Importing Libraries

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

Loading and Viewing Data

```
ld=pd.read_csv(r"C:\Mypythonfiles\Salary_EDA.csv")
```

```
ld
```

```
      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]
```

```
ld.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
```

```
ld.isnull().sum()
```

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

```
ld.dropna(inplace=True)
ld.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 non-null values.

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

```
ld.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  
```

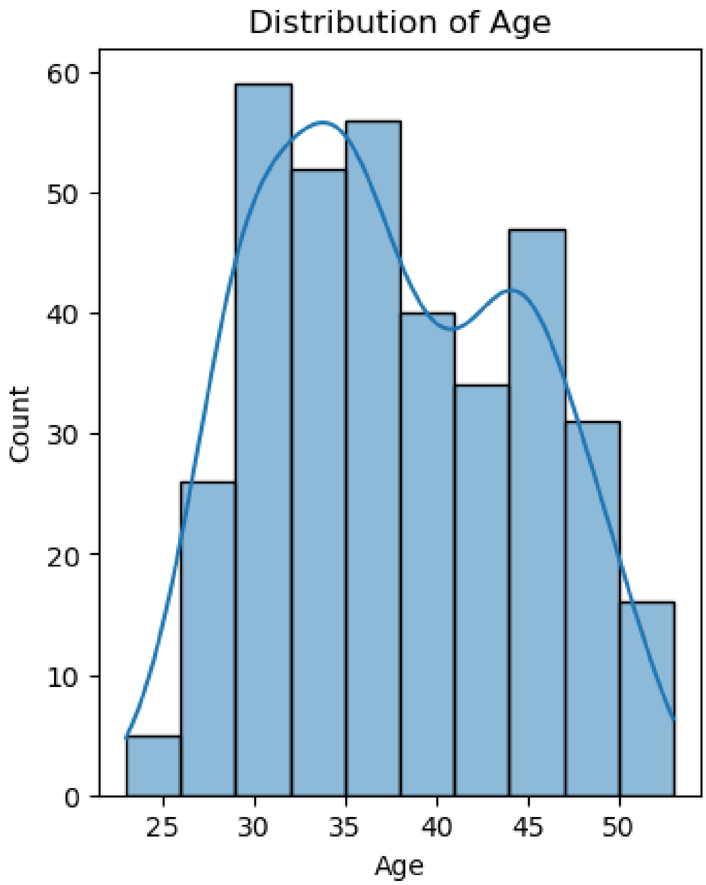
Conclusions:

1. Age,Years of Experience and salary are of datatype float.
2. The average age is approximately 37 years min is 23 and max is 53,Majority range is between 32 and 44.
3. The most frequent gender is Male.
4. The average salary is 100492.
5. The average years of Exprerience is 10 year.
6. NaN-> not applicable for non-numeric values.
7. We have 6 features and 375 rows.
8. salary:there might be outlier.

Visualization

1. Analyse age distribution[Histogram]

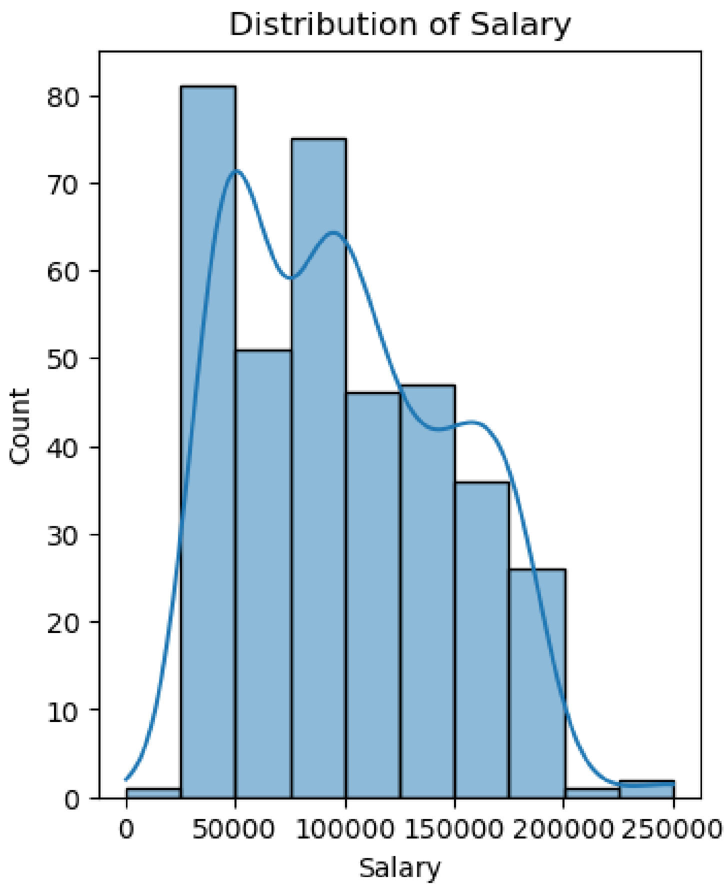
```
plt.figure(figsize=(4,5))
sns.histplot(ld["Age"],kde=True,bins=10)
plt.title("Distribution of Age")
plt.show()
```



Conclusion: majority range is 30. no outerlier exist. minority of people have age of 25. there is a slite positive skew.

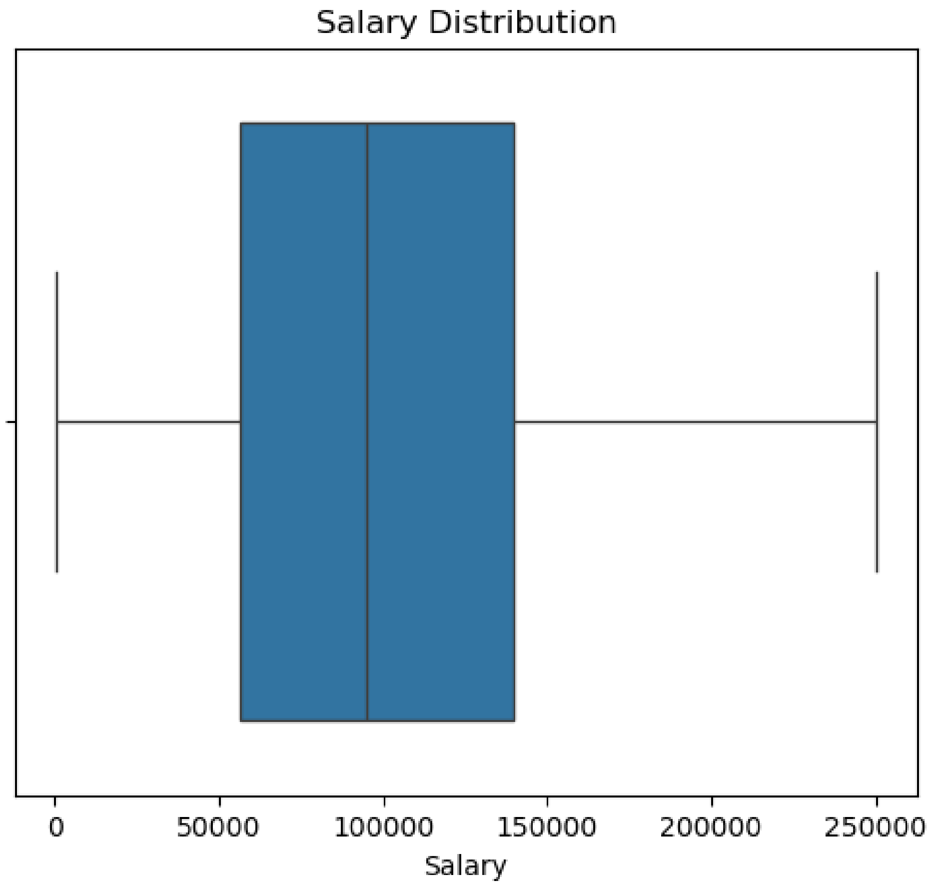
Analyse the distribution of salary usins hist

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



1.positive skew. 2.majority of salary range is 50000. 3.minority of salary range is 250000. 4.no outerlier is exist.

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



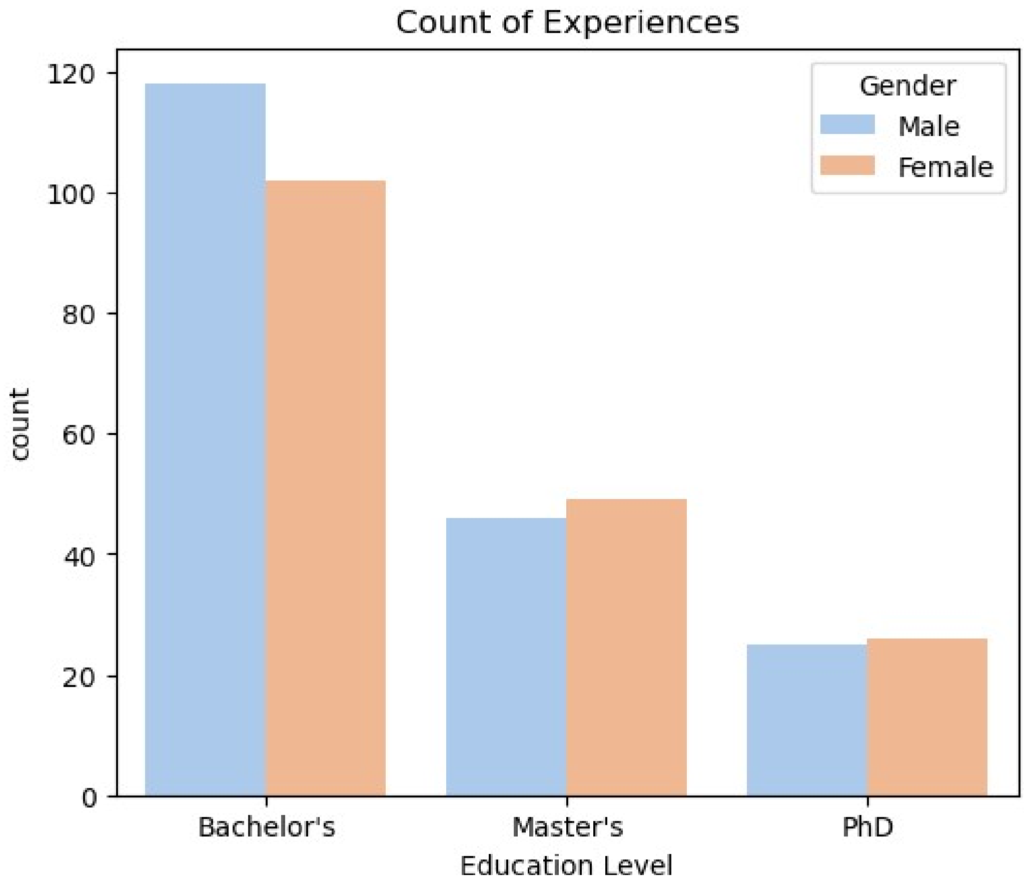
Conclusiopn: 1.No abnormal outerlier is exist. 2.average salary is 100000. 3.upper limit is 150000 and lower limit is 50000.

```
ndf=ld.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

4)

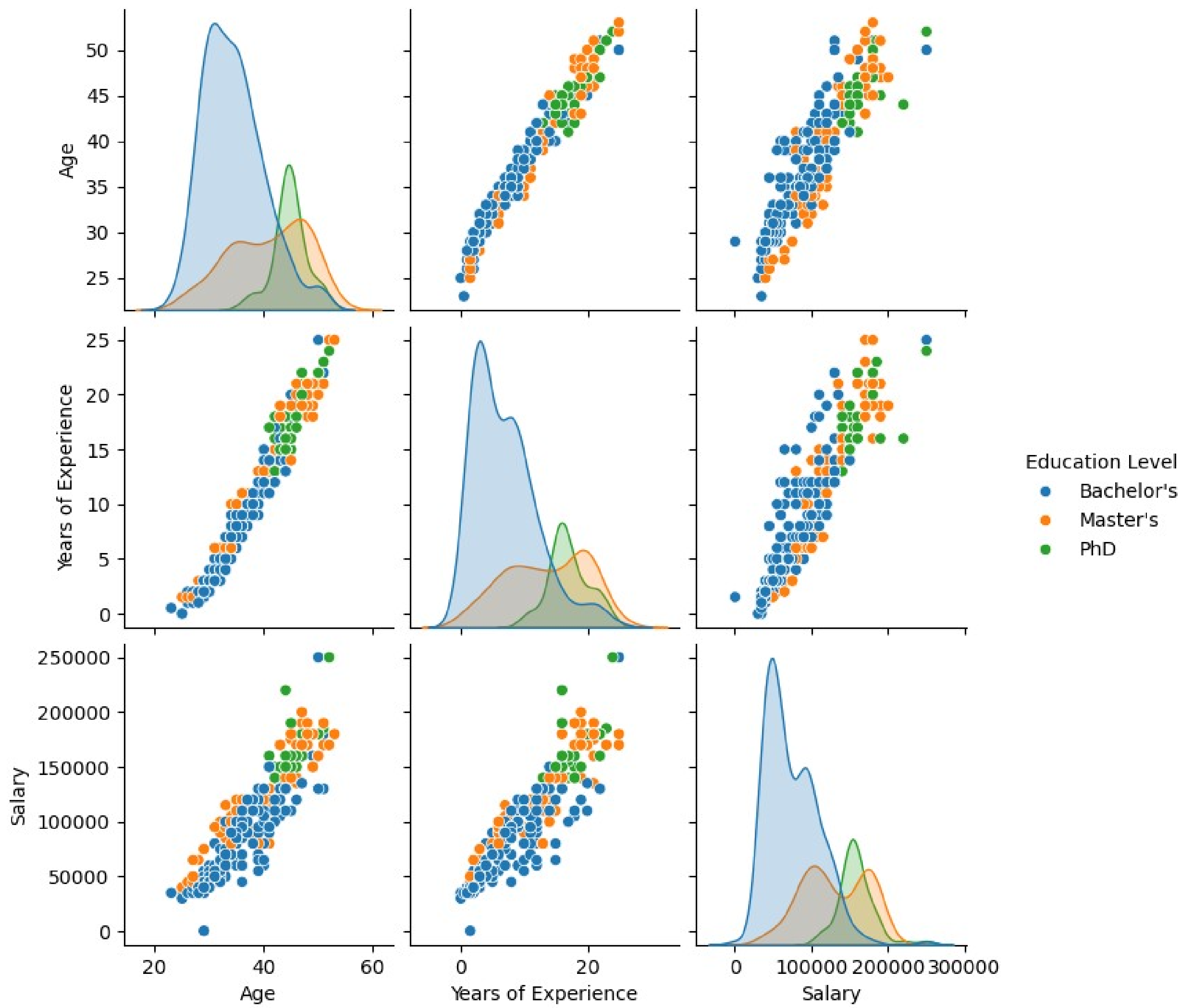
```
plt.figure(figsize=(6,5))
sns.countplot(x=ld["Education
Level"],palette="pastel",hue=ld["Gender"])
plt.title("Count of Experiences")
plt.show()
```



1. It shows the counts of gender snd education level 2.male gender is dominating the Bachuler. 3.Ph.D is the lowest education level,males are dominating.

```
sns.pairplot(ld,hue="Education Level")
```

```
<seaborn.axisgrid.PairGrid at 0x21b40663110>
```



1.It shows the pairplot of the education level of the employee. 2.Age affects the Experince 3.Experience is affects the Salery. 4.Bachelers level is dominating in all the fields of eductaion level. 5.We abserved that age increases 6.The peak salary is ngiven to Bacheler degree people. 7.Employee with bacheler in degree is consistent.

group education level and find average salaery for every category

```
ld.groupby("Education Level")["Salary"].mean()
```

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

```
ld.groupby("Gender")["Salary"].mean()
```

```
Gender
Female     97033.898305
Male      103732.010582
Name: Salary, dtype: float64
```

```
ld.groupby("Age")["Salary"].mean()
```

```
Age
23.0     35000.000000
25.0     35000.000000
26.0     38333.333333
27.0     45000.000000
28.0     41250.000000
29.0     42841.304348
30.0     46666.666667
31.0     54285.714286
32.0     66666.666667
33.0     70625.000000
34.0     90625.000000
35.0     89772.727273
36.0     84545.454545
37.0    103750.000000
38.0    104000.000000
39.0     92916.666667
40.0    103076.923077
41.0    116363.636364
42.0    124090.909091
43.0    141250.000000
44.0    147750.000000
45.0    153529.411765
46.0    151500.000000
47.0    171666.666667
48.0    178125.000000
49.0    170000.000000
50.0    177500.000000
51.0    171000.000000
52.0    210000.000000
53.0    180000.000000
Name: Salary, dtype: float64
```

```
ld.groupby("Years of Experience")["Salary"].mean()
```

```
Years of Experience
0.0      33333.333333
0.5      35000.000000
1.0      36000.000000
1.5      36279.166667
2.0      41833.333333
3.0      51379.310345
4.0      58500.000000
```

```
5.0       63125.000000
6.0       83750.000000
7.0       82000.000000
8.0       88800.000000
9.0      101818.181818
10.0     100555.555556
11.0     100500.000000
12.0     105000.000000
13.0     118000.000000
14.0     125769.230769
15.0     134375.000000
16.0     159411.764706
17.0     143000.000000
18.0     150416.666667
19.0     166333.333333
20.0     166250.000000
21.0     173846.153846
22.0     162222.222222
23.0     177500.000000
24.0     250000.000000
25.0     200000.000000
Name: Salary, dtype: float64
```

```
ld4=ld.select_dtypes(include=["number"])
ld4
```

```
      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
..    ...                  ...       ...
370  35.0                  8.0   85000.0
371  43.0                 19.0  170000.0
372  29.0                  2.0   40000.0
373  34.0                  7.0   90000.0
374  44.0                 15.0  150000.0

[366 rows x 3 columns]
```

```
ld2=ld[(ld["Gender"]=="Female")&(ld["Education Level"]=="Master's")]
ld2["Salary"].mean()
```

```
121020.40816326531
```