```
import pandas as pd
In [1]:
         emp = pd.read_excel(r"C:\Users\Asus.LAPTOP-EMBE8J70\Downloads\Rawdata.xlsx")
In [2]:
         emp
In [3]:
                          Domain
Out[3]:
             Name
                                      Age
                                            Location
                                                        Salary
                                                                  Exp
         0
              Mike
                     Datascience#$ 34 years
                                                       5^00#0
                                             Mumbai
                                                                   2+
         1 Teddy^
                           Testing
                                     45' yr Bangalore 10%%000
                                                                   <3
            Uma#r Dataanalyst^^#
                                                      1$5%000
                                      NaN
                                                NaN
                                                                 4> yrs
              Jane
                       Ana^^lytics
                                      NaN Hyderbad
                                                       2000^0
                                                                  NaN
         3
                          Statistics
            Uttam*
                                     67-yr
                                                NaN
                                                        30000- 5+ year
                              NLP
                                                      6000^$0
               Kim
                                      55yr
                                                Delhi
                                                                  10+
In [4]:
         pd.__version__
         '1.4.2'
Out[4]:
         id(emp)
In [5]:
         1505493363824
Out[5]:
         emp.columns
In [6]:
         Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[6]:
         emp.shape
In [7]:
         (6, 6)
Out[7]:
         emp.head()
In [8]:
Out[8]:
             Name
                          Domain
                                      Age
                                            Location
                                                        Salary
                                                                  Exp
              Mike
                     Datascience#$
                                   34 years
                                             Mumbai
                                                       5^00#0
                                                                   2+
         1 Teddy^
                           Testing
                                     45' yr Bangalore 10%%000
                                                                   <3
                    Dataanalyst^^#
                                                      1$5%000
            Uma#r
                                      NaN
                                                NaN
                                                                 4> yrs
                       Ana^^lytics
                                      NaN Hyderbad
                                                       2000^0
                                                                  NaN
         3
               Jane
            Uttam*
                          Statistics
                                     67-yr
                                                NaN
                                                        30000- 5+ year
         emp.tail()
In [9]:
```

```
Out[9]:
              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
                        Ana^^lytics
         3
               Jane
                                     NaN Hyderbad
                                                       2000^0
                                                                  NaN
                           Statistics
                                                        30000- 5+ year
             Uttam*
                                     67-yr
                                               NaN
          5
                               NLP
                                     55yr
                                               Delhi
                                                      6000^$0
                                                                   10+
                Kim
```

## In [10]: emp.info()

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

#	Column	Non-Null Count	Dtype
0	Name	6 non-null	object
1	Domain	6 non-null	object
2	Age	4 non-null	object
3	Location	4 non-null	object
4	Salary	6 non-null	object
5	Exp	5 non-null	object
dtvp	es: object	(6)	

memory usage: 416.0+ bytes

### emp In [11]:

#### **Domain** Location Out[11]: Name Age Salary Exp 0 Mike Datascience#\$ 34 years 5^00#0 2+ Mumbai Teddy^ Testing 45' yr Bangalore 10%%000 <3 Uma#r Dataanalyst^^# 2 NaN NaN 1\$5%000 4> yrs 3 Jane Ana^^lytics NaN Hyderbad 2000^0 NaN 4 Uttam\* Statistics 67-yr NaN 30000- 5+ year

NLP

55yr

Delhi

6000^\$0

10+

## emp.isnull() In [12]:

5

Out[12]:

Kim

		Name	Domain	Age	Location	Salary	Exp
	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 [13]:
           emp.isna()
Out[13]:
              Name Domain
                               Age Location
                                              Salary
                                                      Exp
          0
               False
                        False
                             False
                                        False
                                               False False
               False
                             False
           1
                        False
                                        False
                                               False False
          2
               False
                        False
                              True
                                               False False
                                        True
          3
               False
                        False
                              True
                                        False
                                               False
                                                      True
           4
               False
                        False False
                                        True
                                               False False
               False
                        False False
                                        False
                                               False False
In [14]:
           emp.isnull().sum()
          Name
Out[14]:
          Domain
                        0
          Age
                        2
          Location
                        2
          Salary
          Exp
          dtype: int64
In [15]:
          emp.columns
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[15]:
```

# DATA CLEANING OR DATA CLEANSING

```
emp['Name']
In [16]:
                 Mike
Out[16]:
          1
               Teddy^
          2
                Uma#r
          3
                 Jane
          4
               Uttam*
          5
                  Kim
          Name: Name, dtype: object
          emp['Name'] = emp['Name'].str.replace(r'\W','', regex=True)
In [17]:
          emp['Name']
                Mike
Out[17]:
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
In [18]:
          emp
```

```
Out[18]:
             Name
                          Domain
                                      Age
                                            Location
                                                        Salary
                                                                  Exp
              Mike
                     Datascience#$
                                  34 years
                                             Mumbai
                                                       5^00#0
                                                                   2+
          0
                                     45' yr
             Teddy
                           Testing
                                           Bangalore
                                                     10%%000
                                                                   <3
                    Dataanalyst^^#
                                                      1$5%000
                                                                4> yrs
          2
              Umar
                                      NaN
                                                NaN
          3
              Jane
                       Ana^^lytics
                                      NaN
                                           Hyderbad
                                                       2000^0
                                                                  NaN
          4 Uttam
                          Statistics
                                     67-yr
                                                NaN
                                                        30000-
                                                               5+ year
          5
               Kim
                              NLP
                                      55yr
                                               Delhi
                                                      6000^$0
                                                                  10+
          emp['Domain']
In [19]:
                 Datascience#$
Out[19]:
          1
                       Testing
          2
               Dataanalyst^^#
          3
                   Ana^^lytics
          4
                    Statistics
          5
                           NLP
          Name: Domain, dtype: object
          emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
In [20]:
          emp['Domain']
In [21]:
               Datascience
Out[21]:
          1
                    Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
          5
                        NLP
          Name: Domain, dtype: object
          emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
In [22]:
          emp['Age']
In [23]:
               34years
Out[23]:
          1
                   45yr
          2
                    NaN
          3
                    NaN
          4
                   67yr
          5
                   55yr
          Name: Age, dtype: object
          emp['Age'] = emp['Age'].str.extract('(\d+)')
In [24]:
          emp['Age']
In [25]:
                 34
Out[25]:
                 45
          1
          2
               NaN
          3
               NaN
          4
                67
          5
                 55
          Name: Age, dtype: object
```

```
In [26]:
          emp
Out[26]:
             Name
                        Domain
                                Age
                                       Location
                                                   Salary
                                                             Ехр
              Mike
                                       Mumbai
                                                  5^00#0
          0
                    Datascience
                                  34
                                                              2+
             Teddy
                                      Bangalore
                                                10%%000
                        Testing
                                  45
                                                              <3
          2
              Umar
                     Dataanalyst NaN
                                          NaN
                                                 1$5%000
                                                           4> yrs
          3
               Jane
                       Analytics NaN
                                      Hyderbad
                                                  2000^0
                                                             NaN
          4 Uttam
                       Statistics
                                  67
                                           NaN
                                                   30000-
                                                          5+ year
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                 6000^$0
                                                             10+
In [27]:
          emp['Location']
                   Mumbai
Out[27]:
          1
                Bangalore
          2
                      NaN
          3
                 Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
In [28]:
          emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
          emp['Location']
In [29]:
                   Mumbai
Out[29]:
          1
                Bangalore
          2
                      NaN
          3
                 Hyderbad
          4
                      NaN
                    Delhi
          Name: Location, dtype: object
In [30]:
Out[30]:
             Name
                        Domain Age
                                       Location
                                                   Salary
                                                             Exp
              Mike
                    Datascience
                                       Mumbai
                                                  5^00#0
                                                              2+
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                10%%000
                                                              <3
          2
              Umar
                     Dataanalyst
                                NaN
                                           NaN
                                                 1$5%000
                                                           4> yrs
               Jane
                       Analytics
                                      Hyderbad
                                                  2000^0
                                                             NaN
          3
                                NaN
                                                          5+ year
          4 Uttam
                                           NaN
                                                   30000-
                       Statistics
                                  67
               Kim
                           NLP
                                  55
                                          Delhi
                                                 6000^$0
                                                             10+
          emp['Salary']
In [31]:
```

```
5^00#0
Out[31]:
               10%%000
          1
          2
               1$5%000
          3
                2000^0
          4
                 30000-
          5
                6000^$0
          Name: Salary, dtype: object
          emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
In [32]:
          emp['Salary']
In [33]:
                 5000
Out[33]:
                10000
          1
          2
                15000
          3
                20000
          4
                30000
          5
                60000
          Name: Salary, dtype: object
In [34]:
          emp
Out[34]:
             Name
                       Domain
                                Age
                                      Location Salary
                                                         Exp
              Mike
                    Datascience
                                 34
                                      Mumbai
                                                 5000
                                                          2+
             Teddy
                        Testing
                                 45
                                     Bangalore
                                                10000
                                                           <3
              Umar
                    Dataanalyst
                               NaN
                                          NaN
                                                15000
                                                        4> yrs
                                                20000
          3
              Jane
                      Analytics
                               NaN
                                     Hyderbad
                                                         NaN
                       Statistics
                                                30000
          4 Uttam
                                 67
                                          NaN
                                                      5+ year
          5
                          NLP
                                 55
                                         Delhi
                                                60000
                                                         10+
               Kim
          emp['Exp']
In [35]:
                     2+
Out[35]:
          1
                     <3
          2
                4> yrs
          3
                    NaN
          4
                5+ year
          5
                    10+
          Name: Exp, dtype: object
In [36]:
          emp['Exp'] = emp['Exp'].str.extract('(\d+)')
          emp['Exp']
In [37]:
                  2
Out[37]:
          1
                  3
          2
                  4
          3
               NaN
          4
                  5
          5
                 10
          Name: Exp, dtype: object
In [38]:
          emp
```

```
Out[38]:
             Name
                       Domain Age
                                       Location Salary
                                                        Exp
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                  5000
                                                          2
          0
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                10000
                                                          3
              Umar
                                                15000
          2
                     Dataanalyst NaN
                                          NaN
                                                          4
          3
               Jane
                       Analytics
                                NaN
                                      Hyderbad
                                                20000
                                                       NaN
                                                30000
                                                          5
          4 Uttam
                       Statistics
                                  67
                                          NaN
                                                60000
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                         10
          clean_data = emp.copy()
In [39]:
In [40]:
          clean_data
Out[40]:
             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
                                      Hyderbad
          3
               Jane
                       Analytics
                                NaN
                                                20000
                                                       NaN
                       Statistics
                                          NaN
                                                30000
                                                          5
          4 Uttam
                                  67
          5
                           NLP
                                  55
                                          Delhi
                                                60000
                Kim
                                                         10
In [41]:
          clean_data.isnull().sum()
                        0
          Name
Out[41]:
          Domain
                        0
                        2
          Age
          Location
                        2
          Salary
                        0
          Exp
          dtype: int64
          import numpy as np
In [42]:
          clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']))
In [43]:
          clean_data['Age']
In [44]:
                   34
Out[44]:
          1
                   45
                50.25
          2
          3
                50.25
          4
                   67
          5
                   55
          Name: Age, dtype: object
          clean data
In [45]:
```

```
Out[45]:
             Name
                       Domain
                                 Age
                                       Location Salary
                                                        Ехр
              Mike
                                        Mumbai
                                                  5000
                                                          2
          0
                    Datascience
                                  34
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                          3
          2
              Umar
                     Dataanalyst 50.25
                                           NaN
                                                 15000
                                                          4
          3
              Jane
                       Analytics 50.25
                                      Hyderbad
                                                 20000
                                                        NaN
                                                          5
          4 Uttam
                       Statistics
                                  67
                                           NaN
                                                 30000
                           NLP
               Kim
                                   55
                                          Delhi
                                                 60000
                                                          10
          clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']))
In [46]:
In [47]:
          clean_data['Exp']
                  2
Out[47]:
                  3
          2
                  4
          3
                4.8
          4
                  5
                 10
          Name: Exp, dtype: object
In [48]:
          clean data
Out[48]:
                                       Location Salary Exp
             Name
                       Domain
                                 Age
              Mike
                    Datascience
                                  34
                                        Mumbai
                                                  5000
                                                          2
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                          3
              Umar
                     Dataanalyst 50.25
                                           NaN
                                                 15000
                                                          4
               Jane
                       Analytics 50.25
                                       Hyderbad
                                                 20000
                                                         4.8
          4 Uttam
                       Statistics
                                  67
                                           NaN
                                                 30000
                                                          5
                           NLP
                                   55
                                          Delhi
                                                 60000
               Kim
                                                         10
          clean data['Location'].isnull().sum()
In [49]:
Out[49]:
          clean_data['Location']
In [50]:
                   Mumbai
Out[50]:
                Bangalore
          2
                      NaN
          3
                 Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode()[@
In [51]:
          clean_data['Location']
In [52]:
```

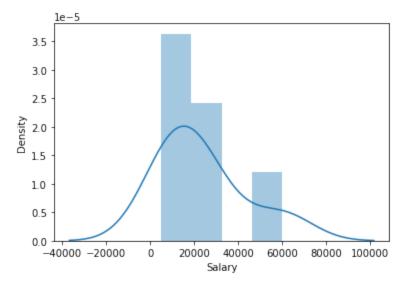
```
Mumbai
Out[52]:
          1
                Bangalore
          2
                Bangalore
          3
                 Hyderbad
          4
                Bangalore
          5
                    Delhi
          Name: Location, dtype: object
In [53]:
          clean_data
Out[53]:
             Name
                        Domain
                                 Age
                                       Location Salary Exp
              Mike
                    Datascience
                                  34
                                        Mumbai
                                                  5000
                                                          2
             Teddy
                        Testing
                                  45
                                       Bangalore
                                                 10000
                                                          3
                                                          4
              Umar
                     Dataanalyst
                                50.25
                                       Bangalore
                                                 15000
          3
               Jane
                       Analytics
                                50.25
                                       Hyderbad
                                                 20000
                                                         4.8
                       Statistics
                                                 30000
                                                          5
          4
             Uttam
                                   67
                                       Bangalore
          5
                Kim
                           NLP
                                   55
                                           Delhi
                                                 60000
                                                          10
In [55]:
          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
                                             object
                           6 non-null
           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: 416.0+ bytes
          clean_data['Age'] = clean_data['Age'].astype(int)
In [56]:
          clean_data
In [57]:
Out[57]:
             Name
                        Domain Age
                                      Location Salary Exp
          0
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                 5000
                                                         2
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                10000
                                                         3
          2
              Umar
                     Dataanalyst
                                     Bangalore
                                                15000
                                                         4
          3
               Jane
                       Analytics
                                  50
                                      Hyderbad
                                                20000
                                                        4.8
                                                         5
          4
             Uttam
                       Statistics
                                  67
                                      Bangalore
                                                30000
                                                60000
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                        10
          clean_data.info()
In [59]:
```

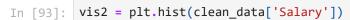
```
<class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
              Column
                         Non-Null Count Dtype
                         -----
          ---
          0
              Name
                         6 non-null
                                         object
          1
              Domain
                         6 non-null
                                         object
          2
              Age
                         6 non-null
                                         int32
          3
              Location 6 non-null
                                         object
              Salary
                         6 non-null
                                         object
          5
              Exp
                         6 non-null
                                         object
         dtypes: int32(1), object(5)
         memory usage: 392.0+ bytes
         clean_data['Exp'] = clean_data['Exp'].astype(int)
In [61]:
          clean_data
In [62]:
            Name
                      Domain Age
                                   Location Salary Exp
Out[62]:
         0
             Mike
                  Datascience
                               34
                                    Mumbai
                                             5000
                                                     2
                                            10000
            Teddy
                      Testing
                               45
                                  Bangalore
                                                     3
         2
             Umar
                   Dataanalyst
                               50
                                  Bangalore
                                            15000
                                                     4
         3
             Jane
                     Analytics
                               50
                                   Hyderbad
                                            20000
                                                     5
         4 Uttam
                                  Bangalore
                                            30000
                     Statistics
                               67
                         NLP
                               55
                                            60000
          5
              Kim
                                      Delhi
                                                    10
In [63]:
         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
                                         int32
              Age
                         6 non-null
              Location 6 non-null
          3
                                         object
          4
              Salary
                         6 non-null
                                         object
          5
                         6 non-null
                                         int32
         dtypes: int32(2), object(4)
         memory usage: 368.0+ bytes
         clean_data['Salary'] = clean_data['Salary'].astype(int)
In [64]:
         clean_data
In [65]:
```

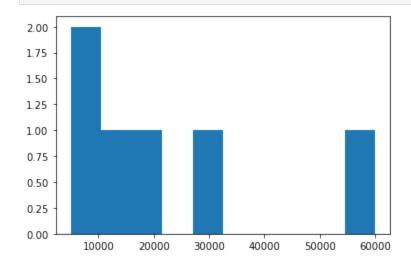
```
Out[65]:
             Name
                       Domain Age
                                     Location Salary Exp
          0
              Mike
                   Datascience
                                 34
                                      Mumbai
                                                5000
                                                        2
             Teddy
                        Testing
                                 45
                                    Bangalore
                                               10000
                                                        3
          2
              Umar
                    Dataanalyst
                                 50
                                     Bangalore
                                               15000
                                                        4
          3
              Jane
                      Analytics
                                     Hyderbad
                                               20000
                                                        5
          4 Uttam
                       Statistics
                                 67
                                     Bangalore
                                               30000
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                       10
          clean_data.info()
In [66]:
          <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
               Domain
                          6 non-null
                                            object
           1
                                            int32
           2
               Age
                          6 non-null
           3
               Location 6 non-null
                                            object
               Salary
                          6 non-null
                                            int32
           5
                          6 non-null
                                            int32
               Exp
          dtypes: int32(3), object(3)
          memory usage: 344.0+ bytes
In [67]:
          clean_data
Out[67]:
             Name
                       Domain Age
                                     Location Salary Exp
          0
              Mike
                    Datascience
                                 34
                                      Mumbai
                                                5000
                                                        2
             Teddy
                                 45
                                    Bangalore
                                               10000
                                                        3
                        Testing
          2
             Umar
                    Dataanalyst
                                    Bangalore
                                               15000
                                                        4
          3
              Jane
                      Analytics
                                 50
                                     Hyderbad
                                               20000
                                                        4
                                    Bangalore
          4 Uttam
                       Statistics
                                 67
                                               30000
                                                        5
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                       10
          clean_data['Name'] = clean_data['Name'].astype('category')
In [70]:
          clean_data.info()
In [71]:
```

```
<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
                                     object
         2
             Age
                      6 non-null
                                     int32
         3
             Location 6 non-null
                                     object
             Salary
                      6 non-null
                                     int32
         5
             Exp
                      6 non-null
                                     int32
        dtypes: category(1), int32(3), object(2)
        memory usage: 522.0+ bytes
In [72]: clean_data['Domain'] = clean_data['Domain'].astype('category')
In [73]: 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
             Domain
         1
                      6 non-null
                                     category
             Age
                      6 non-null
                                     int32
         3
             Location 6 non-null
                                     object
            Salary
                      6 non-null
                                     int32
                      6 non-null
                                     int32
        dtypes: category(2), int32(3), object(1)
        memory usage: 700.0+ bytes
In [74]: clean_data['Location'] = clean_data['Location'].astype('category')
In [75]: clean_data.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
         # Column
                      Non-Null Count Dtype
        ---
                      -----
            Name
                      6 non-null
         0
                                     category
         1
             Domain
                      6 non-null
                                     category
         2
            Age
                      6 non-null
                                     int32
             Location 6 non-null
         3
                                     category
         4
                      6 non-null
             Salary
                                     int32
         5
                      6 non-null
                                     int32
             Exp
        dtypes: category(3), int32(3)
        memory usage: 862.0 bytes
        clean data
In [79]:
```

```
Out[79]:
             Name
                       Domain Age
                                      Location Salary Exp
              Mike
                                  34
                                       Mumbai
                                                  5000
                                                          2
          0
                    Datascience
              Teddy
                         Testing
                                  45
                                      Bangalore
                                                10000
                                                          3
              Umar
                                                15000
          2
                     Dataanalyst
                                  50
                                      Bangalore
                                                          4
          3
               Jane
                       Analytics
                                      Hyderbad
                                                20000
                                                          5
          4 Uttam
                       Statistics
                                  67
                                      Bangalore
                                                30000
           5
               Kim
                           NLP
                                  55
                                          Delhi
                                                60000
                                                         10
           clean_data.to_csv('clean_data.csv')
In [80]:
In [81]:
           import os
           os.getcwd()
           'C:\\Users\\Asus.LAPTOP-EMBE8J70'
Out[81]:
In [82]:
           clean_data
             Name
                                      Location Salary Exp
Out[82]:
                        Domain Age
          0
              Mike
                    Datascience
                                  34
                                       Mumbai
                                                 5000
                                                          2
              Teddy
                                  45
                                      Bangalore
                                                10000
                                                          3
                         Testing
                                      Bangalore
          2
              Umar
                     Dataanalyst
                                  50
                                                15000
                                                          4
                                      Hyderbad
                                                20000
          3
               Jane
                       Analytics
                                  50
                                                         4
          4
             Uttam
                       Statistics
                                  67
                                      Bangalore
                                                30000
                                                          5
          5
                           NLP
                                  55
                                          Delhi
                                                60000
                                                         10
               Kim
           import matplotlib.pyplot as plt
In [85]:
           import seaborn as sns
           import warnings
In [87]:
           warnings.filterwarnings('ignore')
          clean_data['Salary']
In [88]:
                 5000
Out[88]:
                10000
          2
                15000
          3
                20000
          4
                30000
          5
                60000
          Name: Salary, dtype: int32
In [91]:
          vis1 = sns.distplot(clean_data['Salary'])
```





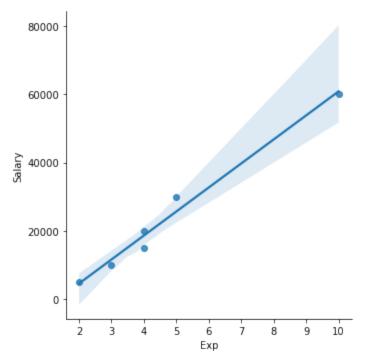


In [94]: clean\_data

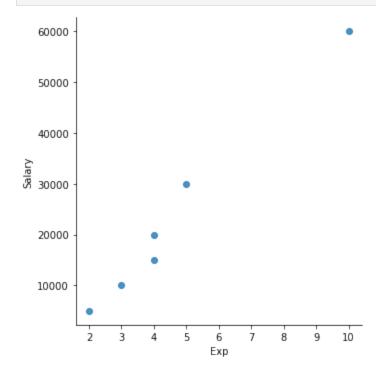
in [5+]. Crean\_date

Out[94]:	: Name		Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	50	Bangalore	15000	4
	3	Jane	Analytics	50	Hyderbad	20000	4
	4	Uttam	Statistics	67	Bangalore	30000	5
	5	Kim	NLP	55	Delhi	60000	10

In [98]: vis4 = sns.lmplot(data = clean\_data, x = 'Exp', y = 'Salary')







In [101... clean\_data[0:6:2]

Out[101]:	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 [102... clean\_data[::-1]

```
Out[102]:
               Name
                          Domain Age
                                          Location
                                                   Salary
                                                           Exp
            5
                              NLP
                                     55
                                             Delhi
                                                    60000
                                                             10
                 Kim
               Uttam
                          Statistics
                                     67
                                         Bangalore
                                                    30000
                                                              5
                                                    20000
            3
                 Jane
                          Analytics
                                     50
                                         Hyderbad
                                                              4
            2
                Umar
                       Dataanalyst
                                     50
                                         Bangalore
                                                    15000
            1
                Teddy
                           Testing
                                     45
                                         Bangalore
                                                    10000
                                                              3
                Mike
                       Datascience
                                          Mumbai
                                                     5000
                                                              2
            clean_data.columns
In [103...
            Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[103]:
            clean_data
In [105...
Out[105]:
               Name
                          Domain Age
                                          Location
                                                   Salary Exp
                                                              2
                Mike
                       Datascience
                                          Mumbai
                                                     5000
                                                    10000
                Teddy
                                     45
                                         Bangalore
                                                              3
                           Testing
                                         Bangalore
            2
                Umar
                       Dataanalyst
                                                    15000
                                                              4
                                         Hyderbad
                                                    20000
            3
                 Jane
                          Analytics
                                                             4
            4
               Uttam
                          Statistics
                                     67
                                         Bangalore
                                                    30000
                                                              5
            5
                  Kim
                             NLP
                                     55
                                             Delhi
                                                    60000
                                                             10
            x_iv = clean_data[['Name','Domain','Age','Location','Exp']]
In [107...
In [108...
            x_iv
Out[108]:
               Name
                          Domain Age
                                          Location Exp
            0
                Mike
                       Datascience
                                     34
                                          Mumbai
                                                      2
               Teddy
                           Testing
                                     45
                                         Bangalore
                                                      3
            2
                Umar
                       Dataanalyst
                                     50
                                         Bangalore
                                                      4
            3
                 Jane
                          Analytics
                                     50
                                         Hyderbad
            4 Uttam
                          Statistics
                                     67
                                         Bangalore
                                                      5
            5
                  Kim
                              NLP
                                     55
                                             Delhi
                                                     10
            y_dv = clean_data[['Salary']]
In [111...
In [112...
            y_dv
```

```
Out[112]:
               Salary
                 5000
            0
                10000
               15000
            2
                20000
                30000
                60000
In [113...
                                                            Ехр
Out[113]:
               Name
                          Domain
                                  Age
                                          Location Salary
                Mike
                       Datascience
                                     34
                                          Mumbai
                                                     5000
                                                              2
                                                    10000
                                         Bangalore
               Teddy
                           Testing
                                     45
                                                              3
            2
                Umar
                       Dataanalyst NaN
                                              NaN
                                                    15000
                                                              4
                 Jane
                         Analytics
                                         Hyderbad
                                                    20000
            3
                                   NaN
                                                           NaN
            4 Uttam
                          Statistics
                                     67
                                              NaN
                                                    30000
                                                              5
                  Kim
                             NLP
                                     55
                                             Delhi
                                                    60000
                                                             10
            clean_data
In [114...
Out[114]:
                                         Location Salary Exp
               Name
                          Domain Age
                Mike
                       Datascience
                                          Mumbai
                                                     5000
                                                             2
               Teddy
                           Testing
                                         Bangalore
                                                    10000
                                                             3
            2
                Umar
                       Dataanalyst
                                         Bangalore
                                                    15000
                                                             4
            3
                 Jane
                         Analytics
                                         Hyderbad
                                                    20000
            4 Uttam
                          Statistics
                                     67
                                         Bangalore
                                                    30000
                                                             5
            5
                 Kim
                             NLP
                                     55
                                             Delhi
                                                    60000
                                                            10
            imputation = pd.get_dummies(clean_data,dtype=int)
In [115...
```

imputation

In [116...

```
Out[116]:
              Age Salary Exp Name Jane Name Kim Name Mike Name Teddy Name Umar Name Uttam I
                                                                                          0
           0
                34
                     5000
                             2
                                         0
                                                    0
                                                                1
                                                                             0
                                                                                                       0
           1
                45
                    10000
                             3
                                         0
                                                    0
                                                                0
                                                                             1
                                                                                          0
                                                                                                       0
                                                    0
                                                                             0
           2
                50
                    15000
                             4
                                         0
                                                                0
                                                                                          1
                                                                                                       0
           3
                50
                    20000
                                         1
                                                                0
                                                                             0
                                                                                          0
                             5
                                         0
                                                    0
                                                                0
                                                                             0
                                                                                          0
           4
                67
                    30000
                                                                                                       1
                    60000
            5
                55
                            10
                                         0
                                                                0
                                                                             0
                                                                                          0
            clean_data
In [117...
Out[117]:
               Name
                        Domain Age
                                       Location Salary Exp
               Mike
                     Datascience
                                   34
                                        Mumbai
                                                  5000
                                                          2
                                                 10000
               Teddy
                                      Bangalore
                                                          3
                         Testing
           2
               Umar
                      Dataanalyst
                                   50
                                      Bangalore
                                                 15000
                                                          4
                                                 20000
           3
                Jane
                        Analytics
                                       Hyderbad
                                                          4
                                      Bangalore
                                                 30000
                                                          5
           4 Uttam
                        Statistics
                                   67
            5
                            NLP
                                          Delhi
                                                 60000
                                                         10
                 Kim
                                   55
           len(clean_data)
In [118...
Out[118]:
            imputation.columns
In [119...
           Index(['Age', 'Salary', 'Exp', 'Name_Jane', 'Name_Kim', 'Name_Mike',
Out[119]:
                   'Name_Teddy', 'Name_Umar', 'Name_Uttam', 'Domain_Analytics',
                   'Domain_Dataanalyst', 'Domain_Datascience', 'Domain_NLP',
                   'Domain_Statistics', 'Domain_Testing', 'Location_Bangalore',
                   'Location_Delhi', 'Location_Hyderbad', 'Location_Mumbai'],
                  dtype='object')
           len(imputation.columns)
In [120...
           19
Out[120]:
  In [ ]:
```