HW3

2024-11-25

#Preparation

```
# 載入資料
library(readr)
data <- read_csv("C:/Users/Ava/Desktop/R/HW3/airline_survey.csv")
```

```
## New names:
## Rows: 103904 Columns: 25
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (5): Gender, Customer Type, Type of Travel, Class, satisfaction dbl (20): ...1,
## id, Age, Flight Distance, Inflight wifi service, Departure/A...
## 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.
## • `` -> `...1`
```

```
View(data)
```

```
# summary
#install.packages("summarytools")
library(summarytools)
```

```
## Warning: 套件 'summarytools' 是用 R 版本 4.4.2 來建造的
```

```
dfSummary(data)
```

```
## Data Frame Summary
## data
## Dimensions: 103904 x 25
## Duplicates: 0
##
## ------------------------------------------------------------------------------------------
--------------------------------------------------
## No   Variable                            Stats / Values                  Freqs (% of Vali
d)       Graph                 Valid      Missing
## ---- ----------------------------------- ------------------------------- -----------------
------- --------------------- ---------- ---------
## 1    ...1                                Mean (sd) : 51951.5 (29994.6)   103904 distinct v
alues   : : : : : : : : : :   103904     0
##      [numeric]                           min < med < max:
: : : : : : : : : :   (100.0%)   (0.0%)
##                                          0 < 51951.5 < 103903
: : : : : : : : : :
##                                          IQR (CV) : 51951.5 (0.6)
: : : : : : : : : :
##
: : : : : : : : : :
##
## 2    id                                  Mean (sd) : 64924.2 (37463.8)   103904 distinct v
alues   : : : : : : : : : :   103904     0
##      [numeric]                           min < med < max:
: : : : : : : : : :   (100.0%)   (0.0%)
##                                          1 < 64856.5 < 129880
: : : : : : : : : :
##                                          IQR (CV) : 64834.5 (0.6)
: : : : : : : : : :
##
: : : : : : : : : :
##
## 3    Gender                              1. Female                       52727 (50.7%)
IIIIIIIII             103904     0
##      [character]                         2. Male                         51177 (49.3%)
IIIIIIII              (100.0%)   (0.0%)
##
## 4    Customer Type                       1. disloyal Customer            18981 (18.3%)
III                   103904     0
##      [character]                         2. Loyal Customer               84923 (81.7%)
IIIIIIIIIIIIIII       (100.0%)   (0.0%)
##
## 5    Age                                 Mean (sd) : 39.4 (15.1)         75 distinct value
s           .   :              103904     0
##      [numeric]                           min < med < max:
: : : : .         (100.0%)   (0.0%)
##                                          7 < 40 < 85
. : : : : :
##                                          IQR (CV) : 24 (0.4)
. : : : : : : : .
##
: : : : : : : : .
##
## 6    Type of Travel                      1. Business travel              71655 (69.0%)
```

```
IIIIIIIIIIIII         103904     0
##      [character]                         2. Personal Travel              32249 (31.0%)
IIIIII                (100.0%)   (0.0%)
##
## 7    Class                               1. Business                     49665 (47.8%)
IIIIIIIII             103904     0
##      [character]                         2. Eco                          46745 (45.0%)
IIIIIIII              (100.0%)   (0.0%)
##                                          3. Eco Plus                      7494 ( 7.2%)
I
##
## 8    Flight Distance                     Mean (sd) : 1189.4 (997.1)      3802 distinct val
ues     :                     103904     0
##      [numeric]                           min < med < max:
: :                   (100.0%)   (0.0%)
##                                          31 < 843 < 4983
: :
##                                          IQR (CV) : 1329 (0.8)
: : . .
##
: : : : : : . .
##
## 9    Inflight wifi service               Mean (sd) : 2.7 (1.3)           0 :  3103 ( 3.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 17840 (17.2%)
III                   (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 25830 (24.9%)
IIII
##                                          IQR (CV) : 2 (0.5)              3 : 25868 (24.9%)
IIII
##                                                                          4 : 19794 (19.1%)
III
##                                                                          5 : 11469 (11.0%)
II
##
## 10   Departure/Arrival time convenient   Mean (sd) : 3.1 (1.5)           0 :  5300 ( 5.1%)
I                     103904     0
##      [numeric]                           min < med < max:                1 : 15498 (14.9%)
II                    (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 17191 (16.5%)
III
##                                          IQR (CV) : 2 (0.5)              3 : 17966 (17.3%)
III
##                                                                          4 : 25546 (24.6%)
IIII
##                                                                          5 : 22403 (21.6%)
IIII
##
## 11   Ease of Online booking              Mean (sd) : 2.8 (1.4)           0 :  4487 ( 4.3%)
103904     0
##      [numeric]                           min < med < max:                1 : 17525 (16.9%)
III                   (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 24021 (23.1%)
IIII
##                                          IQR (CV) : 2 (0.5)              3 : 24449 (23.5%)
IIII
```

```
##                                                                          4 : 19571 (18.8%)
III
##                                                                          5 : 13851 (13.3%)
II
##
## 12   Gate location                       Mean (sd) : 3 (1.3)             0 :     1 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 17562 (16.9%)
III                   (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 19459 (18.7%)
III
##                                          IQR (CV) : 2 (0.4)              3 : 28577 (27.5%)
IIIII
##                                                                          4 : 24426 (23.5%)
IIII
##                                                                          5 : 13879 (13.4%)
II
##
## 13   Food and drink                      Mean (sd) : 3.2 (1.3)           0 :   107 ( 0.1%)
103904     0
##      [numeric]                           min < med < max:                1 : 12837 (12.4%)
II                    (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 21988 (21.2%)
IIII
##                                          IQR (CV) : 2 (0.4)              3 : 22300 (21.5%)
IIII
##                                                                          4 : 24359 (23.4%)
IIII
##                                                                          5 : 22313 (21.5%)
IIII
##
## 14   Online boarding                     Mean (sd) : 3.3 (1.3)           0 :  2428 ( 2.3%)
103904     0
##      [numeric]                           min < med < max:                1 : 10692 (10.3%)
II                    (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 17505 (16.8%)
III
##                                          IQR (CV) : 2 (0.4)              3 : 21804 (21.0%)
IIII
##                                                                          4 : 30762 (29.6%)
IIIII
##                                                                          5 : 20713 (19.9%)
III
##
## 15   Seat comfort                        Mean (sd) : 3.4 (1.3)           0 :     1 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 12075 (11.6%)
II                    (100.0%)   (0.0%)
##                                          0 < 4 < 5                       2 : 14897 (14.3%)
II
##                                          IQR (CV) : 3 (0.4)              3 : 18696 (18.0%)
III
##                                                                          4 : 31765 (30.6%)
IIIIII
##                                                                          5 : 26470 (25.5%)
IIIII
```

```
##
## 16   Inflight entertainment              Mean (sd) : 3.4 (1.3)           0 :    14 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 12478 (12.0%)
II                    (100.0%)   (0.0%)
##                                          0 < 4 < 5                       2 : 17637 (17.0%)
III
##                                          IQR (CV) : 2 (0.4)              3 : 19139 (18.4%)
III
##                                                                          4 : 29423 (28.3%)
IIIII
##                                                                          5 : 25213 (24.3%)
IIII
##
## 17   On-board service                    Mean (sd) : 3.4 (1.3)           0 :     3 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 11872 (11.4%)
II                    (100.0%)   (0.0%)
##                                          0 < 4 < 5                       2 : 14681 (14.1%)
II
##                                          IQR (CV) : 2 (0.4)              3 : 22833 (22.0%)
IIII
##                                                                          4 : 30867 (29.7%)
IIIII
##                                                                          5 : 23648 (22.8%)
IIII
##
## 18   Leg room service                    Mean (sd) : 3.4 (1.3)           0 :   472 ( 0.5%)
103904     0
##      [numeric]                           min < med < max:                1 : 10353 (10.0%)
I                     (100.0%)   (0.0%)
##                                          0 < 4 < 5                       2 : 19525 (18.8%)
III
##                                          IQR (CV) : 2 (0.4)              3 : 20098 (19.3%)
III
##                                                                          4 : 28789 (27.7%)
IIIII
##                                                                          5 : 24667 (23.7%)
IIII
##
## 19   Baggage handling                    Mean (sd) : 3.6 (1.2)           1 :  7237 ( 7.0%)
I                     103904     0
##      [numeric]                           min < med < max:                2 : 11521 (11.1%)
II                    (100.0%)   (0.0%)
##                                          1 < 4 < 5                       3 : 20632 (19.9%)
III
##                                          IQR (CV) : 2 (0.3)              4 : 37383 (36.0%)
IIIIIII
##                                                                          5 : 27131 (26.1%)
IIIII
##
## 20   Checkin service                     Mean (sd) : 3.3 (1.3)           0 :     1 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 12890 (12.4%)
II                    (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 12893 (12.4%)
```

```
II
##                                          IQR (CV) : 1 (0.4)              3 : 28446 (27.4%)
IIIII
##                                                                          4 : 29055 (28.0%)
IIIII
##                                                                          5 : 20619 (19.8%)
III
##
## 21   Inflight service                    Mean (sd) : 3.6 (1.2)           0 :     3 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 :  7084 ( 6.8%)
I                     (100.0%)   (0.0%)
##                                          0 < 4 < 5                       2 : 11457 (11.0%)
II
##                                          IQR (CV) : 2 (0.3)              3 : 20299 (19.5%)
III
##                                                                          4 : 37945 (36.5%)
IIIIIII
##                                                                          5 : 27116 (26.1%)
IIIII
##
## 22   Cleanliness                         Mean (sd) : 3.3 (1.3)           0 :    12 ( 0.0%)
103904     0
##      [numeric]                           min < med < max:                1 : 13318 (12.8%)
II                    (100.0%)   (0.0%)
##                                          0 < 3 < 5                       2 : 16132 (15.5%)
III
##                                          IQR (CV) : 2 (0.4)              3 : 24574 (23.7%)
IIII
##                                                                          4 : 27179 (26.2%)
IIIII
##                                                                          5 : 22689 (21.8%)
IIII
##
## 23   Departure Delay in Minutes          Mean (sd) : 14.8 (38.2)         446 distinct valu
es      :                     103904     0
##      [numeric]                           min < med < max:
:                     (100.0%)   (0.0%)
##                                          0 < 0 < 1592
:
##                                          IQR (CV) : 12 (2.6)
:
##
:
##
## 24   Arrival Delay in Minutes            Mean (sd) : 15.2 (38.7)         455 distinct valu
es      :                     103594     310
##      [numeric]                           min < med < max:
:                     (99.7%)    (0.3%)
##                                          0 < 0 < 1584
:
##                                          IQR (CV) : 13 (2.5)
:
##
:
##
```

```
## 25   satisfaction                        1. neutral or dissatisfied      58879 (56.7%)
IIIIIIIIIII           103904     0
##      [character]                         2. satisfied                    45025 (43.3%)
IIIIIIII              (100.0%)   (0.0%)
## ------------------------------------------------------------------------------------------
--------------------------------------------------
```

```
view(dfSummary(data)) # 看資料的各項分布
```

```
## Switching method to 'browser'
```

```
## Output file written: C:\Users\Ava\AppData\Local\Temp\Rtmp4Q0Bm6\file2a5c745638dc.html
```

```
freq(data)
```

```
## Variable(s) ignored: ...1, id, Age, Flight Distance, Departure Delay in Minutes, Arrival D
elay in Minutes
```

```
## Frequencies
## data$Gender
## Type: Character
##
##                  Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------ -------- --------- -------------- --------- --------------
##       Female    52727     50.75          50.75     50.75          50.75
##         Male    51177     49.25         100.00     49.25         100.00
##         <NA>        0                               0.00         100.00
##        Total   103904    100.00         100.00    100.00         100.00
##
## data$Customer Type
## Type: Character
##
##                             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------------------- -------- --------- -------------- --------- --------------
##       disloyal Customer    18981     18.27          18.27     18.27          18.27
##          Loyal Customer    84923     81.73         100.00     81.73         100.00
##                    <NA>        0                               0.00         100.00
##                   Total   103904    100.00         100.00    100.00         100.00
##
## data$Type of Travel
## Type: Character
##
##                           Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## --------------------- -------- --------- -------------- --------- --------------
##       Business travel    71655     68.96          68.96     68.96          68.96
##       Personal Travel    32249     31.04         100.00     31.04         100.00
##                  <NA>        0                               0.00         100.00
##                 Total   103904    100.00         100.00    100.00         100.00
##
## data$Class
## Type: Character
##
##                    Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## -------------- -------- --------- -------------- --------- --------------
##       Business    49665     47.80          47.80     47.80          47.80
##            Eco    46745     44.99          92.79     44.99          92.79
##       Eco Plus     7494      7.21         100.00      7.21         100.00
##           <NA>        0                               0.00         100.00
##          Total   103904    100.00         100.00    100.00         100.00
##
## data$Inflight wifi service
## Type: Numeric
##
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0     3103      2.99           2.99      2.99           2.99
##           1    17840     17.17          20.16     17.17          20.16
##           2    25830     24.86          45.02     24.86          45.02
##           3    25868     24.90          69.91     24.90          69.91
##           4    19794     19.05          88.96     19.05          88.96
##           5    11469     11.04         100.00     11.04         100.00
##        <NA>        0                               0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
```

```
##
## data$Departure/Arrival time convenient
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0     5300      5.10           5.10      5.10           5.10
##           1    15498     14.92          20.02     14.92          20.02
##           2    17191     16.55          36.56     16.55          36.56
##           3    17966     17.29          53.85     17.29          53.85
##           4    25546     24.59          78.44     24.59          78.44
##           5    22403     21.56         100.00     21.56         100.00
##        <NA>        0                              0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
##
## data$Ease of Online booking
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0     4487      4.32           4.32      4.32           4.32
##           1    17525     16.87          21.18     16.87          21.18
##           2    24021     23.12          44.30     23.12          44.30
##           3    24449     23.53          67.83     23.53          67.83
##           4    19571     18.84          86.67     18.84          86.67
##           5    13851     13.33         100.00     13.33         100.00
##        <NA>        0                              0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
##
## data$Gate location
## Type: Numeric
##
##                Freq     % Valid   % Valid Cum.     % Total   % Total Cum.
## ----------- -------- ----------- -------------- ----------- --------------
##           0        1     0.00096        0.00096     0.00096        0.00096
##           1    17562    16.90214       16.90310    16.90214       16.90310
##           2    19459    18.72786       35.63097    18.72786       35.63097
##           3    28577    27.50327       63.13424    27.50327       63.13424
##           4    24426    23.50824       86.64248    23.50824       86.64248
##           5    13879    13.35752      100.00000    13.35752      100.00000
##        <NA>        0                                0.00000      100.00000
##       Total   103904   100.00000      100.00000   100.00000      100.00000
##
## data$Food and drink
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0      107      0.10           0.10      0.10           0.10
##           1    12837     12.35          12.46     12.35          12.46
##           2    21988     21.16          33.62     21.16          33.62
##           3    22300     21.46          55.08     21.46          55.08
##           4    24359     23.44          78.53     23.44          78.53
##           5    22313     21.47         100.00     21.47         100.00
##        <NA>        0                              0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
```

```
##
## data$Online boarding
## Type: Numeric
##
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0     2428      2.34           2.34      2.34           2.34
##           1    10692     10.29          12.63     10.29          12.63
##           2    17505     16.85          29.47     16.85          29.47
##           3    21804     20.98          50.46     20.98          50.46
##           4    30762     29.61          80.07     29.61          80.07
##           5    20713     19.93         100.00     19.93         100.00
##        <NA>        0                             0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
##
## data$Seat comfort
## Type: Numeric
##
##                 Freq     % Valid   % Valid Cum.     % Total   % Total Cum.
## ----------- -------- ----------- -------------- ----------- --------------
##           0        1     0.00096        0.00096     0.00096        0.00096
##           1    12075    11.62130       11.62227    11.62130       11.62227
##           2    14897    14.33727       25.95954    14.33727       25.95954
##           3    18696    17.99353       43.95307    17.99353       43.95307
##           4    31765    30.57149       74.52456    30.57149       74.52456
##           5    26470    25.47544      100.00000    25.47544      100.00000
##        <NA>        0                               0.00000      100.00000
##       Total   103904   100.00000      100.00000   100.00000      100.00000
##
## data$Inflight entertainment
## Type: Numeric
##
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0       14     0.013          0.013     0.013          0.013
##           1    12478    12.009         12.023    12.009         12.023
##           2    17637    16.974         28.997    16.974         28.997
##           3    19139    18.420         47.417    18.420         47.417
##           4    29423    28.317         75.734    28.317         75.734
##           5    25213    24.266        100.000    24.266        100.000
##        <NA>        0                             0.000        100.000
##       Total   103904   100.000        100.000   100.000        100.000
##
## data$On-board service
## Type: Numeric
##
##                 Freq    % Valid   % Valid Cum.    % Total   % Total Cum.
## ----------- -------- ---------- -------------- ---------- --------------
##           0        3     0.0029         0.0029     0.0029         0.0029
##           1    11872    11.4259        11.4288    11.4259        11.4288
##           2    14681    14.1294        25.5582    14.1294        25.5582
##           3    22833    21.9751        47.5333    21.9751        47.5333
##           4    30867    29.7072        77.2405    29.7072        77.2405
##           5    23648    22.7595       100.0000    22.7595       100.0000
##        <NA>        0                               0.0000       100.0000
##       Total   103904   100.0000       100.0000   100.0000       100.0000
```

```
##
## data$Leg room service
## Type: Numeric
##
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0      472      0.45           0.45      0.45           0.45
##           1    10353      9.96          10.42      9.96          10.42
##           2    19525     18.79          29.21     18.79          29.21
##           3    20098     19.34          48.55     19.34          48.55
##           4    28789     27.71          76.26     27.71          76.26
##           5    24667     23.74         100.00     23.74         100.00
##        <NA>        0                               0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
##
## data$Baggage handling
## Type: Numeric
##
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           1     7237      6.97           6.97      6.97           6.97
##           2    11521     11.09          18.05     11.09          18.05
##           3    20632     19.86          37.91     19.86          37.91
##           4    37383     35.98          73.89     35.98          73.89
##           5    27131     26.11         100.00     26.11         100.00
##        <NA>        0                               0.00         100.00
##       Total   103904    100.00         100.00    100.00         100.00
##
## data$Checkin service
## Type: Numeric
##
##                 Freq     % Valid   % Valid Cum.     % Total   % Total Cum.
## ----------- -------- ----------- -------------- ----------- --------------
##           0        1     0.00096        0.00096     0.00096        0.00096
##           1    12890    12.40568       12.40664    12.40568       12.40664
##           2    12893    12.40857       24.81521    12.40857       24.81521
##           3    28446    27.37719       52.19241    27.37719       52.19241
##           4    29055    27.96331       80.15572    27.96331       80.15572
##           5    20619    19.84428      100.00000    19.84428      100.00000
##        <NA>        0                                0.00000      100.00000
##       Total   103904   100.00000      100.00000   100.00000      100.00000
##
## data$Inflight service
## Type: Numeric
##
##                 Freq    % Valid   % Valid Cum.    % Total   % Total Cum.
## ----------- -------- ---------- -------------- ---------- --------------
##           0        3     0.0029         0.0029     0.0029         0.0029
##           1     7084     6.8178         6.8207     6.8178         6.8207
##           2    11457    11.0265        17.8472    11.0265        17.8472
##           3    20299    19.5363        37.3835    19.5363        37.3835
##           4    37945    36.5193        73.9028    36.5193        73.9028
##           5    27116    26.0972       100.0000    26.0972       100.0000
##        <NA>        0                               0.0000       100.0000
##       Total   103904   100.0000       100.0000   100.0000       100.0000
##
```

```
## data$Cleanliness
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- -------- --------- -------------- --------- --------------
##           0       12     0.012          0.012     0.012          0.012
##           1    13318    12.818         12.829    12.818         12.829
##           2    16132    15.526         28.355    15.526         28.355
##           3    24574    23.651         52.006    23.651         52.006
##           4    27179    26.158         78.163    26.158         78.163
##           5    22689    21.837        100.000    21.837        100.000
##        <NA>        0                              0.000        100.000
##       Total   103904   100.000        100.000   100.000        100.000
##
## data$satisfaction
## Type: Character
##
##                                  Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------------------------- -------- --------- -------------- --------- --------------
##       neutral or dissatisfied    58879     56.67          56.67     56.67          56.67
##                     satisfied    45025     43.33         100.00     43.33         100.00
##                          <NA>        0                               0.00         100.00
##                         Total   103904    100.00         100.00    100.00         100.00
```

```
view(freq(data))
```

```
## Variable(s) ignored: ...1, id, Age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes
```

```
## Output file written: C:\Users\Ava\AppData\Local\Temp\Rtmp4Q0Bm6\file2a5c34f8454c.html
```

```
## Switching method to 'browser'
```

```
## Output file appended: C:\Users\Ava\AppData\Local\Temp\Rtmp4Q0Bm6\file2a5c34f8454c.html
```

```
# 資料清洗
missing_arrival_delay <- data[is.na(data$`Arrival Delay in Minutes`), ]
sum(is.na(data$`Arrival Delay in Minutes`)) # 確認有幾筆缺失值
```

```
## [1] 310
```

```
View(missing_arrival_delay) # 看有缺失值的筆數細項
data_cleaned <- na.omit(data) # 直接刪除有缺失值的資料
View(data_cleaned)
```

Q2.Customer segmentation

```
#install.packages("factoextra")
library(factoextra)
```

```
## Warning: 套件 'factoextra' 是用 R 版本 4.4.2 來建造的
```

```
## 載入需要的套件：ggplot2
```

```
## Warning: 套件 'ggplot2' 是用 R 版本 4.4.2 來建造的
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# 決定最佳群數(使用肘部法則)
# fviz_nbclust(data_cleaned[,9:22], FUNcluster = kmeans, method = "wss", k.max = 20) +
#   labs(title="Elbow Method for K-Means") +
#   geom_vline(xintercept = 3, linetype = 2)
# (我電腦沒有40GB可以跑這行程式碼，所以後面最佳群數用假設的)
```

```
# K-means 分群
k = kmeans(data_cleaned[,9:22], centers=3) # 選取服務滿意度的欄位分群
str(k)
```

```
## List of 9
##  $ cluster     : int [1:103594] 3 1 3 1 2 1 1 2 1 1 ...
##  $ centers     : num [1:3, 1:14] 2.31 3.9 1.91 3.02 3.98 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:14] "Inflight wifi service" "Departure/Arrival time convenient" "Ease of Online booking" "Gate location" ...
##  $ totss       : num 2514603
##  $ withinss    : num [1:3] 738628 508654 555576
##  $ tot.withinss: num 1802858
##  $ betweenss   : num 711745
##  $ size        : int [1:3] 37396 35217 30981
##  $ iter        : int 3
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
k$centers
```

```
##   Inflight wifi service Departure/Arrival time convenient
## 1              2.305059                          3.016285
## 2              3.900048                          3.978703
## 3              1.912075                          2.068720
##   Ease of Online booking Gate location Food and drink Online boarding
## 1               2.587122      2.983929       2.163681        2.555006
## 2               3.889343      3.703439       3.716103        4.062640
## 3               1.674833      2.142959       3.871340        3.166812
##   Seat comfort Inflight entertainment On-board service Leg room service
## 1     2.329233               2.003182         2.728019         2.846401
## 2     4.039867               4.086407         3.831048         3.768606
## 3     4.098092               4.166489         3.662987         3.486718
##   Baggage handling Checkin service Inflight service Cleanliness
## 1         3.106803        2.876885         3.104637    2.106001
## 2         4.000227        3.612517         4.008888    3.888122
## 3         3.846325        3.469933         3.869436    4.027210
```

```
k$withinss
```

```
## [1] 738627.9 508654.1 555576.1
```

```
k$tot.withinss
```

```
## [1] 1802858
```

```
k$size
```

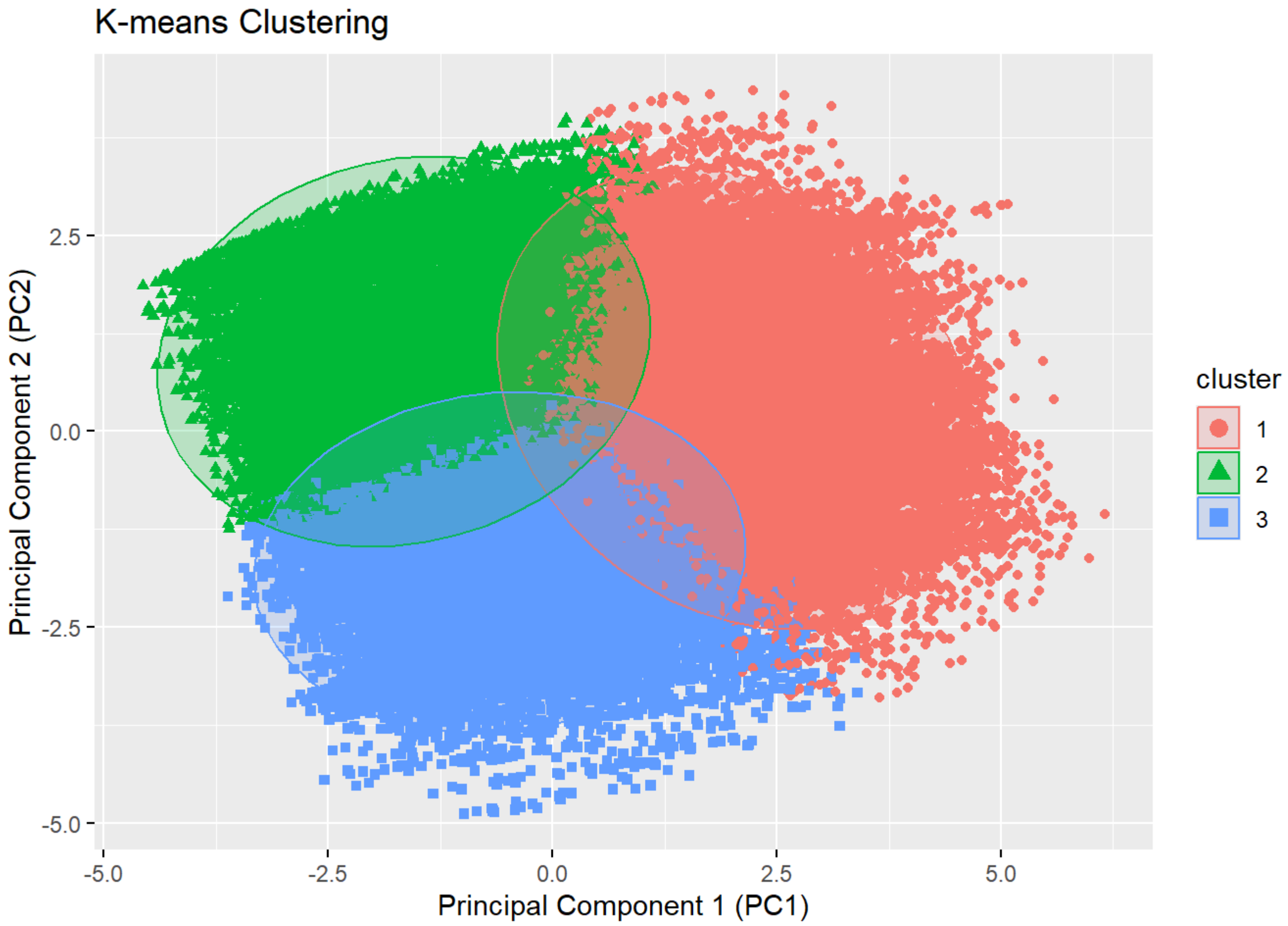
```
## [1] 37396 35217 30981
```

```
data_cleaned$Cluster <- k$cluster
```

```
# 視覺化分群結果
#install.packages("useful")
library(useful)
```

```
## Warning: 套件 'useful' 是用 R 版本 4.4.2 來建造的
```

```
fviz_cluster(k,
             data = data_cleaned[,9:22],
             geom = c("point"),
             ellipse.type = "norm") +
  labs(title = "K-means Clustering",
       x = "Principal Component 1 (PC1)",
       y = "Principal Component 2 (PC2)")
```



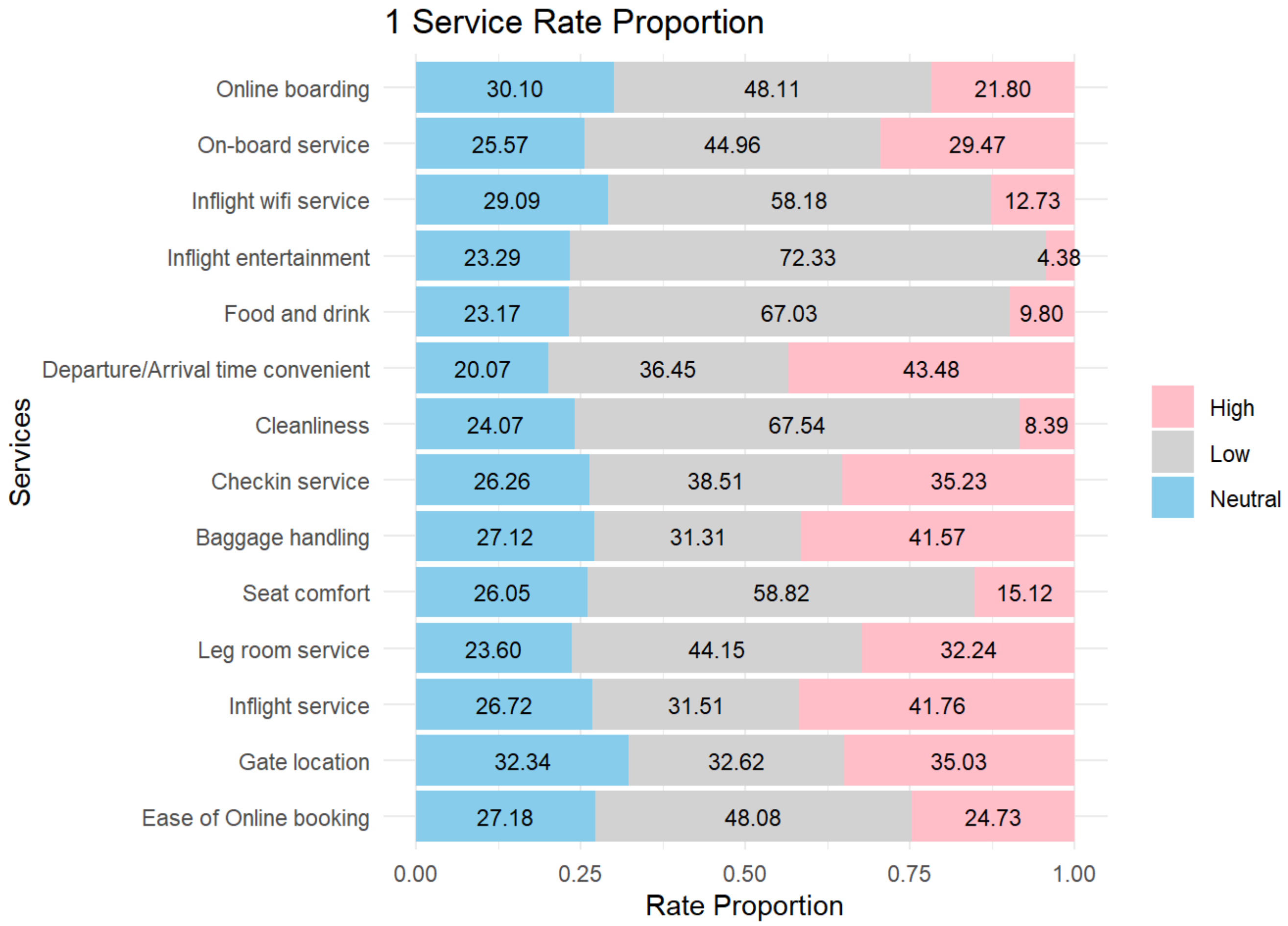
```
# EDA
#install.packages("tidyverse")
library(tidyverse)
```

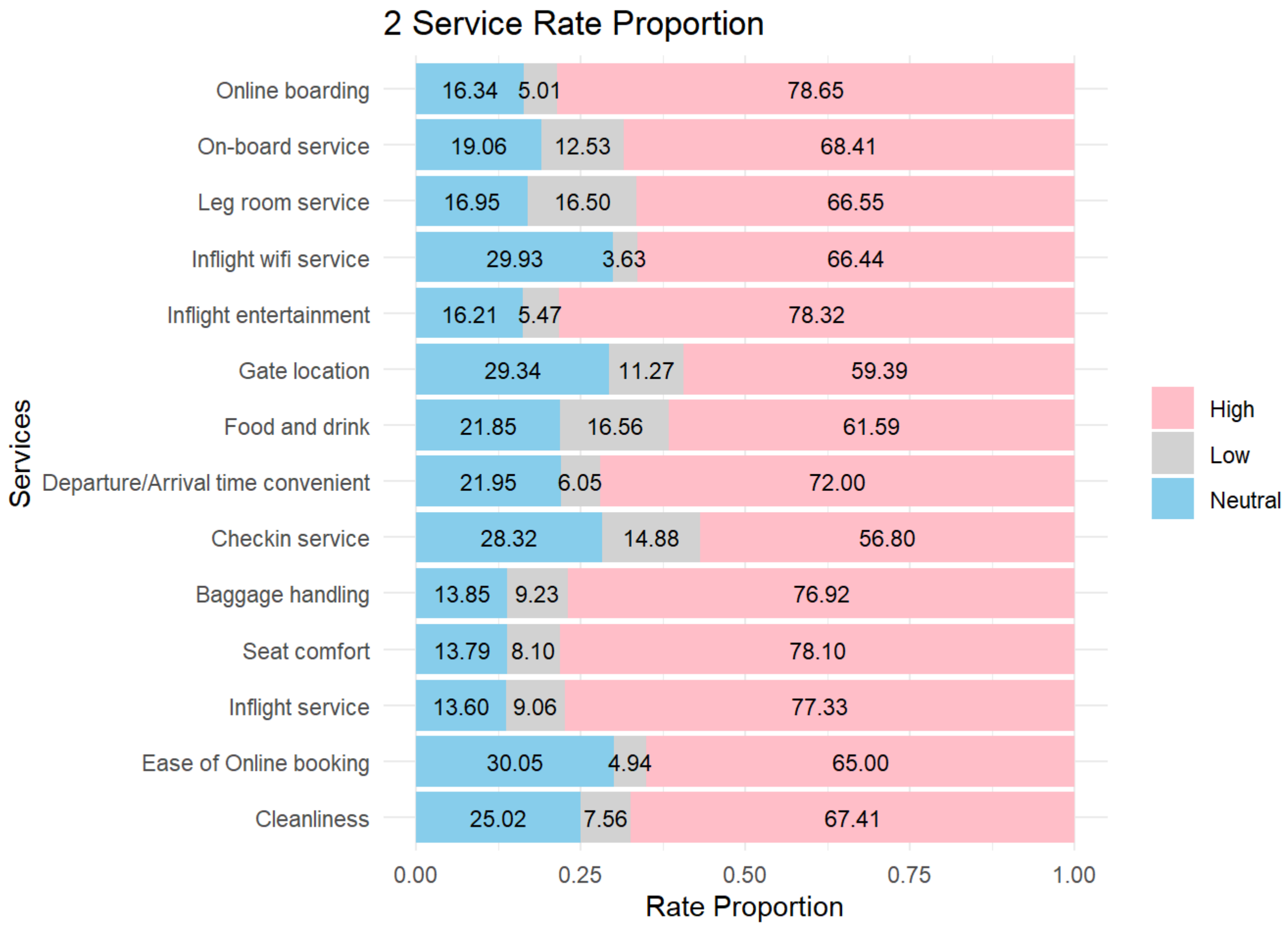
```
## Warning: 套件 'tidyverse' 是用 R 版本 4.4.2 來建造的
```

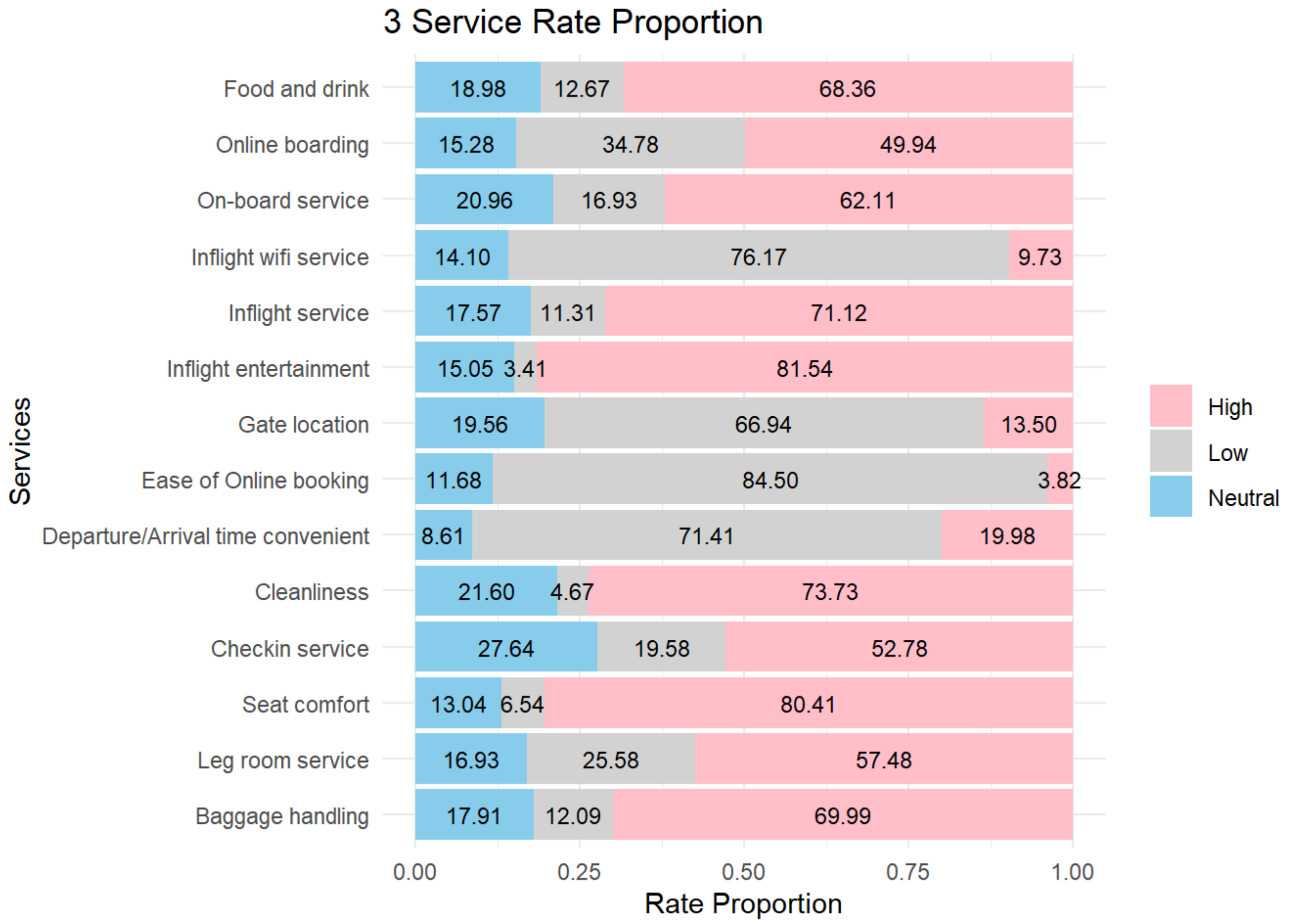
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ stringr   1.5.1
## ✓ forcats   1.0.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ✕ tibble::view()  masks summarytools::view()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
plot_satisfaction_proportion <- function(df, group) {
  service_cols <- names(df)[9:22]
  service_cols <- service_cols[sapply(df[service_cols], is.numeric)]
  # 上面那行是用來確定每一個都是數值型態，理論上應該都是啦，但以防萬一檢查一下

  lowPer <- c()
  neuPer <- c()
  highPer <- c()

  for (col in service_cols) {
    rCount_low <- sum(df[[col]] %in% c(0, 1, 2), na.rm = TRUE)
    rPer_low <- round(rCount_low / nrow(df), 4)
    lowPer <- c(lowPer, rPer_low)

    rCount_neu <- sum(df[[col]] == 3, na.rm = TRUE)
    rPer_neu <- round(rCount_neu / nrow(df), 4)
    neuPer <- c(neuPer, rPer_neu)

    rCount_high <- sum(df[[col]] %in% c(4, 5), na.rm = TRUE)
    rPer_high <- round(rCount_high / nrow(df), 4)
    highPer <- c(highPer, rPer_high)
  }

  df_rate <- data.frame(
    Service = service_cols,
    Low = lowPer,
    Neutral = neuPer,
    High = highPer
  )

  df_rate_long <- df_rate %>%
    pivot_longer(cols = c("Low", "Neutral", "High"), names_to = "Level", values_to = "Proportion")

  plot <- ggplot(df_rate_long, aes(x = reorder(Service, Proportion, sum), y = Proportion, fill = Level)) +
    geom_bar(stat = "identity", position = "stack") +
    coord_flip() +
    labs(x = "Services", y = "Rate Proportion", title = paste(group, "Service Rate Proportion")) +
    scale_fill_manual(values = c("Low" = "lightgrey", "Neutral" = "skyblue", "High" = "pink")) +
    geom_text(aes(label = sprintf("%.2f", Proportion * 100)), position = position_stack(vjust = 0.5), size = 3) +
    theme_minimal() +
    theme(legend.title = element_blank())

  return(plot)
}
```

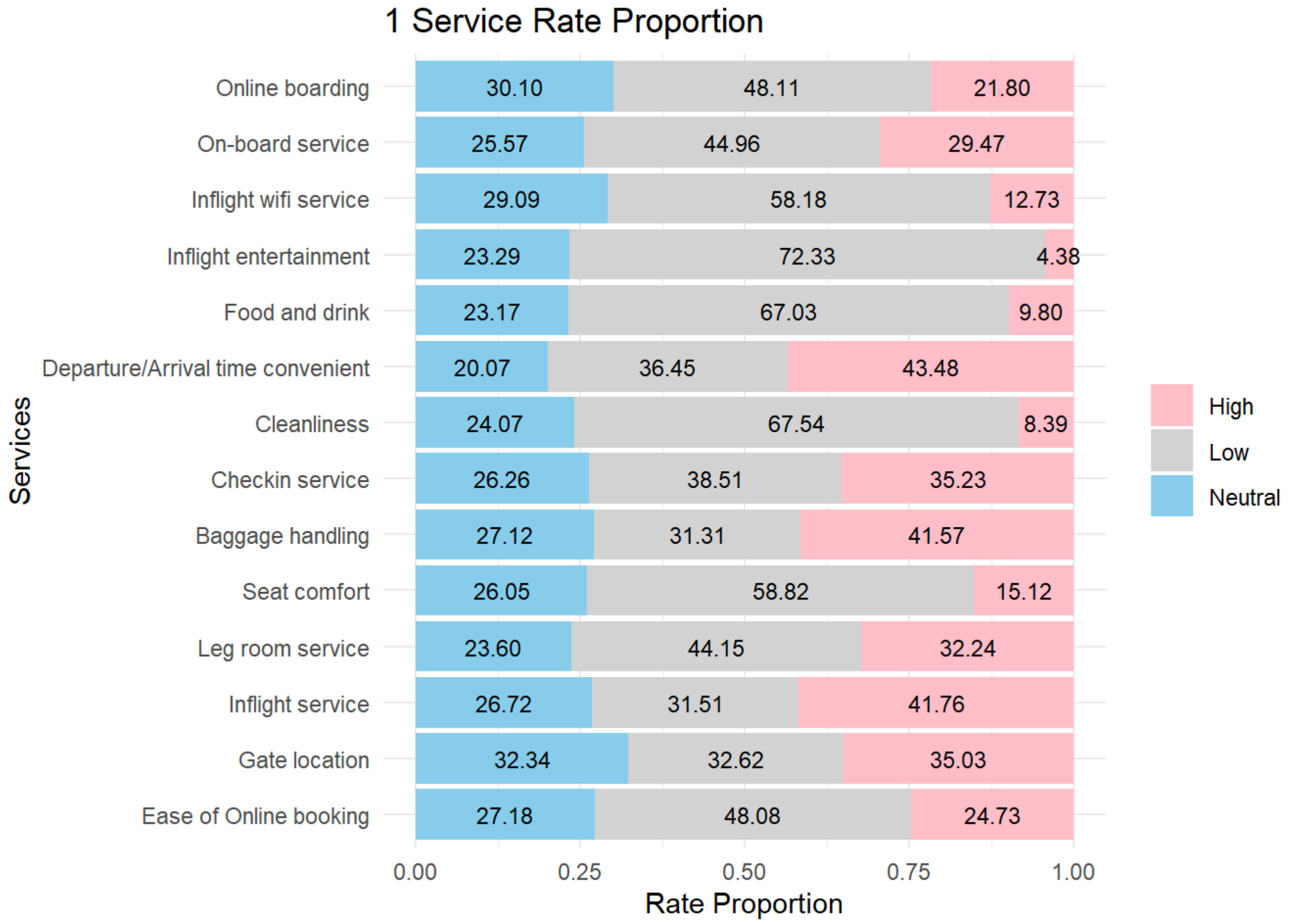
```
# 生成所有群體的圖形
df_split <- split(data_cleaned, data_cleaned$Cluster)
#上面那行是將資料分為三個子資料框，方便之後分開畫三個圖
plots <- lapply(names(df_split), function(Cluster) {
  plot_satisfaction_proportion(df_split[[Cluster]], Cluster)
})

# 顯示圖形（逐一顯示）
lapply(plots, print)
```


1 Service Rate Proportion

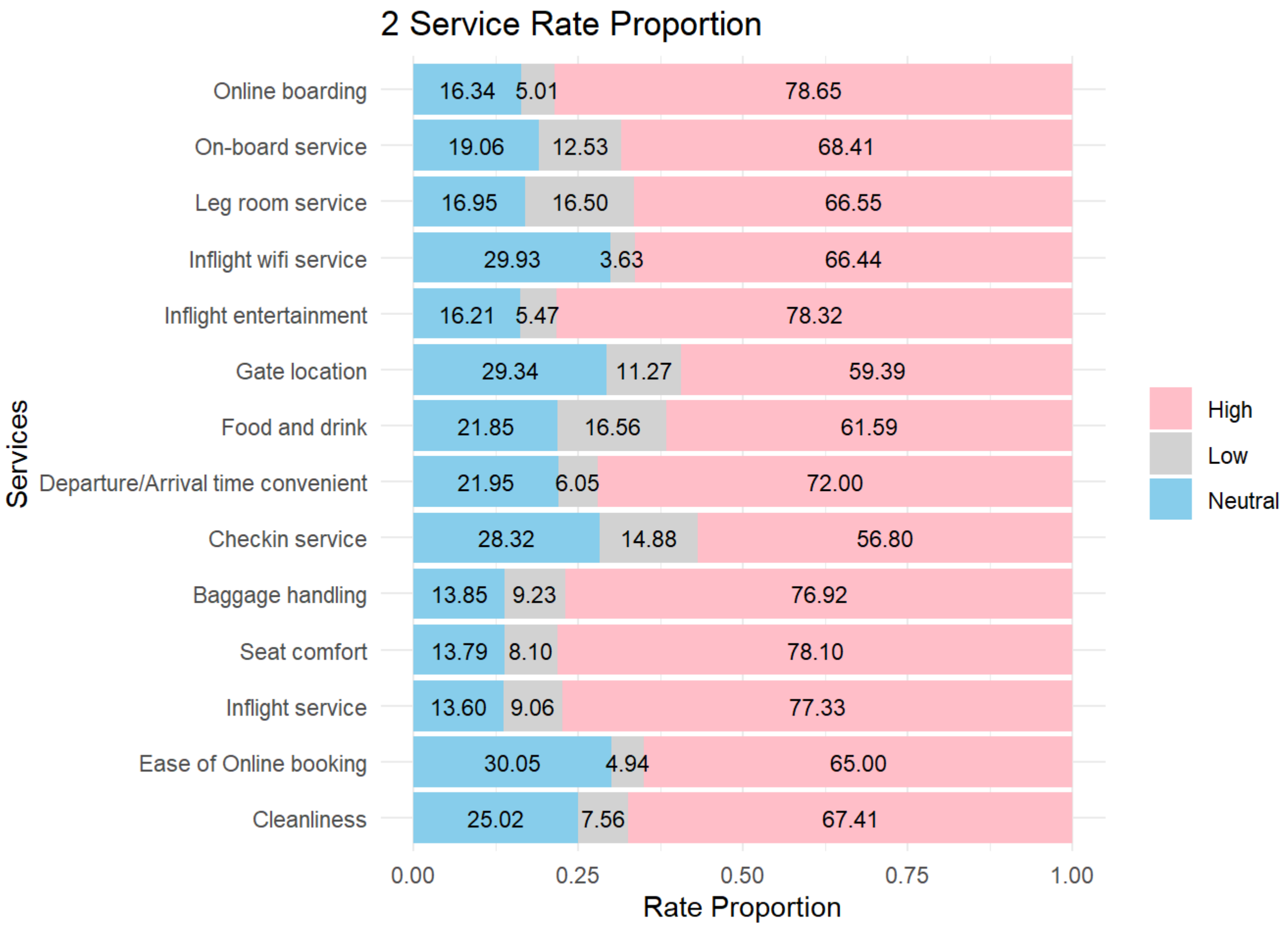

2 Service Rate Proportion


3 Service Rate Proportion
Services
Food and drink 18.98 12.67 68.36
Online boarding 15.28 34.78 49.94
On-board service 20.96 16.93 62.11
Inflight wifi service 14.10 76.17 9.73
Inflight service 17.57 11.31 71.12
Inflight entertainment 15.05 3.41 81.54
Gate location 19.56 66.94 13.50
Ease of Online booking 11.68 84.50 3.82
Departure/Arrival time convenient 8.61 71.41 19.98
Cleanliness 21.60 4.67 73.73
Checkin service 27.64 19.58 52.78
Seat comfort 13.04 6.54 80.41
Leg room service 16.93 25.58 57.48
Baggage handling 17.91 12.09 69.99
0.00
0.25
0.50
0.75
1.00
Rate Proportion
High
Low
Neutral

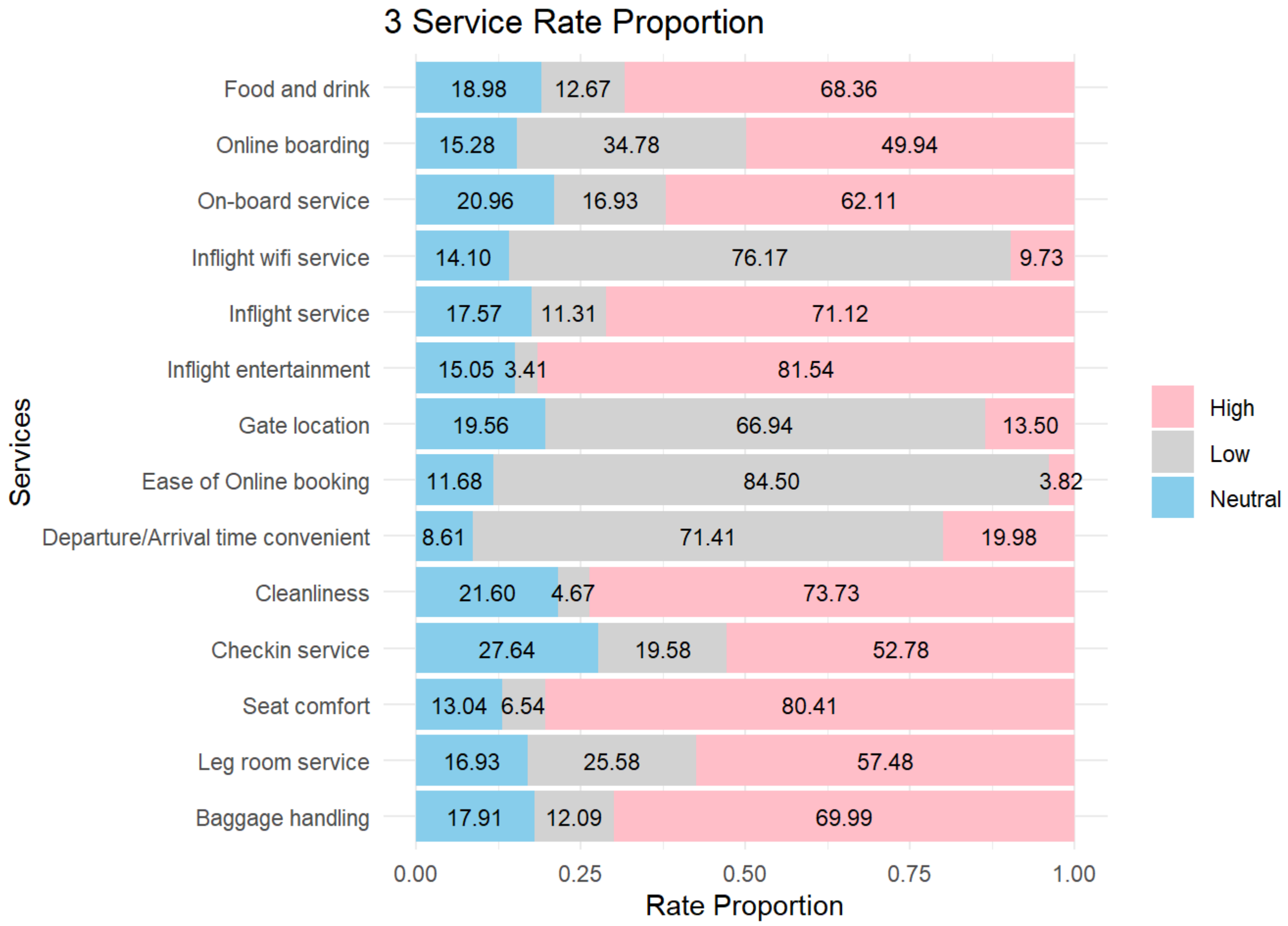
```
## [[1]]
```


1 Service Rate Proportion
Services
Online boarding 30.10 48.11 21.80
On-board service 25.57 44.96 29.47
Inflight wifi service 29.09 58.18 12.73
Inflight entertainment 23.29 72.33 4.38
Food and drink 23.17 67.03 9.80
Departure/Arrival time convenient 20.07 36.45 43.48
Cleanliness 24.07 67.54 8.39
Checkin service 26.26 38.51 35.23
Baggage handling 27.12 31.31 41.57
Seat comfort 26.05 58.82 15.12
Leg room service 23.60 44.15 32.24
Inflight service 26.72 31.51 41.76
Gate location 32.34 32.62 35.03
Ease of Online booking 27.18 48.08 24.73
0.00
0.25
0.50
0.75
1.00
Rate Proportion
High
Low
Neutral

```
##
## [[2]]
```


2 Service Rate Proportion
Services
Online boarding 16.34 5.01 78.65
On-board service 19.06 12.53 68.41
Leg room service 16.95 16.50 66.55
Inflight wifi service 29.93 3.63 66.44
Inflight entertainment 16.21 5.47 78.32
Gate location 29.34 11.27 59.39
Food and drink 21.85 16.56 61.59
Departure/Arrival time convenient 21.95 6.05 72.00
Checkin service 28.32 14.88 56.80
Baggage handling 13.85 9.23 76.92
Seat comfort 13.79 8.10 78.10
Inflight service 13.60 9.06 77.33
Ease of Online booking 30.05 4.94 65.00
Cleanliness 25.02 7.56 67.41
0.00
0.25
0.50
0.75
1.00
Rate Proportion
High
Low
Neutral

```
##
## [[3]]
```



Q1.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)。

找出重要變數：哪些因素影響客戶滿意度。

```
#install.packages("MASS")
library(MASS)
```

```
## Warning: 套件 'MASS' 是用 R 版本 4.4.2 來建造的
```

```
## 
## 載入套件：'MASS'
```

```
## 下列物件被遮斷自 'package:dplyr':
## 
##     select
```

```
#install.packages("randomForest")
library(randomForest)# Q1.Predict passenger satisfaction
```

```
## Warning: 套件 'randomForest' 是用 R 版本 4.4.2 來建造的
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## 載入套件：'randomForest'
```

```
## 下列物件被遮斷自 'package:dplyr':
##
##     combine
```

```
## 下列物件被遮斷自 'package:ggplot2':
##
##     margin
```

```
# 隨機分割訓練集和測試集
# 清理變數名稱
colnames(data_cleaned) <- make.names(colnames(data_cleaned), unique = TRUE)
print(colnames(data_cleaned))
```

```
##  [1] "...1"                              "id"
##  [3] "Gender"                            "Customer.Type"
##  [5] "Age"                               "Type.of.Travel"
##  [7] "Class"                             "Flight.Distance"
##  [9] "Inflight.wifi.service"             "Departure.Arrival.time.convenient"
## [11] "Ease.of.Online.booking"            "Gate.location"
## [13] "Food.and.drink"                    "Online.boarding"
## [15] "Seat.comfort"                      "Inflight.entertainment"
## [17] "On.board.service"                  "Leg.room.service"
## [19] "Baggage.handling"                  "Checkin.service"
## [21] "Inflight.service"                  "Cleanliness"
## [23] "Departure.Delay.in.Minutes"        "Arrival.Delay.in.Minutes"
## [25] "satisfaction"                      "Cluster"
```

```
trainI <- sample(1:nrow(data_cleaned), 51797)
traind <- data_cleaned[trainI,]
testd <- data_cleaned[-trainI,]
```

```
# 選擇需要的特徵欄位，並將滿意度變數轉為因子類型
traind_selected <- traind[, c(9:22, 25)]  # 滿意度在第25欄
testd_selected <- testd[, c(9:22, 25)]

traind_selected$satisfaction <- as.factor(traind_selected$satisfaction)
testd_selected$satisfaction <- as.factor(testd_selected$satisfaction)
```

```
# 確認資料結構
str(traind_selected)
```

```
## tibble [51,797 × 15] (S3: tbl_df/tbl/data.frame)
##  $ Inflight.wifi.service            : num [1:51797] 3 1 4 4 1 3 3 5 2 4 ...
##  $ Departure.Arrival.time.convenient: num [1:51797] 4 1 4 3 1 3 5 3 2 4 ...
##  $ Ease.of.Online.booking           : num [1:51797] 3 1 4 4 1 3 3 3 2 4 ...
##  $ Gate.location                    : num [1:51797] 4 1 4 4 3 3 3 3 2 4 ...
##  $ Food.and.drink                   : num [1:51797] 3 1 2 2 3 2 5 3 2 4 ...
##  $ Online.boarding                  : num [1:51797] 3 4 3 2 1 5 3 5 2 4 ...
##  $ Seat.comfort                     : num [1:51797] 3 4 5 4 1 4 5 2 2 4 ...
##  $ Inflight.entertainment           : num [1:51797] 3 2 5 5 3 4 5 5 2 4 ...
##  $ On.board.service                 : num [1:51797] 4 4 5 5 3 4 4 5 2 4 ...
##  $ Leg.room.service                 : num [1:51797] 5 3 5 5 2 4 5 5 2 3 ...
##  $ Baggage.handling                 : num [1:51797] 4 4 5 5 1 4 5 5 4 4 ...
##  $ Checkin.service                  : num [1:51797] 4 4 5 3 2 3 3 3 2 5 ...
##  $ Inflight.service                 : num [1:51797] 5 4 5 5 3 4 4 5 4 5 ...
##  $ Cleanliness                      : num [1:51797] 3 4 4 4 3 5 5 5 2 4 ...
##  $ satisfaction                     : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1
2 2 1 2 1 2 1 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:310] 214 1125 1530 2005 2109 2486 2631 3622 4
042 4491 ...
##   ..- attr(*, "names")= chr [1:310] "214" "1125" "1530" "2005" ...
```

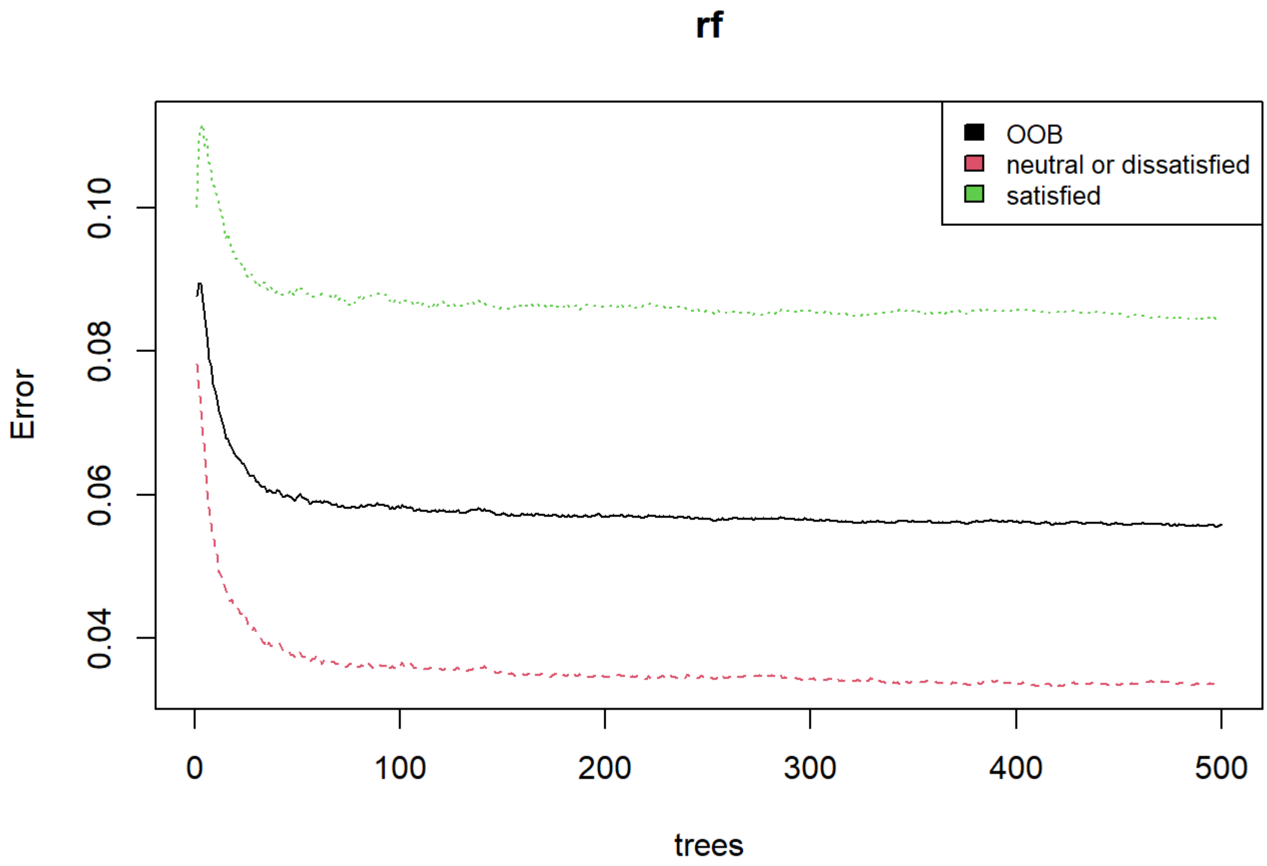
```
str(testd_selected)
```

```
## tibble [51,797 × 15] (S3: tbl_df/tbl/data.frame)
##  $ Inflight.wifi.service            : num [1:51797] 3 2 3 4 1 2 1 4 3 2 ...
##  $ Departure.Arrival.time.convenient: num [1:51797] 2 5 4 3 2 4 4 2 2 1 ...
##  $ Ease.of.Online.booking           : num [1:51797] 3 5 2 4 2 2 4 4 3 2 ...
##  $ Gate.location                    : num [1:51797] 3 5 1 4 2 2 4 3 2 3 ...
##  $ Food.and.drink                   : num [1:51797] 1 2 1 5 4 1 1 4 2 4 ...
##  $ Online.boarding                  : num [1:51797] 3 2 2 5 3 2 1 4 3 2 ...
##  $ Seat.comfort                     : num [1:51797] 1 2 1 5 3 1 1 4 2 1 ...
##  $ Inflight.entertainment           : num [1:51797] 1 2 1 5 1 1 1 4 2 4 ...
##  $ On.board.service                 : num [1:51797] 1 2 3 5 1 1 1 4 4 2 ...
##  $ Leg.room.service                 : num [1:51797] 5 5 4 5 2 2 1 5 3 1 ...
##  $ Baggage.handling                 : num [1:51797] 3 3 4 5 1 5 3 2 2 4 ...
##  $ Checkin.service                  : num [1:51797] 1 1 4 4 4 5 4 2 2 1 ...
##  $ Inflight.service                 : num [1:51797] 4 4 4 5 1 5 4 2 1 3 ...
##  $ Cleanliness                      : num [1:51797] 1 2 1 4 2 1 1 4 2 4 ...
##  $ satisfaction                     : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1
1 2 1 1 1 2 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:310] 214 1125 1530 2005 2109 2486 2631 3622 4
042 4491 ...
##   ..- attr(*, "names")= chr [1:310] "214" "1125" "1530" "2005" ...
```

```
# 建立隨機森林模型
rf <- randomForest(satisfaction ~ ., data = traind_selected, importance = TRUE)
print(rf)
```

```
##
## Call:
##  randomForest(formula = satisfaction ~ ., data = traind_selected,      importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 5.59%
## Confusion matrix:
##                         neutral or dissatisfied satisfied class.error
## neutral or dissatisfied                   28385       990  0.03370213
## satisfied                                  1903     20519  0.08487200
```

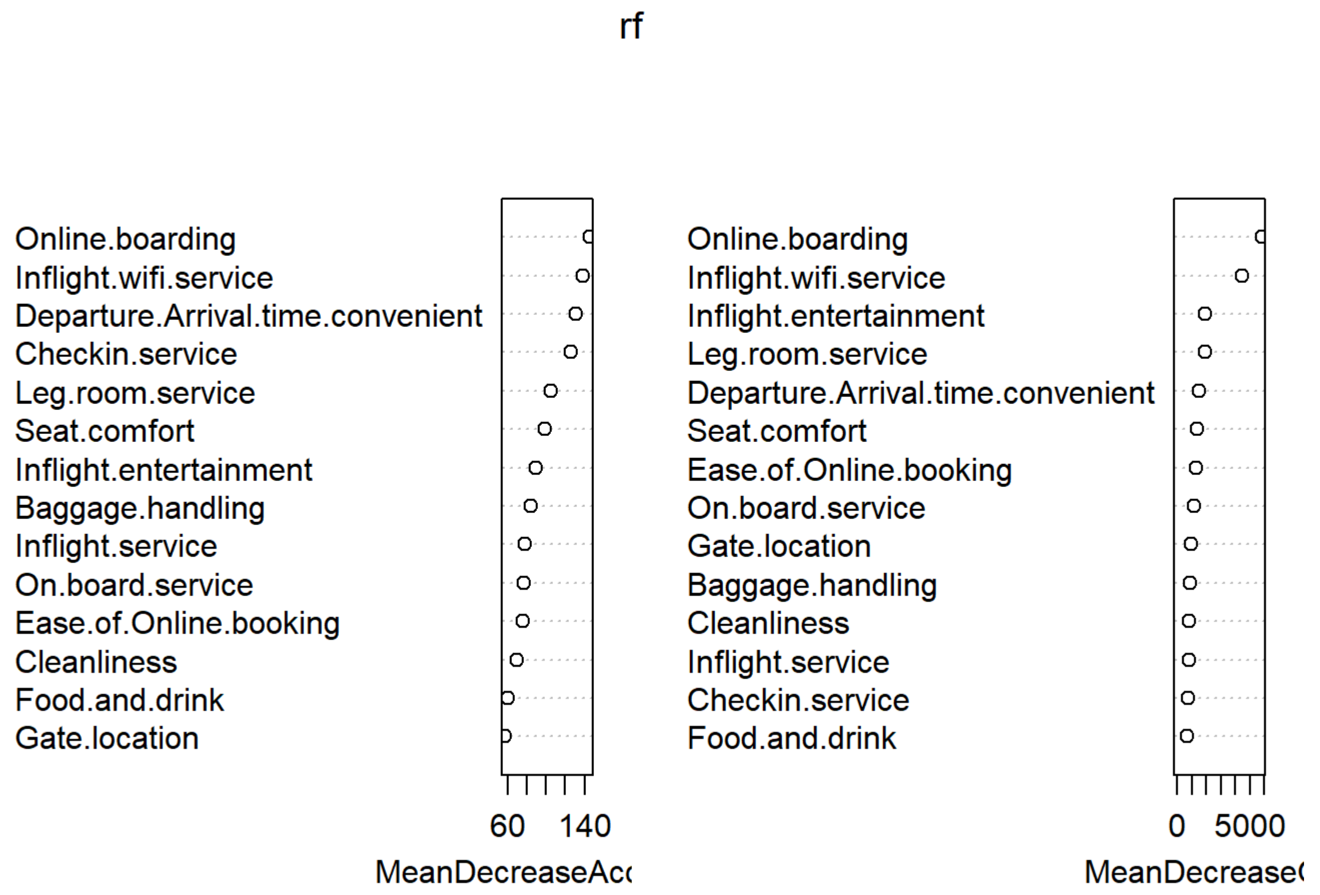
```
# 視覺化錯誤率隨著樹數的變化
plot(rf)
legend("topright", colnames(rf$err.rate), col = 1:4, cex = 0.8, fill = 1:4)
```



```
# 評估變數重要性
importance(rf)
```

```
##                                   neutral or dissatisfied satisfied
## Inflight.wifi.service                           146.33129  98.05745
## Departure.Arrival.time.convenient               107.59545 108.48301
## Ease.of.Online.booking                           66.02662  53.47760
## Gate.location                                    24.60070  55.07099
## Food.and.drink                                   40.53854  51.26370
## Online.boarding                                 114.86419 140.67822
## Seat.comfort                                     88.93986  65.84248
## Inflight.entertainment                           61.91215  77.26274
## On.board.service                                 61.44378  60.55205
## Leg.room.service                                 77.75581  87.71877
## Baggage.handling                                 84.82988  55.60058
## Checkin.service                                 114.33969  52.50394
## Inflight.service                                 74.96390  52.44417
## Cleanliness                                      54.96988  51.41146
##                                   MeanDecreaseAccuracy MeanDecreaseGini
## Inflight.wifi.service                        138.20685        4478.6967
## Departure.Arrival.time.convenient            130.91415        1486.4767
## Ease.of.Online.booking                        75.57719        1291.5536
## Gate.location                                 57.14426         969.5097
## Food.and.drink                                59.83003         694.2901
## Online.boarding                              145.45804        5827.9352
## Seat.comfort                                  98.54041        1360.3793
## Inflight.entertainment                        89.46290        1921.5451
## On.board.service                              76.53113        1134.0046
## Leg.room.service                             104.80274        1880.1505
## Baggage.handling                              83.89769         864.7709
## Checkin.service                              125.61768         722.3460
## Inflight.service                              77.44926         783.1376
## Cleanliness                                   69.53207         828.1477
```

```
varImpPlot(rf)
```



```
#Mean Decrease Accuracy - How much the model accuracy decreases if we drop that variable.
#Mean Decrease Gini - Measure of variable importance based on the Gini impurity index used fo
r the calculation of splits in trees.
```

```
# 預測測試集的滿意度
pred <- predict(rf, newdata = testd_selected)
```

```
# 混淆矩陣：實際值與預測值的對比
conf_matrix <- table(Real = testd_selected$satisfaction, Predict = pred)
```

```
#計算分數
#準確率（Accuracy）
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix) # diag(conf_matrix)是左上那一格
print(paste("Accuracy:", round(accuracy, 4)))
```

```
## [1] "Accuracy: 0.9451"
```

```
#精確率 (Precision) 和 召回率 (Recall)
precision <- diag(conf_matrix) / colSums(conf_matrix)
recall <- diag(conf_matrix) / rowSums(conf_matrix)
print(data.frame(Class = rownames(conf_matrix), Precision = precision, Recall = recall))
```

```
##                                              Class Precision    Recall
## neutral or dissatisfied neutral or dissatisfied 0.9389298 0.9658277
## satisfied                             satisfied 0.9536862 0.9180423
```

```
#F1 分數
f1_score <- 2 * (precision * recall) / (precision + recall)
print(data.frame(Class = rownames(conf_matrix), F1_Score = f1_score))
```

```
##                                              Class  F1_Score
## neutral or dissatisfied neutral or dissatisfied 0.9521888
## satisfied                             satisfied 0.9355248
```

```
#混淆矩陣可視化
#install.packages("caret")
library(caret)
```

```
## Warning: 套件 'caret' 是用 R 版本 4.4.2 來建造的
```

```
## 載入需要的套件：lattice
```

```
## 
## 載入套件：'caret'
```

```
## 下列物件被遮斷自 'package:purrr':
## 
##     lift
```

```
confusionMatrix(pred, testd_selected$satisfaction)
```

```
## Confusion Matrix and Statistics
## 
##                           Reference
## Prediction                 neutral or dissatisfied satisfied
##   neutral or dissatisfied                   28320      1842
##   satisfied                                  1002     20633
##                                                             
##                Accuracy : 0.9451          
##                  95% CI : (0.9431, 0.947)
##     No Information Rate : 0.5661          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.8877          
##                                           
##  Mcnemar's Test P-Value : < 2.2e-16       
##                                           
##             Sensitivity : 0.9658          
##             Specificity : 0.9180          
##          Pos Pred Value : 0.9389          
##          Neg Pred Value : 0.9537          
##              Prevalence : 0.5661          
##          Detection Rate : 0.5467          
##    Detection Prevalence : 0.5823          
##       Balanced Accuracy : 0.9419          
##                                           
##        'Positive' Class : neutral or dissatisfied
## 
```

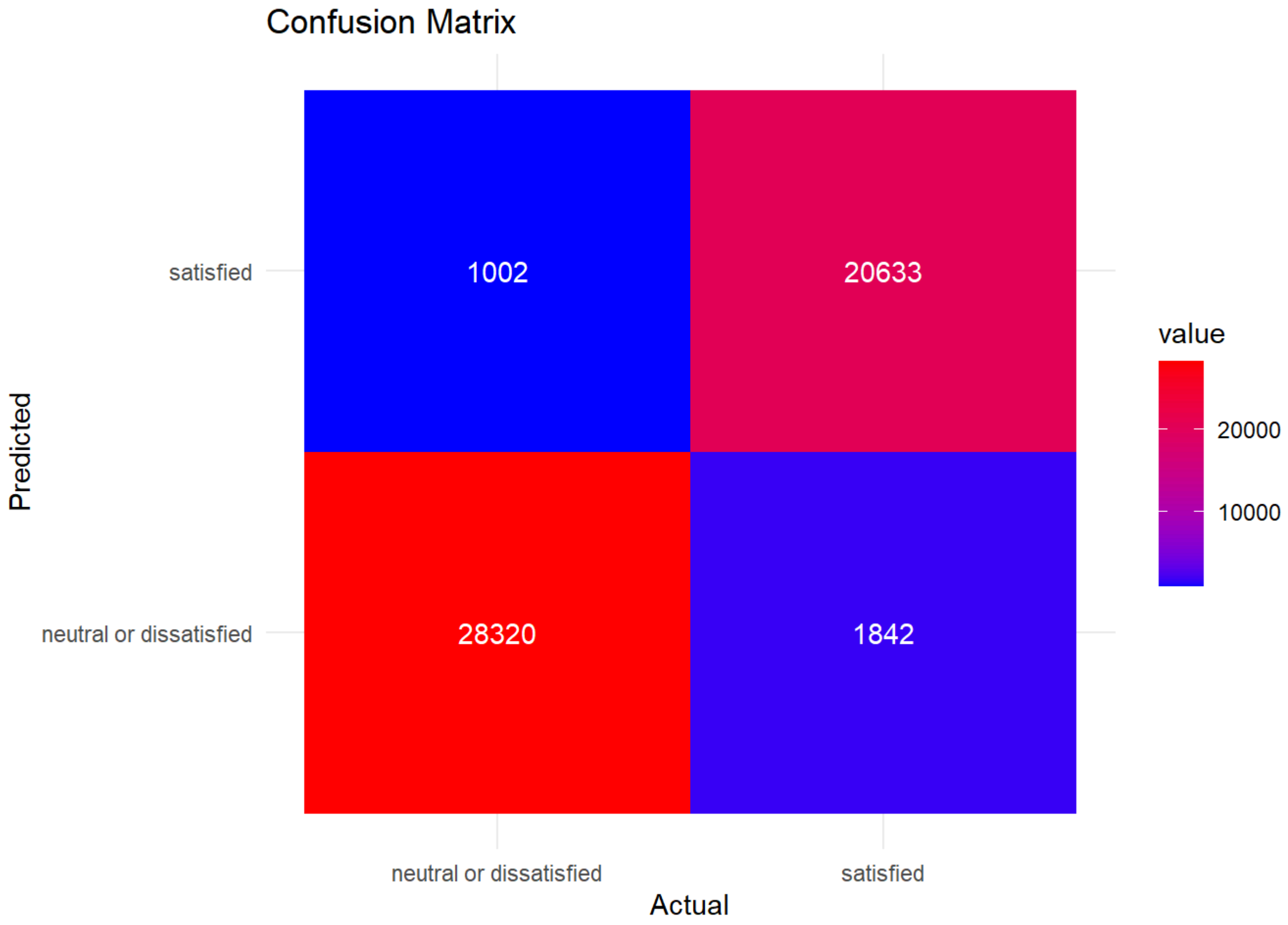
```
library(reshape2)
```

```
## Warning: 套件 'reshape2' 是用 R 版本 4.4.2 來建造的
```

```
## 
## 載入套件：'reshape2'
```

```
## 下列物件被遮斷自 'package:tidyr':
## 
##     smiths
```

```
library(ggplot2)
conf_matrix_melt <- melt(conf_matrix)
ggplot(data = conf_matrix_melt, aes(x = Real, y = Predict, fill = value)) +
  geom_tile() +
  geom_text(aes(label = value), color = "white") +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Confusion Matrix", x = "Actual", y = "Predicted") +
  theme_minimal()
```


Confusion Matrix
Predicted
satisfied
neutral or dissatisfied
1002
20633
28320
1842
neutral or dissatisfied
satisfied
Actual
value
20000
10000