Statistical consulting Homework1

S76134124 何佩勳

2025-02-25

目錄

Question: Summary report for the Titanic dataset

```
setwd("C:/Users/user/Desktop/  0223")
dat <- read.csv("titanic.csv",header = T,fileEncoding = "CP950")
str(dat)
```

```
'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       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "He
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  "" "C85" "" "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```

```
summary(dat)
```

```
  PassengerId       Survived          Pclass          Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character
```

```
 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

     Sex                 Age            SibSp           Parch
 Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
 Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                    Mean   :29.70   Mean   :0.523   Mean   :0.3816
                    3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
                    Max.   :80.00   Max.   :8.000   Max.   :6.0000
                    NA's   :177
    Ticket              Fare           Cabin              Embarked
 Length:891         Min.   :  0.00   Length:891         Length:891
 Class :character   1st Qu.:  7.91   Class :character   Class :character
 Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                    Mean   : 32.20
                    3rd Qu.: 31.00
                    Max.   :512.33
```

共有891個觀測值(891人),排除Passenger Id、Name、Ticket(票號),剩下9個變項,初步先判讀連續變項跟類別變項
連續變項:Age,Fare
類別變項:Survived,Pclass(艙等),Sex,SibSp(兄弟姊妹+夫妻),Parch(父母子女),Cabin(房間號碼),Embarked(出發港口)

連續型變項檢查及分布

```
#check missing
length(which(is.na(dat$Age)))
```

```
[1] 177
```

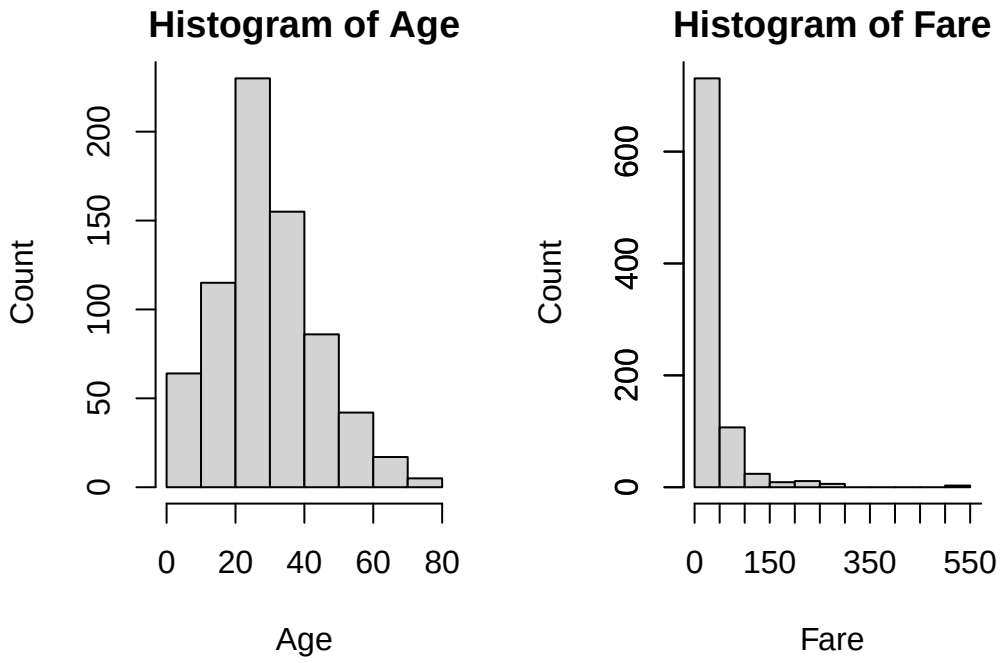
```
length(which(is.na(dat$Fare)))
```

```
[1] 0
```

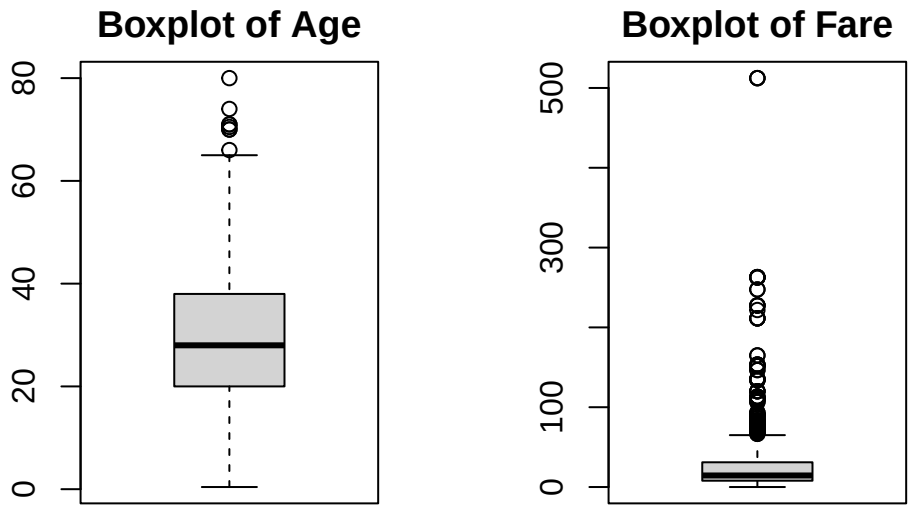
```
#descriptive statistics
Mean <- apply(dat[,c(6,10)], 2, function(x) mean(x, na.rm = TRUE))
Median <- apply(dat[,c(6,10)], 2, function(x) median(x, na.rm = TRUE))
Variance <- apply(dat[,c(6,10)], 2, function(x) var(x, na.rm = TRUE))
Standard_deviation <- apply(dat[,c(6,10)], 2, function(x)sd(x, na.rm = TRUE))
full_range <- apply(dat[,c(6,10)],2,function(x)range(x, na.rm = TRUE))
Range <- full_range[2,]-full_range[1,]
IQR <- apply(dat[,c(6,10)], 2,function(x)IQR(x, na.rm = TRUE))

continuous_table <- data.frame(Mean,Median,Variance,Standard_deviation,Range,IQR)
print(round(continuous_table,2))
```

```
     Mean Median Variance Standard_deviation  Range   IQR
Age  29.7  28.00   211.02              14.53  79.58 17.88
Fare 32.2  14.45  2469.44              49.69 512.33 23.09
```

```r
par(mar = c(4, 4, 2, 2))
par(mfrow = c(1, 2))
#hisrogram of Age
hist(dat$Age,
     main="Histogram of Age",
     xlab="Age",
     ylab="Count",
     yaxt="n")
axis(2, at=seq(0,250, by=50))
#hisrogram of Fare
hist(dat$Fare,
     main="Histogram of Fare",
     xlab="Fare",
     ylab="Count",
     xaxt="n")
axis(1, at=seq(0,600, by=50))
axis(2, at=seq(0,800, by=200))
```




```r
par(mar = c(4, 4, 2, 2))
par(mfrow = c(1, 2))
boxplot(dat$Age,
        main="Boxplot of Age")
boxplot(dat$Fare,
        main="Boxplot of Fare")
```



Age變項:
1. 有177個missing data
2. 排除missing data,平均年齡29.7歲,中位數28歲,標準差14.53歲
3. 排除missing data,20-40歲人數較多最老有80歲,最年輕4個月
Fare變項:
1. 無missing data
2. 平均票價32.2元,中位數14.45元,標準差49.69元
2. 分布非常偏態,幾乎都50元以下,最多有人花512元買票,最少是免費上船

類別型變項檢查及分布

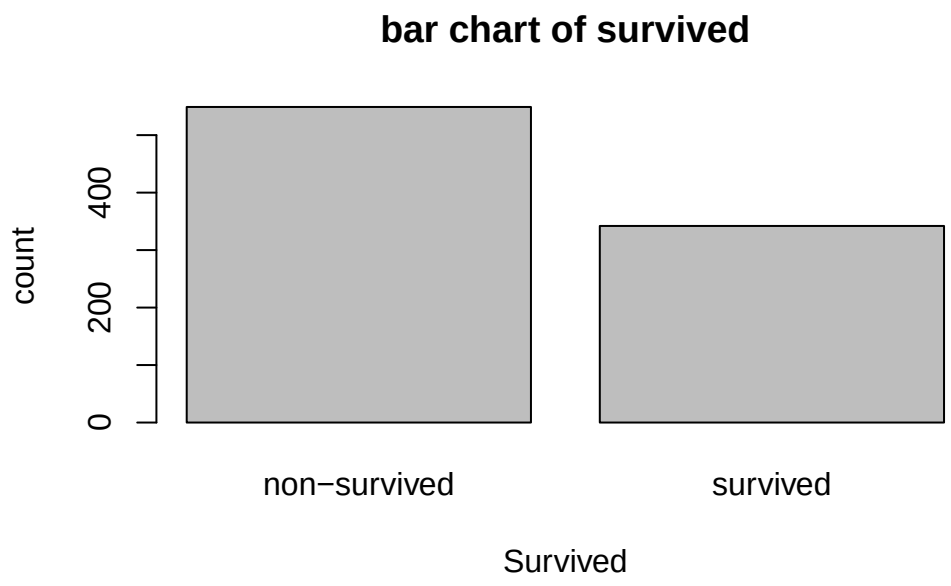
```
##Survived##
table_survived <- table(dat$Survived)
row.names(table_survived) <- c("non-survived","survived")
print(table_survived)
```

```
non-survived     survived
         549          342
```

```
round(prop.table(table_survived) * 100,2)
```

```
non-survived     survived
       61.62        38.38
```

```r
barplot(table_survived,
        main = "bar chart of survived",
        xlab = "Survived",
        ylab = "count")
```

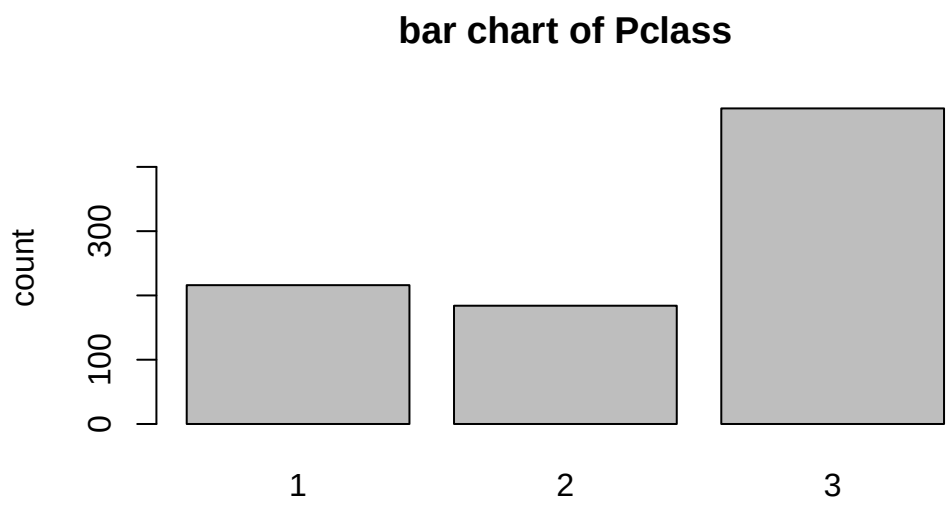


Survived變項:
1. 有342人生還
2. 38.38%生還

```r
##Pclass##
table_Pclass <- table(dat$Pclass)
print(table_Pclass)
```

```
  1   2   3
216 184 491
```

```r
barplot(table_Pclass,
        main = "bar chart of Pclass",
        ylab = "count")
```

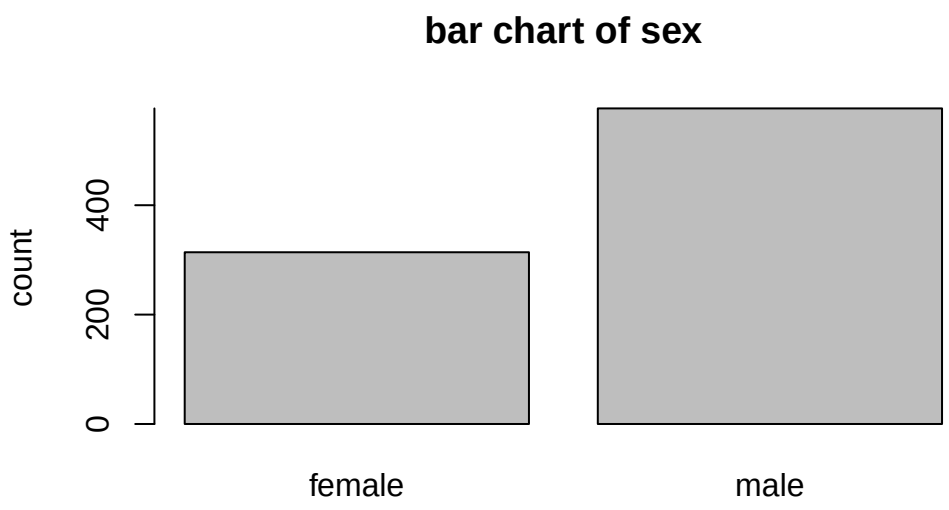


Pclass變項:此船有三種艙等,第3種艙等最多人有491人。

```
##sex##
table_sex <- table(dat$Sex)
print(table_sex)
```

```
female   male
   314    577
```

```
barplot(table_sex,
        main = "bar chart of sex",
        ylab = "count",
        yaxt="n")
axis(2, at=seq(0,800, by=200))
```

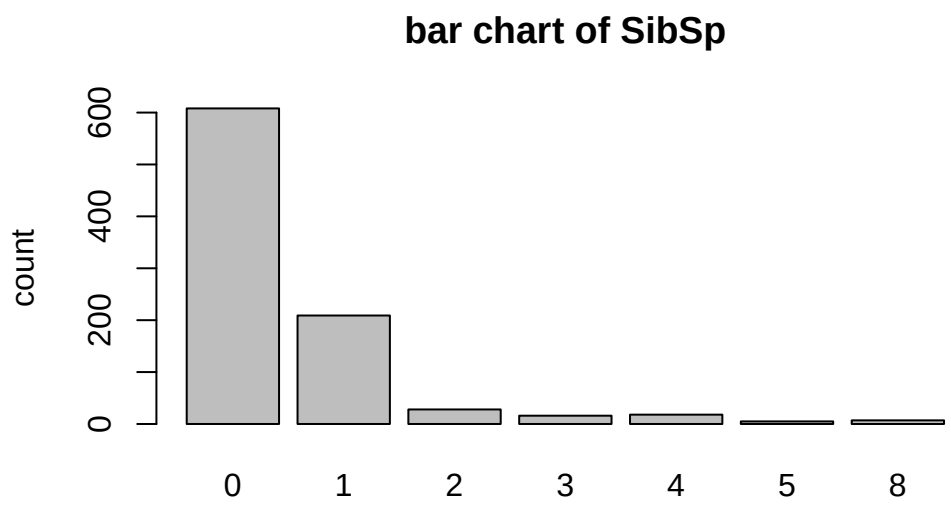


sex變項:女性314人,男性577人。

```
##SibSp##
table_sibsp <- table(dat$SibSp)
print(table_sibsp)
```

```
  0   1   2   3   4   5   8 
608 209  28  16  18   5   7 
```

```
barplot(table_sibsp,
        main = "bar chart of SibSp",
        ylab = "count")
```

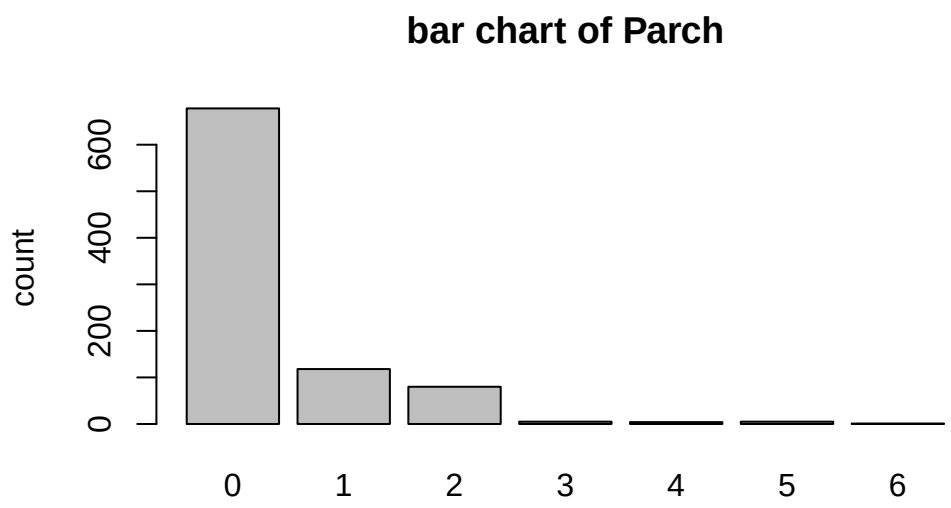


SibSp變項指有多少兄弟姊妹+夫妻一起在船上,大部分都是獨自登船。

```
##Parch##
table_parch <- table(dat$Parch)
print(table_parch)
```

```
  0   1   2   3   4   5   6
678 118  80   5   4   5   1
```

```
barplot(table_parch,
        main = "bar chart of Parch",
        ylab = "count")
```



Parch變項指有多少父母子女一起在船上,大部分都是獨自登船。

```
##Cabin##
table_Cabin <- table(dat$Cabin)
head(table_Cabin)
```

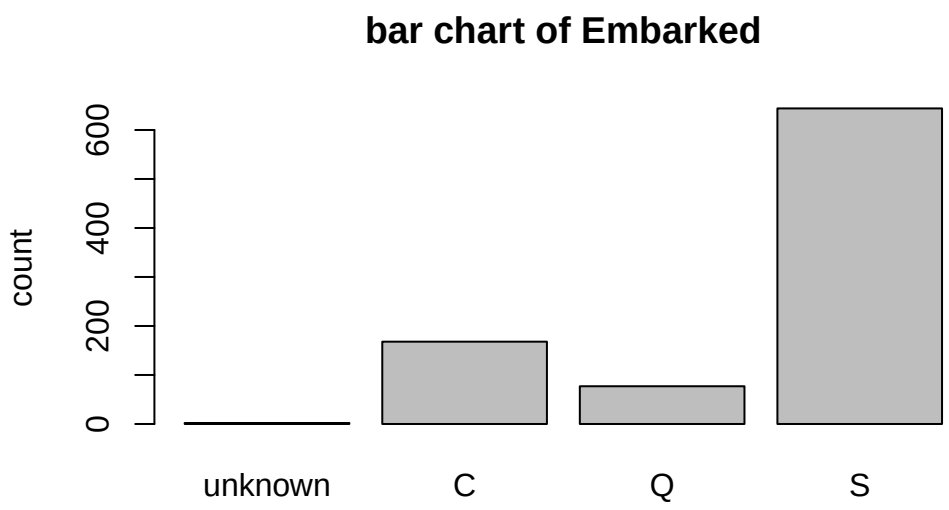
```
    A10 A14 A16 A19 A20
687   1   1   1   1   1
```

Cabin變項指的是入住房間號碼,僅顯示前六筆統計資料,無房間號碼者有687人,太多未知,後續不考慮此變項的分析。

```
##Embarked##
table_embarked <- table(dat$Embarked)
row.names(table_embarked) <- c("unknown","C","Q","S")
print(table_embarked)
```

```
unknown       C       Q       S
      2     168      77     644
```

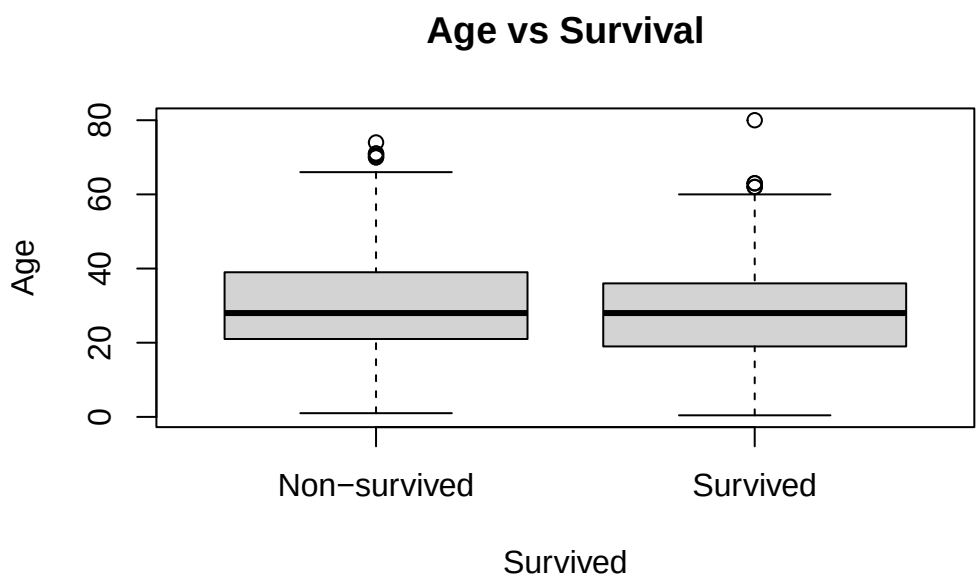
```
barplot(table_embarked,
        main = "bar chart of Embarked",
        ylab = "count")
```



Embarked變項指的是從哪個港口登船,有兩個人不知道從哪裡登船,大部分由S港登船。

各變項與存活之間的情況

```
##Age##
boxplot(dat$Age ~ dat$Survived, data = dat,
        main = "Age vs Survival",
        xlab = "Survived",
        ylab = "Age",
        names = c("Non-survived", "Survived"))
```

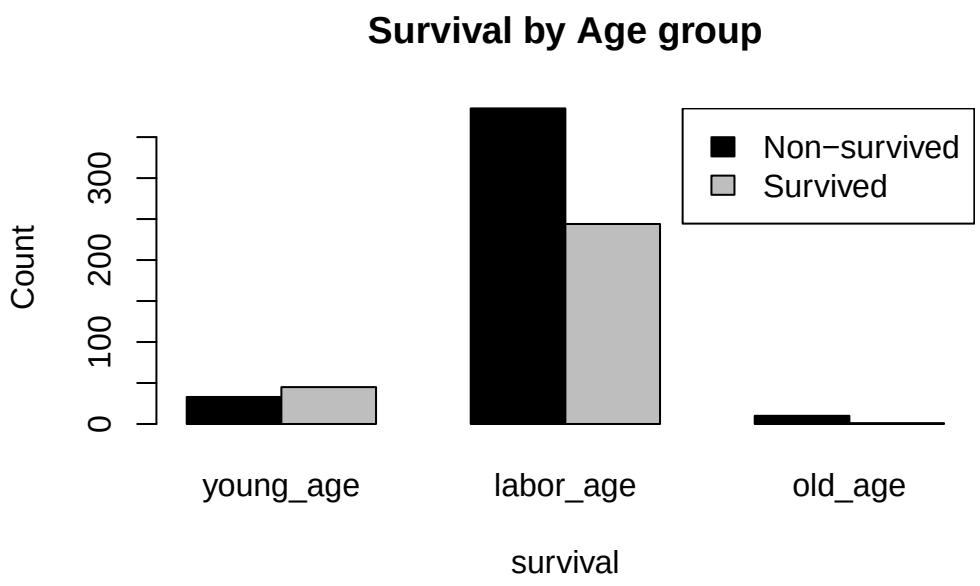


Age在存活的情況來看,乍看似乎存活的平均年齡較小,但極度老的老人也有存活,將年齡分組看一下存活情況。

```
young <- dat[which(dat$Age < 15),]
labor <- dat[which(dat$Age >14 & dat$Age < 66),]
old <- dat[which(dat$Age > 64),]
young_age <- table(young$Survived)
labor_age<- table(labor$Survived)
old_age <- table(old$Survived)
age_group <- rbind(young_age,labor_age,old_age)
age_group <- t(age_group)
row.names(age_group) <- c("non-survived","survived")
round(prop.table(age_group,margin = 2) * 100,2)
```

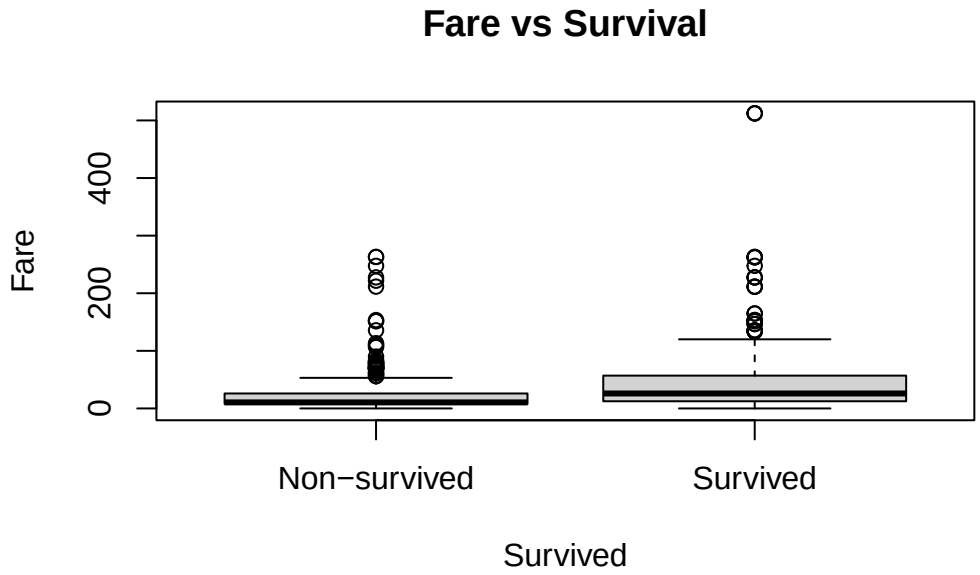
```
             young_age labor_age old_age
non-survived     42.31     61.21   90.91
survived         57.69     38.79    9.09
```

```
barplot(age_group, beside = TRUE,
        main = "Survival by Age group",
        xlab = "survival",
        ylab = "Count",
        col = c("black","gray"))
legend("topright",legend = c("Non-survived", "Survived"),fill = c("black", "gray"))
```



排除missing data,將年齡分成0-14歲為young組,15-64歲為labor組,65歲以上為old組,其中:
14歲以下的半數存活(57.69%),65歲以上九成都死亡(90.91%),15-64僅38.79%存活。

```
##Fare##
boxplot(dat$Fare ~ dat$Survived, data = dat,
        main = "Fare vs Survival",
        xlab = "Survived",
        ylab = "Fare",
        names = c("Non-survived", "Survived"))
```

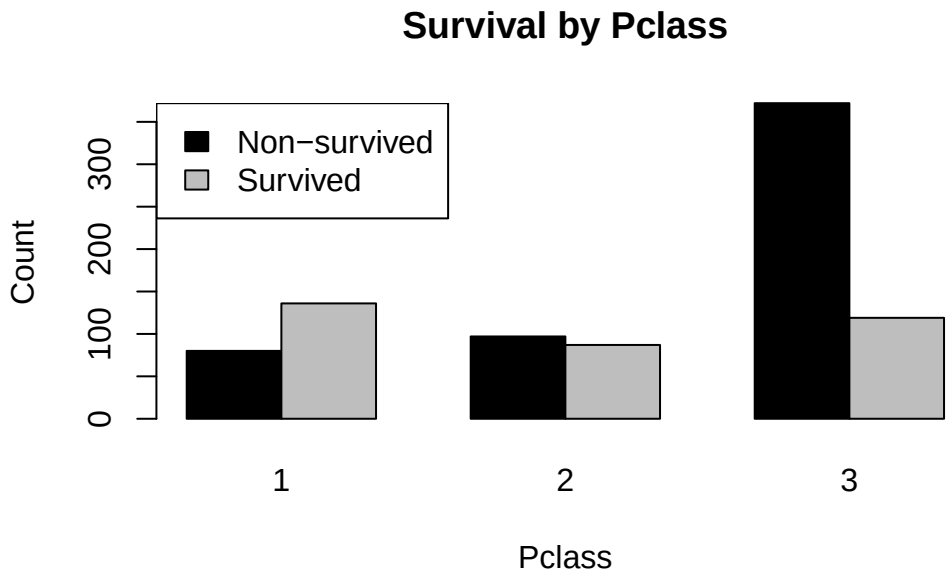


票價最貴的人有存活,整體而言沒有存活的族群票價較集中,似乎也是較便宜的票價。

```r
##Pclass##
table_survival_Pclass <- table(dat$Survived,dat$Pclass)
row.names(table_survival_Pclass) <- c("non-survived","survived")
round(prop.table(table_survival_Pclass,margin = 2) * 100,2)
```

```
                1     2     3
  non-survived 37.04 52.72 75.76
  survived     62.96 47.28 24.24
```

```r
barplot(table_survival_Pclass, beside = TRUE,
        main = "Survival by Pclass",
        xlab = "Pclass",
        ylab = "Count",
        col = c("black","gray"))
legend("topleft",legend = c("Non-survived", "Survived"),fill = c("black", "gray"))
```

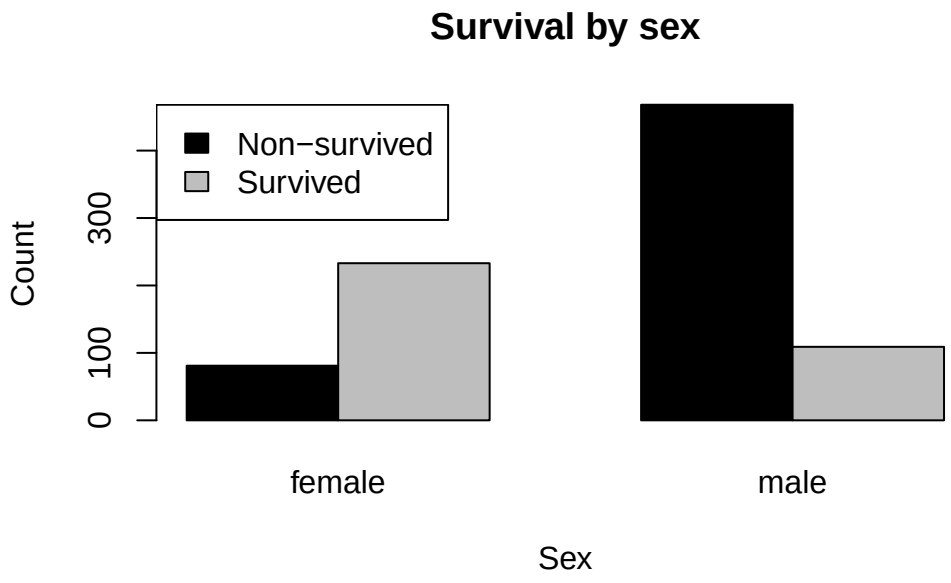


坐第1種艙等的人之中,存活(62.96%)多於死亡(37.04%),坐第3種艙等的人非常多人死亡(75.76%)。

```r
##Sex##
table_survival_sex <- table(dat$Survived,dat$Sex)
row.names(table_survival_sex ) <- c("non-survived","survived")
round(prop.table(table_survival_sex,margin = 2) * 100,2)
```

```
              female  male
  non-survived  25.80 81.11
  survived      74.20 18.89
```

```r
barplot(table_survival_sex, beside = TRUE,
        main = "Survival by sex",
        xlab = "Sex",
        ylab = "Count",
        col = c("black","gray"))
legend("topleft",legend = c("Non-survived", "Survived"),fill = c("black", "gray"))
```

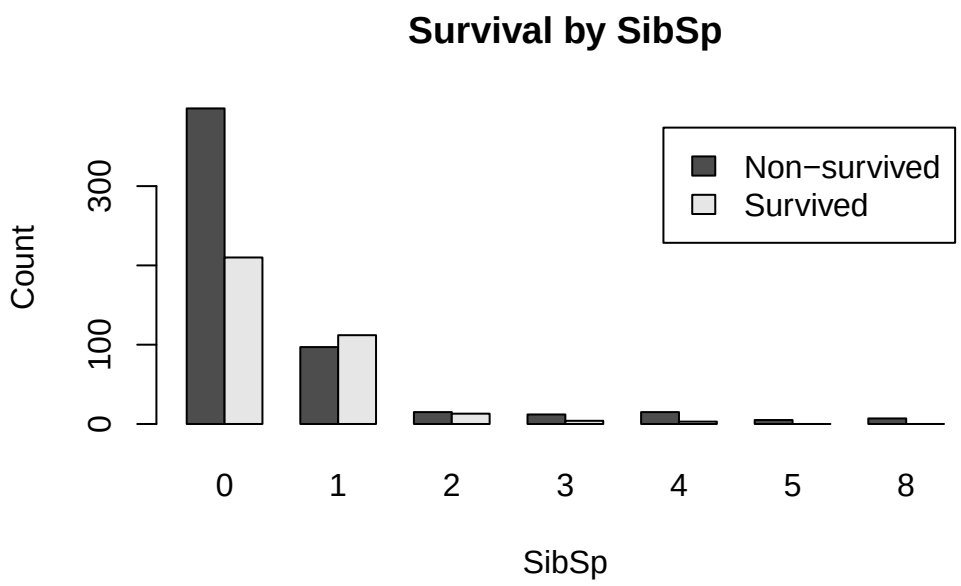


女性存活較多(74.20%都存活),而男性死亡較多(81.11%都死亡)。

```r
##SibSp##
table_survival_SibSp <- table(dat$Survived,dat$SibSp)
row.names(table_survival_SibSp) <- c("non-survived","survived")
round(prop.table(table_survival_SibSp,margin = 2) * 100,2)
```

```
                0     1     2     3     4      5      8
non-survived 65.46 46.41 53.57 75.00 83.33 100.00 100.00
survived     34.54 53.59 46.43 25.00 16.67   0.00   0.00
```

```r
barplot(table_survival_SibSp, beside = TRUE,
        legend = c("Non-survived", "Survived"),
        main = "Survival by SibSp",
        xlab = "SibSp",
        ylab = "Count")
```



有1到2個手足或夫妻的乘客半數都生還,死亡的多為獨自上船者,但有5個以上的手足(夫妻)都死亡。

```
##Parch##
table_survival_Parch <- table(dat$Survived,dat$Parch)
row.names(table_survival_Parch) <- c("non-survived","survived")
round(prop.table(table_survival_Parch ,margin = 2) * 100,2)
```

```
                0     1     2     3      4     5      6
non-survived 65.63 44.92 50.00 40.00 100.00 80.00 100.00
survived     34.37 55.08 50.00 60.00   0.00 20.00   0.00
```

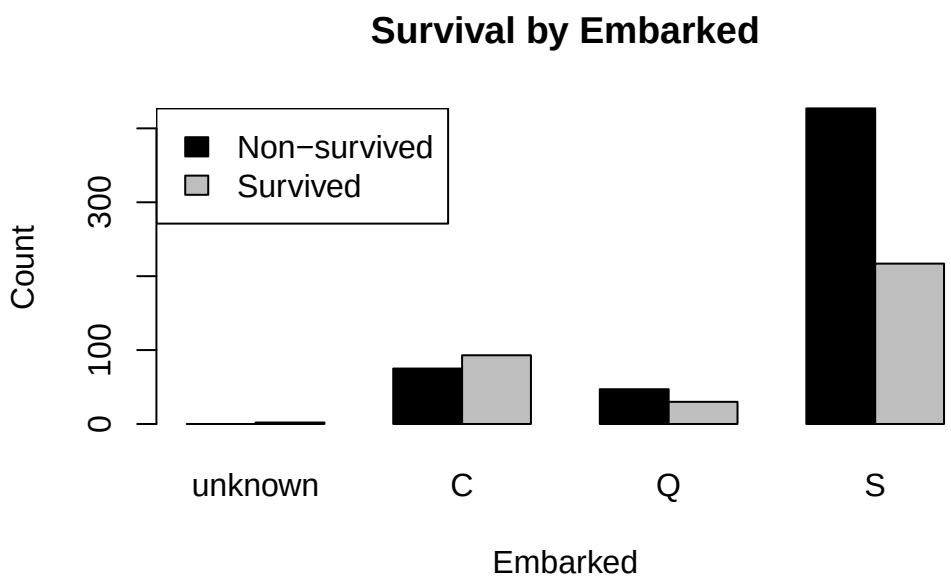
```
barplot(table_survival_Parch, beside = TRUE,
        legend = c("Non-survived", "Survived"),
        main = "Survival by Parch",
        xlab = "Parch",
        ylab = "Count")
```

有1到2個父母子女的乘客半數都生還,死亡的多為獨自上船者,4個以上父母子女的組合幾乎都死亡。

```
##Embarked##
table_survival_Embarked <- table(dat$Survived,dat$Embarked)
colnames(table_survival_Embarked)<- c("unknown","C","Q","S")
row.names(table_survival_Embarked) <- c("non-survived","survived")
round(prop.table(table_survival_Embarked ,margin = 2) * 100,2)
```

```
              unknown      C      Q      S
  non-survived    0.00  44.64  61.04  66.30
  survived      100.00  55.36  38.96  33.70
```

```
barplot(table_survival_Embarked, beside = TRUE,
        main = "Survival by Embarked",
        xlab = "Embarked",
        ylab = "Count",
        col = c("black","gray"))
legend("topleft",legend = c("Non-survived", "Survived"),fill = c("black", "gray"))
```



從C港口出發的族群中生還者佔了55.36%比死亡者多,但S港口坐船族群一半以上的都死亡(66.3%)。

Summary

1.船上大概是那種人?
共有891個觀測值(891人),平均年齡29.7歲,平均花32元買票,乘客多為男性,沒有攜伴上船,大部分都坐第三艙等。
2.存活的狀況?
整船有342人存活,佔38.38%,大部分是女性,14歲以下約一半存活,而老人幾乎都死亡,攜伴1-2人者半數都存活,坐第3種艙等的人以及從S港搭船的人死亡較多。