Data Visualization and Analytic

Haruna Jallow Reg No. MD300-00006/2022

2023-10-30

ASSIGNMENT SOLUTIONS

Load the necessary library for reading Excel files

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load the Excel dataset (replace 'your_data.xlsx' with your file path)
Titanic_data <- read_excel("C:\\Users\\Haruna\\Desktop\\Data Science Materials\\Second_Semester 2023\\Da
attach(Titanic_data)
# View(Titanic_data)
```

QUESTION 1

```
# Calculate the total number of passengers
total_passengers <- nrow(Titanic_data)
cat("Total passengers:", total_passengers)
```

```
## Total passengers: 156
```

```
# Calculate the number of survivors
survivors <- sum(Titanic_data$Survived == 1)
cat("Survivors:", survivors)
```

```
## Survivors: 54
```

```
# Calculate the number of non-survivors
non_survivors <- sum(Titanic_data$Survived == 0)
cat("Non-survivors:", non_survivors)
```

```
## Non-survivors: 102
```

```r
non_survivors1 <- total_passengers - survivors
cat("Non-survivors:", non_survivors1)
```

```
## Non-survivors: 102
```

```r
# Calculate the overall survival rate
survival_rate <- (survivors / total_passengers) * 100
cat("Overall Survival Rate:", survival_rate, "%")
```

```
## Overall Survival Rate: 34.61538 %
```

QUESTION 2

```r
# Calculate the total number of passengers by gender
total_males <- sum(Titanic_data$Sex == "male")
cat("Total Males:", total_males)
```

```
## Total Males: 100
```

```r
total_females <- sum(Titanic_data$Sex == "female")
cat("Total Females:", total_females)
```

```
## Total Females: 56
```

```r
# Calculate the number of passengers by gender within each ticket class
males_by_class <- table(Titanic_data$Pclass[Titanic_data$Sex == "male"])
cat("Number of Males by Ticket Class:")
```

```
## Number of Males by Ticket Class:
```

```r
print(males_by_class)
```

```
##
##  1  2  3
## 21 18 61
```

```r
females_by_class <- table(Titanic_data$Pclass[Titanic_data$Sex == "female"])
cat("Number of Females by Ticket Class:")
```

```
## Number of Females by Ticket Class:
```

```r
print(females_by_class)
```

```
##
##  1  2  3
##  9 12 35
```

QUESTION 3

```r
# Calculate the number of passengers of each sex who survived
male_survivors <- sum(Titanic_data$Survived[Titanic_data$Sex == "male"])
cat("Male Survivors are:", male_survivors)
```

```
## Male Survivors are: 14
```

```r
female_survivors <- sum(Titanic_data$Survived[Titanic_data$Sex == "female"])
cat("Female Survivors are:", female_survivors)
```

```
## Female Survivors are: 40
```

```
# Calculate the number of passengers of each sex who did not survive
male_non_survivors <- sum(Titanic_data$Survived[Titanic_data$Sex == "male"] == 0)
cat("Male Non-Survivors are:", male_non_survivors)
```

```
## Male Non-Survivors are: 86
```

```
female_non_survivors <- sum(Titanic_data$Survived[Titanic_data$Sex == "female"] == 0)
cat("Female Non-Survivors are:", female_non_survivors)
```

```
## Female Non-Survivors are: 16
```

```
# Calculate the survival rate for passengers of each sex
survival_rate_male <- (male_survivors / total_males) * 100
cat("Male Survival Rate is:", survival_rate_male, "%")
```

```
## Male Survival Rate is: 14 %
```

```
survival_rate_female <- (female_survivors / total_females) * 100
cat("Female Survival Rate is:", survival_rate_female, "%")
```

```
## Female Survival Rate is: 71.42857 %
```

QUESTION 4

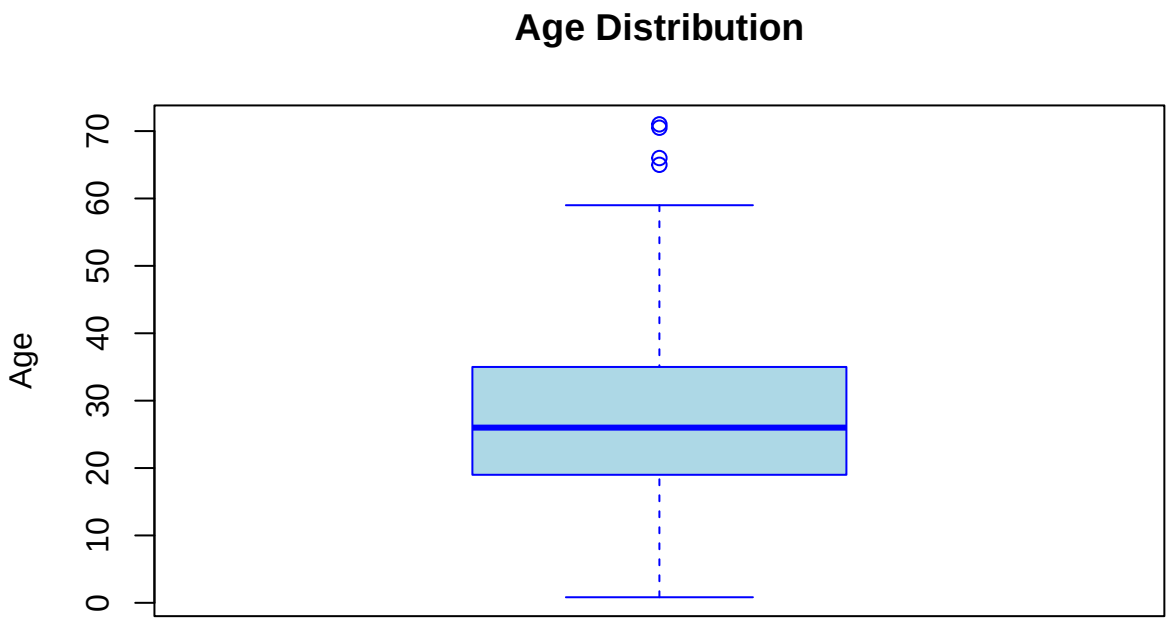
```
# Calculate the number of passengers with age information
passengers_with_age <- sum(!is.na(Titanic_data$Age))
cat("Passengers with Age Information are:", passengers_with_age)
```

```
## Passengers with Age Information are: 126
```

```
# Calculate the number of passengers with missing age information
passengers_missing_age <- sum(is.na(Titanic_data$Age))
cat("Passengers with Missing Age Information are:", passengers_missing_age)
```

```
## Passengers with Missing Age Information are: 30
```

```
# Generate a boxplot to visualize the age distribution
boxplot(Titanic_data$Age[!is.na(Titanic_data$Age)],
        main = "Age Distribution",
        ylab = "Age",
        col = "lightblue",
        border = "blue",
        horizontal = FALSE)
```

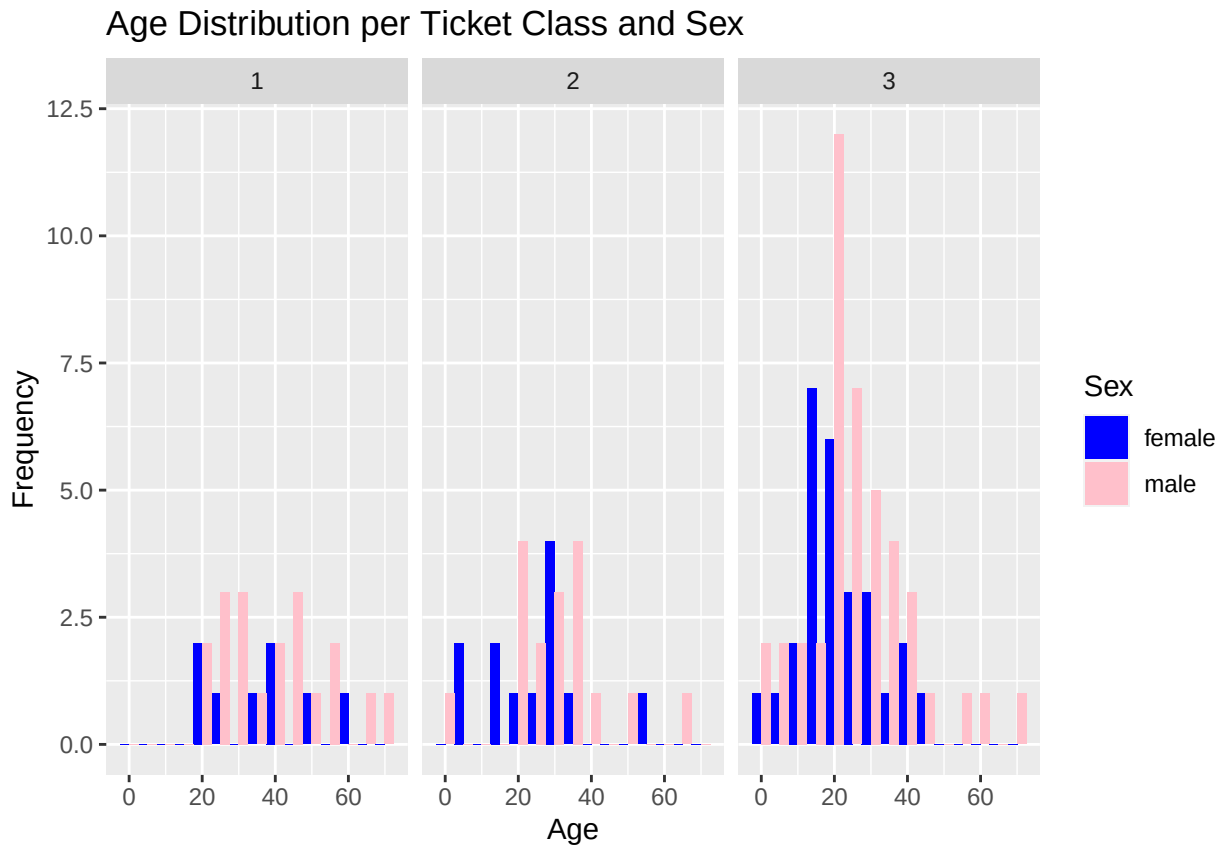


The boxplot shows that the median age of the passengers was around 28 years, and the interquartile range (IQR) was from 20 to 38 years.

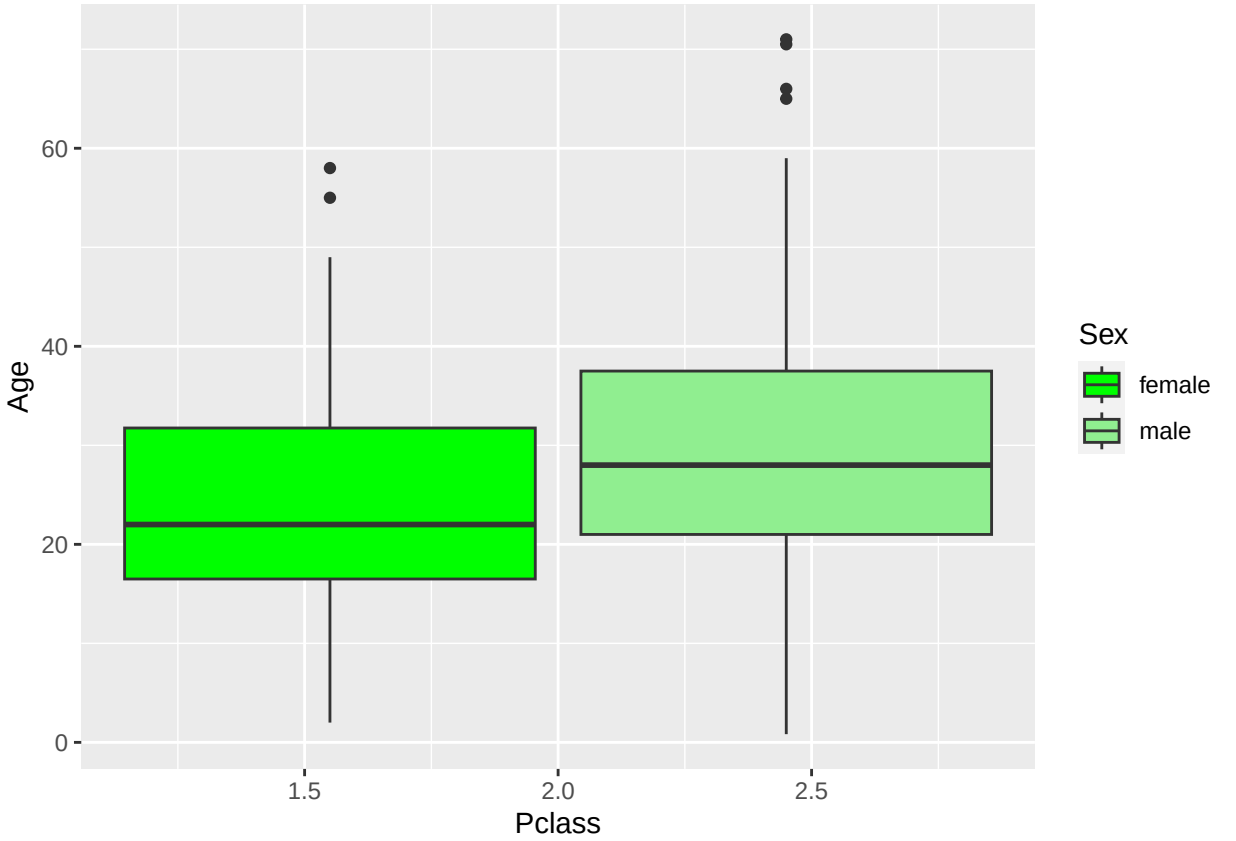
The boxplot shows some outliers at one end of the age distribution, indicating very old passengers on board.

QUESTION 5

```r
# Load the ggplot2 library for creating visualizations
library(ggplot2)

# Create a subset of the data to filter out missing age information
subset_data <- Titanic_data[!is.na(Titanic_data$Age),]

# Create a histogram of age distribution per ticket class and per sex
age_distribution_plot <- ggplot(subset_data, aes(x = Age, fill = Sex)) +
  geom_histogram(binwidth = 5, position = "dodge") +
  facet_wrap(~Pclass) +
  labs(title = "Age Distribution per Ticket Class and Sex",
       x = "Age", y = "Frequency") +
  scale_fill_manual(values = c("blue", "pink"))

# Display the age distribution plot
print(age_distribution_plot)
```




```r
# Create boxplot
ggplot(data = subset_data, mapping = aes(x = Pclass, y = Age, fill = Sex)) +
  geom_boxplot() +
  scale_fill_manual(values = c("green", "lightgreen"))
```



The histogram and boxplot both shows that most passengers were between 20 and 40 years old, with a median age of around 28 years.

The first-class passengers were older than the second and third-class passengers, with a higher median age and more outliers at one end of the age distribution.

The male passengers were older than the female passengers in each class, with a higher median age and more outliers at one end of the age distribution.

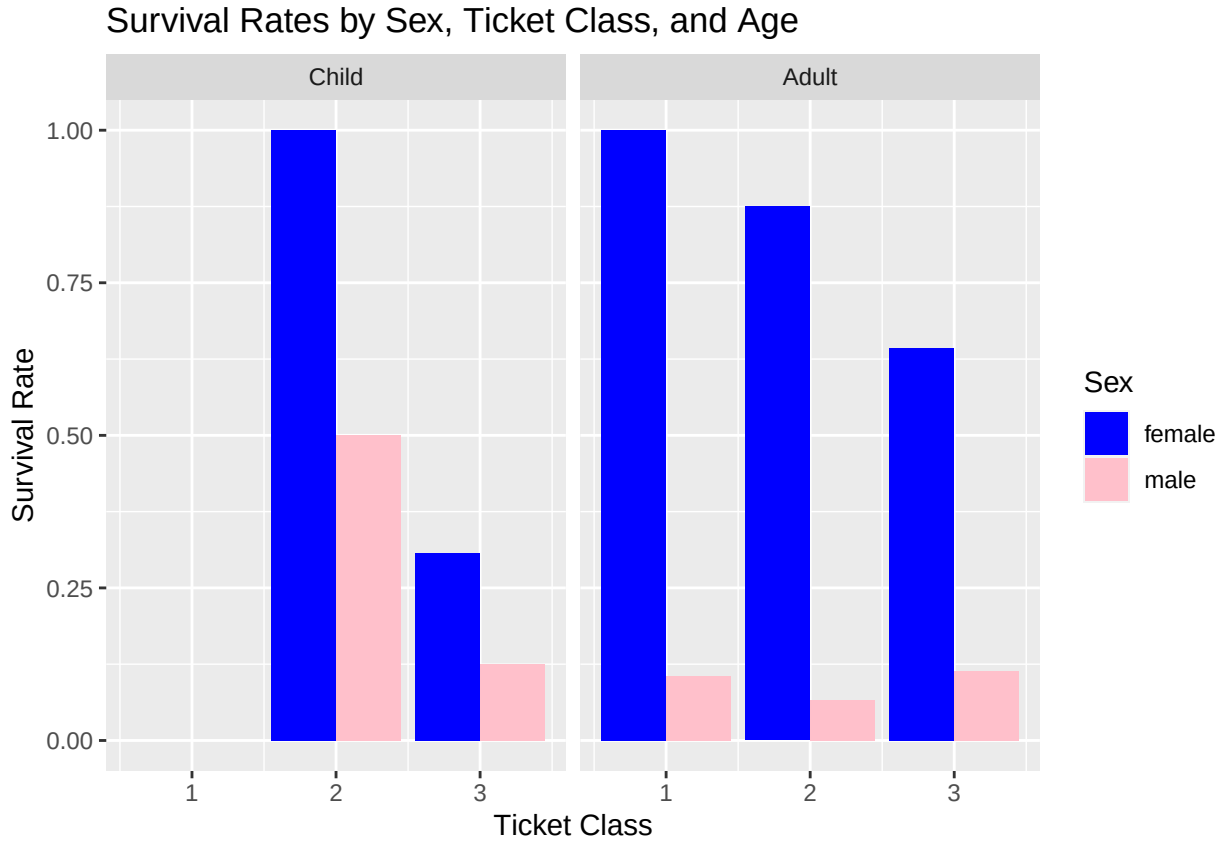
The third class had more passengers than the first and second class but also had more missing values in the age column.

The age distribution was skewed to the right for all groups, meaning that there were more younger passengers than older passengers.

QUESTION 6

```r
# Create a subset of the data to filter out missing age information
subset_data <- Titanic_data[!is.na(Titanic_data$Age),]

# Calculate the survival rates by grouping based on Sex, Pclass, and Age
survival_rates <- aggregate(Survived ~ Sex + Pclass + (Age > 18), data = subset_data, FUN = mean)

# Rename the columns for clarity
colnames(survival_rates) <- c("Sex", "Pclass", "Adult", "SurvivalRate")

# Convert the "Adult" column to a factor
```

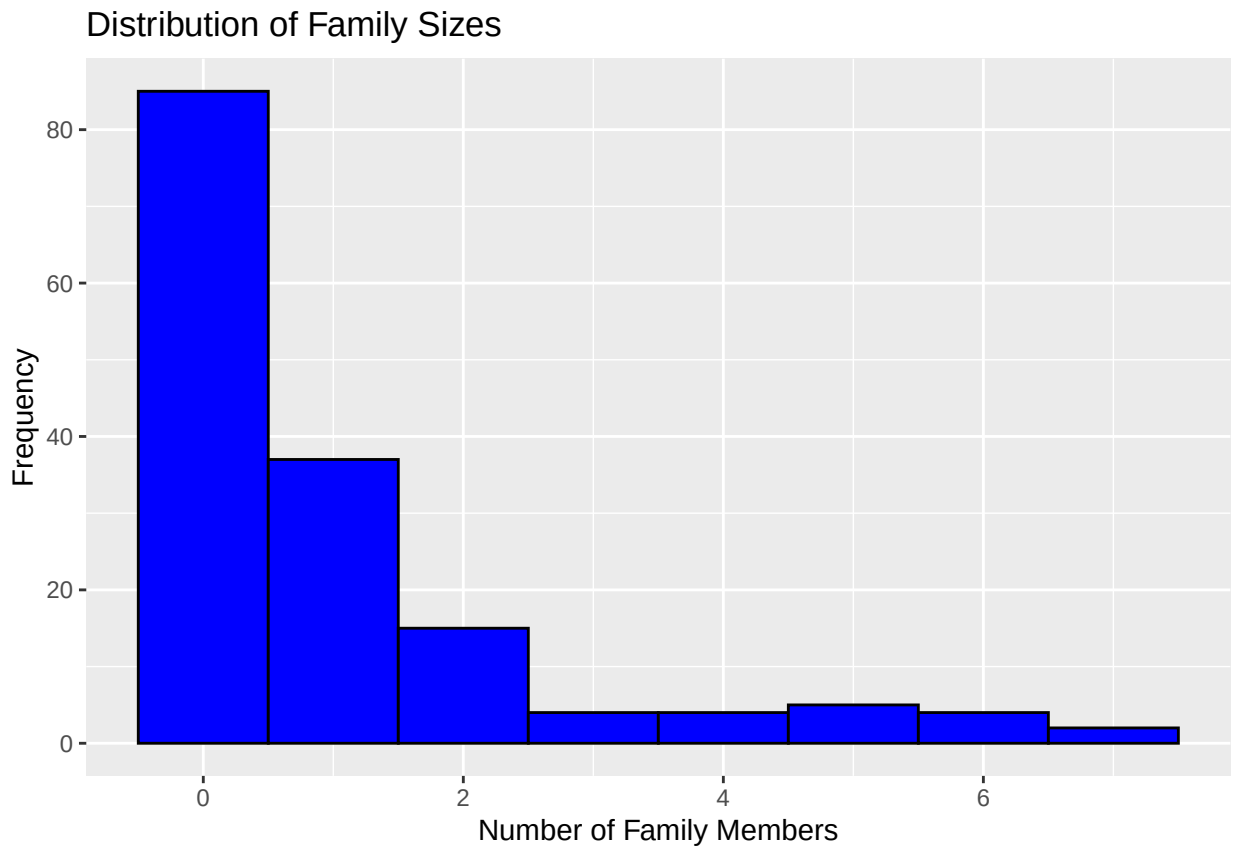
```r
survival_rates$Adult <- factor(survival_rates$Adult, labels = c("Child", "Adult"))

# Create a grouped bar chart
library(ggplot2)
survival_plot <- ggplot(survival_rates, aes(x = Pclass, y = SurvivalRate, fill = Sex)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  facet_wrap(~Adult) +
  labs(title = "Survival Rates by Sex, Ticket Class, and Age",
       x = "Ticket Class", y = "Survival Rate") +
  scale_fill_manual(values = c("blue", "pink"))

# Display the grouped bar chart
print(survival_plot)
```



- The grouped bar chart shows that the survival rate was higher for females than males in each ticket class and age group.
- The survival rate was highest for children in the second class, followed by adults in the first class and adults in the second class and the third class.
- The survival rate was lowest for male adults in both classes.
- There was no survival rate for children in the first class and the survival rate for male child were higher in the second class than in the third class.

QUESTION 7

```r
# Calculate the total number of family members for each passenger
Titanic_data$FamilySize <- Titanic_data$SibSp + Titanic_data$Parch

# Create a histogram to visualize the distribution of family sizes
family_size_plot <- ggplot(Titanic_data, aes(x = FamilySize)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Distribution of Family Sizes",
       x = "Number of Family Members", y = "Frequency")

# Display the family size distribution plot
print(family_size_plot)
```



The histogram shows that most passengers had no family members on board, with a frequency of at least 80. The next most common family size was 1, with a frequency of more than 35. The frequency decreased as the family size increased.

The distribution of family size was skewed to the right, meaning that there were more passengers with smaller family sizes than larger family sizes. The median family size was 0.

The histogram also shows some outliers in the data, such as the passengers with 5, 6, or 7 family members on board. These passengers may have different characteristics or experiences than the rest of the passengers.

QUESTION 8

```r
# Calculate the total number of family members for each passenger
Titanic_data$FamilySize <- Titanic_data$SibSp + Titanic_data$Parch

# Find the largest family sizes
largest_families <- Titanic_data[Titanic_data$FamilySize == max(Titanic_data$FamilySize),]

# Determine the ticket class of the largest families
largest_families_class <- table(largest_families$Pclass)
print(largest_families_class)
```

```
##
## 3
## 2
```

```r
cat("Ticket class of the largest families: 3")
```

```
## Ticket class of the largest families: 3
```

```r
# Find the ticket class with the most passenger
ticket_class_most_passengers <- names(which.max(table(Titanic_data$Pclass)))
cat("Ticket class with the most passengers:", ticket_class_most_passengers)
```

```
## Ticket class with the most passengers: 3
```

```r
# Filter the dataset to only include female passengers
female_passengers <- Titanic_data[Titanic_data$Sex == "female",]

# Calculate the proportion of female passengers who traveled solo in each class
proportion_solo_female <- table(female_passengers$Pclass, female_passengers$FamilySize == 0)
proportion_solo_female <- proportion_solo_female[, "TRUE"] / table(female_passengers$Pclass)

# Find the ticket class with the lowest proportion of solo female passengers
ticket_class_lowest_proportion_solo_female <- names(which.min(proportion_solo_female))
cat("Ticket class with the lowest proportion of solo female passengers:",
    ticket_class_lowest_proportion_solo_female)
```

```
## Ticket class with the lowest proportion of solo female passengers: 1
```

QUESTION 9

```r
# Convert the Fare column to numeric (if it's not already)
Titanic_data$Fare <- as.numeric(as.character(Titanic_data$Fare))

# Count the number of passengers sharing the same ticket number
Titanic_data$TicketCount <- ave(Titanic_data$Ticket, Titanic_data$Ticket, FUN = length)

# Convert columns to numeric (if not already numeric)
Titanic_data$Fare <- as.numeric(Titanic_data$Fare)
Titanic_data$TicketCount <- as.numeric(Titanic_data$TicketCount)

# Calculate the fare per person by dividing the fare by the ticket count
Titanic_data$FarePerPerson <- Titanic_data$Fare/Titanic_data$TicketCount
```

```r
# Calculate the average fare per person
AverageFarePerPerson <- mean(Titanic_data$FarePerPerson, na.rm = TRUE)

# Print the result
cat("The average fare per person on the Titanic was", AverageFarePerPerson)
```

```
## The average fare per person on the Titanic was 23.86071
```

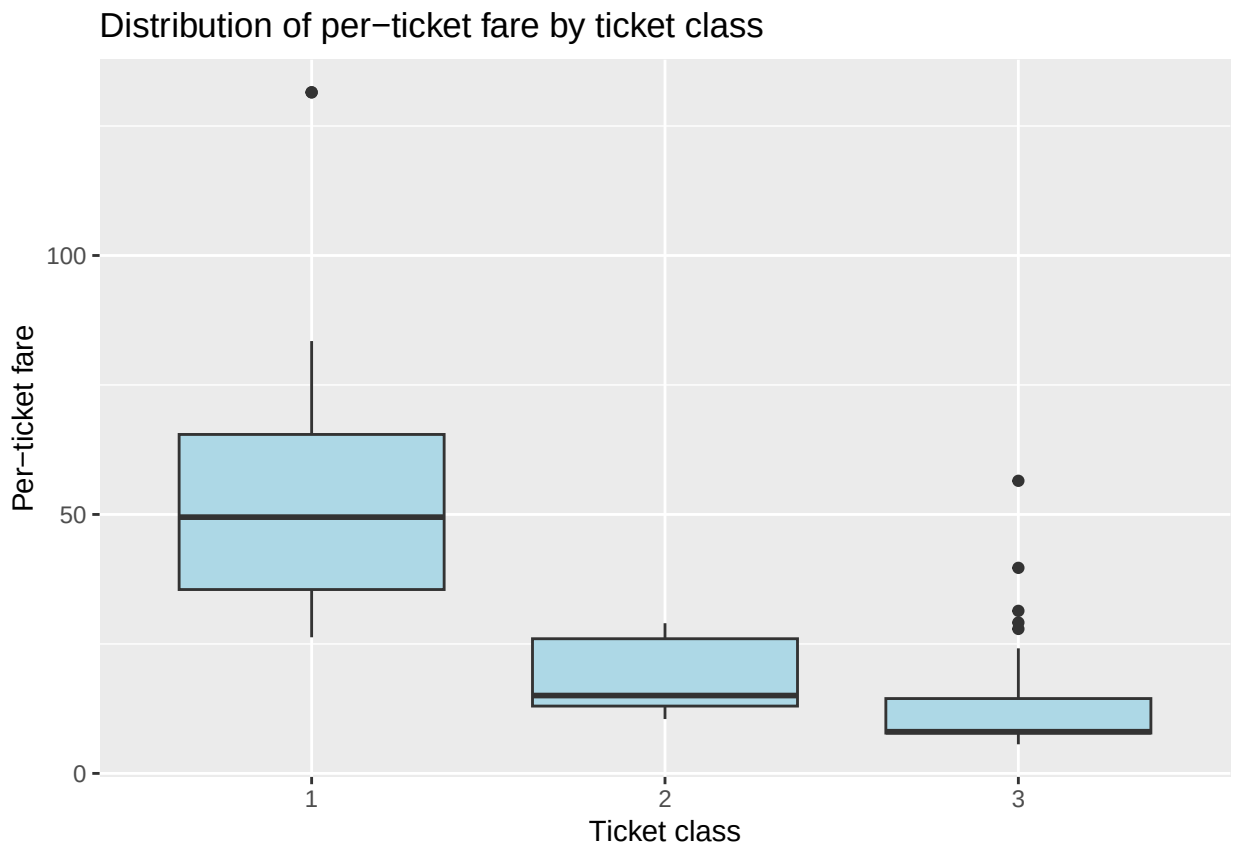
QUESTION 10

```r
# Convert the Ticket column to numeric
Titanic_data$Ticket <- as.numeric(Titanic_data$Ticket)
```

```
## Warning: NAs introduced by coercion
```

```r
# Count the number of passengers sharing the same ticket number
Titanic_data$TicketCount <- ave(Titanic_data$Ticket, Titanic_data$Ticket, FUN = length)

# Calculate the fare per person by dividing the fare by the ticket count
Titanic_data$FarePerPerson <- Titanic_data$Fare/Titanic_data$TicketCount

# Load the ggplot2 library for plotting
library(ggplot2)

# Plot the distribution of the per-ticket fare by ticket class using a boxplot
ggplot(Titanic_data, aes(x = factor(Pclass), y = FarePerPerson)) +
  geom_boxplot(fill = "lightblue") +
  ggtitle("Distribution of per-ticket fare by ticket class") +
  xlab("Ticket class") +
  ylab("Per-ticket fare")
```

```
## Warning: Removed 48 rows containing non-finite values (`stat_boxplot()`).
```



This boxplot shows that the first class has the highest median and range of per-ticket fare, followed by the second class and then the third class.

The first class also has some outliers with very high fares. This suggests that there was a lot of variation in the price of tickets within each class and that some passengers paid much more than others for their tickets.