# 統計諮詢 HW1

R26131060

I-Chuan Hung

2025-02-23

## 目錄

## 一、敘述統計

```r
# R Interface to Python
library(reticulate)
library(Hmisc)

titanic <- read.csv("C:/Users/USER/Desktop/   / /  / /2_21_HW/titanic.csv")

titanic$PassengerId <- as.factor(titanic$PassengerId)
titanic$Survived <- as.factor(titanic$Survived)
titanic$Pclass<- as.factor(titanic$Pclass)
titanic$Age<- as.integer(titanic$Age)
titanic$SibSp <- as.factor(titanic$SibSp)
titanic$Parch <- as.factor(titanic$Parch)

latex(describe(titanic), file="")
```

titanic
12 Variables    891 Observations

**PassengerId**

```
  n   missing   distinct
891         0        891

lowest : 1    2    3    4    5   , highest: 887 888 889 890 891
```

**Survived**

```
  n   missing   distinct
891         0          2

Value          0     1
Frequency    549   342
Proportion 0.616 0.384
```

**Pclass**

```
  n   missing   distinct
891         0          3

Value          1     2     3
Frequency    216   184   491
Proportion 0.242 0.207 0.551
```

### Name

```
    n  missing  distinct
  891        0       891

lowest : Abbing, Mr. Anthony                Abbott, Mr. Rossmore Edward           Abbott, Mrs. Stanton (Rosa Hunt)       Abelson, Mr
highest: Yousseff, Mr. Gerious              Yrois, Miss. Henriette ("Mrs Harbeck") Zabour, Miss. Hileni                  Zabour, Mis
```

### Sex

```
    n  missing  distinct
  891        0         2

Value      female   male
Frequency     314    577
Proportion  0.352  0.648
```

### Age

```
    n  missing  distinct   Info    Mean  pMedian    Gmd   .05   .10   .25   .50   .75   .90   .95
  714      177        71  0.999   29.68       29  16.22     4    14    20    28    38    50    56

lowest :  0  1  2  3  4, highest: 66 70 71 74 80
```

### SibSp

```
    n  missing  distinct
  891        0         7

Value          0     1     2     3     4     5     8
Frequency    608   209    28    16    18     5     7
Proportion 0.682 0.235 0.031 0.018 0.020 0.006 0.008
```

### Parch

```
    n  missing  distinct
  891        0         7

Value          0     1     2     3     4     5     6
Frequency    678   118    80     5     4     5     1
Proportion 0.761 0.132 0.090 0.006 0.004 0.006 0.001
```

### Ticket

```
    n  missing  distinct
  891        0       681

lowest : 110152      110413      110465      110564      110813
highest: W./C. 6608  W./C. 6609  W.E.P. 5734 W/C 14208   WE/P 5735
```

### Fare

```
       n  missing distinct     Info     Mean  pMedian      Gmd      .05      .10      .25
     891        0      248        1     32.2     19.6    36.78    7.225    7.550    7.910
     .50      .75      .90      .95
  14.454   31.000   77.958  112.079

lowest : 0       4.0125  5       6.2375  6.4375 , highest: 227.525 247.521 262.375 263     512.329
```

### Cabin

```
    n  missing  distinct
  204      687       147

lowest : A10 A14 A16 A19 A20, highest: F33 F38 F4  G6  T
```

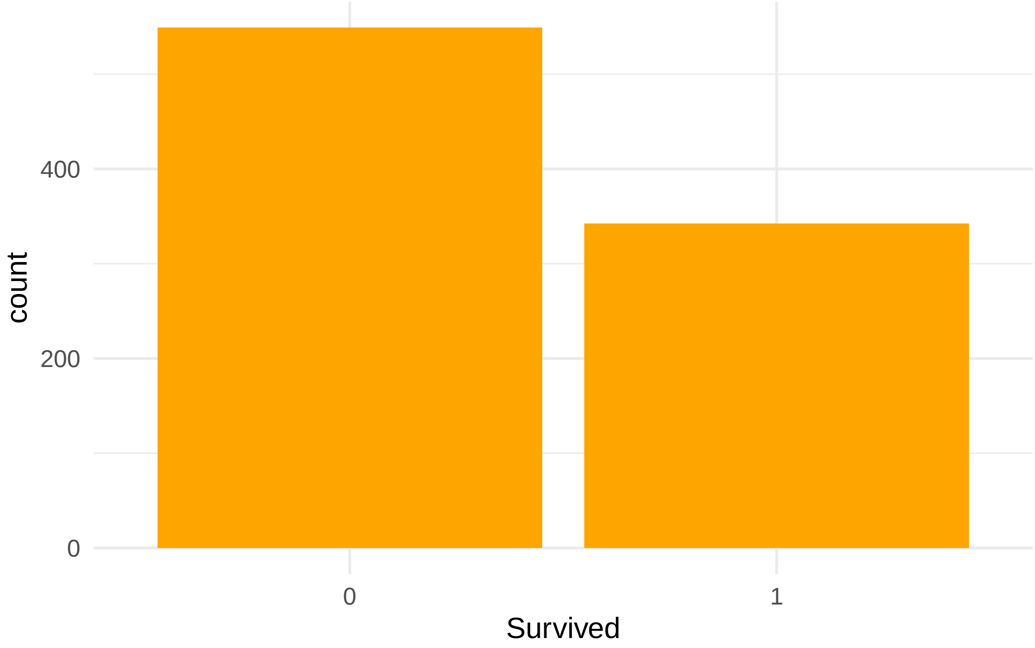
### Embarked

```
    n  missing  distinct
  889        2         3

Value          C     Q     S
Frequency    168    77   644
Proportion 0.189 0.087 0.724
```

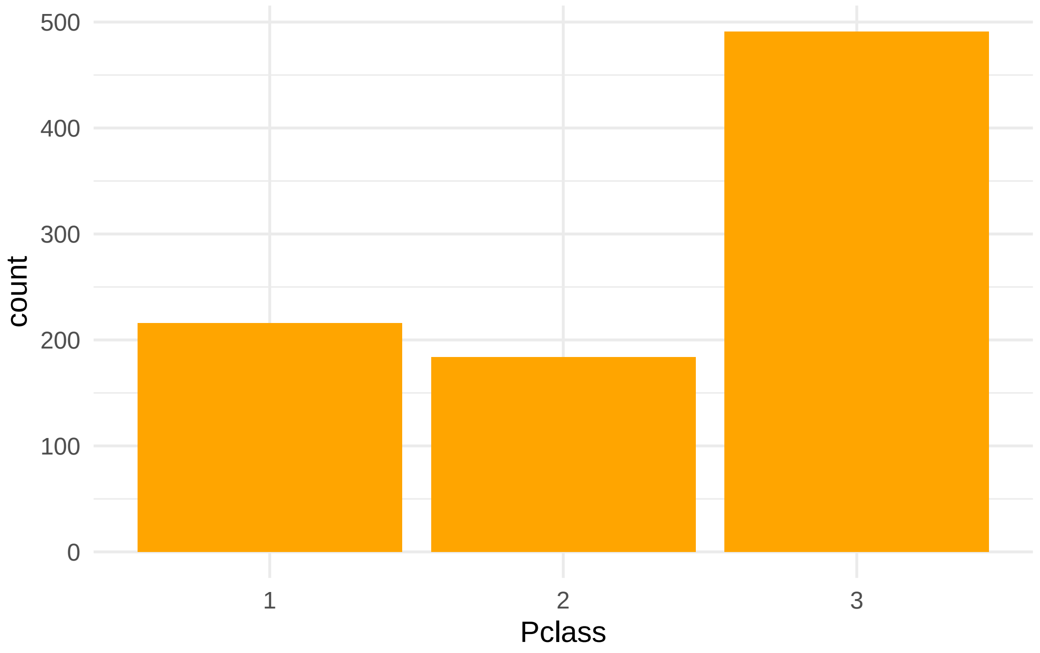
## 二、資料視覺化

### (1) 類別型變數 (Survived, Pclass, Sex, Embarked, SibSp, Parch)

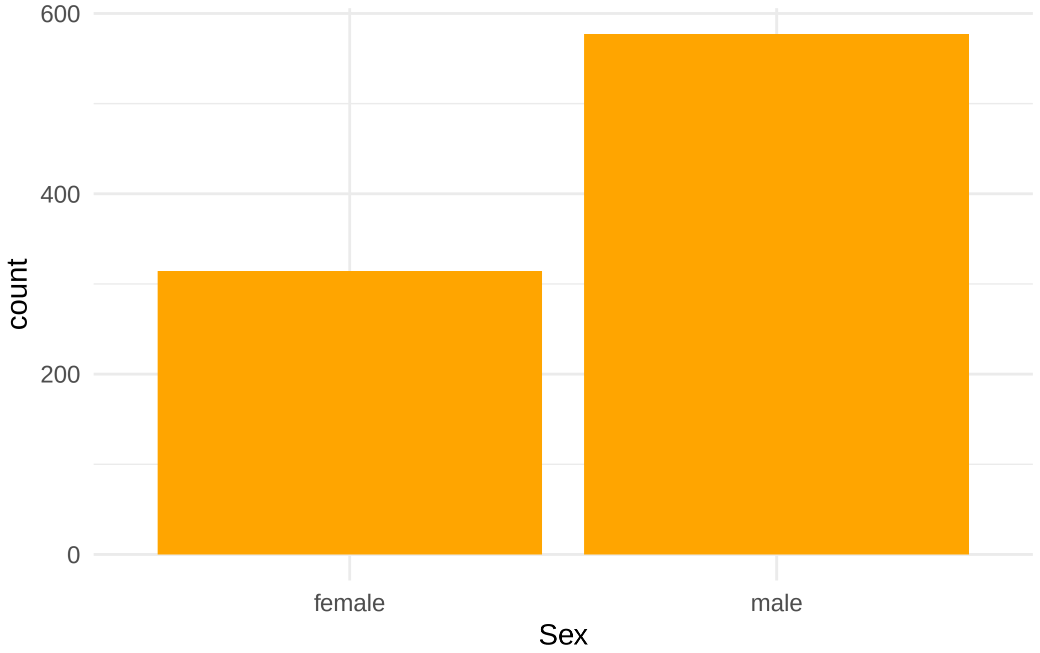
```
library(dplyr)
library(ggplot2)

# Survived
ggplot(titanic, aes(x = Survived)) +
  geom_bar(fill = "orange") +
  theme_minimal()
```



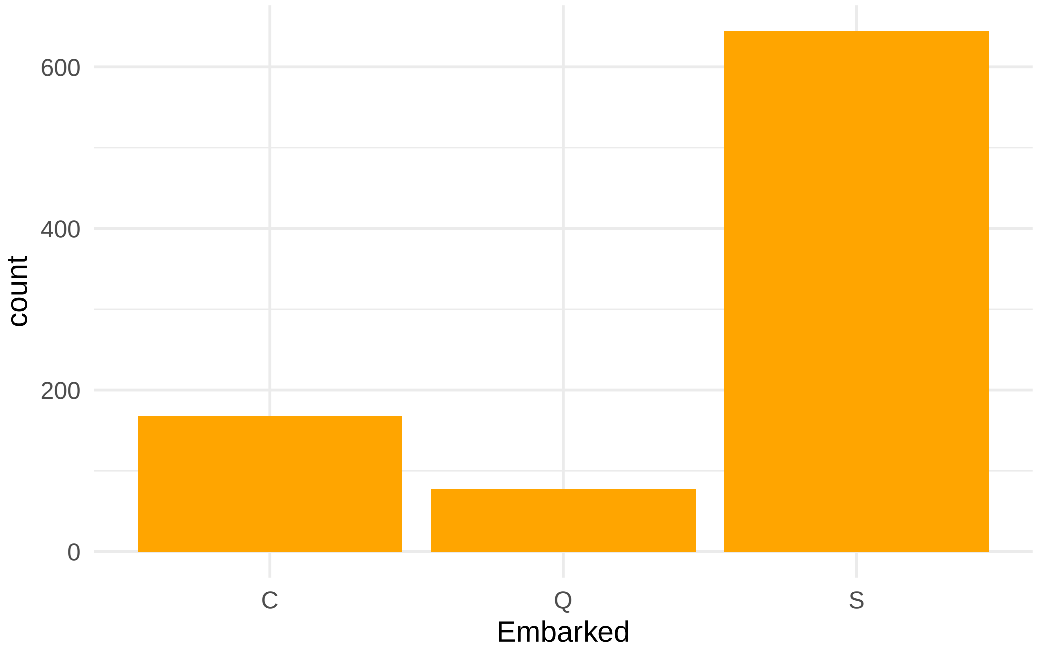
```
# Pclass
ggplot(titanic, aes(x = Pclass)) +
  geom_bar(fill = "orange") +
  theme_minimal()
```



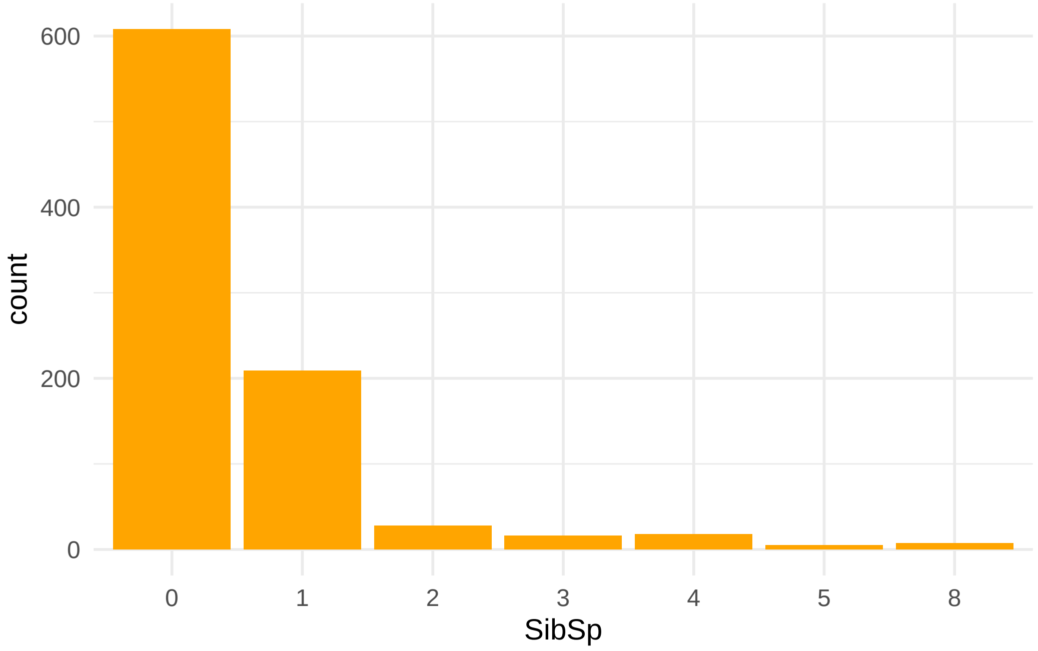
```
# Sex
ggplot(titanic, aes(x = Sex)) +
  geom_bar(fill = "orange") +
  theme_minimal()
```



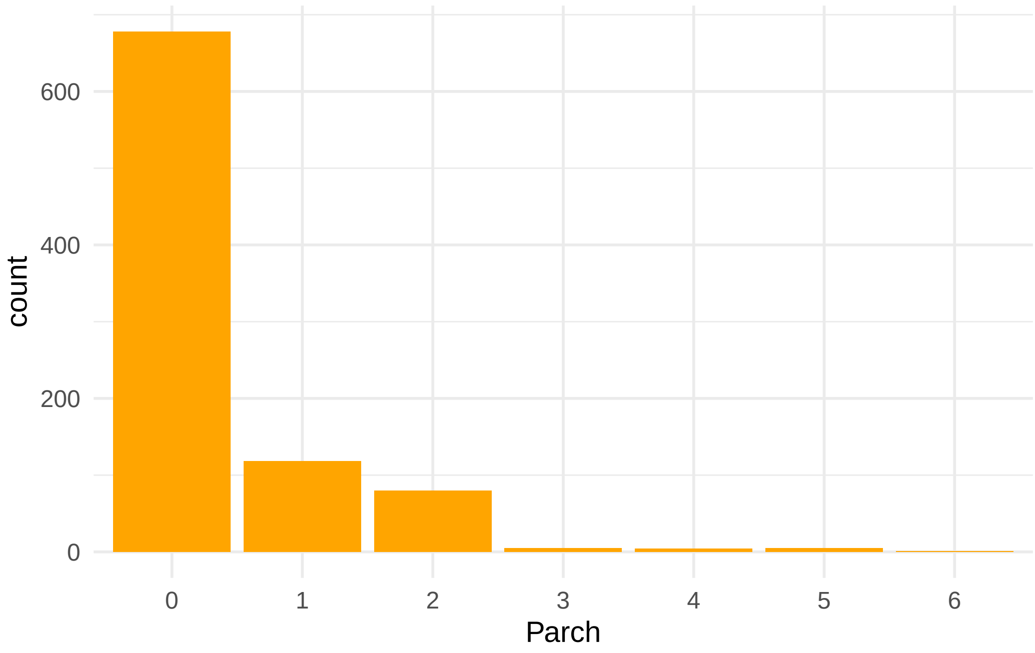
```
# Embarked
dataEmbarked <- titanic[titanic$Embarked != "",]
ggplot(dataEmbarked, aes(x = Embarked)) +
  geom_bar(fill = "orange") +
  theme_minimal()
```



```
# SibSp
ggplot(titanic, aes(x = SibSp)) +
  geom_bar(position = "stack", fill = "orange") +
  theme_minimal()
```

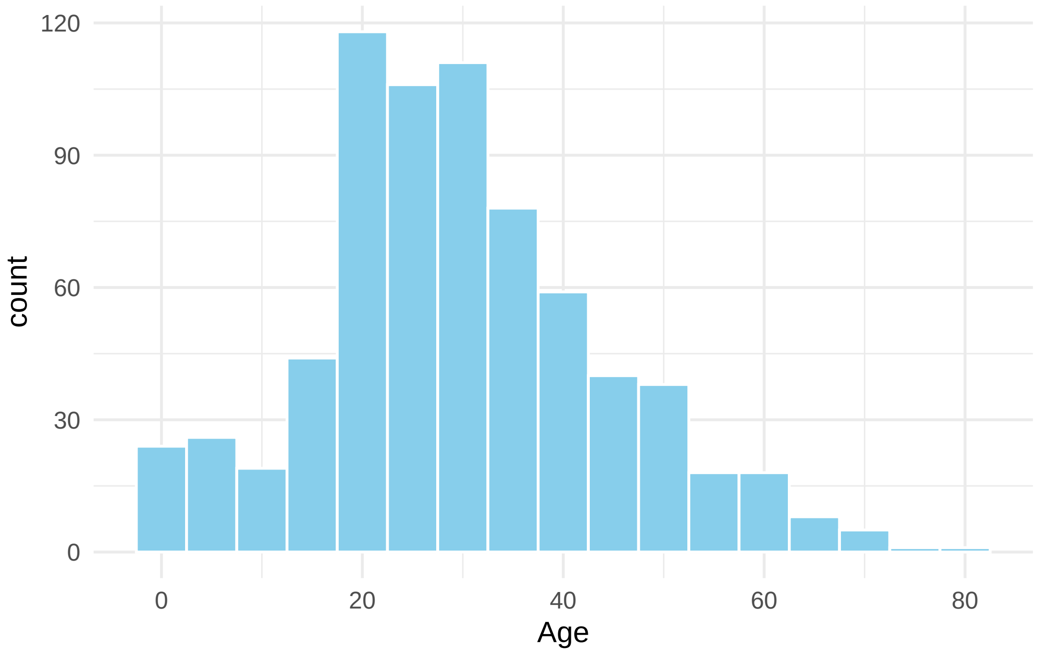


```
# Parch
ggplot(titanic, aes(x = Parch)) +
  geom_bar(position = "stack", fill = "orange") +
  theme_minimal()
```

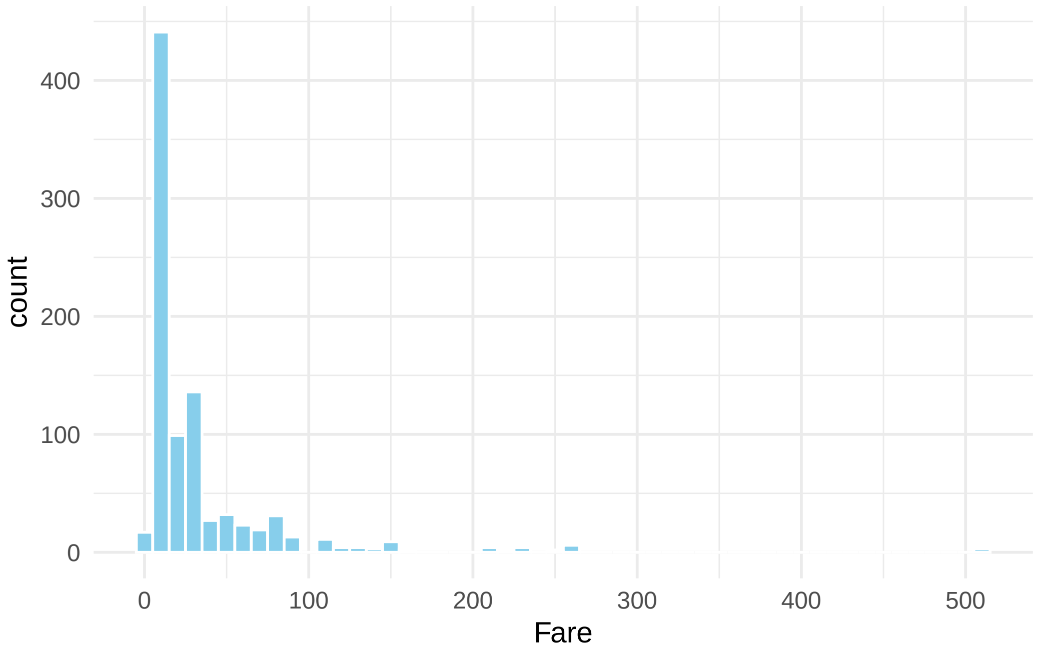

count
600
400
200
0
0
1
2
3
4
5
6
Parch

### (2) 連續型變數 (Age, Fare)

```r
# Age
ggplot(titanic, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "white") +
  theme_minimal()
```



```r
# Fare
ggplot(titanic, aes(x = Fare)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "white") +
  theme_minimal()
```



## 三、結論

Survived(罹難:0,存活:1): 船難罹難人數較多。

Pclass(艙等): 艙等3乘員較多。

Sex(性別): 男性人數約為女性人數的2倍。

Embarked(出發港口): 多數乘員從S港口出發。

SibSp(兄弟姊妹＋老婆丈夫數量): 多數乘員的兄弟姊妹＋老婆丈夫數量為0~1個。

Parch(父母小孩的數量): 多數乘員的父母小孩數量為0~2個。

Age(年齡): 乘員年齡大多介於20~40歲之間，但此變數有177個缺失值。

Fare(票的費用): 大多數成員的船票費用價格位於較低的區間。