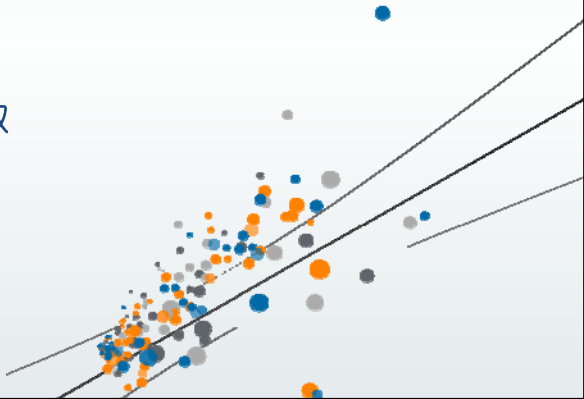


案例1：泰坦尼克号旅客逃生情况分析

- 各种标准/非标准格式本地数据的读取
- 如何对数据作进一步的整理和编辑
- 各类统计表的制作
- 各种基本统计图的制作



案例概况

- 号称永不沉没的泰坦尼克号于1912年在首次航行中误触冰山沉没，共造成了1517人丧生，事发时乘客约有1317人，共498人幸存。数据文件titanic passenger list.csv包括了目前收集到的所有已知1309名乘客的详细数据，包括其姓名、性别、年龄、仓位以及是否生还等信息，现希望分析并展示生还情况和这些变量之间的关系。
 - 结局变量：是否生还
 - 分析变量：性别、年龄、仓位

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin
1	1	Allen, Miss. Elisabeth ...	female	29.0000	0	0	24160	211.338	B5
1	1	Allison, Master. Huds...	male	0.9200	1	2	113781	151.550	C22 C26
1	0	Allison, Miss. Helen L...	female	2.0000	1	2	113781	151.550	C22 C26
1	0	Allison, Mr. Hudson J...	male	30.0000	1	2	113781	151.550	C22 C26
1	0	Allison, Mrs. Hudson ...	female	25.0000	1	2	113781	151.550	C22 C26
1	1	Anderson, Mr. Harry	male	48.0000	0	0	19952	26.550	E12

本地标准格式数据文件的读取

- Tableau中使用的数据结构是标准的关系型数据库中的二维表结构
 - 非二维表结构的数据在读入后都需要转换为对应结构才会加以使用
- 只要是符合二维表格式，即第一行是变量名，第二行起是数据区的数据文件，Tableau都会自动读入相应的变量设定和数值，并为所有变量自动寻找最佳的存储类型



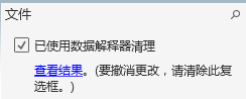
@文彤老师

Tableau案例培训课程

25

非标准数据文档的读取

- 当xls或者文本数据源的头部/尾部存在多余的解释行时，默认情况下他们将会被作为数据内容读入，此时可以使用数据解释器进行自动清理
- Miislog.txt为IIS服务器日志导出的文本文件，前三行均为注释



```
miislog.txt x
1 #Software: Microsoft Internet Information Server 4.0
2 #Version: 1.0
3 #Date: 0000-00-00 00:00:00
4 #Fields: date time c-ip cs-username s-sitename s-computername s-ip cs-method cs-uri-
5 18/02/2001 00:01:20 218.236.234.118 - fictitious_website big-fred 123.45.678.901 GET
6 18/02/2001 00:01:20 218.236.234.118 - fictitious_website big-fred 123.45.678.901 GET
7 18/02/2001 00:01:20 218.236.234.118 - fictitious_website big-fred 123.45.678.901 GET
8 18/02/2001 00:01:20 218.236.234.118 - fictitious_website big-fred 123.45.678.901 GET
9 18/02/2001 00:01:20 218.236.234.118 - fictitious_website big-fred 123.45.678.901 GET
10 18/02/2001 00:01:20 218.236.234.118 - fictitious_website big-fred 123.45.678.901 GET
```

@文彤老师

Tableau案例培训课程

26

数据透视表方式读取

- 重复测量数据的记录格式
 - 宽型：每一个个体被记录为一个Case，所有测量被记录在不同的变量中

id	城市	性别	年龄	index1.200704	index1.200712	index1.200812	index1.200912
1	北京	女	22	109.349	124.971	54.675	85.9174
2	北京	女	19	93.728	101.539	101.539	109.3494
3	上海	女	26	93.728	85.917	109.349	117.1600

- 长型：每一次测量被单独记录为一个Case

#	Abc	Abc	#	Abc	#
CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表	数据透视表	数据透视表
id	城市	性别	年龄	数据透视表字段名称	数据透视表字段值
1	北京	女	22	index1.200704	109.349
2	北京	女	19	index1.200704	93.728
3	上海	女	26	index1.200704	93.728
4	北京	男	29	index1.200704	109.349
5	上海	男	45	index1.200704	85.917
6	北京	女	23	index1.200704	101.539

@文彤老师

Tableau案例培训课程

27

数据透视表方式读取

- 当原始数据为重复测量的宽表形式时，可以使用数据透视表方式进行读取后的长→宽格式转换

#	Abc	Abc	#	#	#	#	#
CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表
id	城市	性别	年龄	index1.200704	index1.200812	index1.200912	
1	北京	女	22	109.349	54.675	85.9174	
2	北京	女	19	93.728	101.539	109.3494	
3	上海	女	26	93.728	109.349	117.1600	
4	北京	男	29	109.349	78.107	101.5387	
5	上海	男	45	85.917			

- 数据：CCSS宽表.xlsx

#	Abc	Abc	#	Abc	#
CCSS宽表	CCSS宽表	CCSS宽表	CCSS宽表	数据透视表	数据透视表
id	城市	性别	年龄	数据透视表字段名称	数据透视表字段值
1	北京	女	22	index1.200704	109.349
2	北京	女	19	index1.200704	93.728
3	上海	女	26	index1.200704	93.728
4	北京	男	29	index1.200704	109.349
5	上海	男	45	index1.200704	85.917
6	北京	女	23	index1.200704	101.539

@文彤老师

Tableau案例培训课程

28

基本的数据整理操作

- 名称与重命名
- 更改数据类型：数值、日期、字符、逻辑
- 字符型变量
 - 别名
 - 数值拆分
- 数值型变量
 - 数值分段（创建级）
- 创建
 - 新变量
 - 数据组
- 隐藏数据列

基本的数据显示设定

  排序字段 数据源顺序

☐ 显示别名 ☐ 显示隐藏字段 1,000 行

- 元数据视图切换
- 字段显示顺序
- 切换显示别名
- 显示隐藏字段
- 设定显示行数
- 案例排序方式

表格的基本框架

- 行 (Row) : 形成表格横行的元素
- 列 (Column) : 形成表格列的元素
 - 行、列元素相交就会形成一个最简单的二维表, 行、列元素不同取值的组合就确定了一个单元格
- 层 (Layer) : 表格中的第三个维度
 - 不妨把此时的表格想象成一个立方体, 行、列、层就对应了该立方体的长、宽和高
 - Tableau表格中未直接提供层维度的操作, 而是用分页方式实现分层效果
- 单元格: 表格呈现数据信息的基本单位, 可以是计数、百分比、均值等任何信息

sex		舱位		
female	male	二等舱	三等舱	头等舱
88.7%	14.6%	49.1%	15.2%	96.5%

@文彤老师

Tableau案例培训课程

31

几种基本表格类型: 叠加表 (Stacking)

- 指在同一张表格中对两个变量进行描述, 或者说表格中有一个维度的元素是由两个以上的变量构成
- 叠加表其实可以被简单的理解为为每个变量分别绘制两个简单的报表, 然后将它们拼接到一起
- 也存在横向拼接的叠加表
- Tableau中未直接提供叠加表, 但可以用仪表板方式得到相同效果

sex		舱位		
female	male	二等舱	三等舱	头等舱
72.7%	19.1%	43.0%	25.5%	61.9%

@文彤老师

Tableau案例培训课程

32

几种基本表格类型：交叉表（Crosstabulation）

- 是观察两个分类变量间联系时最常用的表格技术，它的两个维度都是由分类变量的各类别（及汇总）构成

Tableau 工作表 1 的交叉表视图。列架包含“舱位”，行架包含“sex”。

	舱位		
sex	二等舱	三等舱	头等舱
female	88.7%	49.1%	96.5%
male	14.6%	15.2%	34.1%

Tableau 工作表 1 的交叉表视图，包含总计行和总计列。

	舱位			
sex	二等舱	三等舱	头等舱	总计
female	88.7%	49.1%	96.5%	72.7%
male	14.6%	15.2%	34.1%	19.1%
总计	43.0%	25.5%	61.9%	38.2%

@文彤老师

Tableau案例培训课程

33

几种基本表格类型：嵌套表（Nesting）

- 也可以用于显示两个分类变量间的联系，但是在嵌套表中，这两个变量被放置在同一个表格维度中，即该维度是由两个变量的各种类别组合构成
- 嵌套表并不如交叉表直观。但是当每个单元格内需要呈现的统计指标非常多时，嵌套表更为美观和紧凑

Tableau 工作表 1 的嵌套表视图。列架包含“度量名称”，行架包含“sex”和“舱位”。

sex	舱位	平均值 age	平均值 survived	记录数
female	头等舱	37.0	96.5%	144.0
	二等舱	27.5	88.7%	106.0
	三等舱	22.2	49.1%	216.0
male	头等舱	41.0	34.1%	179.0
	二等舱	30.8	14.6%	171.0
	三等舱	26.0	15.2%	493.0

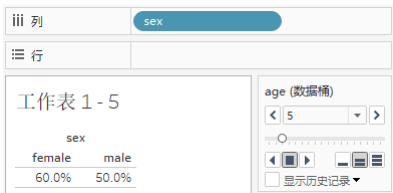
@文彤老师

Tableau案例培训课程

34

几种基本表格类型：多层表（Layers）

- 如果指定了层元素，则表格就由二维扩展到了三维，即多层表。事实上，多层表和嵌套表也非常的类似，只是现在我们只能每次观察到其中一层的数据而已
- 在数据仓库技术中，多层表也被称为数据立方体
- Tableau中主要是通过分页方式实现分层效果



工作表 1 - 5

sex	
female	male
60.0%	50.0%

age (数据桶)

< 5 >

显示历史记录

@文彤老师

Tableau案例培训课程

35

几种基本表格类型：复合表格

- 在实际的工作中，上述表格类型还有可能互相组合，以更好的达到相应的分析目的
- 比如叠加－交叉表（一个维度是分类变量，另一个维度则是两个分类变量的叠加）、嵌套－交叉表（一个维度是分类变量，另一个维度则是两个分类变量的嵌套）等

@文彤老师

Tableau案例培训课程

36

表格绘制的基本步骤

- 确定所需绘制表格的基本结构，如行、列都由什么构成，是否在表格中会出现多个元素的嵌套，有多少种汇总，是否有嵌套汇总等
- 绘制表格的基本结构。这里要将注意力集中在是否已经得到了所需表格结构上。如果结构还不相同，则继续直至完成
- 对细节进行完善，使单元格的输出格式符合要求
- 添加其余变量、统计量到表格中来，使表格中的内容满足相应问题的需求
- 对表格的附加文本和格式进行修饰
- 最后一次审核所绘制的表格，考虑有无需要改进之处

@文彤老师

Tableau案例培训课程

37

维度和度量

- 维度：大致对应着（无序/有序）分类变量，用于对案例进行分组
 - 字符串变量、日期时间变量、布尔（逻辑）变量默认设为维度
 - 也可强行将连续变量拖动设定为维度
 - 数据桶：分段后的数据桶会被作为维度进入分析
 - 度量名称：代表所有度量变量的集合
- 度量：大致对应连续变量，在图表中呈现为原始信息或者汇总信息
 - 数值变量默认设为度量
 - 也可强行将字符串变量拖动设定为度量
 - 记录数：代表符合筛选条件的案例数量
 - 度量值：代表相应度量的具体汇总数值，一般和度量名称等联合使用

@文彤老师

Tableau案例培训课程

38

泰坦尼克数据的表格化分析

- 对性别和舱位的分别考察
 - 数值标记的设定方式
- 联合考察
- 计算行/列合计
- 更改显示格式
- 在单元格内设定多个统计量
- 设定**survived**为维度
- 单元格内指标的不同表达方式
- 浏览数据

@文彤老师

Tableau案例培训课程

39

统计图的分类框架

- 统计图的分类方法有许多种，但作为数据分析和呈现的工具，最好使用和统计学体系最为贴近的方法对其加以分类
- 首先按照其呈现变量的数量，将统计图大致分为单变量图、双变量图、多变量图等
- 随后再根据相应变量的测量尺度进行更细的区分
- **Tableau**中提供了一些新式的图形，他们并不完全符合标准的统计绘图要求，对这些图形的使用应当谨慎，注意不要因此冲淡分析主题

@文彤老师

Tableau案例培训课程

40

统计图的分类框架：单个-分类变量

- 简单条图
 - 按照分类区分直条，直条高度代表频数大小
- 分段条图
 - 按照分类区分颜色，条段大小代表频数/构成比大小
- 饼图
 - 饼块大小代表频数/构成比大小
- 气泡图
 - 用气泡大小代表频数/构成比大小
 - 违背了统计图形应当便于对比数据的基本原则，很好看，但需要控制使用

@文彤老师

Tableau案例培训课程

41

统计图的分类框架：单个-数值变量

- 直方图
 - 对数值进行分组频数汇总，呈现整个取值区间上的数据分布基本特征
 - 在Tableau中是通过对原始数据生成分段变量（数据桶）来实现的
- 箱图
 - 使用百分位数体系刻画整个取值区间
 - 箱体最中间的粗线为P50（中位数），方框上下界为P25和P75（四分位数）
 - 数据用散点的方式加以表示
 - 与四分位数（即方框上下界）的距离超过1.5倍四分位间距（即方框长度）的都会被定义为离群值，相应的界限在图中以线段表示
 - 所有数值均未超界时，该线段就是最大/最小值

@文彤老师

Tableau案例培训课程

42

统计图的分类框架：数值因变量

- 简单条图
 - 呈现分类自变量的影响
 - 点图：基于条图直接衍生而来
- 线图
 - 呈现有序分类自变量（或者时间变量）的影响
 - 双线图：提供两个纵轴尺度，便于对比数值相差较大的两个指标
 - 面积图：基于线图直接衍生而来
- 散点图
 - 呈现连续自变量的影响

@文彤老师

Tableau案例培训课程

43

统计图的分类框架：分类因变量

- 基本都是使用各类条图对数据进行呈现
- 复式条图
 - 呈现两个分类变量各个类别组合情况下的频数分布
- 分段条图
 - 主要突出一个分类变量各类别的频数，并在此基础上表现两个类别的组合频数情况
- 百分条图（马赛克图）
 - 呈现在一个变量不同类别下，另一个变量各类别的百分比变化情况
- 树状图
 - 将两个分类变量置于同等地位，直接显示各个组合单元格所占的百分比

@文彤老师

Tableau案例培训课程

44

其余更复杂的图形

- 多个变量间关系的图形呈现方式
 - 最常见的方式为采用图例对二维图进行扩充
 - 组合统计图：根据实际需要自行设计，最常见的是线图/条图组合
- 统计地图
 - 与地图数据相结合
 - 是Tableau的特色功能之一
 - 可自定义地图数据
- 其他特殊用途的统计图
 - 甘特图：异化的条图，用于反映项目进展是否按照时间计划进行
 - 标靶图：在条图的基础上增加了目标值，可反映任务完成情况
 - 词云：用于直观反映各词汇在语料库中的出现频次

@文彤老师

Tableau案例培训课程

45

注意：图形并非越复杂越好！

- The most common disaster in illustrating is to include too much information in one figure.
- The more points made in an illustration, the more the risk of confusing and discouraging the reviewer.

--Briscoe,1990

@文彤老师

Tableau案例培训课程

46