# Exploratory Data Analysis

# Week 1

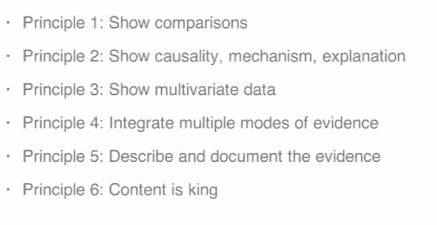
Raw data -> Processing script -> tidy data

## 绘图原则

1. show comparisons
2. 展现可能的原因，解释
3. 展示多维数据：以最简单的方式展示更多的维度信息

Simpson‘s paradox

1. integrate different modes of evidence： 图形展示丰富多样
2. labels，sources等
3. 内容



## Exploratory graphs

**One dimension**

* Five-number summary
* Boxplots
* Histograms
* Density plot
* Barplot

**Two dimensions**

Multiple/overlayed 1-D plots (Lattice/ggplot2)

Scatterplots

Smooth scatterplots

## Ploting system in R

1. base ploting system

空白——plot函数——添加其他元素

其功能主要包含在以下的2个packages内

* graphics：绘图函数
* grDevices：graph devices

1. **Lattice system**

图都是由一个简单函数构成，一次性生成，xyplot

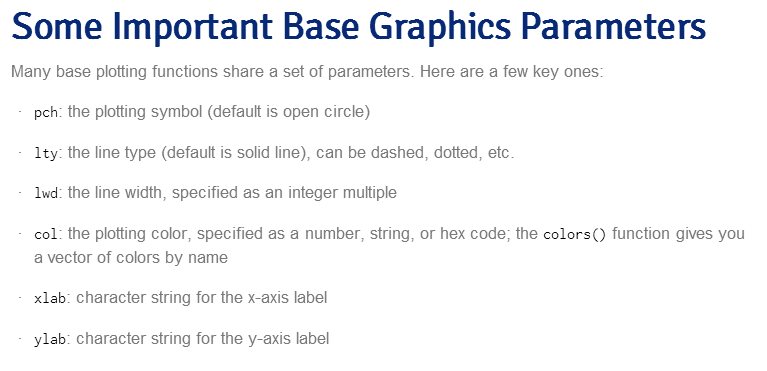
适用于：pannel plot；大量图在一张上

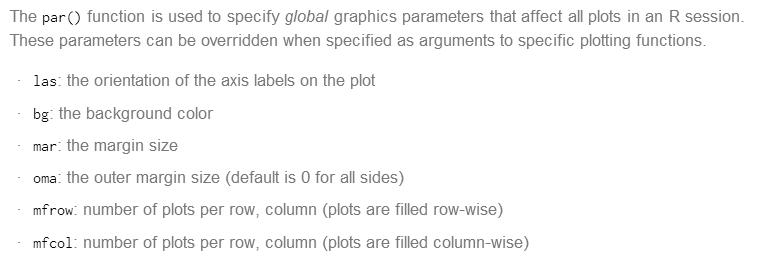
1. **ggplot2 system**

## Base ploting system

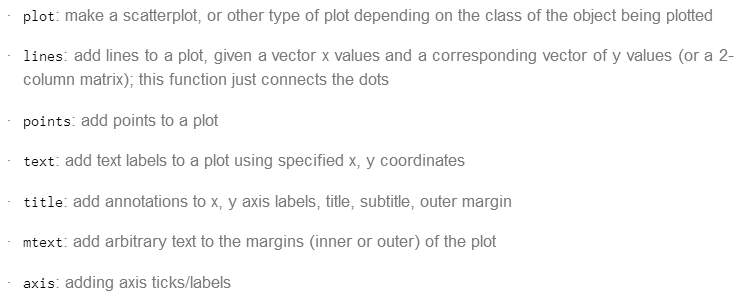
Plot()

很多的参数，可以使用?par

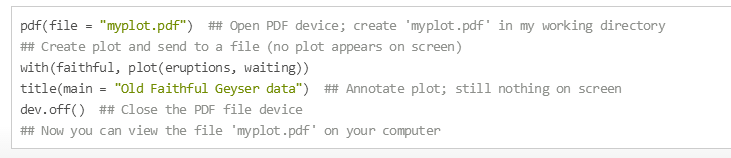




基本的绘图函数：



## Graphic devices



dev.cur() 当前的

dev.set() 设置当前

dev.copy()

dev.copy2pdf()

# Week 2

## Lattice plotting system

一般适合处理高维数据，或者同时做出很多图。Lattice绘图系统主要包括以下两个package:

* lattice: 常用的函数有xyplot，bwplot，levelplot
* grid： 一般很少用

lattice ploting system不像base system一样分为画图，添加其他部分那样分为两步，它用一个函数一次性解决。

Lattice 的绘图函数会返回一个trellis类的对象，其会自动输出。

xyplot：散点图

xyplot(y~x|f\*g,data=)

Bwplot： boxplots

Histogram

Stripplot：和boxplot类似，但是是用的点

Dotplot：

Splom：散点图矩阵，和base系统中的pairs类似

Levelplot，contourplot：绘制image data

Panal function可以自己进行设定：



## 2.ggplot2——qplot

**ggplot2有本书叫做《ggplot2：数据分析与图形艺术》**，中文版上传到群共享了

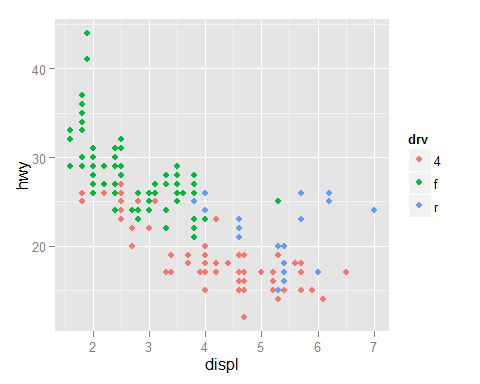
grammar of graphics:  
统计图形可以看做是由数据data 到几何对象属性(颜色、形状、大小)的映射。

* Graph是由aesthetics(size,shape,color)和geoms(points,lines)组成
* 一个基本的函数 qplot （相当于base系统中的plot） \*数据格式必须是data.frame的格式

**qplot()用法**  
qplot即quick plot

###2.1. 散点图  
导入ggplot2自带的一个数据mpg，并且按照drv变量用不同颜色显示

library(ggplot2)  
#str(mpg)  
qplot(displ,hwy,data=mpg,col=drv)



### 2.2. geom对象——二维图

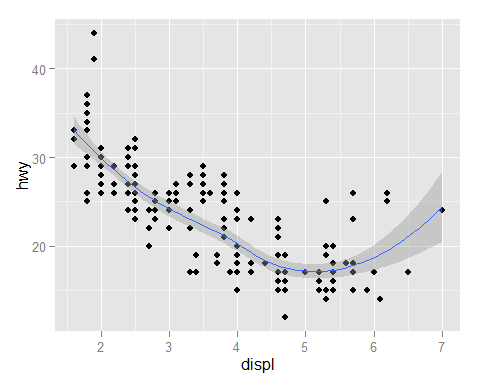
通过改变几何对象(geom),qplot可以变换出其他的图形。

* geom="point" 绘制散点图，在给定x,y时候是默认设置
* geom="smooth" 将拟合一条平滑曲线，并展示置信区间
* geom="boxplot" 箱线图
* geom="path",geom="line" 绘制数据之间的连线

例如point表示散点，smooth表示平滑曲线

qplot(displ,hwy,data=mpg,geom=c("point","smooth"))

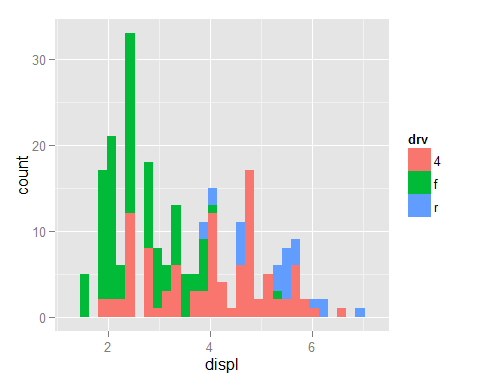
## geom\_smooth: method="auto" and size of largest group is <1000, so using loess. Use 'method = x' to change the smoothing method.



### 2.3. geom对象——一维图

* 对于连续变量
  + geom="histogram" 这是默认
  + geom="density" 绘制密度曲线

qplot(displ,data=mpg,fill=drv)



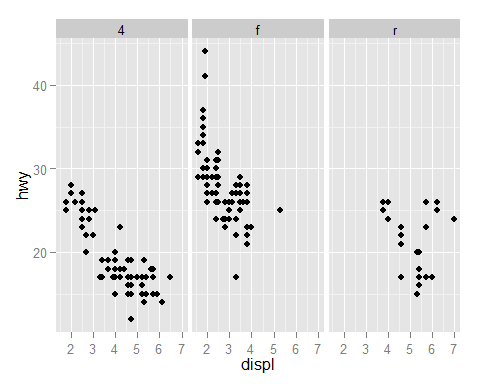
* 离散变量 geom="bar" 条形图

### 2.4. facets

facets和lattice中的pannal类似，将数据分为了若干的子集。主要是根据分类变量绘制多个图qplot()中默认的分面方法是将图形拆成若干个窗格，通过row\_var~col\_var的形式指定。

但是当变量的个数超过两个的时候图形可能会较大，不适合展示。如果只想指定一列或一行，可以使.作为占位符。  
比如row\_var~. 会创建一个单列多行的图形

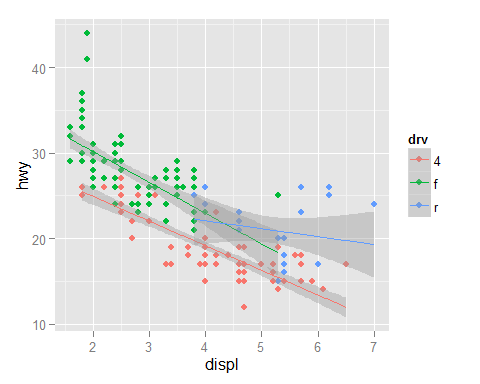
qplot(displ,hwy,data=mpg,facets=.~drv)



### 2.5. 实例

添加回归直线

library(ggplot2)  
qplot(displ,hwy,data=mpg,geom=c("point","smooth"),col=drv,method="lm")



## 3.ggplot()函数

ggplot2的图形的基本元素：

- data frame

- 映射（Mapping）：数据如何映射到颜色、大小形状等几何属性

- 几何对象（Geometric）：点图，线图...

- 统计变换（Statistics）

- 坐标系统（Coordinate）

- 图层（Layer） - 分面（Facet）

qplot()快速绘图，自己设置了很多默认的属性。如果需要自己创建图形的话，就需要使用ggplot()函数，ggplot()是图层+图层的形式。 比如：  
首先需要定义使用的数据，使用aes参数即可，summary()可以查看基本的mapping信息

p<-ggplot(mpg,aes(displ,hwy))  
summary(p)

## data: manufacturer, model, displ, year, cyl, trans, drv, cty, hwy,  
## fl, class [234x11]  
## mapping: x = displ, y = hwy  
## faceting: facet\_null()

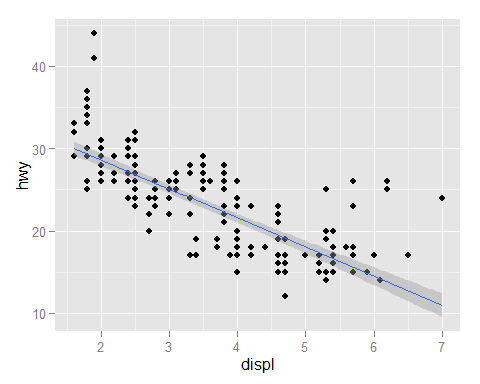
注意：此时是没有图形输出的，因为还没指定图形样式。比如使用散点图显示，用+添加图层即可.  
添加图层的标准方式是

+layer(geom,geom\_par,stat,stat\_par,....)  
e.g. +layer(geom="point")

此外还有一种比较简洁的方式，即

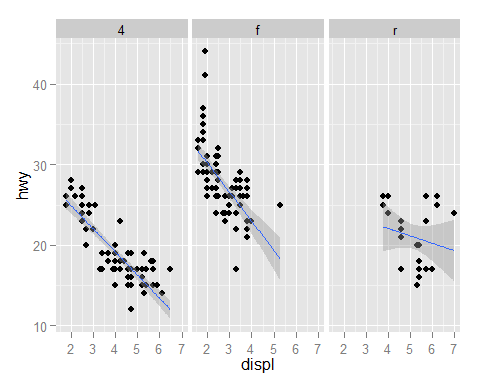
geom\_XXX(geom\_par)

ggplot(mpg,aes(displ,hwy))+  
geom\_point()+  
geom\_smooth(method="lm")



## 3.2添加facets图层

ggplot(mpg,aes(displ,hwy))+geom\_point()+facet\_grid(.~drv)+geom\_smooth(method="lm")



## 3.3 修改其他

修改labs，theme，坐标范围等等，具体的细节查找帮助就可以。

## 