In [1]:

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

%matplotlib inline

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

In [2]:

```
data = pd.read_excel(r'C:\Users\LENOVO\Desktop\Monty Datascien\Rawdata.xlsx')
```

In [3]:

```
data
```

Out[3]:

	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+

Cleaning the Dataset

In [4]:

```
data['Name'] = data['Name'].str.replace(r'\W','')
data['Domain'] = data['Domain'].str.replace(r'\W','')
data['Location'] = data['Location'].str.replace(r'\W','')
```

In [5]:

```
data
```

Out[5]:

	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 [10]:

```
data['Age'] = data['Age'].str.replace(r'\W','')     # Using regular expression
data['Age'] = data['Age'].str.extract('(\d+)')
```

In [7]:

```
data['Salary'] = data['Salary'].str.replace(r'\W','')
```

In [8]:

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

In [11]:

```
data
```

Out[11]:

	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

Filling the Missing Values

In [12]:

```
data['Age'] = data['Age'].fillna(np.mean(pd.to_numeric(data['Age'])))     # with the help of mean
data['Exp'] = data['Exp'].fillna(np.mean(pd.to_numeric(data['Exp'])))
```

In [13]:

```
data
```

Out[13]:

	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	4.8
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [14]:

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

In [15]:

```
New_data = data
```

In [16]:

```
New_data
```

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	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

Updating Datatype

In [17]:

```
New_data.Name = New_data.Name.astype('category')
New_data.Domain = New_data.Domain.astype('category')
New_data.Location = New_data.Location.astype('category')
```

In [18]:

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

In [19]:

```
New_data
```

Out[19]:

	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 [20]:

```
New_data.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      category
 1   Domain    6 non-null      category
 2   Age       6 non-null      int32   
 3   Location  6 non-null      category
 4   Salary    6 non-null      int32   
 5   Exp       6 non-null      int32   
dtypes: category(3), int32(3)
memory usage: 862.0 bytes
```

Saving the cleaned Data

In [21]:

```
New_data.to_csv('New_data.csv')
```

In [22]:

```
import os
os.getcwd()
```

Out[22]:

```
'C:\\Users\\LENOVO'
```

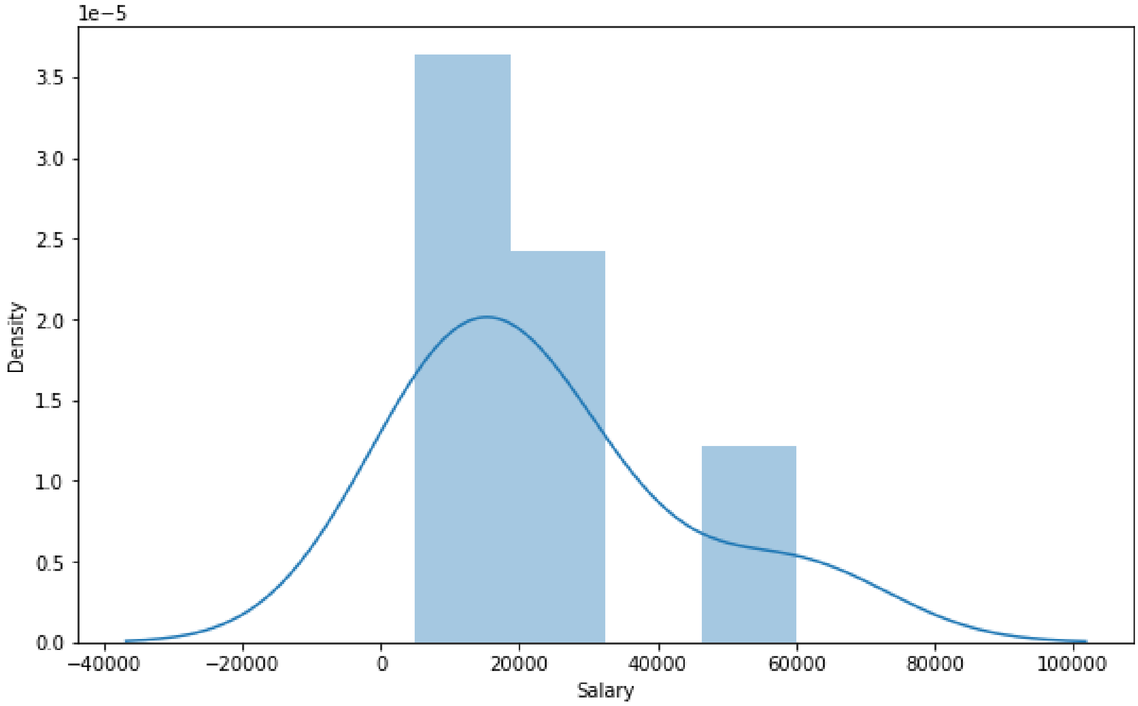
Visualization

In [23]:

```
plt.rcParams['figure.figsize'] = 10,6
```

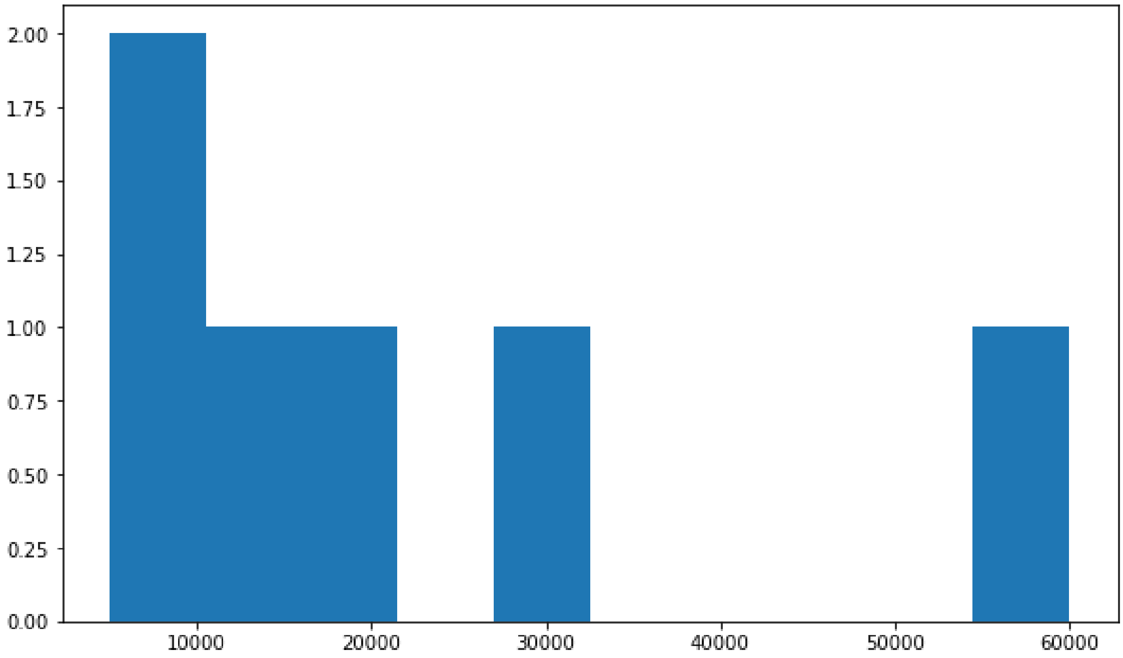
In [24]:

```
vis1 = sns.distplot(New_data['Salary'])
```



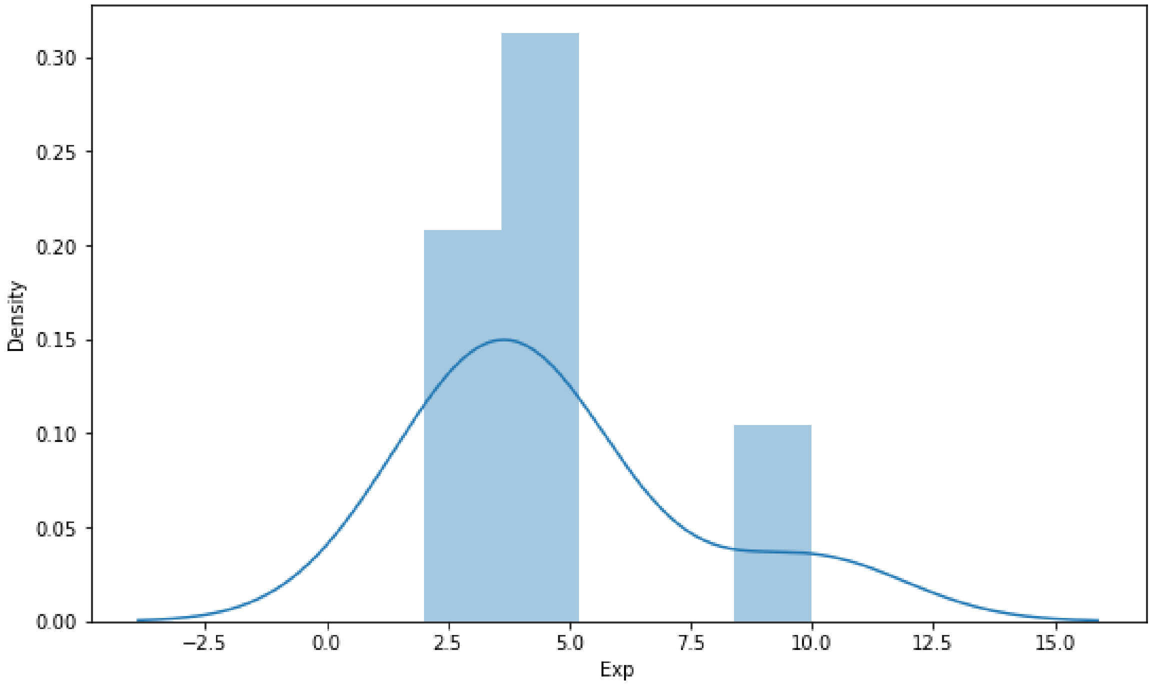
In [25]:

```
vis3 = plt.hist(New_data['Salary'])
```



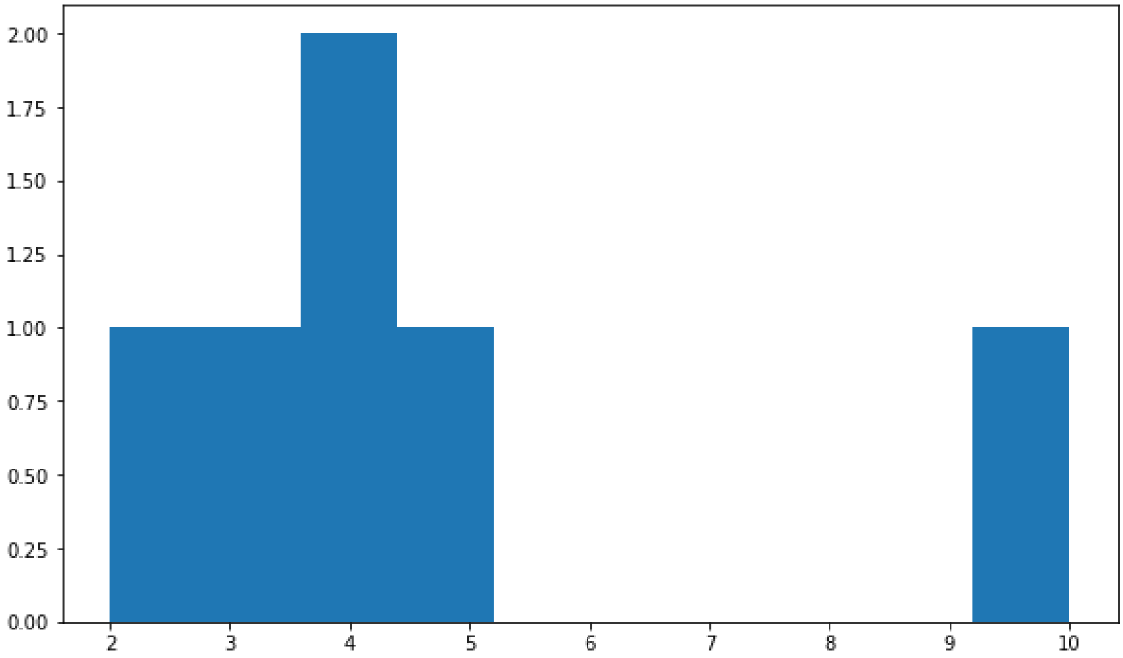
In [26]:

```
vis3 = sns.distplot(New_data['Exp'])
```



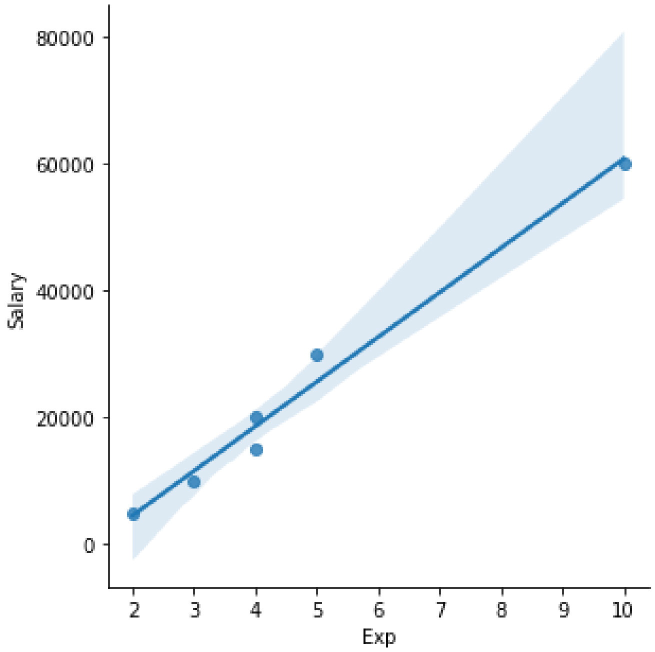
In [28]:

```
vis4 = plt.hist(New_data['Exp'])
```



In [29]:

```
vis5 = sns.lmplot(data = New_data, x = "Exp", y='Salary')
```



Splitting data

In [30]:

```
X = New_data.drop('Salary',axis=1)
```

In [31]:

```
y = New_data['Salary']
```

In [32]:

```
X
```

Out[32]:

	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 [33]:

```
y
```

Out[33]:

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

In [34]:

```
imputation = pd.get_dummies(New_data)
imputation
```

Out[34]:

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Do
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 [ ]: