SAS/R商業資料分析作業三

107508006 歐西四 陳葳芃

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)。Choose one supervised method to predict passenger satisfaction.

Ans:

```
> setwd("~/Downloads/1102 R/HW/hw 3")
> library(readr)
> airline <- read_csv("airline_survey.csv")
Rows: 103904 Columns: 25
 0s── Column specification ────────────────────────────────────────────────────────────────
Delimiter: ","
chr  (5): Gender, Customer Type, Type of Travel, Class, satisfaction
dbl (20): Index, id, Age, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, Ease of...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> airline$Index<-c(1:103904)
> colnames(airline) <- gsub("/", "__", colnames(airline))
> colnames(airline) <- gsub("-", "", colnames(airline))
> colnames(airline) <- gsub(" ", "_", colnames(airline))
> #ls(airline)
> airline<-na.omit(airline)
> 
```

```
> str(airline)
tibble [103,594 × 25] (S3: tbl_df/tbl/data.frame)
 $ Index                            : int [1:103594] 1 2 3 4 5 6 7 8 9 10 ...
 $ id                               : num [1:103594] 70172 5047 110028 24026 119299 ...
 $ Gender                           : chr [1:103594] "Male" "Male" "Female" "Female" ...
 $ Customer_Type                    : chr [1:103594] "Loyal Customer" "disloyal Customer" "Loyal Customer" "Loyal Cust
omer" ...
 $ Age                              : num [1:103594] 13 25 26 25 61 26 47 52 41 20 ...
 $ Type_of_Travel                   : chr [1:103594] "Personal Travel" "Business travel" "Business travel" "Business t
ravel" ...
 $ Class                            : chr [1:103594] "Eco Plus" "Business" "Business" "Business" ...
 $ Flight_Distance                  : num [1:103594] 460 235 1142 562 214 ...
 $ Inflight_wifi_service            : num [1:103594] 3 3 2 2 3 3 2 4 1 3 ...
 $ Departure__Arrival_time_convenient: num [1:103594] 4 2 2 5 3 4 4 3 2 3 ...
 $ Ease_of_Online_booking           : num [1:103594] 3 3 2 5 3 2 2 4 2 3 ...
 $ Gate_location                    : num [1:103594] 1 3 2 5 3 1 3 4 2 4 ...
 $ Food_and_drink                   : num [1:103594] 5 1 5 2 4 1 2 5 4 2 ...
 $ Online_boarding                  : num [1:103594] 3 3 5 2 5 2 2 5 3 3 ...
 $ Seat_comfort                     : num [1:103594] 5 1 5 2 5 1 2 5 3 3 ...
 $ Inflight_entertainment           : num [1:103594] 5 1 5 2 3 1 2 5 1 2 ...
 $ Onboard_service                  : num [1:103594] 4 1 4 2 3 3 3 5 1 2 ...
 $ Leg_room_service                 : num [1:103594] 3 5 3 5 4 4 3 5 2 3 ...
 $ Baggage_handling                 : num [1:103594] 4 3 4 3 4 4 4 5 1 4 ...
 $ Checkin_service                  : num [1:103594] 4 1 4 1 3 4 3 4 4 4 ...
 $ Inflight_service                 : num [1:103594] 5 4 4 4 3 4 5 5 1 3 ...
 $ Cleanliness                      : num [1:103594] 5 1 5 2 3 1 2 4 2 2 ...
 $ Departure_Delay_in_Minutes       : num [1:103594] 25 1 0 11 0 0 9 4 0 0 ...
 $ Arrival_Delay_in_Minutes         : num [1:103594] 18 6 0 9 0 0 23 0 0 0 ...
 $ satisfaction                     : chr [1:103594] "neutral or dissatisfied" "neutral or dissatisfied" "satisfied"
 "neutral or dissatisfied" ...
 - attr(*, "na.action")= 'omit' Named int [1:310] 214 1125 1530 2005 2109 2486 2631 3622 4042 4491 ...
  ..- attr(*, "names")= chr [1:310] "214" "1125" "1530" "2005" ...
> 
```

```
> summary(airline)
     Index              id             Gender          Customer_Type           Age        Type_of_Travel    
 Min.   :     1   Min.   :     1   Length:103594      Length:103594      Min.   : 7.00   Length:103594     
 1st Qu.: 25961   1st Qu.: 32562   Class :character   Class :character   1st Qu.:27.00   Class :character  
 Median : 51956   Median : 64890   Mode  :character   Mode  :character   Median :40.00   Mode  :character  
 Mean   : 51951   Mean   : 64942                                         Mean   :39.38                     
 3rd Qu.: 77926   3rd Qu.: 97370                                         3rd Qu.:51.00                     
 Max.   :103904   Max.   :129880                                         Max.   :85.00                     
    Class           Flight_Distance Inflight_wifi_service Departure__Arrival_time_convenient
 Length:103594      Min.   :  31    Min.   :0.00          Min.   :0.00                      
 Class :character   1st Qu.: 414    1st Qu.:2.00          1st Qu.:2.00                      
 Mode  :character   Median : 842    Median :3.00          Median :3.00                      
                    Mean   :1189    Mean   :2.73          Mean   :3.06                      
                    3rd Qu.:1743    3rd Qu.:4.00          3rd Qu.:4.00                      
                    Max.   :4983    Max.   :5.00          Max.   :5.00                      
 Ease_of_Online_booking Gate_location   Food_and_drink  Online_boarding  Seat_comfort  Inflight_entertainment
 Min.   :0.000          Min.   :0.000   Min.   :0.000   Min.   :0.00    Min.   :0.00   Min.   :0.000         
 1st Qu.:2.000          1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.00   1st Qu.:2.000         
 Median :3.000          Median :3.000   Median :3.000   Median :3.00    Median :4.00   Median :4.000         
 Mean   :2.757          Mean   :2.977   Mean   :3.202   Mean   :3.25    Mean   :3.44   Mean   :3.358         
 3rd Qu.:4.000          3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.00   3rd Qu.:4.000         
 Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.00    Max.   :5.00   Max.   :5.000         
 Onboard_service Leg_room_service Baggage_handling Checkin_service Inflight_service  Cleanliness   
 Min.   :0.000   Min.   :0.000    Min.   :1.000    Min.   :0.000   Min.   :0.000    Min.   :0.000  
 1st Qu.:2.000   1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.000    1st Qu.:2.000  
 Median :4.000   Median :4.000    Median :4.000    Median :3.000   Median :4.000    Median :3.000  
 Mean   :3.383   Mean   :3.351    Mean   :3.632    Mean   :3.304   Mean   :3.641    Mean   :3.286  
 3rd Qu.:4.000   3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:5.000    3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000    Max.   :5.000    Max.   :5.000   Max.   :5.000    Max.   :5.000  
 Departure_Delay_in_Minutes Arrival_Delay_in_Minutes satisfaction      
 Min.   :   0.00            Min.   :   0.00          Length:103594     
 1st Qu.:   0.00            1st Qu.:   0.00          Class :character  
 Median :   0.00            Median :   0.00          Mode  :character  
 Mean   :  14.75            Mean   :  15.18                            
 3rd Qu.:  12.00            3rd Qu.:  13.00                            
 Max.   :1592.00            Max.   :1584.00                            
> 
```

>>說明：圖1：資料預處理，將資料調整為我需要的格式（將空取代為”_”），並剔除掉na值；圖2、3：觀察資料結構並對類別資料加以轉換。

```
> ##將data更改為factor step1
> airline$satisfaction<-as.factor(airline$satisfaction)
> airline$Gender<-as.factor(airline$Gender)
> airline$Customer_Type<-as.factor(airline$Customer_Type)
> airline$Type_of_Travel<-as.factor(airline$Type_of_Travel)
> airline$Class<-as.factor(airline$Class)
> ##將data更改為factor step2
> airline$Gender <- ifelse(airline$Gender == "Male", 1, 0)
> airline$Customer_Type <- ifelse(airline$Customer_Type == "Loyal Customer", 1, 0)
> airline$Type_of_Travel <- ifelse(airline$Type_of_Travel== "Business travel", 1, 0)
> airline$satisfaction <- ifelse(airline$satisfaction == "satisfied", 1, 0)
> ## Use the dummy variable to predict factor variable
> library(fastDummies)
> d2<-dummy_cols(airline)
> #Delete original column and type_Eco
> library(dummies)
> ##dummy 00 = type_Eco
> d2<-d2[,c(-7,-27)]
> colnames(d2) <- gsub(" ", "_", colnames(d2))
> 
```

>>說明：圖4：資料預處理，將類別資料調整為dummy variables，接著刪去其中Class座艙的三種type中的其一: Eco，避免共線性。

```
> ##訓練及測試集
> library(tidyverse)
> train_df <- d2 %>% group_by(satisfaction) %>% sample_frac(0.7)
> test_df  <- anti_join(d2, train_df, by ='Index')
> train_rf <- train_df[,3:26]
> test_rf  <- test_df[,3:26]
> #str(train_rf)
> table(train_df$satisfaction)

    0     1 
41088 31428 
> 
```

>>說明：利用tidyverse套件將資料分割為train and test data，最後移除對分析較無意義的Index以及用戶id兩個變數，重新將train and test data命名為train_rf及test_rf。

```
> ##########logistic regression配適模型
> logit_model <- glm(satisfaction ~ .,
+                    train_rf, family=binomial(link="logit"))
> summary(logit_model)

Call:
glm(formula = satisfaction ~ ., family = binomial(link = "logit"), 
    data = train_rf)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.8237  -0.4933  -0.1760   0.3886   4.0265  

Coefficients:
                                    Estimate Std. Error  z value Pr(>|z|)    
(Intercept)                       -1.132e+01  1.025e-01 -110.421  < 2e-16 ***
Gender                             2.912e-02  2.327e-02    1.252 0.210684    
Customer_Type                      2.067e+00  3.585e-02   57.653  < 2e-16 ***
Age                               -9.524e-03  8.519e-04  -11.179  < 2e-16 ***
Type_of_Travel                     2.714e+00  3.766e-02   72.075  < 2e-16 ***
Flight_Distance                   -3.891e-05  1.345e-05   -2.894 0.003808 ** 
Inflight_wifi_service              3.886e-01  1.369e-02   28.383  < 2e-16 ***
Departure__Arrival_time_convenient -1.185e-01  9.787e-03  -12.110  < 2e-16 ***
Ease_of_Online_booking            -1.509e-01  1.351e-02  -11.170  < 2e-16 ***
Gate_location                      3.159e-02  1.092e-02    2.894 0.003804 ** 
Food_and_drink                    -3.797e-03  1.268e-02   -0.299 0.764661    
Online_boarding                    6.146e-01  1.223e-02   50.259  < 2e-16 ***
Seat_comfort                       7.556e-02  1.336e-02    5.656 1.55e-08 ***
Inflight_entertainment             5.805e-02  1.702e-02    3.412 0.000645 ***
Onboard_service                    3.049e-01  1.219e-02   25.019  < 2e-16 ***
Leg_room_service                   2.573e-01  1.023e-02   25.148  < 2e-16 ***
Baggage_handling                   1.286e-01  1.366e-02    9.410  < 2e-16 ***
Checkin_service                    3.194e-01  1.026e-02   31.136  < 2e-16 ***
Inflight_service                   1.277e-01  1.436e-02    8.893  < 2e-16 ***
Cleanliness                        2.052e-01  1.445e-02   14.197  < 2e-16 ***
Departure_Delay_in_Minutes         4.660e-03  1.183e-03    3.939 8.17e-05 ***
Arrival_Delay_in_Minutes          -9.414e-03  1.167e-03   -8.066 7.25e-16 ***
Class_Business                     7.533e-01  3.073e-02   24.517  < 2e-16 ***
Class_Eco_Plus                    -1.174e-01  4.787e-02   -2.453 0.014157 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99238  on 72515  degrees of freedom
Residual deviance: 48540  on 72492  degrees of freedom
```

```
AIC: 48588

Number of Fisher Scoring iterations: 6

> ##用訓練出的模型來看test data的預測結果(result)
> result <- predict(logit_model,test_rf, type = "response")
> result
           1            2            3            4            5            6            7            8            9 
0.1953713375 0.0152677649 0.0205351588 0.2129458185 0.0211921720 0.9758438826 0.5920361734 0.0630869782 0.0562297547 
          10           11           12           13           14           15           16           17           18 
0.0936733731 0.2912835233 0.7206506677 0.8919306025 0.8711605390 0.0225395456 0.2520578503 0.9557821035 0.9852433453 
          19           20           21           22           23           24           25           26           27 
0.0876708563 0.9757253477 0.0163627395 0.0492029739 0.1675699996 0.4366049755 0.5318634562 0.0227156630 0.1211563093 
          28           29           30           31           32           33           34           35           36 
0.0725212589 0.9850413408 0.0749180356 0.0193048721 0.8822464969 0.8361065001 0.0583160884 0.6155283979 0.4547502271 
          37           38           39           40           41           42           43           44           45 
0.0158274651 0.9737954930 0.0583932401 0.0679573620 0.2362945949 0.3639564211 0.0259921185 0.8860164388 0.2633600770 
          46           47           48           49           50           51           52           53           54 
0.9237097281 0.8943109617 0.1491989327 0.7983966455 0.8767513260 0.0483473978 0.9378497997 0.5352166504 0.2424346036 
          55           56           57           58           59           60           61           62           63 
0.0478758183 0.1165961755 0.0951832677 0.5918177318 0.7971363832 0.9550507200 0.4972939328 0.9736721786 0.0544282700 
          64           65           66           67           68           69           70           71           72 
0.0411061589 0.0135466991 0.9693192887 0.3777524802 0.0419639948 0.0549329066 0.2045974539 0.4763852401 0.8978781118 
          73           74           75           76           77           78           79           80           81 
0.0732670954 0.2523226993 0.9682881353 0.5593846641 0.6548915956 0.8796361297 0.0325260160 0.8727863190 0.8783218299 
          82           83           84           85           86           87           88           89           90 
0.0402852770 0.0120453872 0.5505958383 0.1396620785 0.3509718900 0.9462078741 0.0106394354 0.6100402919 0.8785891939 
          91           92           93           94           95           96           97           98           99 
```

>>說明：利用監督式學習（logistic迴歸配適模型）預設使用全部變數來進行來建模，並進行預測。

```
> ##########模型準確度
> ##計算threshold
> library(InformationValue)
> thres1=optimalCutoff(test_rf$satisfaction, result)
> ##預測結果正確率
> d1 = confusionMatrix(test_rf$satisfaction, result, threshold = thres1)
> (d1[1,1]+d1[2,2])/sum(d1)
[1] 0.8776627
> 
```

>>說明：使用test_rf來比對模型預測正確率，可得出此模型可有87.77%的正確預測率。

- 找出重要變數：哪些因素影響客戶滿意度。What factors are highly correlated to a satisfied (or dissatisfied) passenger?

Ans:

```
> ##########logistic regression配適模型
> logit_model <- glm(satisfaction ~ .,
+                    train_rf, family=binomial(link="logit"))
> summary(logit_model)

Call:
glm(formula = satisfaction ~ ., family = binomial(link = "logit"),
    data = train_rf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8237  -0.4933  -0.1760   0.3886   4.0265

Coefficients:
                                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)                        -1.132e+01  1.025e-01 -110.421  < 2e-16 ***
Gender                              2.912e-02  2.327e-02    1.252 0.210684
Customer_Type                       2.067e+00  3.585e-02   57.653  < 2e-16 ***
Age                                -9.524e-03  8.519e-04  -11.179  < 2e-16 ***
Type_of_Travel                      2.714e+00  3.766e-02   72.075  < 2e-16 ***
Flight_Distance                    -3.891e-05  1.345e-05   -2.894 0.003808 **
Inflight_wifi_service               3.886e-01  1.369e-02   28.383  < 2e-16 ***
Departure__Arrival_time_convenient -1.185e-01  9.787e-03  -12.110  < 2e-16 ***
Ease_of_Online_booking             -1.509e-01  1.351e-02  -11.170  < 2e-16 ***
Gate_location                       3.159e-02  1.092e-02    2.894 0.003804 **
Food_and_drink                     -3.797e-03  1.268e-02   -0.299 0.764661
Online_boarding                     6.146e-01  1.223e-02   50.259  < 2e-16 ***
Seat_comfort                        7.556e-02  1.336e-02    5.656 1.55e-08 ***
Inflight_entertainment              5.805e-02  1.702e-02    3.412 0.000645 ***
Onboard_service                     3.049e-01  1.219e-02   25.019  < 2e-16 ***
Leg_room_service                    2.573e-01  1.023e-02   25.148  < 2e-16 ***
Baggage_handling                    1.286e-01  1.366e-02    9.410  < 2e-16 ***
Checkin_service                     3.194e-01  1.026e-02   31.136  < 2e-16 ***
Inflight_service                    1.277e-01  1.436e-02    8.893  < 2e-16 ***
Cleanliness                         2.052e-01  1.445e-02   14.197  < 2e-16 ***
Departure_Delay_in_Minutes          4.660e-03  1.183e-03    3.939 8.17e-05 ***
Arrival_Delay_in_Minutes           -9.414e-03  1.167e-03   -8.066 7.25e-16 ***
Class_Business                      7.533e-01  3.073e-02   24.517  < 2e-16 ***
Class_Eco_Plus                     -1.174e-01  4.787e-02   -2.453 0.014157 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99238  on 72515  degrees of freedom
Residual deviance: 48540  on 72492  degrees of freedom
```

>>說明：可以從模型的變數顯著來簡易判別，通常***越多代表此變數對於模型來說顯著，也相對重要，可上圖來判別大致有以下這些。

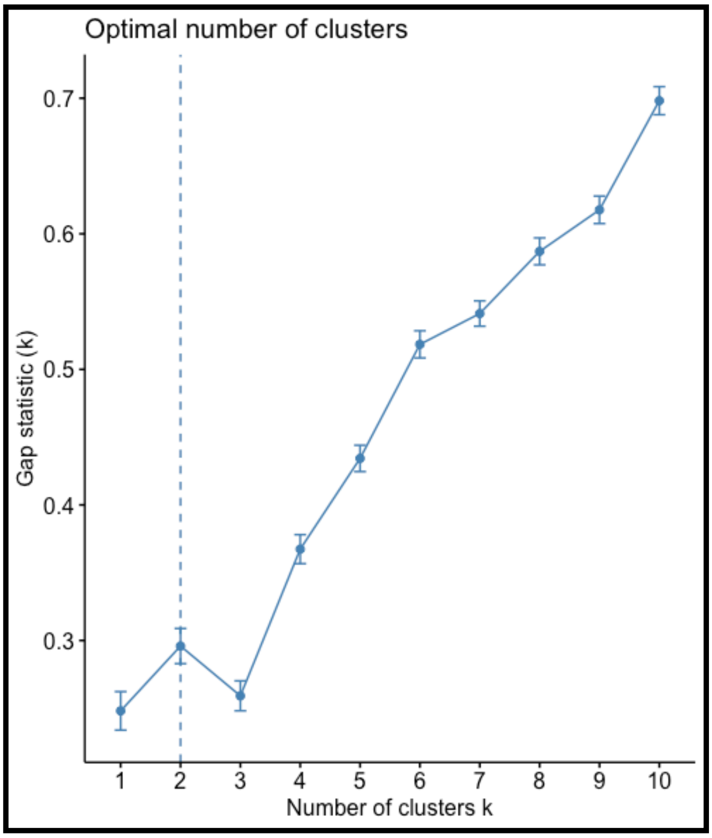
"Arrival_Delay_in_Minutes" "Baggage_handling" "Checkin_service" "Class_Business" "Cleanliness" "Customer_Type" "Departure__Arrival_time_convenient" “Departure_Delay_in_Minutes" "Ease_of_Online_booking" "Flight_Distance" "Gate_location" "Gender" "id" "Index" "Inflight_entertainment" "Inflight_service" "Inflight_wifi_service" "Leg_room_service" "Onboard_service" "Online_boarding" "satisfaction" "Seat_comfort" "Type_of_Travel"

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。Choose one unsupervised method to divide customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

```
> ##########2 非監督式學習
> airline2 <- read_csv("airline_survey.csv")
Rows: 103904 Columns: 25
 0s── Column specification ──────────────────────────────────────────────────────
Delimiter: ","
chr  (5): Gender, Customer Type, Type of Travel, Class, satisfaction
dbl (20): Index, id, Age, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, Ease of Online...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> airline2$Index<-c(1:103904)
> airline2<-na.omit(airline2)
> tinydf <- airline2  %>% sample_frac(0.01)
> ##決定分群k
> library(cluster)
> gap_stat <- clusGap(tinydf[,c(8,23,24)], FUN = kmeans, nstart = 50,
+                     K.max = 10, B = 300)
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 300)  [one "." per sample]:
.................................................. 50
.................................................. 100
.................................................. 150
.................................................. 200
.................................................. 250
.................................................. 300
警告訊息：
1: Quick-TRANSfer stage steps exceeded maximum (= 51800)
2: 10 迭代仍沒有聚合
> fviz_gap_stat(gap_stat)
> |
```



>>說明：利用cluster套件算出推薦k的數量，因而選擇2群。

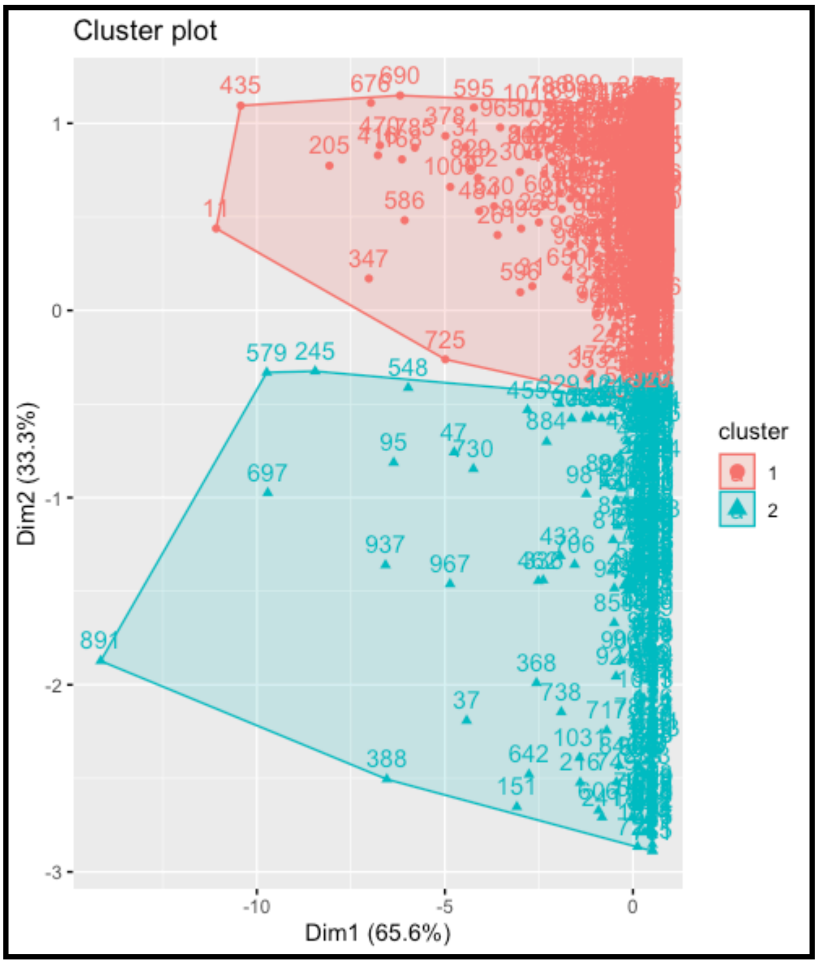
```
> ##
> k = kmeans(tinydf[,c(8,23,24)], centers=2, nstart=50)
> fviz_cluster(k, data = tinydf[,c(8,23,24)])
> table(k$cluster, tinydf$satisfaction)

   neutral or dissatisfied satisfied
  1                     519       244
  2                      93       180
> table(k$cluster, tinydf$`Customer Type`)

   disloyal Customer Loyal Customer
  1               190            573
  2                12            261
> table(k$cluster, tinydf$Class)

   Business Eco Eco Plus
  1      256 444       63
  2      228  40        5
> table(k$cluster, tinydf$`Type of Travel`)

   Business travel Personal Travel
  1             484             279
  2             241              32
> |
```



>>說明：我將客戶分成兩群，其中第一群是在各種不同種類中（不同艙別、忠誠顧客、滿意顧客、出遊目的）均佔有50%以上的客戶（可在程式碼的table比較看出），因為觸及層面多樣，我建議公司可以針對這群顧客做活動推播，因為這群效益較高，受眾且廣（需要主要留住的顧客）。而對第二群的人，則可以增加廣告投入，積極開發此群客戶，來替公司增收益。

附錄: R 程式碼

```
#HW3

setwd("~/Downloads/1102 R/HW/hw 3")

library(readr)
airline <- read_csv("airline_survey.csv")
airline$Index<-c(1:103904)

colnames(airline) <- gsub("/", "__", colnames(airline))
colnames(airline) <- gsub("-", "", colnames(airline))
colnames(airline) <- gsub(" ", "_", colnames(airline))

ls(airline)
airline<-na.omit(airline)

str(airline)
summary(airline)

##將data更改為factor step1
airline$satisfaction<-as.factor(airline$satisfaction)
airline$Gender<-as.factor(airline$Gender)
airline$Customer_Type<-as.factor(airline$Customer_Type)
airline$Type_of_Travel<-as.factor(airline$Type_of_Travel)
airline$Class<-as.factor(airline$Class)
##將data更改為factor step2
airline$Gender <- ifelse(airline$Gender == "Male", 1, 0)
airline$Customer_Type <- ifelse(airline$Customer_Type == "Loyal
Customer", 1, 0)
airline$Type_of_Travel <- ifelse(airline$Type_of_Travel==
"Business travel", 1, 0)
airline$satisfaction <- ifelse(airline$satisfaction ==
"satisfied", 1, 0)

## Use the dummy variable to predict factor variable
library(fastDummies)
d2<-dummy_cols(airline)

#Delete original column and type_Eco
library(dummies)
##dummy 00 = type_Eco
d2<-d2[,c(-7,-27)]

colnames(d2) <- gsub(" ", "_", colnames(d2))

##訓練及測試集
library(tidyverse)
train_df <- d2 %>% group_by(satisfaction) %>% sample_frac(0.7)
test_df  <- anti_join(d2, train_df, by ='Index')
```

```r
train_rf <- train_df[,3:26]
test_rf  <- test_df[,3:26]

str(train_rf)
table(train_df$satisfaction)

##########logistic regression配適模型
logit_model <- glm(satisfaction ~ .,
                   train_rf, family=binomial(link="logit"))
summary(logit_model)
##用訓練出的模型來看test data的預測結果(result)
result <- predict(logit_model,test_rf, type = "response")
result
plot(result)
##########模型準確度
##計算threshold
library(InformationValue)
thres1=optimalCutoff(test_rf$satisfaction, result)

##預測結果正確率
d1 = confusionMatrix(test_rf$satisfaction, result, threshold =
thres1)
(d1[1,1]+d1[2,2])/sum(d1)



##########2 非監督式學習
airline2 <- read_csv("airline_survey.csv")
airline2$Index<-c(1:103904)
airline2<-na.omit(airline2)

tinydf <- airline2  %>% sample_frac(0.01)

##決定分群k
library(cluster)
gap_stat <- clusGap(tinydf[,c(8,23,24)], FUN = kmeans, nstart =
25,
                    K.max = 10, B = 300)
fviz_gap_stat(gap_stat)

##
k = kmeans(tinydf[,c(8,23,24)], centers=2, nstart=25)
fviz_cluster(k, data = tinydf[,c(8,23,24)])


table(k$cluster, tinydf$satisfaction)
table(k$cluster, tinydf$`Customer Type`)
table(k$cluster, tinydf$Class)
table(k$cluster, tinydf$`Type of Travel`)
```