Raw data to clean data conversion using python EDA

In [1]:
```
import pandas as pd
```

In [2]:
```
pd.__version__
```

Out[2]:
```
'2.2.3'
```

In [3]:
```
emp = pd.read_excel(r'C:\Users\Dell\Downloads\Rawdata.xlsx')
```

In [4]:
```
emp
```

Out[4]:

	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 [5]:
```
id(emp)
```

Out[5]:
```
1671923783664
```

In [6]:
```
emp.columns
```

Out[6]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [7]:
```
emp.shape
```

Out[7]:
```
(6, 6)
```

In [8]:
```
emp.head()
```

Out[8]:

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

In [ ]:
```
emp.info()
```

In [ ]:
```
emp
```

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

In [ ]:
```
emp.isna()
```

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

DATA CLEANING OR DATA CLEANSING

In [ ]:
```
emp
```

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

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

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

In [ ]:
```
emp
```

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

In [ ]:
```
emp['Domain']
```

In [ ]:
```
emp
```

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

In [ ]:
```
emp['Location']
```

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

In [ ]:
```
emp['Age']
```

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

In [ ]:
```
emp['Age']
```

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

In [ ]:
```
emp['Salary']
```

In [ ]:
```
emp
```

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

In [ ]:
```
emp['Exp']
```

In [ ]:
```
emp
```

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

In [ ]:
```
clean_data
```

In [ ]:
```
clean_data.isnull().sum()
```

In [ ]:
```
import numpy as np
```

In [ ]:
```
clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'
```

In [ ]:
```
clean_data['Age']
```

In [ ]:
```
clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp'
```

In [ ]:
```
clean_data['Exp']
```

In [ ]:
```
clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mode(
```

In [ ]:
```
clean_data['Location']
```

In [ ]:
```
clean_data
```

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

In [ ]:
```
import os
os.getcwd()
```

In [ ]:
```
import matplotlib.pyplot as plt #visualization
import seaborn as sns
```

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

In [ ]:
```
vis1 = sns.distplot(clean_data['Salary'] , bins = 10)
plt.show(vis1)
```

In [ ]:
```
vis2 = plt.hist(clean_data['Salary'])
plt.show(vis1)
```

In [ ]:
```
vis3 = sns.boxenplot(data = clean_data , x = 'Exp' , y = 'Salary')
```

In [ ]:
```
clean_data["Salary"] = pd.to_numeric(clean_data["Salary"], errors="coerce")
clean_data["Exp"] = pd.to_numeric(clean_data["Exp"], errors="coerce")
```

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

In [ ]:
```
clean_data
```

In [ ]:
```
y_d = clean_data['Salary']
y_d
```

In [ ]:
```
clean_data.columns
```

In [ ]:
```
x_iv = clean_data[['Name', 'Domain', 'Age', 'Location', 'Exp']]
```

In [ ]:
```
x_iv
```

In [ ]:
```
clean_data
```

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

In [ ]:
```
imputation
```

In [ ]:
```
imputation.columns
```

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