**統計諮詢-鐵達尼號**

H24096037張幼澄 H24096011 吳涴棋 H24109513雷子瑩

H24091223陳彥亨H24094035向啟瑤

一、資料特性：

此份鐵達尼號乘客資料共有1309筆，有6個類別變數、2個連續變數以及2個離散變數。由於fare以及embarked的缺失值分別只有1、2筆，因此我們直接把它們刪除，刪除後資料筆數為1306筆。而home.dest由於缺失值較多，除了第2題使用到此變數，其餘問題皆不納入分析。age的部分則用迴歸插補法進行分析，插補過後的各統計量(平均數、中位數、第一與第三四分位數等等)皆與插補前相似。

表一、各變數遺失值個數及比例表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 變數名稱 | pclass, survived  name, sex  sibsp, parch | fare | embarked | age | home.dest |
| 遺失值個數（比例） | 0 (0%) | 1 (0.1%) | 2 (0.2%) | 263(20.1%) | 1. 3.1%) |

* 1. 類別變數之敘述統計

一張含有 文字, 螢幕擷取畫面, 圖表, 圓形 的圖片

自動產生的描述

圖一、存活分佈圖

一張含有 文字, 螢幕擷取畫面, 圖表, Rectangle 的圖片

自動產生的描述

圖二、性別分佈圖

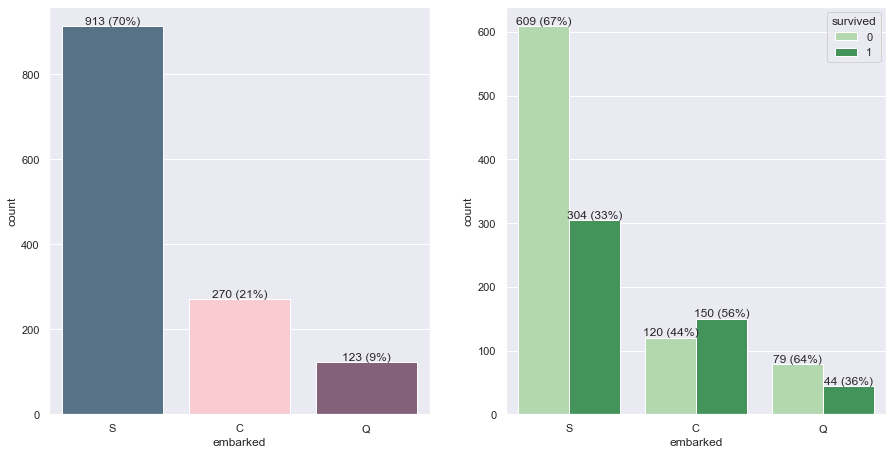
圖三、性別個別存活率

一張含有 文字, 螢幕擷取畫面, 圖表, Rectangle 的圖片

自動產生的描述

圖五、艙等個別存活率

圖四、艙等分佈圖



圖六、登船地點分佈圖

圖七、登船地點個別存活率

由此看出，這份資料中：

* 總體存活率約6成
* 船上的男性比例高於女性，但**女性存活率遠高於男性**。推測可能是與**上救生艇優先順序**有關。
* 各艙等中三等艙佔約5成，頭等艙及二等艙平分剩下5成。而三等艙的死亡率稍高於其他艙等。
* 登船地點則由 S（Southampton）佔比最高，約有7成。而從S（Southampton）登船的人死亡率稍高於其他地點登船的人。**推測登船地點與生存與否較無關係**。

二、問題討論

1. **該資料有哪些變數? 分別代表什麼意義?**

該資料共有10個變數，pclass、survived、name、sex、embarked、home.dest為 類別變數中的名義變數，sibsp、parch為類別變數中的次序變數，age、fare為連續變數各變數意義詳列於下表。

表2、各變數意義統整表

|  |  |  |
| --- | --- | --- |
| 變數 | 定義 | 值或特性 |
| pclass | 艙等 | 1：頭等艙  2：二等艙  3：三等艙 |
| survived | 是否存活 | 0：未存活  1：存活 |
| name | 姓名 | 包含姓氏、名稱及稱謂 |
| sex | 性別 | female：女性  male：男性 |
| age | 年齡 | 浮點數 |
| sibsp | 在船上的兄弟姐妹和配偶人數 | 整數 |
| parch | 在船上家族的父母和小孩人數 | 整數 |
| fare | 船票價格 | 浮點數 |
| embarked | 登船地點 | C：Cherbourg  Q：Queenstown  S：Southampton |
| home.dest | 家鄉/目的地 | 文字 |

1. **請問這組資料，到哪一個目的地(或家鄉)有最多人? 是甚麼地方? 針對死亡的人當中，哪一個年齡層的死亡人數最多?**

在變數home.dest中，有369個目的地（或家鄉），其中**New York, NY**為最多人的目的地（或家鄉），有64個人的目的地（或家鄉）為此。

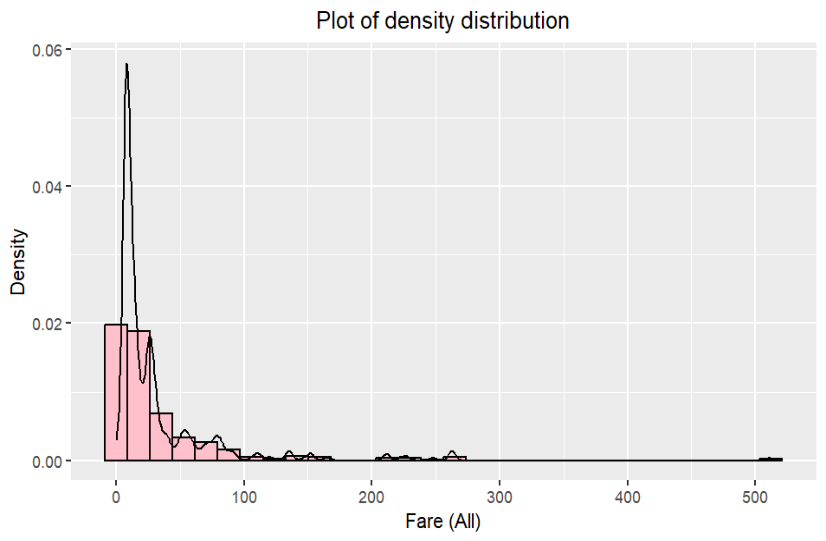
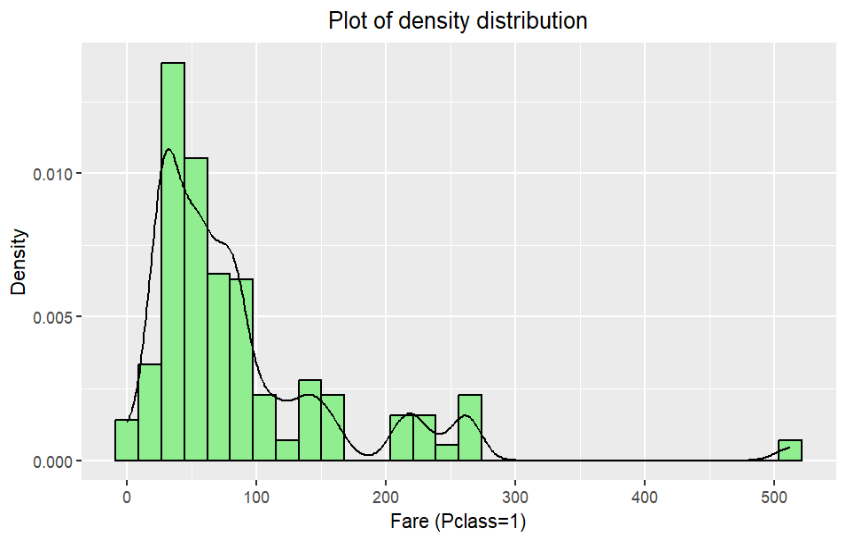
將年齡分為四個年齡層，以20歲為一個單位，分為0~20歲、21~40歲、41~60歲以及61~80歲（本資料最大年齡為80歲）。原始資料在死亡的809人當中，21~40歲的死亡人數最多，為348人(43.0%)。因有部分年齡值缺失，因此我們使用迴歸插補法補齊年齡資料，補齊年齡資料後在死亡809的人當中，21~40歲的死亡人數依然為最多，共504人(62.3%)。

表三、未存活者之年齡分佈

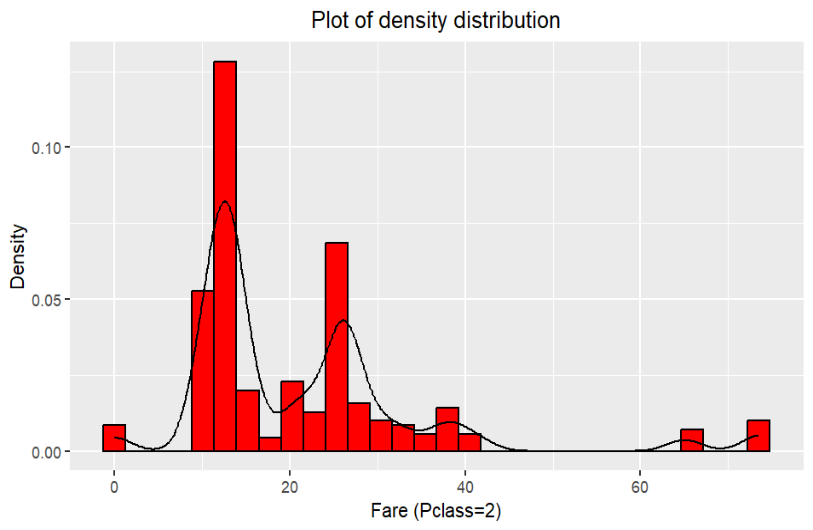
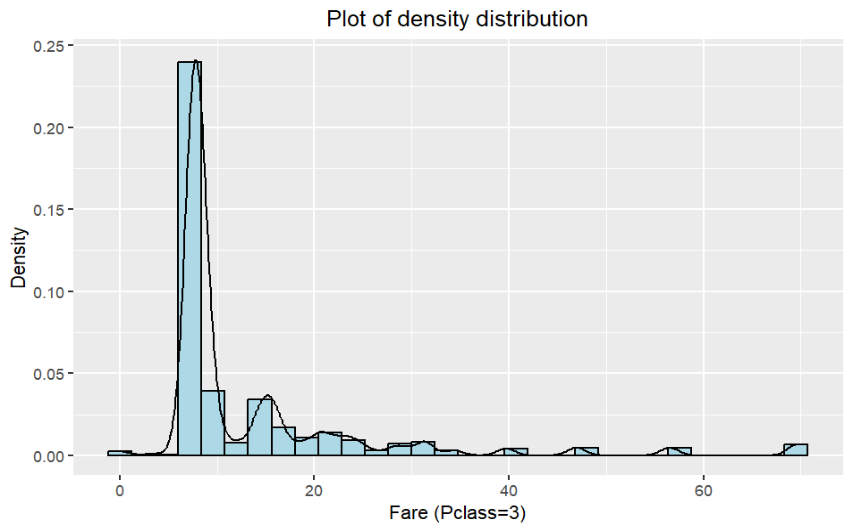
|  |  |  |
| --- | --- | --- |
|  | 原始資料中未存活者之年齡分佈 | 補齊年齡資料後未存活者之年齡分佈 |
| 年齡 | 人數（比例） | |
| 0~20歲 | 134 (16.6%) | 148 (18.3%) |
| 21~40歲 | 348 (43.0%) | 504 (62.3%) |
| 41~60歲 | 112 (13.8%) | 132 (16.3%) |
| 61~80歲 | 25 (3.1%) | 25 (3.1%) |
| 年齡缺失 | 190 (23.5%) | NA |

1. **您覺得票價(fare)在頭等艙/二等艙/三等艙價格分布都一樣嗎?**

票價(fare)整體為右偏分布（圖五），其中票價為0的我們猜測為員工或公關票，因此我們保留票價為0的資料。由頭等艙、二等艙及三等艙票價之敘述性統計量（表4）可知三者皆為右偏分布且為高狹峰，並且頭等艙的票價大致分布在0~300元內，而三等艙的票價集中於30元以下。首先我們先觀察三者之機率密度圖（圖六），我們認為三等艙與其他兩艙等之價格分布應該不相同，接著再使用Kolmogorov-Smirnov Test進行檢定，三個檢定結果皆為顯著（表五），因此我們可以知道這三個艙等之間存在顯著的統計差異，並且三個艙等不為相同分佈。

****Kolmogorov-Smirnov Test檢定的假設：

圖八、票價機率密度圖，整體(左)頭等艙(右)

****

圖八、票價機率密度圖，二等艙(左)三等艙(右)

表4、頭等艙/二等艙/三等艙票價分布之敘述性統計量

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 平均數 | 標準差 | 第一四分位數 | 第三四分位數 | 中位數 |
| 頭等艙 | 87.51 | 80.45 | 30.70 | 107.66 | 60.00 |
| 二等艙 | 21.18 | 13.61 | 13.00 | 26.00 | 15.05 |
| 三等艙 | 13.30 | 11.49 | 7.75 | 15.25 | 8.05 |

表5、各艙等之Kolmogorov-Smirnov Test結果表

|  |  |  |
| --- | --- | --- |
|  | 檢定統計量 | p-value |
| 頭等艙 v.s. 二等艙 | 0.737 | <0.001 |
| 頭等艙 v.s. 三等艙 | 0.883 | <0.001 |
| 二等艙 v.s 三等艙 | 0.651 | <0.001 |

1. **如果是你們組員不幸搭上這艘船，請你用迴歸預測大家的存活率!**

將pclass、sex、age、fare、sibsp、parch、embarked與survived進行羅吉斯迴歸分析，並使用Stepwise with both direction 方法選擇變數，其中fare、parch因AIC值較小，因此我們將其刪除。將pclass、sex、age、embarked與survived重新進行羅吉斯迴歸分析，得到方程式如下：

表6、羅吉斯迴歸模型評估表

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity |
| 羅吉斯迴歸 | 0.799 | 0.860 | 0.703 |

表7、各組員資料與其迴歸預測存活率

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | pclass | sex | age | sibsp | embarked | survived rate (%) | survived |
| 張幼澄 | 3 | Female | 22 | 1 | Q | 90.5 | 1 |
| 吳涴棋 | 1 | Female | 21 | 3 | Q | 99.0 | 1 |
| 向啟瑤 | 1 | Male | 22 | 4 | S | 87.7 | 1 |
| 陳彥亨 | 2 | Male | 21 | 1 | S | 67.3 | 1 |
| 雷子瑩 | 2 | Female | 21 | 1 | C | 96.5 | 1 |

\*存活率>0.5視為存活

1. **如果是你們組員不幸搭上這艘船，請你用機器學習方式預測大家的存活率!**

使用**決策樹**與**隨機森林**兩個機器學習的方式來進行預估並選出較優的模型。

1. 決策樹

首先先將所有變數放入模型內預測，此時的準確度為0.816，可以發現變數sibsp、parch、embarked 對於模型並不是那麼重要。因此我們將三個變數移出模型，將剩下四個變數放入模型後可以得出一個更好的模型，準確度為0.840。

1. 隨機森林

首先先將所有變數放入模型內預測，此時的準確度為0.805，可以發現變數sibsp、parch、embarked 對於模型並不是那麼重要，因此我們將三個變數移出模型。將剩下四個變數放入模型並且將參數max\_depth 設為6後，可以得出一個更好的模型，準確度為0.821。

將兩種方法所得出的Accuracy、Precision、Recall、F1 Score 綜合比較。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| 決策樹 | 0.840 | 0.711 | 0.808 | 0.756 |
| 隨機森林 | 0.821 | 0.890 | 0.820 | 0.861 |

因為隨機森林算出的F1 Score數值較決策樹的高，因此我們決定以**隨機森林**作為接下來的預測模型。

**表8、各組員資料與其預測存活率 (隨機森林)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | pclass | sex | age | fare | 預測存活率(%) | survived |
| 張幼澄 | 3 | Female | 22 | 20 | 71.5 | 1 |
| 吳涴棋 | 1 | Female | 21 | 126 | 98.9 | 1 |
| 向啟瑤 | 1 | Male | 22 | 120 | 28.5 | 0 |
| 陳彥亨 | 2 | Male | 21 | 100 | 11.7 | 0 |
| 雷子瑩 | 2 | Female | 21 | 23 | 88.2 | 1 |

綜合第4、5題的結果，我們發現小組員的預測存活率排名皆為**吳涴棋> 雷子瑩> 張幼澄> 向啟瑤> 陳彥亨**。但準確度卻比想像中的低，或許有更好的模型，另外也需要再更進一步做交叉驗證避免過於擬合。

1. **請每個組員寫一小段這次分析的心得，並點出負責該報告的哪一部分。**

**張幼澄：**

這次報告我負責畫**敘述統計中的圖、第三題的畫圖與檢定、機器學習中的隨機森林**。雖然這次的資料並沒有上次的那麼難處理，但還是感覺比上次的稍為困難一點，理由是太不熟悉機器學習，上次碰到機器學習已經是大二時的資料科學導論，很多都忘記了，也因為整組都沒有人修過機器學習，就需要自己從頭學。但整體而言還是蠻有趣的，大家也比較知道自己應該怎麼做分析，一起討論解決方法等等的。網路上也有看到有人使用Name當中的稱謂來放入變數中，有時稱謂代表了一個人的社會地位，若是時間允許，我認為將這個考慮進分析方法也是一個不錯的選擇。

總而言之，雖然這是一筆大家剛開始學習資料分析或是學習python時常接觸到的資料，但每次分析都會有不同的體驗，也有需要向其他人學習的地方。

**吳涴棋：**

在本次的報告中，我負責對**年齡資料的迴歸插補法以及採用羅吉斯迴歸分析組員存活率**。

這次用鐵達尼號的資料來對組員進行存活分析，雖然在各個課程中很常使用這組資料，但這還是第一次帶入自身資料並進行預測，我覺得很有趣；這次也是我們組第一次找方法處理年齡的missing value，經過討論後決定利用迴歸的方式插補資料，雖然程式不會很複雜，但學習到了一個新的經驗，令我覺得很滿足。總而言之，就目前來說，我很享受此種學習方法及過程。

**向啟瑤：**

我這次成功用R畫出了**長條圖**，大概花了半小時，但澄澄做的圖比我好看很多，所以我的圖就被刪掉了。。。@@ 我要多跟他學習。這次的資料分析起來，我覺得比第一次作業來得有趣，可以從分析完的敘述統計裡知道存活率與其他變數的相關性，也負責**總結前三題的結論**。而最後的機器學習分析的地方，雖然大家都沒有修過相關的課程，但沒關係，眾人一條心，黃土變成金，一起突破難關，主要是我的夥伴們非常厲害，我負責**心態上的穩定**。言而總之，大家還是很順利的完成這次討論，效率感覺有比上次還高！讚讚

**陳彥亨**：

這次報告我負責畫**敘述統計中的圖、機器學習中的決策樹**。鐵達尼號資料其實是個常見的問題，但這次的題目與以往面臨到的都不太一樣，像原本有許多票價0元的案例，經過大家討論後發現這些案例極有可能是船上的員工；又或是男性的生存比例為何遠低於女性等等；甚至還有讓我們自己去預測小組間成員的存活率的問題，算是有種親臨其境的感覺。這次也用了R去做機器學習的決策樹（decision tree)的部分，跟以往用python 稍微有點不太一樣，但一樣也能求出需要的數據，也在比較之下發現隨機森林的模型能更好的預測這次的問題。

總而言之，這次的問題十分有趣，也在過程中嘗試了很多之前並未試過的程式碼，學習到許多！

**雷子瑩**：

這次報告我負責了**問題討論的第一題與第二題**。這次的資料是很有名的鐵達尼號，我是第一次接觸這份資料，這份資料比上次的複雜許多，題目也更難一些。

這次有用迴歸模型與機器學習的方法預測存活率，雖然我的程式能力不足沒能在這部份幫上忙，不過我覺得輸入不同的條件就可以算自己的存活率很有趣，像是我們認為會先讓女生與小孩逃離，所以男性的死亡率相對就高，我負責的第二題中有做到死亡人數中哪個年齡層死亡最多，根據結果來看可以看到20歲到40歲的死亡人數最多，我認為應該也是因為會先讓老人與小孩逃難，所以青壯年的死亡比例才會最高。

透過這份作業將很多所學的應用出來，且由實際例子來分析比起課本上的範例來的有趣多了！