Lesson 5

Multivariate Data

Notes:

Moira Perceived Audience Size Colored by Age

Notes:

Third Qualitative Variable

Notes:

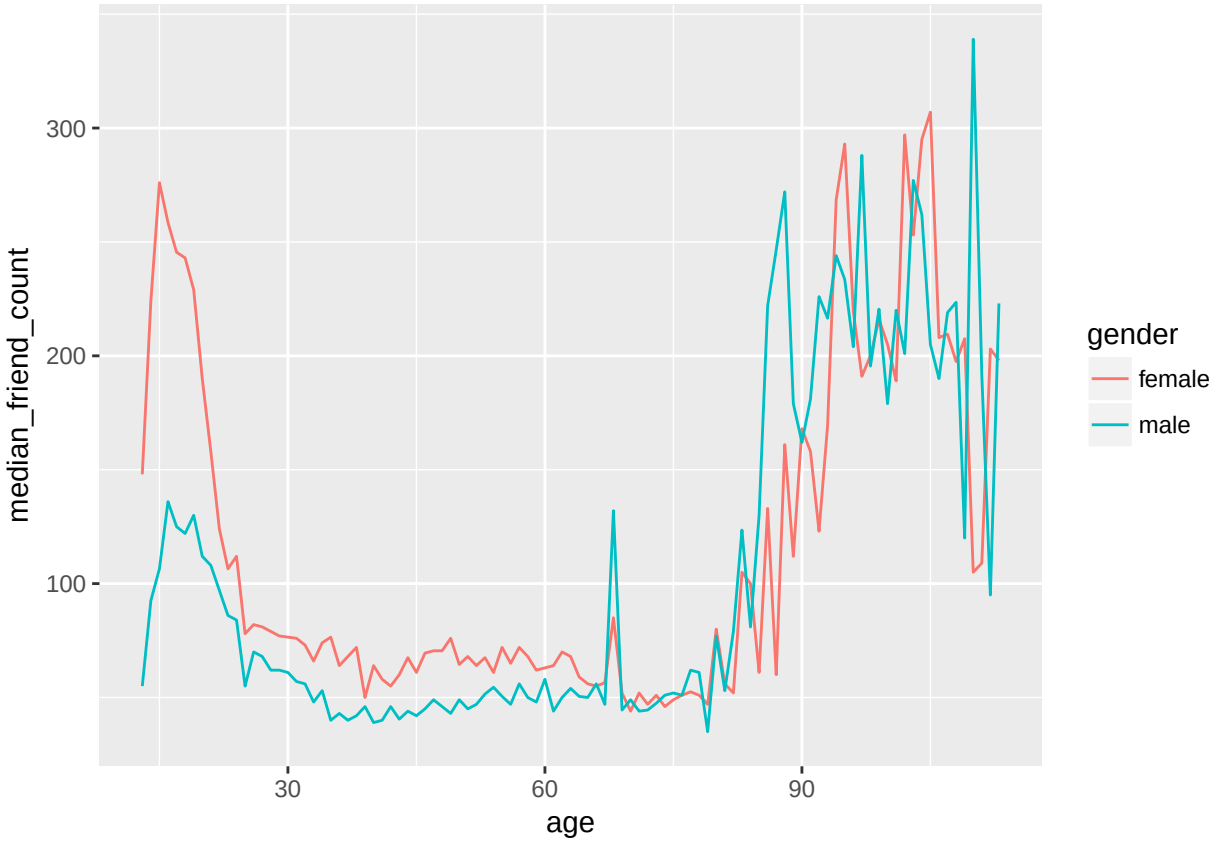
```
pf = read.csv('pseudo_facebook.tsv', sep = '\t')
pf.fc_by_age_gender <- pf %>%
  filter(!is.na(gender)) %>%
  group_by(age, gender) %>%
  summarize(mean_friend_count = mean(friend_count),
            median_friend_count = median(friend_count),
            n = n()) %>%
  ungroup() %>%
  arrange(age)

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.                 55.0   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
```

Plotting Conditional Summaries

Notes:

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



Thinking in Ratios

Notes:

What is the ratio of friends for males vs females

Wide and Long Format

Notes:

Reshaping Data

Notes:

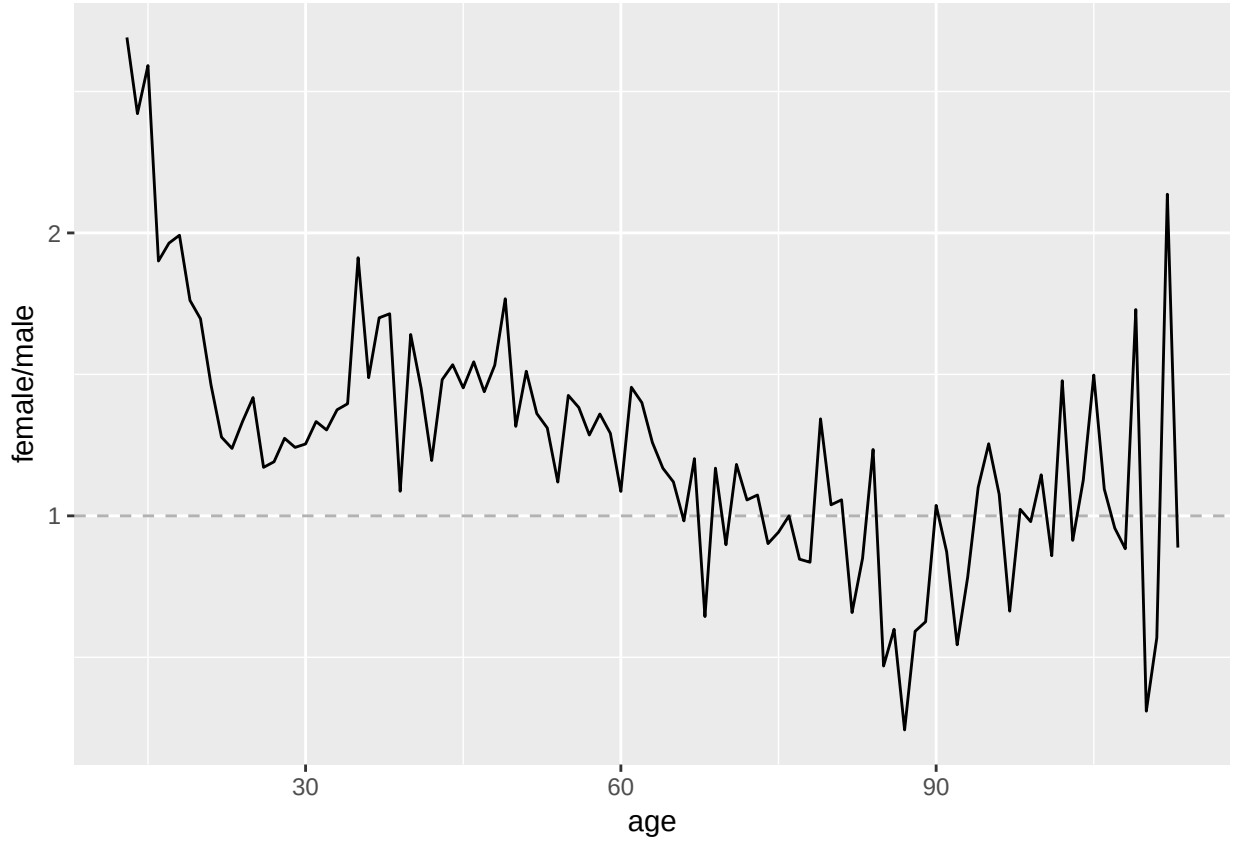
```
#install.packages('reshape2')
pf.fc_by_age_gender.wide <- dcast(pf.fc_by_age_gender,
                                  age ~ gender,
                                  value.var = 'median_friend_count')

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
```

Ratio Plot

Notes:

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



Third Quantitative Variable

Notes:

```
pf$year_joined <- floor(2014 - pf$tenure/365)
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 year_joined
## 1                     0         0                  0        2013
## 2                     0         0                  0        2013
## 3                     0         0                  0        2013
## 4                     0         0                  0        2013
## 5                     0         0                  0        2013
## 6                     0         0                  0        2013
```

Cut a Variable

Notes:

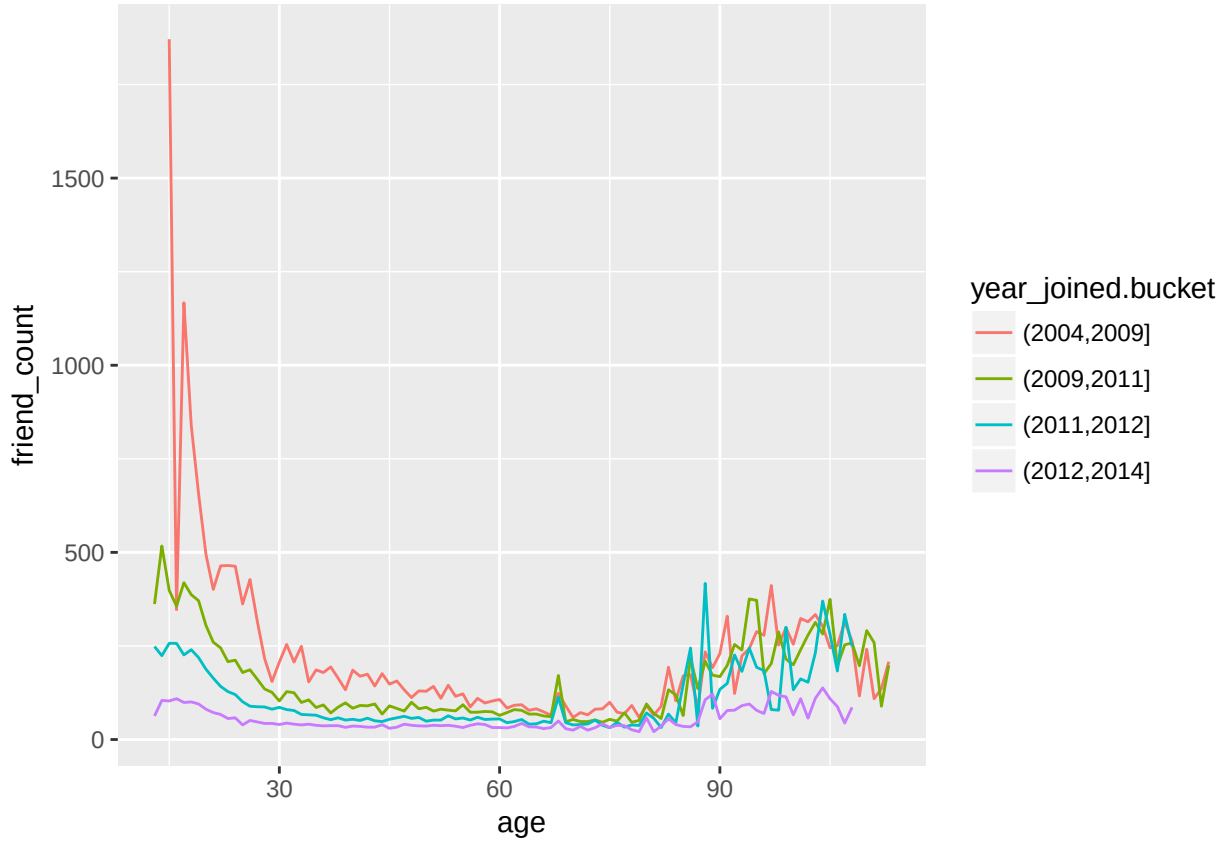
```
pf$year_joined.bucket = cut(pf$year_joined, c(2004, 2009, 2011, 2012, 2014))
table(pf$year_joined.bucket)
```

```
##
## (2004,2009] (2009,2011] (2011,2012] (2012,2014]
##        6669       15308       33366       43658
```

Plotting it All Together

Notes:

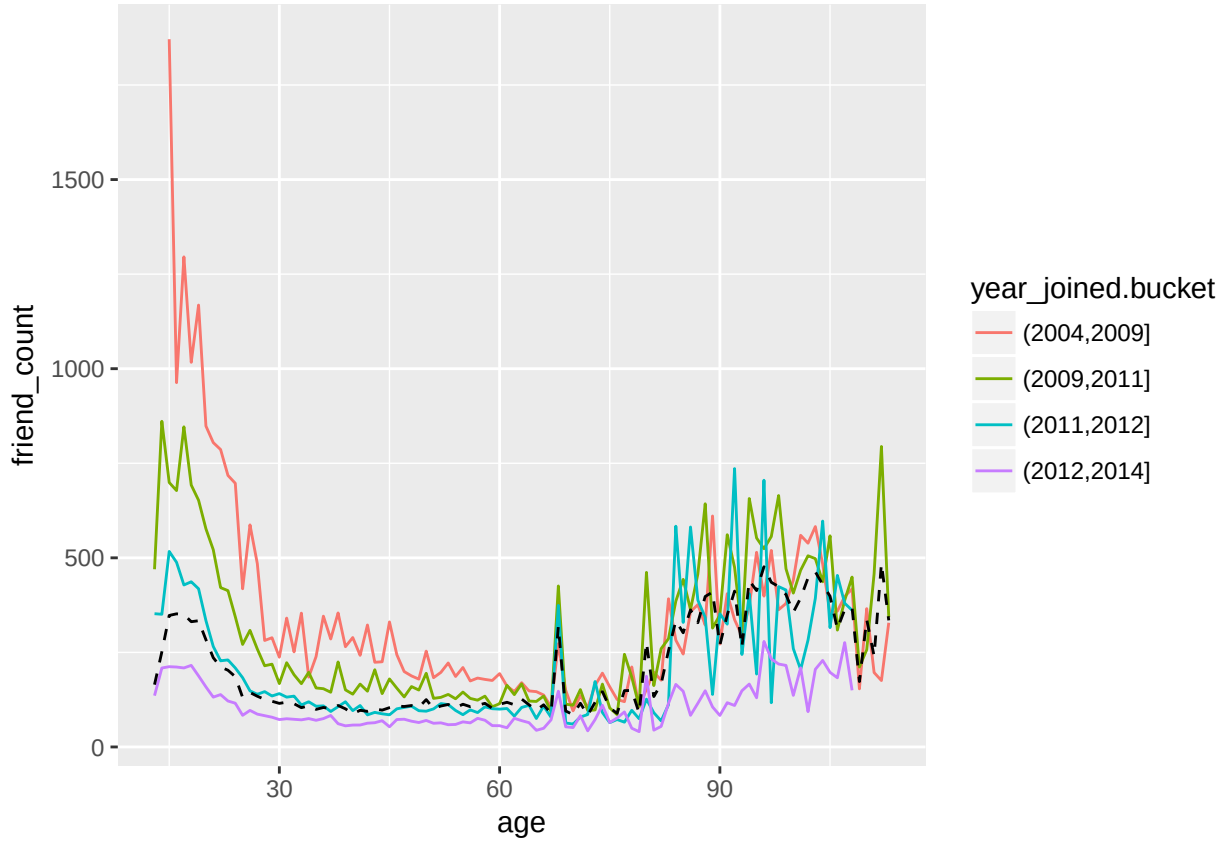
```
ggplot(aes(x = age, y = friend_count), data = subset(pf, !is.na(year_joined.bucket))) +
  geom_line(aes(color = year_joined.bucket), stat='summary', fun.y = median)
```



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='summary', fun.y = mean) +
  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
```

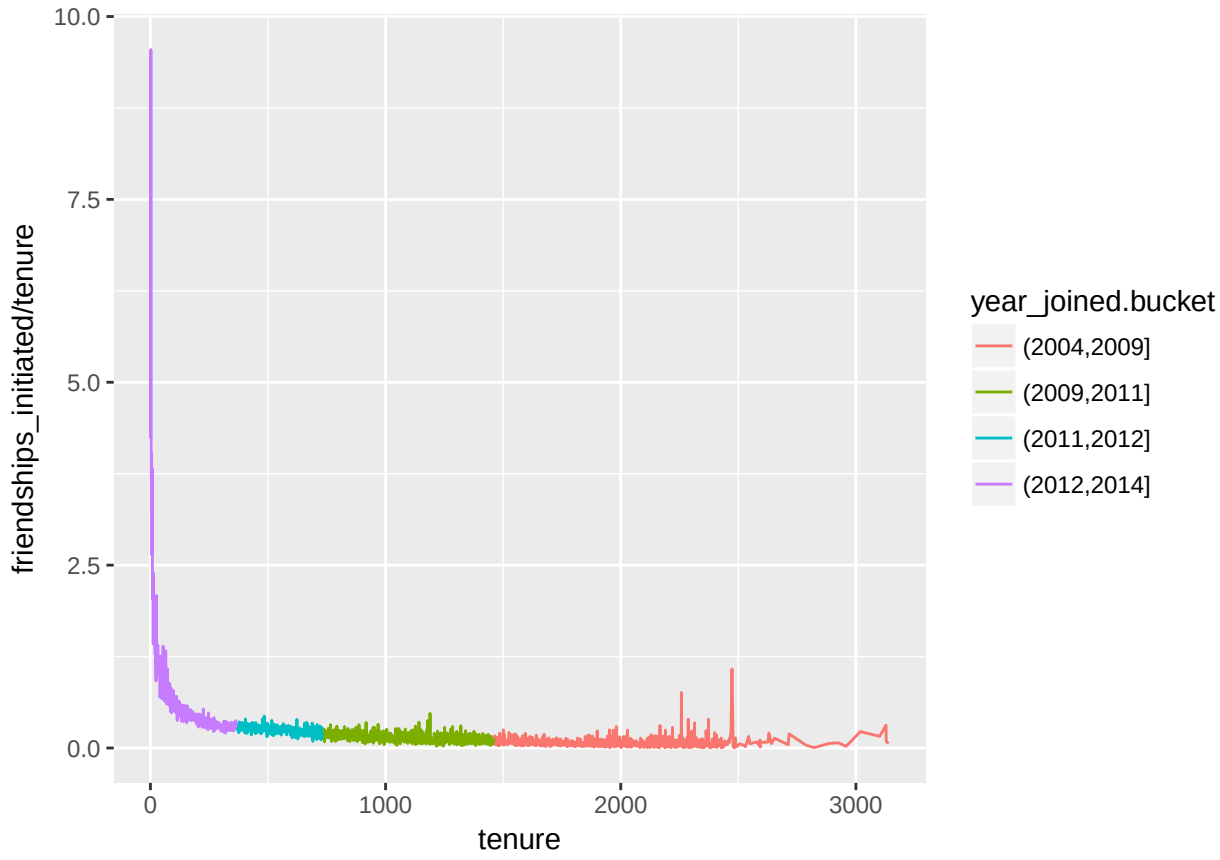
Friendships Initiated

Notes:

What is the median friend rate? 0.2205

What is the maximum friend rate? 417.0

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



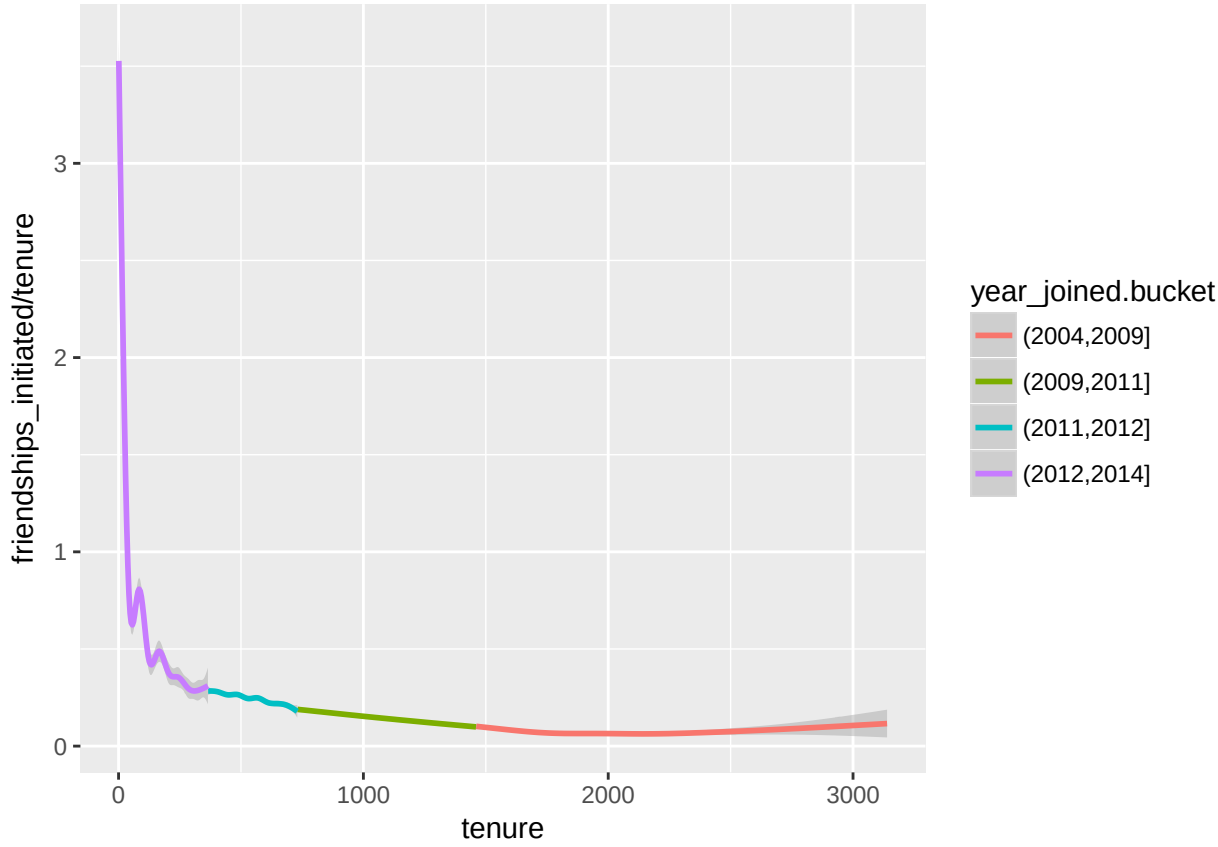
Bias-Variance Tradeoff Revisited

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)
#
#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)

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

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



Sean's NFL Fan Sentiment Study

Notes:

Introducing the Yogurt Data Set

Notes:

Histograms Revisited

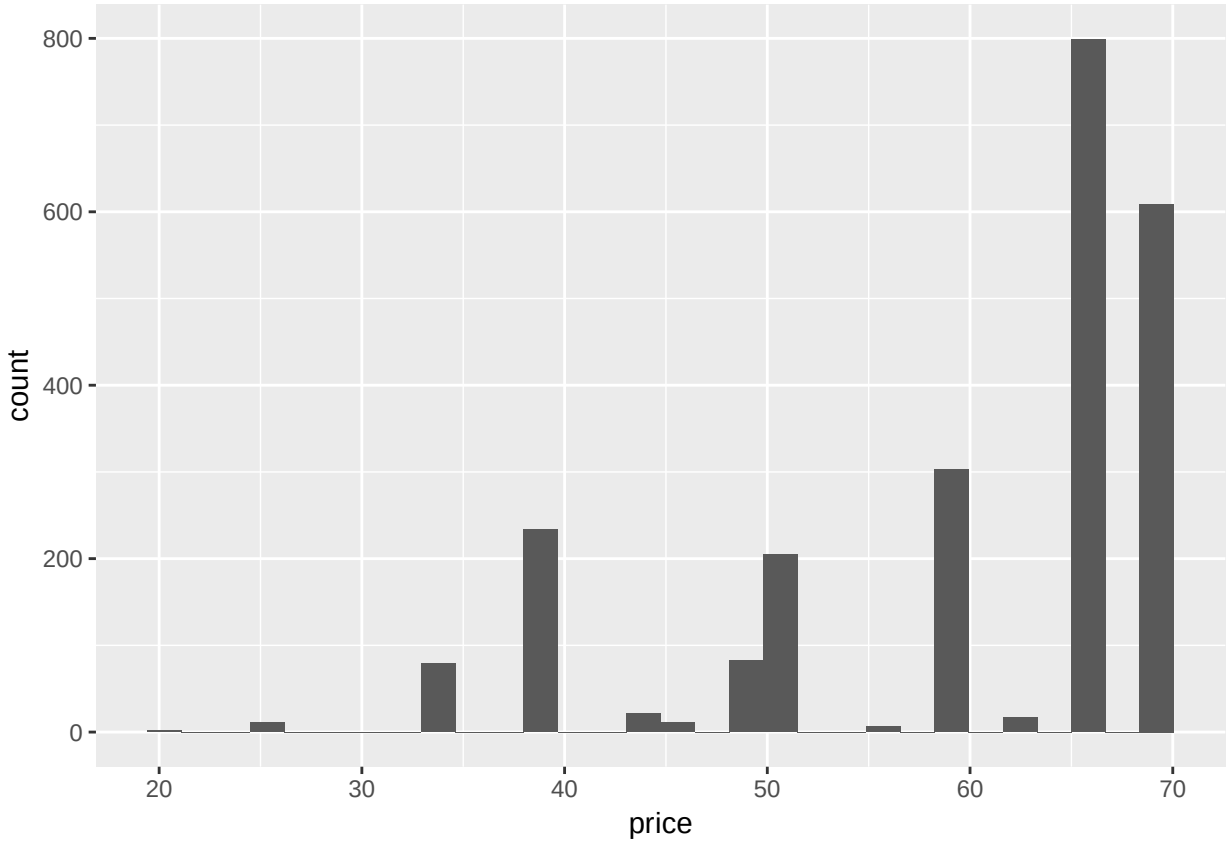
Notes:

```
yo <- read.csv('yogurt.csv')

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

```
##   obs      id  time strawberry blueberry pina.colada plain mixed.berry
## 1   1 2100081  9678          0         0           0     0           1
## 2   2 2100081  9697          0         0           0     0           1
## 3   3 2100081  9825          0         0           0     0           1
## 4   4 2100081  9999          0         0           0     0           1
## 5   5 2100081 10015          1         0           1     0           1
## 6   6 2100081 10029          1         0           2     0           1
##   price
## 1 58.96
## 2 58.96
## 3 65.04
## 4 65.04
## 5 48.96
## 6 65.04
```

```
ggplot(aes(x = price), data = yo) +
  geom_histogram()
```

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



Number of Purchases

Notes:

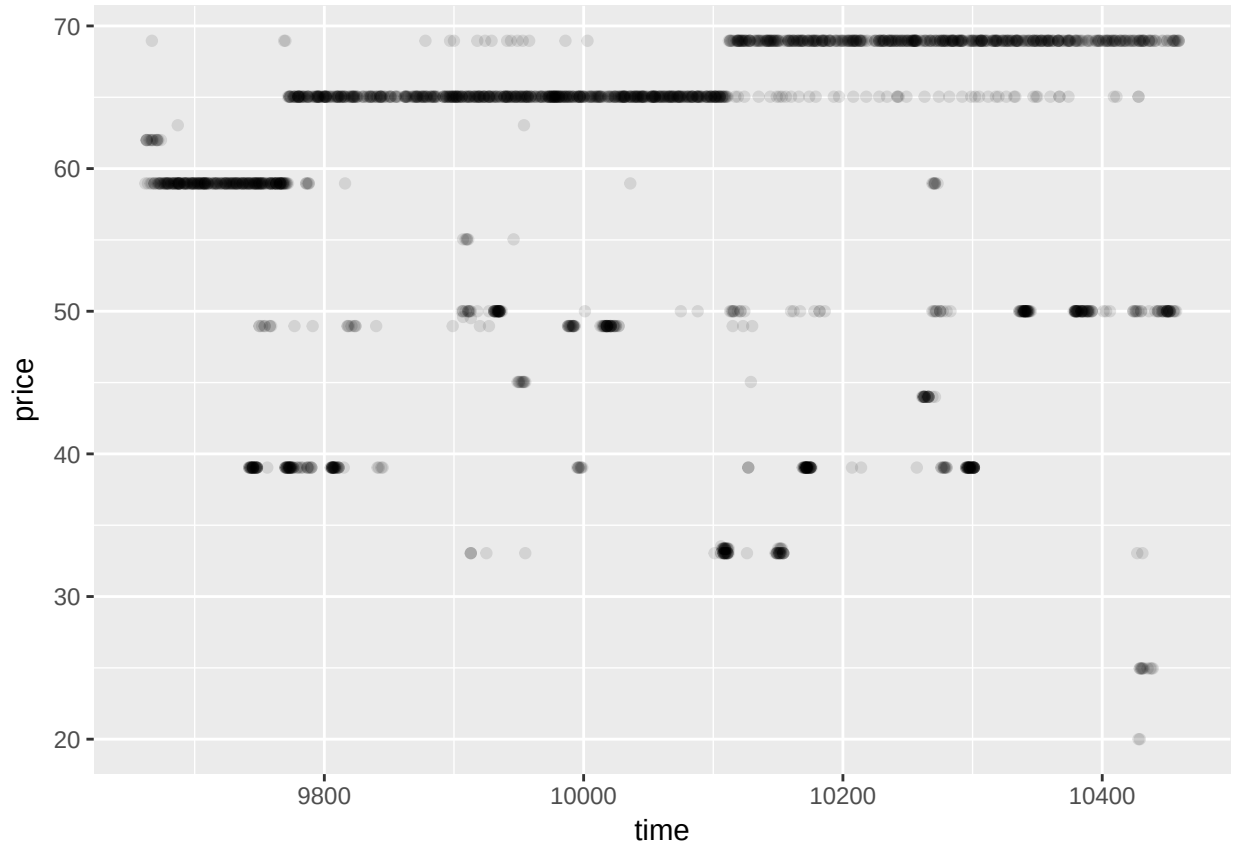
```
yo <- transform(yo, all.purchases = strawberry + blueberry + pina.colada + plain + mixed.berry)
head(yo)
```

```
##   obs      id  time strawberry blueberry pina.colada plain mixed.berry
## 1   1 2100081  9678          0         0           0     0           1
## 2   2 2100081  9697          0         0           0     0           1
## 3   3 2100081  9825          0         0           0     0           1
## 4   4 2100081  9999          0         0           0     0           1
## 5   5 2100081 10015          1         0           1     0           1
## 6   6 2100081 10029          1         0           2     0           1
##   price all.purchases
## 1 58.96             1
## 2 58.96             1
## 3 65.04             1
## 4 65.04             1
## 5 48.96             3
## 6 65.04             4
```

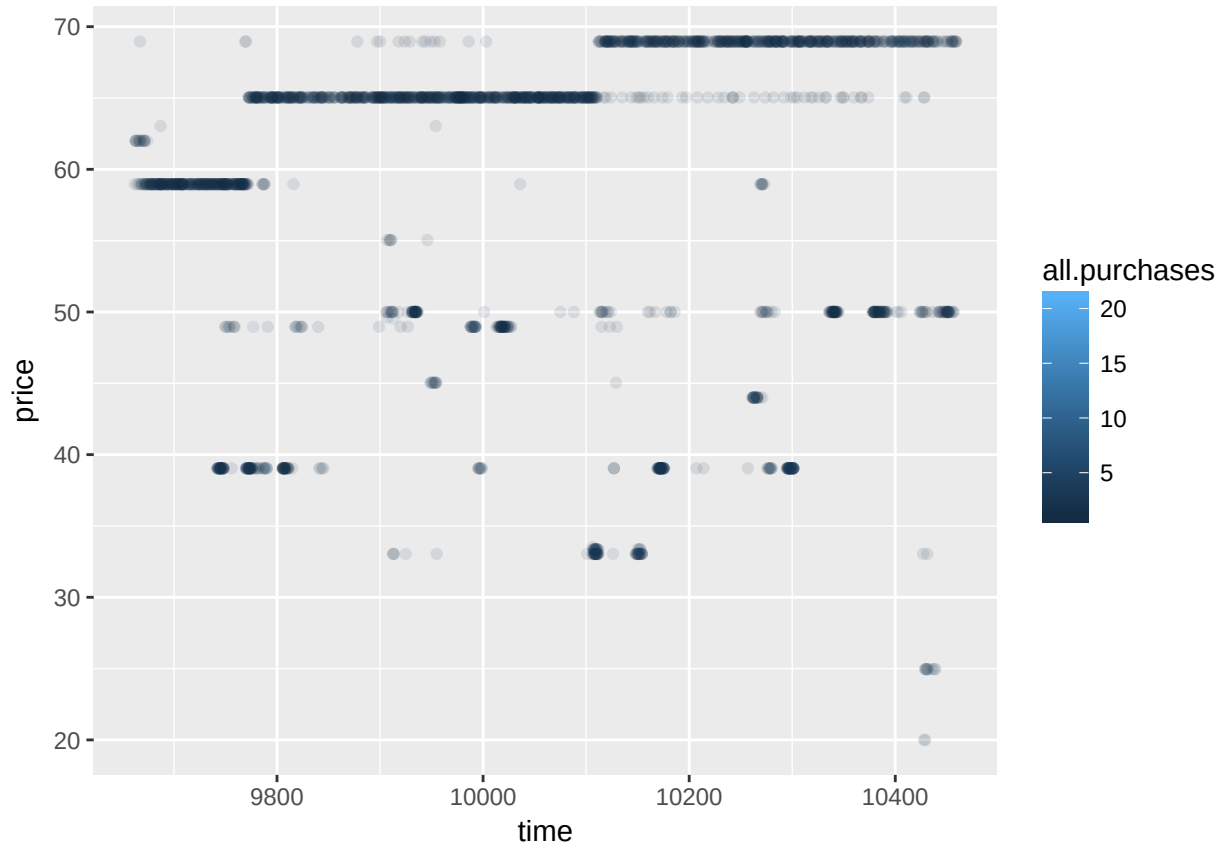
Prices over Time

Notes:

```
ggplot(aes(x = time, y = price), data = yo) +
  geom_point(alpha = 1/10)
```




```
ggplot(aes(x = time, y = price), data = yo) +
  geom_point(alpha = 1/10, aes(color = all.purchases))
```

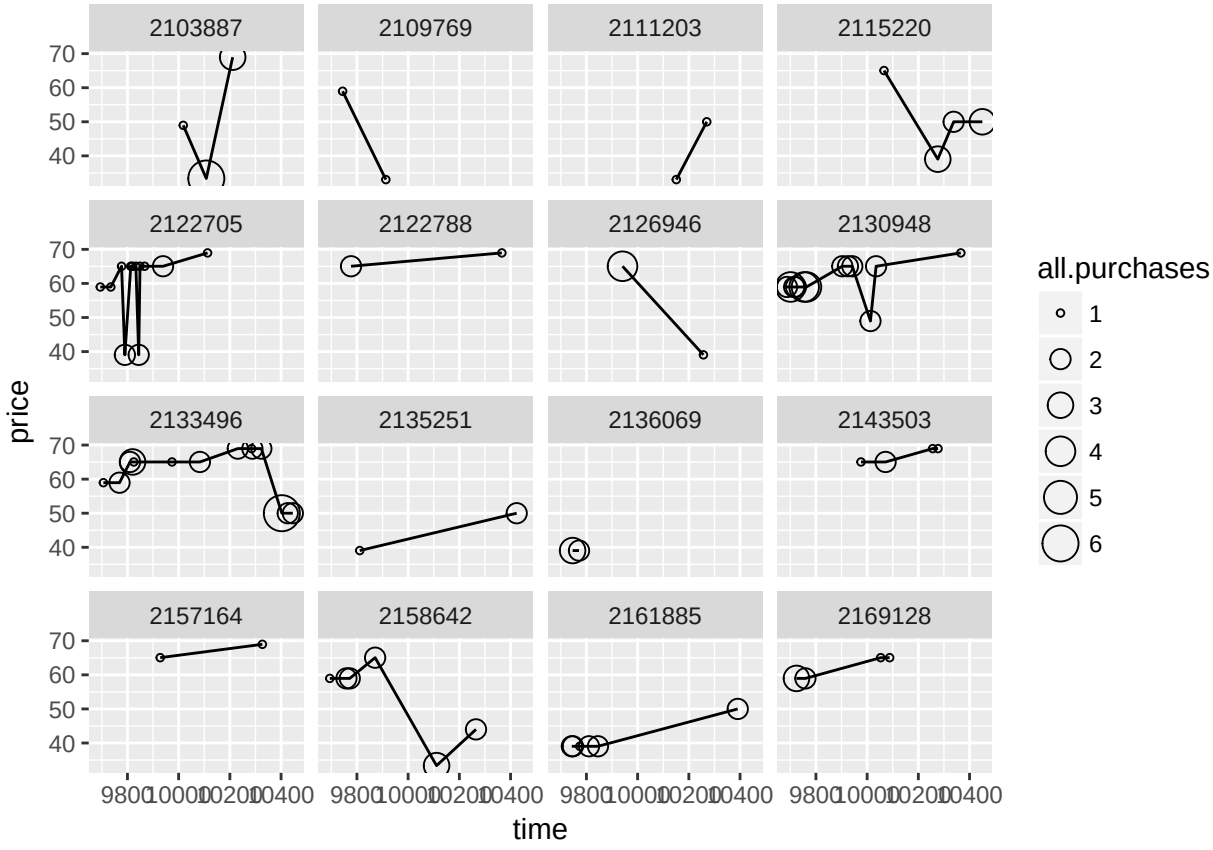


Sampling Observations

Notes:

Looking at Samples of Households

```
set.seed(1)
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:

Scatterplot Matrix

Notes:

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

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
```

```
##     nasa
```

```
theme_set(theme_minimal(20))

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"
```

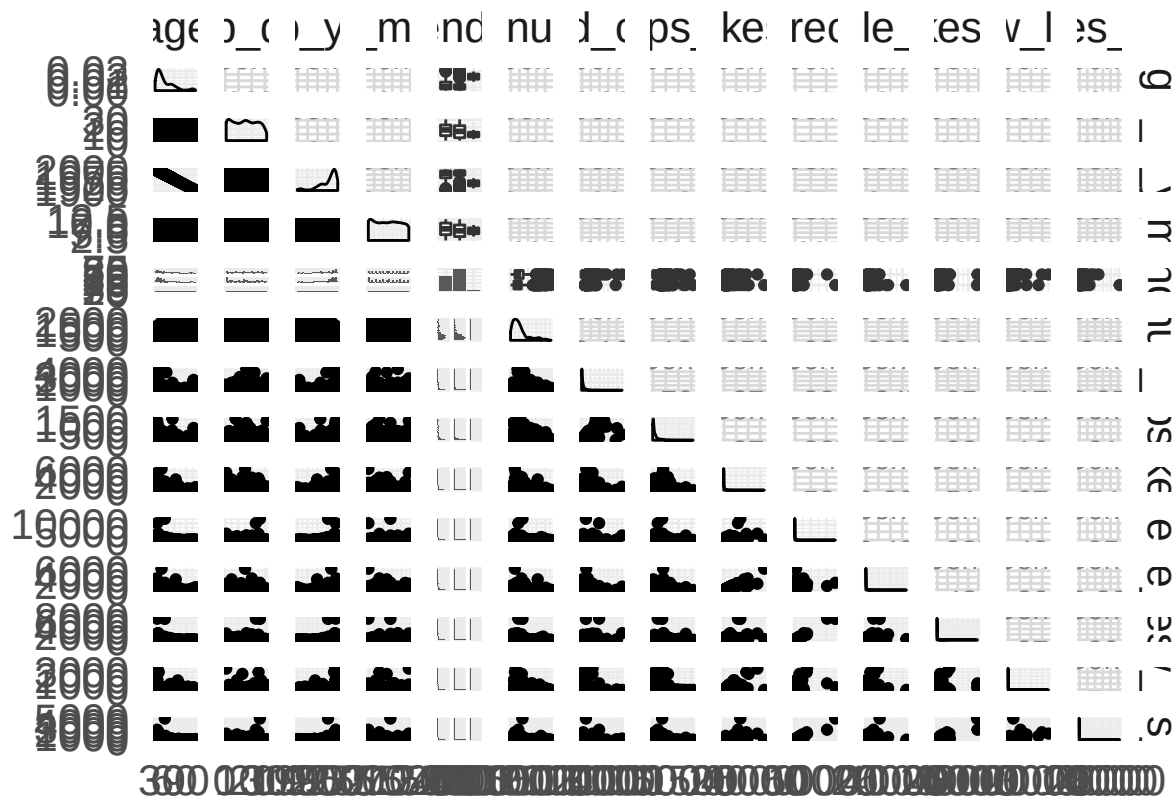
```
ggpairs(pf_subset[sample.int(nrow(pf_subset), 1000), ])
```

```
## `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).
```

```
## `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`.
```



Even More Variables

Notes:

Heat Maps

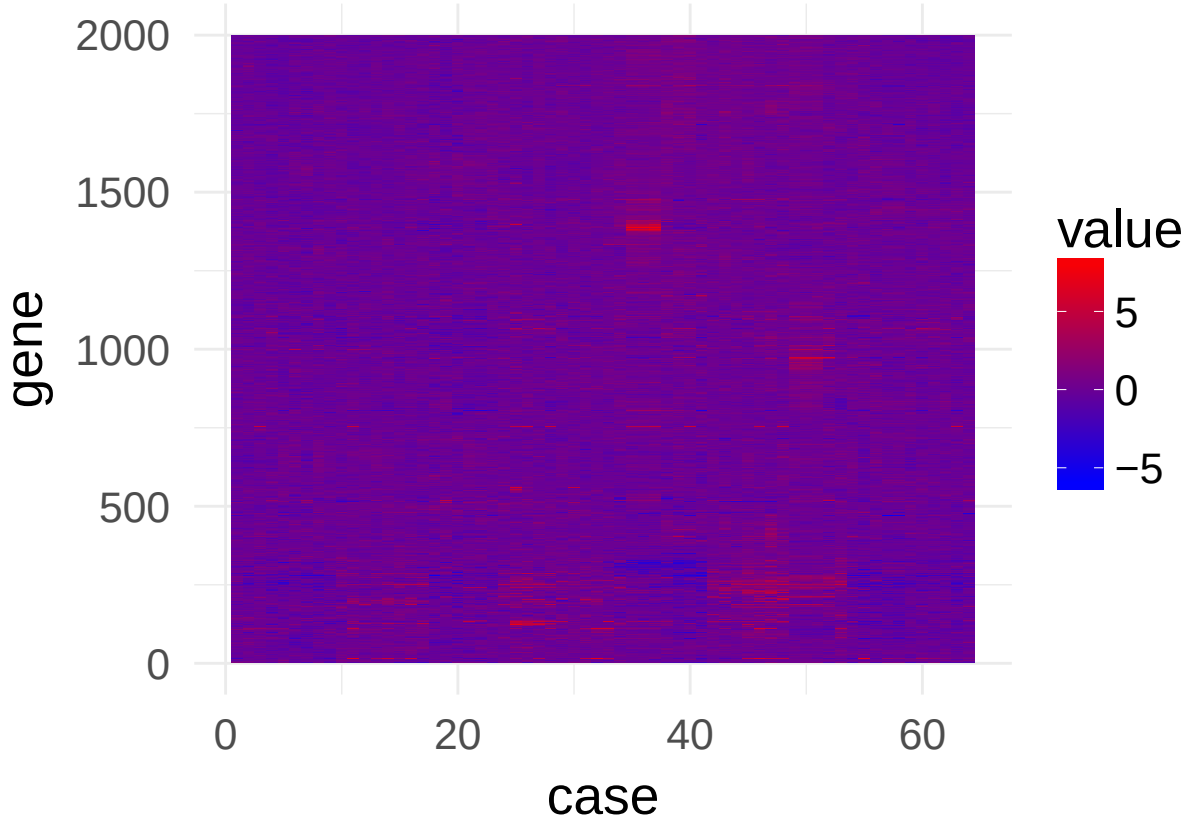
Notes:

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

```
nci.long.samp <- melt(as.matrix(nci[1:2000,]))
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))
```



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!