****

**Titanic案例报告**

学 院： 金融学院

课程名称： 数据挖掘

教学班级： A01

姓 名： 李茂源

学 号： 0172841

目录

[1 数据理解 3](#_Toc9702832)

[1.1 数据集理解 3](#_Toc9702833)

[1.2 数据审核 3](#_Toc9702834)

[2 数据预处理 4](#_Toc9702835)

[2.1 新字段生成 4](#_Toc9702836)

[2.1.1 Partner字段生成 4](#_Toc9702837)

[2.1.2 CabinH字段生成 4](#_Toc9702838)

[2.1.3 Call字段生成 4](#_Toc9702839)

[2.2 相关性分析 4](#_Toc9702840)

[2.2.1 Age，Sex和Survived 4](#_Toc9702841)

[2.2.2 Pclass和Survived 5](#_Toc9702842)

[2.2.3 Fare，Pclass和Embarked 6](#_Toc9702843)

[2.2.4 CarbinH和Pclass 7](#_Toc9702844)

[2.2.5 Partner和Survived 7](#_Toc9702845)

[2.2.6 Call和年龄 8](#_Toc9702846)

[2.2.7 神经网络图 9](#_Toc9702847)

[2.2.8 Apriori算法 9](#_Toc9702848)

[2.3 年龄缺失值填充 9](#_Toc9702849)

[3 模型建立 10](#_Toc9702850)

[3.1 聚类分析 10](#_Toc9702851)

[3.2 C5.0决策树分析 11](#_Toc9702852)

[4 预测和意见 12](#_Toc9702853)

[5 心得体会 12](#_Toc9702854)

# 数据理解

## 数据集理解

泰坦尼克号事件作为重大事故，死亡人数不计其数，可以通过数据挖掘找出存活和一些属性之间的关系，从而达到预测是否存活的效果。

## 数据审核

使用数据审核节点进行数据了解。



图 1‑1

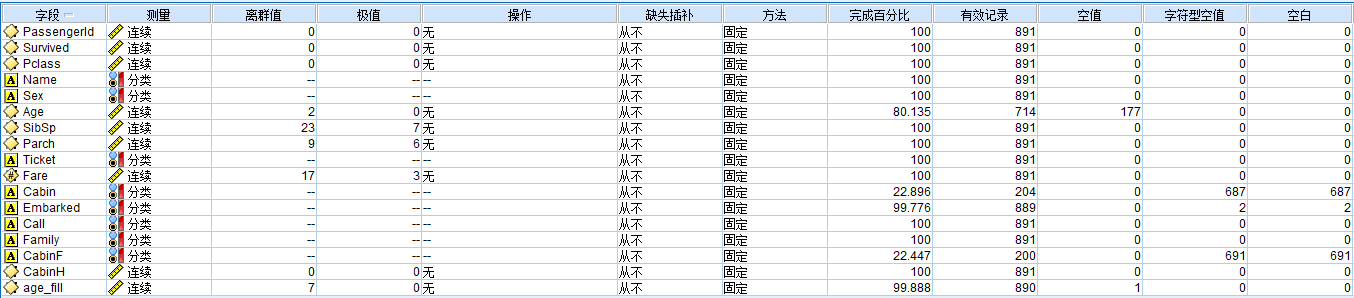


图 1‑2

由图1-1和图1-2可以得出：

* + 本数据集有12个字段，其中年龄的缺失值很多，有177例，对预测有重大影响，需要填充。
  + 默认Pclass和Survived都为连续型数值，要进行属性修改，改为有序和标志
  + Cabin的缺失值也很多，Cabin作为舱号，可能不是没有填，是有些乘客没有Cabin这个舱号。
  + Pclass为3的人数特别多，等级为3的乘客可能地位更加低。

# 数据预处理

## 新字段生成

### Partner字段生成

对于Parch和Sibsp这两个字段，本质上来说，都是伙伴，所以将它们相加导出新字段Partner。

### CabinH字段生成

Cabin代表舱位，而且有舱位的人可能并不多，所以我们导出新字段CarbinH，有舱位的为1，无舱位的为0.

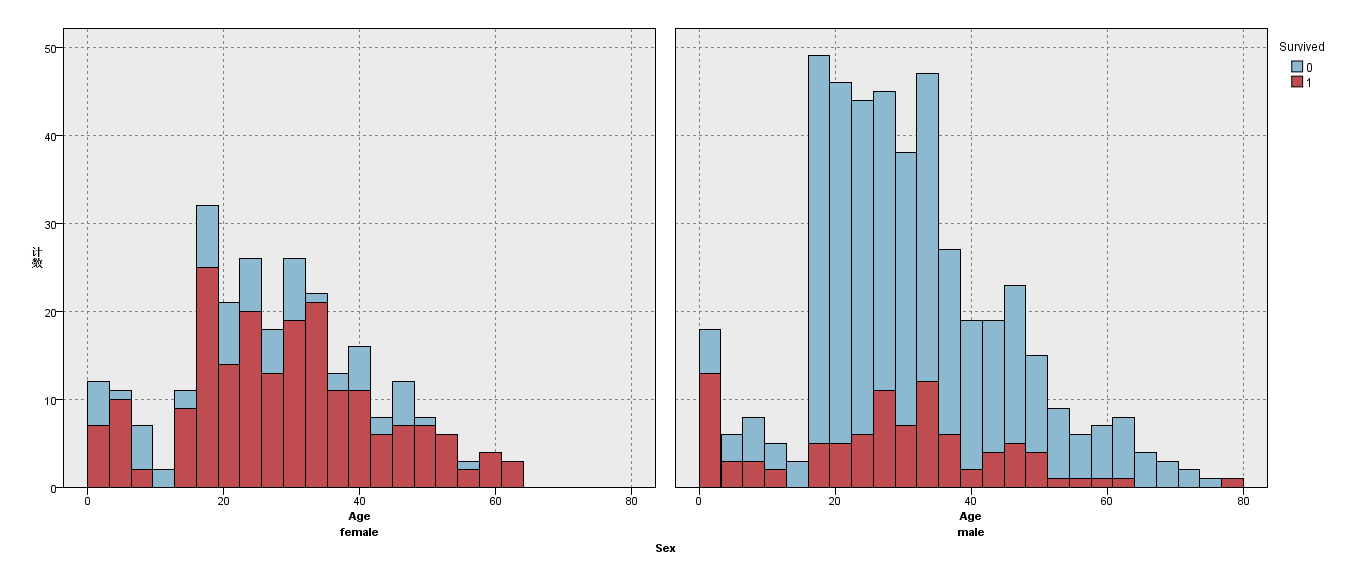
### Call字段生成

名字这个变量不好进入模型，但是称呼和年龄有很大的关系，在填补年龄缺失值有大用处，使用正则表达式提取出称呼，生成Call字段。

## 相关性分析

### Age，Sex和Survived

因为泰坦尼克号中有让小孩和女人先走这个规则，所以年龄，性别和存活肯定有一定联系。



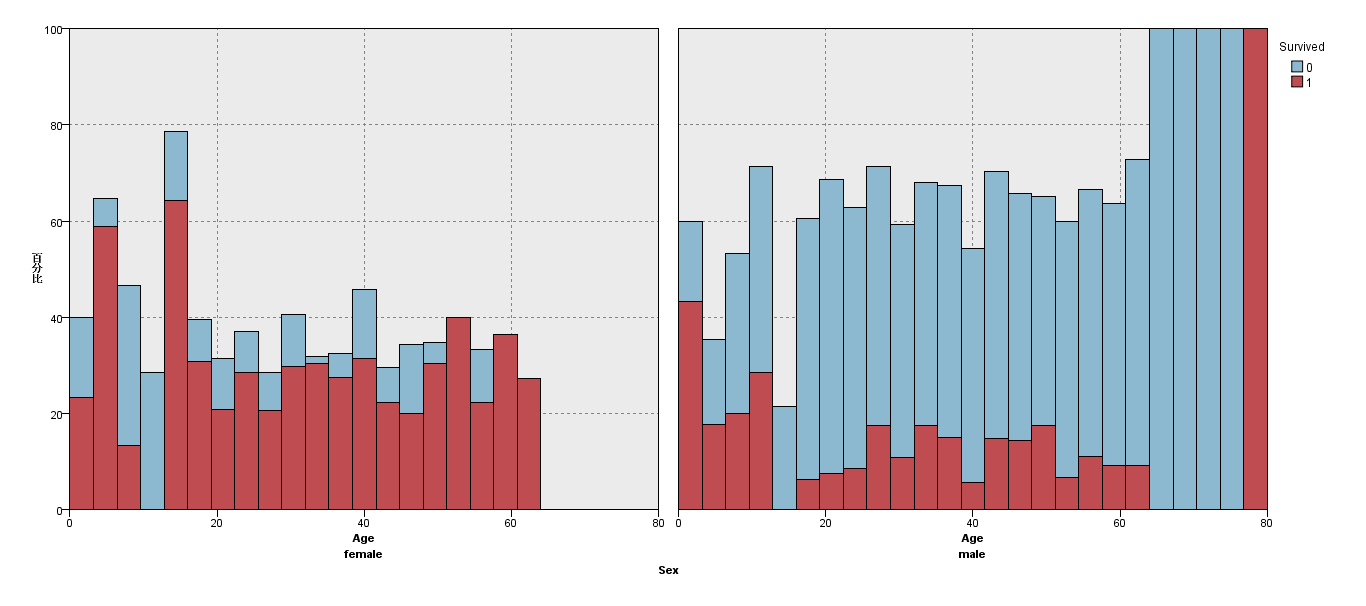


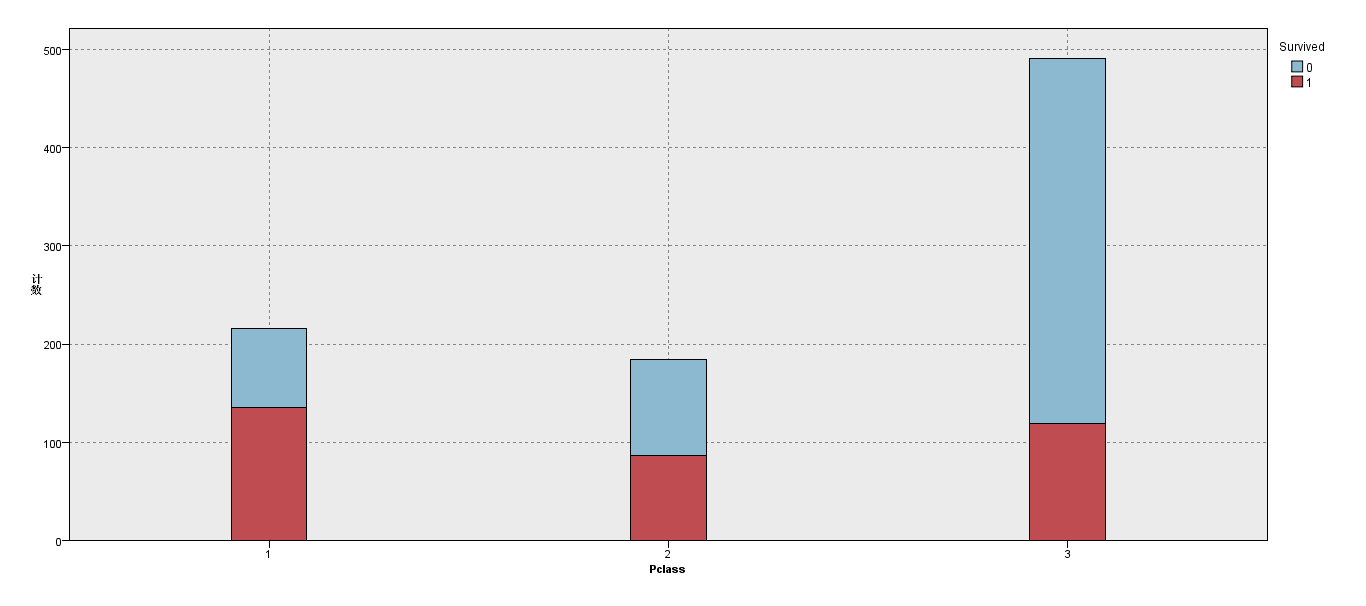
图 2‑1

如图2-1所示：

* 明显可以看出，女性的存活率显著高于男性的存活率。
* 在男性中年龄小的存活率很高，年龄大的存活率很低。
* 年龄小的人存活率高于年龄大的存活率，且在男性中这种更加明显。

### Pclass和Survived

Pclass为乘客等级，反映了人的社会地位，地位高的或许会通过买通等手段存活。



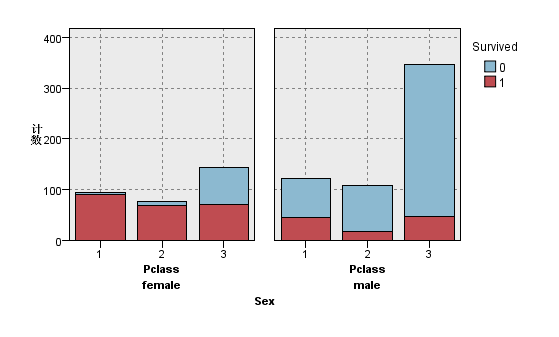


图 2‑2

如图2-2所示：

* Pclass为1的存活的比例显著高于2，而Pclass为2的显著高于3。
* Pclass为1应该就是地位更加高的人。然后逐级递减。
* 女性Pclass为1和2的基本全部存活。

### Fare，Pclass和Embarked

费用和乘客等级很有可能有关，不同港口路程不同，可能也会影响芬妮用。可能会有冗余，进行相关性分析

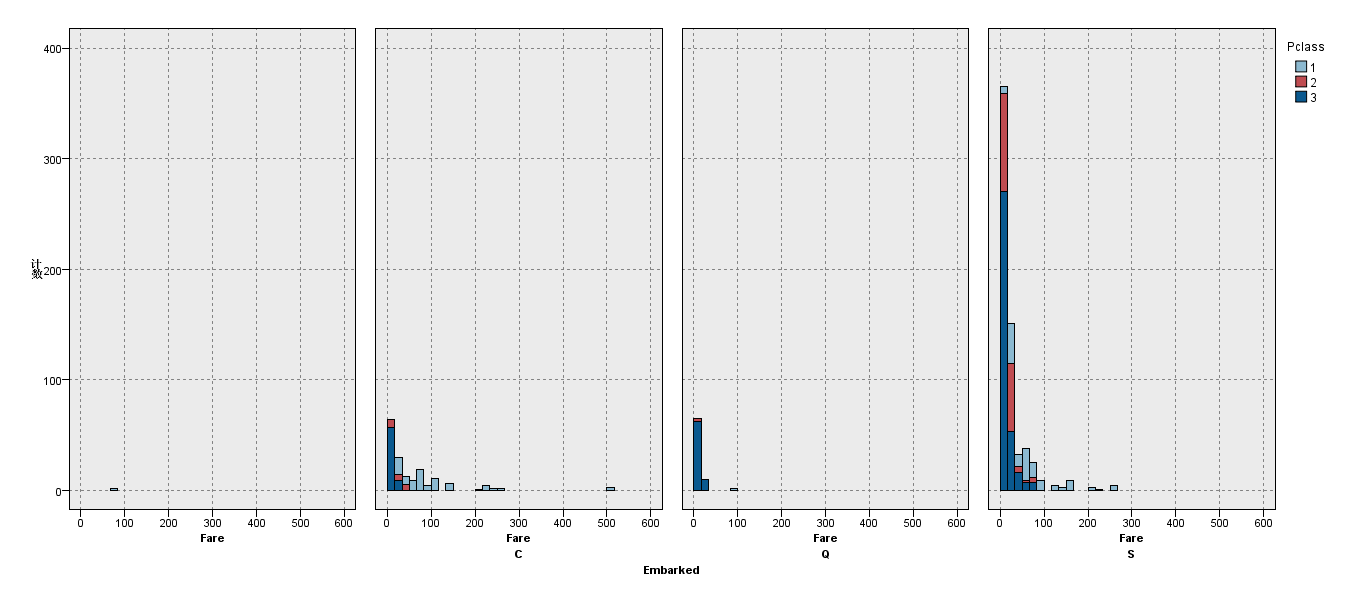


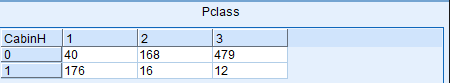
图 2‑3

由图2-3可知：

* Pclass越高，费用也越高
* C港口上的人，等级偏高，Pclass为1的人数占比较大。

### CarbinH和Pclass

只有少数人有舱位，可能等级越高，有舱位可能性越大，采用矩阵和条形图探索它们之间的相关性。



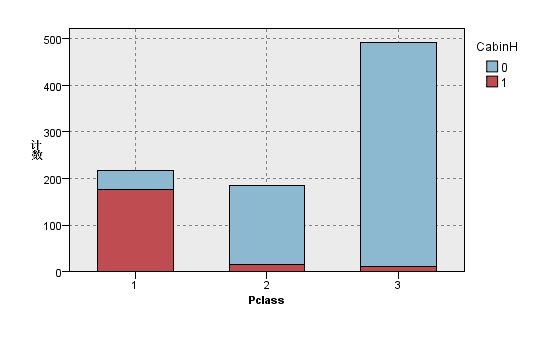


图 2‑4

如图2-4所示：

* Pclass为1的乘客，有舱位的比例明显高于Pclass为2和3的。
* Pclass和Cabin的独立概率为0，说明两个变量之间强相关。

### Partner和Survived

同伴的数量很可能能够影响存活率，探究两变量之间的关系。采用标准化的条形图输出得到

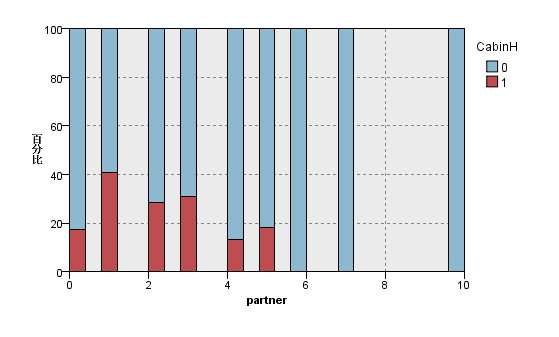


图 2‑5

如图2-5所示：

* 同伴个数在1-3个左右存活率比较高。同伴太多，反而存活率低。

### Call和年龄

年龄的缺失值可以用Call来进行填充，探究两者之间的关系

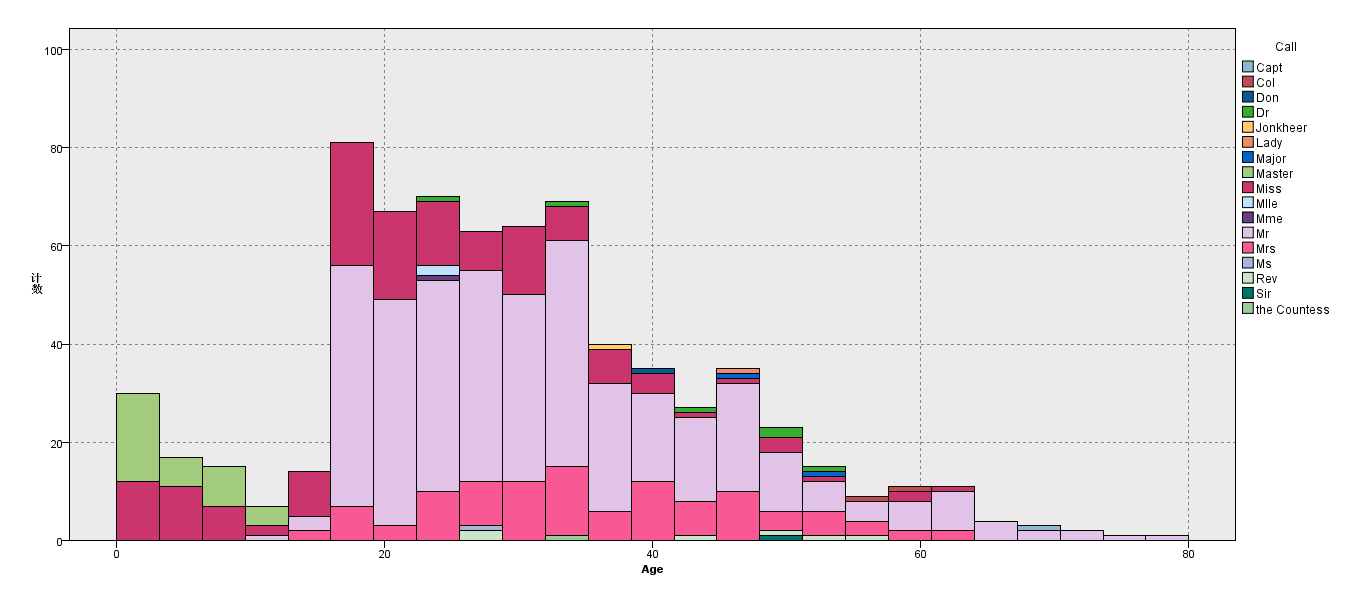


图 2‑6

如图2-6所示：

* 年龄和称呼有关联，比如只有小于12岁才可以叫Master.
* Mr的年龄大于12.
* Miss为年轻女性在20-35之间，Mrs则稍微大一些在25到最后都有称呼。

### 神经网络图

为了探索多个离散变量之间的关系，我们采用神经网络图进行输出。

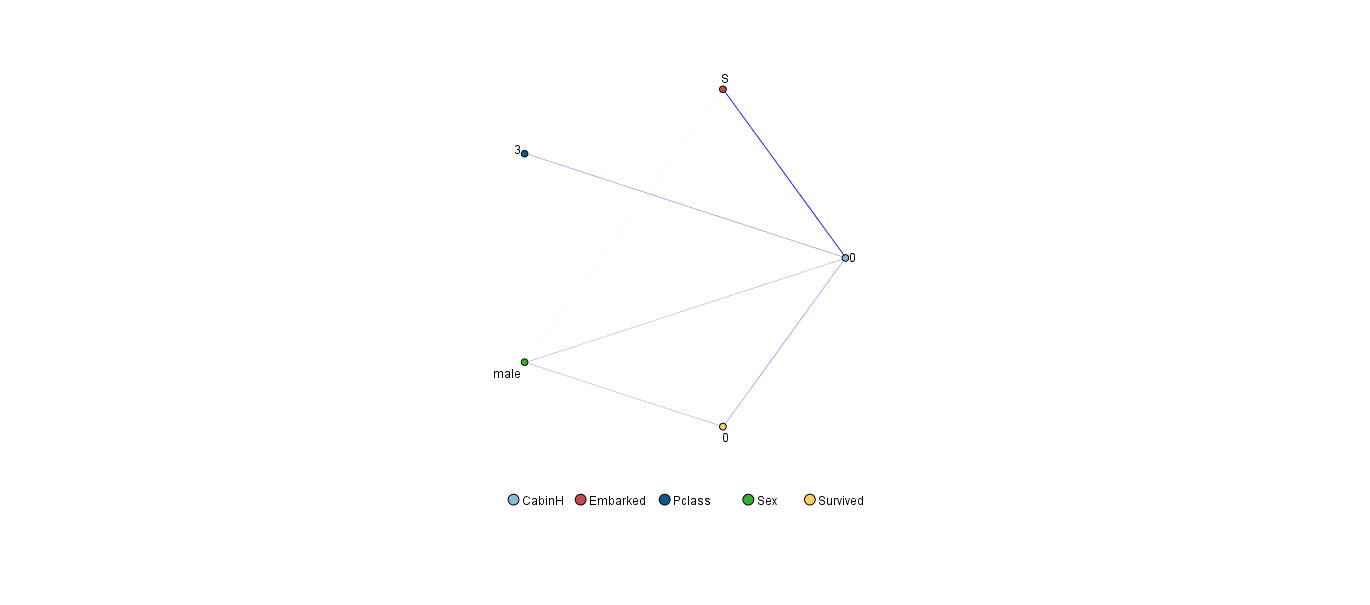


图 2‑7

如图所示：

* 男性，无舱位，无存活。
* Pclass为3，无舱位，也很容易无法存活。

### Apriori算法

为了进一步探究之间的关系，采用Apriori算法来进行相关性分析。输出如下。

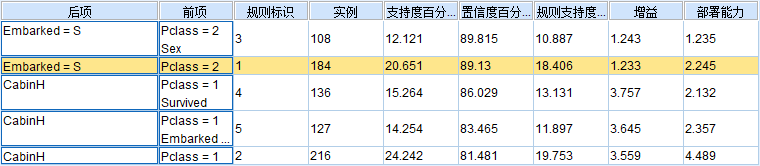


图 2‑8

没有很大的发现，只能说明等级为1的乘客，很可能有舱位。

## 年龄缺失值填充

年龄对是否存活有重大影响，所以要进行缺失值填充。我们按照称呼进行分组，然后利用每组的中位数对年龄的空值进行填充，生成age\_fill字段。

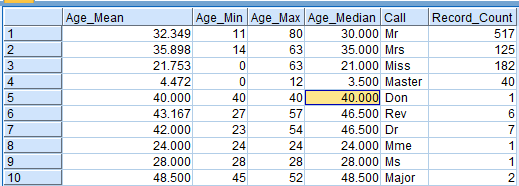


图 2‑9

# 模型建立

## 聚类分析

将Survived,Pclass,Sex,Age,Fare,Embarked,CabinH,age\_fill,partner进行聚类分析，得到如下结果。



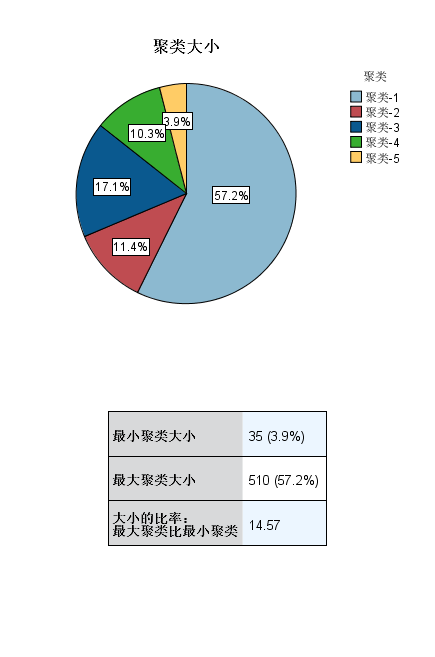


图 3‑1

如图3-1所示：

* 这个聚类分析聚类之间的区分度一般
* 类之间分开的特征是通过性别和社会地位。
* 地位高的男性和年级小的男性在一类，地位低的在一类。
* 地位低和年龄较大的男性占比较大，占了将近一半以上，且这些人大部分死亡。

## C5.0决策树分析

将Survived作为输入变量，Pclass,age\_fill,CabinH,Emarked,Fare,Sex作为输入变量建立C5.0模型，得到输出如下。

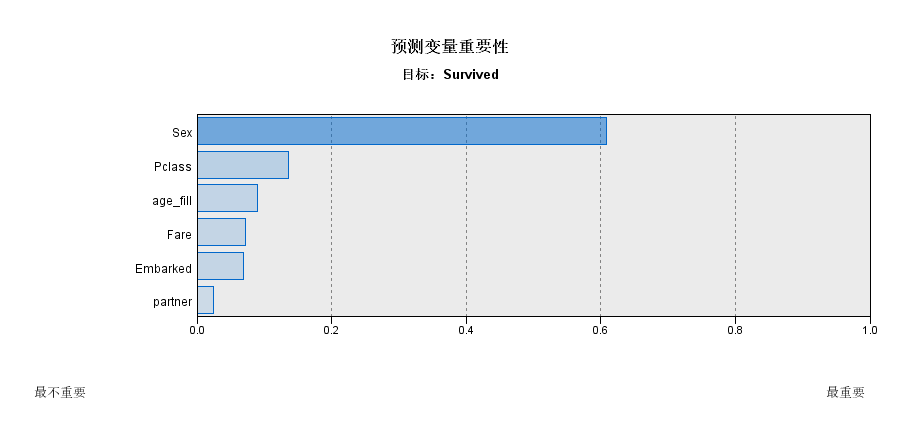


图 3‑2

如图3-2所示：

* Sex是影响存活最大的因素。其次为Pclass，然后是age。
* Partner对存活的影响相对不是很大。

# 预测和意见

* 女性乘客如果Pclass为1，2基本都能存活。
* 男性乘客中，年龄小于13岁，存活可能性比较大。
* 伙伴不能太多，如果伙伴大于3个人，存活率将会比较低。
* 乘客Pclass对存活影响很大，男性如果年龄大于13岁，Pclass不为1的话，存活概率会非常低。
* 存活人的特点，按重要先后分析，首先是Age小于13岁，然后是性别女性，然后是Pclass，Pclass的等级越高活的可能性越大，最后是Partner，一般3个以下存活率高一些。

# 心得体会

* 联想能力还是不强，发现关联能力还是很弱，不能很好把握数据集。对于探索性的东西，我一直都不是擅长，联想不出来。
* 软件操作已经较为熟练，尤其是对绘图操作，已经能够熟练使用图形版。
* 概念虽然很好理解，但是实际数据集处理起来还是有各种各样的问题，尤其是在关联规则挖掘中，很多情况仅仅是因为某个属性的出现次数比较多，占比比较大，不能很好判断关联关系。
* 离散化还是困难，无法很好把握离散化技巧，这个数据集本来想对年龄离散化，结果越离散化准确率越低，最后只能分成未成年人和成年人。
* 文案能力提升很慢，从来都是比较精简回答，也不知道从何说起。我真的很认真写了这个作业。
* 这里其实分析起来很有困难，我个人认为舱位Cabin并没有什么作用，因为有舱位的人并不多，而有分为有舱位和没有舱位又和Pclass有关，当进模型的时候，明显是Pclass更加有用。找来找去还是这些东西，在Ticket的分析时候，因为这些东西都是文本这类东西，分析出来十分困难。所以说来说去还是那点东西，没有发现新的东西。
* Age变量其实也非常不好用，老人太少没有代表性，小孩有一些，只有13岁以下的情况有特别，其他没什么特别。
* 数据量实在太少，无法找到比较有代表性的东西。尤其是在伙伴那一块，有伙伴为3个左右，或者无伙伴的人比较多而已，那种很多伙伴的人死了，很有可能只是偶然，因为这种情况本来就不多。而且在SPSS Modeler中这个分组这一块，不是很熟悉，在堆叠操作时候经常遇到问题。最后还是用的Excel的数据透视图，更加方便些。尤其是填充时候，对Moderler不太熟悉，还是用excel方便。