In [82]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [83]:
```
%matplotlib inline

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

In [84]:
```
df = pd.read_excel('Rawdata.xlsx')
```

In [85]:
```
df
```

Out[85]:

	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 [86]:
```
df.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: 416.0+ bytes
```

In [87]:
```
df['Name'] = df['Name'].str.replace(r'\W','')
```

In [88]:
```
df
```

Out[88]:

	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	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 [89]:
```
df['Domain'] = df['Domain'].str.replace(r'\W','')
```

In [90]:
```
df
```

Out[90]:

	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 [91]:
```
df['Location'] = df['Location'].str.replace(r'\W','')
```

In [92]:
```
df
```

Out[92]:

	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 [93]:
```
df['Salary']=df['Salary'].str.replace(r'\W','')
```

In [94]:
```
df
```

Out[94]:

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

In [95]:
```
df['Age'] = df['Age'].str.replace(r'\W','')
```

In [96]:
```
df
```

Out[96]:

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

In [97]:
```
df['Age'] = df['Age'].str.extract('(\d+)')
```

In [98]:
```
df
```

Out[98]:

	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 [99]:
```
df['Exp'] = df['Exp'].str.extract('(\d+)')
```

In [100]:
```
df
```

Out[100]:

	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 [101]:
```
df.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: 416.0+ bytes
```

In [102]:
```
clean_df = df.copy()
clean_df
```

Out[102]:

	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 [103]:
```
#clean_df['Age'] = clean_df['Age'].astype('int')
clean_df['Salary'] = clean_df['Salary'].astype('int')
#clean_df['Exp'] = clean_df['Exp'].astype('int')
```

In [104]:
```
clean_df.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      int32 
 5   Exp       5 non-null      object
dtypes: int32(1), object(5)
memory usage: 392.0+ bytes
```

In [105]:
```
clean_df['Age'] = clean_df['Age'].fillna(np.mean(pd.to_numeric(clean_df['Age'])))
```

In [106]:
```
clean_df
```

Out[106]:

	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 [107]:
```
clean_df['Age'] = clean_df['Age'].astype('int')
```

In [108]:
```
clean_df['Exp'] = clean_df.Exp.fillna(np.mean(pd.to_numeric(clean_df.Exp)))
```

In [109]:
```
clean_df
```

Out[109]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	NaN	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4.8
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [110]:
```
clean_df['Exp'] = clean_df['Exp'].astype('int')
```

In [111]:
```
clean_df.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  4 non-null      object
 4   Salary    6 non-null      int32 
 5   Exp       6 non-null      int32 
dtypes: int32(3), object(3)
memory usage: 344.0+ bytes
```

In [112]:
```
clean_df['Location'] = clean_df['Location'].fillna(clean_df['Location'].mode()[0]
```

In [113]:
```
clean_df
```

Out[113]:

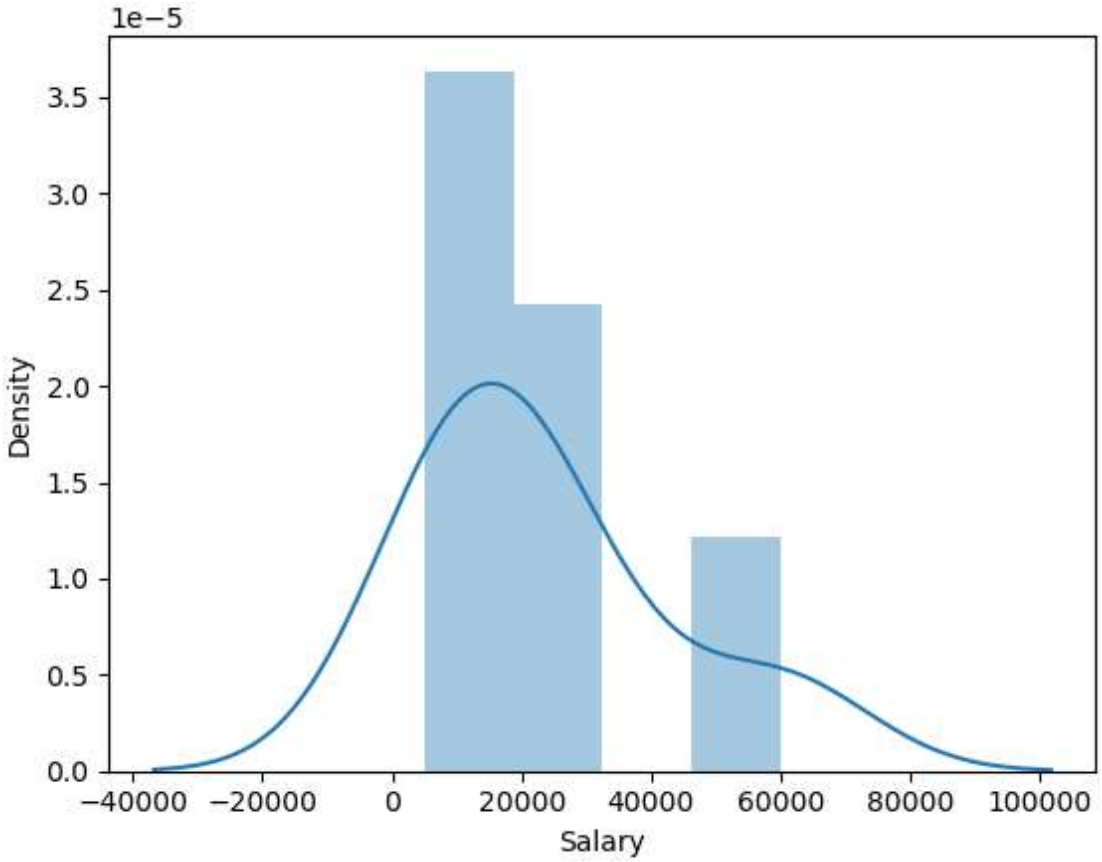
	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
clean_df.to_csv('clean_data.csv')
```

Visualization

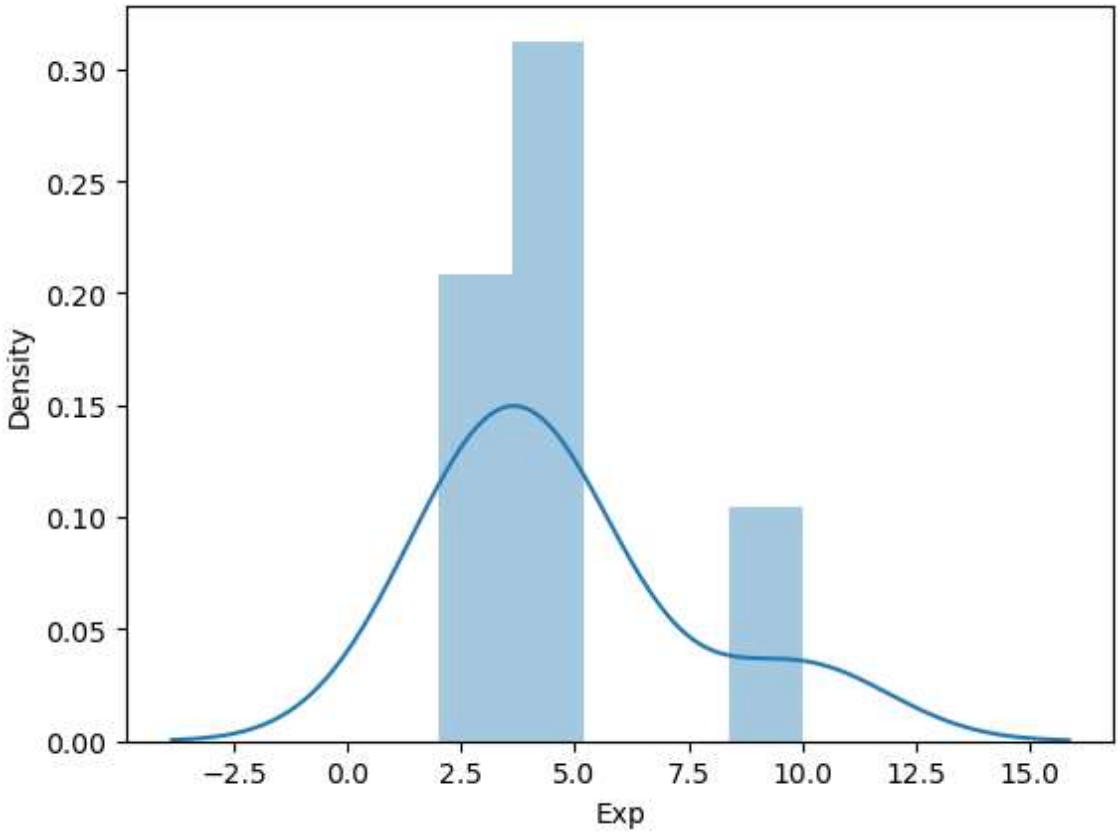
In [116]:
```
sns.distplot(clean_df['Salary'])
```

Out[116]:
```
<Axes: xlabel='Salary', ylabel='Density'>
```



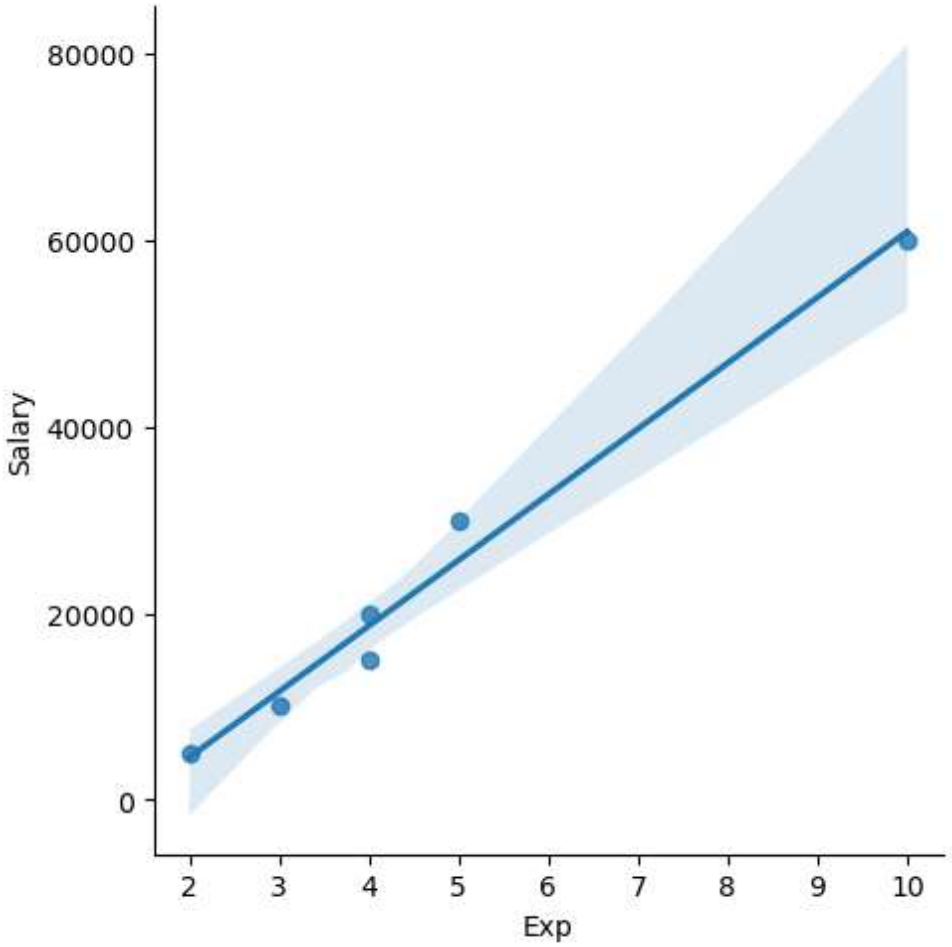
In [117]:
```
sns.distplot(clean_df['Exp'])
```

Out[117]:
```
<Axes: xlabel='Exp', ylabel='Density'>
```



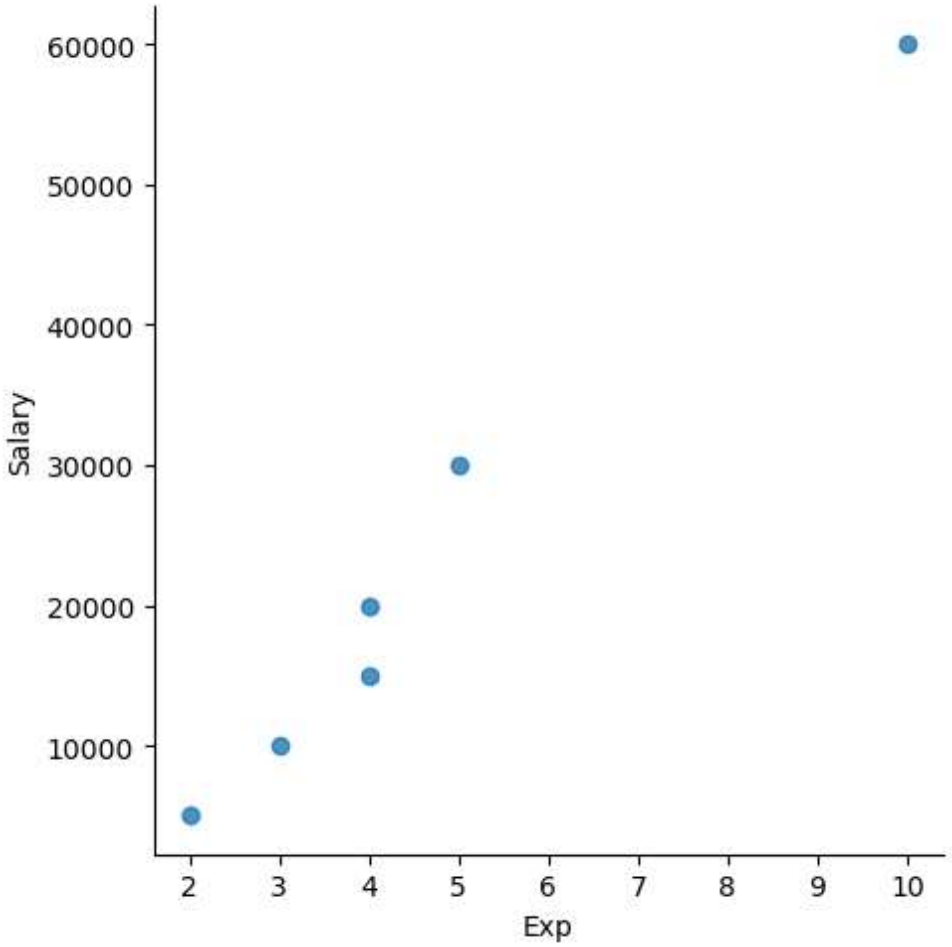
In [120]:
```
sns.lmplot(data = clean_df, y = 'Salary',x = 'Exp')
```

Out[120]:
```
<seaborn.axisgrid.FacetGrid at 0x16b5bef0eb0>
```



In [125]:
```
sns.lmplot(data = clean_df, y = 'Salary',x = 'Exp',fit_reg=False)
```

Out[125]:
```
<seaborn.axisgrid.FacetGrid at 0x16b639620e0>
```



In [126]:
```
clean_df.columns
```

Out[126]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [127]:
```
clean_ind = clean_df[['Name', 'Domain', 'Age', 'Location', 'Exp']]
```

In [128]:
```
clean_ind
```

Out[128]:

	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 [129]:
```
clean_dep = clean_df['Salary']
```

In [130]:
```
clean_dep
```

Out[130]:
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int32
```

In [131]:
```
imputation = pd.get_dummies(clean_df)
imputation
```

Out[131]:

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

In [ ]:
```
```