SEIS 631 - Final Project "Titanic - A True Story"

Adnan Suri

5/10/2022

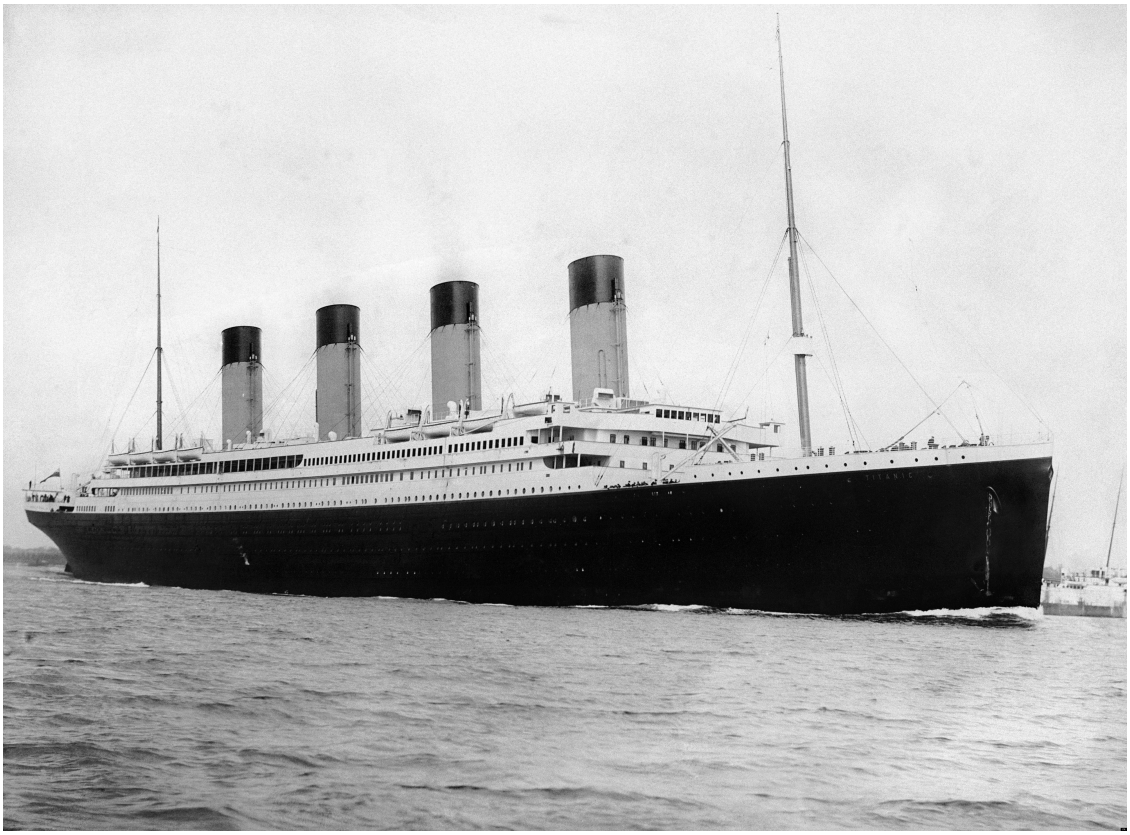


Figure 1: Titanic Ship, 15 April 1912

```
setwd("/Users/Home/Desktop/St Thomas/SEIS 631")
```

```
library(readr);
Titanic <- read_csv("titanic.csv")
```

```
## Rows: 891 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
##
```

```
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Introduction:

WHAT: Ever since I watched the movie “Titanic” in 1997, I have always wanted to learn more the statistical status of the Titanic ship. I hypothesize that more males died vs. females during the event. Also, I would like to know the survival age range and perform further statistical analysis to compare the data, like the overall survival rate.

WHY: It is unfortunate that the Titanic ship sank and people died. People created a dataset to understand statistics and mainly apply those statistics for people’s safety and a better world, which is very similar to the Covid dataset. Because of my interest in Titanic, I will use R language to learn more about titanic data.

HOW: My goal would be to visualize the data by using tidyverse for titanic data manipulation and visualization. I will be doing the data analysis by using the following:

MINIMAL: My goal would be to get more familiar with R, import titanic dataset into R, runs statistics using R functionality, especially ggplot, tidyvers, visuals where I can run the data with graphs, charts formats, and write the report in R Markdown and export in PDF.

AMBITIOUS: My next stretch goal would be to add more visuals and importing the dataset into Power BI (https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-r-scripts). Installing R script into BI desktop and create “Titanic Dashboard” to run different visuals with slicer option.

Topics From Class

Topic 1: Lecture 8: Tidy Data a.k.a Tidyverse approach to R.

```
## 1) Use R functionality to analyze the data using ggplot 2 (data visualization)
```

Topic 2: Chapter 2, Summarizing the data.

```
## 1) Examine the numerical data using scatterplot, histogram, standard deviation, boxplot, quaterlies a
## 2) Analyze data by considering categorical data;
      tables and barplots - class Hw-3
```

Topic 3: Distribution of random variables try with R

```
## 1) Normal distribution - plot(x, qnorm(x))
## 2) Geometric distribution - plot(x, gnorm(x))
```

```
## 3) Binomial distribution - plot(x, bnorm(x))
## 4) Negative binomial distribution - plot(x, pnorm(x))
```

Topic 4: R

```
## 1) Rstudio
## 2) RMarkdown
```

Topic 5: Github and Git

```
## 1) Open Github, fork and README.Md
## 2) Git control version
```

****

```
titanic <- read.table('titanic.csv',header = TRUE, sep = ",")
str(titanic)
```

```
## 'data.frame':    891 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       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Dataset

The titanic.csv file contains information on 891 passengers embarked from three major cities, Southampton (S), Cherbourg (C) and Queestown (Q). The CSV stands for (Comma-Separated-Values, a simple plain text format for storing spreadsheet data). There are 12 variables in this data set include name, passenger Id, gender, age, ticketed class, embarked, whether they survived, etc. I use the data set to explore some of the demographics of the passengers who were aboard the ship, and how their relationship to whether a passenger survived or not.

Libraries

Because I use Tidyvers visuals, I load R libraries (packages) that contain useful functions that will make our analyses quicker and more efficient.

library(tidyverse)

library(ggplot2)

library(dplyr)

library(ggmosaic)

Topic 1: Lecture 8: Tidy Data a.k.a Tidyverse approach to R.

```
## 1) Use R functionality to analyze the data using ggplot 2 (data visualization)
```

What is Tidyverse?

The tidyverse is a collection of R packages (dplyr, ggplot2, tibble, tidyr, forcats, stringr, readr) designed for data science. First, I install the complete tidyverse package with the following command:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

First finding - I would like to know how many passengers classes were in the ship, what cities those passengers came from and which city was the most representative among all the passengers.

The function str(titanic) returns a table, where the columns represent the variables (e.g. sex, age, etc) and the rows represent individuals or observations. 12 varaibles and 891 observations.

ggplot(data = NULL, mapping = aes(), ..., environment = parent.frame())

data as a dataframe from dataset to use for plot. aes = aesthetic mapping to use for plot. aes has many attributes for example, **alpha, colour, fill, group, line, type, size, weight**. I used **fill** property to colour codes. It is important to mention that **fill takes factor data type**. ggplot is smart enough count data for y-variable by default. Just need to input only x variable for ggplot. I converted variable to factor by:

```
titanic$Survived <- as.factor(titanic$Survived)
```

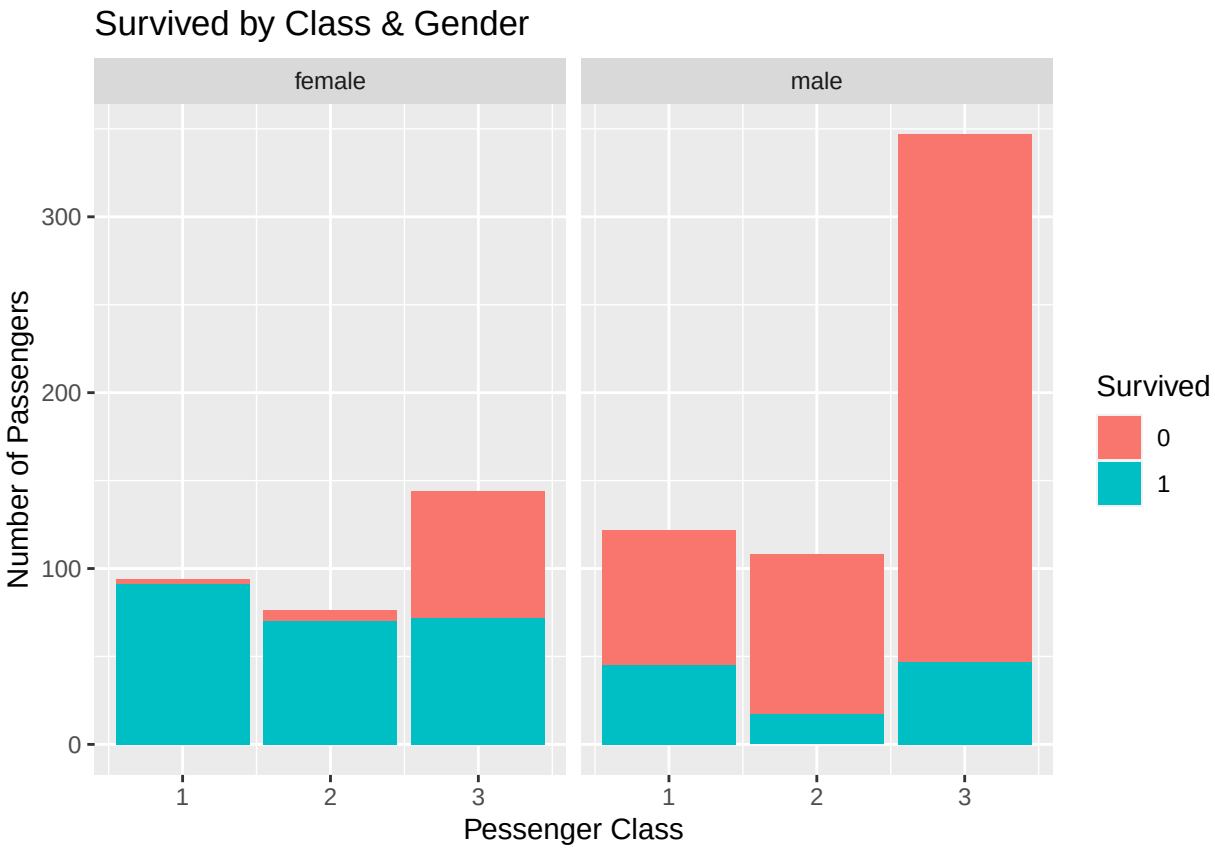
geom_bar() in ggplot - Use the bar graph to analyse passenger class (Pclass) by Embarked (where mounted from). Point noted that variable "Embarked" is a factor so no need to convert it. At the bottom, I used "caption" to write ny name as author.

```
ggplot(titanic, aes(x=Pclass, fill=Embarked)) + geom_bar() + labs( y= "Number of Passenger", x="S = Sou
```

2nd finding - I would like to see survival by passenger class (Pclass).

Introduced "title" and facet_wrap() wraps the panels into rows and columns which fits the number of panels in the layout. The largest death population was from third-class in male caterogry. The numbers of first- and second-class ticketed passengers are fairly similar.

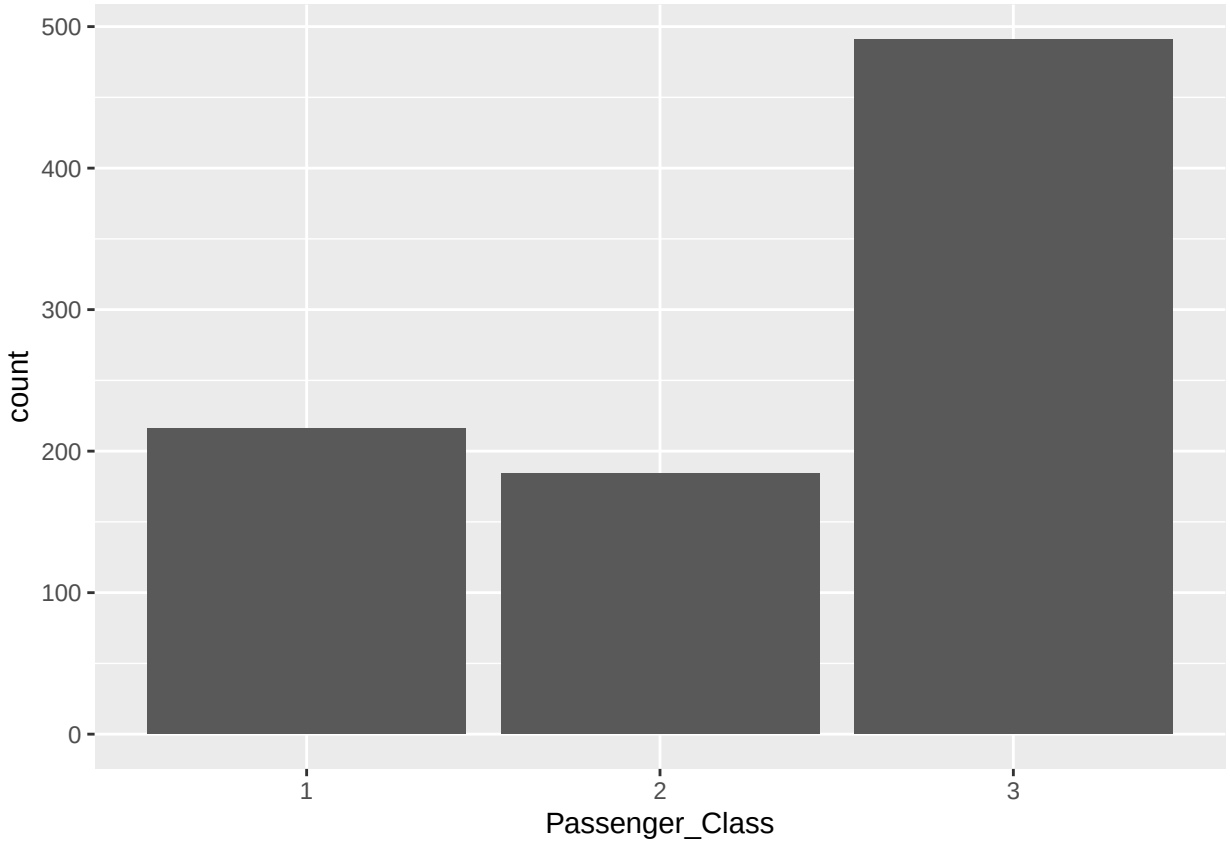
```
titanic$Survived <- as.factor(titanic$Survived)
ggplot(titanic, aes(x=Pclass,fill=Survived)) + geom_bar() + labs( y = "Number of Passengers", x = "Pesse
```



3rd finding - Categorizing Passengers.

Bar graph tell us the largest number of passengers were traveling in thirad-class tickets.

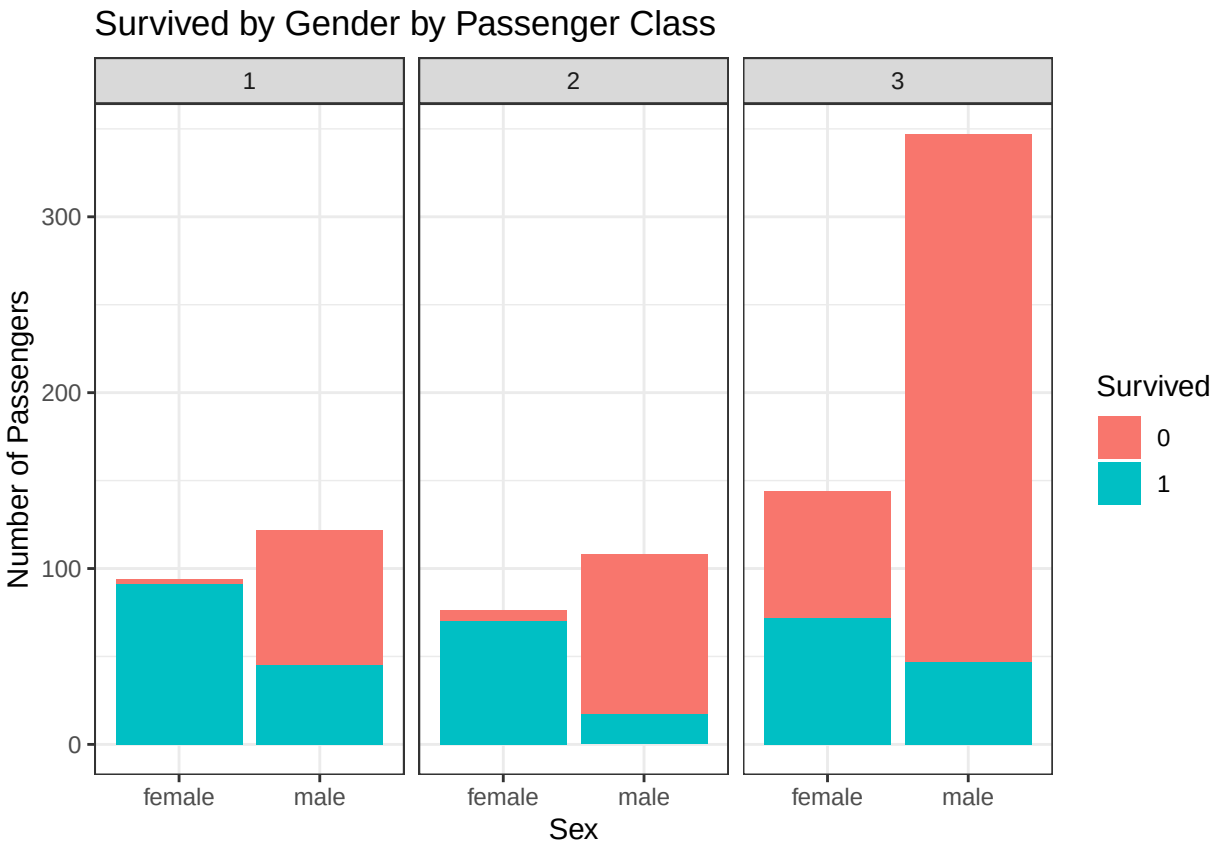
```
Passenger_Class = titanic$Pclass <- as.factor(titanic$Pclass)
ggplot(titanic) + geom_bar(aes(x = Passenger_Class))
```



4th finding - Survival per class by gender.

Bar graph tell us that male death ratio was higher in call three classes especially 3rd class. Females survived more than males in all three caterogries.

```
ggplot(titanic, aes(x=Sex, fill=Survived)) +
  theme_bw() +
  geom_bar() +
  labs( y="Number of Passengers", title="Survived by Gender by Passenger Class")+
  facet_wrap(~Pclass)
```



5th finding (Boxplot) - I want to see the survival rate by age by gender.

Another way to look at the data is to use a summary figure called a "box plot". A box plot depicts information about the median value (thick central line). The females in class of 1, 2 and 3 were at least 24, 20 and 18 years old. Their median age was 37, 28 and 22.

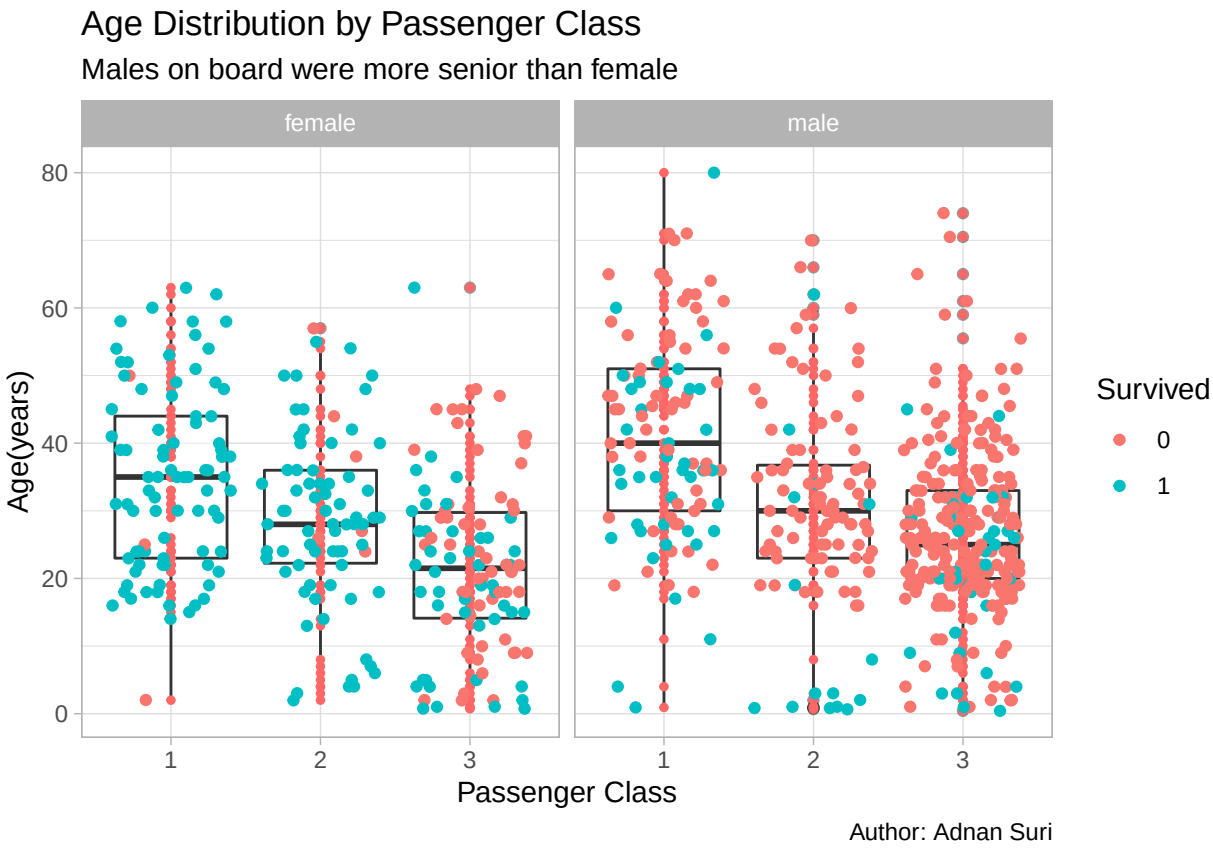
The males in class of 1, 2, and 3 were at least 30, 24 and 20 years old. Their median age was 42, 30 and 25 years old.

geom_point & geom(jitter) - to spread out points using in geom_point. geom_boxplot - distribution of data.

```
ggplot(data=titanic, aes(x=Pclass, y=Age)) + geom_boxplot(alpha=0.5) + geom_point(size =1, color ='#FF6
```

```
## Warning: Removed 177 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
## Removed 177 rows containing missing values (geom_point).
```



6th finding (Mosiac Plot): I want to use ggplot to see the the survival. Install ggmosaic library.

A more sophisticated version of a bar plot is called a "mosaic plot". A mosaic plot need to import the geom_mosaic function from a library called ggmosaic. The mosaic plot shows both that there were more men than women on the Titanic and more women survived. This is strong evidence for the first part of the phrase "Women and children first".

0 - Death

1 - Survived

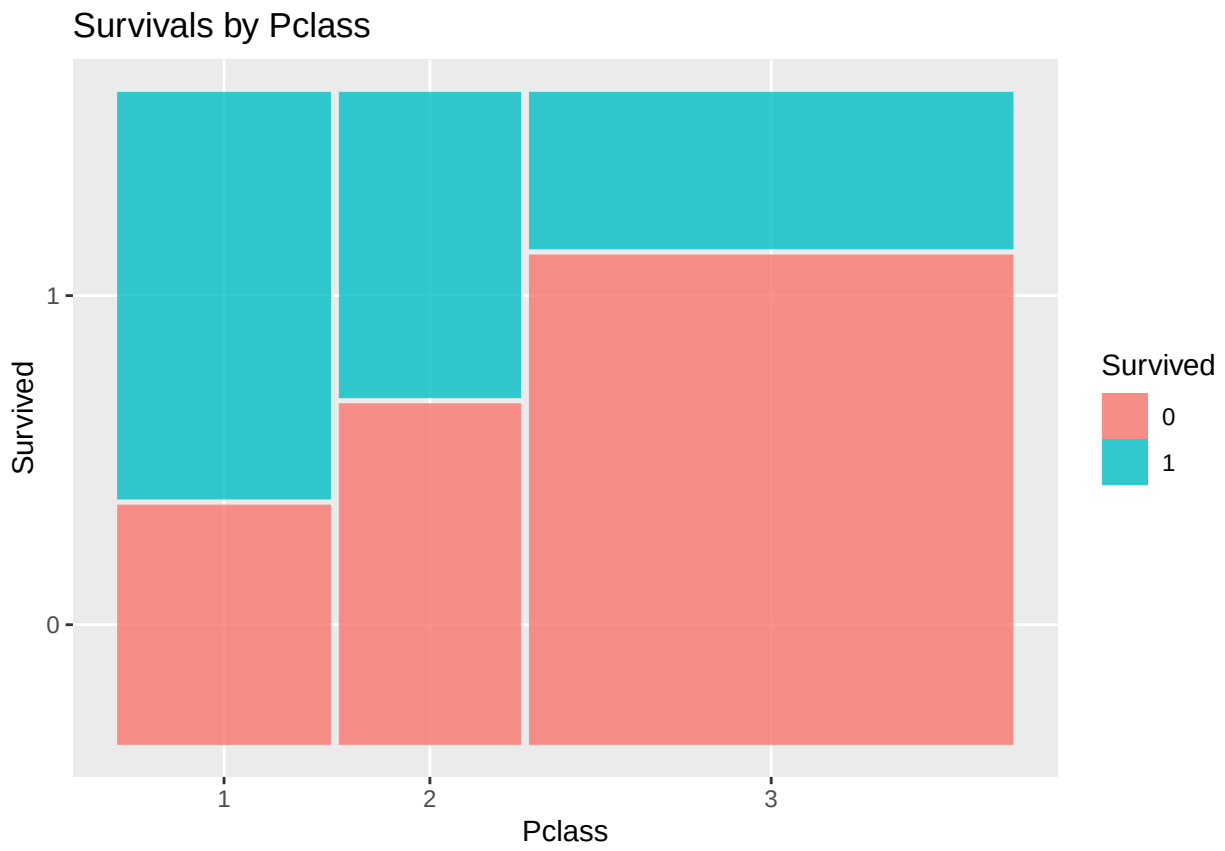
```
library(ggmosaic)
```

```
##
## Attaching package: 'ggmosaic'

## The following object is masked _by_ '.GlobalEnv':
##
##     titanic
```

```
ggplot(titanic) +
  geom_mosaic(aes(x=product(Pclass), fill=Survived)) +
  ggtitle("Survivals by Pclass")
```

```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## Please use `unite()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



Topic 2: Chapter 2, Summarizing the data Examining numerical data Considering categorical data

Variables to our data set, the dimensions, length and column names:

```
dim(titanic)
```

```
## [1] 891  12
```

```
length(dim(titanic))
```

```
## [1] 2
```

```
names(titanic)
```

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

Head & Tail I was curious to see the top six head and bottom of the dataset.

```
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
tail(titanic)
```

```
##     PassengerId Survived Pclass                                      Name    Sex
## 886         886        0      3     Rice, Mrs. William (Margaret Norton) female
## 887         887        0      2                     Montvila, Rev. Juozas   male
## 888         888        1      1              Graham, Miss. Margaret Edith female
## 889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female
## 890         890        1      1                     Behr, Mr. Karl Howell   male
## 891         891        0      3                       Dooley, Mr. Patrick   male
##     Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886  39     0     5     382652 29.125              Q
## 887  27     0     0     211536 13.000              S
## 888  19     0     0     112053 30.000   B42        S
## 889  NA     1     2 W./C. 6607 23.450              S
## 890  26     0     0     111369 30.000  C148        C
## 891  32     0     0     370376  7.750              Q
```

Summary Run summary to learn more about dataset.

```
summary.data.frame(titanic)
```

```
##   PassengerId    Survived Pclass      Name              Sex
##  Min.   :  1.0   0:549    1:216   Length:891         Length:891
##  1st Qu.:223.5   1:342    2:184   Class :character   Class :character
##  Median :446.0            3:491   Mode  :character   Mode  :character
##  Mean   :446.0
##  3rd Qu.:668.5
##  Max.   :891.0
##
##       Age            SibSp           Parch           Ticket
##  Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891
```

```
##  1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character
##  Median :28.00   Median :0.000   Median :0.0000   Mode  :character
##  Mean   :29.70   Mean   :0.523   Mean   :0.3816
##  3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :80.00   Max.   :8.000   Max.   :6.0000
##  NA's   :177
##       Fare           Cabin             Embarked
##  Min.   :  0.00   Length:891         Length:891
##  1st Qu.:  7.91   Class :character   Class :character
##  Median : 14.45   Mode  :character   Mode  :character
##  Mean   : 32.20
##  3rd Qu.: 31.00
##  Max.   :512.33
##
```

Table Survived vs Died

I want to visualize the overall survival rates. We can do this using tables or graphs.

0 - Died 1 - Survived

```
table(titanic$Survived)
```

```
##
##   0   1
## 549 342
```

Male vs Female

```
table(titanic$Sex)
```

```
##
## female   male
##    314    577
```

How many per class

```
table(titanic$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

Percentage of passengers/class

24.24% passengers survived - first Class

20.65% passengers survived - second Class

55.10% passengers survived - third Class

```
table(titanic$Pclass) ->pclass
prop.table(pclass)*100
```

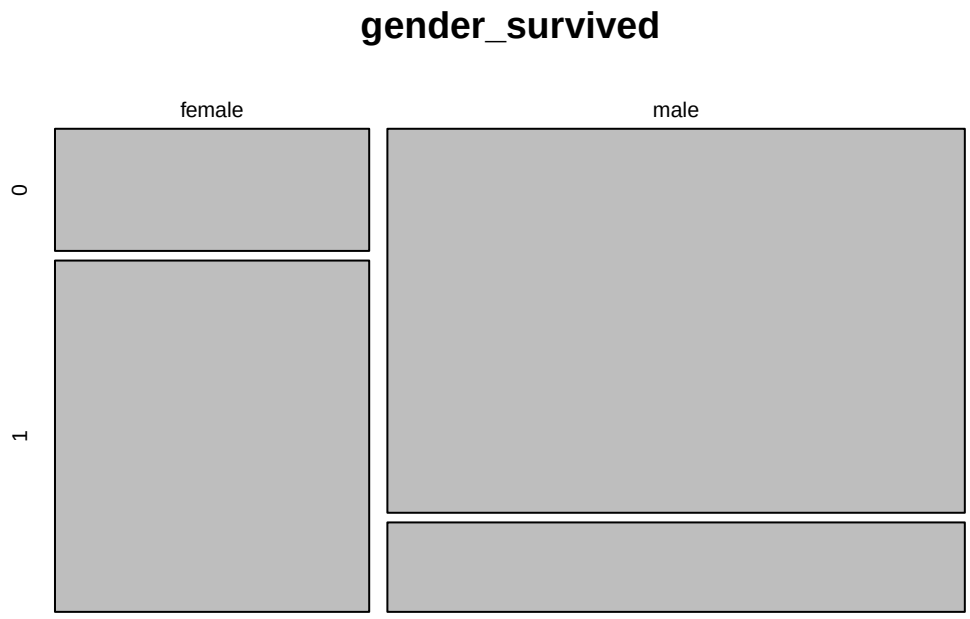
```
##
##        1        2        3
## 24.24242 20.65095 55.10662
```

```
round(pclass, digits=1) ->pclass
pclass
```

```
##
##   1   2   3
## 216 184 491
```

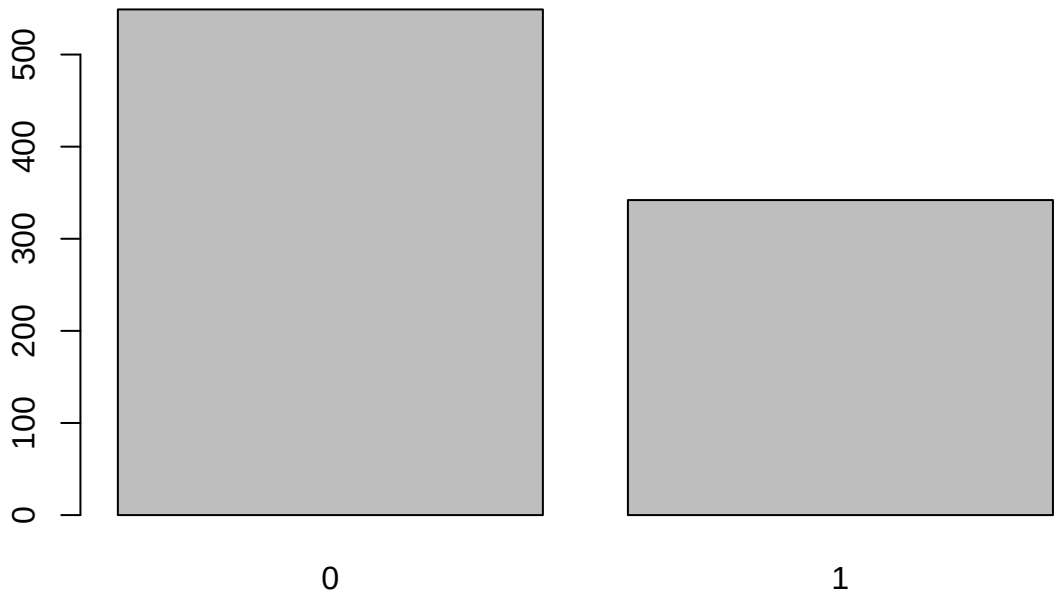
Mosaicplot

```
gender_survived <- table(titanic$Sex, titanic$Survived)
mosaicplot(gender_survived)
```

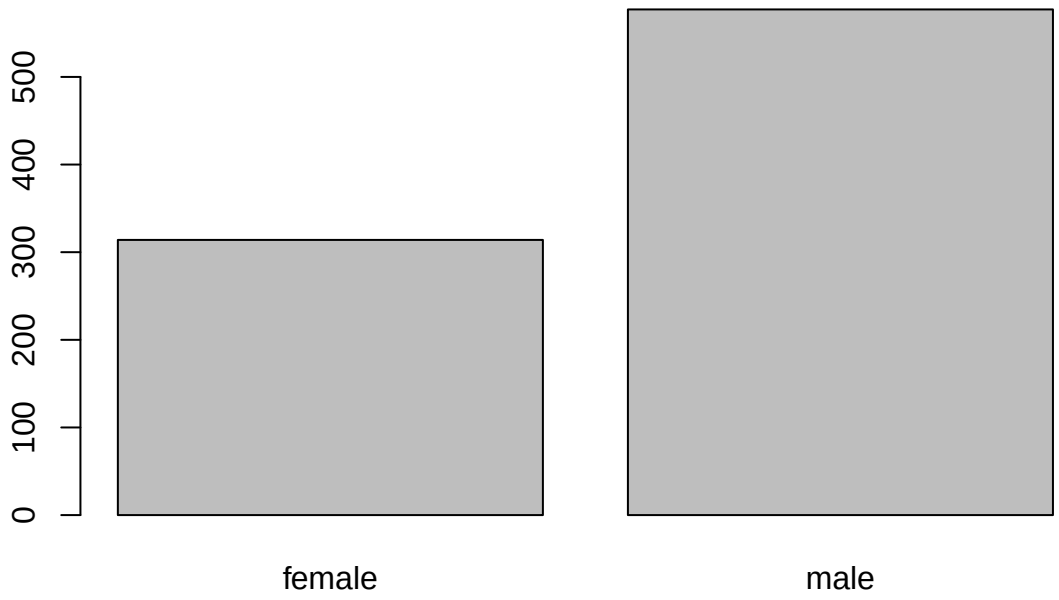


Bar Chart Another visual graph to look data for survival.

```
barplot(table(titanic$Survived))
```




```
barplot(table(titanic$Sex))
```



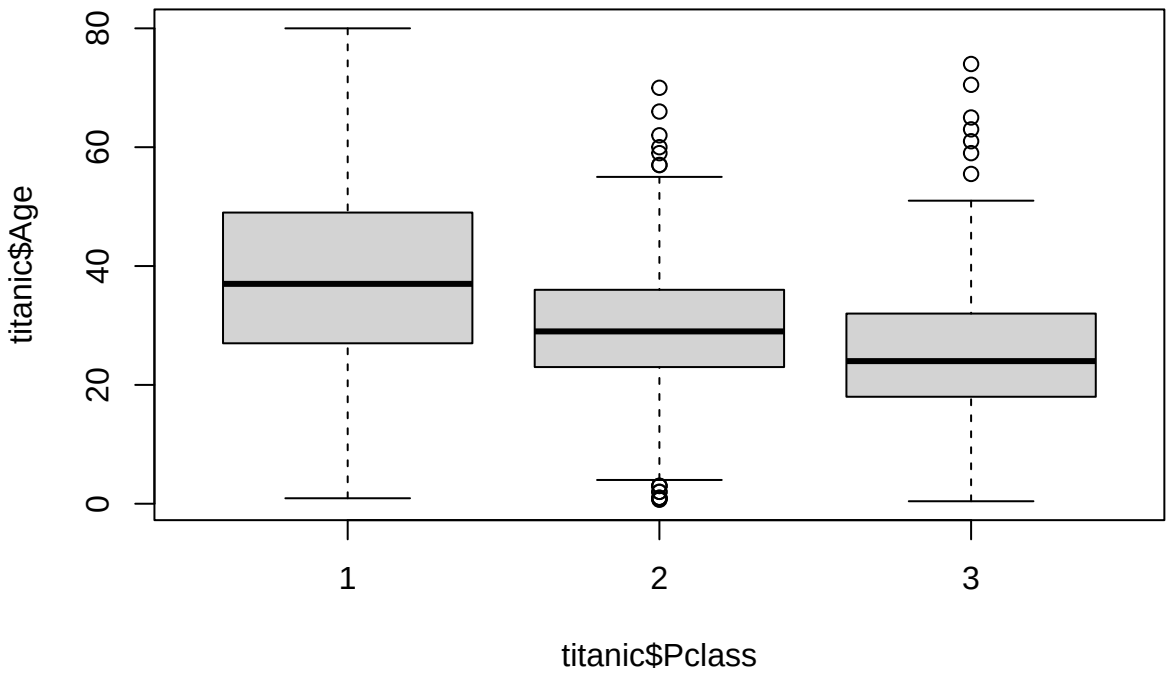
Boxplot for age of men and women in three classes (981 samples)

A distribution showing Age vs Pclass. The median ages in 1st, 2nd and 3rd classes are approximately to 40, 25, 23 respectively.

Majority passenger ages for 1st class is at the approximate range (29,50), 2nd class at (23,37), 3rd class at (18,31).

Survived passengers are younger in 3rd class than in 1st class. There are some outliers in 2nd and 3rd class.

```
boxplot(titanic$Age ~ titanic$Pclass)
```

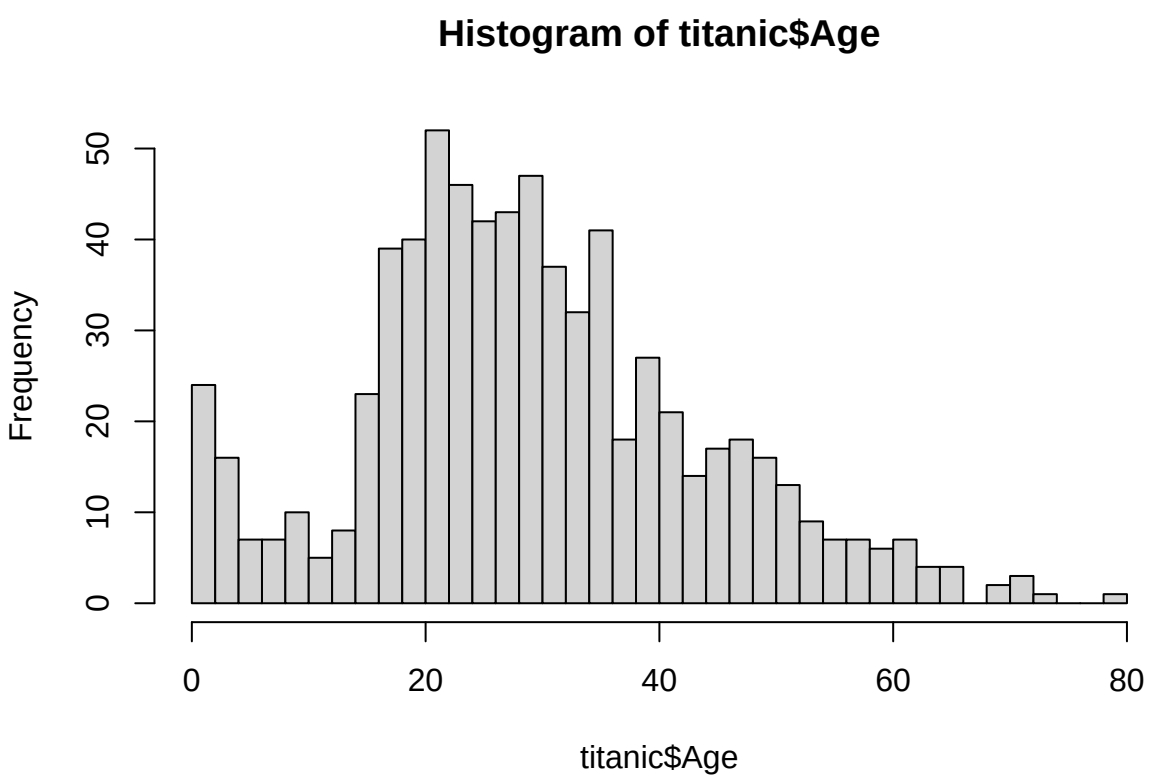


Histogram Passenger Age

Using histogram to visualize the distribution of passengers ages. A histogram is a comman way to visualize the disbribution of continous varaible.

Majority of passengers ages from 20 to 50 years old. There were some seniors ages above 70 - 80 but minimum frequency.

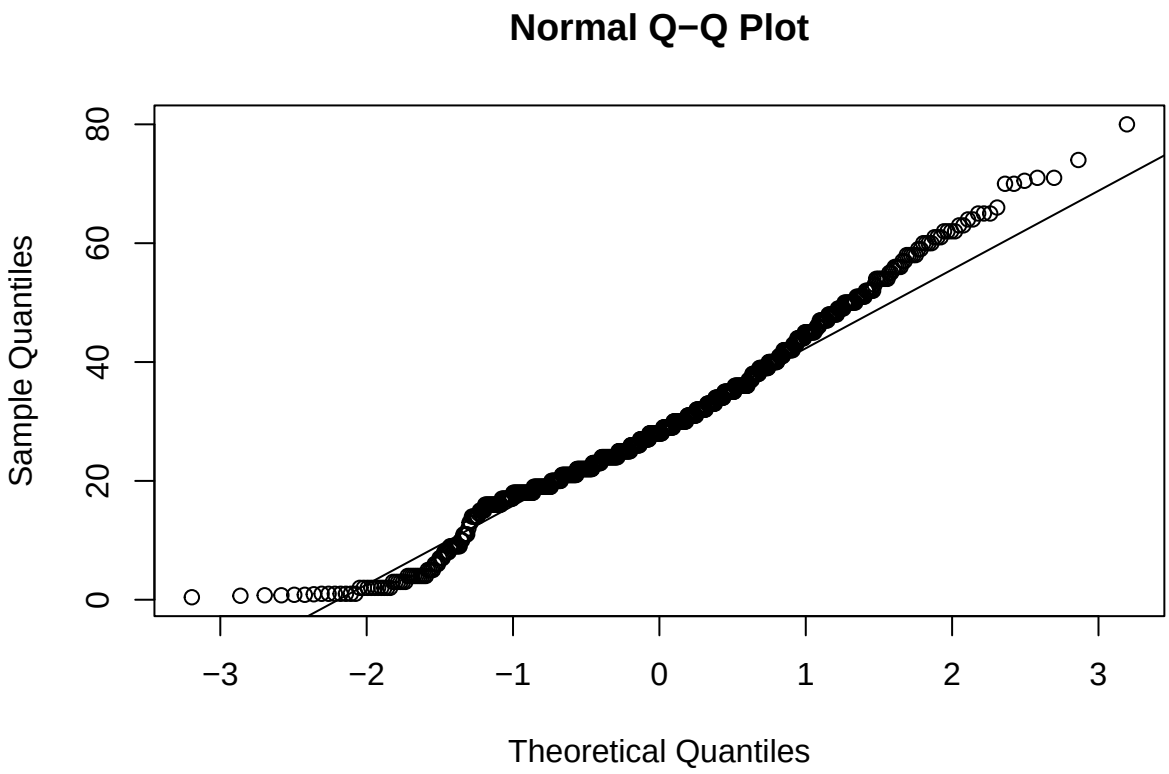
```
hist(titanic$Age, breaks = 50)
```



Topic 3: Chapter 4, Distribution of random variables Q plot

QQ plots can be made in R using a function called qqnorm(). Simply give the vector of data as input and it will draw a QQ plot. (qqline() will draw a line through that Q-Q plot to make the linear relationship easier to see.

```
qqnorm(titanic$Age)
qqline(titanic$Age)
```



Conclusion

My motivation for this interesting dataset project was to explore the data and learn more about the relationship between variables, survival rate, passenger class vs survived population. Also wanted to know how many passengers were on board and survived.

By conditioning on both sex and ticketing class, we gain even more insights into the data and our assesments of relationships between variables can change.

For female passengers, ticketed class appears to have a strong influence on the relationship between age and survival. We see that almost all of the female first-class passengers survived, and the relationship between age and survival is thus flat. The second class female passengers are fairly similar, though with a slight decrease in the probability of survival among the oldest passengers. It's not until we get to the third class passengers that we see a strong indication of the "children first" relationship playing out in terms of survival.

For the male passengers, the "children first" model seems to fit across classes, but note the generally lower probability of survival across ages when comparing first and third class passengers. "Women and children first!"

References:

1. **Titanic Picture**
2. **SEIS-631 HW-3**
3. **Dataset [https://github.com/elisabetta42/dataset_analysis/blob/master/titanic3.csv]**
4. **[https://ggplot2.tidyverse.org]**
5. **(https://github.com/Adam-1792/631-rtopics)**
6. **https://github.com/Adam-1792/Adnan_Suri---631-Final-Project/blob/main/.gitignore**
7. **R Documenation - Help**