Lesson 3

pf <- read.csv('pseudo_facebook.tsv', sep = '')

names(pf)

What to Do First?

Notes:

Pseudo-Facebook User Data

Notes:

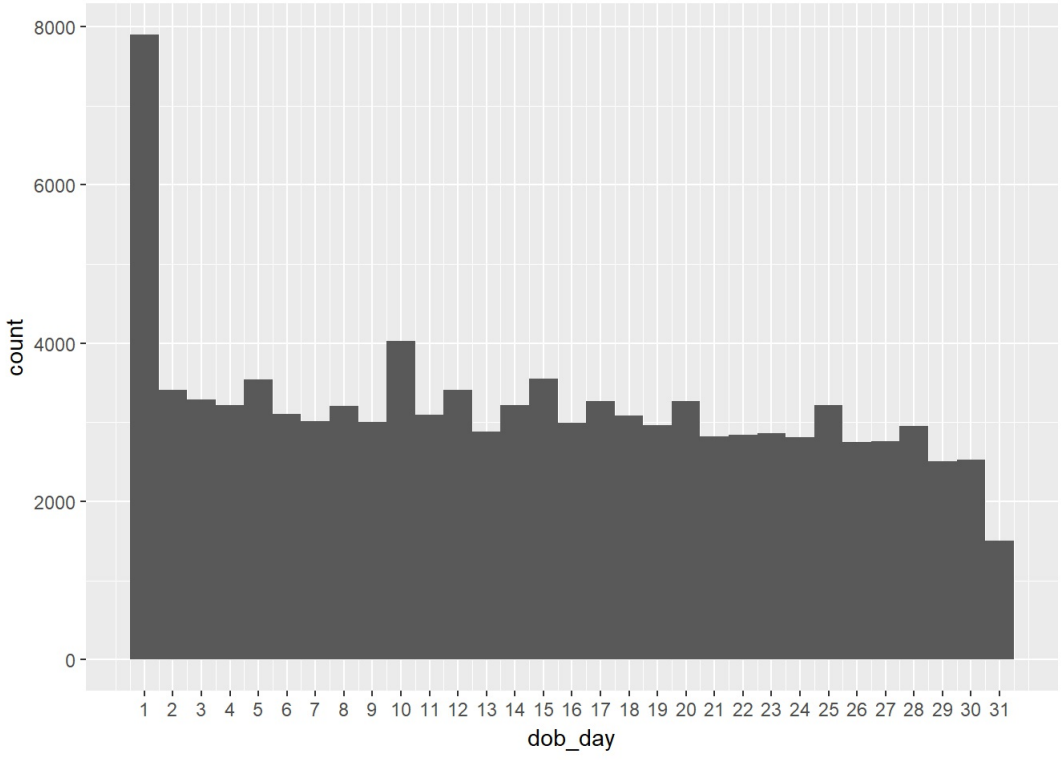
```
pf <- read.csv('pseudo_facebook.tsv', sep = '\t')

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"
```

Histogram of Users' Birthdays

Notes:

```
library(ggplot2)
ggplot(aes(x = dob_day), data = pf) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = 1:31)
```



What are some things that you notice about this histogram?

Response: Day 1 was really high, the 31st was a bit low but all other days were evenly distributed.

Moira's Investigation

Notes: She was checking to see how big of an audience people thought they had for a FB post. ***

Estimating Your Audience Size

Notes: I no longer post to FB. ***

Think about a time when you posted a specific message or shared a photo on Facebook. What was it?

Response: I don't remember #### How many of your friends do you think saw that post? Response: I don't know #### Think about what percent of your friends on Facebook see any posts or comments that you make in a month. What percent do you think that is? Response: No idea. ***

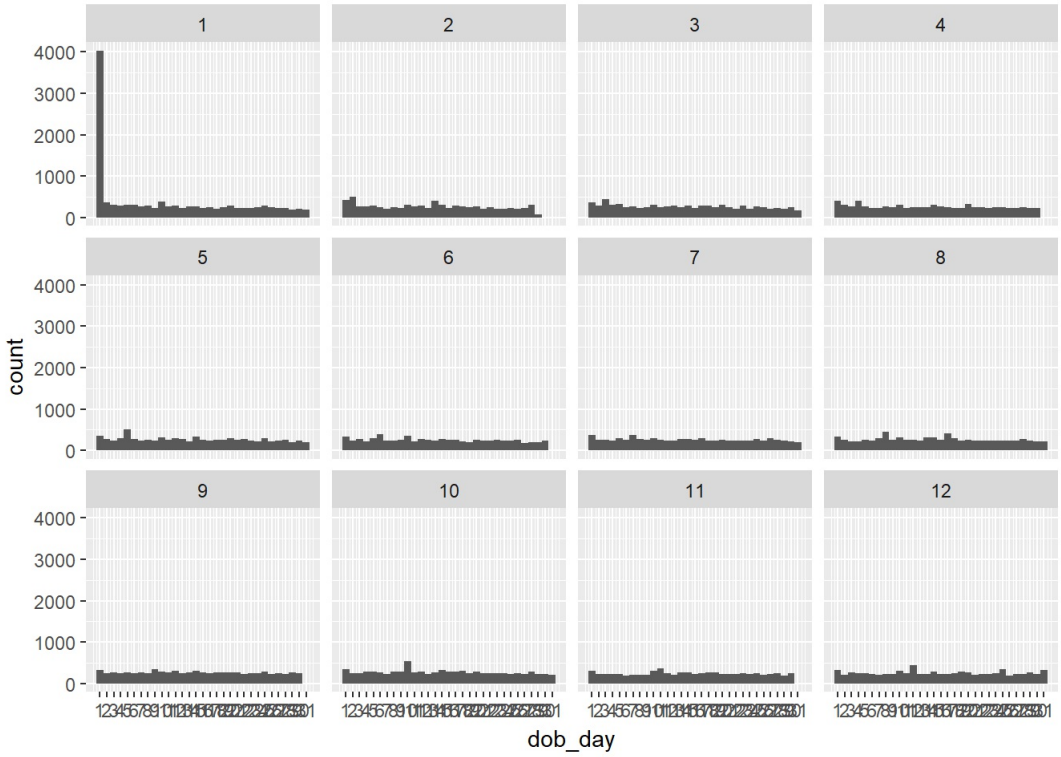
Perceived Audience Size

Notes:

Faceting

Notes: facet_grid(vert ~ horiz) data presentation changes on how the var is set. Use for 2 or more var.

```
ggplot(data = pf, aes(x= dob_day)) +
  geom_histogram(binwidth = 1) +
  scale_x_continuous(breaks = 1:31) +
  facet_wrap(~dob_month)
```



Letâs take another look at our plot. What stands out to you here?

Response: Most entries are for Jan. 1. This could be people accepting defaults. ***

Be Skeptical - Outliers and Anomalies

Notes:

Moira's Outlier

Notes: #### Which case do you think applies to Moiraâs outlier? Response:

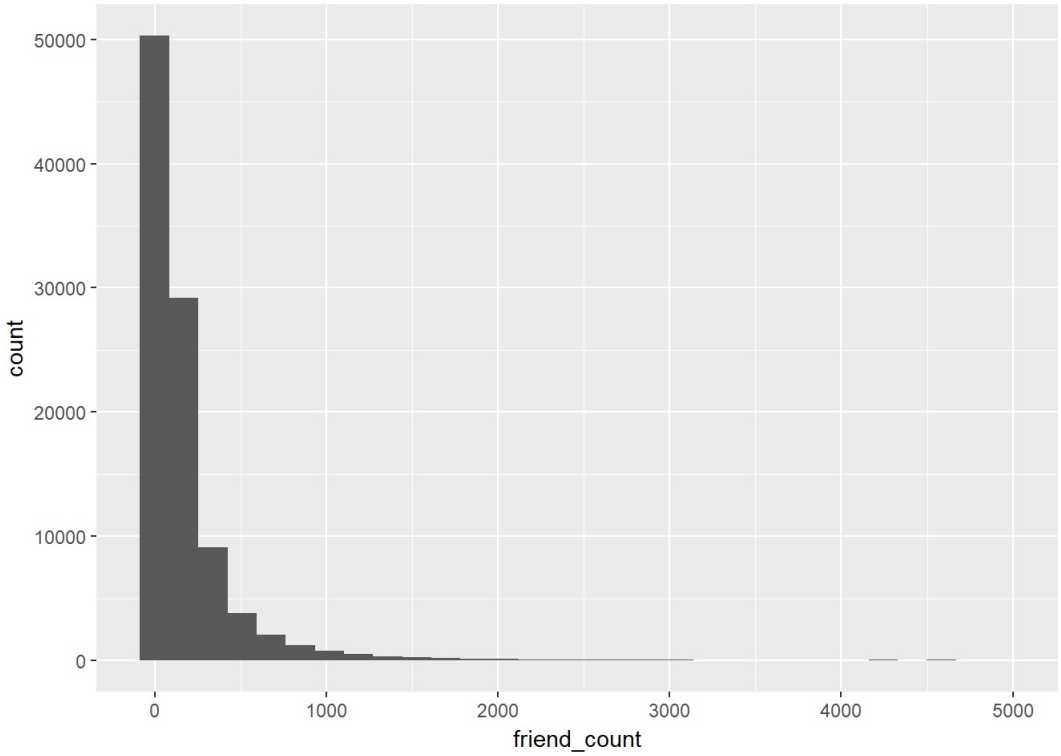
Friend Count

Notes:

What code would you enter to create a histogram of friend counts?

```
library(ggplot2)
ggplot( data = pf, aes(x = friend_count)) +
  geom_histogram()
```

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



How is this plot similar to Moira's first plot?

Response:

Limiting the Axes

Notes:

Exploring with Bin Width

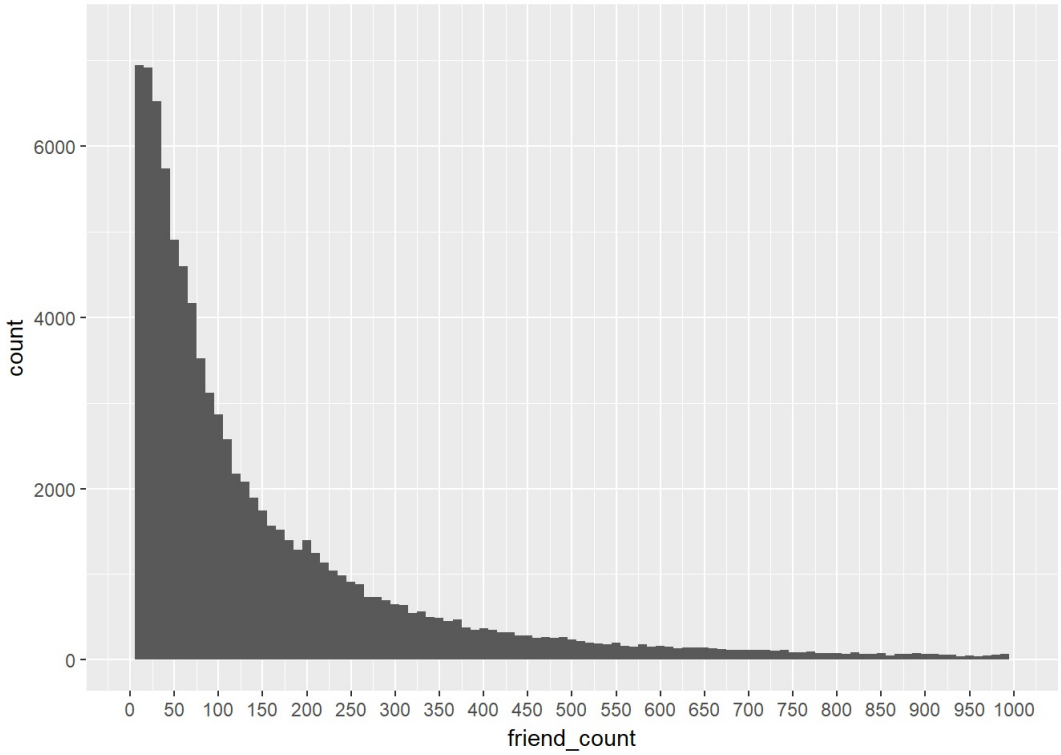
Notes:

Adjusting the Bin Width

Notes: breaks(start, end, interval) ### Faceting Friend Count

```
# What code would you add to create a facet the histogram by gender?
# Add it to the code below.
qplot(x = friend_count, data = pf, binwidth = 10) +
  scale_x_continuous(limits = c(0, 1000),
                     breaks = seq(0, 1000, 50))
```

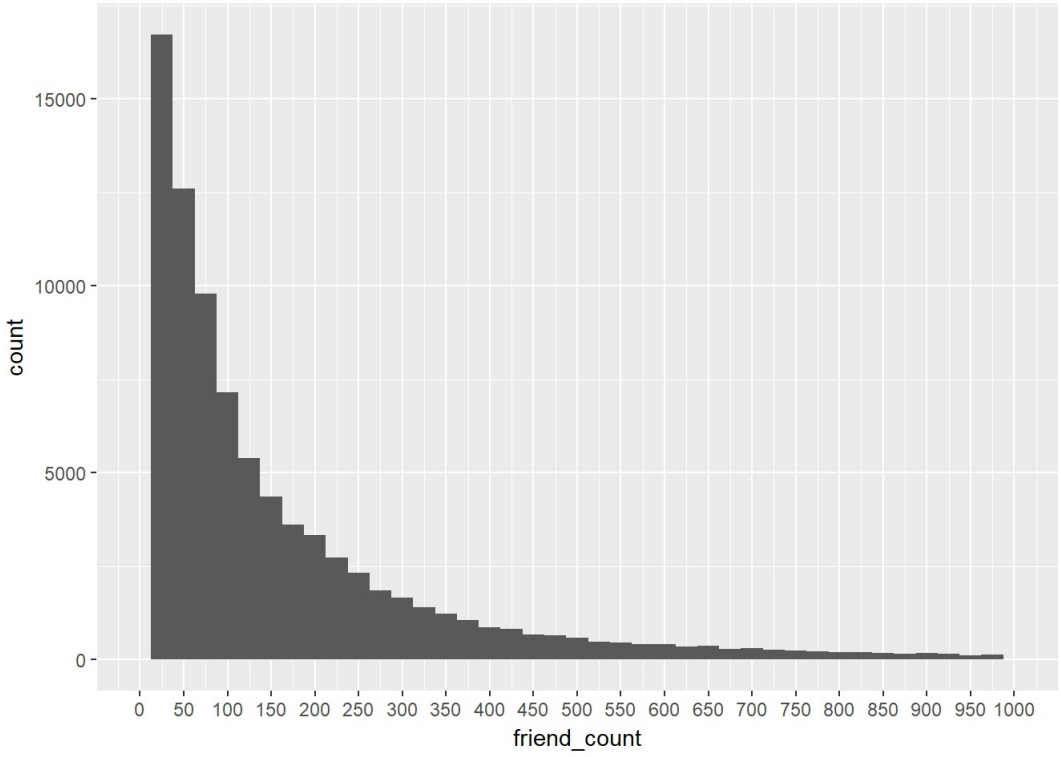
```
## Warning: Removed 2951 rows containing non-finite values (stat_bin).
```



ggplot equivalent

```
ggplot(aes(x = friend_count), data = pf) +
  geom_histogram(binwidth = 25) +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50))
```

```
## Warning: Removed 2951 rows containing non-finite values (stat_bin).
```

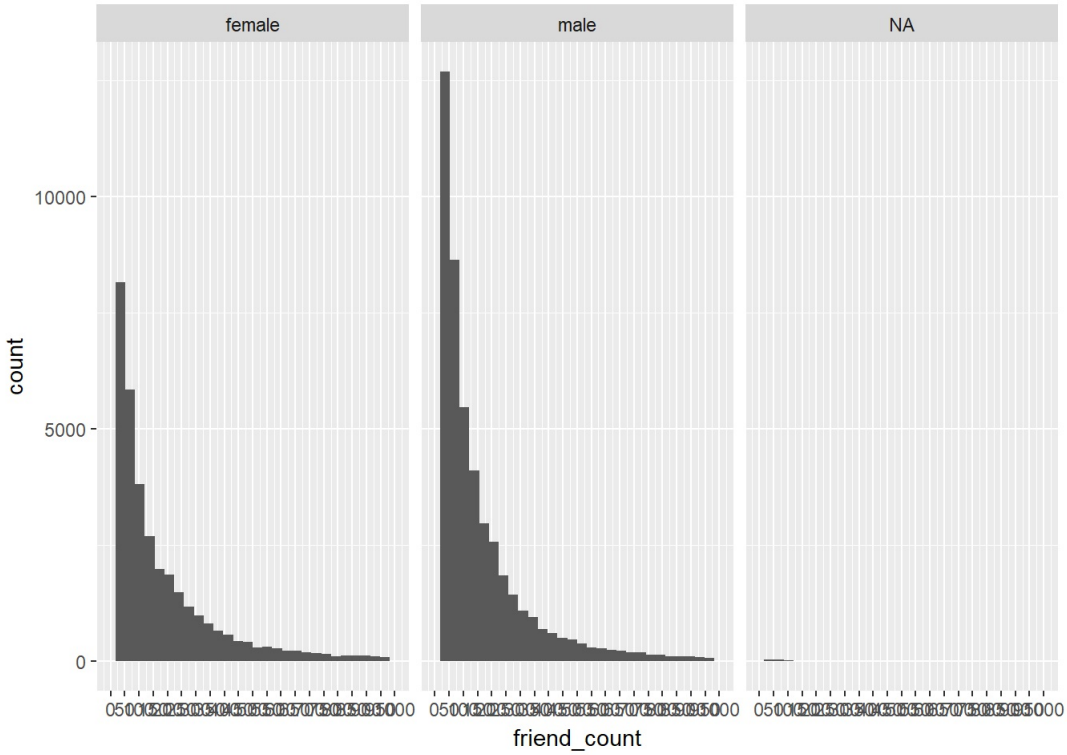


*** ### ggplot with gender var

```
ggplot(aes(x = friend_count), data = pf) +
  geom_histogram() +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50)) +
  facet_wrap(~gender)
```

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

```
## Warning: Removed 2951 rows containing non-finite values (stat_bin).
```



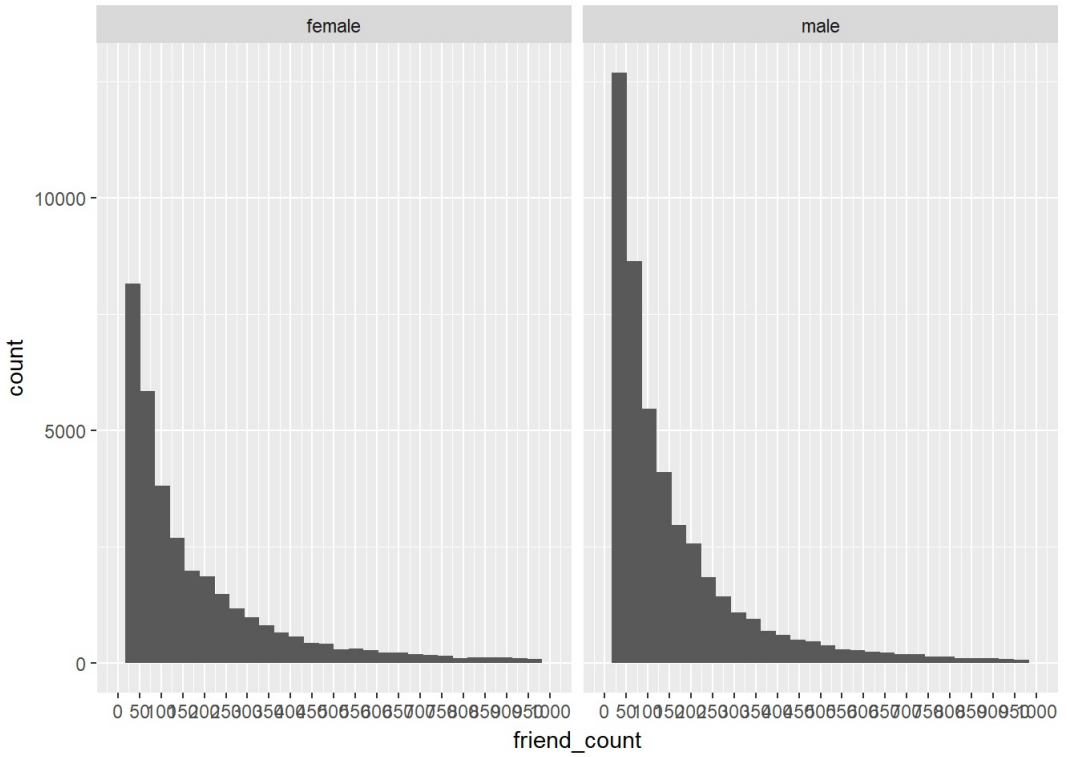
*** ### Omitting NA Values Notes: R

takes data that does not meet criteria and sets it to NA.

```
ggplot(aes(x = friend_count), data = subset(pf, !is.na(gender))) +
  geom_histogram() +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50)) +
  facet_wrap(~gender)
```

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

```
## Warning: Removed 2949 rows containing non-finite values (stat_bin).
```



Statistics ‘by’ Gender

Notes:

```
table(pf$gender)
```

```
##
## female   male
##  40254  58574
```

```
by(pf$friend_count, pf$gender, summary)
```

```
## pf$gender: female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0      37      96     242     244    4923
## ------------------------------------------------------------
## pf$gender: male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0      27      74     165     182    4917
```

Who on average has more friends: men or women?

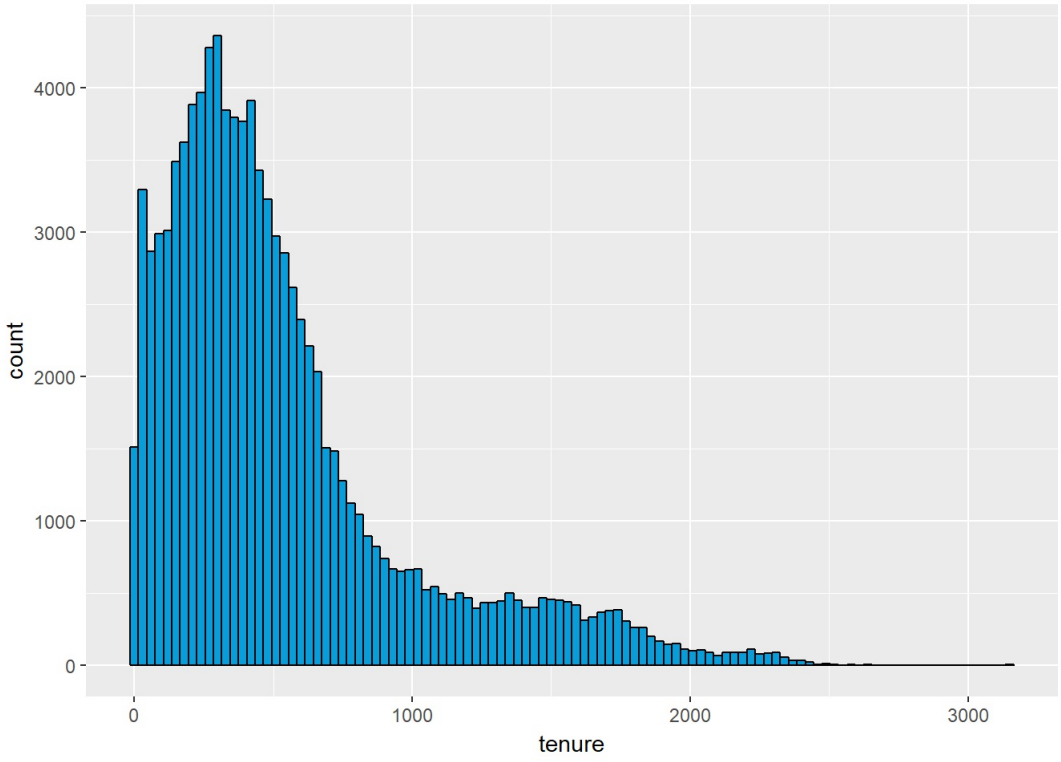
Response: women #### What's the difference between the median friend count for women and men? Response: 22 #### Why would the median be a better measure than the mean? Response: median is the actual middle number while mean is adding it all together and averaging. a few people can skew the mean. ***

Tenure

Notes:

```
ggplot(aes(x = tenure), data = pf) +
  geom_histogram(binwidth = 30, color = 'black', fill = '#099dd9')
```

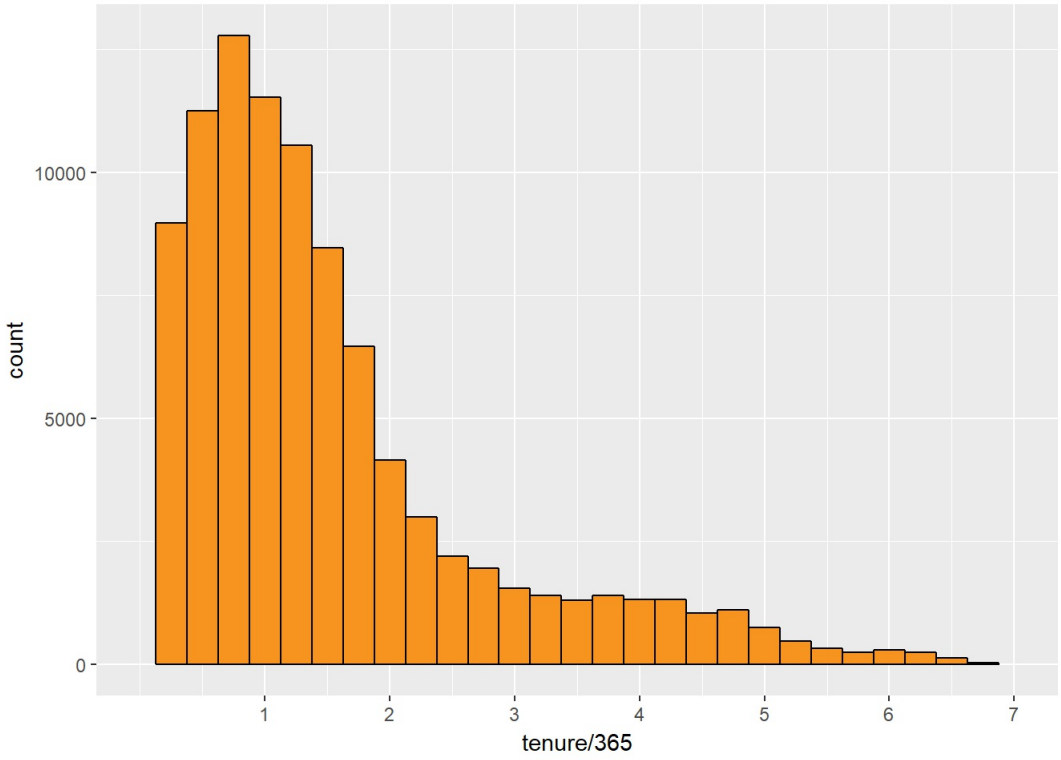
```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



How would you create a histogram of tenure by year?

```
ggplot(aes(x = tenure/365), data = pf) +
  geom_histogram(binwidth = .25, color = 'black', fill = '#f79420') +
  scale_x_continuous(limits = c(0, 7), breaks = seq(1, 7, 1))
```

```
## Warning: Removed 26 rows containing non-finite values (stat_bin).
```

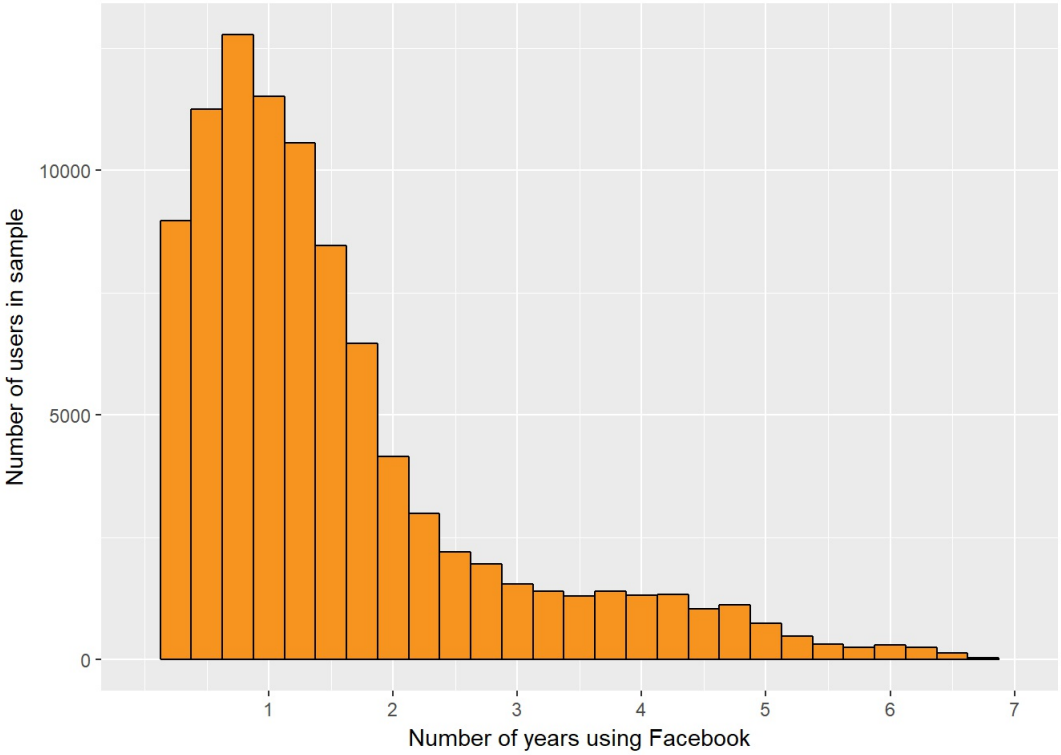


Labeling Plots

Notes:

```
ggplot(aes(x = tenure/365), data = pf) +
  geom_histogram(binwidth = .25, color = 'black', fill = '#f79420') +
  scale_x_continuous(limits = c(0, 7), breaks = seq(1, 7, 1)) +
  xlab('Number of years using Facebook') +
  ylab('Number of users in sample')
```

```
## Warning: Removed 26 rows containing non-finite values (stat_bin).
```

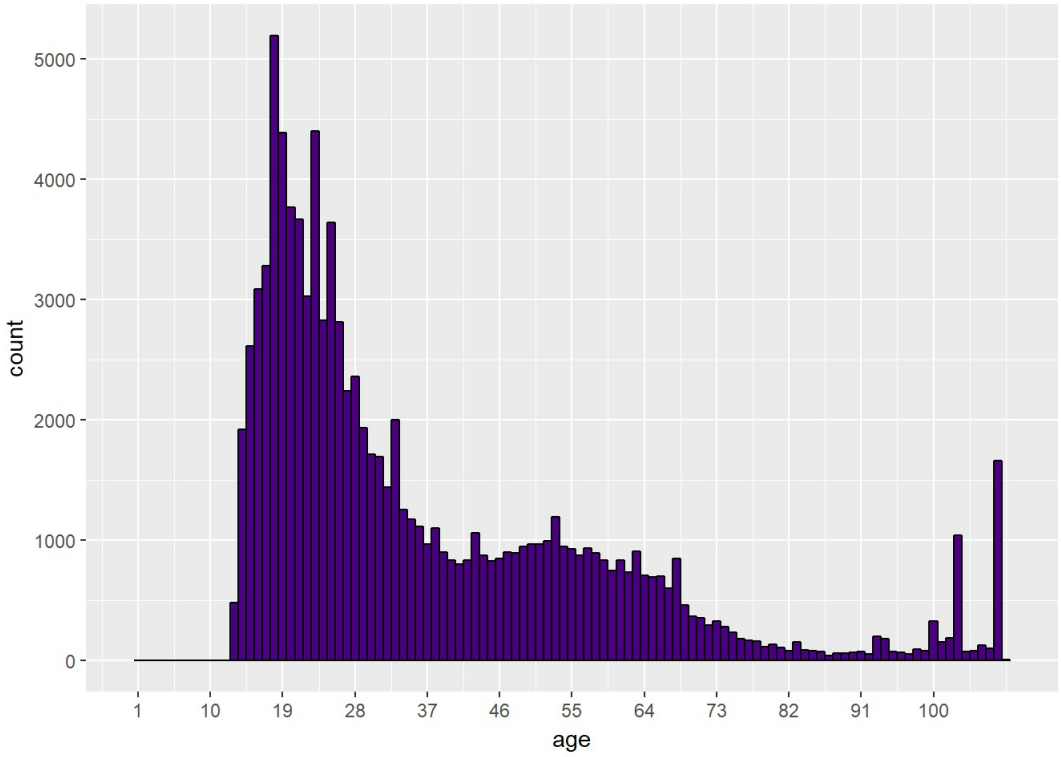


User Ages

Notes:

```
ggplot(aes(x = age), data = pf) +
  geom_histogram(binwidth = 1, color = 'black', fill = '#4b0082') +
  scale_x_continuous(limits = c(0, 110), breaks = seq(1, 100, 9))
```

```
## Warning: Removed 238 rows containing non-finite values (stat_bin).
```



What do you notice?

Response:

The Spread of Memes

Notes:

Lada's Money Bag Meme

Notes:

Transforming Data

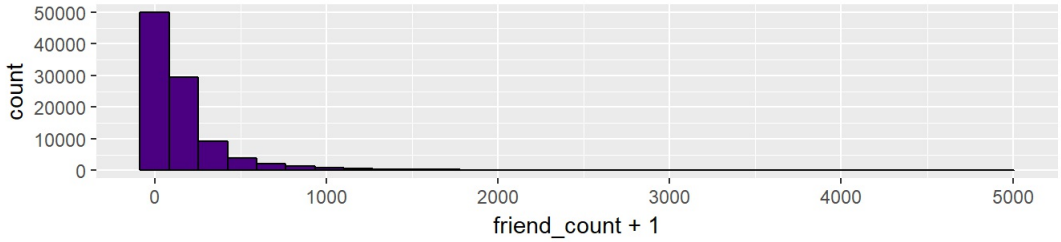
Notes:

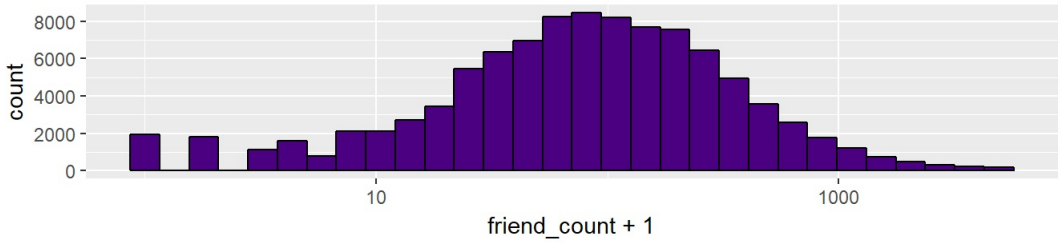
Add a Scaling Layer

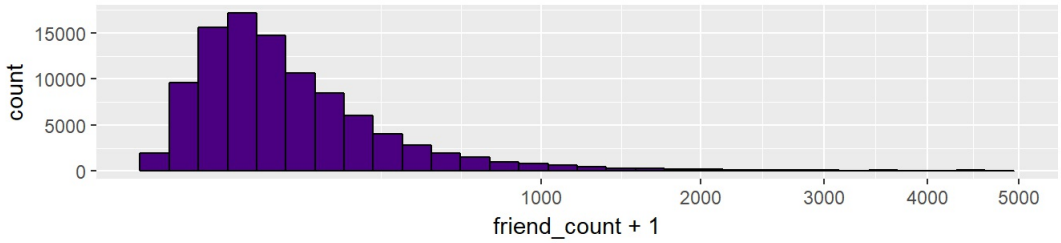
Notes:

```
library(gridExtra)
ap <- ggplot(aes(x = friend_count +1), data = pf) +
  geom_histogram(color = 'black', fill = '#4b0082')
apl <- ap + scale_x_log10()
aps <- ap + scale_x_sqrt()

grid.arrange(ap, apl, aps, ncol = 1)
```

```
## `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`.
```







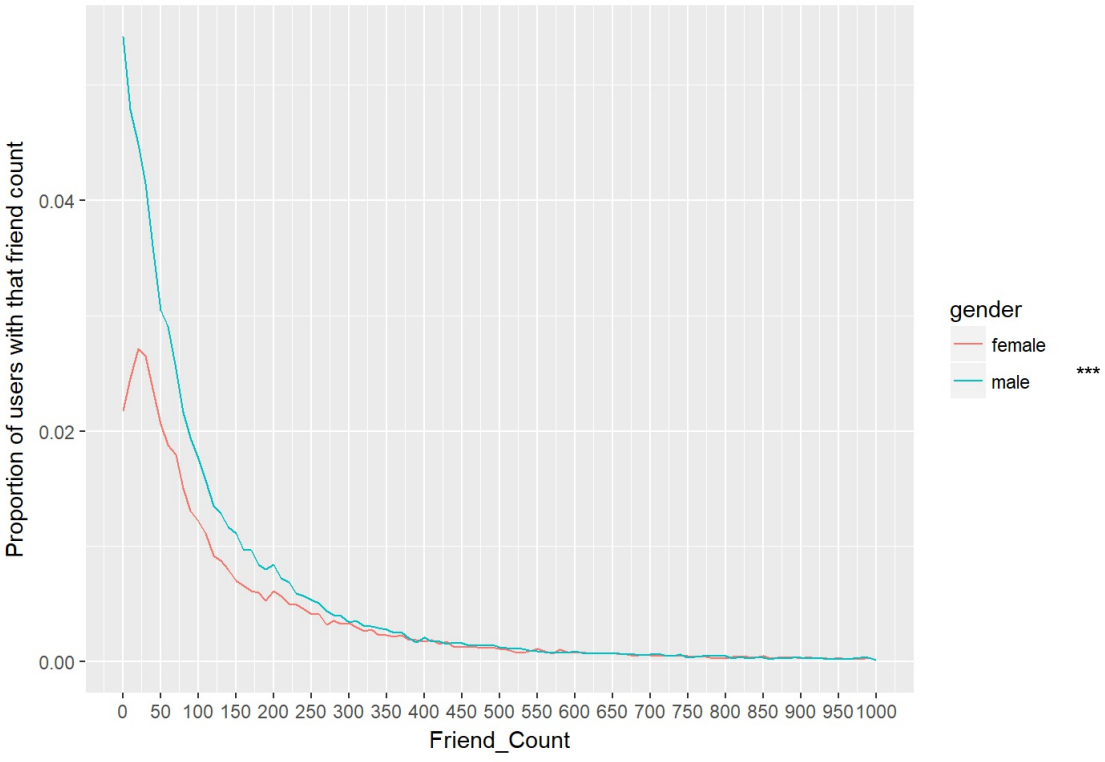
Frequency Polygons

Notes:

```
ggplot(aes(x = friend_count, y = ..count../sum(..count..)), data = subset(pf, !is.na(gender))) +
  geom_freqpoly(aes(color = gender), binwidth = 10) +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50)) +
  xlab('Friend_Count') +
  ylab('Proportion of users with that friend count')
```

```
## Warning: Removed 2949 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 4 rows containing missing values (geom_path).
```

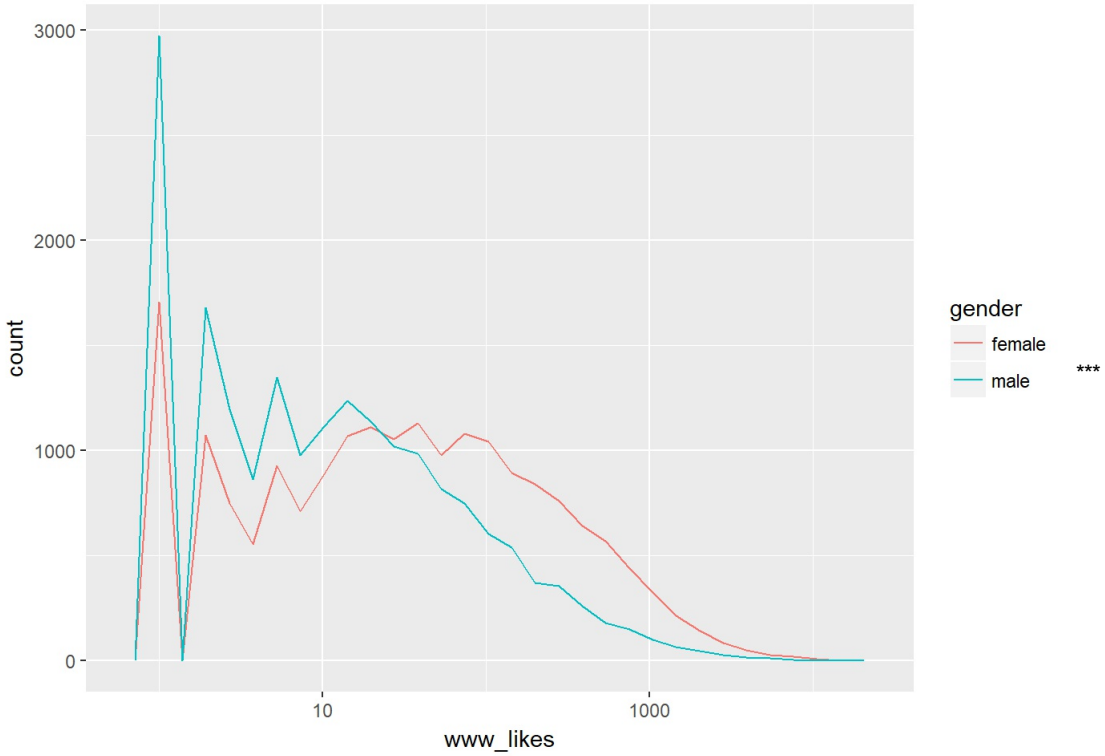


```
ggplot(aes(x = www_likes), data = subset(pf, !is.na(gender))) +
        geom_freqpoly(aes(color = gender)) +
        scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

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

```
## Warning: Removed 60935 rows containing non-finite values (stat_bin).
```



Likes on the Web

Notes:

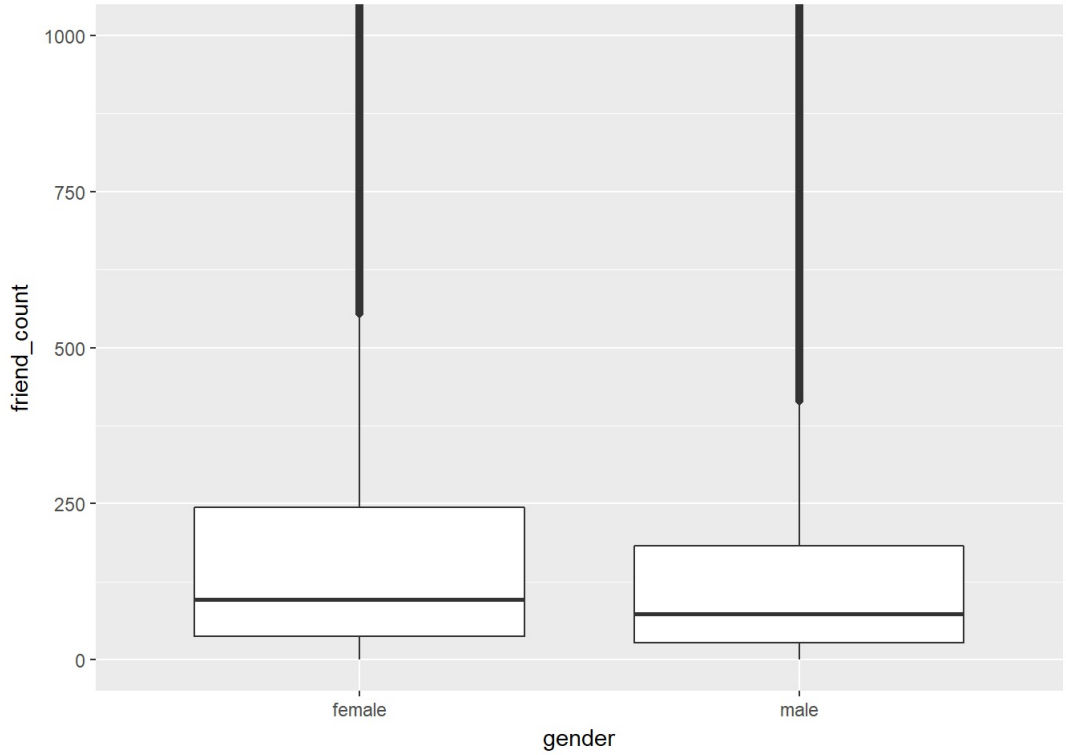
```
by(pf$www_likes, pf$gender, sum)
```

```
## pf$gender: female
## [1] 3507665
## ------------------------------------------------------------
## pf$gender: male
## [1] 1430175
```

Box Plots

Notes:

```
qplot(x = gender, y = friend_count, data = subset(pf, !is.na(gender)), geom = 'boxplot') +
  coord_cartesian(ylim = c(0, 1000))
```



Adjust the code to focus on users who have friend counts between 0 and 1000.

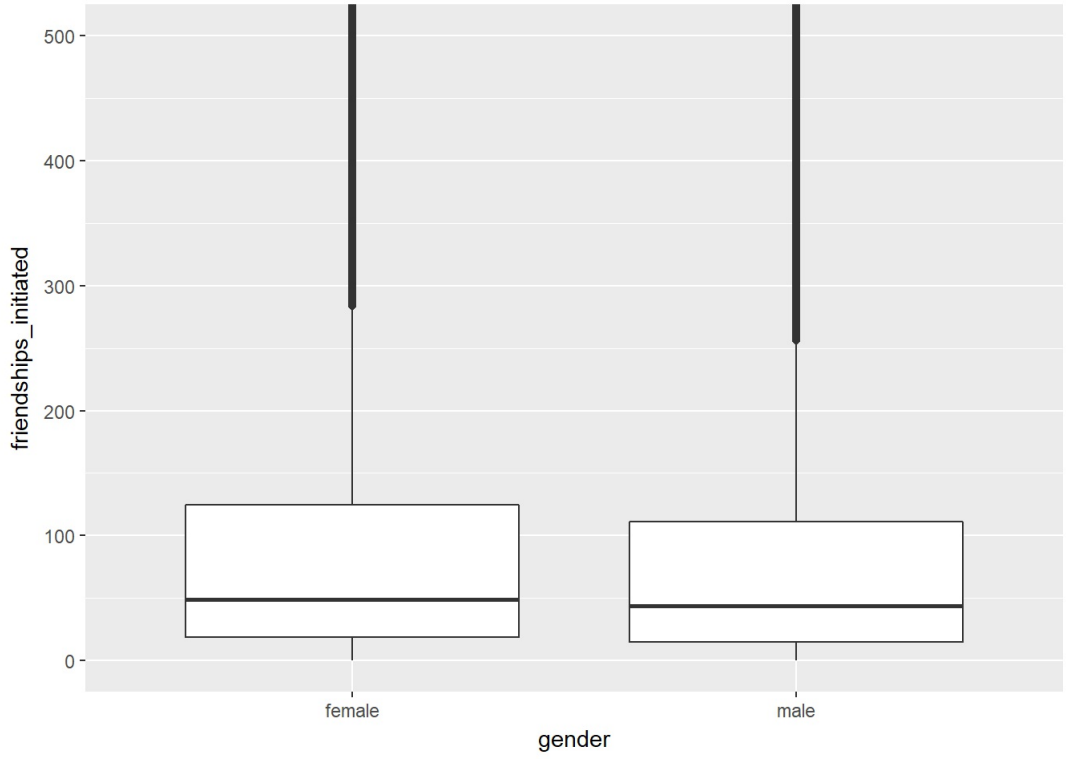
Box Plots, Quartiles, and Friendships

Notes:

On average, who initiated more friendships in our sample: men or women?

Response: #### Write about some ways that you can verify your answer. Response:because the data shows that most friend requests are below 500 we set coord cart limit there.

```
qplot(x = gender, y = friendships_initiated, data = subset(pf, !is.na(gender)), geom = 'boxplot') +
  coord_cartesian(ylim = c(0, 500))
```



Response:

Getting Logical

Notes:

```
summary(pf$mobile_likes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     4.0   106.1    46.0 25111.0
```

```
summary(pf$mobile_likes > 0)
```

```
##    Mode   FALSE    TRUE
## logical   35056   63947
```

```
mobile_check_in <- NA

pf$mobile_check_in <- ifelse(pf$mobile_likes > 0, 1, 0)
pf$mobile_check_in <- factor(pf$mobile_check_in)
summary(pf$mobile_check_in)
```

```
##     0     1
## 35056 63947
```

```
sum(pf$mobile_check_in ==1) / length(pf$mobile_check_in)
```

```
## [1] 0.6459097
```

Response:

Analyzing One Variable

Reflection:

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