encoding====> converting catogarical data into numeric data

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import OneHotEncoder

saldf=pd.read_csv(r"C:\my pythonfiles\Salary_EDA.csv")
saldf.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary
0   90000.0
1   65000.0
2  150000.0
3   60000.0
4   60000.0
```

```
categorical_columns=['Education Level']

encoder=OneHotEncoder(drop=None,sparse_output=False)

encoded_data=encoder.fit_transform(saldf[categorical_columns])
print(encoded_data)
```

```
[[1. 0. 0. 0.]
 [0. 1. 0. 0.]
 [0. 0. 1. 0.]
 ...
 [1. 0. 0. 0.]
 [1. 0. 0. 0.]
 [0. 0. 1. 0.]]
```

the encodec data is in the form of array now we need to convert the encoded feature into a dataframe with catogaries as column names

```
encoded_df=pd.DataFrame(encoded_data,
columns=encoder.get_feature_names_out(categorical_columns))
encoded_df.head()
```

```
   Education Level_Bachelor's  Education Level_Master's  Education
Level_PhD  \
0                          1.0                       0.0
0.0
1                          0.0                       1.0
0.0
2                          0.0                       0.0
1.0
3                          1.0                       0.0
0.0
4                          1.0                       0.0
0.0

   Education Level_nan
0                  0.0
1                  0.0
2                  0.0
3                  0.0
4                  0.0
```

```
encoded_df.drop(['Education Level_nan'],axis=1,inplace=True)
encoded_df.head()
```

```
   Education Level_Bachelor's  Education Level_Master's  Education
Level_PhD
0                          1.0                       0.0
0.0
1                          0.0                       1.0
0.0
2                          0.0                       0.0
1.0
3                          1.0                       0.0
0.0
4                          1.0                       0.0
0.0
```

```
new_df=pd.concat([saldf,encoded_df],axis=1)
new_df.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
```

```
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary  Education Level_Bachelor's  Education Level_Master's  \
0   90000.0                         1.0                       0.0
1   65000.0                         0.0                       1.0
2  150000.0                         0.0                       0.0
3   60000.0                         1.0                       0.0
4   60000.0                         1.0                       0.0

   Education Level_PhD
0                  0.0
1                  0.0
2                  1.0
3                  0.0
4                  0.0
```

```
from sklearn.preprocessing import LabelEncoder

saldf1=pd.read_csv(r"C:\my pythonfiles\Salary_EDA.csv")
saldf1.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary
0   90000.0
1   65000.0
2  150000.0
3   60000.0
4   60000.0
```

```
le=LabelEncoder()
saldf1['Gender_encoded']=le.fit_transform(saldf['Gender'])
saldf1.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
```

```
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary  Gender_encoded
0   90000.0               1
1   65000.0               0
2  150000.0               1
3   60000.0               0
4   60000.0               0
```

```
le1=LabelEncoder()
saldf1['EL_encoded']=le1.fit_transform(saldf['Education Level'])
saldf1.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary  Gender_encoded  EL_encoded
0   90000.0               1           0
1   65000.0               0           1
2  150000.0               1           2
3   60000.0               0           0
4   60000.0               0           0
```

```
from sklearn.preprocessing import MinMaxScaler

saldf2=pd.read_csv(r"C:\my pythonfiles\Salary_EDA.csv")
saldf2.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
```

```
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary
0   90000.0
1   65000.0
2  150000.0
3   60000.0
4   60000.0
```

```
le2=MinMaxScaler()
saldf2['Salary_scalar']=le2.fit_transform(saldf2[['Salary']])
saldf2.head()
```

```
    Age  Gender Education Level          Job Title  Years of
Experience  \
0  32.0    Male      Bachelor's  Software Engineer
5.0
1  28.0  Female        Master's       Data Analyst
3.0
2  45.0    Male             PhD     Senior Manager
15.0
3  36.0  Female      Bachelor's    Sales Associate
7.0
4  36.0  Female      Bachelor's    Sales Associate
7.0

     Salary  Salary_scalar
0   90000.0       0.359103
1   65000.0       0.258963
2  150000.0       0.599439
3   60000.0       0.238935
4   60000.0       0.238935
```

Zscore normalization

```
s=np.array([24,33,23,28,37,35,14,48,43])
print(s.std())
s.mean()
```

```
9.977753031397178

31.666666666666668
```

```
from sklearn.preprocessing import StandardScaler

stdscl=StandardScaler()
saldf2['Salary_scalar']=scaler.fit_transform(saldf2[['Salary']])
saldf2.head()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call
last)
Cell In[52], line 2
      1 stdscl=StandardScaler()
----> 2
saldf2['Salary_scalar']=scaler.fit_transform(saldf2[['Salary']])
      3 saldf2.head()

NameError: name 'scaler' is not defined
```