HW_2

2025-04-14

Question 2

1

```
corona_df <- read.csv("data_corona_world.csv")
head(corona_df)
```

```
##     Country Date.of.reporting Age.group Confirmed.cases Confirmed.fatalities
## 1 Argentina         28-May-20       0-9            1002                    0
## 2 Argentina         28-May-20     10-19            1080                    1
## 3 Argentina         28-May-20     20-29            2813                    1
## 4 Argentina         28-May-20     30-39            3142                    9
## 5 Argentina         28-May-20     40-49            2508                   24
## 6 Argentina         28-May-20     50-59            1812                   54
```

2

```
filterd_df <- subset(corona_df, Country %in% c("Italy", "China"))
head(filterd_df)
```

```
##    Country Date.of.reporting Age.group Confirmed.cases Confirmed.fatalities
## 10   China         17-Feb-20       0-9             416                    0
## 11   China         17-Feb-20     10-19             549                    1
## 12   China         17-Feb-20     20-29            3619                    7
## 13   China         17-Feb-20     30-39            7600                   18
## 14   China         17-Feb-20     40-49            8571                   38
## 15   China         17-Feb-20     50-59           10008                  130
```

3

```
str(filterd_df)
```

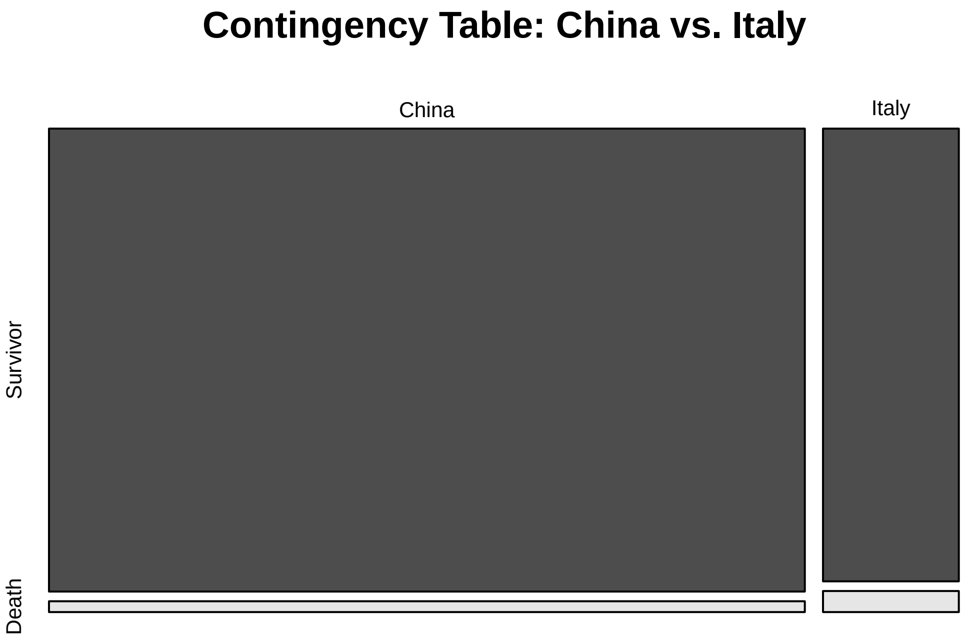
```
## ’data.frame':    18 obs. of  5 variables:
##  $ Country             : chr  "China" "China" "China" "China" ...
##  $ Date.of.reporting   : chr  "17-Feb-20" "17-Feb-20" "17-Feb-20" "17-Feb-20" ...
##  $ Age.group           : chr  "0-9" " 10-19" "20-29" "30-39" ...
##  $ Confirmed.cases     : int  416 549 3619 7600 8571 10008 8583 3918 1408 43 ...
##  $ Confirmed.fatalities: int  0 1 7 18 38 130 309 312 208 0 ...
```

4

```
china_data <- subset(filterd_df, Country == 'China',
                     select = c(Confirmed.cases, Confirmed.fatalities))
china_data <- colSums(china_data)
china_survivors <- china_data["Confirmed.cases"] - china_data["Confirmed.fatalities"]

italy_data <- subset(filterd_df, Country == 'Italy',
                     select = c(Confirmed.cases, Confirmed.fatalities))
italy_data <- colSums(italy_data)
italy_survivors <- italy_data["Confirmed.cases"] - italy_data["Confirmed.fatalities"]

contingency_table <- rbind(
  c(china_survivors, china_data["Confirmed.fatalities"]),
  c(italy_survivors, italy_data["Confirmed.fatalities"])
)

rownames(contingency_table) <- c("China","Italy")
colnames(contingency_table) <- c("Survivor","Death")

print(contingency_table)
```

```
##       Survivor Death
## China    43649  1023
## Italy     7669   357
```

```
mosaicplot(contingency_table, color = TRUE,
           main = "Contingency Table: China vs. Italy")
```



5

```
china_dp <- (contingency_table["China", "Death"] / sum(contingency_table["China", ])) * 100
print(china_dp)
```

```
## [1] 2.290025
```

```
italy_dp <- (contingency_table["Italy", "Death"] / sum(contingency_table["Italy", ])) * 100
print(italy_dp)
```

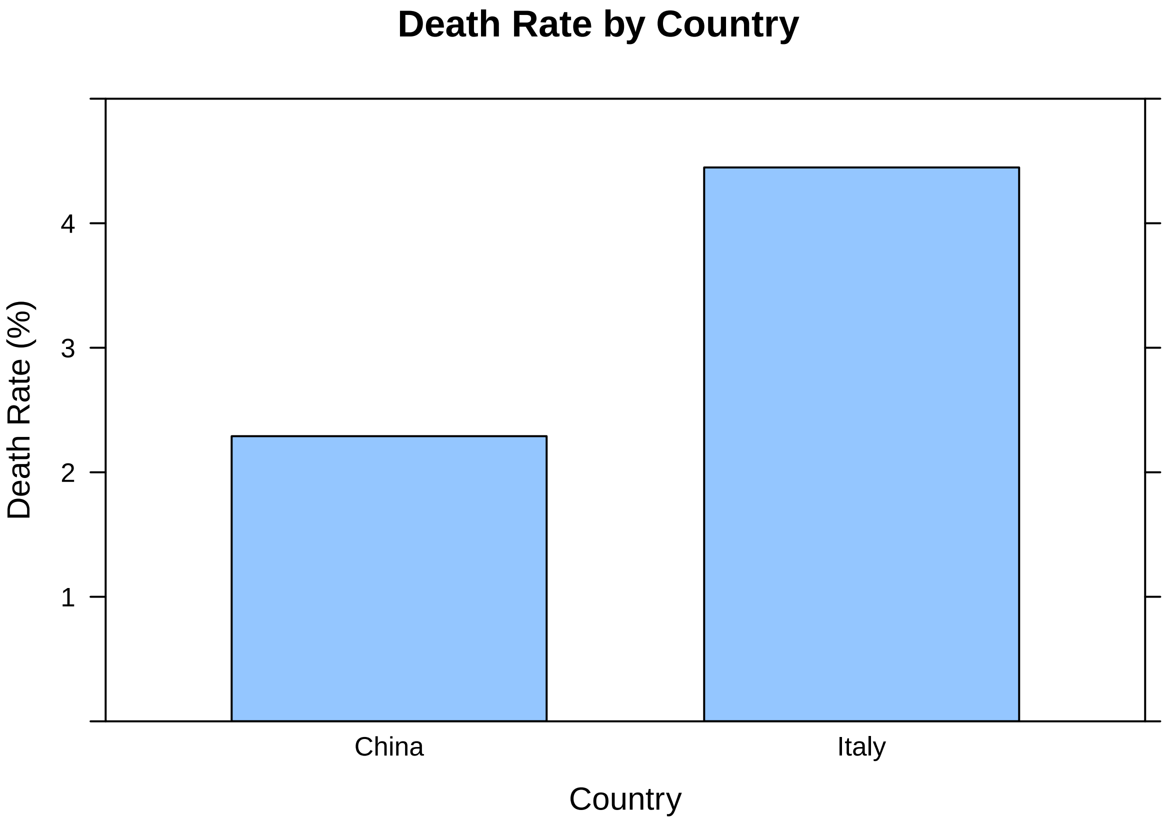
```
## [1] 4.448044
```

It seems that China handled the situation better, since the death rate was lower compared to Italy.

6

```
death_rate_df <- data.frame(
  Country = c("China", "Italy"),
  DeathRate = c(china_dp, italy_dp)
)
```

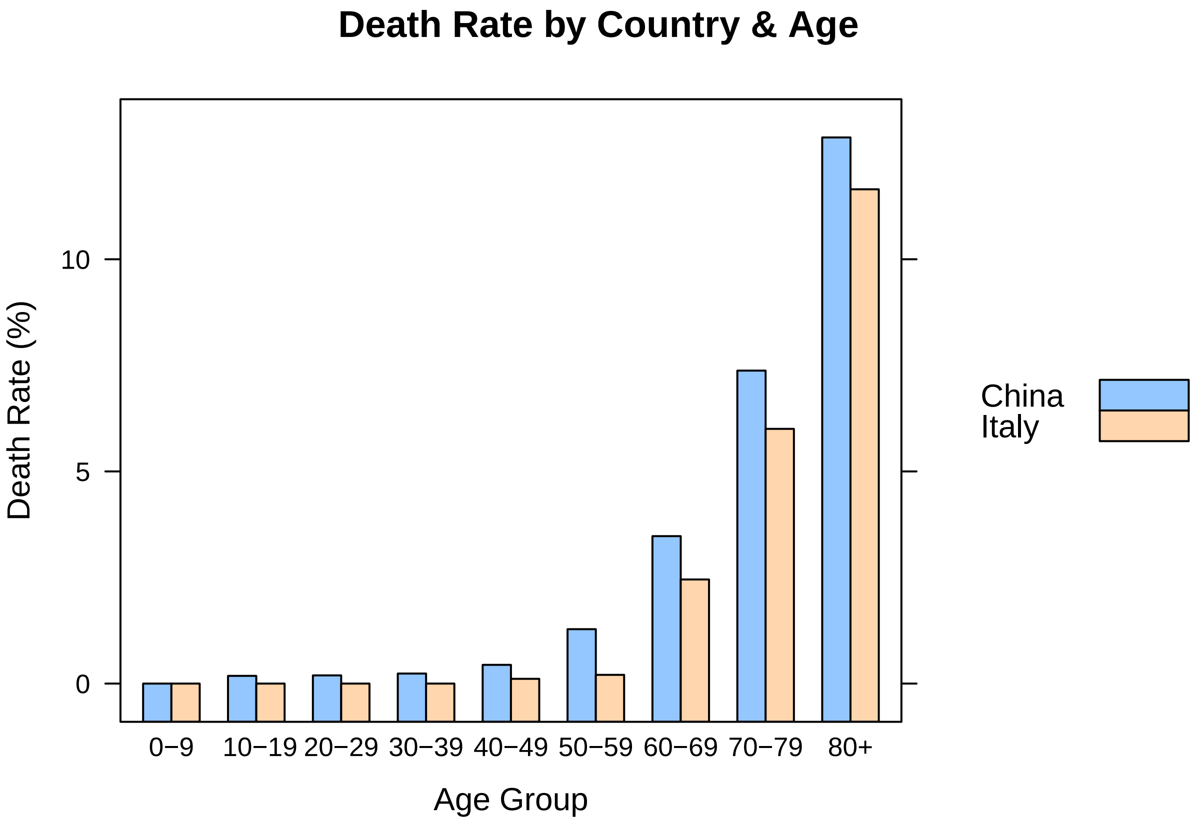
```r
library(lattice)
barchart(DeathRate ~ Country,
         data = death_rate_df,
         main = "Death Rate by Country",
         xlab = "Country",
         ylab = "Death Rate (%)",
         ylim = c(0,5)
         )
```



Based on the graph, being in Italy appears to be more risky.

```r
df_age_china <- subset(filterd_df, Country == 'China',
                 select = c(Age.group, Confirmed.cases, Confirmed.fatalities))

df_age_china$DeathRatio <- (df_age_china$Confirmed.fatalities / (df_age_china$Confirmed.cases + df_age_
df_age_china$Country <- "China"


df_age_italy <- subset(filterd_df, Country == 'Italy',
                       select = c(Age.group, Confirmed.cases, Confirmed.fatalities))

df_age_italy$DeathRatio <- (df_age_italy$Confirmed.fatalities / (df_age_italy$Confirmed.cases + df_age_
df_age_italy$Country <- "Italy"
```

```
death_rate_age <- rbind(df_age_china, df_age_italy)

death_rate_age$Age.group <- factor(
  death_rate_age$Age.group,
  levels = c("0-9", " 10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79", "80+"),
  ordered = TRUE
)


barchart(DeathRatio ~ Age.group,
         group = Country,
         data = death_rate_age,
         auto.key = list(space = "right"),
         main = "Death Rate by Country & Age",
         xlab = "Age Group",
         ylab = "Death Rate (%)"
         )
```



This is an example of Simpson's paradox, because when we look at each age group separately, China's death rate was higher than Italy's.

Question 3

```
df_titanic <- read.csv("titanic.csv")
head(df_titanic)
```

```
##   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
## 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
```

a

```
logic_vec <- !colnames(df_titanic) %in% c("Cabin", "Ticket")

df_titanic <- df_titanic[,logic_vec]
head(df_titanic)
```

```
##   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
## 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
##      Fare Embarked
## 1  7.2500        S
## 2 71.2833        C
## 3  7.9250        S
## 4 53.1000        S
## 5  8.0500        S
## 6  8.4583        Q
```

b

```
completed_rows <- complete.cases(df_titanic)

df_titanic <- df_titanic[completed_rows, ]
head(df_titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 7           7        0      1
##                                                   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
## 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
## 7                              McCarthy, Mr. Timothy J   male  54     0     0
##      Fare Embarked
## 1  7.2500        S
## 2 71.2833        C
## 3  7.9250        S
## 4 53.1000        S
## 5  8.0500        S
## 7 51.8625        S
```

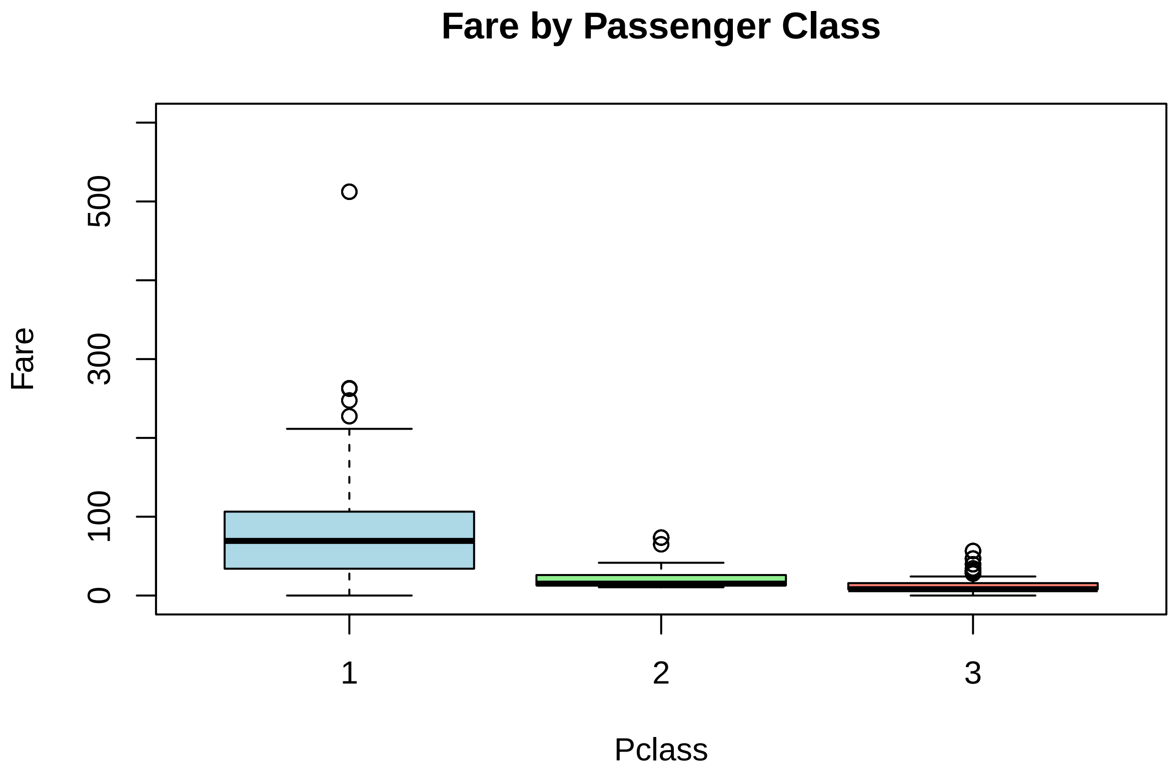
c

```
fare_summary <- aggregate(Fare ~ Pclass, data = df_titanic, FUN = summary)
print(fare_summary)
```

```
##   Pclass Fare.Min. Fare.1st Qu. Fare.Median Fare.Mean Fare.3rd Qu. Fare.Max.
## 1      1   0.00000     34.17915    69.30000  87.96158    106.42500 512.32920
## 2      2  10.50000     13.00000    15.04580  21.47156     26.00000  73.50000
## 3      3   0.00000      7.77500     8.05000  13.22944     15.74170  56.49580
```

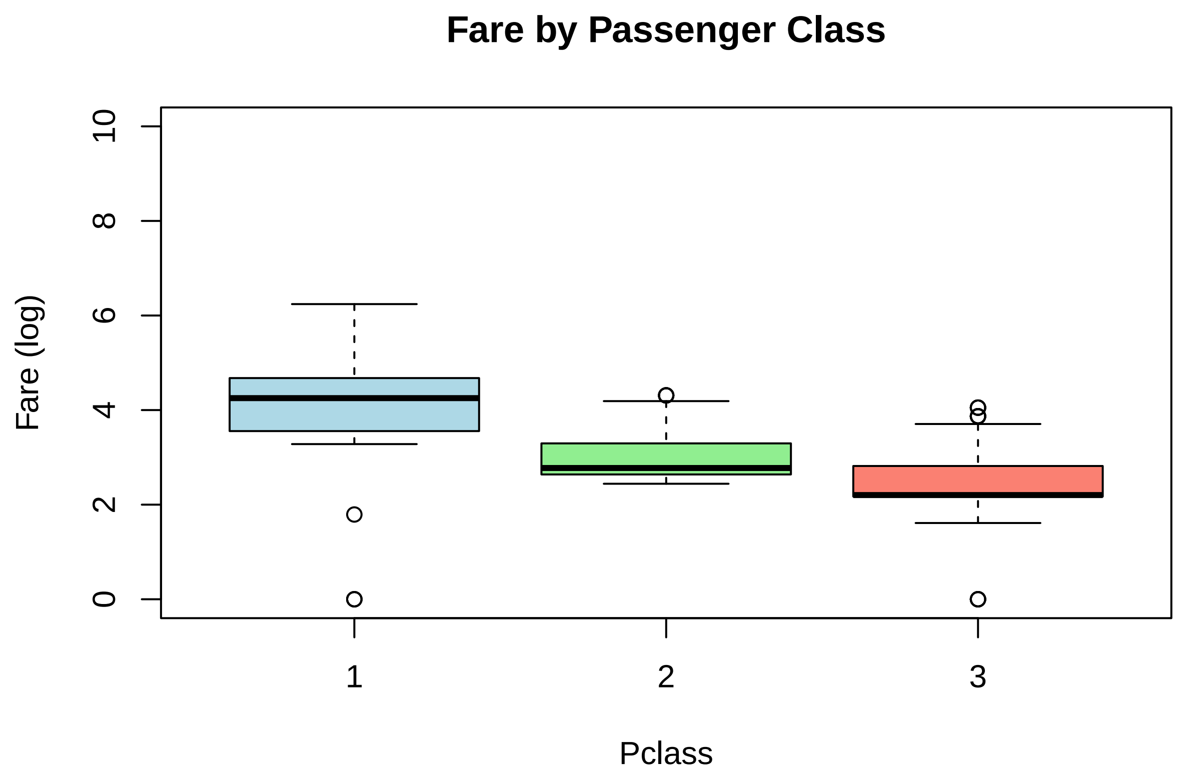
According to the summary, the median and mean for each class show that Class 1 is the most expensive.

```
boxplot(Fare ~ Pclass, data = df_titanic,
        main = "Fare by Passenger Class",
        xlab = "Pclass", ylab = "Fare",
        col = c("lightblue", "lightgreen", "salmon"),
        ylim = c(0,600))
```



We should use a transformation, since there are outliers that make the graph harder to read. We chose a log transformation.

```
df_titanic$logFare <- log(df_titanic$Fare + 1)

boxplot(logFare ~ Pclass, data = df_titanic,
        main = "Fare by Passenger Class",
        xlab = "Pclass", ylab = "Fare (log)",
        col = c("lightblue", "lightgreen", "salmon"),
        ylim = c(0,10)
        )
```



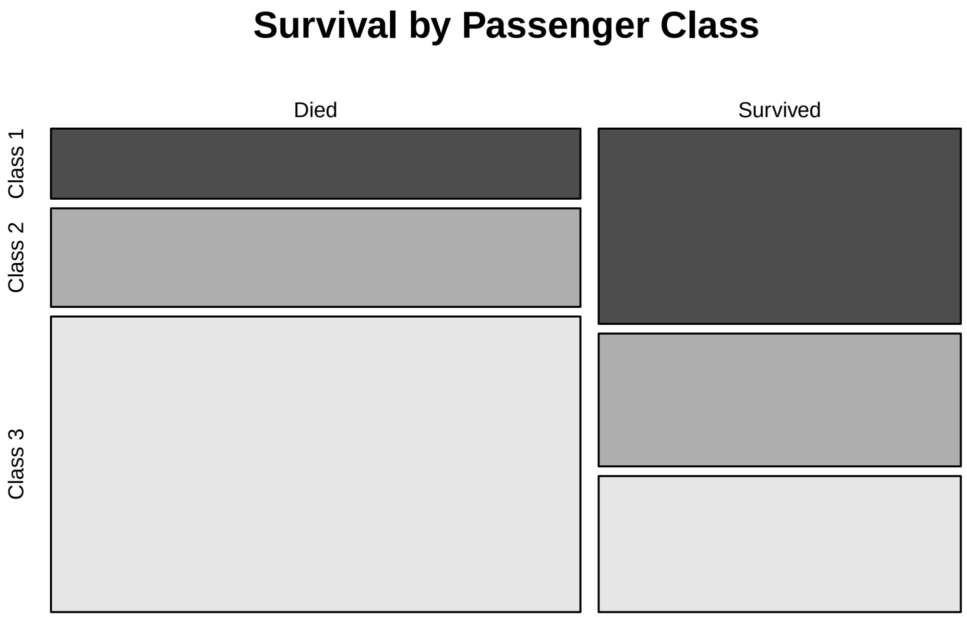
```
variance_by_class <- aggregate(Fare ~ Pclass, data = df_titanic, FUN = var)
print(variance_by_class)
```

```
##   Pclass      Fare
## 1      1 6537.8850
## 2      2  173.9083
## 3      3  100.8650
```

Class 1 shows higher fare variance, meaning there was a wider range in ticket prices.

d

```
table_survival <- table(df_titanic$Survived, df_titanic$Pclass)
rownames(table_survival) <- c("Died", "Survived")
colnames(table_survival) <- c("Class 1", "Class 2", "Class 3")
mosaicplot(table_survival,
           color = TRUE,
           main = "Survival by Passenger Class"
           )
```



d.1

The marginal totals are calculated as

```
colSums(table_survival)
```

```
## Class 1 Class 2 Class 3
##     186     173     355
```

```
rowSums(table_survival)
```

```
##     Died Survived
##      424      290
```

d.2

Expected frequency

```
row_total <- c(424, 290) # Died, Survived
col_total <- c(186, 173, 355) # Class 1,2,3
total <- sum(row_total)

expected <- matrix(0, nrow = 2, ncol = 3)
```

```r
for (i in 1:2) {
  for (j in 1:3) {
    expected[i,j] <- (row_total[i] * col_total[j]) / total
  }
}

rownames(expected) <- c("Died", "Survived")
colnames(expected) <- c("Class 1", "Class 2", "Class 3")

print(expected)
```

```
##            Class 1   Class 2  Class 3
## Died     110.45378 102.73389 210.8123
## Survived  75.54622  70.26611 144.1877
```

d.3

The X^2 statistic is

```r
x <- 0
for (i in 1:2) {
  for (j in 1:3) {
    x <- x + ((table_survival[i,j] - expected[i,j]) ^ 2) / expected[i,j]
  }
}

print(x)
```

```
## [1] 92.90142
```

d.4

e

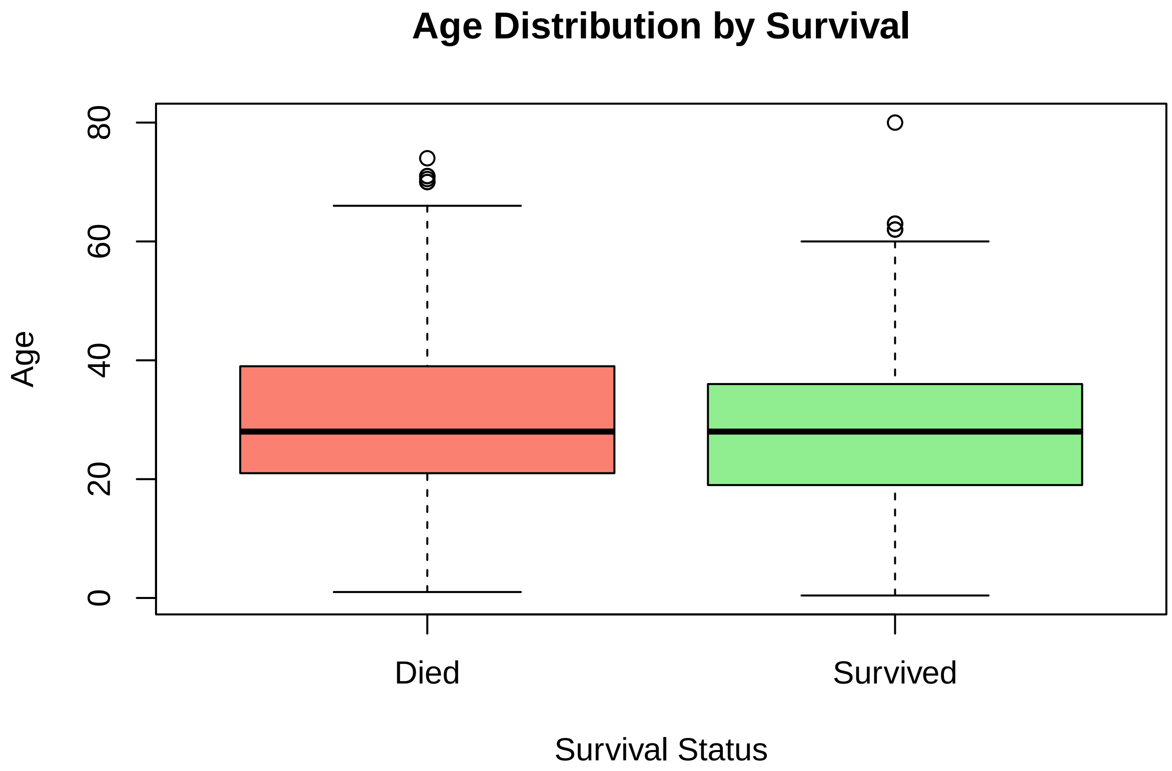
```r
survival_ratio <- table_survival[2, ] / (table_survival[1, ] + table_survival[2, ])
print(survival_ratio)
```

```
##   Class 1   Class 2   Class 3
## 0.6559140 0.4797688 0.2394366
```

The proportions clearly show that Class 1 had the highest chance of survival

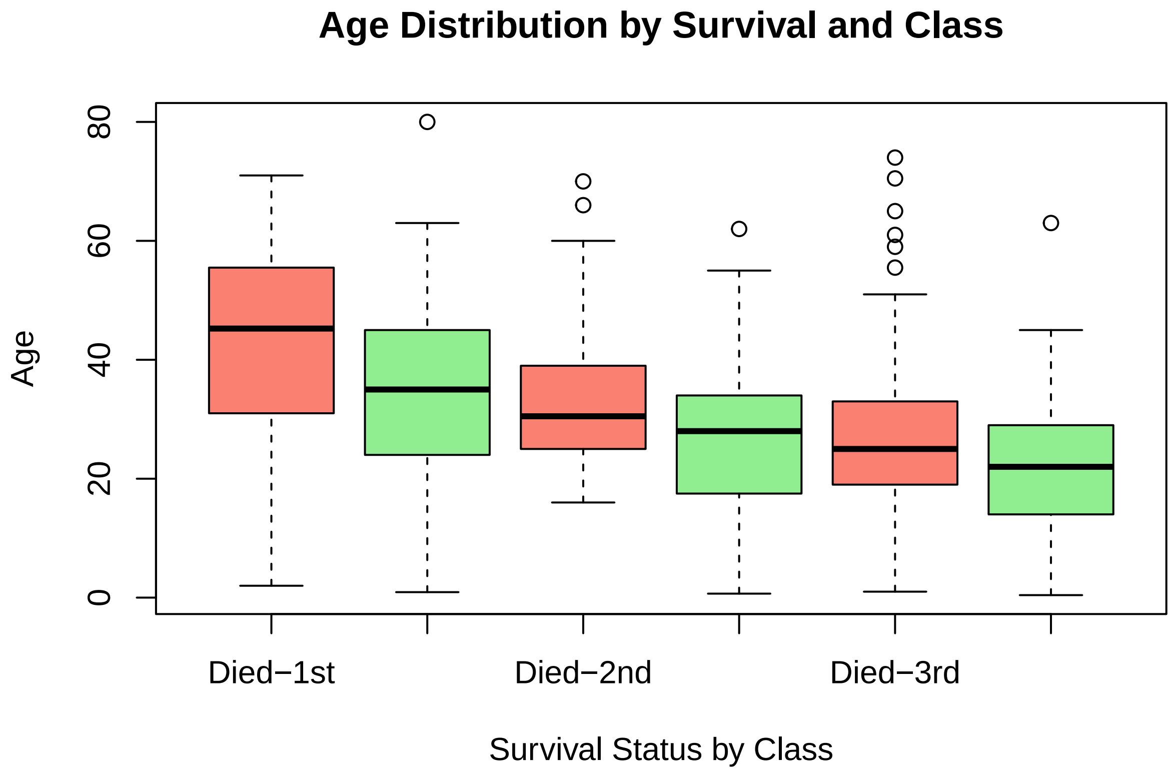
f

```
boxplot(Age ~ Survived,
        names = c("Died", "Survived"),
        data = df_titanic,
        ylab = "Age",
        xlab = "Survival Status",
        col = c("salmon", "lightgreen"),
        main = "Age Distribution by Survival"
        )
```



Based on the graph, there doesn't seem to be a connection between survival and age, since the medians look very similar and the Q1 and Q3 are at about the same level.

```
boxplot(Age ~ Survived + Pclass,
        data = df_titanic,
        names = c("Died-1st", "Survived-1st",
                  "Died-2nd", "Survived-2nd",
                  "Died-3rd", "Survived-3rd"),
        col = c("salmon", "lightgreen"),
        xlab = "Survival Status by Class",
        ylab = "Age",
        main = "Age Distribution by Survival and Class"
        )
```



However, when separating by class, it became apparent that younger passengers were more likely to survive within each class.