Final Project

Titanic Machine Learning from Disaster

蔡儀萱†  
 統計111  
 國立成功大學  
 H24076150

FirstName Surname  
 Department Name  
 Institution/University Name

林少穎  
 統計111  
 國立成功大學  
 H24071273

背景與動機

一張含有 室內, 路面 的圖片

自動產生的描述

第一次接觸到這份資料是老師和助教在上課的過程中皆有提到。此外，我鐵達尼號是一部經典的老電影，容易讓人引起共鳴，並且使人感興趣，而且可我以利用機器學習來預測鐵達尼號發生意外時，船上人員與乘客的存活與否，對我當時的我們來說，是一件很新奇的事情，所以決定用此來作為我們期末報告的我主題。

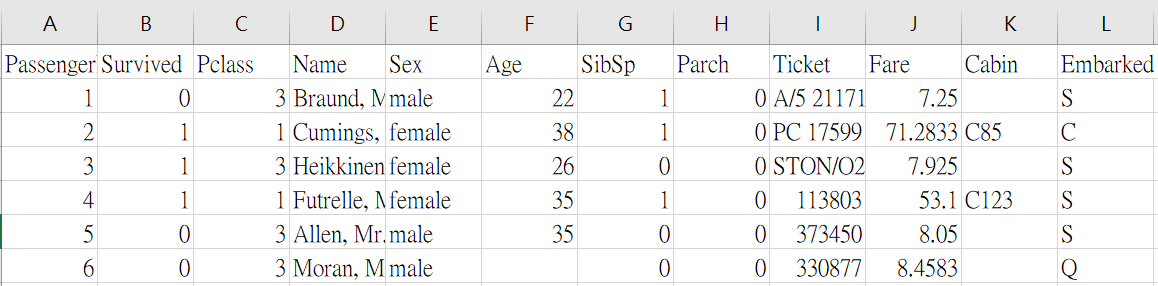
問題敘述

依網上給定的檔案，預測鐵達尼號發生意外時，乘客的存活與否。

潛在應用與價值

資料分析的結果可以驗證社會階層不平等的黑暗面，有權有勢者可能較早被通知發生船難，使其能有較充裕的時間逃難，因此生存率比較高，此外，在過去的社會中，普遍瀰漫著大男人主義，舊傳統的思想而言，男性必須承擔起較多的責任，去拯救老弱婦孺，倘若他們沒有達到社會賦予的期望，自行逃難，極有可能受到社會輿論的譴責，所以男性的生存比例較低，年齡較小者與女性之存活率較高。

資料描述



[A] Passenger - 將乘客的資料編號。

[B] Survived - 預測該欄位，1為存活，0為罹難。

[C] Pclass - 艙等，1為最好，2次之，3則是最差。

[D] Name - 乘客姓名。

[E] Sex - 乘客性別，male為男性，female為女性。

[F] Age - 乘客年齡。

[G] SibSp - 乘客與其兄弟姊妹和夫妻陪同的總人數。

[H] Parch - 乘客與其父母和小孩陪同的總人數。

[I] Ticket - 乘客所持的票根編號。

[J] Fare - 票價。

[K] Cabin - 船艙編號。

[L] Embarked - 乘客登船之港口，C = Cherbourg，

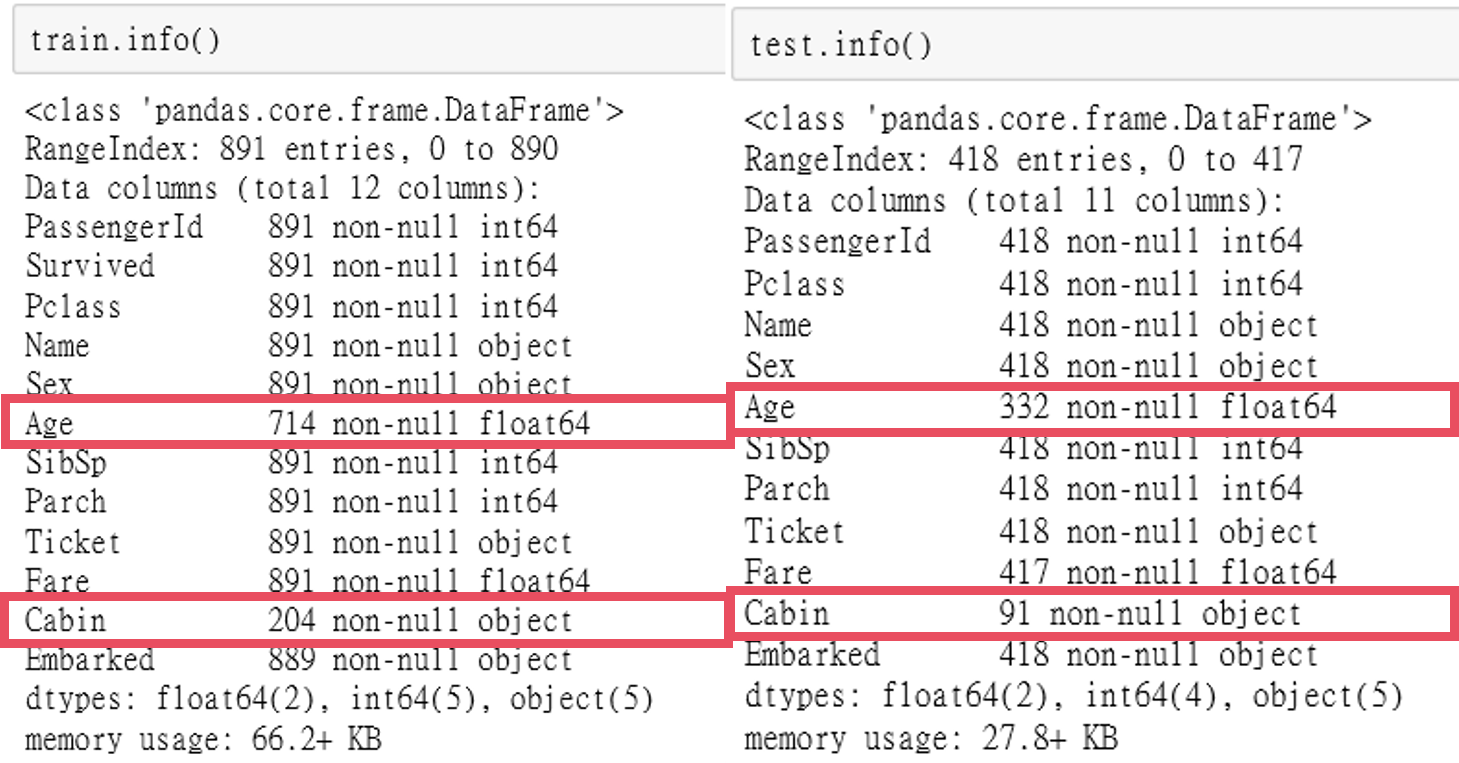
Q = Queenstown，S = Southampton。

具體作法

在上次的proposal後發現該題目已有許多人做過分析，且都有良好的結果，再經過訊問老師的結果，我們決定著重在分析資料上，檢視資料如何影響預測結果，並上網搜尋資料已做參考。

資料分析

**1. 檢查缺失值**



可以發現train檔與test檔的Age和Cabin之欄位的缺失值較多。

**2 .train檔和test檔合併**

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

方便做更準確的資料分析。

**3.初步分析**

3.1乘客生存與否

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

大部分的乘客不幸罹難。

3.2艙等與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

艙等最低者有高比例罹難，相反，艙等最高者生存率最高。可以推測於最高艙等者為有權有勢之人，因此發生船難時，較早被告知，而有較充足的時間逃難。

3.3 性別與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

男性大部分罹難，女性大部分存活。可以推測在過去的傳統觀念中，男性背負著較沉重的責任，若沒有去拯救他人，可能事後會受到社會輿論的譴責，因此在幫助大家時，便錯過黃金逃難時期。

3.4 港口與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

絕大部分之人於S港口登船，同時，高比例發生意外。由上述所提及（3.1），有權有勢者較容易倖存，兩者便可推測，有相同社會地位者居住於較相近的區域，形成社會居住的差異，也從相同的港口登船，才造就罹難人數集中在同個港口的現象（S港口登船之人可能社會階級平均較低）。

3.5 年齡與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

20至40歲者較高比例死亡，年紀較小者容易存活。可推測年齡小容易獲救，青壯年之人因較有能力而選擇幫助尚未獲救的老弱婦孺，錯過可逃難的契機而不幸罹難。

3.6 票價與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

票價高低與存活率呈正相關，前面有提及（3.1），有權有勢者容易存活，通常此種人財力也較為雄厚，所以可以花比較多的錢購買相對艙等高的座位，使得票價愈高存活率愈高。

3.7 父母+小孩的總人數與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

隻身一人登船者存活率較低。

3.8 手足+夫妻的總人數與生存與否的關係

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

隻身一人登船者存活率較低。

※結論

[1] 艙等的生存率 - 1 ＞ 2 ＞ 3。

[2] 性別的生存率 - 女 ＞＞ 男。

[3] 出發港口的生存率 - S ＞ C ＞ Q。

[4] 年齡的生存率 - 年齡小 ＞ 年齡大。

[5] 票價的生存率 - 票價高 ＞ 票價低。

[6] 父母+小孩陪同者的生存率 - 有 ＞ 無。

[7] 手足+夫妻陪同者的生存率 - 有 ＞ 無。

由 [6]、[7] 可看出是否有家屬陪同與生存有明顯地正相關，因此將父母+小孩與手足+夫妻的總人數相加，列為一新欄位（Family\_size），以增加預測準確率。

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

**4. 姓名之欄位**

可將姓名的稱謂提出，以增加判斷每位乘客的年齡與性別。

4.1 提出過程

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Braund, Mr. Owen Harris | → | Mr. Owen Harris | → | Mr. |
| Nasser, Mrs. Nicholas  (Adele Achem) | Mrs. Nicholas  (Adele Achem) | Mrs. |
| Heikkinen, Miss. Laina | Miss. Laina | Miss |

4.2 提出結果

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

提出結果為共有18種稱謂。

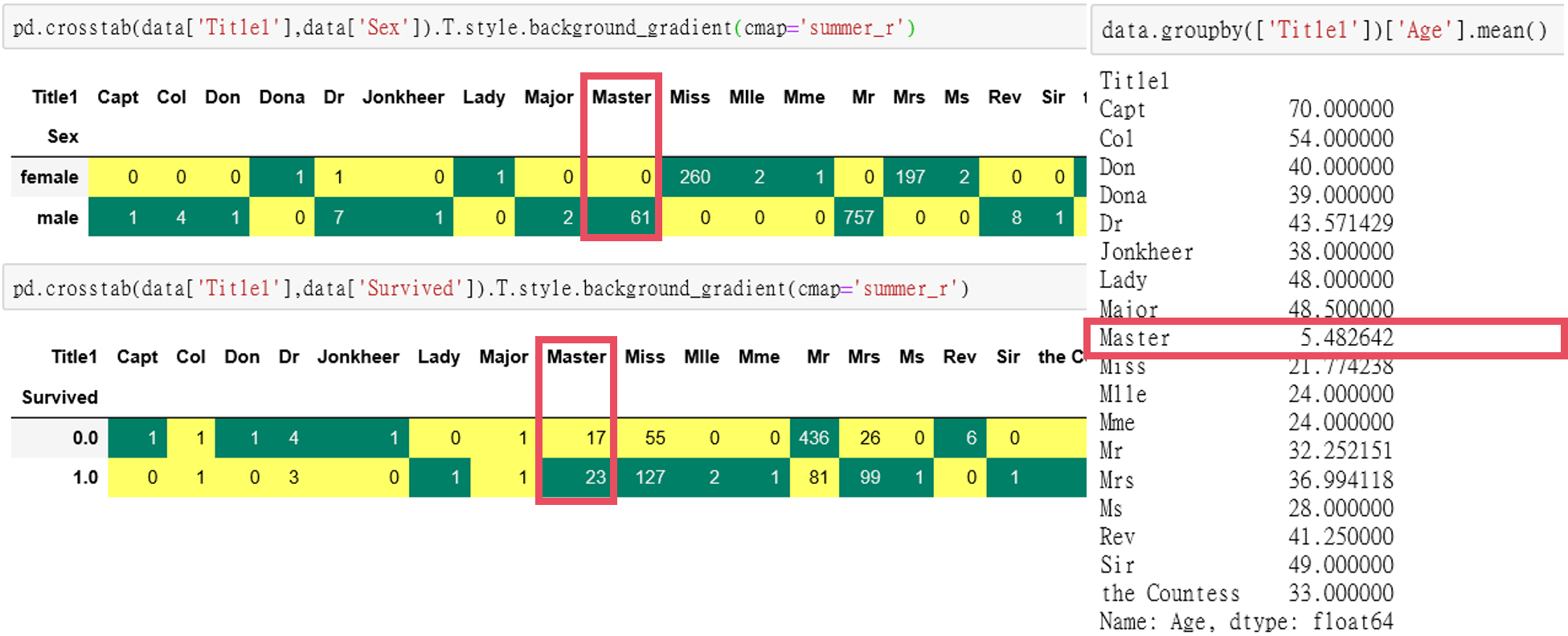
4.3 簡化並合併稱謂

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

仔細觀察後會發現有些稱謂出現的次數過少，若分過細可能將有overfitting的效果，因此將一些稱謂合併，如：lady併成Miss，最後留下的四個稱謂為Master、Miss、Mr. 、Mrs.。

4.4舉Master為例



平均年齡只有5歲，皆為男性，生存率達6成，呼應年齡愈小，存活率愈高。

**5. 填補缺失值**

5.1 船票編號



船票編號為英文加數字組成，由於英文代號可能代表房間位置，故取出作為使用，數字的部分則省略，若船票僅有數字，則用X表示，從上圖亦可發現，有高比例的人集中於某個房間。

5.2 登船港口



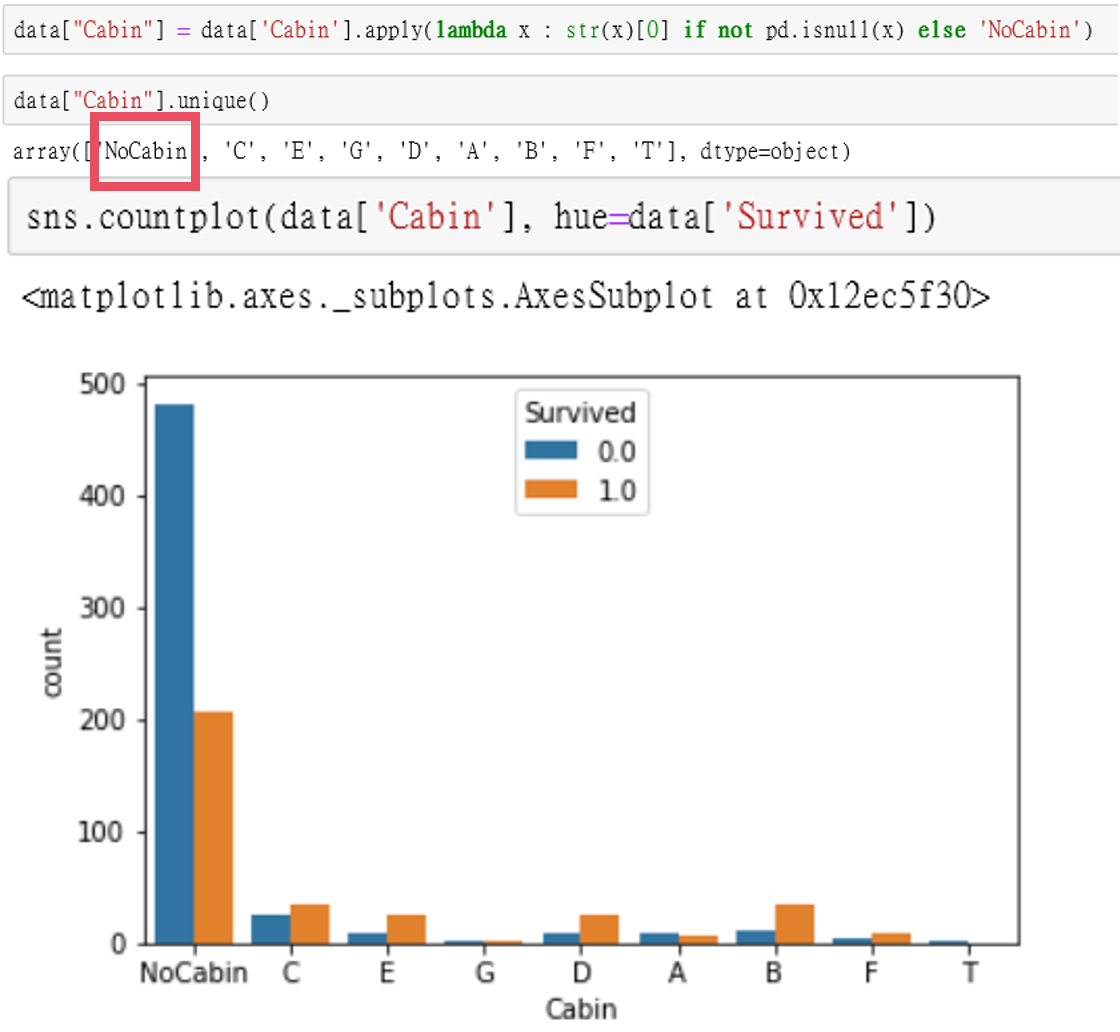
缺失值為少數，所以補上出現次數最多者（S）。

5.3 票價



將缺失值補上票價的平均值。

5.4 船艙編號



船艙編號只取字母，以免分得過細，缺失值補上no cabin，也可發現缺失值者的存活率偏低。

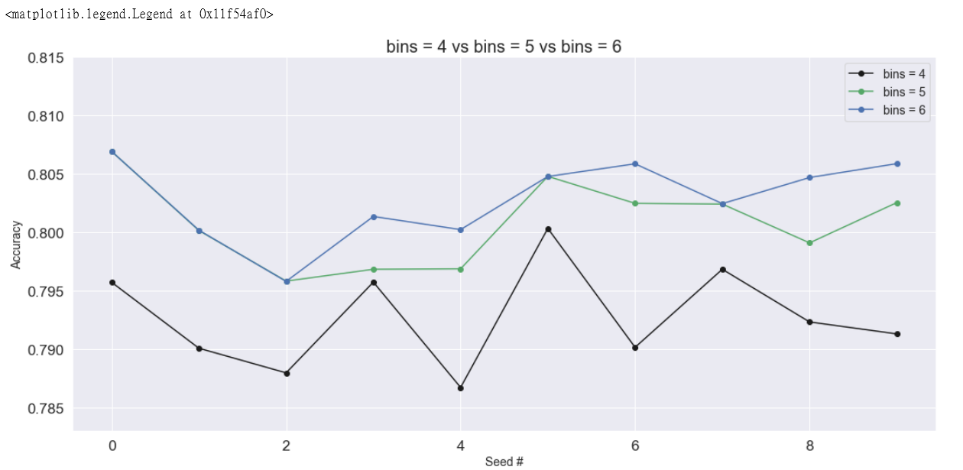
**6. 解決票價overfitting**

6.1 原因

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

船票的價位高達兩百多個且分布廣泛，而票價高低與乘客存活與否成正相關，若不將其分群，票價會被分得過細，將導致預測結果有overfitting的狀況。

6.2 解決方式與分析結果

將票價分成4 ~ 6個區間，再進行預測並計算準確度，結果可見分將票價分成6群的準確度較高。

**7. 相同票根之間的生存關係**

7.1 票根

一張含有 螢幕擷取畫面 的圖片

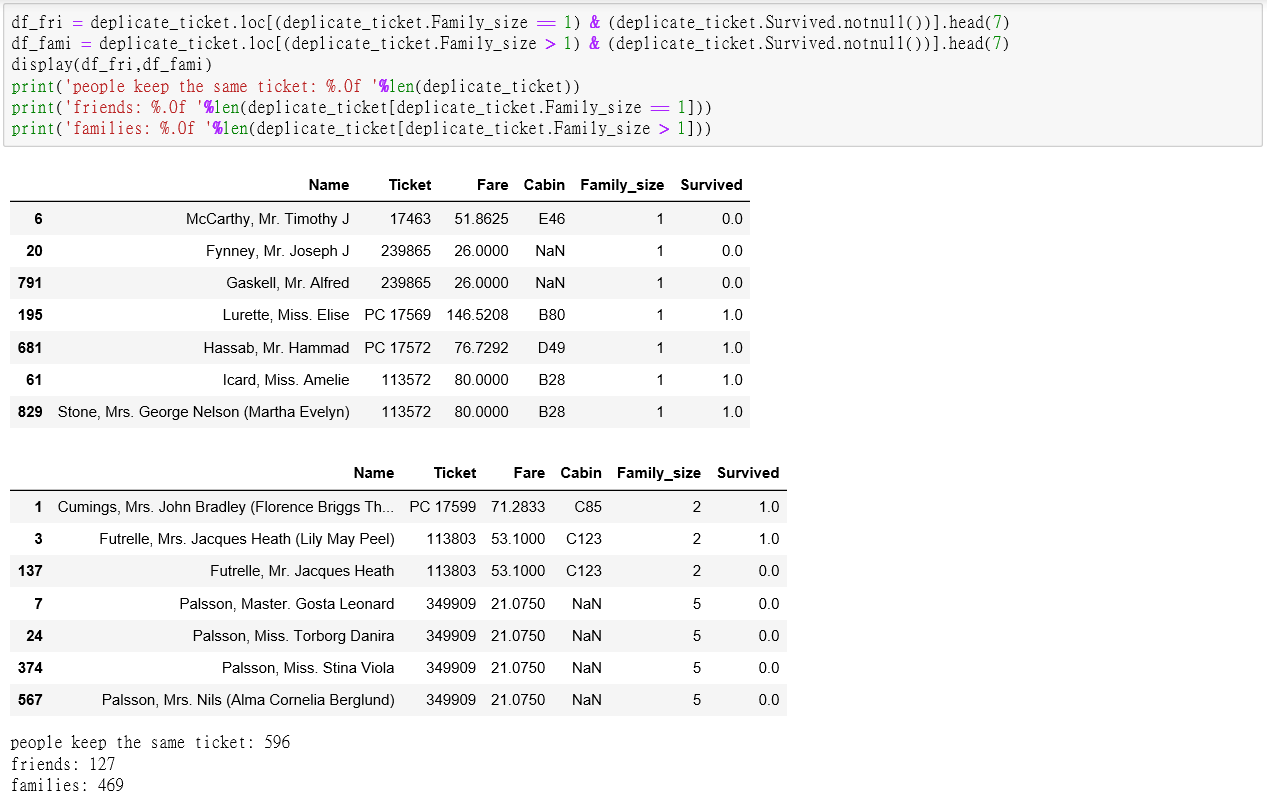
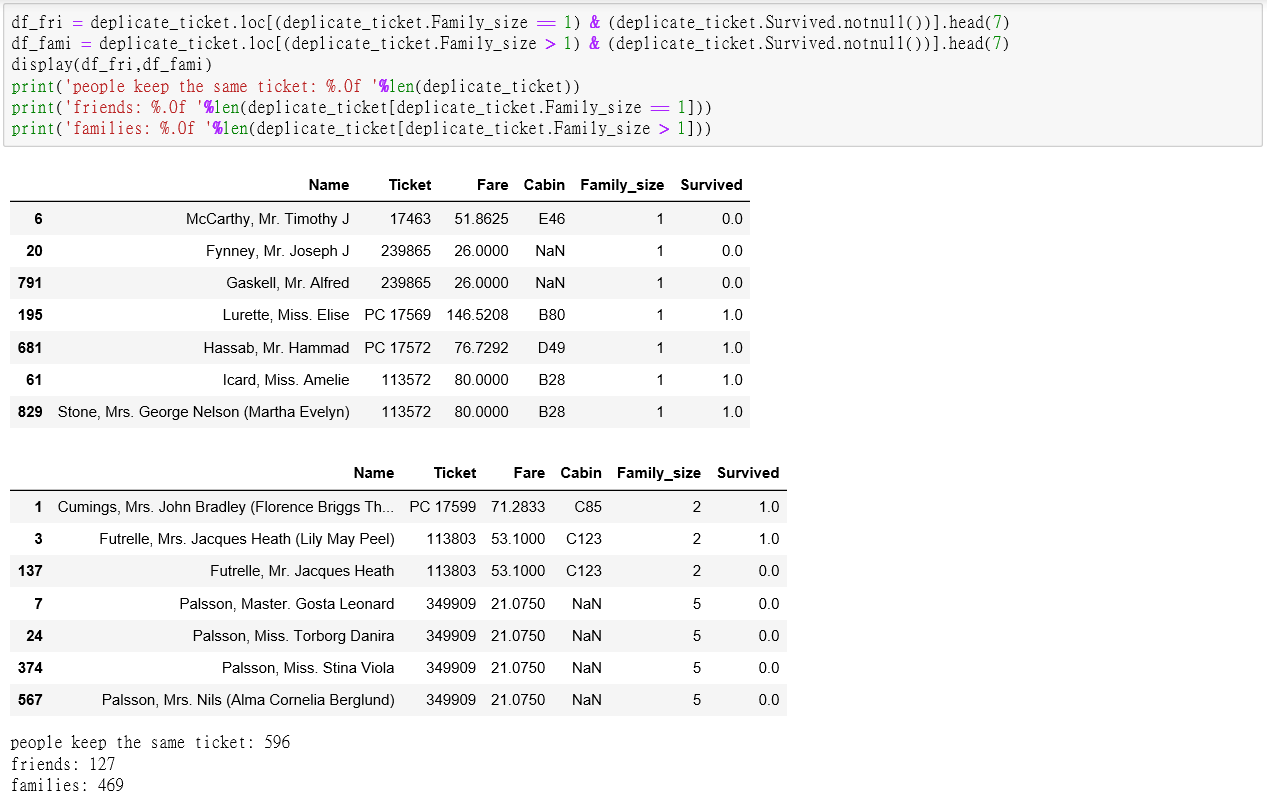
自動產生的描述

分析Ticket之欄位，可見其count為891，unique為681，因此推斷有人持相同票根。

一張含有 螢幕擷取畫面 的圖片

自動產生的描述7.2 計算家庭人數

+1 代表自己



若Family\_size為1，則表示該乘客與朋友同行，大於1則為與親屬同行，而其中是家庭出遊者高達75%，因此相伴而行的人極有可能同生同死，我們以這個想法來幫助提高準確率。

**8.填補年齡缺失值**

8.1 艙等與年齡缺失一張含有 螢幕擷取畫面 的圖片

自動產生的描述

由於年齡的缺失值高達所有資料的20%，而艙等為3的缺失值佔了大部分，如上圖所示。所以我們採用艙等為1和2的年齡資料做填補。

一張含有 螢幕擷取畫面 的圖片

自動產生的描述8.2 年齡與存活與否

由上圖可見，小於16歲者存活率極高，因此我們將其視為重要的資料以填補我缺失值。

一張含有 螢幕擷取畫面 的圖片

自動產生的描述8.3 填補年齡缺失值的方式

依姓名稱謂做為填補年齡缺失值的依據，將各稱謂的缺失值填上其中位數。

結論與心得

這次的 Final Project 我們選擇的題目是「預測鐵達尼號上乘客的生存與否」，因為這個題目被很多人做過，大家也都有很好的預測結果，再與老師討論之後，我們決定將重點擺在分析資料，去看哪些數據會跟乘客的生存與否有什麼關連。

一般我們想到的，可能就是票價、艙等，會認為比較有錢的人，可能在船難發生時，會在第一時間先被通知，較有可能存活下來。

但再仔細去分析資料，會發現性別、年齡、有沒有親友一起同行，都可能影響他們的生存。

所以我們認為，也許我們沒有很多錢財，可以靠關係、靠交換利益，來獲得更多的生存機會，但是如果我們周遭有很忠心、能共患難的親友，那也不一定會活得比較沒有價值。

我們都要好好珍惜身邊每個給予你陪伴跟支持的人，因為有他們，我們的人生才多了很多意義。

REFERENCES

[1] https://www.kaggle.com/c/titanic/overview

[2]hhhttps://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC4-1%E8%AC%9B-kaggle%E7%AB%B6%E8%B3%BD-%E9%90%B5%E9%81%94%E5%B0%BC%E8%99%9F%E7%94%9F%E5%AD%98%E9%A0%90%E6%B8%AC-%E5%89%8D16-%E6%8E%92%E5%90%8D-a8842fea7077?fbclid=IwAR23O0JrYyfgaf2J-uituqk8bZtbq6BVAvCdPMU0P64xHyVBlHHmB1HQ8X8

[3]hhhttps://medium.com/@yulongtsai/https-medium-com-yulongtsai-titanic-top3-8e64741cc11f?fbclid=IwAR2PWYLoJS8eu8L-4\_-aZyVHxW5L-kZj1D26eU95kb06zH7m-5s32ygGyhQ