数据集选择

从以下数据集中选择一个或自己寻找数据集。更多关于寻找数据集的资源见本文档底部。

难度级别	数据集	概述	备注
初级	泰坦尼克号数据	包含泰坦尼克上 2224 名乘客和机组人员的人口统计和乘客信息。关于此数据集的更多信息可在此处查找。	创建一个可视化，展示生还和死亡乘客的人口统计和乘客信息。
初级	棒球数据	一个包含 1,157 名棒球手的数据集，包括他们的用手习惯（左手还是右手）、身高（英寸）、体重（磅）、击球率和全垒得分。	创建一个可视化，展示这些球手的表现差异。
中级	航班	此数据集包含来自 RITA 的美国航班延误和性能信息。你可以直接从 RITA 下载此数据或从“航班”(Flights) 链接下载压缩 csv 文件。 “航班”(Flights) 链接上的文件按年组织，比原文件压缩的更小。关于该数据的更多详情可在此处查找。	调查航班随时间推移的性能或只是查看特定年的数据，然后创建一个图表来展示你的发现。
中级	来自 Proper 的贷款数据 最后一次更新 2014/11/03 此数据字典解释了数据集中的变量。	此数据集包含 113,937 项贷款，每项贷款有 81 个变量，包括贷款金额、借款利率（或利率）、当前贷款状态、借款人收入和很多其他项。	自己对此数据集提问，找出数据中有趣的趋势。

高级	PISA 数据 PISA 数据字典 注：未压缩的 PISA Data csv 文件为 2.75 GB。	国际学生评估项目 (PISA) 是一项对接近完成义务教育的学生进行的技能和知识评估。它并非传统的学校考试，主要目标不是评估学生对学校课程的掌握情况，而是评定其离开学校参与社会的全面素养。 在 2012 年，来自 65 个经济体的约 510,000 名学生代表全球 2800 万名 15 岁学生，参加了 PISA 的阅读、数学和科学评估。在这些经济体中，有 44 个参加了创新问题解决能力评估，18 个参加了理财素养评估。 此数据和调查话题来自 PISA 数据可视化竞赛。在此参阅获胜者和之前的提交，获取更多灵感与示例。	考虑创建一个图表来探索以下话题中的一个。 学校因素在解释学业成绩方面的重要性。 基于性别、位置或学生态度的成绩差异。 基于教师实践和态度的成绩差异。 学业成绩的不均等。
其它 取决于你自己在数据处理方面的经验。	**自己寻找数据集！**	**记住一点，自行寻找和清理数据集会耗费大量时间和精力！** 如果想自行寻找数据集，请参阅下方的检查清单。	自己提出问题并查找数据来解答它。或者，你也可以找到一个数据集，然后提出关于它的问题，直到你找到一些想要分享的有趣发现。

如果你要自己寻找数据集……（见下页）

你最终提交的数据集应该：

- 格式整洁[1]（你可能需要清理和重整数据）
- 采用 dimple.js 或 d3.js 加载数据的常用格式，例如 .csv、.tsv、.txt、.json、.xml 或 .html

以下是寻找数据集的一些资源：

- http://www.pewglobal.org/category/datasets/
- http://databank.worldbank.org/data/home.aspx
- http://www.data.gov/
- http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public
- http://www.inside-r.org/howto/finding-data-internet
- https://www.edsurge.com/n/2014-01-21-education-datapalooza
- 1,001 Data Sets

注释：

1. 整洁数据集是具有特定结构的数据集。可参阅 Hadley Wickham 的论文了解更多关于整洁数据的信息：
http://vita.had.co.nz/papers/tidy-data.pdf

2. 点此下载此文档的英文版本。