HW1

施承峻

2025-02-27

目錄

一、讀取資料與資料初步檢視 1
二、資料缺失值 2
三、敘述統計量與類別個數統整 3
四、繪圖 3

```
# R Interface to Python
library(reticulate)
library(Hmisc)
# latex(describe(mtcars), file="")
```

一、讀取資料與資料初步檢視

```
library(readr)
library(tidyverse)
library(ggplot2)
library(gridExtra)
titanic <- read_csv("titanic.csv")

str(titanic)
```

```
spc_tbl_ [891 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ PassengerId: num [1:891] 1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : num [1:891] 0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : num [1:891] 3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr [1:891] "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thaye
 $ Sex        : chr [1:891] "male" "female" "female" "female" ...
 $ Age        : num [1:891] 22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : num [1:891] 1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : num [1:891] 0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr [1:891] "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num [1:891] 7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr [1:891] NA "C85" NA "C123" ...
 $ Embarked   : chr [1:891] "S" "C" "S" "S" ...
 - attr(*, "spec")=
```

```
  .. cols(
  ..   PassengerId = col_double(),
  ..   Survived = col_double(),
  ..   Pclass = col_double(),
  ..   Name = col_character(),
  ..   Sex = col_character(),
  ..   Age = col_double(),
  ..   SibSp = col_double(),
  ..   Parch = col_double(),
  ..   Ticket = col_character(),
  ..   Fare = col_double(),
  ..   Cabin = col_character(),
  ..   Embarked = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

Titanic資料集有891筆資料，共12個變數，以下是變數的說明與資料類型

變數	說明	類型
PassengerId	ID（一般不考慮放進分析中）	-
Survived	存活與否（0 = 死亡，1 = 存活）	類別
Pclass	艙級（1, 2, 3分別是一、二、三等艙）	類別
Name	乘客姓名（一般不考慮放進分析中）	-
Sex	性別（male, female）	類別
Age	年齡	連續
SibSp	船上的兄弟姊妹或配偶數量	連續
Parch	船上的父母或子女數量	連續
Ticket	票號（一般不考慮放進分析中）	-
Fare	票價	連續
Cabin	艙房號碼（可能考慮不放進分析中）	類別
Embarked	登船港口（C = Cherbourg, Q = Queenstown, S = Southampton）	類別

二、資料缺失值

```
for (col in names(titanic)) {
  missing_rows <- which(is.na(titanic[[col]]))
  if (length(missing_rows) > 0) {
    cat("variable", col, "has missing values, a total of", length(missing_rows),"data entries\n")
  }
}
```

```
variable Age has missing values, a total of 177 data entries
variable Cabin has missing values, a total of 687 data entries
variable Embarked has missing values, a total of 2 data entries
```

發現變數Age、Cabin與Embarked有缺失值

三、敘述統計量與類別個數統整

以下是連續型變數的敘述統計量以及類別型變數各類別個數總結

```
titanic[, c("Survived", "Pclass", "Sex", "Embarked", "Cabin")] <-
  lapply(titanic[, c("Survived", "Pclass", "Sex", "Embarked", "Cabin")], as.factor)

summary(titanic)
```

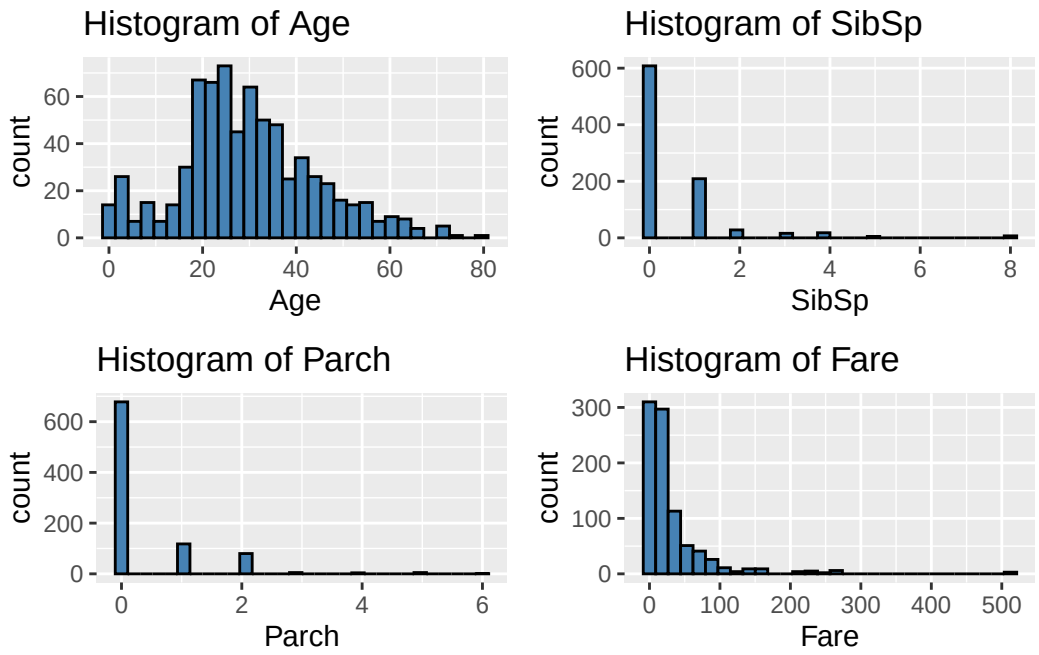
```
  PassengerId    Survived Pclass      Name              Sex
 Min.   :  1.0   0:549    1:216   Length:891         female:314
 1st Qu.:223.5   1:342    2:184   Class :character   male  :577
 Median :446.0            3:491   Mode  :character
 Mean   :446.0
 3rd Qu.:668.5
 Max.   :891.0

      Age            SibSp           Parch           Ticket
 Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891
 1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character
 Median :28.00   Median :0.000   Median :0.0000   Mode  :character
 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
      Fare                Cabin     Embarked
 Min.   :  0.00   B96 B98    :  4   C   :168
 1st Qu.:  7.91   C23 C25 C27:  4   Q   : 77
 Median : 14.45   G6         :  4   S   :644
 Mean   : 32.20   C22 C26    :  3   NA's:  2
 3rd Qu.: 31.00   D          :  3
 Max.   :512.33   (Other)    :186
                  NA's       :687
```

四、繪圖

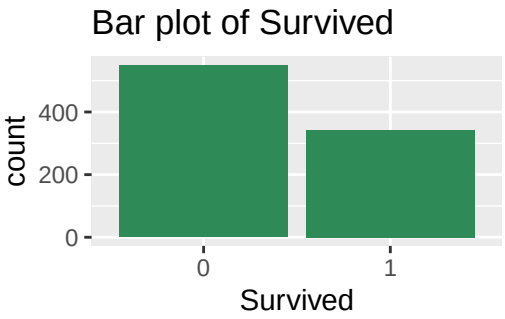
以下是將連續型變數做直方圖以及類別型變數做長條圖

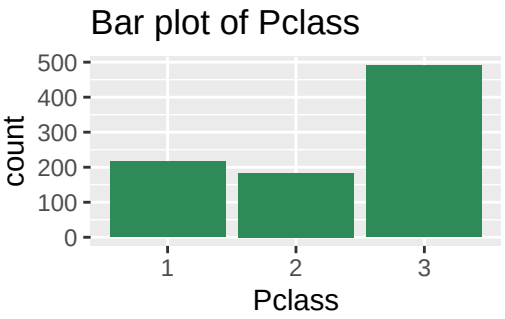
```
# Cont
p1 <- ggplot(titanic, aes(x = Age)) +
  geom_histogram(fill = "steelblue", color = "black") +
  ggtitle("Histogram of Age")

p2 <- ggplot(titanic, aes(x = SibSp)) +
  geom_histogram(fill = "steelblue", color = "black") +
  ggtitle("Histogram of SibSp")

p3 <- ggplot(titanic, aes(x = Parch)) +
  geom_histogram(fill = "steelblue", color = "black") +
  ggtitle("Histogram of Parch")

p4 <- ggplot(titanic, aes(x = Fare)) +
```

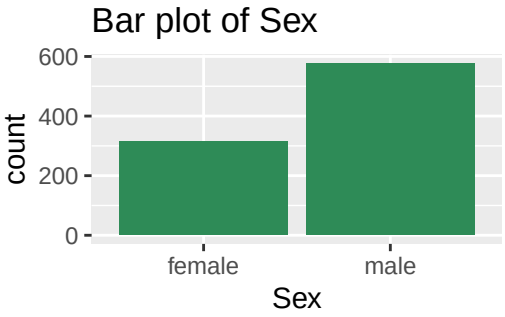
```
  geom_histogram(fill = "steelblue", color = "black") +
  ggtitle("Histogram of Fare")

grid.arrange(p1, p2, p3, p4, ncol = 2)
```

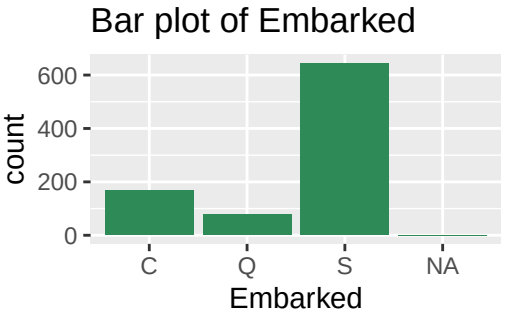



```
# Discrete
p5 <- ggplot(titanic, aes(x = Survived)) +
  geom_bar(fill = "seagreen") + ggtitle("Bar plot of Survived")

p6 <- ggplot(titanic, aes(x = Pclass)) +
  geom_bar(fill = "seagreen") + ggtitle("Bar plot of Pclass")

p7 <- ggplot(titanic, aes(x = Sex)) +
  geom_bar(fill = "seagreen") + ggtitle("Bar plot of Sex")

p8 <- ggplot(titanic, aes(x = Embarked)) +
  geom_bar(fill = "seagreen") + ggtitle("Bar plot of Embarked")

grid.arrange(p5, p6, p7, p8, ncol = 2)
```


Bar plot of Survived
count
400
200
0
0
1
Survived


Bar plot of Pclass
count
500
400
300
200
100
0
1
2
3
Pclass


Bar plot of Sex
count
600
400
200
0
female
male
Sex


Bar plot of Embarked
count
600
400
200
0
C
Q
S
NA
Embarked