

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

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

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

```
In [3]: emp=pd.read_excel(r"C:\Users\admin\Downloads\Rawdata.xlsx")
```

```
In [4]: emp
```

```
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]: id(emp)
```

```
Out[5]: 2242644699728
```

```
In [6]: emp.columns
```

```
Out[6]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

```
In [7]: emp.shape
```

```
Out[7]: (6, 6)
```

```
In [8]: emp.head()
```

```
Out[8]:
```

	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 [9]: emp.tail()
```

Out[9]:

	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 [10]: 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 [11]: emp

Out[11]:

	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 [12]: emp.isnull

Out[12]:

	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 [13]: emp.isnull()

Out[13]:

	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 [14]: `emp.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

In [15]: `emp.isnull().sum()`

Out[15]:

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

In [16]: `emp['Name']`

Out[16]:

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

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

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

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

```
In [19]: emp['Domain']
```

```
Out[19]: 0    Datascience#$
         1         Testing
         2    Dataanalyst^^#
         3         Ana^alytics
         4         Statistics
         5             NLP
         Name: Domain, dtype: object
```

```
In [20]: emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
```

```
In [21]: emp['Domain']
```

```
Out[21]: 0    Datascience
         1         Testing
         2    Dataanalyst
         3         Analytics
         4         Statistics
         5             NLP
         Name: Domain, dtype: object
```

```
In [22]: emp['Age']
```

```
Out[22]: 0    34 years
         1    45' yr
         2         NaN
         3         NaN
         4    67-yr
         5    55yr
         Name: Age, dtype: object
```

```
In [23]: emp['Age']=emp['Age'].str.replace(r'\W','',regex=True)
```

```
In [24]: emp['Age']
```

```
Out[24]: 0    34years
         1    45yr
         2         NaN
         3         NaN
         4    67yr
         5    55yr
         Name: Age, dtype: object
```

```
In [25]: emp['Age']=emp['Age'].str.extract('(\d+)')
```

```
In [26]: emp['Age']
```

```
Out[26]: 0      34
         1      45
         2     NaN
         3     NaN
         4      67
         5      55
         Name: Age, dtype: object
```

```
In [27]: emp['Location']
```

```
Out[27]: 0      Mumbai
         1    Bangalore
         2         NaN
         3    Hyderabad
         4         NaN
         5        Delhi
         Name: Location, dtype: object
```

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

```
In [29]: emp['Location']
```

```
Out[29]: 0      Mumbai
         1    Bangalore
         2         NaN
         3    Hyderabad
         4         NaN
         5        Delhi
         Name: Location, dtype: object
```

```
In [30]: emp['Salary']
```

```
Out[30]: 0      5^00#0
         1     10%%000
         2     1$5%000
         3      2000^0
         4      30000-
         5     6000^$0
         Name: Salary, dtype: object
```

```
In [31]: emp['Salary']=emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
```

```
In [32]: emp['Salary']
```

```
Out[32]: 0      5000
         1     10000
         2     15000
         3     20000
         4     30000
         5     60000
         Name: Salary, dtype: object
```

```
In [33]: emp['Exp']
```

```
Out[33]: 0      2+
        1      <3
        2      4> yrs
        3      NaN
        4      5+ year
        5      10+
        Name: Exp, dtype: object
```

```
In [34]: emp['Exp']=emp['Exp'].str.extract('(\d+)')
```

```
In [35]: emp['Exp']
```

```
Out[35]: 0      2
        1      3
        2      4
        3      NaN
        4      5
        5      10
        Name: Exp, dtype: object
```

```
In [36]: clean_data=emp.copy()
```

```
In [37]: clean_data
```

```
Out[37]:
```

	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 [38]: import numpy as np
```

```
In [39]: clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])))
```

```
In [40]: clean_data['Exp']
```

```
Out[40]: 0      2
        1      3
        2      4
        3      NaN
        4      5
        5      10
        Name: Exp, dtype: object
```

```
In [41]: clean_data['Location'].isnull().sum()
```

Out[41]: np.int64(2)

In [42]: clean_data['Location']

Out[42]: 0 Mumbai
1 Bangalore
2 NaN
3 Hyderabad
4 NaN
5 Delhi
Name: Location, dtype: object

In [43]: clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode(0

In [44]: clean_data['Location']

Out[44]: 0 Mumbai
1 Bangalore
2 NaN
3 Hyderabad
4 NaN
5 Delhi
Name: Location, dtype: object

In [45]: clean_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     object
1   Domain      6 non-null     object
2   Age         6 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 [46]: clean_data['Age'] = clean_data['Age'].astype(int)

In [47]: clean_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      object
1   Domain      6 non-null      object
2   Age         6 non-null      int64
3   Location    4 non-null      object
4   Salary      6 non-null      object
5   Exp         5 non-null      object
dtypes: int64(1), object(5)
memory usage: 420.0+ bytes
```

```
In [48]: clean_data['Salary']=clean_data['Salary'].astype(int)
```

```
In [49]: clean_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      object
1   Domain      6 non-null      object
2   Age         6 non-null      int64
3   Location    4 non-null      object
4   Salary      6 non-null      int64
5   Exp         5 non-null      object
dtypes: int64(2), object(4)
memory usage: 420.0+ bytes
```

```
In [50]: clean_data
```

```
Out[50]:
```

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

```
In [51]: clean_data['Exp']= clean_data['Exp'].fillna(0)
```

```
In [52]: clean_data['Exp']= clean_data['Exp'].astype(int)
```

```
In [53]: clean_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      object
1   Domain      6 non-null      object
2   Age         6 non-null      int64
3   Location    4 non-null      object
4   Salary      6 non-null      int64
5   Exp         6 non-null      int64
dtypes: int64(3), object(3)
memory usage: 420.0+ bytes
```

```
In [54]: clean_data['Location']=clean_data['Location'].fillna(clean_data['Location'].mode()[
```

```
In [55]: clean_data
```

```
Out[55]:
```

	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	0
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [56]: clean_data.to_excel('cleandata.xlsx',index=False)
```

```
In [57]: import os
os.getcwd()
```

```
Out[57]: 'C:\\Users\\admin\\Documents\\Data science\\Python'
```

```
In [58]: clean_data
```

```
Out[58]:
```

	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	0
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

EDA Technique lets Apply

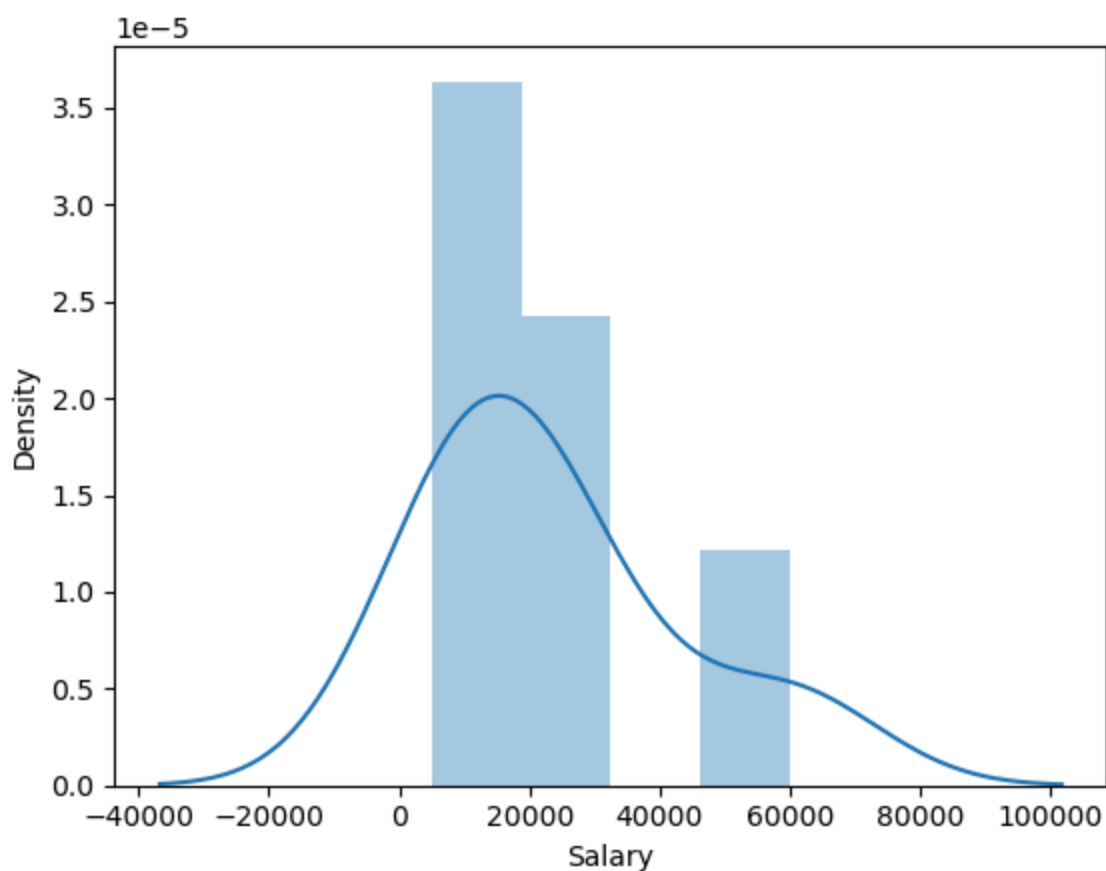
```
In [59]: import matplotlib.pyplot as plt  
import seaborn as sns
```

```
In [60]: import warnings  
warnings.filterwarnings('ignore')
```

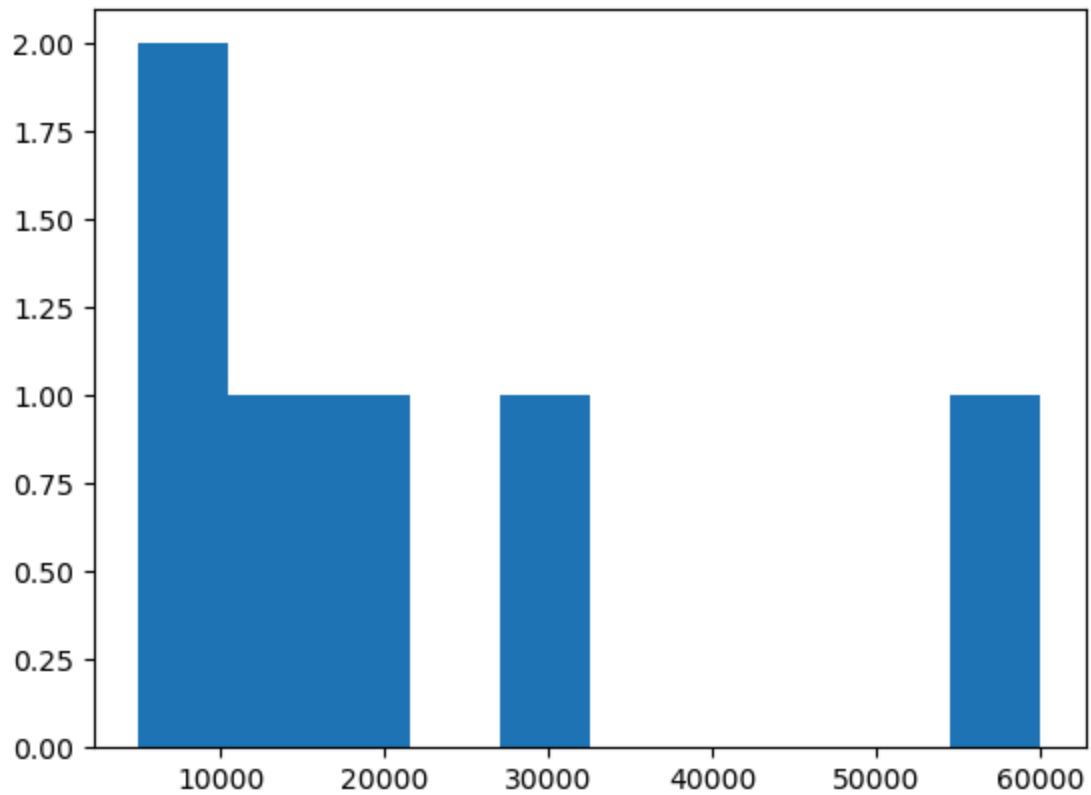
```
In [62]: clean_data['Salary']
```

```
Out[62]: 0    5000  
1    10000  
2    15000  
3    20000  
4    30000  
5    60000  
Name: Salary, dtype: int64
```

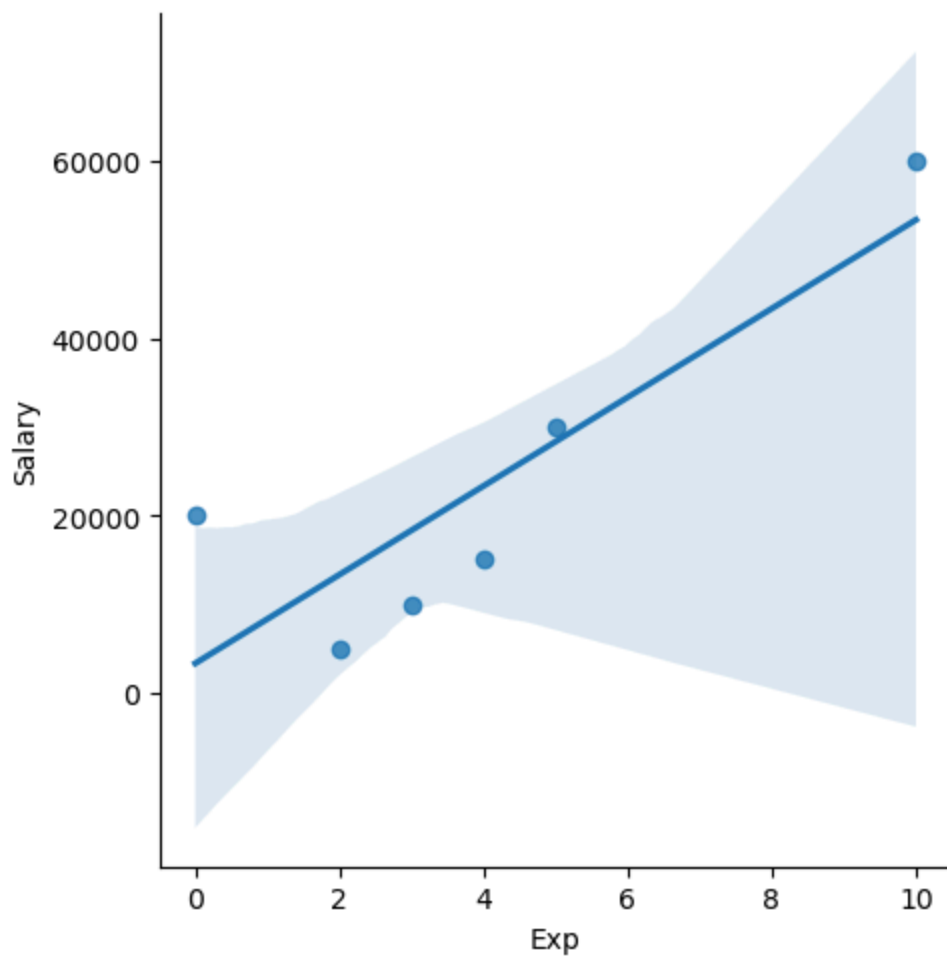
```
In [64]: vis1=sns.distplot(clean_data['Salary'])
```



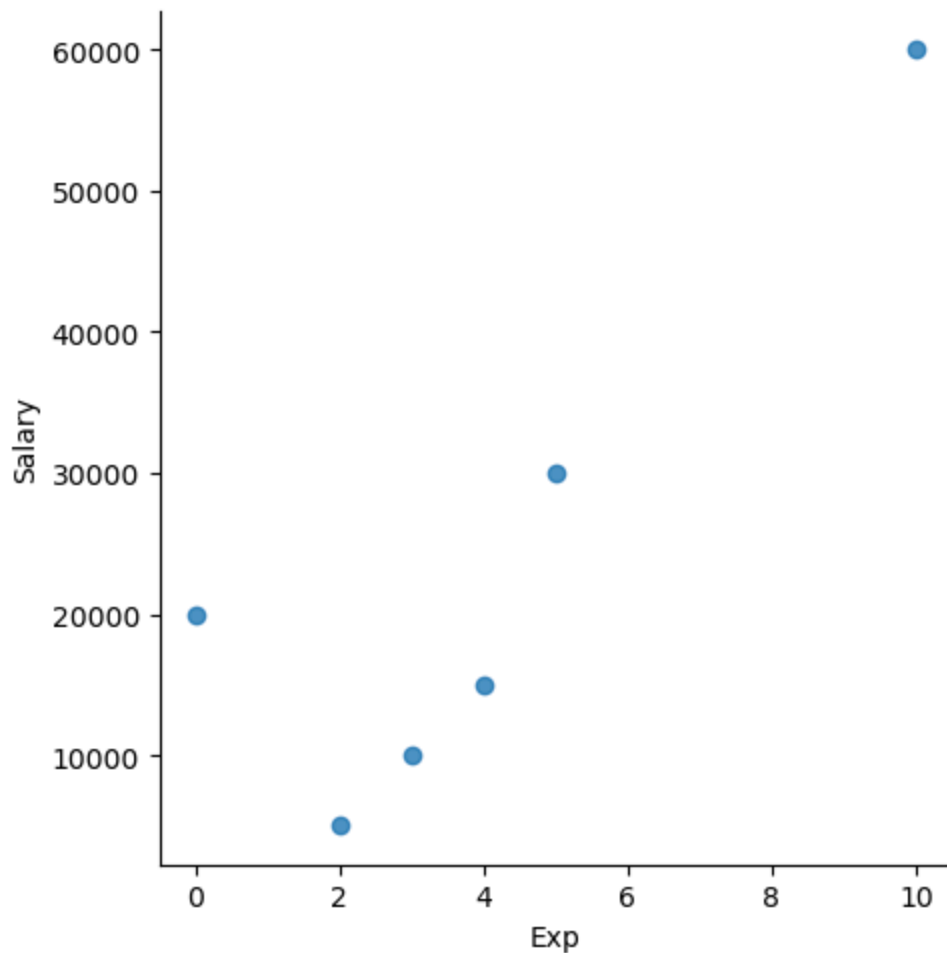
```
In [65]: vis2=plt.hist(clean_data['Salary'])
```



```
In [70]: vis4=sns.lmplot(data=clean_data,x='Exp',y='Salary')
```



```
In [71]: vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)
```



```
In [72]: clean_data[:]
```

```
Out[72]:
```

	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	0
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

```
In [73]: clean_data[0:6:2]
```

Out[73]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
2	Umar	Dataanalyst	50	Bangalore	15000	4
4	Uttam	Statistics	67	Bangalore	30000	5

In [74]: `clean_data[::-1]`

Out[74]:

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

In [75]: `clean_data.columns`

Out[75]: `Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')`

In [76]: `X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]`

In [77]: `X_iv`

Out[77]:

	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	0
4	Uttam	Statistics	67	Bangalore	5
5	Kim	NLP	55	Delhi	10

In [78]: `y_dv = clean_data[['Salary']]`

In [79]: `y_dv`

Out[79]: **Salary****0** 5000**1** 10000**2** 15000**3** 20000**4** 30000**5** 60000

In [80]: emp

Out[80]: **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 [81]: clean_data

Out[81]: **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 0**4** Uttam Statistics 67 Bangalore 30000 5**5** Kim NLP 55 Delhi 60000 10

In [82]: X_iv

Out[82]:

	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	0
4	Uttam	Statistics	67	Bangalore	5
5	Kim	NLP	55	Delhi	10

In [83]: `y_dv`

Out[83]:

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

In [84]: `clean_data`

Out[84]:

	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	0
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [88]: `imputation = pd.get_dummies(clean_data)`

In [86]: `imputation`

Out[86]:

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

In [87]:

clean_data

Out[87]:

	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	0
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10