In [1]:
```
import pandas as pd
```

In [2]:
```
pd.__version__
```

Out[2]:
```
'2.2.3'
```

In [3]:
```
rd=pd.read_excel(r'C:\Users\nlnar\Downloads\Rawdata.xlsx')
```

In [4]:
```
rd
```

Out[4]:

	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 [5]:
```
rd.isnull().sum()
```

Out[5]:
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [6]:
```
id(rd)   # address of yhe memory location
```

Out[6]:
```
1603869464560
```

In [7]:
```
rd.columns
```

Out[7]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [8]:
```
rd.shape
```

Out[8]:
```
(6, 6)
```

In [9]:
```
rd.head()
```

Out[9]:

	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

In [10]:
```
rd.tail()
```

Out[10]:

	Name	Domain	Age	Location	Salary	Exp
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 [11]:
```
rd.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 [12]:
```
rd.isnull()
```

Out[12]:

	Name	Domain	Age	Location	Salary	Exp
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	True	False	False
3	False	False	True	False	False	True
4	False	False	False	True	False	False
5	False	False	False	False	False	False

In [13]:
```
rd.isnull().sum()
```

Out[13]:
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [14]:
```
rd.isna()
```

Out[14]:

	Name	Domain	Age	Location	Salary	Exp
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	True	False	False
3	False	False	True	False	False	True
4	False	False	False	True	False	False
5	False	False	False	False	False	False

data cleaning or cleansing

In [15]:
```
rd
```

Out[15]:

	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 [16]:
```
rd['Name']
```

Out[16]:
```
0      Mike
1    Teddy^
2     Uma#r
3      Jane
4    Uttam*
5       Kim
Name: Name, dtype: object
```

In [17]:
```
rd['Name']=rd['Name'].str.replace(r'\W','',regex=True)
```

In [18]:
```
rd['Name']
```

Out[18]:
```
0     Mike
1    Teddy
2     Umar
3     Jane
4    Uttam
5      Kim
Name: Name, dtype: object
```

In [19]:
```
rd['Domain']=rd['Domain'].str.replace(r'\W','',regex=True)
```

In [20]:
```
rd['Domain']
```

Out[20]:
```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

In [21]:
```
rd
```

Out[21]:

	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	Umar	Dataanalyst	NaN	NaN	1$5%000	4> yrs
3	Jane	Analytics	NaN	Hyderbad	2000^0	NaN
4	Uttam	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^$0	10+

In [22]:
```
rd['Location']=rd['Location'].str.replace(r'\W','',regex=True)
```

In [23]:
```
rd['Location']
```

Out[23]:
```
0       Mumbai
1    Bangalore
2          NaN
3     Hyderbad
4          NaN
5        Delhi
Name: Location, dtype: object
```

In [24]:
```
rd
```

Out[24]:

	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	Umar	Dataanalyst	NaN	NaN	1$5%000	4> yrs
3	Jane	Analytics	NaN	Hyderbad	2000^0	NaN
4	Uttam	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^$0	10+

In [25]:
```
rd['Age']
```

Out[25]:
```
0    34 years
1      45' yr
2         NaN
3         NaN
4       67-yr
5        55yr
Name: Age, dtype: object
```

In [26]:
```
rd['Age']=rd['Age'].str.extract('(\\d+)')
```

In [27]:
```
rd['Age']
```

Out[27]:
```
0     34
1     45
2    NaN
3    NaN
4     67
5     55
Name: Age, dtype: object
```

In [28]:
```
rd['Salary']=rd['Salary'].str.replace(r'\W','',regex=True)
```

In [29]:
```
rd['Salary']
```

Out[29]:
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: object
```

In [30]:
```
rd['Exp']=rd['Exp'].str.extract('(\\d+)')
```

In [31]:
```
rd['Exp']
```

Out[31]:
```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

In [32]:
```
rd
```

Out[32]:

	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 [33]:
```
clean_data=rd.copy()
```

workshop day2

In [34]:
```
clean_data.isnull().sum()
```

Out[34]:
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [35]:
```
import numpy as np
```

In [36]:
```
clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])
```

In [37]:
```
clean_data['Age']
```

Out[37]:
```
0       34
1       45
2    50.25
3    50.25
4       67
5       55
Name: Age, dtype: object
```

In [38]:
```
clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp'])
```

In [39]:
```
clean_data['Exp']
```

Out[39]:
```
0      2
1      3
2      4
3    4.8
4      5
5     10
Name: Exp, dtype: object
```

In [40]:
```
clean_data['Location']=clean_data['Location'].fillna(clean_data['Location'].mode()[
```

In [41]:
```
clean_data['Location']
```

Out[41]:
```
0       Mumbai
1    Bangalore
2    Bangalore
3     Hyderbad
4    Bangalore
5        Delhi
Name: Location, dtype: object
```

In [42]:
```
clean_data
```

Out[42]:

	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 [43]:
```
clean_data.to_csv('clean_data.csv')
```

In [44]:
```
import os
os.getcwd()
```

Out[44]:
```
'C:\\Users\\nlnar'
```

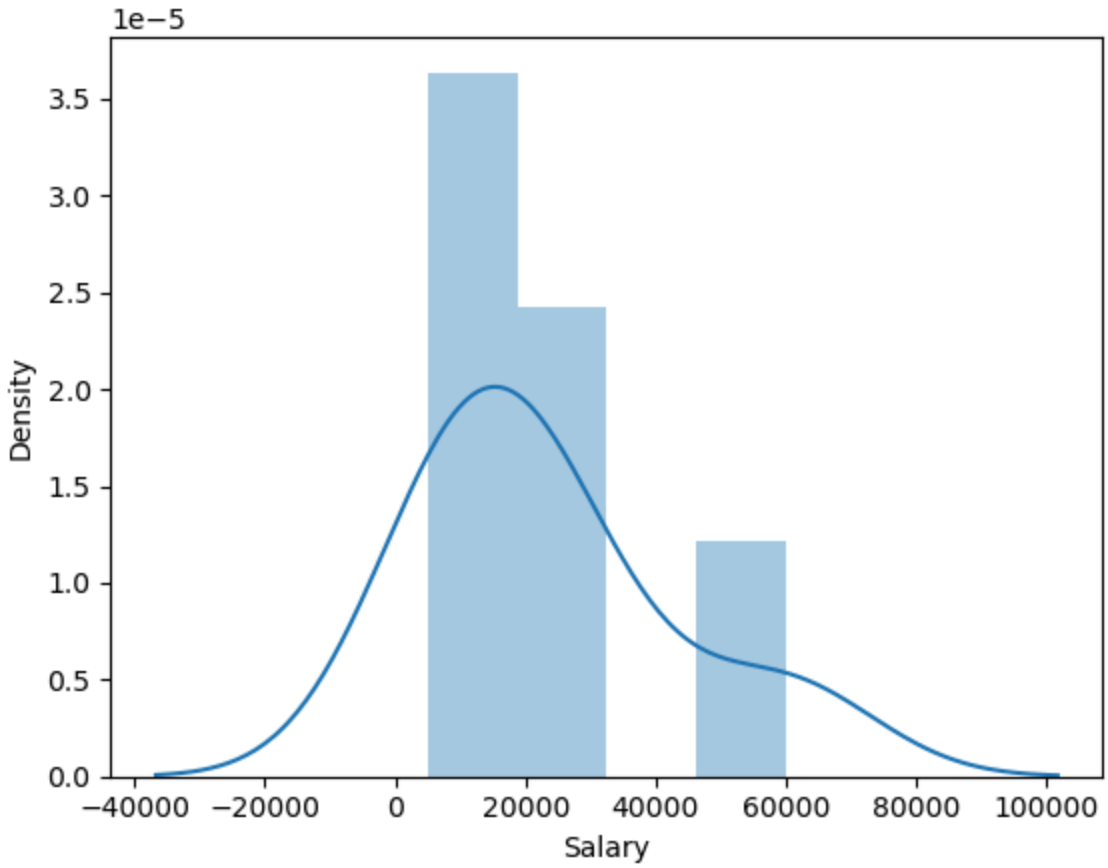
In [45]:
```
import matplotlib.pyplot as plt
```

In [46]:
```
import seaborn as sns
```

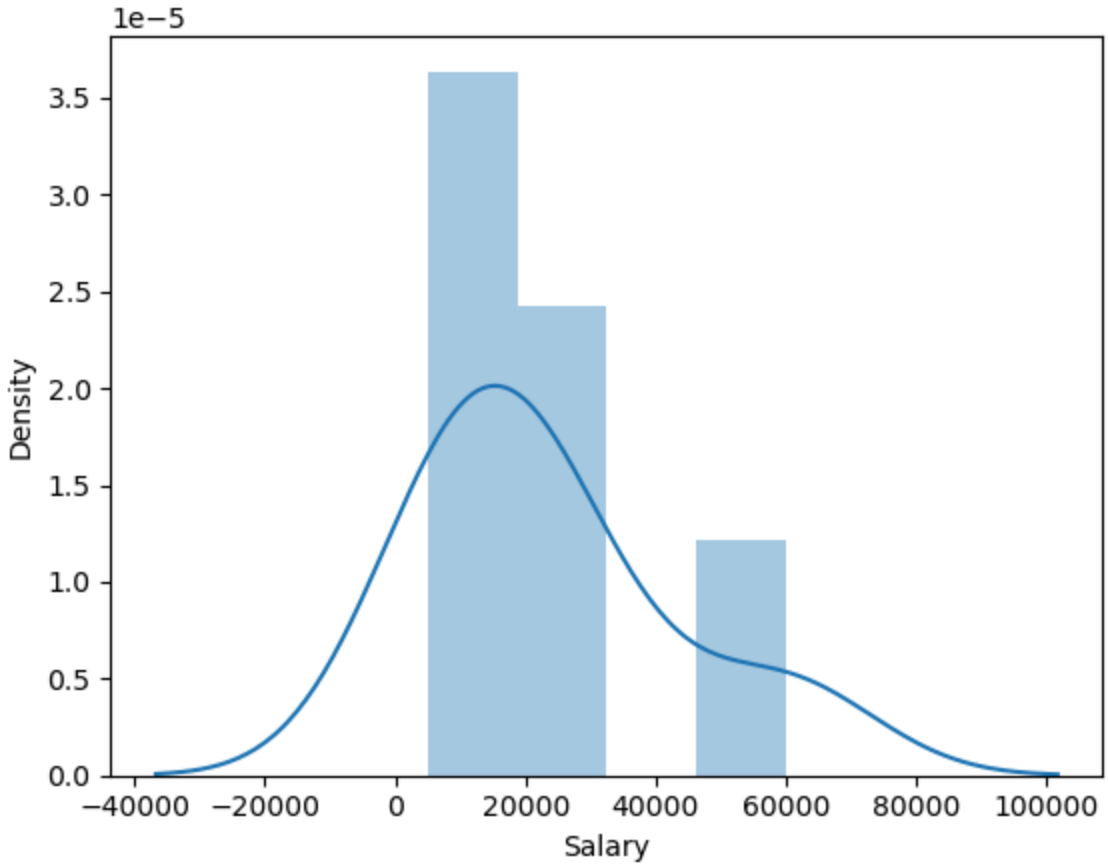
In [47]:
```
import warnings
warnings.filterwarnings('ignore')
```

vis1

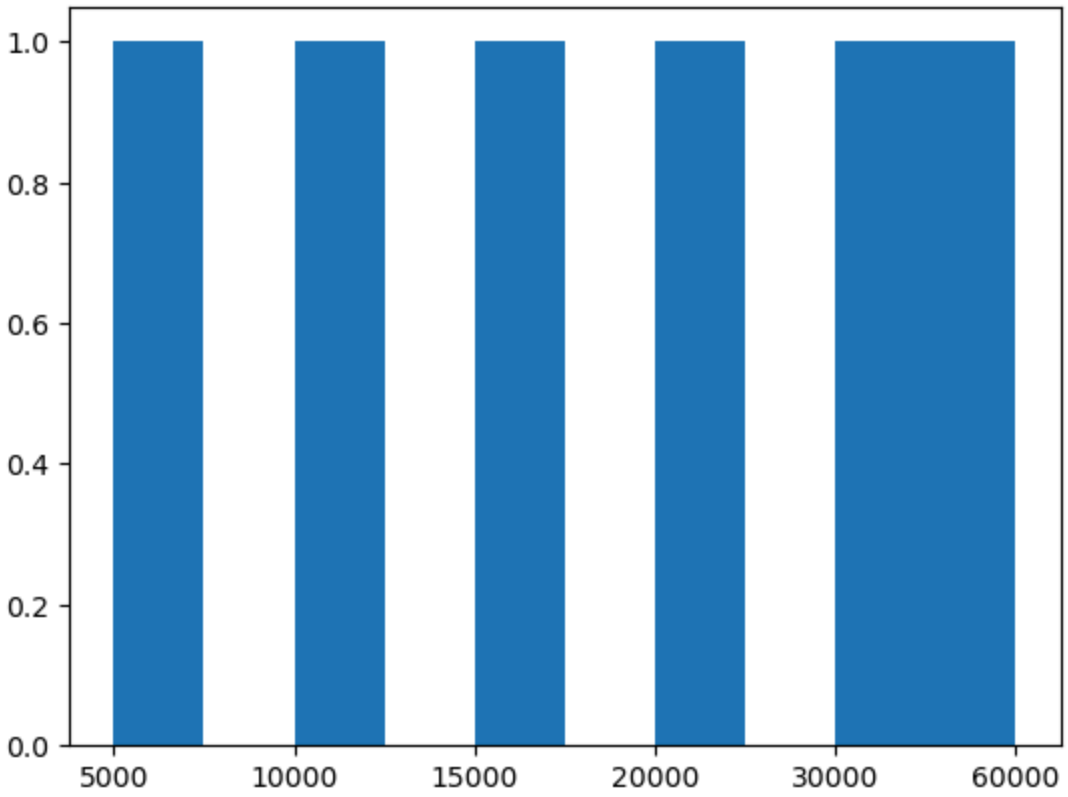
In [48]:
```
vis1=sns.distplot(clean_data['Salary'])
```



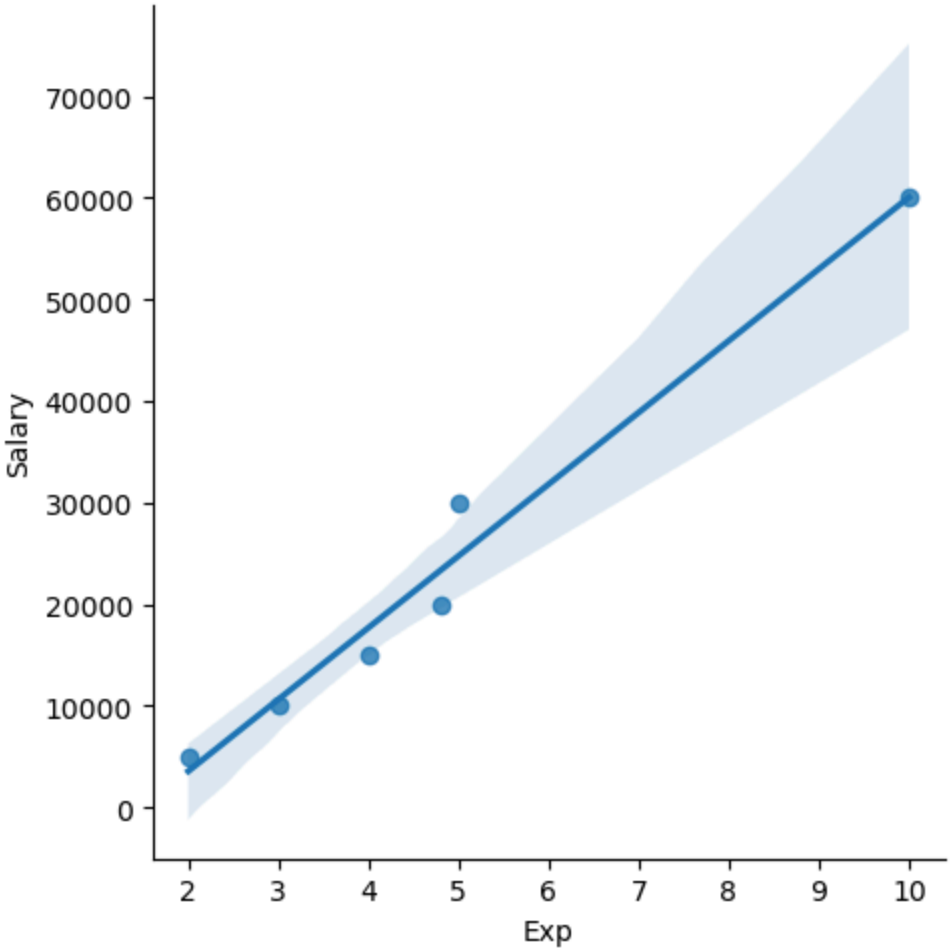
In [49]:
```
vis1=sns.distplot(clean_data['Salary'])
```



In [51]:
```
vis2=plt.hist(clean_data['Salary'])
```



In [74]:
```
vis3=sns.lmplot(data=clean_data,x='Exp',y='Salary')
```



In [62]:
```
clean_data.columns
```

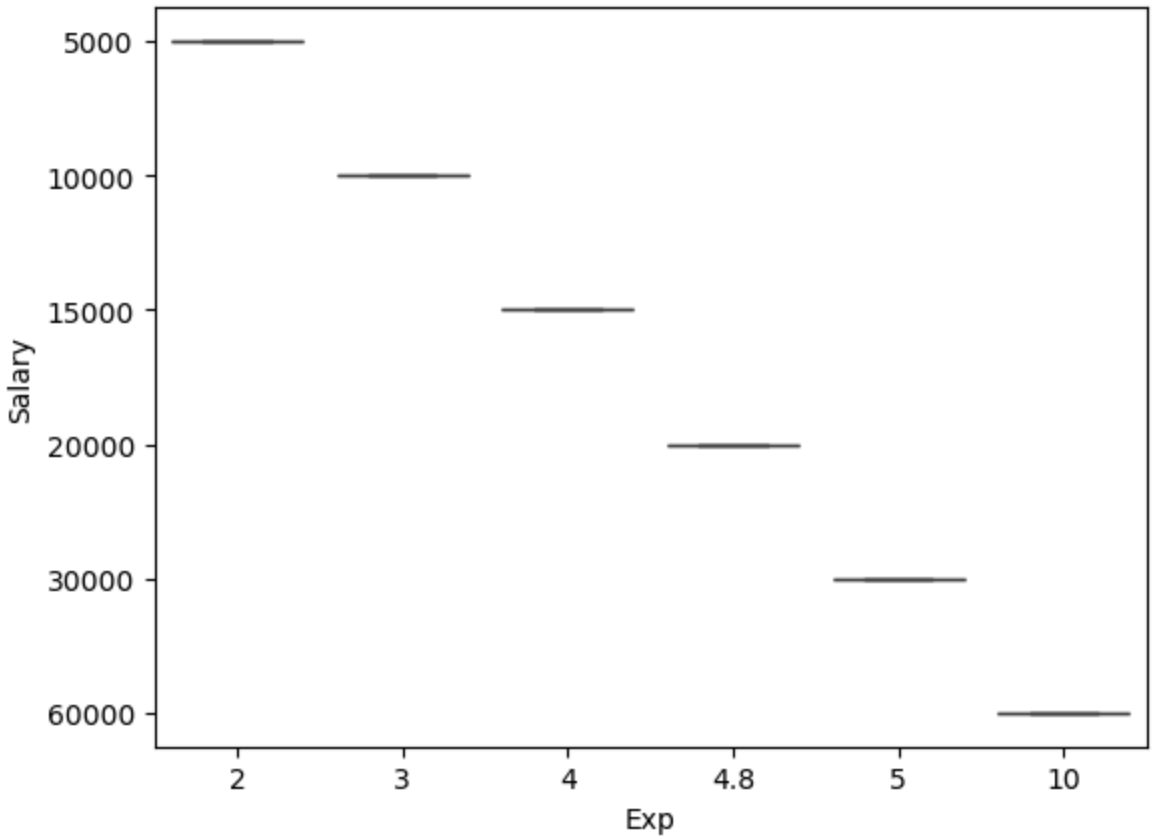
Out[62]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [73]:
```
clean_data.dtypes
```

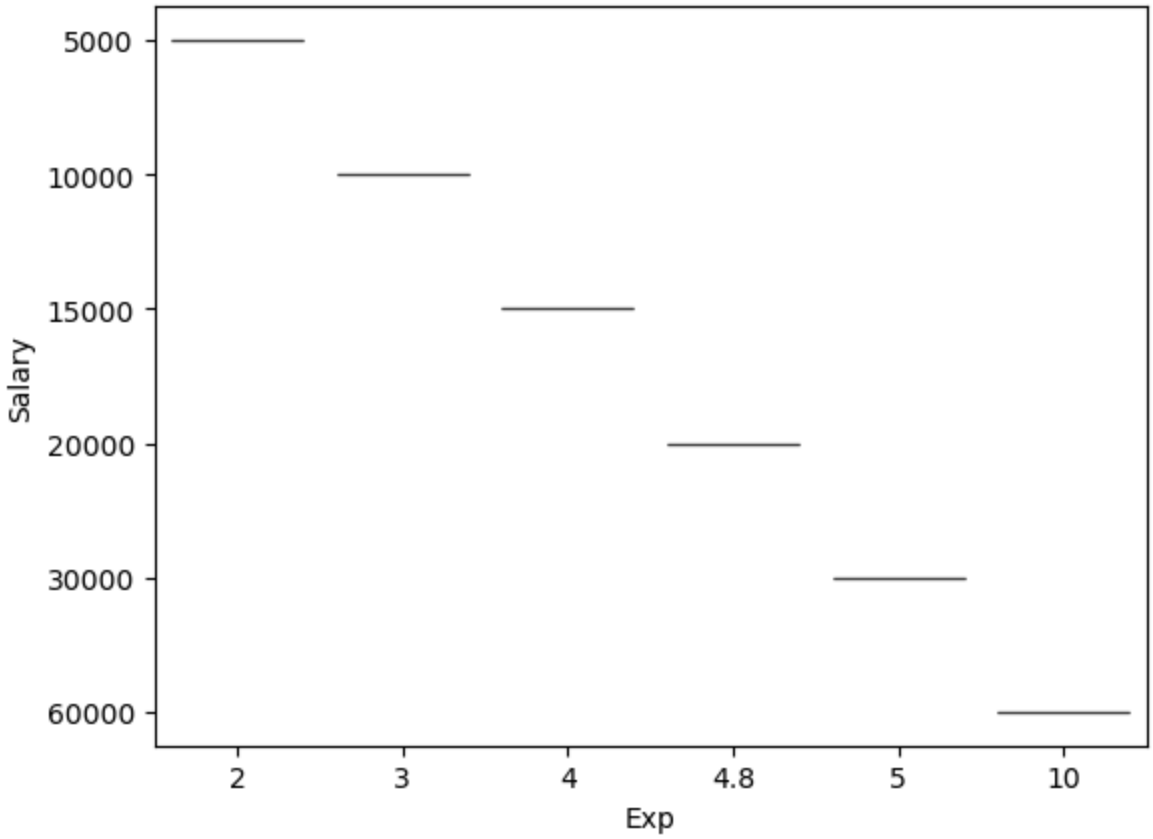
Out[73]:
```
Name         object
Domain       object
Age          object
Location     object
Salary        int64
Exp         float64
dtype: object
```

In [72]:
```
clean_data['Salary']=pd.to_numeric(clean_data['Salary'],errors='coerce')
clean_data['Exp']=pd.to_numeric(clean_data['Exp'],errors='coerce')
```

In [56]:
```
vis3=sns.boxplot(data=clean_data,x='Exp',y='Salary')
```



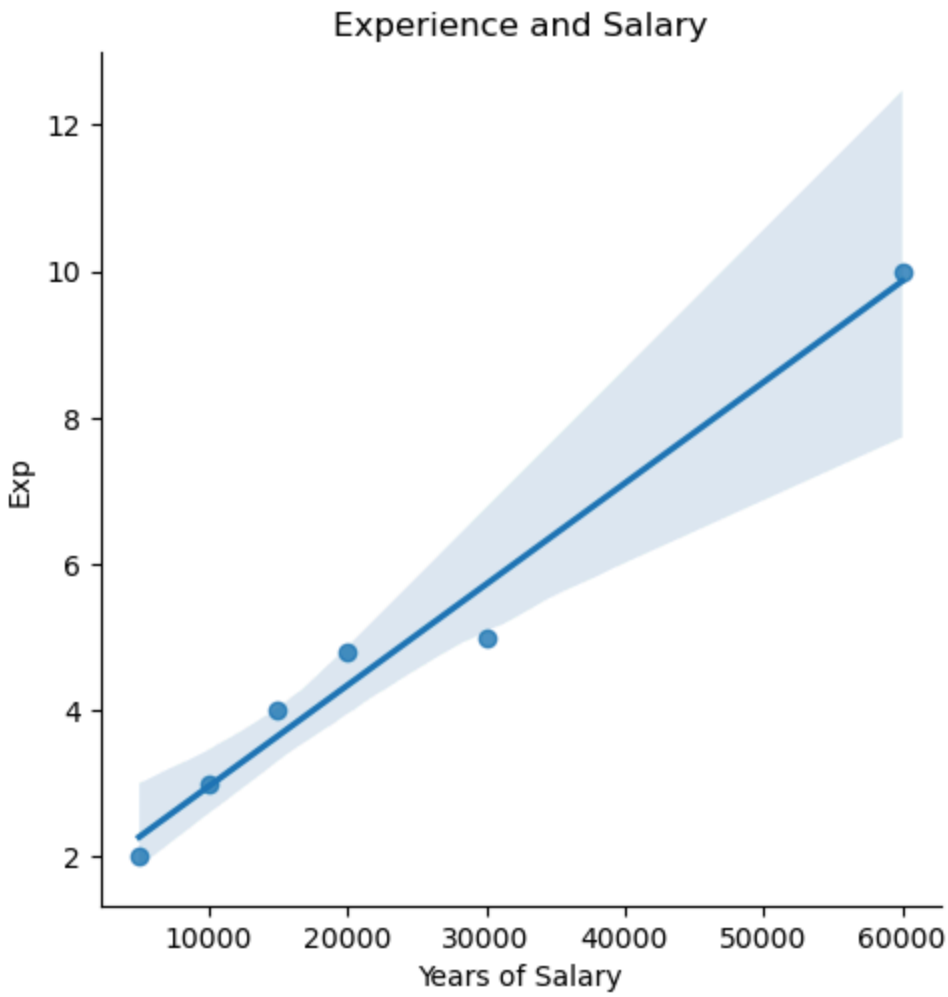
In [57]:
```
vis3=sns.boxenplot(data=clean_data,x='Exp',y='Salary')
```



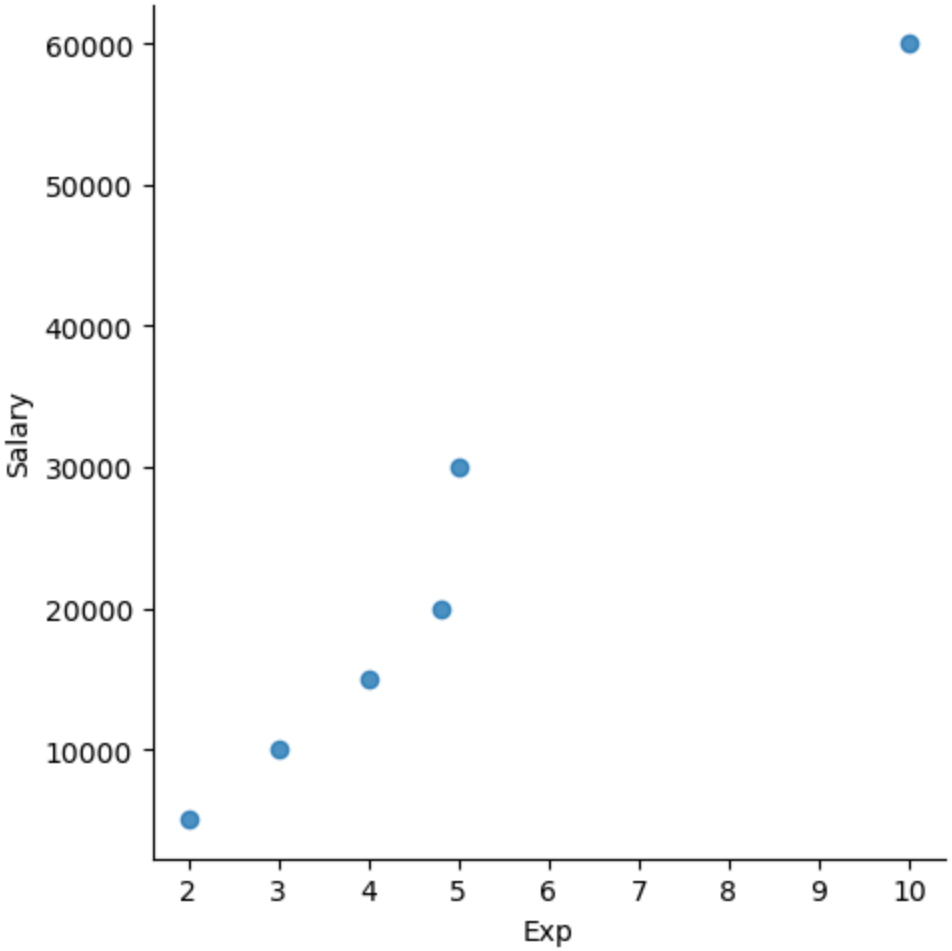
In [ ]:
```

```

In [75]:
```
vis3=sns.lmplot(data=clean_data,x='Salary',y='Exp')
plt.title('Experience and Salary')
plt.xlabel('Years of Salary')
plt.ylabel('Exp')
plt.show()
```



In [76]:
```
vis3=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=False)
```



In [ ]:

In [ ]:
```
import seaborn as sns
```

In [77]:
```
clean_data
```

Out[77]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2.0
1	Teddy	Testing	45	Bangalore	10000	3.0
2	Umar	Dataanalyst	50.25	Bangalore	15000	4.0
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	Bangalore	30000	5.0
5	Kim	NLP	55	Delhi	60000	10.0

In [78]:
```
y=clean_data['Salary']   # dependant var
```

In [79]:
```
y
```

Out[79]:
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int64
```

In [80]:
```
clean_data.columns
```

Out[80]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [81]:
```
x_iv=clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]   # independant var
```

In [82]:
```
x_iv
```

Out[82]:

	Name	Domain	Age	Location	Exp
0	Mike	Datascience	34	Mumbai	2.0
1	Teddy	Testing	45	Bangalore	3.0
2	Umar	Dataanalyst	50.25	Bangalore	4.0
3	Jane	Analytics	50.25	Hyderbad	4.8
4	Uttam	Statistics	67	Bangalore	5.0
5	Kim	NLP	55	Delhi	10.0

In [83]:
```
imputation=pd.get_dummies(clean_data,dtype=int)   # data which ml understands
```

In [84]:
```
imputation
```

Out[84]:

	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Utt
0	5000	2.0	0	0	1	0	0	
1	10000	3.0	0	0	0	1	0	
2	15000	4.0	0	0	0	0	1	
3	20000	4.8	1	0	0	0	0	
4	30000	5.0	0	0	0	0	0	
5	60000	10.0	0	1	0	0	0	

6 rows × 23 columns

In [85]:
```
imputation.columns
```

Out[85]:
```
Index(['Salary', 'Exp', 'Name_Jane', 'Name_Kim', 'Name_Mike', 'Name_Teddy',
       'Name_Umar', 'Name_Uttam', 'Domain_Analytics', 'Domain_Dataanalyst',
       'Domain_Datascience', 'Domain_NLP', 'Domain_Statistics',
       'Domain_Testing', 'Age_50.25', 'Age_34', 'Age_45', 'Age_55', 'Age_67',
       'Location_Bangalore', 'Location_Delhi', 'Location_Hyderbad',
       'Location_Mumbai'],
      dtype='object')
```

In [86]:
```
len(imputation.columns)
```

Out[86]:
```
23
```

In [87]:
```
clean_data[:]
```

Out[87]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2.0
1	Teddy	Testing	45	Bangalore	10000	3.0
2	Umar	Dataanalyst	50.25	Bangalore	15000	4.0
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	Bangalore	30000	5.0
5	Kim	NLP	55	Delhi	60000	10.0

In [88]:
```
clean_data[0:6:2]
```

Out[88]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2.0
2	Umar	Dataanalyst	50.25	Bangalore	15000	4.0
4	Uttam	Statistics	67	Bangalore	30000	5.0

In [89]:
```
clean_data[::-1]
```

Out[89]:

	Name	Domain	Age	Location	Salary	Exp
5	Kim	NLP	55	Delhi	60000	10.0
4	Uttam	Statistics	67	Bangalore	30000	5.0
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
2	Umar	Dataanalyst	50.25	Bangalore	15000	4.0
1	Teddy	Testing	45	Bangalore	10000	3.0
0	Mike	Datascience	34	Mumbai	5000	2.0

In [ ]:
```

```