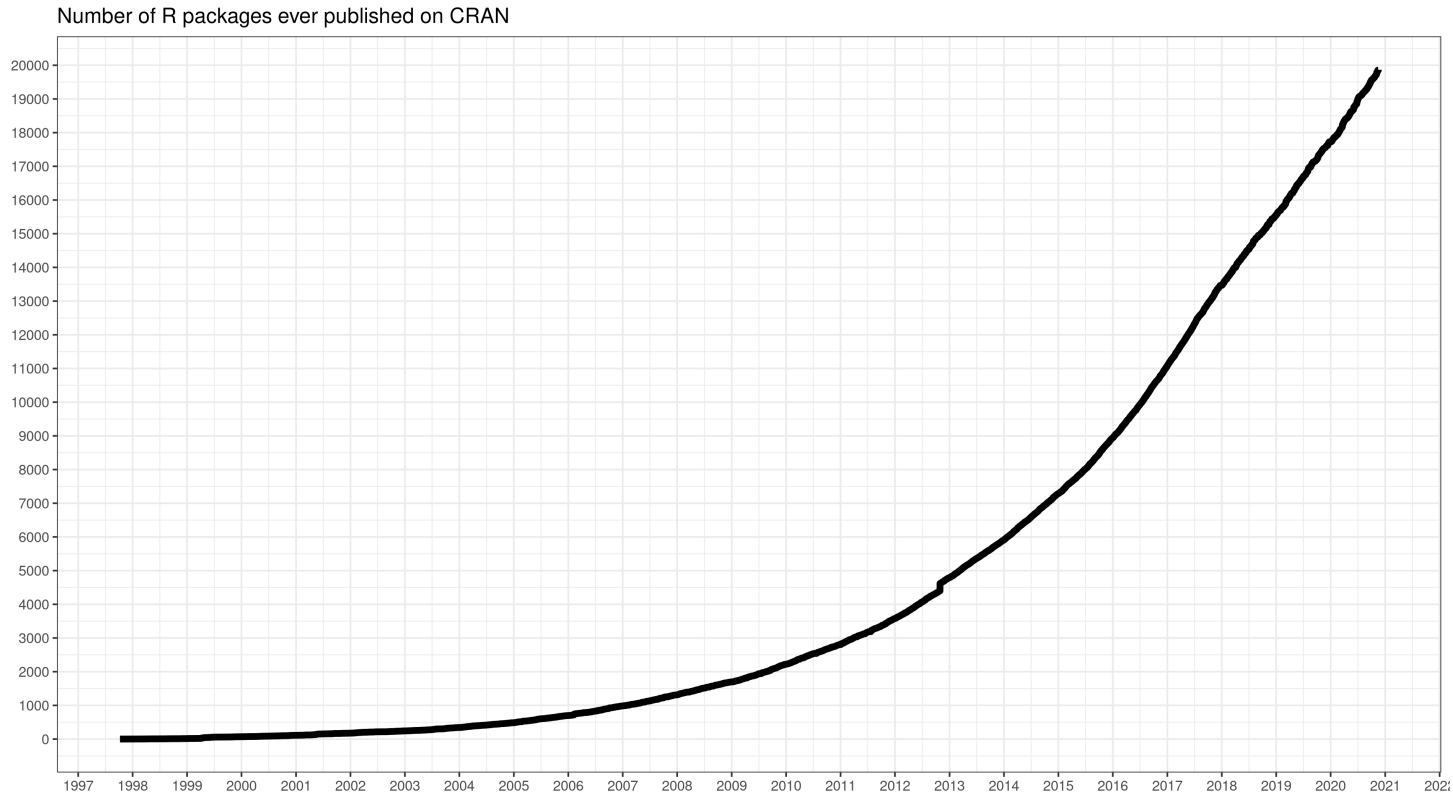
Learning R

Pylone

2021/08/05

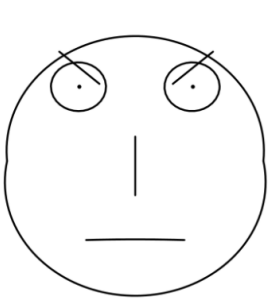
1.Why

1.1开源：相当于手机系统，可以使用任何在此平台上开发的软件，叫做包（**package**）

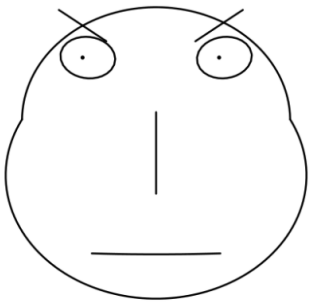


packages number

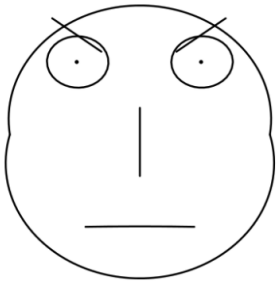
1.2 无所不能的分析及可视化工具，一些可视化例子，浏览以便于激发灵感



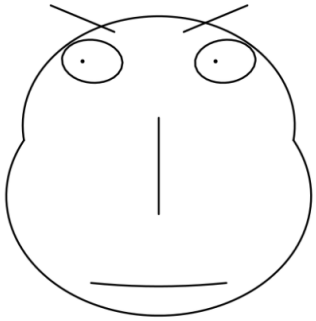
AARONSON,L.H.



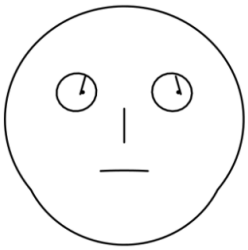
ALEXANDER,J.M.



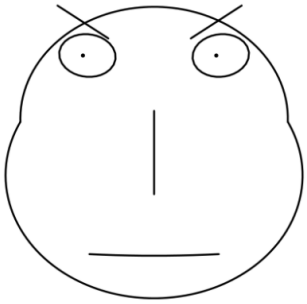
ARMENTANO,A.J.



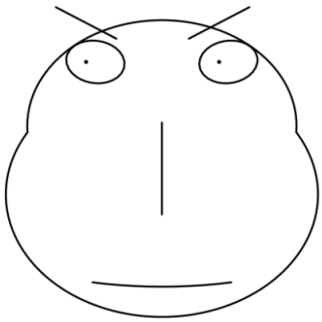
BERDON,R.I.



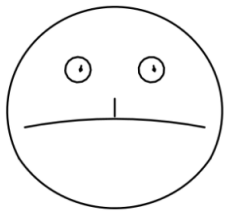
BRACKEN,J.J.



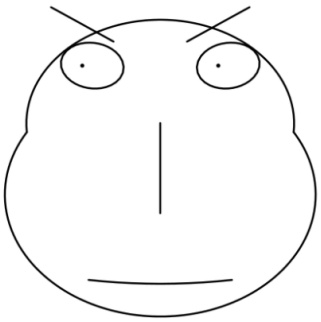
BURNS,E.B.



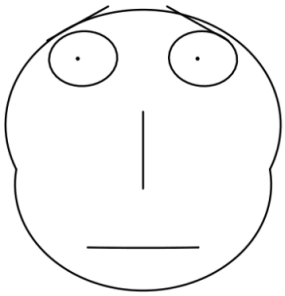
CALLAHAN,R.J.



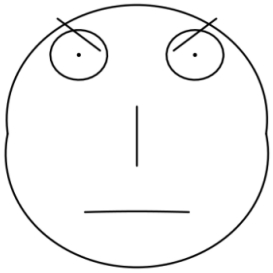
COHEN,S.S.



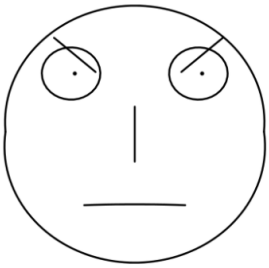
DALY,J.J.



DANNEHY,J.F.



DEAN,H.H.



DEVITA,H.J.

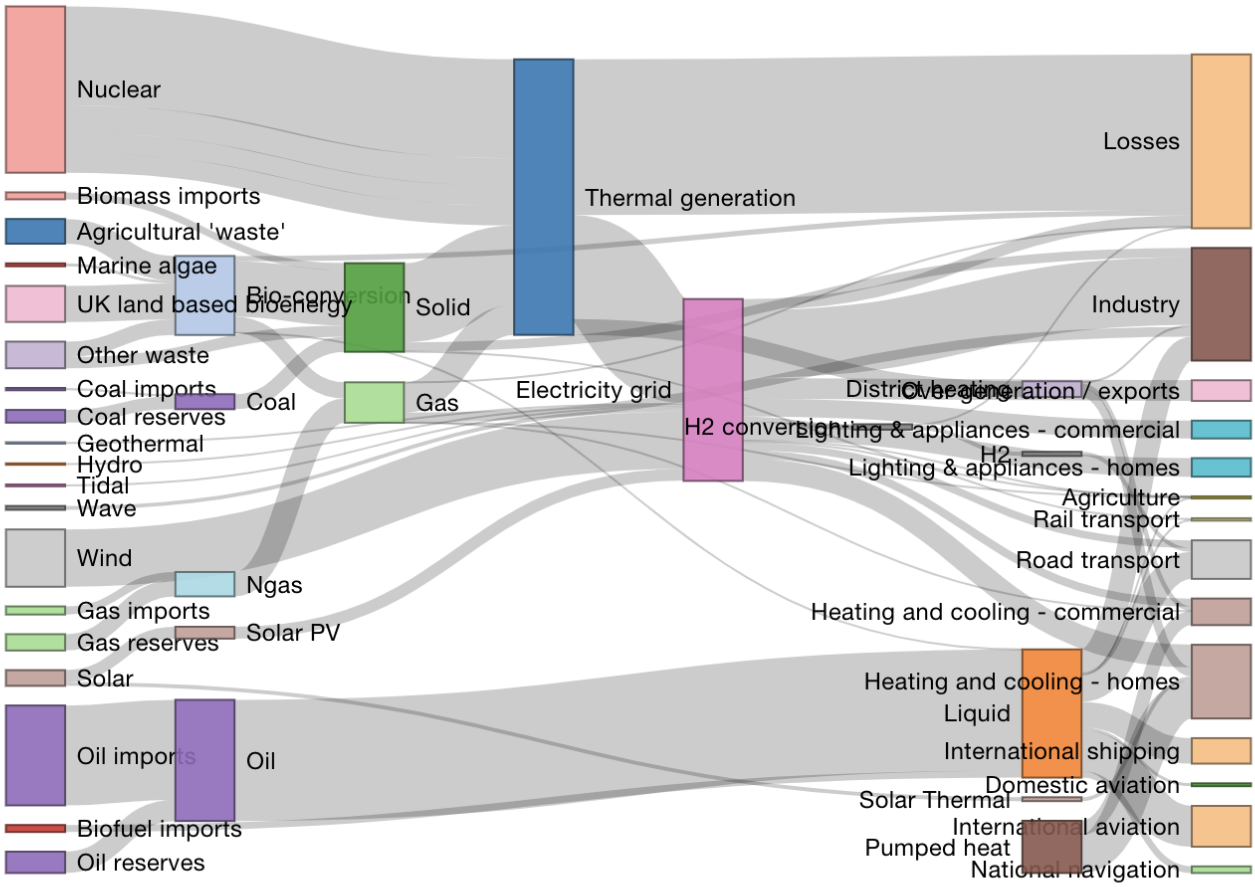
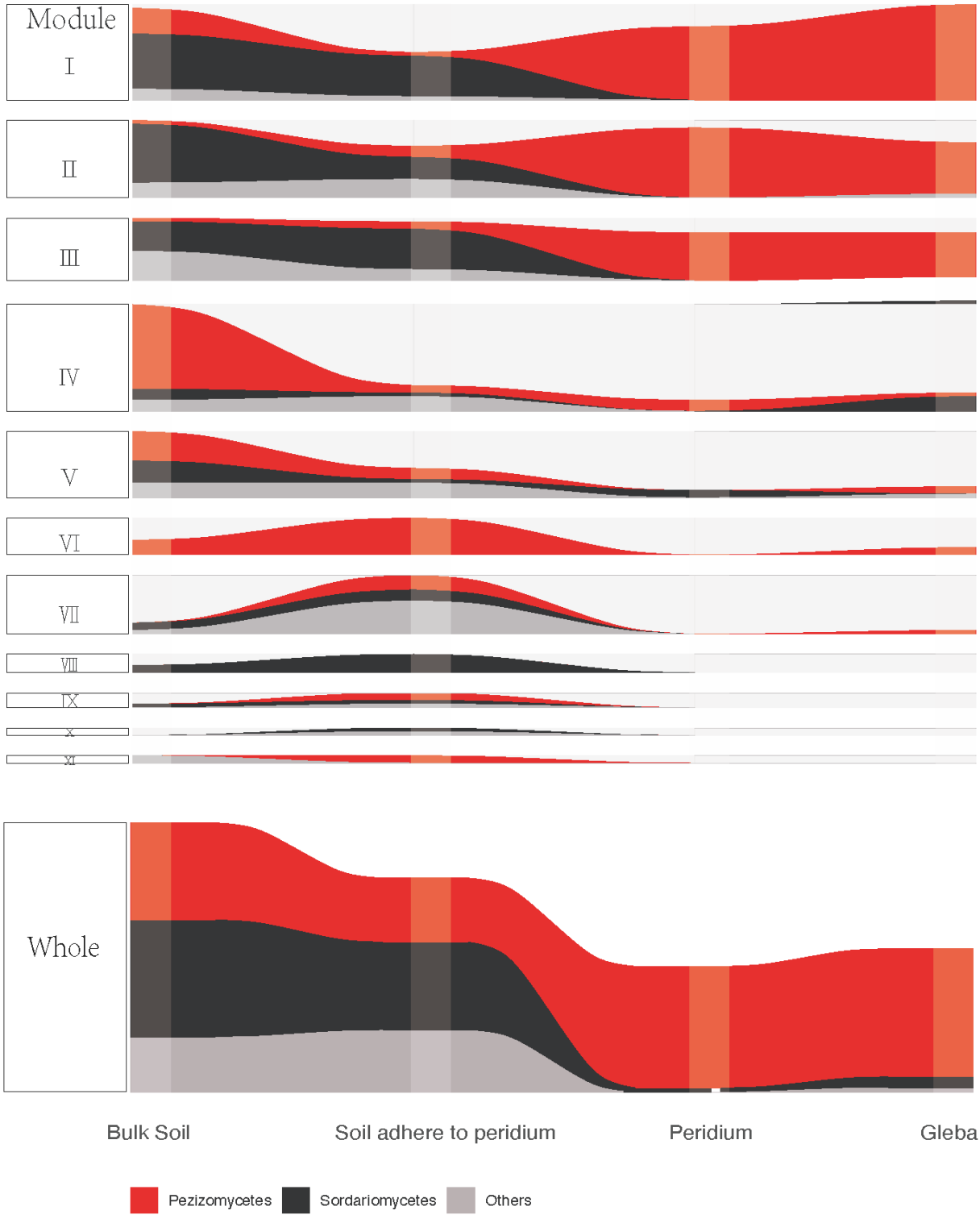
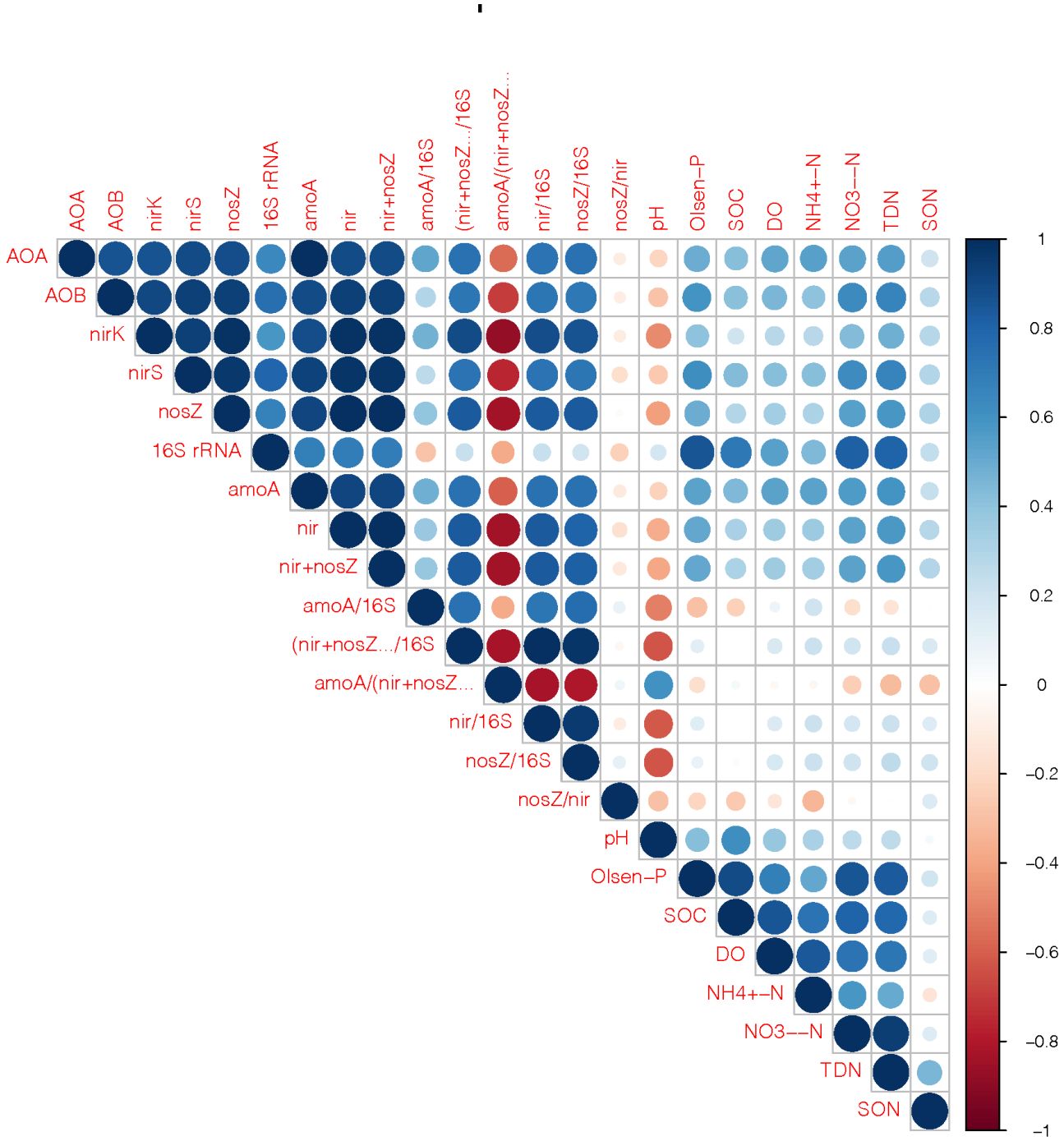


Figure 9.8: Sankey diagram

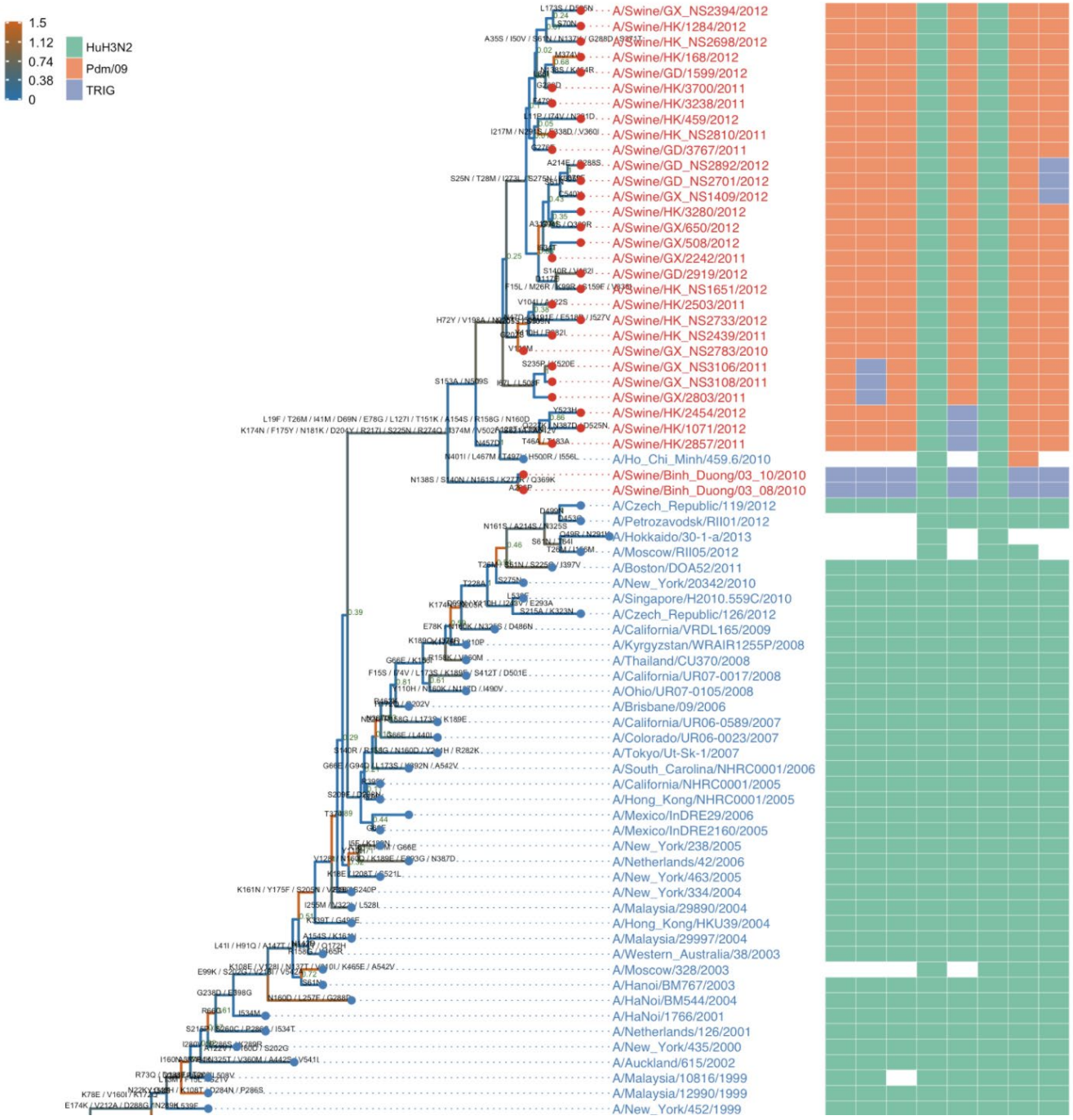
桑基图


Module
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
Whole
Bulk Soil
Soil adhere to peridium
Peridium
Gleba
Pezizomycetes
Sordariomycetes
Others

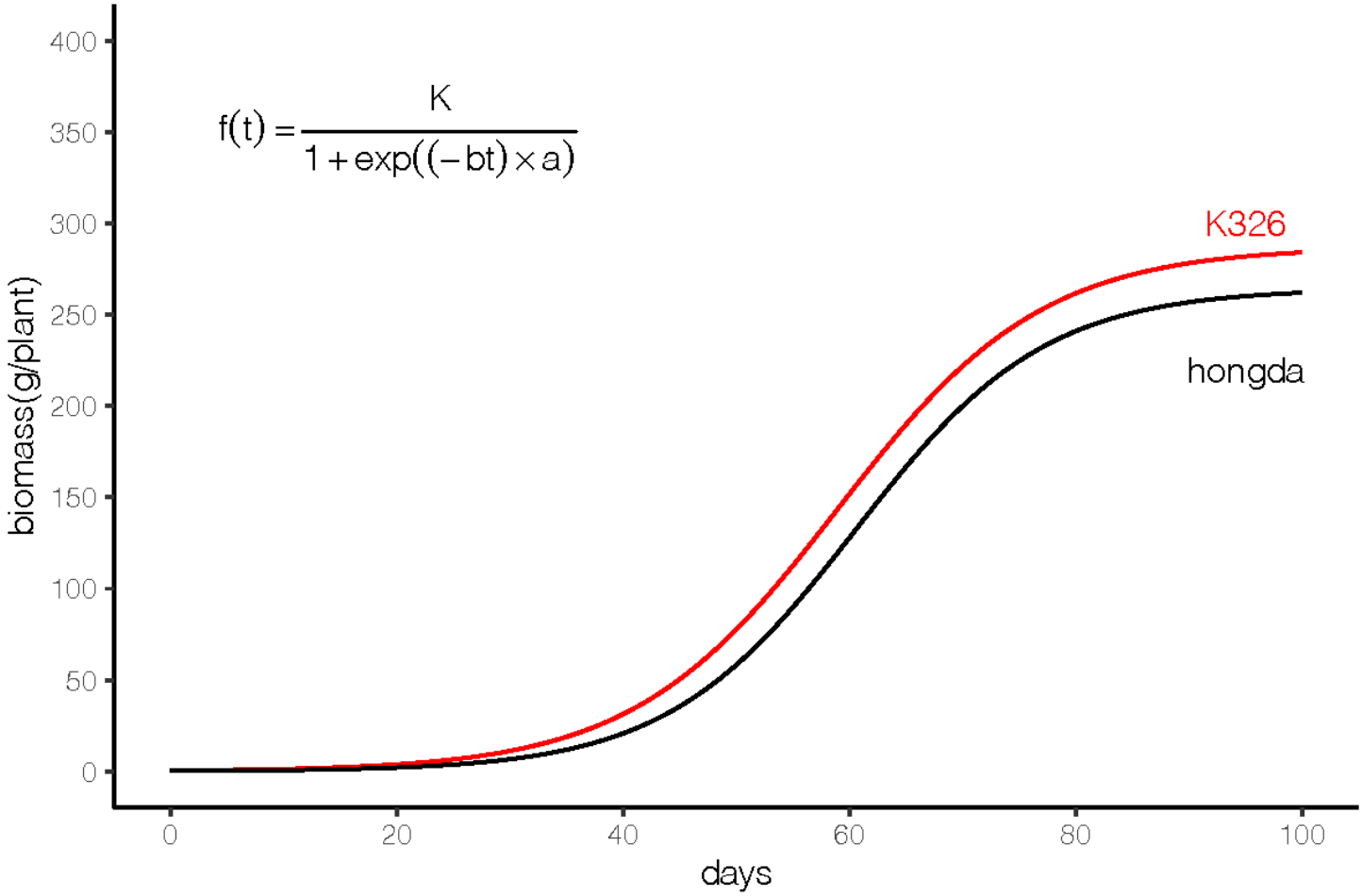
冲积图


AOA
AOB
nirK
nirS
nosZ
16S rRNA
amoA
nir
nir+nosZ
amoA/16S
(nir+nosZ.../16S
amoA/(nir+nosZ...
nir/16S
nosZ/16S
nosZ/nir
pH
Olsen−P
SOC
DO
NH4+−N
NO3−−N
TDN
SON
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

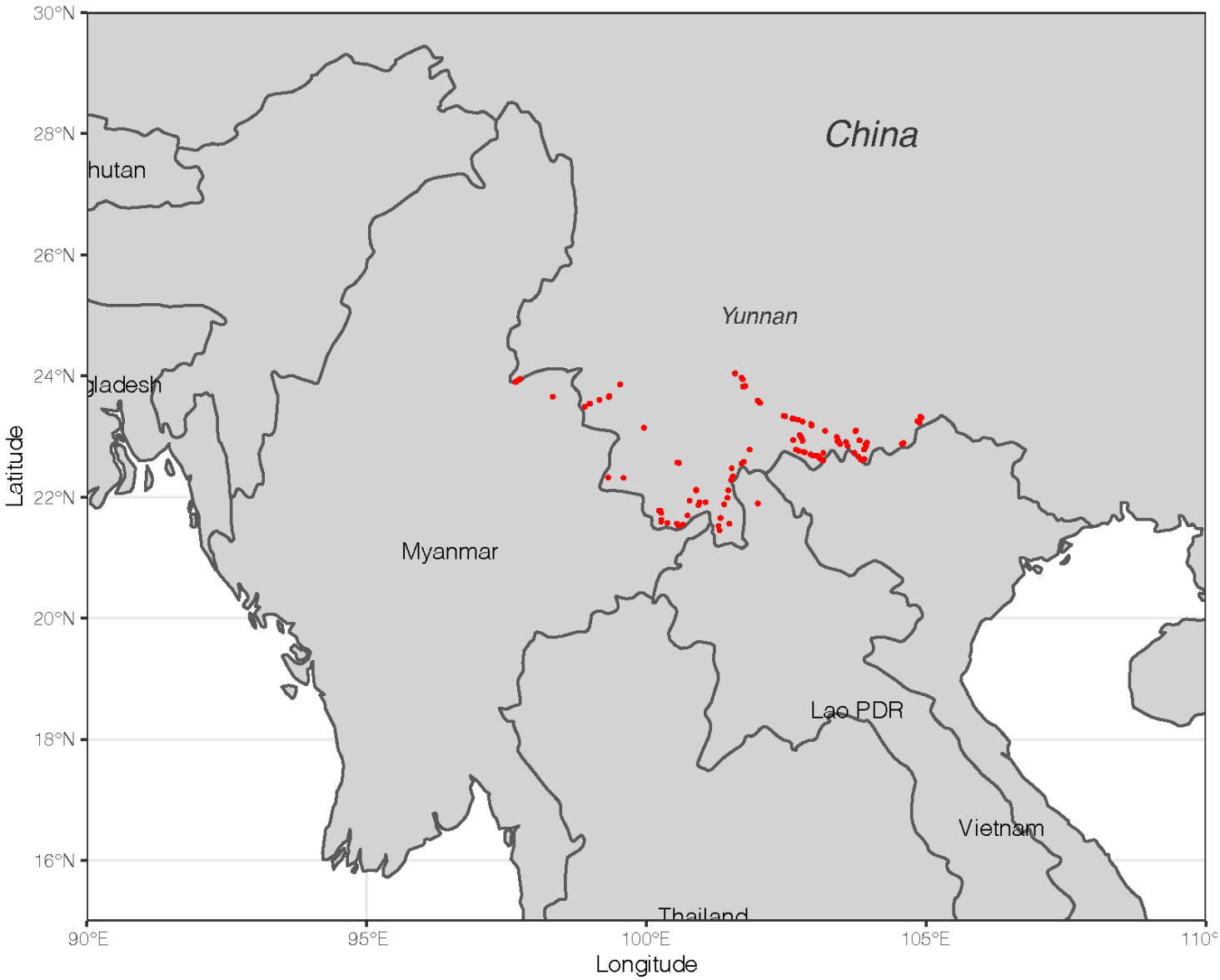
相关性


1.5
1.12
0.74
0.38
0
HuH3N2
Pdm/09
TRIG
A/Swine/GX_NS2394/2012
A/Swine/HK/1284/2012
A/Swine/HK_NS2698/2012
A/Swine/HK/168/2012
A/Swine/GD/1599/2012
A/Swine/HK/3700/2011
A/Swine/HK/3238/2011
A/Swine/HK/459/2012
A/Swine/HK_NS2810/2011
A/Swine/GD/3767/2011
A/Swine/GD_NS2892/2012
A/Swine/GD_NS2701/2012
A/Swine/GX_NS1409/2012
A/Swine/HK/3280/2012
A/Swine/GX/650/2012
A/Swine/GX/508/2012
A/Swine/GX/2242/2011
A/Swine/GD/2919/2012
A/Swine/HK_NS1651/2012
A/Swine/HK/2503/2011
A/Swine/HK_NS2733/2012
A/Swine/HK_NS2439/2011
A/Swine/GX_NS2783/2010
A/Swine/GX_NS3106/2011
A/Swine/GX_NS3108/2011
A/Swine/GX/2803/2011
A/Swine/HK/2454/2012
A/Swine/HK/1071/2012
A/Swine/HK/2857/2011
A/Ho_Chi_Minh/459.6/2010
A/Swine/Binh_Duong/03_10/2010
A/Swine/Binh_Duong/03_08/2010
A/Czech_Republic/119/2012
A/Petrozavodsk/RII01/2012
A/Hokkaido/30-1-a/2013
A/Moscow/RII05/2012
A/Boston/DOA52/2011
A/New_York/20342/2010
A/Singapore/H2010.559C/2010
A/Czech_Republic/126/2012
A/California/VRDL165/2009
A/Kyrgyzstan/WRAIR1255P/2008
A/Thailand/CU370/2008
A/California/UR07-0017/2008
A/Ohio/UR07-0105/2008
A/Brisbane/09/2006
A/California/UR06-0589/2007
A/Colorado/UR06-0023/2007
A/Tokyo/Ut-Sk-1/2007
A/South_Carolina/NHRC0001/2006
A/California/NHRC0001/2005
A/Hong_Kong/NHRC0001/2005
A/Mexico/InDRE29/2006
A/Mexico/InDRE2160/2005
A/New_York/238/2005
A/Netherlands/42/2006
A/New_York/463/2005
A/New_York/334/2004
A/Malaysia/29890/2004
A/Hong_Kong/HKU39/2004
A/Malaysia/29997/2004
A/Western_Australia/38/2003
A/Moscow/328/2003
A/Hanoi/BM767/2003
A/HaNoi/BM544/2004
A/HaNoi/1766/2001
A/Netherlands/126/2001
A/New_York/435/2000
A/Auckland/615/2002
A/Malaysia/10816/1999
A/Malaysia/12990/1999
A/New_York/452/1999

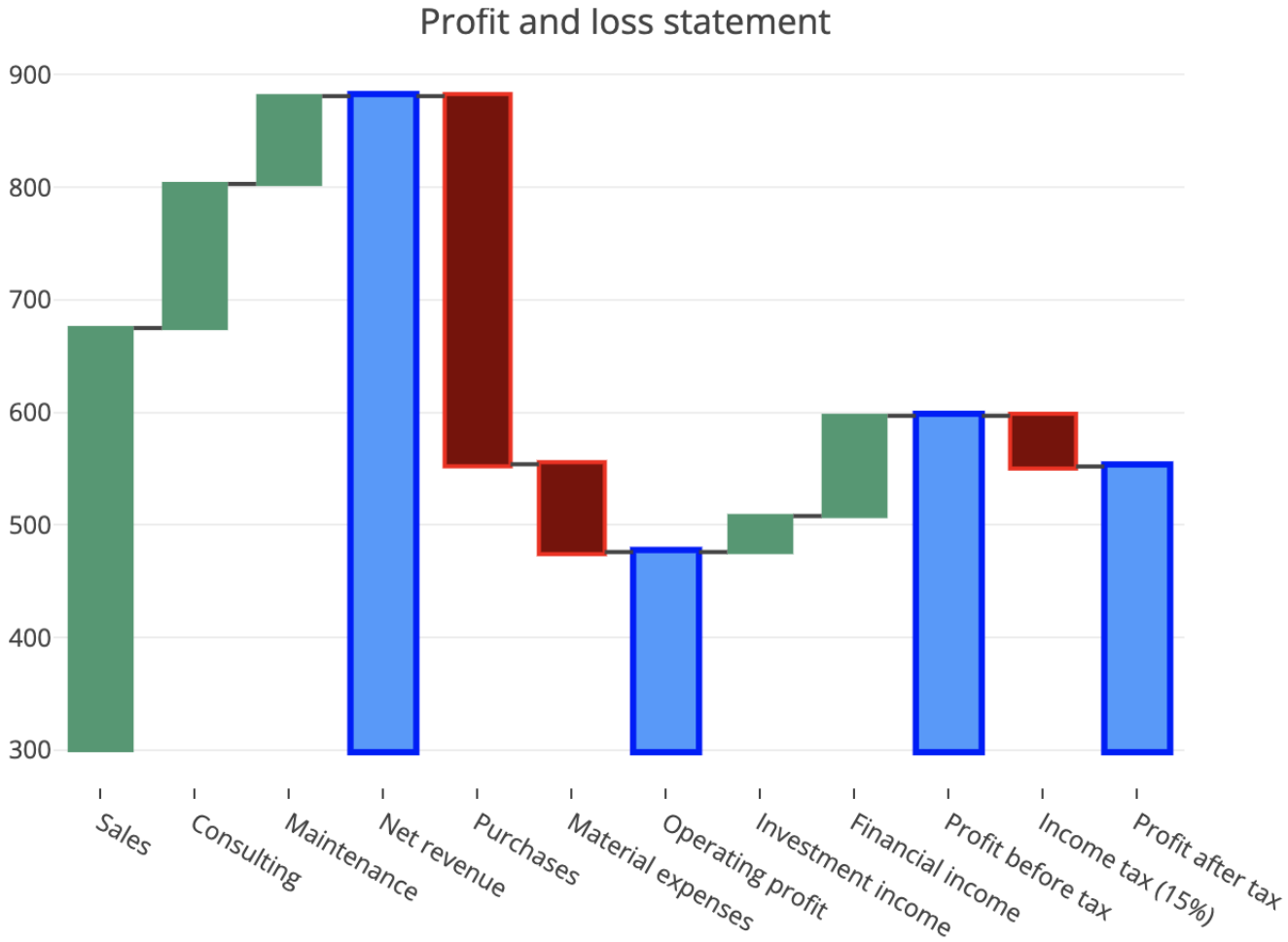
聚类图


$f(t) = \frac{K}{1 + \exp((-bt) \times a)}$
K326
hongda
biomass(g/plant)
days
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100

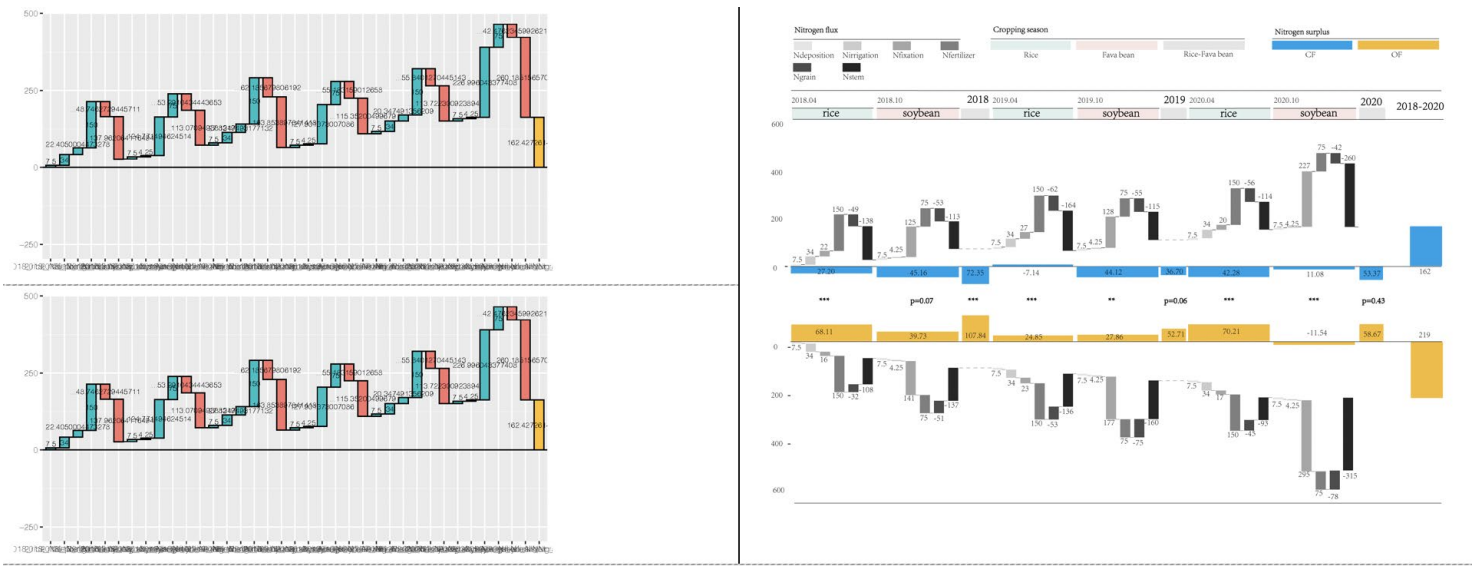
曲线图


30°N
28°N
26°N
24°N
22°N
20°N
18°N
16°N
Latitude
China
hutan
Yunnan
gladesh
Myanmar
Lao PDR
Vietnam
Thailand
90°E
95°E
100°E
105°E
110°
Longitude

map



waterfall



使用R出图，AI加工图

1.3更大魅力在于对数据的自由探索

1.4丰富的功能：word、pdf、ppt、html、blog等

2.How

Q1:怎么学的更快?

带着需求来学习，学以致用

Q2:小白怎么入手?

安装、基础操作：教程，《R语言实战》相当于查询手册

Q3:遇到问题怎么办?

（1）先从教程（查询手册）中找答案，再去百度、谷歌（推荐），再交流、总结、分享

（2）真想学好还是建议翻墙，巨量优质的教程分享

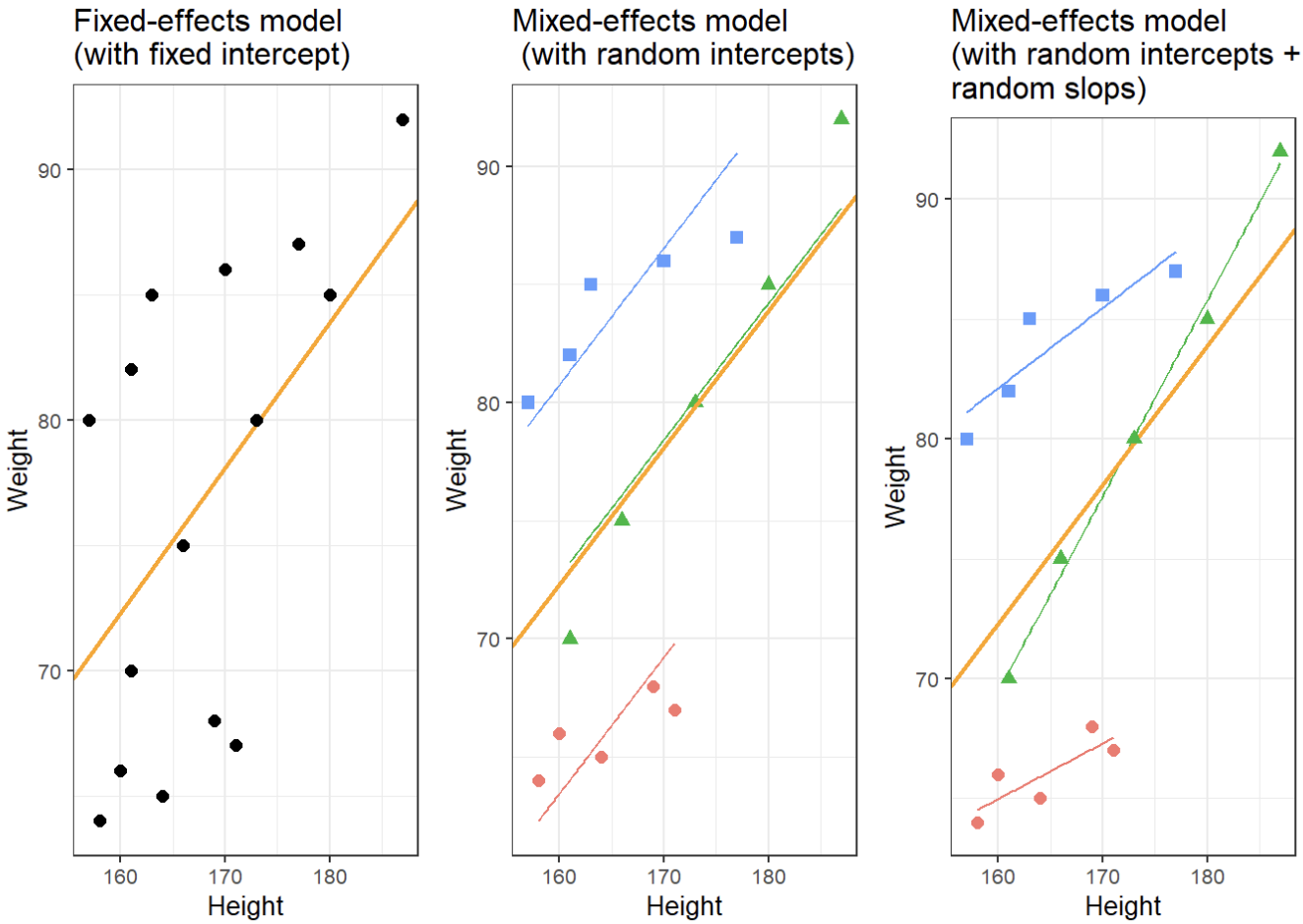
速查表

3.我的数据可以用哪些方法分析？-多元统计

逃不出多元统计分析，去看学习强国的多元统计分析课程。

3.1变量

- 自变量：多元相关cor、主成分分析OCA、因子分析CA
- 自变量与因变量：多元回归、线性模型（广义glm、一般（方差分析）、普通、线性混合效应模型lme）、冗余分析RDA



线性混合效应模型

> 尤其适用于以下情况：
>
> - 数据有缺失值；
> - 不平衡设计，如一个处理3个样本，另一个5个样本；
> - 结构化（nested）数据；
> - 结果变量非连续，比如评分（5，10，15）；
> - 等等。。

3.2样本

- 聚类分析

3.3样本与变量

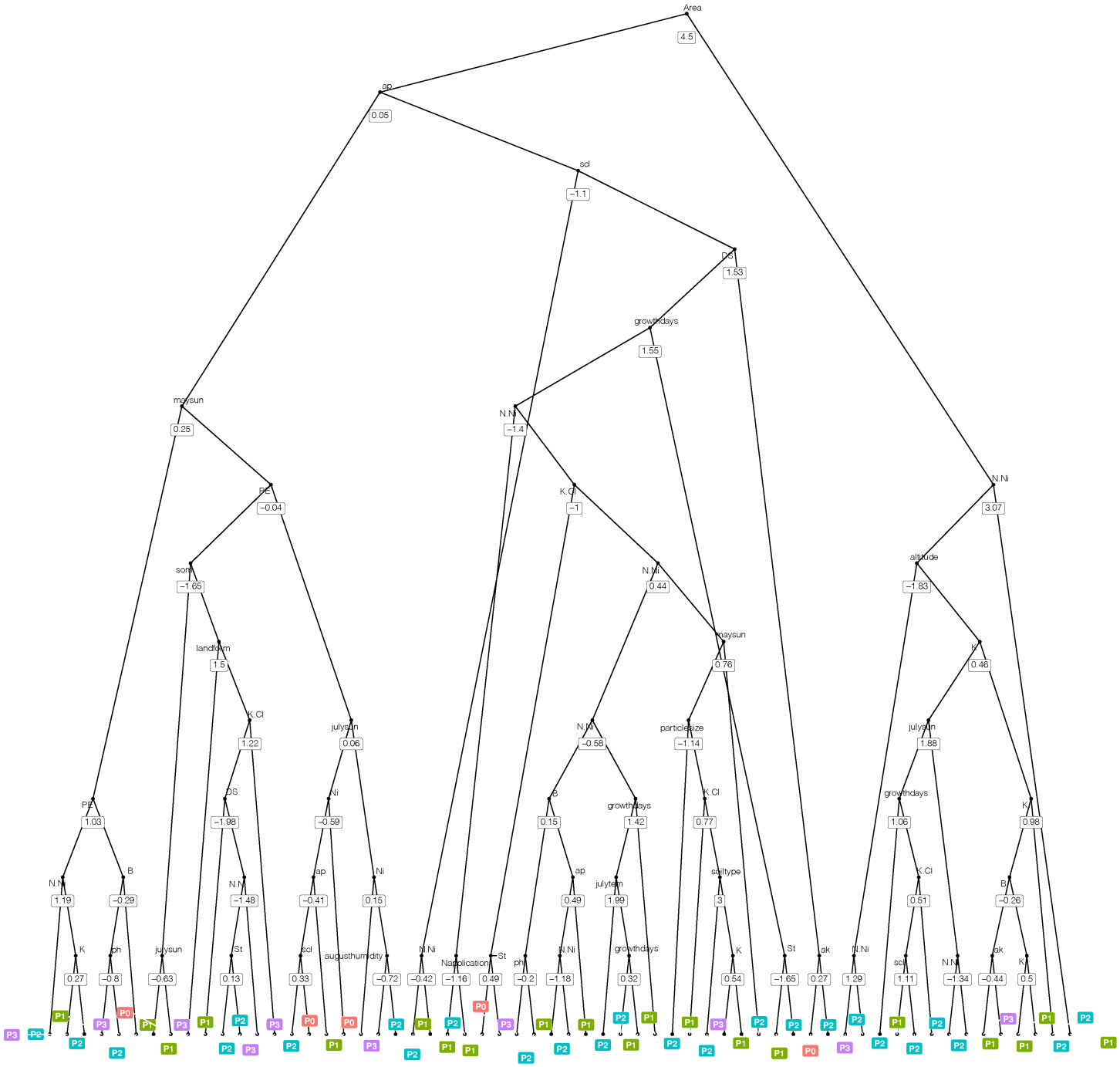
- 样本与自变量：对应分析（CA）
- 样本、自变量与因变量：约束性对应分析（CCA）

4.我想用主成分分析来分析数据，用哪些包？怎么用？

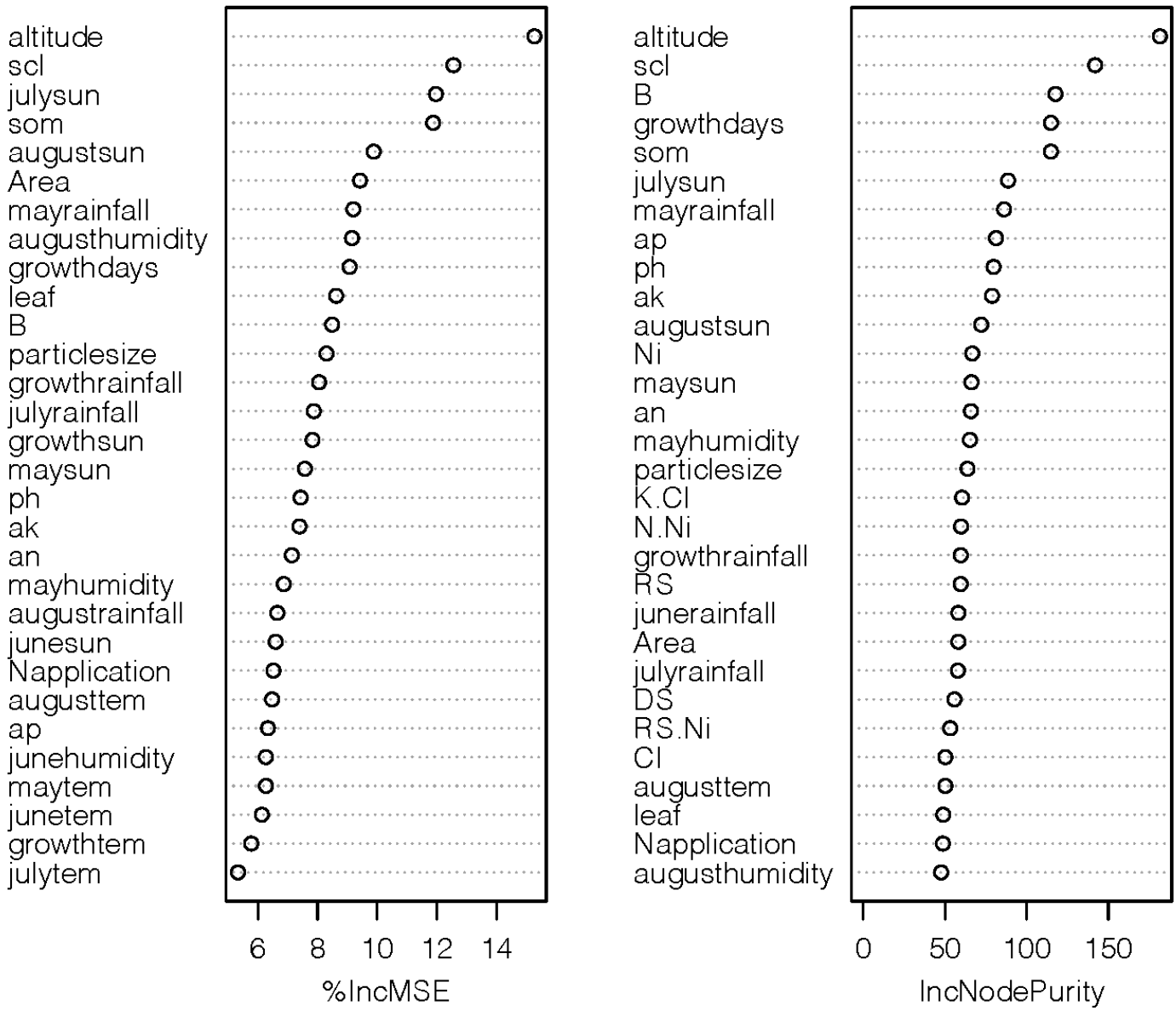
- 每种分析方法最好找对应的包，查询手册-网络教程-线上或线下寻求帮助
- 需要从实践、学习中不断总结，大家分享
- 好评包：dplyr、ggplot等

5.预测模型

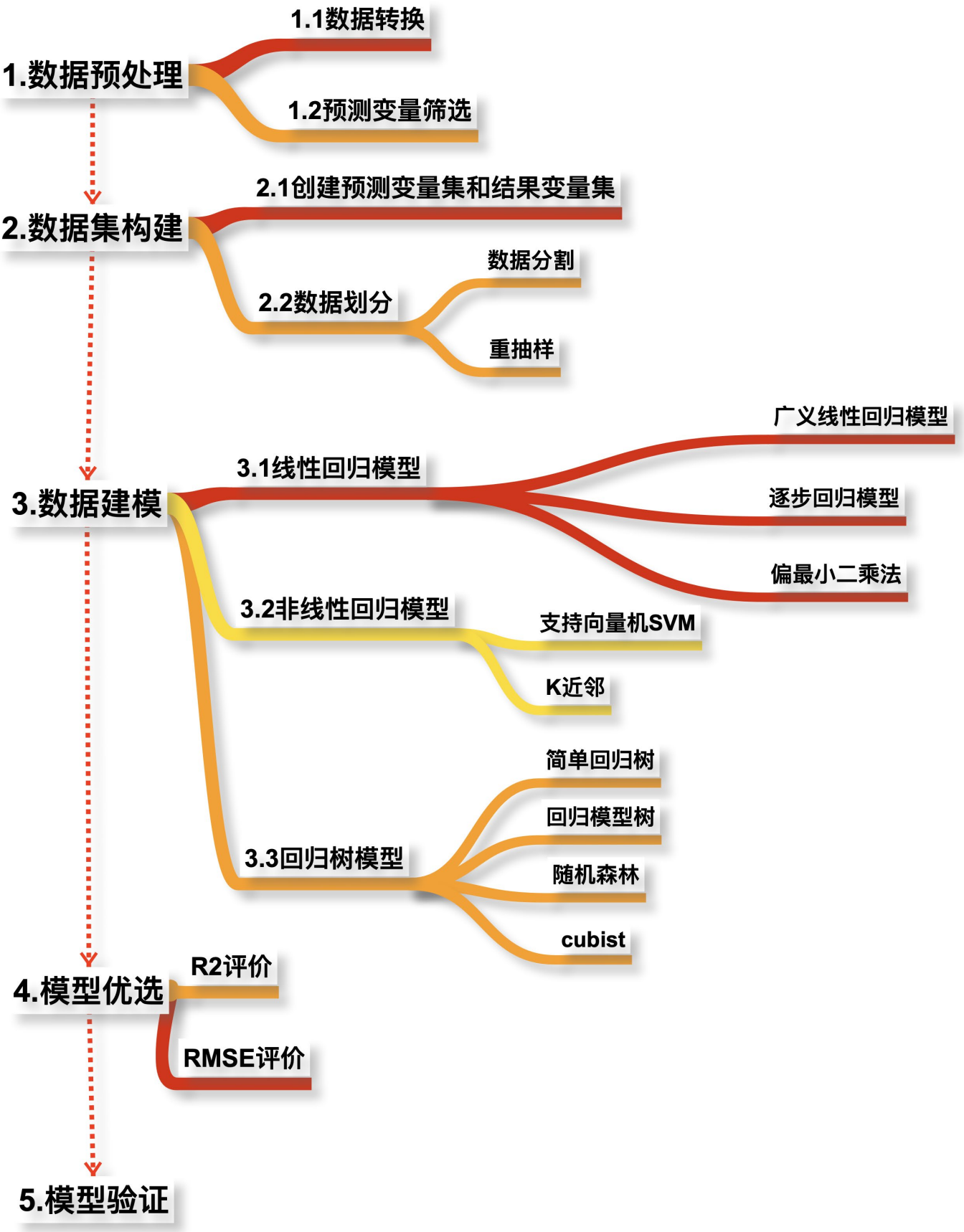
随机森林、神经网络、支持向量机等，大量数据的训练，caret包


Area
4.5
ap
0.05
scl
−1.1
DS
1.53
growthdays
1.55
maysun
0.25
N.Ni
−1.4
N.Ni
3.07
PE
−0.04
K.Cl
−1
altitude
−1.83
soil
−1.65
N.Ni
0.44
K
0.46
landform
1.5
maysun
0.76
K.Cl
1.22
julysun
0.06
N.Ni
−0.58
particlesize
−1.14
julysun
1.88
PE
1.03
DS
−1.98
Ni
−0.59
B
0.15
growthdays
1.42
K.Cl
0.77
growthdays
1.06
K
0.98
N.Ni
1.19
B
−0.29
N.Ni
−1.48
ap
−0.41
Ni
0.15
ap
0.49
julytem
1.99
soiltype
3
K.Cl
0.51
B
−0.26
K
0.27
ph
−0.8
julysun
−0.63
St
0.13
scl
0.33
augusthumidity
−0.72
N.Ni
−0.42
Napplication
−1.16
St
0.49
ph
−0.2
N.Ni
−1.18
growthdays
0.32
K
0.54
St
−1.65
ak
0.27
N.Ni
1.29
scl
1.11
N.Ni
−1.34
ak
−0.44
K
0.5

随机森林预测


forest
altitude
scl
julysun
som
augustsun
Area
mayrainfall
augusthumidity
growthdays
leaf
B
particlesize
growthrainfall
julyrainfall
growthsun
maysun
ph
ak
an
mayhumidity
augustrainfall
junesun
Napplication
augusttem
ap
junehumidity
maytem
junetem
growthtem
julytem
6
8
10
12
14
%IncMSE
altitude
scl
B
growthdays
som
julysun
mayrainfall
ap
ph
ak
augustsun
Ni
maysun
an
mayhumidity
particlesize
K.Cl
N.Ni
growthrainfall
RS
junerainfall
Area
julyrainfall
DS
RS.Ni
Cl
augusttem
leaf
Napplication
augusthumidity
0
50
100
150
IncNodePurity

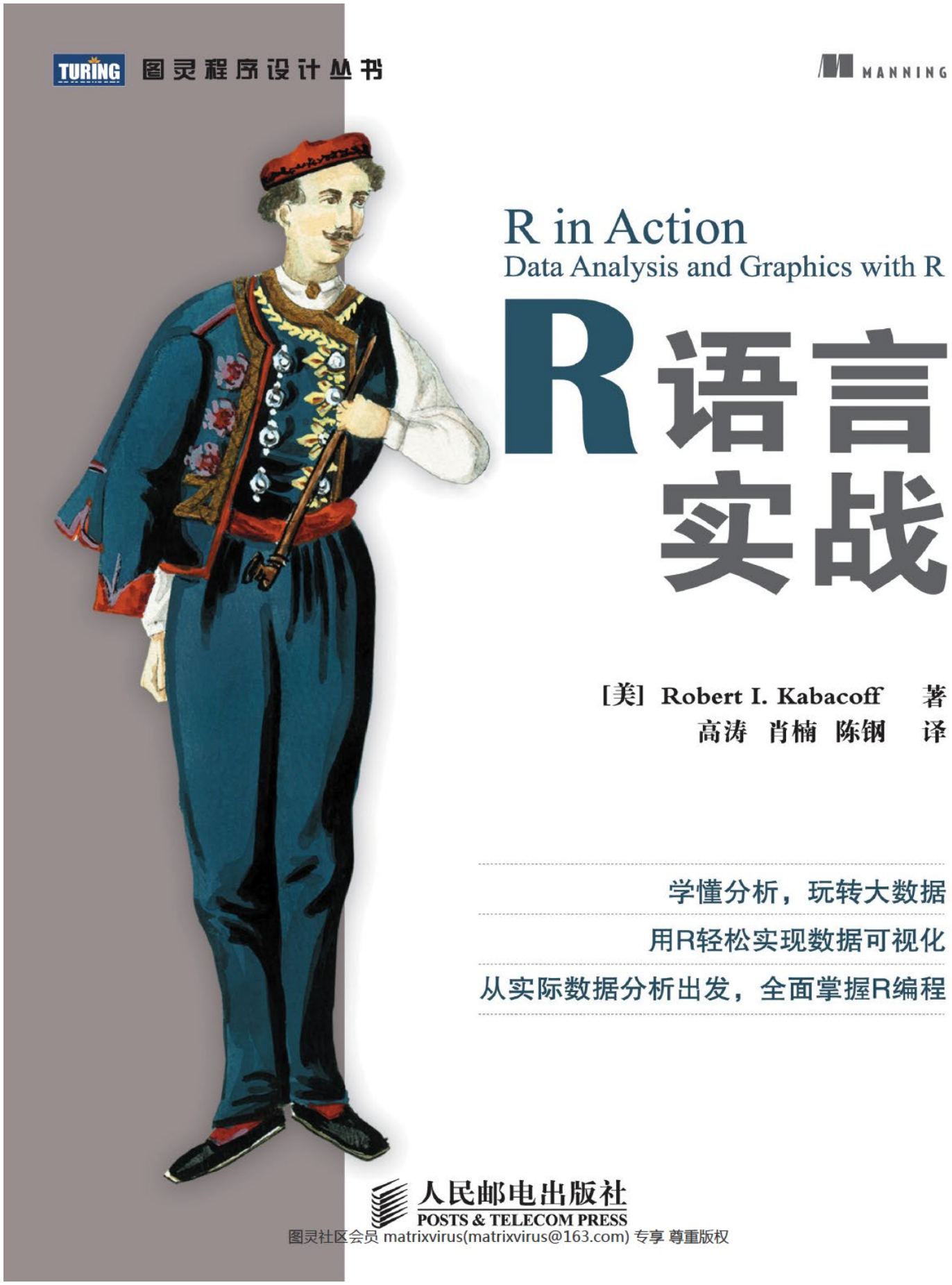
变量重要性排序


1.数据预处理
1.1数据转换
1.2预测变量筛选
2.数据集构建
2.1创建预测变量集和结果变量集
2.2数据划分
数据分割
重抽样
3.数据建模
3.1线性回归模型
广义线性回归模型
逐步回归模型
偏最小二乘法
3.2非线性回归模型
支持向量机SVM
K近邻
3.3回归树模型
简单回归树
回归模型树
随机森林
cubist
4.模型优选
R2评价
RMSE评价
5.模型验证

建模流程

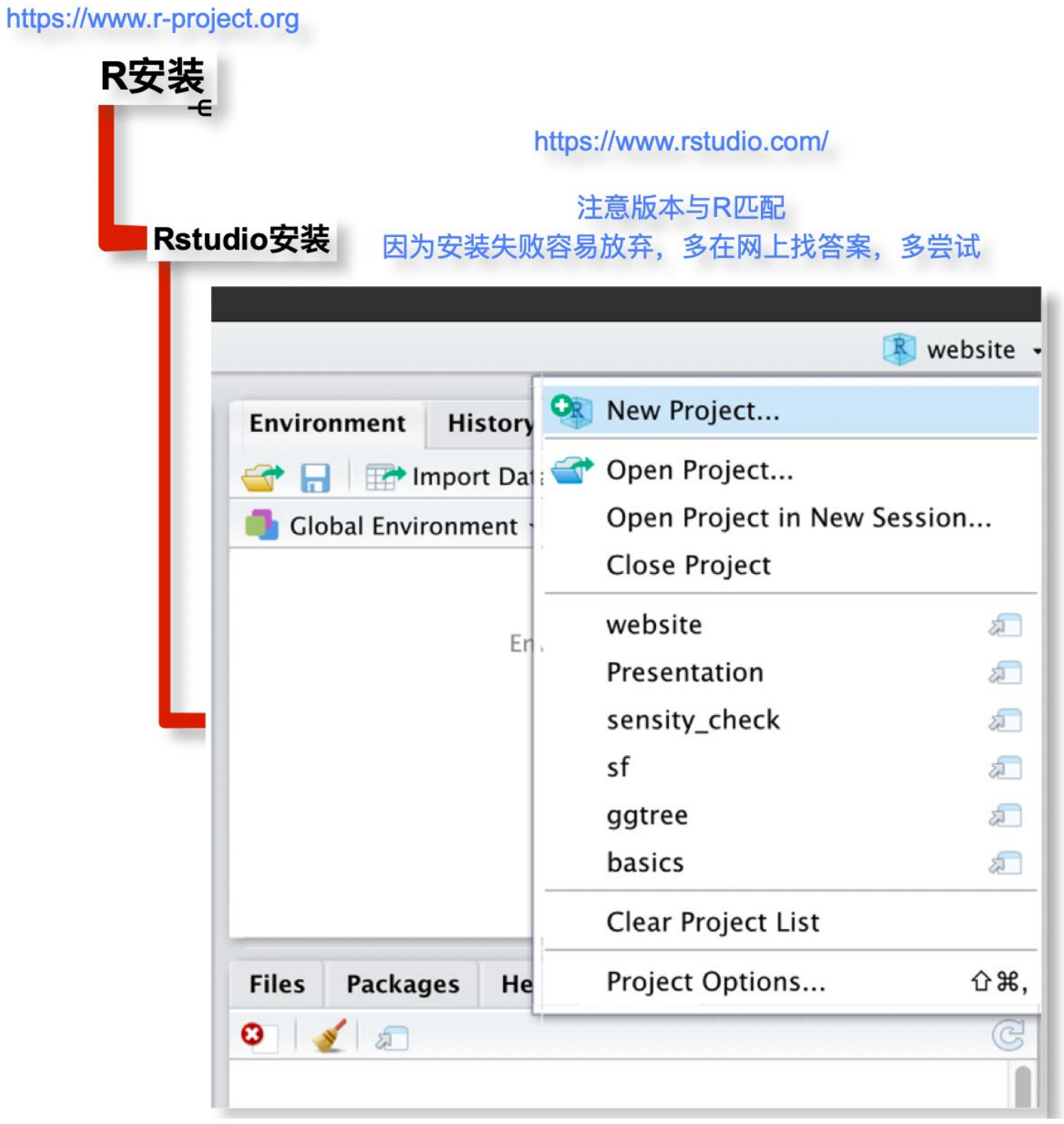
6.Workflow

一切围绕《R语言实战》查询



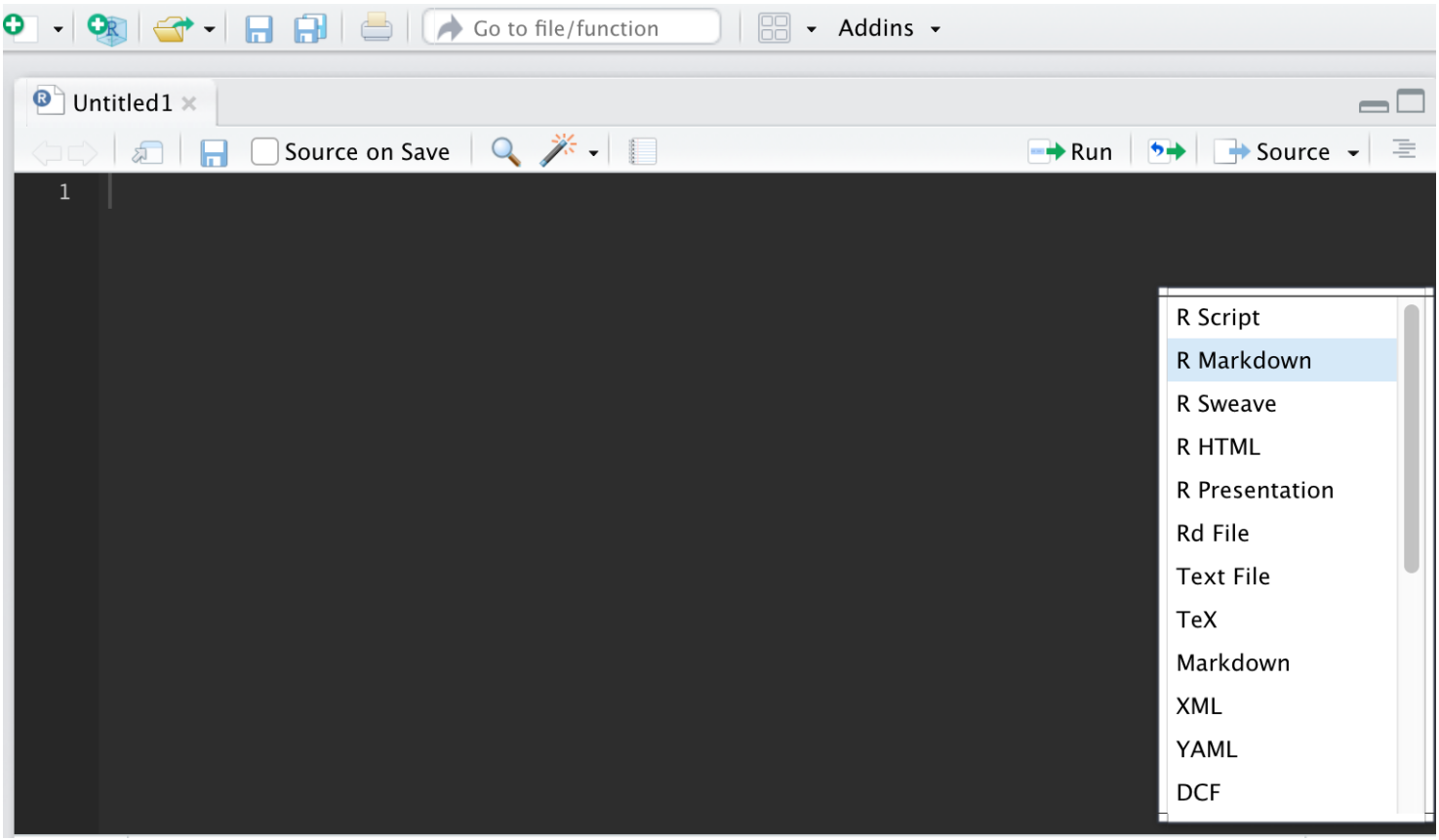
R语言实战

6.1 软件安装



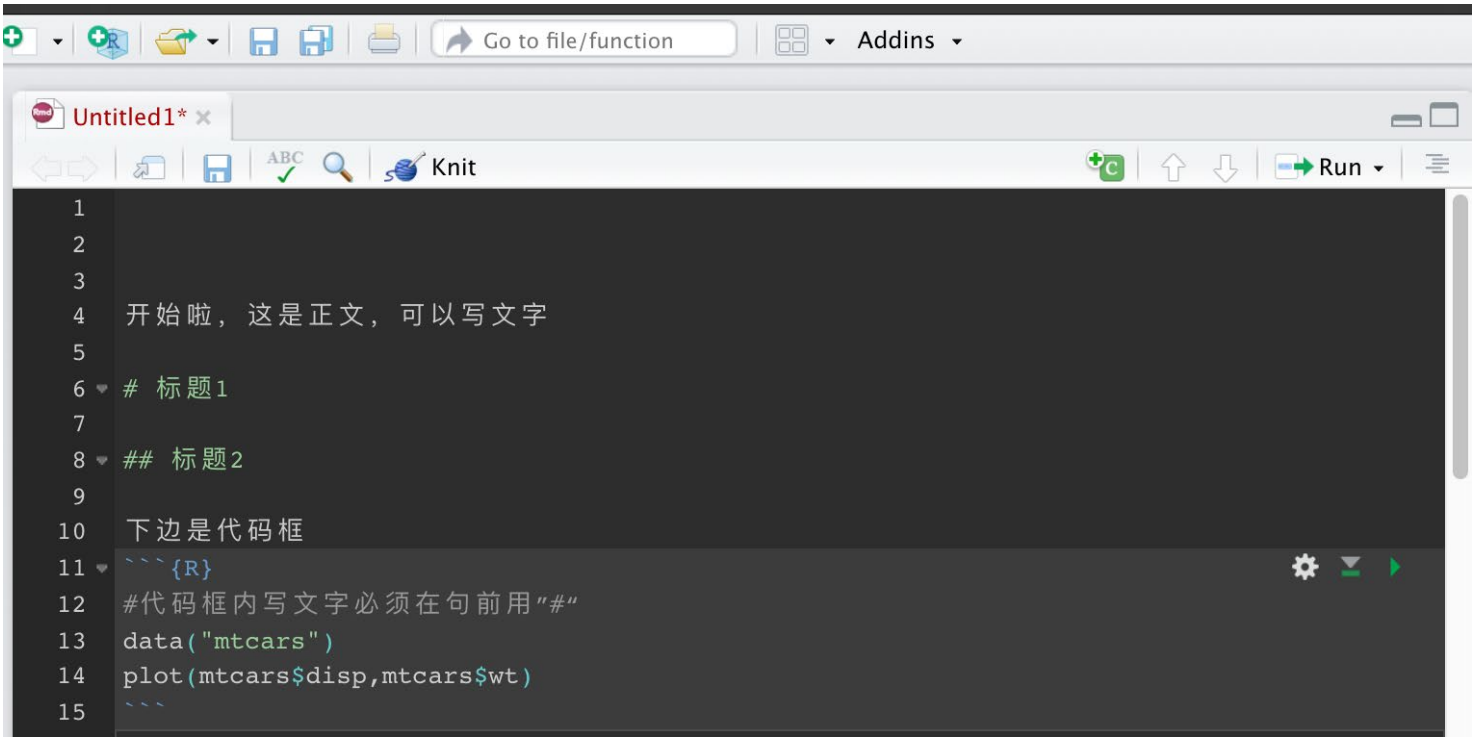
1

6.1 新建项目



2

6.2 使用RMarkdown



3

6.3 数据处理

6.3.1 导入数据

```
#install.packages("readxl")
library(readxl)
cardata <- read_excel("/Users/profits/Rdata/website/cardata.xlsx")
head(cardata)
```

```
## # A tibble: 6 x 12
##   Cartype        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Mazda RX4     21       6   160   110  3.9   2.62  16.5     0     1     4     4
## 2 Mazda RX4 W…  21       6   160   110  3.9   2.88  17.0     0     1     4     4
## 3 Datsun 710    22.8     4   108    93  3.85  2.32  18.6     1     1     4     1
## 4 Hornet 4 Dr…  21.4     6   258   110  3.08  3.22  19.4     1     0     3     1
## 5 Hornet Spor…  18.7     8   360   175  3.15  3.44  17.0     0     0     3     2
## 6 Valiant       18.1     6   225   105  2.76  3.46  20.2     1     0     3     1
```

6.3.3 筛选数据

筛选disp高于120，VS=1的数据

```
library(dplyr)
cardata_filter<-cardata%>%filter(disp>120,vs==1)
cardata_filter
```

```
## # A tibble: 8 x 12
##   Cartype        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Hornet 4 Dr…  21.4     6  258    110  3.08  3.22  19.4     1     0     3     1
## 2 Valiant       18.1     6  225    105  2.76  3.46  20.2     1     0     3     1
## 3 Merc 240D     24.4     4  147.    62  3.69  3.19  20       1     0     4     2
## 4 Merc 230      22.8     4  141.    95  3.92  3.15  22.9     1     0     4     2
## 5 Merc 280      19.2     6  168.   123  3.92  3.44  18.3     1     0     4     4
## 6 Merc 280C     17.8     6  168.   123  3.92  3.44  18.9     1     0     4     4
## 7 Toyota Coro…  21.5     4  120.    97  3.7   2.46  20.0     1     0     3     1
## 8 Volvo 142E    21.4     4  121    109  4.11  2.78  18.6     1     1     4     2
```

选取变量mpg和cyl

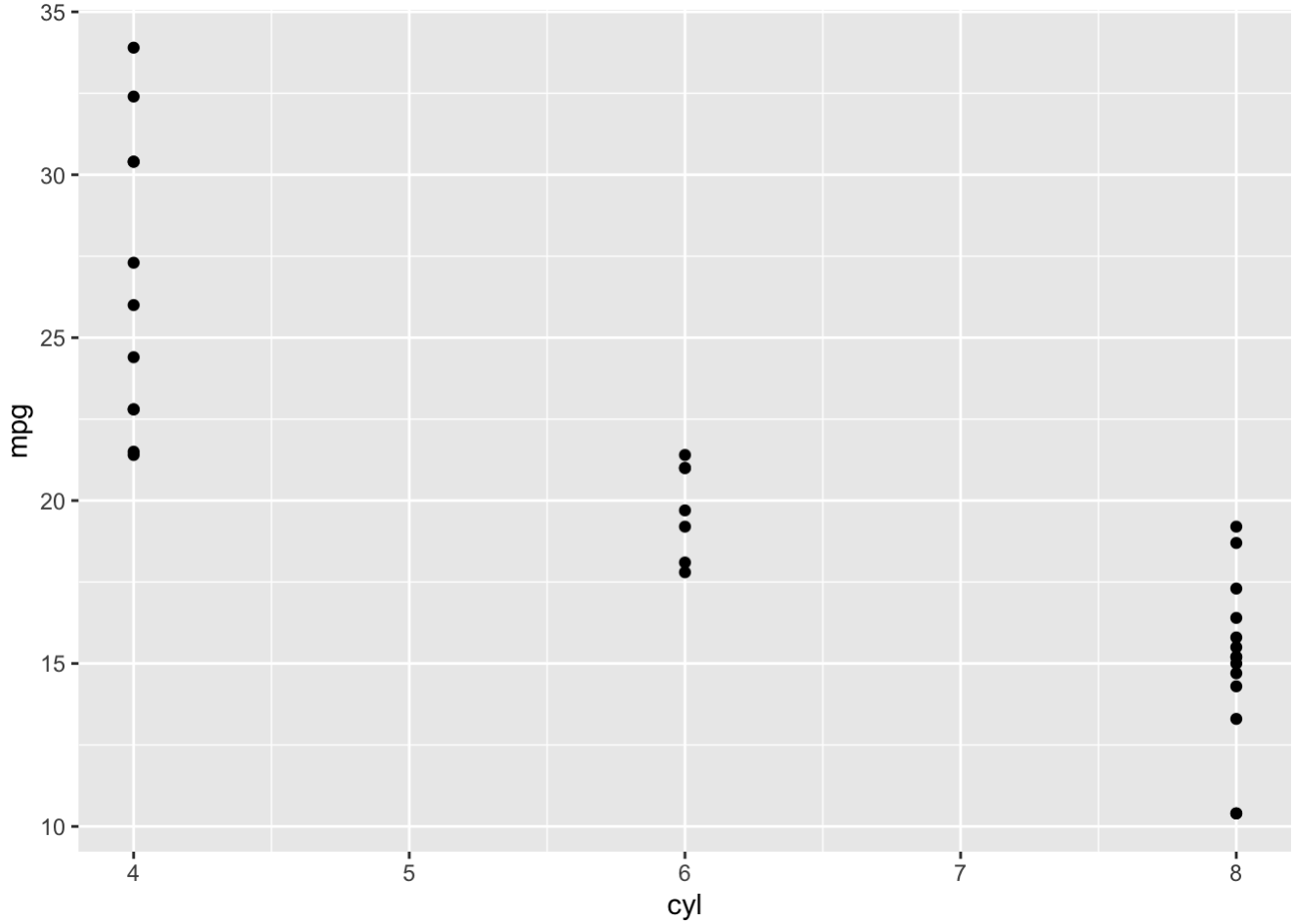
```
cardata_select<-cardata%>%select(mpg,cyl)
cardata_select
```

```
## # A tibble: 32 x 2
##      mpg   cyl
##    <dbl> <dbl>
##  1  21       6
##  2  21       6
##  3  22.8     4
##  4  21.4     6
##  5  18.7     8
##  6  18.1     6
##  7  14.3     8
##  8  24.4     4
##  9  22.8     4
## 10  19.2     6
## # … with 22 more rows
```

6.3.4 数据展示

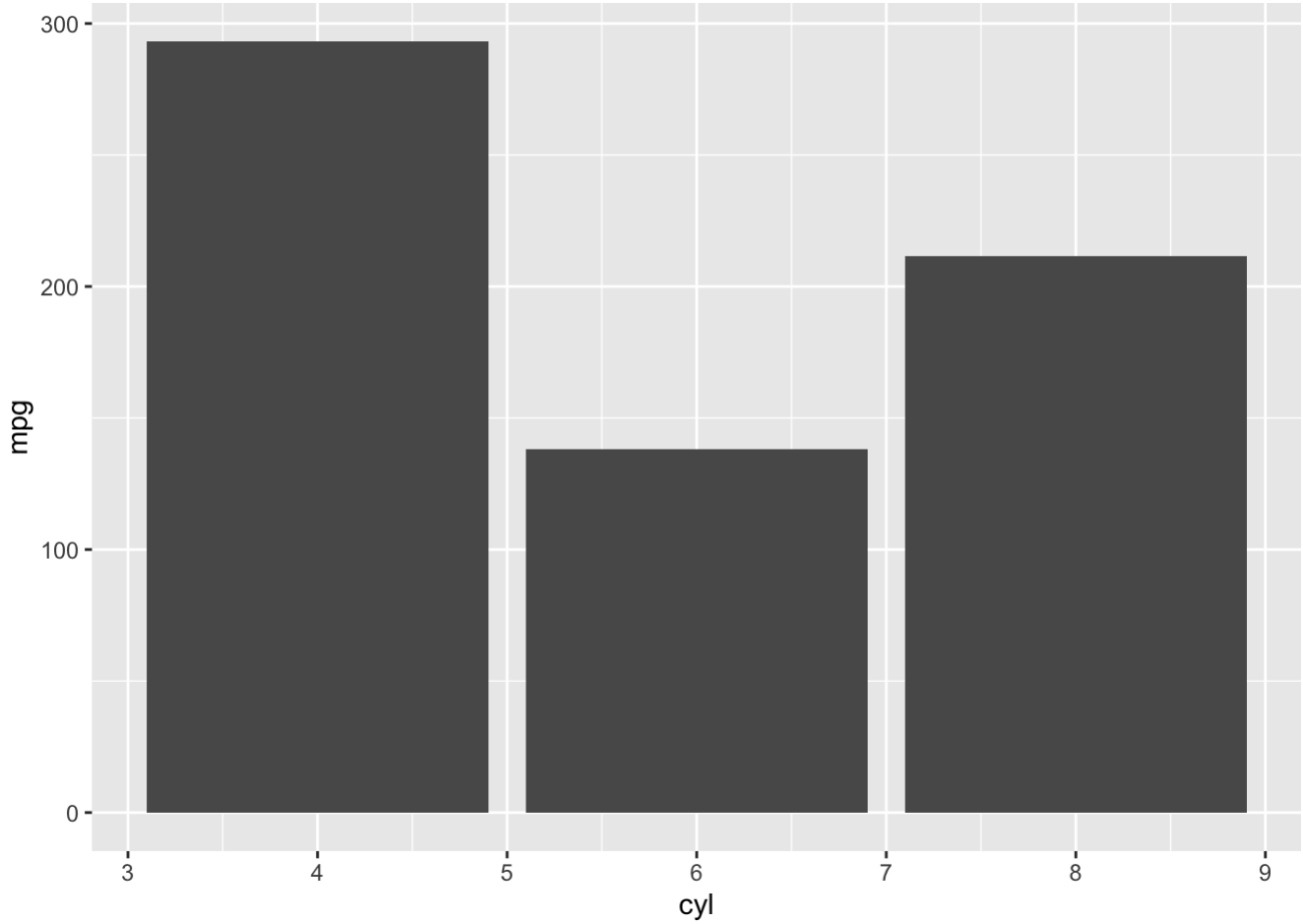
散点图

```
library(ggplot2)
ggplot(cardata,aes(x=cyl,y=mpg))+geom_point()
```



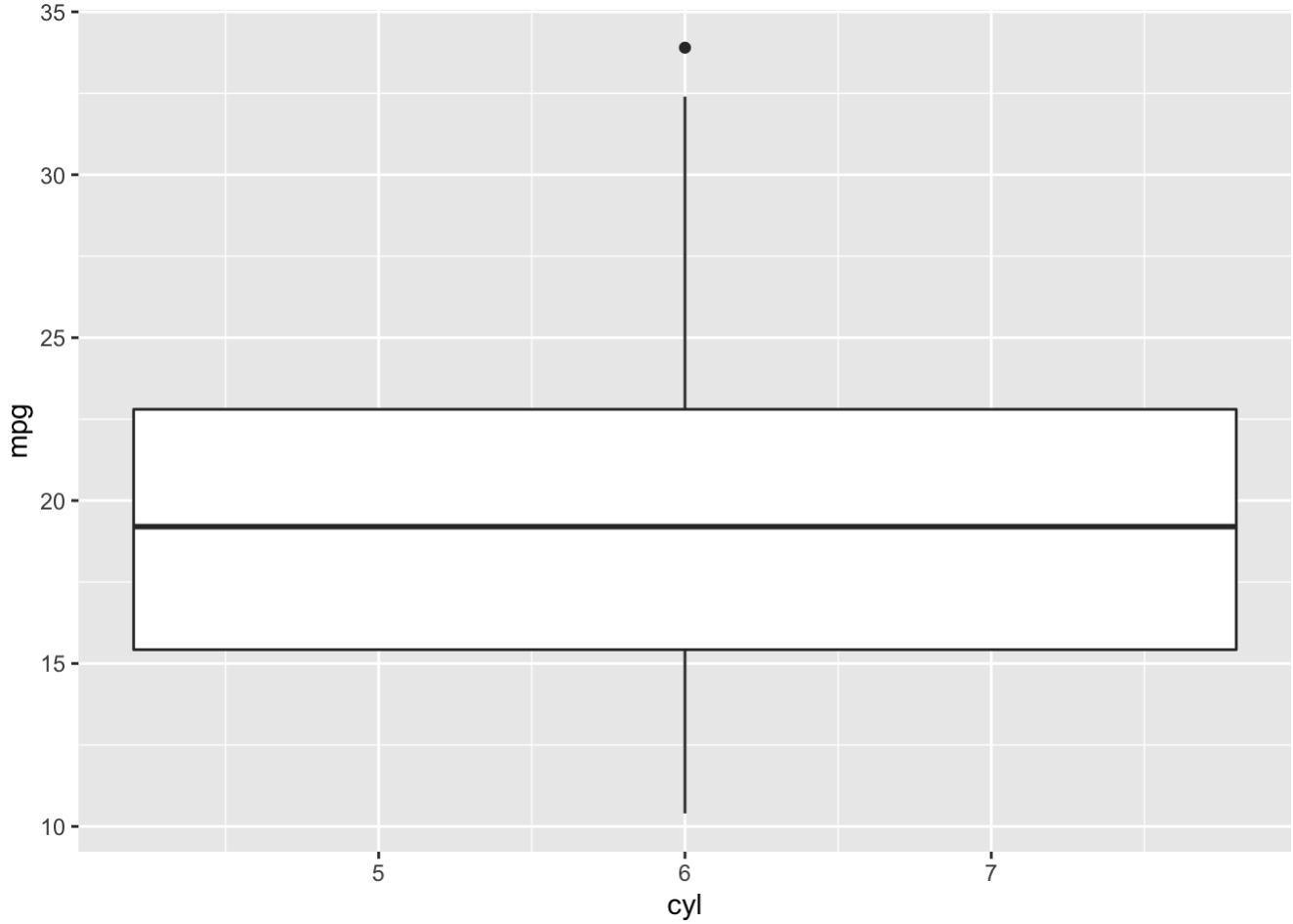
柱状图

```
ggplot(cardata,aes(x=cyl,y=mpg))+geom_col()
```



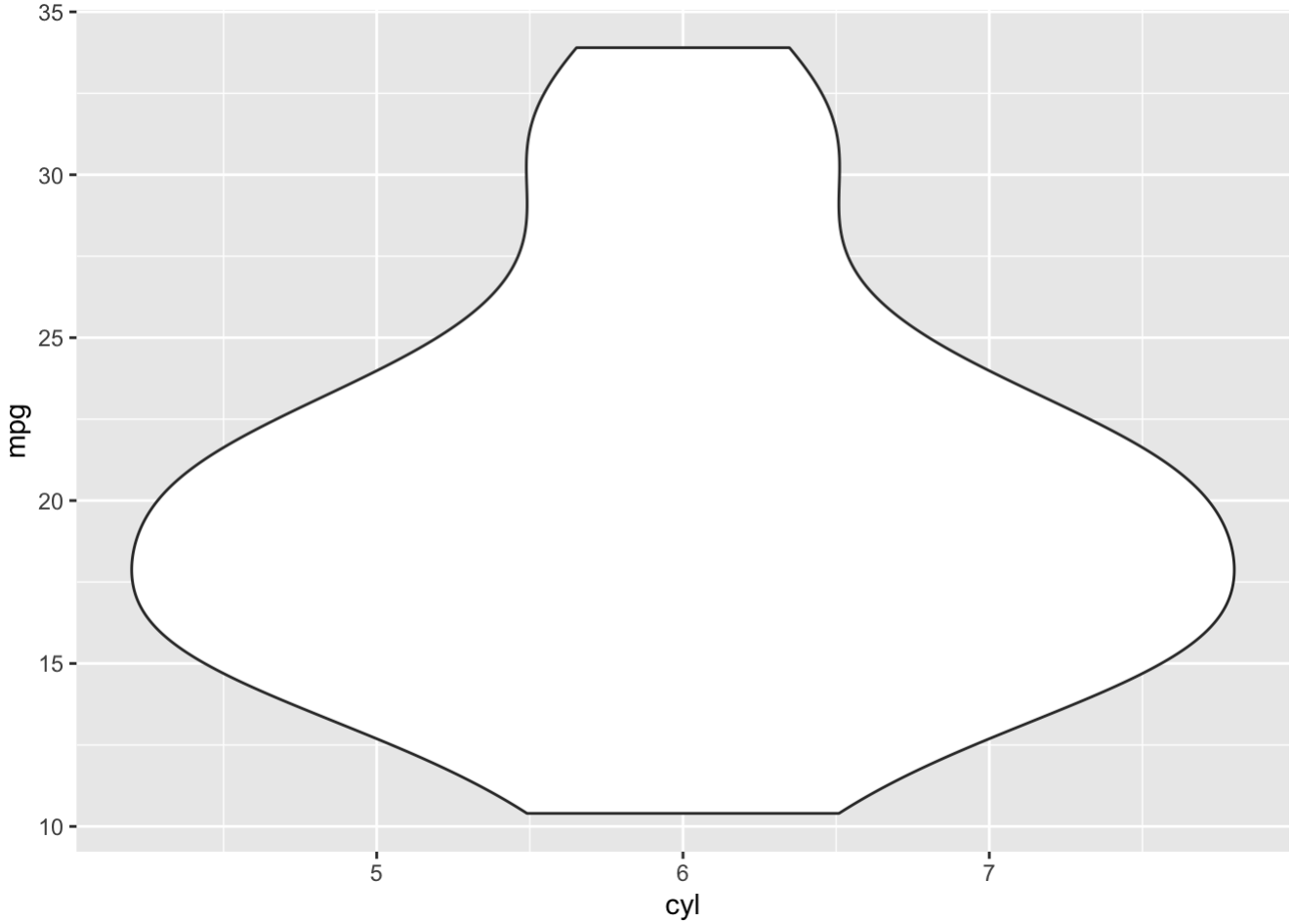
箱式图

```
ggplot(cardata,aes(x=cyl,y=mpg))+geom_boxplot()
```



小提琴图

```
ggplot(cardata,aes(x=cyl,y=mpg))+geom_violin()
```



6.3.5 数据分析

```
#anova
cardata$cyl<-as.factor(cardata$cyl)
anova_fit<-aov(cardata$mpg~cardata$cyl)
summary(anova_fit)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## cardata$cyl   2  824.8   412.4    39.7 4.98e-09 ***
## Residuals    29  301.3    10.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

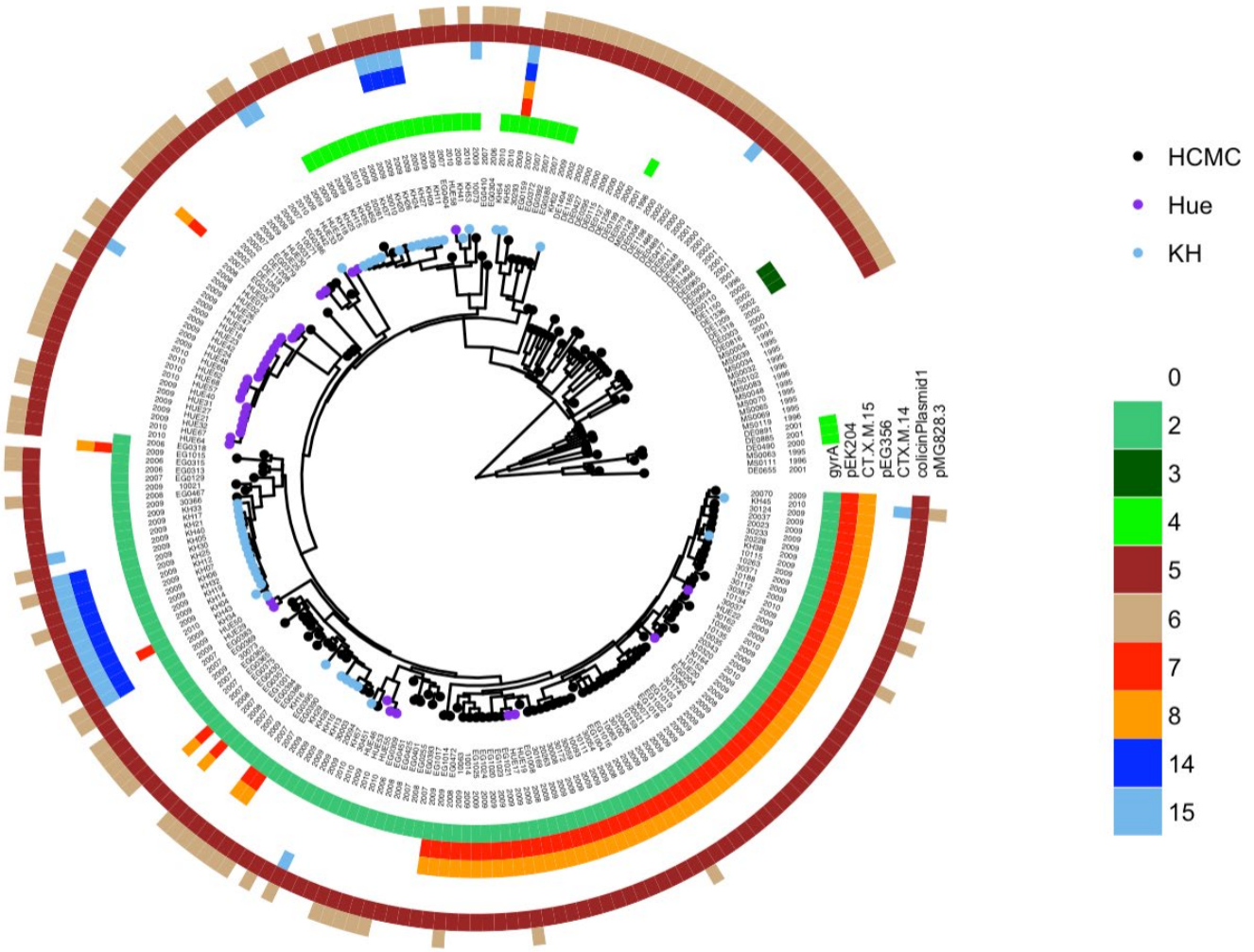
```
#lm_fit
lm_fit<-lm(mpg~cyl,cardata)
summary.aov(lm_fit)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
```

```
## cyl           2  824.8   412.4    39.7 4.98e-09 ***
## Residuals    29  301.3    10.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.4 更进一步的分析?

多重比较、线性混合效应模型、更好看的图，网上查找教程。比如，我想做个这样的聚类图:



网上找 教程

7.不断学习、分享、总结


只争朝夕-科研数据分析学习小组
该二维码7天内(8月12日前)有效，重新进入将更新