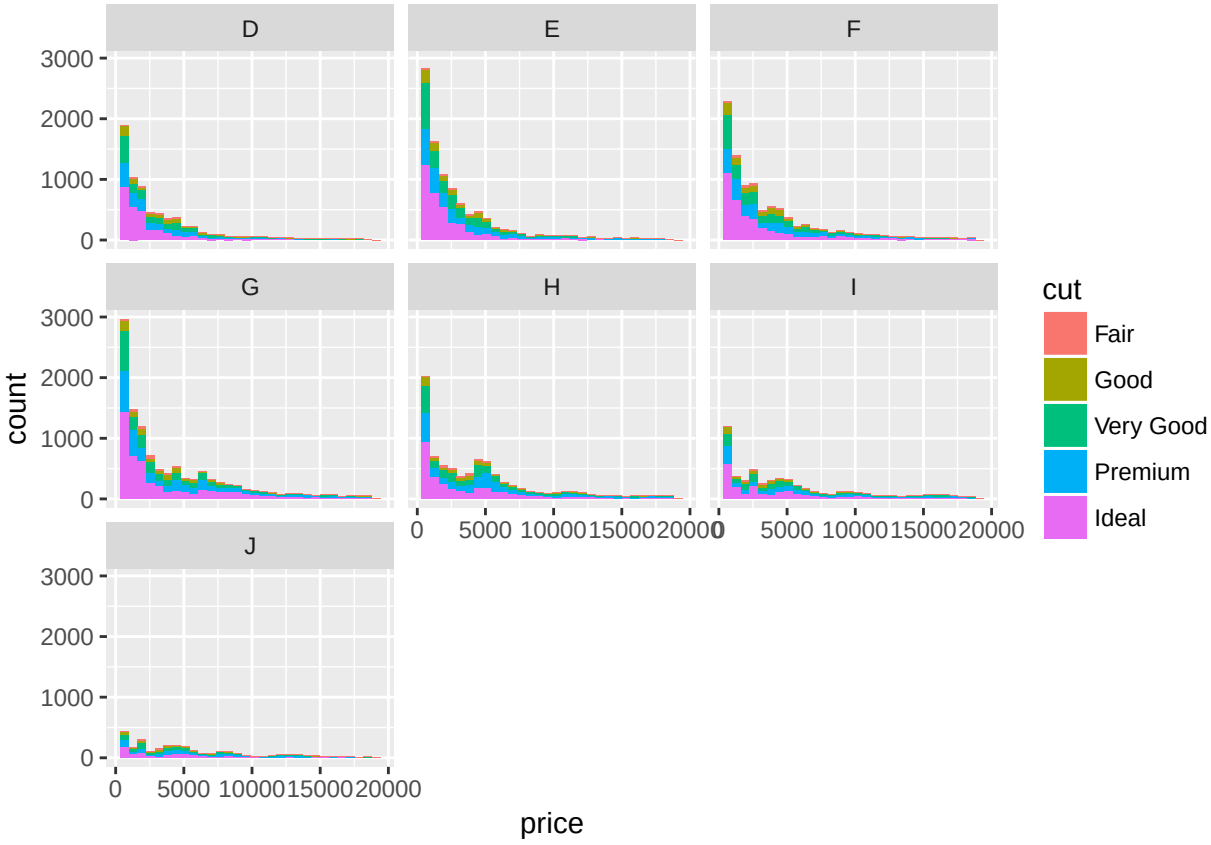
EDA: Using R with Multiple Variables

Show/Hide Nav Price Histograms with Facet and Color

Create a histogram of diamond prices. Facet the histogram by diamond color, and use cut to color the histogram bars.

```
library(ggplot2)
library(dplyr)
```

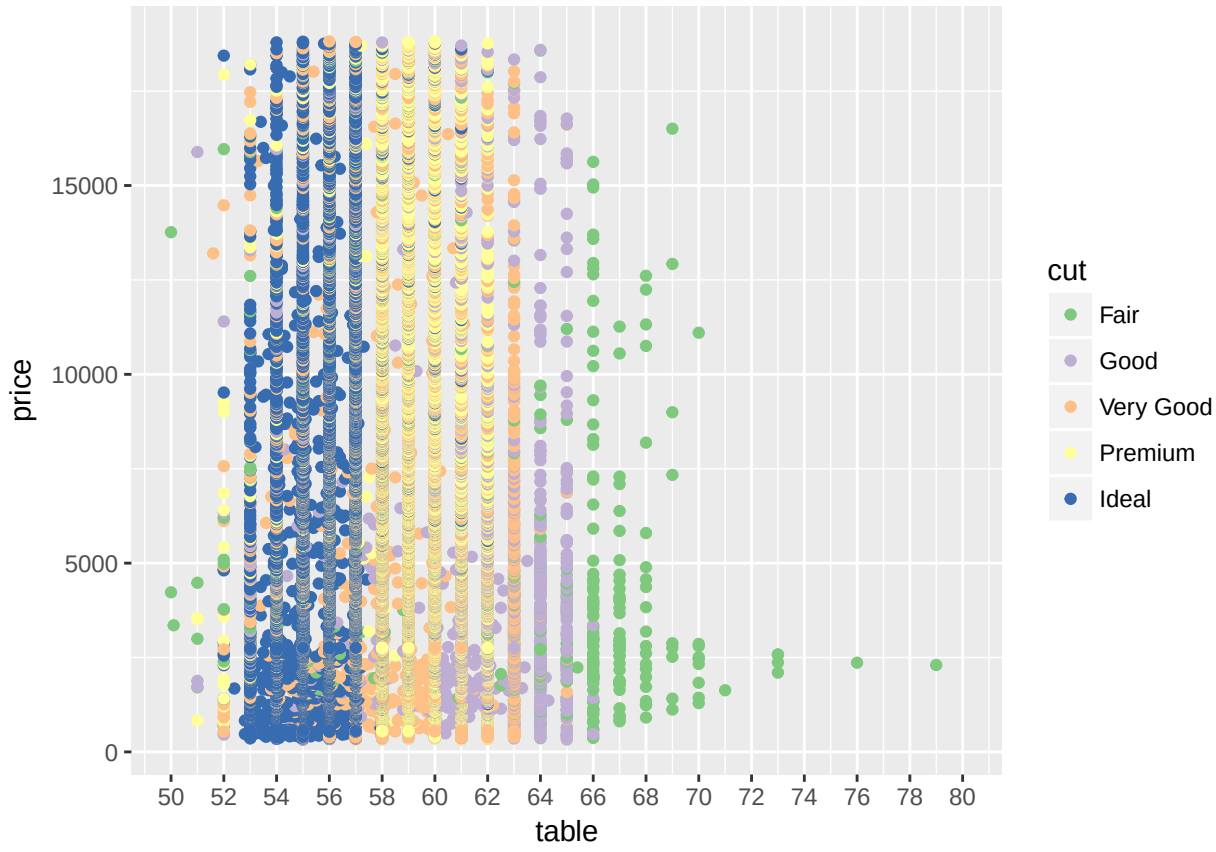
```
diamonds <- diamonds
ggplot(data = diamonds,
       aes(x=price, fill=cut)) +
  facet_wrap(~color) +
  geom_histogram()
```



Price vs. Table Colored by Cut

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

```
ggplot(data = diamonds,
       aes(x=table, y=price, col=cut )) +
  geom_point() +
  scale_x_continuous(limits = c(50,80), breaks = seq(50,80,2)) +
  scale_color_brewer(type = 'qual')
```



Typical Table Value

What is the typical table range for the majority of diamonds of ideal cut?

53 - 57

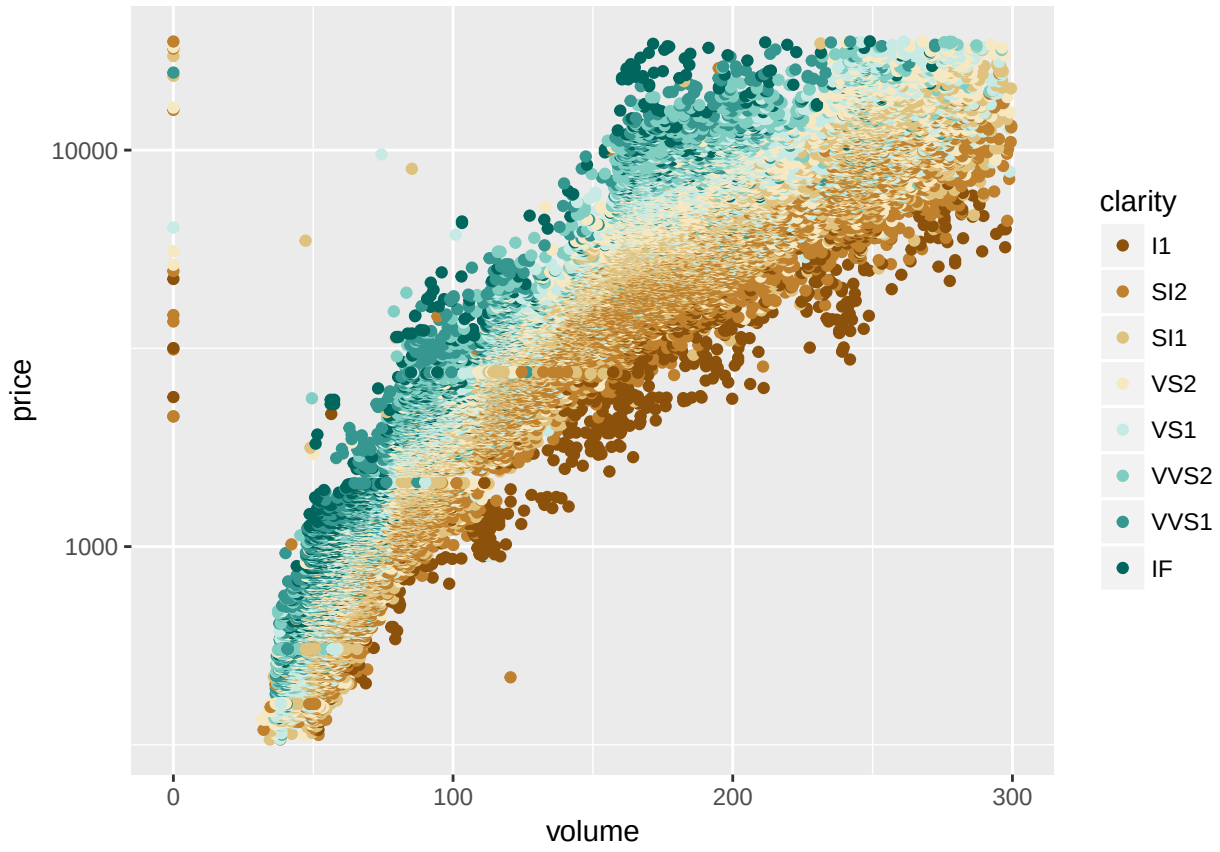
What is the typical table range for the majority of diamonds of premium cut?

58 - 62

Price vs. Volume and Diamond Clarity

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(data = diamonds,
       aes(x=volume, y=price, col=clarity)) +
  geom_point() +
  scale_y_log10() +
  scale_x_continuous(limits = c(0, 300)) +
  scale_color_brewer(type = 'div')
```



Proportion of Friendships Initiated

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')
head(pf)
```

```
##    userid age dob_day dob_year dob_month gender tenure friend_count
## 1 2094382  14      19     1999        11   male    266            0
## 2 1192601  14       2     1999        11 female      6            0
## 3 2083884  14      16     1999        11   male     13            0
## 4 1203168  14      25     1999        12 female     93            0
## 5 1733186  14       4     1999        12   male     82            0
## 6 1524765  14       1     1999        12   male     15            0
##   friendships_initiated likes likes_received mobile_likes
## 1                     0     0              0            0
## 2                     0     0              0            0
## 3                     0     0              0            0
## 4                     0     0              0            0
## 5                     0     0              0            0
## 6                     0     0              0            0
##   mobile_likes_received www_likes www_likes_received
## 1                     0         0                  0
## 2                     0         0                  0
## 3                     0         0                  0
```

```
## 4                     0         0                 0
## 5                     0         0                 0
## 6                     0         0                 0
```

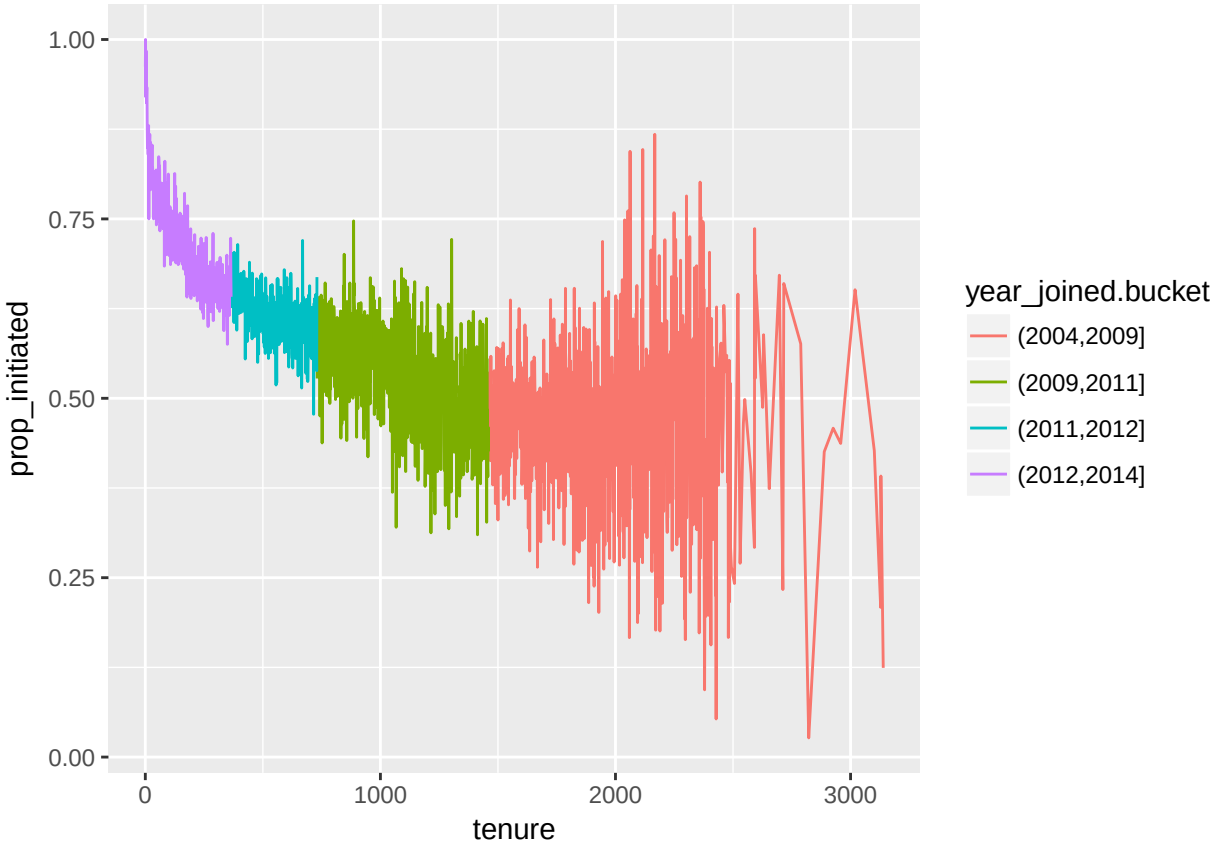
```
pf$prop_initiated <- pf$friendships_initiated/pf$friend_count
```

prop_initiated vs. tenure

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))
na.omit(pf)
```

```
ggplot(data = subset(pf, pf$prop_initiated!="NaN"),
       aes(x = tenure, y = prop_initiated)) +
  geom_line(stat = 'summary', fun.y = median, aes(color = year_joined.bucket))
```

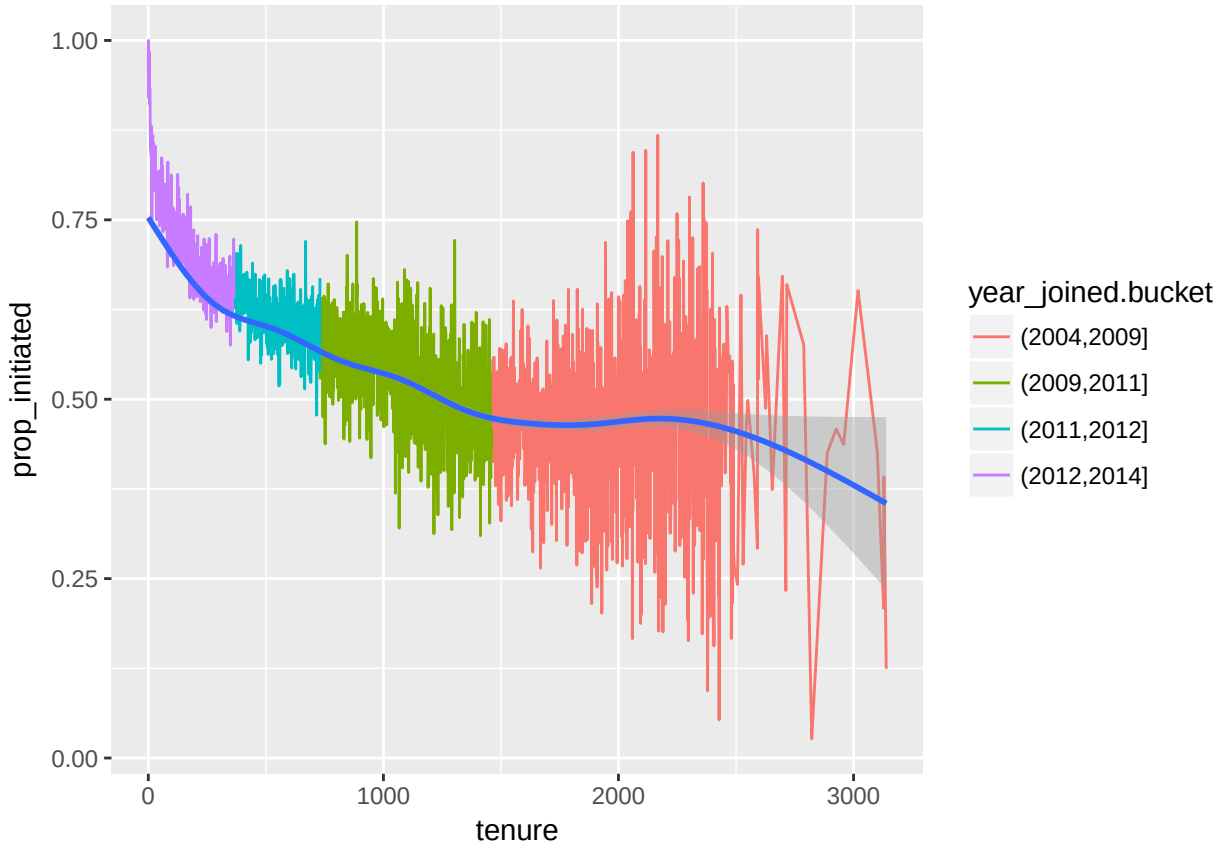


Smoothing prop_initiated vs. tenure

Smooth the last plot you created of 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(data = subset(pf, pf$prop_initiated!="NaN"),
       aes(x = tenure, y = prop_initiated)) +
```

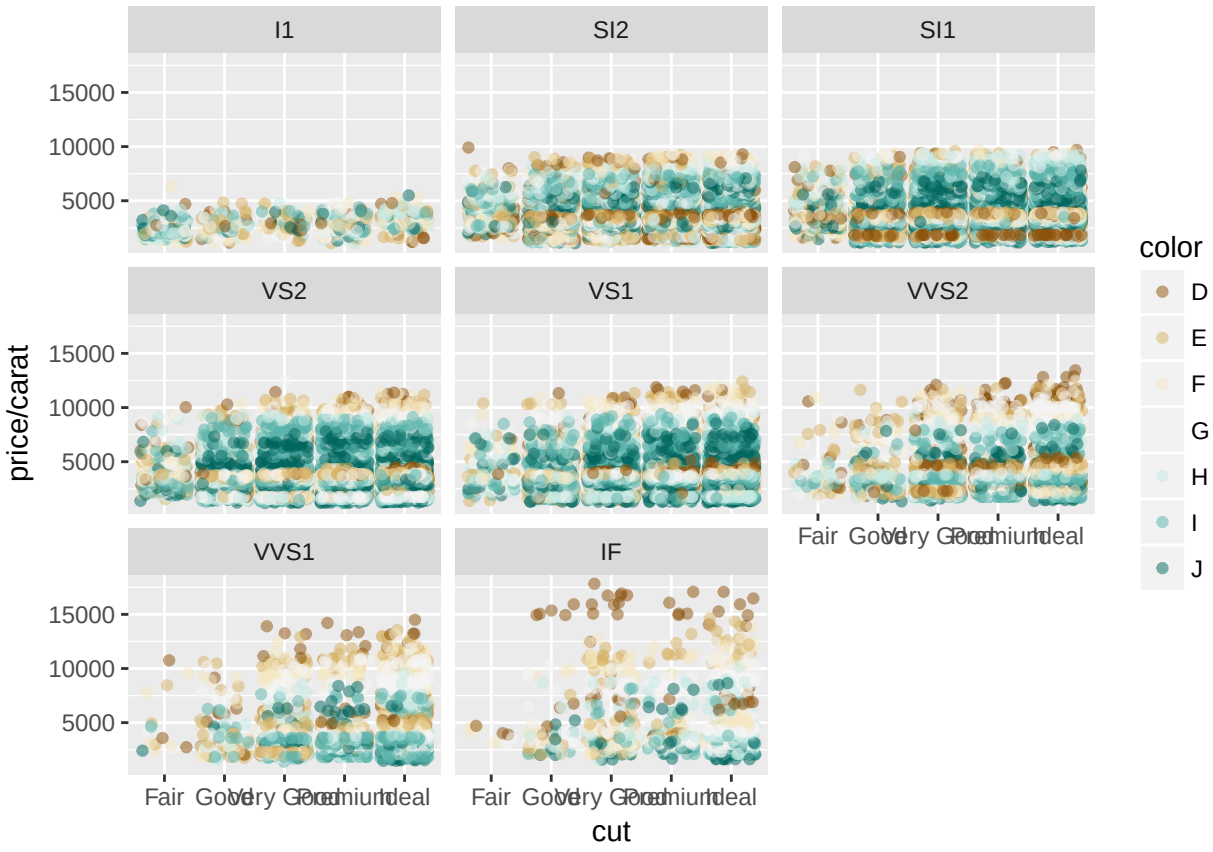
```
  geom_line(stat = 'summary', fun.y = median, aes(color = year_joined.bucket)) +
  geom_smooth()
```



Price/Carat Binned, Faceted, & Color

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(data = diamonds,
       aes(x=cut, y=price/carat)) +
  geom_jitter(alpha=1/2, aes(col=color)) +
  facet_wrap(~clarity) +
  scale_color_brewer(type = 'div')
```


I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
price/carat
15000
10000
5000
Fair
Good
Very Good
Premium
Ideal
cut
color
D
E
F
G
H
I
J