```
import pandas as pd
In [1]:
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^
                                      45' yr
                                            Bangalore
                                                       10%%000
                                                                     <3
                            Testing
         2
            Uma#r
                    Dataanalyst^^#
                                       NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
         3
                        Ana^^lytics
                                       NaN Hyderbad
                                                         2000^0
                                                                    NaN
              Jane
            Uttam*
                          Statistics
                                                 NaN
                                                                 5+ year
                                      67-yr
                                                         30000-
               Kim
                              NLP
                                       55yr
                                                 Delhi
                                                        6000^$0
                                                                    10+
In [5]:
         id(emp)
         2242644699728
Out[5]:
         emp.columns
In [6]:
         Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[6]:
In [7]:
        emp.shape
Out[7]:
         (6, 6)
        emp.head()
In [8]:
Out[8]:
                           Domain
             Name
                                       Age
                                             Location
                                                          Salary
                                                                    Exp
         0
              Mike
                      Datascience#$
                                    34 years
                                              Mumbai
                                                         5^00#0
                                                                     2+
         1 Teddy^
                            Testing
                                      45' yr
                                            Bangalore
                                                       10%%000
                                                                     <3
            Uma#r Dataanalyst^^#
                                       NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
         3
              Jane
                        Ana^^lytics
                                       NaN Hyderbad
                                                         2000^0
                                                                    NaN
         4 Uttam*
                          Statistics
                                                 NaN
                                                         30000- 5+ year
                                      67-yr
        emp.tail()
In [9]:
```

Out[9]:		Name	Domain	Age	Location	Salary	Ехр
	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
dtvn	es: obiect	(6)	

memory usage: 420.0+ bytes

In [11]: emp

Out[11]:		Name	Domain	Age	Location	Salary	Ехр
Out[11]:	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]:	<b< th=""><th>ound met</th><th>hod DataFrame.is</th><th>null of</th><th>Name</th><th>Dom</th><th>ain</th><th>Age</th><th>Location</th></b<>	ound met	hod DataFrame.is	null of	Name	Dom	ain	Age	Location
	Sa	lary	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	Ехр
	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]:	em	p.isna(	)				
Out[14]:		Name	Domain	Age	Location	Salary	Ехр
	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
Tn [15].	o.m.	n fanul	1/) sum/)				
In [15]:			1().sum()				
Out[15]:	Na Do	me main	0 0				
	Ag Lo	e cation	2				
	Sa	lary	0				
	Ex dt	p ype: in	1 nt64				
In [16]:	em	p['Name	']				
Out[16]:	0	Mi	.ke				
	1	Tedd	ly^				
	2 3	Uma Ja	ı#r ıne				
	4 5	Utta K	ım* (im				
			e, dtype	: obje	ct		
In [17]:	em	p['Name	']=emp['N	Name']	.str.repla	ace(r'\W	۱,'','
In [18]:	emi	p['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]:
                Datascience#$
                      Testing
          1
          2
               Dataanalyst^^#
          3
                  Ana^^lytics
          4
                   Statistics
                           NLP
          Name: Domain, dtype: object
         emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
In [20]:
In [21]:
          emp['Domain']
Out[21]:
               Datascience
          1
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
                       NLP
          Name: Domain, dtype: object
In [22]:
         emp['Age']
Out[22]:
               34 years
                 45' yr
          1
          2
                    NaN
          3
                    NaN
          4
                  67-yr
                   55yr
          Name: Age, dtype: object
         emp['Age']=emp['Age'].str.replace(r'\W','',regex=True)
In [23]:
         emp['Age']
In [24]:
Out[24]:
          0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
          5
                  55yr
          Name: Age, dtype: object
          emp['Age']=emp['Age'].str.extract('(\\d+)')
In [25]:
In [26]:
         emp['Age']
```

```
Out[26]: 0
                34
          1
                45
          2
               NaN
          3
               NaN
                67
                55
          Name: Age, dtype: object
         emp['Location']
In [27]:
Out[27]: 0
                  Mumbai
               Bangalore
          1
          2
                     NaN
          3
                Hyderbad
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
         emp['Location']=emp['Location'].str.replace(r'\W','',regex=True)
In [28]:
In [29]:
         emp['Location']
Out[29]:
                  Mumbai
               Bangalore
          1
          2
                     NaN
          3
                Hyderbad
          4
                     NaN
                   Delhi
          Name: Location, dtype: object
In [30]:
         emp['Salary']
Out[30]:
                5^00#0
               10%%000
          1
          2
               1$5%000
                2000^0
          3
          4
                30000-
               6000^$0
          Name: Salary, dtype: object
In [31]:
         emp['Salary']=emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
         emp['Salary']
In [32]:
          0
                5000
Out[32]:
          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
               5+ year
          Name: Exp, dtype: object
In [34]: emp['Exp']=emp['Exp'].str.extract('(\\d+)')
In [35]:
         emp['Exp']
Out[35]:
                 2
                 3
          1
          2
                 4
          3
               NaN
                 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
                                       Mumbai
                                                 5000
                                                          2
                                 34
             Teddy
                        Testing
                                     Bangalore
                                                10000
                    Dataanalyst NaN
                                                15000
                                                          4
          2
             Umar
                                          NaN
          3
              Jane
                      Analytics NaN
                                     Hyderbad
                                                20000
                                                       NaN
          4 Uttam
                       Statistics
                                                30000
                                 67
                                          NaN
                                                          5
               Kim
                          NLP
                                 55
                                          Delhi
                                                60000
                                                         10
In [38]:
         import numpy as np
         clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])
In [39]:
In [40]:
         clean_data['Exp']
Out[40]: 0
                 2
          1
                 3
          2
                 4
          3
               NaN
                 5
                10
          Name: Exp, dtype: object
In [41]: clean_data['Location'].isnull().sum()
```

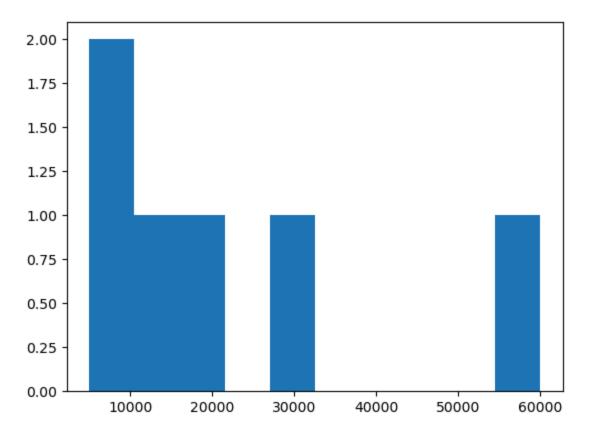
```
Out[41]: np.int64(2)
In [42]:
         clean_data['Location']
Out[42]:
                  Mumbai
               Bangalore
          1
          2
                     NaN
          3
                Hyderbad
          4
                     NaN
                   Delhi
         Name: Location, dtype: object
         clean_data['Location'] =clean_data['Location'].fillna(clean_data['Location'].mode(@
In [44]:
         clean_data['Location']
Out[44]: 0
                  Mumbai
               Bangalore
          1
          2
                     NaN
          3
               Hyderbad
          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):
                       Non-Null Count Dtype
             Column
             -----
         0
             Name
                       6 non-null
                                       object
         1
             Domain
                       6 non-null
                                       object
         2
             Age
                       6 non-null
                                       object
         3
             Location 4 non-null
                                       object
                       6 non-null
         4
             Salary
                                       object
                       5 non-null
         5
             Exp
                                       object
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [46]:
         clean_data['Age']=clean_data['Age'].astype(int)
         clean_data.info()
In [47]:
```

```
<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
                       5 non-null
                                        object
             Exp
        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):
                       Non-Null Count Dtype
             Column
            -----
                       -----
        ---
                                        _ _ _ _
         0
             Name
                       6 non-null
                                        object
             Domain
                       6 non-null
         1
                                        object
         2
             Age
                       6 non-null
                                        int64
         3
             Location 4 non-null
                                        object
         4
             Salary
                       6 non-null
                                        int64
         5
                       5 non-null
                                        object
        dtypes: int64(2), object(4)
        memory usage: 420.0+ bytes
In [50]:
         clean data
Out[50]:
                                     Location Salary
            Name
                      Domain Age
                                                      Exp
         0
              Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                         2
             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
         clean_data['Exp'] = clean_data['Exp'].fillna(0)
In [51]:
In [52]:
         clean_data['Exp'] = clean_data['Exp'].astype(int)
In [53]:
         clean_data.info()
```

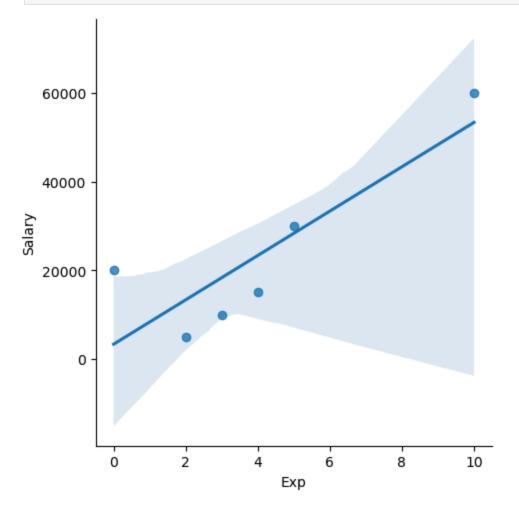
```
<class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
              Column
                        Non-Null Count Dtype
             -----
                                         ____
         0
             Name
                        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
                        6 non-null
                                         int64
              Exp
        dtypes: int64(3), object(3)
        memory usage: 420.0+ bytes
          clean_data['Location']=clean_data['Location'].fillna(clean_data['Location'].mode()[
In [54]:
          clean data
In [55]:
Out[55]:
                       Domain Age
                                       Location Salary Exp
             Name
          0
              Mike Datascience
                                  34
                                       Mumbai
                                                  5000
                                                          2
             Teddy
                        Testing
                                      Bangalore
                                                 10000
                                                          3
          2
              Umar
                    Dataanalyst
                                      Bangalore
                                                 15000
                                                          4
          3
              Jane
                       Analytics
                                      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()
          'C:\\Users\\admin\\Documents\\Data science\\Python'
Out[57]:
In [58]:
          clean_data
Out[58]:
             Name
                       Domain Age
                                       Location Salary
                                                        Exp
          0
              Mike Datascience
                                       Mumbai
                                                  5000
                                  34
                                                          2
             Teddy
                        Testing
                                      Bangalore
                                                 10000
                                                          3
          2
              Umar
                    Dataanalyst
                                  50
                                      Bangalore
                                                 15000
                                                          4
          3
              Jane
                       Analytics
                                  50
                                      Hyderbad
                                                 20000
                                                          0
                                                 30000
                                                          5
          4
             Uttam
                       Statistics
                                      Bangalore
                                  67
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                 60000
                                                         10
```

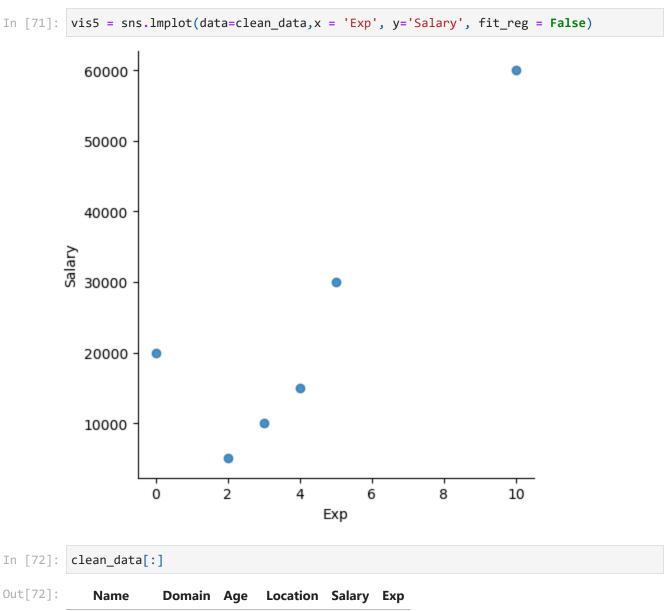
## **EDA Technique lets Apply**

```
In [59]:
         import matplotlib.pyplot as plt
         import seaborn as sns
In [60]:
         import warnings
         warnings.filterwarnings('ignore')
         clean_data['Salary']
In [62]:
Out[62]:
                5000
          1
               10000
               15000
               20000
               30000
               60000
         Name: Salary, dtype: int64
In [64]:
         vis1=sns.distplot(clean_data['Salary'])
                1e-5
           3.5
           3.0
           2.5
           2.0
           1.5
           1.0
           0.5
           0.0
              -40000 -20000
                                   0
                                          20000
                                                   40000
                                                             60000
                                                                      80000
                                                                               100000
                                                Salary
         vis2=plt.hist(clean_data['Salary'])
```



In [70]: vis4=sns.lmplot(data=clean\_data,x='Exp',y='Salary')





In [72]:	CI	ean_dat	a[:]				
Out[72]:		Name	Domain	Age	Location	Salary	Ехр
	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
              Umar
                     Dataanalyst
                                  50
                                      Bangalore
                                                 15000
                                                          4
            Uttam
                       Statistics
                                  67
                                      Bangalore
                                                 30000
                                                           5
         clean_data[::-1]
In [74]:
Out[74]:
                                       Location
             Name
                        Domain Age
                                                Salary
                                                        Exp
          5
                           NLP
                                          Delhi
                                                 60000
               Kim
                                  55
                                                         10
                                      Bangalore
          4 Uttam
                       Statistics
                                                 30000
                                                           5
                                  67
                                      Hyderbad
                                                 20000
          3
               Jane
                       Analytics
                                                          0
                                  50
          2
              Umar
                     Dataanalyst
                                      Bangalore
                                                 15000
                                                          4
                                  50
             Teddy
                         Testing
                                      Bangalore
                                                 10000
                                                           3
                                  45
              Mike
                    Datascience
                                        Mumbai
                                                  5000
                                                           2
                                  34
In [75]:
         clean_data.columns
Out[75]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
         X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [76]:
In [77]: X_iv
Out[77]:
             Name
                        Domain Age
                                       Location Exp
          0
              Mike Datascience
                                                   2
                                  34
                                        Mumbai
             Teddy
                                      Bangalore
                         Testing
                                                   3
                     Dataanalyst
          2
              Umar
                                  50
                                      Bangalore
                                                   4
          3
               Jane
                       Analytics
                                  50
                                      Hyderbad
                                                   0
             Uttam
                       Statistics
                                      Bangalore
                                                   5
          4
                                  67
          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]:

Out[80]:

	Name	Domain	Age	Location	Salary	Ехр
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	Ехр
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	۸۵۵	Location	Eve	
040[02].	_	Mike		34	Mumbai		
	0	Teddy	Datascience Testing	45	Bangalore	3	
	2	Umar	Dataanalyst	50	Bangalore	4	
	3	Jane	•	50	_		
			Analytics		Hyderbad	0	
	4	Uttam	Statistics	67	Bangalore	5	
	5	Kim	NLP	55	Delhi	10	
In [83]:	у_	dv					
Out[83]:							
24 [03].	_	Salary	-				
	0	5000					
	1	10000					
	2	15000					
	3	20000					
	4	30000					
	5	60000					
In [84]:	<b>c</b> 1	ean_dat	a				
	-				1	6.1	
Out[84]:	_	Name	Domain		Location		_
	0		Datascience		Mumbai	5000	
	1	Teddy	Testing	45	Bangalore	10000	
	2	Umar	Dataanalyst	50	Bangalore	15000	
	3	Jane	Analytics	50	Hyderbad	20000	
	4	Uttam	Statistics	67	Bangalore	30000	
	5	Kim	NLP	55	Delhi	60000	
					, -		
In [88]:	im	putatio	n = pd.get_d	dummie	es(clean_da	ata)	

imputation

In [86]:

Out[86]:		Age	Salary	Ехр	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	
	4									•
In [87]:	cl	ean_da	ata							

Out[87]:

	Name	Domain	Age	Location	Salary	Ехр
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