In [285…
```
import pandas as pd
```

RAW DATA

In [286…
```
emp=pd.read_excel(r"D:\NIT\24NOV\23rd - Eda practicle\EDA- Practicle\Rawdata.xls
```

In [287…
```
emp
```

Out[287…

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

In [288…
```
emp.columns
```

Out[288…
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [289…
```
emp.shape
```

Out[289…
```
(6, 6)
```

In [290…
```
emp.head()
```

Out[290…

	Name	Domain	Age	Location	Salary	Exp
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 [291…
```
emp.tail()
```

Out[291...

	Name	Domain	Age	Location	Salary	Exp
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 [292...
```
emp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Name      6 non-null      object
 1   Domain    6 non-null      object
 2   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 [293...
```
emp
```

Out[293...

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

In [294...
```
emp.columns
```

Out[294...
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [295...
```
emp['Domain']
```

Out[295...
```
0     Datascience#$
1           Testing
2    Dataanalyst^^#
3       Ana^^lytics
4        Statistics
5               NLP
Name: Domain, dtype: object
```

In [296...
```
emp['Salary']
```

Out[296…
```
0     5^00#0
1    10%%000
2    1$5%000
3     2000^0
4     30000-
5    6000^$0
Name: Salary, dtype: object
```

In [297…
```
emp.isnull()
```

Out[297…

	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
4	False	False	False	True	False	False
5	False	False	False	False	False	False

In [298…
```
emp.isnull().sum()
```

Out[298…
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [299…
```
emp['Name']
```

Out[299…
```
0      Mike
1    Teddy^
2     Uma#r
3      Jane
4    Uttam*
5       Kim
Name: Name, dtype: object
```

In [300…
```
emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
```

In [301…
```
emp
```

Out[301...

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#$	34 years	Mumbai	5^00#0	2+
1	Teddy	Testing	45' yr	Bangalore	10%%000	<3
2	Umar	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 [302...
```
emp['Name']
```

Out[302...
```
0     Mike
1    Teddy
2     Umar
3     Jane
4    Uttam
5      Kim
Name: Name, dtype: object
```

In [303...
```
emp['Domain']=emp['Domain'].str.replace(r'\W','' ,regex=True)
```

In [304...
```
emp['Domain']
```

Out[304...
```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

In [305...
```
emp['Age']=emp['Age'].str.replace(r'\W','',regex=True)
```

In [306...
```
emp['Age']
```

Out[306...
```
0    34years
1       45yr
2        NaN
3        NaN
4       67yr
5       55yr
Name: Age, dtype: object
```

In [307...
```
emp['Age']=emp['Age'].str.extract("(\d+)")
```

In [308...
```
emp['Age']
```

Out[308...
```
0     34
1     45
2    NaN
3    NaN
4     67
5     55
Name: Age, dtype: object
```

In [309…
```
emp['Salary']=emp['Salary'].str.replace('\W',"",regex=True)
```

In [310…
```
emp["Salary"]
```

Out[310…
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: object
```

In [311…
```
emp
```

Out[311…

	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> yrs
3	Jane	Analytics	NaN	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5+ year
5	Kim	NLP	55	Delhi	60000	10+

In [312…
```
emp['Exp']=emp['Exp'].str.extract('(\d+)')
```

In [313…
```
emp['Exp']
```

Out[313…
```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

In [314…
```
emp
```

Out[314…

	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 [315…
```
Clean_emp2=emp.copy()
```

In [316…
```
Clean_emp2
```

Out[316...

	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 [317...
```
Clean_emp2['Salary']
```

Out[317...
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: object
```

In [318...
```
import numpy as np
```

In [319...
```
Clean_emp2.isnull().sum()
```

Out[319...
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

In [320...
```
Clean_emp2
```

Out[320...

	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 [321...
```
Clean_emp2['Age']=Clean_emp2['Age'].fillna(np.mean(pd.to_numeric(Clean_emp2['Age
```

In [322...
```
Clean_emp2
```

Out[322...

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	NaN	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	NaN
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [323...
```
Clean_emp2['Exp']=Clean_emp2['Exp'].fillna(np.mean(pd.to_numeric(Clean_emp2['Exp
```

In [324...
```
Clean_emp2['Exp']
```

Out[324...
```
0      2
1      3
2      4
3    4.8
4      5
5     10
Name: Exp, dtype: object
```

In [325...
```
Clean_emp2
```

Out[325...

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	NaN	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	NaN	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [326...
```
Clean_emp2['Location']=Clean_emp2['Location'].fillna(Clean_emp2['Location'].mode
```

In [327...
```
Clean_emp2['Location']
```

Out[327...
```
0       Mumbai
1    Bangalore
2    Bangalore
3     Hyderbad
4    Bangalore
5        Delhi
Name: Location, dtype: object
```

In [328...
```
Clean_emp2
```

Out[328…

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50.25	Bangalore	15000	4
3	Jane	Analytics	50.25	Hyderbad	20000	4.8
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [329…
```
Clean_emp2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Name      6 non-null      object
 1   Domain    6 non-null      object
 2   Age       6 non-null      object
 3   Location  6 non-null      object
 4   Salary    6 non-null      object
 5   Exp       6 non-null      object
dtypes: object(6)
memory usage: 420.0+ bytes
```

In [330…
```
Clean_emp2.describe()
```

Out[330…

	Name	Domain	Age	Location	Salary	Exp
count	6	6	6.00	6	6	6
unique	6	6	5.00	4	6	6
top	Mike	Datascience	50.25	Bangalore	5000	2
freq	1	1	2.00	3	1	1

In [331…
```
Clean_emp2.isnull().sum()
```

Out[331…
```
Name        0
Domain      0
Age         0
Location    0
Salary      0
Exp         0
dtype: int64
```

In [332…
```
Clean_emp2.dtypes
```

Out[332…
```
Name        object
Domain      object
Age         object
Location    object
Salary      object
Exp         object
dtype: object
```

In [333…
```
Clean_emp2['Age']=Clean_emp2["Age"].astype(int)
```

In [334…
```
Clean_emp2.dtypes
```

Out[334…
```
Name        object
Domain      object
Age          int32
Location    object
Salary      object
Exp         object
dtype: object
```

In [335…
```
Clean_emp2['Salary']=Clean_emp2['Salary'].astype(int)
```

In [336…
```
Clean_emp2['Exp']=Clean_emp2['Exp'].astype(int)
```

In [337…
```
Clean_emp2
```

Out[337…

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

In [338…
```
Clean_emp2.dtypes
```

Out[338…
```
Name        object
Domain      object
Age          int32
Location    object
Salary       int32
Exp          int32
dtype: object
```

In [339…
```
Clean_emp2.describe()
```

Out[339...

	Age	Salary	Exp
count	6.000000	6.000000	6.000000
mean	50.166667	23333.333333	4.666667
std	10.907184	19916.492328	2.804758
min	34.000000	5000.000000	2.000000
25%	46.250000	11250.000000	3.250000
50%	50.000000	17500.000000	4.000000
75%	53.750000	27500.000000	4.750000
max	67.000000	60000.000000	10.000000

In [340...
```
Clean_emp2.info()
```

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

In [341...
```
Clean_emp2['Name']=Clean_emp2['Name'].astype('category')
```

In [342...
```
Clean_emp2
```

Out[342...

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

In [343...
```
Clean_emp2.dtypes
```

Out[343…
```
Name        category
Domain        object
Age            int32
Location      object
Salary         int32
Exp            int32
dtype: object
```

In [344…
```
Clean_emp2['Location']=Clean_emp2['Location'].astype('category')
```

In [345…
```
Clean_emp2['Domain']=Clean_emp2['Domain'].astype('category')
```

CLEAN DATA

In [346…
```
Clean_emp2
```

Out[346…

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

In [347…
```
Clean_emp2.dtypes
```

Out[347…
```
Name        category
Domain      category
Age            int32
Location    category
Salary         int32
Exp            int32
dtype: object
```

In [348…
```
Clean_emp2.to_csv('Clean_emp2.csv')
```

In [349…
```
import os
os.getcwd()
```

Out[349…
```
'C:\\Users\\arnak\\DATA SCIENCE(NIT)'
```

In [350…
```
Clean_emp2
```

Out[350…

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

In [351…
```
import matplotlib.pyplot as plt
import seaborn as sns
```

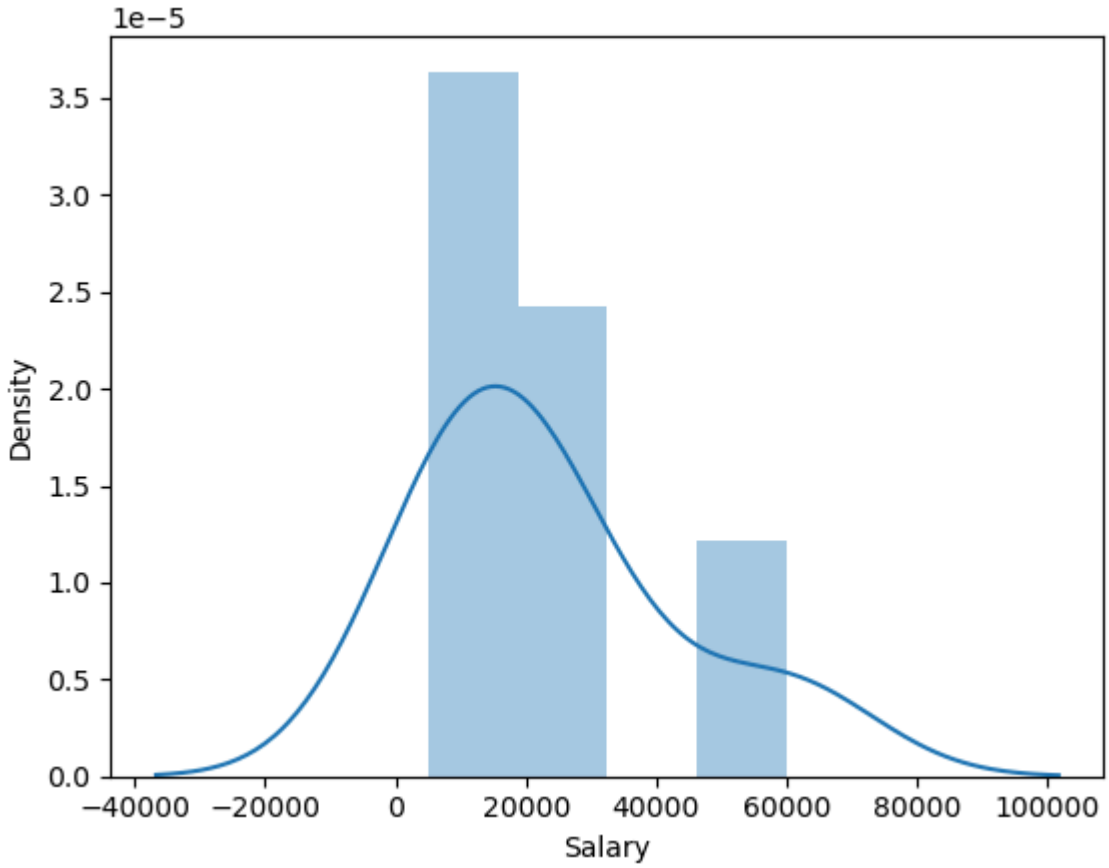
In [352…
```
 %matplotlib inline
```

In [353…
```
import warnings
warnings.filterwarnings('ignore')
```

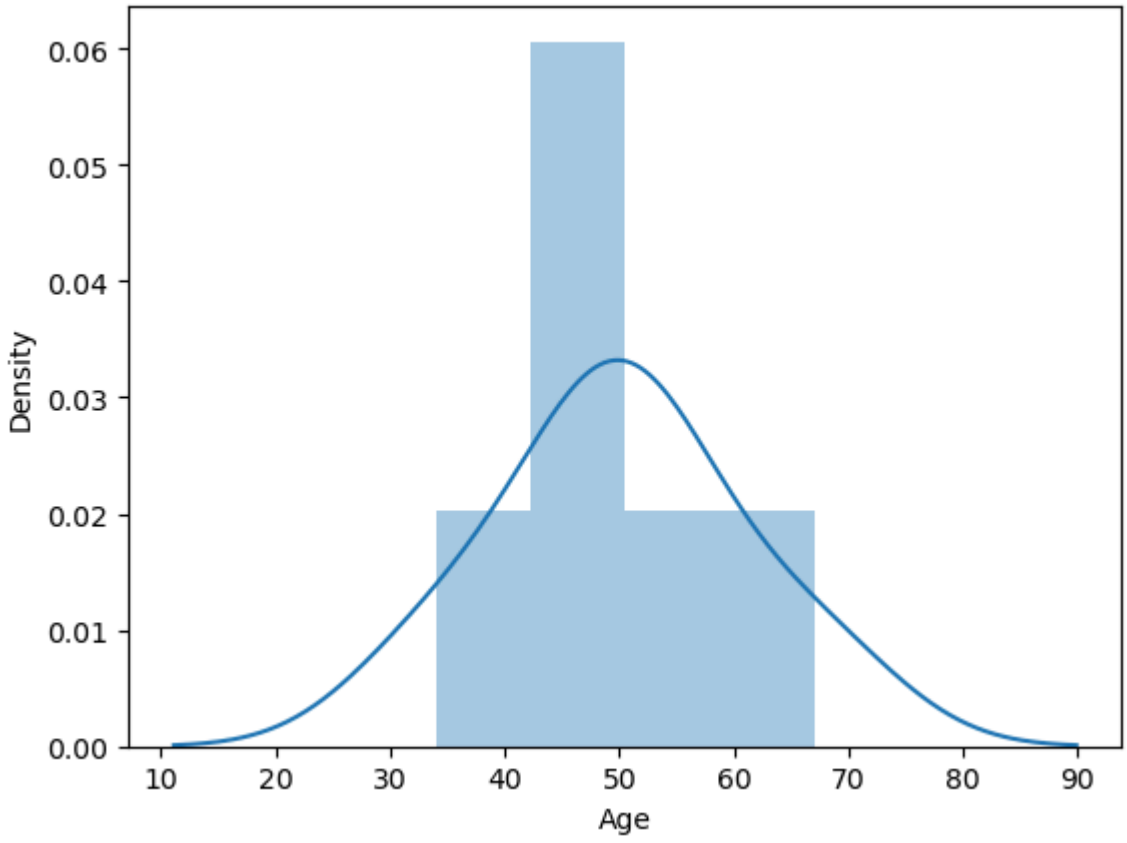
In [354…
```
Clean_emp2['Salary']
```

Out[354…
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int32
```

In [357…
```
vis1=sns.distplot(Clean_emp2['Salary'])
```



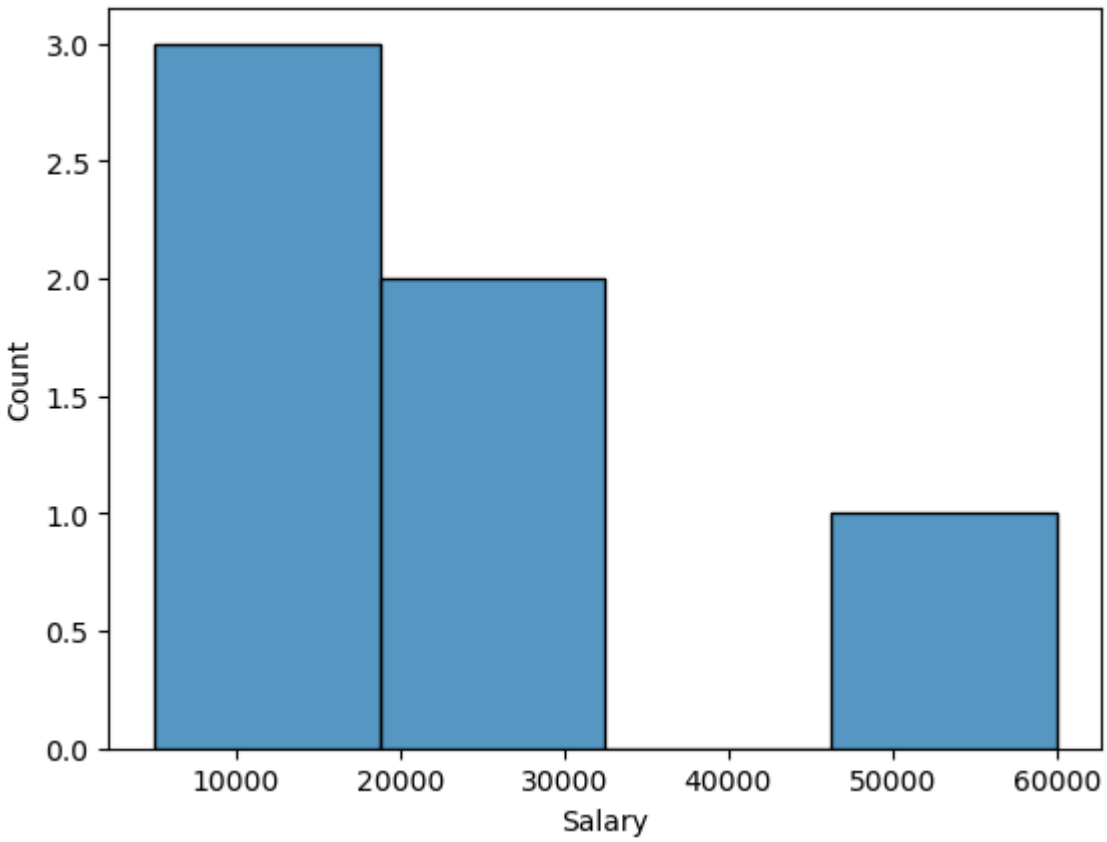
In [358…
```
vis2=sns.distplot(Clean_emp2['Age'])
```



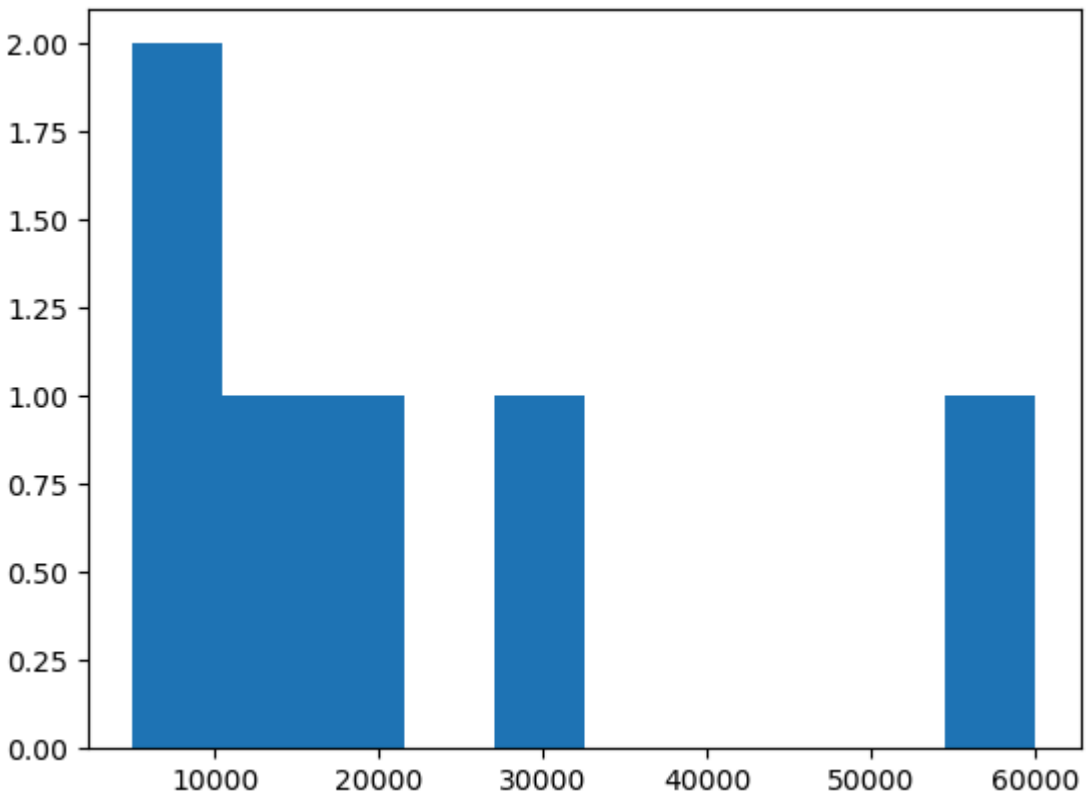
In [361…
```
vis2
```

Out[361…
```
<Axes: xlabel='Age', ylabel='Density'>
```

In [367…
```
vis3=sns.histplot(Clean_emp2['Salary'])
```



In [373…
```
vis3=plt.hist(Clean_emp2['Salary'])
```

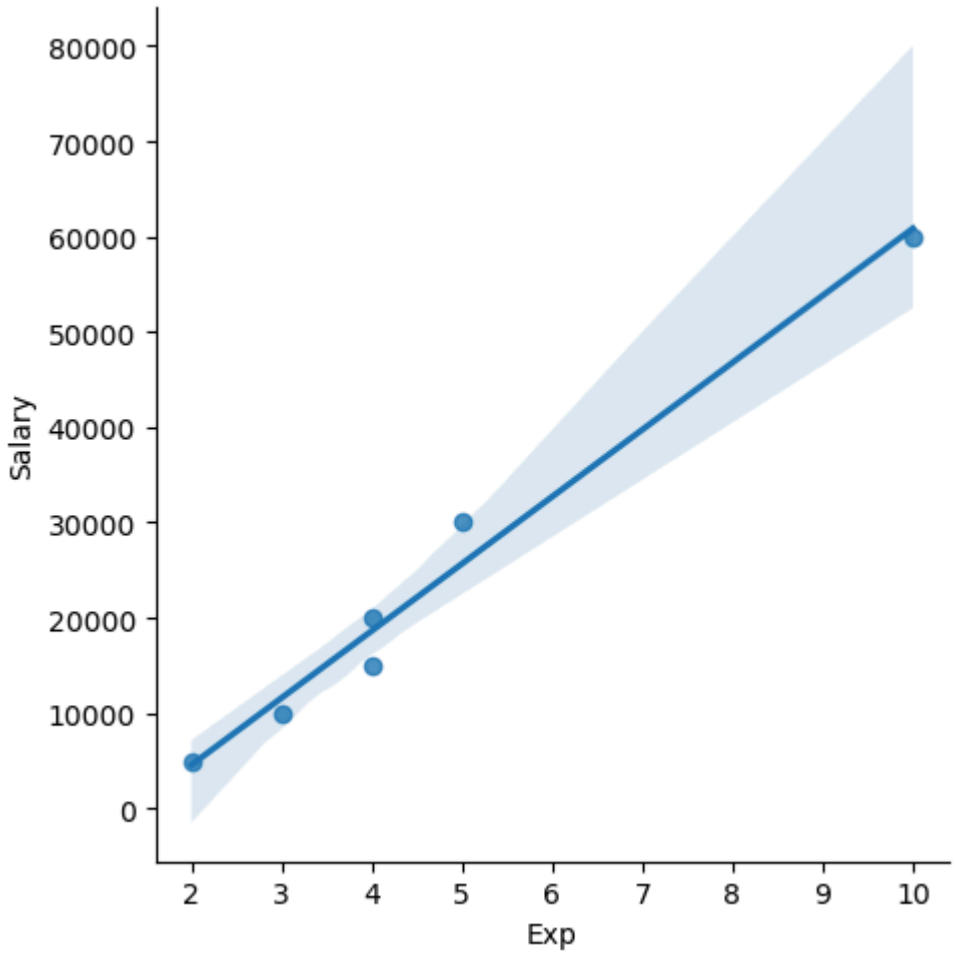


In [375…
```
vis2
```

Out[375…
```
<Axes: xlabel='Age', ylabel='Density'>
```

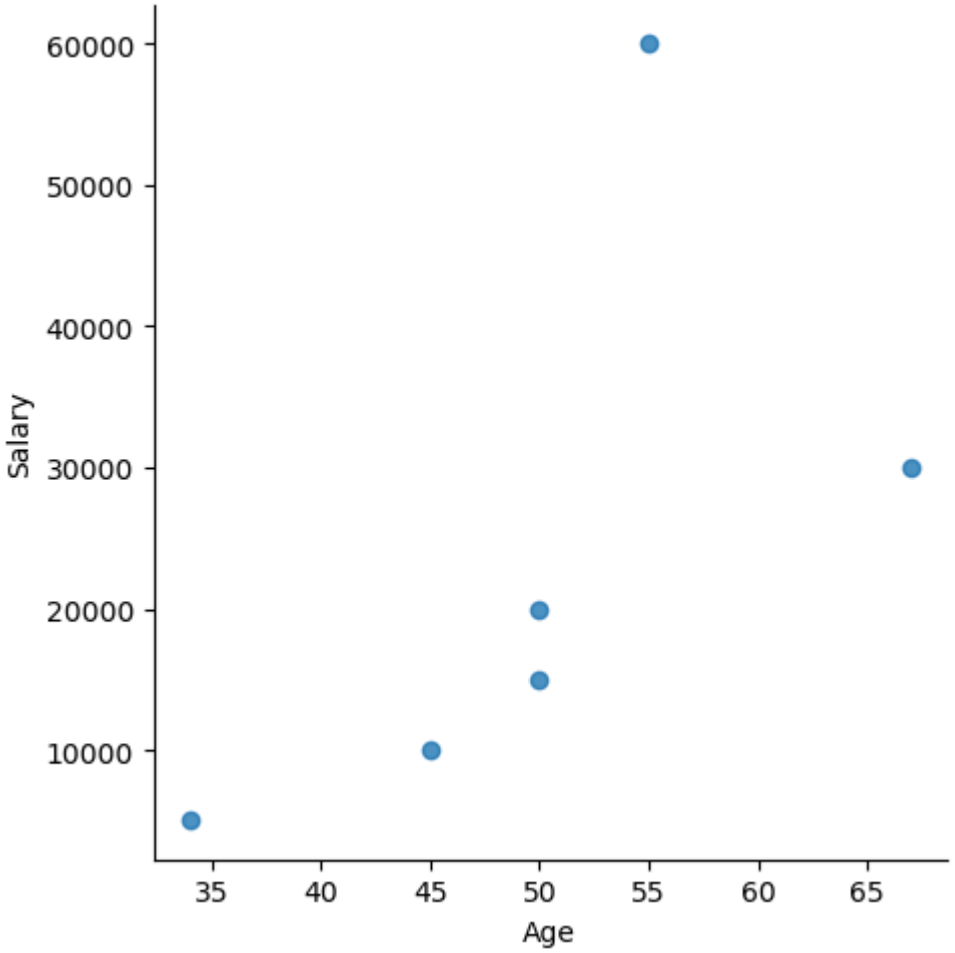
In [378…

```
vis4=sns.lmplot(
    data=Clean_emp2,
    x="Exp",
    y='Salary')
```



In [382…

```
vis4=sns.lmplot(data=Clean_emp2,x='Age',y="Salary",fit_reg=False)
```



In [387…
```
Clean_emp2[:]
```

Out[387…

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

In [388…
```
Clean_emp2[0:6:2]
```

Out[388…

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
2	Umar	Dataanalyst	50	Bangalore	15000	4
4	Uttam	Statistics	67	Bangalore	30000	5

In [389…
```
Clean_emp2[::-1]
```

Out[389...

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

In [390...
```
X_iv =Clean_emp2[['Name', 'Domain', 'Age', 'Location', 'Exp']]
```

In [391...
```
X_iv
```

Out[391...

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

In [392...
```
Clean_emp2
```

Out[392...

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

In [393...
```
imputation=pd.get_dummies(Clean_emp2)
```

In [394...
```
imputation
```

Out[394...

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar
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

In [395...

```
Clean_emp2
```

Out[395...

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

In [396...

```
imputation
```

Out[396...

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar
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

In [ ]: