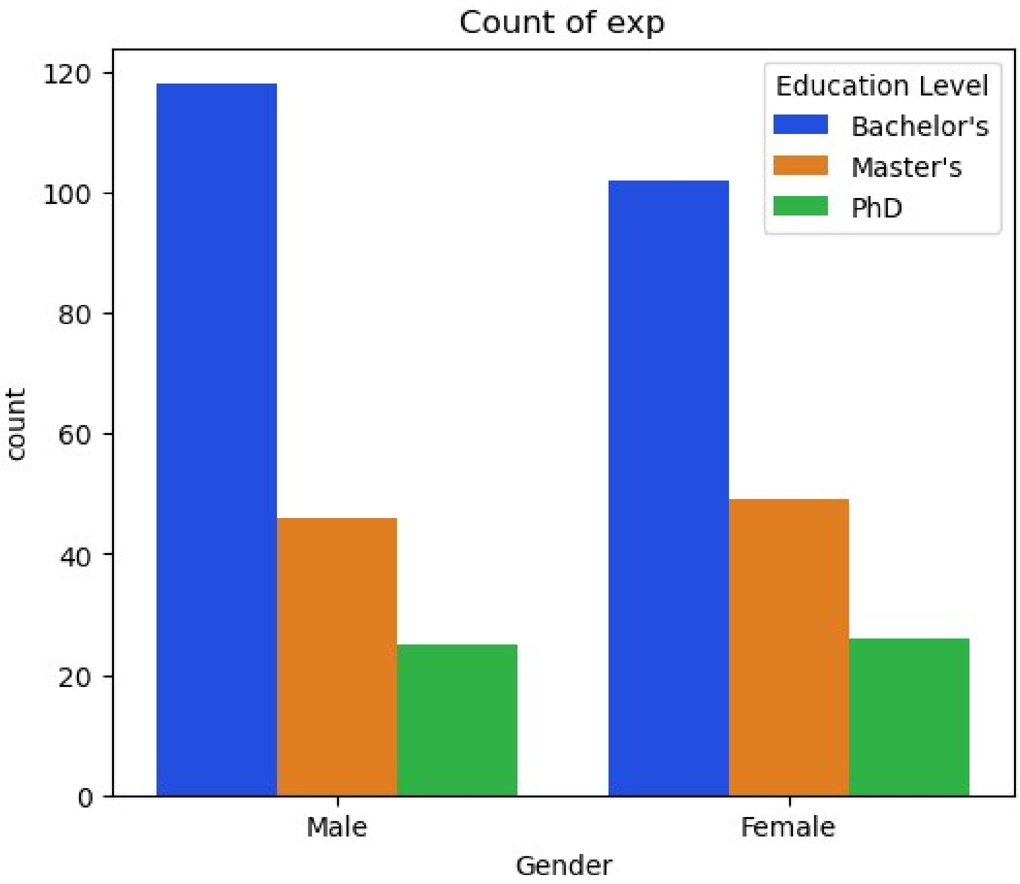
```
plt.title("correlation between age ,exp,sal")
plt.show()
```



- the good corrrelation between the age and experiance
- the poor corrrelation between the age and salary

draw count plot between education and gender

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



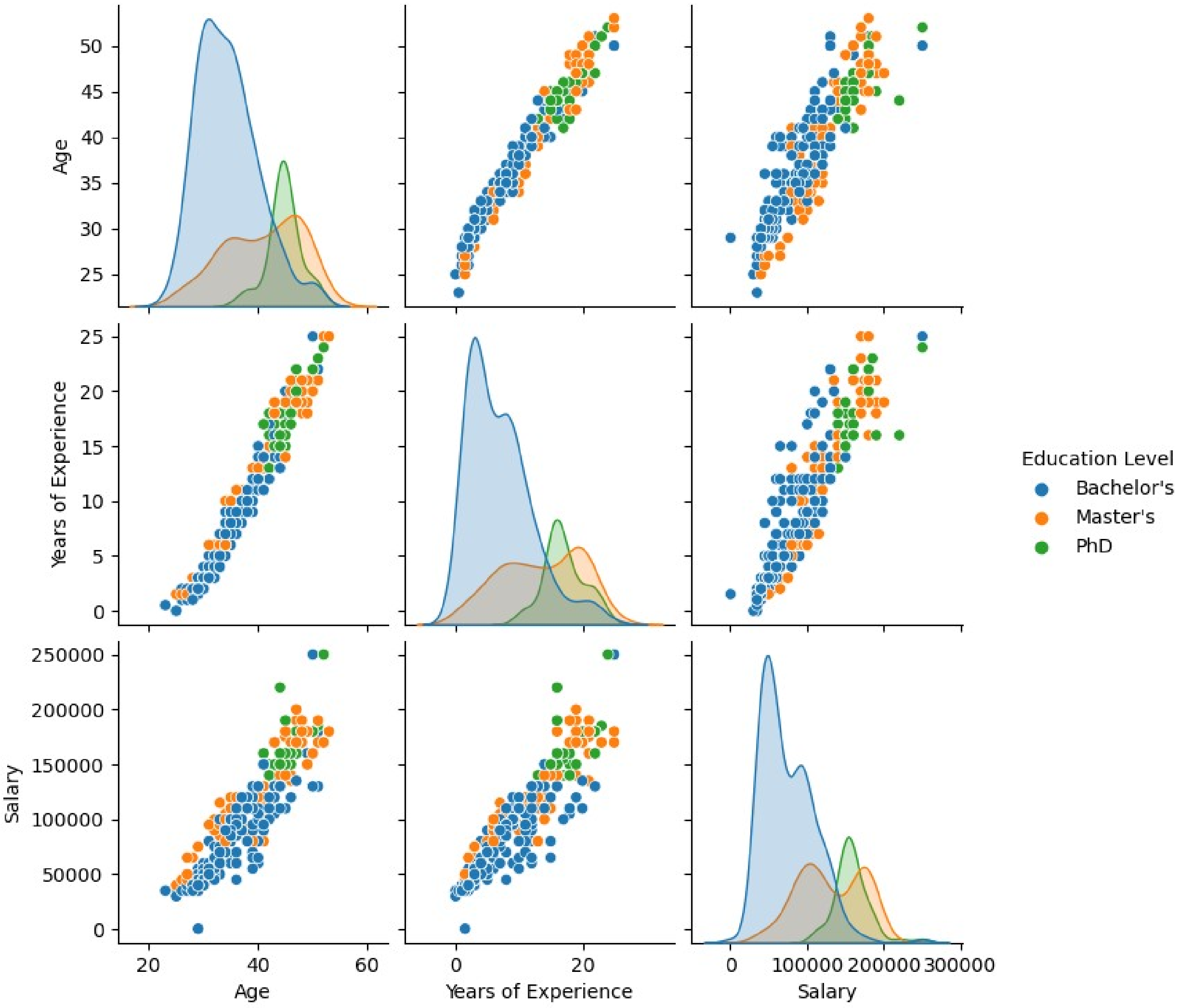
- majority is bachelor dominating and lower is phd
- by taking the girl is low and the male is high(dominating)

construct a pair plot color variation

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

```
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 0x2296b60ff50>
```



min sal- master ,max -bachelor we observed that age increses experiance the pick salary are give two bachelor degree people employee are bachelor degree are consistant in job max year of experiance is bachelor min - masters

# group education level and find avrage salary in every category

# filter dataset in which gender is female and education level is masters and find the average salary on that data set

# filter data set in experiance is more than 20 years and find the avg sal on tha data set

```
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
```

by analyse this bachelor have avgrage salary 74000 masters have avarage salry is 120000 phd AVARAGE SALRY IS 150000

```
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        220  34.368182
Master's           95  40.715789
PhD                51  44.725490
```