数据集选择

## 从以下数据集中选择一个或自己寻找数据集（如果要自己寻找请见下文说明）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | | 概述 | 引导问题 | 时间估计 |
| 红葡萄酒质量 | **1**  文件，它描 数据的收集 | 这个整齐的数据集包含 1,599  种红酒，以及 11 个关于酒的化  学成分的变量。至少 3 名葡萄酒  专家对每种酒的质量进行了评  分，分数在 0（非常差）和 10  （非常好）之间。 | 哪个化学成分影响红葡萄 酒的质量？ | 10 至 20 个  小时 |
| 阅读此文本 述了变量及 方式。 |
| 白葡萄酒质量  阅读此文本文件，它描  述了变量及数据的收集  方式。 | | 这个整齐的数据集包含 4,898  种白葡萄酒，及 11 个量化每  种酒化学成分的变量。至少 3  名葡萄酒专家对每种酒的质量  进行了评分，分数在 0（非常  差）和 10（非常好）之间。 | 哪个化学成分影响白葡萄  酒的质量？ | 10 至 20 个  小时 |
| 国家总统竞选的财政捐  助 | | 从单选按钮中选择一个选举年，  然后按下“导出捐款者数据”按  钮获取可下载的数据集。选择一  个国家然后探索在特定选举年向  总统候选人的财政捐助 | 自己对此数据集提问。你  可以向此数据集添加变  量，例如候选人的性别或  政党。 | 15 至 30 个  小时 |
| 来自 **Prosper** 的贷款数 据  最后一次更新：  2014/11/03  此变量字典解释了数据  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、借  款人就业状态、借款人信用历史  及最新支付信息。 | 自己对此数据集提问。此  数据集中变量众多，你不  可能对它们全部进行探  索。你应该在分析中探索  10 到 15 个变量。 | 15 至 30 个  小时 |
| 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。 | 对数据集提出自己的问  题！ | 30 个小时以  上 |

1 P. Cortez, A. Cerdeira, F.Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236. Available at: [@Elsevier] <http://dx.doi.org/10.1016/j.dss.2009.05.016>[Pre-press (pdf)] <http://www3.dsi.uminho.pt/pcortez/winequalitv09.pdf>[bib] <http://www3.dsi.uminho.pt/pcortez/dss09.bib>

# 如果你在自己寻找数据集……

你的数据集应包括：

□ 至少 1,000 个观察结果

□ 至少包含一个类别变量（你可以创建一个）

□ 至少包含 8 个不同变量

□ 格式整洁 1（你可能需要清理和重整数据）

□ 数据集应采用常用格式，如 .csv、.tsv、.txt 或 .xls

以下是一些寻找数据集的资源：

## http://www.inside-r. org/howto/finding-data-internet（不要使用泰坦尼克数据集）

* <http://opendata.stackexchange.com/>
* <http://www.data.gov/>

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>