Data Analysis with R

Problem Set 3

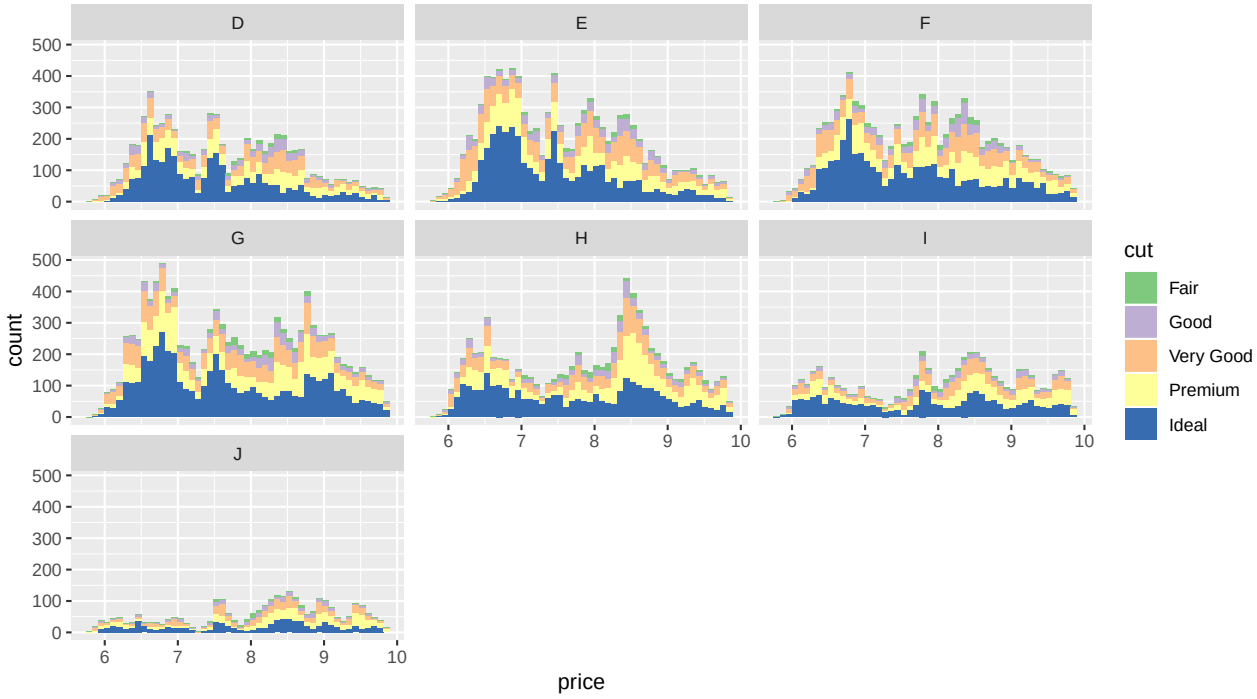
Pramod Duvvuri

4/11/2019

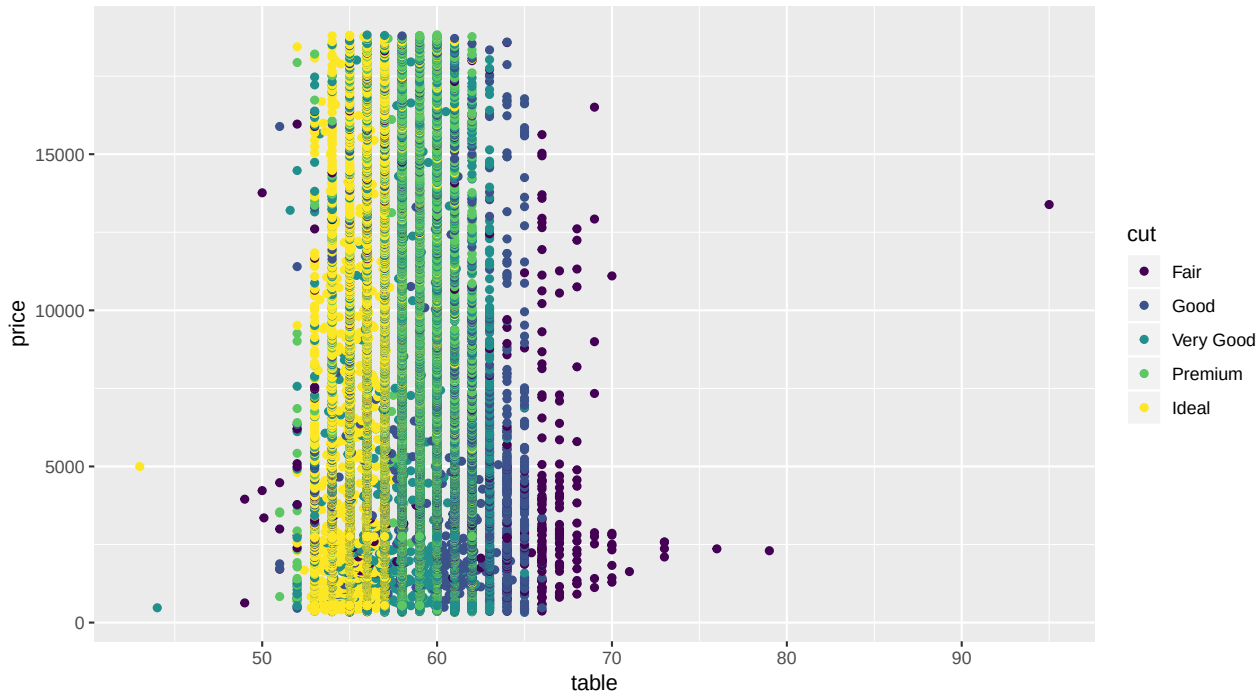
```
library(tidyverse)
data("diamonds")
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 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 0.290 Premium   I     VS2      62.4    58   334  4.2   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
```

```
ggplot(data = diamonds, mapping = aes(x = log(price), fill = cut)) +
  geom_histogram(bins = 50) +  facet_wrap(~color,nrow = 3) +
  scale_fill_brewer(type = 'qual') + xlab('price')
```




```
ggplot(data = diamonds, mapping = aes(x = table, y = price)) +
  scale_fill_brewer(type = 'qual') +
  geom_point(mapping = aes(color = cut))
```




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

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

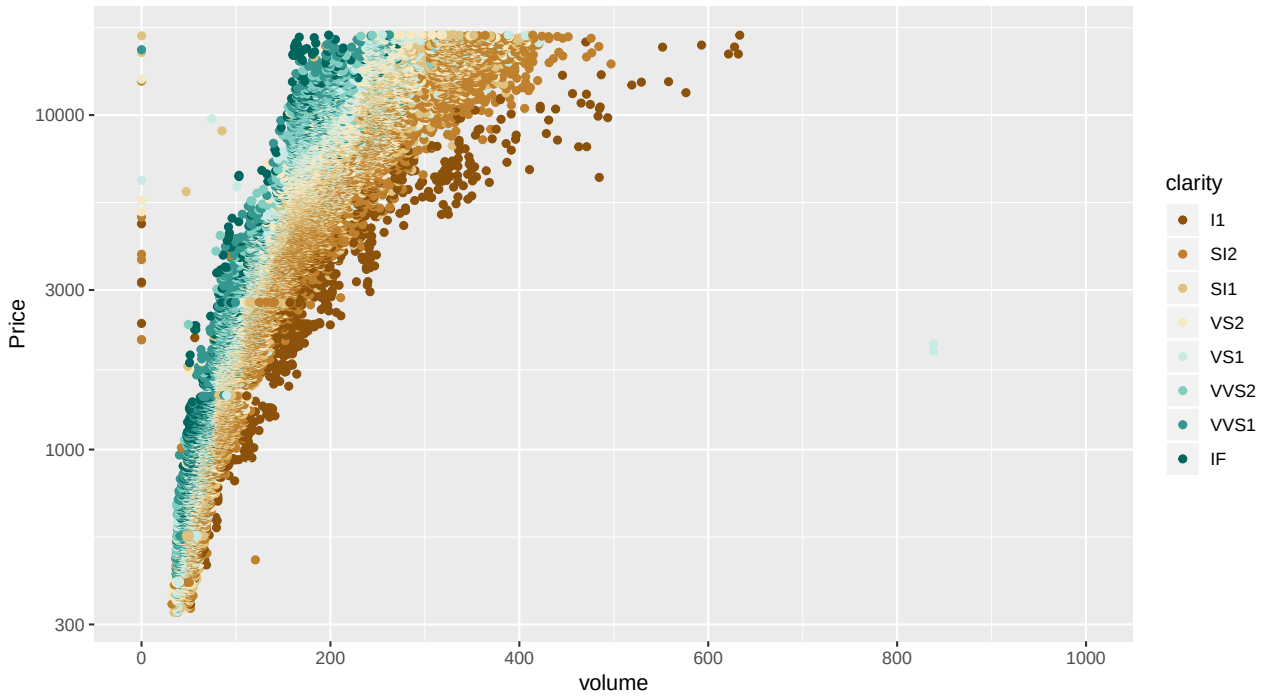
```
range(filter(diamonds, cut == 'Premium')$table)
```

```
## [1] 51 62
```

```
# Adding volume column
diamonds <- mutate(diamonds, volume = round(x * y * z,2))

#summary(diamonds$price)

ggplot(data = subset(diamonds, price < 17378.22) ,
       mapping = aes(y = price, x = volume)) +
  geom_point(mapping = aes(color = clarity)) +
  scale_color_brewer(type = 'div') + scale_y_log10() +
  scale_x_continuous(breaks = seq(0,1000,200), limits = c(0,1000)) +
  ylab('Price')
```

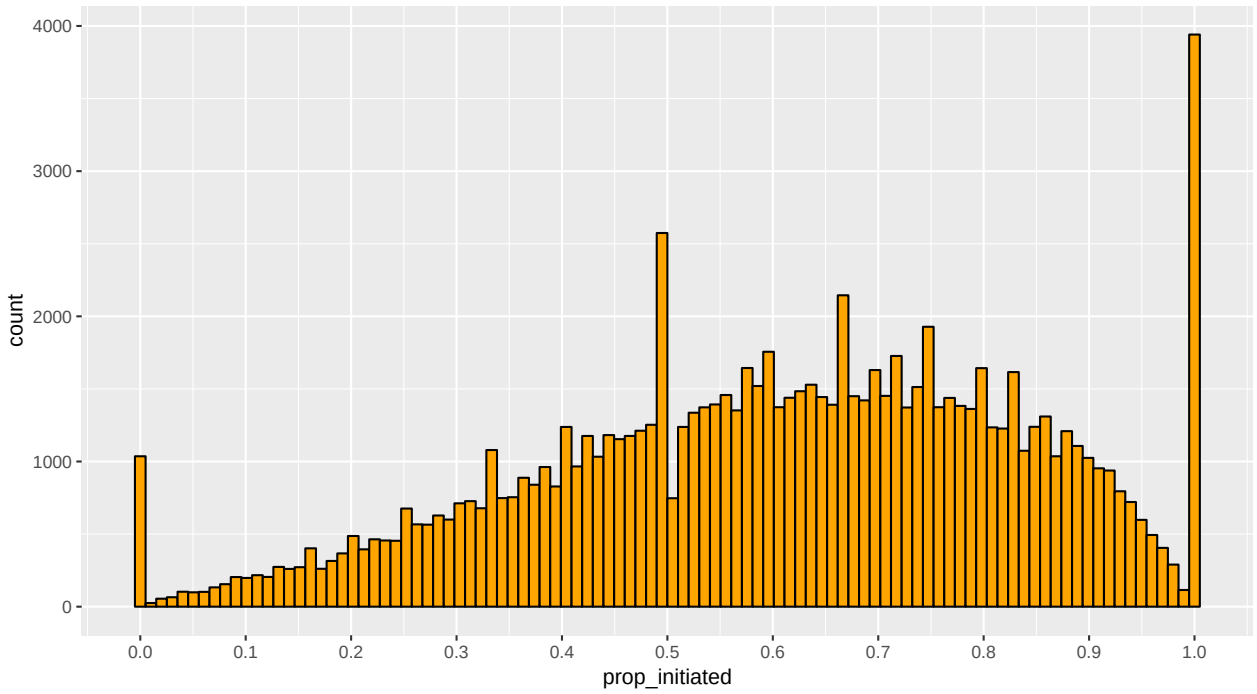



```
pf <- read.delim('pseudo_facebook.tsv')
pf <- na.omit(pf)
names(pf)
```

```
##  [1] "userid"                "age"
##  [3] "dob_day"               "dob_year"
##  [5] "dob_month"             "gender"
##  [7] "tenure"                "friend_count"
##  [9] "friendships_initiated" "likes"
## [11] "likes_received"        "mobile_likes"
## [13] "mobile_likes_received" "www_likes"
## [15] "www_likes_received"
```

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

ggplot(data = pf, mapping = aes(x = prop_initiated)) +
  geom_histogram(na.rm = TRUE, bins = 100, fill = 'orange', color = 'black') +
  scale_x_continuous(breaks = seq(0,1,0.1))
```




```
pf$year_joined <- floor(2014 - pf$tenure/365)
sort(table(pf$year_joined))
```

```
##
##  2005  2006  2014  2007  2008  2009  2010  2011  2012  2013
##     9    14    70   539  1432  4532  5436  9856 33366 43572
```

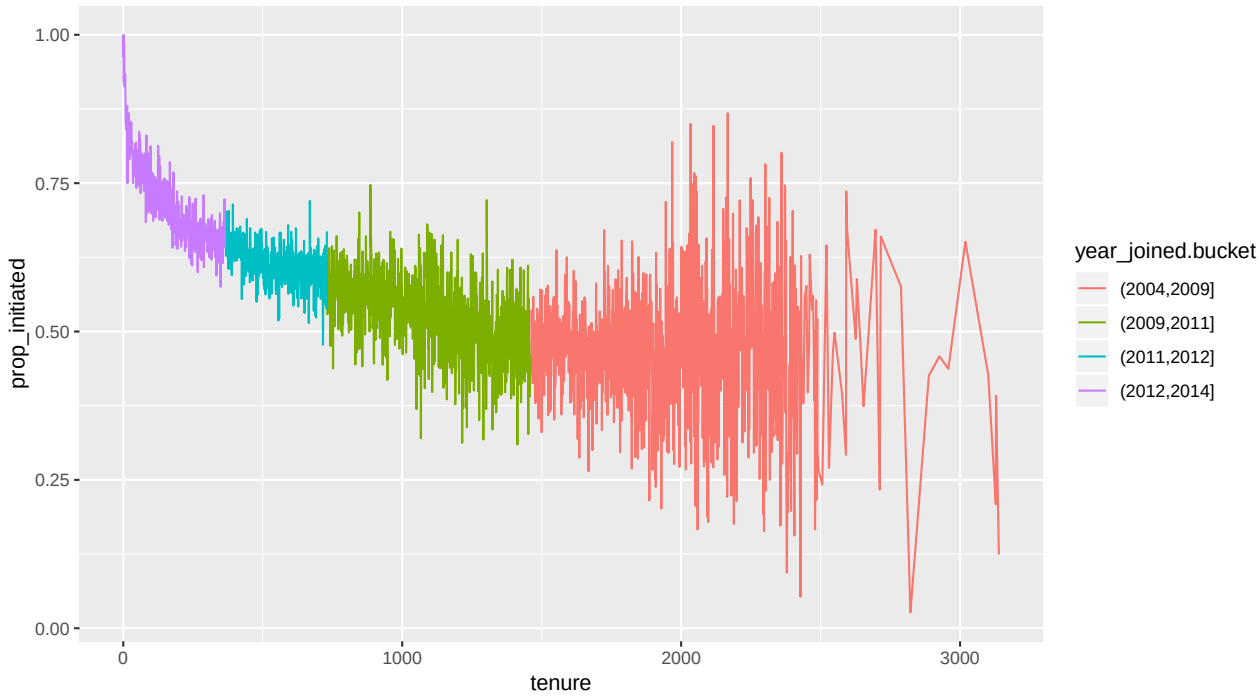
```
?cut

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

table(pf$year_joined.bucket)
```

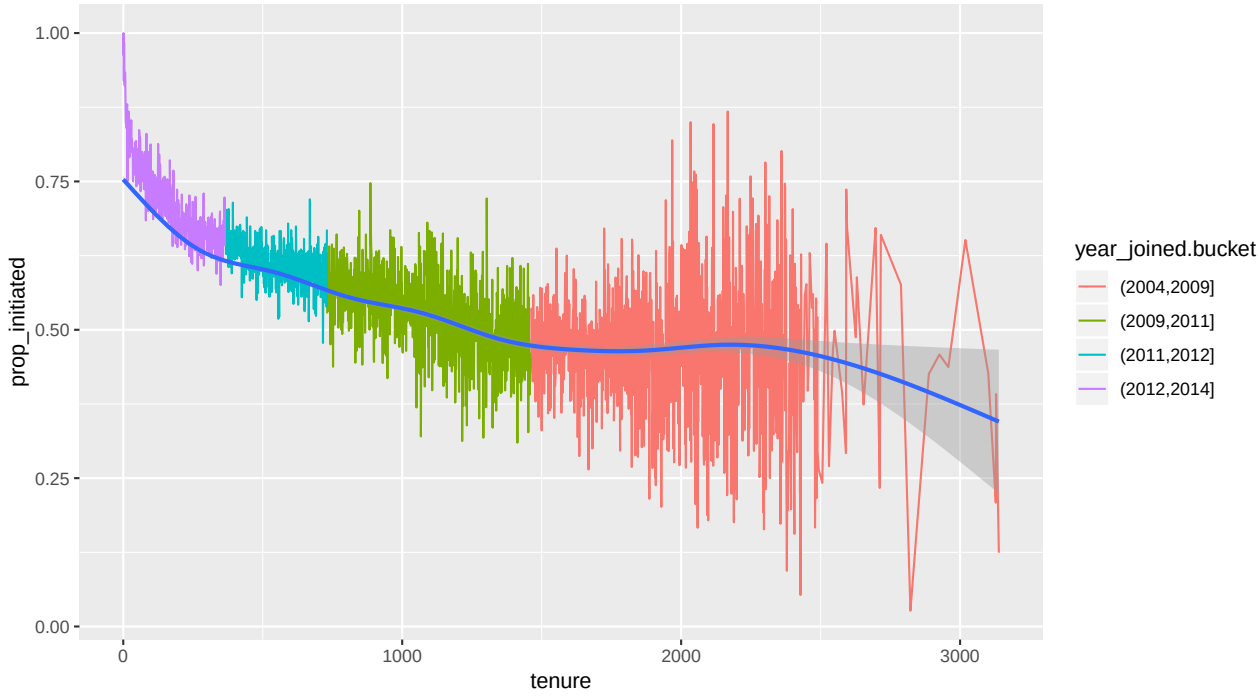
```
##
## (2004,2009] (2009,2011] (2011,2012] (2012,2014]
##        6526       15292       33366       43642
```

```
ggplot(data = na.omit(pf), mapping = aes(x = tenure, y = prop_initiated)) +
  geom_line(stat = 'summary', fun.y = median,
            mapping = aes(color = year_joined.bucket), na.rm = TRUE)
```




```
ggplot(data = na.omit(pf), mapping = aes(x = tenure, y = prop_initiated)) +
  geom_line(stat = 'summary', fun.y = median,
            mapping = aes(color = year_joined.bucket), na.rm = TRUE) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
ggplot(data = diamonds, mapping = aes(x = cut, y = price/carat)) +
  geom_jitter(mapping = aes(color = color)) +
  scale_color_brewer(type = 'div') +
```

```
facet_wrap(~clarity)
```