关于泰坦尼克号数据分析报告

一、生还率分析

自变量：

1、pclass

选取pclass作为自变量的原因是，pclass分为三个等级，依据常识猜测，是否因为选择高等级船票的人较之选择低等船票的人生还几率更大。

1. sex

选取sex作为自变量的原因是，由于男女双方性别而导致身体质素的差异，是否会影响生存几率

1. age

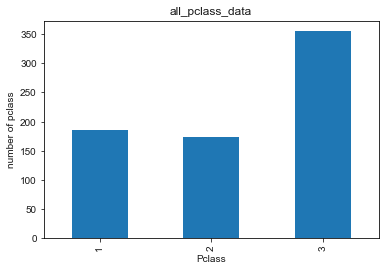
判断是否由于年龄而影响生存几率

因变量：是否生还

二、各项数据分析：

1. **整体数据分析：**

1.1、整体Pclass分析：



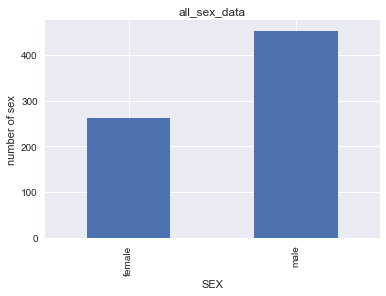
1等仓 186

2等仓 173

3等仓 355

结论：3等仓人数最多，1等仓和2等仓人数基本持平。

1.2、整体性别分析

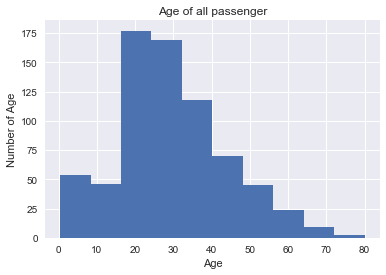


female 261

male 453

结论：船上人员性别男性多于女性

1.3、年龄整体性分析



count 714.000000

mean 29.699118

std 14.526497

min 0.420000

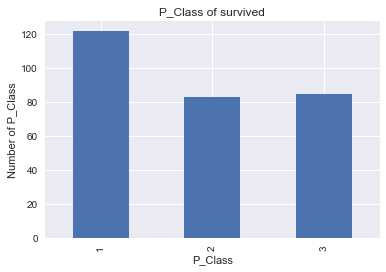
max 80.000000

结论：乘客的年龄分布主要在20-30岁左右，最小年龄0.42岁，最大年龄80岁，由于该数据方差较大，存在数据不准确的情况。

各项分支数据分析：

**2、关于生还者与遇难者pclass分析：**

2.1、生还者数据分析



1等仓 122

2等仓 83

3等仓 85

结论：1等仓生还人数最多，2等仓、3等仓基本相同

2.2、遇难者数据分析



1等仓 64

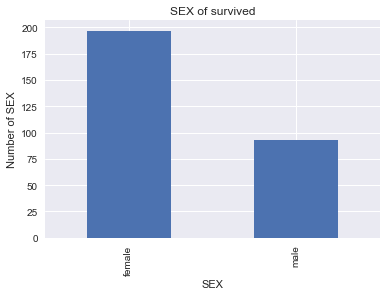
2等仓 90

3等仓 270

结论：1等仓遇难人数最少，3等仓遇难人数最多

**3、生还者遇难者性别分析：**

3.1、生还者性别统计：

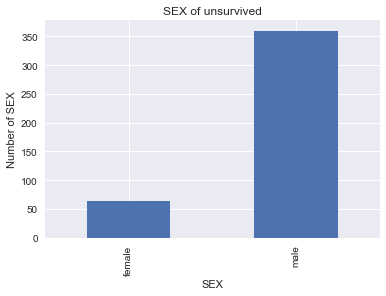


female 197

male 93

结论：女性生还者人数远远大于男性生还者人数

3.2、遇难者性别分析



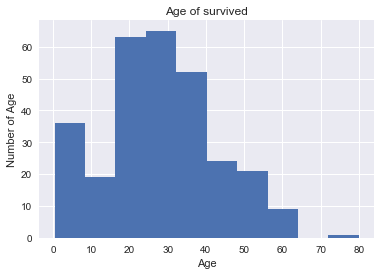
female 64

male 360

结论：女性遇难者人数远远小于男性遇难者

**4、关于生还者与遇难者age分析：**

4.1生还者age分析：



count 290.000000

mean 28.343690

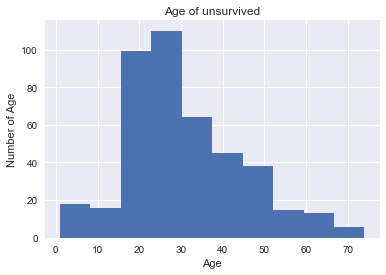
std 14.950952

min 0.420000

max 80.000000

结论：生还者的平均年龄在28.34岁，最小年龄在0.42岁，最大年龄在80岁，由于方差较大，所以此数据可能存在不精确情况。

4.2、遇难者age分析：



count 424.000000

mean 30.626179

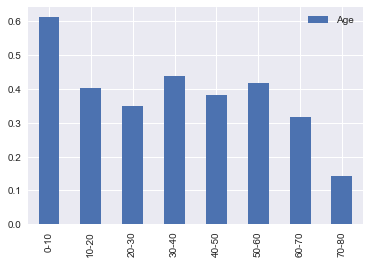
std 14.172110

min 1.000000

max 74.000000

结论：遇难者平局年龄30.62岁最小年龄在1岁，最大年龄74岁由于方差较大，所以此数据可能存在数据不精确情况。

4.3、各组age生还率分析



0-10 0.612903

10-20 0.401961

20-30 0.350000

30-40 0.437126

40-50 0.382022

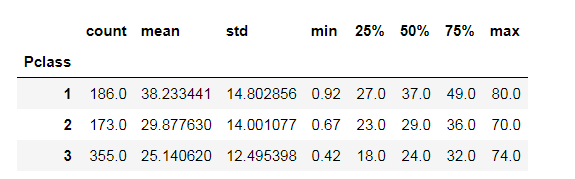
50-60 0.416667

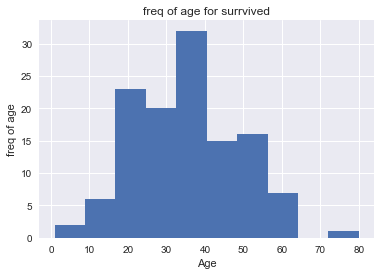
60-70 0.315789

70-80 0.142857

结论:0-10岁生还率最高，但是由于整体数据0-10岁人数较少，该数据可能不精确。

4.4、生还者年龄层与仓位关系分析：





注：此图为1等仓年龄分布

结论：1等仓的平均年龄在38.23岁左右，2等仓的平均年龄在29.87岁左右，3等仓平均年龄在25岁左右。由于30岁到40岁这个年龄在1等仓的人数较多，所以造成30-40岁生还率较高

**三、总体结论**

结论：

1. 仓位对于生存率的影响：在1等仓人数少于其他级别仓位人数时，仍有较高的生还率和较低的遇难率，由此断定仓位对生存率存在影响。即1等仓的乘客生存率比较高。
2. 性别对于生存率的影响 ：女性乘客人数少于男性乘客人数，但是其生存人数比男性乘客高，遇难人数比男性乘客低，由此断定女性乘客生存率较高。
3. 根据图表得出结论0-10岁生还率最高，年龄层与所在仓位有一定关系，38岁左右在1等仓人数较多，所以去除异常数据之后，38岁左右生存率最高。由于源数据的年龄存在估算情况，该项数据方差较大，所以由该项数据得出结论存在误差。
4. 由于生存率是由多个条件共同作用，无法排除其余条件而单独分析某一项因素，同时由于一些数据缺即不精确（比如年龄、船票是否代表本人所在仓位等）失会导致分析结果不够准确，该项目结论只是暂时性的。