```
import pandas as pd
In [1]:
In [2]:
        pd.__version__
Out[2]: '2.1.4'
         emp=pd.read_excel(r"D:\DS_NIT\Oct_24\EDA.xlsx")
In [3]:
In [4]:
        emp
Out[4]:
             Name
                           Domain
                                             Location
                                       Age
                                                          Salary
                                                                    Exp
         0
              Mike
                      Datascience#$
                                    34 years
                                              Mumbai
                                                         5^00#0
                                                                     2+
         1 Teddy^
                                      45' yr
                                            Bangalore
                                                       10%%000
                                                                     <3
                            Testing
         2
            Uma#r
                    Dataanalyst^^#
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
                                       NaN
                       Ana^^lytics
         3
                                       NaN Hyderbad
                                                         2000^0
                                                                    NaN
              Jane
            Uttam*
                          Statistics
                                                 NaN
                                                         30000-
                                                                 5+ year
                                      67-yr
               Kim
                              NLP
                                       55yr
                                                 Delhi
                                                        6000^$0
                                                                    10+
In [5]:
        emp.shape
Out[5]: (6, 6)
        emp.columns
In [6]:
Out[6]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [7]:
        emp.head()
Out[7]:
             Name
                           Domain
                                       Age
                                             Location
                                                          Salary
                                                                    Exp
                                    34 years
                                                         5^00#0
         0
              Mike
                      Datascience#$
                                              Mumbai
                                                                     2+
           Teddy^
                                            Bangalore
                                                       10%%000
                                                                     <3
                           Testing
                                      45' yr
                    Dataanalyst^^#
            Uma#r
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
                                       NaN
                       Ana^^lytics
                                       NaN Hyderbad
                                                         2000^0
              Jane
                                                                    NaN
         4 Uttam*
                          Statistics
                                      67-yr
                                                 NaN
                                                         30000- 5+ year
In [8]: emp.tail()
```

```
Out[8]:
              Name
                           Domain
                                     Age
                                           Location
                                                        Salary
                                                                  Exp
          1 Teddy^
                            Testing
                                    45' yr
                                           Bangalore 10%%000
                                                                   <3
             Uma#r
                     Dataanalyst^^#
                                     NaN
                                               NaN
                                                      1$5%000
                                                                4> yrs
          3
               Jane
                        Ana^^lytics
                                     NaN
                                           Hyderbad
                                                       2000^0
                                                                  NaN
                                                       30000- 5+ year
             Uttam*
                                    67-yr
                           Statistics
                                               NaN
          5
                Kim
                               NLP
                                     55yr
                                               Delhi
                                                      6000^$0
                                                                  10+
 In [9]: emp.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
             Column
                        Non-Null Count Dtype
             -----
                        -----
         0
             Name
                        6 non-null
                                        object
         1
             Domain
                        6 non-null
                                        object
         2
             Age
                        4 non-null
                                        object
         3
             Location 4 non-null
                                        object
         4
             Salary
                        6 non-null
                                        object
         5
             Exp
                        5 non-null
                                        object
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [10]: emp['Domain']
Out[10]: 0
                Datascience#$
                      Testing
          1
          2
               Dataanalyst^^#
                  Ana^^lytics
          3
          4
                   Statistics
                          NLP
          Name: Domain, dtype: object
In [11]: emp.isnull()
Out[11]:
             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
                            True
                                      False
                                              False
                                                   True
              False
                       False False
                                       True
                                              False False
              False
                       False False
                                      False
                                              False False
In [12]: emp.isnull().sum()
```

```
Out[12]: Name
          Domain
                      2
          Age
          Location
                      2
                      0
          Salary
          Exp
          dtype: int64
In [13]:
         emp['Name']
Out[13]: 0
                 Mike
          1
               Teddy^
          2
               Uma#r
          3
                 Jane
               Uttam*
          5
                  Kim
          Name: Name, dtype: object
In [14]: emp['Name']=emp['Name'].str.replace(r'\W','')
         emp['Name']
In [15]:
Out[15]:
                 Mike
               Teddy^
          1
          2
               Uma#r
          3
                 Jane
               Uttam*
          4
          5
                  Kim
          Name: Name, dtype: object
In [16]: emp['Name']=emp['Name'].str.replace(r'\W','',regex=True)
In [17]: emp['Name']
Out[17]:
          0
                Mike
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
In [18]: emp.columns
Out[18]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [19]:
          emp.Domain
Out[19]:
                Datascience#$
          1
                      Testing
          2
               Dataanalyst^^#
                  Ana^^lytics
          3
          4
                   Statistics
                          NLP
          Name: Domain, dtype: object
```

```
In [20]: emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
In [21]: emp['Domain']
Out[21]: 0
               Datascience
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
          5
                       NLP
          Name: Domain, dtype: object
In [22]: emp.columns
Out[22]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [23]:
         emp.Age
Out[23]:
               34 years
                 45' yr
          1
          2
                    NaN
          3
                    NaN
          4
                  67-yr
                   55yr
          Name: Age, dtype: object
In [24]: emp['Age']=emp['Age'].str.extract(r'(\d+)')
In [25]: emp['Age']
Out[25]:
          0
                34
                45
          2
               NaN
          3
               NaN
                67
                55
          Name: Age, dtype: object
In [26]: emp.columns
Out[26]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [27]:
         emp.Location
Out[27]:
          0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
                Hyderbad
          4
                     NaN
                   Delhi
          Name: Location, dtype: object
In [28]: emp.Salary
```

```
Out[28]: 0
                5^00#0
          1
               10%%000
          2
               1$5%000
                2000^0
          3
          4
                30000-
          5
               6000^$0
          Name: Salary, dtype: object
In [29]: emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
In [30]:
         emp['Salary']
Out[30]:
                5000
               10000
          1
          2
               15000
          3
               20000
               30000
          4
          5
               60000
          Name: Salary, dtype: object
In [31]:
         emp.Exp
Out[31]:
                    2+
                     <3
          1
          2
                4> yrs
          3
                   NaN
          4
               5+ year
          5
                   10+
          Name: Exp, dtype: object
In [32]: emp['Exp']=emp['Exp'].str.extract(r'(\d+)')
In [33]: emp['Exp']
                 2
Out[33]:
          0
                 3
          2
                 4
          3
               NaN
                 5
          4
          5
                10
          Name: Exp, dtype: object
In [34]: emp.head()
Out[34]:
             Name
                       Domain Age
                                      Location Salary
                                                        Exp
          0
              Mike Datascience
                                  34
                                       Mumbai
                                                 5000
                                                          2
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                10000
                                                15000
                                                          4
          2
              Umar Dataanalyst NaN
                                          NaN
                       Analytics NaN
              Jane
                                      Hyderbad
                                                20000
                                                       NaN
          4 Uttam
                                                30000
                                                          5
                       Statistics
                                  67
                                          NaN
```

```
clean_data=emp.copy()
In [35]:
         clean_data
In [36]:
Out[36]:
                                                        Ехр
                       Domain Age
                                      Location Salary
             Name
          0
              Mike Datascience
                                       Mumbai
                                                  5000
                                                          2
                                  34
                                                          3
             Teddy
                        Testing
                                      Bangalore
                                                 10000
          2
              Umar
                    Dataanalyst NaN
                                           NaN
                                                15000
                                                          4
          3
              Jane
                       Analytics NaN
                                      Hyderbad
                                                20000
                                                        NaN
                                                 30000
                                                          5
             Uttam
                       Statistics
                                  67
                                           NaN
          5
               Kim
                           NLP
                                  55
                                                          10
                                          Delhi
                                                60000
In [37]:
          import numpy as np
In [38]:
         clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])
In [39]:
          clean_data['Age']
Out[39]:
          0
                  34
          1
                  45
          2
               50.25
               50.25
          3
                  67
          4
          5
                  55
          Name: Age, dtype: object
         clean_data['Location']
In [40]:
Out[40]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
                Hyderbad
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
         clean_data['Location']=clean_data['Location'].fillna(clean_data['Location'].mode()[
In [41]:
In [42]:
          clean_data['Location']
Out[42]:
          0
                  Mumbai
               Bangalore
          1
          2
               Bangalore
          3
                Hyderbad
               Bangalore
          4
          5
                   Delhi
          Name: Location, dtype: object
```

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

```
<class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
            Column Non-Null Count Dtype
        --- -----
                    -----
        0
           Name
                     6 non-null
                                     object
        1
            Domain 6 non-null
                                     object
        2
           Age
                     6 non-null
                                     object
        3
            Location 6 non-null
                                     object
        4
                     6 non-null
                                     object
            Salary
        5
                     6 non-null
                                     object
            Exp
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [48]: clean_data['Age']=clean_data['Age'].astype(int)
         clean_data['Salary']=clean_data['Salary'].astype(int)
         clean_data['Exp']=clean_data['Exp'].astype(int)
         clean_data['Name']=clean_data['Name'].astype('category')
In [49]:
         clean data['Domain']=clean data['Domain'].astype('category')
         clean_data['Location']=clean_data['Location'].astype('category')
In [50]: 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
                                    category
        1
            Domain 6 non-null
                                   category
        2
           Age
                     6 non-null
                                     int32
            Location 6 non-null
        3
                                     category
        4
            Salary
                     6 non-null
                                     int32
        5
                     6 non-null
            Exp
                                     int32
       dtypes: category(3), int32(3)
       memory usage: 866.0 bytes
In [51]: clean_data
Out[51]:
            Name
                     Domain Age
                                   Location Salary Exp
         0
             Mike Datascience
                              34
                                   Mumbai
                                             5000
                                                    2
                                  Bangalore
                                                    3
            Teddy
                      Testing
                              45
                                            10000
         2
            Umar Dataanalyst
                              50 Bangalore
                                                    4
                                            15000
         3
                    Analytics
                                  Hyderbad
                                            20000
             Jane
                                                    4
                                                    5
         4 Uttam
                     Statistics
                                  Bangalore
                                            30000
                              67
              Kim
                        NLP
                              55
                                      Delhi
                                            60000
                                                   10
In [52]: clean_data.to_csv('clean_datapr.csv')
```

```
In [53]:
         import os
         os.getcwd()
Out[53]: "C:\\Users\\evenk\\OneDrive\\Desktop\\DS_NIT\\Oct'24"
In [54]:
         clean_data
Out[54]:
                                      Location Salary Exp
             Name
                       Domain Age
                                                         2
              Mike Datascience
                                 34
                                       Mumbai
                                                 5000
             Teddy
                        Testing
                                 45
                                     Bangalore
                                                10000
                                                         3
          2
             Umar
                    Dataanalyst
                                     Bangalore
                                                15000
                                                         4
          3
              Jane
                      Analytics
                                     Hyderbad
                                                20000
                                                         4
                                                         5
            Uttam
                       Statistics
                                     Bangalore
                                                30000
                          NLP
                                                60000
          5
               Kim
                                 55
                                         Delhi
                                                        10
         LETS APPLY EDA TECHNIQUES
         import matplotlib.pyplot as plt
In [55]:
          import seaborn as sns
In [56]:
         import warnings
         warnings.filterwarnings('ignore')
         clean_data
In [57]:
Out[57]:
             Name
                       Domain Age
                                      Location Salary
                                                       Exp
              Mike Datascience
                                 34
                                       Mumbai
                                                 5000
                                                         2
                                     Bangalore
                                                10000
                                                         3
             Teddy
                        Testing
                    Dataanalyst
                                     Bangalore
                                                15000
                                                         4
             Umar
                                     Hyderbad
                                                20000
              Jane
                      Analytics
                                     Bangalore
                                                30000
                                                         5
            Uttam
                       Statistics
                                 67
          5
               Kim
                          NLP
                                 55
                                          Delhi
                                                60000
                                                        10
```

clean_data['Salary']

In [58]:

```
Out[58]: 0 5000

1 10000

2 15000

3 20000

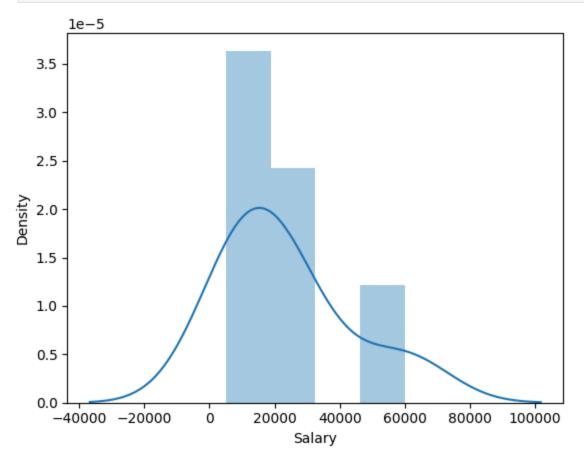
4 30000

5 60000

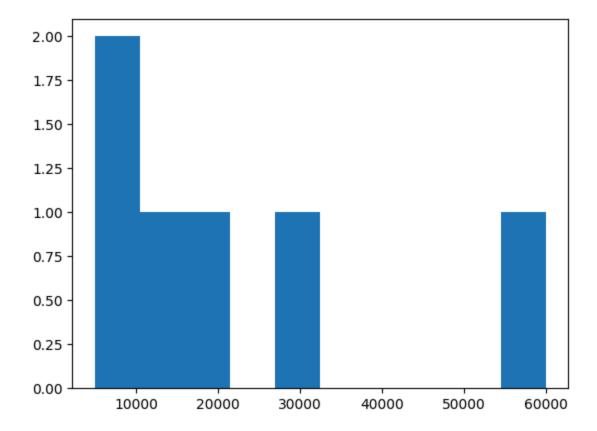
Name: Salary, dtype: int32
```

Visualization

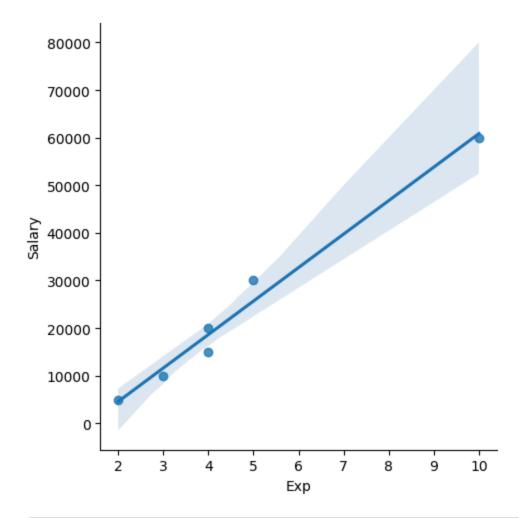
```
In [59]: #univariate
    vis1=sns.distplot(clean_data['Salary'])
```



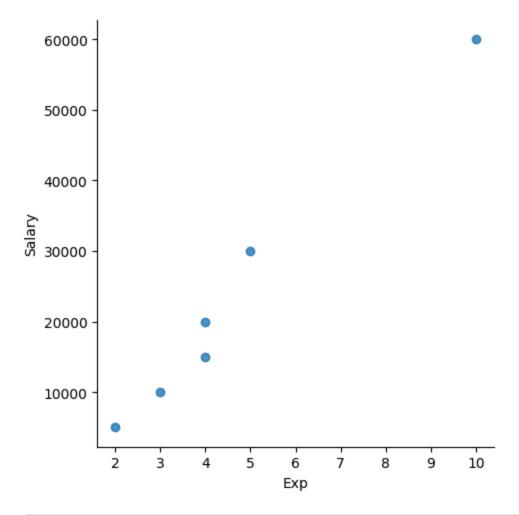
```
In [60]: #outlier
vis2=plt.hist(clean_data['Salary'])
```



In [61]: #bivariate
vis3=sns.lmplot(data=clean_data,x='Exp',y='Salary')



In [62]: vis3=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=False)



In [63]: clean_data

_			_		
\cap	14-	16	$^{\circ}$	- 1	
Uι	<i>.</i>	ΙU	-		

	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 [64]: clean_data[:]

Out[64]:	Name Domain		Ag	Age Location		n S	Salary	E	хр	
	0	Mike	Datascience	3	34	Mumba	ai	5000		2
	1	Teddy	Testing	2	15 E	Bangalor	e ´	10000		3
	2	Umar	Dataanalyst	5	50 E	Bangalor	e ´	15000		4
	3	Jane	Analytics	5	50 I	Hyderbad		20000		4
	4	Uttam	Statistics	6	57 E	Bangalore		30000		5
	5	Kim	NLP	5	55	Delhi		60000		10
In [65]:	cl	ean_dat	a[1:6:2]							
Out[65]:		Name	Domain A	Age	Lo	cation	Sala	ary I	Ехр	
	1	Teddy	Testing	45	Ban	galore	100	000	3	_
	3	Jane	Analytics	50	Нус	lerbad	200	000	4	
	5	Kim	NLP	55		Delhi	600	000	10	
In [66]:	cl	ean_dat	a[2:6]							
Out[66]:		Name	Domain	Ag	e	Location	n S	alary	Ex	кр
	2	Umar	Dataanalyst	5	0 B	angalore	e 1	15000		4
	3	Jane	Analytics	5	0 F	Hyderbad	d 2	20000		4
	4	Uttam	Statistics	6	57 B	angalore	e 3	30000		5
	5	Kim	NLP	5	5	Delh	ni 6	50000		10
In [67]:	cl	ean_dat	a[:]							
Out[67]:		Name	Domain	Ag	je	Locatio	n S	Salary	E	хр
	0	Mike	Datascience	3	34	Mumba	ai	5000		2
	1	Teddy	Testing	4	15 E	Bangalor	e ´	10000		3
	2	Umar	Dataanalyst	5	50 E	Bangalor	e ´	15000		4
	3	Jane	Analytics	5	50 I	Hyderba	d 2	20000		4
	4	Uttam	Statistics	6	57 E	Bangalor	e 3	30000		5
	5	Kim	NLP	5	55	Delh	ni 6	60000		10
In [68]:	cl	ean_dat	a[::-1]							

```
Domain Age
                                       Location Salary
                                                        Exp
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                 60000
                                                         10
                       Statistics
                                      Bangalore
                                                 30000
                                                          5
             Uttam
                                  67
          3
               Jane
                       Analytics
                                  50
                                      Hyderbad
                                                 20000
                                                          4
          2
              Umar
                    Dataanalyst
                                      Bangalore
                                                 15000
                                                          4
          1
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                          3
              Mike Datascience
                                                  5000
                                                          2
                                  34
                                       Mumbai
In [69]: clean_data.columns
Out[69]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [70]: #variable identification(independent variable)
          x_iv=clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [71]: x_iv
Out[71]:
             Name
                       Domain Age
                                       Location Exp
                                  34
                                                   2
          0
              Mike Datascience
                                       Mumbai
             Teddy
                                  45
                                      Bangalore
                                                   3
                        Testing
                    Dataanalyst
              Umar
                                  50
                                      Bangalore
          3
                       Analytics
                                      Hyderbad
               Jane
                                  50
                                      Bangalore
                                                   5
            Uttam
                       Statistics
                                  67
          5
                           NLP
                                  55
                                          Delhi
                                                  10
               Kim
In [72]: #Variable identification(Dependent variable)
         y_iv=clean_data[['Salary']]
In [73]: y_iv
Out[73]:
             Salary
          0
              5000
             10000
             15000
             20000
             30000
          5 60000
```

Out[68]:

Name

In [74]: clean_data

Out[74]:

	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 [75]: **x_iv**

Out[75]:

	Name	Domain	Age	Location	Ехр
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 [76]: **y_iv**

Out[76]: Salary

o 5000

1 10000

2 15000

3 20000

4 30000

5 60000

In [77]: emp

Out[77]:		Name	e D	omain	Age	Locati	on Salar	у Ехр	_			
	0	Mike	e Datas	science	34	Muml	bai 500	0 2				
	1	Teddy	/	Testing	45	Bangalo	ore 1000	0 3				
	2	Umai	r Dataa	analyst	NaN	N	aN 1500	0 4				
	3	Jane	e Ar	nalytics	NaN	Hyderb	ad 2000	0 NaN				
	4	Uttam	n Sta	atistics	67	N	aN 3000	0 5				
	5	Kim	1	NLP	55	De	elhi 6000	0 10				
In [78]:	c1	ean_da	ta									
Out[78]:	-				Λ	l a cati	on Color					
οατ[76].	0	Name Mike		omain science	Age 34	Mumb	on Salar oai 500					
	1	Teddy		Testing	45	Bangalo						
	2	Umai		analyst		Bangalo						
	3	Jane		nalytics	50	Hyderb						
	4	Uttam		atistics	67	Bangalo						
	5	Kim		NLP	55	De						
In [83]:	im	putati	on=pd.	get_du	mmies((clean_d	lata)					
In [84]:	im	putati	.on									
Out[84]:		Age	Salary	Ехр	Name _.	_Jane N	lame_Kim	Name	_Mike	Name_Teddy	Name_Umar	Nan
	0	34	5000	2		False	False		True	False	False	
	1	45	10000	3		False	False		False	True	False	
	2	50	15000	4		False	False		False	False	True	
	3	50	20000	4		True	False		False	False	False	
	4	67	30000	5		False	False		False	False	False	
	5	55	60000	10		False	True		False	False	False	
	4		_									
In [85]:	im	putati	on=pd . ۽	get_du	mmies(clean_d	lata,dtyp	e=int)				
In [86]:	im	putati	.on									

Out[86]:		Age	Salary	Ехр	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Nan
	0	34	5000	2	0	0	1	0	0	
	1	45	10000	3	0	0	0	1	0	
	2	50	15000	4	0	0	0	0	1	
	3	50	20000	4	1	0	0	0	0	
	4	67	30000	5	0	0	0	0	0	
	5	55	60000	10	0	1	0	0	0	
	4		_	-						
In []:										