

410979068_施尚丞_HW4

Shih

2025-03-17

本周作業目標

本次作業同樣使用 SENIC 計畫數據，延續上周的內容，加上3D散佈圖、熱力圖以及K-means分群，觀察哪些變數可能對於分群有較大的影響。

資料說明

根據APPENC該數據集包含來自 113 家醫院的 12 個變數，SENIC 計畫主要在判定感染監事及控制方案是否真能減少醫院內的感染率，資料集中包含了338家醫院中隨機選出的113家，研究期間為1975年至1976年。

以下為12個變數的詳細資訊：

- **編號**：醫院編號
- **停留日數**：病患平均停留日數
- **年齡**：病患平均年齡
- **感染風險**：在醫院內感染之平均推估機率（%）
- **培養率**：培養數對於無院內感染病患數之比率（乘以 100）
- **X 光照射率**：X 光照射數對於無肺炎病患數之比率（乘以 100）
- **病床數**：醫院的平均病床數
- **醫校合作**：1 表示有醫學院合作，2 表示無
- **區域**：1=東北，2=北，3=南，4=西
- **平均病患數**：每日病患人數
- **護士數目**：研究期間平均護士人數
- **可用設備**：35 項設備中醫院可提供之百分比

資料讀取與前處理

```
library(dplyr)

# 讀取資料
data <- read.table("C:/Users/SHI/Desktop/R/113-2/APPENC01.txt", header = FALSE, sep = "", strip.white = TRUE)

# 指定欄位名稱
colnames(data) <- c("編號", "停留日數", "年齡", "感染風險", "培養率", "X光照射率", "病床數", "醫校合作", "區域", "平均病患數", "護士數目", "可用設備")

# 查看資料結構
str(data)
```

```
## 'data.frame':   113 obs. of  12 variables:
## $ 編號       : int  1 2 3 4 5 6 7 8 9 10 ...
## $ 停留日數   : num  7.13 8.82 8.34 8.95 11.2 ...
## $ 年齡       : num  55.7 58.2 56.9 53.7 56.5 50.9 57.8 45.7 48.2 56.3 ...
## $ 感染風險   : num  4.1 1.6 2.7 5.6 5.7 5.1 4.6 5.4 4.3 6.3 ...
## $ 培養率     : num  9 3.8 8.1 18.9 34.5 21.9 16.7 60.5 24.4 29.6 ...
## $ X光照射率  : num  39.6 51.7 74 122.8 88.9 ...
## $ 病床數     : int  279 80 107 147 180 150 186 640 182 85 ...
## $ 醫校合作   : int  2 2 2 2 2 2 2 1 2 2 ...
## $ 區域       : int  4 2 3 4 1 2 3 2 3 1 ...
## $ 平均病患數 : int  207 51 82 53 134 147 151 399 130 59 ...
## $ 護士數目   : int  241 52 54 148 151 106 129 360 118 66 ...
## $ 可用設備   : num  60 40 20 40 40 40 40 60 40 40 ...
```

```
# 檢查缺失值
table(is.na(data))
```

```
##
## FALSE
## 1356
```

```
# 轉換類別變數
data$醫校合作 <- as.factor(data$醫校合作)
data$區域 <- as.factor(data$區域)
```

```
head(data)
```

```
##   編號  停留日數  年齡  感染風險  培養率  x光照射率  病床數  醫校合作  區域  平均病患數
## 1     1      7.13 55.7     4.1     9.0      39.6     279         2     4         207
## 2     2      8.82 58.2     1.6     3.8      51.7      80         2     2          51
## 3     3      8.34 56.9     2.7     8.1      74.0     107         2     3          82
## 4     4      8.95 53.7     5.6    18.9     122.8     147         2     4          53
## 5     5     11.20 56.5     5.7    34.5      88.9     180         2     1         134
## 6     6      9.76 50.9     5.1    21.9      97.0     150         2     2         147
##   護士數目  可用設備
## 1         241         60
## 2          52         40
## 3          54         20
## 4         148         40
## 5         151         40
## 6         106         40
```

```
summary(data)
```

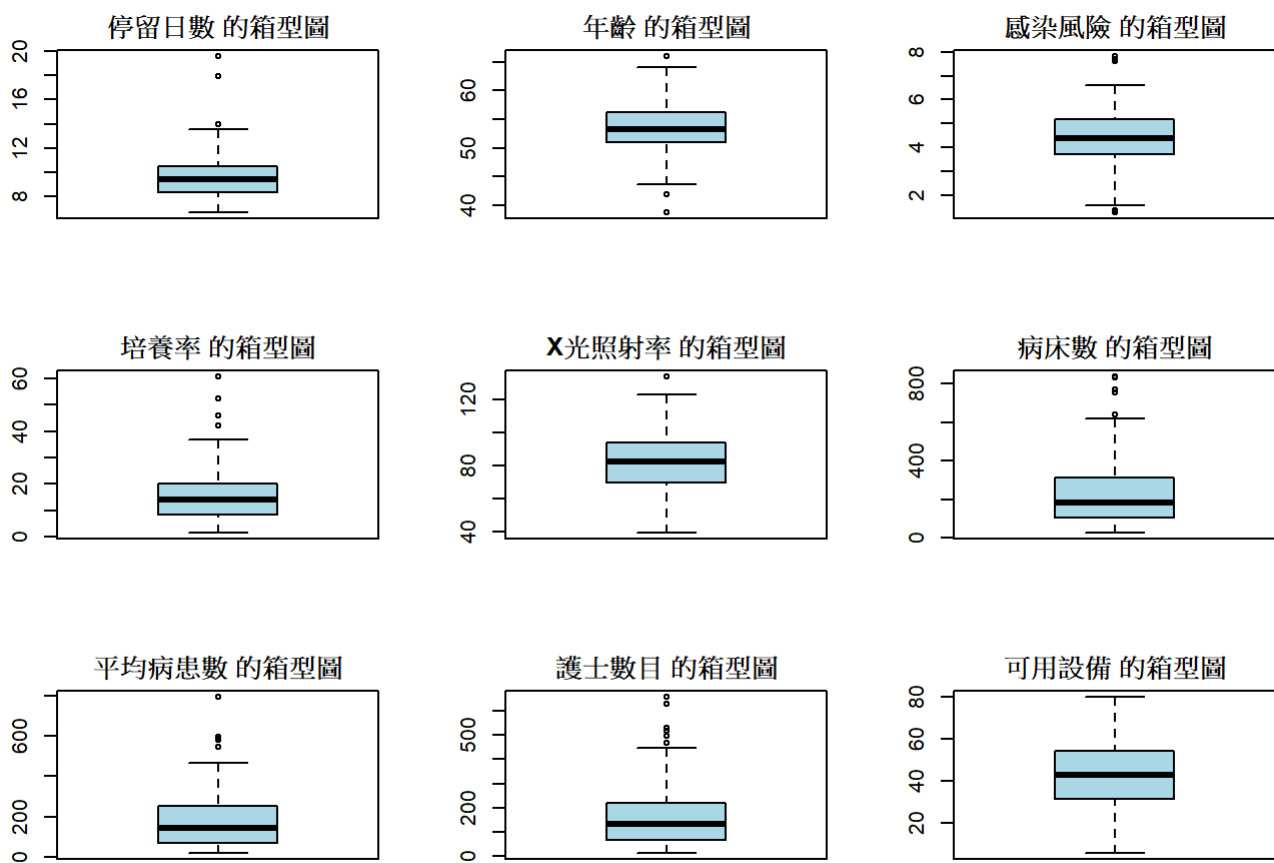
##	編號	停留日數	年齡	感染風險	培養率
##	Min. : 1	Min. : 6.700	Min. :38.80	Min. :1.300	Min. : 1.60
##	1st Qu.: 29	1st Qu.: 8.340	1st Qu.:50.90	1st Qu.:3.700	1st Qu.: 8.40
##	Median : 57	Median : 9.420	Median :53.20	Median :4.400	Median :14.10
##	Mean : 57	Mean : 9.648	Mean :53.23	Mean :4.355	Mean :15.79
##	3rd Qu.: 85	3rd Qu.:10.470	3rd Qu.:56.20	3rd Qu.:5.200	3rd Qu.:20.30
##	Max. :113	Max. :19.560	Max. :65.90	Max. :7.800	Max. :60.50
##	X光照射率	病床數	醫校合作	區域	平均病患數
##	Min. : 39.60	Min. : 29.0	1:17	1:28	Min. : 20.0
##	1st Qu.: 69.50	1st Qu.:106.0	2:96	2:32	1st Qu.: 68.0
##	Median : 82.30	Median :186.0		3:37	Median :143.0
##	Mean : 81.63	Mean :252.2		4:16	Mean :191.4
##	3rd Qu.: 94.10	3rd Qu.:312.0			3rd Qu.:252.0
##	Max. :133.50	Max. :835.0			Max. :791.0
##	護士數目	可用設備			
##	Min. : 14.0	Min. : 5.70			
##	1st Qu.: 66.0	1st Qu.:31.40			
##	Median :132.0	Median :42.90			
##	Mean :173.2	Mean :43.16			
##	3rd Qu.:218.0	3rd Qu.:54.30			
##	Max. :656.0	Max. :80.00			

觀察討論:

這次的資料變數相比上週的作業更多了。特別的是「醫院合作」和「地域」這兩個欄位，雖然是數字，但敘述上卻是類別變數的特徵。 避免模型分析出錯，需要用 `as.factor` 將這兩個欄位的數據變成類別變數。

這邊我們用箱型圖先初步觀察資料的樣態

```
continuous_vars <- c("停留日數", "年齡", "感染風險", "培養率", "X光照射率", "病床數", "平均病患數", "護士數目", "可用設備")
par(mfrow=c(3,3), mar=c(4, 4, 2, 1)) # 設定圖表排列方式，調整邊界
for (var in continuous_vars) {
  boxplot(data[[var]], main=paste(var, "的箱型圖"), col="lightblue", cex.main=1.2)
}
```



```
par(mfrow=c(1,1)) # 恢復單圖模式
```

觀察討論:

- 停留日數、年齡、感染風、可用設備
 - 這些變數的分佈比較集中，沒有明顯極端值，代表住院日數、年齡與感染風險在數據中相對穩定。
- 培養率與 X 光照射率
 - 培養率的離群值較多，顯示有部分醫院進行的檢測次數遠高於平均。
 - X 光照射率也有離群值，可能代表某些醫院對肺炎病患進行了更頻繁的 X 光檢查。
- 病床數、平均病患數、護士數目
 - 多數數據集中在較小範圍，但也有幾個醫院的規模遠超其他醫院，形成明顯的離群點。

連續變數的直方圖

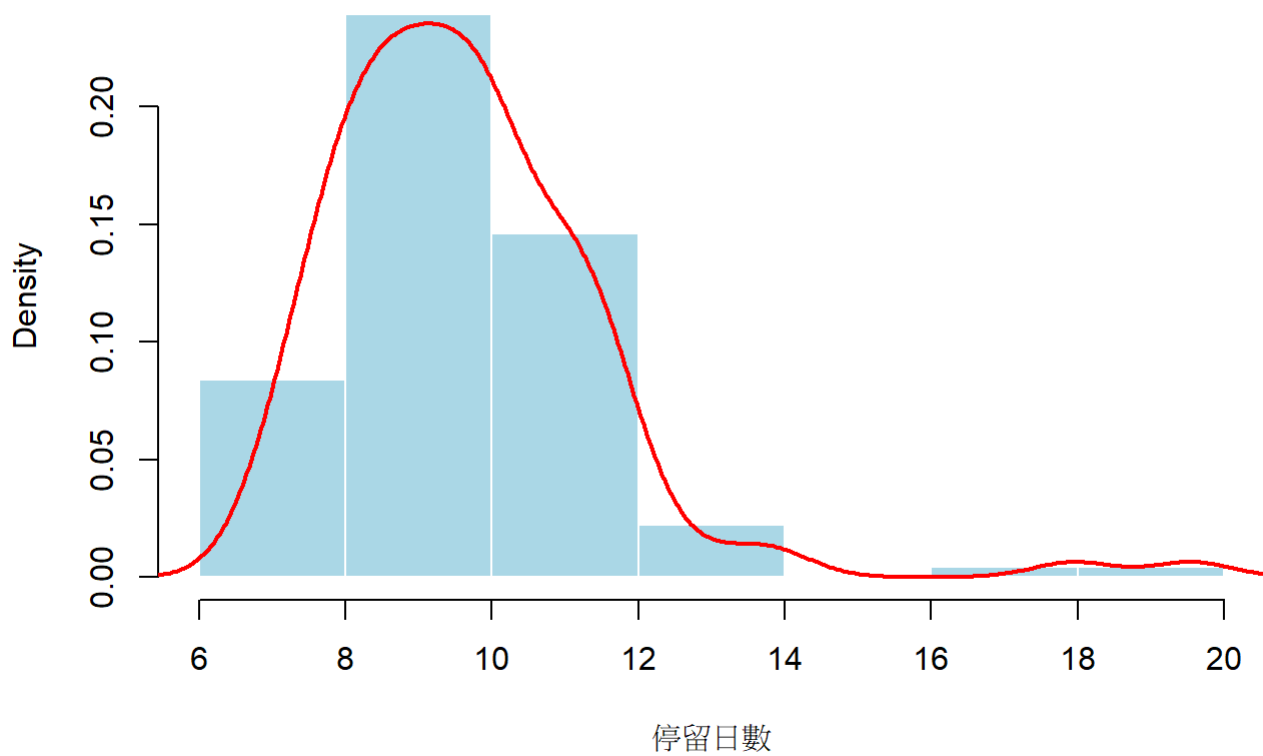
觀察完箱型圖後，稍微了解變數是否有過多的outlier，以及中位數等等數值的可視化。

接下來我想用直方圖來了解數據的樣態。

```
continuous_vars <- c("停留日數", "年齡", "感染風險", "培養率", "X光照射率", "病床數", "平均病患數", "護士數目", "可用設備")

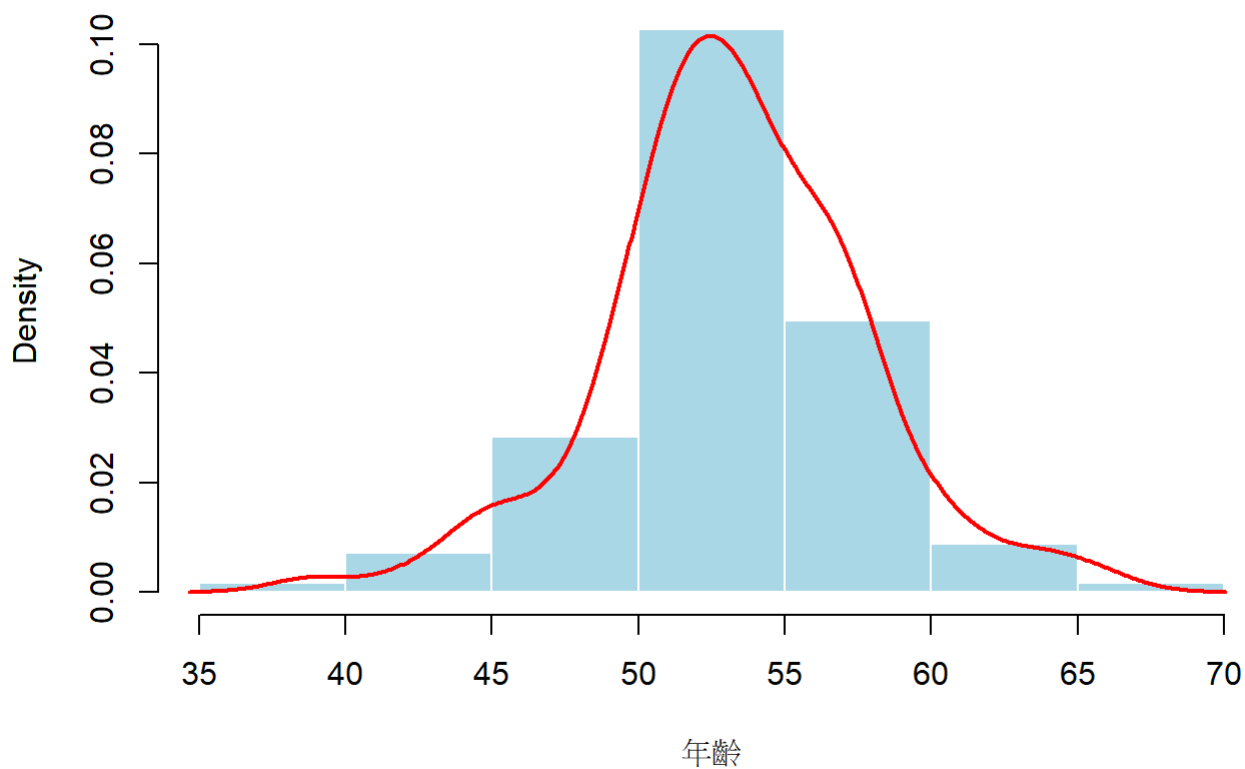
# 依序繪製 9 張直方圖
hist(data$停留日數, main="停留日數的直方圖", xlab="停留日數", col="lightblue", border="white", probability=TRUE)
lines(density(data$停留日數), col="red", lwd=2)
```

停留日數的直方圖

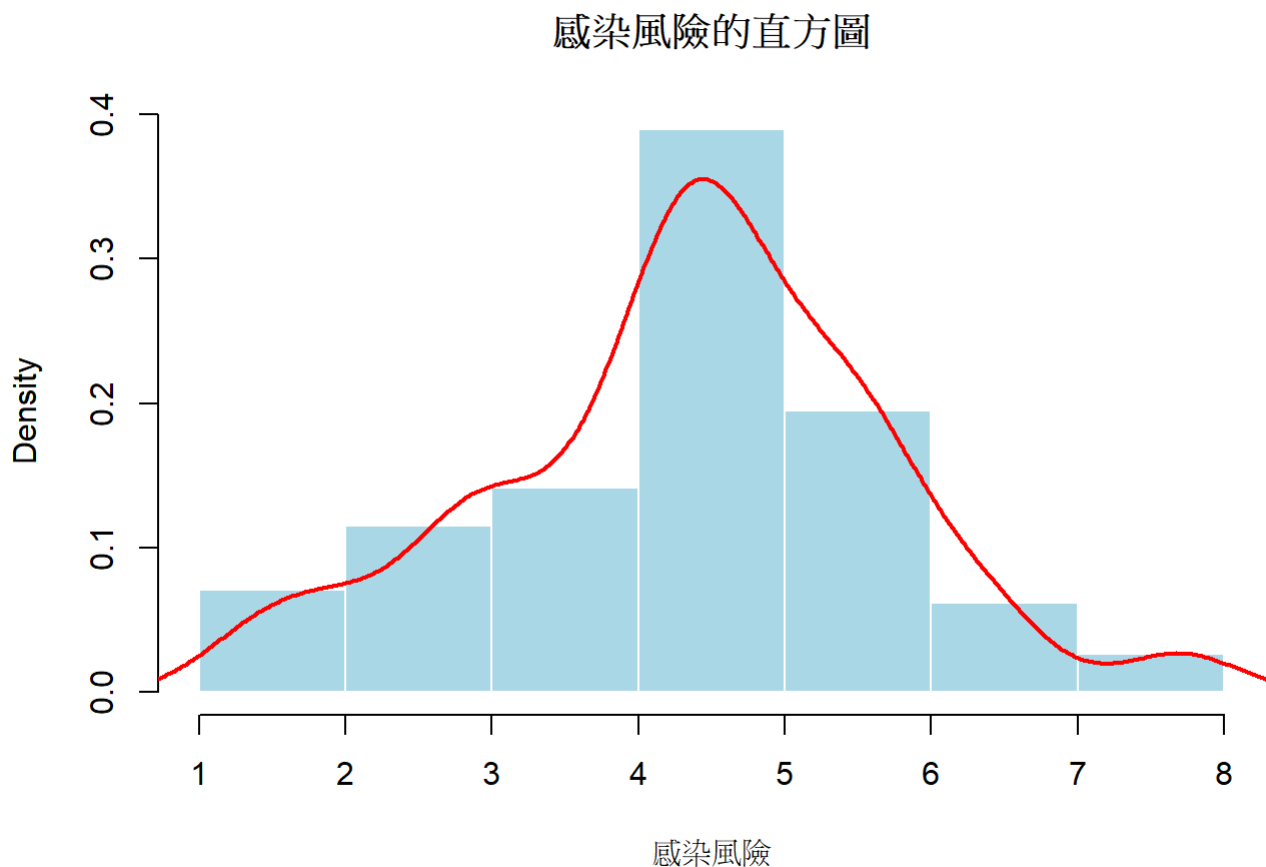


```
hist(data$年齡, main="年齡的直方圖", xlab="年齡", col="lightblue", border="white", probability=TRUE)
lines(density(data$年齡), col="red", lwd=2)
```

年齡的直方圖

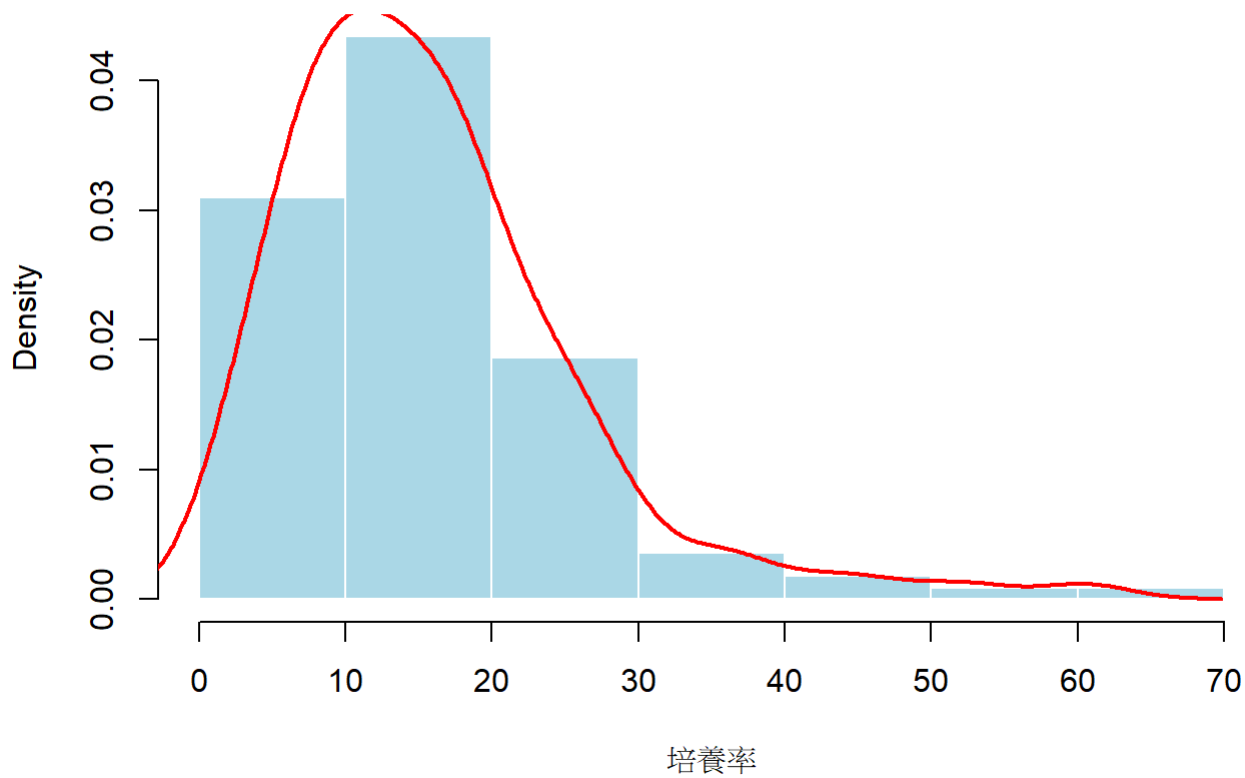


```
hist(data$感染風險, main="感染風險的直方圖", xlab="感染風險", col="lightblue", border="white", probability=TRUE)  
lines(density(data$感染風險), col="red", lwd=2)
```



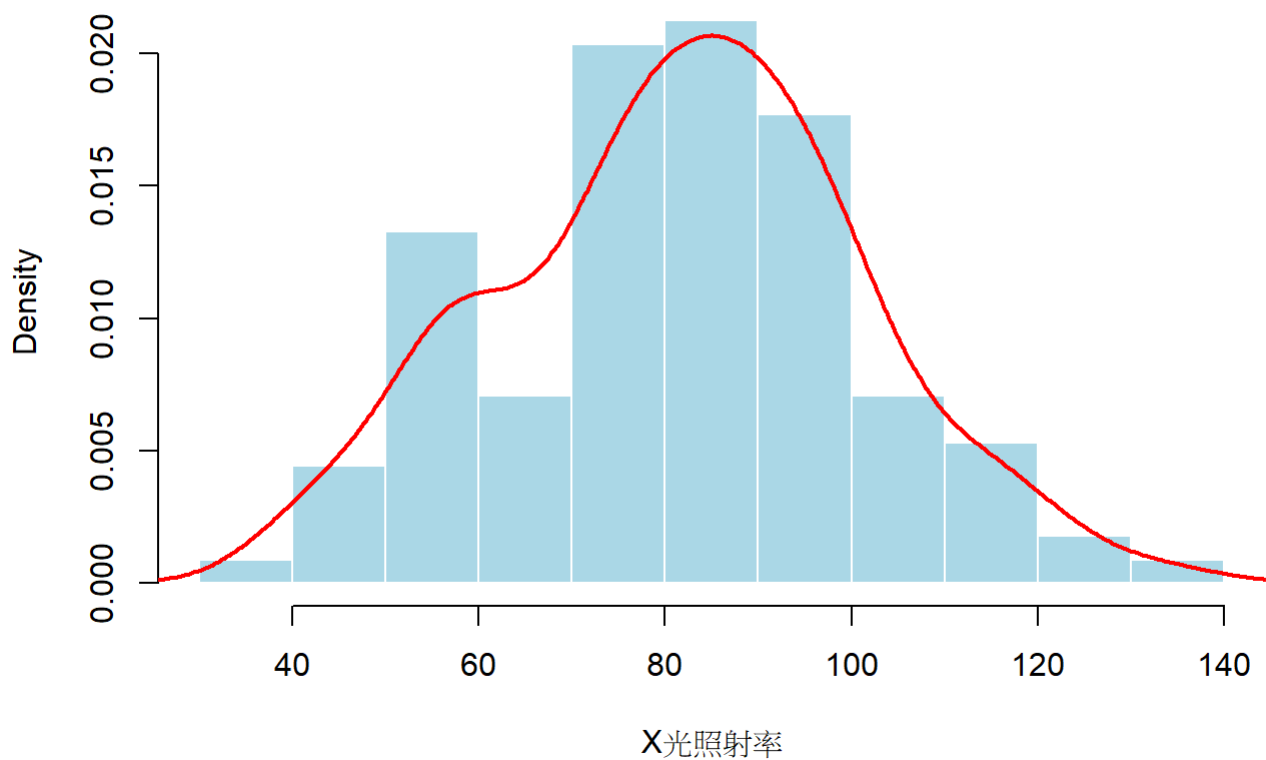
```
hist(data$培養率, main="培養率的直方圖", xlab="培養率", col="lightblue", border="white", probability=TRUE)  
lines(density(data$培養率), col="red", lwd=2)
```

培養率的直方圖



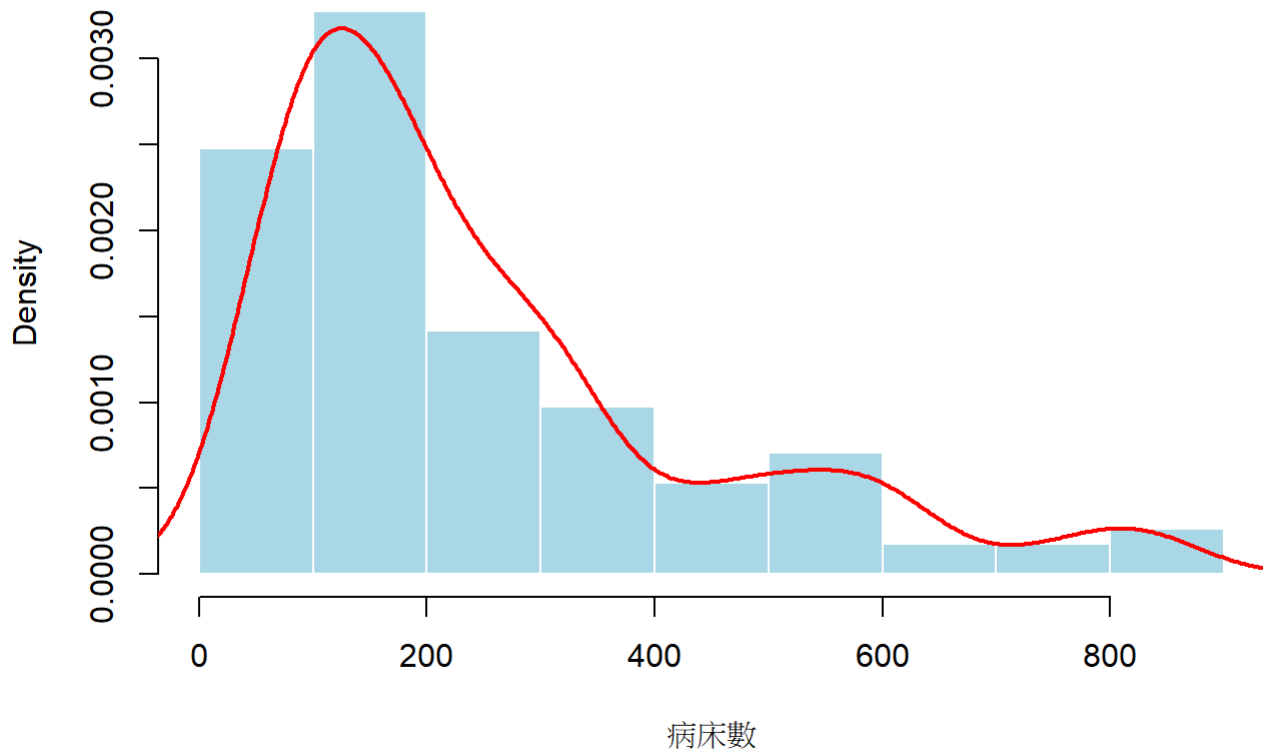
```
hist(data$X光照射率, main="X光照射率的直方圖", xlab="X光照射率", col="lightblue", border="white", probability=TRUE)
lines(density(data$X光照射率), col="red", lwd=2)
```

X光照射率的直方圖



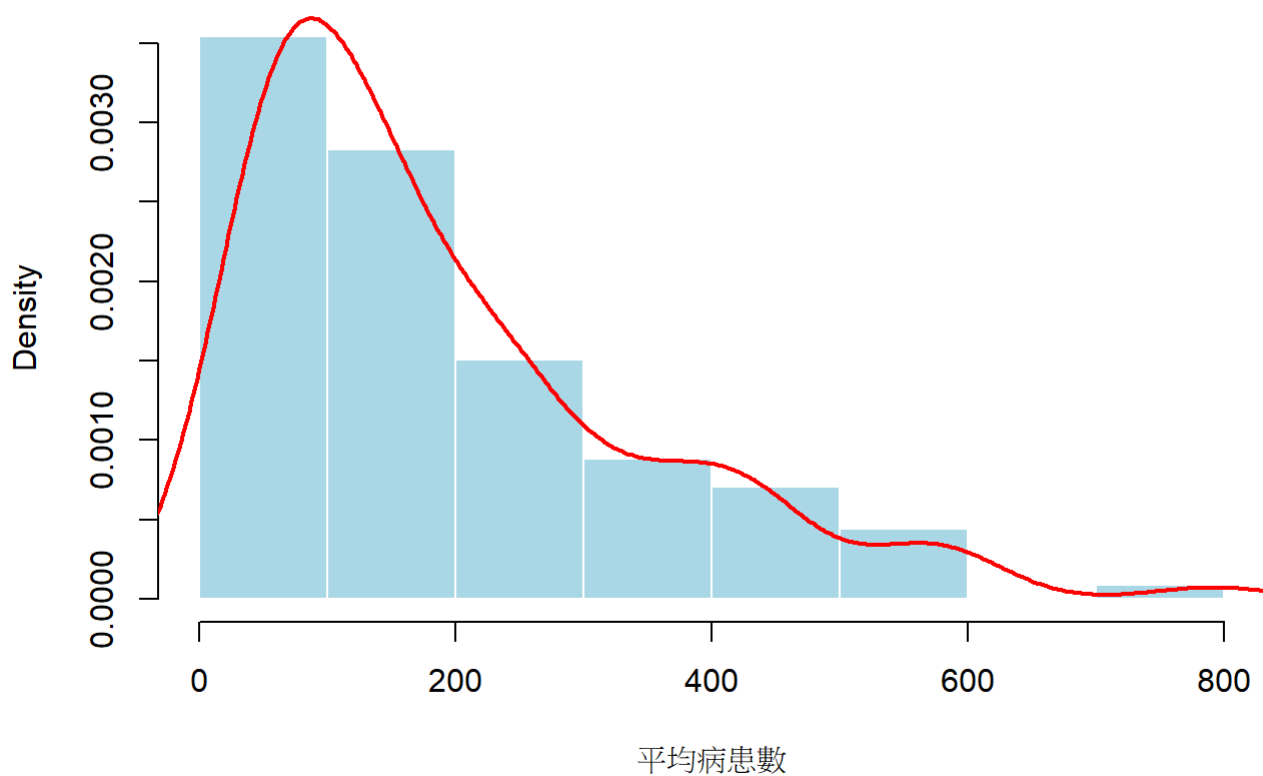
```
hist(data$病床數, main="病床數的直方圖", xlab="病床數", col="lightblue", border="white", probability=TRUE)  
lines(density(data$病床數), col="red", lwd=2)
```

病床數的直方圖



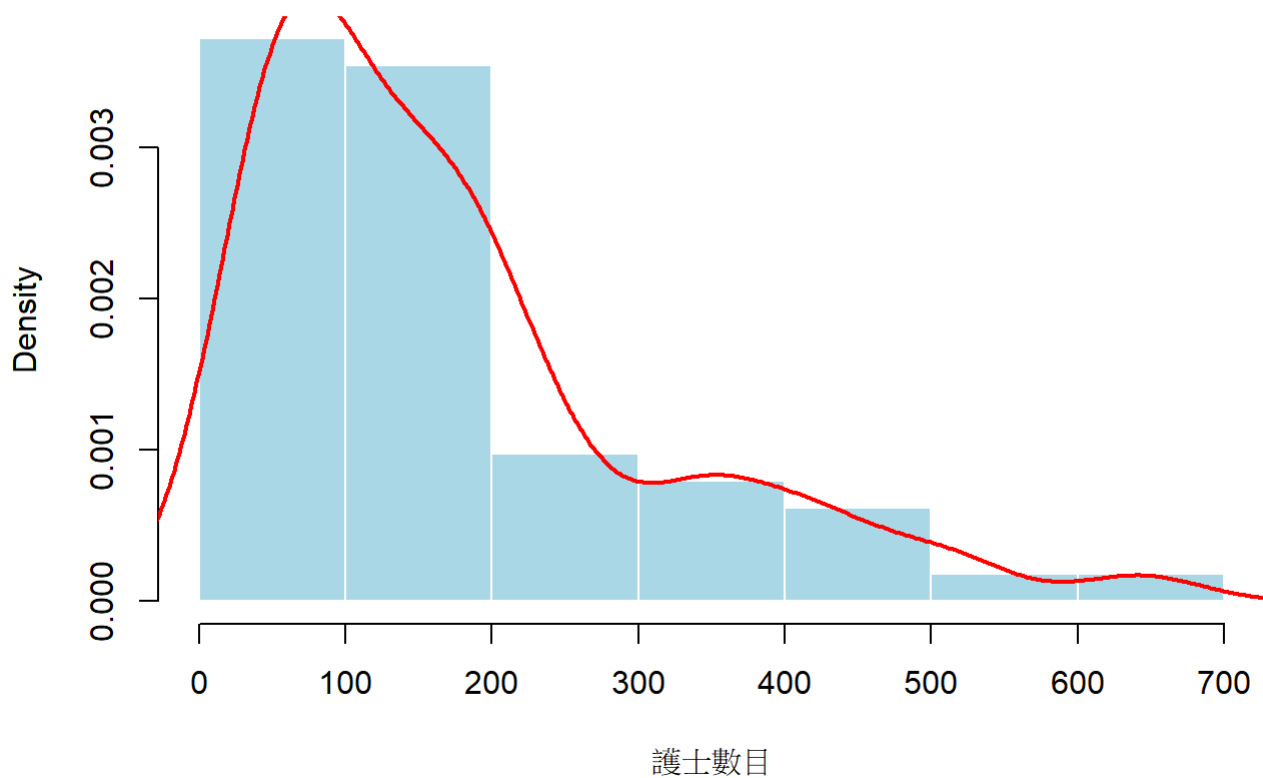
```
hist(data$平均病患數, main="平均病患數的直方圖", xlab="平均病患數", col="lightblue", border="white", probability=TRUE)  
lines(density(data$平均病患數), col="red", lwd=2)
```


平均病患數的直方圖

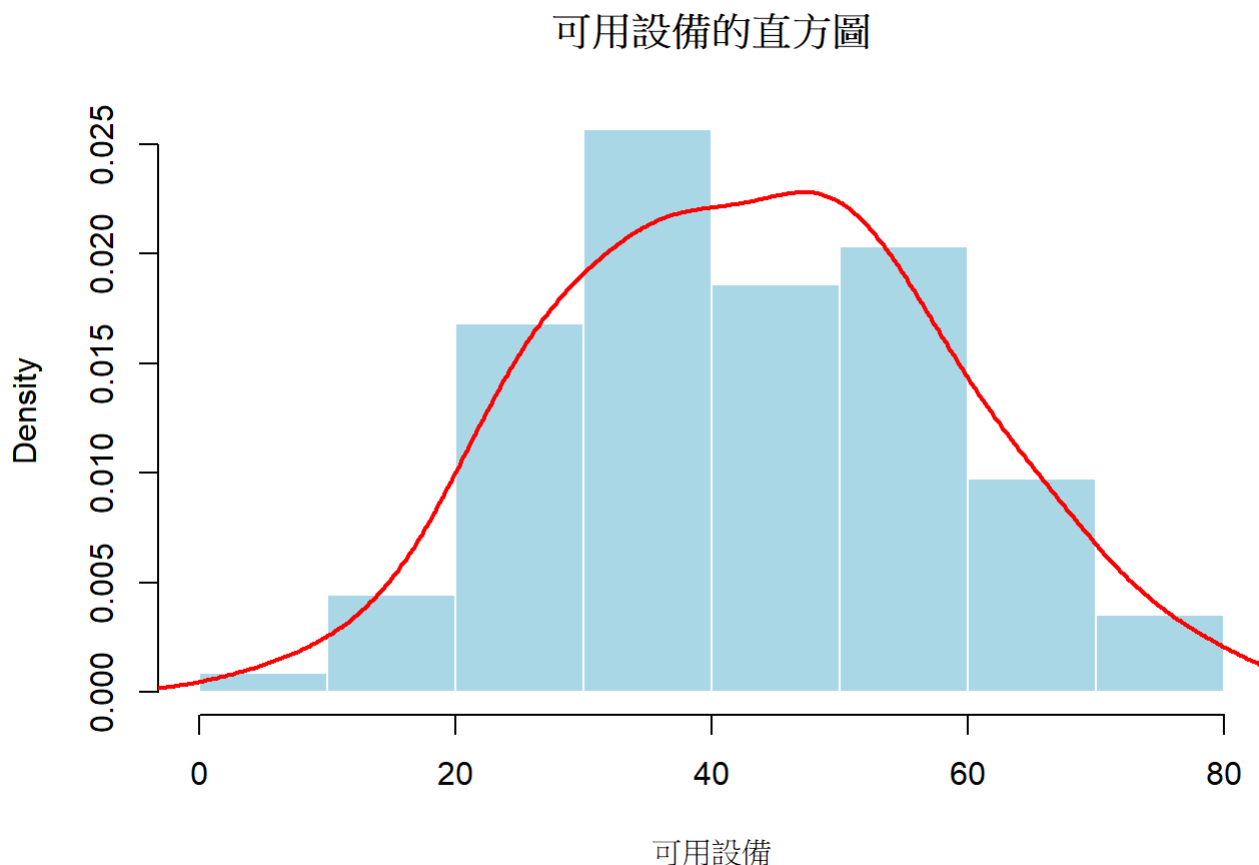


```
hist(data$護士數目, main="護士數目的直方圖", xlab="護士數目", col="lightblue", border="white", probability=TRUE)  
lines(density(data$護士數目), col="red", lwd=2)
```

護士數目的直方圖



```
hist(data$可用設備, main="可用設備的直方圖", xlab="可用設備", col="lightblue", border="white", probability=TRUE)  
lines(density(data$可用設備), col="red", lwd=2)
```



觀察討論:

我原本直方圖也是想要做3*3的子圖，但輸出後觀察到子圖的大小實在是不太好進行判斷跟觀察，因此後來決定每個連續變數都化成各自的直方圖。

停留日數、培養率、病床數、平均病患數以及護士數目有明顯的偏態(右偏)，推測其原因

1. 停留日數（住院天數）

- 多數病人住院時間較短，少數重症患者會極大化住院時間

2. 培養率

- 部分醫院不會頻繁進行培養檢測，只有特定醫院或特定病患才會高頻培養。

3. 病床數

- 醫學中心與區域醫院的病床數差異極大，在這筆資料中小型醫院為大宗，少數大型醫院推高了整體分布。

4. 平均病患數

- 我想應該跟上述原因相同，大型醫院為少數，中小型醫院承接較少但頻率較高的病患術，造成直方圖右偏。

5. 護士數目

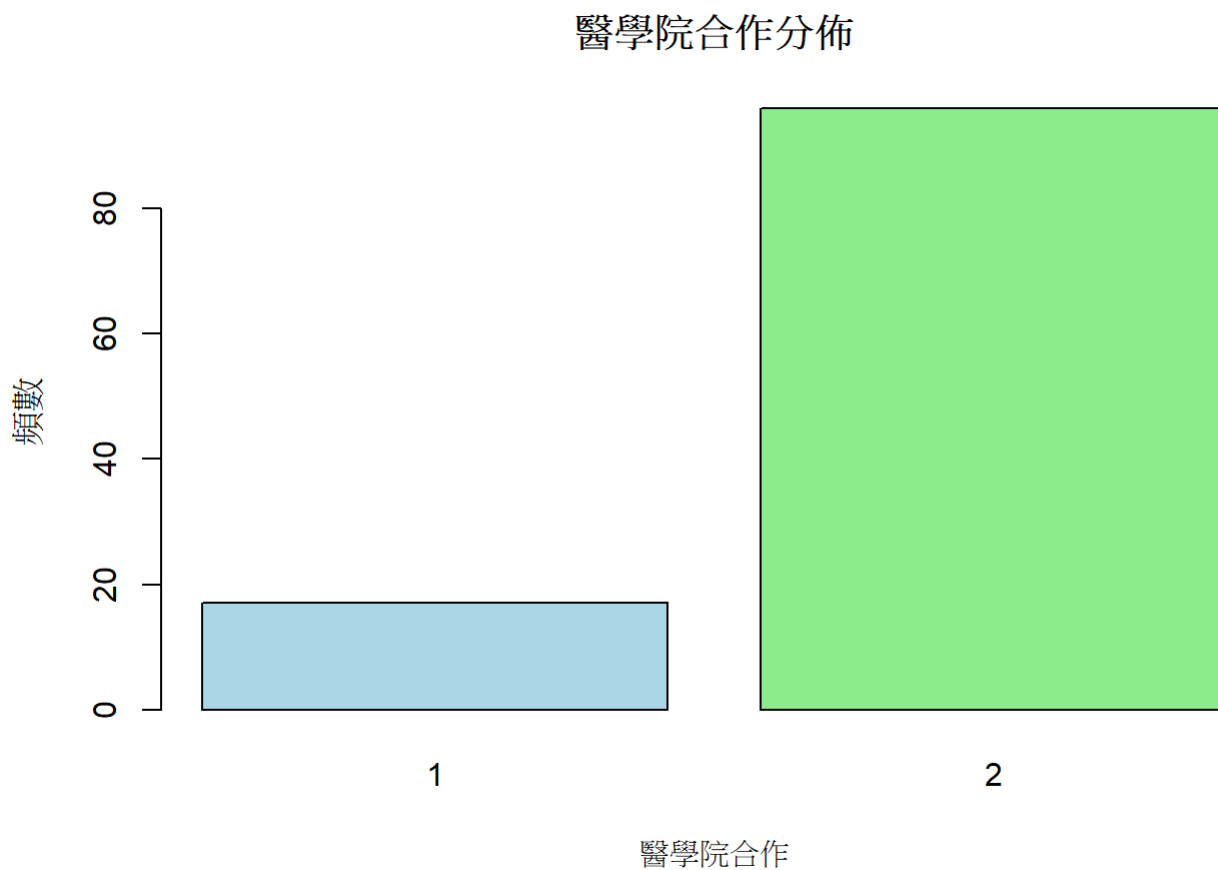
- 多數醫院的護士人數維持在標準範圍內（如 50-200 人），但大型醫學中心擁有數百到上千名護士，造成極端值。

其他變數則看起來符合常態。

類別變數分析

醫校合作的長條圖

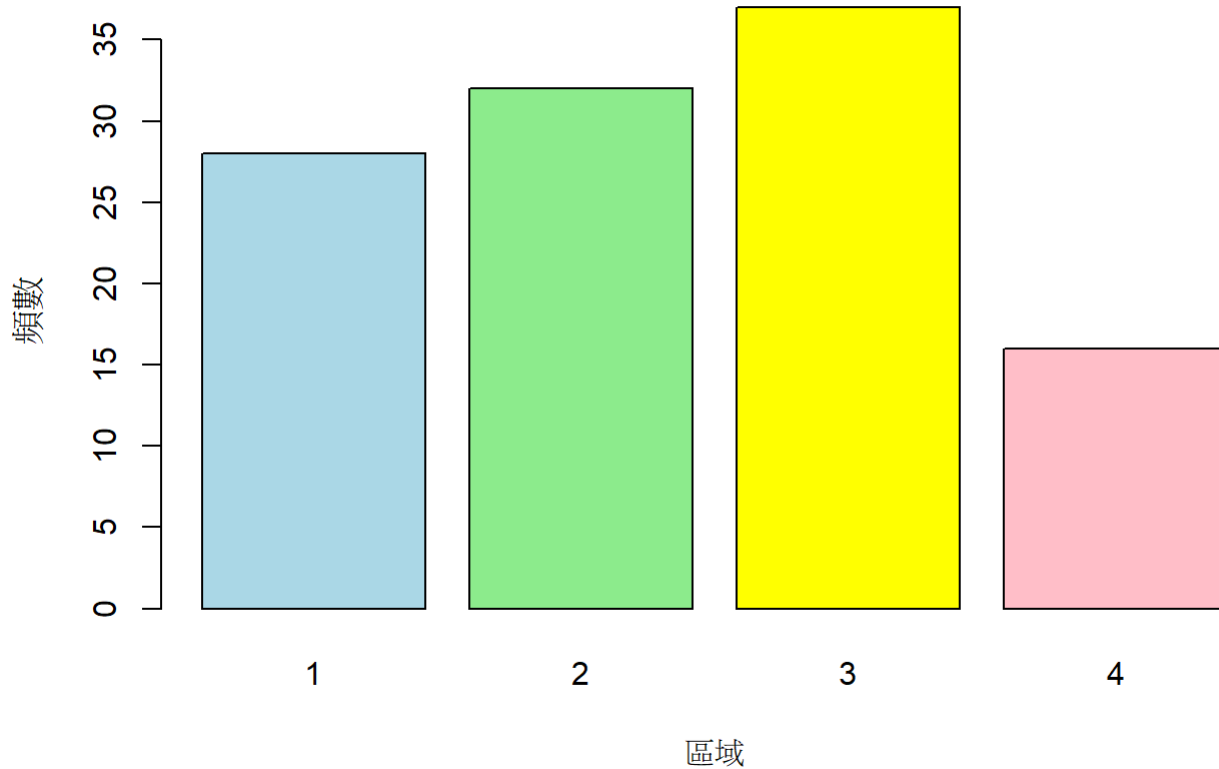
```
barplot(table(data$醫校合作),  
        main="醫學院合作分佈",  
        xlab="醫學院合作",  
        ylab="頻數",  
        col=c("lightblue", "lightgreen"))
```



區域的長條圖

```
barplot(table(data$區域),  
        main="區域分佈",  
        xlab="區域",  
        ylab="頻數",  
        col=c("lightblue", "lightgreen", "yellow", "pink"))
```

區域分佈



醫學院合作與區域的列聯表

```
table1 <- table(data$醫校合作, data$區域)
prop.table(table1) # 總和為 1 的比例表
```

```
##
##           1           2           3           4
##  1 0.04424779 0.06194690 0.02654867 0.01769912
##  2 0.20353982 0.22123894 0.30088496 0.12389381
```

```
prop.table(table1, 1) # 列的和為 1 的比例表
```

```
##
##           1           2           3           4
##  1 0.2941176 0.4117647 0.1764706 0.1176471
##  2 0.2395833 0.2604167 0.3541667 0.1458333
```

```
prop.table(table1, 2) # 欄的和為 1 的比例表
```

```
##
##           1           2           3           4
##  1 0.17857143 0.21875000 0.08108108 0.12500000
##  2 0.82142857 0.78125000 0.91891892 0.87500000
```

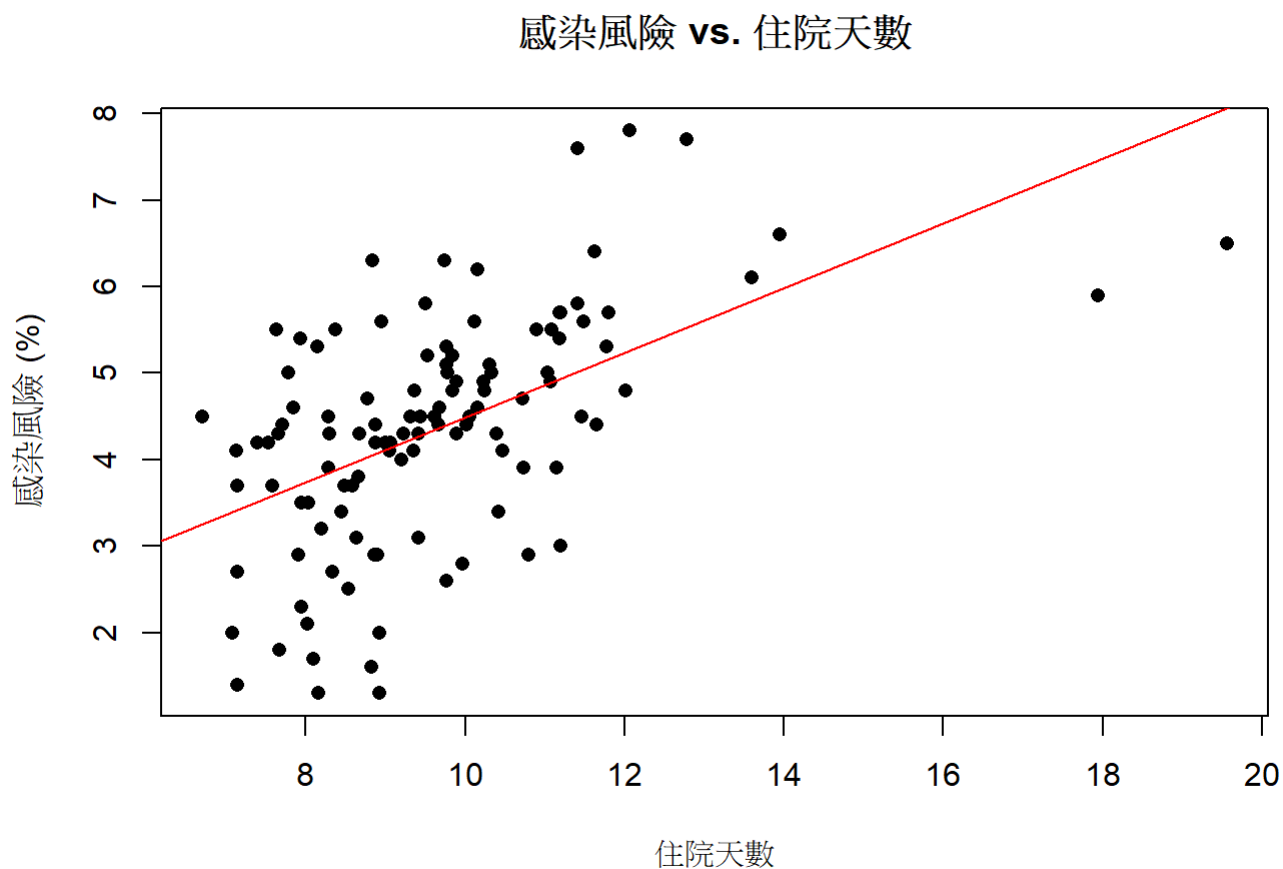
觀察與討論

以醫校合作來說，大多數醫院是沒有合作的。

而地域而言，隨機抽取的四個區域屬於編號4的西部較少，可能該地區較為不發達，或可能是抽樣時的誤差。

感染風險與住院天數

```
plot(data$停留日數, data$感染風險,  
      xlab="住院天數", ylab="感染風險 (%)",  
      main="感染風險 vs. 住院天數", pch=16)  
abline(lm(感染風險 ~ 停留日數, data=data), col="red")
```

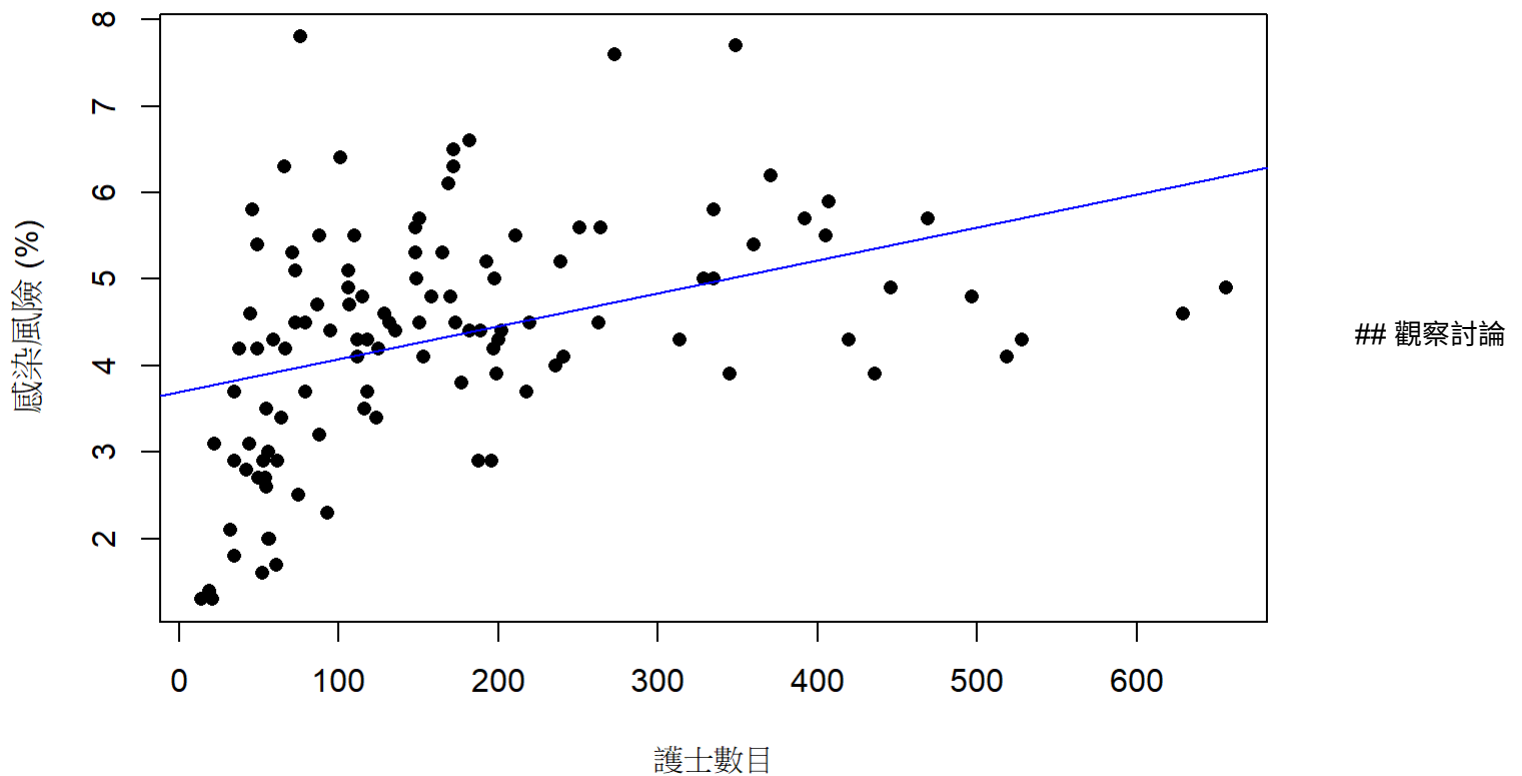


散佈圖

護士數量與感染風險

```
plot(data$護士數目, data$感染風險,  
      xlab="護士數目", ylab="感染風險 (%)",  
      main="護士數目與感染風險", pch=16)  
abline(lm(感染風險 ~ 護士數目, data=data), col="blue")
```

護士數目與感染風險



這邊僅先列出兩個散佈圖做觀察，剛好都是屬於正相關的例子。第一個先看到感染風險跟住院天數的散佈圖，可以看到你住在醫院越久，感染風險越高(畢竟有更多機會遇到更多病)，對於我們的常識來說滿合理的。

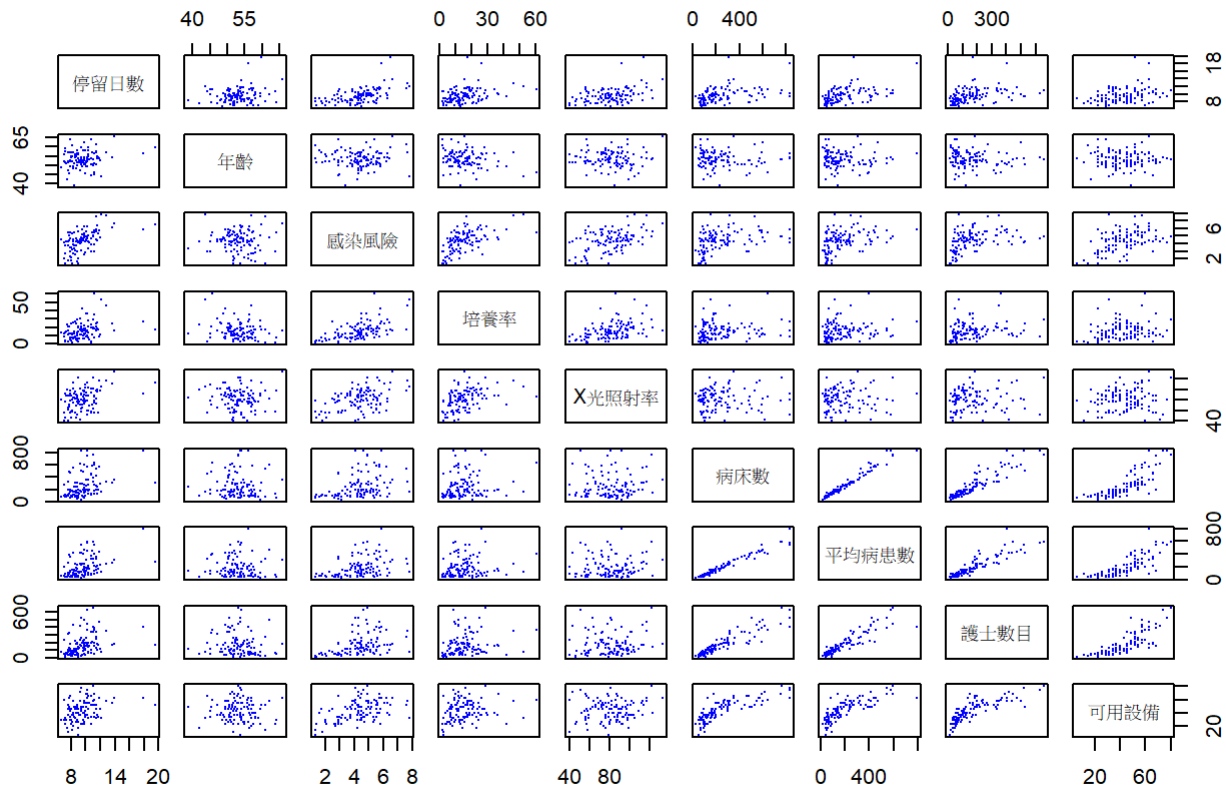
此外，當護士越多時，感染風險越高，剛好呼應到上述的觀察，可能擁有眾多護士數目的醫院都是大型的，並且可能有較多重症患者，因此這個圖算是符合我們上述推論的，但參照到趨勢線來說，正相關的程度相對沒有這麼高。

所有連續變數的散佈圖

這邊我們利用 `pairs` 來一窺所有連續變數的相關程度，就不多做贅述。

```
pairs(data[, c("停留日數", "年齡", "感染風險", "培養率", "X光照射率", "病床數", "平均病患數", "護士數目", "可用設備")],  
      main="連續變數的散佈圖矩陣",  
      col="blue", pch=16, cex=0.1)
```

連續變數的散佈圖矩陣



小小心得

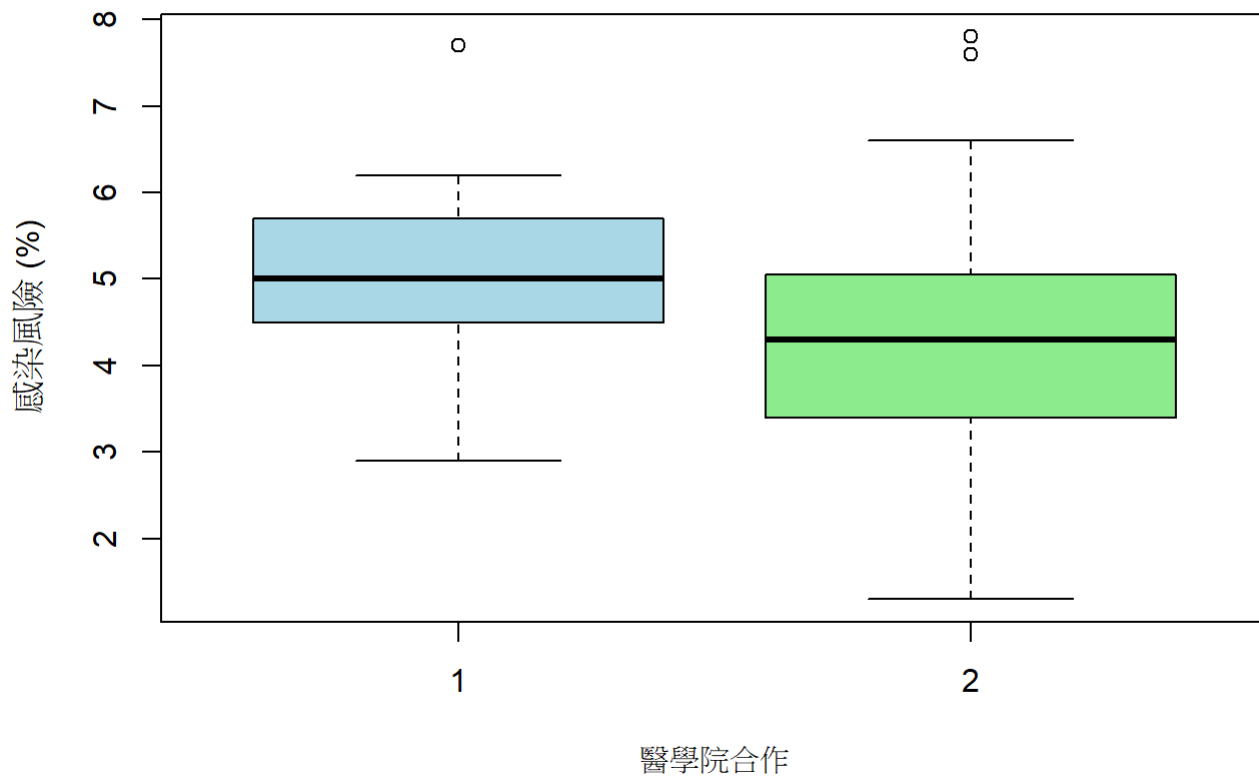
一開始在繪製這張圖表時，點點大到我無法做出推測與分析，後來經過調整 `cex=0.1` 後，才能清楚看到所有散佈圖的展示。

有趣的發現

醫學院合作與感染風險 箱型圖

```
boxplot(感染風險 ~ 醫校合作, data=data,
        xlab="醫學院合作", ylab="感染風險 (%)",
        main="醫學院合作與感染風險", col=c("lightblue", "lightgreen"))
```

醫學院合作與感染風險



觀察討論

居然，沒有與醫學院合作的醫院比較安全？會不會是因為醫學生的經驗不足造成感染風險增加呢？還是容易傳染的病患都喜歡去有跟醫學院合作的醫院呢？這是個滿有趣的現象，否則我還印象中以為像是台大醫院、中國醫真的算是個品牌保證呢。這邊提出來跟大家分享

熱力圖

這邊我想要觀察變數之間的關係。

```
library(ggplot2)
```

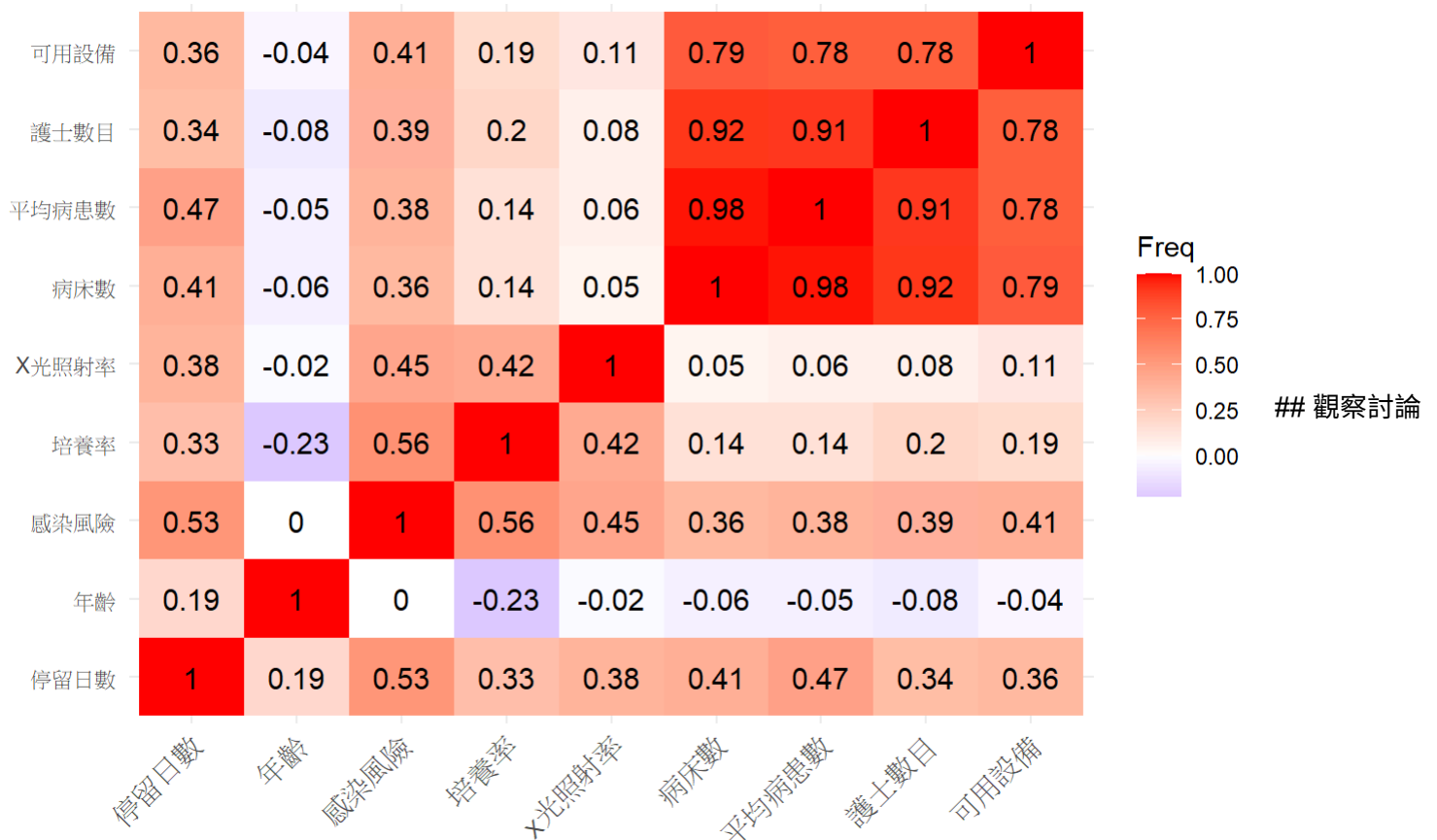
```
## Warning: 套件 'ggplot2' 是用 R 版本 4.4.2 來建造的
```

```
# 計算變數之間的相關矩陣
cor_matrix <- cor(data[, c("停留日數", "年齡", "感染風險", "培養率",
                           "X光照射率", "病床數", "平均病患數", "護士數目", "可用設備")],
                  use = "complete.obs") # 忽略缺失值

# 轉換相關矩陣為數據框
cor_data <- as.data.frame(as.table(cor_matrix))

# 繪製熱力圖（加入數值標籤）
ggplot(cor_data, aes(Var1, Var2, fill = Freq)) +
  geom_tile() +
  geom_text(aes(label = round(Freq, 2)), color = "black", size = 4) + # 加入數值標籤，顯示兩位小數
  scale_fill_gradient2(low="blue", mid="white", high="red", midpoint=0) +
  theme_minimal() +
  labs(title="變數相關性的熱力圖", x="", y="") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


變數相關性的熱力圖



首先觀察到病床數、平均病患數、護士數目、可用設備之間呈高度正相關(最紅的)，這些變數在醫院的規模上有很強的相關性。較大的醫院通常擁有更多的病床、護士，並且配備更多設備。

再來停留日數與感染風險具有正相關，這部分就滿符合常理的。

另外發現X光照射率與病床數高度正相關，可能因為大型醫院可能更常進行 X 光檢查，可能因為這些醫院有更多重症病患所導致。

最後培養率與感染風險呈現負相關 (紫色區域)，這可能代表某些醫院進行較多的微生物培養測試，有助於降低感染風險。

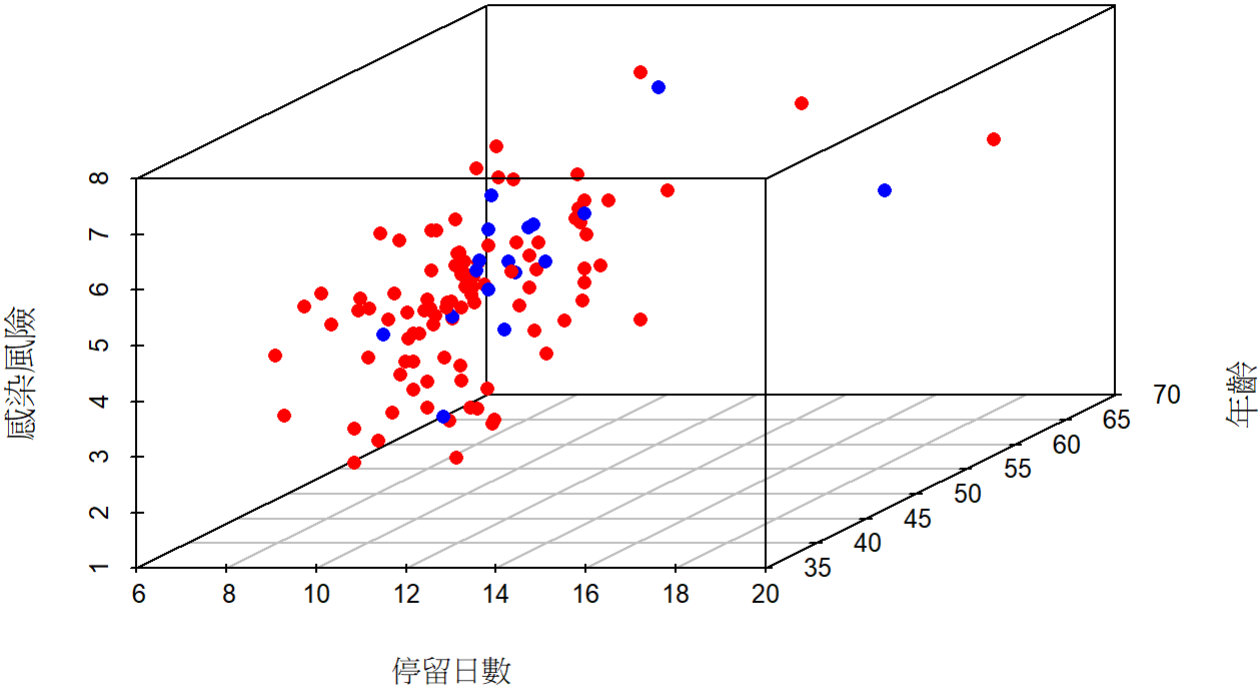
3D 散佈圖練習

```
library(scatterplot3d)

# 設定顏色 (根據醫校合作 1: 藍色, 2: 紅色)
colors <- ifelse(data$醫校合作 == 1, "blue", "red")

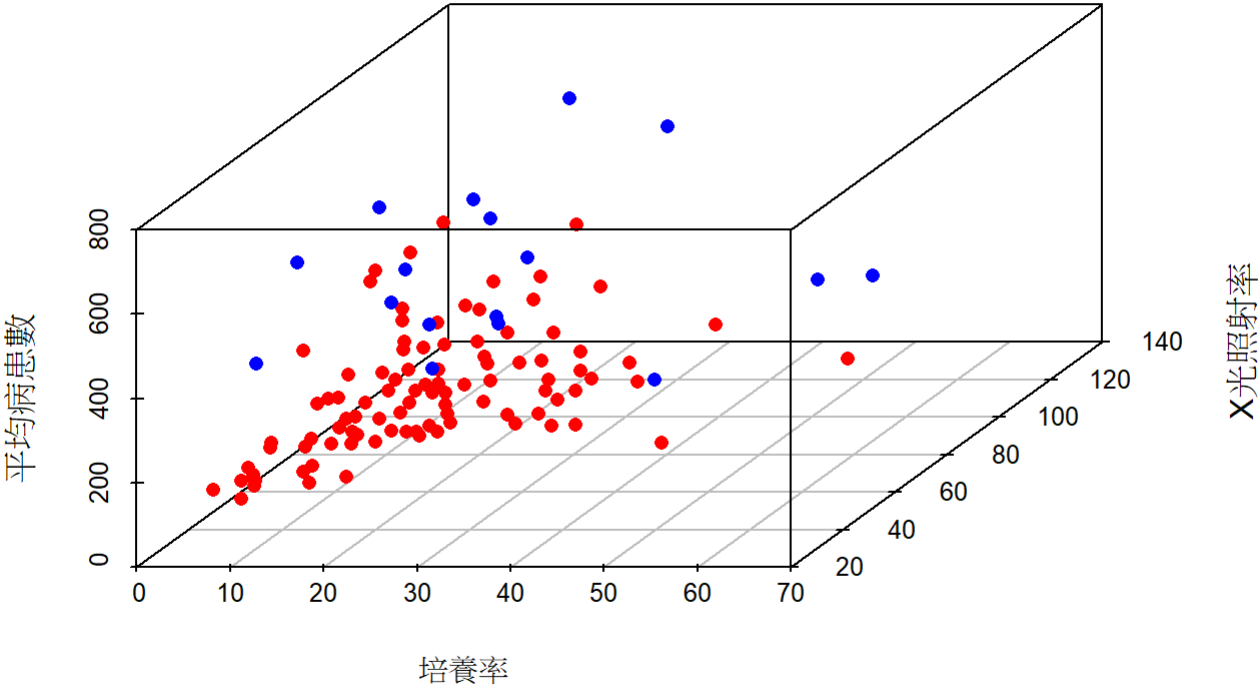
# 3D 散佈圖 1: 停留日數 vs 年齡 vs 感染風險
scatterplot3d(data$停留日數, data$年齡, data$感染風險,
              color=colors, pch=16,
              main="3D散佈圖: 停留日數 vs 年齡 vs 感染風險",
              xlab="停留日數", ylab="年齡", zlab="感染風險")
```

3D散佈圖: 停留日數 vs 年齡 vs 感染風險



```
# 3D 散佈圖 2: 培養率 vs X光照射率 vs 平均病患數
scatterplot3d(data$培養率, data$X光照射率, data$平均病患數,
  color=colors, pch=16,
  main="3D散佈圖: 培養率 vs X光照射率 vs 平均病患數",
  xlab="培養率", ylab="X光照射率", zlab="平均病患數")
```

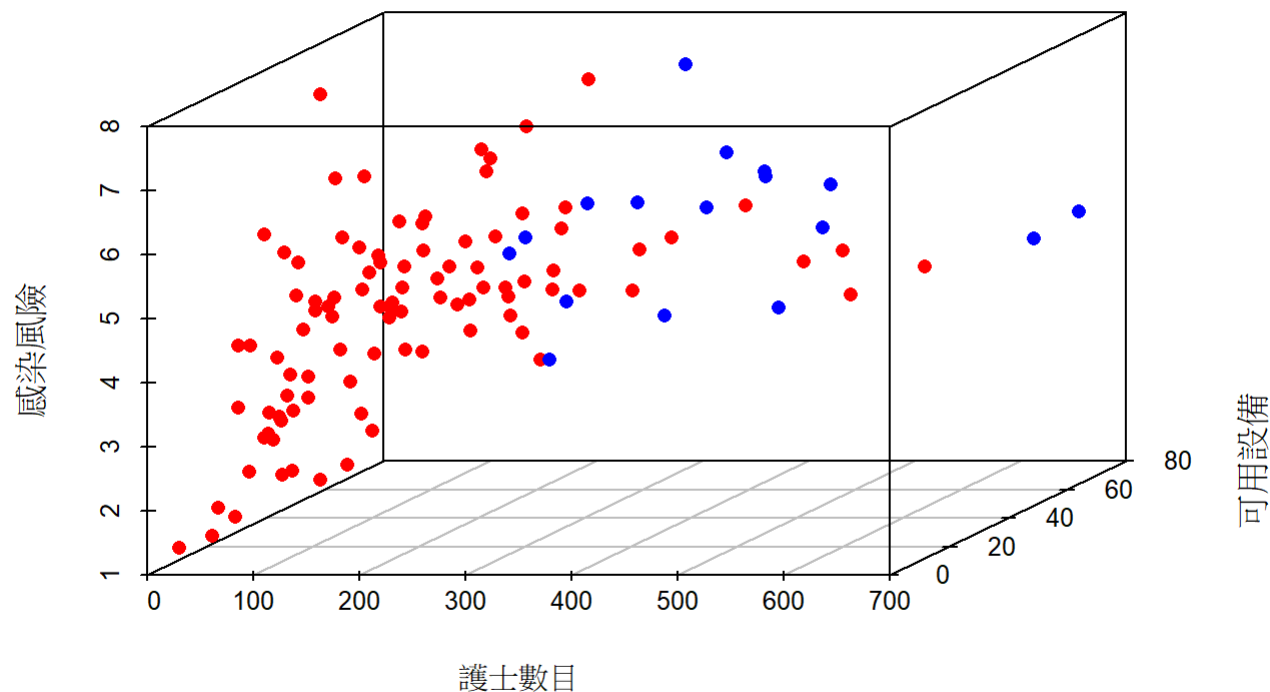
3D散佈圖: 培養率 vs X光照射率 vs 平均病患數



```
# 3D 散佈圖 3: 護士數目 vs 可用設備 vs 感染風險
```

```
scatterplot3d(data$護士數目, data$可用設備, data$感染風險,  
              color=colors, pch=16,  
              main="3D散佈圖: 護士數目 vs 可用設備 vs 感染風險",  
              xlab="護士數目", ylab="可用設備", zlab="感染風險")
```

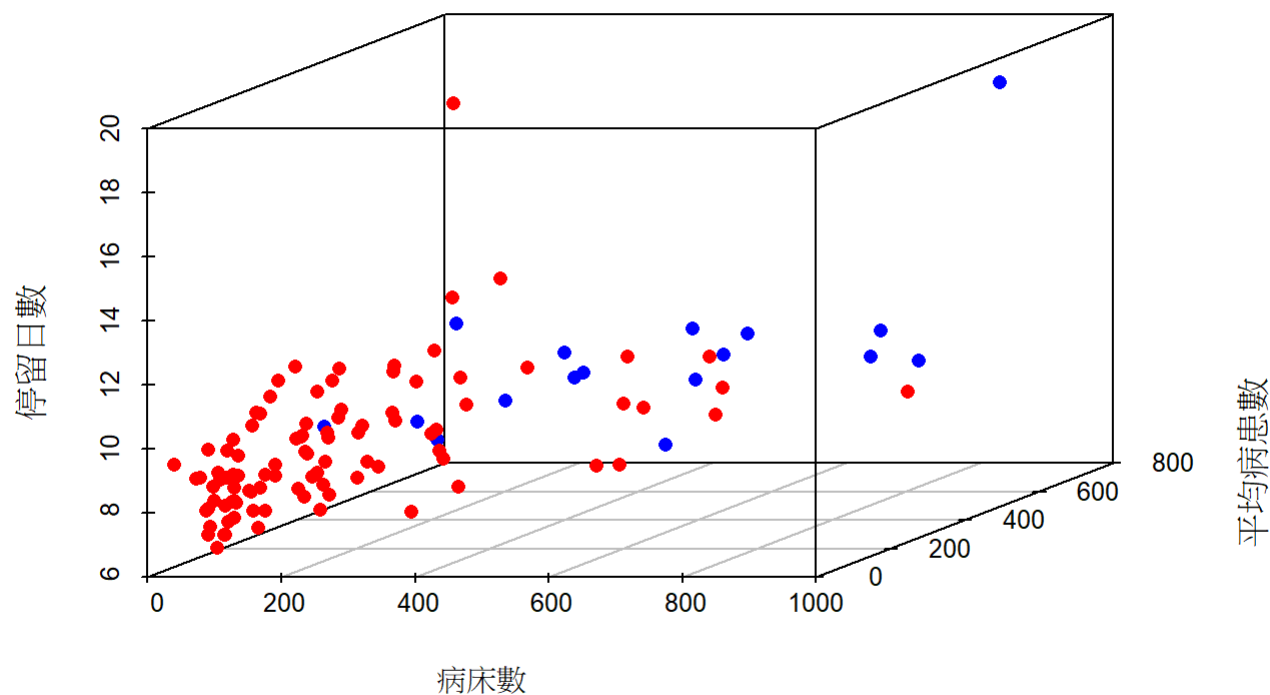
3D散佈圖: 護士數目 vs 可用設備 vs 感染風險



```
# 3D 散佈圖 4: 病床數 vs 平均病患數 vs 停留日數
```

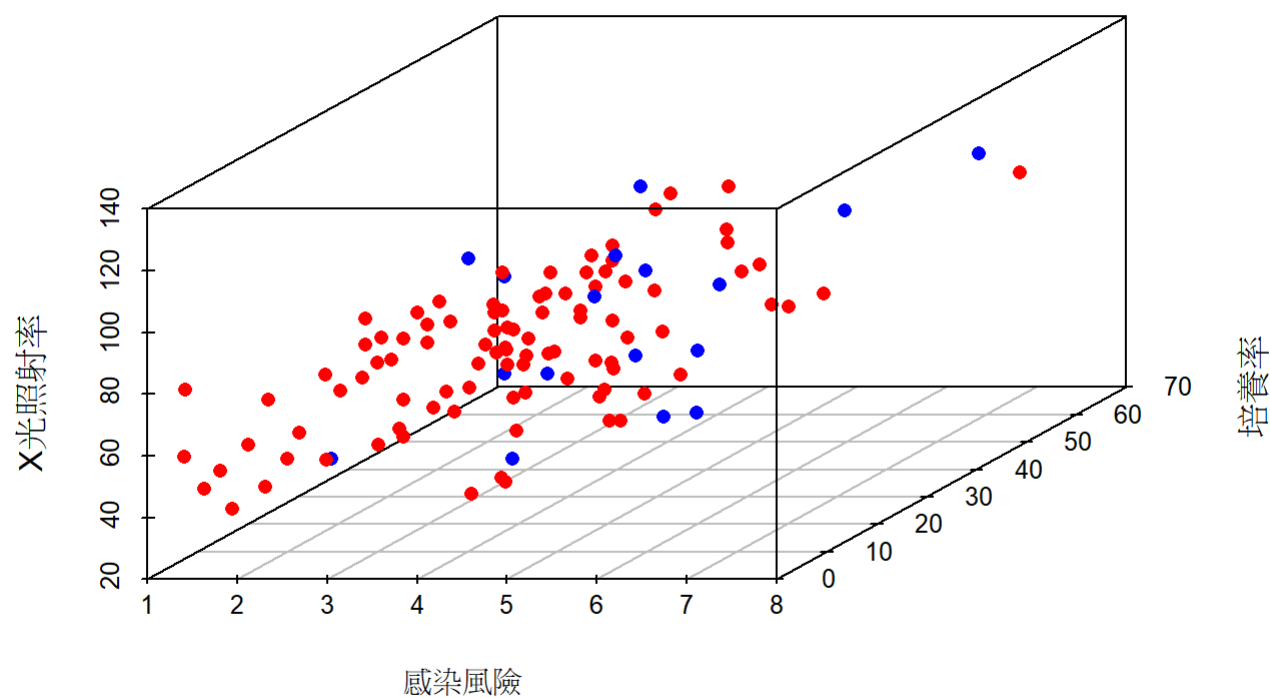
```
scatterplot3d(data$病床數, data$平均病患數, data$停留日數,  
              color=colors, pch=16,  
              main="3D散佈圖: 病床數 vs 平均病患數 vs 停留日數",  
              xlab="病床數", ylab="平均病患數", zlab="停留日數")
```

3D散佈圖: 病床數 vs 平均病患數 vs 停留日數



```
# 3D 散佈圖 5: 感染風險 vs 培養率 vs X光照射率
scatterplot3d(data$感染風險, data$培養率, data$X光照射率,
              color=colors, pch=16,
              main="3D散佈圖: 感染風險 vs 培養率 vs X光照射率",
              xlab="感染風險", ylab="培養率", zlab="X光照射率")
```

3D散佈圖: 感染風險 vs 培養率 vs X光照射率



K-means 分群

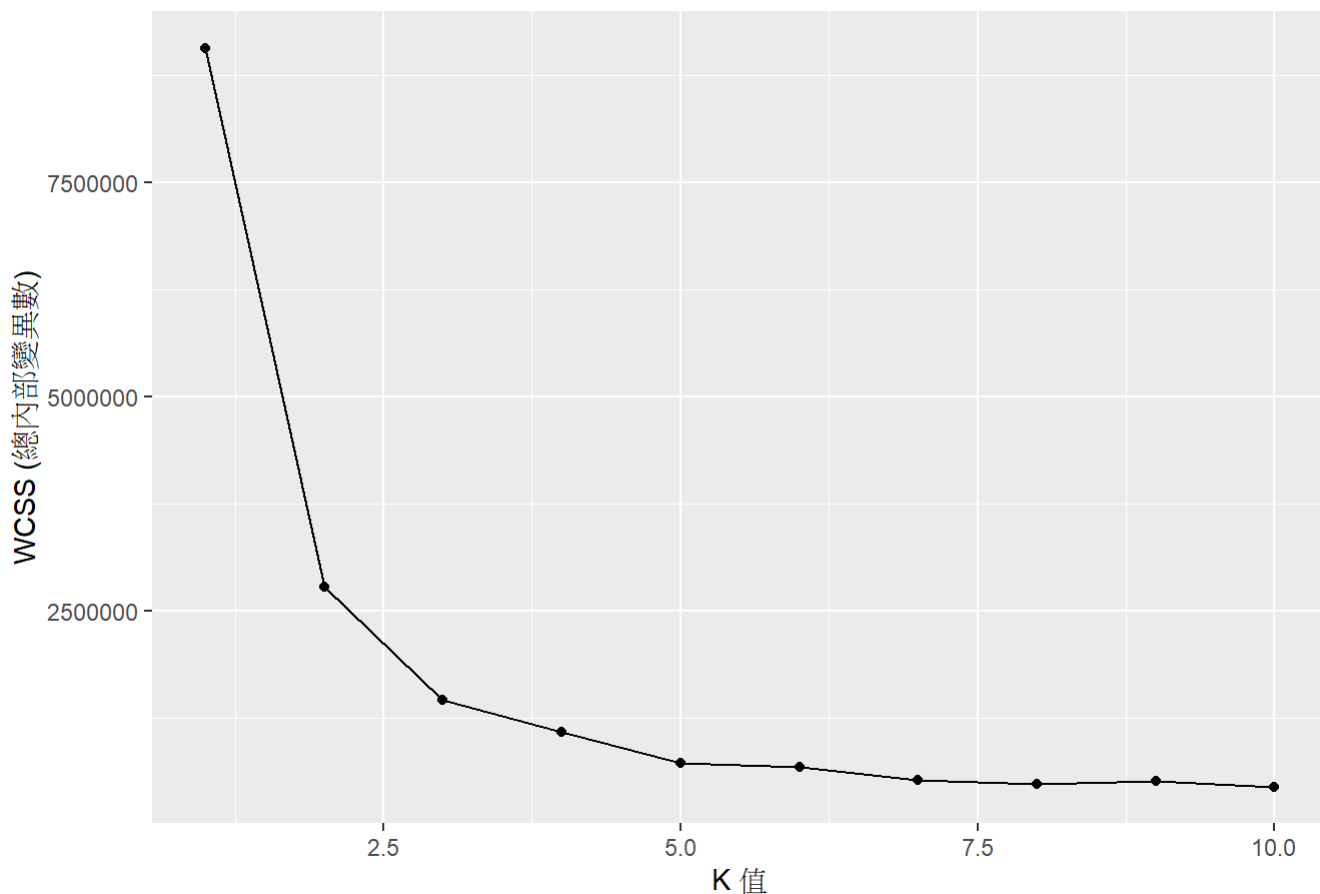
我想去使用K-means對數據中的醫院進行分群，幫助我們了解不同醫院的規模與特性，進而做進一步的醫療資源分析。首先我們要先來看看K值要設定多少比較合理。

```
library(ggplot2)

# 計算不同 K 值的 WCSS
wcss <- c()
for (k in 1:10) {
  kmeans_result <- kmeans(data[, c("停留日數", "年齡", "感染風險", "培養率",
                                    "X光照射率", "病床數", "平均病患數", "護士數目", "可用設備")], centers =
k)
  wcss[k] <- kmeans_result$tot.withinss
}

# 畫出 Elbow Method 圖
ggplot(data.frame(K = 1:10, WCSS = wcss), aes(x = K, y = WCSS)) +
  geom_point() +
  geom_line() +
  ggtitle("Elbow Method 找最佳 K 值") +
  xlab("K 值") +
  ylab("WCSS (總內部變異數)")
```

Elbow Method 找最佳 K 值



我們知道最佳 K 值通常是“彎折點”(elbow point) 在 K = 3 或 4 左右，WCSS 開始趨緩，因此我們可以考慮 K = 3 或 K = 4，我這邊選擇3。

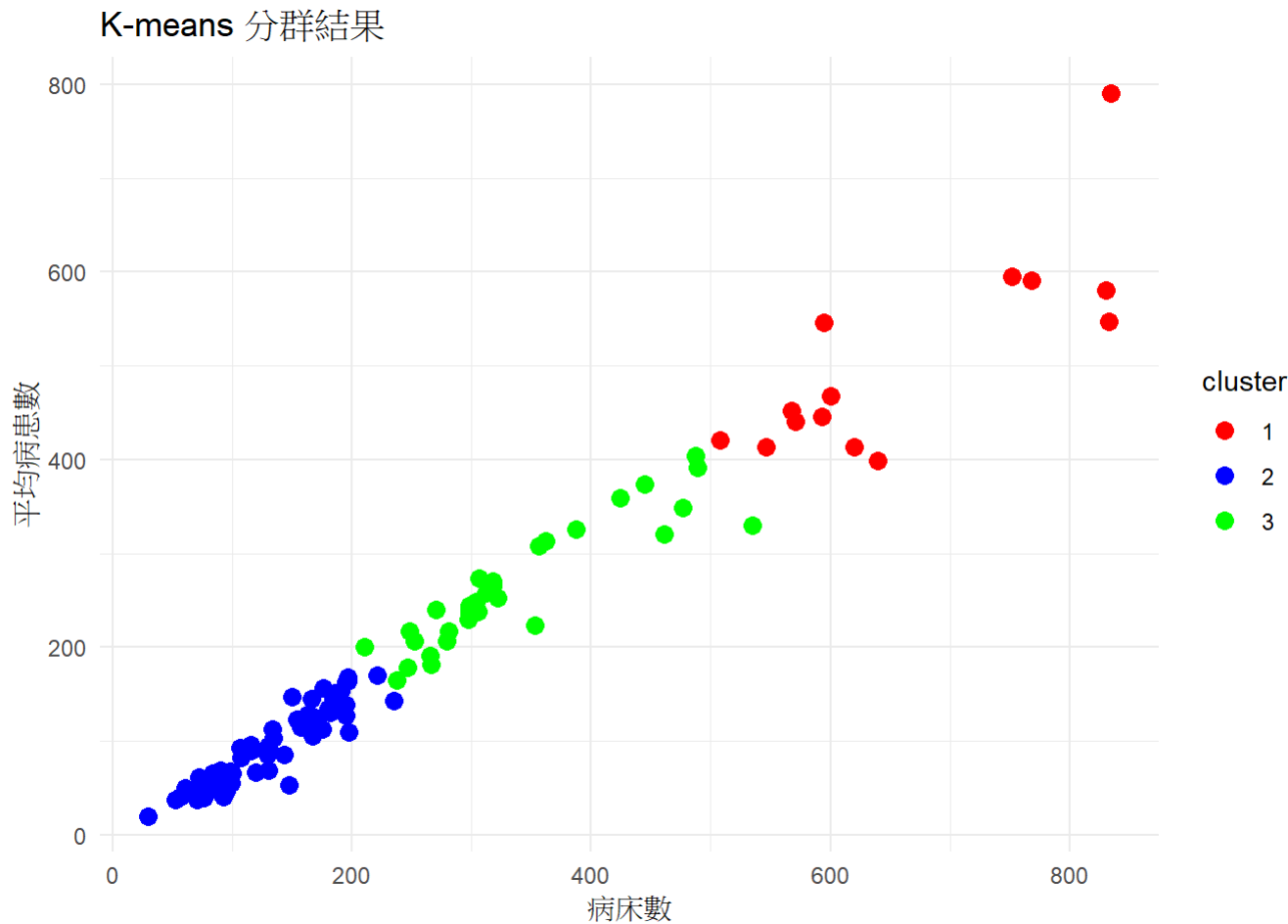
```
library(ggplot2)

# 確保數據格式正確 (移除醫院編號等非數值變數)
kmeans_data <- data[, c("停留日數", "年齡", "感染風險", "培養率",
                        "X光照射率", "病床數", "平均病患數", "護士數目", "可用設備")]

# 執行 K-means 分群 (分 3 群)
set.seed(123) # 設定隨機種子，確保結果一致
kmeans_result <- kmeans(kmeans_data, centers = 3)

# 加入分群結果
data$cluster <- as.factor(kmeans_result$cluster)

# **視覺化分群結果**
ggplot(data, aes(x=病床數, y=平均病患數, color=cluster)) +
  geom_point(size=3) +
  theme_minimal() +
  labs(title="K-means 分群結果", x="病床數", y="平均病患數") +
  scale_color_manual(values=c("red", "blue", "green"))
```



觀察討論

我這邊使用「病床數」與「平均病患數」來分群

這樣的分群可能代表不同規模的醫院，例如：

- Cluster 1 (小型醫院)：病床少、平均病患數較少
- Cluster 2 (中型醫院)：病床數中等、病患數適中
- Cluster 3 (大型醫院)：病床多、每日病患數多

在分群後，我想要利用ANOVA來測試不同分群是否在各變數上有顯著差異。

```
# 確保 cluster 是因子類別變數
data$cluster <- as.factor(data$cluster)
```

```
# 進行 ANOVA 分析
```

```
anova_1 <- aov(停留日數 ~ cluster, data = data)
anova_2 <- aov(感染風險 ~ cluster, data = data)
anova_3 <- aov(年齡 ~ cluster, data = data)
anova_4 <- aov(培養率 ~ cluster, data = data)
anova_5 <- aov(X光照射率 ~ cluster, data = data)
anova_6 <- aov(病床數 ~ cluster, data = data)
anova_7 <- aov(平均病患數 ~ cluster, data = data)
anova_8 <- aov(護士數目 ~ cluster, data = data)
anova_9 <- aov(可用設備 ~ cluster, data = data)
```

```
# 顯示結果
```

```
summary(anova_1)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## cluster         2   69.2    34.62    11.2 3.73e-05 ***
## Residuals     110  340.0     3.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_2)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## cluster         2   25.47   12.735    7.964 0.000589 ***
## Residuals     110 175.91     1.599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_3)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## cluster         2     0.4    0.177    0.009  0.991
## Residuals     110 2229.1   20.265
```

```
summary(anova_4)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## cluster         2    161    80.69    0.767  0.467
## Residuals     110 11571   105.19
```

```
summary(anova_5)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## cluster         2    516   258.1    0.685  0.506
## Residuals     110 41479   377.1
```

```
summary(anova_6)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## cluster      2 3606852 1803426   355.4 <2e-16 ***
## Residuals    110  558237    5075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_7)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## cluster      2 2249876 1124938   310.9 <2e-16 ***
## Residuals    110  398028    3618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_8)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## cluster      2 1736698  868349   219.3 <2e-16 ***
## Residuals    110  435525    3959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_9)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## cluster      2  14908    7454   74.74 <2e-16 ***
## Residuals    110  10971     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

觀察討論

顯著影響分群的變數 ($p < 0.05$)

- 停留日數 ($p = 3.73e-05$)：不同群組的醫院住院天數明顯不同。
- 年齡 ($p = 0.000589$)：不同群組的病患平均年齡有顯著差異。
- 病床數 ($p < 2e-16$)：不同群組的醫院規模 (病床數) 差異極大。
- 平均病患數 ($p < 2e-16$)：分群之間的每日病患數有明顯區別。
- 護士數目 ($p < 2e-16$)：不同群組的護士人數明顯不同，表示醫院規模不同。
- 可用設備 ($p < 2e-16$)：不同群組的醫療設備有顯著差異。### 無顯著影響的變數 ($p > 0.05$)
- 感染風險 ($p = 0.991$)：表示不同群組的醫院 感染風險相似，這可能表示感染風險不一定與醫院規模有關，可能受其他因素影響，如醫療管理或衛生條件。
- 培養率 ($p = 0.467$) 和 X光照射率 ($p = 0.506$) 也沒有顯著差異，可能這些醫療行為在不同醫院間並沒有明顯的區別。

結果與心得

基於 ANOVA 結果，我們可以推測 K-means 分群可能代表醫院的規模 (小型/中型/大型)：

- Cluster 1 (小型醫院)

- 病床數少
- 護士數少
- 設備少
- 住院日數較短
- Cluster 2 (中型醫院)
 - 病床數適中
 - 護士數適中
 - 設備適中
 - 住院日數較長
- Cluster 3 (大型醫院)
 - 病床數多
 - 護士數多
 - 設備多
 - 住院日數更長

上面的圖表我們可以看到 K-means 分群有效區分不同規模的醫院，但與感染風險、培養率等變數的關聯不大。