用数据讲好故事 讲义

Tableau 可视化设计原则

<2018 年 10 月 18 日更新 V1.0>

如果有更新迭代的建议，请发送邮件至 kylie@udacity.com 并抄送 april.chen@udacity.com。谢谢。

「公开课讲解知识点」

* 可视化基础
    - 数据类型、视觉编码
    - Exploratory & Explanary 区别和联系
    - 图表选择、颜色、其他编码的使用原则
* 用数据讲好故事
* Tableau 资源及使用

0 说明

本讲义作为 Udacity 数据分析纳米学位课程的辅助，供 Udacity 数据分析进阶 VIP 班级公开课讲解使用。

公开课以“用可视化讲故事“ 为主线，向大家展示完成最终项目的基本思路和项目要求。讲解中会穿插可视化基础和设计原则。并不包含 Tableau 的实际操作，这部分会提供学习资源，如果有操作相关的问题可以提问助教。

1 为什么要用可视化讲故事

两个原因：

1. 数据分析师通常都有统计分析的背景，但很少有设计方面的训练。这使得绝大部分获取数据、清理整理数据、分析数据以及建立模型的环节都能顺利完成，而在最终沟通展示上力不从心，而展示结果是整个数据分析流程最终受众唯一能够接触到的环节。
2. 将数据转化为“信息”并由此作为驱动做出更好决策不是容易的事情，如果仅仅依赖工具处理和理解数据，而没有遵循清晰的逻辑路径，我们的想法和努力可能并不会取得预期的效果。对很多人而言，可视化等于提取数据并制作出图表，但绝大部分有趣的发现都会以平庸的方式被展现出来，让人难以或无法理解。

为什么我们学习用 Tableau 讲故事：

1. 本课程中可视化工具为 Tableau, 但学到的可视化设计原则适合于所有的可视化工具
2. 如果你是一个非技术人员走上分析岗位，那么利用 Tableau 制作数据可视化并进行沟通也并非难事，整个过程只需要极少（甚至不需要）任何代码，并且学习曲线非常平滑。

2 故事背景的选择 - 选择数据集

在开始前的数据选择阶段，你的重点应该放在思考以下几个问题：

- 我对哪个主题更感兴趣（有兴趣探索的数据才能让你讲出好故事，仅仅为了更快通过项目而选择“简单”的数据显然不是个好主意。
- 对选择的主题我有什么想解答的疑问？
- 通过什么样的结论我可以回答我以上的疑问？

2.1 GitHub 数据集介绍

优达课程侧重传授数据可视化知识，所以为数据准备了了较为干净的数据集（但实际情况中，通常我们需要对原始数据集进行行行清洗）。

简单介绍最受同学欢迎的 3 个数据集：

- 泰坦尼克号数据（初级）
  数据提供了泰坦尼克号上每位乘客的基本信息及生还情况，主要探索方向可以是所有以及生还乘客的社会组成。

- Prosper 贷款数据（中级）
  一个维度较广的数据源，来自美国 P2P 借贷平台 Prosper。探索方向很广阔，结合贷款的利率（预期以及实际值），贷款人的职业，贷款目的，贷款状态，贷款人负债比，贷款人就业状况等可以得出很多有趣的结论。

- 摩拜上海成区用户使用（中级）
  结构较简单，但需要简单清洗的数据。包含了一个月内所有行程的经纬度，难度在于经纬度并不能在 Tableau 中直接转化为地理信息，例如上海的行政辖区或地铁站。如果你对地图类图表（尤其是个性化地图，如 Mapbox）感兴趣，这将是一个很好的练习。

2.2 工作中的数据集

除了完成项目，我们也向同学介绍实际工作中可能用到的几种数据。实际工作中的数据源可以简单分为两类：

- 各式各样的 EXCEL（或 CSV），各个公司的各个部门会使用不同形式的模版表格来记录日常活动，有的模版已经使用了几年以上。如果是供应商或某个特定系统生成的 CSV，一般来说基本可以确保表头的存在，也就是说数据是从第二行开始的，并且每列有自己的列名。如果你需要展示的数据是某个透视表，表格并不开始于 A1，而你只需要静态地展示这些数据（譬如在 Power Point 中）。你需要简单地整理这些数据后再倒入 Tableau：新建一页 excel，确定表头并将数据以纯文本的格式粘贴，简单地验算（如对某个月或某个子项的数据求和）确保没有丢失数据，最后以新建页面作为数据源可以解决多大部分问题。Tableau 自带的数据解释器也可以自动对导入的表格做一些基本的处理。

- 如果你在一个中大型的公司担任数据分析工作，一般会被授予数据库访问权限。常见的是数据来自 CRM，财务，销售，采购，组织架构，产品信息等。组织越是复杂，数据库结构也会变得复杂。一个你需要的数据源往往需要关联多个表（一个 10 列的数据源甚至可能需要 9 次关联），也就是说，像项目中这么理想（干净）的数据源在实际工作中是基本不存在的。你需要和提供元数据的 IT 部门深入配合，理解数据库存储的逻辑，结合实际业务逻辑，利用 SQL 或 Tableau Prep 等工具准备动态或静态的数据。Tableau Prep 在对各类数据的去重、关联、聚合等处理上快速且简单，Prep 使用和 Tableau 相同的语法并支持多种数据源，是个值得尝试的辅助工具。

2.3 数据集的探索，可视化前的准备

拿到数据后，你可能会想粗略了解一下数据的概况。例如数据大概有多少个维度，这些维度大概可以提供什么样的信息。对于一些你关心的维度，数据的分布是什么样的。

举一个简单的例子，我有一个年级期末考试的表格，共有三列，分别是学号，科目和成绩。我可能会想知道：

- 一共有多少学生
- 一共有几门课
- 每个学生都参加了所有课程吗
- 如果不是，每个学生平均参加了几门课
- 哪几门课参加的人数最多，是因为这些是必修课吗
- 每门课的分数分布是如何的，有空值的分数吗
- 满分是 100 吗，及格率是多少
- 哪门课均分比较高
- 课程的受欢迎程度和分数分布或得高分的难易程度有关吗（一个相对有趣的问题被提出）

对于一个非常简单的表格，我尚可以做出大量类似的探索性分析，对于维度更多的项目数据源，这个过程更是必不可少。虽然这些结果只有很少一部分会最后展示在可视化中，但却是必须的。因为不排除所有“正常”的现象，就无法突出你寻找的数据“异常”。如果你还不清楚“异常”可能出现在哪就贸然开始可视化，你又如何突出你想表达的观点呢？

Tableau 或 Tableau Prep 可以让你的这些数据探索既迅速又直观。以刚刚想要进行的探索来说，所有的问题都可以在 5 分钟内通过拖拽回答，不需要任何代码，并且每个问题仅需要不超过 10 次键鼠的操作。如果你没有 Tableau Prep 的使用权限，也可以使用 Python / R 来实现相同的目的。

3 如何讲好故事

3.1 在数据中聚焦于一个具体、清晰的发现

数据可视化的成功并不始于图表的制作，除了对数据的探索，理解上下文应该更占据你的精力。数据可视化就像在牡蛎中寻找珍珠并卖出高价的过程。常见的错误是：在应该进行解释性分析（包装最后的“珍珠”，即结论）时进行探索性分析（简单的展示数据：一百个牡蛎）。向受众展示一切（探索性分析）是非常诱人的，这可以证明你工作及分析的可靠性，但你需要抑制住这个冲动而把注意力集中在解释性分析（将数据抽象为受众能消化的信息）上。试着在组织结论时回答以下问题：

- 你需要观众了解什么或者做什么
- 你在跟谁沟通
  （你的受众越具体，你就越能成功地沟通）
- 沟通时使用什么样的语气
（是轻松还是严肃，是庆祝成功还是鼓励行动）
- 什么样的数据可以用来支持你的观点

如果你只有分钟汇报或展示你的作品，你会说什么？所有的台词将组成你的标签页，作为你可视化视图的文字辅助 （如果你使用 Tableau 的故事功能）。
最终你需要提炼出一句间接的陈述（项目中的限制是四句，已经是相当宽松的要求了），这个中心思想包含三部分：

- 能陈述你的独特的观点
- 切中要害
- 是一个完整的句子

以泰坦尼克号数据为例，在不讨论结论正确性的假设下比较下面两个结论：

- 数据展示了乘客的组成，分析了哪些群体生还率更高（女人，孩子，结伴旅客，富人），证实了电影中妇女儿童优先上救生船的历史。
- 通过控制变量（舱位、亲属人数、年龄段），数据模型证实泰坦尼克号上女性比男性更容易生存（高出 xx%），女童的生存率接近 xx%而老年男女的生还率不受性别影响。

第一个结论看似没有问题，但是过于宽泛，并没有具体比较性别、年龄和所谓“阶级“之间对生还率变量的比重；第二个结论虽然简单，但可以展开更为详实的论证。

从这个故事的结构来看，仅仅平铺式地叙述数据事实肯定不够精彩，常用的故事结构有：

- 总分总
  最简单的作文结构，适用于 3-5 页的故事和较简单的数据

- 前言– A1 – A2 – B1- B2 – 结论
  常用的论文结构，举个例子，A1 和 A2 是 A 的宏观和微观两方面，B1 和 B2 是 B 的正反两面

- WHO-WHERE-WHEN-WHAT-WHY-HOW
  5W1H 回答式地解析你的中心思想

3.2 特定发现得以清楚地传达

3.2.1 图表类型

常见的图表有以下几种，无需要更多“花哨高端“即可满足绝大部分需要

- 简单文本
  只有一两个数据进行分享时，完全不需要任何图表，只是用数字（尽可能突出）配合辅助性的文字是最清晰有效的

- 表格（包括热力图）
  展示多维度的数据，让受众根据需要自行选择关注的点。
  热力图是表格的变种，通过单色（注意不要使用多种颜色）的饱和度突出数值的变化。表格注意淡化边框（窄边框或无边框）

- 散点图
  展示两件事情的关系，观察是否存在及何种关系，往往配合平均线或趋势线进行使用

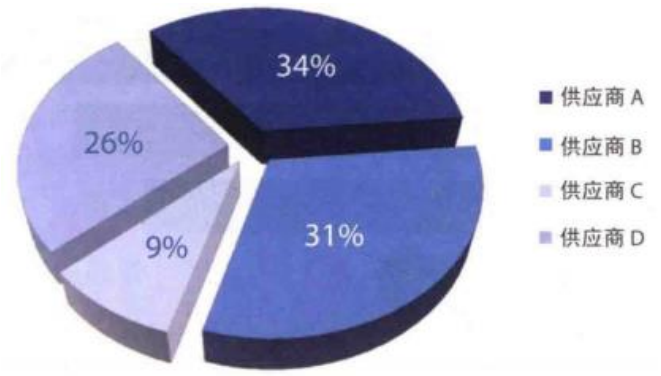
- 折线图
  暗示点之间存在离散数据（一系列数据分割成不同类别）间没有的联系，通常，连续性数据都以时间为单位

- 柱状图
  柱状图的应用广泛，可以代替饼图表示百分比，也可以直观地利用长短比较数据大小
- 斜率图
  适用于两个时间段或两组对比数据点，无需解释线的意义和具体变化是多少而直接展示数据的提升或下降以及变化速度（斜率），唯一的缺点是缺少绝对值

- 瀑布图
  抽离堆叠条形图的一部分进行重点关注，展示最终数据的组成或其中上升下降等变化

- 面积图
  比较极大的数值变化时更实用

注意，这里我们不推荐使用饼图及 3D 图形，理由是有更简单的可视化可以达到目的。人眼不擅长比较饼图的弧度（相对于长度）而 3D 图形使图表变得复杂和扭曲。



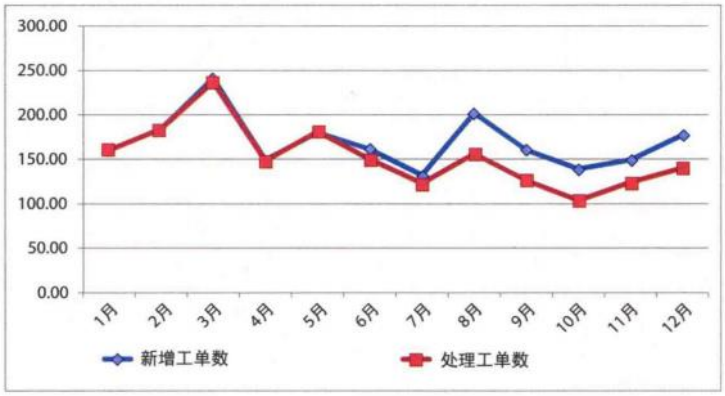
3.2.2 图表细节

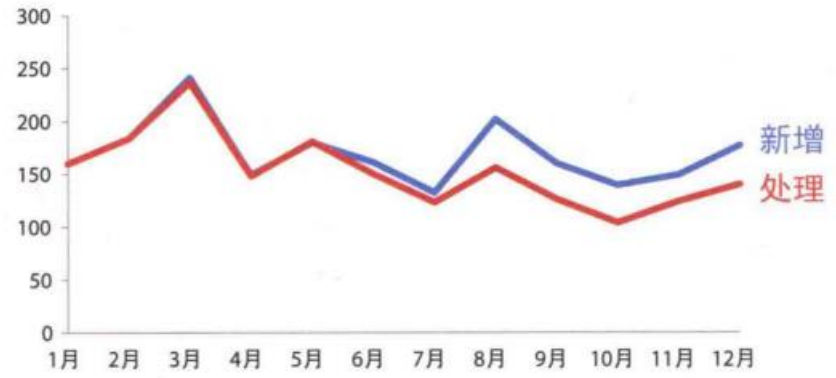
图表细节决定了最终可视化的质量。图表的杂乱消耗了受众的认知负荷，消除不必要的颜色，使用元素对齐，保持留白都是常用的策略。

可视化的一个重要原则是：少即是多。屏幕上保留的内容越少（精炼），受众得以消化和理解的内容越多。这项原则不仅在避免展示探索性分析时有效，对图表及仪表盘的设计同样有重要的指导意义。

比较初稿和终稿的优化过程：

- 取消边框及不必要的参考线或网格线
- 去除数据标记
- 清理坐标轴标签，改变缩进并调整文字的倾斜
- 直接标记数据，利用文字及颜色代替图标



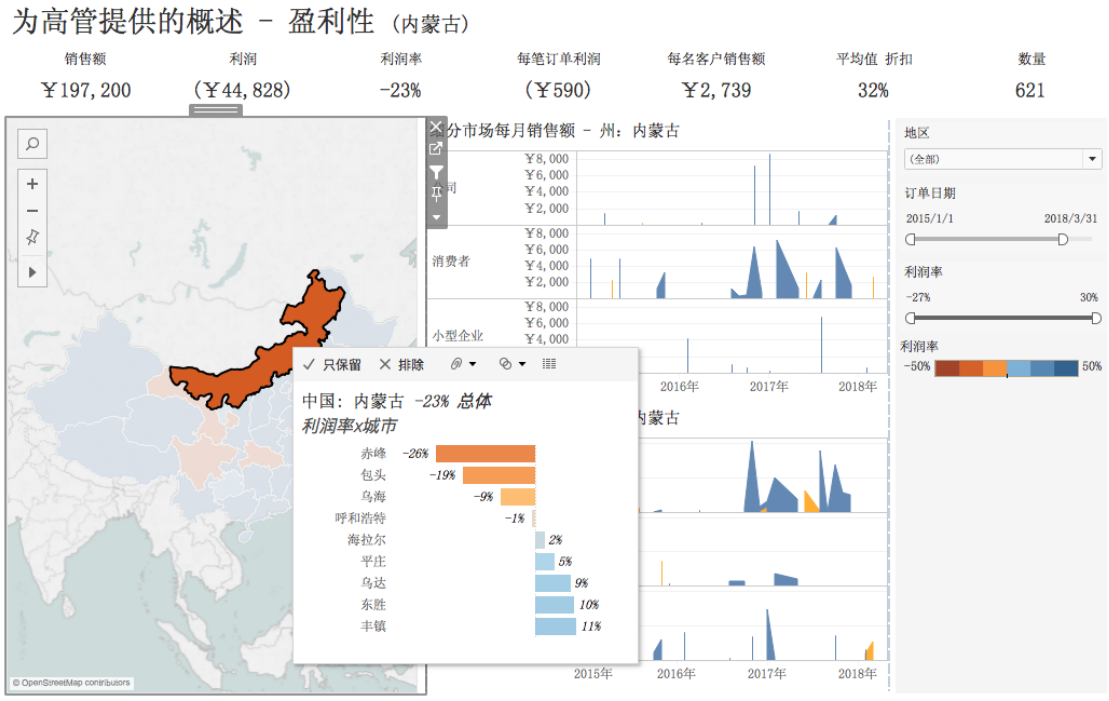


3.3 可视化交互，仪表盘操作

各种交互的添加在于加强表与表之间的联动性，使整个仪表盘更生动。

我们以 Desktop 2018.2 自带的超市示例工作簿（“概述”仪表板）为例。通过点击地图上的内蒙古，地图（省份）被作为整个仪表盘的筛选器，所有表格的输出限制为、为内蒙古。同时，工具提示中给出了（内蒙古的）城市利润率细节。

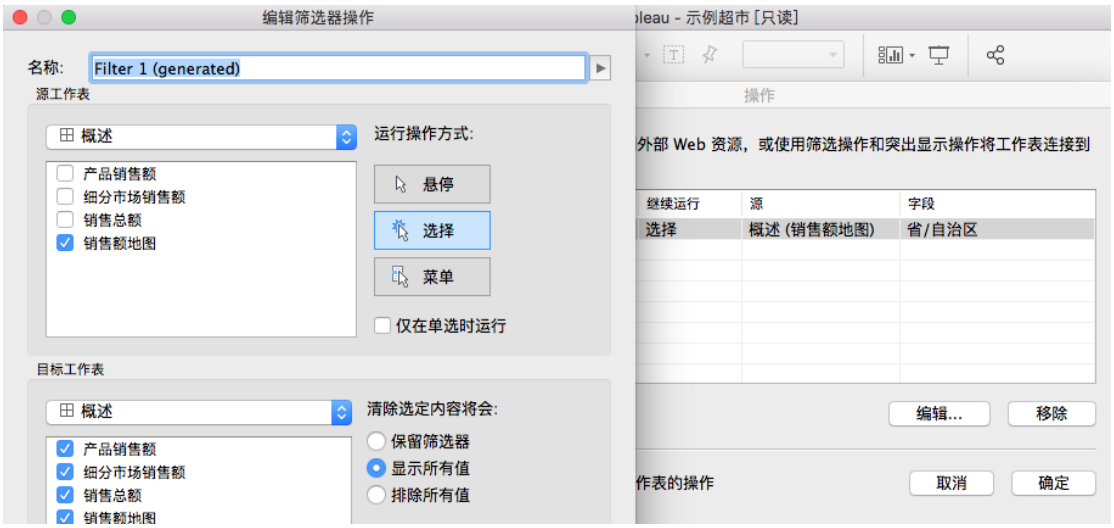
这是一个典型的仪表盘操作， 在展示省级别信息的过程中，地图代替了单独的“省份”筛选器，节约了空间并且引导受众对地图其他省份展开探索。



添加仪表盘操作的方法有两种：

- 直接在仪表盘上点击灰色区域内的漏斗形图标“用作筛选器”
- 依次点击，仪表盘-操作-添加操作

触发方式可以是悬停，点击或菜单；筛选的目标表格可以是同一个仪表盘，也可以是另一个仪表盘（进行页面跳转），可参考“超市”示例工作簿下更多仪表盘的操作设置。



3.4 聚合数据与计算字段

在进行可视化时，我们有时也需要对数据进行一定程度的处理，具有进行汇总或更改数据的粒度的能力。聚合函数不一定通过创建新的计算字段来完成，使用快速计算同样可以达到效果。

聚合计算的规则：

- 任何聚合计算中不得同时包括聚合值和解聚值。例如，SUM(Price)*[Items] 不是有效的表达式，因为 SUM(Price) 已聚合，而 Items 则没有。不过，SUM(Price*Items) 和 SUM(Price)*SUM(Items) 均有效。
- 表达式中的常量可根据情况充当聚合值或解聚值。例如：SUM(Price*7) 和 SUM(Price)*7 均为有效的表达式。
- 所有函数都可用聚合值进行计算。但是，任何给定函数的参数必须或者全部聚合，或者全部解聚。例如：MAX(SUM(Sales),Profit) 不是有效的表达式，因为 Sales 已聚合，而 Profit 则没有。不过，MAX(SUM(Sales),SUM(Profit)) 为有效的表达式。
- 聚合计算的结果始终为度量。

对于自定义计算字段这项中高级的技巧，以下是超出项目必须性的一些扩展。不同场景下遵循的编写原则稍有不同：

- 作品集：练习项目或作品计算字段不受任何约束，利用 Tableau 对数据源只读的特性你可以放心地尝试各种语句。Tableau 的代码语句编写界面十分友好，不仅给出语法和示例，对语句错误类型还会给予提示。

- 比赛：Tableau 的官方比赛及官方认证测试都遵循“快速”的部署原则，在限定的时间内你需要展示特定的结果，意味着你的所有代码编写一般来说限于一行以内（如果需要些大段的代码才能解决，则该解法不是最优的）。有时候，Tableau 希望你用快捷功能代替计算字段来求和或求平均值。但请注意，你需要深刻了解你所使用的数据源结构，及所求均数后的数字意义。

- 工作：业务逻辑往往超乎想象的复杂，而使计算字段的使用变得必须。有时甚至超出 Tableau 可处理的范围：你可能会被要求对某一个产品的销量进行重复统计（既算作 A 销售的也算作 B 销售的），而全公司的总销量不变。如果你的数据源只有 A 和 B 其中一人的记录，你是无法利用 Tableau 增加或减少一行数据的（增减列不受限制）。

常用的命令包括但不限于：

- ATTR(expression): 如果它的所有行都有一个值，则返回该表达式的值。否则返回星号。会忽略 Null 值。

- AVG(expression): 返回表达式中所有值的平均值。AVG 只能用于数字字段。会忽略 Null 值。

- CORR(expression 1, expression2): 返回两个表达式的皮尔森相关系数。 皮尔森相关系数衡量两个变量之间的线性关系。结果范围为 -1 至 +1（包括 -1 和 +1），其中 1 表示精确的正向线性关系，比如一个变量中的正向更改即表示另一个变量中对应量级的正向更改，0 表示方差之间没有线性关系，而 −1 表示精确的反向关系。

- COUNT(expression): 返回组中的项目数。不对 Null 值计数。

- COUNTD(expression): 返回组中不同项目的数量。不对 Null 值计数。

- MAX(expression): 返回表达式在所有记录中的最大值。如果表达式为字符串值，则此函数返回按字母顺序定义的最后一个值。

- MIN(expression): 返回表达式在所有记录中的最小值。如果表达式为字符串值，则此函数返回按字母顺序定义的第一个值。

- SUM(expression): 返回表达式中所有值的总计。SUM 只能用于数字字段。会忽略 Null 值。

- 详细级别表达式, 又成为 LOD，包括 FIXED, INCLUDE 和 EXCLUDE: 简单来说类似与 SQL 里的 WHERE 和 HAVING，我们在这里不做进一步的展开，更多关于 LOD 的教程可参见这里

3.5 撰写报告

本项目中要求你将完成项目的思路撰写为文字的报告。文字报告的提交必须以 PDF 或 MD 格式，报告（开始处）必须包含最初稿及最终稿的 Tableau Public 链接。

- 提交的 ZIP 打包文件中，如果包含了作品原文件，则注意保存为 TBLX 格式而非 TBL 格式

- 数据提取（HYPER 或 TDE）文件不是必须的
- 注意上传至 Public 的文件和 ZIP 中版本的一致性
- 隐藏/删除最终故事以外的所有标签页

文字报告中的 “设计” 部分需要被格外注意。在这个部分中，你需要描述每个图表的设计初衷，也就是 3.2 的部分；以及通过可视化得出的结论。

对于每个仪表盘，你则需要考虑及记录图表的组成及排版。仪表盘并不是多个或者所有图表的堆砌，为何包括或为何不包括你所做的某些图表需要给出理由。不同的图表组合成的仪表盘有着不同的用途，相同图表组合的不同空间位置也会对仪表盘造成影响。

你的“设计”部分并不是帮助审阅者或受众去理解你可视化作品的说明文档，而是帮助你自己梳理和优化设计的工具。你的可视化作品的最终版本需要在脱离你解说、注释的情况下独立地向受众传达和你“设计”部分一致的结论。反之，如果你的可视化结论不加以说明就无法被从视图中提炼，那么该可视化是不成功的。

“反馈”部分也是 Tableau 项目特色的考量部分之一。你需要随着作品版本的更新持续记录他人 （亲戚、朋友、同学或同事）的反馈，以及针对该反馈做出的修改记录。 在展示并征求他们反馈的时候，不要过多地解释你的设计思路和结论，你只需要告诉他们你想回答的问题，看看他们利用你的可视化可以得出哪些结论。

审阅者以外的反馈尤其关键，因为他们往往不是数据爱好者。让一个对数据不敏感的受众很明确地了解到你可视化的意图，比让一个数据分析师理解你的结论更困难。
你针对同一条反馈的修改记录可以不限于一条。反馈的内容有时是十分模糊的：“我不懂这个图什么意思”，“仪表盘有点乱”或者仅仅是“这个颜色不好看”。在你探索解决方案的过程中，持续记录修改方案是必须的，审阅者也可以根据你的记录给出更准确的修改建议。

当然，你也可以对记录的某一条反馈不加以修改但记录理由，毕竟你才是可视化的作者，最了解作品的人一定是你自己。

4 资源及使用

4.1 Tableau 使用教程

Tableau 的具体使用，非常推荐在开始项目前浏览一遍官方的视频教程（https://www.tableau.com/zh-cn/learn/training ）。磨刀不误砍柴工，这个投入绝对是值得的。

4.2 Tableau Public

Tableau Public Gallery （https://public.tableau.com/en-us/s/gallery ），上有大量持续更新的 Tableau 官方选出的特色作品。一部分作品的原文件可以被下载在本地打开，你可以学习他们仪表盘元素的构成、计算字段的使用、颜色的组成等。

对于你感兴趣的作者，你可以持续关注及查看他们所有的往期作品，绝大部分的 Tableau 可视化专家都会活跃于此。

需要注意的是，这些特色作品作为了解 Tableau 作品的多样性及可实现性很好，但是这些作品和实际工作中的需求相差非常远。以 Tableau 官方使用的 Tableau 报表为例，**所有商业中使用视图都以简单、明了以及实际为目的，非常的 “冷淡” 但要求数据 “精准”。**

4.3 Udacity Forum, Tableau Forum

Udacity Forum（中文版：http://discussions.youdaxue.com/ ），优达中国官方论坛，你可以按项目板块提出任何问题。当然提问前先搜索是否有已经回答的相似历史问题，并注意提问的问题描述，一般 48-72 小时内会有论坛导师为你解答。

Tableau Forum（英文：https://community.tableau.com/community/forums ），这里是 Tableau 的官方论坛，绝大部分你遇到的技术性问题都可以在这里找到答案。软件本身的 bug（如连接数据源报错）可以提交工单联系技术支持人员，一般 5-7 个工作日内 Tableau 官方支持会通过 WebEx 方式远程协助你解决。

4.4 外部博客

外部博客也是获取灵感的好地方，资源非常丰富，列举两个：

- 技巧类：举个栗子知乎专栏 （https://zhuanlan.zhihu.com/c_118876582 ）
- 视图类的可以参考：VizWiz （http://www.vizwiz.com/ ）

修订记录

日期	版本	修改人	修改原因
18 年 10 月 18 日	V1 <定稿>	Kylie	修改、定稿
18 年 10 月 9 日	V0.1 <未定稿>	Bill	初稿

感谢以下资深助教对本辅导资料的贡献

Bill Yu (虞振远)

审稿/修改负责人：Kylie（kylie@udacity.com）