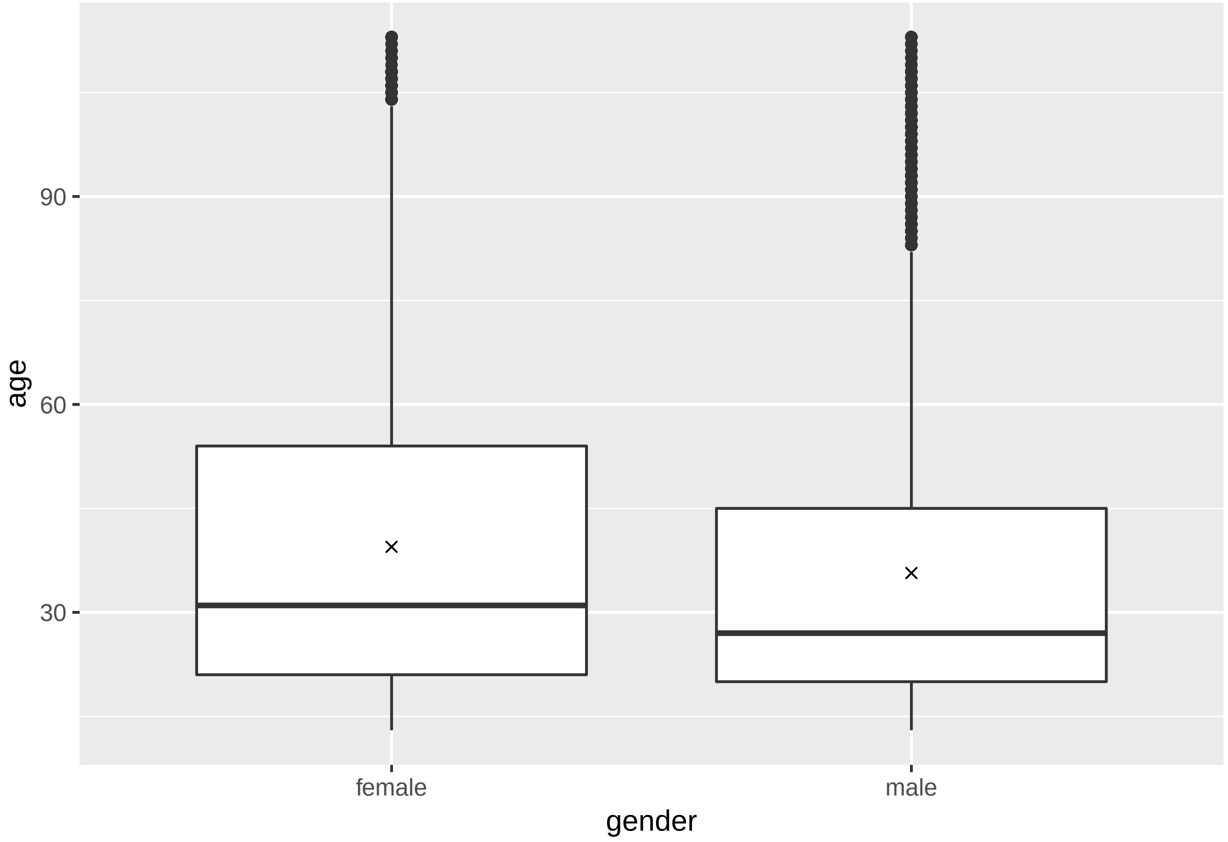
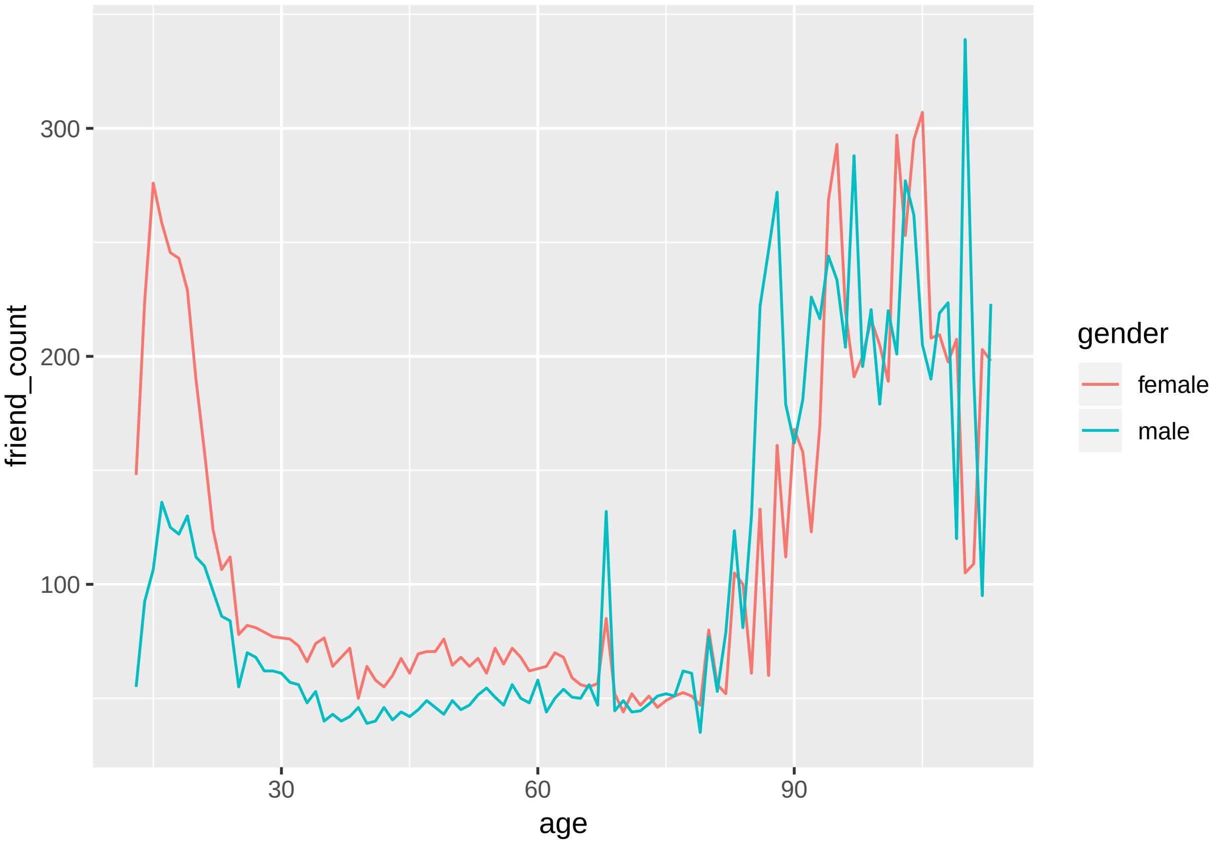
By: Reshu Singh

```
library(ggplot2)
pf <- read.csv('/home/reshu/Desktop/eda/lesson3/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',
```



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



```
library(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
```

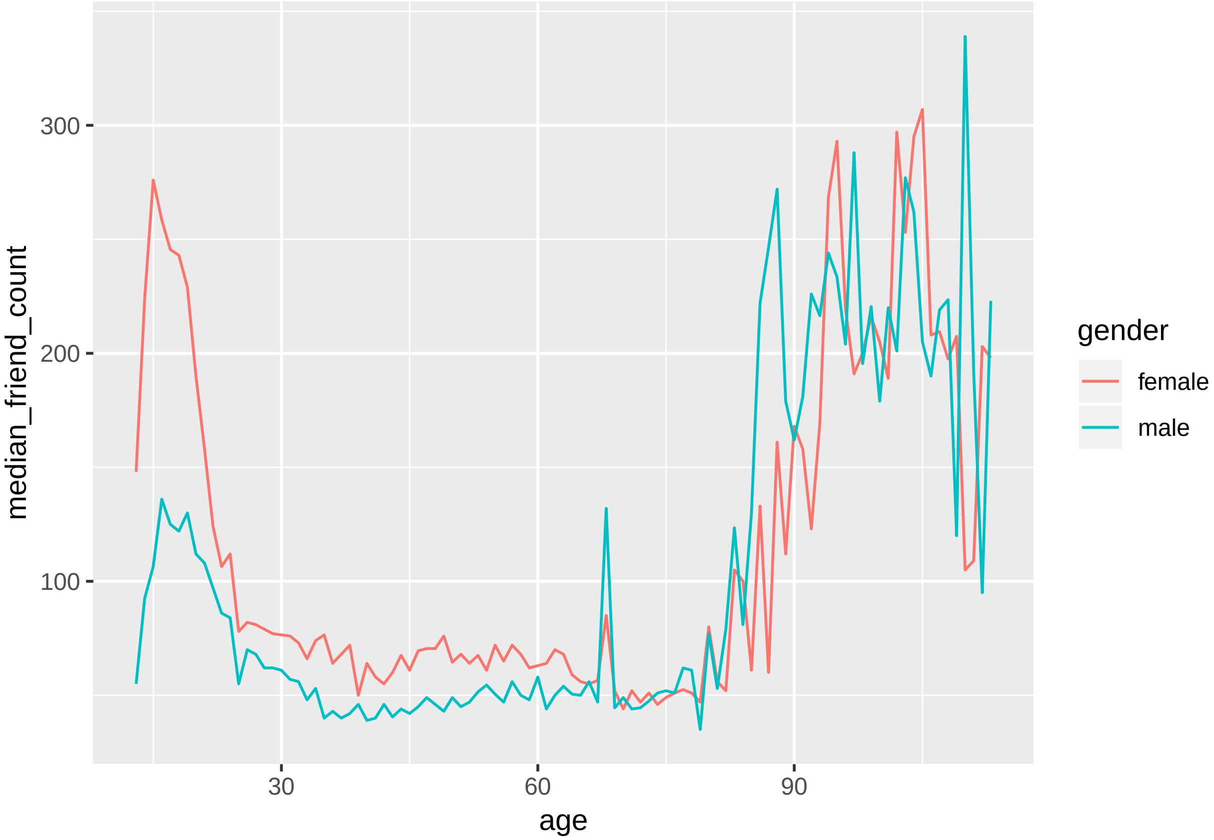
```
#chain functions together %>%
pf.fc_by_age_gender <- pf %>%
  filter(!is.na(gender)) %>%
  group_by(age, gender) %>%
  summarise(mean_friend_count = mean(friend_count),
            median_friend_count = median(friend_count),
            n = n()) %>%
  ungroup() %>%
  arrange(age)

head(pf.fc_by_age_gender)
```

```
## # A tibble: 6 x 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      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
```

Plotting Conditional Summaries

Notes:

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



Wide and Long Format

Notes:

Notes:

```
install.packages("tidyr")
```

```
## Installing package into '/home/reshu/R/x86_64-pc-linux-gnu-library/3.4'
## (as 'lib' is unspecified)
```

```r
library(tidyr)

spread(subset(pf.fc_by_age_gender,
       select = c('gender', 'age', 'median_friend_count')),
       gender, median_friend_count)
```

```
## # A tibble: 101 x 3
##      age female  male
##    <int>  <dbl> <dbl>
##  1    13    148   55
##  2    14    224   92.5
##  3    15    276  106.
##  4    16    258. 136
##  5    17    246. 125
##  6    18    243  122
##  7    19    229  130
##  8    20    190  112
##  9    21    158  108
## 10    22    124   97
## # ... with 91 more rows
```

Reshaping Data

Notes:

```r
install.packages('reshape2')
```

```
## Installing package into '/home/reshu/R/x86_64-pc-linux-gnu-library/3.4'
## (as 'lib' is unspecified)
```

```r
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

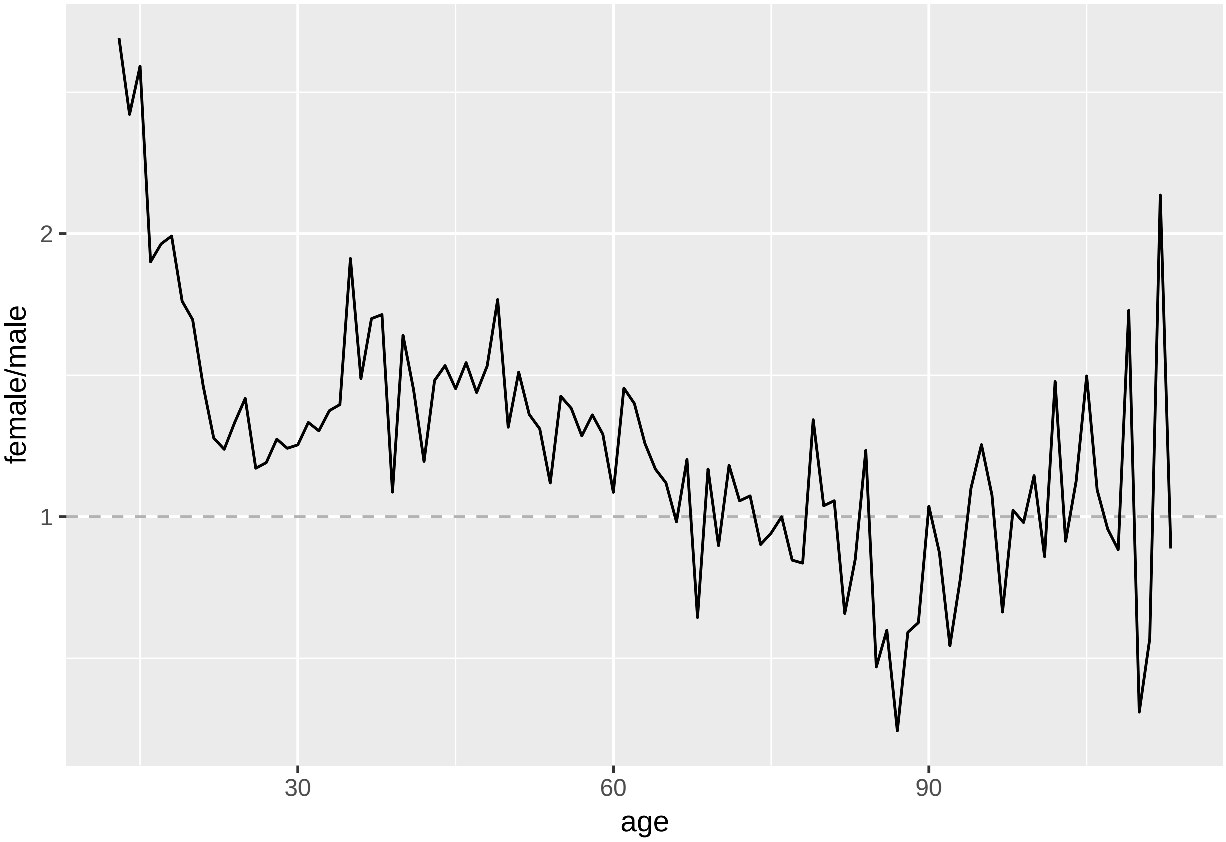
```r
pf.fc_by_age_gender.wide <-
  subset(pf.fc_by_age_gender[c('age', 'gender', 'median_friend_count')],
         !is.na(gender)) %>%
  spread(gender, median_friend_count) %>%
  mutate(ratio = male / female)

head(pf.fc_by_age_gender.wide)
```

```
## # A tibble: 6 x 4
##     age female  male ratio
##   <int>  <dbl> <dbl> <dbl>
## 1    13    148   55   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
```

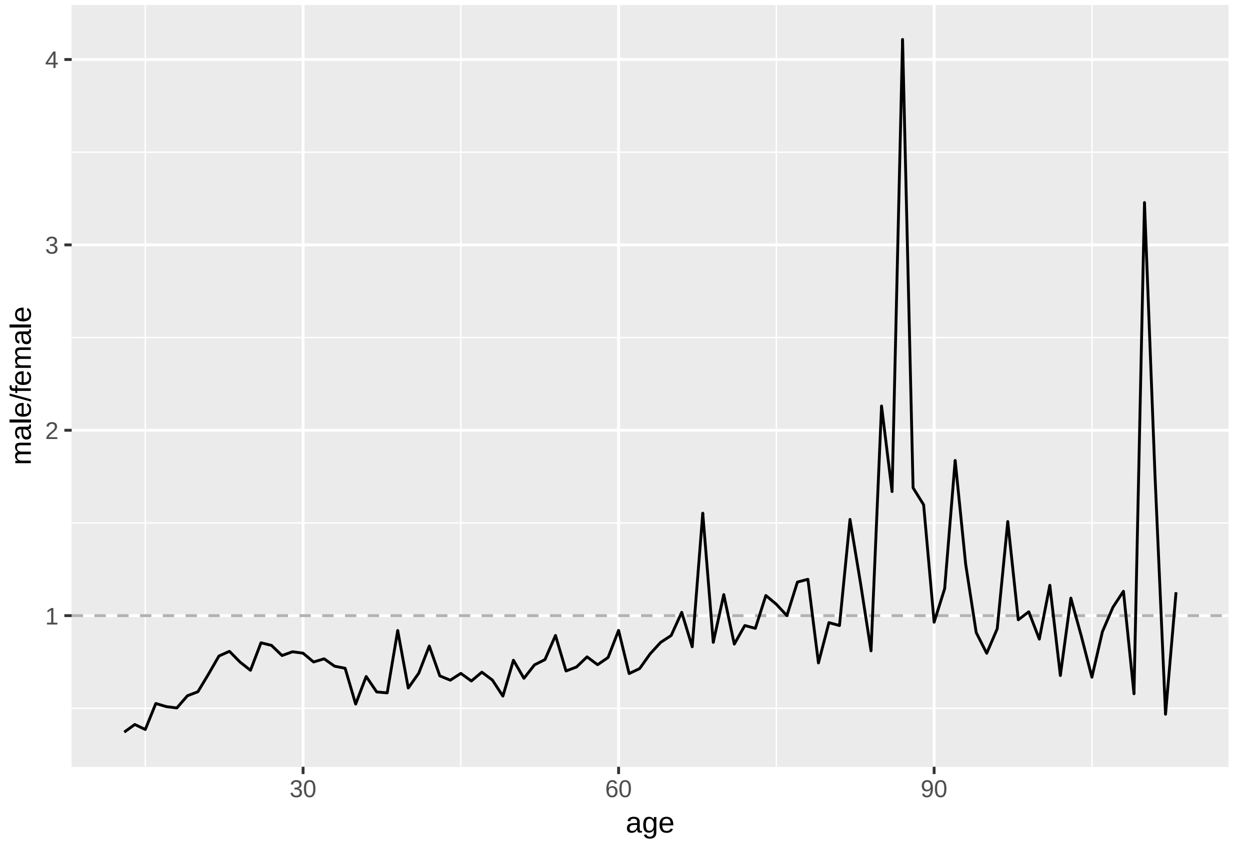
Ratio Plot

Notes:

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



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



***

Third Quantitative Variable

Notes:

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

Cut a Variable

Notes:

```
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
```

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

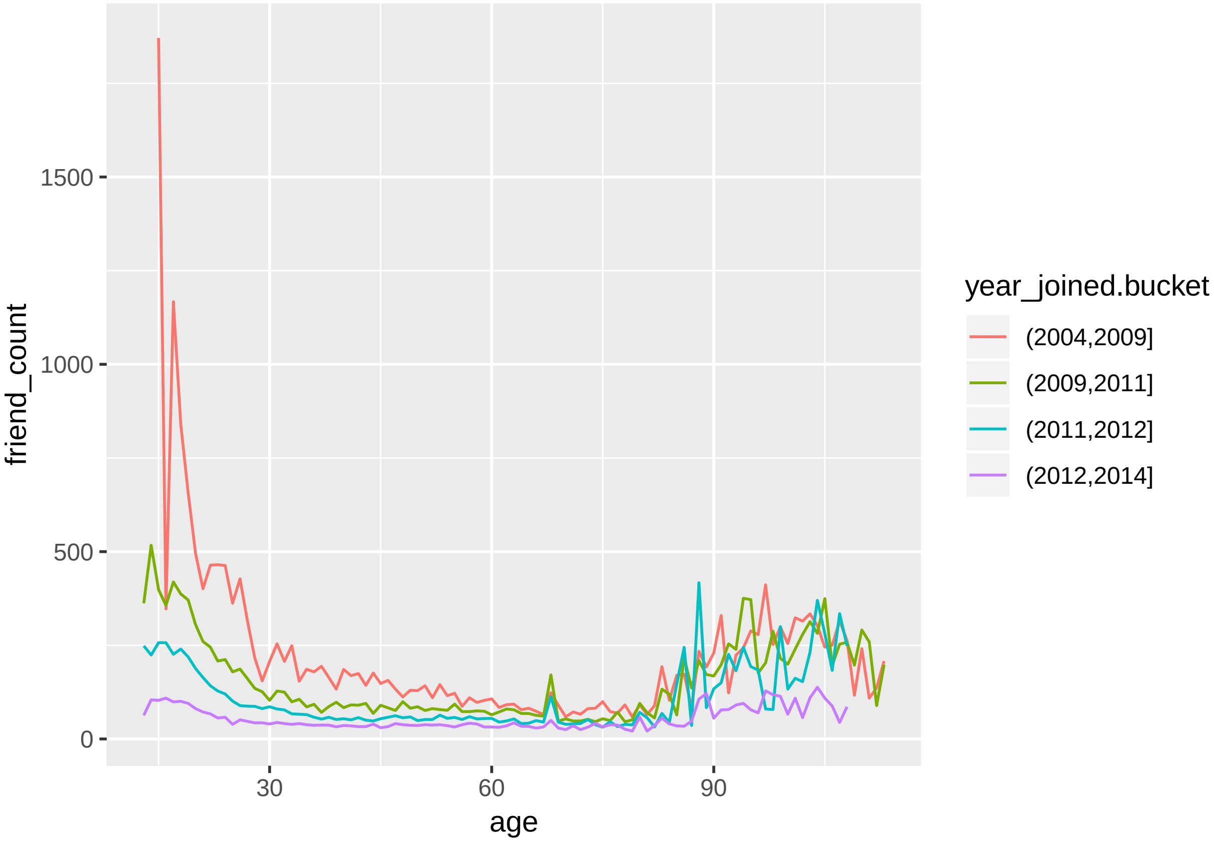
Plotting it All Together

Notes:

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

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

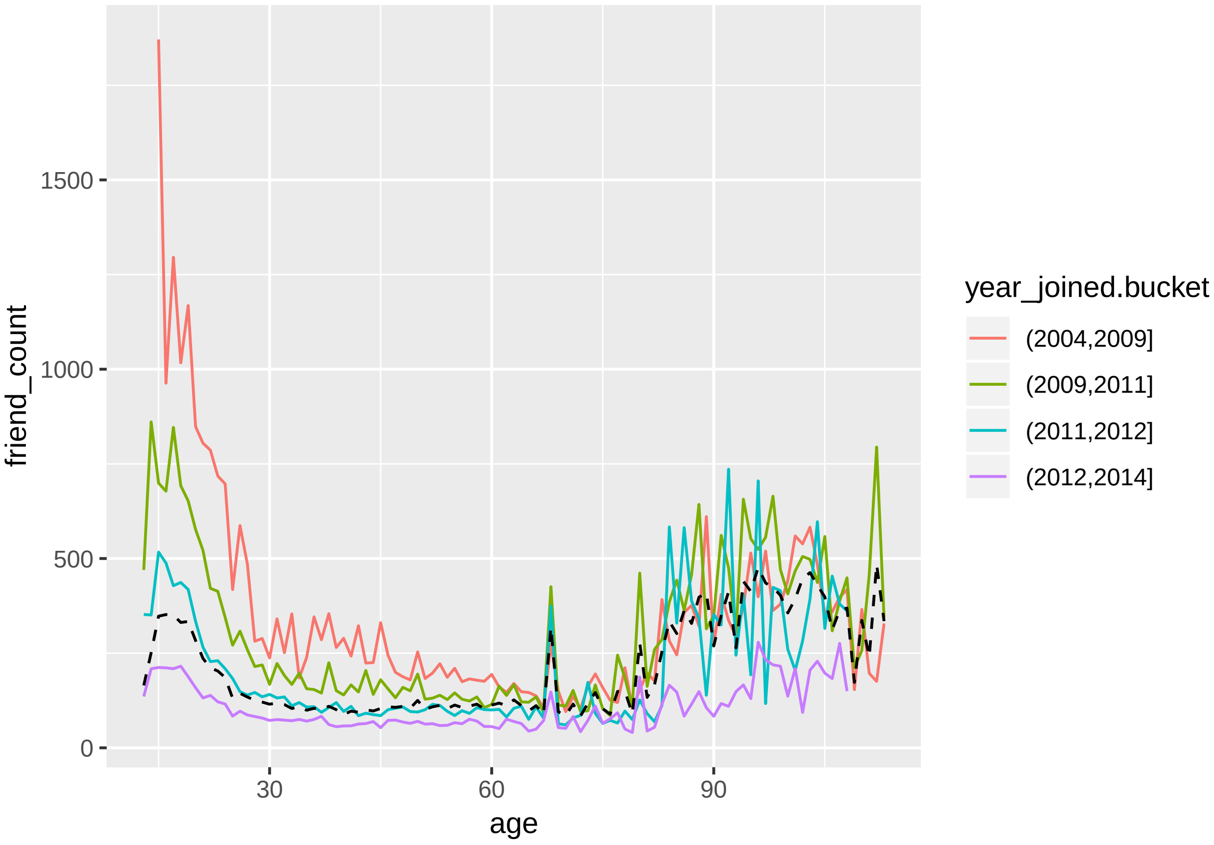
```
ggplot(aes(x = age, y = friend_count),
       data = subset(pf, !is.na(year_joined.bucket))) + geom_line(aes(color =year_joined.bucket), stat =
```



Plot the Grand Mean

Notes:

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



Friending Rate

Notes:

```
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
```

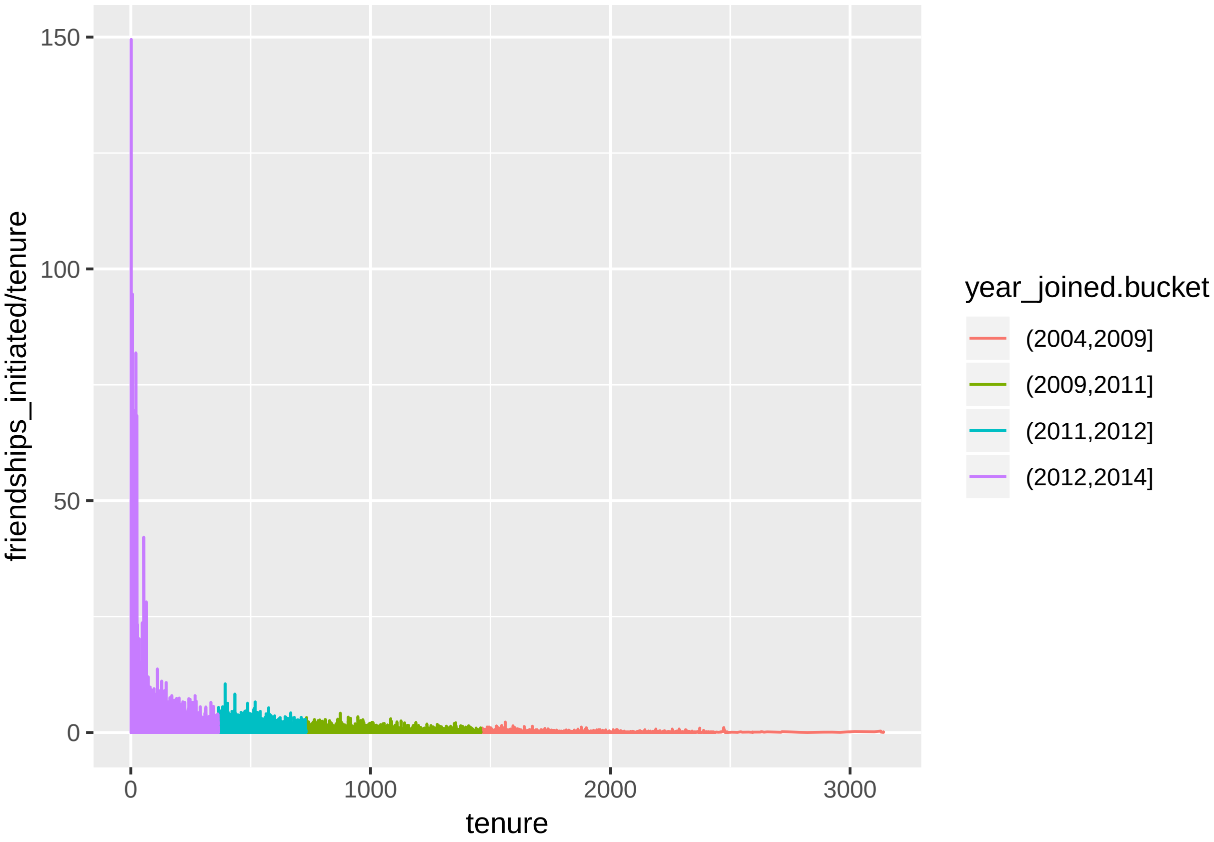
Friendships Initiated

Notes:

What is the median friend rate? 0.2205

What is the maximum friend rate? 417

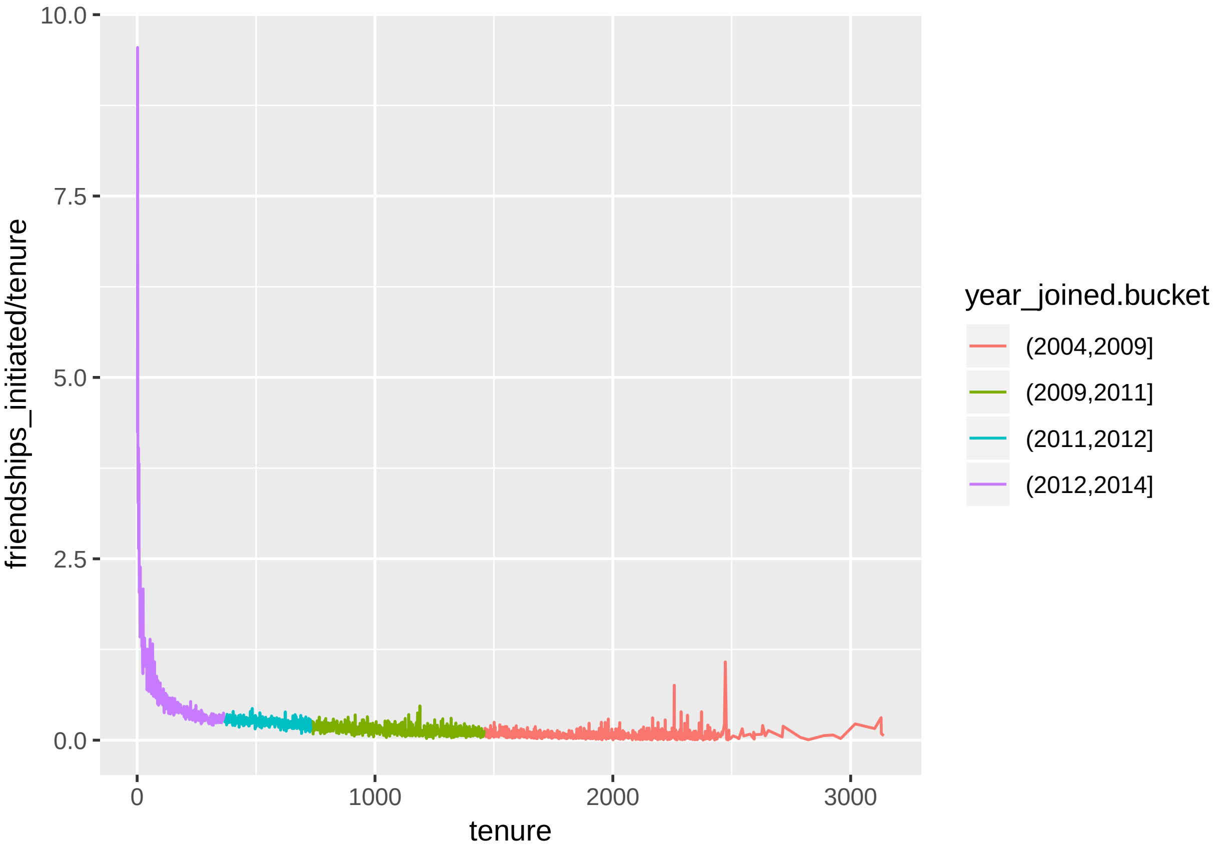
```
ggplot(aes(x = tenure, y = friendships_initiated / tenure),
       data = subset(pf, tenure >= 1)) +
  geom_line(aes(color = year_joined.bucket))
```



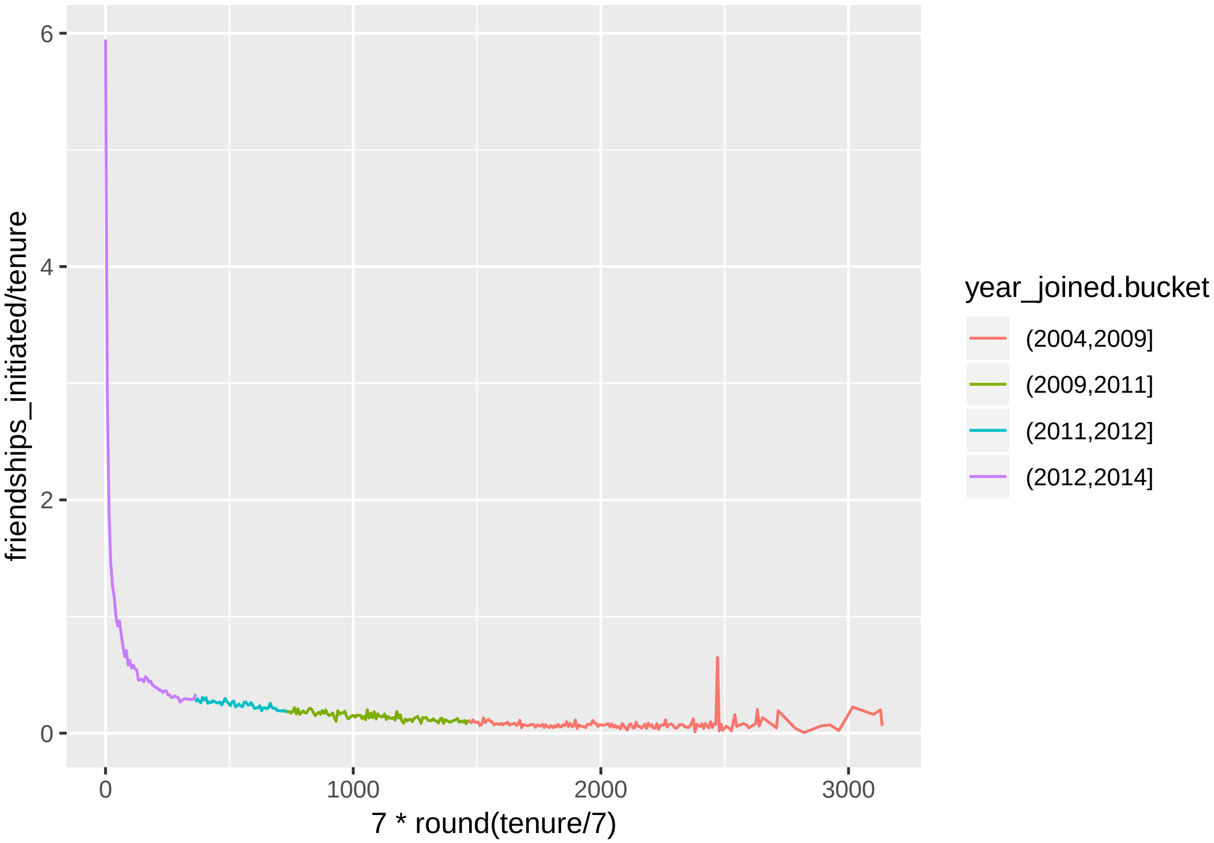
Bias-Variance Tradeoff Revisited

Notes:

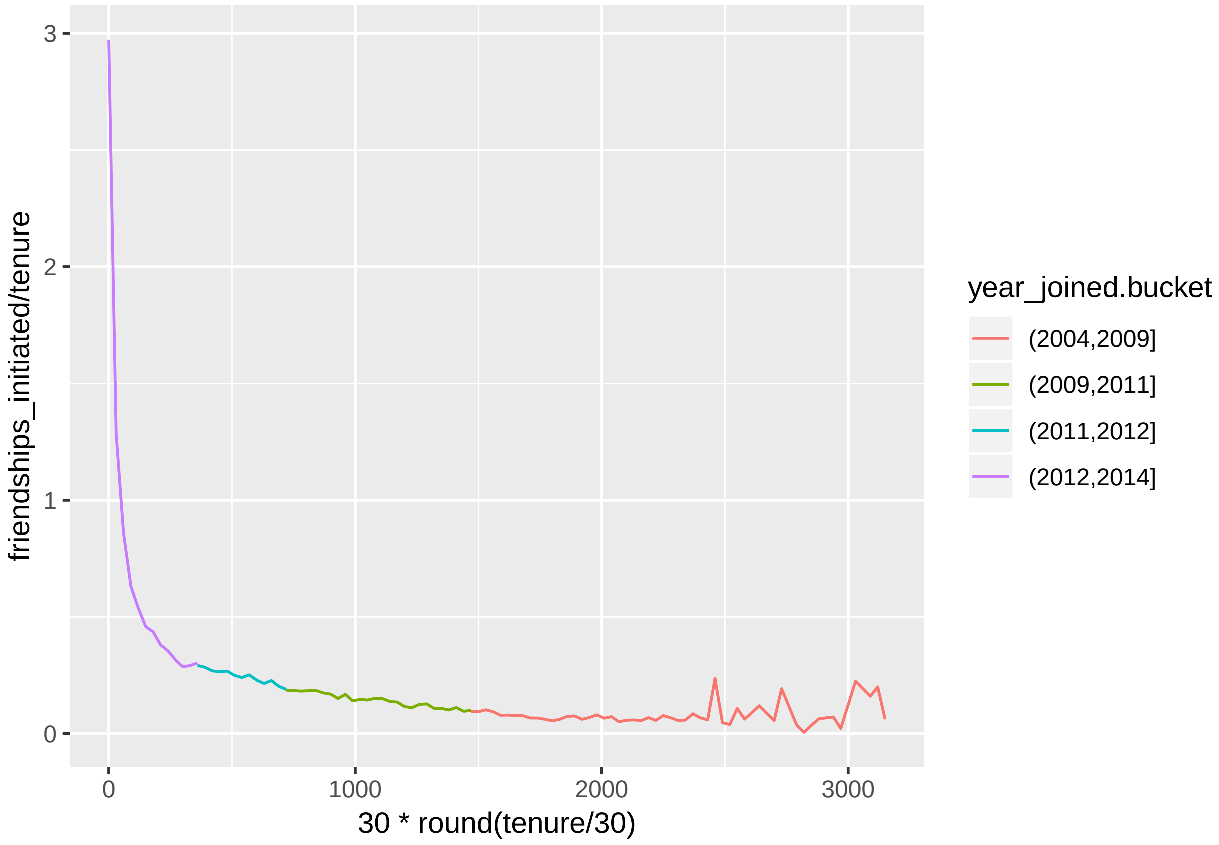
```
ggplot(aes(x = tenure, y = friendships_initiated / tenure),
       data = subset(pf, tenure >= 1)) +
  geom_line(aes(color = year_joined.bucket),
            stat = 'summary',
            fun.y = mean)
```



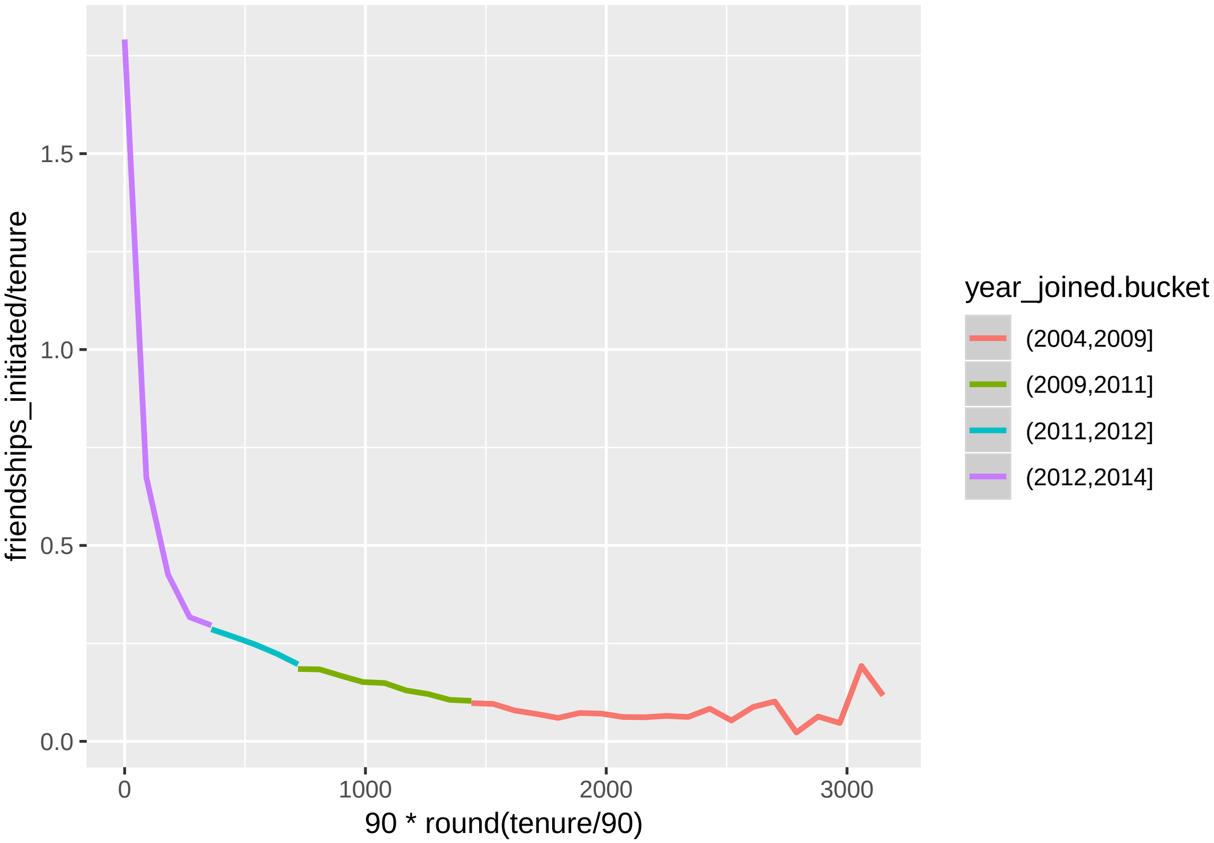
```
ggplot(aes(x = 7 * round(tenure / 7), y = friendships_initiated / tenure),
       data = subset(pf, tenure > 0)) +
  geom_line(aes(color = year_joined.bucket),
            stat = "summary",
            fun.y = mean)
```



```r
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_smooth(aes(color = year_joined.bucket),
              stat = "summary",
              fun.y = mean)
```



Sean’s NFL Fan Sentiment Study

Notes:

Introducing the Yogurt Data Set

Notes:

Histograms Revisited

Notes:

```
yo <- read.csv('/home/reshu/Desktop/eda/lesson5/yogurt.csv')
str(yo)
```

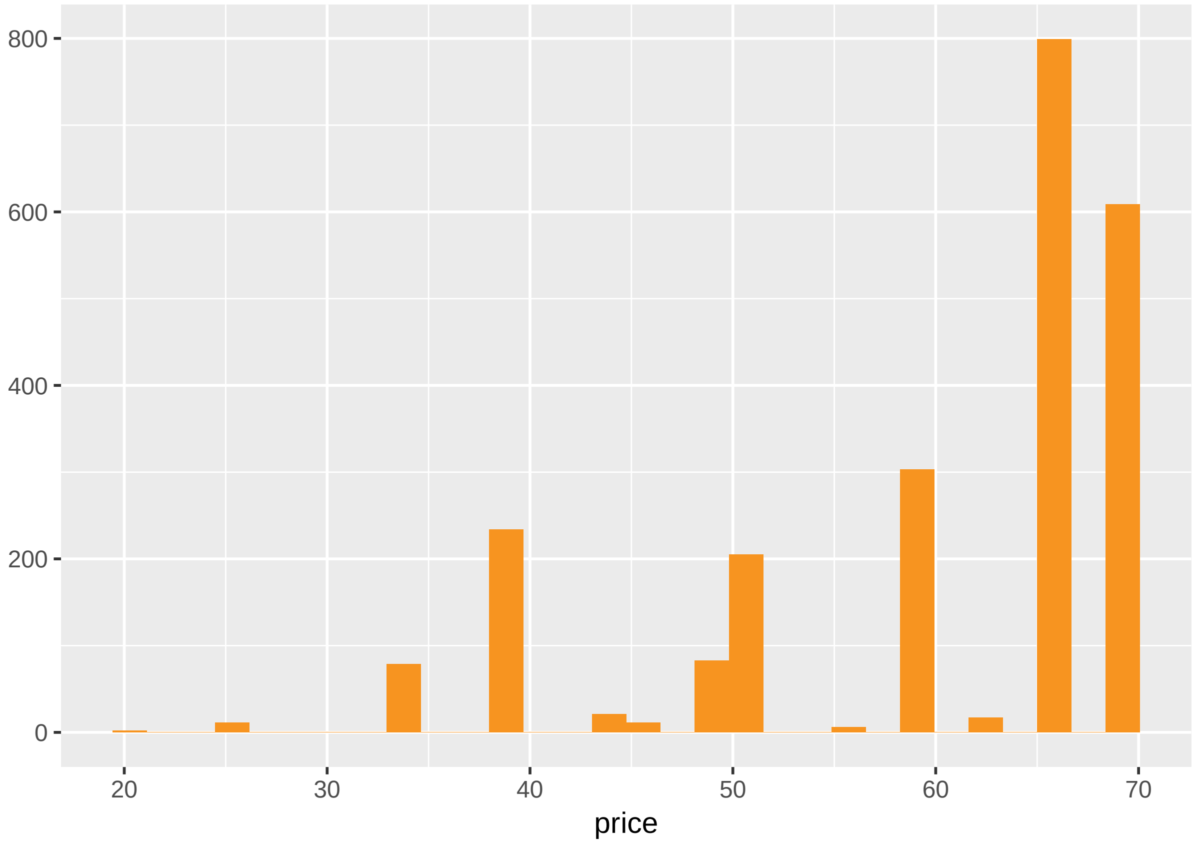
```
## 'data.frame':    2380 obs. of  9 variables:
##  $ obs        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ id         : int  2100081 2100081 2100081 2100081 2100081 2100081 2100081 2100081 2100081 2100081
##  $ time       : int  9678 9697 9825 9999 10015 10029 10036 10042 10083 10091 ...
##  $ strawberry : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ blueberry  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pina.colada: int  0 0 0 0 1 2 0 0 0 0 ...
##  $ plain      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mixed.berry: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ price      : num  59 59 65 65 49 ...
```

```
#Change the id from an int to a factor

yo$id <- factor(yo$id)
str(yo)
```

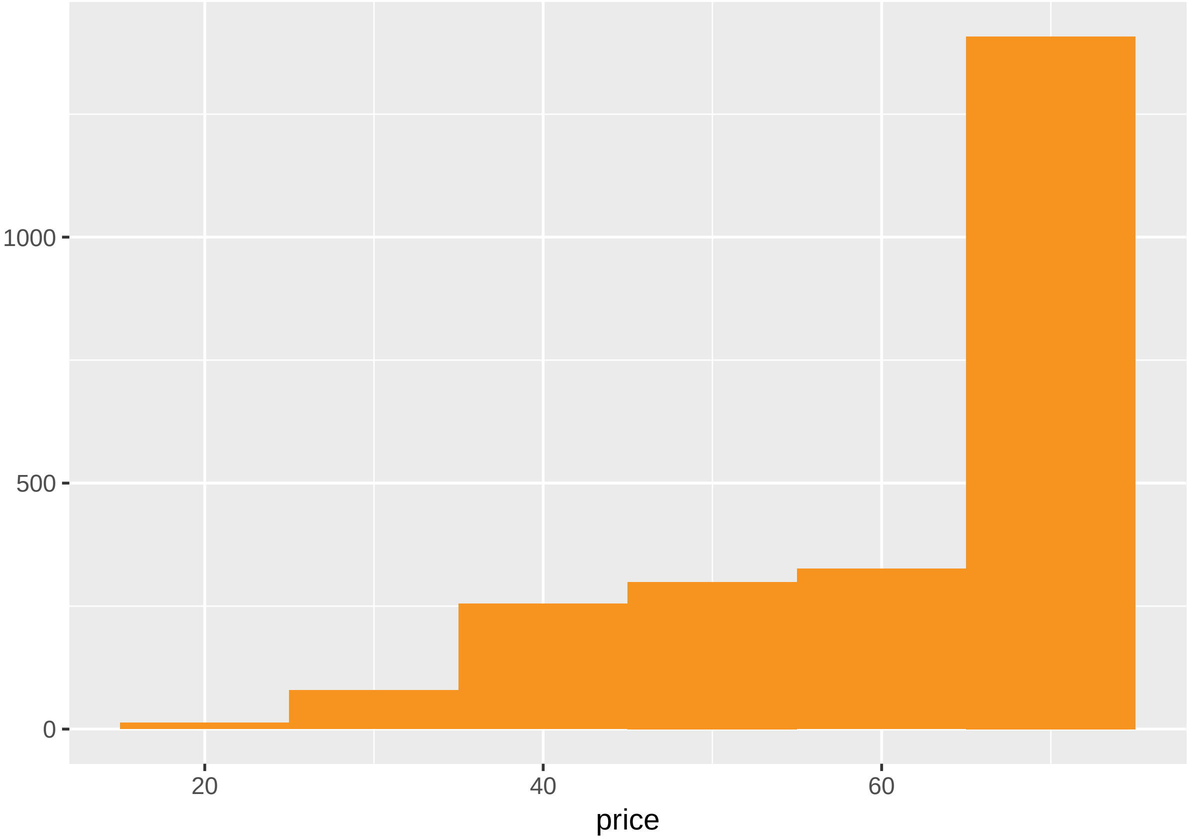
```
## 'data.frame':    2380 obs. of  9 variables:
##  $ obs        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ id         : Factor w/ 332 levels "2100081","2100370",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ time       : int  9678 9697 9825 9999 10015 10029 10036 10042 10083 10091 ...
##  $ strawberry : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ blueberry  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pina.colada: int  0 0 0 0 1 2 0 0 0 0 ...
##  $ plain      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mixed.berry: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ price      : num  59 59 65 65 49 ...
```

```
qplot(data = yo, x = price, fill = I('#F79420'))
```

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



```
qplot(data = yo, x = price, fill = I('#F79420'), binwidth = 10)
```



Number of Purchases

Notes:

```
summary(yo)
```

```
##       obs               id             time         strawberry
##  Min.   :   1.0   2132290:  74   Min.   : 9662   Min.   : 0.0000
##  1st Qu.: 696.5   2130583:  59   1st Qu.: 9843   1st Qu.: 0.0000
##  Median :1369.5   2124073:  50   Median :10045   Median : 0.0000
##  Mean   :1367.8   2149500:  50   Mean   :10050   Mean   : 0.6492
##  3rd Qu.:2044.2   2101790:  47   3rd Qu.:10255   3rd Qu.: 1.0000
##  Max.   :2743.0   2129528:  39   Max.   :10459   Max.   :11.0000
##                   (Other):2061
##    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
##
```

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

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

```r
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
```

```r
str(yo)
```

```
## 'data.frame':    2380 obs. of  9 variables:
##  $ obs        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ id         : Factor w/ 332 levels "2100081","2100370",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ time       : int  9678 9697 9825 9999 10015 10029 10036 10042 10083 10091 ...
##  $ strawberry : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ blueberry  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pina.colada: int  0 0 0 0 1 2 0 0 0 0 ...
##  $ plain      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mixed.berry: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ price      : num  59 59 65 65 49 ...
```

```r
yo$all.purchases
```

```
## NULL
```

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

summary(yo$all.purchases)
```

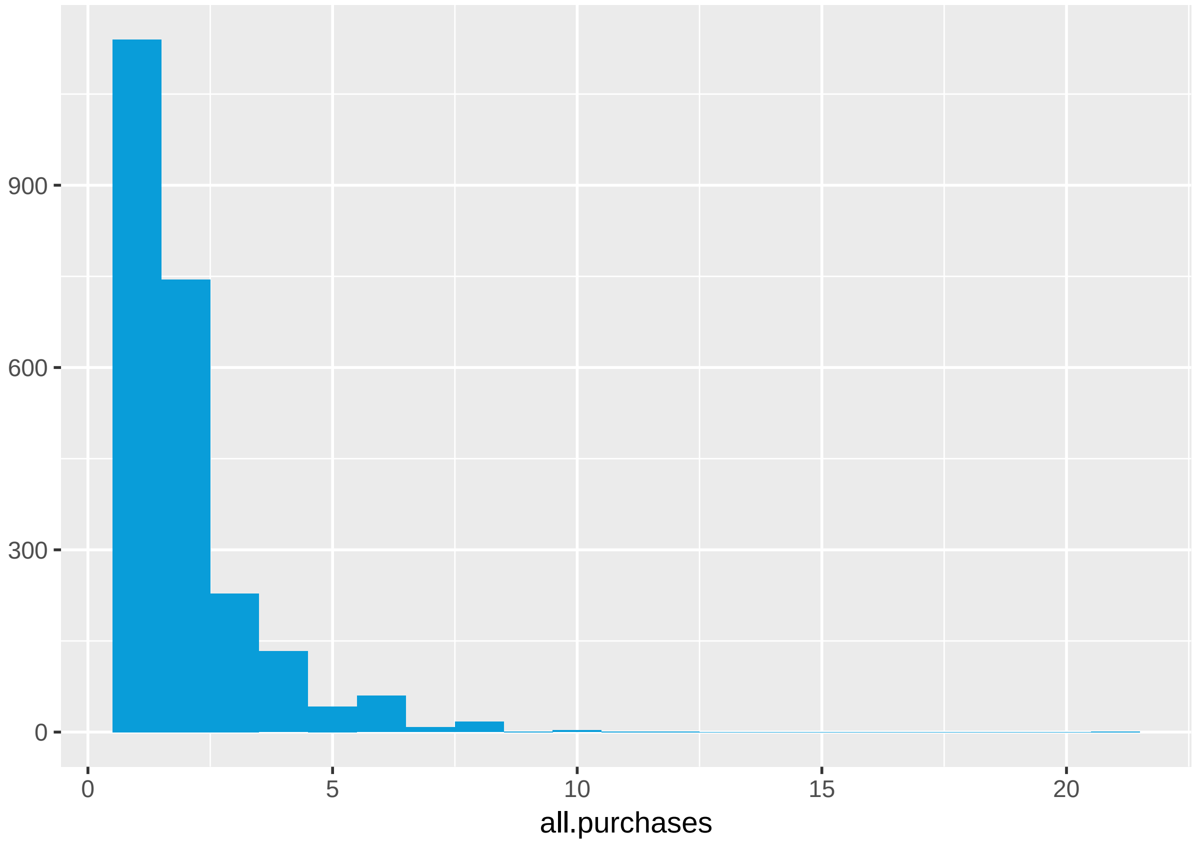
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.971   2.000  21.000
```

```r
#Alternate way

yo$all.purchases <- yo$strawberry + yo$blueberry + yo$pina.colada + yo$plain + yo$mixed.berry
```

Prices over Time

Notes:

```r
qplot(x = all.purchases, data = yo, binwidth = 1, fill = I('#099dd9'))
```

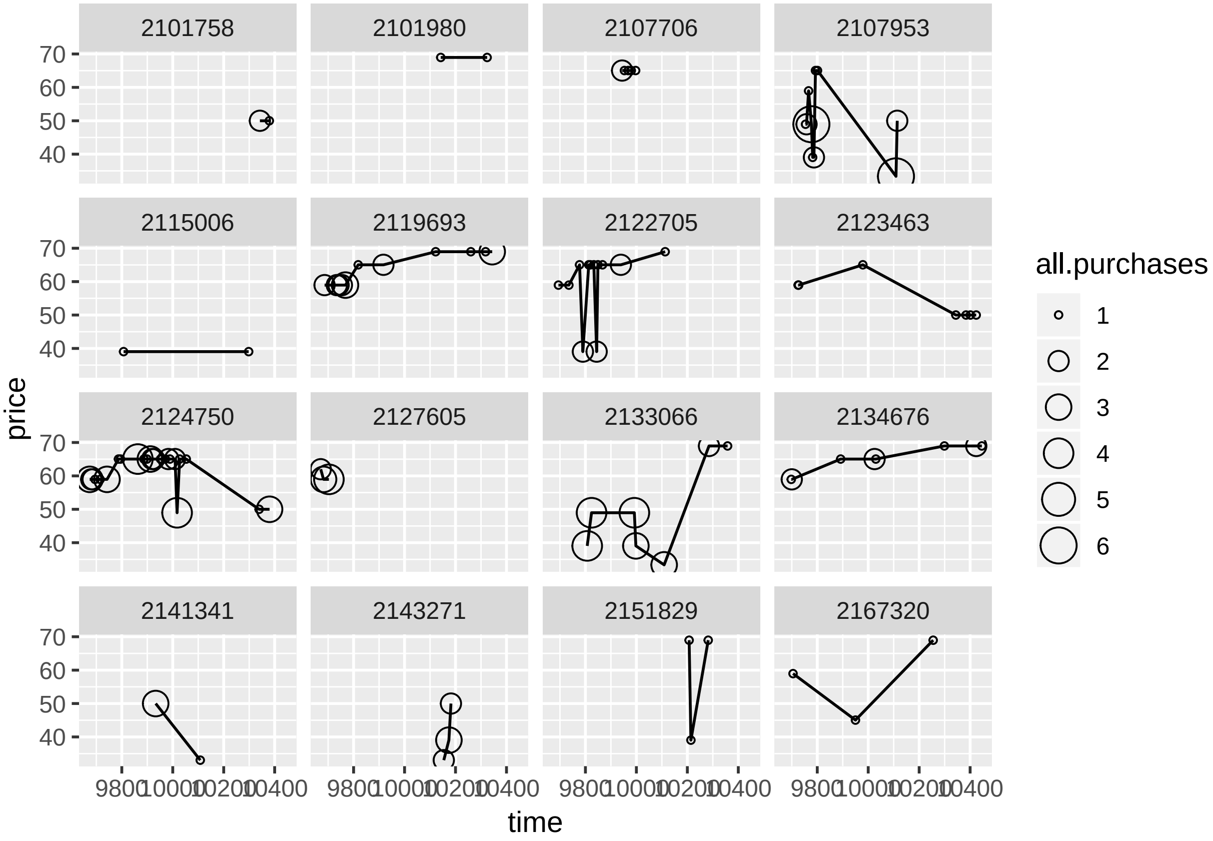


Sampling Observations

Notes:

Looking at Samples of Households

```r
#Set the seed for reproducible reults

set.seed(4230)
sample.ids <- sample(levels(yo$id), 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)
```



The Limits of Cross Sectional Data

Notes:

Many Variables

Notes:

Even More Variables

Notes:

Heat Maps

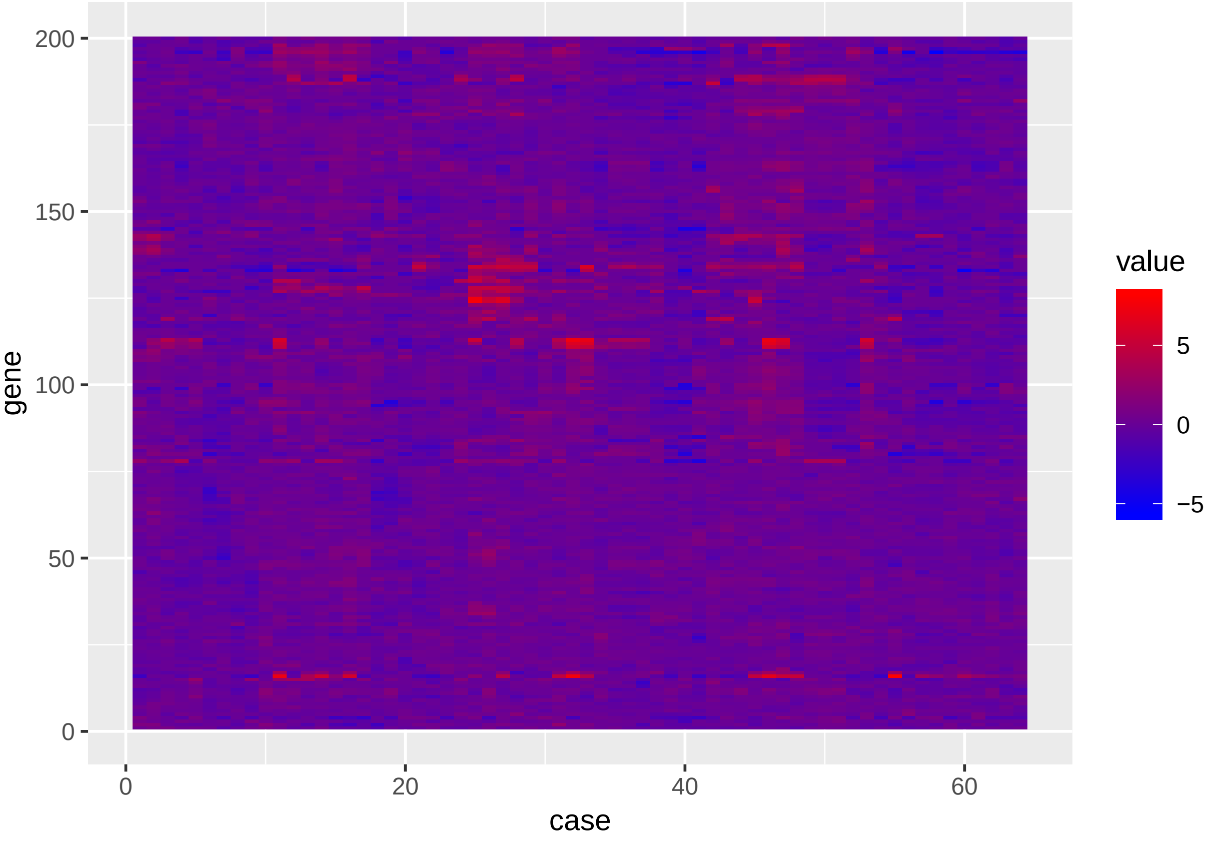
Notes:

```r
nci <- read.table("/home/reshu/Desktop/eda/lesson5/nci.tsv")
colnames(nci) <- c(1:64)
```

```r
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
```

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



Analyzing Three of More Variables

Reflection:

Click **KnitHTML** to see all of your hard work and to have an html page of this lesson, your answers, and your notes!