```
print('hai')
hai
import pandas as pd
emp= pd.read excel(r"C:\Users\ttwrd\Downloads\Rawdata.xlsx")
emp
     Name
                    Domain
                                        Location
                                                   Salary
                                                                Exp
                                 Age
0
     Mike
            Datascience#$
                            34 years
                                          Mumbai
                                                   5^00#0
                                                                 2+
1
  Teddy^
                  Testing
                              45' vr
                                      Bangalore
                                                  10%%000
                                                                 <3
2
           Dataanalyst^^#
                                 NaN
                                                  1$5%000
    Uma#r
                                             NaN
                                                             4> yrs
3
              Ana^^lytics
     Jane
                                 NaN
                                        Hyderbad
                                                   2000^0
                                                                NaN
4
  Uttam*
               Statistics
                               67-yr
                                             NaN
                                                   30000 -
                                                            5+ year
5
      Kim
                       NLP
                                55yr
                                           Delhi
                                                  6000^$0
                                                                10 +
emp.columns
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'],
dtype='object')
emp.shape
(6, 6)
id(emp)
3141807330352
emp.info
<bound method DataFrame.info of</pre>
                                   Name
                                                     Domain
                                                                   Age
Location
           Salary
     Mike
            Datascience#$
                            34 years
                                          Mumbai
                                                   5^00#0
                                                                 2+
1
  Teddy^
                              45' yr
                  Testing
                                      Bangalore
                                                  10%%000
                                                                 <3
2
    Uma#r
           Dataanalyst^^#
                                 NaN
                                             NaN
                                                  1$5%000
                                                             4> yrs
              Ana^^lytics
3
                                                   2000^0
     Jane
                                 NaN
                                        Hyderbad
                                                                NaN
4
  Uttam*
               Statistics
                               67-yr
                                             NaN
                                                   30000 -
                                                            5+ year
5
      Kim
                       NLP
                                55yr
                                           Delhi
                                                  6000^$0
                                                                10+>
emp.head()
     Name
                    Domain
                                 Age
                                        Location
                                                   Salary
                                                                Exp
0
     Mike
            Datascience#$
                            34 years
                                          Mumbai
                                                   5^00#0
                                                                 2+
                              45' yr
1
  Teddv^
                  Testing
                                      Bangalore
                                                  10%%000
                                                                 <3
2
    Uma#r
           Dataanalyst^^#
                                 NaN
                                             NaN
                                                  1$5%000
                                                             4> yrs
3
              Ana^^lytics
                                        Hyderbad
     Jane
                                 NaN
                                                   2000^0
                                                                NaN
  Uttam*
               Statistics
                               67-yr
                                             NaN
                                                   30000-
                                                            5+ year
emp.tail
```

```
<bound method NDFrame.tail of</pre>
                                  Name
                                                 Domain
                                                             Age
Location
           Salary
                      Exp
            Datascience#$ 34 years
    Mike
                                       Mumbai
                                                 5^00#0
                                                             2+
                            45' yr
1
                 Testing
  Teddv^
                                    Bangalore
                                                10%%000
                                                              <3
2
   Uma#r
          Dataanalyst^^#
                               NaN
                                          NaN
                                                1$5%000
                                                          4> yrs
             Ana^^lytics
3
     Jane
                               NaN
                                     Hyderbad
                                                 2000^0
                                                             NaN
4
  Uttam*
               Statistics
                              67-yr
                                          NaN
                                                 30000-
                                                         5+ year
5
     Kim
                     NLP
                               55yr
                                         Delhi
                                                6000^$0
                                                             10+>
emp.isnull()
   Name Domain
                   Age
                        Location
                                   Salary
                                            Exp
   False
          False False
                            False
                                    False
                                           False
   False
          False False
                            False
                                    False
                                           False
1
          False
2
   False
                            True
                                    False
                                           False
                  True
3
   False
          False
                  True
                            False
                                    False
                                           True
   False
          False False
                            True
                                    False
                                           False
5 False
          False False
                            False
                                    False False
emp.isna()
   Name Domain
                   Age
                        Location
                                   Salary
                                            Exp
                                           False
          False False
                                    False
   False
                            False
1
   False
          False False
                            False
                                    False
                                           False
2
   False
          False
                  True
                            True
                                    False
                                           False
3
   False
          False
                            False
                                    False
                  True
                                           True
4
   False
          False False
                            True
                                    False
                                           False
5
  False
          False False
                            False
                                    False False
emp
                                     Location
   Name
                  Domain
                              Age
                                                Salary
                                                            Exp
0
   Mike
          Datascience#$
                          34 years
                                       Mumbai
                                                5^00#0
                                                             2+
1
  Teddy
                            45' yr
                                    Bangalore
                                              10%%000
                                                             <3
                Testing
2
   Umar
         Dataanalyst^^#
                               NaN
                                          NaN
                                               1$5%000
                                                         4> yrs
3
            Ana^^lytics
                              NaN
                                    Hyderbad
   Jane
                                                2000^0
                                                            NaN
4
  Uttam
              Statistics
                             67-yr
                                          NaN
                                                30000-
                                                        5+ vear
5
    Kim
                    NLP
                              55vr
                                        Delhi
                                               6000^$0
                                                            10+
emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
emp['Salary']
0
      5000
1
     10000
2
     15000
3
    20000
4
    30000
    60000
Name: Salary, dtype: object
```

```
emp['Name']=emp['Name'].str.replace(r'\W','',regex=True)
emp['Name']
0
      Mike
1
     Teddy
2
      Umar
3
      Jane
4
     Uttam
5
       Kim
Name: Name, dtype: object
emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
emp['Salary']
0
      5000
1
     10000
2
     15000
3
     20000
4
     30000
5
     60000
Name: Salary, dtype: object
emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
emp['Domain']
0
     Datascience
1
         Testing
2
     Dataanalyst
3
       Analytics
4
      Statistics
5
             NLP
Name: Domain, dtype: object
emp['Age']
0
     34 years
1
       45' yr
2
          NaN
3
          NaN
4
        67-yr
5
         55yr
Name: Age, dtype: object
emp["Age"]=emp["Age"].astype(str)
```

```
emp['Age']=emp['Age'].str.extract(r'(\d+)')
emp['Age']=emp['Age'].str.extract(r'(\d+)')
emp['Age'] = pd.to_numeric(emp['Age'])
emp['Age']
     34.0
0
1
     45.0
2
      NaN
3
      NaN
4
     67.0
5
     55.0
Name: Age, dtype: float64
emp['Exp']
0
          2+
1
          <3
2
      4> yrs
3
         NaN
4
     5+ year
5
         10+
Name: Exp, dtype: object
emp['Exp']=emp['Exp'].str.extract(r'(\d+)')
emp['Exp']
       2
0
1
       3
2
       4
3
     NaN
4
       5
5
      10
Name: Exp, dtype: object
emp
                 Domain
                                 Location
     Name
                          Age
                                            Salary
                                                     Exp
0
     Mike
           Datascience
                         34.0
                                   Mumbai
                                            5^00#0
                                                       2
  Teddy^
                         45.0
                                                       3
1
                Testing
                                Bangalore
                                           10%%000
2
                                                       4
    Uma#r
           Dataanalyst
                          NaN
                                      NaN
                                           1$5%000
3
              Analytics
                          NaN
                                 Hyderbad
                                            2000^0
                                                     NaN
     Jane
                         67.0
4
   Uttam*
            Statistics
                                      NaN
                                             30000 -
                                                       5
5
      Kim
                    NLP
                         55.0
                                    Delhi
                                           6000^$0
                                                      10
emp['Salary']
0
      5000
1
     10000
```

```
2
     15000
3
     20000
4
     30000
5
     60000
Name: Salary, dtype: object
emp
    Name
               Domain
                        Age
                               Location Salary
                                                Exp
    Mike Datascience 34.0
                                 Mumbai
                                          5000
                                                  2
1
  Teddy
              Testing 45.0
                              Bangalore
                                         10000
                                                  3
2
                                                  4
   Umar Dataanalyst
                        NaN
                                    NaN
                                        15000
3
            Analytics
                        NaN
                               Hyderbad
                                         20000
                                                NaN
    Jane
4
  Uttam
           Statistics
                       67.0
                                    NaN
                                         30000
                                                  5
5
                  NLP 55.0
                                  Delhi 60000
                                                 10
     Kim
clean_data=emp.copy()
clean_data
    Name
               Domain
                               Location Salary
                        Age
                                                Exp
0
                                 Mumbai
    Mike
          Datascience
                       34.0
                                          5000
                                                  2
                                                  3
1
  Teddy
              Testing 45.0
                              Bangalore
                                         10000
2
    Umar
                                                  4
          Dataanalyst
                                         15000
                        NaN
                                    NaN
3
                        NaN
    Jane
            Analytics
                               Hyderbad
                                         20000
                                                NaN
4
  Uttam
                       67.0
                                    NaN
                                         30000
                                                  5
           Statistics
5
                  NLP
                                  Delhi
     Kim
                       55.0
                                         60000
                                                 10
clean data['Exp']
0
       2
1
       3
2
       4
3
     NaN
4
       5
5
      10
Name: Exp, dtype: object
import numpy as np
clean data['Exp']=clean data['Exp'].fillna(np.mean(pd.to numeric(clean
data['Exp'])))
clean_data['Exp']
0
       2
1
       3
2
       4
3
     4.8
4
       5
5
      10
Name: Exp, dtype: object
```

```
clean data
    Name
               Domain
                        Age
                              Location Salary
                                               Exp
    Mike
0
          Datascience
                       34.0
                                Mumbai
                                          5000
1
   Teddy
              Testing 45.0
                             Bangalore
                                         10000
                                                  3
2
                                                  4
          Dataanalyst
                                        15000
    Umar
                        NaN
                                   NaN
3
            Analytics
                        NaN
                              Hyderbad
                                         20000 4.8
    Jane
4
           Statistics
                       67.0
                                   NaN
                                                  5
  Uttam
                                         30000
                                 Delhi 60000
5
     Kim
                  NLP 55.0
                                                 10
clean data['Domain']
0
     Datascience
1
         Testing
2
     Dataanalyst
3
       Analytics
4
      Statistics
5
             NLP
Name: Domain, dtype: object
clean data['Location']=clean data['Location'].astype('category')
clean data['Domain']=clean data['Domain'].astype('category')
clean data['Name']=clean_data['Name'].astype('category')
clean data['Location']
clean data['Domain']
0
     Datascience
1
         Testina
2
     Dataanalyst
3
       Analytics
4
      Statistics
5
             NLP
Name: Domain, dtype: category
Categories (6, object): ['Analytics', 'Dataanalyst', 'Datascience',
'NLP', 'Statistics', 'Testing']
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
 1
     Domain
              6 non-null
                               category
 2
               4 non-null
                               float64
     Age
     Location 4 non-null
 3
                               category
 4
     Salary
               6 non-null
                               object
 5
               6 non-null
     Exp
                               object
```

```
dtypes: category(2), float64(1), object(3)
memory usage: 760.0+ bytes
clean data['Location']
0
        Mumbai
1
     Bangalore
2
           NaN
3
      Hyderbad
4
           NaN
5
         Delhi
Name: Location, dtype: category
Categories (4, object): ['Bangalore', 'Delhi', 'Hyderbad', 'Mumbai']
clean data['Location']=clean data['Location'].fillna(clean data['Locat
ion'].mode()[0])
clean data['Location']
        Mumbai
0
1
     Bangalore
2
     Bangalore
3
      Hyderbad
4
     Bangalore
5
         Delhi
Name: Location, dtype: category
Categories (4, object): ['Bangalore', 'Delhi', 'Hyderbad', 'Mumbai']
clean data
    Name
               Domain
                              Location Salary
                        Age
                                               Exp
0
    Mike Datascience 34.0
                                Mumbai
                                         5000
                                                  2
              Testing 45.0
                             Bangalore 10000
                                                  3
1
  Teddy
2
                             Bangalore 15000
                                                  4
    Umar
          Dataanalyst
                        NaN
3
                        NaN
                             Hyderbad 20000 4.8
    Jane
            Analytics
           Statistics
                       67.0
                             Bangalore 30000
                                                  5
  Uttam
5
     Kim
                  NLP 55.0
                                 Delhi 60000
                                                10
clean data['Age']
0
     34.0
1
     45.0
2
      NaN
3
      NaN
4
     67.0
5
     55.0
Name: Age, dtype: float64
clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean
data['Age'])))
clean data['Age']
```

```
0
     34.00
1
     45.00
2
     50.25
3
     50.25
4
     67.00
5
     55.00
Name: Age, dtype: float64
clean data
    Name
               Domain
                         Age
                               Location Salary
                                                 Exp
0
    Mike Datascience 34.00
                                 Mumbai
                                           5000
                                                   2
                                                   3
1
  Teddy
              Testing 45.00
                              Bangalore 10000
2
   Umar
         Dataanalyst 50.25
                              Bangalore
                                         15000
                                                   4
            Analytics 50.25
3
                                                 4.8
    Jane
                              Hyderbad
                                         20000
4
           Statistics
   Uttam
                       67.00
                              Bangalore
                                         30000
                                                   5
5
                                  Delhi
     Kim
                                                  10
                  NLP 55.00
                                         60000
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
               6 non-null
     Name
                               category
1
    Domain
              6 non-null
                               category
 2
               6 non-null
                               float64
     Age
 3
     Location 6 non-null
                               category
4
               6 non-null
     Salary
                               object
 5
               6 non-null
     Exp
                               object
dtypes: category(3), float64(1), object(2)
memory usage: 938.0+ bytes
clean data['Salary']
0
      5000
1
     10000
2
     15000
3
     20000
4
     30000
5
     60000
Name: Salary, dtype: object
clean data['Salary']=clean data['Salary'].astype(int)
clean data['Salary']
0
      5000
1
     10000
2
     15000
3
     20000
```

```
4
     30000
5
     60000
Name: Salary, dtype: int32
clean data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
    Column
              Non-Null Count Dtype
#
- - -
     -----
              6 non-null
 0
     Name
                               category
    Domain 6 non-null
 1
                               category
             6 non-null
 2
                               float64
    Age
 3
    Location 6 non-null
                               category
    Salary
 4
               6 non-null
                               int32
 5
     Exp
               6 non-null
                               object
dtypes: category(3), float64(1), int32(1), object(1)
memory usage: 914.0+ bytes
clean data['Exp']=clean data['Exp'].astype(int)
clean data['Exp']
      2
0
1
      3
2
      4
3
      4
4
     5
5
     10
Name: Exp, dtype: int32
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
                               float64
2
              6 non-null
     Aae
 3
    Location 6 non-null
                               category
 4
     Salary
               6 non-null
                               int32
 5
               6 non-null
                               int32
     Exp
dtypes: category(3), float64(1), int32(2)
memory usage: 890.0 bytes
clean data['Age']=clean data['Age'].astype(int)
```

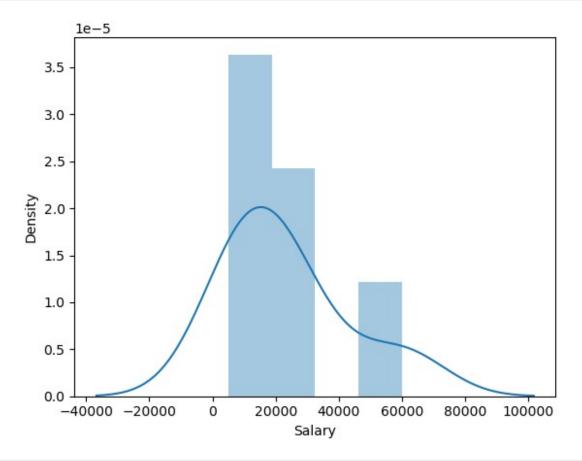
```
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

clean_data.to_csv("clean_data.csv")
import os
os.getcwd()

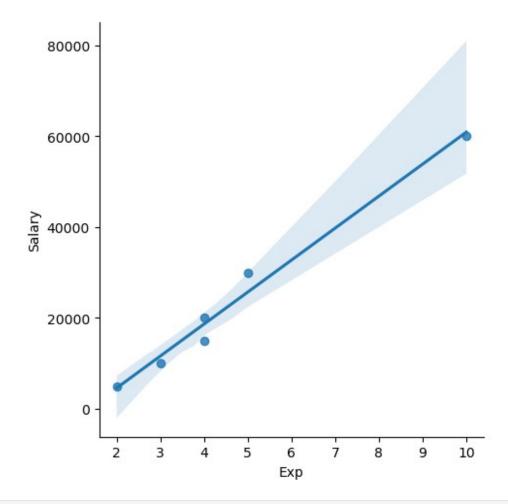
'C:\\Users\\ttwrd\\Downloads'

visl=sns.distplot(clean_data['Salary'])
visl

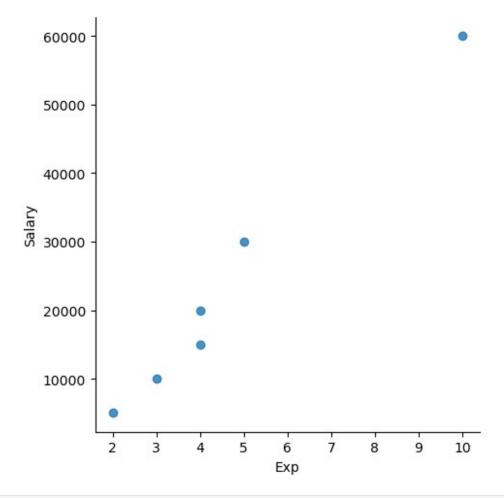
<Axes: xlabel='Salary', ylabel='Density'>
```



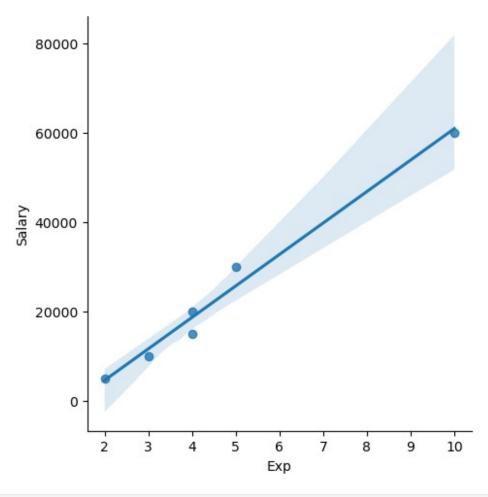
vis2=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=True)



vis2=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=False)



vis2=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=True)



<pre>clean_data[:]</pre>							
0 1 2 3 4 5	Name Mike Teddy Umar Jane Uttam Kim	Domain Datascience Testing Dataanalyst Analytics Statistics NLP	Age 34 45 50 50 67 55	Location Mumbai Bangalore Bangalore Hyderbad Bangalore Delhi	Salary 5000 10000 15000 20000 30000 60000	2 3 4 4 5	
cl	ean_dat	a[<mark>0:6:2</mark>]					
0 2 4	Name Mike Umar Uttam	Domain Datascience Dataanalyst Statistics	Age 34 50 67	Location Mumbai Bangalore Bangalore	Salary 5000 15000 30000	Exp 2 4 5	
cl	ean_dat	a[::- <mark>1</mark>]					
5 4	Name Kim Uttam	Domain NLP Statistics	Age 55 67	Location Delhi Bangalore	Salary 60000 30000	Exp 10 5	

```
3
    Jane
            Analytics
                        50
                             Hyderbad
                                        20000
                                                 4
2
          Dataanalyst
                        50
                            Bangalore
                                                 4
    Umar
                                        15000
1
  Teddy
              Testing
                        45
                            Bangalore
                                        10000
                                                 3
                               Mumbai
                                                 2
0 Mike Datascience
                        34
                                         5000
clean data.columns
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'],
dtype='object')
x indv=clean data[['Name','Domain','Age','Location','Exp']]
type(x indv)
pandas.core.frame.DataFrame
y depvariable=clean data['Salary']
type(y_depvariable)
pandas.core.series.Series
x indv
                                       Exp
    Name
               Domain Age
                             Location
    Mike
                               Mumbai
0
         Datascience
                        34
                                         2
                                         3
1
  Teddy
              Testing
                        45
                            Bangalore
2
                        50 Bangalore
                                         4
  Umar
          Dataanalyst
3
                                         4
    Jane
            Analytics
                        50
                             Hyderbad
4
                            Bangalore
                                         5
  Uttam
           Statistics
                        67
5
     Kim
                  NLP
                        55
                                Delhi
                                        10
y_depvariable
0
      5000
1
     10000
2
     15000
3
     20000
4
     30000
5
     60000
```

Name: Salary, dtype: int32