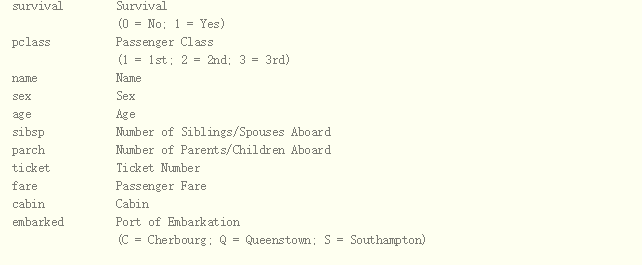
关联规则挖掘

#### 数据集

本次作业选取Titanic存活数据集作为挖掘关联规则的数据集，该数据集总共包含以下属性：

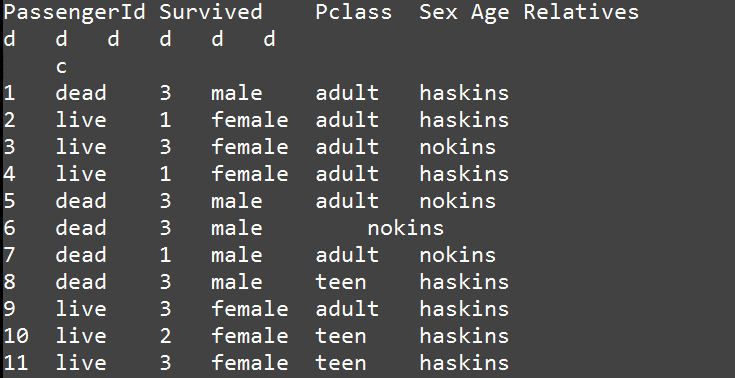


#### 数据预处理

为了方便规则的挖掘，要将数据中的连续值转换成离散值以及剔除一些无用属性，最后的到如下的属性：

1. Pclass : 乘客的级别（1/2/3）
2. Survived: 乘客存活与否（survivid/dead）
3. Sex：乘客的性别（男/女）
4. Age：乘客的年龄（<18岁：teen/18~55岁：middleage/>55岁：old ）
5. Relatives：有无亲属（haskins/nokins）

处理后的文件存在pre.tab文件下，格式如下：



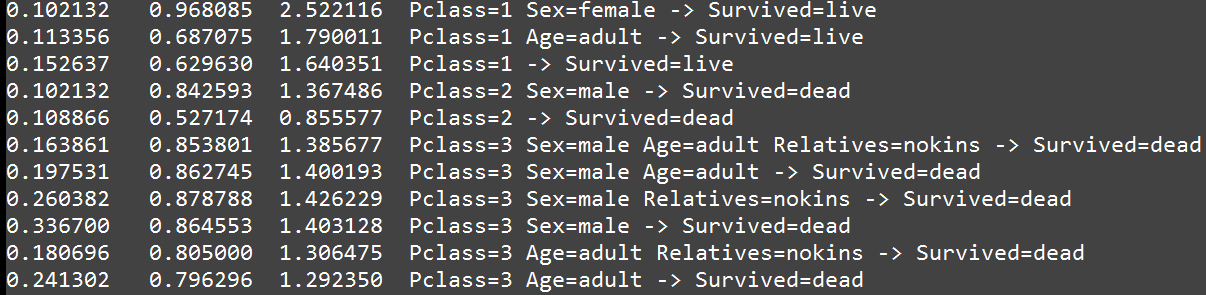
#### 频繁项集挖掘

实验取支持度大于0.1的数据作为频繁项目，用python的数据挖掘包Orange的AssociationRulesInducer接口实现，最终得到75个频繁项目，存储在FreqItems.csv文件下，其条形图如下：

从图中可以看出男性，成年人，死亡出现的最多，说明船上多为男性，并且多为成年人且船上乘客多数没有存活。

#### 关联规则挖掘、计算置信度、评估以及去冗余

通过Orange的AssociationRulesInducer接口获取关联规则，最初得到了28条关联规则，存在文件Associaterule.csv中，格式如下：



其中第一列是支持度，第二列是置信度，第三列是提升值，第四列是关联规则。这些规则当中存在冗余的规则，对于冗余的规则定义如下：

**定义：**若规则B是规则A的衍生规则，并且规则B的lift值小于规则A的lift值，那么规则B就是冗余的。

例如，在初步提取出的关联规则当中，有如下两条规则：

1. 0.995836 Age=adu。lt -> Survived=dead
2. 0.819204 Age=adult Relatives=haskins -> Survived=dead

其中第一列是lift值很明显，第二条规则是第一条规则的衍生规则，但是lift值要小于第一条规则，于是他是冗余的

最终删除了所有的冗余规则，得到了最终的没有冗余的21条关联规则，存储在文件AssociateruleR.csv中。其中各规则lift值的散点图如下所示：

其中lift值最高的规则是Pclass=1 Sex=female -> Survived=live，其说明买头等舱的女性乘客存活的可能性较大。（详见图表.xlsx）