四　川　大　学

博\硕 士研究生课程考试试卷

姓　　名：万科　　学　　号：2019225025117

学院（所、中心）：商学院　　专　　业：工业工程

考试课程名称：分类数据分析	
考试方式　□笔试 □口试 ☑撰写报告 任课教师　李宗敏	考试成绩 考试时间：2020.12.21

四川大学研究生院制

泰坦尼克号数据分析

1 引言

泰坦尼克号的沉没是历史上最臭名昭著的海难之一。1912 年 4 月 15 日，在她的处女航中，被广泛认为的“沉没”RMS 泰坦尼克号与冰山相撞后沉没。不幸的是，船上没有足够的救生艇供所有人使用，导致 2224 名乘客和机组人员中的 1502 人死亡。虽然幸存有一些运气，但似乎有些人比其他人更有可能生存。针对其生存与遇难的人的数据，来分析“什么样的人更有可能生存？”主要使用乘客数据（即年龄，性别，社会经济舱等）来进行分析。

2 研究目的

(1).泰坦尼克号乘客的基本信息分布情况

(2).乘客的信息与生还数据是否有关联

3 数据搜集

3.1 数据来源

数据来自 python seaborn 库自带的泰坦尼克生还者的数据集，该数据集并非泰坦尼克号全部乘客数据，据悉泰坦尼克号上共有 2224 名乘客，此为已处理过的样本数据，共 891 条数据。

3.2 数据预览

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

其中 pclass 和 class，survived 和 alive 分别是字符串和数值类型，因此数据分析时可替换。

Table 1 属性列标的含义

survived,alive	是否获救，1,yes:获救，0,no:遇难		
pclass,class	船舱等级，1,First:一级舱，2,Second:二级舱，3,Third:三级舱		
sex,who	性别	embarked	上船地点简写

age	年龄	embark_town	上船地点全称
sibsp	携带兄弟姐妹的数量	adult_male	是否为成年人
parch	携带父母子女的数量	deck	未知
fare	票价	alone	是否单身乘船

3.3 数据清洗

该数据集共有 15 个属性，age、embarked 和 deck 处存在缺失值，需对缺失值进行处理。

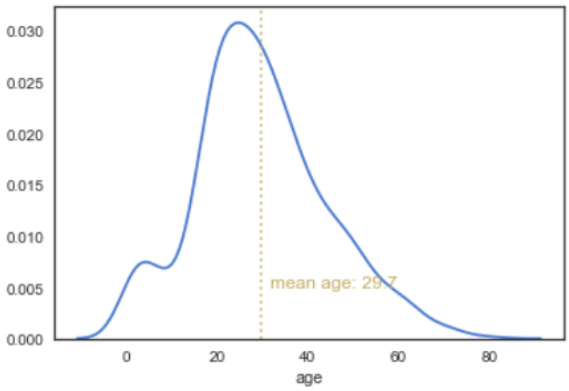
```
titanic.describe()
```

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

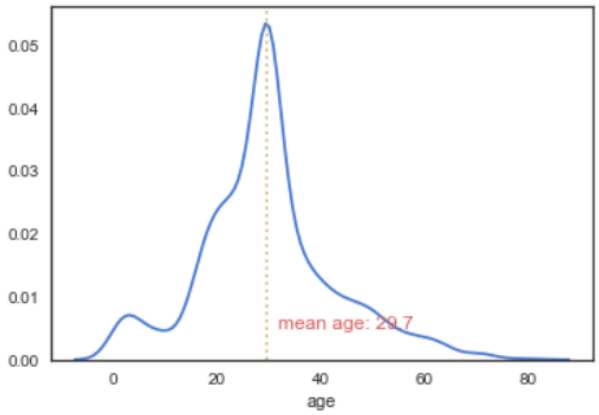
```
titanic.isnull().sum()
```

```
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

(1).对年龄进行缺失值填充



上图为乘客的分布密度图，可以看出年龄呈正态分布，于是用年龄的均值进行缺失值的填充，再进行年龄分布的可视化。



(2). 对 embarked 属性进行缺失值填充：众数填充

```
titanic['age'] = titanic['age'].fillna(titanic['age'].mean())
```

```
sns.distplot(titanic['age'],hist=False,kde=True)
age_mean=titanic['age'].mean()
plt.axvline(age_mean,color='y',linestyle=":",alpha=0.8)
plt.text(age_mean+2,0.005,'mean age: %.1f' % (age_mean), color = 'y')
```

(3). deck、who、adult_malec，embark_town 属性对研究作用较小，因此删除

```
titanic['embarked']=titanic['embarked'].fillna('S')
titanic.isnull().sum()
```

```
survived         0
pclass           0
sex              0
age              0
sibsp            0
parch            0
fare             0
embarked         0
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

```
titanic=titanic.drop(['who','adult_male','deck','embark_town'], axis=1)
titanic.head()
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	no	True

3.4 数据的抽样分布

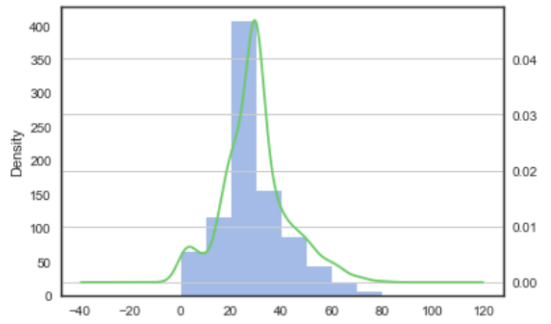
经处理过的数据有 survived/alive,pclass/class,sex,age,sibsp,parch,fare,embarked,alone 共 11 个属性，其中数据的计量尺度如下表所示

属性	计量尺度	属性	计量尺度
survived/alive	定类型	sibsp	定距型
pclass/class	定序型	parch	定距型
sex	定类型	fare	定距型
age	定距型	embarked	定类型
alone	定类型		

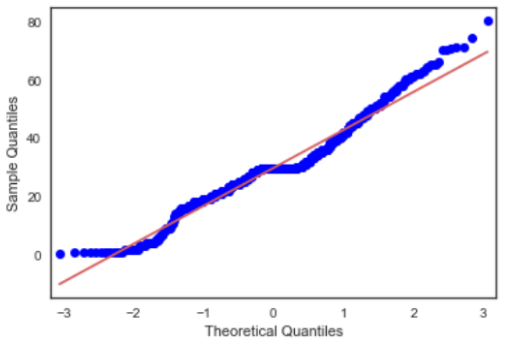
(1). 连续数据的正态性检验

- age

```
titanic['age'].hist(bins=8,alpha = 0.5)
titanic['age'].plot(kind = 'kde', secondary_y=True)
plt.grid()
# 绘制直方图
# 呈现较明显的正太性
```




```
import statsmodels.api as sm
import pylab
sm.qqplot(titanic['age'], line='s')
pylab.show()
```




```
u = titanic['age'].mean()  # 计算均值
std = titanic['age'].std()  # 计算标准差
stats.kstest(titanic['age'], 'norm', (u, std))
# .kstest方法：KS检验，参数分别是：待检验的数据，检验方法（这里设置成norm正态分布），均值与标准差
# 结果返回两个值：statistic → D值，pvalue → P值
# p值小于0.05，为非正态分布
```

```
KstestResult(statistic=0.14845680820991702, pvalue=1.3210991040176237e-17)
```

- fare

```
u = titanic['fare'].mean()  # 计算均值
std = titanic['fare'].std()  # 计算标准差
stats.kstest(titanic['fare'], 'norm', (u, std))
# .kstest方法：KS检验，参数分别是：待检验的数据，检验方法（这里设置成norm正态分布），均值与标准差
# 结果返回两个值：statistic → D值，pvalue → P值
# p值小于0.05，为非正态分布
```

```
KstestResult(statistic=0.2818480409859748, pvalue=4.1796927078903344e-63)
```

由上可知，连续型数据均为非正态数据且不服从泊松分布，在后续采用统计方法进行统计分析时需要注意。

(2). 分类数据二项分布检验

分类数据非二项分布

假设检验汇总

	零假设	检验	显著性	决策者
1	survived = 0 和 1 所定义的类别的发生概率为 0.5 和 0.5。	单样本 Binomial 检验	.000	拒绝零假设。
2	alive = no 和 yes 所定义的类别的发生概率为 0.5 和 0.5。	单样本 Binomial 检验	.000	拒绝零假设。
3	alone = False 和 True 所定义的类别的发生概率为 0.5 和 0.5。	单样本 Binomial 检验	.000	拒绝零假设。
4	sex = male 和 female 所定义的类别的发生概率为 0.5 和 0.5。	单样本 Binomial 检验	.000	拒绝零假设。

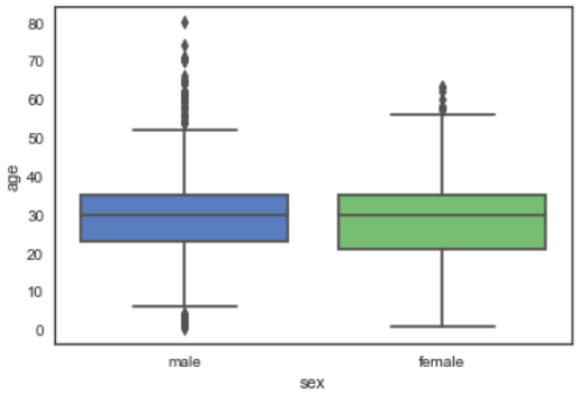
显示渐进显著性。 显著性水平为 .05。

4 数据分析及结果展示

4.1 探索性分析

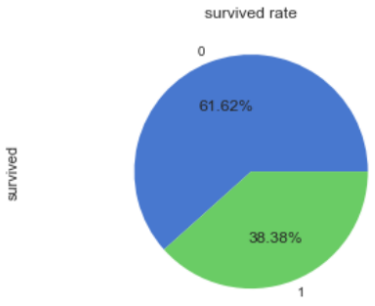
```
sns.boxplot(x='sex', y='age', data=titanic)
```

```
<AxesSubplot:xlabel='sex', ylabel='age'>
```



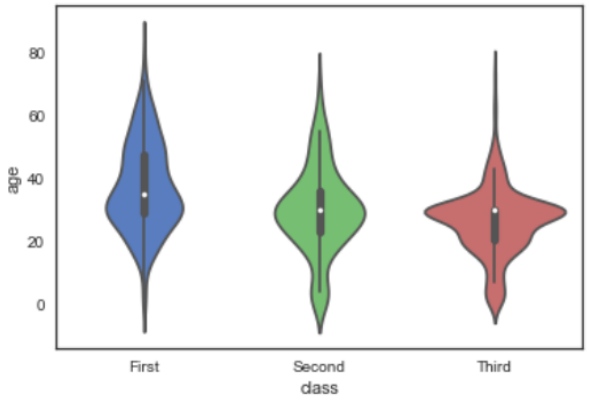

```
plt.axis('equal')
titanic['survived'].value_counts().plot.pie(autopct='%1.2f%%')
plt.title('survived rate')
```

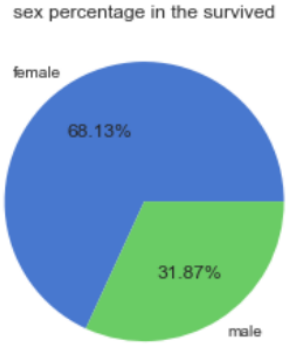
```
Text(0.5, 1.0, 'survived rate')
```




```
sns.violinplot(x="class", y="age", data=titanic)
```

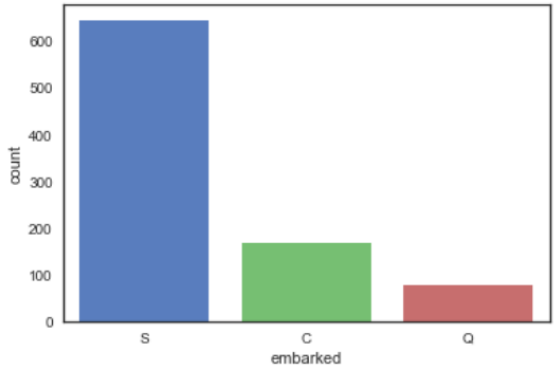
```
<AxesSubplot:xlabel='class', ylabel='age'>
```

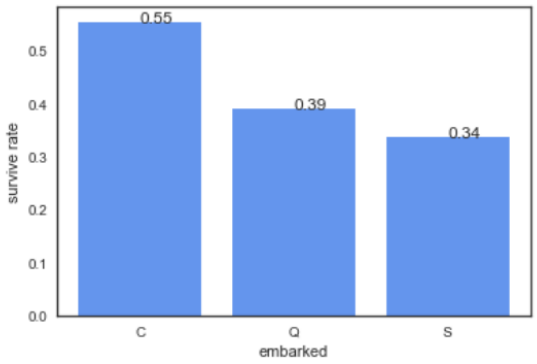




男性的年老者更多，但女性的年龄跨度更加大，891 名乘客中遇难 549 人，获救 342 人，泰坦尼克号生存率大概为 38.38%，死亡率为 61.62%。三种级别的船舱中均是 20-40 岁的人群占比较多。在最终获救的人群中，女性有 233 人，男性有 109 人，女性的占比高达 68.13%,而男性占比仅有 31.67%。

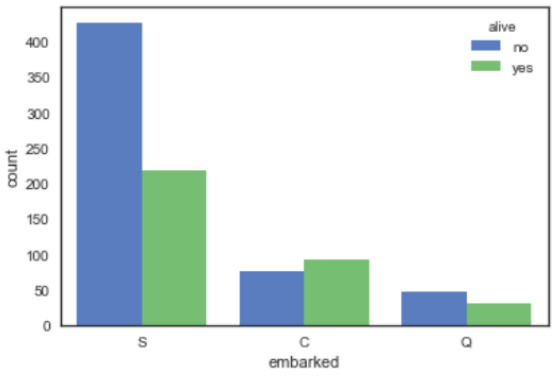
4.1.1 从不同地点上船的人员分布及生存率

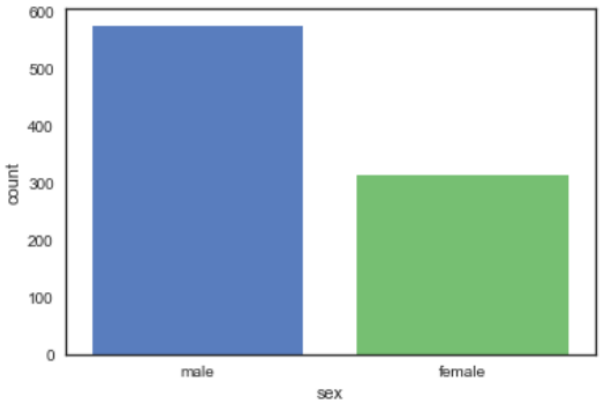





```
sns.countplot(x='embarked', data=titanic,hue="alive")
```

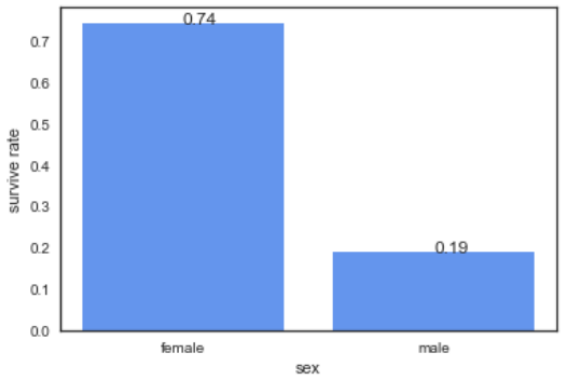
```
<AxesSubplot:xlabel='embarked', ylabel='count'>
```



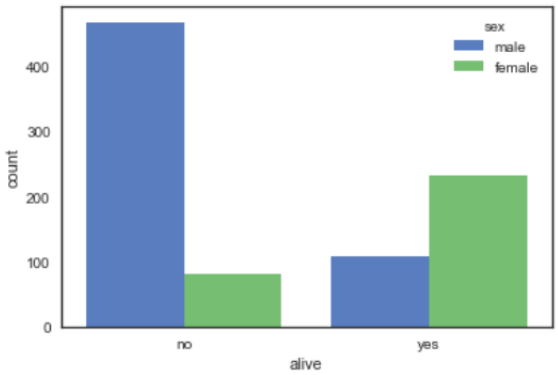


结论：从 S 地上船 646 人，获救 219 人。从 C 地上船 168，获救 93 人。从 Q 地上船 77 人，获救 30 人。从 S 地点上船的乘客最多，但从 C 地点上船的乘客生存率最大

4.1.2 男性与女性的数量分布以及生存率

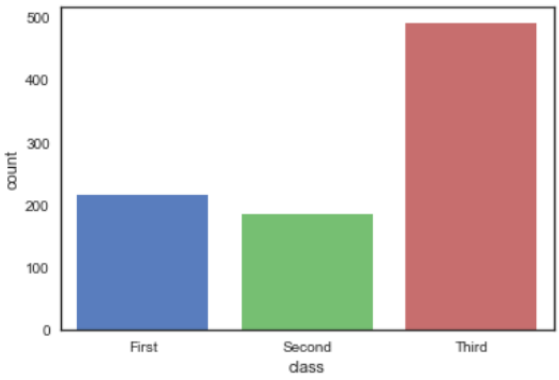


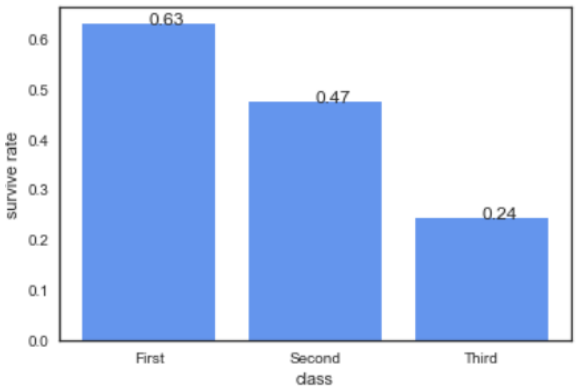

```
sns.countplot(x='alive', hue='sex', data=titanic)
<AxesSubplot:xlabel='alive', ylabel='count'>
```



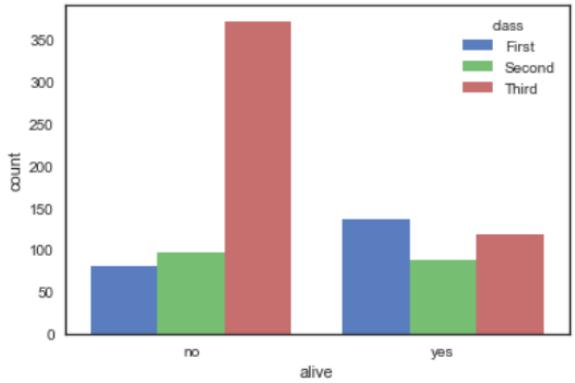
结论：乘客中男性有 577 人，其中获救 109 人。女性 314 人，其中获救 233 人。乘客中男性占比更多，但女性比男性的生存率高且获救的女性更多。

4.1.3 不同等级船舱的人员分布及生存率






```
sns.countplot(x="alive", data=titanic,hue='class')
<AxesSubplot:xlabel='alive', ylabel='count'>
```

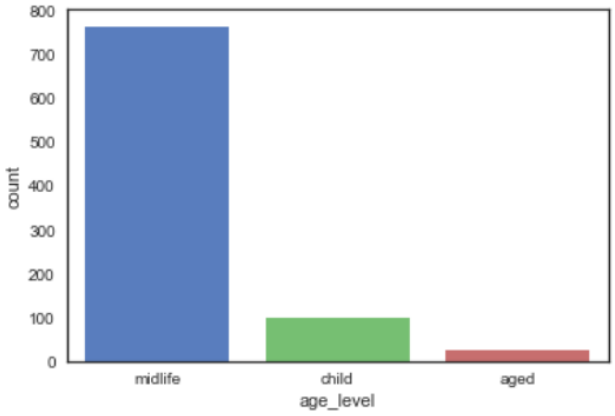


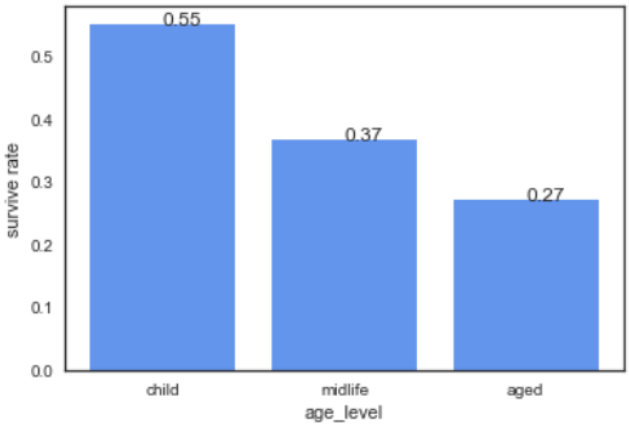
结论：一等舱 216 人，获救 136 人；二等舱 184 人，获救 119 人；三等舱 491 人，获救 87 人。船舱等级越高，生存率也就越高

4.1.4 不同年龄段的人员分布及生存率

年龄为连续型变量，为探索年龄与生存率的关系，对年龄进行分段，小于等于 16 岁的记为 child, 大于等于 60 岁的记为 aged，其余的记为 midlife

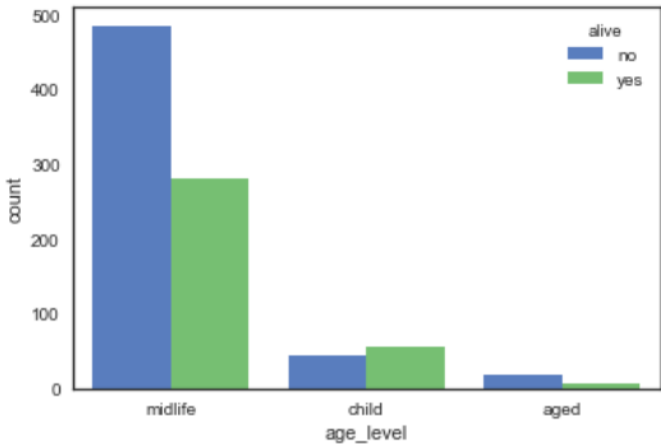
```
def agelevel(age):
    if age <= 16:
        return 'child'
    elif age >= 60:
        return 'aged'
    else:
        return 'midlife'
titanic['age_level'] = titanic['age'].map(agelevel)
```





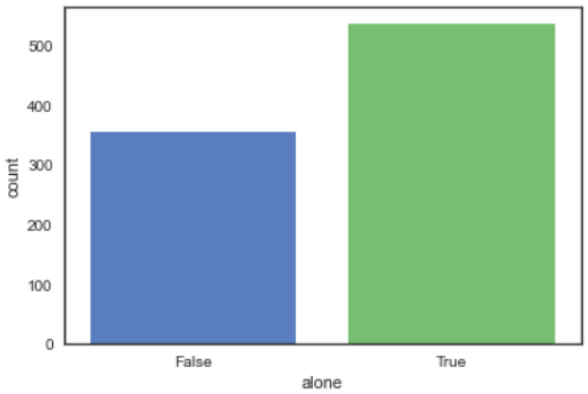

```
sns.countplot(x='age_level',hue='alive', data=titanic)
```

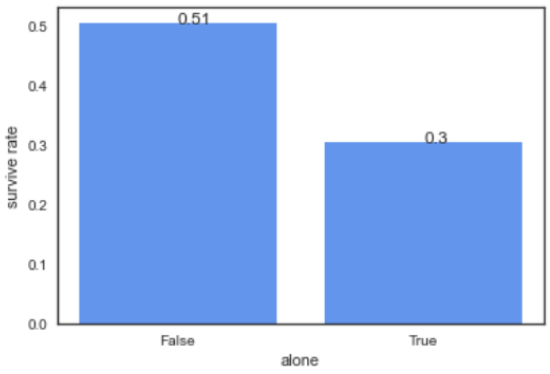
```
<AxesSubplot:xlabel='age_level', ylabel='count'>
```



结论：child 有 100 人，获救 55 人；midlife 有 765 人，获救 280 人；aged 有 26 人，获救 7 人。midlife 即青中年人人数最多，但 child（儿童）的生存率最高，aged（老人）的生存率最低

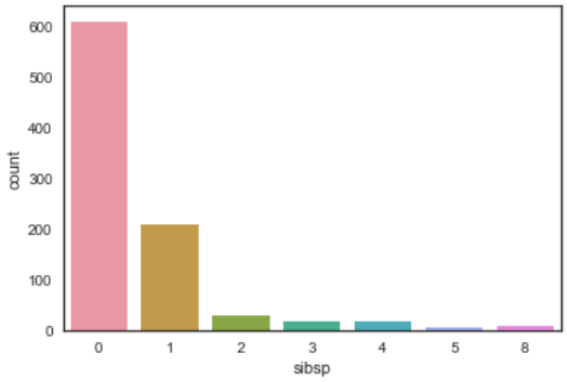
4.1.5 是否单身的人员分布及生存率

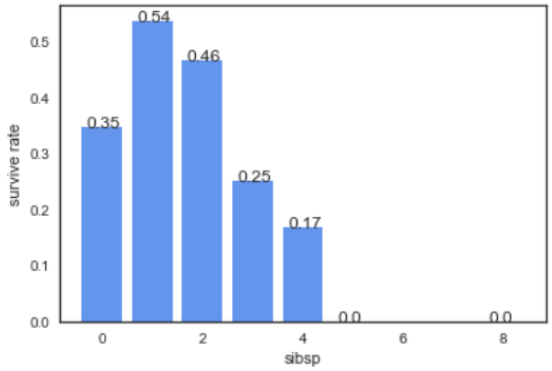




结论：单身乘船 537 人，其中获救 163 人；非单身乘船 354 人，其中获救 179 人。单身人员的生存率略低于携带亲属的人员

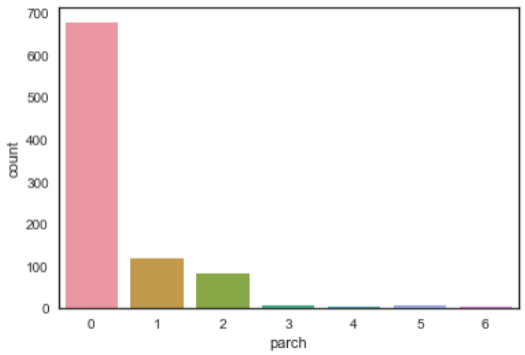
4.1.6 携带兄弟姐妹的人员分布以及生存率

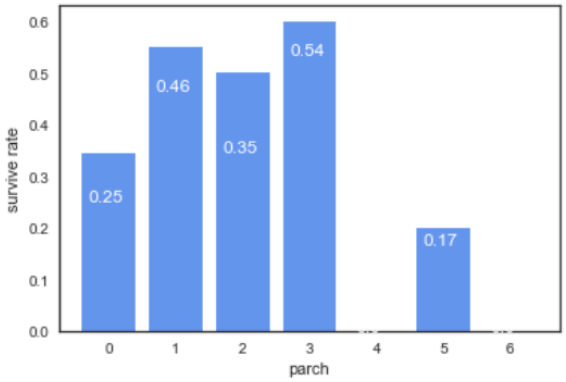




结论：携带兄弟姐妹较少的乘客比不携带兄弟姐妹的乘客生存率要高

4.1.7 携带父母子女的人员分布以及生存率

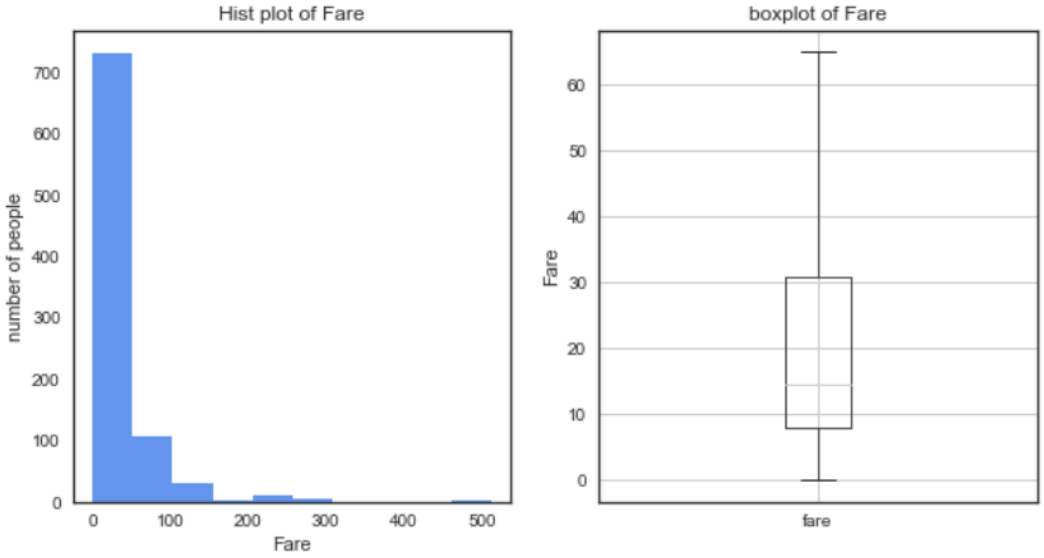




结论：携带父母孩子上船的乘客比没携带父母孩子上船的乘客生存率要大

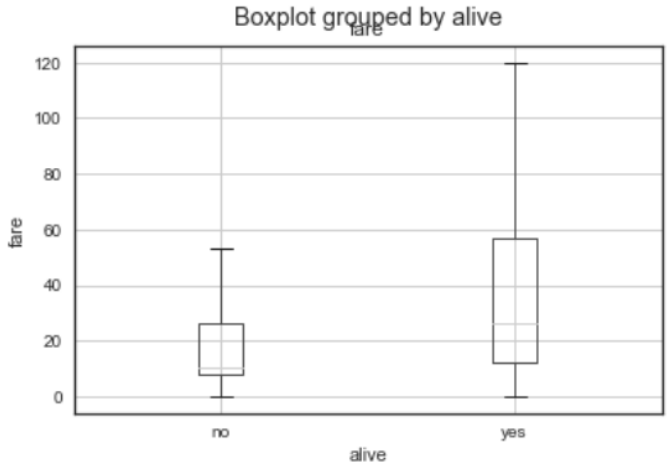
4.1.8 票价分布及其与生存率的关系

```
#创建figure、subplot，用hist做直方图
fig_Fare=plt.figure(figsize=(10,5))
ax_Fare=fig_Fare.add_subplot(1,2,1)
titanic['fare'].hist(bins=10,color='#6495ED',grid=False)
#添加标题，x、y轴标签
ax_Fare.set_title('Hist plot of Fare')
ax_Fare.set_ylabel('number of people')
ax_Fare.set_xlabel('Fare')
#票价箱线图 #作箱线图
plt.subplot(122)
titanic_data.boxplot(column='fare',showfliers=False)
#添加标题、y轴标签
plt.title('boxplot of Fare')
plt.ylabel('Fare')
```




```
titanic.boxplot(column='fare',by='alive',showfliers=False)
#添加y轴标签
plt.ylabel('fare')
```

```
Text(0, 0.5, 'fare')
```

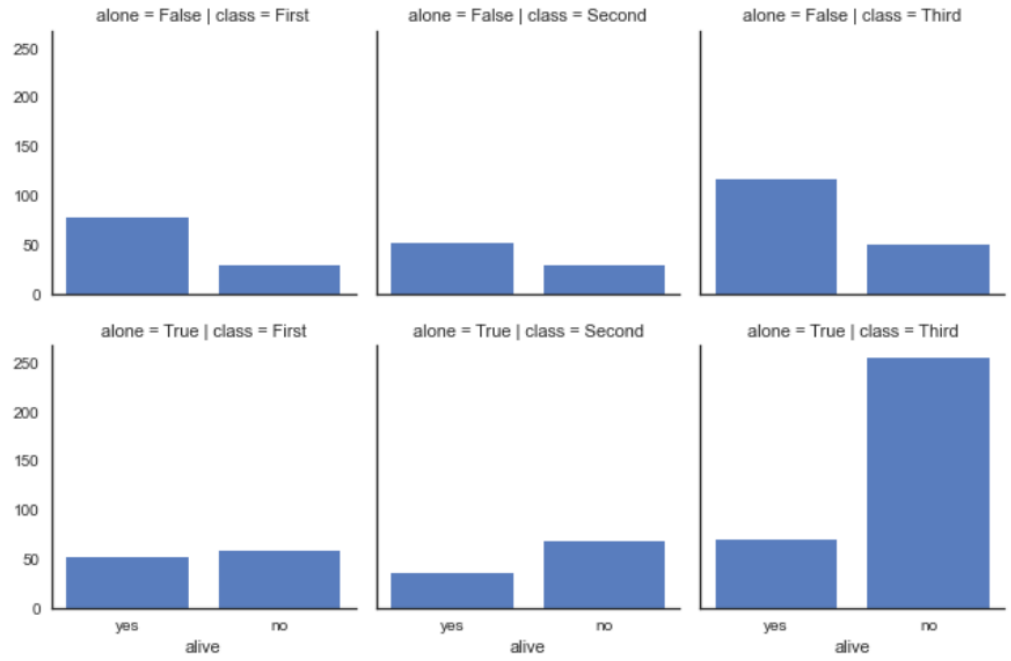



```
titanic['fare'].describe()
```

```
count    891.000000
mean      32.204208
std       49.693429
min        0.000000
25%        7.910400
50%       14.454200
75%       31.000000
max      512.329200
Name: fare, dtype: float64
```

结论：票价与生还有一定相关性，生还者的平均票价要比未生还的高。

4.1.9 结合 class 和 alone 进行分析



结论：船舱等级越高，携带亲属乘船的乘客生还率更高

4.2 深入分析

基于以上分析，提出以下问题，并通过各种检验方法及数据的类型进行深入探讨，本文的置信水平设置为 95%。

(1).获救者与遇难者年龄是否存在显著差异

(2).不同性别获救的比例是否存在显著差异

(3).年龄是否与获救有关。

(4).票价是否与获救有关。

(5).上船的地点是否获救有关。

(6).单身是否与获救有关。

(7).年龄是否与乘坐船舱的等级有关。

(8).性别是否与乘坐船舱的等级有关。

(9).获救比例是否超过 40%

4.2.1 获救者与遇难者年龄是否有显著差异

从以上分析可知，年龄不服从正态分布，总体方差未知，因此不能采取 t 检

验，故采取非参数检验。原假设 H0：未获救者与获救者年龄均值相等；对立假设 H1：未获救者与获救者年龄均值不相等。在 95%的置信水平下进行该检验，

假设检验汇总

	零假设	检验	显著性	决策者
1	在 survived 类别上，age 的分布相同。	独立样本 Mann-Whitney U 检验	.243	保留零假设。

显示渐进显著性。 显著性水平为 .05。

p 值大于 0.05，故接受原假设，获救者与未获救者之间的年龄不存在显著差异， 这与我们的认知存在差异，一般大家均认为灾难时妇女儿童老人优先。

4.2.2 不同性别获救的比例是否有显著差异

构建列联表，进行独立的卡方检验，H0：不同性别获救比例无显著差异；H1：不同性别获救比例有显著差异。

```
import pandas as pd
sc_contingencytable=pd.crosstab(titanic['sex'],titanic['alive'],margins=True)
sc_contingencytable
```

alive	no	yes	All
sex			
female	81	233	314
male	468	109	577
All	549	342	891

```
from  scipy.stats import chi2_contingency
chi2_contingency(sc_contingencytable.iloc[:-1,:-1])
```

```
(260.71702016732104,
 1.1973570627755645e-58,
 1,
 array([[193.47474747, 120.52525253],
        [355.52525253, 221.47474747]]))
```

拒绝原假设，男女获救的比例有显著性差异

p 值小于 0.05，故拒绝原假设，男女获救的比例有显著差异，从以上数据可视化的结果可以看出女性的生存率是远高于男性的，这也间接体现了海难时女士优先。

4.2.3 年龄是否与获救有关

因为 survived 属性为离散变量，年龄为连续型变量，因此采取点二列相关性检验。

```
from scipy import stats
stats.pointbiserialr(titanic['survived'],titanic['age'])
```

```
PointbiserialrResult(correlation=-0.06980851528714312, pvalue=0.03721708372681355)
```

p 值小于 0.05，故拒绝原假设，故年龄与是否获救相关，但相关系数 r=-0.07 极小，所以此处认为年龄与是否获救弱负相关。

4.2.4 票价是否与获救有关

因为 survived 属性为离散型变量，票价为连续型数据，点二列相关不要求数据正态性分布，因此采取点二列相关性检验。

```
#scipy中的stats.pointbiserialr函数可以用于计算该相关系数及其显著性水平
stats.pointbiserialr(titanic['fare'],titanic['survived'])
```

```
PointbiserialrResult(correlation=0.25730652238496243, pvalue=6.120189341917992e-15)
```

p 值小于 0.05，故拒绝原假设，票价与是否获救有正相关关系，相关系数为 r=0.258，票价与船舱级别是成正比的，进一步佐证了船舱级别会影响生存率。

4.2.5 不同上船地点是否与获救有关

构建列联表进行卡方检验。原假设 H0：不同上船地点与获救与否无关；对立假设：不同上船地点与获救与否有关。

```
sc_emb=pd.crosstab(titanic['embarked'],titanic['alive'])
sc_emb
```

alive	no	yes
embarked		
C	75	93
Q	47	30
S	427	219

```
chi2_contingency(sc_emb)
```

```
(25.964452881874784,
 2.3008626481449577e-06,
 2,
 array([[103.51515152,  64.48484848],
        [ 47.44444444,  29.55555556],
        [398.04040404, 247.95959596]]))
```

p 值小于 0.05，故拒绝原假设，不同上船地点与获救与否有关。

4.2.6 单身是否与获救有关

构建列联表进行卡方检验。原假设 H0：单身与获救无关；对立假设：单身与获救有关。

```
a_contingencytable=pd.crosstab(titanic['alone'],titanic['alive'],margins=True)
a_contingencytable
```

alive	no	yes	All
alone			
False	175	179	354
True	374	163	537
All	549	342	891

```
chi2_contingency(a_contingencytable.iloc[:-1,:-1])
```

```
(36.00051446773865,
 1.9726543846517113e-09,
 1,
 array([[218.12121212, 135.87878788],
        [330.87878788, 206.12121212]]))
```

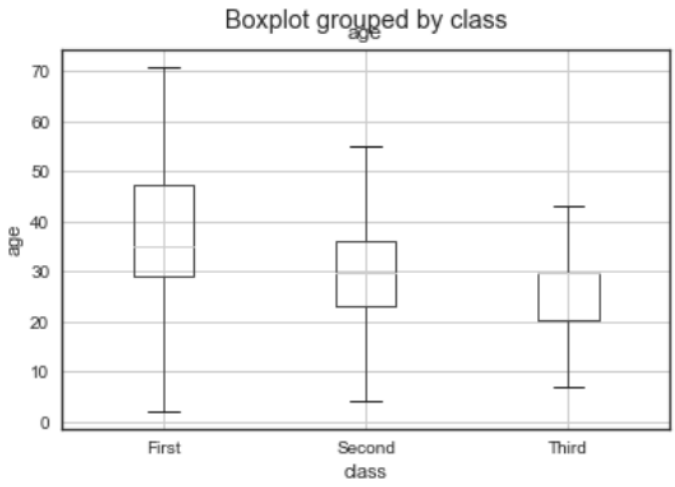
p 值小于 0.05，故拒绝原假设，单身与获救有关。

4.2.7 年龄是否与乘坐船舱的等级有关

由于年龄为连续型变量而船舱等级为三值变量，因此直接绘制年龄关于船舱等级的箱线图进行观察。由下图可知，年龄与乘坐船舱等级之间存在相关关系，乘坐一等舱的乘客大多年龄在 35 岁以上，猜测应该是事业有成的商务人士，而乘坐二等舱的年龄大多在 35 岁以下，属于小有收入的那部分群体，而乘坐三等舱的乘客大多是 20-30 岁的年轻人。

```
titanic.boxplot(column='age',by='class',showfliers=False)
#添加y轴标签
plt.ylabel('age')
```

```
Text(0, 0.5, 'age')
```



4.2.8 性别是否与乘坐船舱的等级有关

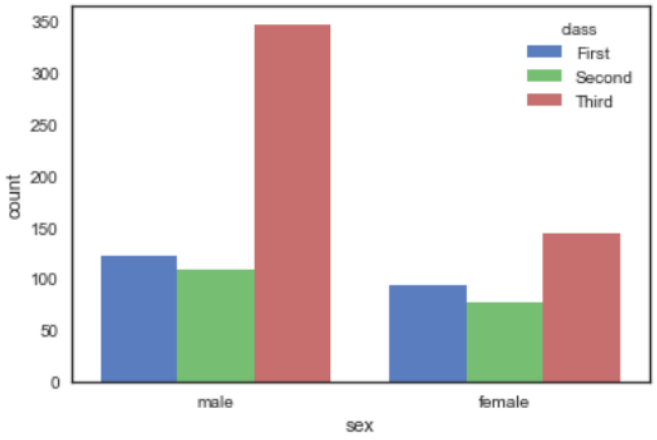
构建列联表进行卡方检验。原假设 H0：性别与乘坐船舱的等级无关；对立假设：性别与乘坐船舱的等级有关。

```
sex_contingencytable=pd.crosstab(titanic['sex'],titanic['class'],margins=True)
sex_contingencytable
```

class	First	Second	Third	All
sex				
female	94	76	144	314
male	122	108	347	577
All	216	184	491	891

```
from  scipy.stats import chi2_contingency
chi2_contingency(sex_contingencytable.iloc[:-1,:-1])
```

```
(16.971499095517114,
 0.00020638864348233114,
 2,
 array([[ 76.12121212,  64.84399551, 173.03479237],
        [139.87878788, 119.15600449, 317.96520763]]))
```



p 值小于 0.05，故拒绝原假设，性别与乘坐船舱的等级有关。由柱状图可知，三等舱的男女比例悬殊极大，其余两个等级的船舱，男女比例无较大差异，这可能与男女的基数以及经济状况有关。

4.2.9 获救的比例是否超过 40%

此时 $p=0.4, n=891, np>5, n(1-p)>5$，可采用正态分布进行检验，故采取总体比例的单侧左尾的假设检验。

```
percentage=342/891
aim=0.4
def percentage_test(d,D,n):
    z=(d-D)/np.sqrt(D*(1-D)/n)
    if np.abs(z)>1.65:
        print('拒绝原假设，获救的比例不超过40%')
    else:
        print('接受原假设，获救的比例超过40%')
percentage_test(percentage,aim,891)
```

```
接受原假设，获救的比例超过40%
```

接受原假设，乘客获救的比例超过 40%。

4.3 二分数据的广义线性模型

二分响应变量 survived 记作 Y，可能的结果为 0（死亡）和 1（获救），获救的概率 $P(Y=1)=\pi$，死亡的概率 $P(Y=0)=1-\pi$，均值 $E(Y)=\pi$，获救活死亡的概率模型为 $\pi(x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$，logistic 回归函数 $\text{logit}[\pi(x)]=\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x$

4.3.1 数据处理

由上可知有几个指标含义上是相同的，因此进行指标的删除。对类目型的特征因子化，且对连续性变量做标准正态化处理。

	sex	age	sibsp	parch	fare	alone	embarked_C	embarked_Q	embarked_S	class_First	class_Second	class_Third
0	0	-0.592481	1	0	-0.502445	1	0	0	1	0	0	1
1	1	0.638789	1	0	0.786845	1	1	0	0	1	0	0
2	1	-0.284663	0	0	-0.488854	1	0	0	1	0	0	1
3	1	0.407926	1	0	0.420730	1	0	0	1	1	0	0
4	0	0.407926	0	0	-0.486337	1	0	0	1	0	0	

4.3.2 模型构建

利用 sklearn 库中的 logisticregression 函数进行模型构建

```
X_train, X_test, y_train, y_test = train_test_split(predictors,prediction,test_size=0.3,random_state=1)
```

```
clf=LogisticRegression(C=1.0, penalty='l1', tol=1e-6)
```

```
clf.fit(X_train, y_train)

clf
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l1', random_state=None, solver='liblinear', tol=1e-06,
          verbose=0, warm_start=False)
```

```
predictions = clf.predict(X_test)
predictions
```

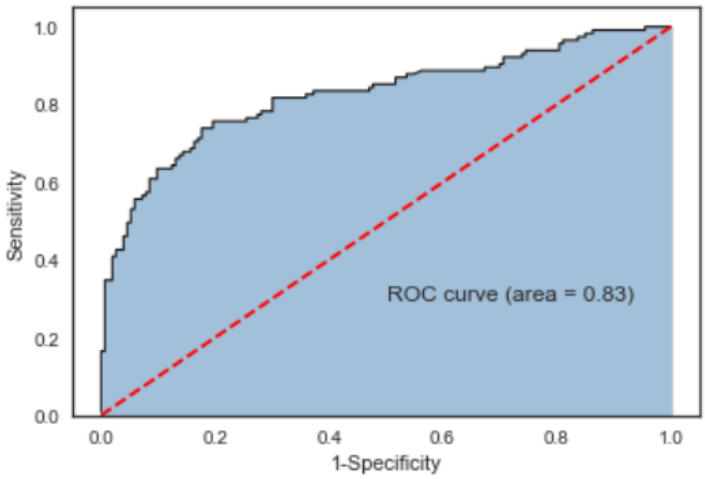
```
array([1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,
       1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0,
       1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
       0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,
       0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1,
       0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0,
       0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1,
       0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,
       0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1,
       0, 0, 1, 0], dtype=int64)
```

4.3.3 模型评估

评估模型的正确率方法就是用预测正确的数目除以总的预测数目，也就是将测试数据的特征 X_test 输入到模型，得到的正确结果数和已知的正确结果的比较。

```
Accuracy = metrics.scorer.accuracy_score(y_test, y_pred)
Sensitivity = metrics.scorer.recall_score(y_test, y_pred)
Specificity = metrics.scorer.recall_score(y_test, y_pred, pos_label=0)
print('模型准确率为%.2f%%:' %(Accuracy*100))
print('正例覆盖率为%.2f%%' %(Sensitivity*100))
print('负例覆盖率为%.2f%%' %(Specificity*100))
```

```
模型准确率为77.99%:
正例覆盖率为66.96%
负例覆盖率为86.27%
```

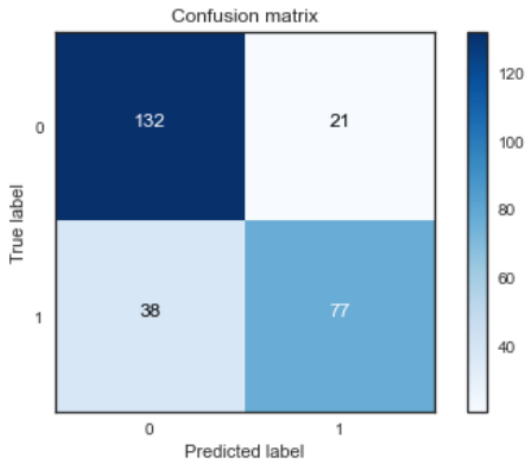



```
clf.intercept_

array([-0.17])
```

模型准确率为 77.99%，正例覆盖率为 66.96%，负例覆盖率为 86.27%。表明模型整体效果还不错，在负例即获救上的识别能力较强。AUC 的面积为 0.83，接近于 1，表明分类预测较好。

召回率： 0.6695652173913044



	columns	coef
0	sex	[2.7471583821369654]
1	age	[-0.5129729163287149]
2	sibsp	[-0.3206041310913266]
3	parch	[0.03985698355525543]
4	fare	[0.00877203429023539]
5	alone	[-0.2919003229837317]
6	embarked_C	[0.0]
7	embarked_Q	[0.0]
8	embarked_S	[-0.48864193942793877]
9	class_First	[0.732284713825075]
10	class_Second	[0.0]
11	class_Third	[-1.4876512505090387]

对于逻辑回归模型得到的 ML 拟合为

$$\text{logit}\left[\pi(x)\right] = -0.17 + 2.75sex - 0.51age - 0.32sibsp + 0.03parch + 0.008fare - 0.29alone$$
$$-0.49embarked\_S + 0.732class\_First - 1.487class\_Third$$

(1). sex 属性，如果是 female 会极大提高最后获救的概率，而 male 会很大程度拉低这个概率。

(2). age 属性为负相关，意味着在我们的模型里，年龄越小，越有获救的优先权

(3). sibsp 属性为负相关，携带兄弟姐妹越多，获救概率越低

(4). parch 属性和船票 fare 有较弱的正相关

(5). alone 属性为负相关，单身的人获救的概率更小

(6). class 属性，1 等舱乘客最后获救的概率会上升，而船舱等级为三等舱会极大的降低获救的概率。

(7). embarked_S 属性为负相关，登船港口 S 会很大程度拉低获救的概率，另外俩港口对获救概率无影响。

5 结论

本次分析主要探寻泰坦尼克号上的生还率和各因素（客舱等级、年龄、性别、上船港口等）的关系，得出了以下结论：

(1). 男女获救的比例存在显著差异，女士和儿童的生存率较高，体现了海难时女士和儿童优先，但 60 岁以上老人的生存率较低，这可能与老人的身体状况以及应急状态有关。

(2). 票价与船舱等级均与生存与否有关，存活下来的乘客票价均值远高于遇难者的票价均值，高票价则对应的是高船舱等级，而一等舱的存活率最高，这显示海难时，一等舱乘客（富人）有优先避难权，船舱等级越高，生存率也越高。

(3). 单身乘船与生存与否存在相关关系，一般而言，相较于携带家属，单身应该更好脱险，但数据分析结果显示，单身的生存率更低一些，这可能有些乘客携带的家属为儿童或女性，在优先他们逃难的同时，这些乘客也顺便上了逃生船。

(4). 携带较少家属的乘客生存率普遍高于单身乘客，而携带较多家属的乘客一般极难脱险。

(5). 获救者与遇难者的年龄不存在显著差异，遇难者年龄多集中于 20-40 岁。

(6). 上船地点与生存与否存在相关关系，在 C 处上船的生存率最高，在 S 处上船的生存率最低，但逻辑回归结果显示，S 处上船的乘客获救率会降低，从从其余地点上船不影响获救率。

(7). 年龄将会影响乘客选择船舱的等级。事业有成的 35 岁左右中高阶层人士将会选择一等船舱，而 20-30 岁正在奋斗的年轻人则会选择三等船舱。基于大的基数以及经济原因，更多的男性乘客将会选择三等船舱，而一二等船舱内男女占比无较大差异。

6 参考文献

[1] https://blog.csdn.net/hpdlzu80100/article/details/78681996

[2] https://blog.csdn.net/han_xiaoyang/article/details/49797143

[3] https://www.cnblogs.com/pythonfl/p/12286742.html

[4] https://wenku.baidu.com/view/c725380b8f9951e79b89680203d8ce2f00666582.html

[5] https://blog.csdn.net/qq_35125180/article/details/107370486

[6] 彭博, 刘丽敏, 张浩苒, 牛文迪, 薛会海, 向修栋. 基于 Logistic 逻辑回归分析的大数据胃癌预测研究[J]. 数码世界, 2020(03):78.

[7] 薄景山, 黄静宜, 张建毅, 王福昌. 基于逻辑回归分析的强震地表破裂预测方法[J]. 地震工程与工程振动, 2019, 39(04):1-7.

[8] H. Bayo Lawal. Categorical Data Analysis With Sas and Spss Applications. 2003,