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 \
  32.0
           Male
                     Bachelor's Software Engineer
5.0
  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
                     Bachelor's
4 36.0 Female
                                   Sales Associate
7.0
     Salary
0
    90000.0
1
    65000.0
2
  150000.0
3
    60000.0
    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
```

Convert the encoded data features into adata names frame 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 \
                           1.0
                                                      0.0
0
0.0
1
                           0.0
                                                      1.0
0.0
                           0.0
                                                      0.0
2
1.0
3
                           1.0
                                                      0.0
0.0
                                                      0.0
4
                           1.0
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
                                                   float64
                                  375 non-null
     Education Level Master's
                                  375 non-null
                                                   float64
 1
 2
     Education Level PhD
                                  375 non-null
                                                   float64
```

```
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
                                    Senior Manager
                            PhD
15.0
3 36.0 Female
                     Bachelor's
                                   Sales Associate
7.0
                     Bachelor's
                                   Sales Associate
4 36.0
        Female
7.0
             Education Level_Bachelor's
                                         Education Level Master's \
     Salary
    90000.0
                                    1.0
                                                              0.0
    65000.0
                                    0.0
                                                              1.0
1
2
                                    0.0
  150000.0
                                                              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
           Female
                         Master's
     28.0
                                                    Data Analyst
2
     45.0
             Male
                              PhD
                                                  Senior Manager
3
                                                 Sales Associate
     36.0
           Female
                       Bachelor's
4
     36.0 Female
                       Bachelor's
                                                 Sales Associate
```

```
Senior Marketing Analyst
370
    35.0
           Female
                       Bachelor's
371
    43.0
             Male
                         Master's
                                          Director of Operations
    29.0
372
           Female
                       Bachelor's
                                          Junior Project Manager
                                   Senior Operations Coordinator
373
    34.0
             Male
                       Bachelor's
                                         Senior Business Analyst
374
    44.0 Female
                              PhD
     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
                           85000.0
370
                     8.0
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
                     Bachelor's Software Engineer
           Male
5.0
                                      Data Analyst
1 28.0
        Female
                       Master's
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
    90000.0
0
1
    65000.0
2
   150000.0
3
    60000.0
    60000.0
le=LabelEncoder()
df1["Gender encoded"]=le.fit transform(df["Gender"])
df1.head()
```

```
Gender Education Level
                                        Job Title Years of
   Age
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
                                  Sales Associate
4 36.0 Female
                    Bachelor's
7.0
            Gender encoded
     Salary
0
   90000.0
1
   65000.0
                         0
2
                          1
  150000.0
3
   60000.0
                          0
   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 \
                    Bachelor's Software Engineer
0 32.0
          Male
5.0
1 28.0 Female
                                     Data Analyst
                      Master's
3.0
2 45.0
          Male
                            PhD
                                   Senior Manager
15.0
                    Bachelor's
3 36.0
        Female
                                  Sales Associate
7.0
4 36.0
        Female
                    Bachelor's
                                  Sales Associate
7.0
            Gender encoded
                            Education encoded
    Salary
   90000.0
0
                          1
                                            0
                                            1
1
   65000.0
                         0
                                            2
2
                          1
  150000.0
3
   60000.0
                         0
                                            0
                          0
   60000.0
le2=LabelEncoder()
df1["Job Title_encoded"]=le2.fit_transform(df["Job Title"])
df1
          Gender Education Level
                                                      Job Title \
     Age
                      Bachelor's
0
    32.0
            Male
                                              Software Engineer
```

1 2 3 4 370 371 372 373 374	28.0 45.0 36.0 36.0 35.0 43.0 29.0 34.0 44.0	Female Female Female Female Male Female Male Female	Ba Ba Ba Ba	Master's PhD chelor's chelor's Master's chelor's chelor's chelor's	Sa Senior Mark Director Junior Pr Senior Operation	Data Analyst enior Manager les Associate les Associate eting Analyst of Operations oject Manager s Coordinator iness Analyst	
\ 0	Years	of Expe	rience 5.0	Salary 90000.0	Gender_encoded	Education_enc	oded 0
1			3.0	65000.0	0		1
2			15.0 7.0	150000.0	1 0		2 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
0 1 2 3 4 370 371 372 373 374		itle_enco	156 17 127 98 98 128 29 67 134 107				
[375	rows	x 9 colur	mns]				

```
df1.head()
    Age Gender Education Level
                                          Job Title Years of
Experience
  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
                     Bachelor's
                                    Sales Associate
         Female
7.0
4 36.0
         Female
                     Bachelor's
                                    Sales Associate
7.0
             Gender encoded
                             Education encoded
                                                 Job Title encoded
     Salary
0
    90000.0
                                                                156
    65000.0
                          0
                                              1
1
                                                                 17
                                              2
2
   150000.0
                           1
                                                                127
3
    60000.0
                           0
                                              0
                                                                 98
                           0
                                                                 98
    60000.0
from sklearn.preprocessing import MinMaxScaler
df2=pd.read csv(r"C:\Mypythonfiles\Salary EDA.csv")
df2
           Gender Education Level
      Age
                                                         Job Title ∖
     32.0
             Male
                       Bachelor's
                                                Software Engineer
0
     28.0
           Female
                         Master's
                                                     Data Analyst
1
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
                                          Senior Business Analyst
                               PhD
     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
                          170000.0
371
                    19.0
                     2.0
372
                            40000.0
```

```
373
                     7.0
                           90000.0
                    15.0 150000.0
374
[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 \
                     Bachelor's Software Engineer
  32.0
          Male
5.0
1 28.0
        Female
                                      Data Analyst
                       Master's
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
                  0.359103
   90000.0
0
   65000.0
                  0.258963
1
2
                  0.599439
   150000.0
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
                    Bachelor's
                                  Sales Associate
3 36.0
        Female
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