In [27]:

```
pip install -U pandasql
```

```
Requirement already satisfied: pandasql in c:\users\user\anaconda3\lib\site-packages (0.
7.3)
Requirement already satisfied: sqlalchemy in c:\users\user\anaconda3\lib\site-packages
(from pandasql) (1.4.39)
Requirement already satisfied: pandas in c:\users\user\anaconda3\lib\site-packages (from
pandasql) (1.4.4)
Requirement already satisfied: numpy in c:\users\user\anaconda3\lib\site-packages (from
pandasql) (1.21.5)
Requirement already satisfied: pytz>=2020.1 in c:\users\user\anaconda3\lib\site-packages
(from pandas->pandasql) (2022.1)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\user\anaconda3\lib\sit
e-packages (from pandas->pandasql) (2.8.2)
Requirement already satisfied: greenlet!=0.4.17 in c:\users\user\anaconda3\lib\site-pack
ages (from sqlalchemy->pandasql) (1.1.1)
Requirement already satisfied: six>=1.5 in c:\users\user\anaconda3\lib\site-packages (fr
om python-dateutil>=2.8.1->pandas->pandasql) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

# Objectives

To perform EDA on the datasets and find:

1. What is the average salary of employees by department?
2. Which department has the highest number of employees?
3. What is the distribution of gender in the company?
4. Is there a correlation between years of experience and salary?
5. Which department has the highest average salary?
6. Other insights

# Overview of the Process Used

1. Used two different linked employee datasets from Kaggle,
2. Formed a data frame required specifically for the first 6 objectives,
3. Used SQL queries using pandasql library to find required results
4. Used various plots to get good Visual Insights from the data

In [28]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import pandasql as psql
```

In [29]:

```
df1=pd.read_csv("C:/Users/User/Downloads/Employee_Salary_Dataset.csv")
df1
```

Out[29]:

| | ID | Experience_Years | Age | Gender | Salary |
|---|---|---|---|---|---|
| **0** | 1 | 5 | 28 | Female | 250000 |
| **1** | 2 | 1 | 21 | Male | 50000 |
| **2** | 3 | 3 | 23 | Female | 170000 |

| | | | | | |
|---|---|---|---|---|---|
| **3** | 4 | 2 | 22 | Male | 25000 |
| **4** | 5 | 1 | 17 | Male | 10000 |
| **5** | 6 | 25 | 62 | Male | 5001000 |
| **6** | 7 | 19 | 54 | Female | 800000 |
| **7** | 8 | 2 | 21 | Female | 9000 |
| **8** | 9 | 10 | 36 | Female | 61500 |
| **9** | 10 | 15 | 54 | Female | 650000 |
| **10** | 11 | 4 | 26 | Female | 250000 |
| **11** | 12 | 6 | 29 | Male | 1400000 |
| **12** | 13 | 14 | 39 | Male | 6000050 |
| **13** | 14 | 11 | 40 | Male | 220100 |
| **14** | 15 | 2 | 23 | Male | 7500 |
| **15** | 16 | 4 | 27 | Female | 87000 |
| **16** | 17 | 10 | 34 | Female | 930000 |
| **17** | 18 | 15 | 54 | Female | 7900000 |
| **18** | 19 | 2 | 21 | Male | 15000 |
| **19** | 20 | 10 | 36 | Male | 330000 |
| **20** | 21 | 15 | 54 | Male | 6570000 |
| **21** | 22 | 4 | 26 | Male | 25000 |
| **22** | 23 | 5 | 29 | Male | 6845000 |
| **23** | 24 | 1 | 21 | Female | 6000 |
| **24** | 25 | 4 | 23 | Female | 8900 |
| **25** | 26 | 3 | 22 | Female | 20000 |
| **26** | 27 | 1 | 18 | Male | 3000 |
| **27** | 28 | 27 | 62 | Female | 10000000 |
| **28** | 29 | 19 | 54 | Female | 5000000 |
| **29** | 30 | 2 | 21 | Female | 6100 |
| **30** | 31 | 10 | 34 | Male | 80000 |
| **31** | 32 | 15 | 54 | Male | 900000 |
| **32** | 33 | 20 | 55 | Female | 1540000 |
| **33** | 34 | 19 | 53 | Female | 9300000 |
| **34** | 35 | 16 | 49 | Male | 7600000 |

In [30]:

```
df2=pd.read_csv("C:/Users/User/Downloads/Department_Dataset.csv")
df2
```

Out[30]:

| | ID | Dept_name | location | travel_required |
|---|---|---|---|---|
| **0** | 1 | HR | Pune | yes |
| **1** | 2 | Finance | Bangalore | no |

|  |  |  |  |  |
|---|---|---|---|---|
| **2** | 3 | Finance | Bangalore | no |
| **3** | 4 | Finance | Pune | no |
| **4** | 5 | Tech | Mumbai | no |
| **5** | 6 | Tech | Pune | no |
| **6** | 7 | Tech | Bangalore | yes |
| **7** | 8 | HR | Bangalore | no |
| **8** | 9 | HR | Pune | no |
| **9** | 10 | HR | Pune | no |
| **10** | 11 | HR | Mumbai | no |
| **11** | 12 | HR | Mumbai | yes |
| **12** | 13 | Finance | Bangalore | yes |
| **13** | 14 | Tech | Bangalore | yes |
| **14** | 15 | Tech | Mumbai | yes |
| **15** | 16 | Tech | Pune | yes |
| **16** | 17 | Tech | Bangalore | no |
| **17** | 18 | Finance | Mumbai | no |
| **18** | 19 | HR | Mumbai | no |
| **19** | 20 | Finance | Bangalore | no |
| **20** | 21 | Tech | Mumbai | no |
| **21** | 22 | Tech | Mumbai | yes |
| **22** | 23 | Tech | Mumbai | no |
| **23** | 24 | Tech | Pune | yes |
| **24** | 25 | Finance | Pune | yes |
| **25** | 26 | HR | Pune | no |
| **26** | 27 | HR | Bangalore | no |
| **27** | 28 | HR | Bangalore | no |
| **28** | 29 | Finance | Bangalore | no |
| **29** | 30 | Finance | Mumbai | no |
| **30** | 31 | Tech | Mumbai | no |
| **31** | 32 | Tech | Pune | yes |
| **32** | 33 | HR | Mumbai | yes |
| **33** | 34 | HR | Bangalore | yes |
| **34** | 35 | Tech | Bangalore | no |

In [31]:

```
df3=df1.merge(df2)
df3
```

Out[31]:

|  | ID | Experience_Years | Age | Gender | Salary | Dept_name | location | travel_required |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 5 | 28 | Female | 250000 | HR | Pune | yes |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 1 | 21 | Male | 50000 | Finance | Bangalore | no |
| **2** | 3 | 3 | 23 | Female | 170000 | Finance | Bangalore | no |
| **3** | 4 | 2 | 22 | Male | 25000 | Finance | Pune | no |
| **4** | 5 | 1 | 17 | Male | 10000 | Tech | Mumbai | no |
| **5** | 6 | 25 | 62 | Male | 5001000 | Tech | Pune | no |
| **6** | 7 | 19 | 54 | Female | 800000 | Tech | Bangalore | yes |
| **7** | 8 | 2 | 21 | Female | 9000 | HR | Bangalore | no |
| **8** | 9 | 10 | 36 | Female | 61500 | HR | Pune | no |
| **9** | 10 | 15 | 54 | Female | 650000 | HR | Pune | no |
| **10** | 11 | 4 | 26 | Female | 250000 | HR | Mumbai | no |
| **11** | 12 | 6 | 29 | Male | 1400000 | HR | Mumbai | yes |
| **12** | 13 | 14 | 39 | Male | 6000050 | Finance | Bangalore | yes |
| **13** | 14 | 11 | 40 | Male | 220100 | Tech | Bangalore | yes |
| **14** | 15 | 2 | 23 | Male | 7500 | Tech | Mumbai | yes |
| **15** | 16 | 4 | 27 | Female | 87000 | Tech | Pune | yes |
| **16** | 17 | 10 | 34 | Female | 930000 | Tech | Bangalore | no |
| **17** | 18 | 15 | 54 | Female | 7900000 | Finance | Mumbai | no |
| **18** | 19 | 2 | 21 | Male | 15000 | HR | Mumbai | no |
| **19** | 20 | 10 | 36 | Male | 330000 | Finance | Bangalore | no |
| **20** | 21 | 15 | 54 | Male | 6570000 | Tech | Mumbai | no |
| **21** | 22 | 4 | 26 | Male | 25000 | Tech | Mumbai | yes |
| **22** | 23 | 5 | 29 | Male | 6845000 | Tech | Mumbai | no |
| **23** | 24 | 1 | 21 | Female | 6000 | Tech | Pune | yes |
| **24** | 25 | 4 | 23 | Female | 8900 | Finance | Pune | yes |
| **25** | 26 | 3 | 22 | Female | 20000 | HR | Pune | no |
| **26** | 27 | 1 | 18 | Male | 3000 | HR | Bangalore | no |
| **27** | 28 | 27 | 62 | Female | 10000000 | HR | Bangalore | no |
| **28** | 29 | 19 | 54 | Female | 5000000 | Finance | Bangalore | no |
| **29** | 30 | 2 | 21 | Female | 6100 | Finance | Mumbai | no |
| **30** | 31 | 10 | 34 | Male | 80000 | Tech | Mumbai | no |
| **31** | 32 | 15 | 54 | Male | 900000 | Tech | Pune | yes |
| **32** | 33 | 20 | 55 | Female | 1540000 | HR | Mumbai | yes |
| **33** | 34 | 19 | 53 | Female | 9300000 | HR | Bangalore | yes |
| **34** | 35 | 16 | 49 | Male | 7600000 | Tech | Bangalore | no |

In [32]:

```
req_df=df3[['ID','Dept_name','Gender','Experience_Years','Salary']]
req_df
```

Out[32]:

| | ID | Dept_name | Gender | Experience_Years | Salary |
|---|---|---|---|---|---|

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| **0** | 1 | HR | Female | 5 | 250000 |
| **1** | 2 | Finance | Male | 1 | 50000 |
| **2** | 3 | Finance | Female | 3 | 170000 |
| **3** | 4 | Finance | Male | 2 | 25000 |
| **4** | 5 | Tech | Male | 1 | 10000 |
| **5** | 6 | Tech | Male | 25 | 5001000 |
| **6** | 7 | Tech | Female | 19 | 800000 |
| **7** | 8 | HR | Female | 2 | 9000 |
| **8** | 9 | HR | Female | 10 | 61500 |
| **9** | 10 | HR | Female | 15 | 650000 |
| **10** | 11 | HR | Female | 4 | 250000 |
| **11** | 12 | HR | Male | 6 | 1400000 |
| **12** | 13 | Finance | Male | 14 | 6000050 |
| **13** | 14 | Tech | Male | 11 | 220100 |
| **14** | 15 | Tech | Male | 2 | 7500 |
| **15** | 16 | Tech | Female | 4 | 87000 |
| **16** | 17 | Tech | Female | 10 | 930000 |
| **17** | 18 | Finance | Female | 15 | 7900000 |
| **18** | 19 | HR | Male | 2 | 15000 |
| **19** | 20 | Finance | Male | 10 | 330000 |
| **20** | 21 | Tech | Male | 15 | 6570000 |
| **21** | 22 | Tech | Male | 4 | 25000 |
| **22** | 23 | Tech | Male | 5 | 6845000 |
| **23** | 24 | Tech | Female | 1 | 6000 |
| **24** | 25 | Finance | Female | 4 | 8900 |
| **25** | 26 | HR | Female | 3 | 20000 |
| **26** | 27 | HR | Male | 1 | 3000 |
| **27** | 28 | HR | Female | 27 | 10000000 |
| **28** | 29 | Finance | Female | 19 | 5000000 |
| **29** | 30 | Finance | Female | 2 | 6100 |
| **30** | 31 | Tech | Male | 10 | 80000 |
| **31** | 32 | Tech | Male | 15 | 900000 |
| **32** | 33 | HR | Female | 20 | 1540000 |
| **33** | 34 | HR | Female | 19 | 9300000 |
| **34** | 35 | Tech | Male | 16 | 7600000 |

In [33]:

```
q1="select * from req_df"
psql.sqldf(q1)
```

Out[33]:

| | ID | Dept_name | Gender | Experience_Years | Salary |
|---|---|---|---|---|---|
| **0** | 1 | HR | Female | 5 | 250000 |
| **1** | 2 | Finance | Male | 1 | 50000 |
| **2** | 3 | Finance | Female | 3 | 170000 |
| **3** | 4 | Finance | Male | 2 | 25000 |
| **4** | 5 | Tech | Male | 1 | 10000 |
| **5** | 6 | Tech | Male | 25 | 5001000 |
| **6** | 7 | Tech | Female | 19 | 800000 |
| **7** | 8 | HR | Female | 2 | 9000 |
| **8** | 9 | HR | Female | 10 | 61500 |
| **9** | 10 | HR | Female | 15 | 650000 |
| **10** | 11 | HR | Female | 4 | 250000 |
| **11** | 12 | HR | Male | 6 | 1400000 |
| **12** | 13 | Finance | Male | 14 | 6000050 |
| **13** | 14 | Tech | Male | 11 | 220100 |
| **14** | 15 | Tech | Male | 2 | 7500 |
| **15** | 16 | Tech | Female | 4 | 87000 |
| **16** | 17 | Tech | Female | 10 | 930000 |
| **17** | 18 | Finance | Female | 15 | 7900000 |
| **18** | 19 | HR | Male | 2 | 15000 |
| **19** | 20 | Finance | Male | 10 | 330000 |
| **20** | 21 | Tech | Male | 15 | 6570000 |
| **21** | 22 | Tech | Male | 4 | 25000 |
| **22** | 23 | Tech | Male | 5 | 6845000 |
| **23** | 24 | Tech | Female | 1 | 6000 |
| **24** | 25 | Finance | Female | 4 | 8900 |
| **25** | 26 | HR | Female | 3 | 20000 |
| **26** | 27 | HR | Male | 1 | 3000 |
| **27** | 28 | HR | Female | 27 | 10000000 |
| **28** | 29 | Finance | Female | 19 | 5000000 |
| **29** | 30 | Finance | Female | 2 | 6100 |
| **30** | 31 | Tech | Male | 10 | 80000 |
| **31** | 32 | Tech | Male | 15 | 900000 |
| **32** | 33 | HR | Female | 20 | 1540000 |
| **33** | 34 | HR | Female | 19 | 9300000 |
| **34** | 35 | Tech | Male | 16 | 7600000 |

In [34]:

```
q2="select Dept_name as Department_Name, ROUND(AVG(Salary),2) as Average_Salary from req
```

```
psql.sqldf(q2)
```

Out[34]:

| | Department_Name | Average_Salary |
|---|---|---|
| **0** | Finance | 2165561.11 |
| **1** | HR | 1958208.33 |
| **2** | Tech | 2077257.14 |

In [35]:

```
q3="select Dept_name, count(*) as Number_of_Employees from req_df group by Dept_name"
psql.sqldf(q3)
```

Out[35]:

| | Dept_name | Number_of_Employees |
|---|---|---|
| **0** | Finance | 9 |
| **1** | HR | 12 |
| **2** | Tech | 14 |

In [36]:

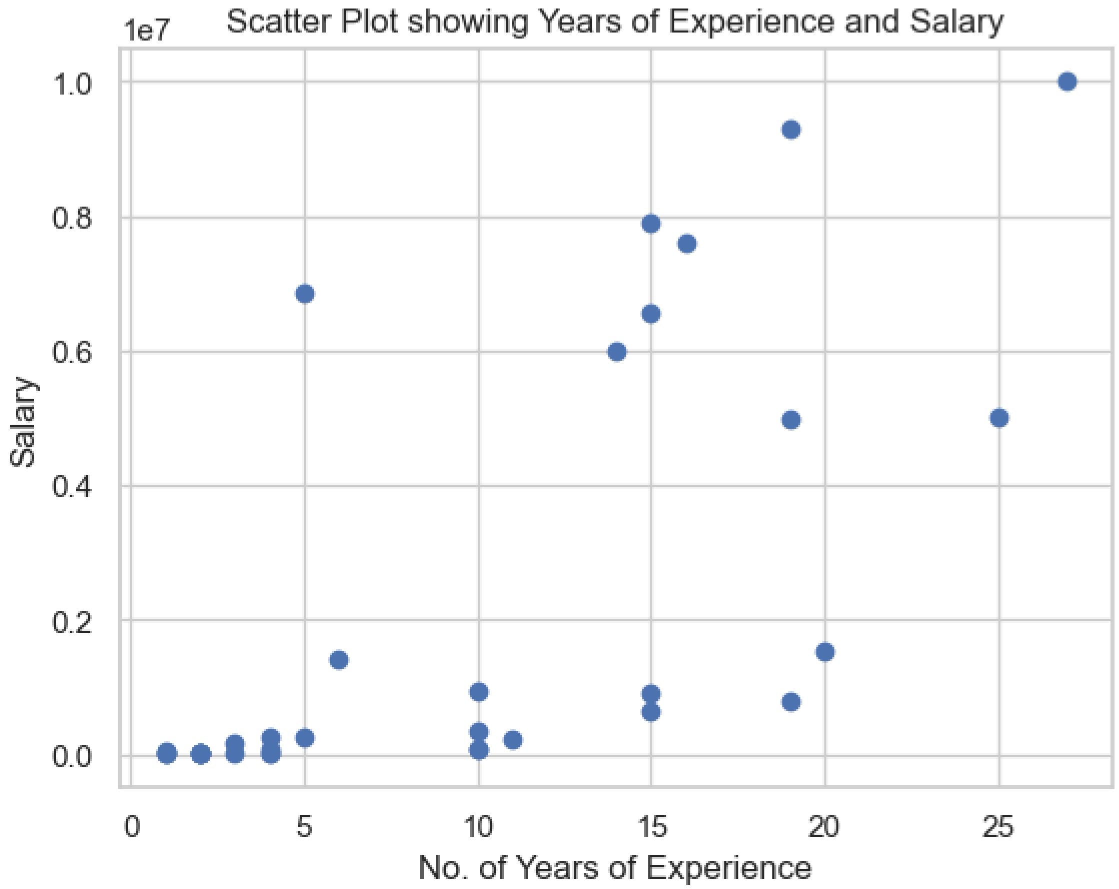
```
q4="select Dept_name, round(avg(Salary),2) as avg_salary from req_df group by Dept_name
psql.sqldf(q4)
```

Out[36]:

| | Dept_name | avg_salary |
|---|---|---|
| **0** | Finance | 2165561.11 |
| **1** | Tech | 2077257.14 |
| **2** | HR | 1958208.33 |

In [37]:

```
plt.scatter(req_df['Experience_Years'],req_df['Salary'])
plt.xlabel('No. of Years of Experience')
plt.ylabel('Salary')
plt.title('Scatter Plot showing Years of Experience and Salary')
plt.show()
```



In [38]:

```
corr=req_df['Experience_Years'].corr(req_df['Salary'])
corr
```

Out[38]:

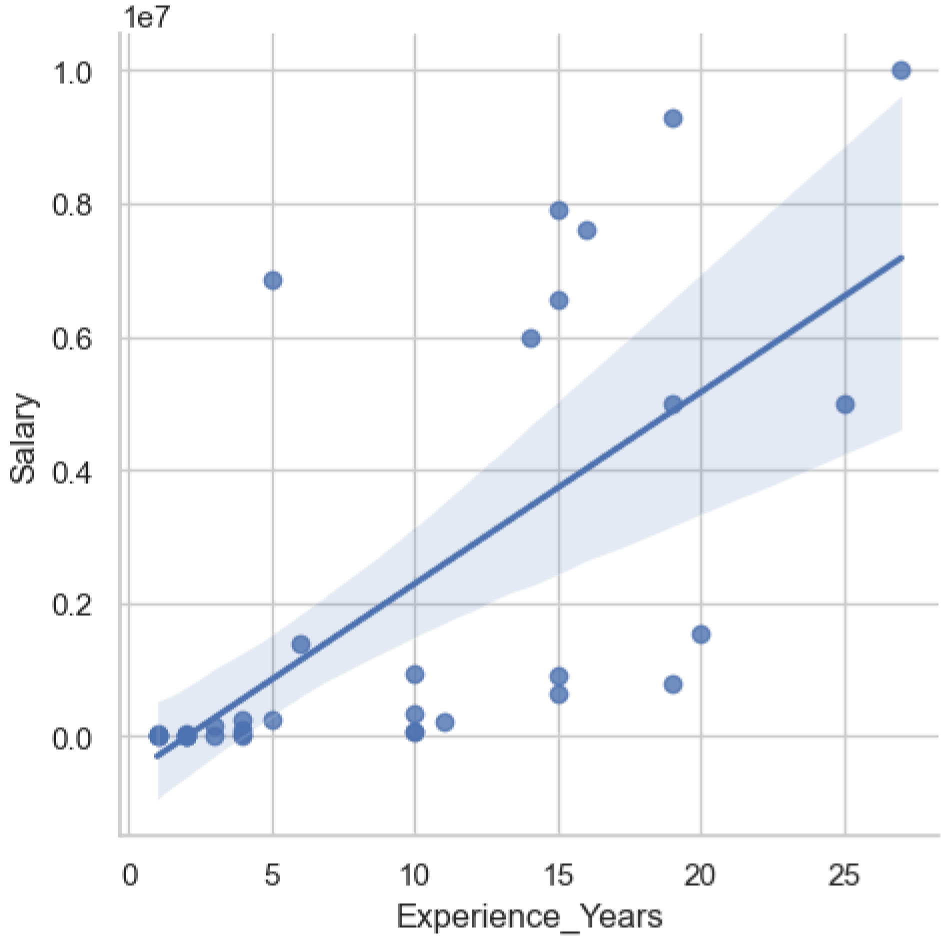
```
0.6855999775494617
```

In [39]:

```
sns.lmplot(x='Experience_Years',y='Salary',data=req_df)
```

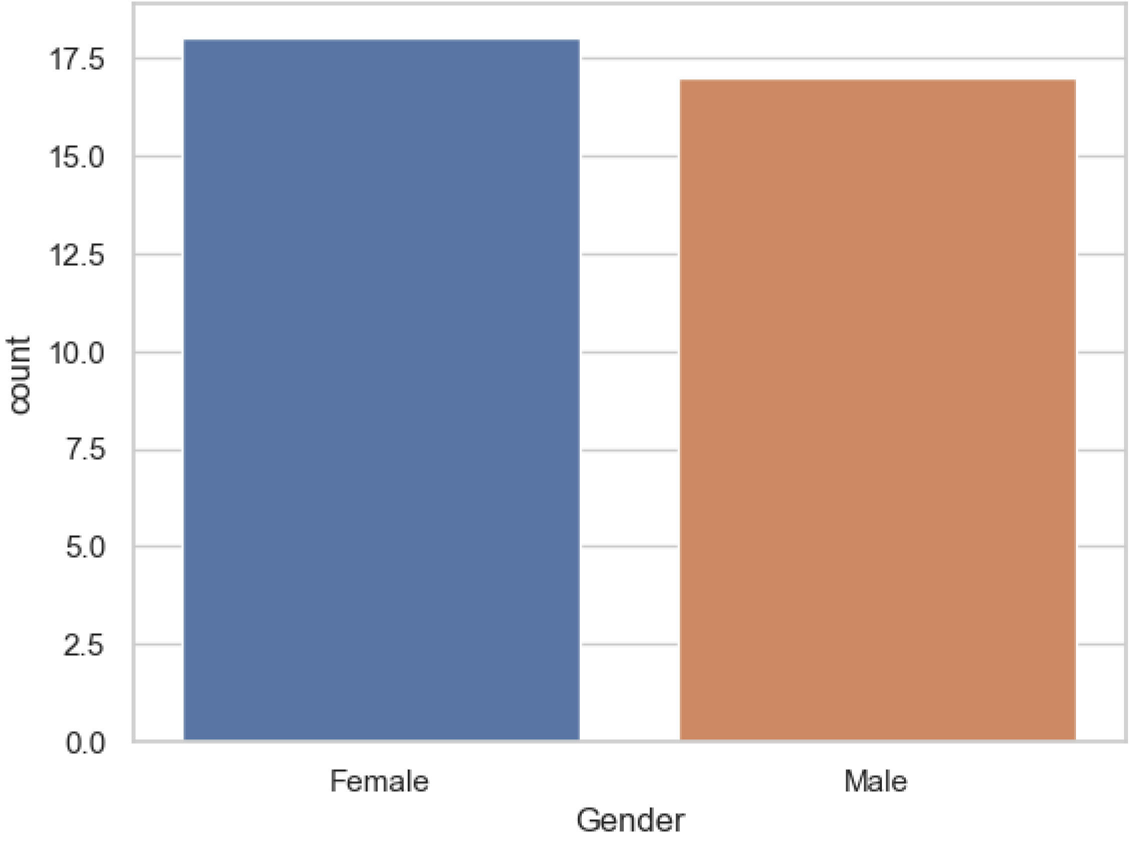
Out[39]:

```
<seaborn.axisgrid.FacetGrid at 0x2139c1c1dc0>
```

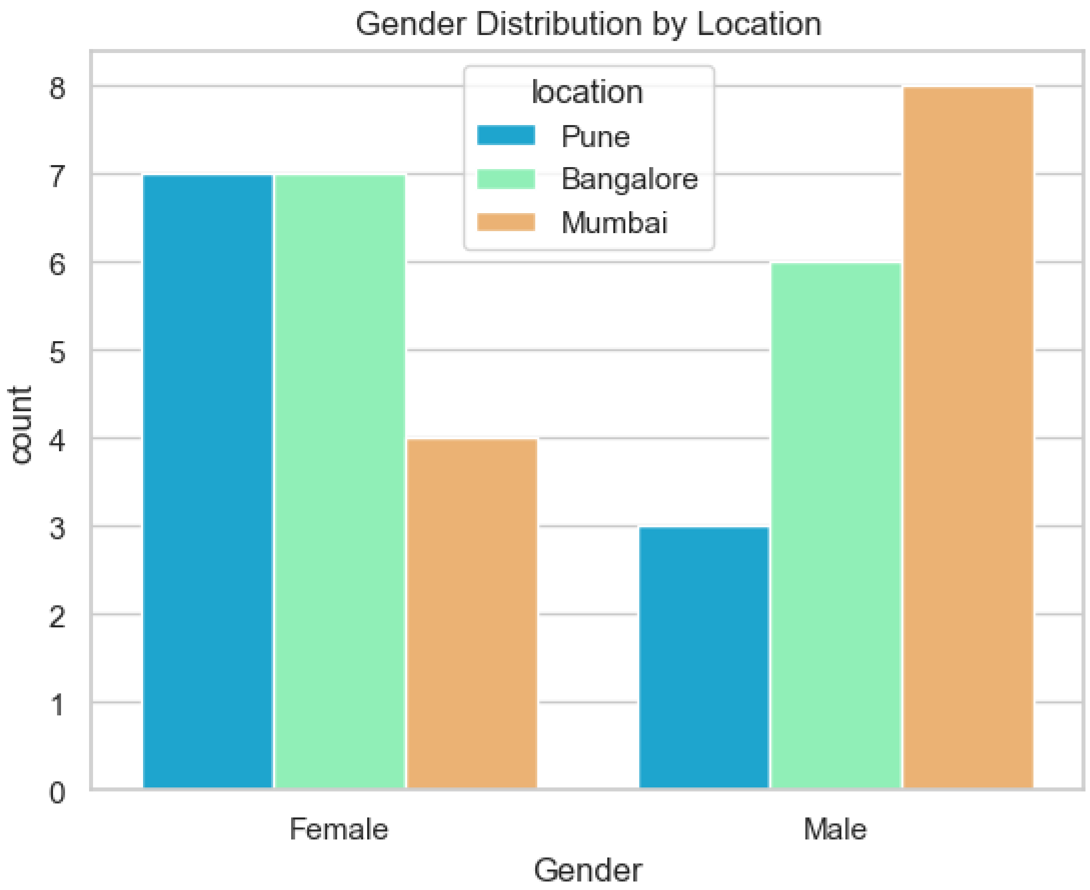


In [40]:
```
sns.countplot(x='Gender',data=req_df)
```

Out[40]:
```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```



In [41]:
```
sns.countplot(x='Gender',hue='location',data=df3, palette='rainbow')
plt.title('Gender Distribution by Location')
plt.show()
```

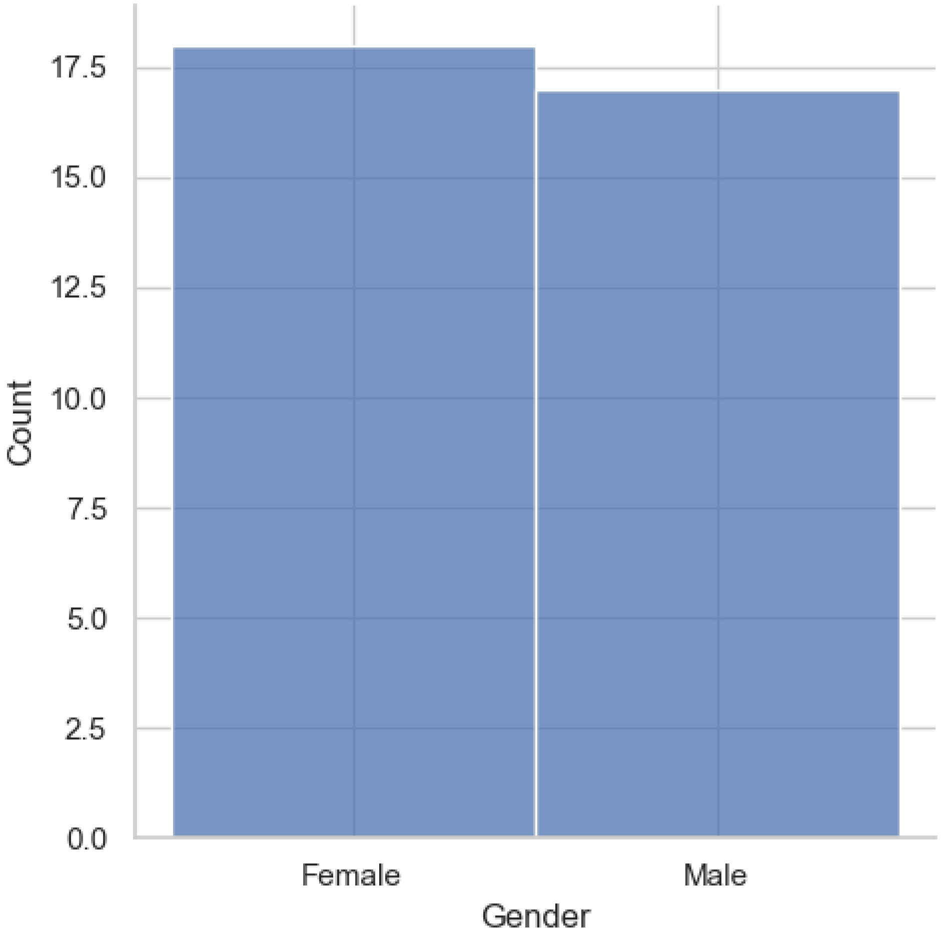


In [42]:

```
sns.displot(req_df['Gender'])
```

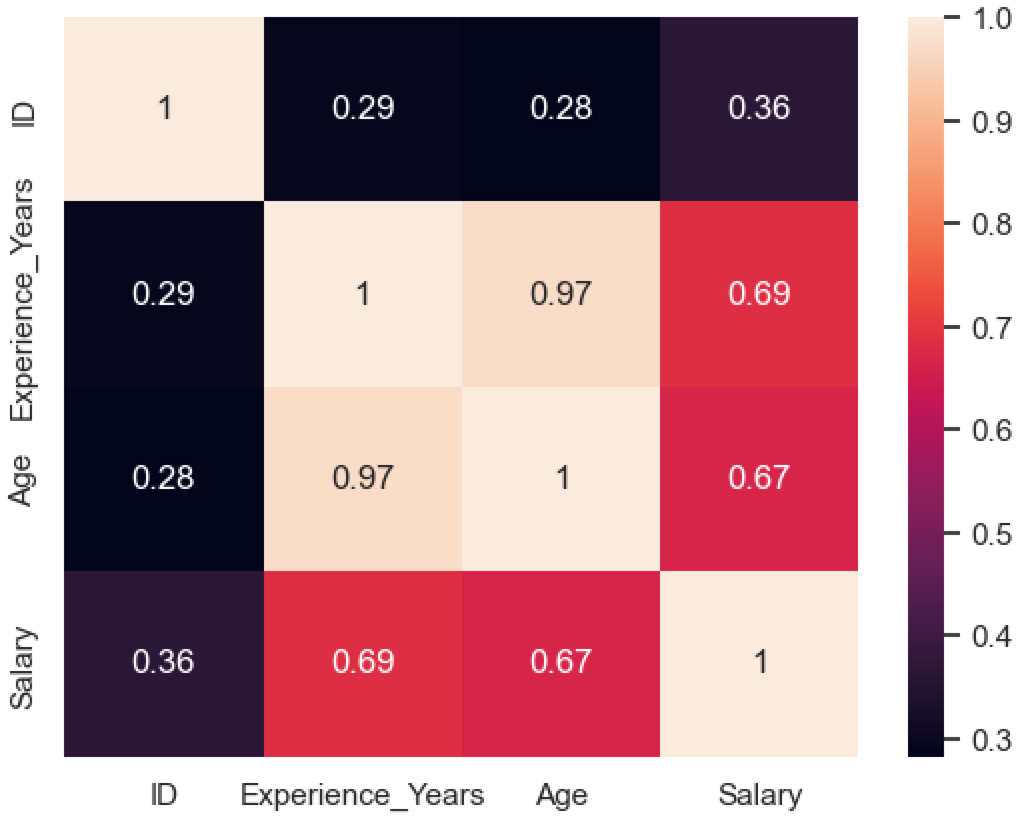
Out[42]:

```
<seaborn.axisgrid.FacetGrid at 0x2139c21c3a0>
```



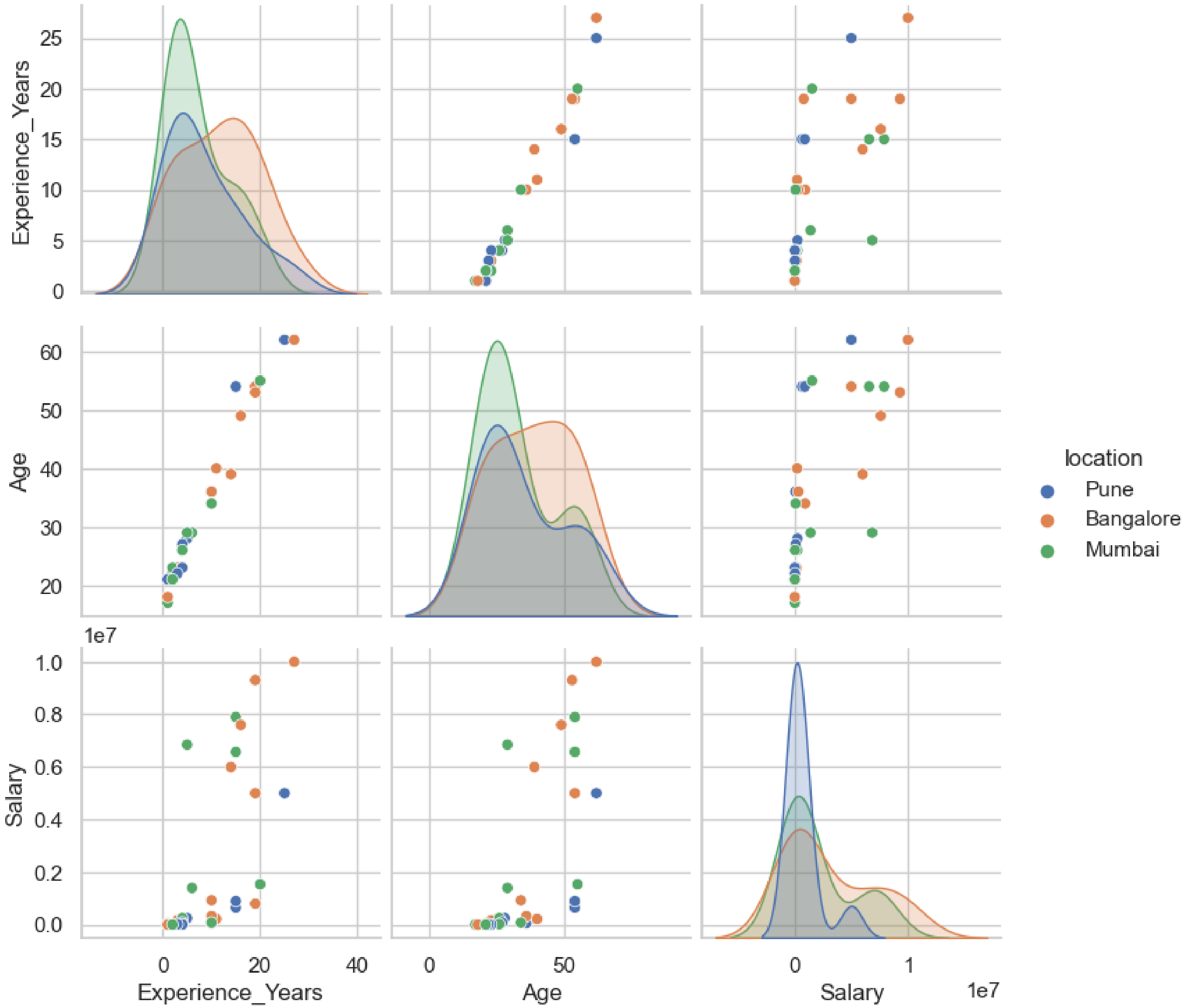
```
In [43]: sns.heatmap(df3.corr(),annot=True)
```

```
Out[43]: <AxesSubplot:>
```



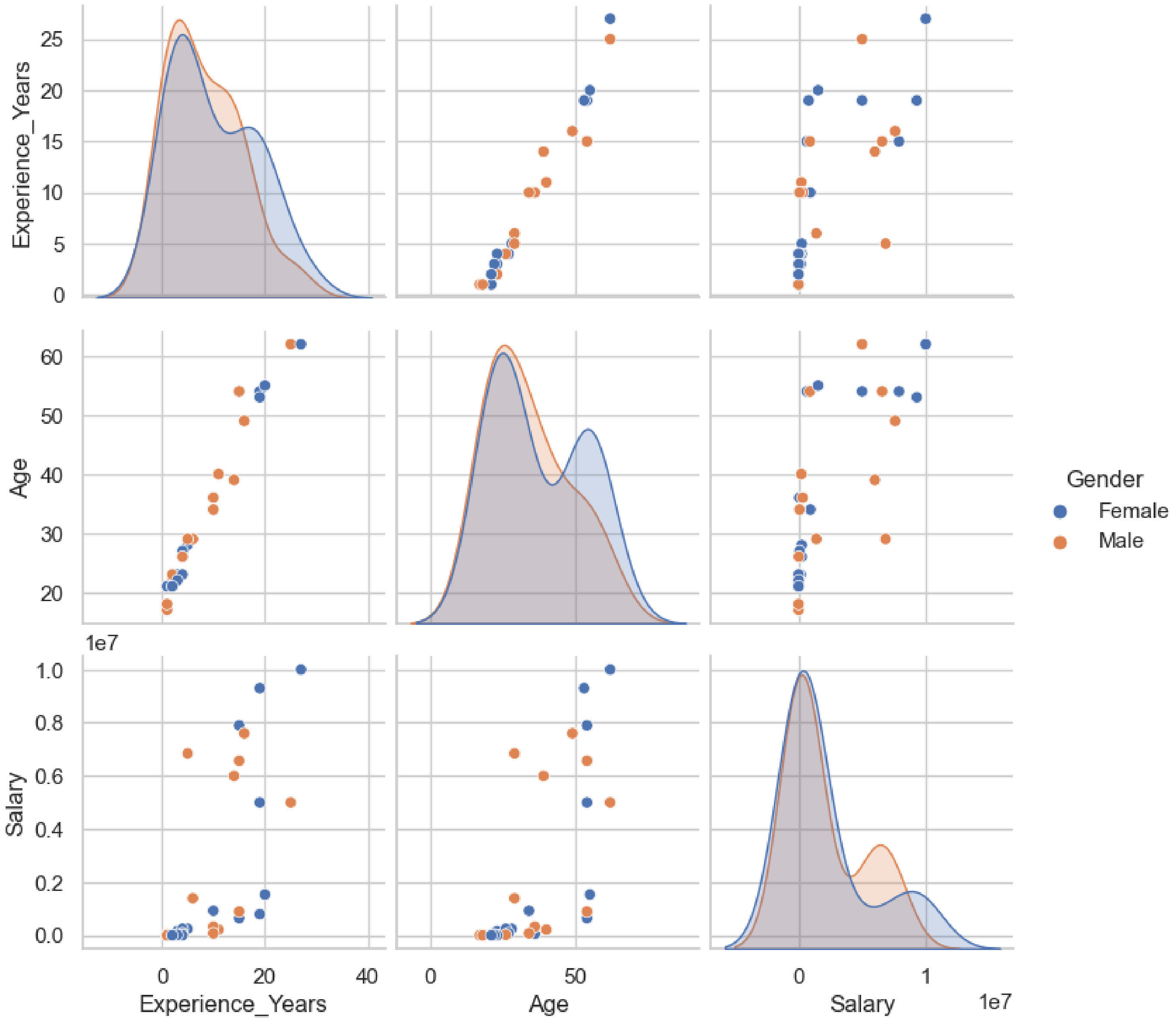
```
In [44]: sns.pairplot(df3[['Experience_Years','Age','Salary','location']],hue='location')
```

```
Out[44]: <seaborn.axisgrid.PairGrid at 0x2139c66b6a0>
```



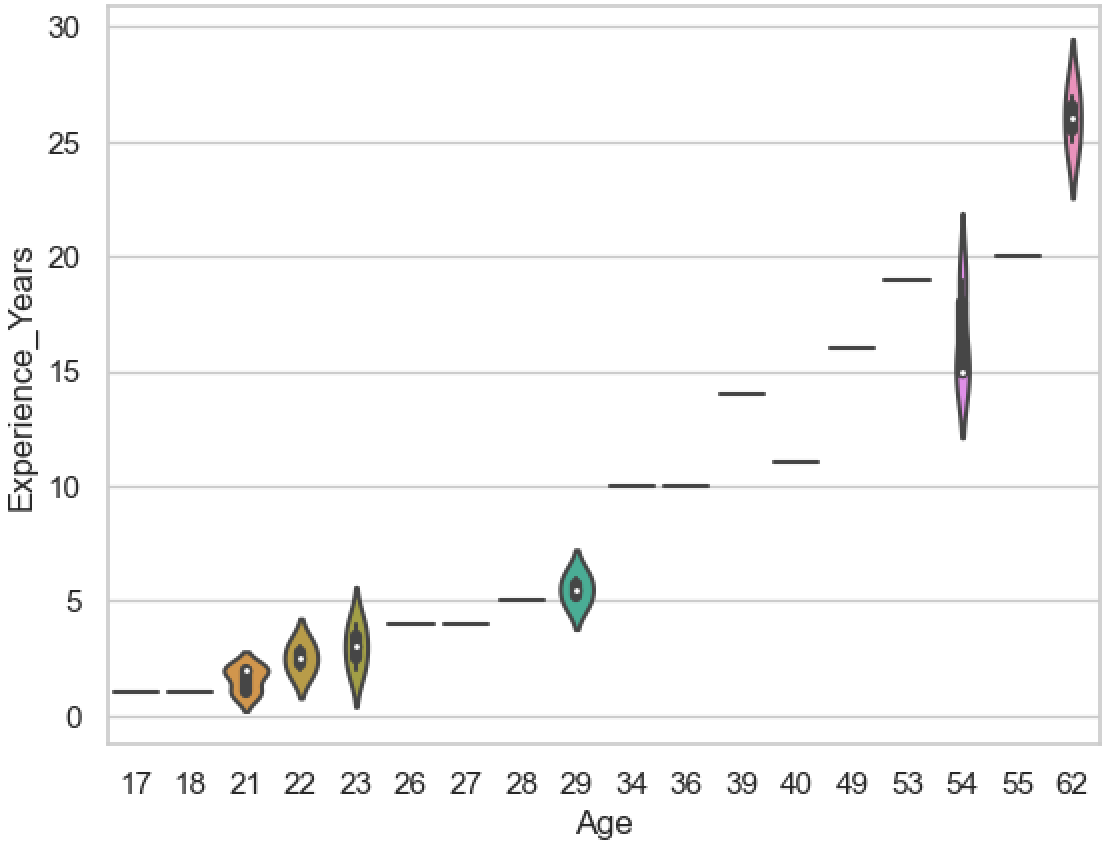
```
In [45]: sns.pairplot(df3[['Experience_Years','Age','Salary','Gender']],hue='Gender')

Out[45]: <seaborn.axisgrid.PairGrid at 0x2139dd0cfa0>
```

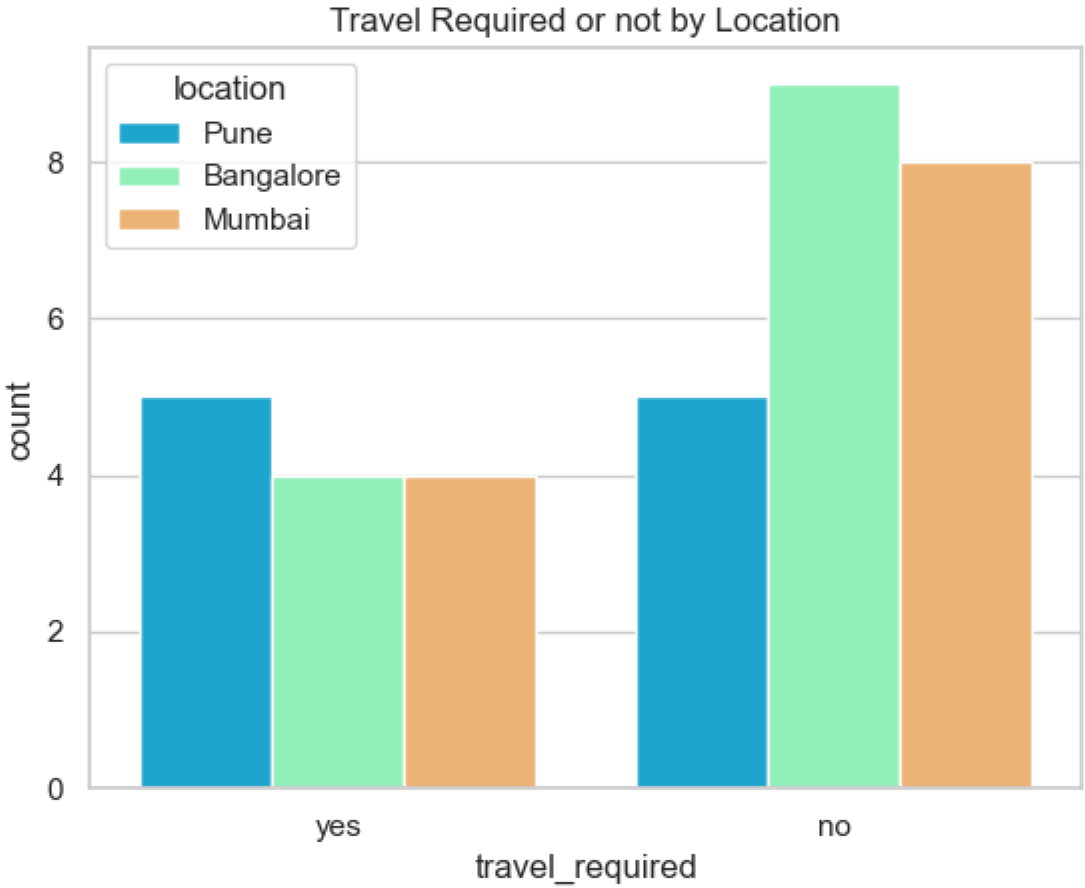


```
In [46]: sns.set(style="whitegrid")
         sns.violinplot(x="Age",y="Experience_Years",data=df3)
```

```
Out[46]: <AxesSubplot:xlabel='Age', ylabel='Experience_Years'>
```



In [47]:
```
sns.countplot(x='travel_required',hue='location',data=df3, palette='rainbow')
plt.title('Travel Required or not by Location')
plt.show()
```



# Conclusion

Findings according to the Exploratory Data Analysis:

1. Average Salary of Finance Department: 2165561.11
2. Average Salary of Tech Department: 2077257.14
3. Average Salary of HR Department: 1958208.33
4. Tech Department has the highest number of Employees (14)
5. There are 18 female employees and 16 male employees in the company. The gender distribution can be considered as more or less equal. Except for Mumbai, in other two locations, the number of females is higher than that of males.
6. The correlation between Years of Experience and Salary is 0.69. This means that the features are quite strongly correlated, and this relation can be used to predict salary.
7. The Finance Department has the highest average salary.
8. Age is highly correlated with Salary and Years of Experience.
9. Except in Pune, for the other two cities, more employees don't require travel. The city with highest number of employees who don't require travel is Bangalore.

# Thank You