```
In [1]: import pandas as pd
In [2]:
        emp=pd.read_excel(r"C:\Users\MANISHA\Downloads\Rawdata.xlsx")
In [3]: emp
Out[3]:
            Name
                          Domain
                                      Age
                                            Location
                                                        Salary
                                                                   Exp
         0
              Mike
                                                       5^00#0
                                                                    2+
                     Datascience#$
                                   34 years
                                             Mumbai
         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
           Uttam*
                          Statistics
                                     67-yr
                                                NaN
                                                        30000- 5+ year
         5
                              NLP
                                                Delhi
                                                      6000^$0
               Kim
                                      55yr
                                                                   10+
In [4]: emp.shape
Out[4]: (6, 6)
        len(emp)
In [5]:
Out[5]: 6
In [6]:
        emp.columns
Out[6]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [7]: len(emp.columns)
Out[7]: 6
In [8]: emp.info()
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
            Column
                      Non-Null Count Dtype
        0
            Name
                      6 non-null
                                       object
        1
            Domain
                      6 non-null
                                       object
        2
                      4 non-null
                                       object
            Age
        3
            Location 4 non-null
                                       object
            Salary
                      6 non-null
                                       object
        5
                      5 non-null
            Exp
                                       object
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [9]: emp['Name']
```

```
Out[9]: 0
                 Mike
          1
               Teddy^
          2
                Uma#r
          3
                 Jane
               Uttam*
          5
                   Kim
          Name: Name, dtype: object
In [10]:
          emp['Domain']
Out[10]: 0
                Datascience#$
          1
                       Testing
               Dataanalyst^^#
          2
                  Ana^^lytics
          3
          4
                    Statistics
          5
                           NLP
          Name: Domain, dtype: object
In [11]: emp['Salary']
Out[11]: 0
                5^00#0
               10%%000
          2
               1$5%000
                2000^0
          3
          4
                30000-
          5
               6000^$0
          Name: Salary, dtype: object
         emp[['Name','Domain','Age','Location','Salary','Exp']]
In [12]:
Out[12]:
              Name
                            Domain
                                        Age
                                              Location
                                                           Salary
                                                                      Exp
          0
               Mike
                       Datascience#$
                                     34 years
                                                          5^00#0
                                                                      2+
                                               Mumbai
             Teddy^
                                              Bangalore
                             Testing
                                       45' yr
                                                        10%%000
                                                                      <3
              Uma#r Dataanalyst^^#
                                        NaN
                                                  NaN
                                                        1$5%000
                                                                    4> yrs
                         Ana^^lytics
                                              Hyderbad
                                                          2000^0
                Jane
                                        NaN
                                                                     NaN
             Uttam*
                            Statistics
                                       67-yr
                                                  NaN
                                                           30000- 5+ year
          5
                               NLP
                                        55yr
                                                  Delhi
                                                         6000^$0
                                                                     10+
                Kim
```

## Cleaning data

```
In [14]: emp['Name']
```

```
Out[14]: 0
          1
               Teddy^
          2
                Uma#r
          3
                 Jane
          4
               Uttam*
          5
                  Kim
          Name: Name, dtype: object
In [15]: emp['Name']=emp['Name'].str.replace(r'\W','',regex=True)
         emp['Name']
In [16]:
Out[16]:
          0
                Mike
               Teddy
          1
                Umar
          2
          3
                Jane
          4
               Uttam
                 Kim
          5
          Name: Name, dtype: object
In [17]:
          emp
Out[17]:
             Name
                          Domain
                                       Age
                                             Location
                                                         Salary
                                                                    Exp
          0
              Mike
                     Datascience#$ 34 years
                                                        5^00#0
                                                                     2+
                                              Mumbai
             Teddy
                                      45' yr
                                            Bangalore 10%%000
                            Testing
                                                                     <3
             Umar Dataanalyst^^#
          2
                                      NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
          3
              Jane
                       Ana^^lytics
                                      NaN Hyderbad
                                                        2000^0
                                                                   NaN
          4
            Uttam
                          Statistics
                                                 NaN
                                                         30000-
                                                                5+ year
                                      67-yr
          5
                              NLP
                                                Delhi
                                                       6000^$0
               Kim
                                       55yr
                                                                    10+
         emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True)
In [18]:
In [19]: emp['Domain']
               Datascience
Out[19]: 0
          1
                   Testing
          2
               Dataanalyst
                 Analytics
          3
                Statistics
          4
                       NLP
          5
          Name: Domain, dtype: object
In [20]:
         emp
```

Mike

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

```
1
                10000
          2
               15000
               20000
          3
          4
                30000
                60000
          5
          Name: Salary, dtype: object
In [28]: emp.head()
Out[28]:
             Name
                       Domain Age
                                       Location Salary
                                                           Exp
          0
              Mike Datascience
                                  34
                                        Mumbai
                                                  5000
                                                            2+
                                      Bangalore
             Teddy
                        Testing
                                  45
                                                 10000
                                                            <3
          2
              Umar
                    Dataanalyst NaN
                                           NaN
                                                 15000
                                                         4> yrs
          3
              Jane
                       Analytics
                                NaN
                                      Hyderbad
                                                 20000
                                                           NaN
          4 Uttam
                       Statistics
                                  67
                                           NaN
                                                 30000 5+ year
In [29]: emp['Exp']=emp['Exp'].str.extract(r'(\d+)')
In [30]: emp['Exp']
Out[30]:
                 2
          1
                  3
          2
                 4
          3
               NaN
                 5
          4
                10
          Name: Exp, dtype: object
In [31]: emp
                       Domain Age
Out[31]:
             Name
                                       Location Salary
                                                         Exp
          0
              Mike Datascience
                                  34
                                        Mumbai
                                                  5000
                                                           2
             Teddy
                                      Bangalore
                                                 10000
                                                           3
                        Testing
                                  45
                    Dataanalyst NaN
                                                 15000
          2
              Umar
                                           NaN
                                                           4
          3
                                      Hyderbad
                                                 20000 NaN
              Jane
                       Analytics
                                NaN
                                                           5
          4
             Uttam
                       Statistics
                                           NaN
                                                 30000
                                  67
                                                 60000
               Kim
                           NLP
                                  55
                                          Delhi
                                                          10
          clean_data=emp.copy()
In [32]:
          clean_data
```

Out[27]: 0

5000

```
Out[32]:
             Name
                        Domain
                                 Age
                                       Location Salary
                                                         Exp
          0
              Mike Datascience
                                  34
                                        Mumbai
                                                   5000
                                                            2
             Teddy
                         Testing
                                  45
                                       Bangalore
                                                  10000
                                                            3
          2
              Umar
                     Dataanalyst NaN
                                           NaN
                                                  15000
                                                            4
                       Analytics
                                 NaN
                                       Hyderbad
                                                 20000 NaN
          3
               Jane
          4
             Uttam
                       Statistics
                                           NaN
                                                  30000
                                                            5
                                   67
                           NLP
               Kim
                                   55
                                           Delhi
                                                  60000
                                                           10
In [66]:
          import numpy as np
          clean_data['Age']= clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']
In [68]:
          clean data
Out[68]:
             Name
                        Domain
                                        Location Salary
                                  Age
                                                          Exp
          0
              Mike Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
             Teddy
                         Testing
                                   45
                                       Bangalore
                                                  10000
                                                             3
          2
              Umar
                     Dataanalyst 50.25
                                            NaN
                                                  15000
                                                            4
          3
               Jane
                       Analytics
                                 50.25
                                        Hyderbad
                                                  20000
                                                         NaN
             Uttam
                       Statistics
                                   67
                                            NaN
                                                  30000
                                                             5
          5
               Kim
                           NLP
                                   55
                                            Delhi
                                                  60000
                                                            10
          clean_data['Exp']= clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']
In [70]:
          clean_data
Out[70]:
                        Domain
                                        Location Salary Exp
             Name
                                  Age
          0
              Mike Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
             Teddy
                         Testing
                                       Bangalore
                                                  10000
                                                            3
                                   45
          2
              Umar
                     Dataanalyst 50.25
                                            NaN
                                                  15000
                                                            4
                       Analytics 50.25
                                        Hyderbad
                                                  20000
          3
               Jane
                                                          4.8
          4 Uttam
                       Statistics
                                                  30000
                                                            5
                                   67
                                            NaN
          5
                           NLP
                                   55
                                            Delhi
                                                  60000
               Kim
                                                           10
          clean data['Location'] = clean data['Location'].fillna(clean data['Location'].mode()
In [72]:
          clean_data
```

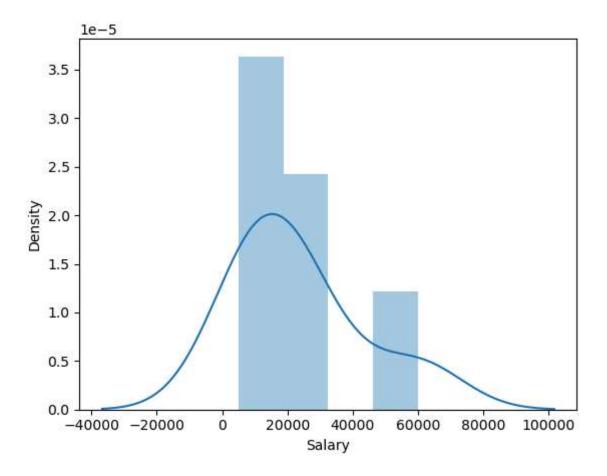
```
Out[72]:
            Name
                      Domain
                                Age
                                      Location Salary Exp
         0
              Mike Datascience
                                 34
                                       Mumbai
                                                 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
            Uttam
                      Statistics
                                     Bangalore
                                                30000
                                                         5
                                  67
               Kim
                          NLP
                                  55
                                          Delhi
                                                60000
                                                        10
         clean_data.isna().sum()
In [74]:
Out[74]:
         Name
                      0
         Domain
                      0
         Age
                      0
          Location
                      0
          Salary
                      0
          Exp
                      0
          dtype: int64
In [76]: 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
                       6 non-null
                                       object
             Age
         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 [78]: clean_data['Name']=clean_data['Name'].astype(str)
         clean_data['Domain']=clean_data['Domain'].astype(str)
         clean_data['Location']=clean_data['Location'].astype(str)
In [80]: 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
                       6 non-null
         5
                                       object
             Exp
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [82]: 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.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
                       6 non-null
             Age
                                       int32
         3
             Location 6 non-null
                                       object
         4
             Salary
                       6 non-null
                                       int32
         5
             Exp
                       6 non-null
                                       int32
        dtypes: int32(3), object(3)
        memory usage: 348.0+ bytes
In [84]: clean_data
Out[84]:
                      Domain Age
                                    Location Salary Exp
            Name
                                34
         0
             Mike Datascience
                                     Mumbai
                                               5000
                                                       2
            Teddy
                                    Bangalore
                                              10000
                       Testing
                                                       3
             Umar Dataanalyst
                                    Bangalore
                                              15000
                                                       4
         3
              Jane
                      Analytics
                                    Hyderbad
                                              20000
         4 Uttam
                      Statistics
                                67
                                    Bangalore
                                              30000
                                                       5
         5
              Kim
                         NLP
                                55
                                        Delhi
                                              60000
                                                      10
         clean data.to csv('clean data.csv')
In [90]:
         clean data.to excel(r"C:\Users\MANISHA\Downloads\20th- EDA Practicle\Rawdata.xlsx")
In [88]:
         import os
In [92]:
         os.getcwd()
```

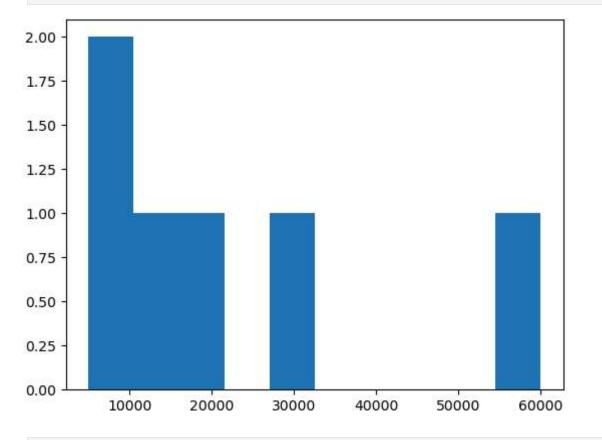
```
Out[92]: 'C:\\Users\\MANISHA'
In [94]: clean_data
                                      Location Salary Exp
Out[94]:
                       Domain Age
             Name
              Mike Datascience
                                                         2
                                 34
                                       Mumbai
                                                 5000
             Teddy
                        Testing
                                 45
                                     Bangalore
                                                10000
                                                         3
                    Dataanalyst
                                     Bangalore
                                                15000
                                                         4
             Umar
          3
                       Analytics
                                     Hyderbad
                                                20000
              Jane
                                 50
                                                         5
                       Statistics
                                     Bangalore
                                                30000
          4 Uttam
          5
                          NLP
                                 55
                                         Delhi
                                                60000
                                                        10
               Kim
```

## **EDA TECHNIQUE LETS APPLY**

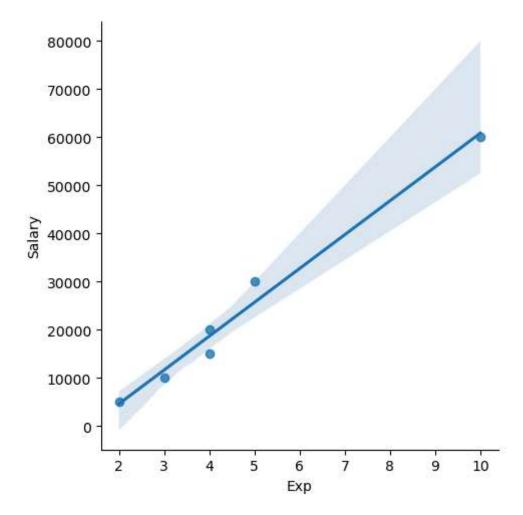
```
In [97]:
           import matplotlib.pyplot as plt
           import seaborn as sns
 In [99]: import warnings
           warnings.filterwarnings('ignore')
In [101...
           clean_data['Salary']
Out[101...
                 5000
           1
                10000
           2
                15000
                20000
           3
                30000
                60000
           Name: Salary, dtype: int32
          vis1=sns.distplot(clean_data['Salary'])
In [103...
```



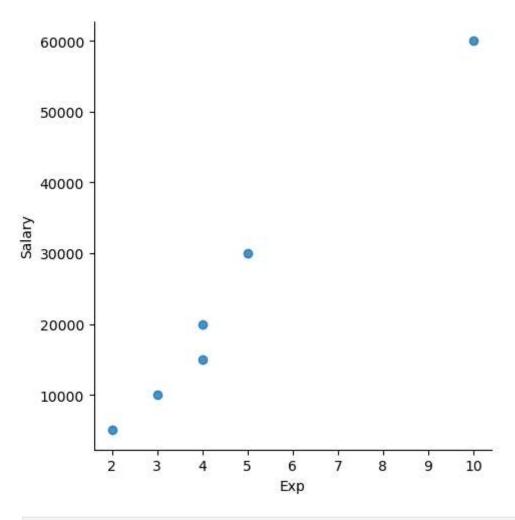
In [105... vis2=plt.hist(clean\_data['Salary'])



In [107... vis3=sns.lmplot(data=clean\_data,x='Exp',y='Salary')



In [109... vis4=sns.lmplot(data=clean\_data,x='Exp',y='Salary',fit\_reg=False)



In [111... clean\_data[:]

Out[111...

	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 [115...

clean\_data[0:6:2]

Out[115...

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

```
In [117...
           clean_data[::-1]
Out[117...
                                        Location Salary Exp
              Name
                         Domain Age
           5
                            NLP
                                            Delhi
                 Kim
                                    55
                                                  60000
                                                           10
              Uttam
                         Statistics
                                   67
                                        Bangalore
                                                  30000
                                                            5
           3
                Jane
                        Analytics
                                   50
                                       Hyderbad
                                                  20000
                                                            4
                      Dataanalyst
           2
               Umar
                                   50
                                       Bangalore
                                                  15000
           1
              Teddy
                          Testing
                                   45
                                       Bangalore
                                                  10000
                                                            3
               Mike Datascience
                                    34
                                         Mumbai
                                                    5000
                                                            2
In [119...
           clean_data.columns
Out[119...
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
           X_iv=clean_data[['Name','Domain','Age','Location','Exp']]
In [121...
In [123...
           X_iv
Out[123...
                                        Location Exp
              Name
                         Domain Age
           0
               Mike Datascience
                                   34
                                         Mumbai
                                                     2
              Teddy
                          Testing
                                   45
                                        Bangalore
                                                     3
           2
               Umar
                      Dataanalyst
                                   50
                                       Bangalore
                                                    4
           3
                        Analytics
                                   50
                                       Hyderbad
                                                    4
                Jane
           4
              Uttam
                         Statistics
                                       Bangalore
                                                    5
                                   67
           5
                            NLP
                                   55
                                                    10
                 Kim
                                            Delhi
           Y_dv=clean_data[['Salary']]
In [125...
In [127...
           Y_dv
Out[127...
              Salary
           0
                5000
              10000
              15000
           2
           3
              20000
               30000
           5 60000
```

In [129...

emp

Out[129...

	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 [131... clean\_data

Out[131...

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

In [133... X\_iv

Out[133...

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

In [135... Y\_dv

In [ ]: