```
import pandas as pd
emp=pd.read excel(r'C:\Users\Admin\Downloads\Rawdata.xlsx')
emp
     Name
                   Domain
                                       Location
                                                  Salary
                                                               Exp
                                 Age
     Mike
                                                  5^00#0
0
            Datascience#$
                            34 years
                                         Mumbai
                                                                2+
1
  Teddv^
                  Testing
                              45' yr
                                      Bangalore
                                                 10%%000
                                                                <3
2
           Dataanalyst^^#
                                                            4> yrs
   Uma#r
                                 NaN
                                            NaN
                                                 1$5%000
3
     Jane
              Ana^^lytics
                                 NaN
                                       Hyderbad
                                                  2000^0
                                                               NaN
4
   Uttam*
               Statistics
                               67-yr
                                            NaN
                                                  30000-
                                                           5+ vear
5
      Kim
                      NLP
                                55yr
                                          Delhi
                                                 6000^$0
                                                               10+
emp.shape ##dimensions of the dataframe
(6, 6)
len(emp)
6
emp.head()
     Name
                   Domain
                                 Age
                                       Location
                                                  Salary
                                                               Exp
                            34 years
0
     Mike
            Datascience#$
                                         Mumbai
                                                  5^00#0
                                                                2+
                             45' yr
1
  Teddy^
                                                                <3
                  Testing
                                      Bangalore
                                                 10%%000
2
           Dataanalyst^^#
                                 NaN
                                                 1$5%000
   Uma#r
                                            NaN
                                                            4> yrs
3
              Ana^^lytics
     Jane
                                 NaN
                                       Hyderbad
                                                  2000^0
                                                               NaN
  Uttam*
               Statistics
                                            NaN
                                                  30000 -
                               67-yr
                                                           5+ year
emp.tail()
                   Domain
                                     Location
     Name
                               Age
                                                Salary
                                                             Exp
                           45' yr
                  Testing
1
  Teddy^
                                    Bangalore
                                               10%%000
                                                              <3
2
           Dataanalyst^^#
                               NaN
    Uma#r
                                          NaN
                                               1$5%000
                                                          4> yrs
              Ana^^lytics
3
     Jane
                               NaN
                                     Hyderbad
                                                2000^0
                                                             NaN
                             67-yr
                                                30000 -
4
  Uttam*
               Statistics
                                          NaN
                                                         5+ year
5
      Kim
                      NLP
                              55yr
                                        Delhi
                                               6000^$0
                                                             10+
emp.columns
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'],
dtype='object')
len(emp.columns)
6
emp.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
```

```
Data columns (total 6 columns):
              Non-Null Count Dtype
 #
     Column
 0
     Name
              6 non-null
                              object
 1
     Domain
              6 non-null
                              object
 2
              4 non-null
                              object
    Age
 3
    Location 4 non-null
                              object
 4
     Salary
              6 non-null
                              object
 5
              5 non-null
     Exp
                              object
dtypes: object(6)
memory usage: 420.0+ bytes
emp.isnull()
                                  Salary
    Name Domain
                   Age
                        Location
                                            Exp
0
   False False False
                           False
                                   False
                                          False
  False
          False False
                           False
                                          False
                                   False
2
                                          False
   False
          False True
                            True
                                   False
3
   False
          False True
                           False
                                   False
                                          True
          False False
4
   False
                           True
                                   False
                                          False
5 False False False
                           False
                                   False False
emp.isnull().sum()
Name
           0
Domain
           0
           2
Age
            2
Location
           0
Salary
            1
Exp
dtype: int64
emp['Name']
0
      Mike
1
     Teddy^
2
     Uma#r
3
       Jane
4
     Uttam*
5
       Kim
Name: Name, dtype: object
emp['Domain']
0
      Datascience#$
1
           Testing
2
     Dataanalyst^^#
3
       Ana^^lytics
4
         Statistics
5
               NLP
Name: Domain, dtype: object
```

```
emp['Age']
0
     34 years
1
       45' yr
2
          NaN
3
          NaN
4
        67-yr
5
         55yr
Name: Age, dtype: object
emp['Location']
0
        Mumbai
1
     Bangalore
2
           NaN
3
      Hyderbad
4
           NaN
5
         Delhi
Name: Location, dtype: object
emp['Salary']
0
      5^00#0
1
     10%%000
2
     1$5%000
3
      2000^0
4
      30000-
5
     6000^$0
Name: Salary, dtype: object
emp['Exp']
0
          2+
1
          <3
2
      4> yrs
3
         NaN
4
     5+ year
5
         10 +
Name: Exp, dtype: object
emp[['Name','Domain']]
                    Domain
     Name
0
     Mike
            Datascience#$
1
  Teddy^
                  Testing
2
  Uma#r
           Dataanalyst^^#
3
     Jane
              Ana^^lytics
               Statistics
4
  Uttam*
5
      Kim
                      NLP
emp[['Name','Domain','Age']]
```

```
Name
                   Domain
                                 Age
0
     Mike
            Datascience#$
                            34 years
1
  Teddy^
                  Testing
                              45' yr
2
    Uma#r
           Dataanalyst^^#
                                 NaN
              Ana^^lytics
3
     Jane
                                 NaN
4
  Uttam*
               Statistics
                               67-yr
5
      Kim
                                55yr
                      NLP
emp[['Name','Domain','Age','Location','Salary','Exp']]
     Name
                   Domain
                                 Age
                                       Location
                                                   Salary
                                                                Exp
     Mike
0
            Datascience#$
                           34 years
                                         Mumbai
                                                   5^00#0
                                                                2+
                              45' yr
1
  Teddy^
                                      Bangalore
                                                  10%%000
                                                                <3
                  Testing
2
   Uma#r
           Dataanalyst^^#
                                 NaN
                                                  1$5%000
                                            NaN
                                                            4> yrs
3
              Ana^^lytics
                                 NaN
                                       Hyderbad
                                                   2000^0
     Jane
                                                               NaN
4
   Uttam*
               Statistics
                               67-yr
                                            NaN
                                                   30000-
                                                           5+ year
5
      Kim
                                          Delhi
                                                  6000^$0
                      NLP
                                55yr
                                                                10 +
```

Data Cleansing

```
emp['Name']
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','')
emp['Name']
0
       Mike
     Teddy^
1
2
      Uma#r
3
       Jane
4
     Uttam*
5
        Kim
Name: Name, dtype: object
emp['Domain']= emp['Domain'].str.replace(r'\W','')
emp['Domain']
0
      Datascience#$
1
            Testing
2
     Dataanalyst^^#
3
        Ana^^lytics
4
         Statistics
```

5 NLP

Name: Domain, dtype: object

EDA Intro

```
import pandas as pd
pd.__version__
'2.2.2'
emp=pd.read excel(r'C:\Users\Admin\Downloads\Rawdata.xlsx')
emp
     Name
                   Domain
                                       Location
                                                   Salary
                                 Age
                                                               Exp
     Mike
            Datascience#$
                                         Mumbai
                                                   5^00#0
                                                                2+
0
                           34 years
                  Testing
                              45' yr
1
  Teddy^
                                      Bangalore
                                                 10%%000
                                                                <3
           Dataanalyst^^#
2
   Uma#r
                                 NaN
                                            NaN
                                                 1$5%000
                                                            4> yrs
              Ana^^lytics
3
                                       Hyderbad
     Jane
                                 NaN
                                                   2000^0
                                                               NaN
4
  Uttam*
               Statistics
                                            NaN
                                                   30000-
                               67-yr
                                                           5+ year
5
      Kim
                      NLP
                                          Delhi
                                                 6000^$0
                                                               10+
                                55yr
emp.isnull().sum()
            0
Name
Domain
            0
            2
Age
            2
Location
Salary
            0
Exp
            1
dtype: int64
id(emp)
1943775782160
emp.columns
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'],
dtype='object')
emp.shape
(6, 6)
emp.head
<bound method NDFrame.head of</pre>
                                    Name
                                                   Domain
                                                                Age
Location
           Salary
                       Exp
     Mike
            Datascience#$ 34 years
                                         Mumbai
                                                   5^00#0
                                                                2+
1 Teddy^
                  Testing
                              45' yr Bangalore
                                                 10%%000
                                                                <3
```

```
2
           Dataanalyst^^#
                                 NaN
                                                  1$5%000
    Uma#r
                                             NaN
                                                             4> vrs
3
              Ana^^lytics
                                 NaN
     Jane
                                        Hyderbad
                                                   2000^0
                                                                NaN
4
   Uttam*
               Statistics
                               67-yr
                                             NaN
                                                   30000 -
                                                            5+ year
5
      Kim
                       NLP
                                55yr
                                           Delhi
                                                  6000^$0
                                                                10+>
emp.tail
<bound method NDFrame.tail of</pre>
                                    Name
                                                   Domain
                                                                 Age
Location
           Salary
                        Exp
                           34 years
                                                   5^00#0
                                                                 2+
     Mike
            Datascience#$
                                          Mumbai
                              45' yr
1
   Teddv^
                  Testing
                                      Bangalore
                                                  10%%000
                                                                 <3
2
           Dataanalyst^^#
    Uma#r
                                 NaN
                                             NaN
                                                  1$5%000
                                                             4> yrs
              Ana^^lytics
3
                                 NaN
                                        Hyderbad
                                                   2000^0
     Jane
                                                                NaN
4
               Statistics
                                             NaN
   Uttam*
                               67-yr
                                                   30000 -
                                                            5+ year
5
                                           Delhi
      Kim
                       NLP
                                55yr
                                                  6000^$0
                                                                10+>
emp.isnull().sum()
            0
Name
Domain
            0
            2
Age
            2
Location
Salary
            0
            1
Exp
dtype: int64
emp.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
#
     Column
               Non-Null Count
                                Dtype
- - -
 0
     Name
               6 non-null
                                object
1
     Domain
               6 non-null
                                object
 2
               4 non-null
     Age
                                object
 3
               4 non-null
     Location
                                object
4
     Salary
               6 non-null
                                object
 5
               5 non-null
                                object
     Exp
dtypes: object(6)
memory usage: 420.0+ bytes
emp.isnull()
    Name
          Domain
                     Age
                          Location
                                    Salary
                                               Exp
   False
           False False
                             False
                                     False
                                             False
1
   False
           False False
                             False
                                     False
                                             False
2
   False
           False
                   True
                              True
                                     False
                                             False
3
           False
                   True
                             False
                                     False
                                              True
   False
4
   False
           False False
                              True
                                     False
                                             False
5
   False
           False False
                             False
                                     False
                                             False
```

```
emp.isna()
   Name Domain
                   Age Location
                                  Salary
                                            Exp
          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
```

Data Cleaning or Data Cleansing

```
emp
                    Domain
                                                    Salary
                                                                 Exp
     Name
                                  Age
                                        Location
            Datascience#$
0
     Mike
                            34 years
                                          Mumbai
                                                    5^00#0
                                                                  2+
                              45' yr
1
  Teddy^
                   Testing
                                       Bangalore
                                                   10%%000
                                                                  <3
2
    Uma#r
           Dataanalyst^^#
                                  NaN
                                                   1$5%000
                                             NaN
                                                              4> yrs
3
              Ana^^lytics
     Jane
                                  NaN
                                        Hyderbad
                                                    2000^0
                                                                 NaN
4
                                             NaN
  Uttam*
                Statistics
                                67-yr
                                                    30000 -
                                                            5+ year
5
      Kim
                       NLP
                                 55yr
                                           Delhi
                                                   6000^$0
                                                                 10 +
emp['Name']
0
       Mike
1
     Teddy^
2
      Uma#r
3
       Jane
4
     Uttam*
5
        Kim
Name: Name, dtype: object
emp['Name']=emp['Name'].str.replace(r'\W','',regex=True)
                                                              #regex
means regular expression non word character
emp['Name']
0
      Mike
1
     Teddy
2
      Umar
3
      Jane
4
     Uttam
5
       Kim
Name: Name, dtype: object
emp
    Name
                   Domain
                                 Age
                                       Location
                                                   Salary
                                                                Exp
0
    Mike
           Datascience#$
                           34 years
                                         Mumbai
                                                   5^00#0
                                                                 2+
1
  Teddv
                                                  10%%000
                                                                 <3
                  Testing
                             45' yr
                                      Bangalore
    Umar
          Dataanalyst^^#
                                 NaN
                                                  1$5%000
                                            NaN
                                                            4> yrs
```

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

```
0
      5000
1
     10000
2
     15000
3
     20000
4
     30000
5
     60000
Name: Salary, dtype: object
emp
                                                         Exp
    Name
                Domain
                            Age
                                   Location Salary
    Mike
         Datascience
                       34years
0
                                     Mumbai
                                              5000
                                                          2+
                                                          <3
1
  Teddy
              Testing
                           45yr
                                 Bangalore
                                             10000
2
   Umar
          Dataanalyst
                            NaN
                                        NaN
                                             15000
                                                      4> yrs
3
    Jane
                            NaN
                                  Hyderbad
                                             20000
            Analytics
                                                         NaN
4
   Uttam
           Statistics
                           67yr
                                        NaN
                                             30000
                                                     5+ year
5
     Kim
                   NLP
                                      Delhi 60000
                           55yr
                                                         10+
emp['Exp']=emp['Exp'].str.replace(r'\W','',regex=True)
emp['Exp']
         2
0
1
         3
2
      4yrs
3
       NaN
4
     5year
5
        10
Name: Exp, dtype: object
emp
    Name
                Domain
                                   Location Salary
                                                       Exp
                            Age
                                                         2
0
    Mike
          Datascience
                        34years
                                     Mumbai
                                              5000
                                                         3
1
  Teddy
              Testing
                           45yr
                                 Bangalore
                                             10000
2
                            NaN
    Umar
          Dataanalyst
                                        NaN
                                             15000
                                                      4yrs
3
    Jane
            Analytics
                            NaN
                                  Hyderbad
                                             20000
                                                       NaN
4
                                        NaN
  Uttam
           Statistics
                           67yr
                                             30000
                                                     5year
5
                   NLP
     Kim
                           55yr
                                      Delhi 60000
                                                        10
emp['Age']=emp['Age'].str.extract('(\\d+)')
emp['Age']
0
      34
1
      45
2
     NaN
3
     NaN
4
      67
5
      55
Name: Age, dtype: object
```

```
emp
    Name
                Domain
                       Age
                               Location Salary
                                                    Exp
    Mike
                                                      2
          Datascience
                          34
                                 Mumbai
                                           5000
1
   Teddy
               Testing
                          45
                              Bangalore
                                          10000
                                                      3
2
    Umar
          Dataanalyst
                                          15000
                         NaN
                                     NaN
                                                   4yrs
3
    Jane
            Analytics
                         NaN
                               Hyderbad
                                          20000
                                                    NaN
4
  Uttam
           Statistics
                          67
                                     NaN
                                          30000
                                                  5year
5
                          55
                                  Delhi
                                          60000
     Kim
                   NLP
                                                     10
emp['Exp']=emp['Exp'].str.extract('(\\d+)')
emp['Exp']
0
       2
1
       3
2
       4
3
     NaN
4
       5
5
      10
Name: Exp, dtype: object
emp
                               Location Salary
    Name
                Domain
                         Age
                                                  Exp
0
    Mike
          Datascience
                          34
                                 Mumbai
                                           5000
                                                    2
                              Bangalore
                                                    3
1
  Teddy
                          45
                                          10000
               Testing
2
                                                    4
    Umar
          Dataanalyst
                         NaN
                                     NaN
                                          15000
3
    Jane
            Analytics
                         NaN
                               Hyderbad
                                          20000
                                                  NaN
4
   Uttam
                          67
                                          30000
                                                    5
            Statistics
                                     NaN
5
     Kim
                   NLP
                          55
                                  Delhi
                                          60000
                                                   10
clean data=emp.copy()
clean data
                               Location Salary
    Name
                Domain
                         Age
                                                  Exp
    Mike
                                                    2
0
          Datascience
                          34
                                 Mumbai
                                           5000
1
  Teddy
               Testing
                          45
                              Bangalore
                                          10000
                                                    3
2
                                                    4
    Umar
          Dataanalyst
                         NaN
                                     NaN
                                          15000
3
            Analytics
                               Hyderbad
                                          20000
                                                  NaN
    Jane
                         NaN
4
   Uttam
            Statistics
                          67
                                     NaN
                                          30000
                                                    5
5
                   NLP
                          55
     Kim
                                  Delhi
                                          60000
                                                   10
clean data['Age']
0
      34
1
      45
2
     NaN
3
     NaN
4
      67
```

```
55
Name: Age, dtype: object
import numpy as np
clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean
data['Age'])))
clean data['Age']
        34
0
1
        45
2
     50.25
3
     50.25
4
        67
5
        55
Name: Age, dtype: object
clean data['Exp']
0
       2
1
       3
2
       4
3
     NaN
4
       5
5
      10
Name: Exp, dtype: object
clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean
_data['Exp'])))
clean data['Exp']
0
       2
       3
1
2
       4
3
     4.8
4
       5
5
Name: Exp, dtype: object
clean_data
                                                  Exp
    Name
               Domain
                          Age
                                Location Salary
    Mike
          Datascience
                           34
                                  Mumbai
                                            5000
                                                    2
                                                    3
1
  Teddy
              Testing
                           45
                               Bangalore
                                           10000
   Umar
          Dataanalyst
                        50.25
                                           15000
                                                    4
                                      NaN
3
            Analytics 50.25
                                Hyderbad
                                           20000
                                                  4.8
    Jane
4
           Statistics
                                      NaN
                                                    5
  Uttam
                           67
                                           30000
5
     Kim
                   NLP
                           55
                                   Delhi
                                           60000
                                                   10
clean_data['Location'].isnull().sum()
```

```
2
clean data['Location']
0
        Mumbai
1
     Bangalore
2
           NaN
3
      Hyderbad
4
           NaN
5
         Delhi
Name: Location, dtype: object
clean data['Location']=clean data['Location'].fillna(clean data['Locat
ion'].mode()[0])
clean_data['Location']
0
        Mumbai
1
     Bangalore
2
     Bangalore
3
      Hyderbad
4
     Bangalore
5
         Delhi
Name: Location, dtype: object
clean data
                                                  Exp
    Name
               Domain
                          Age
                                Location Salary
0
    Mike
         Datascience
                           34
                                  Mumbai
                                            5000
                                                    2
                                          10000
                                                    3
1
  Teddy
              Testing
                           45
                               Bangalore
    Umar Dataanalyst 50.25
2
                               Bangalore
                                           15000
                                                    4
3
                                                  4.8
    Jane
            Analytics 50.25
                               Hyderbad
                                           20000
4
                           67
                               Bangalore
                                                    5
  Uttam
           Statistics
                                           30000
5
     Kim
                  NLP
                           55
                                   Delhi
                                           60000
                                                   10
clean data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #
     Column
               Non-Null Count
                                Dtype
- - -
 0
     Name
               6 non-null
                                object
 1
     Domain
               6 non-null
                                object
 2
     Age
               6 non-null
                                object
 3
     Location 6 non-null
                                object
 4
     Salary
               6 non-null
                                object
 5
     Exp
               6 non-null
                                object
dtypes: object(6)
memory usage: 420.0+ bytes
```

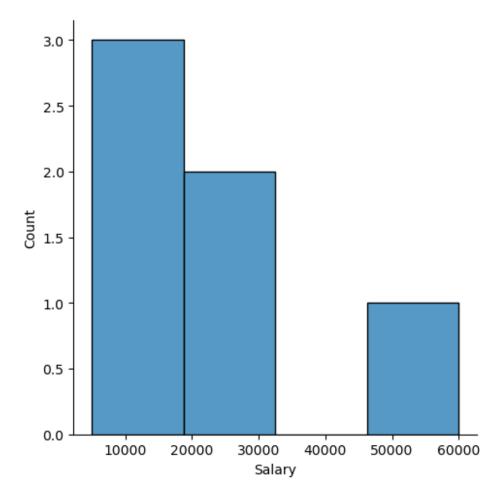
```
clean data['Age']=clean data['Age'].astype(int)
clean data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
#
     Column
               Non-Null Count Dtype
 0
    Name
               6 non-null
                               object
    Domain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

```
3
     Location 6 non-null
                                category
4
                                int32
     Salary
               6 non-null
5
     Exp
               6 non-null
                                int32
dtypes: category(3), int32(3)
memory usage: 866.0 bytes
clean_data
    Name
               Domain Age
                             Location
                                       Salary
                                                Exp
    Mike Datascience
                        34
                               Mumbai
                                          5000
                                                  2
                            Bangalore
                                                  3
1
  Teddv
              Testina
                        45
                                         10000
                                                  4
2
   Umar Dataanalyst
                        50
                            Bangalore
                                         15000
3
                                                  4
           Analytics
                        50
                             Hyderbad
                                         20000
    Jane
4
  Uttam
           Statistics
                            Bangalore
                                         30000
                                                  5
                        67
5
                  NLP
                        55
                                Delhi
                                         60000
                                                 10
     Kim
clean_data.to_csv('clean_data.csv')
import os
os.getcwd()
'C:\\Users\\Admin'
clean_data
                             Location
    Name
               Domain
                       Age
                                       Salary
                                                Exp
0
    Mike Datascience
                        34
                               Mumbai
                                          5000
                                                  2
                                                  3
1
  Teddy
              Testing
                        45
                            Bangalore
                                         10000
2
                                         15000
                                                  4
    Umar
          Dataanalyst
                        50
                            Bangalore
3
                                                  4
    Jane
            Analytics
                        50
                             Hyderbad
                                         20000
                            Bangalore
4
                                                  5
   Uttam
           Statistics
                        67
                                         30000
5
     Kim
                  NLP
                        55
                                Delhi
                                         60000
                                                 10
```

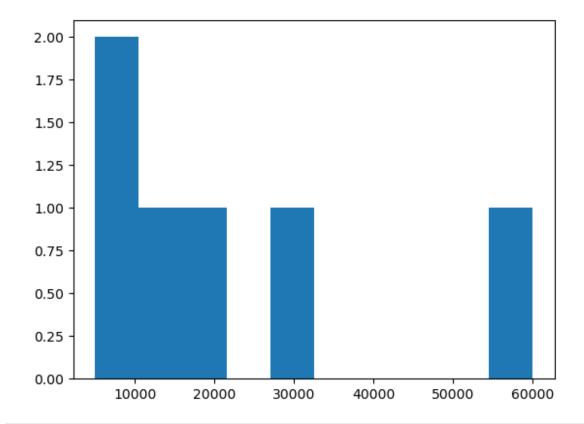
EDA TECHNIQUE LETS APPLY

```
import matplotlib.pyplot as plt
                                   #visualization
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
clean_data['Salary']
0
      5000
1
     10000
2
     15000
3
     20000
4
     30000
5
     60000
Name: Salary, dtype: int32
```

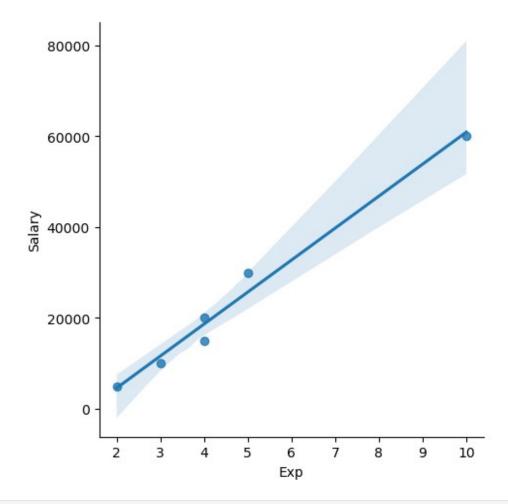
```
clean_data['Salary']
0     5000
1     10000
2     15000
3     20000
4     30000
5     60000
Name: Salary, dtype: int32
vis1=sns.displot(clean_data['Salary'])
```



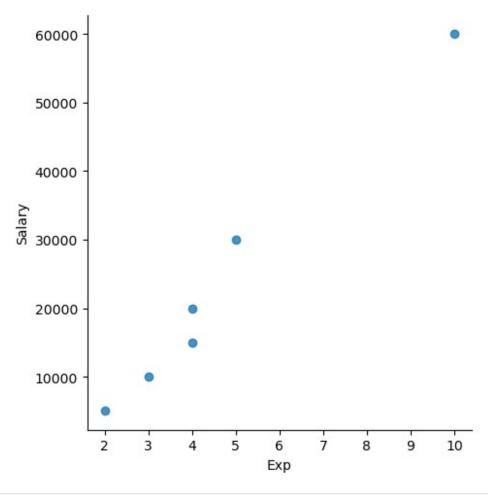
vis2=plt.hist(clean_data['Salary'])



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



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



cl	ean_dat	a[:]					
0 1 2 3 4 5	Name Mike Teddy Umar Jane Uttam Kim	Domain Datascience Testing Dataanalyst Analytics Statistics NLP	Age 34 45 50 50 67 55	Location Mumbai Bangalore Bangalore Hyderbad Bangalore Delhi	Salary 5000 10000 15000 20000 30000 60000	Exp 2 3 4 4 5 10	
cl	ean_dat	a[0:6:2]					
0 2 4	Name Mike Umar Uttam	Domain Datascience Dataanalyst Statistics	Age 34 50 67	Location Mumbai Bangalore Bangalore	Salary 5000 15000 30000	Exp 2 4 5	
cl	ean_dat	a[::- 1]					
5 4	Name Kim Uttam	Domain NLP Statistics	Age 55 67	Location Delhi Bangalore	Salary 60000 30000	Exp 10 5	

```
3
            Analytics
                         50
                              Hyderbad
                                          20000
                                                   4
    Jane
2
          Dataanalyst
                             Bangalore
                                                   4
    Umar
                         50
                                          15000
1
  Teddy
              Testing
                         45
                             Bangalore
                                          10000
                                                   3
                                                   2
    Mike Datascience
                         34
                                Mumbai
                                           5000
clean data.columns
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'],
dtype='object')
X iv=clean data[['Name','Domain','Age','Location','Exp']]
Χiν
    Name
               Domain
                        Age
                              Location
                                         Exp
    Mike
0
          Datascience
                         34
                                Mumbai
                                           2
1
  Teddy
              Testing
                         45
                             Bangalore
                                           3
2
                                           4
   Umar
                             Bangalore
          Dataanalyst
                         50
3
    Jane
            Analytics
                         50
                              Hyderbad
                                           4
4
  Uttam
           Statistics
                         67
                             Bangalore
                                           5
5
     Kim
                   NLP
                         55
                                 Delhi
                                          10
y dv=clean data[['Salary']]
y_dv
   Salary
     5000
0
    10000
1
2
    15000
3
    20000
4
    30000
5
    60000
emp
    Name
               Domain Age
                              Location Salary
                                                Exp
0
    Mike
          Datascience
                         34
                                Mumbai
                                          5000
                                                  2
                                                  3
1
  Teddy
              Testing
                         45
                             Bangalore
                                         10000
2
    Umar
          Dataanalyst
                        NaN
                                    NaN
                                         15000
                                                  4
3
                              Hyderbad
    Jane
            Analytics
                        NaN
                                         20000
                                                NaN
4
                                    NaN
                                                  5
  Uttam
           Statistics
                         67
                                         30000
5
     Kim
                   NLP
                         55
                                 Delhi
                                        60000
                                                 10
clean data
                        Age
                              Location
                                         Salary
    Name
               Domain
                                                 Exp
    Mike
                                           5000
          Datascience
                         34
                                Mumbai
                                                   2
                             Bangalore
                                          10000
                                                   3
1
  Teddy
              Testing
                         45
2
                                                   4
    Umar
          Dataanalyst
                         50
                             Bangalore
                                          15000
3
    Jane
            Analytics
                         50
                              Hyderbad
                                          20000
                                                   4
```

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

clean_data

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

imputation=pd.get_dummies(clean_data)

$\verb"imputation"$

			_		N 161		
	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy
Na	me Um	ar \					
0	- 34	5000	2	False	False	True	False
Fa	lse						
1	45	10000	3	False	False	False	True
Fa	lse						
2	50	15000	4	False	False	False	False
Tr	ue						
3	50	20000	4	True	False	False	False
Fa	lse						
4	67	30000	5	False	False	False	False
Fa	lse						
5	55	60000	10	False	True	False	False
Fa	lse						

		Jttam Doma		lytics	Domai	n_Dataan	alyst	
Doma 0 True	_ F	atascience alse	\	False			False	
1 Fal:	F	alse		False			False	
2 Fal:	F	alse		False			True	
3 Fal:	F	alse		True			False	
4 Fal:		True		False			False	
5 Fal	F	alse		False			False	
		n_NLP Doma Bangalore		tistics	Doma	in_Testi	ng	
0	F	alse		False		Fal	se	False
1	F	alse		False		Tr	ue	True
2	F	alse		False		Fal	se	True
3	F	alse		False		Fal	se	False
4	F	alse		True		Fal	se	True
5		True		False		Fals	se	False
0 1 2 3 4 5	Locati	ion_Delhi False False False False True	Locatio	Fa Fa Fa	rbad alse alse alse True alse	Location _.	_Mumbai True False False False False False	
cle	an_dat	ta						
2 3 4 5	Name Mike Teddy Umar Jane Uttam Kim		nce 34 ing 45 /st 50 ics 50	4 Mi 5 Banga 9 Banga 9 Hyda 7 Banga	ation umbai alore alore erbad alore Delhi	Salary 5000 10000 15000 20000 30000 60000	Exp 2 3 4 4 5 10	
imp	utatio	on						

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy
0	3 4	ar \ 5000	2	False	False	True	False
Fals	45	10000	3	False	False	False	True
Fals 2	50	15000	4	False	False	False	False
True	50	20000	4	True	False	False	False
Fals	67	30000	5	False	False	False	False
Fals 5 Fals	55	60000	10	False	True	False	False
		_Uttam Datascie		in_Analytics	Domain_	Dataanalyst	
0 True	_	False	ence	False		False	
1 Fals		False		False		False	
2 Fals		False		False		True	
3 Fals		False		True		False	
4 Fals		True		False		False	
5 Fals		False		False		False	
				in_Statistics	s Domain	_Testing	
0	CIO	n_Bangal False	Lore	False	9	False	False
1		False		False	9	True	True
2		False		False	9	False	True
3		False		False	9	False	False
4		False		True	9	False	True
5		True		False	9	False	False
0 1 2 3	₋oca [·]	Fal	lse lse lse	F	erbad Lo False False True	cation_Mumba Tru Fals Fals Fals	e e e

|--|