


R语言应用

杨滢

主要内容



R语言基础

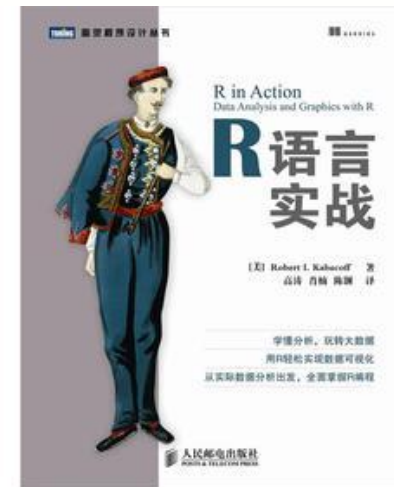
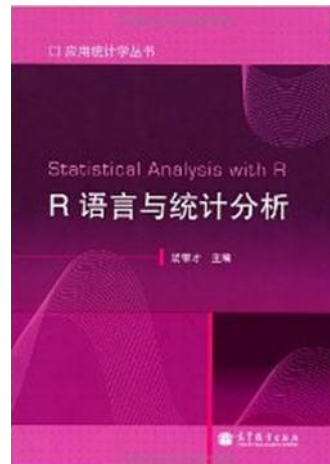


ggplot2绘图

R语言基础

Reference:

- 统计建模与R软件。薛毅。清华大学出版社。
- R语言与统计分析。汤银才。高等教育出版社。
- R语言实战（R in Action）。著：Robert I. Kabacoff；译：高涛/肖楠/陈钢。人民邮电出版社



R语言简介

- R是一种有着强大统计分析和作图功能的软件系统（编程语言）；
- 统计分析能力突出，部分统计功能整合在R语言的底层，但大多数则以包(packages)的形式提供，资源极其丰富；
- R具有强大的数据展示能力，高质量的图像生成，各种现代图像库：graphics, grid, lattice, ggplot2...；
- R的编程语言本质，注定了其强大的拓展和开发能力，可以编制自己的函数，或制作独立的统计分析包，快速实现新算法；
- 灵活，便于与其他工具整合，实现流程化；
- 官方网址：<http://www.r-project.org>；

R软件下载和安装

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

注：

为了后续课程顺利进行，考虑到程序包的兼容性，请安装**2.15.3**版本

程序包(packages)的安装和加载

**library()/
require()**

Bioconductor的包

- `source("http://bioconductor.org/biocLite.R")`
- `biocLite("DESeq")`



非Bioconductor的包

- `install.packages("ggplot2")`

若均无法安装

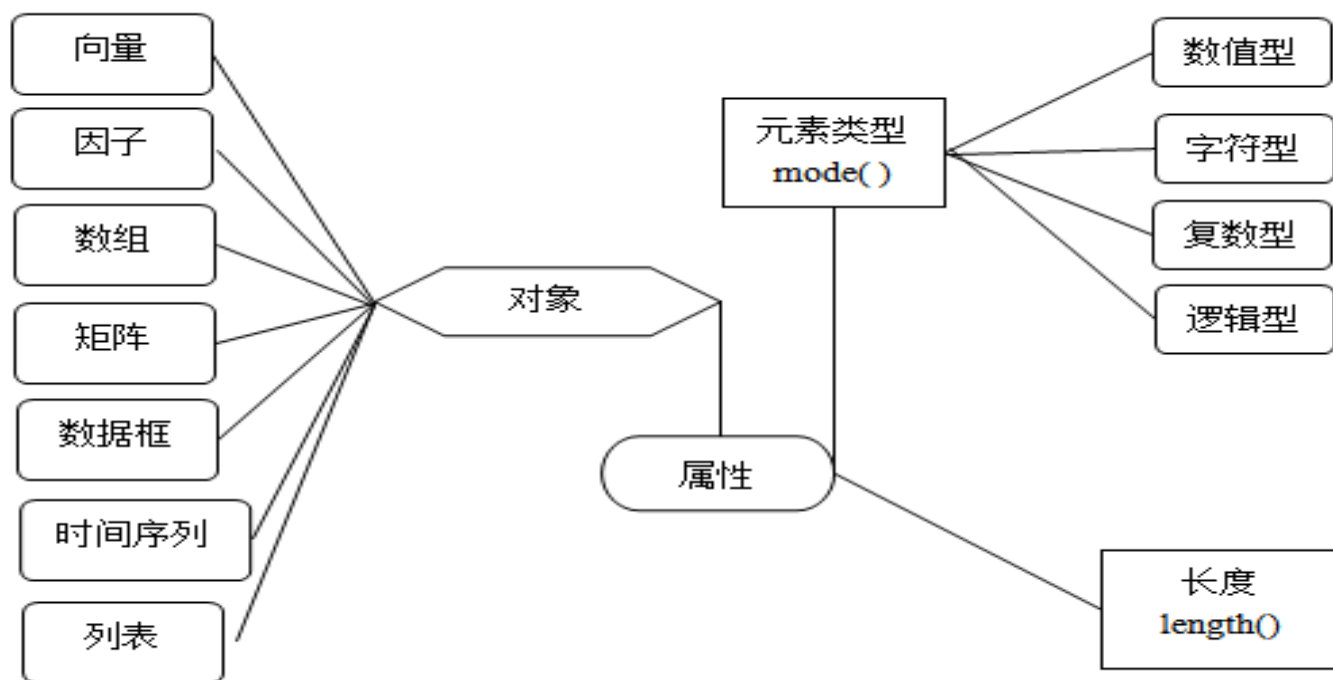
- 下载相应的软件包到本地
- R CMD INSTALL 软件包的压缩文件名称

基本知识

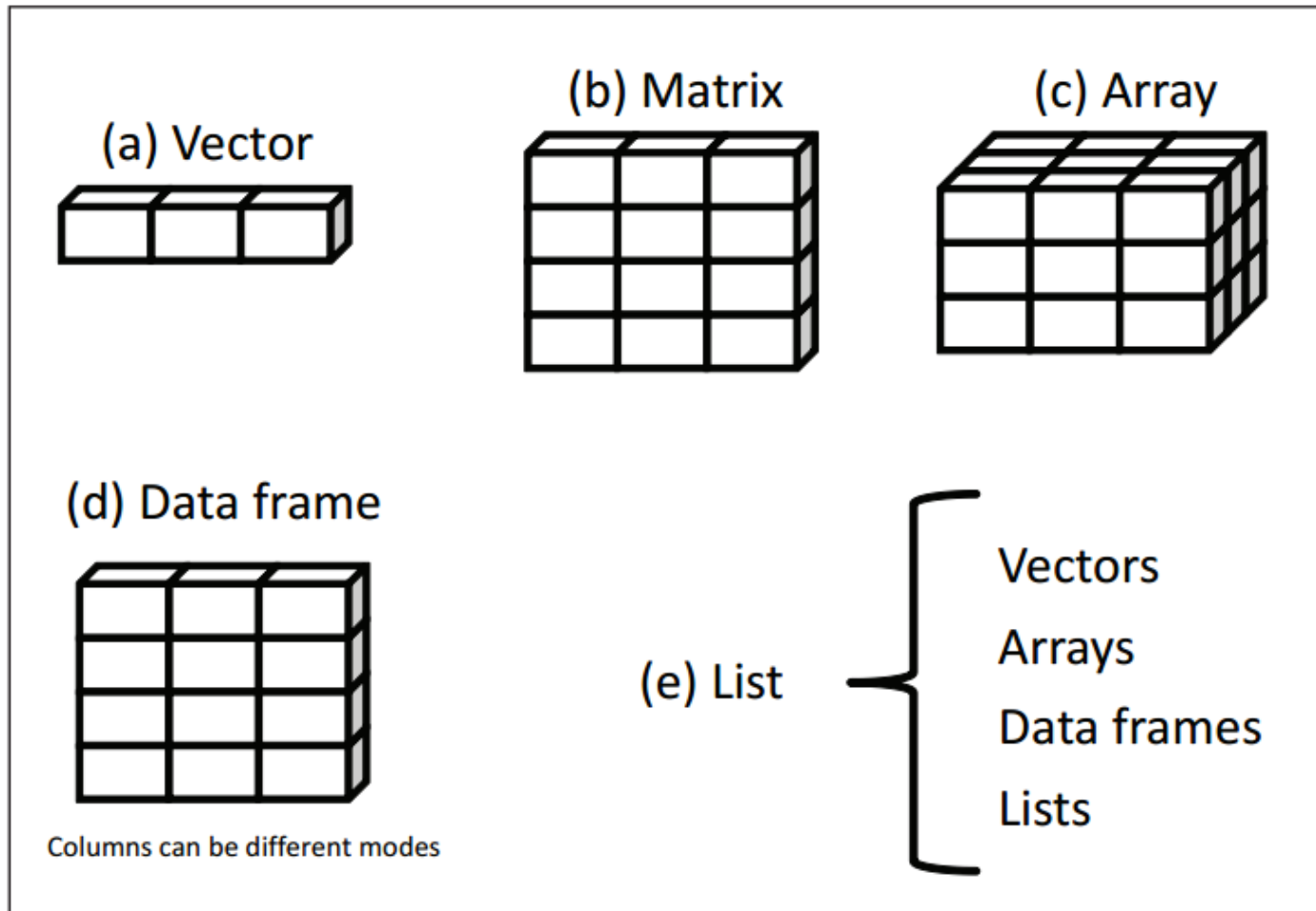
- 符号:
 - >命令或者运算提示符 +续行符
- 基本算术运算
 - +加号 -减号 *乘号 /除号 ^乘方
- 赋值符
 - = 或 <- 或 -> 
- 求助符
 - ? 或 help() 

数据结构与语法

- 在R运行时，所有变量、数据、函数以及结果都是以对象的形式存入计算机的活动内存中的，并冠以相应的变量名。在R中进行的所有操作都是针对存储在活动内存中的对象的。



数据结构



向量

- 定义：用于存储数值型、字符型或逻辑性数据的一维数组。

- **c()**创建向量,eg:

```
a<-c(1,2,3,4,5,6)
```

```
b<-c("one","two","three")
```

```
c<-c(TRUE,FALSE,TRUE,TRUE,FALSE)
```

- **seq(),rep()**创建有规律的数值型标量,eg:

```
x<-1:5
```


```
y<-seq(-5,5, by=1)
```

```
z<-rep(c(1,4,6), times=3)
```

常用统计函数

- `max(x)` 最大值
- `min(x)` 最小值
- `range(x)` 数值的范围
- `which.max(x)` 最大值下标
- `which.min(x)` 最小值下标
- `mean(x)` 均值
- `median(x)` 中位数
- `var(x)` 方差
- `sd(x)` 标准差
- `length(x)` 长度
- `sum(x)` 总和

矩阵

- 定义：二维数组，每个元素拥有相同的模式。实际上是有一个附加属性（维数 dim）的向量。
- **matrix()** 创建矩阵：#默认情况下，**矩阵是按列排列** 
eg1: `mymatrix<-matrix(1:15, nrow=3, ncol=5, byrow=TRUE)`
eg2:
`cells<- c(1,26,24,68)`
`rnames <- c("R1", "R2")`
`cnames<- c("C1", "C2")`
`matrix(cells, nrow=2, ncol=2, byrow=TRUE, dimnames=list(rnames, cnames))`
- 矩阵的下标运算：`mymatrix[1,2]`，`mymatrix[,c(1,3)]`
- 数组：与矩阵类似，但是维度可以大于2. 通过**array()**创建

数据框

- 定义：与矩阵类似，但不同的列包含不同类型的数据

- **data.frame()**创建, eg:

```
df<-data.frame(  
  Name=c("Alice", "Becka", "James", "Jeffrey", "John"),  
  Sex=c("F", "F", "M", "M", "M"),  
  Age=c(13,13,12,13,12),  
  Height=c(56.5,65.3,57.3,62.5,59.0),  
  Weight=c(84,98,83,84,99)  
)
```

- 矩阵转数据框？

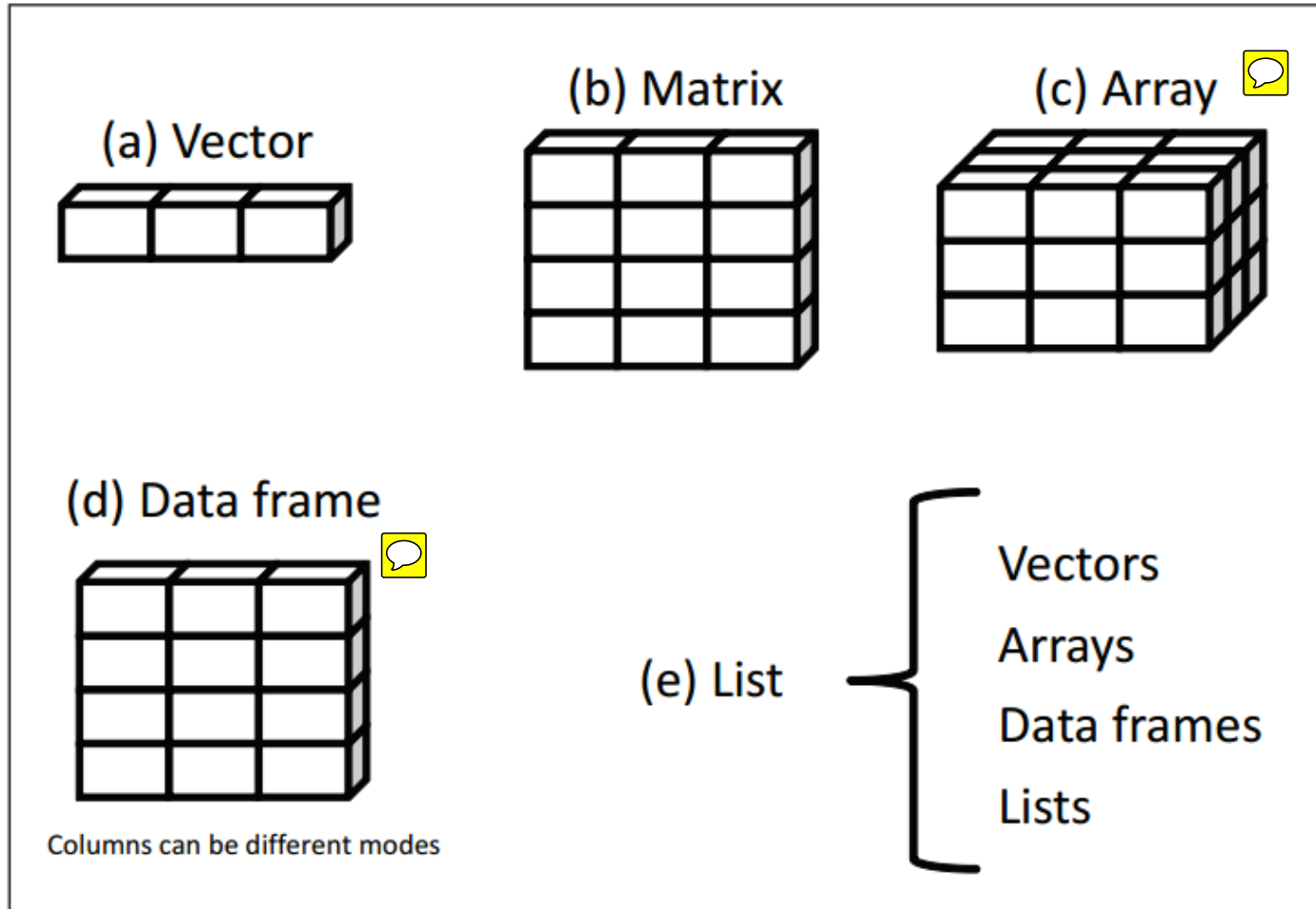
```
x<-data.frame(mymatrix) #采用data.frame()将矩阵转为一个数据框  
new_df<-as.data.frame(mymatrix) #采用as.data.frame()强制转换
```

列表

- 定义：一些对象的有序集合。最为复杂。其各元素类型可以是任意对象，不同元素不必是同一类型。
- **list()**创建，eg:

```
lst<-list(name='fred',wife='mary',no.children=3,child.ages=c(4,7,9))
```

数据结构



常用R函数

函数	功能
head(object)	查看对象的开始部分
tail(object)	查看对象的最后部分
ls()	显示当前的对象列表
length(object)	显示对象中的元素/成分数量
dim(object)	显示对象维度
c(object, object,...)	将对象合并入一个向量
object	输出对象
mode(object)	显示对象的类型
t(object)	转置对象
summary(object)	

高级函数

- **nchar,nzchar**

- 功能说明：
 - nchar: 统计字符的长度
 - nzchar: 统计字符是否为空

example:

- *nchar(c("Novogene","RNA-seq"))*
- *nzchar("")*
- *nzchar("Novogene")*

高级函数

- **substr**

- 功能说明：对字符进行特定区域的提取

- **substr(x, start, stop)**

example:

- $x=c("Novogene")$
- $substr(x,5,9)$

高级函数

- **paste**
- 功能说明：字符串的连接
- **paste (... , sep = " ", collapse = NULL)**
 - sep为向量间连接符
 - collapse为字符内部连接符
- example：产生 “AvsB”
- *paste("a","b",sep="vs")*
- *paste(c("a","b"),collapse="vs")*

高级函数

- **subset**

- 功能说明：提取数据框中满足某一条件的行

example：提取至少有一个样品的rpkm值大于1的行。

- `rpkm<-read.delim("rpkm.xls",row.names=1)`
- `rpkm_1<-subset(rpkm,A1>1 |A2>1 |B1>1 |B2>1)`
- `dim(rpkm_1)`

高级函数

- 添加列
- **merge(datatframeA,dataframeB,by=“ID”)**, 按照ID合并
datatframeA,dataframeB
- **cbind(A,B)**, 不需要索引, 直接合并A,B
- 添加行
- **rbind(A,B)**

数据的读写

- 从带分隔符的文本文件导入数据

- **1. read.table()/read.delim(),eg:**

```
rpkm<-read.table('rpkm.xls',header=T,row.names=1)
```

```
rpkm<-read.delim('rpkm.xls',row.names=1)
```

```
head(rpkm)
```

```
logrpkm<-log10(rpkm+1)
```

```
write.table(logrpkm,'logrpkm.xls',sep='\t',quote=F)
```

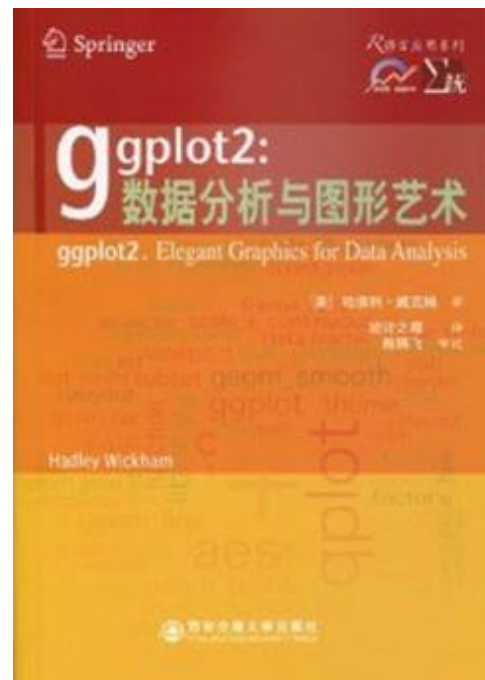
- **2. 利用剪贴板:** 一种最简单的方法是打开Excel中电子表格，选中需要的数据区域，再复制到剪贴板中（使用CTRL+C），然后在R中键入命令

```
mydata<-read.delim("clipboard")
```

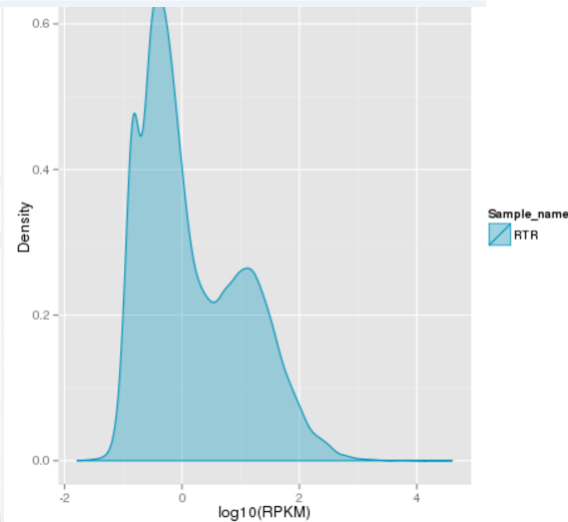
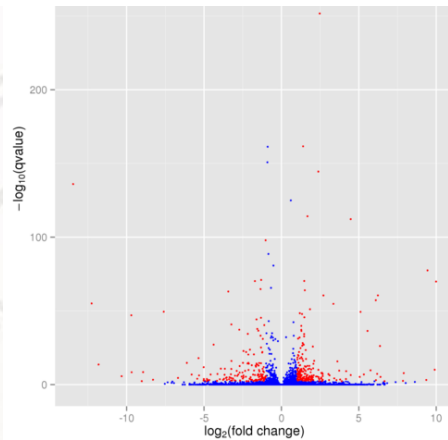
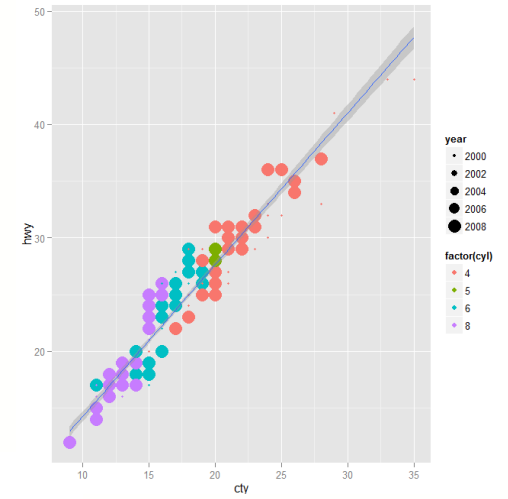
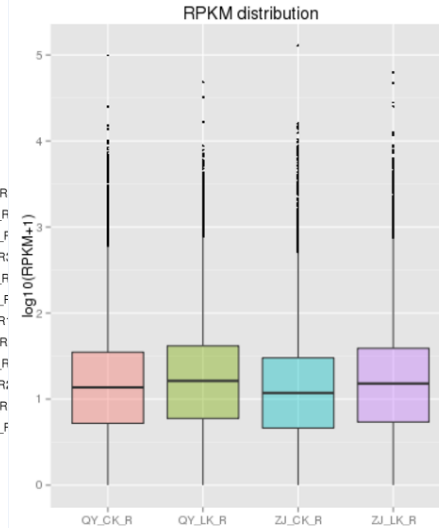
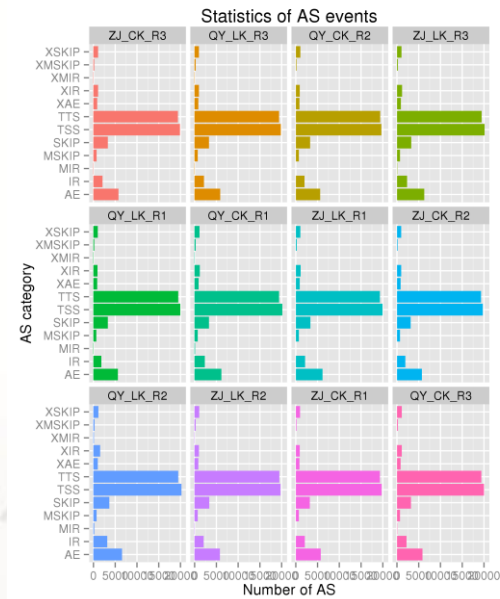
ggplot2绘图

- Reference:
- ggplot2数据分析与图形艺术。

简单、优雅

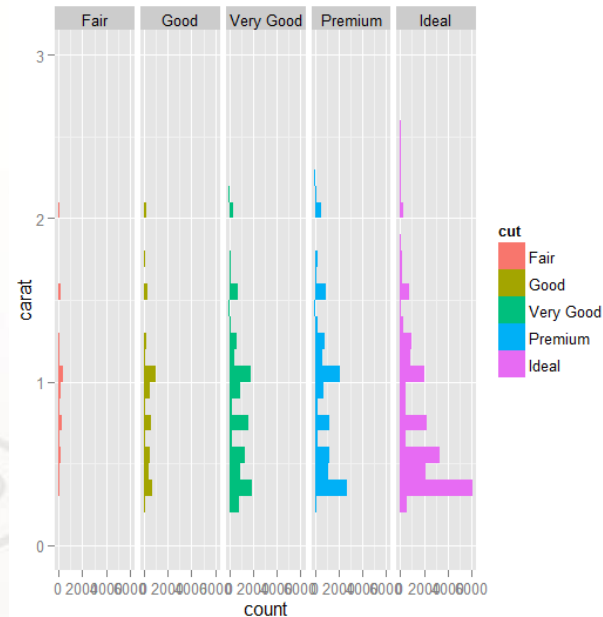


ggplot2绘图示例



快速开始~一个简单的例子

- `library(ggplot2)`
- `data(package='ggplot2')`
- `head(diamonds)`
- `p<-ggplot(diamonds,aes(x=carat))`
- `p+geom_histogram(binwidth=0.1,aes(fill=cut) ,position='dodge')+xlim(0,3)
+coord_flip()+facet_grid(.~cut)`



基本概念

- 数据（Data）和映射（Mapping）
- 标度（Scale）
- 几何对象（Geometric）
- 统计变换（Statistics）
- 坐标系统（Coordinate）
- 图层（Layer）
- 分面（Facet）

数据（Data）和映射（Mapping）

- 将数据中的变量映射到图形属性。映射控制了二者之间的关系。

length	width	depth	trt
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

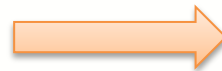


x	y	color
2	3	a
1	2	q
4	5	b
9	10	b

标度 (scale)

- 标度负责控制映射后图形属性的显示方式。具体形式上来看是图例和坐标刻度。Scale和Mapping是紧密相关的概念。

X	Y	Color
2	3	a
1	2	a
4	5	b
9	10	b



X	Y	Color
25	11	red
0	0	red
75	53	blue
200	300	blue

几何对象 (Geometric)

- 几何对象代表我们在图中实际看到的图形元素，如点、线、多边形等。

- ✓ `geom_point()` 绘制散点图
- ✓ `geom_smooth()` 拟合一条平滑曲线
- ✓ `geom_boxplot()` 绘制箱线图
- ✓ `geom_path()` 和 `geom_line()` 绘制数据之间的连线

对于一维连续变量:

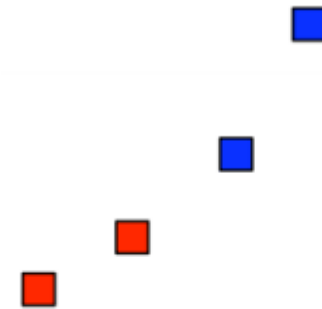
- ✓ `geom_histogram()` 绘制直方图
- ✓ `geom_density()` 绘制密度曲线

对于一维离散变量:

- ✓ `geom_bar()` 绘制条形图

- 一般用法:

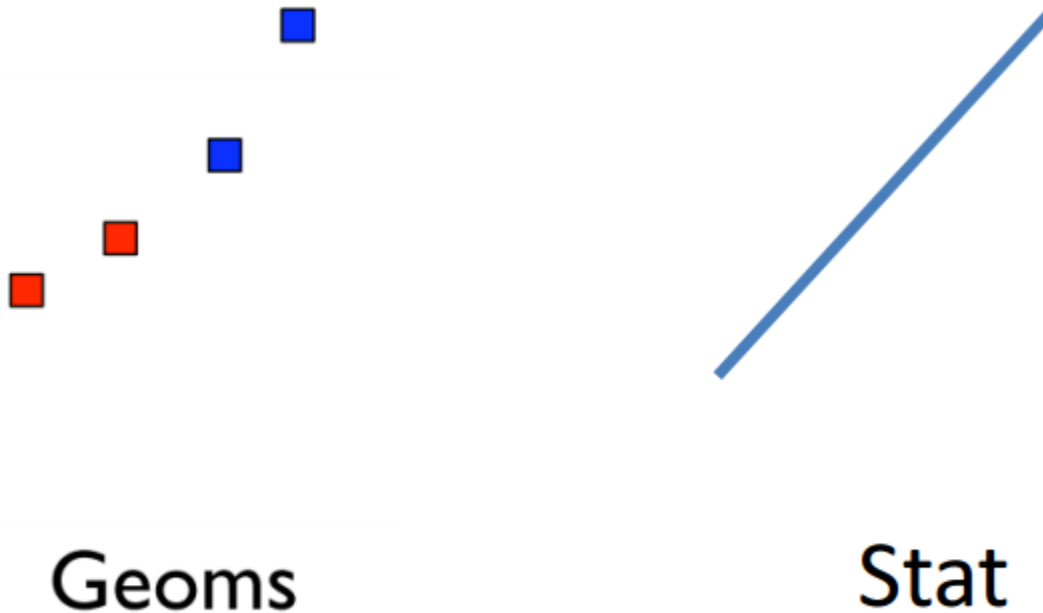
`geom_xxx(mapping, data, binwidth, ..., position)`



Geoms

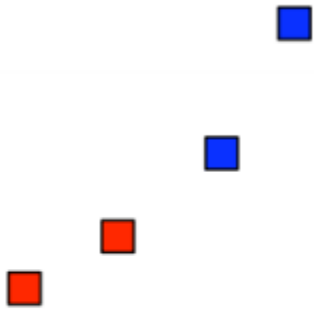
统计变换 (statistics)

- 对原始数据进行某种计算，例如对二元散点图加上一条回归线。



坐标系统（Coordinate）

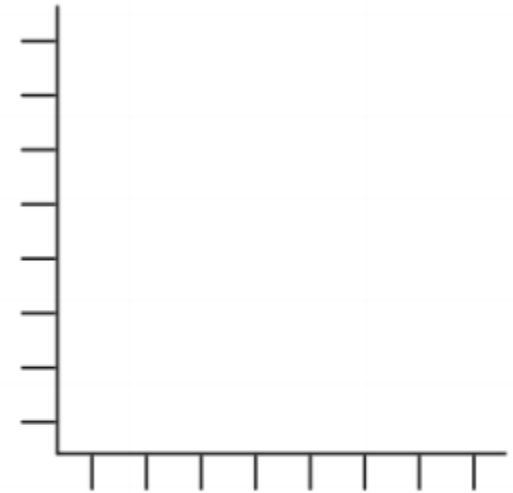
- 坐标系统控制坐标轴并影响所有图形元素，坐标轴可以进行变换以满足不同的需要。



Geoms



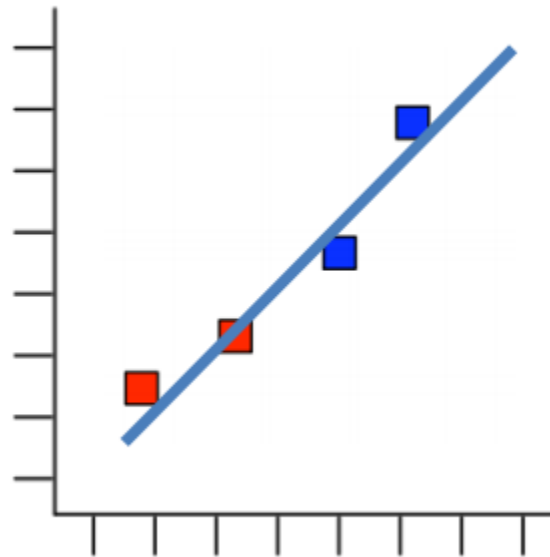
Stat



Coord

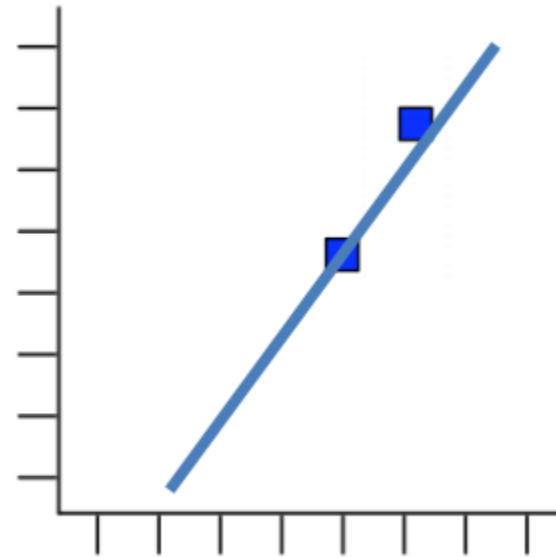
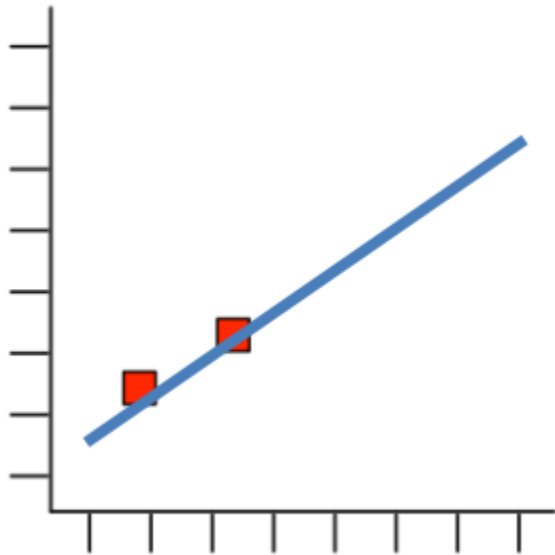
图层 (Layer)

- 数据、映射、几何对象、统计变换等构成一个图层。图层可以允许用户一步步的构建图形，方便单独对图层进行修改。



分面 (Facet)

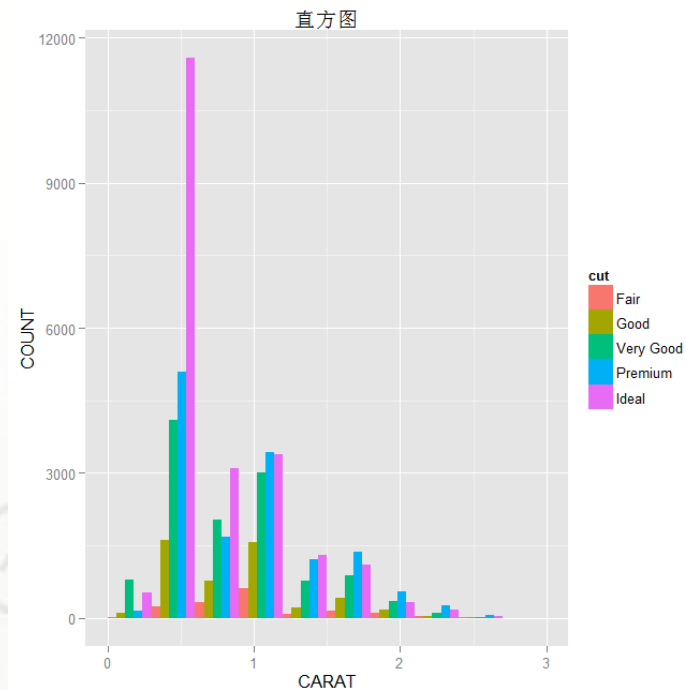
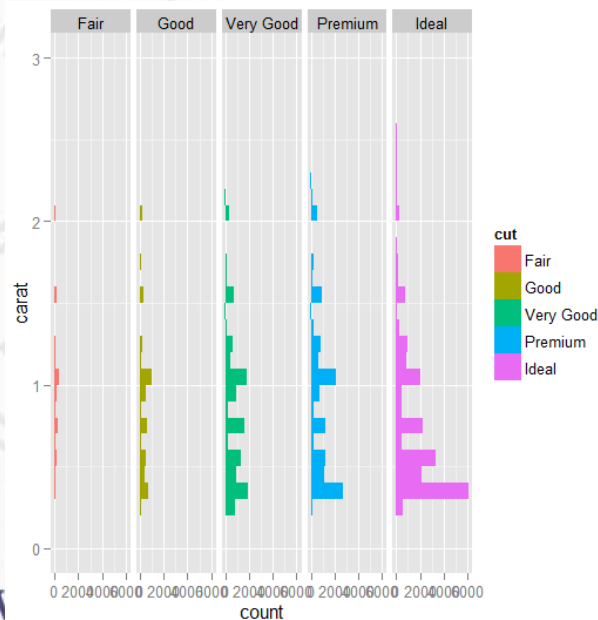
- 条件绘图，将数据按某种方式分组，然后分别绘图。分面就是控制分组绘图的方法和排列形式。



直方图分解步骤

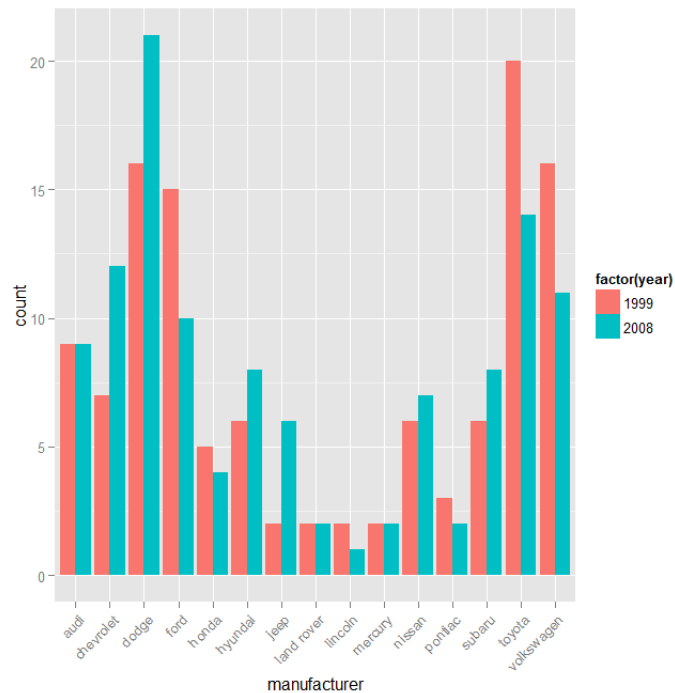
- `library(ggplot2)`
- `data(package='ggplot2')`
- `head(diamonds)`
- `p<-ggplot(diamonds,aes(x=carat))`
- `p+geom_histogram()`
- `p+geom_histogram(binwidth=0.3)`
- `p+geom_histogram(binwidth=0.3,aes(fill=cut))`
- `p+geom_histogram(binwidth=0.3,aes(fill=cut),position='dodge')+xlim(0,3)`
- `p+geom_histogram(binwidth=0.3,aes(fill=cut),position='fill')+xlim(0,3)`
- `p+geom_histogram(binwidth=0.3,aes(fill=cut),position='stack')+xlim(0,3)`
- `##position的三种类型为: 'dodge'(并列);'stack'(堆叠);fill(填充)`

- `p+geom_histogram(binwidth=0.1,aes(fill=cut),position='dodge')+xlim(0,3)+coord_flip()`
- `p+geom_histogram(binwidth=0.1,aes(fill=cut))+xlim(0,3)+coord_flip()+facet_grid(~cut)`
- `p+geom_histogram(binwidth=0.3,aes(fill=cut),position='dodge')+xlim(0,3)+ggtitle('直方图')+xlab('CARAT')+ylab('COUNT')`



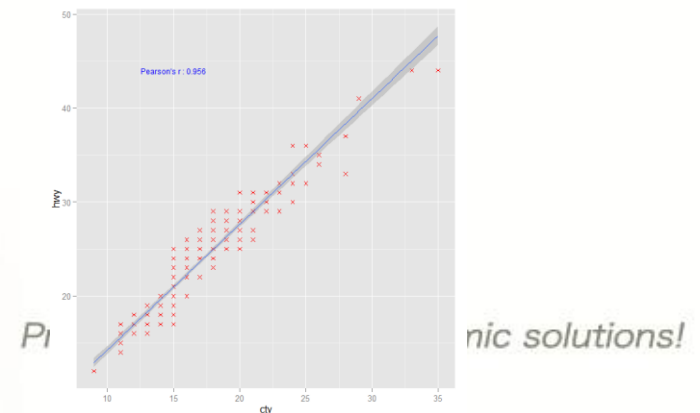
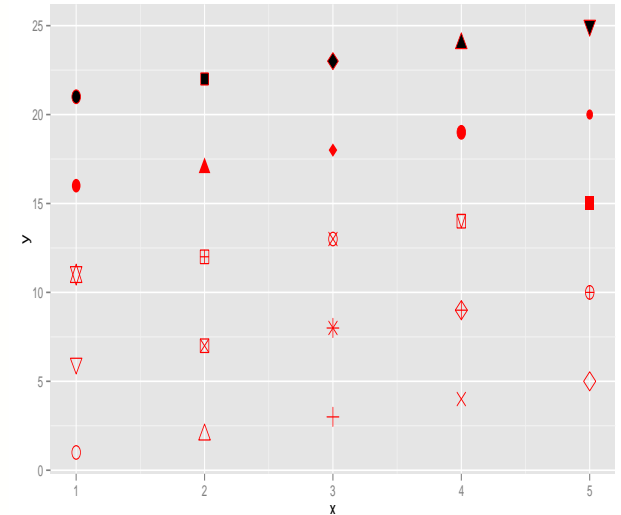
柱形图

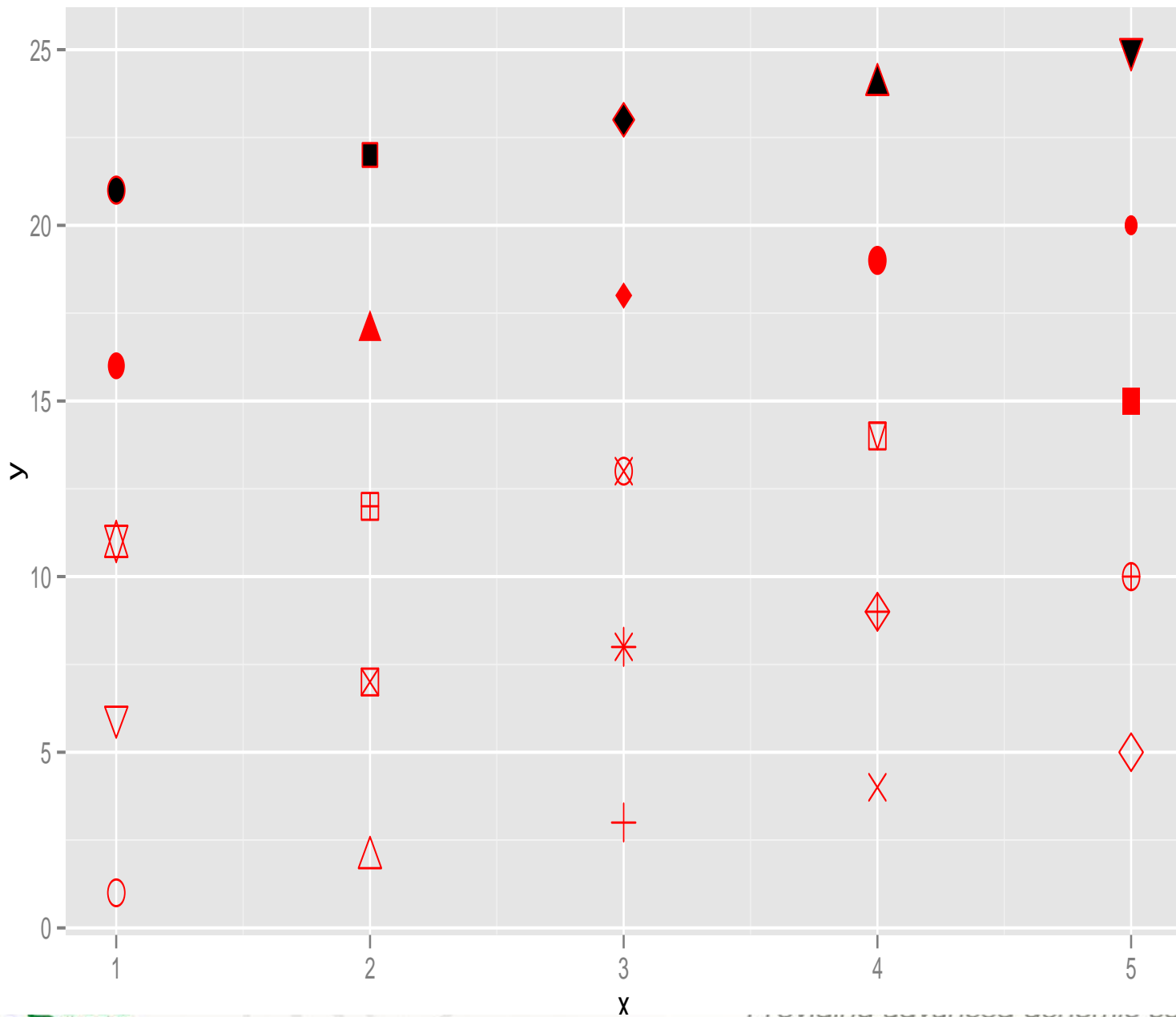
- `p<-ggplot(mpg, aes(x=manufacturer))`
- `p+geom_bar(aes(fill=factor(year)),position="dodge")`
- `p+geom_bar(aes(fill=factor(year)),position="dodge")+theme(axis.text.x=element_text(hjust=1,angle=45))`



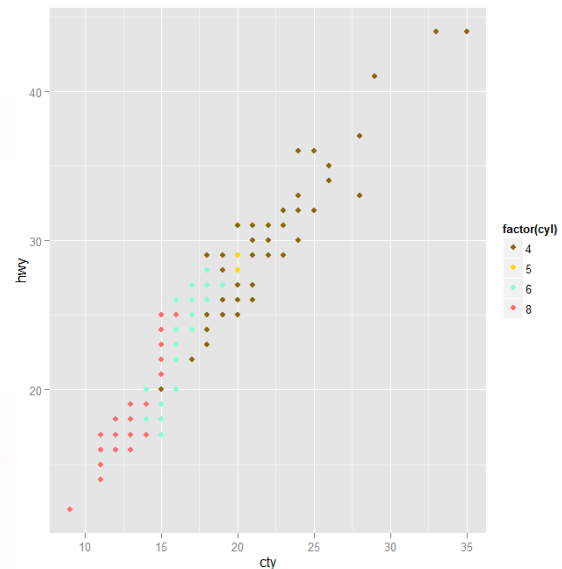
散点图

- `head(mpg)`
- `p<-ggplot(mpg,aes(x=cty,y=hwy))`
- `p+geom_point()`
- `p+geom_point(colour='red')`
- `p+geom_point(colour='red',size=5)`
- `p+geom_point(colour='red',shape=4)`
- `p+geom_point(colour='red',shape=4)+geom_smooth(method='lm')`
- `r<-round(cor(mpgcty,mpghwy),3)`
- `p+geom_point(colour='red',shape=4)+geom_smooth(method='lm')+annotate("text", 15,44, label= paste("Pearson's r :", r), size=3,colour='blue')`

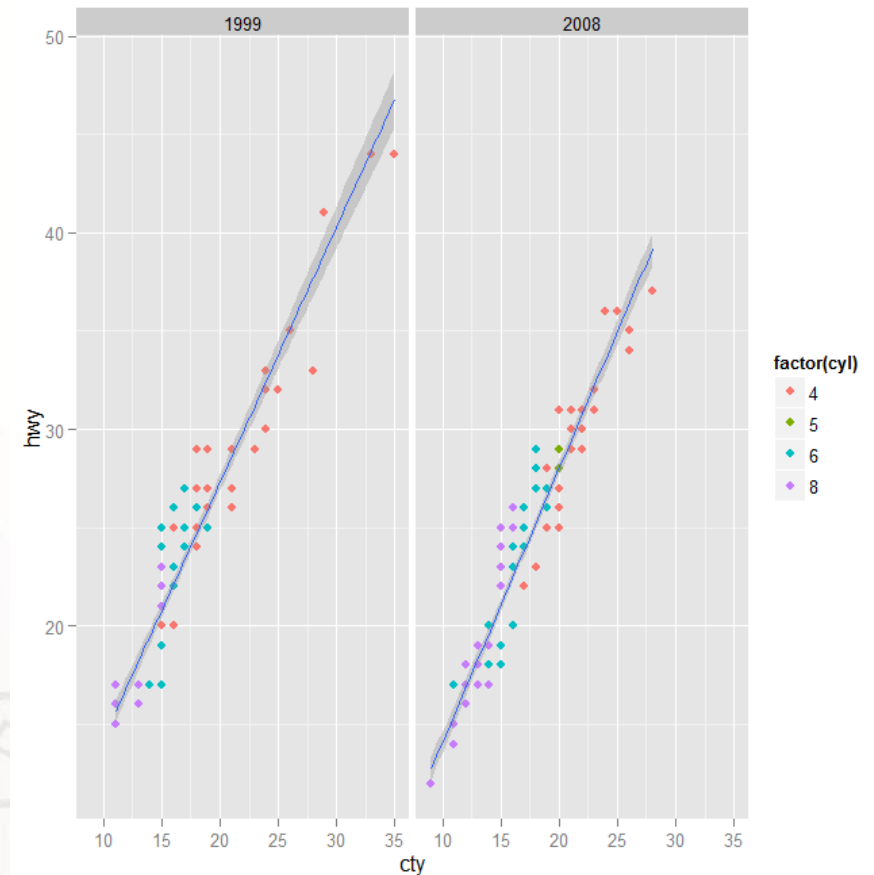
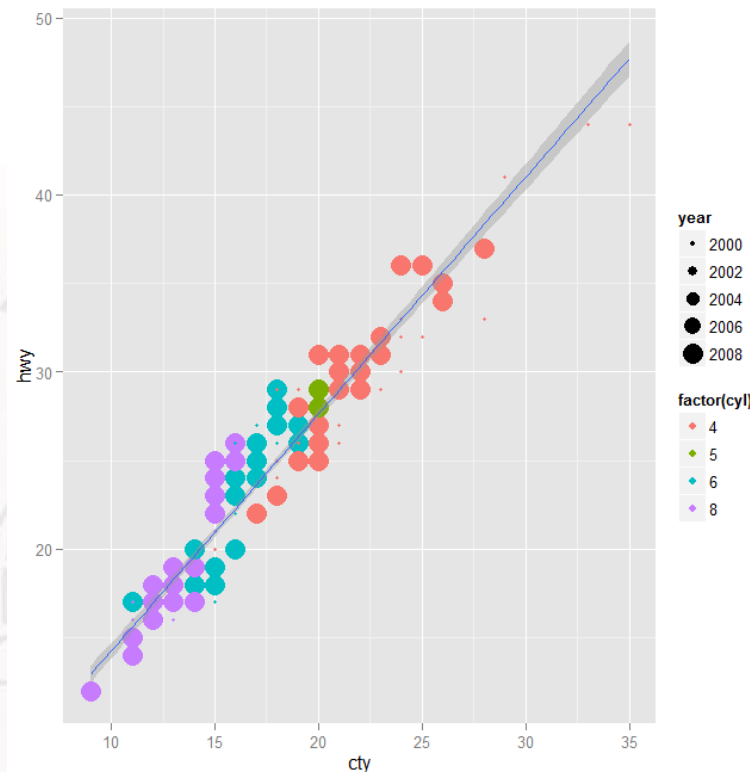




- `p<-ggplot(mpg,aes(x=cty,y=hwy,colour=factor(cyl)))+geom_point()`
- `p`
- `p+stat_smooth(method='lm')`
- `ggplot(mpg,aes(x=cty,y=hwy))+geom_point(aes(colour=factor(cyl)))+stat_smooth(method='lm')`
- `p<-ggplot(mpg,aes(x=cty,y=hwy,colour=factor(cyl)))+geom_point()`
- `p`
- `p+scale_color_manual(values=c("DarkGoldenrod4", "Gold", "#7FFFD4", "#FF6A6A"))`
- (<http://www.cnblogs.com/xianghang123/archive/2012/06/13/2547604.html> RGB配色表)

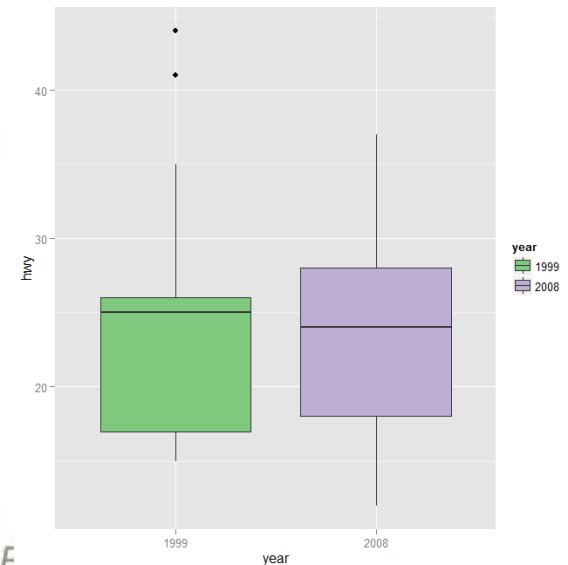
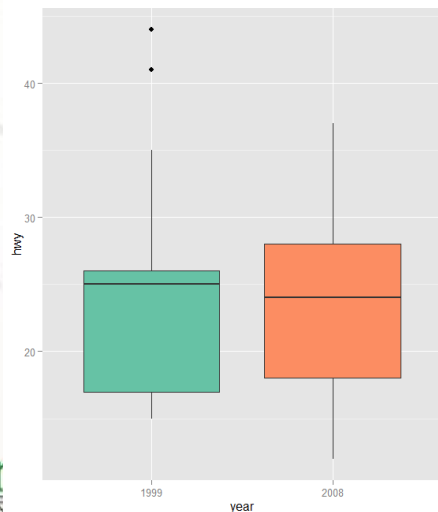


- `ggplot(mpg,aes(x=cty,y=hwy))+geom_point(aes(colour=factor(cyl),size=year))+stat_smooth(method='lm')`
- `ggplot(mpg,aes(x=cty,y=hwy))+geom_point(aes(colour=factor(cyl)))+stat_smooth(method='lm')+facet_grid(~year)`



箱型图

- `ggplot(mpg,aes(x=factor(year),y=hwy))+geom_boxplot(aes(fill=factor(year)))`
- `ggplot(mpg,aes(x=factor(year),y=hwy))+geom_boxplot(aes(fill=factor(year)))+xlab('year')`
- `ggplot(mpg,aes(x=factor(year),y=hwy))+geom_boxplot(aes(fill=factor(year)))+xlab('year')+scale_fill_brewer('year',palette='Set1')`
- `ggplot(mpg,aes(x=factor(year),y=hwy))+geom_boxplot(aes(fill=factor(year)))+xlab('year')+scale_fill_brewer('year',palette='Set2')`
- `ggplot(mpg,aes(x=factor(year),y=hwy))+geom_boxplot(aes(fill=factor(year)))+xlab('year')+scale_fill_brewer('year',palette='Accent')`

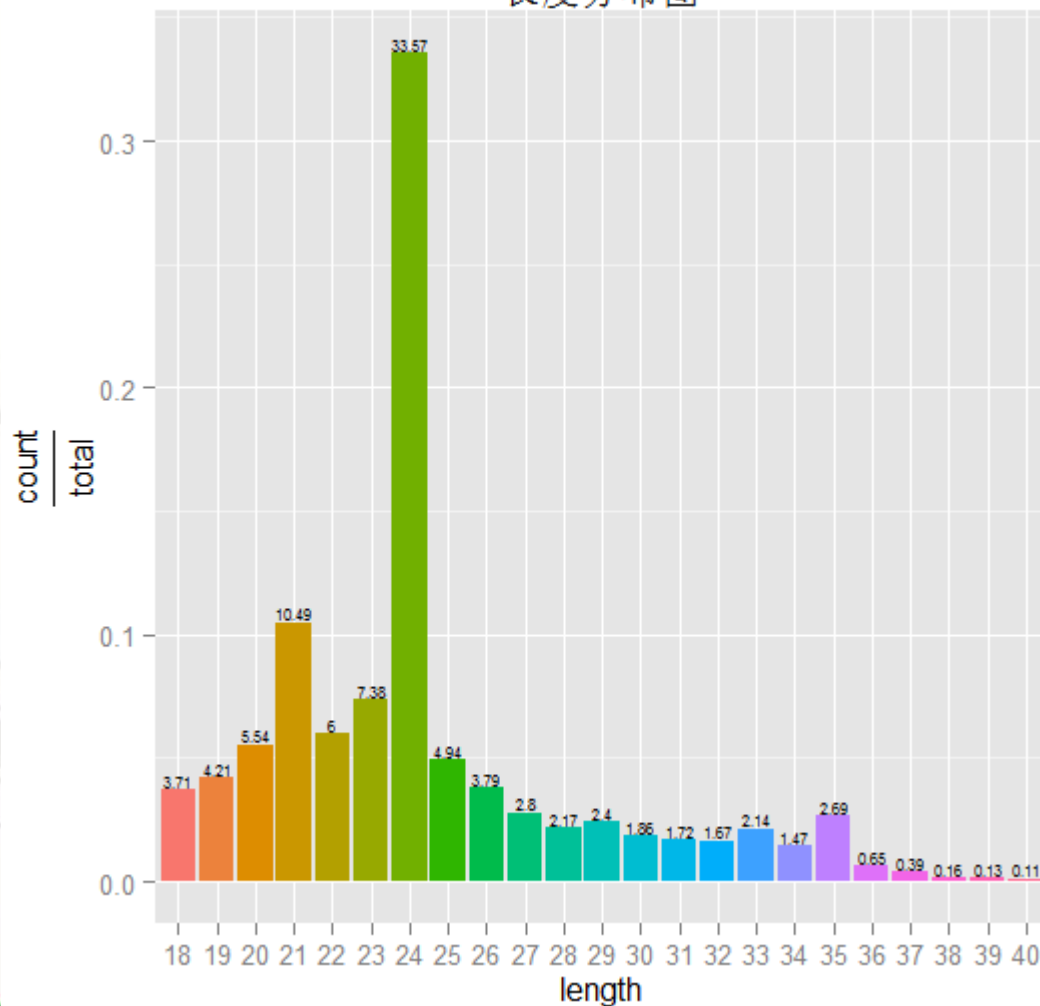


总结

- **R语言基础：**
 - ✓ R的安装及R包的安装加载；
 - ✓ R的基本使用：常用的数据结构、语法、函数
- **ggplot2绘图**
 - ✓ ggplot2的语法结构
 - ✓ ggplot2的常见类型图形的绘制：直方图、柱形图、散点图、箱线图

一个简单的练习

长度分布图



Thanks for your attention!