```
In [1]: import pandas as pd
        emp = pd.read_excel(r'C:\Users\user\Eda and stat\14th - Eda practicle\EDA- Practicl
In [2]:
In [3]: emp
Out[3]:
             Name
                          Domain
                                      Age
                                             Location
                                                         Salary
                                                                    Exp
         0
              Mike
                     Datascience#$
                                   34 years
                                              Mumbai
                                                        5^00#0
                                                                     2+
         1 Teddy^
                           Testing
                                      45' yr
                                            Bangalore
                                                       10%%000
                                                                     <3
            Uma#r
                    Dataanalyst^^#
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
                                      NaN
         3
              Jane
                       Ana^^lytics
                                            Hyderbad
                                                        2000^0
                                                                   NaN
                                      NaN
            Uttam*
                          Statistics
                                      67-yr
                                                 NaN
                                                         30000-
                                                                 5+ year
         5
               Kim
                              NLP
                                      55yr
                                                Delhi
                                                       6000^$0
                                                                    10+
In [4]: emp.columns
Out[4]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [5]:
        emp.shape # Ruturns the shape of the array
Out[5]: (6, 6)
        emp.head() #Diaplay the first few rows
In [6]:
Out[6]:
             Name
                          Domain
                                       Age
                                             Location
                                                         Salary
                                                                    Exp
                     Datascience#$
              Mike
                                   34 years
                                              Mumbai
                                                        5^00#0
                                                                     2+
           Teddy^
                                            Bangalore
                                                                     <3
                           Testing
                                      45' yr
                                                       10%%000
                    Dataanalyst^^#
            Uma#r
                                      NaN
                                                 NaN
                                                       1$5%000
                                                                  4> yrs
                       Ana^^lytics
                                            Hyderbad
                                                        2000^0
         3
              Jane
                                      NaN
                                                                   NaN
            Uttam*
                          Statistics
                                      67-yr
                                                 NaN
                                                         30000-
                                                                5+ year
In [7]: emp.tail() # The function is usd to view the last few rows of a data frame
```

```
Out[7]:
            Name
                          Domain
                                   Age
                                         Location
                                                     Salary
                                                               Exp
        1 Teddy^
                          Testing
                                  45' yr
                                        Bangalore 10%%000
                                                                <3
                   Dataanalyst^^#
            Uma#r
                                   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 [8]: emp.info() #Display the summary of the dataframe

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5

Data columns (total 6 columns):

Column Non-Null Count Divne

#	COTUIIII	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 [9]: emp

Out[9]:	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+	

In [12]: emp.isnull() # To Detecting the missing values

```
Out[12]:
             Name Domain
                             Age Location Salary
                                                       Exp
          0
               False
                        False
                              False
                                        False
                                                False False
               False
                              False
                                        False
                                                False False
          1
                        False
          2
               False
                        False
                              True
                                        True
                                                False False
                                                     True
          3
               False
                        False
                              True
                                        False
                                                False
          4
               False
                        False False
                                        True
                                                False False
               False
                        False False
                                        False
                                                False False
          emp.isnull().sum() # In Pandas, emp.isnull().sum() is used to identify and count mi
In [13]:
Out[13]:
          Name
          Domain
          Age
                       2
          Location
                       2
          Salary
                       0
          Exp
                       1
          dtype: int64
In [14]:
         emp.columns
Out[14]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [15]:
          emp
Out[15]:
              Name
                             Domain
                                         Age
                                                Location
                                                            Salary
                                                                        Exp
          0
                Mike
                                      34 years
                                                                        2+
                       Datascience#$
                                                Mumbai
                                                            5^00#0
             Teddy^
                                               Bangalore
                              Testing
                                        45' yr
                                                          10%%000
                                                                         <3
              Uma#r
                      Dataanalyst^^#
                                         NaN
                                                    NaN
                                                          1$5%000
                                                                     4> yrs
          3
                Jane
                         Ana^^lytics
                                               Hyderbad
                                                            2000^0
                                                                       NaN
                                         NaN
          4
             Uttam*
                            Statistics
                                         67-yr
                                                    NaN
                                                            30000-
                                                                    5+ year
          5
                 Kim
                                NLP
                                         55yr
                                                   Delhi
                                                          6000^$0
                                                                       10+
          emp['Name']
In [16]:
Out[16]:
          0
                  Mike
                Teddy^
          1
          2
                 Uma#r
          3
                  Jane
                Uttam*
          4
          5
                   Kim
          Name: Name, dtype: object
In [18]: emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
```

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

```
34 years
Out[25]: 0
                 45' yr
          1
          2
                    NaN
          3
                    NaN
          4
                  67-yr
          5
                   55yr
          Name: Age, dtype: object
In [26]: emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
         emp['Age']
In [27]:
Out[27]:
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
                  55yr
          Name: Age, dtype: object
In [28]: emp['Age'] = emp['Age'].str.extract('(\d+)')
In [29]: emp['Age']
Out[29]:
          0
                34
                45
          2
               NaN
          3
               NaN
                67
          5
                55
          Name: Age, dtype: object
In [30]: emp
Out[30]:
             Name
                       Domain Age
                                      Location
                                                  Salary
                                                             Exp
          0
              Mike Datascience
                                 34
                                       Mumbai
                                                 5^00#0
                                                              2+
             Teddy
                        Testing
                                 45
                                     Bangalore
                                                10%%000
                                                              <3
          2
             Umar
                    Dataanalyst NaN
                                          NaN
                                                1$5%000
                                                           4> yrs
              Jane
                      Analytics NaN
                                     Hyderbad
                                                 2000^0
                                                            NaN
          4 Uttam
                       Statistics
                                 67
                                          NaN
                                                  30000-
                                                         5+ year
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                                6000^$0
                                                             10+
In [31]: emp['Location']
```

```
Out[31]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
               Hyderbad
          4
                     NaN
                   Delhi
          5
          Name: Location, dtype: object
In [32]: emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
         emp['Location']
In [33]:
Out[33]:
                  Mumbai
               Bangalore
          1
          2
                     NaN
          3
               Hyderbad
          4
                     NaN
                   Delhi
          Name: Location, dtype: object
In [34]: emp['Salary']
Out[34]:
                5^00#0
               10%%000
          1
          2
               1$5%000
          3
                2000^0
                30000-
          4
          5
               6000^$0
          Name: Salary, dtype: object
In [35]: emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
In [36]: emp['Salary']
                5000
Out[36]: 0
               10000
          1
          2
               15000
               20000
          3
               30000
          4
          5
               60000
          Name: Salary, dtype: object
In [37]: emp['Exp']
Out[37]: 0
                    2+
          1
                    <3
          2
                4> yrs
          3
                   NaN
          4
               5+ year
          5
                   10+
          Name: Exp, dtype: object
In [38]: emp['Exp']
```

```
Out[38]: 0
                    2+
          1
                    <3
          2
              4> yrs
          3
                   NaN
               5+ year
          5
                   10+
          Name: Exp, dtype: object
In [39]: emp['Exp'] = emp['Exp'].str.replace(r'\W','',regex=True)
In [40]:
         emp['Exp']
Out[40]:
                   2
                   3
          1
          2
                4yrs
          3
                 NaN
          4
               5year
                  10
          Name: Exp, dtype: object
In [41]: emp['Exp'] = emp['Exp'].str.extract('(\d+)')
In [43]: emp['Exp']
Out[43]:
         0
                 2
          1
                 3
          2
                 4
          3
               NaN
                 5
          5
                10
          Name: Exp, dtype: object
In [44]: emp
Out[44]:
             Name
                       Domain Age
                                      Location Salary
                                                       Exp
          0
              Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                         2
             Teddy
                       Testing
                                 45
                                     Bangalore
                                               10000
          2
             Umar Dataanalyst NaN
                                         NaN
                                               15000
                                                         4
          3
              Jane
                      Analytics NaN
                                     Hyderbad
                                               20000 NaN
          4 Uttam
                      Statistics
                                 67
                                         NaN
                                               30000
                                                         5
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                        10
In [45]: clean_data = emp.copy()
```

. Till now we have rawdata we use regex to to clean the data and removed all noises characted from the dataset

.You can also work in same thing in sql query as well

EDA TECHNIQUE LETS APPPLY

. Missing values treatment for numeric data

[46]:	cl	ean_dat	a					
t[46]:		Name	Domain	Age	Location	Salary	Ехр	
	0	Mike	Datascience	34	Mumbai	5000	2	
	1	Teddy	Testing	45	Bangalore	10000	3	
	2	Umar	Dataanalyst	NaN	NaN	15000	4	
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN	
	4	Uttam	Statistics	67	NaN	30000	5	
	5	Kim	NLP	55	Delhi	60000	10	
[47]:	cl	ean dat	a['Age']					
t[47]:			e, dtype: ob	ject				
[49]:			mpy as np			1 (11)		
[50]:	CT	ean_ɑat	a['Age'] = (cıean_	data[Age].+111n	ıa(np.r	mean(pd.to_numeric(clean_data[
[51]:	cl	ean_dat	a['Age']					
t[51]:	0 1 2 3 4 5 Na	4 50.2 50.2 6 5		ject				
[52]:	em	_						

```
Out[52]:
                                                                                Domain Age
                                             Name
                                                                                                                                     Location Salary
                                                                                                                                                                                                  Ехр
                                  0
                                                 Mike Datascience
                                                                                                                     34
                                                                                                                                        Mumbai
                                                                                                                                                                           5000
                                                                                                                                                                                                         2
                                              Teddy
                                                                                    Testing
                                                                                                                     45
                                                                                                                                   Bangalore
                                                                                                                                                                    10000
                                                                                                                                                                                                         3
                                  2
                                               Umar
                                                                       Dataanalyst NaN
                                                                                                                                                   NaN
                                                                                                                                                                       15000
                                                                                                                                                                                                         4
                                                  Jane
                                                                              Analytics NaN
                                                                                                                                   Hyderbad
                                                                                                                                                                       20000
                                                                                                                                                                                                NaN
                                  3
                                  4
                                             Uttam
                                                                               Statistics
                                                                                                                     67
                                                                                                                                                   NaN
                                                                                                                                                                        30000
                                                                                                                                                                                                         5
                                                    Kim
                                                                                            NLP
                                                                                                                     55
                                                                                                                                                 Delhi
                                                                                                                                                                        60000
                                                                                                                                                                                                      10
In [53]: clean_data
Out[53]:
                                                                                                                                        Location Salary
                                             Name
                                                                                Domain
                                                                                                                  Age
                                                                                                                                                                                                    Exp
                                                 Mike Datascience
                                                                                                                       34
                                                                                                                                          Mumbai
                                                                                                                                                                              5000
                                                                                                                                                                                                           2
                                             Teddy
                                                                                                                                     Bangalore
                                                                                                                                                                          10000
                                                                                                                                                                                                            3
                                                                                   Testing
                                                                                                                       45
                                  2
                                               Umar
                                                                       Dataanalyst 50.25
                                                                                                                                                     NaN
                                                                                                                                                                        15000
                                                                                                                                                                                                           4
                                  3
                                                  Jane
                                                                              Analytics 50.25
                                                                                                                                     Hyderbad
                                                                                                                                                                         20000
                                                                                                                                                                                                 NaN
                                            Uttam
                                                                               Statistics
                                                                                                                                                     NaN
                                                                                                                                                                          30000
                                                                                                                                                                                                            5
                                                                                                                       67
                                                    Kim
                                                                                            NLP
                                                                                                                       55
                                                                                                                                                                          60000
                                                                                                                                                   Delhi
                                                                                                                                                                                                         10
                               clean_data['Exp']
In [54]:
Out[54]:
                                                            2
                                                            3
                                   2
                                                            4
                                   3
                                                    NaN
                                                            5
                                   5
                                                         10
                                  Name: Exp, dtype: object
In [55]: clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp'])).fillna(np
                               clean_data['Exp']
In [56]:
Out[56]:
                                  0
                                                            2
                                                            3
                                   1
                                   2
                                                            4
                                   3
                                                    4.8
                                   4
                                                            5
                                                         10
                                  Name: Exp, dtype: object
In [57]: clean_data
```

```
Out[57]:
             Name
                        Domain
                                  Age
                                        Location Salary Exp
          0
              Mike
                     Datascience
                                   34
                                         Mumbai
                                                    5000
                                                            2
                                        Bangalore
              Teddy
                         Testing
                                   45
                                                   10000
                                                            3
          2
              Umar
                     Dataanalyst 50.25
                                            NaN
                                                   15000
                                                            4
          3
               Jane
                       Analytics 50.25
                                        Hyderbad
                                                   20000
                                                           4.8
          4
             Uttam
                        Statistics
                                   67
                                            NaN
                                                   30000
                                                            5
                Kim
                           NLP
                                   55
                                            Delhi
                                                   60000
                                                           10
In [58]:
          clean_data['Location'].isnull().sum()
Out[58]: 2
          clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode(
In [59]:
          clean_data['Location']
In [60]:
Out[60]:
                   Mumbai
          1
                Bangalore
          2
                Bangalore
                 Hyderbad
          3
          4
                Bangalore
                    Delhi
          Name: Location, dtype: object
          clean_data
In [61]:
Out[61]:
             Name
                        Domain
                                  Age
                                        Location Salary Exp
              Mike Datascience
                                   34
                                         Mumbai
                                                    5000
                                                            2
              Teddy
                                   45
                                        Bangalore
                                                   10000
                                                            3
                         Testing
          2
                     Dataanalyst 50.25
                                       Bangalore
                                                   15000
              Umar
                                                            4
                       Analytics 50.25
                                        Hyderbad
                                                   20000
          3
               Jane
                                                           4.8
                                                   30000
                                                            5
             Uttam
                        Statistics
                                   67
                                        Bangalore
                           NLP
                                                   60000
          5
                Kim
                                   55
                                            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
       --- ----- -----
           Name
        0
                  6 non-null
                                   object
        1
           Domain 6 non-null
                                   object
                6 non-null object object object
        2
           Age
        3
           Location 6 non-null
        4
           Salary
                    6 non-null
                                   object
        5
                    6 non-null
           Exp
                                   object
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [65]: clean_data['Age'] = clean_data['Age'].astype(int)
In [66]: 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
           Domain 6 non-null
        1
                                 object
        2
                   6 non-null
                                   int32
           Age
        3
           Location 6 non-null
                                 object
        4
                    6 non-null
           Salary
                                   object
        5
                    6 non-null
           Exp
                                   object
       dtypes: int32(1), object(5)
       memory usage: 396.0+ bytes
        clean_data['Salary'] = clean_data['Salary'].astype(int)
In [69]:
        clean_data['Exp'] = clean_data['Exp'].astype(int)
In [71]:
In [72]: 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
                    6 non-null
        2
                                   int32
           Age
        3
           Location 6 non-null
                                   object
        4
           Salary
                    6 non-null
                                   int32
        5
           Exp
                    6 non-null
                                   int32
       dtypes: int32(3), object(3)
       memory usage: 348.0+ bytes
In [74]: 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 [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
             Domain
                       6 non-null
         1
                                        category
         2
             Age
                       6 non-null
                                        int32
         3
             Location 6 non-null
                                        category
                                        int32
             Salary
                       6 non-null
         5
                       6 non-null
             Exp
                                        int32
        dtypes: category(3), int32(3)
        memory usage: 866.0 bytes
In [76]: clean_data
Out[76]:
             Name
                       Domain Age
                                     Location Salary Exp
          0
              Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                        2
             Teddy
                        Testing
                                     Bangalore
                                               10000
          2
             Umar
                    Dataanalyst
                                     Bangalore
                                               15000
                                                        4
          3
              Jane
                      Analytics
                                     Hyderbad
                                               20000
                                                        4
          4 Uttam
                      Statistics
                                 67
                                     Bangalore
                                               30000
                                                        5
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                       10
In [77]:
         clean_data.to_csv('clean_data.csv')
         import os
In [78]:
         os.getcwd()
Out[78]: 'C:\\Users\\user\\Eda and stat\\14th - Eda practicle\\EDA- Practicle'
In [79]:
         clean_data
Out[79]:
             Name
                       Domain Age
                                     Location Salary
                                                      Exp
          0
              Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                        2
                                     Bangalore
                                               10000
             Teddy
                        Testing
                                                        3
          2
             Umar Dataanalyst
                                 50
                                     Bangalore
                                               15000
                                                        4
          3
              Jane
                      Analytics
                                     Hyderbad
                                               20000
                                                        4
                                               30000
                                                        5
          4
           Uttam
                      Statistics
                                    Bangalore
                                 67
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                       10
```

```
import matplotlib.pyplot as plt #Visualization
In [80]:
          import seaborn as sns
In [81]: import warnings
          warnings.filterwarnings('ignore')
In [83]:
         clean_data['Salary']
                5000
Out[83]:
               10000
               15000
          2
               20000
               30000
               60000
          Name: Salary, dtype: int32
In [85]: vis1 = sns.distplot(clean_data['Salary'])
                1e-5
           3.5
           3.0
           2.5
           2.0
           1.5
           1.0
```

```
In [86]: vis2 = plt.hist(clean_data['Salary'])
```

20000

40000

Salary

80000

100000

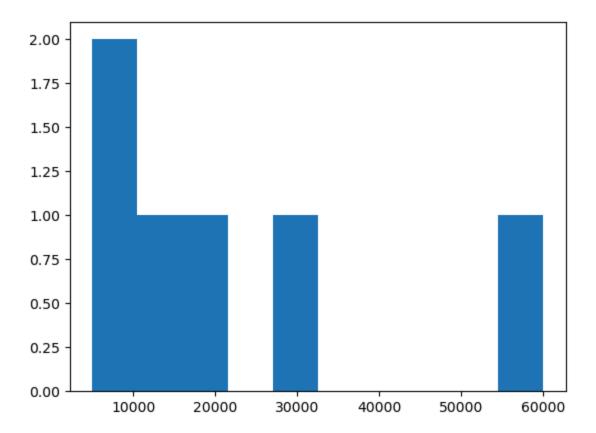
60000

Ó

0.5

0.0

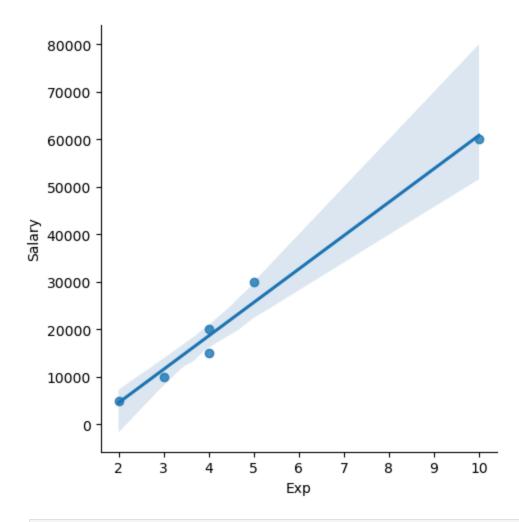
-40000 -20000



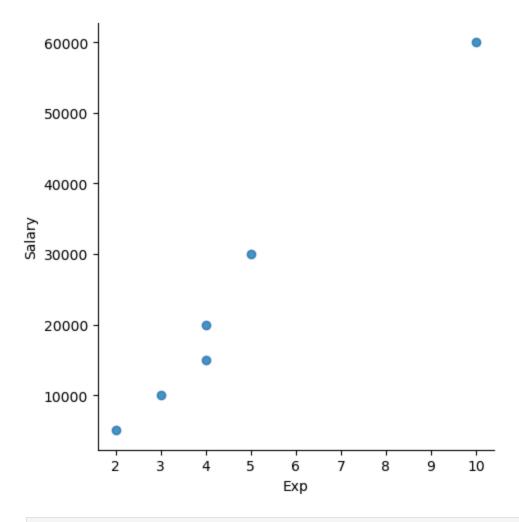
In [87]: clean_data

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



In [90]: vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary',fit_reg=False)



In [91]: clean_data[:]

_			-
\cap	144	101	
Vι	J L	1 フエ	

	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 [92]: clean_data[0:6:2]

Out[92]:

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

```
In [93]:
          clean_data[::-1]
Out[93]:
              Name
                         Domain Age
                                         Location
                                                   Salary
                                                          Exp
           5
                 Kim
                             NLP
                                    55
                                            Delhi
                                                   60000
                                                            10
              Uttam
                         Statistics
                                        Bangalore
                                                   30000
                                                             5
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
                                                   20000
                                                             4
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                             4
           1
              Teddy
                          Testing
                                    45
                                        Bangalore
                                                   10000
                                                             3
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
          clean_data.columns
 In [94]:
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
 In [95]:
           clean_data
 Out[95]:
              Name
                         Domain Age
                                         Location Salary
                                                          Exp
                Mike Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
               Teddy
                                        Bangalore
                                                   10000
                                                             3
                          Testing
                                    45
           2
               Umar
                      Dataanalyst
                                        Bangalore
                                                   15000
                                                             4
                                    50
                                        Hyderbad
                                                   20000
           3
                Jane
                        Analytics
                                                             4
              Uttam
                         Statistics
                                        Bangalore
                                                   30000
                                                             5
                                    67
           5
                             NLP
                                                   60000
                                                            10
                 Kim
                                    55
                                            Delhi
In [100...
           X_iv = clean_data[['Name','Domain','Age','Location','Salary','Exp']]
          X_iv
In [97]:
Out[97]:
                         Domain Age
                                         Location Salary
                                                          Exp
              Name
                Mike Datascience
                                          Mumbai
                                                    5000
           0
                                    34
                                                             2
               Teddy
                          Testing
                                        Bangalore
                                                   10000
                                                             3
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                             4
           3
                                        Hyderbad
                                                   20000
                Jane
                        Analytics
                                                             4
           4
                                        Bangalore
                                                   30000
                                                             5
              Uttam
                         Statistics
                                    67
           5
                 Kim
                             NLP
                                    55
                                                   60000
                                                            10
                                            Delhi
 In [98]: y_dv = clean_data['Salary']
```

```
In [99]:
          y_dv
Out[99]:
                  5000
                 10000
           1
           2
                 15000
           3
                 20000
           4
                 30000
                 60000
           5
           Name: Salary, dtype: int32
In [101...
           emp
Out[101...
              Name
                         Domain Age
                                         Location Salary
                                                            Ехр
                Mike Datascience
                                          Mumbai
                                                     5000
                                                              2
           0
                                    34
                                         Bangalore
                                                    10000
               Teddy
                          Testing
                                    45
                                                              3
                                                    15000
                                                              4
           2
               Umar
                      Dataanalyst NaN
                                             NaN
                         Analytics NaN
                                                    20000
           3
                                         Hyderbad
                                                           NaN
                Jane
           4
              Uttam
                         Statistics
                                    67
                                             NaN
                                                    30000
                                                              5
           5
                                                    60000
                 Kim
                             NLP
                                    55
                                             Delhi
                                                             10
In [102...
           clean_data
Out[102...
              Name
                         Domain Age
                                         Location Salary
                                                           Ехр
           0
                Mike Datascience
                                    34
                                          Mumbai
                                                     5000
                                                             2
                                        Bangalore
                                                    10000
                                                             3
               Teddy
                          Testing
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                             4
           3
                                                   20000
                Jane
                         Analytics
                                        Hyderbad
                                                             4
                                        Bangalore
                                                   30000
                                                             5
              Uttam
                         Statistics
                                    67
                                                    60000
                 Kim
                             NLP
                                    55
                                             Delhi
                                                            10
```

In [103...

X_iv

Out[103		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 [104	у_	dv					
Out[104	0 1 2 3 4	500 1000 1500 2000 3000	90 90 90				
	5 Na	6000		int3	2		
In [105	Na	6000	ary, dtype:	int3	2		
In [105 Out[105	Na	6000 nme: Sal	ary, dtype:		2 Location	Salary	Exp
	Na	6000 nme: Sal ean_dat	oo ary, dtype: a			Salary 5000	Exp 2
	cl	6000 nme: Sal ean_dat Name	Domain	Age	Location		
	cl 0	6000 nme: Sal ean_dat Name Mike	Domain Datascience	Age 34	Location Mumbai	5000	2
	0 1	6000 nme: Sal ean_dat Name Mike Teddy	Domain Datascience Testing	Age 34 45	Location Mumbai Bangalore	5000	2
	0 1 2	6000 nme: Sal ean_dat Name Mike Teddy Umar	Domain Datascience Testing Dataanalyst	Age 34 45 50	Location Mumbai Bangalore Bangalore	5000 10000 15000	2 3 4
	0 1 2	6000 nme: Sal ean_dat Name Mike Teddy Umar Jane	Domain Datascience Testing Dataanalyst Analytics	Age 34 45 50 50	Location Mumbai Bangalore Bangalore Hyderbad	5000 10000 15000 20000	2 3 4 4
	0 1 2 3	6000 mme: Sal ean_dat Name Mike Teddy Umar Jane Uttam	Domain Datascience Testing Dataanalyst Analytics Statistics	Age 34 45 50 50 67	Location Mumbai Bangalore Bangalore Hyderbad Bangalore	5000 10000 15000 20000 30000	2 3 4 4 5

imputation

In [108...

Out[108		Age	Salary	Ехр	Name_	_Jane	Name	e_Kim	Name	_Mike	Name_Teddy	Name_Umar	Nan	
	0	34	5000	2		0		0		1	0	0		
	1	45	10000	3		0		0		0	1	0		
	2	50	15000	4		0		0		0	0	1		
	3	50	20000	4		1		0		0	0	0		
	4	67	30000	5		0		0		0	0	0		
	5	55	60000	10		0		1		0	0	0		
In [109	cle	ean_da	ta											
Out[109		Name	. De	omain	Age	Locat	ion	Salary	Ехр					
	0	Mike	Datas	cience	34	Mum	nbai	5000	2					
	1	Teddy	, 1	Testing	45	Bangal	ore	10000	3					
	2	Umar	Dataa	analyst	50	Bangal	ore	15000	4					
	3	Jane	e An	alytics	50	Hyderk	bad	20000	4					
	4	Uttam	Sta	atistics	67	Bangal	ore	30000	5					
	5	Kim	l	NLP	55	De	elhi	60000	10					
In [110	lei	n(clea	n_data)										
Out[110	6	•												
-														
In [111	im	putati	on.colu	umns										
Out[111	<pre>Index(['Age', 'Salary', 'Exp', 'Name_Jane', 'Name_Kim', 'Name_Mike',</pre>													
In [112	lei	n(impu	tation	colum	ıns)									
Out[112	19													
In []:														