HW1

Carrie

```
library(Hmisc)
```

```
Warning: package 'Hmisc' was built under R version 4.4.2
```

```

Attaching package: 'Hmisc'

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

    format.pval, units
```

```
library(vtable)
```

```
Warning: package 'vtable' was built under R version 4.4.2

Loading required package: kableExtra

Warning: package 'kableExtra' was built under R version 4.4.2
```

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.2
```

```
df = read.csv("titanic.csv")
```

The Titanic dataset contains the details of 891 passengers on board with 12 variables, which are as follows

1.PassengerId : 1 - 891.
2.Survived : “0” respresens not survived,“1” represents survived.
3.Pclass : 1 - 3
4.Name : It’s characters.
5.Sex : “female” “male”
6.Age : 0.42 - 80
7.SibSp : 0 - 8
8.Parch : 0 - 6
9.Ticket :It’s characters.
10.Fare : 0 - 512.33
11.Cabin :It’s characters.
12.Embarked : “S”,“C”,“Q”

From the table ,we can know that:

- There are 177 missing in variable “Age” and lots of missing value in variable “Cabin”.
- There are some odd value, such as A/5 21171,PC 17599 that needs to be well-defined in the future.

```
describe(df)
```

```
df

 12  Variables      891  Observations
--------------------------------------------------------------------------------
PassengerId
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     891        0      891        1      446      446    297.3     45.5
     .10      .25      .50      .75      .90      .95
    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
     891        0        2     0.71      342   0.3838

--------------------------------------------------------------------------------
Pclass
```

```
       n  missing distinct     Info     Mean  pMedian      Gmd
     891        0        3     0.81    2.309      2.5   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
highest: Youseff, Mr. Gerious                   Yrois, Miss. Henriette ("Mrs Harbeck") Zabour
--------------------------------------------------------------------------------
Sex
       n  missing distinct
     891        0        2

Value      female   male
Frequency     314    577
Proportion  0.352  0.648
--------------------------------------------------------------------------------
Age
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     714      177       88    0.999     29.7       29    16.21     4.00
     .10      .25      .50      .75      .90      .95
   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  pMedian      Gmd
     891        0        7    0.669    0.523      0.5    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  pMedian      Gmd
     891        0        7    0.556   0.3816        0   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  pMedian      Gmd      .05
     891        0      248        1     32.2     19.6    36.78    7.225
     .10      .25      .50      .75      .90      .95
   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
--------------------------------------------------------------------------------
```

```
st(df)
```

Table 1: Summary Statistics

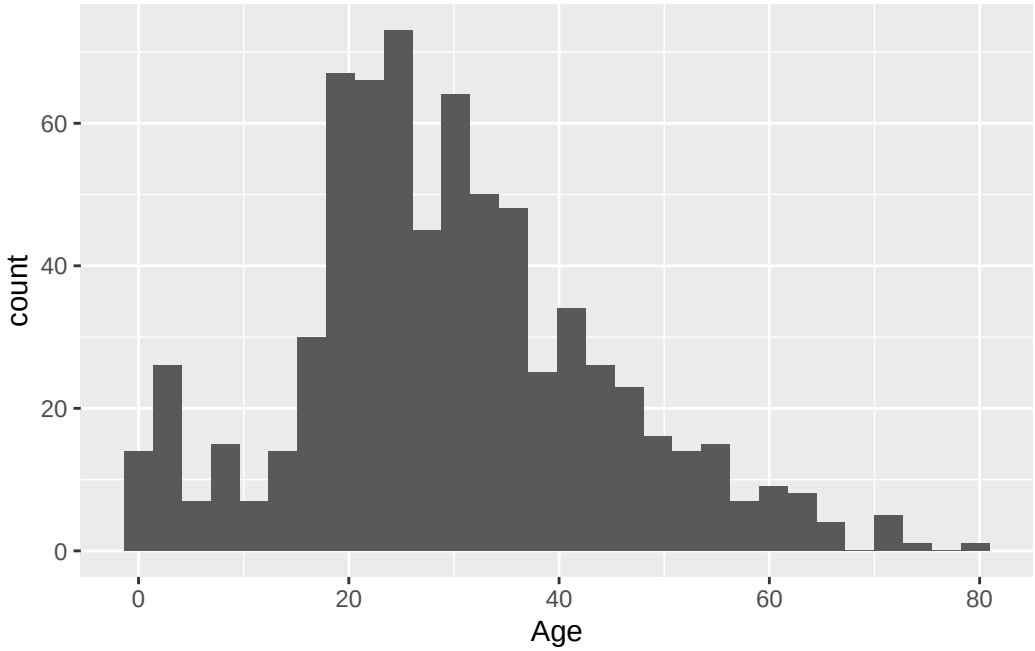
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
PassengerId	891	446	257	1	224	668	891
Survived	891	0.38	0.49	0	0	1	1
Pclass	891	2.3	0.84	1	2	3	3
Sex	891						
... female	314	35%					
... male	577	65%					
Age	714	30	15	0.42	20	38	80
SibSp	891	0.52	1.1	0	0	1	8
Parch	891	0.38	0.81	0	0	0	6
Fare	891	32	50	0	7.9	31	512
Embarked	891						
...	2	0%					
... C	168	19%					
... Q	77	9%					
... S	644	72%					

From the histgram ,we can know that most poeple's age is between 20-40.

```
ggplot(df,aes(x = Age)) + geom_histogram()
```

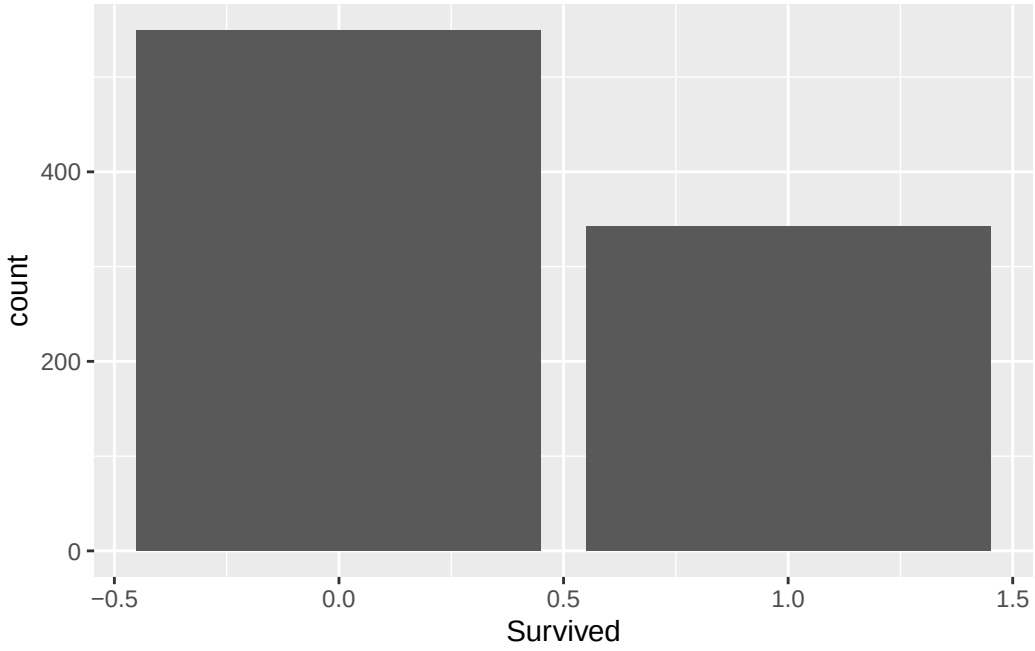
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 177 rows containing non-finite outside the scale range
(`stat_bin()`).
```



And from the barchart, we can know that survuved people is less than the not-survuved ones.

```
ggplot(data = df) + geom_bar(mapping = aes(x = Survived))
```


400
200
0
count
−0.5
0.0
0.5
1.0
1.5
Survived