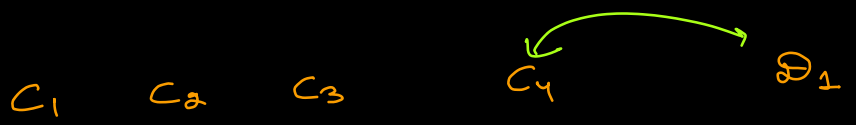
EDA projects



dependant variable

uni → 1
bi → 2
multi → 2+

```python
from statsmodels.stats.outliers_influence import variance_inflation_factor
def compute_vif(considered_features, df):

    X = df[considered_features]
    # the calculation of variance inflation requires a constant
    X['intercept'] = 1

    # create dataframe to store vif values
    vif = pd.DataFrame()
    vif["Variable"] = X.columns
    vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
    vif = vif[vif['Variable']!='intercept']
    return vif
```

```python
compute_vif(num_features, df)
```

	Variable	VIF
0	vehicle_age	1.406352
1	km_driven	1.212640
2	mileage	1.945103
3	engine	6.244006
4	max_power	5.952622
5	seats	2.245733
6	selling_price	2.680638

HN

min

Q_1 25%.

Q_2 50%.

Q_3 75%.

max

IQR = $Q_3 - Q_3$

M → Male → 0

F → Female → 1

A → 0

B → 1

C → 2

M → 1

F → 2 → ⇒ 2 ⇐

O → 0

income

1000

2000 →

500

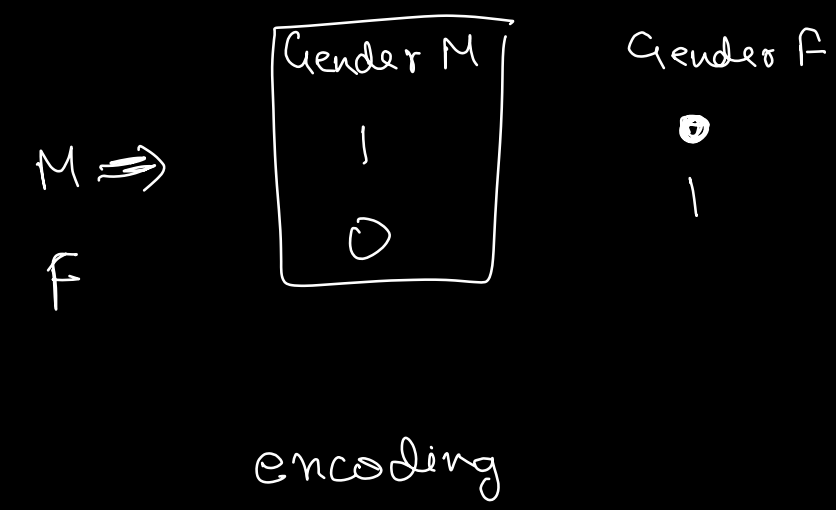
Genders	Gender
M	0
F	1
O	2

Gender_M	Gender_F	Gender_O
1	0	0
0	1	0
0	0	1

⇓

	Gender_F	Gender_O
⇒	0	0
	1	0
	0	1

	M	F	O
M	1	0	0
F	0	1	0
O	0	0	1



encoding