IMT 573: Lab 2 - Exploring Data

Tanu Mitra

Thursday, October 08, 2020

Objectives

In this demo we will dive right in and explore a found dataset. Our aim in this demo is to practice getting to know our data. We will follow the steps of exploratory data analysis in this endeavor. This demo will give you an introduction to the veyr popular data visualization package `ggplot`. We will start with the basics today, and see more of this particular tool later on in the course.

```r
# Load some helpful libraries for this course
library(tidyverse)
```

Data Background

The sinking of the RMS Titanic[1] is a notable historical event. The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning of 15 April 1912, after colliding with an iceberg during her maiden voyage from Southampton to New York City. Of the 2,224 passengers and crew aboard, more than 1,500 died in the sinking, making it one of the deadliest commercial peacetime maritime disasters in modern history.

The disaster was greeted with worldwide shock and outrage at the huge loss of life and the regulatory and operational failures that had led to it. Public inquiries in Britain and the United States led to major improvements in maritime safety. One of their most important legacies was the establishment in 1914 of the International Convention for the Safety of Life at Sea (SOLAS)[2], which still governs maritime safety today. Additionally, several new wireless regulations were passed around the world in an effort to learn from the many missteps in wireless communications—which could have saved many more passengers.

The data we will explore in this lab were originally collected by the British Board of Trade in their investigation of the sinking. You can download these data in CSV format from Canvas. Researchers should note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost.

Formulating a Question

Today, we will consider two questions in our exploration:

- Who were the Titanic passengers? What characteristics did they have?
- What passenger characteristics or other factors are associated with survival?

Read and Inspect Data

To begin, we need to load the Titanic dataset into R. You can do so by executing the following code.

[1] https://en.wikipedia.org/wiki/RMS_Titanic

[2] https://en.wikipedia.org/wiki/International_Convention_for_the_Safety_of_Life_at_Sea

```r
titanic_dataset <- read.csv("../../data/titanic.csv")
class(titanic_dataset)
```

```
## [1] "data.frame"
```

```r
#titanic_dataset
```

Next, we want to inspect our data. We don't want to assume that our data is in exactly as we expect it to be after reading it into R. It is helpful to inspect the data object, confirming to looks as expected.

Try editing to following code chunk to look at the top and bottom of your data frame. Perform any other inspection operations you deem necessary. Do you observe anything concerning?

```r
head(titanic_dataset) # Look at the first few rows of the data frame
```

```
##   pclass survived                                             name    sex
## 1      1        1                    Allen, Miss. Elisabeth Walton female
## 2      1        1                   Allison, Master. Hudson Trevor   male
## 3      1        0                     Allison, Miss. Helen Loraine female
## 4      1        0             Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                              Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S          NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S         135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S          NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                    St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                    New York, NY
```

```r
tail(titanic_dataset) # Look at the last few rows of the data frame
```

```
##      pclass survived                      name    sex  age sibsp parch ticket
## 1304      3        0     Yousseff, Mr. Gerious   male   NA     0     0   2627
## 1305      3        0      Zabour, Miss. Hileni female 14.5     1     0   2665
## 1306      3        0     Zabour, Miss. Thamine female   NA     1     0   2665
## 1307      3        0 Zakarian, Mr. Mapriededer   male 26.5     0     0   2656
## 1308      3        0       Zakarian, Mr. Ortin   male 27.0     0     0   2670
## 1309      3        0        Zimmerman, Mr. Leo   male 29.0     0     0 315082
##         fare cabin embarked boat body home.dest
## 1304 14.4583             C        NA
## 1305 14.4542             C       328
## 1306 14.4542             C        NA
## 1307  7.2250             C       304
## 1308  7.2250             C        NA
## 1309  7.8750             S        NA
```

```r
summary(titanic_dataset) # Use the summary function to inspect variables
```

```
##      pclass         survived         name                 sex
##  Min.   :1.000   Min.   :0.000   Length:1309        Length:1309
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character
##  Median :3.000   Median :0.000   Mode  :character   Mode  :character
##  Mean   :2.295   Mean   :0.382
##  3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :3.000   Max.   :1.000
##
##       age              sibsp            parch          ticket
##  Min.   : 0.1667   Min.   :0.0000   Min.   :0.000   Length:1309
##  1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000   Class :character
##  Median :28.0000   Median :0.0000   Median :0.000   Mode  :character
##  Mean   :29.8811   Mean   :0.4989   Mean   :0.385
##  3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :80.0000   Max.   :8.0000   Max.   :9.000
##  NA's   :263
##       fare            cabin             embarked             boat
##  Min.   :  0.000   Length:1309        Length:1309        Length:1309
##  1st Qu.:  7.896   Class :character   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 33.295
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
##       body        home.dest
##  Min.   :  1.0   Length:1309
##  1st Qu.: 72.0   Class :character
##  Median :155.0   Mode  :character
##  Mean   :160.8
##  3rd Qu.:256.0
##  Max.   :328.0
##  NA's   :1188
```

Solution: *Inspecting the data frame top and bottom reveals data that appears consistent with our expectations. We see that the dataset contains various characteristics about individual passengers inluding the class of their name, sex, and age. We also see that there is a variable called 'survived' which is likely to contain data about whether that person survived the ship's crash. We also note that some variables have missing data, represented by NAs.*

Think about the variables in this data as they are defined. Which variables might you want to re-cast to be the appropriate data type in R?

Solution: *The summary function also reveals that the 'survived' variables is being treated as a numeric variable in R. This characteristic is more appropriately a categorical variable and therefore we will re-cast it as a factor. The same goes for 'pclass'.*

Transform the data type of varibles you identify as improperly cast.

```
# Re-cast categorical variables to be factor data types
titanic_dataset$pclass <- as.factor(titanic_dataset$pclass)
titanic_dataset$survived <- as.factor(titanic_dataset$survived)
```

Trying the Easy Solution First

First, we want to explore who the passengers aboard the Titanic were. There are many ways we might go about this.

Consider for example trying to understand the ages of passengers. We can create a basic visualization to help us understand the distributions of age for Titanic passengers.

```
#hist(titanic_dataset$age)
#hist(titanic_dataset$pclass)
```

```
ggplot(data = titanic_dataset, aes(age)) +
  geom_histogram(fill="blue")
```

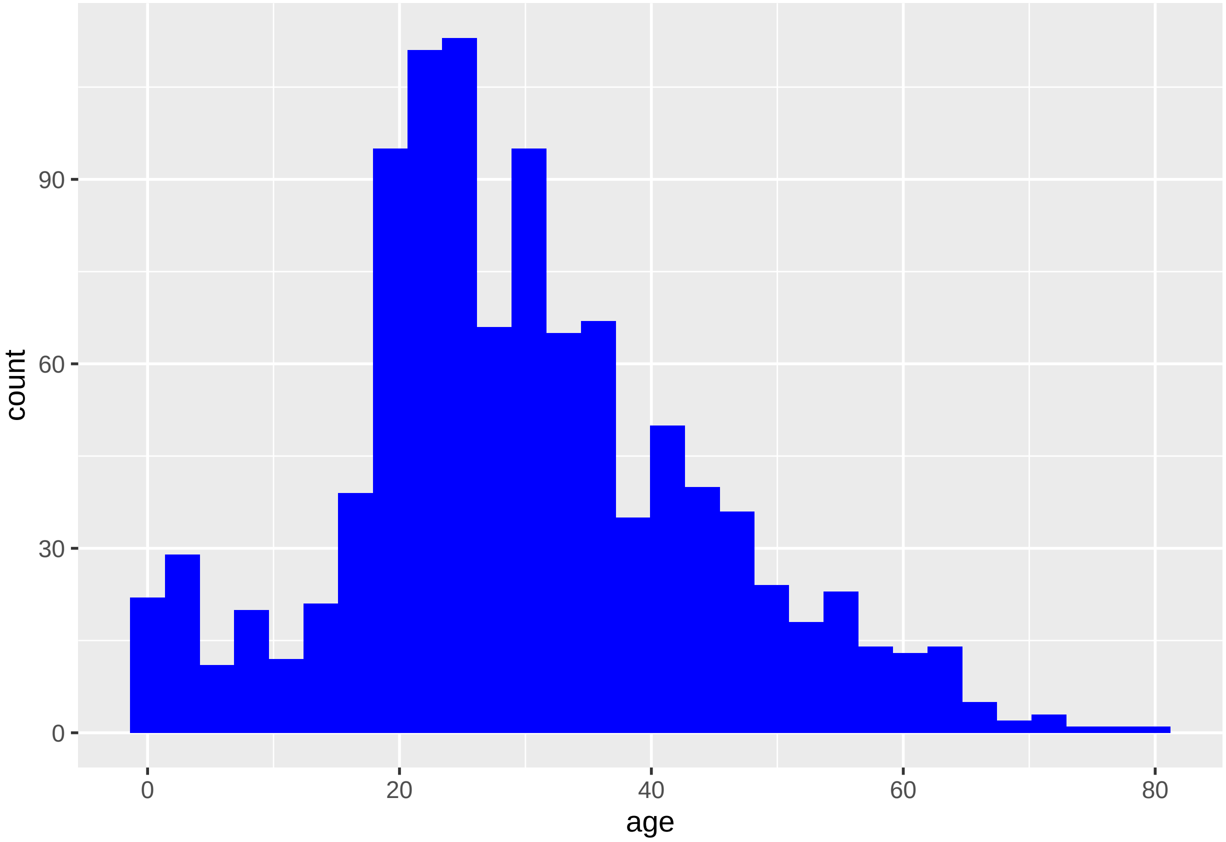

Figure 1: Age of Passenders Aboard the Titanic

We might go further to look at how passenger age might be related to survival. (2nd question: What passenger characteristics or other factors are associated with survival?)

```
# Figure to show age distribution by survival
ggplot(data = titanic_dataset, aes(age, survived)) +
  geom_point(size=2, alpha=0.5, color="red")
```

Do you like the above figure? Why or why not? Produce a new figure that you think does a better job of helping you explore the association between passenger age and survival.

Solution: *The first figure does not do a good job of displaying the data. Points are overlapping and there is a lot of blank space. In particular, it does not help us understand the distribution of ages by survival. A better plot would be a boxplot to show the age distribution for each value of the survival variable. We don't see any striking differences in the age of passengers in each category.*

```
ggplot(data = titanic_dataset, aes(survived, age)) +
  geom_boxplot()
```

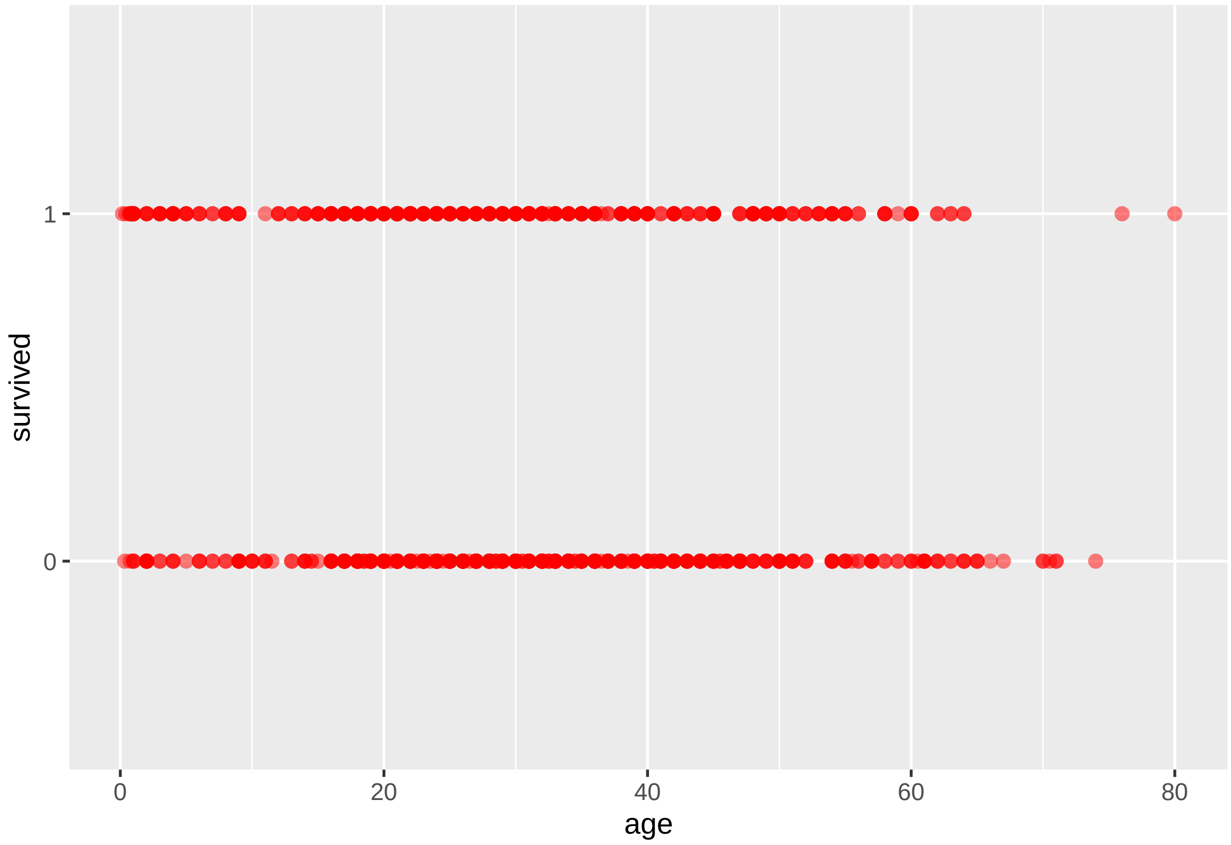

Figure 2: Survival and Passenger Age

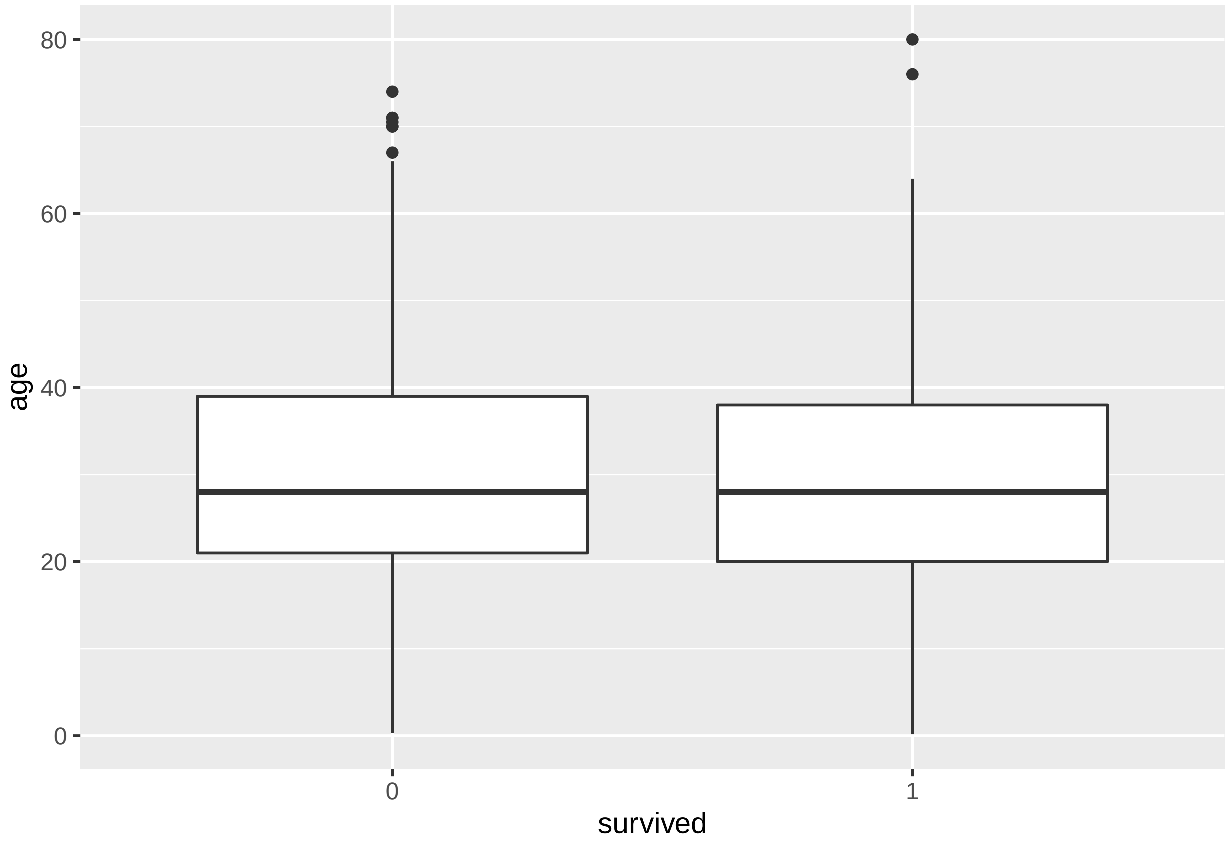

Figure 3: Survival and Passenger Age

```r
#both work
ggplot(data = titanic_dataset) +
  geom_boxplot(aes(survived, age))
```

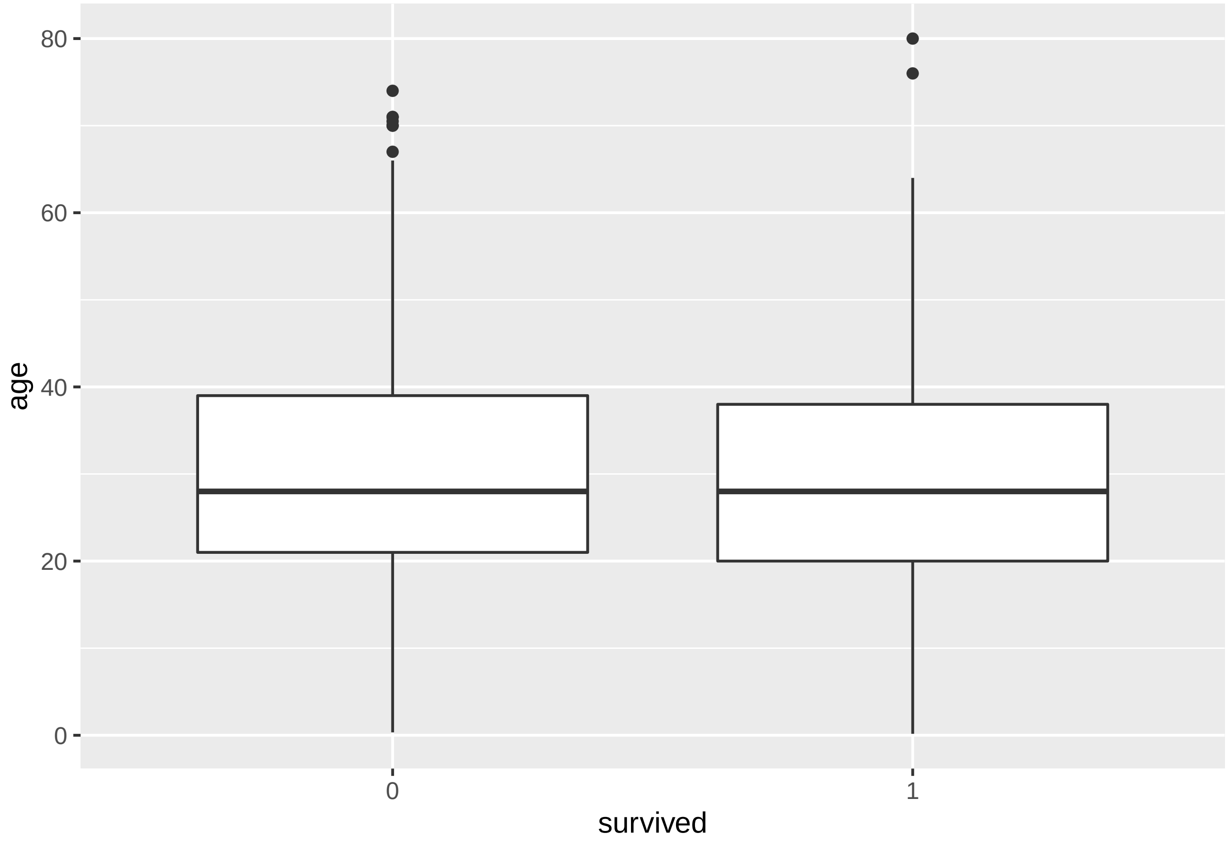

Figure 4: Survival and Passenger Age

Identify one additional data feature you want to explore. Produce one visualization that explore this feature. Describe why you think this is interesting and what you find.

Solution: *We want to look at the relationship between survival and passenger class to determine if evidence suggests high surivial rates for upper class passengers. In the following figure we see not only how many passengers fall into each class, relatively, but also what proportion survived. Data suggests that passengers in 1st and 2nd class cabins had higher rates of surivival, compared the 3rd class passengers.*

```r
mosaicplot(titanic_dataset$pclass ~ titanic_dataset$survived,
           main="Passenger Fate by Traveling Class",
           shade=FALSE,
           color=TRUE, xlab="Pclass", ylab="Survived")
```

What Next?

Consider the exploratory analysis we just completed in the demo. What would you do next?

Solution: *We might want to build a statistical model to compare the relative influence of each factor on survival or help predict the survival of a passenger.*

Statisti
cal

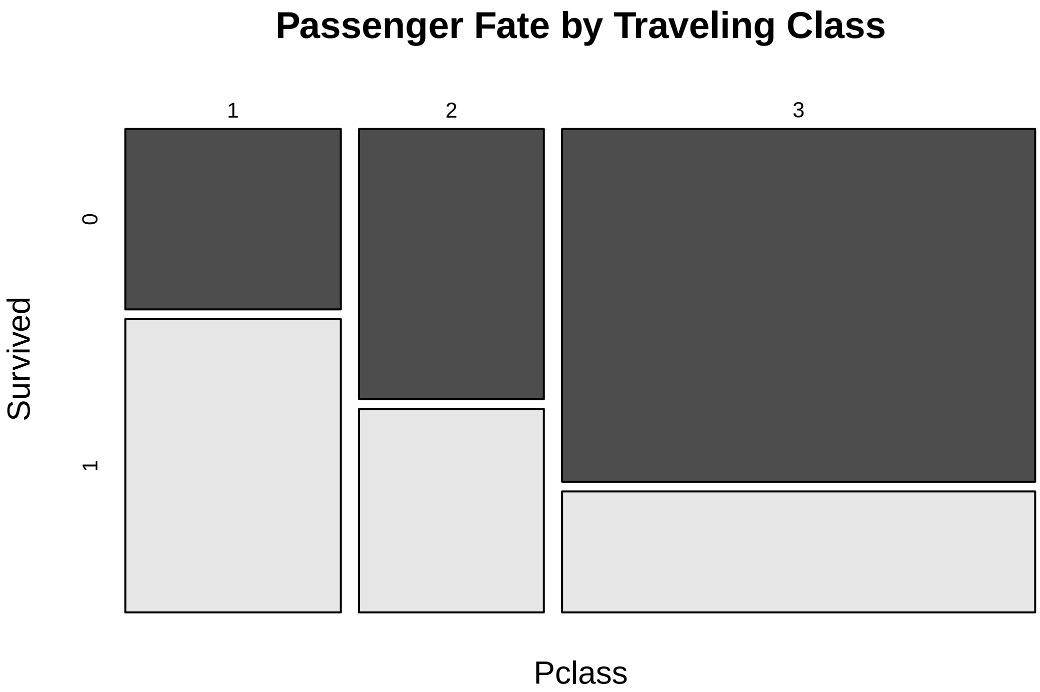

Figure 5: Survival and Passenger Class