練習一

請用上述方法觀察 rules 資料表，並從中找出你認為最有用的規則，並解釋原因

選擇指標

我認為最有用的規則，應為適用範圍最廣且最具實際價值的規則，因此選擇以 **lift(提升度)** 與 **support(支持度)** 作為指標。

lift (提升度)：最重要指標

- 公式為 $\text{lift}(\text{lhs} \rightarrow \text{rhs}) = \frac{\text{confidence}(\text{lhs} \rightarrow \text{rhs})}{\text{support}(\text{rhs})}$
- 衡量規則中 lhs 和 rhs 是否真正具有關聯性。(評估價值的核心指標)
- lift > 1 表示 lhs 和 rhs 具有正向關聯。(lift = 1 表示規則只是隨機共現，無實際意義)
- 提升度越高，關聯性越強。

support (支持度)：次重要指標

- 公式為 $\text{support}(\text{lhs} \rightarrow \text{rhs}) = \frac{\text{count}(\text{lhs} \cap \text{rhs})}{\text{total transactions}}$
- 數據中滿足規則的項目數占總數的比例。(反映規則的適用範圍)
- 支持度越高，規則越具有代表性。
- 支持度過低可能表示規則只適用於小眾情況，實用價值有限。

取出最有用的規則

```
inspect(rules)
rules_with_score <- apriori(Adult, parameter = list(support = 0.5, confidence = 0.9))
quality(rules_with_score)$Score <- quality(rules_with_score)$support * 0.3 + quality(rules_with_score)$lift * 0.7
rules_with_score_sorted <- sort(rules_with_score, by = "Score")
inspect(rules_with_score_sorted)
```

觀察 rules 資料表(52項規則)後，我將相同的內容存進 rules_with_score (避免動到原 rules 資料)，再以 **support 權重 0.3、lift 權重 0.7** 得出的分數進行排序，最後 inspect 完整資料表。如前述， lift > 1 才表示 lhs 和 rhs 具有正向關聯，而資料中**第一個 lift > 1** 的規則為第六項。

```
     lhs                                rhs                                support   confidence  coverage   lift     count score
[1]  {}                              => {capital-loss=None}                0.9532779  0.9532779 1.0000000 1.0000000 46560 0.9859834
[2]  {}                              => {capital-gain=None}                0.9173867  0.9173867 1.0000000 1.0000000 44807 0.9752160
[3]  {capital-loss=None}             => {capital-gain=None}                0.8706646  0.9133376 0.9532779 0.9955863 42525 0.9581098
[4]  {capital-gain=None}             => {capital-loss=None}                0.8706646  0.9490705 0.9173867 0.9955863 42525 0.9581098
[5]  {native-country=United-States} => {capital-loss=None}                 0.8548380  0.9525461 0.8974243 0.9992323 41752 0.9559140
[6]  {race=White}                    => {native-country=United-States}     0.7881127  0.9217231 0.8550428 1.0270761 38493 0.9553871
[7]  {native-country=United-States} => {capital-gain=None}                 0.8219565  0.9159062 0.8974243 0.9983862 40146 0.9454573
[8]  {race=White}                    => {capital-loss=None}                0.8136849  0.9516307 0.8550428 0.9982720 39742 0.9428959
[9]  {race=White,
      capital-loss=None}             => {native-country=United-States}     0.7490480  0.9205626 0.8136849 1.0257830 36585 0.9427625
[10] {race=White,
      capital-gain=None}             => {native-country=United-States}     0.7194628  0.9202807 0.7817862 1.0254689 35140 0.9336670
```

我取出最有用的規則為**「在種族為白人的條件下，其出生地為美國」**。

練習二

嘗試調整最低 support 與 confidence 產生新規則，並從中找尋有用的規則

調整後有用的規則

```
rules_new <- apriori(Adult, parameter = list(support = 0.45, confidence = 0.85))
inspect(rules_new)
quality(rules_new)$Score <- quality(rules_new)$support * 0.3 + quality(rules_new)$lift * 0.7
rules_new_sorted <- sort(rules_new, by = "Score")
inspect(rules_new_sorted)
```

我將 **support 調低成 0.45**、**confidence 調低成 0.85** 後，觀察 rules_new 資料表(116項規則)，並一樣以練習一的公式 **support 權重 0.3**、 **lift 權重 0.7** 得出的分數進行排序，最後 inspect 完整資料表。我把**前20項**規則中 **lift > 1** 的六項視為有用的規則。

```
[9]   {race=White}                    => {native-country=United-States} 0.7881127  0.9217231 0.8550428 1.0270761 38493 0.9553871
[10]  {native-country=United-States} => {race=White}                    0.7881127  0.8781940 0.8974243 1.0270761 38493 0.9553871
[15]  {race=White,
       capital-loss=None}             => {native-country=United-States} 0.7490480  0.9205626 0.8136849 1.0257830 36585 0.9427625
[16]  {capital-loss=None,
       native-country=United-States} => {race=White}                    0.7490480  0.8762454 0.8548380 1.0247972 36585 0.9420724
[17]  {race=White,
       capital-gain=None}             => {native-country=United-States} 0.7194628  0.9202807 0.7817862 1.0254689 35140 0.9336670
[18]  {capital-gain=None,
       native-country=United-States} => {race=White}                    0.7194628  0.8753051 0.8219565 1.0236975 35140 0.9324271
```

我取出有用的規則分別為**「在種族為白人的條件下，其出生地為美國」、「在出生地為美國的條件下，其種族為白人」、「在種族為白人且資本虧損為零的條件下，其出生地為美國」、「在資本虧損為零且出生地為美國的條件下，其種族為白人」、「在種族為白人且資本收益為零的條件下，其出生地為美國」、「在資本收益為零且出生地為美國的條件下，其種族為白人」**。