Exploratory Data Analysis

```
import warnings
In [1]:
         warnings.filterwarnings('ignore')
         import pandas as pd
In [2]:
         emp = pd.read_excel("Rawdata.xlsx")
In [3]:
In [4]:
         emp
Out[4]:
             Name
                          Domain
                                           Location
                                                       Salary
                                     Age
                                                                 Exp
         0
              Mike
                                  34 years
                                                      5^00#0
                                                                  2+
                     Datascience#$
                                            Mumbai
         1 Teddy^
                           Testing
                                          Bangalore
                                                    10%%000
                                                                  <3
                                    45' yr
            Uma#r
                   Dataanalyst^^#
                                                     1$5%000
                                     NaN
                                               NaN
                                                               4> yrs
                       Ana^^lytics
         3
              Jane
                                     NaN
                                          Hyderbad
                                                      2000^0
                                                                 NaN
            Uttam*
                          Statistics
                                                       30000-
                                                              5+ year
                                     67-yr
                                               NaN
               Kim
                             NLP
                                     55yr
                                               Delhi
                                                     6000^$0
                                                                 10+
In [5]:
         emp.shape
         (6, 6)
Out[5]:
         len(emp)
In [6]:
Out[6]:
In [7]:
         emp.columns
         Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[7]:
         len(emp.columns)
In [8]:
Out[8]:
In [9]:
         emp.info()
         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 6 entries, 0 to 5
         Data columns (total 6 columns):
          #
              Column
                        Non-Null Count Dtype
                                          object
          a
             Name
                         6 non-null
              Domain
                         6 non-null
                                          object
                         4 non-null
                                          object
              Age
              Location 4 non-null
                                          object
          3
              Salary
                         6 non-null
                                          object
          5
              Exp
                         5 non-null
                                          object
         dtypes: object(6)
         memory usage: 420.0+ bytes
```

```
emp
In [10]:
Out[10]:
              Name
                            Domain
                                              Location
                                                          Salary
                                        Age
                                                                    Exp
               Mike
                       Datascience#$
                                              Mumbai
                                                         5^00#0
                                                                     2+
                                    34 years
             Teddy^
                            Testing
                                       45' yr
                                             Bangalore
                                                       10%%000
                                                                     <3
              Uma#r
                     Dataanalyst^^#
                                       NaN
                                                 NaN
                                                        1$5%000
          2
                                                                  4> yrs
          3
                         Ana^^lytics
                                       NaN Hyderbad
                                                         2000^0
                                                                   NaN
                Jane
                                                                 5+ year
             Uttam*
                           Statistics
                                       67-yr
                                                 NaN
                                                         30000-
                               NLP
                                                        6000^$0
                                                                    10+
                Kim
                                        55yr
                                                 Delhi
          emp['Name']
In [11]:
                  Mike
Out[11]:
          1
                Teddy^
          2
                Uma#r
          3
                  Jane
          4
                Uttam*
          5
                   Kim
          Name: Name, dtype: object
          emp['Domain']
In [12]:
                Datascience#$
Out[12]:
          1
                       Testing
          2
                Dataanalyst^^#
          3
                   Ana^^lytics
          4
                    Statistics
                            NLP
          Name: Domain, dtype: object
In [13]:
          emp['Age']
                34 years
Out[13]:
          1
                  45' yr
          2
                     NaN
          3
                     NaN
          4
                   67-yr
          5
                    55yr
          Name: Age, dtype: object
          emp['Location']
In [14]:
                   Mumbai
Out[14]:
          1
                Bangalore
          2
                      NaN
          3
                Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
          emp['Salary']
In [15]:
                 5^00#0
Out[15]:
          1
                10%%000
          2
                1$5%000
          3
                 2000^0
          4
                 30000-
                6000^$0
          5
          Name: Salary, dtype: object
```

```
emp['Exp']
In [16]:
Out[16]:
                      <3
          2
                 4> yrs
          3
                     NaN
                5+ year
                     10+
          Name: Exp, dtype: object
           emp[['Name','Domain']]
In [17]:
Out[17]:
               Name
                            Domain
           0
                       Datascience#$
                Mike
             Teddy^
                             Testing
              Uma#r Dataanalyst^^#
           3
                Jane
                          Ana^^lytics
              Uttam*
                            Statistics
           5
                                NLP
                 Kim
           emp[['Name','Domain','Age']]
In [18]:
Out[18]:
               Name
                            Domain
                                         Age
                Mike
                       Datascience#$ 34 years
             Teddy^
                             Testing
                                        45' yr
                      Dataanalyst^^#
              Uma#r
                                         NaN
                          Ana^^lytics
                                         NaN
                Jane
           4
              Uttam*
                            Statistics
                                        67-yr
           5
                                NLP
                 Kim
                                         55yr
           emp[['Name','Domain','Age','Location','Salary','Exp']]
In [19]:
Out[19]:
               Name
                             Domain
                                         Age
                                               Location
                                                            Salary
                                                                       Ехр
           0
                Mike
                       Datascience#$
                                     34 years
                                                           5^00#0
                                                                        2+
                                                Mumbai
           1 Teddy^
                             Testing
                                        45' yr
                                              Bangalore
                                                         10%%000
                                                                        <3
                      Dataanalyst^^#
              Uma#r
                                                          1$5%000
                                        NaN
                                                   NaN
                                                                    4> yrs
           3
                          Ana^^lytics
                                                           2000^0
                Jane
                                        NaN
                                               Hyderbad
                                                                      NaN
                                        67-yr
              Uttam*
                            Statistics
                                                   NaN
                                                           30000-
                                                                   5+ year
           5
                 Kim
                                NLP
                                         55yr
                                                   Delhi
                                                          6000^$0
                                                                       10+
```

Data Cleansing

```
In [20]: emp['Name']
```

```
Mike
          0
Out[20]:
          1
               Teddy^
          2
                Uma#r
          3
                 Jane
          4
               Uttam*
          5
                  Kim
          Name: Name, dtype: object
          emp['Name'] = emp['Name'].str.replace(r'\W','')
In [21]:
In [22]:
          emp['Name']
                Mike
Out[22]:
          1
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
          5
                 Kim
          Name: Name, dtype: object
          emp['Domain'] = emp['Domain'].str.replace(r'\W','')
In [23]:
          emp['Domain']
In [24]:
               Datascience
Out[24]:
          1
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
          5
                       NLP
          Name: Domain, dtype: object
          emp['Age'] = emp['Age'].str.replace(r'\W','')
In [25]:
          emp['Age']
In [26]:
               34years
Out[26]:
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
          5
                  55yr
          Name: Age, dtype: object
          emp['Age'] = emp['Age'].str.extract('(\d+)')
In [27]:
          emp['Age']
In [28]:
                34
Out[28]:
                45
          2
               NaN
          3
               NaN
          4
                67
          5
                55
          Name: Age, dtype: object
In [29]:
          emp
```

Domain

Age

Location

Salary

Ехр

Name

Out[29]:

```
0
              Mike
                    Datascience
                                                5^00#0
                                                            2+
                                 34
                                      Mumbai
          1 Teddy
                        Testing
                                 45 Bangalore 10%%000
                                                            <3
          2
                    Dataanalyst NaN
                                               1$5%000
              Umar
                                         NaN
                                                         4> yrs
                                                2000^0
          3
              Jane
                      Analytics NaN
                                     Hyderbad
                                                           NaN
          4
             Uttam
                       Statistics
                                 67
                                         NaN
                                                 30000-
                                                        5+ year
          5
                          NLP
                                 55
                                         Delhi
                                               6000^$0
               Kim
                                                           10+
          emp['Location'] = emp['Location'].str.replace(r'\W','')
In [30]:
          emp['Location']
In [31]:
                   Mumbai
Out[31]:
          1
               Bangalore
          2
                      NaN
          3
                Hyderbad
          4
                      NaN
          5
                    Delhi
          Name: Location, dtype: object
In [32]: emp['Salary'] = emp['Salary'].str.replace(r'\W','')
          emp['Salary']
In [33]:
                5000
Out[33]:
          1
               10000
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: object
          emp['Exp'] = emp['Exp'].str.extract('(\d+)')
In [34]:
In [35]:
          emp['Exp']
                  2
Out[35]:
          1
                 3
          2
                 4
          3
               NaN
          4
                 5
          5
                10
          Name: Exp, dtype: object
In [36]:
          emp
```

Domain Age Location Salary Exp

Out[36]:

Name

	0	Mike	Datascience	34	Mumbai	5000	2		
	1	Teddy	Testing	45	Bangalore	10000	3		
	2	Umar	Dataanalyst	NaN	NaN	15000	4		
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN		
	4	Uttam	Statistics	67	NaN	30000	5		
	5	Kim	NLP	55	Delhi	60000	10		
In [37]:	cl	.ean_da	ta = emp.co	ppy()					
In [38]:	clean_data								
L 1		.can_aa	Ld						
Out[38]:		Name	Domain	Age	Location	Salary	Ехр		
	0	Name		Age 34	Location Mumbai	Salary 5000	Exp 2		
		Name Mike	Domain	34	Mumbai				
	0	Name Mike Teddy	Domain Datascience	34 45	Mumbai	5000	2		
	0	Name Mike Teddy	Domain Datascience Testing Dataanalyst	34 45 NaN	Mumbai Bangalore	5000	2 3 4		
	0 1 2	Name Mike Teddy Umar	Domain Datascience Testing Dataanalyst	34 45 NaN	Mumbai Bangalore NaN	5000 10000 15000	2 3 4		
	0 1 2	Mame Mike Teddy Umar Jane	Domain Datascience Testing Dataanalyst Analytics	34 45 NaN NaN	Mumbai Bangalore NaN Hyderbad	5000 10000 15000 20000	2 3 4 NaN		

Missing Value Treatment

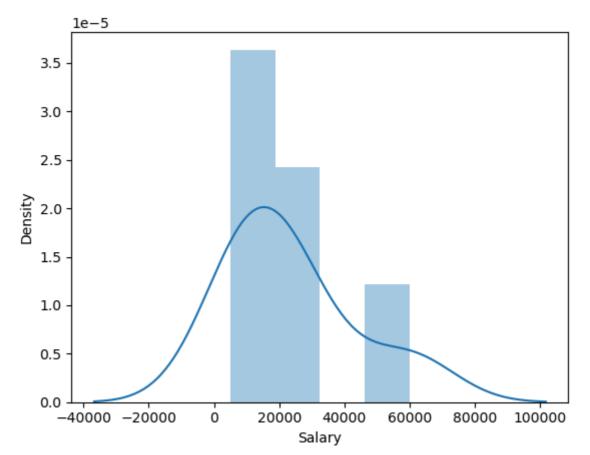
In [39]:	clean_data									
Out[39]:	Name Domain Age Location Salary Exp									
	0	Mike	Datascience	34	Mumbai	5000	2			
	1	Teddy	Testing	45	Bangalore	10000	3			
	2	Umar	Dataanalyst	NaN	NaN	15000	4			
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN			
	4	Uttam	Statistics	67	NaN	30000	5			
	5	Kim	NLP	55	Delhi	60000	10			
In [40]:	<pre>clean_data.info()</pre>									

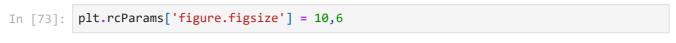
```
<class 'pandas.core.frame.DataFrame'>
                               RangeIndex: 6 entries, 0 to 5
                               Data columns (total 6 columns):
                                                                              Non-Null Count Dtype
                                               Column
                                ---
                                              -----
                                                                                -----
                                  0
                                               Name
                                                                                6 non-null
                                                                                                                                     object
                                              Domain
                                                                               6 non-null
                                                                                                                                     object
                                  1
                                  2
                                                                               4 non-null
                                                                                                                                     object
                                               Location 4 non-null
                                                                                                                                     object
                                  4
                                                                                6 non-null
                                                                                                                                     object
                                               Salary
                                  5
                                               Exp
                                                                                5 non-null
                                                                                                                                     object
                               dtypes: object(6)
                              memory usage: 420.0+ bytes
In [41]:
                               import numpy as np
In [42]:
                               clean_data
Out[42]:
                                        Name
                                                                      Domain
                                                                                                                  Location Salary
                                                                                                Age
                                                                                                                                                                     Exp
                                           Mike Datascience
                                                                                                     34
                                                                                                                     Mumbai
                                                                                                                                                  5000
                                                                                                                                                                            2
                                                                                                                                                10000
                                                                                                                                                                            3
                                         Teddy
                                                                         Testing
                                                                                                    45
                                                                                                                Bangalore
                               2
                                          Umar
                                                             Dataanalyst
                                                                                               NaN
                                                                                                                              NaN
                                                                                                                                                15000
                                                                                                                                                                            4
                               3
                                            Jane
                                                                     Analytics
                                                                                               NaN
                                                                                                                 Hyderbad
                                                                                                                                                20000 NaN
                                        Uttam
                                                                     Statistics
                                                                                                     67
                                                                                                                              NaN
                                                                                                                                                30000
                                                                                                                                                                            5
                                              Kim
                                                                                NLP
                                                                                                     55
                                                                                                                            Delhi
                                                                                                                                                60000
                                                                                                                                                                          10
In [43]:
                               clean_data['Age']
                                                   34
Out[43]:
                                                  45
                               2
                                               NaN
                               3
                                               NaN
                               4
                                                   67
                               5
                                                   55
                              Name: Age, dtype: object
In [44]: clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.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['A
                               clean_data['Age']
In [45]:
                                                         34
Out[45]:
                                                         45
                               1
                               2
                                               50.25
                               3
                                               50.25
                               4
                                                         67
                               5
                                                         55
                               Name: Age, dtype: object
In [46]:
```

```
Out[46]:
                                     Name
                                                                Domain
                                                                                       Age
                                                                                                        Location Salary
                                                                                                                                                       Exp
                             0
                                        Mike
                                                        Datascience
                                                                                            34
                                                                                                          Mumbai
                                                                                                                                     5000
                                                                                                                                                             2
                                                                                                      Bangalore
                             1
                                     Teddy
                                                                   Testing
                                                                                            45
                                                                                                                                10000
                                                                                                                                                            3
                                                                                                                                  15000
                                                                                                                                                             4
                             2
                                      Umar
                                                        Dataanalyst
                                                                                       NaN
                                                                                                                   NaN
                             3
                                         Jane
                                                              Analytics
                                                                                       NaN
                                                                                                       Hyderbad
                                                                                                                                  20000
                                                                                                                                                    NaN
                                                                                                                                                            5
                             4
                                     Uttam
                                                               Statistics
                                                                                            67
                                                                                                                   NaN
                                                                                                                                   30000
                             5
                                                                         NLP
                                                                                                                                                          10
                                           Kim
                                                                                            55
                                                                                                                 Delhi
                                                                                                                                   60000
                             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.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.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fil
In [48]:
In [49]:
                             clean_data['Exp']
                                                 2
Out[49]:
                             1
                                                 3
                             2
                                                 4
                             3
                                           4.8
                             4
                                                 5
                             5
                                              10
                            Name: Exp, dtype: object
                             clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode(
In [50]:
                             clean_data['Location']
In [51]:
                                                    Mumbai
Out[51]:
                             1
                                           Bangalore
                             2
                                           Bangalore
                             3
                                              Hyderbad
                             4
                                           Bangalore
                                                       Delhi
                             5
                            Name: Location, dtype: object
                             clean data
In [52]:
Out[52]:
                                     Name
                                                                Domain
                                                                                                          Location
                                                                                                                                   Salary
                                                                                          Age
                                                                                                                                                       Exp
                             0
                                       Mike
                                                       Datascience
                                                                                              34
                                                                                                            Mumbai
                                                                                                                                       5000
                                                                                                                                                             2
                                                                                                        Bangalore
                                                                                                                                     10000
                                     Teddy
                                                                  Testing
                                                                                             45
                                                                                                                                                             3
                             2
                                                                                       50.25
                                      Umar
                                                        Dataanalyst
                                                                                                        Bangalore
                                                                                                                                     15000
                                                                                                                                                             4
                             3
                                         Jane
                                                               Analytics 50.25
                                                                                                         Hyderbad
                                                                                                                                    20000
                                                                                                                                                         4.8
                                                                                                                                                             5
                             4
                                     Uttam
                                                               Statistics
                                                                                              67
                                                                                                         Bangalore
                                                                                                                                     30000
                             5
                                                                         NLP
                                                                                                                                    60000
                                           Kim
                                                                                              55
                                                                                                                   Delhi
                                                                                                                                                          10
                             clean_data['Age'] = clean_data['Age'].astype(int)
In [53]:
                             clean_data['Salary'] = clean_data['Salary'].astype(int)
In [54]:
                             clean_data['Exp'] = clean_data['Exp'].astype(int)
In [55]:
                             clean_data
In [56]:
```

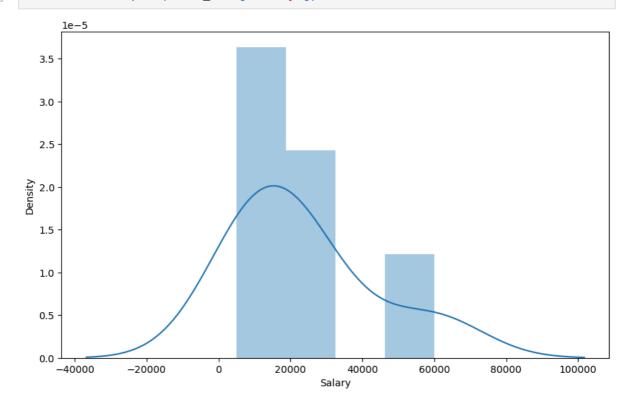
```
Name
                                  Location Salary Exp
Out[56]:
                     Domain Age
             Mike Datascience
                                            5000
         0
                              34
                                   Mumbai
         1 Teddy
                      Testing
                              45 Bangalore
                                          10000
                                                   3
                                 Bangalore
         2
            Umar
                  Dataanalyst
                              50
                                           15000
                                                   4
         3
             Jane
                    Analytics
                              50
                                  Hyderbad
                                           20000
                                                   4
         4 Uttam
                     Statistics
                              67
                                 Bangalore
                                           30000
                                                   5
                        NLP
                                           60000
         5
              Kim
                              55
                                     Delhi
                                                  10
        clean_data.info()
In [57]:
         <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
                      6 non-null
                                        int32
          2
              Age
          3 Location 6 non-null
                                        object
          4
              Salary 6 non-null
                                        int32
          5
                        6 non-null
                                        int32
              Exp
         dtypes: int32(3), object(3)
         memory usage: 348.0+ bytes
         clean_data['Name'] = clean_data['Name'].astype('category')
In [58]:
         clean data['Domain'] = clean data['Domain'].astype('category')
         clean_data['Location'] = clean_data['Location'].astype('category')
In [59]:
         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 Name
                        6 non-null
                                        category
              Domain
                       6 non-null
          1
                                        category
          2
              Age
                        6 non-null
                                        int32
              Location 6 non-null
          3
                                        category
          4
              Salary
                        6 non-null
                                        int32
          5
                        6 non-null
              Exp
                                        int32
         dtypes: category(3), int32(3)
         memory usage: 866.0 bytes
In [60]:
         clean_data
```

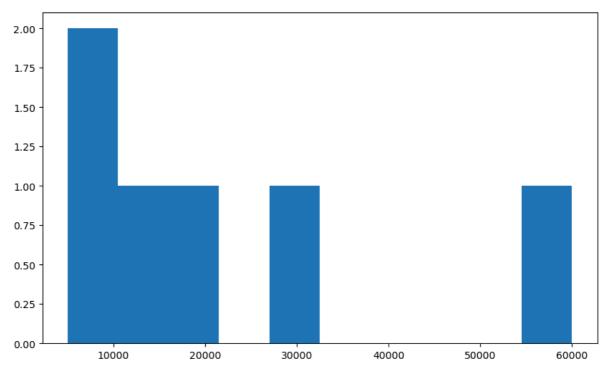
```
Out[60]:
             Name
                                      Location Salary Exp
                       Domain
                                Age
               Mike
                    Datascience
                                                 5000
                                                         2
          0
                                  34
                                       Mumbai
              Teddy
                         Testing
                                  45
                                      Bangalore
                                                10000
                                                         3
                                                15000
                                                         4
          2
              Umar
                     Dataanalyst
                                  50
                                      Bangalore
          3
               Jane
                       Analytics
                                  50
                                      Hyderbad
                                                20000
                                                         4
                                                         5
          4
             Uttam
                       Statistics
                                  67
                                      Bangalore
                                                30000
          5
                           NLP
                                  55
                                                60000
                                                         10
                Kim
                                          Delhi
          clean_data.to_csv('clean_data.csv')
In [61]:
In [62]:
           import os
           os.getcwd()
           'C:\\Users\\JANHAVI\\NIT'
Out[62]:
In [68]:
           import matplotlib.pyplot as plt # visualization
           import seaborn as sns # Advanced visualization
          clean_data.columns
In [69]:
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[69]:
           clean_data
In [70]:
Out[70]:
             Name
                       Domain
                                      Location Salary Exp
                                Age
               Mike
                    Datascience
                                                         2
          0
                                  34
                                       Mumbai
                                                 5000
              Teddy
                                  45
                                      Bangalore
                                                10000
                                                         3
                         Testing
          2
              Umar
                     Dataanalyst
                                      Bangalore
                                                15000
                                                         4
          3
               Jane
                       Analytics
                                  50
                                      Hyderbad
                                                20000
                                                         4
                                                30000
                                                         5
          4
             Uttam
                       Statistics
                                  67
                                      Bangalore
          5
                           NLP
                                  55
                                          Delhi
                                                60000
                                                         10
                Kim
          clean_data['Salary']
In [71]:
                 5000
Out[71]:
          1
                10000
          2
                15000
          3
                20000
          4
                30000
                60000
          Name: Salary, dtype: int32
          vis1 = sns.distplot(clean_data['Salary'])
In [72]:
```

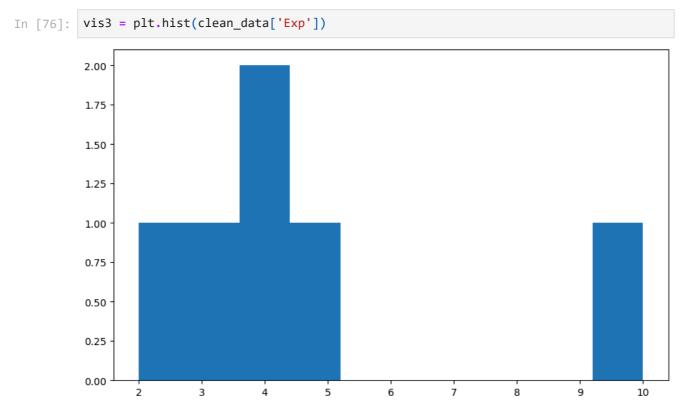




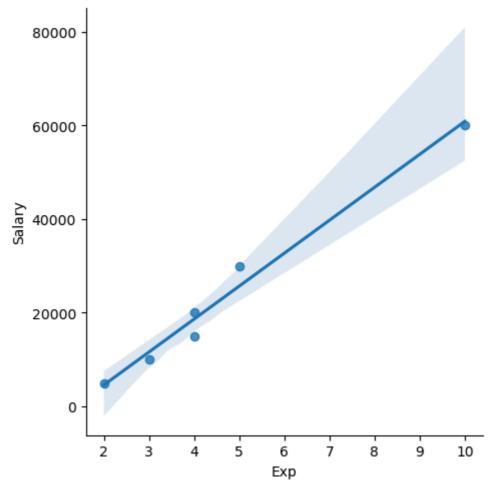
In [74]: vis1 = sns.distplot(clean_data['Salary'])

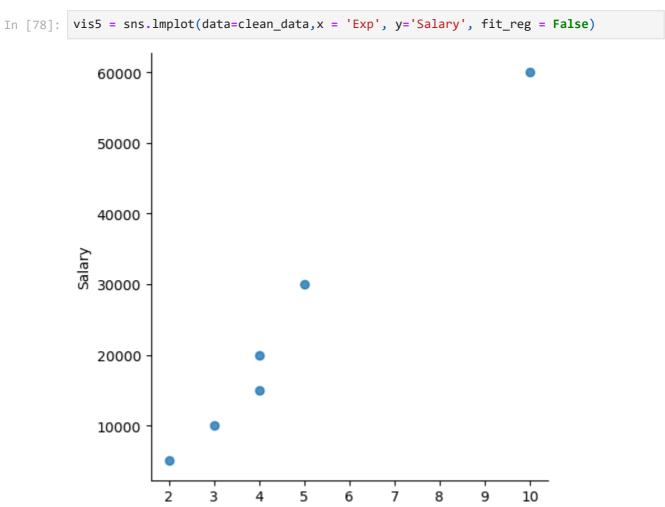






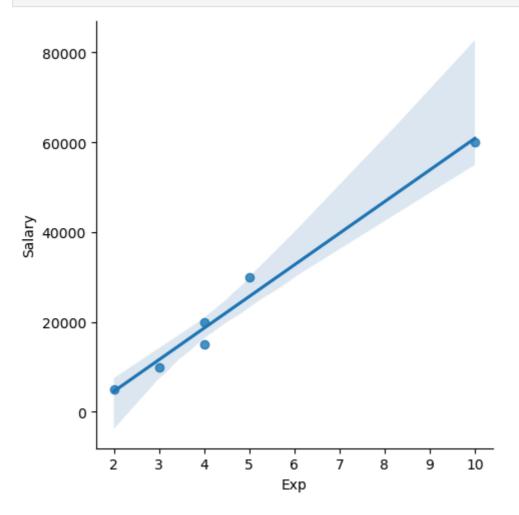
In [77]: vis4 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary')





Exp

In [79]: vis6 = sns.lmplot(data=clean_data,x = 'Exp', y='Salary', fit_reg = True)



In [80]: clean_data

Out[80]:		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 [81]: clean_data[:]

Out[81]:		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

```
clean_data[:2]
In [82]:
              Name
                                        Location Salary
Out[82]:
                        Domain Age
                                                         Exp
               Mike Datascience
                                         Mumbai
                                                   5000
                                                            2
                                   34
                                                            3
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                  10000
           clean_data[2:]
In [83]:
Out[83]:
                                                 Salary Exp
              Name
                        Domain
                                 Age
                                        Location
              Umar Dataanalyst
                                   50
                                       Bangalore
                                                  15000
                                                           4
           3
               Jane
                        Analytics
                                   50
                                       Hyderbad
                                                  20000
                                                           4
                                                           5
           4
                                                  30000
              Uttam
                        Statistics
                                   67
                                       Bangalore
           5
                Kim
                            NLP
                                   55
                                           Delhi
                                                  60000
                                                          10
           clean_data[:]
In [84]:
Out[84]:
              Name
                        Domain
                                  Age
                                        Location
                                                 Salary
                                                         Ехр
           0
               Mike Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                  10000
                                                            3
           2
              Umar
                     Dataanalyst
                                   50
                                       Bangalore
                                                  15000
                                                            4
           3
               Jane
                        Analytics
                                       Hyderbad
                                                  20000
           4
                        Statistics
                                                  30000
                                                            5
              Uttam
                                   67
                                       Bangalore
           5
                            NLP
                                   55
                                                  60000
                Kim
                                           Delhi
                                                           10
           clean_data[0:1]
In [85]:
Out[85]:
                                       Location
              Name
                        Domain Age
                                                 Salary
                                                        Exp
           0
               Mike Datascience
                                   34
                                        Mumbai
                                                  5000
                                                           2
           clean_data
In [88]:
Out[88]:
              Name
                        Domain
                                        Location
                                 Age
                                                 Salary
               Mike Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                  10000
                                                            3
           2
              Umar
                     Dataanalyst
                                   50
                                       Bangalore
                                                  15000
                                                            4
           3
               Jane
                        Analytics
                                   50
                                       Hyderbad
                                                  20000
           4
                                       Bangalore
                                                  30000
                                                            5
              Uttam
                        Statistics
                                   67
           5
                            NLP
                                   55
                                                  60000
                Kim
                                           Delhi
                                                           10
           x_iv = clean_data.drop(['Salary'],axis=1)
In [89]:
In [90]:
           x_iv
```

```
Location Exp
Out[90]:
              Name
                        Domain Age
           0
               Mike
                     Datascience
                                        Mumbai
                                  34
              Teddy
                         Testing
                                  45
                                      Bangalore
           2
                     Dataanalyst
                                  50
                                      Bangalore
                                                   4
              Umar
                                  50
           3
                                       Hyderbad
                                                   4
               Jane
                       Analytics
                                                   5
           4
              Uttam
                        Statistics
                                  67
                                      Bangalore
           5
                Kim
                            NLP
                                  55
                                           Delhi
                                                  10
In [91]:
          y_dv = clean_data.drop(['Name', 'Domain', 'Age', 'Location', 'Exp'], axis=1)
In [92]:
           y_dv
Out[92]:
              Salary
               5000
              10000
           2
              15000
              20000
              30000
              60000
           clean_data
In [93]:
Out[93]:
                                       Location Salary Exp
              Name
                        Domain
                                 Age
               Mike
                     Datascience
                                  34
                                        Mumbai
                                                  5000
                                                           2
              Teddy
                         Testing
                                  45
                                      Bangalore
                                                 10000
                                                           3
           2
              Umar
                     Dataanalyst
                                  50
                                      Bangalore
                                                 15000
                                                           4
               Jane
                       Analytics
                                  50
                                       Hyderbad
                                                 20000
                                                           4
           4
              Uttam
                                                 30000
                                                           5
                        Statistics
                                  67
                                      Bangalore
                            NLP
                                                 60000
                Kim
                                  55
                                           Delhi
                                                          10
           imputation = pd.get_dummies(clean_data)
In [96]:
In [97]:
           imputation
```

Out[97]:		Age	Salary	Ехр	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttar
	0	34	5000	2	0	0	1	0	0	
	1	45	10000	3	0	0	0	1	0	
	2	50	15000	4	0	0	0	0	1	
	3	50	20000	4	1	0	0	0	0	
	4	67	30000	5	0	0	0	0	0	
	5	55	60000	10	0	1	0	0	0	
-		-	-	-						•