Ex2: PCA - sklearn

- Cho dữ liệu student.xlsx.
- Đọc dữ liệu vào dataframe.
- Thực hiện giảm chiều dữ liệu với sklearn.PCA
- Trực quan hóa dữ liệu

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.decomposition import PCA
```

In [2]:

```
data = pd.read_excel("student.xlsx", index_col=0)
data.head()
```

Out[2]:

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

In [3]:

```
pca = PCA(2)
pca.fit(data)
```

Out[3]:

```
PCA(copy=True, iterated_power='auto', n_components=2, random_state=None,
  svd_solver='auto', tol=0.0, whiten=False)
```

In [4]:

```
print(pca.components_)
print(pca.components_.shape)
print(pca.explained_variance_)
print(pca.explained_variance_.shape)
```

```
[[-0.59862919 -0.51336438 -0.61489845]
 [ 0.47005554  0.39643891 -0.78859621]]
(2, 3)
[605.64181179 313.26463747]
(2,)
```

In [5]:

```
B = pca.transform(data)
B[0:5]
```

Out[5]:

```
array([[-28.71093503, -11.33365494],
       [ -7.21795959,  47.87528492],
       [  7.69489417,  -1.77743486],
       [-10.75205928, -25.43532109],
       [ 59.50165485,  -4.11438216]])
```

In [6]:

```
pca.explained_variance_ratio_
```

Out[6]:

```
array([0.57863867, 0.29929742])
```

In [7]:

```
principalDf = pd.DataFrame(data = B
             , columns = ['principal component 1', 'principal component 2'])
principalDf.head()
```

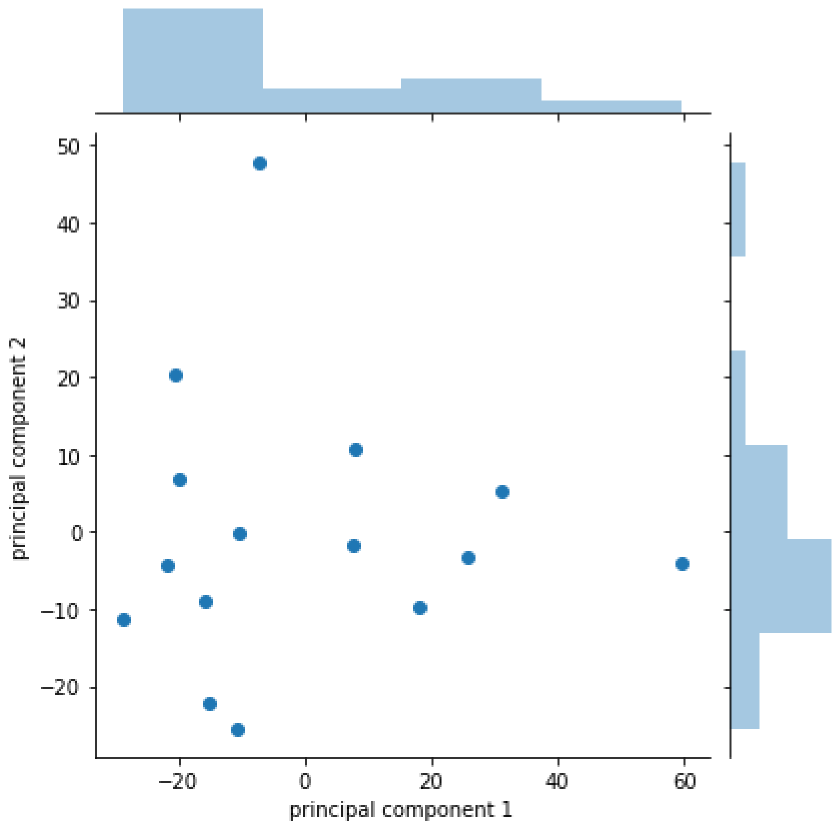
Out[7]:

	principal component 1	principal component 2
0	-28.710935	-11.333655
1	-7.217960	47.875285
2	7.694894	-1.777435
3	-10.752059	-25.435321
4	59.501655	-4.114382

In [8]:

```
plt.figure(figsize=(8,6))
sns.jointplot(x='principal component 1', y='principal component 2', data = princip
plt.show()
```

```
<Figure size 576x432 with 0 Axes>
```


principal component 2
principal component 1