```
import pandas as pd
In [1]:
In [2]: emp = pd.read_excel(r'C:\Users\rohit\Downloads\Rawdata.xlsx')
In [3]:
        emp
Out[3]:
             Name
                          Domain
                                      Age
                                            Location
                                                         Salary
                                                                   Exp
         0
              Mike
                     Datascience#$ 34 years
                                                        5^00#0
                                                                    2+
                                             Mumbai
           Teddy^
                                     45' yr Bangalore
                           Testing
                                                      10%%000
                                                                    <3
         2
            Uma#r
                    Dataanalyst^^#
                                                NaN
                                      NaN
                                                       1$5%000
                                                                 4> yrs
         3
                       Ana^^lytics
                                      NaN Hyderbad
              Jane
                                                        2000^0
                                                                   NaN
         4
            Uttam*
                          Statistics
                                      67-yr
                                                NaN
                                                        30000-
                                                                5+ year
         5
                              NLP
               Kim
                                      55yr
                                                Delhi
                                                       6000^$0
                                                                   10+
In [4]:
        id(emp)
Out[4]: 1506001507856
In [5]:
        emp.columns
Out[5]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [6]: emp.info()
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
            Column
                      Non-Null Count Dtype
                      6 non-null
           Name
                                       object
        0
        1
            Domain
                      6 non-null
                                       object
        2
            Age
                      4 non-null
                                       object
        3
            Location 4 non-null
                                       object
            Salary
                      6 non-null
                                       object
                      5 non-null
                                       object
            Exp
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [7]:
        emp.shape
Out[7]: (6, 6)
In [8]: emp.head()
```

24, 17:16					Raw_data	to Clean_c	data co	nversion us	ing python l	ED.
Out[8]:		Name	Dom	ain	Age	e Loca	tion	Salar	у Ех	ιp
	0	Mike	Datascienc	e#\$	34 years	s Mur	nbai	5^00#	0 2	<u>?</u> +
	1	Teddy^	Tes	ting	45' y	r Banga	lore	10%%00	0 <	<3
	2	Uma#r	Dataanalyst ⁷	^#	NaN	J I	NaN	1\$5%00	0 4> y	rs
	3	Jane	Ana^^ly	tics	NaN	l Hydei	bad	2000^	0 Na	Ν
	4	Uttam*	Statis	tics	67-y	r l	NaN	30000)- 5+ yea	ar
In [9]:	em	p.tail()							
Out[9]:		Name	Dom	ain	Age	Locatio	n	Salary	Ехр	
	1	Teddy^	Tes	ting	45' yr	Bangalor	e 10	0%%000	<3	
	2	Uma#r	Dataanalyst /	^#	NaN	Na	N 1	\$5%000	4> yrs	
	3	Jane	Ana^^ly	tics	NaN	Hyderba	d	2000^0	NaN	
	4	Uttam*	Statis	tics	67-yr	Na	N	30000-	5+ year	
	5	Kim	1	NLP	55yr	Dell	ni 6	5000^\$0	10+	
In [10]:	em	p.isnul	1()							
Out[10]:		Name	Domain Ag	je	Location	Salary	Ex	р		
	0	False	False Fal	se	False	False	Fals	se		
	1	False	False Fal	se	False	False	Fals	se		
	_		- I -		-					

	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 [11]: emp.isna()

Out[11]:

	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 [12]: emp.isnull().sum()

```
Out[12]: Name 0
Domain 0
Age 2
Location 2
Salary 0
Exp 1
dtype: int64
```

Data cleaning or Data Cleansing

```
In [14]:
         emp['Name']
Out[14]:
          0
                  Mike
               Teddy^
          1
          2
                 Uma#r
          3
                  Jane
          4
               Uttam*
          5
                   Kim
          Name: Name, dtype: object
          emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True) # non word charactere
In [15]:
In [16]:
          emp['Name']
Out[16]:
                 Mike
          1
               Teddy
          2
                 Umar
          3
                 Jane
          4
               Uttam
          5
                  Kim
          Name: Name, dtype: object
In [17]:
Out[17]:
             Name
                           Domain
                                               Location
                                        Age
                                                           Salary
                                                                      Exp
          0
              Mike
                      Datascience#$
                                                          5^00#0
                                                                       2+
                                    34 years
                                               Mumbai
                                              Bangalore
              Teddy
                            Testing
                                       45' yr
                                                        10%%000
                                                                       <3
                     Dataanalyst^^#
          2
              Umar
                                        NaN
                                                  NaN
                                                         1$5%000
                                                                    4> yrs
          3
                        Ana^^lytics
                                              Hyderbad
                                                          2000^0
               Jane
                                        NaN
                                                                     NaN
                           Statistics
             Uttam
                                       67-yr
                                                  NaN
                                                           30000-
                                                                   5+ year
                               NLP
                                        55yr
                                                         6000^$0
               Kim
                                                  Delhi
                                                                      10+
In [18]:
          emp['Domain']
Out[18]:
          0
                 Datascience#$
          1
                       Testing
          2
               Dataanalyst^^#
                   Ana^^lytics
          3
          4
                    Statistics
          Name: Domain, dtype: object
```

```
emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
In [19]:
In [20]:
         emp['Domain']
Out[20]: 0
               Datascience
          1
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
          5
                       NLP
          Name: Domain, dtype: object
In [21]:
          emp
Out[21]:
             Name
                       Domain
                                   Age
                                          Location
                                                      Salary
                                                                 Exp
              Mike
                                                     5^00#0
          0
                    Datascience
                                34 years
                                          Mumbai
                                                                  2+
             Teddy
                                  45' yr
                                         Bangalore
                                                   10%%000
                                                                  <3
          1
                        Testing
          2
             Umar
                    Dataanalyst
                                              NaN
                                                    1$5%000
                                                               4> yrs
                                   NaN
                                         Hyderbad
          3
              Jane
                       Analytics
                                   NaN
                                                     2000^0
                                                                NaN
          4
             Uttam
                       Statistics
                                   67-yr
                                              NaN
                                                      30000-
                                                             5+ year
          5
                           NLP
                                   55yr
                                             Delhi
                                                    6000^$0
                                                                 10+
               Kim
In [22]: emp['Age']
Out[22]: 0
               34 years
                 45' yr
          2
                    NaN
          3
                    NaN
          4
                  67-yr
          5
                   55yr
          Name: Age, dtype: object
         emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
In [23]:
In [24]:
         emp['Age']
Out[24]: 0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
                  55yr
          Name: Age, dtype: object
In [25]: emp['Age'] = emp['Age'].str.extract('(\d+)')
        <>:1: SyntaxWarning: invalid escape sequence '\d'
        <>:1: SyntaxWarning: invalid escape sequence '\d'
        C:\Users\rohit\AppData\Local\Temp\ipykernel_14268\1884116463.py:1: SyntaxWarning:
        invalid escape sequence '\d'
          emp['Age'] = emp['Age'].str.extract('(\d+)')
In [26]: emp['Age']
```

```
Out[26]:
                34
                45
          1
          2
               NaN
          3
               NaN
                67
          4
          5
                55
          Name: Age, dtype: object
In [27]:
                                                   Salary
Out[27]:
             Name
                       Domain
                                 Age
                                       Location
                                                              Exp
              Mike
                                                   5^00#0
          0
                    Datascience
                                  34
                                        Mumbai
                                                               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
                           NLP
          5
               Kim
                                  55
                                           Delhi
                                                  6000^$0
                                                              10+
In [28]:
          emp['Location']
Out[28]:
          0
                  Mumbai
          1
               Bangalore
          2
                      NaN
          3
                Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
          emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
In [29]:
In [30]:
          emp['Location']
Out[30]:
          0
                  Mumbai
          1
               Bangalore
          2
                      NaN
          3
                Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
In [31]:
          emp
```

```
Out[31]:
             Name
                       Domain Age
                                      Location
                                                  Salary
                                                             Exp
          0
              Mike Datascience
                                 34
                                       Mumbai
                                                  5^00#0
                                                              2+
             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
            Uttam
                       Statistics
                                          NaN
                                                  30000- 5+ year
                                 67
          5
               Kim
                          NLP
                                  55
                                          Delhi
                                                 6000^$0
                                                             10+
         emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
In [32]:
In [33]:
         emp['Salary']
Out[33]: 0
                5000
               10000
          1
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: object
In [34]:
         emp['Exp']
                    2+
Out[34]:
          0
          1
                    <3
          2
                4> yrs
          3
                   NaN
               5+ year
                   10+
          Name: Exp, dtype: object
In [35]:
         emp['Exp'] = emp['Exp'].str.replace(r'\W','',regex=True)
         emp['Exp']
In [36]:
                   2
Out[36]:
          0
          1
                   3
          2
                4yrs
          3
                 NaN
               5year
          4
                  10
          Name: Exp, dtype: object
         emp['Exp'] = emp['Exp'].str.extract('(\d+)')
In [37]:
        <>:1: SyntaxWarning: invalid escape sequence '\d'
        <>:1: SyntaxWarning: invalid escape sequence '\d'
        C:\Users\rohit\AppData\Local\Temp\ipykernel_14268\3836251810.py:1: SyntaxWarning:
        invalid escape sequence '\d'
          emp['Exp'] = emp['Exp'].str.extract('(\d+)')
In [38]: emp['Exp']
```

Out[38]: 0 2 1 3 2 4 3 NaN 4 5 5 10 Name: Exp, dtype: object

In [39]: emp

Out[39]:

	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 [40]: clean_data = emp.copy()

In [41]: clean_data

Out[41]:

	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

EDA TECHNIQUE

In [43]: clean_data

Out[43]:

	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

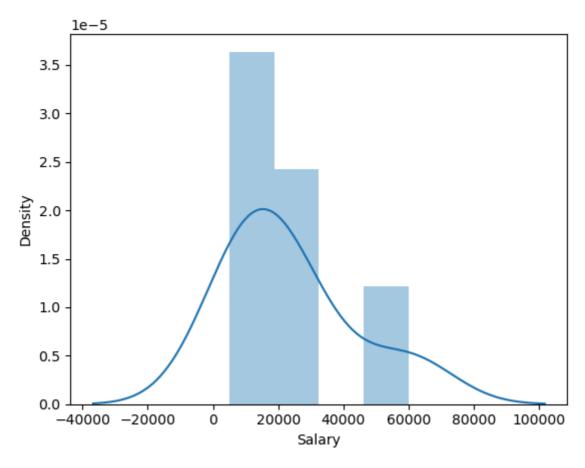
```
In [44]:
                                                             clean_data.isnull().sum()
Out[44]: Name
                                                                                                                                                    0
                                                                   Domain
                                                                                                                                                    0
                                                                                                                                                   2
                                                                   Age
                                                                   Location
                                                                                                                                                   2
                                                                   Salary
                                                                                                                                                   0
                                                                   Exp
                                                                   dtype: int64
In [45]: clean_data['Age']
Out[45]: 0
                                                                                                           34
                                                                   1
                                                                                                           45
                                                                   2
                                                                                                    NaN
                                                                   3
                                                                                                     NaN
                                                                                                           67
                                                                   4
                                                                                                           55
                                                                   Name: Age, dtype: object
In [46]: import numpy as np
In [47]: clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(clean_data['Age'].fillna(np.to_numeric(c
In [48]: clean_data['Age']
Out[48]: 0
                                                                                                                          34
                                                                                                                         45
                                                                   2
                                                                                                   50.25
                                                                                                    50.25
                                                                   3
                                                                                                                         67
                                                                   4
                                                                                                                          55
                                                                   Name: Age, dtype: object
In [49]: clean_data['Exp']
Out[49]:
                                                                  0
                                                                                                                   2
                                                                   1
                                                                                                                   3
                                                                                                                   4
                                                                   2
                                                                   3
                                                                                                     NaN
                                                                   4
                                                                                                                   5
                                                                                                            10
                                                                   Name: Exp, dtype: object
In [50]: clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp'])).fillna(np
In [51]: clean_data['Exp']
Out[51]: 0
                                                                                                                   2
                                                                   1
                                                                                                                   3
                                                                   2
                                                                                                                   4
                                                                   3
                                                                                                     4.8
                                                                   4
                                                                                                                   5
                                                                                                            10
                                                                   Name: Exp, dtype: object
In [52]: clean_data
```

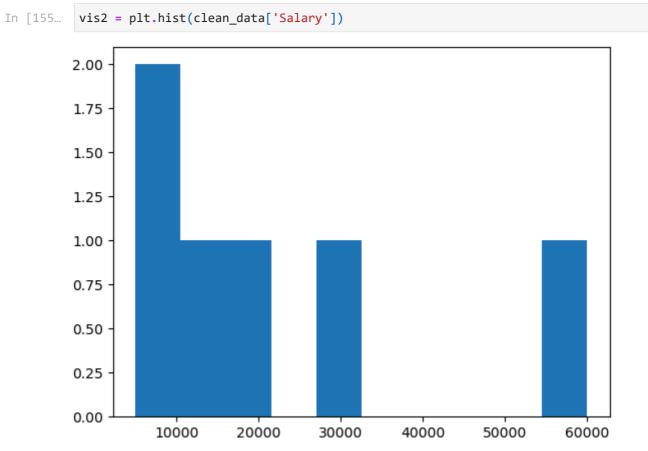
```
Out[52]:
              Name
                        Domain
                                  Age
                                         Location Salary Exp
          0
              Mike Datascience
                                    34
                                          Mumbai
                                                    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
          clean_data['Location'].isnull().sum()
In [53]:
Out[53]: 2
In [54]:
          clean_data['Location']
          0
Out[54]:
                   Mumbai
          1
                Bangalore
          2
                      NaN
          3
                 Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mc
In [55]:
          clean_data['Location']
In [56]:
Out[56]:
          0
                   Mumbai
          1
                Bangalore
          2
                Bangalore
          3
                 Hyderbad
          4
                Bangalore
          5
                    Delhi
          Name: Location, dtype: object
          clean_data
In [57]:
Out[57]:
              Name
                        Domain
                                         Location Salary Exp
                                  Age
          0
              Mike
                     Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
              Teddy
                                    45
                                        Bangalore
                                                   10000
                                                             3
                         Testing
          2
              Umar
                     Dataanalyst
                                 50.25
                                        Bangalore
                                                   15000
                                                             4
          3
               Jane
                                 50.25
                                        Hyderbad
                                                   20000
                        Analytics
                                                           4.8
             Uttam
                                        Bangalore
                                                             5
          4
                        Statistics
                                    67
                                                   30000
          5
                Kim
                            NLP
                                    55
                                                   60000
                                                            10
                                            Delhi
In [58]:
          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
          Salary 6 non-null
                                   object
        5
           Exp
                   6 non-null
                                   object
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [59]:
        clean_data['Age'] = clean_data['Age'].astype(int)
In [60]: 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 int32
3 Location 6 non-null object
        4 Salary 6 non-null
                                   object
                   6 non-null
        5
            Exp
                                   object
       dtypes: int32(1), object(5)
       memory usage: 396.0+ bytes
In [61]: clean_data['Salary'] = clean_data['Salary'].astype(int)
In [62]: 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
                                 obiect
           Age 6 non-null
Location 6 non-null
        2 Age
                                   int32
        3
                                   object
        4 Salary 6 non-null
                                   int32
           Exp
        5
                   6 non-null
                                   object
       dtypes: int32(2), object(4)
       memory usage: 372.0+ bytes
In [63]: clean data['Exp'] = clean data['Exp'].astype(int)
In [64]: clean data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
            Column
                    Non-Null Count Dtype
        --- -----
                      -----
                      6 non-null
        0
            Name
                                     object
           Domain 6 non-null
        1
                                     object
        2 Age
                     6 non-null
                                     int32
           Location 6 non-null
        3
                                      object
        4
            Salary
                      6 non-null
                                      int32
        5
                      6 non-null
                                      int32
            Exp
        dtypes: int32(3), object(3)
        memory usage: 348.0+ bytes
         clean_data['Name'] = clean_data['Name'].astype('category')
In [65]:
         clean_data['Domain'] = clean_data['Domain'].astype('category')
         clean_data['Location'] = clean_data['Location'].astype('category')
In [66]: clean_data.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
            Column
                    Non-Null Count Dtype
                      _____
        ---
                     6 non-null
        0
            Name
                                     category
        1
            Domain 6 non-null
                                     category
        2
            Age
                    6 non-null
                                      int32
        3
            Location 6 non-null
                                      category
        4
                     6 non-null
                                      int32
            Salary
                      6 non-null
                                      int32
        dtypes: category(3), int32(3)
        memory usage: 866.0 bytes
         clean_data
In [67]:
Out[67]:
            Name
                     Domain Age
                                   Location Salary Exp
         0
             Mike Datascience
                               34
                                    Mumbai
                                              5000
                                                     2
            Teddy
                      Testing
                               45
                                   Bangalore
                                             10000
                                                     3
         2
            Umar
                   Dataanalyst
                               50
                                   Bangalore
                                             15000
                                                     4
         3
             Jane
                     Analytics
                                   Hyderbad
                                             20000
           Uttam
                     Statistics
                               67
                                   Bangalore
                                             30000
                                                     5
                         NLP
         5
              Kim
                               55
                                       Delhi
                                            60000
                                                    10
In [68]:
         clean data.to csv('clean data.csv')
In [69]:
         import os
         os.getcwd()
Out[69]: 'C:\\Users\\rohit'
In [70]:
         clean data
```

Out[70]:		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	4
	4	Uttam	Statistics	67	Bangalore	30000	5
	5	Kim	NLP	55	Delhi	60000	10
<pre>In [71]: In [72]: In [73]: In [74]: Out[74]:</pre>	im wa cl 0 1 2 3 4 5	port se port wa rnings. ean_dat 500 1000 1500 2000 3000 6000	filterwarnir a['Salary'] 0 0 0	s ngs('i	ignore')		
In [153	vi	s1 = sn	s.distplot(:lean_	_data['Sala	ary'])	

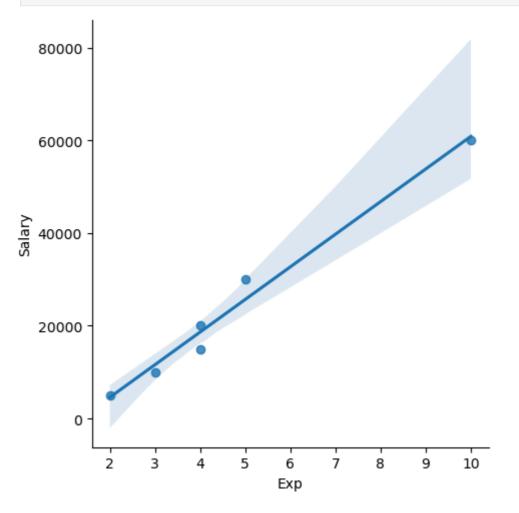


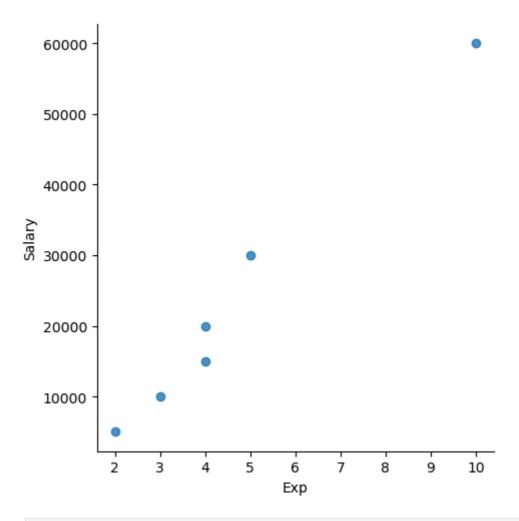


clean_data

In [77]:

Out[77]:		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	4
	4	Uttam	Statistics	67	Bangalore	30000	5
	5	Kim	NLP	55	Delhi	60000	10





In [161... clean_data[:]

\cap		-1	\sim	1	
υu	니	щ	O	Т	

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

In [163... clean_data[0:6:2]

Out[163...

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

In [165...

clean_data[::-1]

5		Name	Domain	Age	Location	Salary	Ехр
	5	Kim	NLP	55	Delhi	60000	10
	4	Uttam	Statistics	67	Bangalore	30000	5
	3	Jane	Analytics	50	Hyderbad	20000	4
	2	Umar	Dataanalyst	50	Bangalore	15000	4
	1	Teddy	Testing	45	Bangalore	10000	3
	0	Mike	Datascience	34	Mumbai	5000	2
L67	cl	ean_dat	a.columns				
167	In	idex(['N	lame', 'Doma	in',	'Age', 'Lo	cation'	, 'Sa
174	x_	iv = cl	ean_data[['I	Name',	, 'Domain',	'Age'	, 'Loc
176	Х	iv					
176	_	Name	Domain	Δαρ	Location	Evn	
	0	Mike	Datascience	Age 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	
178	у_	dv = cl	ean_data[['	Salary	/']]		
180	у_	dv					
180	7 —	Salary					
	0	5000	_				
	1	10000					
	2						
		15000					
	3	20000					
	4	30000					
	5	60000					
[182	em	D					
L	5						

O		Г - п	0	\neg	
υu	ı	1	ŏ	Z	

	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 [184...

clean_data

Out[184...

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

In [186... X_iv

Out[186...

	Name	Domain	Age	Location	Ехр
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 [188... y_dv

Out[188... Salary

0 5000

1 10000

2 15000

3 20000

4 30000

5 60000

In [190...

clean_data

Out[190...

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

In [202... imputation = pd.get_dummies(clean_data, dtype=int)

In [204...

imputation

Out[204...

	Age	Salary	Ехр	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar
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
4								•

In [198...

clean_data

Out[198		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	4
	4	Uttam	Statistics	67	Bangalore	30000	5
	5	Kim	NLP	55	Delhi	60000	10

In [206...

imputation

Out[206...

1 4 2 5 3 5	4 5000 5 10000 0 15000 0 20000	3 4	0 0 0	0 0 0	1 0 0	0 1 0	0 0 1
2 5 3 5	0 15000	4				1 0	0
3 5			0	0	0	0	1
	0 20000						
	0 20000	4	1	0	0	0	0
4 6	7 30000	5	0	0	0	0	0
5 5	5 60000	10	0	1	0	0	0
4							>

In []: