In [1]:

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

In [2]:

```python
emp=pd.read_excel(r"D:\Full Stack Data Science\14 Aug\11th,14th\EDA- Pract
emp
```

Out[2]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#$	34 years	Mumbai	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^$0	10+

In [3]:

```python
emp.shape
```

Out[3]:

```
(6, 6)
```

In [4]:

```python
emp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Name      6 non-null      object
 1   Domain    6 non-null      object
 2   Age       4 non-null      object
 3   Location  4 non-null      object
 4   Salary    6 non-null      object
 5   Exp       5 non-null      object
dtypes: object(6)
memory usage: 420.0+ bytes
```

In [5]:

```python
emp.columns
```

Out[5]:

```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [6]:

```
emp[['Name','Domain','Age','Location','Salary','Exp']]
```

Out[6]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#$	34 years	Mumbai	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^$0	10+

In [7]:

```
emp['Name']
```

Out[7]:

```
0      Mike
1    Teddy^
2     Uma#r
3      Jane
4    Uttam*
5       Kim
Name: Name, dtype: object
```

In [8]:

```
emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
```

In [9]:

```
emp['Name']
```

Out[9]:

```
0     Mike
1    Teddy
2     Umar
3     Jane
4    Uttam
5      Kim
Name: Name, dtype: object
```

In [10]:

```
emp['Domain'] = emp['Domain'] .str.replace(r'\W','',regex=True)
```

In [11]:

```
emp['Domain']
```

Out[11]:

```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

In [12]:

```
emp['Age'] = emp['Age'] .str.replace(r'\W','',regex=True)
```

In [13]:

```
emp['Age']
```

Out[13]:

```
0    34years
1       45yr
2        NaN
3        NaN
4       67yr
5       55yr
Name: Age, dtype: object
```

In [14]:

```
emp['Age'] = emp['Age'] .str.extract('(\d+)')
```

In [15]:

```
emp
```

Out[15]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5^00#0	2+
1	Teddy	Testing	45	Bangalore	10%%000	<3
2	Umar	Dataanalyst	NaN	NaN	1$5%000	4> yrs
3	Jane	Analytics	NaN	Hyderbad	2000^0	NaN
4	Uttam	Statistics	67	NaN	30000-	5+ year
5	Kim	NLP	55	Delhi	6000^$0	10+

In [16]:

```
emp['Salary']=emp['Salary'].str.replace('\W','',regex=True)
emp
```

Out[16]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2+
1	Teddy	Testing	45	Bangalore	10000	<3
2	Umar	Dataanalyst	NaN	NaN	15000	4> yrs
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5+ year
5	Kim	NLP	55	Delhi	60000	10+

In [17]:

```
import re
```

In [18]:

```python
emp['Exp']=emp['Exp'].str.replace('\W','',regex=True)
emp['Exp']
```

Out[18]:

```
0        2
1        3
2     4yrs
3      NaN
4    5year
5       10
Name: Exp, dtype: object
```

In [19]:

```python
emp['Exp']=emp['Exp'].str.extract('(\d+)')
emp['Exp']

```

Out[19]:

```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

In [20]:

```python
emp
```

Out[20]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	NaN	NaN	15000	4
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [21]:

```python
clean=emp.copy()
```

Missing Value Treatment

In [22]:

```
clean
```

Out[22]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	NaN	NaN	15000	4
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [23]:

```
clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Name      6 non-null      object
 1   Domain    6 non-null      object
 2   Age       4 non-null      object
 3   Location  4 non-null      object
 4   Salary    6 non-null      object
 5   Exp       5 non-null      object
dtypes: object(6)
memory usage: 420.0+ bytes
```

In [24]:

```
clean.isnull().sum()
```

Out[24]:

```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [25]:

```
clean['Age']
```

Out[25]:

```
0     34
1     45
2    NaN
3    NaN
4     67
5     55
Name: Age, dtype: object
```

In [26]:

```
clean['Age']=clean['Age'].fillna(np.mean(pd.to_numeric(clean['Age'])))
```

In [27]:

```
clean['Age']
```

Out[27]:

```
0       34
1       45
2    50.25
3    50.25
4       67
5       55
Name: Age, dtype: object
```

In [28]:

```
clean
```

Out[28]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	NaN	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [29]:

```
clean['Location']
```

Out[29]:

```
0       Mumbai
1    Bangalore
2          NaN
3     Hyderbad
4          NaN
5        Delhi
Name: Location, dtype: object
```

In [30]:

```
clean['Location']=clean['Location'].fillna(clean['Location'].mode()[0])
clean['Location']
```

Out[30]:

```
0       Mumbai
1    Bangalore
2    Bangalore
3     Hyderbad
4    Bangalore
5        Delhi
Name: Location, dtype: object
```

In [31]:

```
clean
```

Out[31]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	Bangalore	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	NaN
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [32]:

```
clean['Exp']
```

Out[32]:

```
0       2
1       3
2       4
3     NaN
4       5
5      10
Name: Exp, dtype: object
```

In [33]:

```
clean['Exp']=clean['Exp'].fillna(np.mean(pd.to_numeric(clean['Exp'])))
```

In [34]:

```
clean['Exp']
```

Out[34]:

```
0       2
1       3
2       4
3     4.8
4       5
5      10
Name: Exp, dtype: object
```

In [35]:

```
clean
```

Out[35]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	Bangalore	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [36]:

```
clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Name      6 non-null      object
 1   Domain    6 non-null      object
 2   Age       6 non-null      object
 3   Location  6 non-null      object
 4   Salary    6 non-null      object
 5   Exp       6 non-null      object
dtypes: object(6)
memory usage: 420.0+ bytes
```

In [37]:

```
clean['Age']=clean['Age'].astype(int)
clean['Salary']=clean['Salary'].astype(int)
clean['Exp']=clean['Exp'].astype(int)
```

In [38]:

```
clean.info()
```

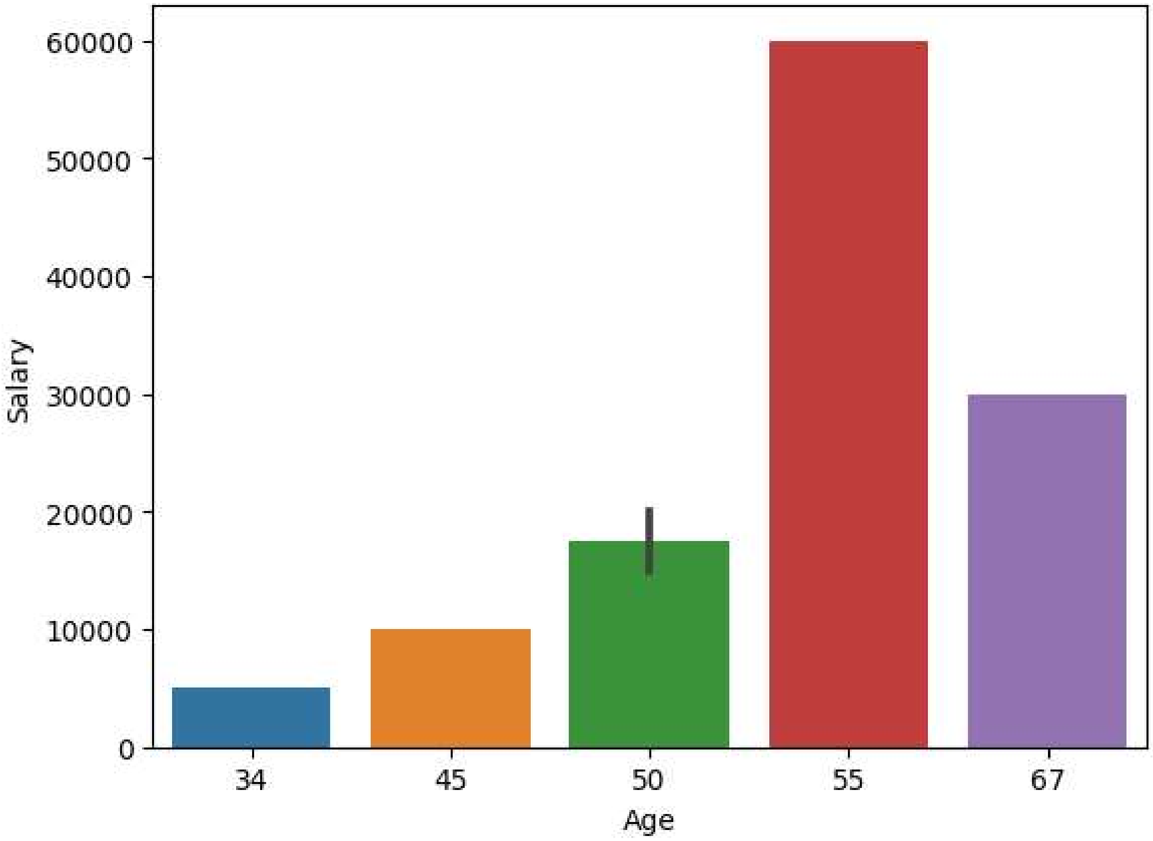
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Name      6 non-null      object
 1   Domain    6 non-null      object
 2   Age       6 non-null      int32 
 3   Location  6 non-null      object
 4   Salary    6 non-null      int32 
 5   Exp       6 non-null      int32 
dtypes: int32(3), object(3)
memory usage: 348.0+ bytes
```

In [39]:

```
sns.barplot(data=clean,x='Age',y='Salary')
```

Out[39]:

```
<Axes: xlabel='Age', ylabel='Salary'>
```

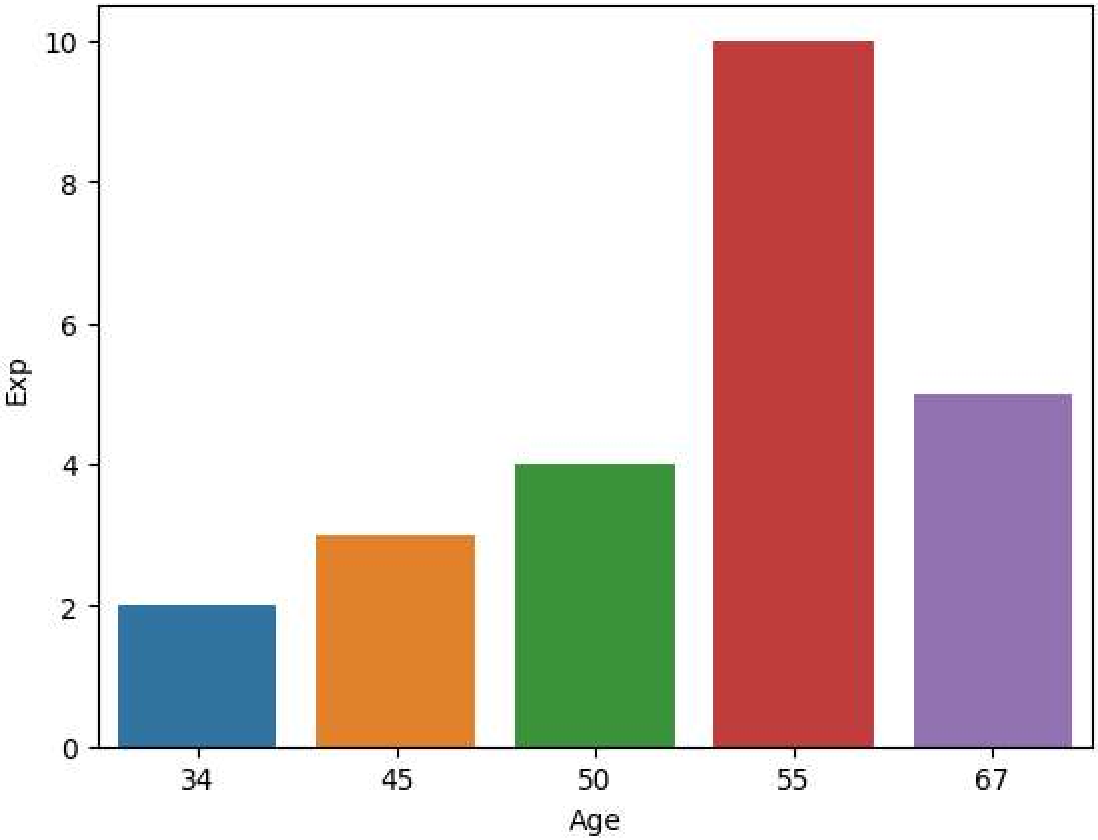


In [40]:

```
sns.barplot(data=clean,x='Age',y='Exp')
```

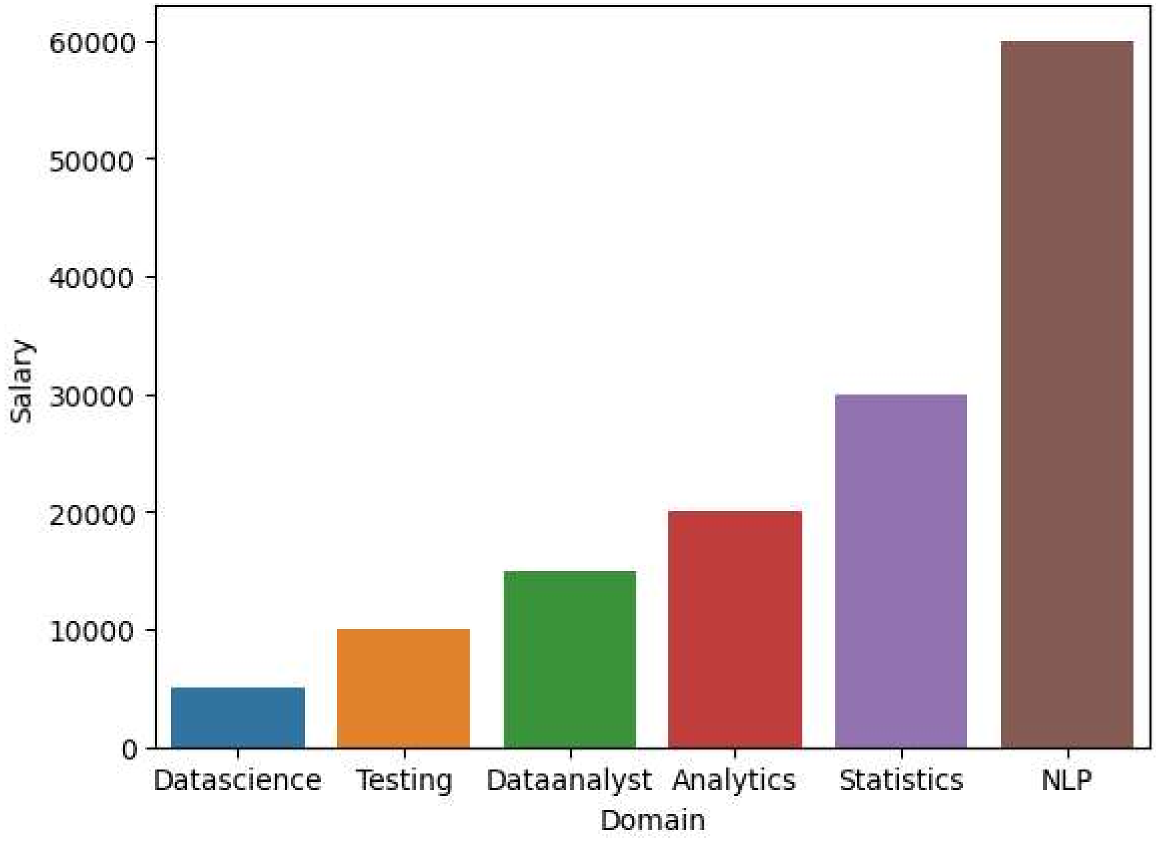
Out[40]:

```
<Axes: xlabel='Age', ylabel='Exp'>
```



In [41]:
```
sns.barplot(data=clean,x='Domain',y='Salary')
```

Out[41]:
```
<Axes: xlabel='Domain', ylabel='Salary'>
```

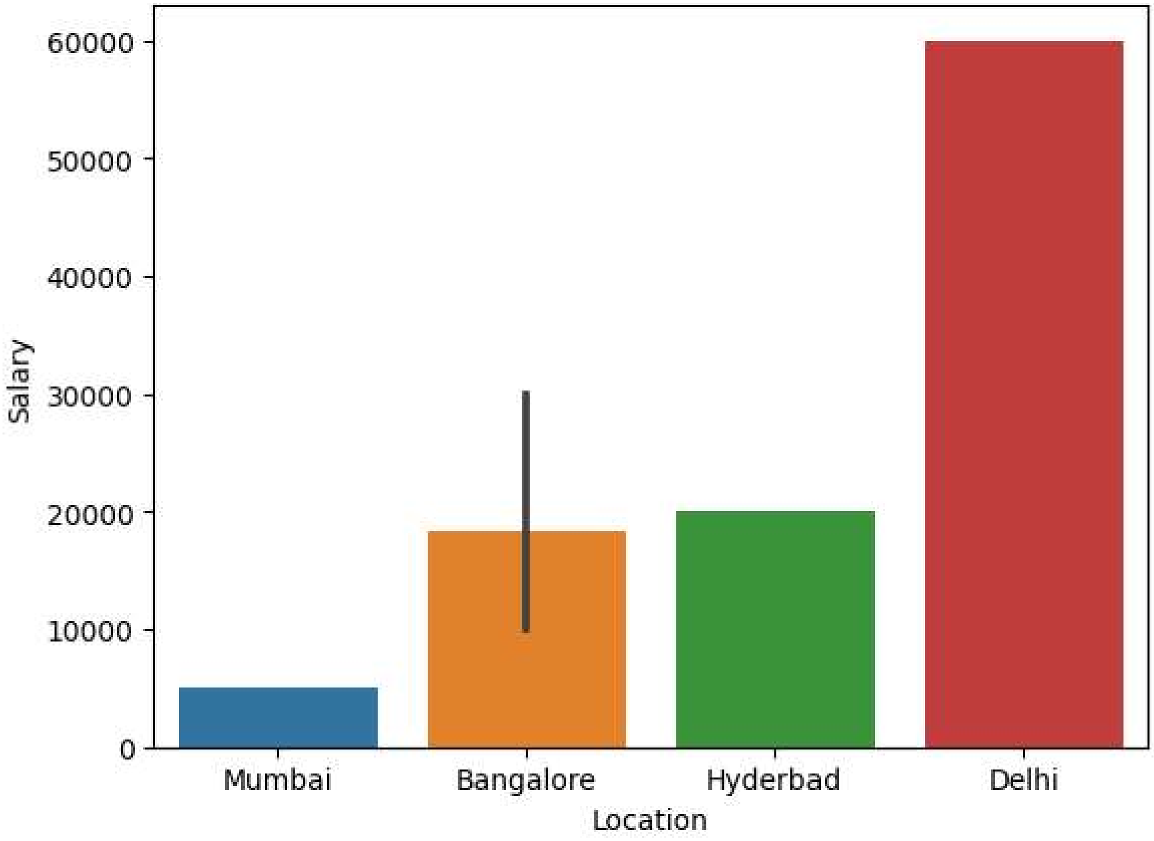


In [42]:

```python
sns.barplot(data=clean,x='Location',y='Salary')
```

Out[42]:

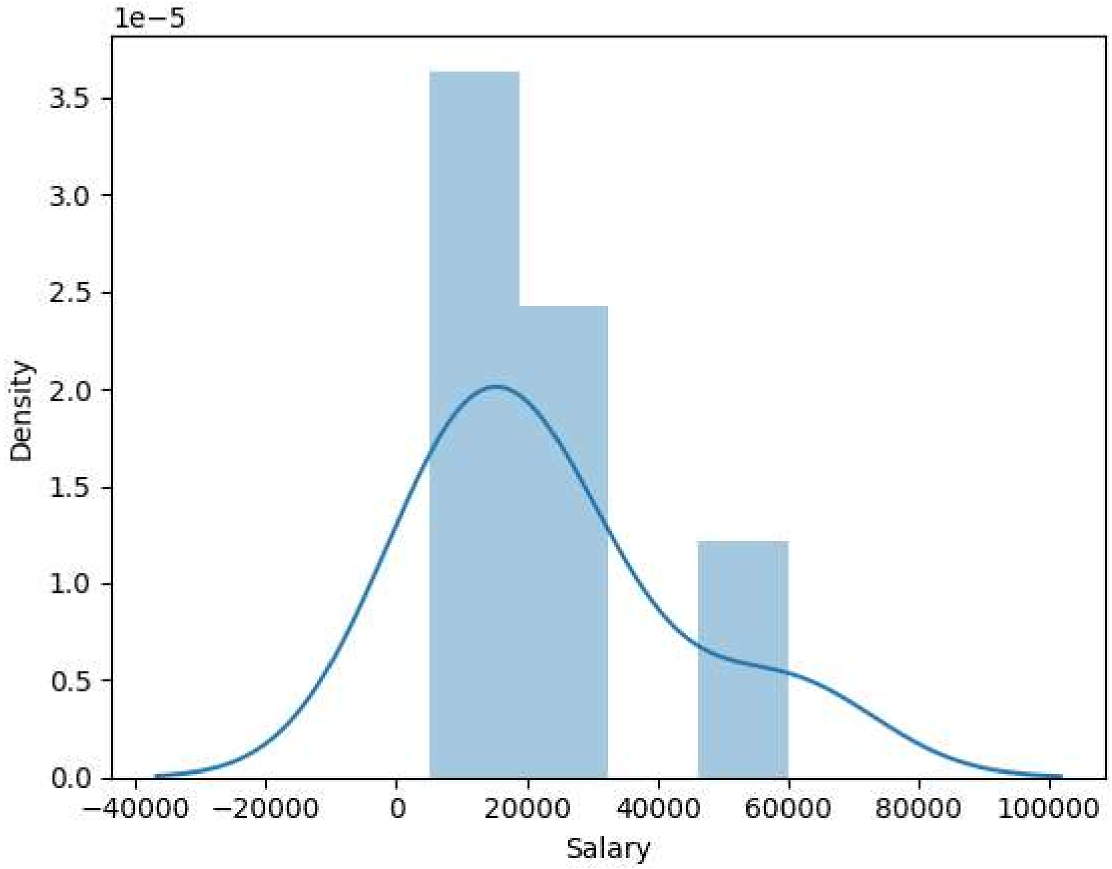
```
<Axes: xlabel='Location', ylabel='Salary'>
```



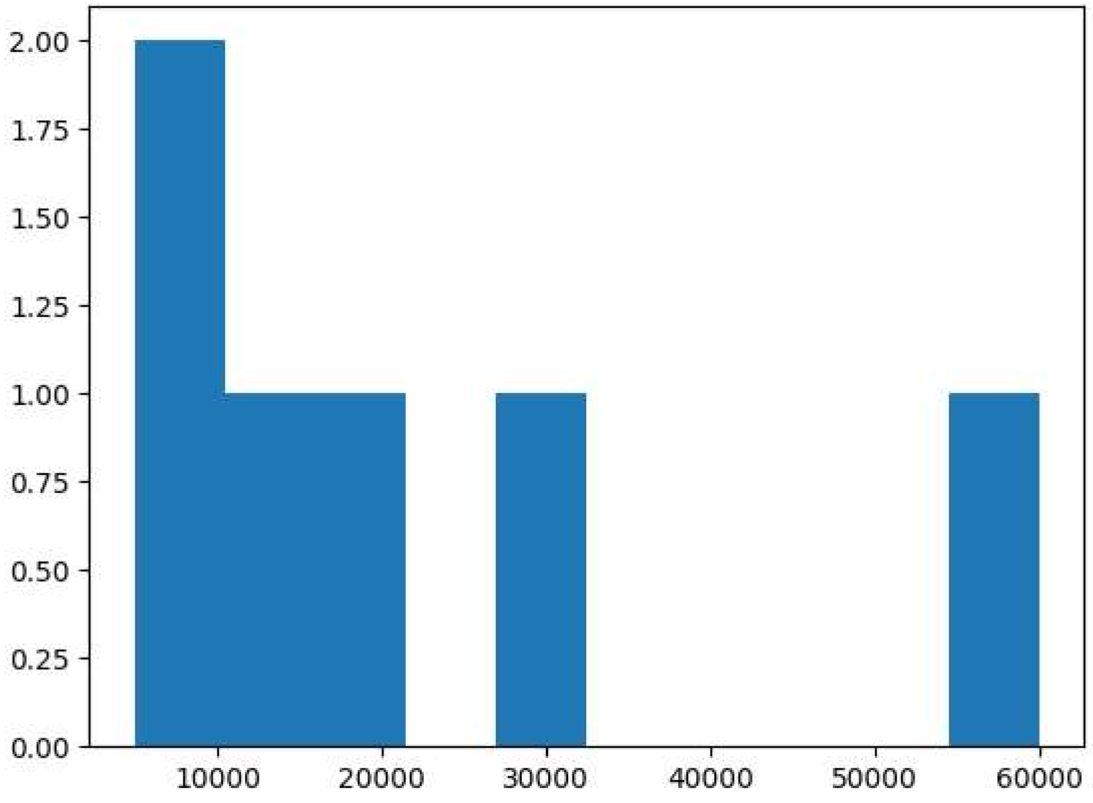
In [43]:

```python
import warnings
warnings.filterwarnings('ignore')
```

```
In [44]: vis1=sns.distplot(clean['Salary'])
```



```
In [45]: vis2=plt.hist(clean['Salary'])
```



In [46]:

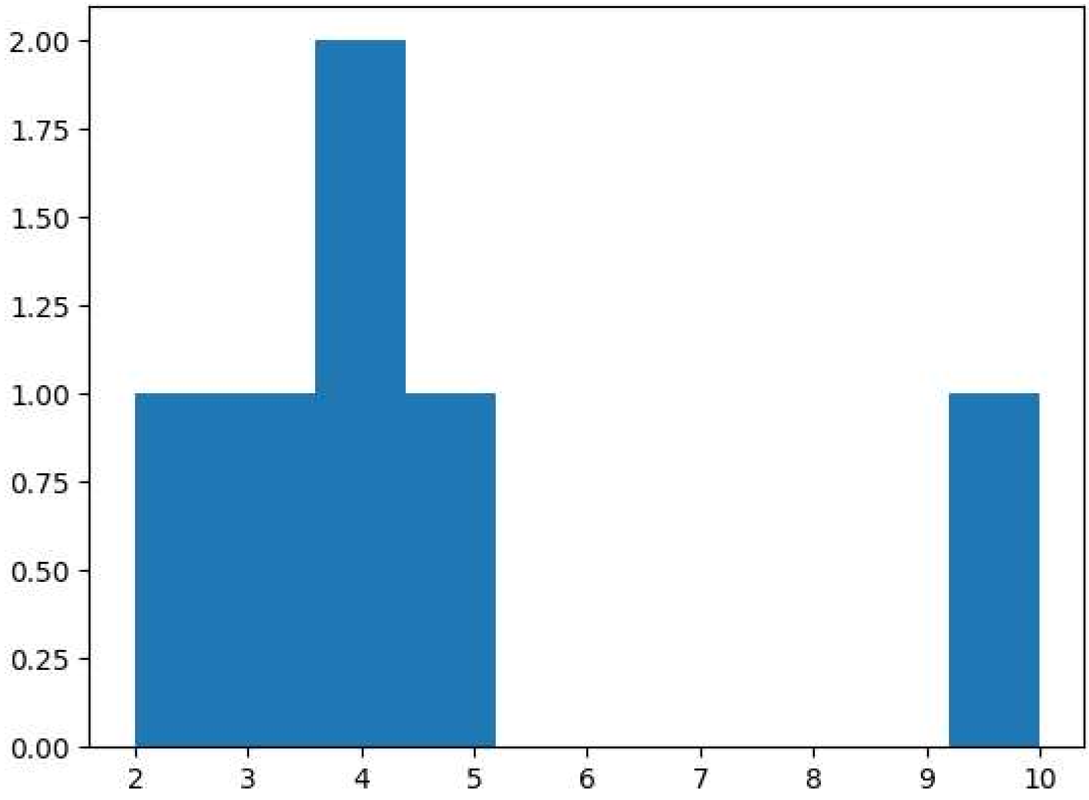
```
clean['Exp']
```

Out[46]:

```
0     2
1     3
2     4
3     4
4     5
5    10
Name: Exp, dtype: int32
```

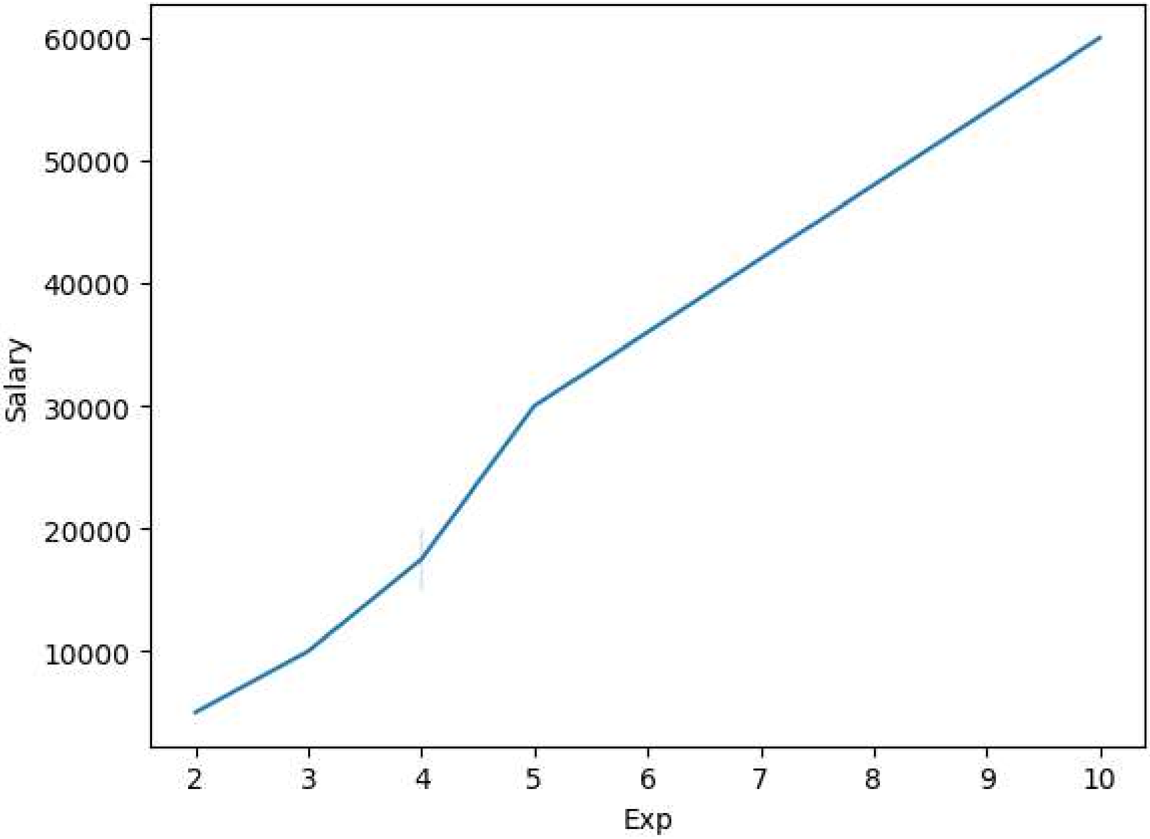
In [47]:

```
vis3=plt.hist(clean['Exp'])
```



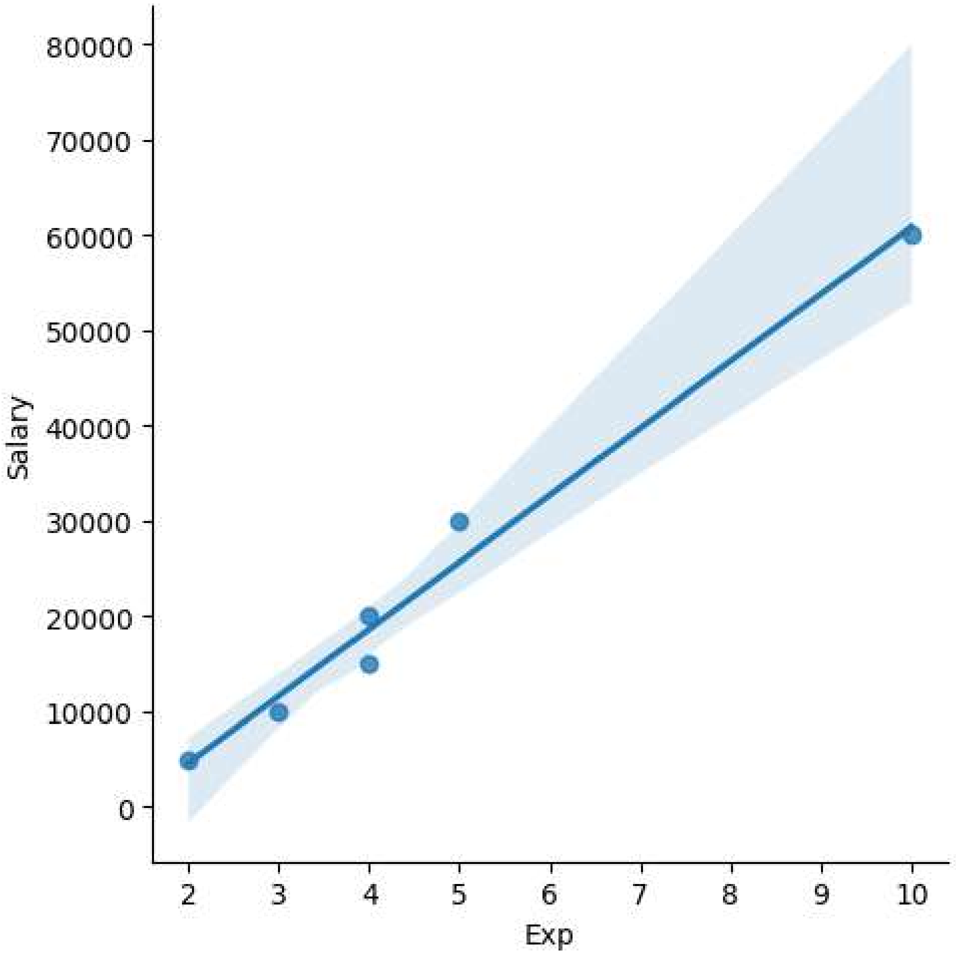
In [48]:

```
vis4=sns.lineplot(data=clean,x='Exp',y='Salary')
```



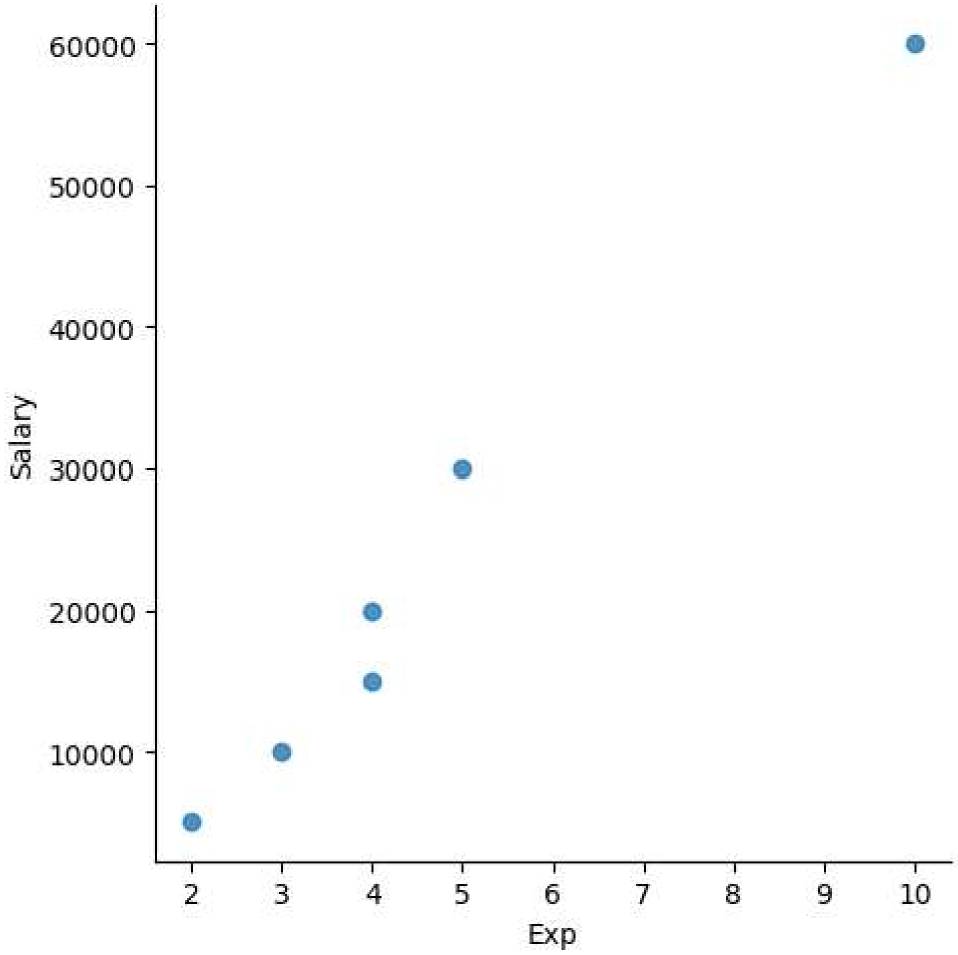
In [49]:

```python
vis4=sns.lmplot(data=clean,x='Exp',y='Salary')
```



In [50]:

```python
vis5=sns.lmplot(data=clean,x='Exp',y='Salary',fit_reg=False)
```



In [51]:

```python
clean[:]
```

Out[51]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [52]:

```
x_iv=clean.drop(['Salary'],axis=1)
x_iv
```

Out[52]:

	Name	Domain	Age	Location	Exp
0	Mike	Datascience	34	Mumbai	2
1	Teddy	Testing	45	Bangalore	3
2	Umar	Dataanalyst	50	Bangalore	4
3	Jane	Analytics	50	Hyderbad	4
4	Uttam	Statistics	67	Bangalore	5
5	Kim	NLP	55	Delhi	10

In [53]:

```
clean
```

Out[53]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [54]:

```
y_dv = clean.drop(['Name', 'Domain', 'Age', 'Location','Exp'],axis=1)
y_dv
```

Out[54]:

	Salary
0	5000
1	10000
2	15000
3	20000
4	30000
5	60000

In [55]:

```python
x_iv
```

Out[55]:

	Name	Domain	Age	Location	Exp
0	Mike	Datascience	34	Mumbai	2
1	Teddy	Testing	45	Bangalore	3
2	Umar	Dataanalyst	50	Bangalore	4
3	Jane	Analytics	50	Hyderbad	4
4	Uttam	Statistics	67	Bangalore	5
5	Kim	NLP	55	Delhi	10

In [56]:

```python
clean
```

Out[56]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [57]:

```python
imputation=pd.get_dummies(clean).astype(int)
imputation
```

Out[57]:

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Utt
0	34	5000	2	0	0	1	0	0	
1	45	10000	3	0	0	0	1	0	
2	50	15000	4	0	0	0	0	1	
3	50	20000	4	1	0	0	0	0	
4	67	30000	5	0	0	0	0	0	
5	55	60000	10	0	1	0	0	0	

In [ ]:

```python

```