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

sdf=pd.read_csv("C:\mypythonfiles\Salary_EDA.csv")
sdf.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
```

```
sdf.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
```

Conclusions: 1.age,exp,salary are in foat datatype 2.gender,educational level,job title are in object datatype 3.Null-values exist

Handling null values

```
sdf.isnull().sum()
```

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

```
sdf.dropna(inplace=True)
sdf.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
```

Conclusion: all the null values are dropped.now feature have no null values

Summary statistics

```
sdf.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
```

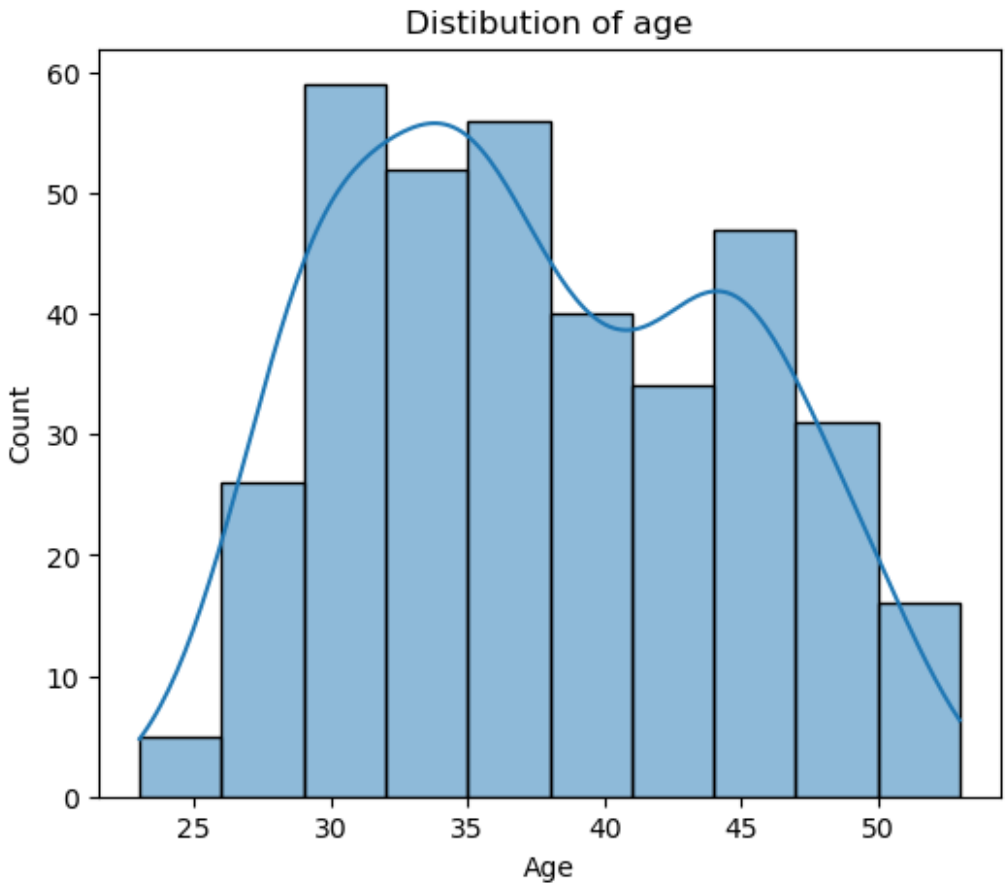
```
       Years of Experience         Salary
count            366.000000     366.000000
unique                  NaN            NaN
top                     NaN            NaN
freq                    NaN            NaN
mean              10.045082  100492.759563
std                6.517102   48013.732434
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
```

visualisation

1.analyze age distribution[histogram]

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

```
C:\Users\DELL\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):
```



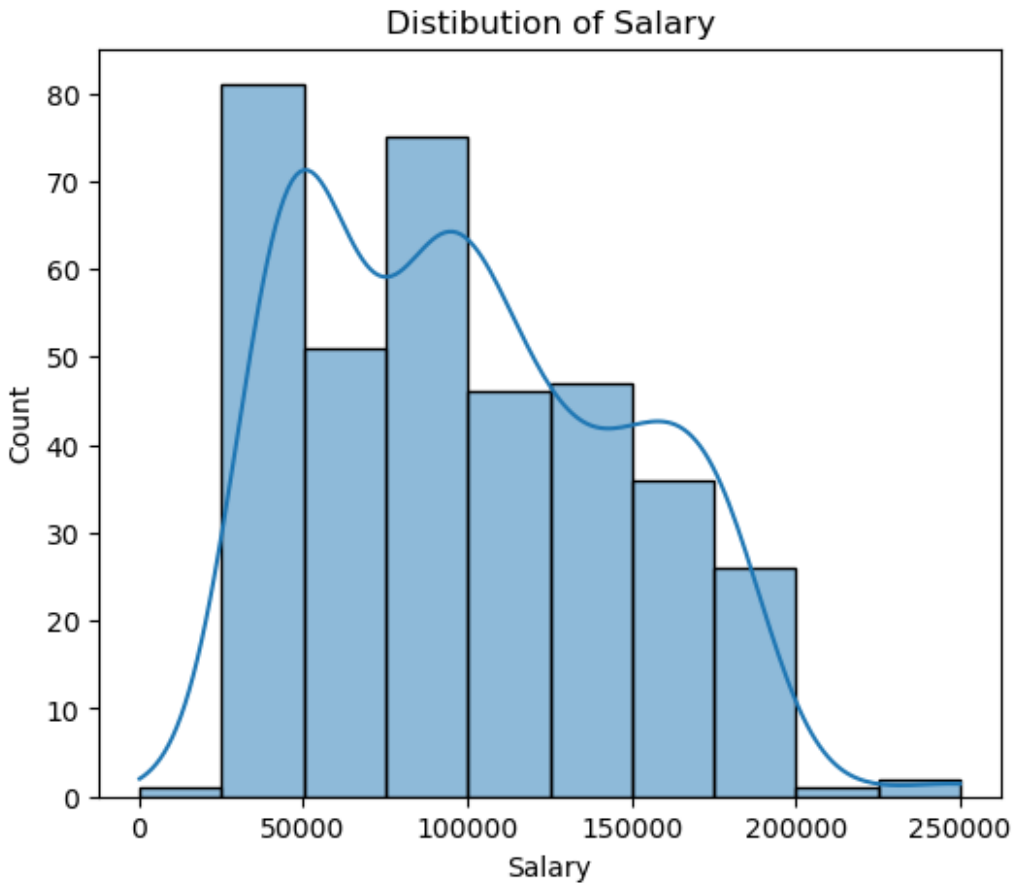
Conclusion:

- there is slight positive skew
- majority of age is between 30
- there is no outliers

analyze the distribution of salary

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

```
C:\Users\DELL\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):
```

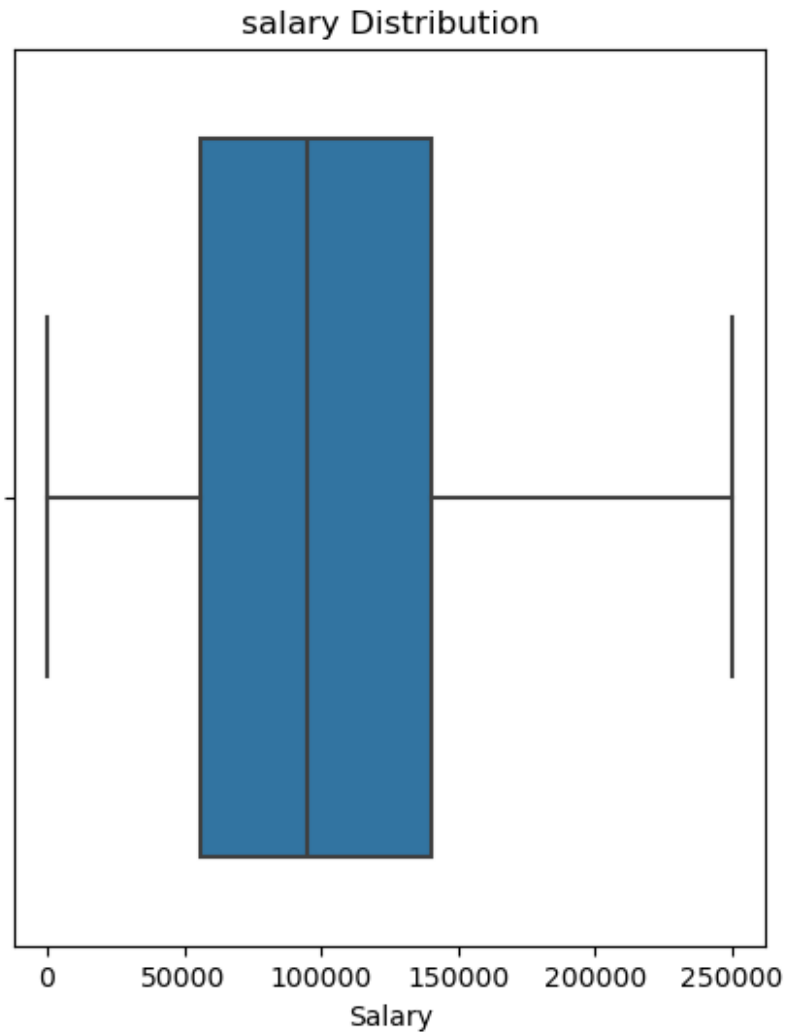


Conclusion:

- the avg salary is 50000

analyse salary distribution using boxplot

```
plt.figure(figsize=(5,6))
sns.boxplot(x= sdf['Salary'])
plt.title('salary Distribution')
plt.show()
```



Conclusion:

- there is no outlier
- there is upperbound
- there is lowerbound
- the average value salary is 90000 to 1lakh

find the correlation matrix

```
#step1: filter the numerical value
nsdf=sdf.select_dtypes(include =['number'])
nsdf.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(figsize=(6,5))
sns.heatmap(nsdf.corr(), cmap='coolwarm', annot=True)
plt.title('Correlation between sal,age,exp')
plt.show()
```

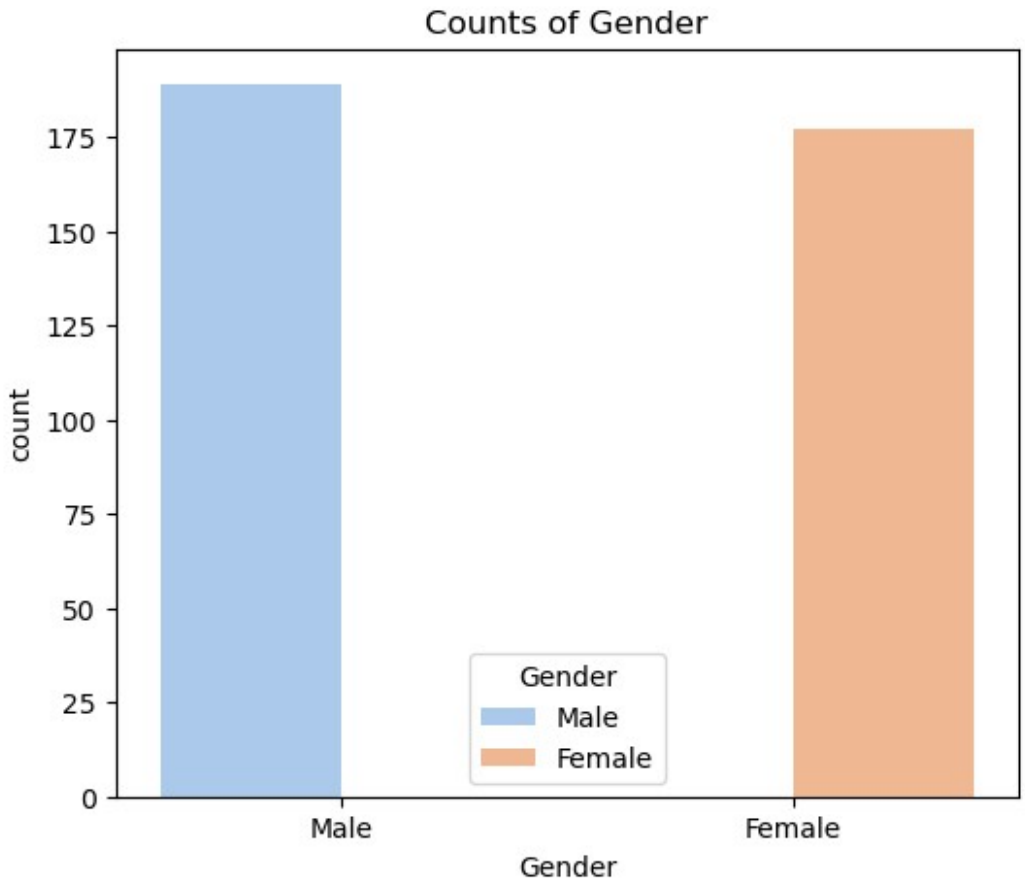
conclusion:

- salary and experiance correlated

draw a countplot for the feature gender

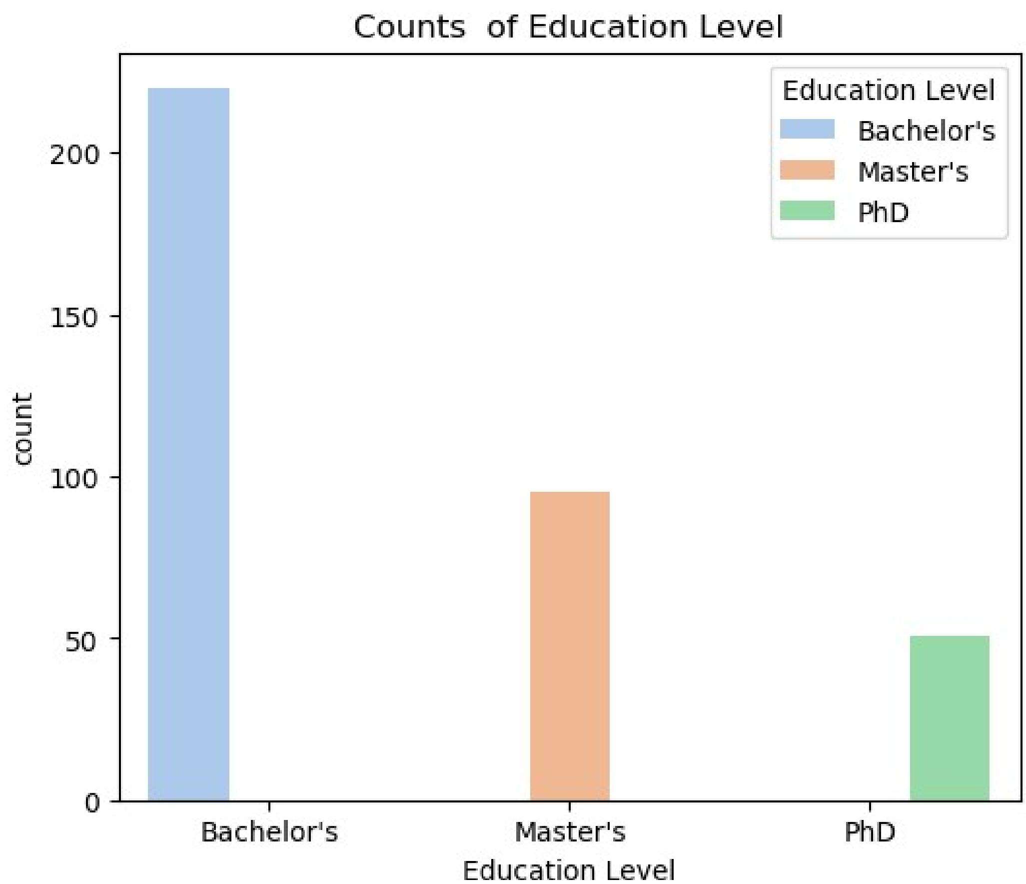
draw a countplot on education level

```
plt.figure(figsize=(6,5))
sns.countplot(x=sdf['Gender'],palette='pastel',hue=sdf['Gender'])
plt.title('Counts of Gender')
plt.show()
```



conclusion: -there are more counts of male than female

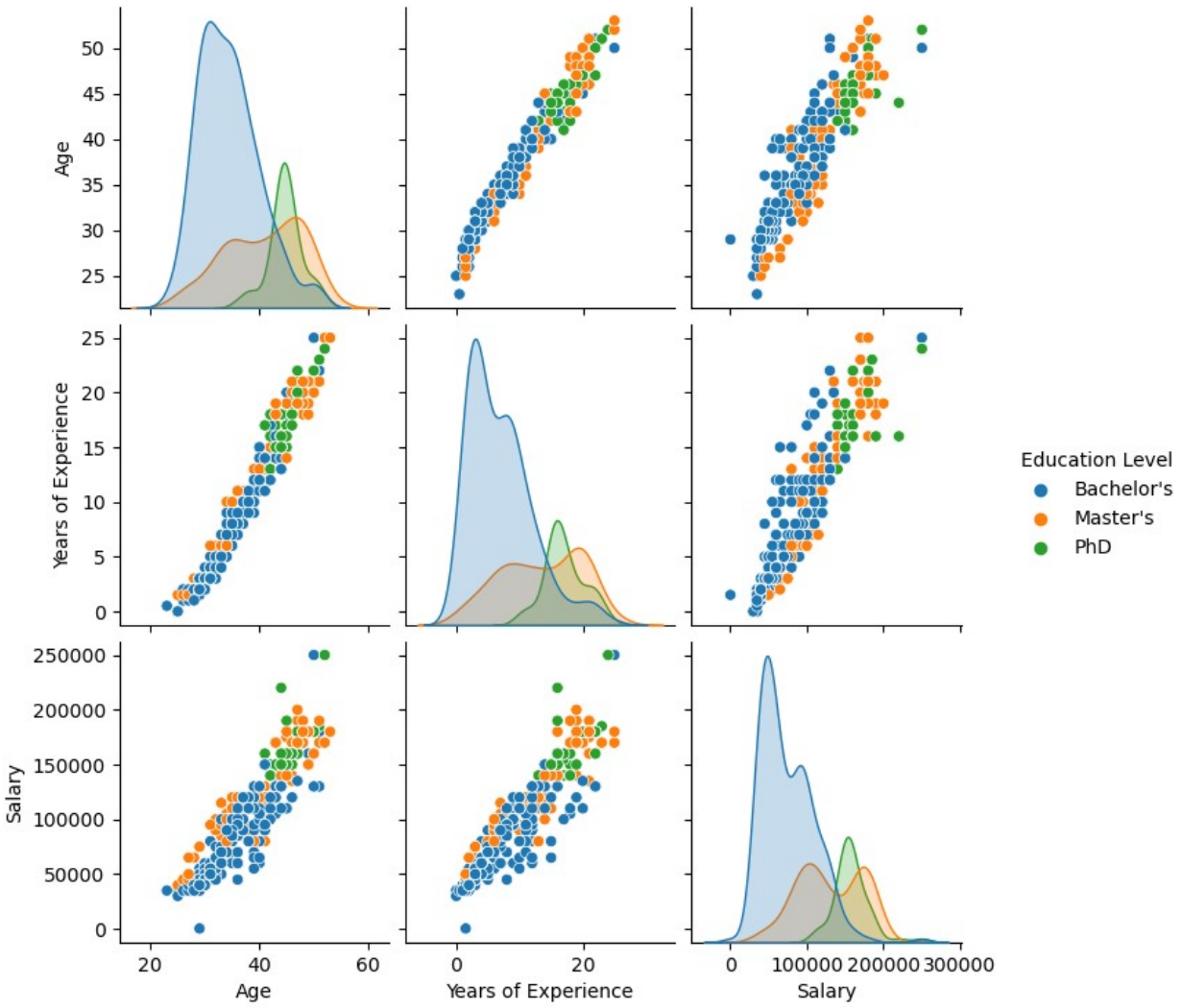
```
plt.figure(figsize=(6,5))
sns.countplot(x=sdf['Education
Level'],palette='pastel',hue=sdf['Education Level'])
plt.title('Counts  of Education Level')
plt.show()
```




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

C:\Users\DELL\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:\Users\DELL\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:\Users\DELL\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 0x2885dde95d0>
```



conclusion:

- we observed that if age increases the experiance is also increased
- the peak salary is given to bachelor degree people
- employee in bachelor degree consistly
- salary is also excpe

-group the education level and find average of all the categories -filter dataset in which gender is female and education level is masters and also find the average of the data -filter dataset in which ecperiance is more than 20 years and find the averge salary on dataset

```
sdf.groupby('Education Level')['Salary'].mean()

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

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

```
121020.40816326531
```

```
Exp20=sdf[sdf['Years of Experience']>20]
Exp20['Salary'].mean()
```

```
175892.85714285713
```

```
sdf.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