```
In [56]: import pandas as pd
In [57]: emp=pd.read_excel(r'C:\Users\yogay\OneDrive\Documents\python\Rawdata.xlsx')
In [58]: emp
Out[58]:
              Name
                          Domain
                                     Age
                                          Location
                                                      Salary
                                                               Ехр
                                                     5^00#0
                                                                2+
          0
               Mike
                   Datascience#$ 34 years
                                           Mumbai
          1 Teddy<sup>^</sup>
                          Testing
                                    45' yr
                                         Bangalore
                                                   10%%000
                                                                <3
                    Dataanalyst^^#
                                                    1$5%000
          2
             Uma#r
                                     NaN
                                              NaN
                                                             4> yrs
          3
               Jane
                       Ana^^lytics
                                     NaN
                                          Hyderbad
                                                     2000^0
                                                               NaN
          4
             Uttam*
                         Statistics
                                    67-yr
                                              NaN
                                                      30000-
                                                            5+ year
                Kim
                            NLP
                                     55yr
                                             Delhi
                                                    6000^$0
                                                               10+
In [59]: emp.shape
          (6, 6)
Out[59]:
In [60]:
          len(emp)
In [61]:
          emp.columns
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[61]:
         len(emp.columns)
In [62]:
Out[62]:
In [63]: emp.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
           #
               Column
                          Non-Null Count Dtype
           0
               Name
                          6 non-null
                                            object
           1
               Domain
                          6 non-null
                                            object
           2
                Age
                           4 non-null
                                            object
           3
                Location 4 non-null
                                            object
           4
                           6 non-null
                                            obiect
                Salary
           5
               Exp
                           5 non-null
                                            object
          dtypes: object(6)
          memory usage: 416.0+ bytes
In [64]: emp
Out[64]:
              Name
                          Domain
                                     Age
                                          Location
                                                      Salary
                                                               Exp
          0
              Mike
                    Datascience#$ 34 years
                                           Mumbai
                                                     5^00#0
                                                                2+
          1 Teddy<sup>^</sup>
                          Testing
                                    45' yr
                                         Bangalore
                                                   10%%000
                                                                <3
          2
             Uma#r
                    Dataanalyst^^#
                                     NaN
                                              NaN
                                                    1$5%000
                                                              4> yrs
          3
                       Ana^^lytics
                                     NaN
                                          Hyderbad
                                                     2000^0
                                                               NaN
               Jane
          4 Uttam*
                         Statistics
                                    67-yr
                                              NaN
                                                      30000-
                                                            5+ year
          5
                Kim
                            NLP
                                     55yr
                                              Delhi
                                                    6000^$0
In [65]: emp['Name']
          0
                  Mike
Out[65]:
                Teddy^
          2
                Uma#r
          3
                  Jane
          4
                Uttam*
                  Kim
          Name: Name, dtype: object
In [66]: emp['Domain']
          0
                 Datascience#$
Out[66]:
                       Testing
                Dataanalyst^^#
          2
                   Ana^^lytics
          3
          4
                    Statistics
                           NLP
```

Name: Domain, dtype: object

```
In [67]: emp['Age']
Out[67]: 0
                34 years
                  45' yr
                     NaN
          3
                     NaN
          4
                   67-yr
                    55yr
          5
          Name: Age, dtype: object
In [68]: emp['Location']
          0
                 Mumbai
Out[68]:
              Bangalore
                      NaN
          2
          3
                Hyderbad
                     NaN
          5
                    Delhi
          Name: Location, dtype: object
In [69]: emp['Salary']
Out[69]: 0
                 5^00#0
                10%%000
                1$5%000
          2
          3
                 2000^0
                 30000-
              6000^$0
          5
          Name: Salary, dtype: object
In [70]: emp['Exp']
Out[70]:
                     <3
                 4> yrs
          2
          3
                    NaN
                5+ year
                    10+
          5
          Name: Exp, dtype: object
In [71]: emp[['Name','Domain']]
                          Domain
Out[71]:
              Name
               Mike
                    Datascience#$
          1 Teddy<sup>^</sup>
                          Testing
          2 Uma#r Dataanalyst^^#
               Jane
                        Ana^^lytics
             Uttam*
                         Statistics
                             NLP
                Kim
In [72]: emp[['Name','Domain','Age']]
Out[72]:
              Name
                          Domain
                                     Age
               Mike Datascience#$ 34 years
          1 Teddy<sup>^</sup>
                          Testing
                                    45' yr
          2 Uma#r Dataanalyst^^#
                                     NaN
                        Ana^^lytics
                                     NaN
          3
               Jane
            Uttam*
                         Statistics
                                     67-yr
                Kim
                             NLP
                                     55yr
In [73]: emp[['Name','Domain','Age','Location','Salary','Exp']]
                                                                Ехр
Out[73]:
                          Domain
                                     Age Location
                                                      Salary
              Name
                                                                 2+
               Mike
                     Datascience#$ 34 years
                                            Mumbai
                                                      5^00#0
                                    45' yr Bangalore 10%%000
                                                                 <3
          1 Teddy<sup>^</sup>
                          Testing
             Uma#r Dataanalyst^^#
                                     NaN
                                               NaN
                                                    1$5%000
                                                              4> yrs
               Jane
                        Ana^^lytics
                                     NaN
                                          Hyderbad
                                                      2000^0
                                                                NaN
                         Statistics
                                     67-yr
                                                      30000- 5+ year
            Uttam*
                                               NaN
                             NLP
                Kim
                                     55yr
                                              Delhi
                                                     6000^$0
                                                                10+
In [74]: emp['Name']
```

```
Mike
Out[74]:
                Teddy^
                 Uma#r
                  Jane
          4
                Uttam*
          5
                   Kim
          Name: Name, dtype: object
In [75]:
          import warnings
          warnings.filterwarnings('ignore')
In [76]: emp['Name']=emp['Name'].str.replace(r'\W','')
In [77]: emp
                                                               Ехр
             Name
                        Domain
                                    Age Location
                                                     Salary
Out[77]:
              Mike
                   Datascience#$ 34 years
                                                     5^00#0
                                                                2+
                                           Mumbai
                                                                <3
                         Testing
                                   45' yr Bangalore 10%%000
          1 Teddy
          2
             Umar Dataanalyst^^#
                                    NaN
                                              NaN
                                                   1$5%000
                                                             4> yrs
                       Ana^^lytics
                                    NaN Hyderbad
                                                     2000^0
                                                              NaN
              Jane
                                   67-yr
                                                     30000- 5+ year
          4 Uttam
                        Statistics
                                             NaN
                            NLP
               Kim
                                    55yr
                                             Delhi
                                                    6000^$0
                                                               10+
```

data cleaning

```
In [78]: emp['Domain']=emp['Domain'].str.replace(r'\W','')
In [79]:
          emp['Domain']
               Datascience
Out[79]:
                   Testing
               Dataanalyst
          3
                 Analytics
          4
                Statistics
                       NLP
          Name: Domain, dtype: object
In [80]: emp
Out[80]:
                                                 Salary
            Name
                      Domain
                                 Age Location
                                                          Exp
             Mike Datascience 34 years
                                                 5^00#0
                                       Mumbai
                                                            <3
          1 Teddy
                                45' yr Bangalore 10%%000
                       Testing
          2
             Umar Dataanalyst
                                NaN
                                          NaN
                                               1$5%000
                     Analytics
                                NaN Hyderbad
                                                 2000^0
                                                          NaN
              Jane
                                                 30000- 5+ year
          4 Uttam
                                67-yr
                                          NaN
                     Statistics
                                         Delhi
              Kim
                        NLP
                                 55yr
                                                6000^$0
                                                           10+
In [81]: emp['Age']=emp['Age'].str.replace(r'\W','')
In [82]: emp['Age']
               34years
Out[82]:
          1
                  45yr
                   NaN
          2
                   NaN
          3
          4
                  67yr
                  55yr
          Name: Age, dtype: object
In [83]: emp['Age']=emp['Age'].str.extract('(\d+)')
In [84]: emp['Age']
                34
Out[84]:
                45
          2
               NaN
          3
               NaN
          4
                67
                55
          Name: Age, dtype: object
In [85]: emp
```

```
Name
                       Domain Age
                                      Location
                                                  Salary
                                                            Ехр
Out[85]:
              Mike Datascience
                                       Mumbai
                                                 5^00#0
                                                             2+
           1 Teddy
                        Testing
                                 45 Bangalore
                                               10%%000
                                                             <3
                                                1$5%000
                                                          4> yrs
              Umar
                    Dataanalyst NaN
                                         NaN
               Jane
                       Analytics NaN
                                     Hyderbad
                                                 2000^0
                                                           NaN
                       Statistics
                                                  30000- 5+ year
              Uttam
                                          NaN
                          NLP
                                 55
                                         Delhi
                                                6000^$0
               Kim
                                                            10+
In [86]: emp['Salary'] = emp['Salary'].str.replace(r'\W','')
In [87]:
           emp['Salary']
                 5000
           0
Out[87]:
                10000
                15000
                20000
                30000
                60000
           Name: Salary, dtype: object
In [88]:
Out[88]:
             Name
                       Domain Age
                                      Location Salary
                                                         Ехр
                                                          2+
               Mike Datascience
                                       Mumbai
                                                5000
                                                          <3
           1 Teddy
                        Testing
                                 45
                                     Bangalore
                                               10000
                                               15000
              Umar
                    Dataanalyst NaN
                                          NaN
                                                       4> yrs
               Jane
                       Analytics
                               NaN
                                     Hyderbad
                                               20000
                                                         NaN
              Uttam
                       Statistics
                                 67
                                          NaN
                                               30000
                                                      5+ year
                          NLP
               Kim
                                 55
                                         Delhi
                                               60000
                                                         10+
In [89]: emp['Exp']=emp['Exp'].str.extract('(\d+)')
In [90]:
           emp
                                      Location Salary
Out[90]:
             Name
                       Domain Age
                                                      Exp
               Mike
                    Datascience
                                       Mumbai
                                                5000
                                                         2
                                                         3
           1 Teddy
                        Testing
                                 45
                                     Bangalore
                                               10000
                    Dataanalyst NaN
                                               15000
                                                         4
              Umar
                                          NaN
               Jane
                       Analytics
                               NaN
                                     Hyderbad
                                               20000
                                                      NaN
              Uttam
                       Statistics
                                 67
                                               30000
                                          NaN
                          NLP
                                                        10
               Kim
                                 55
                                         Delhi
                                               60000
In [91]: clean_data = emp.copy()
In [92]: clean_data
                       Domain Age
Out[92]:
             Name
                                      Location Salary
                                                     Exp
               Mike Datascience
                                       Mumbai
           1 Teddy
                        Testing
                                 45
                                     Bangalore
                                               10000
                                                         3
              Umar
                    Dataanalyst
                               NaN
                                          NaN
                                               15000
               Jane
                       Analytics NaN
                                     Hyderbad
                                               20000
             Uttam
                       Statistics
                                 67
                                          NaN
                                               30000
                                                        5
               Kim
                          NLP
                                 55
                                               60000
                                                        10
                                         Delhi
```

Handling Missing Values

In [93]: clean_data

```
Out[93]:
             Name
                      Domain Age
                                    Location Salary Exp
             Mike Datascience
                                34
                                     Mumbai
                                              5000
                                                      2
          1 Teddy
                       Testing
                                45
                                   Bangalore
                                             10000
                                                      3
                                                      4
          2
             Umar
                   Dataanalyst NaN
                                        NaN
                                             15000
          3
              Jane
                      Analytics
                              NaN
                                    Hyderbad
                                             20000
                                                    NaN
                                67
             Uttam
                      Statistics
                                        NaN
                                             30000
                         NLP
                                55
                                       Delhi
                                             60000
               Kim
                                                     10
In [94]: 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
                Age
                           4 non-null
                                            object
           3
                Location 4 non-null
                                            object
           4
                Salary
                           6 non-null
                                            object
               Exp
                           5 non-null
                                            object
          dtypes: object(6)
          memory usage: 416.0+ bytes
In [95]: import numpy as np
In [96]: clean_data.head(1)
                       Domain Age Location Salary
Out[96]:
                                                   Ехр
             Mike Datascience
                                             5000
                                                     2
                                34
                                    Mumbai
In [97]: clean_data['Age']
                 34
Out[97]:
                 45
          2
                NaN
          3
                NaN
                 67
                 55
          Name: Age, dtype: object
In [98]: clean data['Age'] = clean data['Age'].fillna(np.mean(pd.to numeric(clean data['Age'])))
In [99]: clean_data['Age']
                   34
Out[99]:
                   45
          2
                50.25
          3
                50.25
                   67
          5
                   55
          Name: Age, dtype: object
In [100... emp
Out[100]:
              Name
                       Domain Age Location Salary
                                                     Exp
               Mike
                    Datascience
                                      Mumbai
                                               5000
                                                       2
           1 Teddy
                        Testing
                                 45 Bangalore
                                              10000
                                                       3
           2
                                              15000
                                                       4
              Umar
                    Dataanalyst
                               NaN
                                         NaN
           3
               Jane
                       Analytics
                               NaN
                                     Hyderbad
                                              20000
                                                     NaN
           4 Uttam
                       Statistics
                                 67
                                         NaN
                                              30000
                                                       5
                          NLP
                                 55
                                              60000
                Kim
                                        Delhi
                                                      10
          clean_data
In [101...
                                      Location Salary
Out[101]:
              Name
                       Domain
                                Age
                                                      Exp
               Mike
                    Datascience
                                                5000
                                                        2
                                  34
                                       Mumbai
           1 Teddy
                        Testing
                                  45
                                               10000
                                                        3
                                     Bangalore
           2 Umar
                     Dataanalyst 50.25
                                          NaN
                                               15000
                                                        4
                                               20000
           3
               Jane
                       Analytics 50.25
                                     Hyderbad
                                                     NaN
```

4 Uttam

Kim

Statistics

NLP

67

55

NaN

Delhi

30000

60000

5

10

```
clean data['Exp'] = clean data['Exp'].fillna(np.mean(pd.to numeric(clean data['Exp'])))
In [103...
          clean_data['Exp']
In [104...
                  2
Out[104]:
                  4
           2
           3
                4.8
           4
                  5
                 10
           Name: Exp, dtype: object
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode()[0])
In [105...
In [106...
          clean_data['Location']
                   Mumbai
Out[106]:
                Bangalore
           2
                Bangalore
           3
                 Hyderbad
                Bangalore
                    Delhi
           Name: Location, dtype: object
In [107... clean_data
Out[107]:
             Name
                       Domain
                               Age
                                     Location Salary Exp
                                                      2
           0 Mike Datascience
                                 34
                                              5000
                                      Mumbai
           1 Teddy
                       Testing
                                 45 Bangalore
                                              10000
                                                      3
           2 Umar
                    Dataanalyst 50.25 Bangalore
                                              15000
                                                      4
                      Analytics 50.25 Hyderbad
                                              20000
                                                     48
           3
              Jane
           4 Uttam
                      Statistics
                                 67 Bangalore
                                              30000
                                                      5
               Kim
                          NLP
                                        Delhi
                                              60000
                                                     10
          change types of objects
In [108...
          clean_data['Age'] = clean_data['Age'].astype(int)
In [109...
          clean_data['Salary'] = clean_data['Salary'].astype(int)
          clean_data['Exp'] = clean_data['Exp'].astype(int)
In [110...
          clean_data['Domain'] = clean_data['Domain'].astype('category')
In [113...
In [114... clean_data['Name'] = clean_data['Name'].astype('category')
In [115... clean_data['Location'] = clean_data['Location'].astype('category')
In [116... 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
                                           category
           1
               Domain
                          6 non-null
                                           category
           2
                                           int32
               Age
                          6 non-null
           3
               Location 6 non-null
                                           category
           4
               Salary
                          6 non-null
                                           int32
               Exp
                          6 non-null
                                           int32
          dtypes: category(3), int32(3)
          memory usage: 862.0 bytes
In [117... clean data
             Name
                       Domain Age Location Salary Exp
           0 Mike Datascience
                               34
                                     Mumbai
                                              5000
           1 Teddy
                       Testing
                               45 Bangalore
                                             10000
                                                     3
           2 Umar
                    Dataanalyst
                               50 Bangalore
                                             15000
                                                     4
              Jane
                      Analytics
                                50
                                   Hyderbad
                                             20000
                                                     4
                                             30000
           4 Uttam
                      Statistics
                                67 Bangalore
                                                     5
                          NLP
                                       Delhi
                                             60000
In [120... clean_data.to_csv('clean_data.csv') #get dataset into the system
```

```
In [118... import os
          os.getcwd()
           'C:\\Users\\yogay'
Out[118]:
          clean_data.columns
In [121...
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [122...
          import matplotlib.pyplot as plt # visualization
          import seaborn as sns # Advanced visualization
          import warnings
In [123...
          warnings.filterwarnings('ignore')
In [124... clean_data
Out[124]:
             Name
                       Domain Age Location Salary Exp
           0 Mike
                   Datascience
                                     Mumbai
                                              5000
                                                     2
                       Testing
                               45 Bangalore 10000
           1 Teddy
                                                     3
           2 Umar
                   Dataanalyst
                               50 Bangalore
                                            15000
                                                     4
              Jane
                      Analytics
                                   Hyderbad
                                             20000
                                                     4
                                             30000
           4 Uttam
                      Statistics
                                67 Bangalore
                                                     5
                                                    10
               Kim
                          NLP
                                55
                                       Delhi
                                             60000
In [125_ clean_data['Salary']
                 5000
Out[125]:
                10000
                15000
           3
                20000
                30000
                60000
           Name: Salary, dtype: int32
In [126... vis1 = sns.distplot(clean_data['Salary'])
                  1e-5
             3.5
             3.0
             2.5
          Density
             2.0
             1.5
             1.0
             0.5
```

```
In [127... plt.rcParams['figure.figsize'] = 8,5
In [128... vis1 = sns.distplot(clean_data['Salary'])
```

60000

80000

100000

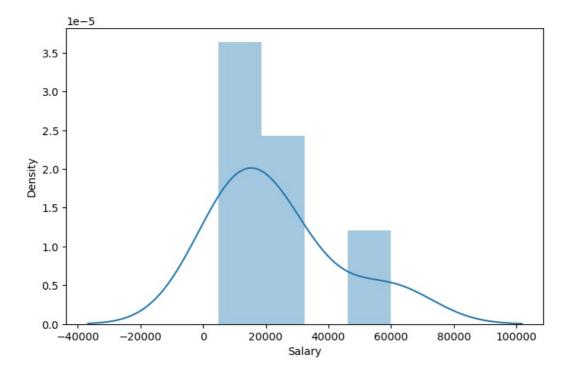
-40000 -20000

0

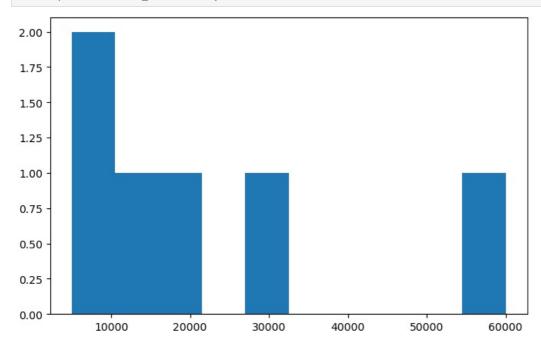
20000

40000

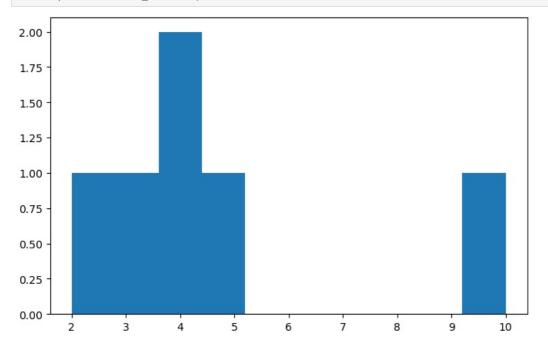
Salary

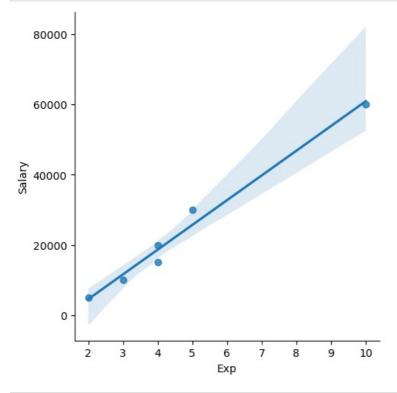


In [129... vis2 = plt.hist(clean_data['Salary'])

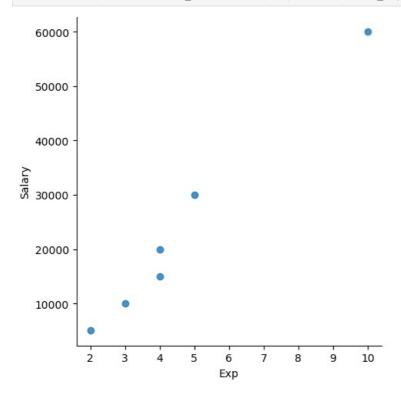


in [130... vis3 = plt.hist(clean_data['Exp'])

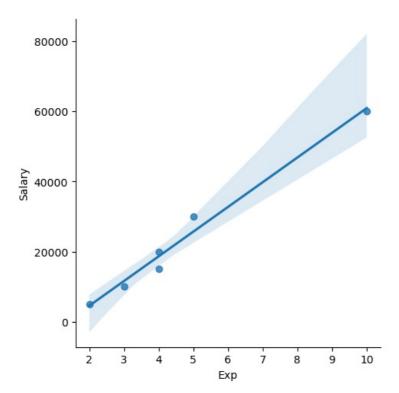




In [132_ vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)



In [135... vis6 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = True)



In [136... clean_data

Out[136]:

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

In [137... clean_data[:]

Out[137]:

	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 [138... clean_data[:2]

Out[138]:

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

In [139... clean_data[2:]

Out[139]:

		Name	Domain	Age	Location	Salary	Exp
	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 [140... clean_data[0:1]

Out[140]

:		Name	Domain	Age	Location	Salary	Exp
	0	Mike	Datascience	34	Mumbai	5000	2

v iv = clean data dron(['Salary'] avis=

```
In [142… clean_data
              Name
                        Domain Age Location Salary Exp
Out[142]:
               Mike
                    Datascience
                                       Mumbai
                                                5000
                                                        2
           1 Teddy
                                 45 Bangalore
                                              10000
                         Testing
           2 Umar
                                               15000
                     Dataanalyst
                                 50 Bangalore
                                                        4
           3
               Jane
                       Analytics
                                 50
                                     Hyderbad
                                               20000
                                                        4
            4 Uttam
                       Statistics
                                 67 Bangalore
                                               30000
                                                        5
                           NLP
                                         Delhi
                                               60000
                Kim
                                 55
                                                       10
In [143... x_iv
                                              Exp
                                    Location
Out[143]:
              Name
                        Domain Age
                                                 2
               Mike
                    Datascience
                                       Mumbai
           1 Teddy
                        Testing
                                 45 Bangalore
                                                 3
           2 Umar
                     Dataanalyst
                                 50 Bangalore
                                                 4
               Jane
                       Analytics
                                     Hyderbad
            4 Uttam
                       Statistics
                                 67 Bangalore
                                                 5
                                                10
                Kim
                           NLP
                                 55
                                         Delhi
In [144... x_iv.columns
           Index(['Name', 'Domain', 'Age', 'Location', 'Exp'], dtype='object')
Out[144]:
In [145... clean_data.columns
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[145]:
In [146... y_dv = clean_data.drop(['Name', 'Domain', 'Age', 'Location', 'Exp'],axis=1)
In [147... y_dv
Out[147]:
           Salary
            0 5000
           1 10000
           2 15000
           3 20000
           4 30000
           5 60000
In [148... clean_data
                                    Location Salary Exp
Out[148]:
              Name
                        Domain Age
           0 Mike
                                                5000
                    Datascience
                                       Mumbai
                                               10000
           1 Teddy
                                 45 Bangalore
                                                        3
                        Testing
           2 Umar
                     Dataanalyst
                                 50
                                     Bangalore
                                               15000
                                                        4
               Jane
                                               20000
                       Analytics
                                     Hyderbad
            4 Uttam
                       Statistics
                                               30000
                                 67 Bangalore
                                                        5
                Kim
                           NLP
                                 55
                                         Delhi
                                               60000
                                                       10
In [149... X_iv
                                              Exp
              Name
                                     Location
Out[149]:
                        Domain Age
               Mike
                     Datascience
                                       Mumbai
                                                 2
                                                 3
           1 Teddy
                        Testing
                                 45 Bangalore
           2 Umar
                     Dataanalyst
                                 50 Bangalore
                                                 4
               Jane
                       Analytics
                                 50
                                     Hyderbad
                                                 4
            4 Uttam
                       Statistics
                                 67 Bangalore
                                                 5
                           NLP
                Kim
                                 55
                                         Delhi
                                                10
```

TIL [141 | v_tv - crean uara urop([Sarary],axts-t/

In [150... y_dv

```
Salary
                5000
             1
                10000
             2
                15000
             3
                20000
             4
                30000
                60000
In [151...
            imputation = pd.get dummies(clean data)
In [152...
            imputation
                            Ехр
                Age Salary
                                  Name_Jane Name_Kim Name_Mike Name_Teddy Name_Umar Name_Uttam Domain_Analytics Domain_Dataanalyst
             0
                 34
                      5000
                               2
                                           0
                                                       0
                                                                   1
                                                                                0
                                                                                             0
                                                                                                           0
                                                                                                                             0
                                                                                                                                                 0
                 45
                     10000
                               3
                                           0
                                                       0
                                                                   0
                                                                                             0
                                                                                                           0
                                                                                                                             0
                                                                                                                                                 0
                                                       0
                                                                                0
                                                                                                           0
                                                                                                                             0
             2
                 50
                     15000
                               4
                                           0
                                                                   0
                                                                                             1
                                                                                                                                                 1
                                                       0
                                                                   0
                                                                                0
                                                                                                           0
                                                                                                                             1
             3
                 50
                     20000
                               4
                                                                                             0
                                                                                                                                                 0
                 67
                     30000
                               5
                                           0
                                                       0
                                                                   0
                                                                                0
                                                                                             0
                                                                                                           1
                                                                                                                             0
                                                                                                                                                 0
                 55
                     60000
                              10
                                           0
                                                                                             0
                                                                                                           0
                                                                                                                             0
In [153...
            clean_data.shape
             (6, 6)
Out[153]:
In [154...
            clean data.columns
             Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[154]:
            imputation.columns
In [155...
            'Domain_Dataanalyst', 'Domain_Datascience', 'Domain_NLP', 'Domain_Statistics', 'Domain_Testing', 'Location_Bangalore', 'Location_Delhi', 'Location_Hyderbad', 'Location_Mumbai'],
                    dtype='object')
           imputation.describe()
In [156...
                                                                                                        Name_Umar Name_Uttam Domain_Analytics
Out[156]:
                         Age
                                     Salary
                                                  Exp Name_Jane
                                                                   Name_Kim Name_Mike Name_Teddy
                                                          6.000000
                                                                                                                                           6.000000
             count
                    6.000000
                                   6.000000
                                             6.000000
                                                                     6.000000
                                                                                  6.000000
                                                                                               6.000000
                                                                                                            6.000000
                                                                                                                         6.000000
                                                                                  0.166667
             mean 50.166667 23333.333333
                                             4 666667
                                                          0.166667
                                                                     0.166667
                                                                                               0.166667
                                                                                                            0.166667
                                                                                                                         0.166667
                                                                                                                                           0.166667
               std
                    10.907184 19916.492328
                                             2.804758
                                                          0.408248
                                                                     0.408248
                                                                                  0.408248
                                                                                               0.408248
                                                                                                            0.408248
                                                                                                                         0.408248
                                                                                                                                           0.408248
                                             2.000000
                                                          0.000000
                                                                     0.000000
                                                                                 0.000000
                                                                                               0.000000
                                                                                                           0.000000
                                                                                                                         0.000000
                                                                                                                                           0.000000
               min
                    34.000000
                               5000.000000
                                                                                  0.000000
              25%
                    46.250000
                              11250.000000
                                             3.250000
                                                          0.000000
                                                                     0.000000
                                                                                               0.000000
                                                                                                            0.000000
                                                                                                                         0.000000
                                                                                                                                           0.000000
              50%
                    50.000000
                              17500.000000
                                             4.000000
                                                          0.000000
                                                                     0.000000
                                                                                  0.000000
                                                                                               0.000000
                                                                                                            0.000000
                                                                                                                         0.000000
                                                                                                                                           0.000000
                                                          0.000000
              75%
                    53.750000 27500.000000
                                             4.750000
                                                                     0.000000
                                                                                  0.000000
                                                                                               0.000000
                                                                                                            0.000000
                                                                                                                         0.000000
                                                                                                                                           0.000000
                   67.000000
                              60000.000000
                                            10.000000
                                                          1.000000
                                                                     1.000000
                                                                                  1.000000
                                                                                               1.000000
                                                                                                            1.000000
                                                                                                                         1.000000
                                                                                                                                           1.000000
              max
4
```

imputation.info()

In [157...

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 19 columns):
                            Non-Null Count Dtype
# Column
     Age 6 non-null
Salary 6 non-null
Exp 6 non-null
Name_Jane 6 non-null
Name_Kim 6 non-null
Name_Mike 6 non-null
Name_Teddy 6 non-null
Name_Umar 6 non-null
Name_Uttam 6 non-null
Domain_Analytics 6 non-null
Domain_Dataanalyst 6 non-null
                                 -----
0
                                                       int32
                                                       int32
 2
                                                       int32
 3
                                                      uint8
 4
                                                       uint8
 5
                                                       uint8
 6
                                                      uint8
 7
                                                       uint8
 8
                                                       uint8
 9
                                                       uint8
 10 Domain_Dataanalyst 6 non-null
                                                       uint8
 11 Domain Datascience 6 non-null
                                                       uint8
 12 Domain NLP
                                 6 non-null
                                                       uint8
13 Domain_Statistics 6 non-null
                                                       uint8
 14 Domain_Testing
                                 6 non-null
                                                       uint8
 15 Location Bangalore 6 non-null
                                                       uint8
16 Location_Delhi 6 non-null
17 Location_Hyderbad 6 non-null
18 Location_Mumbai 6 non-null
                                                       uint8
                                                       uint8
                                                       uint8
dtypes: int32(3), uint8(16)
memory usage: 296.0 bytes
```

In []:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js