```
import numpy as np
In [70]:
           Rawdata= pd.read_excel('Rawdata.xlsx')
           Rawdata
Out[70]:
                Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                     Exp
            0
                      Datascience#$
                                     34 years
                                                          5^00#0
                                                                      2+
                Mike
                                                Mumbai
               Teddy^
                                       45' yr Bangalore
                                                        10%%000
                             Testing
                                                                      <3
               Uma#r Dataanalyst^^#
                                                         1$5%000
                                        NaN
                                                  NaN
                                                                   4> yrs
            3
                          Ana^^lytics
                                        NaN
                                              Hyderbad
                                                          2000^0
                                                                     NaN
                Jane
               Uttam*
                           Statistics
                                        67-yr
                                                  NaN
                                                           30000- 5+ year
            5
                               NLP
                                                  Delhi
                                                         6000^$0
                                                                     10+
                 Kim
                                        55yr
In [71]:
           emp = Rawdata.copy()
           emp
Out[71]:
                Name
                            Domain
                                              Location
                                        Age
                                                           Salary
                                                                     Exp
            0
                Mike
                      Datascience#$
                                     34 years
                                                          5^00#0
                                                                      2+
                                                Mumbai
               Teddy<sup>^</sup>
                                       45' yr Bangalore
                                                        10%%000
                                                                      <3
                             Testing
               Uma#r Dataanalyst^^#
                                        NaN
                                                  NaN
                                                        1$5%000
                                                                   4> yrs
            3
                          Ana^^lytics
                                                          2000^0
                Jane
                                        NaN
                                              Hyderbad
                                                                     NaN
               Uttam*
                           Statistics
                                                  NaN
                                                          30000- 5+ year
            4
                                        67-yr
            5
                 Kim
                               NLP
                                        55yr
                                                  Delhi
                                                         6000^$0
                                                                     10+
In [72]:
           id(emp)
Out[72]: 2150525672912
In [73]:
          emp.columns
Out[73]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='obje
           ct')
In [74]:
           emp.shape
Out[74]: (6, 6)
```

In [69]:

import pandas as pd

```
In [75]:
           emp.head(7)
Out[75]:
                Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                     Exp
            0
                Mike
                      Datascience#$
                                     34 years
                                                Mumbai
                                                          5^00#0
                                                                      2+
               Teddy^
                                       45' yr Bangalore
                                                        10%%000
                                                                      <3
                             Testing
               Uma#r Dataanalyst^^#
                                        NaN
                                                  NaN
                                                         1$5%000
                                                                   4> yrs
            3
                Jane
                          Ana^^lytics
                                        NaN
                                              Hyderbad
                                                          2000^0
                                                                     NaN
               Uttam*
                           Statistics
                                        67-yr
                                                  NaN
                                                           30000- 5+ year
            5
                 Kim
                               NLP
                                        55yr
                                                  Delhi
                                                         6000^$0
                                                                     10+
In [76]:
           emp.tail(7)
Out[76]:
                Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                     Exp
            0
                Mike
                      Datascience#$
                                    34 years
                                                Mumbai
                                                          5^00#0
                                                                      2+
               Teddy^
                                       45' yr Bangalore
                                                        10%%000
                                                                      <3
            1
                             Testing
                      Dataanalyst^^#
            2
               Uma#r
                                        NaN
                                                  NaN
                                                         1$5%000
                                                                   4> yrs
            3
                Jane
                          Ana^^lytics
                                        NaN
                                              Hyderbad
                                                          2000^0
                                                                     NaN
               Uttam*
                           Statistics
                                                           30000-
            4
                                        67-yr
                                                  NaN
                                                                  5+ year
            5
                 Kim
                               NLP
                                        55yr
                                                  Delhi
                                                         6000^$0
                                                                     10+
In [77]:
           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
                             4 non-null
                                                object
                 Age
            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 [78]:
           emp.isnull()
Out[78]:
               Name
                     Domain
                                    Location Salary
                               Age
                                                       Exp
            0
               False
                        False
                              False
                                        False
                                               False
                                                     False
               False
            1
                        False
                              False
                                        False
                                               False
                                                     False
               False
            2
                        False
                               True
                                         True
                                               False
                                                     False
```

5

False

False

False

True

False

False

True

False

False

False False

False False

True

False

False

False False

```
In [79]:
           emp.isnull().sum()
Out[79]:
           Name
                          0
           Domain
                          0
                          2
           Age
           Location
                          2
           Salary
                          0
           Exp
                          1
           dtype: int64
           Data cleaning
In [80]:
           emp
Out[80]:
                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
                          Ana^^lytics
                                                Hyderbad
                                                             2000^0
                 Jane
                                         NaN
                                                                       NaN
               Uttam*
                            Statistics
                                         67-yr
                                                    NaN
                                                             30000- 5+ year
            5
                  Kim
                                NLP
                                          55yr
                                                    Delhi
                                                           6000^$0
                                                                        10+
           emp['Name']
In [81]:
Out[81]:
           0
                    Mike
           1
                 Teddy^
           2
                   Uma#r
           3
                    Jane
           4
                 Uttam*
           5
                     Kim
           Name: Name, dtype: object
In [82]:
           emp['Name'] = emp['Name'].str.replace(r'\W','',regex = True)
           emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex = True)
emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex = True)
           emp['Exp'] = emp['Exp'].str.replace(r'\W','',regex = True)
Out[82]:
               Name
                          Domain
                                       Age
                                             Location Salary
                                                                Exp
                                  34 years
            0
                Mike
                      Datascience
                                              Mumbai
                                                        5000
                                                                  2
            1
               Teddy
                           Testing
                                      45' yr
                                            Bangalore
                                                       10000
                                                                  3
                                                       15000
            2
               Umar
                       Dataanalyst
                                      NaN
                                                 NaN
                                                                4yrs
                                                       20000
            3
                Jane
                         Analytics
                                      NaN
                                            Hyderbad
                                                               NaN
               Uttam
                         Statistics
                                      67-yr
                                                 NaN
                                                       30000
                                                              5year
```

Kim

NLP

55yr

Delhi

60000

10

```
In [83]: emp['Age'] = emp['Age'].str.extract('(\d+)')
emp['Exp'] = emp['Exp'].str.extract('(\d+)')
emp
```

Out[83]:

	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 [84]: clean_data = emp.copy()
    clean_data
```

Out[84]:

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

Till now we have raw data we use

EDA

```
In [85]: | clean_data.isnull().sum()
Out[85]: Name
                      0
         Domain
                      0
         Age
                      2
                      2
         Location
         Salary
                      0
                      1
         Exp
         dtype: int64
In [86]: clean_data['Age']
Out[86]: 0
                34
         1
               45
         2
              NaN
         3
              NaN
               67
               55
         Name: Age, dtype: object
```

```
In [87]:
          clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data))
          clean_data
Out[87]:
             Name
                      Domain
                               Age
                                    Location
                                             Salary Exp
          0
              Mike
                   Datascience
                                34
                                      Mumbai
                                              5000
                                                      2
             Teddy
                                45
                                    Bangalore
                                             10000
                                                      3
                       Testing
             Umar
                    Dataanalyst 50.25
                                             15000
                                                      4
                                        NaN
           3
              Jane
                      Analytics 50.25
                                    Hyderbad
                                             20000
                                                    4.8
             Uttam
                      Statistics
                                67
                                        NaN
                                             30000
                                                      5
           4
           5
               Kim
                         NLP
                                55
                                        Delhi
                                             60000
                                                     10
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location')]
In [88]:
          clean_data['Location']
Out[88]:
         0
                  Mumbai
          1
               Bangalore
          2
               Bangalore
          3
                Hyderbad
          4
               Bangalore
          5
                   Delhi
          Name: Location, dtype: object
In [89]:
          clean_data
Out[89]:
             Name
                      Domain
                               Age
                                    Location
                                             Salary Exp
          0
              Mike
                   Datascience
                                34
                                              5000
                                                      2
                                      Mumbai
             Teddy
                                45
                                    Bangalore
                                             10000
                       Testing
                                                      3
           2
             Umar
                    Dataanalyst 50.25
                                    Bangalore
                                             15000
                                                      4
           3
              Jane
                      Analytics
                              50.25
                                    Hyderbad
                                             20000
                                                    4.8
                                             30000
           4
             Uttam
                      Statistics
                                67
                                    Bangalore
                                                      5
                         NLP
                                             60000
           5
                                55
                                        Delhi
                                                     10
               Kim
In [90]:
          cleaned emp = clean data.copy()
          cleaned_emp
Out[90]:
```

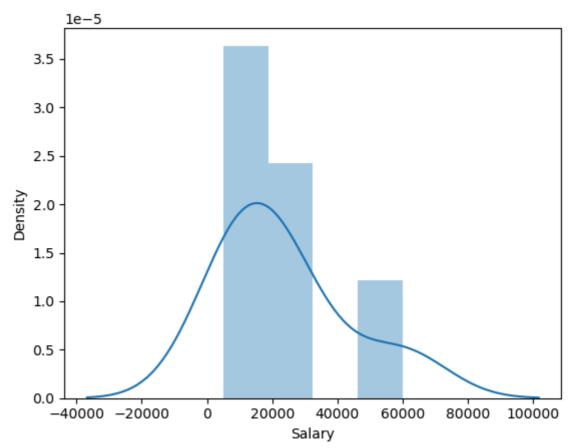
	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	Bangalore	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

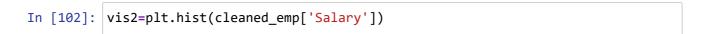
```
In [91]: cleaned_emp.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          #
              Column
                         Non-Null Count Dtype
          0
                         6 non-null
                                         object
              Name
          1
              Domain
                         6 non-null
                                         object
                         6 non-null
          2
              Age
                                         object
          3
              Location 6 non-null
                                         object
          4
              Salary
                         6 non-null
                                         object
                         6 non-null
          5
              Exp
                                         object
         dtypes: object(6)
         memory usage: 420.0+ bytes
In [92]:
         cleaned_emp.describe()
Out[92]:
                 Name
                          Domain
                                   Age
                                        Location Salary Exp
           count
                     6
                                  6.00
                                                         6
          unique
                     6
                                  5.00
                                              4
                                                    6
                                                         6
                                                         2
             top
                  Mike Datascience 50.25 Bangalore
                                                  5000
            freq
                     1
                                  2.00
                                              3
                                                    1
                                                         1
In [93]:
         cleaned_emp['Age'] = cleaned_emp['Age'].astype(int)
         cleaned_emp['Salary'] = cleaned_emp['Salary'].astype(int)
         cleaned_emp['Exp'] = cleaned_emp['Exp'].astype(int)
         cleaned_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
                         6 non-null
                                         int32
              Age
          3
              Location 6 non-null
                                         object
          4
                         6 non-null
                                         int32
              Salary
          5
                         6 non-null
                                         int32
         dtypes: int32(3), object(3)
         memory usage: 348.0+ bytes
```

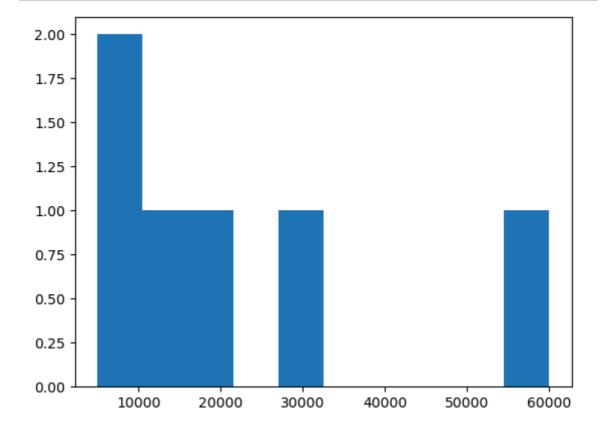
```
In [94]:
          cleaned_emp.describe()
Out[94]:
                      Age
                                 Salary
                                             Exp
                  6.000000
                               6.000000
                                         6.000000
           count
                 50.166667
                           23333.333333
                                         4.666667
           mean
                 10.907184
                           19916.492328
                                         2.804758
             std
                 34.000000
                            5000.000000
                                         2.000000
            min
            25%
                 46.250000
                           11250.000000
                                         3.250000
            50%
                 50.000000
                          17500.000000
                                         4.000000
            75%
                 53.750000
                           27500.000000
                                         4.750000
                67.000000 60000.000000
                                        10.000000
In [95]:
          cleaned_emp['Name'] = cleaned_emp['Name'].astype('category')
          cleaned_emp['Domain'] = cleaned_emp['Domain'].astype('category')
          cleaned_emp['Location'] = cleaned_emp['Location'].astype('category')
          cleaned_emp
Out[95]:
              Name
                                     Location Salary Exp
                       Domain Age
           0
              Mike
                    Datascience
                                 34
                                      Mumbai
                                               5000
                                                       2
             Teddy
                        Testing
                                 45
                                    Bangalore
                                              10000
                                                       3
           2
              Umar
                    Dataanalyst
                                 50
                                    Bangalore
                                              15000
                                                       4
           3
              Jane
                       Analytics
                                 50
                                     Hyderbad
                                              20000
                                                       4
             Uttam
                       Statistics
                                    Bangalore
                                              30000
                                                       5
           5
                          NLP
                                 55
                                        Delhi
                                              60000
                                                      10
               Kim
In [96]:
          cleaned emp.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
                                            category
           1
               Domain
                          6 non-null
                                            category
           2
               Age
                          6 non-null
                                            int32
           3
               Location 6 non-null
                                            category
               Salary
           4
                           6 non-null
                                            int32
           5
                           6 non-null
                                            int32
          dtypes: category(3), int32(3)
          memory usage: 866.0 bytes
In [97]:
          cleaned_emp.to_excel('cleaned_emp.xlsx')
In [98]:
          import os
          os.getcwd()
Out[98]: 'C:\\Users\\Sandeep\\OneDrive\\Desktop\\Coching\\Institute\\Projects'
```

EDA Techniques

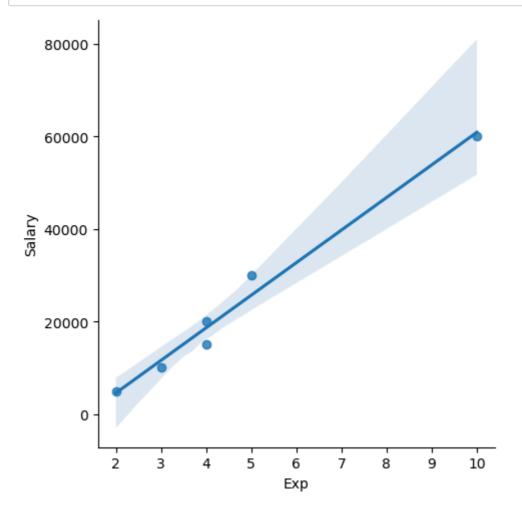
```
In [99]:
          import matplotlib.pyplot as plt
          import seaborn as sns
          import warnings
          warnings.filterwarnings('ignore')
In [100]: cleaned_emp['Salary']
Out[100]:
                5000
               10000
          2
               15000
          3
               20000
               30000
          4
               60000
          Name: Salary, dtype: int32
In [101]: vis1=sns.distplot(cleaned_emp['Salary'])
```



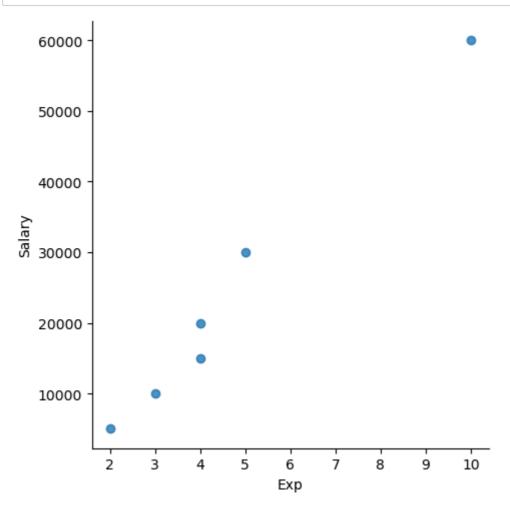




In [103]: vis3=sns.lmplot(data=cleaned_emp, x ='Exp', y='Salary')



In [104]: vis4=sns.lmplot(data=cleaned_emp,x='Exp',y='Salary',fit_reg=False)



In [105]: cleaned_emp[:]

Out	1105	
out	[TOD]	

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

In [106]: cleaned_emp[0:6:2]

Out[106]:

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

```
In [107]: cleaned_emp[::-1]
Out[107]:
               Name
                         Domain Age
                                       Location Salary Exp
            5
                 Kim
                            NLP
                                   55
                                           Delhi
                                                 60000
                                                         10
               Uttam
                         Statistics
                                   67
                                       Bangalore
                                                 30000
                                                          5
            3
                Jane
                        Analytics
                                   50
                                       Hyderbad
                                                 20000
                                                          4
            2
               Umar
                      Dataanalyst
                                       Bangalore
                                                 15000
                                                          4
               Teddy
                          Testing
                                   45
                                       Bangalore
                                                 10000
                Mike Datascience
                                                  5000
                                                          2
                                   34
                                        Mumbai
In [108]: | cleaned_emp.columns
Out[108]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='obje
            ct')
In [109]:
           x_iv = cleaned_emp[['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp']]
           x_iv
Out[109]:
               Name
                         Domain Age
                                       Location Salary Exp
            0
                Mike
                      Datascience
                                   34
                                        Mumbai
                                                  5000
                                                          2
               Teddy
                          Testing
                                   45
                                       Bangalore
                                                 10000
                                                          3
                                                 15000
               Umar
                      Dataanalyst
                                   50
                                       Bangalore
                                                          4
                Jane
                        Analytics
                                       Hyderbad
                                                 20000
                                                          4
                                                 30000
               Uttam
                         Statistics
                                   67
                                       Bangalore
                                                          5
                            NLP
                                   55
                                                 60000
                                                         10
                 Kim
                                           Delhi
In [110]:
           y_dv=cleaned_emp[['Salary']]
           y_dv
Out[110]:
               Salary
               5000
               10000
               15000
               20000
            3
```

```
In [111]:
             Rawdata
Out[111]:
                  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
                 Uttam*
                                                                30000- 5+ year
              4
                               Statistics
                                            67-yr
                                                       NaN
              5
                                   NLP
                    Kim
                                            55yr
                                                       Delhi
                                                               6000^$0
                                                                            10+
In [112]:
             emp
Out[112]:
                 Name
                            Domain
                                     Age
                                            Location
                                                      Salary
                                                              Exp
              0
                  Mike
                                                        5000
                        Datascience
                                       34
                                             Mumbai
                                                                 2
                                           Bangalore
                 Teddy
                                                       10000
                                                                 3
              1
                             Testing
                                       45
                         Dataanalyst
                                     NaN
                                                       15000
                                                                 4
              2
                 Umar
                                                NaN
              3
                  Jane
                           Analytics
                                     NaN
                                            Hyderbad
                                                       20000
                                                              NaN
              4
                 Uttam
                           Statistics
                                       67
                                                NaN
                                                       30000
                                                                 5
              5
                   Kim
                               NLP
                                       55
                                                Delhi
                                                      60000
                                                                10
In [113]:
             clean_data
Out[113]:
                 Name
                            Domain
                                      Age
                                             Location Salary Exp
              0
                                                                 2
                  Mike
                        Datascience
                                        34
                                                         5000
                                              Mumbai
                 Teddy
                             Testing
                                        45
                                                        10000
                                                                 3
              1
                                            Bangalore
              2
                 Umar
                         Dataanalyst 50.25
                                            Bangalore
                                                        15000
                                                                 4
              3
                  Jane
                           Analytics
                                     50.25
                                             Hyderbad
                                                       20000
                                                                4.8
                 Uttam
                           Statistics
                                        67
                                            Bangalore
                                                       30000
                                                                  5
              4
              5
                               NLP
                                                       60000
                   Kim
                                        55
                                                 Delhi
                                                                 10
In [114]:
             cleaned_emp
Out[114]:
                 Name
                            Domain
                                     Age
                                            Location
                                                      Salary
                                                             Exp
              0
                  Mike
                        Datascience
                                       34
                                             Mumbai
                                                       5000
                                                                2
              1
                 Teddy
                             Testing
                                       45
                                           Bangalore
                                                      10000
                                                                3
              2
                 Umar
                         Dataanalyst
                                       50
                                           Bangalore
                                                      15000
                                                                4
```

4

5

Jane

Uttam

Kim

Analytics

Statistics

NLP

67

55

Hyderbad

Bangalore

Delhi

20000

30000

60000

4

5

10

```
In [115]:
             x_iv
Out[115]:
                Name
                          Domain
                                  Age
                                        Location
                                                  Salary Exp
                 Mike
                      Datascience
                                    34
                                          Mumbai
                                                   5000
                                                            2
             0
                Teddy
                                    45
                                        Bangalore
                                                   10000
                                                            3
                           Testing
             2
                Umar
                       Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                            4
             3
                 Jane
                         Analytics
                                        Hyderbad
                                                  20000
                                                            4
                Uttam
                         Statistics
                                    67
                                        Bangalore
                                                  30000
                                                            5
                             NLP
             5
                 Kim
                                    55
                                            Delhi
                                                  60000
                                                           10
In [116]:
            imputation = pd.get_dummies(cleaned_emp)
            imputation
Out[116]:
                Age
                    Salary Exp
                                Name_Jane
                                             Name_Kim Name_Mike Name_Teddy
                                                                                  Name_Umar
                                                                                              Name
                               2
             0
                 34
                      5000
                                       False
                                                   False
                                                               True
                                                                            False
                                                                                        False
             1
                 45
                     10000
                               3
                                       False
                                                   False
                                                               False
                                                                             True
                                                                                        False
             2
                     15000
                                                                                         True
                 50
                               4
                                       False
                                                   False
                                                               False
                                                                            False
             3
                 50
                     20000
                                        True
                                                   False
                                                               False
                                                                            False
                                                                                        False
                               4
                     30000
                               5
                                                                                        False
                 67
                                       False
                                                   False
                                                               False
                                                                            False
             4
             5
                 55
                     60000
                                                               False
                                                                            False
                                                                                        False
                              10
                                       False
                                                   True
            imputation1 = pd.get_dummies(cleaned_emp).sum()
In [118]:
            imputation1
Out[118]:
            Age
                                         301
                                      140000
            Salary
            Exp
                                           28
            Name Jane
                                            1
                                            1
            Name_Kim
            Name Mike
                                            1
            Name_Teddy
                                            1
            Name_Umar
                                            1
            Name_Uttam
                                            1
            Domain_Analytics
                                            1
            Domain_Dataanalyst
                                            1
            Domain_Datascience
                                            1
            Domain_NLP
                                            1
            Domain_Statistics
                                            1
                                            1
            Domain_Testing
                                            3
            Location_Bangalore
            Location Delhi
                                            1
            Location_Hyderbad
                                            1
            Location_Mumbai
                                            1
            dtype: int64
```

In []:

In []:	
In []:	