# 泰坦尼克的存活分析

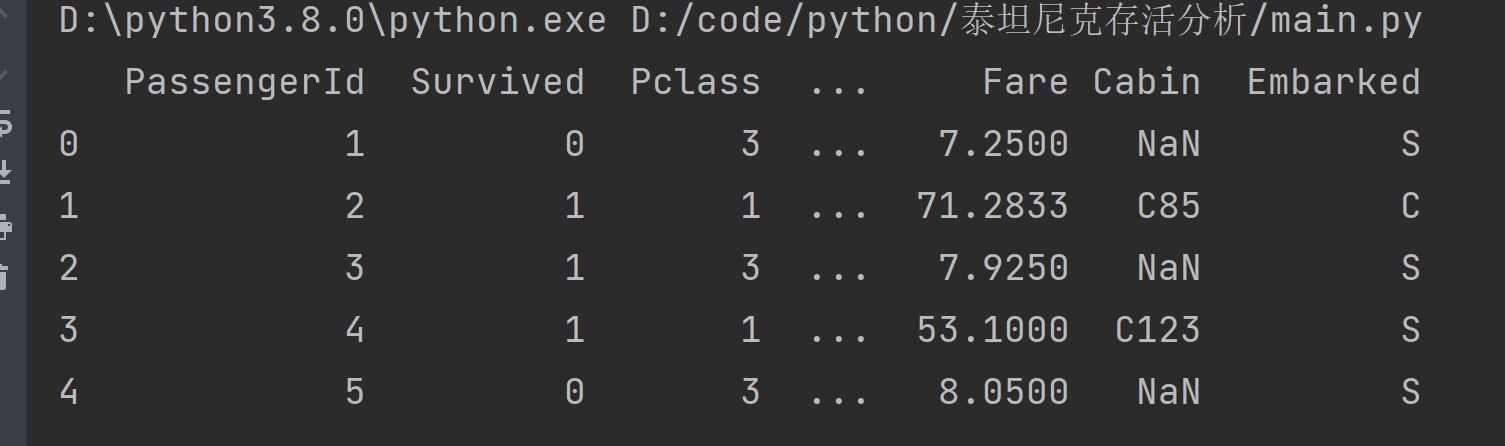
陈宇

数据字段表：

|  |  |
| --- | --- |
| 字段 | 解释 |
| PassengerId | 乘客编号 |
| Survived | 是否获救1-获救 |
| Pclass | 船仓等级 |
| Name | 名字 |
| Sex | 性别 |
| Age | 年龄 |
| SibSp | 兄弟姐妹个数 |
| Parch | 老人孩子个数 |
| Ticket | 船票编号 （没什么用） |
| Fare | 船票价格 |
| Cabin | 船仓编号 （没什么用） |
| Embarked | 上岸码头 |

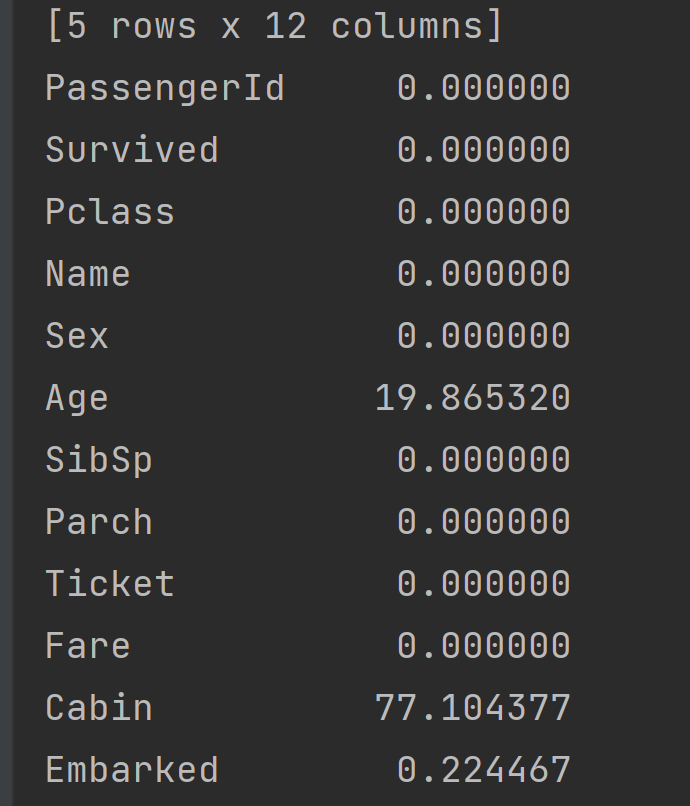
## 数据读取

### 读取数据

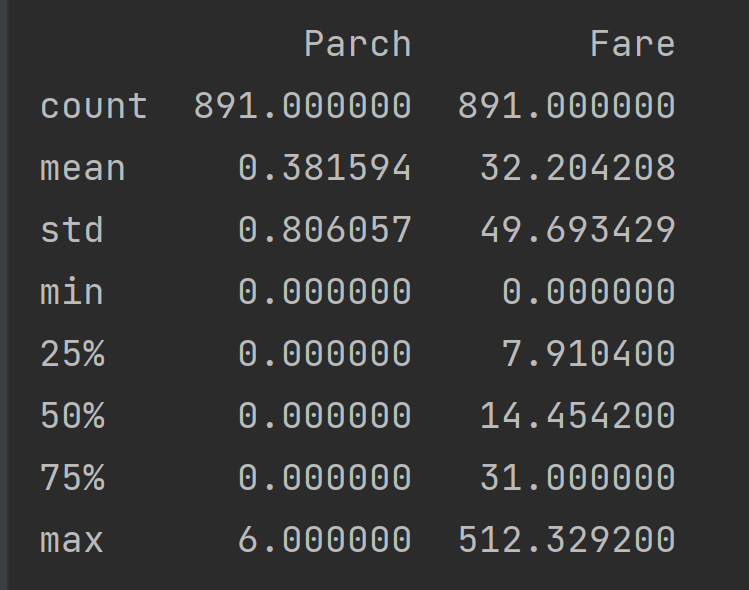
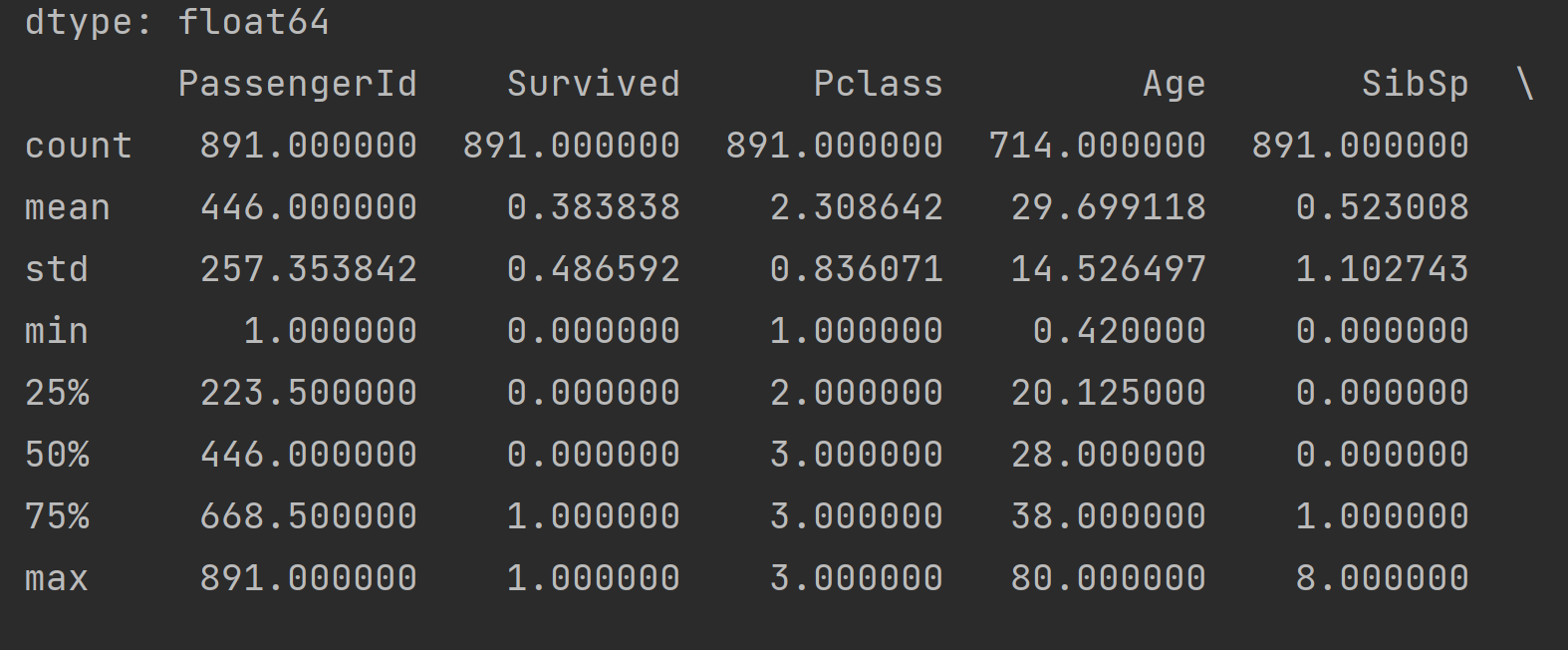


### 统计数据各项指标

1. **缺失值比例统计：**



1. **数值型数据初步统计**

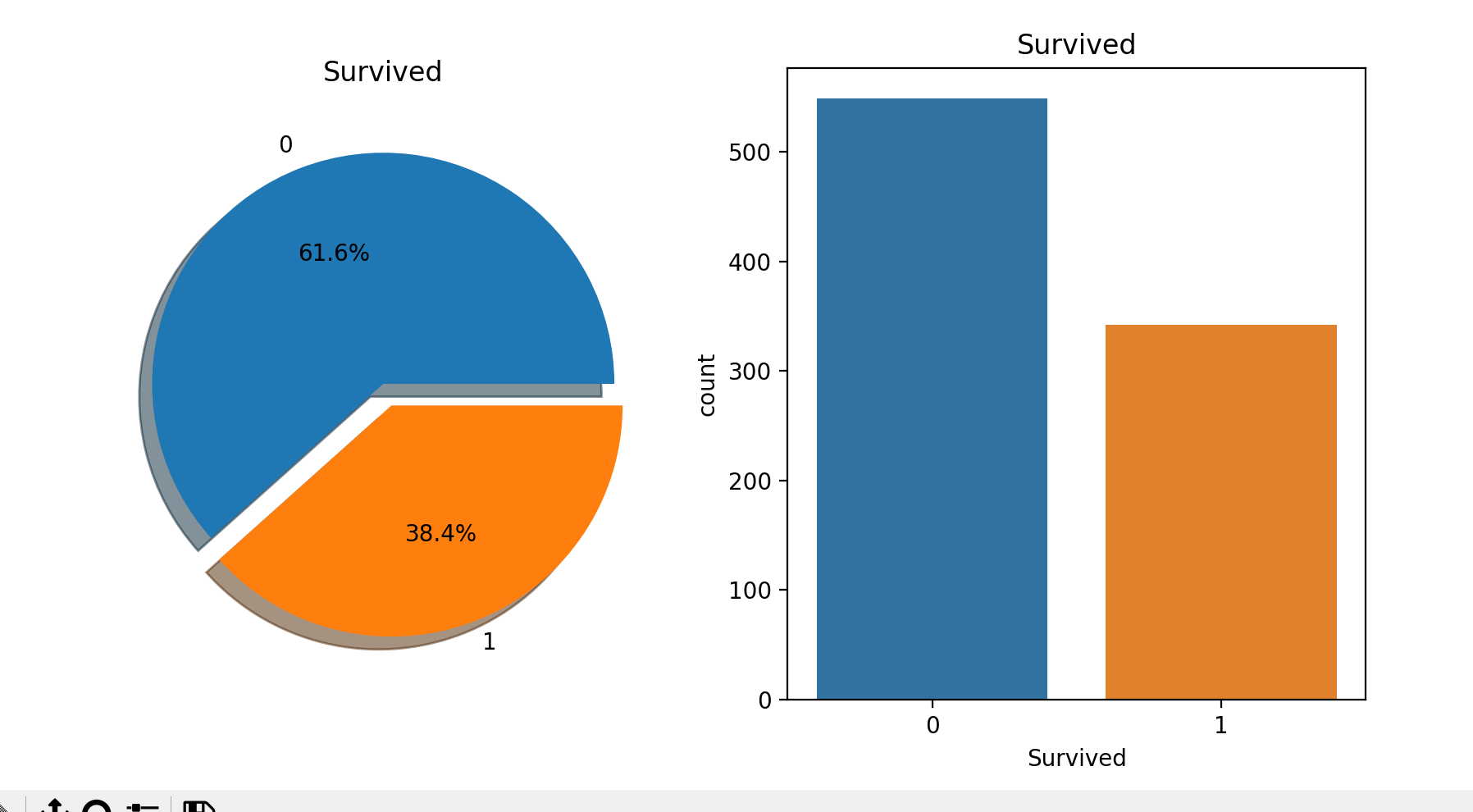


### 明确数据规模和任务

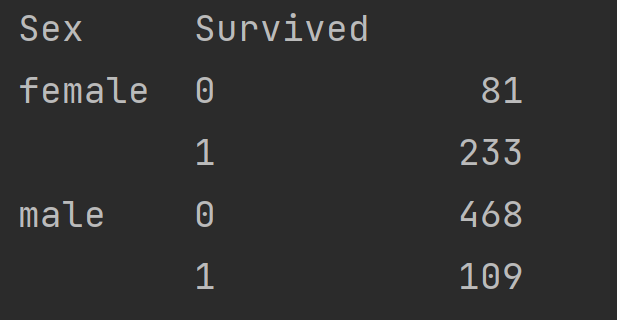
## 特征理解分析

### 单特征理解分析

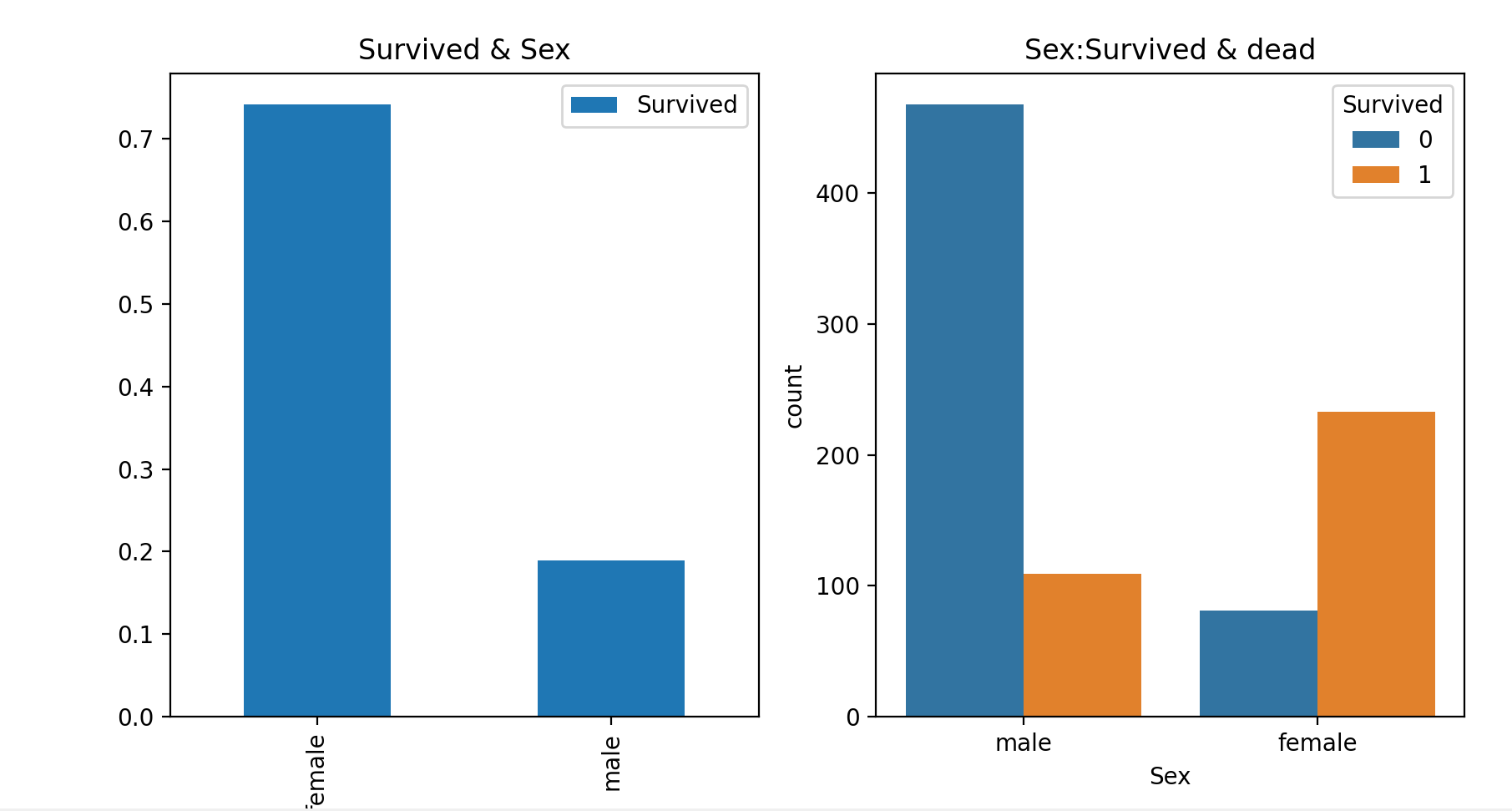
1. **标签结果统计**

****

1. **性别统计**

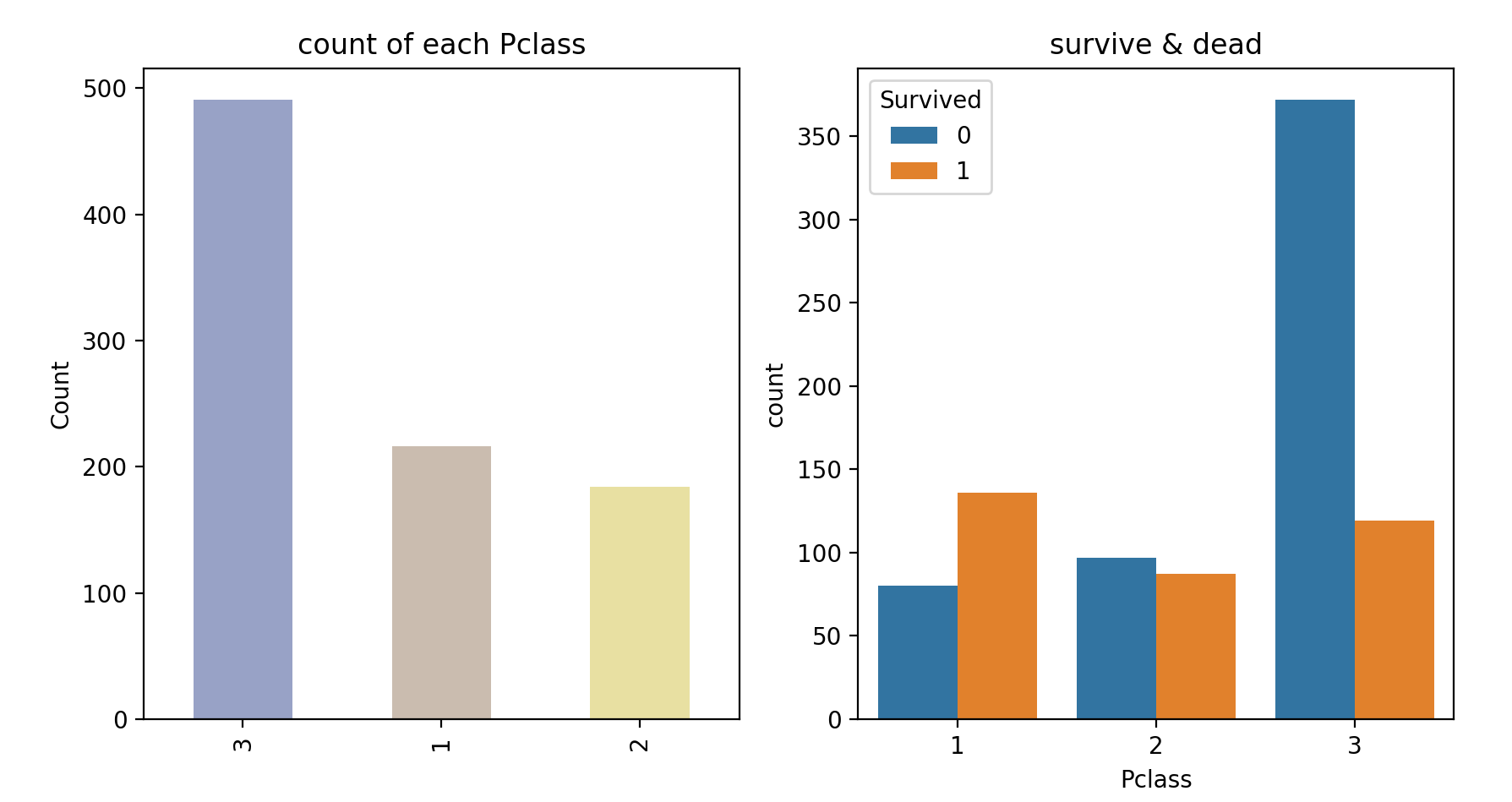


**性别与存活率初步分析：**

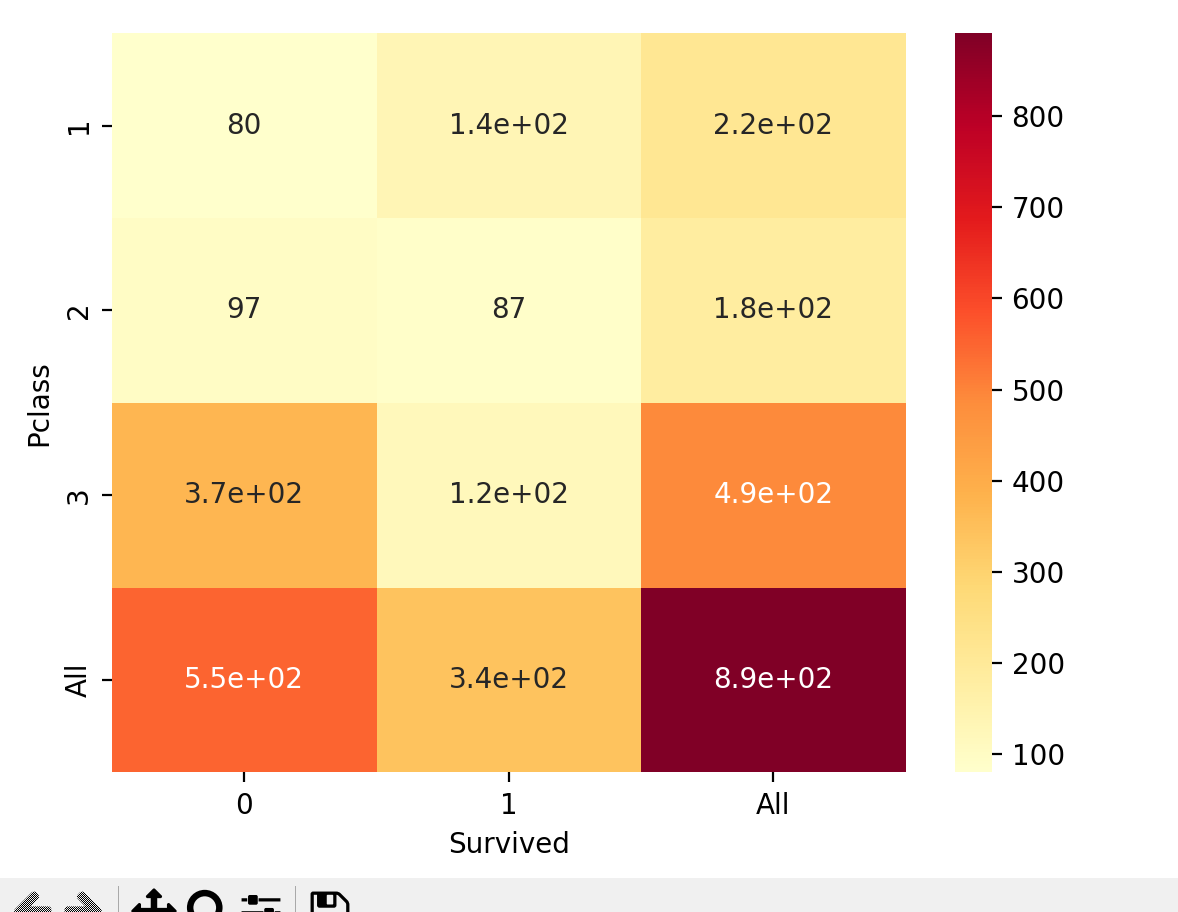


1. **船舱等级初步分析**

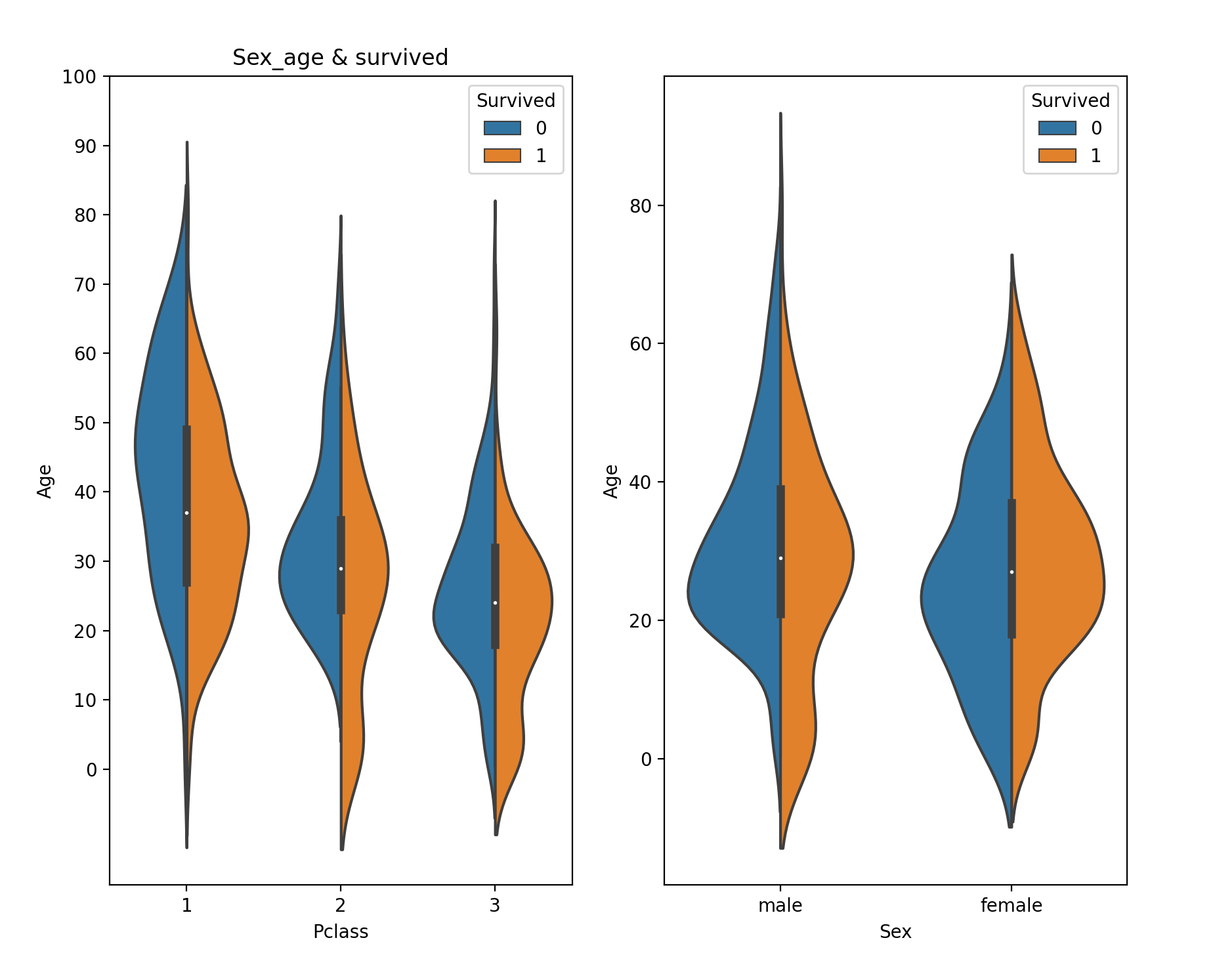
**统计图：**

****

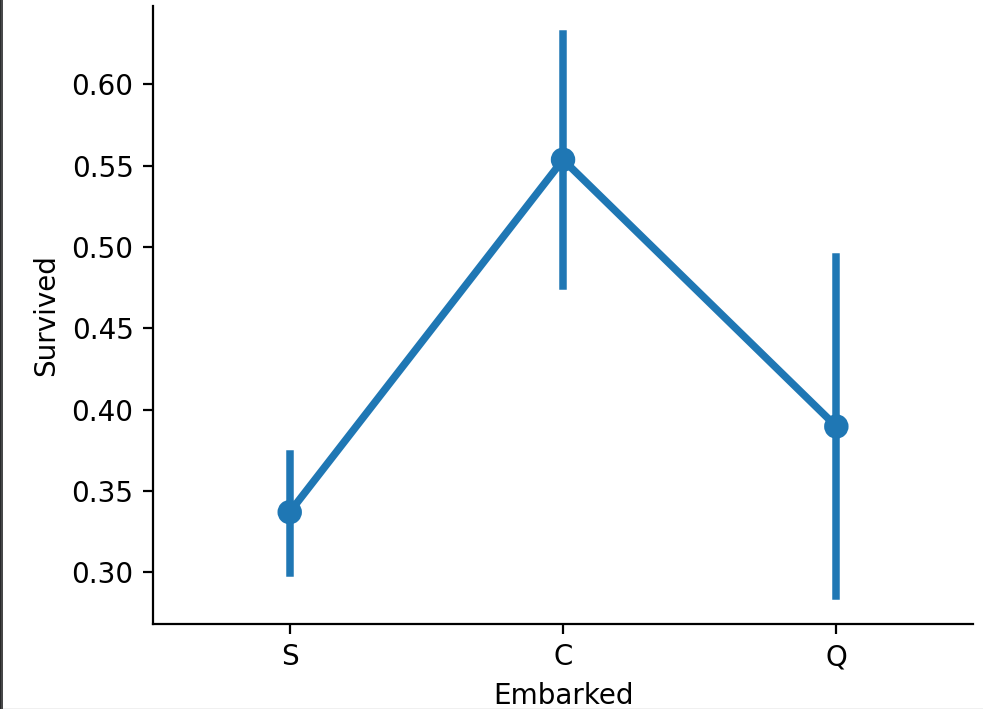
**交叉表热图可视化**

****

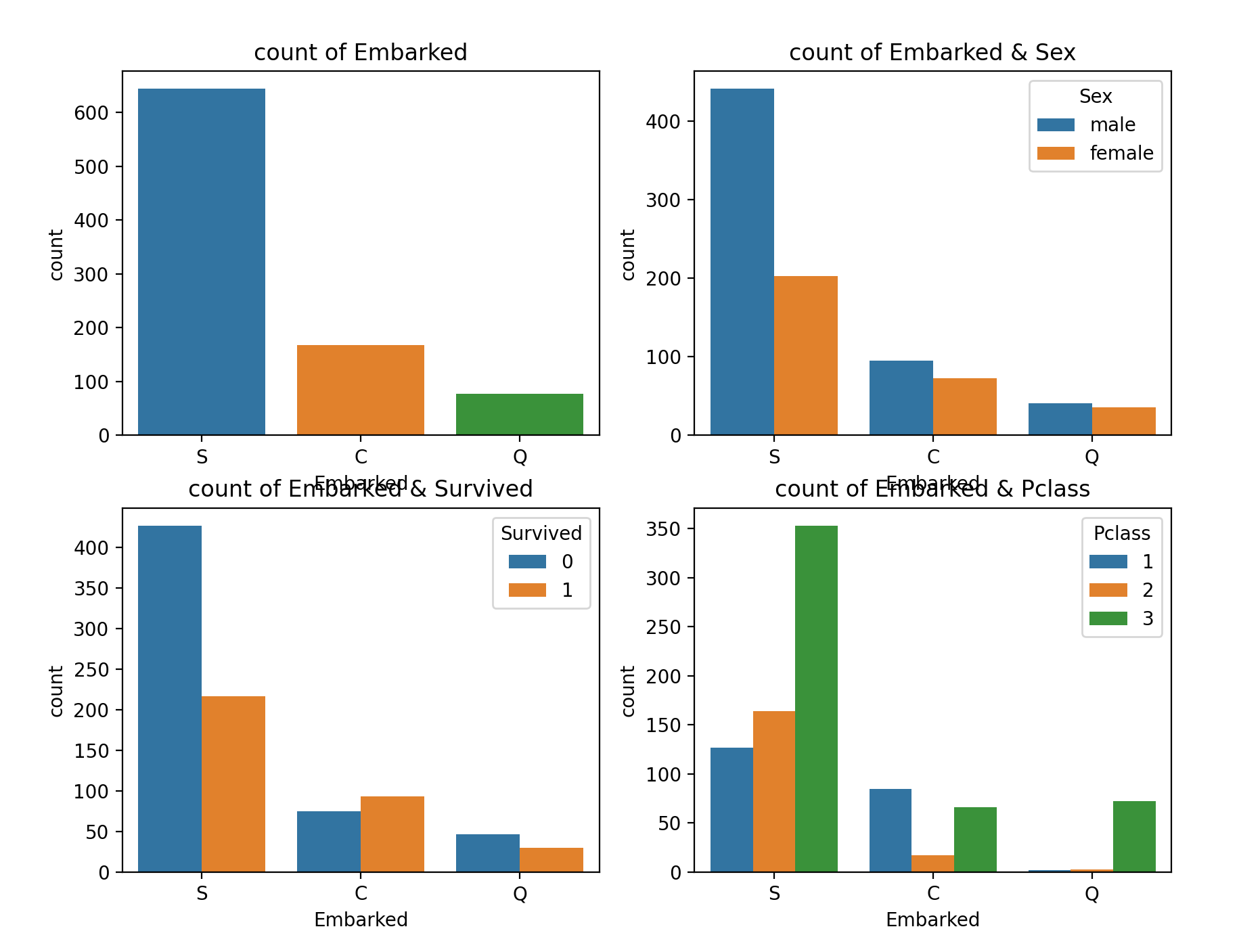
1. **年龄存活率分析:**

**小提琴图：**

1. **登船地点存活率统计：**

****

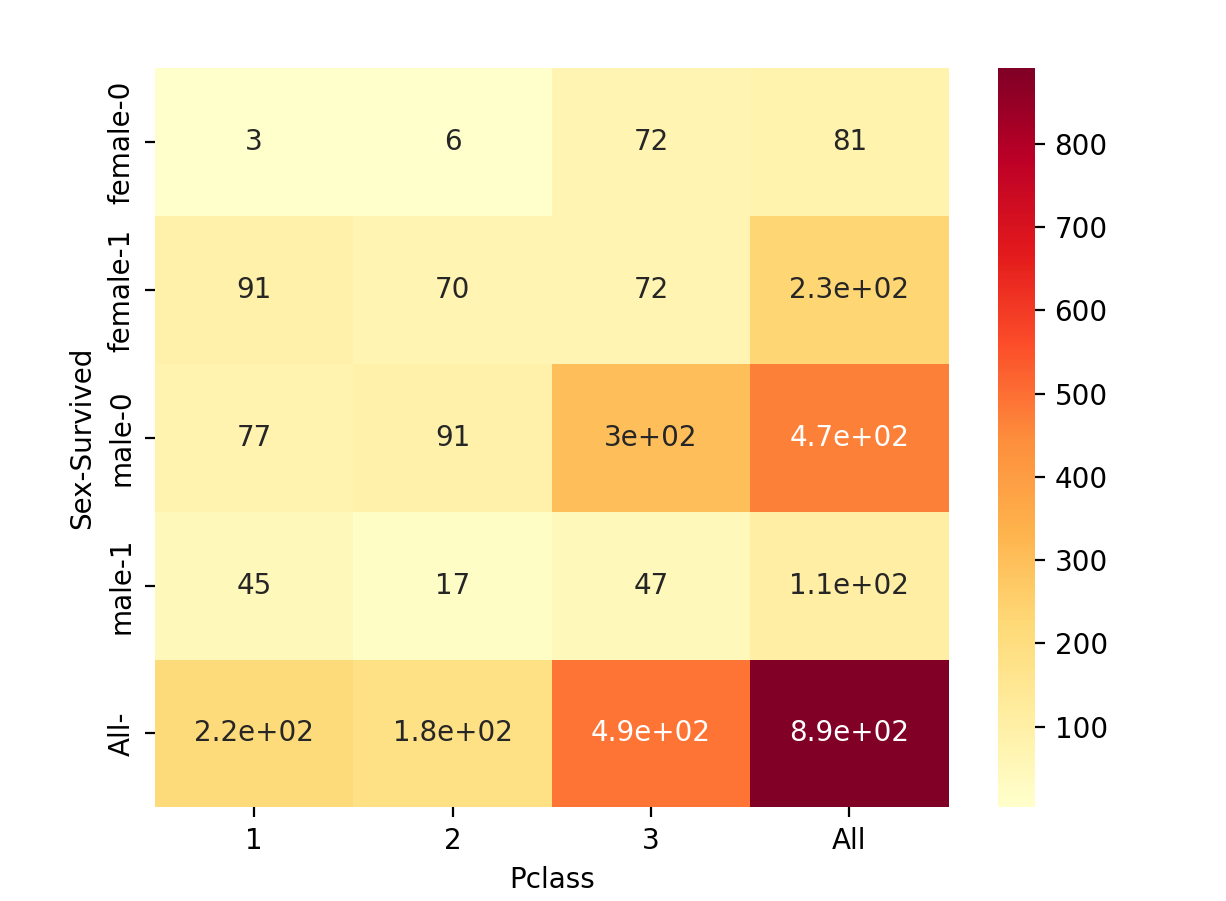
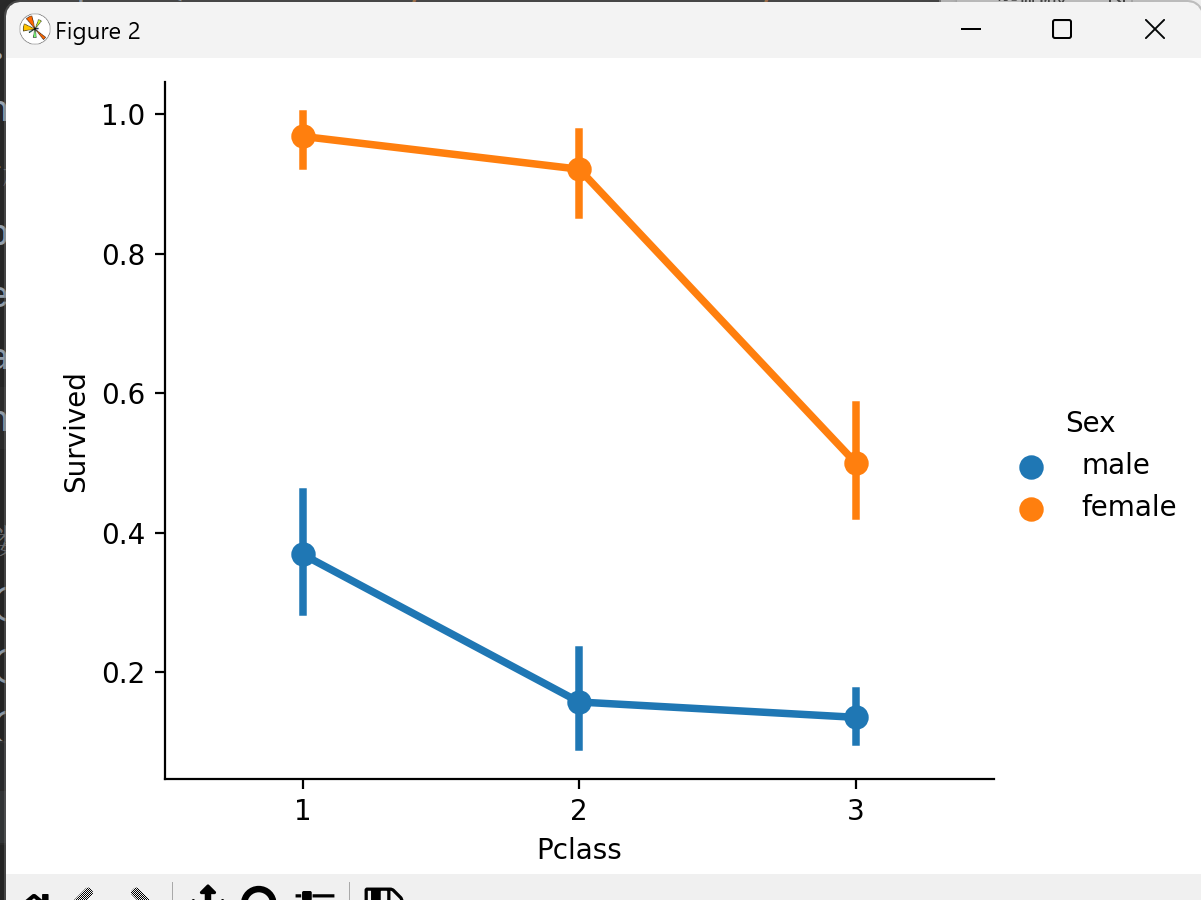
**各种数据统计：**

****

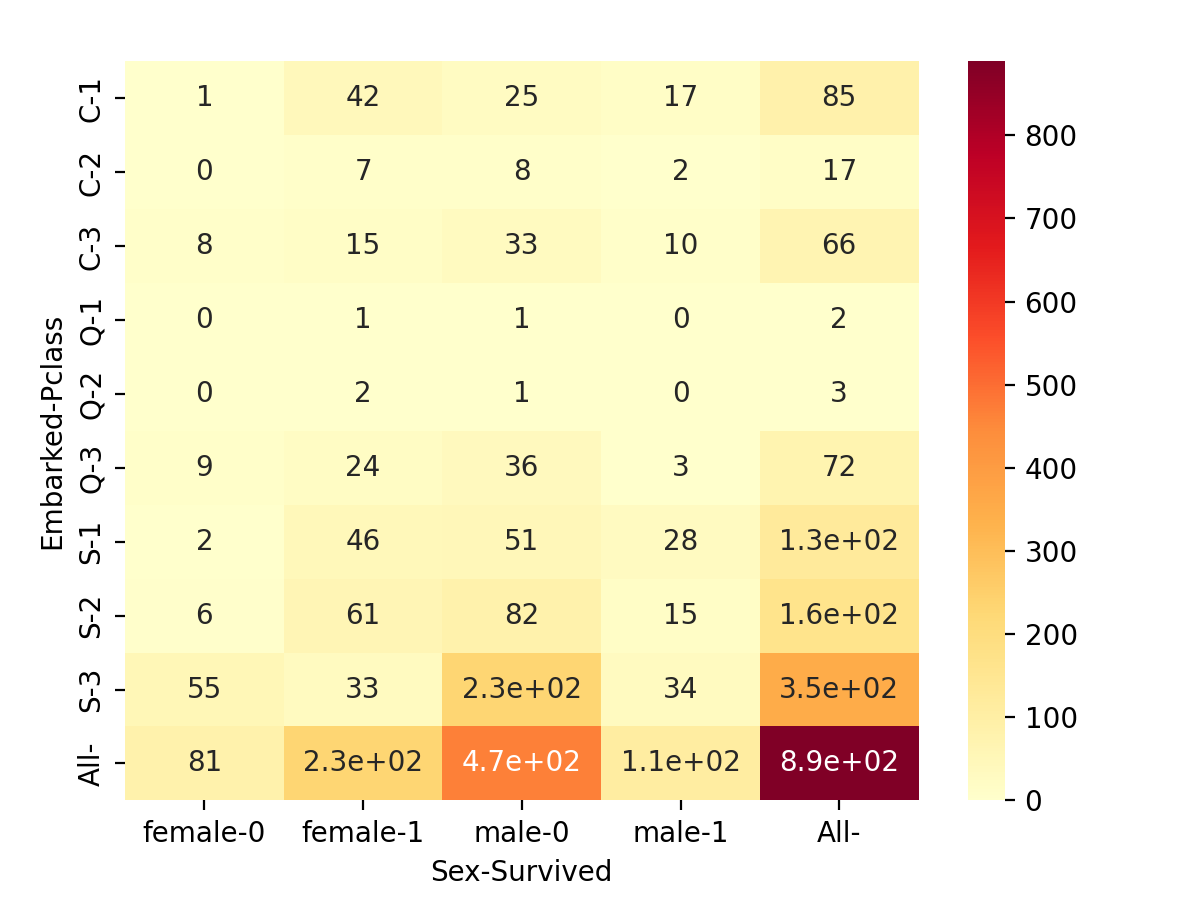
### 多变量统计分析

1. **船舱等级与性别存活率分析**

**热图： 比例折线：**

**** ****

1. **登船码头与船舱等级+性别统计热图：**

****

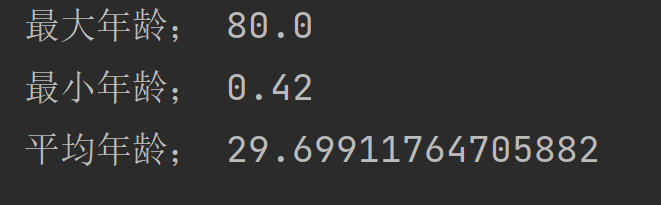
### 统计绘图得出结论

## 数据清洗与预处理

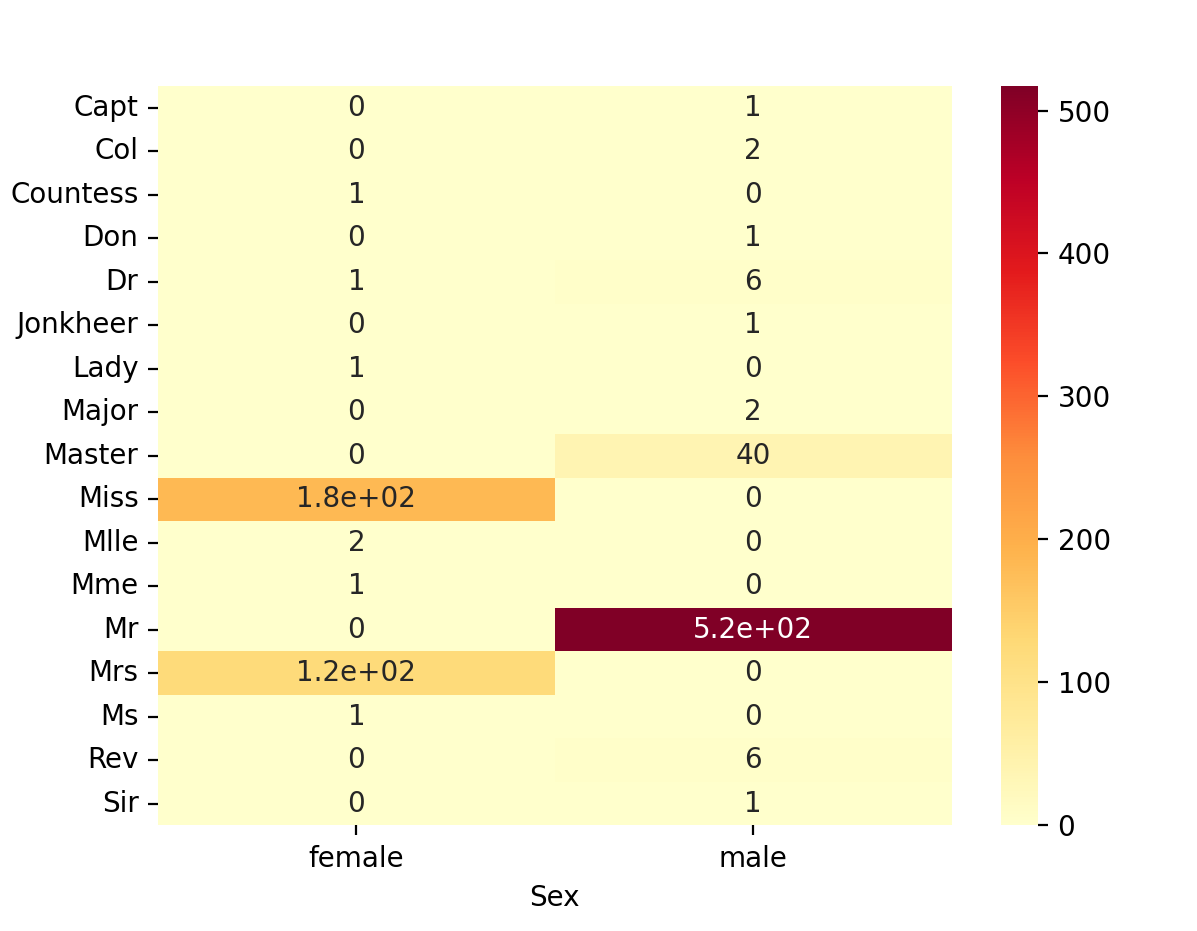
### 缺失值填充

1. **年龄缺失值处理：**

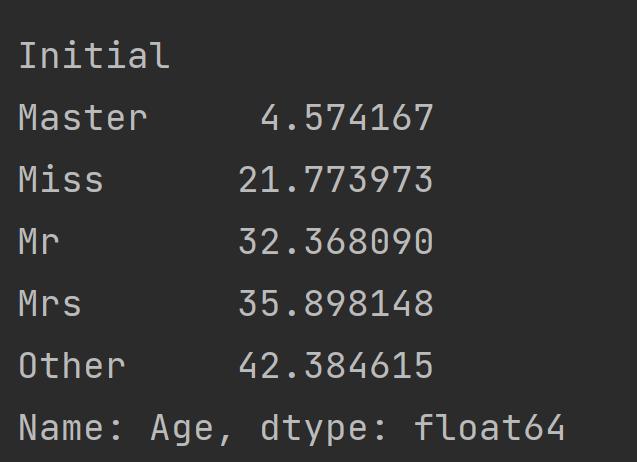
**年龄统计：**

****

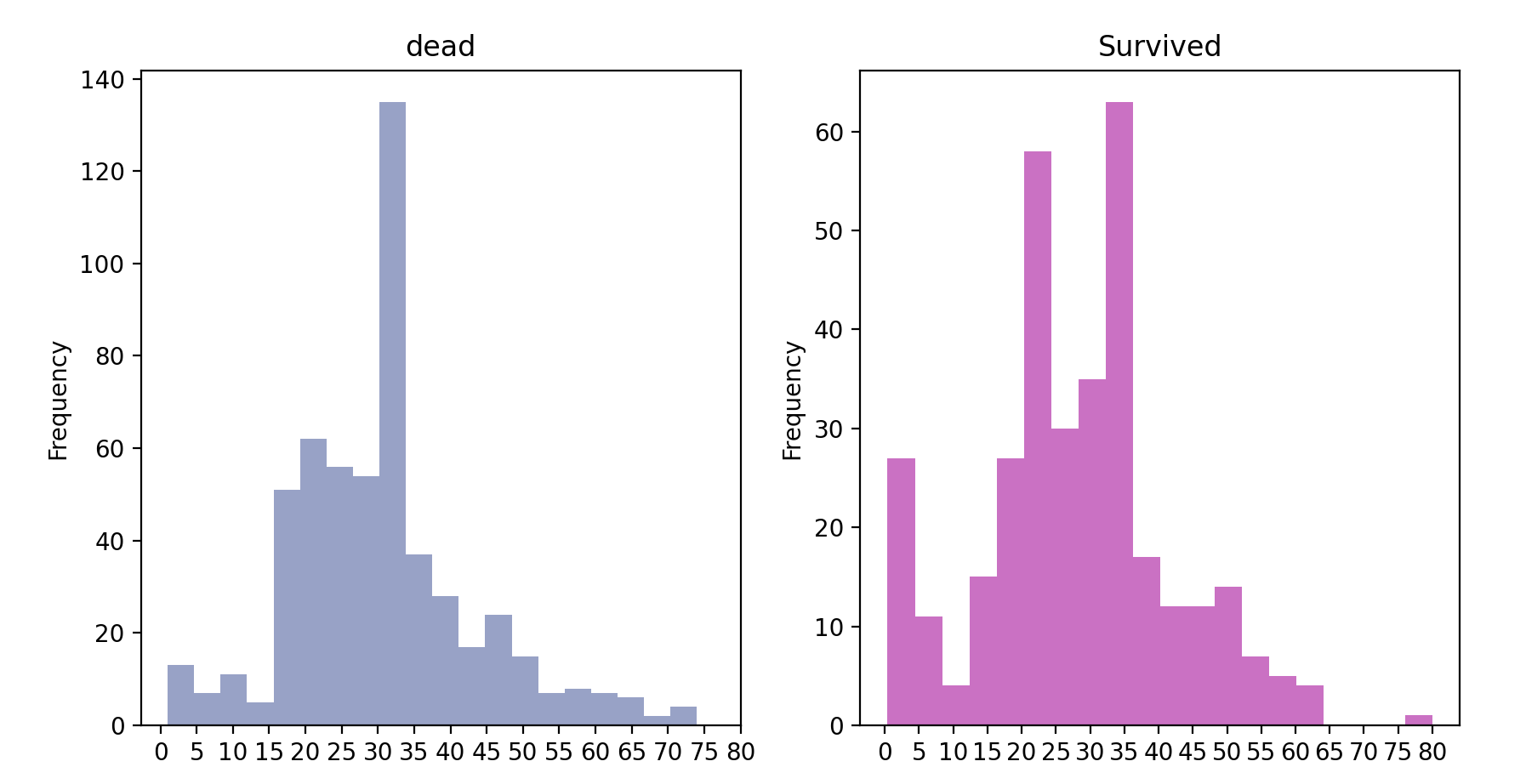
**使用名字称呼统计各个年龄段：**

****

**计算各个称呼年龄平均值：**

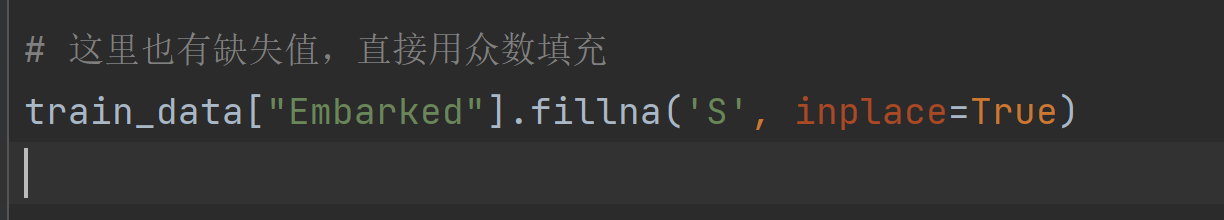
****

**最终各年龄存活统计图：**

****

1. **登船码头缺失值处理：**

**直接众数填充**

****

### 特征归一化

1. **对离散值：Pclass归一化**
2. **对离散值：Embarked 归一化**

### 筛选有价值特征

### 分析特征之间的相关性

## 建立模型与标签准备

### 特征数据与标签准备

### 数据集切分

### 多种建模算法对比

1. 使用keras搭建一个深度神经网络

网络特征：两层隐含层，第一层为63个结点，第二层使用32结点

隐含层激活函数：relu

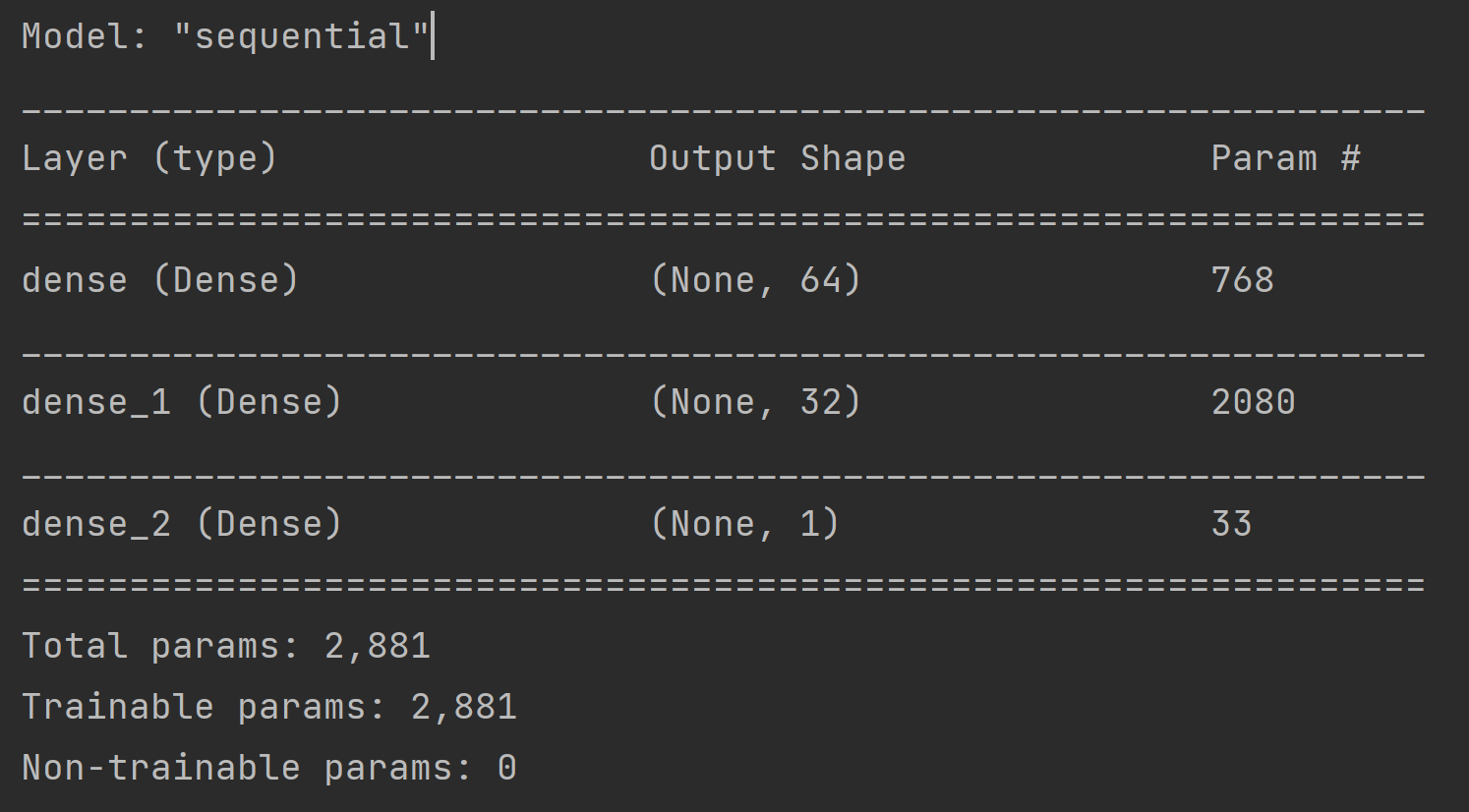
输出层激活函数：sigmid

损失函数：ce

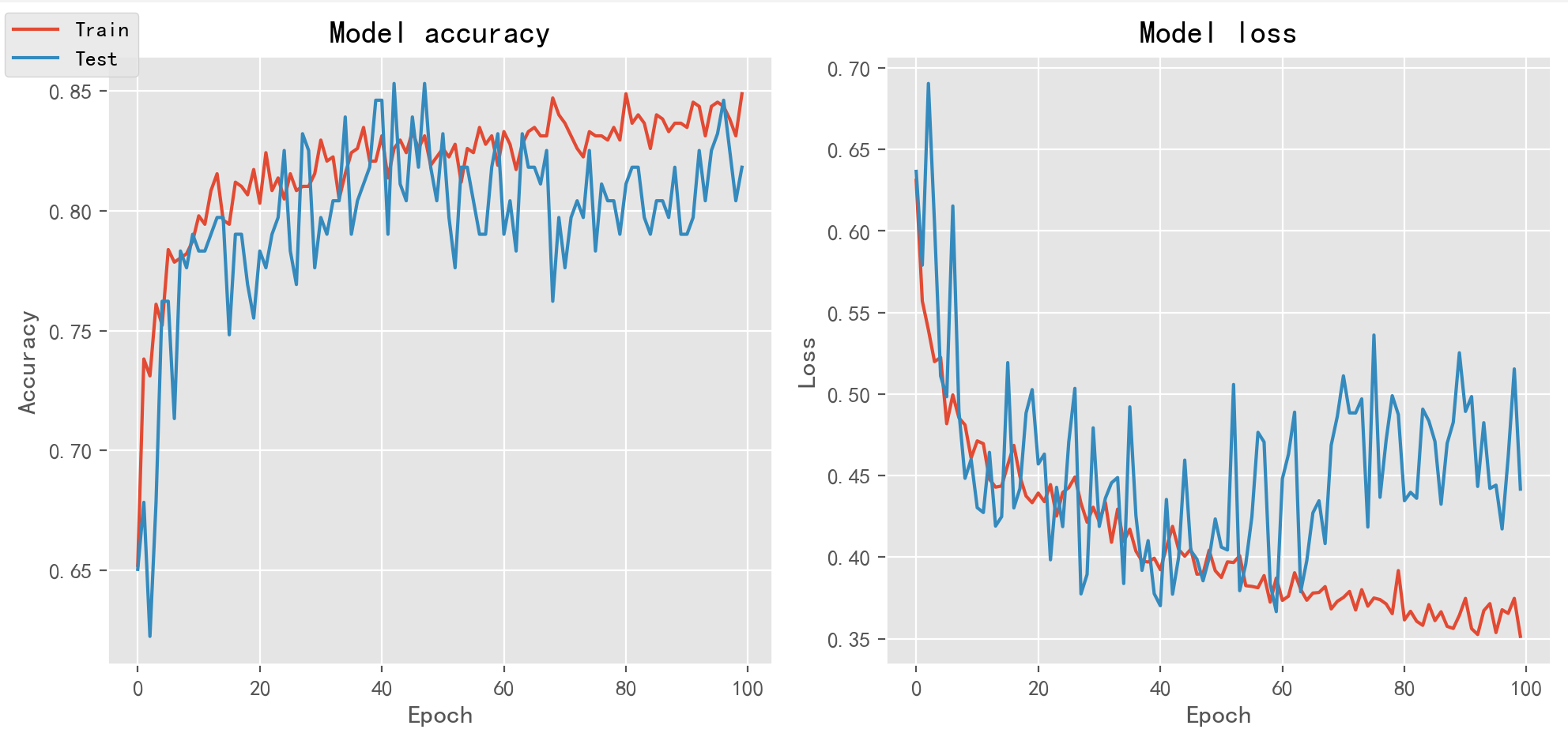
验证集占比：0.2

Epoch：100

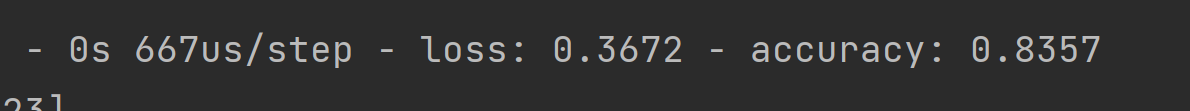
Batch：10



**训练结果：**



**测试结果:**

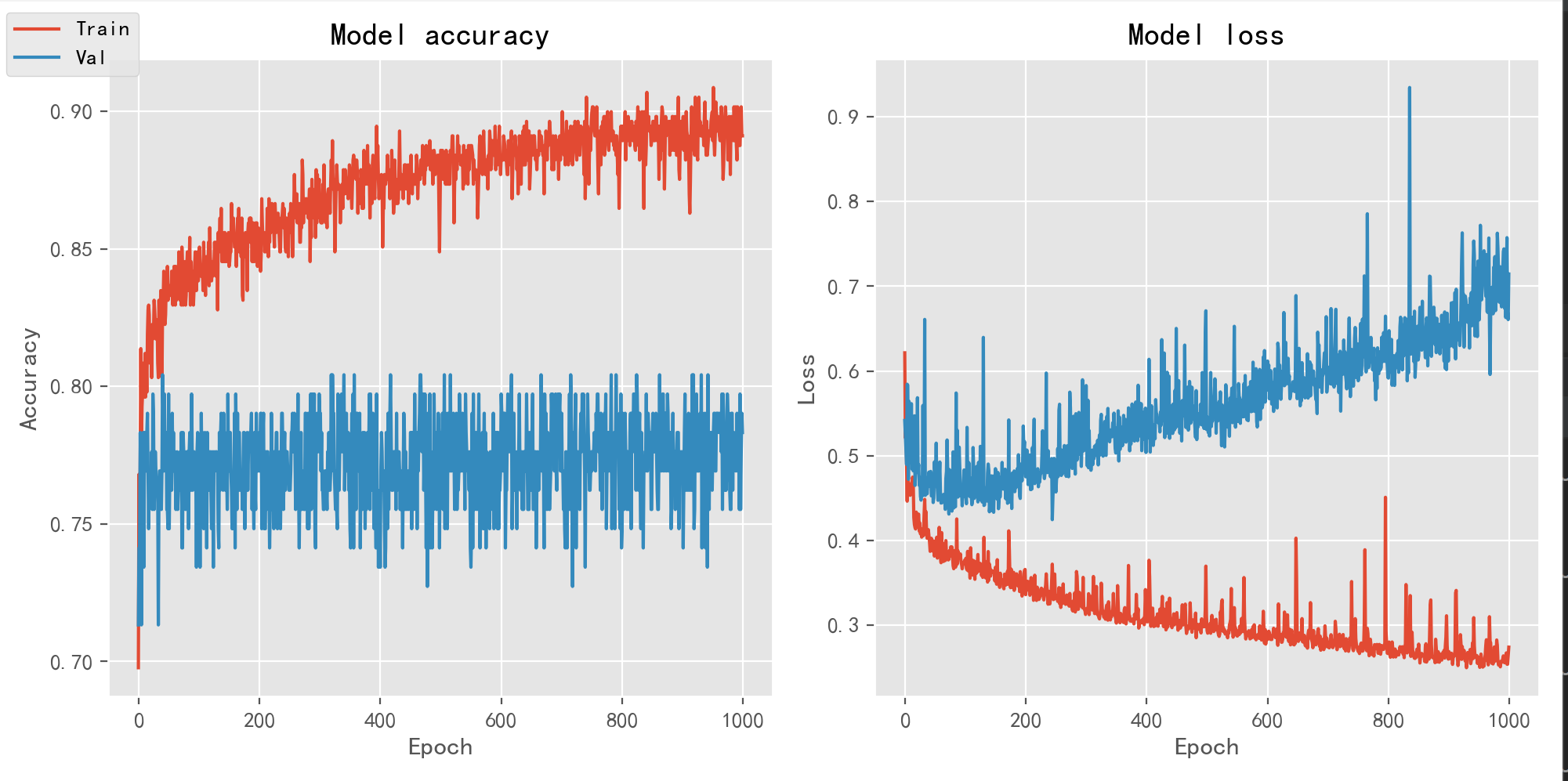
****

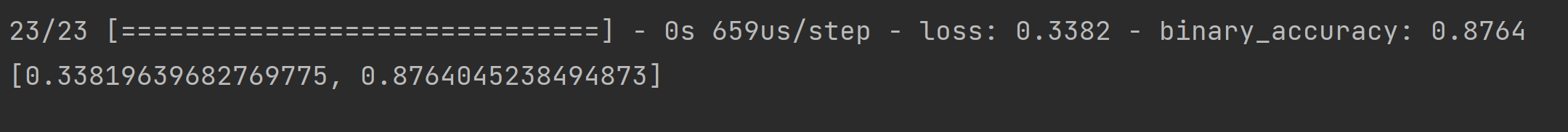
对于一般简单的数据集，一两层隐藏层通常就足够了。

通常，**对所有隐藏层使用相同数量的神经元就足够了**。对于某些数据集，拥有较大的第一层并在其后跟随较小的层将导致更好的性能，因为第一层可以学习很多低阶的特征，这些较低层的特征可以馈入后续层中，提取出较高阶特征。

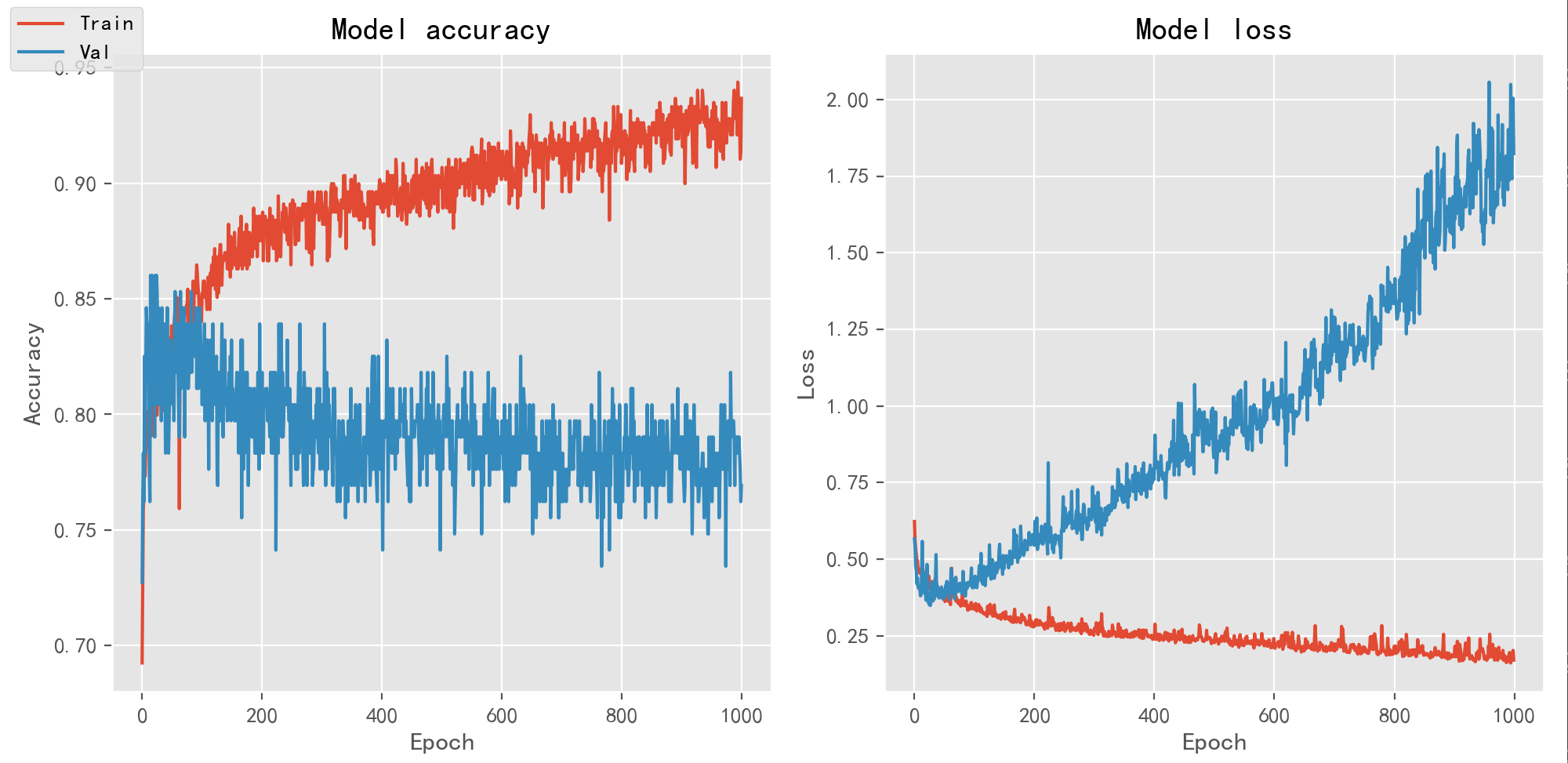
基于此，先对神经网络的层数和每层的节点数进行探究

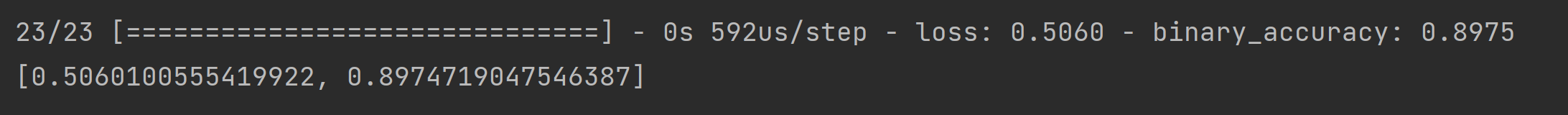
**一层：**

****

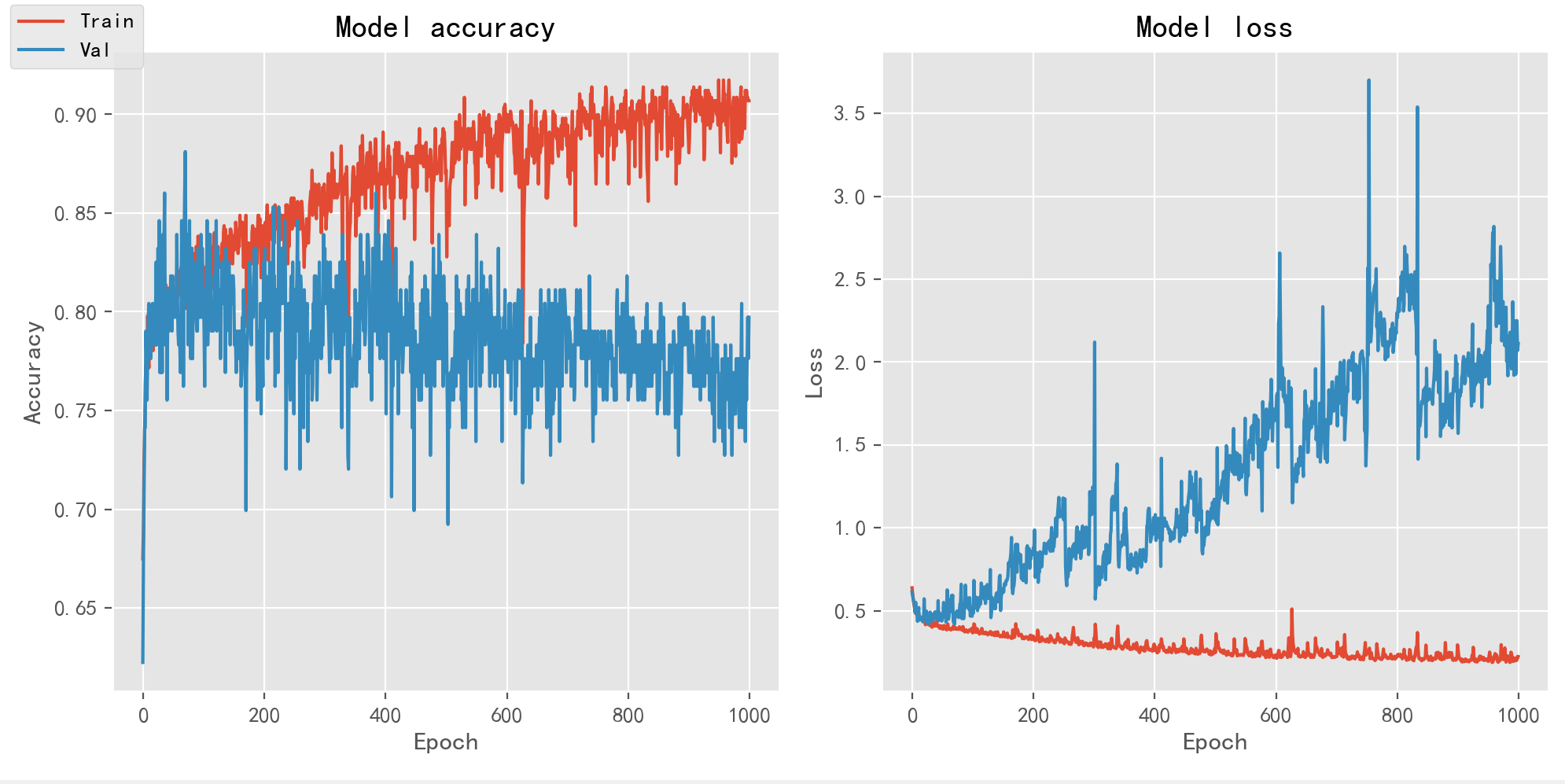
****

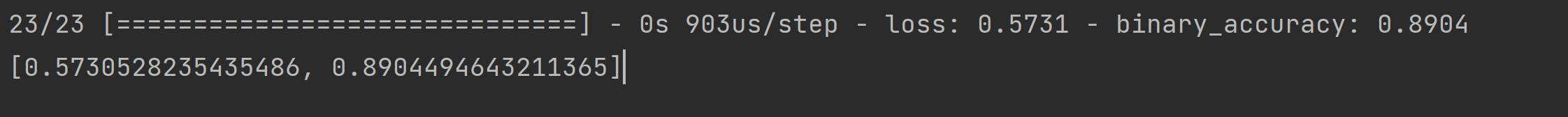
**第二层：**

****

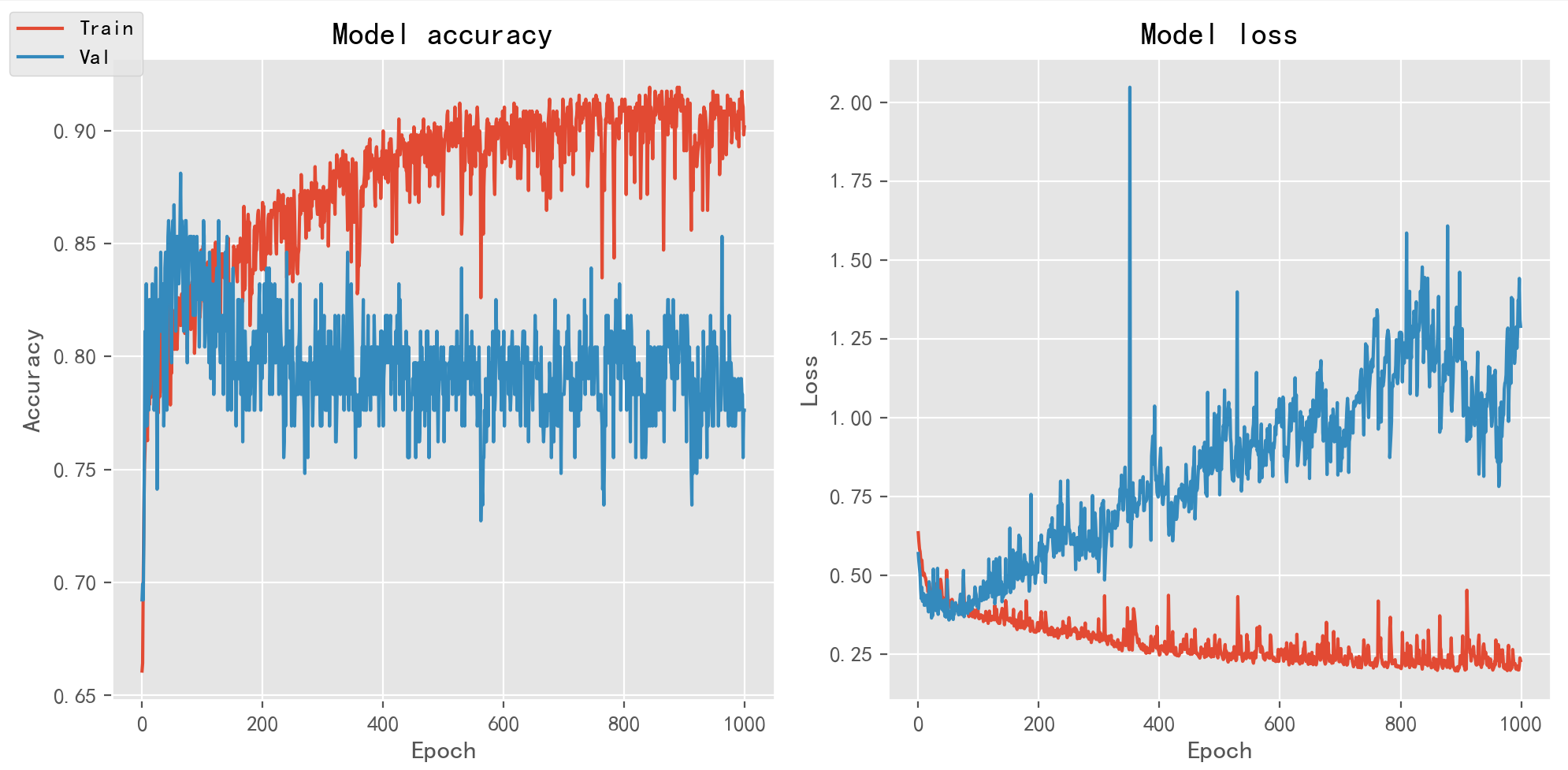
****

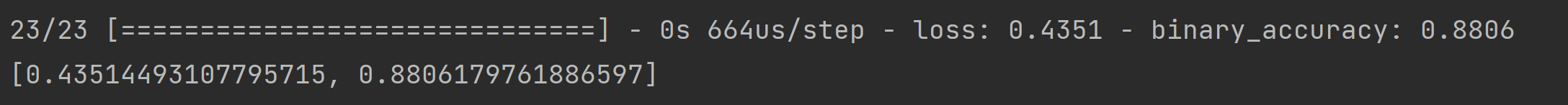
**三层：**

****

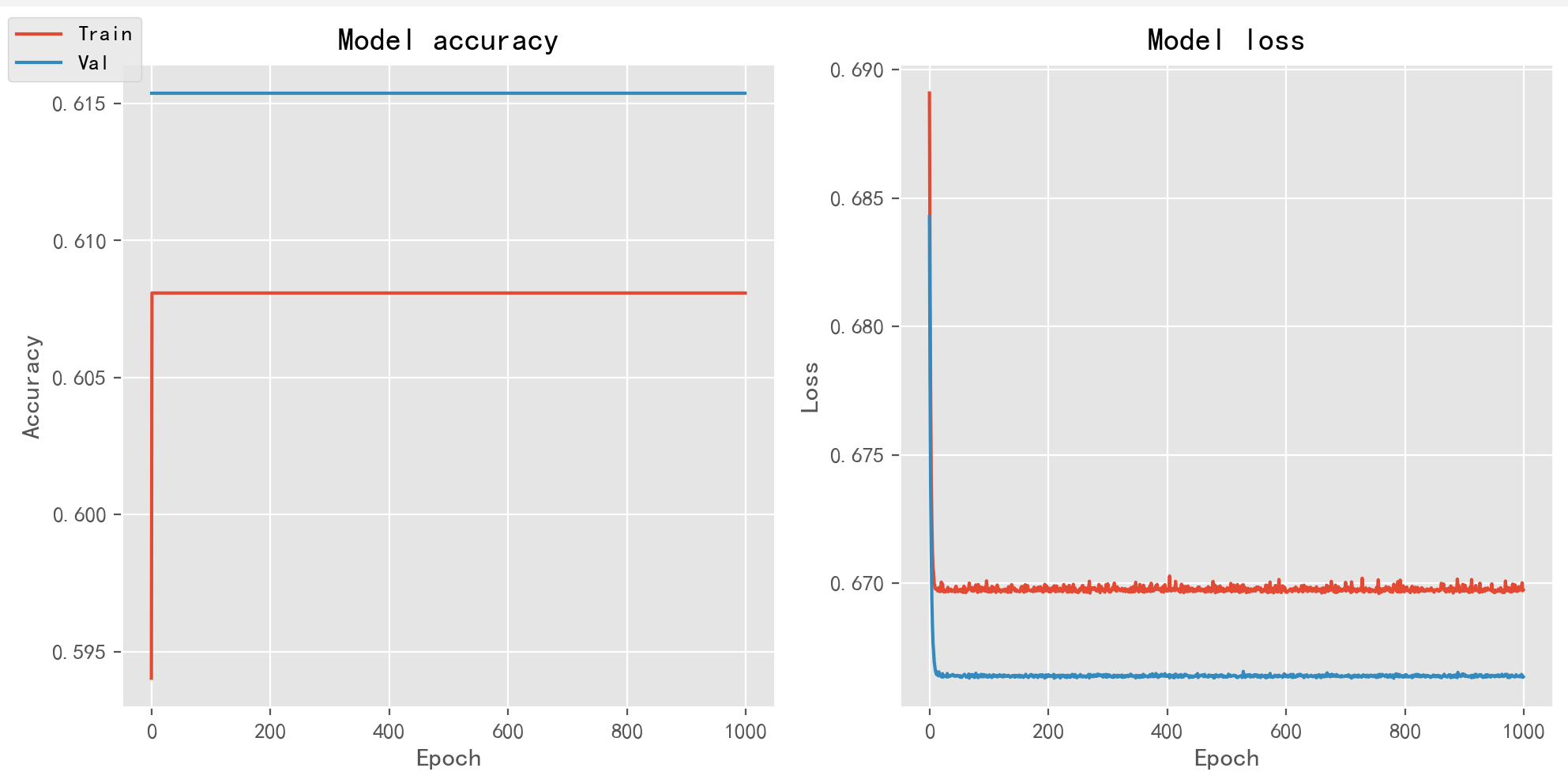
****

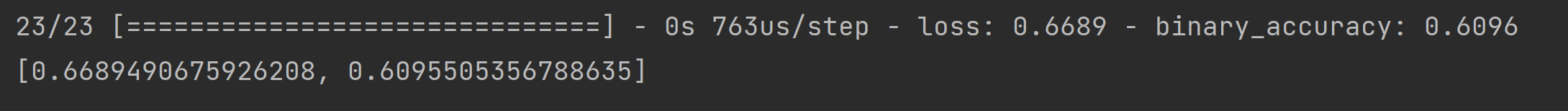
**四层：**

****

****

**五层：**

****

****

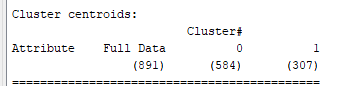
### 集成策略等方案改进

## 数据降维：

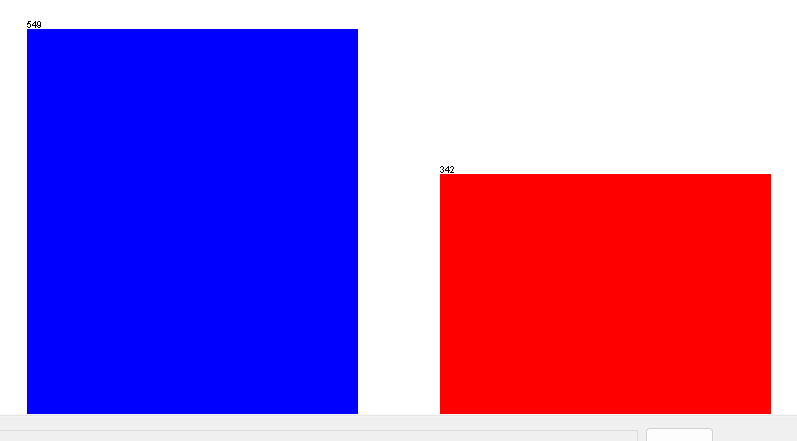
使用weka软件通过简单的KMeans无监督聚类方法对输入属性评估：

簇数：2

聚类结果：

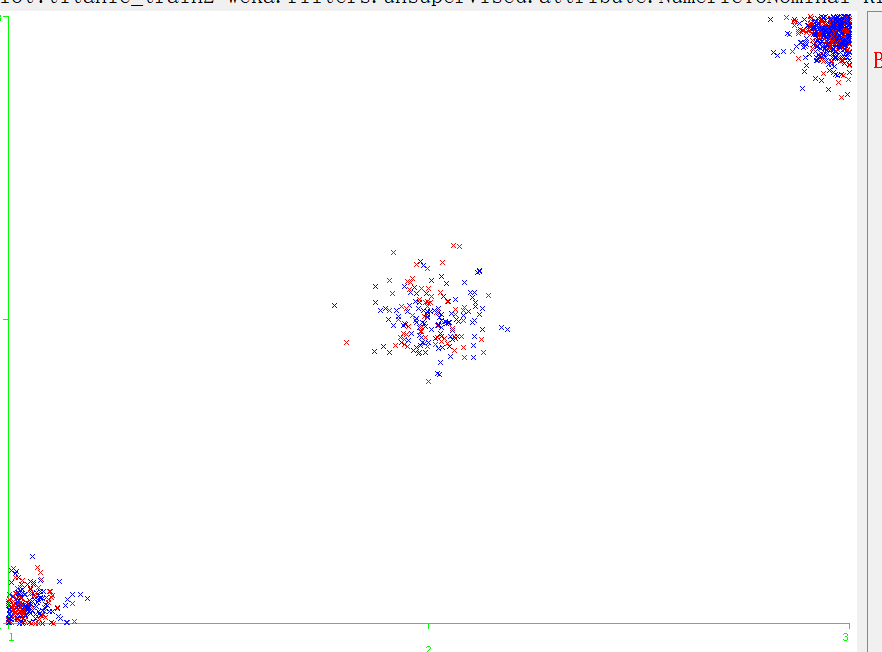


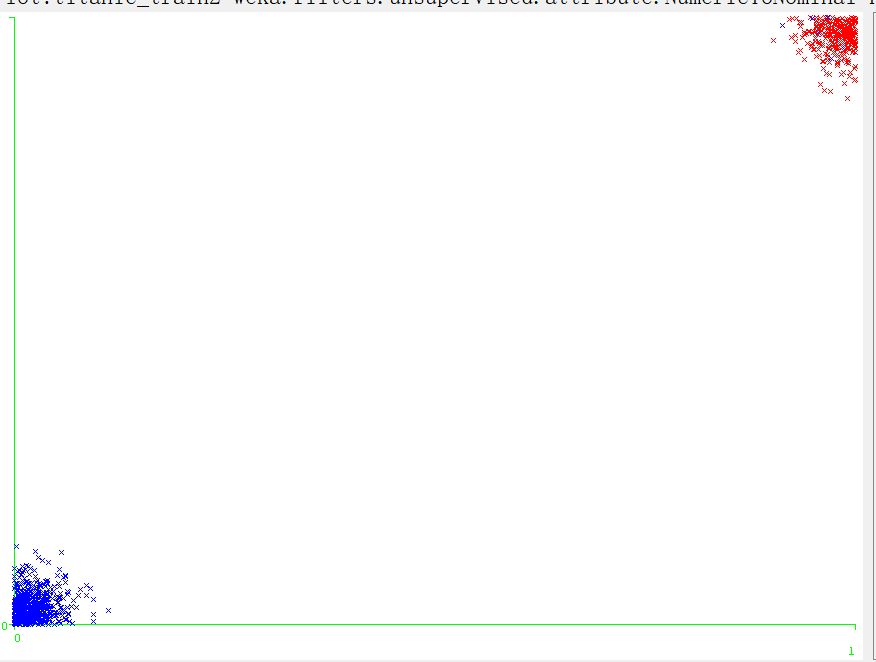
实际分类：

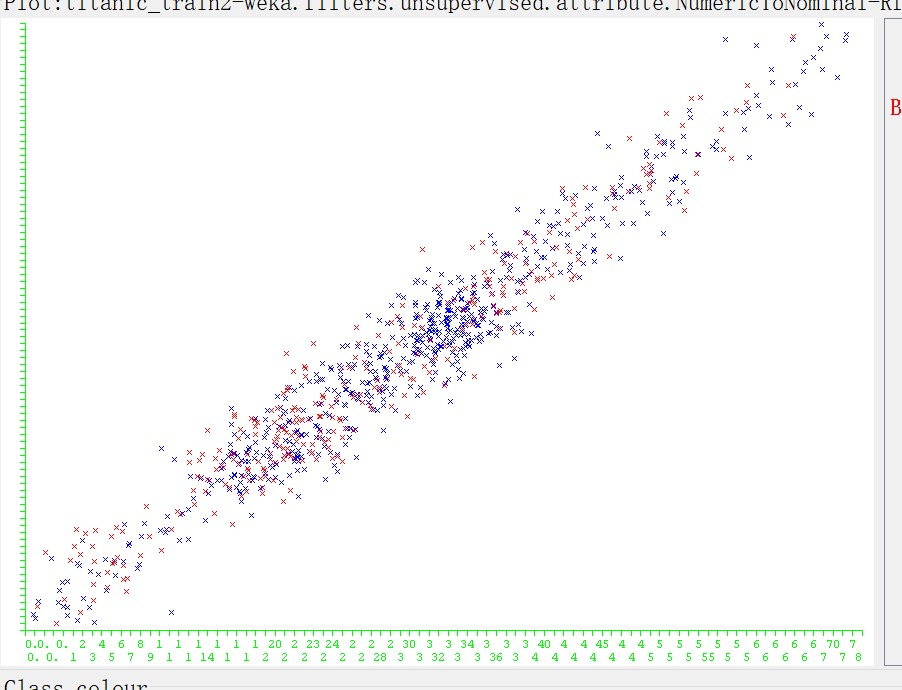


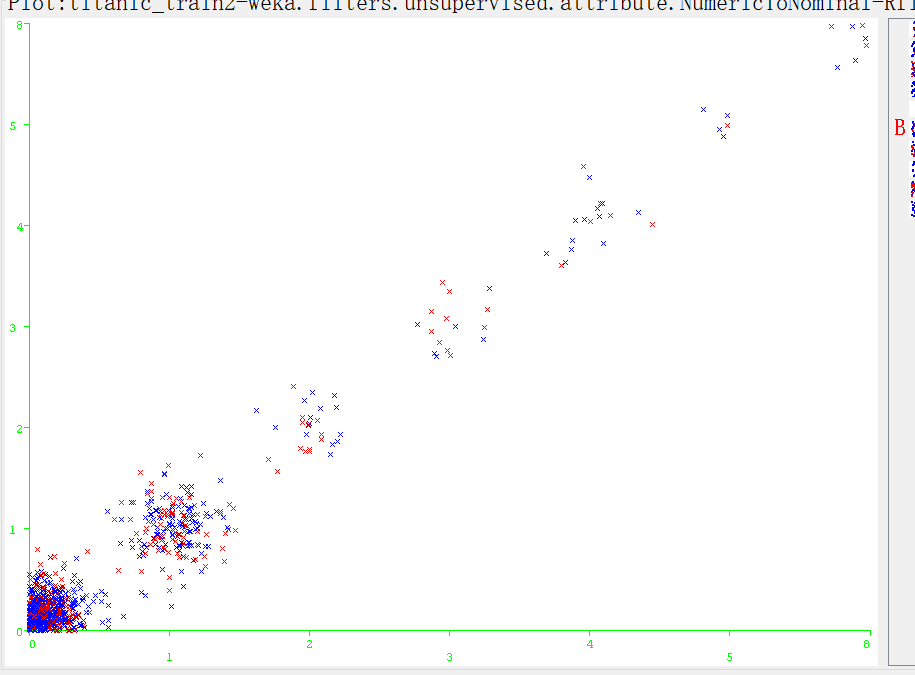
形成了大小与实际分类近似相等的簇，可以认为聚类所形成的簇具有较高的质量，初步判定输入属性对于实际分类能力应该是较强的。

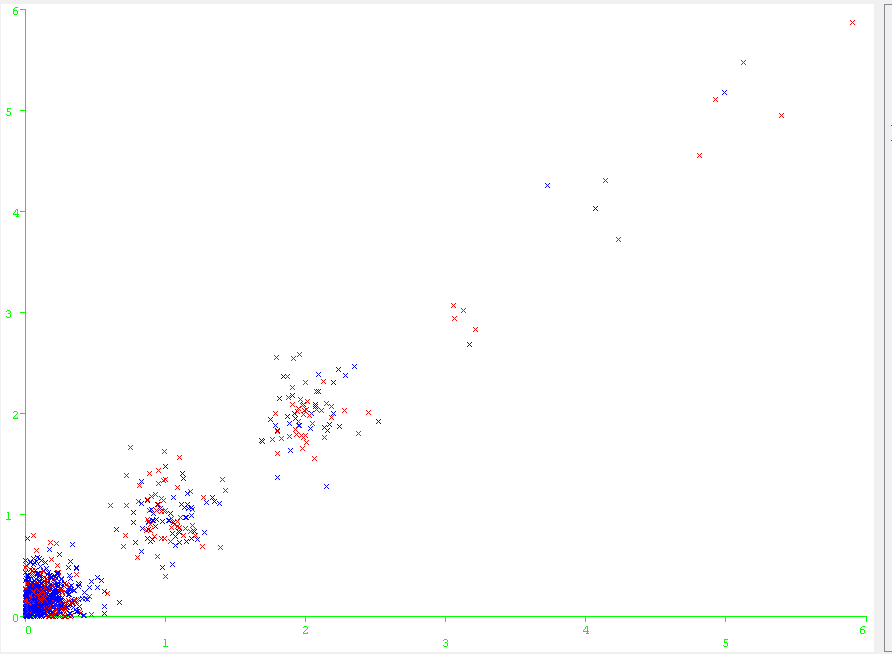
查看所有属性的分类能力：

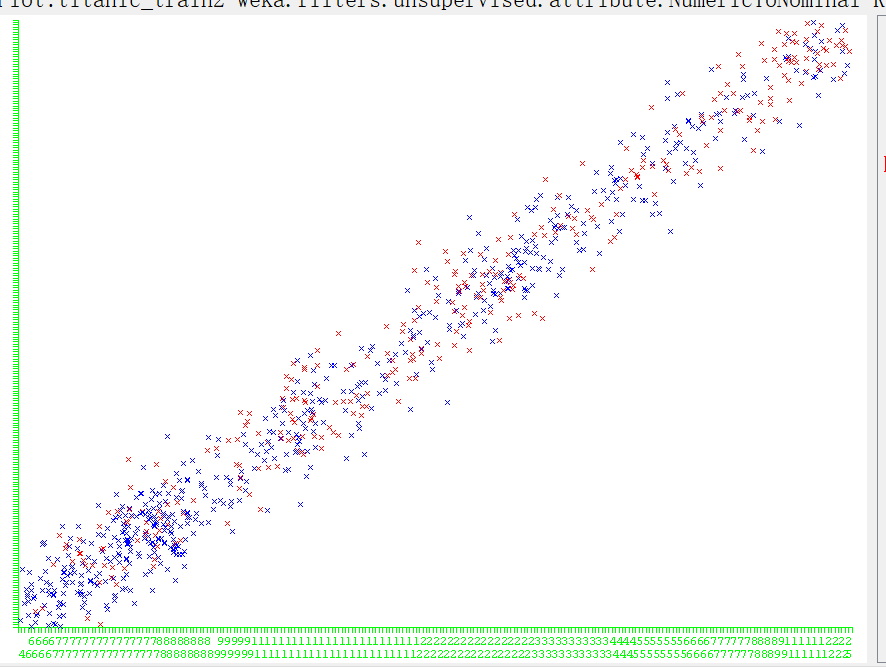
Pclass：

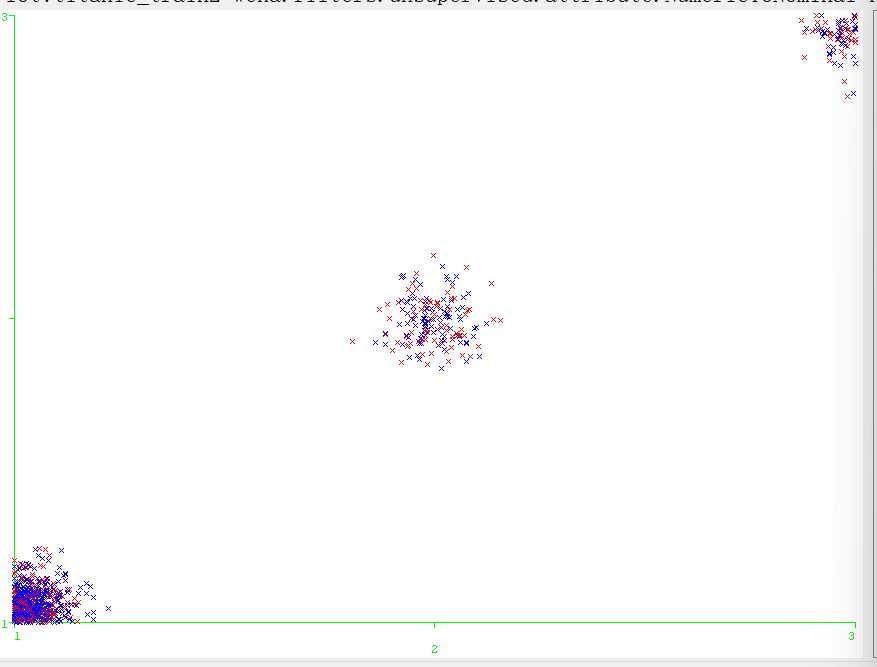
Sex: 

Age：

SibSp: 

Parch：

Fare：

Embarked: 

在这里可以看出，除了性别之外，其它属性分类能力都不咋好。

在Fare、Age有一丢丢的分类能力，SibSp和Parch很相似。

所以我只保留sex、age、Fare、SibSp和parch其中一个。