ASSIGNMENT 4

Chunyao Wang(chw132),
Wenxing Li(wel85),
Jie Rong(jir18)

Table of Contents

1. Overview

This project analyzes Titanic Data (train.csv) in kaggle.com. In this assignment, we use R to program and draw 5 different Graphics (Whisker-plot, Histogram, Face grid, Violin Plot and Heatmap) to present relationships between two or more attributes. In the file, there are 12 attributes which present passenger's information. Below is the information we summarized of train.csv file:

- 891 passengers in total
- 342 survived and 549 did not;
- 314 were women and 577 were men;
- 216 strayed in 1st class, 184 in second class and 491 in third class;

As preparation, we set up ggplot2 package into RStudio and then import Titanic data.

R code:

```
data=read.csv(file = "/Users/chunyao/Desktop/train.csv",header = TRUE,sep = ",")
df=data.frame(data)
library(ggplot2)
```

2. Questions and Answers

2.1 Whisker-plot

In this part, we draw a Whisker-plot Graphic using sex and fare. Whisker plot is a diagram showing statistical distribution of a data set. This plot uses five statistics: Minimum value, Second quartile, Median value, Third quartile and Maximum value.

According to the result, it is obviously that women pay more money for the fare than men. The maximum value is almost the same between men and women. But the range of majority of women are much larger then men, and their corresponding fare is also higher than men. The cause of this result maybe the majority of women want a more comfortable place to stay. Or, it is also possible that the women in this ship are richer than the men.

R code:

```
ggplot(df,aes(factor(Sex),Fare))+geom_boxplot(outlier.colour = "red",outlier.size = 3)+ggtitle("Sex in Comparsion with the Fare")
```

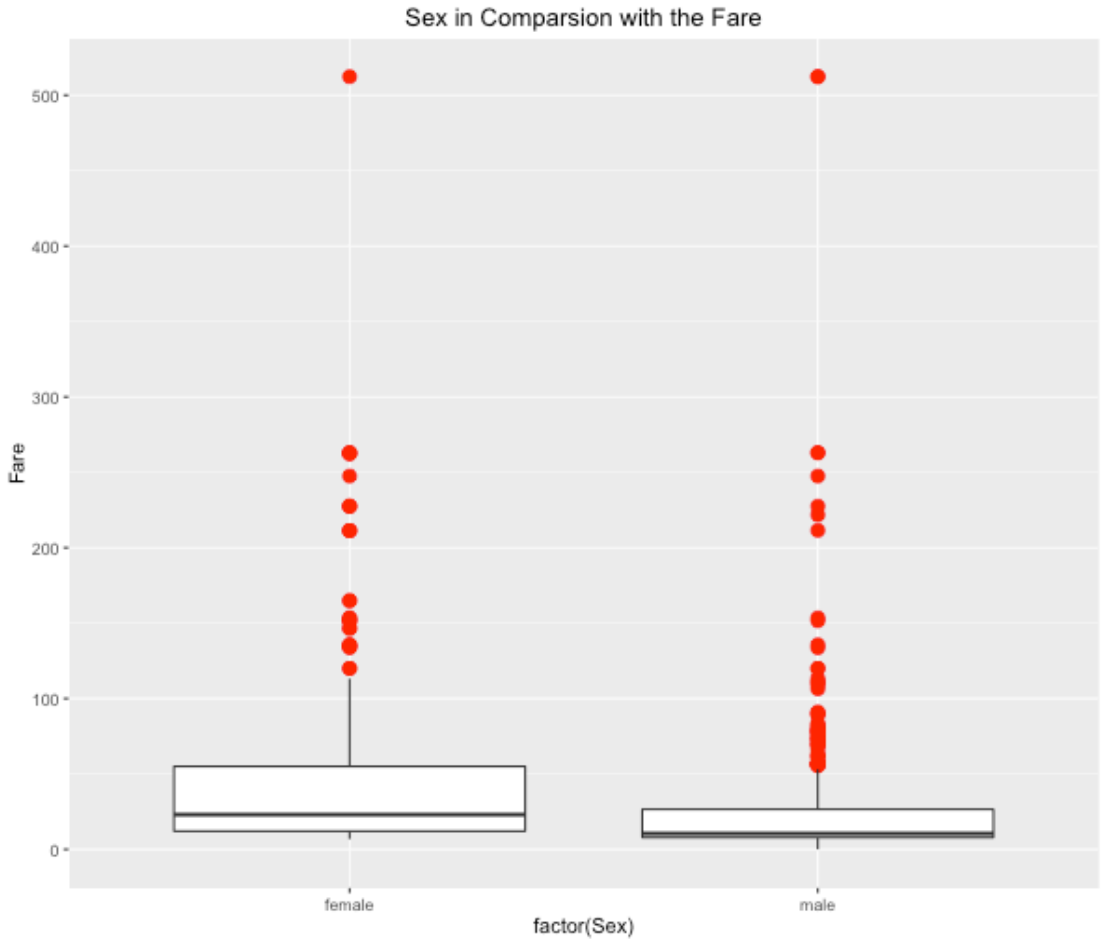

Fig2.1 Whisker-plot

2.2 Histogram

The Histogram graphically summarize the distribution of the data. In this Histogram, we want to know the how many men and women survived in Titanic. As we can see in Histogram, the total number of men in Titanic is about half more than women. However, more women survived than men, no only in proportion but also in total number (the orange bar shows they did not survive, the green bar shows they survived). The reason of this result maybe that the men have a sense of responsibility to protect women, and they left survive opportunity to women.

R code:

```
ggplot(df,aes(factor(Sex),fill=factor(Survived)))+geom_bar()+ggtitle("Sex in Survived")
```

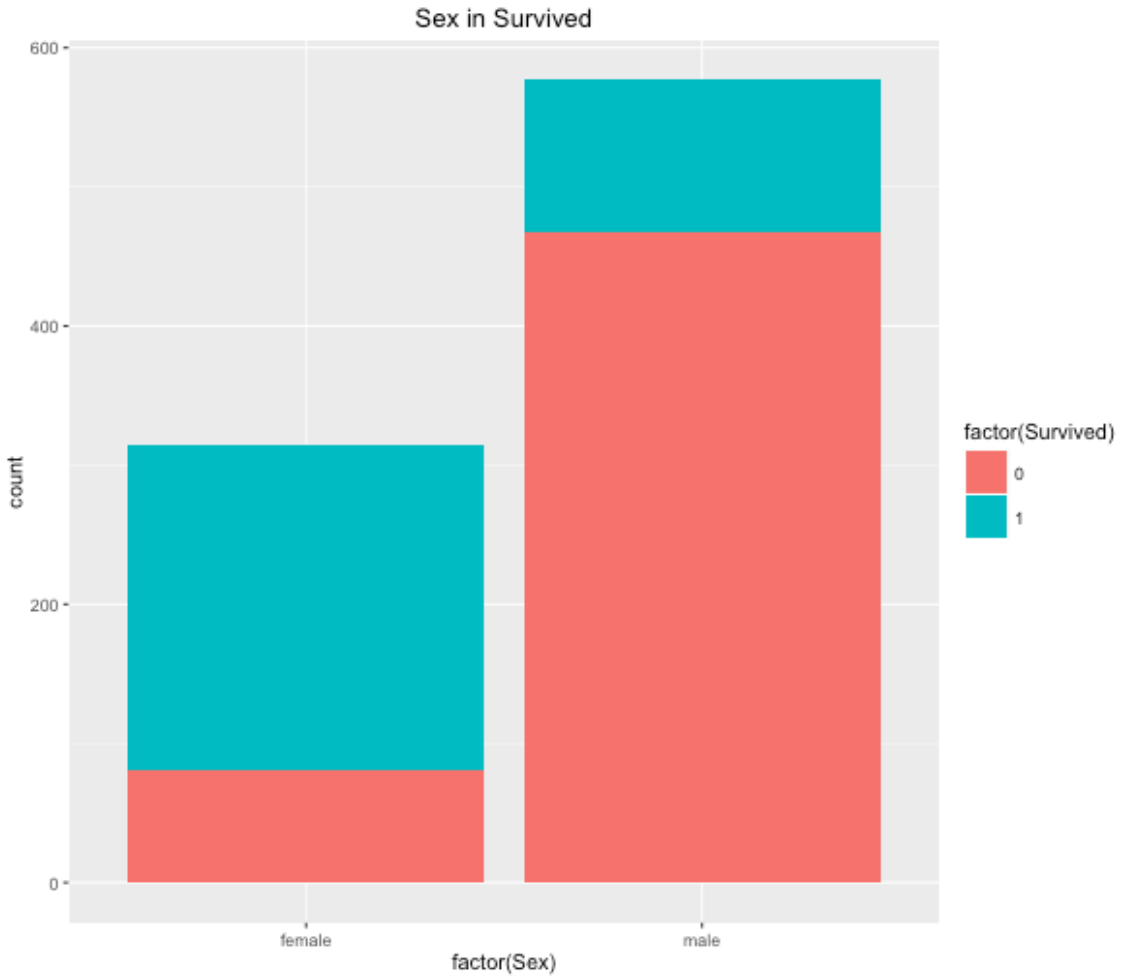

Fig2.2 Histogram

2.3 Facet grid

We use Facet grid to figure out the relation among sex, age, class and Survived or not. In this part we use four attributes in one graphic. First of all, we can see more orange dots in male line than women. This is as same as the 2.2 result, more women survived than men. Secondly, we can see the survived distribution in different class. We can see that in the first class there are the least number men and women not survived. The 2rd class died more people and the third class died most. Thirdly, we can see the youngest female died in the first class, the older died more in the second class of women. In men, the distribution that not survived in age are evenly.

```
ggplot(df,aes(Age,Sex,colour=factor(Survived)))+geom_point()+facet_grid(~Pclass)+ggtitle("In different classes, male and female at different age survived situation")
```

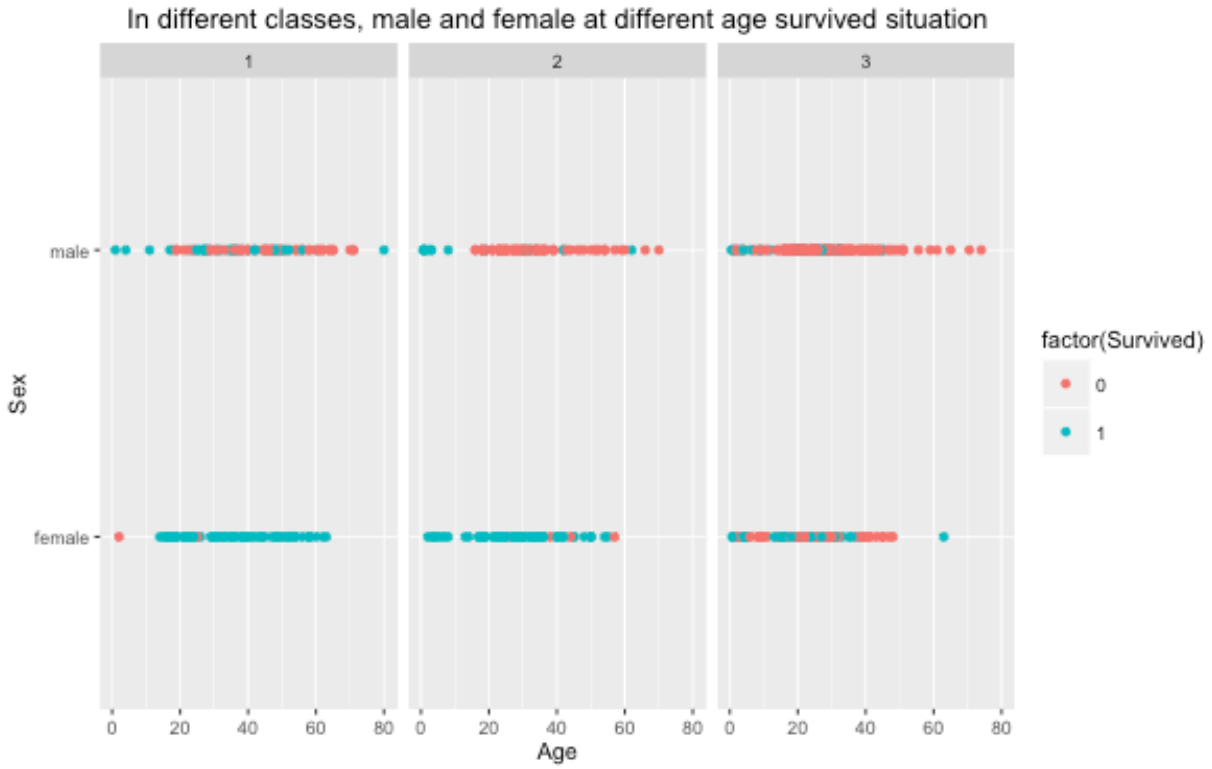

Fig2.3 Facet grid

2.4 Violin plot

Figure 2.4 shows the age distribution of saved and unsaved people based on sex. "1" represents saved people. "0" represents unsaved people. According to the figure, more people survived than those didn't. As it can be seen, male is more than female; the age range of male is wider than that of female. The violin plot indicates most people are between 20 to 40 years' old.

```
ggplot(df,aes(factor(Survived),Age,colour=factor(Survived)))+geom_violin()+ggtitle("Age distribution of saved and unsaved male and female")
```

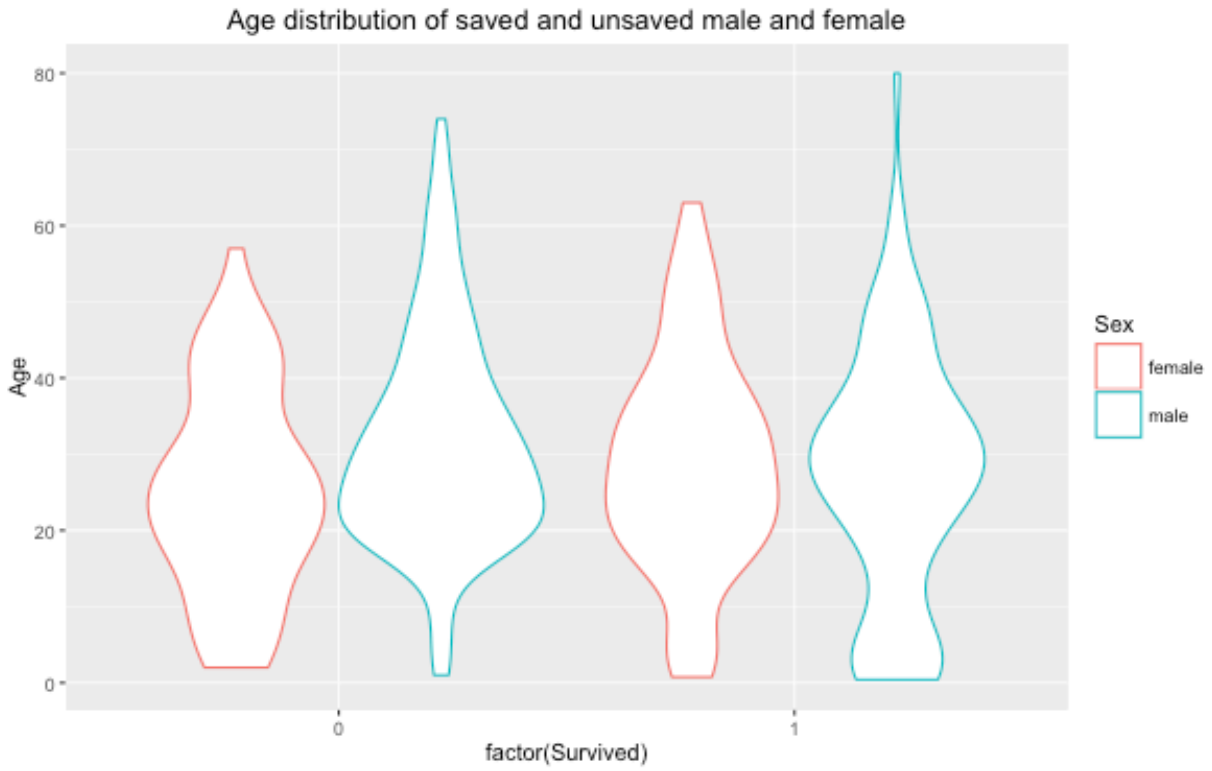

Fig2.4 Violin plot

2.5 Heatmap

Figure2.5 demonstrates the age distribution of saved and unsaved people based on the class type. According to the plot, most people belongs to class 3; Most young people belongs to class 3. As it can be seen, the percentage of saved people in class 1 ranks 1st while the percentage of saved people in class 3 is the lowest. What is more, it can be figure out that young people have higher possibility to survive than elder people.

R code:

```
ggplot(df,aes(Sex,Age))+geom_tile(aes(fill=Survived))+scale_fill_gradient(low="steel blue",high = "Red")+ggtitle("The distribution of age of saved and unsaved passengers in different classes ")
```

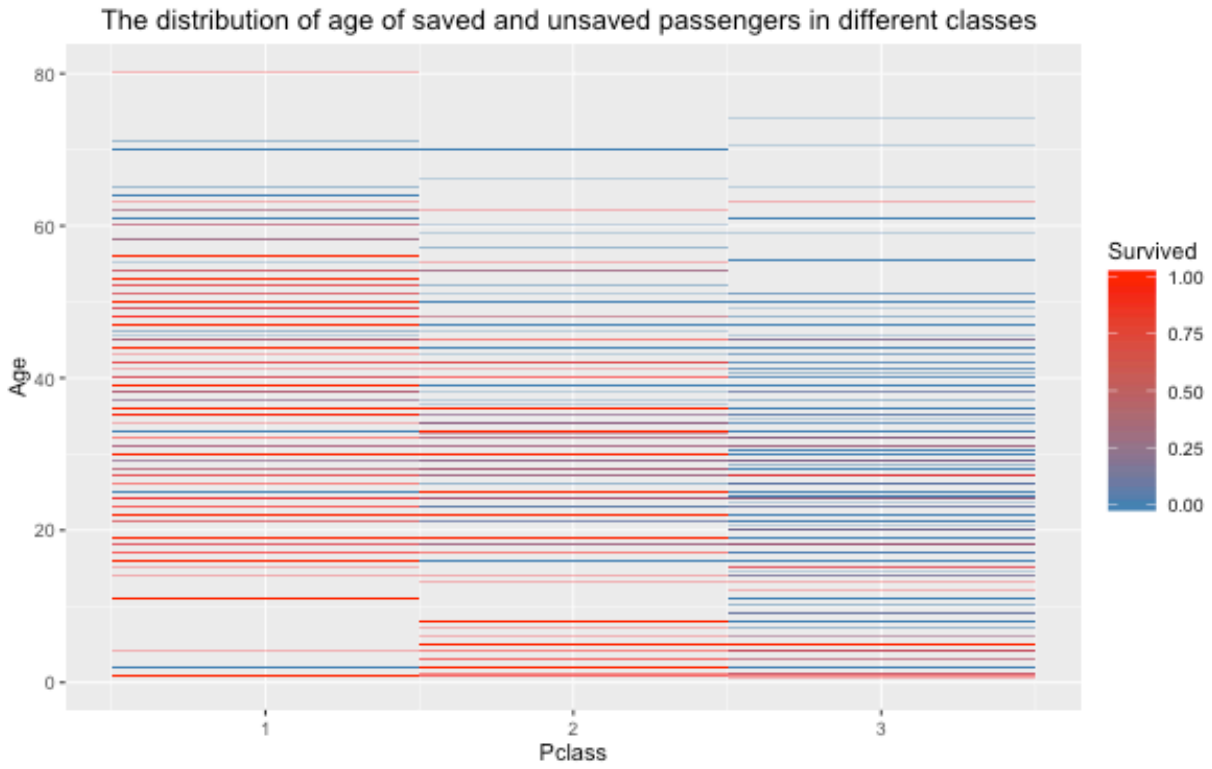

Fig2.5 Heatmap