Assignment 2_Data Visualization

Samaun Sarwar Khan

COLUMN CHART

```
library(ggplot2)
# Column chart for class distribution
ggplot(data, aes(x = factor(Pclass))) +
  geom_bar(aes(fill = factor(Pclass)), position = "dodge", width = 0.7) +
  scale_fill_manual(values = c("steelblue", "aquamarine", "#009999"),
                    name = "Passenger Class",
                    labels = c("1" = "1st Class", "2" = "2nd Class", "3" = "3rd Class")) +
  labs(title = "Figure 1. Titanic Passenger Class Distribution",
       x = "Passenger Class",
       y = "Number of Passengers") +
  theme_minimal()
```

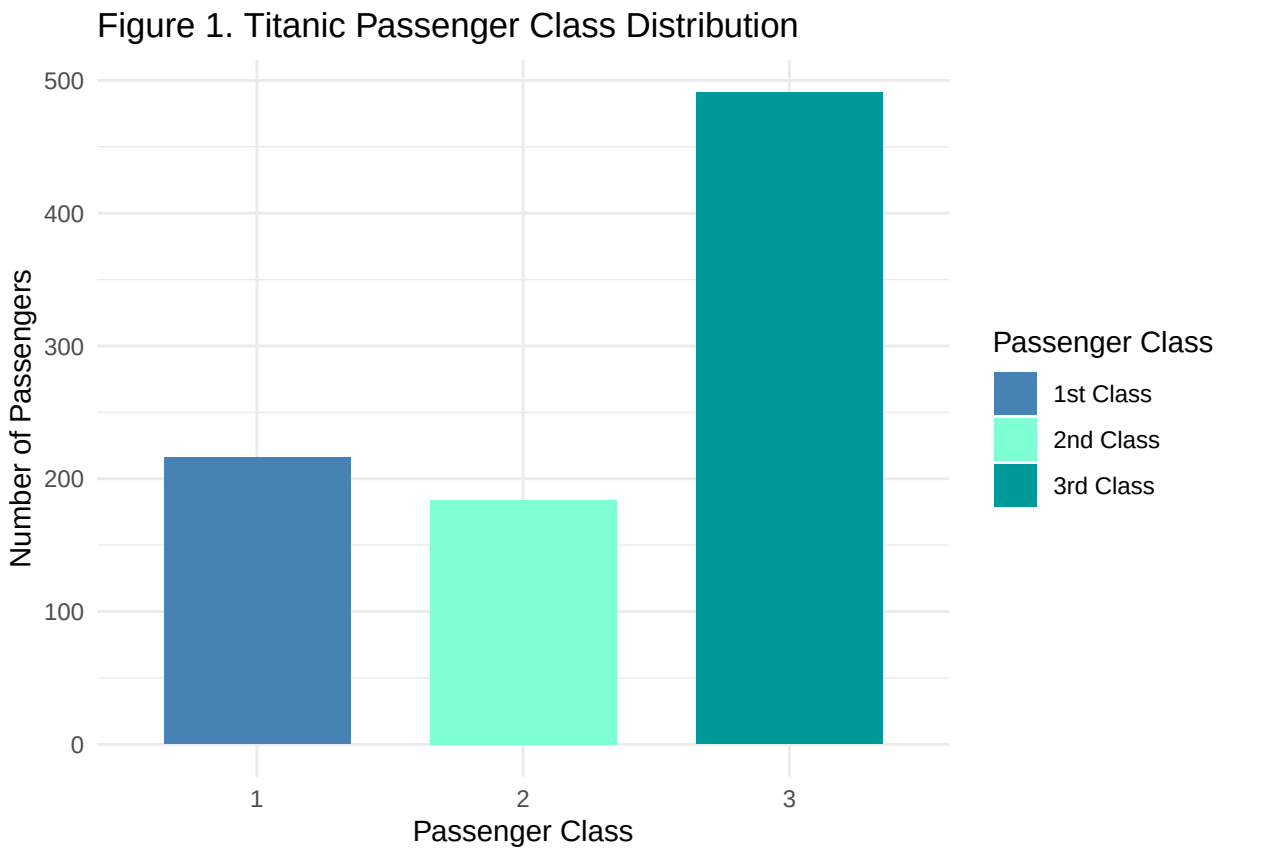


Figure 1 illustrates the Titanic's passenger distribution across classes: first class, second class, and third class. First and second class passengers each represent about a quarter of the sample, while third class passengers make up more than half, indicating their majority onboard.

BAR CHART (Oriented Horizontally)

```
library(ggplot2)
# A horizontal bar chart for passenger class
ggplot(data, aes(y = factor(Pclass), fill = factor(Pclass))) +
  geom_bar() +
  labs(title = "Figure 2. Titanic Passenger Class Distribution",
       y = "Passenger Class",
       x = "Number of Passengers") +
  scale_fill_manual(values = c("#0000FF", "#336699", "#66CCFF"),
                    labels = c("1" = "1st Class", "2" = "2nd Class", "3" = "3rd Class"),
                    name = "Passenger Class") +  # Colors and legend labels
  theme(axis.text.x = element_text(angle = 0))
```

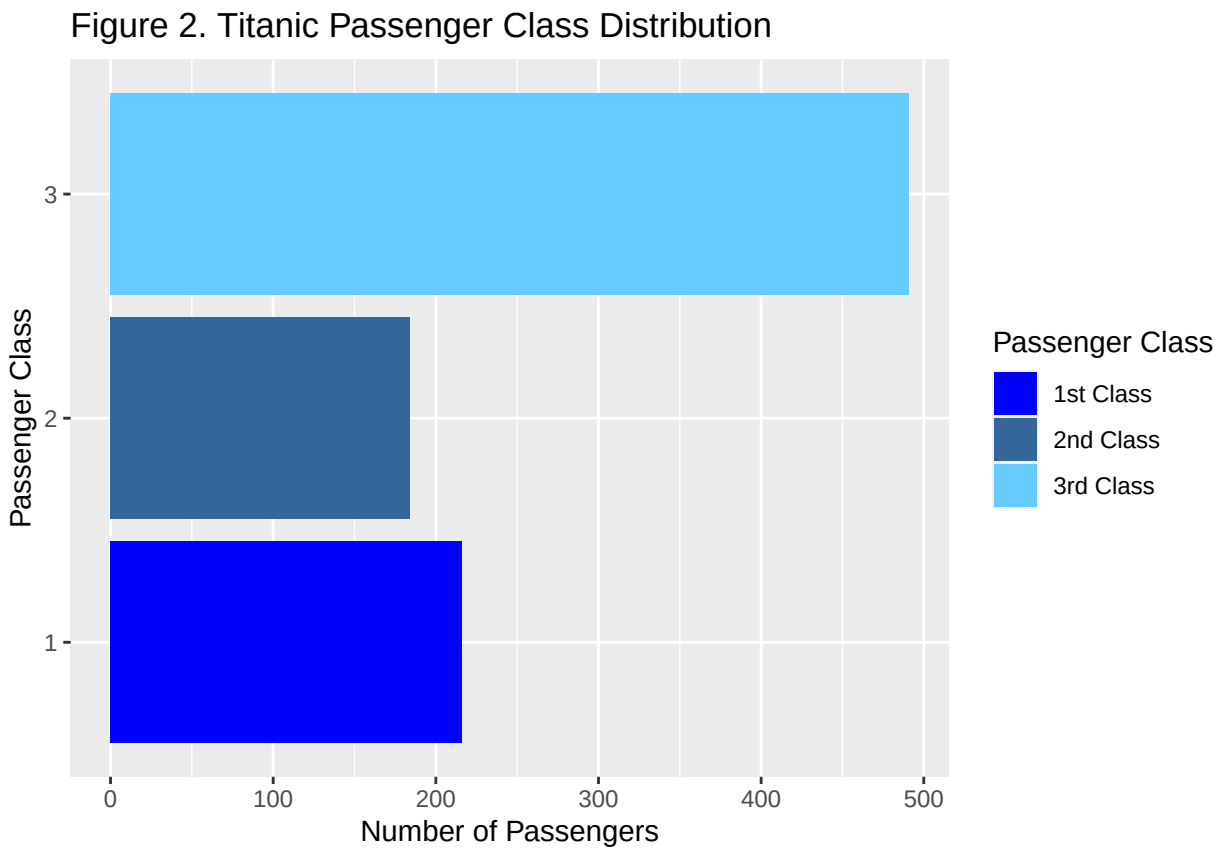


Figure 2 demonstrates the same as the previous column chart. The number of third class passengers are more than first and second class passengers combined indicating the occupancy according to passenger class for the Titanic.

PIE CHART

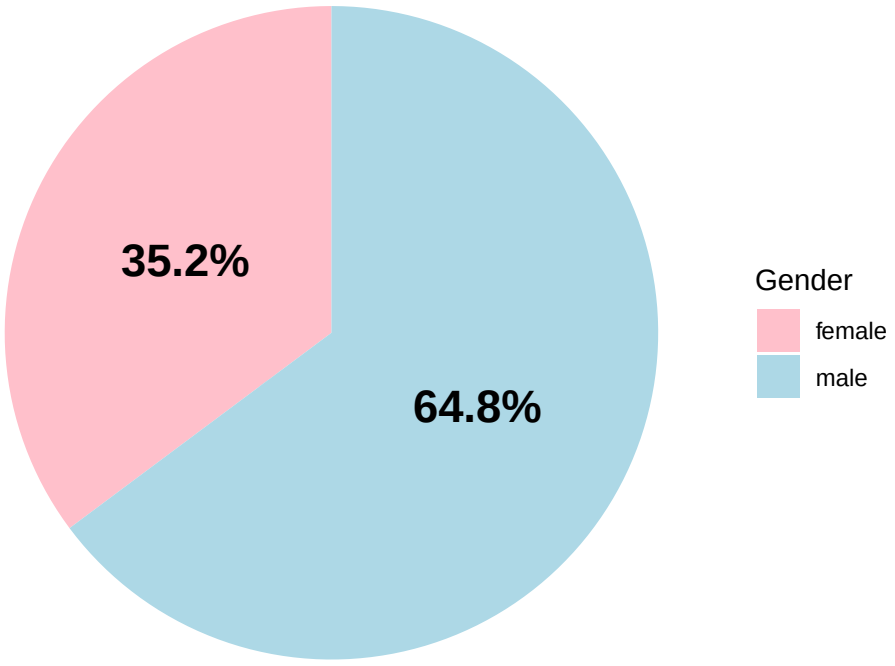
```r
# Load necessary libraries
library(ggplot2)

# Calculate sex distribution percentages with one digit after decimal
sex_counts <- table(data$Sex)
sex_percentages <- prop.table(sex_counts) * 100

# Create a data frame for the percentages with one digit after decimal
percentage_data <- data.frame(Sex = names(sex_percentages),
                              Percentage = sprintf("%.1f", sex_percentages))

# Create a pie chart for sex distribution with specified customization
ggplot(percentage_data, aes(x = "", y = as.numeric(Percentage), fill = Sex)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  theme_void() +  # Remove background and box
  geom_text(aes(label = paste0(Percentage, "%")),
            position = position_stack(vjust = 0.5), color = "black", size = 6,
            fontface = "bold") +
  labs(title = "Figure 3. Titanic Gender Distribution") +
  scale_fill_manual(values = c("pink", "lightblue"),
                    labels = c("female", "male"),
                    name = "Gender")
```

Figure 3. Titanic Gender Distribution



According to figure 3, among the passengers in the Titanic 35.2% were female and 64.8% were male, which indicates that the lions share of passengers (almost double the female) were male.

HISTOGRAM

```
# Create histogram
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, color = "white", fill = "lightblue") +
  stat_bin(aes(y=..count.., label=..count..),
           geom="text", binwidth=4, vjust=-2) +
  xlab("Age") +
  ylab("Number of Passengers") +
  ggtitle("Figure 4. Age Distribution of Titanic Passengers")
```

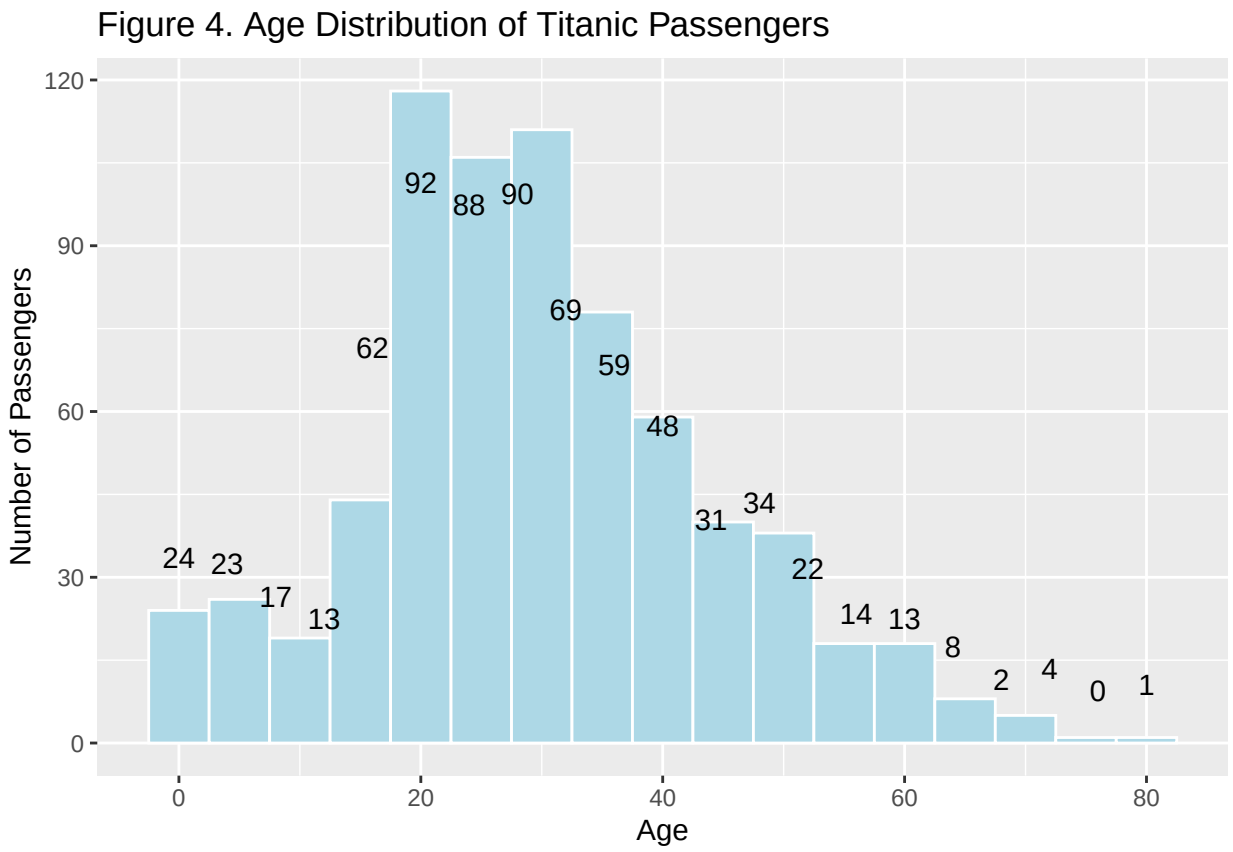


Figure 4 shows the distribution of passengers according to age. Most passengers belonged to 20-40 years old. We can conclude that most of the Titanic passengers were in their youth. However, Titanic had passengers from variety of ages ranging from 1-80 years old, where most passengers were middle aged.

BOX PLOT

```
# Load necessary libraries
library(ggplot2)

# Create a box plot for Age variable
ggplot(data, aes(y = Age)) +
  geom_boxplot(fill = "#1f78b4", color = "black", alpha = 0.7) +
  labs(title = "Figure 5. Titanic Age Distribution",
       y = "Age") +
  theme_minimal()
```

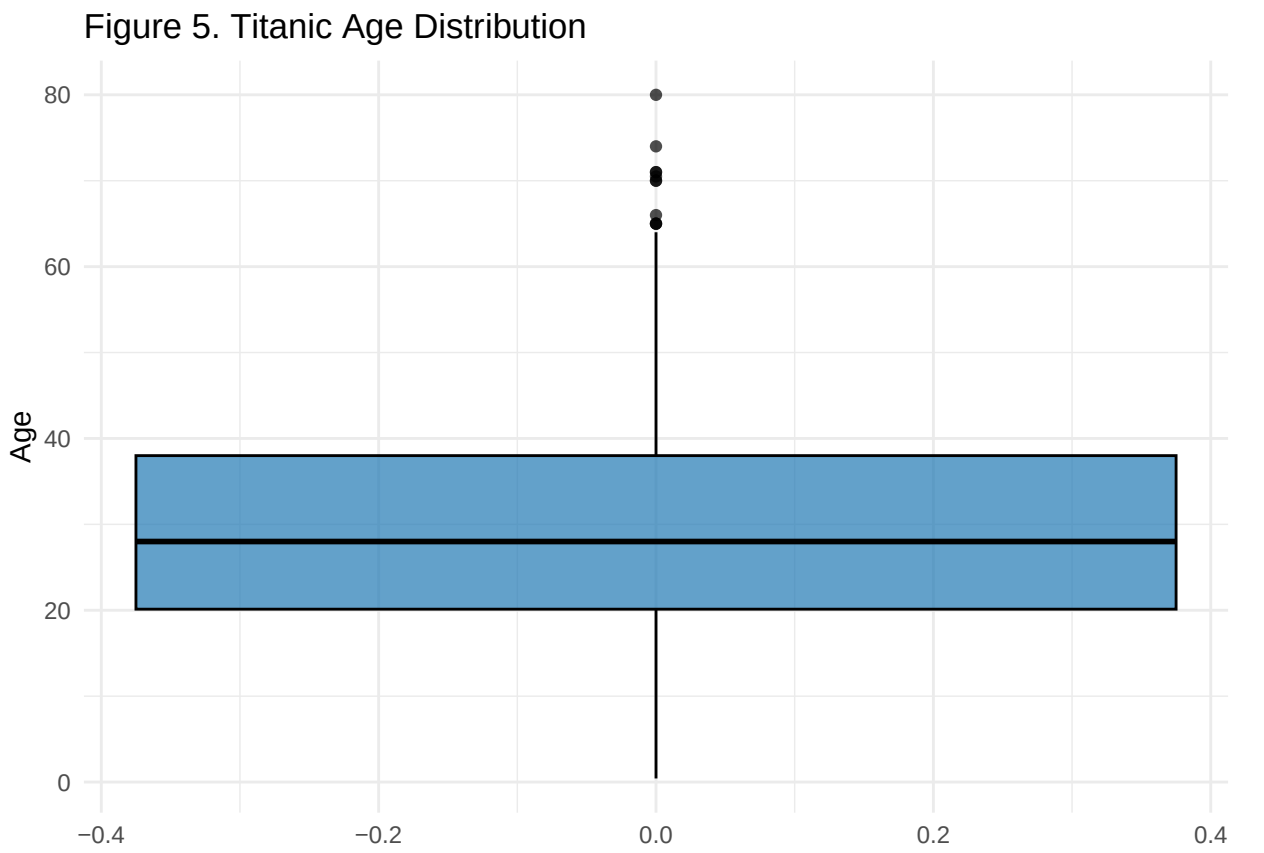


Figure 5 shows most passengers belonged to 20-40 years old with a median od 27 years old. We can also see having some outliers which indicates that some passengers were way older than avergae age of the passengers.

DENSITY PLOT

```r
# Load necessary libraries
library(ggplot2)

# Create a density plot for Fare variable
ggplot(data, aes(x = Fare)) +
  geom_density(fill = "#1f78b4", color = "black", alpha = 0.7) +
  labs(title = "Figure 6. Titanic Fare Density Plot",
       x = "Fare",
       y = "Density") +
  theme_minimal()
```

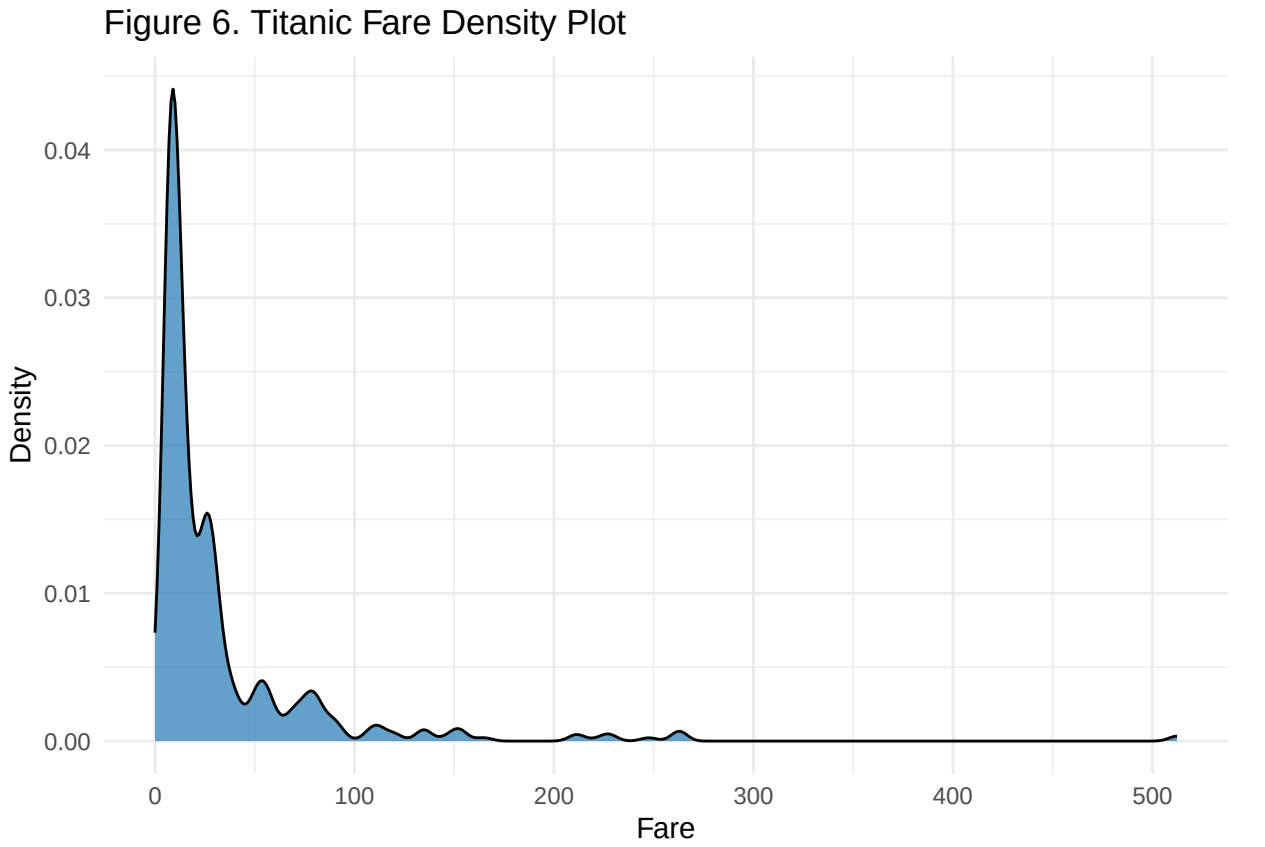


Figure 6 shows that most fare for the Titanic ranged from 1-100 dollars but as the plot shows extension till 500 we can conclude that some fares were higher than the regular fares.

VIOLIN PLOT

```r
# Load necessary libraries
library(ggplot2)

# Create a violin plot for the "Fare" variable with custom legend labels and title
ggplot(data, aes(x = factor(Pclass), y = Fare, fill = factor(Pclass))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.7) +
  labs(title = "Figure 7. Fare Distribution according to Passenger class",
       x = "Passenger Class",
       y = "Fare") +
  scale_fill_manual(values = c("#1f78b4", "#33a02c", "#e31a1c"),
                    labels = c("1" = "1st Class", "2" = "2nd Class", "3" = "3rd Class"),
                    name = "Passenger Class") +
  theme_minimal()
```

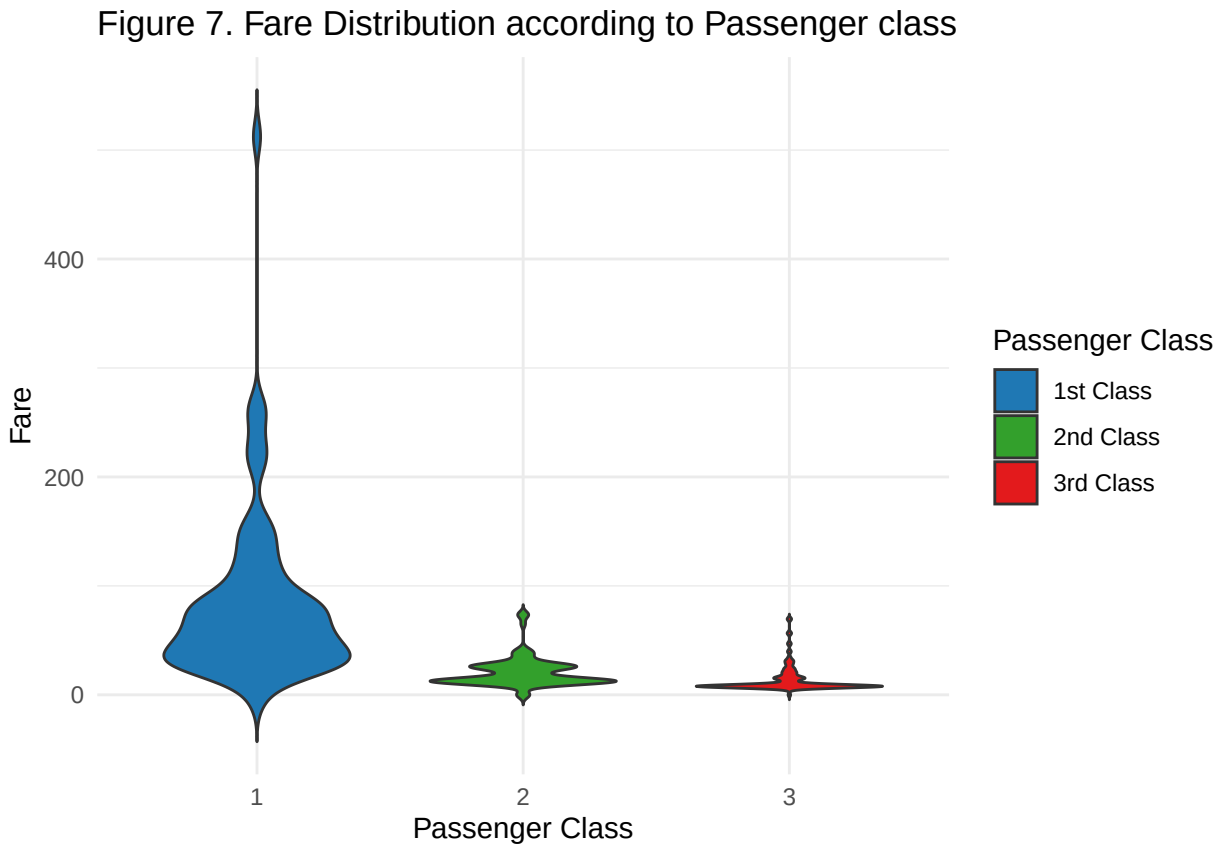


Figure 7 reveal distinct fare patterns among Titanic classes. First-class fares show wide-ranging diversity, indicating varied amenities and accommodations. Second-class fares, while narrower, still exhibit variability, suggesting differing cabin options. Third-class fares are comparatively uniform, suggesting standardized accommodations.

BEESWARM PLOT

```r
# Load necessary libraries
library(ggplot2)
library(ggbeeswarm)

# Filter data for first-class passengers (Pclass = 1)
first_class_data <- subset(data, Pclass == 1)

# Create a bee swarm plot for Age against first-class passengers
ggplot(first_class_data, aes(x = "", y = Age, color = factor(Pclass))) +
  geom_beeswarm(size = 3, alpha = 0.7) +
  labs(title = "Figure 8. Age Distribution of First Class Passengers",
       x = "First Class Passengers",
       y = "Age") +
  theme_minimal() +
  scale_color_manual(name = "Passenger Class",
                     values = c("1" = "steelblue"),
                     labels = c("1" = "1st Class"))
```

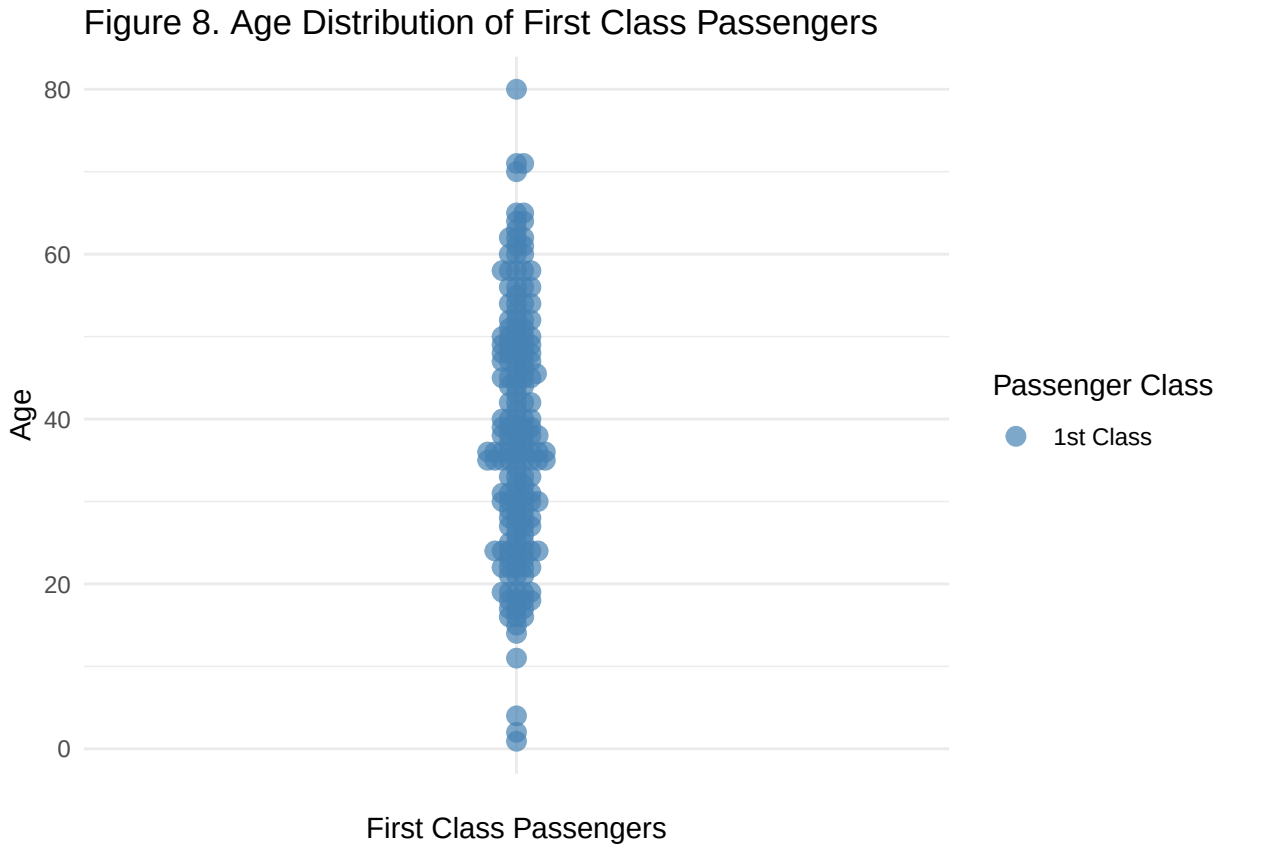


Figure 8 shows the age distribution of first class passengers, which depicts that passengers aging 10-65 years old was the age of majority passengers belonging from first class.

Reference- https://www.kaggle.com/datasets/yasserh/titanic-dataset