R-Analysis

Tom Blain, Daniel Gardner, Xin Guan, Xinyu Li

05/10/2022

```
#if(!require("readr")) install.packages("readr")
#if(!require("ggplot2")) install.packages("ggplot2")
#if(!require("gridExtra")) install.packages("gridExtra")
#if(!require("dplyr")) install.packages("dplyr")
#if(!require("ggplots")) install.packages("ggplots")
#if(!require("plotrix")) install.packages("plotrix")
library("fs") # for cross-platform directories (path_wd)
library("readr") #For read_csv
library("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
```

```
library("knitr") # For kable
library("ggplot2") # For plots
library("gridExtra")
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library("gplots")
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library("plotrix")# For general stacked histogram
```

```
##
## Attaching package: 'plotrix'
```

```
## The following object is masked from 'package:gplots':
```

```
##
##     plotCI
```

```
data <- read_csv(path_wd("01-Data.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.
```

```
data<-as.data.frame(data)
```

Analysing the Data

Let's start with looking at first few rows of data.

```
head(data)
```

```
##   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  <NA>        S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250  <NA>        S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500  <NA>        S
## 6           330877  8.4583  <NA>        Q
```

Here is a brief summary of the data set.

```
summary(data)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
```

```
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     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
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

This data set consists of our binary survival variable we are interested, as well as 9 other co-variates that may influence survival. We can start by encoding our categorical variables as factors, with the ticket class variable being ordered in this case.

```
data$Sex <- as.factor(data$Sex)
data$Survived <- as.factor(data$Survived)
data$Pclass <- as.ordered(data$Pclass)
```

Then we can check the incompleteness of the data,

```
sapply(data, function(x) {sum(is.na(x))})
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

We can see that for the most part this data set is complete. Importantly, key variables like Survived, Pclass and Sex have all been recorded, besides the 891 NA observations we are using for the test data. However there are quite large gaps when it comes to Age (Missing 177) and Cabin number (Missing 687).

We can deal with missing values in each column by fill a suitable substitution such as mode, mean or median.

```
data$Embarked[is.na(data$Embarked)]<-mode(data$Embarked)
#Use mode for the missing "Embarked" values.
data$Age[is.na(data$Age)]<-mean(data$Age,na.rm = T)
#Use average of the existing age values for the missing "Age" values.
```

Manipulating the data

We can also try to deconstruct some of the variables to infer more information. For example, the passenger name is very difficult to use in any kind of analysis, so we can try and turn it into a factor variable based off the title of each person.

```
data$Title <- sapply(data$Name, function(x) {strsplit(x, split='[,.]')[[1]][2]})
data$Title <- sub(' ', '', data$Title) #removing spaces before title
kable(table(data$Sex, data$Title))
```

	Capt	Col	Don	Dr	Jonkheer	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Rev	Sir
female	0	0	0	1	0	1	0	0	182	2	1	0	125	1	0	0
male	1	2	1	6	1	0	2	40	0	0	0	517	0	0	6	1

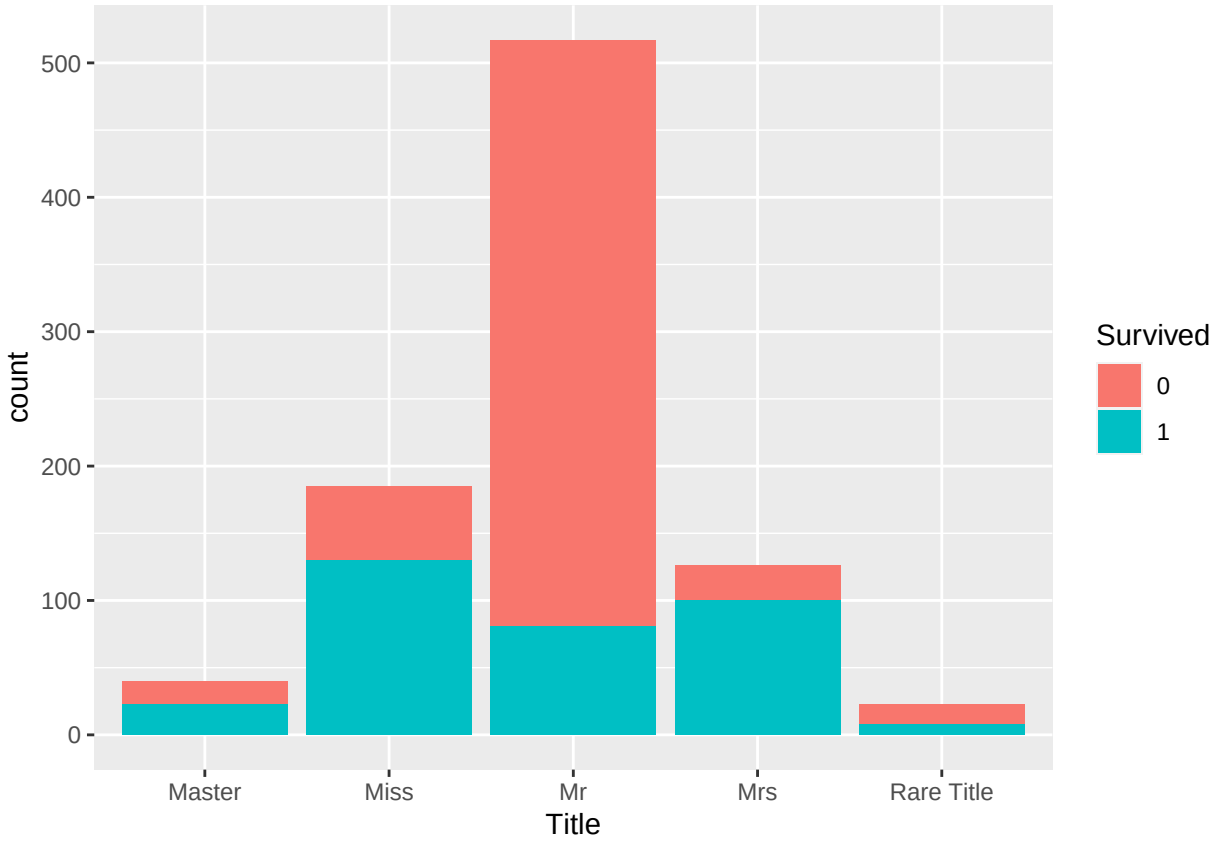
This is better than having over 1700 different names, although we can still reduce the number of dimensions of this new title variable. One way of doing this is grouping similar titles together, such as 'Ms','Miss' and 'Mlle'(Mademoiselle), all referring to young or unmarried women. Similarly we can grouo 'Mrs' and 'Mme'(Madame) as these refer to married women. Anything beside these two categories along with Mr and Master we can group in one 'Rare titles' category as miscellaneous titles.

```
data$Title[data$Title %in% c("Mlle", "Ms")] <- "Miss"
data$Title[data$Title== "Mme"] <- "Mrs"
data$Title[!(data$Title %in% c('Master', 'Miss', 'Mr', 'Mrs'))] <- "Rare Title"
data$Title <- as.factor(data$Title)
kable(table(data$Sex, data$Title))
```

	Master	Miss	Mr	Mrs	Rare Title
female	0	185	0	126	3
male	40	0	517	0	20

Now we have a much more clear table we can plot this to get an idea of how Title affects survival

```
ggplot(data[!is.na(data$Survived),], aes(x = Title, fill = Survived)) +
  geom_bar(stat='count', position='stack') +
  labs(x = 'Title') +theme_grey()
```



Create a variable "Familysize" which is the sum of variables "SibSp" and "Parch". Analysing family size seems to make more sense than analyse siblings and parents separately.

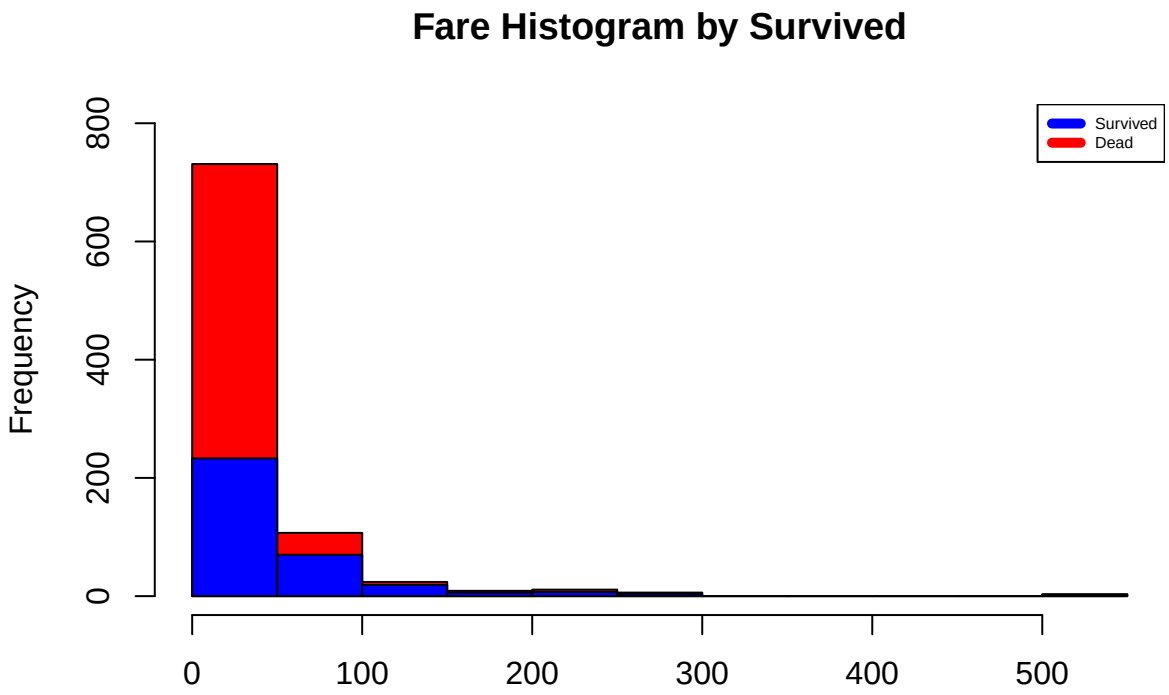
```
data$Familysize=data$SibSp+data$Parch
```

Visualising the data

We can then start visualizing the amount of deaths and who died due to the factor variables via bar charts

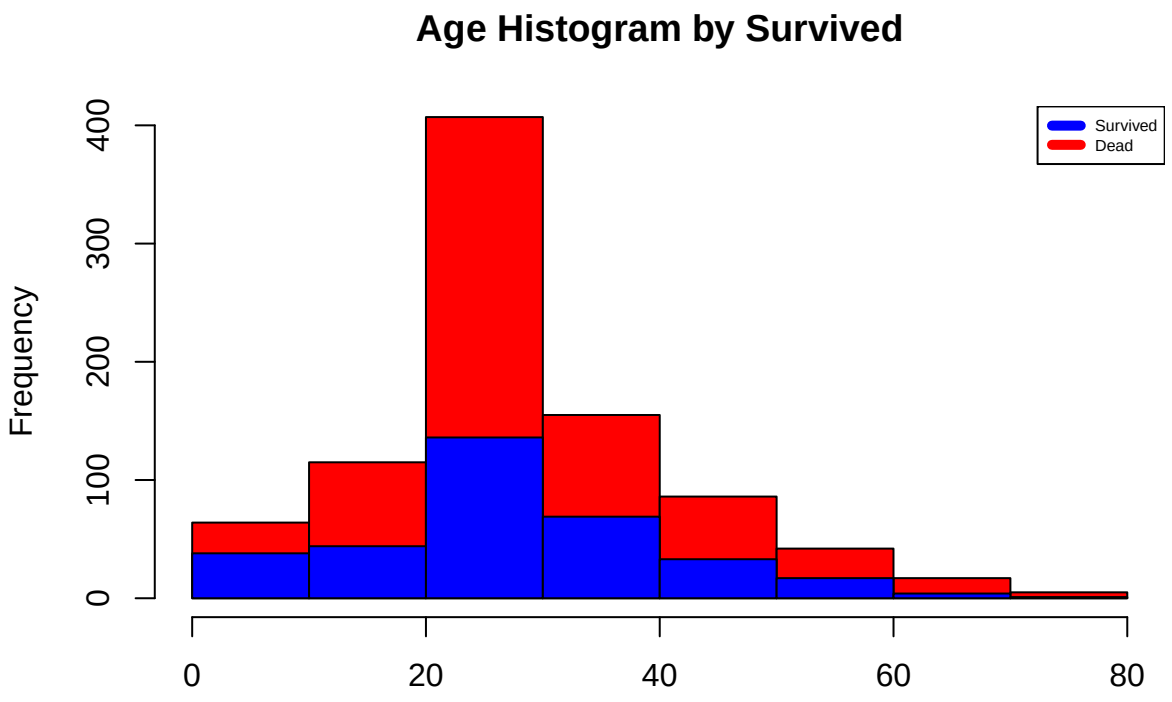
Let's firstly look at the most basic stacked histogram: We see that the survival rate of female is a lot higher than male, this is because that the men on board the ocean liner gave women and children priority access to the lifeboats.

```
histStack(x=data$Fare,z=factor(data$Survived),col=c("red","blue"),main = "Fare Histogram by Survived",yl
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5) #Add legend to indicat
```

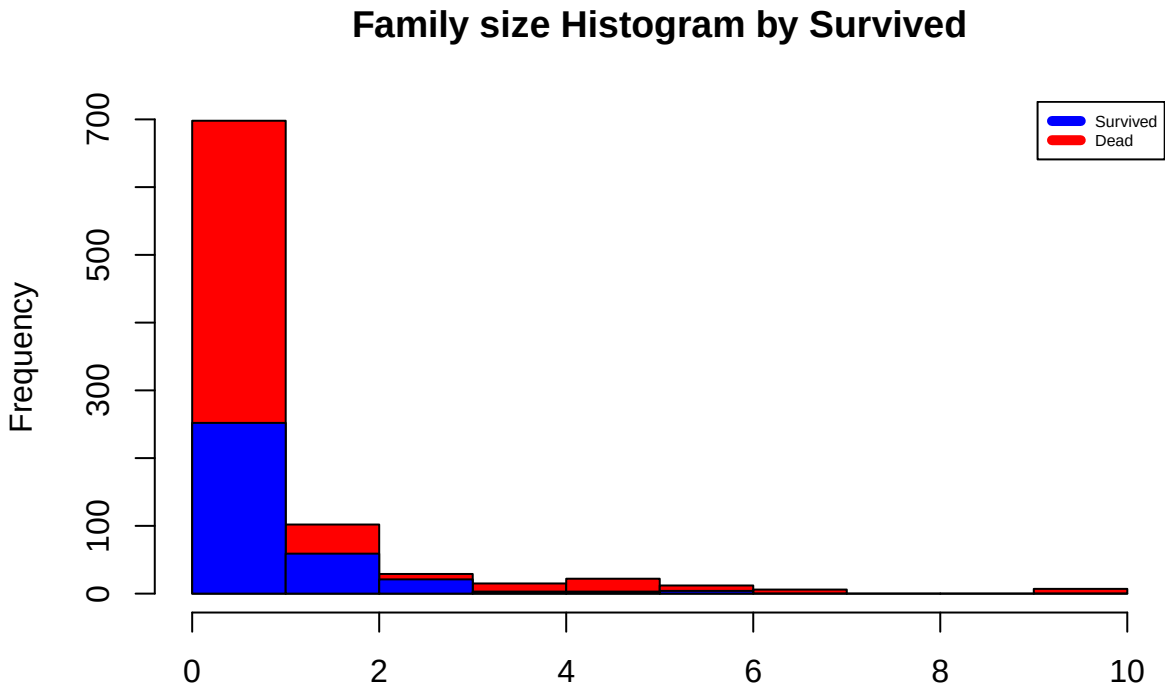


People between 20-40 years old has significantly lower survival rate.

```
histStack(x=data$Age,z=factor(data$Survived),col=c("red","blue"),main = "Age Histogram by Survived",ylir
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5)
```

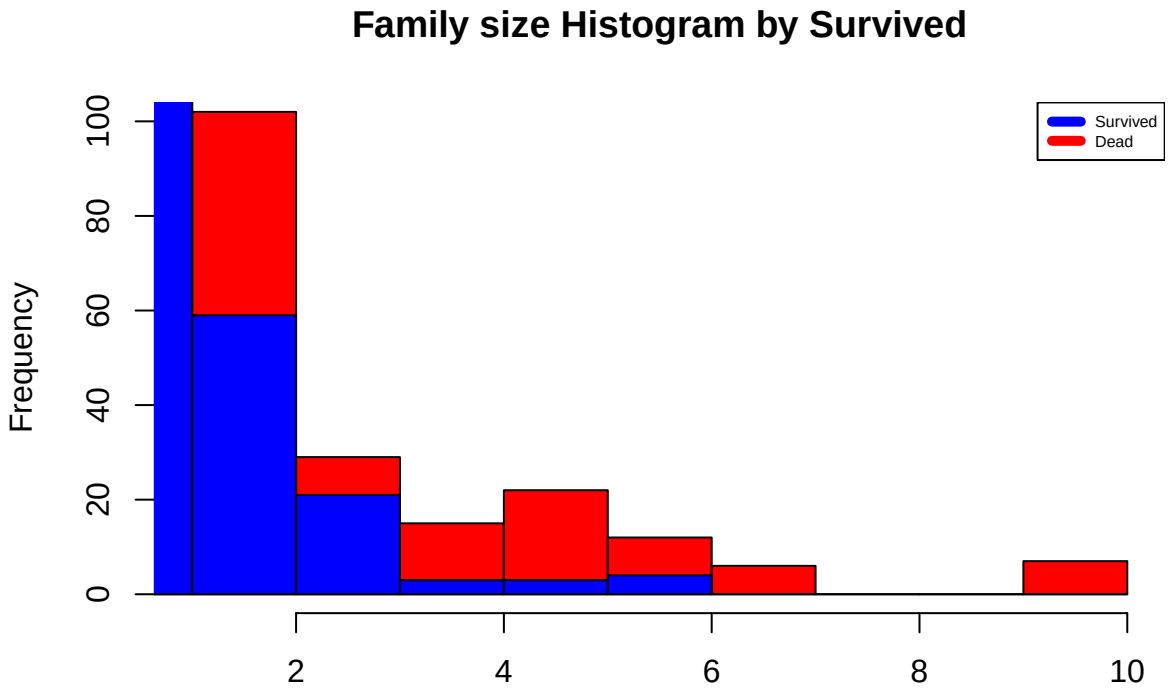



```
histStack(x=data$Familysize,z=factor(data$Survived),col=c("red","blue"),main = "Family size Histogram by
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5)
```



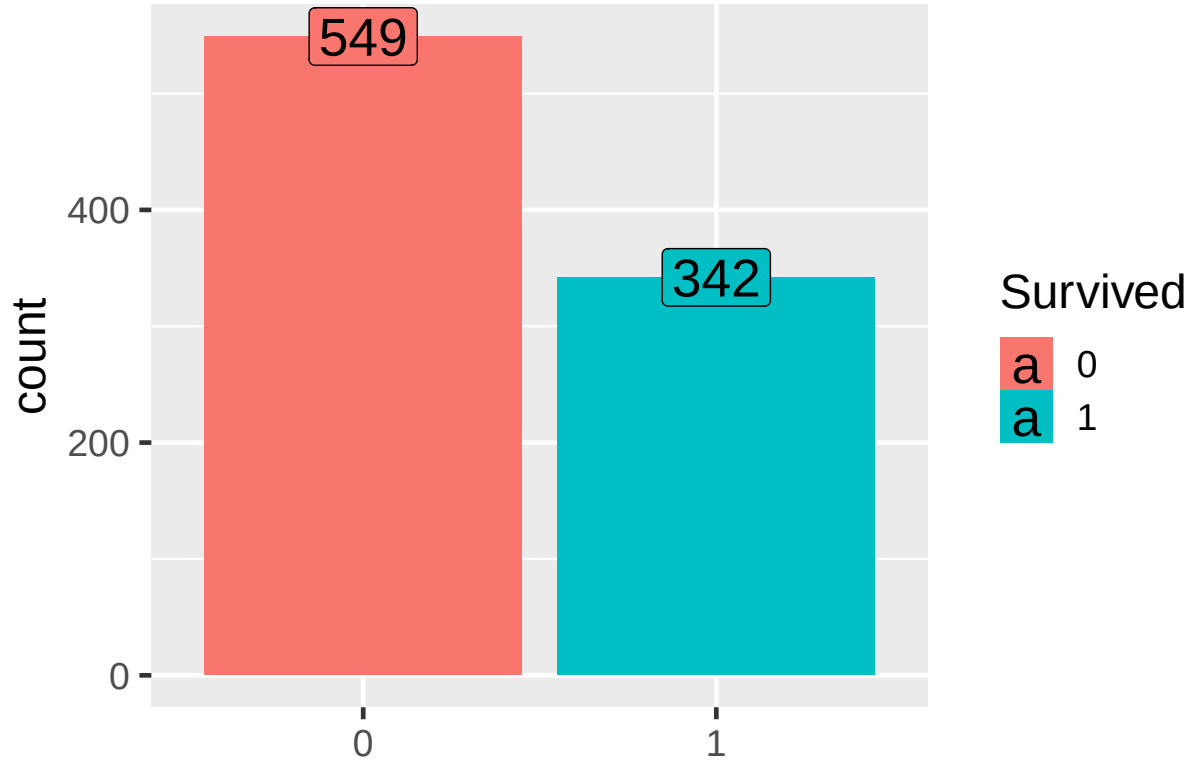
If we zoom in a little bit, we can see that the family size of 2 or 3 has significantly higher survival rate than family of other sizes.

```
histStack(x=data$Familysize,z=factor(data$Survived),col=c("red","blue"),main = "Family size Histogram by
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5)
```



There are other options such as ggplot2, which plots prettier graphs, and contains more graphical features.

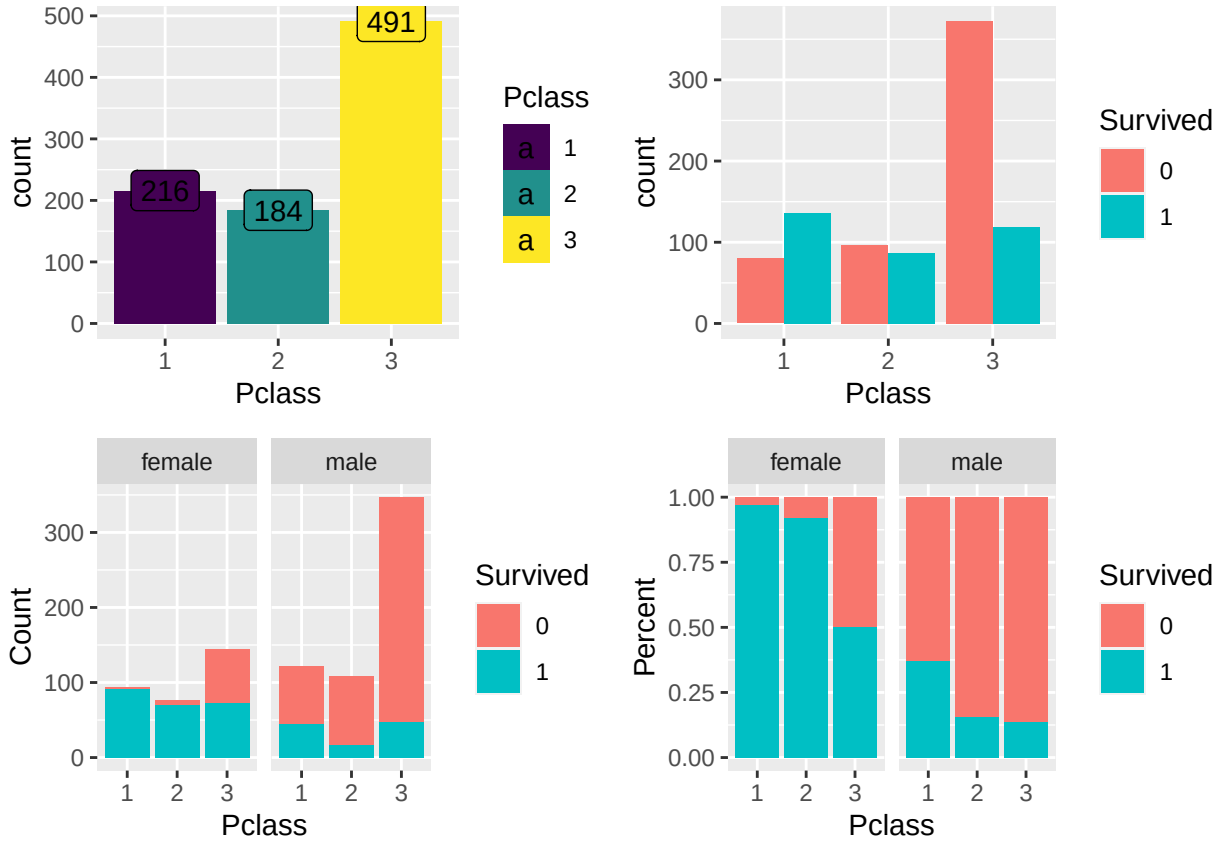
```
ggplot(data[!is.na(data$Survived),], aes(x = Survived, fill = Survived)) +
  geom_bar(stat='count') +
  labs(x = 'How many people died and survived on the Titanic?') +
        geom_label(stat='count',aes(label=..count..), size=7) +
        theme_grey(base_size = 18)
```



How many people died and survived on the Titanic?

And we can do further analysis on how class affects your chances of survival

```
p3 <- ggplot(data, aes(x = Pclass, fill = Pclass)) +
  geom_bar(stat='count', position='dodge') +
  labs(x = 'Pclass') + geom_label(stat='count', aes(label=..count..)) +
   theme(legend.position="none") + theme_grey()
p4 <- ggplot(data[!is.na(data$Survived),], aes(x = Pclass, fill = Survived)) +
  geom_bar(stat='count', position='dodge') + labs(x = 'Pclass') +
        theme(legend.position="none") + theme_grey()
p5 <- ggplot(data[!is.na(data$Survived),], aes(x = Pclass, fill = Survived)) +
  geom_bar(stat='count', position='stack') +
  labs(x = 'Pclass', y= "Count") + facet_grid(.~Sex) +
        theme(legend.position="none") + theme_grey()
p6 <- ggplot(data[!is.na(data$Survived),], aes(x = Pclass, fill = Survived)) +
  geom_bar(stat='count', position='fill') +
  labs(x = 'Pclass', y= "Percent") + facet_grid(.~Sex) +
        theme(legend.position="none") + theme_grey()
grid.arrange(p3, p4, p5, p6, ncol=2)
```




```
#if(!require("readr")) install.packages("readr")
#if(!require("ggplot2")) install.packages("ggplot2")
#if(!require("gridExtra")) install.packages("gridExtra")
#if(!require("dplyr")) install.packages("dplyr")
#if(!require("ggplots")) install.packages("ggplots")
#if(!require("plotrix")) install.packages("plotrix")
library("fs") # for cross-platform directories (path_wd)
library("readr") #For read_csv
library("dplyr")
library("knitr") # For kable
library("ggplot2") # For plots
library("gridExtra")
library("gplots")
library("plotrix")# For general stacked histogram
```

```
data <- read_csv(path_wd("01-Data.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.
```

```
data<-as.data.frame(data)
```

Analysing the Data

Let's start with looking at first few rows of data.

```
head(data)
```

```
##   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  <NA>        S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250  <NA>        S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500  <NA>        S
## 6           330877  8.4583  <NA>        Q
```

Here is a brief summary of the data set.

```
summary(data)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     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
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
```

```
##                     Max.   :512.33
##
```

This data set consists of our binary survival variable we are interested, as well as 9 other co-variates that may influence survival. We can start by encoding our categorical variables as factors, with the ticket class variable being ordered in this case.

```
data$Sex <- as.factor(data$Sex)
data$Survived <- as.factor(data$Survived)
data$Pclass <- as.ordered(data$Pclass)
```

Then we can check the incompleteness of the data,

```
sapply(data, function(x) {sum(is.na(x))})
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

We can see that for the most part this data set is complete. Importantly, key variables like Survived, Pclass and Sex have all been recorded, besides the 891 NA observations we are using for the test data. However there are quite large gaps when it comes to Age (Missing 177) and Cabin number (Missing 687).

We can deal with missing values in each column by fill a suitable substitution such as mode, mean or median.

```
data$Embarked[is.na(data$Embarked)]<-mode(data$Embarked)
#Use mode for the missing "Embarked" values.

data$Age[is.na(data$Age)]<-mean(data$Age,na.rm = T)
#Use average of the existing age values for the missing "Age" values.
```

Manipulating the data

We can also try to deconstruct some of the variables to infer more information. For example, the passenger name is very difficult to use in any kind of analysis, so we can try and turn it into a factor variable based off the title of each person.

```
data$Title <- sapply(data$Name, function(x) {strsplit(x, split='[,.]')[[1]][2]})
data$Title <- sub(' ', '', data$Title) #removing spaces before title
kable(table(data$Sex, data$Title))
```

	Capt	Col	Don	Dr	Jonkheer	Lady	Major	Master	Miss	Mlle	Mme	Mr	Mrs	Ms	Rev	Sir
female	0	0	0	1	0	1	0	0	182	2	1	0	125	1	0	0
male	1	2	1	6	1	0	2	40	0	0	0	517	0	0	6	1

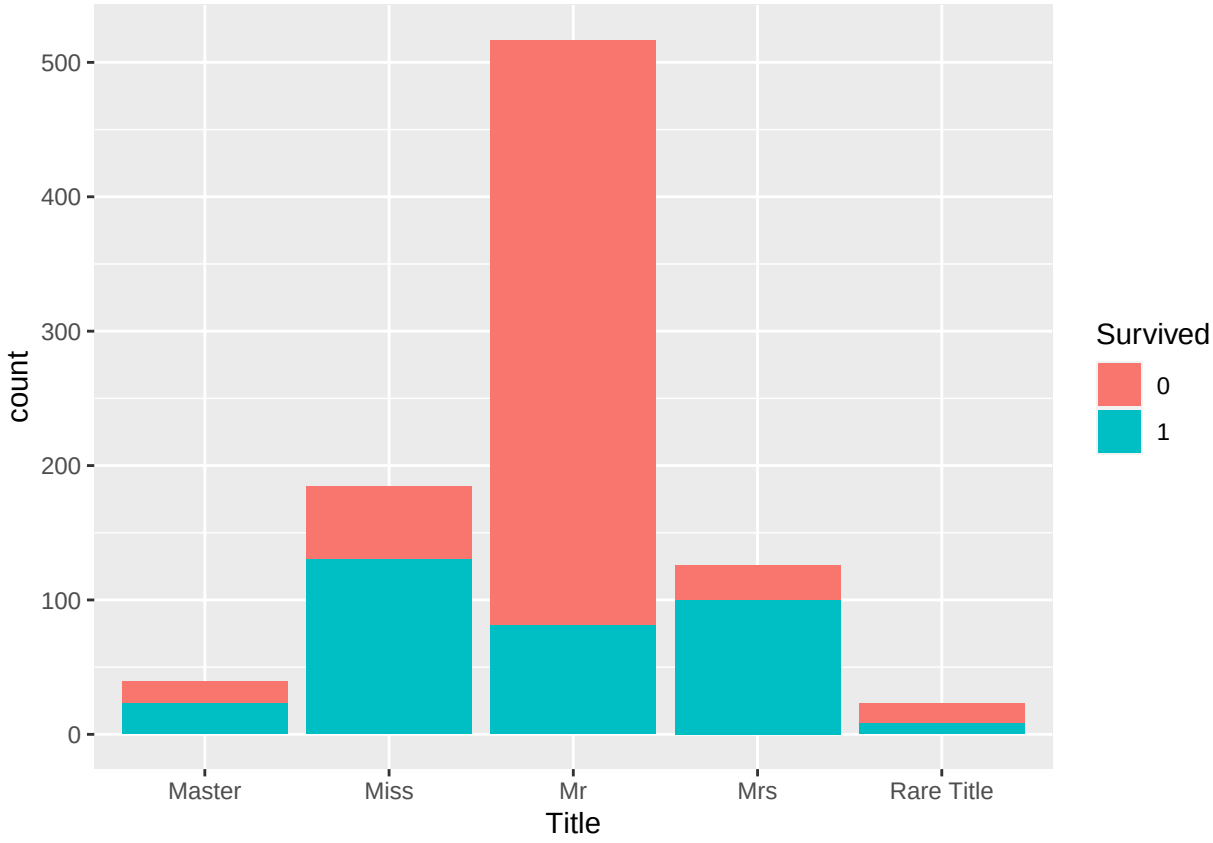
This is better than having over 1700 different names, although we can still reduce the number of dimensions of this new title variable. One way of doing this is grouping similar titles together, such as 'Ms','Miss' and 'Mlle'(Mademoiselle), all referring to young or unmarried women. Similarly we can grouo 'Mrs' and 'Mme'(Madame) as these refer to married women. Anything beside these two categories along with Mr and Master we can group in one 'Rare titles' category as miscellaneous titles.

```
data$Title[data$Title %in% c("Mlle", "Ms")] <- "Miss"
data$Title[data$Title== "Mme"] <- "Mrs"
data$Title[!(data$Title %in% c('Master', 'Miss', 'Mr', 'Mrs'))] <- "Rare Title"
data$Title <- as.factor(data$Title)
kable(table(data$Sex, data$Title))
```

	Master	Miss	Mr	Mrs	Rare Title
female	0	185	0	126	3
male	40	0	517	0	20

Now we have a much more clear table we can plot this to get an idea of how Title affects survival

```
ggplot(data[!is.na(data$Survived),], aes(x = Title, fill = Survived)) +
  geom_bar(stat='count', position='stack') +
  labs(x = 'Title') +theme_grey()
```



Create a variable "Familysize" which is the sum of variables "SibSp" and "Parch". Analysing family size seems to make more sense than analyse siblings and parents separately.

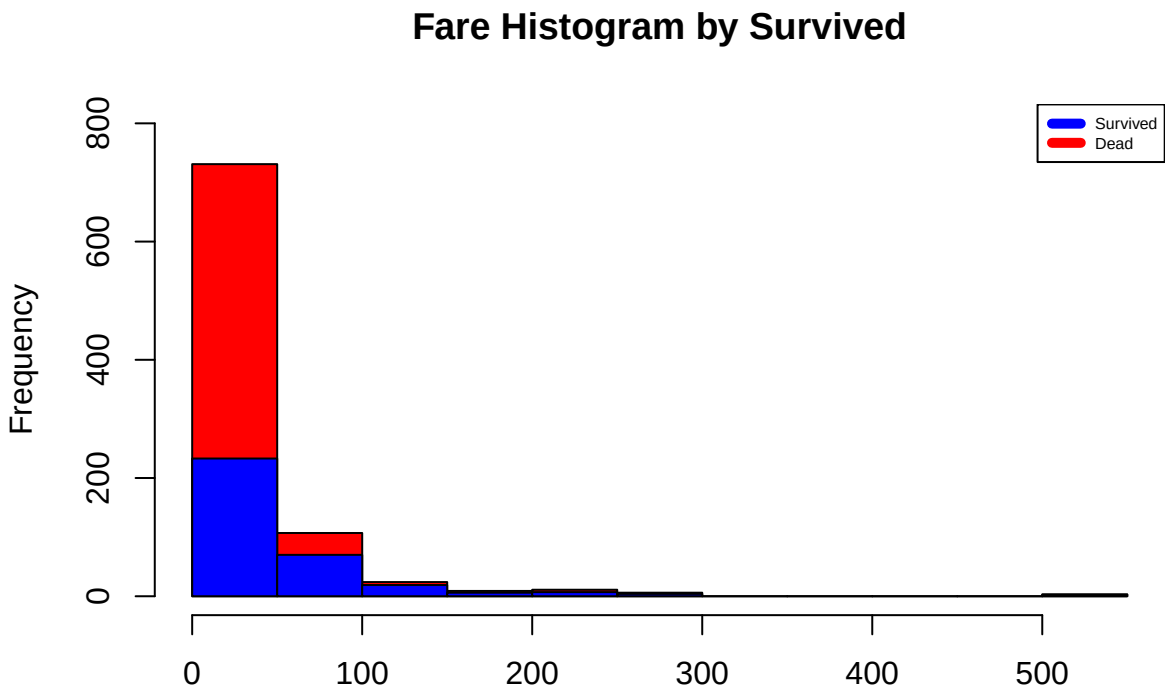
```
data$Familysize=data$SibSp+data$Parch
```

Visualising the data

We can then start visualizing the amount of deaths and who died due to the factor variables via bar charts

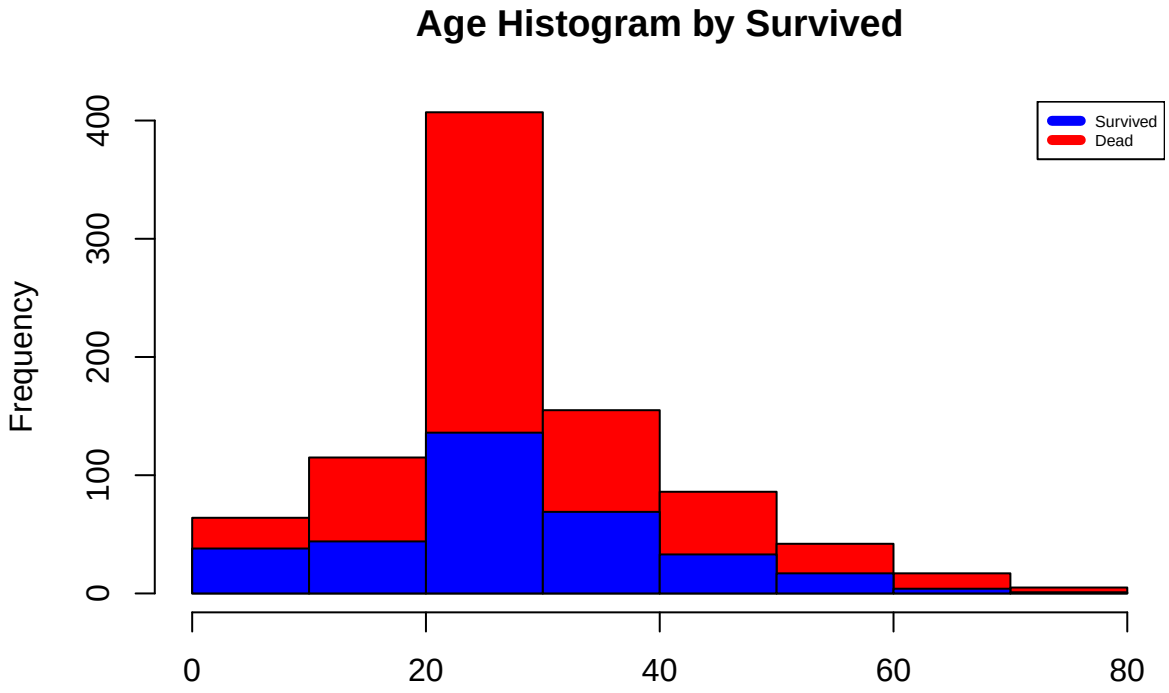
Let's firstly look at the most basic stacked histogram: We see that the survival rate of female is a lot higher than male, this is because that the men on board the ocean liner gave women and children priority access to the lifeboats.

```
histStack(x=data$Fare,z=factor(data$Survived),col=c("red","blue"),main = "Fare Histogram by Survived",y
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5) #Add legend to indicat
```

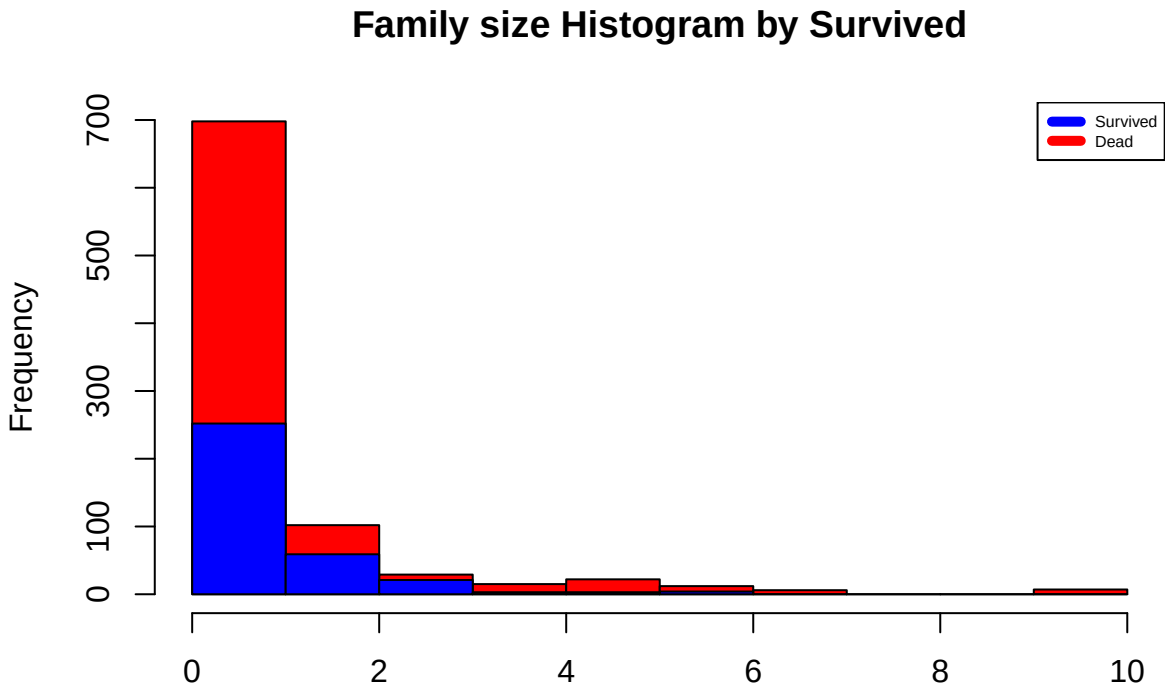


People between 20-40 years old has significantly lower survival rate.

```
histStack(x=data$Age,z=factor(data$Survived),col=c("red","blue"),main = "Age Histogram by Survived",yli
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5)
```

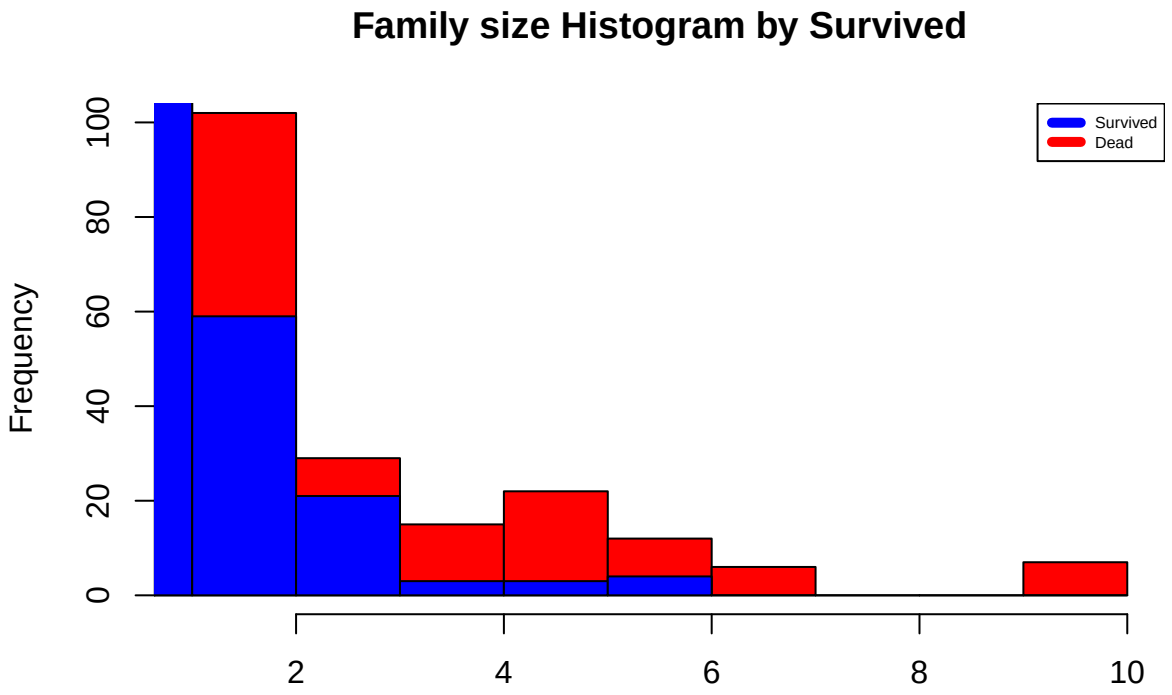



```
histStack(x=data$Familysize,z=factor(data$Survived),col=c("red","blue"),main = "Family size Histogram by
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5)
```



If we zoom in a little bit, we can see that the family size of 2 or 3 has significantly higher survival rate than family of other sizes.

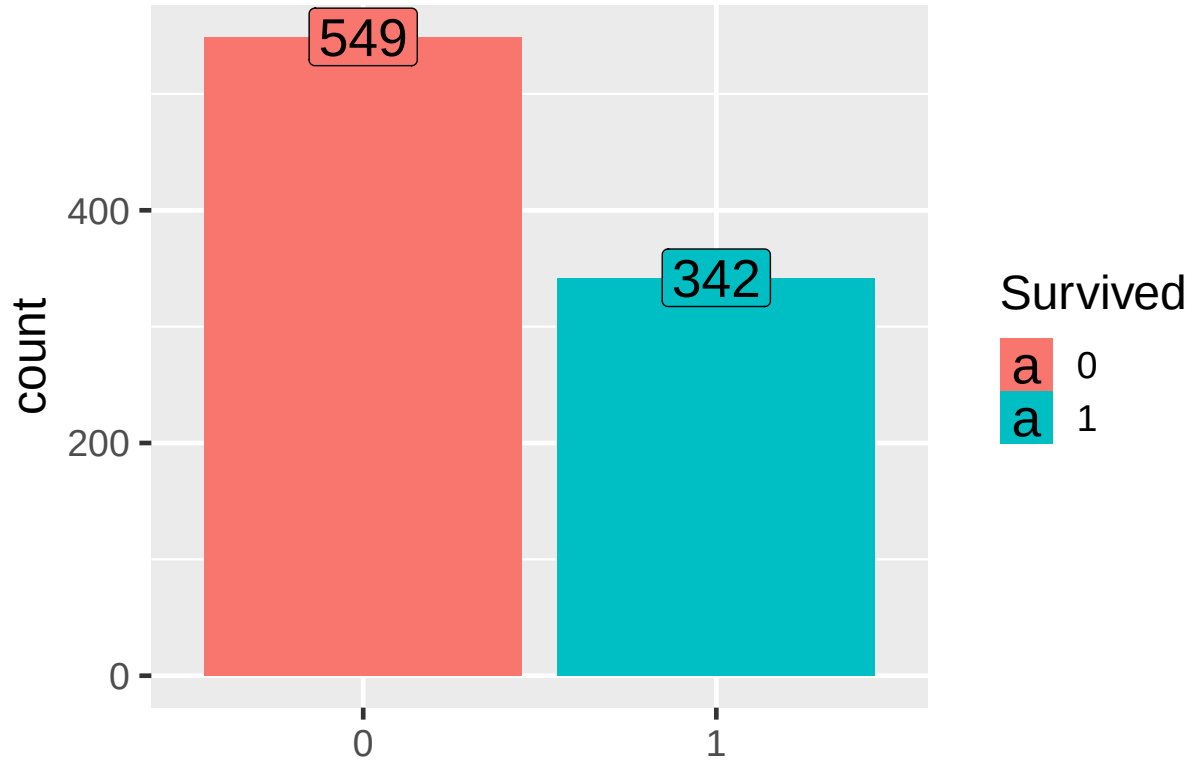
```
histStack(x=data$Familysize,z=factor(data$Survived),col=c("red","blue"),main = "Family size Histogram by
legend("topright",legend=c("Survived","Dead"),col=c("blue","red"),lwd=5,cex=0.5)
```



There are other options such as ggplot2, which plots prettier graphs, and contains more graphical features.

```
ggplot(data[!is.na(data$Survived),], aes(x = Survived, fill = Survived)) +
  geom_bar(stat='count') +
  labs(x = 'How many people died and survived on the Titanic?') +
```

```
        geom_label(stat='count',aes(label=..count..), size=7) +
        theme_grey(base_size = 18)
```

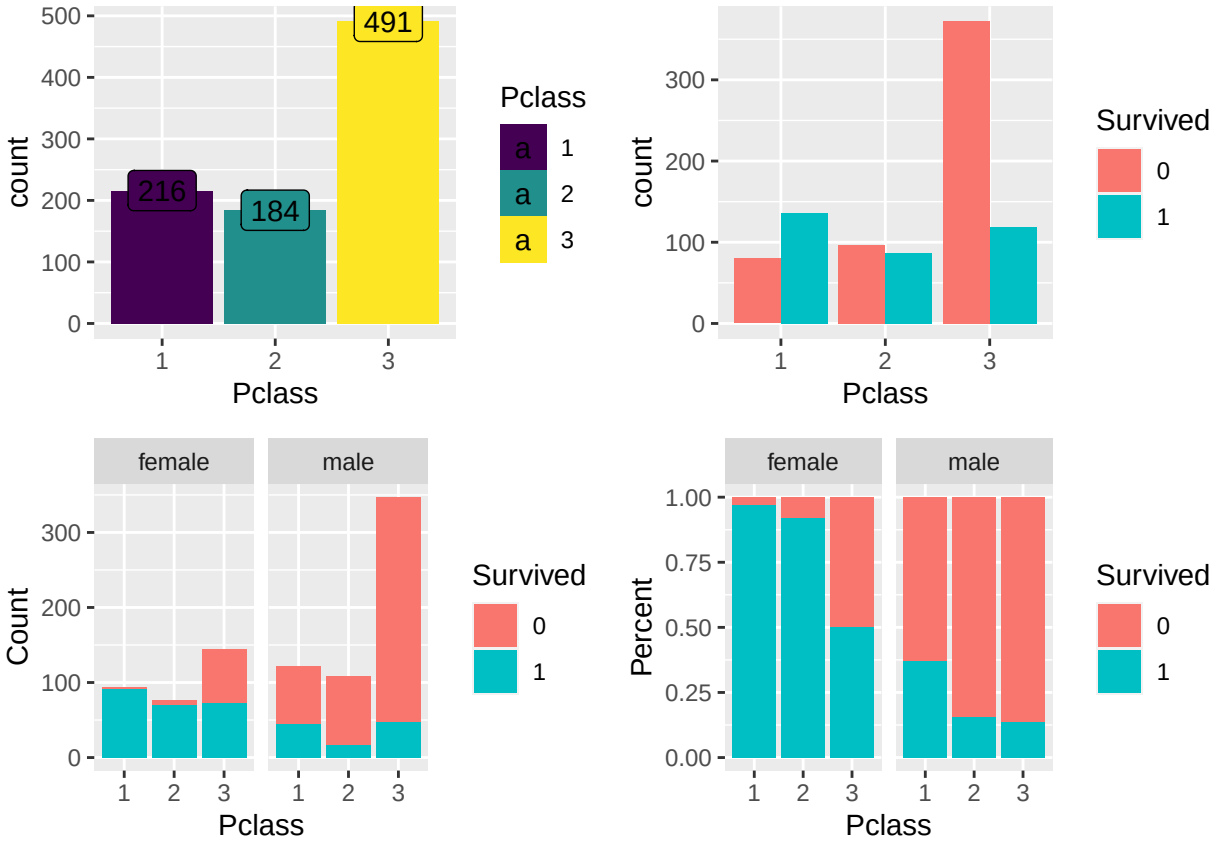


How many people died and survived on the Titanic?

And we can do further analysis on how class affects your chances of survival

```
p3 <- ggplot(data, aes(x = Pclass, fill = Pclass)) +
  geom_bar(stat='count', position='dodge') +
  labs(x = 'Pclass') + geom_label(stat='count', aes(label=..count..)) +
   theme(legend.position="none") + theme_grey()
p4 <- ggplot(data[!is.na(data$Survived),], aes(x = Pclass, fill = Survived)) +
  geom_bar(stat='count', position='dodge') + labs(x = 'Pclass') +
        theme(legend.position="none") + theme_grey()
p5 <- ggplot(data[!is.na(data$Survived),], aes(x = Pclass, fill = Survived)) +
  geom_bar(stat='count', position='stack') +
  labs(x = 'Pclass', y= "Count") + facet_grid(.~Sex) +
        theme(legend.position="none") + theme_grey()
p6 <- ggplot(data[!is.na(data$Survived),], aes(x = Pclass, fill = Survived)) +
  geom_bar(stat='count', position='fill') +
  labs(x = 'Pclass', y= "Percent") + facet_grid(.~Sex) +
        theme(legend.position="none") + theme_grey()

grid.arrange(p3, p4, p5, p6, ncol=2)
```


500
400
300
200
100
0
count
216
184
491
1
2
3
Pclass
Pclass
a
a
a
1
2
3
300
200
100
0
count
1
2
3
Pclass
Survived
0
1
female
male
300
200
100
0
Count
1
2
3
1
2
3
Pclass
Survived
0
1
female
male
1.00
0.75
0.50
0.25
0.00
Percent
1
2
3
1
2
3
Pclass
Survived
0
1