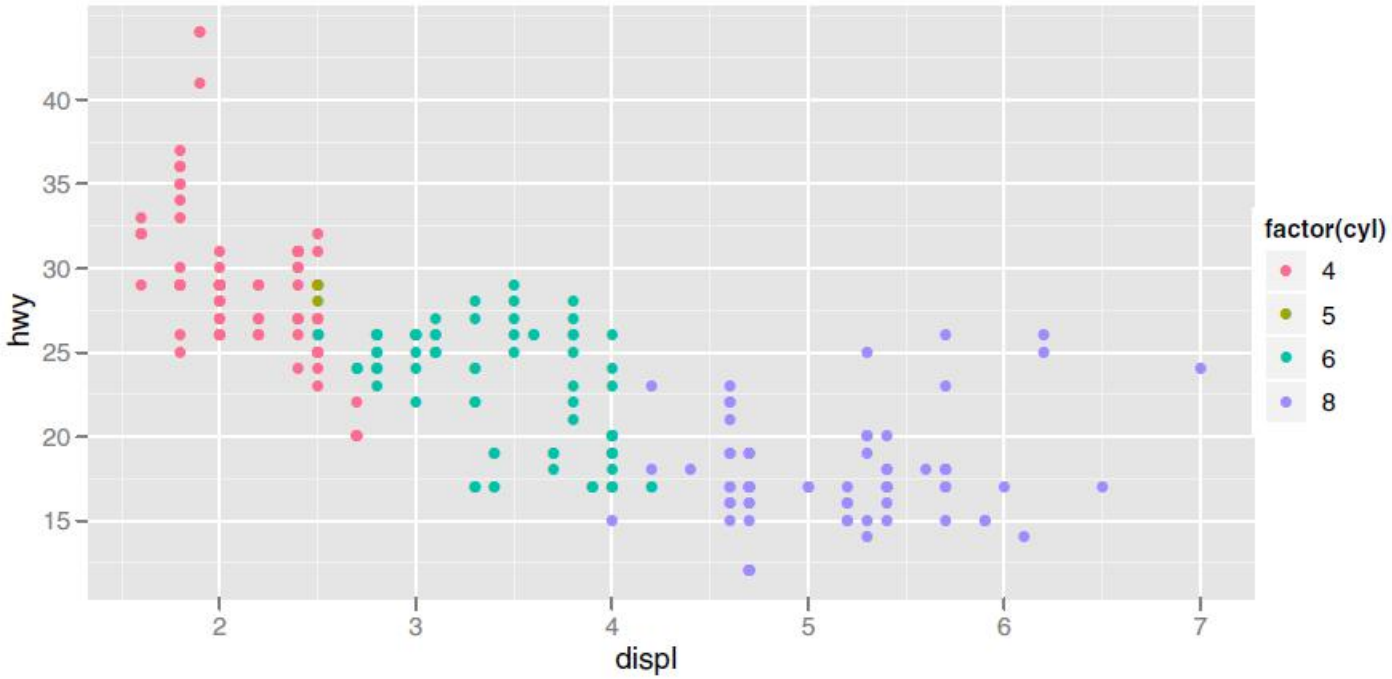
ggplot2理论基础：图层语法（The layered grammar of graphics）

数据集

- R自带数据集：mpg
- mpg记录了美国1999年和2008年部分汽车的制造商、型号、类别、引擎大小、传动系和耗油量等信息。
- 该数据集包含38种型号的汽车。
- 有趣的问题：
  - 引擎大小和耗油量有什么关系？
  - 是不是某些制造商比其他制造商更关注耗油量？
  - 耗油量在过去十年中有没有明显增加？

绘制散点图

- qplot(displ, hwy, data = mpg, colour = factor(cyl))



这是一个含有两个连续型变量的散点图：发动机排量和每加仑行驶英里数。点的颜色由第三个变量气缸数量决定。

图形属性与数据的映射（1/2）

- 散点图中，每个观测数据都用一个点（.）来表示，点的位置有两个变量的值决定。
- 每个点的属性有横坐标、纵坐标、大小、颜色、形状，这些属性称之为**图形属性（aesthetics，直译为“美学”）**。
- 每个图形属性都可以映射为一个变量或者设定成一个常数。

图形属性与数据的映射（2/2）

- 点（point）、线（line）、条（bar）都是**几何对象**的具体形式，称作**geom**。
- 几何对象决定了图形的“类型”（type）。

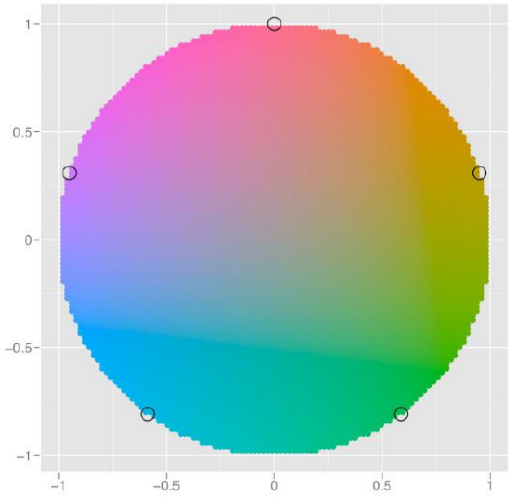
	Named plot	Geom	Other features
散点图	scatterplot	point	
气泡图	bubblechart	point	size mapped to a variable
条形图	barchart	bar	
箱线图	box-and-whisker plot	boxplot	
折线图	line chart	line	

标度变换（1/2）

- 从数据单位（如英里每加仑、汽缸数）转换成电脑可以识别的物理单位（如像素和颜色），这个转换过程称之为**标度变换**（**scaling**）。
- 根据点的位置（x和y）来确定它在图中的位置，是由坐标系统决定的，被称作**coord**。

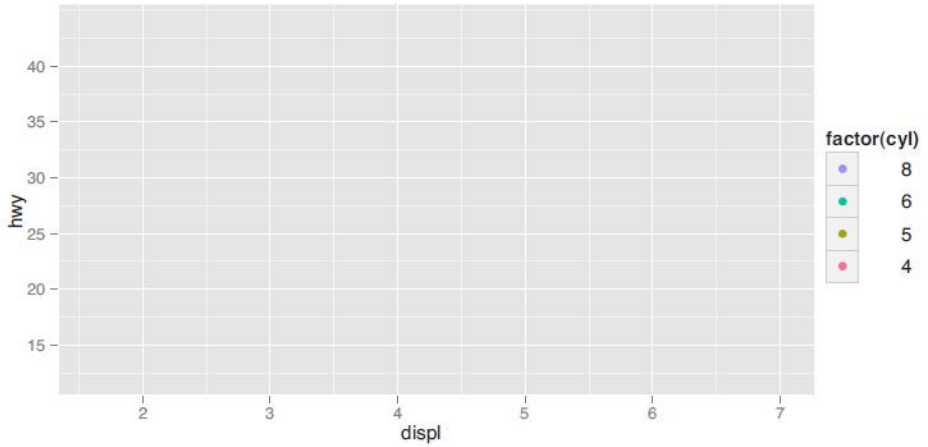
标度变换（2/2）

- 颜色变换略复杂，因为我们需要得到一个非数字的结果。颜色可以看做由三种组件组成，与人眼中识别颜色的三种细胞对应。这三种细胞建立了一个三维颜色空间。
- 颜色的标度转换就是将数据的值映射到这个空间中。
- 映射方法很多
  - cyl是分类变量，
    - 可以将其等距地映射到色轮上
    - （见右图）



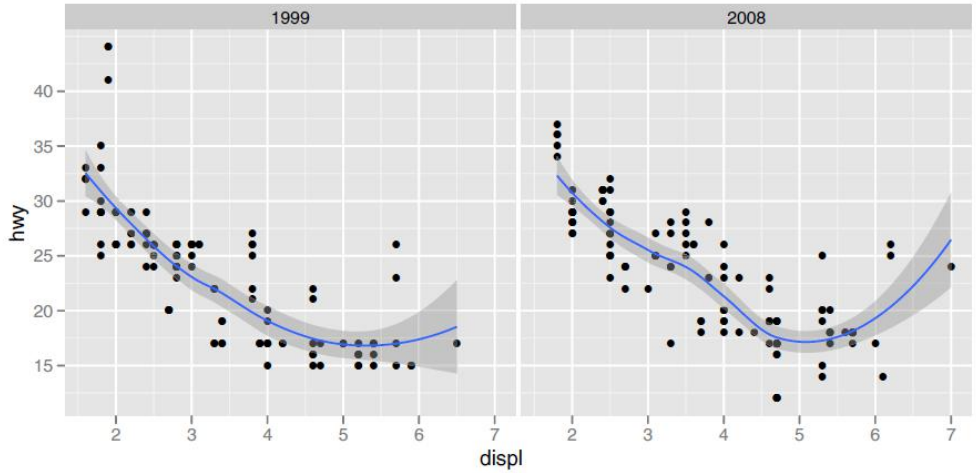
图形对象（graphical objects）

- 绘制一个完整的图形，需要组合三类图形对象：
  - **数据**：由点来表示。
  - **标度和坐标系**：用来生成坐标轴和图例，通过它们我们才能读出图中蕴含的信息。
  - **图形注释**：如背景和标题。
- 下图只留下了标度和图形注释。



添加更多组件（1/2）

- qplot(displ, hwy, data=mpg, **facets =** . ~ year) + geom_smooth()



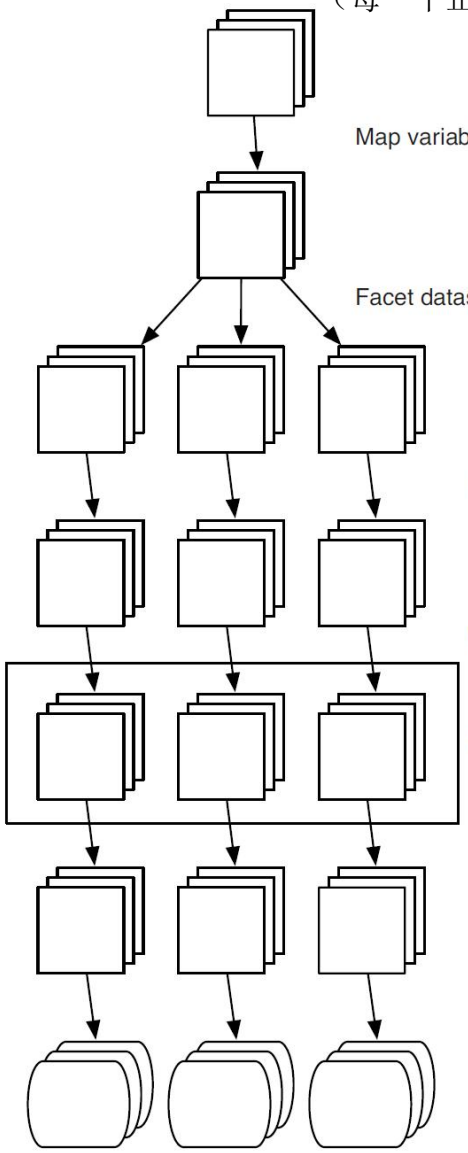
- 这幅图添加了三种新组件：分面、多个图层、统计量。
- 分面和图层将原数据切割成多个数据集：
  - 可以将其想象为一个三维矩阵：**分面面板形成了一个二维网格，图层在第三维的方向上叠加**。

添加更多组件（2/2）

- **平滑曲线层**与散点层的不同点在于它没有展示原始数据，而是展示了统计变换后的数据。
- 添加平滑曲线层需要在之前的流程中再添加一步：将数据映射到图形属性后，需要对其进行统计变换。

ggplot2绘图过程图解

（每一个正方形代表一个图层，本图展示了一个含有三个分面面板和三个图层的图）



Map variables to aesthetics 将变量映射到图形属性

Facet datasets 对数据进行**分面**处理

Transform scales **标度**变换：标度变换先于统计变换，因此统计量都是基于标度变换后的数据计算的。

Compute aesthetics 计算图形属性

Train scales **标度训练**：计算完统计量之后，所有分面和图层的数据集中的每一个标度都会被训练。标度训练将根据小数据集里数据的范围得到整体数据的范围。如果没有这一步，当不同的层叠加到一起时，它们的位置就会错乱。

Map scales

标度映射：标度映射将数据映射到图形属性中。这是一个局部操作：每个数据集里的变量都映射给相应的图形属性值。

Render geoms

渲染几何对象

图层语法的组件

图层语法的组件

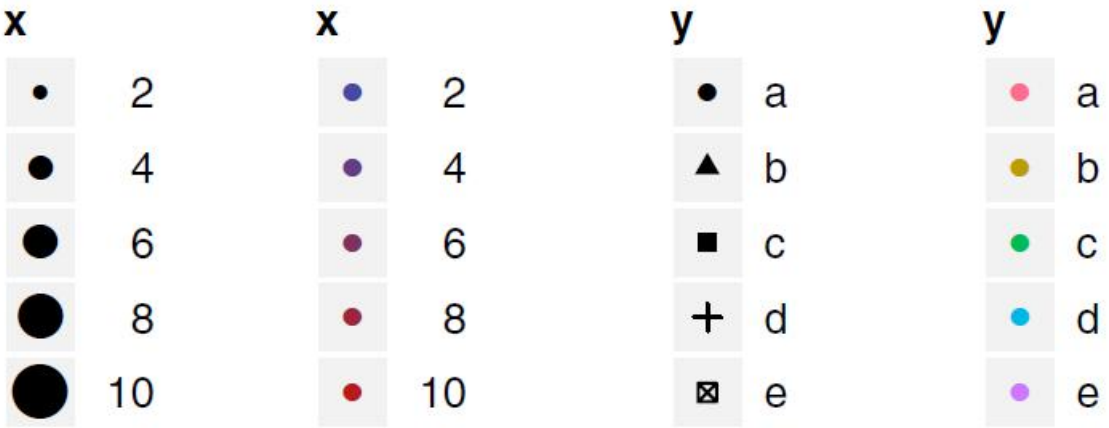
- 图层语法（layered grammar)将一张图定义为以下组件的组合：
  - 一个默认数据集和一组从变量到图形属性（aesthetics）的映射。
  - 一个或多个图层。每个图层都由一个几何对象、一个统计转换、一个位置调整，以及一个可选的从变量到图形属性的映射。
  - 一个标度：每个图形属性映射都对应一个标度。
  - 一个坐标系统。
  - 分面设定。

图层（layers）

- 图层的作用是在图上生成我们能理解的对象。
- 一个图层由四部分组成：
  - 数据和图形属性的映射
  - 统计转换
  - 一种几何对象
  - 一个位置调整（即一种位置调整的方式）

标度（Scales）（1/2）

- **标度**控制数据到图形属性的映射。图上每一个图形属性都对应着一个标度。
- 每个标度都作用于图中所有的数据，以确保数据到图形属性映射的一致性。
- 图为四种不同scales的图例。



标度（Scales）（2/2）

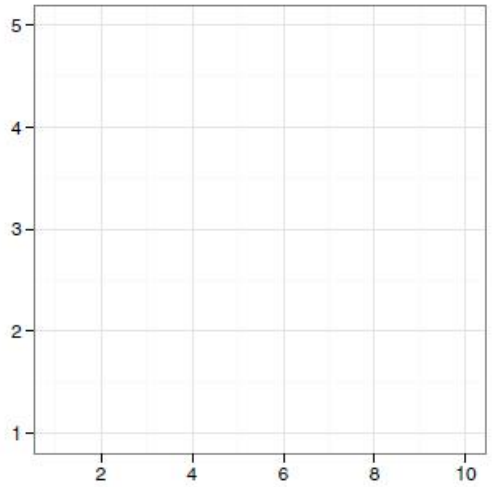
- 一个**标度**就是含有一组参数的函数。
- 其逆函数被用来绘制**参照对象**，通过参照对象你才能读出图中的隐含信息。
  - 参照对象可以是**坐标轴**（位置标度），或者是
  - **图例**。

坐标系（coord）（1/2）

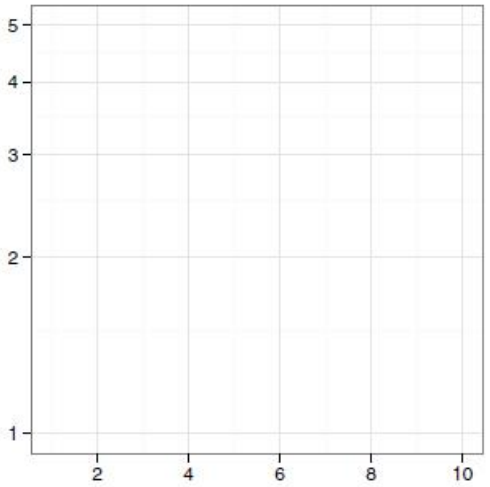
- 坐标系（coordinate system）简称为coord，用于将对象的位置映射到图形的平面上。
- 位置通常由两个坐标(x,y)决定，三维及以上尚未在ggplot2中实现。
- 笛卡尔坐标系是最常用的二维坐标系，极坐标系和地图投影用的相对较少。

坐标系（coord）（2/2）

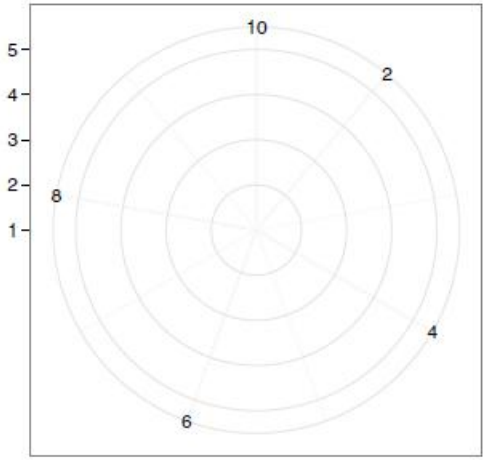
- 坐标系影响所有的位置变量，此外，坐标系还可以改变集合对象的外观。
- 标度变换是在统计变换之前进行，而坐标变换是在之后进行。
- 坐标系控制着坐标轴和网格线的绘制，见下图。



笛卡尔（Cartesian）



半对数（semi-log）



极坐标（polar）

分面（faceting）

- 通过分面可以方便的展示数据的不同子集。特别是验证在不同条件下模型是否一致时，分面是非常强大的工具。
- 分面可以设定哪些变量可用来分割数据，以及设定是否应该对位置标度加以限制。

数据结构

- 在ggolot2中，一个图形对象就是一个包含数据、映射、图层、标度、坐标和分面的列表。此外，图形对象还有一个options组件（稍后讨论）。

绘图方式

- 绘图有两种方式
  - 一步到位式：利用qplot()
  - 利用ggplot()函数和图层函数逐步作图。

图形输出（1/2）

- print() 将图形呈现到屏幕上。
- ggsave() 将图形保存在磁盘上。
- summary() 简单查看图形的结构。
  - 首先给出图形的默认设置，然后给出每个图层的信息。
- save() 把图形的缓存副本（一个图形对象的完整副本）保存到磁盘，稍后可用load()来重现该图。

图形输出（2/2）

```
p <- qplot(displ, hwy, data = mpg, colour = factor(cyl))
summary(p)
# Save plot object to disk
save(p, file = "plot.rdata")
rm(p) #删除p
# Load from disk
load("plot.rdata")
print(p)
# Save png to disk
ggsave("plot.png", width = 5, height = 5)
```