

Lab 1: Python数据分析实践 实验报告

孙俊晖(231880101)

1 引言

1.1 背景

1912年4月14日23时40分左右，当时世界上体积最庞大、内部设施最豪华的客运轮船—泰坦尼克号与一座冰山相撞后沉没，此次沉没事故为和平时期死伤人数最为惨重的一次海难，其残骸直至1985年才被再度发现，受到联合国教育、科学及文化组织的保护。Titanic数据集是一个关于此次沉没事故遇难者生存的数据集，包含乘客的性别、年龄、所在船舱、登船港口、票号、是否生还等信息。

1.2 目标

利用Python语言以及Python库对Titanic数据集进行读取和分析，如计算乘客的性别和年龄分布，分析不同因素对生还概率的影响等，并对此进行绘图可视化以更好的展示数据分析结果。

2 方法

2.1 技术栈

开发环境：Visual Studio Code 1.93.0

编程语言：Python 3.7.9 64-bit

2.2 数据读取、数据分析以及数据分析结果展示

2.2.1 导入所需的Python库

利用pandas库读取数据，利用matplotlib库进行绘图可视化

```
1 | import pandas as pd  
2 | import matplotlib.pyplot as plt
```

2.2.2 利用pandas库从titanic.csv文件中读取数据

利用pandas库读取数据后，为了便于在绘图时展示数据分析结果，将读取的数据的英文关键字替换为中文。

```

1 # 从CSV文件中读取Titanic数据集
2 data = pd.read_csv("titanic.csv", encoding="utf-8")
3 # 设置支持中文的字体
4 plt.rcParams['font.sans-serif'] = ['SimHei']
5 # 重命名列
6 data = data.rename(columns={'Survived': '结果'})
7 data = data.rename(columns={'Sex': '性别'})
8 data = data.rename(columns={'Pclass': '舱位'})
9 data = data.rename(columns={'Age': '年龄'})
10 # 重命名列中的数据
11 data['性别'] = data['性别'].replace({'female': '女性', 'male': '男性'})
12 data['结果'] = data['结果'].replace({0: '死亡', 1: '幸存'})
13 data['舱位'] = data['舱位'].replace({1: '一等舱', 2: '二等舱', 3: '三等舱'})

```

2.2.3 计算乘客性别、年龄分布

计算乘客性别分布时只需要计算男乘客和女乘客的人数即可，但计算乘客年龄分布时得先将年龄分为不同的区间，再分别计算不同年龄区间上的乘客人数。性别分布用gender_distribution表示，年龄分布用age_distribution表示。

```

1 #计算乘客性别分布
2 gender_distribution = data['性别'].value_counts()
3 #计算乘客年龄分布
4 bins = [0, 10, 20, 30, 40, 50, 60, 70, 80]
5 labels = ['0-10', '11-20', '21-30', '31-40', '41-50', '51-60', '61-70', '71+']
6 data['年龄区间'] = pd.cut(data['年龄'], bins=bins, labels=labels, right=False)
7 age_distribution = data['年龄区间'].value_counts()

```

2.2.4 分析性别、船舱和年龄对生存概率的影响

以船舱这个因素为例，先计算不同舱位等级的乘客的生还人数、死亡人数，再用（不同舱位等级的生还人数/该舱位等级的乘客总数）来刻画不同舱位的乘客的生还率。在分析性别和年龄对生存概率的影响时采用同样的方法，但在分析年龄对生存概率的影响时，与计算年龄分布时相同，分析的是不同年龄区间上乘客的生还人数与生还率。不同性别的生还人数和死亡人数用survival_by_sex表示，不同性别的生还率用survival_rate_by_sex表示；不同舱位等级的生还人数和死亡人数用survival_by_pclass表示，不同舱位等级的生还率用survival_rate_by_pclass表示；不同年龄区间的生还人数和死亡人数用survival_by_age表示，不同年龄区间的生还率用survival_rate_by_age表示；

```

1 # 分析性别对生存概率的影响
2
3 # 计算每个性别的生还人数
4 survival_by_sex = data.groupby(['性别', '结果']).size().unstack(fill_value=0)
5
6 # 计算每个性别的生还率
7 survival_rate_by_sex = survival_by_sex.div(survival_by_sex.sum(axis=1),
8 axis=0) * 100
9 # 分析船舱（舱位等级）对生存概率的影响
10 survival_by_pclass = data.groupby(['舱位', '结
果']).size().unstack(fill_value=0)
11 survival_rate_by_pclass =
12 survival_by_pclass.div(survival_by_pclass.sum(axis=1), axis=0) * 100
13
14 # 分析年龄对生存概率的影响
15 survival_by_age = data.groupby('年龄区间').size()
16 survival_rate_by_age = survival_by_age.div(survival_by_age.sum())
17
18 # 分析多个因素对生存概率的影响
19 survival_by_all = data.groupby(['舱位', '年龄区间', '性别']).size()
20 survival_rate_by_all = survival_by_all.div(survival_by_all.sum())
21
22 # 分析不同因素对生存概率的影响
23 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
24 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
25
26 # 分析不同因素对生存概率的影响
27 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
28 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
29
30 # 分析不同因素对生存概率的影响
31 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
32 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
33
34 # 分析不同因素对生存概率的影响
35 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
36 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
37
38 # 分析不同因素对生存概率的影响
39 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
40 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
41
42 # 分析不同因素对生存概率的影响
43 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
44 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
45
46 # 分析不同因素对生存概率的影响
47 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
48 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
49
50 # 分析不同因素对生存概率的影响
51 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
52 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
53
54 # 分析不同因素对生存概率的影响
55 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
56 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
57
58 # 分析不同因素对生存概率的影响
59 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
60 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
61
62 # 分析不同因素对生存概率的影响
63 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
64 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
65
66 # 分析不同因素对生存概率的影响
67 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
68 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
69
70 # 分析不同因素对生存概率的影响
71 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
72 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
73
74 # 分析不同因素对生存概率的影响
75 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
76 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
77
78 # 分析不同因素对生存概率的影响
79 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
80 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
81
82 # 分析不同因素对生存概率的影响
83 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
84 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
85
86 # 分析不同因素对生存概率的影响
87 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
88 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
89
90 # 分析不同因素对生存概率的影响
91 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
92 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
93
94 # 分析不同因素对生存概率的影响
95 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
96 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())
97
98 # 分析不同因素对生存概率的影响
99 survival_by_factor = data.groupby(['舱位', '年龄区间']).size()
100 survival_rate_by_factor = survival_by_factor.div(survival_by_factor.sum())

```

```
11 #分析年龄对生存概率的影响
12 bins = [0, 12, 18, 60, 100]
13 labels = ['儿童(0-12)', '青少年(12-18)', '青年人(18-60)', '老年人(60-100)']
14 data['年龄区间'] = pd.cut(data['年龄'], bins=bins, labels=labels, right=False)
15 survival_by_age = data.groupby(['年龄区间', '结果']).size().unstack(fill_value=0)
16 survival_rate_by_age = survival_by_age.div(survival_by_age.sum(axis=1), axis=0) * 100
```

2.2.5 展示数据分析结果

首先将计算得到的数据（性别分布、年龄分布、不同性别/舱位等级/年龄区间的生存率）直接输出

```
1 print(gender_distribution)
2 print(age_distribution)
3 print(survival_rate_by_sex)
4 print(survival_rate_by_pclass)
5 print(survival_rate_by_age)
```

输出的数据如下图所示（幸存和死亡这两列的数字表示的是幸存/死亡的乘客的百分比）：

男性	577	
女性	314	
Name: 性别, dtype: int64		
21-30	220	
31-40	167	
11-20	102	
41-50	89	
0-10	62	
51-60	48	
61-70	19	
71+	6	
Name: 年龄区间, dtype: int64		
结果	幸存	死亡
性别		
女性	74.203822	25.796178
男性	18.890815	81.109185
结果	幸存	死亡
舱位		
一等舱	62.962963	37.037037
二等舱	47.282609	52.717391
三等舱	24.236253	75.763747
结果	幸存	死亡
年龄区间		
儿童(0-12)	57.352941	42.647059
青少年(12-18)	48.888889	51.111111
青年人(18-60)	38.608696	61.391304
老年人(60-100)	26.923077	73.076923

接着，使用matplotlib库进行绘图可视化，这里我选择采用柱状图的形式展示数据，因为柱状图易于理解，可以直观地比较不同类别的数据，并清晰地多维度展示数据的分布情况。

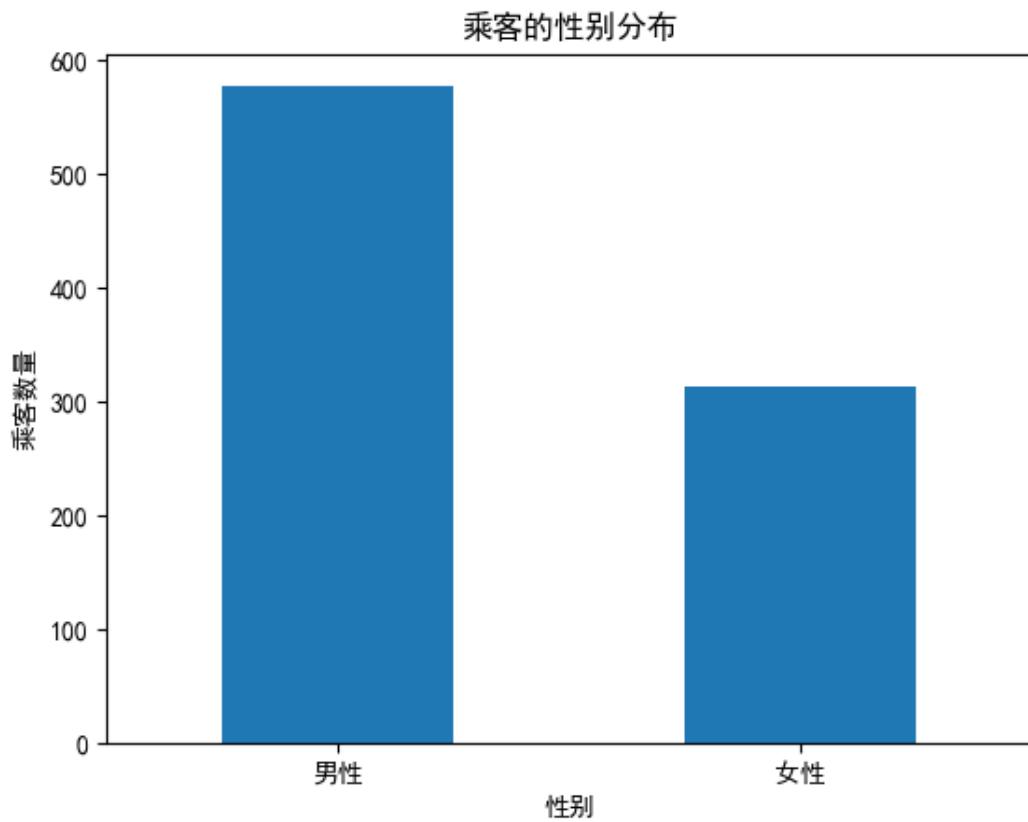
```

1 # 可视化性别分布
2 gender_distribution.plot(kind='bar')
3 plt.title("乘客的性别分布")
4 plt.xlabel("性别")
5 plt.ylabel("乘客数量")
6 plt.xticks(rotation=0)

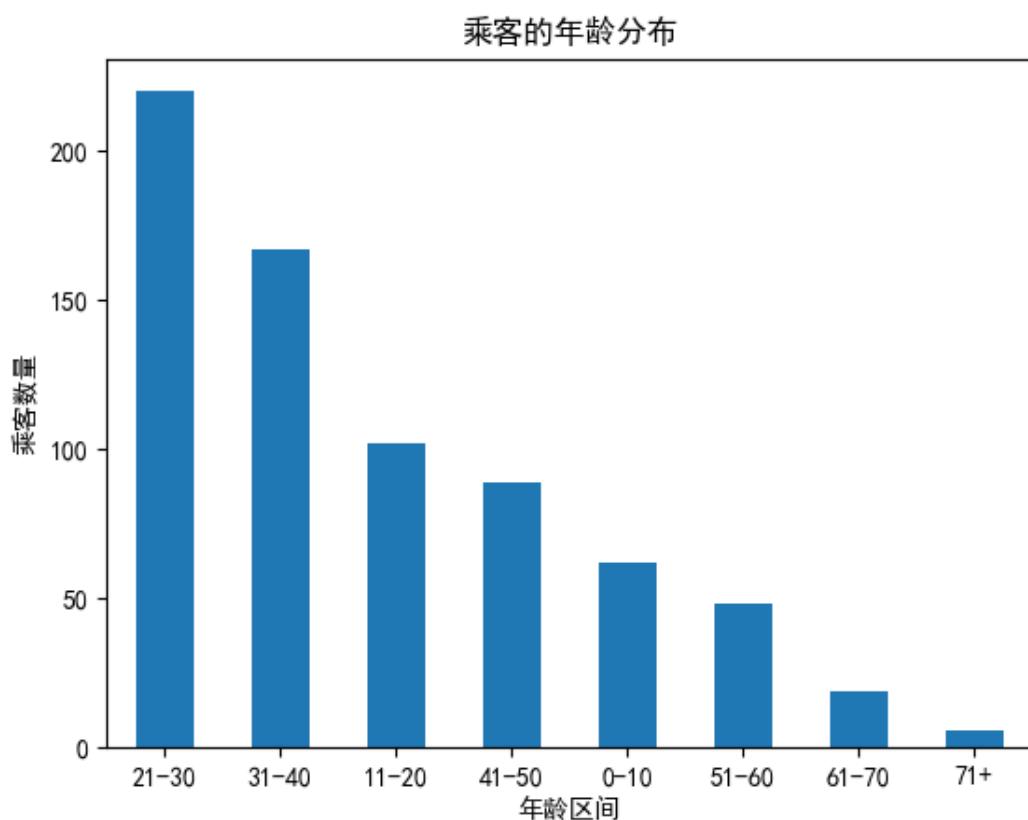
```

```
7 plt.show()
8 # 可视化年龄分布
9 age_distribution.plot(kind='bar')
10 plt.title('乘客的年龄分布')
11 plt.xlabel('年龄区间')
12 plt.ylabel('乘客数量')
13 plt.xticks(rotation=0)
14 plt.show()
15 #可视化性别对生还概率的影响
16 survival_rate_by_sex.plot(kind='bar', stacked=False, figsize=(10, 6))
17 plt.title('性别对生还概率的影响')
18 plt.xlabel('性别')
19 plt.ylabel('幸存/死亡比例 (%)')
20 plt.xticks(rotation=0)
21 plt.legend(title='是否存活', labels=['幸存', '死亡'])
22 plt.show()
23 #可视化船舱对生还概率的影响
24 survival_rate_by_pclass.plot(kind='bar', stacked = False, figsize=(10, 6))
25 plt.title('船舱对生还概率的影响')
26 plt.xlabel('舱位等级(Pclass)')
27 plt.ylabel('幸存/死亡比例 (%)')
28 plt.xticks(rotation=0)
29 plt.legend(title='是否存活', labels=['幸存', '死亡'])
30 plt.show()
31 #可视化年龄对生还概率的影响
32 plt.title('年龄对生还概率的影响')
33 plt.xlabel('年龄区间')
34 plt.ylabel('生存率 (%)')
35 plt.xticks(rotation=0)
36 plt.legend(title='是否存活', labels=['幸存', '死亡'])
37 plt.show()
```

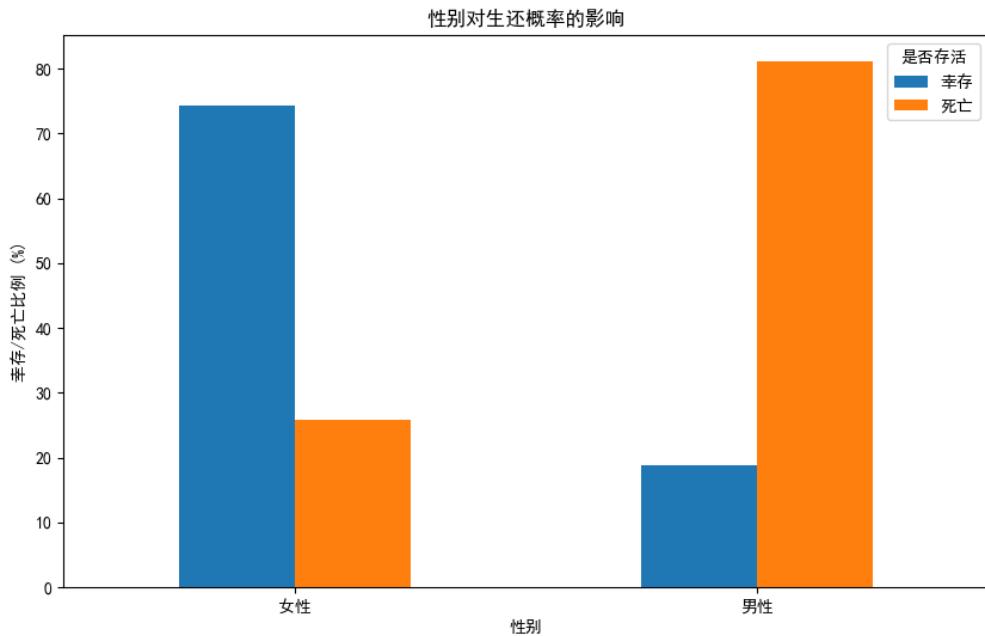
用matplotlib可视化性别分布如图所示：



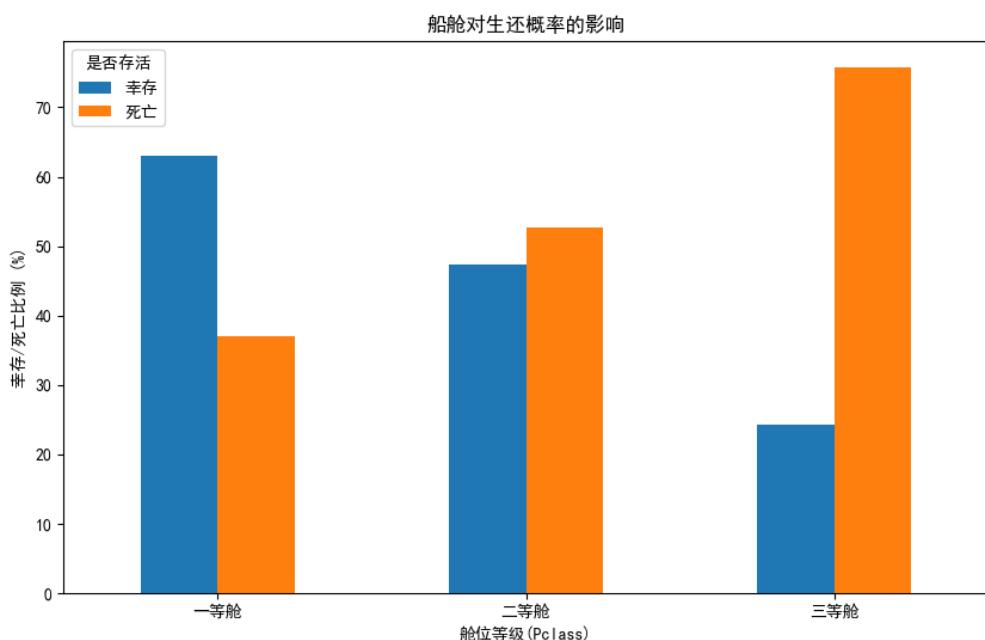
用matplotlib可视化性别分布如图所示：



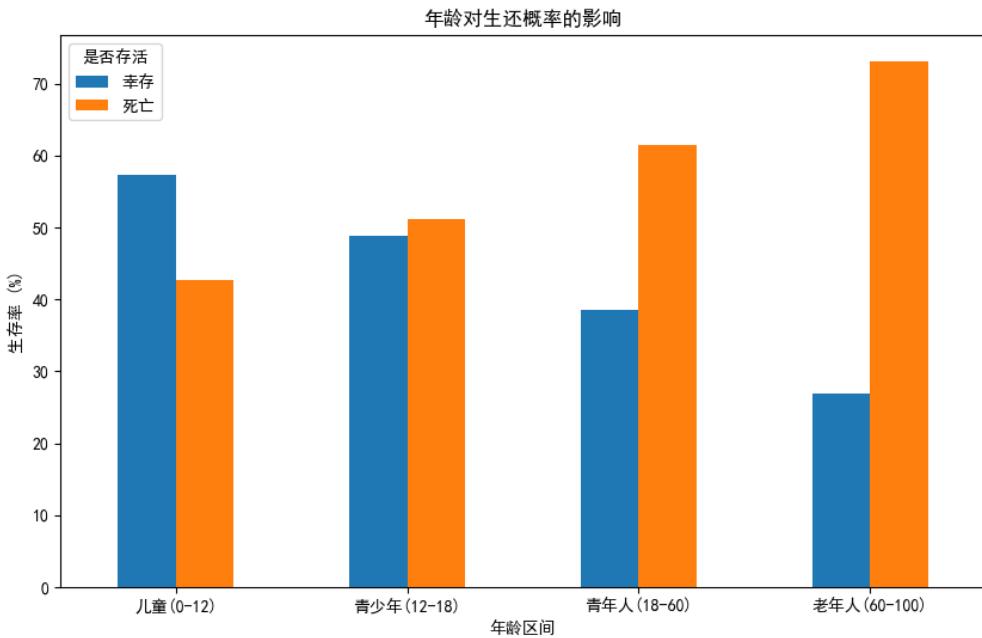
用matplotlib可视化性别对生还概率的影响如图所示：



用matplotlib可视化船舱对生还概率的影响如图所示：



用matplotlib可视化年龄对生还概率的影响如图所示：



3 总结

经过上述对Titanic的数据的分析后，可以从展示的数据分析结果得到以下结论：

1. 男性乘客的人数(577)明显多于女性人数(314)
2. 乘客的年龄主要集中在21-50
3. 女性乘客的生还概率(74.2%)显著高于男性乘客的生还概率(18.8%)
4. 舱位等级越高，乘客的生还概率越大(一等舱： 63.0% 二等舱： 47.3% 三等舱： 24.2%)
5. 儿童的生还概率最高(57.4%)，青少年次之(48.9%)，青年人更低(38.6%)，老年人最低(26.9%)

以下是我个人基于上述结论的一些推测：

1. 事故发生后，泰坦尼克号上可能坚持的救援原则是先救女人和未成年人(儿童+青年人)，这才导致人数占比较低的女性乘客的生还率显著高于男性乘客，儿童和青年人的生还率显著高于青年人和老年人。
2. 事故发生后，舱位等级越高的船舱的乘客能更及时地得到救援，这才导致舱位等级越高，乘客的生还率越大。

4 参考文献

- 1.<https://www.w3schools.com/python/default.asp>
- 2.https://blog.csdn.net/2401_85291273/article/details/139389100
- 3.https://blog.csdn.net/2201_75791084/article/details/139370223