

===== 原始資料（未降維）分析 =====

Silhouette = 0.4540

===== PCA 降維分析 =====

PCA 1D → silhouette = 0.6118

PCA 2D → silhouette = 0.4540

===== 最佳 PCA 維度 =====

最佳維度：1D(silhouette = 0.6118)

↑注：重新進行 K-means ($n=3$) 分群，計算降維後的 Silhouette score。
具體怎麼做 用說的 不用程式

下面我用「老師想聽得懂、你能在作業或口頭簡報上講得清楚」的方式說明整個流程，完全不用程式碼。

整體流程概念

你要做的事情分成兩大部分：

1. 原始資料 → K-means → 算 silhouette score
2. PCA 降維 → K-means → 算 silhouette → 選出最好的維度

這兩個步驟做完，就是完整的 PCA + K-means 分群分析。

1 原始資料分析（未降維）

你的 students_200.csv 有兩個特徵：

- 國文分數
- 英文分數

所以它本質上是「2 維資料」。

你要做的事情是：

(1) 準備資料

把所有學生的 (chinese, english) 組成資料矩陣 X。

(2) 執行 K-means (K=3)

因為分 3 群，所以 K=3。

你最後會得到：

複製

```
原始資料 silhouette score = S_raw
```

2 PCA 降維分析

你要測試多個降維版本 (例如 PCA 1 維 / 2 維 / 3 維)。

但你的資料只有兩個欄位，所以最多只能做到 2 維：

- PCA 1 維
- PCA 2 維 (其實 = 原始資料)

PCA 過程 (概念)：

PCA 會做兩件事：

1. 找出資料最有變化方向 (主成分)
2. 投影到較低維空間

3 PCA 之後重新做 K-means

每一個 PCA 綴度，你都要重做一次 K-means 分群。

例如：

PCA → 1 維

- 原本 (chinese, english) 變成一個數字
- 用這個 1 維資料做 K-means
- 算 silhouette score → S1



訪問作者問題

- 算 silhouette score → S1

PCA → 2 維

- 其實跟原始一樣 (因為本來就兩欄)

- 得到 silhouette score → S2

如果你有 3 維資料 (例如三科成績) , 就會測試到 PCA=3 。

4 比較所有 silhouette score

你將得到 :

降維維度	silhouette score
原始資料	S_raw
PCA 1 維	S1
PCA 2 維	S2

選擇分數最高的那一個，那個就是最佳 PCA 維度。

✓ 你的最後分析要回答三件事 :

(A) 原始資料 (未降維) 結果如何 ?

例如 :

- K-means 後分群合理嗎 ?
- silhouette score 是多少 ?

(B) 降維後效果有改善嗎 ?



- K-means 後分群合理嗎？
- silhouette score 是多少？

(B) 降維後效果有改善嗎？

例如：

- PCA 1 維可能把資料壓扁 → silhouette score 下降
- PCA 2 維與原本相同 → score 不變

(C) 最佳 PCA 維度是哪一個？

通常會是：

- 如果資料本來只有 2 個欄位 → 最好維度是 2 (等於不降維)
- 如果資料維度很高 → 常常是 2 或 3 維最好

你會得到一句結論，例如：

| 「在本資料中，PCA 降到 1 維會損失太多資訊，使 K-means 分群效果變差，因此最佳維度為 2 維，與原始資料一致。」

✗ 要不要我幫你寫成「可交作業版本」的正式文字？

我可以替你寫成：

- 報告格式
- 有條列式、小標題
- 有 PCA & silhouette score 的概念圖
- 有結果解讀

從原本的 `kmean` 讀資料的程式下去改
先從原始資料（`2d`）且 `k=3` 去算 Silhouette
之後投影 `2d` 和 `1d` 去算降維後的資料分界度