```
In [1]:
         import pandas as pd
In [2]:
         pd.__version__
Out[2]: '2.2.2'
        emp=pd.read_excel(r'/Users/bhavanichimmili/Downloads/Rawdata.xlsx')
In [3]:
In [4]:
         emp
Out[4]:
                           Domain
             Name
                                       Age
                                             Location
                                                         Salary
                                                                    Exp
                                                        5^00#0
              Mike
                     Datascience#$
                                   34 years
                                              Mumbai
                                                                     2+
         0
         1
            Teddy^
                           Testing
                                     45' yr
                                            Bangalore 10%%000
                                                                     <3
                    Dataanalyst^^#
         2
            Uma#r
                                       NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
         3
                       Ana^^lytics
                                            Hyderbad
                                                        2000^0
              Jane
                                       NaN
                                                                    NaN
                         Statistics
                                      67-yr
         4
            Uttam*
                                                 NaN
                                                         30000-
                                                                 5+ year
                              NLP
         5
               Kim
                                      55yr
                                                Delhi
                                                       6000^$0
                                                                    10+
        id(emp)
In [5]:
Out[5]:
         5027204656
In [6]:
         emp.columns
Out[6]:
         Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='obj
         ect')
In [7]:
         emp.shape
Out[7]:
         (6, 6)
In [8]:
         emp.head()
Out[8]:
             Name
                           Domain
                                       Age
                                             Location
                                                         Salary
                                                                    Exp
         0
              Mike
                     Datascience#$
                                   34 years
                                              Mumbai
                                                        5^00#0
                                                                     2+
            Teddy^
                           Testing
                                     45' yr
                                            Bangalore
                                                      10%%000
                                                                     <3
                    Dataanalyst^^#
         2
            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
In [9]:
        emp.tail()
```

Out[9]:	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
		Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
		Uttam*	Statistics	67-yr	NaN	30000-	5+ year
	5	Kim	NLP	55yr	Delhi	6000^\$0	10+

# In [10]: emp.info() # information of the dataframe

<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
d+vr	ace object	(6)	

dtypes: object(6)

memory usage: 420.0+ bytes

In [11]: emp

Out[11]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^\$0	10+

# In [12]: emp.isnull()

Out[12]:

	Name	Domain	Age	Location	Salary	Ехр
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	True	False	False
3	False	False	True	False	False	True
4	False	False	False	True	False	False
5	False	False	False	False	False	False

In [13]: emp.isna()

```
Out[13]:
             Name Domain
                              Age Location Salary
                                                     Exp
          0
              False
                       False False
                                      False
                                              False False
              False
                       False False
                                      False
                                              False False
          1
          2
              False
                      False
                             True
                                       True
                                              False False
          3
              False
                      False
                                              False
                                                   True
                             True
                                      False
          4
              False
                      False False
                                       True
                                              False False
          5
              False
                       False False
                                      False
                                              False False
In [14]: emp.isnull().sum()
Out[14]:
          Name
                       0
          Domain
                       2
          Age
          Location
                       2
          Salary
                       0
                       1
          Exp
          dtype: int64
In [15]: emp.columns
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='obj
Out[15]:
          ect')
In [16]:
Out[16]:
              Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                      Exp
          0
               Mike
                      Datascience#$ 34 years
                                               Mumbai
                                                          5^00#0
                                                                       2+
             Teddy^
                            Testing
                                       45' yr Bangalore 10%%000
                                                                       <3
          2
              Uma#r
                     Dataanalyst^^#
                                        NaN
                                                        1$5%000
                                                  NaN
                                                                   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+
          DATA CLEANING
In [18]: emp['Name']
Out[18]:
          0
                  Mike
          1
                Teddy^
          2
                 Uma#r
          3
                  Jane
          4
                Uttam*
          5
                   Kim
          Name: Name, dtype: object
          emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
In [19]:
          emp['Name']
In [20]:
```

```
Out[20]: 0
                Mike
          1
               Teddy
          2
                Umar
          3
                 Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
In [21]: emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
In [22]:
          emp['Domain']
Out[22]:
          0
               Datascience
          1
                    Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
          5
                        NLP
          Name: Domain, dtype: object
In [23]:
          emp
Out[23]:
                                          Location
             Name
                       Domain
                                    Age
                                                      Salary
                                                                Exp
          0
              Mike Datascience 34 years
                                           Mumbai
                                                     5^00#0
                                                                  2+
          1 Teddy
                        Testing
                                  45' yr
                                         Bangalore
                                                   10%%000
                                                                  <3
             Umar
                    Dataanalyst
                                              NaN
          2
                                   NaN
                                                    1$5%000
                                                              4> yrs
          3
              Jane
                       Analytics
                                   NaN
                                         Hyderbad
                                                     2000^0
                                                                NaN
          4
            Uttam
                      Statistics
                                   67-yr
                                              NaN
                                                     30000-
                                                             5+ year
          5
               Kim
                           NLP
                                   55yr
                                             Delhi
                                                    6000^$0
                                                                 10+
In [24]: emp['Age'] = emp['Age'].str.replace(r'\\\','',regex=True)
In [25]:
          emp['Age']
Out[25]:
          0
               34years
          1
                   45yr
          2
                   NaN
          3
                    NaN
          4
                   67yr
          5
                   55yr
          Name: Age, dtype: object
In [26]: emp['Age'] = emp['Age'].str.extract('(\\d+)')
In [27]:
          emp['Age']
Out[27]:
                34
          0
          1
                 45
          2
               NaN
          3
               NaN
          4
                67
          5
                 55
          Name: Age, dtype: object
```

```
In [28]: emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
In [29]: emp['Location']
Out[29]:
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
               Hyderbad
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
In [30]: emp['Salary'] = emp['Salary'].str.replace(r'\W','', regex=True)
In [31]: emp['Salary']
Out[31]:
                5000
          1
               10000
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: object
In [32]:
         emp.head()
Out[32]:
             Name
                       Domain Age
                                     Location Salary
                                                        Exp
          0
             Mike Datascience
                                34
                                      Mumbai
                                               5000
                                                         2+
          1 Teddy
                       Testing
                                45
                                    Bangalore 10000
                                                         <3
          2
             Umar
                   Dataanalyst NaN
                                         NaN 15000
                                                      4> yrs
             Jane
                      Analytics NaN Hyderbad 20000
                                                        NaN
          4 Uttam
                      Statistics
                                67
                                         NaN 30000 5+ year
In [33]: emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
         emp['Exp']
In [34]:
Out[34]:
          0
                 2
                 3
          1
          2
                 4
          3
               NaN
          4
                 5
          5
                10
          Name: Exp, dtype: object
In [35]:
         emp
```

Out[35]:		Name	Domain	Age	Location	Salary	Exp
	<b>0</b> 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 [36]: clean\_data = emp.copy()

In [37]: clean\_data

#### Out[37]: Name Domain Age **Location Salary** Exp 0 Mike Datascience 34 Mumbai 5000 2 Teddy Bangalore 10000 Testing 45 3 2 Dataanalyst NaN 15000 4 Umar NaN 3 Analytics NaN Hyderbad 20000 NaN Jane 4 Uttam **Statistics** 67 NaN 30000 5 5 NLP 55 Delhi 60000 Kim 10

### EDA technique

In [39]: clean\_data

Out[39]: Name Domain Age **Location Salary** Exp 5000 0 Mike Datascience 34 Mumbai 2 Teddy Testing 45 Bangalore 10000 3 2 Umar Dataanalyst NaN 15000 4 NaN 3 Analytics NaN Hyderbad 20000 NaN Jane Statistics 30000 4 Uttam 67 NaN 5 5 Kim NLP 55 Delhi 60000 10

In [40]: clean\_data.isnull().sum()

Out[40]: Name 0
Domain 0
Age 2
Location 2
Salary 0
Exp 1
dtype: int64

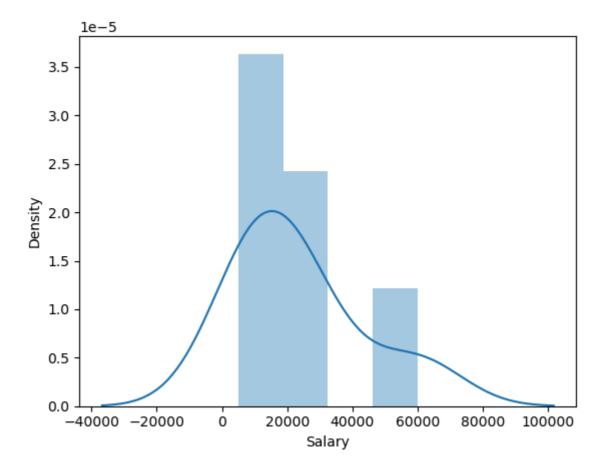
In [41]: clean\_data['Age']

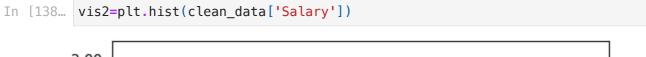
```
Out[41]: 0
                34
          1
                45
          2
               NaN
          3
               NaN
          4
                67
          5
                55
          Name: Age, dtype: object
In [42]: import numpy as np
In [43]: clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_
In [44]: clean_data['Age']
Out[44]: 0
                  34
          1
                  45
          2
               50.25
          3
               50.25
          4
                  67
          5
                  55
          Name: Age, dtype: object
In [45]: clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_
In [46]: clean_data['Exp']
Out[46]:
          0
                 2
                 3
          1
          2
                 4
          3
               4.8
                 5
                10
          Name: Exp, dtype: object
In [47]: clean_data
Out[47]:
             Name
                       Domain
                                Age
                                      Location Salary Exp
                                                         2
          0
              Mike Datascience
                                 34
                                       Mumbai
                                                5000
          1 Teddy
                                 45 Bangalore
                       Testing
                                               10000
                                                         3
                    Dataanalyst 50.25
          2 Umar
                                          NaN
                                               15000
                                                         4
              Jane
                      Analytics 50.25
                                     Hyderbad 20000
                                                       4.8
          4 Uttam
                      Statistics
                                  67
                                          NaN 30000
                                                         5
                                 55
                                                        10
                          NLP
                                         Delhi 60000
          5
              Kim
In [48]: clean_data['Location']=clean_data['Location'].fillna(clean_data['Location')]
In [49]: clean_data['Location']
```

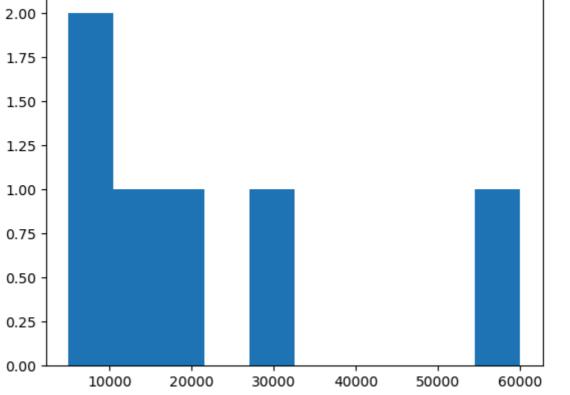
```
Out[49]:
          0
                  Mumbai
               Bangalore
          1
          2
               Bangalore
          3
                Hyderbad
          4
               Bangalore
          5
                   Delhi
          Name: Location, dtype: object
In [50]:
         clean_data
Out [50]:
             Name
                       Domain
                                      Location Salary Exp
                                 Age
          0
              Mike Datascience
                                  34
                                       Mumbai
                                                 5000
                                                         2
            Teddy
                        Testing
                                  45 Bangalore
                                                10000
                                                         3
          2
             Umar
                    Dataanalyst 50.25
                                     Bangalore
                                                         4
                                                15000
          3
              Jane
                      Analytics 50.25
                                      Hyderbad
                                               20000
                                                       4.8
          4
            Uttam
                      Statistics
                                      Bangalore
                                                30000
                                                         5
                                  67
                          NLP
                                          Delhi 60000
          5
               Kim
                                  55
                                                        10
In [51]: clean_data.info()
        <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
         0
             Domain
                        6 non-null
                                         object
         1
         2
             Age
                        6 non-null
                                         object
         3
             Location 6 non-null
                                         object
         4
             Salary
                        6 non-null
                                         object
                        6 non-null
         5
             Exp
                                         object
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [52]: emp.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
         #
             Column
                        Non-Null Count Dtype
         0
             Name
                        6 non-null
                                         object
         1
             Domain
                        6 non-null
                                         object
         2
                        4 non-null
             Age
                                         object
         3
                        4 non-null
             Location
                                         object
         4
             Salary
                        6 non-null
                                         object
         5
             Exp
                        5 non-null
                                         object
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [53]:
         clean_data['Age'] = clean_data['Age'].astype(int)
         clean_data['Age']
In [54]:
```

```
Out[54]: 0
              34
              45
         1
         2
              50
         3
              50
         4
              67
         5
              55
         Name: Age, dtype: int64
In [55]: 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
                       6 non-null
         1
            Domain
                                       object
         2
            Age
                       6 non-null
                                       int64
         3
            Location 6 non-null
                                       object
             Salary
                       6 non-null
                                       object
         5
                       6 non-null
                                       object
             Exp
        dtypes: int64(1), object(5)
        memory usage: 420.0+ bytes
In [56]: clean_data['Salary'] = clean_data['Salary'].astype(int)
In [57]: clean_data['Exp'] = clean_data['Exp'].astype(int)
In [58]: 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
                       6 non-null
         1
            Domain
                                       object
         2
                       6 non-null
                                       int64
             Age
         3
             Location 6 non-null
                                       object
                       6 non-null
                                       int64
         4
             Salary
         5
                       6 non-null
                                       int64
             Exp
        dtypes: int64(3), object(3)
        memory usage: 420.0+ bytes
In [59]: 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 [60]: clean_data.info()
```

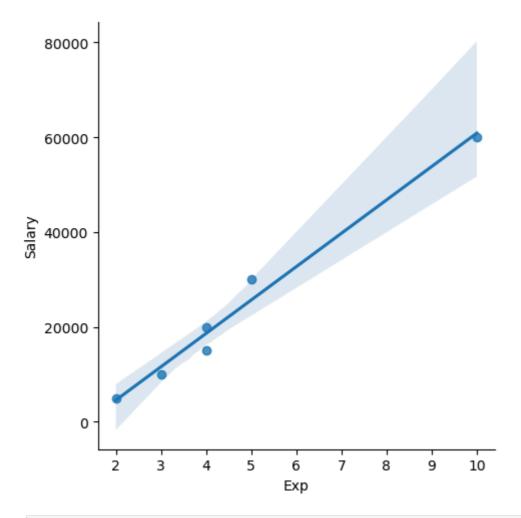
```
<class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
             Column
                       Non-Null Count Dtype
         0
             Name
                        6 non-null
                                        category
         1
             Domain
                      6 non-null
                                        category
         2
                        6 non-null
             Age
                                        int64
         3
             Location 6 non-null
                                        category
         4
             Salary
                        6 non-null
                                        int64
         5
                        6 non-null
                                        int64
             Exp
        dtypes: category(3), int64(3)
        memory usage: 938.0 bytes
In [61]: clean_data.to_csv('clean_data.csv')
In [62]:
         import os
         os.getcwd()
Out[62]: '/Users/bhavanichimmili'
In [126...
         clean data
Out [126...
             Name
                                     Location Salary Exp
                       Domain Age
          0
             Mike Datascience
                                34
                                     Mumbai
                                               5000
                                                       2
          1 Teddy
                       Testing
                                45 Bangalore 10000
          2
             Umar
                   Dataanalyst
                                    Bangalore
                                             15000
                                                       4
          3
             Jane
                      Analytics
                                50
                                    Hyderbad 20000
           Uttam
                      Statistics
                                67
                                    Bangalore 30000
                                                       5
               Kim
                          NLP
                                55
                                        Delhi 60000
                                                      10
In [128...
         import matplotlib.pyplot as plt
         import seaborn as sns
In [130...
         import warnings
         warnings.filterwarnings('ignore')
In [132... clean_data['Salary']
Out [132...
          0
                5000
               10000
          1
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: int64
In [134... vis1=sns.distplot(clean_data['Salary'])
```



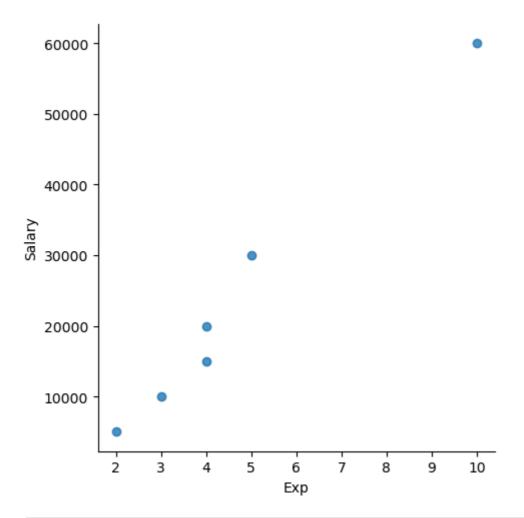




In [140... vis4=sns.lmplot(data=clean\_data,x='Exp',y='Salary')



In [142... vis5=sns.lmplot(data=clean\_data,x='Exp',y='Salary',fit\_reg=False)



In [144... clean\_data[:]

Ο.		Γ α	Л	Л	
UI	JT	1.1	4	4	

	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 [146... clean\_data[0:6:2]

# Out[146...

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

# In [148...

clean\_data[::-1]

```
Out [148...
             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
                                    Bangalore 15000
                                50
                                                        4
             Teddy
                       Testing
                                45
                                    Bangalore 10000
                                                        2
              Mike
                   Datascience
                                 34
                                      Mumbai
                                              5000
In [150... clean_data.columns
Out[150... Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='obj
          ect')
In [156... X_iv= clean_data[['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp']]
In [158... X_iv
Out[158...
             Name
                       Domain Age
                                     Location Salary Exp
          0
             Mike Datascience
                                34
                                      Mumbai
                                              5000
                                                        2
          1 Teddy
                       Testing
                                45
                                    Bangalore 10000
                                                        3
          2 Umar
                    Dataanalyst
                                     Bangalore
                                              15000
                                                        4
                                50
          3
              Jane
                      Analytics
                                 50
                                     Hyderbad 20000
          4 Uttam
                      Statistics
                                     Bangalore 30000
                                 67
                                                        5
          5
               Kim
                          NLP
                                 55
                                        Delhi 60000
                                                       10
In [160... y_dv = clean_data[['Salary']]
In [162... y_dv
Out [162...
             Salary
          0
             5000
          1 10000
          2 15000
          3 20000
          4 30000
          5 60000
In [164... emp
```

5, 11:22 AM					E	DA TECHNI	IQUES
Out[164		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	NaN	NaN	15000	4
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN
	4	Uttam	Statistics	67	NaN	30000	5
	5	Kim	NLP	55	Delhi	60000	10
In [166	cl	ean dat	ta				
Out[166		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
	.,						
	X_	1V					
Out[168		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 [170	У_	dv					
Out [170	7_	Salary					
00.0 [ 2 / 0	0	5000	_				
	1	10000					
	2	15000					
	3	20000					
	4	30000					

**5** 60000

In	[	]:	<pre>imputation = p</pre>
In	[	]:	
In	[	]:	