Titanic EDA analysis

Omar Abdulmaqsoud

2022-05-27

* load in libraries

```r
library(tidyverse)
```

* Load in Data

```r
titanic <- read_csv("C:\\Users\\User\\Downloads\\Titanic-dataset.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.
```

```r
head(titanic)
```

PassengerId	Survived	Pclass
<dbl>	<dbl>	<dbl>
1	0	3
2	1	1
3	1	3
4	1	1
5	0	3
6	0	3

6 rows | 1-3 of 12 columns

```r
tail(titanic)
```

PassengerId	Survived	Pclass
<dbl>	<dbl>	<dbl>
886	0	3
887	0	2
888	1	1

889	0	3
890	1	1
891	0	3

6 rows | 1-3 of 12 columns

* Preparing data for exploration

```
glimpse(titanic)
```

```
## Rows: 891
## Columns: 12
## $ PassengerId <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ Survived    <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1~
## $ Pclass      <dbl> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3~
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Fl~
## $ Sex         <chr> "male", "female", "female", "female", "male", "male", "mal~
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, ~
## $ SibSp       <dbl> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0~
## $ Parch       <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0~
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "37~
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625,~
## $ Cabin       <chr> NA, "C85", NA, "C123", NA, NA, "E46", NA, NA, NA, "G6", "C~
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S"~
```

* Convert some variables into factors

```
titanic$Survived=as.factor(titanic$Survived)
titanic$Pclass=as.factor(titanic$Pclass)
titanic$Sex=as.factor(titanic$Sex)
titanic$Embarked=as.factor(titanic$Embarked)
titanic$Cabin=as.factor(titanic$Cabin)
```

* Missing values

```
summary(titanic)
```

```
##   PassengerId    Survived Pclass      Name                Sex
##  Min.   :  1.0   0:549    1:216   Length:891         female:314
##  1st Qu.:223.5   1:342    2:184   Class :character   male  :577
##  Median :446.0            3:491   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   B96 B98    :  4   C   :168
##  1st Qu.:  7.91   C23 C25 C27:  4   Q   : 77
##  Median : 14.45   G6         :  4   S   :644
##  Mean   : 32.20   C22 C26    :  3   NA's:  2
##  3rd Qu.: 31.00   D          :  3
##  Max.   :512.33   (Other)    :186
##                   NA's       :687
```

```
table(titanic$Embarked)
```

```
##
##   C   Q   S
## 168  77 644
```

```
titanic$Embarked[is.na(titanic$Embarked)]<-"S"
```

```
titanic$Age[is.na(titanic$Age)]<-median(titanic$Age,na.rm = T)
```

```
summary(titanic)
```

```
##   PassengerId    Survived Pclass      Name               Sex
##  Min.   :  1.0   0:549    1:216   Length:891         female:314
##  1st Qu.:223.5   1:342    2:184   Class :character   male  :577
##  Median :446.0            3:491   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.:22.00   1st Qu.:0.000   1st Qu.:0.0000   Class :character
##  Median :28.00   Median :0.000   Median :0.0000   Mode  :character
##  Mean   :29.36   Mean   :0.523   Mean   :0.3816
##  3rd Qu.:35.00   3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :80.00   Max.   :8.000   Max.   :6.0000
##
##       Fare                 Cabin     Embarked
##  Min.   :  0.00   B96 B98    :  4   C:168
##  1st Qu.:  7.91   C23 C25 C27:  4   Q: 77
##  Median : 14.45   G6         :  4   S:646
##  Mean   : 32.20   C22 C26    :  3
##  3rd Qu.: 31.00   D          :  3
##  Max.   :512.33   (Other)    :186
##                   NA's       :687
```

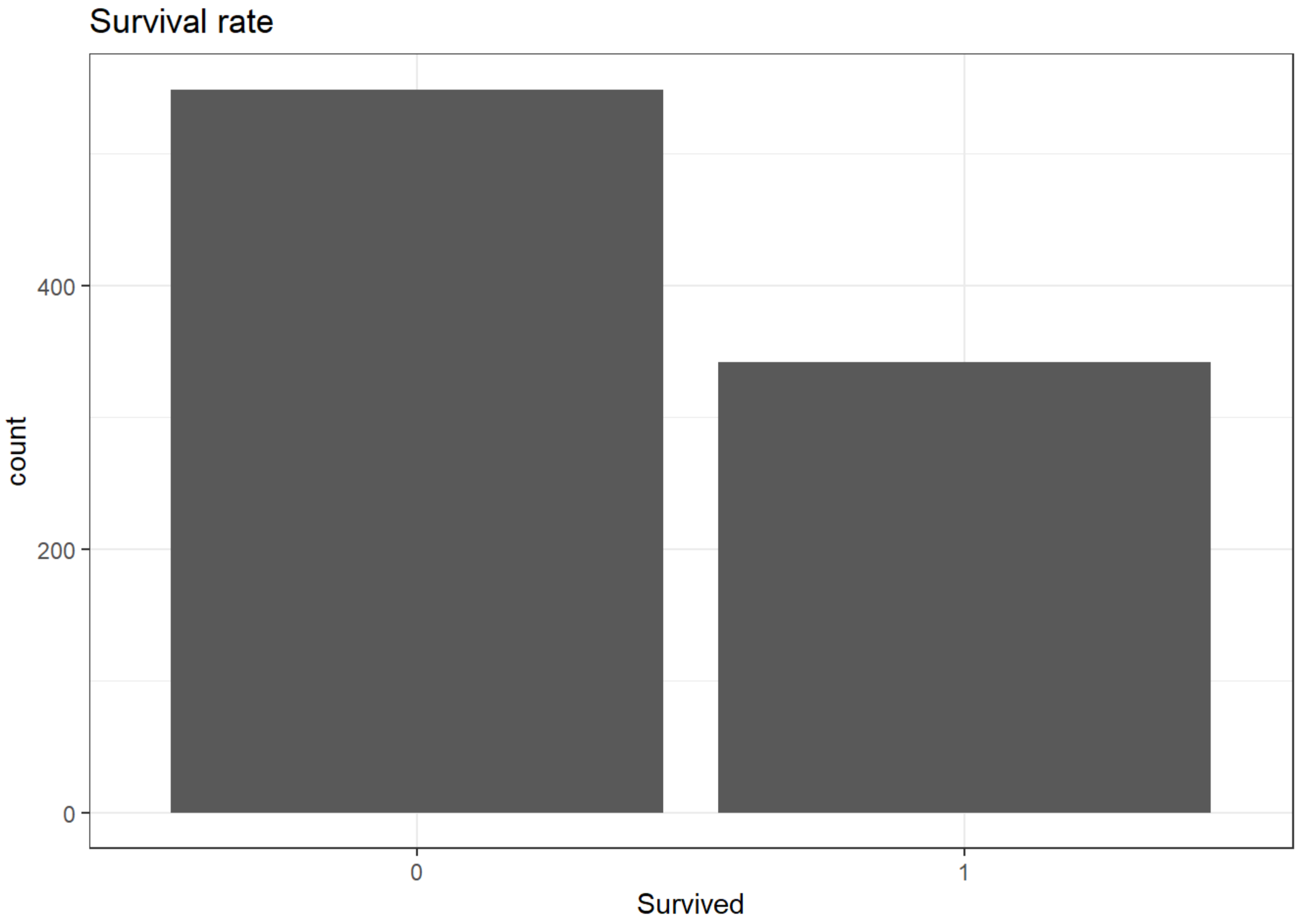
* Creating a new column called Familysize

```
titanic$Familysize<- 1+titanic$SibSp+titanic$Parch
```

Exploratory analysis

*What is the survival rate?

```
ggplot(titanic) + geom_bar(mapping = aes(x=Survived)) + labs(title = "Survival rate") + theme_bw()
```



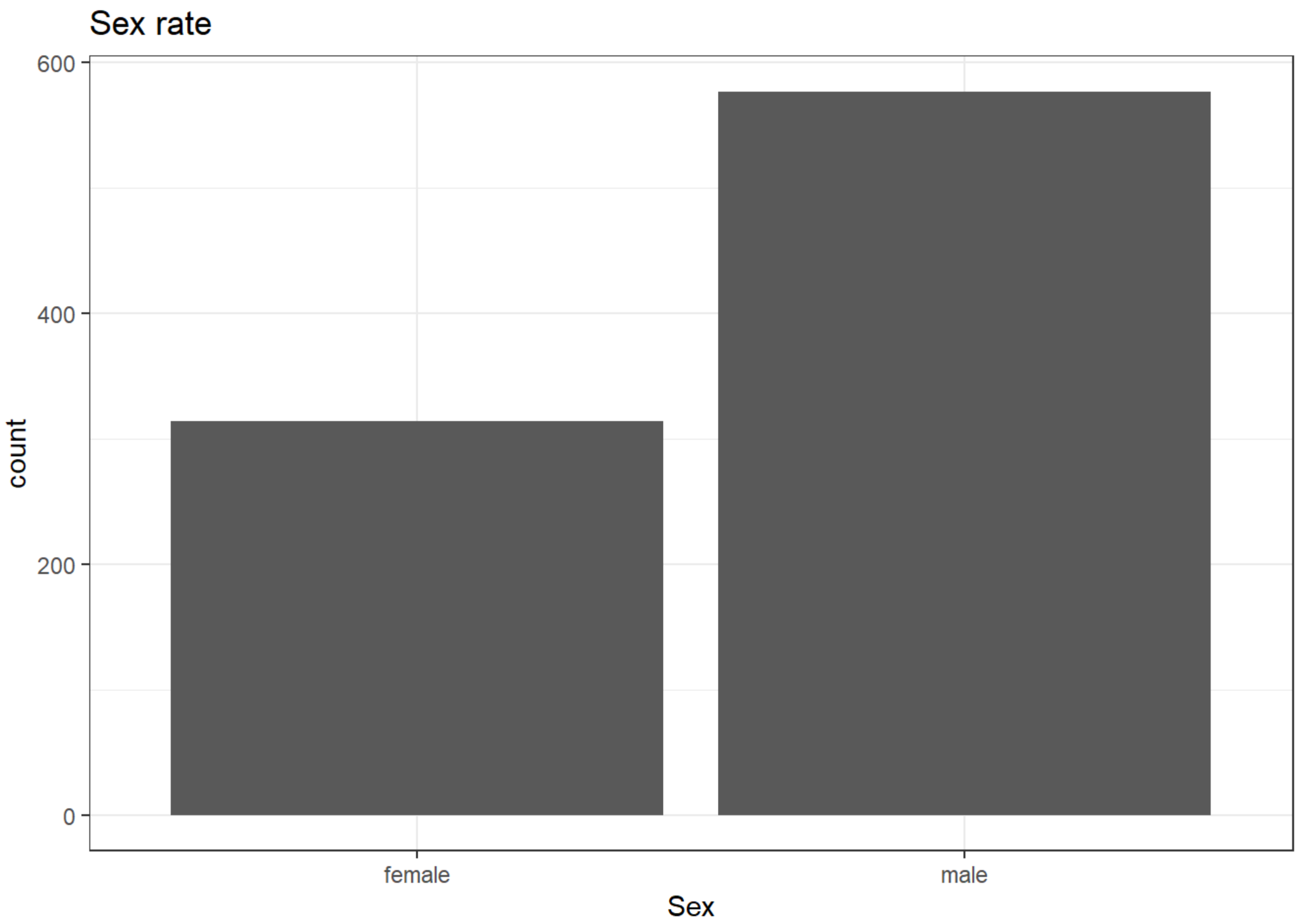
```
table(titanic$Survived)
```

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

There are 549 deaths and 342 survivors

* What is the gender rate?

```
ggplot(titanic) + geom_bar(mapping = aes(x=Sex)) + labs(title = "Sex rate") + theme_bw()
```



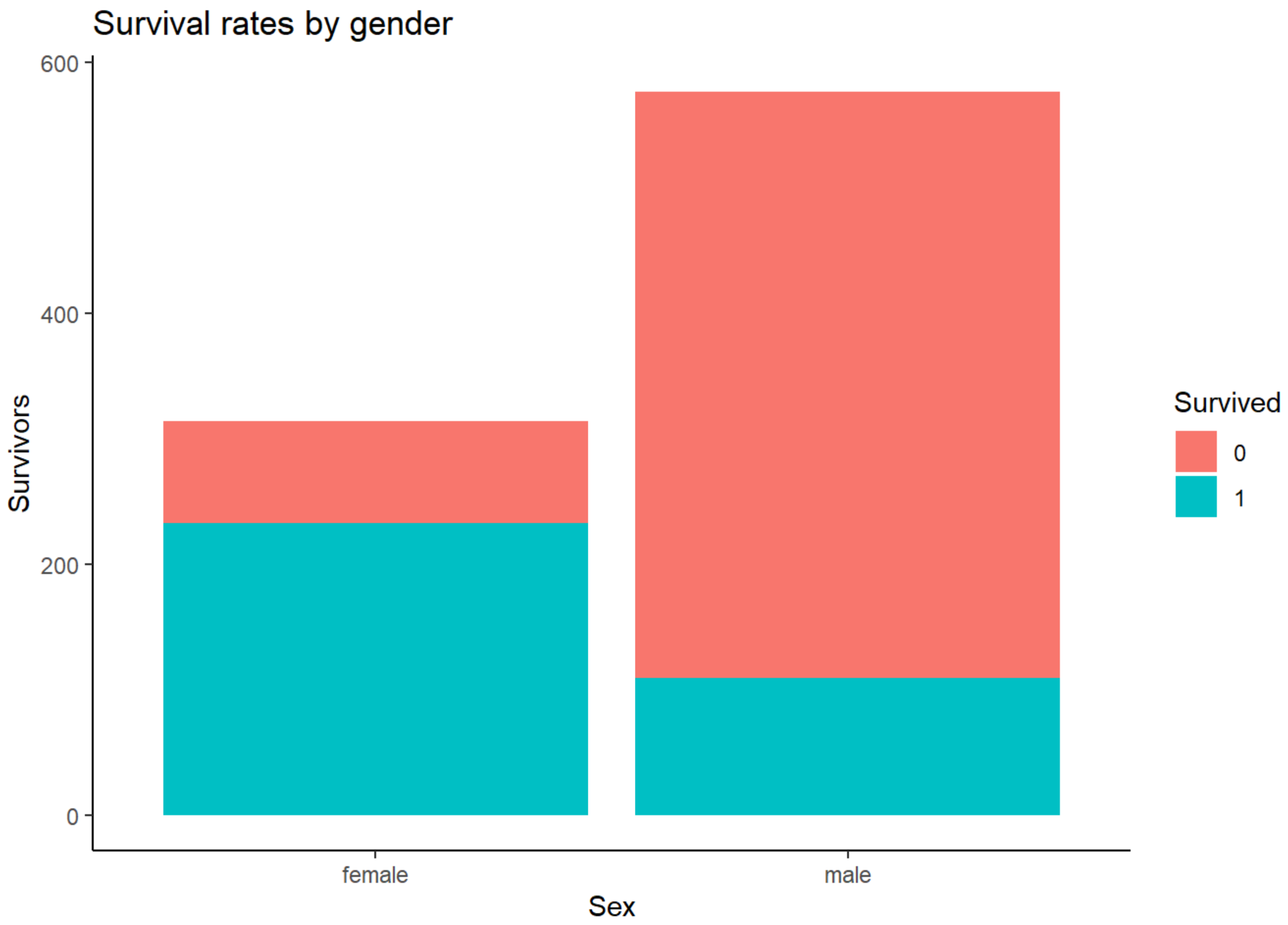
```
table(titanic$Sex)
```

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

There are 577 males and 314 females

*** What is the survival rate by gender?**

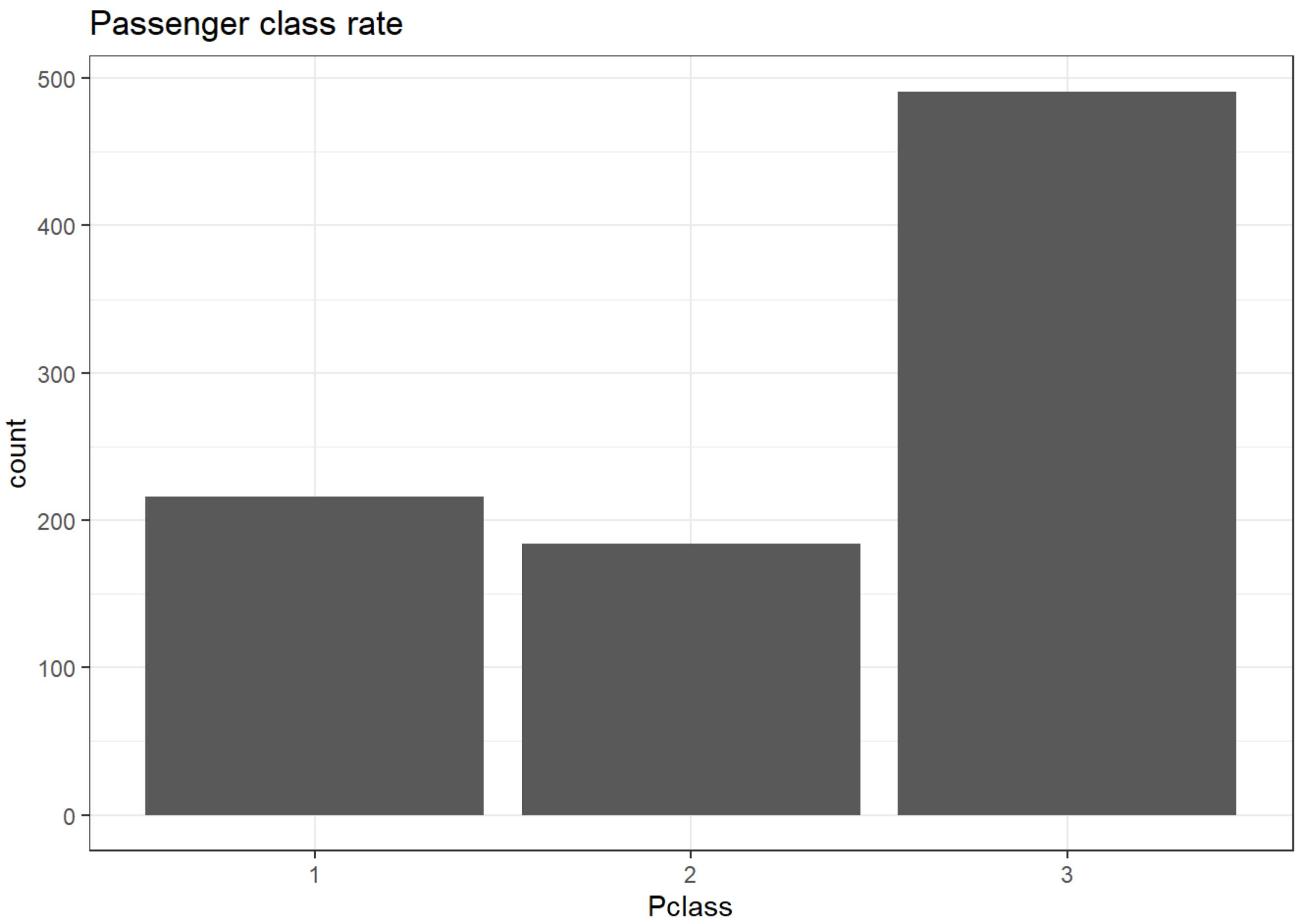
```
ggplot(titanic , aes(x= Sex, fill = Survived)) +
  geom_bar() +
  theme_classic() + labs(title = "Survival rates by gender",y="Survivors")
```



- Females are more likely to survive than men

* What is the Pclass rate?

```
ggplot(titanic) + geom_bar(mapping = aes(x=Pclass)) + labs(title = "Passenger class rate") + t
heme_bw()
```



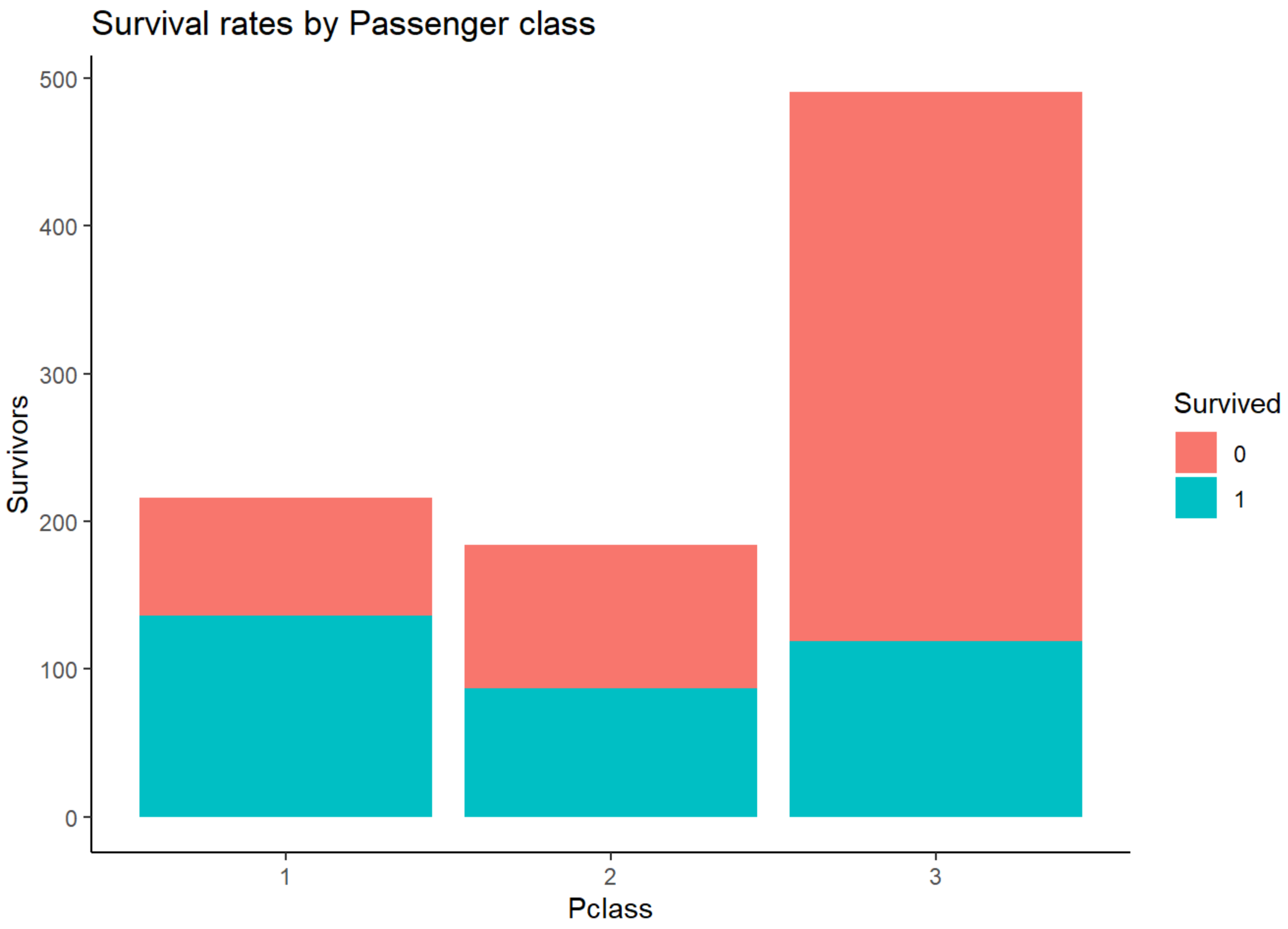
```
table(titanic$Pclass)
```

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

- Third class has the highest number of passengers (491 passenger). Then first class(216 passenger) and second class (184 passenger)

*** What is the survival rates by Pclass?**

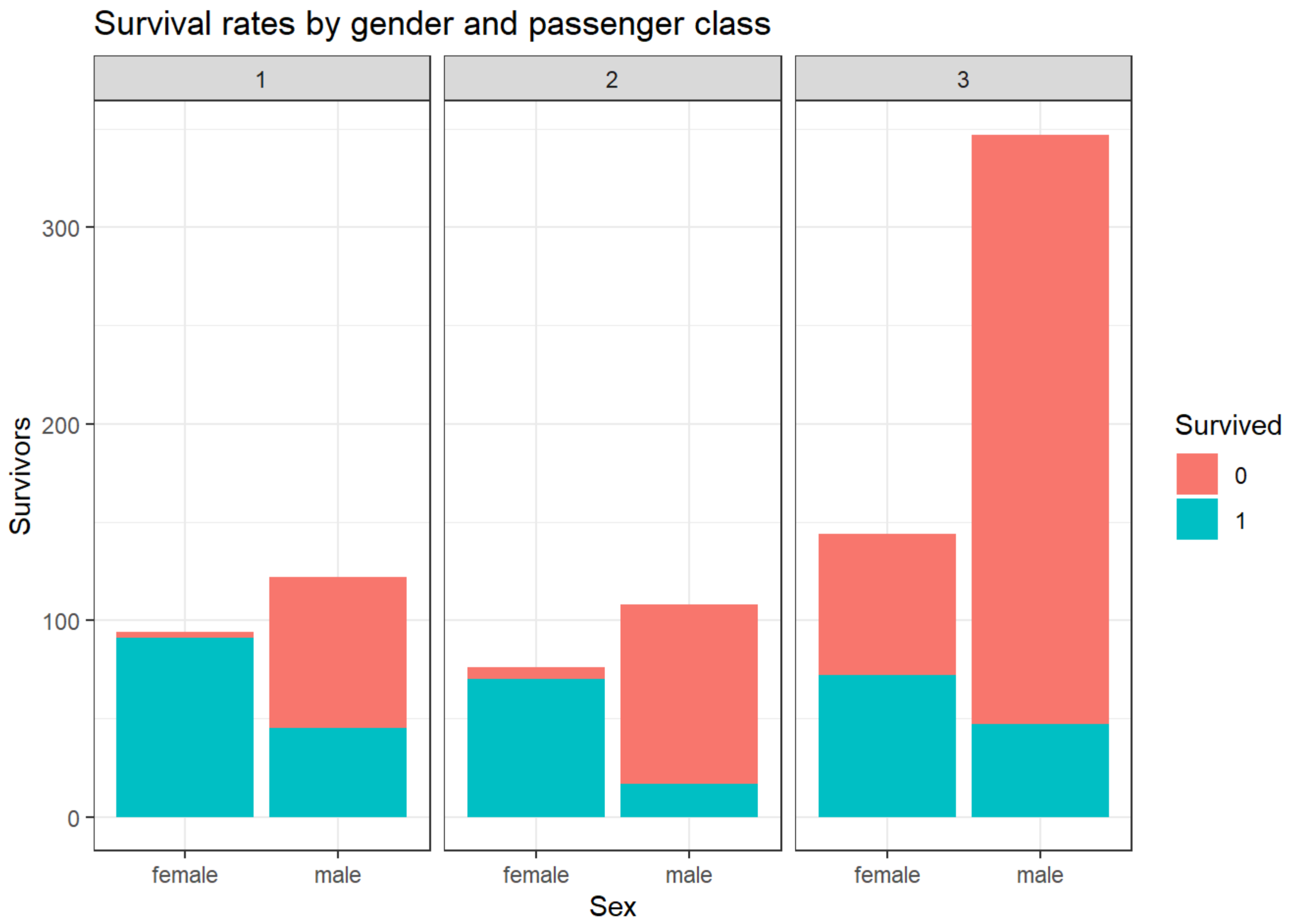
```
ggplot(titanic, aes(x=Pclass, fill = Survived)) +
  geom_bar() +
  theme_classic() + labs(title = "Survival rates by Passenger class",y="Survivors")
```



- First class passengers are more likely to survive than second and third class passengers

* What is the survival rate by gender and Pclass?

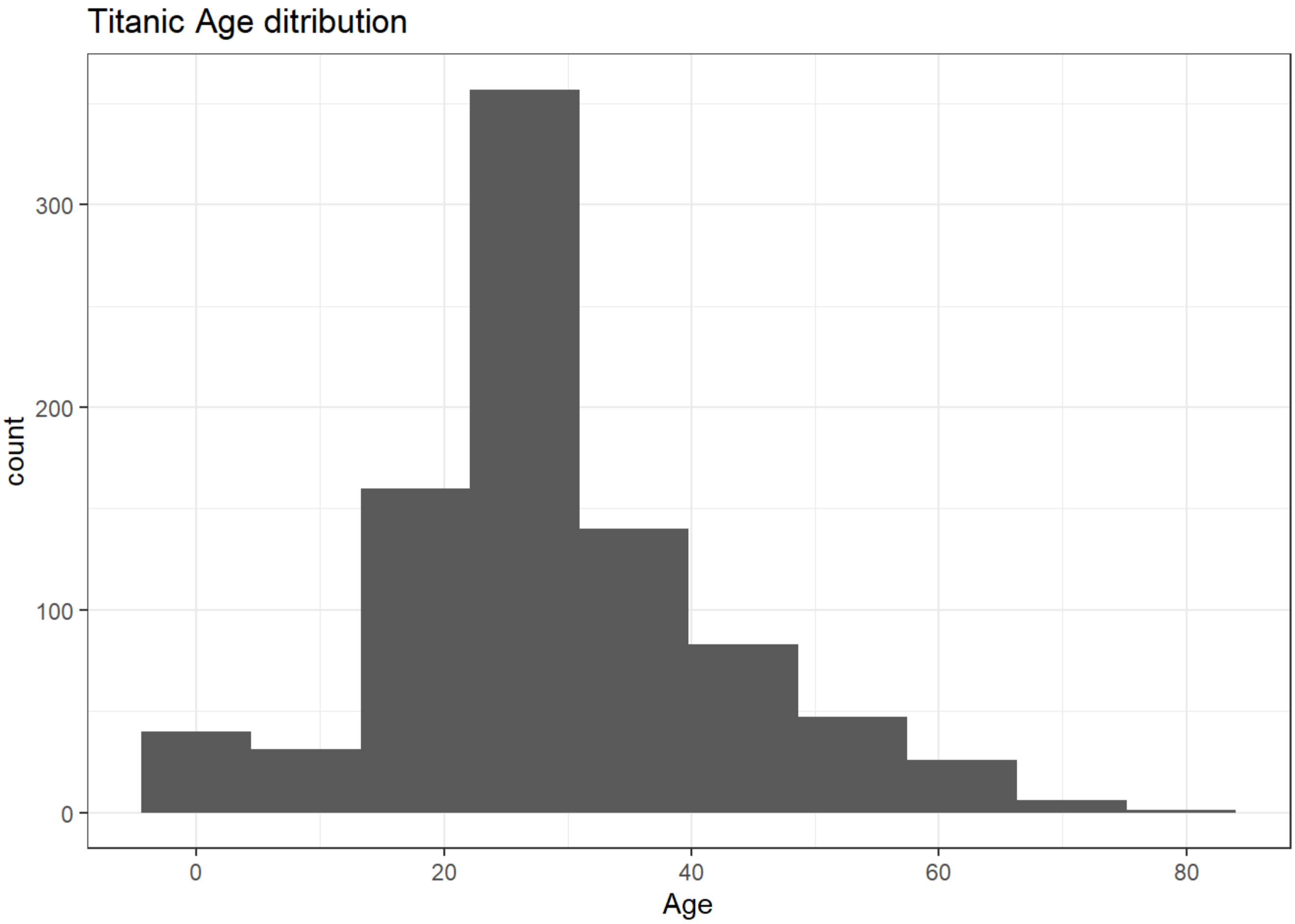
```
ggplot(titanic, aes(x=Sex,fill=Survived))+geom_bar()+theme_bw()+facet_wrap(~Pclass)+
  labs(title = "Survival rates by gender and passenger class",y="Survivors")
```



- First class Males and Females are more likely to survive than other classes passengers

*** How does the age distributes?**

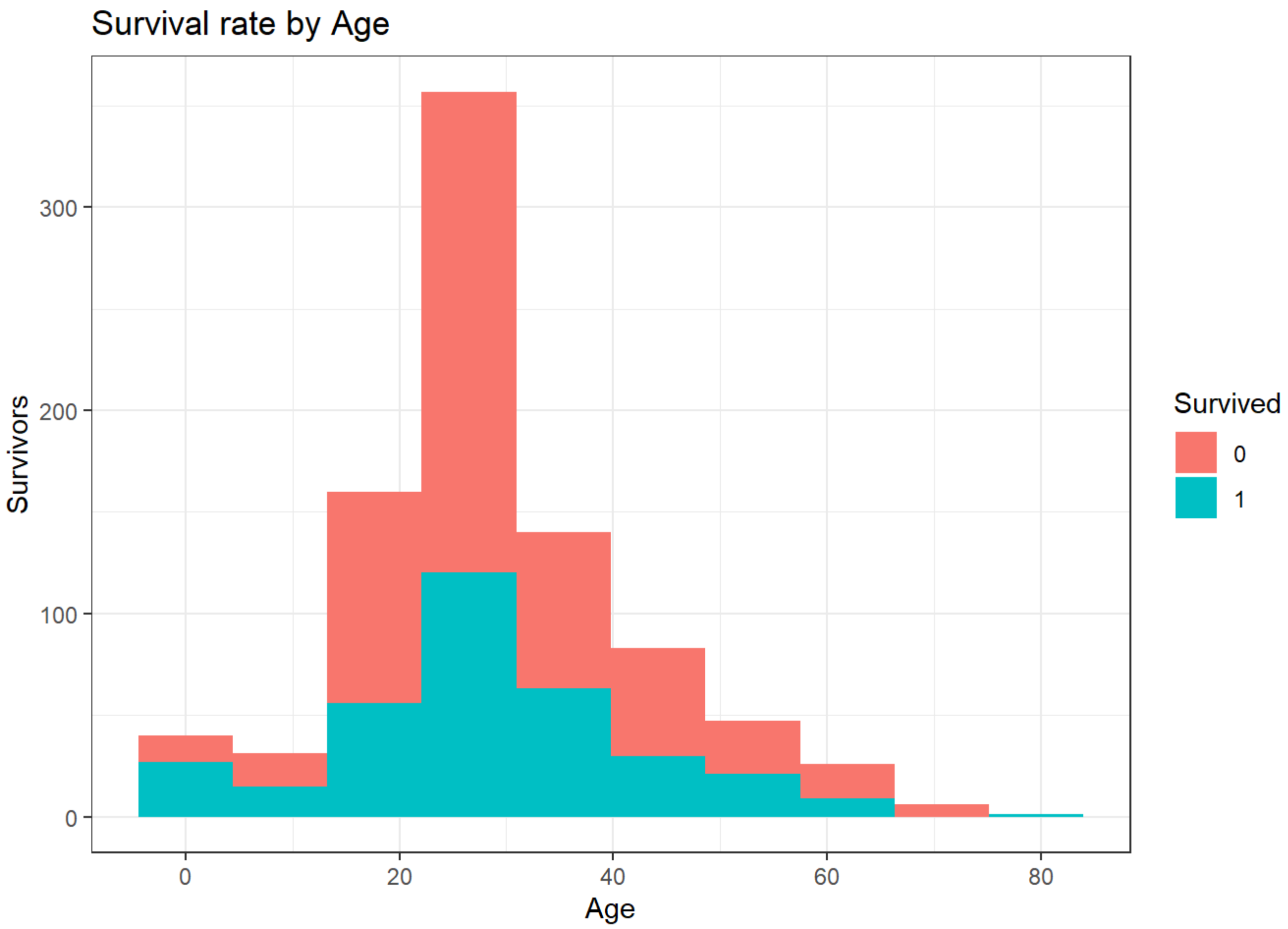
```
ggplot(titanic, aes(x=Age))+geom_histogram(bins = 10)+theme_bw()+
  labs(title = "Titanic Age ditribution")
```



- Passengers between the ages of 20 and 35 are the highest age group

* What is the survival rate by age?

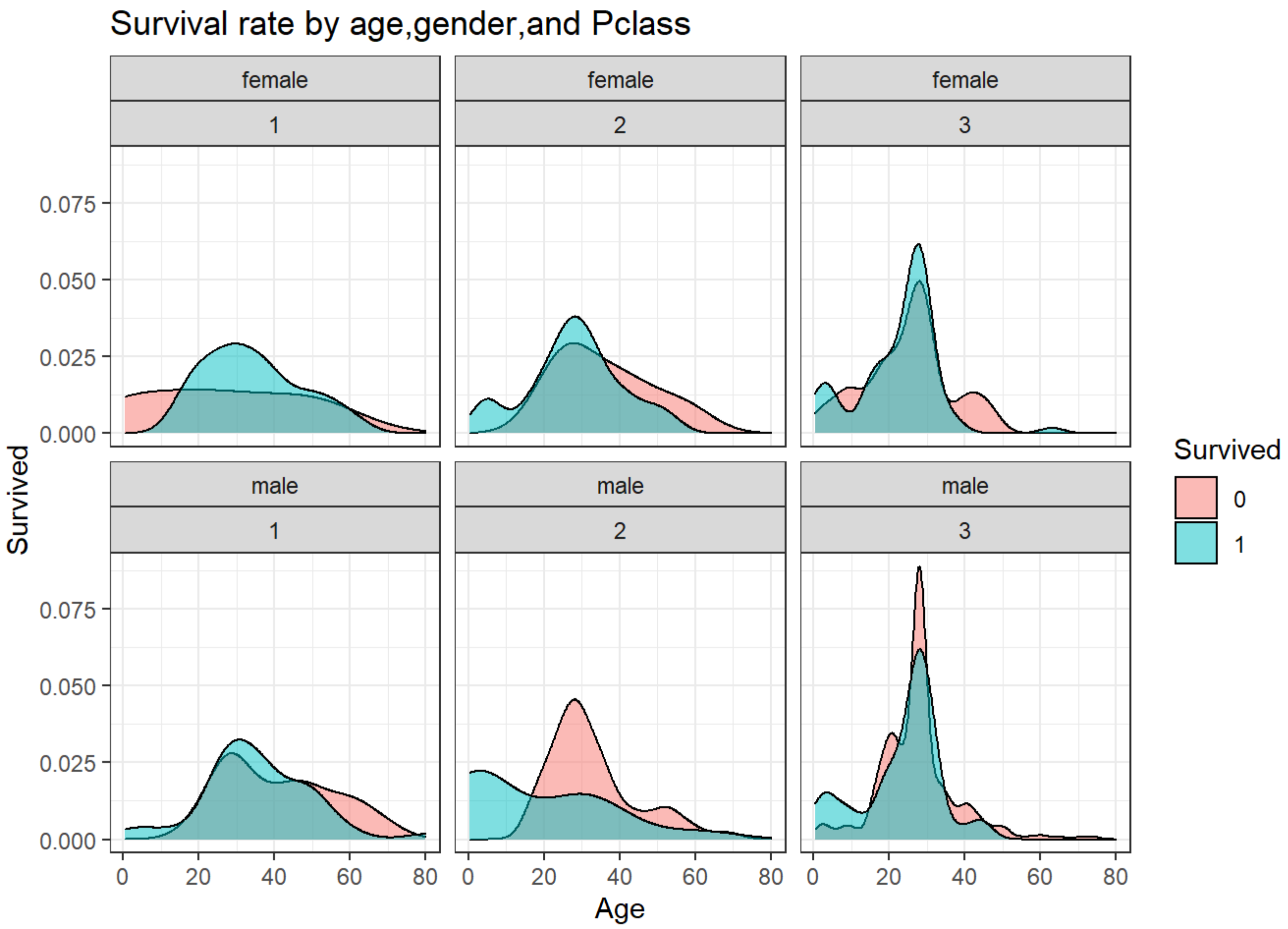
```
ggplot(titanic, aes(x=Age, fill=Survived))+geom_histogram(bins = 10)+theme_bw()+
  labs(title = "Survival rate by Age",y="Survivors")
```



- Ages between 23 and 32 are more likely to die than other ages

*** What is the Survival rate by Age, sex, and Pclass?**

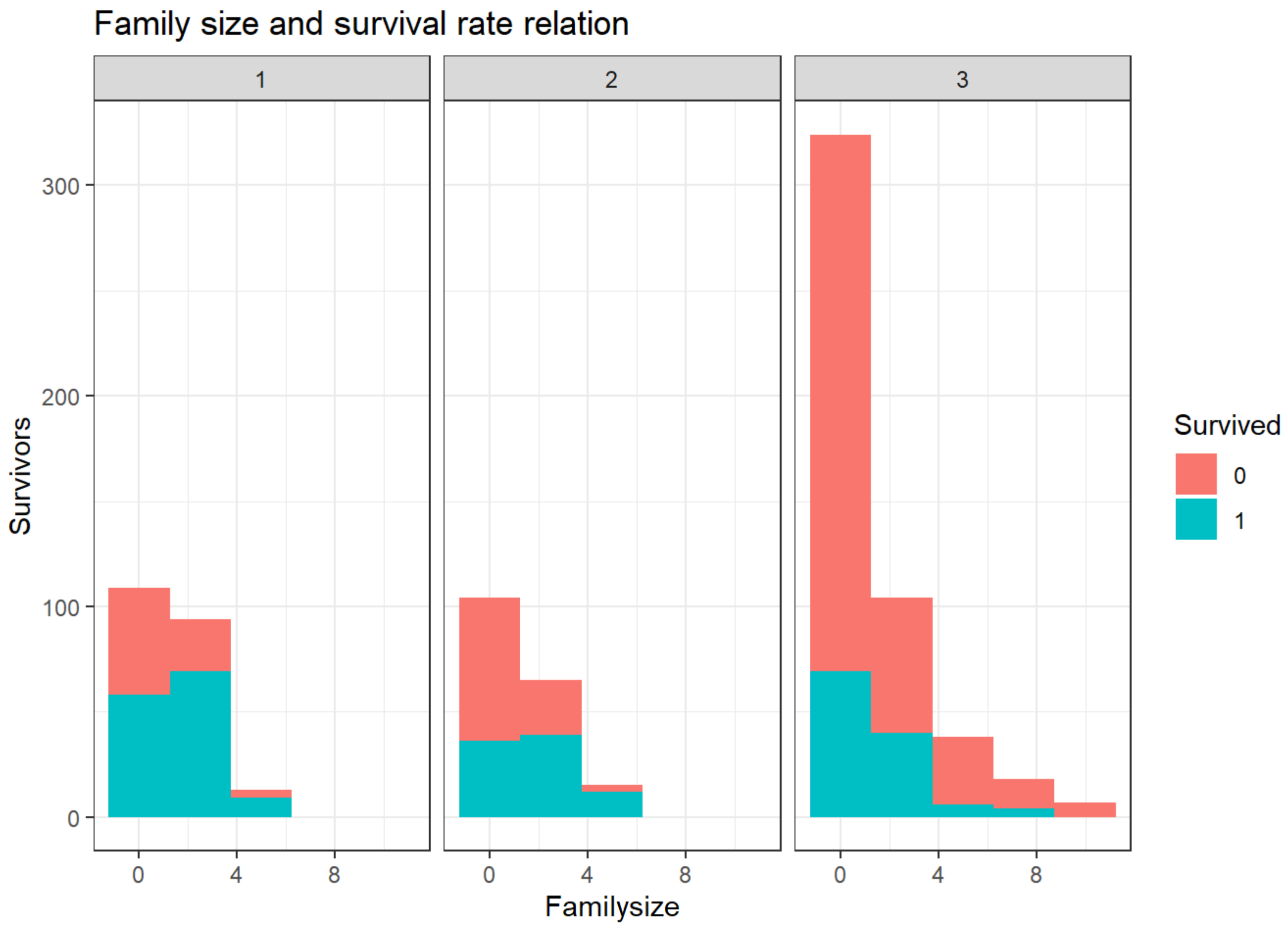
```
ggplot(titanic,aes(x=Age, fill=Survived))+geom_density(alpha=0.5)+theme_bw()+facet_wrap(Sex~Pc
lass)+
  labs(title = "Survival rate by age,gender,and Pclass",y="Survived")
```



- Children have better chances of survival

* Is there a relation between Familysize and survival rate?

```
ggplot(titanic,aes(x=Familysize,fill=Survived))+geom_histogram(bins = 5)+
  theme_bw()+
  facet_grid(~Pclass)+labs(title = "Family size and survival rate relation",y="Survivors")
```



- Families between between 1 and 3 members have better chances of survival

Insights

- - There are 549 deaths and 342 survivor
  - There are 577 males and 314 females
  - Children and Females are more likely to survive than men
  - Third class has the highest number of passengers (491 passenger). Then first class(216 passenger) and second class (184 passenger)
  - First class passengers are more likely to survive than second and third class passengers
  - Passengers between the ages of 20 and 35 are the highest age group
  - Ages between 23 and 32 are more likely to die than other ages
  - Children have better chances of survival
  - Families between between 1 and 3 members have better chances of survival