R Notebook

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

plot(cars)

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

library(ggplot2) # includes ggplot2 package  
library(dplyr) # includes 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

ship <- read.csv("/Users/ajaykommineni/Downloads/train\_and\_test2.csv") # file path is set to variable and reads the file  
  
str(ship)

## 'data.frame': 1309 obs. of 28 variables:  
## $ Passengerid: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : num 22 38 26 35 35 28 54 2 27 14 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Sex : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ sibsp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ zero : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.2 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.3 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.4 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.5 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.6 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ zero.7 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.8 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.9 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.10 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.11 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.12 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.13 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.14 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ zero.15 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.16 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Embarked : int 2 0 2 2 2 1 2 2 2 0 ...  
## $ zero.17 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ zero.18 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ X2urvived : int 0 1 1 1 0 0 0 0 1 1 ...

# To print the structure of the data set  
head(ship)

## Passengerid Age Fare Sex sibsp zero zero.1 zero.2 zero.3 zero.4 zero.5  
## 1 1 22 7.2500 0 1 0 0 0 0 0 0  
## 2 2 38 71.2833 1 1 0 0 0 0 0 0  
## 3 3 26 7.9250 1 0 0 0 0 0 0 0  
## 4 4 35 53.1000 1 1 0 0 0 0 0 0  
## 5 5 35 8.0500 0 0 0 0 0 0 0 0  
## 6 6 28 8.4583 0 0 0 0 0 0 0 0  
## zero.6 Parch zero.7 zero.8 zero.9 zero.10 zero.11 zero.12 zero.13 zero.14  
## 1 0 0 0 0 0 0 0 0 0 0  
## 2 0 0 0 0 0 0 0 0 0 0  
## 3 0 0 0 0 0 0 0 0 0 0  
## 4 0 0 0 0 0 0 0 0 0 0  
## 5 0 0 0 0 0 0 0 0 0 0  
## 6 0 0 0 0 0 0 0 0 0 0  
## Pclass zero.15 zero.16 Embarked zero.17 zero.18 X2urvived  
## 1 3 0 0 2 0 0 0  
## 2 1 0 0 0 0 0 1  
## 3 3 0 0 2 0 0 1  
## 4 1 0 0 2 0 0 1  
## 5 3 0 0 2 0 0 0  
## 6 3 0 0 1 0 0 0

# displays the top few columns of the data set  
summary(ship)

## Passengerid Age Fare Sex   
## Min. : 1 Min. : 0.17 Min. : 0.000 Min. :0.000   
## 1st Qu.: 328 1st Qu.:22.00 1st Qu.: 7.896 1st Qu.:0.000   
## Median : 655 Median :28.00 Median : 14.454 Median :0.000   
## Mean : 655 Mean :29.50 Mean : 33.281 Mean :0.356   
## 3rd Qu.: 982 3rd Qu.:35.00 3rd Qu.: 31.275 3rd Qu.:1.000   
## Max. :1309 Max. :80.00 Max. :512.329 Max. :1.000   
##   
## sibsp zero zero.1 zero.2 zero.3 zero.4   
## Min. :0.0000 Min. :0 Min. :0 Min. :0 Min. :0 Min. :0   
## 1st Qu.:0.0000 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0.0000 Median :0 Median :0 Median :0 Median :0 Median :0   
## Mean :0.4989 Mean :0 Mean :0 Mean :0 Mean :0 Mean :0   
## 3rd Qu.:1.0000 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :8.0000 Max. :0 Max. :0 Max. :0 Max. :0 Max. :0   
##   
## zero.5 zero.6 Parch zero.7 zero.8 zero.9   
## Min. :0 Min. :0 Min. :0.000 Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0.000 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0.000 Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0.385 Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0.000 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :9.000 Max. :0 Max. :0 Max. :0   
##   
## zero.10 zero.11 zero.12 zero.13 zero.14 Pclass   
## Min. :0 Min. :0 Min. :0 Min. :0 Min. :0 Min. :1.000   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.:2.000   
## Median :0 Median :0 Median :0 Median :0 Median :0 Median :3.000   
## Mean :0 Mean :0 Mean :0 Mean :0 Mean :0 Mean :2.295   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.:3.000   
## Max. :0 Max. :0 Max. :0 Max. :0 Max. :0 Max. :3.000   
##   
## zero.15 zero.16 Embarked zero.17 zero.18   
## Min. :0 Min. :0 Min. :0.000 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:1.000 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :2.000 Median :0 Median :0   
## Mean :0 Mean :0 Mean :1.493 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:2.000 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :2.000 Max. :0 Max. :0   
## NA's :2   
## X2urvived   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.2613   
## 3rd Qu.:1.0000   
## Max. :1.0000   
##

#shows all the statics of the data

# 1. Survival rate by passenger class

# Survival rate by passenger class  
  
# Create the survival rate by passenger class plot  
ggplot(ship, aes(x = factor(Pclass), fill = factor(X2urvived))) +   
 # plotting Pclass as x-axis and X2urvived as fill  
 geom\_bar(position = "fill") +   
 # Creates bars with proportions (0 to 1)  
 scale\_y\_continuous(labels = scales::percent) + # Convert y-axis proportions to percentages  
 scale\_fill\_manual(values = c("0" = "red", "1" = "green"),   
 # Custom colors for survival status  
 labels = c("Did Not Survive", "Survived")) +   
 # Add labels for the legend  
 labs(title = "Survival Rate by Passenger Class",   
 # Adding chart title and labels of X and Y axis  
 x = "Passenger Class",  
 y = "Percentage",   
 fill = "Survival Status") +  
 theme\_minimal()

#Findings  
# The survival rate for second-class passengers is lower than that of first-class passengers but higher than that of third-class passengers. # A far greater percentage of first-class passengers survived than those in second and third classes.  
# The percentage of third-class passengers who survived was far lower.  
# During the evacuation, third-class passengers were probably given the least priority because they were situated on the lower decks, further away from the lifeboats.

# 2. Survival rate by age of the passenger

# Age distribution by survival  
ggplot(ship, aes(x = Age, fill = factor(X2urvived))) +  
 # ggplot with data and uses aes mapping x-axis to age and the fills accordingly with factor(X2urvived)  
 geom\_histogram(binwidth = 5, position = "identity", alpha = 0.6) +  
 # Create histogram with binwidth of 5, overlapping bars, and transparency  
 scale\_fill\_manual(values = c("0" = "red", "1" = "green"),   
 # colors for survival status   
 labels = c("Did Not Survive", "Survived")) +  
   
 #Add labels for the legend  
 labs(title = "Age Distribution by Survival",  
 # adding labels to the x-axis, y-axis and the title  
 x = "Age",  
 y = "Count",  
 fill = "Survival Status") +  
 theme\_bw()

#Findings  
# The age distribution of the survivors is displayed in the plot. The individuals who survived were obviously between the ages of 20 and 40. Additionally, most infants between the ages of 0 and 10 survive. It demonstrates that they received top priority. This graphic is often useful for understanding the distribution of population age versus survival.

#3 show the counts of survival by number of parents/children aboard

# Convert the 'Survived' column to a factor with meaningful labels  
ship$Survived <- factor(ship$X2urvived, labels = c("No", "Yes"))  
# Create the plot  
ggplot(ship, aes(x = Parch, fill = Survived)) +  
 # Use a bar plot to show counts of survival by number of parents/children aboard  
 geom\_bar(position = "stack") +  
 # Add labels and title to the plot  
 labs(  
 title = "Survival by Number of Parents/Children Aboard",  
 x = "Number of Parents/Children Aboard (Parch)",  
 y = "Count",  
 fill = "Survived"  
 ) +  
 # Use a minimal theme for a clean look  
 theme\_minimal()

#findings  
#The plot indicates that survival rates were significantly impacted by family size, or the number of parents and children on board and Traveling Alone has Lower survival rates.Traveling with Family: Higher survival rates, especially for larger families. This is consistent with historical reports of the Titanic disaster, which show that families traveling together had greater access to lifeboats and that women and children were given priority during evacuation.

#4 Comprehensive Facet Grid of Survival Factors

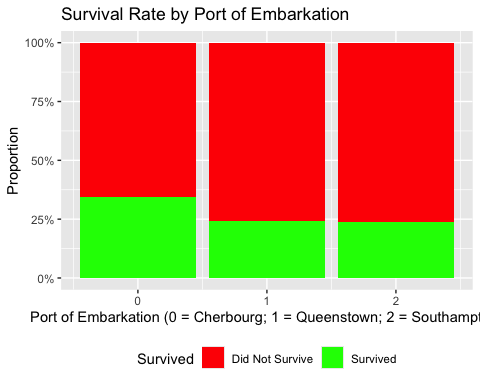
# Convert the 'X2urvived' column to a factor with levels for survival status  
ship$urvived <- factor(ship$X2urvived, labels = c("No", "Yes"))  
  
# Create the plot  
ggplot(ship) +   
 # x = Age, y = Fare, color = Survival status  
 geom\_point(aes(x = Age, y = Fare, color = urvived), alpha = 0.7) +  
 # Create a facet grid: Rows = Sex, Columns = Pclass  
 facet\_grid(Sex ~ Pclass) +  
 # Adding plot labels and title to the plot  
 labs(  
 x = "Age",   
 y = "Fare",   
 color = "Survived",   
 title = "Survival, Age, and Fare Across Gender and Passenger Class"  
 ) +  
 theme(  
 strip.background = element\_rect(fill = "gray"),   
 # Light gray background for facet labels  
 strip.text = element\_text(color = "black")   
 # Black text for facet labels  
 )

#findings  
# I made use of every variable in the dataframe in this plot. The x-axis represents the fare variable, while the y-axis represents the age variable. The survival in all variables format is revealed by charting the survival variable using the plot's color, and all plots are ultimately separated according to classes.  
#The survival rate was higher for females than for males.Passenger Class: The survival rates of the first and second classes were higher than those of the third class. The survival rate was lower for senior passengers and greater for children.Higher survival rates were associated with higher fares, which indicate a higher socioeconomic position.

#5 Create the survival rate by port of embarkation plot

# Creating the survival rate by port of embarkation plot  
ggplot(ship, aes(x = Embarked, fill = factor(X2urvived))) +   
 # Use Embarked as x-axis and Survived as fill  
 # Use geom\_bar with position = "fill" to create a proportional stacked bar chart  
 geom\_bar(position = "fill") +   
 # Create stacked bars with proportions (0 to 1)  
 scale\_fill\_manual(values = c("0" = "red", "1" = "green"), # Custom colors for survival status  
 labels = c("Did Not Survive", "Survived")) +   
 # Add labels for the legend  
 scale\_y\_continuous(labels = scales::percent) + # Convert y-axis proportions to percentages  
 labs(title = "Survival Rate by Port of Embarkation", x = "Port of Embarkation (0 = Cherbourg; 1 = Queenstown; 2 = Southampton)", y = "Proportion", fill = "Survived") + # Label for the legend  
 theme(legend.position = "bottom") # Move legend to the bottom

## Warning: Removed 2 rows containing non-finite outside the scale range  
## (`stat\_count()`).



#findings  
#plot clearly shows that survival doesn't depend on the port of embarkation  
#THe Queenstown, Southampton, and Cherbourg had the lowest survival rates. This implies that variations in passenger demographics, ticket class distribution, or other factors may have affected survival outcomes at the port of embarkation.  
#plot clearly shows that survival doesn't depend on the port of embarkation