# Raw Data to Clean Data conversion using python EDA

```
import pandas as pd
         pd.__version__
In [2]:
Out[2]:
         '2.2.3'
In [3]: # pip install --upgrade openpyxl
        emp=pd.read_excel(r"C:\Users\shaik\Downloads\Rawdata.xlsx")
In [5]:
         emp
Out[5]:
             Name
                           Domain
                                        Age
                                              Location
                                                           Salary
                                                                      Exp
         0
              Mike
                      Datascience#$
                                    34 years
                                               Mumbai
                                                          5^00#0
                                                                       2+
            Teddy^
                            Testing
                                       45' yr Bangalore
                                                        10%%000
                                                                       <3
                    Dataanalyst^^#
                                                  NaN
             Uma#r
                                       NaN
                                                         1$5%000
                                                                    4> yrs
         3
               Jane
                        Ana^^lytics
                                       NaN
                                            Hyderbad
                                                          2000^0
                                                                     NaN
                                                                  5+ year
            Uttam*
                           Statistics
                                       67-yr
                                                  NaN
                                                          30000-
                                                         6000^$0
               Kim
                               NLP
                                        55yr
                                                  Delhi
                                                                      10+
In [6]:
         id(emp)
Out[6]: 2168099765760
In [7]:
         emp.head()
Out[7]:
             Name
                           Domain
                                        Age
                                              Location
                                                           Salary
                                                                      Exp
              Mike
                      Datascience#$
                                    34 years
                                               Mumbai
                                                          5^00#0
                                                                       2+
            Teddy^
                            Testing
                                       45' yr
                                             Bangalore
                                                        10%%000
                                                                       <3
                    Dataanalyst^^#
             Uma#r
                                       NaN
                                                  NaN
                                                         1$5%000
                                                                    4> yrs
                        Ana^^lytics
                                                          2000^0
         3
               Jane
                                       NaN
                                              Hyderbad
                                                                     NaN
                                                                  5+ year
            Uttam*
                           Statistics
                                       67-yr
                                                          30000-
                                                  NaN
In [8]: emp.tail()
```

Out[8]:	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 [9]: 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 in 6 non-null 4 non-null 1 Domain object 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

### In [10]: emp

#### Out[10]:

	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 [11]: emp.isnull()

### Out[11]:

	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 [12]:
           emp.isna()
Out[12]:
              Name
                      Domain
                                       Location
                                 Age
                                                 Salary
                                                           Exp
           0
                False
                          False
                                False
                                           False
                                                   False
                                                          False
                False
                          False
                                False
                                           False
                                                   False
                                                          False
           2
               False
                          False
                                 True
                                           True
                                                   False
                                                          False
           3
               False
                          False
                                True
                                           False
                                                   False
                                                          True
               False
                          False
                                False
                                                   False False
           4
                                           True
           5
                False
                          False
                                False
                                           False
                                                   False False
           emp.isnull().sum()
In [13]:
Out[13]:
           Name
                         0
           Domain
                         0
                         2
           Age
           Location
                         2
           Salary
           Exp
                         1
           dtype: int64
```

# Data cleaning or cleansing

```
emp
In [14]:
Out[14]:
              Name
                             Domain
                                                Location
                                                             Salary
                                         Age
                                                                        Exp
          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
              Uttam*
                            Statistics
                                         67-yr
                                                    NaN
                                                            30000-
                                                                    5+ year
          5
                 Kim
                                NLP
                                         55yr
                                                   Delhi
                                                           6000^$0
                                                                        10+
In [15]:
          emp['Name']
Out[15]:
                  Mike
          1
                Teddy^
          2
                 Uma#r
          3
                  Jane
                Uttam*
                   Kim
          Name: Name, dtype: object
          emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True) # non word character.
In [16]:
          emp['Name']
In [17]:
```

```
Out[17]: 0
                Mike
          1
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
In [18]:
          emp
Out[18]:
             Name
                           Domain
                                       Age
                                              Location
                                                          Salary
                                                                     Exp
                      Datascience#$ 34 years
              Mike
          0
                                              Mumbai
                                                         5^00#0
                                                                      2+
          1
             Teddy
                            Testing
                                      45' yr
                                             Bangalore
                                                       10%%000
                                                                      <3
                    Dataanalyst^^#
          2
              Umar
                                       NaN
                                                  NaN
                                                        1$5%000
                                                                   4> yrs
          3
              Jane
                        Ana^^lytics
                                       NaN
                                             Hyderbad
                                                         2000^0
                                                                    NaN
                                                                 5+ year
          4
             Uttam
                           Statistics
                                      67-yr
                                                  NaN
                                                          30000-
                              NLP
          5
               Kim
                                       55yr
                                                 Delhi
                                                        6000^$0
                                                                     10+
In [19]:
          emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
In [20]:
          emp['Domain']
Out[20]:
          0
               Datascience
          1
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
                        NLP
          Name: Domain, dtype: object
          emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
In [21]:
In [22]:
          emp['Age']
Out[22]:
          0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
                  55yr
          Name: Age, dtype: object
          emp['Age'] = emp['Age'].str.extract('(\\d+)') # r(r'(\\d+)')
In [23]:
In [24]:
          emp['Age']
Out[24]:
          0
                34
          1
                45
          2
               NaN
          3
               NaN
          4
                67
          5
                55
          Name: Age, dtype: object
```

```
In [25]:
          emp
Out[25]:
              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
          4
              Uttam
                        Statistics
                                   67
                                            NaN
                                                     30000-
                                                             5+ year
          5
                            NLP
                Kim
                                   55
                                            Delhi
                                                   6000^$0
                                                                10+
          emp['Location'] = emp['Location'].str.replace(r'\W','',regex=True)
In [26]:
In [27]:
          emp['Location']
                   Mumbai
Out[27]:
          1
                Bangalore
          2
                      NaN
          3
                 Hyderbad
          4
                      NaN
                    Delhi
          5
          Name: Location, dtype: object
In [28]:
          emp
Out[28]:
              Name
                        Domain
                                  Age
                                        Location
                                                     Salary
                                                                Exp
          0
               Mike
                     Datascience
                                   34
                                         Mumbai
                                                    5^00#0
                                                                 2+
              Teddy
                                                                 <3
          1
                         Testing
                                   45
                                        Bangalore
                                                  10%%000
          2
              Umar
                     Dataanalyst
                                 NaN
                                            NaN
                                                   1$5%000
                                                              4> yrs
                                        Hyderbad
                                                    2000^0
          3
               Jane
                        Analytics
                                 NaN
                                                                NaN
          4
              Uttam
                        Statistics
                                   67
                                            NaN
                                                     30000-
                                                             5+ year
                            NLP
          5
                Kim
                                   55
                                            Delhi
                                                   6000^$0
                                                                10+
          emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
In [29]:
In [30]:
          emp['Salary']
Out[30]:
          0
                 5000
          1
                10000
          2
                15000
          3
                20000
          4
                30000
          5
                60000
          Name: Salary, dtype: object
          emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
In [31]:
In [32]:
          emp['Exp']
```

```
Out[32]: 0
          3
                NaN
                  5
                 10
          Name: Exp, dtype: object
In [33]:
Out[33]:
              Name
                        Domain
                                  Age
                                         Location Salary
                                                           Exp
               Mike
                     Datascience
                                    34
                                                    5000
                                                              2
                                         Mumbai
              Teddy
                         Testing
                                        Bangalore
                                                   10000
                                                              3
                                   45
          2
                     Dataanalyst
                                             NaN
                                                   15000
                                                              4
              Umar
                                  NaN
          3
                        Analytics
                                  NaN
                                        Hyderbad
                                                   20000 NaN
               Jane
             Uttam
                        Statistics
                                    67
                                             NaN
                                                   30000
                                                              5
                Kim
                            NLP
                                    55
                                            Delhi
                                                   60000
                                                            10
In [34]:
          clean_data = emp.copy()
In [35]:
          clean_data
Out[35]:
              Name
                        Domain
                                  Age
                                        Location Salary
                                                           Exp
                                                              2
               Mike Datascience
                                   34
                                         Mumbai
                                                    5000
              Teddy
                          Testing
                                    45
                                        Bangalore
                                                   10000
                                                              3
                                  NaN
                     Dataanalyst
                                                   15000
                                                              4
              Umar
                                             NaN
          3
               Jane
                        Analytics
                                  NaN
                                        Hyderbad
                                                   20000
             Uttam
                        Statistics
                                    67
                                             NaN
                                                   30000
                                                              5
                            NLP
                Kim
                                    55
                                            Delhi
                                                   60000
                                                             10
```

### Till now we have raw data we use regex to clean the data and removed all noice charected from the dataset

you can also work in same thing in sql query as well

## **EDA Technique Lets Apply**

# Missing Value Treatment For Numerical Data

```
In [36]:
                                  clean_data
Out[36]:
                                              Name
                                                                                 Domain
                                                                                                               Age
                                                                                                                                      Location Salary
                                                                                                                                                                                                   Exp
                                                 Mike
                                                                                                                     34
                                                                                                                                                                            5000
                                                                                                                                                                                                           2
                                                                      Datascience
                                                                                                                                         Mumbai
                                               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
                                                                                             NLP
                                   5
                                                     Kim
                                                                                                                     55
                                                                                                                                                  Delhi
                                                                                                                                                                        60000
                                                                                                                                                                                                       10
                                  clean_data.isnull().sum()
In [37]:
Out[37]:
                                   Name
                                                                              0
                                   Domain
                                                                              0
                                   Age
                                                                              2
                                   Location
                                                                              2
                                   Salary
                                                                              0
                                                                               1
                                   Exp
                                   dtype: int64
In [38]:
                                 clean_data['Age']
Out[38]:
                                                         34
                                                         45
                                   1
                                   2
                                                     NaN
                                   3
                                                     NaN
                                                         67
                                                         55
                                   Name: Age, dtype: object
In [39]:
                                  import numpy as np
In [40]: 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['Age'].fillna(np.to_numeric(clean_dat
                                  clean_data['Age']
In [41]:
Out[41]:
                                                                34
                                   1
                                                                45
                                   2
                                                     50.25
                                                     50.25
                                   3
                                   4
                                                                67
                                                                55
                                   Name: Age, dtype: object
In [42]: clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mc
In [43]: clean_data['Location']
```

```
Out[43]: 0
                                                      Mumbai
                              1
                                             Bangalore
                              2
                                             Bangalore
                              3
                                               Hyderbad
                              4
                                             Bangalore
                              5
                                                         Delhi
                              Name: Location, dtype: object
                            clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fil
                             clean_data['Exp']
In [45]:
Out[45]:
                             0
                                                   2
                              1
                                                   3
                              2
                                                   4
                              3
                                             4.8
                              4
                                                   5
                                                10
                              Name: Exp, dtype: object
In [46]:
                             clean_data
Out[46]:
                                                                                                                    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
                                                                                                                 Bangalore
                                                                                                                                                                           4
                                                                                                                                                15000
                                                                   Analytics 50.25
                                                                                                                 Hyderbad
                              3
                                          Jane
                                                                                                                                                20000
                                                                                                                                                                       4.8
                                     Uttam
                                                                   Statistics
                                                                                                                 Bangalore
                                                                                                                                                30000
                                                                                                                                                                           5
                              4
                                                                                                     67
                                             Kim
                                                                              NLP
                                                                                                                             Delhi
                                                                                                                                                60000
                              5
                                                                                                     55
                                                                                                                                                                        10
In [47]:
                            clean_data.info()
                         <class 'pandas.core.frame.DataFrame'>
                         RangeIndex: 6 entries, 0 to 5
                         Data columns (total 6 columns):
                                        Column
                           #
                                                                      Non-Null Count Dtype
                          _ _ _
                                       -----
                                                                      _____
                            0
                                        Name
                                                                      6 non-null
                                                                                                                       object
                            1
                                       Domain
                                                                      6 non-null
                                                                                                                       object
                                                                      6 non-null
                                                                                                                       object
                            2
                                        Age
                                        Location 6 non-null
                                                                                                                       object
                            4
                                        Salary
                                                                      6 non-null
                                                                                                                       object
                            5
                                        Exp
                                                                       6 non-null
                                                                                                                       object
                         dtypes: object(6)
                         memory usage: 420.0+ bytes
                             clean_data['Age'] = clean_data['Age'].astype(int)
In [49]: clean_data['Age']
```

```
Out[49]: 0
               34
          1
               45
          2
               50
          3
               50
          4
               67
               55
          Name: Age, dtype: int64
         clean_data['Salary'] = clean_data['Salary'].astype(int)
         clean_data
In [51]:
Out[51]:
             Name
                       Domain Age
                                     Location Salary Exp
              Mike Datascience
                                 34
                                      Mumbai
                                                5000
                                                        2
             Teddy
                        Testing
                                 45
                                     Bangalore
                                               10000
                                 50
          2
             Umar
                    Dataanalyst
                                     Bangalore
                                               15000
                                                        4
                                               20000
          3
              Jane
                      Analytics
                                    Hyderbad
                                 50
                                                       4.8
            Uttam
                      Statistics
                                     Bangalore
                                               30000
                                                        5
                                 67
               Kim
                          NLP
                                 55
                                         Delhi 60000
                                                        10
In [52]:
         clean data['Exp'] = clean data['Exp'].astype(int)
         clean_data['Exp']
In [53]:
Out[53]: 0
                2
          1
                3
          2
                4
          3
                4
          4
                5
               10
          Name: Exp, dtype: int64
In [54]: clean_data.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
                       Non-Null Count Dtype
         #
             Column
            Name
                       6 non-null
                                        object
         0
         1
             Domain
                       6 non-null
                                        object
                       6 non-null
                                        int64
             Age
             Location 6 non-null
                                        object
         3
         4
             Salary
                       6 non-null
                                        int64
                       6 non-null
                                        int64
             Exp
        dtypes: int64(3), object(3)
        memory usage: 420.0+ bytes
In [55]: 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 [56]: clean_data.info()
```

<class 'pandas.core.frame.DataFrame'>

Non-Null Count Dtype

RangeIndex: 6 entries, 0 to 5 Data columns (total 6 columns):

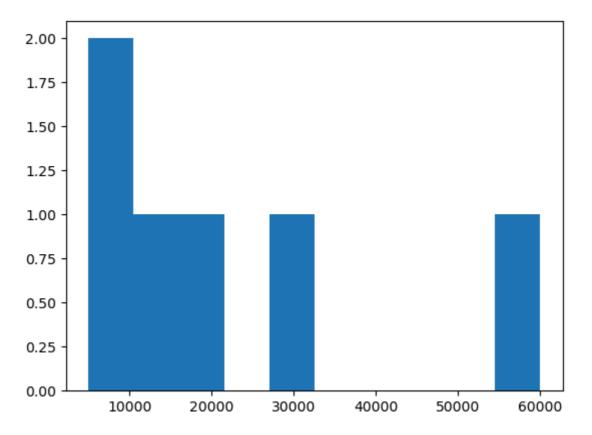
Column

```
_____
         0
             Name
                        6 non-null
                                         category
         1
             Domain
                       6 non-null
                                        category
                        6 non-null
                                         int64
             Age
         3
             Location 6 non-null
                                         category
             Salary
                        6 non-null
                                         int64
         5
                        6 non-null
                                         int64
             Exp
        dtypes: category(3), int64(3)
        memory usage: 938.0 bytes
In [57]:
         clean_data
Out[57]:
             Name
                       Domain
                                Age
                                      Location Salary Exp
              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
                                      Bangalore
                                                30000
                                                          5
                                  67
          5
               Kim
                           NLP
                                  55
                                          Delhi
                                                60000
                                                         10
          clean_data.to_csv('clean_data.csv')
In [58]:
In [59]:
          import os
          os.getcwd() # from os give the saved current working directly.
Out[59]:
         'C:\\Users\\shaik'
In [60]:
         clean_data
Out[60]:
             Name
                       Domain
                                Age
                                      Location
                                                Salary Exp
          0
              Mike
                   Datascience
                                       Mumbai
                                                 5000
                                                          2
                                  34
                                                10000
             Teddy
                        Testing
                                  45
                                      Bangalore
                                                          3
          1
          2
                    Dataanalyst
                                  50
                                      Bangalore
                                                15000
             Umar
          3
              Jane
                       Analytics
                                  50
                                      Hyderbad
                                                20000
                                                          5
             Uttam
                       Statistics
                                      Bangalore
                                                30000
                                  67
                           NLP
                                  55
          5
               Kim
                                          Delhi
                                                60000
                                                         10
```

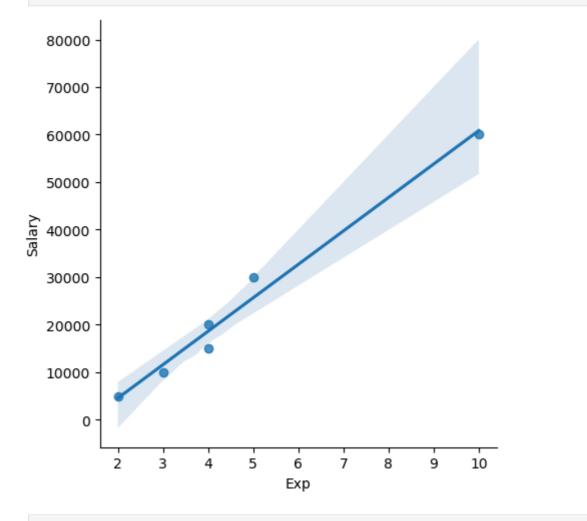
## **EDA TECHNIQUE LETS APPLY**

In [71]: import matplotlib.pyplot as plt # visualization
import seaborn as sus

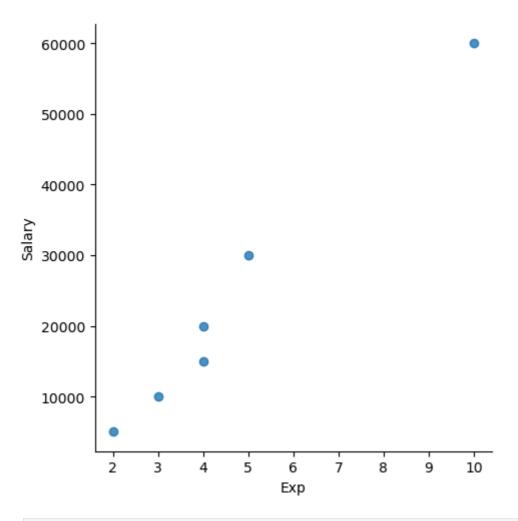
```
In [72]:
         import warnings
         warnings.filterwarnings('ignore')
In [75]:
         clean_data['Salary']
Out[75]:
                5000
          1
               10000
          2
               15000
          3
               20000
               30000
               60000
          Name: Salary, dtype: int64
In [84]:
         import seaborn as sns
         import matplotlib.pyplot as plt
         vis1 = sns.distplot(clean_data['Salary'])
In [90]:
                1e-5
           3.5
           3.0
           2.5
           2.0
           1.5
           1.0
           0.5
           0.0
              -40000 -20000
                                   Ó
                                          20000
                                                   40000
                                                             60000
                                                                      80000
                                                                               100000
                                                Salary
        vis2 = plt.hist(clean_data['Salary'])
```



In [91]: vis4 = sns.lmplot(data=clean\_data,x = 'Exp', y='Salary')



In [92]: vis5 = sns.lmplot(data=clean\_data,x = 'Exp', y='Salary', fit\_reg = False)



In [94]: clean\_data[:]

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

In [96]: clean\_data[0:6:2]

Out[96]:		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 [97]: clean\_data[::-1]

```
Domain Age
 Out[97]:
              Name
                                         Location Salary Exp
           5
                 Kim
                             NLP
                                    55
                                            Delhi
                                                   60000
                                                            10
             Uttam
                         Statistics
                                    67
                                        Bangalore
                                                   30000
                                                             5
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
                                                   20000
                                                             4
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
           1
               Teddy
                          Testing
                                    45
                                        Bangalore
                                                   10000
                                                             3
                Mike
                      Datascience
                                    34
                                          Mumbai
                                                    5000
                                                             2
 In [98]:
          clean_data.columns
Out[98]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [100...
           x_iv = clean_data[['Name','Domain','Age','Location','Exp']]
In [101...
           x_iv
Out[101...
              Name
                         Domain Age
                                       Location Exp
           0
               Mike
                      Datascience
                                    34
                                          Mumbai
                                                     2
               Teddy
                          Testing
                                    45
                                        Bangalore
                      Dataanalyst
           2
               Umar
                                    50
                                        Bangalore
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
           4
              Uttam
                         Statistics
                                    67
                                        Bangalore
                                                     5
           5
                 Kim
                             NLP
                                    55
                                            Delhi
                                                    10
In [102...
           y_dv = clean_data[['Salary']]
In [103...
           y_dv
Out[103...
              Salary
           0
                5000
               10000
               15000
           2
           3
               20000
               30000
               60000
In [105...
           emp
```

0		Г и	$\cap$	$\Gamma$	
U	шт		И	$\neg$	
$\sim$	~ ~	_	$\overline{}$		

alary Exp
5000 2
10000 3
15000 4
20000 NaN
30000 5
50000 10

In [106...

clean\_data

Out[106...

	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 [108... x\_iv

Out[108...

	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 [109... y\_dv

Out[109... Salary 0 5000 10000 15000 2 20000 30000 60000 In [110... clean\_data Out[110... Name Domain Age **Location Salary Exp** 0 Mike Datascience Mumbai 5000 2 34 Bangalore Teddy 45 10000 Testing 3 Dataanalyst Bangalore 15000 4 2 Umar 50 Hyderbad 20000 3 Jane Analytics 50 Uttam 5 Statistics 67 Bangalore 30000 NLP 5 Kim 55 Delhi 60000 10 In [111... impurtation = pd.get\_dummies(clean\_data) In [112... impurtation Out[112...

	Age	Salary	Ехр	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar
0	34	5000	2	False	False	True	False	False
1	45	10000	3	False	False	False	True	False
2	50	15000	4	False	False	False	False	True
3	50	20000	4	True	False	False	False	False
4	67	30000	5	False	False	False	False	False
5	55	60000	10	False	True	False	False	False
4			-					•

In [113...

clean\_data

Out[113		Name	D	omain	Age	Loca	tion	Salary	Ехр				
	0	Mike Datascience		34	Mui	mbai	5000	2					
	1	Teddy	Teddy Testing		45	Banga	alore	10000	3				
	2	Umar	Dataa	Dataanalyst  Analytics		Banga	alore	15000	4				
	3	Jane	An			50 Hydei		20000	4				
	<b>4</b> Uttam Statistics		67	Banga	alore	30000	5						
	5	Kim		NLP	55	Delhi		60000	10				
In [114	im	purtat	ion										
Out[114		Age Salary Exp		Name	_Jane	Nam	ne_Kim	Name	_Mike	Name_1	Гeddy	Name_Umar	
	0	34	5000 2			False		False		True		False	False
	1	45	10000	3		False		False		False		True	False
	2	50	15000	4		False		False		False		False	True

True

False

False

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

**False** 

**False** 

True

False

**False** 

False

**False** 

**False** 

**False** 

**False** 

**False** 

**False** 

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

In [ ]:

20000

30000

60000

5

10

3

50