一、Data Wranging数据整理

数据整理可归纳为以下三步：

1. 数据收集（Gather）
2. 数据评估（Assess）
3. 数据清理（Clean）

**1.数据收集（Gather）**

收集数据的方式有很多，最简单、最常见的是下载现成的数据。比如从kaggle上下载数据集。

但为了可扩展性（Scalability）和重复性（Reproducibility），有时需要以编程的（Programmatically）方式下载。比如需要下载的文件量较大，有成百上千个，甚至位于不同页面。

从网上爬取数据。比如爬知乎，爬豆瓣。

从各种API获得数据。比如电影数据API，股票数据API，Twitter数据API，等等。

**2.数据评估（Assess）**

数据评估可以从两方面进行：质量（quality），整洁度（tidiness）

**质量（Quality）**

低质量的数据常被称为**脏数据（dirty data），**比如：

-数据丢失，缺值

-数据无效

-数据不准确

-数据不一致，比如使用不同的长度单位

**整洁度（Tidiness）**

不整洁数据被称为杂乱数据（messy data），

A dataset is messy or tidy depending on how rows, columns, and tables are matched up with observations, variables, and types. In tidy data:

Each variable forms a column.

Each observation forms a row.

Each type of observational unit forms a table.

**3.数据清理（Clean）**

分为手工清理和程序清理。

**程序清理：**

Define: convert our assessments into defined cleaning tasks. These definitions also serve as an instruction list so others (or yourself in the future) can look at your work and reproduce it.

Code: convert those definitions to code and run that code.

Test: test your dataset, visually or with code, to make sure your cleaning operations worked.

Always make copies of the original pieces of data before cleaning!

**Reassess and Iterate（重新评估和迭代）**

After cleaning, always reassess and iterate on any of the data wrangling steps if necessary.

**Store (Optional 可选)**

Store data, in a file or database for example, if you need to use it in the future.

**参考文档：**

1. Kaggle是国外著名的数据竞赛分析平台，对新手十分友好，这里主要记录我在这个小挑战中学到一些数据清洗方法。

kaggle链接：<https://www.kaggle.com>

1. python数据清洗的例子： <https://blog.csdn.net/weixin_40683253/article/details/81665896>
2. 关于数据清洗的步骤与方法理解https://blog.csdn.net/wyqwilliam/article/details/84801095