1 信息数据可视化

1.1 绘图基础

绘图 (plot) 是数据可视化的基本工具, 它通过将数据点在二维或三维坐标系中以图形化的方式呈现, 展示数据的关系, 趋势和分布, 从而帮助用户理解数据的结构和模式.

从我们见过的各种图中可以得出两点重要结论: 首先, 绘图并没有明确的标准和范式, 不同的绘图方式适用于不同类型的数据和分析需求; 其次, 绘图的设计是一个迭代的过程, 需要根据数据的特点和用户的需求不断调整和优化, 不能一劳永逸地确定.

1.1.1 提升绘图质量的原则

尽管绘图方法需要依具体情况而定, 人们也已经提出了一些普遍的绘图原则可以被用作指导, 以提高绘图能传达出有用信息的可能性. 下面我们将介绍 William S. Cleveland 提出的提升绘图质量的两条路线和相应原则.

提升视觉效果 提升视觉效果的目标是提高图表的可读性, 使其更加清晰直观, 避免过于复杂的元素使用户感到疑惑和疲劳. 具体原则如下:

1. **减少杂乱, 突出数据.**
   绘图应聚焦于数据本身, 移除所有干扰或分散注意力的多余元素. 例如, 可以去掉不必要的网格线, 背景颜色或装饰性图案, 以确保数据点和趋势线清晰可见.
2. **使用视觉突出元素展示数据.**
   连接点的线条不应遮挡数据点, 确保数据点清晰可见. 使用颜色, 形状或大小等视觉属性来区分不同类别或组的数据点, 以增强图表的可读性和信息传达效果.
3. **使用简洁的参考线, 标签, 注释和图例**
   参考线应仅用于标示数据中的重要阈值, 并且应尽量简洁, 只在必要时使用. 避免过多的参考线, 标签或注释, 以免干扰数据本身的展示.
4. **确保叠加数据集中不同组的符号可分辨, 同组的符号在视觉上易于组合**
   应当确保不同组的数据点在视觉上易于区分, 而同组的数据点在视觉上易于组合.
5. **使用适当的比例刻度和数据边距**
   坐标轴需要清晰地界定数据区域. 适当的数据边距有助于突出图表中的核心内容, 避免数据被挤压在边缘.

提升理解效果 提升理解效果的目标是确保图表能够有效传达数据的含义和关系, 帮助用户理解数据背后的逻辑. 具体原则如下:

1. **提供清晰的解释和结论**
   图形是验证假设或传达结果的工具. 为了有效传达信息, 需要通过文字 (如注释或标题) 描述一切所需信

息. 文字重点应放在解释主要特点和阐述结论上.

2. **充分利用可用空间**
   尽量填充满绘图区域, 避免无意义的空白. 这有助于确保数据的可见性和图表的紧凑性.
3. **适当使用对数刻度**
   对数刻度适用于数据范围跨度较大的情况, 可以帮助更好地展示数据的分布和关系.
4. **使用合适的纵横比**
   优化图表的纵横比使得图形中的线条方向更加清晰可辨. 通常, 45° 的倾斜角有助于提高图表的可读性.
5. **确保图标对齐且比例一致**对齐并列图表时, 确保它们的比例一致且对齐, 一遍用户能方便地对图标进行对比分析, 避免视觉不一致造成的混淆.

1.1.2 基础绘图技术

折线图 **折线图 (line plot)** 也被称作**符号连接图 (connected symbol plot)**, 它通过连接数据点的线条来展示数据的变化趋势和关系. 折线图适用于展示时间序列数据或连续变量之间的关系, 适用于识别数据中的模式和趋势.

点图 **点图 (dot plot)** 是一种通过点的分布来展示数据的图表类型, 和柱状图的原理类似. 点图适用于展示分类数据或离散变量之间的关系, 适用于比较不同类别的数据分布和频率.

散点图 **散点图 (scatter plot)** 是一种通过点的分布来展示两个变量之间关系的图表类型. 散点图适用于需要观察数据之间的分布与关联度的场合, 可以用于直观地识别集群, 离群点和线性/非线性关系.

1.2 高维数据的可视化

1.2.1 多维尺度法

为了显示包含在高维数据中的距离信息, 我们可以采用**多维尺度法 (Multidimensional Scaling, MDS)** 的降维形式, 通过线性/非线性的映射将高维数据映射到低维空间, 同时尽量保持数据点之间的距离关系.

在降维过程中, 由于原高维空间的复杂性, 数据的距离关系无法完全被保留或准确反映, 甚至可能出现冲突.

1.2.2 平行坐标法

经典可视化中假设不同维度的坐标轴相互正交, 因此人们能够直观理解并可视化的正交坐标系最多不超过三维. 当数据维度超过了三维后, 基于正交坐标系的可视化技术就失效了, 于是人们考虑放弃“正交”约束, 改将各维度坐标轴画成平面内的一系列平行线, 每个数据点的各维度分量对应有各坐标轴的取值, 相连起来可形成一条折线. 这种方法被称为**平行坐标法 (Parallel Coordinates)**.

平行坐标空间和数据所处的原有的笛卡尔空间有着有趣的对应关系. 在笛卡尔空间中正相关的数据构成

的直线对应于平行坐标空间中的一组不相交的折线; 而在笛卡尔空间中负相关的数据构成的直线对应于平行坐标空间中的一组相交的折线. 此外, 笛卡尔空间中的圆和椭圆映射成平行坐标空间中的双曲线; 笛卡尔空间中的旋转对应于平行坐标空间的平移, 反之亦然; 笛卡尔空间中的拐点对应于平行坐标空间中的顶点.

1.3 文字可视化

文本可视化通常涉及将文本数据映射到图形, 图表或其他视觉元素上, 以便直观地呈现文本的特征和结构. 不同的文本可视化技术可以根据数据和分析任务的需求选择和定制. 以下是一些常见的例子.

1.3.1 词云

1.3.2 TIARA

1.4 图可视化

定义 1.1 图

简单而言, 图是由顶点和连接顶点的边构成的数学结构, 通常可以记作 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, 其中 $\mathcal{V}$ 和 $\mathcal{E}$ 分别是顶点集合和边集合.

根据边是否有确定的指向, 可以将图分为有向图和无向图.

通常, 组织架构, 文件目录, 社交网络等数据具有图的形式. 图可视化的目的是帮助分析人员理解图中的结构特征, 节点之间的关系以及信息传播的路径.

1.4.1 力导向布局

力导向布局 (Force-Directed Layout) 是一种常用于图数据可视化的布局算法, 通过模拟物理力学中的相互作用力来优化节点位置.

力导向布局的主要原理是维持节点之间的斥力和边连接的节点之间的引力的平衡. 这可以通过将边抽象成具有一定原长的弹簧来实现[1]. 我们可以按照物理仿真中的弹簧质点系统的模拟方法对图进行可视化.

[1]如果节点之间靠得太近, 距离小于弹簧原长, 它们之间将远离, 这就模拟了节点之间的斥力; 如果节点之间离得太远, 距离大于弹簧原长, 它们之间将靠近, 这就模拟了节点之间的引力.