

HW1

顧以恩 RE6131066

1 Titanic dataset

1.1 Import packages and read data

```
#install.packages("titanic")
library(titanic)
library(dplyr)
library(ggplot2)
library(GGally)

data(titanic_train)
head(titanic_train)
```

	PassengerId	Survived	Pclass
1	1	0	3
2	2	1	1
3	3	1	3
4	4	1	1
5	5	0	3
6	6	0	3

	Name	Sex	Age	SibSp	Parch
1	Braund, Mr. Owen Harris	male	22	1	0
2	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0

1 TITANIC DATASET

2

```

3                               Heikkinen, Miss. Laina female 26    0    0
4      Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35    1    0
5                               Allen, Mr. William Henry   male 35    0    0
6                               Moran, Mr. James           male NA    0    0

      Ticket    Fare Cabin Embarked
1      A/5 21171  7.2500           S
2      PC 17599 71.2833    C85      C
3 STON/O2. 3101282  7.9250           S
4      113803 53.1000    C123      S
5      373450  8.0500           S
6      330877  8.4583           Q

```

1.2 Variables description

```
summary(titanic_train)
```

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		

1.3 Variable description

- Survival: Survival status (0 = No 38%, 1 = Yes 62%)
- Pclass: Ticket class (Ordinal variable: 1 = 1st, 2 = 2nd, 3 = 3rd)
- Sex: Gender (Nominal variable: male = Male 65%, female = Female 35%)
- Age: Age (Continuous variable: 0.42 – 80, 177 missing values)
- SibSp: Number of siblings or spouses aboard the ship (Ordinal variable: 0 – 8)
- Parch: Number of parents or children aboard the ship (Ordinal variable: 0 – 6)
- Ticket: Ticket number
- Fare: Ticket price (Continuous variable: 0 – 512, no missing values)
- Cabin: Cabin number
- Embarked: Port of embarkation (C = Cherbourg 19%, Q = Queenstown 9%, S = Southampton 72%, 2 missing values)

1.4 Transformation

```
# Sex should be binary categorical variable
# Therefore transform it into factor
print(length(table(titanic_train$Sex)))
```

```
[1] 2
```

```
titanic_train$Sex = as.factor(titanic_train$Sex)

#Embarked is also categorical variable, transform it into factor
print(length(table(titanic_train$Embarked)))
```

```
[1] 4
```

```
titanic_train$Embarked = as.factor(titanic_train$Embarked)
titanic_train = titanic_train %>% mutate(Survived = as.factor(Survived),
                                         Sex = as.factor(Sex),
                                         #SibSp = as.factor(SibSp),
                                         #Parch = as.factor(Parch),
                                         Embarked = ifelse(Embarked=="",NA,Embarked) %>%
                                         )

# Name, Ticket and Cabin variables are all charactors
head(titanic_train$Name,5)
```

```
[1] "Braund, Mr. Owen Harris"
[2] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
[3] "Heikkinen, Miss. Laina"
[4] "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
[5] "Allen, Mr. William Henry"
```

```
head(titanic_train$Ticket,5)
```

```
[1] "A/5 21171"      "PC 17599"      "STON/O2. 3101282" "113803"
[5] "373450"
```

```
head(titanic_train$Cabin,5)
```

```
[1] ""      "C85"  ""      "C123" ""
```

```
titanic_train = titanic_train %>% select(-c(PassengerId, Name, Ticket, Cabin))
```

```
summary(titanic_train)
```

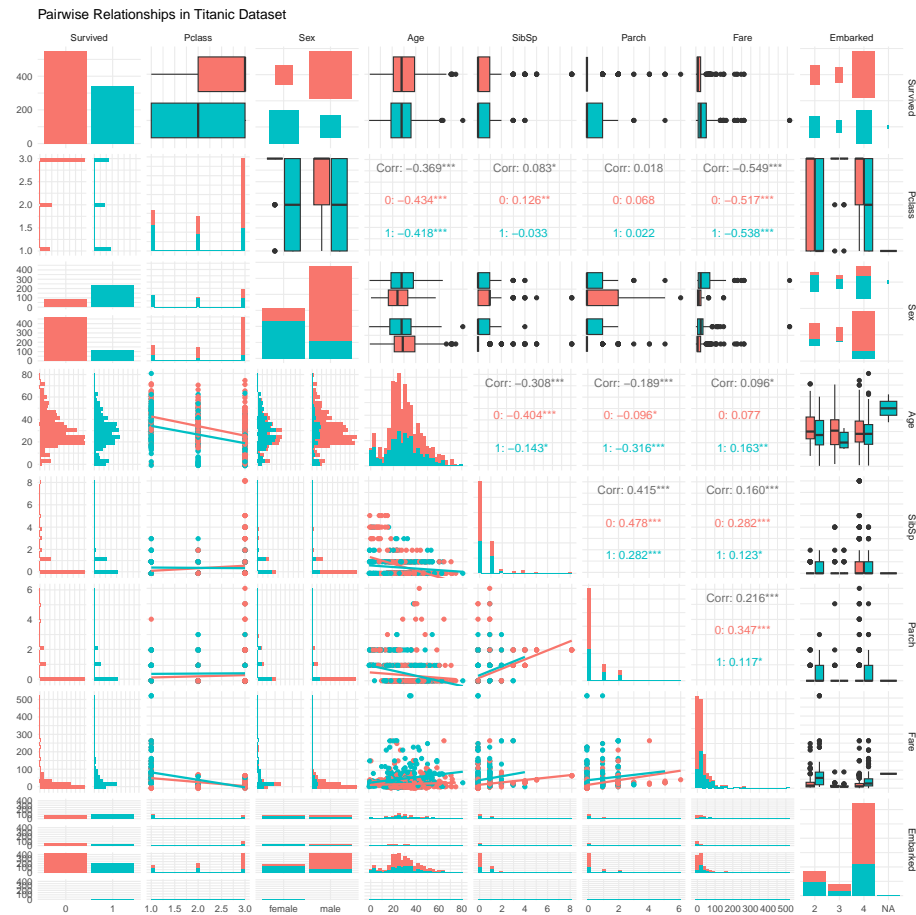
Survived	Pclass	Sex	Age	SibSp
0:549	Min. :1.000	female:314	Min. : 0.42	Min. :0.000
1:342	1st Qu.:2.000	male :577	1st Qu.:20.12	1st Qu.:0.000
	Median :3.000		Median :28.00	Median :0.000
	Mean :2.309		Mean :29.70	Mean :0.523
	3rd Qu.:3.000		3rd Qu.:38.00	3rd Qu.:1.000
	Max. :3.000		Max. :80.00	Max. :8.000
			NA's :177	
Parch	Fare	Embarked		
Min. :0.0000	Min. : 0.00	2 :168		
1st Qu.:0.0000	1st Qu.: 7.91	3 : 77		
Median :0.0000	Median : 14.45	4 :644		
Mean :0.3816	Mean : 32.20	NA's: 2		
3rd Qu.:0.0000	3rd Qu.: 31.00			
Max. :6.0000	Max. :512.33			

1.5 Visuallization

```

ggpairs(
  titanic_train,
  title = "Pairwise Relationships in Titanic Dataset",
  lower = list(continuous = wrap("smooth", method = "lm", se = FALSE)), # 下三角加入回歸
  upper = list(continuous = wrap("cor", size = 4)), # 上三角顯示相關係數
  diag = list(continuous = "barDiag"), # 對角線繪製直方圖
  mapping = aes(color = Survived) # 顏色區分存活
) +
  theme_minimal()

```

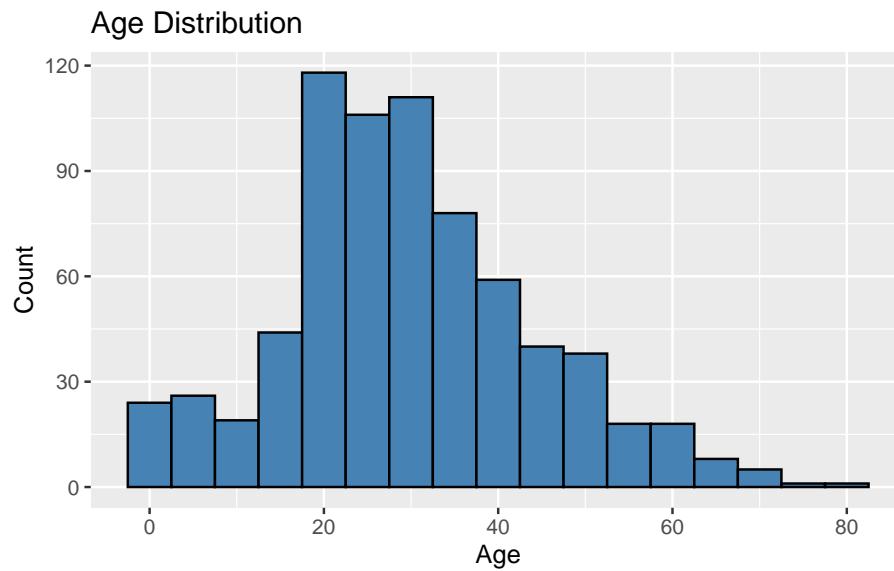


1.6 Discussion

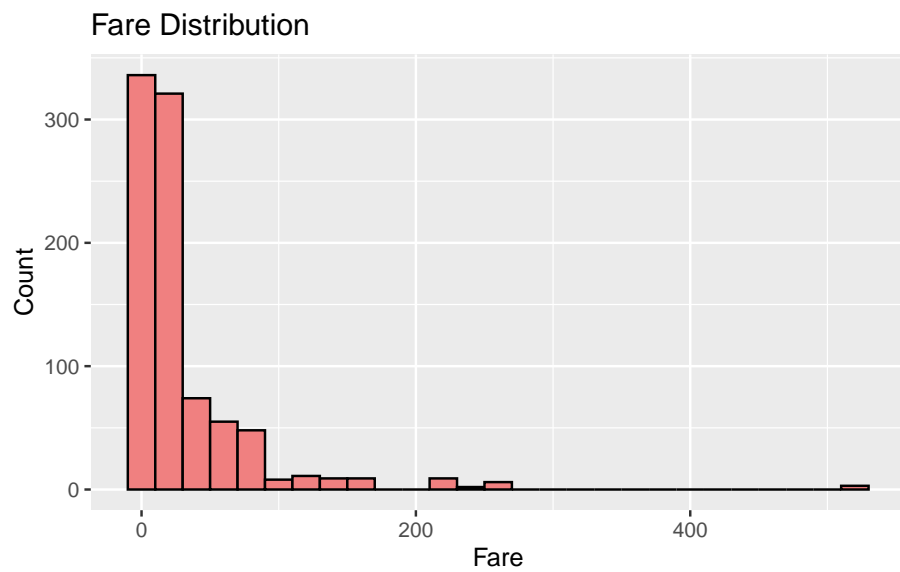
- a) Survival Analysis (Survived) More people did not survive (Survived = 0, shown in red). The survival rate seems to vary significantly with Sex and Pclass.
- b) Passenger Class (Pclass) First-class (Pclass = 1) passengers had a higher fare and survival rate. Third-class (Pclass = 3) passengers had the lowest survival rate.
- c) Gender Influence (Sex) Female passengers had a higher survival rate than males. Males had a lower chance of survival, as indicated by the red proportion in the Sex vs. Survived plots.
- d) Age Distribution (Age) Young children had a higher survival rate. The majority of passengers were adults. There is a negative correlation between Age and Survival (older passengers had a lower chance of survival).
- e) Family Members (SibSp and Parch) Having more siblings/spouses (SibSp) or parents/children (Parch) slightly increases survival probability, but very high numbers reduce survival chances. Large families had a lower survival rate compared to single passengers.
- f) Fare and Survival (Fare) Higher Fare is positively correlated with survival (Corr = 0.538***), indicating that wealthier passengers had better chances. This aligns with the observation that first-class passengers survived more.
- g) Embarkation Port (Embarked) Most passengers embarked from Southampton (S). Higher survival rates are observed among passengers from Cherbourg (C), likely due to a higher proportion of first-class passengers.

1.7 Appendix

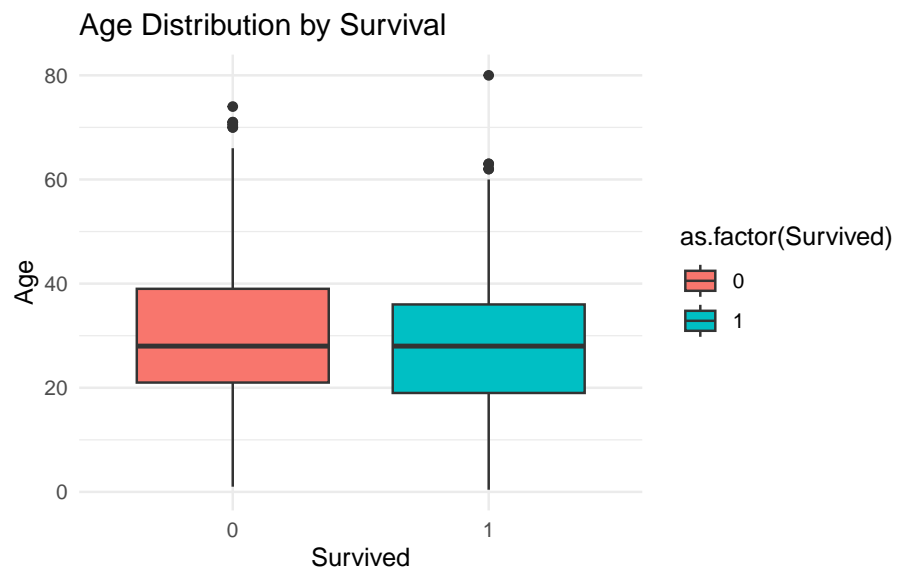
```
ggplot(titanic_train, aes(x = Age)) +  
  geom_histogram(binwidth = 5, fill = "steelblue", color = "black") +  
  labs(title = "Age Distribution", x = "Age", y = "Count")
```



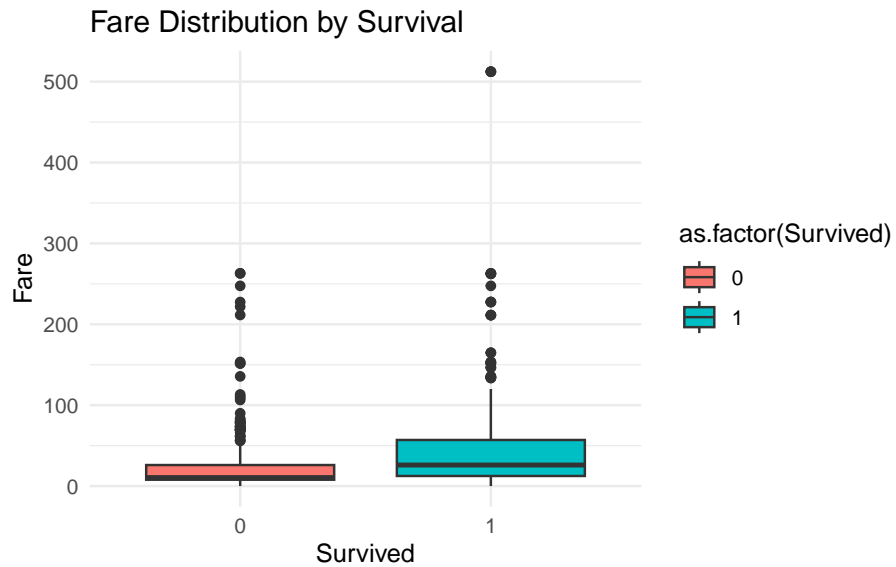
```
ggplot(titanic_train, aes(x = Fare, fill = as.factor(Survived))) +  
  geom_histogram(binwidth = 20, fill = "lightcoral", color = "black") +  
  labs(title = "Fare Distribution", x = "Fare", y = "Count")
```

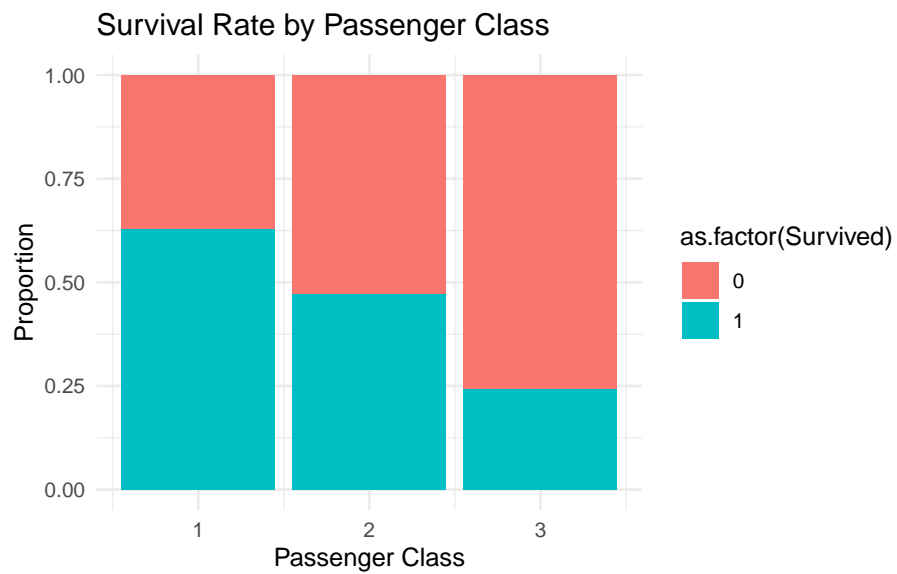
```
ggplot(titanic_train, aes(x = as.factor(Survived), y = Age, fill = as.factor(Survived)))  
  geom_boxplot() +  
  labs(title = "Age Distribution by Survival", x = "Survived", y = "Age") +  
  theme_minimal()
```



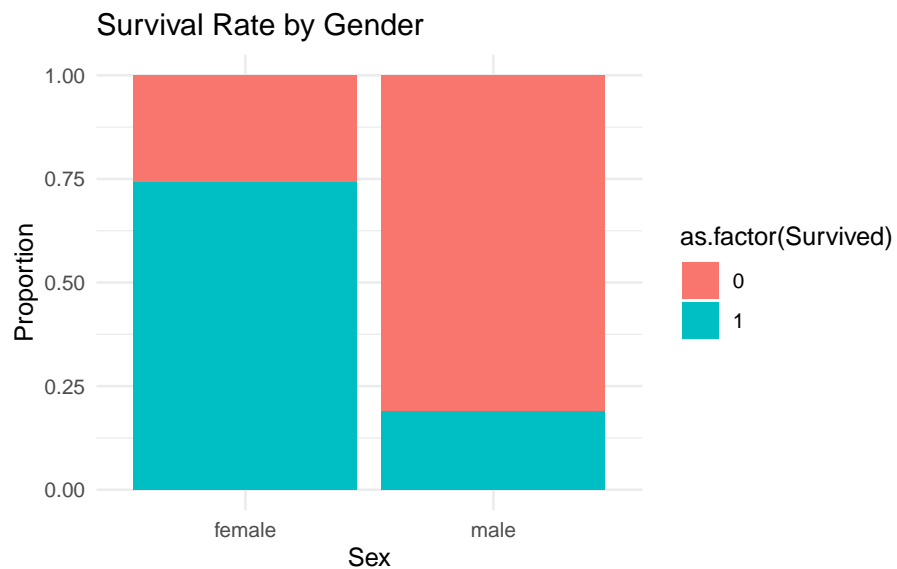
```
ggplot(titanic_train, aes(x = as.factor(Survived), y = Fare, fill = as.factor(Survived))) +  
  geom_boxplot() +  
  labs(title = "Fare Distribution by Survival", x = "Survived", y = "Fare") +  
  theme_minimal()
```



```
ggplot(titanic_train, aes(x = Pclass, fill = as.factor(Survived))) +  
  geom_bar(position = "fill") +  
  labs(title = "Survival Rate by Passenger Class", x = "Passenger Class", y = "Proportion") +  
  theme_minimal()
```



```
ggplot(titanic_train, aes(x = Sex, fill = as.factor(Survived))) +  
  geom_bar(position = "fill") +  
  labs(title = "Survival Rate by Gender", x = "Sex", y = "Proportion") +  
  theme_minimal()
```



```
library(ggcorrplot)

cor_matrix <- cor(titanic_train[, sapply(titanic_train, is.numeric)], use = "complete.obs")
ggcorrplot(cor_matrix, lab = TRUE)
```

