R ggplot2 Assignment 2

Yan Li

Aug 8th 2015

We first have to load the Required packages.

#Load required packages in the user-defined library location  
.libPaths("C:/work/Rpackages/R3.2.1")  
suppressMessages(library(ggplot2, quietly = T))  
suppressMessages(library(dplyr, quietly = T))  
suppressMessages(library(mosaicData, quietly = T))  
  
#turn off warning messages  
options(warn=-1)

The data used for this assignment is mpg or diamonds from ggplot2 package.

#Show the top 6 rows of the mpg/diamonds data to know the data information  
head(mpg)

## manufacturer model displ year cyl trans drv cty hwy fl class  
## 1 audi a4 1.8 1999 4 auto(l5) f 18 29 p compact  
## 2 audi a4 1.8 1999 4 manual(m5) f 21 29 p compact  
## 3 audi a4 2.0 2008 4 manual(m6) f 20 31 p compact  
## 4 audi a4 2.0 2008 4 auto(av) f 21 30 p compact  
## 5 audi a4 2.8 1999 6 auto(l5) f 16 26 p compact  
## 6 audi a4 2.8 1999 6 manual(m5) f 18 26 p compact

head(diamonds)

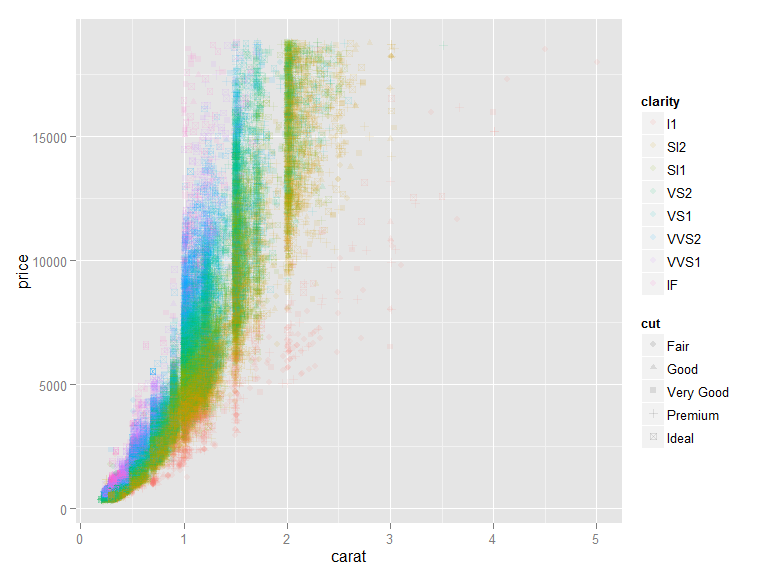
## carat cut color clarity depth table price x y z  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.29 Premium I VS2 62.4 58 334 4.20 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75  
## 6 0.24 Very Good J VVS2 62.8 57 336 3.94 3.96 2.48

This assignment is to create 3 plots, each to answer one question and with 3-5 iterations to show the process.

The first question is about the diamonds data: what is the price change along with the carat for different cut and clarity.

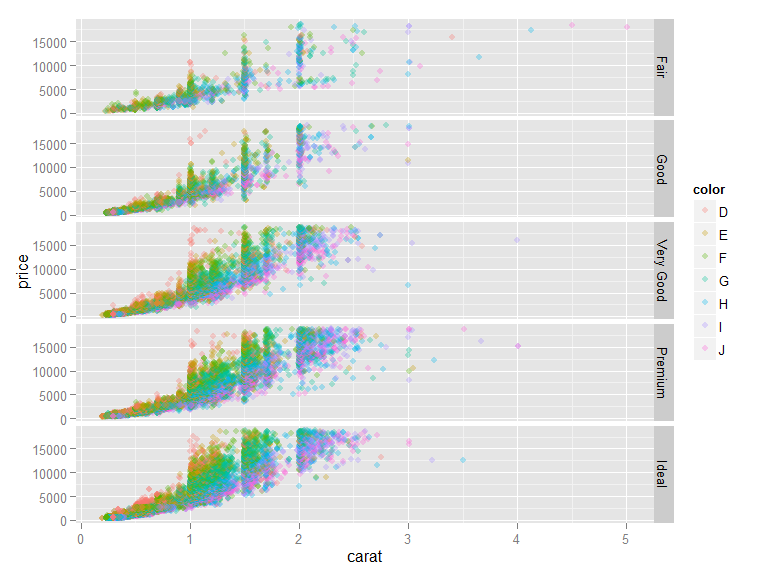
First we see the price change with shape = cut and color = clarity:

#Basic Model  
p <- ggplot( data=diamonds, aes(x=carat, y=price, shape = cut, color= clarity) )   
#Scatter plot   
 p+ geom\_point( alpha=0.1 )

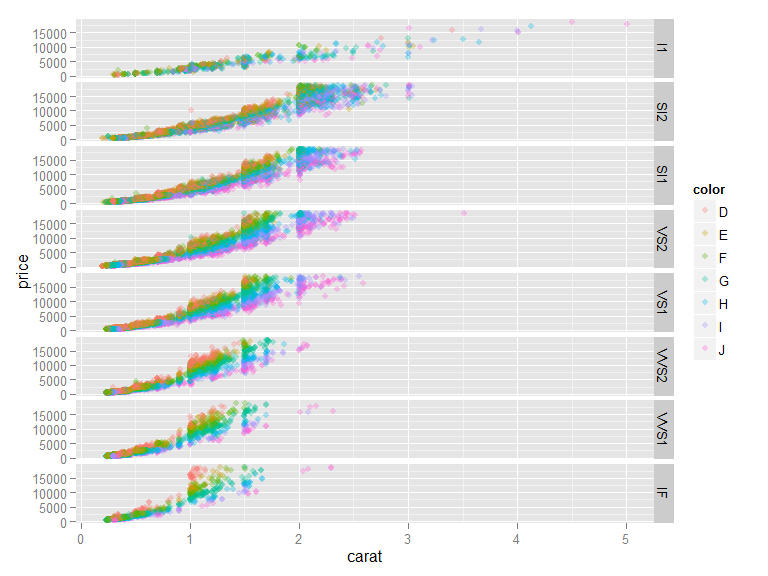


It is not very clear to see the price change if the customer has a clear requirement of clarity and cut. And it takes a long time to get the plots (over 5 mins). So we can subset the data into different group based on clarity and cut:

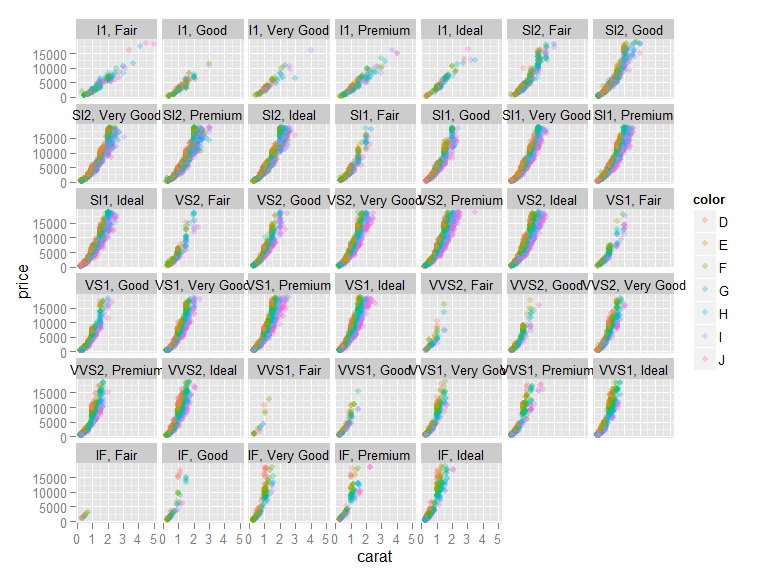
#Delete the shape to speed up the generation of the plot   
p2 <- ggplot( data=diamonds, aes(x=carat, y=price, color= color ) )  
  
#Subset data to group based on clarity and cut  
#First we check facet = cut~.  
p2+ geom\_point( alpha=0.3, size = I(2))+  
 facet\_grid(cut~. )



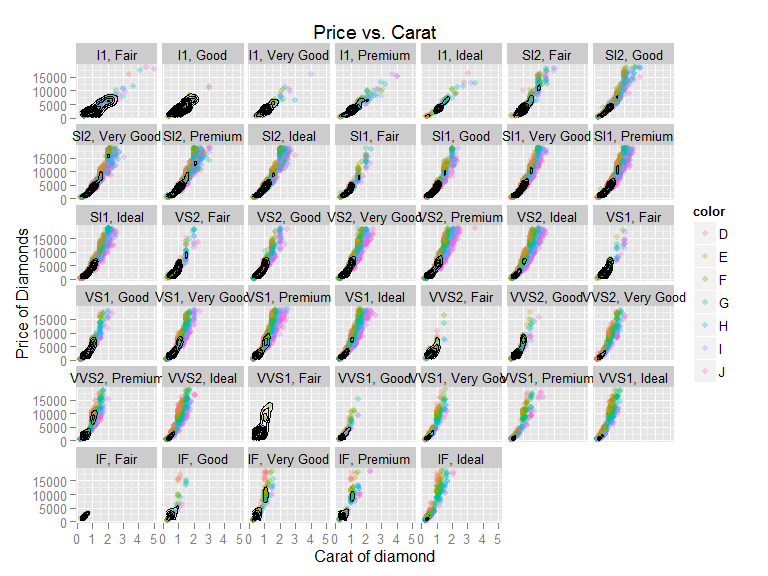
#Second we check facet = clarity ~.  
p2+ geom\_point( alpha=0.3, size = I(2))+  
 facet\_grid(clarity~. )



#Last facet = clarity ~cut  
p2+ geom\_point( alpha=0.3, size = I(2))+  
 facet\_wrap(clarity~cut )



#Add 2d density  
#Add title  
 p2+ geom\_point( alpha=0.3, size = I(2))+  
 facet\_wrap(clarity~cut )+stat\_density2d(color = "black") +  
 labs(title="Price vs. Carat", x="Carat of diamond", y="Price of Diamonds")

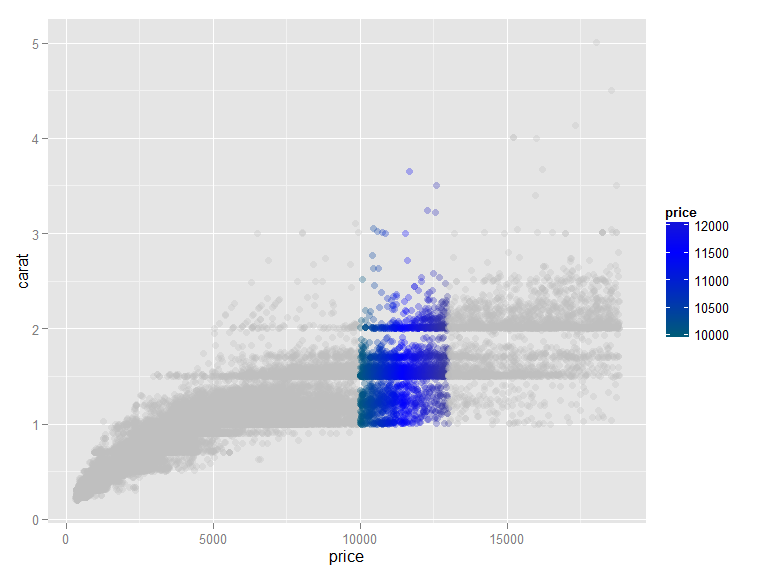


From the above plots, customers can easily find any combination of clarity and cut subgroup. Then they can further examine the price change.

Second question is: How many choices do I have if my bugget of buying a diamond is $10,000-$12,000 (the range of the diamonds data for price is $326-$18,823)?

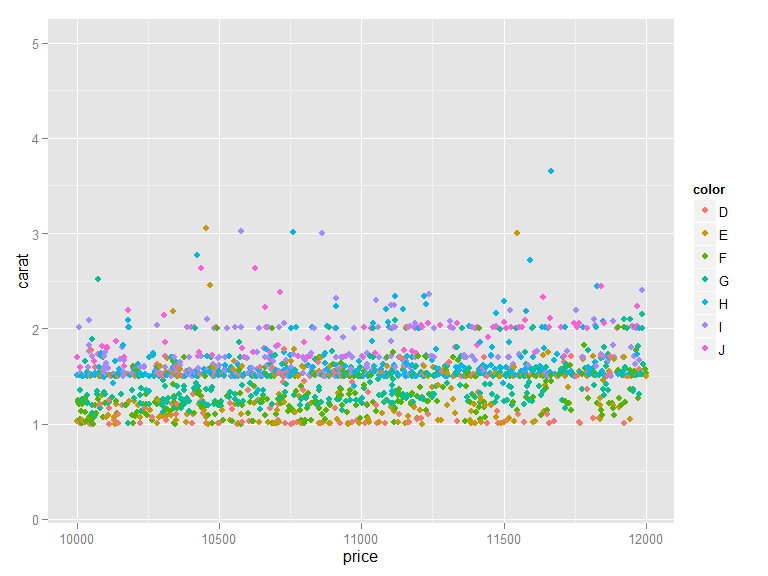
This needs to limit the price range.

##limit data range and set 'outlier' color  
require(scales)  
p3 <- ggplot( data=diamonds, aes(price, carat) )   
  
p3+  
geom\_point(aes(color=price), size=2.5,alpha = 0.3) +  
 scale\_color\_gradient2(low=muted("light blue"), high=muted("dark blue"), mid="blue",  
 midpoint = 11500, limits=c(10000,12000), na.value="gray")

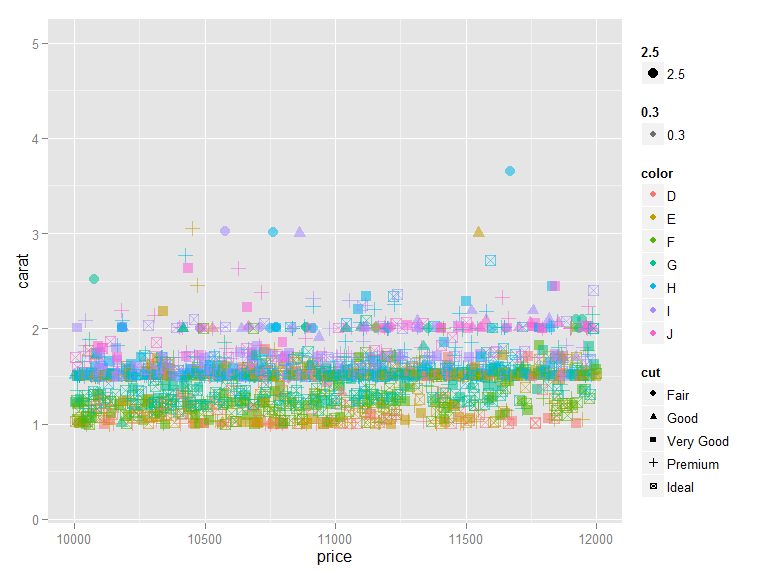


Now We focus on the the price range(10000,12000).

p3+geom\_point(aes(colour = color))+  
 scale\_x\_continuous(limits=c(10000,12000))

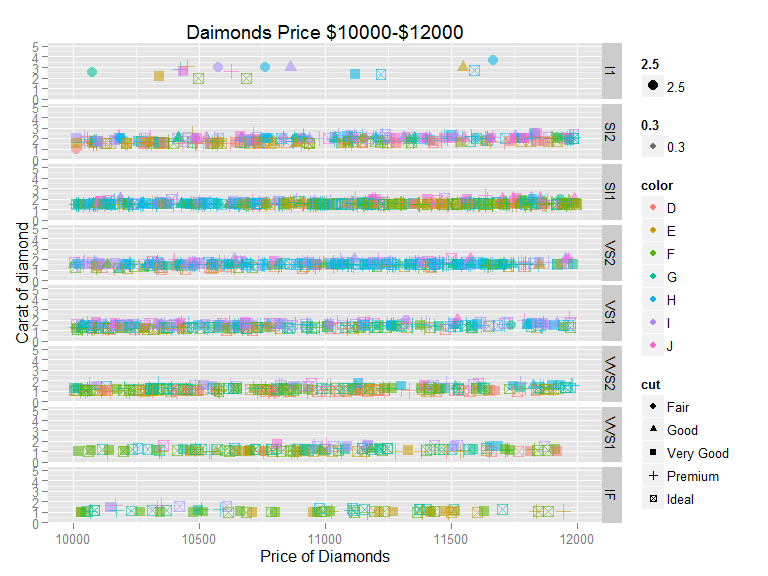


#Add shape, size and alpha to make a better looking plot  
p3+geom\_point(aes(colour = color, shape= cut, size = 2.5, alpha = 0.3))+  
 scale\_x\_continuous(limits=c(10000,12000))



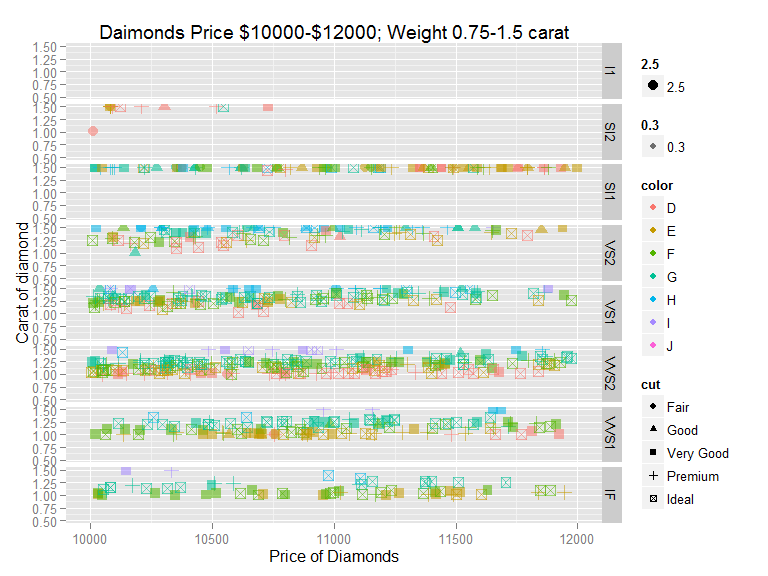
It is not easy to find cut,clarity features. So subsetting the data by clarity.

p3+geom\_point(aes(colour = color, shape = cut,size = 2.5, alpha = 0.3))+  
 facet\_grid(clarity~.)+  
 scale\_x\_continuous(limits=c(10000,12000))+  
 labs(title="Daimonds Price $10000-$12000", x="Price of Diamonds", y="Carat of diamond")



Now we may want to limit the weight of the diamonds too:

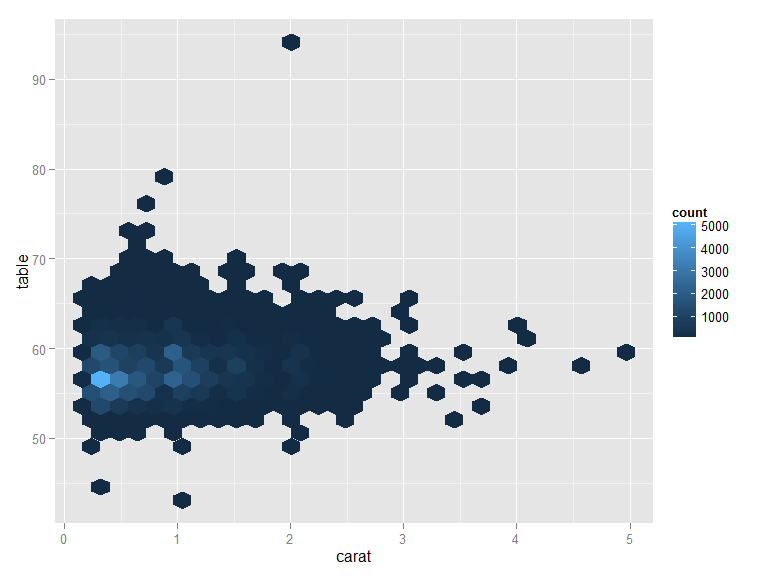
#Limit the carat  
p3+geom\_point(aes(colour = color, shape = cut,size = 2.5, alpha = 0.3))+  
 facet\_grid(clarity~.)+  
 scale\_y\_continuous(limits=c(0.5,1.5))+  
 scale\_x\_continuous(limits=c(10000,12000))+  
 labs(title="Daimonds Price $10000-$12000; Weight 0.75-1.5 carat", x="Price of Diamonds", y="Carat of diamond")



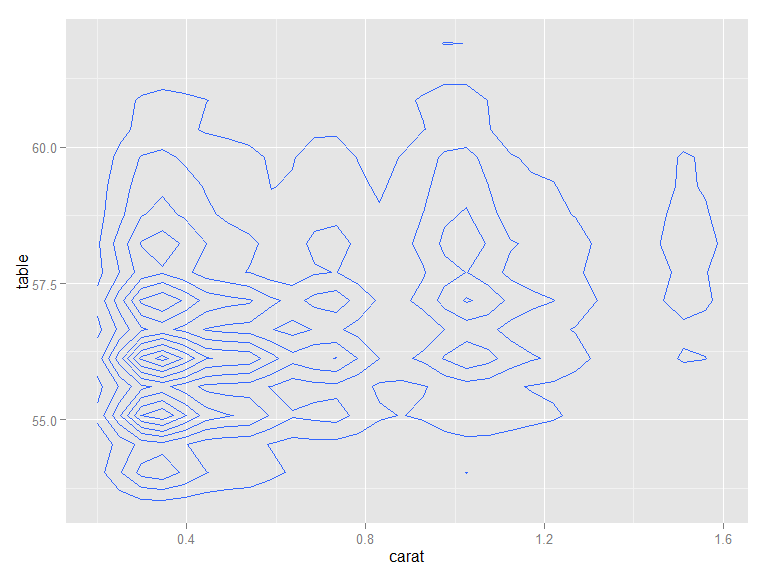
Now we can choose the diamonds based on our buget.

The third plot is about: How does the geometric measures(table) of diamonds relate to the weight (carat)

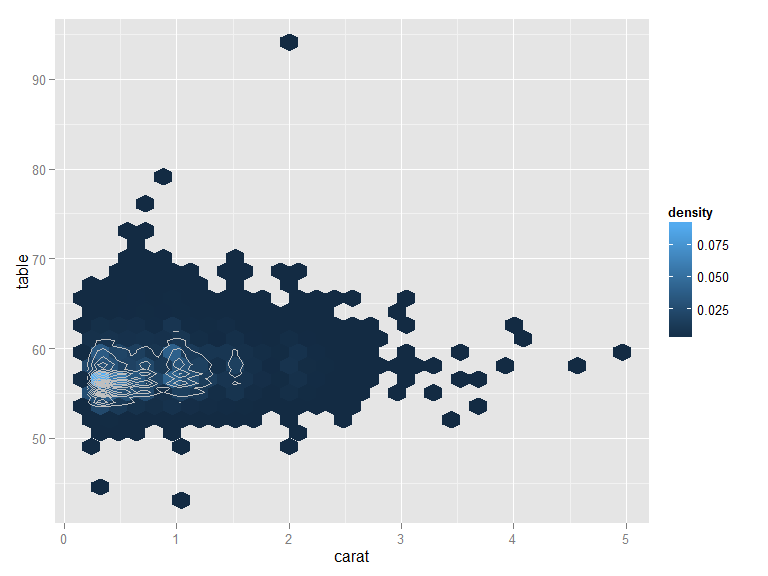
#2d histogram using hexagonal shape  
require(hexbin)  
p4 <- ggplot( diamonds, aes(x = carat, y=table) )  
  
p4 + stat\_binhex()



#2d kernel density  
p4 + stat\_density2d()



#combine the above two  
p4 +   
 stat\_binhex(aes(fill=..density..)) +  
 stat\_density2d(color="gray75", size=.3)



#Use tile showing the density  
p4 +xlim(1,3)+  
 # stat\_binhex(aes(fill=..density..)) +  
 stat\_density2d(geom = "tile",aes(fill=..density..),contour=F)+  
scale\_fill\_gradient(limit = c(1e-5,0.45))

