复习与提示
DataFrame
关系代数制图
宽表到长表
R 版总评计算
dbplyr
后备资料

DataFrame

续本达

清华大学 工程物理系

2023-08-01 清华

安装软件 R 的 ggplot2 与 dbplyr

Debian

```
apt install r-cran-ggplot2 r-cran-dbplyr r-cran-dplyr r-cran-reshape2 r-cran-rsqlite
```

最佳工具

- Dataframe 流派的关系代数系统 全局最佳工具是 GNU R
  - Python 只是小圈子的最佳工具
- Grammar of graphics 的全局最佳工具是 R ggplot2
  - seaborn 只是眼前的妥协，未来 Python 小圈子的最佳工具可能是 plotnine

关系代数复习

- 集合运算：$\cap, \cup, \setminus$
  - SQL `INTERSECT` `UNION` `EXCEPT`
- 线性运算：$\otimes, \Pi, \sigma$
  - SQL `SELECT` `WHERE`
- 关系运算：连接 $\bowtie$
  - SQL `JOIN`
- 拓展运算：GroupBy（分组） $\mathcal{G}$
  - SQL `GROUP BY`

[66%] 少年爱迪生想

- ☒ 比较物理系和工物系的男女比例
- ☒ 算出大家的总评成绩：
  - ☒ 小作业权相等，总体占 65% 的成绩
  - ☒ 大作业占 30% 的成绩
  - ☒ 划分出不同的分数段，给出某分数段同学的手机号
- ☐ 画出各班平均小作业成绩的变化曲线

```
SELECT 学号, sum(权重*分数) AS 总评 FROM
(SELECT 学号, CASE 大作业
  WHEN 0 THEN 0.65
  WHEN 1 THEN 0.3
  END AS 权重, avg(分数) AS 分数
  FROM longscores GROUP BY 学号, 大作业)
GROUP BY 学号
HAVING 总评 > 105
```

学号	总评
49	105.16375
55	105.70625

DataFrame

由 R 语言提出的数据表格形式：与 SQL 一样都脱胎于关系代数思想

SQL 的区别

- SQL 是描述性语言，与函数式编程更接近。
  - 生成的表格不能随意改动。改动只是通过创建新表格实现。
- DataFrame 是动态结构
  - 可以更灵活地更改；
  - 但过于灵活可能带来调试上的麻烦。
- SQL 可以处理大于内存容量的数据， DataFrame 必须借助 Spark 等大数据平台才能实现。

GNU R 语言

- R 是著名的统计学语言，应用广泛
- 起源于 1970 年代的 S 语言
  - 在 fortran 库的基础上交互地处理数据进行统计分析
    调用 fortran 库提供交互环境是 S, Matlab 和 Numpy 的共同起源
  - 提供交互式的绘图工具，以便于快速迭代
  - 受 LISP 语言的影响
  - DataFrame 概念影响了几乎所有统计工具
    例如 S-PLUS, SAS, SPSS 应用广泛
- GNU R 重新使用 Scheme 实现了 S 语言，并以自由软件形式发布
  - 自由软件和开源运动是为科学研究提供了可 复现 的工具

为什么使用 R

- 大部分统计学算法都有 R 语言的实现
  站在数理统计学天才们的肩膀上。
- 非常适合快速试验并找到统计方法上的 最佳工具

`dplyr` 数据整理的利器，简洁的语法表达关系代数

`dbplyr` 与 SQL 浑然连成一体，统一 DataFrame 与 SQL

`ggplot2` 科学制图利器，易用性远好于 matplotlib

数据类型

呼叫传送门

numeric 数值型：分成整型、浮点型和复数型

character 字符型

factor 枚举型

```
# 获得帮助
help(help)
help(numeric)
```

数值型

- 数组
- 数组的索引
- 字符型
- 字符串操作
- 布尔型
- 枚举型

程序结构

选择结构

循环结构

函数

各类概率分布

- 正态分布

DataFrame 用法

- 简单制图
- 线性回归
- 枚举类型

DataFrame 里的关系代数基本操作

- 集合基本运算：交、并、差
- 线性代数运算：笛卡尔积、投影、选择
- 连接
- GroupBy, SortBy

图形语法 Grammar of Graphics

- 在表格与图形要素之间建立映射关系
- 表格对称的同等层次的列，都可以映射到各类图形要素中
  1. x 轴坐标
  2. y 轴坐标
  3. 颜色
  4. 点形状（圈、三角、方块等）
  5. 线形状（实线、虚线、点划线等）
  6. 点的大小
  7. 线的粗细
  8. 透明度
  9. 子图位置
- 在绘图时，通过调整映射关系来找到最佳的展现方式

参考书

- Leland Wilkinson, The Grammar of Graphics （沉痛缅怀 Wilkinson 教授）
- Hadley Wickham, A Layered Grammar of Graphics

ggplot

呼叫传送门 Merge-GroupBy.slides.html

```
require(ggplot2)
data(mtcars)
p <- ggplot(mtcars, aes(x=wt, y=hp)) + geom_point()
print(p)
```

```
p <- p + aes(color=am)
print(p)
```

ggplot 枚举型制图

```
mtcars$am <- factor(mtcars$am, levels=c(0, 1), labels=c("automatic", "manual"))
p <- p %+% mtcars
print(p)
```

表格的等价变换

呼叫传送门 Merge-GroupBy.slides.html

- R 和 Pandas DataFrame 可以将长表与宽表相互转换
  - SQLite 不具备此功能
- melt 函数

melt

- `id_vars` 指定作为索引的列
- `value_vars` 指定作为数值的列
- `var_name` 指定长表中组变量的列名
- `value_name` 指定长表中值的列名

统计班级平均分

[100%] 少年爱迪生想

- ☒ 比较物理系和工物系的男女比例
- ☒ 算出大家的总评成绩：
  - ☒ 小作业权相等，总体占 65% 的成绩
  - ☒ 大作业占 30% 的成绩
  - ☒ 划分出不同的分数段，给出某分数段同学的手机号
- ☒ 画出各班平均小作业成绩的变化曲线

课堂练习

SQL 的 DataFrame 接口

```
library(dplyr)
library(dbplyr)
con <- DBI::dbConnect(RSQLite::SQLite(), dbname = "dataframe-practice/people.db")
classes <- tbl(con, "classes")

classes %>% filter(院系=='物理')
```

基科 71　物理
物理 72　物理
物理 71　物理
基科 72　物理

查看 SQL 语句

```
sql_render(classes %>% filter(院系=='物理'))
```

```
<SQL> SELECT *
FROM `classes`
WHERE (`院系` = '物理')
```

关系代数

- 关系：$\{(r,s)|r \in R, s \in S\}$
- 关系代数：在集合基础上定义关系运算的封闭系统
  - 封闭系统：运算作用于一个或多个关系上来生成一个关系
- 围绕关系代数设计的关系数据库是存储海量数据的标准
  - 代表：Structure Query Language (SQL) 语言
- 关系代数的思想具有一般性： 管理、添加和分析数据

直观理解：一切都是表格

Event	Channel	Time	Weight
0	0	1	1.1
0	0	1.1	1.15
0	2	1.2	1.3
1	3	0.8	0.9

Event	Channel	Wave
0	0	[0,0.1,...,0]
0	2	[0,0.2,...,0]

基本动机

- 关系代数设计师 Todd，图灵奖工作
- 数据都应该自我描述
  - 即使数据的存储形式变了，对程序进行操作的程序也不应该改变
  - 反例：链表
  - 反例：随意写成的 Excel 表格
- 方便扩展到大规模的数据库中

实用价值

- 引擎优化与应用分工
- 引擎：自动 out-of-core computing (超出内存的运算)
- 引擎：自动并行计算