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

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

Out[2]:
```
'2.2.2'
```

In [3]:
```
emp=pd.read_excel(r'C:\Users\S SHYAMILI\OneDrive\Desktop\data science\Rawdata.xlsx'
```

In [4]:
```
emp.head()
```

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

In [5]:
```
emp.columns
```

Out[5]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

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

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

In [7]:
```
id(emp)
```

Out[7]:
```
1828105542096
```

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

Out[8]:

	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 [9]:
```
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 [10]:
```
emp.isnull().sum()
```

Out[10]:
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [11]:
```
emp['Name']
```

Out[11]:
```
0      Mike
1    Teddy^
2     Uma#r
3      Jane
4    Uttam*
5       Kim
Name: Name, dtype: object
```

In [12]:
```
emp['Name']=emp['Name'].str.replace(r'\W','',regex=True)
```

In [13]:
```
emp['Name']
```

Out[13]:
```
0     Mike
1    Teddy
2     Umar
3     Jane
4    Uttam
5      Kim
Name: Name, dtype: object
```

In [16]:
```
emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
```

In [17]:
```
emp['Domain']
```

Out[17]:
```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

In [18]:
```
emp['Age']=emp['Age'].str.replace(r'\W','',regex=True)
```

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

Out[19]:
```
0    34years
1       45yr
2        NaN
3        NaN
4       67yr
5       55yr
Name: Age, dtype: object
```

In [20]:
```
emp['Age']=emp['Age'].str.extract(r'(\d+)')
```

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

Out[21]:
```
0     34
1     45
2    NaN
3    NaN
4     67
5     55
Name: Age, dtype: object
```

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

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

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

In [36]:
```
emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
```

In [43]:
```
emp['Salary']
```

Out[43]:
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: object
```

In [45]:
```
emp['Exp']=emp['Exp'].str.extract(r'(\d+)')
```

In [47]:
```
emp['Exp']
```

Out[47]:
```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

In [50]:
```
clean_data=emp.copy()
```

In [57]:
```
clean_data.isnull().sum()
```

Out[57]:
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [59]:
```
import numpy as np
```

In [61]:
```
clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])
```

In [63]:
```
clean_data['Age']
```

Out[63]:
```
0       34
1       45
2    50.25
3    50.25
4       67
5       55
Name: Age, dtype: object
```

In [71]:
```
clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp'])
```

In [104…
```
clean_data['Exp']
```

Out[104…
```
0      2
1      3
2      4
3    4.8
4      5
5     10
Name: Exp, dtype: object
```

In [65]:
```
clean_data['Location']=clean_data['Location'].fillna((clean_data['Location'])).mode
```

In [67]:
```
clean_data['Location']
```

Out[67]:
```
0    Bangalore
1    Bangalore
2    Bangalore
3    Bangalore
4    Bangalore
5    Bangalore
Name: Location, dtype: object
```

In [73]:
```
clean_data
```

Out[73]:

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

In [75]:
```
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  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 [112…
```
clean_data.to_csv('clean_data.csv')
```

In [77]:
```
clean_data['Name']=clean_data['Name'].astype('category')
```

In [81]:
```
clean_data['Domain']=clean_data['Domain'].astype('category')
```

In [79]:
```
clean_data['Location']=clean_data['Location'].astype('category')
```

In [85]:
```
clean_data['Age']=clean_data['Age'].astype('int')
```

In [87]:
```
clean_data['Salary']=clean_data['Salary'].astype('int')
```

In [89]:
```
clean_data['Exp']=clean_data['Exp'].astype('int')
```

In [91]:
```
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      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: 778.0 bytes
```

In [101…
```
clean_data.to_csv('clean_data.csv')
```

In [103…
```
import os
```

In [105…
```
os.getcwd()
```

Out[105…
```
'C:\\Users\\S SHYAMILI'
```

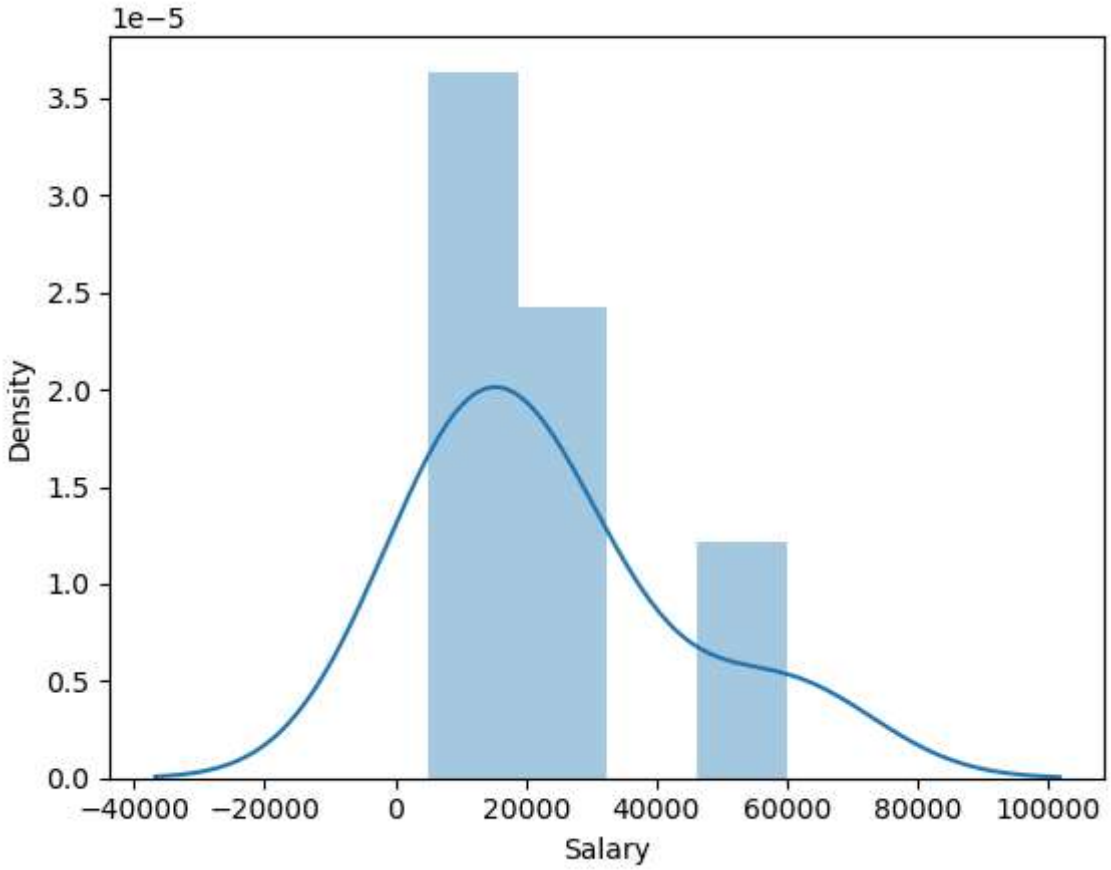
In [107…
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [109…
```
import warnings
warnings.filterwarnings('ignore')
```

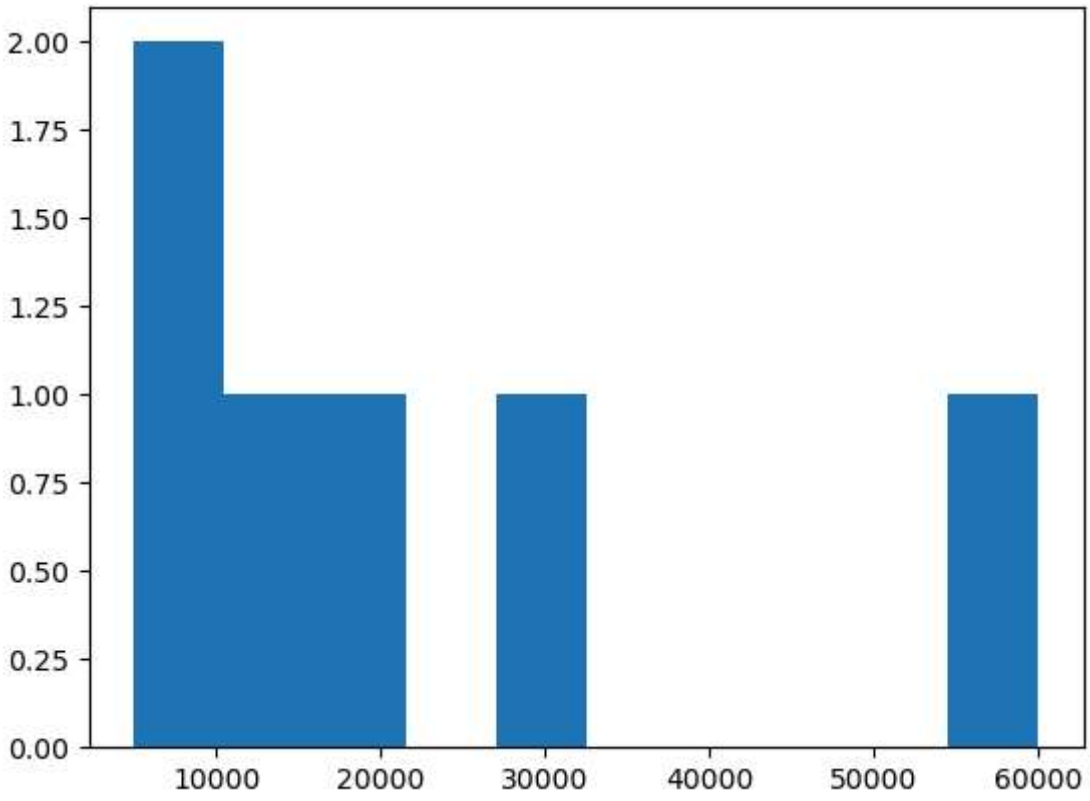
In [111…
```
clean_data['Salary']
```

Out[111…
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int32
```

In [113…
```
vis1=sns.distplot(clean_data['Salary'])
```



In [117…
```
viss2=plt.hist(clean_data['Salary'])
```



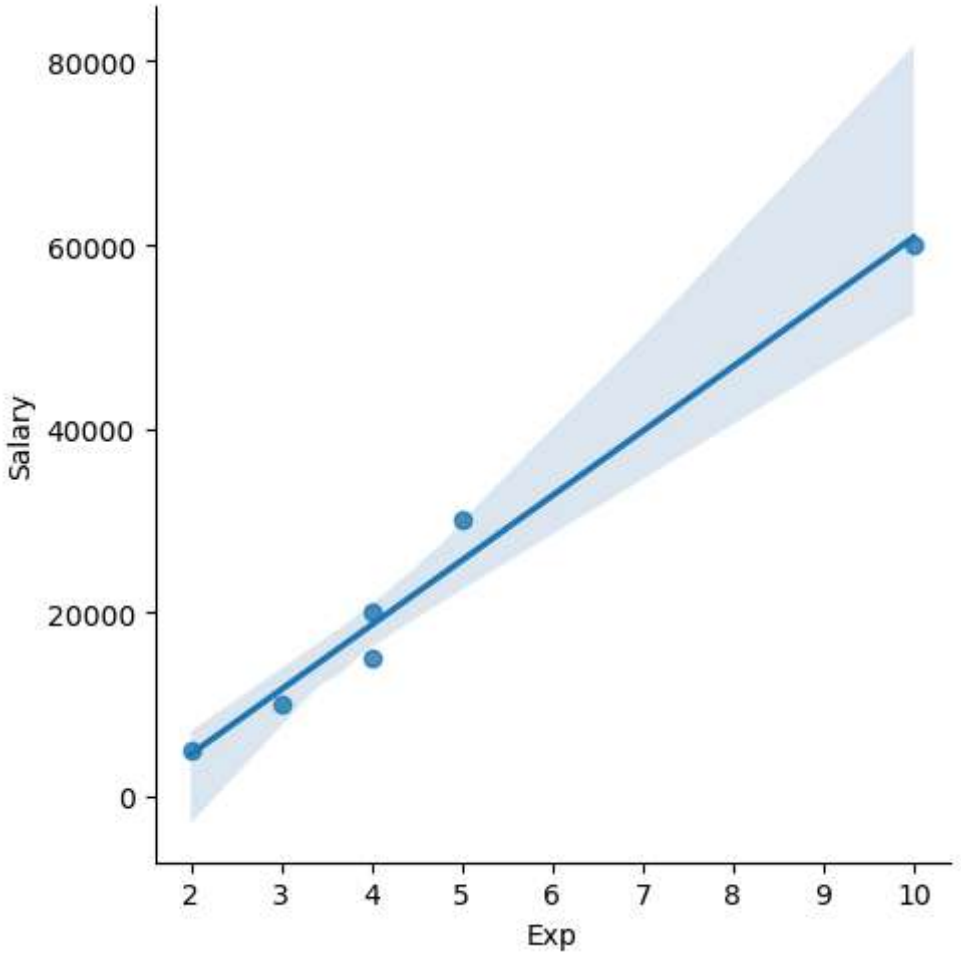
In [119…
```
clean_data
```

Out[119…

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

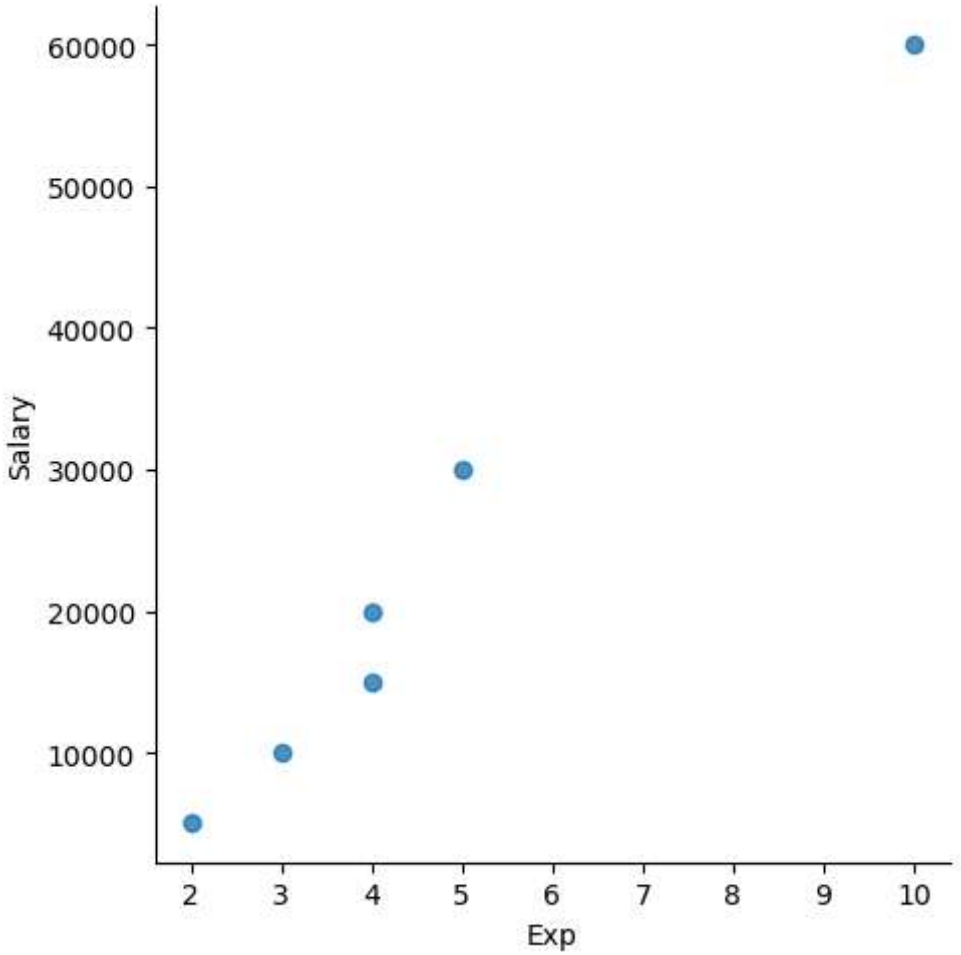
In [125…

```
vis3=sns.lmplot(data=clean_data,x='Exp',y='Salary')
```



In [129…

```
vis3=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=False)
```



In [131…
```
clean_data[::]
```

Out[131…

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

In [133…
```
clean_data[:2:]
```

Out[133…

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Bangalore	5000	2
1	Teddy	Testing	45	Bangalore	10000	3

In [135…
```
clean_data[-2:5:]
```

Out[135…

	Name	Domain	Age	Location	Salary	Exp
4	Uttam	Statistics	67	Bangalore	30000	5

In [137…
```
clean_data
```

Out[137…

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

In [145…
```
x_iv=clean_data[['Name','Domain','Age','Location','Exp']]
```

In [149…
```
y_dv=clean_data['Salary']
```

In [153…
```
imputations=pd.get_dummies(clean_data)
```

In [155…
```
imputations
```

Out[155…

	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	4	True	False	False	False	False	
4	67	30000	5	False	False	False	False	False	
5	55	60000	10	False	True	False	False	False	

In [157…
```
imputations=pd.get_dummies(clean_data,dtype=int)
```

In [159…
```
imputations
```

Out[159…

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Nan
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 [163…

```
len(imputations.columns)
```

Out[163… 16

In [ ]: