RE6124019 HomeWork1(Titanic)

Table of contents

Import Package

```
library(Hmisc)
```

```
Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, units
```

```
library(table1)
```

```
Attaching package: 'table1'

The following objects are masked from 'package:Hmisc':

    label, label<-, units
```

```
The following objects are masked from 'package:base':

    units, units<-
```

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

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" ...
```

```
latex(describe(train), file = "", caption.placement = "top")
```

train
12 Variables 891 Observations

PassengerId

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
891	0	891	1	446	297.3	45.5	90.0	223.5	446.0	668.5	802.0	846.5

lowest : 1 2 3 4 5, highest: 887 888 889 890 891

Survived

n	missing	distinct	Info	Sum	Mean	Gmd
891	0	2	0.71	342	0.3838	0.4735

Pclass

n	missing	distinct	Info	Mean	Gmd
891	0	3	0.81	2.309	0.8631

Value	1	2	3
Frequency	216	184	491
Proportion	0.242	0.207	0.551

For the frequency table, variable is rounded to the nearest 0

Name

n	missing	distinct
891	0	891

lowest : Abbing, Mr. Anthony Abbott, Mr. Rossmore Edward Abbott, Mrs. Stanton (Rosa Hunt)
highest: Youseff, Mr. Gerious Yrois, Miss. Henriette ("Mrs Harbeck") Zabour, Miss. Hileni

Sex

n	missing	distinct
891	0	2

Value	female	male
Frequency	314	577
Proportion	0.352	0.648

Age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
714	177	88	0.999	29.7	16.21	4.00	14.00	20.12	28.00	38.00	50.00	56.00

```
lowest : 0.42 0.67 0.75 0.83 0.92, highest: 70   70.5 71   74   80
```

SibSp

n	missing	distinct	Info	Mean	Gmd
891	0	7	0.669	0.523	0.823

Value	0	1	2	3	4	5	8
Frequency	608	209	28	16	18	5	7
Proportion	0.682	0.235	0.031	0.018	0.020	0.006	0.008

For the frequency table, variable is rounded to the nearest 0

Parch

n	missing	distinct	Info	Mean	Gmd
891	0	7	0.556	0.3816	0.6259

Value	0	1	2	3	4	5	6
Frequency	678	118	80	5	4	5	1
Proportion	0.761	0.132	0.090	0.006	0.004	0.006	0.001

For the frequency table, variable is rounded to the nearest 0

Ticket

n	missing	distinct
891	0	681

```
lowest : 110152      110413      110465      110564      110813
highest: W./C. 6608  W./C. 6609  W.E.P. 5734 W/C 14208   WE/P 5735
```

Fare

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
891	0	248	1	32.2	36.78	7.225	7.550	7.910	14.454	31.000	77.958	112.079

```
lowest : 0       4.0125  5       6.2375  6.4375 , highest: 227.525 247.521 262.375 263     512.329
```

Cabin

n	missing	distinct
204	687	147

```
lowest : A10 A14 A16 A19 A20, highest: F33 F38 F4  G6  T
```

Embarked

```
   n missing distinct
 889       2        3

Value          C     Q     S
Frequency    168    77   644
Proportion 0.189 0.087 0.724
```

```
table1(~ Pclass+Sex+Age+SibSp+Parch+Fare+Embarked| Survived, data=train)
```

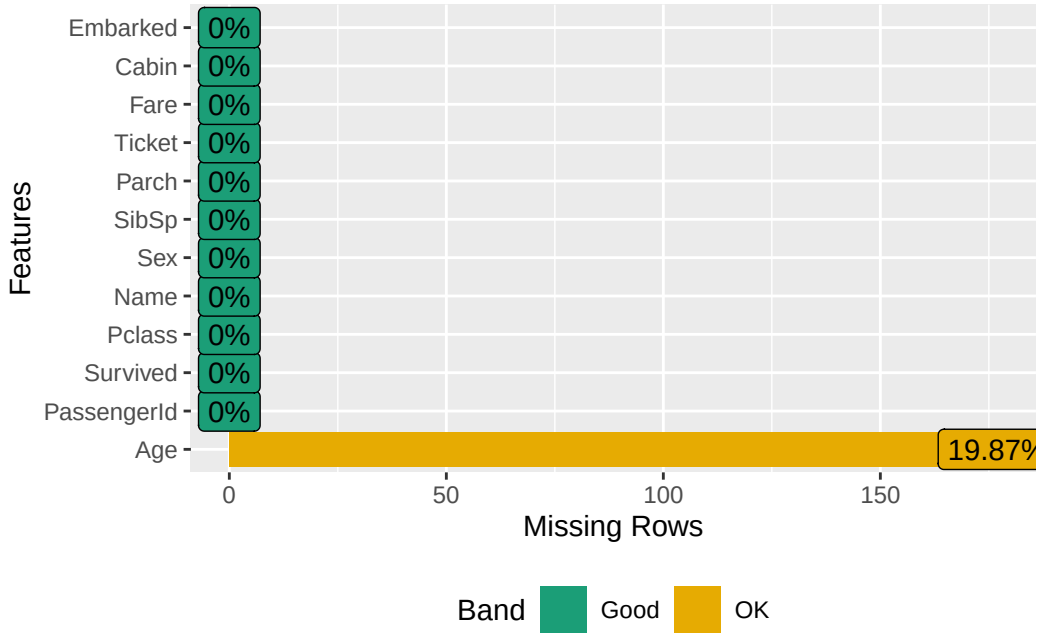
```
Warning in table1.formula(~Pclass + Sex + Age + SibSp + Parch + Fare + Embarked
| : Terms to the right of '|' in formula 'x' define table columns and are
expected to be factors with meaningful labels.

Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	0	1	Overall
	(N=549)	(N=342)	(N=891)
Pclass			
Mean (SD)	2.53 (0.736)	1.95 (0.863)	2.31 (0.836)
Median [Min, Max]	3.00 [1.00, 3.00]	2.00 [1.00, 3.00]	3.00 [1.00, 3.00]
Sex			
female	81 (14.8%)	233 (68.1%)	314 (35.2%)
male	468 (85.2%)	109 (31.9%)	577 (64.8%)
Age			
Mean (SD)	30.6 (14.2)	28.3 (15.0)	29.7 (14.5)
Median [Min, Max]	28.0 [1.00, 74.0]	28.0 [0.420, 80.0]	28.0 [0.420, 80.0]
Missing	125 (22.8%)	52 (15.2%)	177 (19.9%)
SibSp			
Mean (SD)	0.554 (1.29)	0.474 (0.709)	0.523 (1.10)
Median [Min, Max]	0 [0, 8.00]	0 [0, 4.00]	0 [0, 8.00]
Parch			
Mean (SD)	0.330 (0.823)	0.465 (0.772)	0.382 (0.806)
Median [Min, Max]	0 [0, 6.00]	0 [0, 5.00]	0 [0, 6.00]
Fare			
Mean (SD)	22.1 (31.4)	48.4 (66.6)	32.2 (49.7)
Median [Min, Max]	10.5 [0, 263]	26.0 [0, 512]	14.5 [0, 512]
Embarked			
C	75 (13.7%)	93 (27.2%)	168 (18.9%)
Q	47 (8.6%)	30 (8.8%)	77 (8.6%)

	0	1	Overall
S	427 (77.8%) 0 (0%)	217 (63.5%) 2 (0.6%)	644 (72.3%) 2 (0.2%)

```
plot_missing(train)
```



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
```

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

```
[1] 0
```

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

```
[1] 0
```

```
#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
```

```
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
```

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

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

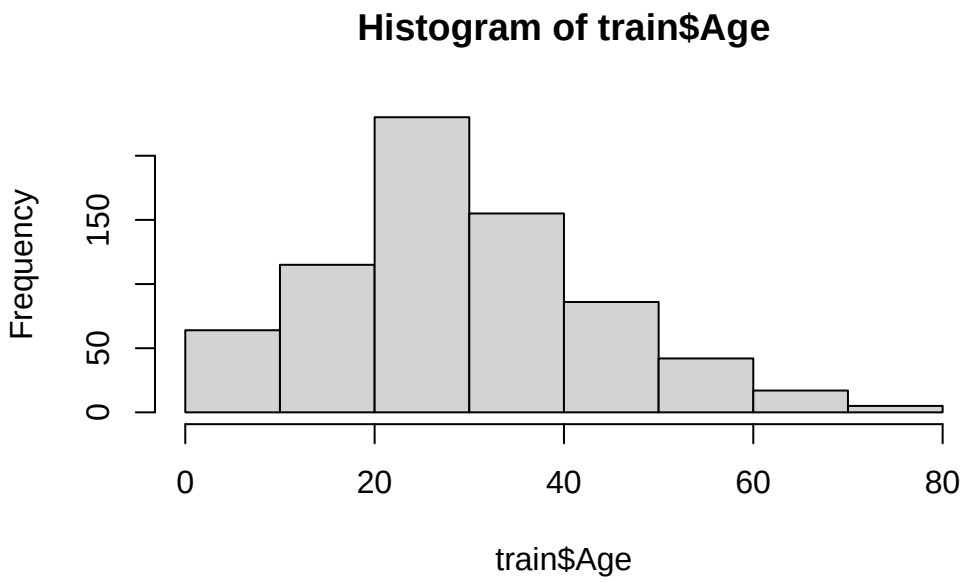
```
#Sex
table(train$Sex)
```

```
female   male
   314    577
```

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

```
female   male
   152    266
```

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




```
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)
```

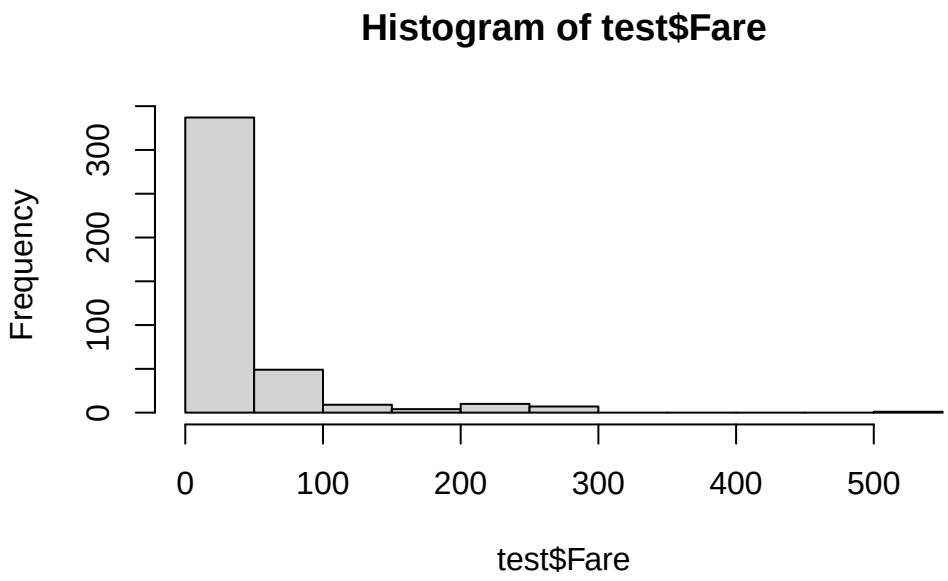
```
  0   1   2   3   4   5   6
678 118  80   5   4   5   1
```

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

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

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


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



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

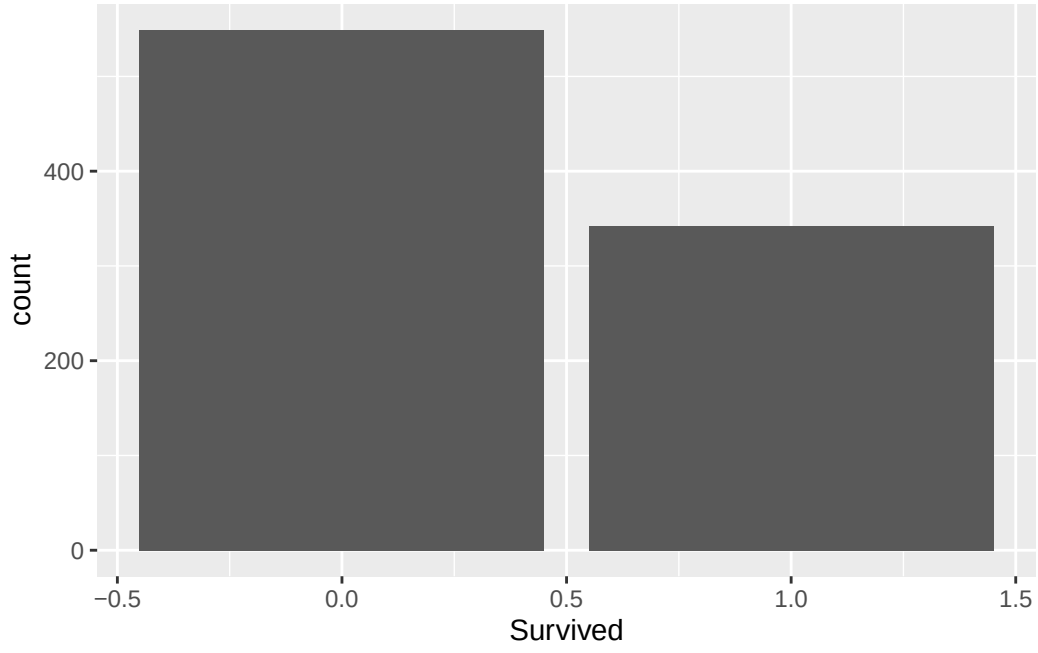
```
      C   Q   S
  2 168  77 644
```

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

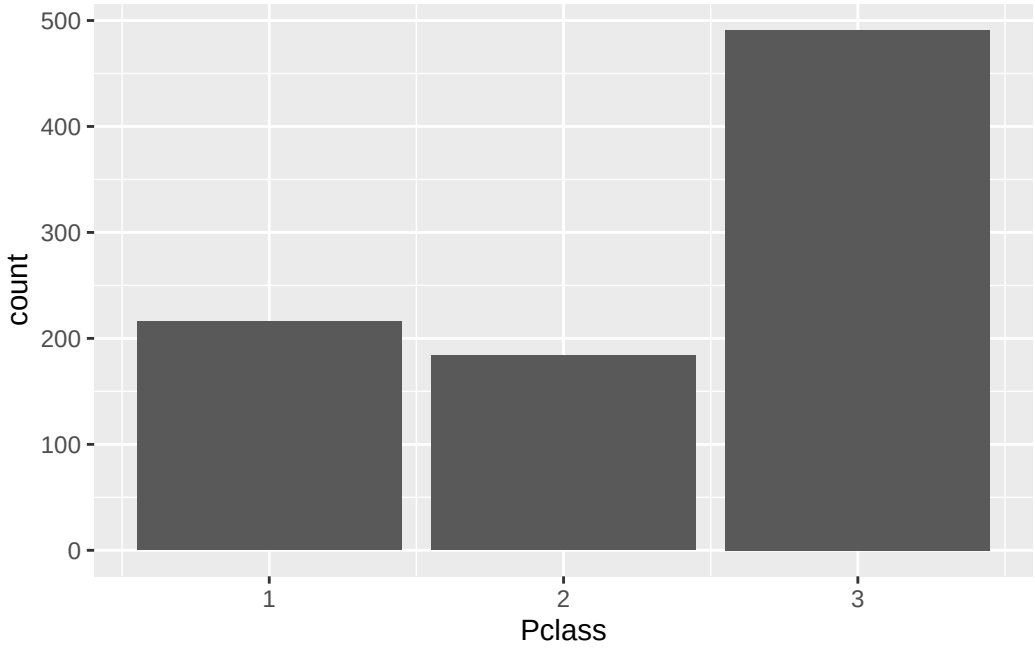
```
  C   Q   S
102  46 270
```

Plot (GGPLOT2)

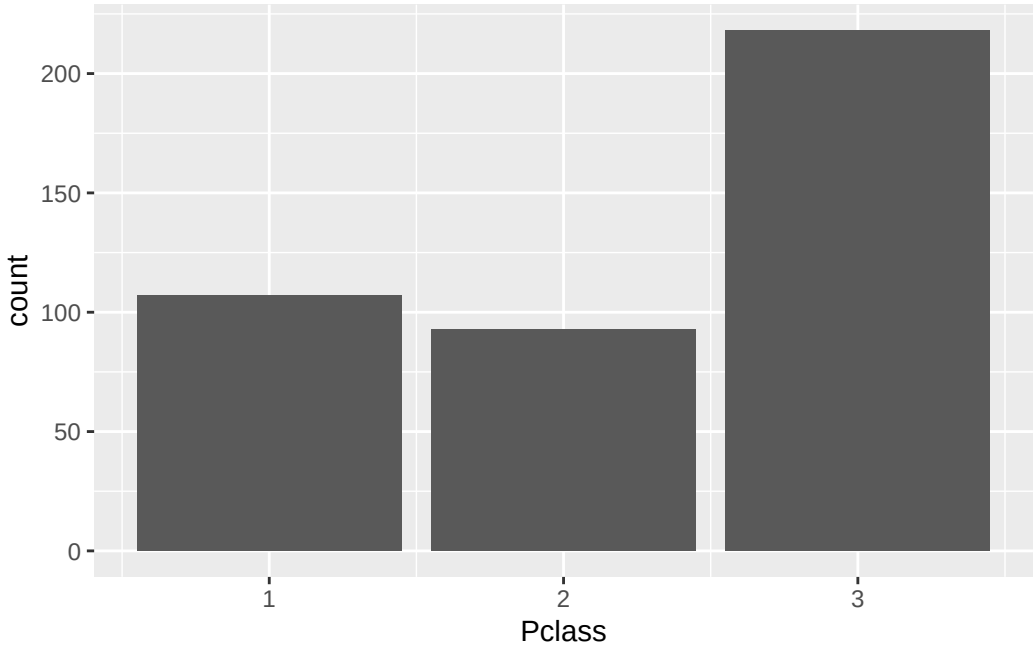
```
#Survived
ggplot(train, aes(Survived)) + geom_bar()
```



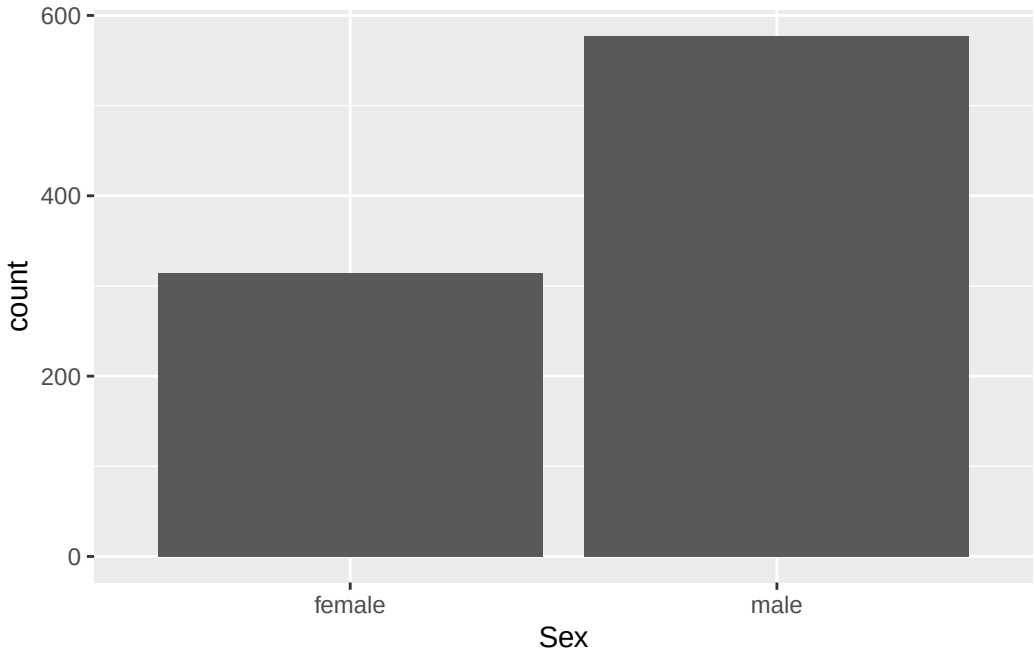
```
#Pclass
ggplot(train, aes(Pclass)) + geom_bar()
```



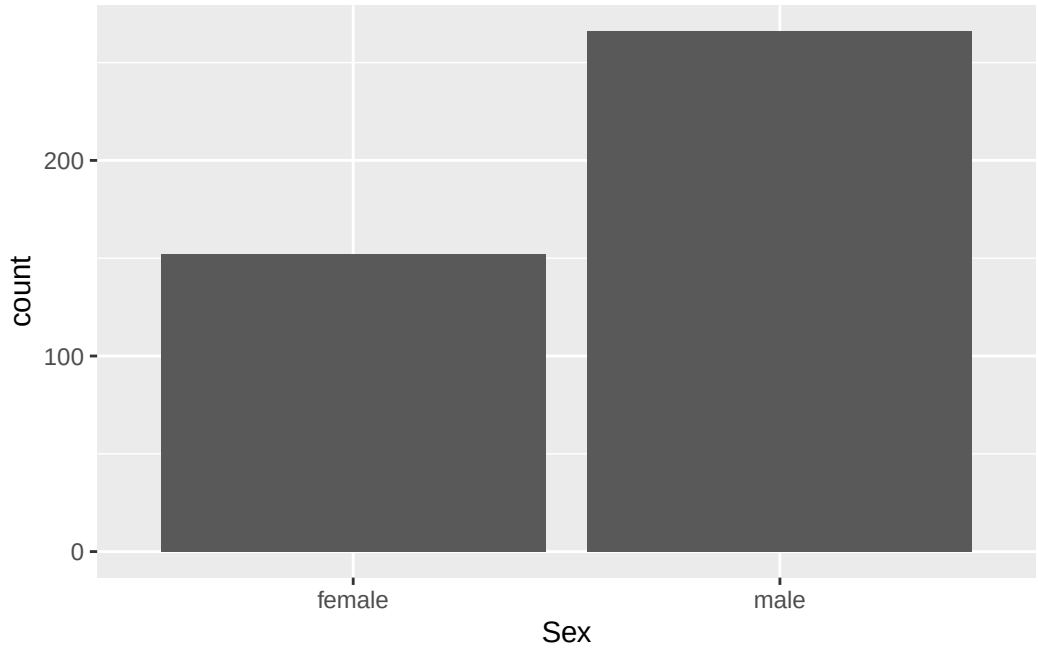

```
ggplot(test, aes(Pclass)) + geom_bar()
```




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



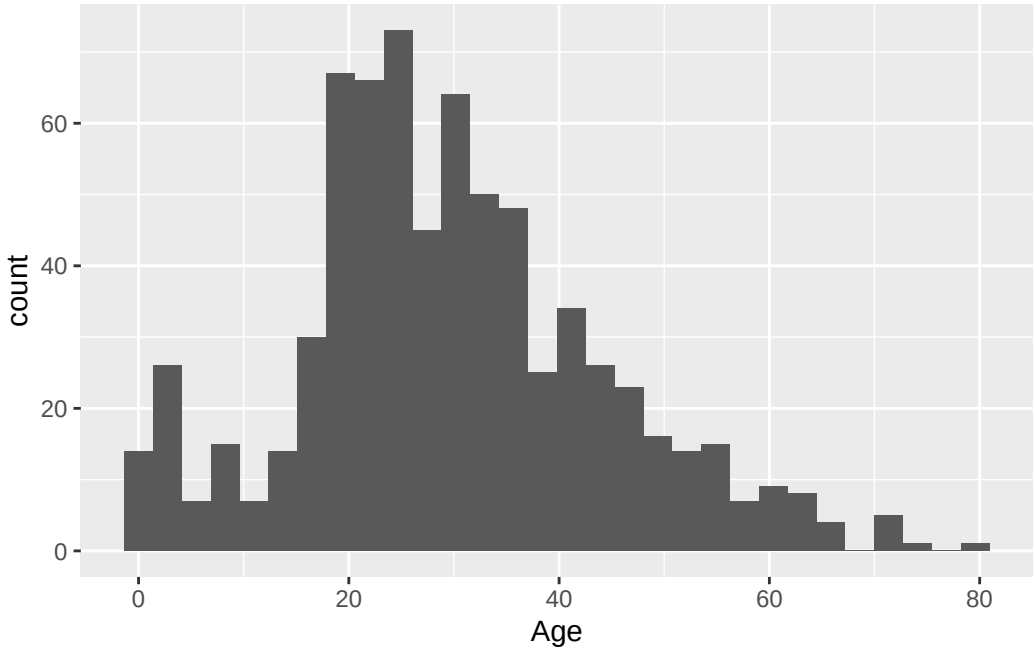
```
ggplot(test, aes(Sex)) + geom_bar()
```




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

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

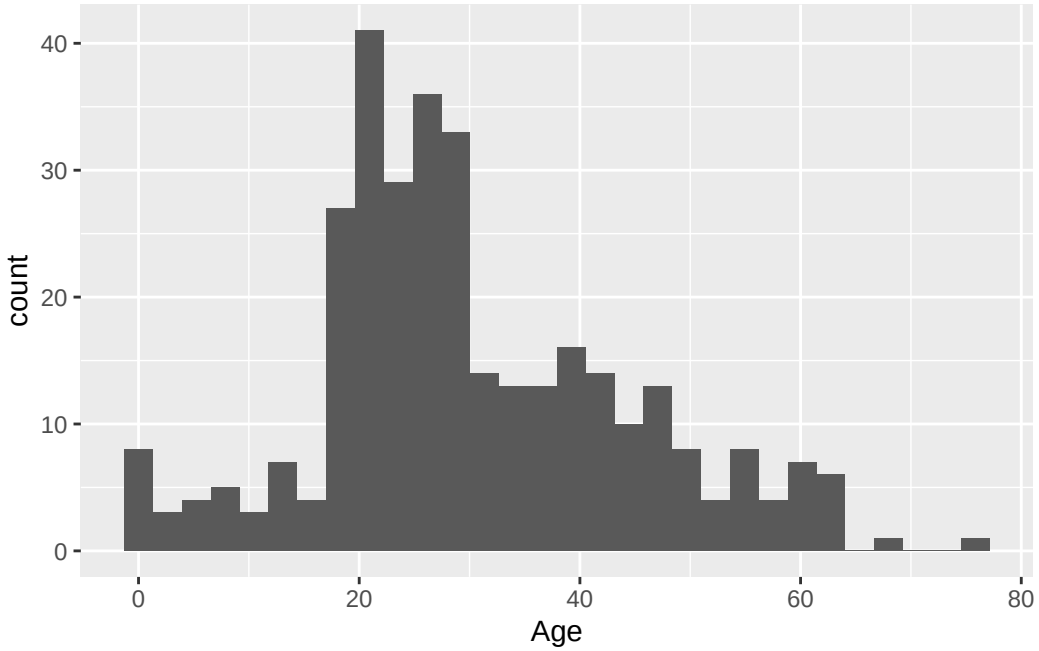
```
Warning: Removed 177 rows containing non-finite values (`stat_bin()`).
```



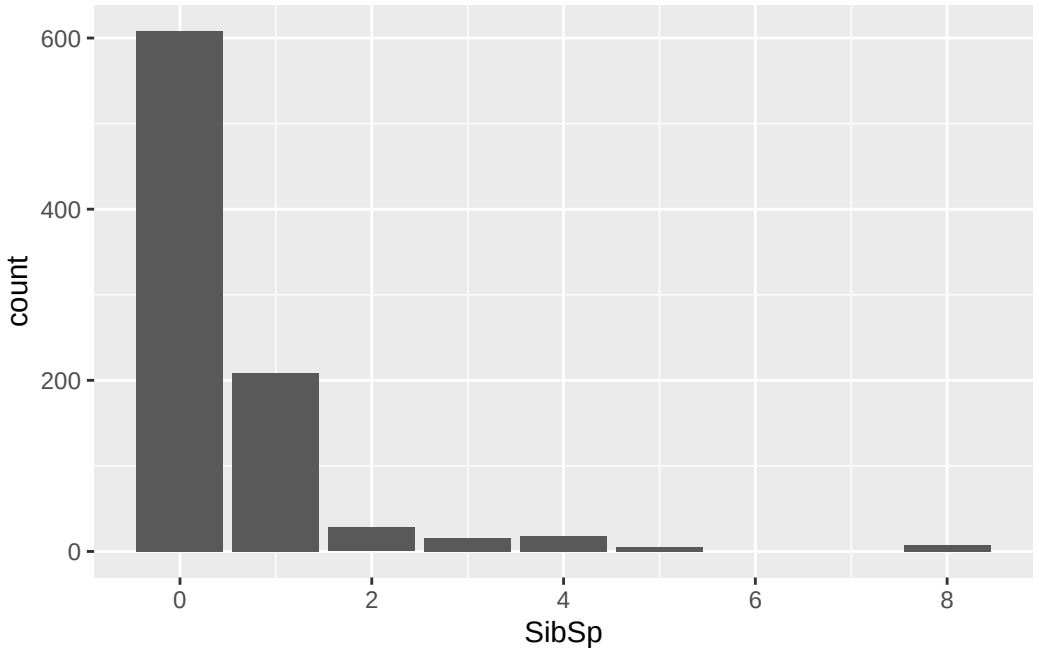

```
ggplot(test, aes(Age)) + geom_histogram()
```

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

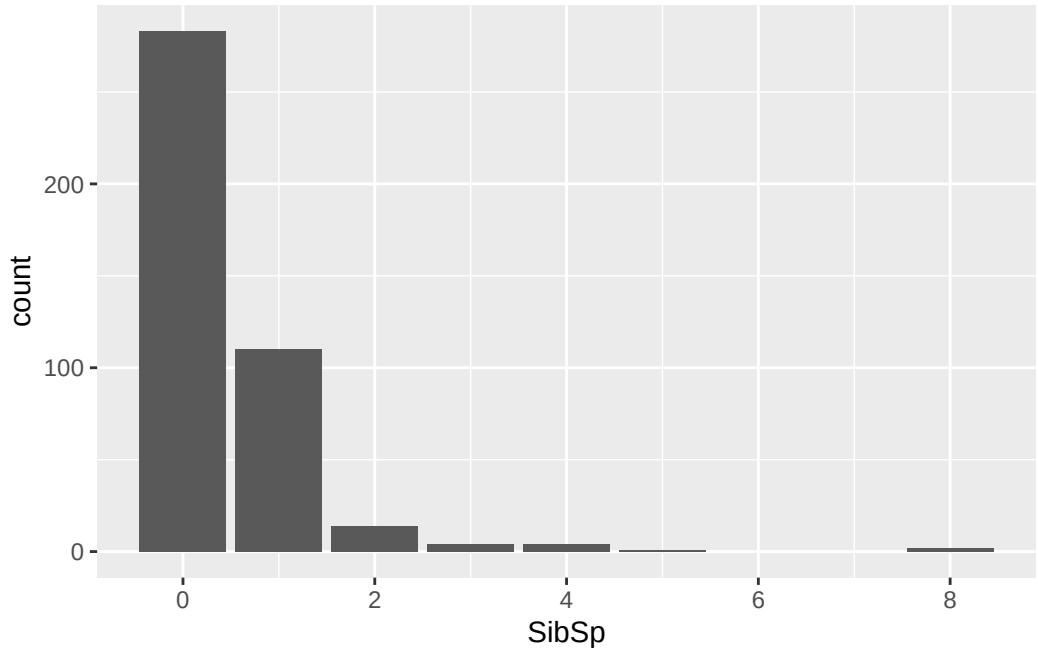
```
Warning: Removed 86 rows containing non-finite values (`stat_bin()`).
```



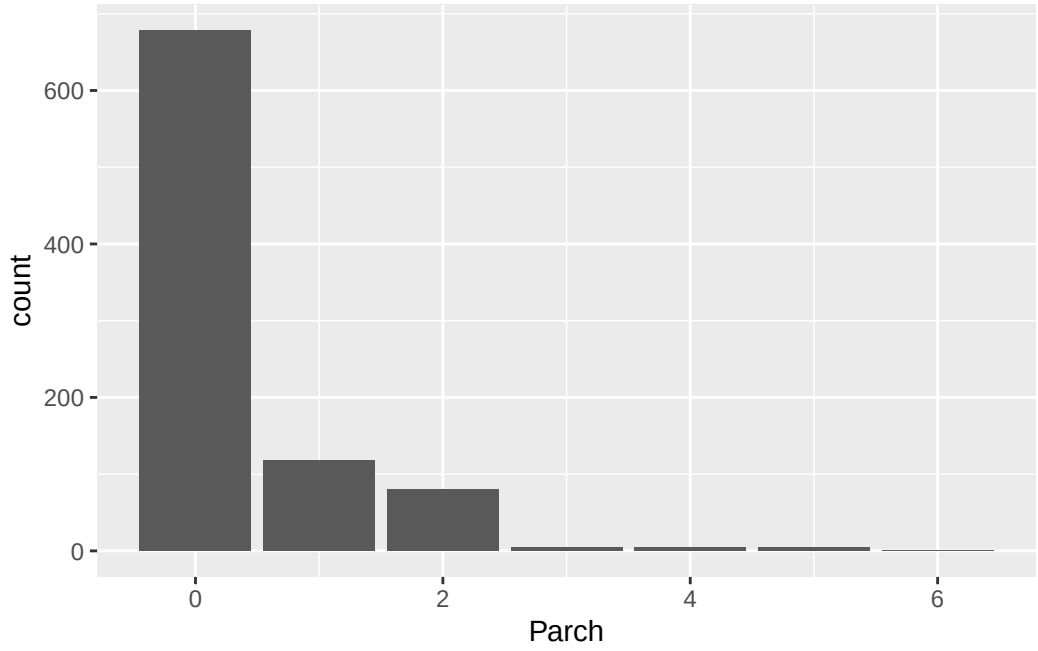

```
#SibSp
ggplot(train, aes(SibSp)) + geom_bar()
```



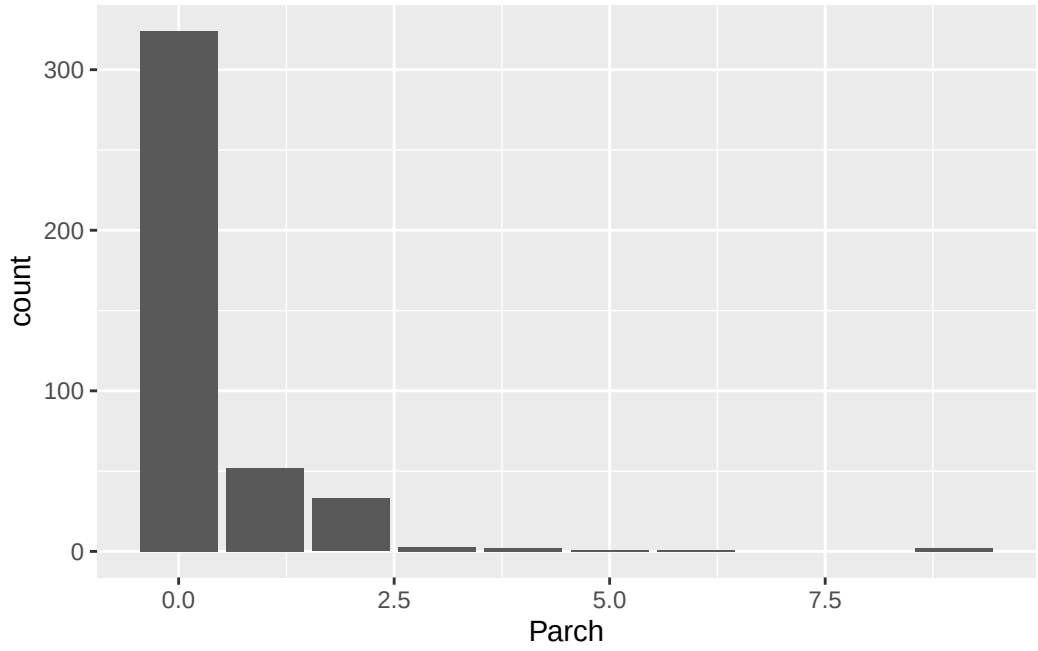
```
ggplot(test, aes(SibSp)) + geom_bar()
```




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

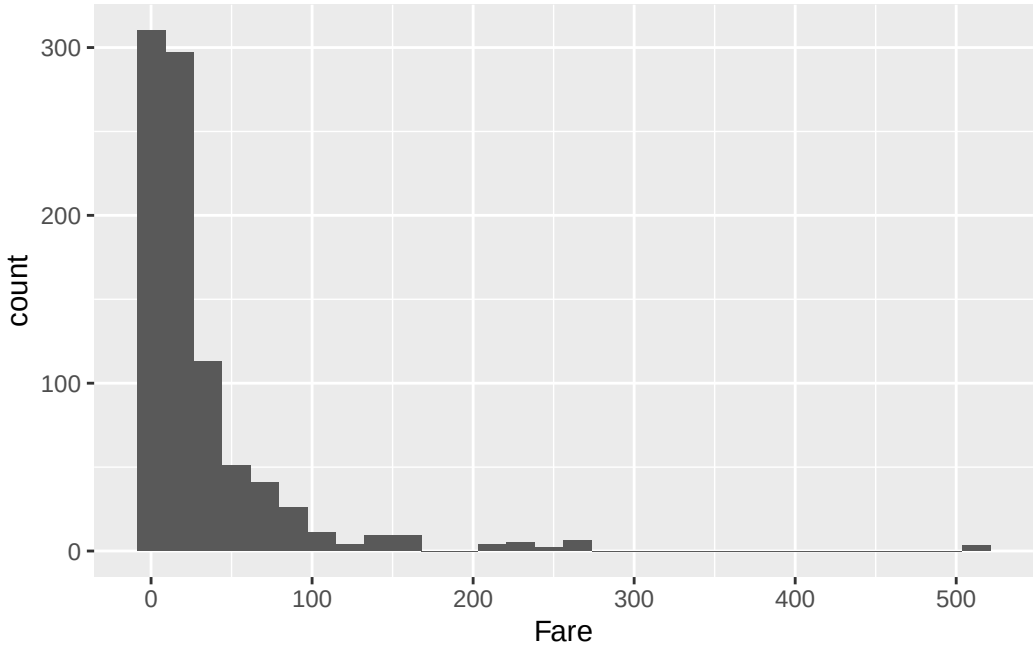



```
ggplot(test, aes(Parch)) + geom_bar()
```



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

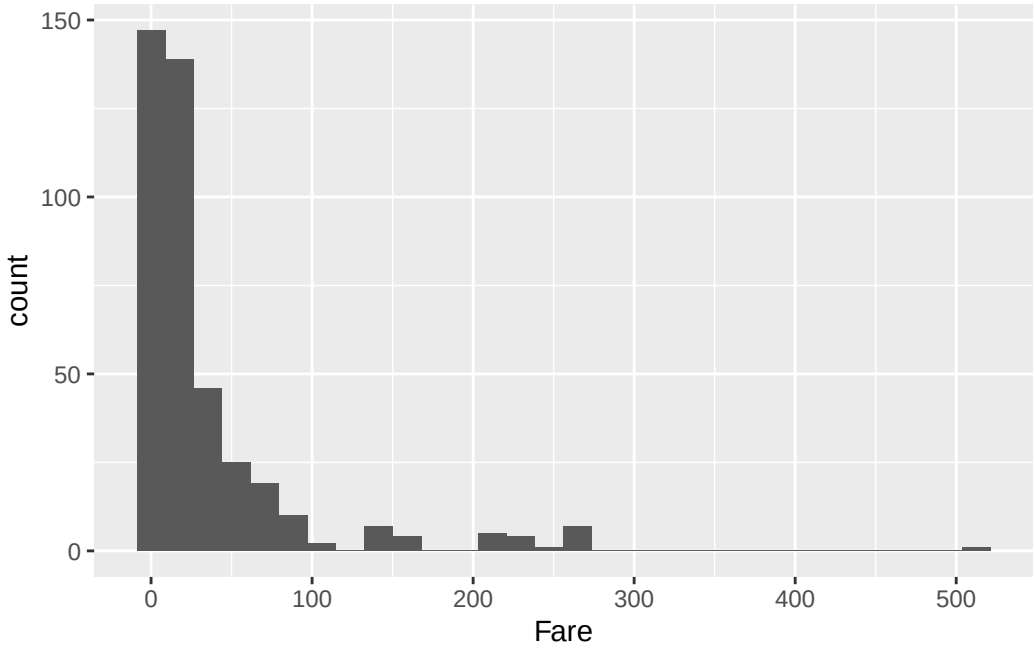
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



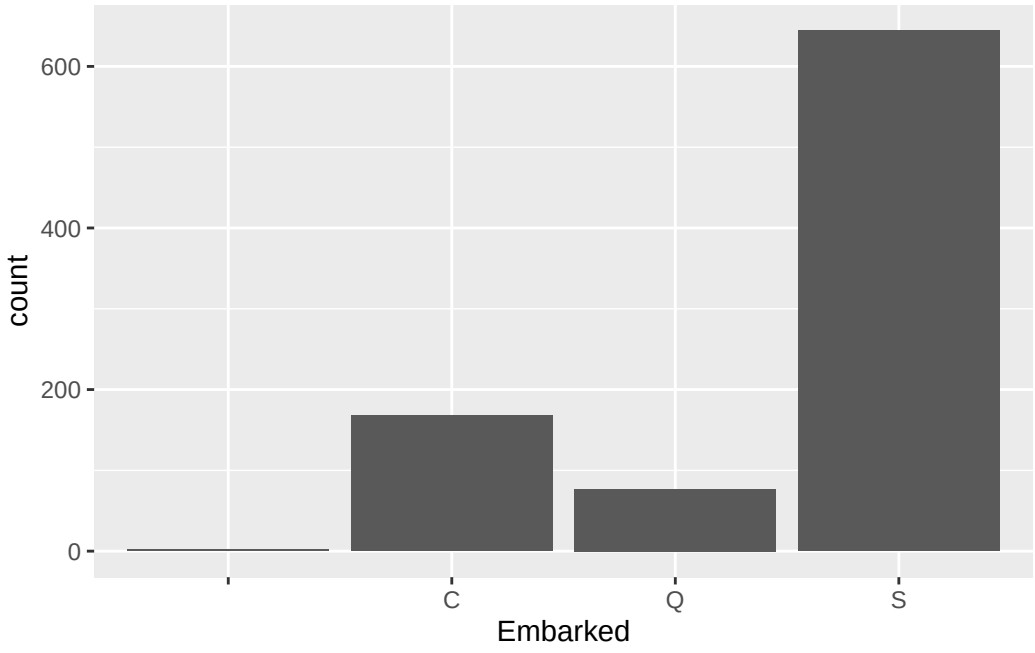

```
ggplot(test, aes(Fare)) + geom_histogram()
```

```
`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()
```