Seaborn

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

 mydata={ 'Names' : ['Ram','Sam','Raj','Ullas'],
        'Age' : [22,23,19,20],
        'Salary' : [20000,22000,25000,42000],
         'Exc': [2,2,1,3]
      }

df=pd.DataFrame(mydata)
df.head()
```

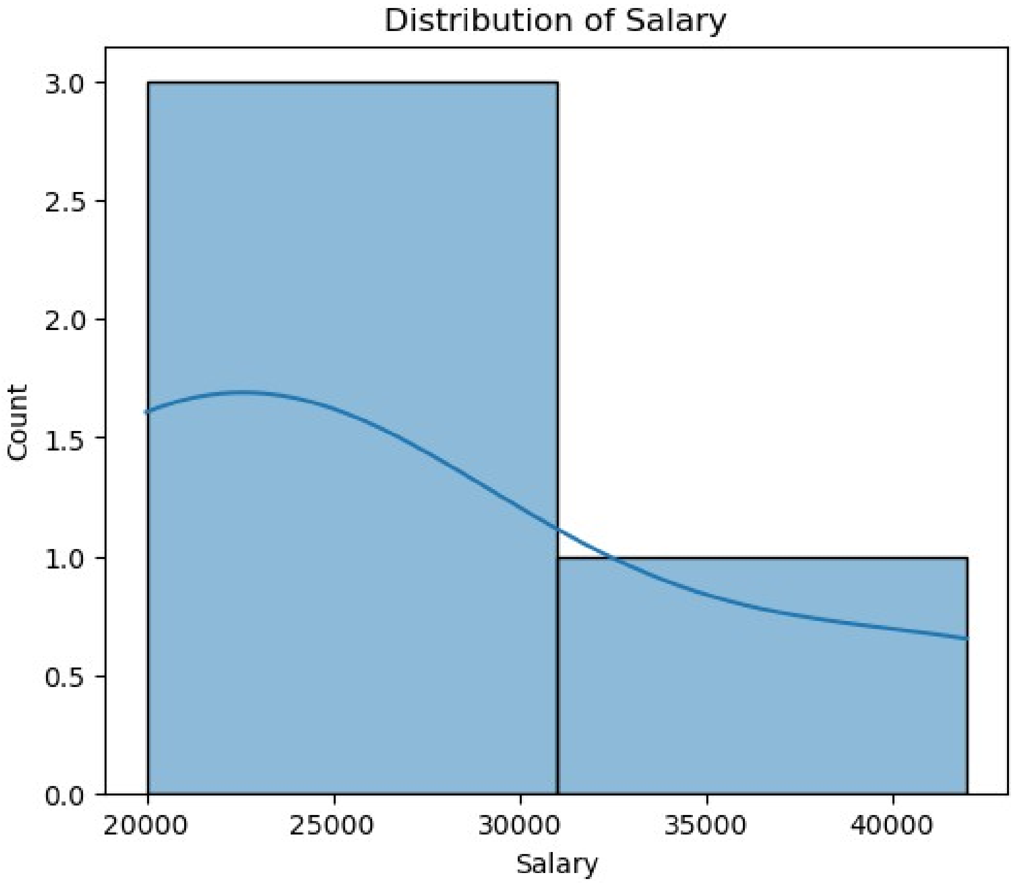
```
   Names  Age  Salary  Exc
0    Ram   22   20000    2
1    Sam   23   22000    2
2    Raj   19   25000    1
3  Ullas   20   42000    3
```

1.Histogram

1.Positive skew, Large salary value 2.No outlier detected 3.Average salary is about 10000 4.Majority salary are between

```
plt.figure(figsize=(6,5))
sns.histplot(df["Salary"],kde=True,bins=2)
plt.title("Distribution 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):
```



1.Positive skew, large salary value

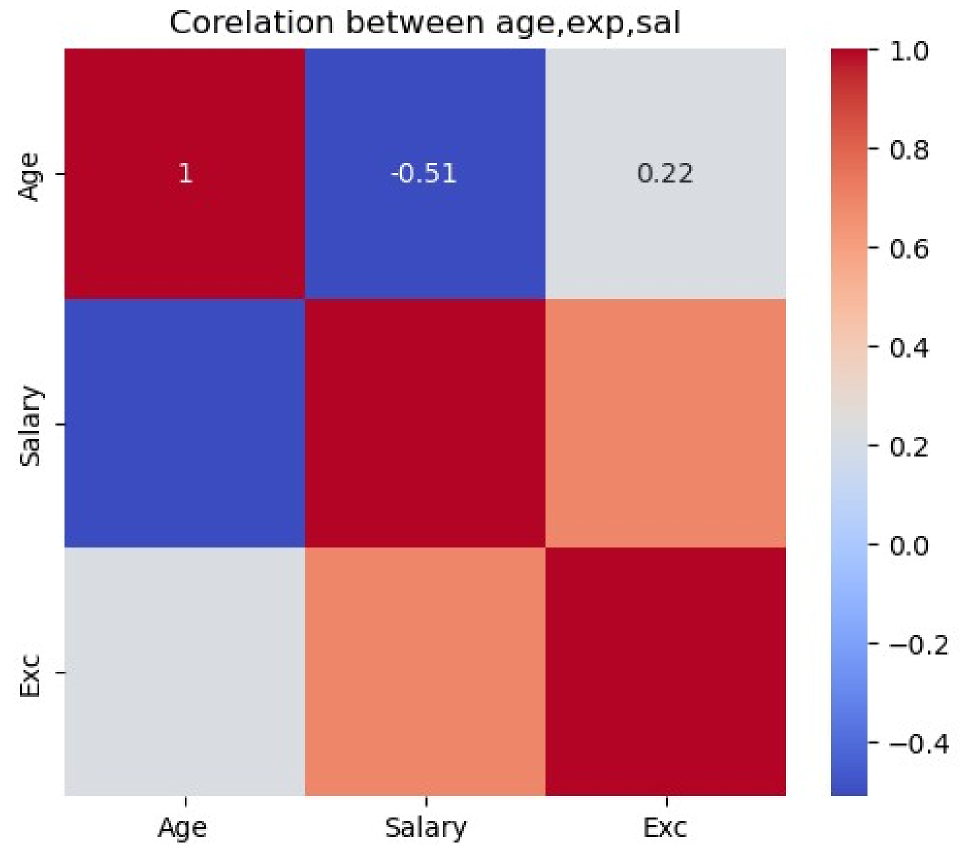
1. No outlier detected 3.Average salary is about 22000 4.Majority salary are between 20000 to 30000

Corelation matrix(Heat map)

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

```
   Age  Salary  Exc
0   22   20000    2
1   23   22000    2
2   19   25000    1
3   20   42000    3
```

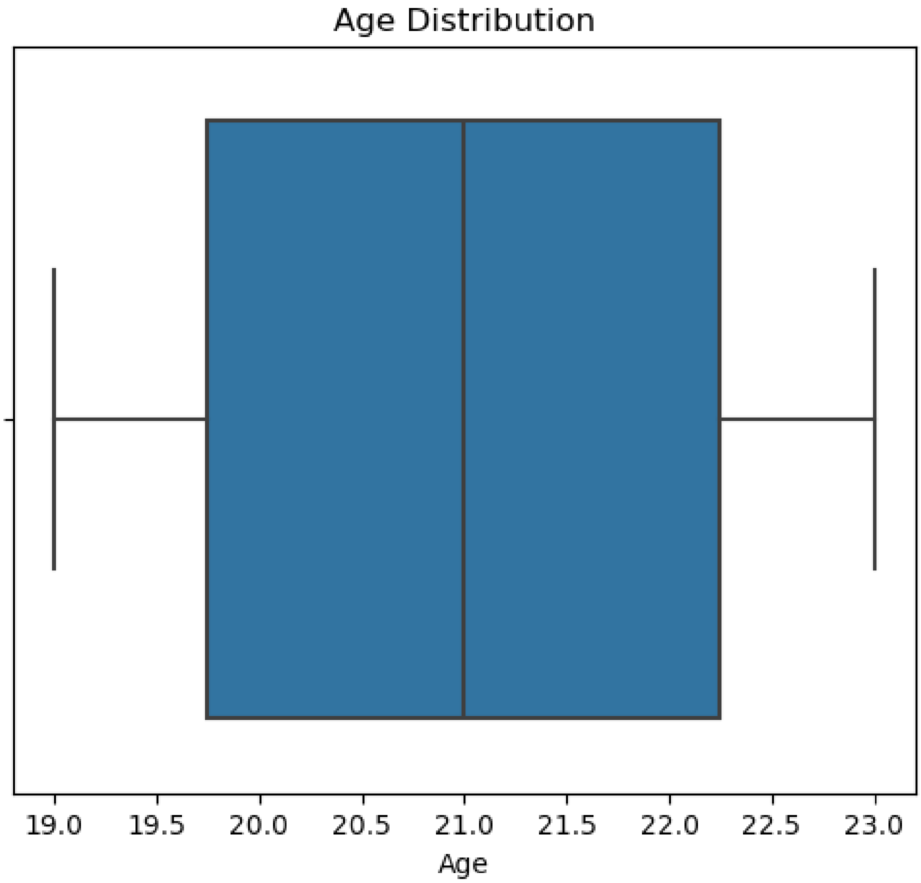
```
plt.figure(figsize=(6,5))
sns.heatmap(ndf.corr(),cmap='coolwarm',annot=True)
plt.title("Corelation between age,exp,sal")
plt.show()
```



1.Dark area is more Correlated 2.light color area are less corelated

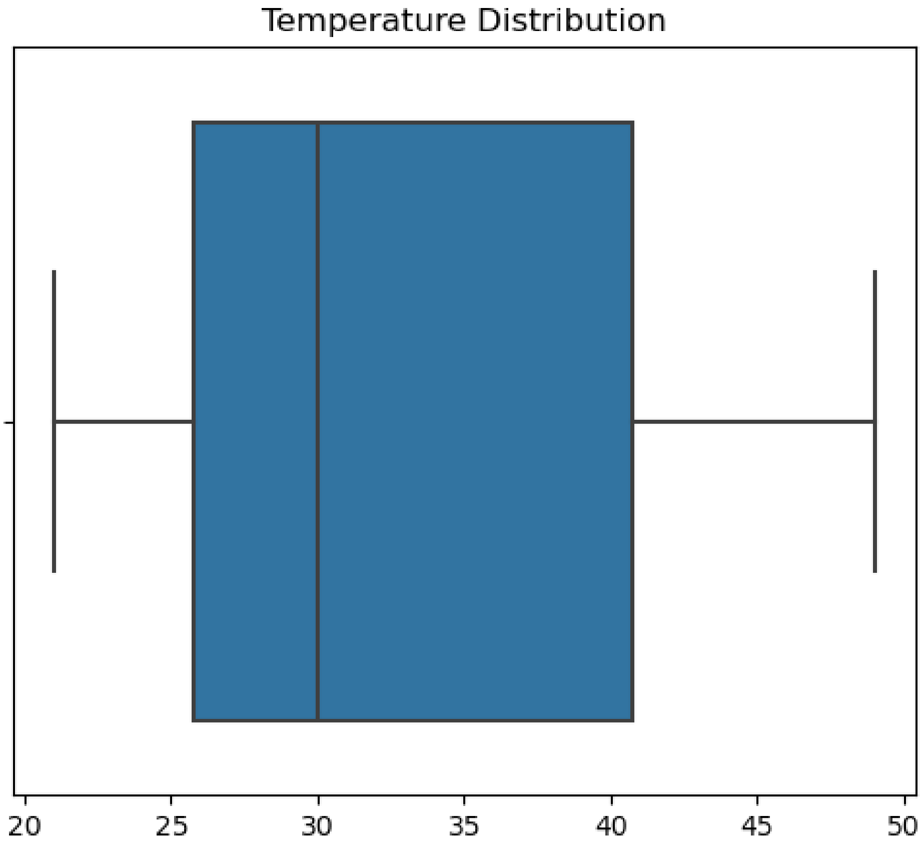
Box plot

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



1.The average age is 21 2.The abnormal value is around 23

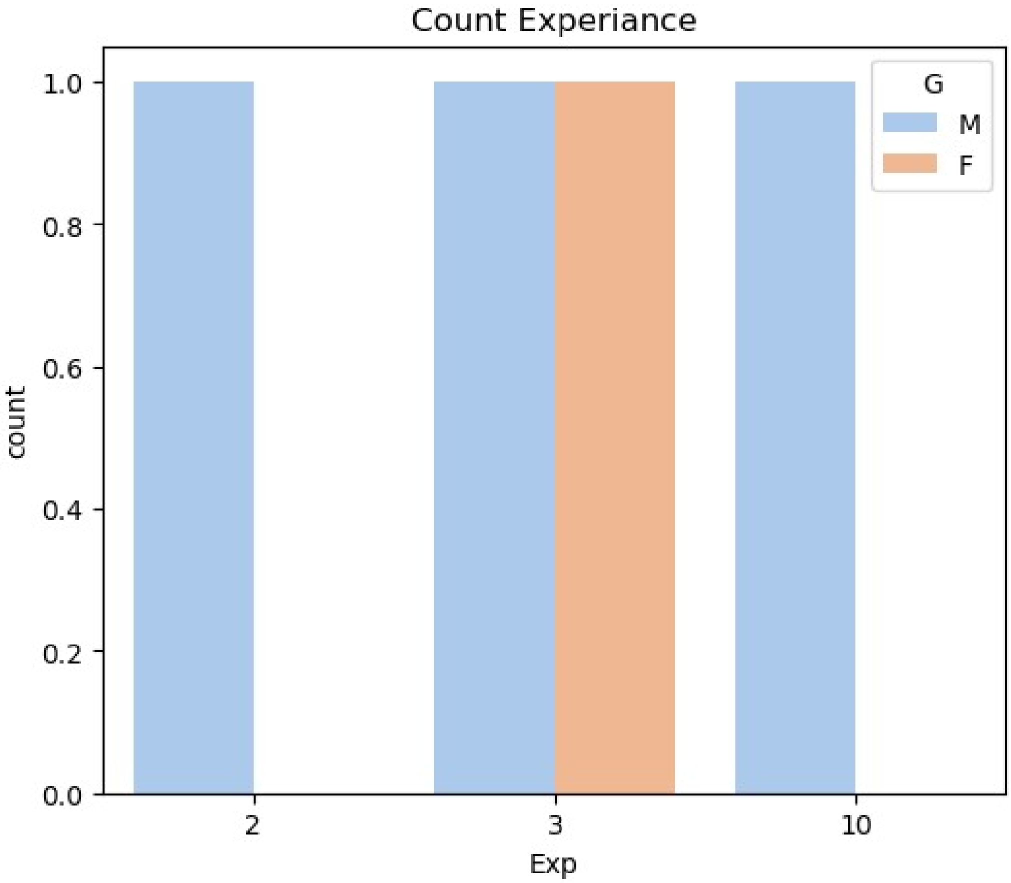
```
temp=[21,47,39,22,31,33,29,26,27,25,49,46]
plt.figure(figsize=(6,5))
sns.boxplot(x=temp)
plt.title("Temperature Distribution")
plt.show()
```



1.Average age value is 30 2.there is no abnormal value 3.The lower bound is around 21 and upper bound is around 49

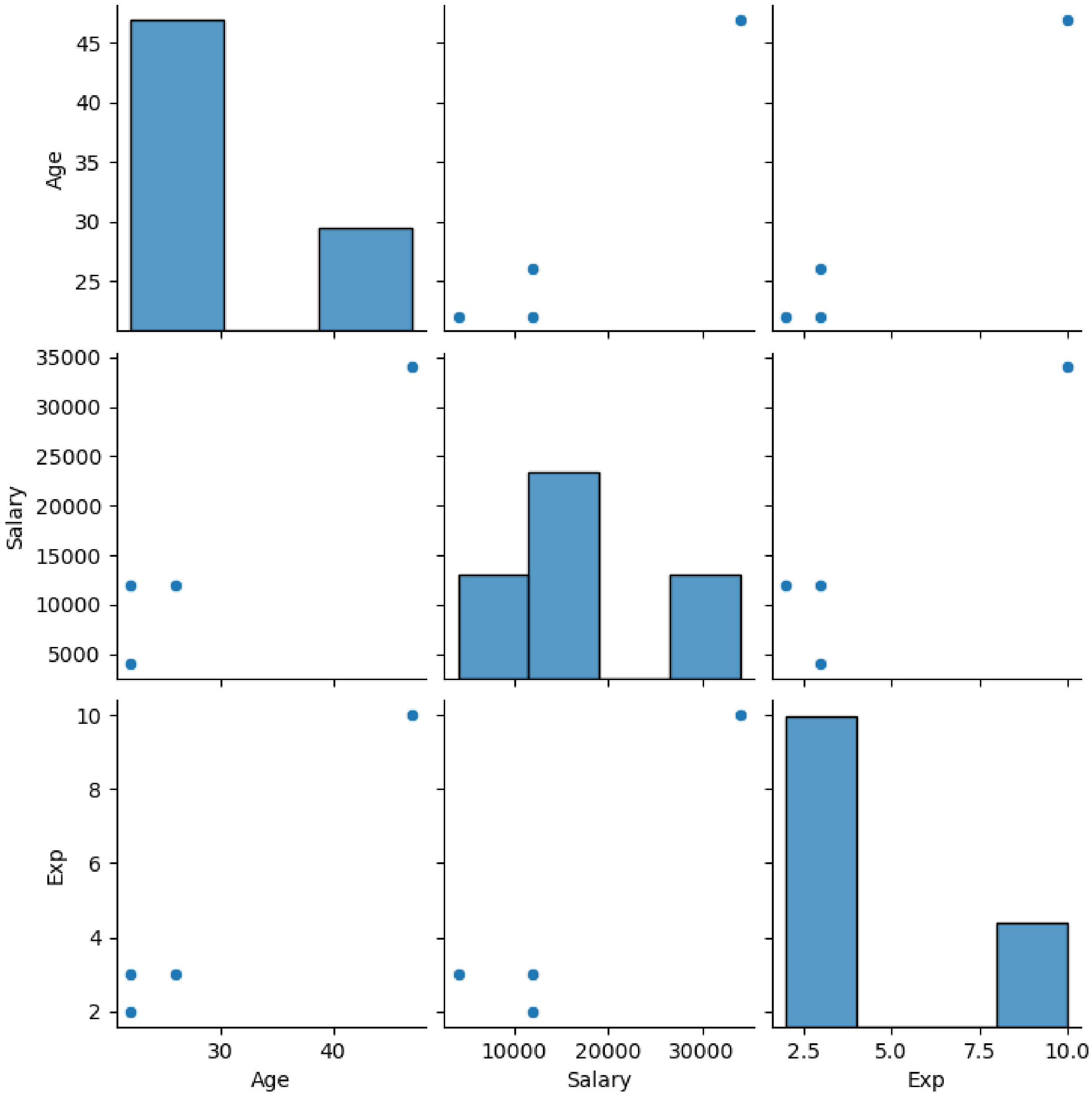
Countplot

```
 mydata={ 'Names' : ['Ram','Sam','Raj','Ullas'],
        'Age' : [22,22,26,47],
        'Salary' : [12000,4000,12000,34000],
         'Exp': [2,3,3,10],
         'G' : ['M','F','M','M']
       }
df1=pd.DataFrame(mydata)

plt.figure(figsize=(6,5))
sns.countplot(x = df1['Exp'],palette='pastel',hue=df1['G'])
plt.title("Count Experiance")
plt.show()
```



Pair plot

```
sns.pairplot(df1)

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 0x24693322710>
```



```
df=pd.read_csv(r"C:\Users\DELL\Downloads\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
```

conclusion:

1. Age, Experience and Salary have float datatype
2. Gender, Education,job title have object datatyp
3. Null-valuese

Handling Nullvalues

```
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 droped. Now the features have non-null

```
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, Maximuum age is 53, average age is 37.4 -Majority of age falls between 32 and 34 -few entries from 50s

1. Gender

   -There are two unique value male and female -Among 366, 189 entries are male and 177 entries are female. So we can say male is slightly dominating 3.Education level -Most of the data concentrates on bachelor's(dominating) 4.Job title -Among 366, 12 times director of marketing is repeated. Others are repeated less than 12 times which means no job title is dominating in the dataset 5.Years of Experience -Minimum experiance is 0, Maximum experiance is 25, Average experiance is 25. -Majority of people have exoeriance between 4 and 15 6.Salary -Minimum salary is 350,maximum experiance is 250000, Average salary is 1L -Majority of salary between 56000 and 1L

   - Their might be outliers, min-350,avg-1L,There is lot difference(error,part-time)

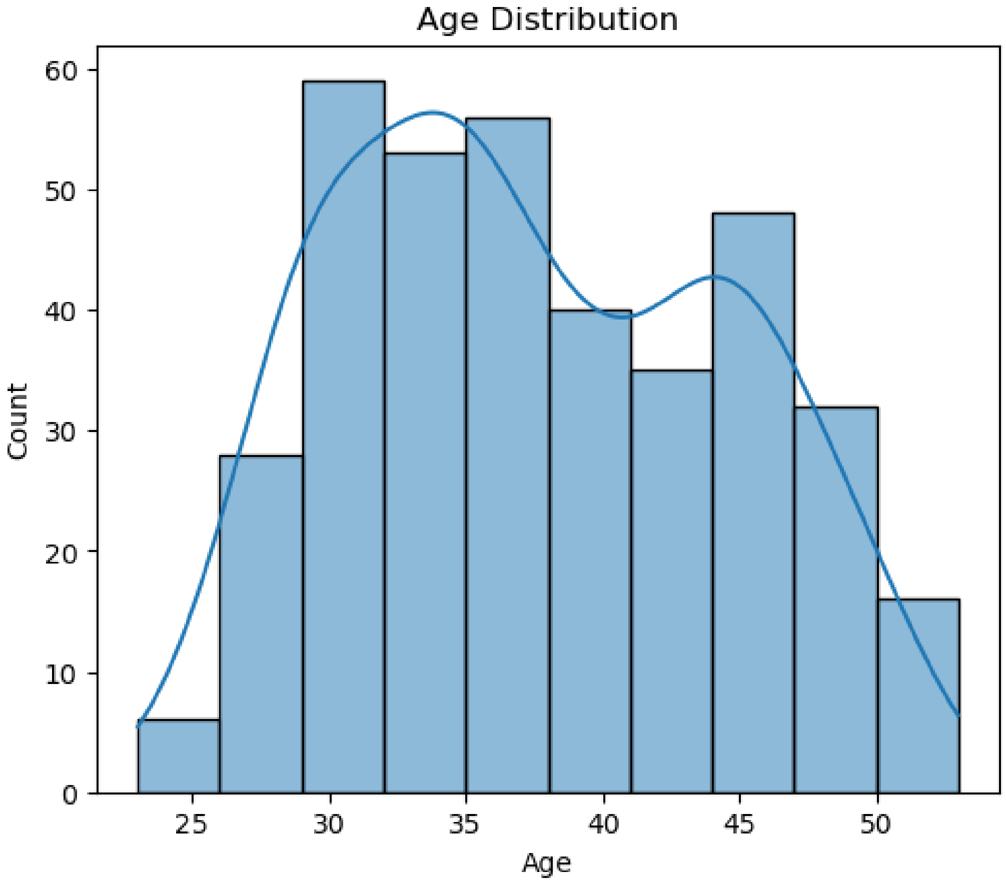
Vidsualizations

```
1.Analyze age distribution [Histogram]

df=pd.read_csv(r"C:\Users\DELL\Downloads\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]
```

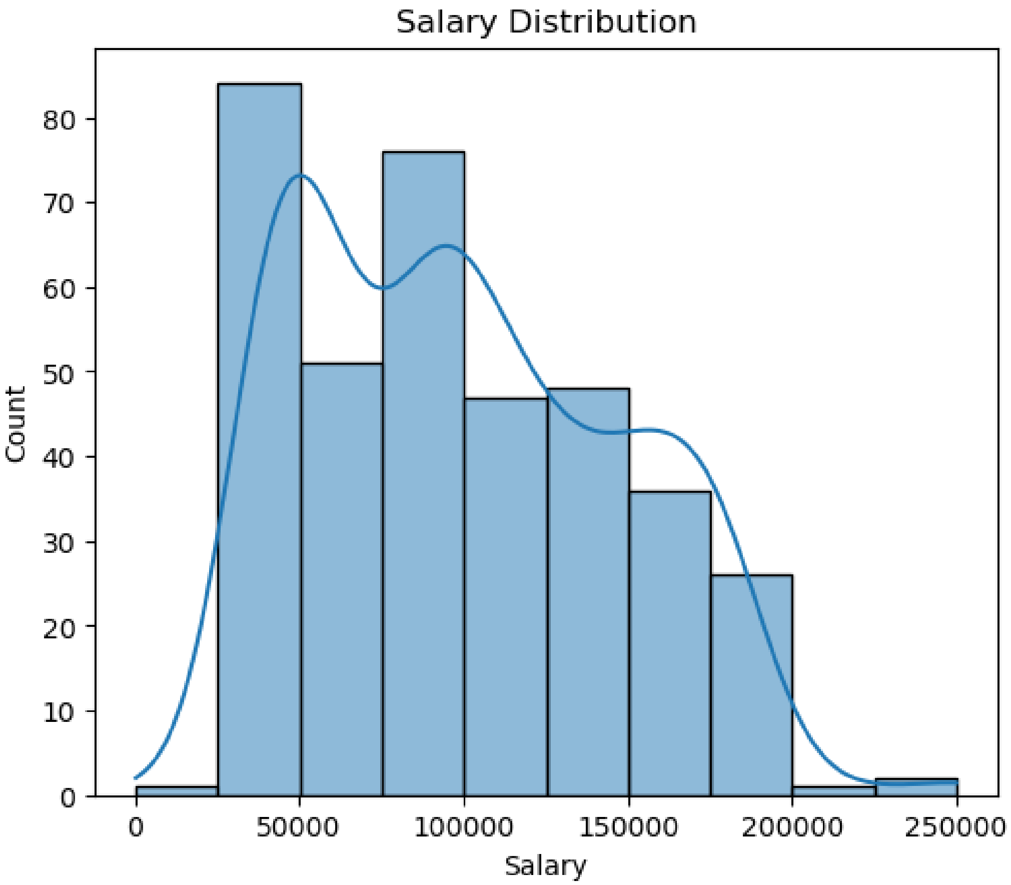
```
plt.figure(figsize=(6,5))
sns.histplot(df["Age"],kde=True,bins=10)
plt.title("Age Distribution")
plt.show()
```

```
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):
```



```
plt.figure(figsize=(6,5))
sns.histplot(df["Salary"],kde=True,bins=10)
plt.title("Salary Distribution")
plt.show()
```

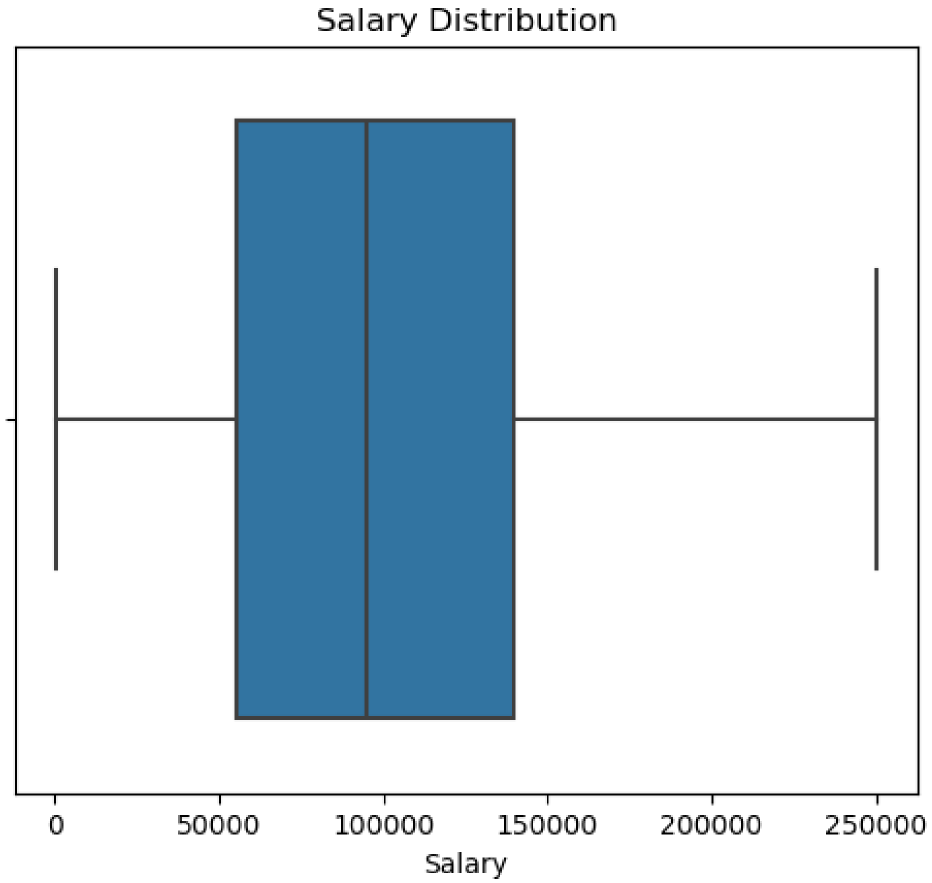
```
C:\Users\DELL\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



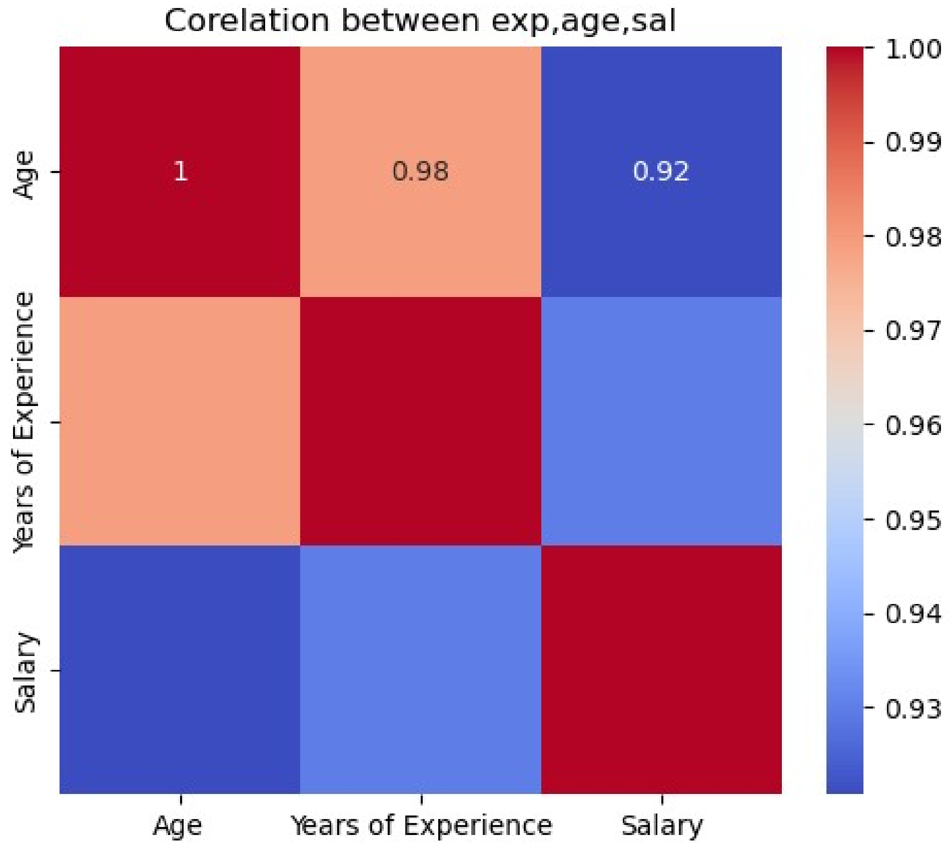
Conclusion -Minimum salary is 350 -maximum experiance is 250000, Average salary is 1L -Majority of salary between 56000 and 1L

- Their might be outliers, min-350,avg-1L,There is lot difference(error,part-time)

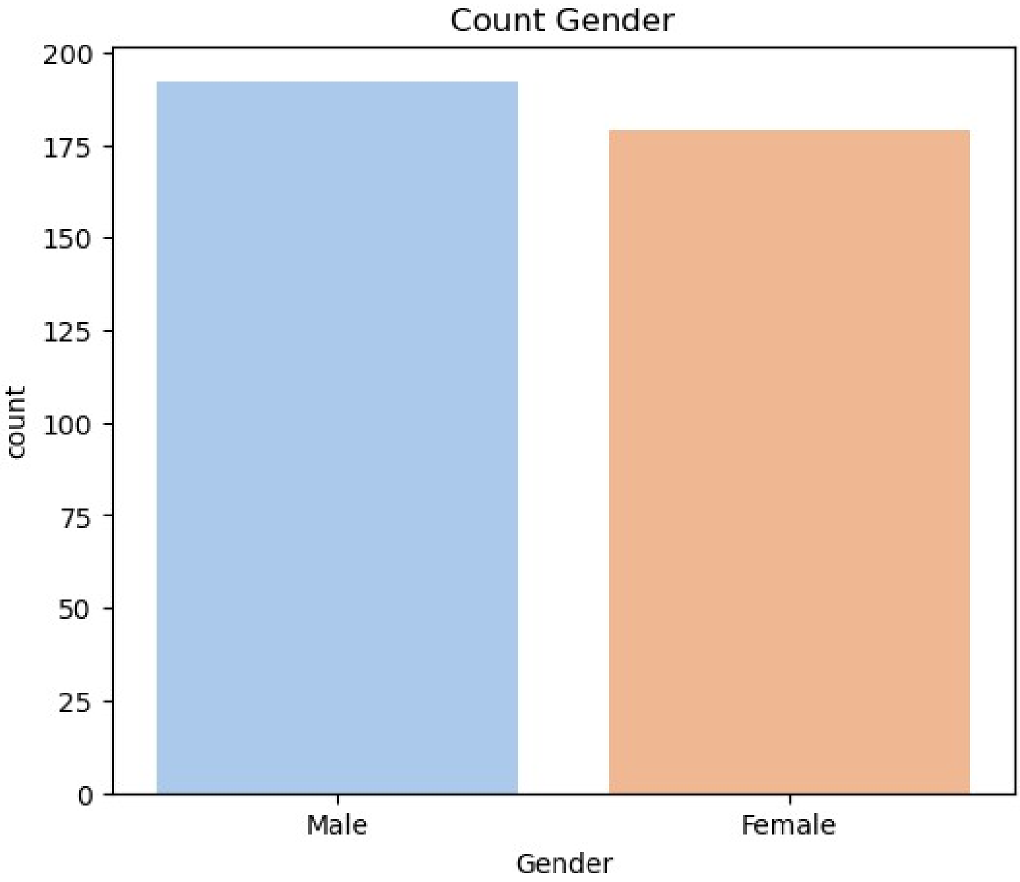
```
plt.figure(figsize = (6,5))
sns.boxplot(x = df["Salary"])
plt.title("Salary Distribution")
plt.show()
```



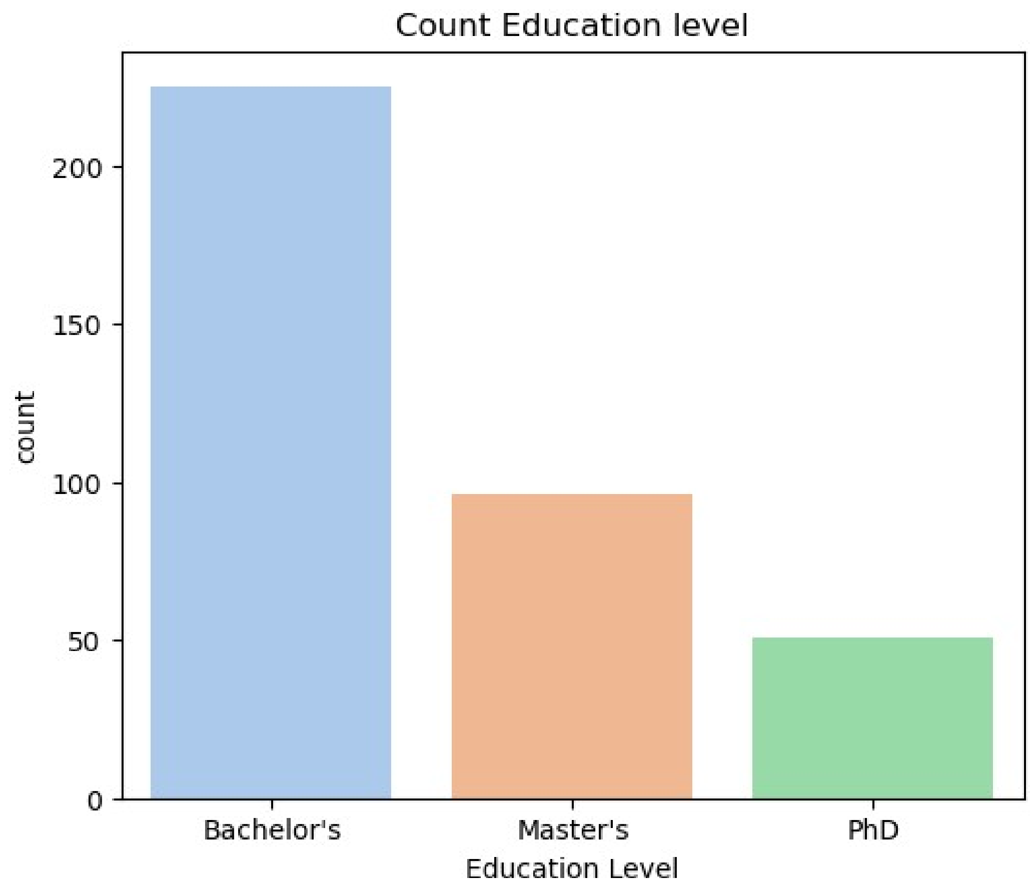
```
ndf=df.select_dtypes(include = ["number"])
ndf.head()

    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(figsize=(6,5))
sns.heatmap(ndf.corr(),cmap='coolwarm',annot=True)
plt.title("Corelation between exp,age,sal")
plt.show()
```



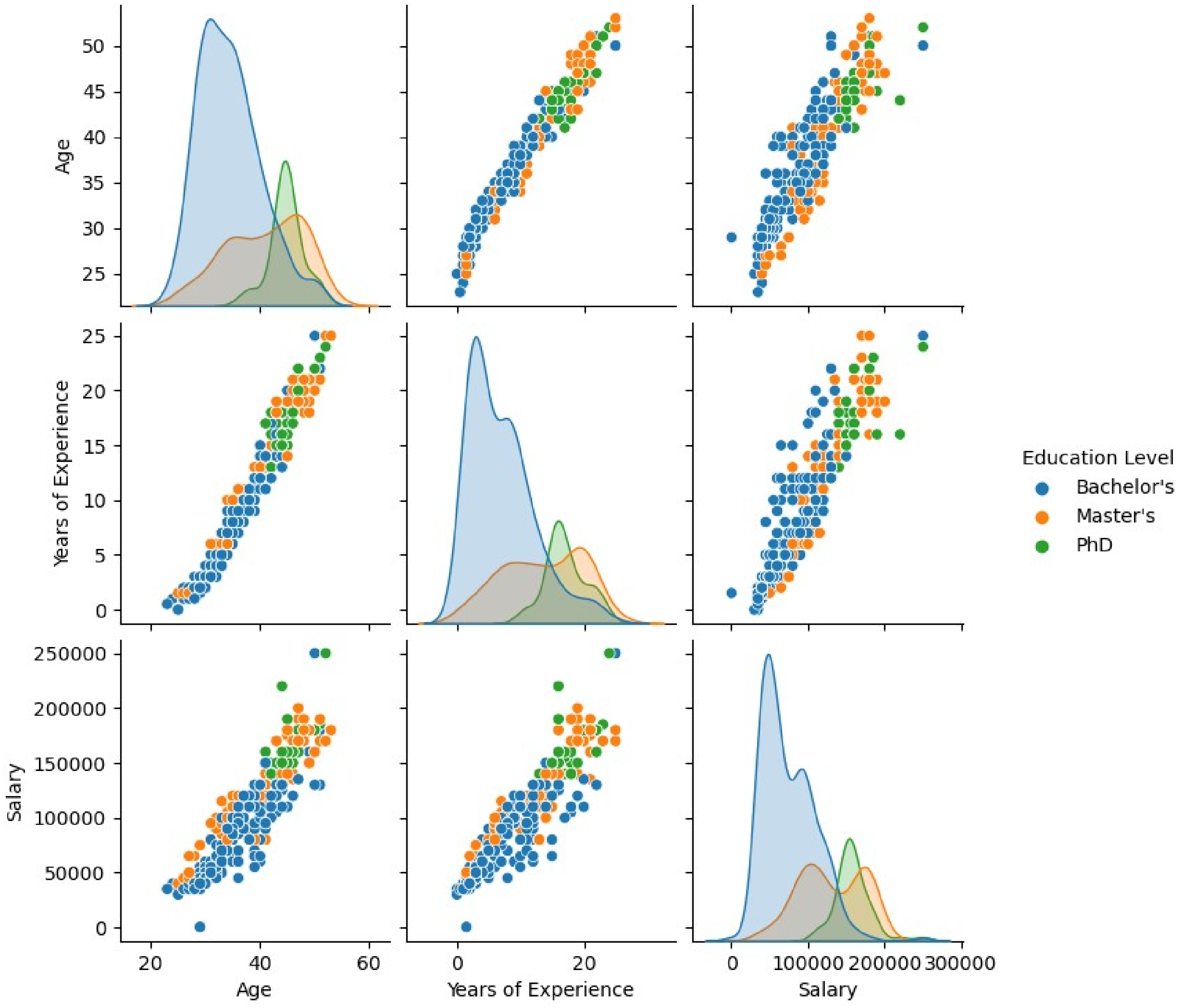
```
plt.figure(figsize=(6,5))
sns.countplot(x = df['Gender'],palette='pastel')
plt.title("Count Gender")
plt.show()
```



```python
plt.figure(figsize=(6,5))
sns.countplot(x = df['Education Level'],palette='pastel')
plt.title("Count Education level")
plt.show()
```



```
sns.pairplot(df,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 0x275a5132750>
```



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

gf

Education Level
Bachelor's      74465.848214
Master's       129583.333333
PhD            157843.137255
Name: Salary, dtype: float64

f=df[(df["Years of Experience"]>20)]
f['Salary'].mean()

175892.85714285713
```