



**数据挖掘**

**上机实验报告**

|  |  |
| --- | --- |
| 专 业： | 人工智能 |
| 班 级： | 192204 |
| 学 号： | 19220432 |
| 姓 名： | 陆昊宇 |

2024年09月14日

## 一、实验目的

1. 掌握使用python探索数据集属性的方法。

## 二、实验内容

#### 1. 泰坦尼克号数据集分析实验

（1） 自由选择可视化方法，可视化数据集中的Pclass属性，Sex属性，Age属性，SibSp属性，Fare属性和Embarked属性；

（2）回答以下问题 , 可考虑使用中心趋势度量和离散度度量进行分析：

1）年龄与生存情况之间的关系是什么？

1. 性别与生存情况之间的关系是什么？
2. 社会地位与生存情况之间的关系是什么？
3. 为什么有些票的价格很高？票价的高低与哪些属性有关？

**注**：对于每个问题给出理由，有图表数据支撑更好。

**提交入口：**

<https://send2me.cn/ozvY6cGJ/Rv69hFiyffA4JQ>

## 三、实验报告评分标准

1. 完成实验（1）（10分）

2. 完成实验（2）（60分）

4. 实验过程与实验结论的记录完整（10分）

5. 表述逻辑清晰（10分）

6. 排版工整自洽，图表标题准确（10分）

## 四、实验过程

本实验使用pandas库处理数据，matplotlib库绘图。

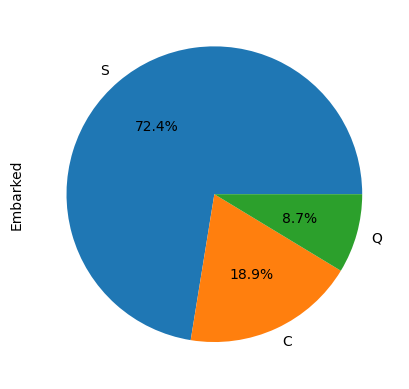
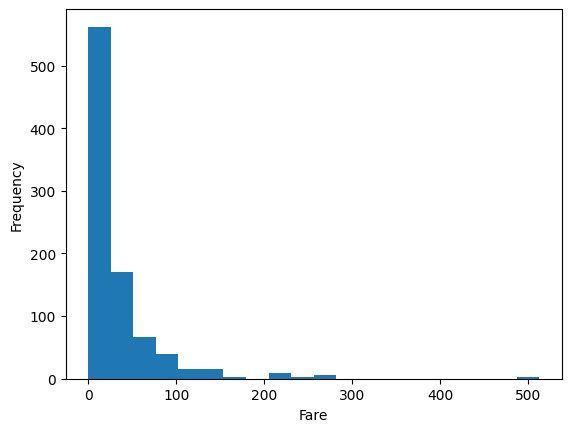
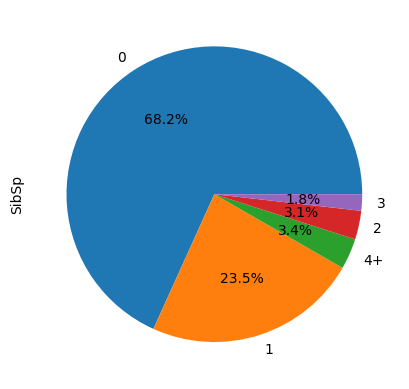
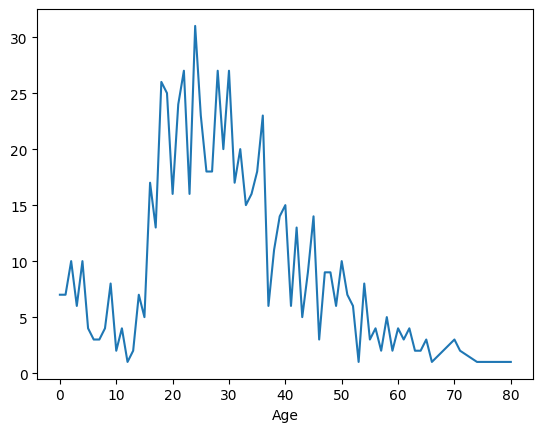
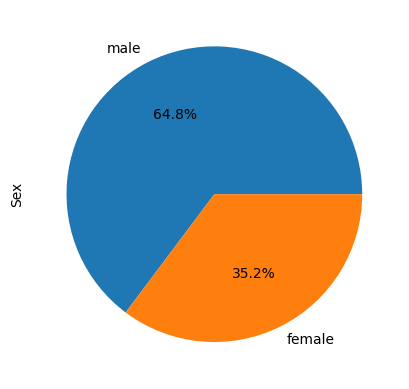
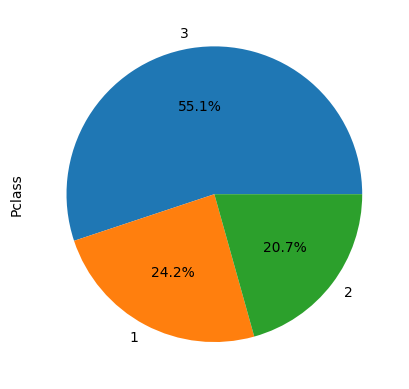
读入数据后，可视化了数据集中的Pclass，Sex，Age，SibSp，Fare和Embarked属性。

为了探究各属性与生存情况之间的关系，使用条件索引及value\_counts函数计算各部分的情况，缺失值使用dropna处理。对Sex, Age等列进行了必要的整理，比如把字符串编码为数值数据，浮点型整理到区间。

本实验通过绘图的手段得到如下文的直观结果。具体代码参考 [github链接](https://github.com/Mr-LUHAOYU/DataMining/blob/main/Project1/main.ipynb).

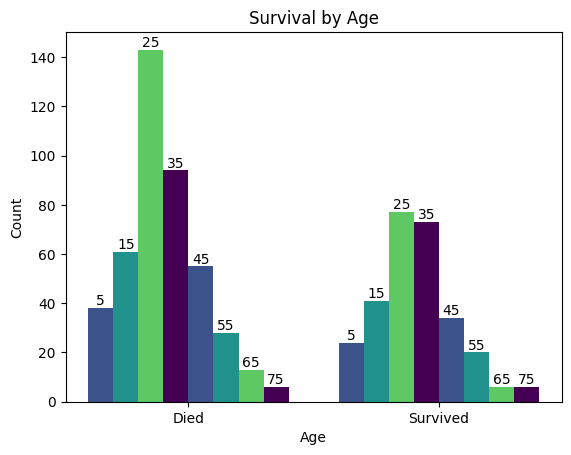
## 五、实验结论

（1） 自由选择可视化方法，可视化数据集中的Pclass属性，Sex属性，Age属性，SibSp属性，Fare属性和Embarked属性；

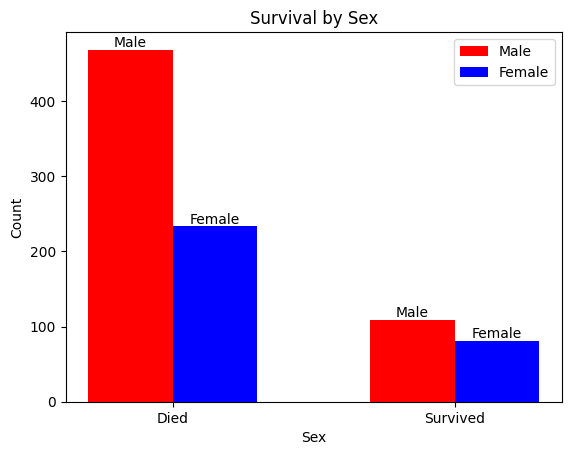


（2）回答以下问题 , 可考虑使用中心趋势度量和离散度度量进行分析：

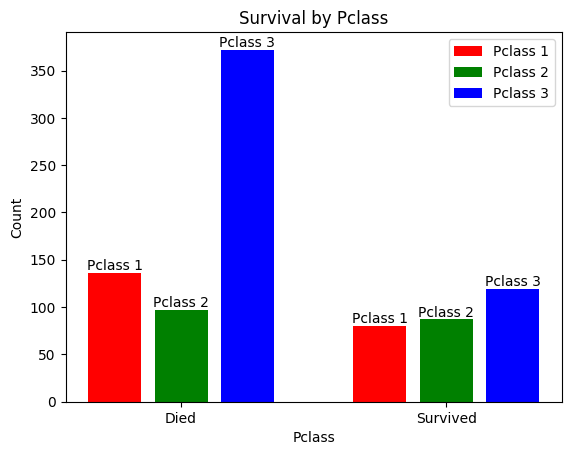
1）年龄与生存情况之间的关系是什么？答：没有显著关系



1. 性别与生存情况之间的关系是什么？答：女性存活率更高



1. 社会地位与生存情况之间的关系是什么？答：Pclass3的死亡率显著高于其它，说明社会地位高的人更可能存活。



1. 为什么有些票的价格很高？票价的高低与哪些属性有关？ 答：与社会地位、家庭结构以及性别有较强的相关性。

