**泰坦尼克生存概率预测**

项目背景：本项目从数据分析的角度，寻找与生存率项目相关的生存因素，建立逻辑回归模型预测tantic人员生存情况。

**1.提出问题：找出最优模型来预测titanic乘客生存结果**

项目概况：Titanic号的沉没是历史上最惨痛的沉船事故之一。在1912年4月15号，Titanic首次航行便因撞击冰山而沉没。在2224名乘客和船员中，有1502人因事而死亡。这场事故令世界为之震惊，随之便引起了人们对船舶安全的重视。

没有足够多的救生艇是该事故吞噬如此多生命的原因之一。虽然有一些幸运的因素使人们能够幸存，但是似乎某一类人拥有更多的生存率，如妇女，孩子和头等舱室的人。

在该项目中，我将分析哪类人群更容易幸存，更主要的是，我将通过机器学习工具来预测哪些人在事故中幸存下来。

**2.收集数据**：数据来源于[Kaggle's Titanic: Machine Learning from Disaster](https://www.kaggle.com/c/titanic/data)。

**3.数据处理**：

3.1初探数据

首先我们必须对数据有一个大致的了解，该数据有哪些属性，是什么类型的数据，和我们要预测的目标有什么关系

1）Survived是我们要的预测的变量，它是二分类变量，0代表死亡，1代表幸存。其它都是潜在的自变量。必须说明：只有对的变量才能对建立更好的模型，有时候变量越多并不意味着模型越好

2）PassengerId和Ticket变量为随机的唯一的标记值，这对结果并没有影响，应排除在模型之外

3）Pclass是一个代表船票等级的有序变量，1代表头等舱，2代表2等舱，3代表3等舱

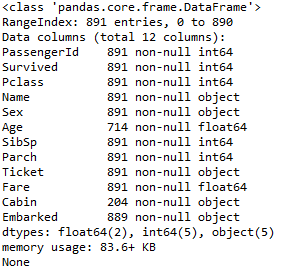
4）Name是一个分类变量，可以从变量Name中看出性别，家庭规模，职位头衔。如果这些信息存在的话，我们可以用于模型中

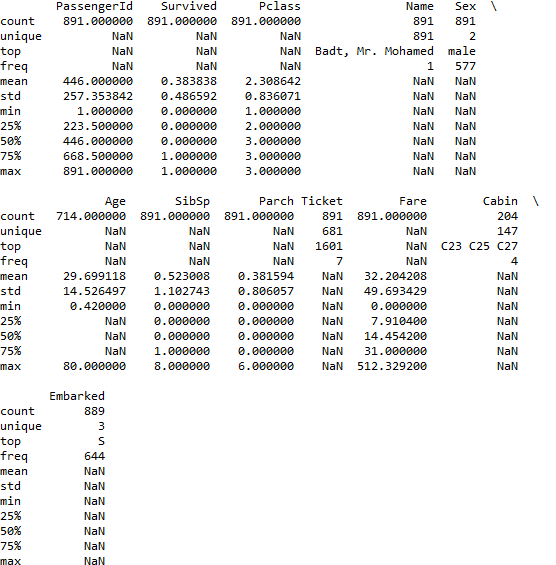
5）Sex和Embarked是分类变量，他们可以通过数学方法转化为虚拟变量

6）Age和Fare是连续型变量

7）SibSp代表船上同辈人数的数量，Parch代表父辈和孩子的数量，都是离散型变量，将他们组合后可以重新变成家庭规模的变量

8）Cabin为分类变量，他们用来显示其在船舶的位置或甲板位置，因为他们有太多缺失值，我们可以将其从数据模型中排除。

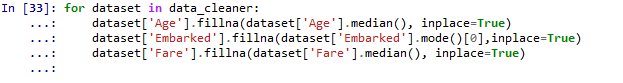




总结：训练集共有891行，其中Age有177空缺值， Cabin共有681个空缺值，Embarked共有2个空值。存活率在25%左右。

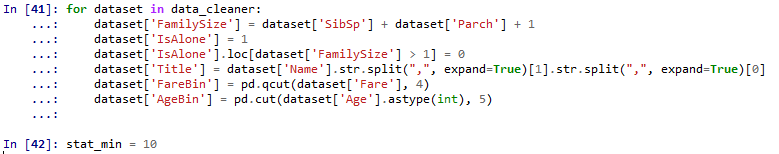
3.2数据清洗

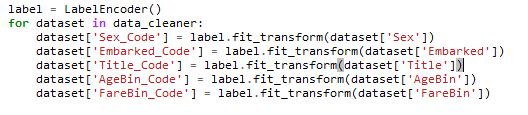
处理噪声数据，处理缺失值，创造新的特征值，转化特征值



用Age的中位数填充空缺值，用Embarked的众数填充空缺值，用Fare的中位数填充空缺值，其中Cabin的空缺值太多，填充难度太大。

创建新的属性值，IsAlone，FareBin, AgeBin.

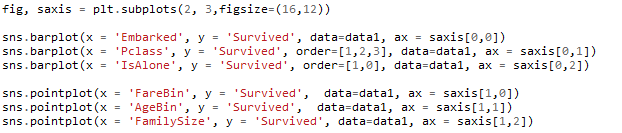


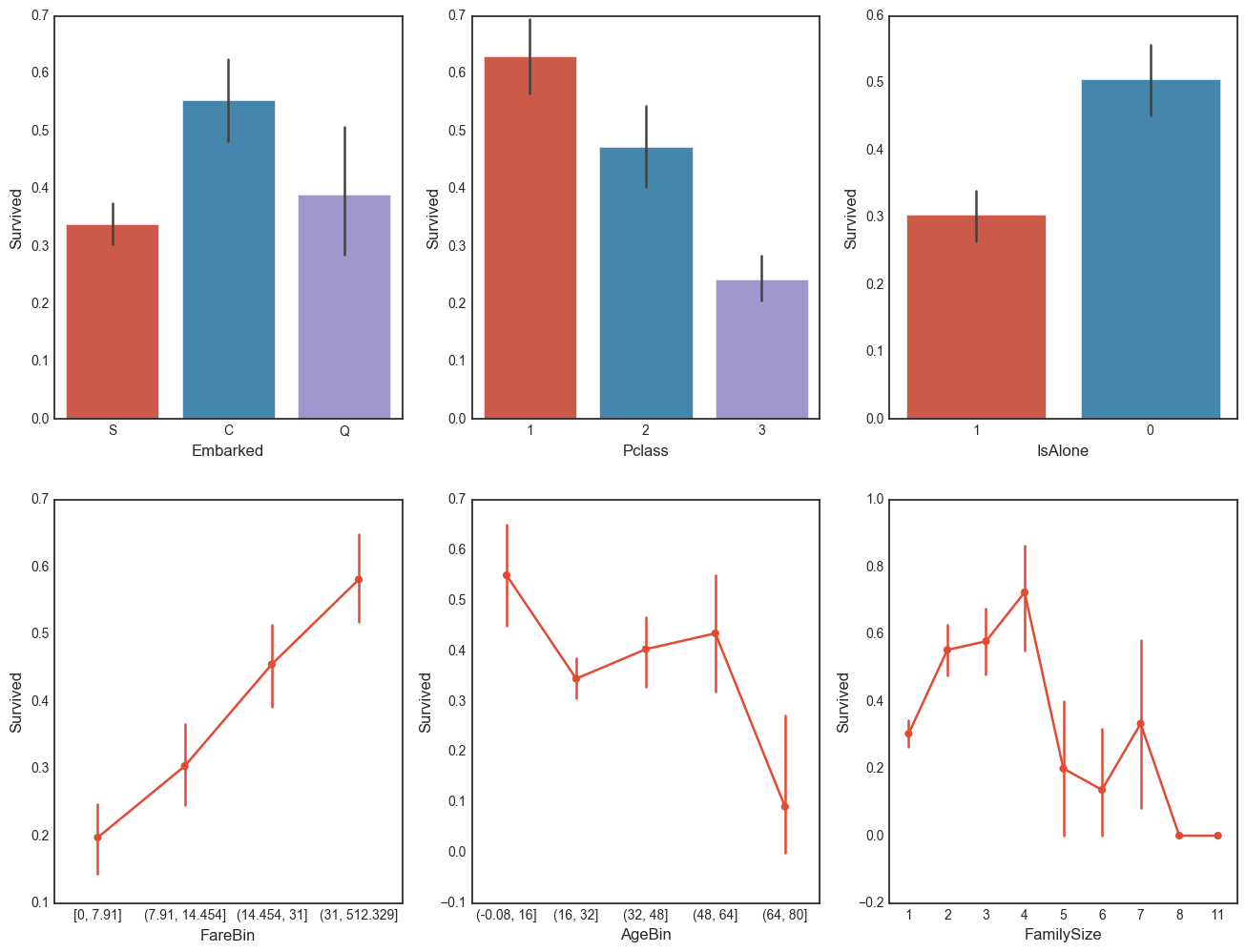


将分类型数据转化为多个二分类数据，有利于模型的的建立。

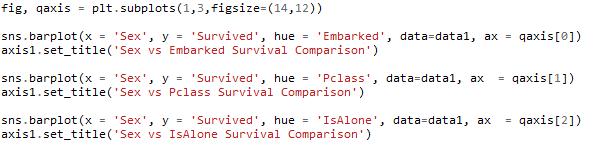
**4.探索数据，寻找特征与目标之间的初步关系，为建立模型做准备**

完成数据清洗的工作后，下面将对数据进行分析和汇总，寻找特征与目标之间的相关关系。

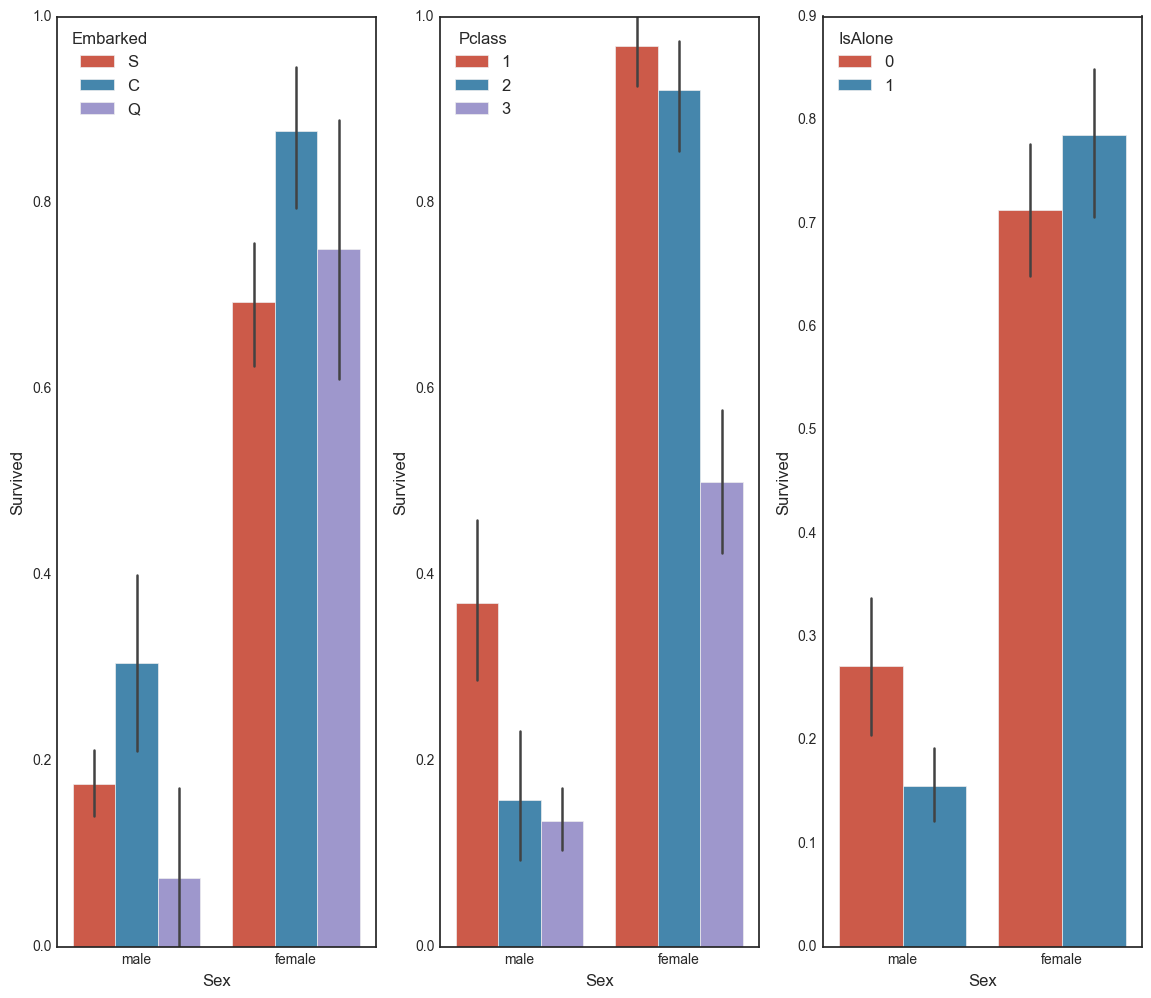


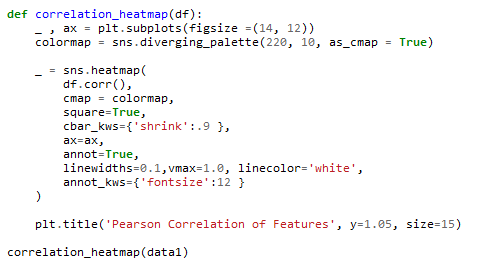


从上图可以看出Embarked（登船港口）C开头的生存率较高 ；从舱室等级来看，舱室登记越高，生存率越大，因为根据船舱布置来说，高等舱室布置在船舶较好的位置，并且救助设施也比较完善，事故发生时，可以快速逃生。结合FamilySize与IsAlone与Survived关系图来看，当家庭规模在2-4之间时，生存率较高，2-4人一起行动时，较为灵活，同时相互帮助，而当规模超过数量时，团队活动受限，不利于逃生。从FareBin来看，工资越高，逃生概率越大，因为FareBin越高，所住的舱室越好，逃生概率越大。从AgeBIn来看，在逃生过程中，婴幼儿受到了较好的照顾，而老人则由于行动不便，逃生概率小。

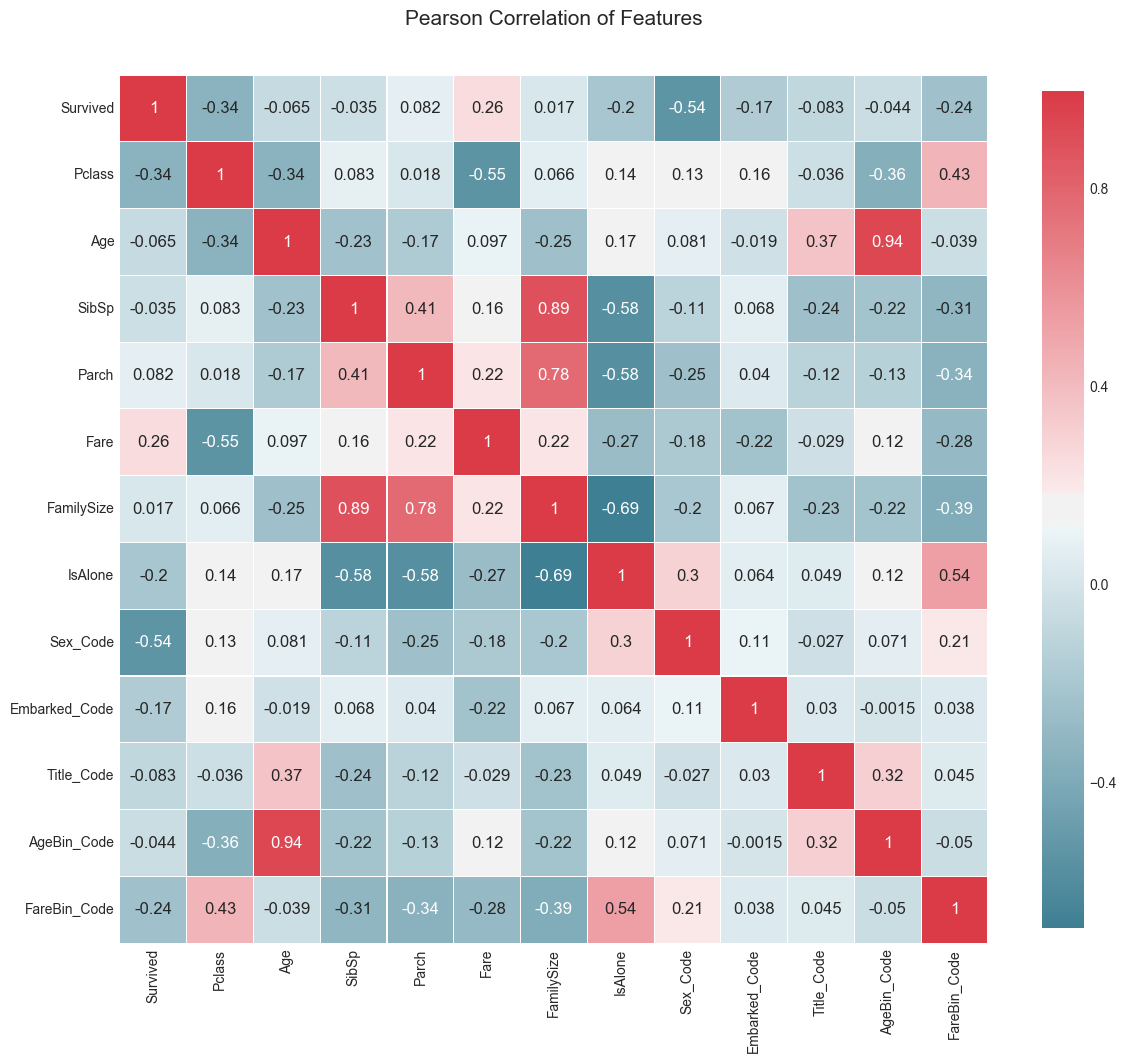


无论是在哪种情况下，女性的逃生概率都高于男性；在事故发生时，女性受到了较好的保护，这是人性的闪光点。





相关关系的热图展示了特征与特征之间以及特征与预测目标之间的相关关系。



**5.建立模型**

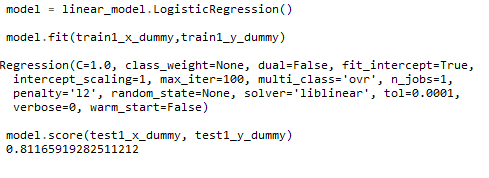
数据科学是一个由数学，计算机科学，以及商业管理结合而成的学科，大多数的数据科学家来源于三者之一。他们总是偏向于他们之前的学科。但是，数据科学好比一个三条腿走路的工具，三者缺一不可。但是不要太过于担心，我们只需要了解核心即可。

正如其名字表达的那样，机器学习是告诉机器怎们做，而不是做什么，大数据这个话题已经发展了十几年，而且越来越受欢迎，因为进入门槛正在变低。这种变化利弊均有，优点是，更多的人可以利用算法来解决更多的现实问题（因为门槛变低），而缺点则是：人们并不知道有时候他们利用这些工具得出的结果是不准确的。

机器学习的目的在于解决人们所遇到的问题，机器学习可以分为监督学习，无监督学习，强化学习。监督学习是人们训练模型的数据包括正确的结果，而无监督学习则不包括。强化学习强调如何基于环境而行动，以取得最大化的预期利益。

当人们进行数据建模时，总喜欢问，最好的机器学习算法是什么，事实上，并没有最好的，只有最适合的。

在这里，我们使用sklearn包里的LogisticRegression模型，模型的正确率为0.81165919282511212



**6.总结**

在该项目中，我通过-提出问题—收集数据—数据清洗—数据初步探索—数据模型等步骤完成了对Titanic逃生率的预测。正确率达到了0.8。总的来说，达到了最初的预期。