Converting categorical data into numerical data

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

df=pd.read_csv(r"C:\Mypythonfiles\Salary_EDA.csv")

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

Filter Categorical features

```
categorical_cols=["Education Level"]
```

Define and apply Encoder

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

encoded_data=encoder.fit_transform(df[categorical_cols])

encoded_data
```

```
array([[1., 0., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 1., 0.],
       ...,
       [1., 0., 0., 0.],
       [1., 0., 0., 0.],
       [0., 0., 1., 0.]])
```

Convert the encoded data features intoa data namesframe eith the categories as column

- The encoded data is in the form of array. now we need to coinvert the encoded featured itno dataframe with categories as column names.

```
encode_df=pd.DataFrame(encoded_data,columns=encoder.get_feature_names_
out(categorical_cols))
encode_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
```

```
encode_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 375 entries, 0 to 374
Data columns (total 4 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Education Level_Bachelor's  375 non-null    float64
 1   Education Level_Master's    375 non-null    float64
 2   Education Level_PhD         375 non-null    float64
```

```
 3   Education Level_nan         375 non-null    float64
dtypes: float64(4)
memory usage: 11.8 KB
```

```
f_df=pd.concat([df,encode_df],axis=1)
```

```
f_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  Education Level_nan
0                  0.0                  0.0
1                  0.0                  0.0
2                  1.0                  0.0
3                  0.0                  0.0
4                  0.0                  0.0
```

Label Encooding

```
from sklearn.preprocessing import LabelEncoder
```

```
df1=pd.read_csv(r"C:\Mypythonfiles\Salary_EDA.csv")
```

```
df1
```

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

```
..    ...     ...             ...                                ...
370  35.0  Female      Bachelor's       Senior Marketing Analyst
371  43.0    Male        Master's         Director of Operations
372  29.0  Female      Bachelor's         Junior Project Manager
373  34.0    Male      Bachelor's  Senior Operations Coordinator
374  44.0  Female             PhD        Senior Business Analyst

     Years of Experience    Salary
0                    5.0   90000.0
1                    3.0   65000.0
2                   15.0  150000.0
3                    7.0   60000.0
4                    7.0   60000.0
..                   ...       ...
370                  8.0   85000.0
371                 19.0  170000.0
372                  2.0   40000.0
373                  7.0   90000.0
374                 15.0  150000.0

[375 rows x 6 columns]
```

```
df1.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()
```

```
df1["Gender_encoded"]=le.fit_transform(df["Gender"])
```

```
df1.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()
df1["Education_encoded"]=le1.fit_transform(df["Education Level"])

df1.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  Education_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
```

```
le2=LabelEncoder()
df1["Job Title_encoded"]=le2.fit_transform(df["Job Title"])
df1
```

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

```
1    28.0  Female        Master's                   Data Analyst
2    45.0    Male             PhD                Senior Manager
3    36.0  Female      Bachelor's               Sales Associate
4    36.0  Female      Bachelor's               Sales Associate
..    ...     ...             ...                           ...
370  35.0  Female      Bachelor's       Senior Marketing Analyst
371  43.0    Male        Master's         Director of Operations
372  29.0  Female      Bachelor's         Junior Project Manager
373  34.0    Male      Bachelor's  Senior Operations Coordinator
374  44.0  Female             PhD        Senior Business Analyst

     Years of Experience    Salary  Gender_encoded  Education_encoded
\
0                    5.0   90000.0               1                  0

1                    3.0   65000.0               0                  1

2                   15.0  150000.0               1                  2

3                    7.0   60000.0               0                  0

4                    7.0   60000.0               0                  0

..                   ...       ...             ...                ...

370                  8.0   85000.0               0                  0

371                 19.0  170000.0               1                  1

372                  2.0   40000.0               0                  0

373                  7.0   90000.0               1                  0

374                 15.0  150000.0               0                  2


     Job Title_encoded
0                  156
1                   17
2                  127
3                   98
4                   98
..                 ...
370                128
371                 29
372                 67
373                134
374                107

[375 rows x 9 columns]
```

```
df1.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  Education_encoded  Job Title_encoded
0   90000.0               1                  0                156
1   65000.0               0                  1                 17
2  150000.0               1                  2                127
3   60000.0               0                  0                 98
4   60000.0               0                  0                 98
```

```
from sklearn.preprocessing import MinMaxScaler

df2=pd.read_csv(r"C:\Mypythonfiles\Salary_EDA.csv")

df2
```

```
      Age  Gender Education Level                       Job Title  \
0    32.0    Male      Bachelor's              Software Engineer
1    28.0  Female        Master's                   Data Analyst
2    45.0    Male             PhD                 Senior Manager
3    36.0  Female      Bachelor's                Sales Associate
4    36.0  Female      Bachelor's                Sales Associate
..    ...     ...             ...                             ...
370  35.0  Female      Bachelor's       Senior Marketing Analyst
371  43.0    Male        Master's         Director of Operations
372  29.0  Female      Bachelor's         Junior Project Manager
373  34.0    Male      Bachelor's  Senior Operations Coordinator
374  44.0  Female             PhD        Senior Business Analyst

     Years of Experience    Salary
0                    5.0   90000.0
1                    3.0   65000.0
2                   15.0  150000.0
3                    7.0   60000.0
4                    7.0   60000.0
..                   ...       ...
370                  8.0   85000.0
371                 19.0  170000.0
372                  2.0   40000.0
```

```
373                   7.0   90000.0
374                  15.0  150000.0

[375 rows x 6 columns]
```

```
Scale=MinMaxScaler()

df2["Salary_SCaled"]=Scale.fit_transform(df[["Salary"]])
df2.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_SCaled
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
```

Z-Score Normalization

```
from sklearn.preprocessing import StandardScaler
Sts=StandardScaler()
df2["Salary_SCaled"]=Sts.fit_transform(df[["Salary"]])
df2.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_SCaled
0   90000.0      -0.211488
1   65000.0      -0.733148
2  150000.0       1.040496
3   60000.0      -0.837480
4   60000.0      -0.837480
```