# 多變量期末報告

# 主題：Airline Passenger Satisfaction Prediction

組員：110354013 陳韋豫

110354024 林韋成

110354029 陳槐廷

110354030 程長磊

# 一、 動機與目的：

乘客在選擇航空公司時，會考慮很多因素，像是飛機上的餐點、座椅舒適度、服務態度等，找出哪些因素和乘客對航空公司的整體滿意度有很大的相關，透過調整這些相關因素提升乘客對航空公司整體滿意度，吸引更多乘客前來搭乘。

找出哪些乘客在滿意度調查表中雖是勾選滿意此家航空公司，但是預測出的結果是中立或著是不滿意，針對此類的乘客去做進一步的了解，是否是預測錯誤還是為潛在不滿意的乘客，如此一來，航空公司可以做出相對應的改進，並確實照顧到每一位乘客的需求。

# 二、 資料描述：

資料來源：Kaggle

資料網址：https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

資料內容：103904 筆訓練集，25976 筆測試集，總共 23 個變數。

<u>變數介紹</u>

預測變數：

Gender：乘客性別（女、男）

Customer Type：乘客類型（忠實乘客、非忠實乘客）

Age：乘客的實際年齡

Type of Travel：乘客的飛行目的（個人旅行、商務旅行）

Class：乘客艙別（商務、經濟、豪華經濟）

Flight Distance：飛行距離

（0：NA、1-5：1 為非常不滿意、5 為非常滿意）

Inflight wifi service：機上 wifi 服務滿意度

Departure/Arrival time convenient：出發/到達時間方便的滿意度

Ease of Online booking：在線預訂的滿意度

Gate location：登機門位置的滿意度

Food and drink：食品和飲料的滿意度

Online boarding：線上登機報到的滿意度

Seat comfort：座椅舒適度滿意度

Inflight entertainment：機上娛樂滿意度

On-board service：機上服務滿意度

Leg room service：腿部空間滿意度

Baggage handling：行李處理滿意度

Check-in service：登機服務滿意度

Inflight service：機上服務滿意度

Cleanliness：清潔度滿意度

Departure Delay in Minutes：出發時延遲分鐘數

Arrival Delay in Minutes：到達時延遲分鐘數

反應變數：

Satisfaction：航空公司的滿意度（滿意、中立或不滿意）

# 三、 研究方法：

透過探索性資料分析觀察乘客對航空公司滿意度比例，並檢視不同艙等下或不同客戶類型下乘客對航空公司滿意度比例是否有差異，做資料的初步分析。另外找出有哪些因素對航空公司滿意度有很大的影響。

MCA(Multiple Corresponding Analysis) 是透過指定觀察值和類別的數值來量化類別資料，因此相同類別的個體會緊密在一起，而不同類別的個體則會分開。這裡我們研究哪些滿意度與總體滿意度有較大的關聯，藉由將各個滿意度的變數降維並投影到各維度上面，並找出圖中投影到各個維度上的變數哪些與總滿意度較為靠近，以觀察出影響總體滿意度最有關係之滿意度變數。

MDS(Multidimensional Scaling) 是藉由觀察值之間的距離計算相似度(similarity)或相異度(dissimilarity)，在保留資料原始關係的情況下，將高維度資料投影到低維度空間進行分類或定位。在此，我們的資料中大多數為滿意度資料，計算順序型資料 (ordinal data)時的相異度無法滿足三角不等式，代表我們無法使用歐式距離(Euclidean distances)。我們將使用 MDS 方法將資料做分群或者效仿 PCA 方法中的 biplot 觀察資料之間是否存在某些關係，其中使用的變數為飛行距離和所有項目的滿意度(不包含乘客對航空公司的滿意度)，共 15 個變數，當中飛行距離為連續型變數，其餘為類別變數。由於 training data 筆數非常多，約有 10 萬筆，故我們直接將含有遺失值的資料刪除，並且從中隨機抽取 2000 筆來代表整體資料。此外，為了去除飛行距離單位，將其除上自己的標準差，以免單位影響分析的結果。

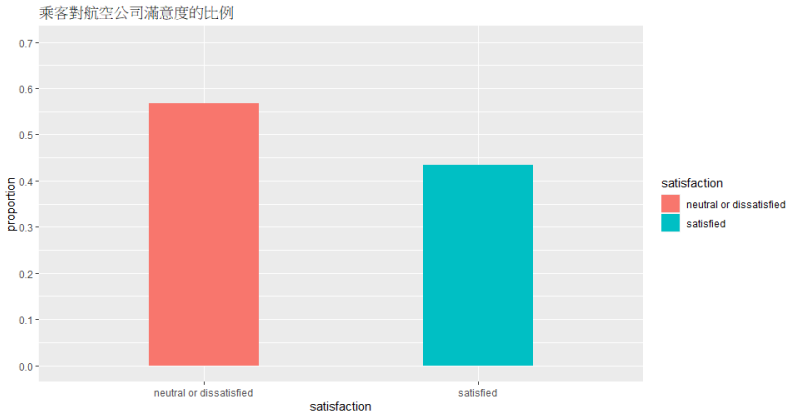
利用除了年齡以外的所有變數建立隨機森林模型預測乘客對航空公司的滿意度，並利用 OOB error 選擇決策樹的數量，進而不斷的迭代得出分類準則。依照隨

機森林給出的各變數 importance，判斷哪些變數對航空公司的滿意度影響較大。

# 四、 研究結果：

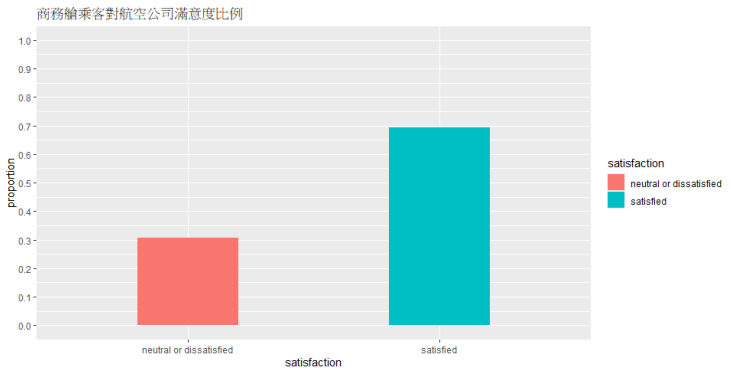
## 探索性資料分析

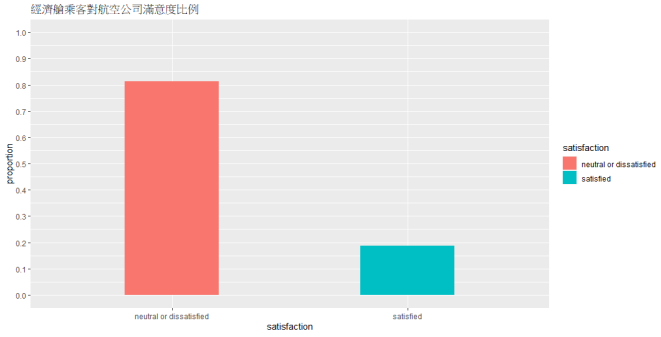
乘客對航空公司滿意度比例：(藍色為滿意比例，紅色為中立或不滿意的比例)

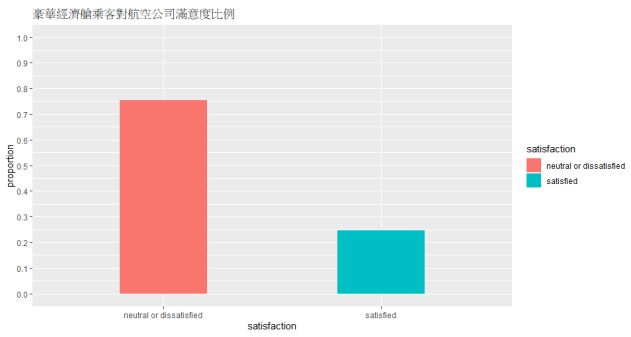



由上圖可以看到乘客對航空公司滿意的比例不高且不到一半，此顯示航空公司可能存在某些問題需要解決。

不同艙等下，乘客對航空公司滿意度比例：

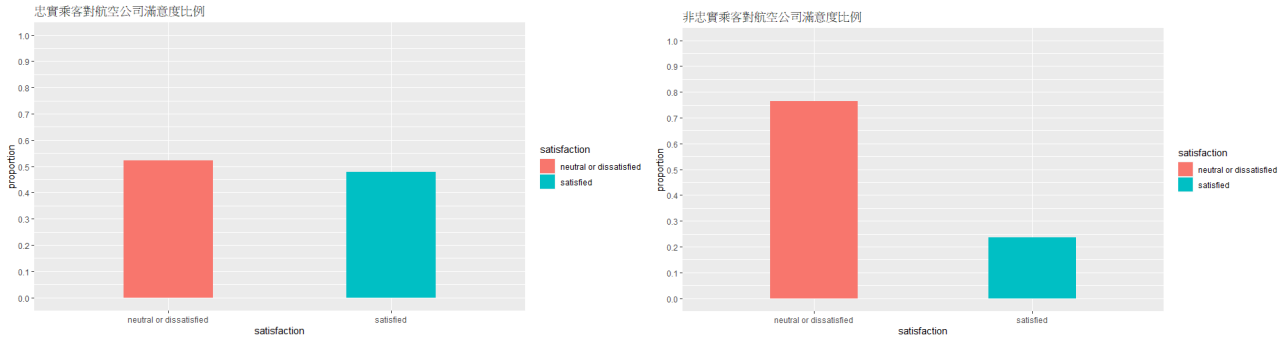


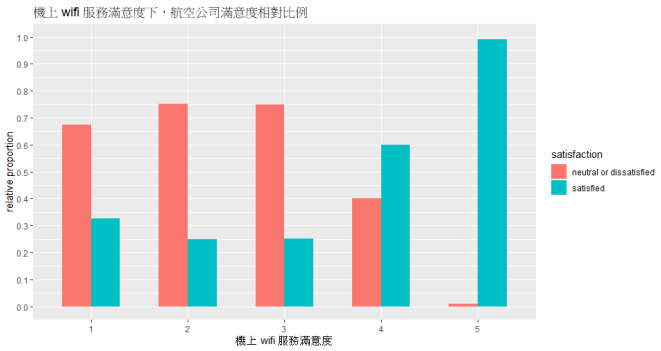
由上到下由左到右分別為商務艙乘客、經濟艙乘客和豪華經濟艙乘客對航空公司滿意度比例，因商務艙乘客對航空公司滿意比例接近 0.7，表示航空公司對於乘坐商務艙的乘客提供不錯的服務和環境，但經濟艙或豪華經濟艙的乘客對航空公司滿意比例明顯很低，故航空公司可能需要針對該客群提供的服務或環境做出改善。

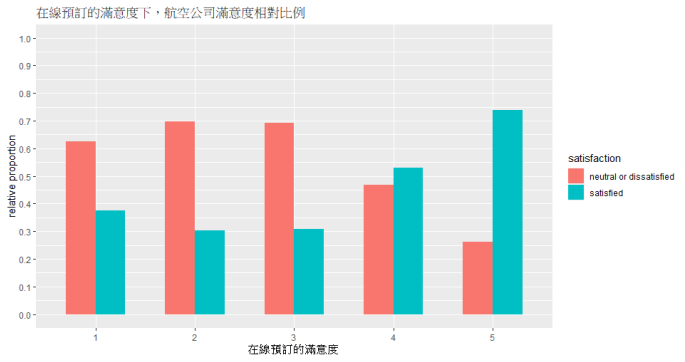
不同乘客類型下，乘客對航空公司滿意度比例：

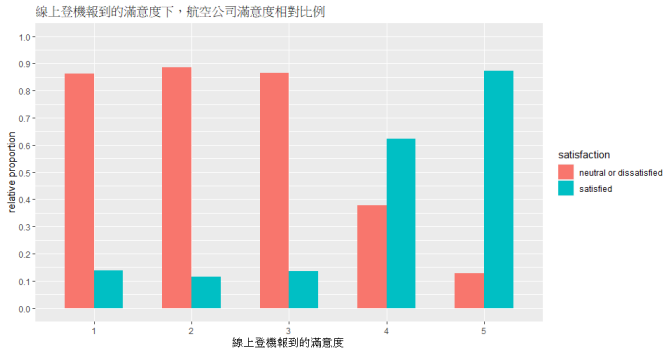



左圖中忠實乘客滿意比例並沒有太高甚至比中立或不滿意的比例還低；右圖中非忠實乘客滿意比例比中立或不滿意的比例低上許多，故航空公司須想辦法提升忠實乘客滿意度，才不會造成忠實乘客流失，並且吸引非忠實乘客且提升非忠實客戶的服務體驗，幫助航空公司獲得更多忠實客群。

不同項目滿意度下，乘客對航空公司滿意度比例：

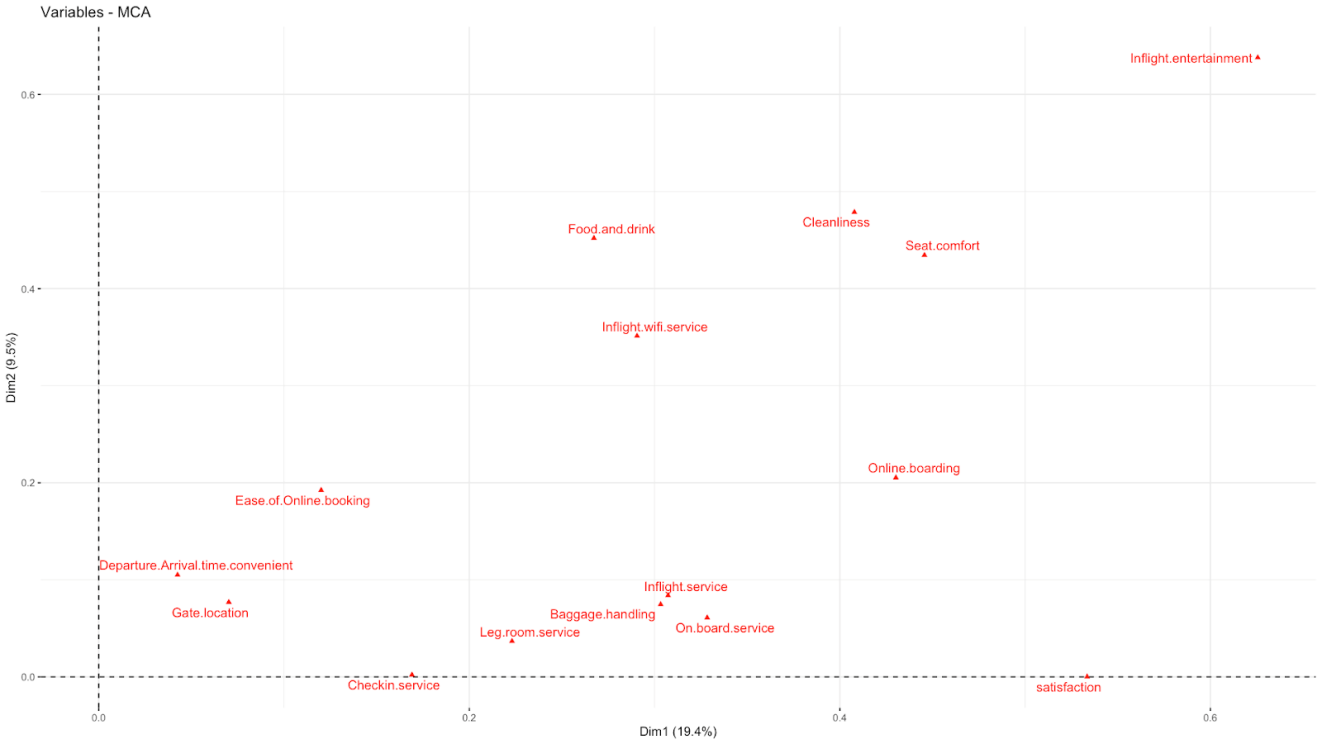




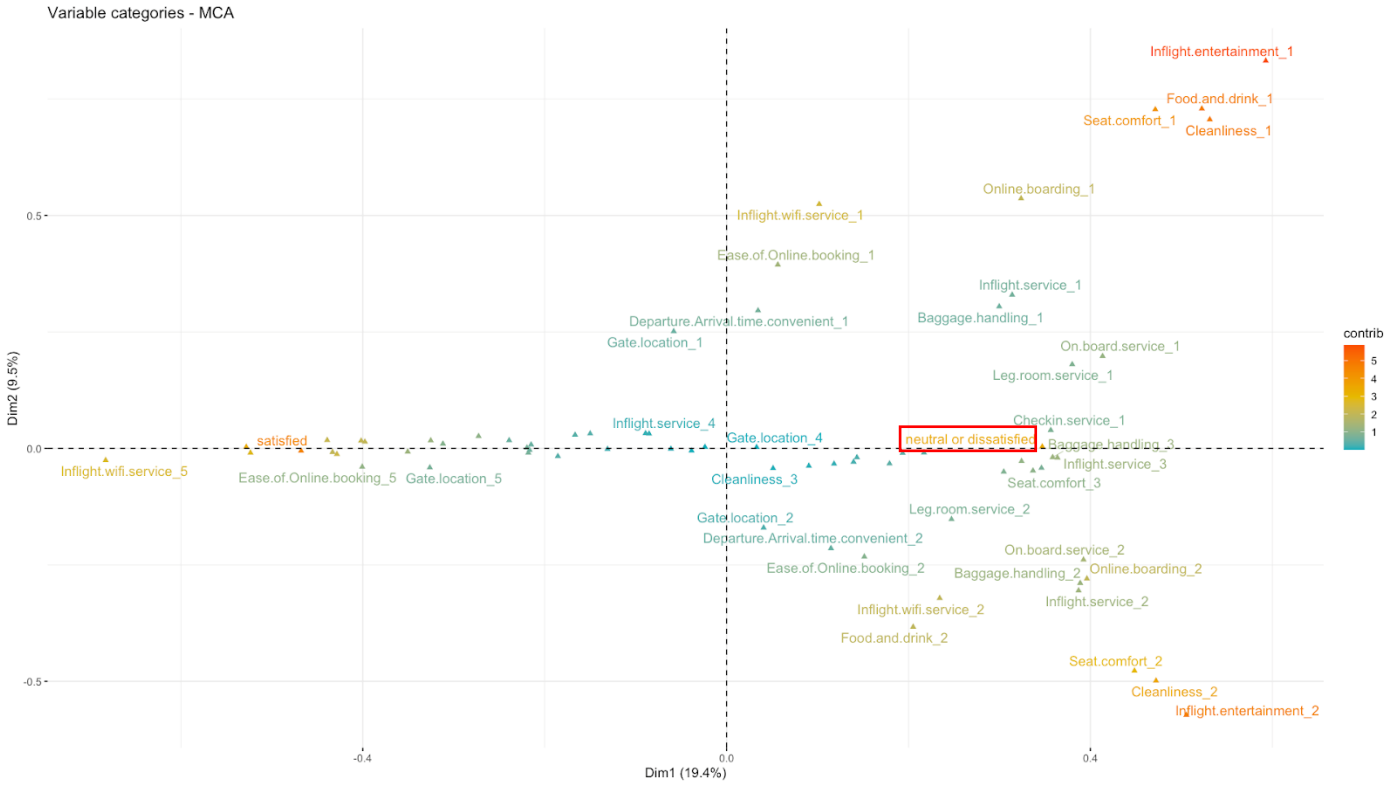

上面為在不同項目的滿意度下，乘客對航空公司滿意度的相對比例，滿意度項目由上到下由左到右分別為機上 wifi 服務滿意度、在線預訂的滿意度和線上登機報到的滿意度，可以看到當滿意度在 1 或 2 時，中立或不滿意的相對比例明顯高於滿意的相對比例，但隨著滿意度的上升滿意的相對比例有上升的現象，且當滿意度為 5 時，滿意的相對比例為 0.7 以上，故推測這三個項目的滿意度對航空公司滿意度有一定的影響。

## MCA




觀察各類別變數間的關係，哪些滿意度對乘客對飛機的整體滿意度的關係最

大。由上圖可以看到第一維度解釋比例比第二維度高出許多，故像 Online boarding、Inflight entertainment、Seat comfort 和 Cleanliness 因在第一維度離 satisfaction 較近，故推論這些項目的滿意度和乘客對航空公司滿意度有較大關係。




上圖可以看到在 neutral or dissatisfied 旁邊圍繞很多滿意度選 3 的項目，也就是選擇中間滿意度。故必須更加注意在各項滿意度選很多 3 的乘客，因其可能對航空公司的滿意度偏中立或不滿意的機率偏高，需要去了解其是否有不滿意的地方，不滿意的原因為何。

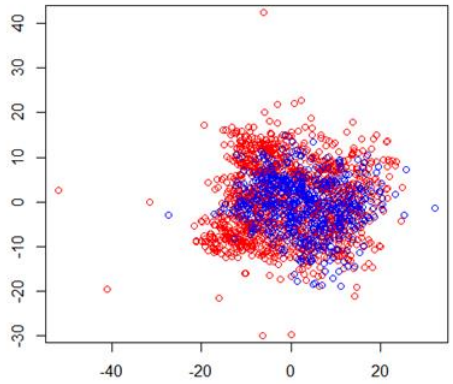
MDS

| Isotonic | Isotonic | Sammon | Sammon |
|---|---|---|---|
| Manhattan | Minkowski | Manhattan | Minkowski |
| 28.83387 | 24.18047 | 0.1464366 | 0.1931711 |

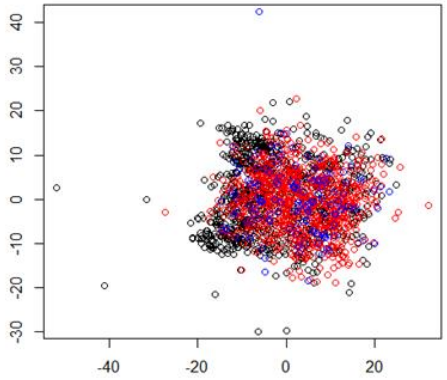
由以上表格可以發現在 Sammon mapping 之下 Manhattan distance 的 Stress 較低。接

著，我們針對乘客的飛行目的、乘客飛機上的旅行等級，將降維後的資料分別標上不同顏色並畫出圖形來做比較。

不同飛行目的

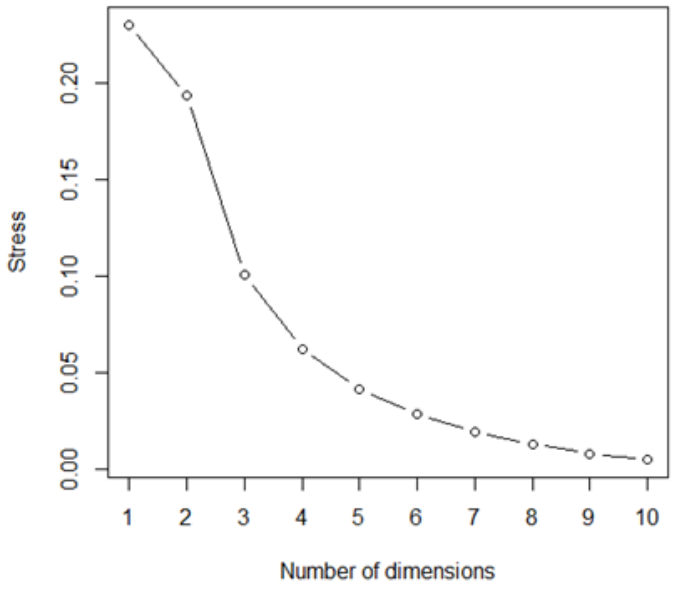



不同旅行等級




左圖中藍色點為商務旅行，紅色點為私人旅行。其中可以發現商務旅行的資料較集中，這可能代表商務旅行對於各評分項目的滿意度較一致，私人旅行的座標點較分散可能代表著有離群值。

右圖中黑色點為商務艙旅客、紅色為經濟艙旅客、藍色為豪華經濟艙旅客。圖中大多數黑點有較明顯集中於一側的趨勢，但紅點以及藍點則較難看出差異。

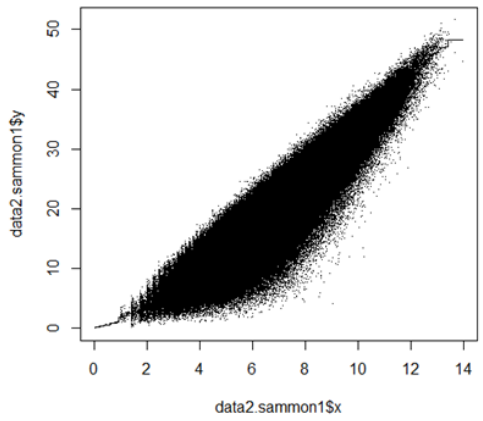



由上圖可以發現在維度=4、5 的時候出現拐角，所以接下來我們更改維度再來

執行一次 MDS。

| 維度 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| Sammon | 0.0663995 | 0.03857278 | 0.02360394 | 0.01742175 |

從上表可以發現 Sammon 在維度 4 到 5 的時候減少的幅度變少大。最終我們認為 Sammon 方法應該取維度=4。我們可以畫出 Shepard diagram 來確認我們在該維度下的擬合度好不好。



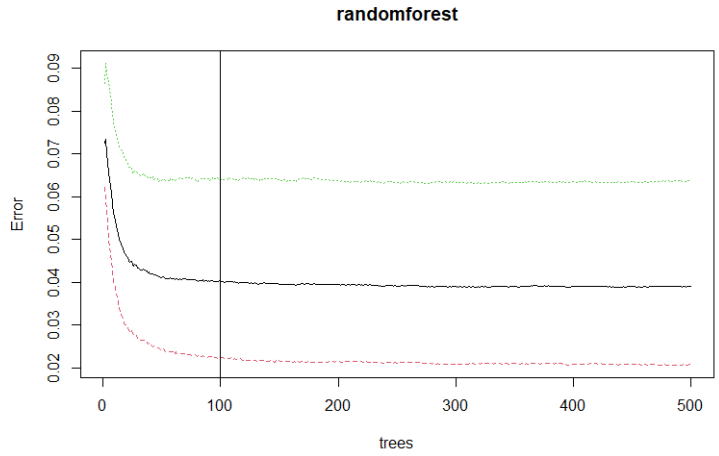

從圖中可以看出資料的擴散量仍然很大，這代表擬合不盡完美。這可能是因為我們選用的 Manhattan distance 在順序型資料上或許並不是那麼適合，若能找到更適合用來計算順序型資料的距離或是使用旋轉、鏡射、平移改變投影的方式，或許能使得 Stress 降低更多。

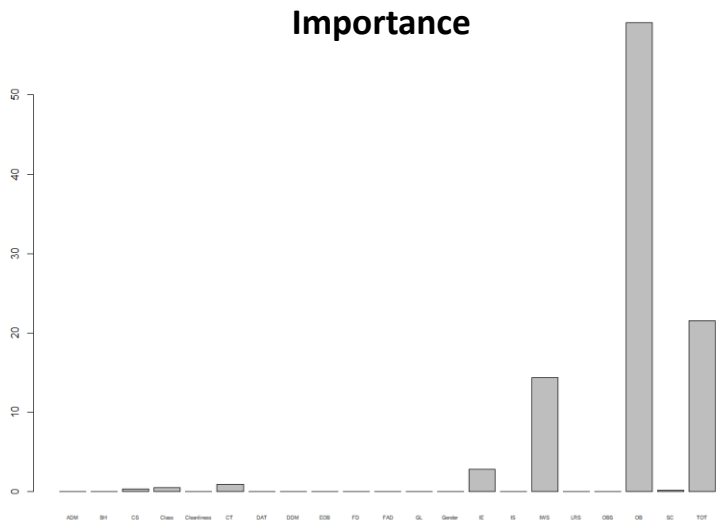
## 隨機森林

Bagging:

設置 rpart.control 的參數 minisplit=10、minbucket=3、cp=0.01。樹的數量選取是用 randomForest 函數來判斷在多少樹之下可得到比較低的 error，雖然 trees=458

時有最低的 error，但 trees=100 後趨於平穩故選擇 100。




使用 bagging 建立隨機森林模型，並利用測試集判斷是否過度擬和，train error rate：0.1190232、test error rate：0.1204375，結果相近，故沒有過度擬合的問題，所以我們進一步探討變數對整體滿意度的重要性，下方圖可知，前三高的值分別為 online boarding、type of travel、inflight wifi service。

Boosting：

選取樹的數量與上方的 bagging 相同，使用 boosting 建立隨機森林模型，對訓練集的樣本加權重，並進行迭代，誤差較小的弱分類器權重越大。下方為模型對訓練集和測試集的 error rate 和 confusion matrix。發現對訓練集和測試集預測的

error rate 分別為 0.0456355 和 0.04577656，差異不大。

Test Confusion

| | Observed Class | |
|---|---|---|
| Predicted Class | neutral or dissatisfied | satisfied |
| neutral or dissatisfied | 13214 | 713 |
| satisfied | 376 | 9560 |

Train Confusion

| | Observed Class | |
|---|---|---|
| Predicted Class | neutral or dissatisfied | satisfied |
| neutral or dissatisfied | 53523 | 2957 |
| satisfied | 1424 | 37800 |

比較 bagging 和 boosting 隨機森林模型測試集的 error rate，選擇較低錯誤率的 boosting 隨機森林作為我們最後的預測模型。

bagging 和 boosting 顯示前三名重要的預測變數皆為 Online boarding、Inflight wifi service、type of travel，其中兩個變數和探索性資料分析所找出與乘客對航空公司的整體滿意度有較大相關的變數一致。

# 五、 結論：

乘客對航空公司滿意度的比例並不高，進一步的分析發現乘坐經濟艙和豪華經濟艙的乘客對航空公司滿意度比例很低，航空公司應該針對乘坐經濟艙和豪華經濟艙的客群所提供的服務和環境做進一步的改善。

利用 MDS 做出來的結果分類效果不佳，且資料的擴散量仍然很大，代表擬合不夠好，故我們無法從中獲得資訊。

利用探索性資料分析、MCA 和隨機森林觀察到哪些項目的滿意度與乘客對航空公司滿意度有較大相關，其中三個方法皆認為乘客對機上 wifi 服務滿意度和乘客對航空公司滿意度有較大相關，探索性資料分析和隨機森林分析結果認為，線上登機報到的滿意度和乘客對航空公司滿意度有較大相關，推測飛機上的 wifi 不穩定會影響乘客乘機體驗，更進一步發現不滿意(滿意度為 1 或 2)機上 wifi 的乘客中以商務目的的乘客佔多數，故航空公司需要多加注意以商務目的來往的乘客在機上 wifi 的使用體驗，並加以改善。

由於現代人對於網路的依賴度增加，如果常因為網站的卡頓或繁瑣的報到流程導致嚴重影響使用者體驗，乘客對該公司網站的容忍度也會隨之降低，所以建議航空公司需要時常維護網站的品質，好讓使用者不會因網站的問題影響到乘客對航空公司的滿意度。

利用 boosting 隨機森林預測乘客對航空公司滿意度，其 test error rate 為 0.04577656，透過此預測模型幫助航空公司找出哪些乘客在滿意度調查表中雖是勾選滿意此家航空公司，但是預測出的結果是中立或著是不滿意，針對此類的乘客去做進一步的了解，是否是預測錯誤還是為潛在不滿意的乘客，如此一來，航空公司可以做出相對應的改進，並確實照顧到每一位乘客的需求。