HW1

A summary report for the Titanic dataset

Hung-Tse Hsu

2025-02-23

目錄

一、讀取資料

```
# R Interface to Python
library(readxl)
library(reticulate)
library(Hmisc)
titanic.df <- read.csv("/Users/xuhongze/Statistical_Consulting/HW1/titanic.csv")
latex(describe(titanic.df), file="")
```

titanic.df
12 Variables 891 Observations

PassengerId

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
891	0	891	1	446	297.3	45.5	90.0	223.5	446.0	668.5	802.0	846.5

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

Survived

n	missing	distinct	Info	Sum	Mean	Gmd
891	0	2	0.71	342	0.3838	0.4735

Pclass

n	missing	distinct	Info	Mean	Gmd
891	0	3	0.81	2.309	0.8631

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

For the frequency table, variable is rounded to the nearest 0

Name

n	missing	distinct
891	0	891

lowest : Abbing Mr. Anthony Abbott Mr. Rossmore Edward Abbott Mrs. Stanton (Rosa Hunt) Abelson Mr. Sa
highest: Youssef Mr. Gerious Yrois Miss. Henriette ("Mrs Harbeck") Zabour Miss. Hileni Zabour Miss. T

Sex

```
   n    missing    distinct
 891          0           2

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

Age

```
   n    missing    distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
 714        177         88    0.999     29.7    16.21     4.00    14.00    20.12    28.00    38.00    50.00    56.00

lowest : 0.42 0.67 0.75 0.83 0.92, highest: 70   70.5 71   74   80
```

SibSp

```
   n    missing    distinct     Info     Mean      Gmd
 891          0           7    0.669    0.523    0.823

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

For the frequency table, variable is rounded to the nearest 0
```

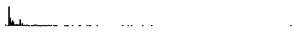
Parch

```
   n    missing    distinct     Info     Mean      Gmd
 891          0           7    0.556   0.3816   0.6259

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

For the frequency table, variable is rounded to the nearest 0
```

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      Gmd      .05      .10      .25      .50      .75      .90      .95
 891          0        248        1     32.2    36.78    7.225    7.550    7.910   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
```

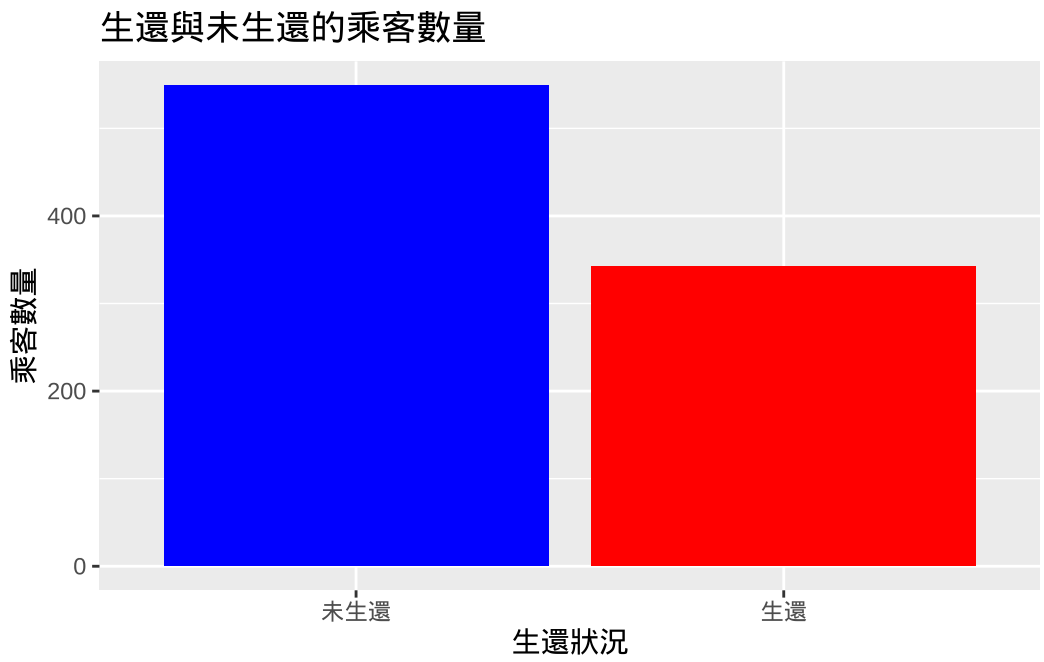
Embarked

```
   n    missing    distinct
 889          2           3

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

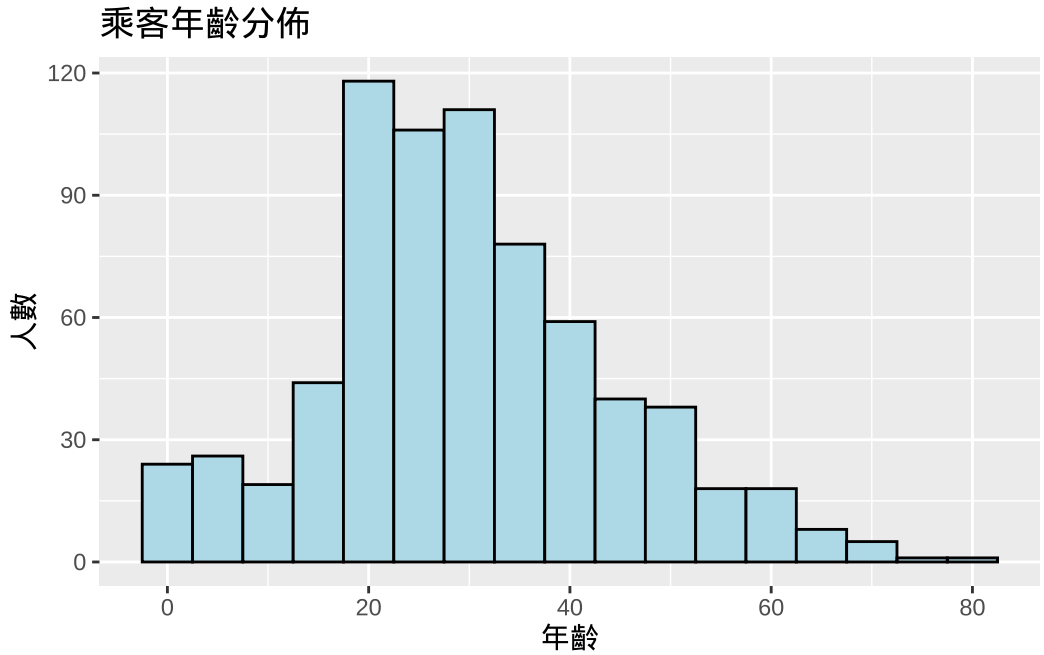
二、資料視覺化

```
#
library(ggplot2)
library(dplyr)
library(showtext)
showtext_auto()
theme_set(theme_gray(base_family = "PingFang TC"))

# 1. Barplot of Survived
```

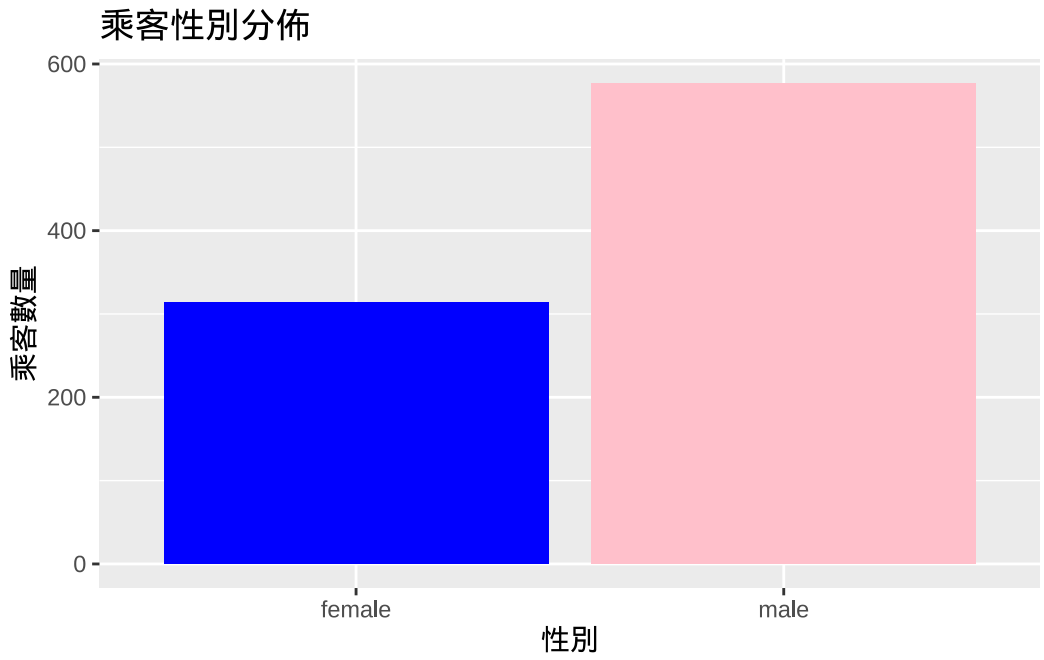
```
ggplot(titanic.df, aes(x = factor(Survived))) +
  geom_bar(fill = c("blue", "red")) +
  labs(x = "   ", y = "   ", title = "      ") +
  scale_x_discrete(labels = c("  ", " "))
```



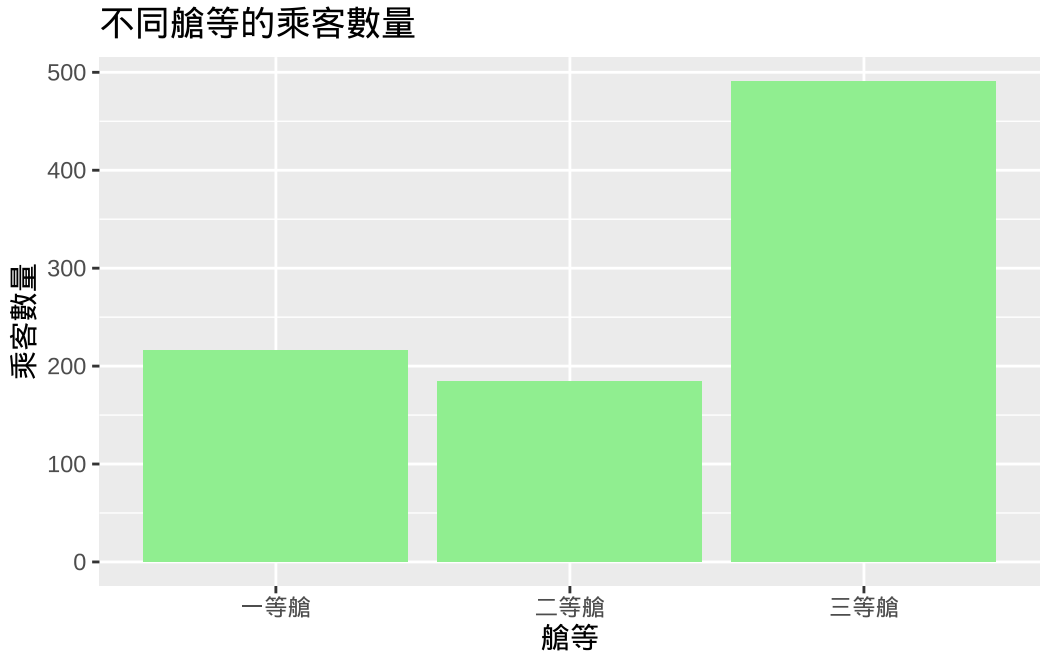

```
# 2. Histogram of Age
ggplot(titanic.df, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(x = " ", y = " ", title = "    ")
```



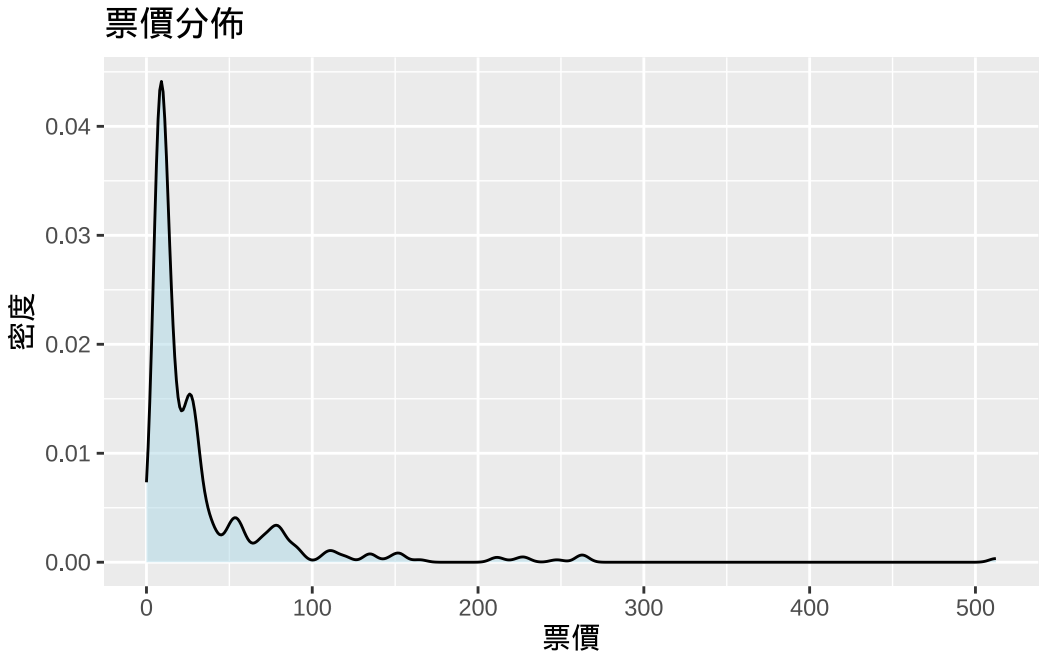
```
# 3. Barplot of Sex
ggplot(titanic.df, aes(x = Sex)) +
  geom_bar(fill = c("blue", "pink")) +
  labs(x = " ", y = "   ", title = "    ")
```



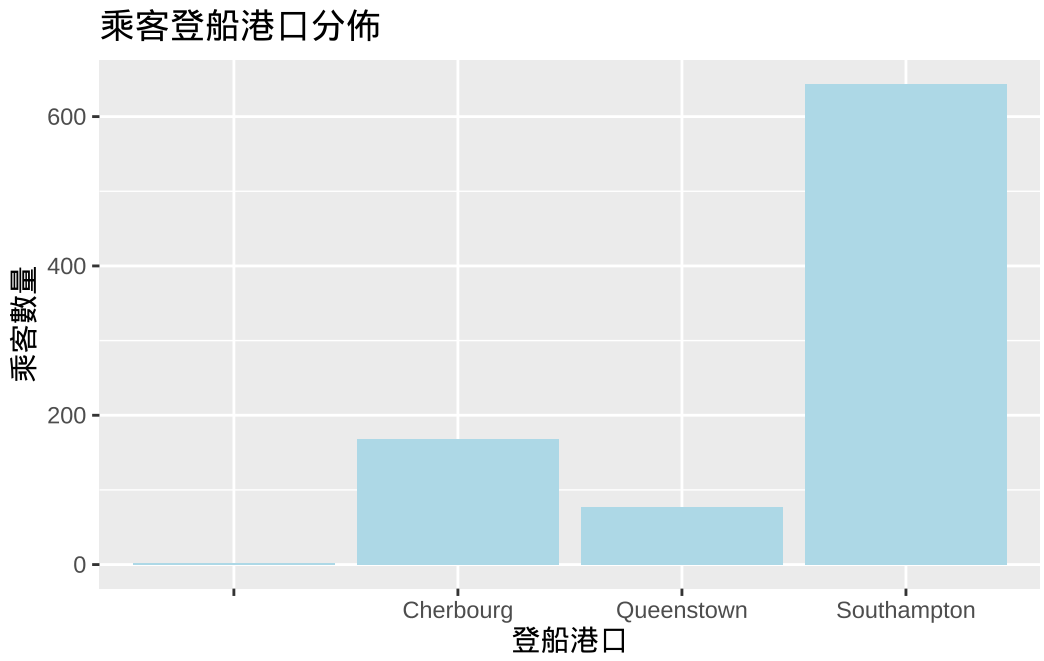

```
# 4. Barplot of Pclass
ggplot(titanic.df, aes(x = factor(Pclass))) +
  geom_bar(fill = "lightgreen") +
  labs(x = " ", y = "   ", title = "      ") +
  scale_x_discrete(labels = c("  ", "  ", "  "))
```



```r
# # 5. Barplot of Fare
# ggplot(titanic.df, aes(y = Fare)) +
#   geom_boxplot(fill = "lightcoral") +
#   labs(y = " ", title = "   ")

# 5. Density plot of Far
ggplot(titanic.df, aes(x = Fare)) +
  geom_density(fill = "lightblue", alpha = 0.5) +
  labs(x = " ", y = " ", title = "   ")
```




```r
# 6. Barplot of Embarked
ggplot(titanic.df, aes(x = Embarked)) +
  geom_bar(fill = "lightblue") +
  labs(x = "   ", y = "   ", title = "     ") +
  scale_x_discrete(labels = c("C" = "Cherbourg", "Q" = "Queenstown", "S" = "Southampton"))
```



三、資料摘要

1. Passenger Id

總共有 891 筆資料，無缺失值，所有資料的 Passenger Id 皆不重複。

2. Survived (生還狀況)

總共有 891 筆資料，無缺失值，生還狀況有兩種分類。

0（未生還）：342 人、1（生還）：549 人。

生還率約為 38.38%。

3. Pclass (艙等)

總共有 891 筆資料，無缺失值，艙等有 3 種分類。

一等艙：216 人、二等艙：184 人、三等艙：491 人。

三等艙的比例最大，佔總數的 55.1%。

4. Name (姓名)

總共有 891 筆資料，無缺失值。

姓名有 891 種不同的名稱。

5. Sex (性別)

總共有 891 筆資料，無缺失值，性別有兩種分類。

女性：314 人（35.2%）、男性：577 人（64.8%）。

6. Age (年齡)

總共有 714 筆資料（177 筆缺失值）

最小值：4個月

最大值：80歲

平均年齡：29.7 歲

年齡的標準差：16.21

年齡分佈的中位數：28 歲

四分位數範圍從 20.12 歲到 50 歲

7. SibSp (兄弟姊妹/配偶數量)

總共有 891 筆資料，無缺失值。

608 位乘客有 0 名兄弟姊妹/配偶，209 位乘客有 1 名。

8. Parch (父母/子女數量) 總共有 891 筆資料，無缺失值。

678 位乘客有 0 名父母或子女，118 位乘客有 1 名。

9. Ticket (票號)

總共有 891 筆資料，無缺失值。

10. Fare (票價)

總共有 891 筆資料，無缺失值，票價範圍從 0 到 512.33 美元。

平均票價：32.2 美元

標準差：36.78 美元

票價的中位數：14.45 美元

四分位數範圍從 7.91 美元到 77.96 美元

11. Cabin (艙房號碼)

總共有 204 筆資料，缺失值有 687 筆，艙房號碼有 147 種不同的標識。

部分艙房號碼如 A10、A14、F33 等，顯示出不均勻的分佈。

12. Embarked (登船港口)

總共有 889 筆資料，缺失值 2 筆

登船港口有 3 種（C = Cherbourg，Q = Queenstown，S = Southampton）。

S：644 人（72.4%）、C：168 人（18.9%）、Q：77 人（8.7%）