泰坦尼克乘客生还预测数据集:数据分析报告

# 一、数据概览

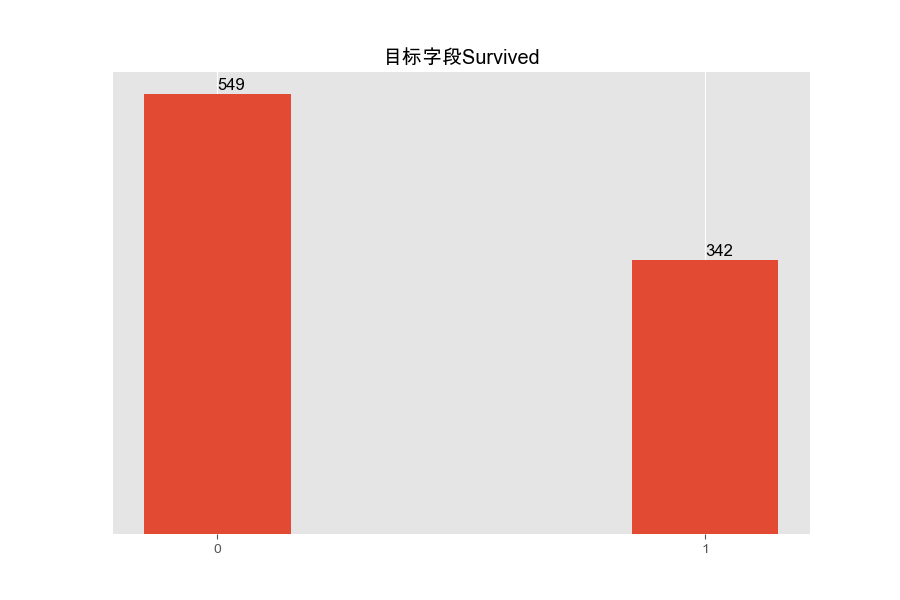
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 891 entries, 0 to 890  
Data columns (total 11 columns):  
Survived 891 non-null int64  
Pclass 891 non-null int64  
Name 891 non-null object  
Sex 891 non-null object  
Age 714 non-null float64  
SibSp 891 non-null int64  
Parch 891 non-null int64  
Ticket 891 non-null object  
Fare 891 non-null float64  
Cabin 204 non-null object  
Embarked 889 non-null object  
dtypes: float64(2), int64(4), object(5)  
memory usage: 76.6+ KB

# 二、缺失值分析

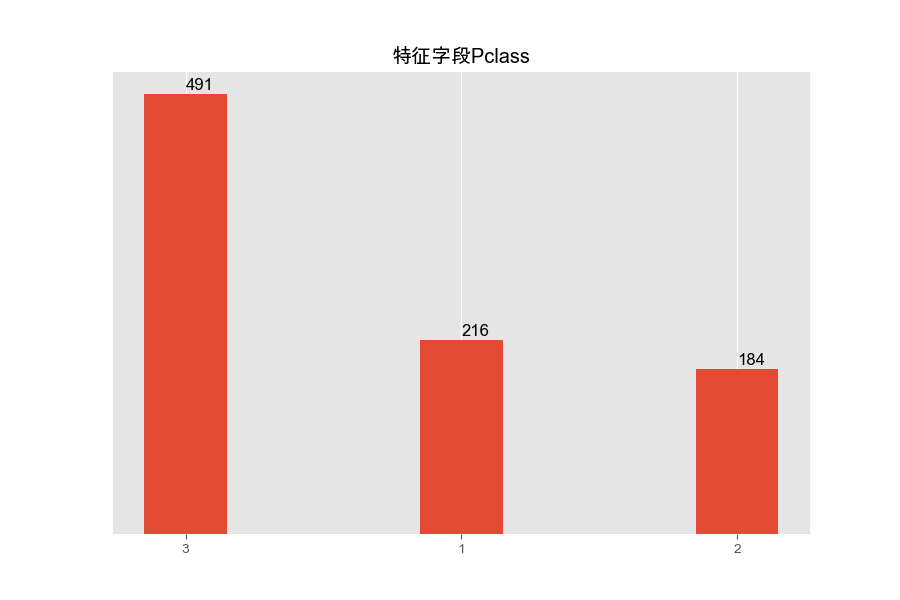
含有缺失值的特征：  
[('Cabin', '缺失率：0.771'), ('Age', '缺失率：0.199'), ('Embarked', '缺失率：0.002')]

# 三、每个字段的值的分布

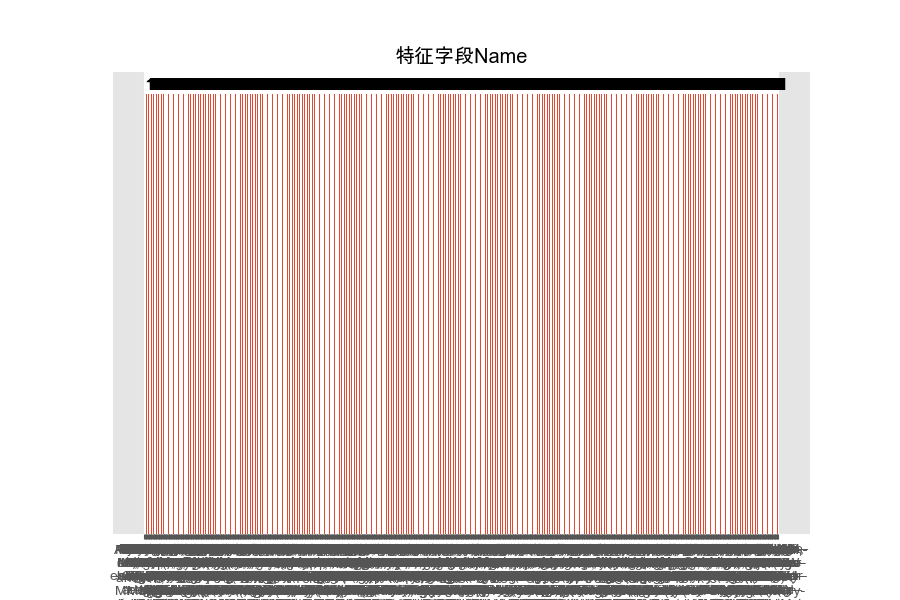
1. 特征字段：Survived



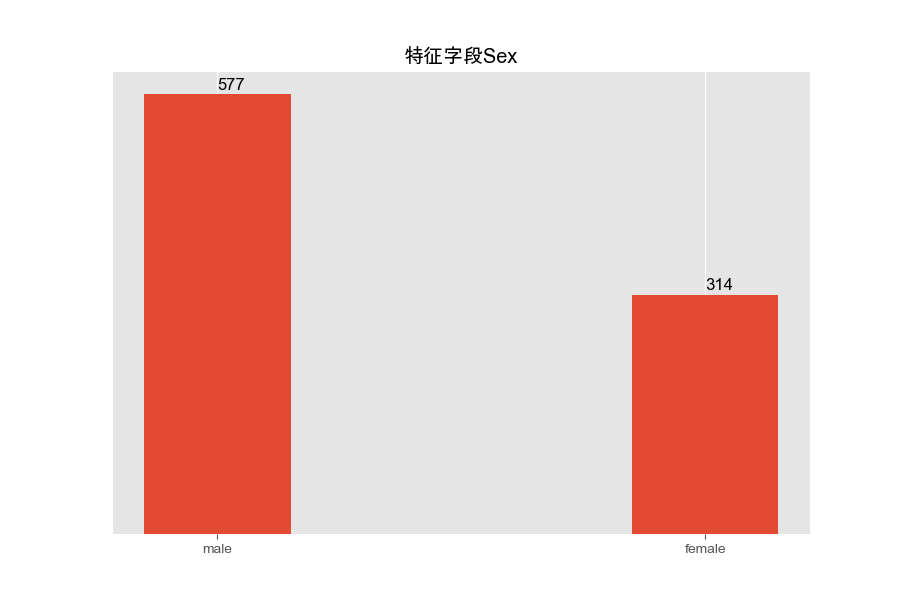
1. 特征字段：Pclass



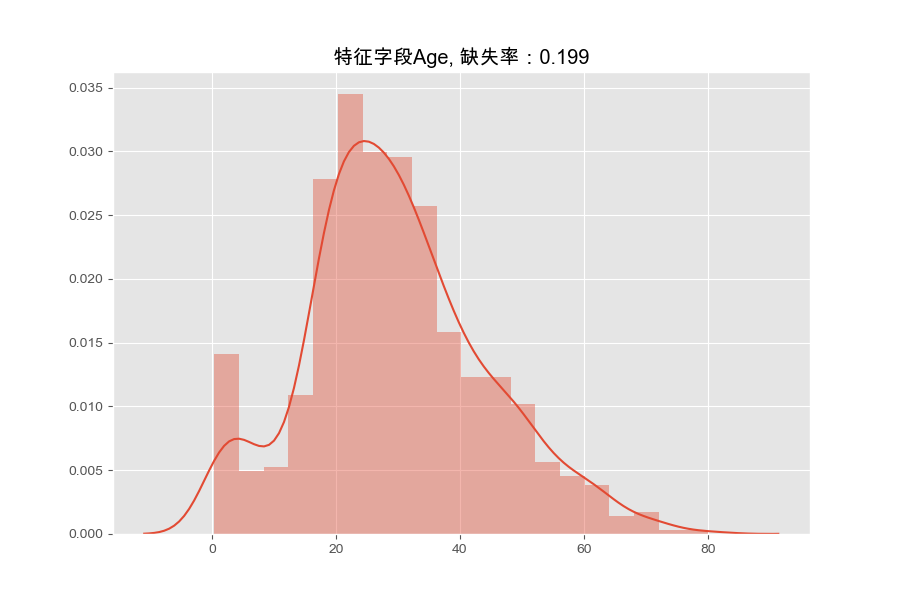
1. 特征字段：Name



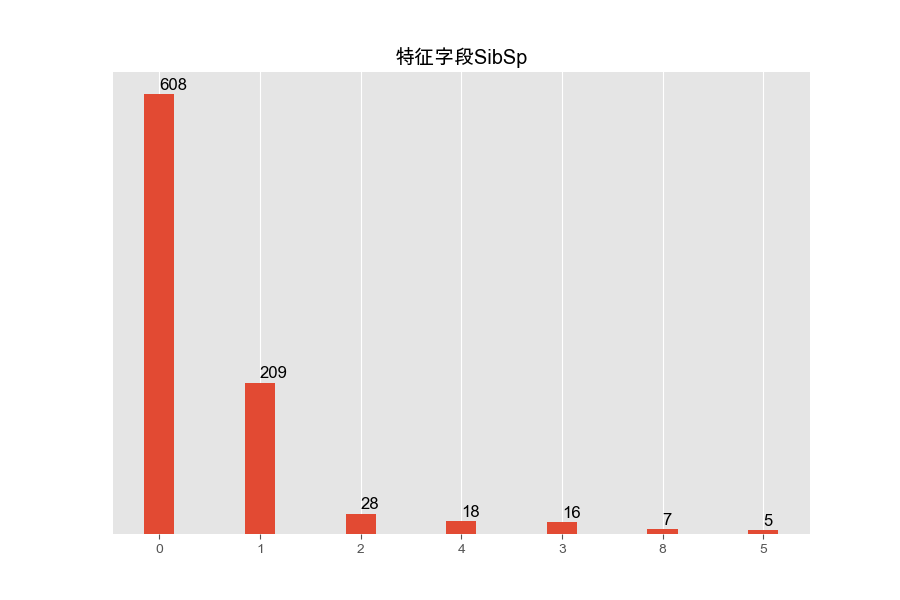
1. 特征字段：Sex



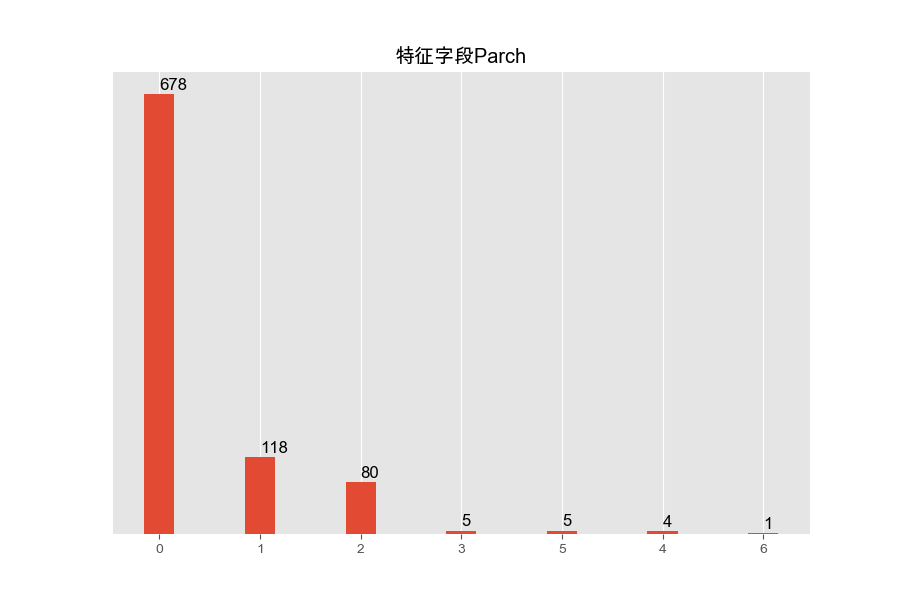
1. 字段：Age



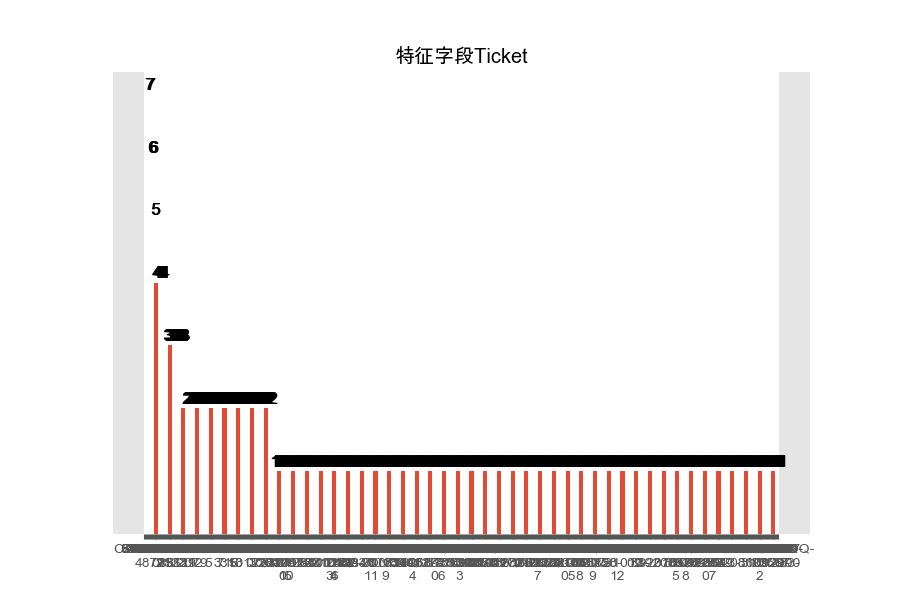
1. 特征字段：SibSp



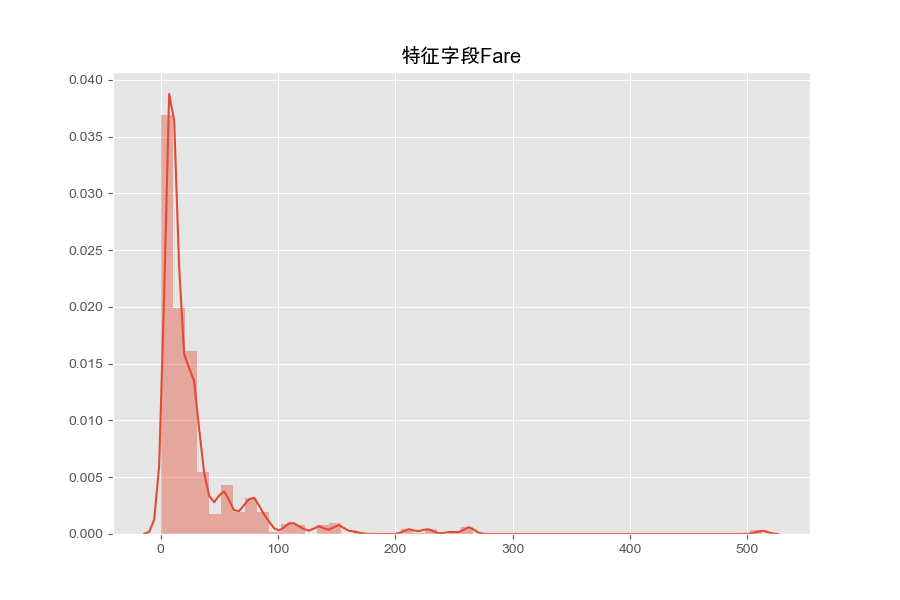
1. 特征字段：Parch



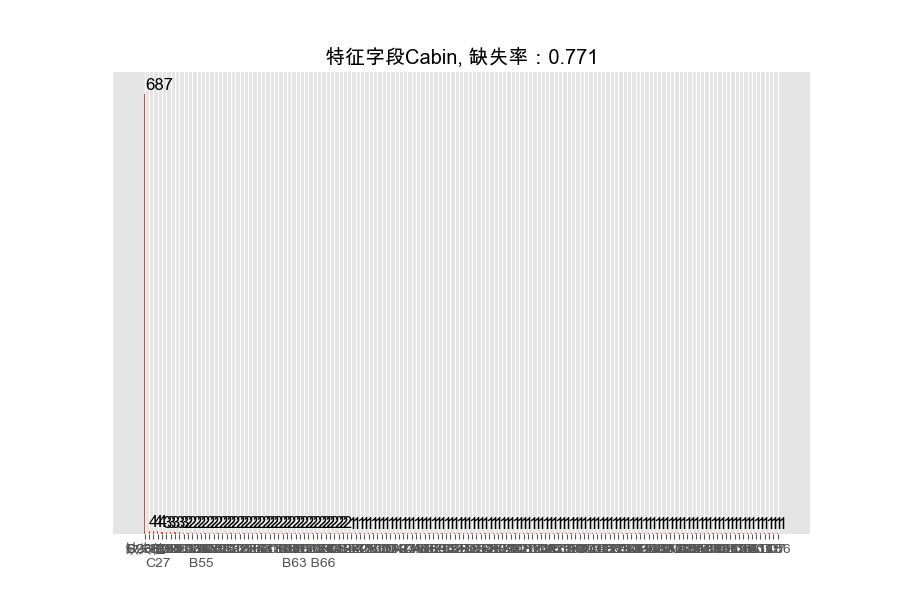
1. 特征字段：Ticket



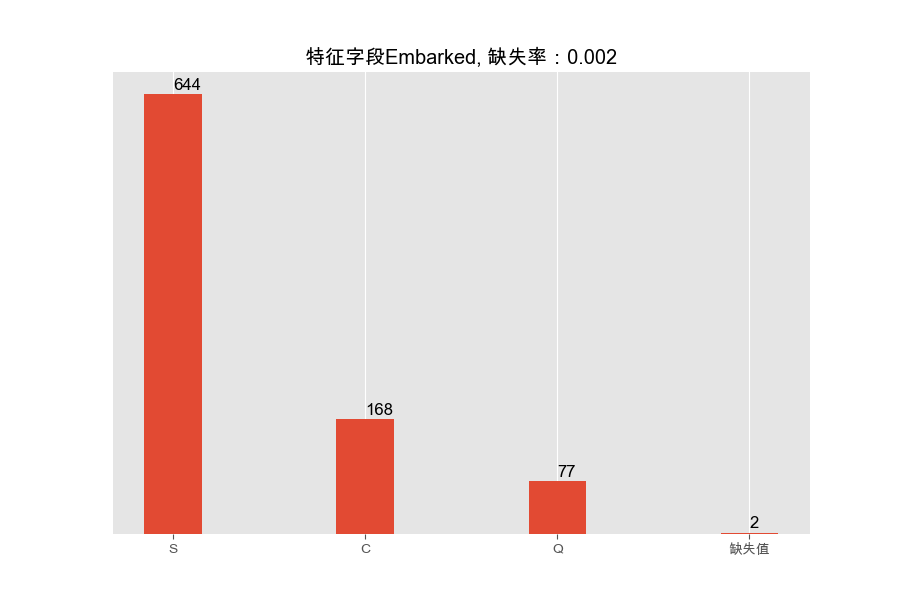
1. 字段：Fare



1. 特征字段：Cabin

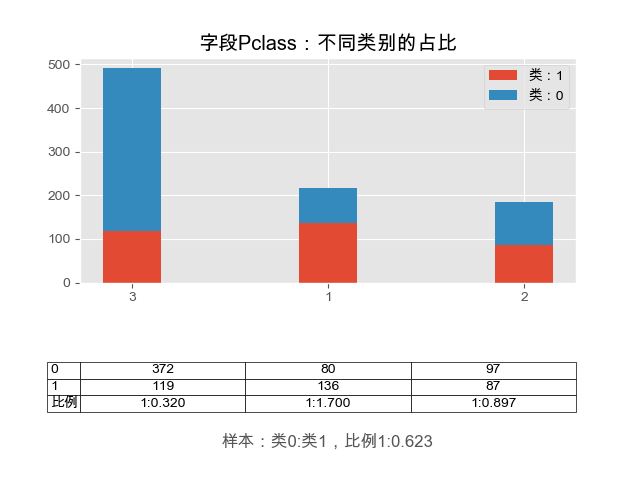


1. 特征字段：Embarked

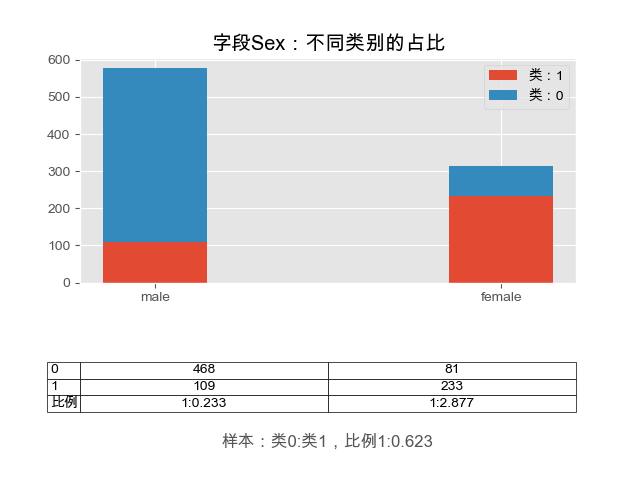


# 四、每个特征字段与目标字段之间的关系

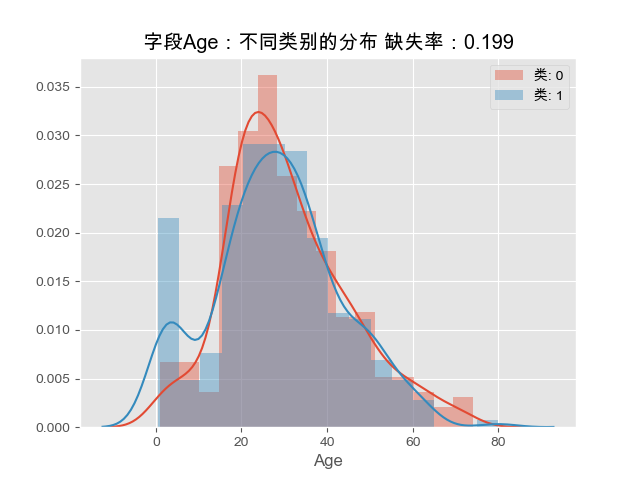
* 关系: Pclass VS Survived



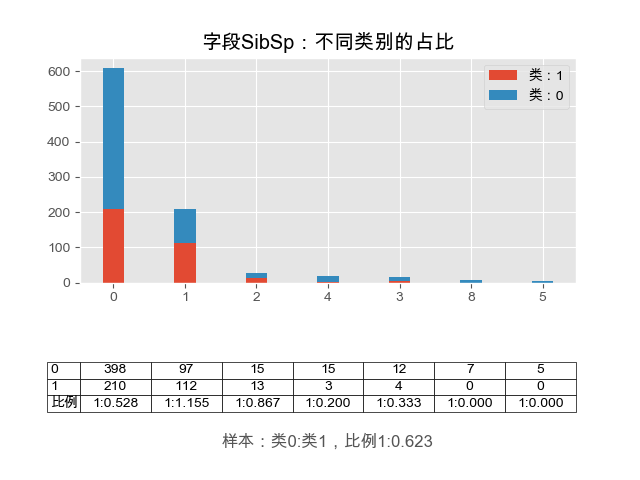
* 关系: Sex VS Survived



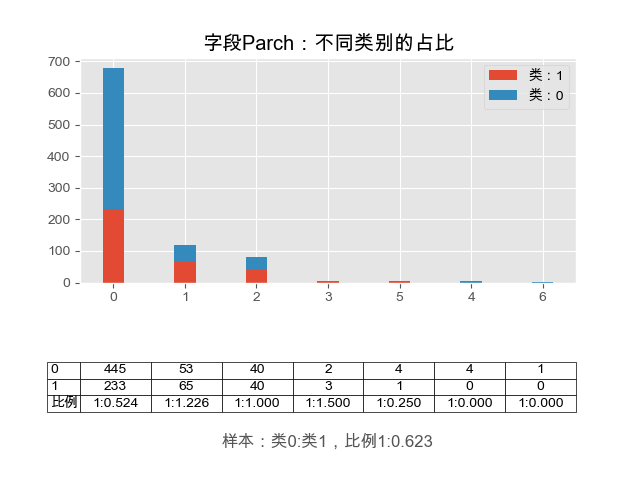
* 关系: Age VS Survived



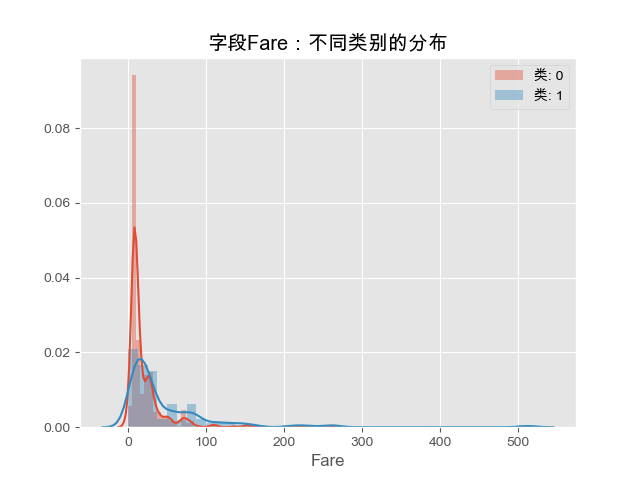
* 关系: SibSp VS Survived



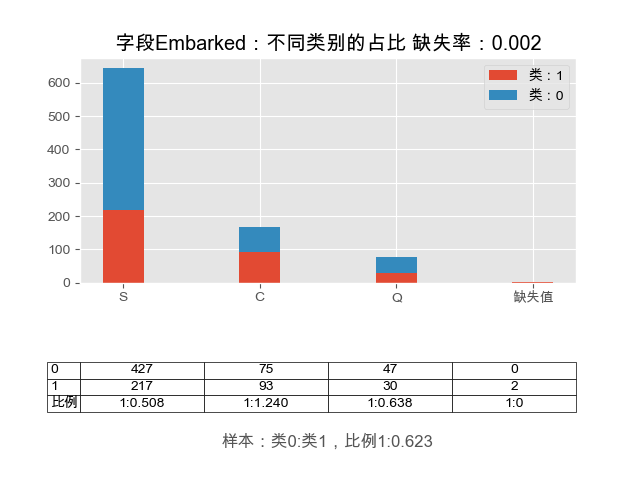
* 关系: Parch VS Survived



* 关系: Fare VS Survived



* 关系: Embarked VS Survived



# 五、其他

类别型特征中，不同值较多的字段：

* 字段Name，不同值的个数：891
* 字段Ticket，不同值的个数：681
* 字段Cabin，不同值的个数：148