数据分析基础

Introduction of Data Analytics

Mr. Black

目录

- 大神的工具箱
- R基础数据处理
- R基础绘图

大神的工具箱

大神 Hadley Wickham

Hadley Wickham

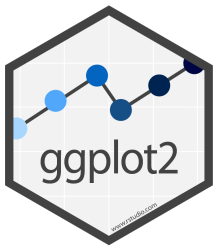
Chief Scientist at **RStudio**

Adjunct Professor of Statistics at the University of Auckland, Stanford University, and Rice University

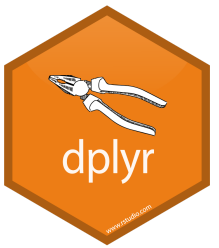
R packages: ggplot2, dplyr, tidyr, stringr, lubridate, readr...

Website: http://hadley.nz

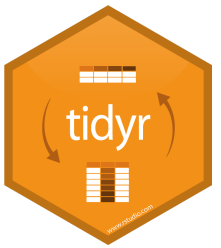
大神的工具箱 Tidyverse



ggplot2
数据可视化



dplyr
数据操作



tidyr
数据整理

readr
数据读写

purrr
函数式编程

tibble
现代Data Frame

forcats
因子数据

lubridate
日期时间数据

stringr
字符数据

marittr
管道操作符

R 基础数据处理

文件读取和保存

R语言可以很容易的从结构化的文本文件中获取数据。用户可以利用read.table()函数将结构化的文本文件读入R，并转化成数据框。read.table()函数的定义如下：

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
  dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
  row.names, col.names, as.is = !stringsAsFactors,
  na.strings = "NA", colClasses = NA, nrows = -1,
  skip = 0, check.names = TRUE, fill = !blank.lines.skip,
  strip.white = FALSE, blank.lines.skip = TRUE,
  comment.char = "#", allowEscapes = FALSE, flush = FALSE,
  stringsAsFactors = default.stringsAsFactors(),
  fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)
```

```
read.csv(file, header = TRUE, sep = ",", quote = "\"", ...)
```

```
read.delim(file, header = TRUE, sep = "\t", quote = "\"", ...)
```

文件读取和保存

参数	默认值	说明
file	无	文件名（可以为URL地址等）
header	FALSE	文件的第一行是否为变量的名称
sep	空	数据列之间的分隔符
quote	" 或 '	数据值的引用符号（例如："Male"）
row.names	无	行名称
col.names	无	列名称
na.strings	NA	表示NA值的字符串
comment.char	#	注释行起始字符
stringsAsFactors	default.stringsAsFactors()	字符串是否作为因子

文件读取和保存

用户也可以将对象保存成文本文件，在R中可以利用write.table()函数实现。write.table()函数定义如下：

```
write.table(x, file = "",
  append = FALSE,
  quote = TRUE,
  sep = " ", eol = "\n",
  na = "NA", dec = ".",
  row.names = TRUE,
  col.names = TRUE,
  qmethod = c("escape",
              "double"),
  fileEncoding = "")
```

参数	默认值	说明
x	无	被保存的对象
file	无	文件名
append	FALSE	是否从文件尾部追加写入
quote	TRUE	是否使用双引号括起数据值
sep	空格	数据列之间的分隔符
eol	\n	行尾结束符
na	NA	表示NA值的字符串
row.names	TURE	是否保存行名称
col.names	TURE	是否保存列名称

文件读取和保存

除了系统提供的读取文本文件的函数外，readr包提供了一个更加快速和友好的方式将列表文件读入R。read_delim()函数定义如下：

```
read_delim(file, delim, quote = "\"", escape_backslash = FALSE,
  escape_double = TRUE, col_names = TRUE, col_types = NULL,
  locale = default_locale(), na = c("", "NA"), quoted_na = TRUE,
  comment = "", trim_ws = FALSE, skip = 0, n_max = Inf,
  guess_max = min(1000, n_max), progress = show_progress())
```

函数	说明
read_delim(), read_csv(), read_tsv(), read_csv2()	读取分隔符文件
read_fwf(), read_table()	读取固定宽度文件
read_lines()	按行读取文件
read_file()	读取整个文件

文件读取和保存

在读取较大文件时，readr包中的函数会提供一个进度条供用户观察文件读取进度。readr包中的函数相比R中自带的函数读取文件速度要快，因此当用户希望读取较大文件时建议使用readr中的函数。

```
poker.hand.testing <- read_csv("poker-hand-testing.data.txt",
                               col_names=FALSE)
|==================================================| 100%   23 MB
```

```
system.time(read_csv("poker-hand-testing.data.txt"))
# 用户 系统 流逝
# 1.36 0.02 1.41
system.time(read.csv("poker-hand-testing.data.txt"))
# 用户 系统 流逝
# 4.94 0.01 4.97
```

数据变换

dplyr常用函数

函数	功能	函数	功能	函数	功能
glimpse	描述	filter	过滤	distinct	去重
sample_n, sample_frac	采样	slice, top_n	选择行	select	选择列
summarise, summarise_each	概括	count, first, last	-	n, n_distinct	-
group_by	分组	xxx_join	关联	intersect	交集
union	并集	setdiff	差集	setequal	异同
mutate, transmute	衍生	gather	列转行	spread	行转列

数据变换

常用窗口函数

函数	功能	函数	功能
row_number	行号	min_rank	rank(ties.method = "min")
dense_rank	无缝排序	percent_rank	将min_rank归一化到[0, 1]
cume_dist	累积分布	ntile	划分为n等份
lead	数据提前n个，后面补NA	lag	数据滞后n个，前面补NA
cumall	累积all函数	cumany	累积any函数
cummean	累积mean函数	cumsum	累积sum函数
cumprod	累积prod函数	cummax	累积max函数
cummin	累积min函数		

apply函数族及其扩展

在R语言中我们倡导尽可能少的使用循环结构，因为大量的使用for和while语句会使得代码不够简洁，同时也失去了R作为函数式编程语言的一些优势。因此我们可以利用一些R自带的函数去替换for和while等循环结构，这就是R中的apply函数族。当然apply函数族并不能够完全的替代循环结构，尤其是当每次循环之间存在复杂的关系时，apply函数族只是部分循环的一个更优的替代方案。

能不用for就不用for！

apply, lapply, sapply, tapply, mapply

apply函数族及其扩展

plyr扩展包常用函数

输入/输出	array	data frame	list	舍弃
array	aaply	adply	alply	a_ply
data frame	daply	ddply	dlply	d_ply
list	laply	ldply	llply	l_ply

R 可视化

ggplot2

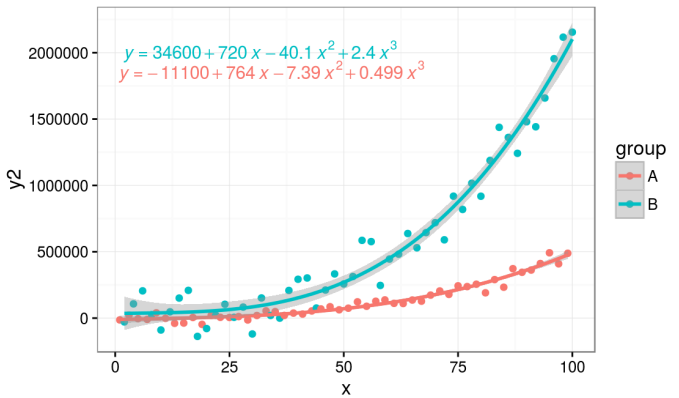
ggplot2是R中一套基于特定图形语法[1]的绘图系统。从ggplot2的观点出发，一张统计图包括从数据（**Data**）到几何对象（**Geometric Objects**，描述了绘制的图像类型）的一个映射（**Aesthetic Mapping**）。图中还可能包括不同的图层（**Layer**），数据的统计变换（**Statistics Transformation**），用于控制数据和映射的标度（**Scales**），不同的坐标系统（**Coordinate Systems**）以及子图像分面（**Facet**）。除此之外，还可能包括一些相关的主题（**Themes**）和注释（**Annotation**）。本章节仅对ggplot2绘图系统做出基本讲解，大部分内容参考了ggplot2的官方文档[2]，更加详细全面的使用请参见《ggplot2: elegant graphics for data analysis》[3]。

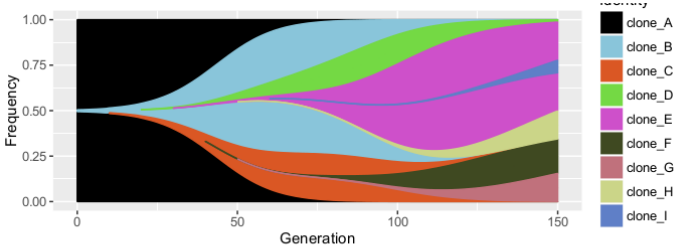
[1] L. Wilkinson, The grammar of graphics. Springer Science & Business Media, 2006.

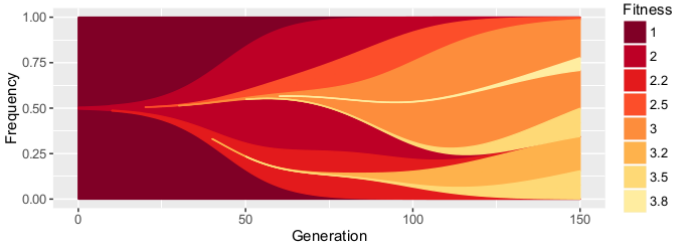
[2] http://ggplot2.tidyverse.org/

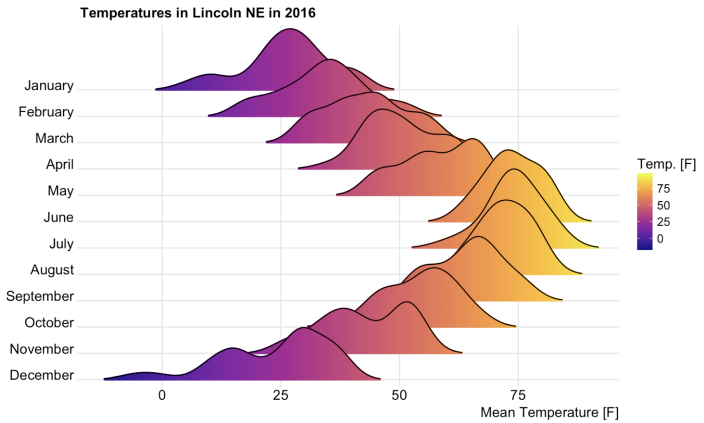
[3] H. Wickham, ggplot2: elegant graphics for data analysis. Springer Science & Business Media, 2009.

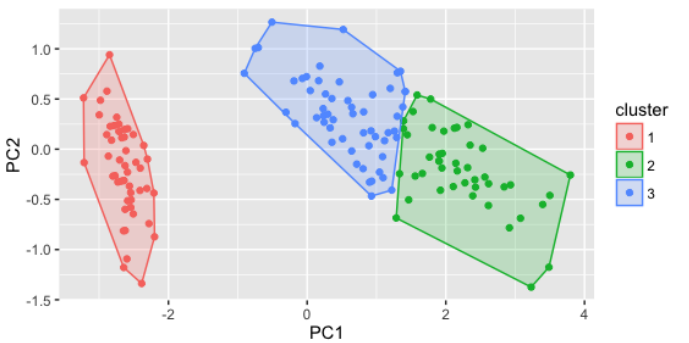
ggplot2


$y = 34600 + 720\,x - 40.1\,x^2 + 2.4\,x^3$
$y = -11100 + 764\,x - 7.39\,x^2 + 0.499\,x^3$
2000000
1500000
1000000
500000
0
y2
0
25
50
75
100
x
group
A
B


1.00
0.75
0.50
0.25
0.00
Frequency
0
50
100
150
Generation
clone_A
clone_B
clone_C
clone_D
clone_E
clone_F
clone_G
clone_H
clone_I


1.00
0.75
0.50
0.25
0.00
Frequency
0
50
100
150
Generation
Fitness
1
2
2.2
2.5
3
3.2
3.5
3.8


Temperatures in Lincoln NE in 2016
January
February
March
April
May
June
July
August
September
October
November
December
0
25
50
75
Mean Temperature [F]
Temp. [F]
75
50
25
0

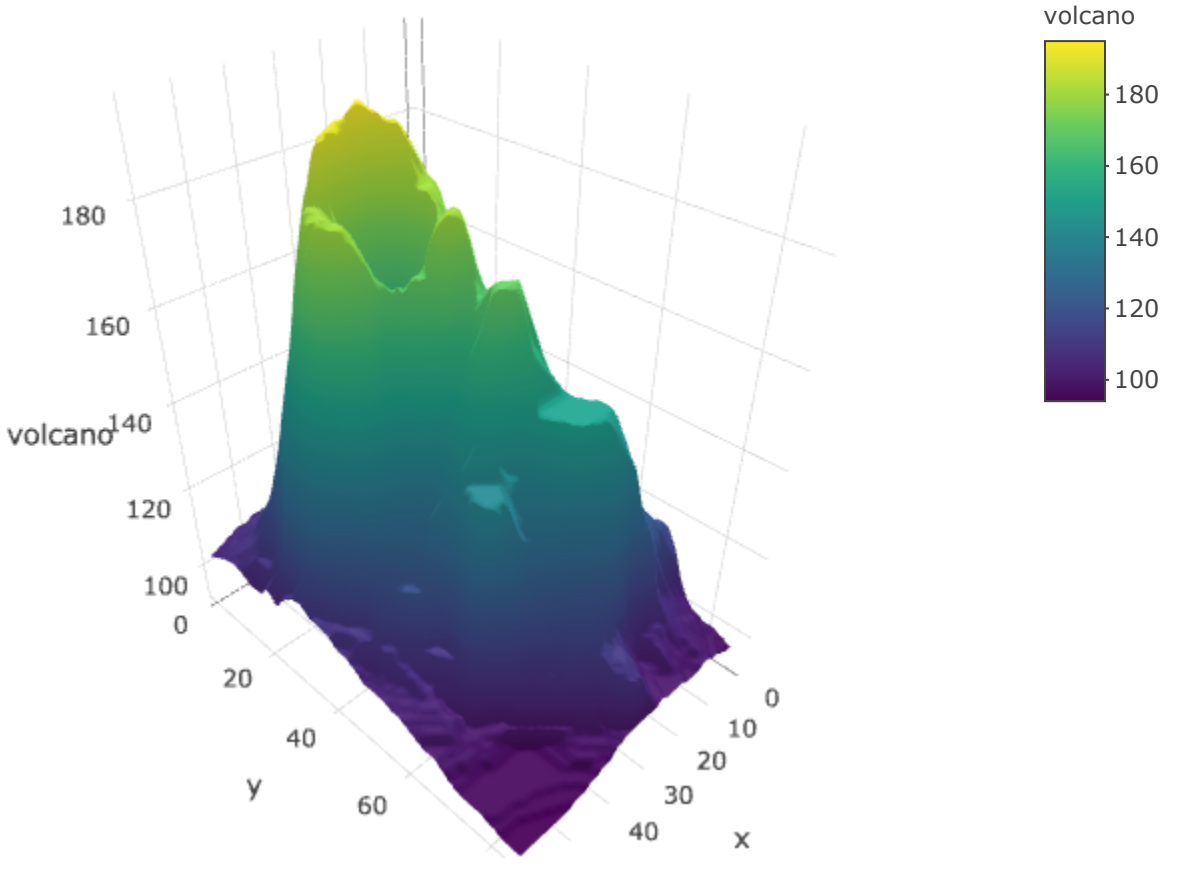

1.0
0.5
0.0
-0.5
-1.0
-1.5
PC2
-2
0
2
4
PC1
cluster
1
2
3

ggplot2

ggplot2 官网: http://ggplot2.tidyverse.org/

ggplot2 扩展: http://www.ggplot2-exts.org/

Plotly


volcano
180
160
140
120
100
volcano
180
160
140
120
100
0
20
40
60
y
0
10
20
30
40
x

Thanks