目录

[泰坦尼克号幸存者分析报告 1](#_Toc6830292)

[一．项目背景 1](#_Toc6830293)

[二．报告内容 1](#_Toc6830294)

[三．数据源处理 1](#_Toc6830295)

[1.数据一览 1](#_Toc6830296)

[2.对空值的处理 1](#_Toc6830297)

[四．数据分析 1](#_Toc6830298)

[1.乘客存活率与性别的关联 1](#_Toc6830299)

[2.存活率与年龄的关联 2](#_Toc6830300)

[3. sibSp对生存率的影响 2](#_Toc6830301)

[4.Parch对生存率的影响 2](#_Toc6830302)

[5.舱位等级对生存率的影响 2](#_Toc6830303)

[6.登船口对生存率的影响 2](#_Toc6830304)

# 泰坦尼克号幸存者分析报告

## 一．项目背景

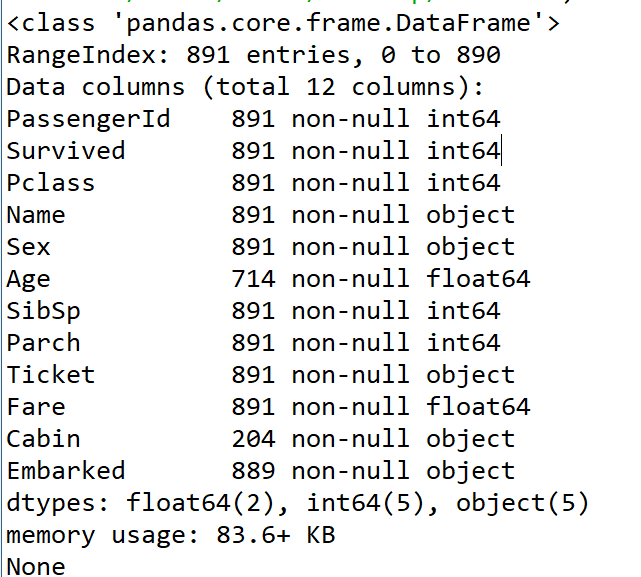
泰坦尼克号在19世纪初是世界上体积最大、内部设施最为豪华的客运游轮，有着“永不沉没”的美誉，然而就是这样一个号称“永不沉没”的邮轮，在它的处女航行之中就遭受厄运。1912年4月14日23时40分左右，泰坦尼克号与一座冰山相撞，造成右舷船艏至船中部破裂，五间水密舱进水。次日凌晨2时20分左右，泰坦尼克船体断裂成两截后沉入大西洋底3700米处。

## 二．报告内容

该报告主要分析的数据集摘自<https://www.kaggle.com/c/titanic/data>，主要阐述乘客的固有属性对最后生存率的影响，并利用python的机器学习库对程序进行数据训练，并展示机器学习后的预测结果与实际结果对比。

## 三．数据源处理

### 1.数据一览

在开始对数据进行分析前，需要先查看数据集中的列属性字段，弄清它们的关系。

PassengerId代表了乘客的编号。

Survived的值代表存活状况，1为存活，0为未存活。

Pclass代表乘客的舱位等级，一共有三个等级的舱位，1级为最好，舱位的豪华程度随数值递增而递减。

Name代表乘客的名字，隐式包含了乘客的社会地位信息。

Sex代表乘客的性别。

Age代表了乘客的年龄。

SibSp的值代表乘客的随行同伴数量。

Parch代表船上父母子女数量。

Ticket为船票的编号。

Fare代表该乘客的船票价格。

Cabin为船舱号。

Embarked为乘客的登船点。

### 2.对空值的处理

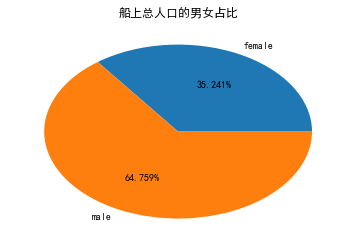
Age字段有一定的缺失，由于年龄对最后的分析结果有着高相关度，因此需要填充空值。使用sklearn的随机森林模型对数值字段进行分析，可以较为准确的填充年龄数值。

Cabin字段的缺失量非常大，并且数据样本的数量不足够多，所以在此不对空值进行处理，在后续的数据分析中不会将该部分的数据计入考虑。

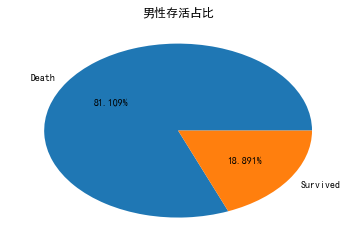
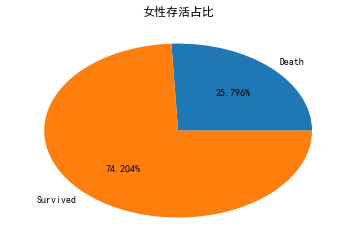
Embarked字段的缺失数量足够微小，对最后的数据分析结果起到的影响不大，可以不用对空值进行填充处理。

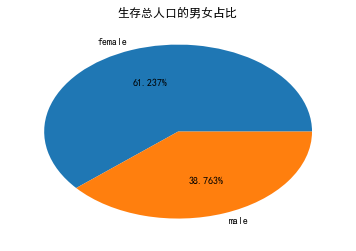
## 四．数据分析

### 1.乘客存活率与性别的关联

首先是船上总人口的男女占比。

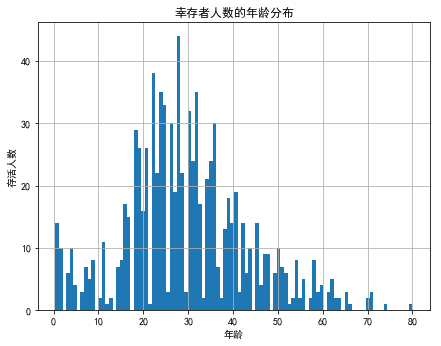
由此图，我们可以很明显的得到船上的男性远大于女性的结论。接下来观察一下男女分别的存活率。

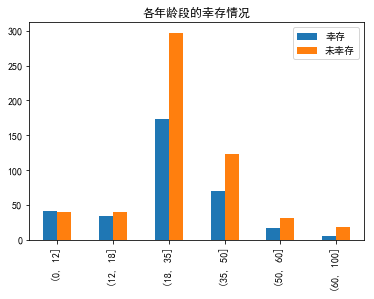


由上述两张饼状图，不难发现女性的存活率要远远高于男性，在史实上确实是船长下令先让妇孺登上救生艇，从而导致了女性的生存率大于男性。接下来是幸存的男女人数对比。

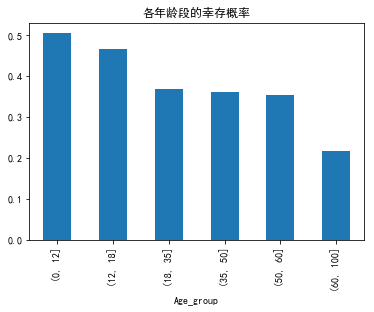
即便是男性人口的基数大于女性，但最终仍然是女性存活的人数居多，可以得到结论，作为女性会在泰坦尼克事件中具有更大的生存可能性。

### 2.存活率与年龄的关联



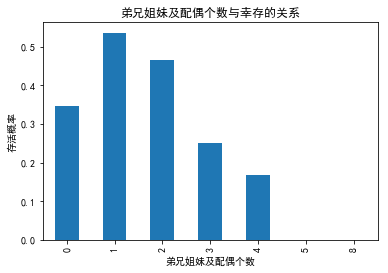
从“幸存者人数的年龄分布”图上的显示，得知存活者的年龄集中在20到35之间，但由于此年龄段的人口基数较多，无法得到年龄与生存率相关的准确结论，将连续的年龄离散，进行分段式观察。

从该柱状图可以很清楚的看到儿童的生存率远高于其他年龄段的人，在条状图中对幸存的概率进行量化，更进一步地体现了不同人群的生存率差异，由此可以得出年龄确实对幸存率起到了一定的影响。

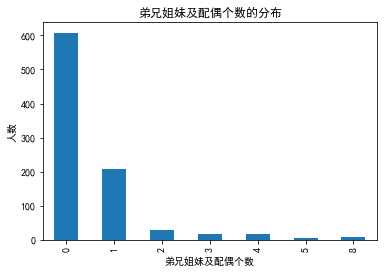


### 3. sibSp对生存率的影响

俗话说人多力量大，那么是不是同伴的数量越多，就更有可能在这场危机中存活下来呢？



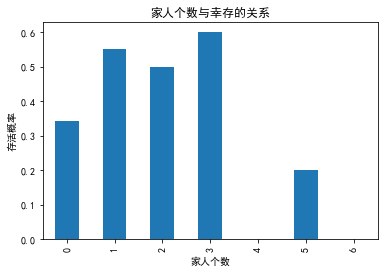
经过对数据的统计量化后，上述图表直观的体现了弟兄姐妹及配偶个数与幸存概率之间的关系，不难发现sibSp的数量对存活的确有一定的帮助。

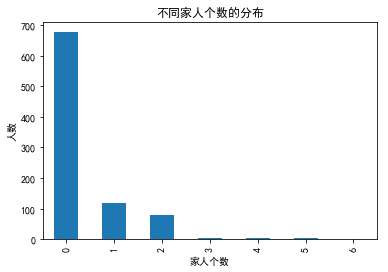


由上述柱状图图我们可以看到sibSp个数多的乘客并不是很多，所以我们由上图得到的sibSp个数多会对生存有所帮助的结论可能并不是十分准确，但还是能够知道当sibSp数量较少时，生存率会比其他情况高。

### 4.Parch对生存率的影响

证实了sibSp会对生存率起到影响后，是否Parch数量也会对生存率起到影响呢？

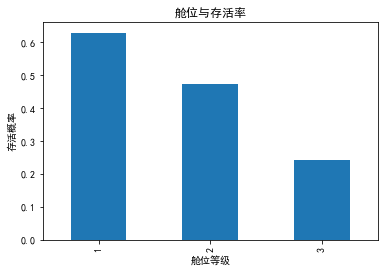
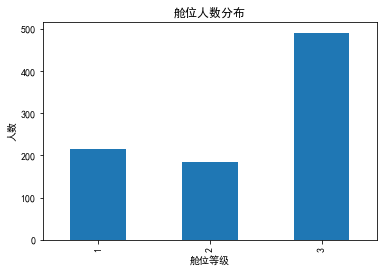




在此我们得到了与sibSp情况相似的结论，但是要注意的是，Parch个数相对于sibSp个数而已，数值为3以上的数据更为稀少。

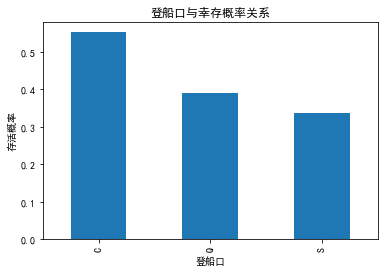
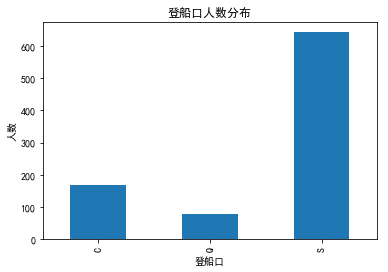
### 5.舱位等级对生存率的影响

舱位等级隐藏的代表了乘客的身份地位，当时泰坦尼克号上在一等舱的人物几乎都是当时的社会精英和各种地方政要。猜测舱位越豪华，生存率将会越高。



首先分析不同舱位等级人数的分布，三等舱的人数要较多一些，一等舱和二等舱的人数要较少一些，不过数据样本的量足够多，以至于能够做出判断。将存活率和舱位等级的数据统计绘图之后，发现得到了和假设相同的结论，即舱位越豪华，存活率越高。

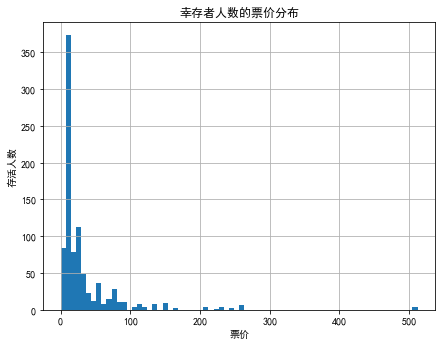
### 6.登船口对生存率的影响



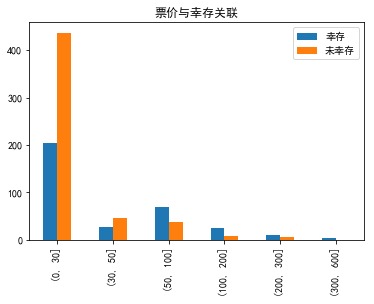
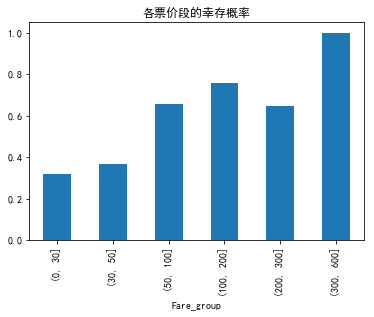
泰坦尼克号途径三个港口，我们发现登船口的不同也会导致生存率差异。

### 7.票价对生存率的影响

票价和舱位等级一样，体现了一个人的社会地位。



由上述分布图得知，即使在泰坦尼克号这样一艘豪华邮轮上，乘客们也并非全是清一色的富裕人士，还是有着大量的中低阶层的民众。



观察上面两幅条状图，不难判断，富裕阶层的人拥有着比普通民众更加高的存活率。这个结果与分析Pclass后所得到的结果是吻合的。