0. 匯入要用到的資料和套件 並確認檔案內容

```
library(class)
library(tidyverse)
library(randomForest)
library(ggplot2)
library(factoextra)
library(dplyr)

setwd("C:\\Users\\ASUS\\Desktop\\五 234 R\\HW3") #放你的路徑
airline <- read.csv("airline_survey.csv")
airline<-na.omit(airline)
set.seed(100)
```

1. 辨認出滿意與不滿意客戶
   任選 1 種監督式學習方法配適模型，預測滿意度 satisfaction (2 類：滿意、中立或不滿意)。

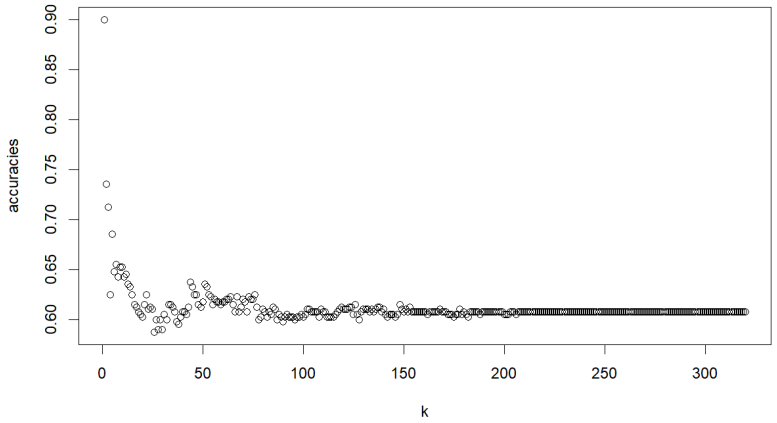
選 KNN 作為監督式學習方法配式模型

```
# 調整變數型態（文字轉為數字級距）
airline$satisfaction <- as.factor(ifelse(airline$satisfaction == "satisfied", 1, 0))
airline$Gender<- as.factor(ifelse(airline$Gender=="Male",1,0))
airline$Customer.Type<- as.factor(ifelse(airline$Customer.Type=="Loyal Customer",1,0))
airline$Type.of.Travel<- as.factor(ifelse(airline$Type.of.Travel=="Business travel",1,0))
airline$Class<-
as.factor(ifelse(airline$Class=="Business",2,ifelse(airline$Class=="Eco",1,0)))
airline<- airline %>% mutate(
  Flight.Distance = (Flight.Distance - min(Flight.Distance)) / (max(Flight.Distance) -
min(Flight.Distance)),
  Departure.Delay.in.Minutes= (Departure.Delay.in.Minutes - min(Departure.Delay.in.Minutes)) /
(max(Departure.Delay.in.Minutes) - min(Departure.Delay.in.Minutes)),
  Arrival.Delay.in.Minutes= (Arrival.Delay.in.Minutes -
min(Arrival.Delay.in.Minutes)) / (max(Arrival.Delay.in.Minutes) -
min(Arrival.Delay.in.Minutes))
)

# training and test data
airline.1<-airline[sample(1:nrow(airline),2000),]
traind<-airline.1[sample(1:nrow(airline.1),1600),]
testd<-airline.1[sample(1:nrow(airline.1),400),]

### KNN ###
pred=knn(traind[,2:24], testd[,2:24], cl=traind[,25], k = 6) #prob=T
table(real=testd[,25], pred)
```

右圖為程式部分截圖結果

```
      pred
real   0  1
   7   3  1
   8   2  4
   9   3  0
   10  0  3
   11  1  2
   12  3  1
   13  1  1
   14  3  1
   15  2  2
   16  5  1
   17  2  1
   18  2  1
   19  1  2
   20  5  4
   21  3  1
   22  2  2
   23  1  1
   24  5  4
   25  5  4
   26 10  3
```

```r
#choose k
range <- 1:round(0.2 * nrow(traind)) #通常 k 的上限為訓練樣本數的 20%
accuracies <- rep(NA, length(range))

for (i in range) {
  test_predicted <- knn(train = traind[,2:24], test = testd[,2:24], cl = traind[,25], k = i)
  conf_mat <- table(testd$satisfaction, test_predicted)
  accuracies[i] <- sum(diag(conf_mat))/sum(conf_mat)
}

##視覺化上面結果
plot(range, accuracies, xlab = "k")
```

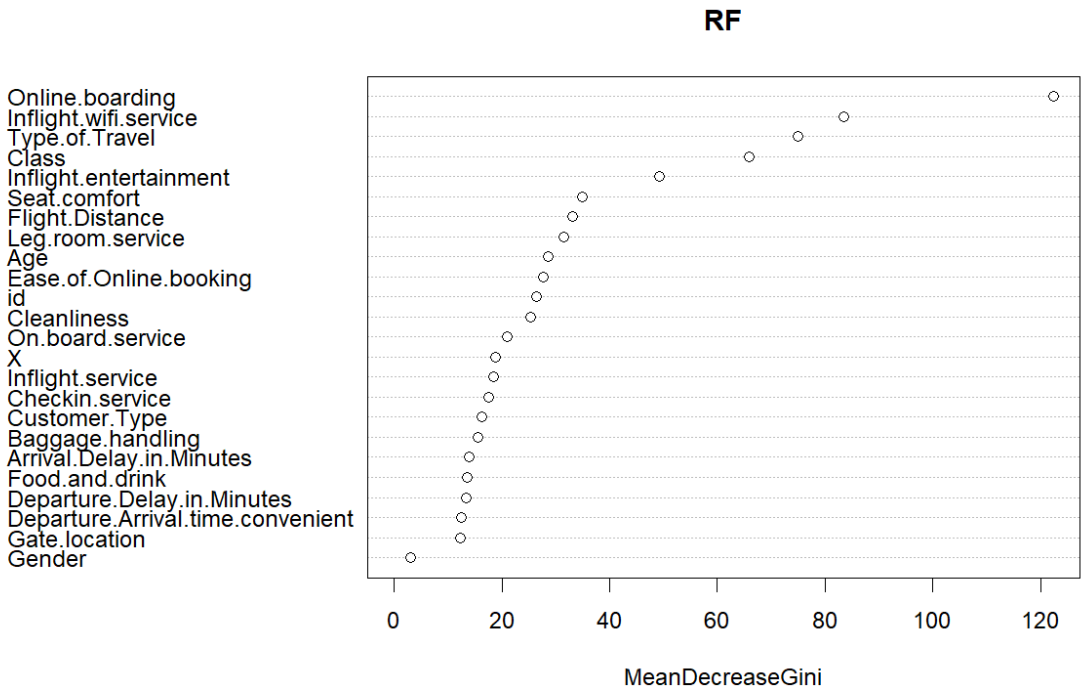



```r
which.max(accuracies) #k
```

```
[1] 1
```

可得知當 K=1 的時候有最好的預測結果

```r
# 找出重要變數：哪些因素影響客戶滿意度。
RF <- randomForest(satisfaction~. ,data=traind, importane = T, na.action = na.omit)
importance(RF)
varImpPlot(RF)
```



數值越大代表越重要

可以看出 Online.boarding 最重要且遠高於其他變數

接下來 Inflight.wifi.service、Type.of.Travel、Class、 Inflight.entertainment 這四項都有達到 40 以上，明顯高過其他變數

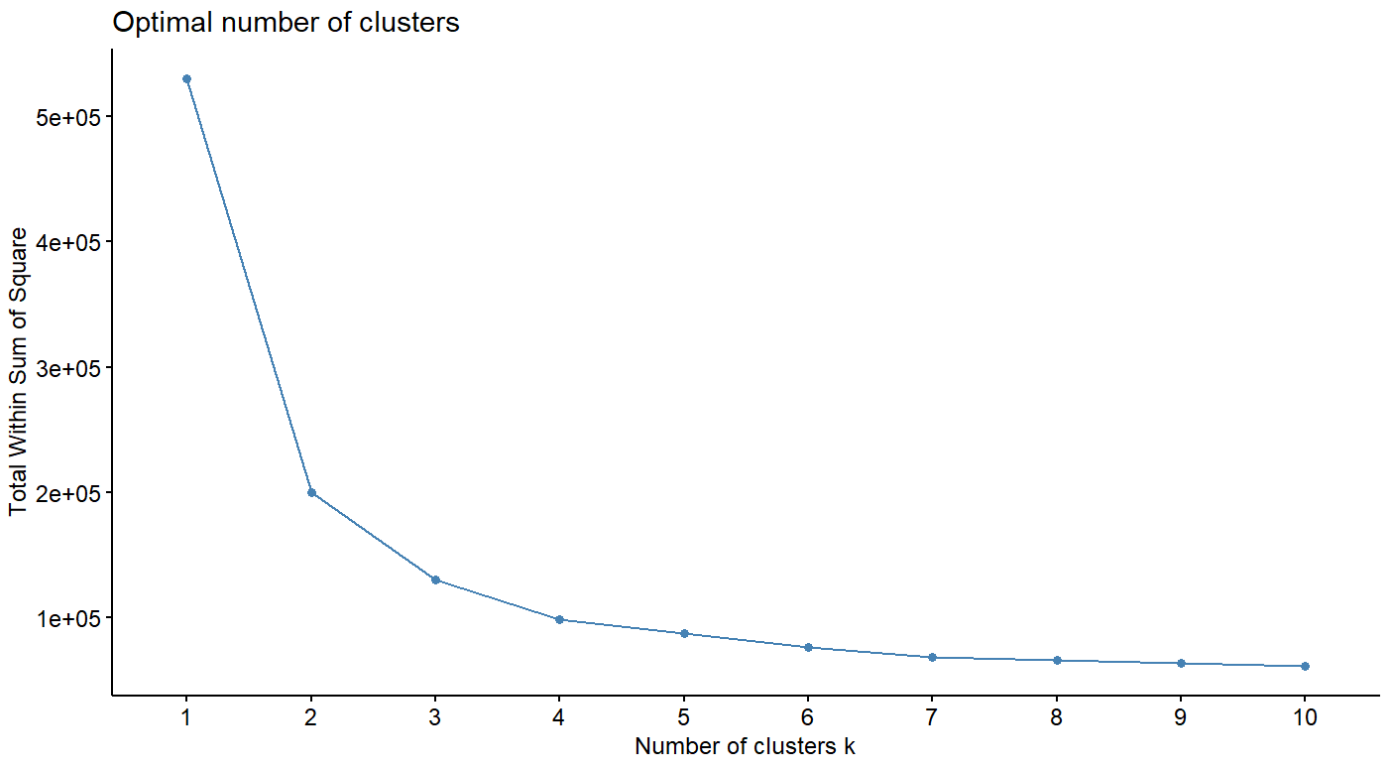
而 Gender 則趨近於 0，可看出其不是重要變數

2. 描述客戶

任選 1 種非監督式方法，將客戶分群，介紹你分出來的群，對於這些不同的客戶群集提出給該航空業的商業策略建議。

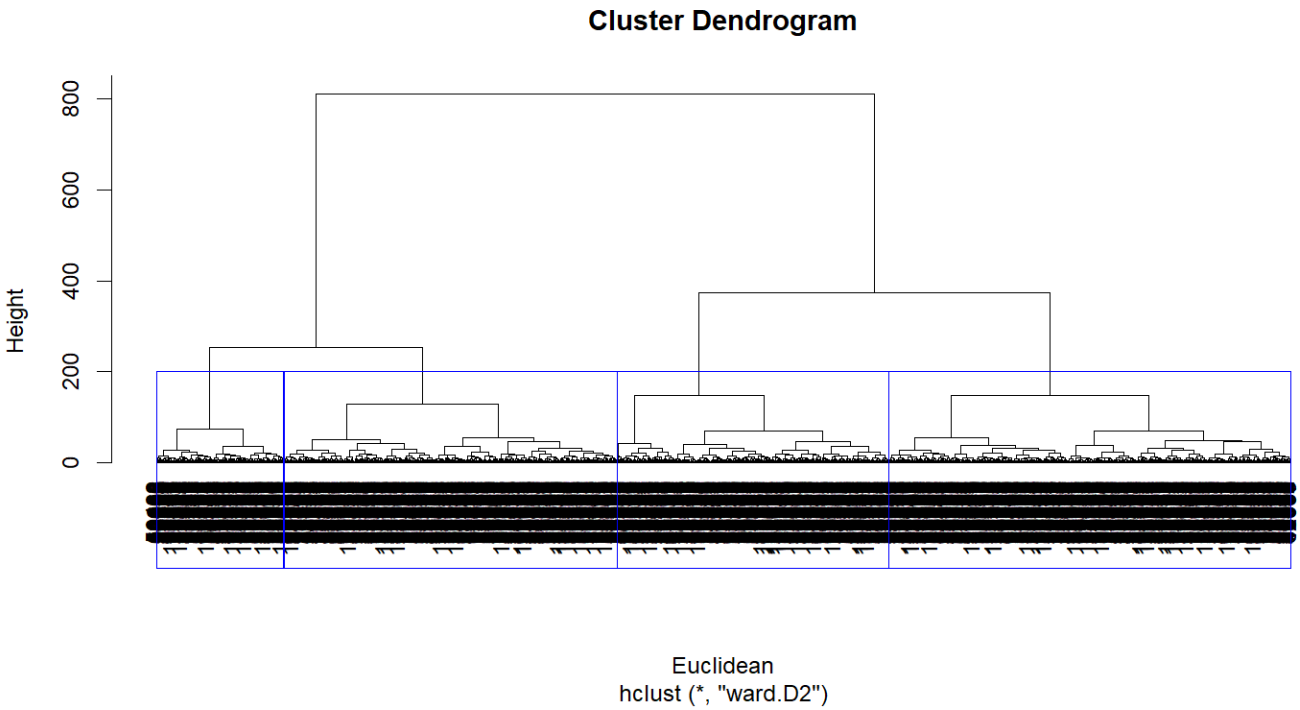
註：不需使用所有變數，可以先篩選你覺得有用的變數再去做分析。不需做訓練集、測試集。若電腦無法讀取大資料，可任選部分資料讀取。

```
airline.2<-airline.1[,-c(1:2)]
airline.2<-na.omit(airline.2)

# 歐式距離
E.dist <- dist(airline.2, method="euclidean")
tree1 <- hclust(E.dist, method="ward.D2")
fviz_nbclust(airline.2, FUN = hcut, method = "wss")

plot(tree1, xlab="Euclidean",h=-1) #h=-1
```

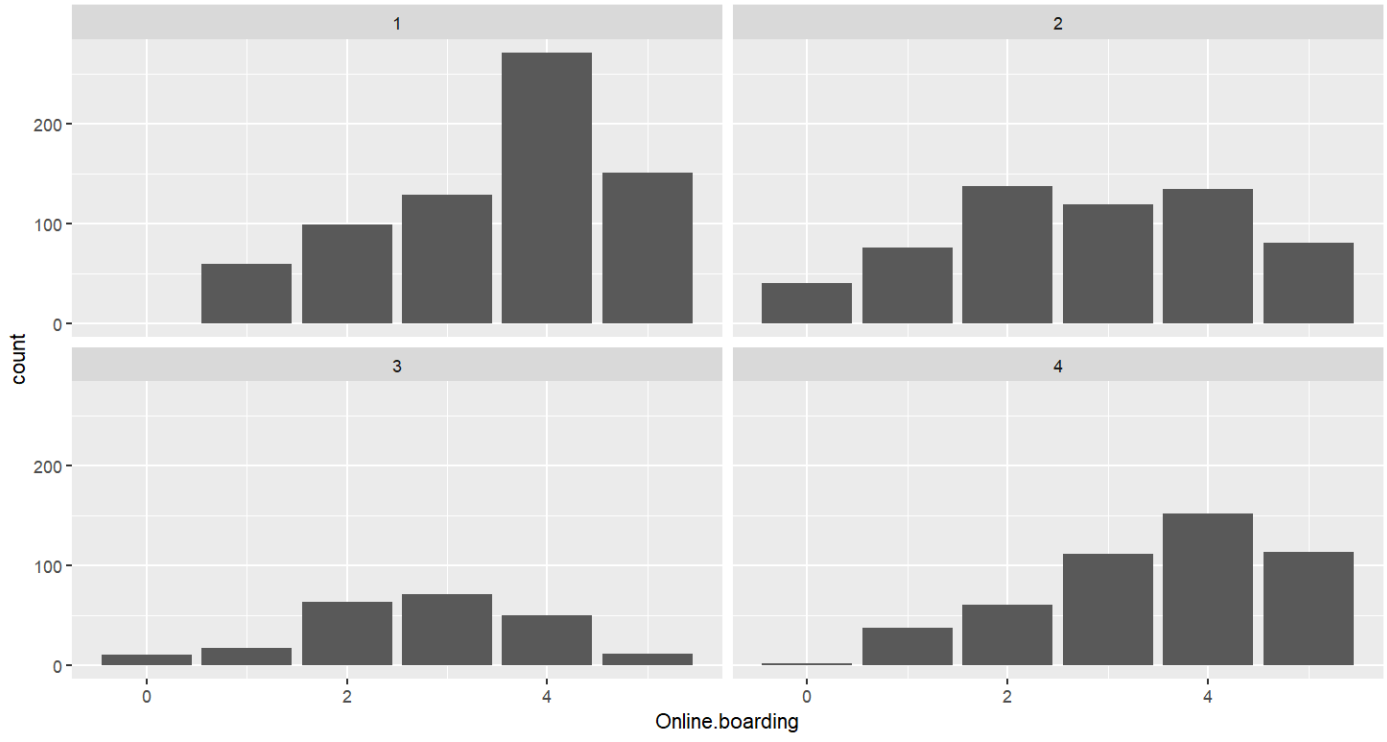



```
#觀察過 tree 之後決定分成四群
rect.hclust(tree1,k=4,border="blue")

cluster <- cutree(tree1, k=4)
```

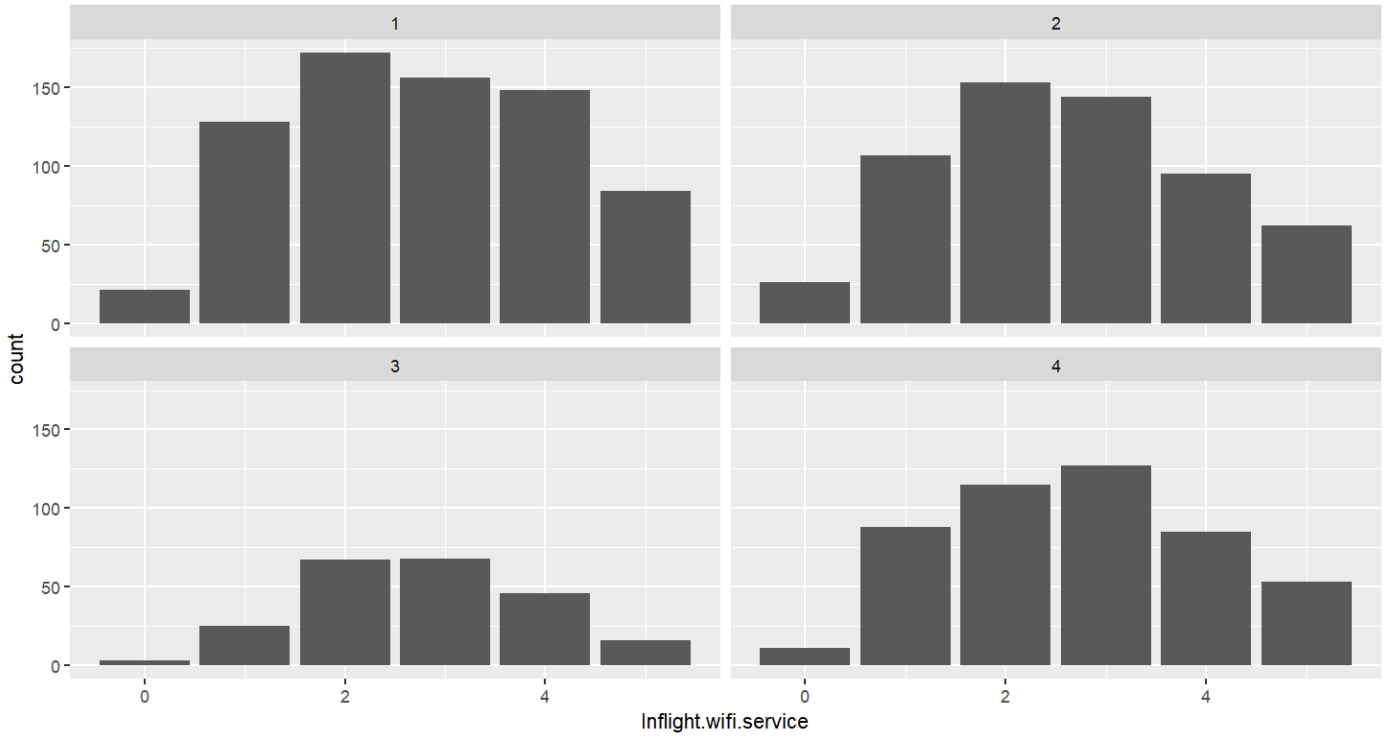
```
#看分群狀況
table(cluster)
```



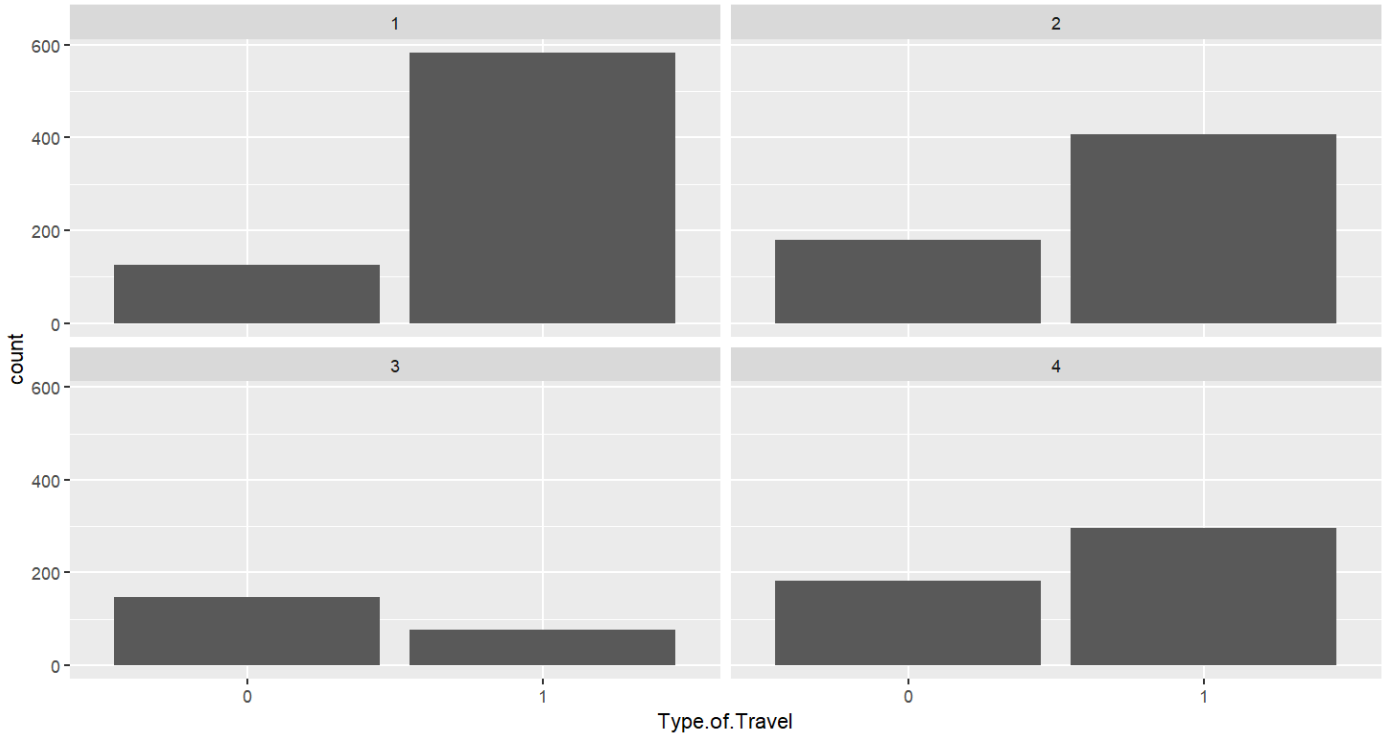

```
airline.2=cbind(airline.2,cluster)

# 依照前面找出的重要程度依序畫圖
# Oline.boarding
ggplot( data= airline.2) +
  geom_bar( aes( x = Online.boarding)) +
  facet_wrap( ~ cluster)
```



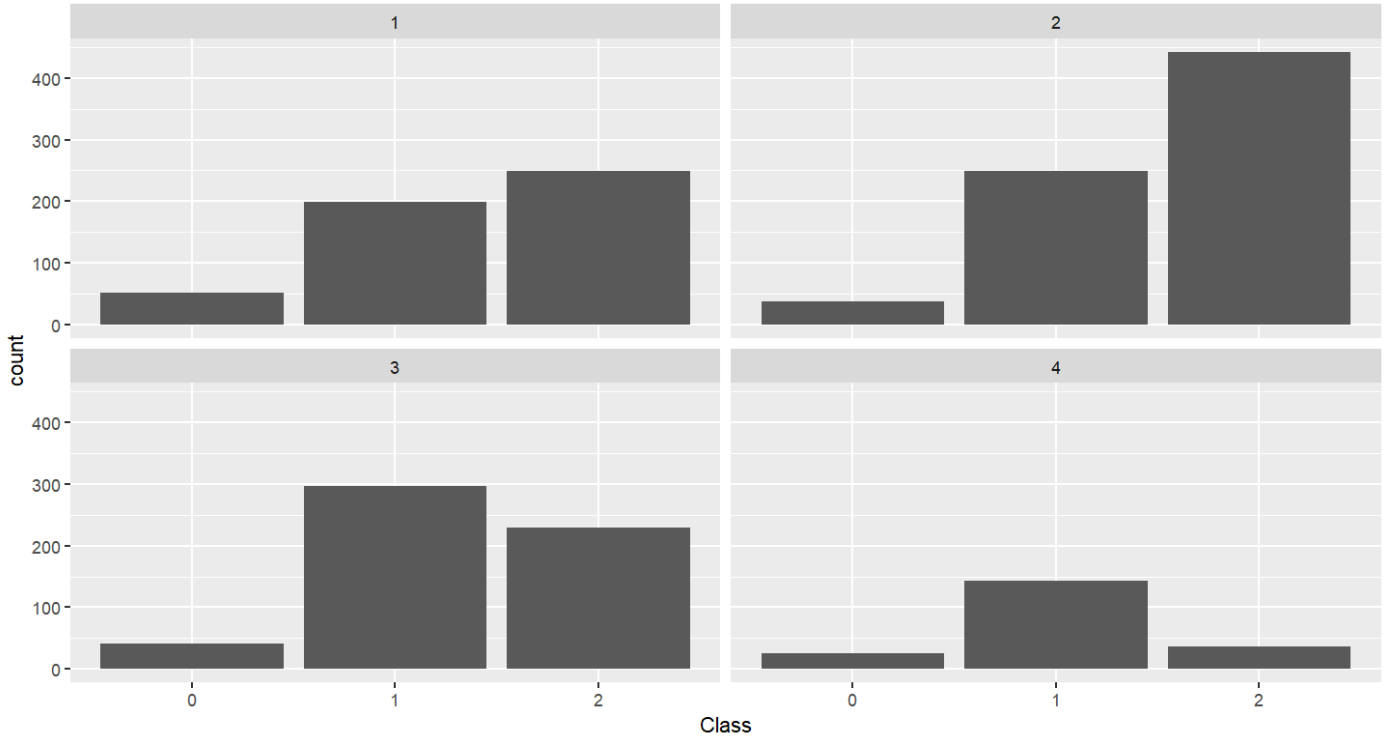
```
# Inflight.wifi.service
ggplot( data =airline.2) +
  geom_bar( aes( x = Inflight.wifi.service)) +
  facet_wrap( ~ cluster)
```



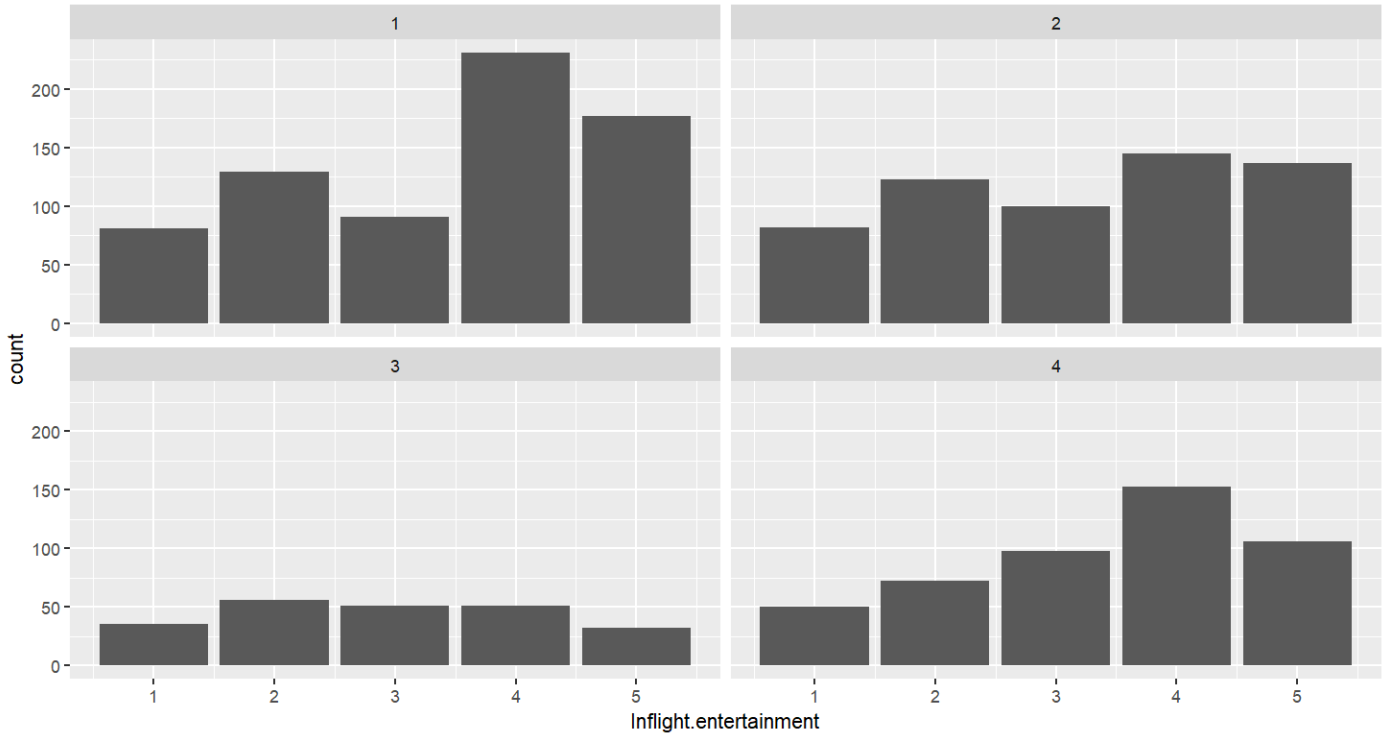

```
# Type.of.Travel
ggplot( data =airline.2) +
  geom_bar( aes( x = Type.of.Travel)) +
  facet_wrap( ~ cluster)
```



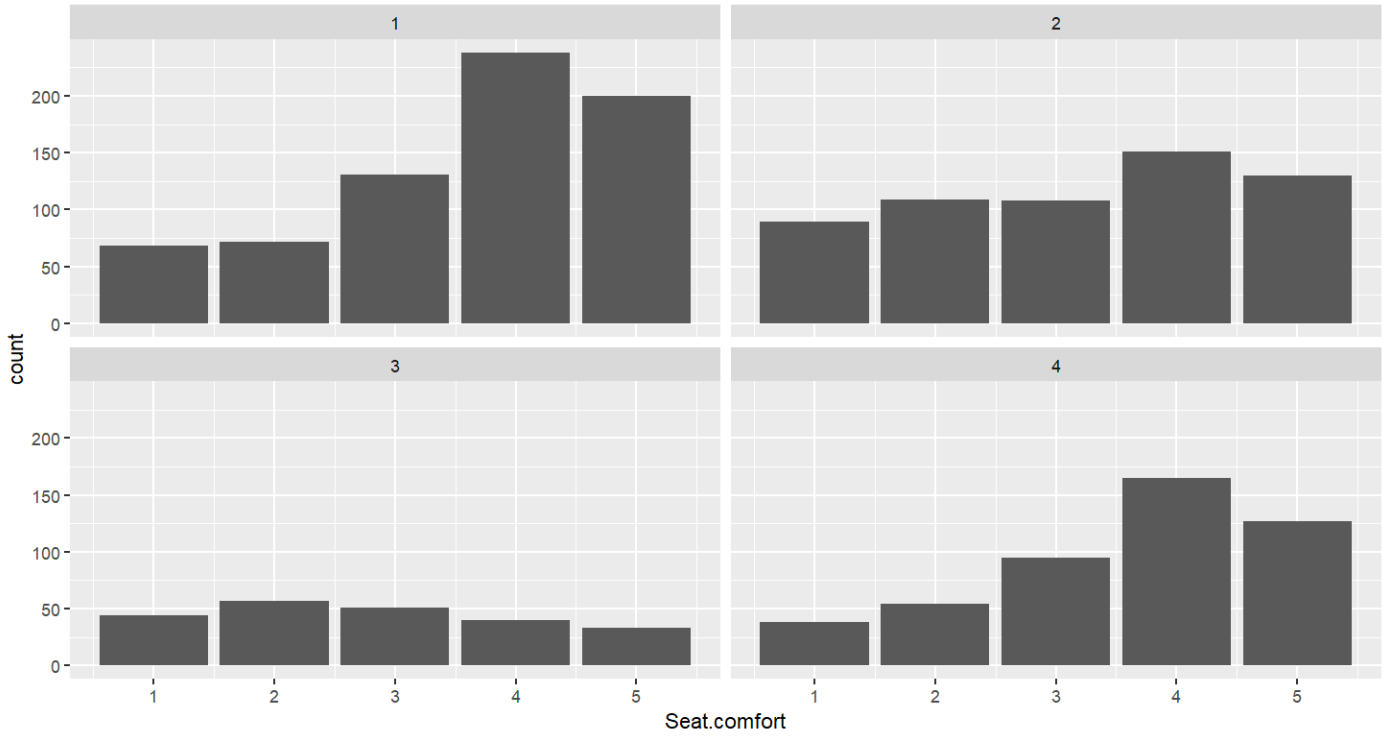
```
# 再多畫幾組
# Class
ggplot( data =airline.2) +
  geom_bar( aes( x = Class)) +
  facet_wrap( ~ cluster)
```



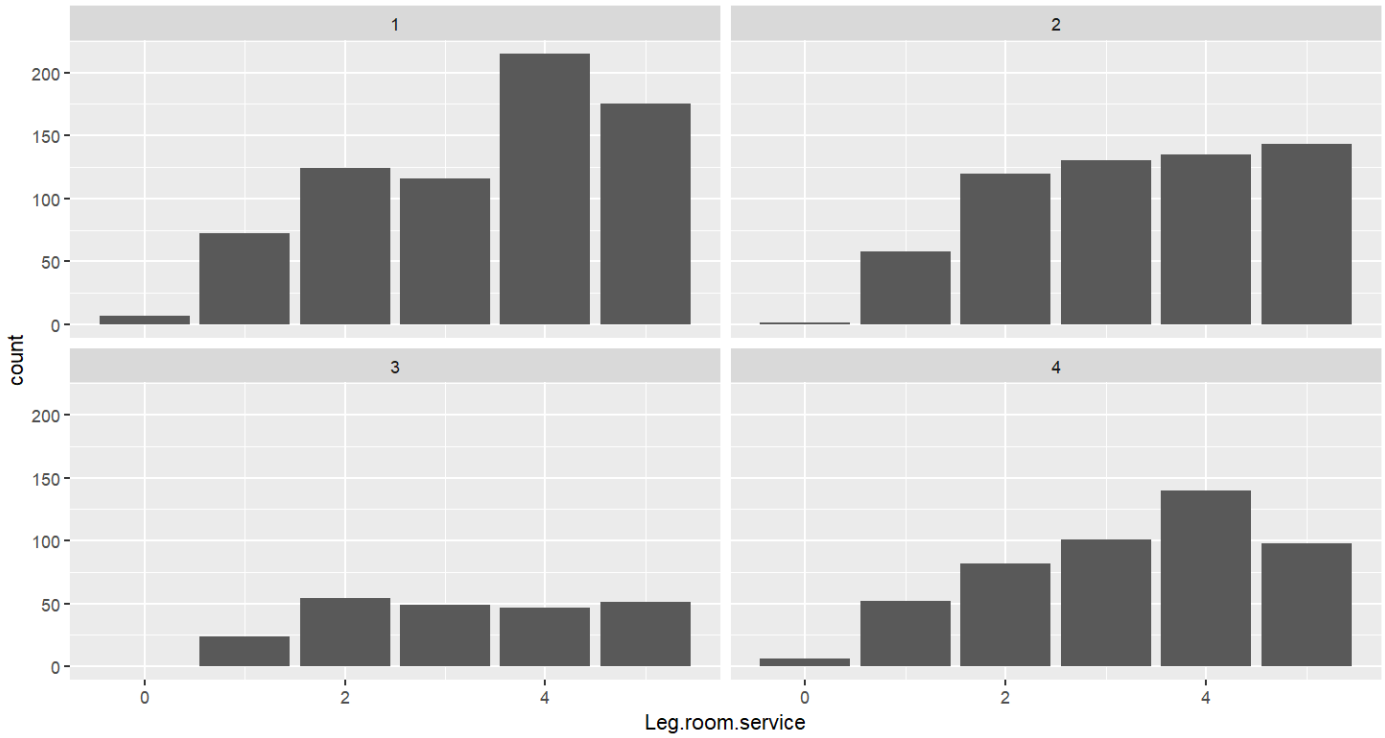

```
# Inflight.entertainment
ggplot( data =airline.2) +
  geom_bar( aes( x = Inflight.entertainment)) +
  facet_wrap( ~ cluster)
```



```
# Seat.comfort
ggplot( data =airline.2) +
  geom_bar( aes( x = Seat.comfort)) +
  facet_wrap( ~ cluster)
```




```
# Leg.room.service
ggplot( data =airline.2) +
  geom_bar( aes( x = Leg.room.service)) +
  facet_wrap( ~ cluster)
```



1. Oline.boarding：第一、第四群客戶較在乎
2. Inflight.wifi.service：第四群客戶較在乎網路速度、第一二群需求量大
3. Type.of.Travel：第一、二、四群客戶皆為商務旅行，其中第一群最為明顯
4. Class：第二群較多商務艙、第三群商務經濟艙都有但仍以 Eco 為主
5. Inflight.entertainment：第一、四群非常在乎娛樂設施
6. Seat.comfort：第一、二、四群都很在乎坐位，其中第一群最為明顯
7. Leg.room.service：第一、二、四群都很在乎腳的空間，其中第一群最為明顯

   故可從第一和四群著手，兩者皆為商旅人士，明顯較在乎 Oline.boarding 和 Inflight.entertainment、Seat.comfort、Leg.room.service

   第二群消費者和一四群相似，但較不在乎 Oline.boarding

   此外也可以從網速著手，可以稍微調高一二群的網路費率以賺取更多營收（其需求量大，故需求彈性小，即便調高價格，可能還是會購買此項服務），並提供第四群更快的網速以提高購買意願

   最後從 Class 可以看出，Eco Plus 很少人使用，使用者第一群較多一些，可以考慮取消 Eco Plus，再配合前述策略，就有機會將客戶導至 Business 以賺取更多營收（第一群的客戶較多商旅人士）