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

Out[1]:
```
'2.2.2'
```

In [2]:
```
#pip install --upgrade openpyxl
```

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

In [4]:
```
df
```

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(df)
```

Out[5]:
```
1593733166000
```

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

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

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

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

In [8]:
```
df.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]:

```
df.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]:

```
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: 420.0+ bytes
```

In [11]:

```python
df.isnull()
```

Out[11]:

	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 [12]:

```python
df.isna()
```

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]:

```python
df.isnull().sum()  # gives the count of null values
```

Out[13]:

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

Data Cleaning

```
In [14]: df
```

Out[14]:

	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 [15]: df['Name']
```

```
Out[15]: 0      Mike
         1    Teddy^
         2     Uma#r
         3      Jane
         4    Uttam*
         5       Kim
         Name: Name, dtype: object
```

```
In [16]: df['Name'] = df['Name'].str.replace(r'\W','',regex=True)   # nonword character
```

```
In [17]: df['Name']
```

```
Out[17]: 0     Mike
         1    Teddy
         2     Umar
         3     Jane
         4    Uttam
         5      Kim
         Name: Name, dtype: object
```

```
In [18]: df['Domain']
```

```
Out[18]: 0     Datascience#$
         1           Testing
         2    Dataanalyst^^#
         3       Ana^^lytics
         4        Statistics
         5               NLP
         Name: Domain, dtype: object
```

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

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

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

```
In [21]: df['Location']
```

```
Out[21]: 0       Mumbai
         1    Bangalore
         2          NaN
         3     Hyderbad
         4          NaN
         5        Delhi
         Name: Location, dtype: object
```

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

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

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

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

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

```
In [25]: df['Age'] =df['Age'].str.extract('(\\d+)')   # \\d is used to exract the string
```

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

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

```
In [27]: df
```

Out[27]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5^00#0	2+
1	Teddy	Testing	45	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	NaN	30000-	5+ year
5	Kim	NLP	55	Delhi	6000^$0	10+

In [28]:

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

In [29]:

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

In [30]:

```
df
```

Out[30]:

	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 [31]:

```
clean_data = df.copy()
```

In [32]:

```
clean_data
```

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

EDA techniques

In [33]:

```
clean_data.isnull().sum()
```

Out[33]:

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

In [34]:

```
import numpy as np
```

In [35]:

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

In [36]:

```
clean_data['Age']
```

Out[36]:

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

In [37]:

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

In [38]:

```
clean_data['Exp']
```

Out[38]:

```
0      2
1      3
2      4
3    4.8
4      5
5     10
Name: Exp, dtype: object
```

In [39]:
```
clean_data
```

Out[39]:

	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 [40]:
```
clean_data['Location'] =clean_data['Location'].fillna(clean_data['Location'].mode()[0])
```

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

get the code from memory to system to share

In [43]:

```
clean_data.to_csv('clean_data2.csv')
```

In [44]:

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

Out[44]:

```
'C:\\Users\\ADMIN\\vs code projects'
```

In [45]:

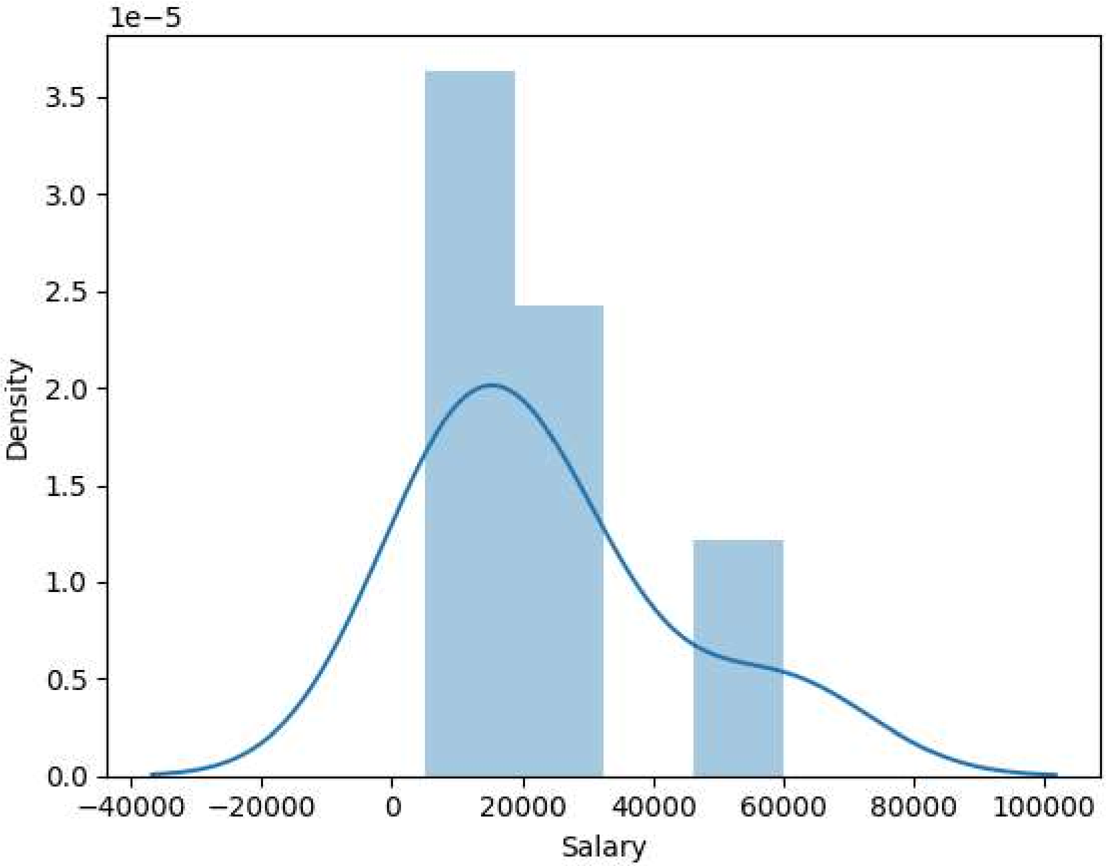
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [46]:

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

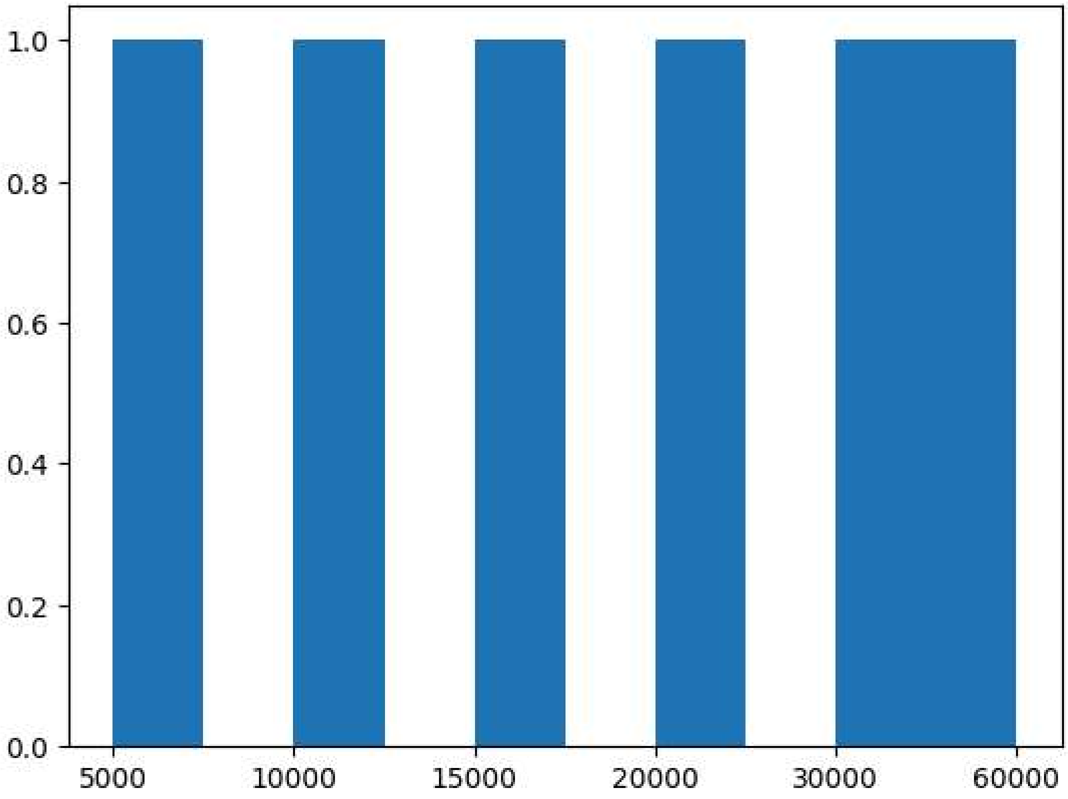
In [47]:

```
vis1 = sns.distplot(clean_data['Salary'])  # univarient
plt.show()
```



In [48]:

```
vis2 = plt.hist(clean_data['Salary'])
plt.show()
```



```
In [49]: clean_data
```

Out[49]:

	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 [50]:

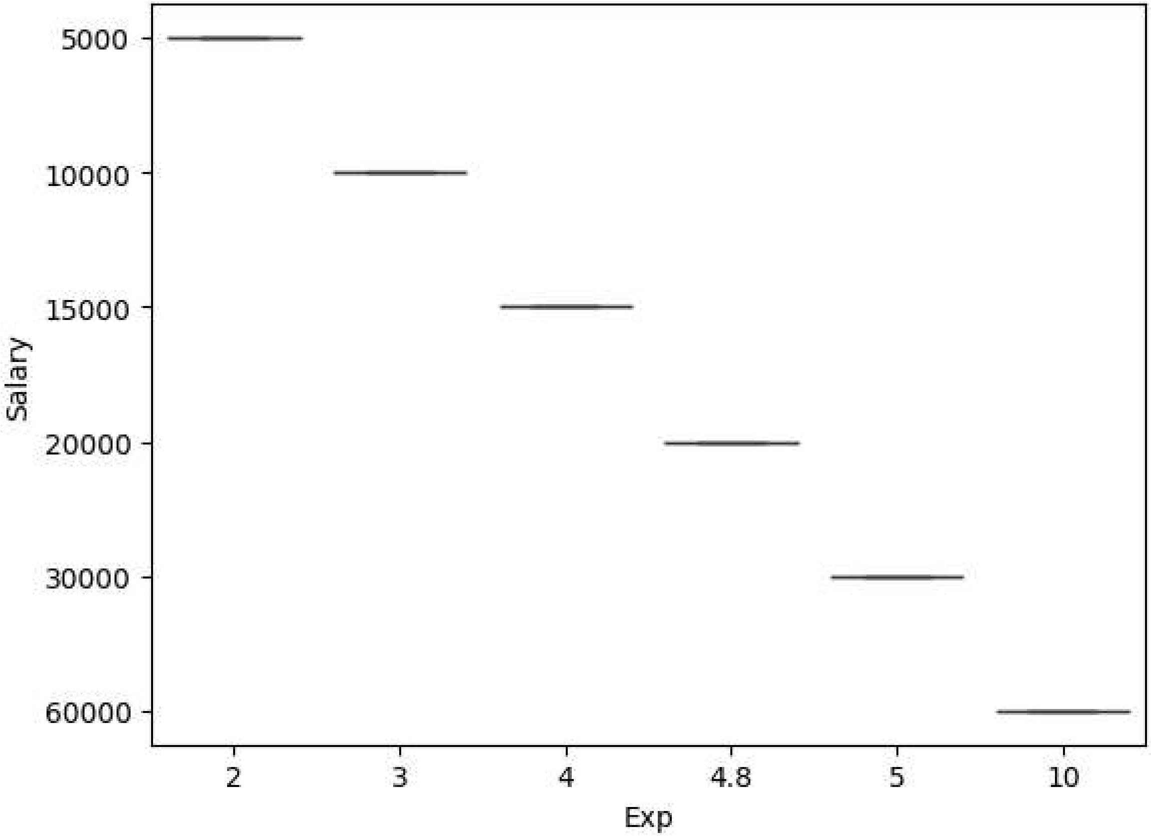
```
clean_data
```

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	50.25	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

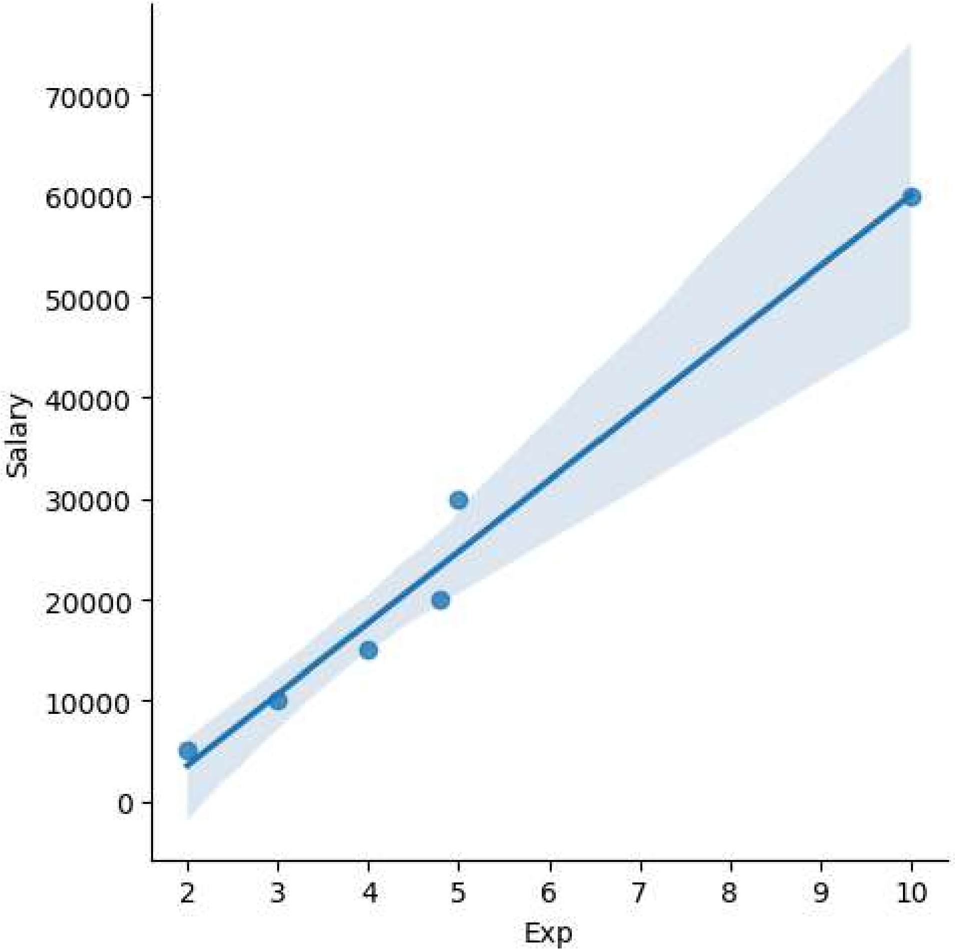
In [51]:

```
vis3 = sns.boxplot(data=clean_data,x='Exp',y='Salary')
plt.show()
```



```
In [52]: clean_data['Exp'] = pd.to_numeric(clean_data['Exp'], errors='coerce')
         clean_data['Salary'] = pd.to_numeric(clean_data['Salary'], errors='coerce')
         clean_data = clean_data.dropna(subset=['Exp', 'Salary'])
```

```
In [53]: vis4 = sns.lmplot(data=clean_data,x='Exp',y='Salary')
         plt.show()
```



In [54]:
```
clean_data
```

Out[54]:

	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 [55]:

```
y = clean_data['Salary']
```

In [56]:

```
y
```

Out[56]:

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

In [57]:

```
clean_data.columns
```

Out[57]:

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

In [58]:

```
x=clean_data[['Name', 'Domain', 'Age', 'Location','Exp']]
```

In [59]:

```
x
```

Out[59]:

	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 [60]:

```
clean_data
```

Out[60]:

	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 [61]:

```
imputation=pd.get_dummies(clean_data,dtype=int)
```

In [62]:

```
imputation
```

Out[62]:

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

6 rows × 23 columns

In [63]:

```
len(imputation.columns)
```

Out[63]:

```
23
```