# 数据挖掘分析报告

学号：2620170053 姓名：夏铭泽

### 1. 数据源

从大作业一的两个数据集中任选一个进行分析。

### 2. 要求

1. 对数据集进行处理，转换成适合关联规则挖掘的形式；
2. 找出频繁项集；
3. 导出关联规则，计算其支持度和置信度
4. 对规则进行评价，可使用Lift，也可以使用教材中所提及的其它指标

### 3. 提交的内容

* 对数据集进行处理的源程序
* 关联规则挖掘的源程序
* 挖掘结果及分析
* 挖掘过程的报告

### 4. 数据集处理

### 4.1数据集选取

数据集选取的是“Building\_Permits”

### 4.2 数据集处理

代码见Datapre

对于二值属性将原始数据空白部分直接设为N，对于数值属性和标称属性，选取了取值较少的几种用于挖掘。

处理结果保存在test.csv中

### 5. 数据挖掘分析

### 5.1 读取数据

tr=read.transactions("test.csv",format="basket",sep=",",rm.duplicates=TRUE)

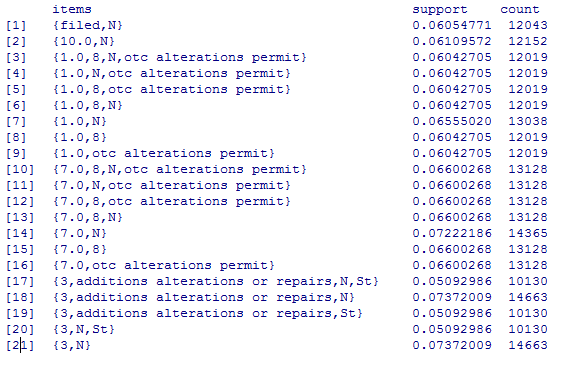
### 5.2 求频繁项集

求频繁项集

frequentsets=eclat(tr,parameter=list(support=0.05,maxlen=8))

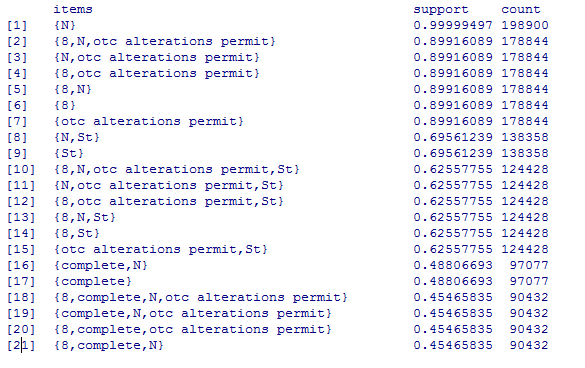
察看求得的频繁项集，结果如图：

inspect(frequentsets[1:10])



根据支持度对求得的频繁项集排序并察看，如图：

inspect(sort(frequentsets))

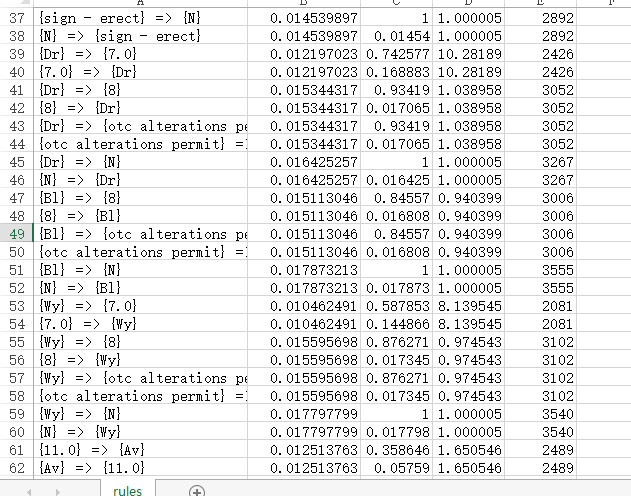


### 5.3 关联分析

求关联规则

rules=apriori(tr,parameter=list(support=0.01,confidence=0.01))

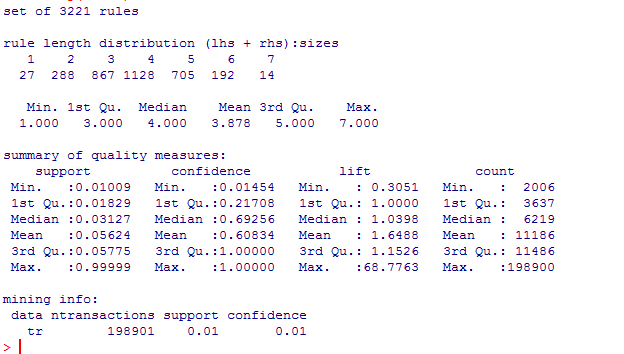
导出关联规则（rules.csv）。



write(rules, file = "rules.csv", sep = ",", quote=TRUE, row.names=TRUE)

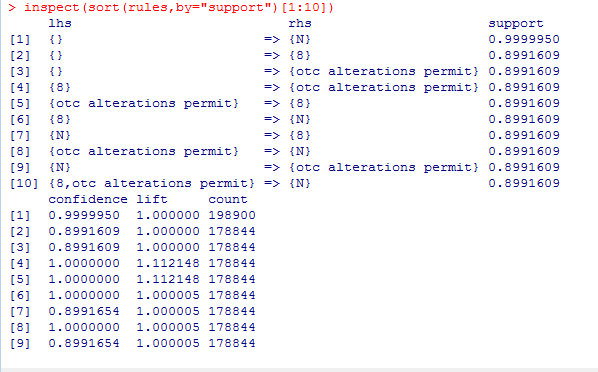
察看求得的关联规则之摘要，如图：

summary(rules)



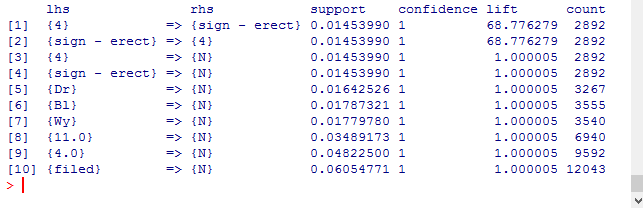
按支持度查看前10条规则，如图：

inspect(sort(rules,by="support")[1:10])



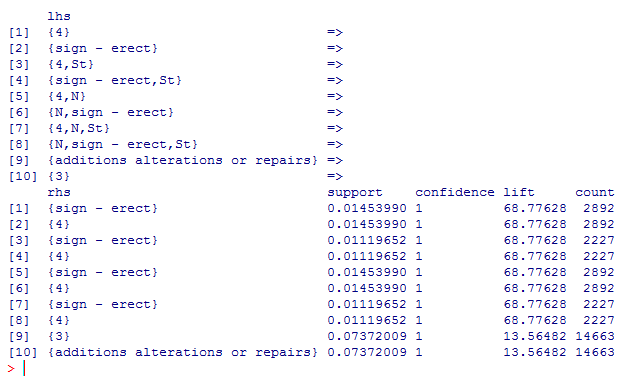
按置信度查看前10条规则，如图：

inspect(sort(rules,by="confidence")[1:10])



按lift查看前10条规则，如图：

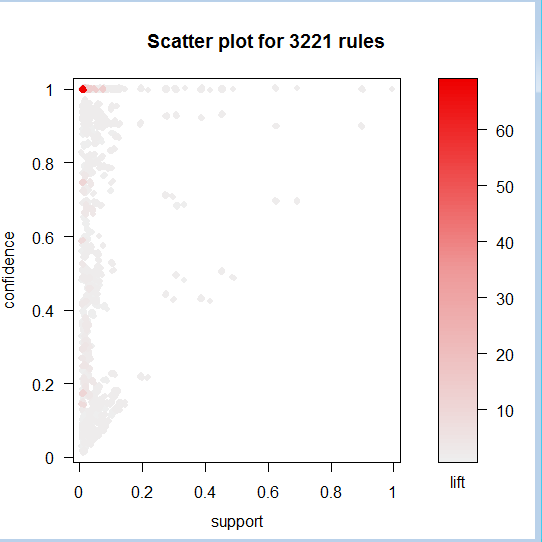
inspect(sort(rules,by="lift")[1:10])



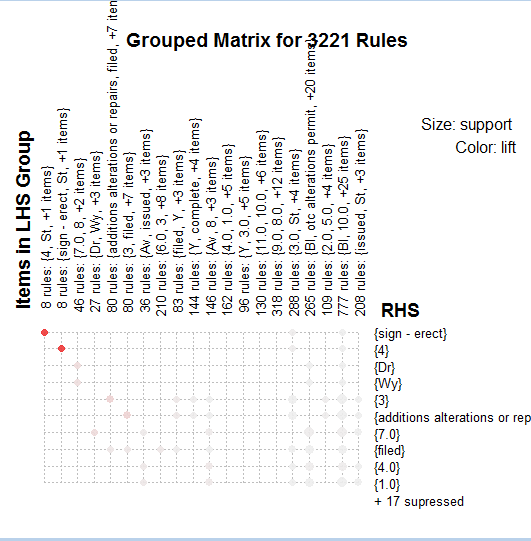
由图可知，支持度较高的一般置信度也较高，但是置信度高的可能支持度较低。

### 5.4 对规则进行可视化

使用散点图对关联规则进行可视化。图中每个点对应于相应的支持度和置信度值，分别由图形的横纵轴显示，其中关联规则点的颜色深浅由lift值的高低决定。



使用气泡图来展现关联规则，提升度lift是圈的颜色深浅，圈的大小表示支持度support的大小。LHS的个数和分组中最频繁项集显示在列的标签里。Lift从左上角到右下角逐渐减少。



也可以通过箭头和圆圈来表示关联规则，利用项点表示项集，边表示规则中关系。圆圈越大表示支持度support越大，颜色越深表示提升度lift越大。

