Raw data to clean data conversion using python EDA

```
In [1]: import pandas as pd
        emp = pd.read_excel(r'E:\19th - EDA Practicle\19th - EDA Practicle\EDA- Practicle\Rawdata.xlsx')
In [2]:
In [3]: emp
Out[3]:
            Name
                         Domain
                                           Location
                                      Age
                                                        Salary
                                                                  Exp
                    Datascience#$ 34 years
             Mike
                                            Mumbai
                                                       5^00#0
        0
                                                                   2+
        1 Teddy^
                                    45' yr Bangalore 10%%000
                          Testing
                                                                   <3
        2 Uma#r Dataanalyst^^#
                                                     1$5%000
                                     NaN
                                               NaN
                                                                4> yrs
                      Ana^^lytics
                                     NaN Hyderbad
                                                      2000^0
        3
              Jane
                                                                 NaN
                                                       30000- 5+ year
        4 Uttam*
                         Statistics
                                     67-yr
                                               NaN
                                                     6000^$0
        5
              Kim
                             NLP
                                     55yr
                                               Delhi
                                                                  10+
In [4]:
        emp.columns
Out[4]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [5]: emp.shape
Out[5]: (6, 6)
In [6]: emp.head()
```

Out[6]:		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
	1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
	2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
	3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
	4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year

In [7]: emp.tail()

Out[7]:

	Name	Domain	Age	Location	Salary	Ехр
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^\$0	10+

In [8]: emp.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):

	(•
#	Column	Non-Null Count	Dtype
0	Name	6 non-null	object
1	Domain	6 non-null	object
2	Age	4 non-null	object
3	Location	4 non-null	object
4	Salary	6 non-null	object
5	Exp	5 non-null	object
1.4		(6)	

dtypes: object(6)

memory usage: 420.0+ bytes

```
In [9]: emp
Out[9]:
                                       Age Location
             Name
                           Domain
                                                         Salary
                                                                    Ехр
         0
              Mike
                      Datascience#$ 34 years
                                              Mumbai
                                                        5^00#0
                                                                     2+
         1 Teddy^
                            Testing
                                      45' yr Bangalore 10%%000
                                                                     <3
         2 Uma#r Dataanalyst^^#
                                      NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
               Jane
                        Ana^^lytics
                                      NaN Hyderbad
                                                        2000^0
                                                                   NaN
         3
                                                         30000- 5+ year
         4 Uttam*
                          Statistics
                                      67-yr
                                                 NaN
               Kim
                                                       6000^$0
         5
                              NLP
                                       55yr
                                                 Delhi
                                                                    10+
         emp['Domain']
In [10]:
               Datascience#$
Out[10]: 0
          1
                     Testing
              Dataanalyst^^#
          2
          3
                 Ana^^lytics
          4
                  Statistics
          5
                         NLP
         Name: Domain, dtype: object
         emp.isnull()
In [11]:
Out[11]:
            Name Domain Age Location Salary
                                                   Ехр
             False
                      False False
                                             False False
                                     False
             False
                      False False
                                             False False
                                     False
         2
             False
                      False
                           True
                                             False False
                                      True
             False
                      False True
                                             False True
         3
                                     False
```

False False

False False

True

False

False

5 False

False False

False False

```
In [12]: emp.isnull().sum()
Out[12]: Name
                     0
         Domain
                     0
         Age
         Location
         Salary
                     0
         Exp
                     1
         dtype: int64
         emp['Name']
In [13]:
Out[13]: 0
                Mike
              Teddy^
         1
               Uma#r
         2
          3
                Jane
         4
              Uttam*
          5
                 Kim
         Name: Name, dtype: object
In [14]: emp['Name'] = emp['Name'].str.replace(r'\W','', regex=True)
         emp['Name']
In [15]:
Out[15]: 0
               Mike
         1
              Teddy
         2
               Umar
          3
               Jane
         4
              Uttam
         5
                Kim
         Name: Name, dtype: object
         emp['Domain']
In [16]:
Out[16]: 0
               Datascience#$
                     Testing
         1
              Dataanalyst^^#
         2
                 Ana^^lytics
          3
         4
                  Statistics
                         NLP
         Name: Domain, dtype: object
```

```
emp['Domain'] = emp['Domain'].str.replace(r'\W','', regex=True)
         emp['Domain']
In [18]:
Out[18]: 0
               Datascience
          1
                  Testing
              Dataanalyst
          2
                Analytics
          3
          4
                Statistics
          5
                      NLP
         Name: Domain, dtype: object
         emp
In [19]:
Out[19]:
                                        Location
                                  Age
                                                     Salary
                                                               Ехр
            Name
                      Domain
             Mike Datascience 34 years
                                         Mumbai
                                                    5^00#0
                                                                2+
         1 Teddy
                       Testing
                                 45' yr Bangalore
                                                  10%%000
                                                                <3
         2 Umar Dataanalyst
                                  NaN
                                            NaN
                                                   1$5%000
                                                             4> yrs
             Jane
                      Analytics
                                  NaN Hyderbad
                                                    2000^0
                                                               NaN
          3
                                                     30000- 5+ year
         4 Uttam
                      Statistics
                                  67-yr
                                             NaN
              Kim
                          NLP
                                                   6000^$0
                                                               10+
          5
                                  55yr
                                            Delhi
In [20]:
         emp['Age']
Out[20]: 0
               34 years
          1
                45' yr
          2
                   NaN
          3
                   NaN
          4
                 67-yr
          5
                  55yr
         Name: Age, dtype: object
         emp['Age'] = emp['Age'].str.replace(r'\W','', regex=True)
In [21]:
In [22]: emp['Age']
```

```
Out[22]: 0
              34years
                 45yr
         1
                  NaN
         2
          3
                  NaN
          4
                 67yr
                 55yr
          5
         Name: Age, dtype: object
In [23]:
         emp['Age'] = emp['Age'].str.extract('(\\d+)')
         emp['Age']
In [24]:
Out[24]: 0
               34
               45
         1
          2
              NaN
          3
              NaN
         4
               67
          5
               55
         Name: Age, dtype: object
In [25]:
         emp
Out[25]:
            Name
                      Domain Age
                                     Location
                                                 Salary
                                                           Exp
            Mike Datascience
                                34
                                     Mumbai
                                                5^00#0
                                                            2+
         1 Teddy
                       Testing
                                45 Bangalore
                                              10%%000
                                                            <3
         2 Umar Dataanalyst NaN
                                        NaN
                                               1$5%000
                                                         4> yrs
             Jane
                     Analytics NaN Hyderbad
                                                2000^0
                                                          NaN
         3
                                                30000- 5+ year
         4 Uttam
                     Statistics
                                67
                                         NaN
              Kim
                         NLP
                                55
                                        Delhi
                                               6000^$0
         5
                                                           10+
         emp['Location'] = emp['Location'].str.replace(r'\W','')
In [26]:
In [28]: emp['Location']
```

```
Out[28]: 0
                 Mumbai
              Bangalore
         1
                    NaN
         2
               Hyderbad
          3
          4
                    NaN
          5
                  Delhi
         Name: Location, dtype: object
In [29]: emp['Salary']
               5^00#0
Out[29]: 0
         1
              10%%000
          2
              1$5%000
               2000^0
          3
               30000-
          4
              6000^$0
         Name: Salary, dtype: object
In [30]: emp['Salary'] = emp['Salary'].str.replace(r'\W','', regex=True)
In [31]: emp['Salary']
Out[31]: 0
               5000
              10000
         1
              15000
          2
          3
              20000
              30000
          4
              60000
         Name: Salary, dtype: object
In [32]:
         emp
```

```
Out[32]:
           Name
                    Domain Age Location Salary
                                                     Ехр
        0 Mike Datascience
                                   Mumbai
                             34
                                            5000
                                                      2+
                              45 Bangalore 10000
        1 Teddy
                     Testing
                                                      <3
        2 Umar Dataanalyst NaN
                                      NaN 15000
                                                   4> yrs
                    Analytics NaN Hyderbad 20000
        3 Jane
                                                     NaN
         4 Uttam
                    Statistics
                              67
                                      NaN 30000 5+ year
         5 Kim
                       NLP
                              55
                                      Delhi 60000
                                                     10+
In [33]: emp['Exp']
Out[33]: 0
                  2+
                  <3
         1
         2
              4> yrs
         3
                 NaN
         4
             5+ year
                 10+
        Name: Exp, dtype: object
In [34]: emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
In [35]: emp['Exp']
Out[35]: 0
               2
         1
               3
         2
               4
             NaN
         3
         4
               5
              10
        Name: Exp, dtype: object
In [36]:
```

```
Out[36]:
                                   Location Salary Exp
            Name
                     Domain Age
            Mike Datascience
                               34
                                    Mumbai
                                              5000
                                                      2
         1 Teddy
                      Testing
                               45 Bangalore
                                             10000
            Umar
                  Dataanalyst NaN
                                       NaN
                                             15000
                     Analytics NaN Hyderbad
                                             20000 NaN
             Jane
         4 Uttam
                     Statistics
                               67
                                             30000
                                                      5
                                       NaN
             Kim
                        NLP
                                       Delhi 60000
                               55
                                                     10
```

In [37]: clean_data = emp.copy()

missing values treatment for numerical data

```
clean_data
In [38]:
Out[38]:
            Name
                      Domain Age
                                    Location Salary
             Mike Datascience
                                     Mumbai
                                               5000
                                                       2
                                45 Bangalore
         1 Teddy
                       Testing
                                              10000
                   Dataanalyst NaN
            Umar
                                        NaN
                                              15000
                     Analytics NaN Hyderbad
             Jane
                                              20000 NaN
         4 Uttam
                     Statistics
                                67
                                        NaN
                                              30000
                                                       5
              Kim
                         NLP
                                55
                                        Delhi
                                              60000
                                                      10
         clean_data['Age']
```

```
Out[39]: 0
               34
               45
          1
          2
              NaN
          3
              NaN
          4
               67
               55
          5
         Name: Age, dtype: object
In [40]: import numpy as np
In [41]:
         clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])))
In [42]: clean_data['Age']
Out[42]: 0
                  34
                 45
          1
          2
              50.25
          3
              50.25
                 67
          4
                 55
          5
         Name: Age, dtype: object
In [43]: clean_data['Exp']
Out[43]: 0
                2
          1
                3
          2
                4
          3
              NaN
                5
          4
          5
               10
         Name: Exp, dtype: object
         clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp'])))
In [45]:
         clean_data['Exp']
```

```
Out[45]: 0
                2
          1
                 3
          2
                4
          3
              4.8
          4
                5
               10
         Name: Exp, dtype: object
In [46]:
         clean_data
Out[46]:
                                Age Location Salary Exp
                      Domain
            Name
             Mike Datascience
                                       Mumbai
                                                 5000
                                                         2
                                                10000
         1 Teddy
                       Testing
                                     Bangalore
                                 45
             Umar Dataanalyst 50.25
                                                15000
                                          NaN
                                                         4
             Jane
                      Analytics 50.25 Hyderbad
                                                20000
                                                       4.8
         4 Uttam
                      Statistics
                                 67
                                          NaN
                                                30000
                                                         5
              Kim
                          NLP
                                 55
                                          Delhi
                                                60000
                                                        10
         clean_data['Location'].isnull().sum()
In [47]:
Out[47]: 2
         clean_data['Location']
In [48]:
Out[48]: 0
                 Mumbai
              Bangalore
         1
          2
                    NaN
          3
               Hyderbad
          4
                    NaN
                  Delhi
         Name: Location, dtype: object
         clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode()[0])
In [49]:
         clean_data['Location']
In [50]:
```

```
Out[50]: 0
                  Mumbai
          1
               Bangalore
               Bangalore
          2
          3
               Hyderbad
          4
               Bangalore
          5
                   Delhi
          Name: Location, dtype: object
         clean_data
In [51]:
Out[51]:
                                     Location Salary Exp
             Name
                       Domain
                                Age
                                       Mumbai
              Mike Datascience
                                                  5000
                                                          2
                                                 10000
         1 Teddy
                        Testing
                                      Bangalore
                                                          3
             Umar Dataanalyst 50.25
                                      Bangalore
                                                 15000
                                                          4
                      Analytics 50.25 Hyderbad
                                                 20000
          3
              Jane
                                                        4.8
          4 Uttam
                      Statistics
                                      Bangalore
                                                 30000
                                                          5
              Kim
          5
                          NLP
                                  55
                                          Delhi
                                                 60000
                                                         10
         clean_data.info()
In [52]:
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
             Column
                       Non-Null Count Dtype
             Name
                       6 non-null
                                        object
                                       object
         1
             Domain
                       6 non-null
         2
             Age
                       6 non-null
                                       object
             Location 6 non-null
                                       object
             Salary
         4
                       6 non-null
                                        object
                                       object
             Exp
                       6 non-null
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [53]: clean_data['Age'] = clean_data['Age'].astype(int)
```

```
In [54]: clean_data.info()
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
                      Non-Null Count Dtype
            Column
           -----
            Name
                      6 non-null
                                     object
            Domain 6 non-null
                                    object
        2
                      6 non-null
                                    int32
            Age
            Location 6 non-null
                                    object
                      6 non-null
            Salary
                                     object
        5
                      6 non-null
            Exp
                                      object
       dtypes: int32(1), object(5)
       memory usage: 396.0+ bytes
In [55]: clean_data['Salary'] = clean_data['Salary'].astype(int)
         clean_data['Exp'] = clean_data['Exp'].astype(int)
         clean_data.info()
In [56]:
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
                      Non-Null Count Dtype
            Column
            Name
                      6 non-null
                                     object
            Domain 6 non-null
        1
                                   object
            Age
                      6 non-null
                                    int32
            Location 6 non-null
                                    object
            Salary
                      6 non-null
                                     int32
                      6 non-null
            Exp
                                      int32
       dtypes: int32(3), object(3)
       memory usage: 348.0+ bytes
         clean data['Name'] = clean_data['Name'].astype('category')
In [57]:
         clean_data['Domain'] = clean_data['Domain'].astype('category')
         clean_data['Location'] = clean_data['Location'].astype('category')
         clean_data.info()
In [58]:
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
              Non-Null Count Dtype
    Column
    -----
              6 non-null
 0
    Name
                              category
    Domain 6 non-null
                              category
 1
    Age
              6 non-null
                             int32
    Location 6 non-null
                              category
    Salary
              6 non-null
                              int32
    Exp
              6 non-null
 5
                              int32
dtypes: category(3), int32(3)
memory usage: 866.0 bytes
```

In [59]: clean_data

_		
\cap	150	
Uut	.))	

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

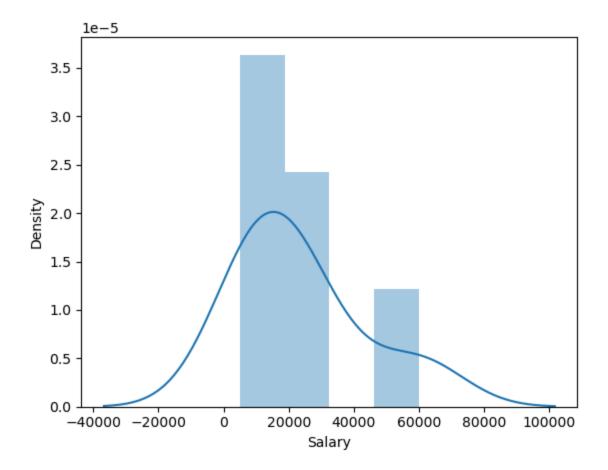
```
In [60]: clean_data.to_csv('clean_data.csv')
```

```
In [61]: import os
    os.getcwd()
```

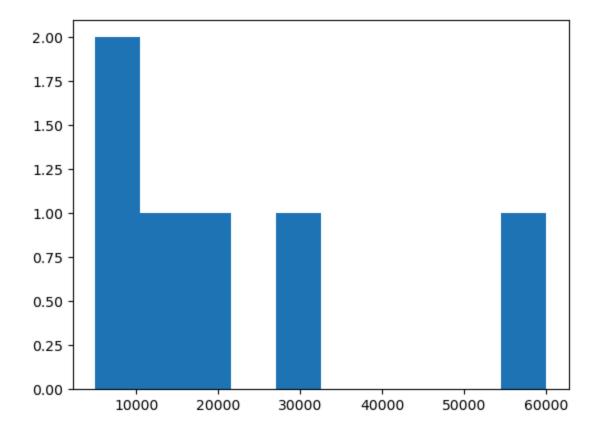
Out[61]: 'C:\\Users\\91911'

In [62]: clean_data

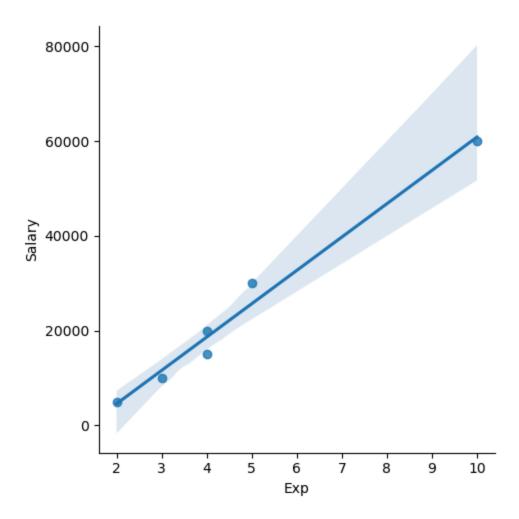
```
Out[62]:
                     Domain Age Location Salary Exp
            Name
             Mike Datascience
                                    Mumbai
                                              5000
         0
                               34
                                                      2
                               45 Bangalore
                                             10000
         1 Teddy
                       Testing
                                                      3
            Umar Dataanalyst
                               50 Bangalore
                                             15000
                                                      4
                               50 Hyderbad 20000
         3
             Jane
                     Analytics
                                                      4
         4 Uttam
                     Statistics
                               67 Bangalore
                                             30000
                                                      5
              Kim
         5
                         NLP
                               55
                                       Delhi 60000
                                                     10
         import matplotlib.pyplot as plt # visualization
In [64]:
         import seaborn as sns
         import warnings
In [65]:
         warnings.filterwarnings('ignore')
In [66]:
         clean_data['Salary']
Out[66]: 0
               5000
         1
              10000
              15000
         2
         3
              20000
         4
              30000
              60000
         Name: Salary, dtype: int32
In [67]: vis1 = sns.distplot(clean_data['Salary'])
```



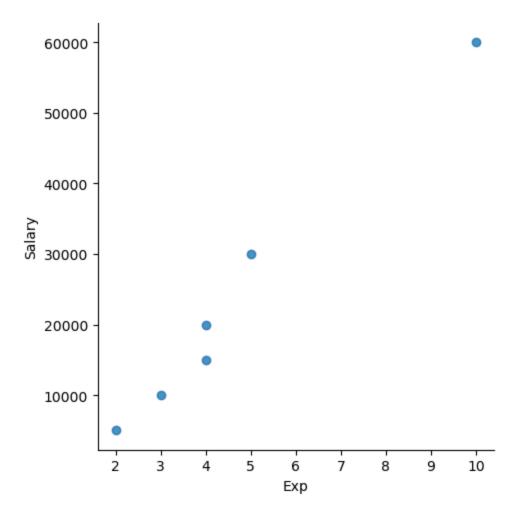
In [68]: vis2 = plt.hist(clean_data['Salary'])



In [69]: vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary')



In [70]: vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)



In [71]: clean_data[:]

Out[71]:		Name	Domain	Δαρ	Location	Salary	Ехр
00.0[, =],		Itallic	Domain	Age	Location	Salai y	LVA
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
2 Umar		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 [72]:	cl	ean_dat	a[0:6:2]				
Out[72]:		Name	Domain	Age	Location	Salary	Ехр

Mike Datascience Mumbai 5000 Dataanalyst 50 Bangalore 15000 **2** Umar 67 Bangalore 30000 4 Uttam Statistics 5

clean_data[::-1] In [73]:

Out[73]: Name Domain Age Location Salary Exp

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

clean_data.columns In [74]:

```
Out[74]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [75]: X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [76]: X_iv
Out[76]:
                     Domain Age Location Exp
           Name
         0 Mike Datascience
                             34 Mumbai
                                             2
         1 Teddy
                              45 Bangalore
                     Testing
                                             3
         2 Umar Dataanalyst
                              50 Bangalore
         3 Jane
                    Analytics
                              50 Hyderbad
         4 Uttam
                    Statistics
                              67 Bangalore
                                             5
                                           10
         5 Kim
                       NLP
                              55
                                     Delhi
In [77]: y_dv = clean_data[['Salary']]
In [78]: y_dv
Out[78]:
           Salary
         0
           5000
         1 10000
         2 15000
         3 20000
         4 30000
         5 60000
In [79]:
        emp
```

Out[79]:		Name	Domain	Age	Location	Salary	Ехр
Out[79]:	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	NaN	NaN	15000	4
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN
	4	Uttam	Statistics	67	NaN	30000	5
	5	Kim	NLP	55	Delhi	60000	10

In [80]: clean_data

Out[80]:

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

In [81]: **X_iv**

Out[81]:		Name	Domain	Age	Location	Ехр
Out[81]:	0	Mike	Datascience	34	Mumbai	2
	1	Teddy	Testing	45	Bangalore	3
2		Umar	Dataanalyst	50	Bangalore	4
	3	Jane	Analytics	50	Hyderbad	4
	4	Uttam	Statistics	67	Bangalore	5
	5	Kim	NLP	55	Delhi	10

In [82]: **y_dv**

Out[82]:

Salary

- **o** 5000
- **1** 10000
- **2** 15000
- **3** 20000
- **4** 30000
- **5** 60000

In [83]: clean_data

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

In [84]: imputation = pd.get_dummies(clean_data)

In [85]: imputation

Out[85]:

:		Age	Salary	Ехр	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Domain_Analytics	Domain_
	0	34	5000	2	False	False	True	False	False	False	False	
	1	45	10000	3	False	False	False	True	False	False	False	
	2	50	15000	4	False	False	False	False	True	False	False	
	3	50	20000	4	True	False	False	False	False	False	True	
	4	67	30000	5	False	False	False	False	False	True	False	
	5	55	60000	10	False	True	False	False	False	False	False	

In [86]: clean_data

Out[86]:		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	50	Bangalore	15000	4
	3	Jane	Analytics	50	Hyderbad	20000	4
	4	Uttam	Statistics	67	Bangalore	30000	5

55

Delhi 60000 10

NLP

5 Kim