泰坦尼克号数据分析报告

背景介绍

“泰坦尼克号”的沉没是历史上最为惨重的一次海难。1912年4月15日，泰坦尼克号在处女航中与冰山相撞后沉没，2224名乘客和机组人员中有1502人死亡。这场耸人听闻的悲剧震惊了国际社会，从而为以后建造的船只带来了更好的安全条例。这次沉船事故造成人员伤亡的原因之一是救生艇不够。虽然在沉船事件有部分幸存者，但还是有些人比其他人更有可能幸免于难，比如妇女、孩子和上流社会。在这个项目中，我将创建一个可视化，显示那些幸存的和死亡的之间的乘客信息的差别。

1 选择数据集

数据集来自于 Udacity网站提供（泰塔尼克号数据）。在这个数据集中，它包含来自泰坦尼克号上的2224名乘客和机组人员的人员的信息。

2 数据分析探索

2.1 数据描述

数据变量包括：乘客ID（Passenger ID）,幸存（Survived），乘客等级<高/中/低>（Pclass），性别（Sex），年龄（Age)，堂兄弟/妹个数(SibSp)，父母与小孩个数(Parch)，船票信息(Ticket)，票价(Fare)，客舱(Cabin)，登船港口(Embarked)。

1. Pclass: Passenger Class (1 = 1st (Upper); 2 = 2nd (Middle); 3 = 3rd (Lower))
2. Survived: 0 = No; 1 = Yes

Age: age is in Years; Fractional if Age less than One (1); If the Age is estimated, it is in the form xx.5
SibSp: Number of Siblings/Spouses Aboard
Parch: Number of Parents/Children Aboard
Fare: Passenger Fare (British pound), is in Pre-1970 British Pounds; Conversion Factors: 1 = 12s =240dand1s=20d
Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
参考信息来自于Kaggle：*https://www.kaggle.com/c/titanic/data*

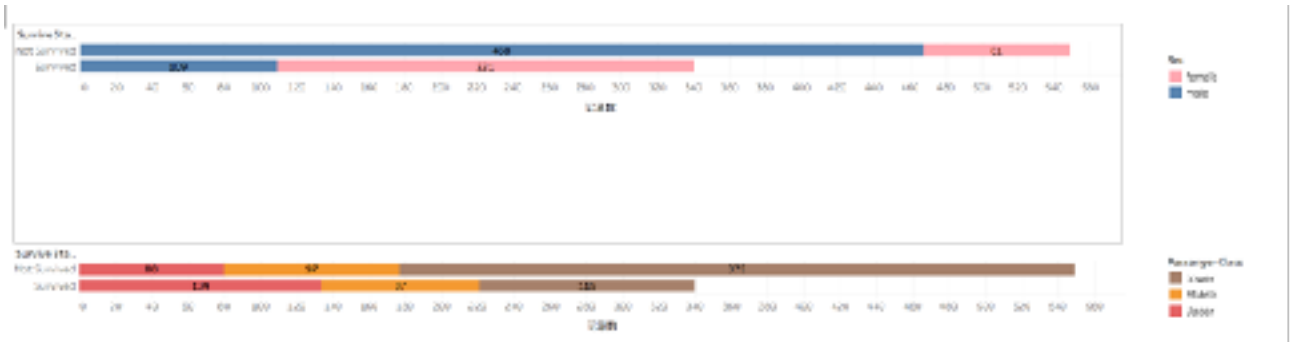
2.2 数据清理

在目测数据集后，我用Python清洗了数据，我清洗掉了任何空的或者重复的乘客ID，然后用以下方法修改了数据的列，并将清理后的数据集保存为“clean.csv”为csv文件里。

1. Pclass-column:用“Upper,Middle,Lower”替换数据集中的数字“1，2，3”。
2. 用“Not survived,Survived”替换数据集中的数字“0，1”.
3. Age column:将年龄划分为7个不同的范围（0-10、11-20、21-30、31-40、41-50、51-60、61-70、71-80）。
4. Embarked column: 用全称"Cherbourg, Queenstown, Southampton" 替换了数据集中的"C, Q, S”符合。

3 创建可视化

在本节中，使用Tableau创建了可视化，用以解释并应道读者认识理解数据集中的关键信息。共有两个版本，一个是初始版本，一个是通过分析反馈改进的修改版。



Survive Status	Sex	记录数	沿着 区(向下)的 记录数 的总计 %
Not Survived	female	81.0	14.75%
	male	468.0	85.25%
Survived	female	231.0	67.94%
	male	109.0	32.06%

Survive Sta..	Passenger ..	记录数	沿着 区(向下)的 记录数 的总计 %
Not Survived	Lower	372.0	67.76%
	Middle	97.0	17.67%
	Upper	80.0	14.57%
Survived	Lower	119.0	35.00%
	Middle	87.0	25.59%
	Upper	134.0	39.41%

Survive Sta..	Embarked	记录数	沿着 区(向下)的 记录数 的总计 %
Not Survived	Cherbourg	75.0	13.66%
	Queenstown	47.0	8.56%
	southampton	427.0	77.78%
Survived	Cherbourg	93.0	27.35%
	Queenstown	30.0	8.82%
	southamptor	217.0	63.82%

图形 **1:** 幸存者和未幸存者在性别，舱位，登船港口信息

通过读者的反馈后，最终我把乘客的信息加入到了我的故事里。

将乘客信息如“生还情况、性别、船票信息、登船港口”。这是为了让读者更容易筛选找到乘客，然后我用条形图显示幸存者和未幸存者在性别上的差异，在340名幸存者中，女性乘客占67.94%男性乘客占32.06%，在549名未幸存的乘

客中男性占85.25%女性占14.75%，幸存乘客中头等舱高于中等舱和下等舱。登船人数最多的港口是南安普顿港口。[图形1]

Sex	Age Range	平均值 Age	记录数	沿着 区(向下) 的 记录数 的总计 %
female	0-10	4.6	31.0	9.94%
	11-20	16.8	46.0	14.74%
	21-30	25.4	81.0	25.96%
	31-40	35.3	54.0	17.31%
	41-50	45.5	31.0	9.94%
	51-60	55.1	14.0	4.49%
	61-70	63.0	2.0	0.64%
	None		53.0	16.99%
male	0-10	4.0	33.0	5.72%
	11-20	17.7	69.0	11.96%
	21-30	25.4	149.0	25.82%
	31-40	34.9	100.0	17.33%
	41-50	45.3	55.0	9.53%
	51-60	54.8	28.0	4.85%
	61-70	64.1	14.0	2.43%
	71-80	73.3	5.0	0.87%
	None		124.0	21.49%

图形 **2:** 幸存和未幸存乘客的年龄范围信息

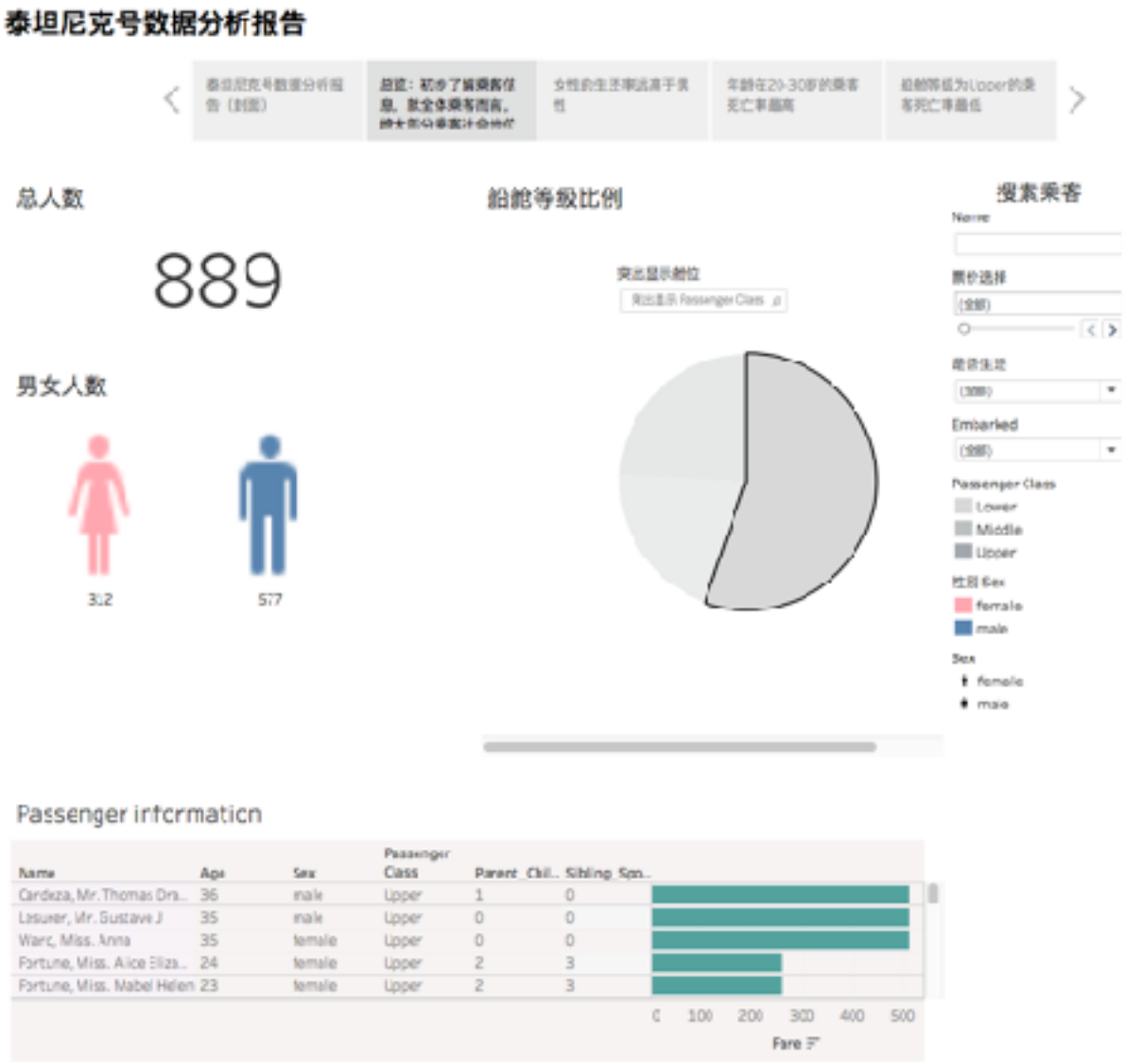
在所有乘客中，女性大多在*11-40*岁之间，其中*21-30*岁年龄段所占比例最高，为*25.96%*；*10*岁以下和*60*岁以上年龄段所占比例较小。男性大多在*11-40*岁之间，*21-30*岁年龄段所占比例最高，为*25.82%*，*10*岁以下和*60*岁以上年龄段所占比将相对较少。*[*图形 *2]*

然后用直方图来显示不同性别和不同舱位之间的平均票价。因为对于读者来说，看到最高、中等和最低的平均票价更加明显，而且读者可以清楚进行比较。不

同等级的女性平均票价高于男性，男女头等舱的乘客平均票价高于其他两类乘客。

大多数旅客带着孩子，其中有22.10%的乘客和10岁以下的乘客没有生还。

有27.3%和21.60%个兄弟姐妹在不到10岁的乘客没有幸存下来。



4 获取反馈后改善的可视化

[获取反馈后的改进后的可视化图表链接]

4.1 获取反馈

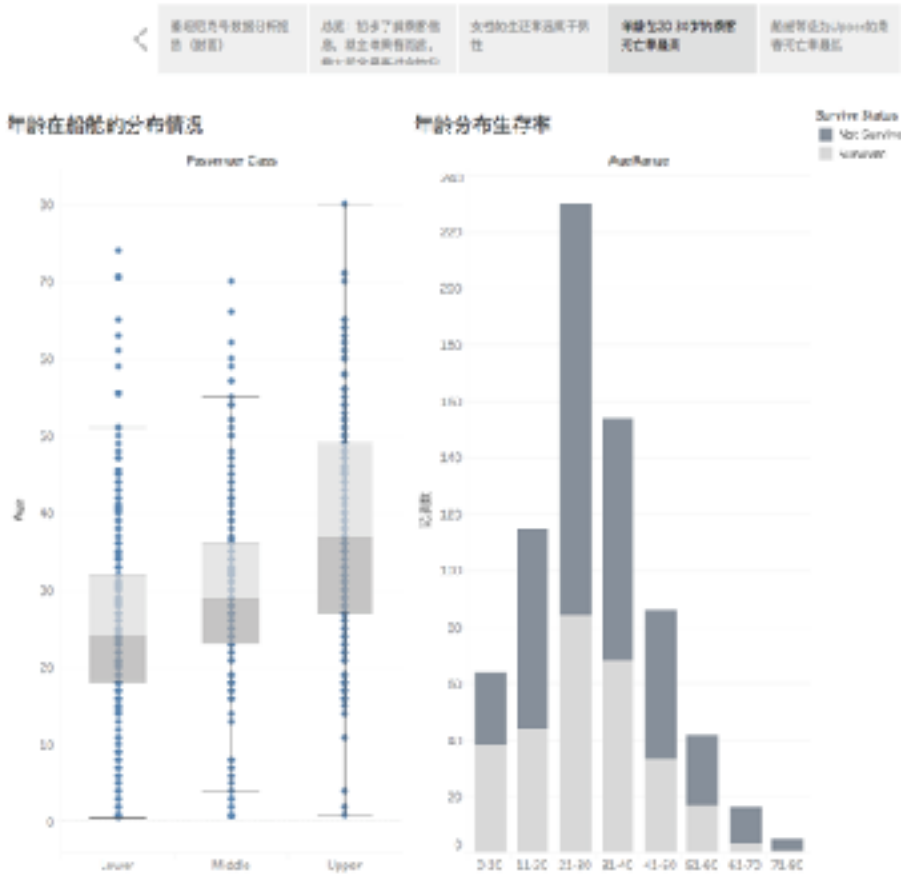
1. 优点：初始版本的图标能准确展示各类分析数据，

2. 缺点：应该给每个所展示的图表做一个结论，图表凌乱，图表表现形式可以更丰富些，可视化图表缺少交互，
3. 反馈意见：故事版未串联起来，读者无法得到清晰的主旨；可视化最大优点就是用图形证明所论述的观点（让读者看图便可得出结论，无需读者再做过多的计算）；个别论点叙述不够清晰；做一些表与表的交换使得读者有更好的体验；多尝试不同类型的图形，但颜色不易过多，使用恰当的图形类型更有利与证明你的观点；尝试编辑一些聚合数据。

4.2 改善可视化

根据所获取的反馈，我从新设计了我的图形。通过对比其他变量与存活率的关系做可视化，并且得到了一下结论：

总体来讲，生还率主要受三大因素影响：性别、年龄和社会阶层，女性高于男性，年龄小的高于年龄大的，社会阶层高的大于社会阶层低的。由此可见，虽然此事故在历史上有著名的“让妇女和孩子先走”的经典故事，但其真实情况也任然建立在社会阶层的基础上。



结论：通过对对数据做可视化分析得出结论：女性和儿童（尤其是上流社会的）生还率更高（女性生还率为74%，全员平均生还率38%），较富裕阶层（舱别更高的）的乘客也更有生还的可能。