# 数据挖掘作业2

**数据集：**<https://www.kaggle.com/c/titanic/download/train.csv>

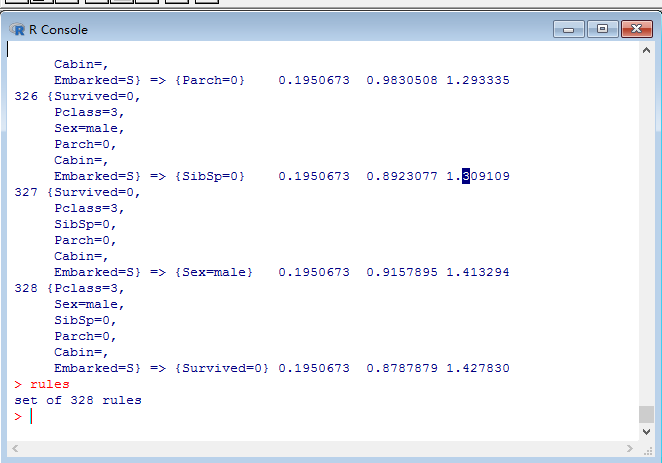
**数据集预处理说明：**因为用的是R语言编程工具，其中她的库arules的apriori算法不支持数字，但是网站给的数据集中1、2、3、6、7、8、10列中是数字格式，为了转换成字符，我在数据集的复制了一行数据，然后改成这种格式：

|  |
| --- |
| 1/,0/,3/,"Braund, Mr. Owen Harris",male,22/,1/,0/,A/5 21171,7.25/,,S |

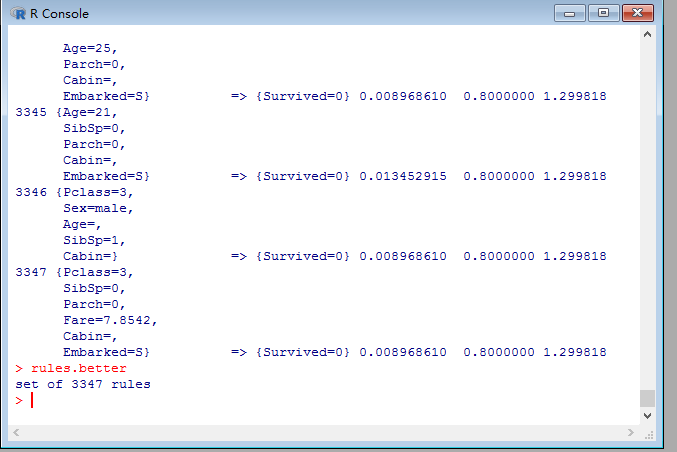
这样用R语言读取数据的时候全部是字符，可以用apriori算法来进行处理，同事也不会影响数据集原来的属性。

**环境：**win10+R语言3.3.1

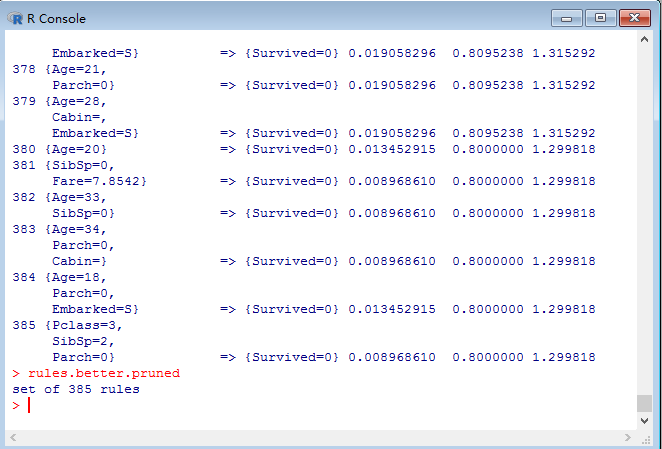
1. 首先加载库：arules，用apriori算法进行关联分析



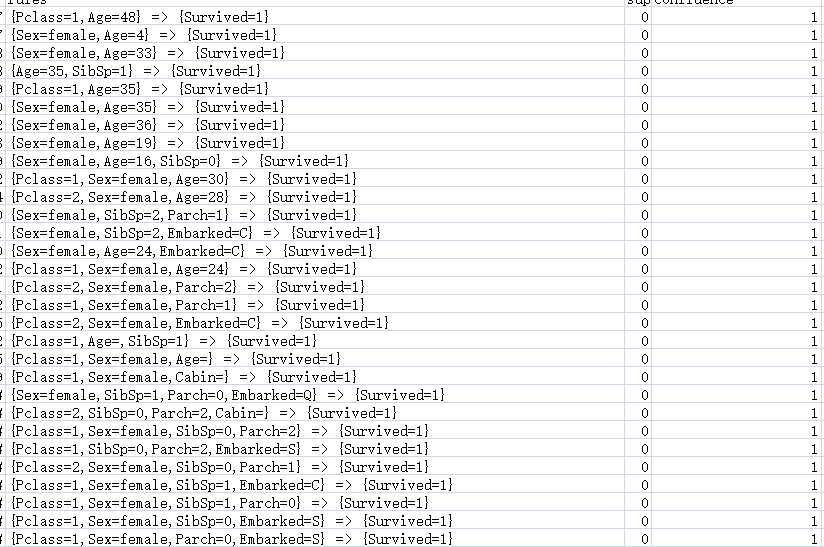
1. 然后设置条件，比如说我想看生存情况，那我的条件就设置为右边的集合rhs为rhs=c("Survived=0","Survived=1")，然后最小关联项、支持度、置信度分别设置为2、0.005、0.8，结果如图：



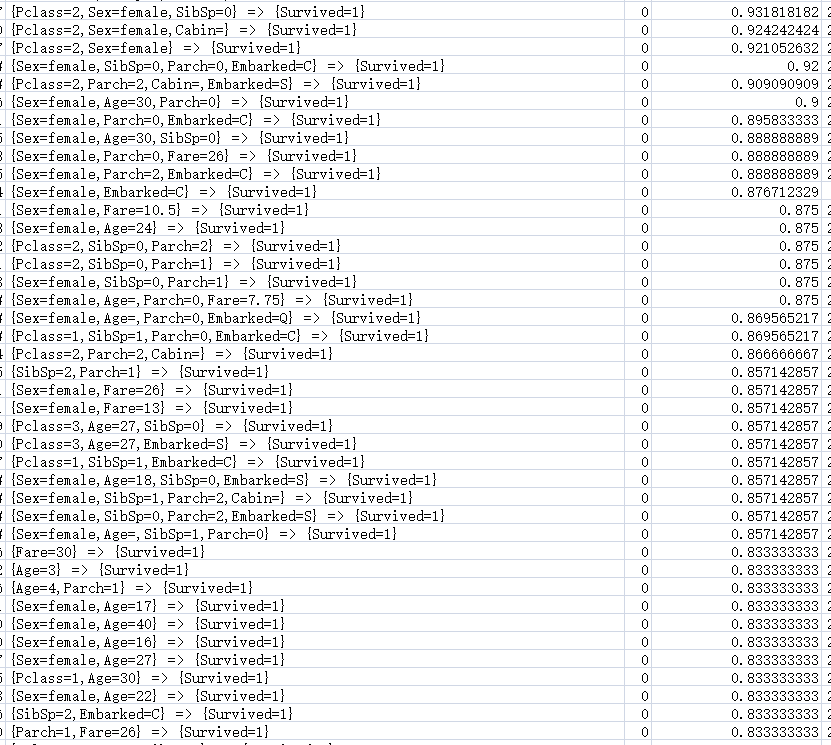
1. 然后去除冗余的规则，结果如图：



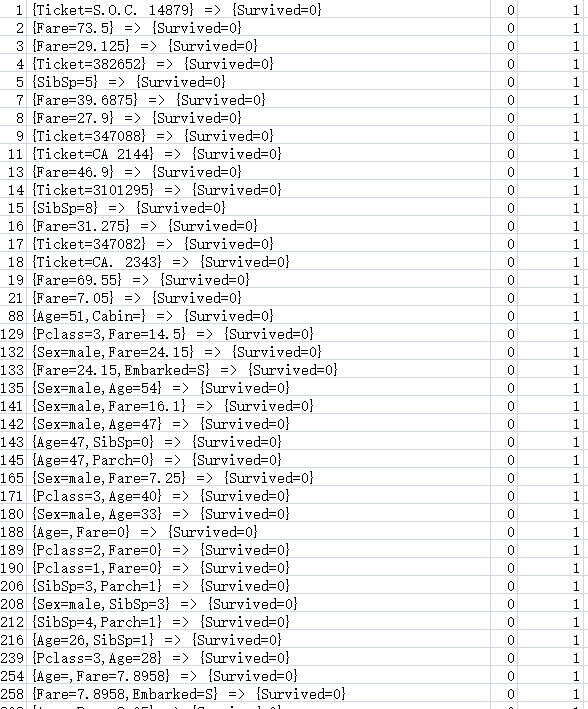
1. 对结果进行评价，因为有385条相关的规则，所以先把结果导出方便分析，从挖出的规则来分析，如图：

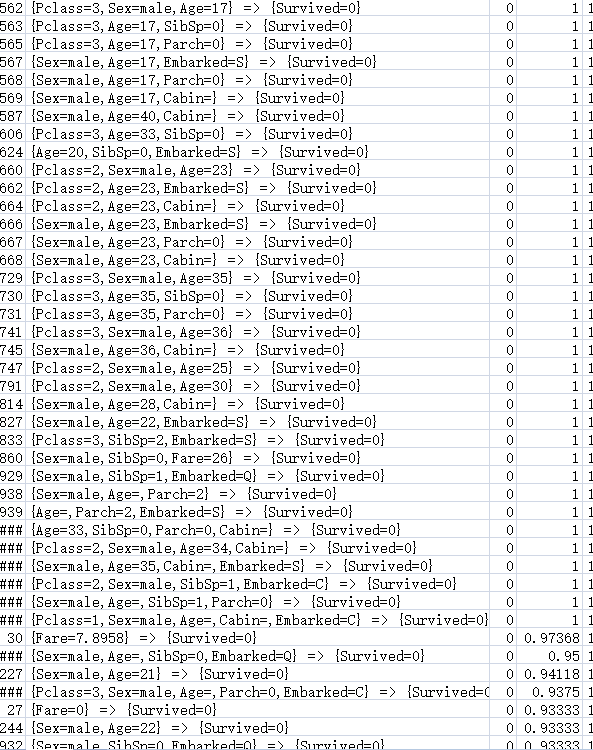


可以看出最明显的结果是置信度为1的情况下，生存下来的人的条件有这么几个特征：Pclass=1、Sex=female，通俗点讲就是头等舱的女乘客生存几率最大，再来看看下一张图片：



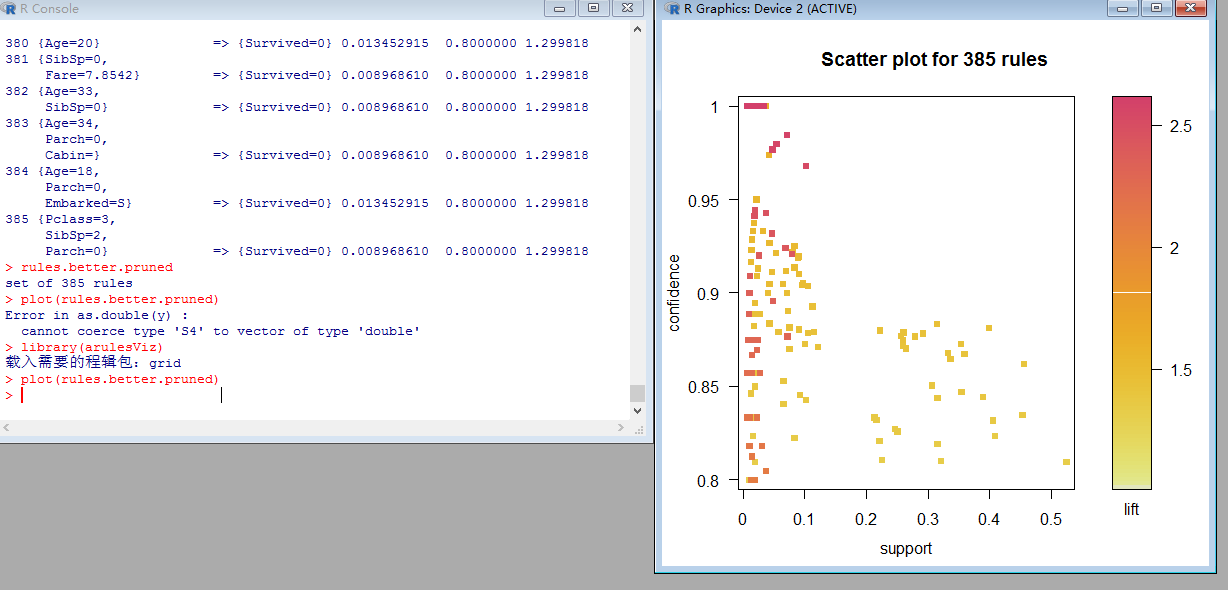
结合上一张，结果表明确实整体女性生存几率较大。接下来看看死亡者的特征：





基本上都是些2、3等仓的男性，这个故事告诉我们旅行要安全就得买头等舱。。。

1. 画图需要加载arulesViz库，结果如图：



1. 参考：

<http://cos.name/2013/02/association-rules-with-r-and-sas/>

<http://blog.csdn.net/sunbow0/article/details/41962185>