```
In [234...
           import pandas as pd
In [235...
           pd.__version__
Out[235...
            '2.3.0'
           emp = pd.read_excel(r"C:\Users\hp\Downloads\Rawdata.xlsx")
In [236...
           emp
Out[236...
                              Domain
                Name
                                           Age
                                                 Location
                                                              Salary
                                                                          Exp
           0
                 Mike
                         Datascience#$
                                       34 years
                                                  Mumbai
                                                             5^00#0
                                                                          2+
              Teddy^
                                          45' yr
                                                 Bangalore
                                                                           <3
           1
                               Testing
                                                           10%%000
           2
               Uma#r
                       Dataanalyst^^#
                                          NaN
                                                     NaN
                                                            1$5%000
                                                                       4> yrs
           3
                          Ana^^lytics
                                          NaN
                                                 Hyderbad
                                                             2000^0
                 Jane
                                                                         NaN
               Uttam*
                              Statistics
                                                                      5+ year
           4
                                          67-yr
                                                     NaN
                                                              30000-
           5
                                           55yr
                                                     Delhi
                                                            6000^$0
                  Kim
                                  NLP
                                                                         10+
In [237...
           id(emp)
Out[237...
           3057123670544
In [238...
           emp.columns
Out[238...
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [239...
           emp.shape
Out[239...
           (6, 6)
In [240...
           emp.head()
Out[240...
                Name
                              Domain
                                                 Location
                                           Age
                                                              Salary
                                                                          Exp
           0
                 Mike
                        Datascience#$
                                       34 years
                                                             5^00#0
                                                                           2+
                                                  Mumbai
              Teddy^
                               Testing
                                          45' yr Bangalore
                                                            10%%000
                                                                           <3
           2
               Uma#r
                       Dataanalyst^^#
                                          NaN
                                                     NaN
                                                            1$5%000
                                                                       4> yrs
                           Ana^^lytics
           3
                 Jane
                                          NaN
                                                 Hyderbad
                                                             2000^0
                                                                         NaN
               Uttam*
                                          67-yr
                              Statistics
                                                     NaN
                                                              30000-
                                                                      5+ year
In [241...
           emp.tail()
```

Out[241...

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

In [242...

emp.info()

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

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

dtypes: object(6)

memory usage: 420.0+ bytes

In [243...

emp.isnull()

Out[243...

	Name	Domain	Age	Location	Salary	Ехр
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	True	False	False
3	False	False	True	False	False	True
4	False	False	False	True	False	False
5	False	False	False	False	False	False

In [244...

emp.isna()

Out[244...

	Name	Domain	Age	Location	Salary	Ехр
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	True	True	False	False
3	False	False	True	False	False	True
4	False	False	False	True	False	False
5	False	False	False	False	False	False

```
emp.isnull().sum()
In [245...
Out[245...
           Name
                        0
           Domain
                        0
                        2
           Age
           Location
                        2
           Salary
                        0
           Exp
                        1
           dtype: int64
           DATA CLEANING OR DATA CLEANSING
In [246...
           emp
Out[246...
               Name
                              Domain
                                          Age
                                                 Location
                                                              Salary
                                                                        Exp
           0
                Mike
                        Datascience#$
                                       34 years
                                                            5^00#0
                                                                          2+
                                                  Mumbai
              Teddy^
                                         45' yr Bangalore
                                                           10%%000
                                                                          <3
           1
                               Testing
               Uma#r Dataanalyst^^#
           2
                                                     NaN
                                                           1$5%000
                                          NaN
                                                                      4> yrs
           3
                          Ana^^lytics
                                          NaN
                                                Hyderbad
                                                            2000^0
                 Jane
                                                                        NaN
           4
              Uttam*
                             Statistics
                                         67-yr
                                                             30000-
                                                                     5+ year
                                                     NaN
           5
                                 NLP
                                          55yr
                                                    Delhi
                                                           6000^$0
                                                                        10+
                 Kim
In [247...
           emp['Name']
Out[247...
           0
                   Mike
           1
                 Teddy^
           2
                 Uma#r
           3
                   Jane
           4
                 Uttam*
           5
                    Kim
           Name: Name, dtype: object
In [248...
           emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
           emp['Name']
In [249...
Out[249...
           0
                 Mike
           1
                 Teddy
           2
                  Umar
           3
                  Jane
           4
                 Uttam
           5
                   Kim
           Name: Name, dtype: object
In [250...
           emp
```

```
Out[250...
               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
               Umar
                      Dataanalyst^^#
                                          NaN
                                                     NaN
                                                           1$5%000
                                                                       4> yrs
            3
                Jane
                          Ana^^lytics
                                          NaN
                                                Hyderbad
                                                             2000^0
                                                                        NaN
                                                             30000- 5+ year
            4
               Uttam
                             Statistics
                                         67-yr
                                                     NaN
            5
                 Kim
                                 NLP
                                          55yr
                                                    Delhi
                                                            6000^$0
                                                                         10+
In [251...
           emp['Domain']
Out[251...
            0
                  Datascience#$
            1
                         Testing
            2
                 Dataanalyst^^#
            3
                    Ana^^lytics
                     Statistics
            4
                             NLP
            Name: Domain, dtype: object
In [252...
           emp['Domain']= emp['Domain'].str.replace(r'\W','',regex=True)
In [253...
           emp['Domain']
Out[253...
            0
                 Datascience
            1
                     Testing
            2
                 Dataanalyst
            3
                   Analytics
            4
                  Statistics
                          NLP
            Name: Domain, dtype: object
In [254...
           emp
Out[254...
               Name
                         Domain
                                      Age
                                             Location
                                                          Salary
                                                                     Ехр
            0
                Mike
                      Datascience
                                                         5^00#0
                                                                      2+
                                   34 years
                                              Mumbai
               Teddy
                           Testing
                                     45' yr
                                            Bangalore
                                                       10%%000
                                                                      <3
            2
               Umar
                       Dataanalyst
                                      NaN
                                                 NaN
                                                        1$5%000
                                                                   4> yrs
            3
                Jane
                         Analytics
                                      NaN
                                            Hyderbad
                                                         2000^0
                                                                     NaN
            4
               Uttam
                         Statistics
                                      67-yr
                                                 NaN
                                                          30000-
                                                                  5+ year
            5
                 Kim
                             NLP
                                      55yr
                                                 Delhi
                                                        6000^$0
                                                                     10+
In [255...
           emp['Age']
```

```
Out[255...
                34 years
                   45' yr
           1
           2
                      NaN
           3
                      NaN
           4
                    67-yr
           5
                     55yr
           Name: Age, dtype: object
In [256...
           emp['Age']= emp['Age'].str.extract('(\\d+)')
In [257...
           emp['Age']
Out[257...
           0
                  34
           1
                  45
           2
                 NaN
           3
                 NaN
           4
                  67
                  55
           Name: Age, dtype: object
In [258...
           emp
Out[258...
               Name
                         Domain
                                  Age
                                         Location
                                                      Salary
                                                                 Exp
           0
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                     5^00#0
                                                                  2+
               Teddy
                                        Bangalore
           1
                          Testing
                                    45
                                                   10%%000
                                                                  <3
               Umar
                      Dataanalyst NaN
                                             NaN
                                                    1$5%000
           2
                                                               4> yrs
                                        Hyderbad
                                                     2000^0
                                                                 NaN
           3
                Jane
                         Analytics NaN
                         Statistics
                                                      30000-
                                                             5+ year
           4
              Uttam
                                    67
                                             NaN
                 Kim
                             NLP
                                             Delhi
                                                    6000^$0
                                                                 10+
           5
                                    55
In [259...
           emp['Location']
Out[259...
           0
                    Mumbai
           1
                 Bangalore
           2
                       NaN
           3
                  Hyderbad
           4
                       NaN
           5
                     Delhi
           Name: Location, dtype: object
           emp['Location']= emp['Location'].str.replace(r'\W','',regex=True)
In [260...
In [261...
           emp['Location']
Out[261...
           0
                    Mumbai
           1
                 Bangalore
           2
                       NaN
           3
                  Hyderbad
           4
                       NaN
           5
                     Delhi
           Name: Location, dtype: object
In [262...
           emp
```

```
Out[262...
               Name
                         Domain Age
                                         Location
                                                      Salary
                                                                 Ехр
           0
                Mike Datascience
                                    34
                                          Mumbai
                                                     5^00#0
                                                                  2+
           1
               Teddy
                          Testing
                                    45 Bangalore
                                                   10%%000
                                                                   <3
           2
               Umar
                      Dataanalyst
                                  NaN
                                             NaN
                                                    1$5%000
                                                               4> yrs
           3
                Jane
                         Analytics
                                  NaN
                                         Hyderbad
                                                     2000^0
                                                                 NaN
           4
              Uttam
                         Statistics
                                    67
                                             NaN
                                                      30000- 5+ year
           5
                 Kim
                             NLP
                                    55
                                             Delhi
                                                    6000^$0
                                                                 10+
In [263...
           emp['Salary']
Out[263...
           0
                  5^00#0
           1
                 10%%000
           2
                 1$5%000
           3
                  2000^0
                  30000-
           4
                 6000^$0
           Name: Salary, dtype: object
In [264...
           emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
In [265...
           emp['Salary']
Out[265...
           0
                  5000
                 10000
           1
           2
                 15000
           3
                 20000
           4
                 30000
                 60000
           Name: Salary, dtype: object
In [266...
           emp
Out[266...
               Name
                         Domain
                                   Age
                                         Location
                                                   Salary
                                                               Exp
           0
                Mike
                      Datascience
                                    34
                                                     5000
                                                               2+
                                          Mumbai
                                                    10000
               Teddy
                          Testing
                                    45
                                         Bangalore
                                                                <3
           2
               Umar
                      Dataanalyst
                                   NaN
                                             NaN
                                                    15000
                                                            4> yrs
           3
                Jane
                         Analytics
                                   NaN
                                         Hyderbad
                                                    20000
                                                              NaN
           4
               Uttam
                         Statistics
                                    67
                                             NaN
                                                    30000
                                                           5+ year
           5
                 Kim
                             NLP
                                    55
                                             Delhi
                                                    60000
                                                               10+
In [267...
           emp['Exp']
```

```
Out[267...
           0
                      2+
           1
                      <3
           2
                4> yrs
           3
                     NaN
           4
              5+ year
                     10+
           Name: Exp, dtype: object
           emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
In [268...
In [269...
           emp['Exp']
Out[269...
           0
                   2
           1
                   3
           2
                   4
           3
                NaN
           4
                   5
                  10
           Name: Exp, dtype: object
In [270...
           clean_data = emp.copy()
In [271...
           clean_data
Out[271...
              Name
                         Domain Age
                                        Location Salary
                                                          Exp
                                                             2
           0
               Mike Datascience
                                    34
                                         Mumbai
                                                    5000
              Teddy
                                    45 Bangalore
                                                   10000
                                                             3
           1
                          Testing
           2
                      Dataanalyst
                                                   15000
                                                             4
               Umar
                                  NaN
                                            NaN
           3
                                        Hyderbad
                                                   20000 NaN
                Jane
                        Analytics NaN
                                                   30000
           4
              Uttam
                        Statistics
                                    67
                                            NaN
                                                             5
                            NLP
           5
                Kim
                                    55
                                            Delhi
                                                   60000
                                                            10
In [272...
           clean_data.isnull().sum()
Out[272...
                        0
           Name
           Domain
                        0
                        2
           Age
                        2
           Location
                        0
           Salary
                        1
           Exp
           dtype: int64
In [273...
           clean_data['Age']
           0
Out[273...
                  34
                  45
           1
           2
                NaN
           3
                NaN
           4
                  67
                  55
           Name: Age, dtype: object
In [274...
           import numpy as np
```

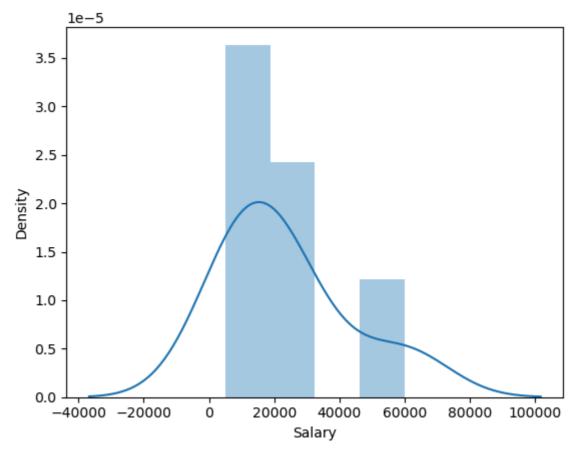
```
In [275...
                                     clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_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.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Ag
                                     clean_data['Age']
In [276...
Out[276...
                                                                   34
                                      1
                                                                   45
                                      2
                                                        50.25
                                                       50.25
                                      3
                                      4
                                                                   67
                                      5
                                                                   55
                                      Name: Age, dtype: object
                                     clean_data['Exp']
In [277...
Out[277...
                                                               2
                                      1
                                                               3
                                      2
                                                               4
                                      3
                                                       NaN
                                      4
                                                               5
                                                            10
                                      Name: Exp, dtype: object
In [278...
                                     clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']))
                                     clean_data['Exp']
In [279...
Out[279...
                                      0
                                                               2
                                      1
                                                               3
                                      2
                                                               4
                                      3
                                                       4.8
                                      4
                                                               5
                                                            10
                                      Name: Exp, dtype: object
In [280...
                                     clean_data
Out[280...
                                                Name
                                                                                   Domain
                                                                                                                     Age
                                                                                                                                          Location Salary Exp
                                      0
                                                   Mike Datascience
                                                                                                                          34
                                                                                                                                             Mumbai
                                                                                                                                                                                5000
                                                                                                                                                                                                            2
                                                                                                                                       Bangalore
                                                 Teddy
                                                                                       Testing
                                                                                                                          45
                                                                                                                                                                             10000
                                                                                                                                                                                                            3
                                      2
                                                                          Dataanalyst 50.25
                                                                                                                                                        NaN
                                                                                                                                                                             15000
                                                                                                                                                                                                            4
                                                  Umar
                                      3
                                                     Jane
                                                                                  Analytics 50.25
                                                                                                                                         Hyderbad
                                                                                                                                                                             20000
                                                                                                                                                                                                       4.8
                                                                                                                                                                                                            5
                                                Uttam
                                                                                  Statistics
                                                                                                                          67
                                                                                                                                                       NaN
                                                                                                                                                                             30000
                                                                                                NLP
                                      5
                                                        Kim
                                                                                                                           55
                                                                                                                                                      Delhi
                                                                                                                                                                            60000
                                                                                                                                                                                                         10
In [281...
                                     clean_data['Location'].isnull().sum()
Out[281...
                                      np.int64(2)
In [282...
                                     clean_data['Location']
```

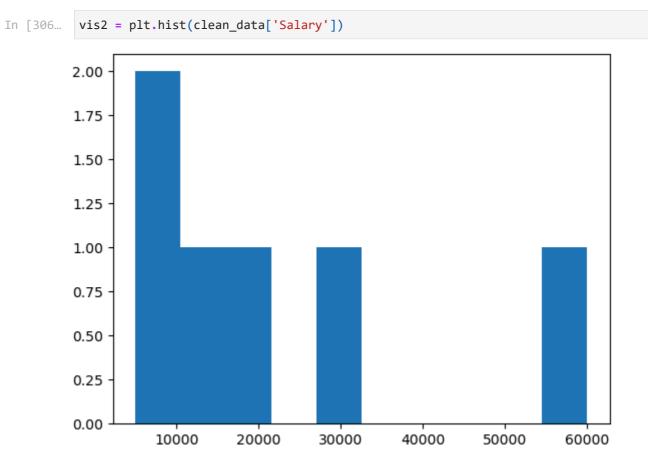
```
Out[282...
                   Mumbai
           1
                Bangalore
           2
                      NaN
           3
                 Hyderbad
           4
                      NaN
           5
                    Delhi
           Name: Location, dtype: object
           clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mod
In [283...
In [284...
           clean_data['Location']
Out[284...
           0
                   Mumbai
           1
                Bangalore
           2
                Bangalore
           3
                 Hyderbad
           4
                Bangalore
                    Delhi
           Name: Location, dtype: object
In [285...
           clean_data
Out[285...
                                        Location Salary Exp
              Name
                        Domain
                                  Age
           0
               Mike Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
              Teddy
                                       Bangalore
                                                  10000
           1
                         Testing
                                   45
                                                            3
           2
              Umar Dataanalyst
                                 50.25
                                                            4
                                       Bangalore
                                                  15000
                        Analytics 50.25
           3
               Jane
                                        Hyderbad
                                                  20000
                                                          4.8
           4
             Uttam
                        Statistics
                                        Bangalore
                                                  30000
                                                            5
                                   67
                                            Delhi
                Kim
                            NLP
                                                  60000
           5
                                    55
                                                           10
In [286...
           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
                         4 non-null
                                          object
          2
              Age
              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 [287...
           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
         0 Name 6 non-null object
1 Domain 6 non-null object
         2 Age 6 non-null
                                      object
         3 Location 6 non-null
                                       object
                       6 non-null
         4
            Salary
                                       object
         5
             Exp
                      6 non-null
                                       object
         dtypes: object(6)
        memory usage: 420.0+ bytes
In [288...
          clean_data['Age']=clean_data['Age'].astype(int)
          clean_data['Age']
In [289...
Out[289...
               34
          1
               45
          2
               50
          3
               50
          4
               67
          5
               55
          Name: Age, dtype: int64
In [290...
          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
            Name
         2 Age 6 non-null
3 Location 6 non-null
                                       int64
                                       object
            Salary 6 non-null
         4
                                       object
                      6 non-null
         5
             Exp
                                       object
         dtypes: int64(1), object(5)
        memory usage: 420.0+ bytes
In [291...
          clean data['Salary'] = clean data['Salary'].astype(int)
In [292...
          clean_data['Salary']
Out[292...
          0
                5000
          1
               10000
          2
               15000
          3
               20000
          4
               30000
               60000
          Name: Salary, dtype: int64
In [293...
          clean_data['Exp'] = clean_data['Exp'].astype(int)
In [294...
          clean data['Exp']
```

```
Out[294...
                2
          1
                3
          2
                4
          3
                4
          4
                5
          5
               10
          Name: Exp, dtype: int64
In [295...
          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
                        6 non-null
                                        object
             Name
          1
              Domain
                        6 non-null
                                        object
          2
             Age
                        6 non-null
                                        int64
          3
              Location 6 non-null
                                        object
                        6 non-null
                                        int64
              Salary
          5
                        6 non-null
                                        int64
              Exp
         dtypes: int64(3), object(3)
         memory usage: 420.0+ bytes
In [296...
          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 [297...
          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
                        6 non-null
          2
             Age
                                        int64
              Location 6 non-null
          3
                                        category
          4
                        6 non-null
                                        int64
              Salary
          5
              Exp
                        6 non-null
                                        int64
         dtypes: category(3), int64(3)
         memory usage: 938.0 bytes
In [298...
          clean data
Out[298...
             Name
                       Domain Age
                                      Location Salary Exp
          0
              Mike
                    Datascience
                                                 5000
                                                         2
                                 34
                                       Mumbai
              Teddy
                        Testing
                                 45
                                     Bangalore
                                                10000
          2
              Umar
                    Dataanalyst
                                     Bangalore
                                                15000
                                                         4
          3
               Jane
                       Analytics
                                 50
                                    Hyderbad
                                                20000
                                                         5
             Uttam
                       Statistics
                                 67
                                     Bangalore
                                                30000
          5
               Kim
                           NLP
                                 55
                                         Delhi
                                                60000
                                                        10
```

```
clean_data.to_csv('clean_data.csv')
In [299...
           import os
In [300...
           os.getcwd()
Out[300...
            'C:\\Users\\hp'
In [301...
           clean_data
Out[301...
              Name
                         Domain
                                  Age
                                        Location Salary Exp
               Mike Datascience
                                                    5000
           0
                                    34
                                         Mumbai
                                                             2
                                                   10000
              Teddy
                          Testing
                                    45
                                        Bangalore
                                                             3
           1
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                             4
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
                                                   20000
           4
              Uttam
                         Statistics
                                        Bangalore
                                                   30000
                                                             5
                                    67
                             NLP
           5
                Kim
                                    55
                                            Delhi
                                                   60000
                                                            10
In [302...
           import matplotlib.pyplot as plt
           import seaborn as sns
In [303...
           import warnings
           warnings.filterwarnings('ignore')
In [304...
           clean_data['Salary']
Out[304...
           0
                  5000
           1
                 10000
           2
                 15000
           3
                 20000
           4
                 30000
           5
                 60000
           Name: Salary, dtype: int64
           vis1 = sns.distplot(clean_data['Salary'])
In [305...
```





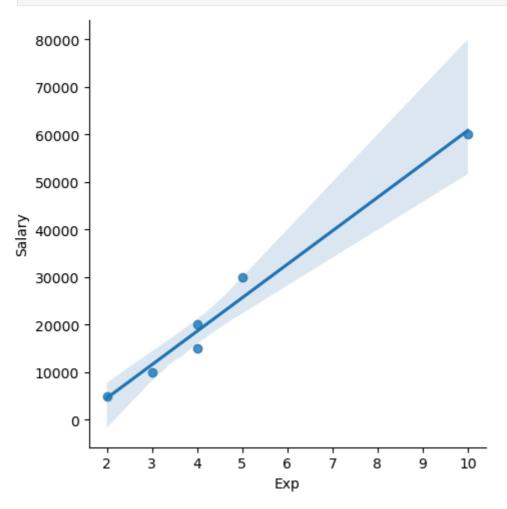
clean_data

In [307...

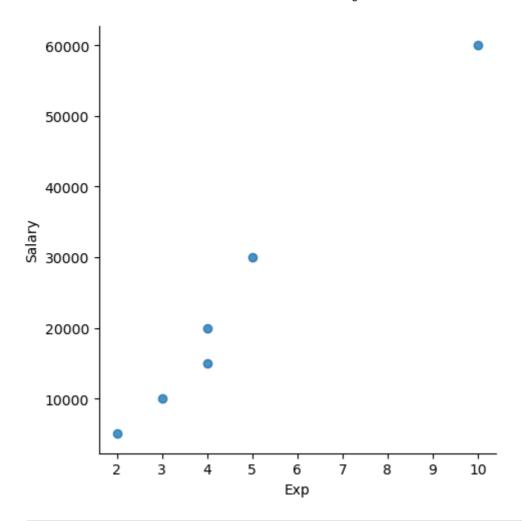
Out[307...

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



In [309... vis5 = sns.lmplot(data=clean_data,x ='Exp', y='Salary',fit_reg = False)



In [310... clean_data[:]

Out[310...

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

Out[311...

	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 [312... clean_data[::-1]

8/8/25, 12:45 AM EDA Cleaning Data

```
Out[312...
```

	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	4
1	Teddy	Testing	45	Bangalore	10000	3
0	Mike	Datascience	34	Mumbai	5000	2

In [313... clean_data.columns

Out[313... Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')

In [314... clean_data

Out[314...

	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 [315...

emp

Out[315...

	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 [316...

clean_data

```
Out[316...
```

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

```
In [320... x_iv = clean_data[['Name','Domain','Age','Salary','Exp']]
    x_iv
```

```
Out[320...
```

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

```
In [321... y_dv = clean_data[['Salary']]
y_dv
```

Out[321...

Salary

- **o** 5000
- **1** 10000
- **2** 15000
- **3** 20000
- **4** 30000
- **5** 60000

In [322...

imputation = pd.get_dummies(clean_data, dtype=int)
imputation

Out[322		Age	Salary	Ехр	Name_	Jane	Nam	e_Kim	Name	_Mike	Name_Tedo	dy N	lame_Umar
	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
	4												•
In [323	clean_data												
Out[323		Name	. De	omain	Age	Loca	tion	Salary	Ехр				
	0	Mike	Datas	cience	34	Mur	nbai	5000	2				
	1	Teddy	, Т	esting	45	Banga	alore	10000	3				
	2	Umar	Dataa	nalyst	50	Banga	alore	15000	4				
	3	Jane	e An	alytics	50	Hyde	rbad	20000	4				
	4	Uttam	Sta	atistics	67	Banga	alore	30000	5				
	5	Kim	l	NLP	55	[Delhi	60000	10				
In [324	ler	n(clea	n_data))									
Out[324	6												
			-										
In [325	·												
Out[325	<pre>Index(['Age', 'Salary', 'Exp', 'Name_Jane', 'Name_Kim', 'Name_Mike',</pre>												
In [326	id	(emp)											
Out[326	30	571236	570544										
In [327	emp	emp.columns											
Out[327	In	<pre>Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')</pre>											
In [328	emp	emp.shape											
Out[328	(6	, 6)											
In []:													

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