```
In [1]:
        import pandas as pd
        emp=pd.read_excel(r"C:\Users\rahee\Downloads\Rawdata.xlsx")
In [2]:
Out[2]:
            Name
                          Domain
                                            Location
                                      Age
                                                        Salary
                                                                   Exp
                     Datascience#$ 34 years
              Mike
                                                        5^00#0
                                                                    2+
                                             Mumbai
           Teddy^
                                     45' yr Bangalore
                           Testing
                                                      10%%000
                                                                    <3
                                                      1$5%000
         2
            Uma#r
                    Dataanalyst^^#
                                      NaN
                                                NaN
                                                                 4> yrs
         3
                       Ana^^lytics
                                     NaN Hyderbad
                                                        2000^0
                                                                   NaN
              Jane
                                                        30000-
                                                                5+ year
         4
            Uttam*
                          Statistics
                                     67-yr
                                                NaN
         5
               Kim
                              NLP
                                      55yr
                                                Delhi
                                                       6000^$0
                                                                   10+
In [3]:
        id(emp) #memory address
Out[3]: 2197987376496
In [4]: emp.columns
Out[4]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [5]:
        emp.shape
Out[5]: (6, 6)
        emp.head()
In [6]:
Out[6]:
             Name
                          Domain
                                      Age
                                            Location
                                                         Salary
                                                                    Ехр
         0
              Mike
                     Datascience#$ 34 years
                                                        5^00#0
                                                                    2+
                                             Mumbai
         1 Teddy^
                                     45' yr Bangalore
                           Testing
                                                      10%%000
                                                                    <3
            Uma#r Dataanalyst^^#
                                    NaN
                                                NaN
                                                       1$5%000
                                                                 4> yrs
                       Ana^^lytics
                                    NaN Hyderbad
                                                        2000^0
              Jane
                                                                   NaN
                                                        30000-
                                                                5+ year
           Uttam*
                          Statistics
                                     67-yr
                                                NaN
```

emp.tail()

In [7]:

```
Out[7]:
            Name
                          Domain
                                   Age
                                         Location
                                                     Salary
                                                                Exp
         1 Teddy^
                           Testing 45' yr Bangalore 10%%000
                                                                 <3
            Uma#r Dataanalyst^^#
                                   NaN
                                                              4> yrs
                                             NaN
                                                   1$5%000
        3
                       Ana^^lytics
                                   NaN Hyderbad
                                                     2000^0
                                                               NaN
              Jane
            Uttam*
                          Statistics
                                   67-yr
                                             NaN
                                                     30000-
                                                             5+ year
         5
                                             Delhi
                                                    6000^$0
                                                                10+
              Kim
                             NLP
                                   55yr
```

In [8]: emp.info()

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

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

dtypes: object(6)

memory usage: 420.0+ bytes

In [9]: emp

Out[9]: Name Domain Age Loc

	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 [10]: emp.isnull()

Out[10]:

	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 [11]:
          emp.isna()
Out[11]:
                     Domain Age Location
                                              Salary
             Name
                                                       Exp
              False
                        False
                              False
                                        False
                                                False False
              False
                        False
                              False
                                        False
                                                False False
          2
              False
                        False
                              True
                                        True
                                                False False
              False
                        False
                              True
                                        False
                                                False
                                                      True
          4
              False
                        False
                              False
                                         True
                                                False False
              False
                        False
                              False
                                        False
                                                False False
          emp.isnull().sum()
In [12]:
Out[12]:
          Name
                       0
          Domain
                       0
                       2
          Age
          Location
                       2
          Salary
          Exp
          dtype: int64
In [13]:
          emp.columns
Out[13]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [14]:
          emp
Out[14]:
              Name
                             Domain
                                                Location
                                         Age
                                                            Salary
                                                                        Ехр
          0
               Mike
                       Datascience#$
                                      34 years
                                                 Mumbai
                                                            5^00#0
                                                                         2+
             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
                                NLP
                 Kim
                                         55yr
                                                   Delhi
                                                           6000^$0
                                                                        10+
```

Data Cleaning or Data Cleansing

In [16]: emp

```
Out[16]:
              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
             Uttam*
                           Statistics
                                       67-yr
                                                 NaN
                                                         30000- 5+ year
          5
                Kim
                               NLP
                                       55yr
                                                 Delhi
                                                        6000^$0
                                                                    10+
In [17]:
         emp['Name']
Out[17]: 0
                 Mike
          1
               Teddy^
          2
                Uma#r
          3
                 Jane
          4
               Uttam*
                  Kim
          Name: Name, dtype: object
In [18]: emp['Name']=emp['Name'].str.replace(r'\W','',regex=True) #nan word change
In [19]:
         emp['Name']
         0
Out[19]:
                Mike
          1
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
In [20]:
         emp['Domain']
Out[20]: 0
                Datascience#$
          1
                      Testing
               Dataanalyst^^#
          2
          3
                  Ana^^lytics
          4
                   Statistics
          5
                          NLP
          Name: Domain, dtype: object
In [21]: emp['Domain']=emp['Domain'].str.replace(r'\W','',regex=True) #nan word change
          emp['Domain']
In [22]:
Out[22]:
               Datascience
          1
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
                       NLP
          5
          Name: Domain, dtype: object
         emp
In [23]:
```

```
Out[23]:
             Name
                       Domain
                                   Age
                                       Location
                                                     Salary
                                                               Ехр
          0
             Mike Datascience 34 years
                                         Mumbai
                                                    5^00#0
                                                                2+
                                 45' yr Bangalore
             Teddy
                        Testing
                                                 10%%000
                                                                <3
             Umar
          2
                   Dataanalyst
                                  NaN
                                             NaN
                                                   1$5%000
                                                             4> yrs
          3
              Jane
                      Analytics
                                  NaN Hyderbad
                                                    2000^0
                                                               NaN
            Uttam
                      Statistics
                                  67-yr
                                             NaN
                                                    30000- 5+ year
               Kim
                          NLP
                                   55yr
                                            Delhi
                                                   6000^$0
                                                               10+
In [24]:
         emp['Age']
Out[24]: 0
               34 years
                45' yr
          2
                   NaN
          3
                    NaN
          4
                  67-yr
                  55yr
          Name: Age, dtype: object
In [25]: emp['Age']=emp['Age'].str.replace(r'\W','',regex=True) #nan word change
In [26]:
         emp['Age']
Out[26]: 0
               34years
          1
                  45yr
          2
                  NaN
          3
                   NaN
          4
                  67yr
                  55yr
          Name: Age, dtype: object
In [27]: emp['Age']=emp['Age'].str.extract(r'(\d+)')
         emp['Age']
In [28]:
Out[28]: 0
                34
          1
                45
          2
               NaN
          3
               NaN
          4
                67
                55
          Name: Age, dtype: object
In [29]:
         emp
```

```
Out[29]:
             Name
                      Domain Age
                                    Location
                                                 Salary
                                                            Exp
          0
             Mike Datascience
                                 34
                                      Mumbai
                                                 5^00#0
                                                             2+
             Teddy
                        Testing
                                 45 Bangalore
                                              10%%000
                                                             <3
          2
             Umar
                   Dataanalyst NaN
                                         NaN
                                               1$5%000
                                                          4> yrs
          3
              Jane
                      Analytics NaN
                                     Hyderbad
                                                 2000^0
                                                           NaN
            Uttam
                      Statistics
                                         NaN
                                                 30000- 5+ year
                                 67
               Kim
                          NLP
                                 55
                                         Delhi
                                                6000^$0
                                                            10+
         emp['Location']
In [59]:
Out[59]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
               Hyderbad
          4
                     NaN
                   Delhi
          Name: Location, dtype: object
In [61]: emp['Location']=emp['Location'].str.replace(r'\W','',regex=True)
In [63]: emp['Location']
Out[63]: 0
                  Mumbai
          1
              Bangalore
          2
                     NaN
          3
              Hyderbad
          4
                     NaN
                   Delhi
          Name: Location, dtype: object
In [65]: emp['Salary']=emp['Salary'].str.replace(r'\W','',regex=True)
In [67]: emp['Salary']
Out[67]: 0
                5000
          1
               10000
          2
               15000
               20000
          3
               30000
          4
          5
               60000
          Name: Salary, dtype: object
In [69]:
         emp['Exp']
                    2+
Out[69]:
          1
                    <3
          2
                4> yrs
          3
                   NaN
          4
               5+ year
          5
                   10+
          Name: Exp, dtype: object
In [77]: emp['Exp']=emp['Exp'].str.extract(r'(\d+)')
```

```
emp['Exp']
In [79]:
                 2
Out[79]: 0
                 3
          2
                 4
          3
               NaN
                 5
                10
          Name: Exp, dtype: object
In [81]:
          emp
Out[81]:
                       Domain Age
             Name
                                       Location Salary
              Mike Datascience
                                  34
                                                           2
                                        Mumbai
                                                  5000
                        Testing
                                  45 Bangalore
                                                 10000
             Teddy
                    Dataanalyst
                                NaN
                                                 15000
                                                           4
          2
              Umar
                                           NaN
                                      Hyderbad
          3
              Jane
                       Analytics
                                NaN
                                                 20000 NaN
             Uttam
                                                           5
                       Statistics
                                  67
                                           NaN
                                                 30000
               Kim
                           NLP
                                  55
                                          Delhi
                                                 60000
                                                          10
In [85]:
          clean_data=emp.copy()
         clean_data
In [87]:
Out[87]:
             Name
                       Domain
                                Age
                                       Location Salary
                                                         Exp
          0
              Mike Datascience
                                  34
                                        Mumbai
                                                  5000
                                                           2
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                 10000
                                                           3
          2
              Umar
                     Dataanalyst
                                NaN
                                           NaN
                                                 15000
                                                           4
          3
              Jane
                       Analytics
                                NaN
                                      Hyderbad
                                                 20000
                                                        NaN
          4
             Uttam
                       Statistics
                                  67
                                           NaN
                                                 30000
                                                           5
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                 60000
                                                          10
```

Missing value Treatment

In [90]: clean_data

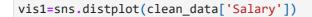
```
Out[90]:
                       Domain Age Location Salary
             Name
                                                      Exp
          0
              Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                         2
             Teddy
                        Testing
                                 45 Bangalore
                                               10000
                                                         3
          2
              Umar
                    Dataanalyst NaN
                                         NaN
                                               15000
                                                         4
          3
              Jane
                      Analytics NaN
                                     Hyderbad
                                               20000 NaN
            Uttam
                       Statistics
                                 67
                                         NaN
                                               30000
                                                         5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                        10
In [92]: 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
                     6 non-null
         1 Domain
                                       object
                  in 6 non-null
4 non-null
         2 Age
                                        object
         3 Location 4 non-null
                                        object
         4
             Salary
                       6 non-null
                                        object
         5
              Exp
                       5 non-null
                                        object
         dtypes: object(6)
         memory usage: 420.0+ bytes
          import numpy as np
In [94]:
In [96]:
          clean_data
Out[96]:
             Name
                       Domain
                                Age
                                      Location
                                               Salary
                                                      Exp
                                                5000
          0
              Mike Datascience
                                 34
                                      Mumbai
                                                         2
             Teddy
                        Testing
                                 45 Bangalore
                                               10000
                                                         3
          1
          2
                    Dataanalyst
                               NaN
                                               15000
                                                         4
              Umar
                                         NaN
              Jane
                                     Hyderbad
                                               20000
                                                      NaN
          3
                      Analytics
                               NaN
            Uttam
                                               30000
                                                         5
                       Statistics
                                 67
                                         NaN
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                        10
In [98]:
          clean_data.head(1)
Out[98]:
             Name
                       Domain
                               Age Location
                                              Salary Exp
                                                       2
              Mike Datascience
                                 34
                                     Mumbai
                                               5000
In [100...
          clean data['Age']
```

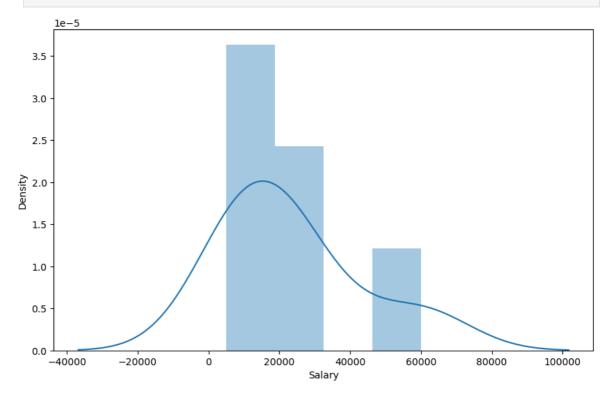
```
Out[100...
                  34
                  45
           1
           2
                 NaN
           3
                 NaN
           4
                  67
           5
                  55
           Name: Age, dtype: object
In [102...
           clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])
In [104...
           clean_data['Age']
Out[104...
           0
                    34
           1
                    45
           2
                 50.25
           3
                 50.25
           4
                    67
                    55
           Name: Age, dtype: object
In [106...
           emp
Out[106...
               Name
                         Domain
                                   Age
                                         Location Salary
                                                            Exp
           0
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                     5000
                                                              2
                                         Bangalore
                                                    10000
               Teddy
                          Testing
                                    45
                                                              3
           1
               Umar
                      Dataanalyst
                                             NaN
                                                    15000
                                                              4
           2
                                  NaN
                                         Hyderbad
                                                    20000
           3
                Jane
                         Analytics
                                  NaN
                                                           NaN
                                             NaN
                                                    30000
           4
              Uttam
                         Statistics
                                    67
                                                              5
                             NLP
                                             Delhi
                                                    60000
           5
                 Kim
                                    55
                                                             10
In [108...
           clean_data
Out[108...
               Name
                         Domain
                                    Age
                                          Location
                                                    Salary
                                                             Exp
           0
                Mike
                                     34
                                                      5000
                                                               2
                      Datascience
                                           Mumbai
                                         Bangalore
               Teddy
                          Testing
                                     45
                                                     10000
                                                               3
                                                     15000
           2
               Umar
                      Dataanalyst
                                   50.25
                                              NaN
                                                               4
           3
                Jane
                         Analytics
                                  50.25
                                          Hyderbad
                                                     20000 NaN
                                                               5
           4
              Uttam
                         Statistics
                                     67
                                              NaN
                                                     30000
                             NLP
                                     55
                                                     60000
           5
                 Kim
                                              Delhi
                                                              10
           clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp
In [110...
In [112...
           clean_data['Exp']
```

```
Out[112...
                   2
            1
                   3
            2
                   4
            3
                 4.8
                   5
            4
            5
                  10
            Name: Exp, dtype: object
In [114...
           clean_data
Out[114...
               Name
                          Domain
                                           Location Salary Exp
                                    Age
            0
                Mike
                      Datascience
                                      34
                                            Mumbai
                                                       5000
                                                               2
            1
               Teddy
                           Testing
                                      45
                                          Bangalore
                                                     10000
                                                               3
            2
               Umar
                       Dataanalyst
                                   50.25
                                               NaN
                                                     15000
                                                               4
            3
                Jane
                         Analytics
                                   50.25
                                          Hyderbad
                                                     20000
                                                              4.8
            4
               Uttam
                         Statistics
                                      67
                                               NaN
                                                     30000
                                                               5
            5
                 Kim
                             NLP
                                      55
                                              Delhi
                                                     60000
                                                              10
In [120...
           clean_data['Location']=clean_data['Location'].fillna(clean_data['Location'].mode
In [122...
           clean_data['Location']
            0
Out[122...
                    Mumbai
            1
                 Bangalore
            2
                 Bangalore
            3
                  Hyderbad
            4
                 Bangalore
            5
                      Delhi
            Name: Location, dtype: object
In [124...
           clean_data
Out[124...
               Name
                          Domain
                                    Age
                                           Location
                                                     Salary Exp
            0
                Mike
                                                       5000
                                                               2
                      Datascience
                                      34
                                            Mumbai
               Teddy
                           Testing
                                      45
                                          Bangalore
                                                     10000
                                   50.25
            2
               Umar
                       Dataanalyst
                                          Bangalore
                                                     15000
                                                               4
            3
                Jane
                         Analytics
                                   50.25
                                          Hyderbad
                                                     20000
                                                              4.8
                                                               5
            4
               Uttam
                         Statistics
                                      67
                                          Bangalore
                                                     30000
                              NLP
            5
                 Kim
                                      55
                                              Delhi
                                                     60000
                                                              10
           clean_data['Age']=clean_data['Age'].astype(int)
In [126...
In [128...
           clean_data['Salary']=clean_data['Salary'].astype(int)
           clean_data['Exp']=clean_data['Exp'].astype(int)
In [130...
           clean_data
```

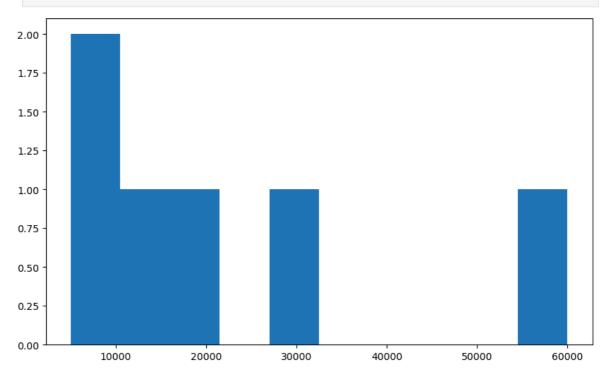
```
Out[130...
            Name
                      Domain Age Location Salary Exp
          0
             Mike Datascience
                               34
                                    Mumbai
                                             5000
                                                     2
             Teddy
                               45 Bangalore
                                             10000
                                                     3
                       Testing
          2
             Umar
                  Dataanalyst
                               50
                                   Bangalore
                                            15000
                                                     4
          3
                     Analytics
                                  Hyderbad
                                            20000
              Jane
                               50
            Uttam
                                   Bangalore
                                            30000
                                                     5
                     Statistics
                               67
              Kim
                         NLP
                               55
                                            60000
                                       Delhi
                                                    10
In [132...
         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
                                   object
         1 Domain 6 non-null
         2 Age 6 non-null
                                     int32
         3 Location 6 non-null
                                     object
         4
            Salary
                      6 non-null
                                     int32
         5
             Exp
                     6 non-null
                                     int32
        dtypes: int32(3), object(3)
        memory usage: 348.0+ bytes
In [134...
         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 [136...
         clean_data.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
            Column Non-Null Count Dtype
                      -----
        ---
                      6 non-null
         0 Name
                                     category
            Domain 6 non-null
         1
                                     category
                     6 non-null
                                     int32
         2 Age
         3
             Location 6 non-null
                                     category
         4
                      6 non-null
             Salary
                                     int32
         5
             Exp
                      6 non-null
                                     int32
        dtypes: category(3), int32(3)
        memory usage: 866.0 bytes
In [138...
         clean_data
```

```
Out[138...
               Name
                         Domain Age
                                        Location Salary Exp
           0
               Mike Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
               Teddy
                          Testing
                                    45
                                        Bangalore
                                                   10000
                                                             3
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
                                                             4
           3
                         Analytics
                                    50
                                        Hyderbad
                                                   20000
                Jane
                                                             4
              Uttam
                         Statistics
                                        Bangalore
                                                   30000
                                                             5
                                    67
           5
                 Kim
                             NLP
                                    55
                                             Delhi
                                                   60000
                                                            10
In [140...
           clean_data.to_csv('clean_data.csv')
In [144...
           import os
           os.getcwd()
            'C:\\Users\\rahee\\FSDS AI GEN AI'
Out[144...
           clean_data.columns
In [146...
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[146...
           import matplotlib.pyplot as plt #visualization
In [148...
           import seaborn as sns #Advanced Visualization
In [150...
           import warnings
           warnings.filterwarnings('ignore')
           clean_data
In [152...
Out[152...
               Name
                         Domain
                                  Age
                                         Location Salary Exp
           0
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
                                                   10000
                                                             3
           1
               Teddy
                          Testing
                                    45
                                        Bangalore
                                                   15000
                                                             4
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
           3
                                                   20000
                Jane
                         Analytics
                                    50
                                        Hyderbad
           4
              Uttam
                         Statistics
                                                   30000
                                                             5
                                    67
                                        Bangalore
           5
                             NLP
                                    55
                                                   60000
                                                            10
                 Kim
                                             Delhi
In [154...
           clean_data['Salary']
           0
Out[154...
                  5000
           1
                 10000
           2
                 15000
           3
                 20000
           4
                 30000
                 60000
           Name: Salary, dtype: int32
In [170...
           plt.rcParams['figure.figsize']=10,6
```

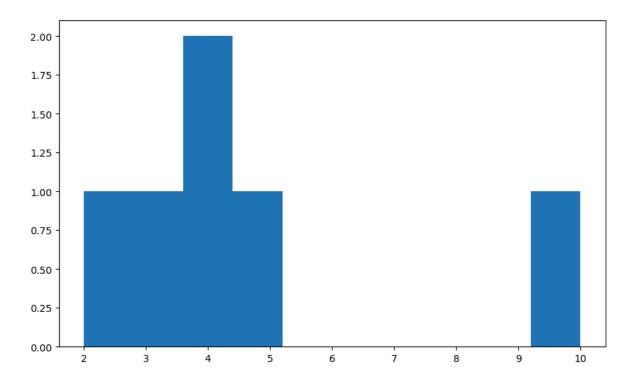




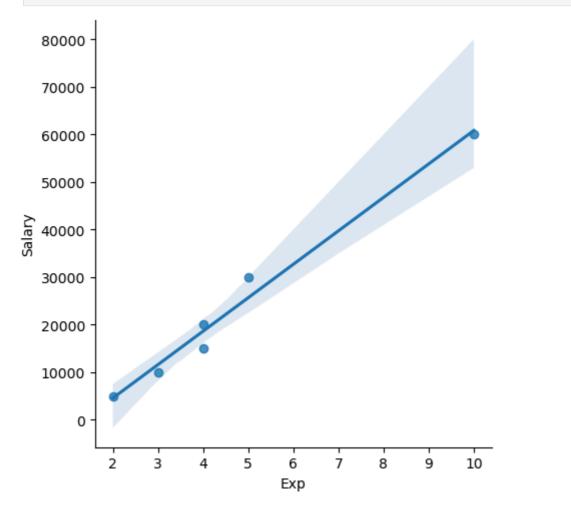
In [172... vis2=plt.hist(clean_data['Salary'])



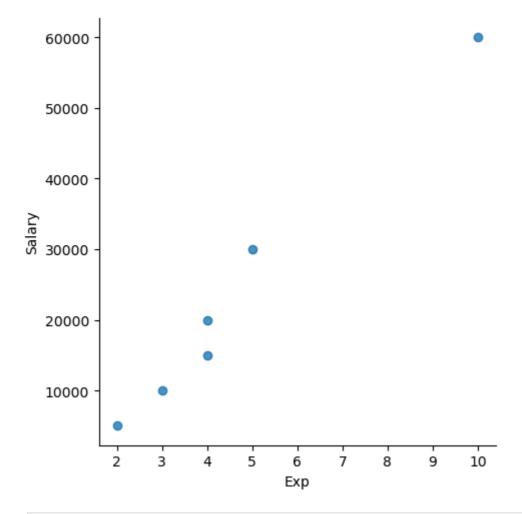
In [174... vis3=plt.hist(clean_data['Exp'])



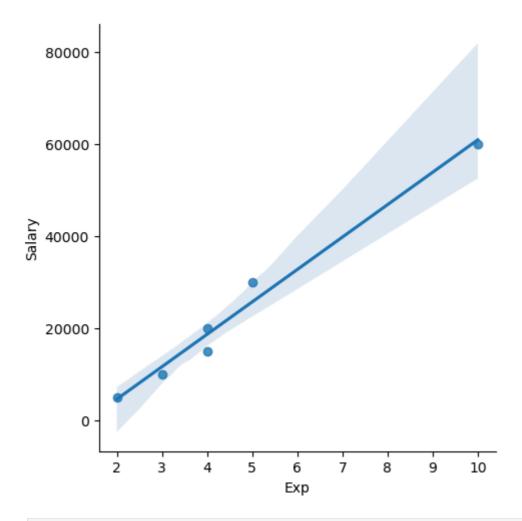
In [176... vis4=sns.lmplot(data=clean_data,x='Exp',y='Salary')



In [178... vis5=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=False)



In [180... vis6=sns.lmplot(data=clean_data,x='Exp',y='Salary',fit_reg=True)



In [182...

clean_data

Out[182...

	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 [184...

clean_data[:]

_		Name	Domain	Age	Location	Salary	Ехр		
	0	Mike	Datascience	34	Mumbai	5000	2		
	1	Teddy	Testing	45	Bangalore	10000	3		
	2	Umar	Dataanalyst	50	Bangalore	15000	4		
	3	Jane	Analytics	50	Hyderbad	20000	4		
	4	Uttam	Statistics	67	Bangalore	30000	5		
	5	Kim	NLP	55	Delhi	60000	10		
In [186	cl	ean_dat	a[:2]						
Out[186	Name		Domain	Age	Location	Salary	Ехр		
	0	Mike	Datascience	34	Mumbai	5000	2		
	1	Teddy	Testing	45	Bangalore	10000	3		
In [188	<pre>clean_data[2:]</pre>								
			a[2:]						
Out[188		Name	Domain	Age	Location	Salary	Ехр		
Out[188	2			Age 50	Location Bangalore	Salary 15000	Exp 4		
Out[188	2	Name	Domain						
Out[188		Name Umar	Domain Dataanalyst	50	Bangalore	15000	4		
Out[188	3	Name Umar Jane	Domain Dataanalyst Analytics	50 50	Bangalore Hyderbad	15000	4		
Out[188	3	Name Umar Jane Uttam	Domain Dataanalyst Analytics Statistics	50 50 67	Bangalore Hyderbad Bangalore	15000 20000 30000	4 4 5		

```
KeyError
                                         Traceback (most recent call last)
File ~\Anaconda\Lib\site-packages\pandas\core\indexes\base.py:3805, in Index.get_
loc(self, key)
  3804 try:
-> 3805
            return self._engine.get_loc(casted_key)
   3806 except KeyError as err:
File index.pyx:167, in pandas._libs.index.IndexEngine.get_loc()
File index.pyx:196, in pandas._libs.index.IndexEngine.get_loc()
File pandas\\_libs\\hashtable_class_helper.pxi:7081, in pandas._libs.hashtable.Py
ObjectHashTable.get_item()
File pandas\\_libs\\hashtable_class_helper.pxi:7089, in pandas._libs.hashtable.Py
ObjectHashTable.get_item()
KeyError: (0, 3)
The above exception was the direct cause of the following exception:
KeyError
                                         Traceback (most recent call last)
Cell In[190], line 1
----> 1 clean_data[0,3]
File ~\Anaconda\Lib\site-packages\pandas\core\frame.py:4102, in DataFrame.__getit
em__(self, key)
  4100 if self.columns.nlevels > 1:
  4101
            return self._getitem_multilevel(key)
-> 4102 indexer = self.columns.get_loc(key)
  4103 if is_integer(indexer):
   4104
            indexer = [indexer]
File ~\Anaconda\Lib\site-packages\pandas\core\indexes\base.py:3812, in Index.get_
loc(self, key)
  3807
          if isinstance(casted_key, slice) or (
  3808
               isinstance(casted_key, abc.Iterable)
               and any(isinstance(x, slice) for x in casted_key)
  3809
   3810
               raise InvalidIndexError(key)
  3811
-> 3812
          raise KeyError(key) from err
  3813 except TypeError:
          # If we have a listlike key, _check_indexing_error will raise
  3814
          # InvalidIndexError. Otherwise we fall through and re-raise
  3815
  3816
          # the TypeError.
   3817
          self._check_indexing_error(key)
KeyError: (0, 3)
```

```
Out[192...
               Name
                          Domain Age
                                         Location
                                                    Salary Exp
            0
                Mike
                      Datascience
                                     34
                                           Mumbai
                                                      5000
                                                               2
                                                     10000
               Teddy
                                     45
                                         Bangalore
                                                               3
            1
                           Testing
            2
                Umar
                       Dataanalyst
                                     50
                                         Bangalore
                                                     15000
                                                               4
                Jane
                         Analytics
                                     50
                                          Hyderbad
                                                     20000
            3
                                                               4
               Uttam
                          Statistics
                                         Bangalore
                                                     30000
                                                               5
                                     67
            5
                 Kim
                              NLP
                                     55
                                              Delhi
                                                     60000
                                                              10
           x_iv=clean_data.drop(['Salary'],axis=1)
In [194...
In [196...
           clean_data
Out[196...
               Name
                          Domain Age
                                          Location
                                                    Salary Exp
                                                      5000
                                                               2
            0
                Mike
                      Datascience
                                     34
                                           Mumbai
                                         Bangalore
                                                     10000
               Teddy
                                     45
                                                               3
            1
                           Testing
            2
                Umar
                       Dataanalyst
                                     50
                                         Bangalore
                                                     15000
                                                               4
            3
                         Analytics
                                     50
                                          Hyderbad
                                                     20000
                Jane
                                                               4
               Uttam
                          Statistics
                                         Bangalore
                                                     30000
                                                               5
            4
                                     67
            5
                              NLP
                                     55
                                              Delhi
                                                     60000
                                                              10
                 Kim
In [198...
           x_iv
Out[198...
               Name
                          Domain
                                   Age
                                          Location
                                                    Exp
                Mike
                                                       2
            0
                       Datascience
                                     34
                                           Mumbai
               Teddy
                                         Bangalore
            1
                                     45
                                                       3
                           Testing
            2
                Umar
                       Dataanalyst
                                     50
                                         Bangalore
                                                       4
            3
                Jane
                                     50
                                          Hyderbad
                         Analytics
                                                       4
            4
               Uttam
                          Statistics
                                     67
                                         Bangalore
                                                       5
                              NLP
            5
                 Kim
                                     55
                                              Delhi
                                                      10
In [200...
           x_iv.columns
            Index(['Name', 'Domain', 'Age', 'Location', 'Exp'], dtype='object')
Out[200...
In [202...
           clean_data.columns
            Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[202...
In [204...
           clean_data
```

```
Out[204...
               Name
                         Domain Age Location Salary Exp
                                                              2
           0
                Mike
                      Datascience
                                     34
                                           Mumbai
                                                     5000
                                         Bangalore
               Teddy
                           Testing
                                     45
                                                     10000
                                                              3
                       Dataanalyst
            2
               Umar
                                     50
                                         Bangalore
                                                    15000
                                                              4
            3
                Jane
                         Analytics
                                     50
                                         Hyderbad
                                                    20000
                                                              4
               Uttam
                         Statistics
                                     67
                                         Bangalore
                                                    30000
                                                              5
            5
                 Kim
                             NLP
                                     55
                                             Delhi
                                                     60000
                                                             10
In [206...
           y_dv=clean_data.drop(['Name', 'Domain', 'Age', 'Location', 'Exp'],axis=1)
In [208...
           y_dv
Out[208...
               Salary
                5000
            0
               10000
            1
               15000
            2
               20000
            3
               30000
            4
               60000
In [210...
           clean_data
Out[210...
               Name
                          Domain Age
                                          Location Salary Exp
            0
                Mike
                      Datascience
                                     34
                                           Mumbai
                                                     5000
                                                              2
               Teddy
                                     45
                                         Bangalore
                                                     10000
                                                              3
            1
                           Testing
            2
               Umar
                       Dataanalyst
                                     50
                                         Bangalore
                                                    15000
                                                              4
            3
                Jane
                                     50
                                         Hyderbad
                                                    20000
                         Analytics
                                                              4
            4
               Uttam
                                         Bangalore
                                                     30000
                                                              5
                         Statistics
                                     67
            5
                 Kim
                             NLP
                                     55
                                             Delhi
                                                     60000
                                                             10
```

In [212...

x_iv

Out[212		Name	Domain	Age	Location	Ехр	
	0	Mike	Datascience	34	Mumbai	2	
	1	Teddy	Testing	45	Bangalore	3	
	2	Umar	Dataanalyst	50	Bangalore	4	
	3	Jane	Analytics	50	Hyderbad	4	
	4	Uttam	Statistics	67	Bangalore	5	
	5	Kim	NLP	55	Delhi	10	
In [214	у_	dv					
Out[214		Salary	-				
	0	5000					
	1	10000					
	2	15000					
	3	20000					
	4	30000					
	5	60000					
	_						
In [216	cl	ean_dat	a				
Out[216		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

67 Bangalore 30000

Delhi 60000

5

10

4 Uttam

imputation

5

In [222...

In [224...

Kim

Statistics

NLP

55

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

Out[224		Age	Salary	Ехр	Name_Jan	e Nan	ne_Kim	Name	_Mike	Name_Teddy	Name_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 [226	cl	ean_da	ata								
Out[226		Nam	e Do	omain	Age Lo	ocation	Salary	Ехр			
	0	Mik	e Datas	cience	34 N	⁄lumbai	5000	2			
	1	Tedd	у 1	Testing	45 Ba	ngalore	10000	3			
	2	Umar Dataanalyst 50		50 Bai	ngalore	15000	4				
3		3 Jane Analytics		50 Hy	derbad	20000	4				
	4 Utta		n Sta	atistics	67 Bai	ngalore	30000	5			
	5	Kin	n	NLP	55	Delhi	60000	10			
In [228	im	putat:	ion								
Out[228			Salary	Ехр	Namo lan	o Nan	ao Kim	Namo	Miko	Name_Teddy	Name_Umar
040[220	0	34	5000	2	ivallie_Jali	0	0	INAIIIE	_iviike 	0	0
	1	45	10000	3		0	0		0	1	0
	2	50	15000	4		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			_							b

raw data with lot of regex, missing, uncleandata regex, clean

fill missing numerical & cateigroica clean_dataset (data cleaning) 3 month - 5mont

outlier treatement, univati, bivariate, corelation split the data into $x_i.v & y_dv$ impute cateogrica data to numerical eda part complete

Next step

- we splitn x_iv -- x_train, x_test
- we split y_dv -- y_train, y_test
- build the ml model with x_train & y_train

In []: