```
import pandas as pd
In [1]:
In [2]: pd.__version__
Out[2]: '2.2.2'
In [3]: pip install --upgrade openpyxl
       Requirement already satisfied: openpyxl in c:\users\hanshu\anaconda3\lib\site-pac
       kages (3.1.5)
       Requirement already satisfied: et-xmlfile in c:\users\hanshu\anaconda3\lib\site-p
       ackages (from openpyxl) (1.1.0)
       Note: you may need to restart the kernel to use updated packages.
In [4]: pd.__version__
Out[4]: '2.2.2'
In [5]:
        emp = pd.read_excel(r"C:\Users\Hanshu\Desktop\excel data\Rawdata.xlsx")
        emp
Out[5]:
            Name
                          Domain
                                      Age
                                            Location
                                                        Salary
                                                                   Exp
         0
             Mike
                     Datascience#$ 34 years
                                             Mumbai
                                                       5^00#0
                                                                   2+
         1 Teddy^
                                     45' yr Bangalore
                                                                   <3
                           Testing
                                                     10%%000
            Uma#r Dataanalyst^^#
                                     NaN
                                                      1$5%000
         2
                                                NaN
                                                                4> yrs
                       Ana^^lytics
                                     NaN Hyderbad
                                                       2000^0
         3
              Jane
                                                                  NaN
            Uttam*
                          Statistics
                                     67-yr
                                                NaN
                                                        30000-
                                                               5+ year
         5
               Kim
                             NLP
                                      55yr
                                               Delhi
                                                      6000^$0
                                                                  10+
In [6]:
        id(emp)
Out[6]: 2456864696192
In [7]: emp.columns
Out[7]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [8]: emp.shape
Out[8]: (6, 6)
```

In [9]: emp.head()

Out[9]:	Name		Domain	Age	e Location	n Salar	у Ехр
	0	Mike	Datascience#\$	34 years	s Mumba	i 5^00#	0 2+
	1	Teddy^	Testing	45' y	r Bangalore	e 10%%00	0 <3
	2	Uma#r	Dataanalyst^^#	NaN	N NaN	J 1\$5%00	0 4> yrs
	3	Jane	Ana^^lytics	NaN	N Hyderbad	2000^0	0 NaN
	4	Uttam*	Statistics	67-y	•		- 5+ year
In [10]:	em	p.tail()					
Out[10]:		Name	Domain	Age	Location	Salary	Ехр
	1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
	2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
	3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
	4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
	5	Kim	NLP	55yr	Delhi	6000^\$0	10+
In [11]:	em	p.info()					
	Rang Data # 0 1 2 3 4 5 dtyp	geIndex: a columns Column Name Domain Age Locatio Salary Exp pes: obje	6 non-null 5 non-null	o 5 mns): unt Dt ob ob ob ob	me'> ype ject ject ject ject ject		
In [12]:	em	р					
Out[12]:		Name	Domain	Age	e Location	n Salar	у Ехр
	0	Mike	Datascience#\$	34 years	s Mumba	i 5^00#	0 2+
	1	Teddy^	Testing	45' y	r Bangalore	e 10%%00	0 <3
	2	Uma#r	Dataanalyst^^#	NaN	N NaN	1\$5%00	0 4> yrs
	3	Jane	Ana^^lytics	NaN	N Hyderbac	2000^	0 NaN
	4	Uttam*	Statistics	67-y	r NaN	30000	- 5+ year
	5	Kim	NLP	55y	r Delh	i 6000^\$(0 10+

In [13]: emp.isnull()

Out[13]:		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 [14]:	em	p.isna()				
Out[14]:		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 False False False True False False False False False False False False

In [16]: emp

Out	[1	6	
			_

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

```
Out[17]: 0
               Mike
            Teddy^
         1
         2
              Uma#r
         3
                Jane
         4
              Uttam*
         5
                 Kim
         Name: Name, dtype: object
In [18]: emp['Domain']
Out[18]: 0
               Datascience#$
         1
                     Testing
         2
             Dataanalyst^^#
         3
                Ana^^lytics
         4
                  Statistics
         5
                         NLP
         Name: Domain, dtype: object
In [19]: emp['Age']
Out[19]: 0
              34 years
                45' yr
         1
         2
                   NaN
         3
                   NaN
                 67-yr
         5
                  55yr
         Name: Age, dtype: object
In [20]: emp['Salary']
Out[20]: 0
               5^00#0
         1
              10%%000
         2
              1$5%000
         3
               2000^0
         4
              30000-
              6000^$0
         Name: Salary, dtype: object
In [21]: emp['Exp']
Out[21]: 0
                   2+
         1
                   <3
         2
               4> yrs
         3
                  NaN
         4
             5+ year
         5
                  10+
         Name: Exp, dtype: object
In [22]:
         emp
```

```
Out[22]:
             Name
                           Domain
                                       Age
                                             Location
                                                      Salary
                                                                   Exp
                      Datascience#$ 34 years
                                                                    2+
         0
              Mike
                                             Mumbai
                                                        5^00#0
          1 Teddy^
                                    45' yr Bangalore 10%%000
                          Testing
                                                                    <3
             Uma#r Dataanalyst^^#
                                    NaN
                                                NaN
                                                      1$5%000
                                                                 4> yrs
                                    NaN Hyderbad
                        Ana^^lytics
                                                        2000^0
         3
               Jane
                                                                   NaN
             Uttam*
                                      67-yr
                                                                5+ year
                          Statistics
                                                NaN
                                                      30000-
               Kim
                              NLP
                                       55yr
                                                Delhi
                                                       6000^$0
                                                                   10+
In [23]: emp[['Name','Domain']]
Out[23]:
             Name
                           Domain
              Mike
                      Datascience#$
          0
          1 Teddy^
                           Testing
             Uma#r Dataanalyst^^#
          2
                       Ana^^lytics
         3
               Jane
             Uttam*
                          Statistics
          5
               Kim
                              NLP
In [24]: emp[['Name','Domain','Age']]
Out[24]:
             Name
                           Domain
                                       Age
                      Datascience#$ 34 years
          0
              Mike
          1 Teddy^
                           Testing
                                      45' yr
             Uma#r Dataanalyst^^#
          2
                                     NaN
                        Ana^^lytics
                                    NaN
          3
               Jane
             Uttam*
                           Statistics
                                      67-yr
          5
                              NLP
                                       55yr
               Kim
```

In [25]: emp[['Name','Domain','Age','Location']]

Out[25]:		Name	Domain	Age	Location
	0	Mike	Datascience#\$	34 years	Mumbai
	1	Teddy^	Testing	45' yr	Bangalore
	2	Uma#r	Dataanalyst^^#	NaN	NaN
	3	Jane	Ana^^lytics	NaN	Hyderbad
	4	Uttam*	Statistics	67-yr	NaN
	5	Kim	NLP	55yr	Delhi

DATA CLEANING & DATA CLEANSING

```
In [26]: emp['Name']
Out[26]: 0
                 Mike
              Teddy^
          2
               Uma#r
          3
                 Jane
              Uttam*
          5
                  Kim
          Name: Name, dtype: object
         emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True) # remove caps and...
In [27]:
         emp['Name']
Out[27]: 0
               Mike
          1
             Teddy
          2
               Umar
                Jane
          4
              Uttam
                 Kim
          Name: Name, dtype: object
In [28]: emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
         emp['Domain']
Out[28]: 0
              Datascience
          1
                   Testing
          2
             Dataanalyst
                Analytics
          3
          4
                Statistics
                       NLP
          Name: Domain, dtype: object
In [29]:
         emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
         emp['Age']
Out[29]: 0
               34years
          1
                 45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
                  55yr
          Name: Age, dtype: object
```

```
In [30]:
          emp['Age'] = emp['Age'].str.extract('(\\d+)') # extract used for remove categori
          emp['Age']
Out[30]:
          0
                34
          1
                45
          2
               NaN
          3
               NaN
          4
                67
          5
                55
          Name: Age, dtype: object
In [31]: emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
          emp['Location']
Out[31]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
               Hyderbad
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
In [32]: emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
          emp['Salary']
Out[32]:
          0
                5000
               10000
          1
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: object
          emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
In [33]:
          emp['Exp']
Out[33]:
                 2
          1
                 3
          2
                 4
          3
               NaN
          4
                 5
          5
                10
          Name: Exp, dtype: object
In [34]:
          emp
Out[34]:
             Name
                                                        Exp
                       Domain
                                Age
                                       Location Salary
          0
              Mike
                    Datascience
                                  34
                                        Mumbai
                                                  5000
                                                           2
             Teddy
                        Testing
                                      Bangalore
                                                 10000
                                                           3
          2
              Umar
                    Dataanalyst
                                NaN
                                           NaN
                                                 15000
                                                           4
                                                 20000
          3
              Jane
                       Analytics
                                NaN
                                      Hyderbad
                                                 30000
                                                           5
          4
             Uttam
                       Statistics
                                  67
                                           NaN
          5
                           NLP
                                  55
                                          Delhi
                                                 60000
               Kim
                                                          10
```

```
In [35]:
         clean_data = emp.copy()
         clean_data
Out[35]:
            Name
                      Domain Age Location Salary
                                                      Exp
             Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                         2
         1 Teddy
                       Testing
                                 45 Bangalore
                                               10000
                                                         3
            Umar Dataanalyst NaN
                                                         4
         2
                                         NaN
                                              15000
                      Analytics NaN Hyderbad
         3
                                               20000 NaN
              Jane
           Uttam
                      Statistics
                                         NaN
                                               30000
                                                         5
                                 67
                          NLP
              Kim
                                 55
                                         Delhi 60000
                                                        10
```

EDA TECHNIQUES

Missing Value Treatement

```
In [36]: clean_data.isnull().sum()
Out[36]:
                        0
          Name
           Domain
                        0
                        2
           Age
           Location 2
           Salary
           Exp
           dtype: int64
In [37]: 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 4 non-null object
          2 Age 4 non-null
3 Location 4 non-null
4 Salary 6 non-null
                                           object
                                            object
                                            object
                          5 non-null
         dtypes: object(6)
         memory usage: 420.0+ bytes
          import numpy as np
In [38]:
In [39]: clean_data.head()
```

```
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
                                 3
                                               Jane
                                                                           Analytics
                                                                                                          NaN
                                                                                                                            Hyderbad
                                                                                                                                                               20000 NaN
                                        Uttam
                                                                            Statistics
                                                                                                               67
                                                                                                                                            NaN
                                                                                                                                                               30000
                                                                                                                                                                                                5
In [40]:
                                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_numeric
                                clean_data['Age']
In [41]:
Out[41]: 0
                                                             34
                                 1
                                                            45
                                 2
                                                  50.25
                                                  50.25
                                 3
                                 4
                                                             67
                                 5
                                                             55
                                 Name: Age, dtype: object
                                 clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp
In [42]:
                                 clean_data['Exp']
Out[42]:
                                 0
                                                         2
                                                         3
                                 2
                                                         4
                                 3
                                                  4.8
                                                         5
                                 4
                                                      10
                                 Name: Exp, dtype: object
In [43]:
                                 clean_data
Out[43]:
                                            Name
                                                                             Domain
                                                                                                                                  Location Salary Exp
                                                                                                             Age
                                 0
                                              Mike
                                                                  Datascience
                                                                                                                 34
                                                                                                                                    Mumbai
                                                                                                                                                                     5000
                                                                                                                                                                                                2
                                            Teddy
                                                                                                                               Bangalore
                                                                                                                                                                  10000
                                                                                Testing
                                                                                                                 45
                                                                                                                                                                                                3
                                                                   Dataanalyst
                                                                                                          50.25
                                                                                                                                                                  15000
                                 2
                                             Umar
                                                                                                                                              NaN
                                                                                                                                                                                                4
                                                                                                         50.25
                                                                                                                               Hyderbad
                                                                                                                                                                  20000
                                 3
                                               Jane
                                                                           Analytics
                                                                                                                                                                                            4.8
                                                                                                                                                                  30000
                                                                                                                                                                                                5
                                 4
                                          Uttam
                                                                            Statistics
                                                                                                                                              NaN
                                                                                                                 67
                                                                                        NLP
                                 5
                                                                                                                  55
                                                                                                                                                                  60000
                                                                                                                                                                                             10
                                                  Kim
                                                                                                                                            Delhi
                                clean_data['Location'].isnull().sum
```

Out[39]:

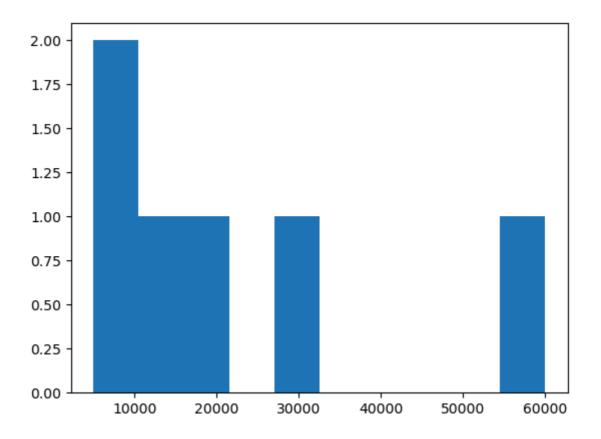
Name

```
Out[44]: <bound method Series.sum of 0
                                            False
          1
               False
          2
                True
          3
               False
          4
               True
          5
               False
          Name: Location, dtype: bool>
In [45]: clean_data['Location'].isnull().sum()
Out[45]: 2
In [46]:
         clean_data['Location']
Out[46]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
                Hyderbad
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
         clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mc
In [47]:
          clean_data['Location']
Out[47]:
         0
                  Mumbai
               Bangalore
          1
          2
               Bangalore
          3
                Hyderbad
               Bangalore
          4
          5
                   Delhi
          Name: Location, dtype: object
In [48]:
         clean_data
Out[48]:
             Name
                       Domain
                                 Age
                                       Location Salary Exp
          0
              Mike Datascience
                                  34
                                        Mumbai
                                                  5000
                                                          2
             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
          5
               Kim
                           NLP
                                   55
                                           Delhi
                                                 60000
                                                          10
In [49]: emp.info()
```

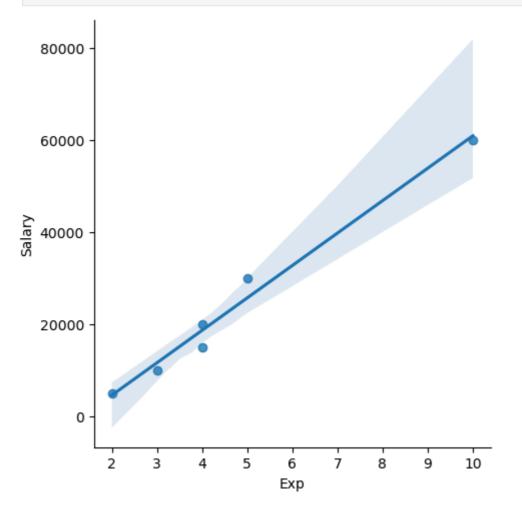
```
<class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          # Column Non-Null Count Dtype
         --- ----- -----
          0 Name 6 non-null object
1 Domain 6 non-null object
2 Age 4 non-null object
3 Location 4 non-null object
4 Salary 6 non-null object
5 Exp 5 non-null object
         dtypes: object(6)
         memory usage: 420.0+ bytes
In [50]: 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 [51]: clean_data['Age']=clean_data['Age'].astype(int)
          clean_data['Age']
Out[51]: 0
                34
               45
           1
           2
               50
           3 50
           4 67
           5
               55
           Name: Age, dtype: int32
In [52]: clean_data['Salary']=clean_data['Salary'].astype(int)
          clean_data['Salary']
Out[52]: 0
                5000
           1
               10000
           2
               15000
           3 20000
           4 30000
                60000
           Name: Salary, dtype: int32
In [53]: clean_data['Exp']=clean_data['Exp'].astype(int)
          clean_data['Exp']
```

```
Out[53]: 0
                2
          1
                3
          2
                4
          3
                4
          4
                5
          5
               10
          Name: Exp, dtype: int32
In [54]:
         clean_data
Out[54]:
             Name
                       Domain Age
                                      Location Salary Exp
          0
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                  5000
                                                          2
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                          3
          1
          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
In [55]:
          clean_data['Name']=clean_data['Name'].astype('category')
          clean_data['Domain']=clean_data['Domain'].astype('category')
          clean_data['Location']=clean_data['Location'].astype('category')
In [56]: 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
                                         category
         1
             Domain
                        6 non-null
                                         category
                        6 non-null
         2
             Age
                                         int32
                                         category
         3
             Location 6 non-null
         4
             Salary
                        6 non-null
                                         int32
         5
                        6 non-null
                                         int32
        dtypes: category(3), int32(3)
        memory usage: 866.0 bytes
In [57]:
          clean_data
Out[57]:
                                      Location Salary Exp
             Name
                       Domain Age
                                                          2
          0
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                  5000
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                          3
          1
          2
                    Dataanalyst
                                  50
                                                 15000
                                                          4
             Umar
                                      Bangalore
          3
                       Analytics
                                  50
                                      Hyderbad
                                                 20000
              Jane
                                                          4
                                                          5
                                                 30000
          4
             Uttam
                       Statistics
                                  67
                                      Bangalore
                           NLP
          5
               Kim
                                  55
                                          Delhi
                                                60000
                                                         10
```

```
In [58]:
        clean_data.to_csv('clean_data.csv')
In [59]: import os
         os.getcwd() # from os to get saved current working directory
Out[59]: 'C:\\Users\\Hanshu\\basics'
         EDA TECNIQUES APPLYING
In [60]:
         import matplotlib.pyplot as plt
                                          # visualiztion
         import seaborn as sns
In [61]:
         import warnings
         warnings.filterwarnings('ignore')
        clean_data['Salary']
In [62]:
Out[62]: 0
               5000
         1
              10000
         2
              15000
         3
              20000
         4
              30000
              60000
         Name: Salary, dtype: int32
In [63]: vis1 = sns.distplot(clean_data['Salary'])
               1e-5
           3.5
           3.0
           2.5
           2.0
           1.5
           1.0
           0.5
           0.0
             -40000 -20000
                                        20000
                                                 40000
                                                                  80000
                                                                          100000
                                  0
                                                         60000
                                             Salary
In [64]: vis2 = plt.hist(clean_data['Salary'])
```



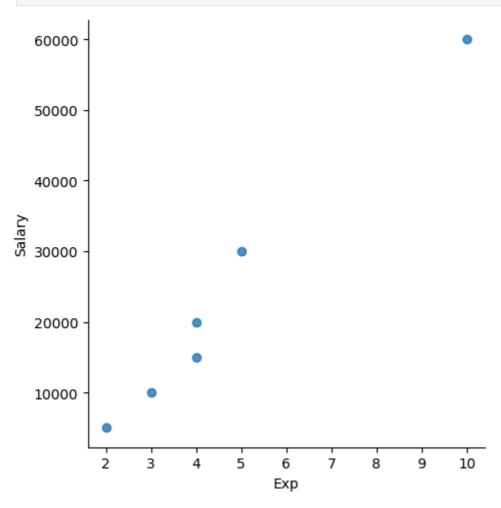
In [65]: vis4 = sns.lmplot(data=clean_data,x='Exp',y='Salary')



In [66]: clean_data

Out[66]:		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 [67]: vis5=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=False)



In [68]: clean_data[:]

```
Out[68]:
             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
               Kim
                           NLP
                                  55
                                          Delhi
                                                 60000
                                                         10
In [69]:
          clean_data[0:6:2]
Out[69]:
             Name
                       Domain Age
                                       Location
                                                Salary Exp
          0
              Mike Datascience
                                  34
                                       Mumbai
                                                  5000
                                                          2
                                      Bangalore
              Umar
                    Dataanalyst
                                  50
                                                 15000
             Uttam
                       Statistics
                                  67
                                      Bangalore
                                                 30000
                                                          5
          clean_data[::-1]
In [70]:
Out[70]:
             Name
                       Domain Age
                                       Location
                                                Salary Exp
                           NLP
          5
               Kim
                                  55
                                          Delhi
                                                 60000
                                                         10
            Uttam
                                      Bangalore
                                                 30000
                       Statistics
                                                          5
                                  67
                       Analytics
                                      Hyderbad
                                                 20000
          3
              Jane
                                  50
                                                          4
                                      Bangalore
          2
              Umar
                    Dataanalyst
                                                 15000
                                  50
                                                          4
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                          3
              Mike
                    Datascience
                                  34
                                        Mumbai
                                                  5000
                                                          2
In [71]: clean_data.columns
Out[71]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [73]: x_iv = clean_data[['Name','Domain','Age','Location','Exp']]
In [75]: x_iv
                       # variable identification
```

```
Out[75]:
             Name
                        Domain Age
                                       Location Exp
          0
                     Datascience
                                         Mumbai
                                                    2
               Mike
                                   34
              Teddy
                                       Bangalore
                                                    3
                         Testing
                                   45
              Umar
                     Dataanalyst
                                       Bangalore
                                                    4
          2
                                   50
                                       Hyderbad
          3
               Jane
                        Analytics
                                   50
                                                    4
                                       Bangalore
             Uttam
                        Statistics
                                   67
                                                    5
          5
                Kim
                            NLP
                                   55
                                            Delhi
                                                   10
          y_dv = clean_data[['Salary']]
In [76]:
Out[76]:
             Salary
               5000
          0
              10000
              15000
          2
              20000
              30000
              60000
In [77]:
          emp
Out[77]:
              Name
                        Domain
                                  Age
                                        Location
                                                  Salary
                                                           Exp
          0
               Mike
                     Datascience
                                   34
                                         Mumbai
                                                    5000
                                                             2
              Teddy
                                   45
                                        Bangalore
                                                   10000
                                                             3
                         Testing
          2
              Umar
                     Dataanalyst
                                  NaN
                                            NaN
                                                   15000
                                                             4
                                  NaN
                                        Hyderbad
                                                   20000 NaN
          3
               Jane
                        Analytics
          4
             Uttam
                        Statistics
                                   67
                                            NaN
                                                   30000
                                                             5
                            NLP
                                   55
                                            Delhi
                                                   60000
                                                            10
                Kim
```

In [78]:

clean_data

Out[78]:		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 [79]:	x_	iv					
Out[79]:		Name	Domain	Age	Location	Ехр	
	0	Mike	Datascience	34	Mumbai	2	
	1	Teddy	Testing	45	Bangalore	3	
	2	Umar	Dataanalyst	50	Bangalore	4	
	3	Jane	Analytics	50	Hyderbad	4	
	4	Uttam	Statistics	67	Bangalore	5	
	5	Kim	NLP	55	Delhi	10	
In [80]:	У_	dv					
Out[80]:		Salary					
	0	5000					
	1	10000					
	2	15000					
	3	20000					
	4	30000					
	5	60000					

In [81]: clean_data

Out[81]:		Name	e D	omain	Age	Locat	ion	Salary	Ехр			
	0	Mike	e Datas	cience	34	Mun	nbai	5000	2			
	1	Teddy	/	Гesting	45	Banga	lore	10000	3			
	2	Uma	r Dataa	analyst	50	Banga	lore	15000	4			
	3	Jane	e Ar	alytics	50	Hyder	bad	20000	4			
	4	Uttan	n St	atistics	67	Banga	lore	30000	5			
	5	Kin	า	NLP	55	D	elhi	60000	10			
In [82]:	im	putati	ion = po	d.get_	dummie	es(clea	n_da	ata)				
In [83]:	im	putati	Lon									
Out[83]:		Age		Ехр	Name	_Jane	Nam	ne_Kim	Name	_Mike	Name_Teddy	Name_Umar
	0	34	5000	2		False		False		True	False	False
	1	45	10000	3		False		False		False	True	False
	2	50	15000	4		False		False		False	False	True
	3	50	20000	4		True		False		False	False	False
	4	67	30000	5		False		False		False	False	False
	5	55	60000	10		False		True		False	False	False
	4											•
In []:	cl	ean_										
In []:												
In []:												
In []:												
In []:												
In []:												
In []:												
In []:												
In []:												
In []:												
In []:												
In []:												

In []:	
In [[]:	
In [[]:	
In []:	
In []:	
In [[]:	
In []:	
In [[]:	
In [[]:	
In []:	
In [[]:	
In []:	
In [[]:	
In [[]:	
In [[]:	
In []:	
In [[]:	
In []:	
In [[]:	

In []:	
In []:	
In []:	