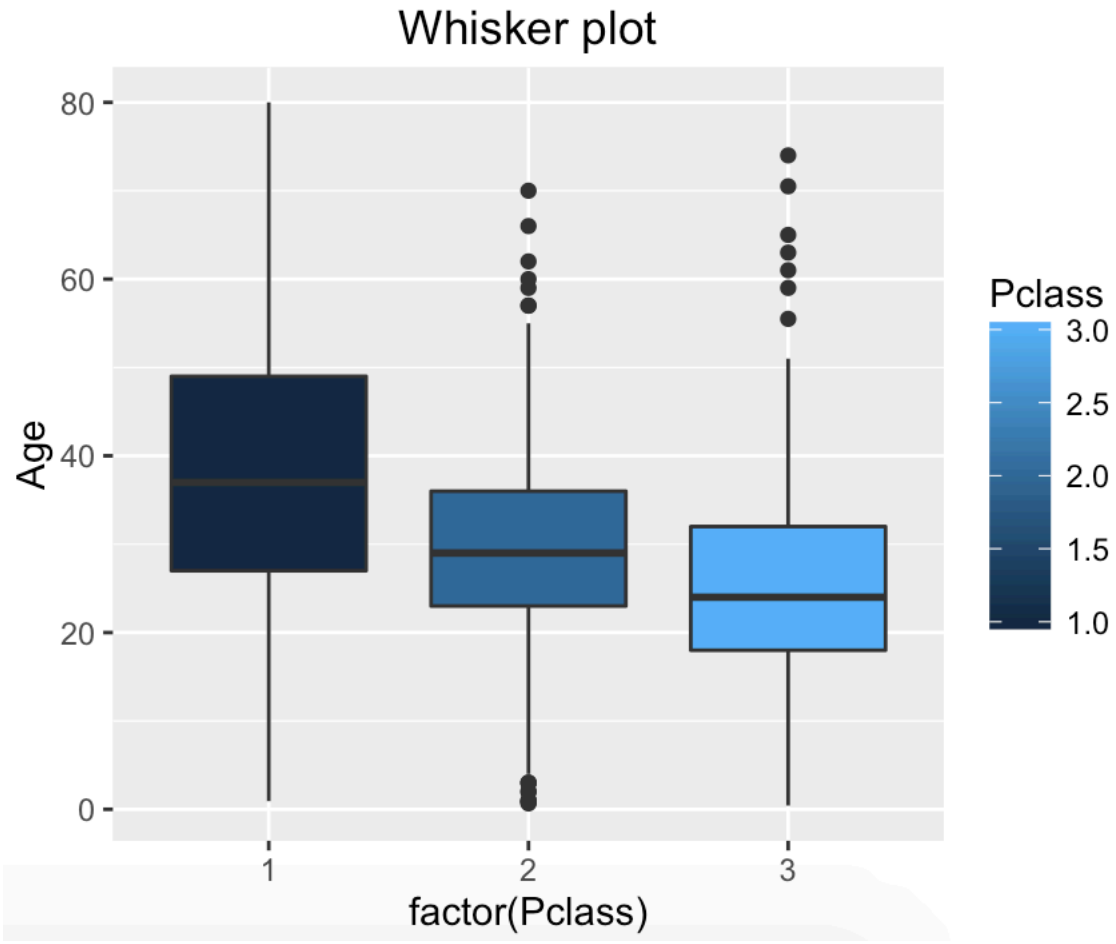
Data Analytics Assignment 4
Haitao Zhou HAZ59@pitt.edu
Github: htz92111
Xin Jin XIJ21@pitt.edu
Github: navis09
Yingzhi Yang YIY50@pitt.edu
Github: HenryYang0914

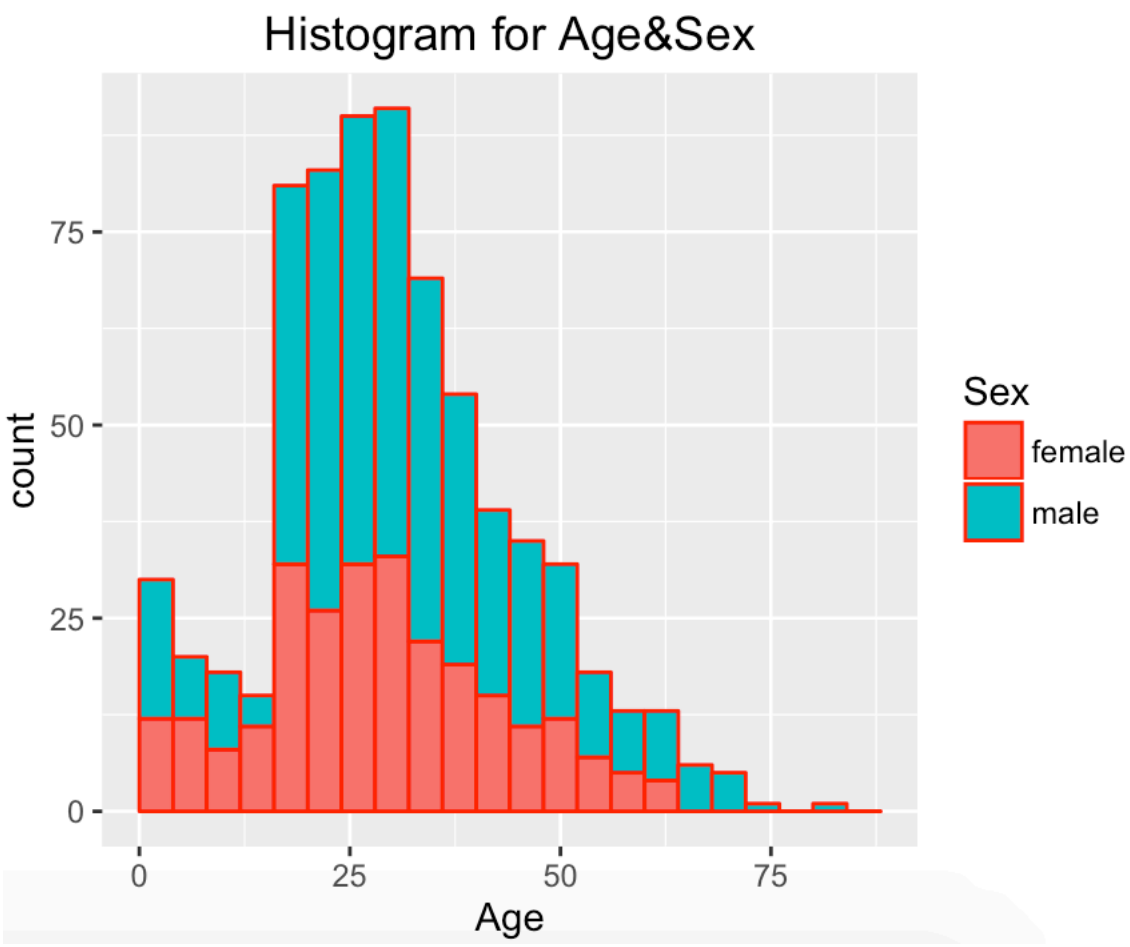
1.install package ggplot2
install.packages(“ggplot2”)
library("ggplot2", lib.loc="/Library/Frameworks/R.framework/Versions/3.2/Resources/library")

2.load data train.csv
dat <- read.csv("/Users/JasonKing/Downloads/train.csv")
summary(dat)

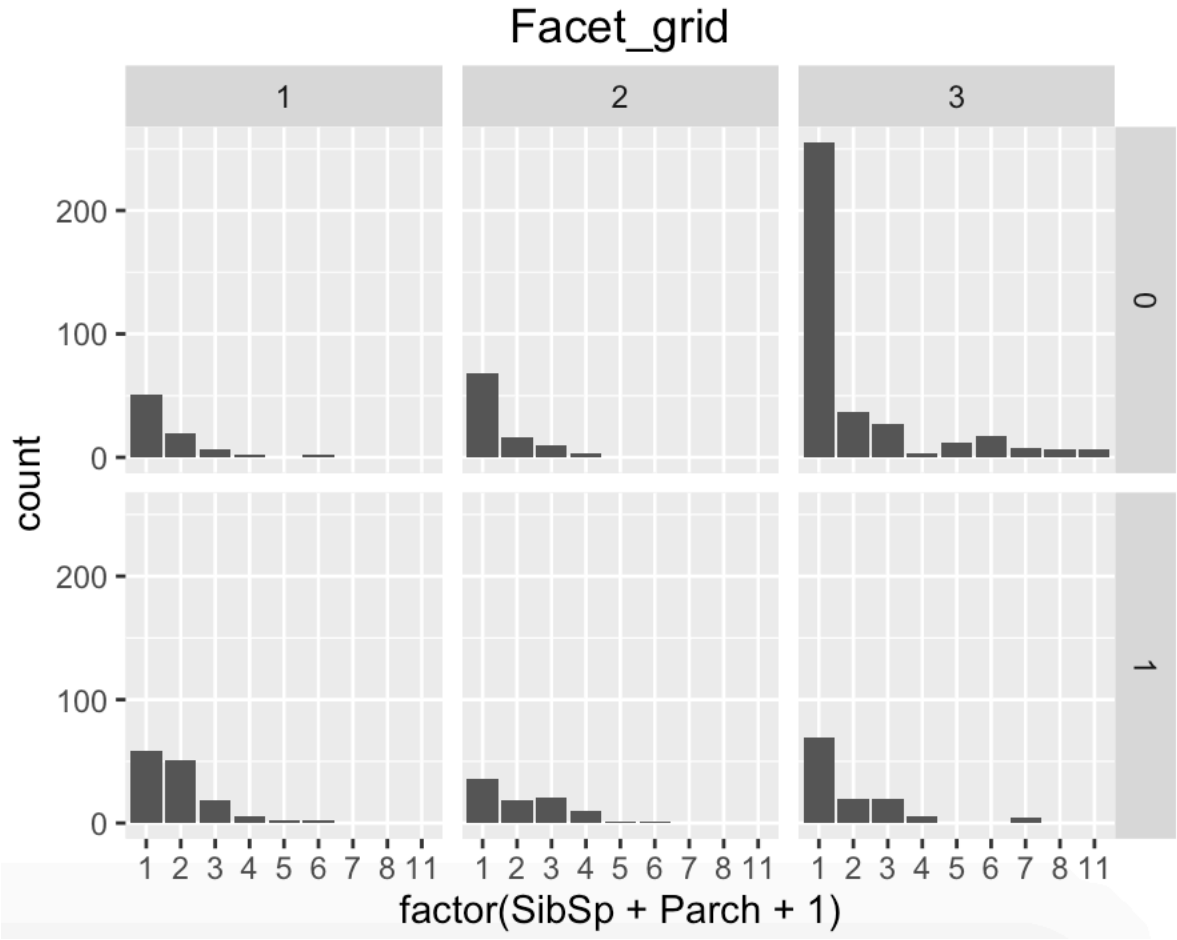
3.R programming and screen shot of result
1. ggplot(dat,aes(factor(Pclass),Age))+geom_boxplot(aes(fill = Pclass))+labs(title="Whisker plot")



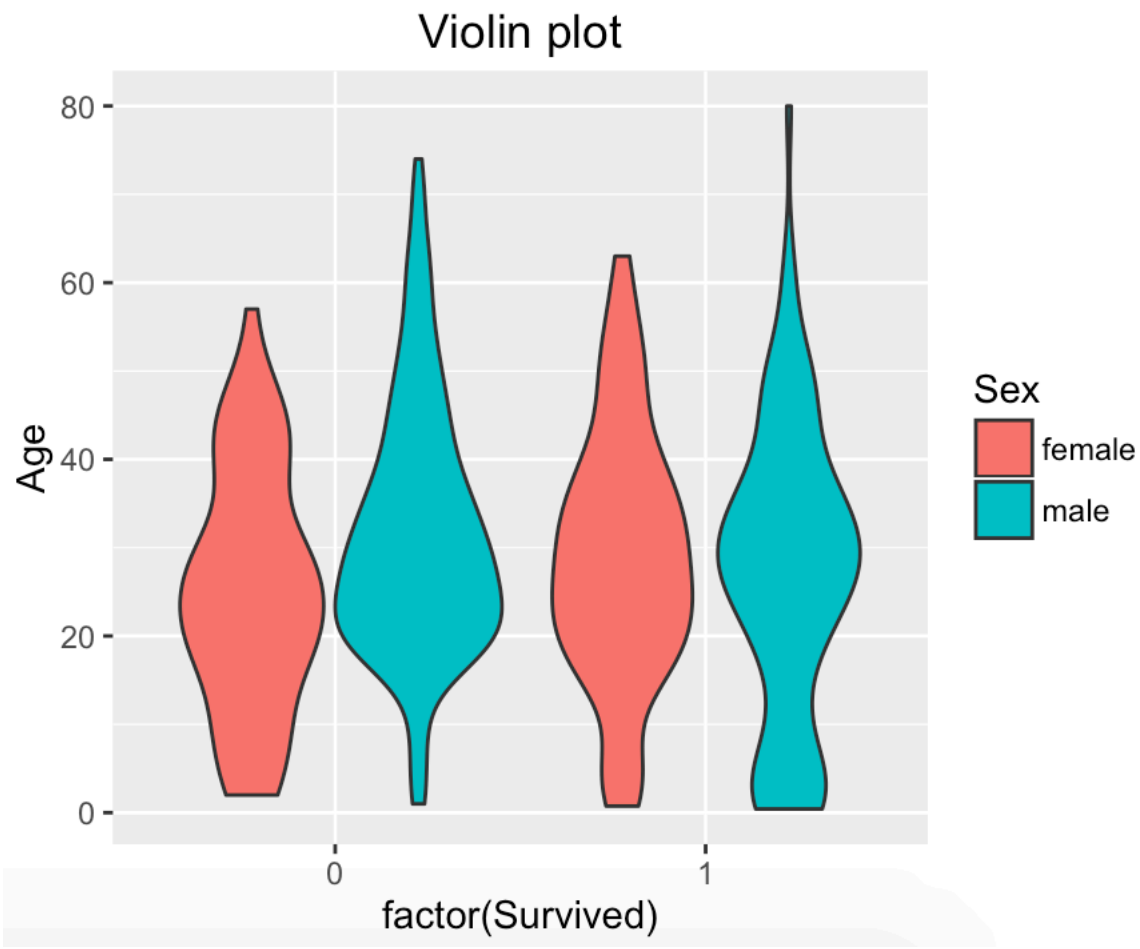
2.ggplot(dat)+geom_histogram(aes(x=Age,fill=Sex),breaks=seq(0, 90, by =4), col="red")+labs(title="Histogram for Age&Sex")



3.ggplot(dat)+geom_bar(aes(factor(SibSp+Parch+1)))+facet_grid(Survived~Pclass)+labs(title="Facet_grid")

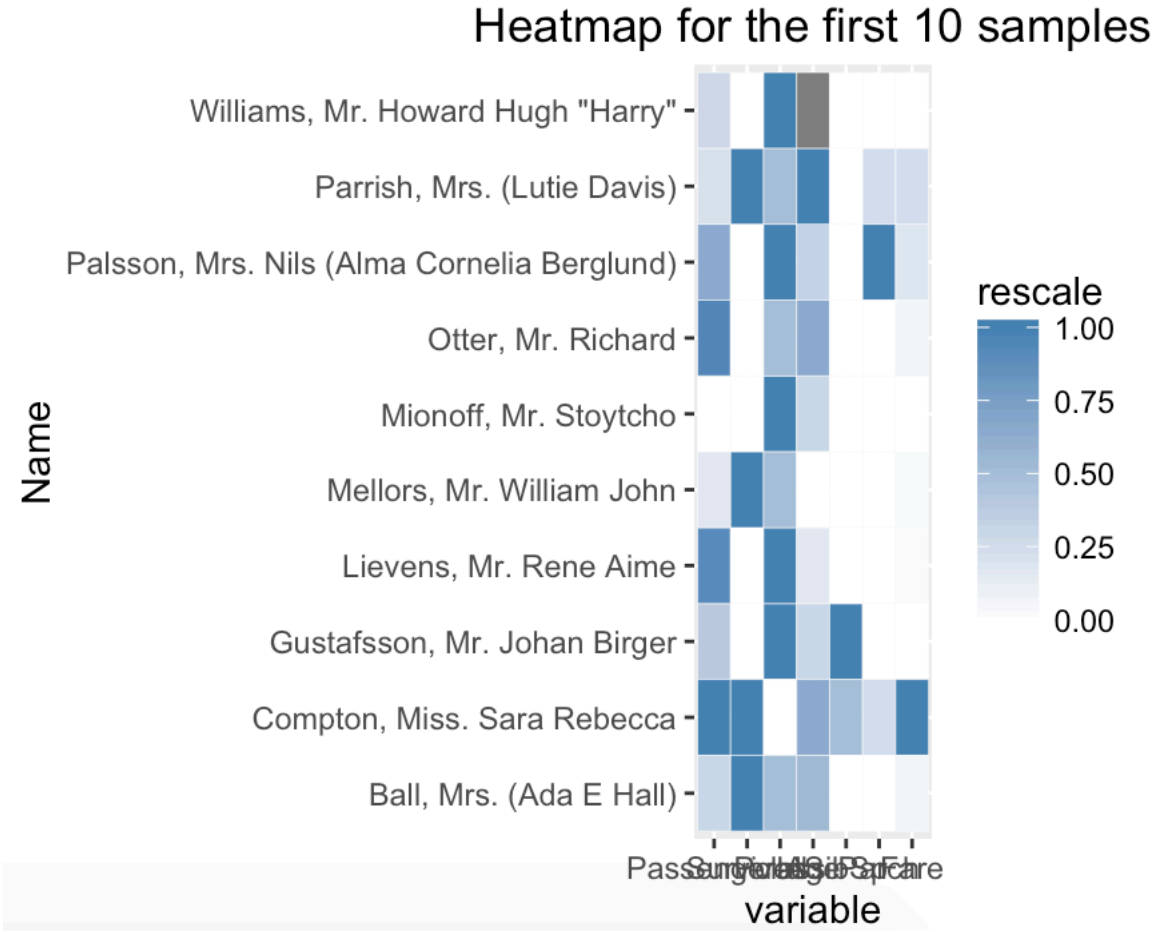


4.ggplot(dat,aes(factor(Survived),Age))+geom_violin(aes(fill=Sex))+labs(title="Violin plot")



5. small <- dat[sample(nrow(dat),10),]
small.m <- melt(small)

```
small.m <- ddply(small.m, .(variable),transform,rescale=rescale(value))
ggplot(small.m,aes(variable,Name))+geom_tile(aes(fill=rescale),color='white')+scale_fill_gradient(low="white",high="steelblue")
```



In this assignment, we use R language to analysis the passengers of Titanic：
Firstly, we created a whisker plot of the passengers' ages with Pclass. We can easily see the distribution of passengers' age from 3 Class:
1.Passengers' age in class one is very concentrated around 40
2.Age in class two is lower than class one ,which is concentrate around 30, and have several particular values
3.Passengers in class three are always younger, which around 25 years old and have more particular values.

Secondly, we used a histogram to analysis the age and sex of the passengers. We can easily see:
1. male passengers are much more than female passengers.
2.the peak of both male's and female's age is around 30.
3. There are rarely female passengers who are older than 60 years old.

Thirdly, we draw a facet grid, it tells the sum of passengers with different SibSP and Parch. We can see the relation in "class", "passenger is survive or dead", "sib of passengers":

1.Dead people are much more than survived people in every class.
2.The amount of people in class 3 is largest, but most of them are dead.
3.Most of people only have 1 sib (no matter from which class, no matter survived or dead)
4.The amount of people in class 1 is smallest, but the survived rate is highest in all 3 classes.

Fourthly, we mixed the Survived factor into age & sex & survived or dead, using a violin plot:
1.The age range of female is smaller than that of male(no matter survived or dead)
2. Most of survived people are around 20 years old(no matter male or female)
3. Most of boys (young male) are survived.
4. Most of girls (young female) are dead.
5.There are some really old male, who is almost 80 years old, are survived.

The final diagram is a heatmap. We selected10 passengers in the data file to analysis, and we show all attributes of these ten passengers in colors. The deeper the color is, the larger the value is. This diagram is not only easy to figure out one specific passenger's attributes , but also easy to compare one's attributes to another's. Furthermore, color is a very directly way to see the degree.

From the R assignment, we can see that R language is very strong for data analysis, and we can choose the right diagram for different problems. Each of the diagrams is helpful. Some of them can represent the relationship and numerical values directly and clearly. While the others may contain more information in one diagram. So when we try to program in R to solve problem or represent some data to others, we need to select the right diagram to the problem.