Statistical Consulting

Homework 1

蕭鈺承 R26134042

2025-02-28

目錄

一、讀取資料、安裝包下載

```
library(reticulate)
library(Hmisc)
setwd("C:/Users/chars/Desktop/2025_Statistical_Consulting/Homework1")
titanic <- read.csv("titanic.csv")
```

二、資料介紹

本資料共有 12 個變數，共 891 個觀測值。
其中包含 8 個離散型變數（包含 Nominal 及 Ordinal）：

- PassengerId：乘客編號
- Survived：是否生還（0 = No, 1 = Yes）
- Pclass：票務艙等（1 = 貴賓艙, 2 = 中等艙, 3 = 經濟艙）
- Name：乘客姓名
- Sex：性別
- Ticket：票號
- Cabin：艙房號碼
- Embarked：登船港口（C = Cherbourg, Q = Queenstown, S = Southampton）

以及 4 個連續型變數：

- Age：年齡
- SibSp：同行兄弟姊妹、配偶數量
- Parch：同行父母、子女數量
- Fare：票價

三、資料前處理及摘要

將資料中缺失的欄位紀錄為 NA，並將部分以數字或字串格式紀錄的變數轉為類別型變數。

```
# str(titanic)
titanic[titanic==""] <- NA
titanic$PassengerId <- as.factor(titanic$PassengerId)
titanic$Survived <- as.factor(titanic$Survived)
titanic$Pclass <- as.factor(titanic$Pclass)
titanic$Sex <- as.factor(titanic$Sex)
titanic$SibSp <- as.factor(titanic$SibSp)
titanic$Parch <- as.factor(titanic$Parch)
titanic$Ticket <- as.factor(titanic$Ticket)
titanic$Embarked <- as.factor(titanic$Embarked)
```

```
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: Youssef, 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	88	0.999	29.7	29	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
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

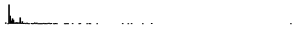
Value	0	1	2	3	4	5	6
Frequency	678	118	80	5	4	5	1
Proportion	0.761	0.132	0.090	0.006	0.004	0.006	0.001

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

四、分析方向簡介

本資料是統計領域常見的經典數據集，經常用於分類問題。若以比較為分析目的，則可以比較不同艙等、性別或登船港口之間的生還率差異。若以配飾模型為分析目的，則可以將變數 Survived 視為反應變數，進而篩選出影響生還率的變數。若以關聯性為分析目的，則可以透過皮爾森相關係數衡量連續型變數之間的相關，或透過卡方檢定測試離散型變數之間是否存在顯著相關。