```
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
from matplotlib import pyplot as plt

import kagglehub

# Download latest version
path = kagglehub.dataset_download("harrywang/wine-dataset-for-
clustering")

print("Path to dataset files:", path)
```

```
Downloading from
https://www.kaggle.com/api/v1/datasets/download/harrywang/wine-
dataset-for-clustering?dataset_version_number=1...

100%|██████████| 4.36k/4.36k [00:00<00:00, 7.50MB/s]

Extracting files...
Path to dataset files: /root/.cache/kagglehub/datasets/harrywang/wine-
dataset-for-clustering/versions/1
```

```
x=pd.read_csv("/root/.cache/kagglehub/datasets/harrywang/wine-dataset-
for-clustering/versions/1/wine-clustering.csv")
print(x)
```

```
     Alcohol  Malic_Acid   Ash  Ash_Alcanity  Magnesium  Total_Phenols
\
0      14.23        1.71  2.43          15.6        127           2.80

1      13.20        1.78  2.14          11.2        100           2.65

2      13.16        2.36  2.67          18.6        101           2.80

3      14.37        1.95  2.50          16.8        113           3.85

4      13.24        2.59  2.87          21.0        118           2.80

..       ...         ...   ...           ...        ...            ...

173    13.71        5.65  2.45          20.5         95           1.68

174    13.40        3.91  2.48          23.0        102           1.80

175    13.27        4.28  2.26          20.0        120           1.59

176    13.17        2.59  2.37          20.0        120           1.65

177    14.13        4.10  2.74          24.5         96           2.05
```

```
     Flavanoids  Nonflavanoid_Phenols  Proanthocyanins
Color_Intensity   Hue  \
0            3.06                  0.28             2.29
5.64  1.04
1            2.76                  0.26             1.28
4.38  1.05
2            3.24                  0.30             2.81
5.68  1.03
3            3.49                  0.24             2.18
7.80  0.86
4            2.69                  0.39             1.82
4.32  1.04
..            ...                   ...              ...                ..
.   ...
173          0.61                  0.52             1.06
7.70  0.64
174          0.75                  0.43             1.41
7.30  0.70
175          0.69                  0.43             1.35
10.20  0.59
176          0.68                  0.53             1.46
9.30  0.60
177          0.76                  0.56             1.35
9.20  0.61

     OD280  Proline
0     3.92     1065
1     3.40     1050
2     3.17     1185
3     3.45     1480
4     2.93      735
..     ...      ...
173   1.74      740
174   1.56      750
175   1.56      835
176   1.62      840
177   1.60      560

[178 rows x 13 columns]
```

```
kmeans=KMeans(n_clusters=4,random_state=0)
kmeans.fit(x)
labels=kmeans.labels_
print(labels)
```

```
[1 1 1 3 2 3 1 1 1 1 3 1 3 1 3 3 1 1 3 2 2 2 1 1 2 2 1 1 2 1 1 3 1 1 1
2 2
 1 1 2 2 1 1 2 2 1 1 1 1 1 1 1 1 3 1 1 1 1 1 0 2 0 2 0 0 2 0 0 2 2 2 0
```

```
0 1
 2 0 0 0 2 0 0 2 2 0 0 0 0 0 2 2 0 0 0 0 0 1 2 0 2 0 2 0 0 0 2 0 0 0 0
2 0
 0 2 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 2 0 0 2 2 2 2 0 0 2 2 2 0 0 2 2
0 2
 2 0 0 0 0 2 2 2 0 2 2 2 0 2 0 2 2 0 2 2 2 2 0 0 2 2 2 2 2 0]
```

```
plt.scatter(x.iloc[:,6], x.iloc[:,12], c=labels)
plt.scatter(
    kmeans.cluster_centers_[:,6],
    kmeans.cluster_centers_[:,12],
    marker='x',
    s=200
)
plt.xlabel("Magnesium")
plt.ylabel("Proline")
plt.title("K-Means Clustering (K=4)")
plt.show()
```