Summarize and Visualize One Categorical Variable

Step 1: Load the data set and define colors.

Step 2: Change Survived to a factor.

Use the factor() function to change `titanic$Survived` into a vector called SurvivedFactor, then replace the original `titanic$Survived` variable with the new vector SurvivedFactor.

```
str(titanic)
```

```
## 'data.frame':    1313 obs. of  5 variables:
##  $ Name    : Factor w/ 1310 levels "Abbing, Mr Anthony",..: 22 25 26 27 24 31 45 46 50 54 ...
##  $ PClass  : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : num  29 2 30 25 0.92 47 63 39 58 71 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 1 1 2 1 2 2 1 2 1 2 ...
##  $ Survived: int  1 0 0 0 1 1 1 0 1 0 ...
```

```
SurvivedFactor <- factor(titanic$Survived,
                         levels = c("0", "1"),
                         labels = c("No", "Yes"))
titanic$Survived <- SurvivedFactor # Note that here we're replacing the Survived column with the SurvivedFactor c
olumn, instead of adding it to the data set as a new column

head(titanic, n = 10)
```

```
##                                          Name PClass    Age    Sex Survived
## 1                    Allen, Miss Elisabeth Walton    1st 29.00 female      Yes
## 2                     Allison, Miss Helen Loraine    1st  2.00 female       No
## 3               Allison, Mr Hudson Joshua Creighton    1st 30.00   male       No
## 4  Allison, Mrs Hudson JC (Bessie Waldo Daniels)    1st 25.00 female       No
## 5                  Allison, Master Hudson Trevor    1st  0.92   male      Yes
## 6                             Anderson, Mr Harry    1st 47.00   male      Yes
## 7                 Andrews, Miss Kornelia Theodosia    1st 63.00 female      Yes
## 8                             Andrews, Mr Thomas, jr    1st 39.00   male       No
## 9   Appleton, Mrs Edward Dale (Charlotte Lamson)    1st 58.00 female      Yes
## 10                            Artagaveytia, Mr Ramon    1st 71.00   male       No
```

```
str(titanic)
```

```
## 'data.frame':    1313 obs. of  5 variables:
##  $ Name    : Factor w/ 1310 levels "Abbing, Mr Anthony",..: 22 25 26 27 24 31 45 46 50 54 ...
##  $ PClass  : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : num  29 2 30 25 0.92 47 63 39 58 71 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 1 1 2 1 2 2 1 2 1 2 ...
##  $ Survived: Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 1 2 1 ...
```

Look at the first 11 passengers in the titanic data:

```
##                                          Name PClass    Age    Sex Survived
## 1                    Allen, Miss Elisabeth Walton    1st 29.00 female      Yes
## 2                     Allison, Miss Helen Loraine    1st  2.00 female       No
## 3               Allison, Mr Hudson Joshua Creighton    1st 30.00   male       No
## 4  Allison, Mrs Hudson JC (Bessie Waldo Daniels)    1st 25.00 female       No
## 5                  Allison, Master Hudson Trevor    1st  0.92   male      Yes
## 6                             Anderson, Mr Harry    1st 47.00   male      Yes
## 7                 Andrews, Miss Kornelia Theodosia    1st 63.00 female      Yes
## 8                             Andrews, Mr Thomas, jr    1st 39.00   male       No
## 9   Appleton, Mrs Edward Dale (Charlotte Lamson)    1st 58.00 female      Yes
## 10                            Artagaveytia, Mr Ramon    1st 71.00   male       No
## 11                          Astor, Colonel John Jacob    1st 47.00   male       No
```

Step 3: Construct a frequency table.

Use the table() command in R to construct a *frequency table*. Note that here we're only working with the first 11 passengers in the data set.

```
table(head(titanic$Survived, n=11)) # Make a table of the number of passengers that didn't survive and the number
that did, from only the first 11 passengers.
```

```
##
##  No Yes
##   6   5
```

Step 4: Make a relative frequency table.

Count *relative frequency* (proportion) of survivors and non-survivors in the data.

```
# Manually check the proportion of survivors and non-survivors
6/(6+5) # Proportion that didn't survive
```

```
## [1] 0.5454545
```

```
5/(6+5) # Proportion that survived
```

```
## [1] 0.4545455
```

```
prop.table(table(head(titanic$Survived, n=11))) # Use the prop.table() function to do this automatically
```

```
##
##        No       Yes
## 0.5454545 0.4545455
```

Step 5: Make frequency and relative frequency tables of all passenger survival.

Next, calculate the frequency and relative frequency for the entire titanic data set.

```
tbl.titanic <- table(titanic$Survived)
tbl.titanic
```

```
##
##  No Yes
## 863 450
```

```
prop.table(tbl.titanic)
```

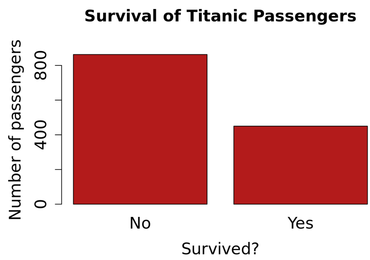
```
##
##        No       Yes
## 0.6572734 0.3427266
```

Step 6: Make a barplot of the survival of Titanic passengers.

Use the table you create of passenger survival to create a barplot with the command barplot().

```
counts = table(titanic$Survived) # counts is a frequency table of the titanic$Survived column

par(mar=2+c(5.1,4.1,4.1,2.1)) # Set the margin around the plot
barplot(counts, # counts tells R the number of bars to make and the height of those bars
        main="Survival of Titanic Passengers",
        col = crimson,
        ylab="Number of passengers", xlab="Survived?",
        cex.axis=1.5, cex.main=1.5,
        cex.names=1.5, cex.lab=1.5)
```



Step 7: Make a barplot of the survival proportions of Titanic passengers.

Plot the proportion of survivors by dividing counts by the total number of passengers, then making a new barplot.

```
totnum = sum(counts)
par(mar=2+c(5.1,4.1,4.1,2.1))
barplot(counts / totnum,
        main="Survival of Titanic Passengers", col = crimson,
        ylab="Proportion of passengers", xlab="Survived?",
        ylim=c(0,1),cex.axis=1.5, cex.main=1.5,
        cex.names=1.5, cex.lab=1.5)
```