Categorical variable Encoding Dummy variales & One-Hot Encoding

In [2]:
```
import pandas as pd
```

In [7]:
```
tips_df=pd.read_csv(r"tips.csv")
tips_df
```

Out[7]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [17]:
```
tips_df.head()
```

Out[17]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

In [39]:
```
dummy_df= pd.get_dummies(tips_df)  # k dummy variable
dummy_df
```

Out[39]:

	total_bill	tip	size	sex_Female	sex_Male	smoker_No	smoker_Yes	day_Fri	day_Sat	day_Sun	day_Thu
0	16.99	1.01	2	1	0	1	0	0	0	1	
1	10.34	1.66	3	0	1	1	0	0	0	1	
2	21.01	3.50	3	0	1	1	0	0	0	1	
3	23.68	3.31	2	0	1	1	0	0	0	1	
4	24.59	3.61	4	1	0	1	0	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	...
239	29.03	5.92	3	0	1	1	0	0	1	0	
240	27.18	2.00	2	1	0	0	1	0	1	0	
241	22.67	2.00	2	0	1	0	1	0	1	0	
242	17.82	1.75	2	0	1	1	0	0	1	0	
243	18.78	3.00	2	1	0	1	0	0	0	0	

244 rows × 13 columns

```
pd.get_dummies(
    data,
    prefix=None,
    prefix_sep='_',
    dummy_na=False,
    columns=None,
    sparse=False,
```

```
    drop_first=False, #k-1 maatlab 1 hi dummay variable lena
    dtype=None,
```

In [41]:
```
dummy_df= pd.get_dummies(tips_df,drop_first=True)  # k dummy variable k-1
dummy_df
```

Out[41]:

	total_bill	tip	size	sex_Male	smoker_Yes	day_Sat	day_Sun	day_Thur	time_Lunch
0	16.99	1.01	2	0	0	0	1	0	0
1	10.34	1.66	3	1	0	0	1	0	0
2	21.01	3.50	3	1	0	0	1	0	0
3	23.68	3.31	2	1	0	0	1	0	0
4	24.59	3.61	4	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...
239	29.03	5.92	3	1	0	1	0	0	0
240	27.18	2.00	2	0	1	1	0	0	0
241	22.67	2.00	2	1	1	1	0	0	0
242	17.82	1.75	2	1	0	1	0	0	0
243	18.78	3.00	2	0	0	0	0	1	0

244 rows × 9 columns

One-Hot- encoding with sklearn

In [42]:
```
from sklearn.preprocessing import OneHotEncoder
```

```
# OneHotEncoder(
    *,
    categories='auto',
```

```
    drop=None,
    sparse=True,
    dtype=<class 'numpy.float64'>,
    handle_unknown='error',
)
```

In [43]:
```
oh_enc=OneHotEncoder(sparse=False,drop="first")
```

In [44]:
```
oh_enc_arr=oh_enc.fit_transform(tips_df[['sex','smoker','day','time']])
oh_enc_arr
```

Out[44]:
```
array([[0., 0., 0., 1., 0., 0.],
       [1., 0., 0., 1., 0., 0.],
       [1., 0., 0., 1., 0., 0.],
       ...,
       [1., 1., 1., 0., 0., 0.],
       [1., 0., 1., 0., 0., 0.],
       [0., 0., 0., 0., 1., 0.]])
```

In [45]:
```
dummy_df.keys()
```

Out[45]:
```
Index(['total_bill', 'tip', 'size', 'sex_Male', 'smoker_Yes', 'day_Sat',
       'day_Sun', 'day_Thur', 'time_Lunch'],
      dtype='object')
```

In [48]:
```
oh_enc_df=pd.DataFrame(oh_enc_arr,columns=['sex_Male', 'smoker_Yes', 'day_Sat',
       'day_Sun', 'day_Thur', 'time_Lunch'])
```

In [50]:
```
oh_enc_df
```

Out[50]:

	sex_Male	smoker_Yes	day_Sat	day_Sun	day_Thur	time_Lunch
0	0.0	0.0	0.0	1.0	0.0	0.0
1	1.0	0.0	0.0	1.0	0.0	0.0
2	1.0	0.0	0.0	1.0	0.0	0.0
3	1.0	0.0	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	1.0	0.0	0.0
...	...	...	...	...	...	...
239	1.0	0.0	1.0	0.0	0.0	0.0
240	0.0	1.0	1.0	0.0	0.0	0.0
241	1.0	1.0	1.0	0.0	0.0	0.0
242	1.0	0.0	1.0	0.0	0.0	0.0
243	0.0	0.0	0.0	0.0	1.0	0.0

244 rows × 6 columns