# 数据挖掘大作业二：关联规则挖掘

张霖 2120151063

## 一、数据集的选择及处理

### 1.1 数据说明

本次实验选择的数据集为Titanic数据集，数据来源为<https://www.kaggle.com/c/titanic/data>。数据集内包含PassengerId，Survived，Pclass，Name，Sex，Age，SibSp，Parch，Ticket，Fare，Cabin，Embarked十二项，它们的含义如下：

|  |  |
| --- | --- |
| 变量名 | 变量描述 |
| PassengerId | 乘客的编号 |
| Survived | 是否存活（0代表没有存活，1代表存活） |
| Pclass | 客舱等级（1,2,3分别代表一二三登记） |
| Name | 姓名 |
| Sex | 性别 |
| Age | 年龄 |
| SibSp | 兄弟姐妹或者配偶同在船上的数目 |
| Parch | 父母和孩子同在船上的数目 |
| Ticket | 船票类型 |
| Fare | 票的费用 |
| Cabin | 客舱号 |
| Embarked | 登船地点(C = Cherbourg; Q = Queenstown; S = Southampton) |

### 1.2 数据预处理

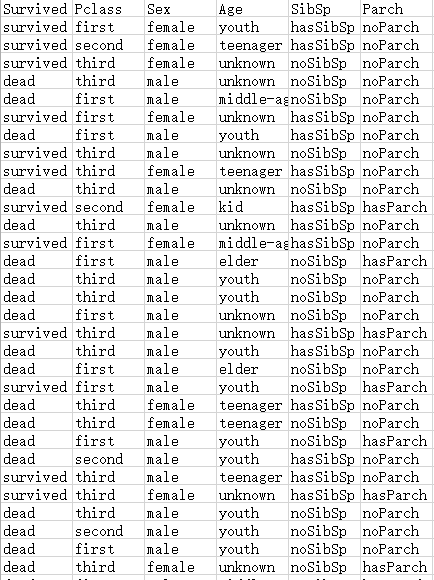
实验只选取了其中的6种类型的数据：Survived，Pclass，Sex，Age，SibSp，Parch；并使用R语言对数据做了以下处理：

1. Survived类型中，使用“survived”代替数据1，使用“dead”代替数据0。
2. Pclass类型中，使用“first”代替数据1，使用“second”代替数据2，使用“third”代替数据3。
3. Age类型中，年龄为空的为“unkown”,年龄小于10的为“kid”，大于10小于20的为“teenager”，大于20小于40的为“youth”，大于40小于60的为“middle-ager”，大于60的为“elder”。
4. SibSp类型中，其中兄弟姐妹或者配偶同在船上的为“hasSibSp”，没有的为“noSibSp”。
5. Parch类型中，其中父母和孩子同在船上的为“hasSibSp”，没有的为“noSibSp”。

代码如下：

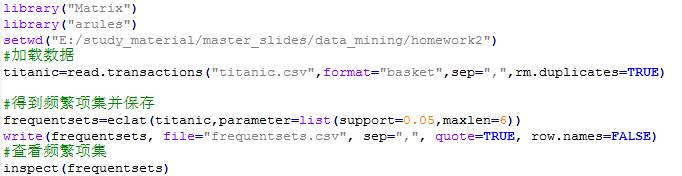


处理后的数据如下图所示：

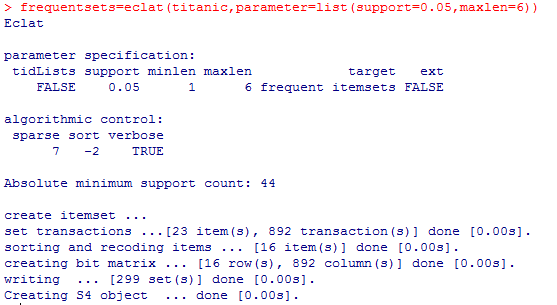


## 二、找出频繁项集

本次实验使用R语言，找出频繁项集并保存的代码如下：

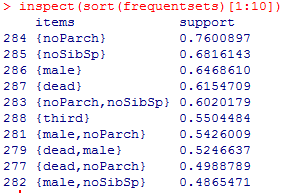


结果如下：



符合条件的共有299条频繁项。

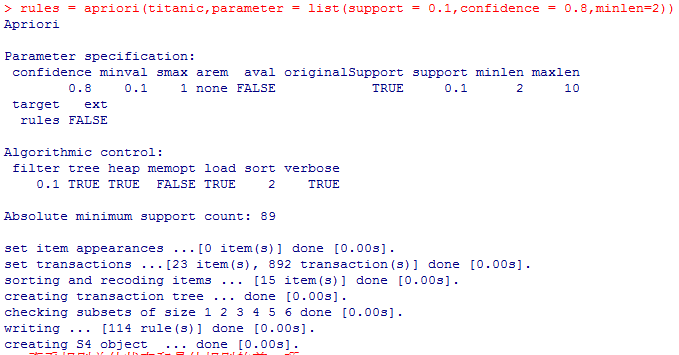
查看频繁项集的前十名，如下图：



从结果中我们可以看到父母和孩子同时在船上的较少，不足三成，兄弟姐妹或配偶同时来的也并不多，没有亲人来的大概6成，除此之外大约64%的人为男性，撞冰山之后的死亡人数超过六成等。

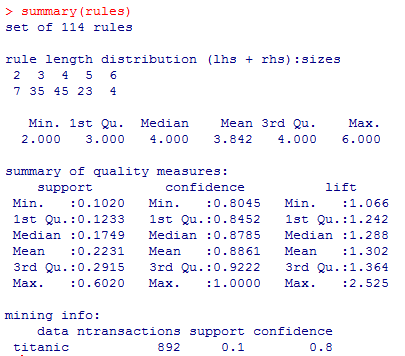
## 三、导出关联规则

### 3.1 使用apriori函数进行规则挖掘



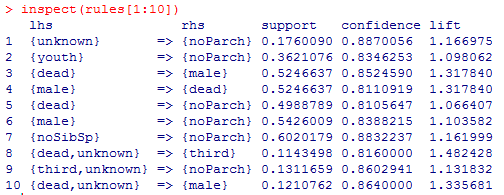
从返回的结果看，总共有114条规则生成。

### 3.2 查看规则的汇总信息



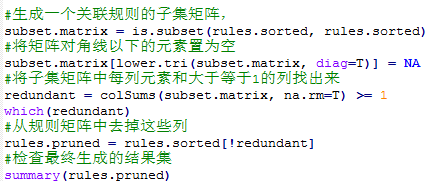
有结果可知，长度为4的规则最多。

### 3.3 查看具体关联规则

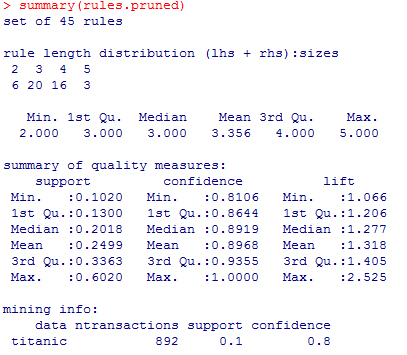


## 四、去除冗余规则

相关代码：



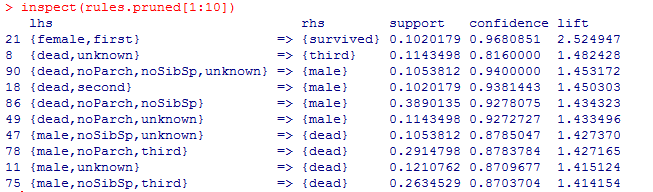
结果如下：



原规则共有114条，去除冗余规则后剩余45条规则，故可知相当多的规则是冗余的。

## 五、对规则进行评价

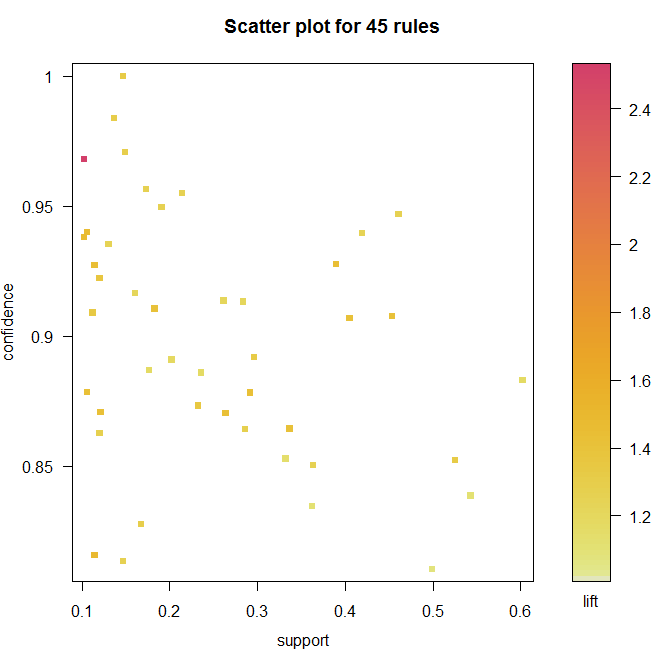
这里使用lift指标对规则进行评价，由于去除冗余规则时使用lift指标进行排序，故再次查看时不需要重新排序。



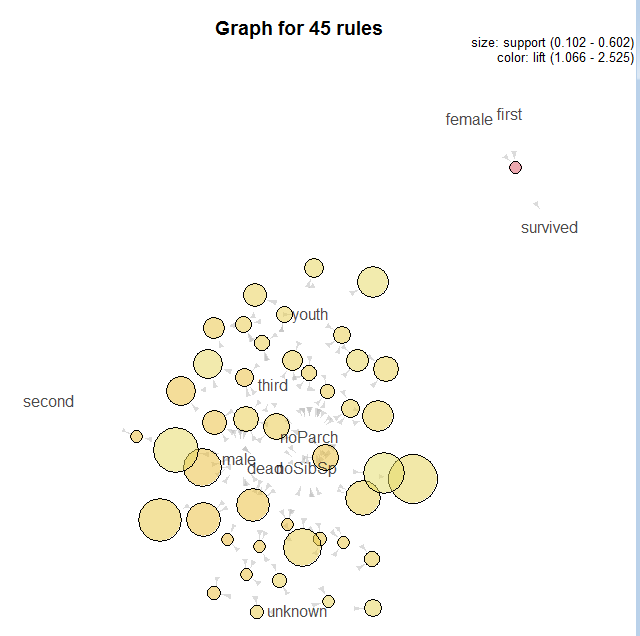
从中可以得到比较有用的规则。如，一等舱的女性有存活下来的概率很大，可能是因为她们被优先救助。而不知道年龄的死亡人员大都来自于三等舱

## 六、对规则进行展示

### 5.1 散点图



### 5.2 泡泡图



### 5.3 平行坐标

