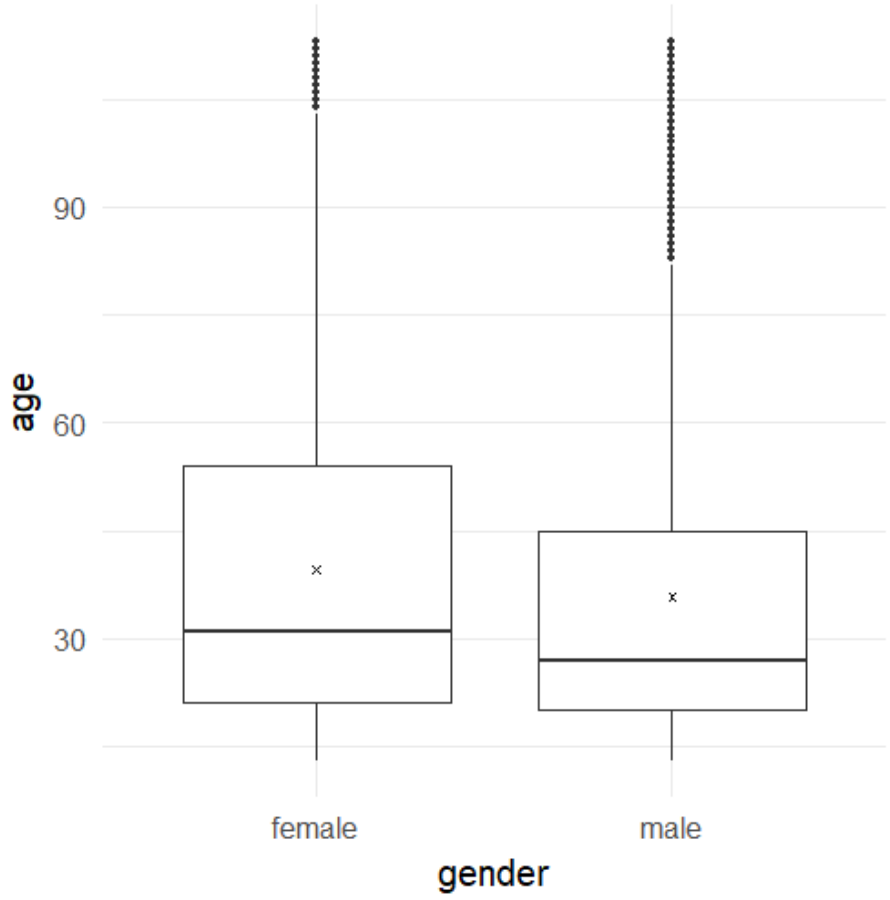
Lesson 7: Explore Many Variables

#title: "Udacity_Lesson7_ExploreManyVariables" #author: "Armir Kaçabeti"

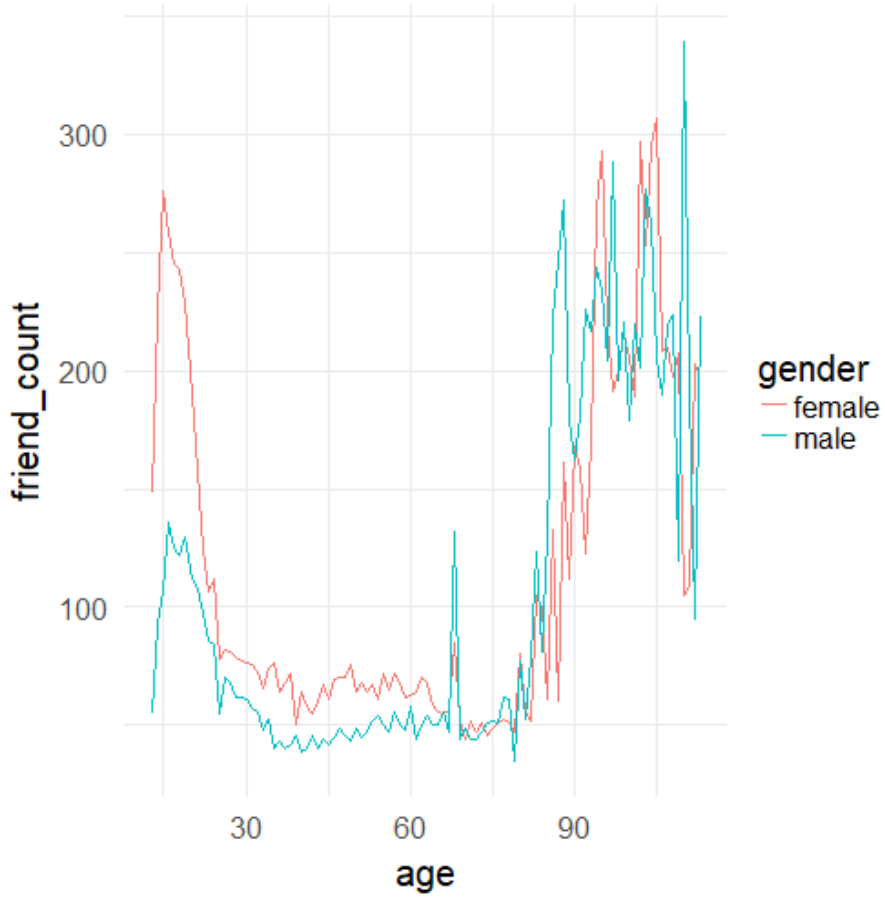
Lesson 7: Explore Many Variables

###Third Qualitative Variable

```
library(ggplot2)
pf <- read.csv('pseudo_facebook.tsv', sep = '\t')
ggplot(aes(x = gender, y = age),
       data = subset(pf, !is.na(gender))) + geom_boxplot() +
  stat_summary(fun.y = mean, geom = 'point', shape = 4)
```




```
ggplot(aes(x = age, y = friend_count),
       data = subset(pf, !is.na(gender))) +
  geom_line(aes(color = gender), stat = 'summary', fun.y = median)
```



#Q1

```
library(dplyr)

age_gender_group <- group_by(pf, age, gender)
age_gender_group <- filter(age_gender_group, !is.na(gender))
pf.fc_by_age_gender <- summarise(age_gender_group,
                                 mean_friend_count = mean(friend_count),
                                 median_friend_count = median(friend_count),
                                 n = n())
arrange(pf.fc_by_age_gender, age)
```

```
## # A tibble: 202 x 5
## # Groups:   age [101]
##      age gender mean_friend_count median_friend_count     n
##    <int> <fct>              <dbl>               <dbl> <int>
##  1    13 female               259               148     193
##  2    13 male                 102                55.0   291
```

```
##  3    14 female              362               224     847
##  4    14 male                164                92.5  1078
##  5    15 female              539               276    1139
##  6    15 male                201               106    1478
##  7    16 female              520               258    1238
##  8    16 male                240               136    1848
##  9    17 female              539               246    1236
## 10    17 male                236               125    2045
## # ... with 192 more rows
```

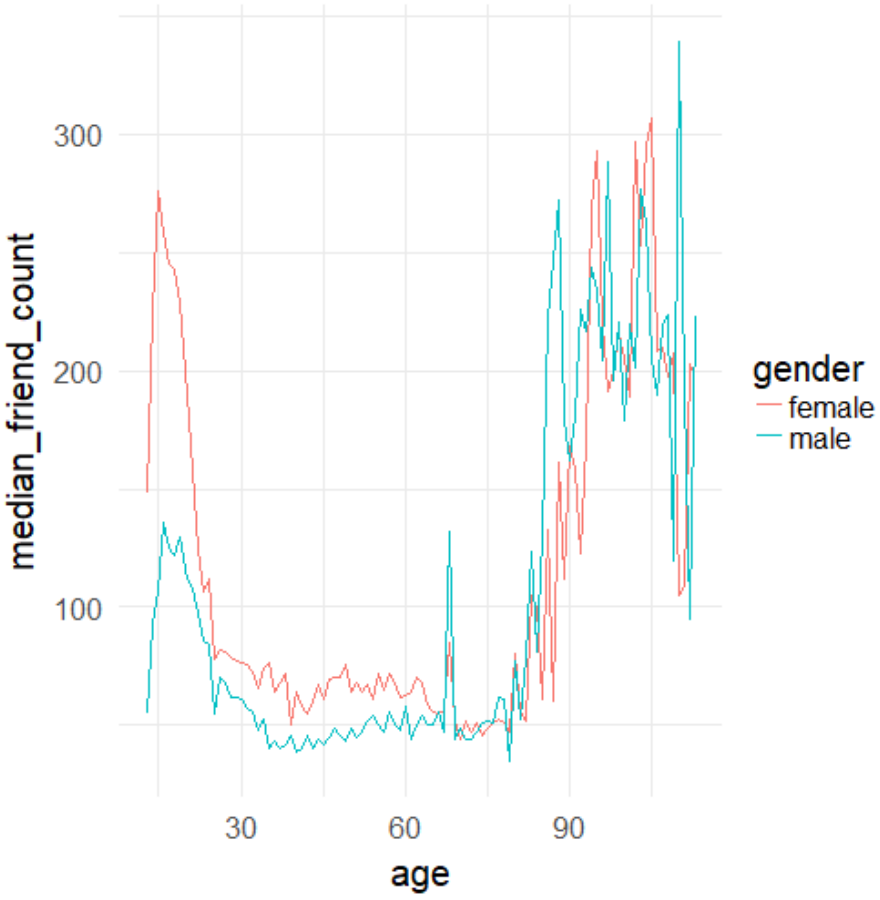
```
head(pf.fc_by_age_gender, 10)
```

```
## # A tibble: 10 x 5
## # Groups:   age [5]
##      age gender mean_friend_count median_friend_count     n
##    <int> <fct>              <dbl>               <dbl> <int>
##  1    13 female               259               148     193
##  2    13 male                 102                55.0   291
##  3    14 female               362               224     847
##  4    14 male                 164                92.5  1078
##  5    15 female               539               276    1139
##  6    15 male                 201               106    1478
##  7    16 female               520               258    1238
##  8    16 male                 240               136    1848
##  9    17 female               539               246    1236
## 10    17 male                 236               125    2045
```

#Q2

Plotting Conditional Summaries

```
ggplot(data = pf.fc_by_age_gender, aes(x = age, y = median_friend_count)) + geom_line(aes(co
```



Reshaping Data

```
#install.packages('reshape2')
library(reshape2)

pf.fc_by_age_gender.wide <- dcast(pf.fc_by_age_gender,
                                  age ~ gender,
                                  value.var = 'median_friend_count')

head(pf.fc_by_age_gender.wide)
```

```
##   age female  male
## 1  13  148.0  55.0
## 2  14  224.0  92.5
## 3  15  276.0 106.5
```

```
## 4  16  258.5 136.0
## 5  17  245.5 125.0
## 6  18  243.0 122.0
```

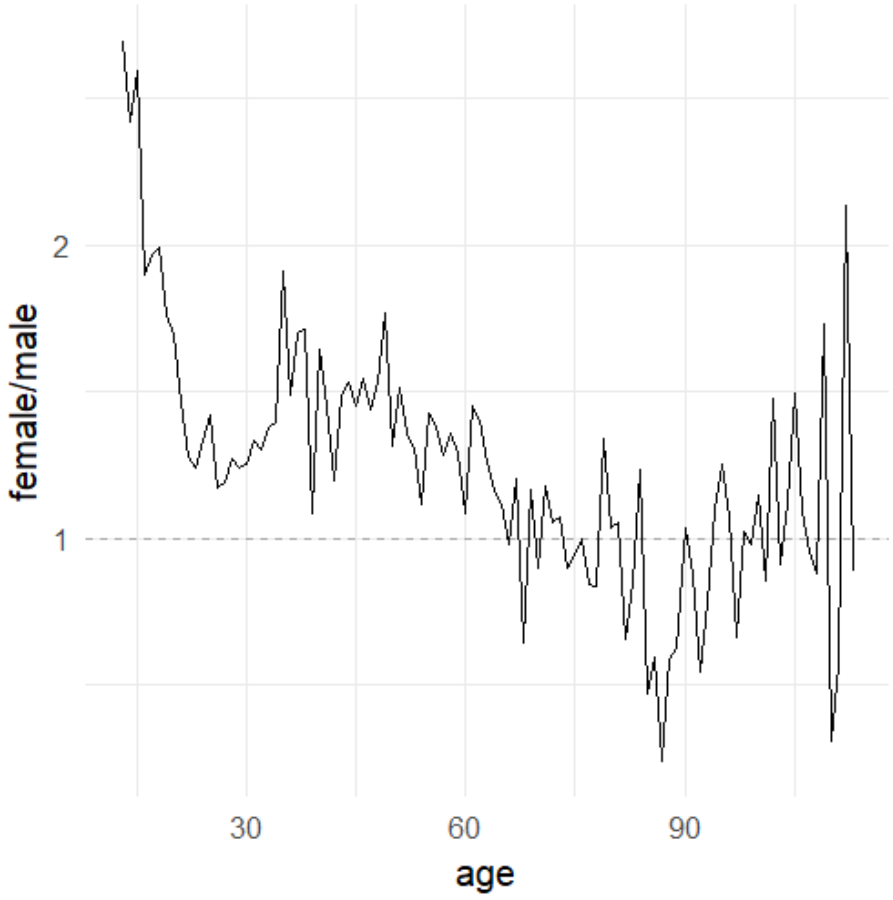
```
### Alternative code with dplyr and tidyr
library(dplyr)
#install.packages('tidyr')
library(tidyr)
pf.fc_by_age_gender.wide <-     subset(pf.fc_by_age_gender[c('age', 'gender', 'median_friend
    spread(gender, median_friend_count) %>%
    mutate(ratio = male / female)

head(pf.fc_by_age_gender.wide)
```

```
## # A tibble: 6 x 4
## # Groups:   age [6]
##     age female  male ratio
##   <int>  <dbl> <dbl> <dbl>
## 1    13    148  55.0 0.372
## 2    14    224  92.5 0.413
## 3    15    276 106   0.386
## 4    16    258 136   0.526
## 5    17    246 125   0.509
## 6    18    243 122   0.502
```

#Q3

Ratio Plot

```
ggplot(data = pf.fc_by_age_gender.wide, aes(x = age, y = female / male)) +
  geom_line() +
  geom_hline(yintercept = 1, alpha = 0.3, linetype = 2)
```



#Q4

Third Quantitative Variable

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

Cut a Variable

```
summary(pf$year_joined)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2005    2012    2012    2012    2013    2014       2
```

```
table(pf$year_joined)
```

```
##
##  2005  2006  2007  2008  2009  2010  2011  2012  2013  2014
```

```
##     9    15   581  1507  4557  5448  9860 33366 43588    70
```

```
#?cut
```

#Q5

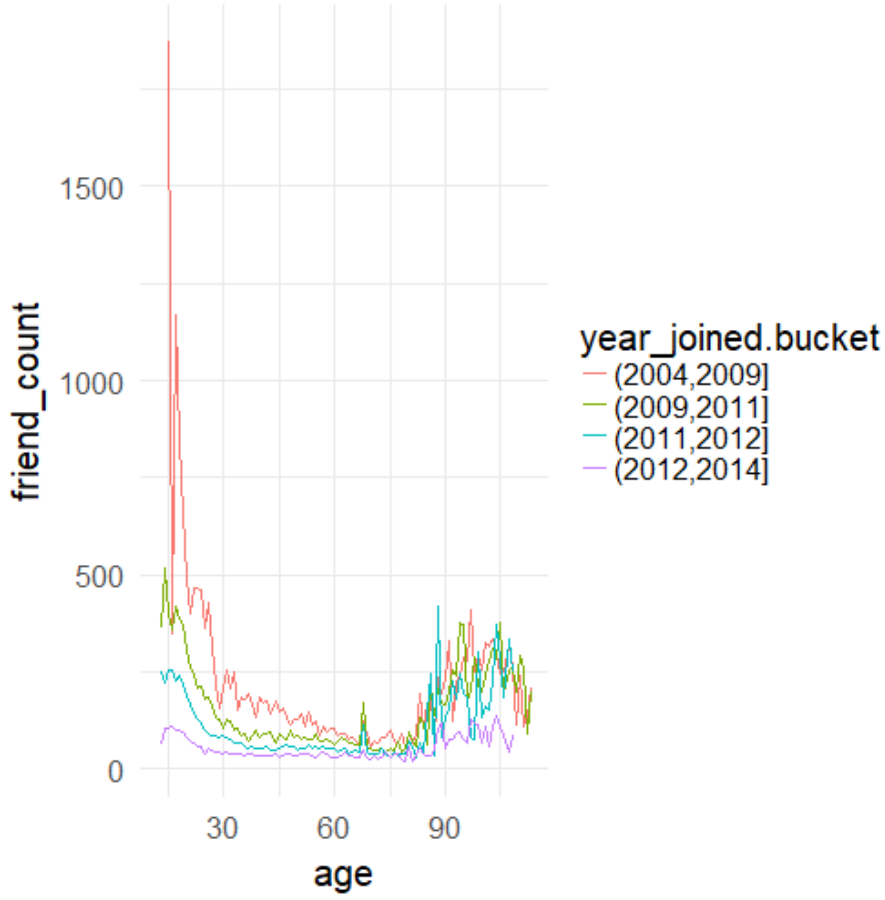
```
pf$year_joined.bucket <- cut(pf$year_joined, c(2004, 2009, 2011, 2012, 2014))
```

Plotting it All Together

```
table(pf$year_joined.bucket, useNA = 'ifany')
```

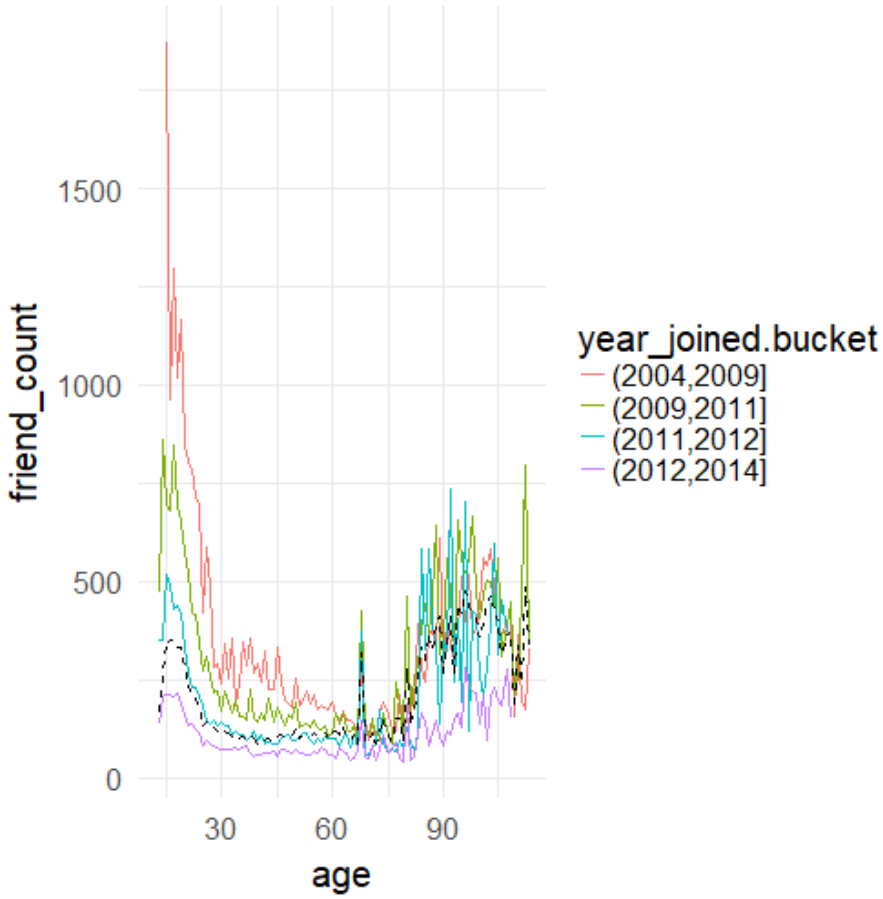
```
##
## (2004,2009] (2009,2011] (2011,2012] (2012,2014]        <NA>
##        6669       15308       33366       43658           2
```

```
ggplot(data = subset(pf, !is.na(year_joined.bucket)), aes(x= age, y=friend_count)) +
  geom_line(aes(color = year_joined.bucket), stat = 'summary', fun.y = median)
```



Plot the Grand Mean

```
ggplot(data = subset(pf, !is.na(year_joined.bucket)), aes(x= age, y=friend_count)) +
  geom_line(aes(color = year_joined.bucket), stat = 'summary', fun.y = mean) +
  geom_line(stat = 'summary', fun.y = mean, linetype = 2)
```



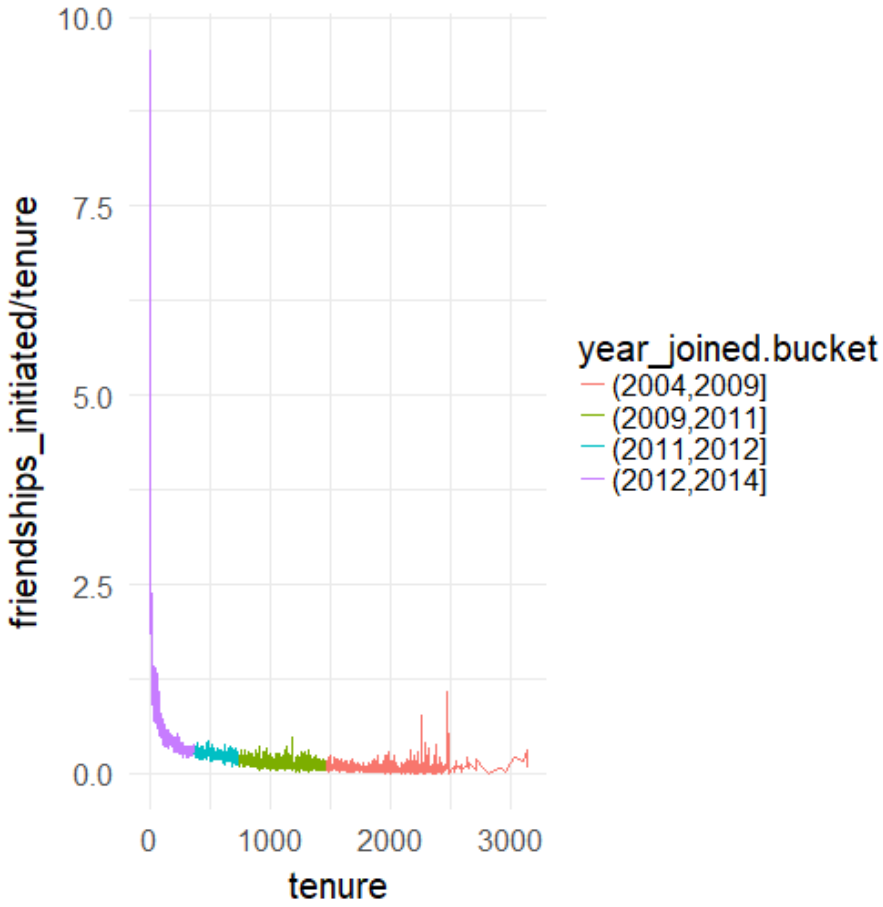
Friending Rate

```
with(subset(pf, tenure >= 1), summary(friend_count / tenure))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0775   0.2205   0.6096   0.5658 417.0000
```

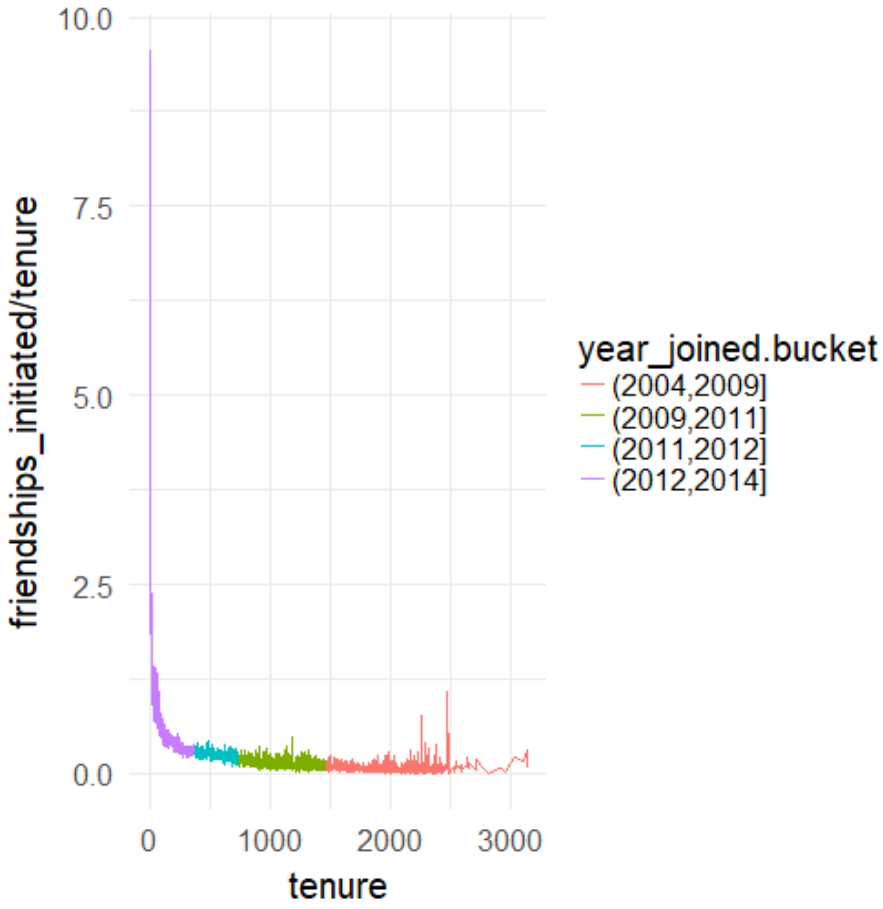
Friendships Initiated

```
ggplot(data = subset(pf, tenure >= 1), aes(x= tenure, y=friendships_initiated / tenure)) +
  geom_line(aes(color = year_joined.bucket), stat = 'summary', fun.y = mean)
```



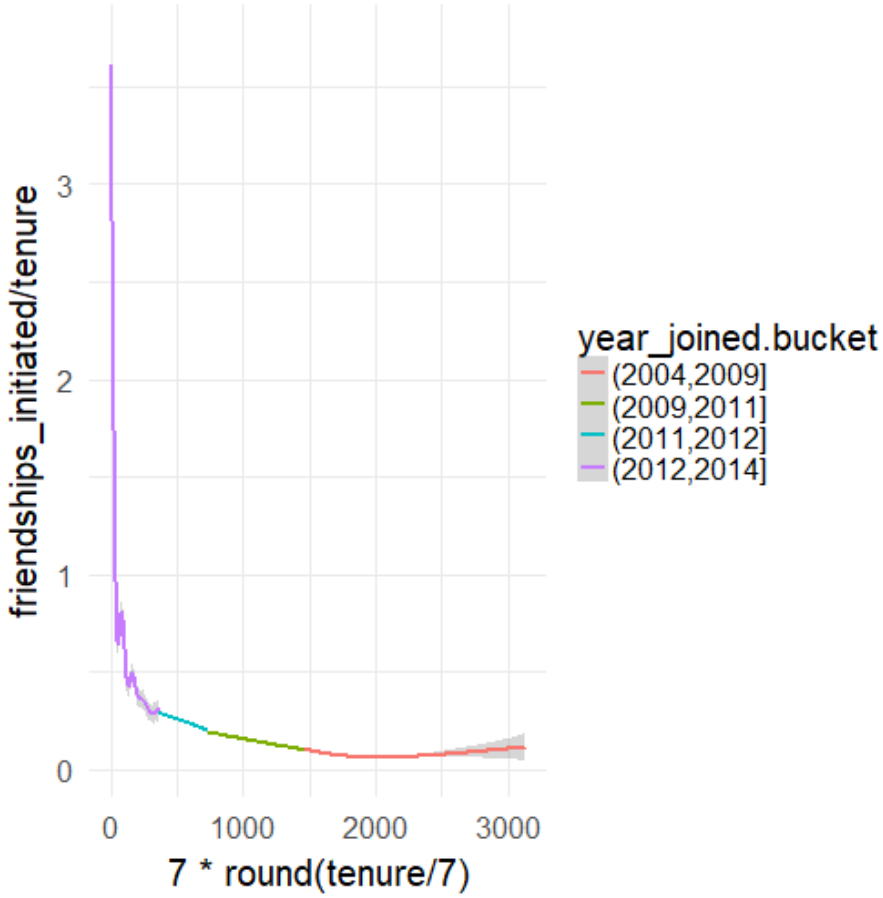
Bias-Variance Tradeoff Revisited

```
ggplot(aes(x = tenure, y = friendships_initiated / tenure),
       data = subset(pf, tenure >= 1)) +
  geom_line(aes(color = year_joined.bucket),
            stat = 'summary',
            fun.y = mean)
```

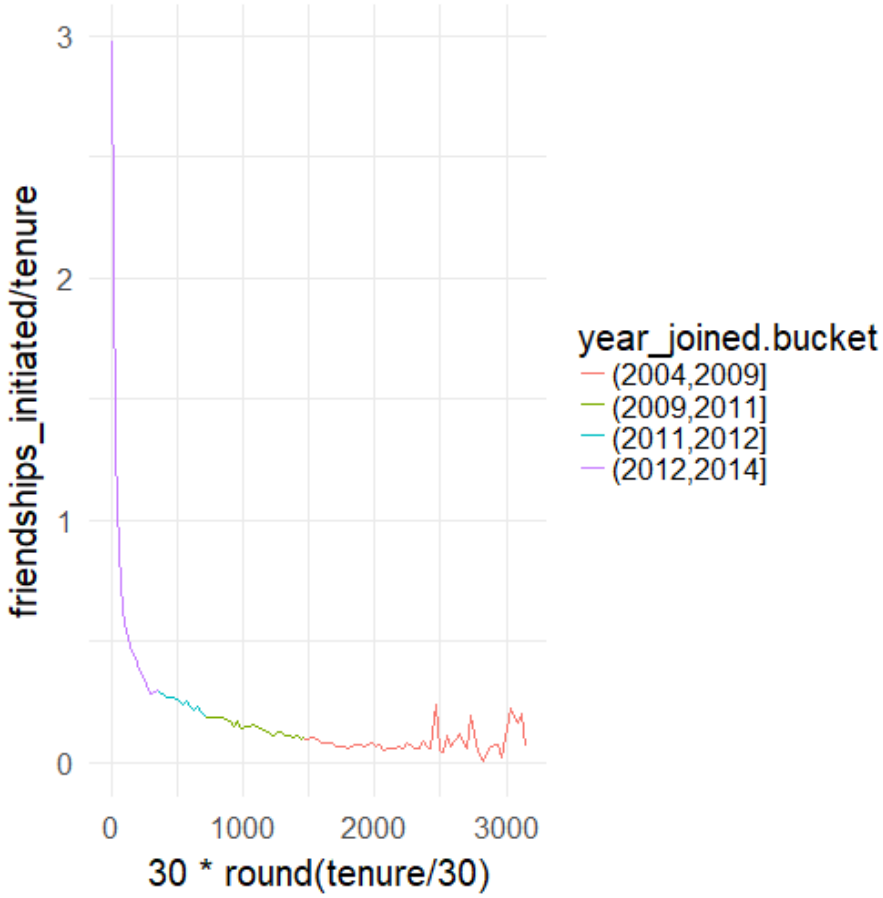



```
ggplot(aes(x = 7 * round(tenure / 7), y = friendships_initiated / tenure),
       data = subset(pf, tenure > 0)) +
  geom_smooth(aes(color = year_joined.bucket))
```

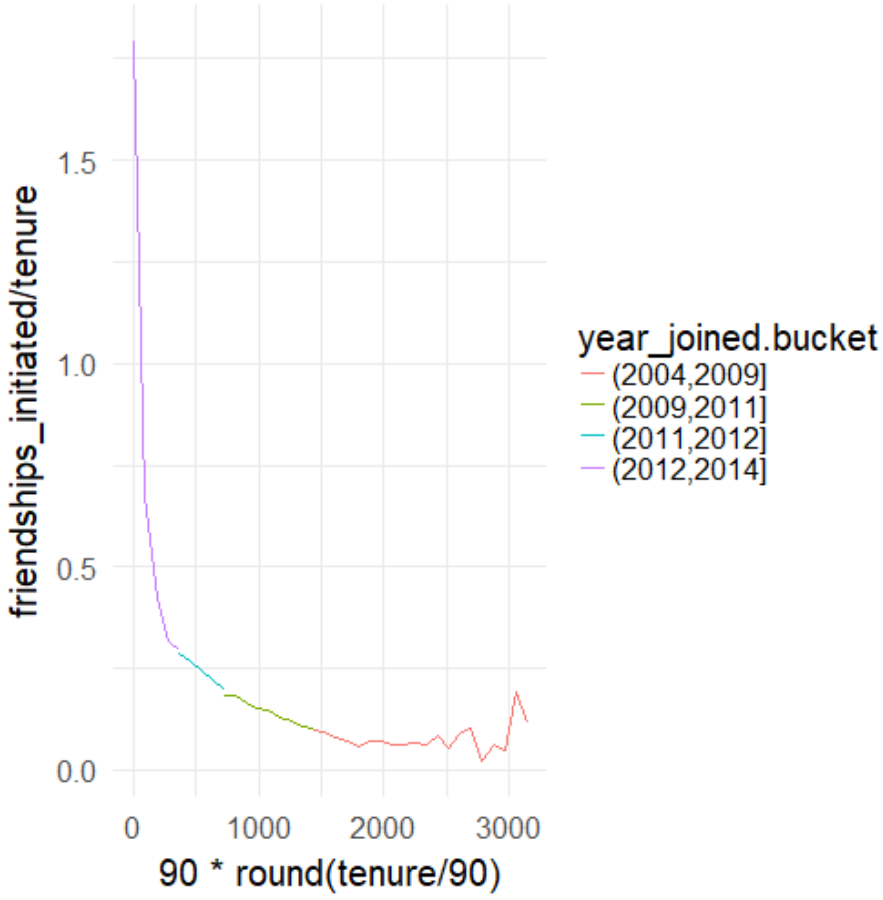
```
## `geom_smooth()` using method = 'gam'
```




```
ggplot(aes(x = 30 * round(tenure / 30), y = friendships_initiated / tenure),
       data = subset(pf, tenure > 0)) +
  geom_line(aes(color = year_joined.bucket),
            stat = "summary",
            fun.y = mean)
```



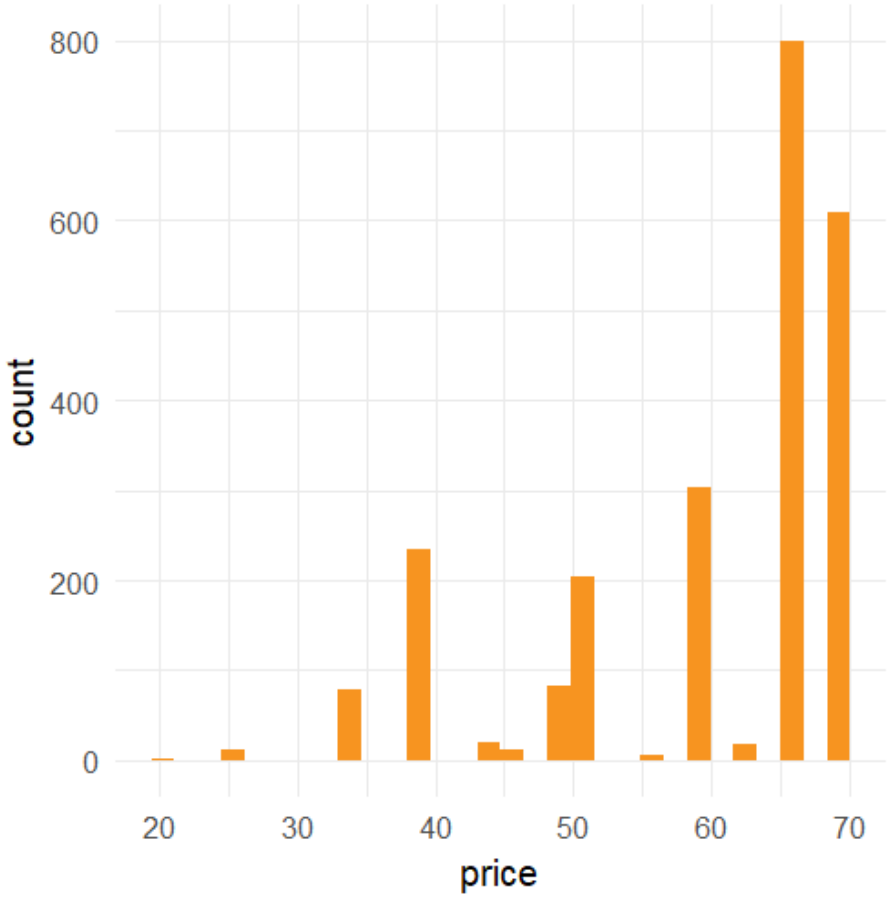

```
ggplot(aes(x = 90 * round(tenure / 90), y = friendships_initiated / tenure),
       data = subset(pf, tenure > 0)) +
  geom_line(aes(color = year_joined.bucket),
            stat = "summary",
            fun.y = mean)
```



Histograms Revisited

```
yo <- read.csv("yogurt.csv")
qplot(data = yo, x = price, fill = I('#F79420'))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Number of Purchases

```
summary(yo)
```

```
##       obs              id              time         strawberry
##  Min.   :   1.0   Min.   :2100081   Min.   : 9662   Min.   : 0.0000
##  1st Qu.: 696.5   1st Qu.:2114348   1st Qu.: 9843   1st Qu.: 0.0000
##  Median :1369.5   Median :2126532   Median :10045   Median : 0.0000
##  Mean   :1367.8   Mean   :2128592   Mean   :10050   Mean   : 0.6492
##  3rd Qu.:2044.2   3rd Qu.:2141549   3rd Qu.:10255   3rd Qu.: 1.0000
##  Max.   :2743.0   Max.   :2170639   Max.   :10459   Max.   :11.0000
##    blueberry        pina.colada          plain          mixed.berry
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 0.0000   Median : 0.0000   Median :0.0000   Median :0.0000
```

```
##  Mean   : 0.3571   Mean   : 0.3584   Mean   :0.2176   Mean   :0.3887
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :12.0000   Max.   :10.0000   Max.   :6.0000   Max.   :8.0000
##      price
##  Min.   :20.00
##  1st Qu.:50.00
##  Median :65.04
##  Mean   :59.25
##  3rd Qu.:68.96
##  Max.   :68.96
```

```
unique(yo$price)
```

```
##  [1] 58.96 65.04 48.96 68.96 39.04 24.96 50.00 45.04 33.04 44.00 33.36
## [12] 55.04 62.00 20.00 49.60 49.52 33.28 63.04 33.20 33.52
```

```
length(unique(yo$price))
```

```
## [1] 20
```

```
table(yo$price)
```

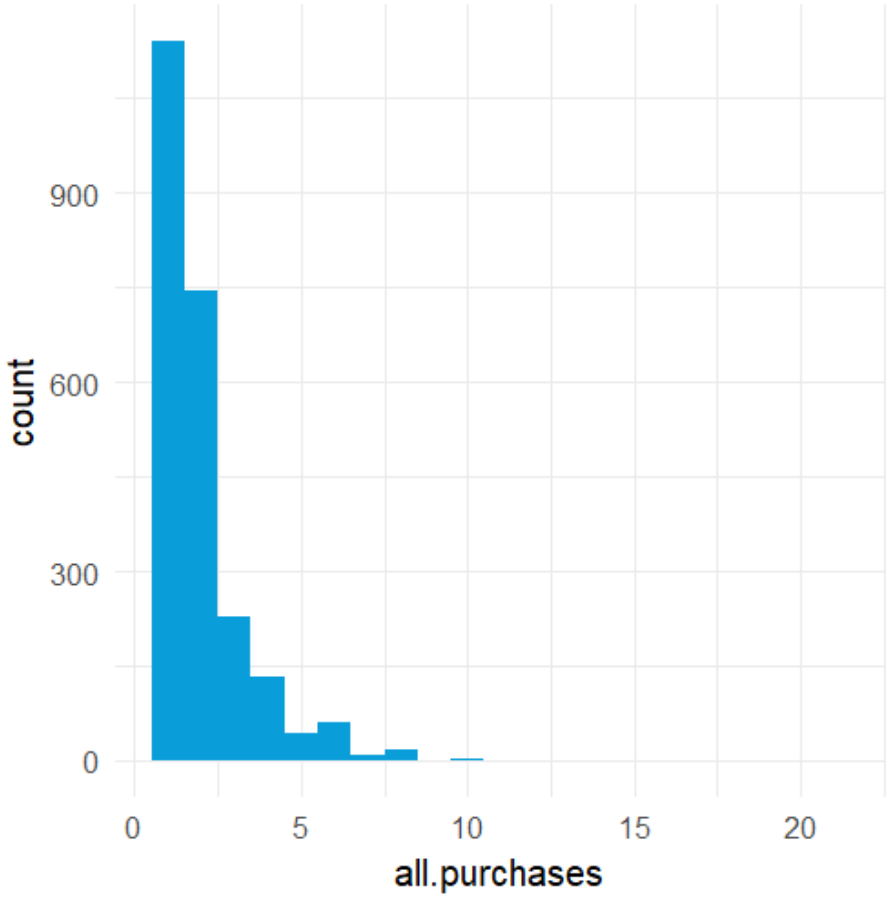
```
##
##    20 24.96 33.04  33.2 33.28 33.36 33.52 39.04    44 45.04 48.96 49.52
##     2    11    54     1     1    22     1   234    21    11    81     1
##  49.6    50 55.04 58.96    62 63.04 65.04 68.96
##     1   205     6   303    15     2   799   609
```

```
yo <- transform(yo, all.purchases = strawberry + blueberry +
                  pina.colada + plain + mixed.berry)

summary(yo$all.purchases)
```

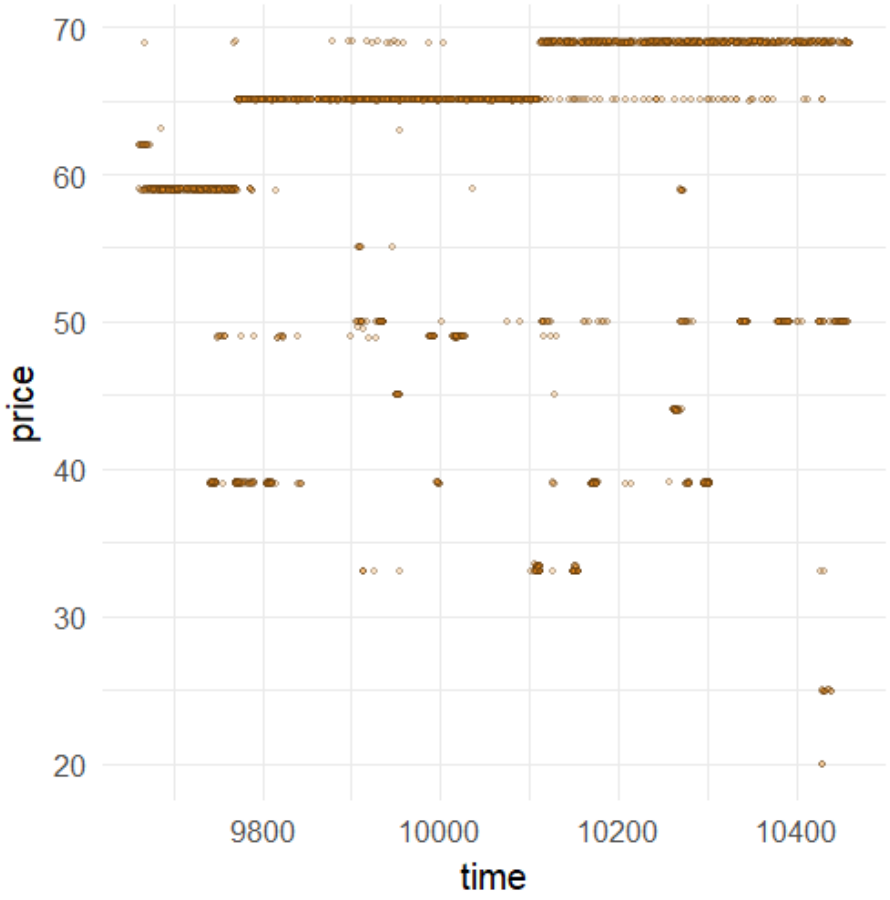
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.971   2.000  21.000
```

Prices over Time

```
qplot(x = all.purchases, data = yo, binwidth = 1,
      fill = I('#099DD9'))
```

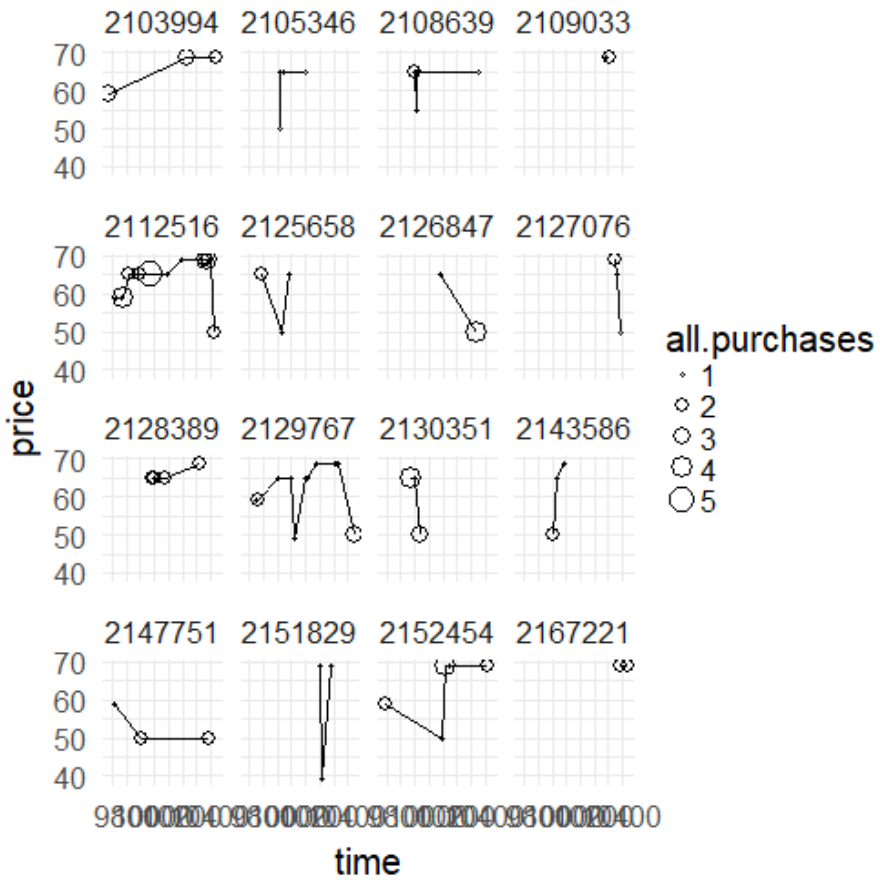



```
ggplot(data = yo, aes(x = time, y = price)) +
  geom_jitter(alpha = 1/4, shape = 21, fill = I('#F79420'))
```



Looking at Samples of Households

```
set.seed(2056)
sample_id <- unique(yo$id)
sample.ids <- sample(x = sample_id, size = 16)

ggplot(aes(x = time, y = price),
       data = subset(yo, id %in% sample.ids)) +
  facet_wrap( ~ id) +
  geom_line() +
  geom_point(aes(size = all.purchases), pch = 1)
```



Scatterplot Matrix

```
#install.packages("GGally")
library(GGally)

theme_set(theme_minimal(20))

# set the seed for reproducible results
set.seed(1836)
pf_subset <- pf[, c(2:15)]
names(pf_subset)
```

```
##  [1] "age"                   "dob_day"
##  [3] "dob_year"              "dob_month"
##  [5] "gender"                "tenure"
```

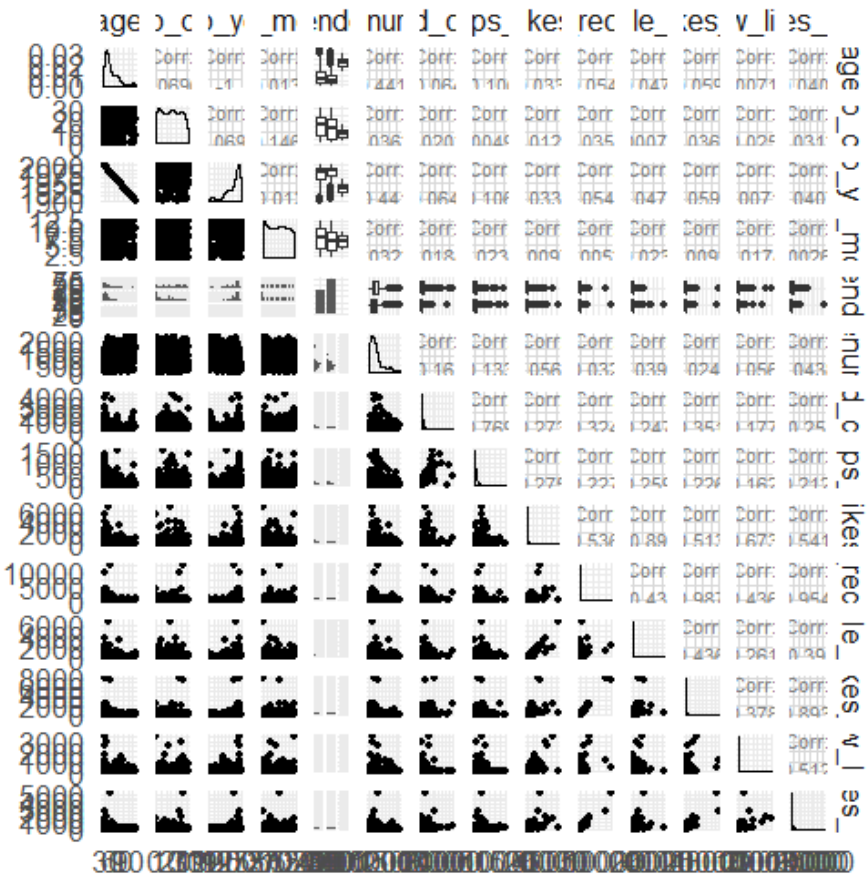
```
##  [7] "friend_count"          "friendships_initiated"
##  [9] "likes"                 "likes_received"
## [11] "mobile_likes"          "mobile_likes_received"
## [13] "www_likes"             "www_likes_received"
```

```
ggpairs(pf_subset[sample.int(nrow(pf_subset), 1000), ])
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).

## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Heat Maps

```
nci <- read.table("nci.tsv")
colnames(nci) <- c(1:64)

nci.long.samp <- melt(as.matrix(nci[1:200,]))
names(nci.long.samp) <- c("gene", "case", "value")
head(nci.long.samp)
```

```
##   gene case  value
## 1    1    1  0.300
## 2    2    1  1.180
## 3    3    1  0.550
## 4    4    1  1.140
## 5    5    1 -0.265
```

```
## 6    6    1 -0.070
```

```
ggplot(aes(y = gene, x = case, fill = value),
  data = nci.long.samp) +
  geom_tile() +
  scale_fill_gradientn(colours = colorRampPalette(c("blue", "red"))(100))
```