```
In [128...
           import pandas as pd
           pd.__version__
In [129...
Out[129...
            '2.2.2'
In [130...
           dataFile = pd.read_excel(r'C:\Python_Learning\Week5_EDA\20thFeb\RawData.xlsx')
In [131...
           dataFile
Out[131...
               Name
                              Domain
                                           Age
                                                 Location
                                                              Salary
                                                                         Exp
           0
                 Mike
                         Datascience#$
                                       34 years
                                                  Mumbai
                                                             5^00#0
                                                                          2+
           1 Teddy^
                                                Bangalore
                               Testing
                                          45' yr
                                                           10%%000
                                                                          <3
               Uma#r
                       Dataanalyst^^#
                                                     NaN
                                                            1$5%000
                                                                       4> yrs
           2
                                          NaN
                                               Hyderbad
                           Ana^^lytics
           3
                 Jane
                                          NaN
                                                             2000^0
                                                                        NaN
              Uttam*
                             Statistics
                                          67-yr
                                                     NaN
                                                              30000-
                                                                     5+ year
                  Kim
                                  NLP
                                           55yr
                                                     Delhi
                                                            6000^$0
                                                                         10+
In [132...
           id(dataFile)
Out[132...
           1455084775776
In [133...
           dataFile.shape
Out[133...
            (6, 6)
In [134...
           len(dataFile)
Out[134...
           6
In [135...
           dataFile.columns
Out[135...
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [136...
           dataFile.describe()
Out[136...
                    Name
                                 Domain
                                             Age Location
                                                              Salary Exp
                        6
                                       6
                                                4
                                                                        5
            count
                                                          4
                                                                   6
           unique
                        6
                                                4
                                                          4
                                                                        5
                     Mike Datascience#$
                                          34 years
                                                    Mumbai 5^00#0
                                                                       2+
               top
                                                1
              freq
                        1
                                                                        1
```

```
In [137... dataFile.head()
```

Out[137...

	Name	Domain	Age	Location	Salary	Ехр
0	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year

In [138...

dataFile.tail()

Out[138...

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

In [139...

dataFile.info()

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

Ducu	COTAMILIS (cocar o coramiis,	•
#	Column	Non-Null Count	Dtype
0	Name	6 non-null	object
1	Domain	6 non-null	object
2	Age	4 non-null	object
3	Location	4 non-null	object
4	Salary	6 non-null	object
5	Exp	5 non-null	object

dtypes: object(6)

memory usage: 420.0+ bytes

In [140...

dataFile.isnull()

```
Out[140...
               Name Domain Age Location Salary
                                                           Exp
            0
                False
                          False
                                 False
                                           False
                                                   False False
            1
                False
                          False
                                 False
                                           False
                                                   False False
            2
                False
                          False
                                 True
                                            True
                                                   False False
            3
                False
                          False
                                 True
                                           False
                                                   False
                                                         True
                False
                          False False
                                            True
                                                   False False
                False
                          False False
                                           False
                                                    False False
In [141...
            dataFile.isnull().sum()
Out[141...
            Name
            Domain
                          0
            Age
                          2
            Location
                          2
            Salary
                          0
            Exp
                          1
            dtype: int64
In [142...
            dataFile.isna()
Out[142...
                                 Age Location Salary
               Name Domain
                                                           Ехр
                False
                                 False
                                           False
                                                   False
                                                          False
            0
                          False
                False
                                 False
            1
                          False
                                           False
                                                   False
                                                          False
            2
                False
                          False
                                 True
                                            True
                                                   False False
            3
                False
                          False
                                 True
                                           False
                                                   False
                                                         True
            4
                False
                                False
                                                   False False
                          False
                                            True
                False
                          False False
                                           False
                                                    False False
            dataFile['Name']
In [143...
Out[143...
            0
                    Mike
            1
                  Teddy^
            2
                   Uma#r
            3
                    Jane
            4
                  Uttam*
                     Kim
            5
            Name: Name, dtype: object
            dataFile['Name'] = dataFile['Name'].str.replace(r'[^a-zA-Z]', '', regex=True)
In [144...
In [145...
            dataFile['Name']
```

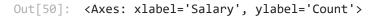
```
Out[145...
                 Mike
           1
                Teddy
           2
                 Umar
           3
                  Jane
                Uttam
           4
                   Kim
           Name: Name, dtype: object
           dataFile['Domain'] = dataFile['Domain'].str.replace(r'\W', '', regex=True)
In [146...
           dataFile
In [147...
Out[147...
              Name
                         Domain
                                           Location
                                     Age
                                                        Salary
                                                                   Exp
               Mike Datascience 34 years
                                                       5^00#0
                                                                    2+
           0
                                            Mumbai
              Teddy
                          Testing
                                    45' yr
                                           Bangalore
                                                     10%%000
                                                                    <3
           2
               Umar
                      Dataanalyst
                                     NaN
                                               NaN
                                                      1$5%000
                                                                 4> yrs
           3
                Jane
                        Analytics
                                     NaN Hyderbad
                                                       2000^0
                                                                  NaN
           4
              Uttam
                        Statistics
                                    67-yr
                                               NaN
                                                        30000-
                                                                5+ year
           5
                Kim
                            NLP
                                     55yr
                                               Delhi
                                                      6000^$0
                                                                   10+
In [148...
           dataFile['Age'] = dataFile['Age'].str.replace(r'[^0-9]', '', regex=True)
In [149...
           dataFile
Out[149...
              Name
                         Domain Age
                                        Location
                                                     Salary
                                                                Exp
           0
               Mike Datascience
                                   34
                                         Mumbai
                                                    5^00#0
                                                                 2+
              Teddy
                          Testing
                                   45
                                       Bangalore
                                                  10%%000
                                                                 <3
                                                              4> yrs
           2
               Umar
                      Dataanalyst NaN
                                            NaN
                                                   1$5%000
           3
                Jane
                        Analytics NaN
                                        Hyderbad
                                                    2000^0
                                                               NaN
                                                            5+ year
           4 Uttam
                        Statistics
                                   67
                                            NaN
                                                    30000-
           5
                            NLP
                Kim
                                   55
                                            Delhi
                                                   6000^$0
                                                                10+
           dataFile['Age'] = dataFile['Age'].str.extract(r'(\d+)')
In [150...
           dataFile['Location'] = dataFile['Location'].str.replace(r'\W', '', regex=True)
In [151...
           dataFile['Salary'] = dataFile['Salary'].str.replace(r'\W', '', regex=True)
           dataFile['Exp'] = dataFile['Exp'].str.extract(r'(\d+)')
In [152...
In [153...
           dataFile
```

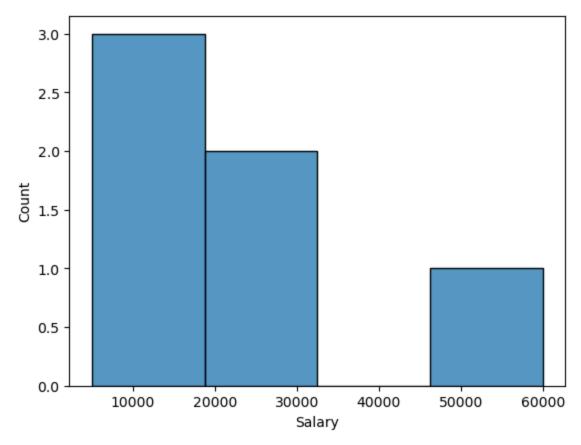
```
Out[153...
              Name
                         Domain Age
                                         Location Salary
                                                           Ехр
           0
               Mike Datascience
                                    34
                                         Mumbai
                                                    5000
                                                             2
              Teddy
                                    45
                                        Bangalore
                                                   10000
                                                             3
                          Testing
           2
               Umar
                      Dataanalyst NaN
                                             NaN
                                                   15000
                                                             4
                        Analytics NaN
                                        Hyderbad
                                                   20000 NaN
           3
                Jane
              Uttam
                        Statistics
                                             NaN
                                                   30000
                                                             5
                                    67
                            NLP
                                    55
                                                   60000
                                                            10
                Kim
                                            Delhi
           clean_data = dataFile.copy()
In [154...
In [155...
           clean_data
Out[155...
              Name
                                         Location Salary
                         Domain
                                  Age
                                                           Exp
           0
               Mike Datascience
                                    34
                                         Mumbai
                                                    5000
                                                             2
              Teddy
                          Testing
                                        Bangalore
                                                   10000
                                                             3
                                    45
                      Dataanalyst NaN
                                                             4
           2
               Umar
                                             NaN
                                                   15000
           3
                        Analytics NaN
                                        Hyderbad
                                                   20000
                                                          NaN
                Jane
           4
              Uttam
                        Statistics
                                                   30000
                                                             5
                                    67
                                             NaN
                Kim
                            NLP
                                    55
                                            Delhi
                                                   60000
                                                            10
In [156...
           # EDA TECHNIQUES
           clean_data.isnull().sum()
Out[156...
           Name
           Domain
                        2
           Age
                        2
           Location
           Salary
                        0
           Exp
                        1
           dtype: int64
In [125...
           clean_data['Age'].mean()
Out[125...
           50.16666666666664
 In [32]:
           import numpy as np
           # FILL MISSING VALS
           clean_data['Age'] = pd.to_numeric(clean_data['Age'])
In [159...
           clean_data['Age'] = clean_data['Age'].fillna(clean_data['Age'].mean())
In [160...
           clean_data['Age']
```

```
Out[160...
                                                     34.00
                                   0
                                    1
                                                    45.00
                                    2
                                                     50.25
                                    3
                                                     50.25
                                    4
                                                     67.00
                                                     55.00
                                    5
                                    Name: Age, dtype: float64
                                   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.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']).fillna(np.to_numeric(clean_data['Exp'])).fillna(np.to_numer
   In [35]:
   In [36]:
                                   clean_data
  Out[36]:
                                             Name
                                                                              Domain
                                                                                                              Age
                                                                                                                                  Location Salary Exp
                                   0
                                                 Mike Datascience
                                                                                                                   34
                                                                                                                                     Mumbai
                                                                                                                                                                      5000
                                                                                                                                                                                                2
                                                                                                                                Bangalore
                                                                                                                                                                   10000
                                              Teddy
                                                                                  Testing
                                                                                                                   45
                                                                                                                                                                                                3
                                   2
                                                Umar
                                                                     Dataanalyst 50.25
                                                                                                                                               NaN
                                                                                                                                                                   15000
                                                                                                                                                                                                4
                                                                                                                                                                  20000
                                   3
                                                  Jane
                                                                             Analytics 50.25
                                                                                                                                Hyderbad
                                                                                                                                                                                            4.8
                                                                                                                                                                  30000
                                                                                                                                                                                                5
                                   4
                                            Uttam
                                                                             Statistics
                                                                                                                   67
                                                                                                                                               NaN
                                    5
                                                    Kim
                                                                                          NLP
                                                                                                                   55
                                                                                                                                             Delhi
                                                                                                                                                                  60000
                                                                                                                                                                                             10
   In [37]:
                                  clean_data['Location'].isnull().sum()
  Out[37]: 2
   In [38]:
                                   clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode(
   In [39]:
                                   clean_data
   Out[39]:
                                             Name
                                                                              Domain
                                                                                                              Age
                                                                                                                                  Location Salary Exp
                                   0
                                                 Mike Datascience
                                                                                                                   34
                                                                                                                                     Mumbai
                                                                                                                                                                      5000
                                                                                                                                                                                                2
                                              Teddy
                                                                                                                                Bangalore
                                                                                                                                                                   10000
                                                                                                                                                                                                3
                                                                                  Testing
                                                                                                                   45
                                   2
                                                Umar
                                                                     Dataanalyst 50.25
                                                                                                                                Bangalore
                                                                                                                                                                   15000
                                                                                                                                                                                                4
                                   3
                                                                                                        50.25
                                                                                                                                Hyderbad
                                                                                                                                                                  20000
                                                  Jane
                                                                             Analytics
                                                                                                                                                                                            4.8
                                                                                                                                Bangalore
                                                                                                                                                                   30000
                                                                                                                                                                                                5
                                   4
                                            Uttam
                                                                             Statistics
                                                                                                                   67
                                                                                          NLP
                                                                                                                                                                  60000
                                    5
                                                    Kim
                                                                                                                   55
                                                                                                                                             Delhi
                                                                                                                                                                                             10
   In [40]:
                                  clean_data.info()
```

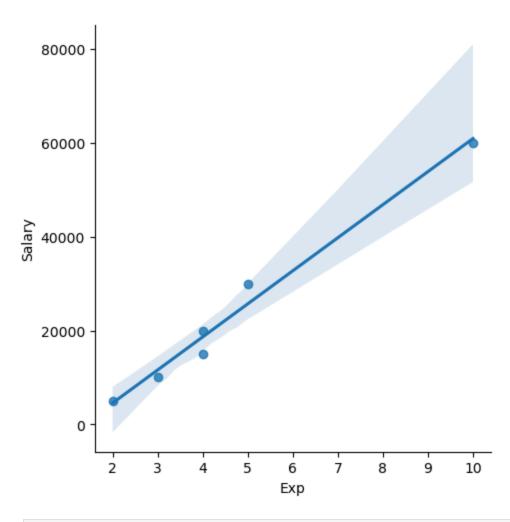
```
<class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
                      Non-Null Count Dtype
            Column
        --- -----
                      -----
                                     ----
        0
            Name
                      6 non-null
                                     object
        1
            Domain
                      6 non-null
                                     object
        2
            Age
                      6 non-null
                                     object
            Location 6 non-null
        3
                                     object
        4
                      6 non-null
            Salary
                                     object
        5
                      6 non-null
            Exp
                                     object
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [41]: clean_data['Age'] = clean_data['Age'].astype(int)
         clean_data['Salary'] = clean_data['Salary'].astype(int)
         clean_data['Exp'] = clean_data['Exp'].astype(int)
In [42]: 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
                                     object
        1
            Domain
                      6 non-null
                                     object
        2
            Age
                      6 non-null
                                     int32
            Location 6 non-null
        3
                                     object
        4
            Salary
                      6 non-null
                                     int32
        5
            Exp
                      6 non-null
                                     int32
       dtypes: int32(3), object(3)
       memory usage: 348.0+ bytes
In [43]: clean_data['Name'] = clean_data['Name'].astype('category')
         clean_data['Domain'] = clean_data['Domain'].astype('category')
         clean_data['Location'] = clean_data['Location'].astype('category')
In [44]: clean_data.info()
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 6 entries, 0 to 5
       Data columns (total 6 columns):
            Column
                      Non-Null Count Dtype
                      -----
        0
            Name
                      6 non-null
                                     category
        1
            Domain
                      6 non-null
                                     category
        2
            Age
                      6 non-null
                                     int32
        3
            Location 6 non-null
                                     category
        4
            Salary
                      6 non-null
                                     int32
        5
                      6 non-null
                                     int32
            Exp
       dtypes: category(3), int32(3)
       memory usage: 866.0 bytes
In [45]:
        clean_data.to_csv(r'C:\Python_Learning\Week5_EDA\20thFeb\CleanData.xlsx')
```

```
In [46]:
        import os
        os.getcwd()
Out[46]: 'C:\\Users\\bmittipa'
In [47]:
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        warnings.filterwarnings('ignore')
In [48]:
        # Visualization
        clean_data['Salary']
Out[48]: 0
              5000
             10000
         1
         2
             15000
             20000
             30000
             60000
        Name: Salary, dtype: int32
        sns.distplot(clean_data['Salary'])
In [49]:
Out[49]: <Axes: xlabel='Salary', ylabel='Density'>
              1e-5
          3.5
          3.0
          2.5
       Density
          2.0
          1.5
          1.0
          0.5
          0.0
            -40000 -20000
                                      20000
                                              40000
                                0
                                                      60000
                                                               80000
                                                                       100000
                                           Salary
In [50]:
        sns.histplot(clean_data['Salary'])
```





Out[51]: <seaborn.axisgrid.FacetGrid at 0x152c665d3d0>



```
In [62]:
         clean_data.columns
Out[62]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [55]: X_vars = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
In [61]: X_vars
Out[61]:
             Name
                                      Location Exp
                       Domain Age
                    Datascience
                                                  2
          0
              Mike
                                  34
                                       Mumbai
             Teddy
                        Testing
                                  45
                                      Bangalore
          2
              Umar
                     Dataanalyst
                                  50
                                      Bangalore
              Jane
                       Analytics
                                      Hyderbad
             Uttam
                       Statistics
                                      Bangalore
                                                   5
               Kim
                           NLP
                                  55
                                          Delhi
                                                  10
```

Y vars

In [59]:

In [60]:

Y_vars = clean_data[['Salary']]

Out[60]:		Salar	у							
	0	500	0							
	1	1000	0							
	2	1500	0							
	3	2000	0							
	4	3000	0							
	5	6000	0							
In [63]:	<pre># IMPUTATION imputation = pd.get_dummies(clean_data)</pre>									
In [64]:	imputation									
Out[64]:		۸۵۵		_	Name Issa	Name Kim	Name Mike	Name Teddy	Name_Umar	Nan
		Age	Salary	Exp	Name_Jane	rtanic_itani		rtuine_ready		
	0	34	Salary 5000	Exp 2	False	False	True	False	False	
	0									
		34	5000	2	False	False	True	False	False	
	1	34 45	5000	2	False False	False False	True False	False True	False False	
	1 2	34 45 50	5000 10000 15000	2 3 4	False False	False False	True False False	False True False	False False True	
	1 2 3	34 45 50 50	5000 10000 15000 20000	2 3 4 4	False False True	False False False	True False False False	False False False	False False True False	
	1 2 3 4	34 45 50 50 67	5000 10000 15000 20000 30000	2 3 4 4 5	False False True False	False False False False	True False False False	False False False False	False False True False False	•