```
import pandas as pd

emp=pd.read_excel(r'C:\Users\Admin\Downloads\Rawdata.xlsx')

emp
```

```
     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+
```

```
emp.shape  ##dimensions of the dataframe
```

```
(6, 6)
```

```
len(emp)
```

```
6
```

```
emp.head()
```

```
     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
```

```
emp.tail()
```

```
     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+
```

```
emp.columns
```

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

```
len(emp.columns)
```

```
6
```

```
emp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
```

```
Data columns (total 6 columns):
 #   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
```

```
emp.isnull()
```

	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

```
emp.isnull().sum()
```

```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

```
emp['Name']
```

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

```
emp['Domain']
```

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

```
emp['Age']
```

```
0    34 years
1      45' yr
2         NaN
3         NaN
4       67-yr
5        55yr
Name: Age, dtype: object
```

```
emp['Location']
```

```
0       Mumbai
1    Bangalore
2          NaN
3     Hyderbad
4          NaN
5        Delhi
Name: Location, dtype: object
```

```
emp['Salary']
```

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

```
emp['Exp']
```

```
0         2+
1         <3
2     4> yrs
3        NaN
4    5+ year
5        10+
Name: Exp, dtype: object
```

```
emp[['Name','Domain']]
```

```
     Name          Domain
0    Mike   Datascience#$
1  Teddy^         Testing
2   Uma#r  Dataanalyst^^#
3    Jane     Ana^^lytics
4  Uttam*      Statistics
5     Kim             NLP
```

```
emp[['Name','Domain','Age']]
```

```
     Name          Domain       Age
0    Mike   Datascience#$  34 years
1  Teddy^         Testing    45' yr
2   Uma#r  Dataanalyst^^#       NaN
3    Jane     Ana^^lytics       NaN
4  Uttam*      Statistics     67-yr
5     Kim             NLP      55yr
```

```
emp[['Name','Domain','Age','Location','Salary','Exp']]
```

```
     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+
```

Data Cleansing

```
emp['Name']
```

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

```
emp['Name']= emp['Name'].str.replace(r'\W','')
```

```
emp['Name']
```

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

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

```
emp['Domain']
```

```
0     Datascience#$
1           Testing
2    Dataanalyst^^#
3       Ana^^lytics
4        Statistics
```

```
5                NLP
Name: Domain, dtype: object
```

EDA Intro

```
import pandas as pd
```

```
pd.__version__
```

```
'2.2.2'
```

```
emp=pd.read_excel(r'C:\Users\Admin\Downloads\Rawdata.xlsx')
```

```
emp
```

```
     Name          Domain       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+
```

```
emp.isnull().sum()
```

```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

```
id(emp)
```

```
1943775782160
```

```
emp.columns
```

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

```
emp.shape
```

```
(6, 6)
```

```
emp.head
```

```
<bound method NDFrame.head of      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+>
```

```
emp.tail
```

```
<bound method NDFrame.tail of      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+>
```

```
emp.isnull().sum()
```

```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

```
emp.info()
```

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

```
emp.isnull()
```

```
    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
```

```
emp.isna()
```

	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

Data Cleaning or Data Cleansing

```
emp
```

	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+

```
emp['Name']
```

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

```
emp['Name']=emp['Name'].str.replace(r'\W','',regex=True)   #regex means regular expression non word character
```

```
emp['Name']
```

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

```
emp
```

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#$	34 years	Mumbai	5^00#0	2+
1	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+
```

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

```
emp['Domain']
```

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

```
emp
```

```
    Name       Domain       Age   Location   Salary      Exp
0   Mike  Datascience  34 years     Mumbai   5^00#0       2+
1  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+
```

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

```
emp['Age']
```

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

```
emp
```

```
    Name       Domain      Age   Location   Salary      Exp
0   Mike  Datascience  34years     Mumbai   5^00#0       2+
1  Teddy      Testing     45yr  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     67yr        NaN   30000-  5+ year
5    Kim          NLP     55yr      Delhi  6000^$0      10+
```

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

```
emp['Salary']
```

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

```
emp
```

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

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

```
emp['Exp']
```

```
0        2
1        3
2     4yrs
3      NaN
4    5year
5       10
Name: Exp, dtype: object
```

```
emp
```

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

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

```
emp['Age']
```

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

```
emp
```

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

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

```
emp['Exp']
```

```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

```
emp
```

```
    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
```

```
clean_data=emp.copy()
```

```
clean_data
```

```
    Name       Domain  Age   Location Salary  Exp
0   Mike  Datascience   34     Mumbai   5000    2
1  Teddy      Testing   45  Bangalore  10000    3
2   Umar  Dataanalyst  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
```

```
clean_data['Age']
```

```
0     34
1     45
2    NaN
3    NaN
4     67
```

```
5     55
Name: Age, dtype: object
```

```
import numpy as np
```

```
clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean
_data['Age'])))
```

```
clean_data['Age']
```

```
0       34
1       45
2    50.25
3    50.25
4       67
5       55
Name: Age, dtype: object
```

```
clean_data['Exp']
```

```
0      2
1      3
2      4
3    NaN
4      5
5     10
Name: Exp, dtype: object
```

```
clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean
_data['Exp'])))
```

```
clean_data['Exp']
```

```
0      2
1      3
2      4
3    4.8
4      5
5     10
Name: Exp, dtype: object
```

```
clean_data
```

```
    Name       Domain    Age   Location Salary  Exp
0   Mike  Datascience     34     Mumbai   5000    2
1  Teddy      Testing     45  Bangalore  10000    3
2   Umar  Dataanalyst  50.25        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
```

```
clean_data['Location'].isnull().sum()
```

```
2
```

```
clean_data['Location']
```

```
0       Mumbai
1    Bangalore
2          NaN
3     Hyderbad
4          NaN
5        Delhi
Name: Location, dtype: object
```

```
clean_data['Location']=clean_data['Location'].fillna(clean_data['Locat
ion'].mode()[0])
```

```
clean_data['Location']
```

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

```
clean_data
```

	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

```
clean_data.info()
```

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

```
clean_data['Age']=clean_data['Age'].astype(int)

clean_data.info()
```

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

```
clean_data['Salary']=clean_data['Salary'].astype(int)
clean_data['Exp']=clean_data['Exp'].astype(int)

clean_data.info()
```

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

```
clean_data['Name']=clean_data['Name'].astype('category')
clean_data['Domain']=clean_data['Domain'].astype('category')
clean_data['Location']=clean_data['Location'].astype('category')

clean_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 6 columns):
 #   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   Exp       6 non-null      int32
dtypes: category(3), int32(3)
memory usage: 866.0 bytes
```

```
clean_data
```

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

```
clean_data.to_csv('clean_data.csv')
```

```
import os
os.getcwd()
```

```
'C:\\Users\\Admin'
```

```
clean_data
```

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

EDA TECHNIQUE LETS APPLY

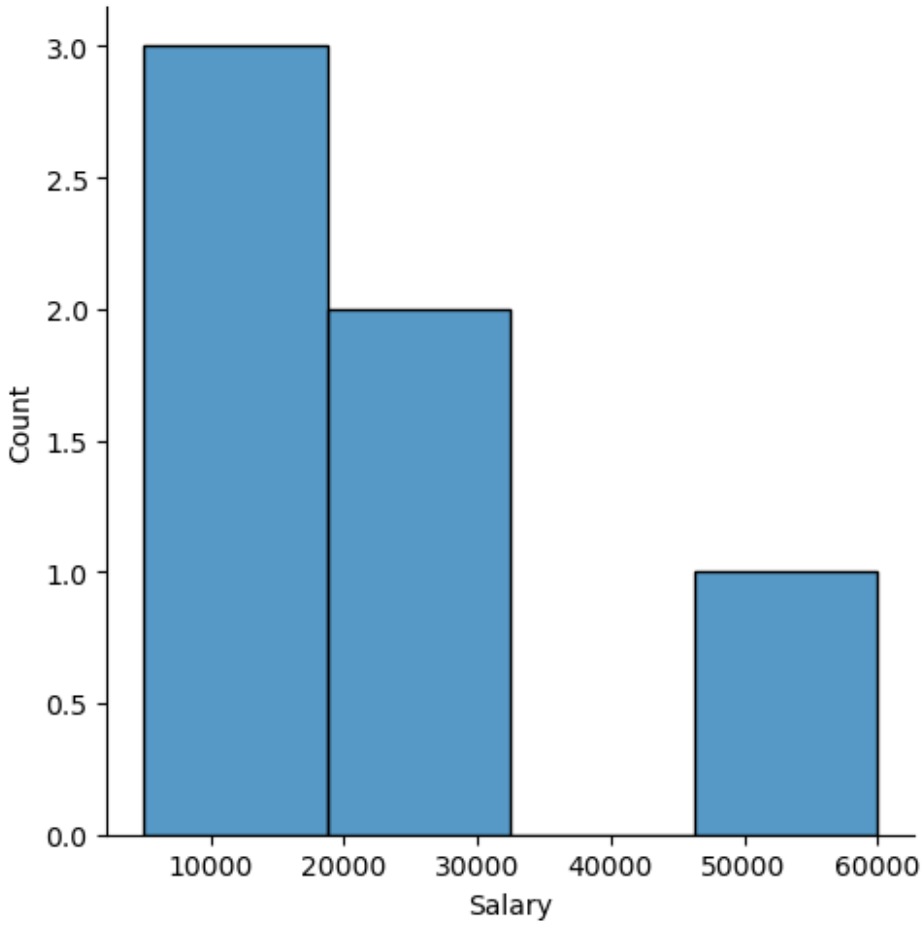
```
import matplotlib.pyplot as plt   #visualization
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')

clean_data['Salary']
```

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

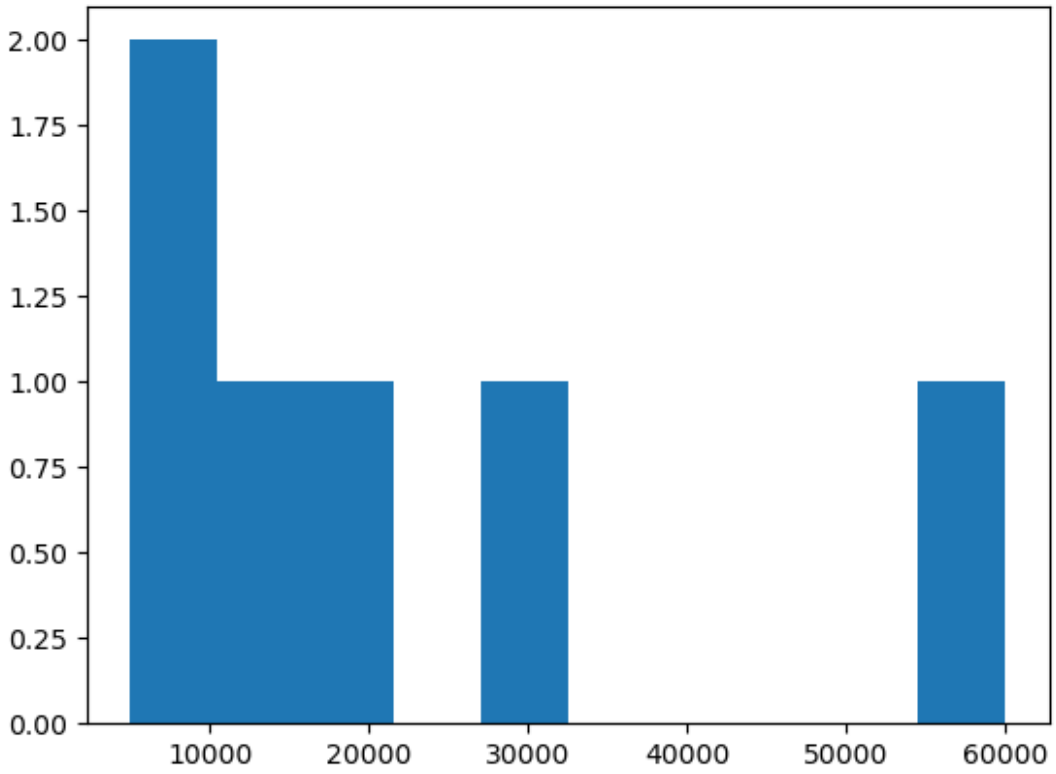
```
clean_data['Salary']
```

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

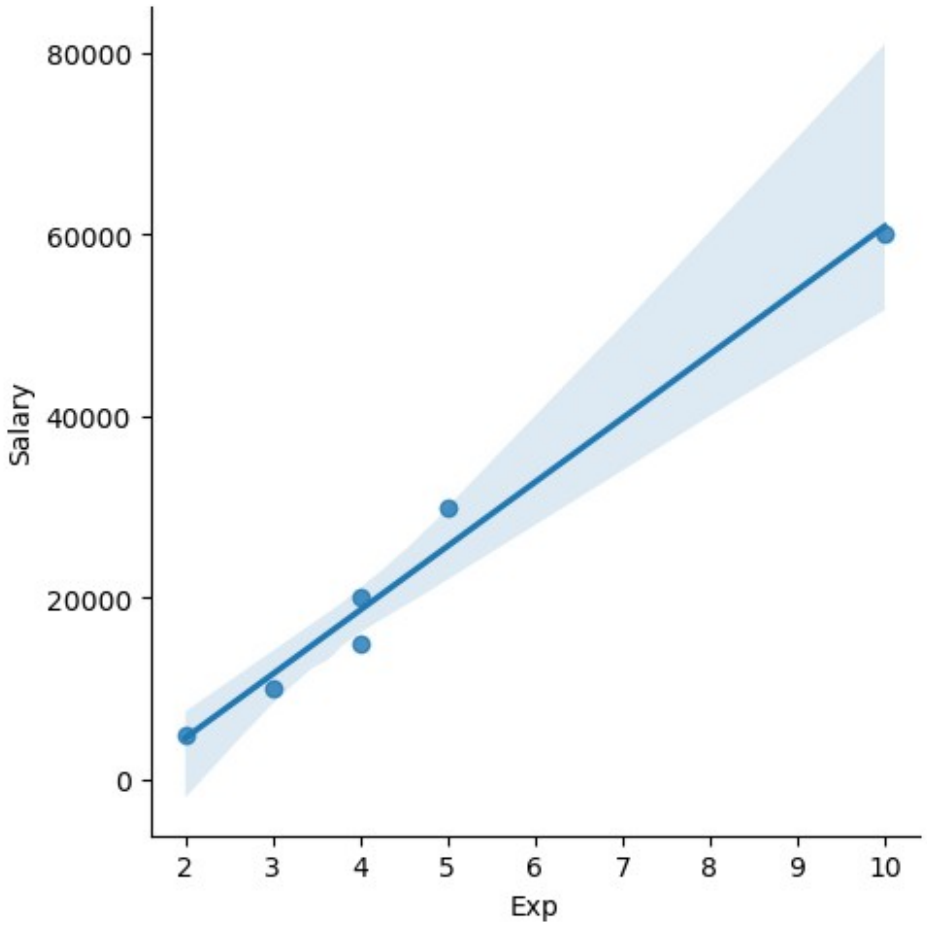
```
vis1=sns.displot(clean_data['Salary'])
```



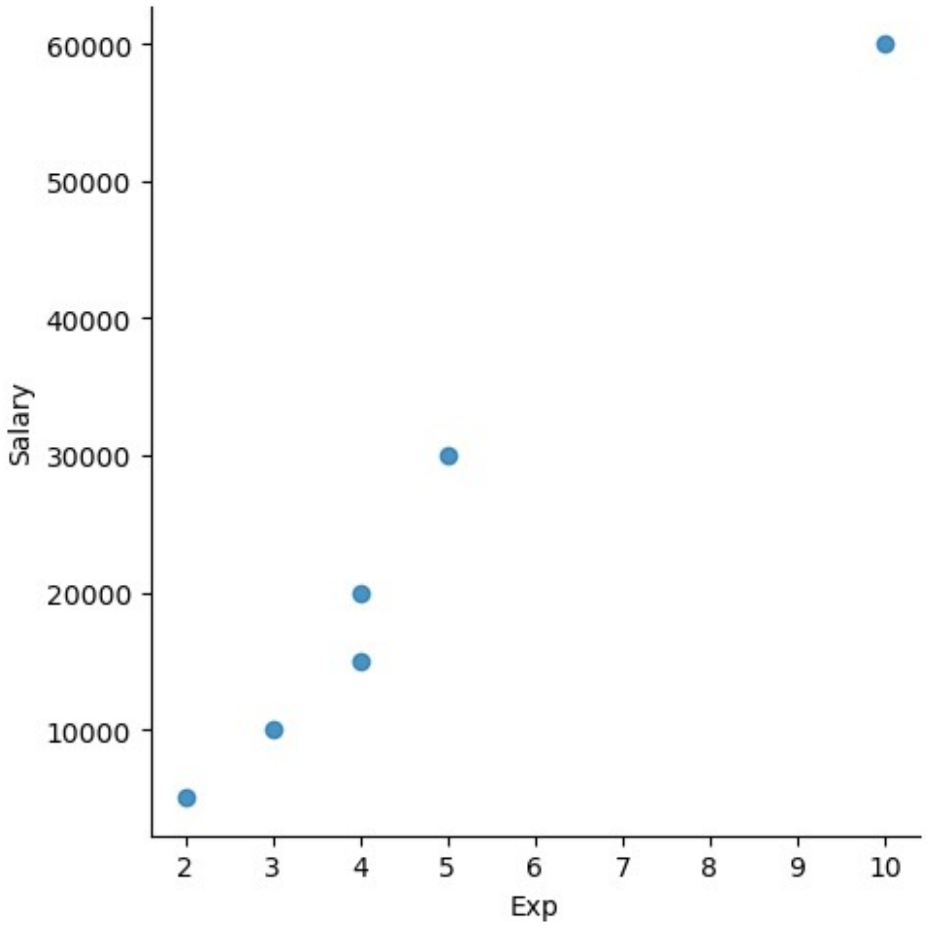

```
vis2=plt.hist(clean_data['Salary'])
```




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




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




```
clean_data[:]
```

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

```
clean_data[0:6:2]
```

	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

```
clean_data[::-1]
```

	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
```

```
clean_data.columns
```

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

```
X_iv=clean_data[['Name','Domain','Age','Location','Exp']]
```

```
X_iv
```

```
    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
```

```
y_dv=clean_data[['Salary']]
```

```
y_dv
```

```
   Salary
0    5000
1   10000
2   15000
3   20000
4   30000
5   60000
```

```
emp
```

```
    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
```

```
clean_data
```

```
    Name       Domain  Age   Location  Salary  Exp
0   Mike  Datascience   34     Mumbai    5000    2
1  Teddy      Testing   45  Bangalore   10000    3
2   Umar  Dataanalyst   50  Bangalore   15000    4
3   Jane    Analytics   50   Hyderbad   20000    4
```

```
4  Uttam   Statistics   67  Bangalore   30000    5
5    Kim          NLP   55      Delhi   60000   10
```

X_iv

```
    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
```

y_dv

```
   Salary
0    5000
1   10000
2   15000
3   20000
4   30000
5   60000
```

clean_data

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

```
imputation=pd.get_dummies(clean_data)
```

imputation

```
   Age  Salary  Exp  Name_Jane  Name_Kim  Name_Mike  Name_Teddy
Name_Umar  \
0   34    5000    2      False     False       True       False
False
1   45   10000    3      False     False      False        True
False
2   50   15000    4      False     False      False       False
True
3   50   20000    4       True     False      False       False
False
4   67   30000    5      False     False      False       False
False
5   55   60000   10      False      True      False       False
False
```

```
   Name_Uttam  Domain_Analytics  Domain_Dataanalyst
Domain_Datascience  \
0       False             False               False
True
1       False             False               False
False
2       False             False                True
False
3       False              True               False
False
4        True             False               False
False
5       False             False               False
False

   Domain_NLP  Domain_Statistics  Domain_Testing
Location_Bangalore  \
0       False              False           False                False

1       False              False            True                 True

2       False              False           False                 True

3       False              False           False                False

4       False               True           False                 True

5        True              False           False                False


   Location_Delhi  Location_Hyderbad  Location_Mumbai
0           False              False             True
1           False              False            False
2           False              False            False
3           False               True            False
4           False              False            False
5            True              False            False
```

clean_data

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

imputation

```
   Age  Salary  Exp  Name_Jane  Name_Kim  Name_Mike  Name_Teddy
Name_Umar  \
0   34    5000    2      False     False       True       False
False
1   45   10000    3      False     False      False        True
False
2   50   15000    4      False     False      False       False
True
3   50   20000    4       True     False      False       False
False
4   67   30000    5      False     False      False       False
False
5   55   60000   10      False      True      False       False
False

   Name_Uttam  Domain_Analytics  Domain_Dataanalyst
Domain_Datascience  \
0       False             False               False
True
1       False             False               False
False
2       False             False                True
False
3       False              True               False
False
4        True             False               False
False
5       False             False               False
False

   Domain_NLP  Domain_Statistics  Domain_Testing
Location_Bangalore  \
0       False              False           False               False

1       False              False            True                True

2       False              False           False                True

3       False              False           False               False

4       False               True           False                True

5        True              False           False               False


   Location_Delhi  Location_Hyderbad  Location_Mumbai
0           False              False             True
1           False              False            False
2           False              False            False
3           False               True            False
```

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