### Create a histogram of diamond prices.

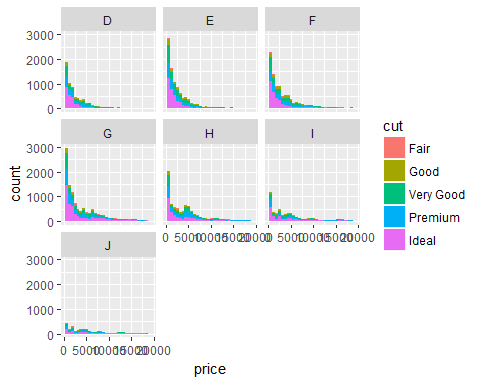
### Facet the histogram by diamond color and use cut to color the histogram bars.

library(ggplot2)  
data(diamonds)  
names(diamonds)

## [1] "carat" "cut" "color" "clarity" "depth" "table" "price"   
## [8] "x" "y" "z"

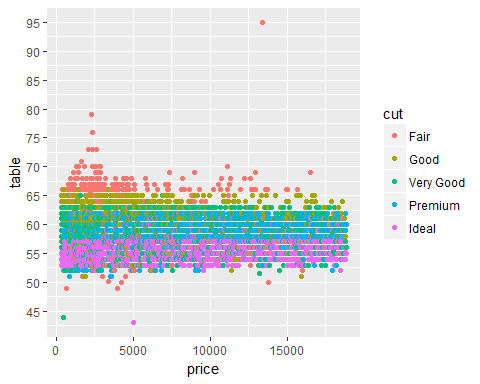
ggplot(aes(x = price, fill = cut), data = diamonds)+  
 geom\_histogram()+  
 facet\_wrap( ~ color, ncol = 3)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Create a scatterplot of diamond price vs. table and color the points by the cut of the diamond.

ggplot(aes(x = price, y = table), data = diamonds)+  
 geom\_point(aes(color = cut))+  
 scale\_y\_continuous(breaks = seq(40, 95, 5))



# 没解决的问题

What is the typical table range for the majority of diamonds of ideal cut? What is the typical table range for the majority of diamonds of remium cut? to Use the graph that you created from the previous exercise to see the answer. You do not need to run summaries.

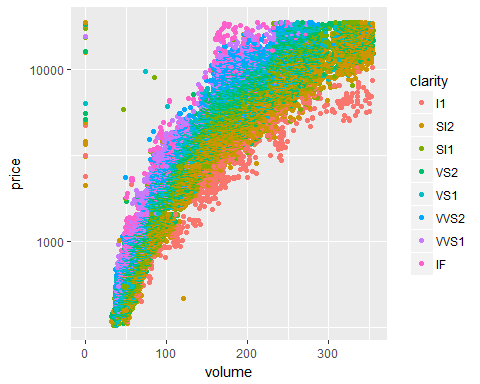
summary(  
 diamonds$table)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 43.00 56.00 57.00 57.46 59.00 95.00

### Create a scatterplot of diamond price vs. volume (x \* y \* z) and color the points by the clarity of diamonds. Use scale on the y-axis to take the log10 of price. You should also omit the top 1% of diamond volumes from the plot.

diamonds$volume <- diamonds$x \* diamonds$y \* diamonds$z  
ggplot(aes(x = volume, y = price), data = diamonds)+  
 geom\_point(aes(color = clarity))+  
 scale\_y\_log10()+  
 xlim(0, quantile(diamonds$volume, 0.99))

## Warning: Removed 540 rows containing missing values (geom\_point).



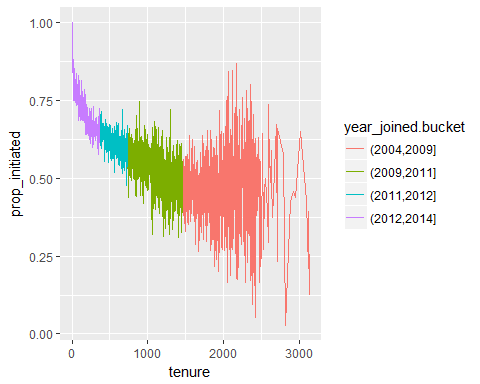
### Your task is to create a new variable called ‘prop\_initiated’ in the Pseudo-Facebook data set. The variable should contain the proportion of friendships that the user initiated.

pf <- read.csv('pseudo\_facebook.tsv', sep = '\t')  
pf$prop\_initiated <- pf$friendships\_initiated / pf$friend\_count

### Create a line graph of the median proportion of friendships initiated (‘prop\_initiated’) vs. tenure and color the line segment by year\_joined.bucket.

pf$year\_joined <- floor(2014 - pf$tenure/365)  
pf$year\_joined.bucket <- cut(pf$year\_joined, c(2004, 2009, 2011, 2012, 2014))  
ggplot(aes(x = tenure, y = prop\_initiated),   
 data = subset(pf, !is.na(year\_joined.bucket)))+  
 geom\_line(aes(color = year\_joined.bucket),   
 stat = 'summary',  
 fun.y = median)

## Warning: Removed 1962 rows containing non-finite values (stat\_summary).

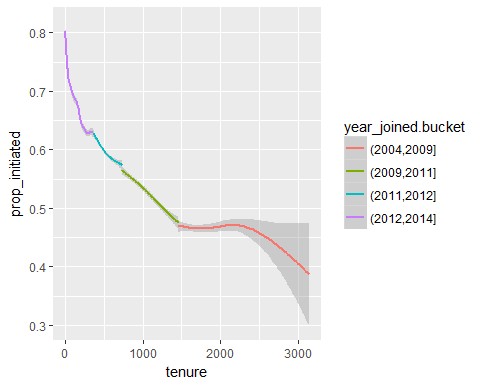


### Smooth the last plot you created of prop\_initiated vs tenure colored by year\_joined.bucket. You can bin together ranges of tenure or add a smoother to the plot.

ggplot(aes(x = tenure, y = prop\_initiated),  
 data = subset(pf, !is.na(year\_joined.bucket)))+  
 geom\_smooth(aes(color = year\_joined.bucket))

## `geom\_smooth()` using method = 'gam'

## Warning: Removed 1962 rows containing non-finite values (stat\_smooth).

 ###On average, which group initiated the greatest proportion of its Facebook friendships? The plot with the smoother that you created in the last exercise can help you answer this question. My answer: People who joined after 2012

### For the group with the largest proportion of friendships initiated, what is the group’s average (mean) proportion of friendships initiated?

My answer: 0.6654

with(subset(pf, year\_joined > 2012), summary(prop\_initiated))

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.5115 0.7018 0.6654 0.8490 1.0000 1468

### Why do you think this group’s proportion of friendships initiated is higher than the others?

I think it is because they are in the era of Internet.

### Create a scatter plot of the price/carat ratio of diamonds. The variable x should be assigned to cut. The points should be colored by diamond color, and the plot should be faceted by clarity.

ggplot(aes(x = cut, y = price/carat), data = diamonds)+  
 geom\_point(aes(color = color))+  
 facet\_wrap( ~ clarity, ncol = 2)

