【项目15】 泰坦尼克号获救问题

数据来源：Kaggle数据集 → 共有1309名乘客数据，其中891是已知存活情况（train.csv），剩下418则是需要进行分析预测的（test.csv）

字段意义：

PassengerId: 乘客编号

Survived :存活情况（存活：1 ; 死亡：0）

Pclass : 客舱等级

Name : 乘客姓名

Sex : 性别

Age : 年龄

SibSp : 同乘的兄弟姐妹/配偶数

Parch : 同乘的父母/小孩数

Ticket : 船票编号

Fare : 船票价格

Cabin :客舱号

Embarked : 登船港口

目的：通过已知获救数据，预测乘客生存情况

研究问题：

1、整体来看，存活比例如何？

要求：

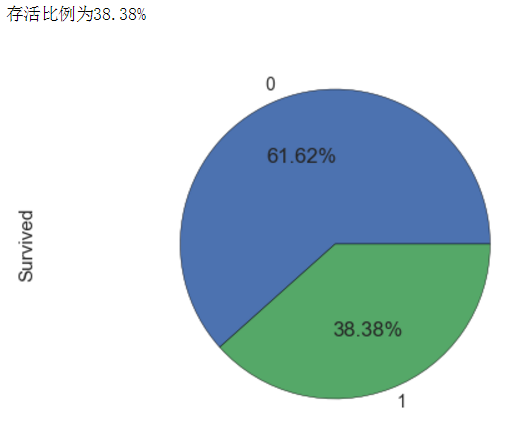
① 读取已知生存数据train.csv

② 查看已知存活数据中，存活比例如何？

提示：

① 注意过程中筛选掉缺失值之后再分析

② 这里用seaborn制图辅助研究



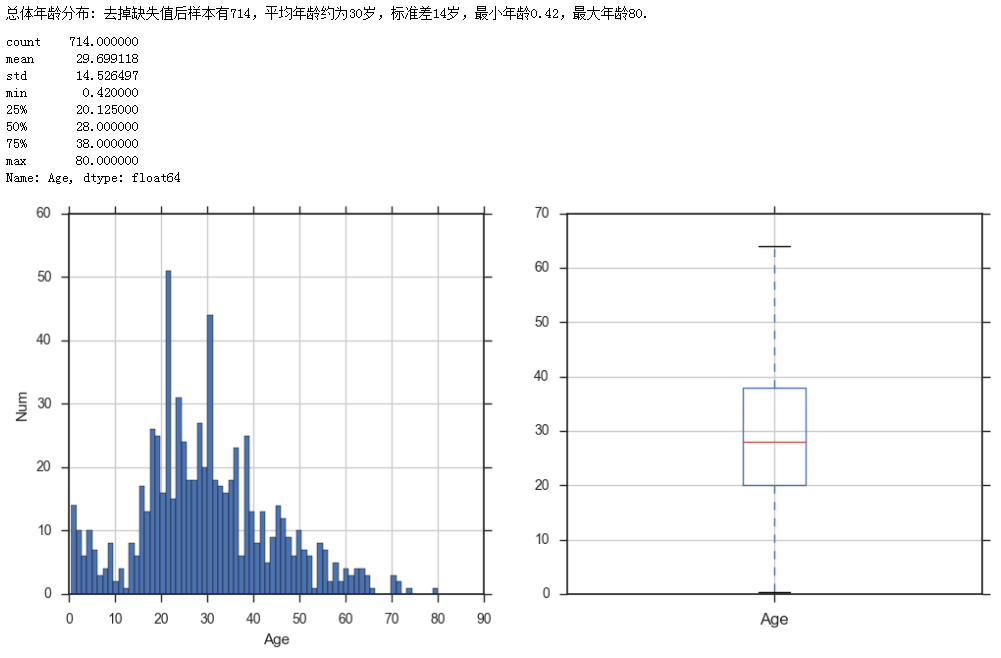
2、结合性别和年龄数据，分析幸存下来的人是哪些人？

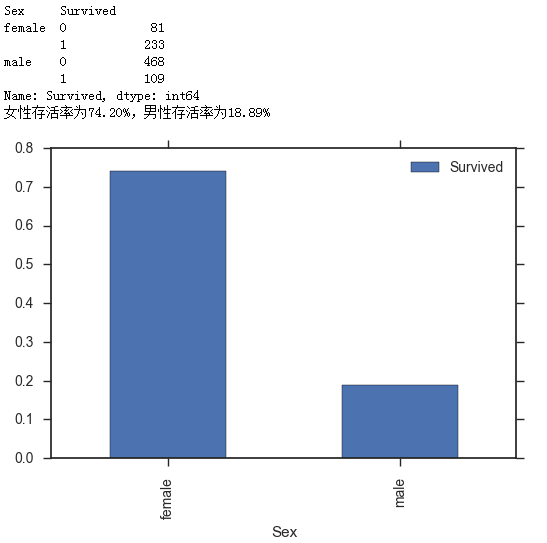
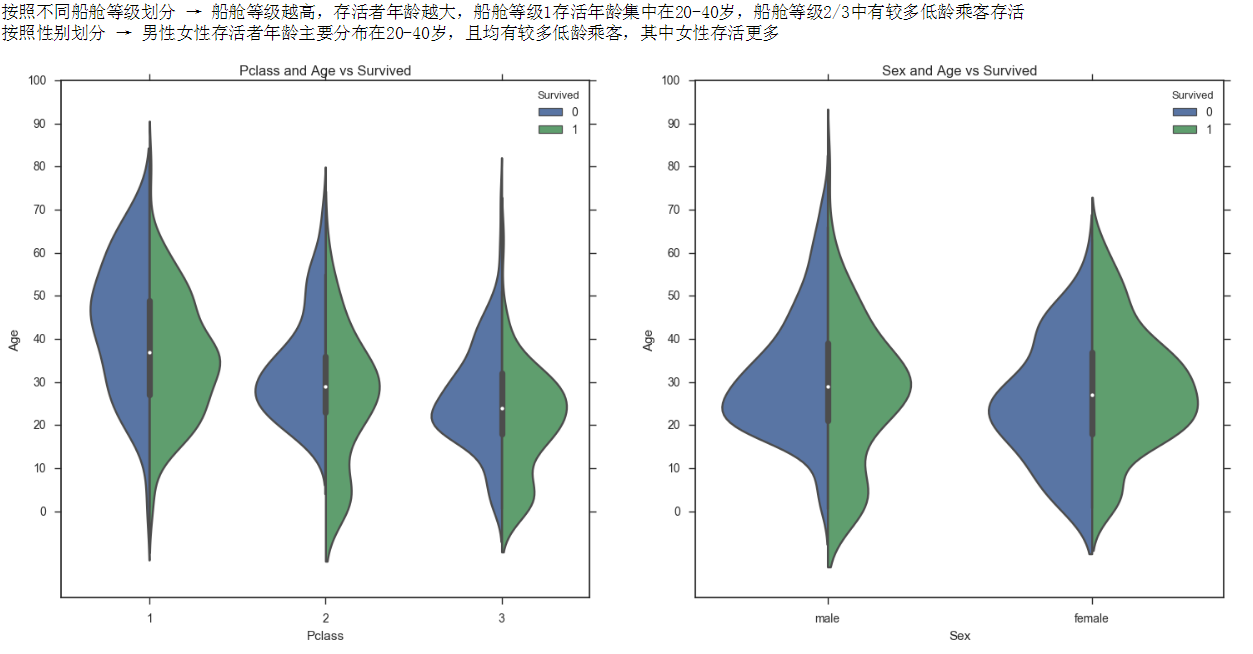
要求：

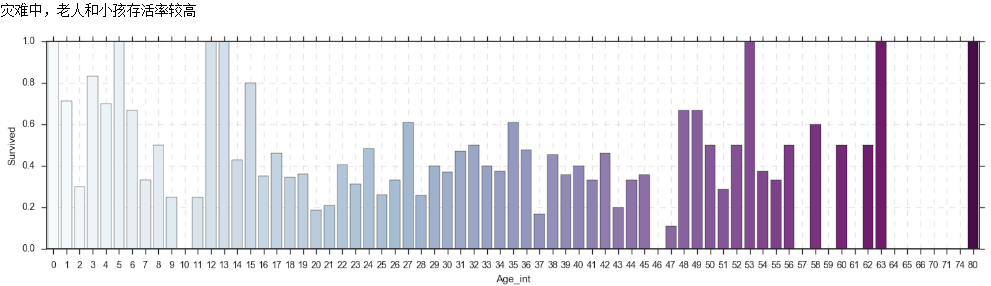
① 年龄数据的分布情况

② 男性和女性存活情况

③ 老人和小孩存活情况



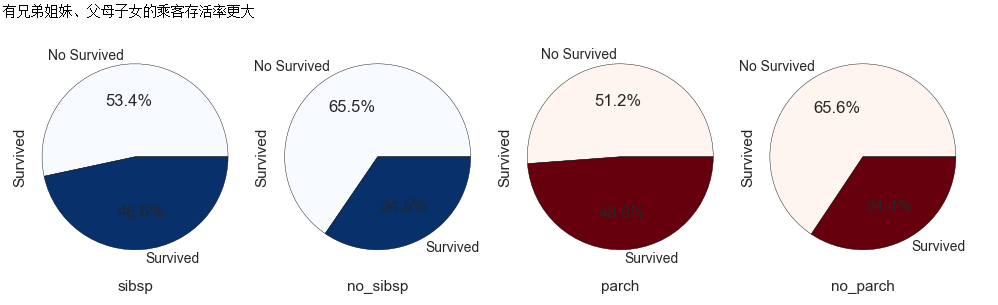


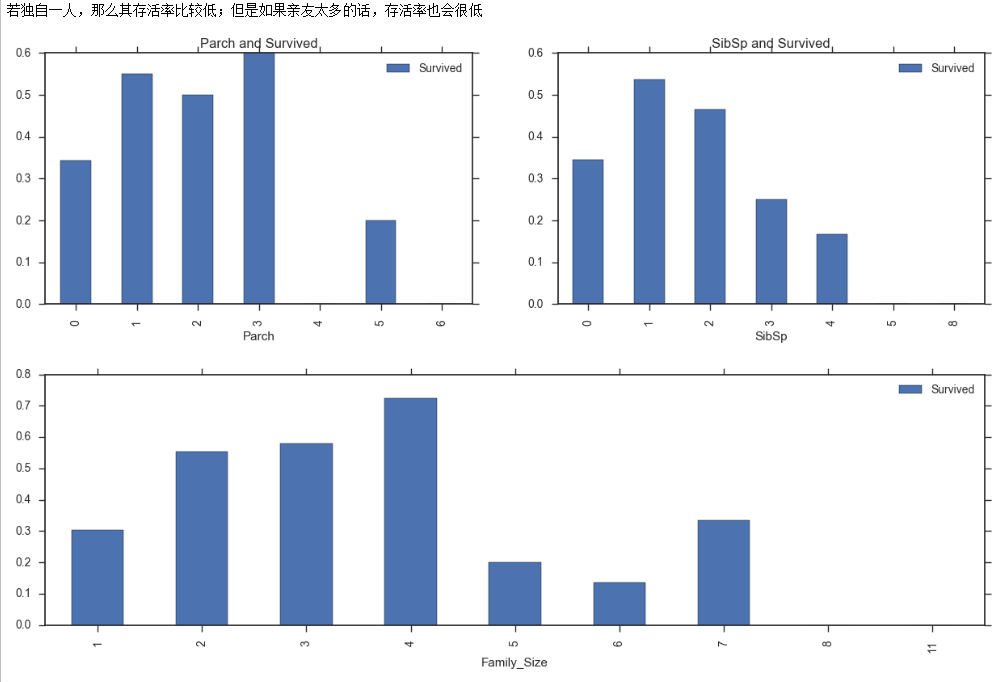
3、结合 SibSp、Parch字段，研究亲人多少与存活的关系

要求：

① 有无兄弟姐妹/父母子女和存活与否的关系

② 亲戚多少与存活与否的关系



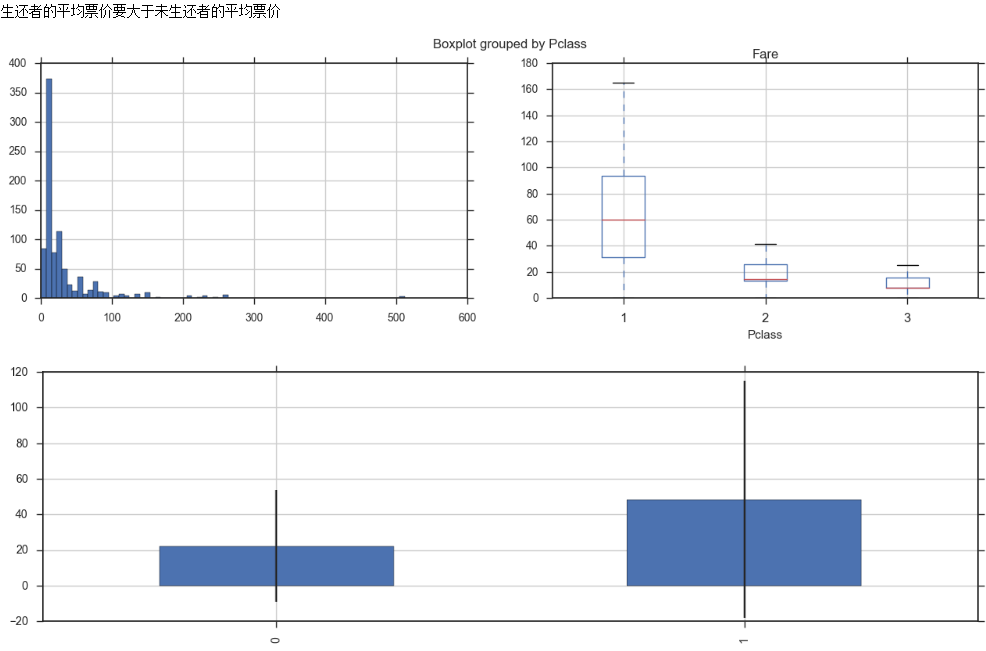


4、结合票的费用情况，研究票价和存活与否的关系

要求：

① 票价分布和存活与否的关系

② 比较研究生还者和未生还者的票价情况



5、利用KNN分类模型，对结果进行预测

要求：

① 模型训练字段：'Survived','Pclass','Sex','Age','Fare','Family\_Size'

② 模型预测test.csv样本数据的生还率

提示：

① 训练数据集中，性别改为数字表示 → 1代表男性，0代表女性