



ggplot2:让数据说话

徐洲更

植物生理生态研究所 王佳伟课题组



为什么我们需要作图

为什么我们需要作图

文字太过苍白

拟南芥一年生细弱草本，高20-35厘米，被单毛与分枝毛。茎不分枝或自中上部分枝，下部有时为淡紫白色，茎上常有纵槽，上部无毛，下部被**单毛**，偶杂有2叉毛。基生叶莲座状，倒卵形或匙形，长1-5厘米，宽3-15毫米，顶端钝圆或略急尖，基部渐窄成柄，边缘有少数不明显的齿，两面均有2-3叉毛；茎生叶无柄，**披针形、条形、长圆形或椭圆形**，长5-15 (-50) 毫米，宽1-2 (-10) 毫米。花序为疏松的总状花序，结果时可伸长达20厘米；萼片长圆卵形，长约1.5毫米，顶端钝、外轮的基部成囊状，外面无毛或有少数单毛；花瓣白色，长圆条形，长2-3毫米，先端钝圆，基部线形。角果长10-14毫米，宽不到1毫米，果瓣两端钝或钝圆，有1中脉与稀疏的网状脉，多为桔黄色或淡紫色；果梗伸展，长3-6毫米。

为什么我们需要作图

容易理解
容易记忆



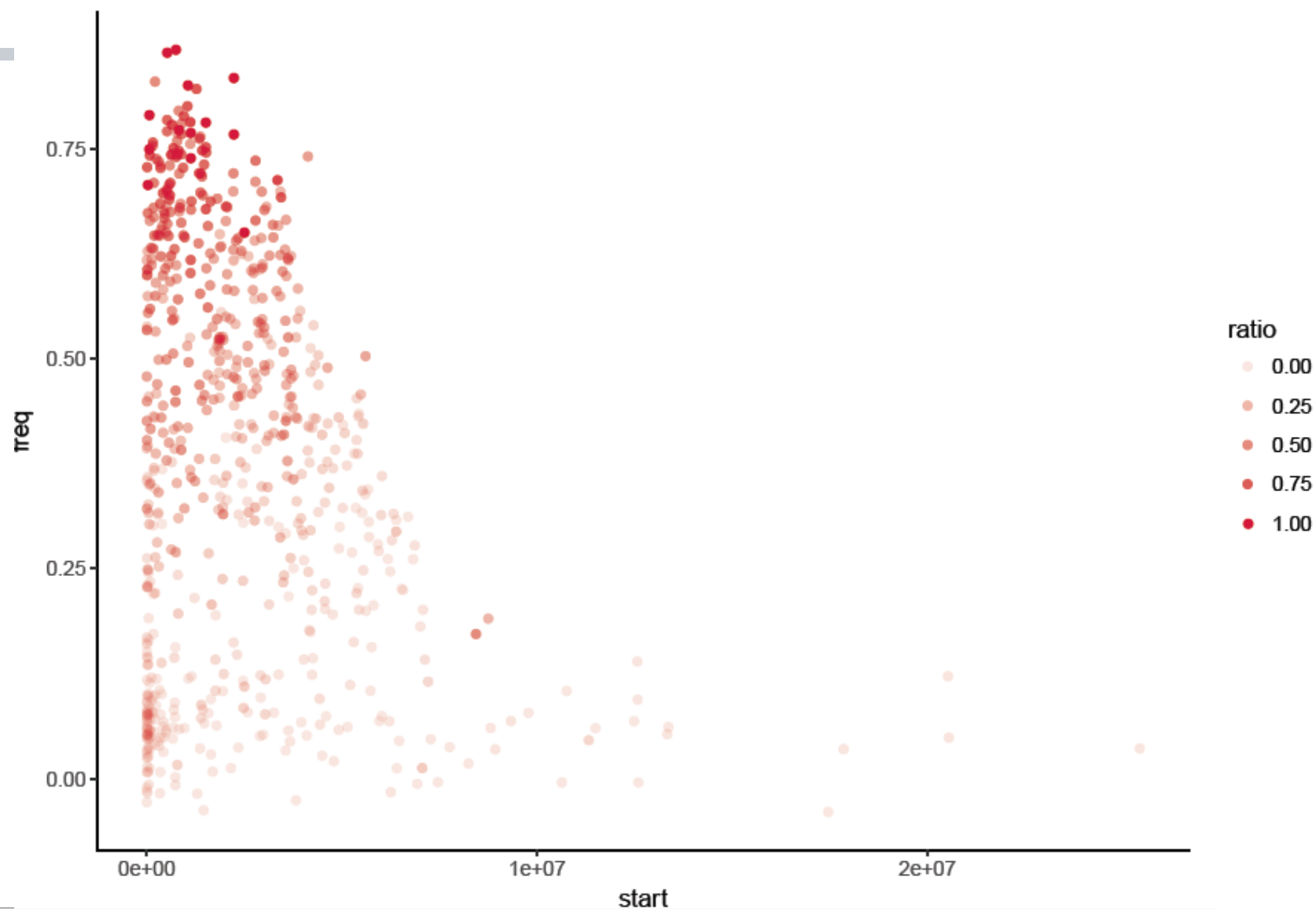
为什么我们需要作图

文字太过苍白

qry	chrom	start	end	freq	length	count	ratio
ctg_10007	Chr1	3837	20315687	0.0695767762	25332	5	0.013
ctg_10027	Chr2	4229483	17464078	0.1050741281	25205	4	0.004
ctg_10065	Chr3	1557701	8345434	0.0692522328	25141	1	0.003
ctg_10129	Chr3	10140423	12456958	0.0642596187	25049	3	0.003
ctg_10373	Chr1	222306	20100141	0.5320894742	24674	76	0.284
ctg_104245	Chr1	1661106	1662377	0.2070945987	1395	3	0.176
ctg_104345	Chr3	9807690	9814663	0.1534232030	1392	1	0.167
ctg_104423	Chr1	2944870	2946258	0.3476775707	1389	1	0.125
ctg_10463	Chr1	2029452	18431990	0.4059111658	24523	2	0.010
ctg_10507	Chr3	3226138	23612647	0.1245147233	24474	3	0.005
ctg_10515	Chr1	56416	19574686	0.1648044292	24464	2	0.007
ctg_10527	Chr2	32230	15312612	-0.0007113784	24440	1	0.005
ctg_105361	Chr1	1868462	1869829	0.5229219727	1364	5	0.556
ctg_105689	Chr1	2785765	2787034	0.6644014384	1355	3	0.600
ctg_106404	Chr1	830511	831404	0.7200007906	1337	4	0.500
ctg_107956	Chr1	3442147	3443444	0.6918304141	1298	3	0.750
ctg_10807	Chr1	6253640	12391446	-0.0161764440	24088	2	0.013
ctg_1081	Chr3	177282	21156973	0.0364846205	70774	1	0.001
ctg_10811	Chr4	7687573	19335517	0.0774379815	24176	1	0.004
ctg_108476	Chr1	783176	784460	0.0162210218	1286	1	0.091
ctg_10867	Chr1	4429519	11013894	0.0949277013	24097	7	0.042
ctg_1087	Chr1	70195	19869538	0.0122903658	70623	1	0.001
ctg_10015	Chr2	155506	15561524	0.0630251702	24025	3	0.004

为什么我们需要作图

人类拥有最强的
模式识别能力





为什么要用R/ggplot2

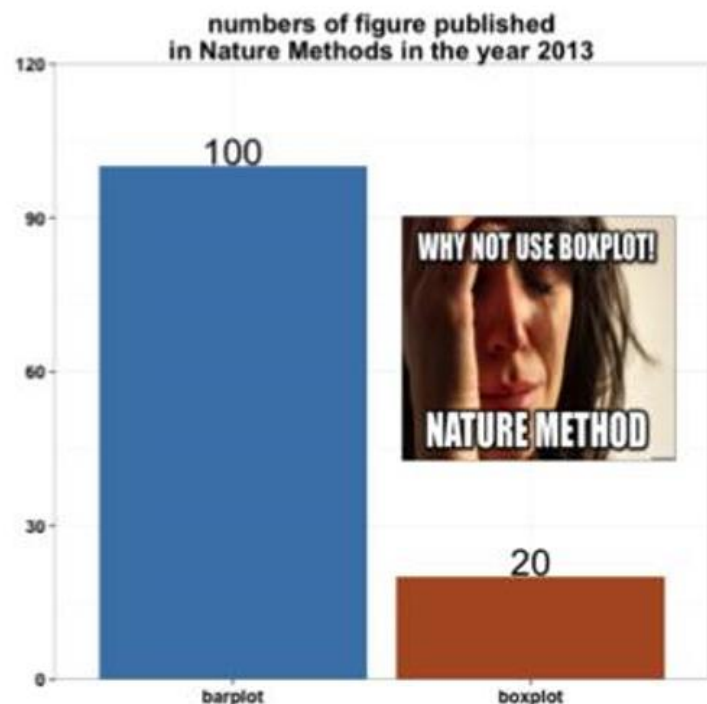
R是开源免费软件



- 1.R 是一个有着统计分析功能及强大作图功能的软件系统，是由新西兰奥克兰大学统计学系的 Ross Ihaka 和 Robert Gentleman 共同创立，可能是因为两人首字母都是R所以就叫R语言。
- 2.R语言是一门统计语言，天生就不同于其他的编程语言。R语言封装了各种基础学科的计算函数，我们在R语言编程的过程中只需要调用这些计算函数，就可以构建出面向不同领域、不同业务的、复杂的数学模型。

Excel是**收费**软件

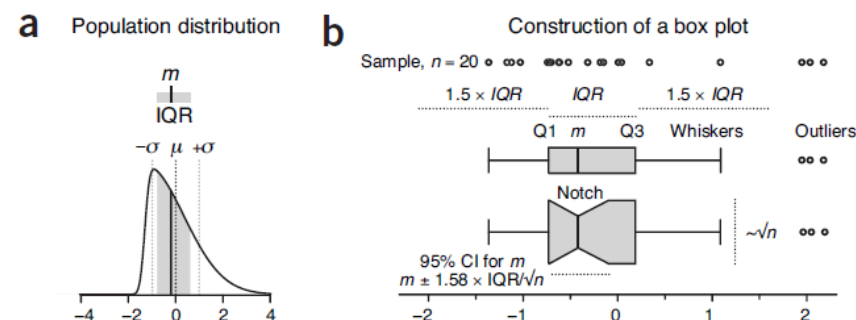
Excel限制了你的想象力



Visualizing samples with box plots

Use box plots to illustrate the spread and differences of samples.

Visualization methods enhance our understanding of sample data and help us make comparisons across samples. Box plots are a simple but powerful graphing tool that can be used in place of histograms to address both goals. Whereas histograms require a sample size of at least 30 to be useful, box plots require a sample size of only 5, provide more detail in the tails of the distribution and are more readily compared across three or more samples. Several enhancements to the basic box plot can render it even more informative.



MENU **nature|methods**

Access provided by Shanghai Institutes for Biological Sciences, CAS

Altmetric: 147 Citations: 125

[More detail >>](#)

Correspondence

BoxPlotR: a web tool for generation of box plots

Michaela Spitzer, Jan Wildenhain, Juri Rappsilber & Mike Tyers

<http://shiny.chemgrid.org/boxplotr/>



是不是稍微
引起了你的注意力



如何姿势正确的入门R语言

什么**最**重要？ —————→ 解决问题最重要

什么**更**重要？ —————→ 怎么解决更重要

入门R语言的核心技巧

1. 安装R和Rstudio
2. 打开Rstudio
3. 导入Excel数据
4. 搜索R包并安装
5. 查看R包使用说明
6. 调用R包清洗数据
7. 调用R包可视化作图
8. 导出PDF
9. 保存代码，关闭Rstudio
10. 用PS AI修图

R和Rstudio不是一回事

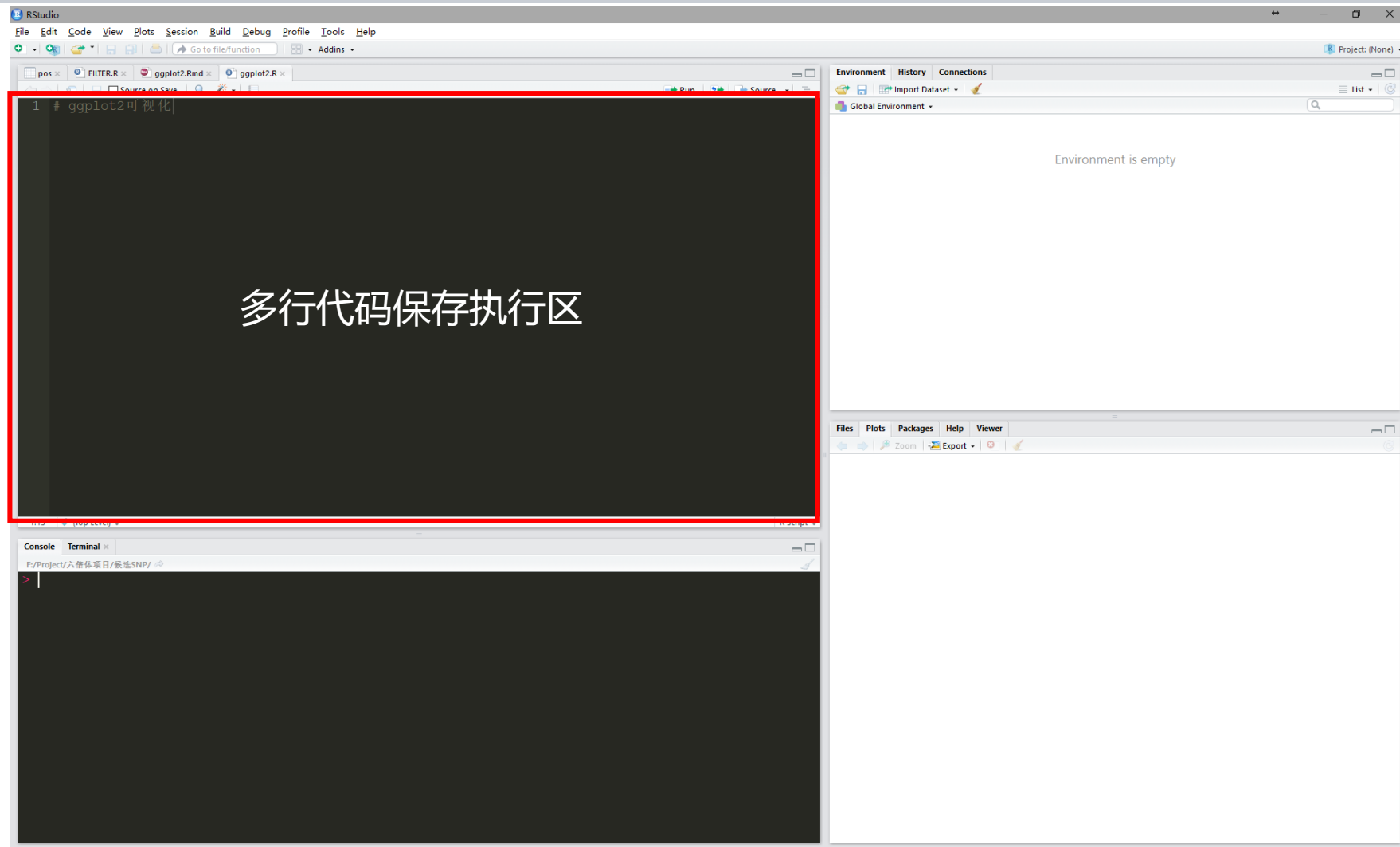


R是一门编程语言，执行R代码的唯一地方
<https://www.r-project.org/>

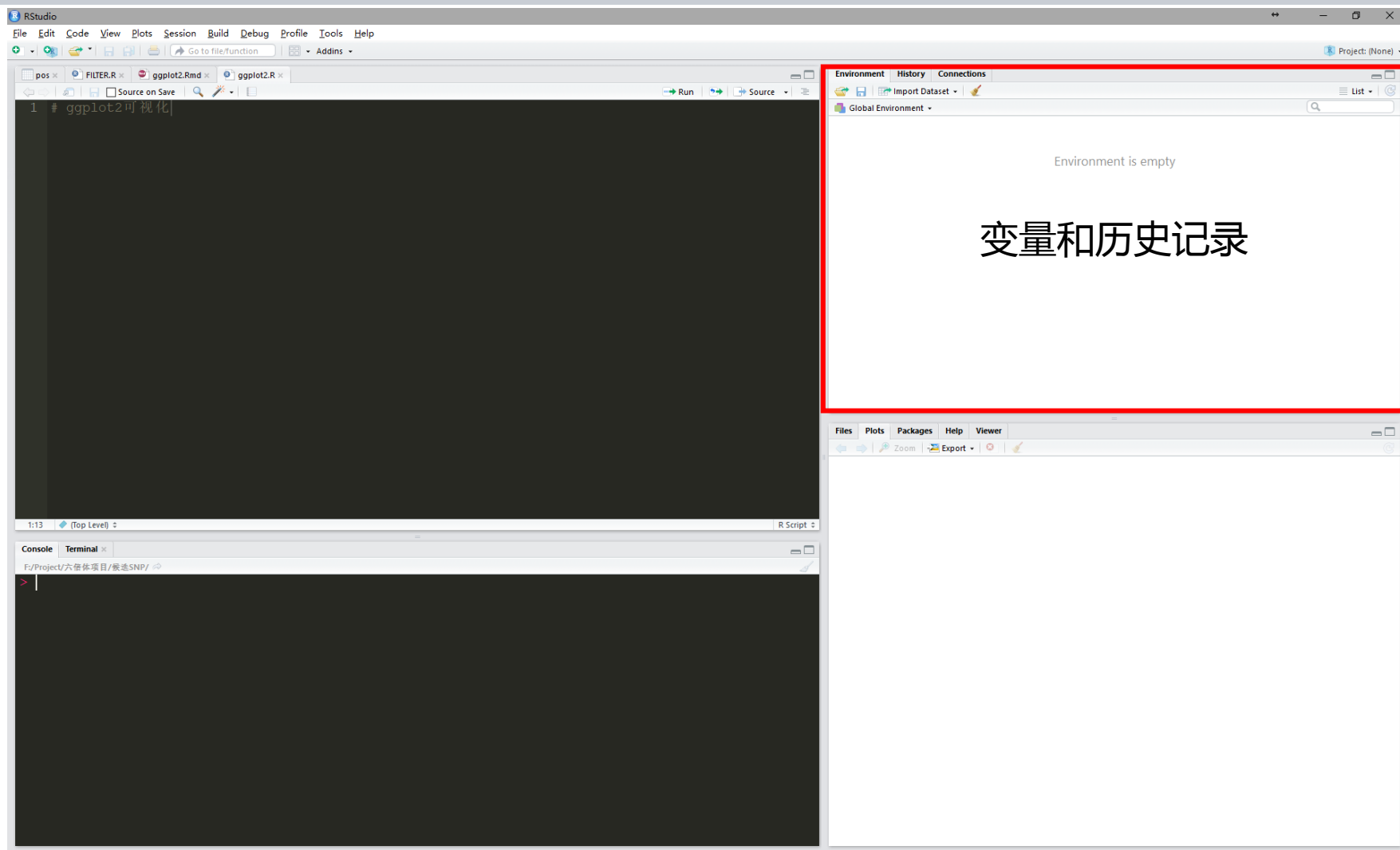


Rstudio是方便写代码的地方之一
<https://www.rstudio.com/>

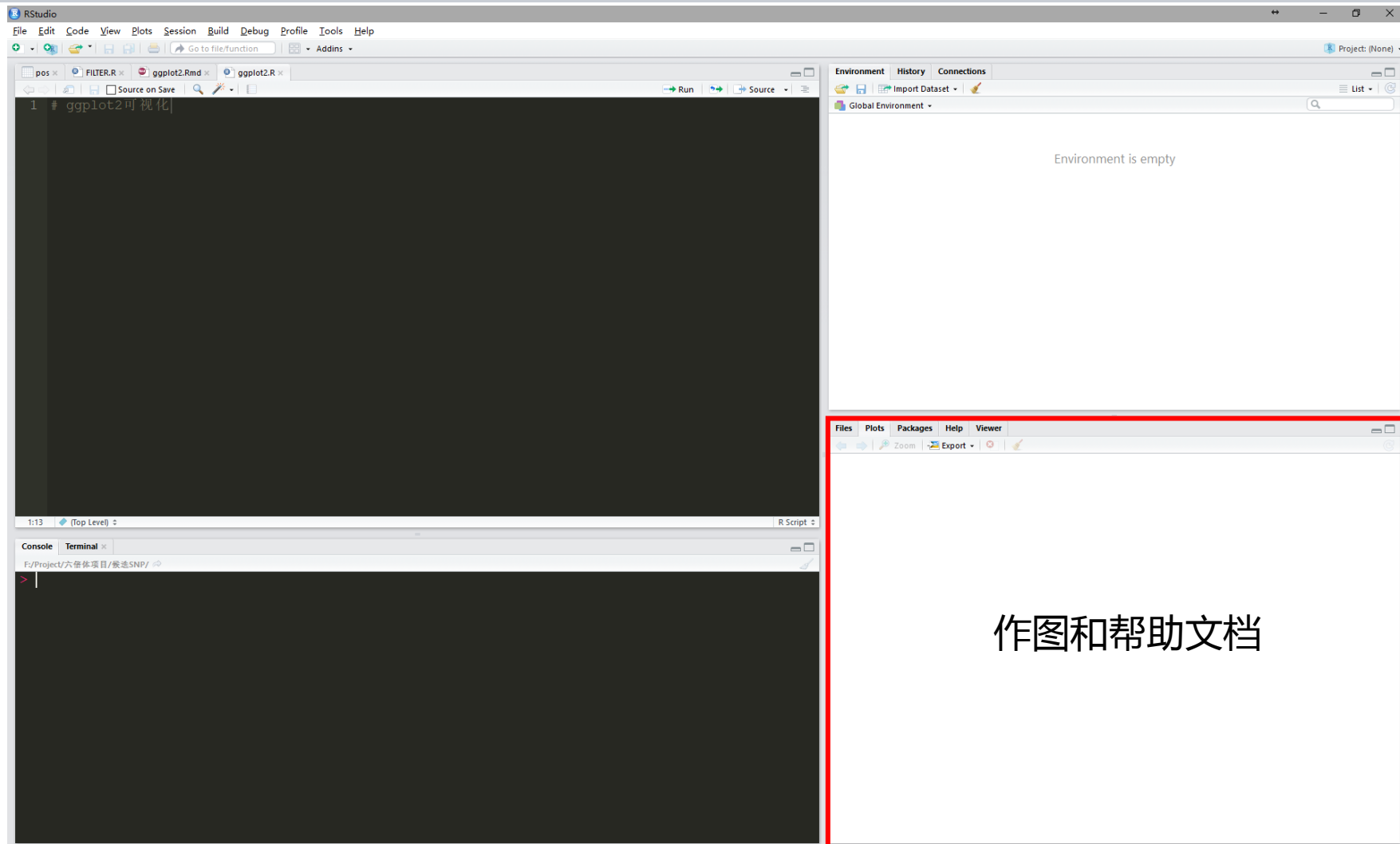
Rstudio界面



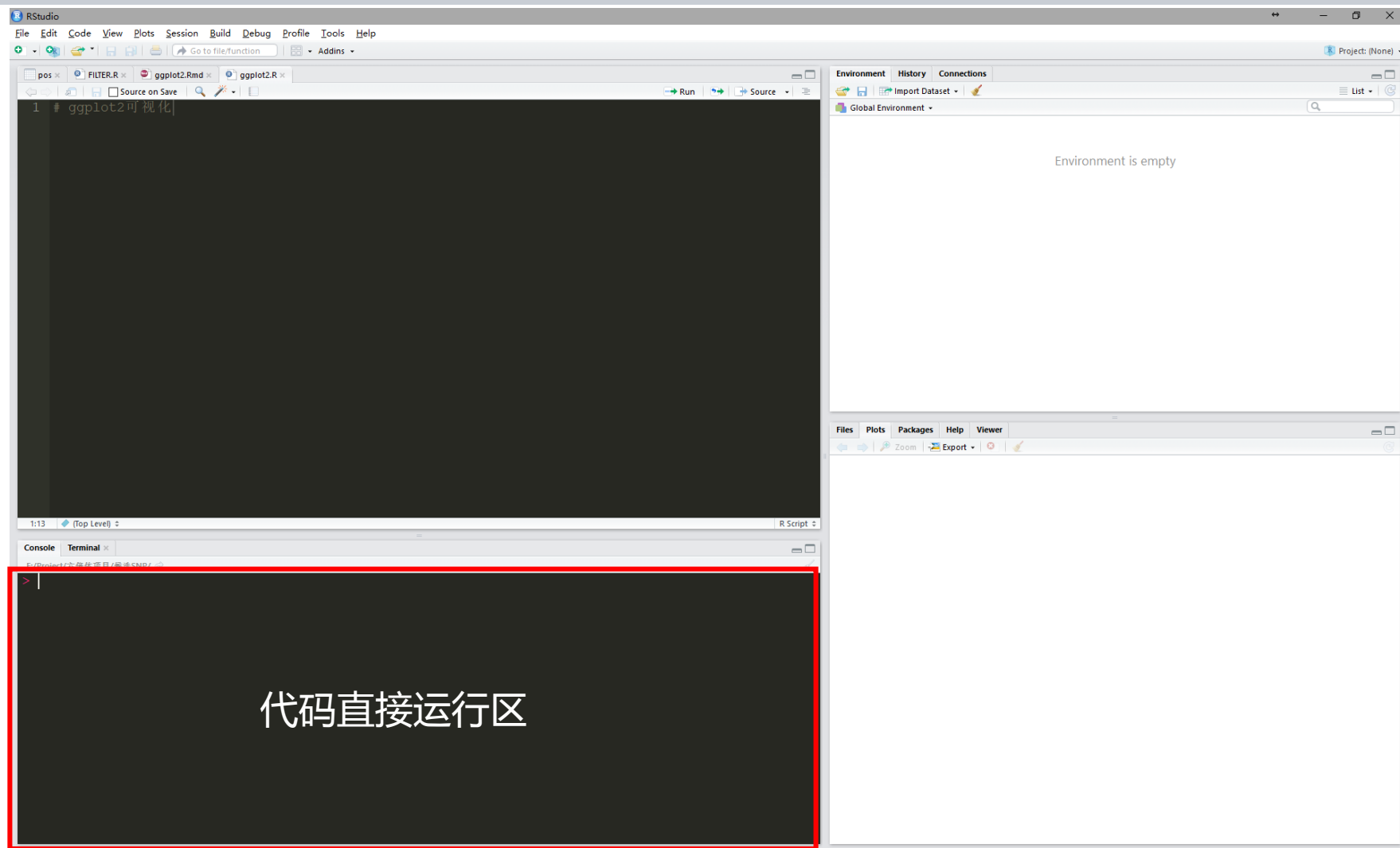
Rstudio界面



Rstudio界面



Rstudio界面



如何导入数据

```
df <- read.csv(file = "dataset.csv", header = TRUE)
```

```
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 = "\"",  
  dec = ".", fill = TRUE, comment.char = "", ...)
```

```
read.csv2(file, header = TRUE, sep = ";", quote = "\"",  
  dec = ",", fill = TRUE, comment.char = "", ...)
```

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

```
read.delim2(file, header = TRUE, sep = "\t", quote = "\"",  
  dec = ",", fill = TRUE, comment.char = "", ...)
```

搜索R包并安装



 [百度一下](#)

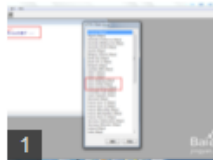
[网页](#) [新闻](#) [贴吧](#) [知道](#) [音乐](#) [图片](#) [视频](#) [地图](#) [文库](#) [更多»](#)

百度为您找到相关结果约2,280,000个 [搜索工具](#)

[R语言如何导入Excel的数据? - 知乎](#)
3个回答 · 提问时间: 2013年04月14日
比较传统的方法是把excel文件保存成csv格式, 即逗号分隔文件 然后用R里面的read.csv读取
其实现在有很多package提供直接读取excel功能 比如说gdata 的read.xls
<https://www.zhihu.com/question...> [V2](#) - 百度快照

[R语言导入EXCEL数据 - CSDN博客](#)
2017年6月9日 - 今天找了好多包来快速读取数量比较大的excel包,但是总是出现
OutOfMemoryError (Java): GC overhead limit exceeded,内存溢出报错。找到了两个R包可...
blog.csdn.net/qq_35242... - 百度快照

[用R语言导入Excel xlsx及xls格式的数据 百度经验](#)



1

首先, 导入R语言需要
加载xlsx包, 没有...



2

安装好xlsx包后, 接下
来导入存放在"C:...



3

我们上面的代码只是把
test.xlsx导入了R...



4

那么如何查看Mydata
里面的数据呢? 只需...

[显示全部](#) 

[jingyan.baidu.com](#) 

搜索R包并安装



任坤

45 人赞同了该回答

以下扩展包可以读取xlsx文件:

- [CRAN - Package xlsx](#)
- [awalker89/openxlsx · GitHub](#)
- [CRAN - Package RODBC](#) 通过 `odbcConnectExcel2007` 连接到xlsx表格, 然后通过SQL选择表格中的数据
- [hadley/readxl · GitHub](#) 最新的一个扩展包, 可以跨平台、无依赖地读取xls, xlsx中的表格数据。

搜索R包并安装

安装R包的两个策略

```
install.packages("tidyverse")  
install.packages("devtools")
```

```
devtools::install_github("tidyverse/readxl")
```

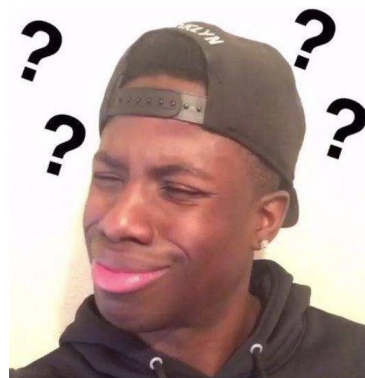


R包安装不求人

查看R包使用说明

载入R包:

```
library(tidyr)  
library(dplyr)  
library(ggplot2)  
....
```



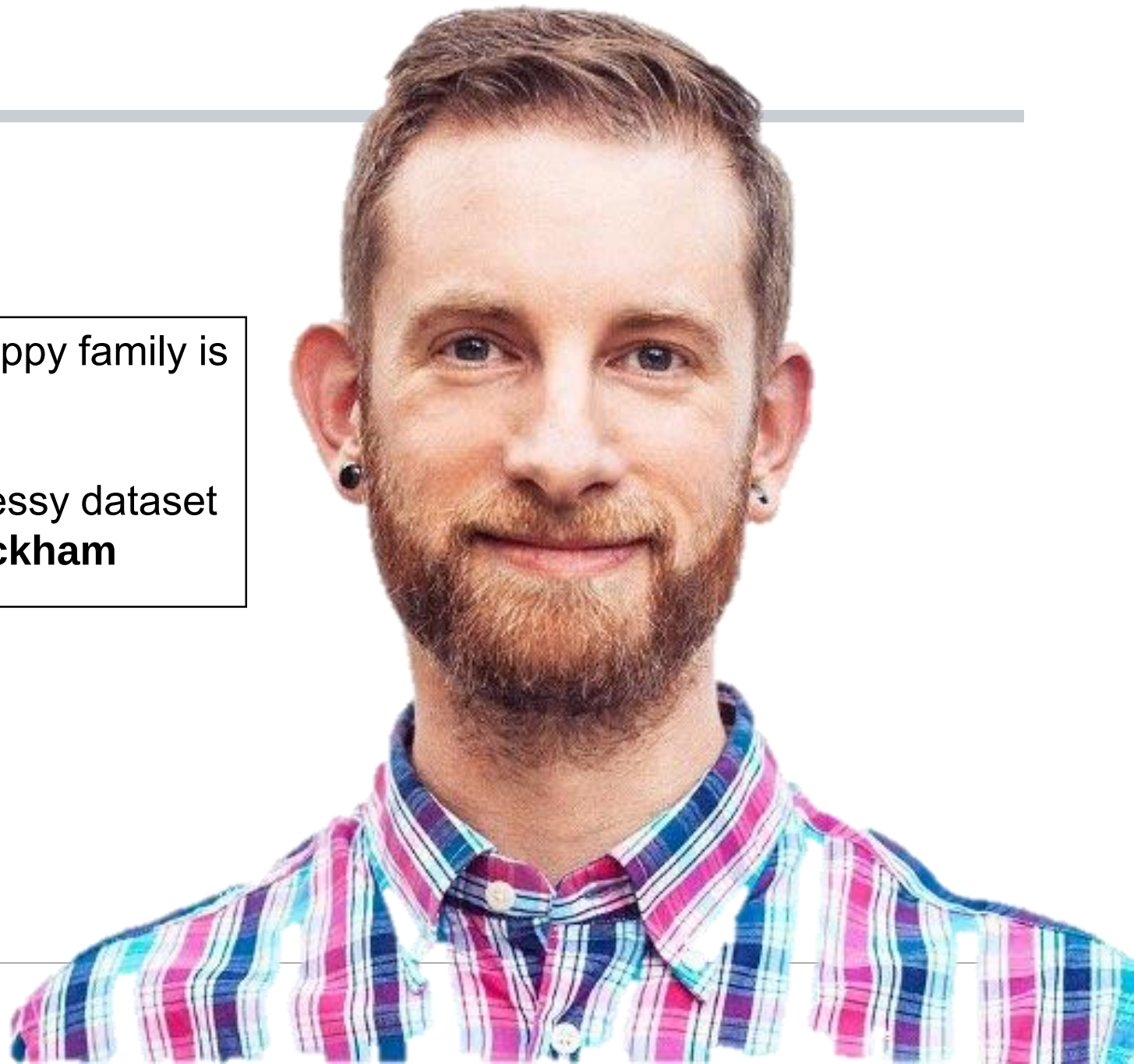
查看帮助

```
?ggplot2  
help(ggplot2)  
??(ggplot2)  
vignette(ggplot2)  
....
```

调用R包清洗数据

“Happy families are all alike; every unhappy family is unhappy in its own way.” — Leo Tolstoy

“Tidy datasets are all alike, but every messy dataset is messy in its own way.” — **Hadley Wickham**



调用R包清洗数据

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

observations

country	year	cases	population
Afghanistan	99	745	19987071
Afghanistan	00	2666	20595360
Brazil	99	37737	172006362
Brazil	00	80488	174504898
China	99	212258	127291272
China	00	213766	128042583

values

- 1.每个变量都要单独成列
- 2.每个观察必须单独成行
- 3.每个位置都只能有一个值

调用R包清洗数据

	country	year	cases	population
	<chr>	<int>	<int>	<int>
1	Afghanistan	1999	745	19987071
2	Afghanistan	2000	2666	20595360
3	Brazil	1999	37737	172006362
4	Brazil	2000	80488	174504898
5	China	1999	212258	1272915272
6	China	2000	213766	1280428583

```
# A tibble: 6 x 3
```

	country	year	rate
*	<chr>	<int>	<chr>
1	Afghanistan	1999	745/19987071
2	Afghanistan	2000	2666/20595360
3	Brazil	1999	37737/172006362
4	Brazil	2000	80488/174504898
5	China	1999	212258/1272915272
6	China	2000	213766/1280428583

	country	year	type	count
	<chr>	<int>	<chr>	<int>
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases	212258
10	China	1999	population	1272915272
11	China	2000	cases	213766
12	China	2000	population	1280428583

如何学会用R清理数据



R的主要数据结构



R语言的数据管理



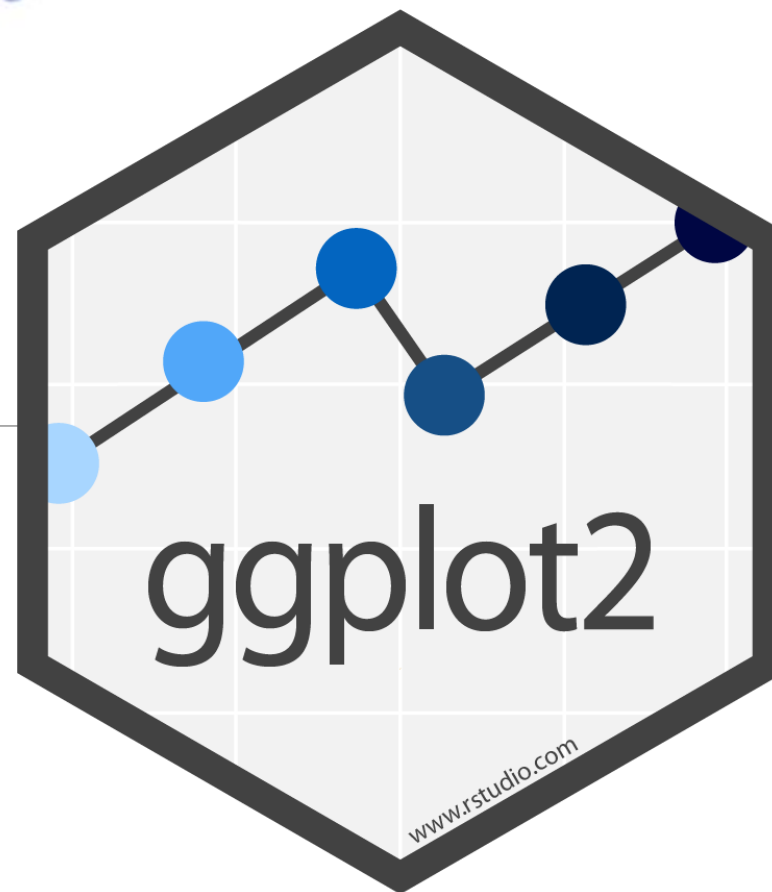
常用函数小抄



使用ggplot2进行数据可视化

一张统计图形就是从**数据**到**几何对象**
的**图形属性**的一个映射。

此外，图形中还可能包含数据的**统计变换**，
最后绘制在**坐标系**中，
而**分面**则可以用来生成数据不同子集的图形



图层就象是含有文字或图形等元素的胶片，一张张按顺序叠放在一起，组合起来形成页面的最终效果

主题

黑白，灰色，彩色

坐标系

笛卡尔坐标系，极坐标

数据变换

线性变换，趋势线

分面

大图分小图

几何对象

点,线,面,框

图形属性

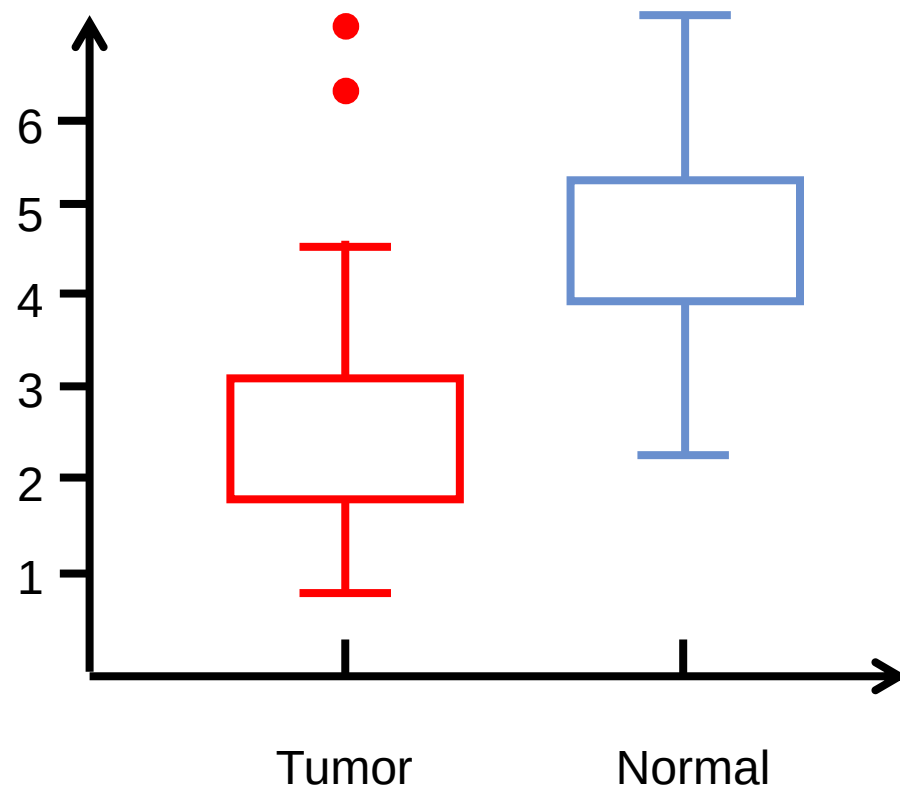
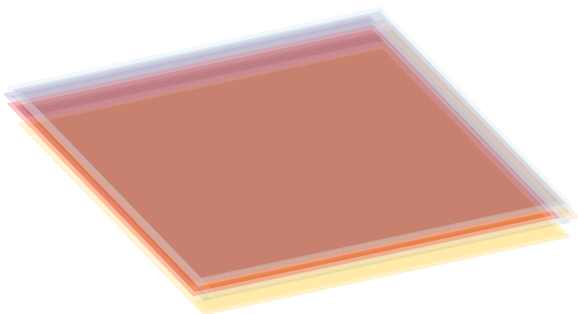
形状,大小,颜色

数据

数据框



图层就象是含有文字或图形等元素的胶片，一张张按顺序叠放在一起，组合起来形成页面的最终效果



ggplot2使用方法

- 定义要绘图的数据，并用ggplot2生成一个空的绘图对象
- 确定图形形状，或者用来展示数据的几何对象
- 用aes函数指定映射关系
- 修缮

ggplot2使用方法

形状

统计变换

坐标系

分面

标度

主题

数据

ggplot (data = **<DATA>**) +

<GEOM_FUNCTION> (mapping = aes(**<MAPPINGS>**),

stat = **<STAT>** , position = **<POSITION>**) +

<COORDINATE_FUNCTION> + 位置

<FACET_FUNCTION> +

<SCALE_FUNCTION> +

<THEME_FUNCTION>

属性映射

required

Not
required,
sensible
defaults
supplied

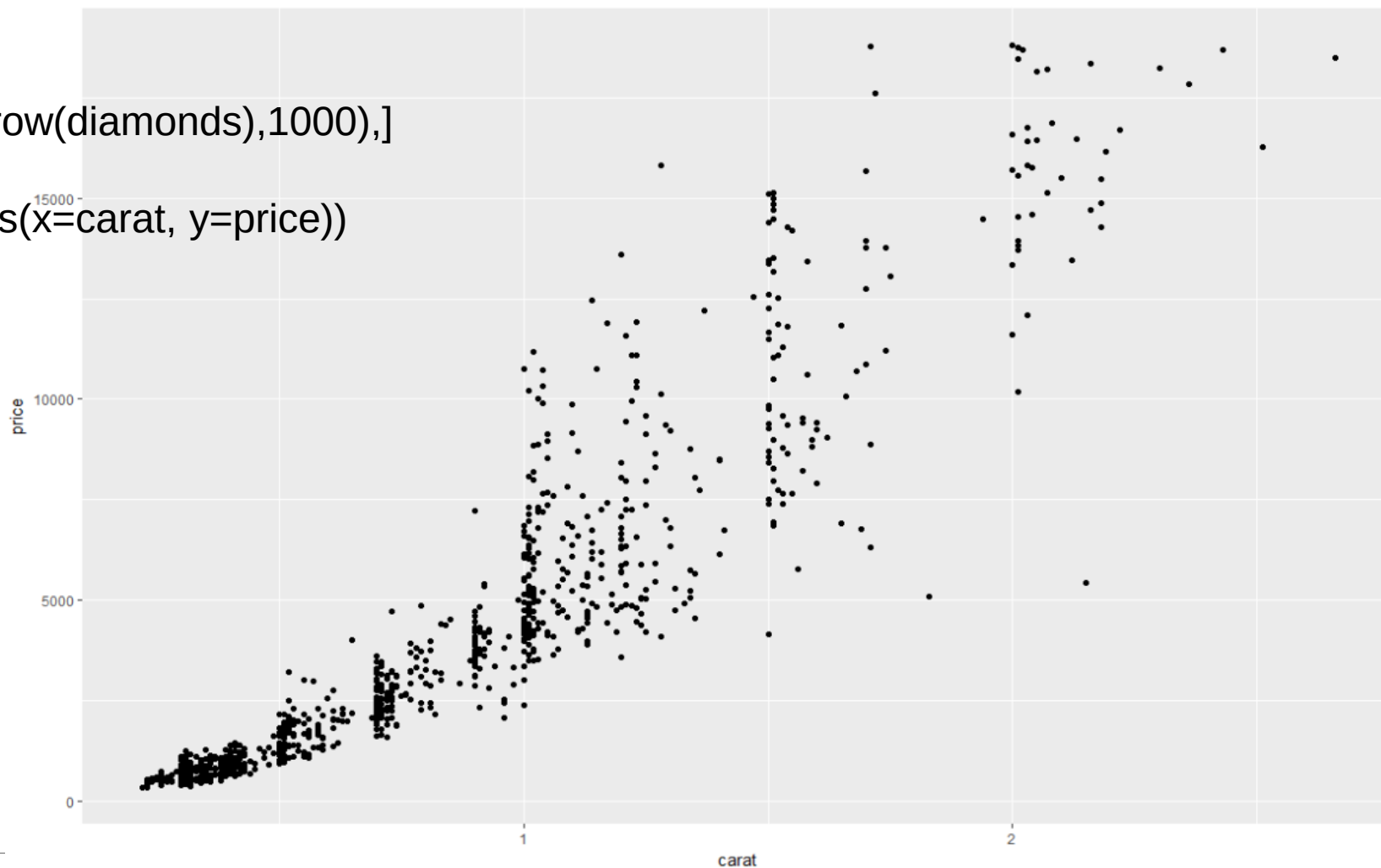
数据（Data）和映射（Mapping）

克拉	切工	颜色	净度	深度	钻面宽度	价格	x	y	z
0.71	Very Good	H	SI1	62.5	60	2096	5.68	5.75	3.57
0.79	Premium	H	SI1	61.8	59	2275	5.97	5.91	3.67
1.03	Ideal	F	SI1	62.4	57	6178	6.48	6.44	4.03
0.5	Ideal	E	VS2	62.2	54	1624	5.08	5.11	3.17
0.27	Ideal	E	VS1	61.6	56	470	4.14	4.17	2.56
0.3	Premium	E	VS2	61.7	58	658	4.32	4.34	2.67
0.56	Very Good	G	SI2	62.8	58	1091	5.19	5.23	3.27
1.05	Very Good	J	SI1	62.5	58	4199	6.47	6.52	4.06
0.32	Ideal	E	VS2	62.1	54	900	4.42	4.4	2.74

diamond数据集

数据 (Data) 和映射 (Mapping)

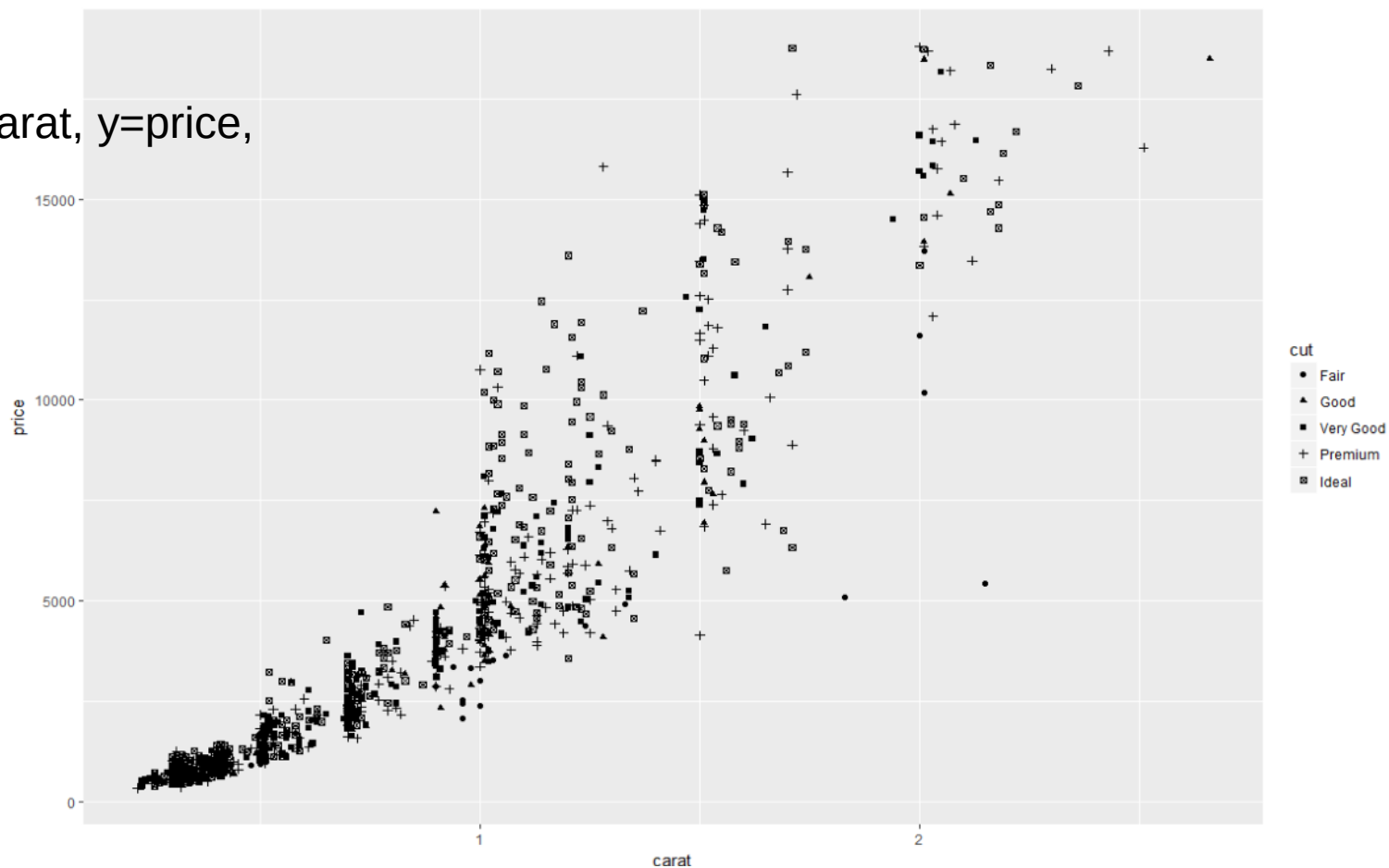
```
data("diamonds")  
set.seed(42)  
small <- diamonds[sample(nrow(diamonds),1000),]  
p <- ggplot(data = small)  
p + geom_point(mapping=aes(x=carat, y=price))
```



将克拉映射到X轴，价格映射到Y轴

数据 (Data) 和映射 (Mapping)

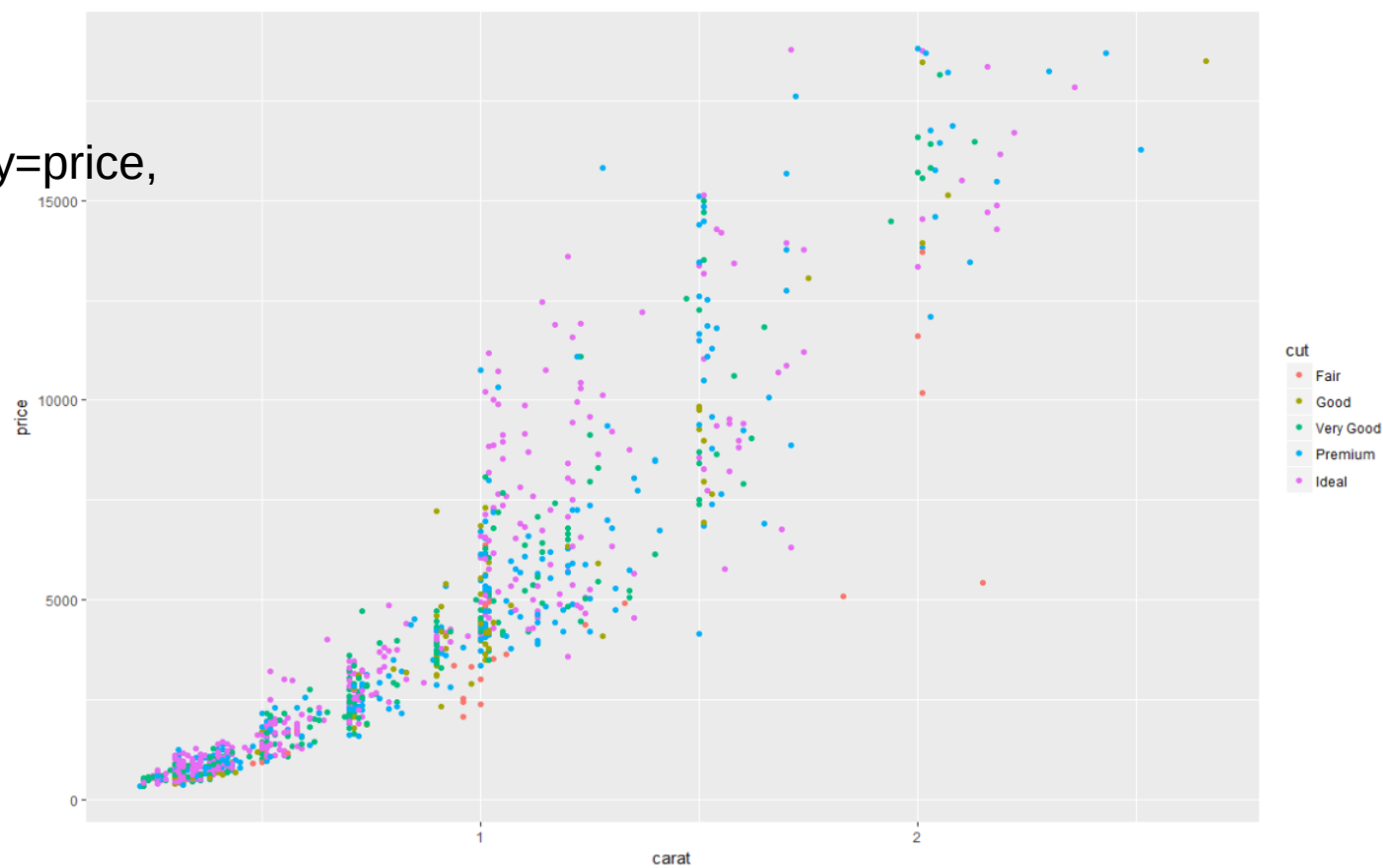
```
p <- ggplot(data = small)
p + geom_point(mapping=aes(x=carat, y=price,
  shape=cut))
```



进一步将切工映射到点的形状

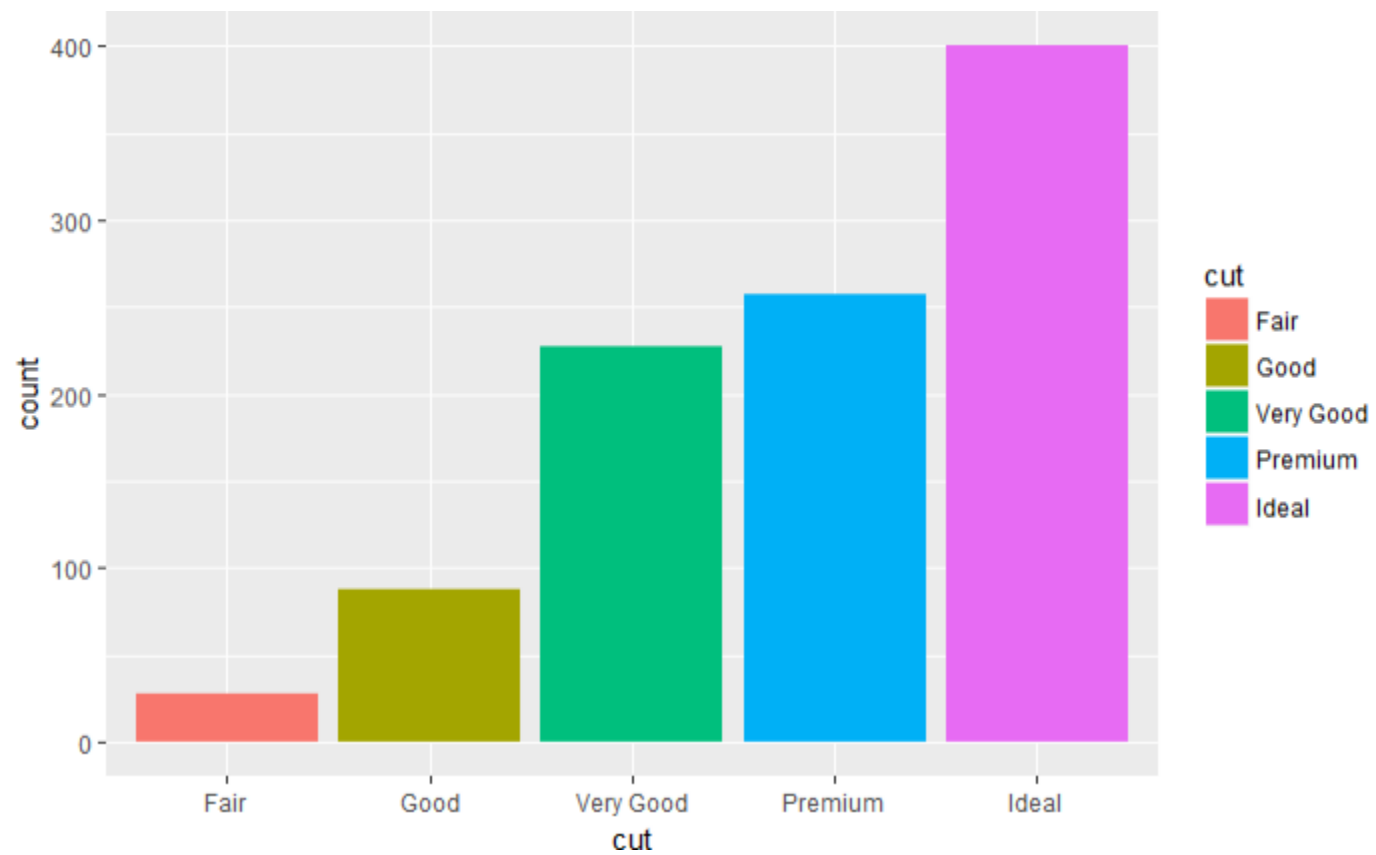
数据 (Data) 和映射 (Mapping)

```
p <- ggplot(data = small)
p + geom_point(mapping=aes(x=carat, y=price,
colour=cut))
```



也可以将切工映射到点的颜色

几何对象



几何对象 (Geometric)

图形类型	几何对象	函数
散点图	点	geom_point
折线图	线	geom_line
柱状图	柱 (分类变量)	geom_bar
直方图	直方 (连续变量)	geom_histogram
箱线图	箱 (小数据)	geom_boxplot

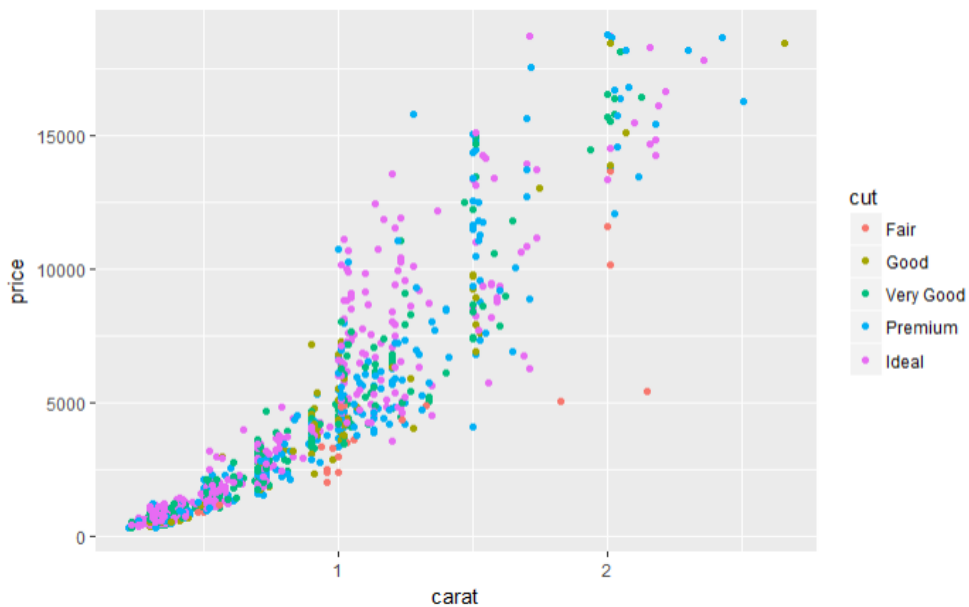
geom_abline geom_area geom_bar geom_bin2d geom_blank geom_boxplot geom_contour
geom_crossbar geom_density geom_density2d geom_dotplot geom_errorbar
geom_errorbarh geom_freqpoly geom_hex geom_histogram geom_hline geom_jitter
geom_line geom_linerange geom_map geom_path geom_point geom_pointrange
geom_polygon geom_quantile geom_raster geom_rect geom_ribbon geom_rug
geom_segment geom_smooth geom_step geom_text geom_tile geom_violin geom_vline

标度(scale)

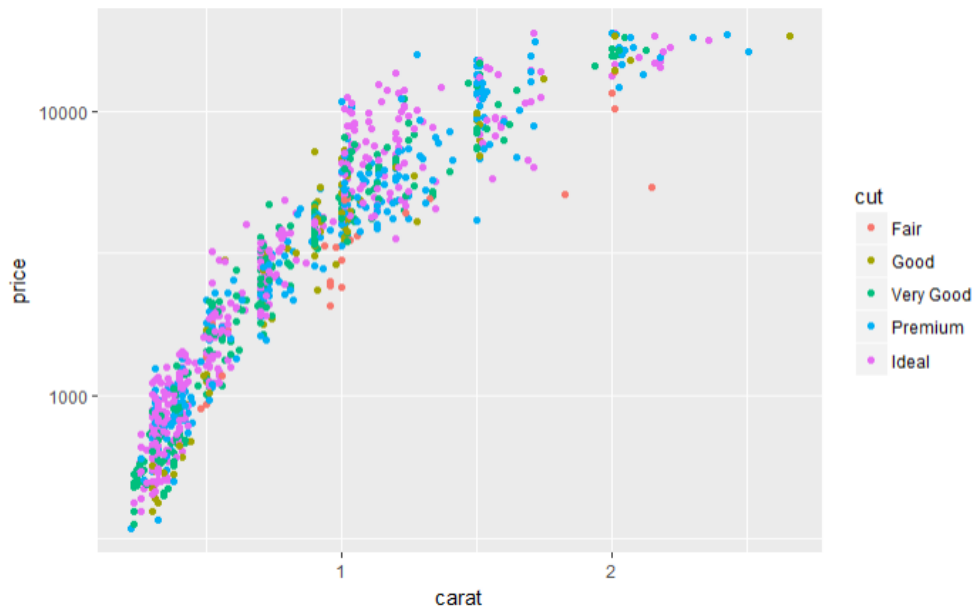
标度控制着数据到图形属性的映射。

标度将我们的数据转换为视觉上可以**感知**的东西：例如大小，颜色，位置或性状。

标尺也为我们提供了读图时所使用的工具：坐标轴和图例



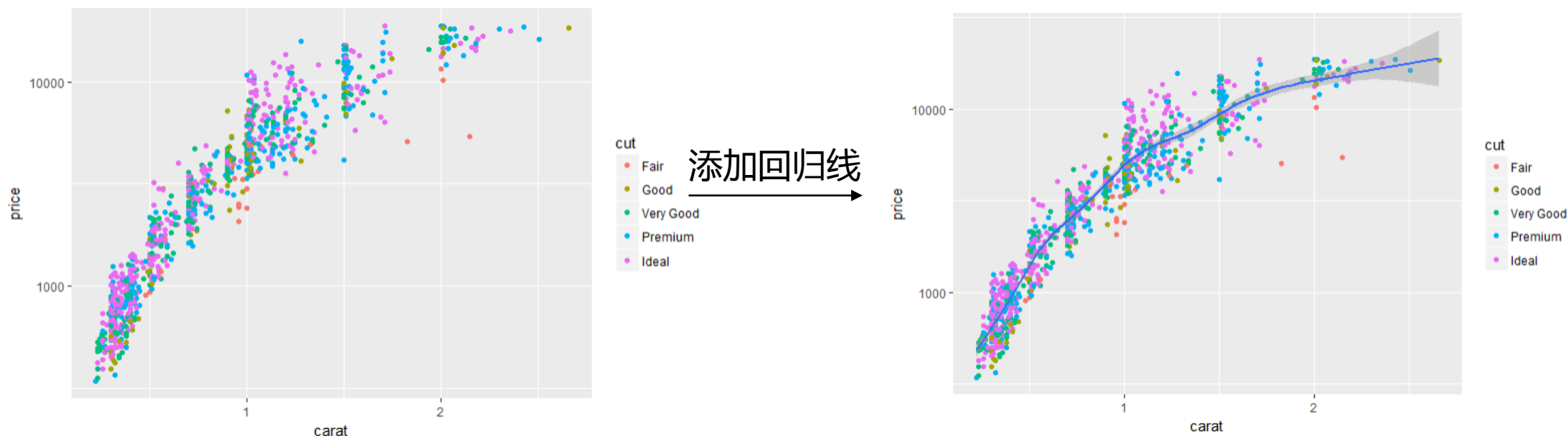
普通的Y轴



对数化的Y轴

统计变换

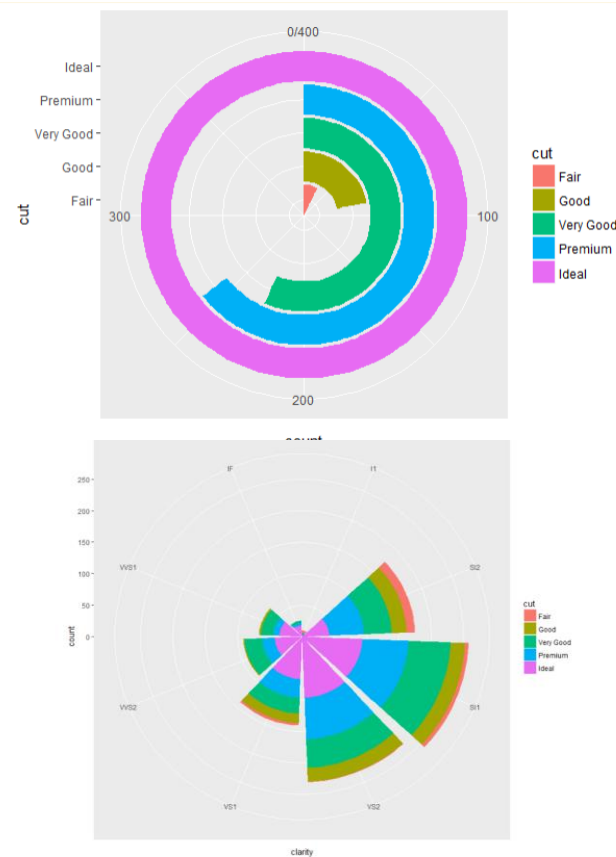
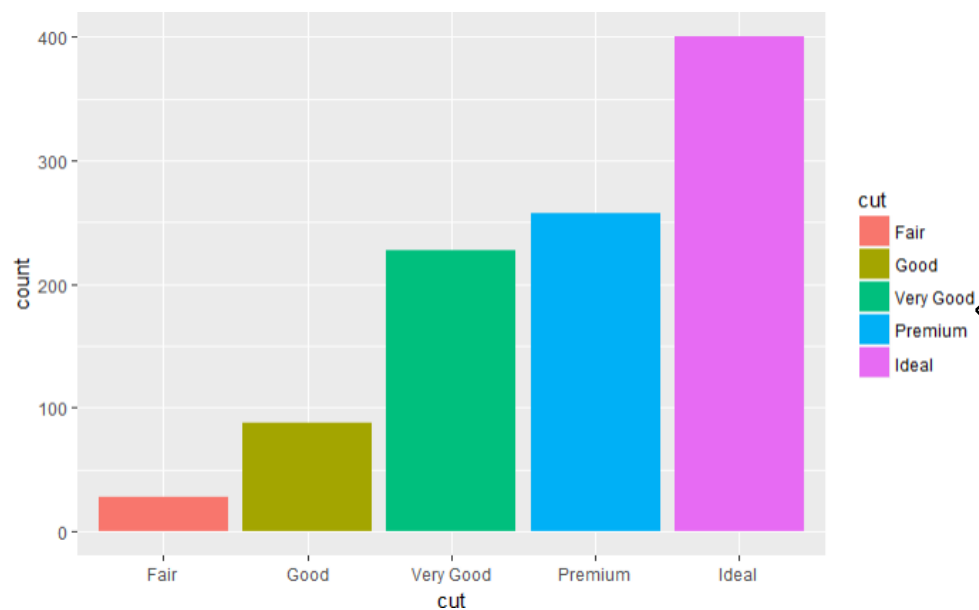
统计变换，通常指的是以某种方式对数据信息进行汇总。



stat_abline stat_contour stat_identity stat_summary stat_bin stat_density
stat_qq stat_summary2d stat_bin2d stat_density2d stat_quantile
stat_summary_hex stat_bindot stat_ecdf stat_smooth stat_unique stat_binhex
stat_function stat_spoke stat_vline stat_boxplot stat_hline stat_sum
stat_ydensity

坐标系统 (Coordinante)

坐标系统控制坐标轴，可以进行变换，例如XY轴翻转，笛卡尔坐标和极坐标转换，以满足我们的各种需求。



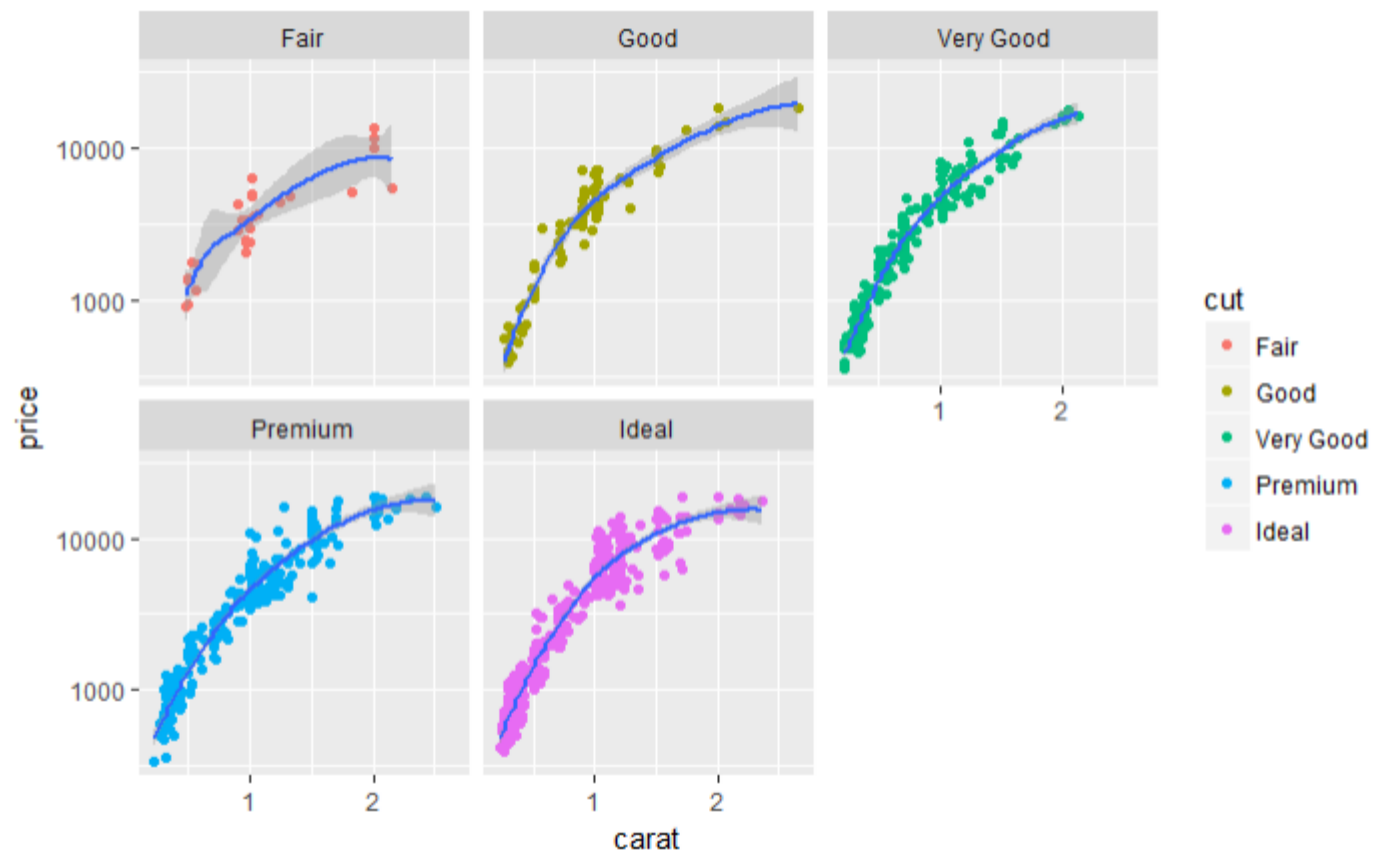
笛卡尔坐标

极坐标

分面(facet)

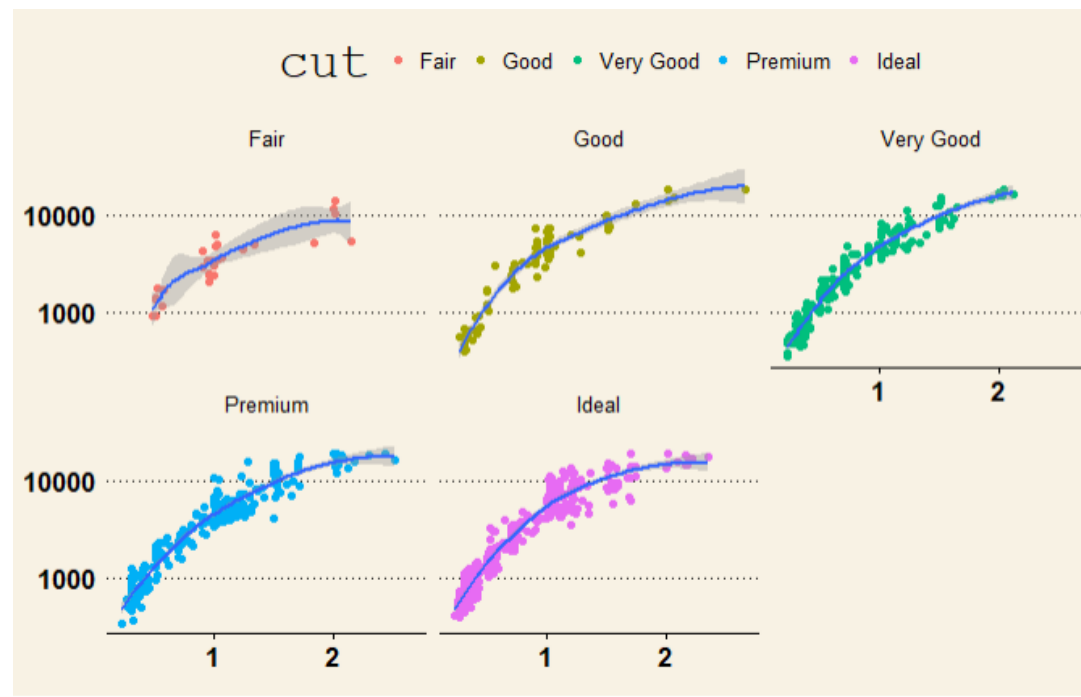
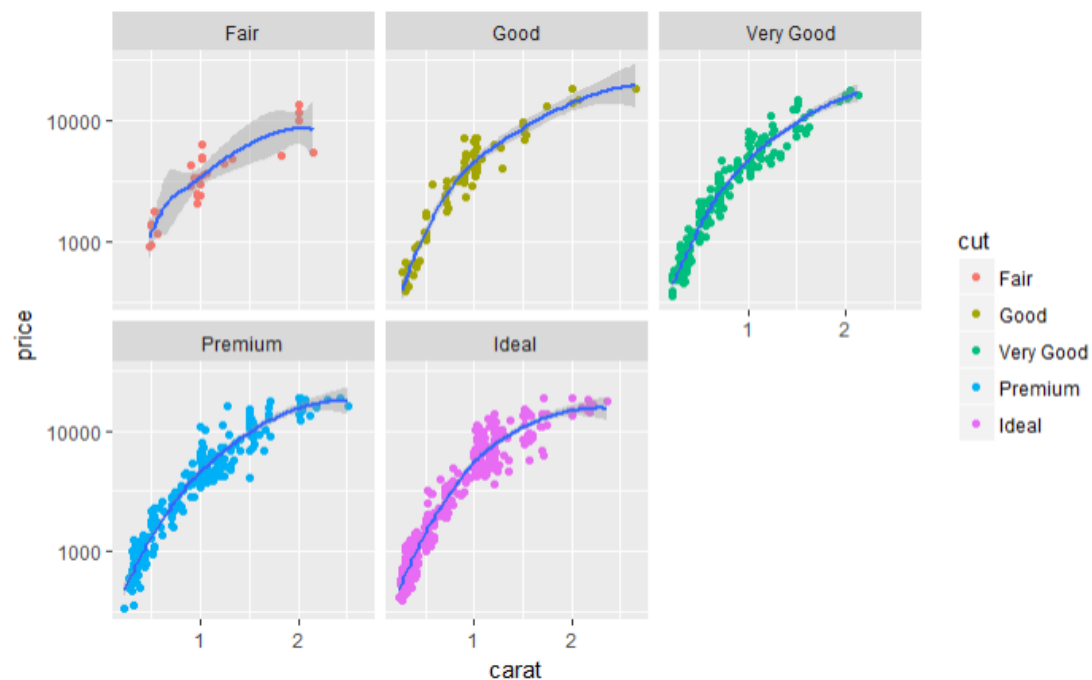
分面可以让我们按照某种给定的条件，对数据进行分组，然后分别画图。

```
ggplot(small, aes(x=carat, y=price))  
+geom_point(aes(colour=cut))  
+scale_y_log10()  
+facet_wrap(~cut)+stat_smooth()
```



主题(theme): 让图更好看

一张好看的图还需要对图进行定制，如改变字体，字体大小，坐标轴，背景等各种元素



ggtheme: 提供更多的主题选择

ggsci: 给你的配色提供参考

如何使用ggplot2画柱状图+误差图

```
Normal <- c(0.83, 0.79, 0.99, 0.69)
Cancer <- c(0.56, 0.56, 0.64, 0.52)
m <- c(mean(Normal), mean(Cancer))
s <- c(sd(Normal), sd(Cancer))
d <- data.frame(V=c("Normal", "Cancer"), mean=m, sd=s)
d$V <- factor(d$V, levels=c("Normal", "Cancer"))
p <- ggplot(d, aes(V, mean, fill=V, width=.5))
p <- p+geom_errorbar(aes(ymin=mean, ymax=mean+sd, width=.2),
position=position_dodge(width=.8))
p <- p + geom_bar(stat="identity", position=position_dodge(width=.8), colour="black")
p <- p + scale_fill_manual(values=c("grey80", "white"))
p <- p + theme_bw() +theme(legend.position="none") + xlab("") + ylab("")
p <- p + theme(axis.text.x = element_text(face="bold", size=12), axis.text.y =
element_text(face="bold", size=12))
p <- p+scale_y_continuous(expand=c(0,0), limits=c(0, 1.2), breaks=seq(0, 1.2, by=.2))
p <- p+geom_segment(aes(x=1, y=.98, xend=1, yend=1.1))
p <- p+geom_segment(aes(x=2, y=.65, xend=2, yend=1.1))
p <- p+geom_segment(aes(x=1, y=1.1, xend=2, yend=1.1))
p <- p + annotate("text", x=1.5, y=1.08, label="*")
```

保存成PDF

```
ggsave("test.pdf",p)
```

进一步的入门:如何通过Google来使用ggplot2可视化

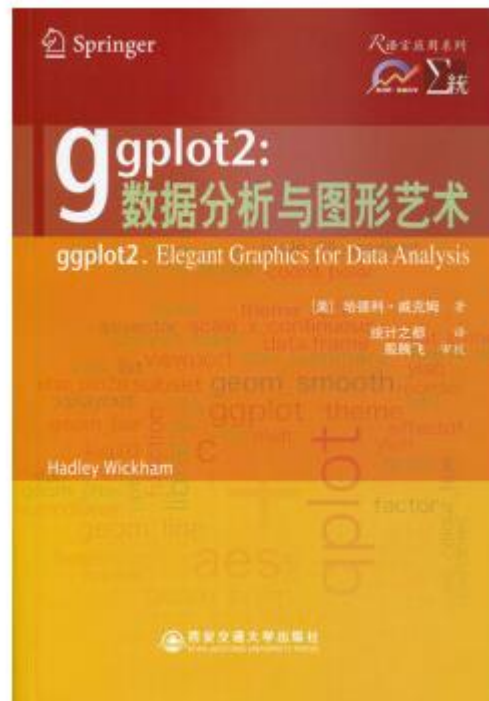
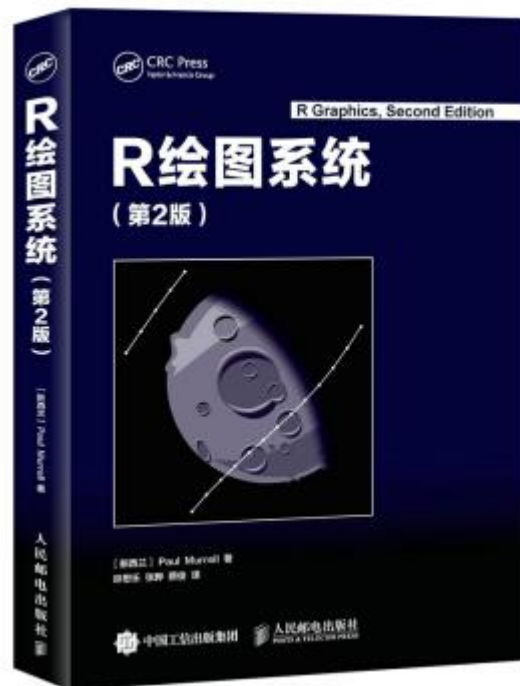
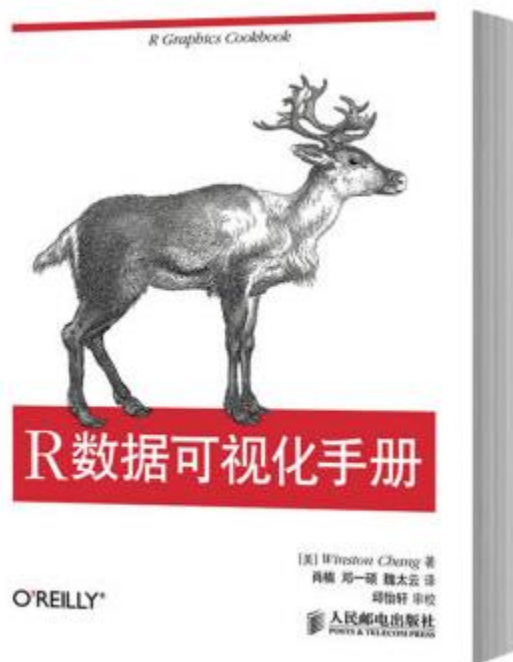


上篇



下篇

进一步的入门



biobabble



Thanks.

徐洲更

公众号：生信媛

简书:hoptop

做到

微信扫一扫加入星球

 知识星球

