```
In [1]:
        import pandas as pd
In [2]: pd.__version__
Out[2]: '2.0.3'
        emp = pd.read_excel("Rawdata.xlsx")
In [3]:
In [4]: emp
Out[4]:
                          Domain
                                            Location
            Name
                                      Age
                                                        Salary
                                                                   Exp
         0
              Mike
                     Datascience#$ 34 years
                                             Mumbai
                                                        5^00#0
                                                                    2+
                                     45' yr Bangalore
           Teddy^
                           Testing
                                                      10%%000
                                                                    <3
            Uma#r Dataanalyst^^#
                                   NaN
         2
                                                NaN
                                                      1$5%000
                                                                 4> yrs
                                   NaN Hyderbad
         3
              Jane
                       Ana^^lytics
                                                        2000^0
                                                                  NaN
            Uttam*
                          Statistics
                                     67-yr
                                                NaN
                                                        30000-
                                                                5+ year
               Kim
                              NLP
                                      55yr
                                                Delhi
                                                      6000^$0
                                                                   10+
In [5]: id(emp)
Out[5]: 1677872730384
In [6]: emp.columns
Out[6]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
        emp.shape
In [7]:
Out[7]: (6, 6)
In [8]: emp.head()
Out[8]:
             Name
                          Domain
                                      Age
                                            Location
                                                        Salary
                                                                   Exp
         0
              Mike
                     Datascience#$ 34 years
                                             Mumbai
                                                       5^00#0
                                                                    2+
         1 Teddy^
                           Testing
                                     45' yr
                                           Bangalore
                                                      10%%000
                                                                    <3
           Uma#r Dataanalyst^^#
         2
                                      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
                                                                Exp
         1 Teddy^
                           Testing 45' yr Bangalore 10%%000
                                                                <3
            Uma#r Dataanalyst^^#
                                   NaN
                                                              4> yrs
                                             NaN
                                                   1$5%000
                       Ana^^lytics
                                   NaN Hyderbad
                                                     2000^0
                                                               NaN
        3
              Jane
         4 Uttam*
                          Statistics
                                  67-yr
                                             NaN
                                                     30000- 5+ year
         5
                                             Delhi
                                                   6000^$0
                                                               10+
              Kim
                             NLP
                                   55yr
```

In [10]: emp.info()

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

#	Column	Non-Null Count	Dtype
0	Name	6 non-null	object
1	Domain	6 non-null	object
2	Age	4 non-null	object
3	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 [11]: emp.isnull()

Out[11]: Name Domain Age Location Salary

	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 [12]: emp.isna()

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

```
emp.isnull().sum()
In [13]:
Out[13]: Name
                      0
          Domain
                     0
                     2
          Age
          Location 2
          Salary
                     0
          Exp
                      1
          dtype: int64
In [14]: emp['Name']
Out[14]: 0
                Mike
             Teddy^
          2
               Uma#r
          3
                 Jane
          4
              Uttam*
                  Kim
          Name: Name, dtype: object
In [15]: emp['Domain']
Out[15]: 0
                Datascience#$
          1
                      Testing
          2
            Dataanalyst^^#
          3
                 Ana^^lytics
          4
                   Statistics
          5
                          NLP
          Name: Domain, dtype: object
In [16]:
         emp['Age']
Out[16]: 0
               34 years
          1
                45' yr
          2
                    NaN
          3
                    NaN
          4
                  67-yr
          5
                   55yr
          Name: Age, dtype: object
In [17]: emp['Location']
Out[17]: 0
                 Mumbai
          1
              Bangalore
          2
                     NaN
              Hyderbad
          3
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
In [18]:
         emp['Salary']
Out[18]: 0
                5^00#0
          1
              10%%000
          2
              1$5%000
          3
              2000^0
          4
                30000-
          5
               6000^$0
          Name: Salary, dtype: object
```

```
emp[['Name','Domain']]
In [19]:
Out[19]:
                             Domain
               Name
           0
                Mike
                        Datascience#$
              Teddy^
                              Testing
           2
              Uma#r
                      Dataanalyst^^#
                Jane
                          Ana^^lytics
              Uttam*
                             Statistics
           5
                                 NLP
                 Kim
          emp[['Name','Domain','Age']]
Out[20]:
               Name
                             Domain
                                          Age
           0
                Mike
                        Datascience#$
                                      34 years
              Teddy^
                              Testing
                                         45' yr
                      Dataanalyst^^#
              Uma#r
                                          NaN
           3
                          Ana^^lytics
                                          NaN
                Jane
                                         67-yr
              Uttam*
                             Statistics
           5
                 Kim
                                 NLP
                                          55yr
In [21]:
          emp[['Name','Domain','Age','Location','Exp']]
Out[21]:
               Name
                             Domain
                                          Age
                                                 Location
                                                              Exp
           0
                Mike
                        Datascience#$
                                      34 years
                                                               2+
                                                  Mumbai
             Teddy^
                                                               <3
                              Testing
                                         45' yr
                                                Bangalore
                      Dataanalyst^^#
           2
              Uma#r
                                          NaN
                                                     NaN
                                                            4> yrs
           3
                          Ana^^lytics
                                          NaN
                                                Hyderbad
                Jane
                                                              NaN
           4
              Uttam*
                             Statistics
                                         67-yr
                                                     NaN
                                                           5+ year
           5
                                 NLP
                 Kim
                                          55yr
                                                    Delhi
                                                              10+
```

Data Cleaning Or Data Cleansing

In [22]: emp['Name'] #regex = #\$%^

```
Out[22]: 0
               Mike
          1
              Teddy^
          2
               Uma#r
          3
                 Jane
          4
               Uttam*
          5
                  Kim
          Name: Name, dtype: object
In [23]: emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
In [24]:
         emp['Name']
Out[24]: 0
               Mike
               Teddy
          2
               Umar
          3
                Jane
          4
               Uttam
                 Kim
          Name: Name, dtype: object
In [25]:
         emp
Out[25]:
             Name
                          Domain
                                      Age
                                             Location
                                                        Salary
                                                                   Exp
                     Datascience#$ 34 years
          0
              Mike
                                             Mumbai
                                                        5^00#0
                                                                    2+
                                            Bangalore 10%%000
             Teddy
                           Testing
                                     45' yr
                                                                    <3
                    Dataanalyst^^#
                                                NaN
                                                      1$5%000
          2
             Umar
                                      NaN
                                                                 4> yrs
                       Ana^^lytics
                                            Hyderbad
                                                        2000^0
                                                                   NaN
          3
              Jane
                                      NaN
                          Statistics
                                                        30000-
          4
             Uttam
                                      67-yr
                                                NaN
                                                                5+ year
                              NLP
                                      55yr
                                                       6000^$0
          5
               Kim
                                                Delhi
                                                                   10+
In [26]:
         emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
In [27]:
         emp['Domain']
Out[27]: 0
               Datascience
          1
                   Testing
          2
             Dataanalyst
          3
               Analytics
          4
                Statistics
          5
                       NLP
          Name: Domain, dtype: object
In [28]: emp['Age']=emp['Age'].str.replace(r'\W','',regex=True)
In [29]:
         emp['Age']
Out[29]: 0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
          5
                  55yr
          Name: Age, dtype: object
```

```
In [30]: emp['Age']=emp['Age'].str.extract('(\\d+)') #r(r'(\\d+)')
In [31]: emp['Age']
Out[31]: 0
               34
         1
               45
         2
              NaN
         3
              NaN
         4
               67
         5
               55
         Name: Age, dtype: object
In [32]: emp['Location']=emp['Location'].str.replace(r'\W','',regex=True)
In [33]:
        emp['Location']
Out[33]: 0
                Mumbai
              Bangalore
         1
         2
                    NaN
         3
              Hyderbad
                    NaN
                   Delhi
         Name: Location, dtype: object
In [34]: emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
         emp['Salary']
In [35]:
Out[35]:
         0
               5000
         1
              10000
         2
              15000
         3
              20000
         4
              30000
         5
              60000
         Name: Salary, dtype: object
In [36]: emp['Exp']=emp['Exp'].str.replace(r'\W','',regex=True)
In [37]:
         emp['Exp']
Out[37]: 0
                   2
         1
                   3
         2
               4yrs
         3
               NaN
               5year
         5
                 10
         Name: Exp, dtype: object
In [38]:
        emp['Exp']=emp['Exp'].str.extract('(\d+)')
In [39]: emp['Exp']
```

```
Out[39]: 0 2
            NaN
              5
               10
         Name: Exp, dtype: object
In [40]: clean_data = emp.copy()
In [41]: clean_data
Out[41]:
                      Domain Age
            Name
                                   Location Salary
            Mike Datascience
                                34
                                     Mumbai
                                               5000
                                                       2
                                45 Bangalore
         1 Teddy
                       Testing
                                              10000
             Umar
                   Dataanalyst NaN
                                              15000
                                        NaN
                     Analytics NaN
                                    Hyderbad
                                              20000 NaN
         3
             Jane
            Uttam
                     Statistics
                                        NaN
                                              30000
                                                       5
                                67
                         NLP
                                55
                                       Delhi
                                              60000
                                                      10
              Kim
```

Clean missing value treatment

In [42]:	em	р					
Out[42]:		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 [43]:	cl	ean_dat	a				

```
Out[43]:
                                    Name
                                                               Domain Age
                                                                                                     Location Salary
                                                                                                                                                         Exp
                            0
                                      Mike Datascience
                                                                                            34
                                                                                                           Mumbai
                                                                                                                                       5000
                                                                                                                                                               2
                                    Teddy
                                                                                            45 Bangalore
                                                                                                                                     10000
                                                                                                                                                               3
                                                                   Testing
                            2
                                     Umar
                                                      Dataanalyst NaN
                                                                                                                                    15000
                                                                                                                                                               4
                                                                                                                    NaN
                            3
                                                              Analytics NaN
                                                                                                       Hyderbad
                                                                                                                                    20000 NaN
                                       Jane
                                  Uttam
                                                               Statistics
                                                                                                                    NaN
                                                                                                                                    30000
                                                                                                                                                               5
                                                                                            67
                                          Kim
                                                                         NLP
                                                                                             55
                                                                                                                   Delhi
                                                                                                                                    60000
                                                                                                                                                             10
In [44]:
                          clean_data.isnull().sum()
Out[44]:
                                                              0
                           Name
                            Domain
                                                              2
                            Age
                                                             2
                            Location
                            Salary
                                                             0
                            Exp
                                                              1
                            dtype: int64
In [45]: clean_data.info()
                       <class 'pandas.core.frame.DataFrame'>
                       RangeIndex: 6 entries, 0 to 5
                       Data columns (total 6 columns):
                         #
                                     Column
                                                            Non-Null Count Dtype
                                    ----
                                                                 -----
                          0
                                  Name
                                                             6 non-null
                                                                                                               object
                          1 Domain 6 non-null
                                                                                                              object
                                                              4 non-null
                                                                                                              object
                          2
                                  Age
                          3
                                   Location 4 non-null
                                                                                                               object
                                     Salary
                                                                 6 non-null
                                                                                                               object
                          5
                                     Exp
                                                                 5 non-null
                                                                                                               object
                       dtypes: object(6)
                       memory usage: 420.0+ bytes
In [46]: import numpy as np
                           clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age'])).fillna(np.to_numer
In [47]:
                          clean_data['Age']
In [48]:
Out[48]: 0
                                                  34
                            1
                                                  45
                            2
                                         50.25
                            3
                                          50.25
                            4
                                                  67
                            5
                                                  55
                            Name: Age, dtype: object
In [49]: | clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']
In [50]: clean_data['Exp']
```

```
3
                  4
          3
                4.8
                  5
          4
                 10
          Name: Exp, dtype: object
         clean_data
In [51]:
Out[51]:
                                         Location Salary Exp
             Name
                        Domain
                                  Age
              Mike
          0
                     Datascience
                                    34
                                         Mumbai
                                                    5000
                                                             2
                                                   10000
             Teddy
                         Testing
                                    45
                                        Bangalore
                                                             3
                                 50.25
          2
              Umar
                     Dataanalyst
                                             NaN
                                                   15000
                                                             4
          3
               Jane
                        Analytics 50.25
                                        Hyderbad
                                                   20000
                                                           4.8
          4
             Uttam
                        Statistics
                                    67
                                             NaN
                                                   30000
                                                             5
                            NLP
                                                   60000
          5
               Kim
                                    55
                                            Delhi
                                                            10
In [52]:
          clean_data['Location'].isnull().sum()
Out[52]: 2
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mod
In [53]:
In [54]:
          clean_data
Out[54]:
             Name
                        Domain
                                  Age
                                         Location Salary Exp
          0
              Mike
                     Datascience
                                    34
                                         Mumbai
                                                    5000
                                                             2
          1
              Teddy
                         Testing
                                    45
                                        Bangalore
                                                   10000
                                                             3
          2
              Umar
                     Dataanalyst
                                 50.25
                                        Bangalore
                                                   15000
                                                             4
          3
               Jane
                        Analytics
                                 50.25
                                        Hyderbad
                                                   20000
                                                           4.8
          4
             Uttam
                        Statistics
                                    67
                                        Bangalore
                                                   30000
                                                             5
          5
                            NLP
                                    55
                                                   60000
                Kim
                                            Delhi
                                                            10
          emp.info()
In [55]:
```

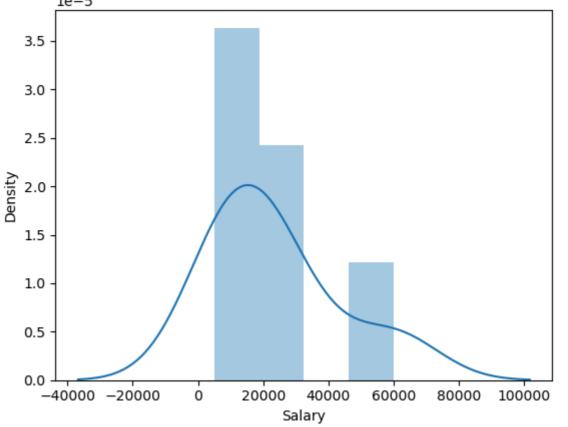
Out[50]: 0

2

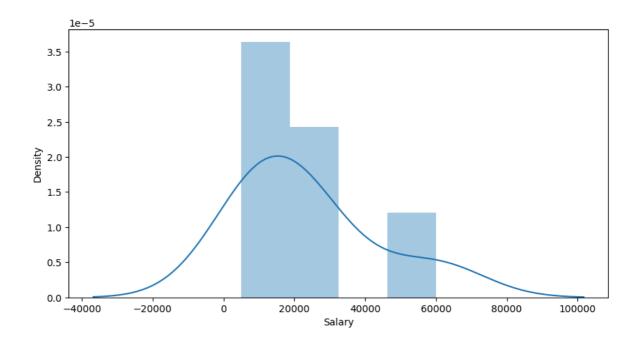
```
<class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          # Column Non-Null Count Dtype
          --- ----- -----
          0 Name 6 non-null object
1 Domain 6 non-null object
2 Age 4 non-null object
3 Location 4 non-null object
4 Salary 6 non-null object
5 Exp 5 non-null object
          dtypes: object(6)
         memory usage: 420.0+ bytes
In [56]: clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          # Column Non-Null Count Dtype
          --- -----
          0 Name 6 non-null object
1 Domain 6 non-null object
2 Age 6 non-null object
3 Location 6 non-null object
4 Salary 6 non-null object
          5 Exp
                          6 non-null
                                             object
          dtypes: object(6)
         memory usage: 420.0+ bytes
In [57]: clean_data['Age']= clean_data['Age'].astype(int)
In [58]: clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          # Column Non-Null Count Dtype
          0 Name 6 non-null object
1 Domain 6 non-null object
2 Age 6 non-null int32
3 Location 6 non-null object
4 Salary 6 non-null object
5 Exp 6 non-null object
          --- -----
                          -----
          dtypes: int32(1), object(5)
          memory usage: 396.0+ bytes
In [59]: clean_data['Salary']= clean_data['Salary'].astype(int)
In [60]: clean data['Exp']= clean data['Exp'].astype(int)
In [61]: clean_data.info()
```

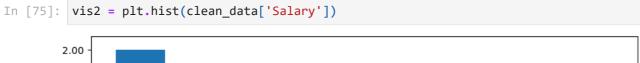
```
<class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
           Column Non-Null Count Dtype
        ---
                     -----
         0 Name 6 non-null object
1 Domain 6 non-null object
         2 Age 6 non-null
                                      int32
         3 Location 6 non-null
                                       object
         4
            Salary 6 non-null
                                       int32
         5
                                       int32
             Exp
                      6 non-null
        dtypes: int32(3), object(3)
        memory usage: 348.0+ bytes
In [62]: clean_data['Name']= clean_data['Name'].astype('category')
In [63]: clean data['Domain']= clean data['Domain'].astype('category')
In [64]: clean_data['Location']= clean_data['Location'].astype('category')
In [65]: 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 Age 6 non-null int32
3 Location 6 non-null category
            Salary
                      6 non-null
                                       int32
         5
                                       int32
             Exp
                       6 non-null
        dtypes: category(3), int32(3)
        memory usage: 866.0 bytes
In [66]: clean_data
Out[66]:
            Name
                      Domain Age Location Salary Exp
             Mike Datascience
                                                5000
                                                        2
         0
                                34
                                      Mumbai
            Teddy
                       Testing
                                45
                                    Bangalore
                                               10000
                                                        3
         1
         2
             Umar
                   Dataanalyst
                                50
                                    Bangalore
                                               15000
                                                        4
                      Analytics
                                 50
                                    Hyderbad
                                               20000
         3
              Jane
                                               30000
                                                        5
         4
            Uttam
                      Statistics
                                67
                                    Bangalore
                          NLP
                                55
         5
              Kim
                                        Delhi
                                              60000
                                                       10
         clean data.to csv('clean data.csv')
In [67]:
In [68]:
         import os
         os.getcwd()
Out[68]: 'C:\\Users\\Prachi\\FSDS SENAPATI SIR\\Projects'
```

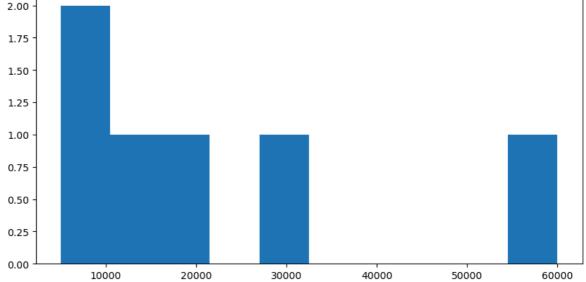
```
In [69]:
         import matplotlib.pyplot as plt
                                                 #visualization
         import seaborn as sns
         import warnings
In [70]:
         warnings.filterwarnings('ignore')
In [71]:
        clean_data['Salary']
               5000
Out[71]: 0
          1
              10000
          2
              15000
          3
              20000
              30000
              60000
          Name: Salary, dtype: int32
In [72]: vis1 = sns.distplot(clean_data['Salary'])
                1e-5
```



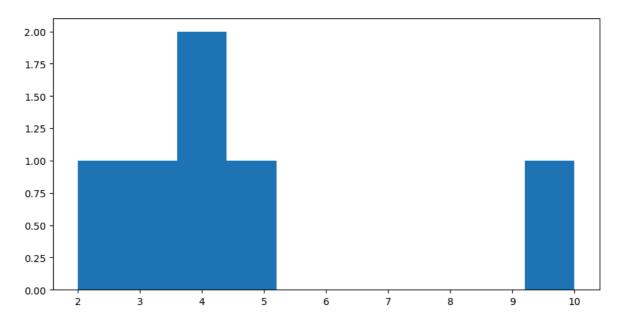
```
In [73]: plt.rcParams['figure.figsize']=10,5
In [74]: vis1 = sns.distplot(clean_data['Salary'])
```







In [76]: vis3 = plt.hist(clean_data['Exp'])

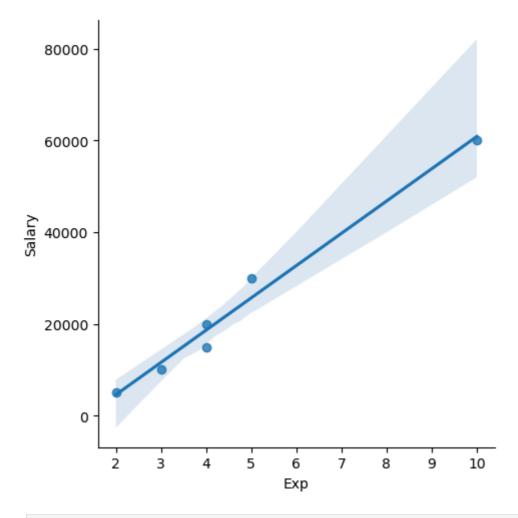


In [77]: clean_data

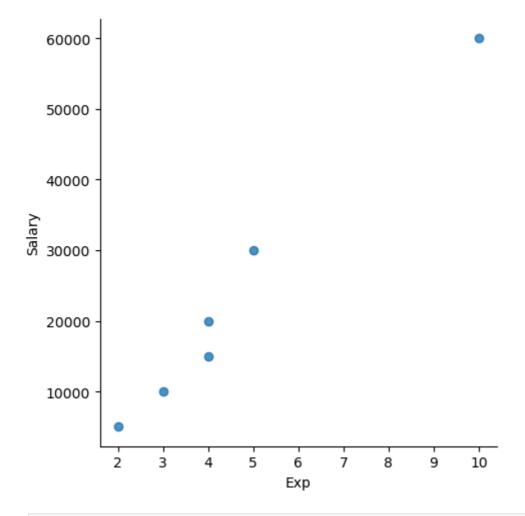
() :	1	1 7	7	
UΠ	J L	1 /	/	١.
		L "		٠.

	Name	Domain	Age	Location	Salary	Ехр
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	y 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

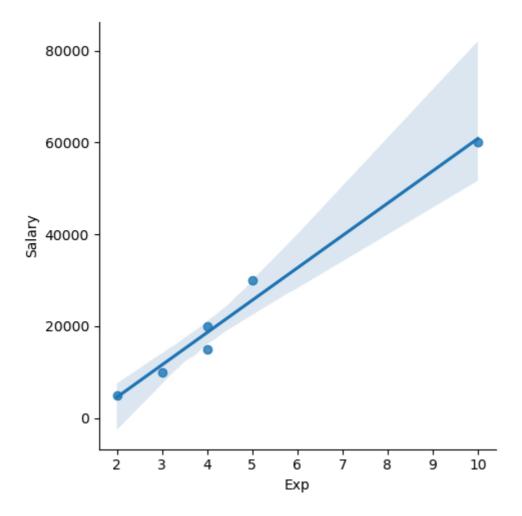
```
In [80]: vis4 = sns.lmplot(data=clean_data,x='Exp',y='Salary')
```



In [83]: vis5 = sns.lmplot(data=clean_data, x='Exp', y='Salary',fit_reg = False)



In [84]: vis6 = sns.lmplot(data=clean_data,x='Exp', y='Salary', fit_reg=True)



In [85]: clean_data

0 1		
()	1 2 5	
Out	00	۰

	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 [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
	5	Kim	NLP	55	Delhi	60000	10
In [87]:	cl	ean_dat	a[:2]				
Out[87]:		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
In [88]:	cl	ean_dat	a[:]				
Out[88]:		Name	Domain	Age	Location	Salary	Ехр
Out[88]:	0	Name Mike	Domain Datascience	Age 34	Location Mumbai	Salary 5000	Exp 2
Out[88]:	0						
Out[88]:		Mike	Datascience	34	Mumbai	5000	2
Out[88]:	1	Mike Teddy	Datascience Testing	34 45	Mumbai Bangalore	5000	2
Out[88]:	1	Mike Teddy Umar	Datascience Testing Dataanalyst	34 45 50	Mumbai Bangalore Bangalore	5000 10000 15000	2 3 4
Out[88]:	1 2 3	Mike Teddy Umar Jane	Datascience Testing Dataanalyst Analytics	34 45 50 50 67	Mumbai Bangalore Bangalore Hyderbad Bangalore	5000 10000 15000 20000	2 3 4 4 5
Out[88]:	1 2 3 4	Mike Teddy Umar Jane Uttam	Datascience Testing Dataanalyst Analytics Statistics	34 45 50 50 67	Mumbai Bangalore Bangalore Hyderbad Bangalore	5000 10000 15000 20000 30000	2 3 4 4 5
Out[88]: In [89]:	1 2 3 4 5	Mike Teddy Umar Jane Uttam	Datascience Testing Dataanalyst Analytics Statistics NLP	34 45 50 50 67	Mumbai Bangalore Bangalore Hyderbad Bangalore	5000 10000 15000 20000 30000	2 3 4 4 5
	1 2 3 4 5	Mike Teddy Umar Jane Uttam Kim	Datascience Testing Dataanalyst Analytics Statistics NLP	34 45 50 50 67 55	Mumbai Bangalore Bangalore Hyderbad Bangalore	5000 10000 15000 20000 30000 60000	2 3 4 4 5 10
In [89]:	1 2 3 4 5	Mike Teddy Umar Jane Uttam Kim ean_dat	Datascience Testing Dataanalyst Analytics Statistics NLP	34 45 50 50 67 55	Mumbai Bangalore Bangalore Hyderbad Bangalore Delhi Location	5000 10000 15000 20000 30000 60000	2 3 4 4 5 10
In [89]:	1 2 3 4 5	Mike Teddy Umar Jane Uttam Kim ean_dat	Datascience Testing Dataanalyst Analytics Statistics NLP a[0:1] Domain Datascience	34 45 50 50 67 55	Mumbai Bangalore Bangalore Hyderbad Bangalore Delhi Location	5000 10000 15000 20000 30000 60000	2 3 4 4 5 10

```
Out[90]:
              Name
                        Domain Age
                                       Location Salary Exp
          0
              Mike
                     Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                  10000
                                                            3
          2
              Umar
                     Dataanalyst
                                   50
                                       Bangalore
                                                  15000
                                                            4
          3
               Jane
                       Analytics
                                   50
                                       Hyderbad
                                                  20000
                                                            4
             Uttam
                        Statistics
                                   67
                                       Bangalore
                                                  30000
                                                            5
          5
                Kim
                            NLP
                                   55
                                           Delhi
                                                  60000
                                                           10
          x_iv = clean_data.drop(['Salary'],axis=1)
In [91]:
In [92]:
          x_iv
Out[92]:
             Name
                                        Location
                        Domain Age
                                                 Exp
              Mike
                     Datascience
                                   34
                                         Mumbai
                                                    2
          0
              Teddy
                                   45
                                       Bangalore
                                                    3
                         Testing
              Umar
                     Dataanalyst
                                       Bangalore
          2
                                   50
                                                    4
          3
               Jane
                       Analytics
                                   50
                                       Hyderbad
                                                    4
             Uttam
          4
                        Statistics
                                   67
                                       Bangalore
                                                    5
          5
                            NLP
                Kim
                                   55
                                           Delhi
                                                   10
In [93]: x_iv.columns
Out[93]: Index(['Name', 'Domain', 'Age', 'Location', 'Exp'], dtype='object')
In [94]:
          clean_data
Out[94]:
                                        Location
                                                 Salary Exp
              Name
                        Domain Age
              Mike
                                                   5000
                                                            2
          0
                     Datascience
                                   34
                                         Mumbai
                                                  10000
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                            3
          2
              Umar
                     Dataanalyst
                                   50
                                       Bangalore
                                                  15000
                                                            4
          3
               Jane
                       Analytics
                                   50
                                       Hyderbad
                                                  20000
          4
             Uttam
                        Statistics
                                   67
                                       Bangalore
                                                  30000
                                                            5
                            NLP
          5
                Kim
                                   55
                                           Delhi
                                                  60000
                                                           10
         y_dv = clean_data.drop(['Name', 'Domain', 'Age', 'Location', 'Exp'],axis=1)
In [96]: y_dv
```

In [97]: clean_data

Out[97]:

	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 [98]: **x_iv**

Out[98]:

	Name	Domain	Age	Location	Ехр
0 Mike Dataso		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 [100... y_dv

Out[100... Salary 0 5000 10000 1 15000 2 20000 3 30000 60000 In [101... clean_data Out[101... Name Domain Age **Location Salary Exp** Mumbai 0 Mike 2 Datascience 34 5000 Bangalore 1 Teddy **Testing** 45 10000 3 Dataanalyst Bangalore 2 Umar 50 15000 4 3 Jane Analytics 50 Hyderbad 20000 5 4 Uttam Statistics 67 Bangalore 30000 60000 5 Kim NLP 55 Delhi 10 In [103... imputation = pd.get_dummies(clean_data) In [104... imputation Out[104... Salary Exp Age Name_Jane Name_Kim Name_Mike Name_Teddy Name_Umar 0 34 5000 2 False False True False False 1 45 10000 3 False False False False True 2 50 15000 4 False False False False True

False

False

False

False

False

False

3

4

5

50

67

55

20000

30000

60000

4

5

10

True

False

False

False

False

True

False

False

False