39hlhgox8

February 20, 2025

```
[177]: import pandas as pd
```

```
[179]: pd.__version__
```

```
[179]: '2.2.2'
```

```
[181]: emp = pd.read_excel(r'C:\Users\mikim\Downloads\Rawdata.xlsx')
```

```
[183]: emp
```

```
[183]:      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+
```

```
[185]: id(emp)
```

```
[185]: 2352588227120
```

```
[187]: emp.columns
```

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

```
[189]: emp.shape
```

```
[189]: (6, 6)
```

```
[191]: emp.head()
```

```
[191]:      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
```

```
[193]: emp.tail()
```

```
[193]:      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+
```

```
[195]: 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
```

```
[197]: emp
```

```
[197]:      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+
```

```
[199]: emp.isnull()
```

```
[199]:     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
```

```
[201]: emp.isna() #isnull and isna both are same
```

```
[201]:     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
```

```
[203]: emp.isnull().sum()
```

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

Data Cleaning

```
[206]: emp['Name']
```

```
[206]: 0      Mike
       1    Teddy^
       2     Uma#r
       3      Jane
       4    Uttam*
       5       Kim
       Name: Name, dtype: object
```

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

```
[210]: emp['Name']
```

```
[210]: 0     Mike
       1    Teddy
       2     Umar
       3     Jane
       4    Uttam
       5      Kim
       Name: Name, dtype: object
```

```
[212]: emp
```

```
[212]:     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+
```

[214]:
```
emp['Domain']
```

[214]:
```
0     Datascience#$
1           Testing
2    Dataanalyst^^#
3       Ana^^lytics
4        Statistics
5               NLP
Name: Domain, dtype: object
```

[216]:
```
emp['Domain'] = emp['Domain'].str.replace(r'\W','', regex=True)
```

[218]:
```
emp['Domain']
```

[218]:
```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

[220]:
```
emp
```

[220]:
```
    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+
```

[222]:
```
emp['Age'] = emp['Age'].str.replace(r'\W','', regex=True)
```

[224]:
```
emp['Age']
```

[224]:
```
0    34years
1       45yr
2        NaN
3        NaN
4       67yr
5       55yr
```

```
Name: Age, dtype: object
```

[226]:

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

```
<>:1: SyntaxWarning: invalid escape sequence '\d'
<>:1: SyntaxWarning: invalid escape sequence '\d'
C:\Users\mikim\AppData\Local\Temp\ipykernel_3648\1884116463.py:1: SyntaxWarning:
invalid escape sequence '\d'
  emp['Age'] = emp['Age'].str.extract('(\d+)')
```

[228]:

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

[228]:

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

[230]:

```python
emp
```

[230]:

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

[232]:

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

[234]:

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

[234]:

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

[236]:

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

[238]:

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

```
[238]: 0     5000
       1    10000
       2    15000
       3    20000
       4    30000
       5    60000
       Name: Salary, dtype: object
```

```
[240]: emp
```

```
[240]:     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+
```

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

```
<>:1: SyntaxWarning: invalid escape sequence '\d'
<>:1: SyntaxWarning: invalid escape sequence '\d'
C:\Users\mikim\AppData\Local\Temp\ipykernel_3648\3836251810.py:1: SyntaxWarning:
invalid escape sequence '\d'
  emp['Exp'] = emp['Exp'].str.extract('(\d+)')
```

```
[246]: emp['Exp']
```

```
[246]: 0      2
       1      3
       2      4
       3    NaN
       4      5
       5     10
       Name: Exp, dtype: object
```

```
[248]: emp
```

```
[248]:     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
```

```
[250]: clean_data = emp.copy()
```

[252]:

```
clean_data
```

[252]:

```
    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
```

[254]:

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

[254]:

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

[256]:

```
import numpy as np
```

[258]:

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

[260]:

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

[260]:

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

[262]:

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

[264]:

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

[264]:

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

[266]:

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

[268]:

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

[268]:

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

[270]:

```
clean_data
```

[270]:

```
    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
```

[272]:

```
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
```

[274]:

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

[276]:

```
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
```

[278]:

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

[284]:

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

[286]:

```
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
```

[288]:

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

[290]:

```
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
```

[292]:
```
clean_data
```

[292]:
```
    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
```

[294]:
```
clean_data.to_csv('clean_data.csv')
```

[298]:
```
import os
os.getcwd() #from the os give the saved current working directly
```

[298]:
```
'C:\\Users\\mikim'
```

[300]:
```
clean_data
```

[300]:
```
    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
```

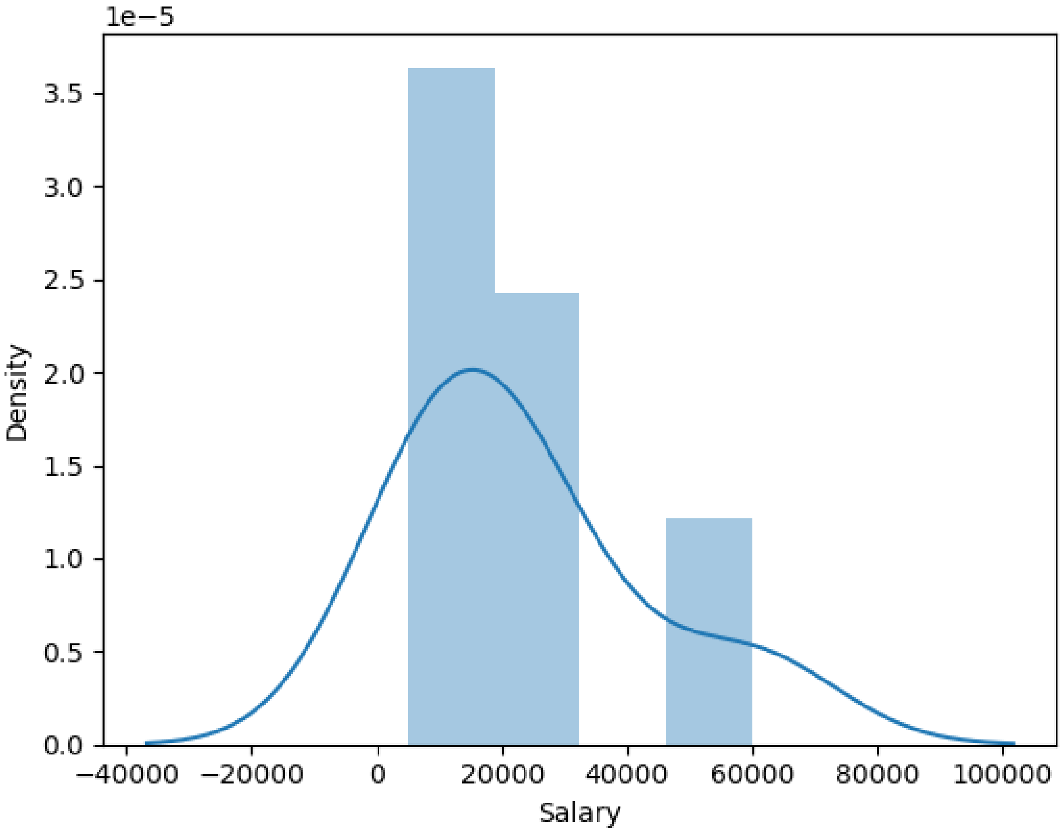
[302]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

[303]:
```
import warnings
warnings.filterwarnings('ignore')
```

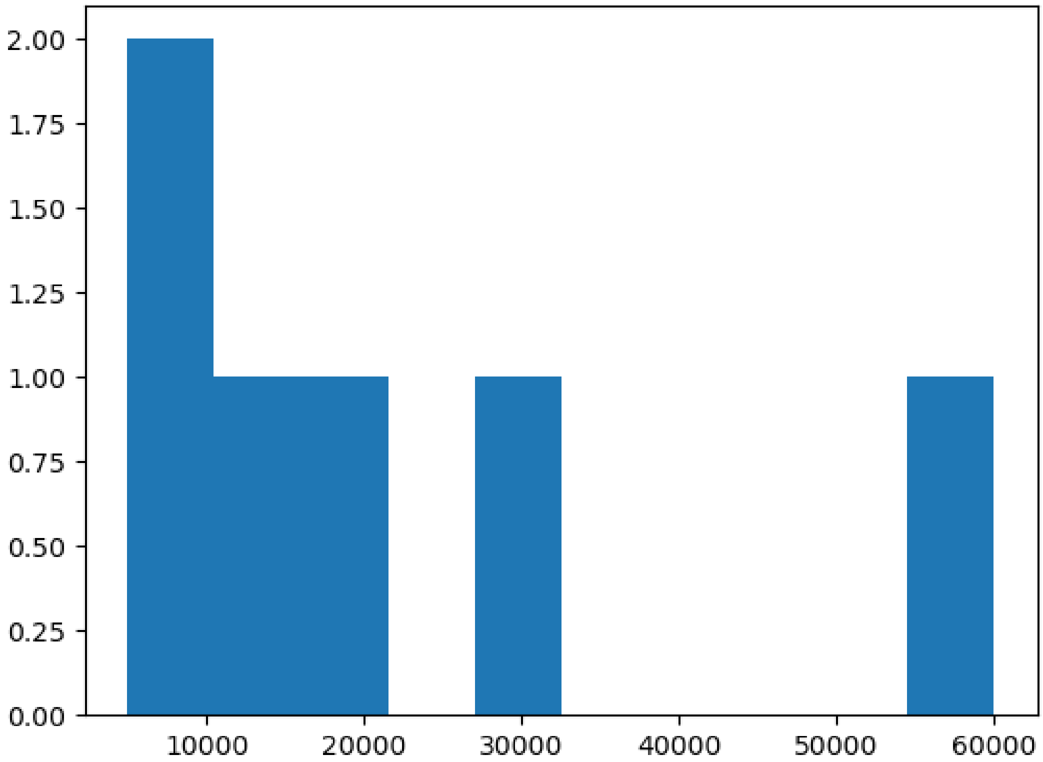
[304]:
```
clean_data['Salary']
```

[304]:
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int32
```

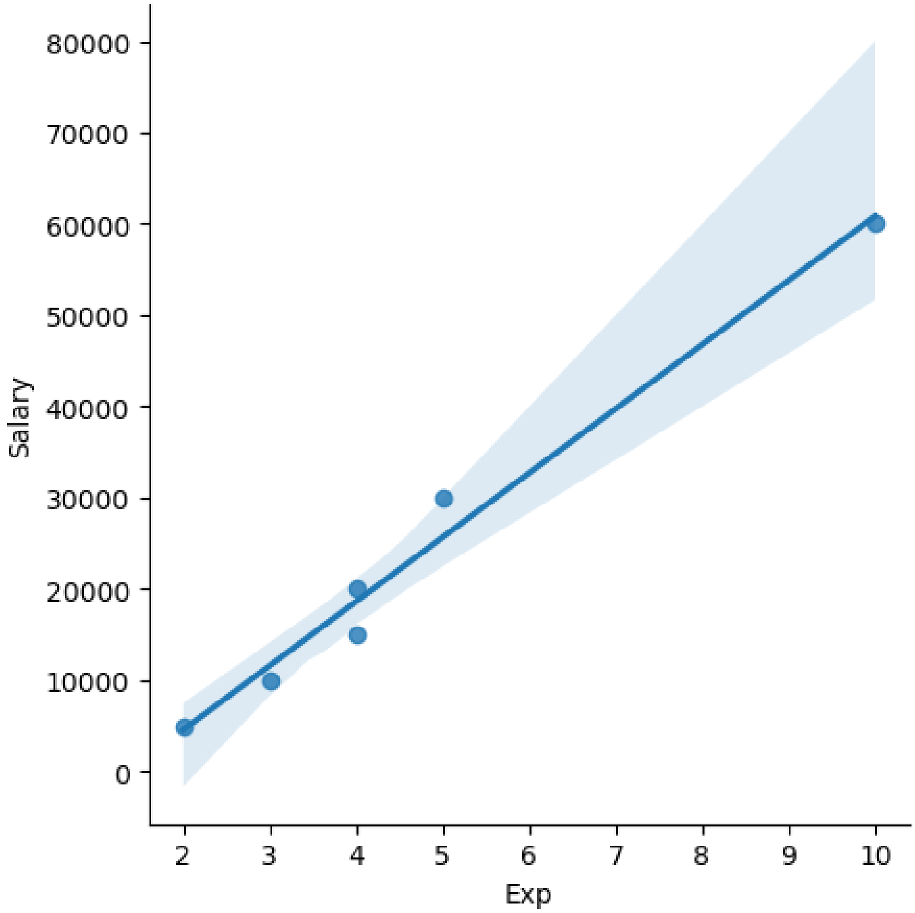
[308]:
```
vis1 = sns.distplot(clean_data['Salary'])
plt.show()
```



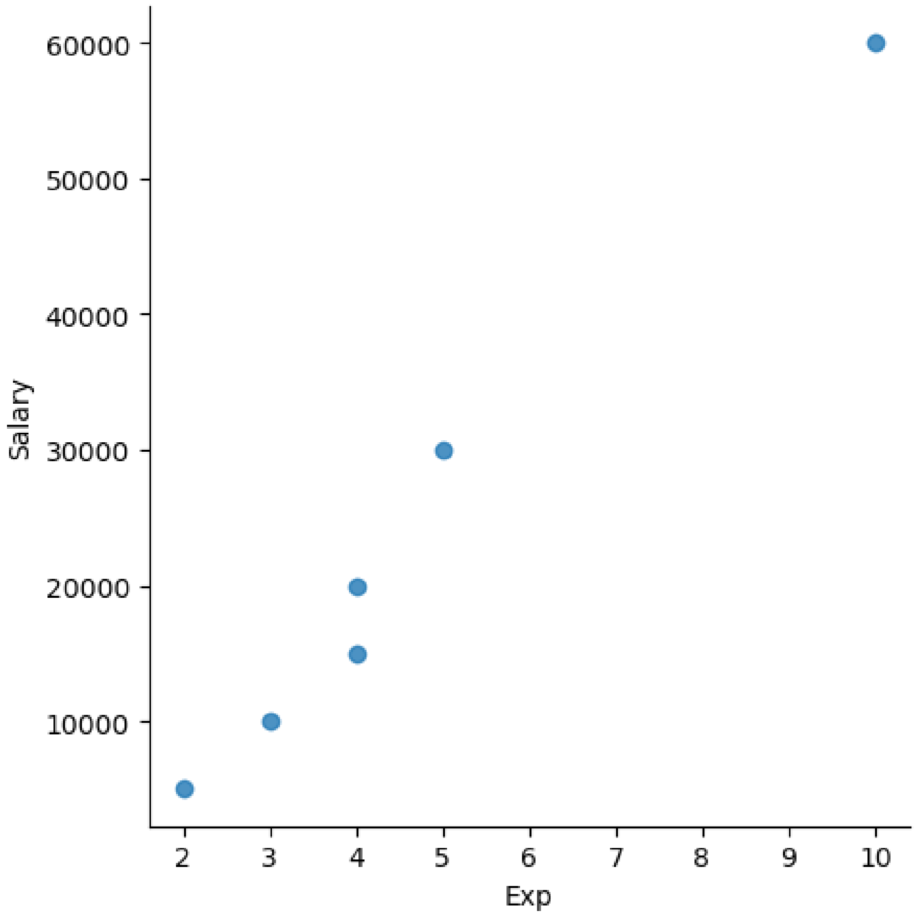
[310]:
```
vis2 = plt.hist(clean_data['Salary'])
plt.show()
```



[312]:
```
vis4= sns.lmplot(data=clean_data, x='Exp', y='Salary')
```



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



```
[316]: clean_data
```

```
[316]:     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
```

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

```
[320]: X_iv #independent variable
```

```
[320]:     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
```

[324]:
```
y_dv = clean_data['Salary']
```

[326]:
```
y_dv #dependent variable
```

[326]:
```
0     5000
1    10000
2    15000
3    20000
4    30000
5    60000
Name: Salary, dtype: int32
```

[328]:
```
clean_data
```

[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  Bangalore   15000    4
3   Jane    Analytics   50   Hyderbad   20000    4
4  Uttam   Statistics   67  Bangalore   30000    5
5    Kim          NLP   55      Delhi   60000   10
```

[330]:
```
imputation = pd.get_dummies(clean_data,dtype=int)
```

[332]:
```
imputation
```

[332]:
```
   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

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

   Domain_NLP  Domain_Statistics  Domain_Testing  Location_Bangalore  \
0           0                   0              0                   0
1           0                   0              1                   1
2           0                   0              0                   1
3           0                   0              0                   0
4           0                   1              0                   1
5           1                   0              0                   0

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

```
[334]: clean_data
```

```
[334]:     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
```

```
[336]: len(clean_data)
```

```
[336]: 6
```

```
[338]: imputation.columns
```

```
[338]: Index(['Age', 'Salary', 'Exp', 'Name_Jane', 'Name_Kim', 'Name_Mike',
              'Name_Teddy', 'Name_Umar', 'Name_Uttam', 'Domain_Analytics',
              'Domain_Dataanalyst', 'Domain_Datascience', 'Domain_NLP',
              'Domain_Statistics', 'Domain_Testing', 'Location_Bangalore',
              'Location_Delhi', 'Location_Hyderbad', 'Location_Mumbai'],
             dtype='object')
```

```
[340]: len(imputation.columns)
```

```
[340]: 19
```

```
[ ]:
```