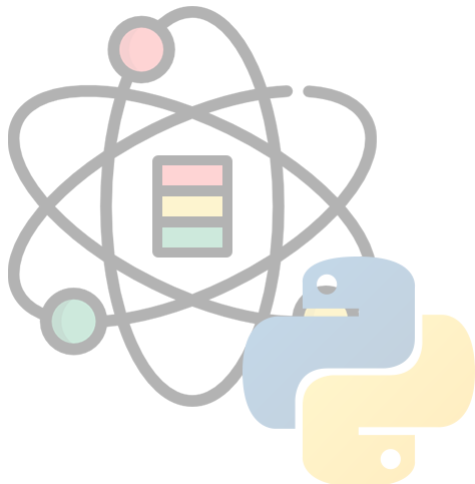
Python 数据科学导论
Data Science Introduction with Python

数据可视化
Data Visualization

范叶亮

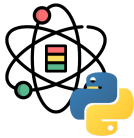


目录

- 数据可视化
- Matplotlib & Seaborn
- plotnine
- 基于 Web 的绘图库

数据可视化

数据可视化



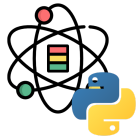
数据可视化（Data Visualization） [1] 被许多学科视为与视觉传达含义相同的现代概念，它涉及到数据的可视化表示的创建和研究，数据可视化既是一门**艺术**也是一门**科学**。

为了清晰有效地传递信息，数据可视化使用统计图形、图表、信息图表和其他工具。可以使用点、线或条对数字数据进行编码，以便在视觉上传达定量信息。 有效的可视化可以帮助用户分析和推理数据和证据。它使复杂的数据更容易理解和使用。用户可能有特定的分析任务（如进行比较或理解因果关系），以及该任务要遵循的图形设计原则。表格通常用于用户查找特定的度量，而各种类型的图表用于显示一个或多个变量的数据中的模式或关系。

数据可视化主要旨在借助于图形化手段，清晰有效地传达与沟通信息。但是，这并不就意味着，数据可视化就一定因为要实现其功能用途而令人感到枯燥乏味，或者是为了看上去绚丽多彩而显得极端复杂。为了有效地传达思想概念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，从而实现对于相当稀疏而又复杂的数据集的深入洞察。然而，设计人员往往并不能很好地把握设计与功能之间的平衡，从而创造出华而不实的数据可视化形式，无法达到其主要目的，也就是传达与沟通信息。

[1] https://zh.wikipedia.org/wiki/数据可视化

数据可视化



一图胜千言

One look is worth a thousand words.

A picture is worth a thousand words.

One Look Is Worth A Thousand Words--

One look at our line of Republic, Firestone, Miller and United States tires can tell you more than a hundred personal letters or advertisements.

WE WILL PROVE THEIR VALUE BEFORE YOU INVEST ONE DOLLAR IN THEM.

Ever consider buying Supplies from a catalog?

What's the use! Call and see what you are buying. One look at our display of automobile and motorcycle accessories will convince you of the fact.

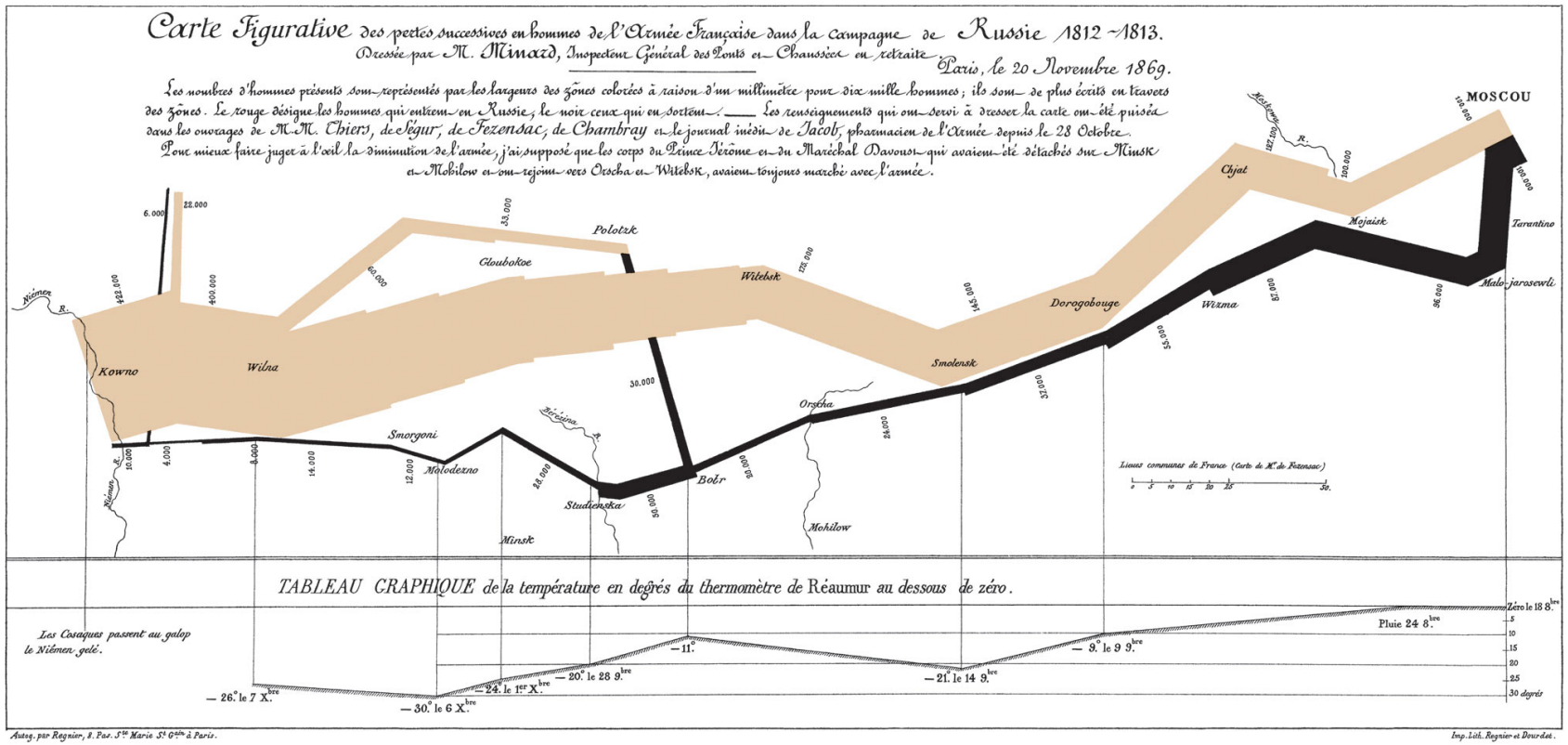
THAT WE HAVE EVERYTHING FOR THE AUTO

Piqua Auto Supply House

133 N. Main St.—Piqua, O.

[1] https://zh.wikipedia.org/wiki/一畫勝千言

数据可视化

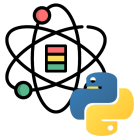


法国工程师查尔斯·约瑟夫·密纳德于 1861 年绘制的关于拿破仑帝国入侵俄罗斯的信息图

数据可视化

图片来源：https://cloud.baidu.com/product/sugar.html

The Grammar Of Graphics



Wilkinson [1] 创建了一套用来描述所有统计图形深层特性的语法规则，该语法回答了“什么是统计图形”这一个问题。一张统计图有如下独立的图形部件所组成 [2]:

- 最基础的部分是你想要可视化的**数据（data）**以及一系列数据中的变量对应到图形属性的**映射（mapping）**；
- **几何对象（geom）**代表你在图中实际看到的图形元素，如点、线、多边形等；
- **统计变换（stats）**是对数据进行的某种汇总。例如：将数据分组技术以创建直方图，或将一个二维的关系利用线性模型进行解释。统计变换是可选的，但通常非常有用；
- **标度（scale）**的作用是将数据的取值映射到图形空间，例如用颜色，大小或形状来表示不同的取值。展现标度的常见做法是绘制图例和坐标轴，它们实际上是从图形到数据的一个映射，使读者可以从图形中读取原始的数据。
- **坐标系（coord）**描述了数据是如何映射到图形所在的平面的，它同时提供了看图所需的坐标轴和网格线。我们通常使用的是笛卡尔座标系，但也可以将其变换为其他类型，如极坐标和地图投影。
- **分面（facet）**描述了如何将数据分解为各个子集，以及如何对子集作图并联合进行展示。分面可叫做条件作图或网格作图。

[1] Wilkinson, Leland. "The grammar of graphics." *Handbook of Computational Statistics*. Springer, Berlin, Heidelberg, 2012. 375-414.
[2] Wickham, Hadley. *ggplot2: elegant graphics for data analysis*. Springer, 2016.

The Grammar Of Graphics

- R 实现

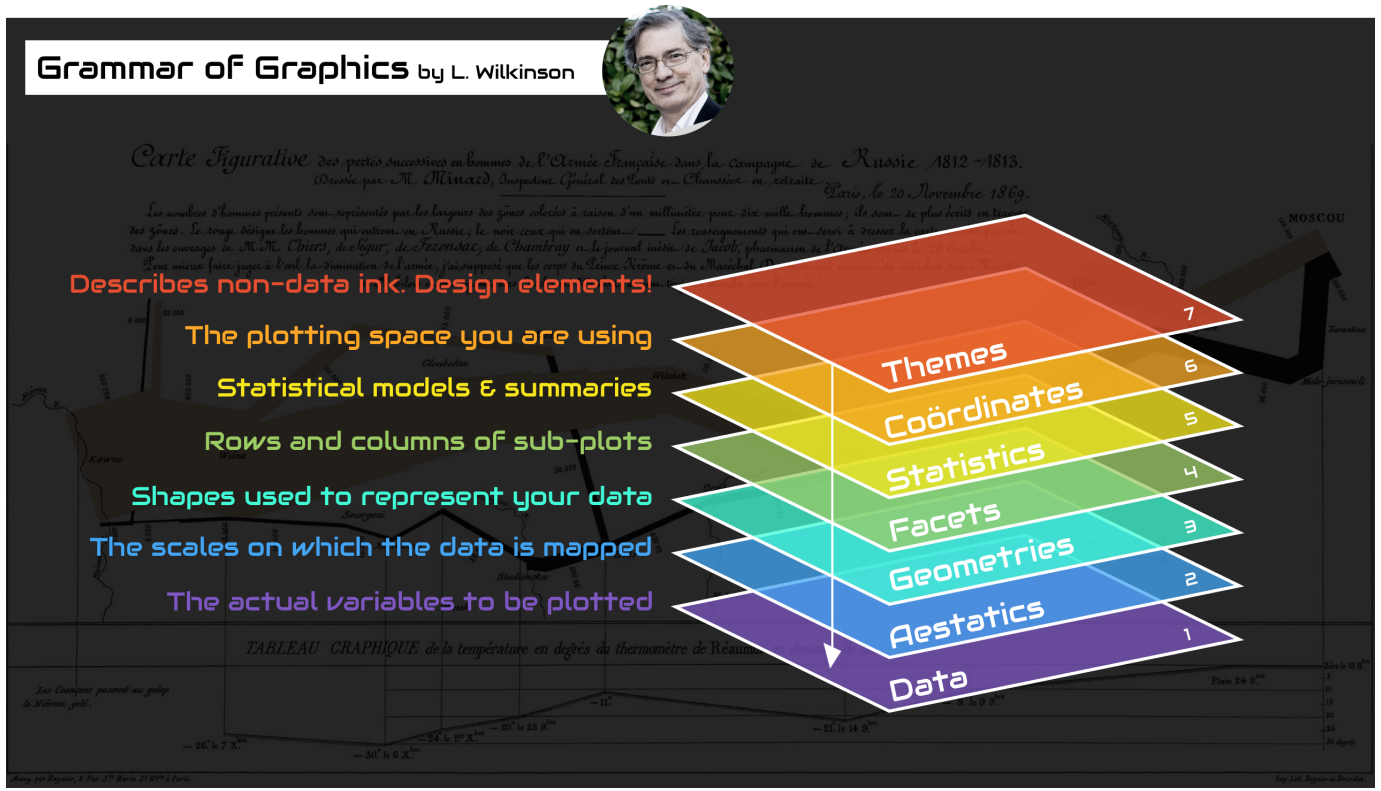
ggplot2

- Python 实现

plotnine

- JavaScript 实现

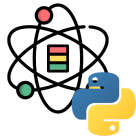
AntV | G2



图片来源：https://medium.com/@TdeBeus

Matplotlib & Seaborn

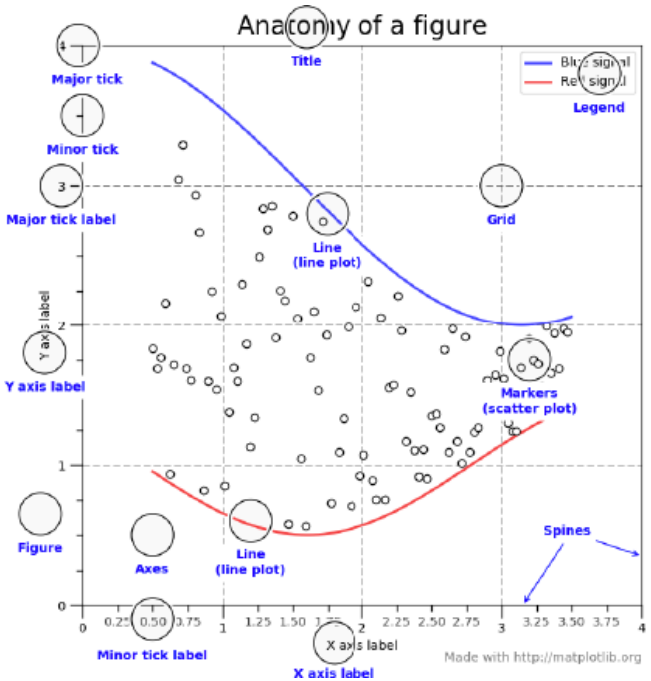
Matplotlib



Matplotlib [1] 是 Python 的一个 2D 绘图库，其可以生成多种格式的出版物质量的图形并提供跨平台的交互环境。Matplotlib 可用于 Python 脚本，Python 和 IPython 命令行环境，Jupyter Notebook，Web 应用和多种用户交互工具包中。

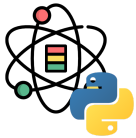
`pyplot` 提供了以类似与 MATLAB 交互的简易绘图接口，尤其是当用户在使用 IPython 时。对于高级用户来说，可以通过类似于 MATLAB 的一系列函数控制线型，字体，轴等属性。

matplotlib



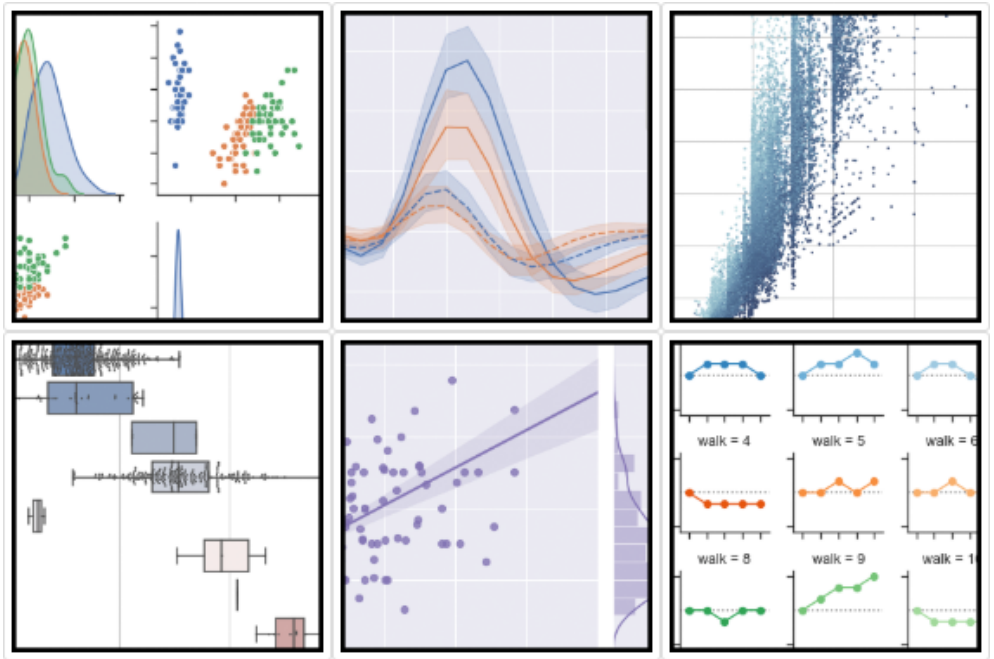
[1] 官网：https://matplotlib.org/
[2] Cheatsheet：https://github.com/matplotlib/cheatsheets

Seaborn



Seaborn [1] 是一个统计绘图库，它构建在 Matplotlib 基础之上，同时适用于 Pandas 数据结构。

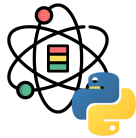
Seaborn 旨在将数据可视化作为数据理解和探索的核心。其以数据集为导向的绘图函数可以操作 Dataframe 和 Array，通过必要的语义映射和统计汇总生成提供有用信息的图形。



[1] 官网：https://seaborn.pydata.org/
[2] Cheatsheet：https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Python_Seaborn_Cheat_Sheet.pdf

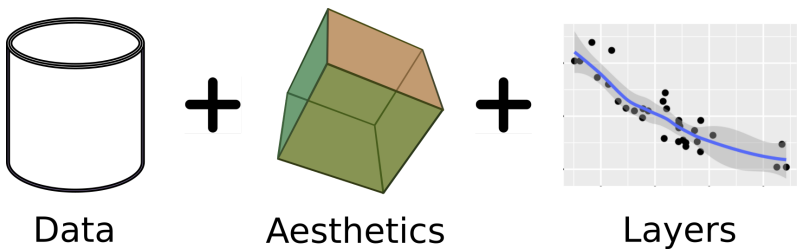
plotnine

plotnine



plotnine [1] 是 Grammar of Graphic 的 Python 实现，其基于 ggplot2 [2] 构建。图形语法允许用户通过显式地将数据映射到可视化的对象来生成图形。

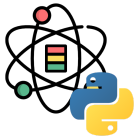
利用图形语法绘图是很强大的，它可以让使用者轻松地思考并构建自定义的图形。



[1] 官网：https://plotnine.readthedocs.io/
[2] ggplot2：https://ggplot2.tidyverse.org/
[3] Cheatsheet：https://github.com/EasyChart/Beautiful-Visualization-with-python/blob/master/Plotnine学习手册.pdf

基于 Web 的绘图库

Bokeh

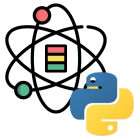


Bokeh [1] 是一个用于现在 Web 浏览器的交互式可视化库。它提供了一个用于大型或流式数据集的优雅简明的可视化方案。Bokeh 可以帮助大家轻松快速地构建交互式的绘图，仪表盘和数据应用。

[1] 官网：https://docs.bokeh.org/

[2] Cheatsheet：https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Python_Bokeh_Cheat_Sheet.pdf

Plotly



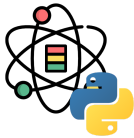
Plotly 的 Python 库（plotly.py）[1] 是一个开源的交互式绘图库，它支持 40 多种不同类型的图表，涵盖了统计、金融、地理、科学和 3 维等多种用例。

plotly.py 建立在 Plotly JavaScript 库的基础之上，这使得 Python 用户可以构建精美的交互式可视化，这些可视化效果可以显示在 Jupyter Notebook，保存至独立的 HTML 文件，或者使用 Python 构建的 Web 应用中。

通过深度集成 `orca` 图像导出工具，plotly.py 还提供了强大的非 Web 环境支持，包括：桌面编辑器（例如：QtConsole，Spyder，PyCharm）和静态文档发布（例如：导出 Notebook 至带有高质量矢量图片的 PDF）。

[1] 官网：https://plot.ly/python/
[2] Cheatsheet：https://images.plot.ly/plotly-documentation/images/python_cheat_sheet.pdf

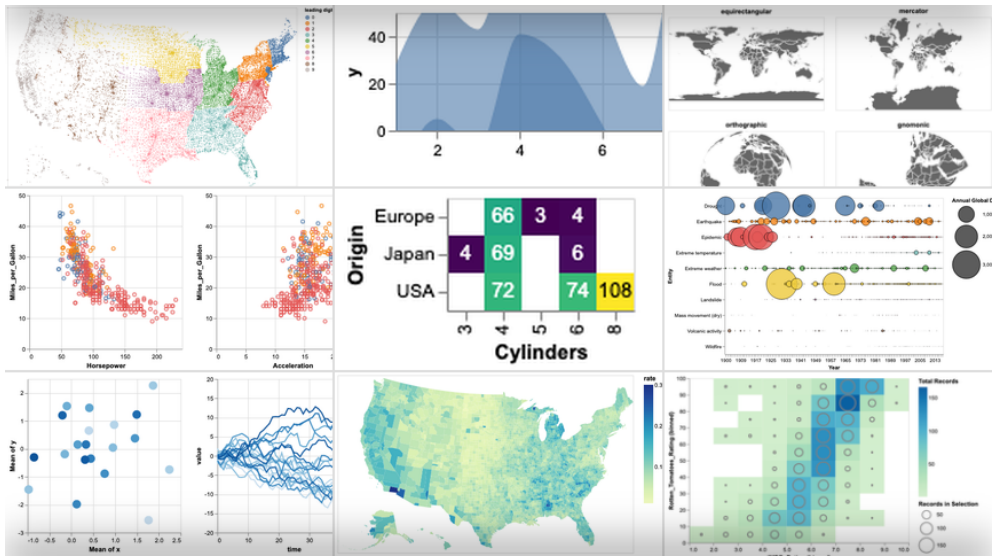
Altair



Altair [1] 是一个 Python 的 declarative statistical 可视化库，其基于 Vega 和 Vega-Lite。

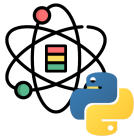
利用 Altair，你可以有更多的时间去理解你的数据及其含义。Altair 的 API 简单，友好，一致，并构建于强大的 Vega-Lite 可视化语法之上。这种优雅的简单性使得利用最少的代码就可以产生了漂亮而有效的可视化效果。

Altair 是由 Jake Vanderplas 和 Brian Granger 以及 UW Interactive Data Lab 密切合作开发。



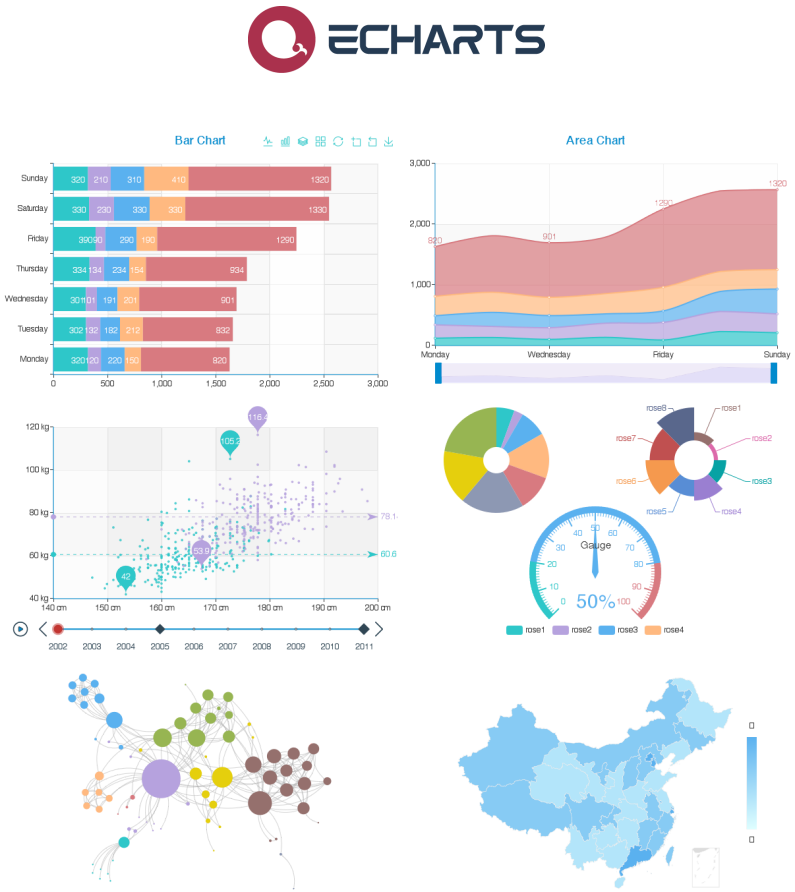
[1] 官网：https://altair-viz.github.io/

Echarts & pyecharts



ECharts [1] 是一个使用 JavaScript 实现的开源可视化库，可以流畅的运行在 PC 和移动设备上，兼容当前绝大部分浏览器（IE8/9/10/11，Chrome，Firefox，Safari 等），底层依赖矢量图形库 ZRender，提供直观，交互丰富，可高度个性化定制的数据可视化图表。

ECharts 提供了常规的折线图、柱状图、散点图、饼图、K 线图，用于统计的盒形图，用于地理数据可视化的地图、热力图、线图，用于关系数据可视化的关系图、treemap、旭日图，多维数据可视化的平行坐标，还有用于 BI 的漏斗图，仪表盘，并且支持图与图之间的混搭。pyecharts [2] 是 Echarts 的 Python 绑定。



[1] Echarts 官网：https://echarts.apache.org/
[2] pyecharts 官网：https://pyecharts.org/

感谢倾听