Data Analytics

Assignment 4

29/02/2016

- **Assignment 4: R Programming**

- **Group members:**

  - Nanxun Xie
  - Yifan Zhao

- **Using System** : RStudio

- **Using Programming language** : R

preparation

STEP 1: Import the train.csv file into RStudio.

R code:

```
# import the csv file "train" into a data frame
train<-read.csv("train.csv",header = T)
```

Comment: First we have to import original data file into RStudio before processing. we use the command above in R console to read the train.csv into the train set.

STEP 2: Loading some necessary package in RStudio

Pre-requested jars:

```
# package ggplot2 is used to draw various type pictures.
library(ggplot2)
```

Comment: To draw pictures, we have to use the package ggplot2 in R studio.

```
# package reshape and plyr are used to process the data in file “train”.
library(reshape)
library(plyr)
library(scales)
```

Comment: To process the data later, we also have to use these three packages above.

```
# package lattice, grid, DMwR are used to fill in the missing value
library(lattice)
library(grid)
library(DMwR)
```

Comment: To deal with the missing value in the data, we also have to use these three packages above.

STEP 3: Fill in the missing value

R code:

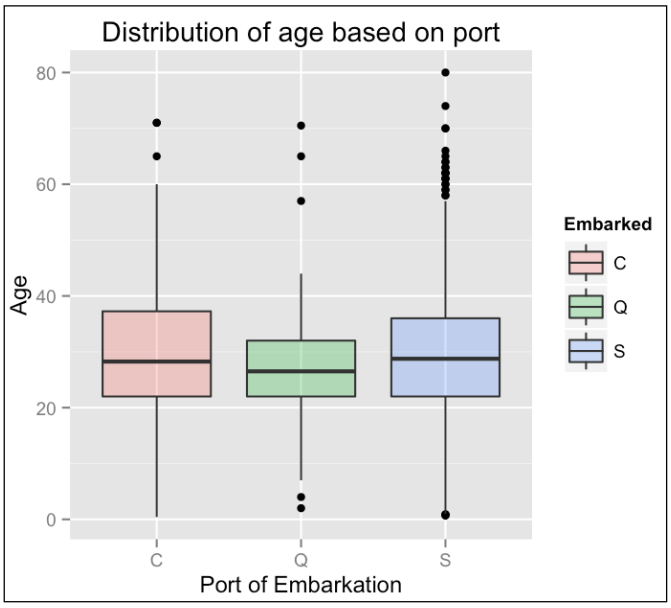
```
# If there are any defaults in ‘Embarked’, using ’S’ to fill in the missing value
train$Embarked[train$Embarked==""] = “S”
```

Comment: Assigning the value “s” to the column Embarked in the data of train if there are any defaults.

```
# using knnImputation function to fill the missing value in age
train<-knnImputation(train,k=10,meth="median")
```

Comment: As we used before, using the knnImputation function to assign to the train data when there are any missing values.

STEP 4:Create different types plot to describe the passengers of Titanic



Whisker-plot

R code:

```
#Create Whisker plot about Port of Embarkation and age of passenger
whisker<-
ggplot(train,aes(Embarked,Age,fill=Embarked))
whisker+geom_boxplot(alpha=0.3)+labs(x='Port of Embarkation',title='Distribution of age based on port')
```

Comment: Using ggplot function to plot the distribution of the age based on the port of embarkation.
the result is shown below:

This Whisper plot shows the distribution of age which is based on port.

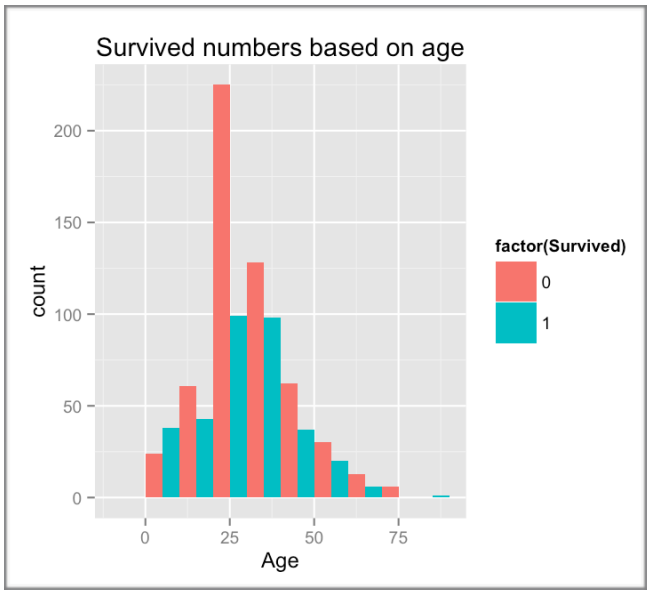
According to the plot, we can know that most passenger are between the age of 20 to 40 among the three different port. And also the median of age are all near to 30.

The range of age of S(Southampton) are the largest among three ports.

Histogram

R code:

```
#Creating the Histogram plot about the survival probability of all ages
hist<-ggplot(train,aes(Age,fill=factor(Survived)))
hist+geom_histogram(position = 'dodge',binwidth=10)+labs(title="Survived numbers based on age")
```



Comment: Using ggplot to portray the histogram of the survival probability of any ages.

the result is shown below:

This Histogram plot shows the number of survivor and victims at all ages.

According the plot, when the age is less than 10 or more than 75, we can see the number of survivor are more than victims. So, we know the children or aged people are more likely to survival. Maybe there are some policies state that the aged people or children are first to save.

Besides, the age between 20 to 30, the number of survivor are much less than victim. So, we know the young people are more likely to die. Maybe they are the last crowds to get away from the ship.

Facet grid

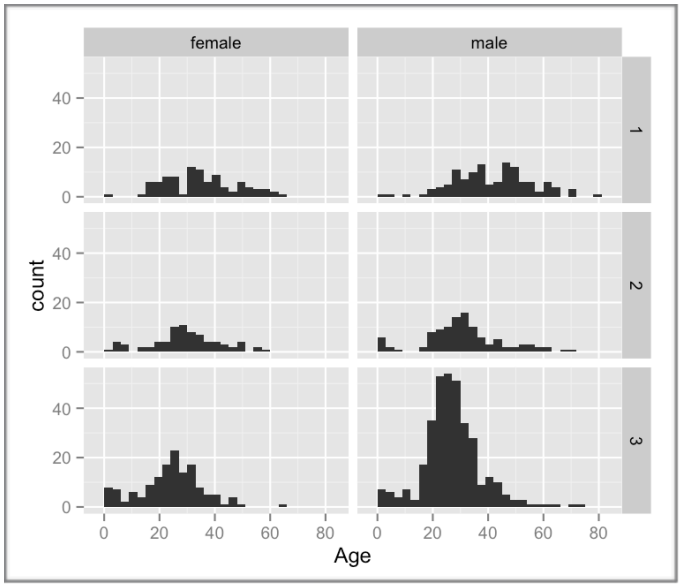
R code:

```
#Create Facet grid plot about the distribution of age based on sex and Pclass(Passenger class)

grid<-ggplot(train,aes(Age))
grid+geom_histogram(binwidth=3)+facet_grid(Pclass~Sex)
```

Comment: With the use of ggplot, we define the grid plot which is used to present the distribution of age based on both the sex and the Pclass.

the result is shown below:



This Facet grid plot shows the distribution of age based on sex and Pclass.

According to the plot, we can know that the amount of passenger who belongs to class three are the most.

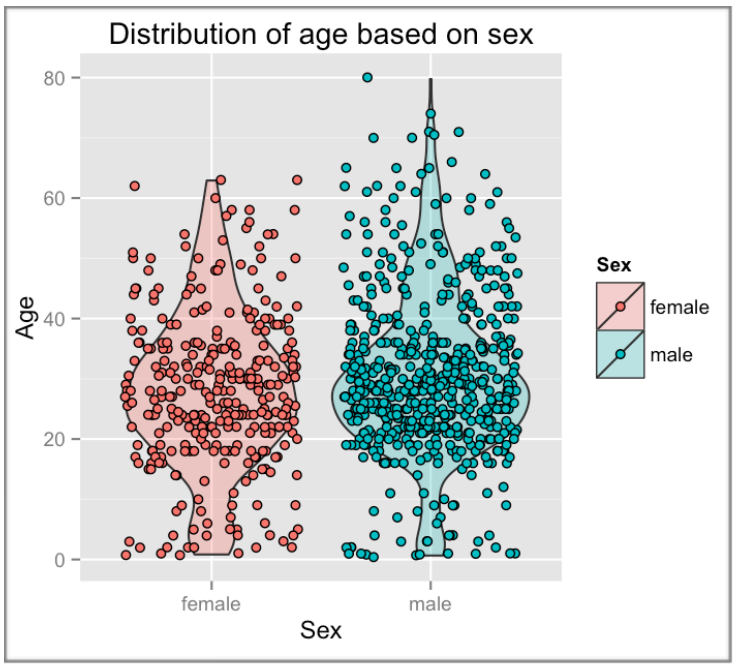
Besides, the number of males are more than female.

Moreover, the number of male passenger and belongs to the class three is extremely higher than others.

Violin plot

R code:

```
#Create the violin plot to describe the distribution of age based on sex
violin<-ggplot(train,aes(Sex,Age,fill=Sex))
```



violin+geom_violin(alpha=0.3,width=0.9)+geom_jitter(shape=21)+
labs(title='Distribution of age based on sex')

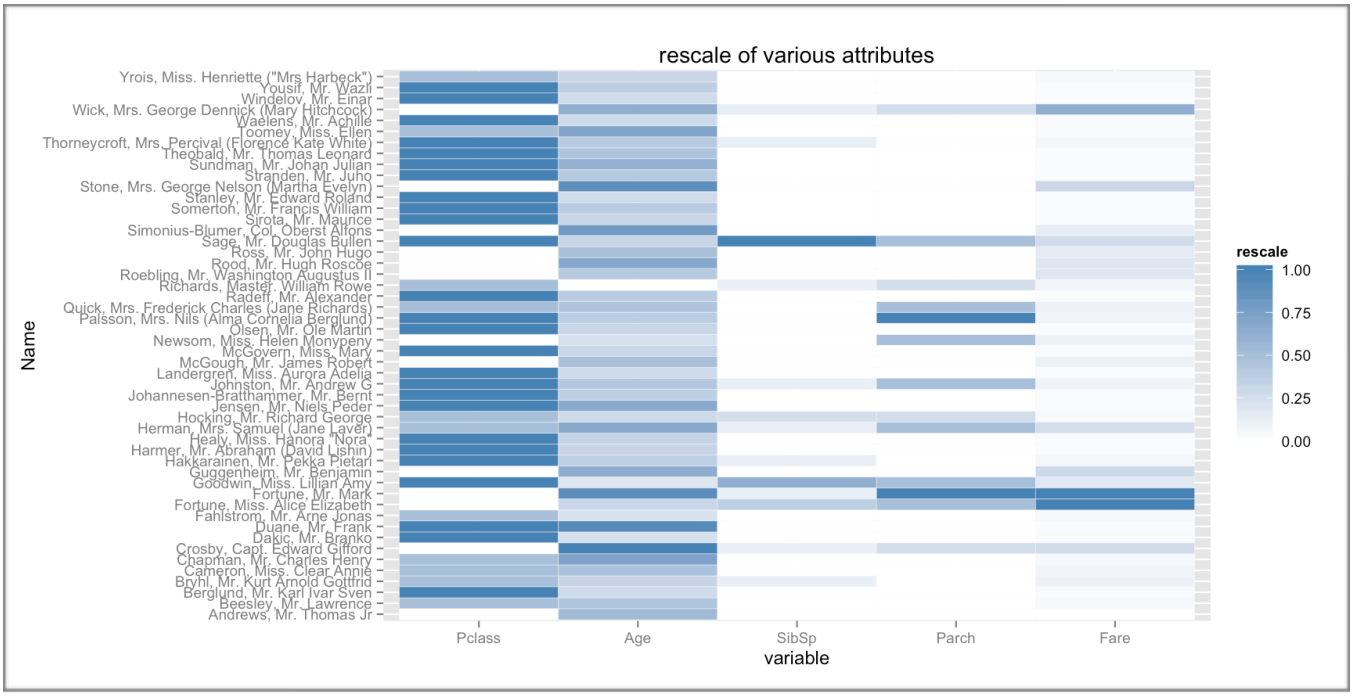
Comment: With the use of ggplot function in the R studio, we describe the violin plot which is used to present the distribution of age based on the sex only.

the result is shown below:

This violin plot demonstrates the distribution of age based on sex.

According to this plot we can see the range of male's age are much larger than it of female.

And most people are between the age of 20 to 40 both in male and female.



Heatmap plot

R code:

```
#Create heat map plot to justify the level of each passenger about Pclass, SibSp,Age,Parch,Fare in a group which pick up 50 sample from train dataset randomly.
trainsmall<-train[sample(nrow(train),50),]
train.m <-
melt(trainsmall,id=c("PassengerId","Survived","Name","Sex","Ticket","Cabin","Embarked"))
train.m <- ddply(train.m, .(variable), transform, rescale = rescale(value))
heat<-ggplot(train.m,aes(x=variable,y=Name,fill=rescale))
heat+geom_tile(aes(fill = rescale),colour = "white") + scale_fill_gradient(low = "white", high = "steelblue")+labs(title='rescale of various attributes')
```

the result is shown below:

This heat map plot shows that the level of each passenger about Pclass, SibSp,Age,Parch,Fare in a group which are made up with 50 samples taken from the train dataset randomly.

According to the plot, we can find that most passenger are belong to class 3, that is to say, most passenger in train dataset should belong to lower class.

Moreover, most passenger have no relatives (which SinSp and Parch are rescale to 0).

Besides, most passenger buy tickets with a lower price, which is also proved by the class level of most passenger, the lower class is normally with a lower price, cheaper.

Summary: After observation, we found that ladies and children are almost alive, and the young man died most, besides, the age, the sex and the Pclass are almost the most important factors of survival.