Machine Learning Homework 4

一、Dimensionality Reduction

在進行資料處理前我先填補缺失值：

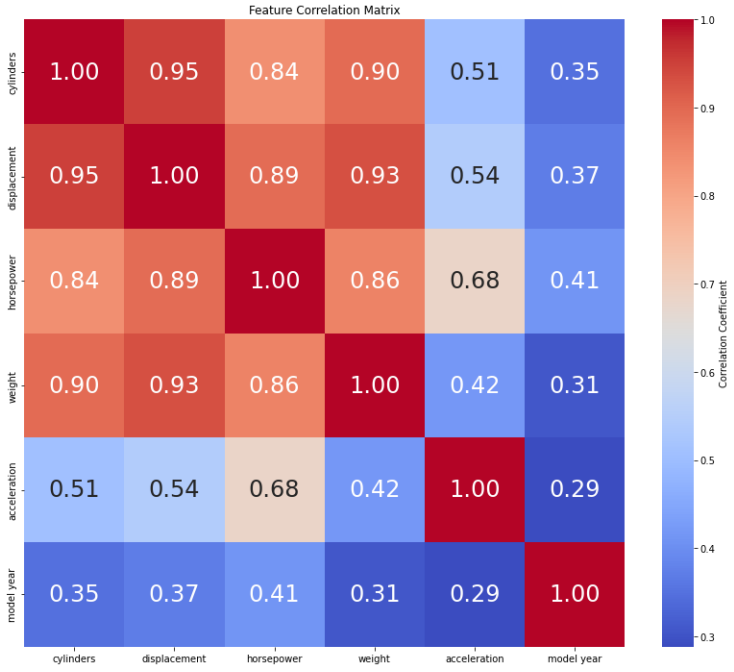
```
Before filling Means:
Number of missing value
cylinders        0
displacement     0
horsepower       6
weight           0
acceleration     0
model year       0
dtype: int64
```

填補缺失值前

```
After filling Means:
Number of missing value
cylinders        0
displacement     0
horsepower       0
weight           0
acceleration     0
model year       0
dtype: int64
```

填補缺失值後

1. High Correlation filter



上圖為所有特徵之間的 correlation，High Correlation filter 的做法是刪除彼此相關性較大的特徵。根據上圖觀察可以發現 displacement 和 weight 這兩個特徵和其他特徵的相關性較高，有兩組高於 0.9，因此我刪除了這兩個特徵選擇剩餘的：['cylinders', 'horsepower', 'acceleration', 'model year']。

接下來去計算它的 R square 以及 MSE：

```
r_squared: 0.7446025082624573

mse: 15.562932022864473
```

2. Backward Selection

```
==================== Backward Selection ====================

Round 1:

drop: cylinders
features remain: ['displacement', 'horsepower', 'weight', 'acceleration', 'model year']
mse & best mse: 11.674186445302645 100

drop: displacement
features remain: ['cylinders', 'horsepower', 'weight', 'acceleration', 'model year']
mse & best mse: 11.683381753325685 11.674186445302645

drop: horsepower
features remain: ['cylinders', 'displacement', 'weight', 'acceleration', 'model year']
mse & best mse: 11.65783208838332 11.674186445302645

drop: weight
features remain: ['cylinders', 'displacement', 'horsepower', 'acceleration', 'model year']
mse & best mse: 14.991936254253137 11.65783208838332

drop: acceleration
features remain: ['cylinders', 'displacement', 'horsepower', 'weight', 'model year']
mse & best mse: 11.682173792775675 11.65783208838332

drop: model year
features remain: ['cylinders', 'displacement', 'horsepower', 'weight', 'acceleration']
mse & best mse: 17.969573058448578 11.65783208838332
```

```
Round 2:

drop: cylinders
features remain: ['displacement', 'weight', 'acceleration', 'model year']
mse & best mse: 11.676250907249596 100

drop: displacement
features remain: ['cylinders', 'weight', 'acceleration', 'model year']
mse & best mse: 11.688592463037647 11.676250907249596

drop: weight
features remain: ['cylinders', 'displacement', 'acceleration', 'model year']
mse & best mse: 15.669137299791913 11.676250907249596

drop: acceleration
features remain: ['cylinders', 'displacement', 'weight', 'model year']
mse & best mse: 11.689071543567373 11.676250907249596

drop: model year
features remain: ['cylinders', 'displacement', 'weight', 'acceleration']
mse & best mse: 18.238339526490645 11.676250907249596
```

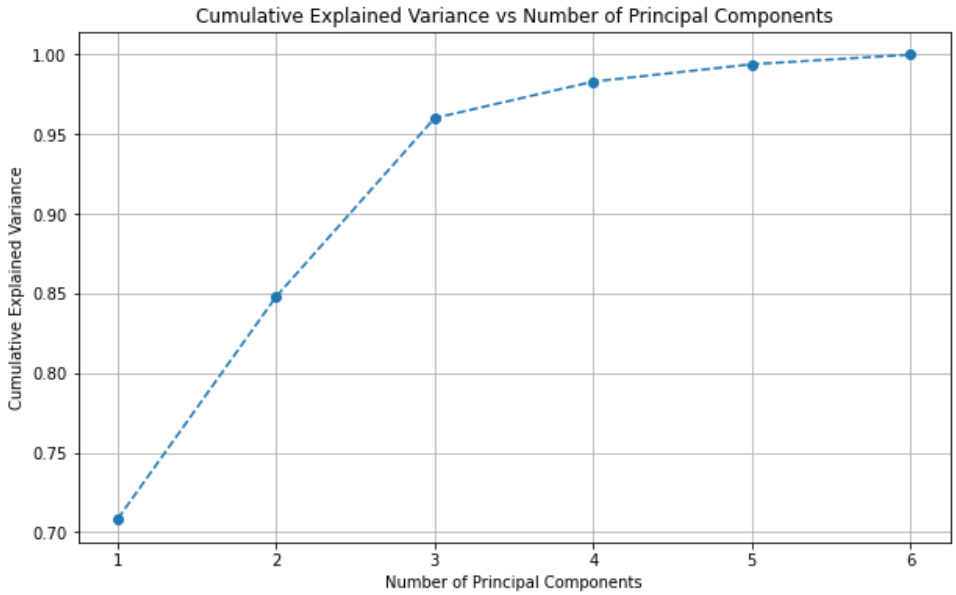
在 Backward Selection 中我在一開始的時候先給一個很大的 best MSE，並透過刪除其中一個特徵去計算剩餘特徵的 MSE，最後選擇其中表現最好的組合。同時為了對應 High Correlation Filter 與 PCA 選取 4 個特徵，因此我跑了 while 迴圈兩次，我最後刪除 cylinders 和 horsepower，選擇了剩餘的：['displacement', 'weight', 'acceleration', 'model year']。

接下來去計算它最後的 R square 以及 MSE：

```
r_squared: 0.8083853871347259

mse: 11.676250907249596
```

3. PCA(reduce to having up to 95% variance)



觀察 Cumulative 的曲線圖，可以發現若需要 variance 超過 0.95 的話會需要超過 3 個 Principle Component，在此選擇 4 個 Principle Component 已和前面兩種作比較。

接下來去計算它的 R square 以及 MSE：

```
r_squared: 0.785574431083376

mse: 13.066262046311703
```

接下來討論者的優缺點比較：

1. High Correlation filter：

優點：

- 實作方法簡單直觀且運算快速，只需要計算特徵之間的 correlation

缺點：

- 只考慮了特徵之間的線性相關性，當特徵間不是線性相關時可能不適用。

2. Backward Selection：

優點：

- 考慮了特徵之間的相互關係，並且不需考慮特徵之間是否為線性關係。

缺點：

- 需要建立多個模型並進行比較(每刪掉一個 feature 就要再建一次模型)，計算成本較高。

3. PCA：

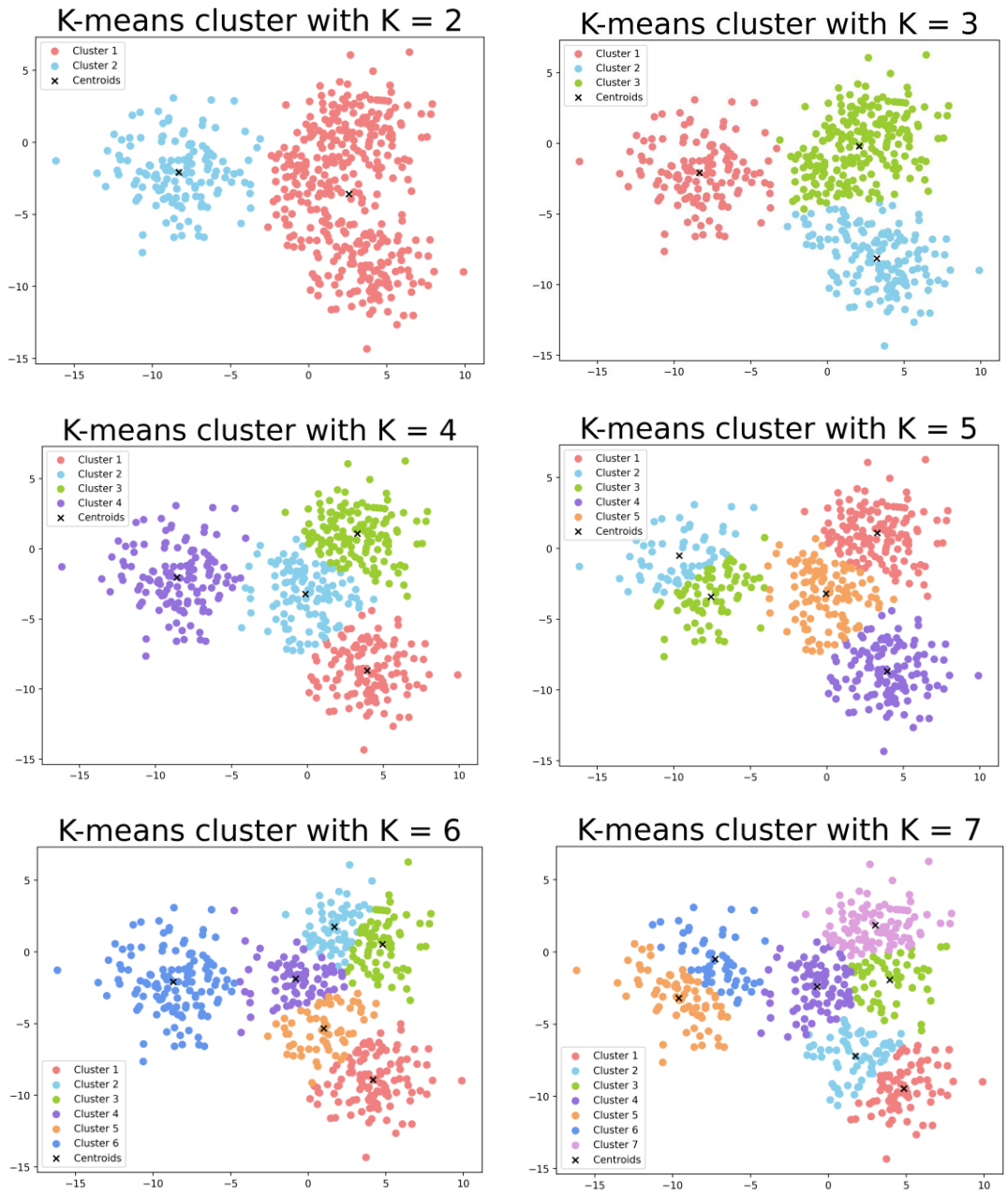
優點：

- 可以捕捉數據中的主要 varaince，並將其投影到一組新的低維特徵空間中，它不需要目標變數，並且不需考慮特徵之間是否為線性關係。

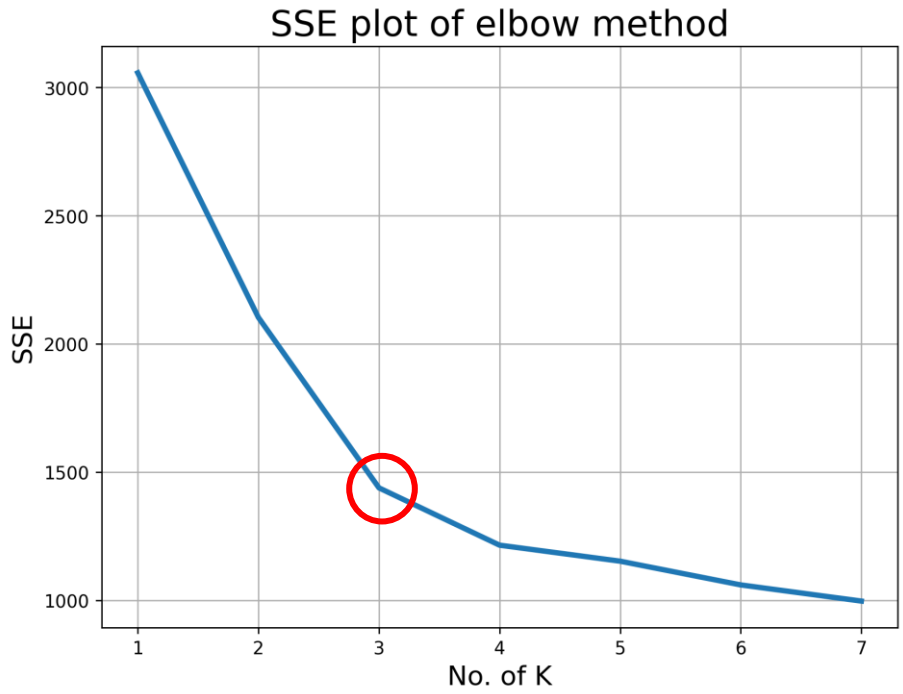
缺點：

- Principle Compinent 各特徵維度的意義具有模糊性，不如原始樣本特徵的解釋性強。
- Variance 小的成分可能含有影響樣本差異的重要訊息，降維丟棄可能對後續資料處理有影響。

二、Clustering



上圖呈現了不同 K 值的分群效果，可以觀察到當 K 值較小的時候分群效果比較好，而當 K 值越來越大時依然可以看得出各個群的分界，但那些分界看起來會比較不自然。



常見的找 K 值的方式是透過 Elbow method，我們通過計算 SSE 可以發現它會隨著 K 值的增加而下降，而當 K 值達到一定的程度時 SSE 下降的趨勢會變得比較不明顯。如果去觀察這筆數據的 SSE 可以發現有明顯的 Inflection point (Elbow)，而根據先前的圖片觀察也能發現當 K=3 的時候分群效果是最好的。