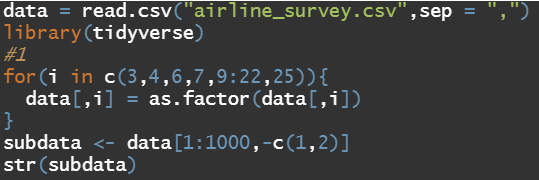
商業分析：SAS / R HW3

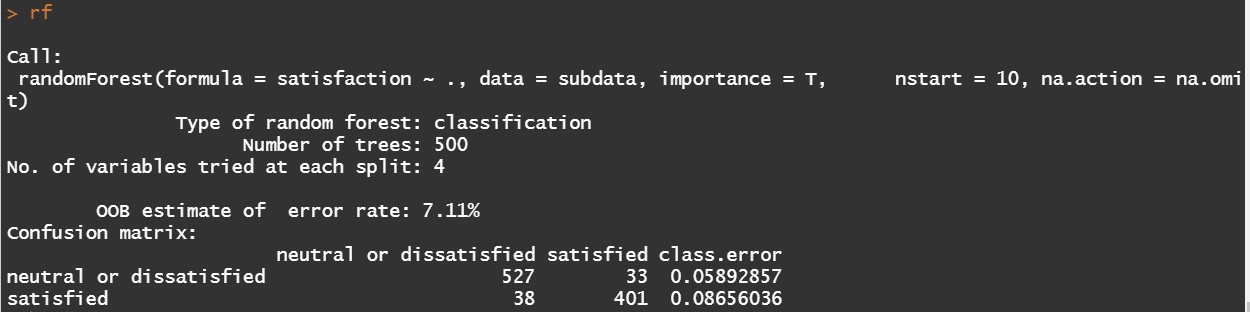
108208004 經濟三 白植允

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

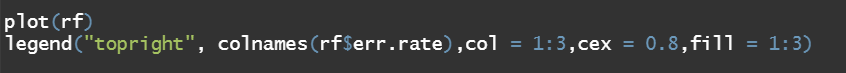
* 任選 1 種監督式學習方法配適模型，預測滿意度 satisfaction (2 類：滿意、中立或不滿意)。

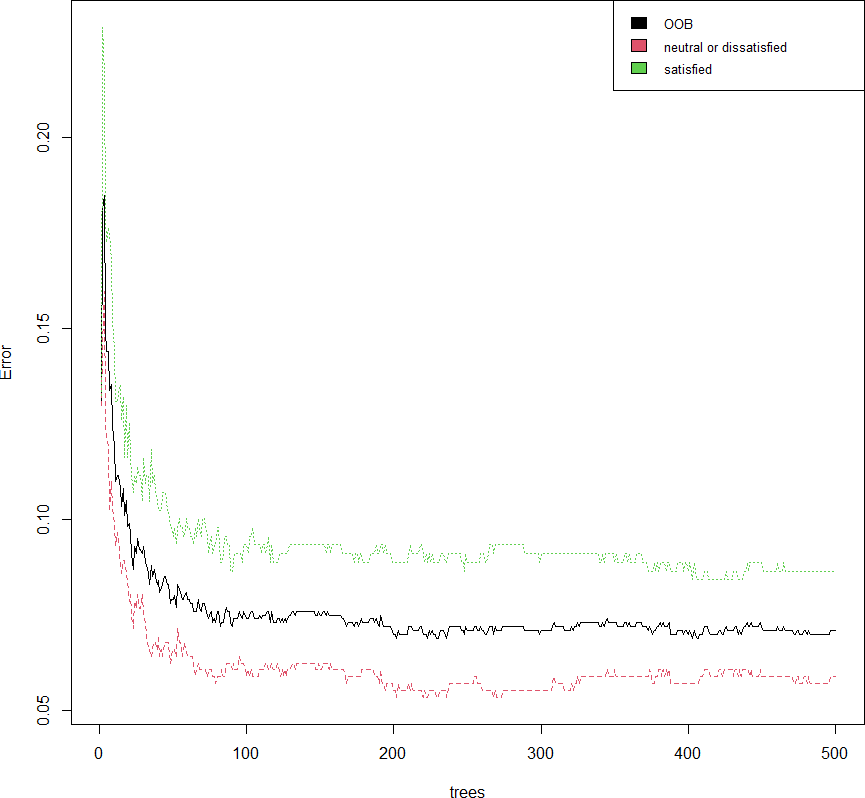


=>先將資料分類，並且只取前 1000 項(電腦容量的關係)



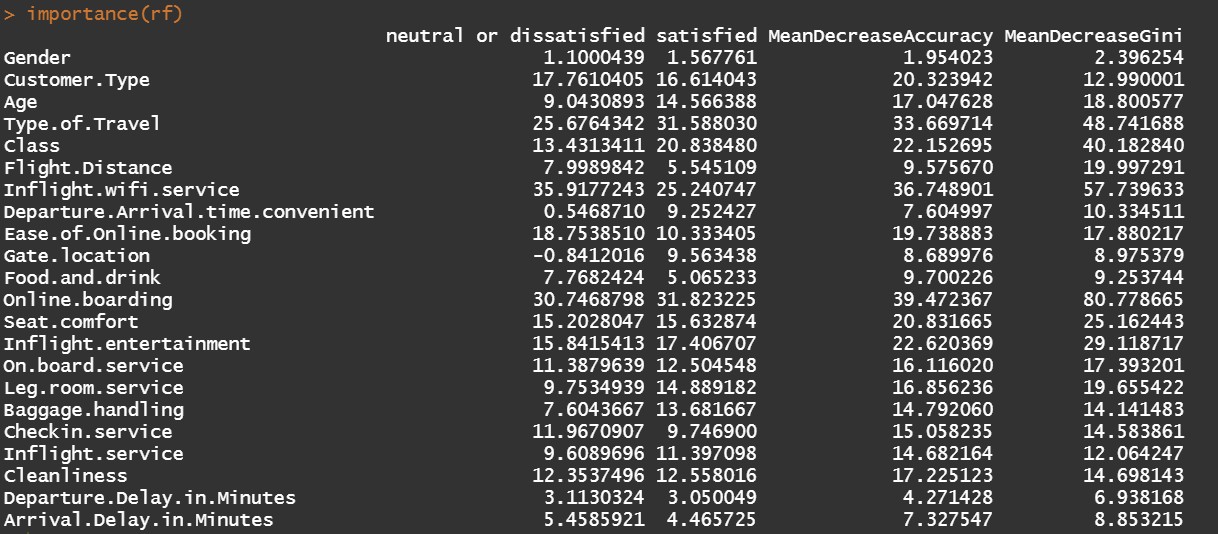
=>做 random forest，總共做了 500 棵樹且用了 4 個變數來分類，最後的袋外錯誤率為 7.11%。

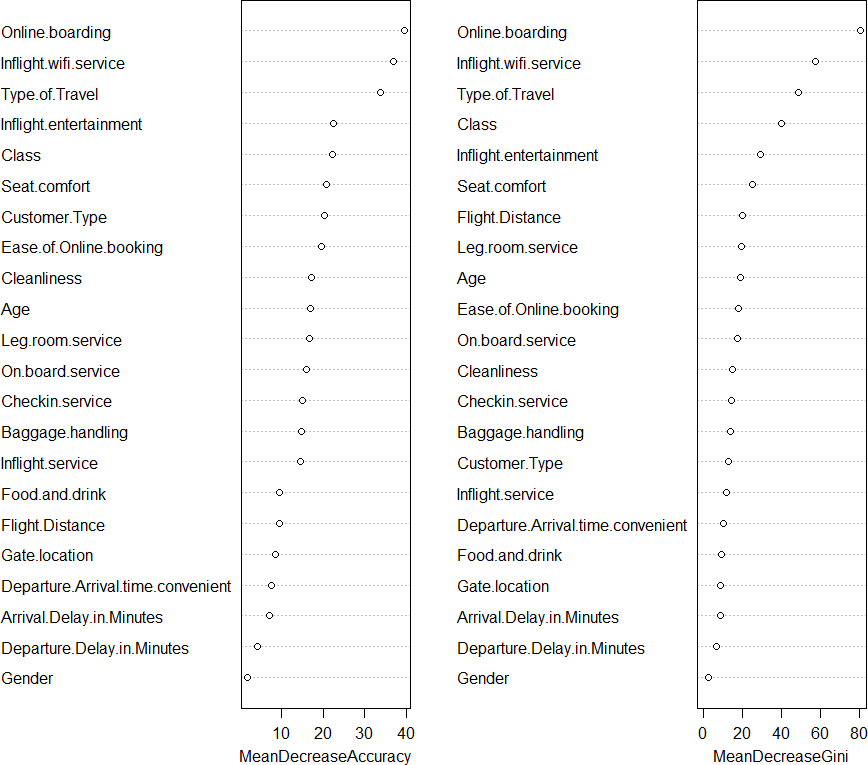




=>大概在第 100 棵樹以後錯誤率漸趨平穩

* 找出重要變數：哪些因素影響客戶滿意度。



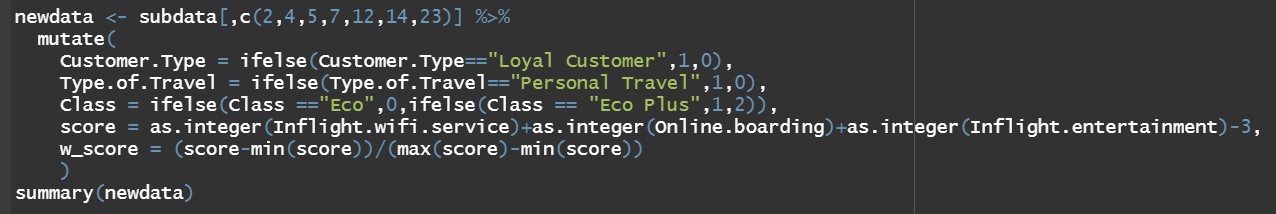


=>mean decrease accuracy: 將該變數變成隨機變數對預測準確性的降低程度

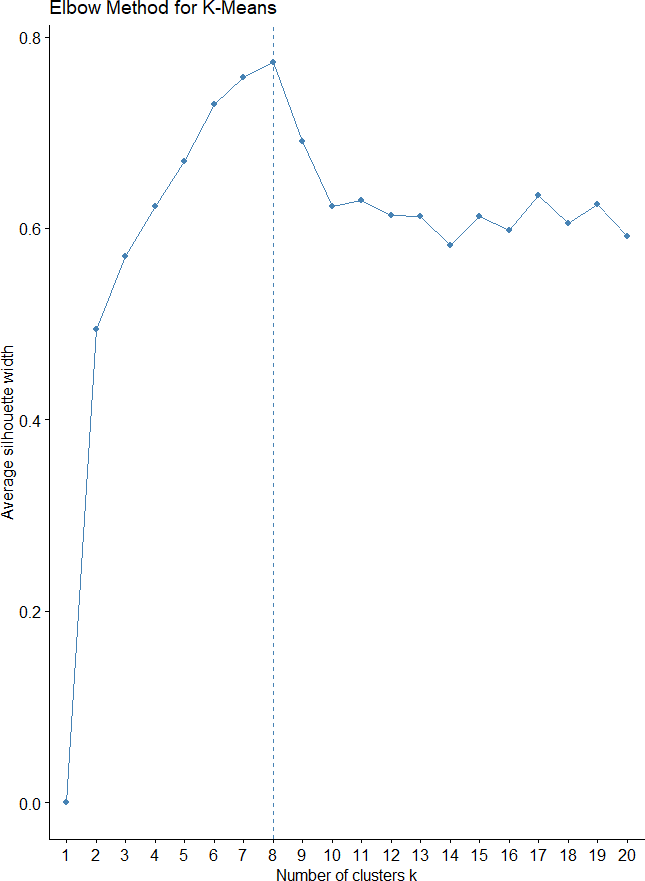
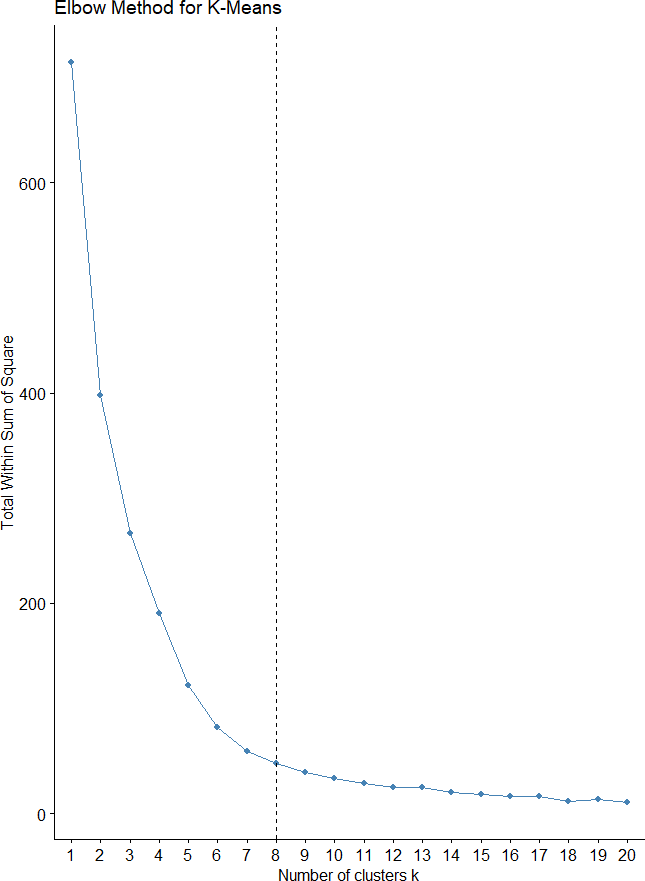
mean decrease Gini: 該變量對分類樹上觀測值的異質性的影響。

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。

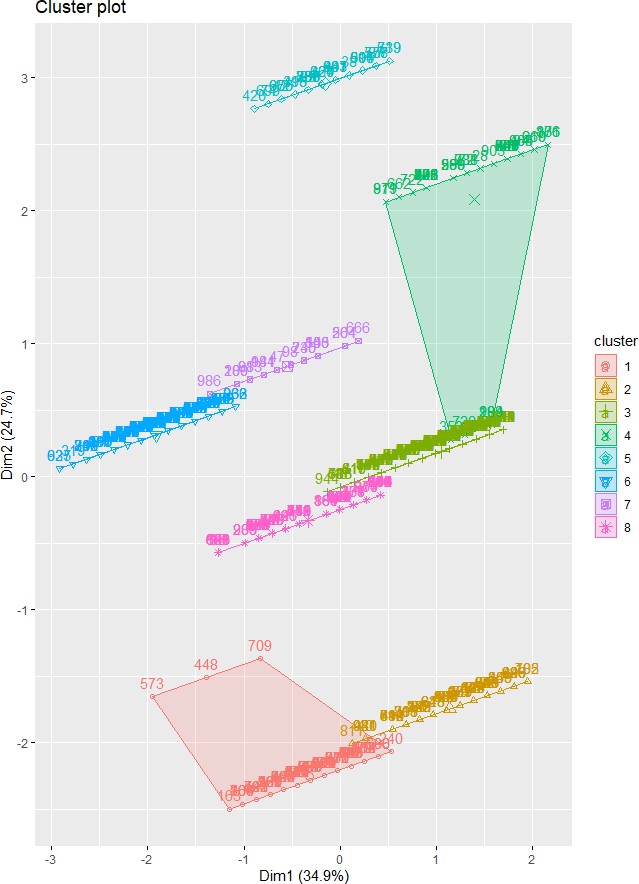
1. 描述客戶 Customer segmentation

* 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

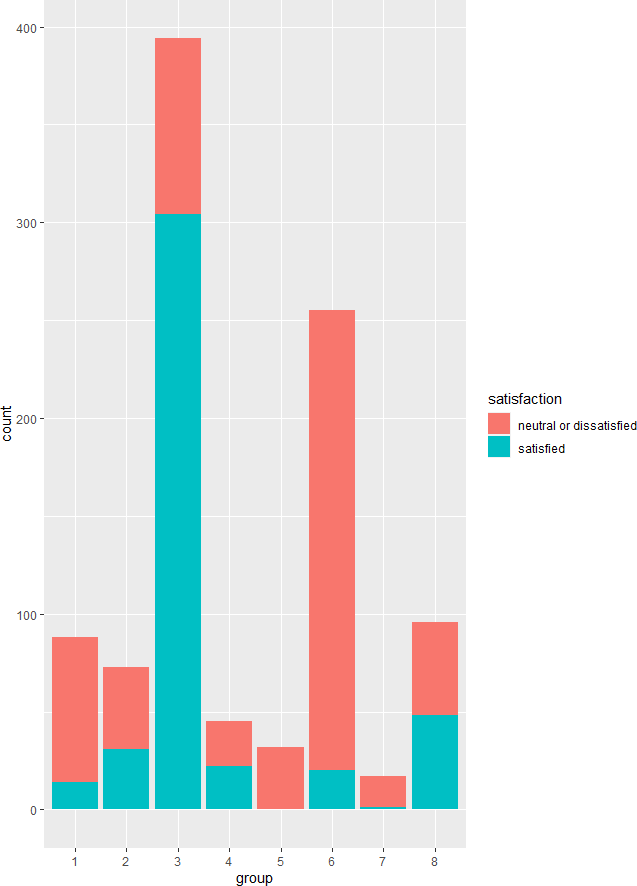
=>先將資料整理，類別資料轉換成 0,1,2，數值資料做(x-min)/range，其中 score 是指 inflight.wifi.service、online.boarding、inflight.entertainment 的分數加總， w\_score 則是再對 score 做整理



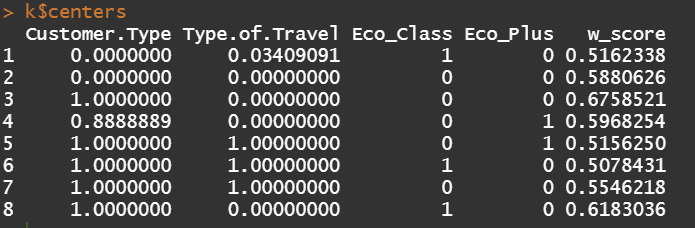
=>根據兩種不同的 elbow method(wss,silhouette)，決定將資料分成八組



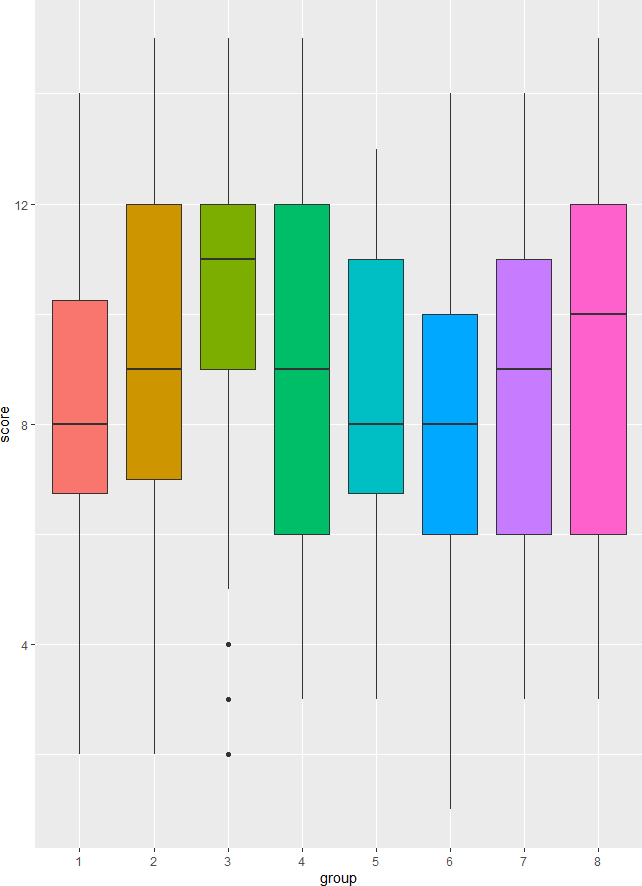
=>八組的分類還蠻清楚的，其中第六組的分群範圍較大，散佈在外圍的點可以當作 Outlier，以探討主要分布那群為主



=>圖中可看出第 1,2,5,6,7 組的滿意度相當低(都低於 50%)



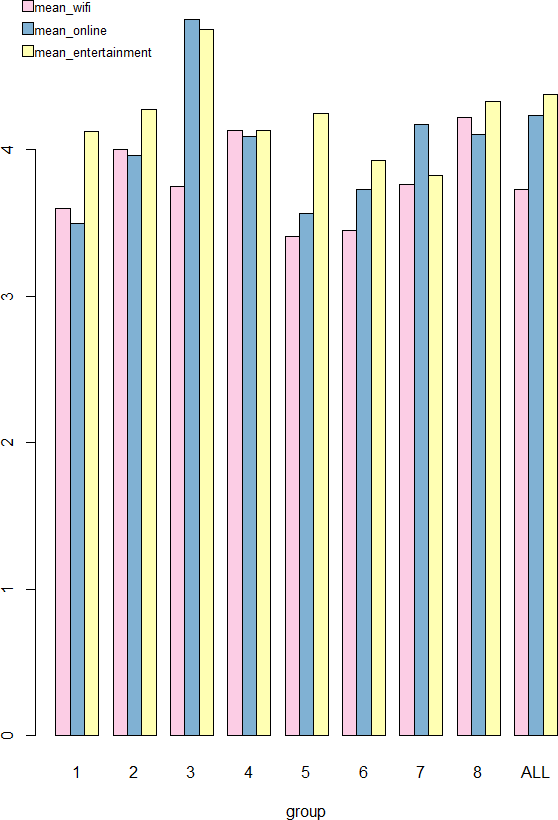
=>第 2 組客人的這三項分數並沒有與平均差距很大，也許是其他因素影響滿意度，需要透過其他變數探討滿意度較低的原因。



=>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)