```
In [1]:
         import pandas as pd
 In [3]: pd.__version__
Out[3]: '2.2.2'
 In [5]: emp=pd.read_excel(r'C:\Users\DELL\Downloads\Rawdata.xlsx')
 In [7]: emp
Out[7]:
             Name
                          Domain
                                           Location
                                                      Salary
                                     Age
                                                                Exp
                     Datascience#$ 34 years
         0
              Mike
                                            Mumbai
                                                      5^00#0
                                                                 2+
         1 Teddy^
                                    45' yr Bangalore 10%%000
                           Testing
                                                                  <3
            Uma#r Dataanalyst^^#
                                  NaN
                                               NaN
                                                    1$5%000
                                                               4> yrs
                                  NaN Hyderbad
                                                    2000^0
         3
              Jane
                       Ana^^lytics
                                                                NaN
            Uttam*
                         Statistics
                                    67-yr
                                               NaN
                                                    30000- 5+ year
               Kim
                             NLP
                                     55yr
                                              Delhi
                                                     6000^$0
                                                                10+
 In [9]: id(emp)
Out[9]: 2139623741232
In [32]: emp.shape
Out[32]: (6, 6)
In [36]: len(emp)
Out[36]: 6
In [26]: emp.columns
Out[26]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [38]: len(emp.columns)
Out[38]: 6
In [40]: 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
         0 Name 6 non-null
         3 Location 4 non-null
4 Salary 6 non-null
5 Exp 5 non-null
                                        object
                                         object
                                         object
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [42]:
         emp
Out[42]:
              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
          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^$0
                                                                      10+
          emp['Name']
In [44]:
Out[44]: 0
                 Mike
               Teddy^
          1
          2
               Uma#r
          3
                 Jane
          4
               Uttam*
          5
          Name: Name, dtype: object
In [46]: emp['Domain']
Out[46]: 0
                Datascience#$
          1
                      Testing
               Dataanalyst^^#
          2
          3
                Ana^^lytics
                  Statistics
          4
                           NLP
          Name: Domain, dtype: object
In [48]: emp['Age']
Out[48]: 0
               34 years
          1
                 45' yr
          2
                   NaN
          3
                    NaN
          4
                  67-yr
          5
                   55yr
          Name: Age, dtype: object
```

```
In [50]: emp['Location']
Out[50]: 0
                Mumbai
         1
             Bangalore
         2
                    NaN
         3 Hyderbad
                    NaN
         5
                  Delhi
         Name: Location, dtype: object
In [52]: emp['Salary']
Out[52]: 0
              5^00#0
             10%%000
         2
             1$5%000
              2000^0
         3
         4
              30000-
             6000^$0
         Name: Salary, dtype: object
In [54]: emp['Exp']
Out[54]: 0
                   2+
         1
                   <3
         2 4> yrs
                NaN
             5+ year
         4
         5
                  10+
         Name: Exp, dtype: object
In [56]: emp[['Name','Domain']]
Out[56]:
             Name
                         Domain
         0
              Mike
                     Datascience#$
         1 Teddy^
                          Testing
           Uma#r Dataanalyst^^#
         2
         3
                       Ana^^lytics
             Jane
                         Statistics
            Uttam*
                             NLP
         5
               Kim
In [60]: emp[['Name','Domain','Age']]
```

Out[60]:		Name	Domain	Age			
	0	Mike	Datascience#\$	34 years			
	1	Teddy^	Testing	45' yr			
	2	Uma#r	Dataanalyst^^#	NaN			
	3	Jane	Ana^^lytics	NaN			
	4	Uttam*	Statistics	67-yr			
	5	Kim	NLP	55yr			
In [62]:	em	p[['Name	','Domain','Age	e','Locat	ion','Sala	rv','Exp']	1
Out[62]:		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 [64]:	em	р					
In [64]: Out[64]:	em	p Name	Domain	Age	Location	Salary	Ехр
	em 0		Domain Datascience#\$		Location Mumbai	Salary 5^00#0	Exp 2+
		Name Mike			Mumbai		
	0	Name Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
	0	Name Mike Teddy^	Datascience#\$	34 years 45' yr	Mumbai Bangalore	5^00#0 10%%000	2+
	0 1 2	Name Mike Teddy^ Uma#r	Datascience#\$ Testing Dataanalyst^^#	34 years 45' yr NaN	Mumbai Bangalore NaN	5^00#0 10%%000 1\$5%000	2+ <3 4> yrs
	0 1 2 3	Name Mike Teddy^ Uma#r Jane	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics	34 years 45' yr NaN NaN	Mumbai Bangalore NaN Hyderbad	5^00#0 10%%000 1\$5%000 2000^0	2+ <3 4> yrs NaN
Out[64]:	0 1 2 3 4 5	Name Mike Teddy^ Uma#r Jane Uttam* Kim	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP	34 years 45' yr NaN NaN 67-yr	Mumbai Bangalore NaN Hyderbad NaN	5^00#0 10%%000 1\$5%000 2000^0 30000-	2+ <3 4> yrs NaN 5+ year
Out[64]: In [28]:	0 1 2 3 4 5	Name Mike Teddy^ Uma#r Jane Uttam* Kim p.head()	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP	34 years 45' yr NaN NaN 67-yr 55yr	Mumbai Bangalore NaN Hyderbad NaN Delhi	5^00#0 10%%000 1\$5%000 2000^0 30000- 6000^\$0	2+ <3 4> yrs NaN 5+ year 10+
Out[64]:	0 1 2 3 4 5	Name Mike Teddy^ Uma#r Jane Uttam* Kim p.head() Name	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP Domain	34 years 45' yr NaN NaN 67-yr 55yr	Mumbai Bangalore NaN Hyderbad NaN Delhi Location	5^00#0 10%%000 1\$5%000 2000^0 30000- 6000^\$0	2+ <3 4> yrs NaN 5+ year 10+
Out[64]: In [28]:	0 1 2 3 4 5	Name Mike Teddy^ Uma#r Jane Uttam* Kim p.head() Name Mike	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP Domain Datascience#\$	34 years 45' yr NaN NaN 67-yr 55yr	Mumbai Bangalore NaN Hyderbad NaN Delhi Location Mumbai	5^00#0 10%%000 1\$5%000 2000^0 30000- 6000^\$0 Salary 5^00#0	2+ <3 4> yrs NaN 5+ year 10+ Exp 2+
Out[64]: In [28]:	0 1 2 3 4 5 em	Name Mike Teddy^ Uma#r Jane Uttam* Kim p.head() Name Mike Teddy^	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP Domain Datascience#\$ Testing	34 years 45' yr NaN NaN 67-yr 55yr Age 34 years 45' yr	Mumbai Bangalore NaN Hyderbad NaN Delhi Location Mumbai Bangalore	5^00#0 10%%000 1\$5%000 2000^0 30000- 6000^\$0 Salary 5^00#0 10%%000	2+ <3 4> yrs NaN 5+ year 10+ Exp 2+ <3
Out[64]: In [28]:	0 1 2 3 4 5 em	Name Mike Teddy^ Uma#r Jane Uttam* Kim p.head() Name Mike Teddy^ Uma#r	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP Domain Datascience#\$ Testing Dataanalyst^^#	34 years 45' yr NaN NaN 67-yr 55yr Age 34 years 45' yr NaN	Mumbai Bangalore NaN Hyderbad NaN Delhi Location Mumbai Bangalore NaN	5^00#0 10%%000 1\$5%000 2000^0 30000- 6000^\$0 Salary 5^00#0 10%%000 1\$5%000	2+
Out[64]: In [28]:	0 1 2 3 4 5 em	Name Mike Teddy^ Uma#r Jane Uttam* Kim p.head() Name Mike Teddy^	Datascience#\$ Testing Dataanalyst^^# Ana^^lytics Statistics NLP Domain Datascience#\$ Testing	34 years 45' yr NaN NaN 67-yr 55yr Age 34 years 45' yr	Mumbai Bangalore NaN Hyderbad NaN Delhi Location Mumbai Bangalore	5^00#0 10%%000 1\$5%000 2000^0 30000- 6000^\$0 Salary 5^00#0 10%%000	2+ <3 4> yrs NaN 5+ year 10+ Exp 2+ <3

```
In [15]: emp.tail()
Out[15]:
              Name
                            Domain
                                      Age
                                             Location
                                                         Salary
                                                                    Exp
             Teddy^
                             Testing
                                     45' yr
                                            Bangalore
                                                       10%%000
                                                                     <3
                      Dataanalyst^^#
                                                       1$5%000
              Uma#r
                                      NaN
                                                 NaN
                                                                  4> yrs
          3
                         Ana^^lytics
                                            Hyderbad
                                                         2000^0
               Jane
                                      NaN
                                                                    NaN
                                                         30000-
                                                                 5+ year
             Uttam*
                            Statistics
                                      67-yr
                                                 NaN
          5
                Kim
                                NLP
                                      55yr
                                                Delhi
                                                       6000^$0
                                                                    10+
In [17]:
          emp.isnull()
Out[17]:
             Name Domain Age Location Salary
                                                      Exp
              False
          0
                        False
                              False
                                        False
                                               False False
                       False
                              False
                                               False False
          1
              False
                                        False
          2
                       False
                                                     False
              False
                              True
                                        True
                                               False
                       False
          3
              False
                              True
                                        False
                                               False
                                                      True
          4
                       False
              False
                              False
                                        True
                                               False
                                                     False
          5
              False
                        False
                              False
                                        False
                                               False False
In [19]: emp.isna()
Out[19]:
             Name Domain Age Location Salary
```

			_		_	
0	False	False	False	Fals	e False	False
1	False	False	False	Fals	e False	False
2	False	False	True	Tru	e False	False
3	False	False	True	Fals	e False	True
4	False	False	False	Tru	e False	False
5	False	False	False	Fals	e False	False

In [21]: emp

```
Out[21]:
             Name
                           Domain
                                       Age
                                             Location
                                                         Salary
                                                                    Ехр
          0
               Mike
                      Datascience#$ 34 years
                                              Mumbai
                                                        5^00#0
                                                                    2+
          1 Teddy^
                                      45' yr Bangalore 10%%000
                           Testing
                                                                     <3
             Uma#r Dataanalyst^^#
          2
                                     NaN
                                                 NaN
                                                       1$5%000
                                                                 4> yrs
                                    NaN Hyderbad
          3
               Jane
                        Ana^^lytics
                                                        2000^0
                                                                   NaN
             Uttam*
                           Statistics
                                      67-yr
                                                 NaN
                                                         30000- 5+ year
          5
                Kim
                              NLP
                                       55yr
                                                Delhi
                                                       6000^$0
                                                                   10+
In [23]:
         emp.isnull().sum()
Out[23]:
          Name
                      0
          Domain
                      0
          Age
                      2
          Location
                      2
          Salary
          Exp
          dtype: int64
         Cleaning
In [66]:
         emp
Out[66]:
             Name
                           Domain
                                       Age
                                             Location
                                                         Salary
                                                                   Exp
                      Datascience#$ 34 years
          0
               Mike
                                              Mumbai
                                                        5^00#0
                                                                    2+
          1 Teddy^
                                      45' yr Bangalore
                                                      10%%000
                            Testing
                                                                     <3
             Uma#r Dataanalyst^^#
                                                       1$5%000
          2
                                     NaN
                                                 NaN
                                                                 4> yrs
                        Ana^^lytics
                                     NaN Hyderbad
                                                        2000^0
          3
               Jane
                                                                   NaN
             Uttam*
                           Statistics
                                      67-yr
                                                         30000-
                                                                5+ year
                                                 NaN
          5
                Kim
                              NLP
                                       55yr
                                                Delhi
                                                       6000^$0
                                                                   10+
In [68]:
         emp['Name']
Out[68]: 0
                 Mike
               Teddy^
          1
          2
               Uma#r
                 Jane
          3
          4
               Uttam*
                  Kim
          Name: Name, dtype: object
         emp['Name'] = emp['Name'].str.replace(r'\W', '', regex=True)
In [70]:
In [72]: emp['Name']
```

```
Out[72]: 0
                Mike
          1
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
In [74]: emp['Domain'] = emp['Domain'].str.replace(r'\W', '', regex=True)
In [76]:
         emp['Domain']
Out[76]: 0
               Datascience
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
                       NLP
          Name: Domain, dtype: object
         emp['Age'] = emp['Age'].str.replace(r'\W', '', regex=True)
In [78]:
In [80]:
         emp['Age']
Out[80]: 0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
                  55yr
          Name: Age, dtype: object
In [82]:
         emp['Age'] = emp['Age'].str.extract('(\\d+)')
In [84]:
         emp['Age']
          0
Out[84]:
                34
          1
                45
          2
               NaN
          3
               NaN
                67
          4
          5
                55
          Name: Age, dtype: object
In [86]:
          emp
Out[86]:
             Name
                       Domain
                                Age
                                       Location
                                                   Salary
                                                              Exp
          0
              Mike
                                                  5^00#0
                                                              2+
                    Datascience
                                  34
                                       Mumbai
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                10%%000
                                                               <3
                                                 1$5%000
          2
             Umar
                    Dataanalyst
                                NaN
                                          NaN
                                                            4> yrs
                                                  2000^0
          3
              Jane
                       Analytics
                                NaN
                                      Hyderbad
                                                             NaN
          4
             Uttam
                       Statistics
                                  67
                                          NaN
                                                   30000-
                                                           5+ year
                           NLP
                                  55
                                                 6000^$0
          5
               Kim
                                          Delhi
                                                              10+
```

```
In [88]: emp['Location']
Out[88]:
                   Mumbai
           0
           1
                Bangalore
           2
                       NaN
           3
                Hyderbad
           4
                       NaN
           5
                     Delhi
           Name: Location, dtype: object
           emp['Salary']
 In [90]:
Out[90]: 0
                 5^00#0
                10%%000
           2
                1$5%000
           3
                 2000^0
           4
                 30000-
                6000^$0
           Name: Salary, dtype: object
          emp['Salary'] = emp['Salary'].str.replace(r'\W', '',regex=True)
In [96]:
 In [98]:
           emp['Salary']
           0
                 5000
Out[98]:
                10000
           1
           2
                15000
           3
                20000
           4
                30000
                60000
           Name: Salary, dtype: object
In [100...
           emp
Out[100...
              Name
                        Domain
                                  Age
                                        Location Salary
                                                             Exp
           0
               Mike
                                   34
                                                   5000
                                                              2+
                     Datascience
                                         Mumbai
                                                   10000
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                              <3
           2
                                                   15000
               Umar
                      Dataanalyst
                                  NaN
                                            NaN
                                                           4> yrs
               Jane
                        Analytics
                                  NaN
                                        Hyderbad
                                                   20000
                                                            NaN
              Uttam
                        Statistics
                                   67
                                            NaN
                                                  30000
                                                          5+ year
                Kim
                            NLP
                                    55
                                            Delhi
                                                  60000
                                                             10+
In [102...
           emp['Exp']
Out[102...
                      2+
           1
                      <3
           2
                 4> yrs
           3
                    NaN
           4
                5+ year
           5
                    10+
           Name: Exp, dtype: object
          emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
In [104...
```

```
In [106...
            emp['Exp']
Out[106...
            0
                    2
            1
                    3
            2
                    4
            3
                  NaN
            4
                    5
            5
                   10
            Name: Exp, dtype: object
In [108...
            emp
Out[108...
               Name
                          Domain
                                    Age
                                           Location Salary
            0
                                                                 2
                Mike
                       Datascience
                                      34
                                            Mumbai
                                                       5000
            1
               Teddy
                            Testing
                                      45
                                           Bangalore
                                                       10000
                                                                 3
                                                                 4
            2
                Umar
                       Dataanalyst
                                    NaN
                                                NaN
                                                       15000
            3
                 Jane
                          Analytics
                                    NaN
                                           Hyderbad
                                                       20000 NaN
            4
               Uttam
                          Statistics
                                      67
                                                NaN
                                                       30000
                                                                 5
            5
                  Kim
                              NLP
                                      55
                                               Delhi
                                                       60000
                                                                10
In [110...
            clean_data=emp.copy()
In [112...
            emp
Out[112...
               Name
                          Domain
                                    Age
                                            Location
                                                      Salary
                                                               Exp
            0
                Mike
                       Datascience
                                      34
                                            Mumbai
                                                       5000
                                                                 2
               Teddy
                                      45
                                           Bangalore
                                                       10000
                                                                 3
            1
                            Testing
            2
                Umar
                                                      15000
                                                                 4
                       Dataanalyst
                                    NaN
                                                NaN
            3
                 Jane
                                           Hyderbad
                                                      20000
                                                              NaN
                          Analytics
                                    NaN
            4
                                                NaN
                                                      30000
                                                                 5
               Uttam
                          Statistics
                                      67
            5
                              NLP
                                      55
                                               Delhi
                                                                10
                 Kim
                                                       60000
In [169...
            clean_data
Out[169...
               Name
                          Domain
                                    Age
                                           Location
                                                      Salary Exp
            0
                Mike
                       Datascience
                                      34
                                            Mumbai
                                                       5000
                                                                2
                                                      10000
               Teddy
                            Testing
                                      45
                                          Bangalore
                                                                3
            2
                Umar
                       Dataanalyst
                                      50
                                          Bangalore
                                                      15000
                                                                4
            3
                 Jane
                                      50
                                           Hyderbad
                                                      20000
                          Analytics
                                                      30000
                                                                5
            4
               Uttam
                          Statistics
                                      67
                                          Bangalore
            5
                  Kim
                              NLP
                                      55
                                               Delhi
                                                      60000
                                                               10
```

```
clean_data
In [114...
Out[114...
                                         Name
                                                                      Domain
                                                                                               Age
                                                                                                                 Location Salary
                                                                                                                                                                    Exp
                                           Mike Datascience
                                0
                                                                                                    34
                                                                                                                    Mumbai
                                                                                                                                                 5000
                                                                                                                                                                           2
                                         Teddy
                                                                         Testing
                                                                                                    45 Bangalore
                                                                                                                                               10000
                                                                                                                                                                           3
                                1
                                2
                                          Umar
                                                            Dataanalyst NaN
                                                                                                                             NaN
                                                                                                                                              15000
                                                                                                                                                                           4
                                3
                                            Jane
                                                                    Analytics NaN Hyderbad
                                                                                                                                              20000
                                                                                                                                                                  NaN
                                4
                                       Uttam
                                                                     Statistics
                                                                                                                             NaN
                                                                                                                                              30000
                                                                                                                                                                           5
                                                                                                    67
                                5
                                               Kim
                                                                                NLP
                                                                                                    55
                                                                                                                            Delhi
                                                                                                                                              60000
                                                                                                                                                                        10
In [171...
                               clean_data.info()
                           <class 'pandas.core.frame.DataFrame'>
                           RangeIndex: 6 entries, 0 to 5
                           Data columns (total 6 columns):
                             #
                                         Column
                                                                  Non-Null Count Dtype
                                        -----
                            ---
                                                                       -----
                              0
                                        Name
                                                                   6 non-null
                                                                                                                        category
                              1
                                       Domain 6 non-null
                                                                                                                        category
                              2 Age
                                                                    6 non-null
                                                                                                                        int32
                                         Location 6 non-null
                                                                                                                        category
                              4
                                          Salary
                                                                       6 non-null
                                                                                                                        int32
                              5
                                                                        6 non-null
                                                                                                                        int32
                                          Exp
                           dtypes: category(3), int32(3)
                           memory usage: 866.0 bytes
In [116...
                               clean_data['Age']
Out[116...
                                0
                                                  34
                                1
                                                 45
                                2
                                               NaN
                                3
                                               NaN
                                4
                                                  67
                                                  55
                                Name: Age, dtype: object
In [118...
                               import numpy as np
                               clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age'])).fillna(np.to_numer
In [120...
In [122...
                               clean_data['Age']
Out[122...
                                0
                                                        34
                                1
                                                        45
                                2
                                               50.25
                                3
                                               50.25
                                4
                                                        67
                                5
                                                        55
                                Name: Age, dtype: object
In [124...
                            clean_data['Exp']
```

```
2
Out[124...
                                      1
                                                                3
                                      2
                                                               4
                                      3
                                                        NaN
                                                               5
                                      4
                                      5
                                                            10
                                      Name: Exp, dtype: object
                                     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.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']).fil
In [126...
In [128...
                                     clean_data['Exp']
Out[128...
                                      0
                                                               2
                                      1
                                                                3
                                      2
                                                                4
                                      3
                                                        4.8
                                      4
                                                                5
                                                            10
                                      Name: Exp, dtype: object
In [130...
                                     clean_data
Out[130...
                                                 Name
                                                                                    Domain
                                                                                                                      Age
                                                                                                                                            Location Salary Exp
                                      0
                                                    Mike Datascience
                                                                                                                           34
                                                                                                                                              Mumbai
                                                                                                                                                                                  5000
                                                                                                                                                                                                              2
                                                  Teddy
                                                                                                                                         Bangalore
                                                                                                                                                                              10000
                                      1
                                                                                       Testing
                                                                                                                           45
                                                                                                                                                                                                              3
                                                                          Dataanalyst 50.25
                                                                                                                                                        NaN
                                                                                                                                                                              15000
                                                                                                                                                                                                              4
                                      2
                                                   Umar
                                                                                  Analytics 50.25
                                                                                                                                          Hyderbad
                                                                                                                                                                              20000
                                      3
                                                     Jane
                                                                                                                                                                                                         4.8
                                                                                   Statistics
                                                                                                                                                        NaN
                                                                                                                                                                              30000
                                                                                                                                                                                                              5
                                      4
                                                Uttam
                                                                                                                           67
                                                        Kim
                                                                                                NLP
                                                                                                                           55
                                                                                                                                                       Delhi
                                                                                                                                                                             60000
                                      5
                                                                                                                                                                                                           10
                                     clean_data['Location'].isnull().sum()
In [134...
Out[134...
                                     clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mc
In [136...
In [138...
                                     clean_data['Location']
Out[138...
                                      0
                                                                   Mumbai
                                      1
                                                        Bangalore
                                      2
                                                        Bangalore
                                      3
                                                           Hyderbad
                                      4
                                                        Bangalore
                                      5
                                                                       Delhi
                                      Name: Location, dtype: object
In [140...
                                     clean_data
```

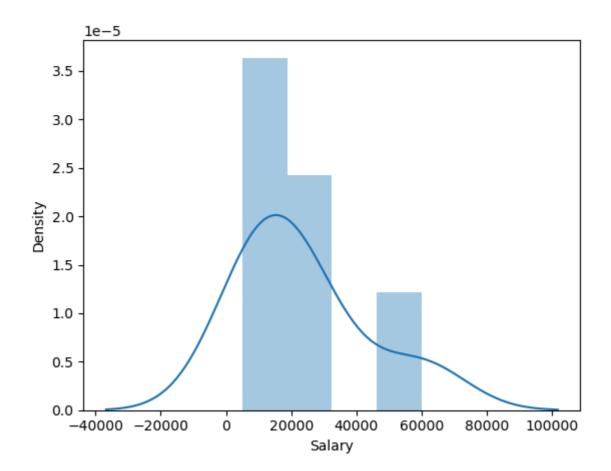
```
Out[140...
             Name
                       Domain
                                Age Location Salary Exp
          0
              Mike Datascience
                                  34
                                       Mumbai
                                                 5000
                                                         2
             Teddy
                                  45 Bangalore
                                                10000
                                                         3
                        Testing
          2
              Umar
                   Dataanalyst 50.25
                                      Bangalore
                                                15000
                                                         4
          3
                       Analytics 50.25
                                      Hyderbad
              Jane
                                                20000
                                                        4.8
             Uttam
                                      Bangalore
                                                30000
                                                         5
                       Statistics
                                  67
               Kim
                          NLP
                                  55
                                                60000
                                                        10
                                          Delhi
In [142...
          clean_data.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
             Column Non-Null Count Dtype
          0 Name 6 non-null
                                        object
          1 Domain 6 non-null
                                        object
          2 Age 6 non-null
                                        object
          3 Location 6 non-null
                                        object
          4
              Salary
                        6 non-null
                                        object
          5
              Exp
                        6 non-null
                                        object
         dtypes: object(6)
         memory usage: 420.0+ bytes
In [144...
          clean_data['Age'] = clean_data['Age'].astype(int)
In [146...
          clean_data['Age']
Out[146...
               34
          1
               45
          2
               50
          3
               50
          4
               67
               55
          Name: Age, dtype: int32
In [148...
          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
Domain 6 non-null
            Name
                       6 non-null
                                        object
                                        object
          1
                      6 non-null
                                        int32
          2
             Age
              Location 6 non-null
          3
                                        object
          4
              Salary
                        6 non-null
                                        object
          5
              Exp
                        6 non-null
                                        object
         dtypes: int32(1), object(5)
         memory usage: 396.0+ bytes
In [150...
          clean_data['Salary'] = clean_data['Salary'].astype(int)
          clean_data['Exp'] = clean_data['Exp'].astype(int)
```

```
In [152...
          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
          2
             Age
                      6 non-null
                                        int32
          3
             Location 6 non-null
                                        object
          4
             Salary 6 non-null
                                        int32
                        6 non-null
                                        int32
          5
              Exp
         dtypes: int32(3), object(3)
         memory usage: 348.0+ bytes
In [154...
          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 [156...
          clean_data.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          #
              Column
                      Non-Null Count Dtype
             ----
                        -----
          0
             Name
                      6 non-null
                                        category
          1
             Domain 6 non-null
                                        category
          2
                      6 non-null
             Age
                                        int32
          3
             Location 6 non-null
                                        category
          4
              Salary
                       6 non-null
                                        int32
          5
                        6 non-null
                                        int32
              Exp
         dtypes: category(3), int32(3)
         memory usage: 866.0 bytes
In [158...
          clean_data
Out[158...
             Name
                       Domain Age
                                     Location Salary Exp
          0
                                                5000
                                                        2
              Mike
                   Datascience
                                 34
                                      Mumbai
          1
             Teddy
                                 45
                                     Bangalore
                                               10000
                                                        3
                        Testing
          2
              Umar
                    Dataanalyst
                                 50
                                     Bangalore
                                               15000
                                                        4
          3
               Jane
                       Analytics
                                 50
                                     Hyderbad
                                               20000
                                                        5
          4
             Uttam
                       Statistics
                                 67
                                     Bangalore
                                               30000
                           NLP
          5
               Kim
                                 55
                                         Delhi
                                               60000
                                                       10
In [160...
          clean_data.to_csv('clean_data.csv')
In [162...
          import os
          os.getcwd()
Out[162...
          'C:\\Users\\DELL'
```

```
In [164...
           clean_data
Out[164...
               Name
                         Domain
                                  Age
                                         Location Salary Exp
           0
               Mike
                      Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
           1
               Teddy
                          Testing
                                    45
                                        Bangalore
                                                   10000
                                                             3
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                             4
           3
                Jane
                         Analytics
                                    50
                                        Hyderbad
                                                   20000
                                                             4
                                                   30000
                                                             5
           4
              Uttam
                         Statistics
                                    67
                                        Bangalore
           5
                             NLP
                 Kim
                                    55
                                             Delhi
                                                  60000
                                                            10
           EDA Techniques Let's Apply
In [173...
           import matplotlib.pyplot as plt
           import seaborn as sns
In [175...
           import warnings
           warnings.filterwarnings('ignore')
In [177...
           clean_data
Out[177...
              Name
                         Domain Age
                                         Location Salary Exp
           0
                                                             2
               Mike
                      Datascience
                                    34
                                          Mumbai
                                                    5000
               Teddy
                                    45
                                        Bangalore
                                                    10000
                                                             3
                          Testing
           2
                      Dataanalyst
                                    50
                                                   15000
                                                             4
               Umar
                                        Bangalore
           3
               Jane
                         Analytics
                                    50
                                        Hyderbad
                                                   20000
                                                             4
                                                             5
           4
              Uttam
                         Statistics
                                    67
                                        Bangalore
                                                   30000
           5
                             NLP
                 Kim
                                    55
                                             Delhi
                                                   60000
                                                            10
In [179...
           clean_data['Salary']
Out[179...
           0
                  5000
                 10000
           1
           2
                 15000
           3
                 20000
           4
                 30000
           5
                 60000
           Name: Salary, dtype: int32
```

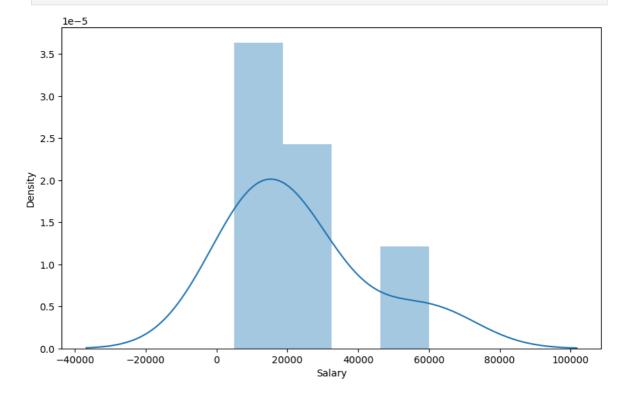
vis1 = sns.distplot(clean_data['Salary'])

In [185...

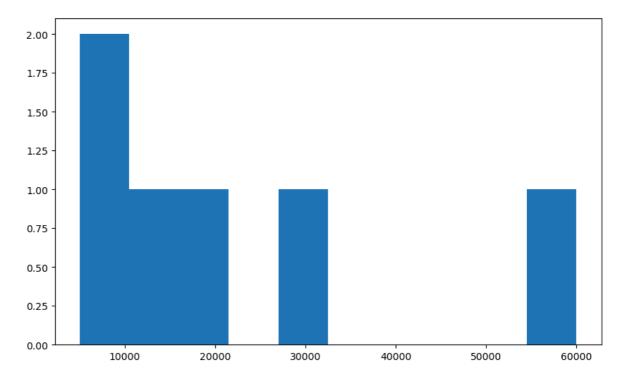


```
In [187... plt.rcParams['figure.figsize'] = 10,6
```

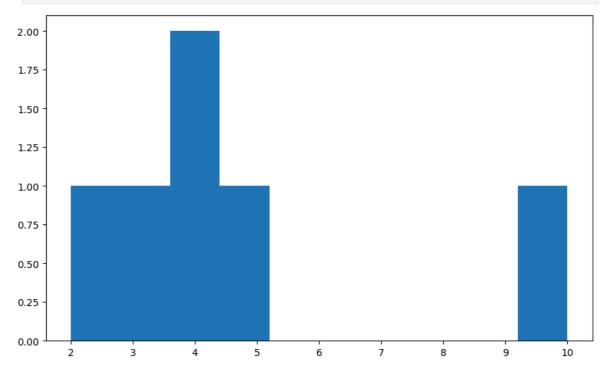
In [189... vis1 = sns.distplot(clean_data['Salary'])



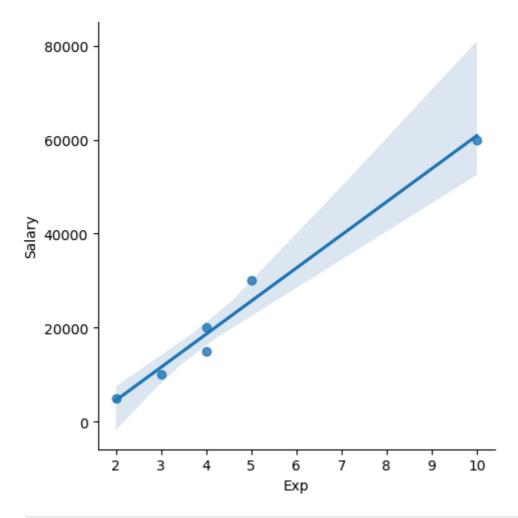
```
In [191... vis2 = plt.hist(clean_data['Salary'])
```



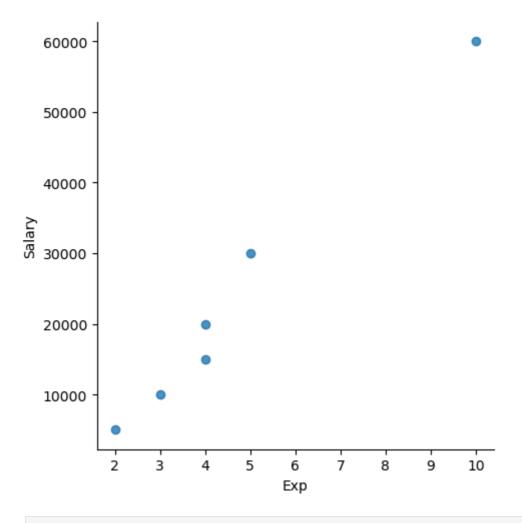
In [193... vis3 = plt.hist(clean_data['Exp'])



In [195... vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary')



In [197... vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)



In [199... clean_data[:]

Out[199...

	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 [201... clean_data[0:6:2]

Out[201...

	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 [203... clean_data[::-1]

```
Out[203...
              Name
                        Domain Age Location Salary Exp
           5
                Kim
                            NLP
                                    55
                                            Delhi
                                                  60000
                                                           10
           4 Uttam
                        Statistics
                                    67
                                        Bangalore
                                                   30000
                                                            5
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
                                                   20000
                                                            4
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
               Teddy
                          Testing
                                   45
                                        Bangalore
                                                   10000
                                                            3
               Mike
                      Datascience
                                    34
                                         Mumbai
                                                    5000
                                                            2
In [205...
          clean_data.columns
Out[205...
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
          X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [207...
          X_iv
In [211...
Out[211...
                         Domain Age
              Name
                                       Location Exp
               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
In [213...
          y_dv = clean_data[['Salary']]
In [215...
           y_dv
Out[215...
              Salary
                5000
           0
               10000
               15000
           2
           3
               20000
               30000
               60000
In [217...
           emp
```

Out[217		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	NaN	NaN	15000	4
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN
	4	Uttam	Statistics	67	NaN	30000	5
	5	Kim	NLP	55	Delhi	60000	10
In [219	cl	ean_dat	a				
Out[219		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 [221	X_	iv					
Out[221		Name	Domain	Age	Location	Ехр	
	0	Mike	Datascience	34	Mumbai	2	
	1	Teddy	Testing	45	Bangalore	3	
	2	Umar	Dataanalyst	50	Bangalore	4	
	3	Jane	Analytics	50	Hyderbad	4	
	4	Uttam	Statistics	67	Bangalore	5	

Kim

In [225... y_dv

NLP

55

Delhi

10

Out[225... Salary 0 5000 10000 1 15000 2 20000 30000 60000 In [227... clean_data Out[227... Name Domain Age **Location Salary Exp** 0 2 Mike Datascience 34 Mumbai 5000 Bangalore Teddy **Testing** 45 10000 3 Dataanalyst Bangalore 2 Umar 50 15000 4 3 Jane Analytics 50 Hyderbad 20000 Uttam Statistics 67 Bangalore 30000 5 5 Kim NLP 55 Delhi 60000 10 In [229... imputation = pd.get_dummies(clean_data) In [231... imputation Out[231... Age Salary Exp Name_Jane Name_Kim Name_Mike Name_Teddy Name_Umar 0 34 5000 2 False False True False False 1 45 10000 3 False False False False True 2 50 15000 4 False False False True False 3 50 20000 False True False False False 4 30000 5 False False 67 False False False 5 55 60000 10 False False True False False

In [233...

clean_data

Out[233		Name	, D	omain	Δαρ	Locati	ion	Salary	Evn			
046[233												
	0	Mike	Datas	cience	34	Mum	bai	5000	2			
	1	Teddy	, 7	Testing	45	Bangal	ore	10000	3			
	2	Umar	Dataa	analyst	50	Bangal	ore	15000	4			
	3	Jane	e An	alytics	50	Hyderk	oad	20000	4			
	4	Uttam	Sta	atistics	67	Bangal	ore	30000	5			
	5	Kim	l	NLP	55	De	elhi	60000	10			
In [235	imputation											
Out[235		Age	Salary	Ехр	Name _.	_Jane	Nam	e_Kim	Name	_Mike	Name_Teddy	Name_Umar
	0	34	5000	2		False		False		True	False	False
	1	45	10000	3		False		False		False	True	False
	2	50	15000	4		False		False		False	False	True
	3	50	20000	4		True		False		False	False	False
	4	67	30000	5		False		False		False	False	False
	5	55	60000	10		False		True		False	False	False

In []: