Project Part 1

Description of Data Set

Background Information

The data set I am using for my project is called "Survival of Passengers on the Titanic." The data set contains information on the survival status, sex, age, and passenger class of 1,309 passengers in the Titanic disaster of 1912. In total, there were 2,240 passengers on the Titanic and more than 1,500 died. The Titanic sinking is a very historic, disastrous event. It would be interesting to explore what factors played a role in a passenger's likelihood of survival. This data frame contains 1,309 observations with several variables, including name, survived status, sex, age, and passenger class. The data set was begun by various researchers. The main source of this data is from the Encyclopedia Titanica, an online source of information about the Titanic. The data set was begun by various researchers. One of the original sources is Eaton & Haas (1994) "Titanic: Triumph and Tragedy", Patrick Stephens Ltd, which includes a passenger list created by many researchers and edited by Michael A. Findlay. Other researchers have updated and improved the titanic data frame to include previously missing data and new variables. For the purposes of my project, however, I will be using the older, smaller data set, TitanicSurvival.

```
library(ggplot2)
Titanic <- read.csv("~/Downloads/TitanicSurvival.csv")
```

```
## Reference 1
## Reference 2
```

Explanation of Rows and Variables

TitanicSurvival is a data frame with 1309 observations on the following 4 variables: survived (no or yes), sex (female or male), age (in years–and for some children, fractions of a year), passengerClass (1st, 2nd, or 3rd).

For the purposes of my project, I will be focusing on the variables: survived, sex, and passengerClass. I want to explore which variables impacted a passenger's chance at survival. Each observation or row represents a different passenger that contains information about the associated passenger. The data set contains each of their names if applicable.

Data Collection

This data set is a population of 1309 passengers on the Titanic. Although it is a subset of a larger Titanic passenger population, it is still its own population. For the purposes of my project, I want to draw conclusions about the entire population of 1309 passengers. The data was collected from various researchers who were interested in finding information about passengers on the Titanic. The data was collected from the online database, Encyclopedia Titanica, which has since grown. One of the original sources is Eaton & Haas (1994) "Titanic: Triumph and Tragedy", Patrick Stephens Ltd, which includes a passenger list created by many researchers and edited by Michael A. Findlay. I will be using all 1309 observations/passengers in my data exploration.

Potential Issues

There are some potential issues with this data set. There are missing values for the age variable for some passengers, as this information was unknown. Because of this, I will not be exploring the age variable in my data exploration.

Numerical Representation

```
# Subsetting data to just one variable (sex and passengerClass)
Sex <- Titanic['sex']
class <- Titanic['passengerClass']

# Finding female proportion that survived
Female <- Titanic[which(Titanic$sex=='female' &
          Titanic$survived=='yes'),]
Femaleprop <- nrow(Female)/sum(Sex=='female')

# Finding male proportion that survived
Male <-Titanic[which(Titanic$sex=='male' &
```

```
       Titanic$survived=='yes'),]
Maleprop <- nrow(Male)/sum(Sex=='male')

# Finding 1st class proportion that survived
FirstClass <- Titanic[which(Titanic$passengerClass=='1st' &
              Titanic$survived=='yes'),]
FirstClassprop <- nrow(FirstClass)/sum(class=='1st')

# Finding 2nd class proportion that survived
SecondClass <- Titanic[which(Titanic$passengerClass=='2nd' &
               Titanic$survived=='yes'),]
SecondClassprop <- nrow(SecondClass)/sum(class=='2nd')

# Finding 3rd class proportion that survived
ThirdClass <- Titanic[which(Titanic$passengerClass=='3rd' &
              Titanic$survived=='yes'),]
ThirdClassprop <- nrow(ThirdClass)/sum(class=='3rd')

# Combining proportions into matrix and transforming them into percentages
Props <- round(matrix(c(Femaleprop,Maleprop,FirstClassprop,
         SecondClassprop,ThirdClassprop)*100),1)

# Adding row and column names
rownames(Props)<- c("Female", "Male", "1st Class", "2nd Class", "3rd Class")
colnames(Props) <- c("Survival Rate Percentage")

Props
```

```
##           Survival Rate Percentage
## Female                        72.7
## Male                          19.1
## 1st Class                     61.9
## 2nd Class                     43.0
## 3rd Class                     25.5
```

For my numerical representation, I wanted to explore the different variables and what impact they would have on a passenger's survival percentage. I created a matrix that contains the

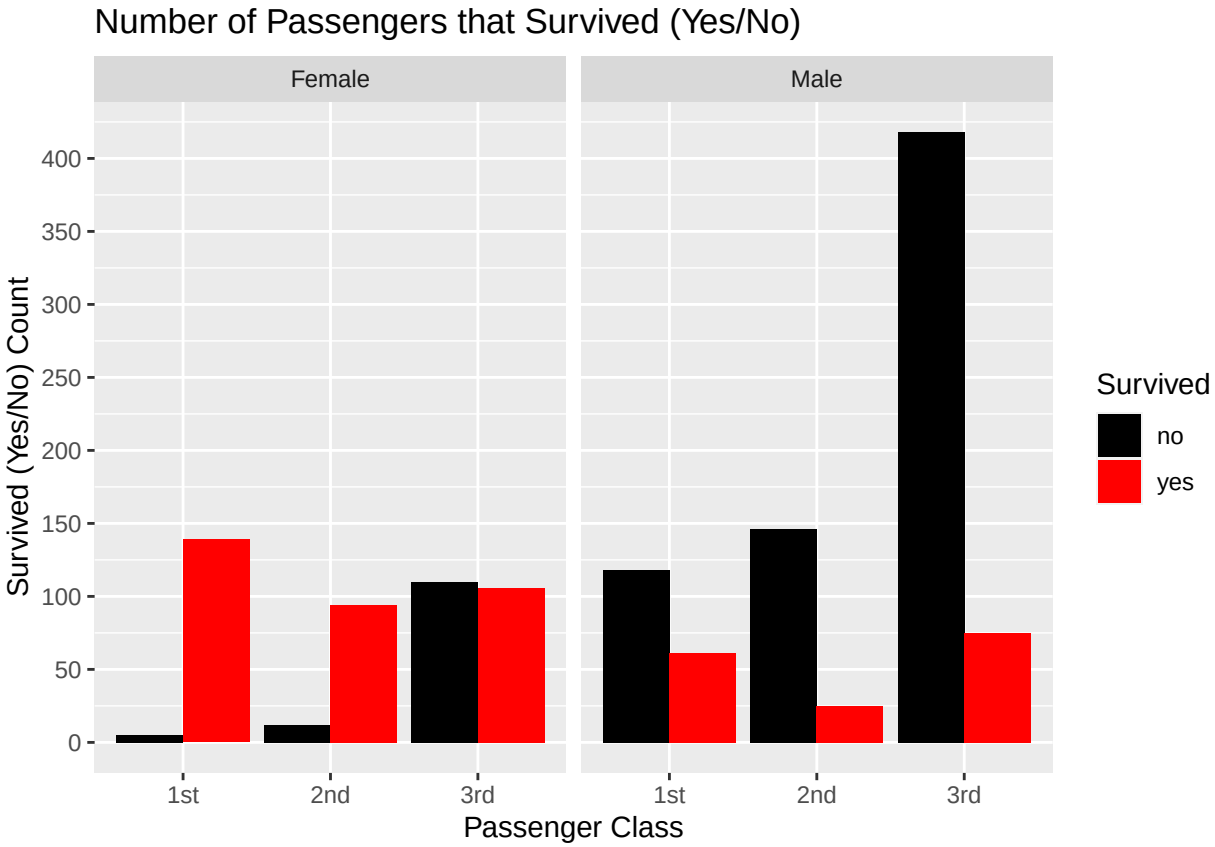
percentage of passengers who survived based on whether the individual was Female, Male, 1st Class, 2nd Class, or 3rd Class. I wanted to see what percentage rates for these variables would be higher than the others. Overall, we see that a passenger's survival rate was much higher if the individual was a female (72.7%). A passenger that was 1st class also had a higher survival rate (61.9%) than those of 2nd (43%) or 3rd (25.5%) class. Men, overall, had a very low survival rate (19.1%). These survival rates are not unexpected. It is common knowledge that women and children were prioritized in the rescue attempt after the ship began sinking. First class passengers were also prioritized over second and third class passengers. This numerical analysis helps the researcher see the disparities between the survival rates of passengers based on certain information about them. For further exploration, a researcher could even try to build a model to predict a passenger's chance of survival based on information about the passenger.

Visual Representation

```
# Creating the new labels for female and male on graph
labels <- c(female="Female", male="Male")

# Creating the histogram
Graphic1 <- ggplot(Titanic, aes(x=passengerClass, fill=survived)) +
  geom_bar(position='dodge') + facet_grid(.~sex, labeller=labeller(sex=labels)) +
  labs(title="Number of Passengers that Survived (Yes/No)",
  x="Passenger Class", y="Survived (Yes/No) Count") +
  scale_y_continuous(breaks=seq(0,600,by=50)) +
  scale_fill_manual(breaks=c("no","yes"),
  values=c("black","red")) + labs(fill="Survived")

Graphic1
```



For my visual representation, I wanted to show the different counts of passengers that survived and did not survive between the different passenger classes. I also separated the graphs by sex. Based on this graph, it is obvious that a first class passenger had a better chance at survival than a third class passenger. This is definitely obvious within the male passengers. Females, overall, had a higher count of survival no matter what passenger class they were in. Among the female passengers, almost all of the women in first or second class survived. For the third class, however, about as many women survived as did not survive. Males in the third class overwhelmingly did not survive. Overall, more women survived than did not. This is not the case with men, as more men died than survived. This graph helps the researcher see both how the variables of sex and class play a role in a passenger's chance at survival.

References

1. https://www.history.com/topics/early-20th-century-us/titanic
2. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html