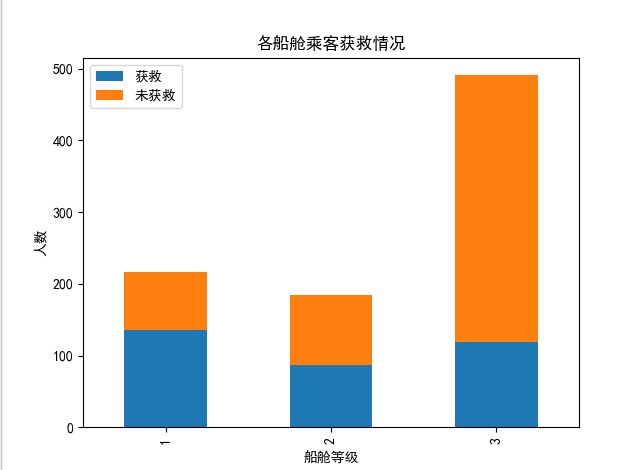
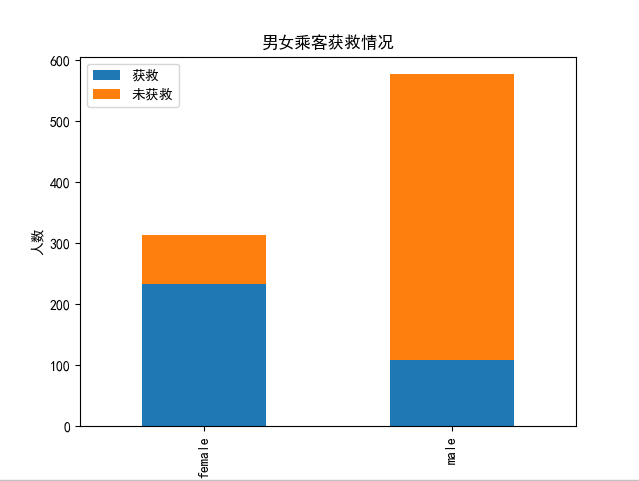
1. 数据可视化



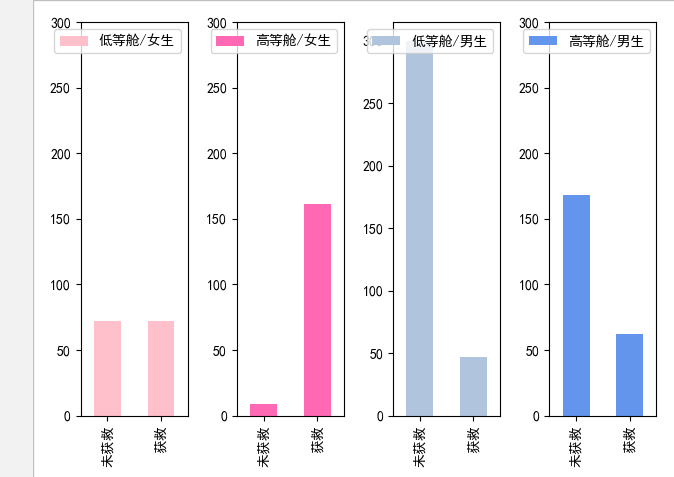
各船舱乘客获救情况：获救情况与船舱等级有关

船舱等级越高，获救几率越大



男女乘客获救情况：获救情况与男女性别有关

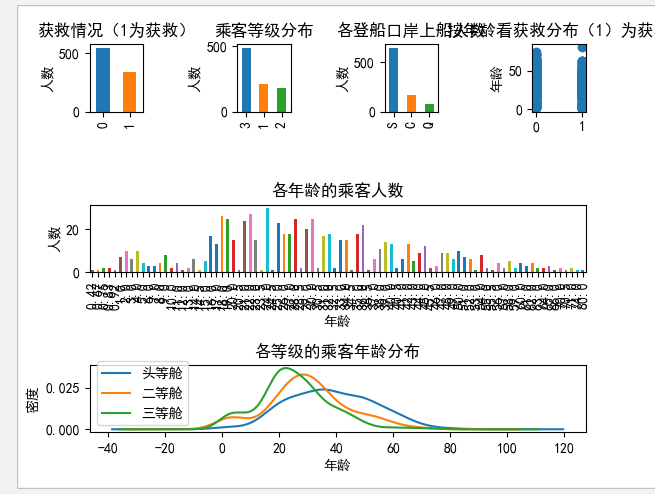
女性比男性更容易获救。



各船舱级别下男女获救情况：论证上述两个观点

①获救情况与性别有关，女性比男性更易获救。

②获救情况与船舱等级有关，船舱等级越高越容易获救。



高等舱的平均年龄大于低等舱

获救几率与年龄无关（从年龄看获救分布的图中，0/1两竖形图分布差距不大）

代码见data\_view.py

1. 数据预处理

**发现Age(年龄)缺失的数据较少，所以采用插值法来处理。用到的插值方法是随机森林  
而Cabin（船舱号）则缺失了约3/4，而船舱号码暂时看不出来对获救有什么影响  
所以我选择了对其进行数值化，有船舱号的值记为为1，没有船舱号的记为0  
同样，为了方便后面的建模，对非数值的属性Sex（性别），Embarked（登舱口），Name（名字），Ticket（票号）进行同样的数值化处理。这里用到了pandas库中的dummies函数。最后将将Age和Fare两个属性进行零——均值标准化。**

具体代码见data\_pre\_deal.py

1. 建模
2. 逻辑回归做拟合：0.7676

具体代码见data\_reg\_model.py

1. svm模型拟合：0.77511

具体代码见data\_svm\_model.py

1. 流程图

