Final Project - Math 110 Intro to Data Science Programming

Larine Hamied

12/15/2018

```
library(ggplot2)
```

Titanic Dataset

I got this dataset from Kaggle at https://www.kaggle.com/naresh31/titanic-machine-learning-from-disaster/data (https://www.kaggle.com/naresh31/titanic-machine-learning-from-disaster/data). It is from a current Competition on the Kaggle site to use machine learning with the dataset to predict information about a specific passenger and if they survived or not after the ship sank. They provided a “train”, “test”, and “gender_submission” dataset. They “train” dataset has all the categories we are going to analyze. The “test”" dataset has all of the categories EXCEPT for if the person survived or not. The “gender_submission” dataset is two columns: the passenger ID and if the person survived or not. Below, I merge all of the datasets to have one “Titanic” dataset to manipulate to look at the data and do some visualization to show concepts from earlier in the course.

I am creating a function that makes a Titanic Object that merges all the data talked about above.

```
TitanicObject <- function () {
  Titanic1 <- read.csv("train.csv", header = T, sep = ",")
  Titanic2 <- read.csv("test.csv", header = T, sep = ",")
  Titanic2Answers <- read.csv("gender_submission.csv", header = T, sep = ",")
  Titanic2$Survived <- Titanic2Answers$Survived
  Titanic <- rbind(Titanic1, Titanic2)
  return(Titanic)
}
```

```
t <- TitanicObject()
names(t)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
str(t)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 929 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Here are the types and meaning of each category: **PassengerId** - integers, number of which passenger it is. **Survived** - 0 if they did not survive, 1 if they did survive. **Pclass** - Class of Travel where 1 is the highest and 3 is the lowest. **Name** - the Name of the passenger. **Sex** - male or female. **Age** - if a decimal less than 1, it is an infant. **SibSpNumber** - number of siblings or spouses aboard. **ParchNumber** - number of parents or child aboard. **Ticket** - ticket code/number. **Fare** - amount paid for the fare. **Cabin**- cabin number. **Embarked** - The port in which a passenger has embarked, where C stands for Cherbourg, S stands for Southampton, and Q stands for Queenstown

Data Manipulation

```
createAgeGroups <- function(t) {
  t$AgeGroup <- 'NA'
  t$AgeGroup[t$Age >= 65] <- 'Elderly'
  t$AgeGroup[t$Age < 65 & t$Age >= 18] <- 'Adult'
  t$AgeGroup[t$Age > 2 & t$Age < 18] <- 'Child'
  t$AgeGroup[t$Age < 2] <- 'Infant'
  return(t)
}
```

Creates a column to the dataset that filters people’s ages in the following groups: Less than 2 years old = Infant Ages 2 - 18 = Child Ages 18 - 64 = Adult Ages 65 + = Elderly

```
createFamilySize <- function(t) {
  t$FamilySize <- t$SibSp + t$Parch + 1
  return(t)
}
```

Creates a column that adds up the number of siblings, spouses, parents or child on board. If a person did not have anyone reported, then family size is 1.

```
summary(t$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   0.000   7.896  14.450  33.300  31.280 512.300       1
```

Looking at a summary of the fares, I came up with the following numbers to create groups for the fare prices below.

```
createFareRange <- function(t) {
  t$FareRange <- '30+'
  t$FareRange[t$Fare < 30 & t$Fare >= 20] <- '20-30'
  t$FareRange[t$Fare < 20 & t$Fare >= 10] <- '10-20'
  t$FareRange[t$Fare < 10] <- '<10'
  return(t)
}
```

This can potentially give us an idea of social class.

```
createFirstClass <- function(t) {
  t$FirstClass <- 0
  t$FirstClass[t$Pclass == 1] <- 1
  t$FirstClass <- as.factor(t$FirstClass)
  return(t)
}
```

This creates an additional category that I can append later of passengers in first class or not, to also get an idea of social class.

```
head(t$Name)
```

```
## [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
## [6] Moran, Mr. James
## 1307 Levels: Abbing, Mr. Anthony ... Zakarian, Mr. Ortin
```

```
t$Title <- sapply(as.character(t$Name), FUN=function(x) {strsplit(x, split='[,.]')[[1]][2]})
t$Title <- sub(' ', '', t$Title)
unique(sort(t$Title))
```

```
##  [1] "Capt"         "Col"          "Don"          "Dona"
##  [5] "Dr"           "Jonkheer"     "Lady"         "Major"
##  [9] "Master"       "Miss"         "Mlle"         "Mme"
## [13] "Mr"           "Mrs"          "Ms"           "Rev"
## [17] "Sir"          "the Countess"
```

```
sort(table(t$Title), decreasing = T)
```

```
## 
##           Mr         Miss          Mrs       Master           Dr 
##          757          260          197           61            8 
##          Rev          Col        Major         Mlle           Ms 
##            8            4            2            2            2 
##         Capt          Don         Dona     Jonkheer         Lady 
##            1            1            1            1            1 
##          Mme          Sir the Countess 
##            1            1            1
```

Looking at the different names in the dataset, they include a “Title” for each passenger. I looked at the unique titles in the dataset and wanted to get an idea of the number of each title. As one would guess, the most common titles are Mr, Miss and Mrs.

```r
addTitle <- function(t) {
  t$Title <- sapply(as.character(t$Name), FUN=function(x) {strsplit(x, split='[,.]')[[1]][2]})
  t$Title <- sub(' ', '', t$Title)
  table(t$Title)

  t$Title[t$Title %in% c('Miss', 'Mrs', 'Ms','Mme', 'Mlle')] <- 'Miss'
  t$Title[t$Title %in% c('Capt', 'Don', 'Major', 'Sir', 'Rev', 'Col', 'Dr', 'Master', 'Jonkheer')] <- 'Sir'
  t$Title[t$Title %in% c('Dona', 'Lady', 'the Countess')] <- 'Lady'
  t$Title <- factor(t$Title)
  return(t)
}
```

This function makes categories for the titles: Mr I left as its own category, where **Mr** is a common man’s title. Miss, Mrs, Ms, Mme & Mmle I grouped together as **Miss** or a common woman’s title. Capt, Don, Major, Sir, Rev, Col, Dr, Master & Jonkheer I grouped as **Sir** or an elite man’s title. Dona, Lady, and the Countess I grouped as **Lady** or an elite woman’s title.

Functions for Data Analysis

```
SurvivalByVariable <- function(t, survived, variable){
  data <- as.data.frame(aggregate(survived ~ variable, data=t , FUN=sum))
  xLabel <- deparse(substitute(variable))
  g <- ggplot(data, aes(x= variable, y = survived, fill = variable)) + geom_bar(stat = "identity") + labs(x=xLabel, y="Number of Survivors", fill = xLabel)
  return(print(g))
}

SurvivalByVariablePercentage <- function(t, survived, variable){
  asPercentage <- as.data.frame(aggregate(survived ~ variable, data=t , FUN=function(x) {sum(x)/length(x)}))
  xLabel <- deparse(substitute(variable))
  g <- ggplot(asPercentage, aes(x= variable, y = survived, fill = variable)) + geom_bar(stat = "identity") + labs(x=xLabel, y="Percentage of Survivors", fill = xLabel)
  return(print(g))
}

SurvivalTwoVariables <- function(t, survived, variable1, variable2){
  data <- as.data.frame(aggregate(survived ~ variable1 + variable2, data=t , FUN=sum))
  xLabel <- deparse(substitute(variable1))
  legendLabel <- deparse(substitute(variable2))
  g <- ggplot(data, aes(x= variable1, y = survived, fill = variable2)) + geom_bar(stat = "identity") + labs(x=xLabel, "Number of Survivors", fill = legendLabel)
  return(print(g))
}

SurvivalTwoVariablesPercentage <- function(t, survived, variable1, variable2){
  asPercentage <- as.data.frame(aggregate(survived ~ variable1 + variable2, data=t , FUN=function(x) {sum(x)/length(x)}))
  xLabel <- deparse(substitute(variable1))
  legendLabel <- deparse(substitute(variable2))
  g <- ggplot(asPercentage, aes(x= variable1, y = survived, fill = variable2)) + geom_bar(stat = "identity", position=position_dodge()) +  labs(x=xLabel, y="Percentage of Survivors", fill = legendLabel)
  return(print(g))
}
```

Code

Data Manipulation using functions from above

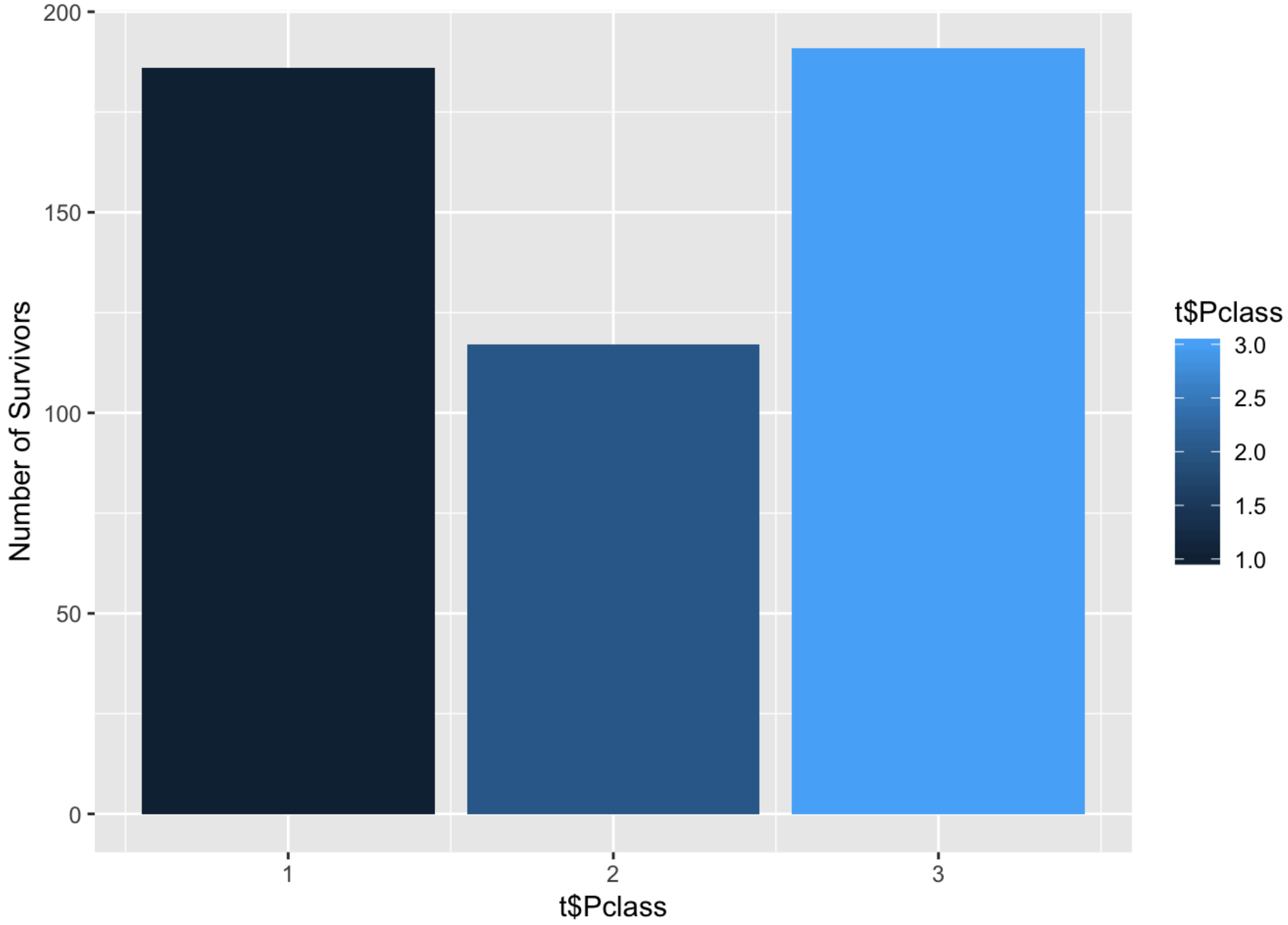
```
t <- createAgeGroups(t)
t <- createFamilySize(t)
t <- createFareRange(t)
t <- createFirstClass(t)
t <- addTitle(t)
```

Data Analysis using functions from above

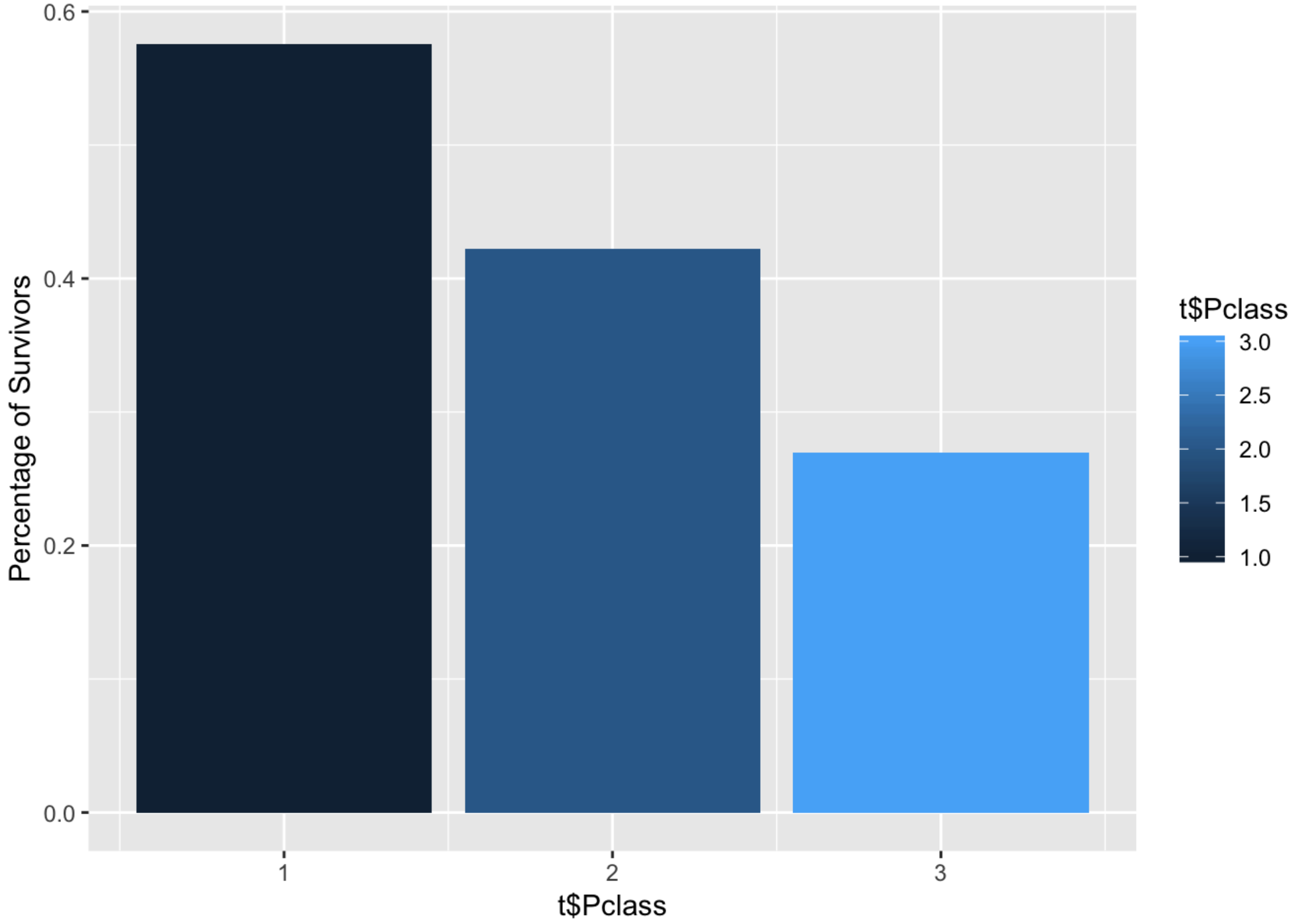
Social Class

Looking at the three cabins, where 1 is the best/most expensive and 3 is the worst, we see that the overall percentage of survivors decreased as the cabin was lower (First Class had a nearly 60% survival rate while the second class had about 40% and the third class had less than 30%)

```
SurvivalByVariable(t, t$Survived, t$Pclass)
```

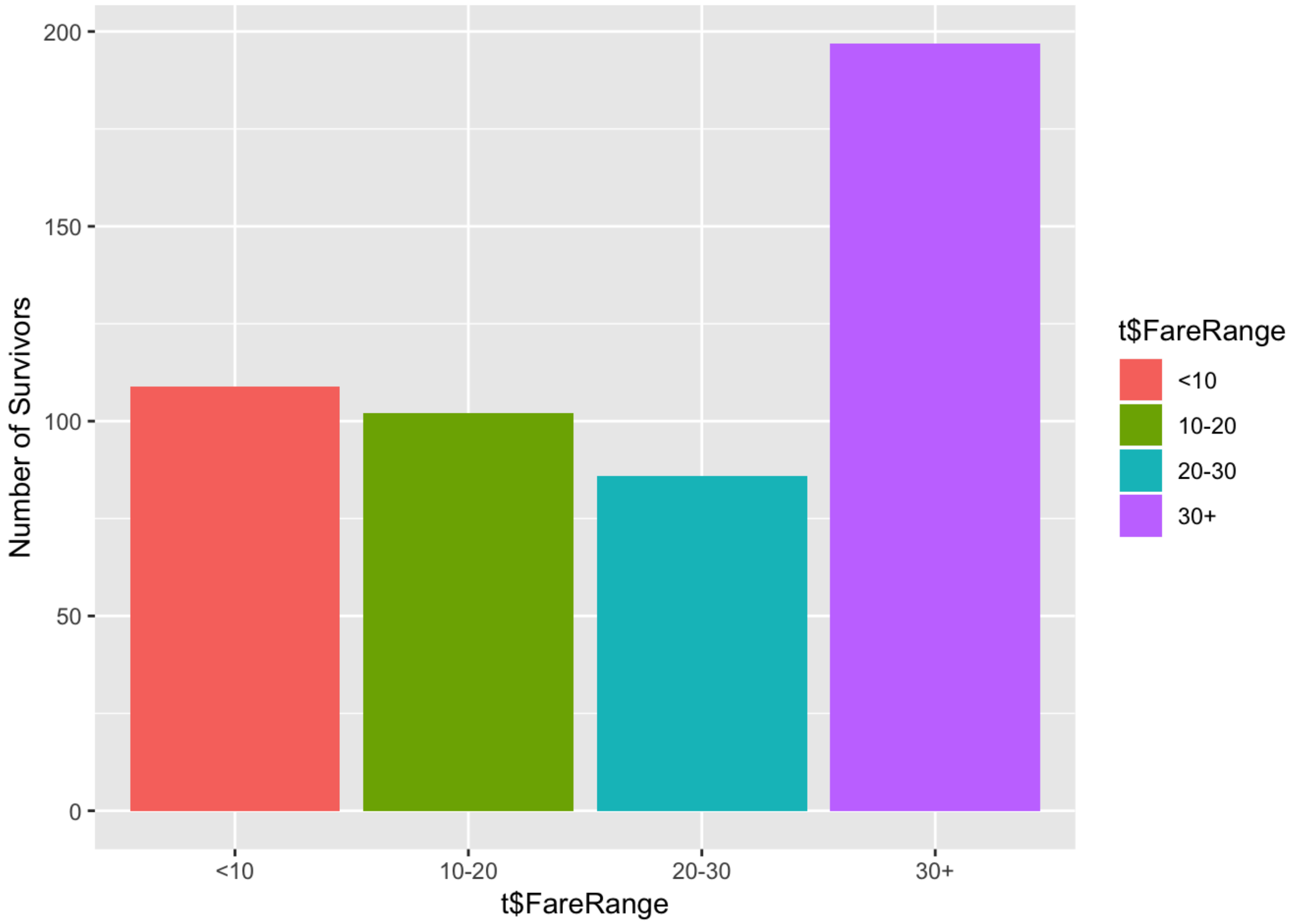


```
SurvivalByVariablePercentage(t, t$Survived, t$Pclass)
```

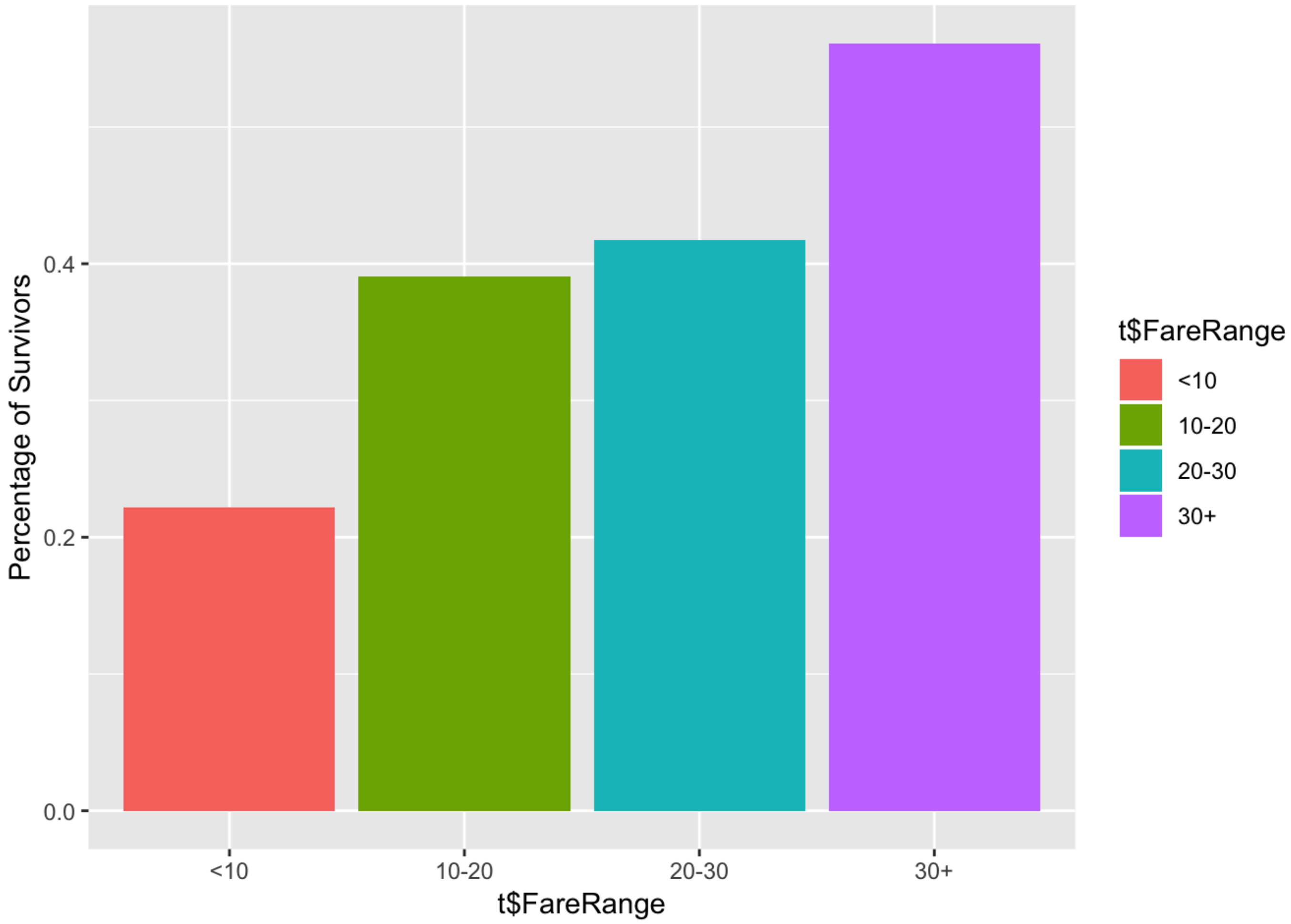


Below I will look at the same thing but with the fare paid:

```
SurvivalByVariable(t, t$Survived, t$FareRange)
```

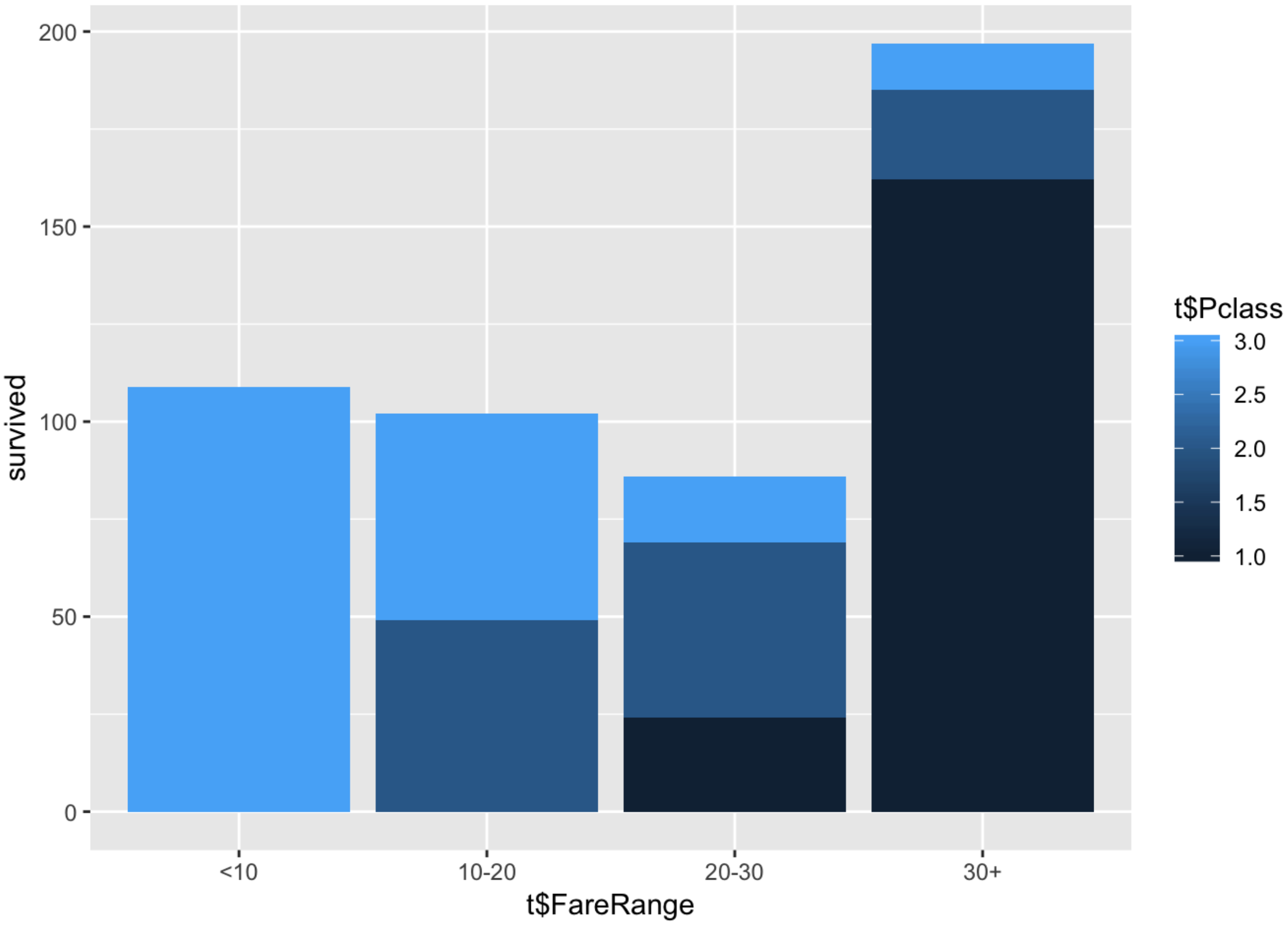


```
SurvivalByVariablePercentage(t, t$Survived, t$FareRange)
```

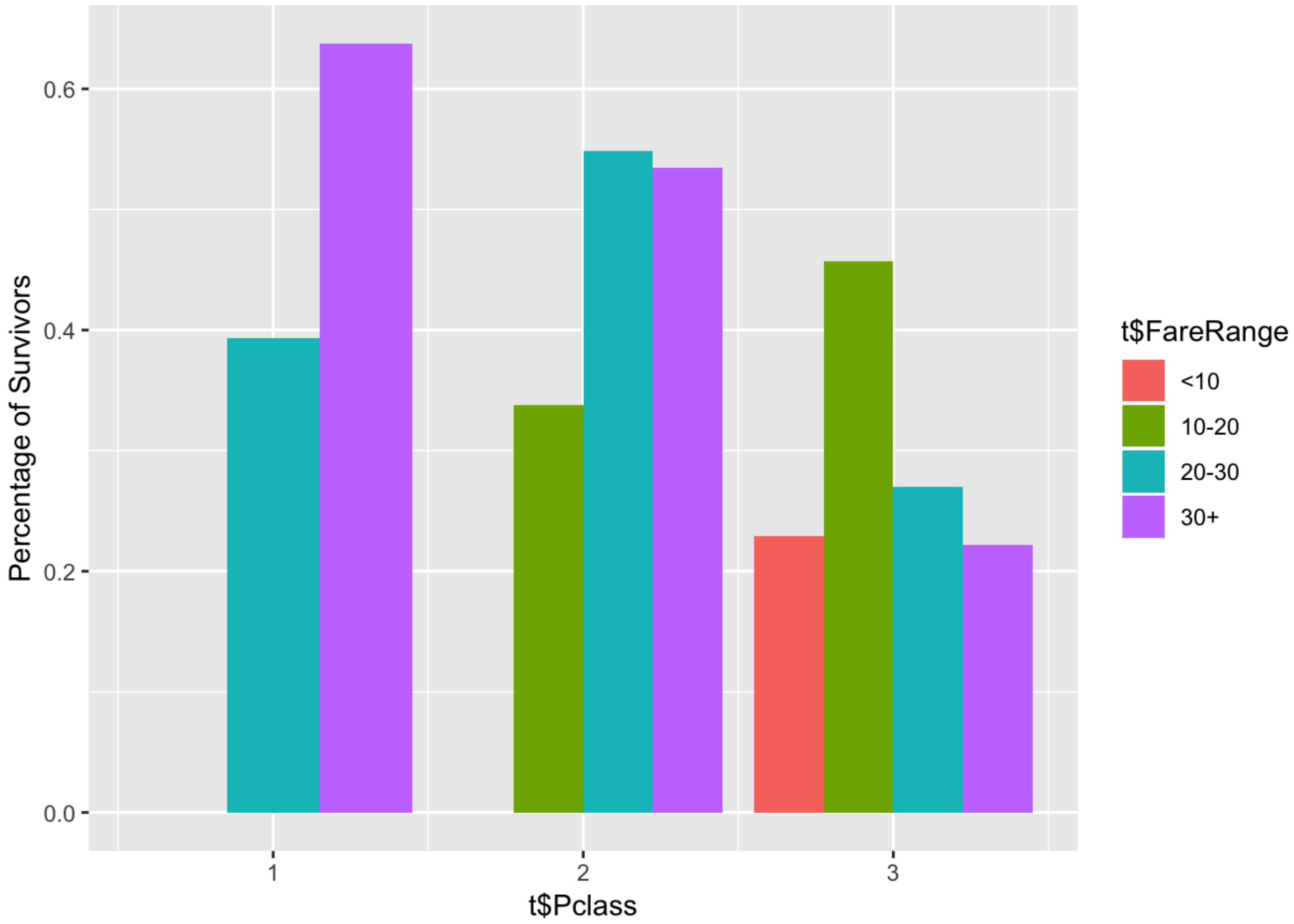


Below is a comparison of Fare Range and the Class to confirm the assumption that higher fare meant higher class ticket.

```
SurvivalTwoVariables(t, t$Survived, t$FareRange, t$Pclass)
```



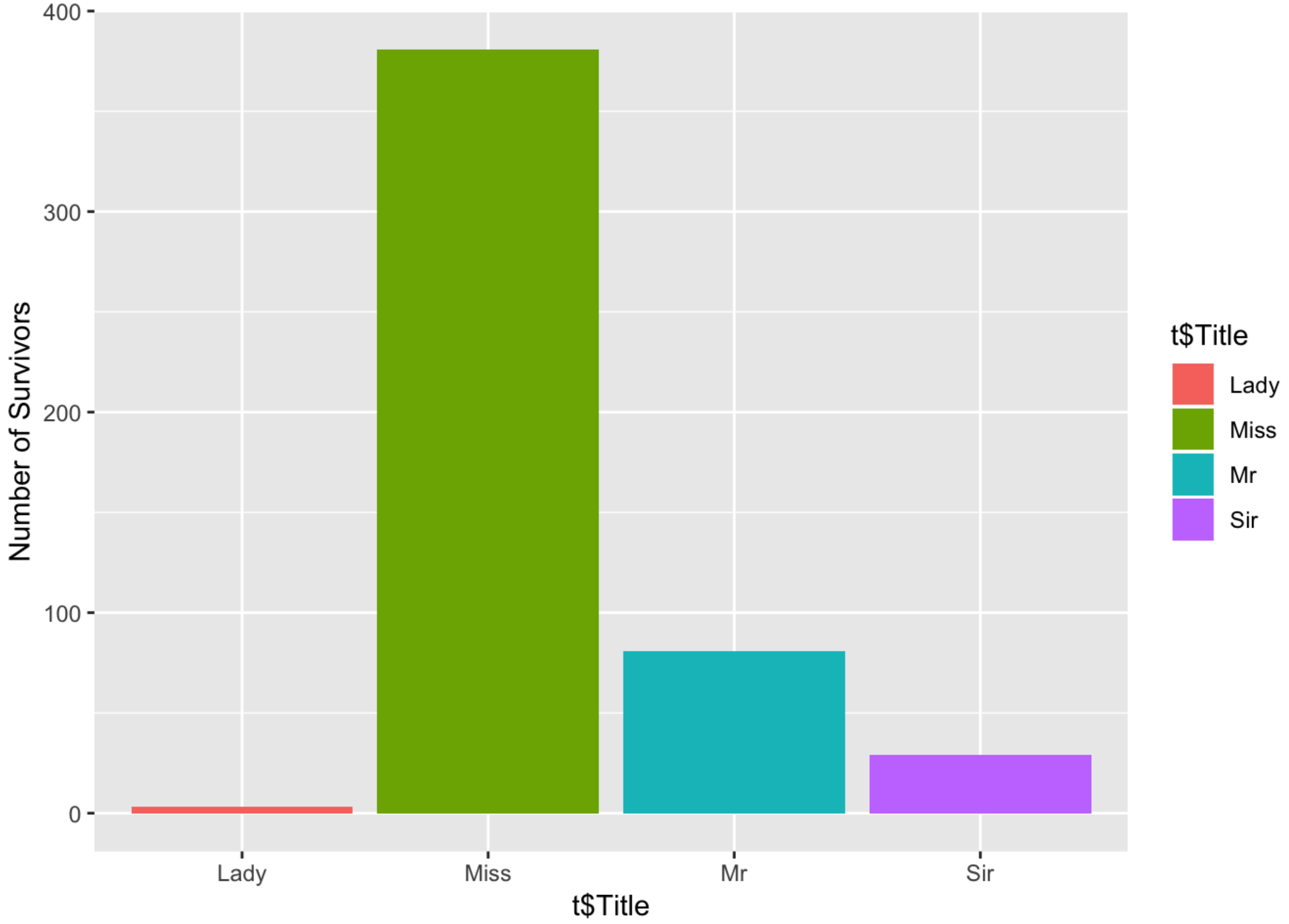
```
SurvivalTwoVariablesPercentage(t, t$Survived, t$Pclass, t$FareRange)
```



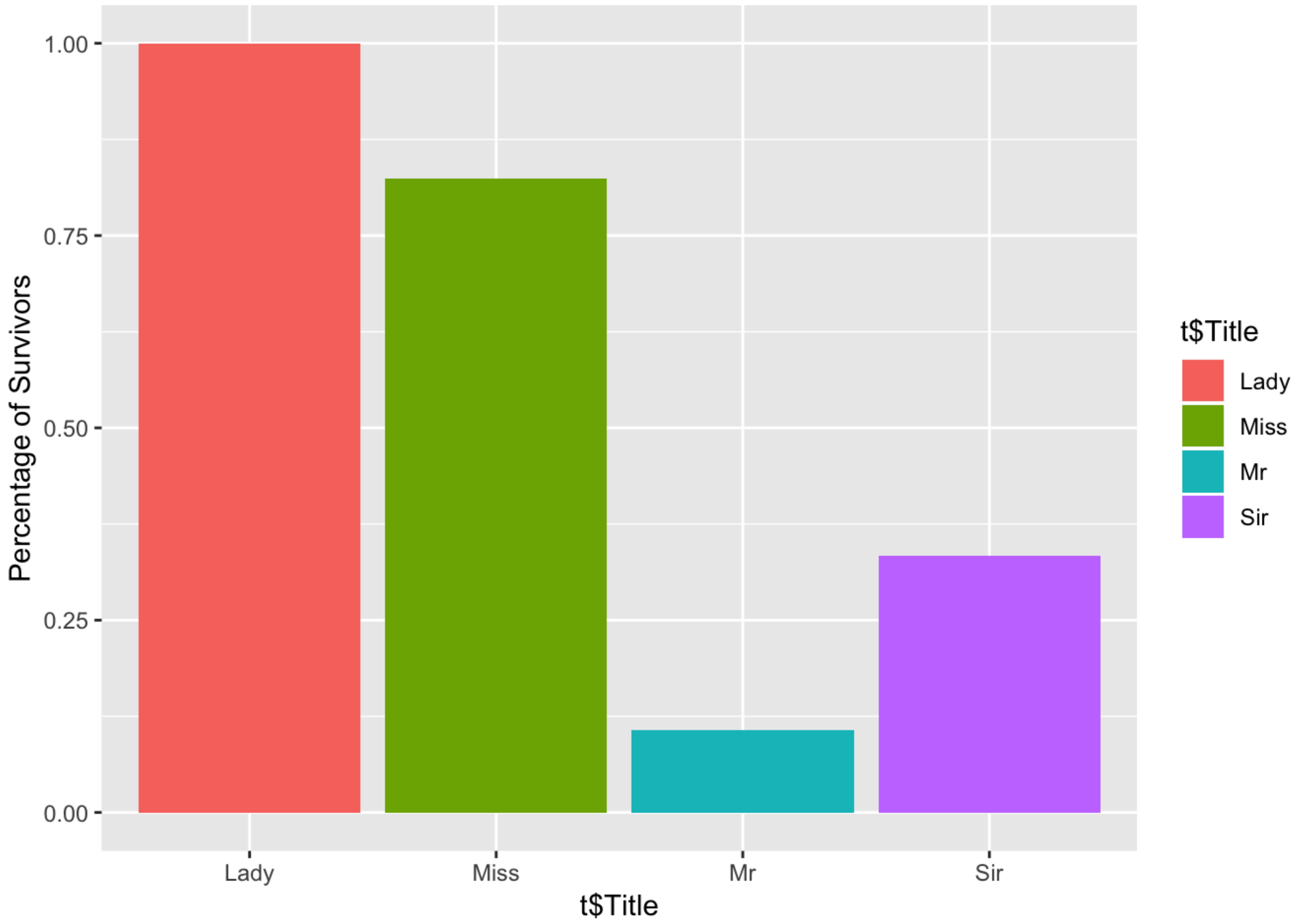
I don’t believe this shows anything conclusive about a higher priced ticket meaning you are more likely to be saved, however. A higher class/cabin, though did have a difference.

Below I will see if the elite title that I sparsed out earlier makes any difference:

```
SurvivalByVariable(t, t$Survived, t$Title)
```



```
SurvivalByVariablePercentage(t, t$Survived, t$Title)
```

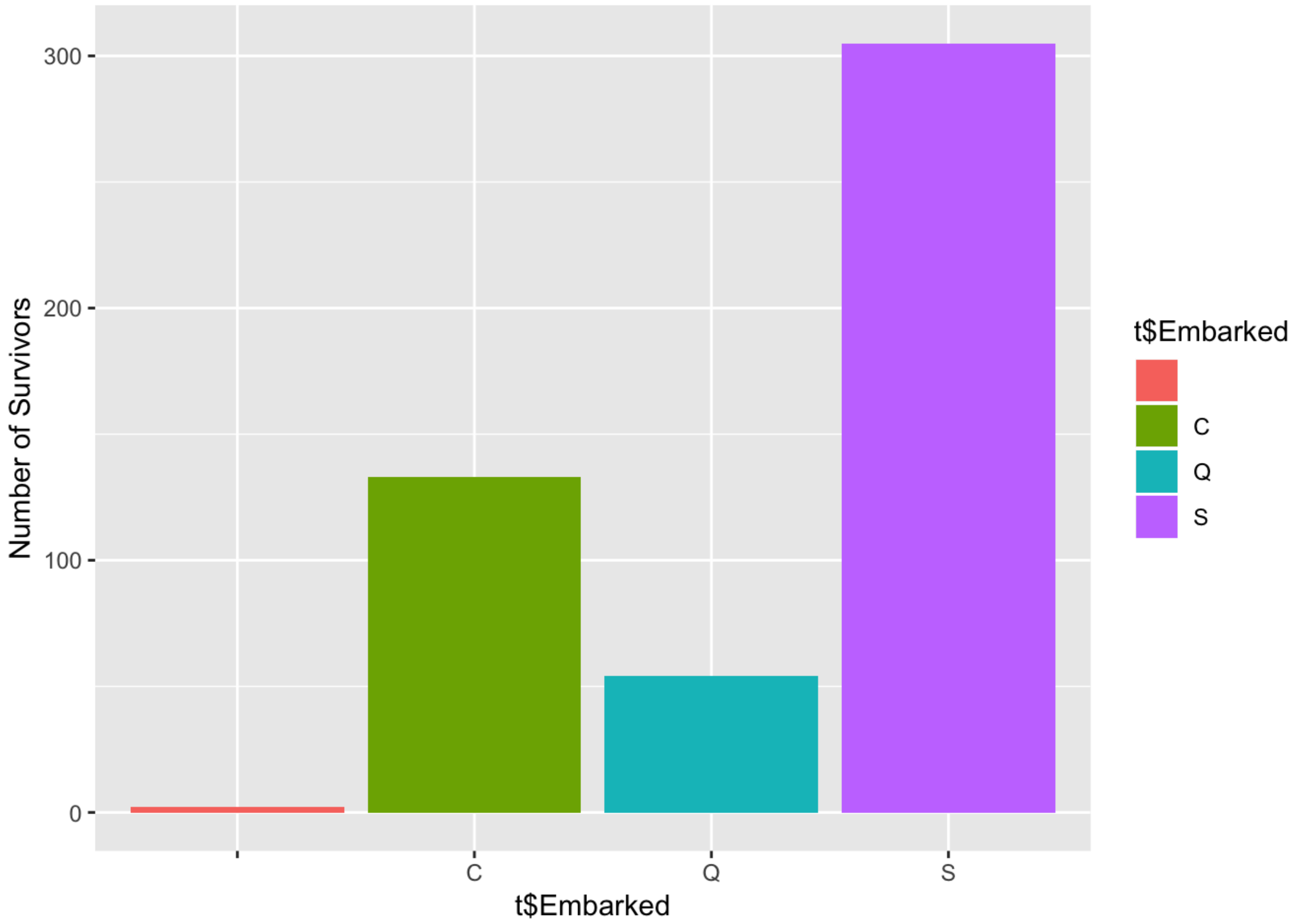


The more elite titles had a higher survival percentage rate. Clearly the limitation on this is that the denominator of the “Lady”or “Sir” group is much lower than “Miss” or “Mr”, but it is interesting that those with fancier titles were technically more likely to be saved.

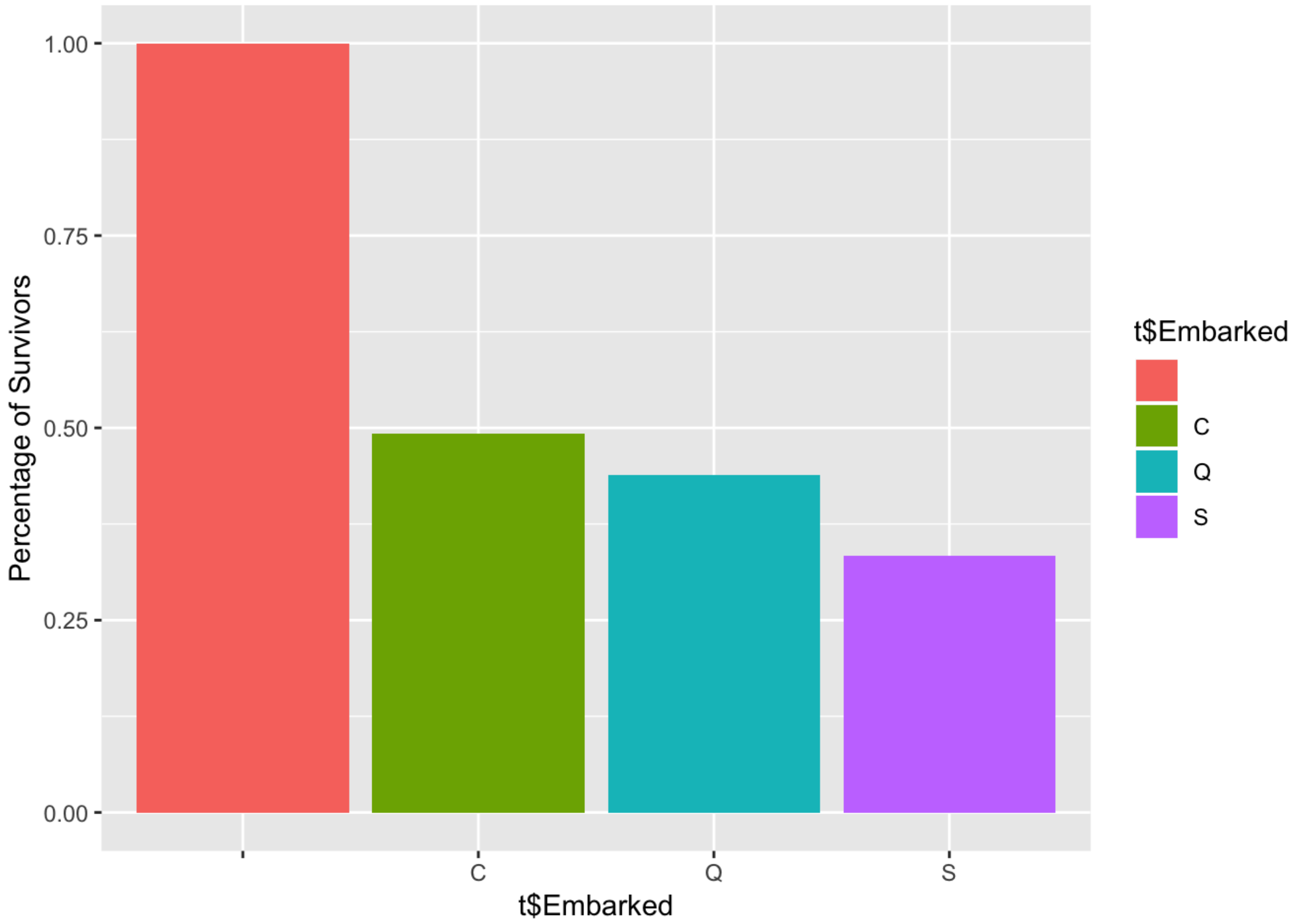
Geographic Location

I am trying to see if people who embarked from a certain port were more likely to be saved.

```
SurvivalByVariable(t, t$Survived, t$Embarked)
```

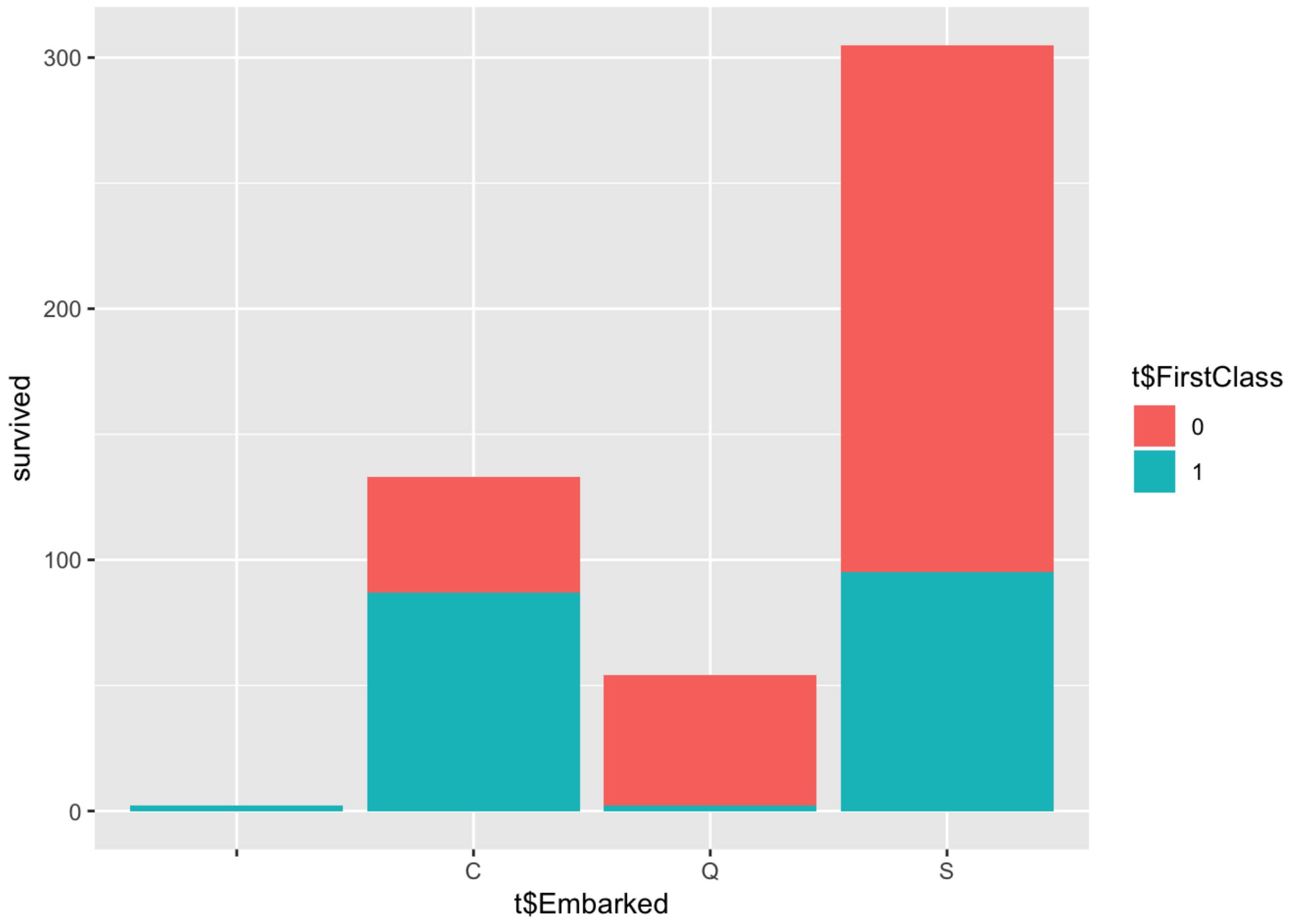


```
SurvivalByVariablePercentage(t, t$Survived, t$Embarked)
```

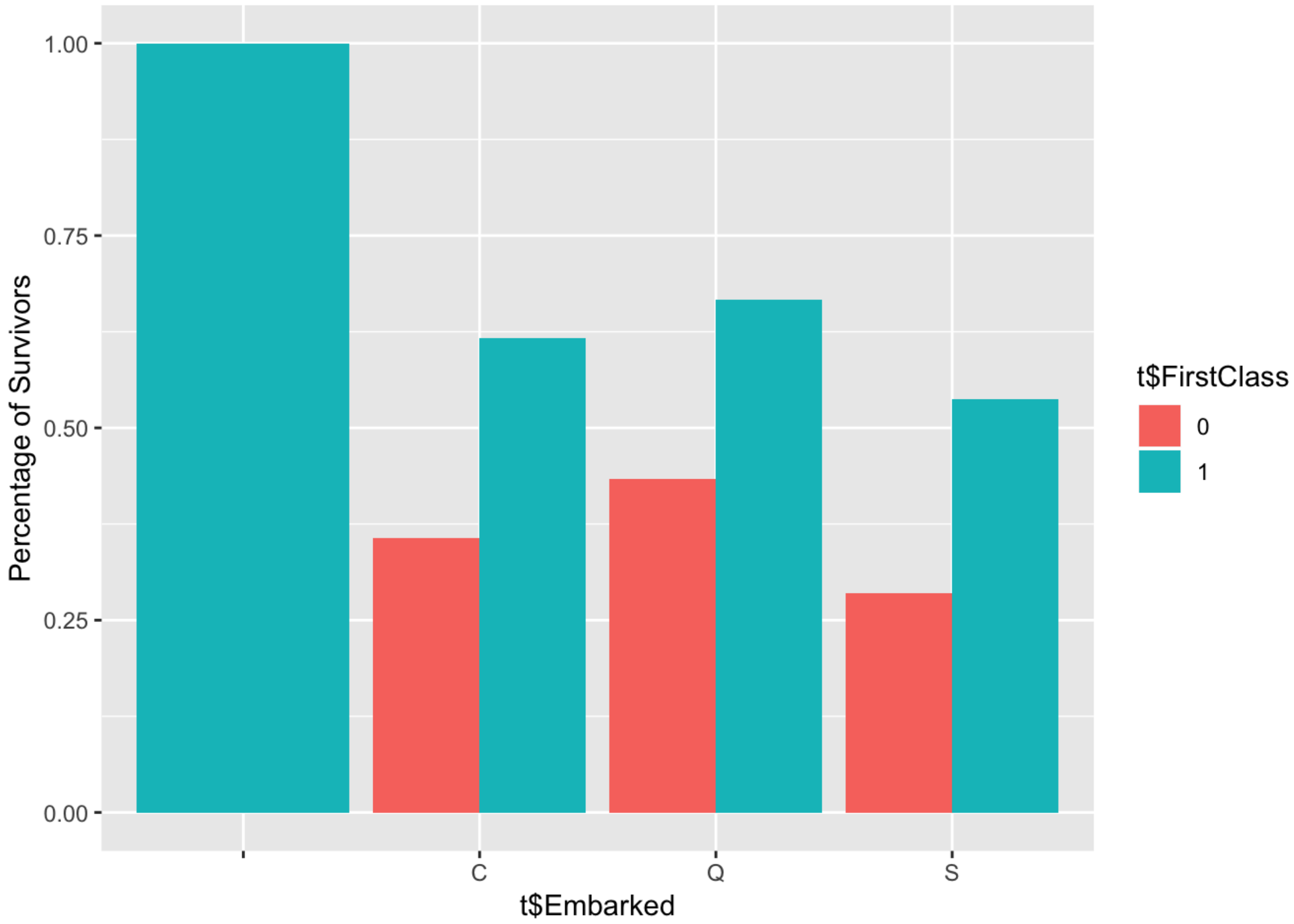


Looking at percentages, clearly most people left from Southampton. The percentage of survivals was not drastically higher. It is interesting to see if there is anything related to which social class may have be leaving from each port. Using the “first class” variable:

```
SurvivalTwoVariables(t, t$Survived, t$Embarked, t$FirstClass)
```



```
SurvivalTwoVariablesPercentage(t, t$Survived, t$Embarked, t$FirstClass)
```

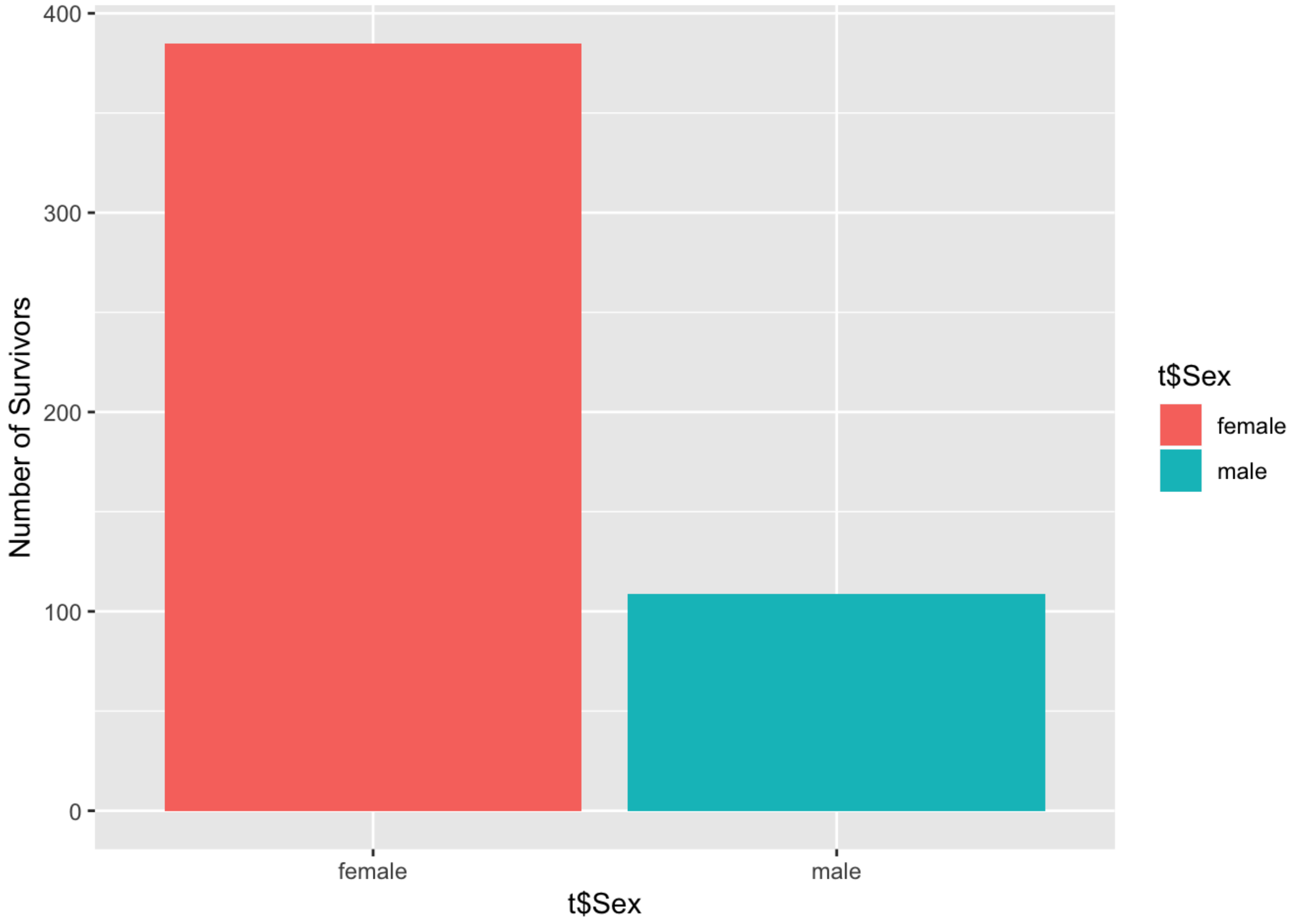


Here we see that the most people left from Southampton port, but the most people from First Class left from the Cherbourg port. Which port you left from did not make you that much more likely to survive.

Women and Children First?

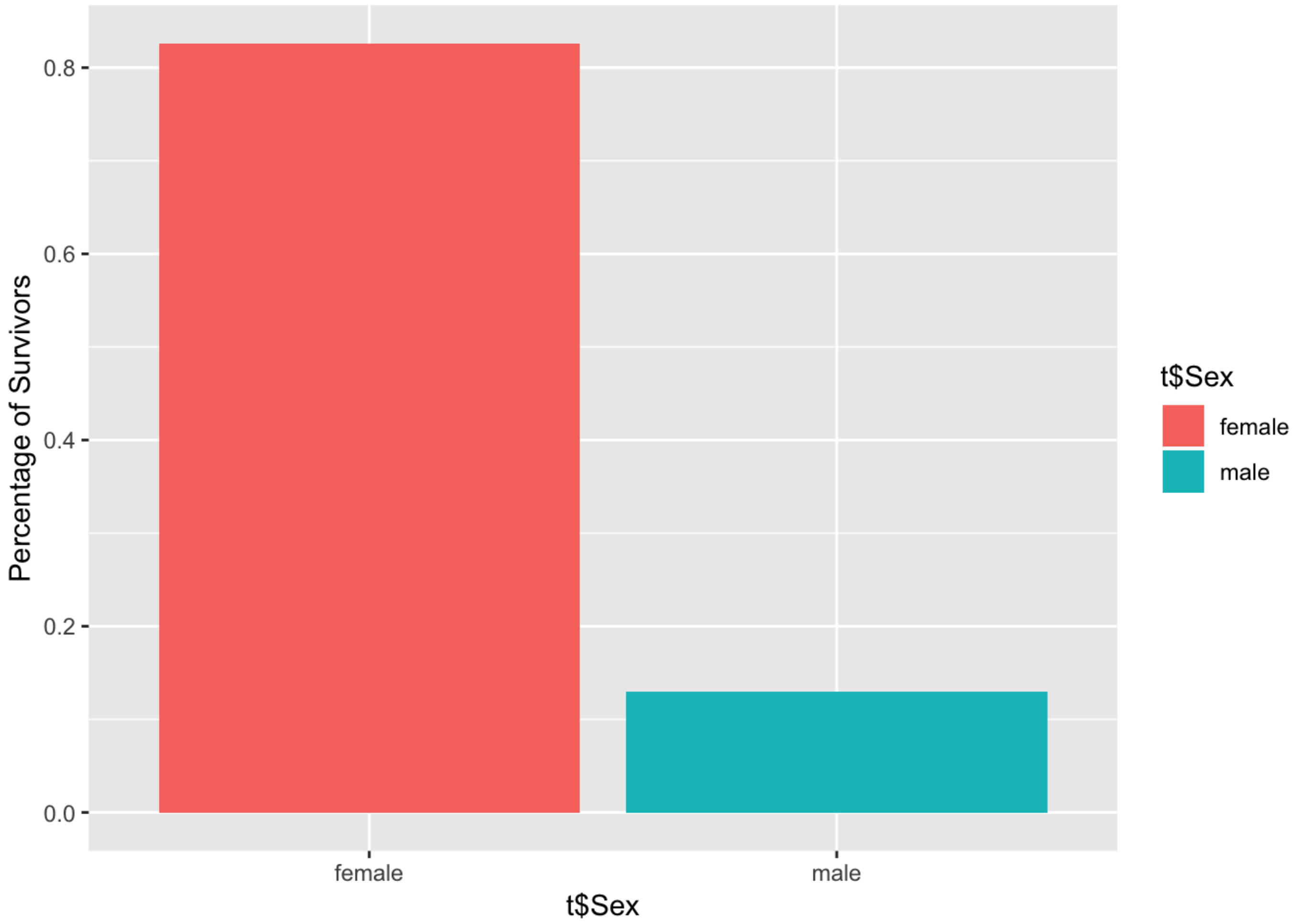
First, let’s look at women:

```
SurvivalByVariable(t, t$Survived, t$Sex)
```



Far more women survived than men. To make sure this isn’t a product of more women on the boat than men, let’s look at it as a function of percentage:

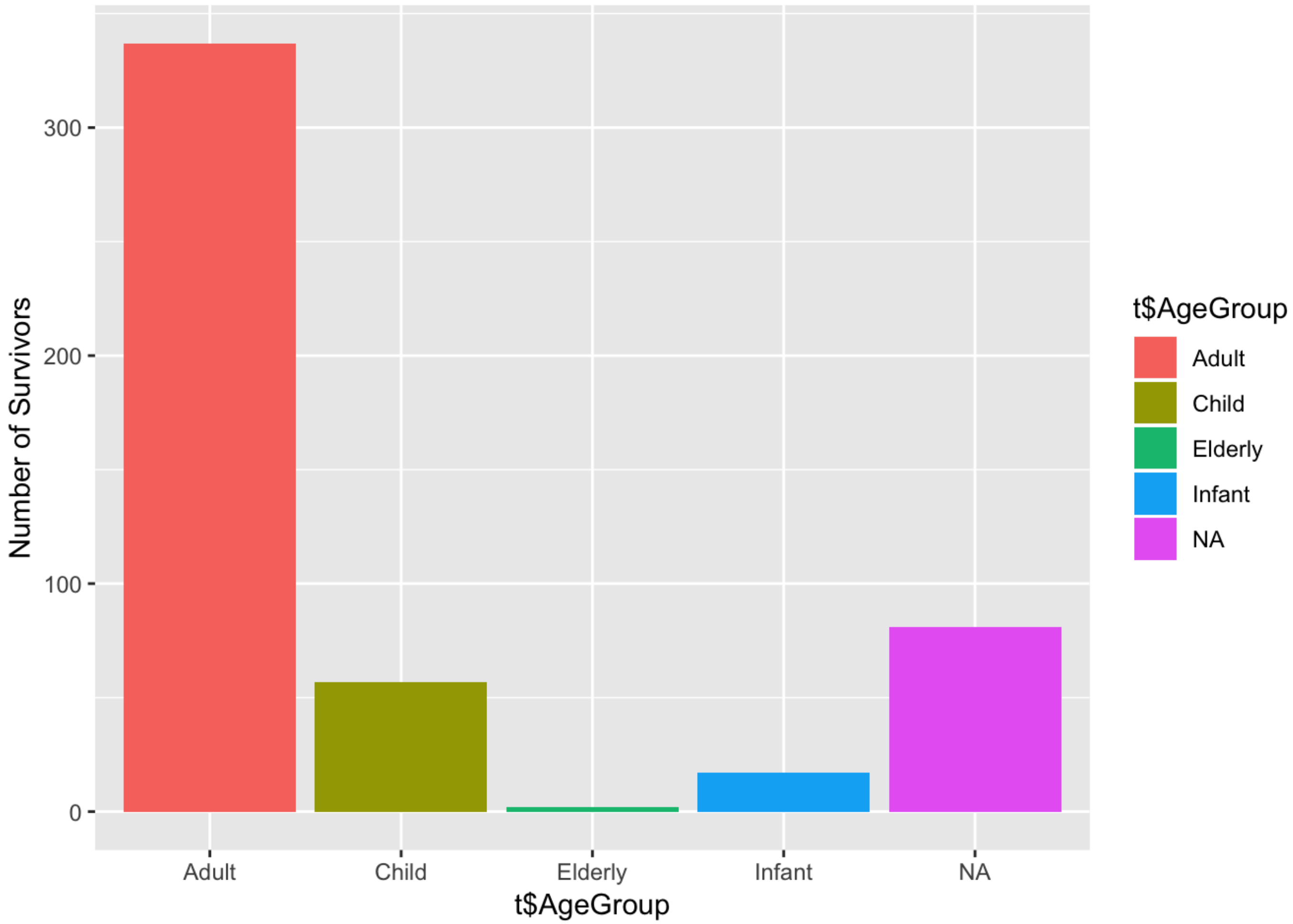
```
SurvivalByVariablePercentage(t, t$Survived, t$Sex)
```



Here we see over 80% of women survived while nearly the inverse is true for men (less than 20% survived).

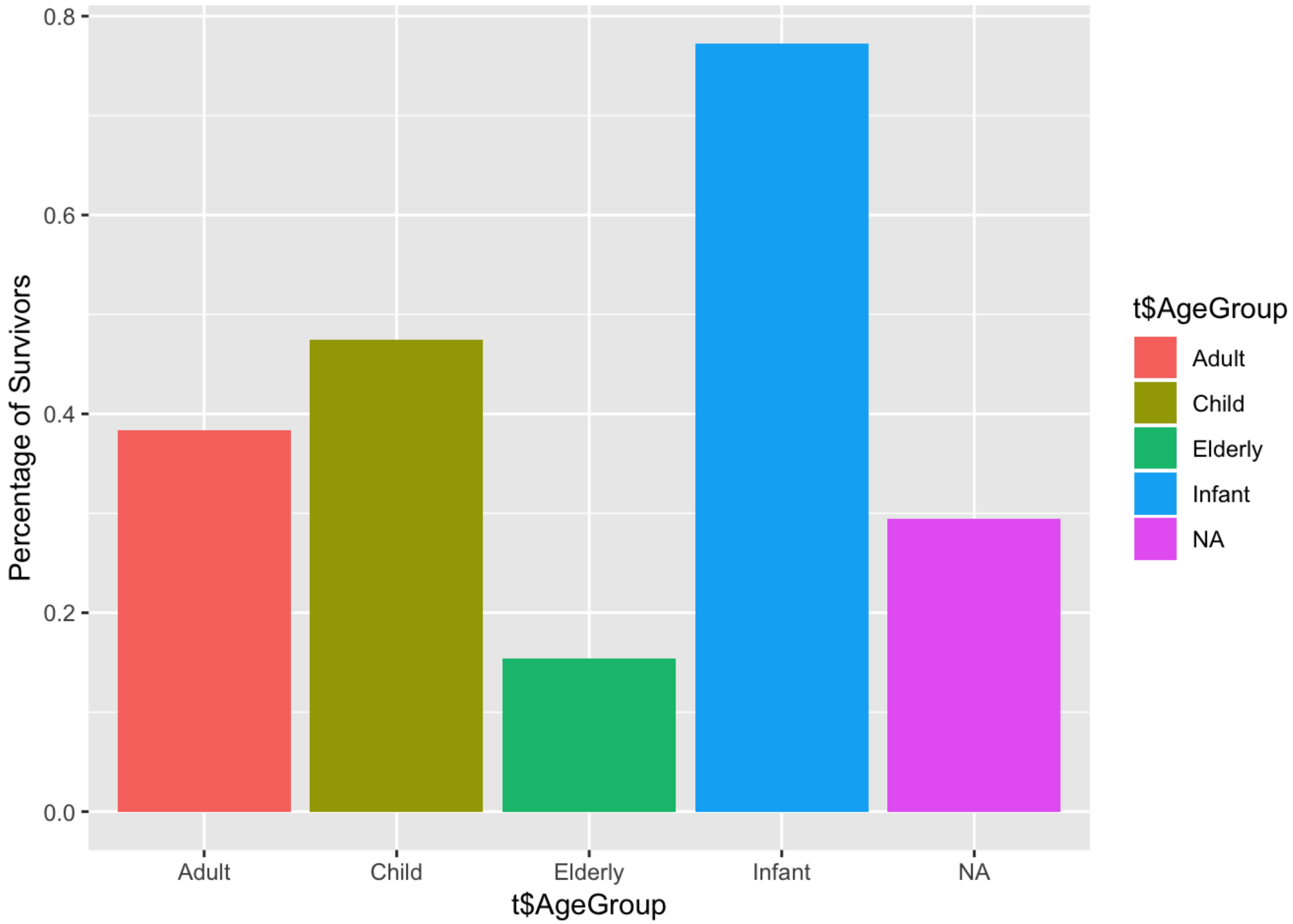
Now, let's look at age:

```
SurvivalByVariable(t, t$Survived, t$AgeGroup)
```



We get a better idea of what this means by looking at the percentage.

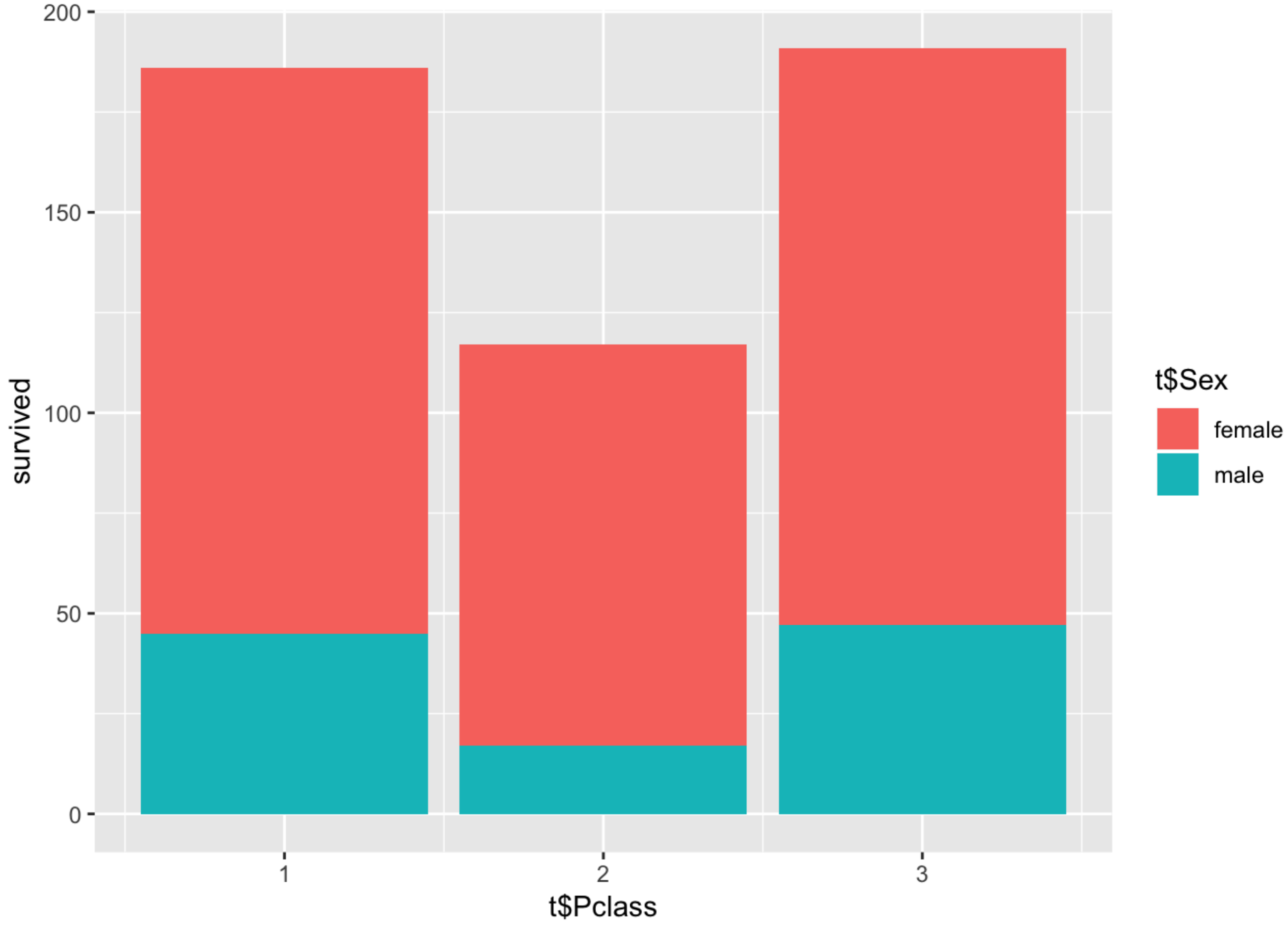
```
SurvivalByVariablePercentage(t, t$Survived, t$AgeGroup)
```



Children and infants had a higher percentage of survival.

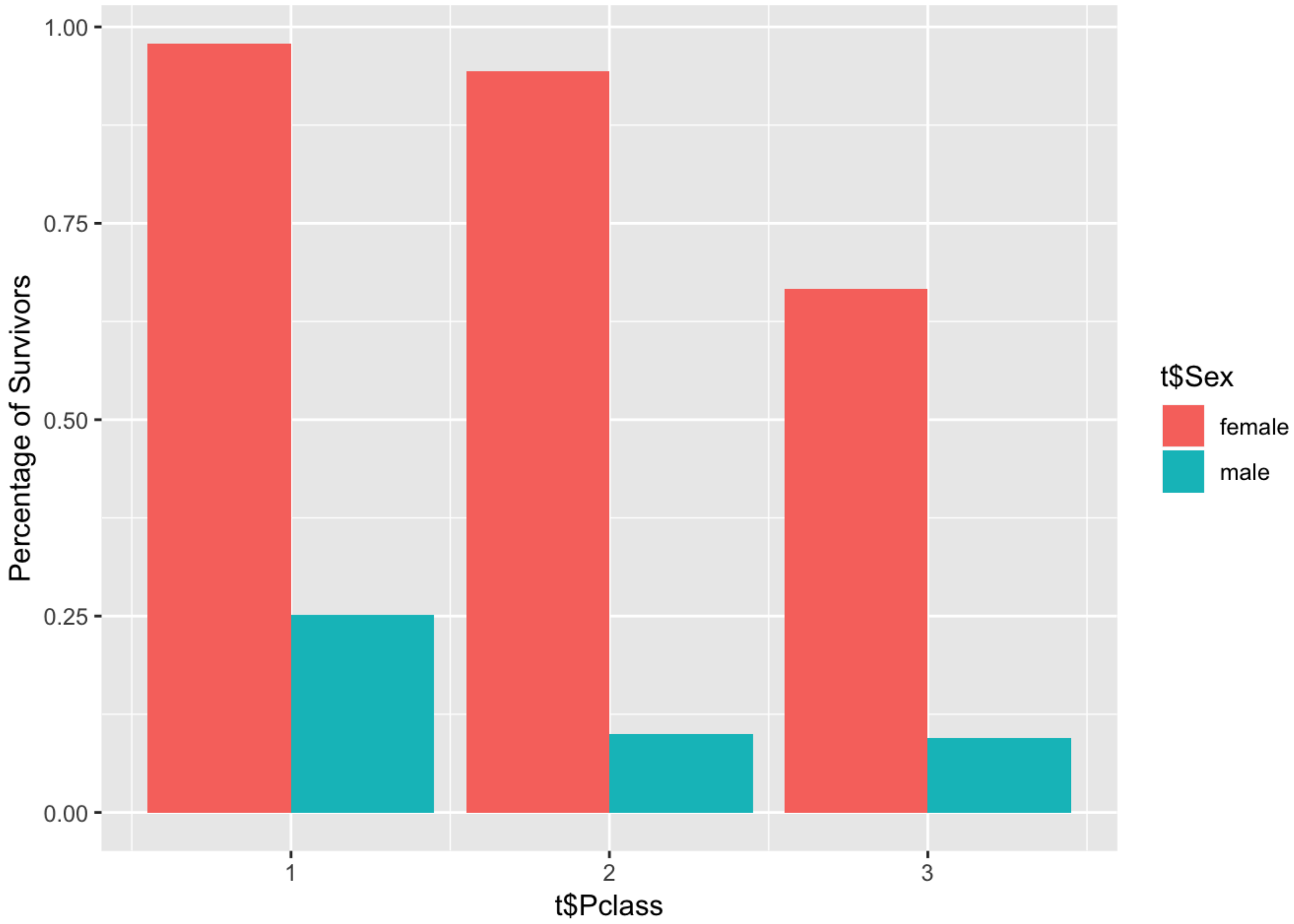
Rich Women and Children First?

```
SurvivalTwoVariables(t, t$Survived, t$Pclass, t$Sex)
```



The trend of women as a majority of survivors seems to still stand, despite class. But when we look at it as percentage:

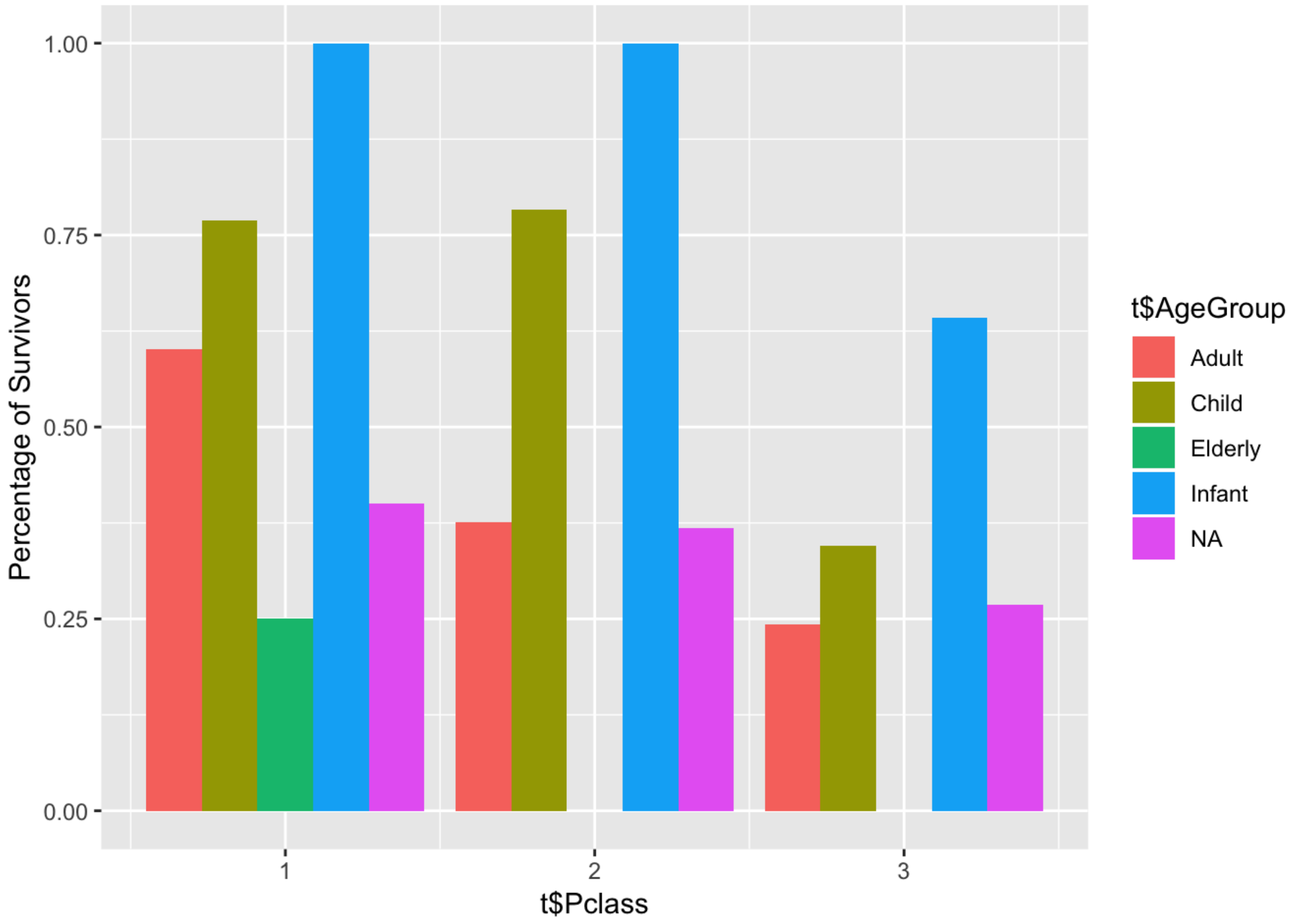
```
SurvivalTwoVariablesPercentage(t, t$Survived, t$Pclass, t$Sex)
```



A higher percentage of women in higher classes survived, and there is over a 20% drop between the second class and third class of precentage of women survived.

Let’s look at the same thing with children:

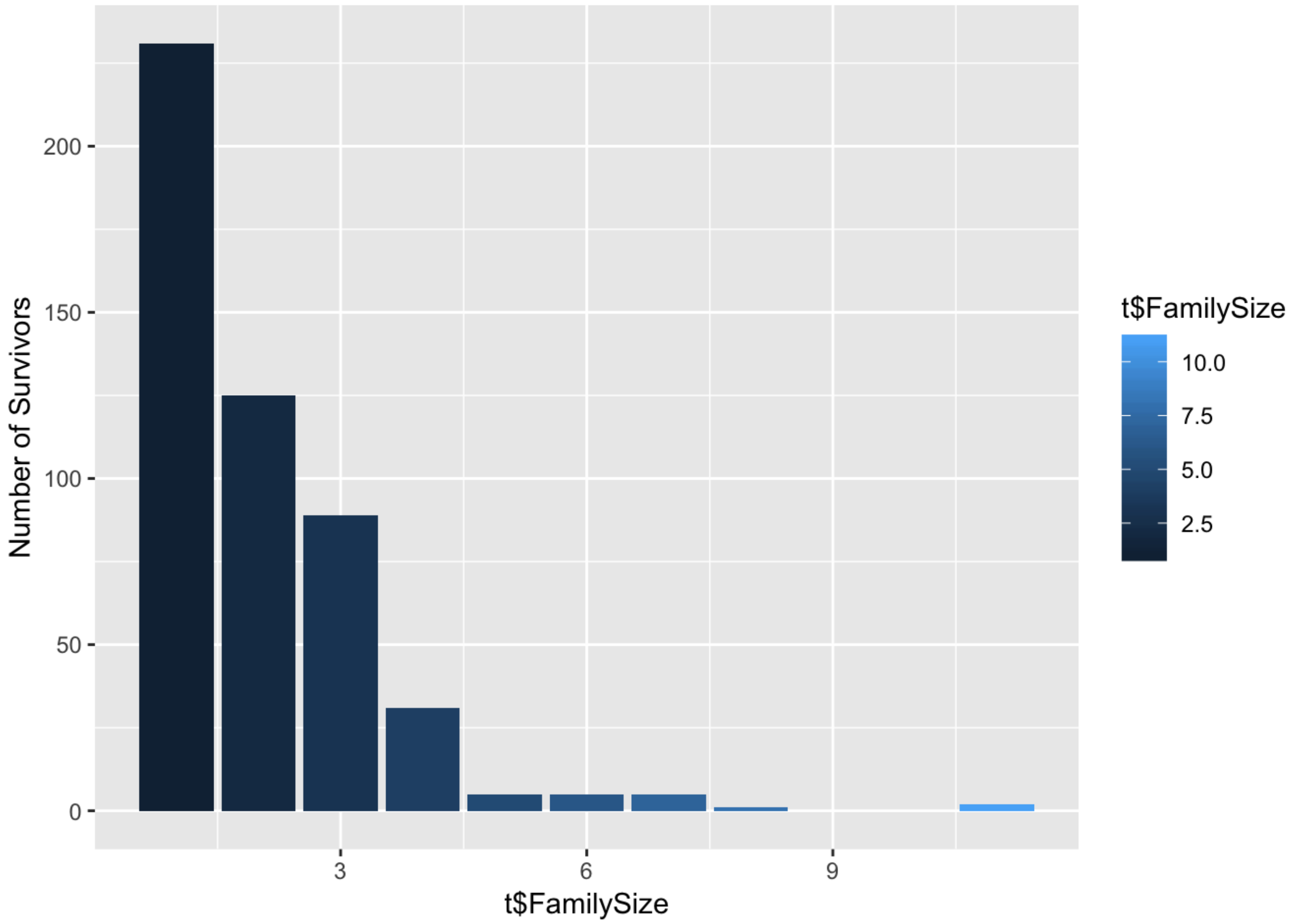
```
SurvivalTwoVariablesPercentage(t, t$Survived, t$Pclass, t$AgeGroup)
```



There is a similar trend to the graph of women and social class above. While infants and children were prioritized in all classes, there was clear drop off after the second class for the third class.

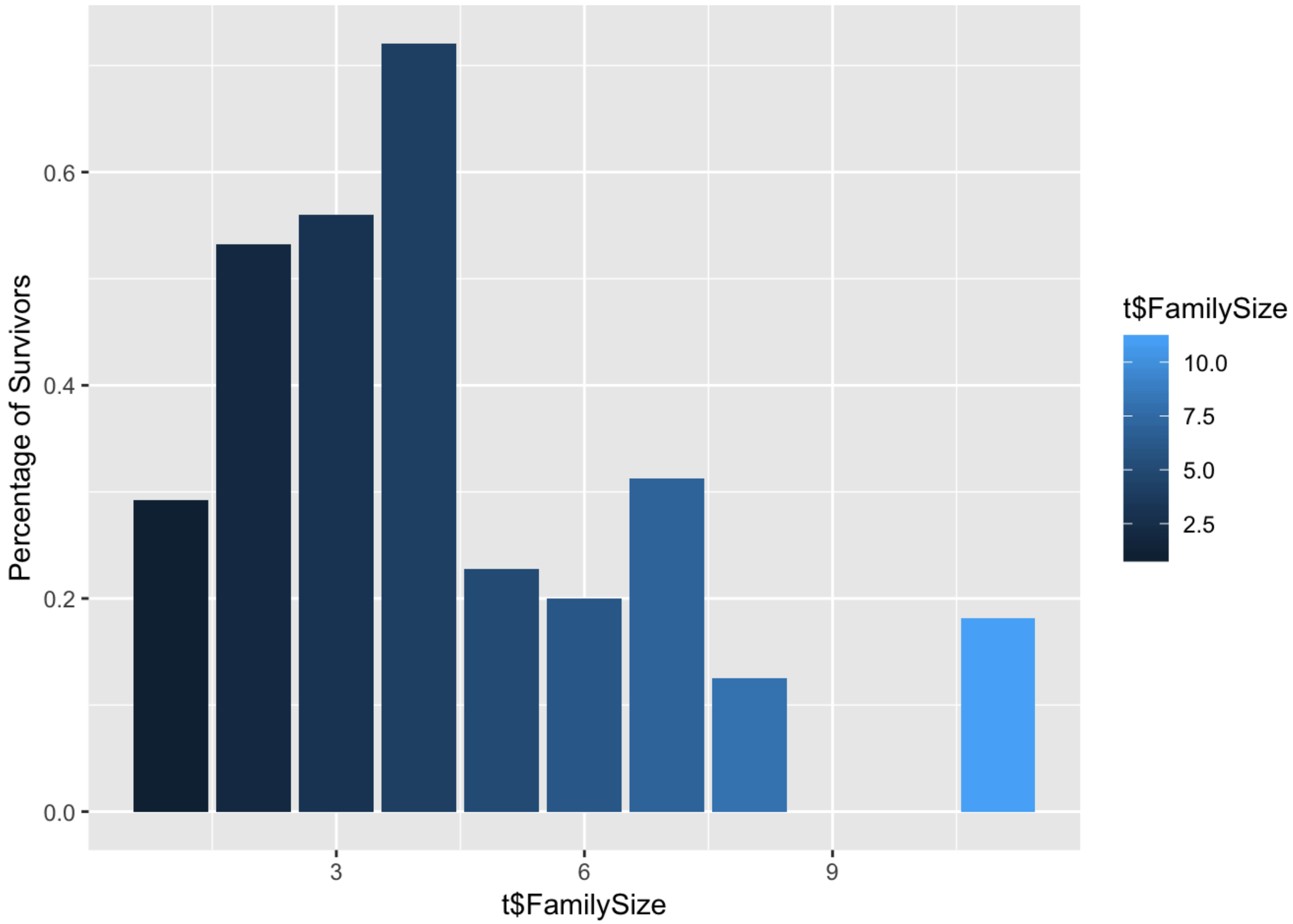
Families

```
SurvivalByVariable(t, t$Survived, t$FamilySize)
```



Here we might guess that if you were travelling alone, it may have been easier to save yourself, rather than worry about other family members. Let’s see if percentages work out for this, too:

```
SurvivalByVariablePercentage(t, t$Survived, t$FamilySize)
```



Families of size 2,3 and 4 had a higher percentage of survival. Maybe this has to do with women and children being paired together and therefore more likely to be saved? And the bigger the familiy, the harder to keep together?

Overall, there are limitations to this dataset and it is only around 1,300 entries, while there were nearly 1000 more on board (2,222). However, this is some interesting data to play with in terms of assumptions we make about trying to survive such a catastophre.