Handle Categorical Features

One Hot Encoding

In [14]:

```
import pandas as pd
```

In [16]:

```
df=pd.read_csv('titanic.csv')
df
```

Out[16]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

891 rows × 12 columns

In [21]:

```
df=pd.read_csv('titanic.csv',usecols=['Sex'])
df
```

Out[21]:

	Sex
0	male
1	female
2	female
3	female
4	male
...	...
886	male
887	female
888	female
889	male
890	male

891 rows × 1 columns

In [22]:

```
df.head()
```

Out[22]:

	Sex
0	male
1	female
2	female
3	female
4	male

In [24]:

```
pd.get_dummies(df)
```

Out[24]:

	Sex_female	Sex_male
0	0	1
1	1	0
2	1	0
3	1	0
4	0	1
...	...	...
886	0	1
887	1	0
888	1	0
889	0	1
890	0	1

891 rows × 2 columns

In [ ]:

In [ ]:

In [9]:

```
pd.get_dummies(df,drop_first=True).head()
```

Out[9]:

	Sex_male
0	1
1	0
2	0
3	0
4	1

In [25]:

```
df=pd.read_csv('titanic.csv',usecols=['Embarked'])
df
```

Out[25]:

	Embarked
0	S
1	C
2	S
3	S
4	S
...	...
886	S
887	S
888	S
889	C
890	Q

891 rows × 1 columns

In [14]:

```
df['Embarked'].unique()
```

Out[14]:

```
array(['S', 'C', 'Q', nan], dtype=object)
```

In [16]:

```
df.dropna(inplace=True)
```

In [19]:

```
pd.get_dummies(df,drop_first=True).head()
```

Out[19]:

	Embarked_Q	Embarked_S
0	0	1
1	0	0
2	0	1
3	0	1
4	0	1

In [20]:

```
#### Onehotencoding with many categories in a feature
```

In [28]:

```
df=pd.read_csv('C:/Users/91920/Downloads/Compressed/Feature-Engineering-Live-sessions-maste
df
```

Out[28]:

	ID	y	X0	X1	X2	X3	X4	X5	X6	X8	...	X375	X376	X377	X378	X379	X38
0	0	130.81	k	v	at	a	d	u	j	o	...	0	0	1	0	0	
1	6	88.53	k	t	av	e	d	y	l	o	...	1	0	0	0	0	
2	7	76.26	az	w	n	c	d	x	j	x	...	0	0	0	0	0	
3	9	80.62	az	t	n	f	d	x	l	e	...	0	0	0	0	0	
4	13	78.02	az	v	n	f	d	h	d	n	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4204	8405	107.39	ak	s	as	c	d	aa	d	q	...	1	0	0	0	0	
4205	8406	108.77	j	o	t	d	d	aa	h	h	...	0	1	0	0	0	
4206	8412	109.22	ak	v	r	a	d	aa	g	e	...	0	0	1	0	0	
4207	8415	87.48	al	r	e	f	d	aa	l	u	...	0	0	0	0	0	
4208	8417	110.85	z	r	ae	c	d	aa	g	w	...	1	0	0	0	0	

4209 rows × 378 columns

In [4]:

```
df=pd.read_csv('C:/Users/91920/Downloads/Compressed/Feature-Engineering-Live-sessions-maste
```

In [5]:

```
df.head()
```

Out[5]:

	X0	X1	X2	X3	X4	X5	X6
0	k	v	at	a	d	u	j
1	k	t	av	e	d	y	l
2	az	w	n	c	d	x	j
3	az	t	n	f	d	x	l
4	az	v	n	f	d	h	d

In [30]:

```
df['X0'].unique()
```

Out[30]:

```
array(['k', 'az', 't', 'al', 'o', 'w', 'j', 'h', 's', 'n', 'ay', 'f', 'x',
       'y', 'aj', 'ak', 'am', 'z', 'q', 'at', 'ap', 'v', 'af', 'a', 'e',
       'ai', 'd', 'aq', 'c', 'aa', 'ba', 'as', 'i', 'r', 'b', 'ax', 'bc',
       'u', 'ad', 'au', 'm', 'l', 'aw', 'ao', 'ac', 'g', 'ab'],
      dtype=object)
```

In [10]:

```
df['X0'].value_counts()
```

Out[10]:

```
z     360
ak    349
y     324
ay    313
t     306
x     300
o     269
f     227
n     195
w     182
j     181
az    175
aj    151
s     106
ap    103
h      75
d      73
al     67
v      36
af     35
m      34
ai     34
e      32
ba     27
at     25
a      21
ax     19
am     18
i      18
aq     18
u      17
aw     16
l      16
ad     14
au     11
k      11
b      11
as     10
r      10
bc      6
ao      4
c       3
q       2
aa      2
ab      1
ac      1
g       1
Name: X0, dtype: int64
```

In [32]:

```
for i in df.columns:
    print(len(df[i].unique()))
```

```
47
27
44
7
4
29
12
```

KDD Orange Cup Compition

In [11]:

```
df.X1.value_counts().sort_values(ascending=False).head(10)
```

Out[11]:

```
aa    833
s     598
b     592
l     590
v     408
r     251
i     203
a     143
c     121
o      82
Name: X1, dtype: int64
```

In [12]:

```
lst_10=df.X1.value_counts().sort_values(ascending=False).head(10).index
lst_10=list(lst_10)
```

In [13]:

```
lst_10
```

Out[13]:

```
['aa', 's', 'b', 'l', 'v', 'r', 'i', 'a', 'c', 'o']
```

In [42]:

```
import numpy as np
for categories in lst_10:
    df[categories]=np.where(df['X1']==categories,1,0)
```

In [49]:

```
lst_10.append('X1')
```

In [50]:

```
df[lst_10]
```

Out[50]:

	aa	s	b	l	v	r	i	a	c	o	X1
0	0	0	0	0	1	0	0	0	0	0	v
1	0	0	0	0	0	0	0	0	0	0	t
2	0	0	0	0	0	0	0	0	0	0	w
3	0	0	0	0	0	0	0	0	0	0	t
4	0	0	0	0	1	0	0	0	0	0	v
5	0	0	1	0	0	0	0	0	0	0	b
6	0	0	0	0	0	1	0	0	0	0	r
7	0	0	0	1	0	0	0	0	0	0	l
8	0	1	0	0	0	0	0	0	0	0	s
9	0	0	1	0	0	0	0	0	0	0	b
10	0	0	0	0	0	1	0	0	0	0	r
11	0	0	0	0	0	1	0	0	0	0	r
12	0	0	1	0	0	0	0	0	0	0	b
13	0	0	0	0	0	1	0	0	0	0	r
14	0	1	0	0	0	0	0	0	0	0	s
15	0	0	0	1	0	0	0	0	0	0	l
16	0	0	0	0	0	1	0	0	0	0	r
17	1	0	0	0	0	0	0	0	0	0	aa
18	0	0	0	0	0	0	0	0	1	0	c
19	0	0	0	0	0	0	0	1	0	0	a
20	0	1	0	0	0	0	0	0	0	0	s
21	1	0	0	0	0	0	0	0	0	0	aa
22	0	0	0	0	0	1	0	0	0	0	r
23	0	0	1	0	0	0	0	0	0	0	b
24	0	0	0	0	0	1	0	0	0	0	r
25	0	1	0	0	0	0	0	0	0	0	s
26	0	0	0	1	0	0	0	0	0	0	l
27	1	0	0	0	0	0	0	0	0	0	aa
28	0	1	0	0	0	0	0	0	0	0	s
29	0	0	1	0	0	0	0	0	0	0	b
...	...	...	...	...	...	...	...	...	...	...	...
4179	0	0	0	0	0	0	0	1	0	0	a
4180	0	0	0	0	1	0	0	0	0	0	v
4181	0	0	0	0	0	1	0	0	0	0	r
4182	0	0	0	0	1	0	0	0	0	0	v

	aa	s	b	l	v	r	i	a	c	o	X1
4183	1	0	0	0	0	0	0	0	0	0	aa
4184	0	1	0	0	0	0	0	0	0	0	s
4185	0	0	0	0	1	0	0	0	0	0	v
4186	0	0	0	0	0	0	0	0	0	0	f
4187	0	0	0	1	0	0	0	0	0	0	l
4188	0	1	0	0	0	0	0	0	0	0	s
4189	0	1	0	0	0	0	0	0	0	0	s
4190	0	1	0	0	0	0	0	0	0	0	s
4191	0	0	0	0	1	0	0	0	0	0	v
4192	0	0	0	1	0	0	0	0	0	0	l
4193	0	0	0	0	1	0	0	0	0	0	v
4194	0	0	0	0	0	0	0	0	0	1	o
4195	0	0	0	1	0	0	0	0	0	0	l
4196	0	0	0	0	0	0	0	0	0	1	o
4197	0	0	0	0	1	0	0	0	0	0	v
4198	0	0	0	0	0	0	0	1	0	0	a
4199	1	0	0	0	0	0	0	0	0	0	aa
4200	1	0	0	0	0	0	0	0	0	0	aa
4201	0	0	0	0	1	0	0	0	0	0	v
4202	0	0	0	1	0	0	0	0	0	0	l
4203	0	1	0	0	0	0	0	0	0	0	s
4204	0	1	0	0	0	0	0	0	0	0	s
4205	0	0	0	0	0	0	0	0	0	1	o
4206	0	0	0	0	1	0	0	0	0	0	v
4207	0	0	0	0	0	1	0	0	0	0	r
4208	0	0	0	0	0	1	0	0	0	0	r

4209 rows × 11 columns

In [31]:

```
pwd
```

Out[31]:

```
'C:\\Users\\91920\\Machine Learning\\EDA'
```

In [ ]:

```

```