Assignment 4

Dan Sun(das225) Yue Su(yus55)

Pre-process

1.Import file

```
train <- read.csv("~/Desktop/train.csv",header=T)
```

2.Import Library

```
library(ggplot2)
```

3.Check Missing Data

```
any(is.na(train))
```

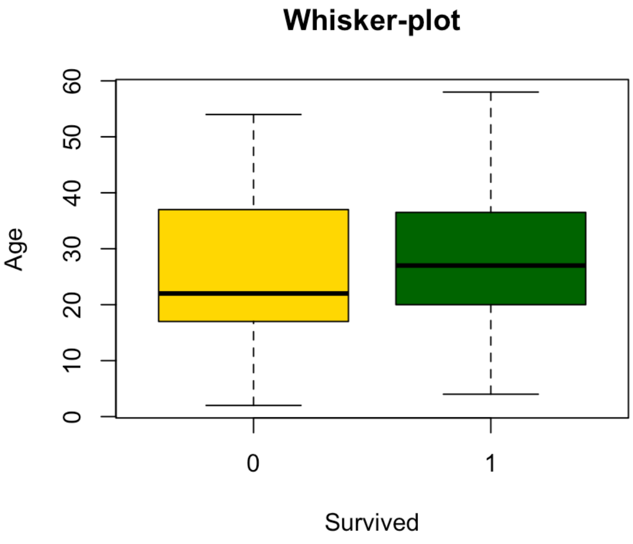
4.Delete invalid items

```
newtrain <- train[!(train$Age==""), ]
```

Whisker-plot

Code:

```
boxplot(Age~Survived, data=newtrain, main="Whisker-plot", xlab="Survived", ylab="Age", col=(c("gold","darkgreen")))
```

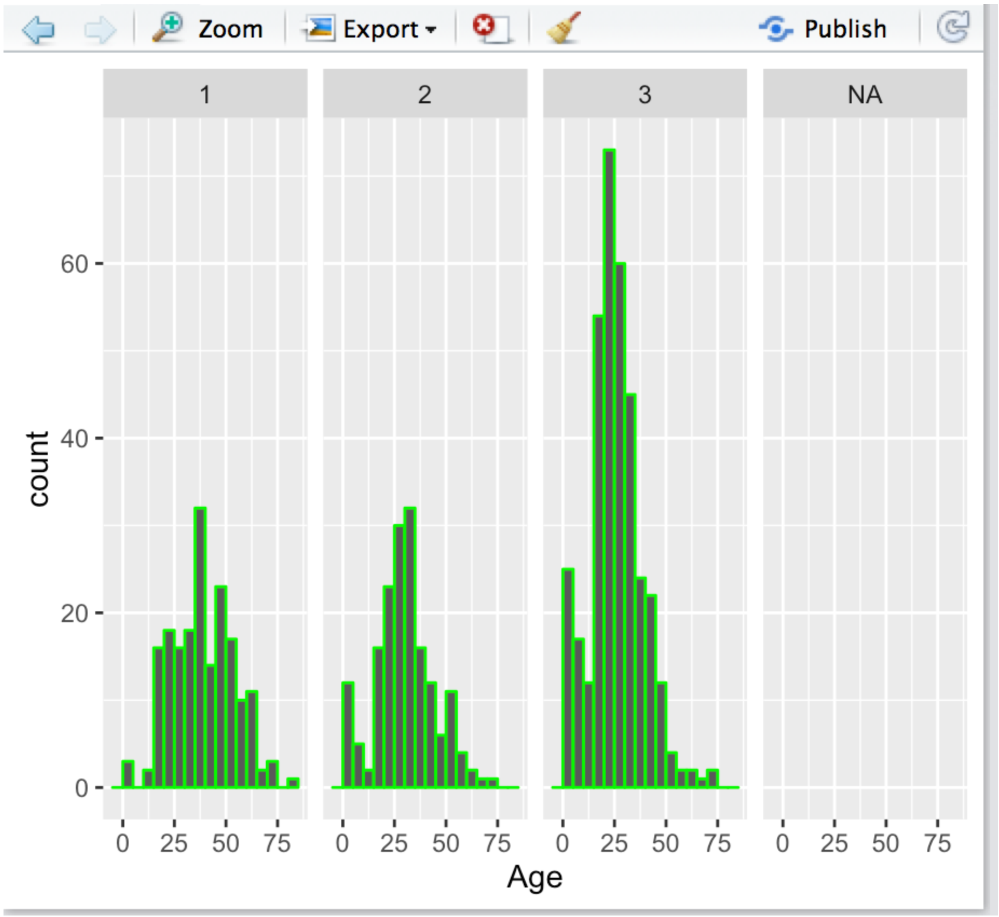


Conclusion:According to these two plots, we are able to know that most passengers are between 20 to 40.Also, we can find out the median of the passengers who are not survived is close to 20 and the median of the passengers who are survived is close to 25.

Histogram + Facet

Code:

ggplot(newtrain,aes(Age))+geom_histogram(aes(fill = Survived), binwidth=5, color="green")+facet_grid(.~Pclass)

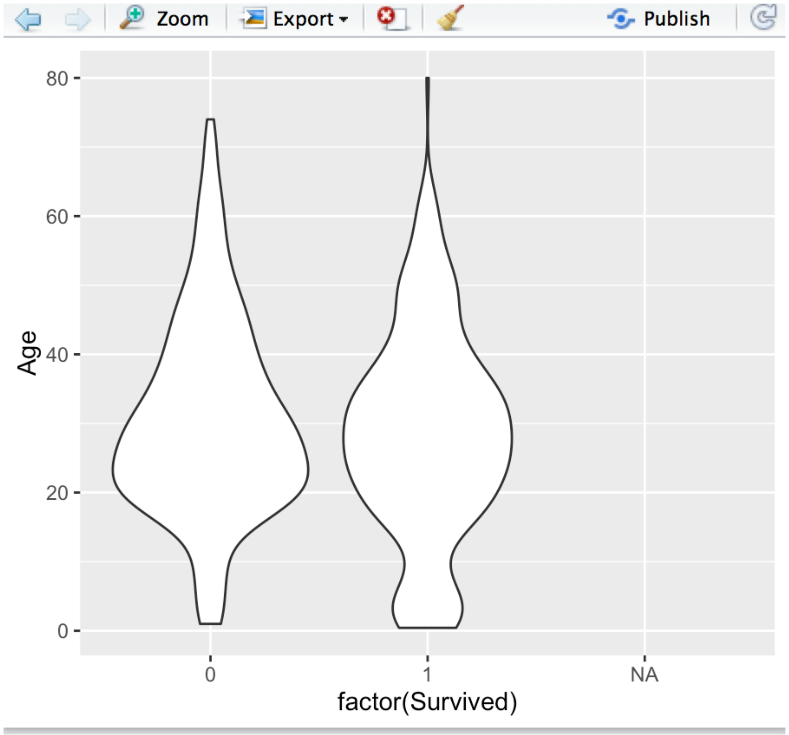


Conclusion: We combine the Histogram plot with the Facet which can generate a overview plot for us. We generate 3 histogram plots based on Pclass . And for each histogram, we are able to come out with different conclusion: For Pclass 1&2, we are able to find out that these two pclass are very similar: for age from 25- 35 ,both of them own most amount of people.However, for Pcalss 3, we can figure out that most people are 25 year old in Pclass 3. Also, Pclass 3 owns much more people.

Violin

Code:

```
ggplot(newtrain,aes(factor(Survived),Age))+geom_violin()
```



Conclusion: According to the violin plot, we are able to find out the distribution of age as well as sex based on survived passengers in a single plot.In this plot , we can see that male passengers own much longer age range than female passengers.Also , both of male and female passengers are collected in range of 20 - 40.

Heatmap

Code:

```
newtrain<-newtrain[1:15,]
train_matrix<-data.matrix(newtrain)
heatmap(train_matrix, main = "Correlation", notecol="black", density.info="none", trace="none", margins =c(12,9), col=my_palette, breaks=col_breaks, dendrogram="row", Colv="NA")
```

Conlcusion:

Connot generate a plot;

Error message:

```
Error in hclustfun(distfun(x)) :
  NA/NaN/Inf in foreign function call (arg 11)
```