Jessie Wong
January 6, 2023
Independent/Self-Learning Project

Titanic Dataset: Wrangling, Visualization, and Decision Tree

Examining Data

```
rm(list = ls())
library(datasets)
Titanic
head(Titanic)
str(Titanic)   table
is.table(Titanic)   Yes
typeof(Titanic)   double precision
```

Loading Packages

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(datasets, tidyr, party, pacman, magrittr, rio, tidyverse, ggplot2)
library(tidyverse)
install.packages("ggplot2")
library(ggplot2)
```

Wrangling Data

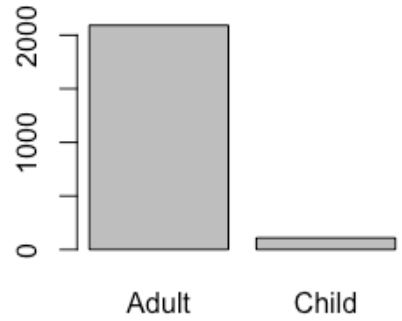
```
df <  Titanic %>%          Convert table to rows
  as_tibble() %>%              Convert to tibble with rows
  uncount(n) %>%
  mutate_all(as_factor) %>%
  mutate_all(fct_infreq) %>%   reordering bargraphs in decreasing manner
  print()

str(df)  converted to a tibble
dim(df)  2201 x 4 (4 2 2 2)
colors()
```

Visualizing Data

Histogram of Age

```
names(df)  "Class"  "Sex"  "Age"  "Survived"
df %>%
  select(Age) %>%
```

```
plot()
```



There are a lot more adults than children. Let's see the proportions of them.

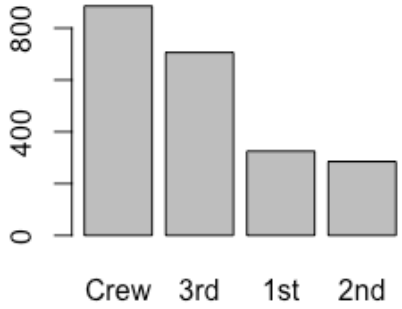
```
df1 <- prop.table(table(df$Age))
df1
```

In the Titanic, 95% are adults and 5% are children.

```
    Adult     Child
0.95047706 0.04952294
```

Histogram of Class

```
df %>%
  select(Class) %>%
  plot()
```



There are more crew members than the 3 classes.

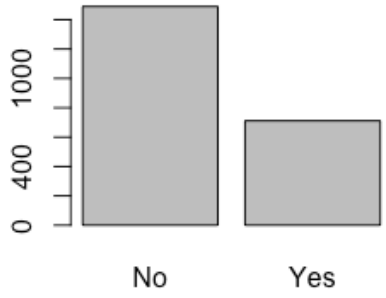
```
df2 <- prop.table(table(df$Class))
df2
```

Crew members take up 40% of all those onboard, while 1st class, 2nd class, and 3rd class take up 14.8%, 13%, and 32.1%, respectively.

```
     Crew       3rd       1st       2nd
0.4020900 0.3207633 0.1476602 0.1294866
```

Survival Histogram

```
df %>%
  select(Survived) %>%
  plot()
```



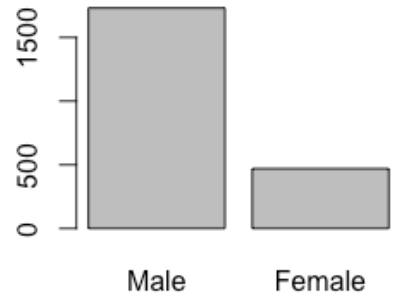

```
df3 <  prop.table(table(df$Survived))
df3
```

32% survived while 68% died. Therefore, the overall survival rate is less than a third.

```
       No       Yes
0.676965 0.323035
```

Histogram of Sex

```
df %>%
  select(Sex) %>%
  plot()
```




```
df4 <  prop.table(table(df$Sex))
df4
```

```
     Male    Female
0.7864607 0.2135393
```

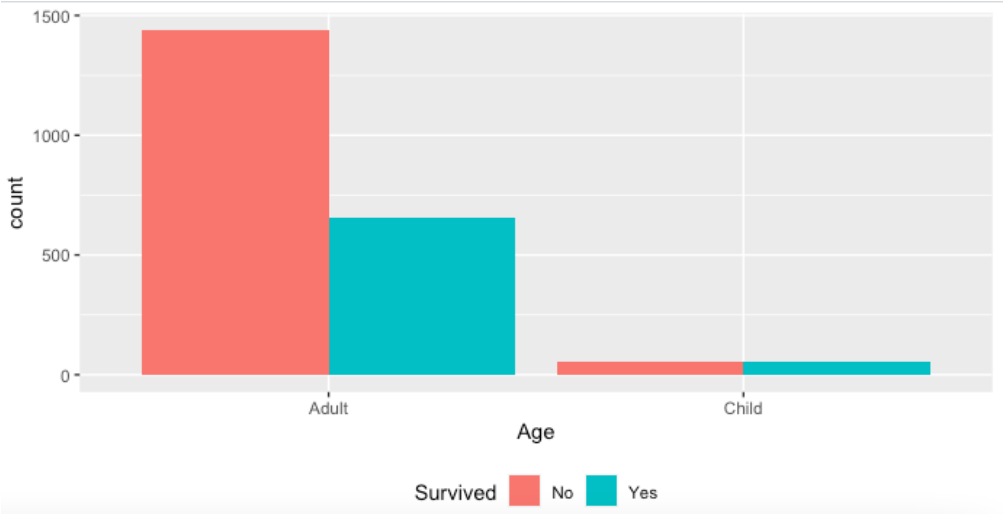
Graph of Adult and Children that Survived

```
table(df$Age)
```

```
Adult Child
 2092   109
```

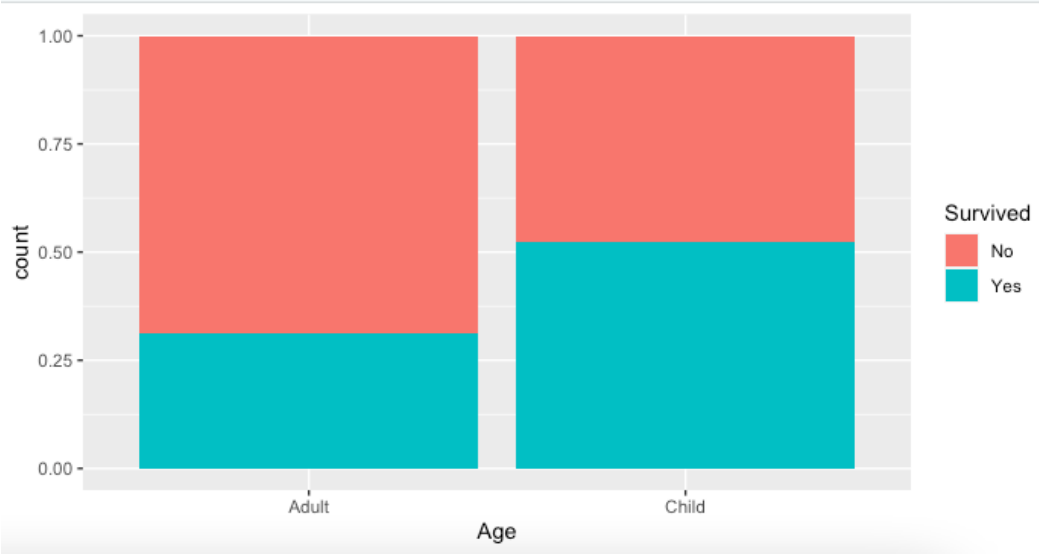
```
par(mfrow = c(1, 2))
df %>%
  ggplot(aes(Age, fill=Survived))+
  geom_bar(position = position_dodge())+
  theme(legend.position = "bottom")
```

This graph shows that over half of adults died, while a little over half of children survived.




```
df %>%
  ggplot(aes(Age, fill = Survived)) +
  geom_bar(position = "fill")
```

The graph below demonstrates that over 60% of adults died, while over half of children survived. This demonstrates that adults may have prioritized the safety of children before their own.



Make a Matrix of the Proportion of Adults and Children who Survived/Died

```
dfa <-  prop.table(table(df$Age,df$Survived))
dfa
```

```
              No       Yes
Adult 65.333939 29.713766
Child  2.362562  2.589732
```

These results show that a little over a half of children in the Titanic survived, while over 60% of adults died.

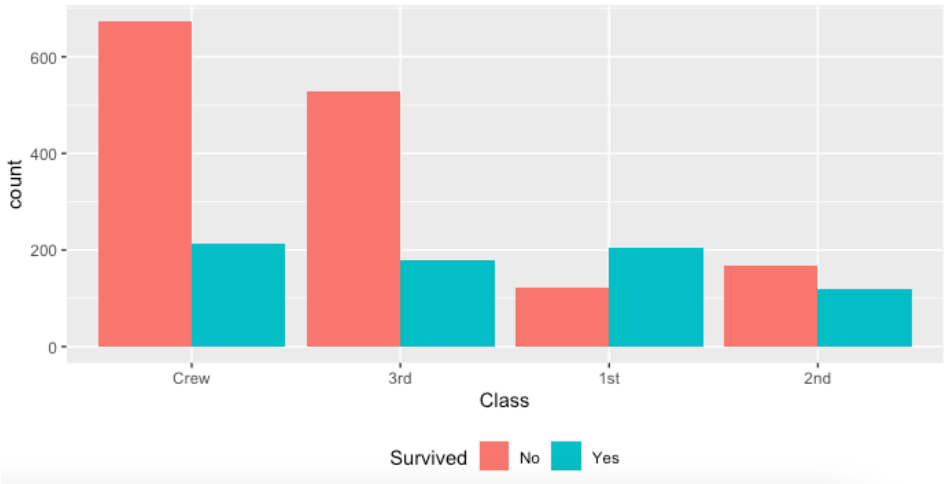
Plot Class with those who Survived/Died

Names: "3rd" "1st" "2nd" "Crew"

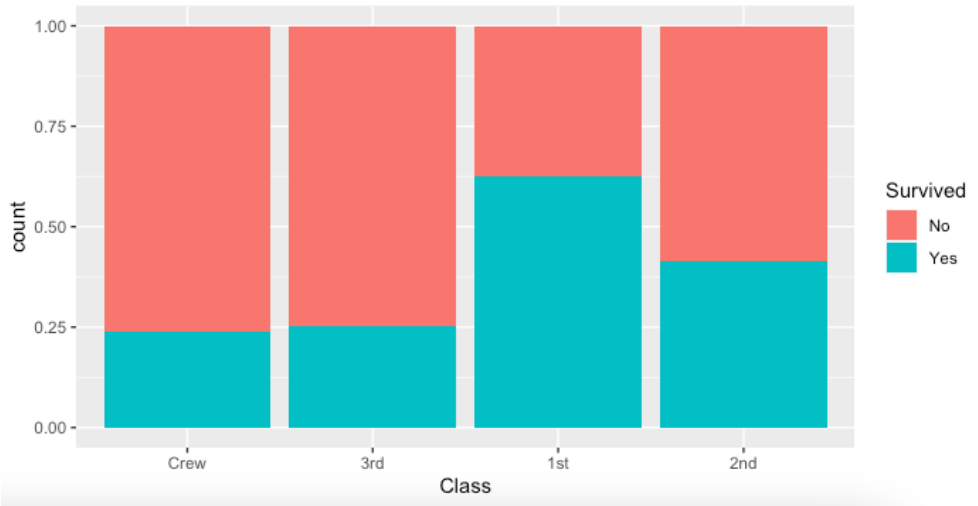
```
table(df$Class)
```

```
Crew  3rd  1st  2nd
 885  706  325  285
```

```
df %>%
  ggplot(aes(Class, fill=Survived))+
  geom_bar(position = position_dodge())+
  theme(legend.position = "bottom")
```




```
df %>%
  ggplot(aes(Class, fill = Survived)) +
  geom_bar(position = "fill")
```



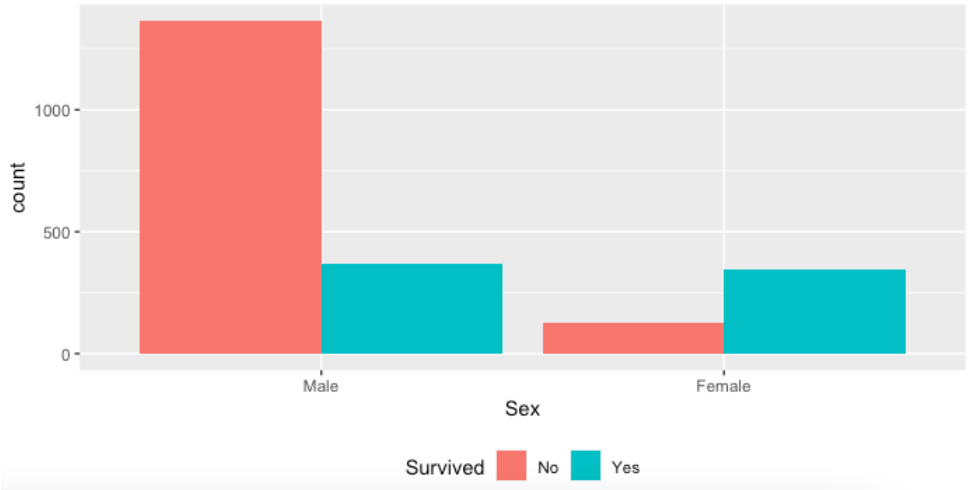
The above plots show that those in crew and 3rd class approximately died in similar proportions, while those in 2nd and 1st class had a higher survival rate. Additionally, over half of those in 1st class survived while about 40% of those in 2nd class survived. Overall, this demonstrates that the safety of the wealthy are prioritized.

Male and Female Survival Plot

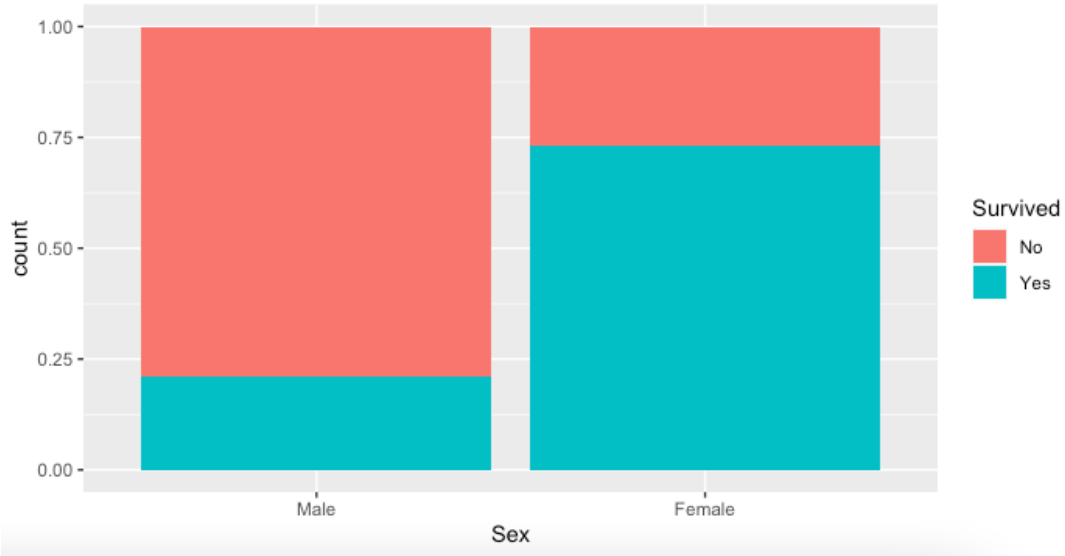
```
table(df$Sex)
```

```
  Male Female
  1731    470
```

```
df %>%
  ggplot(aes(Sex, fill=Survived))+
  geom_bar(position = position_dodge())+
  theme(legend.position = "bottom")
```




```
df %>%
  ggplot(aes(Sex, fill = Survived)) +
  geom_bar(position = "fill")
```



The above graphs show that more females survived than men, which tells us that women might've been prioritized over males.

Let's make a matrix.

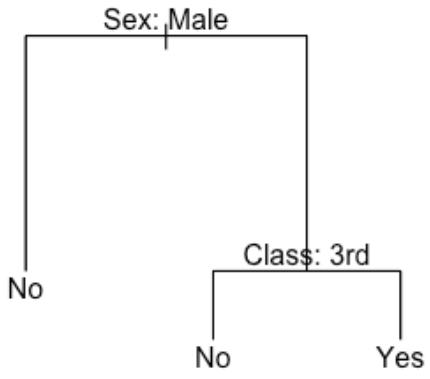
```
dfg <  prop.table(table(df$Sex,df$Survived))
dfg
```

```
                No        Yes
Male   0.61971831 0.16674239
Female 0.05724671 0.15629259
```

Decision Tree

Make a Decision Tree to Predict Who Won't Survive, and Find its Prediction Accuracy.

```
dfd=as.data.frame(df)
install.packages("tree")
library(tree)
Survived <-  (df$Survived)

set.seed(1)
train=sample(nrow(df),nrow(df)*0.8)
tree.model=tree(Survived~.,dfd,subset=train)
dfd.test=dfd[ train,]
Survived.test=Survived[ train]
cv.model=cv.tree(tree.model,K=10,FUN=prune.misclass)
cv.model

prune.model=prune.tree(tree.model,best=5)
plot(prune.model)
text(prune.model,pretty=0)
```



This decision tree tells us that the 2 most important factors in predicting survival rates is Sex and Class. To be more specific, Sex:Male is the most important predictor in this dataset. If the sex is male, then they won't survive. If the sex isn't male but instead female,

and they're in 3rd class, then they won't survive. Otherwise, if they're female and not in 3rd class, then they'll survive.
Overall, for better chances of survival in the Titanic, you'll have to be female and either in Crew, 1st class, or 2nd class.

Model Evaluation

```
prunetree.pred=predict(prune.model,(dfd.test),type="class")
table(prunetree.pred,as.factor(Survived.test))
```

```
prunetree.pred  No Yes
           No  289  82
           Yes   5  65
```

Recall: 18.36%. 18.36% is a small proportion, so the model isn't able to find all relevant cases.
Precision: 44.22%. This means that 44.22% of positive identifications was actually correct.

```
mean(prunetree.pred==Survived.test)
```

```
[1] 0.8027211
```

We can conclude that about 80.27% time, we can predict survival correctly.