Homework 1

Summary of Titanic Dataset

Tsai, Bing-Yan

2025-02-21

目錄

一、安裝套件及讀取資料

```
# R Interface to Python
library(reticulate)
library(Hmisc)
library(tinytex)
library(dplyr)
library(ggplot2)
library(gridExtra)
titanic <- read.csv("titanic.csv")
```

二、變數類型

```
str(titanic)
```

```
'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" ...
```

Nomimal variables：PassengerId, Survived, Name, Sex, Ticket, Cabin, Embarked
Ordinal variables：Pclass

Numeric variables：Age, Sibsp, Parch, Fare

三、缺失值

```
titanic_cleaned <- titanic
titanic_cleaned[titanic_cleaned == ""] <- NA
missing_values <- colSums(is.na(titanic_cleaned))
print(missing_values)
```

```
PassengerId    Survived      Pclass        Name         Sex         Age
          0           0           0           0           0         177
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0         687           2
```

891筆資料中，缺失值年齡有177筆、艙等有687筆、登船港口有2筆，其餘欄位皆無缺失值。

四、類別變數描述

```
# Survived
survival_table <- titanic %>%
  group_by(Survived) %>%
  summarise(Count = n()) %>%
  mutate(Count_Percentage = round(Count / nrow(titanic) * 100, 2)) %>%
  mutate(Percentage = round(Count / sum(Count) * 100, 2))
print(survival_table)
```

```
# A tibble: 2 x 4
  Survived Count Count_Percentage Percentage
     <int> <int>            <dbl>      <dbl>
1        0   549             61.6       61.6
2        1   342             38.4       38.4
```

```
# Pclass
pclass_table <- titanic %>%
  group_by(Pclass) %>%
  summarise(Count = n(),
            Count_Percentage = round(Count / nrow(titanic) * 100, 2),
            Survival_Rate = round(mean(Survived) * 100, 2))
print(pclass_table)
```

```
# A tibble: 3 x 4
  Pclass Count Count_Percentage Survival_Rate
   <int> <int>            <dbl>         <dbl>
1      1   216             24.2          63.0
2      2   184             20.6          47.3
3      3   491             55.1          24.2
```

```
# Sex
sex_table <- titanic %>%
  group_by(Sex) %>%
  summarise(Count = n(),
            Count_Percentage = round(Count / nrow(titanic) * 100, 2),
            Survival_Rate = round(mean(Survived) * 100, 2))
print(sex_table)
```

```
# A tibble: 2 x 4
  Sex    Count Count_Percentage Survival_Rate
  <chr>  <int>            <dbl>         <dbl>
1 female   314             35.2          74.2
2 male     577             64.8          18.9
```

```
# Embarked
embarked_table <- titanic %>%
  group_by(Embarked) %>%
  summarise(Count = n(),
            Count_Percentage = round(Count / nrow(titanic) * 100, 2),
            Survival_Rate = round(mean(Survived, na.rm = TRUE) * 100, 2))
print(embarked_table)
```

```
# A tibble: 4 x 4
  Embarked Count Count_Percentage Survival_Rate
  <chr>    <int>            <dbl>         <dbl>
1 ""           2             0.22         100
2 "C"        168            18.9           55.4
3 "Q"         77             8.64          39.0
4 "S"        644            72.3           33.7
```

表一為存活(1)及死亡(0)的人數及比例
表二為各艙等人數、比例及存活率
表三為男女性人數、比例及存活率
表四為各登船港口人數、比例及存活率

五、數值變數描述

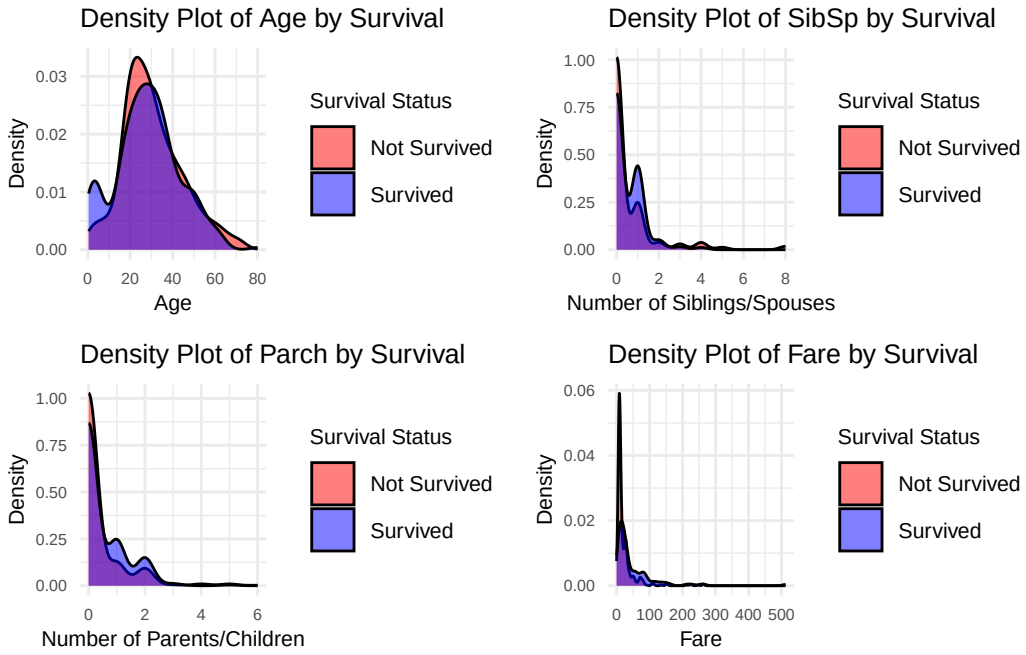
```
#Summary of Age, SibSp, Parch, Fare
numeric_summary <- summary(select(titanic, c(Age, SibSp, Parch, Fare)))
print(numeric_summary)
```

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

```
p1 <- ggplot(titanic, aes(x = Age, fill = as.factor(Survived))) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("red", "blue"),
                    labels = c("Not Survived", "Survived")) +
  labs(title = "Density Plot of Age by Survival",
       x = "Age", y = "Density", fill = "Survival Status") +
  theme_minimal()+
  theme(
    plot.title = element_text(size = 10),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    axis.text.x = element_text(size = 6),
    axis.text.y = element_text(size = 6),
```

```
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8),
  )

p2 <- ggplot(titanic, aes(x = SibSp, fill = as.factor(Survived))) +
  geom_density(alpha = 0.5, adjust = 1.5) +
  scale_fill_manual(values = c("red", "blue"),
                    labels = c("Not Survived", "Survived")) +
  labs(title = "Density Plot of SibSp by Survival",
       x = "Number of Siblings/Spouses", y = "Density", fill = "Survival Status") +
  theme_minimal()+
  theme(
    plot.title = element_text(size = 10),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    axis.text.x = element_text(size = 6),
    axis.text.y = element_text(size = 6),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8),
  )

p3 <- ggplot(titanic, aes(x = Parch, fill = as.factor(Survived))) +
  geom_density(alpha = 0.5, adjust = 1.5) +
  scale_fill_manual(values = c("red", "blue"),
                    labels = c("Not Survived", "Survived")) +
  labs(title = "Density Plot of Parch by Survival",
       x = "Number of Parents/Children", y = "Density", fill = "Survival Status") +
  theme_minimal()+
  theme(
    plot.title = element_text(size = 10),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    axis.text.x = element_text(size = 6),
    axis.text.y = element_text(size = 6),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8),
  )

p4 <- ggplot(titanic, aes(x = Fare, fill = as.factor(Survived))) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("red", "blue"),
                    labels = c("Not Survived", "Survived")) +
  labs(title = "Density Plot of Fare by Survival",
       x = "Fare", y = "Density", fill = "Survival Status") +
  theme_minimal()+
  theme(
    plot.title = element_text(size = 10),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    axis.text.x = element_text(size = 6),
    axis.text.y = element_text(size = 6),
    legend.title = element_text(size = 8),
    legend.text = element_text(size = 8),
```

```
  )

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



表五為年齡、同行兄弟姊妹或配偶數量、同行父母或子女數量、票價的五數摘要。
圖為各數值變數，依存活和死亡別的密度圖，
左上為年齡、右上為同行兄弟姊妹或配偶數量、
左下為同行父母或子女數量、右下為票價。