Case: Survival Prediction of Titanic Disaster

qinluoao

2017/8/20

1912年泰坦尼克号邮轮在其处女启航时触礁冰山而沉没，电影《泰坦尼克号》中穷画家Jack和贵族女Rose抛弃世俗的偏见坠入爱河，最终Jack把生命的机会让给了Rose。下面我们用数据来理性看待电影中的情节是否是灾难面前一个个动人故事的缩影。

1.数据预处理及探索性分析

安装包的语句：install.packages(c("tidyverse","mnormt","xml2","VIM"， "mice","randomForest"))

```
library(tidyverse) #数据科学工具包
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(VIM)  #缺失值绘图
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

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

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issu
es
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(mice) #多重插补包
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete
```

```
library(scales) #绘图设置坐标轴美元符号
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(caret) #算法包，交叉验证
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(randomForest)
```

```
## randomForest 4.6-12
```

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

```
##
## Attaching package: 'randomForest'
```

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

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

以上操作完成分析需要的各种包安装,接着导入数据：

```
setwd("F:/Code/Titanic")
train <- read.csv("train.csv",stringsAsFactor=F,na.strings = "")
#na.strings = ""如果没写，会导致后面缺失值判断有误
test <- read.csv("test.csv",stringsAsFactor=F,na.strings = "")
```

1.1探索分析

```
head(train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                             Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                              Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                            Allen, Mr. William Henry   male  35     0
## 6                                    Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500  <NA>        S
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250  <NA>        S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500  <NA>        S
## 6     0           330877  8.4583  <NA>        Q
```

```
str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  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        : chr  "male" "female" "female" "female" ...
##  $ 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   : chr  "S" "C" "S" "S" ...
```

```
summary(train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.    :  1.0   Min.    :0.0000   Min.    :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     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
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
attach(train)
```

1.1.1分类变量sex,pclass,embarked与Survived的关系

```
prop.table(table(Sex, Survived),1)  #单个因素中存活百分比查看
```

```
##         Survived
## Sex              0         1
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

```
prop.table(table(Pclass, Survived),1)
```

```
##       Survived
## Pclass         0         1
##      1 0.3703704 0.6296296
##      2 0.5271739 0.4728261
##      3 0.7576375 0.2423625
```

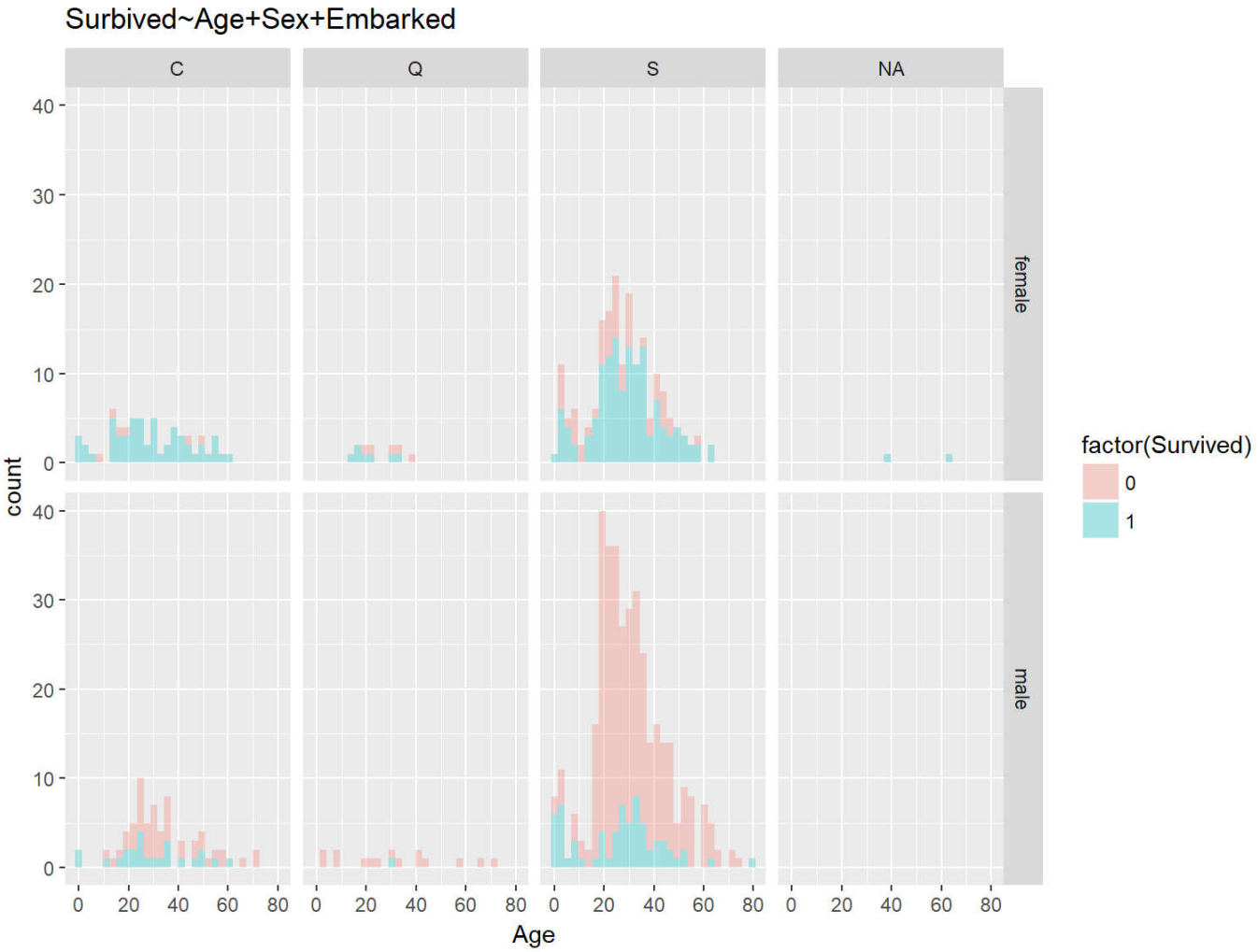
```
prop.table(table(Embarked, Survived),1)
```

```
##          Survived
## Embarked         0         1
##        C 0.4464286 0.5535714
##        Q 0.6103896 0.3896104
##        S 0.6630435 0.3369565
```

1.1.2连续变量Age与Survived的关系

```
ggplot(data=train, aes(x=Age, fill=factor(Survived)))+
geom_histogram(alpha=.3) +
facet_grid(factor(Sex)~factor(Embarked))+
labs(title = "Surbived~Age+Sex+Embarked")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



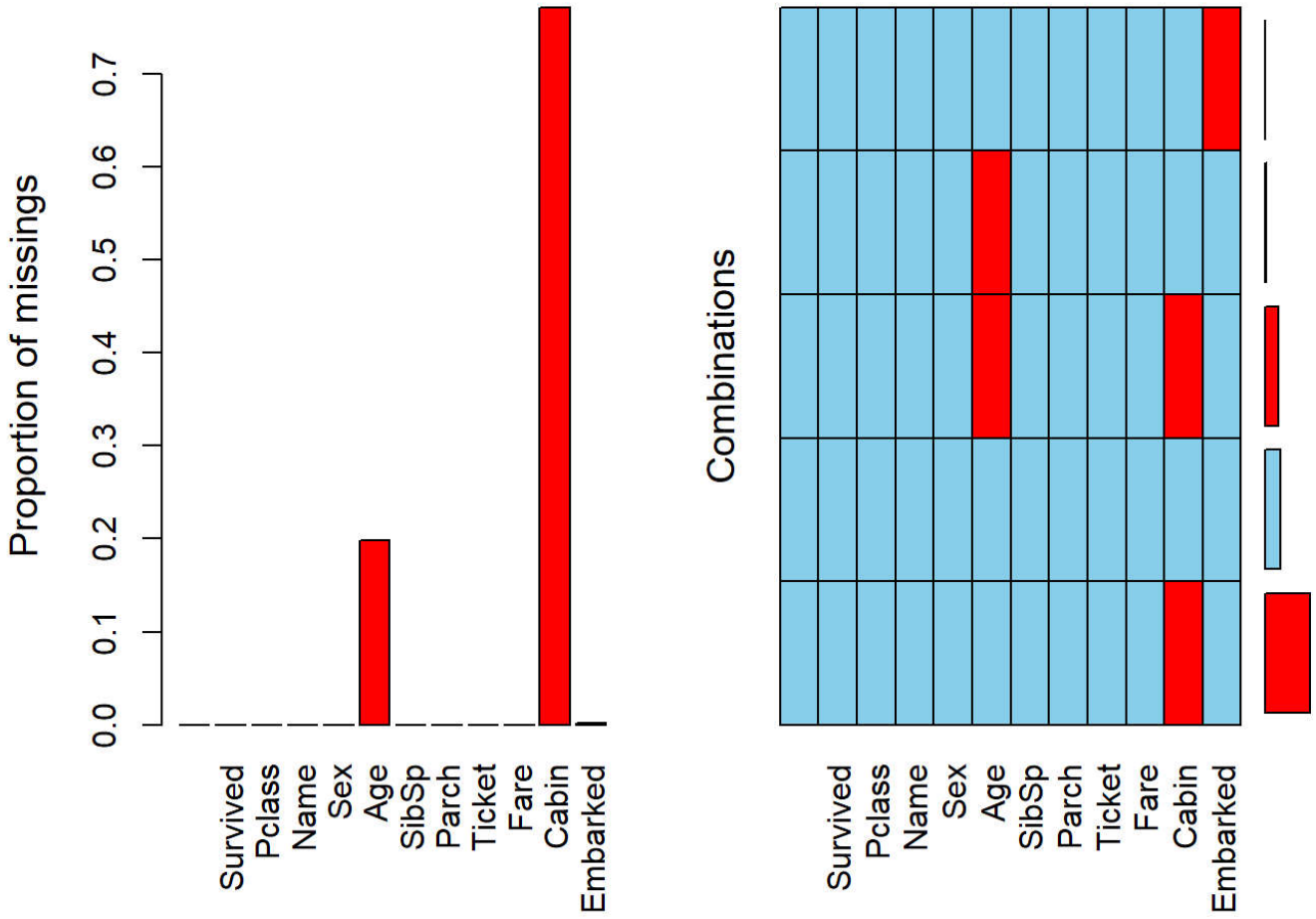
存活与各变量的初步结论: 性别：女性>男性;票等级：1>2>3;登船地点：C>Q>S。 灾难总存活率为38%，女性存活率高达74%，男性存活率仅19%。灾难面前妇女和儿童优先的绅士精神让人感动。

1.2数据缺失情况探索

```
md.pattern(train) #查看缺失情况
```

```
##     PassengerId Survived Pclass SibSp Parch Fare Age Ticket Name Sex Cabin
## 521           1        1      1     1     1    1   1      1    0   0     0
## 140           1        1      1     1     1    1   0      1    0   0     0
## 193           1        1      1     1     1    1   1      0    0   0     0
##  37           1        1      1     1     1    1   0      0    0   0     0
##               0        0      0     0     0    0 177    230  891 891   891
##     Embarked
## 521        0    4
## 140        0    5
## 193        0    5
##  37        0    6
##          891 3971
```

```
aggr(train)    #绘制缺失图表
```




```
mean(is.na(Age)) #求缺失百分比
```

```
## [1] 0.1986532
```

```
mean(is.na(Cabin))
```

```
## [1] 0.7710438
```

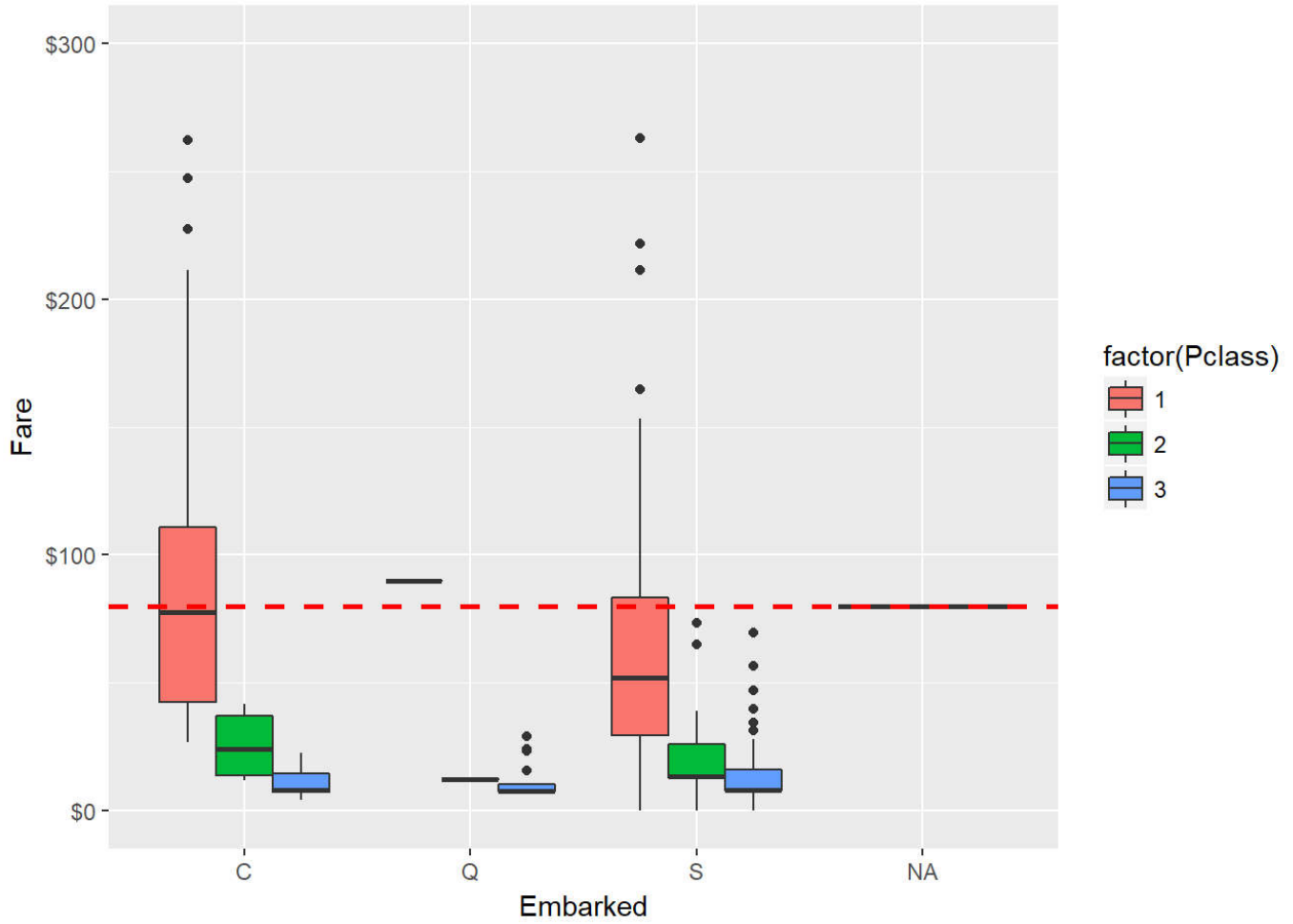
通过以上计算得缺失情况结论：891行*12列数据，其中数据缺失情况为 Embarked：2个，Age:177个缺失比例20%，Cabin：687个缺失比例77%，共计866个缺失值。

1.3缺失值处理

常用方法有行删除、成对删除（cor(sleep, use="pairwise.complete.obs")）、简单插补、多重插补等。

1.3.1Embarked的缺失值处理：简单插补，利用统计量代替缺失值

```
ggplot(train, aes(x = Embarked, y = Fare, fill = factor(Pclass))) +
geom_boxplot() +
geom_hline(aes(yintercept=80), colour='red', linetype='dashed', lwd=1) +
scale_y_continuous(labels=dollar_format(),limits = c(0,300))+
labs(tile="Embarked与Pclass、Fare的关系")
```




```
train[c(62,830),12] <-'C' #train$Embarked [c(62,830)]
Embarked <- factor(Embarked)
```

1.3.2Age的缺失值处理：用mice包多重插补数值类型或因子类型数据

mice插补原理

缺失值的插补通过Gibbs抽样完成。每个包含缺失值的变量都默认可通过数据集中的其他变量预测得来，于是这些预测方程便可用来预测缺失数据的有效值，默认方法ppm(Predictive mean matching)。

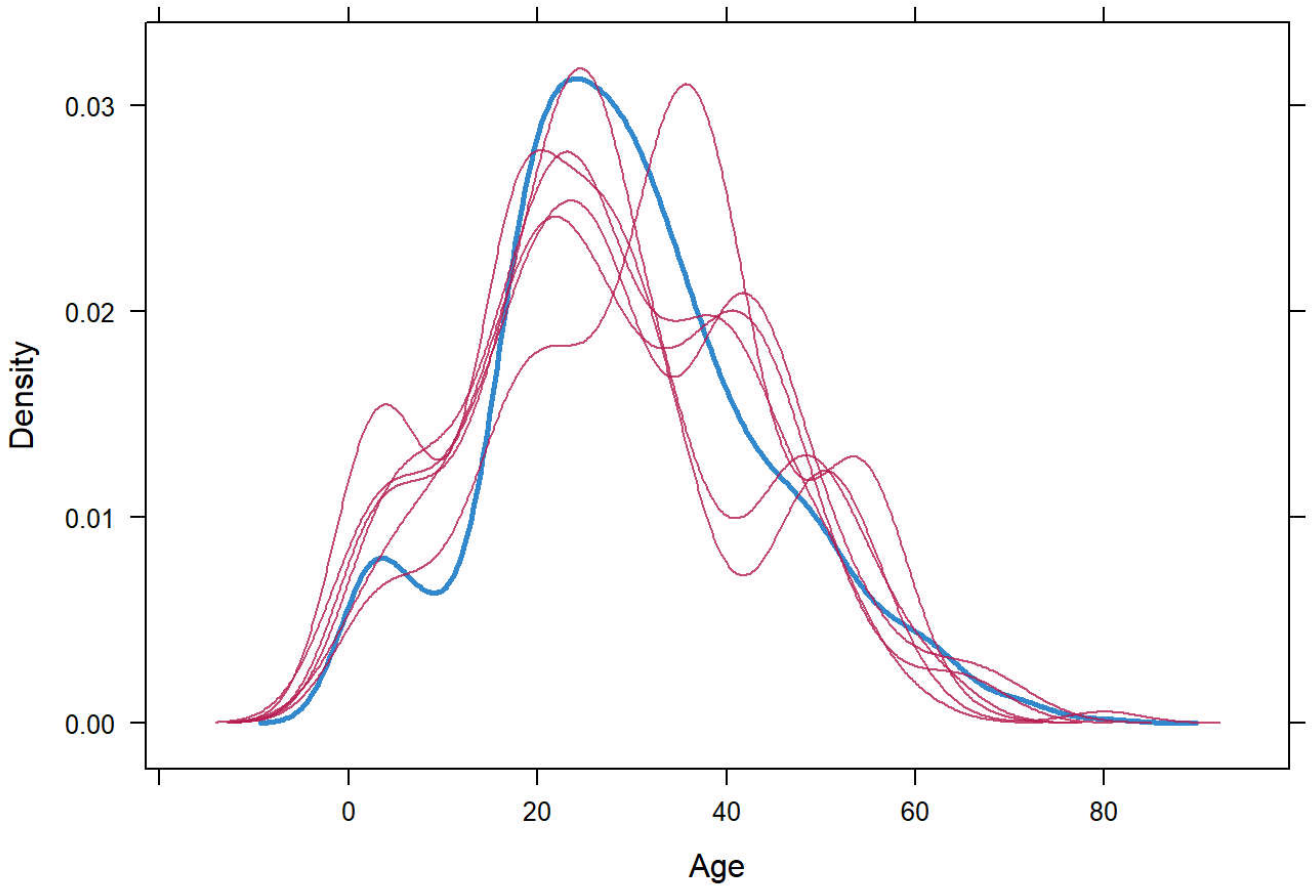
插补过程

缺失数据集——>mice估计插补成多个数据集,imp包含m个插补数据集的列表，以及完成插补过程的VisitSequence、PredictorMatrix——>插补分析：with对每个数据集应用统计模型插补（glm、lm模型）,fit是一个包含m个统计分析结果的列表——>pool将这些单独模型整合到一起，pooled是一个包含这m个统计分析平均结果的列表——>summary评价插补模型优劣（F-TEST）——>complete观察m个插补数据集中的任意一个

```
imp <- mice(train[,-11],m=6,seed = 1234) #删除第11列Cabin数据后对Age插补
```

```
##
##  iter imp variable
##   1   1  Age
##   1   2  Age
##   1   3  Age
##   1   4  Age
##   1   5  Age
##   1   6  Age
##   2   1  Age
##   2   2  Age
##   2   3  Age
##   2   4  Age
##   2   5  Age
##   2   6  Age
##   3   1  Age
##   3   2  Age
##   3   3  Age
##   3   4  Age
##   3   5  Age
##   3   6  Age
##   4   1  Age
##   4   2  Age
##   4   3  Age
##   4   4  Age
##   4   5  Age
##   4   6  Age
##   5   1  Age
##   5   2  Age
##   5   3  Age
##   5   4  Age
##   5   5  Age
##   5   6  Age
```

```
library(lattice)
densityplot(imp) #原数据与插补数据分布对比
```



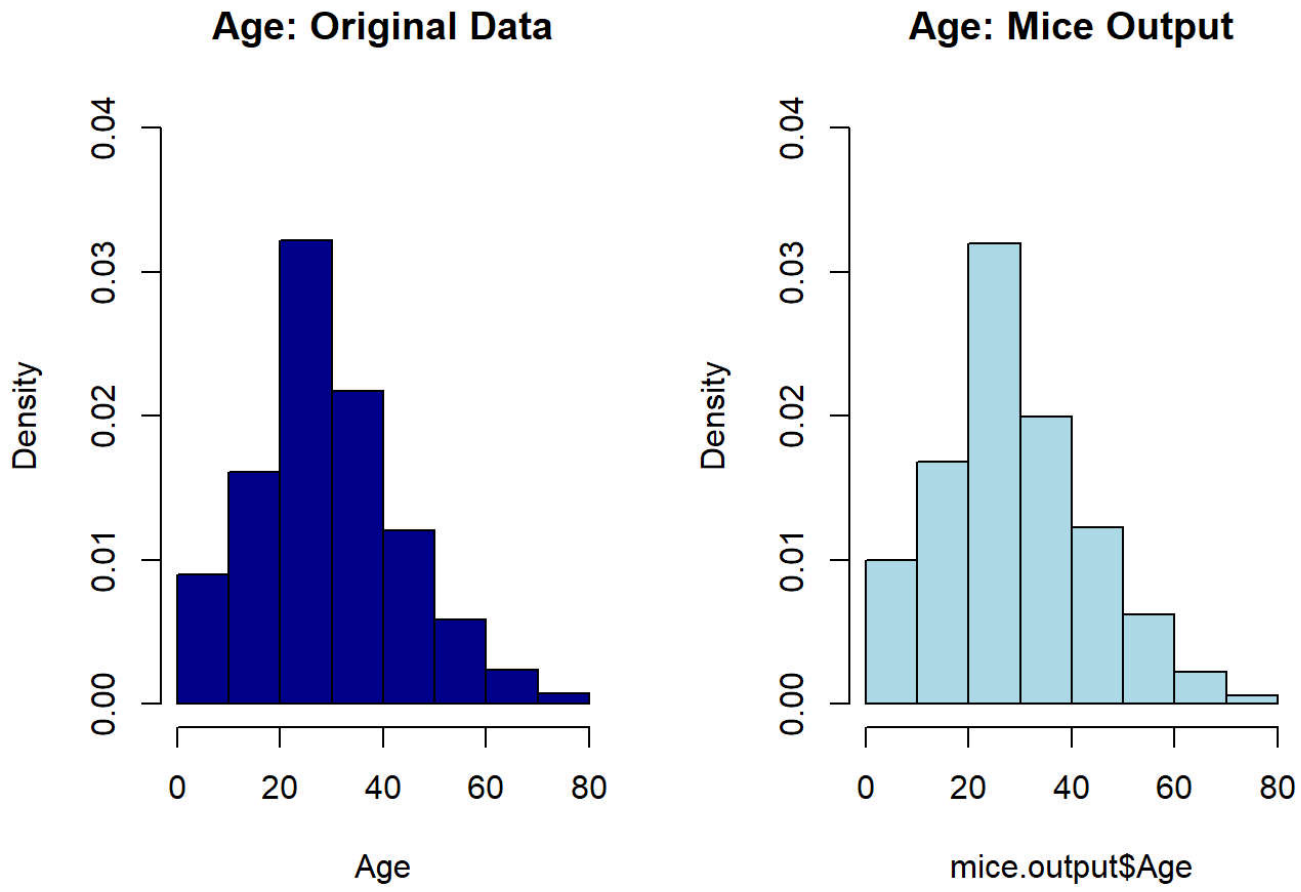

```
#m值增大，可以提高精度，减少随机误差
summary(imp)
```

```
## Multiply imputed data set
## Call:
## mice(data = train[, -11], m = 6, seed = 1234)
## Number of multiple imputations:  6
## Missing cells per column:
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare    Embarked
##           0           0           0           0           0
## Imputation methods:
## PassengerId    Survived      Pclass        Name         Sex         Age
##          ""          ""          ""          ""          ""       "pmm"
##       SibSp       Parch      Ticket        Fare    Embarked
##          ""          ""          ""          ""          ""
## VisitSequence:
## Age
##   6
## PredictorMatrix:
##             PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket
## PassengerId           0        0      0    0   0   0     0     0      0
## Survived              0        0      0    0   0   0     0     0      0
## Pclass                0        0      0    0   0   0     0     0      0
## Name                  0        0      0    0   0   0     0     0      0
## Sex                   0        0      0    0   0   0     0     0      0
## Age                   1        1      1    0   0   0     1     1      0
## SibSp                 0        0      0    0   0   0     0     0      0
## Parch                 0        0      0    0   0   0     0     0      0
## Ticket                0        0      0    0   0   0     0     0      0
## Fare                  0        0      0    0   0   0     0     0      0
## Embarked              0        0      0    0   0   0     0     0      0
##             Fare Embarked
## PassengerId    0        0
## Survived       0        0
## Pclass         0        0
## Name           0        0
## Sex            0        0
## Age            1        0
## SibSp          0        0
## Parch          0        0
## Ticket         0        0
## Fare           0        0
## Embarked       0        0
## Random generator seed value:  1234
```

```
mice.output <- complete(imp,1)
```

初始数据和插补后的数据分布对比

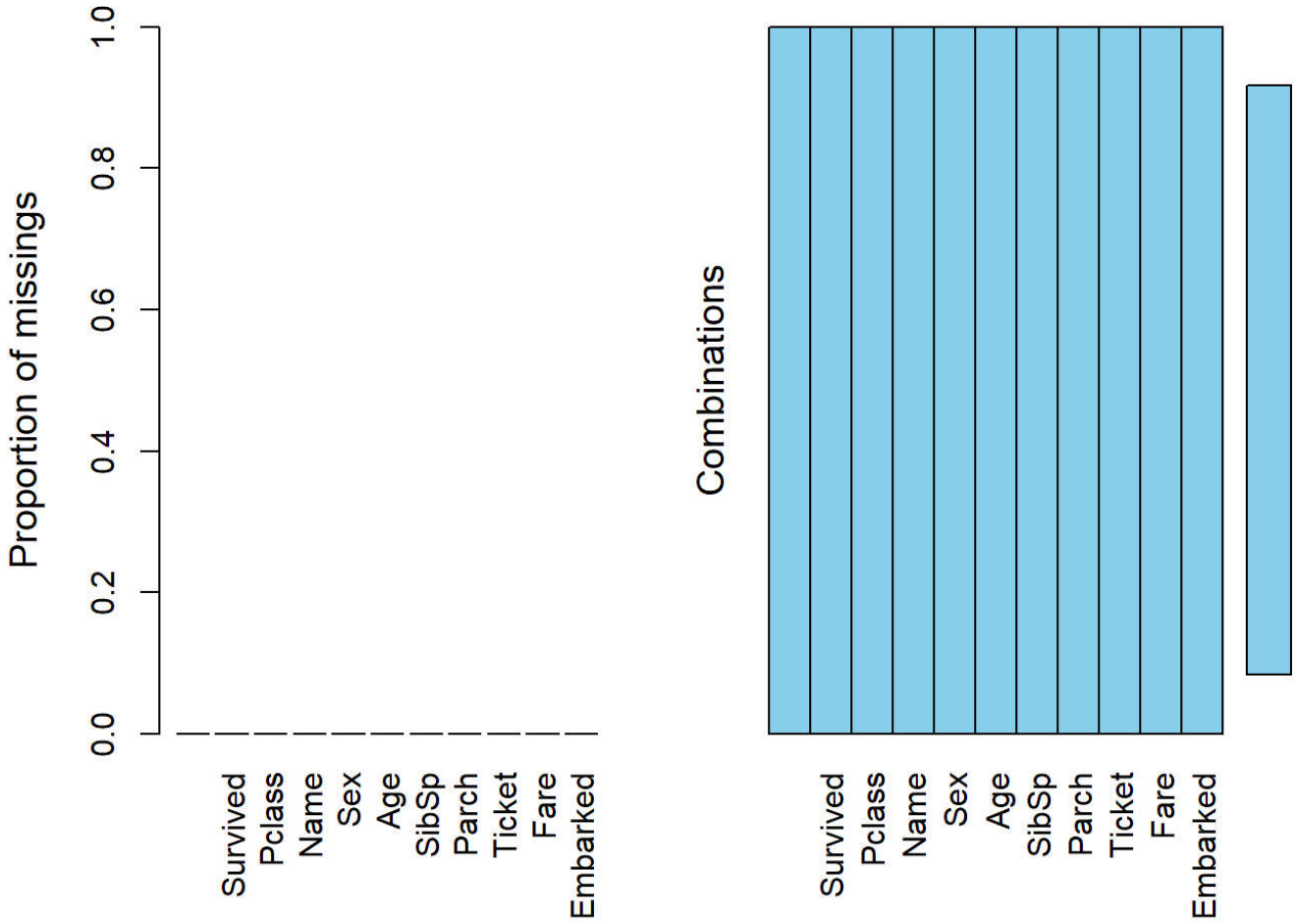
```
par(mfrow=c(1,2))
hist(Age, freq=F, main='Age: Original Data', col='darkblue', ylim=c(0,0.04))
hist(mice.output$Age, freq=F, main='Age: Mice Output', col='lightblue', ylim=c(0,0.04))
```




```
#插补后求缺失百分比
md.pattern(mice.output) #查看缺失情况
```

```
##     PassengerId Survived Pclass Age SibSp Parch Fare Ticket Name Sex
## 661           1        1      1   1     1     1    1      1    0   0
## 230           1        1      1   1     1     1    1      0    0   0
##               0        0      0   0     0     0    0    230  891 891
##     Embarked
## 661        0    3
## 230        0    4
##          891 2903
```

```
aggr(mice.output)
```



1.4相关性探索

```
library(corrplot)
trainCor=select(train,-c(PassengerId,Name,Age,Sex,Ticket,Cabin,Embarked))
str(trainCor)
```

```
## 'data.frame':    891 obs. of  5 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

```
descrCor = cor(trainCor) #数值变量相关性探讨
corrplot(descrCor,order="FPC",addCoef.col = "grey")
```

2.建模预测

2.1交叉验证

```
control <-trainControl(method = "cv",number = 10)
metric="Accuracy"
detach(train)
attach(mice.output)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     Embarked
```

```
Survived <- factor(Survived)
```

2.2构建模型

随机森林算法

```
set.seed(123)
rf.model<-train(factor(Survived)~ Pclass + Sex  + SibSp + Parch  + Embarked,data=mice.output,method="rf",trControl=control,metric=metric)  #test$Age有缺失值未纳入模型导致预测精度低
prediction <- predict(rf.model, test)
model_output<-read.csv("F:/Code/Titanic/gender_submission.csv")
model_output$Surived <-prediction  # 76.555%
```

3.彩蛋：可视化数据挖掘工具rattel安装

```
#install.packages(c("RGtk2","rattle","RGtk2Extras"))
library(RGtk2)
library(rattle)
```

```
## Rattle: A free graphical interface for data mining with R.
## XXXX 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## 键入'rattle()'去轻摇、晃动、翻滚你的数据。
```

```
rattle() #启动rattle
```

第一次运行报错“Error in method(obj, …) : Invalid root element: ‘requires’” 这主要是RGtk2的问题，install.packages(“RGtk2”) #自动下载安装的RGtk2版本为2.20.33 需要版本降级为2.20.31，更换方式 https://cran.r-project.org/web/packages/RGtk2/index.html 上下载后2.20.31版本，解压到 “我的R安装目标-3.4.1_temp2”，重启Rstudio后RGtk2版本降为2.20.31。