# 先確定下面四個套件已經安裝並在user library有勾選

require(formatR)

require(highlight)

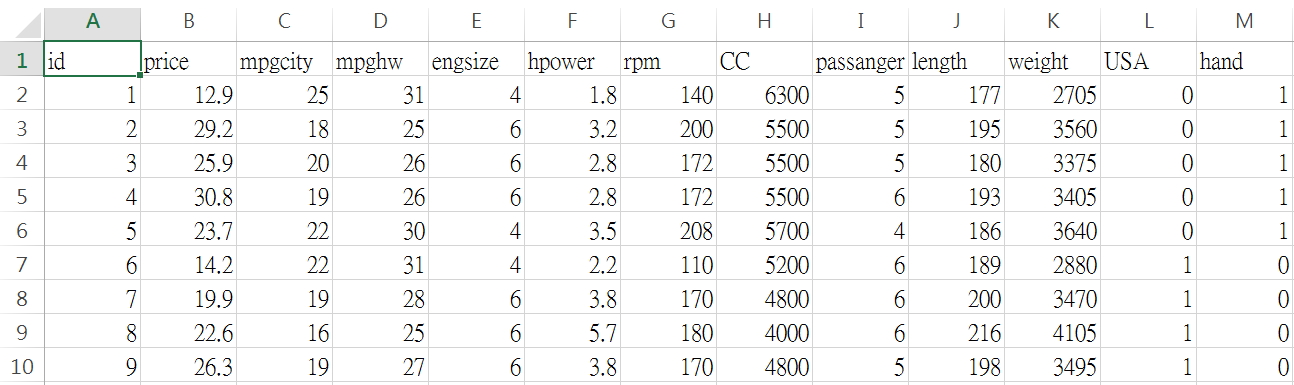
require(xtable)

require(ggplot2)

# 設定工作目錄

setwd("C:/RData")

# 確定EXCEL檔car2013test.csv在目錄C:/RData下

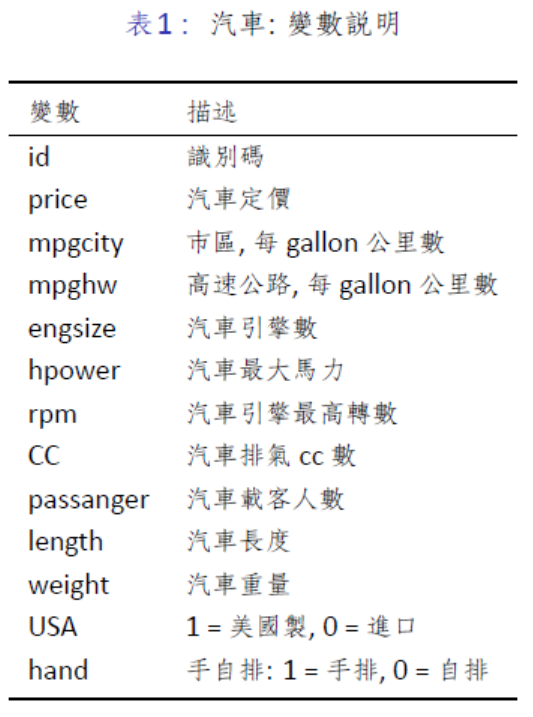


# 將EXCEL檔案讀入R, colClasses行變數的類型, EXCEL檔案中有變數名稱, 所以設定header=TRUE, 列名稱設為無row.names=NULL, 檔案中小數點的符號為dec="."

car2013test <- read.csv("car2013test.csv", header=TRUE, row.names=NULL, dec=".")

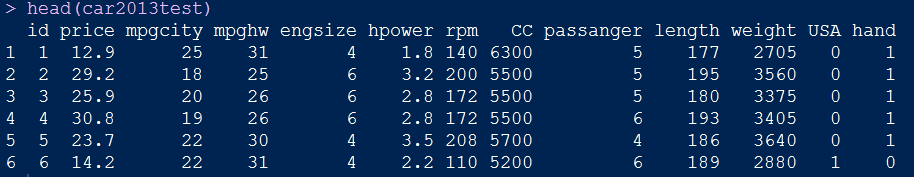
# 檢查資料的dimension, 就是檢查有幾筆, 變數有幾個

 # 有93筆資料, 有13個變數



# 印出汽車這個data frame 的第一部分(前6筆)

head(car2013test)



# 瞭解qplot函數的用法

help(qplot)

### Usage

qplot(x, y = NULL, ..., data, facets = NULL, margins = FALSE,

geom = "auto", stat = list(NULL), position = list(NULL), xlim = c(NA,

NA), ylim = c(NA, NA), log = "", main = NULL,

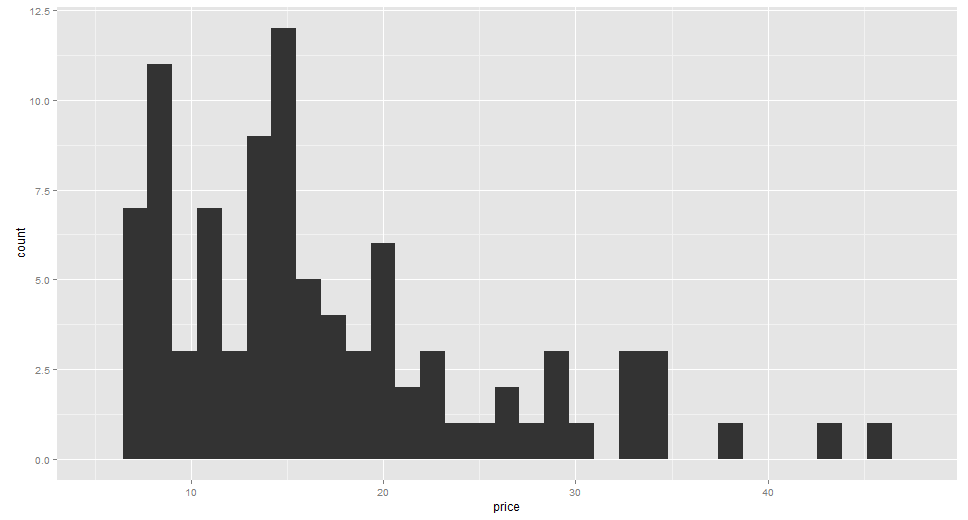
xlab = deparse(substitute(x)), ylab = deparse(substitute(y)), asp = NA)

### Arguments

|  |  |
| --- | --- |
| x | x values |
| y | y values |
| ... | other aesthetics passed for each layer |
| data | data frame to use (optional). If not specified, will create one, extracting vectors from the current environment. |
| facets | faceting formula to use. Picks [facet\_wrap](http://127.0.0.1:25408/help/library/ggplot2/help/facet_wrap) or [facet\_grid](http://127.0.0.1:25408/help/library/ggplot2/help/facet_grid) depending on whether the formula is one sided or two-sided |
| margins | whether or not margins will be displayed |
| geom | character vector specifying geom to use. Defaults to "point" if x and y are specified, and "histogram" if only x is specified. |
| stat | character vector specifying statistics to use |
| position | character vector giving position adjustment to use |
| xlim | limits for x axis |
| ylim | limits for y axis |
| log | which variables to log transform ("x", "y", or "xy") |
| main | character vector or expression for plot title |
| xlab | character vector or expression for x axis label |
| ylab | character vector or expression for y axis label |
| asp | the y/x aspect ratio |

# 畫汽車定價的直方圖

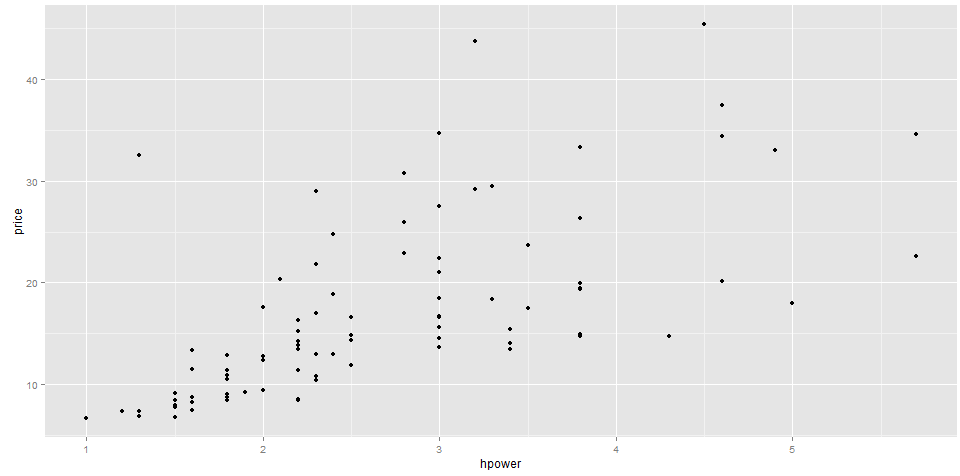
qplot(x=price, data=car2013test)



# 汽車定價的直方圖可知汽車定價的分配是右偏

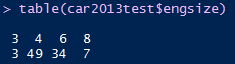
# 畫汽車最大馬力與定價的散佈圖

qplot(x=hpower, y=price, data=car2013test)



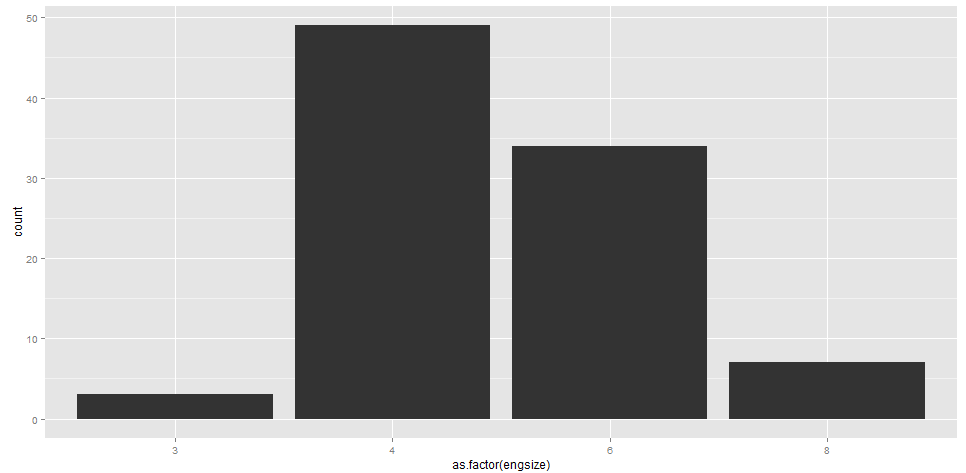
# 由散佈圖可知汽車最大馬力與定價有正相關, 但隨著最大馬力增加, 汽車的定價的變異愈大

# 先將汽車引擎數變數類型設成因子(類別)



# 畫汽車引擎數的直方圖

qplot(x=as.factor(engsize), data=car2013test)

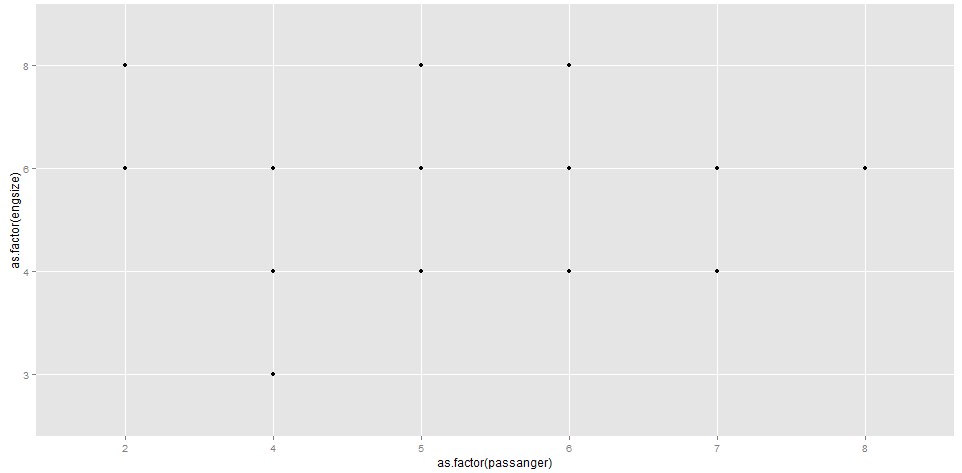


# 由直方圖可知, 引擎數大多數是4或6, 引擎數是3或8的車較少

# 先將汽車引擎數與載客人數兩變數類型都設成因子(類別)

# 再畫汽車載客人數與汽車引擎數的散佈圖

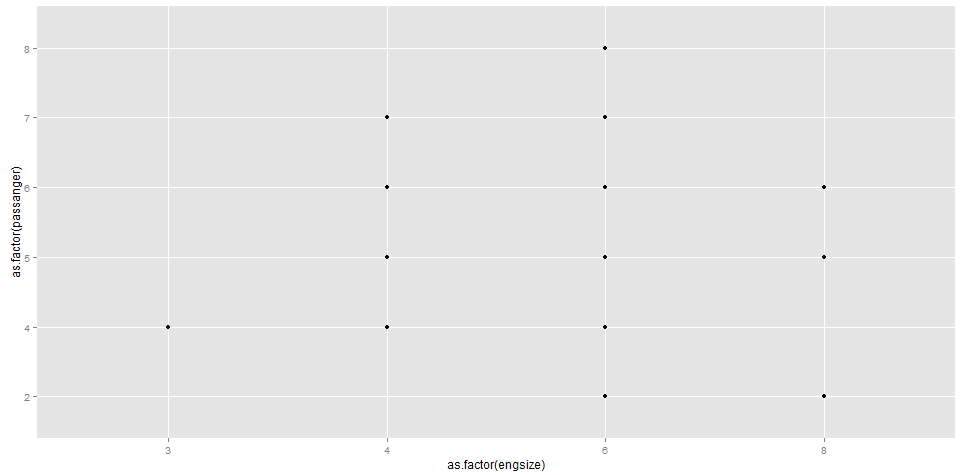
qplot(x=as.factor(passanger), y=as.factor(engsize), data=car2013test)



# 引擎數為6時, 載客人數從2到8人的車都有

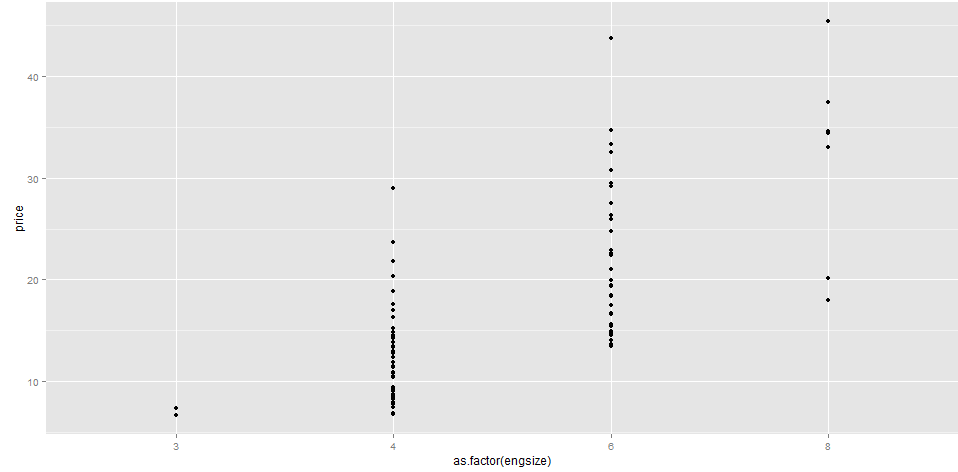
# 再畫汽車引擎數與載客人數的散佈圖

qplot(x=as.factor(engsize), y=as.factor(passanger), data=car2013test)



# 畫汽車引擎數與定價的散佈圖

qplot(x=as.factor(engsize), y=price, data=car2013test)

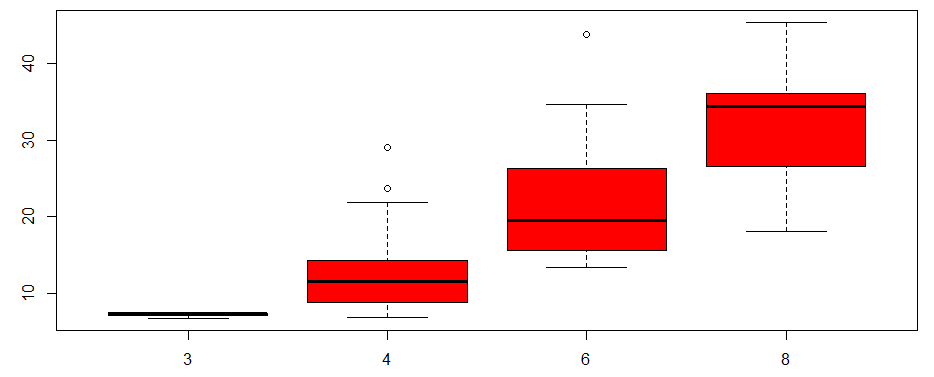


# 可看出汽車引擎數與定價有正相關

# 用盒狀圖表示也很清楚

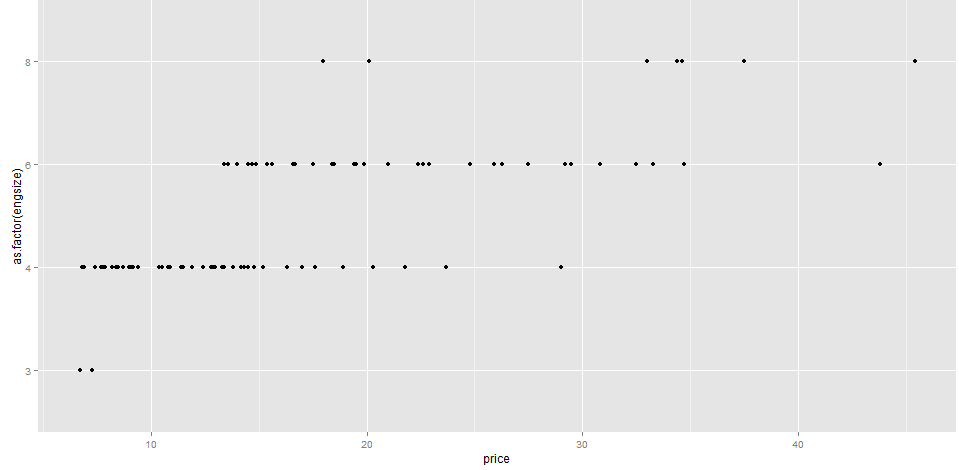
# 畫按引擎數分組的 定價的盒狀圖, 顏色用紅色

boxplot(price ~ engsize, data = car2013test, col = "red")



# 畫汽車定價與引擎數的散佈圖

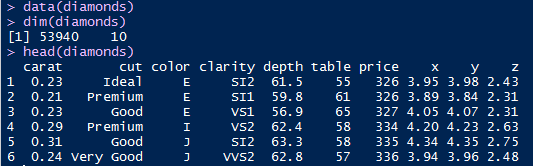
qplot(x=price, y=as.factor(engsize), data=car2013test)



# 引擎數=3時, 汽車定價的變異很小, 隨著引擎數增加, 定價也增加

# 下面的data frame是diamond

# <http://www.diamondse.info/>

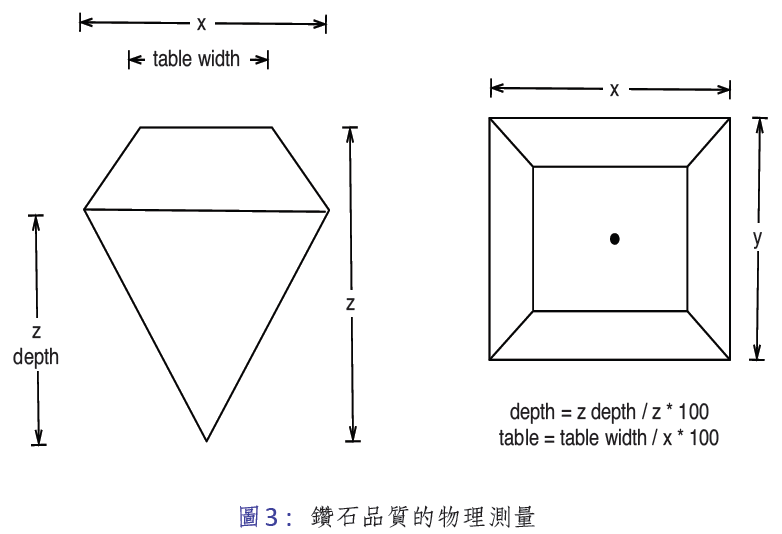


# 共53940筆鑽石資料, 有10個變數

# 鑽石的價值(price)與品質與物理測量值有關

# 品質是指4c(carat, cut, color, clarity)(克拉，切工，顏色，淨度)

# 物理測量有depth, table, x, y, z

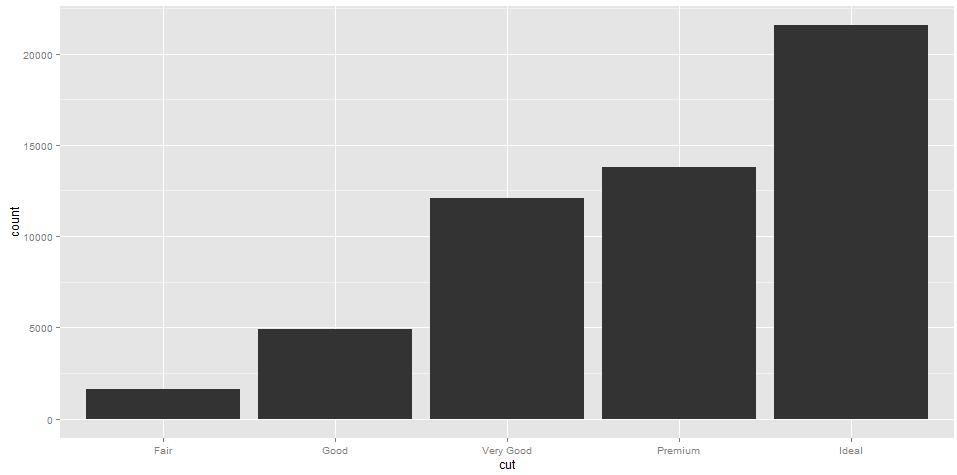


# 切工有5種等級, 以下是次數分配



# 畫切工的長條圖

qplot(cut, data = diamonds, geom = "bar")



# 在53940個鑽石中, 切工為理想的最多(有21551個)

**直方圖(Histogram)**

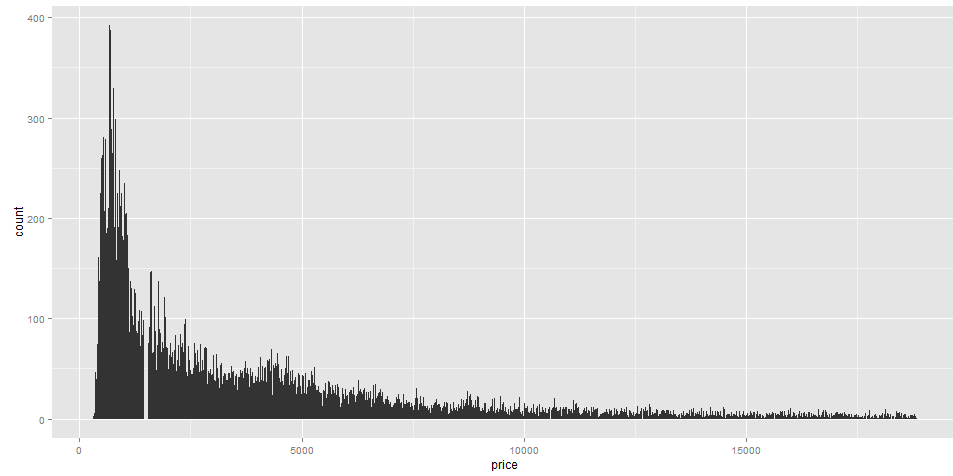
Histogram是最簡單也最常用的密度函數的估計方法，Histogram的估計方法是先將資料分組，每組等寬，寬度(bin width)為，而第一組的起始值(origin)是，分組後的區間為、、，然後再計算落在每組資料的個數。因此的histogram定義如下：



在建立histogram時，必須先選擇起始點及組寬，組寬愈大時組數愈少，組寬愈小時組數愈多，組數太多或太少都不好，資料多組數可以多些，一般組數介於5~20之間較恰當。

# 畫價值的直方圖, 組寬binwidth = 10

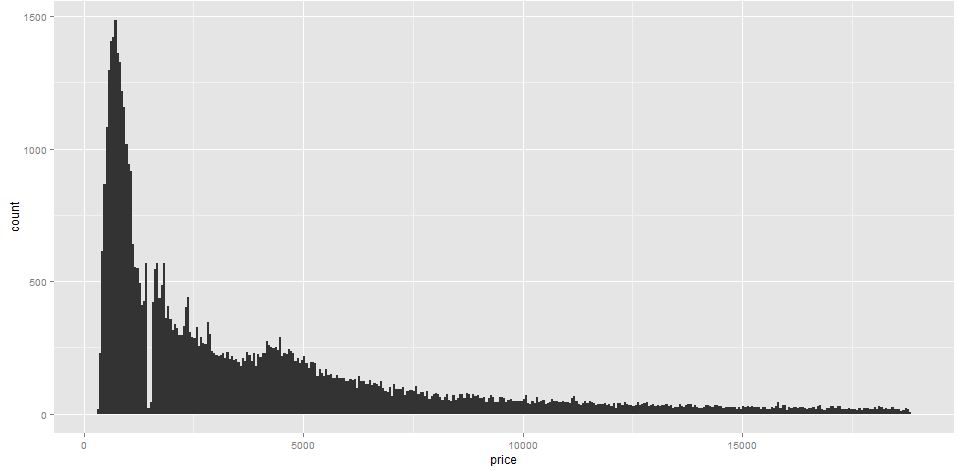
qplot(price, data = diamonds, geom = "histogram", binwidth = 10)



# 由直方圖可知, 鑽石的價值的分配極為右偏

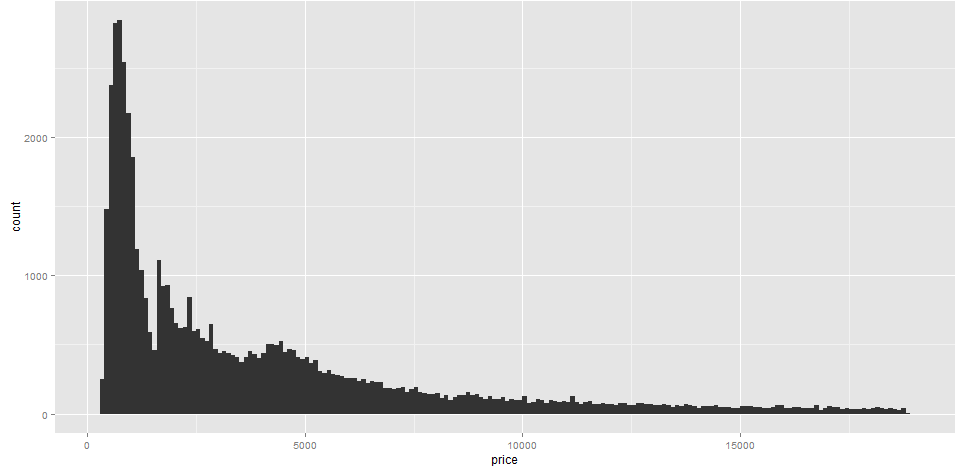
# 畫價值的直方圖, 組寬binwidth = 50

qplot(price, data = diamonds, geom = "histogram", binwidth = 50)



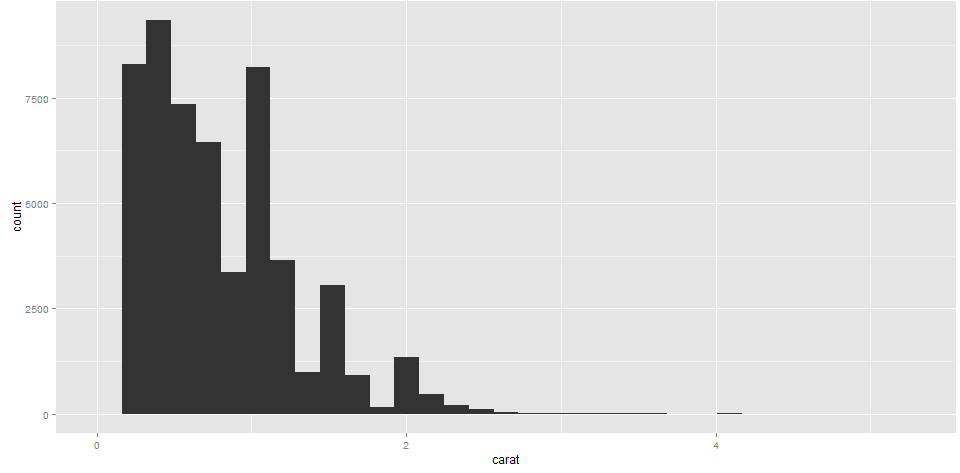
# 畫價值的直方圖, 組寬binwidth = 100

qplot(price, data = diamonds, geom = "histogram", binwidth = 100)



# 畫克拉的直方圖

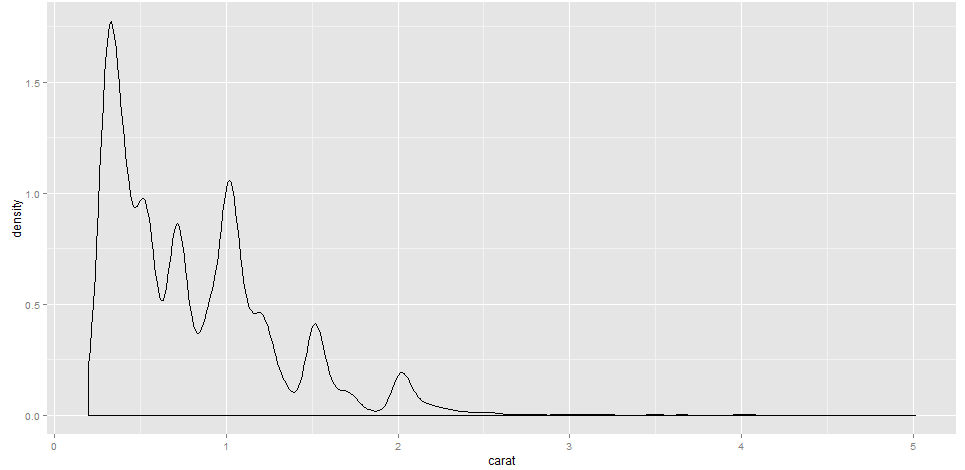
qplot(carat, data = diamonds, geom = "histogram")



# 由直方圖可知, 鑽石的克拉的分配也是右偏

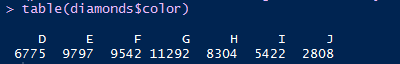
# 畫克拉的估計的密度函數圖

qplot(carat, data = diamonds, geom = "density")



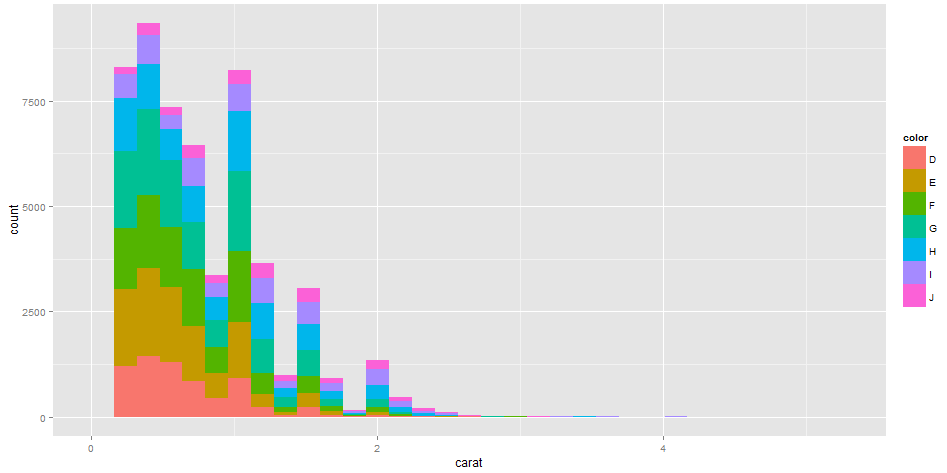
# 由密度函數圖可知, 鑽石的克拉的分配也是右偏, 克拉愈大的鑽石愈少, 有5個高峰, 高峰集中在0.25, 0.75, 1, 1.5, 2克拉的地方

# 鑽石的顏色的次數分配



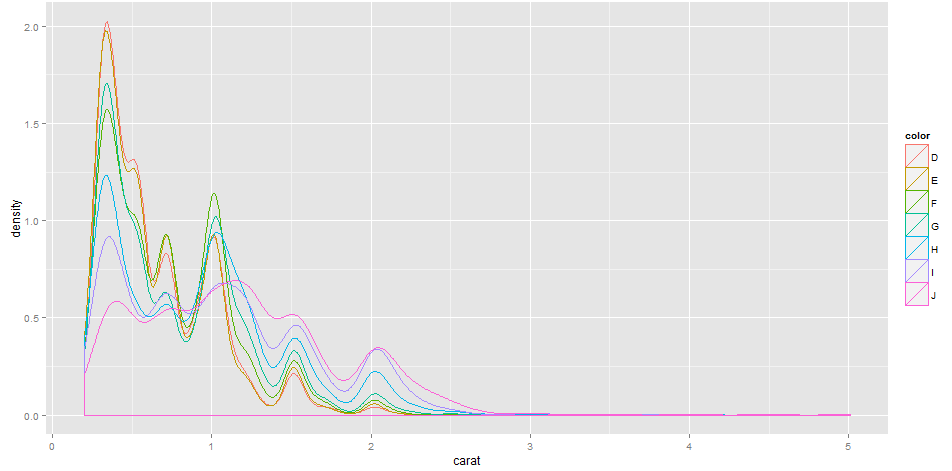
# 畫克拉的直方圖, 直方圖的長條用鑽石的顏色填滿

qplot(carat, data = diamonds, geom = "histogram", fill = color)

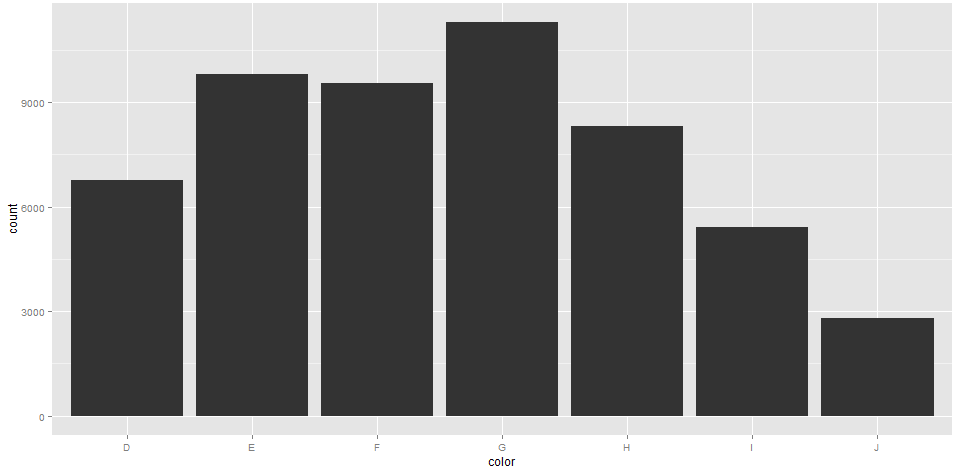


# 畫按鑽石的顏色分組克拉的估計的密度函數圖

qplot(carat, data = diamonds, geom = "density", colour = color)



# 畫鑽石的顏色的長條圖



# 按鑽石的顏色分組算克拉的平均數存到a

> a<-tapply(diamonds$carat, diamonds$color, mean)

> a

D E F G H I J

0.6577948 0.6578667 0.7365385 0.7711902 0.9117991 1.0269273 1.1621368

# 鑽石的次數分配

> table(diamonds$color)

D E F G H I J

6775 9797 9542 11292 8304 5422 2808

# 鑽石的次數分配存到b

> b<-table(diamonds$color)

# 計算顏色次數的加權值(用鑽石平均克拉做為權重)

> a\*b

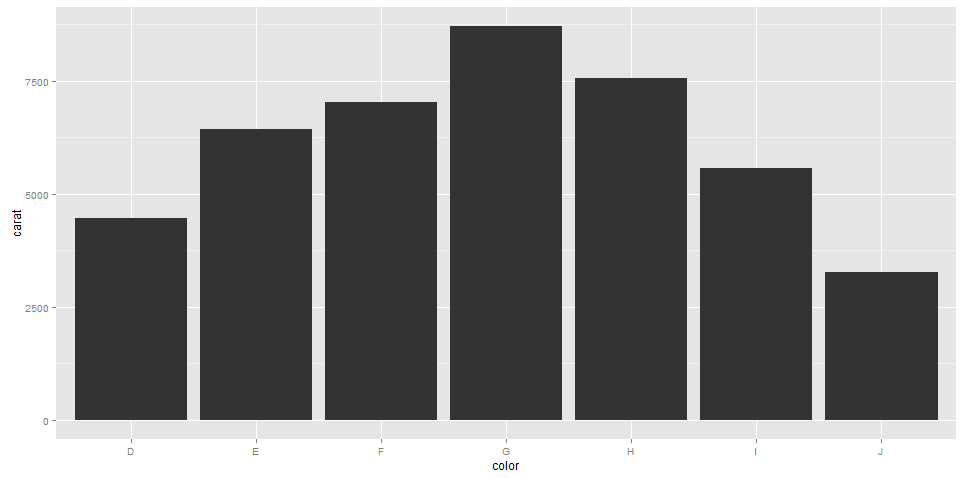
D E F G H I J

4456.56 6445.12 7028.05 8708.28 7571.58 5568.00 3263.28

# 畫用鑽石克拉加權的顏色的長條圖, 指令scale\_y\_continuous("carat")是指將軸的標記改為carat

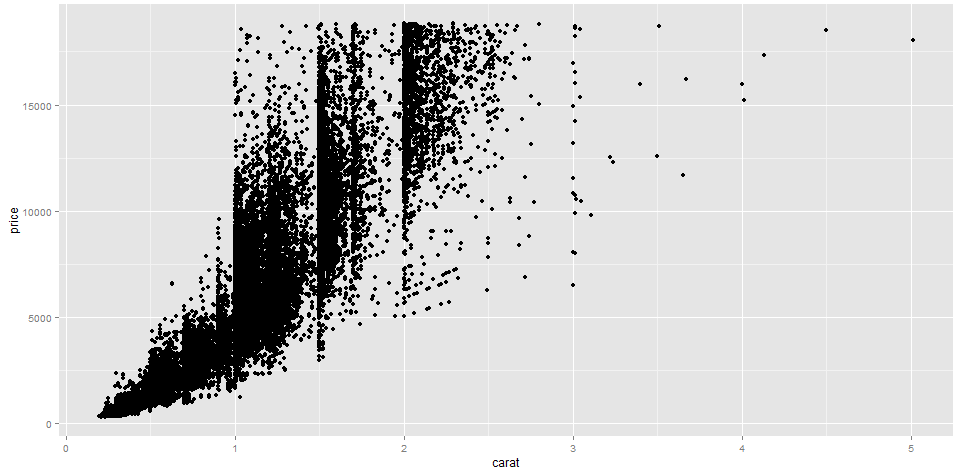
qplot(color, data = diamonds, geom = "bar", weight = carat) +

scale\_y\_continuous("carat")



# 畫鑽石克拉與價值的散佈圖

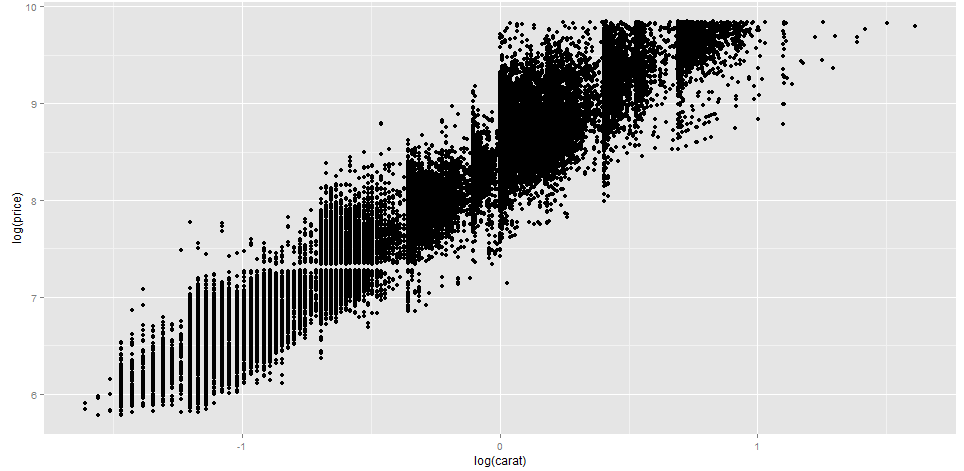
qplot(carat, price, data = diamonds)



# 鑽石克拉與價值的關係是非線性的

# 畫鑽石log(克拉)與log(價值)的散佈圖

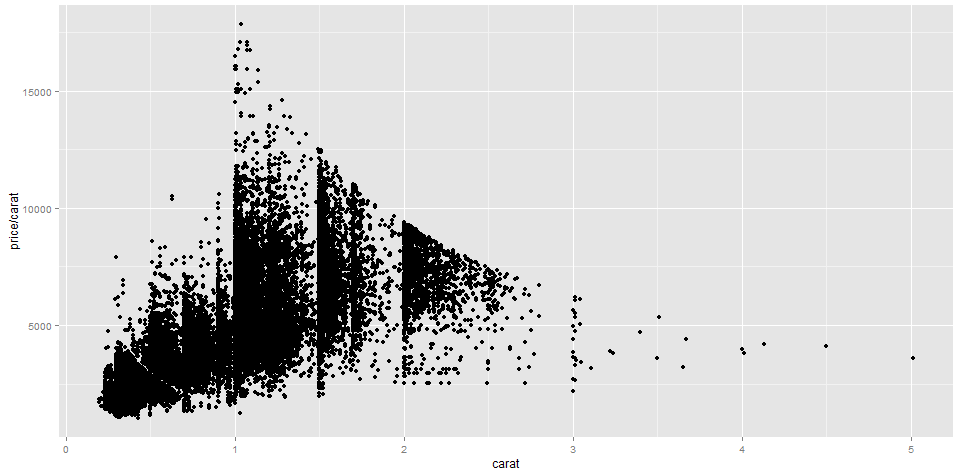
qplot(log(carat), log(price), data = diamonds)



# 鑽石log(克拉)與log(價值)的關係是線性的

# 畫鑽石克拉與每克拉單價的散佈圖

qplot(carat, price/carat, data = diamonds)



# 一克拉以上的鑽石, 克拉愈大, 每克拉的單價愈低

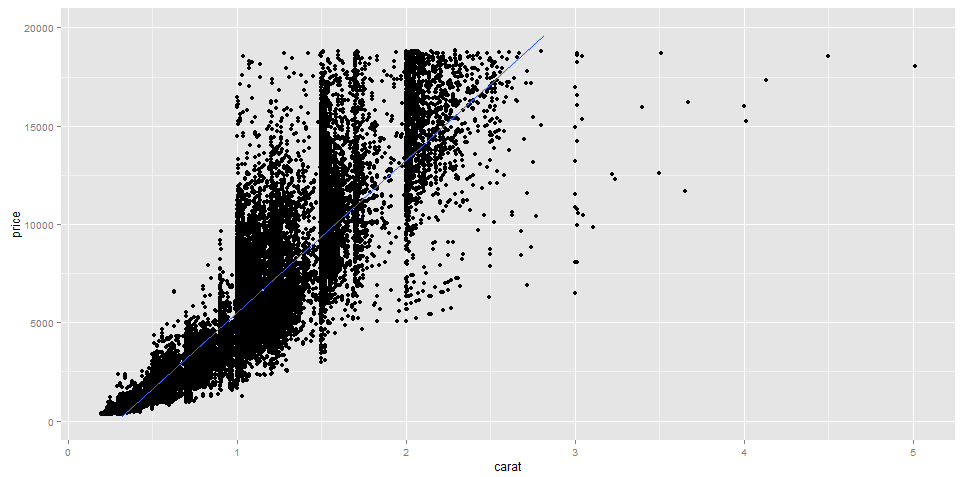
# 畫鑽石克拉與價值的散佈圖, ylim = c(0, 20000) 將y軸範圍改成0到20000,

圖上畫點與平滑線, 線是直線

qplot(carat, price, data = diamonds,

geom = c("point", "smooth"),

method = lm, ylim = c(0, 20000))



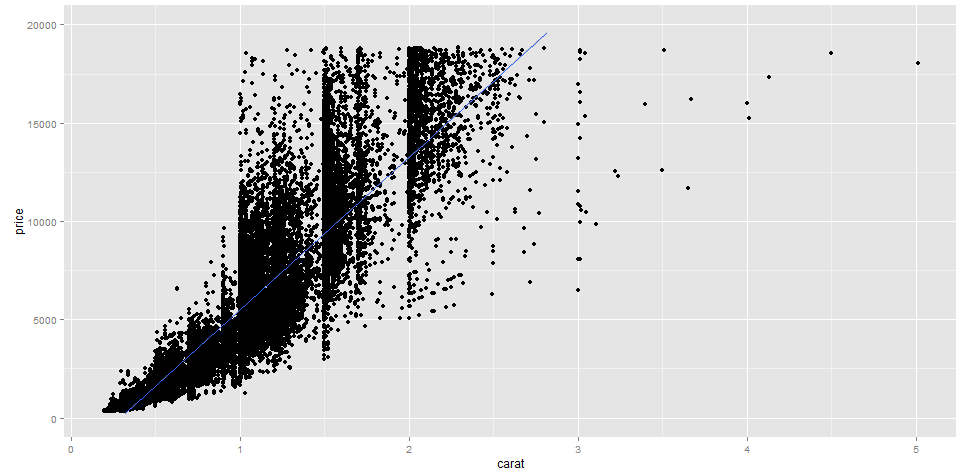
# 另一種方法

# 畫鑽石克拉與價值的散佈圖, ylim = c(0, 20000) 將y軸範圍改成0到20000,

圖上畫平滑直線

qplot(carat, price, data = diamonds,

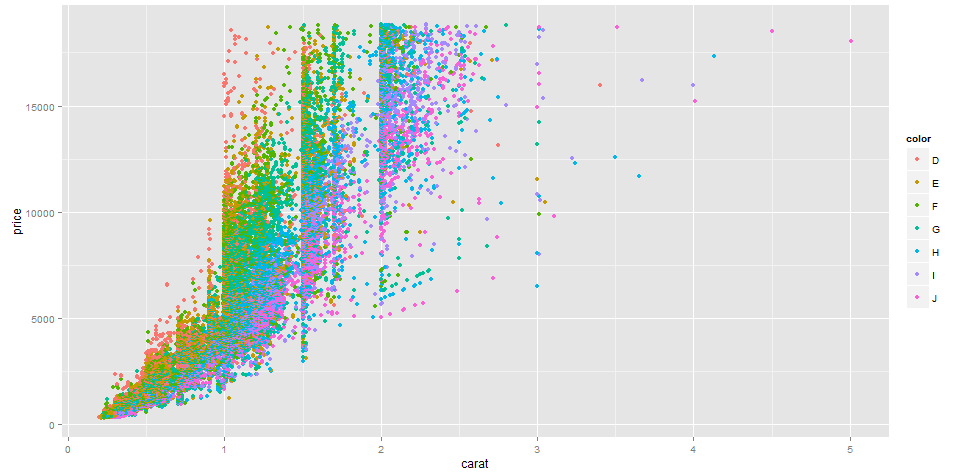
ylim = c(0, 20000)) + geom\_smooth(method = lm)



# 增加美畫

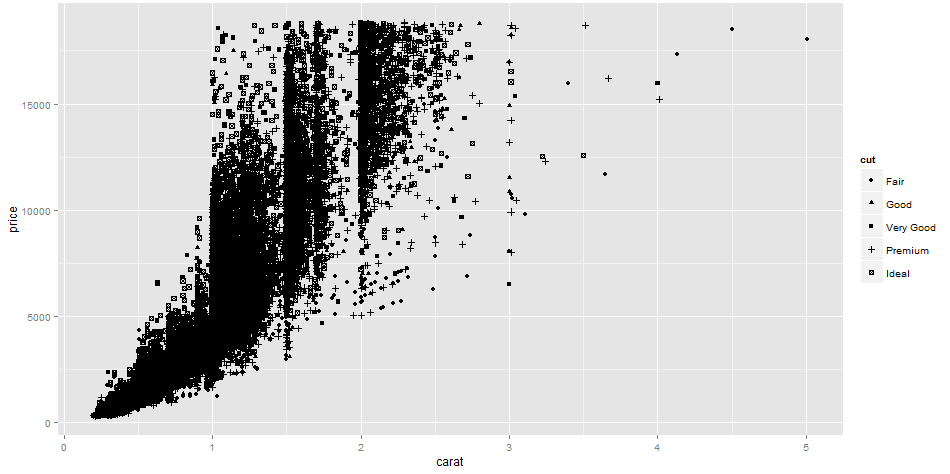
# 畫按鑽石的顏色分組的鑽石克拉與價值的散佈圖, 點顏色按鑽石顏色

qplot(carat, price, data = diamonds, colour = color)



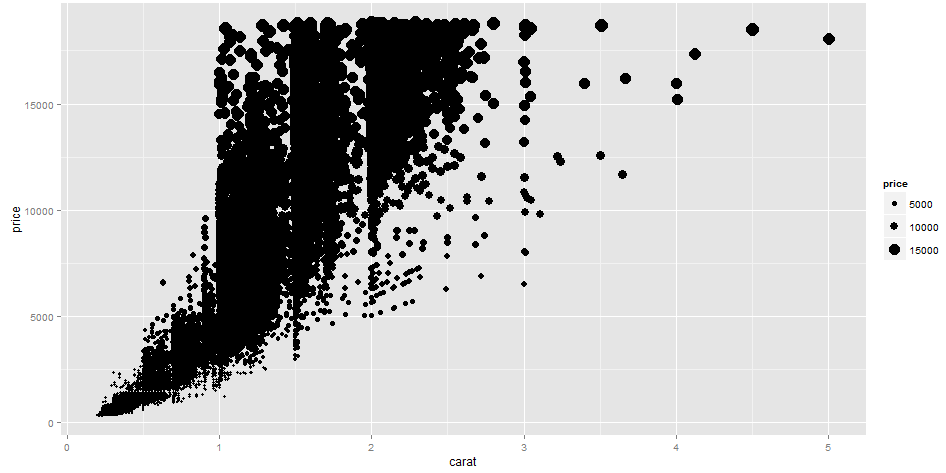
# 畫按鑽石的切割分組鑽石克拉與價值的散佈圖, 點的形狀按鑽石切割

qplot(carat, price, data = diamonds, shape = cut)



# 畫按鑽石的價值分組鑽石克拉與價值的散佈圖, 點的大小按鑽石價值

qplot(carat, price, data = diamonds, size = price)



library(scales)

# 上面圖資料點太多, 很多重疊, 將點縮小

alpha("black")

[1] "#000000FF"

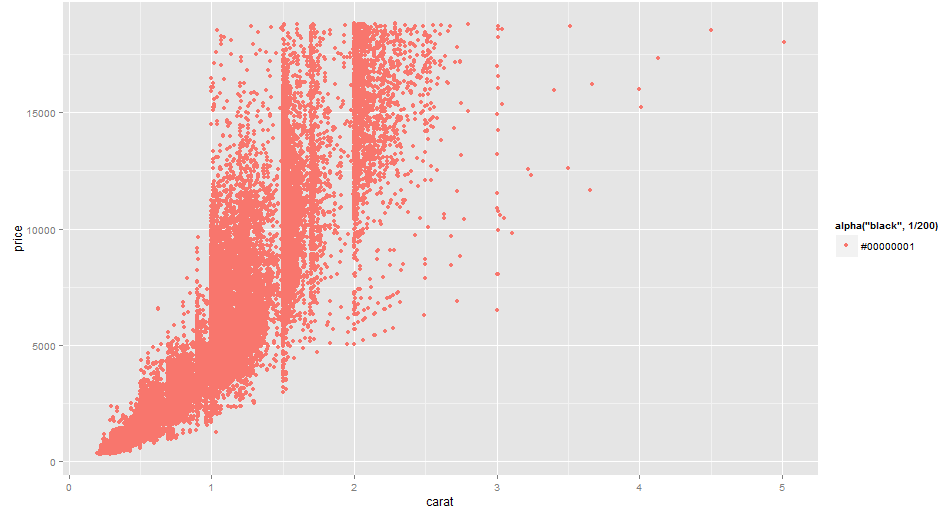
alpha("black", 1/200)

[1] "#00000001"

# 畫鑽石克拉與價值的散佈圖

qplot(carat, price, data = diamonds,

colour = alpha("black", 1/200))



# 資料點太多, 很多重疊, 修改顏色的透明度為1/200。modify the fill colour

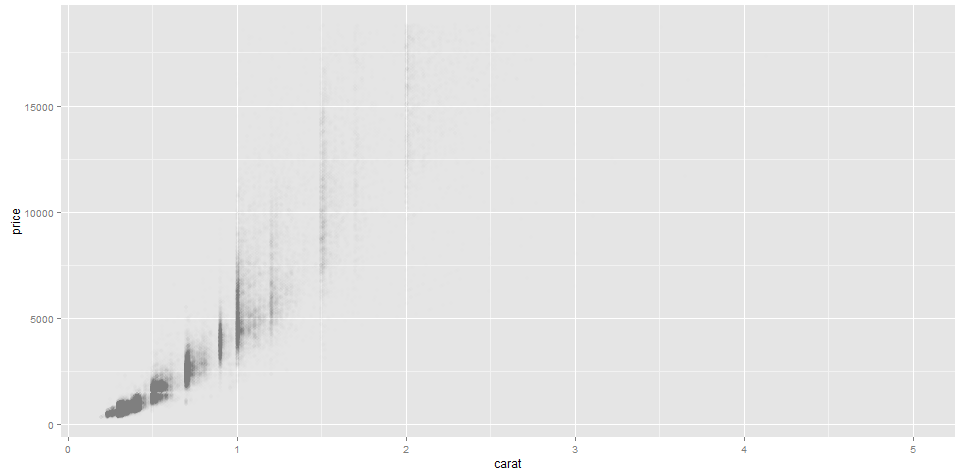
I(alpha("black", 1/200))

[1] "#00000001"#

# 畫鑽石克拉與價值的散佈圖

qplot(carat, price, data = diamonds,

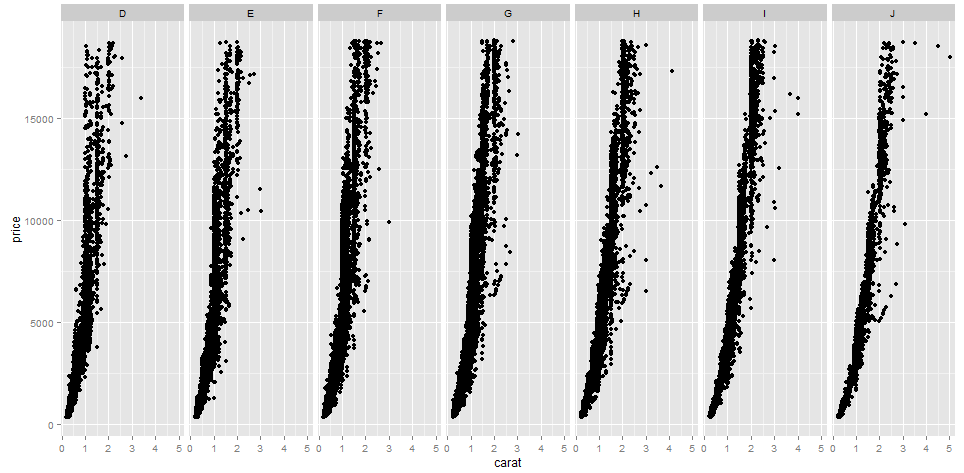
colour = I(alpha("black", 1/200)))



# 分成小面

# 畫按顏色分組的各別的鑽石克拉與價值的散佈圖

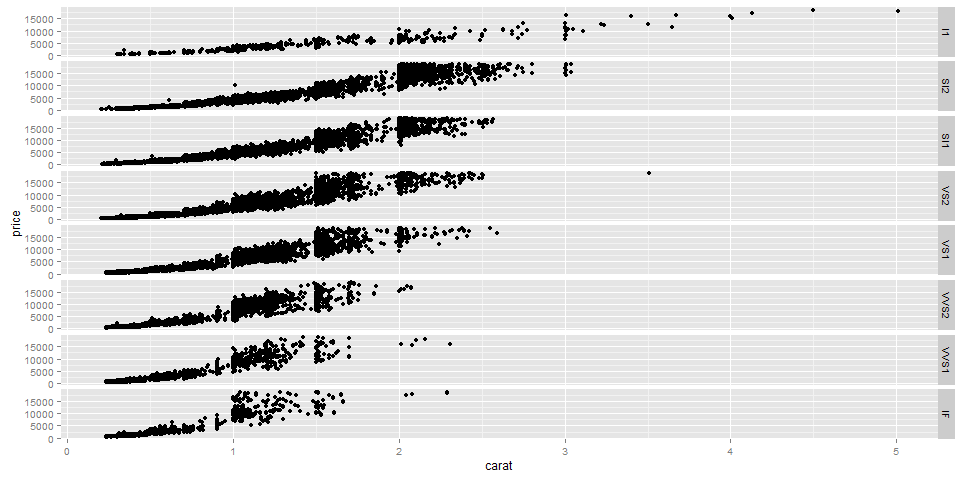
qplot(carat, price, data = diamonds, facets = . ~ color) # 顏色在x軸



# 固定顏色時, 克拉愈高價值也愈高

# 畫按淨度分組的各別的鑽石克拉與價值的散佈圖

qplot(carat, price, data = diamonds, facets = clarity ~ .) # 淨度在y軸

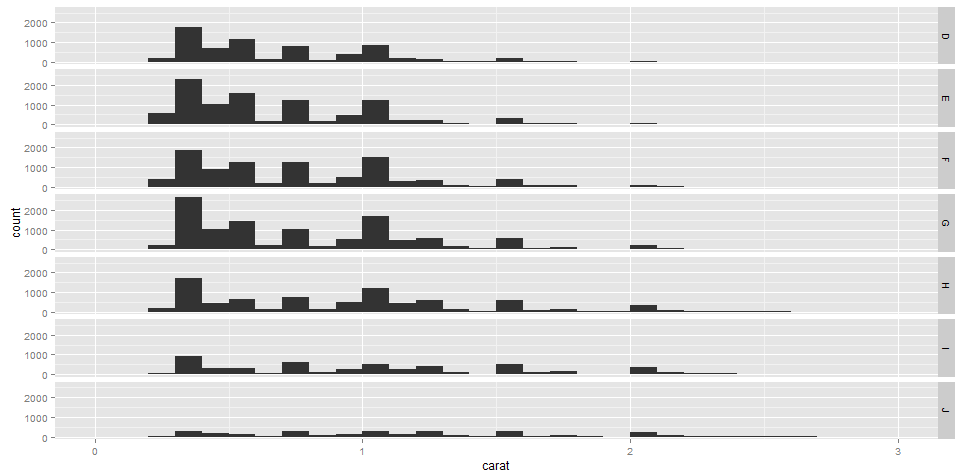


# 固定淨度時, 克拉愈高價值也愈高

# 畫按顏色分組的各別的鑽石克拉的直方圖

qplot(carat, data = diamonds, facets = color ~ .,

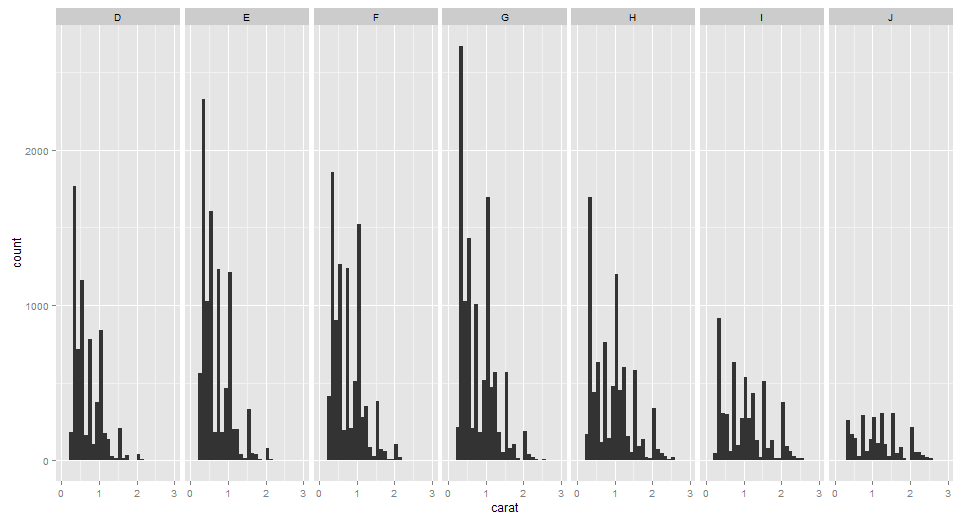
geom = "histogram", binwidth = 0.1, xlim = c(0, 3)) # 顏色在y軸



# 畫按顏色分組的各別的鑽石克拉的直方圖

qplot(carat, data = diamonds, facets = . ~ color,

geom = "histogram", binwidth = 0.1, xlim = c(0, 3)) # 顏色在x軸



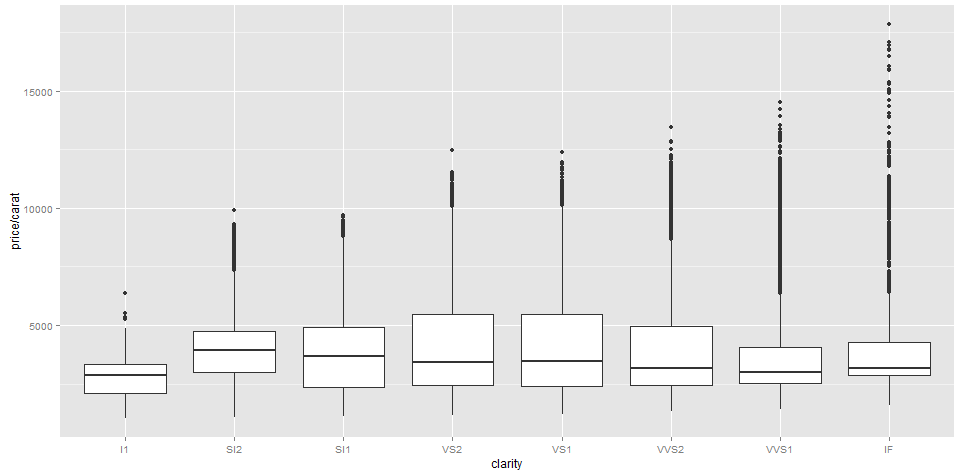
# 不論哪種顏色, 鑽石克拉小的都比較多

# 淨度的次數分配



# 畫按淨度分組的每克拉單價的盒型圖

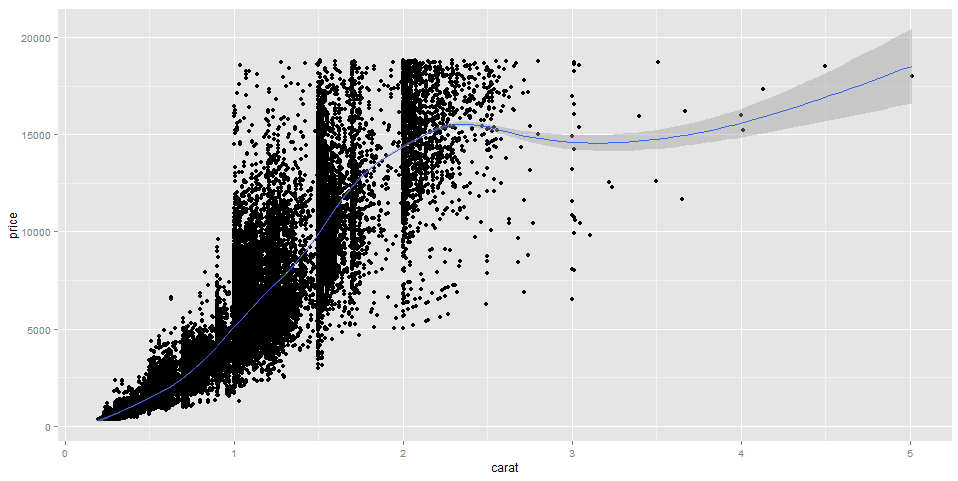
qplot(clarity, price/carat, data = diamonds, geom = "boxplot")



# 不論哪種淨度, 每克拉單價有相當多的極端值

# 畫鑽石克拉與價值的散佈圖, 畫點與平滑的估計曲線加信賴界限

qplot(carat, price, data = diamonds, geom = c("point", "smooth"))



# 信賴界限愈來愈寬因為那裡資料很少

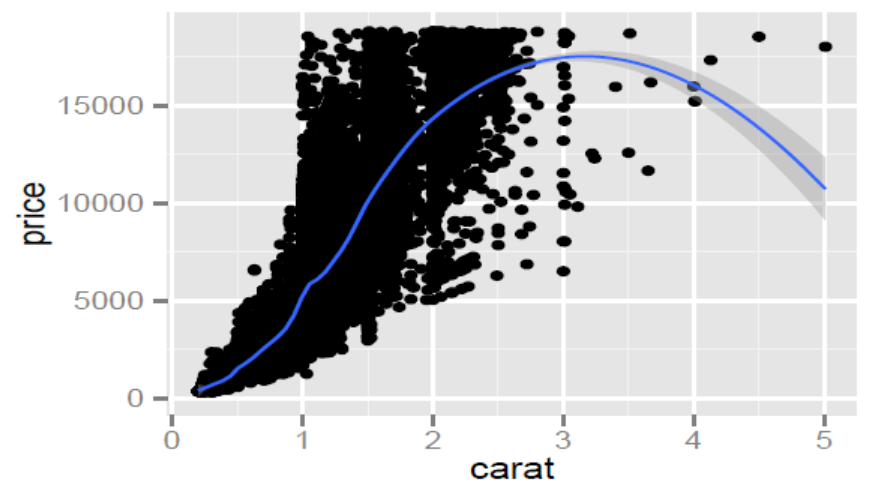
# loess 是Local Polynomial Regression Fitting的縮寫

# 改用另一種平滑法

# 畫鑽石克拉與價值的散佈圖, 畫點與平滑的估計曲線加信賴界限

qplot(carat, price, data = diamonds, geom = c("point", "smooth"),

method = "loess", span = 0.2, se = TRUE)



# 結果完成不同

require(mgcv)

# Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation and GAMMs by REML/PQL

mgcv provides functions for generalized additive modelling (gam and bam) and generalized additive mixed modelling (gamm, and random.effects). The term GAM is taken to include any GLM estimated by quadratically penalized (possibly quasi-) likelihood maximization.

Particular features of the package are facilities for automatic smoothness selection, and the provision of a variety of smooths of more than one variable. User defined smooths can be added. A Bayesian approach to confidence/credible interval calculation is provided. Linear functionals of smooths, penalization of parametric model terms and linkage of smoothing parameters are all supported. Lower level routines for generalized ridge regression and penalized linearly constrained least squares are also available.

# 再用另一種平滑法

# 畫鑽石克拉與價值的散佈圖, 畫點與平滑的估計曲線加信賴界限

qplot(carat, price, data = diamonds, geom = c("point", "smooth"),

method = "gam", formula = y ~ s(x, bs = "cs"))

