Raw Data To Clean Data Conversion Using Python EDA

Data cleaning how to implement eda technique on dataset to build ml model the steps which we follow today any interviewer ask the datacleaning .isna() - check missing value .fillna() - fill missing value

2]:	import pandas as pd								
4]:	pdversion								
4]:	'2.2.2'								
6]:	<pre>emp = pd.read_excel(r'E:\Data Science & AI\Dataset files\Rawdata.xlsx')</pre>								
8]:	етр								
8]:		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+		
0]:	id(emp)							
0]:	25!	56826072	608						
2]:	emp	.column	S						
2]:	Ind	dex(['Na	me', 'Domain',	'Age', '	Location',	'Salary',	, 'Exp'],	dtype='obje	
4]:	emp	.shape							
4]:	(6	, 6)							
6]:	emp	head()							
6]:_		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		

```
In [18]: emp.tail()
Out[18]:
                            Domain
                                     Age
                                            Location
                                                        Salary
              Name
                                                                  Exp
             Teddy^
                            Testing
                                    45' yr
                                           Bangalore
                                                     10%%000
                                                                    <3
             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
                                                      6000^$0
                                                                  10+
                Kim
                               NLP
                                     55yr
                                               Delhi
In [20]: emp.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 6 entries, 0 to 5
        Data columns (total 6 columns):
             Column
                      Non-Null Count Dtype
                                        object
         0 Name
                      6 non-null
         1 Domain 6 non-null
                                        object
                      4 non-null
n 4 non-null
6 non-null
         2 Age
                                        object
         3 Location 4 non-null
                                        object
             Salary
                                        object
         5
                        5 non-null
             Exp
                                        object
        dtypes: object(6)
        memory usage: 420.0+ bytes
In [24]: emp
Out[24]:
              Name
                            Domain
                                        Age
                                              Location
                                                          Salary
                                                                     Ехр
          0
               Mike
                      Datascience#$ 34 years
                                                         5^00#0
                                                                      2+
                                               Mumbai
          1 Teddy^
                                       45' yr Bangalore
                                                                      <3
                            Testing
                                                       10%%000
             Uma#r Dataanalyst^^#
                                                        1$5%000
          2
                                       NaN
                                                  NaN
                                                                   4> yrs
          3
                        Ana^^lytics
                                       NaN Hyderbad
                                                         2000^0
                                                                    NaN
               Jane
                                                                  5+ year
          4
             Uttam*
                           Statistics
                                       67-yr
                                                  NaN
                                                          30000-
          5
                               NLP
                                        55yr
                                                        6000^$0
                                                                     10+
                Kim
                                                 Delhi
         emp['Domain']
In [22]:
Out[22]: 0
                Datascience#$
          1
                      Testing
          2
               Dataanalyst^^#
          3
                  Ana^^lytics
          4
                   Statistics
                          NLP
          Name: Domain, dtype: object
In [26]: emp.isnull()#emp.isna()
```

```
Out[26]:
                Name Domain Age Location Salary
                                                        Exp
            0
                 False
                          False False
                                          False
                                                 False False
             1
                 False
                          False False
                                          False
                                                 False False
             2
                 False
                          False
                                True
                                          True
                                                 False False
             3
                 False
                          False
                                          False
                                True
                                                 False
                                                        True
             4
                 False
                          False False
                                                 False False
                                          True
                 False
                          False False
                                          False
                                                 False False
            emp.isnull().sum()
  In [28]:
  Out[28]: Name
                         0
             Domain
                         0
             Age
                         2
             Location
                         2
             Salary
                         0
             Exp
             dtype: int64
Data Cleaning or Data Cleansing
  In [30]:
            emp['Name']
  Out[30]: 0
                    Mike
             1
                  Teddy^
             2
                  Uma#r
             3
                    Jane
             4
                  Uttam*
             5
                     Kim
             Name: Name, dtype: object
           emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)#nonword charac
  In [34]:
            emp['Name']
  In [36]:
  Out[36]: 0
                   Mike
             1
                  Teddy
             2
                  Umar
             3
                   Jane
             4
                  Uttam
             5
                    Kim
             Name: Name, dtype: object
            emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
  In [42]:
  In [44]:
            emp['Domain']
  Out[44]: 0
                  Datascience
             1
                      Testing
                  Dataanalyst
             2
             3
                  Analytics
             4
                   Statistics
             5
                          NLP
             Name: Domain, dtype: object
```

```
In [46]: emp['Age'] = emp['Age'].str.replace(r'\W','',regex=True)
In [48]:
         emp['Age']
Out[48]: 0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
          5
                  55yr
          Name: Age, dtype: object
In [52]: emp['Age'] = emp['Age'].str.extract(r'(\d+)')
In [54]:
         emp['Age']
Out[54]:
                34
          0
                45
          1
          2
               NaN
          3
               NaN
          4
                67
          5
                55
          Name: Age, dtype: object
In [56]:
         emp
Out[56]:
             Name
                       Domain Age
                                      Location
                                                  Salary
                                                             Exp
          0
              Mike
                    Datascience
                                 34
                                       Mumbai
                                                  5^00#0
                                                              2+
             Teddy
                                 45
                                      Bangalore
                                                10%%000
                                                              <3
                        Testing
          2
             Umar
                    Dataanalyst NaN
                                          NaN
                                                1$5%000
                                                           4> yrs
                                                  2000^0
          3
              Jane
                      Analytics
                                NaN
                                      Hyderbad
                                                            NaN
                                                  30000- 5+ year
          4
             Uttam
                       Statistics
                                 67
                                          NaN
                          NLP
                                  55
                                                 6000^$0
          5
               Kim
                                          Delhi
                                                             10+
In [58]: emp['Location'] = emp['Location'].str.replace(r'\W','')
In [60]:
         emp['Location']
Out[60]:
          0
                  Mumbai
          1
               Bangalore
          2
                     NaN
          3
                Hyderbad
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
In [66]:
         emp['Salary'] = emp['Salary'].str.replace(r'\W','',regex=True)
In [68]: emp['Salary']
```

```
1
                   10000
              2
                    15000
              3
                   20000
              4
                    30000
              5
                   60000
              Name: Salary, dtype: object
  In [70]:
             emp
  Out[70]:
                                      Age
                                                      Salary
                 Name
                            Domain
                                             Location
                                                                   Exp
              0
                  Mike
                         Datascience
                                        34
                                              Mumbai
                                                         5000
                                                                    2+
                 Teddy
                             Testing
                                        45
                                            Bangalore
                                                        10000
                                                                    <3
              1
              2
                 Umar
                         Dataanalyst
                                      NaN
                                                 NaN
                                                        15000
                                                                 4> yrs
              3
                  Jane
                            Analytics
                                      NaN
                                            Hyderbad
                                                        20000
                                                                  NaN
              4
                 Uttam
                            Statistics
                                        67
                                                 NaN
                                                        30000
                                                               5+ year
              5
                   Kim
                                NLP
                                        55
                                                 Delhi
                                                        60000
                                                                   10+
             emp['Exp'] = emp['Exp'].str.extract(r'(\d+)')
  In [72]:
  In [74]:
             emp['Exp']
                      2
  Out[74]:
              0
                      3
              1
              2
                      4
              3
                   NaN
              4
                      5
                    10
              Name: Exp, dtype: object
  In [76]:
             emp
  Out[76]:
                 Name
                            Domain
                                             Location
                                                       Salary
                                      Age
                                                                Exp
                                                                   2
              0
                  Mike
                         Datascience
                                        34
                                              Mumbai
                                                         5000
                 Teddy
                             Testing
                                        45
                                            Bangalore
                                                        10000
              2
                 Umar
                         Dataanalyst
                                      NaN
                                                 NaN
                                                        15000
                                                                   4
              3
                  Jane
                            Analytics
                                      NaN
                                            Hyderbad
                                                        20000
                                                               NaN
              4
                 Uttam
                            Statistics
                                        67
                                                 NaN
                                                        30000
                                                                   5
              5
                   Kim
                                NLP
                                        55
                                                 Delhi
                                                        60000
                                                                 10
  In [78]:
             clean_data = emp.copy()
>>>>>Till now we have raw data we use regex to clean the data and removed all noise characted from the dataset
```

>>>>>you can also work in same things in sql query as well>Missing Values Treatment for Numerical data

Out[68]:

In [80]:

clean_data

5000

```
Out[80]:
                                                                           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
                                                                       Uttam
                                                                                                                                   Statistics
                                                                                                                                                                                                                                                 NaN
                                                                                                                                                                                                                                                                                   30000
                                                                                                                                                                                                                                                                                                                                            5
                                                                                                                                                                                                67
                                                                                       Kim
                                                                                                                                                         NLP
                                                                                                                                                                                                 55
                                                                                                                                                                                                                                               Delhi
                                                                                                                                                                                                                                                                                    60000
                                                                                                                                                                                                                                                                                                                                       10
In [82]: clean_data['Age']
Out[82]: 0
                                                                                             34
                                                                                            45
                                                          2
                                                                                      NaN
                                                          3
                                                                                      NaN
                                                          4
                                                                                            67
                                                                                             55
                                                          Name: Age, dtype: object
In [84]: import numpy as np
In [86]: 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.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_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'].f
In [88]: clean_data['Age']
Out[88]: 0
                                                                                                         34
                                                                                                     45
                                                          2
                                                                               50.25
                                                          3
                                                                                50.25
                                                          4
                                                                                                        67
                                                                                                         55
                                                          Name: Age, dtype: object
In [90]: clean_data['Exp']
Out[90]: 0
                                                                                                   2
                                                          1
                                                                                                   3
                                                          2
                                                                                                   4
                                                          3
                                                                                      NaN
                                                          4
                                                                                                   5
                                                                                             10
                                                          Name: Exp, dtype: object
In [92]: 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.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp'])).fillna(np
In [94]: clean_data['Exp']
Out[94]: 0
                                                                                                    2
                                                                                                    3
                                                          1
                                                                                                   4
                                                          2
                                                          3
                                                                                       4.8
                                                          4
                                                                                                    5
                                                          Name: Exp, dtype: object
```

```
In [96]:
           clean_data
Out[96]:
                         Domain
               Name
                                           Location Salary Exp
                                    Age
                                                               2
           0
                Mike
                      Datascience
                                     34
                                           Mumbai
                                                      5000
               Teddy
                           Testing
                                     45
                                          Bangalore
                                                     10000
                                                               3
                                   50.25
           2
               Umar
                      Dataanalyst
                                               NaN
                                                     15000
                                                               4
           3
                Jane
                         Analytics
                                   50.25
                                          Hyderbad
                                                     20000
                                                              4.8
                                                               5
           4
               Uttam
                         Statistics
                                     67
                                               NaN
                                                     30000
           5
                             NLP
                 Kim
                                      55
                                              Delhi
                                                     60000
                                                              10
           clean_data['Location'].isnull().sum()
 In [98]:
Out[98]: 2
In [100...
           clean_data['Location']
           0
Out[100...
                    Mumbai
           1
                 Bangalore
           2
                        NaN
           3
                  Hyderbad
           4
                        NaN
                     Delhi
           Name: Location, dtype: object
In [102...
           clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mo
In [104...
           clean_data['Location']
Out[104...
                    Mumbai
           1
                 Bangalore
           2
                 Bangalore
           3
                  Hyderbad
           4
                 Bangalore
           5
                     Delhi
           Name: Location, dtype: object
In [106...
           clean_data
Out[106...
               Name
                         Domain
                                    Age
                                           Location Salary Exp
                                                               2
           0
                Mike
                                     34
                                                      5000
                      Datascience
                                           Mumbai
                                                     10000
                                                               3
               Teddy
                           Testing
                                     45
                                          Bangalore
           2
                                   50.25
                                                     15000
                                                               4
               Umar
                      Dataanalyst
                                          Bangalore
           3
                Jane
                         Analytics
                                   50.25
                                          Hyderbad
                                                     20000
                                                              4.8
                                                               5
                                                     30000
           4
               Uttam
                         Statistics
                                     67
                                          Bangalore
           5
                             NLP
                                      55
                 Kim
                                              Delhi
                                                     60000
                                                              10
In [108...
           clean_data.info()
```

```
RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
           # Column Non-Null Count Dtype
          --- ----- -----
          0 Name 6 non-null object
1 Domain 6 non-null object
2 Age 6 non-null object
           3 Location 6 non-null object
4 Salary 6 non-null object
5 Exp 6 non-null object
          dtypes: object(6)
          memory usage: 420.0+ bytes
In [110...
           clean_data['Age'] = clean_data['Age'].astype(int)
In [112...
          clean data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
           # Column Non-Null Count Dtype
          --- -----
           0 Name 6 non-null
                                          object
          1 Domain 6 non-null object
2 Age 6 non-null int32
3 Location 6 non-null object
4 Salary 6 non-null object
5 Exp 6 non-null object
          dtypes: int32(1), object(5)
          memory usage: 396.0+ bytes
In [114...
           clean_data['Salary'] = clean_data['Salary'].astype(int)
           clean_data['Exp'] = clean_data['Exp'].astype(int)
In [116...
          clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
           # Column Non-Null Count Dtype
          --- -----
          0 Name 6 non-null object
1 Domain 6 non-null object
2 Age 6 non-null int32
3 Location 6 non-null object
           4 Salary 6 non-null
                                           int32
           5
                         6 non-null
               Exp
                                            int32
          dtypes: int32(3), object(3)
          memory usage: 348.0+ bytes
In [118...
           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 [120...
          clean_data.info()
```

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

RangeIndex: 6 entries, 0 to 5 Data columns (total 6 columns): Column Non-Null Count Dtype 0 Name 6 non-null category 1 Domain 6 non-null category 6 non-null int32 Age Location 6 non-null 3 category 4 Salary 6 non-null int32 int32 5 Exp 6 non-null dtypes: category(3), int32(3) memory usage: 866.0 bytes In [122... clean_data Out[122... Name Domain Age Location Salary Exp 0 Mike Datascience 34 Mumbai 5000 2 Teddy **Testing** 45 Bangalore 10000 3 1 2 Umar Dataanalyst 50 Bangalore 15000 4 3 Analytics 50 Hyderbad 20000 4 Jane 4 Statistics Bangalore 30000 5 Uttam 67 5 NLP 55 Delhi 60000 Kim 10 In [124... clean_data.to_csv('clean_data.csv') import os In [126... os.getcwd() 'C:\\Users\\roy62\\Data Science & AI' Out[126... In [128... clean_data Out[128... Name **Domain** Age Location Salary Exp 5000 2 0 Mike Datascience 34 Mumbai Teddy 45 Bangalore 10000 3 Testing Dataanalyst Bangalore 15000 2 Umar 50 4 Jane 50 Hyderbad 20000 3 Analytics 5 Bangalore 30000 Uttam Statistics 67 NLP 55 5 Delhi 60000 Kim 10

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

EDA TECHNIQUE LETS APPLY

import matplotlib.pyplot as plt # visualization
import seaborn as sns

```
In [131...
           import warnings
           warnings.filterwarnings('ignore')
In [134...
           clean_data['Salary']
Out[134...
           0
                 5000
           1
                10000
           2
                15000
           3
                20000
           4
                30000
                60000
           Name: Salary, dtype: int32
          vis1 = sns.distplot(clean_data['Salary'])
In [136...
                 1e-5
             3.5
             3.0
             2.5
             2.0
             1.5
             1.0
             0.5
```

In [138... vis2 = plt.hist(clean_data['Salary'])

20000

40000

Salary

60000

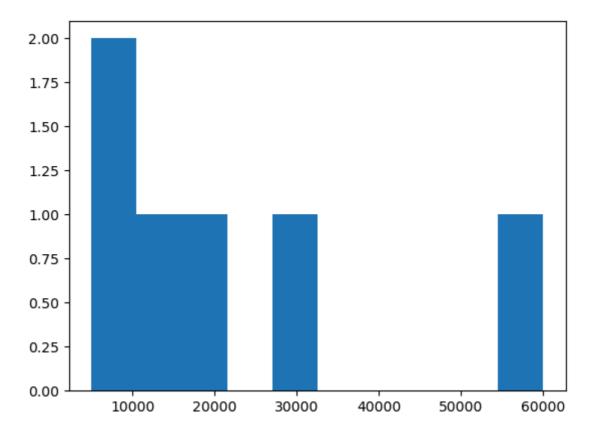
80000

100000

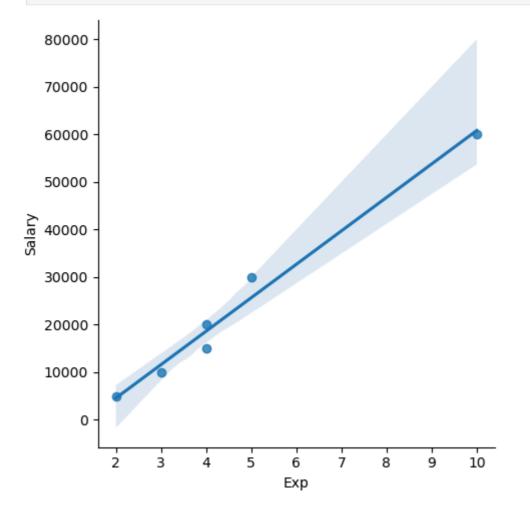
ò

0.0

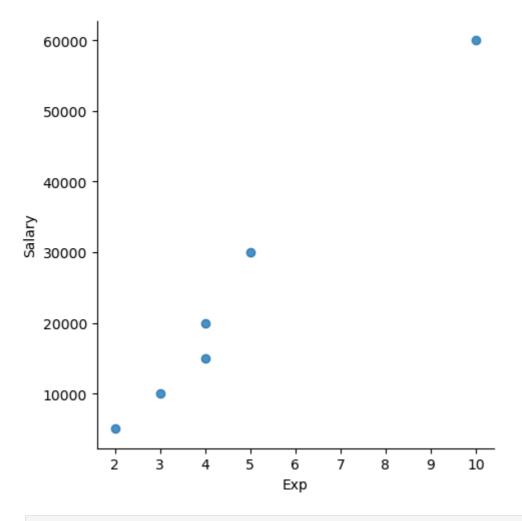
-40000 -20000



In [140... vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary')



In [142... vis5 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = False)



In [144... clear

clean_data[:]

Out[144...

	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 [146...

clean_data[0:6:2]

Out[146...

	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 [148...

clean_data[::-1]

```
Out[148...
              Name
                        Domain Age Location Salary Exp
           5
                Kim
                            NLP
                                    55
                                            Delhi
                                                   60000
                                                           10
           4 Uttam
                        Statistics
                                    67
                                        Bangalore
                                                   30000
                                                            5
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
                                                   20000
                                                            4
           2
               Umar
                      Dataanalyst
                                    50
                                        Bangalore
                                                   15000
               Teddy
                          Testing
                                   45
                                        Bangalore
                                                   10000
                                                            3
                Mike
                      Datascience
                                    34
                                         Mumbai
                                                    5000
                                                            2
In [150...
           clean_data.columns
Out[150...
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
          X_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [152...
          X_iv
In [154...
Out[154...
                         Domain Age
              Name
                                       Location Exp
               Mike
                      Datascience
                                    34
                                         Mumbai
                                                    2
               Teddy
                          Testing
                                   45
                                        Bangalore
                                                    3
           2
               Umar
                      Dataanalyst
                                   50
                                        Bangalore
                                                    4
           3
                Jane
                        Analytics
                                    50
                                        Hyderbad
           4
              Uttam
                        Statistics
                                    67
                                        Bangalore
                                                    5
           5
                Kim
                            NLP
                                    55
                                            Delhi
                                                    10
In [156...
          y_dv = clean_data[['Salary']]
In [158...
           y_dv
Out[158...
              Salary
                5000
           0
               10000
               15000
           2
           3
               20000
               30000
               60000
In [160...
           emp
```

Out[160		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	NaN	NaN	15000	4
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN
	4	Uttam	Statistics	67	NaN	30000	5
	5	Kim	NLP	55	Delhi	60000	10
T [162	- 7		-				
In [162	CI	ean_dat	а				
Out[162		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	20000	5
					Bangaiore	30000	3
	5	Kim	NLP	55	Delhi	60000	10
					_		
In [164	5				_		
In [164 Out[164				55	_	60000	
		iv	NLP	55	Delhi	60000	
	X_	iv Name	NLP Domain	55 Age	Delhi	60000 Exp	
	x_ 0	iv Name Mike	NLP Domain Datascience	55 Age 34	Delhi Location Mumbai	Exp 2	

5

10

In [166... y_dv

Statistics

NLP

Uttam

Kim

5

67

55

Bangalore

Delhi

Out[166... Salary 5000 0 10000 15000 2 20000 30000 60000 In [168... clean_data Out[168... Name Domain Age **Location Salary Exp** 0 2 Mike Datascience 34 Mumbai 5000 Bangalore 1 Teddy **Testing** 45 10000 3 Dataanalyst Bangalore 2 Umar 50 15000 4 3 Jane Analytics 50 Hyderbad 20000 Uttam Statistics 67 Bangalore 30000 5 5 Kim NLP 55 Delhi 60000 10 imputation = pd.get_dummies(clean_data) In [170... In [172... imputation Out[172... Age Salary Exp Name_Jane Name_Kim Name_Mike Name_Teddy Name_Umar 0 34 5000 2 False False True False False 1 45 10000 3 False False False False True 2 50 15000 4 False False False True False 3 50 20000 False True False False False 4 30000 5 False False 67 False False False 5 55 60000 10 False False True False False

In [174...

clean_data

$\cap \cup +$	17/
Uul	1/4

	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 [176...

imputation

Out[176...

	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

#--->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 []: