HW

summary of titanic

Wang Hao Yu - RE6131032

2025-02-26

目錄

```r
# options(repos = c(CRAN = "https://cran.rstudio.com"))
# install.packages("rmarkdown")

# install.packages("tinytex")
# tinytex::install_tinytex()
```

載入套件

```r
library(ggplot2)
library(dplyr)
library(gridExtra)
```

載入資料

```r
titanic <- read.csv("titanic.csv", stringsAsFactors = TRUE)
```

基本資訊

```r
head(titanic)
```

```
  PassengerId Survived Pclass
1           1        0      3
2           2        1      1
3           3        1      3
4           4        1      1
5           5        0      3
6           6        0      3
                                                 Name    Sex Age SibSp Parch
1                             Braund, Mr. Owen Harris   male  22     1     0
2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3                              Heikkinen, Miss. Laina female  26     0     0
4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5                            Allen, Mr. William Henry   male  35     0     0
6                                    Moran, Mr. James   male  NA     0     0
            Ticket    Fare Cabin Embarked
1        A/5 21171  7.2500              S
2         PC 17599 71.2833   C85        C
3 STON/O2. 3101282  7.9250              S
4           113803 53.1000  C123        S
5           373450  8.0500              S
6           330877  8.4583              Q
```

```r
str(titanic)
```

```
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ..
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

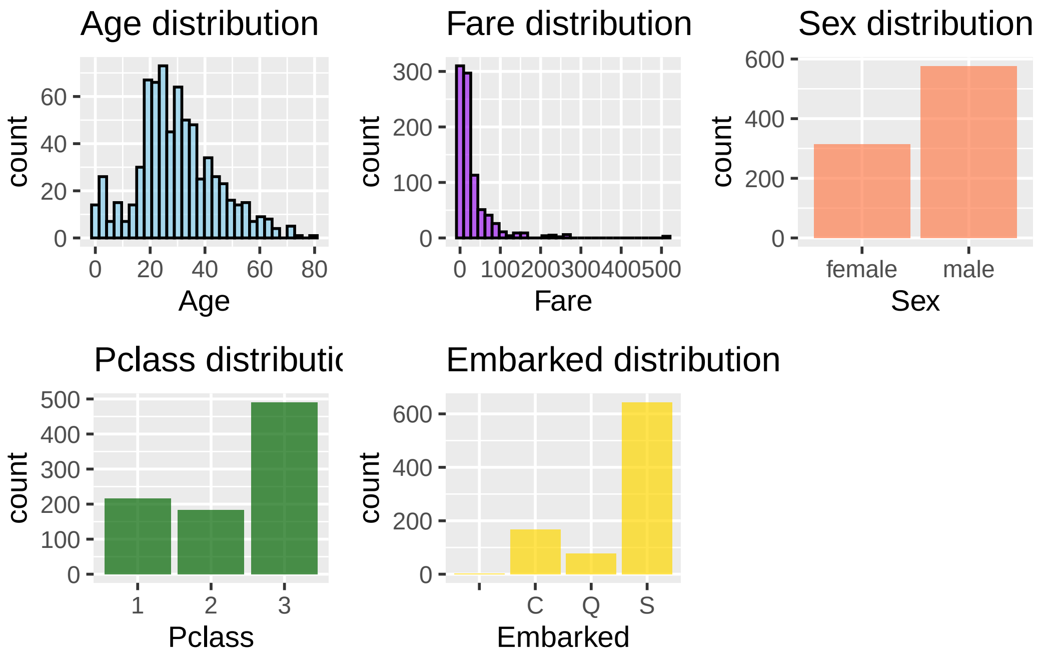
```r
summary(titanic)
```

```
  PassengerId       Survived          Pclass     
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000  
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000  
 Median :446.0   Median :0.0000   Median :3.000  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309  
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :891.0   Max.   :1.0000   Max.   :3.000  
```

```
                                   Name         Sex           Age
Abbing, Mr. Anthony                  :  1   female:314   Min.   : 0.42
Abbott, Mr. Rossmore Edward          :  1   male  :577   1st Qu.:20.12
Abbott, Mrs. Stanton (Rosa Hunt)     :  1                Median :28.00
Abelson, Mr. Samuel                  :  1                Mean   :29.70
Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
Adahl, Mr. Mauritz Nils Martin       :  1                Max.   :80.00
(Other)                              :885                NA's   :177
    SibSp           Parch             Ticket          Fare
Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
                                 (Other) :852
        Cabin     Embarked
           :687    :  2
B96 B98    :  4   C:168
C23 C25 C27:  4   Q: 77
G6         :  4   S:644
C22 C26    :  3
D          :  3
(Other)    :186
```

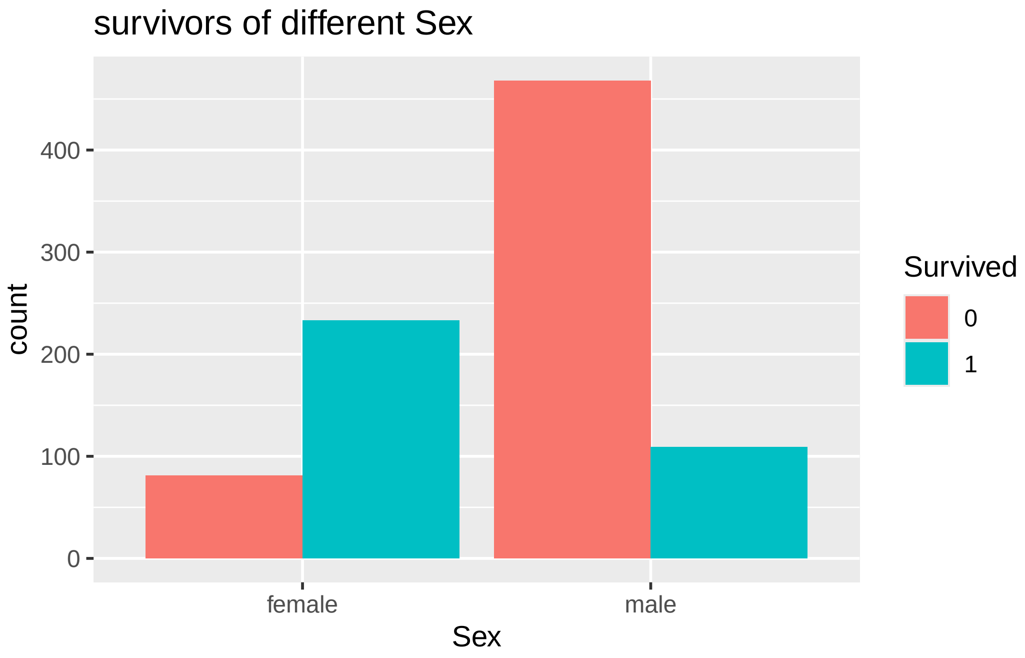
各features分布情况

```r
p1 <- ggplot(titanic, aes(x = Age)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Age distribution", x = "Age", y = "count")

p2 <- ggplot(titanic, aes(x = Fare)) +
  geom_histogram(bins = 30, fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Fare distribution", x = "Fare", y = "count")

p3 <- ggplot(titanic, aes(x = Sex)) +
  geom_bar(fill = "coral", alpha = 0.7) +
  labs(title = "Sex distribution", x = "Sex", y = "count")

p4 <- ggplot(titanic, aes(x = factor(Pclass))) +
  geom_bar(fill = "darkgreen", alpha = 0.7) +
  labs(title = "Pclass distribution", x = "Pclass", y = "count")

p5 <- ggplot(titanic, aes(x = Embarked)) +
  geom_bar(fill = "gold", alpha = 0.7) +
  labs(title = "Embarked distribution", x = "Embarked", y = "count")

grid.arrange(p1, p2, p3, p4, p5, ncol = 3)
```



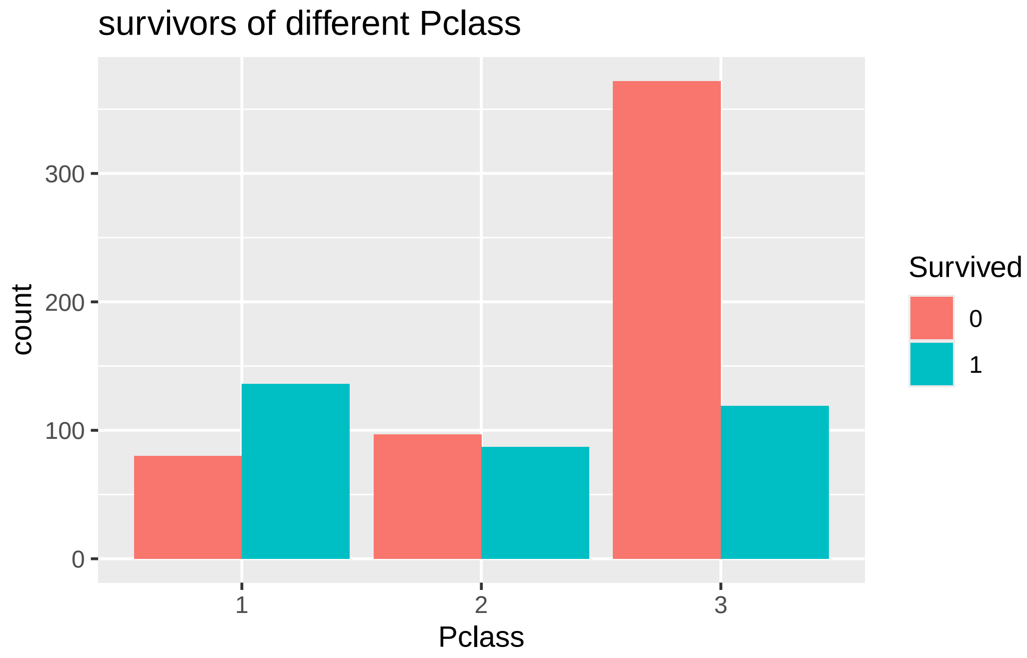
不同性别存活率

```
ggplot(titanic, aes(x = Sex, fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  labs(title = "survivors of different Sex", x = "Sex", y = "count", fill = "Survived")
```



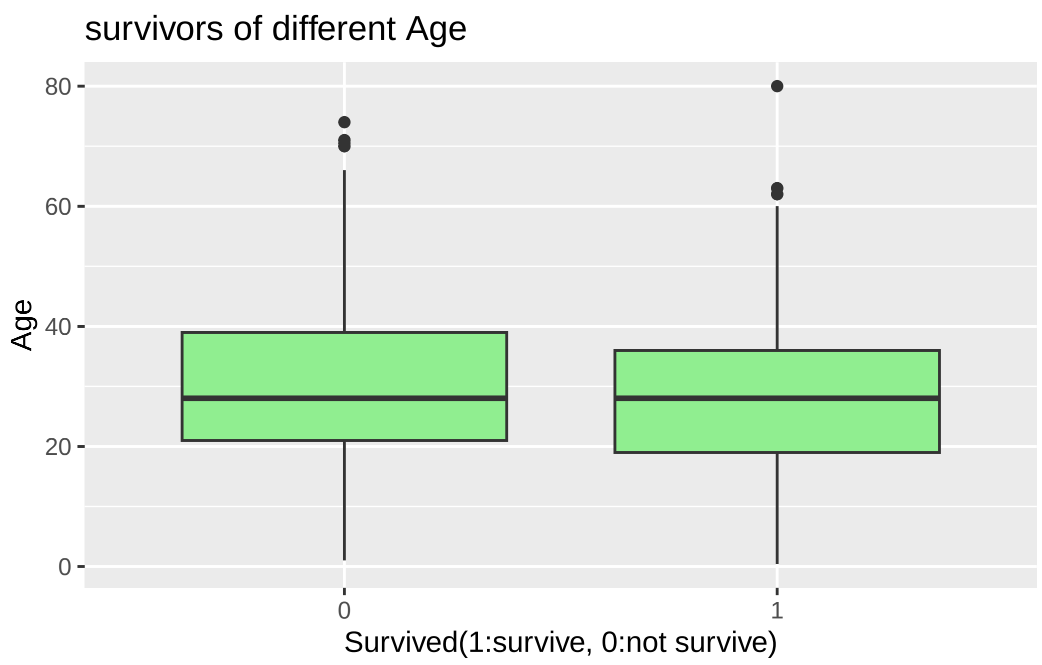
不同艙位存活率

```
# Cell 5:
ggplot(titanic, aes(x = factor(Pclass), fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  labs(title = "survivors of different Pclass", x = "Pclass", y = "count", fill = "Survived")
```



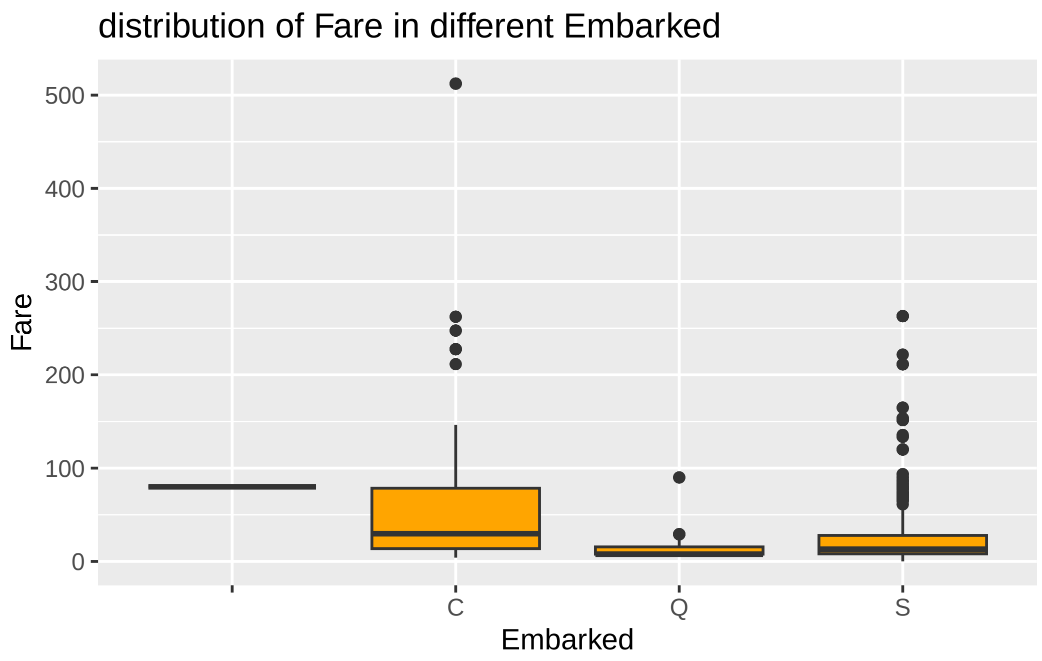
存活與年齡的關係

```
ggplot(titanic, aes(x = factor(Survived), y = Age)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "survivors of different Age", x = "Survived(1:survive, 0:not survive)", y = "Age")
```



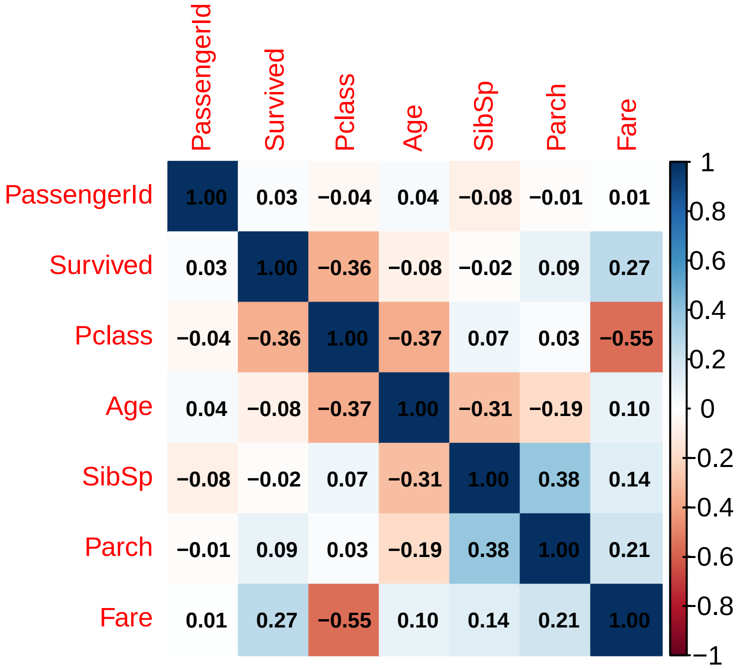
不同 Embarked 票價的分佈

```
ggplot(titanic, aes(x = Embarked, y = Fare)) +
  geom_boxplot(fill = "orange") +
  labs(title = "distribution of Fare in different Embarked", x = "Embarked", y = "Fare")
```



相關係數 heatmap

```r
num_data <- titanic %>% select_if(is.numeric)
cor_mat <- cor(num_data, use = "complete.obs")

if (!requireNamespace("corrplot", quietly = TRUE)) {
  install.packages("corrplot", repos = "https://cran.rstudio.com")
}
library(corrplot)

corrplot(cor_mat,
         method = "color",
         addCoef.col = "black",
         tl.cex = 0.8,
         number.cex = 0.7)
```


PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
PassengerId 1.00 0.03 −0.04 0.04 −0.08 −0.01 0.01
Survived 0.03 1.00 −0.36 −0.08 −0.02 0.09 0.27
Pclass −0.04 −0.36 1.00 −0.37 0.07 0.03 −0.55
Age 0.04 −0.08 −0.37 1.00 −0.31 −0.19 0.10
SibSp −0.08 −0.02 0.07 −0.31 1.00 0.38 0.14
Parch −0.01 0.09 0.03 −0.19 0.38 1.00 0.21
Fare 0.01 0.27 −0.55 0.10 0.14 0.21 1.00