UDA651ProblemSet5

Mallesh

December 28, 2015

Load the diamonds dataset and required libraries.

```
require(dplyr)
```

```
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

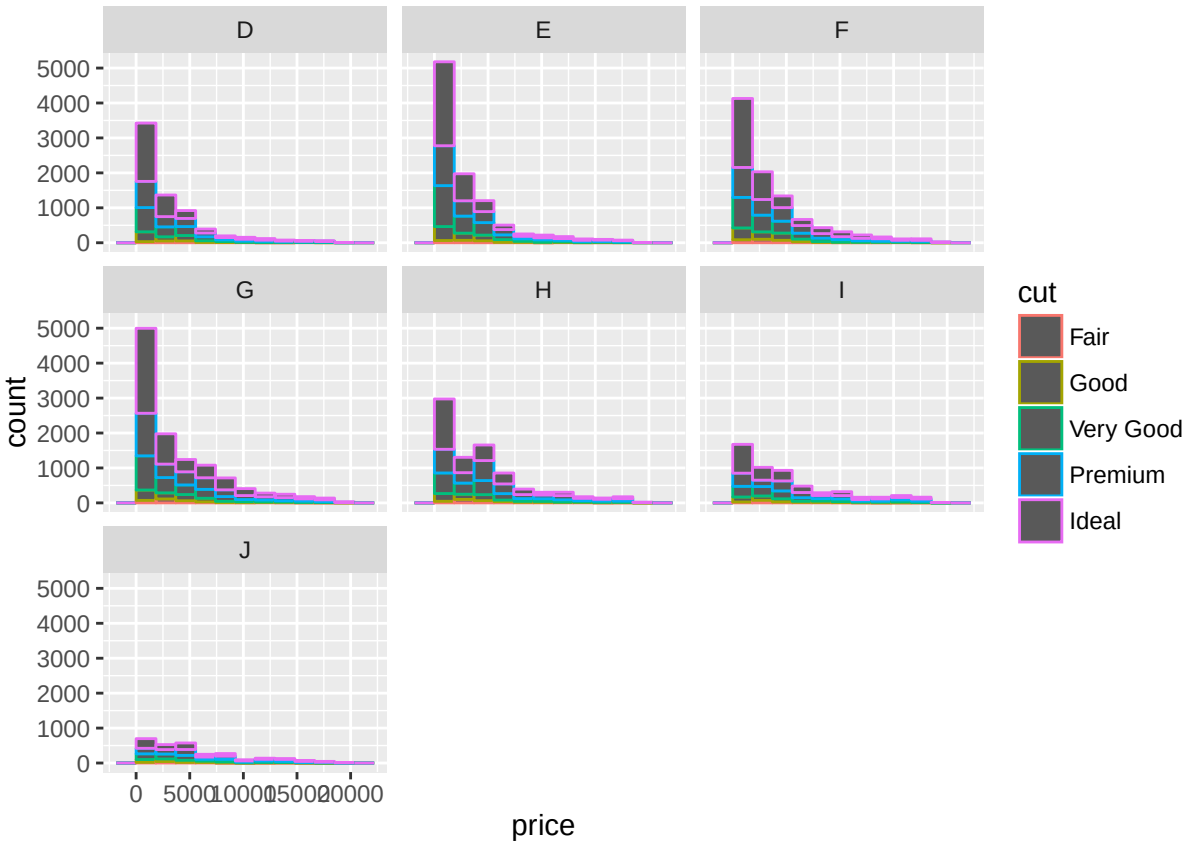
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
data("diamonds")
```

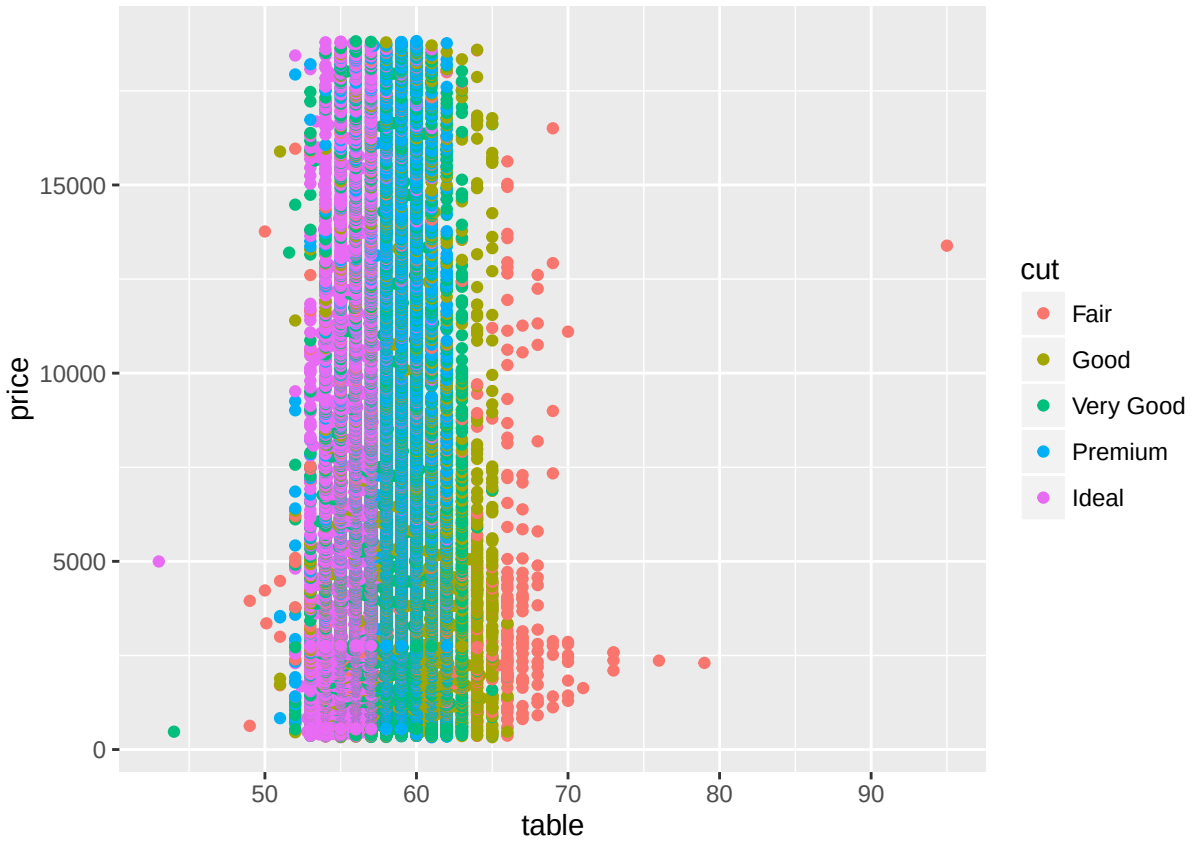
Create a histogram of diamond prices, Facet the histogram by diamond color and use cut to color the histogram bars.

```
qplot(data=diamonds, price, color=cut, bins=10) + facet_wrap(~color) +
  scale_fill_brewer(type = 'qual')
```



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)) + geom_point(aes(color=cut))
```



What is the typical table range for cut IDEAL.

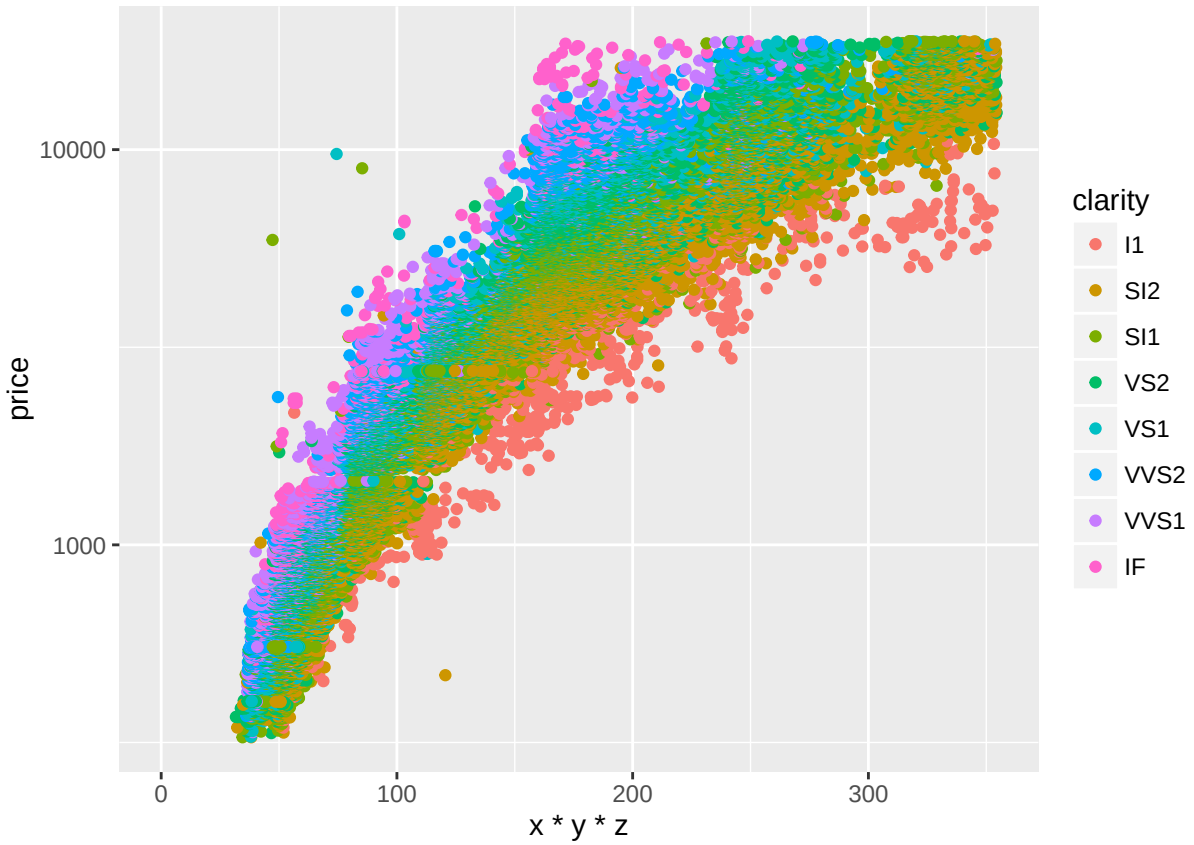
```
range(filter(diamonds, cut=="Ideal")$table)
```

```
## [1] 43 63
```

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.

```
ggplot(data=filter(diamonds, x*y*z >0), aes(x=x * y * z, y=price)) +
  xlim(0, quantile(diamonds$x * diamonds$y * diamonds$z, 0.99)) +
  geom_point(aes(color=clarity)) +
  scale_y_log10()
```

```
## Warning: Removed 540 rows containing missing values (geom_point).
```



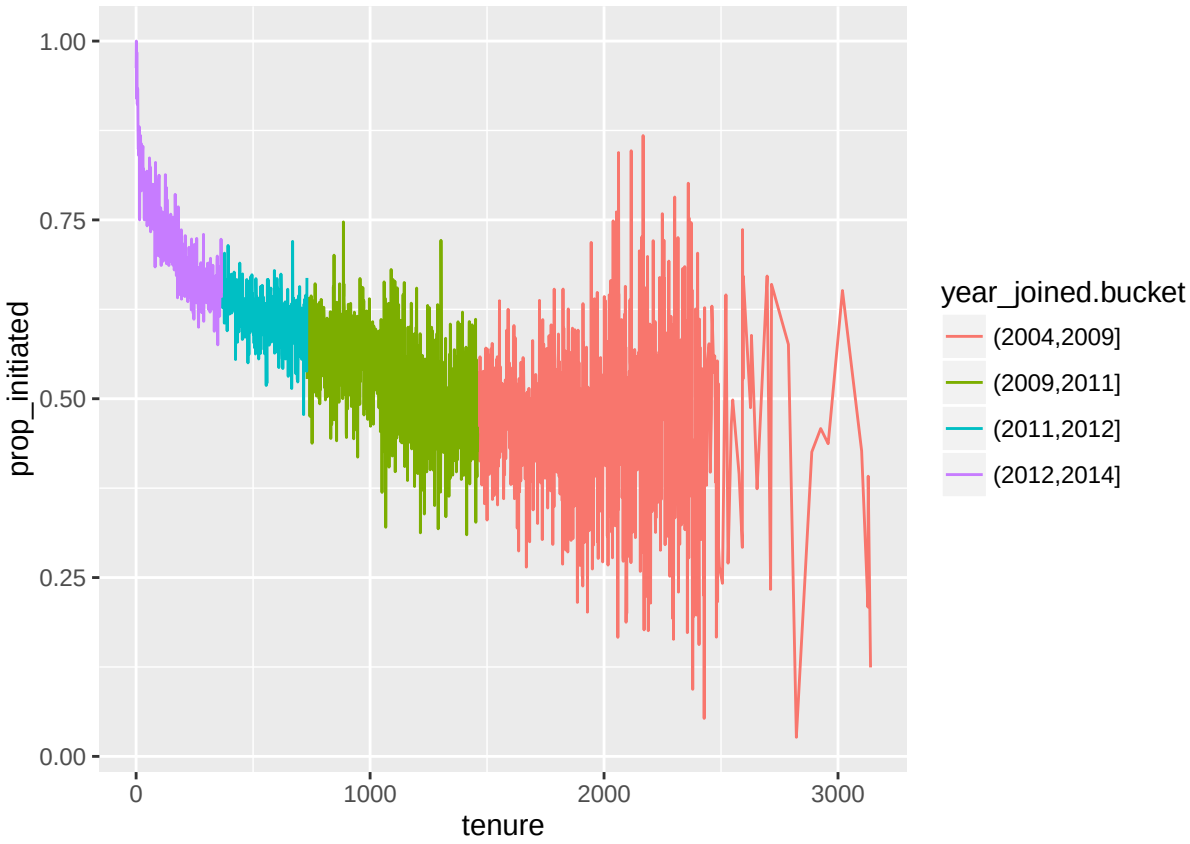
create a new variable called 'prop_initiated' in the Pseudo-Facebook data set. The variable should contain the proportion of friendships that the user initiated.

```
pf <- read.csv("pseudo_facebook.tsv", sep="\t")

pf <- pf %>%
  mutate(prop_initiated = ifelse(friend_count > 0 ,friendships_initiated/friend_count,NA))

#pf$prop_initiated <- ifelse(is.finite(pf$prop_initiated),pf$prop_initiated,0)
```

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 <- pf %>% mutate(year_joined = floor(2014 - (tenure/365) ))

pf$year_joined.bucket <- cut(pf$year_joined,
                             breaks=c(2004, 2009,2011, 2012, 2014))

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

```
## Warning: Removed 1964 rows containing non-finite values (stat_summary).
```

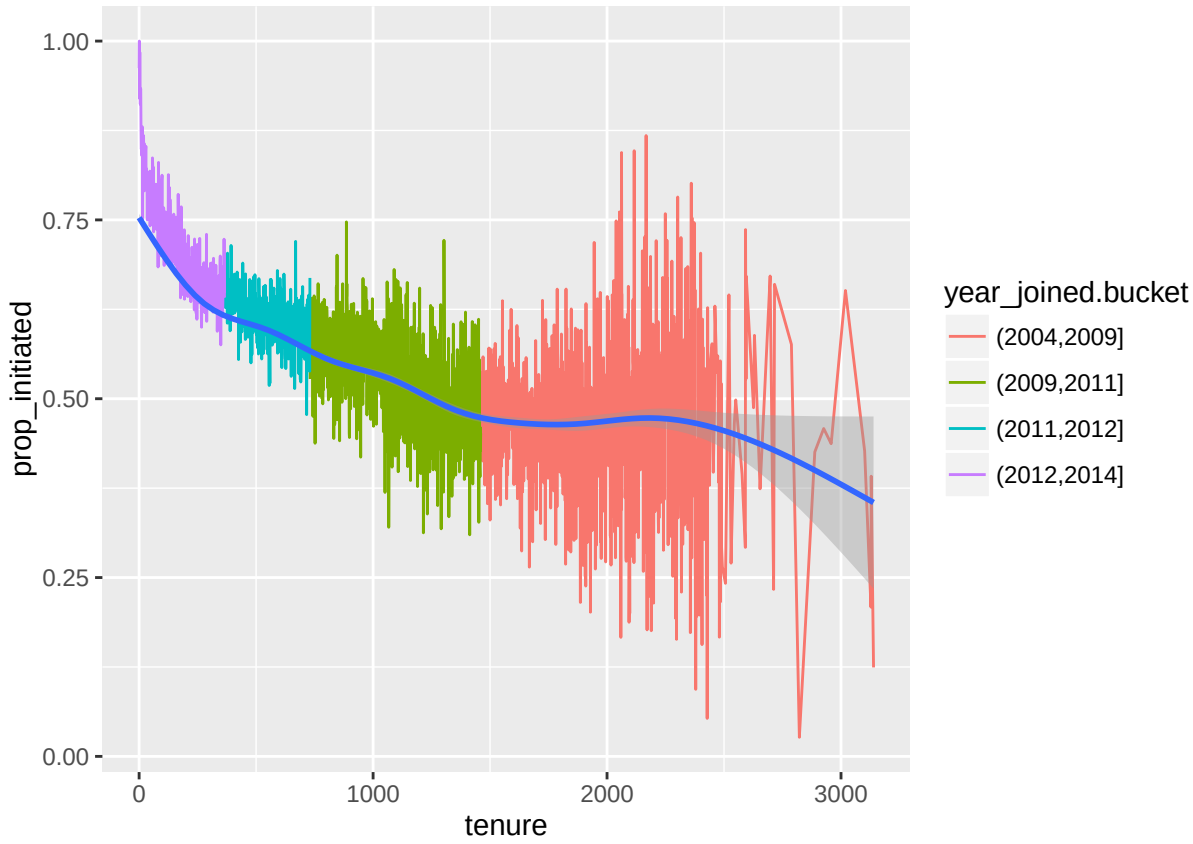


Smooth the last plot you created of of prop_initiated vs tenure colored by year_joined.bucket. You can bin together ranges of tenure or add a smoother to the plot.

```
ggplot(data=pf, aes(x=tenure, y=prop_initiated)) +
  geom_line(aes(color=year_joined.bucket), stat="summary", fun.y=median) +
  #geom_line(aes(color=year_joined.bucket), stat="summary", fun.y=mean)
  geom_smooth()
```

```
## Warning: Removed 1964 rows containing non-finite values (stat_summary).
```

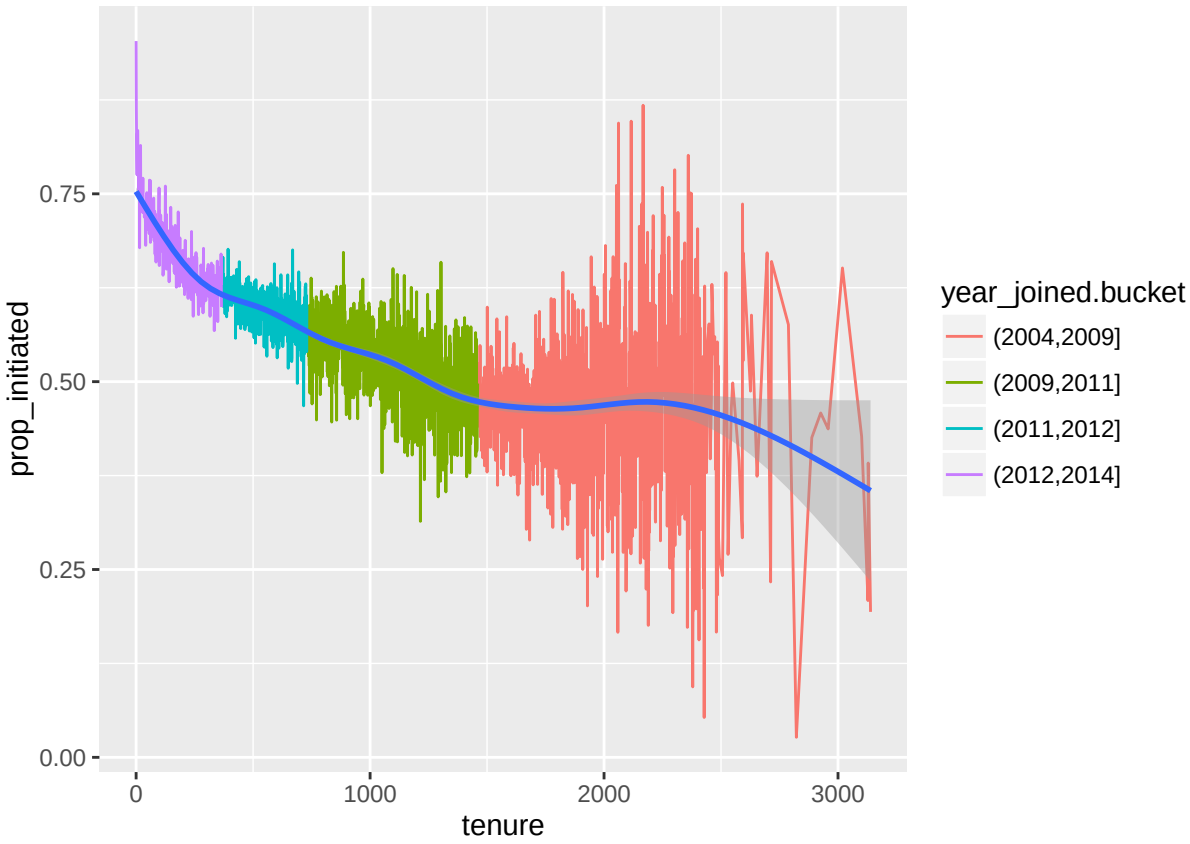
```
## Warning: Removed 1964 rows containing non-finite values (stat_smooth).
```




```
ggplot(data=pf, aes(x=tenure, y=prop_initiated)) +
  #geom_line(aes(color=year_joined.bucket), stat="summary", fun.y=median) +
  geom_line(aes(color=year_joined.bucket), stat="summary", fun.y=mean)  +
  geom_smooth()
```

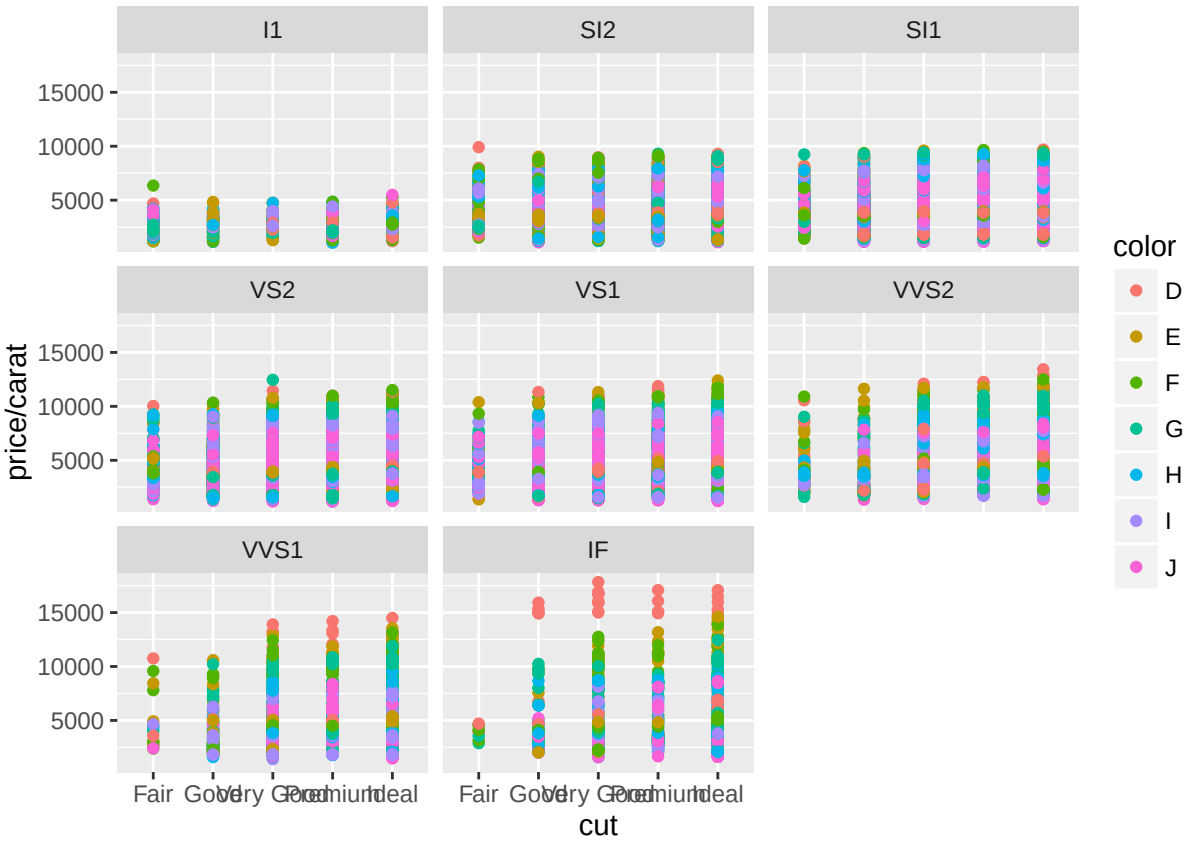
```
## Warning: Removed 1964 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 1964 rows containing non-finite values (stat_smooth).
```



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_point(aes(color=color)) + facet_wrap(~clarity)
```



Below is the final exercise where a dataset from Gapminder is downloaded.

Save the Unemployment data set for 15+ year olds for several countries for last 25+ years. The data is avaiable in Excel which was converted into CSV for the ease of reading into R.

```
require(tidyr)
```

```
## Loading required package: tidyr
```

```
unm <- read.csv("unemployment_15.csv")

head(unm)
```

```
##   Total.15..unemployment.... X1981 X1982 X1983 X1984 X1985 X1986 X1987
## 1                  Australia    NA    NA    NA    NA    NA   8.1   8.0
## 2                     Canada   7.6  11.0  11.9  11.3  10.6   9.6   8.8
## 3                 Czech Rep.    NA    NA    NA    NA    NA    NA    NA
## 4                    Estonia    NA    NA    NA    NA    NA    NA    NA
## 5                    Finland   4.8   5.3   5.4   5.0   4.9   5.2   5.0
## 6                     France   7.4   8.1   8.4   9.8  10.2  10.4  10.5
##   X1988 X1989 X1990 X1991 X1992 X1993 X1994 X1995 X1996 X1997 X1998 X1999
## 1   7.2   6.1   6.9   9.5  10.4  10.5   9.4   8.2   8.2   8.2   7.7   6.9
## 2   7.7   7.5   8.1  10.3  11.1  11.3  10.3   9.4   9.6   9.1   8.3   7.6
## 3    NA    NA    NA    NA    NA   4.3   4.3   4.0   3.9   4.8   6.4   8.7
## 4    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 5   4.4   3.1   3.1   6.5  11.6  16.1  16.4  15.2  14.4  12.6  11.3  10.2
## 6  10.0   9.4   8.9   9.4  10.2  11.5  12.1  11.4  12.0  12.1  11.5  10.8
```

```
##   X2000 X2001 X2002 X2003 X2004 X2005  X
## 1   6.2   6.7   6.4   6.0   5.5   5.1 NA
## 2   6.8   7.2   7.6   7.6   7.2   6.7 NA
## 3   8.8   8.2   7.3   7.8   8.3   8.3 NA
## 4  13.7  12.5  10.2  10.0   9.5   7.9 NA
## 5   9.7   9.1   9.0   9.0   8.8   8.3 NA
## 6   9.5   8.7   9.0   9.8  10.0   9.9 NA
```

```
names(unm) <- c("country",1981:2005,"dummy")
```

reshape the data, convert the data from columns to rows. Filter the NAs after the conversion.

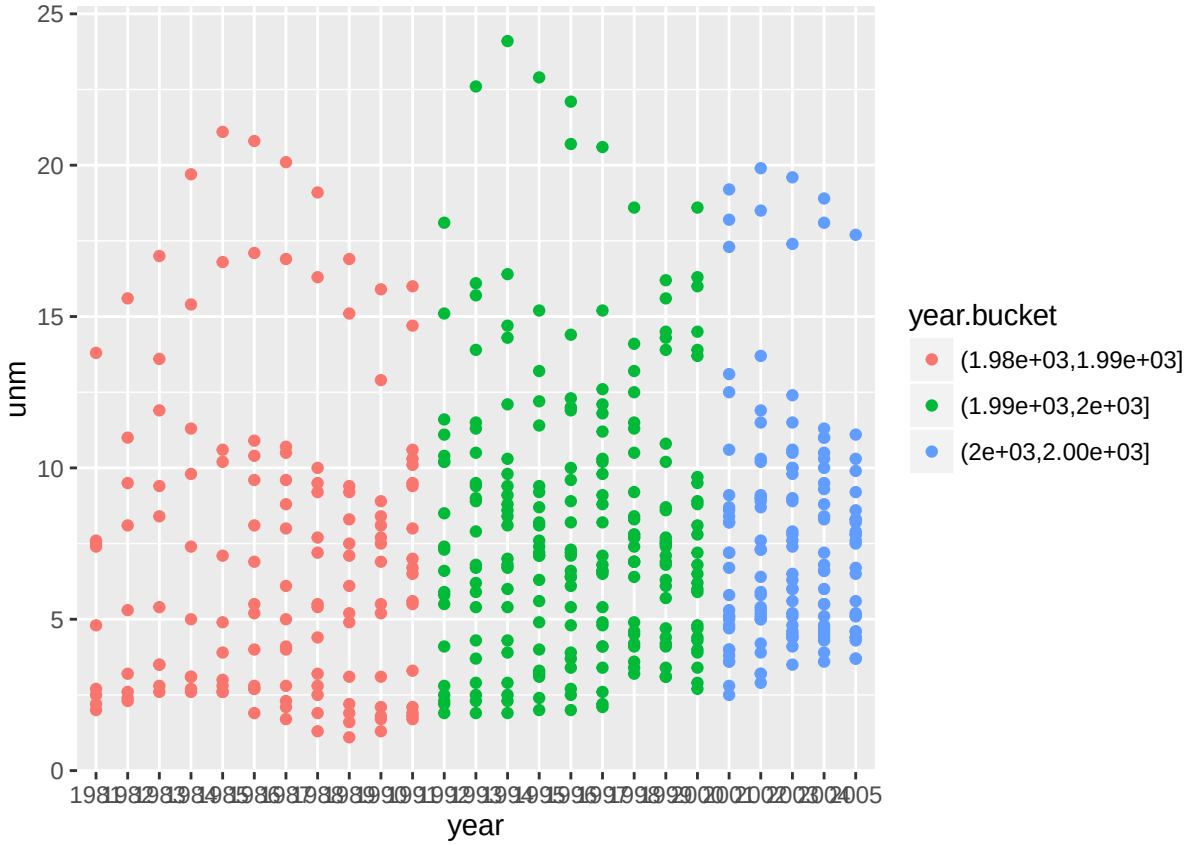
```
unmg <- gather(unm, 'year',"unm",2:27)

unmg <- filter(unmg, !is.na(unm))
```

create bins for each decade:

```
unmg <- unmg %>%
  mutate(year_num = as.integer(year) + 1980)


unmg$year.bucket <- cut(unmg$year_num,
                            breaks=c(1980,1991,2000,2005), right=T)
```

plot unemplyment mean across the years:

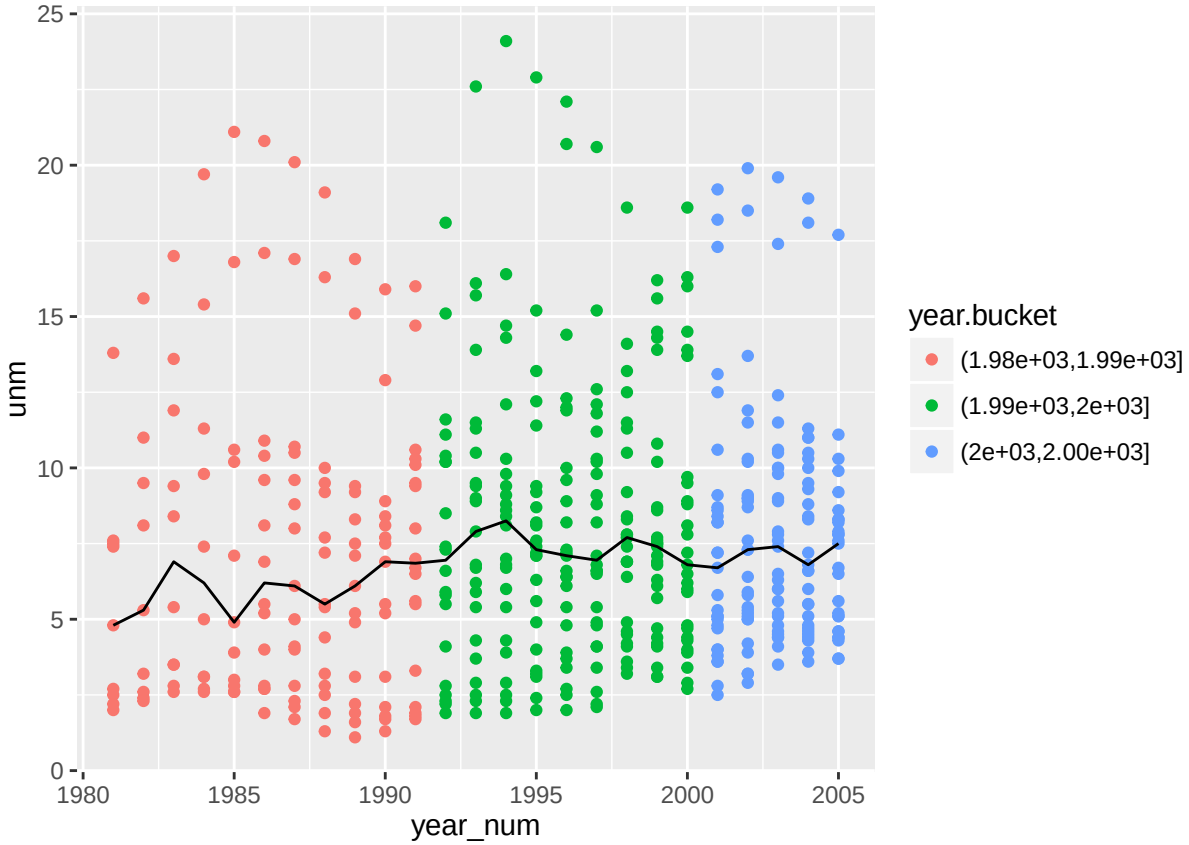
```
ggplot(unmg, aes(x=year, y=unm)) + geom_point(aes(color=year.bucket))
```



Scatter plot doesn't really show any patterns except that the minimum employment rate increased across the decades.

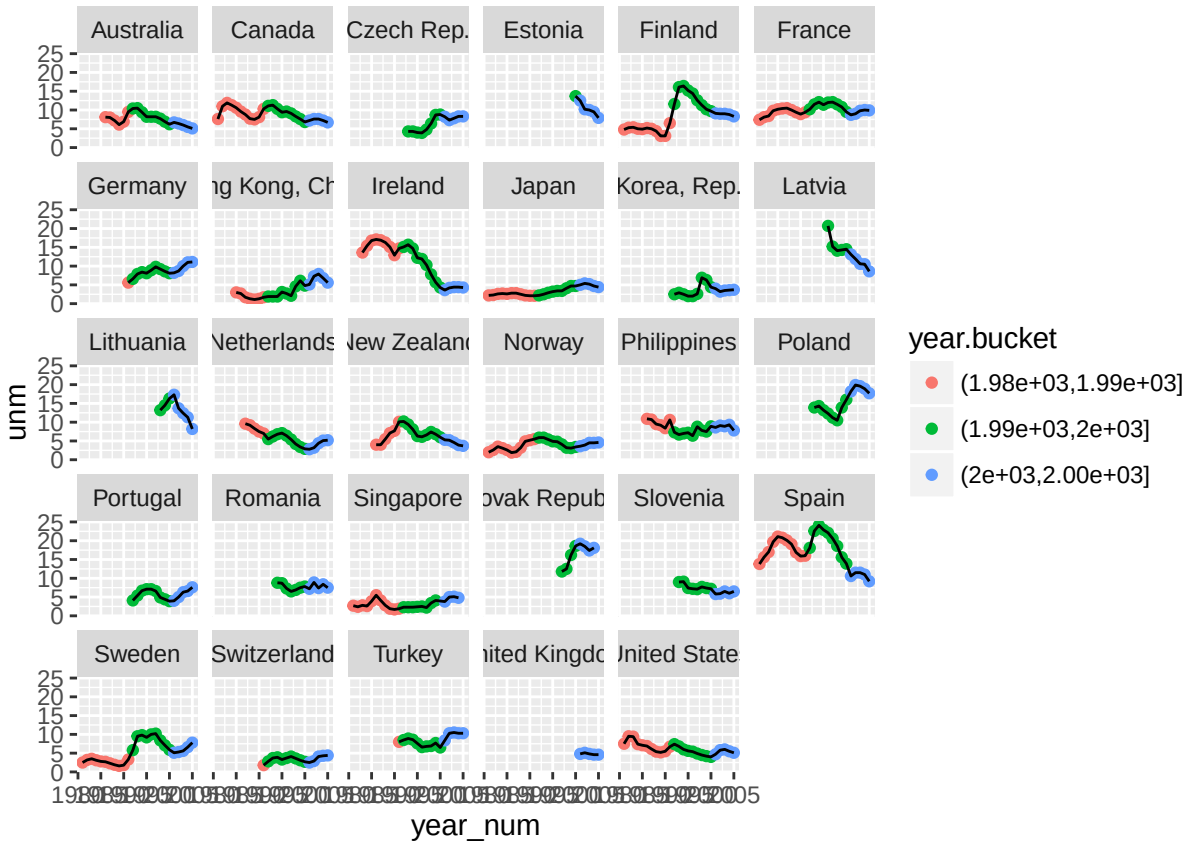
Let us confirm our observation by pointing the median unemplyment rate.

```
ggplot(unmg, aes(x=year_num, y=unm)) + geom_point(aes(color=year.bucket)) +

  geom_line(stat="summary", fun.y=median)
```



lets add facet wrap for different countries:

```
ggplot(unmg, aes(x=year_num, y=unm)) + geom_point(aes(color=year.bucket)) +

  geom_line(stat="summary", fun.y=median) + facet_wrap(~country)
```



This graph indicates trend across countries. Advanced countries like Australia, Canada, United states show a downward trend, while some of the countries like China indicates an upward trend in unemplyment rate.