```
In [1]:
         import pandas as pd
In [2]:
         emp = pd.read_excel(r'F:\Internship project ml\projet dataset\Rawdata.xlsx')
In [3]:
         emp
Out[3]:
              Name
                           Domain
                                       Age
                                             Location
                                                         Salary
                                                                   Exp
          0
               Mike
                     Datascience#$
                                   34 years
                                              Mumbai
                                                         5^00#0
                                                                     2+
             Teddy^
                                      45' yr
                                            Bangalore
                                                      10%%000
           1
                           Testing
                                                                     <3
           2
             Uma#r
                     Dataanalyst^^#
                                       NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
                        Ana^^lytics
           3
                                       NaN
                                            Hyderbad
                                                         2000^0
                                                                   NaN
               Jane
             Uttam*
                          Statistics
                                      67-yr
                                                 NaN
                                                         30000-
                                                                5+ year
           5
                Kim
                              NLP
                                       55yr
                                                Delhi
                                                        6000^$0
                                                                   10+
In [4]:
         id(emp)
Out[4]: 1353077694912
In [5]:
         emp.columns
Out[5]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [6]:
         emp.shape
Out[6]: (6, 6)
         emp.head()
In [7]:
Out[7]:
              Name
                           Domain
                                       Age
                                             Location
                                                         Salary
                                                                   Exp
          0
                     Datascience#$
                                   34 years
                                                         5^00#0
                                                                     2+
               Mike
                                              Mumbai
             Teddy^
                                            Bangalore
                                                      10%%000
                           Testing
                                      45' yr
                                                                     <3
             Uma#r
                     Dataanalyst^^#
                                                       1$5%000
           2
                                       NaN
                                                 NaN
                                                                  4> yrs
           3
               Jane
                        Ana^^lytics
                                       NaN
                                            Hyderbad
                                                         2000^0
                                                                   NaN
             Uttam*
                          Statistics
                                      67-yr
                                                 NaN
                                                         30000- 5+ year
```

```
In [8]:
          emp.tail()
 Out[8]:
               Name
                            Domain
                                      Age
                                            Location
                                                        Salary
                                                                   Exp
              Teddy^
                             Testing
                                     45' yr
                                           Bangalore
                                                     10%%000
                                                                    <3
                      Dataanalyst^^#
               Uma#r
                                     NaN
                                                NaN
                                                      1$5%000
                                                                 4> yrs
            3
                Jane
                         Ana^^lytics
                                     NaN
                                           Hyderbad
                                                        2000^0
                                                                  NaN
               Uttam*
                           Statistics
                                     67-yr
                                                NaN
                                                        30000- 5+ year
            4
            5
                 Kim
                               NLP
                                     55yr
                                               Delhi
                                                       6000^$0
                                                                   10+
 In [9]:
          emp.head(3)
 Out[9]:
               Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                    Exp
            0
                Mike
                       Datascience#$ 34 years
                                               Mumbai
                                                          5^00#0
                                                                     2+
            1
              Teddy^
                             Testing
                                       45' yr
                                              Bangalore
                                                        10%%000
                                                                     <3
              Uma#r Dataanalyst^^#
                                        NaN
                                                  NaN
                                                         1$5%000 4> yrs
In [10]:
          emp.info()
           <class 'pandas.core.frame.DataFrame'>
           RangeIndex: 6 entries, 0 to 5
           Data columns (total 6 columns):
            #
                 Column
                            Non-Null Count
                                               Dtype
                             -----
            0
                 Name
                             6 non-null
                                               object
            1
                 Domain
                             6 non-null
                                               object
            2
                 Age
                            4 non-null
                                               object
            3
                                                object
                 Location
                            4 non-null
            4
                             6 non-null
                 Salary
                                               object
            5
                 Exp
                             5 non-null
                                               object
           dtypes: object(6)
           memory usage: 416.0+ bytes
In [11]:
           emp
Out[11]:
               Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                     Exp
            0
                Mike
                      Datascience#$
                                    34 years
                                                          5^00#0
                                                                      2+
                                               Mumbai
               Teddy<sup>^</sup>
                                              Bangalore
            1
                             Testing
                                       45' yr
                                                        10%%000
                                                                      <3
            2
              Uma#r
                      Dataanalyst^^#
                                        NaN
                                                  NaN
                                                         1$5%000
                                                                   4> yrs
            3
                         Ana^^lytics
                                              Hyderbad
                Jane
                                        NaN
                                                          2000^0
                                                                     NaN
               Uttam*
                           Statistics
                                        67-yr
                                                  NaN
                                                           30000-
                                                                  5+ year
            4
```

5

Kim

NLP

55yr

Delhi

6000^\$0

10+

```
In [12]: emp.isnull()
Out[12]:
               Name Domain
                               Age Location
                                             Salary
                                                      Exp
               False
                        False
                              False
                                        False
                                               False
            0
                                                     False
            1
               False
                        False False
                                        False
                                               False False
            2
               False
                        False
                               True
                                        True
                                               False False
               False
                        False
                               True
                                        False
                                                      True
            3
                                               False
               False
                        False False
                                        True
                                               False False
               False
                        False False
                                        False
                                               False False
In [13]:
           emp.isnull().sum()
Out[13]: Name
                         0
           Domain
                         0
                         2
           Age
           Location
                         2
           Salary
                         0
                         1
           Exp
           dtype: int64
In [14]: emp.columns
Out[14]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [15]:
           emp
Out[15]:
               Name
                            Domain
                                        Age
                                              Location
                                                          Salary
                                                                     Exp
            0
                Mike
                      Datascience#$ 34 years
                                                          5^00#0
                                                                      2+
                                               Mumbai
               Teddy^
                                             Bangalore
                                                                      <3
            1
                             Testing
                                       45' yr
                                                       10%%000
                                                        1$5%000
               Uma#r
                      Dataanalyst^^#
            2
                                        NaN
                                                  NaN
                                                                   4> yrs
            3
                         Ana^^lytics
                                                          2000^0
                Jane
                                        NaN
                                              Hyderbad
                                                                    NaN
               Uttam*
                           Statistics
                                       67-yr
                                                  NaN
                                                          30000-
                                                                 5+ year
            5
                 Kim
                               NLP
                                        55yr
                                                 Delhi
                                                         6000^$0
                                                                     10+
In [16]:
          emp['Name']
Out[16]:
          0
                   Mike
           1
                 Teddy^
           2
                  Uma#r
           3
                   Jane
           4
                 Uttam*
           5
                    Kim
           Name: Name, dtype: object
```

```
In [17]: emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
In [18]: emp['Name']
Out[18]: 0
                Mike
          1
               Teddy
          2
                Umar
          3
                 Jane
          4
               Uttam
          5
                  Kim
          Name: Name, dtype: object
In [19]:
          emp
Out[19]:
             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
              Umar 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
                            NLP
                                             Delhi
               Kim
                                    55yr
                                                    6000^$0
                                                               10+
         emp['Domain']
In [20]:
Out[20]: 0
                Datascience#$
          1
                       Testing
          2
               Dataanalyst^^#
          3
                   Ana^^lytics
          4
                    Statistics
          5
                           NLP
          Name: Domain, dtype: object
In [21]: emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
In [22]: emp['Domain']
Out[22]: 0
               Datascience
                    Testing
          1
          2
               Dataanalyst
          3
                  Analytics
          4
                Statistics
          5
                        NLP
          Name: Domain, dtype: object
```

```
In [23]: emp
Out[23]:
              Name
                         Domain
                                     Age
                                           Location
                                                       Salary
                                                                 Exp
            0
                     Datascience 34 years
                                                       5^00#0
                Mike
                                            Mumbai
                                                                   2+
                                          Bangalore
                                                    10%%000
            1
               Teddy
                          Testing
                                    45' yr
                                                                   <3
            2
               Umar
                      Dataanalyst
                                     NaN
                                               NaN
                                                     1$5%000
                                                                4> yrs
                Jane
                        Analytics
                                     NaN
                                          Hyderbad
                                                       2000^0
                                                                 NaN
            3
            4
               Uttam
                        Statistics
                                    67-yr
                                               NaN
                                                       30000-
                                                              5+ year
            5
                            NLP
                                              Delhi
                                                      6000^$0
                Kim
                                     55yr
                                                                  10+
In [24]:
          emp['Age']=emp['Age'].str.replace(r'\W',' ',regex=True)
In [26]: emp['Age']
Out[26]: 0
                 34 years
           1
                   45 yr
           2
                      NaN
           3
                      NaN
           4
                    67 yr
           5
                     55yr
           Name: Age, dtype: object
          emp['Age'] = emp['Age'].str.extract('(\d+)')
In [27]:
In [28]:
          emp['Age']
Out[28]:
                  34
                  45
           1
           2
                 NaN
           3
                 NaN
           4
                  67
           5
                  55
           Name: Age, dtype: object
In [29]:
           emp
Out[29]:
              Name
                         Domain Age
                                       Location
                                                    Salary
                                                              Exp
            0
                Mike
                     Datascience
                                   34
                                         Mumbai
                                                    5^00#0
                                                               2+
            1
               Teddy
                          Testing
                                   45
                                       Bangalore
                                                 10%%000
                                                               <3
                      Dataanalyst NaN
                                                  1$5%000
            2
               Umar
                                           NaN
                                                            4> yrs
                                       Hyderbad
                                                   2000^0
            3
               Jane
                        Analytics NaN
                                                              NaN
               Uttam
                        Statistics
                                   67
                                           NaN
                                                    30000-
                                                           5+ year
            5
                Kim
                            NLP
                                   55
                                           Delhi
                                                  6000^$0
                                                              10+
```

```
In [30]: emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
In [31]: emp['Location']
Out[31]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
                Hyderbad
          3
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
In [32]: emp['Salary']
Out[32]: 0
                5^00#0
          1
               10%%000
          2
               1$5%000
          3
                2000^0
          4
                30000-
          5
               6000^$0
          Name: Salary, dtype: object
In [35]: emp['Salary']=emp['Salary'].str.replace(r'\W', '', regex=True)
In [36]: emp['Salary']
Out[36]: 0
                5000
          1
               10000
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: object
In [37]: emp.head()
Out[37]:
             Name
                                   Location Salary
                      Domain Age
                                                      Exp
          0
             Mike Datascience
                                     Mumbai
                                              5000
                                                       2+
           1 Teddy
                       Testing
                                45
                                   Bangalore
                                             10000
                                                       <3
                    Dataanalyst NaN
                                            15000
             Umar
                                       NaN
                                                    4> yrs
                                             20000
           3
              Jane
                      Analytics NaN
                                   Hyderbad
                                                     NaN
           4 Uttam
                      Statistics
                                67
                                       NaN
                                             30000 5+ year
```

```
In [38]: emp['Exp']
Out[38]: 0
                     2+
                     <3
          2
                4> yrs
          3
                    NaN
          4
               5+ year
          5
                    10+
          Name: Exp, dtype: object
In [39]: emp['Exp']=emp['Exp'].str.replace(r'\W', '', regex=True)
In [40]: emp['Exp']
Out[40]: 0
                    2
          1
                    3
          2
                4yrs
          3
                 NaN
          4
               5year
          5
                   10
          Name: Exp, dtype: object
In [41]: | emp['Exp'] = emp['Exp'].str.extract('(\d+)')
In [42]: emp['Exp']
Out[42]: 0
                  2
          1
                  3
          2
                  4
          3
               NaN
                  5
          5
                10
          Name: Exp, dtype: object
In [43]: emp
Out[43]:
             Name
                       Domain Age Location Salary
                                                     Exp
           0
              Mike
                    Datascience
                                      Mumbai
                                               5000
             Teddy
                        Testing
                                45
                                    Bangalore
                                              10000
                                                       3
                                              15000
                                                       4
              Umar
                    Dataanalyst NaN
                                        NaN
              Jane
                      Analytics NaN
                                    Hyderbad
                                              20000 NaN
                      Statistics
                                67
                                        NaN
                                              30000
                                                       5
             Uttam
                          NLP
                                              60000
           5
               Kim
                                55
                                        Delhi
                                                      10
In [44]: | clean_data = emp.copy()
```

```
In [45]: clean_data
Out[45]:
             Name
                       Domain Age
                                    Location Salary Exp
                                34
                                      Mumbai
                                               5000
           0
              Mike Datascience
                                                       2
                                    Bangalore
                                              10000
                                                       3
             Teddy
                        Testing
                                45
                                              15000
                                                       4
              Umar
                    Dataanalyst NaN
                                        NaN
              Jane
                      Analytics NaN
                                    Hyderbad
                                              20000 NaN
                                              30000
             Uttam
                      Statistics
                                67
                                        NaN
                                                       5
                                              60000
           5
               Kim
                          NLP
                                55
                                        Delhi
                                                      10
          clean_data.isnull().sum()
In [46]:
Out[46]: Name
                       0
          Domain
                       0
                       2
          Age
          Location
                       2
                       0
          Salary
                       1
          Exp
          dtype: int64
In [47]: clean_data['Age']
Out[47]: 0
                 34
                 45
          1
          2
               NaN
          3
               NaN
          4
                 67
          5
                 55
          Name: Age, dtype: object
In [48]: import numpy as np
In [49]: | clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data[
In [50]: clean_data['Age']
Out[50]: 0
                   34
                   45
          1
          2
               50.25
          3
               50.25
          4
                   67
          5
                   55
          Name: Age, dtype: object
```

```
In [51]: clean_data['Exp']
Out[51]: 0
                  2
          1
                  3
          2
                  4
          3
                NaN
          4
                  5
          5
                 10
          Name: Exp, dtype: object
In [52]: | clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data[
In [53]: clean_data['Exp']
Out[53]: 0
                  2
          1
                  3
          2
                  4
          3
                4.8
          4
                  5
          5
                 10
          Name: Exp, dtype: object
In [54]:
          clean_data
Out[54]:
              Name
                       Domain
                                Age
                                      Location Salary Exp
           0
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                5000
                                                        2
              Teddy
                                     Bangalore
                                               10000
                                                        3
                        Testing
                    Dataanalyst 50.25
                                               15000
                                                        4
              Umar
                                          NaN
                       Analytics 50.25
                                     Hyderbad
                                               20000
              Jane
                                                       4.8
              Uttam
                       Statistics
                                  67
                                          NaN
                                               30000
                                                        5
                          NLP
                                               60000
           5
               Kim
                                  55
                                         Delhi
                                                       10
In [55]: clean_data['Location'].isnull().sum()
Out[55]: 2
In [56]: | clean_data['Location']
Out[56]: 0
                   Mumbai
          1
                Bangalore
          2
                      NaN
          3
                 Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
In [57]: clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].r
```

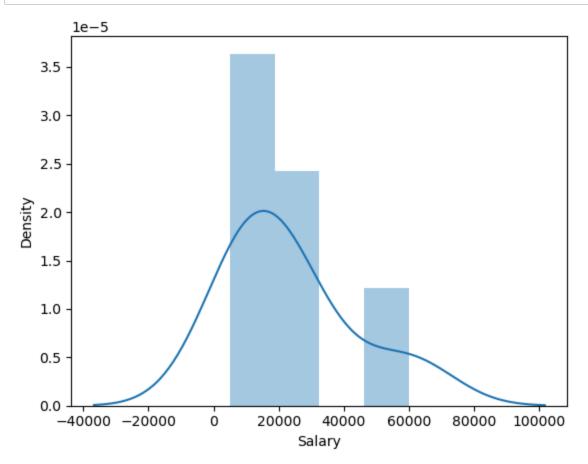
```
In [59]: clean_data['Location']
Out[59]: 0
                  Mumbai
               Bangalore
          1
          2
               Bangalore
          3
                Hyderbad
          4
               Bangalore
          5
                    Delhi
          Name: Location, dtype: object
In [60]:
         clean_data
Out[60]:
             Name
                       Domain
                                Age
                                     Location Salary Exp
                                                5000
                                                       2
           0
              Mike Datascience
                                 34
                                       Mumbai
             Teddy
                                               10000
                                                       3
           1
                        Testing
                                 45 Bangalore
           2
              Umar
                    Dataanalyst 50.25 Bangalore
                                               15000
                                                       4
           3
              Jane
                      Analytics 50.25
                                     Hyderbad
                                               20000
                                                      4.8
             Uttam
                      Statistics
                                     Bangalore
                                               30000
                                                       5
                          NLP
                                               60000
                                                      10
           5
               Kim
                                 55
                                         Delhi
In [61]: 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
                                            object
               Domain
                          6 non-null
           2
               Age
                          6 non-null
                                            object
           3
               Location 6 non-null
                                            object
           4
               Salary
                          6 non-null
                                            object
           5
                          6 non-null
               Exp
                                            object
          dtypes: object(6)
          memory usage: 416.0+ bytes
```

In [62]: clean_data['Age'] = clean_data['Age'].astype(int)

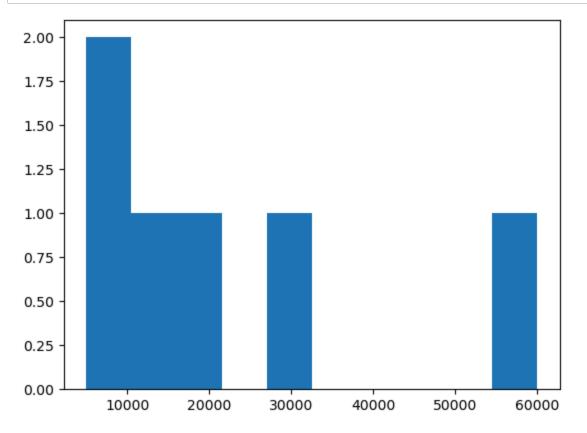
```
In [63]: clean_data.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
                        Non-Null Count Dtype
              Column
          0
              Name
                        6 non-null
                                        object
          1
            Domain
                        6 non-null
                                        object
          2
             Age
                        6 non-null
                                        int32
          3
             Location 6 non-null
                                        object
          4
              Salary
                        6 non-null
                                        object
          5
              Exp
                        6 non-null
                                        object
         dtypes: int32(1), object(5)
         memory usage: 392.0+ bytes
In [64]: | clean_data['Salary'] = clean_data['Salary'].astype(int)
In [65]: | clean_data['Exp'] = clean_data['Exp'].astype(int)
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
         --- -----
          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
                                        int32
          5
              Exp
                        6 non-null
                                        int32
         dtypes: int32(3), object(3)
         memory usage: 344.0+ bytes
In [67]: clean_data['Name'] = clean_data['Name'].astype('category')
         clean_data['Domain'] = clean_data['Domain'].astype('category')
         clean_data['Location'] = clean_data['Location'].astype('category')
```

```
In [68]: 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
                                            category
           1
               Domain
                          6 non-null
                                            category
           2
               Age
                          6 non-null
                                            int32
           3
                Location
                          6 non-null
                                             category
           4
                Salary
                          6 non-null
                                            int32
           5
                Exp
                           6 non-null
                                             int32
          dtypes: category(3), int32(3)
          memory usage: 862.0 bytes
In [69]:
          clean_data
Out[69]:
                                     Location Salary Exp
              Name
                       Domain Age
           0
               Mike Datascience
                                 34
                                      Mumbai
                                               5000
                                                       2
                                 45 Bangalore
                                              10000
           1
              Teddy
                        Testing
                                                       3
                                              15000
           2
              Umar Dataanalyst
                                50 Bangalore
                                                       4
                                              20000
           3
              Jane
                       Analytics
                                 50
                                    Hyderbad
                                                       4
              Uttam
                      Statistics
                                67
                                    Bangalore
                                              30000
                                                       5
           5
               Kim
                          NLP
                                 55
                                        Delhi 60000
                                                      10
In [70]: clean_data.to_csv('clean_data.csv')
In [71]:
          import os
          os.getcwd() #from the os give the saved current working directly
Out[71]: 'C:\\Users\\NEHA'
In [72]:
          clean_data
Out[72]:
                                     Location Salary Exp
              Name
                       Domain Age
           0
               Mike Datascience
                                 34
                                      Mumbai
                                               5000
                                                       2
           1
              Teddy
                        Testing
                                45 Bangalore
                                              10000
                                                       3
                    Dataanalyst
                                    Bangalore
                                              15000
           2
              Umar
                                 50
                                                       4
                                              20000
              Jane
                      Analytics
                                 50
                                    Hyderbad
                                                       4
              Uttam
                       Statistics
                                 67
                                    Bangalore
                                              30000
                                                       5
                                        Delhi 60000
           5
               Kim
                          NLP
                                 55
                                                      10
```

```
In [73]: import matplotlib.pyplot as plt # visualization
         import seaborn as sns
In [74]: import warnings
         warnings.filterwarnings('ignore')
In [75]: clean_data['Salary']
Out[75]: 0
               5000
         1
              10000
         2
              15000
         3
              20000
         4
              30000
         5
              60000
         Name: Salary, dtype: int32
In [76]: vis1 = sns.distplot(clean_data['Salary'])
```



In [77]: vis2 = plt.hist(clean_data['Salary'])

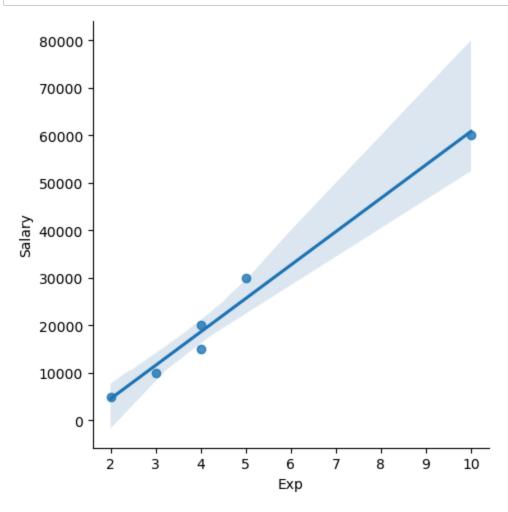


In [78]: clean_data

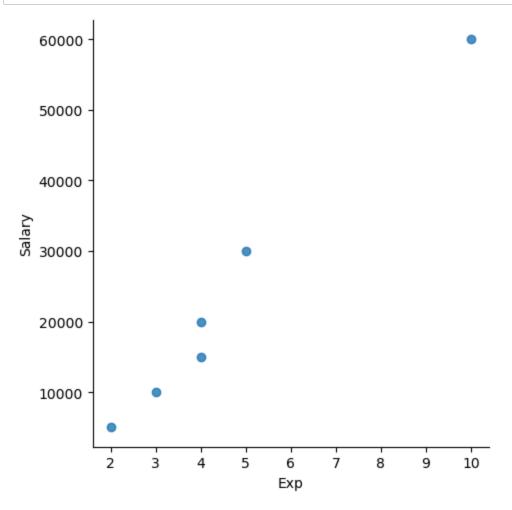
Out[78]:

	Name	Domain	Age	Location	Salary	Exp
() Mike	Datascience	34	Mumbai	5000	2
1	l Teddy	Testing	45	Bangalore	10000	3
2	2 Umar	Dataanalyst	50	Bangalore	15000	4
3	3 Jane	Analytics	50	Hyderbad	20000	4
4	U ttam	Statistics	67	Bangalore	30000	5
Ę	K im	NLP	55	Delhi	60000	10

In [79]: vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary')



In [80]: vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)



In [81]: clean_data[:]

Out[81]:

	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 [82]:
          clean_data[::]
Out[82]:
               Name
                          Domain Age
                                         Location
                                                   Salary Exp
            0
                                                     5000
                                                             2
                      Datascience
                                    34
                                          Mumbai
                Mike
                                                    10000
            1
               Teddy
                           Testing
                                    45
                                        Bangalore
                                                             3
                                                    15000
            2
                Umar
                       Dataanalyst
                                    50
                                        Bangalore
                                                             4
            3
                Jane
                         Analytics
                                    50
                                         Hyderbad
                                                   20000
                                                             4
            4
               Uttam
                         Statistics
                                    67
                                        Bangalore
                                                   30000
                                                             5
                             NLP
                                                   60000
            5
                 Kim
                                    55
                                             Delhi
                                                            10
           clean_data[1:3]
In [83]:
Out[83]:
               Name
                          Domain Age
                                        Location Salary Exp
               Teddy
                           Testing
                                    45
                                        Bangalore
                                                   10000
                                                             3
                Umar Dataanalyst
                                        Bangalore
                                                   15000
                                                             4
                                    50
In [84]:
           clean_data[1:3:4]
Out[84]:
               Name
                      Domain Age
                                      Location
                                               Salary
               Teddy
                       Testing
                                 45
                                     Bangalore
                                                10000
                                                          3
           clean_data[0:6:2]
In [85]:
Out[85]:
               Name
                                         Location
                                                   Salary Exp
                          Domain Age
            0
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                     5000
                                                             2
                                                    15000
            2
                Umar
                       Dataanalyst
                                    50
                                        Bangalore
                                                             4
               Uttam
                         Statistics
                                        Bangalore
                                                   30000
                                                             5
In [86]:
           clean_data[::-1]
Out[86]:
               Name
                          Domain Age
                                         Location
                                                   Salary Exp
            5
                                                   60000
                 Kim
                             NLP
                                    55
                                             Delhi
                                                            10
                                                   30000
            4
               Uttam
                         Statistics
                                    67
                                        Bangalore
                                                             5
            3
                Jane
                         Analytics
                                    50
                                         Hyderbad
                                                   20000
                                                             4
                Umar
                       Dataanalyst
                                        Bangalore
                                                    15000
            2
                                    50
                                                             4
                                                    10000
            1
               Teddy
                           Testing
                                    45
                                        Bangalore
                                                             3
                                                     5000
            0
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                             2
```

```
In [87]: clean_data[:]
Out[87]:
               Name
                         Domain Age
                                        Location
                                                  Salary Exp
                      Datascience
                                                    5000
                                                            2
            0
                                    34
                                         Mumbai
                Mike
                                                  10000
            1
               Teddy
                          Testing
                                    45
                                        Bangalore
                                                            3
                                                  15000
            2
               Umar
                      Dataanalyst
                                   50
                                        Bangalore
                                                            4
                         Analytics
                                    50
                                        Hyderbad
                                                  20000
            3
                Jane
                                                            4
            4
               Uttam
                         Statistics
                                    67
                                        Bangalore
                                                  30000
                                                            5
                            NLP
                                                  60000
            5
                 Kim
                                    55
                                            Delhi
                                                           10
In [88]:
           clean_data.columns
Out[88]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [89]:
           clean_data
Out[89]:
               Name
                          Domain Age
                                        Location
                                                  Salary Exp
                                                            2
            0
                Mike
                      Datascience
                                    34
                                         Mumbai
                                                    5000
               Teddy
                                        Bangalore
                                                  10000
            1
                          Testing
                                   45
                                                            3
                      Dataanalyst
                                                  15000
            2
               Umar
                                   50
                                        Bangalore
                                                            4
                                        Hyderbad
                                                  20000
            3
                Jane
                         Analytics
                                    50
                                                            4
               Uttam
                         Statistics
                                    67
                                        Bangalore
                                                  30000
                                                            5
                                                  60000
            5
                 Kim
                            NLP
                                    55
                                            Delhi
                                                           10
In [90]: X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [91]:
           X_iv
Out[91]:
               Name
                         Domain Age
                                        Location
                                                  Exp
                      Datascience
            0
                Mike
                                    34
                                         Mumbai
                                                    2
                                        Bangalore
            1
               Teddy
                          Testing
                                   45
                                                    3
            2
               Umar
                      Dataanalyst
                                        Bangalore
                                                    4
                                    50
                         Analytics
                                        Hyderbad
            3
                Jane
                                    50
                                                    4
               Uttam
                                                    5
                         Statistics
                                    67
                                        Bangalore
                            NLP
            5
                                    55
                                            Delhi
                                                   10
                 Kim
In [92]: y_dv = clean_data[['Salary']]
```

```
In [93]:
           y_dv
Out[93]:
                Salary
                 5000
             0
             1
                10000
                15000
             2
             3
                20000
             4
                30000
             5
                60000
In [94]:
            emp
Out[94]:
                Name
                                    Age
                                                              Exp
                           Domain
                                           Location
                                                     Salary
             0
                 Mike
                       Datascience
                                      34
                                            Mumbai
                                                       5000
                                                                2
             1
                                                      10000
                                                                3
                Teddy
                            Testing
                                      45
                                          Bangalore
                        Dataanalyst NaN
                                                      15000
                                                                4
             2
                Umar
                                               NaN
                                           Hyderbad
                                                      20000
             3
                 Jane
                          Analytics NaN
                                                             NaN
                                      67
                                                      30000
                                                                5
             4
                Uttam
                          Statistics
                                               NaN
             5
                  Kim
                              NLP
                                      55
                                               Delhi
                                                      60000
                                                               10
In [95]:
            clean_data
Out[95]:
                Name
                                    Age
                                           Location
                                                     Salary
                                                             Exp
                           Domain
             0
                       Datascience
                                                                2
                 Mike
                                      34
                                            Mumbai
                                                       5000
             1
                Teddy
                            Testing
                                      45
                                          Bangalore
                                                      10000
                                                                3
             2
                Umar
                        Dataanalyst
                                      50
                                          Bangalore
                                                      15000
                                                                4
             3
                          Analytics
                                           Hyderbad
                                                      20000
                                                                4
                 Jane
                                      50
                                          Bangalore
                                                      30000
                                                                5
             4
                Uttam
                          Statistics
                                      67
                                                      60000
             5
                  Kim
                              NLP
                                      55
                                               Delhi
                                                               10
In [96]:
           X_iv
Out[96]:
                Name
                           Domain
                                    Age
                                           Location
                                                     Exp
             0
                                                        2
                       Datascience
                                      34
                 Mike
                                            Mumbai
             1
                Teddy
                            Testing
                                      45
                                          Bangalore
                                                        3
             2
                        Dataanalyst
                                      50
                                                        4
                Umar
                                          Bangalore
             3
                 Jane
                          Analytics
                                      50
                                           Hyderbad
                                                        4
                                                        5
             4
                Uttam
                          Statistics
                                      67
                                          Bangalore
```

10

Delhi

5

Kim

NLP

55

```
In [97]: y_dv
 Out[97]:
                Salary
                 5000
             0
                10000
                15000
             2
                20000
             3
                30000
                60000
 In [98]:
            clean_data
 Out[98]:
                Name
                          Domain Age
                                        Location Salary Exp
                                                    5000
                                                            2
             0
                 Mike
                      Datascience
                                    34
                                          Mumbai
             1
                Teddy
                                    45
                                        Bangalore
                                                   10000
                                                            3
                           Testing
                                                   15000
                Umar
                       Dataanalyst
                                        Bangalore
                                                            4
             2
                                    50
                                        Hyderbad
                                                   20000
                                                            4
             3
                 Jane
                         Analytics
                                    50
                                                   30000
                                        Bangalore
                                                            5
                Uttam
                         Statistics
                                    67
             5
                  Kim
                             NLP
                                    55
                                            Delhi
                                                   60000
                                                           10
            imputation = pd.get_dummies(clean_data, dtype=int)
 In [99]:
In [100]:
            imputation
Out[100]:
                Age Salary Exp
                                 Name_Jane Name_Kim Name_Mike Name_Teddy Name_Umar Name_Uttai
                      5000
                                                                                            0
             0
                 34
                               2
                                           0
                                                      0
                                                                  1
                                                                                0
                      10000
                                                                  0
             1
                 45
                               3
                                           0
                                                      0
                                                                                1
                                                                                            0
             2
                 50
                      15000
                               4
                                           0
                                                      0
                                                                  0
                                                                                0
                                                                                            1
                 50
                     20000
                               4
                                           1
                                                      0
                                                                  0
                                                                                0
                                                                                            0
             3
             4
                 67
                     30000
                               5
                                           0
                                                      0
                                                                  0
                                                                                0
                                                                                            0
                     60000
                                           0
                                                      1
                                                                  0
                                                                                0
                                                                                            0
             5
                 55
                              10
```

```
In [101]: clean_data
Out[101]:
               Name
                          Domain Age
                                        Location Salary Exp
                                                   5000
                Mike Datascience
                                    34
                                         Mumbai
                                                           2
                                                  10000
             1 Teddy
                          Testing
                                   45 Bangalore
                                                           3
                Umar
                      Dataanalyst
                                   50 Bangalore
                                                  15000
                                                           4
                         Analytics
                                        Hyderbad
                                                  20000
                                                           4
                Jane
               Uttam
                         Statistics
                                   67
                                       Bangalore
                                                  30000
                                                           5
                             NLP
                                            Delhi 60000
                 Kim
                                    55
                                                          10
In [102]: len(clean_data)
Out[102]: 6
In [103]: imputation.columns
Out[103]: Index(['Age', 'Salary', 'Exp', 'Name_Jane', 'Name_Kim', 'Name_Mike',
                    'Name_Teddy', 'Name_Umar', 'Name_Uttam', 'Domain_Analytics',
                    'Domain_Dataanalyst', 'Domain_Datascience', 'Domain_NLP',
                    'Domain_Statistics', 'Domain_Testing', 'Location_Bangalore', 'Location_Delhi', 'Location_Hyderbad', 'Location_Mumbai'],
                   dtype='object')
In [104]: len(imputation.columns)
Out[104]: 19
  In [ ]:
```