泰坦尼克号生存预测模型分析报告

完成人：袁泓博

引言

本项目基于历史上著名的泰坦尼克号沉船事故数据，通过统计分析和机器学习方法建立了一个预测乘客生存可能性的模型。通过对数据的详细预处理、分析和模型建立，旨在回答“什么样的人更有可能存活下来？”这一问题。

一、数据导入与预处理

数据集包括891名乘客的信息，如社会经济状态、性别、年龄等。首先读取数据集：

```
titanic_data <- read.csv("titanic.csv")
```

然后对数据进行预处理，在该阶段，首先对数据类型进行了转换将其转换为因子和数字方便供后续分析处理：

```
#转换数据类型
titanic_data$Survived<-as.factor(titanic_data$Survived)
```

```
titanic_data$Sex<-as.factor(titanic_data$Sex)
titanic_data$SibSp<-as.numeric(titanic_data$SibSp)
titanic_data$Parch<-as.numeric(titanic_data$Parch)
```

然后我们需要查找数据中的缺失值和空值：

```
#查找缺失值
sapply(titanic_data, function(x) sum(is.na(x)))
#查找空值
sapply(titanic_data, function(x) sum(x==""))
```

发现Age变量有177个缺失值；Cabin有687个空值，Embarked有2个空值。

(1).处理Embarked空值

处理策略：根据客舱等级与费用推测两个乘客的进港口

```
#确定空值对应行数的变量Pclass,Fare
titanic_data[which(titanic_data$Embarked==""),c("Pclass","Fare")]
##     Pclass Fare
## 62       1   80
## 830      1   80
```

发现Embarked为空值第62、830行的乘客，都是乘坐一等舱，费用为80美元。接下来，将一等舱乘客的 Fare、Pclass 、Embarked筛选出来。再根据乘坐一等舱从不同的进港口的平均费用多少推测出两个空值的对应进港口。

```
filter1<-
titanic_data[which(titanic_data$Pclass=="1"),c("Fare","Pclass","Embarked")]
S_fare<-median(filter1$Fare[which(filter1$Embarked=="S")])
C_fare<-median(filter1$Fare[which(filter1$Embarked=="C")])
Q_fare<-median(filter1$Fare[which(filter1$Embarked=="Q")])
S_fare
## [1] 52
C_fare
## [1] 78.2667
Q_fare
## [1] 90
```

发现从C进港口的费用与80美元接近，所以将两个Embarked缺失值赋值为C

```
titanic_data$Embarked[c(62,830)]<-"C"
```

(2).处理Cabin空值

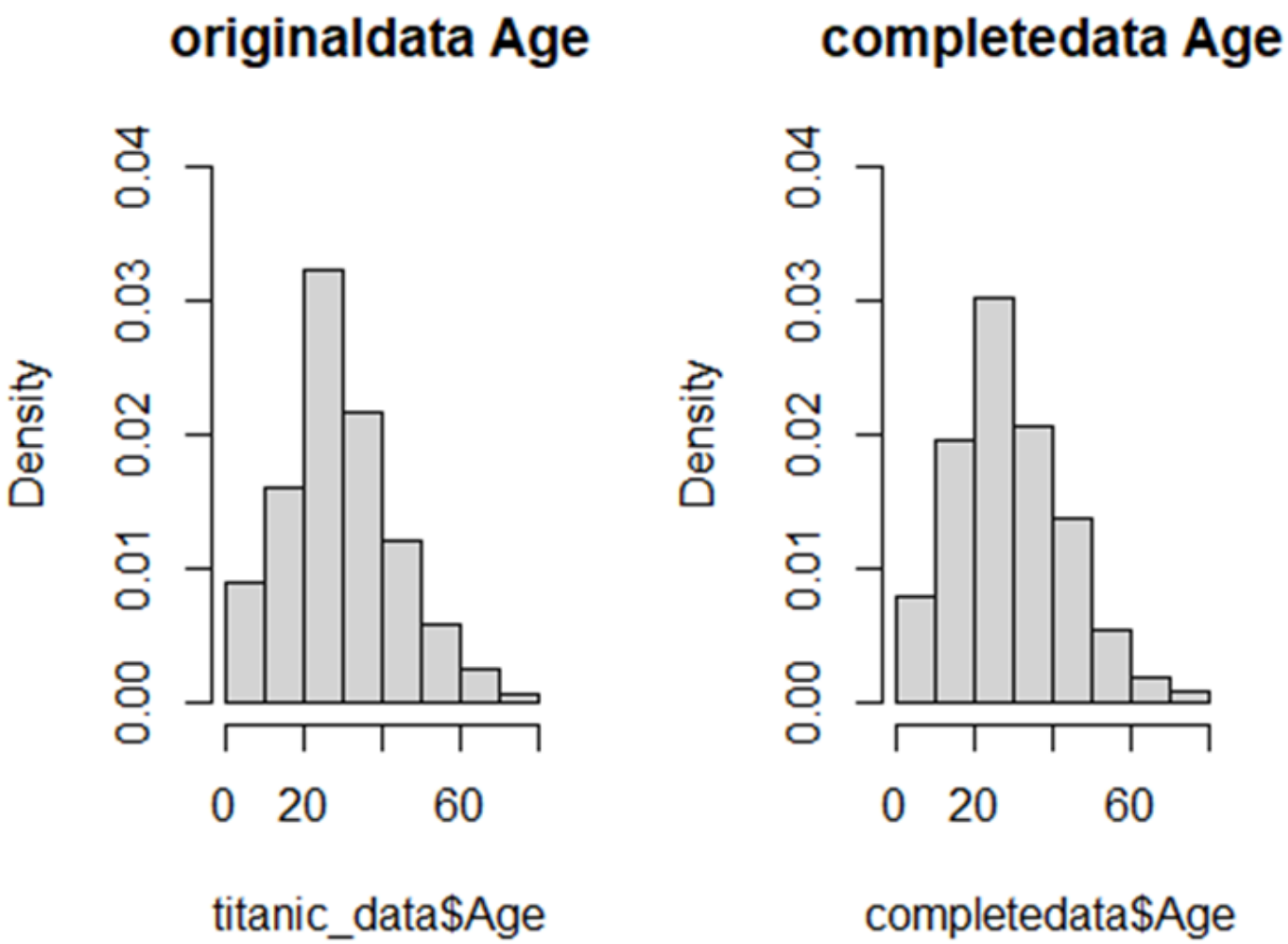
处理策略：由于总共891条数据，Cabin存在687条空值，无法代表整体数据，所以直接把该变量剔除。

```
titanic_data$Cabin<-NULL
```

(3).处理Age缺失值

处理策略：由于缺失值个数较多，运用mice包，进行系统自动根据数据特征自动填补缺失值。先利用mice()函数建模，再利用complete()函数生成完整的数据。对 Age 变量的177个缺失值，采用 mice 包的多重插补方法，以保留数据的完整性和代表性：

```
#插补缺失值
tempdata<-
mice(titanic_data[,!names(titanic_data)%in%c('PassengerId','Ticket','Survived',
'Name')],method ="pmm")
#作图比较数据特征
par(mfrow=c(1,2))  #设置图形的放置
hist(titanic_data$Age,freq = F,main='originaldata Age',ylim=c(0,0.04))
hist(completedata$Age,freq = F,main='completedata Age',ylim=c(0,0.04) )
```



通过图形可以发现，原先的full中Age的密度分布图和填补过后的completefull中Age的密度分布图没有多大差异，所以该填补缺失值是成功的。

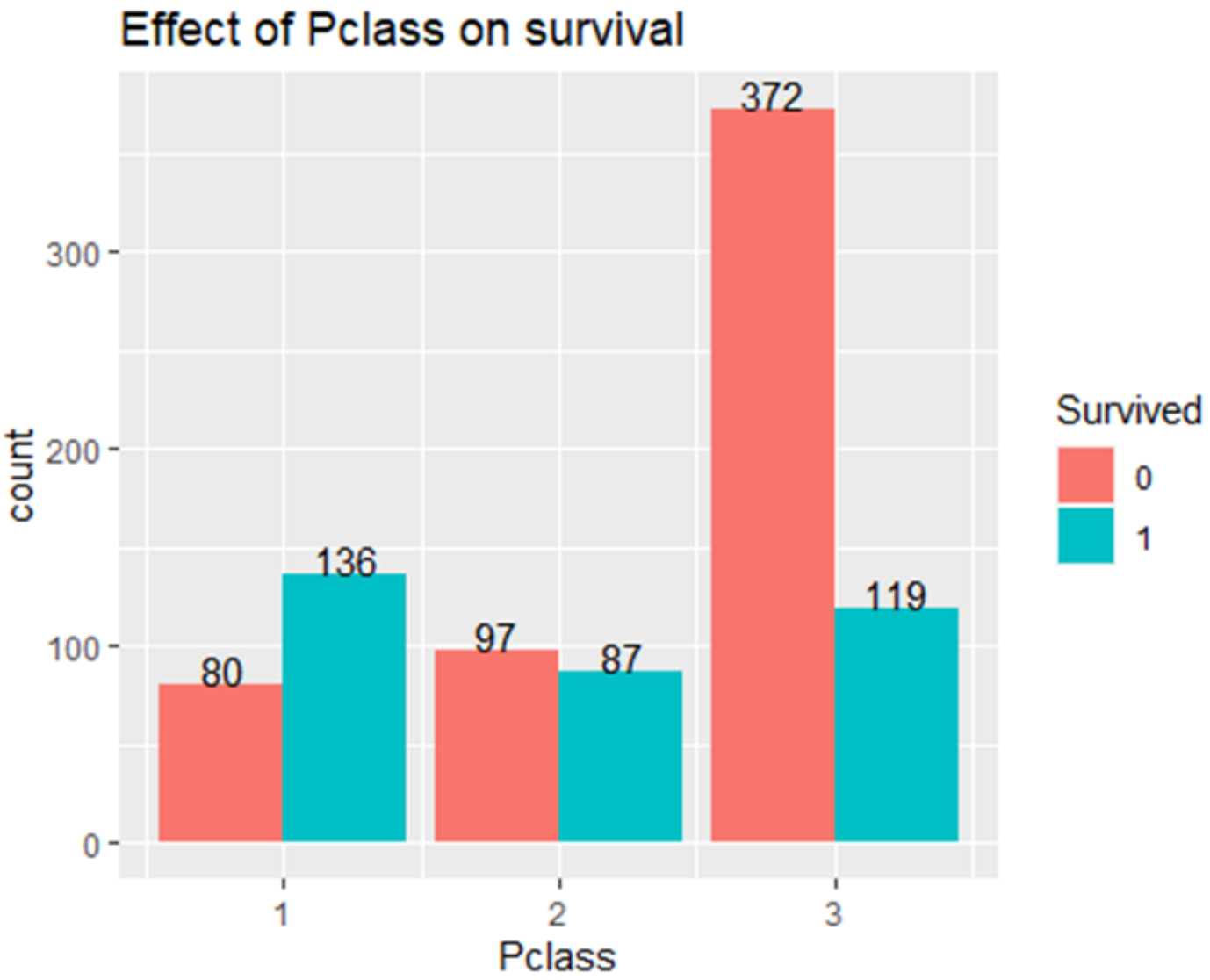
```
titanic_data$Age<-completedata$Age
```

二、数据分析

(1) 舱位等级（Pclass）

通过条形图直观的对比分析（后续分析只需要把x换成对应需要分析的因子）：

```
ggplot(titanic_data,mapping = aes(x=Pclass,y=after_stat(count),fill=Survived))+
  geom_bar(stat = "count",position = "dodge")+
  labs(title="Effect of Pclass on survival",x="Pclass",y="count")+
  geom_text(stat="count",aes(label=after_stat(count)),position
=position_dodge(0.9),vjust=0)
```



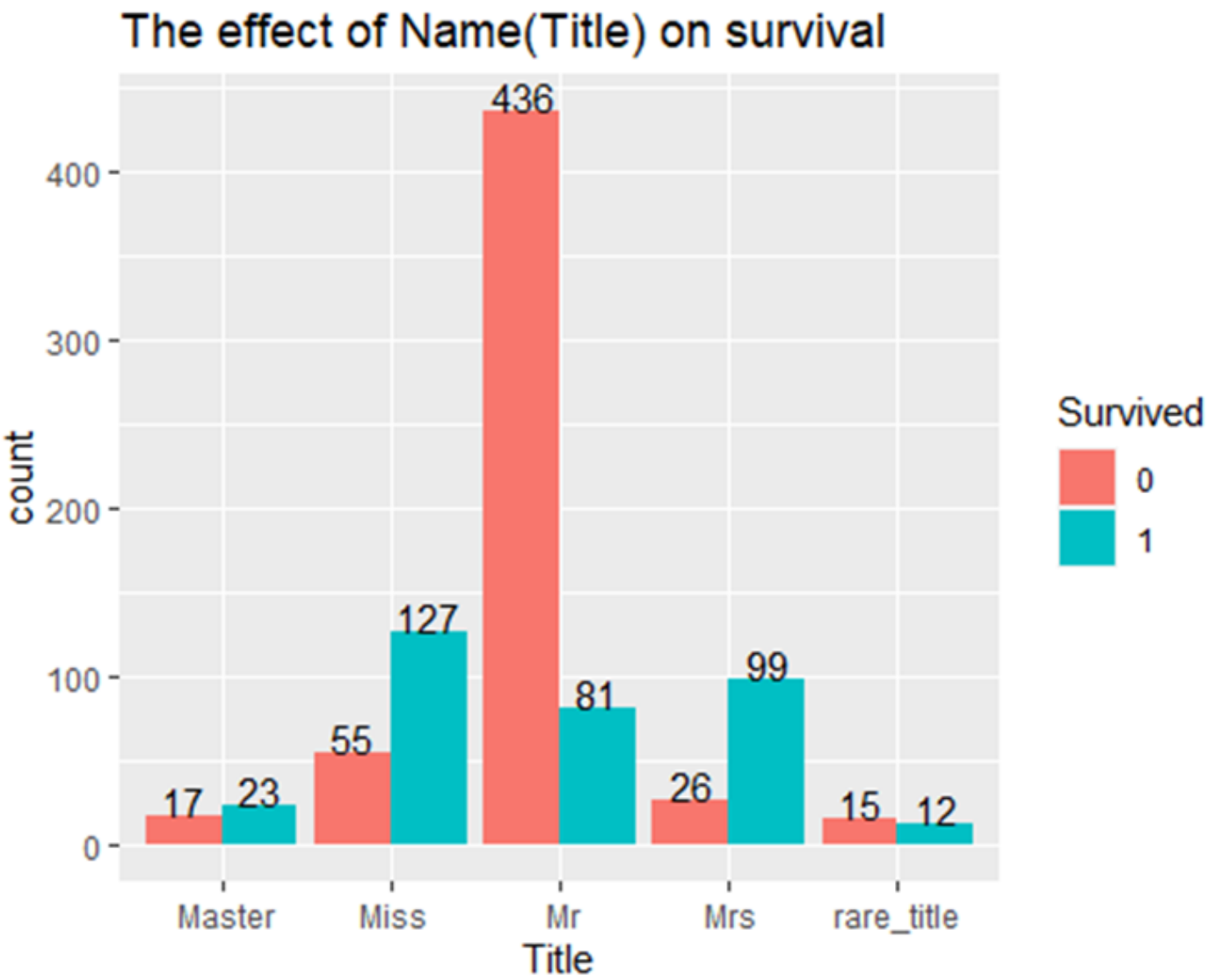
分析显示一等舱乘客的生存率最高（62.96%），其次是二等舱（47.28%），而三等舱生存率最低（24.24%）。所以，可以发现舱位等级越高，生存率越高。

(2) 称呼（Title）

Name中含有Mrs，Miss等称呼，可以作为头衔标签对其讨论

```
titanic_data$Title<-gsub("(.*,)|(\\..*)","",titanic_data$Name) #利用正则表达式把姓名中的头衔分列出来
rare<-names(which(table(titanic_data$Title)<10))#由于较少的头衔不具有代表性，所以将头衔个数小于10的筛选出来
titanic_data$Title[titanic_data$Title %in% rare]<-"rare_title" #同一头衔个数小于10的，全部改为rare_title
titanic_data$Title<-as.factor(titanic_data$Title)
summary(titanic_data$Title)
##     Master       Miss         Mr        Mrs rare_title
```

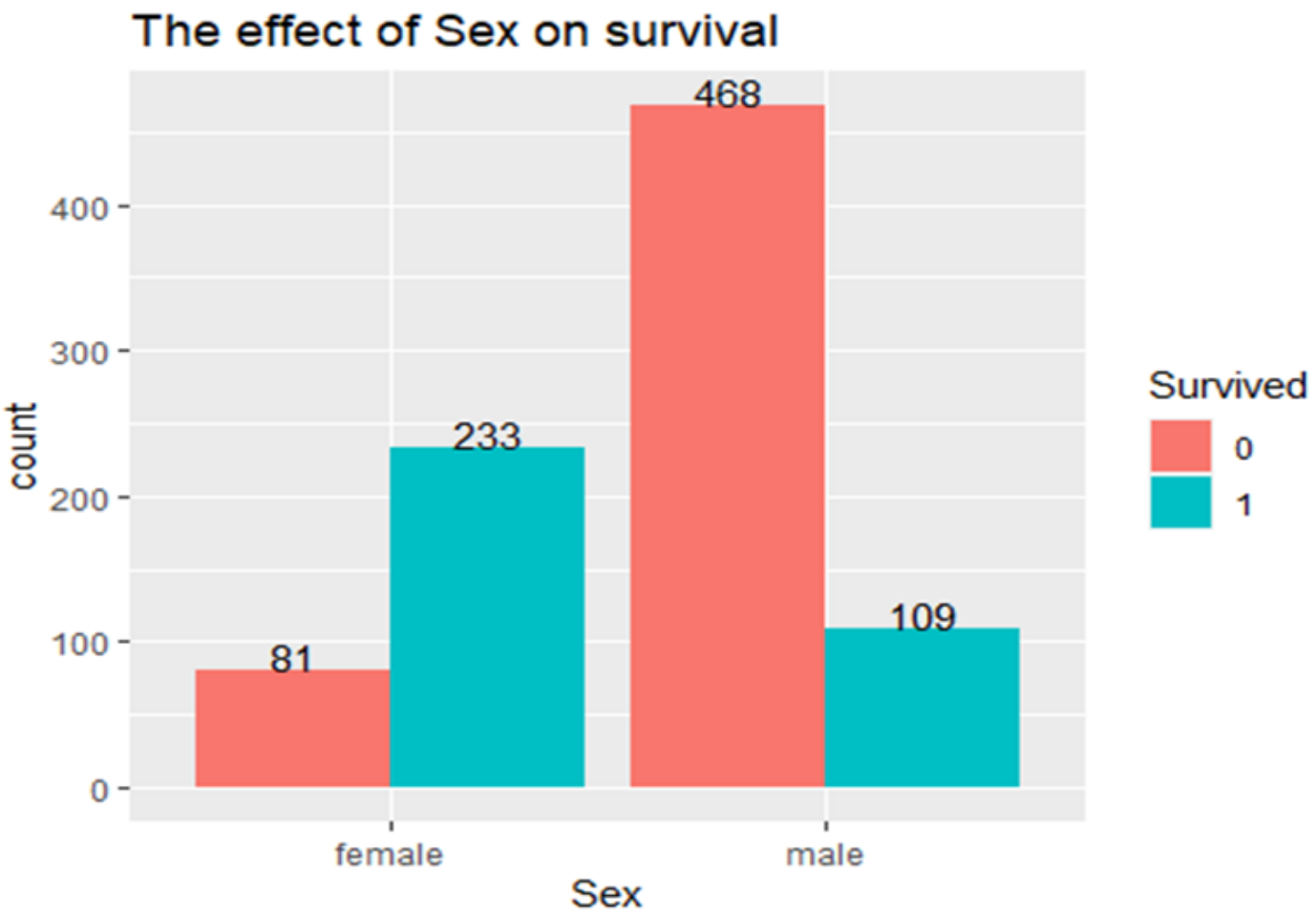
```
7 ##         40        182        517        125         27
```



从乘客的姓名中提取出的称呼显示，Miss 和 Mrs 称呼的乘客生存率较高，而 Mr 称呼的乘客生存率最低。

(3) 性别（Sex）

与上述因素同样使用条形图分析：

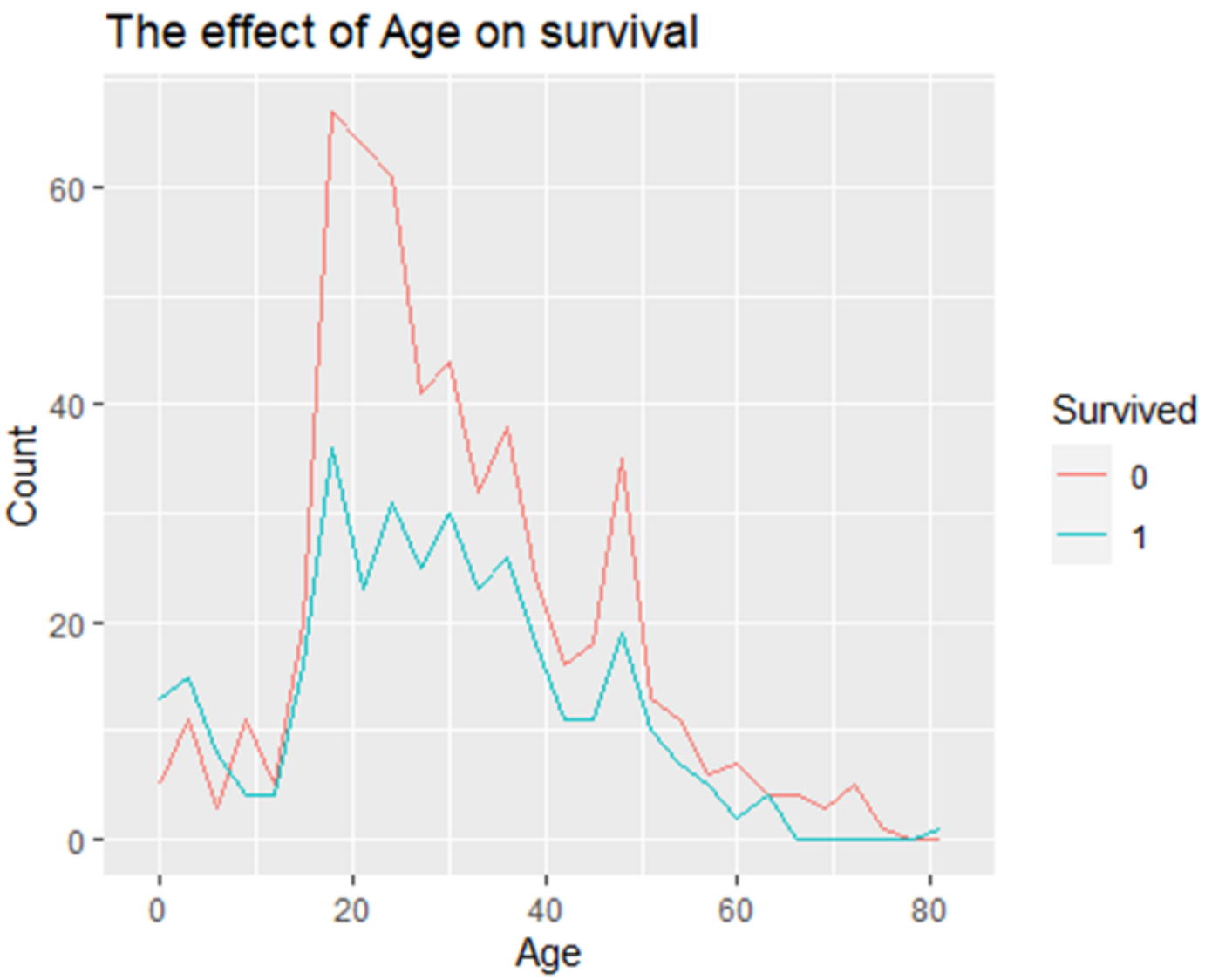


发现女性乘客的生存率（74.20%）显著高于男性（18.89%）。

(4) 年龄折线图（Age）

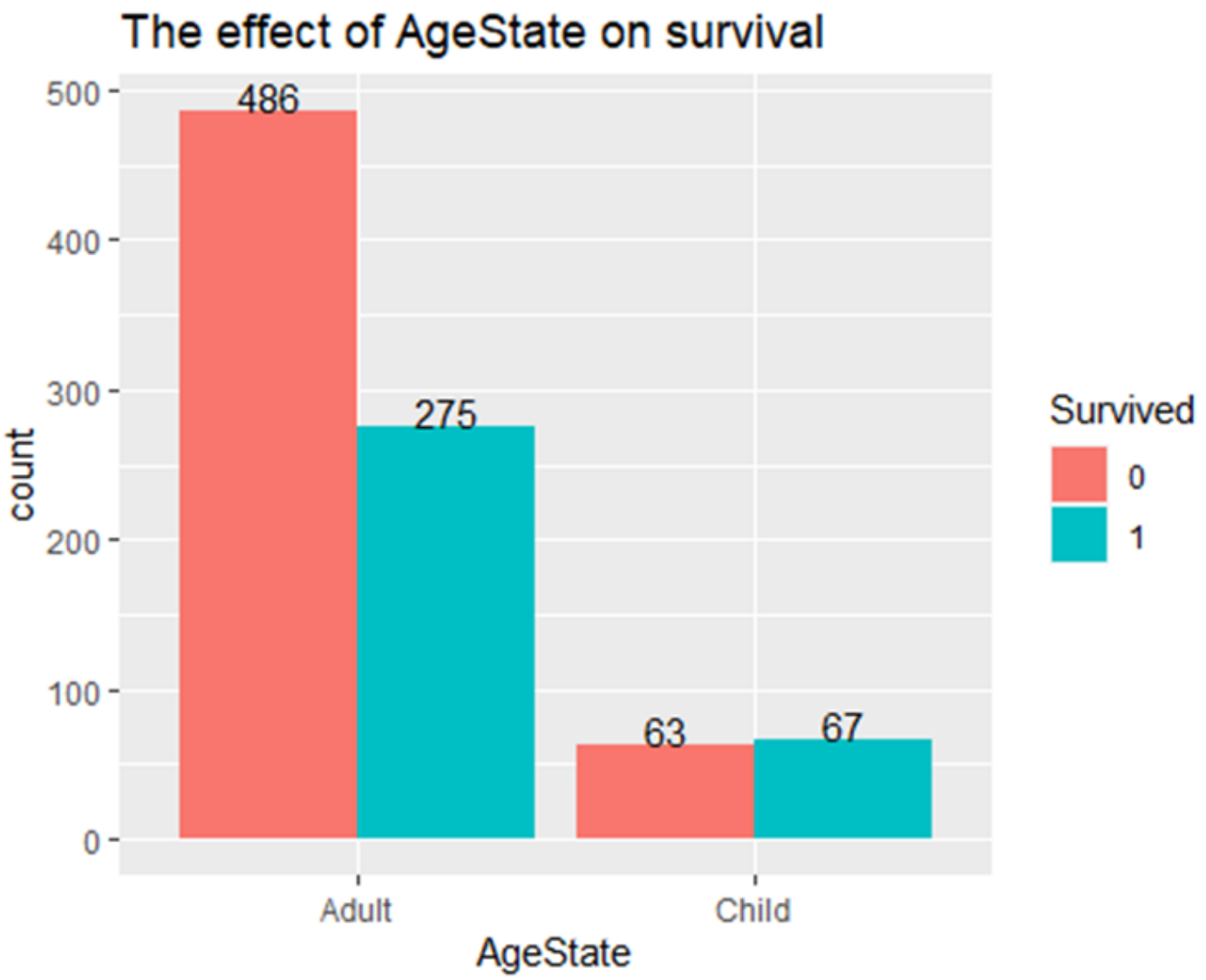
年龄做为较为连续的变量可以使用折线图对生存率分析：

```
ggplot(titanic_data,mapping = aes(x=Age,y=after_stat(count),color=Survived))+
    geom_line(aes(label=after_stat(count)), stat = 'bin', binwidth=3)  +
    labs(title = "The effect of Age on survival", x = "Age", y = "Count")
```



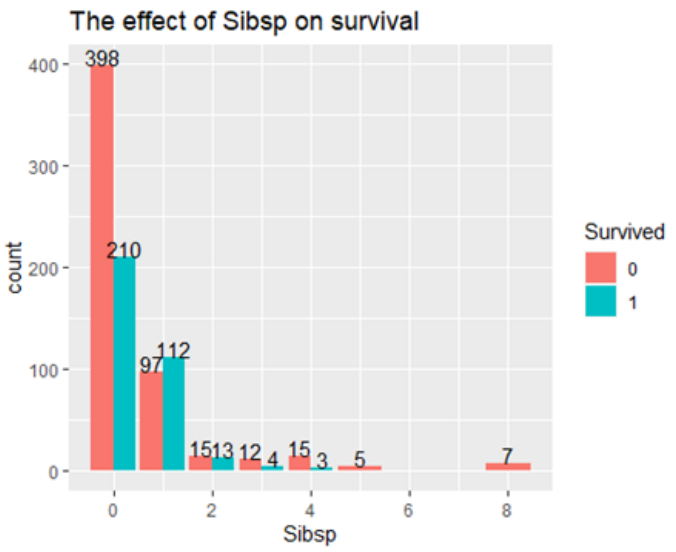
观察折线图可以发现年龄对生存率有影响，未成年人的生存率高于成年人，为了验证可以划分为未成年人和成年人再画条形图进行比较：

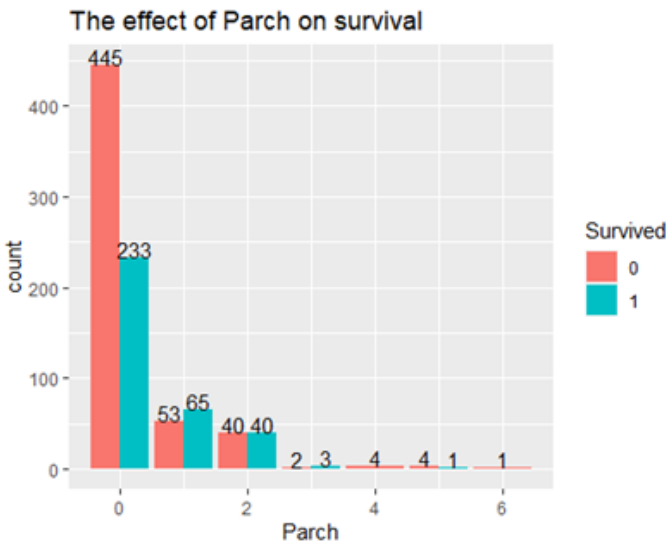
```
titanic_data$AgeState[titanic_data$Age<18]<-"Child"
titanic_data$AgeState[titanic_data$Age>=18]<-"Adult"
```



证明未成年幸存率高于成年人。

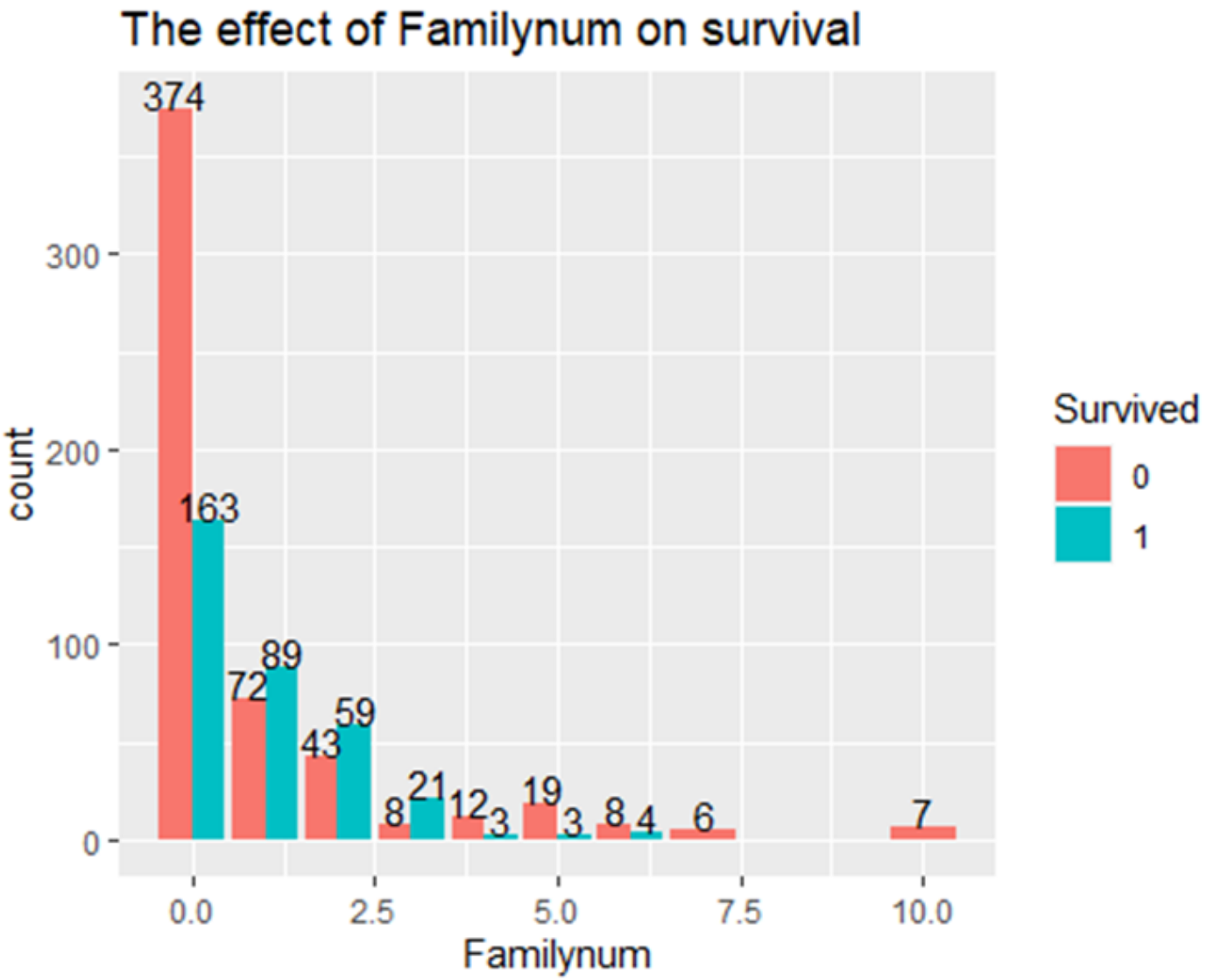
(5) 家庭成员（SibSp 和 Parch）





发现家庭成员数量与存活率相关，为了进一步验证建立新变量家属个数Familynum：

```
titanic_data$Familynum<-titanic_data$SibSp+titanic_data$Parch
```

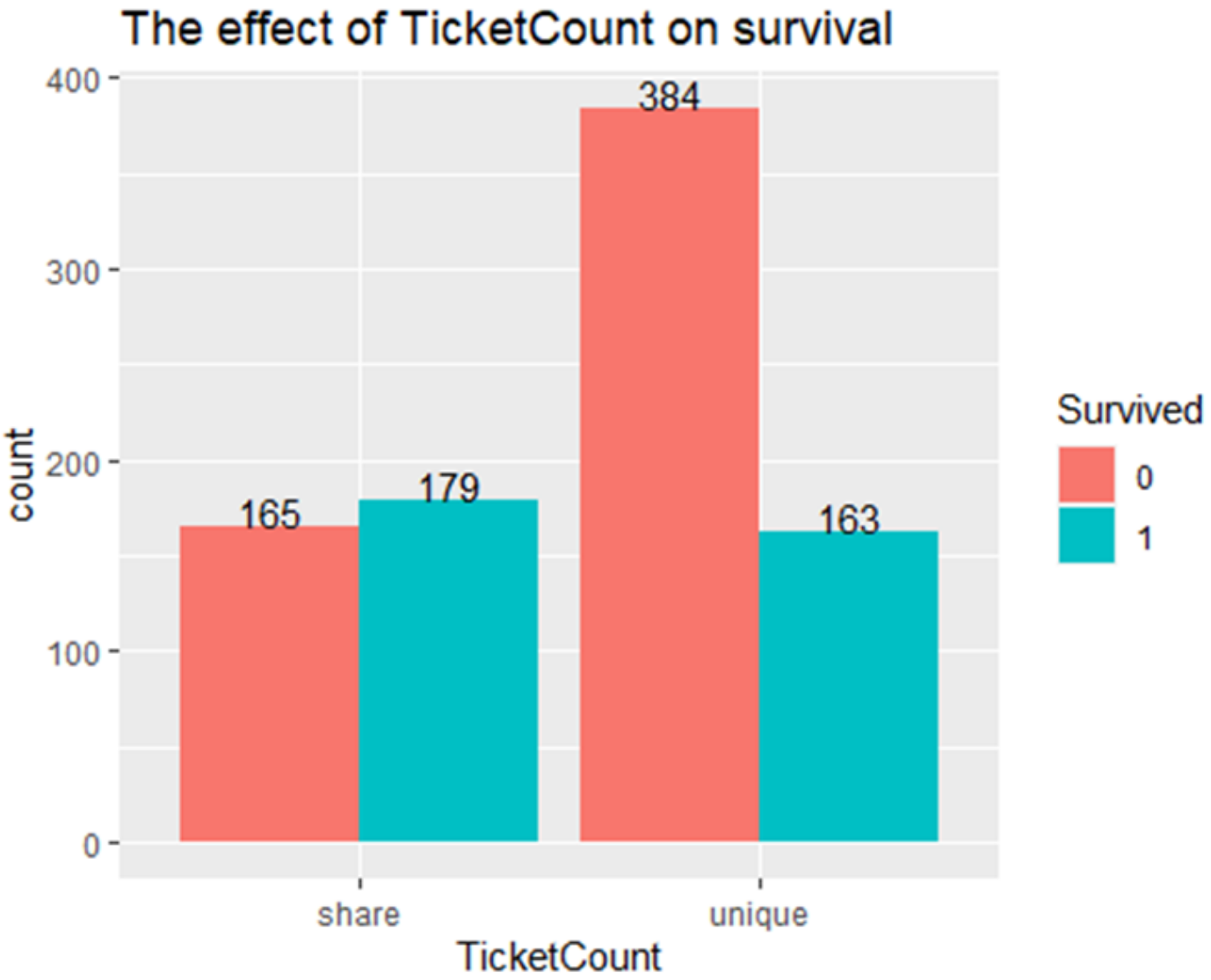


最后发现家庭成员数量在1-3个的乘客生存率较高，家属个数为零或者大于3个的幸存率较低。

(6) 船票（Ticket）

由于船票号重复率低，所以研究使用共同船票和单独船票的乘客与生存率的关系：

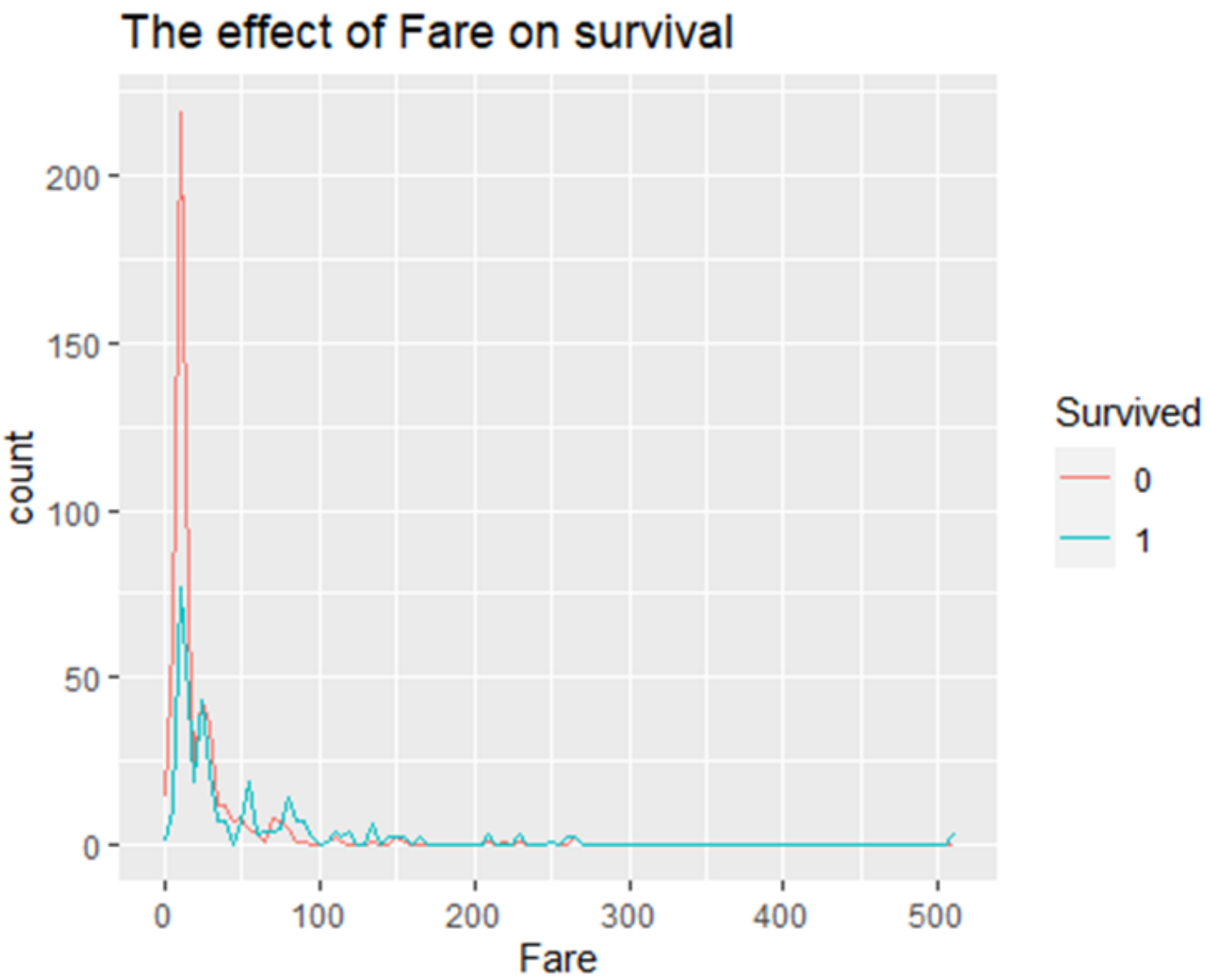
```
ticket.count<-
aggregate(titanic_data$Ticket,by=list(titanic_data$Ticket),function(x)
sum(!is.na(x)))
titanic_data$TicketCount <- apply(titanic_data, 1, function(x)
ticket.count[which(ticket.count[, 1] == x['Ticket']), 2])
titanic_data$TicketCount<-factor(sapply(titanic_data$TicketCount,function(x)
ifelse(x>1,"share","unique")))
```



发现用船票的乘客生存率（52.03%）高于独立船票的乘客（29.80%）。

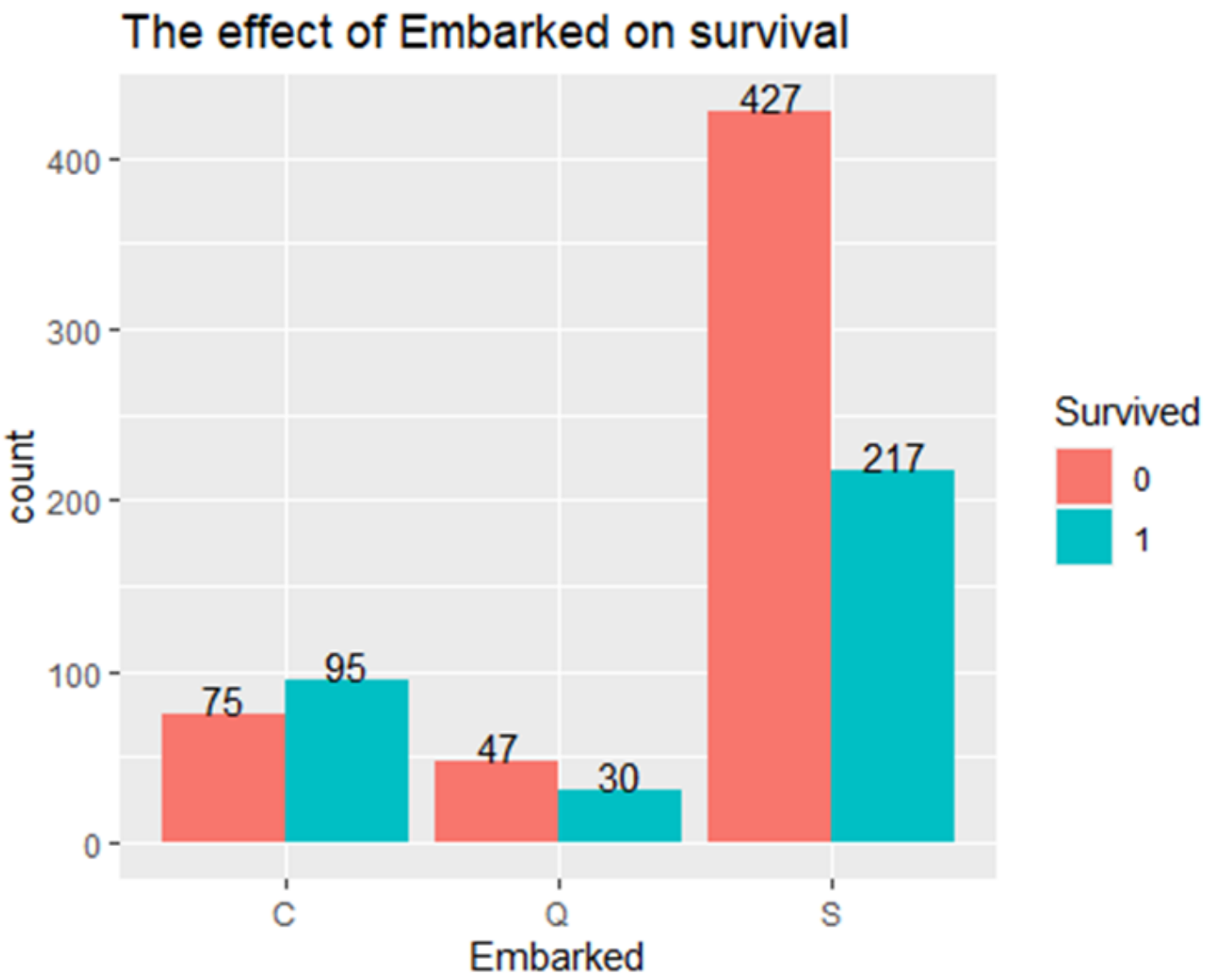
(7) 票价（Fare）

票价也属于连续性的变量可以使用折线图：



票价较高的乘客生存率也较高。

(8) 登船港口（Embarked）



从Cherbourg港口上船的乘客生存率最高。

三、数据建模与评估

（1）建模

本作业采用逻辑回归模型进行建模，建立是否存活(Survived)的预测模型。

```
#选择特征用来建模
cols_selected<-c("Survived","Pclass","Sex","Age","Fare","Embarked","AgeState","Familynum","TicketCount")
train_data<-titanic_data[,colnames(titanic_data) %in% cols_selected]
#建模
fit1=glm(Survived~.,data=train_data,family =binomial())
summary(fit1)
##
## Call:
## glm(formula = Survived ~ ., family = binomial(), data = train_data)
```

```
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         5.105985   0.625640   8.161 3.32e-16 ***
## Pclass             -1.097467   0.148155  -7.408 1.29e-13 ***
## Sexmale            -2.738730   0.203604 -13.451  < 2e-16 ***
## Age                -0.025646   0.008461  -3.031  0.00244 **
## Fare                0.001555   0.002464   0.631  0.52785
## EmbarkedQ           0.005782   0.389810   0.015  0.98817
## EmbarkedS          -0.375288   0.240751  -1.559  0.11904
## AgeStateChild       0.847040   0.351512   2.410  0.01597 *
## Familynum          -0.368743   0.083171  -4.434 9.27e-06 ***
## TicketCountunique -0.387387   0.235149  -1.647  0.09947 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  775.31  on 881  degrees of freedom
## AIC: 795.31
##
## Number of Fisher Scoring iterations: 5
```

根据第一次建模进行调整，发现Pclass、Sex、Age、Familynum 和 TicketCount 与存活率的相关性高，故选取了 Pclass、Sex、Age、Familynum 和 TicketCount 特征再次进行建模。

```
#经过调整模型变量
fit2=glm(Survived~Pclass+Sex+Age+Familynum+TicketCount,data=train_data,family=binomial())

summary(fit2)
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Familynum + TicketCount,
##     family = binomial(), data = train_data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         5.475725   0.500098  10.949  < 2e-16 ***
## Pclass             -1.157482   0.126444  -9.154  < 2e-16 ***
## Sexmale            -2.740682   0.199603 -13.731  < 2e-16 ***
## Age                -0.036211   0.007326  -4.943 7.69e-07 ***
## Familynum          -0.321745   0.077005  -4.178 2.94e-05 ***
## TicketCountunique -0.492365   0.223636  -2.202   0.0277 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  785.92  on 885  degrees of freedom
## AIC: 797.92
##
## Number of Fisher Scoring iterations: 5
```

（2）模型质量评估

```
#30%的训练集数据用来验证模型
set.seed(123)
par <- createDataPartition(titanic_data$Survived,times = 1,p=0.3,list = F)
test_data <- titanic_data[par,]
####预测准确率
pred <- predict(fit2,test_data[1:nrow(test_data),],type = "response")
fittedresults<-ifelse(pred>0.5,1,0);
Error<-sum(factor(fittedresults)!=test_data$Survived)/nrow(test_data)
Error
## [1] 0.1940299
#计算模型系数的指数，即各变量的几率比。
exp(coef(fit2))
##       (Intercept)            Pclass           Sexmale               Age
##      238.82354247        0.31427655        0.06452635        0.96443656
##         Familynum TicketCountunique
##         0.72488314        0.61117933
```

模型分析表明，舱位等级、性别、年龄和家庭成员数量是影响生存率的重要因素。且预测错误率为20%左右，也就是说80%的记录均预测正确，显示了较高的预测效力。

四、结论

综上所述，泰坦尼克号上生还率较高的群体特征为：一等舱乘客、女性、年轻乘客和家庭成员数量在1-3个的乘客。此外，从Cherbourg港口上船和购买较高票价的乘客也显示出较高的生存率。