RE6124019 HomeWork1(Titanic)

Import Package

```
library(ggplot2) #Plot
library(titanic) #Dataset
```

Load Dataset

```
train <- titanic_train
test <- titanic_test

str(train)
```

```
'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 T
 $ 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" ...
```

```
str(test)
```

```
'data.frame':   418 obs. of  11 variables:
 $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
 $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
 $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas
 $ Sex        : chr  "male" "female" "male" "male" ...
 $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
 $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
 $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
 $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
 $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
 $ Cabin      : chr  "" "" "" "" ...
 $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

Exploratory Data Analysis (EDA)

```
#Type
sapply(train, class)
```

```
PassengerId    Survived      Pclass        Name         Sex         Age
  "integer"   "integer"   "integer" "character" "character"   "numeric"
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
  "integer"   "integer" "character"   "numeric" "character" "character"
```

```
sapply(test, class)
```

```
PassengerId      Pclass        Name         Sex         Age       SibSp
  "integer"   "integer" "character" "character"   "numeric"   "integer"
      Parch      Ticket        Fare       Cabin    Embarked
  "integer" "character"   "numeric" "character" "character"
```

```
#Missing Value
sum(is.na(train))
```

```
[1] 177
```

```
sum(is.na(test))
```

```
[1] 87
```

```r
#Duplicate Value
sum(duplicated(train))
```

```
[1] 0
```

```r
sum(duplicated(test))
```

```
[1] 0
```

```r
#Summary
summary(train)
```

```
  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
```

```r
summary(test)
```

```
 PassengerId         Pclass          Name               Sex
Min.   : 892.0   Min.   :1.000   Length:418         Length:418
1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
Median :1100.5   Median :3.000   Mode  :character   Mode  :character
Mean   :1100.5   Mean   :2.266
3rd Qu.:1204.8   3rd Qu.:3.000
Max.   :1309.0   Max.   :3.000

      Age             SibSp            Parch           Ticket
Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
Mean   :30.27   Mean   :0.4474   Mean   :0.3923
3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
Max.   :76.00   Max.   :8.0000   Max.   :9.0000
NA's   :86
      Fare            Cabin              Embarked
Min.   :  0.000   Length:418         Length:418
1st Qu.:  7.896   Class :character   Class :character
Median : 14.454   Mode  :character   Mode  :character
Mean   : 35.627
3rd Qu.: 31.500
Max.   :512.329
NA's   :1
```

Data Analysis (Every Variable)

```
#Survived
table(train$Survived)
```

```
  0   1
549 342
```

```
#Pclass
table(train$Pclass)
```

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

```r
table(test$Pclass)
```

```
  1   2   3
107  93 218
```

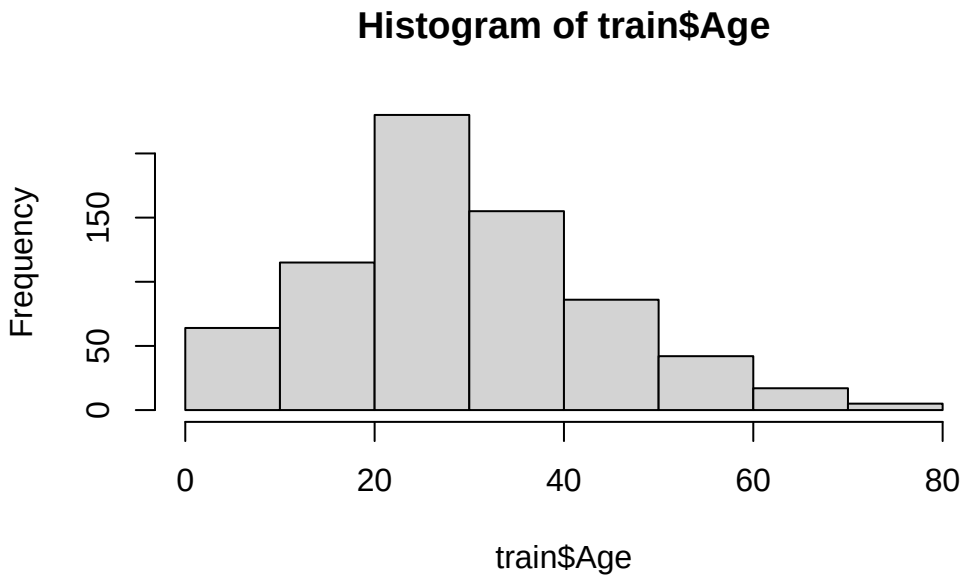
```r
#Sex
table(train$Sex)
```

```
female   male
   314    577
```

```r
table(test$Sex)
```

```
female   male
   152    266
```

```r
#Age
hist(train$Age)
```



```r
hist(test$Age)
```


```r
#SibSp
table(train$SibSp)
```

```
  0   1   2   3   4   5   8
608 209  28  16  18   5   7
```

```r
table(test$SibSp)
```

```
  0   1   2   3   4   5   8
283 110  14   4   4   1   2
```

```r
#Parch
table(train$Parch)
```

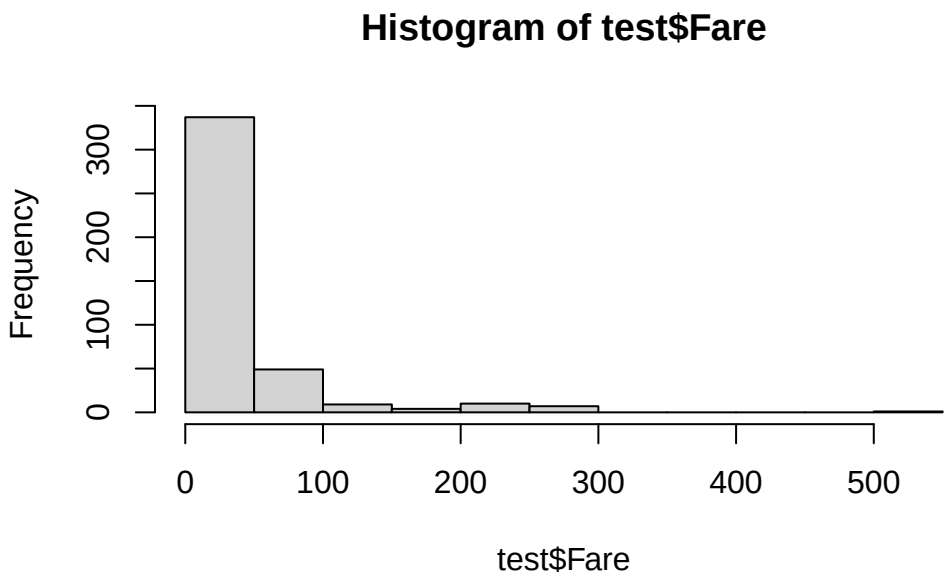
```
  0   1   2   3   4   5   6
678 118  80   5   4   5   1
```

```
table(test$Parch)
```

```
  0   1   2   3   4   5   6   9
324  52  33   3   2   1   1   2
```

```
#Fare
hist(train$Fare)
```


```
hist(test$Fare)
```




```
#Embarked
table(train$Embarked)
```

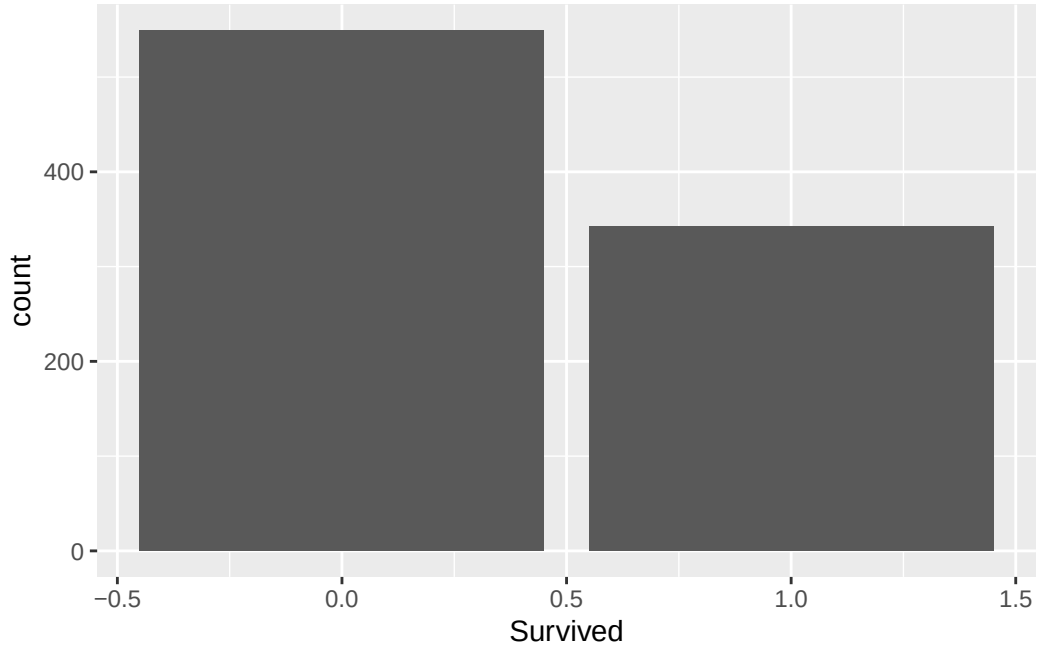
```
     C   Q   S
  2 168  77 644
```

```
table(test$Embarked)
```

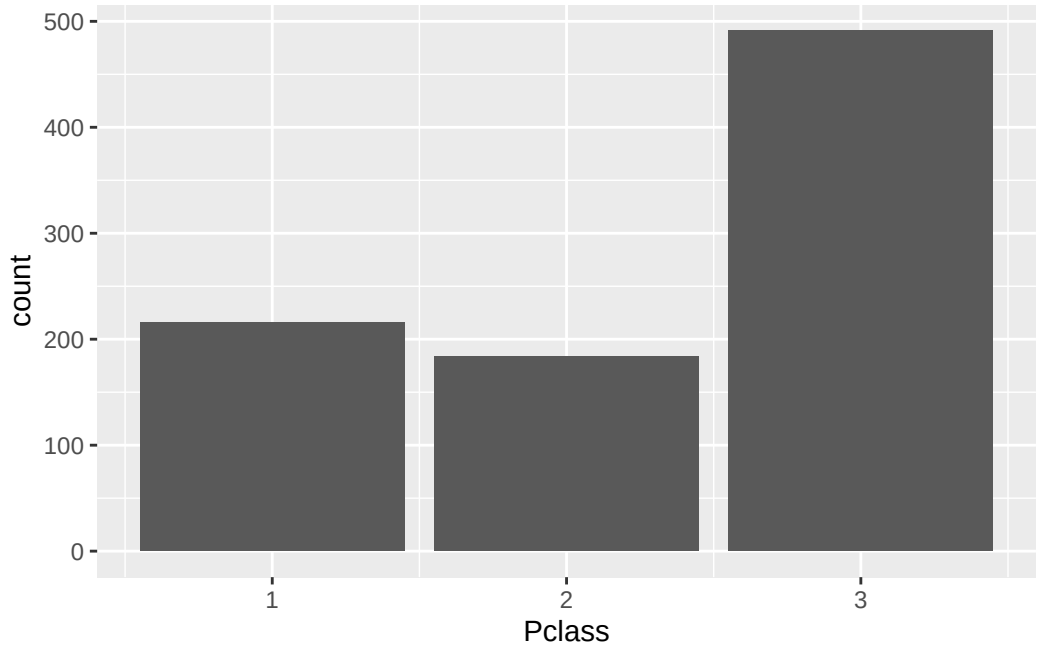
```
  C   Q   S
102  46 270
```

Plot (GGPLOT2)

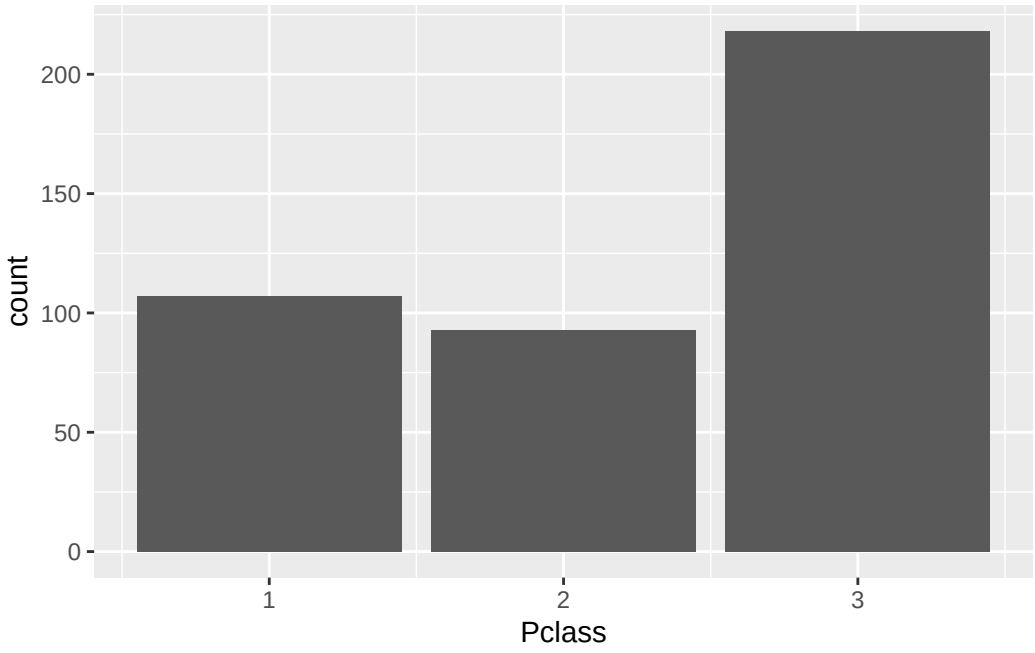
```
#Survived
ggplot(train, aes(Survived)) + geom_bar()
```



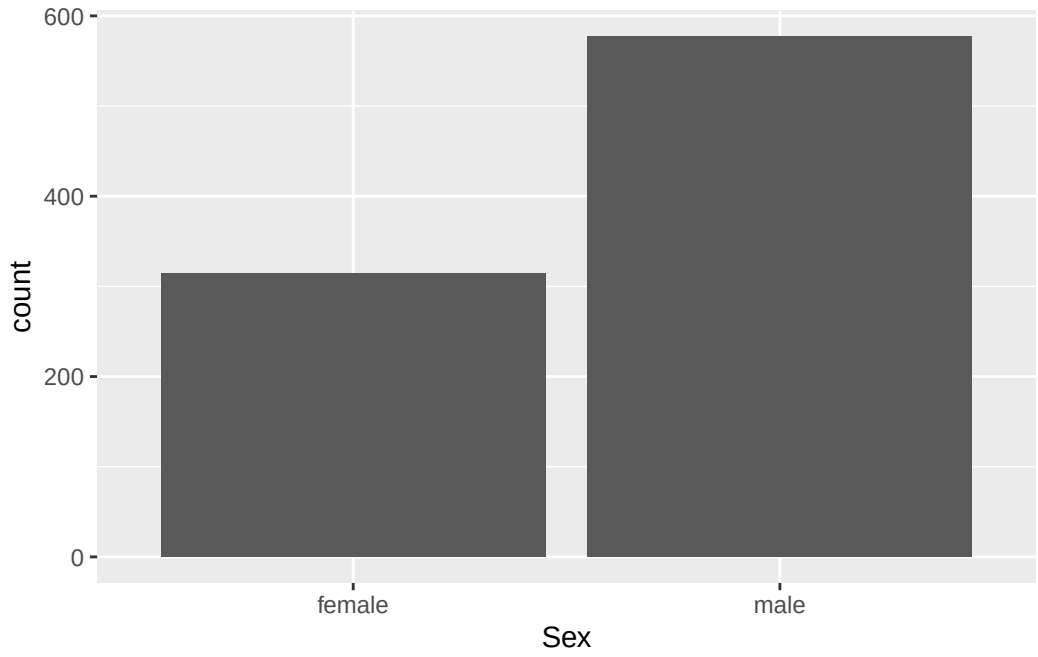

```
#Pclass
ggplot(train, aes(Pclass)) + geom_bar()
```



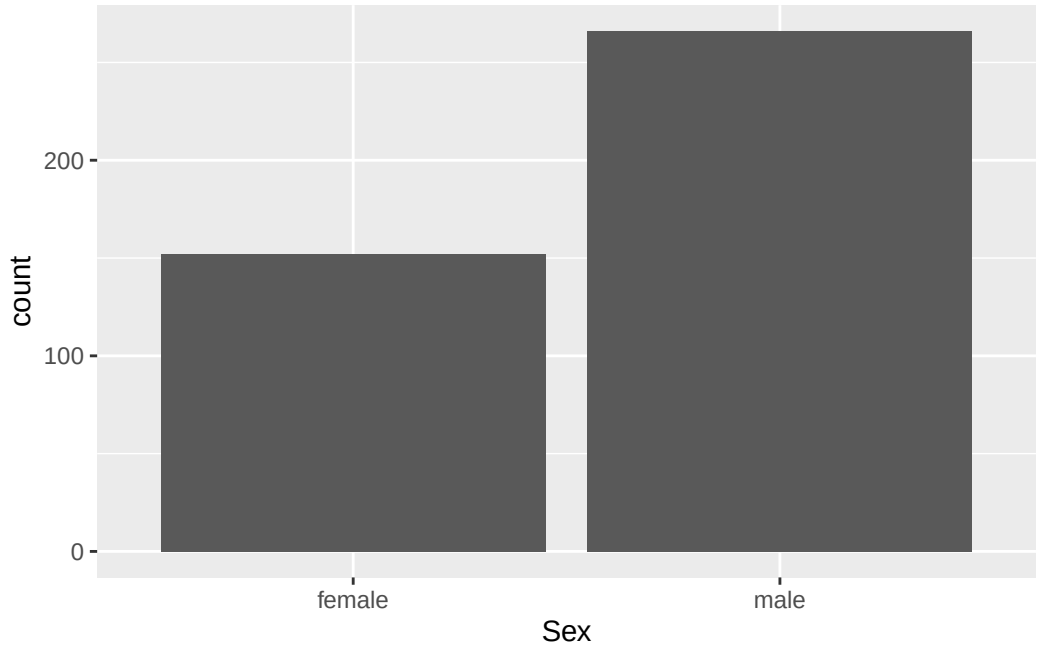
```
ggplot(test, aes(Pclass)) + geom_bar()
```




```
#Sex
ggplot(train, aes(Sex)) + geom_bar()
```



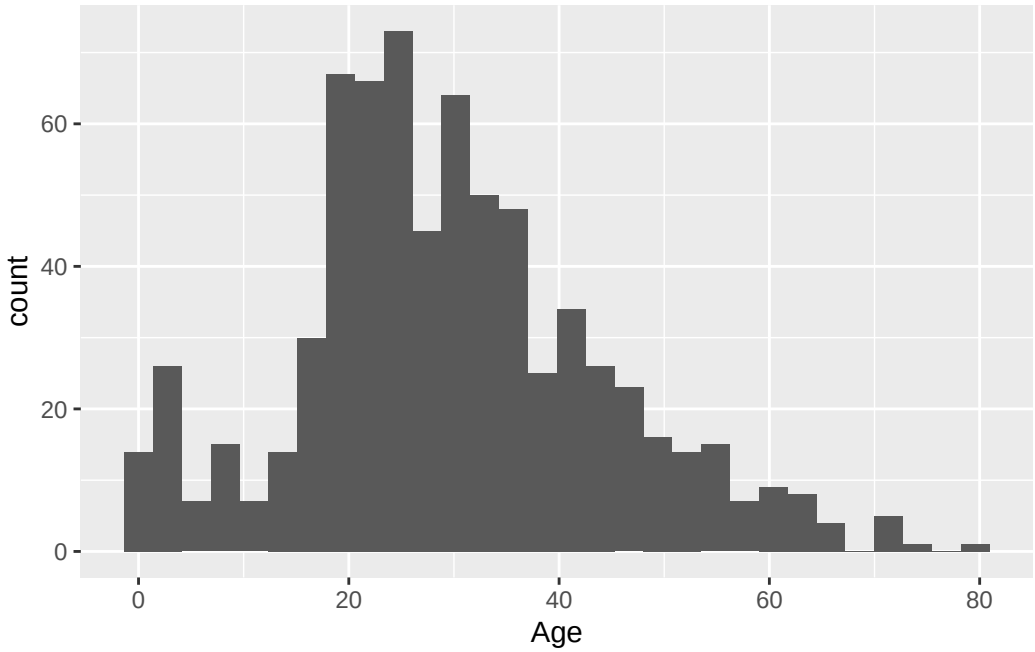

```
ggplot(test, aes(Sex)) + geom_bar()
```



```
#Age
ggplot(train, aes(Age)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

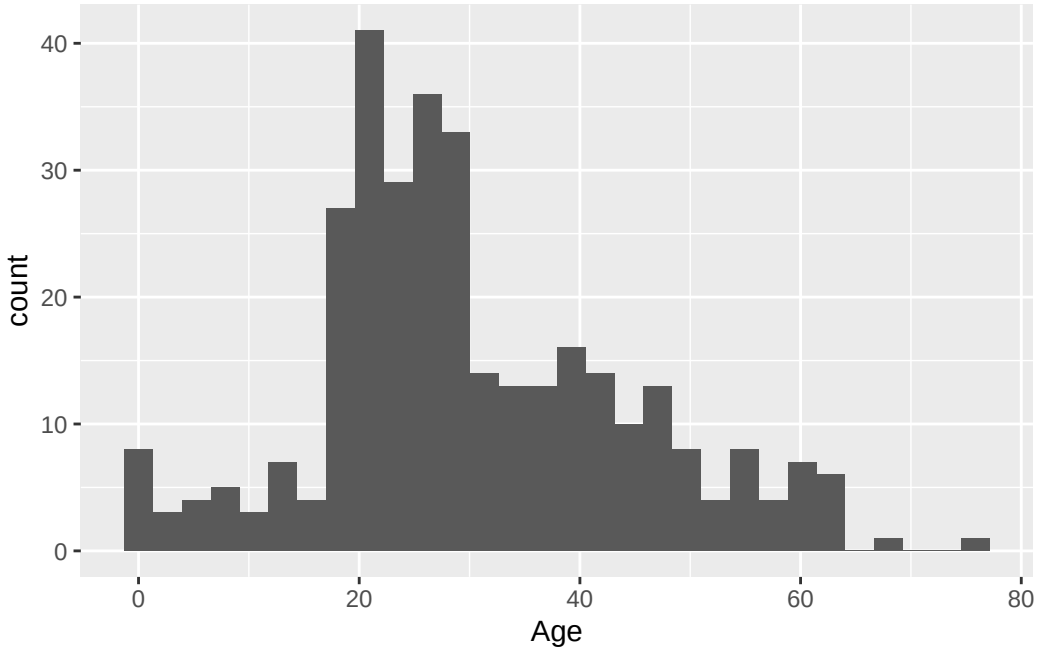
```
Warning: Removed 177 rows containing non-finite values (`stat_bin()`).
```



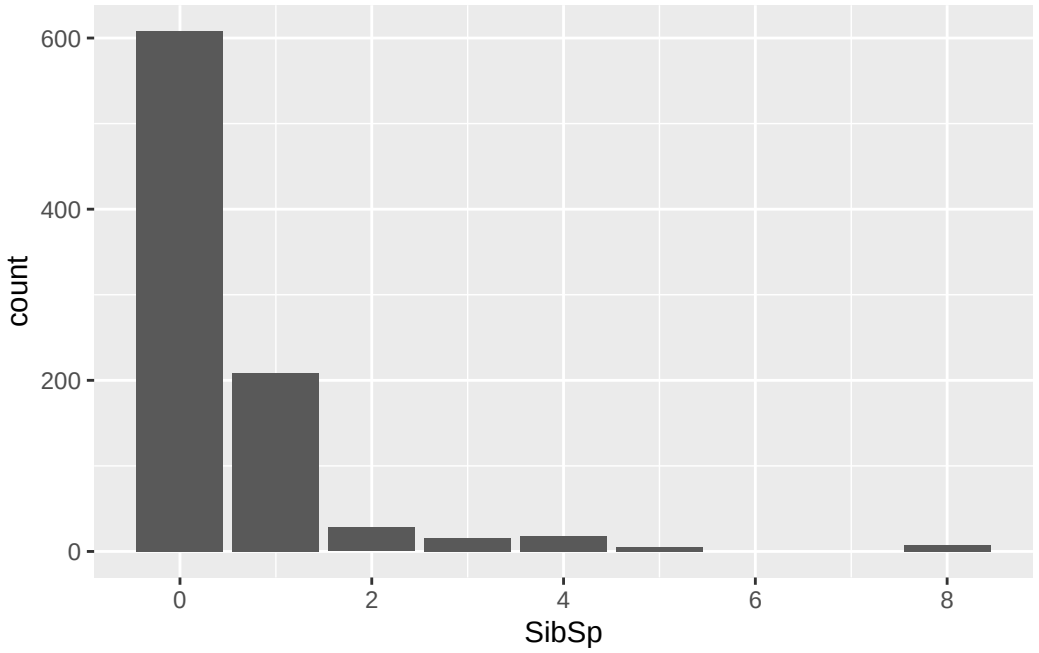

```
ggplot(test, aes(Age)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

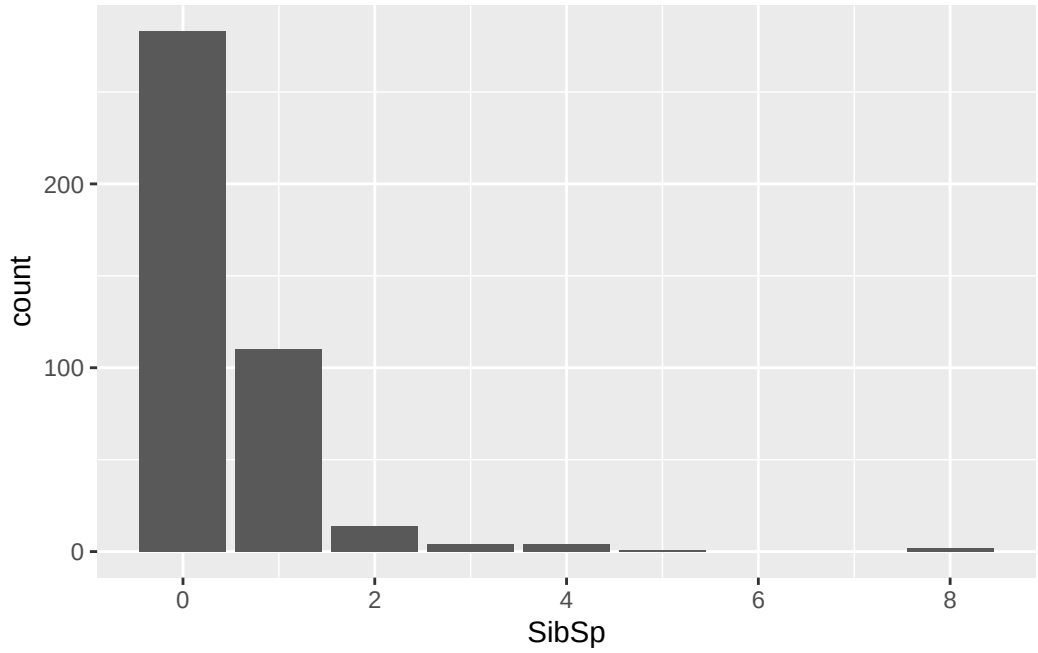
```
Warning: Removed 86 rows containing non-finite values (`stat_bin()`).
```



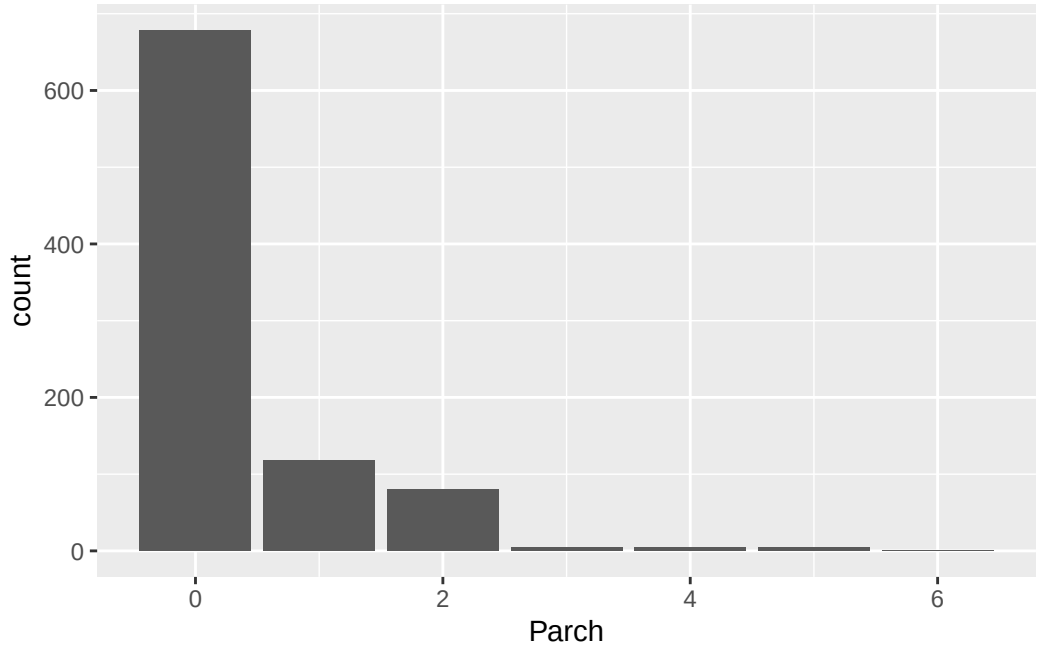

```
#SibSp
ggplot(train, aes(SibSp)) + geom_bar()
```



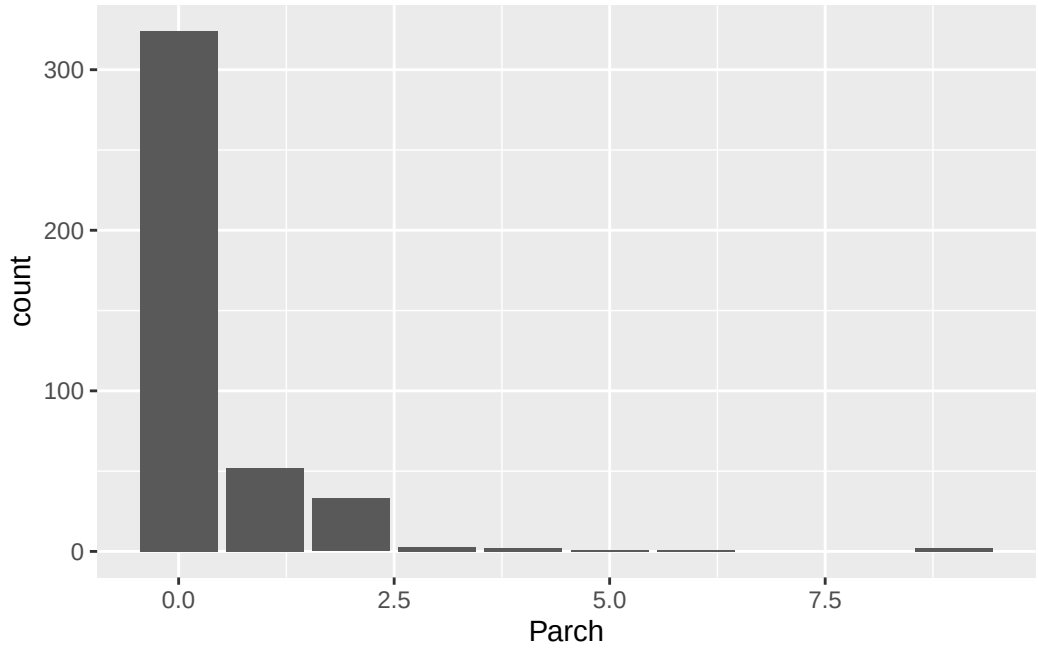
```
ggplot(test, aes(SibSp)) + geom_bar()
```




```
#Parch
ggplot(train, aes(Parch)) + geom_bar()
```

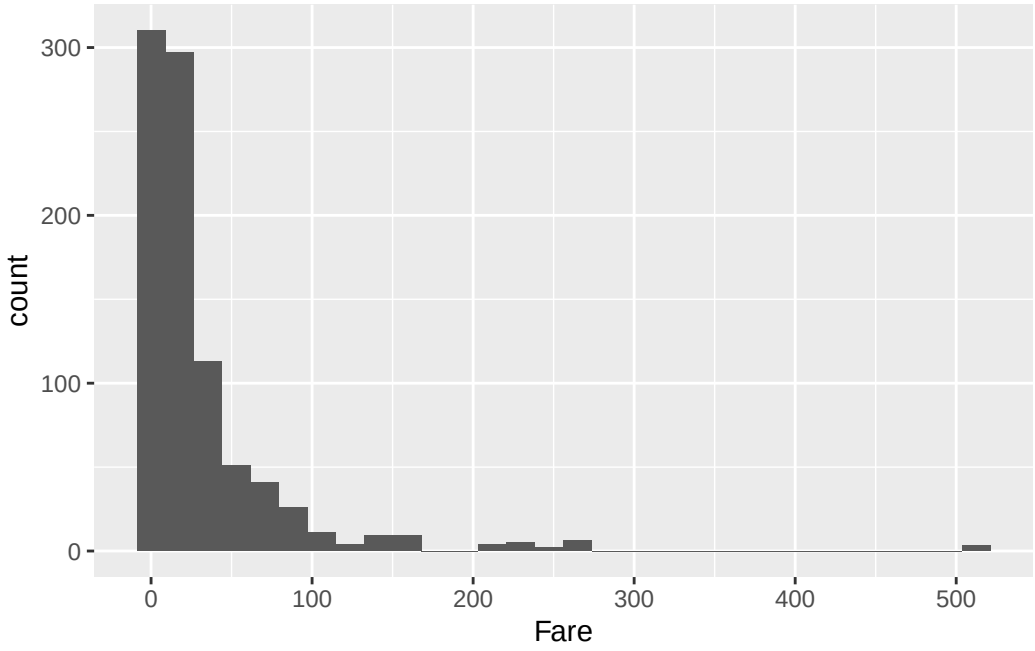



```
ggplot(test, aes(Parch)) + geom_bar()
```



```
#Fare
ggplot(train, aes(Fare)) + geom_histogram()
```

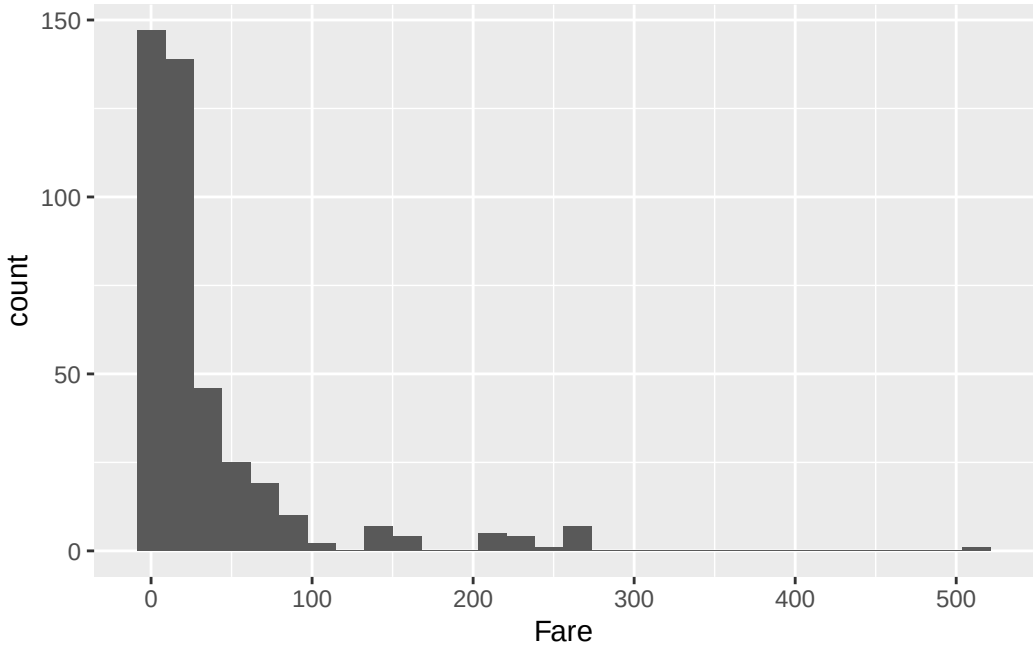
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



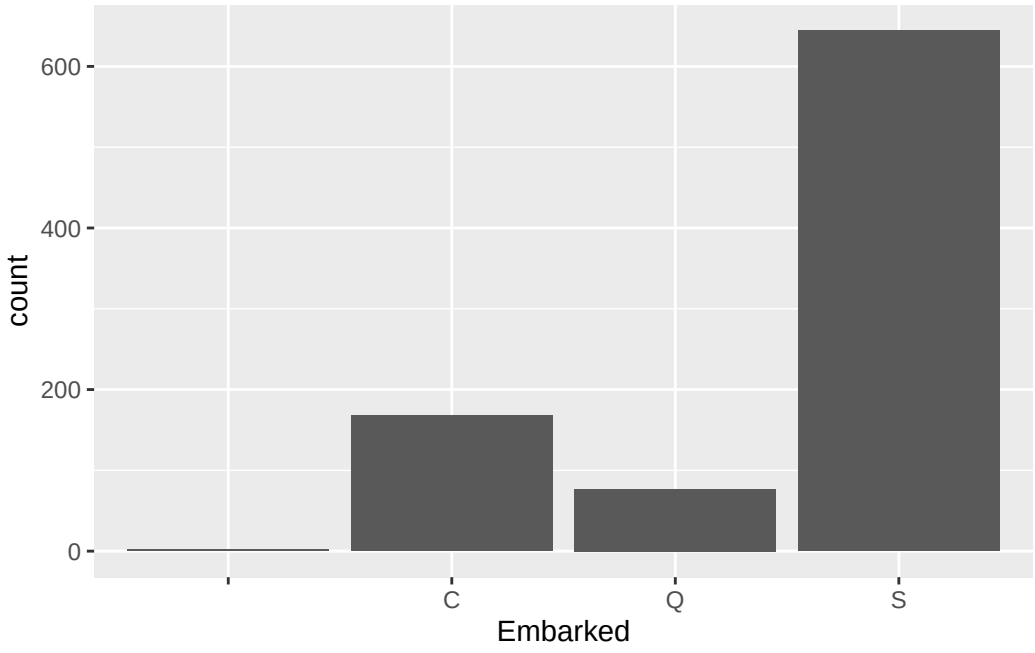

```
ggplot(test, aes(Fare)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```




```
#Embarked
ggplot(train, aes(Embarked)) + geom_bar()
```



```
ggplot(test, aes(Embarked)) + geom_bar()
```