410979068_施尚丞_HW2

施尚丞

2025-03-03

本周作業目標

- 根據資料的變數，製作多張圖表。
- 繪製一張有明顯效果的圖表，並說明；
- 繪製一張以為有效果，但是沒有的圖表，並說明。

資料說明

本數據來源為州立大學入學資料，目的是研究新鮮人之 GPA 是否與其高中排名與 ACT 測驗成績有關聯，資料時間範圍為 1996 年至 2000 年，包含以下欄位：

- **編號 (Index)**：學生編號 (1-705)
- **GPA**：大學第一學年的平均成績
- **高中排名**：高中時期班上成績排名
- **ACT**：ACT 入學測驗分數
- **學年 (Year)**：新鮮人入學年份

資料讀取與前處理

```
df <- read.table("C:/Users/SHI/Desktop/R/113-2/APPENC04.txt", header = FALSE, sep = "", strip.white = T
RUE)
colnames(df) <- c("Index", "GPA", "High_School_Rank", "ACT", "Year")
head(df)
```

```
##   Index  GPA High_School_Rank ACT Year
## 1     1 0.98               61  20 1996
## 2     2 1.13               84  20 1996
## 3     3 1.25               74  19 1996
## 4     4 1.32               95  23 1996
## 5     5 1.48               77  28 1996
## 6     6 1.57               47  23 1996
```

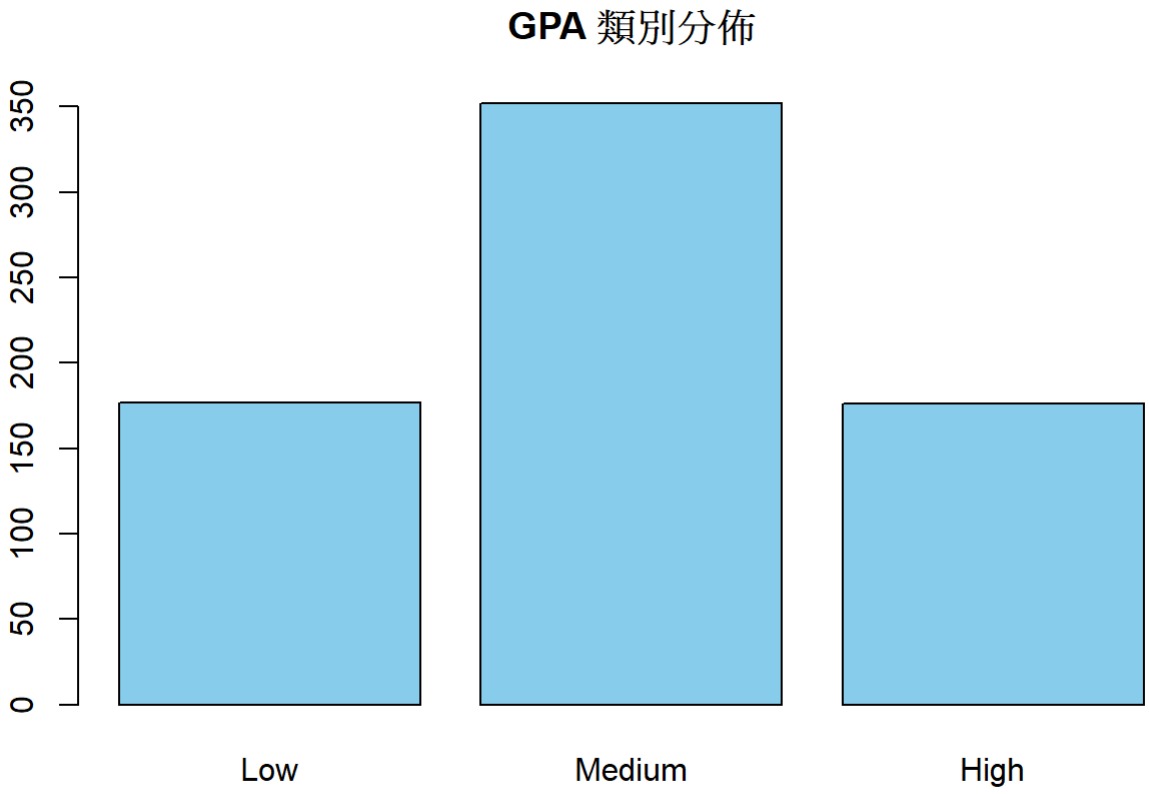
```
summary(df)
```

```
##      Index          GPA        High_School_Rank      ACT             Year
##  Min.   :  1   Min.   :0.510   Min.   : 4.00    Min.   :13.00   Min.   :1996
##  1st Qu.:177   1st Qu.:2.609   1st Qu.:68.00    1st Qu.:22.00   1st Qu.:1997
##  Median :353   Median :3.050   Median :81.00    Median :25.00   Median :1998
##  Mean   :353   Mean   :2.977   Mean   :76.95    Mean   :24.54   Mean   :1998
##  3rd Qu.:529   3rd Qu.:3.470   3rd Qu.:92.00    3rd Qu.:28.00   3rd Qu.:1999
##  Max.   :705   Max.   :4.000   Max.   :99.00    Max.   :35.00   Max.   :2000
```

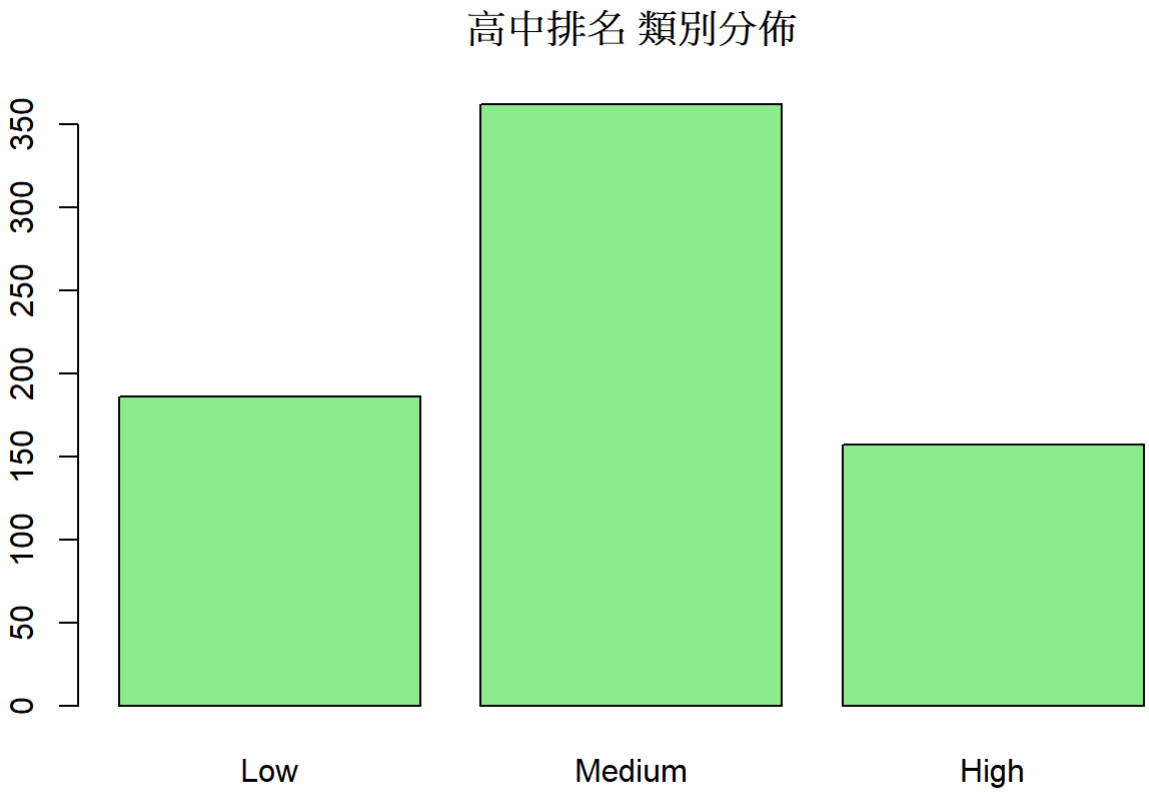
```
df$GPA_cat <- cut(df$GPA, breaks = quantile(df$GPA, probs = c(0, 0.25, 0.75, 1), na.rm = TRUE),
                  include.lowest = TRUE, labels = c("Low", "Medium", "High"))
df$rank_cat <- cut(df$High_School_Rank, breaks = quantile(df$High_School_Rank, probs = c(0, 0.25, 0.75,
1), na.rm = TRUE),
                   include.lowest = TRUE, labels = c("Low", "Medium", "High"))
df$ACT_cat <- cut(df$ACT, breaks = quantile(df$ACT, probs = c(0, 0.25, 0.75, 1), na.rm = TRUE),
                  include.lowest = TRUE, labels = c("Low", "Medium", "High"))
```

單一類別變數的長條圖

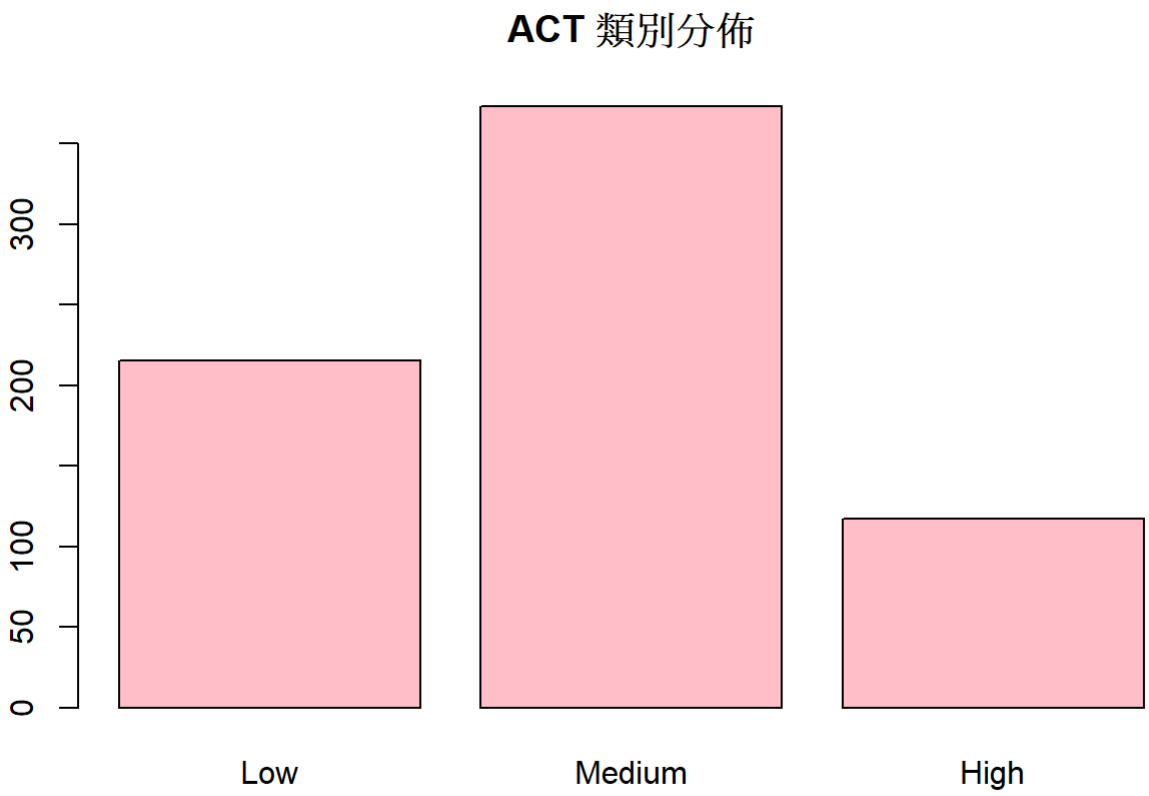
```
barplot(table(df$GPA_cat), main="GPA 類別分佈", col="skyblue", border="black")
```




```
barplot(table(df$rank_cat), main="高中排名 類別分佈", col="lightgreen", border="black")
```




```
barplot(table(df$ACT_cat), main="ACT 類別分佈", col="pink", border="black")
```



觀察討論:

GPA 的分佈以 Medium 類別為主，顯示大部分學生的 GPA 在中等範圍；高中排名的分佈顯示多數學生的排名位於中等區間，且高排名 (Low) 的比例相對較少；ACT 成績的分佈中，高分 (High) 的比例較低，多數集中於 Medium 類別。

```
# 兩個類別變數的列聯表與比例表

## GPA_cat 和 rank_cat
tab1 <- table(df$GPA_cat, df$rank_cat)
cat("GPA 與 高中排名 的列聯表:")
```

```
## GPA 與 高中排名 的列聯表:
```

```
print(tab1)
```

```
##
##          Low Medium High
##   Low     73     95    9
##   Medium  95    200   57
##   High    18     67   91
```

```
print(prop.table(tab1))
```

```
##
##                 Low     Medium       High
##   Low    0.10354610 0.13475177 0.01276596
##   Medium 0.13475177 0.28368794 0.08085106
##   High   0.02553191 0.09503546 0.12907801
```

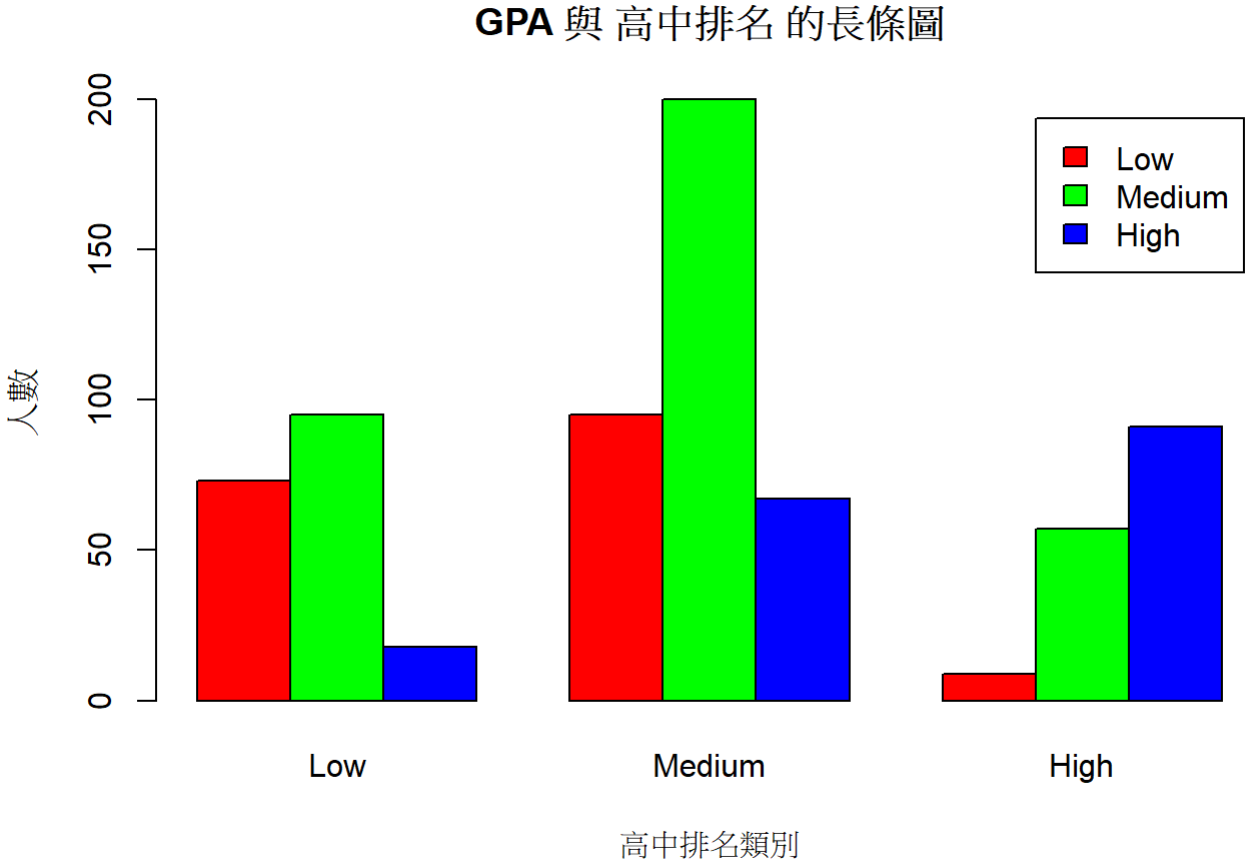
```
print(prop.table(tab1, margin = 1))
```

```
##
##                 Low     Medium       High
##   Low    0.41242938 0.53672316 0.05084746
##   Medium 0.26988636 0.56818182 0.16193182
##   High   0.10227273 0.38068182 0.51704545
```

```
print(prop.table(tab1, margin = 2))
```

```
##
##                 Low     Medium       High
##   Low    0.39247312 0.26243094 0.05732484
##   Medium 0.51075269 0.55248619 0.36305732
##   High   0.09677419 0.18508287 0.57961783
```

```
barplot(tab1, beside=TRUE, legend = TRUE,
        main="GPA 與 高中排名 的長條圖",
        col=c("red", "green", "blue"),
        xlab="高中排名類別",
        ylab="人數")  # 增加X、Y軸標籤
```




```
## GPA_cat 和 ACT_cat
tab2 <- table(df$GPA_cat, df$ACT_cat)
cat("GPA 與 ACT 的列聯表:")
```

```
## GPA 與 ACT 的列聯表:
```

```
print(tab2)
```

```
##
##          Low Medium High
##   Low     78     88   11
##   Medium 118    189   45
##   High    19     96   61
```

```
print(prop.table(tab2))
```

```
##
##                 Low     Medium       High
##   Low    0.11063830 0.12482270 0.01560284
##   Medium 0.16737589 0.26808511 0.06382979
##   High   0.02695035 0.13617021 0.08652482
```

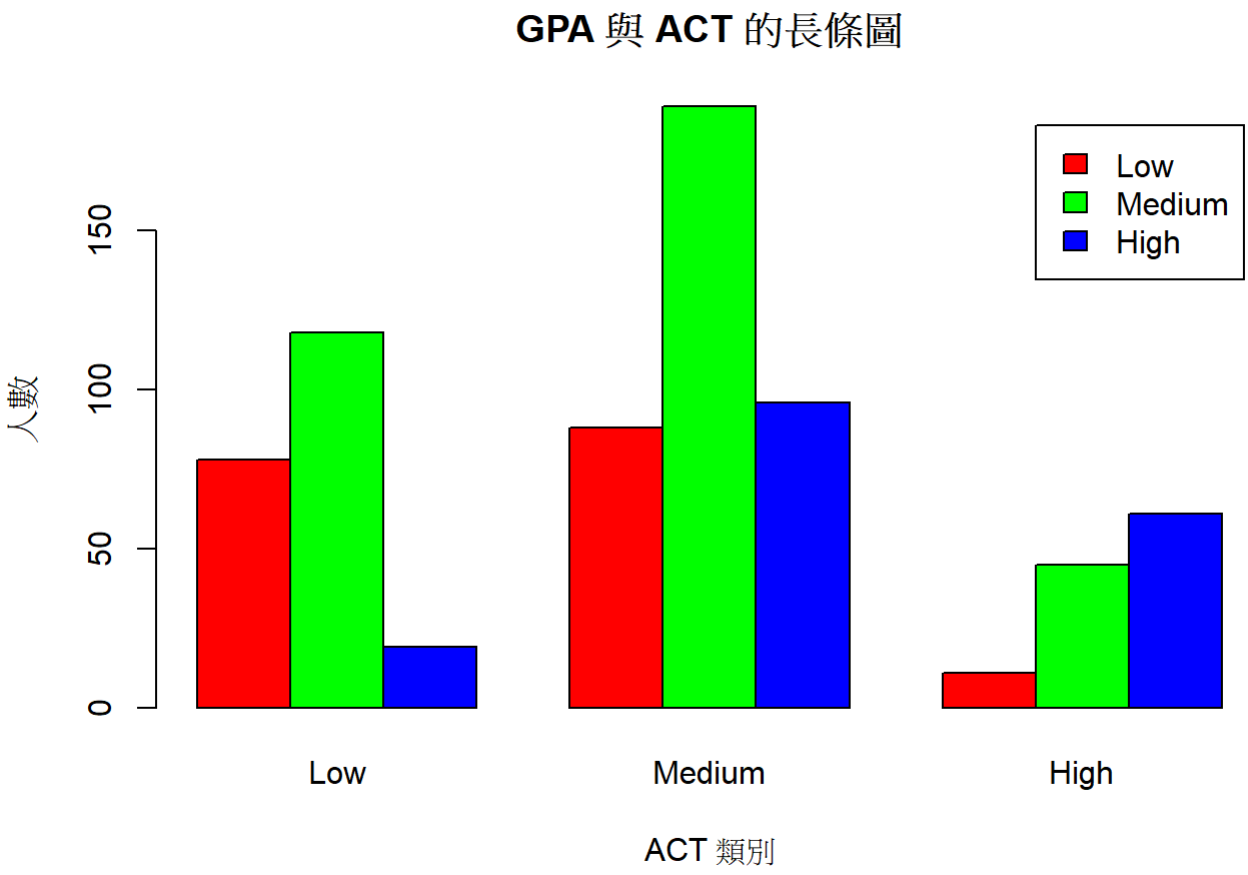
```
print(prop.table(tab2, margin = 1))
```

```
##
##                 Low     Medium       High
##   Low    0.44067797 0.49717514 0.06214689
##   Medium 0.33522727 0.53693182 0.12784091
##   High   0.10795455 0.54545455 0.34659091
```

```
print(prop.table(tab2, margin = 2))
```

```
##
##                 Low     Medium       High
##   Low    0.36279070 0.23592493 0.09401709
##   Medium 0.54883721 0.50670241 0.38461538
##   High   0.08837209 0.25737265 0.52136752
```

```
barplot(tab2, beside=TRUE, legend = TRUE,
        main="GPA 與 ACT 的長條圖",
        col=c("red", "green", "blue"),
        xlab="ACT 類別",
        ylab="人數")  # 增加X、Y軸標籤
```



觀察討論

ACT 成績越高，不一定 GPA 也較高，滿多人雖然GPA高但考ACT時成績卻在中等，這些人可能不太擅長考試。")

```
## rank_cat 和 ACT_cat
tab3 <- table(df$rank_cat, df$ACT_cat)
cat("高中排名 與 ACT 的列聯表:")
```

```
## 高中排名 與 ACT 的列聯表:
```

```
print(tab3)
```

```
##
##          Low Medium High
##   Low     87     89   10
##   Medium 109    204   49
##   High    19     80   58
```

```
print(prop.table(tab3))
```

```
##
##                 Low     Medium       High
##   Low    0.12340426 0.12624113 0.01418440
##   Medium 0.15460993 0.28936170 0.06950355
##   High   0.02695035 0.11347518 0.08226950
```

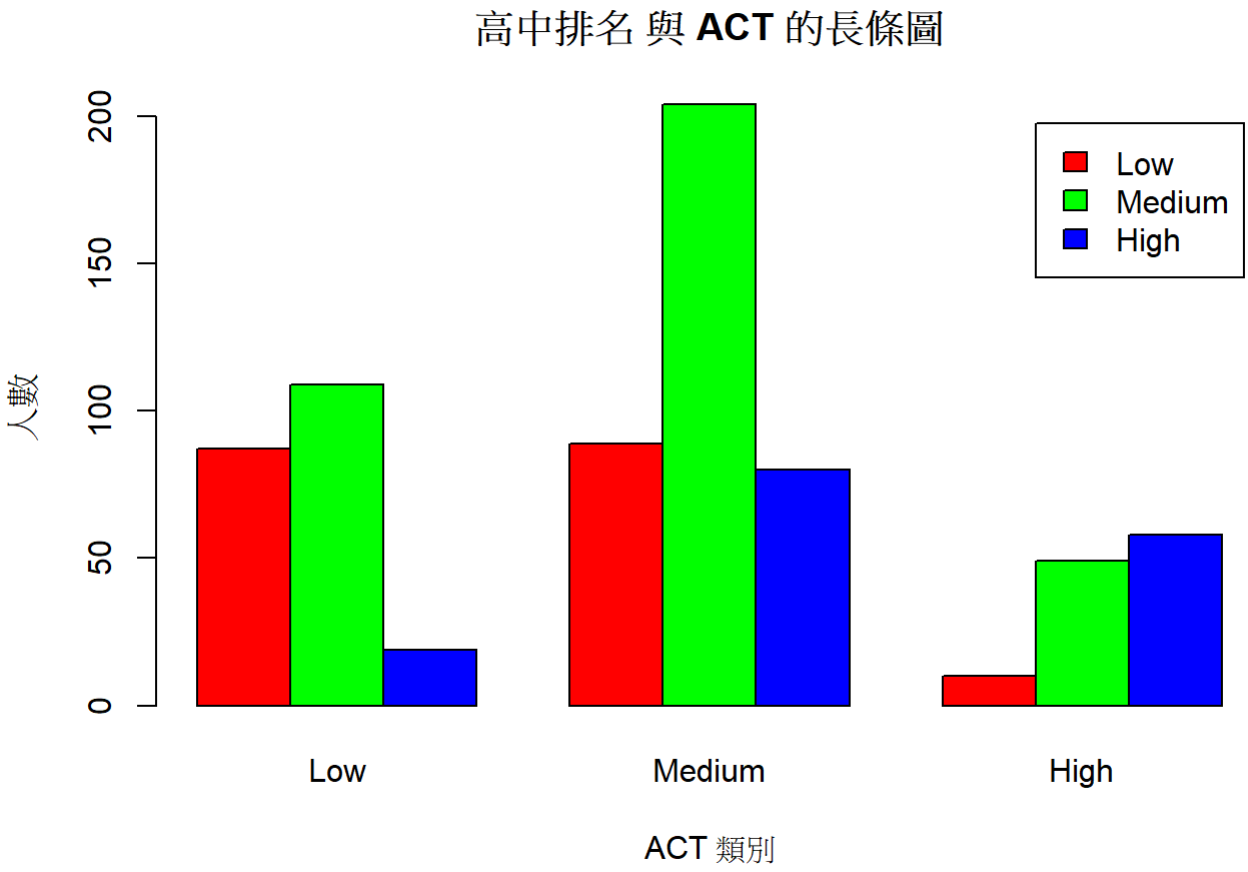
```
print(prop.table(tab3, margin = 1))
```

```
##
##                 Low     Medium       High
##   Low    0.46774194 0.47849462 0.05376344
##   Medium 0.30110497 0.56353591 0.13535912
##   High   0.12101911 0.50955414 0.36942675
```

```
print(prop.table(tab3, margin = 2))
```

```
##
##                 Low     Medium       High
##   Low    0.40465116 0.23860590 0.08547009
##   Medium 0.50697674 0.54691689 0.41880342
##   High   0.08837209 0.21447721 0.49572650
```

```
barplot(tab3, beside=TRUE, legend = TRUE,
        main="高中排名 與 ACT 的長條圖",
        col=c("red", "green", "blue"),
        xlab="ACT 類別",
        ylab="人數")  # 增加X、Y軸標籤
```

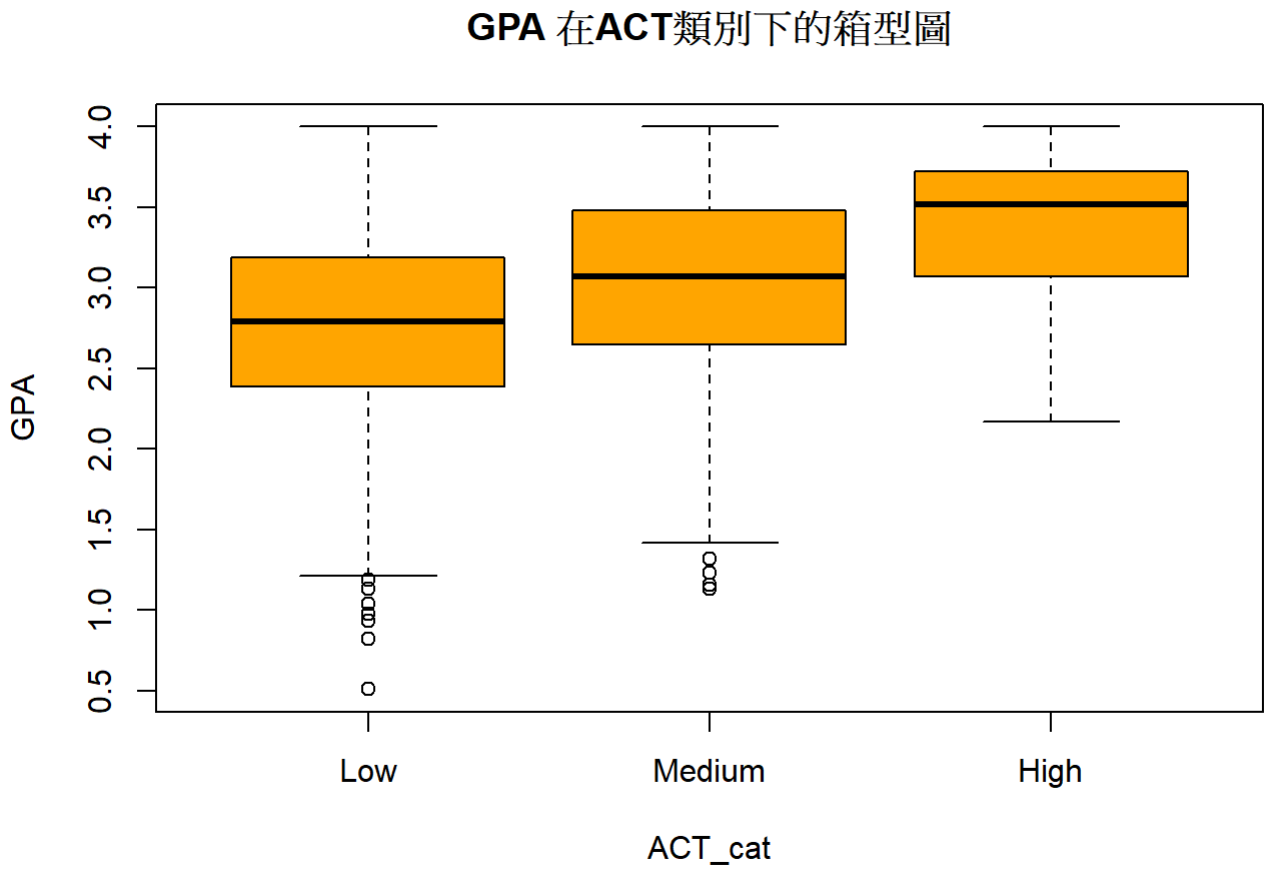


觀察討論:

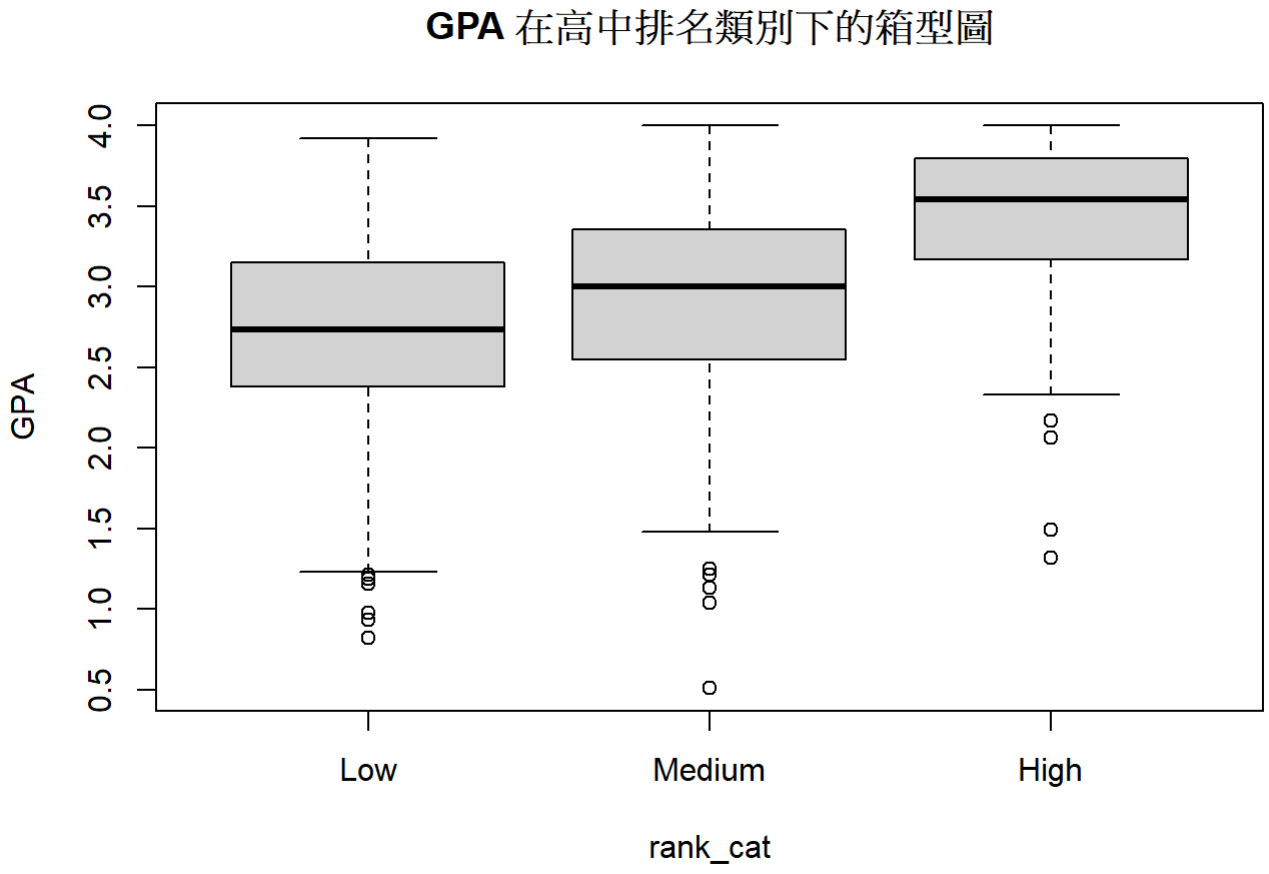
這邊可以觀察到圖表的型態跟上一個圖表類似，可以發現有一部分的人雖然高中排名高，但不擅長考試，畢竟GPA會影響到高中排名也是相對直觀的。

單一連續變數在類別變數下的表現

```
boxplot(GPA ~ ACT_cat, data=df, main="GPA 在ACT類別下的箱型圖", col="orange")
```



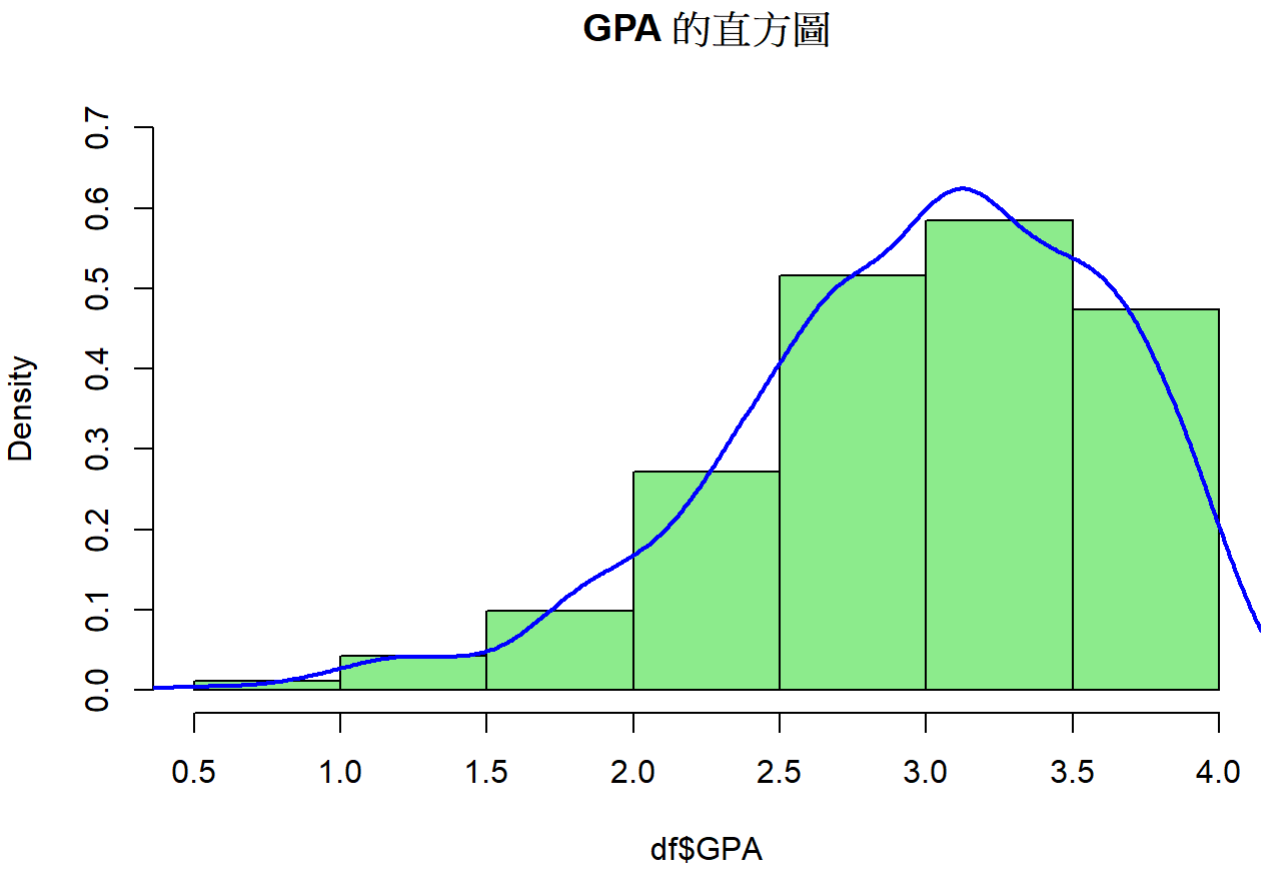

```
boxplot(GPA ~ rank_cat, data=df, main="GPA 在高中排名類別下的箱型圖", col="lightgray")
```



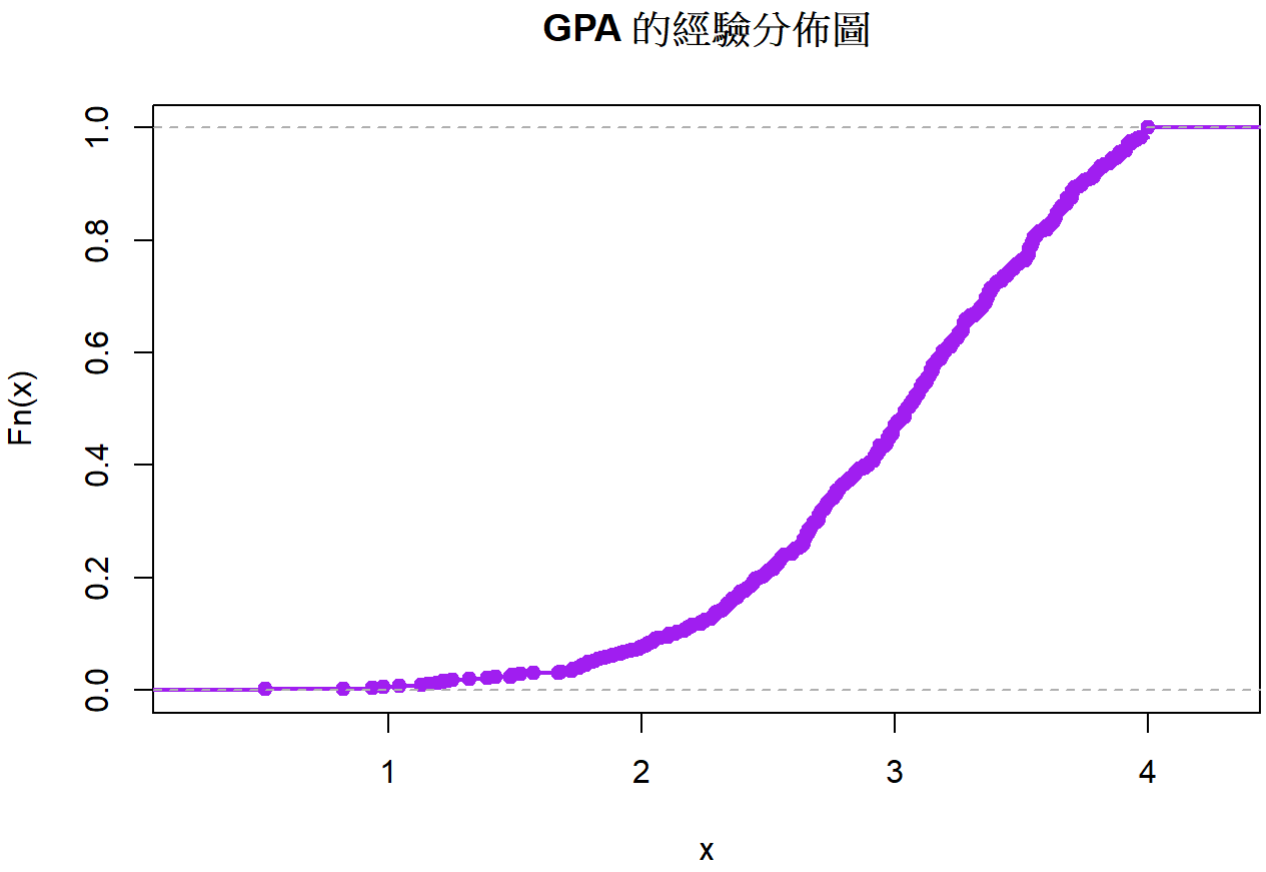
觀察討論:

ACT 分數越高，GPA 的中位數也越高，但高 ACT 組別的 GPA 分佈較分散；高中排名越高，GPA 的中位數也越高，並且 GPA 的分佈較為集中，顯示一致的趨勢。

```
hist(df$GPA, probability = TRUE, main="GPA 的直方圖", col="lightgreen", ylim=c(0, 0.7))
lines(density(df$GPA, na.rm = TRUE), col="blue", lwd=2)
```




```
plot(ecdf(df$GPA), main="GPA 的經驗分佈圖", col="purple", lwd=2)
```

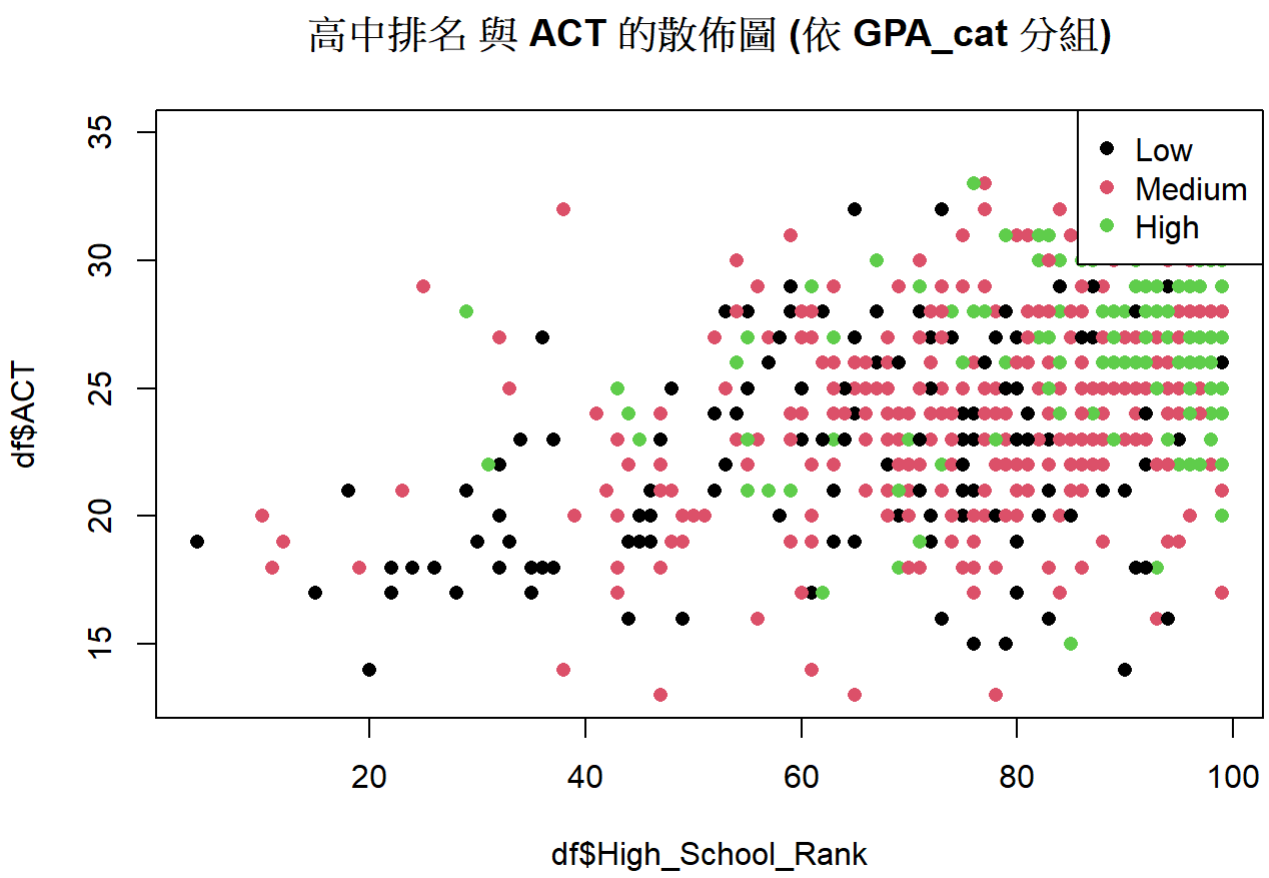


觀察討論:

GPA 的分佈接近常態分佈，但右尾稍長，顯示有少部分學生的 GPA 高於大多數人；而GPA 的累積分佈圖顯示大約 50% 的學生 GPA 低於中位數，分佈相對均勻，但右尾較長。

所有連續變數的散佈圖 (依 GPA_cat 分組)

```
plot(df$High_School_Rank, df$ACT, col=as.numeric(df$GPA_cat), main="高中排名 與 ACT 的散佈圖 (依 GPA_cat 分組)", pch=16)
legend("topright", legend = levels(df$GPA_cat), col=1:3, pch=16)
```

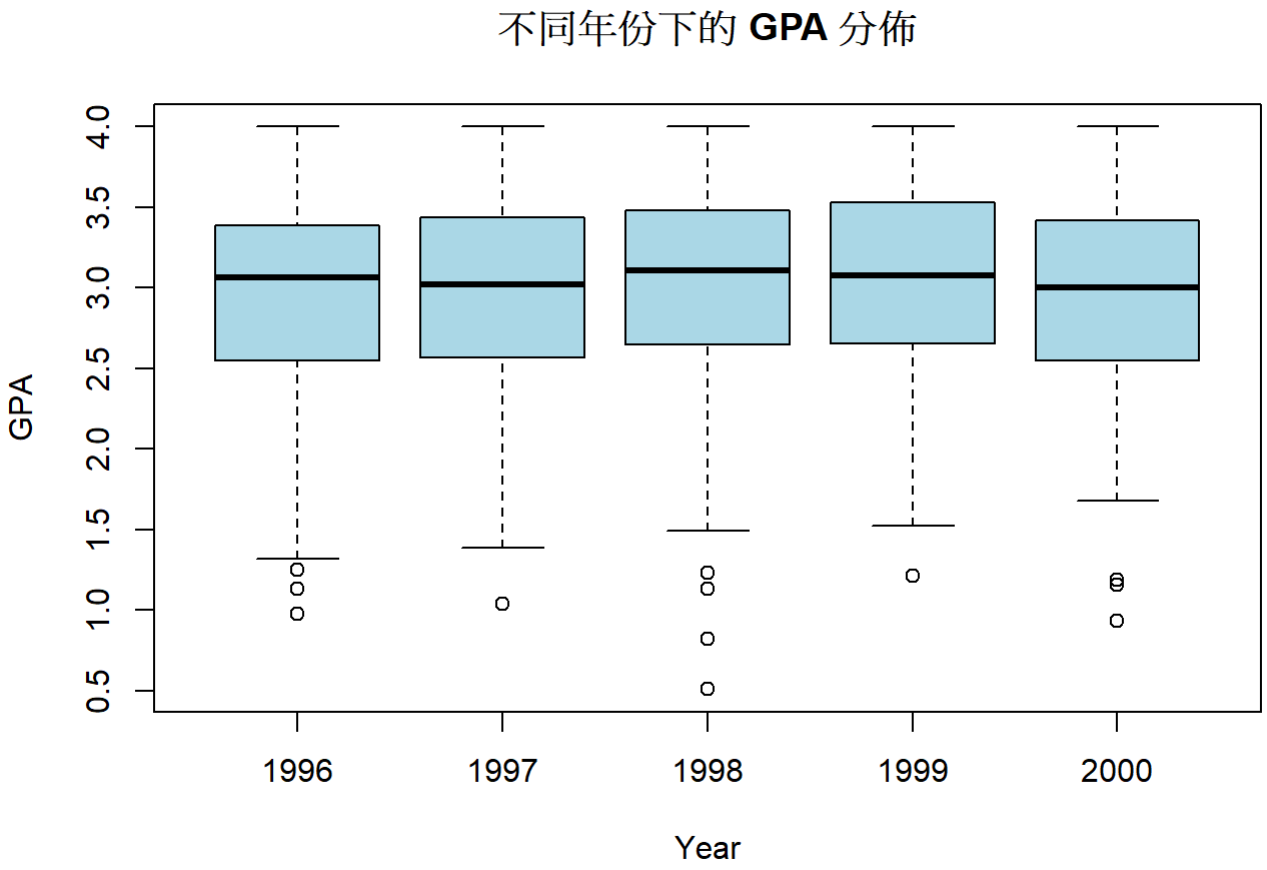


觀察討論:

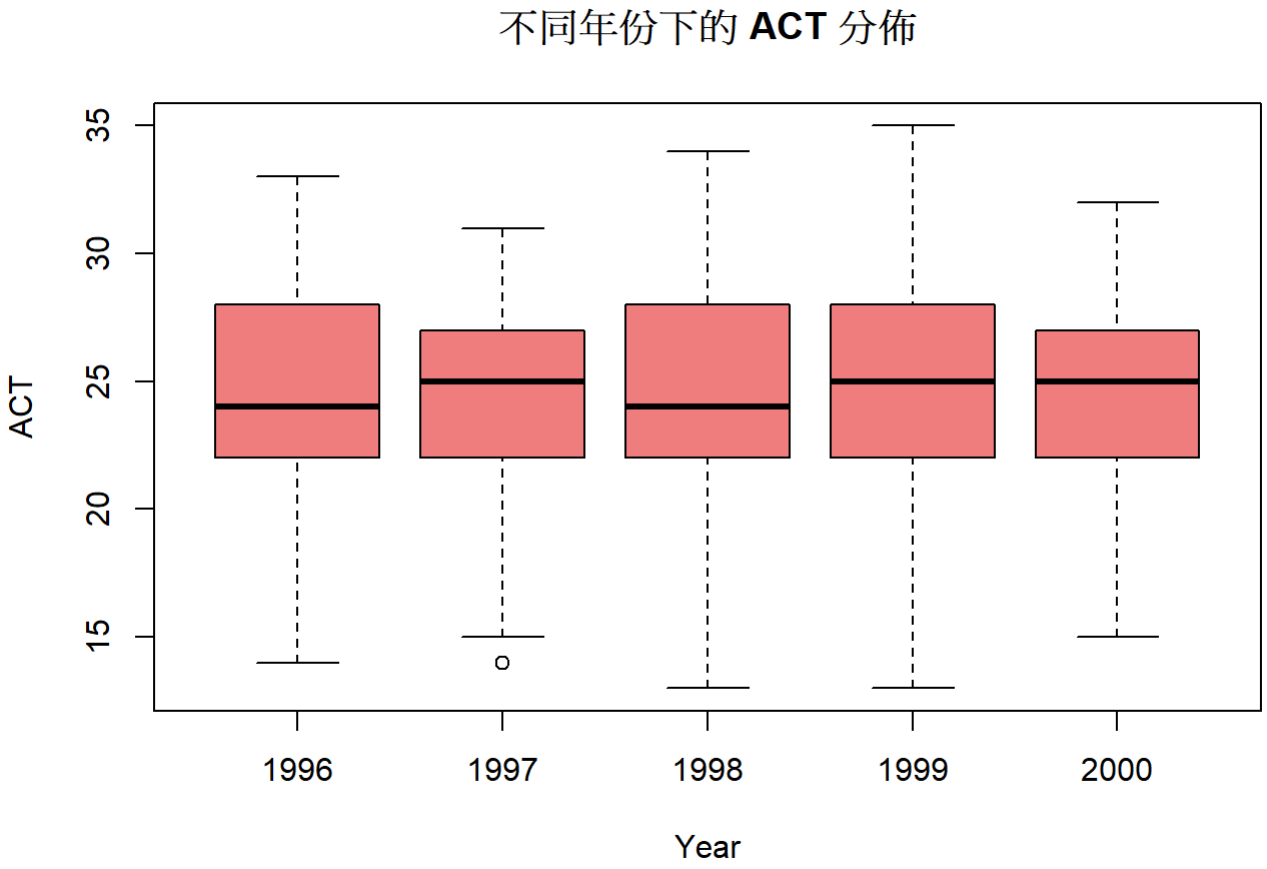
GPA 分組對高中排名與 ACT 的關係有明顯影響，高 GPA 組別集中於高排名與高 ACT 區域。

年份 (Year) 的分析

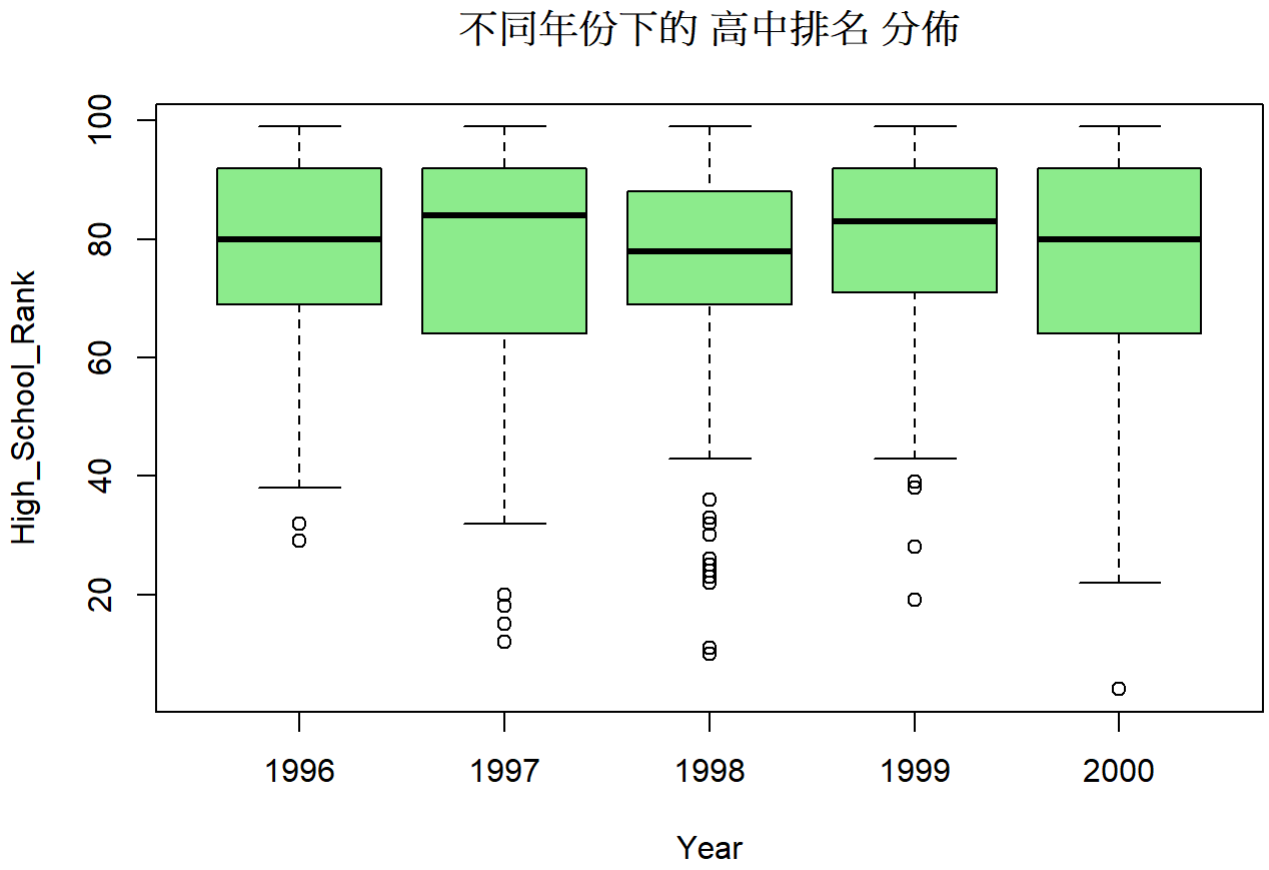
```
# GPA 按年份的箱型圖
boxplot(GPA ~ Year, data=df, main="不同年份下的 GPA 分佈", col="lightblue")
```




```
# ACT 按年份的箱型圖
boxplot(ACT ~ Year, data=df, main="不同年份下的 ACT 分佈", col="lightcoral")
```




```
# 高中排名按年份的箱型圖
boxplot(High_School_Rank ~ Year, data=df, main="不同年份下的 高中排名 分佈", col="lightgreen")
```



觀察討論

這五年學生GPA的表現沒有相差太多，甚至中位數與全距的表現也極為類似，可以看出學術評分標準較一致。就ACT和高中排名的箱型圖而言，雖然有些許的波動，但整體而言沒有相差太多。

結果與觀察

根據上述分析，我發現：

GPA與高中排名不見得會全然反映在ACT的表現上，但通常好的GPA伴隨著好的高中排名，反之如此。

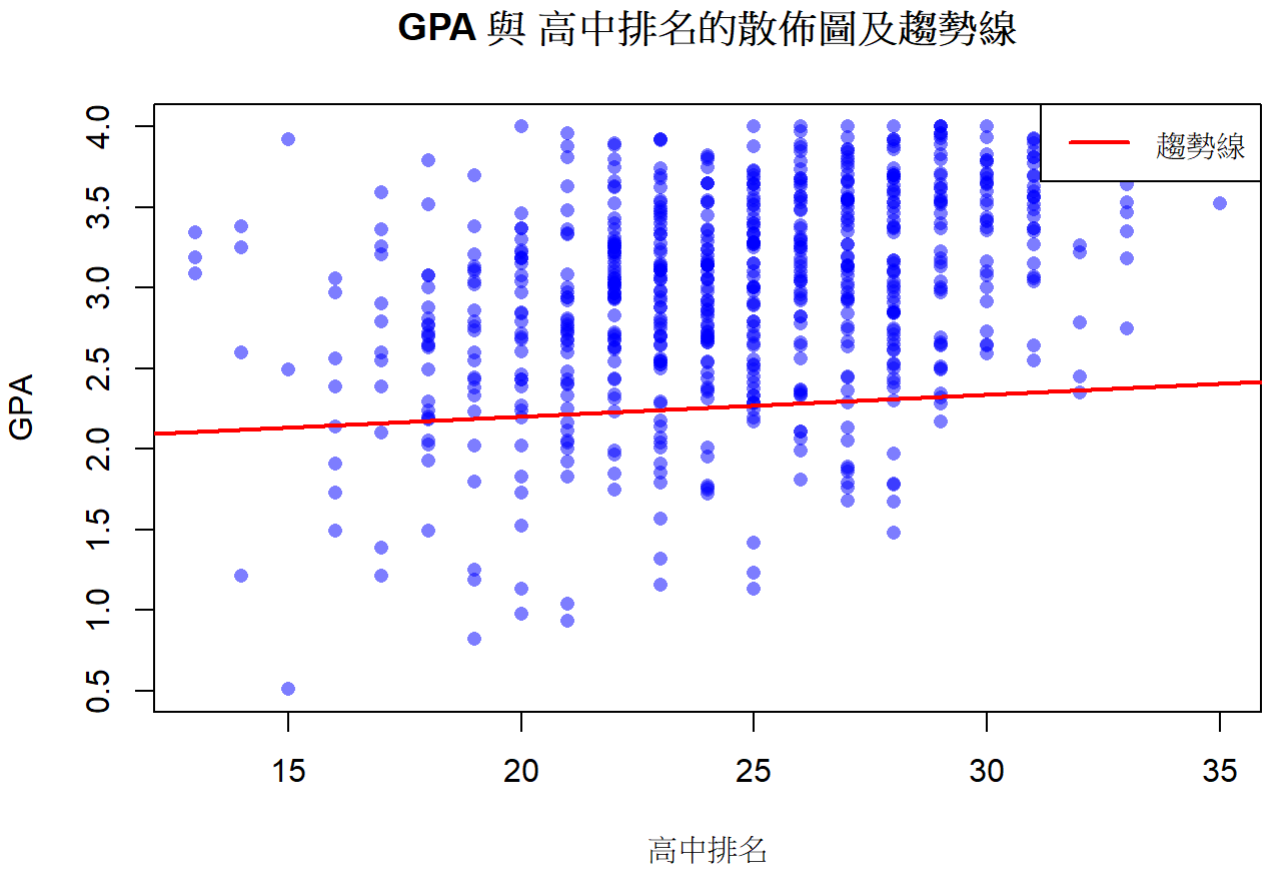
有趣的發現

我自己在做這份作業時，發現了一個我自己第一眼無法解讀的資訊: 「為什麼ACT高時GPA高的累積人數不高，但箱型圖的趨勢卻是ACT高的那群人GPA中位數高?」 後來經過更深入的解讀後，發現箱型圖中「ACT 越高的學生，通常GPA也較高」這個現象僅反映「每個群體內部」的分佈特徵，並不表示整體人數。 而長條圖中的High ACT組別的GPA中位數很高，但該組的整體人數較少，因此在長條圖上，藍色部分 (高GPA) 的總數自然不會太多。

看似第一時間矛盾的現象，實際上反映著「高分群體小而精」的特徵。

有明顯效果的圖表：GPA vs. 高中排名的散佈圖

```
plot(df$ACT, df$GPA, main="GPA 與 高中排名的散佈圖及趨勢線",
     xlab="高中排名", ylab="GPA", pch=16, col=rgb(0, 0, 1, 0.5))
model_rank <- lm(GPA ~ High_School_Rank, data=df)
abline(model_rank, col="red", lwd=2)
legend("topright", legend=c("趨勢線"), col="red", lwd=2)
```



觀察討論

雖然目前課程還沒有上到迴歸分析，但我覺得可以很好的完成這次作業的要求，因此納入其中。

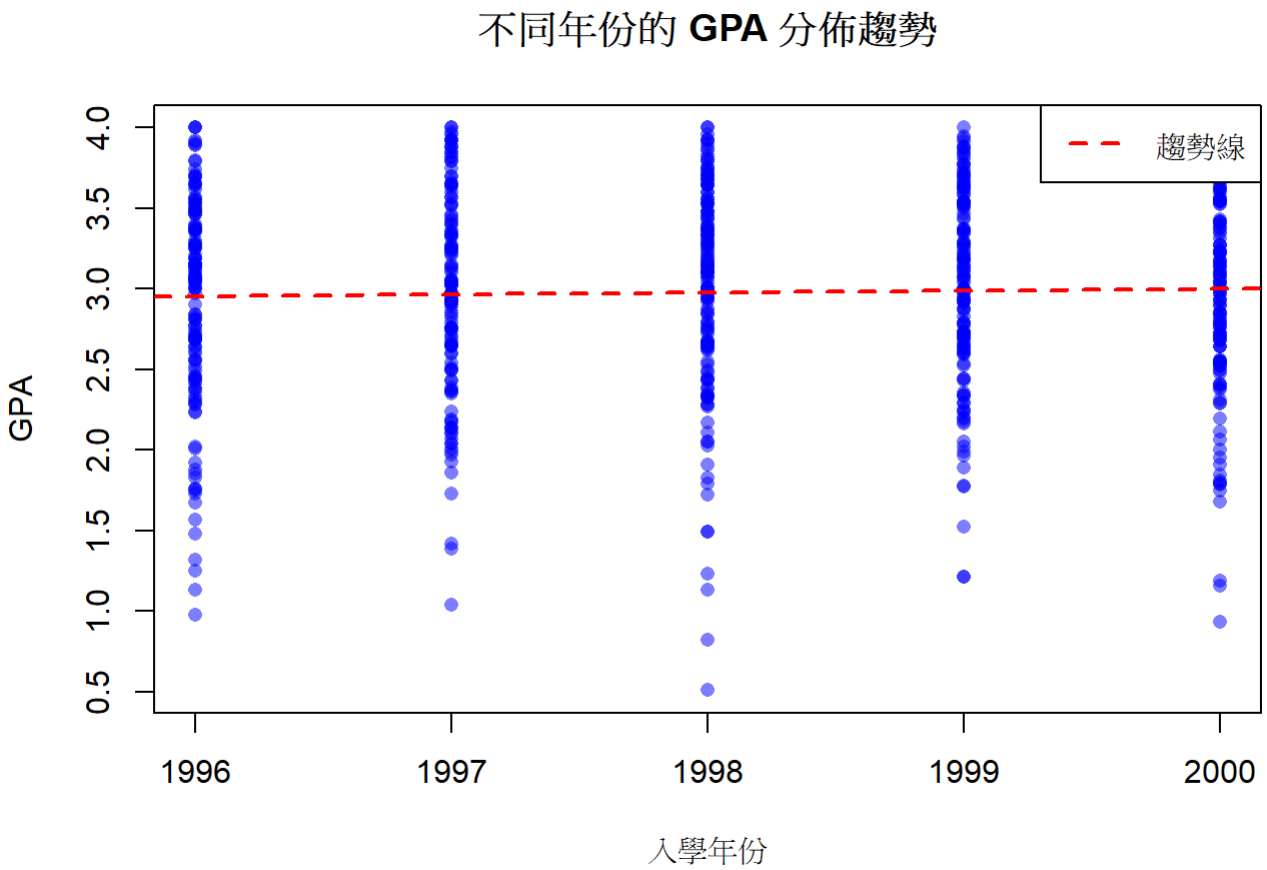
根據這個圖表跟趨勢線可以看到，高中排名與 GPA 呈現明顯的正相關趨勢，我在猜這邊的高中排名比較類似PR值，表示高中排名較高的學生GPA 也相對較高。這與我們的直覺一致，因為高中成績通常能反映到學科的學術能力表現。

不符合預設想法的圖表: GPA vs. 入學年份的散佈圖

```
# 繪製 GPA vs. 入學年份的散佈圖
plot(df$Year, df$GPA,
     main="不同年份的 GPA 分佈趨勢",
     xlab="入學年份", ylab="GPA",
     pch=16, col=rgb(0, 0, 1, 0.5))

# 建立回歸模型
model_year <- lm(GPA ~ Year, data=df)

# 加上回歸線
abline(model_year, col="red", lwd=2, lty=2)

# 顯示回歸結果
summary(model_year)
```

```
##
## Call:
## lm(formula = GPA ~ Year, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46731 -0.36004  0.08082  0.50269  1.04441
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.73006   34.04354  -0.550    0.582
## Year          0.01086    0.01704   0.638    0.524
##
## Residual standard error: 0.6348 on 703 degrees of freedom
## Multiple R-squared:  0.000578,   Adjusted R-squared:  -0.0008436
## F-statistic: 0.4066 on 1 and 703 DF,  p-value: 0.5239
```

```
# 加上圖例
legend("topright", legend=c("趨勢線"), col="red", lwd=2, lty=2)
```



觀察討論

原本想說有沒有可能隨著年分的推移，GPA有成績通膨的問題(近期台大就有在討論這個議題)。

但從這個圖表很明顯可以觀察到，趨勢線基本上是水平的，意味著GPA不會因為每一年而有太多的變動，像是每隔一年持續變高或是變低，基本上滿穩定的，GPA機制沒有太大問題。