STATISTICAL CONSULTING HW1

R26131086

Sheng-Yue Chang

2025-02-22

Table of contents

1. Data checking and fixing

```
datatit <- read.csv("C:/Users/r2613/Rstudio/StatCons_hw/HW1/titanic.csv")


library(Hmisc)

latex(describe(datatit),file="")
```

datatit
12 Variables 891 Observations

PassengerId 

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
891	0	891	1	446	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
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, Mrs. Stanton (Rosa Hunt)       Abelson, Mr
highest: Youseff, Mr. Gerious                 Yrois, Miss. Henriette ("Mrs Harbeck") Zabour, Miss. Hileni                  Zabour, Mis
```

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	.10	.25	.50	.75	.90	.95
714	177	88	0.999	29.7	29	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	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	.10	.25
891	0	248	1	32.2	19.6	36.78	7.225	7.550	7.910

.50	.75	.90	.95
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

```
datatit$Sex <- as.factor(datatit$Sex)
datatit$Pclass <- as.factor(datatit$Pclass)
datatit$Age <-  as.integer(datatit$Age)
datatit$Survived <-  as.factor(datatit$Survived)

latex(describe(datatit),file="")
```

datatit
12 Variables 891 Observations

PassengerId

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
891	0	891	1	446	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
891	0	2

Value	0	1
Frequency	549	342
Proportion	0.616	0.384

Pclass

n	missing	distinct
891	0	3

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

Name

n	missing	distinct
891	0	891

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

---

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	.10	.25	.50	.75	.90	.95
714	177	71	0.999	29.68	29	16.22	4	14	20	28	38	50	56

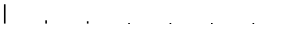
```
lowest :  0  1  2  3  4, highest: 66 70 71 74 80
```

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	.10	.25
891	0	248	1	32.2	19.6	36.78	7.225	7.550	7.910

.50	.75	.90	.95
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
```

```r
summary(datatit)
```

```
  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
```

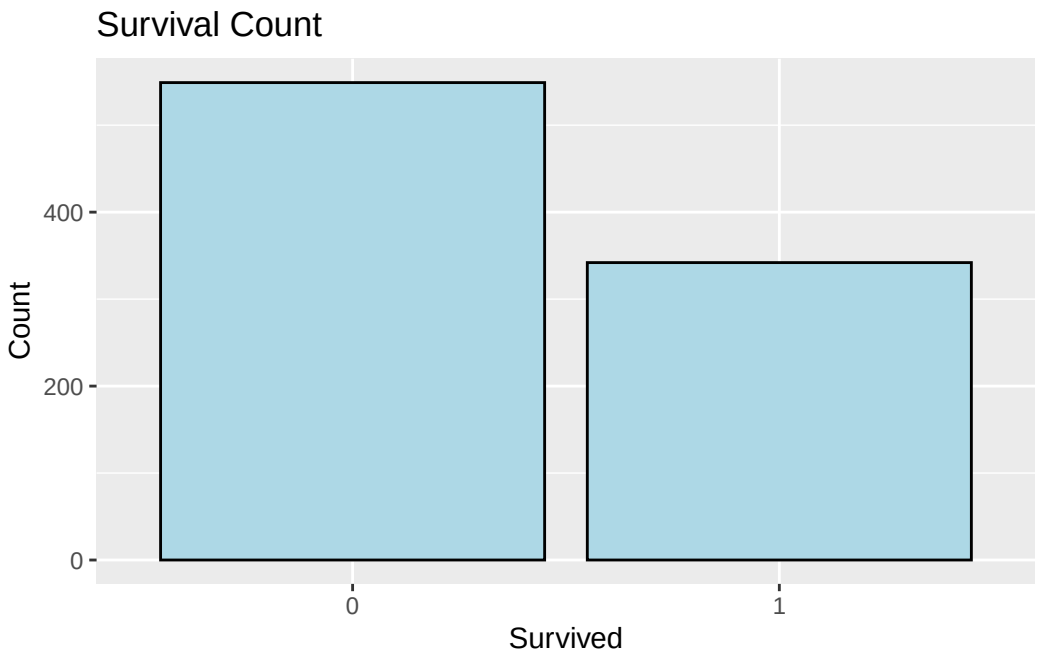
```
Median :446.0            3:491   Mode  :character
Mean   :446.0
3rd Qu.:668.5
Max.   :891.0

     Age            SibSp           Parch           Ticket
Min.   : 0.00   Min.   :0.000   Min.   :0.0000   Length:891
1st Qu.:20.00   1st Qu.:0.000   1st Qu.:0.0000   Class :character
Median :28.00   Median :0.000   Median :0.0000   Mode  :character
Mean   :29.68   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   Length:891         Length:891
1st Qu.:  7.91   Class :character   Class :character
Median : 14.45   Mode  :character   Mode  :character
Mean   : 32.20
3rd Qu.: 31.00
Max.   :512.33
```

2. Visualize some variables

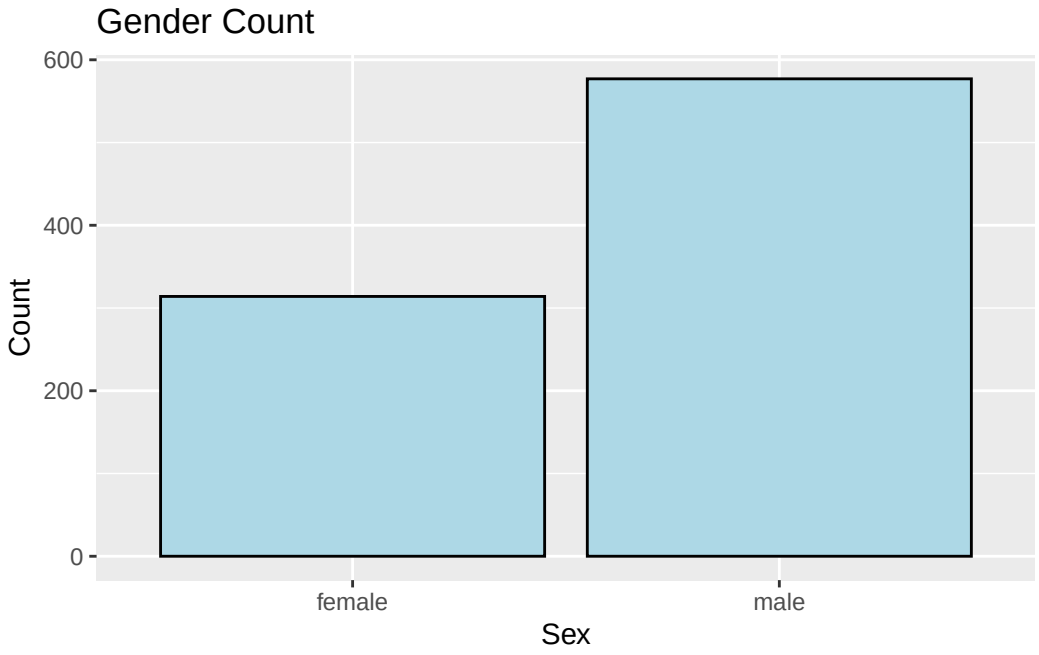
May conclude : Survived, Pclass, Sex, Age, SibSp, Parch, Fare

#(1) Survived, Pclass, Sex

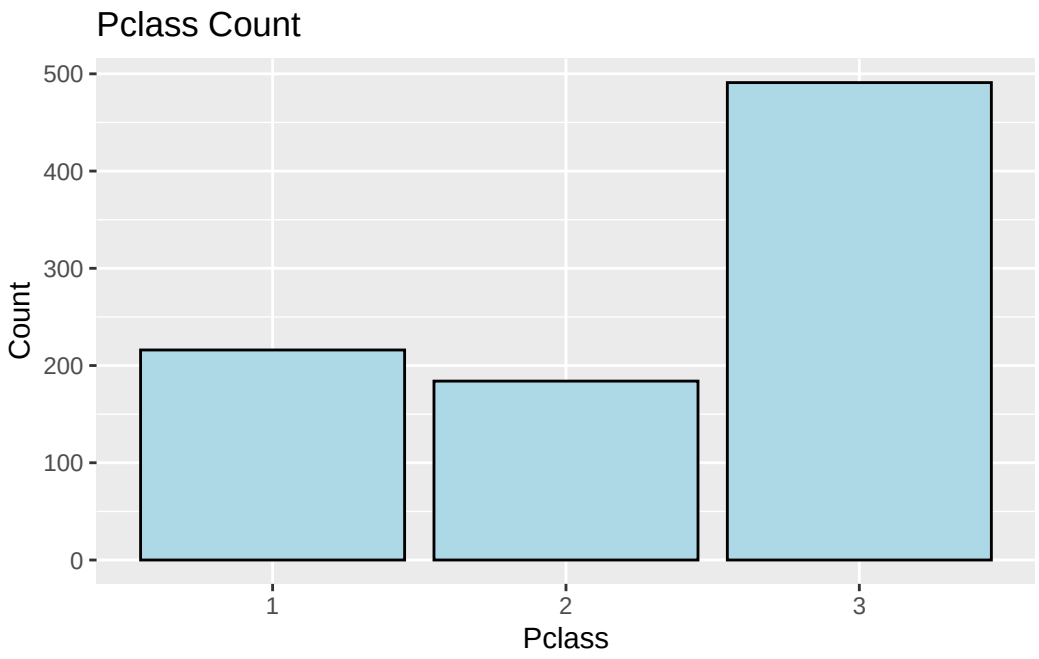
```r
library(ggplot2)

ggplot(datatit, aes(x = Survived)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(x = "Survived", y = "Count", title = "Survival Count")
```




```
ggplot(datatit, aes(x = Sex)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(x = "Sex", y = "Count", title = "Gender Count")
```

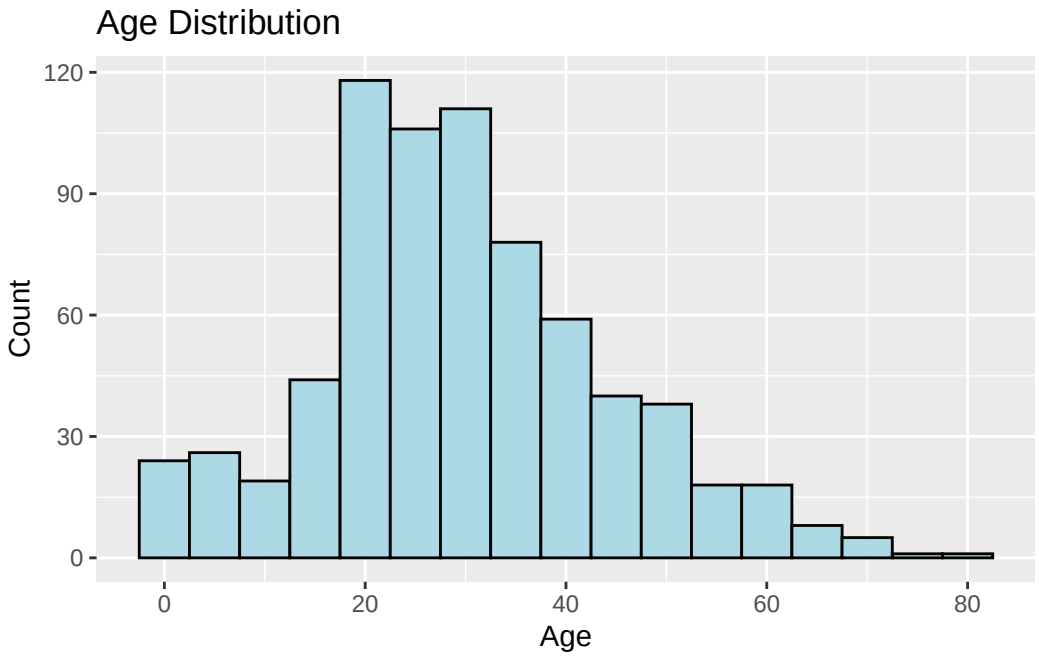



```
ggplot(datatit, aes(x = Pclass)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(x = "Pclass", y = "Count", title = "Pclass Count")
```

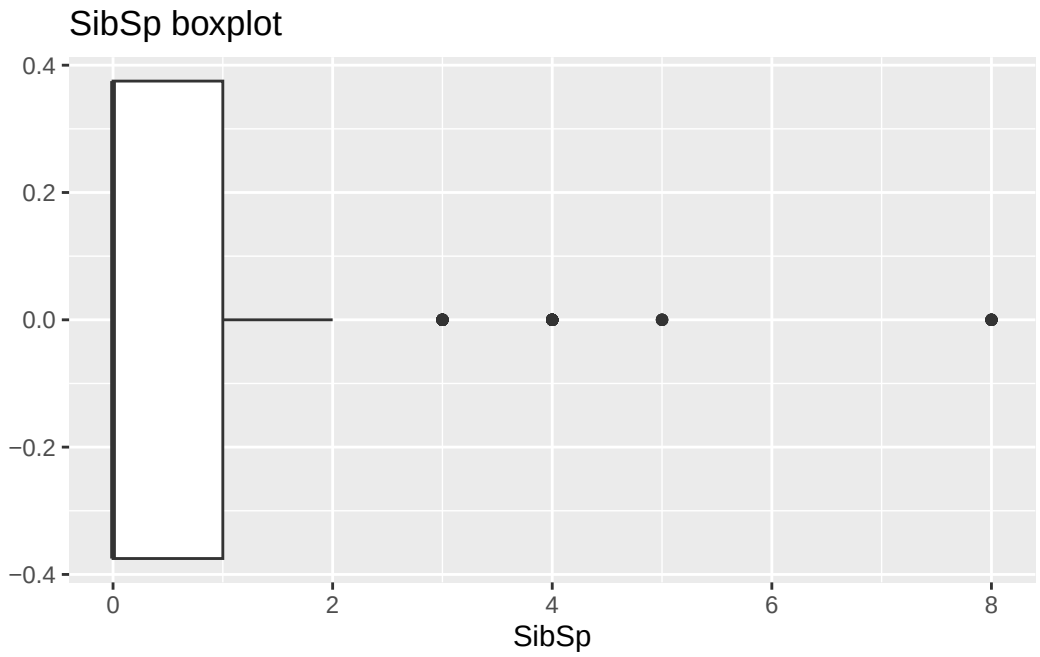


#(2) Age, SibSp, Parch, Fare

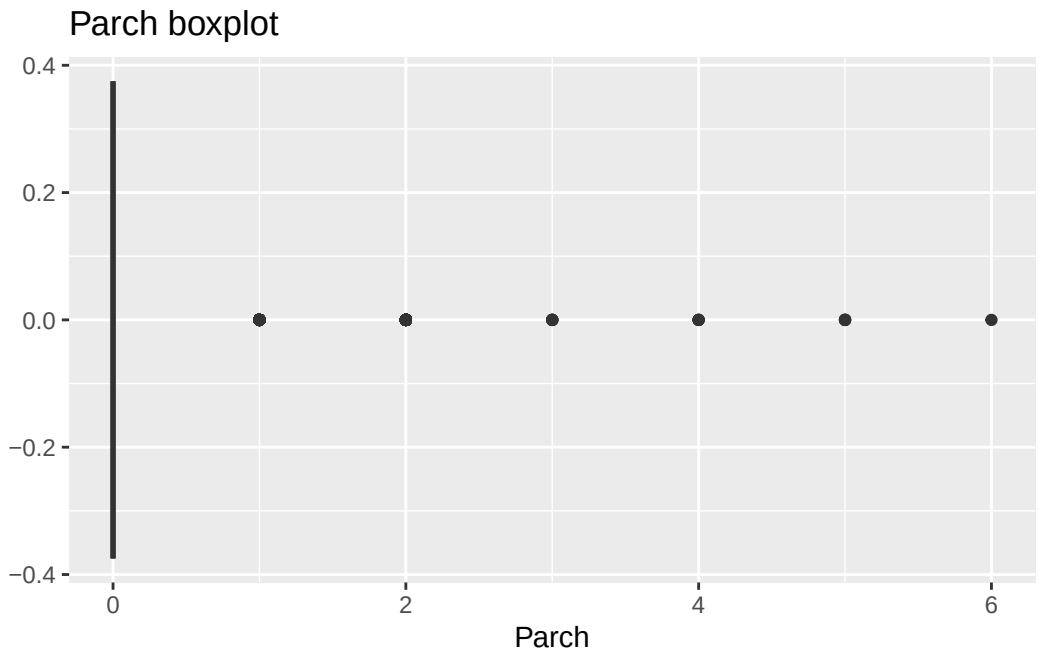
```
datatit_cleanage <- datatit[complete.cases(datatit$Age), ]

ggplot(datatit_cleanage, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(x = "Age", y = "Count", title = "Age Distribution")
```



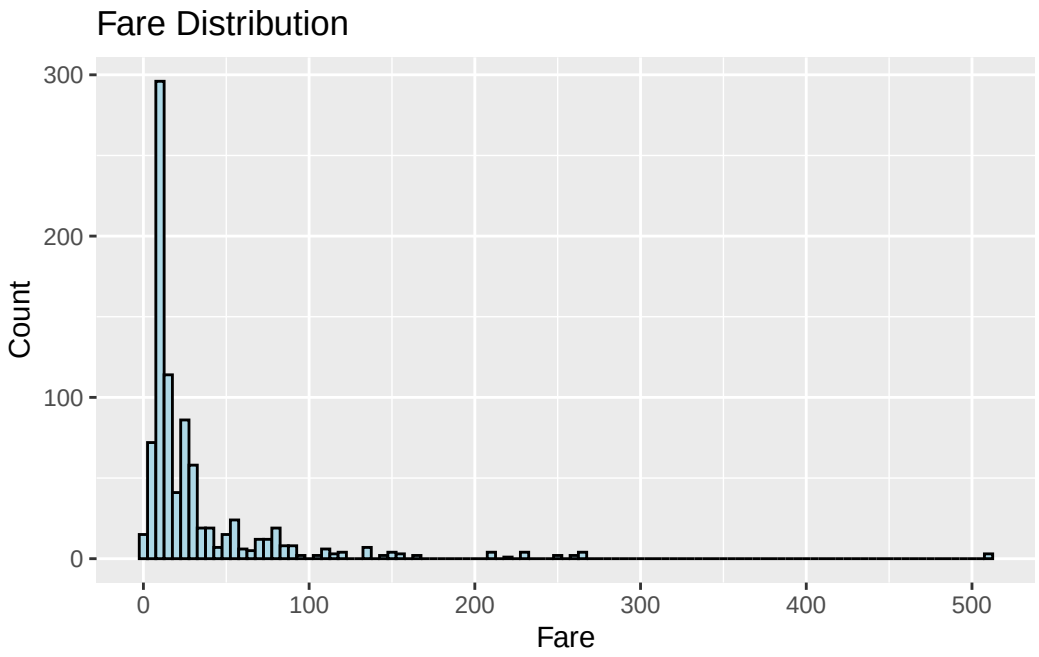
```
ggplot(datatit, aes(x = SibSp)) +
  geom_boxplot() +
  labs(x = "SibSp", title = "SibSp boxplot")
```




```
ggplot(datatit, aes(x = Parch)) +
  geom_boxplot() +
  labs(x = "Parch", title = "Parch boxplot")
```



```r
ggplot(datatit, aes(x = Fare)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(x = "Fare", y = "Count", title = "Fare Distribution")
```



3. Conlusion

(1)

- Fewer passengers survived.
- Most of the passengers are male.
- The third class has the largest number of passengers.

(2)

- There are some missing data in Age column, and most of the passengers are between 20 to 40 years old.
- Most passengers have few number of family members.
- Fares are mainly concentrated in the lower range.