

Titanic Data Analysis (Homework1)

I-Hsuan Wu

2025-02-27

目錄

1. 讀取 Titanic 數據集，並查看數據結構	1
2. 視覺化	3
2.1 乘客人口統計分析	3
2.1.1 乘客年齡分布	3
2.1.2 年齡與性別的關聯性	4
2.2 生存率分析	5
2.2.1 艙等與生存率	5
2.2.2 性別與生存率	6
2.2.3 艙等、性別與生存率	7
2.2.4 年齡與生存率	7
2.3 其他影響因素	8
2.3.1 艙等與票價分布	8
2.3.2 登船港口與生存率	9
3. 統計描述與結論	10

□ 作業 GitHub 連結：[Statistical-Consulting_HW1](#)

1. 讀取 Titanic 數據集，並查看數據結構

```
df <- read.csv("C:/Users/user/Desktop/ /HW1/titanic.csv")
```

下表顯示了 Titanic 數據集的結構和基本統計資訊，包括變數的類型與連續型變數的數值分布。

```
str(df)
```

```
'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" ...
```

```
summary(df)
```

PassengerId	Survived	Pclass	Name
Min. : 1.0	Min. :0.0000	Min. :1.000	Length:891
1st Qu.:223.5	1st Qu.:0.0000	1st Qu.:2.000	Class :character
Median :446.0	Median :0.0000	Median :3.000	Mode :character
Mean :446.0	Mean :0.3838	Mean :2.309	
3rd Qu.:668.5	3rd Qu.:1.0000	3rd Qu.:3.000	
Max. :891.0	Max. :1.0000	Max. :3.000	

Sex	Age	SibSp	Parch
Length:891	Min. : 0.42	Min. :0.000	Min. :0.0000
Class :character	1st Qu.:20.12	1st Qu.:0.000	1st Qu.:0.0000
Mode :character	Median :28.00	Median :0.000	Median :0.0000
	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		

Ticket	Fare	Cabin	Embarked
Length:891	Min. : 0.00	Length:891	Length:891
Class :character	1st Qu.: 7.91	Class :character	Class :character
Mode :character	Median : 14.45	Mode :character	Mode :character
	Mean : 32.20		
	3rd Qu.: 31.00		
	Max. :512.33		

下面顯示了各類別變數的類別分布情況，例如生還人數、不同艙等乘客比例、性別比例等，以瞭解乘客的基本分佈。

```
table(df$Survived)
```

```
0 1
549 342
```

```
table(df$Pclass)
```

```
1 2 3
216 184 491
```

```
table(df$Sex)
```

```
female male
314 577
```

```
table(df$Embarked)
```

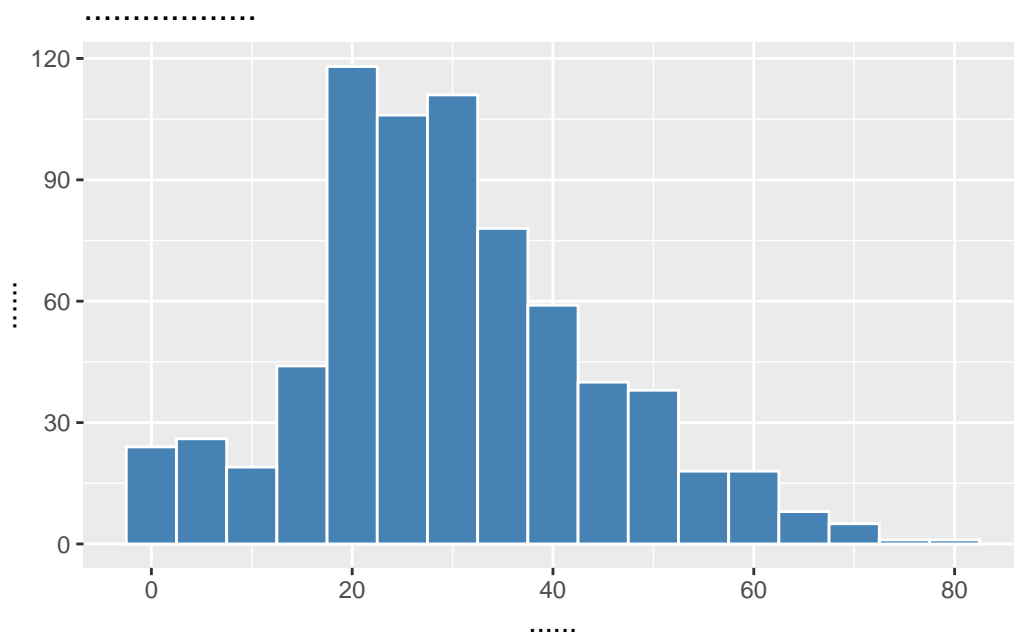
```
C Q S
2 168 77 644
```

2.視覺化

2.1乘客人口統計分析

2.1.1 乘客年齡分布

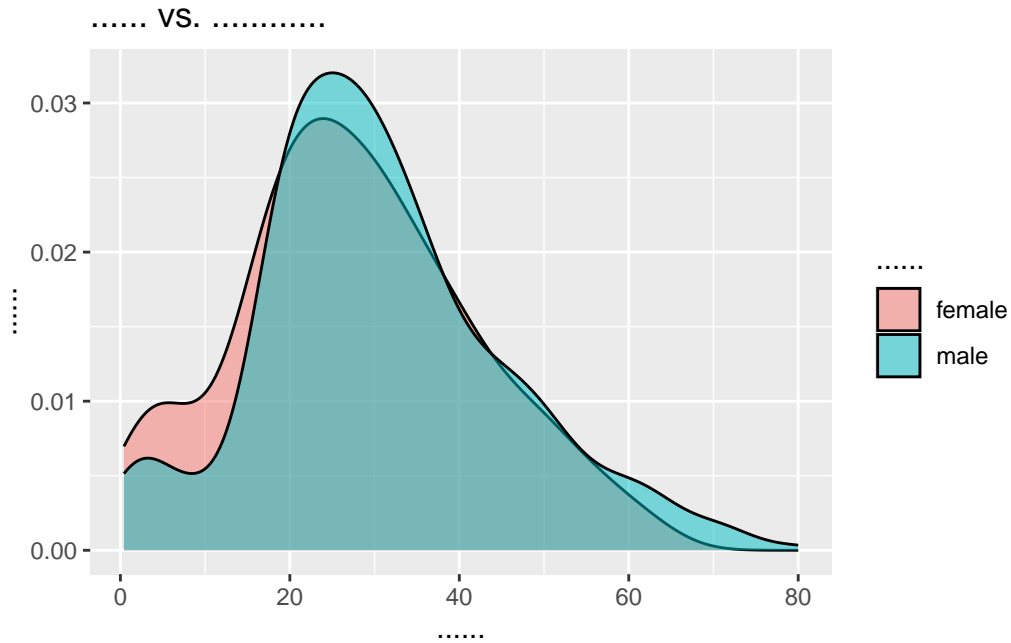
```
ggplot(df, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = " ", x = " ", y = " ")
```



大部分乘客的年齡集中在20-30歲之間，表示該年齡層佔多數。

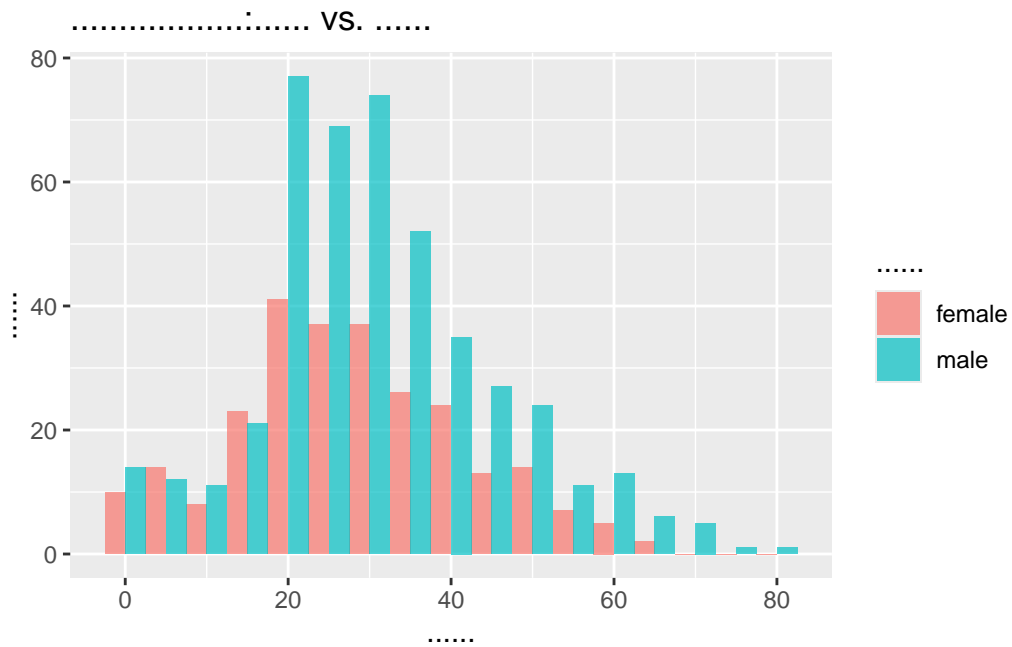
2.1.2 年齡與性別的關聯性

```
ggplot(df, aes(x = Age, fill = Sex)) +  
  geom_density(alpha = 0.5) +  
  labs(title = "    vs.    ", x = " ", y = " ", fill = " ")
```



此圖顯示男性與女性的年齡分布。若女性年齡整體較小，則可能影響其生存率較高的結果。

```
ggplot(df, aes(x = Age, fill = Sex)) +  
  geom_histogram(binwidth = 5, position = "dodge", alpha = 0.7) +  
  labs(title = "    vs.    ", x = " ", y = " ", fill = " ")
```

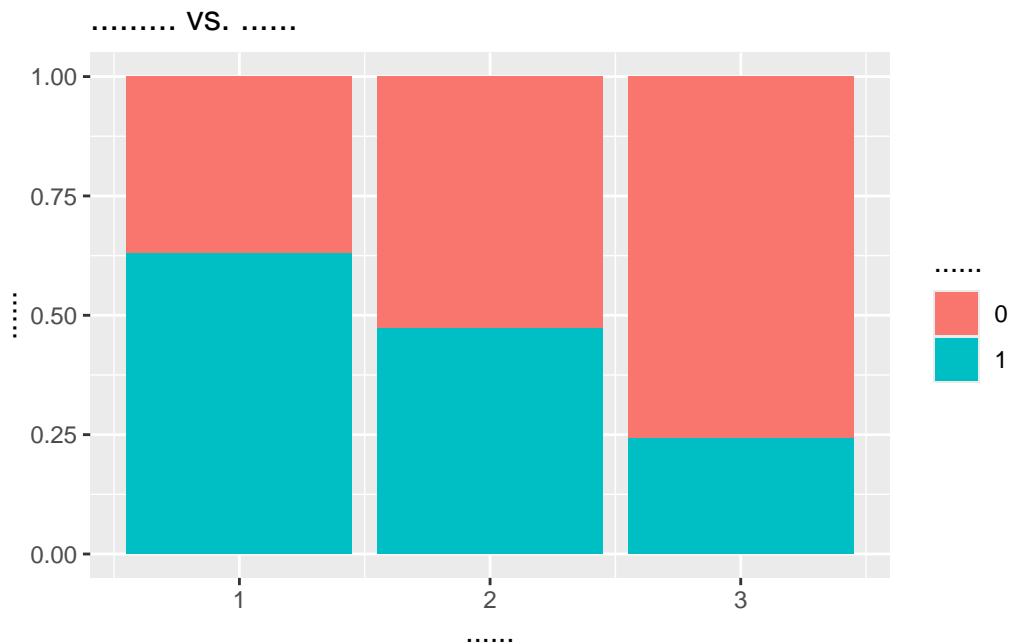


此圖比較不同年齡層的男女乘客人數，在20歲以後，男性的人數明顯高於女性，且在65歲以上，幾乎沒有女性。。

2.2 生存率分析

2.2.1 艙等與生存率

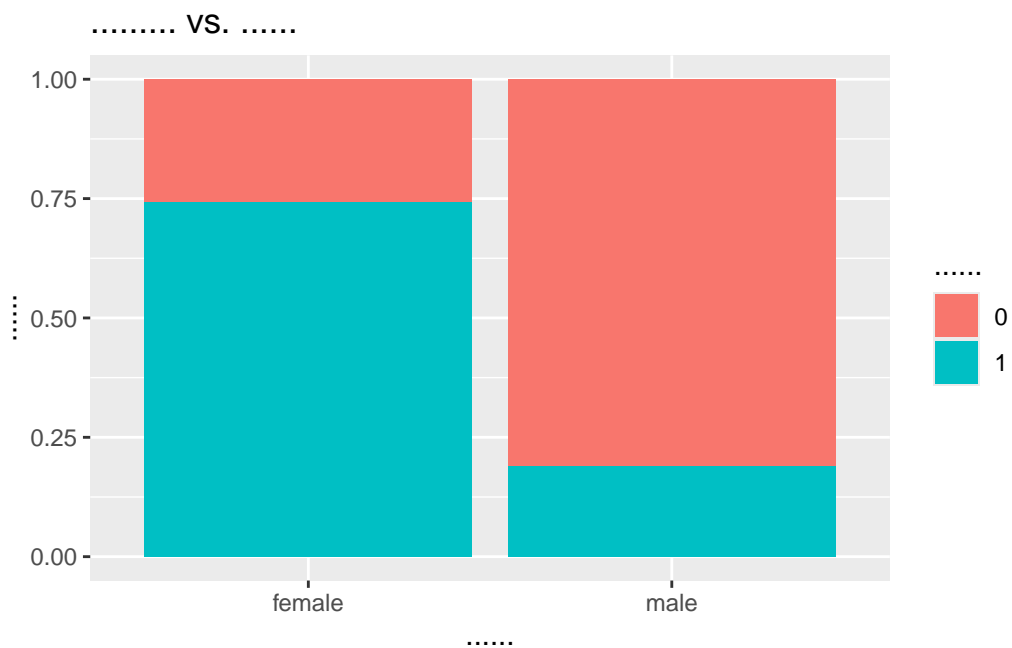
```
ggplot(df, aes(x = Pclass, fill = as.factor(Survived))) +
  geom_bar(position = "fill") +
  labs(title = " vs. ", y = " ", x = " ", fill = " ")
```



一等艙乘客的生存率最高，而三等艙的生存率最低，顯示艙等與生存機率有顯著關聯。

2.2.2 性別與生存率

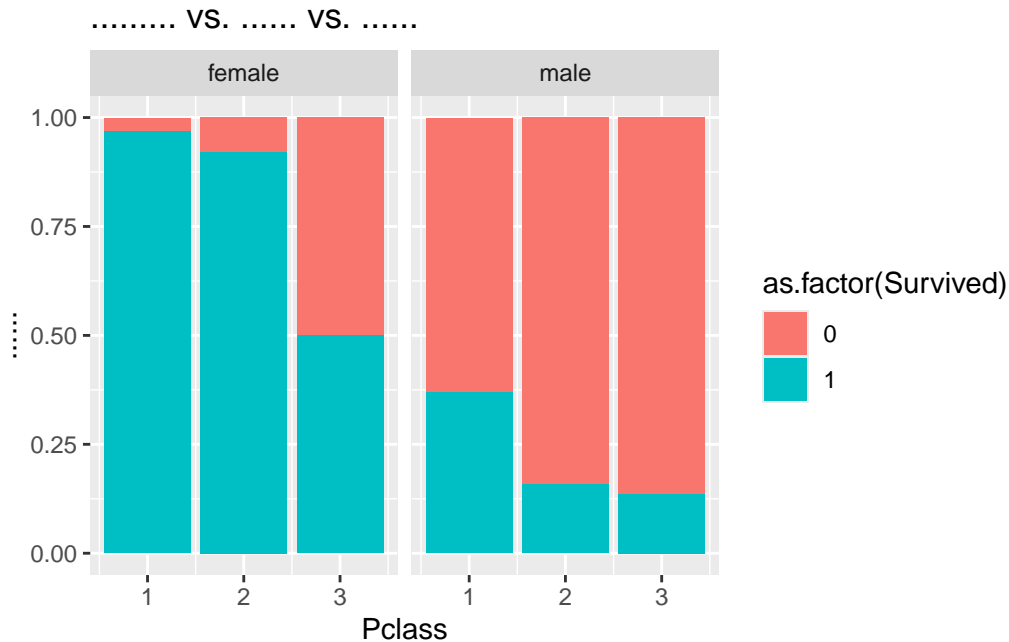
```
ggplot(df, aes(x = Sex, fill = as.factor(Survived))) +
  geom_bar(position = "fill") +
  labs(title = "    vs. ", y = " ", x = " ", fill = " ")
```



女性的生存率明顯高於男性。

2.2.3 艙等、性別與生存率

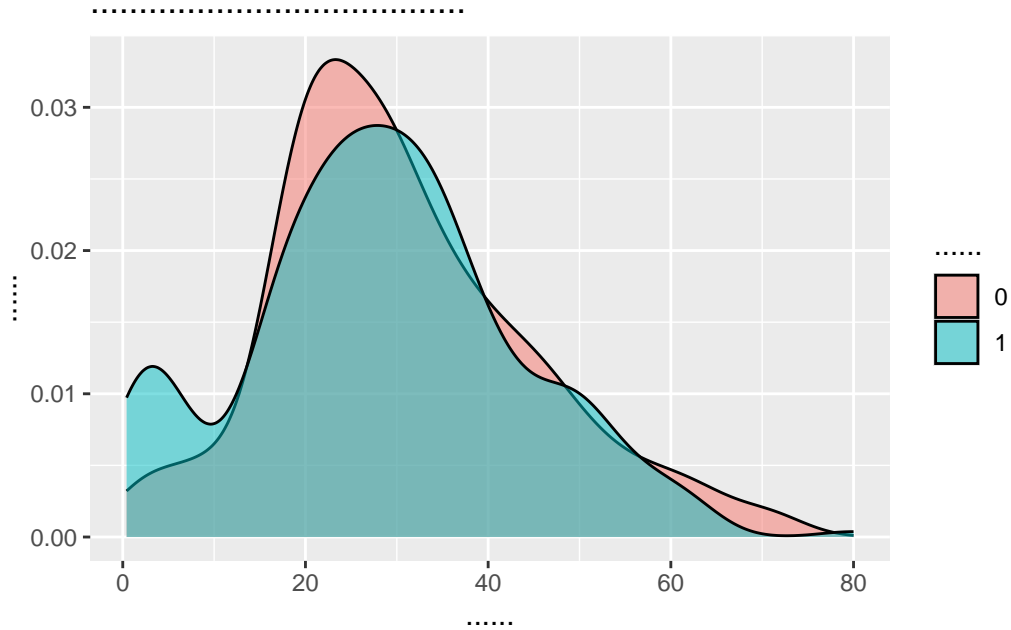
```
ggplot(df, aes(x = Pclass, fill = as.factor(Survived))) +  
  geom_bar(position = "fill") +  
  facet_wrap(~Sex) +  
  labs(title = "      vs.      vs.      ", y = " ")
```



不同性別與艙等的生存率顯示：女性在所有艙等的生存率都明顯高於男性，特別是一等艙女性生存率最高，而三等艙男性生存率最低。

2.2.4 年齡與生存率

```
ggplot(df, aes(x = Age, fill = as.factor(Survived))) +  
  geom_density(alpha = 0.5) +  
  labs(title = "      ", x = " ", y = " ", fill = " ")
```

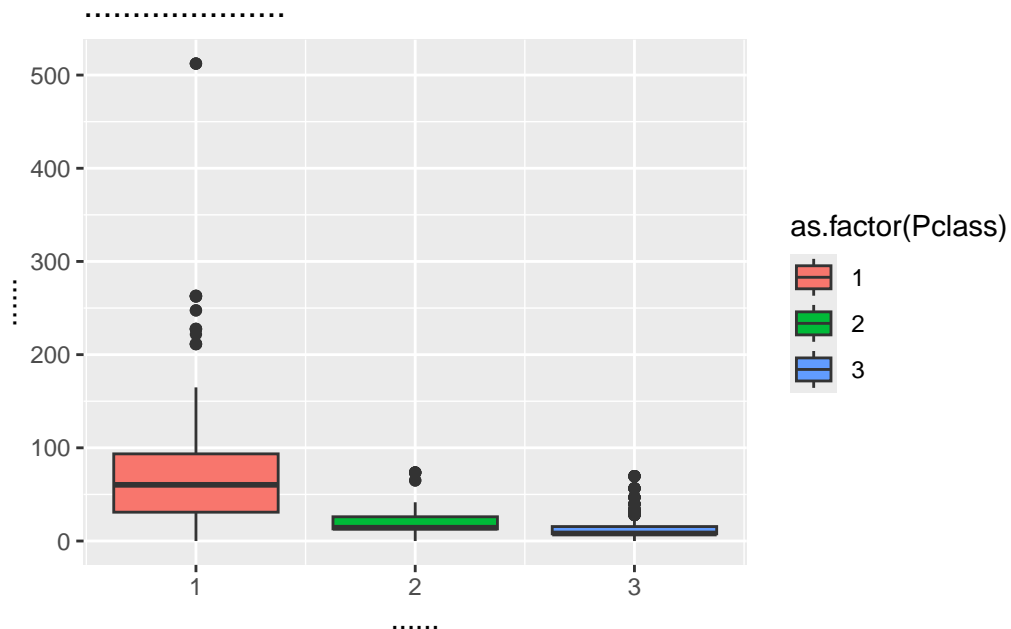


年齡與生存率的關聯顯示：年齡較小的乘客生存率較高，而年齡較大的乘客生存率則較低。

2.3 其他影響因素

2.3.1 艙等與票價分布

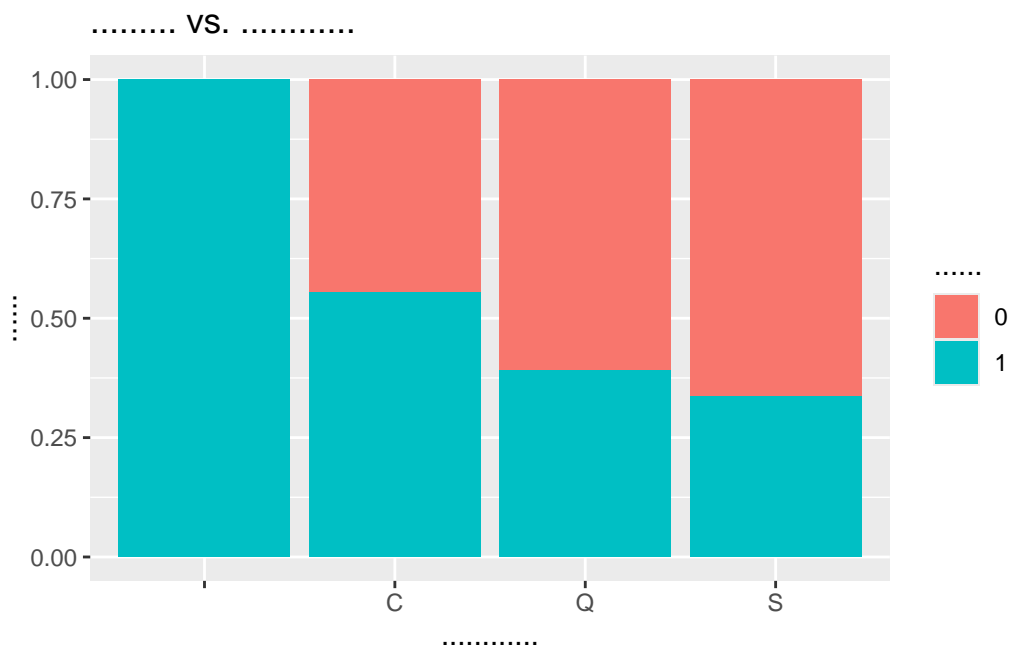
```
ggplot(df, aes(x = Pclass, y = Fare, fill = as.factor(Pclass))) +
  geom_boxplot() +
  labs(title = " ", x = " ", y = " ")
```

票價的箱型圖顯示，一等艙的票價高於二、三等艙，顯示艙等與票價之間的差異。

2.3.2 登船港口與生存率

```
ggplot(df, aes(x = Embarked, fill = as.factor(Survived))) +
  geom_bar(position = "fill") +
  labs(title = " vs. ", y = " ", x = " ", fill = " ")
```



從登船港口來看，從C港 (Cherbourg) 登船的乘客生存率最高。(因為未知的登船口資料從前面的類別結構來看只有兩筆，所以先忽略)

3. 統計描述與結論

- 年齡分布：大多數乘客年齡集中在20-30歲，且65歲以上幾乎都是男性乘客，孩童的生存率較高。
- 性別影響：女性的生存率顯著高於男性。
- 艙等影響：一等艙乘客的生存率遠高於二、三等艙。
- 票價與艙等：一等艙的票價高於二、三等艙。
- 登船港口影響：C港 (Cherbourg) 登船的乘客生存率最高。

以上結果顯示 Titanic 事故的生存機率可能受到多種因素影響，包括性別、艙等、年齡、票價與登船港口等。