#load the data set

train <- read.csv("C:/Users/jiaji/Desktop/train.csv",header = T,stringsAsFactors = F)  
test <- read.csv("C:/Users/jiaji/Desktop/test.csv",header = T,stringsAsFactors = F)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data = bind\_rows(train,test)  
data[,c(1,2,3,5,12)] = lapply(data[,c(1,2,3,5,12)], FUN = as.factor)

#basic describes

train = data[1:891,]  
test = data[892:1309,]  
summary(train)

## PassengerId Survived Pclass Name Sex   
## 1 : 1 0:549 1:216 Length:891 female:314   
## 2 : 1 1:342 2:184 Class :character male :577   
## 3 : 1 3:491 Mode :character   
## 4 : 1   
## 5 : 1   
## 6 : 1   
## (Other):885   
## Age SibSp Parch Ticket   
## Min. : 0.42 Min. :0.000 Min. :0.0000 Length:891   
## 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000 Class :character   
## Median :28.00 Median :0.000 Median :0.0000 Mode :character   
## Mean :29.70 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 C :168   
## 1st Qu.: 7.91 Class :character Q : 77   
## Median : 14.45 Mode :character S :644   
## Mean : 32.20 NA's: 2   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

colnames(train)

## [1] "PassengerId" "Survived" "Pclass" "Name" "Sex"   
## [6] "Age" "SibSp" "Parch" "Ticket" "Fare"   
## [11] "Cabin" "Embarked"

str(train)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: Factor w/ 1309 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr NA "C85" NA "C123" ...  
## $ Embarked : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...

library("pastecs")

##   
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':  
##   
## first, last

stat.desc(train)

## PassengerId Survived Pclass Name Sex Age SibSp  
## nbr.val NA NA NA NA NA 7.140000e+02 891.00000000  
## nbr.null NA NA NA NA NA 0.000000e+00 608.00000000  
## nbr.na NA NA NA NA NA 1.770000e+02 0.00000000  
## min NA NA NA NA NA 4.200000e-01 0.00000000  
## max NA NA NA NA NA 8.000000e+01 8.00000000  
## range NA NA NA NA NA 7.958000e+01 8.00000000  
## sum NA NA NA NA NA 2.120517e+04 466.00000000  
## median NA NA NA NA NA 2.800000e+01 0.00000000  
## mean NA NA NA NA NA 2.969912e+01 0.52300786  
## SE.mean NA NA NA NA NA 5.436405e-01 0.03694329  
## CI.mean NA NA NA NA NA 1.067328e+00 0.07250613  
## var NA NA NA NA NA 2.110191e+02 1.21604308  
## std.dev NA NA NA NA NA 1.452650e+01 1.10274343  
## coef.var NA NA NA NA NA 4.891222e-01 2.10846437  
## Parch Ticket Fare Cabin Embarked  
## nbr.val 891.00000000 NA 891.000000 NA NA  
## nbr.null 678.00000000 NA 15.000000 NA NA  
## nbr.na 0.00000000 NA 0.000000 NA NA  
## min 0.00000000 NA 0.000000 NA NA  
## max 6.00000000 NA 512.329200 NA NA  
## range 6.00000000 NA 512.329200 NA NA  
## sum 340.00000000 NA 28693.949300 NA NA  
## median 0.00000000 NA 14.454200 NA NA  
## mean 0.38159371 NA 32.204208 NA NA  
## SE.mean 0.02700393 NA 1.664792 NA NA  
## CI.mean 0.05299881 NA 3.267377 NA NA  
## var 0.64972824 NA 2469.436846 NA NA  
## std.dev 0.80605722 NA 49.693429 NA NA  
## coef.var 2.11234407 NA 1.543073 NA NA

#check NA

sapply(train,function(x)sum(is.na(x)))

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 177   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 687 2

sapply(test,function(x)sum(is.na(x)))

## PassengerId Survived Pclass Name Sex Age   
## 0 418 0 0 0 86   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 1 327 0

#fill column embarked

which(is.na(train$Embarked))

## [1] 62 830

#发现是62和830号乘客缺少embarked信息,通过查看他们的票价和船舱，发现都是一等舱乘客，票价均为80.

train[c(62,830),c("Survived","Pclass","Sex","Age","Fare")]

## Survived Pclass Sex Age Fare  
## 62 1 1 female 38 80  
## 830 1 1 female 62 80

library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v readr 1.3.1  
## v tibble 2.1.3 v purrr 0.3.2  
## v tidyr 1.0.0 v stringr 1.4.0  
## v ggplot2 3.2.1 v forcats 0.4.0

## -- Conflicts ---------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x tidyr::extract() masks pastecs::extract()  
## x dplyr::filter() masks stats::filter()  
## x pastecs::first() masks dplyr::first()  
## x dplyr::lag() masks stats::lag()  
## x pastecs::last() masks dplyr::last()

data2=filter(train,Pclass==1)  
summarise(group\_by(data2,Embarked),mfare = median(Fare),n())

## Warning: Factor `Embarked` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

## # A tibble: 4 x 3  
## Embarked mfare `n()`  
## <fct> <dbl> <int>  
## 1 C 78.3 85  
## 2 Q 90 2  
## 3 S 52 127  
## 4 <NA> 80 2

data[c(62,830),c("Embarked")]="C"

#填补fare

which(is.na(data$Fare))

## [1] 1044

f=filter(data,Embarked=="S",Pclass==3)  
data[1044,c("Fare")]=round(mean(na.omit(f$Fare)),2)

age缺失值太多了，考虑用回归预测的方法进行填补，先把需要进行回归预测的变量变成factor

data[,c(1,2,3,5,12)] = lapply(data[,c(1,2,3,5,12)], FUN = as.factor)

#回归预测很重要，要掌握

data\_no\_age=data[which(is.na(data$Age)==TRUE),]  
data\_age=data[which(is.na(data$Age)==FALSE),]  
lm=lm(Age~Pclass+Sex+SibSp+Parch+Fare+Embarked,data=data\_age)  
age\_predict=predict(lm,newdata=data\_no\_age)  
data[which(is.na(data$Age)==TRUE),"Age"]=round(age\_predict,0)

#choose useful columns

datause=subset(train,select=c("Survived","Pclass","Sex","Age","SibSp","Parch","Fare","Embarked"))

#visulization

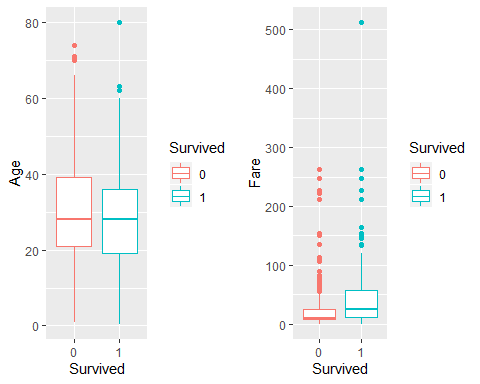
q1=qplot(Survived, Age, data = datause,colour = Survived,geom="boxplot")  
q2=qplot(Survived, Fare, data = datause,colour = Survived,geom="boxplot")  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

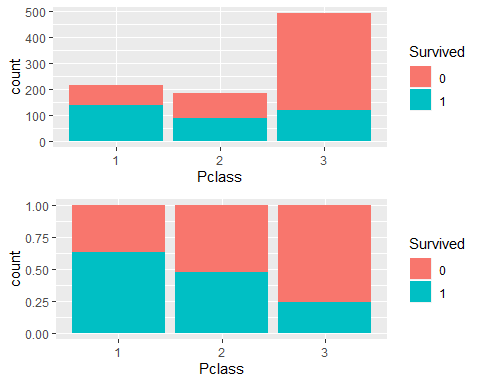
## The following object is masked from 'package:dplyr':  
##   
## combine

grid.arrange(q1,q2,ncol=2)

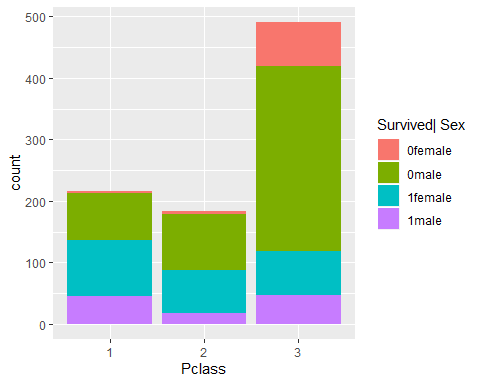
## Warning: Removed 177 rows containing non-finite values (stat\_boxplot).



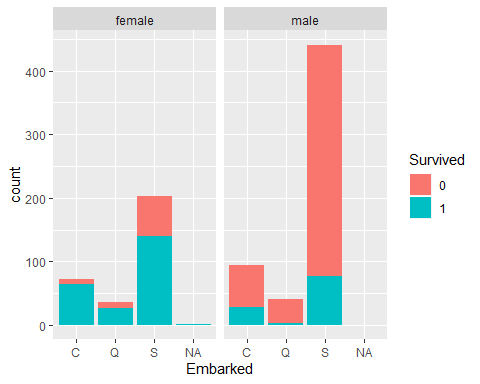
g1=ggplot(data=datause,aes(x=Pclass,fill=Survived))+geom\_bar()  
g2=ggplot(data=datause,aes(x=Pclass,fill=Survived))+geom\_bar(position="fill")  
library(gridExtra)#combine two plots  
grid.arrange(g1,g2,nrow=2)



library(stringr)  
ggplot(data=datause,aes(x=Pclass,fill=str\_c(Survived,Sex)))+geom\_bar()+scale\_fill\_discrete("Survived| Sex")

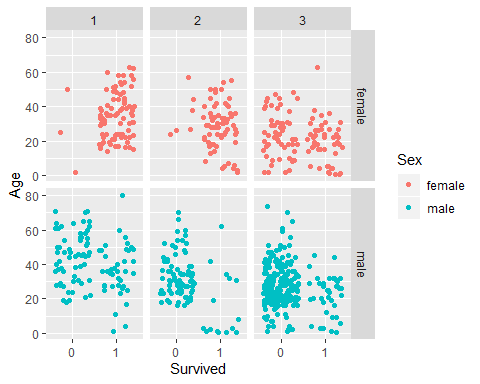


ggplot(data=datause,aes(x=Embarked,fill=Survived))+geom\_bar()+facet\_wrap(~Sex)



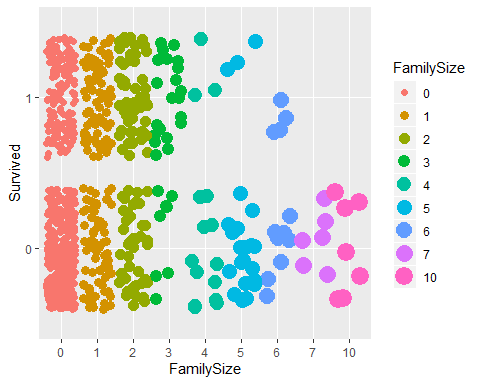
ggplot(data = datause,aes(x=Survived,y=Age))+geom\_jitter(aes(color=Sex))+facet\_grid(Sex~Pclass)

## Warning: Removed 177 rows containing missing values (geom\_point).



ggplot(data=datause,aes(x =as.factor(SibSp + Parch), y = Survived))+labs(x="FamilySize")+  
geom\_jitter(aes(color =(as.factor(SibSp + Parch)), size = (as.factor(SibSp + Parch))))+scale\_color\_discrete("FamilySize") +scale\_size\_discrete("FamilySize")

## Warning: Using size for a discrete variable is not advised.



train = data[1:891,]  
test = data[892:1309,]  
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

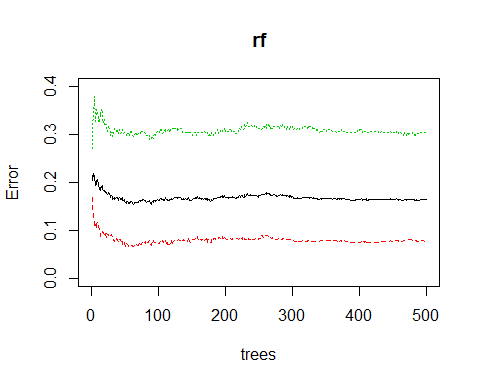
##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

## The following object is masked from 'package:dplyr':  
##   
## combine

rf = randomForest(Survived~Sex+Age+Pclass+SibSp+Parch+Fare+Embarked,data =train)  
plot(rf,ylim = c(0,0.4))



for(i in 80:90){  
 mtry\_fit\_2=randomForest(Survived~Sex+Age+Pclass+SibSp+Parch+Fare+Embarked,data =train,mtry=3,ntree=i)  
 err=mean(mtry\_fit\_2$err.rate)  
 print(err)  
 }

## [1] 0.188787  
## [1] 0.193994  
## [1] 0.1906157  
## [1] 0.1870306  
## [1] 0.1882639  
## [1] 0.189527  
## [1] 0.1909723  
## [1] 0.1869315  
## [1] 0.190761  
## [1] 0.1896108  
## [1] 0.1908315

rf = randomForest(Survived~Sex+Age+Pclass+SibSp+Parch+Fare+Embarked,data = train,ntree = 500,mtry=3,importance=TRUE)  
prediction = predict(object = rf,newdata = test)  
result = data.frame(test$PassengerId,prediction)