商業分析：SAS / R HW3

108208004 經濟三 白植允

1. 辨認出滿意與不滿意客戶 Predict passenger satisfaction.

- 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)。

```
data = read.csv("airline_survey.csv",sep = ",")
library(tidyverse)
#1
for(i in c(3,4,6,7,9:22,25)){
  data[,i] = as.factor(data[,i])
}
subdata <- data[1:1000,-c(1,2)]
str(subdata)
```

=>先將資料分類，並且只取前 1000 項(電腦容量的關係)

```
library(randomForest)
rf <- randomForest(satisfaction ~.,data = subdata, importance=T ,na.action = na.omit,nstart = 10)
```

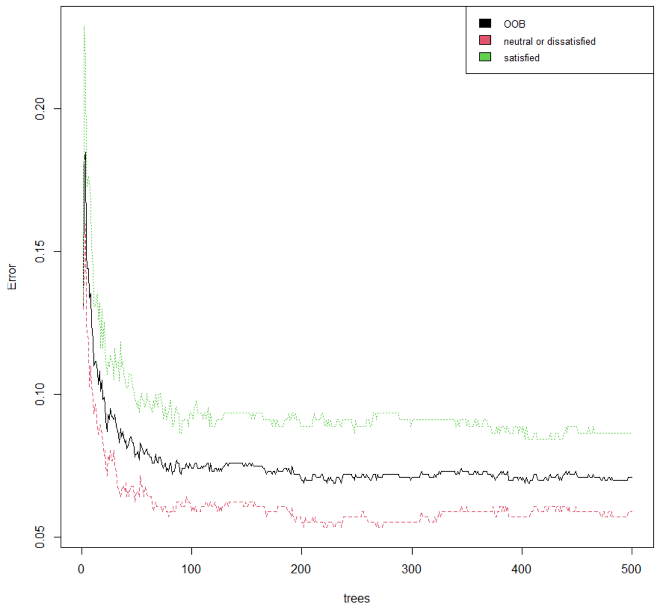
```
> rf

Call:
 randomForest(formula = satisfaction ~ ., data = subdata, importance = T,      nstart = 10, na.action = na.omi
t)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 7.11%
Confusion matrix:
                        neutral or dissatisfied satisfied class.error
neutral or dissatisfied                     527        33  0.05892857
satisfied                                    38       401  0.08656036
```

=>做 random forest，總共做了 500 棵樹且用了 4 個變數來分類，最後的袋外錯誤率為 7.11%。

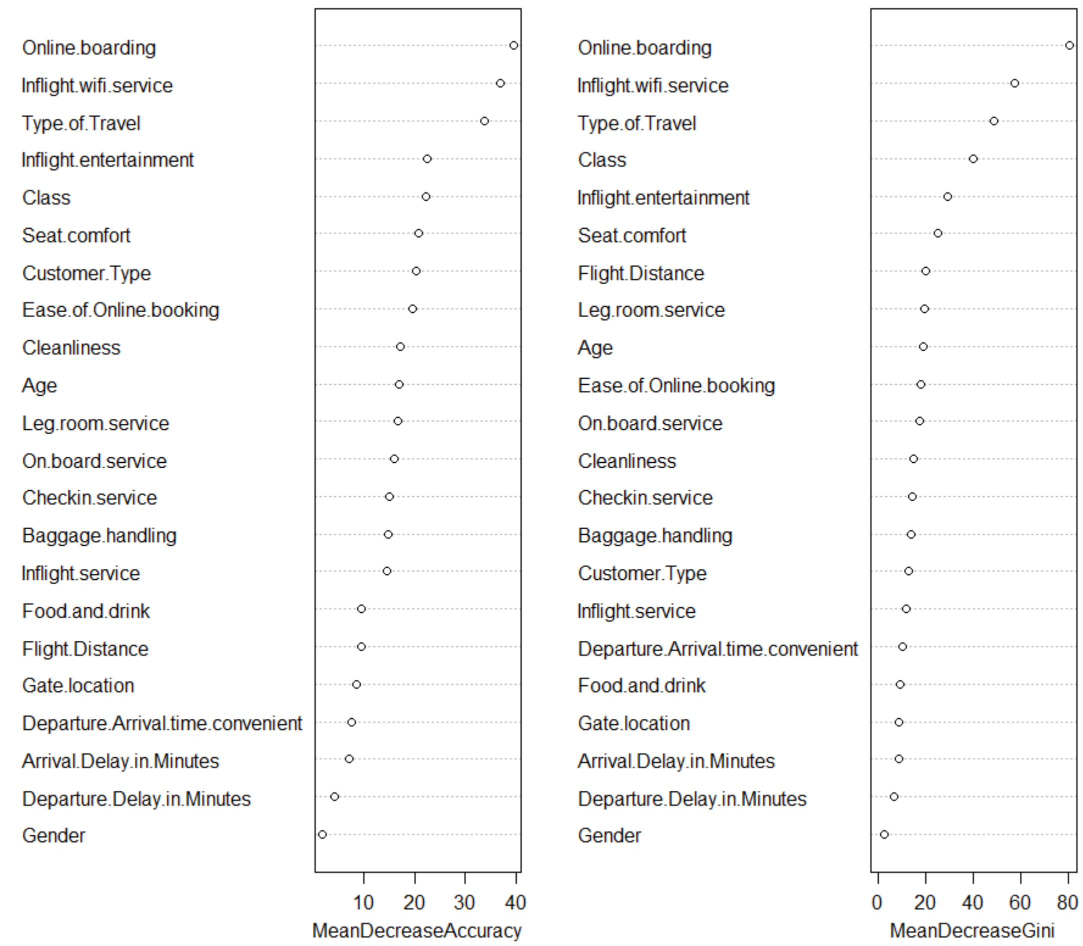
```
plot(rf)
legend("topright", colnames(rf$err.rate),col = 1:3,cex = 0.8,fill = 1:3)
```



=>大概在第 100 棵樹以後錯誤率漸趨平穩

● 找出重要變數：哪些因素影響客戶滿意度。

```
> importance(rf)
                                  neutral or dissatisfied satisfied MeanDecreaseAccuracy MeanDecreaseGini
Gender                                          1.1000439  1.567761             1.954023         2.396254
Customer.Type                                  17.7610405 16.614043            20.323942        12.990001
Age                                             9.0430893 14.566388            17.047628        18.800577
Type.of.Travel                                 25.6764342 31.588030            33.669714        48.741688
Class                                          13.4313411 20.838480            22.152695        40.182840
Flight.Distance                                 7.9989842  5.545109             9.575670        19.997291
Inflight.wifi.service                          35.9177243 25.240747            36.748901        57.739633
Departure.Arrival.time.convenient               0.5468710  9.252427             7.604997        10.334511
Ease.of.Online.booking                         18.7538510 10.333405            19.738883        17.880217
Gate.location                                  -0.8412016  9.563438             8.689976         8.975379
Food.and.drink                                  7.7682424  5.065233             9.700226         9.253744
Online.boarding                                30.7468798 31.823225            39.472367        80.778665
Seat.comfort                                   15.2028047 15.632874            20.831665        25.162443
Inflight.entertainment                         15.8415413 17.406707            22.620369        29.118717
On.board.service                               11.3879639 12.504548            16.116020        17.393201
Leg.room.service                                9.7534939 14.889182            16.856236        19.655422
Baggage.handling                                7.6043667 13.681667            14.792060        14.141483
Checkin.service                                11.9670907  9.746900            15.058235        14.583861
Inflight.service                                9.6089696 11.397098            14.682164        12.064247
Cleanliness                                    12.3537496 12.558016            17.225123        14.698143
Departure.Delay.in.Minutes                      3.1130324  3.050049             4.271428         6.938168
Arrival.Delay.in.Minutes                        5.4585921  4.465725             7.327547         8.853215
```



=>mean decrease accuracy: 將該變數變成隨機變數對預測準確性的降低程度

mean decrease Gini: 該變量對分類樹上觀測值的異質性的影響。

由這兩個指標可以發現，Online.boarding、Inflight.wifi.service、Type.of.travel、Class、Inflight.Entertainment 這五項對於顧客滿意度有顯著影響。

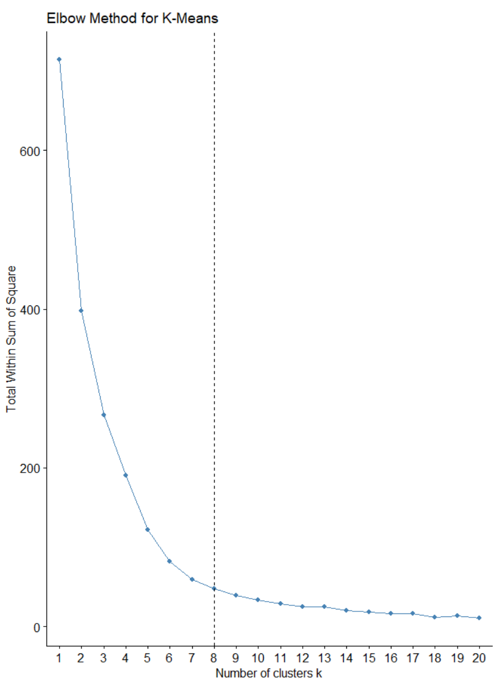
2. 描述客戶 Customer segmentation

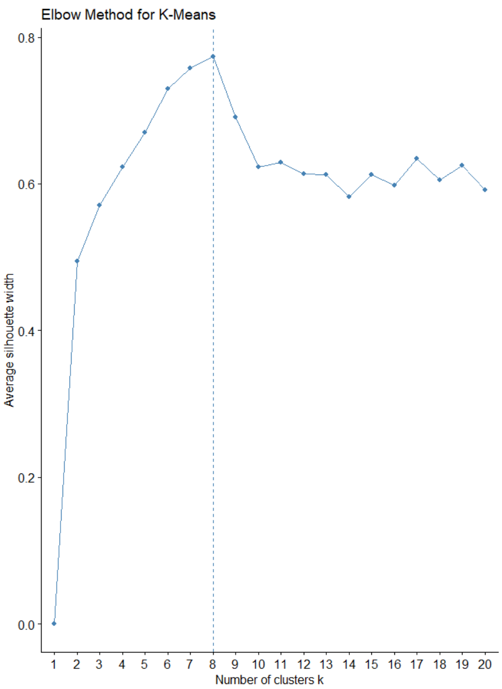
- 任選 1 種非監督式方法，將客戶分群，介紹你分出來的群，對於這些不同的客戶群集提出給該航空業的商業策略建議。

➢ 我想針對上述重要變數做分群，並探討群集特徵與滿意度的關係

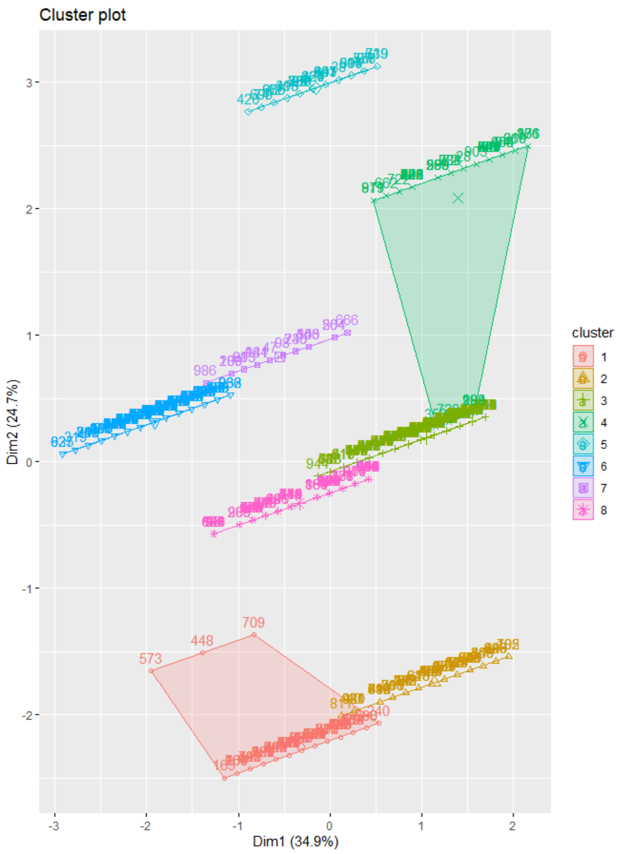
```
newdata <- subdata[,c(2,4,5,7,12,14,23)] %>%
  mutate(
    Customer.Type = ifelse(Customer.Type=="Loyal Customer",1,0),
    Type.of.Travel = ifelse(Type.of.Travel=="Personal Travel",1,0),
    Class = ifelse(Class =="Eco",0,ifelse(Class == "Eco Plus",1,2)),
    score = as.integer(Inflight.wifi.service)+as.integer(Online.boarding)+as.integer(Inflight.entertainment)-3,
    w_score = (score-min(score))/(max(score)-min(score))
    )
summary(newdata)
```

=>先將資料整理，類別資料轉換成 0,1,2，數值資料做(x-min)/range，其中 score 是指 inflight.wifi.service、online.boarding、inflight.entertainment 的分數加總，w_score 則是再對 score 做整理

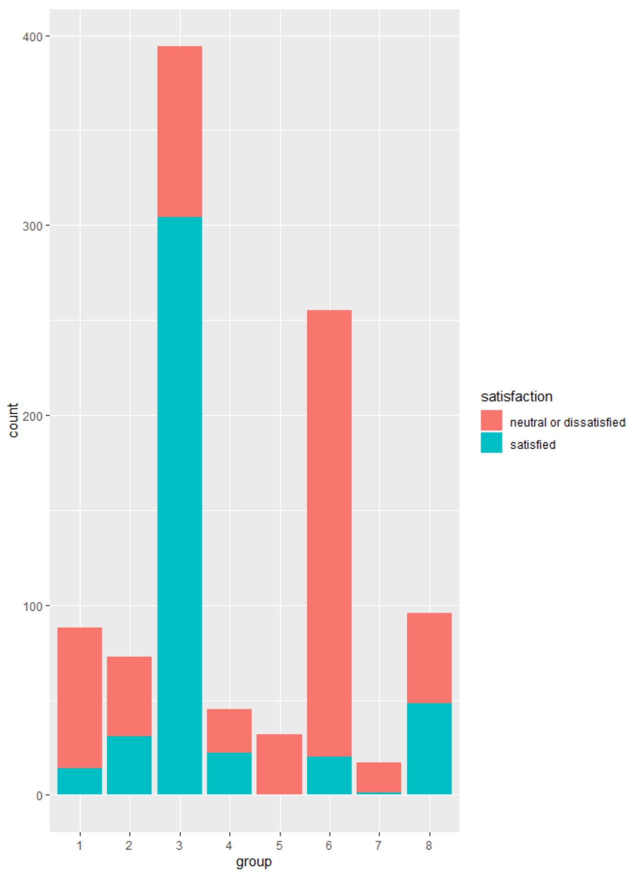




=>根據兩種不同的 elbow method(wss,silhouette)，決定將資料分成八組



=>八組的分類還蠻清楚的，其中第六組的分群範圍較大，散佈在外圍的點可以當作 Outlier，以探討主要分布那群為主

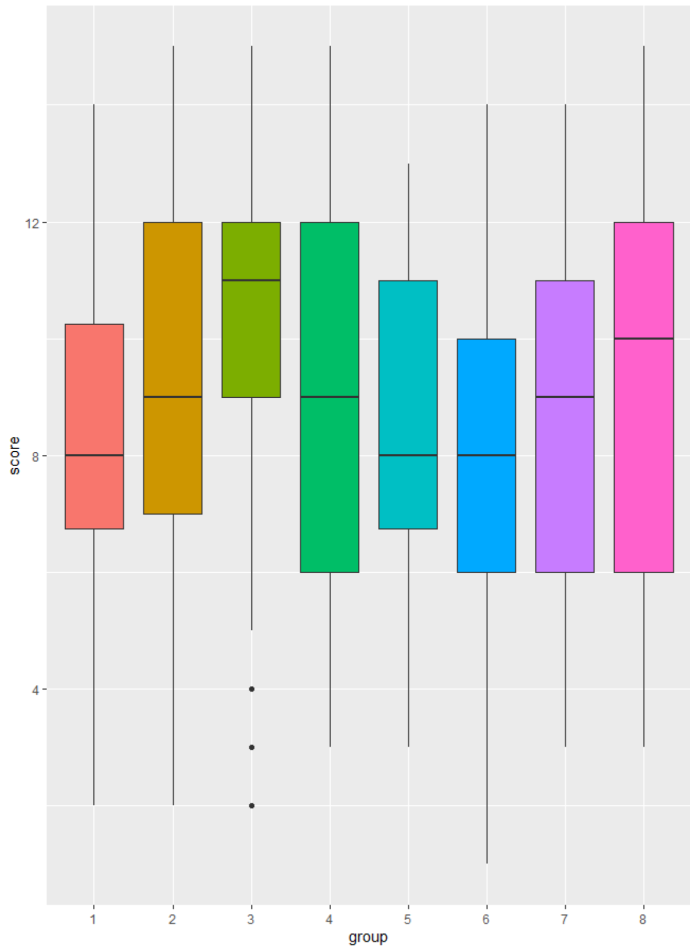


=>圖中可看出第 1,2,5,6,7 組的滿意度相當低(都低於 50%)

```
> k$centers
  Customer.Type Type.of.Travel Eco_Class Eco_Plus   w_score
1     0.0000000     0.03409091         1        0 0.5162338
2     0.0000000     0.00000000         0        0 0.5880626
3     1.0000000     0.00000000         0        0 0.6758521
4     0.8888889     0.00000000         0        1 0.5968254
5     1.0000000     1.00000000         0        1 0.5156250
6     1.0000000     1.00000000         1        0 0.5078431
7     1.0000000     1.00000000         0        0 0.5546218
8     1.0000000     0.00000000         1        0 0.6183036
```

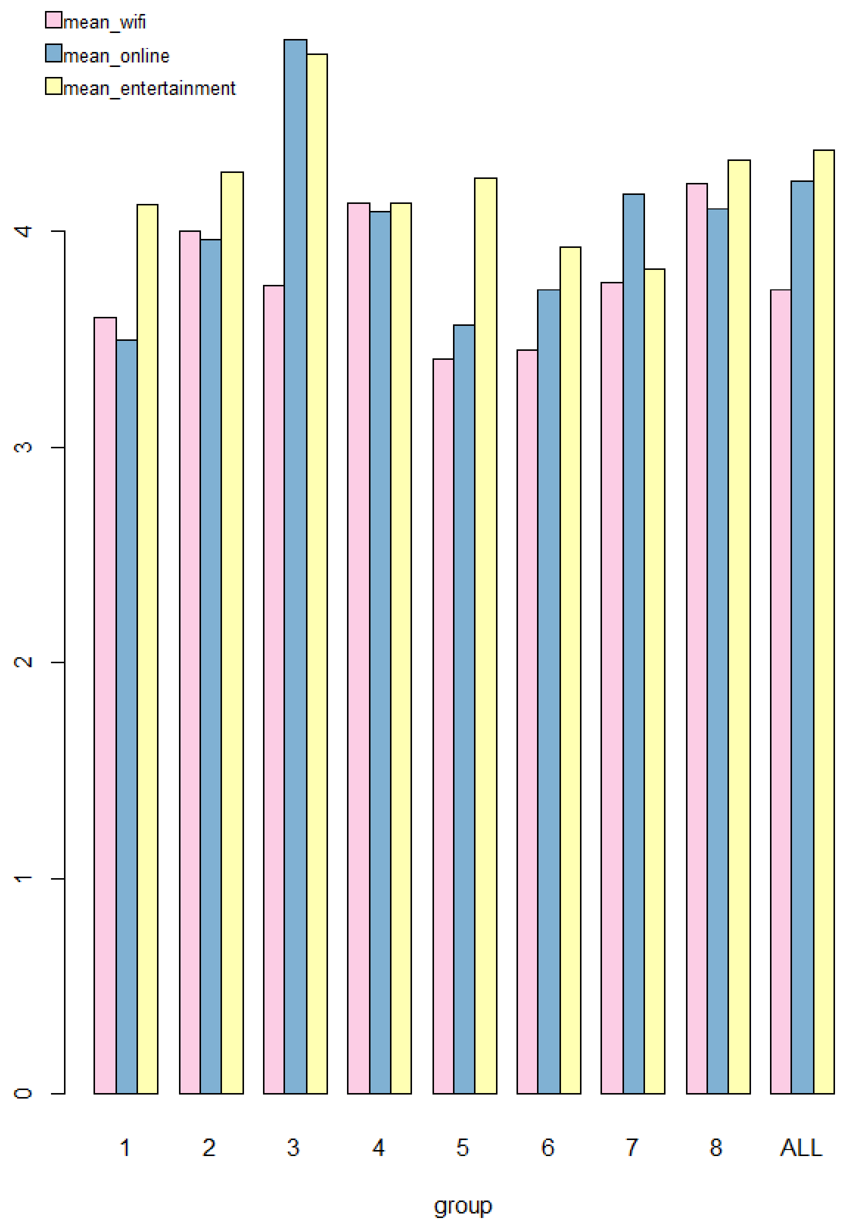
```
> mean(newdata$w_score)
[1] 0.5962857
```

=>第 2 組客人的這三項分數並沒有與平均差距很大，也許是其他因素影響滿意度，需要透過其他變數探討滿意度較低的原因。



=>綜合第 5,6,7 組客人可看出不管是乘坐哪種艙別，Loyal Customer 對於 Personal Travel 的服務都不是很滿意。

=>策略 1:公司針對 Loyal Customer 祭出優惠，例如與旅遊業者合作，只要是 Loyal Customer 參加旅遊行程能享有機票折扣，或是在飛機上享有專屬服務(更多樣的電影可以看或是能免費使用 WIFI)



=>大部分客群都對於 Online.Boarding 不滿意，公司應盡速完善網路預辦登機的服務。

=>第 1 組客人可看成 disloyal Customer、Business Travel、Eco Class，應該就是一般出差的旅客，我認為他們的行程比較緊湊且較常搭乘飛機，因此除了改善 Online.Boarding 的服務外，也許公司可以提出積點服務，像是搭乘多少次 Eco class 的班機能讓他們晉升商務艙，或是針對這類客人辦理快速通關的服務。

附錄:R 程式碼

```
data = read.csv("airline_survey.csv",sep = ",")
library(tidyverse)
#1
for(i in c(3,4,6,7,9:22,25)){
```

```
  data[,i] = as.factor(data[,i])
}
subdata <- data[1:1000,-c(1,2)]
str(subdata)

library(randomForest)
rf <- randomForest(satisfaction ~.,data = subdata,
importance=T ,na.action = na.omit,nstart = 10)
rf

plot(rf)
legend("topright", colnames(rf$err.rate),col = 1:3,cex = 0.8,fill =
1:3)

importance(rf)
varImpPlot(rf)

#2

newdata <- subdata[,c(2,4,5,7,12,14,23)] %>%
  mutate(
    Customer.Type = ifelse(Customer.Type=="Loyal Customer",1,0),
    Type.of.Travel = ifelse(Type.of.Travel=="Personal Travel",1,0),
    Eco_Class = ifelse(Class =="Eco",1,0),
    Eco_Plus = ifelse(Class == "Eco Plus",1,0),
    score =
as.integer(Inflight.wifi.service)+as.integer(Online.boarding)+as.inte
ger(Inflight.entertainment)-3,
    w_score = (score-min(score))/(max(score)-min(score))
    )
summary(newdata)

library(factoextra)
fviz_nbclust(newdata[,c(1,2,8,9,11)],
             FUNcluster = kmeans, #k-Means
             nstart =20,
             method ="wss",  #total within sum of square
             k.max = 20 #max number of clusters to consider
```

```
          )+
  labs(title = "Elbow Method for K-Means")+
  geom_vline(xintercept = 8,linetype =2)

fviz_nbclust(newdata[,c(1,2,8,9,11)],
             FUNcluster = kmeans, #k-Means
             nstart =20,
             method ="silhouette",  #total within sum of square
             k.max = 20 #max number of clusters to consider
)+
  labs(title = "Elbow Method for K-Means")

k = kmeans(newdata[,c(1,2,8,9,11)],centers = 8, nstart = 20)
k$cluster
fviz_cluster(k,data = newdata[,c(1,2,8,9,11)])
newdata$group = as.factor(k$cluster)
str(newdata)

ggplot(newdata,aes(group,fill = satisfaction))+
  geom_bar()

k$centers
mean(newdata$w_score)



ggplot(newdata,aes(x=group,y=score,fill = group)) +
geom_boxplot()+theme(legend.position = "none")

group <- newdata %>%
  group_by(group) %>%
  summarise(mean_wifi = mean(as.integer(Inflight.wifi.service)),
            mean_online = mean(as.integer(Online.boarding)),
            mean_entertainment =
mean(as.integer(Inflight.entertainment))) %>%
  mutate(group = as.character(group))

summarise1 <- newdata %>%
  summarise(mean_wifi = mean(as.integer(Inflight.wifi.service)),
```

```
          mean_online = mean(as.integer(Online.boarding)),
          mean_entertainment =
mean(as.integer(Inflight.entertainment)))
group[nrow(group)+1,] <-
list("ALL",summarise1$mean_wifi,summarise1$mean_online,summarise1$mea
n_entertainment)

library(RColorBrewer)
display.brewer.all()
rcols <- sample(brewer.pal(12,name = "Set3"),3)
barplot(data =
group,cbind(mean_wifi,mean_online,mean_entertainment)~group,beside
=T,col = rcols)
legend("topleft",inset = c(-0.05,-0.05), legend =
c("mean_wifi","mean_online","mean_entertainment"),fill =
rcols,cex=0.8,bty = "n",xpd=T,x.intersp = 0.1)
```