泰坦尼克号数据集可视化

总结

该项目是泰坦尼克号数据可视化，此数据集包含了泰坦尼克号上的人口统计数据和人员信息。重要的变量包含性别，年龄，仓位，是否生存等。通过可视化，发现女性存活率远大于男性存活率。 头等舱存活率大于二等舱和三等舱。 10 岁以下的儿童生存率比大部分人高。其中头等舱女性的存活率最高。二等舱和三等舱女性的存活率也要高于头等舱男性存活率头等舱二等舱的儿童存活率比头等舱的青中年老年人都高。三等舱的老人存活率最低。结论：在发生灾难时，泰坦尼克号的人员把生存的几率让给了女性，孩子以及富人。

设计

第一版设计总结如下：

数据清理，对年龄的缺失值用中位数做补充，乘客登船码头 Embarked 用众数替代缺失值。数据清理后首先是想看看这个数据集里面包含哪些变量。哪些变量与最终生还情况有关。

看看变量的分布情况：

船上船舱有多少？我们选用气泡图，从颜色和大小上看分了三个舱位，其中三等舱的数量最多。

每个舱位里，男女比例如何？这里我们选用了百分比堆积条形图。此图更直观的展示出在每个舱位里，男性的比例都要比女性多。

哪些变量与最终生还情况有关：

想要哪些变量影响了最终的生还情况。不同性别，不同年龄，不同仓位下的存活率有没有区别？哪一个更高？对性别和舱位分别考察，联合考察。不同年龄阶段，生存有多大的差异？

年龄的生还情况：

船上人员年龄大概是多少岁？哪个年龄段的人比较多？针对这个做了一个年龄直方图。更直观的展示了年龄的分布，显示出 20 到 30 岁的人数最多。把年龄与生存比例做条图。

性别和舱位的生还情况：

把性别，舱位交叉表列出来，比较直观的反映出不同舱位，不同性别的人的存活率。把总的存活率也显示出来了。

条形图加上颜色分类之后，更明显的展示出头等舱的女性存活率最高。

版本 1

第一版反馈

反馈问题：

1 最好加一个封面更能表明此项目
2 年龄的数据桶可以再调整一下，加一个参数
3 色号有些乱，分不清
4 文字描述部分较少，有些图不是太理解

5 能不能多些酷炫图形

第一版反馈后修改

版本 2

第二版的设计如下：
针对反馈 1，此版本添加了封面和结尾页面
针对反馈 2，对年龄添加参数，有利于交互，查看不同年龄段的年龄分布。
针对反馈 3，把颜色重新修改利于辨认。
针对反馈 4，在故事页面里添加了一些说明。
针对反馈 5，添加了一些酷炫图，利用变量 embarked 采取树状图，希望可以反映船上人员从哪个港口上来的多？哪个港口的票价最高。儿童老人中青年人的票价是否一样。对变量 SibSp 和 Parch 进行分类，直观展示出相关的存活率。

第二版反馈

1 故事逻辑不缜密，展示内容与结论有些不相关。
2 结论有误：老年人并没有获得很高的生存率

第二版反馈后修改

最终版：titanic 故事

最终版的设计如下：
针对反馈 1，删除与结论不相关的 embarked 树状图，删除了 SibSp 和 Parch 分类条形图，专注于观察性别，舱位，年龄和生存之间的关系。
针对反馈 2，仔细观察了年龄的条形图和年龄生还情况，发现 80 岁的生化率是 100%，但他只是一个人，不具有代表性。甚至通过观察，发现老年人的存活率并不高。

除此之外，我还增加了
1 舱位和年龄交叉分析情况，采用树状图可以清晰的查看不同舱位不同年龄人员的存活率，并且列出相关的文字表，方便查看具体情况。
2 性别、舱位、年龄交叉分析生还情况，通过条形图清楚展示。