Contents

Lesson 3

###Reading in Data

```
getwd()
```

```
## [1] "/Users/xiaonili"
```

```
list.files()
```

```
##  [1] "0325.Rhistory"                  "ABTest"
##  [3] "Applications"                   "Desktop"
##  [5] "Documents"                      "Downloads"
##  [7] "env"                            "Facets-ggplots-.html"
##  [9] "Facets-ggplots-.log"            "Facets(ggplots).Rmd"
## [11] "gitskills"                      "learngit"
## [13] "lesson3_student_files"          "lesson3_student.html"
## [15] "lesson3_student.md"             "lesson3_student.rmd"
## [17] "Library"                        "Movies"
## [19] "Music"                          "myproject"
## [21] "nba-players-histograms.R"       "nba-players.csv"
## [23] "opt"                            "Pictures"
## [25] "pseudo_facebook.tsv"            "Public"
## [27] "QEMU"                           "Read and Use Histogram in R.Rmd"
## [29] "Read and Use Histograms.Rmd"    "Read-and-Use-Histogram-in-R.html"
## [31] "Udacity"                        "VirtualBox VMs"
```

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

```
##  [1] "userid"                "age"                   "dob_day"
##  [4] "dob_year"              "dob_month"             "gender"
##  [7] "tenure"                "friend_count"          "friendships_initiated"
## [10] "likes"                 "likes_received"        "mobile_likes"
## [13] "mobile_likes_received" "www_likes"             "www_likes_received"
```

Notes:

Pseudo-Facebook User Data

Notes:

Histogram of Users' Birthdays

Notes:

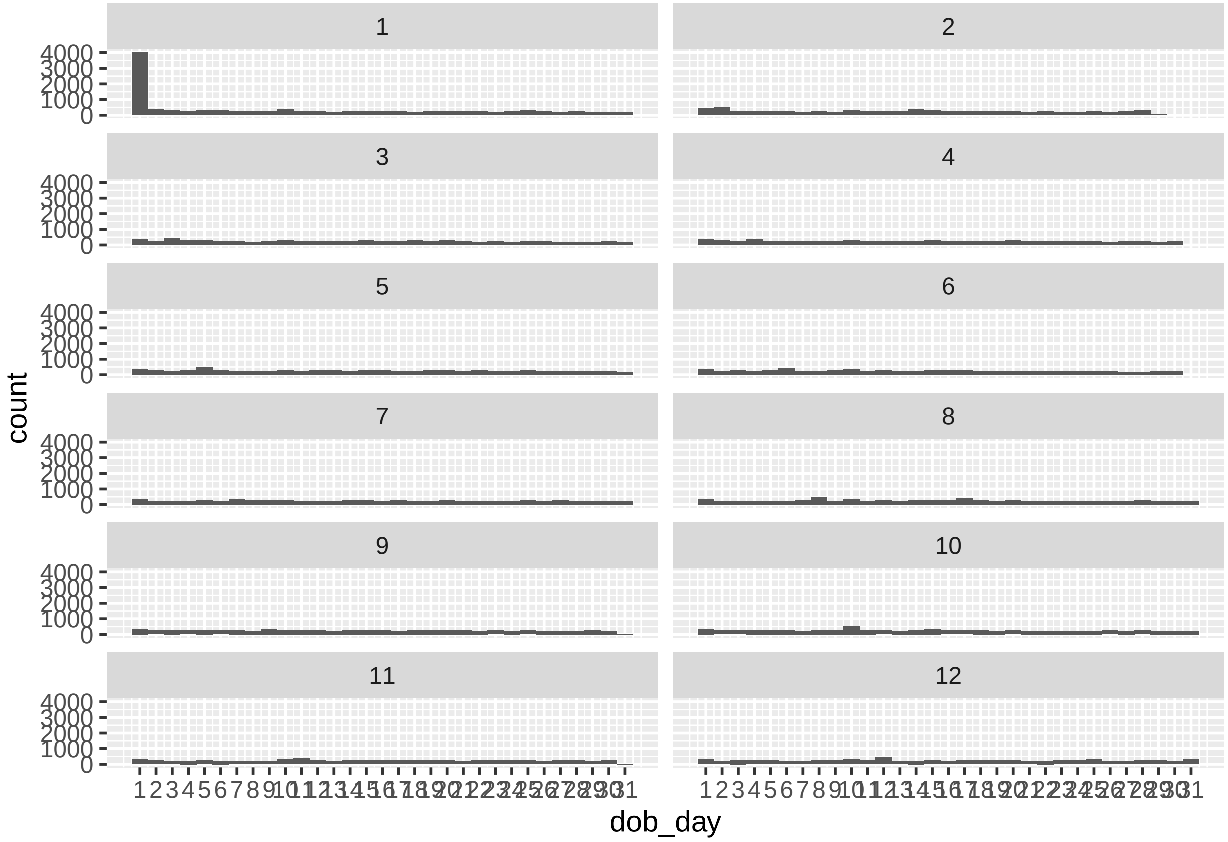
```
#install.packages('ggplot2')
library(ggplot2)
#names(pf)
#for old version:
#qplot(x=dob_day,data=pf)+
    scale_x_discrete(breaks=1:31)
```

```
## <ggproto object: Class ScaleDiscretePosition, ScaleDiscrete, Scale, gg>
##     aesthetics: x xmin xmax xend
##     axis_order: function
##     break_info: function
##     break_positions: function
##     breaks: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 ...
##     call: call
##     clone: function
##     dimension: function
##     drop: TRUE
##     expand: waiver
##     get_breaks: function
##     get_breaks_minor: function
##     get_labels: function
##     get_limits: function
##     guide: waiver
##     is_discrete: function
##     is_empty: function
##     labels: waiver
##     limits: NULL
##     make_sec_title: function
##     make_title: function
##     map: function
##     map_df: function
##     n.breaks.cache: NULL
##     na.translate: TRUE
##     na.value: NA
##     name: waiver
##     palette: function
##     palette.cache: NULL
##     position: bottom
##     range: <ggproto object: Class RangeDiscrete, Range, gg>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeDiscrete, Range, gg>
##     range_c: <ggproto object: Class RangeContinuous, Range, gg>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeContinuous, Range, gg>
##     rescale: function
##     reset: function
##     scale_name: position_d
##     train: function
##     train_df: function
##     transform: function
##     transform_df: function
##     super:  <ggproto object: Class ScaleDiscretePosition, ScaleDiscrete, Scale, gg>
```

```
#qplot(x=dob_day,data=pf)+
    scale_x_continuous(breaks=1:31)
```

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --    1
```

```
ggplot(aes(x=dob_day),data=pf)+
    geom_histogram(binwidth=1)+
    scale_x_continuous(breaks=1:31)+
    facet_wrap(~dob_month,ncol=2)
```



What are some things that you notice about this histogram?

Response:

```
#big difference for Jan.
```

Moira's Investigation

Notes:

Estimating Your Audience Size

Notes:

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

Response:

How many of your friends do you think saw that post?

Response:

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:

Perceived Audience Size

Notes:

Faceting

Notes:

```
# facet_wrap(~dob_month)
```

Let's take another look at our plot. What stands out to you here?

Response:

Be Skeptical - Outliers and Anomalies

Notes:

Moira's Outlier

Notes: #### Which case do you think applies to Moira's outlier? Response: bad data. ***

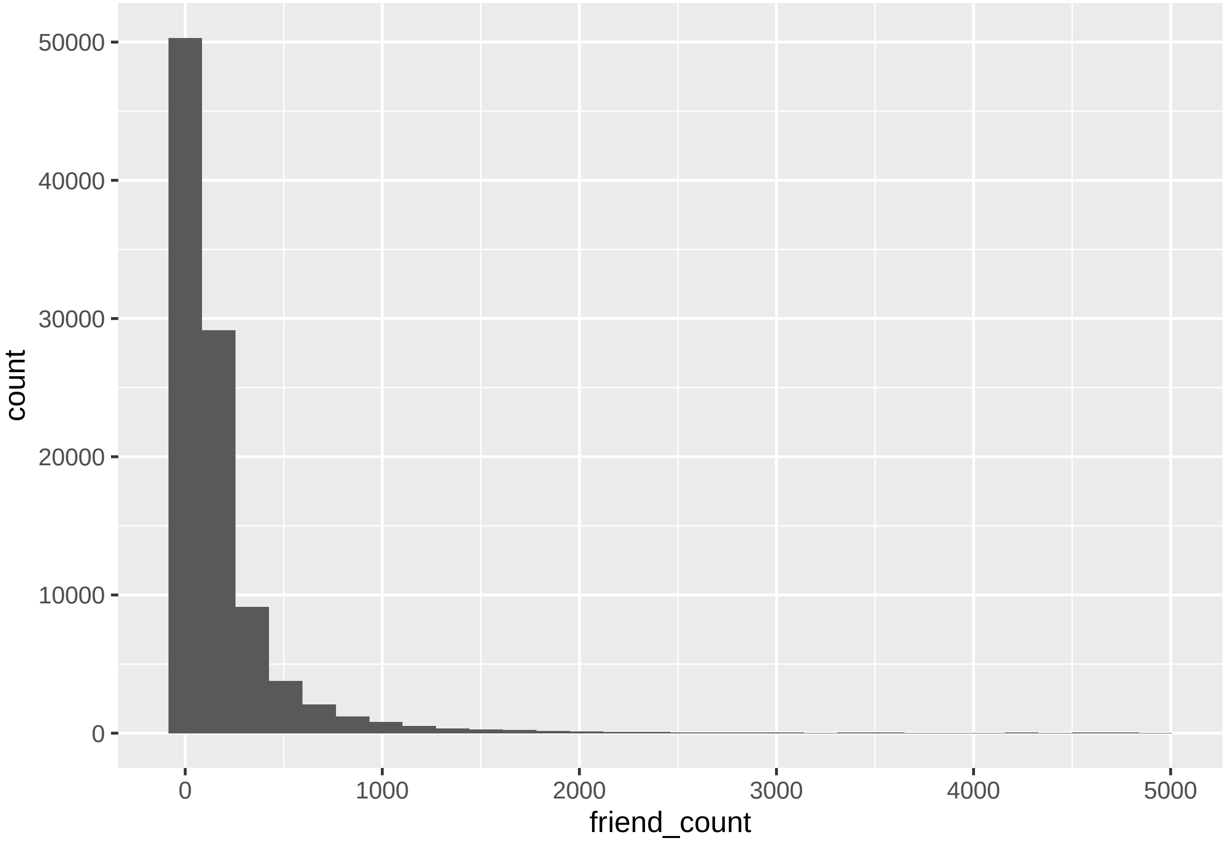
Friend Count

Notes:

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

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

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



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

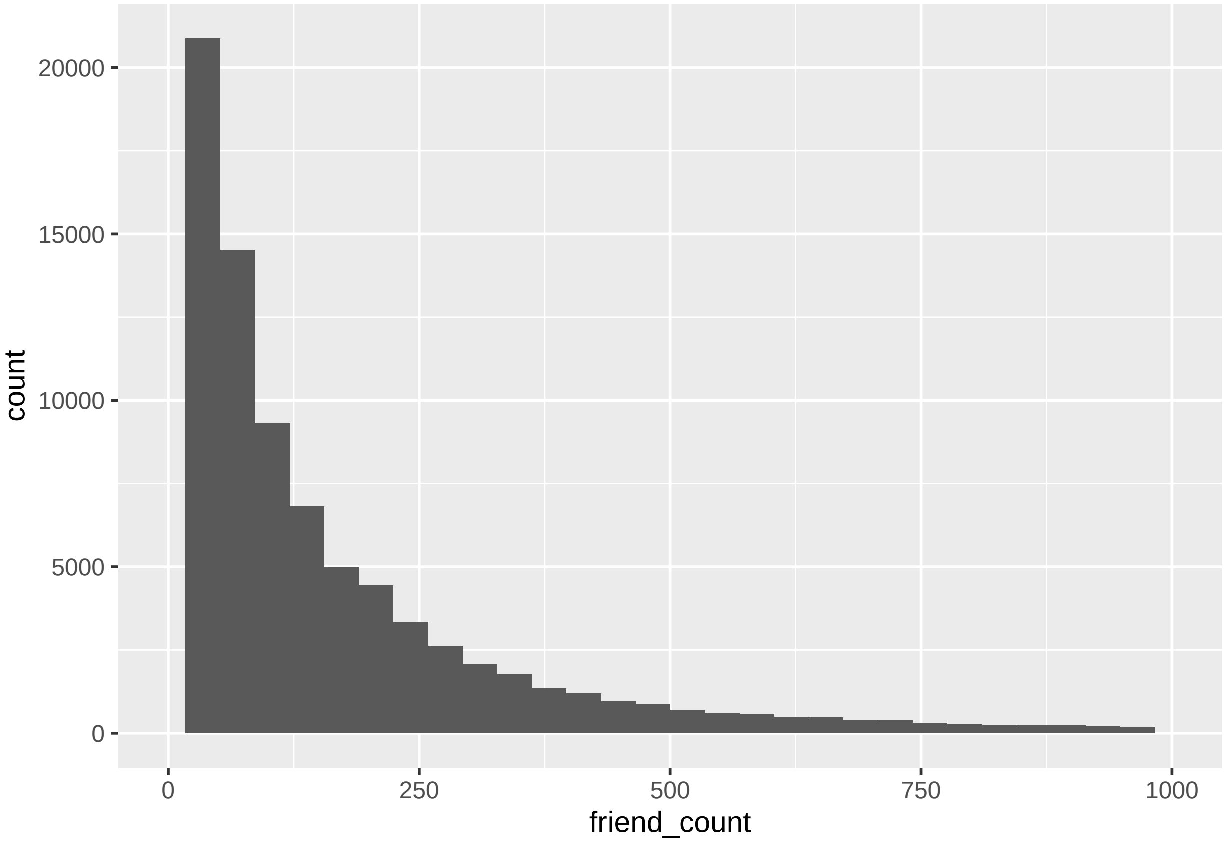
Response:

Limiting the Axes

Notes:

```
ggplot(aes(x=friend_count),data=pf) +geom_histogram()+scale_x_continuous(limits=c(0,1000))
```

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

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

## Warning: Removed 2 rows containing missing values (geom_bar).
```



Exploring with Bin Width

Notes:

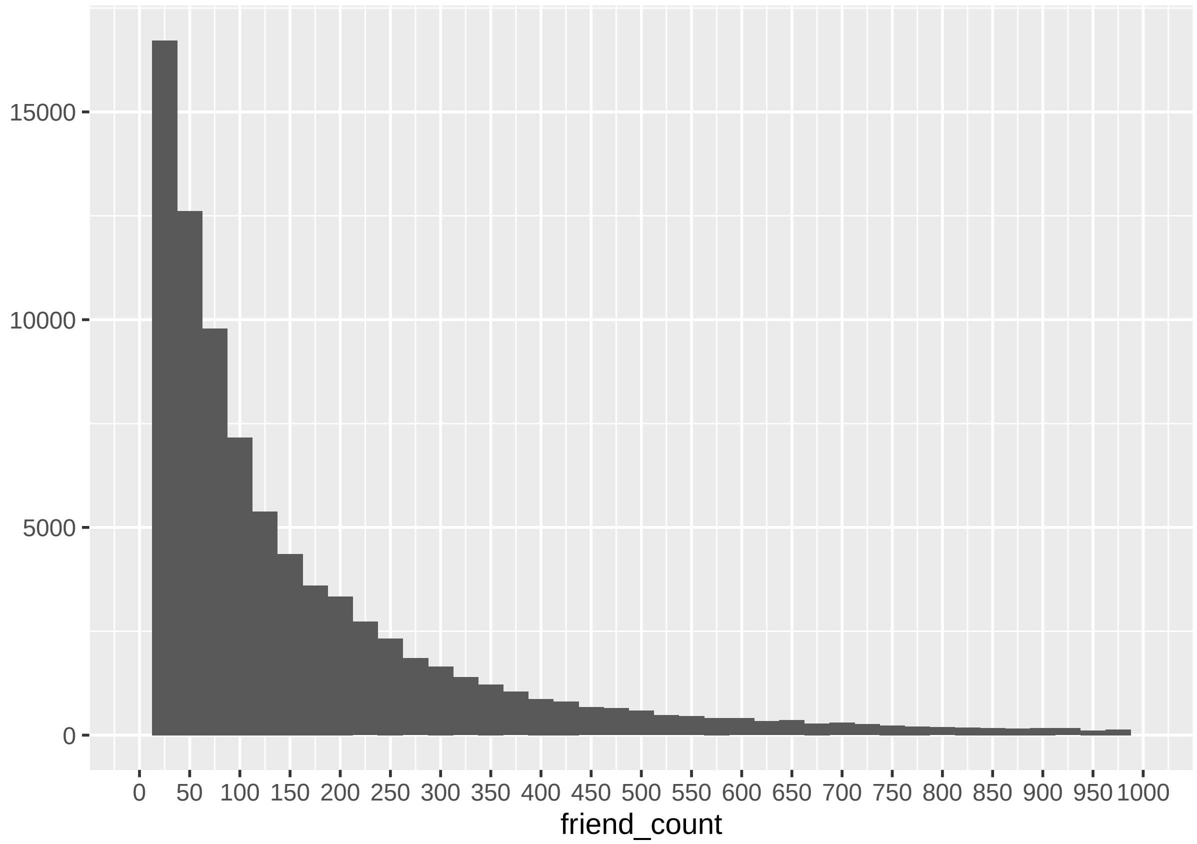
Adjusting the Bin Width

Notes:

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

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

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

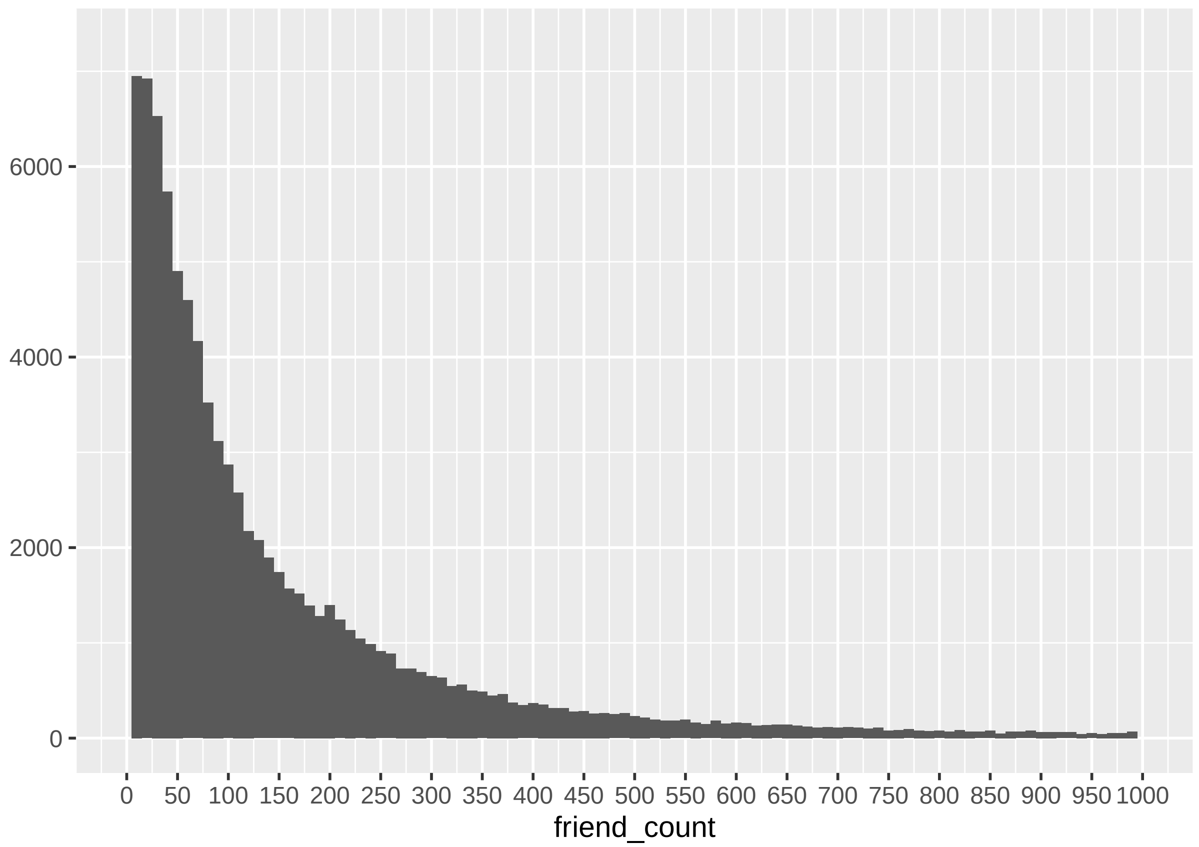


Faceting Friend Count

```r
# 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).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Omitting NA Values

Notes:

Statistics 'by' Gender

Notes:

```
library(ggplot2)
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).
```

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

[](lesson3_student_files/figure-latex/Statistics 'by' Gender-1.pdf)

Who on average has more friends: men or women?

Response:

```
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
```

What's the difference between the median friend count for women and men?

Response:

```
96-74
```

```
## [1] 22
```

Why would the median be a better measure than the mean?

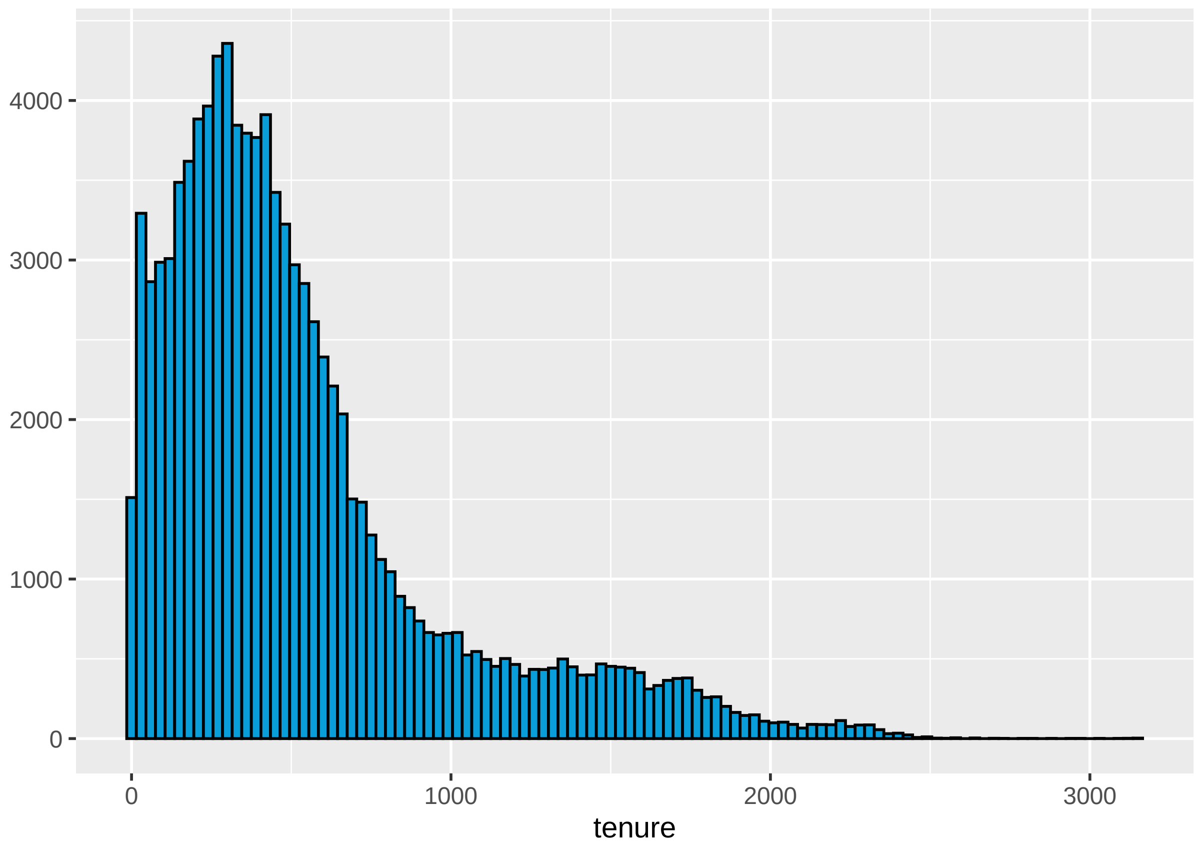
Response: median is more robust statistic. ***

Tenure

Notes:

```
qplot(x=tenure,data=pf,binwidth=30,
            color=I('black'),fill=I('#099DD9'))
```

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

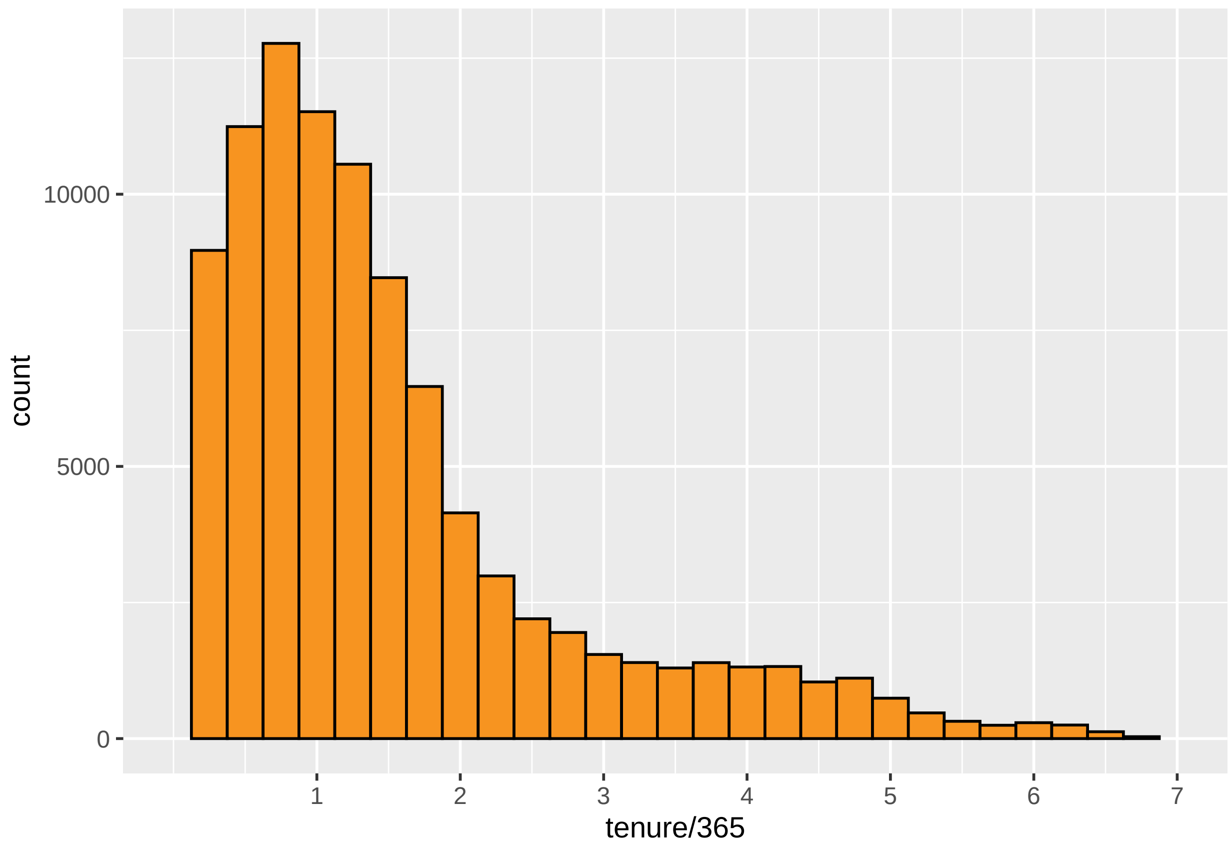


How would you create a histogram of tenure by year?

```
ggplot(aes(x=tenure/365),data=pf)+
    geom_histogram(binwidth = .25,color='black',fill='#F79420')+
    scale_x_continuous(breaks=seq(1,7,1),limits=c(0,7))
```

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

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



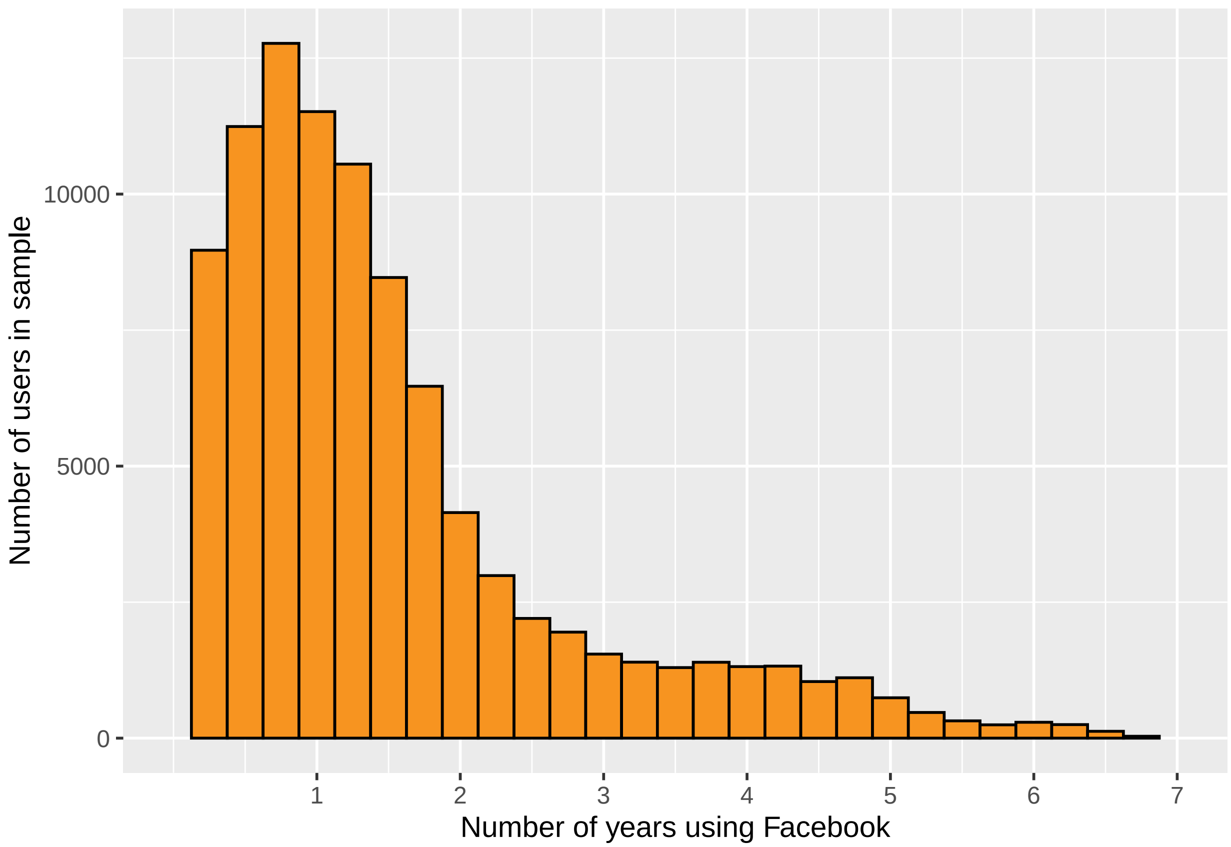
Labeling Plots

Notes:

```
ggplot(aes(x=tenure/365),data=pf)+
    geom_histogram(binwidth = .25,color='black',fill='#F79420')+
    scale_x_continuous(breaks=seq(1,7,1),limits=c(0,7))+
     xlab('Number of years using Facebook')+ylab('Number of users in sample')
```

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

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

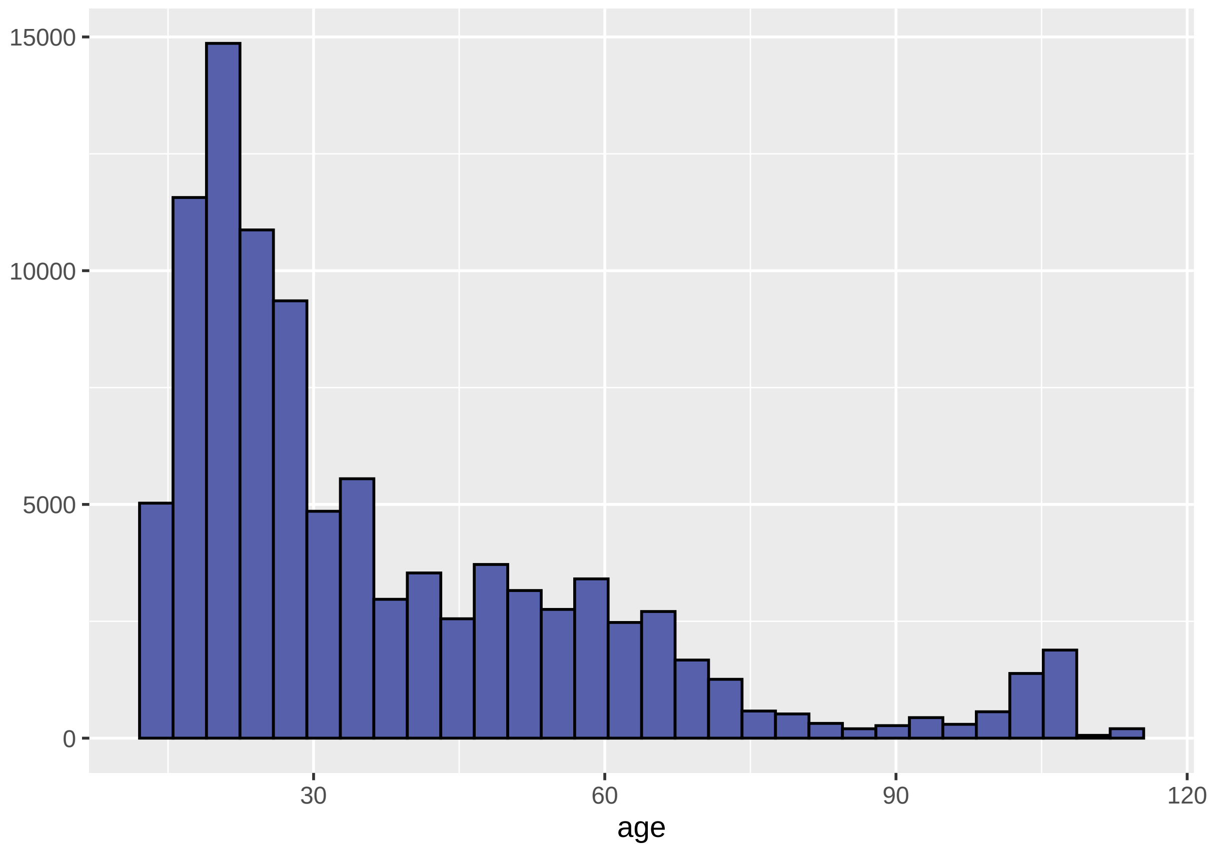


User Ages

Notes:

```
#ggplot(aes(x=age),data=pf)+geom_histogram(binwidth=1,fill='#5760AB')+
#   scale_x_continuous(breaks=seq(0,113,5))

qplot(x=age,data=pf,
            color=I('black'),fill=I('#5760AB'))
```

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



What do you notice?

Response:

The Spread of Memes

Notes:

Lada's Money Bag Meme

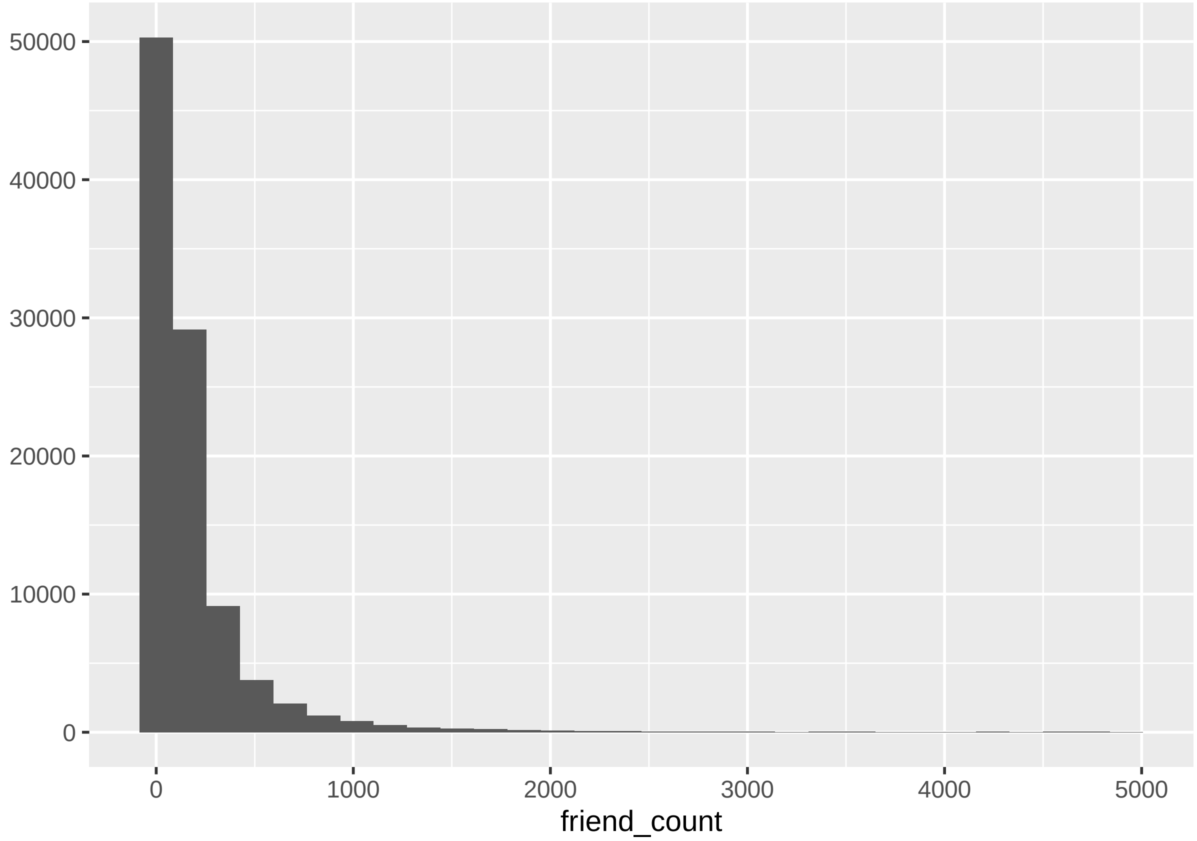
Notes:

Transforming Data

Notes:

```r
qplot(x=friend_count,data=pf)
```

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



```r
summary(pf$friend_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    31.0    82.0   196.4   206.0  4923.0
```

```r
summary(log10(pf$friend_count+1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   1.505   1.919   1.868   2.316   3.692
```

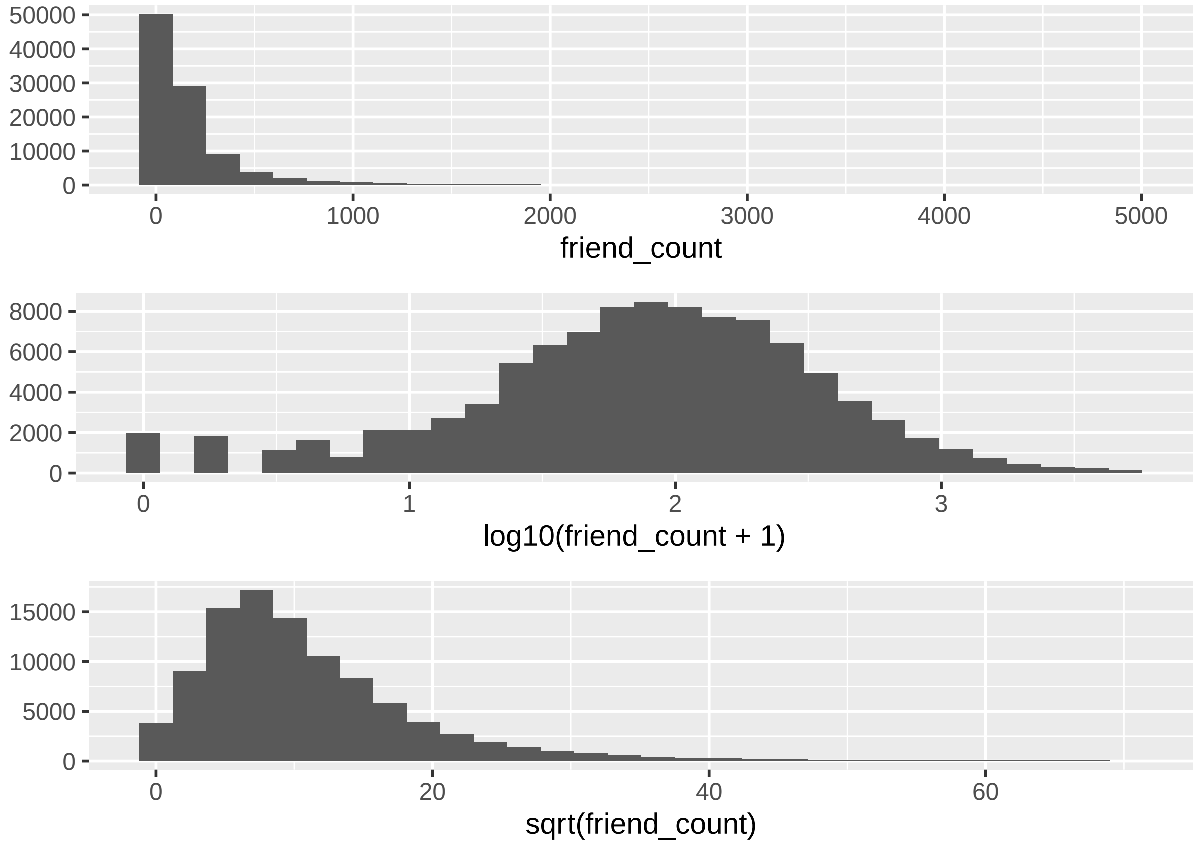
```r
summary(sqrt(pf$friend_count))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   5.568   9.055  11.088  14.353  70.164
```

Transforming Data solution

```r
library(gridExtra)
p1=qplot(x=friend_count,data=pf)
p2=qplot(x=log10(friend_count+1),data=pf)
p3=qplot(x=sqrt(friend_count),data=pf)
grid.arrange(p1,p2,p3,ncol=1)
```

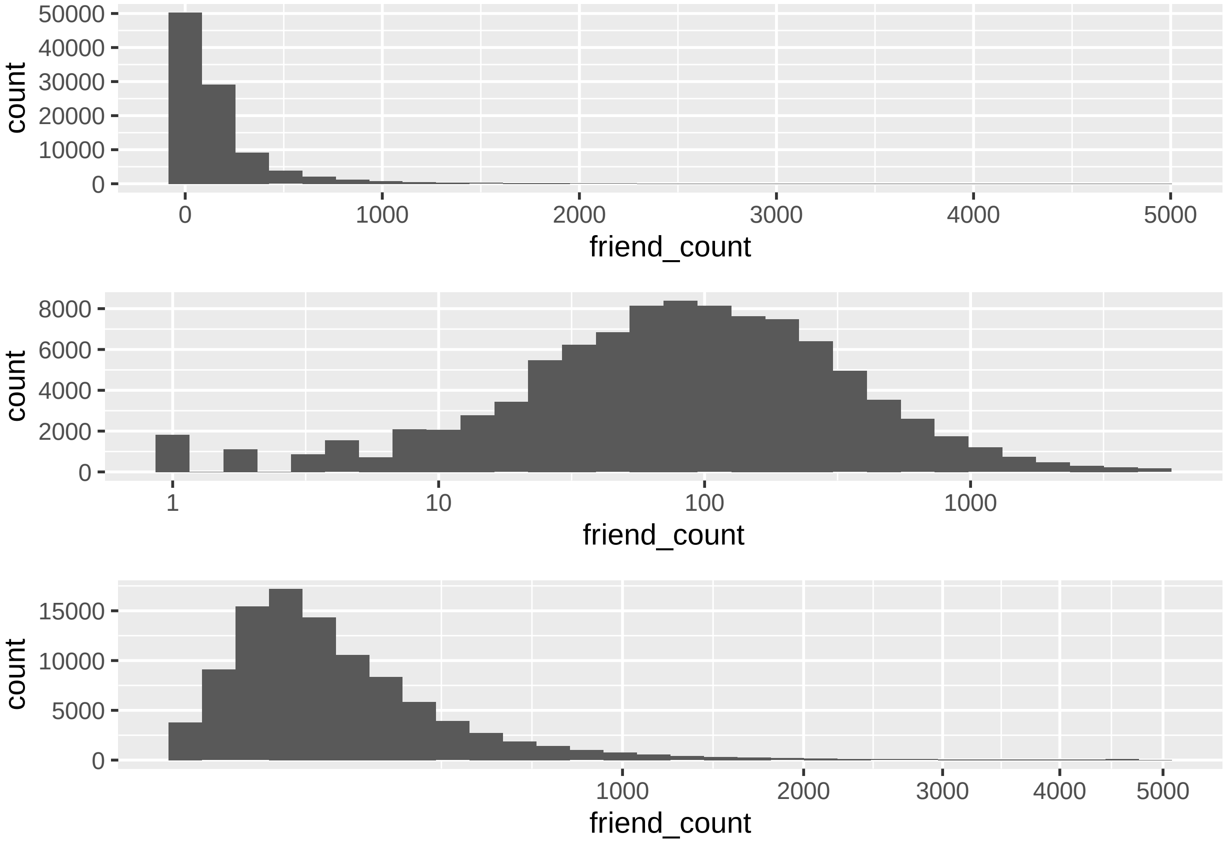
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



### Transforming Data alternate solution

```r
p1=ggplot(aes(x=friend_count),data=pf)+geom_histogram()
p2=p1+scale_x_log10()
p3=p1+scale_x_sqrt()
grid.arrange(p1,p2,p3,ncol=1)
```

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

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

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

## Warning: Removed 1962 rows containing non-finite values (stat_bin).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

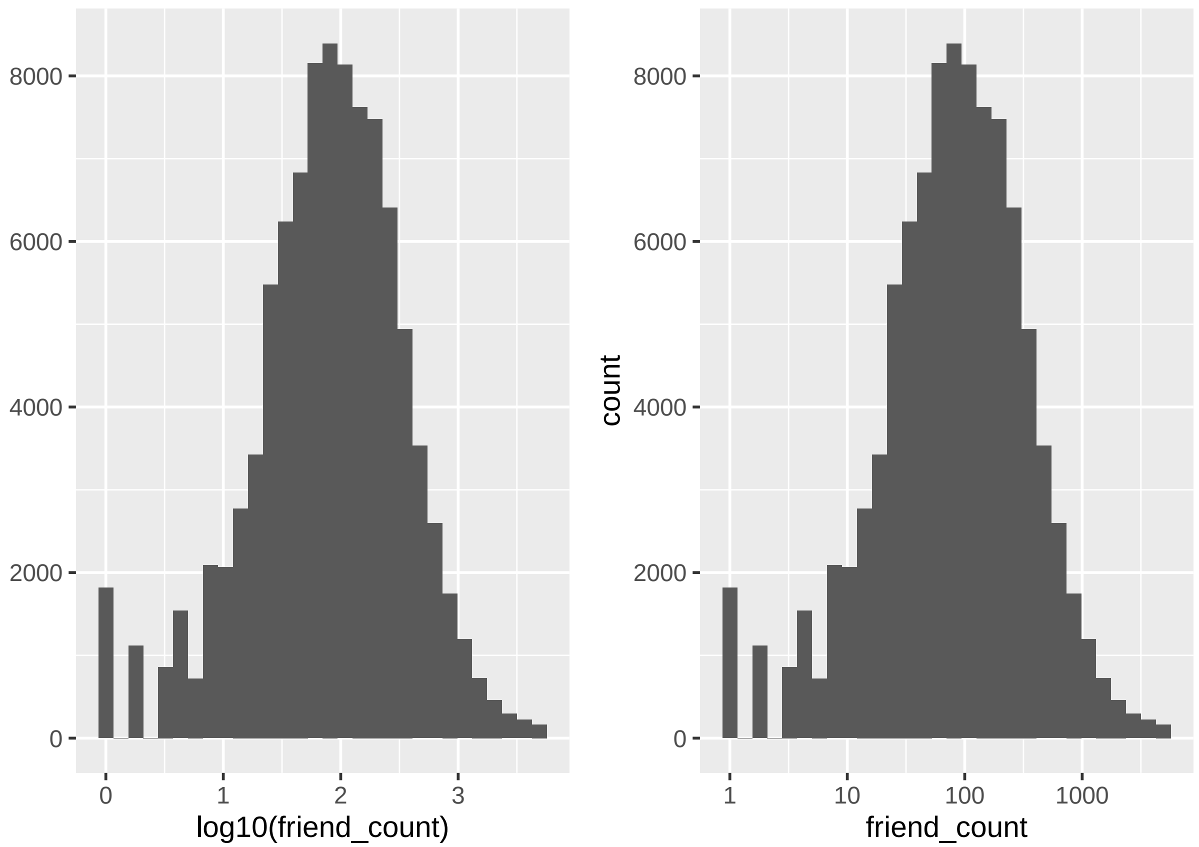


Add a Scaling Layer

Notes:

```
logScale=qplot(x=log10(friend_count),data=pf)

countScale=ggplot(aes(x=friend_count),data=pf)+geom_histogram()+scale_x_log10()

grid.arrange(logScale,countScale,ncol=2)
```

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

## Warning: Removed 1962 rows containing non-finite values (stat_bin).

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

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

## Warning: Removed 1962 rows containing non-finite values (stat_bin).
```

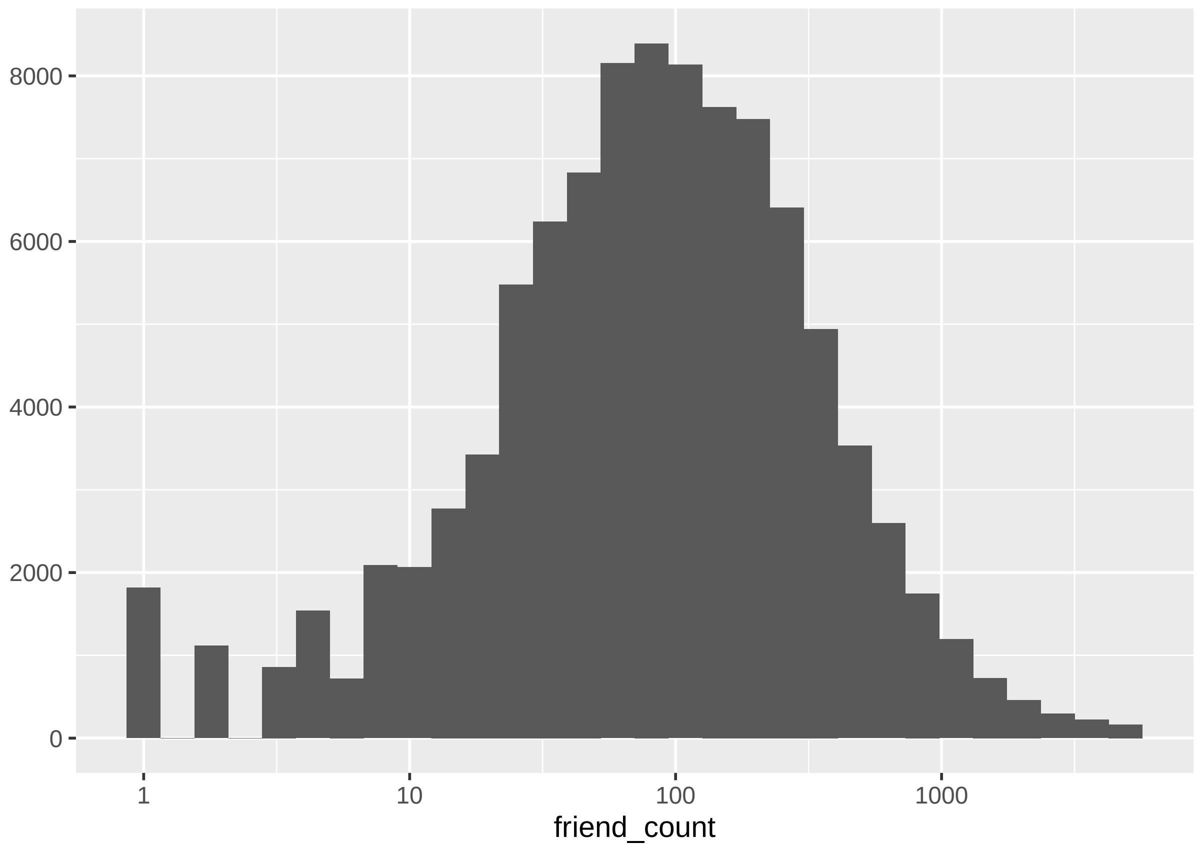


```
qplot(x=friend_count,data=pf)+
    scale_x_log10()
```

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

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

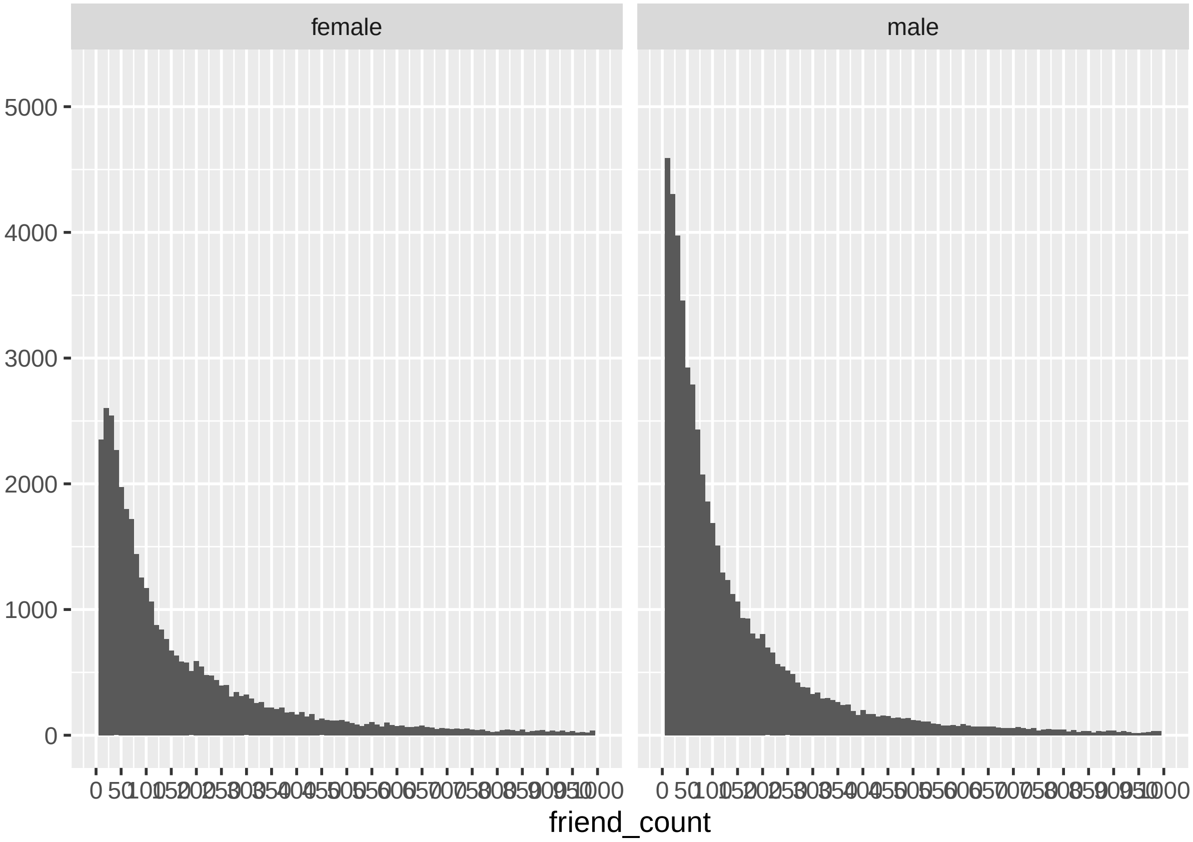
```
## Warning: Removed 1962 rows containing non-finite values (stat_bin).
```



Frequency Polygons

```
qplot(x=friend_count,data=subset(pf,!is.na(gender)),binwidth=10)+
    scale_x_continuous(lim=c(0,1000),breaks = seq(0,1000,50))+facet_wrap(~gender)
```

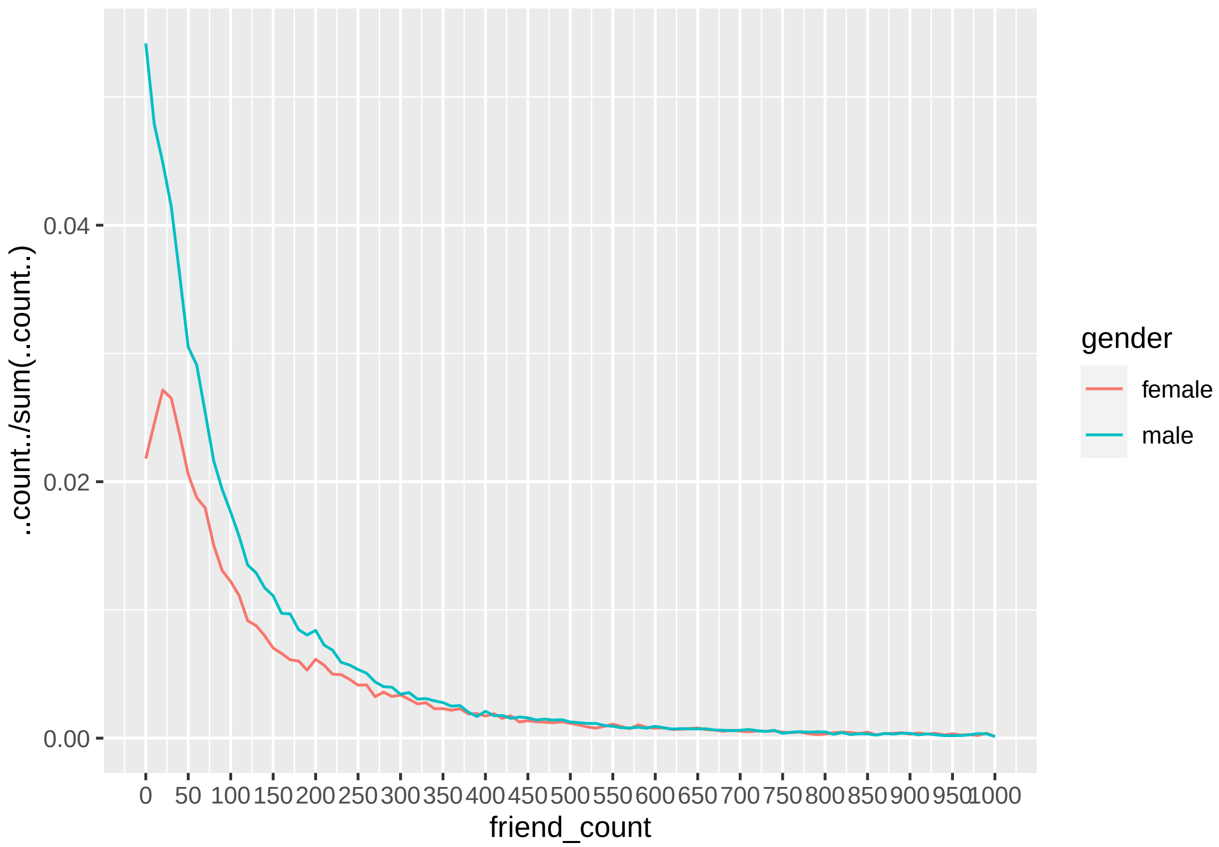
```
## Warning: Removed 2949 rows containing non-finite values (stat_bin).

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



```
qplot(x=friend_count,y=..count../sum(..count..),data=subset(pf,!is.na(gender)),binwidth=10,geom='freqpol
    scale_x_continuous(lim=c(0,1000),breaks = seq(0,1000,50))
```

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

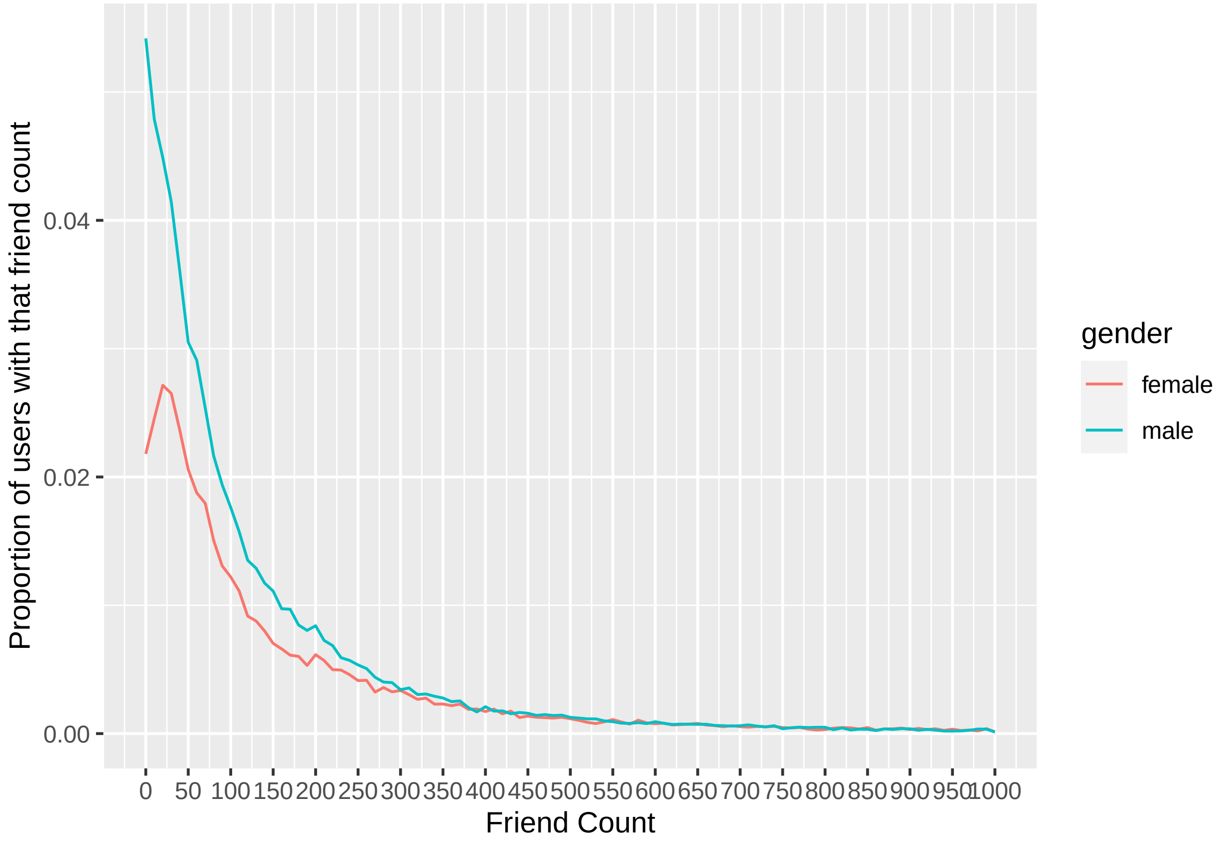
```
## Warning: Removed 4 row(s) containing missing values (geom_path).
```



```
#alternative
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 row(s) containing missing values (geom_path).
```



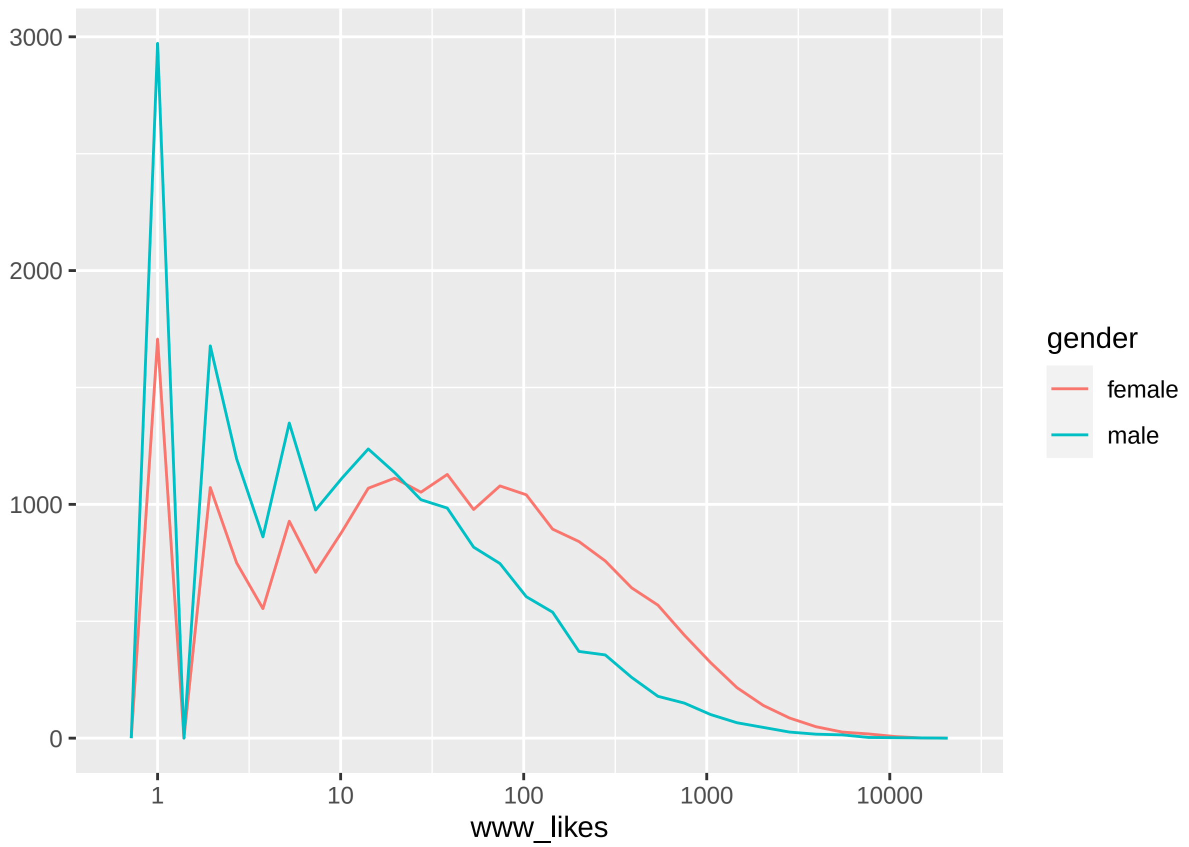
```
qplot(x=www_likes,data = subset(pf,!is.na(gender)),
            geom = 'freqpoly',color=gender)+
            scale_x_continuous()+
            scale_x_log10()
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## 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:

```
names(pf)
```

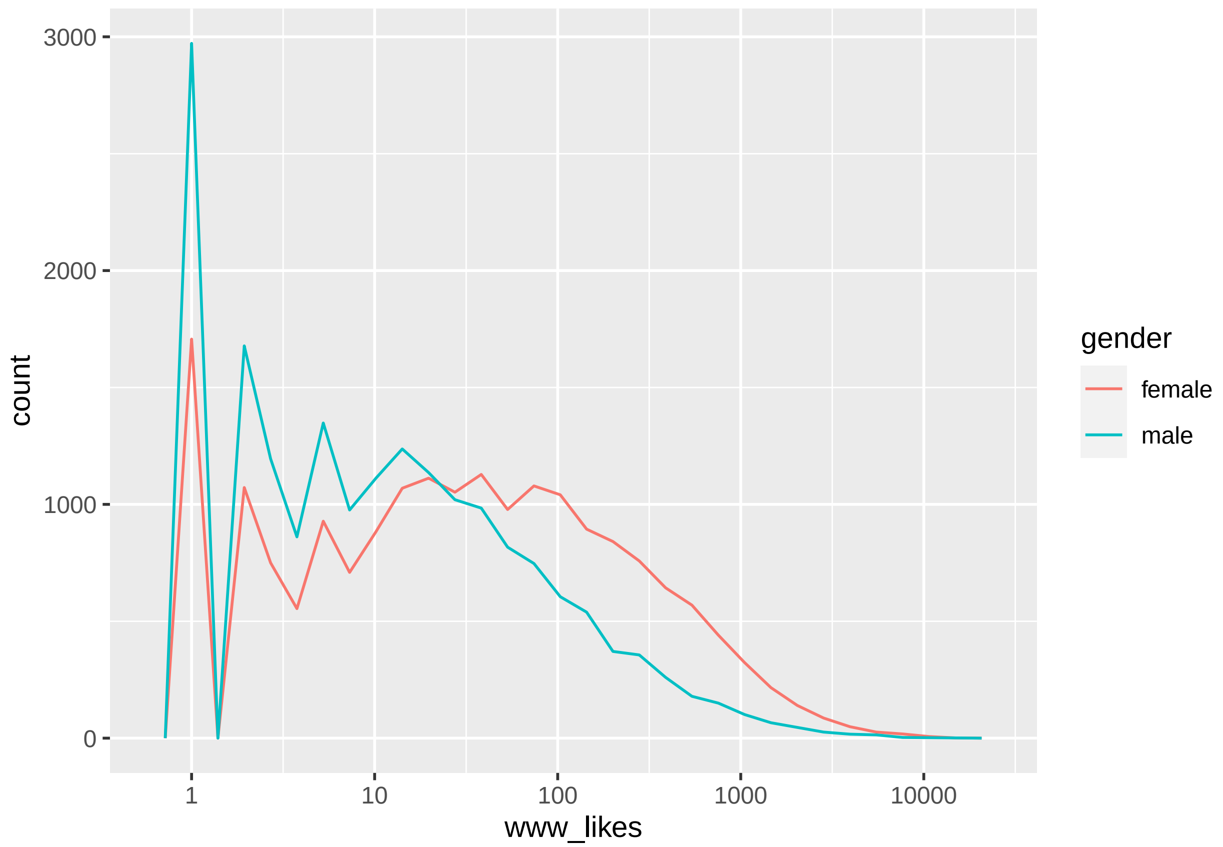
```
##  [1] "userid"                "age"                   "dob_day"
##  [4] "dob_year"              "dob_month"             "gender"
##  [7] "tenure"                "friend_count"          "friendships_initiated"
## [10] "likes"                 "likes_received"        "mobile_likes"
## [13] "mobile_likes_received" "www_likes"             "www_likes_received"
```

```
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).
```



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

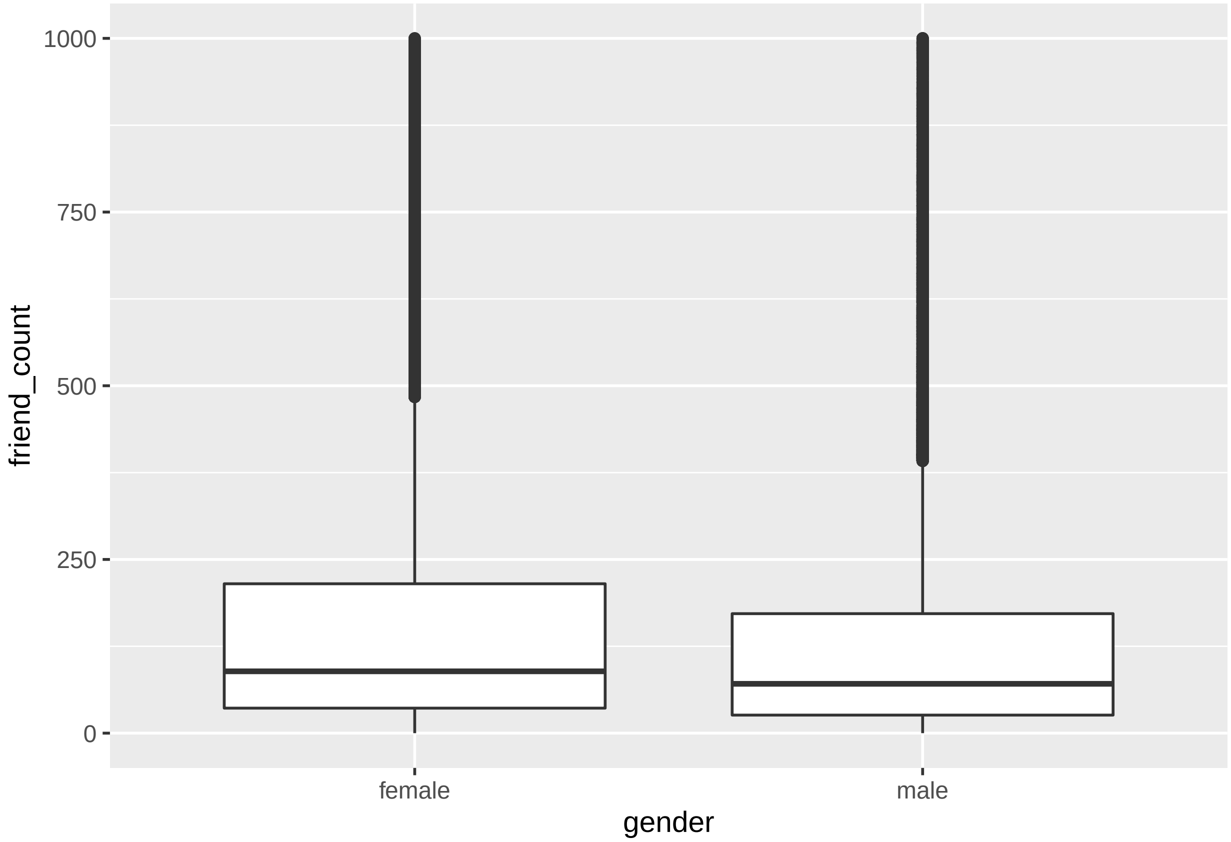
```
## pf$gender: female
## [1] 3507665
## ------------------------------------------------------------
## pf$gender: male
## [1] 1430175
```

Box Plots

Notes:

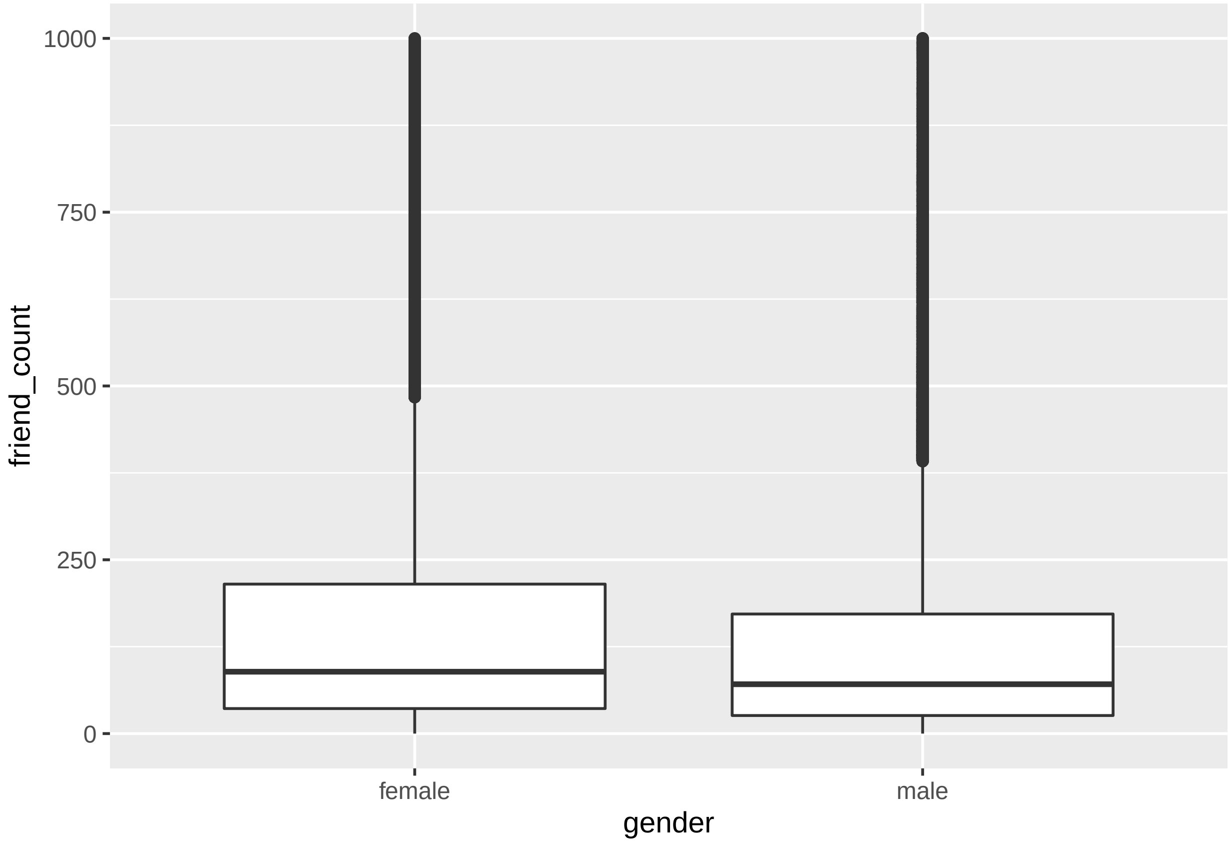
```
# ylim()

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

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



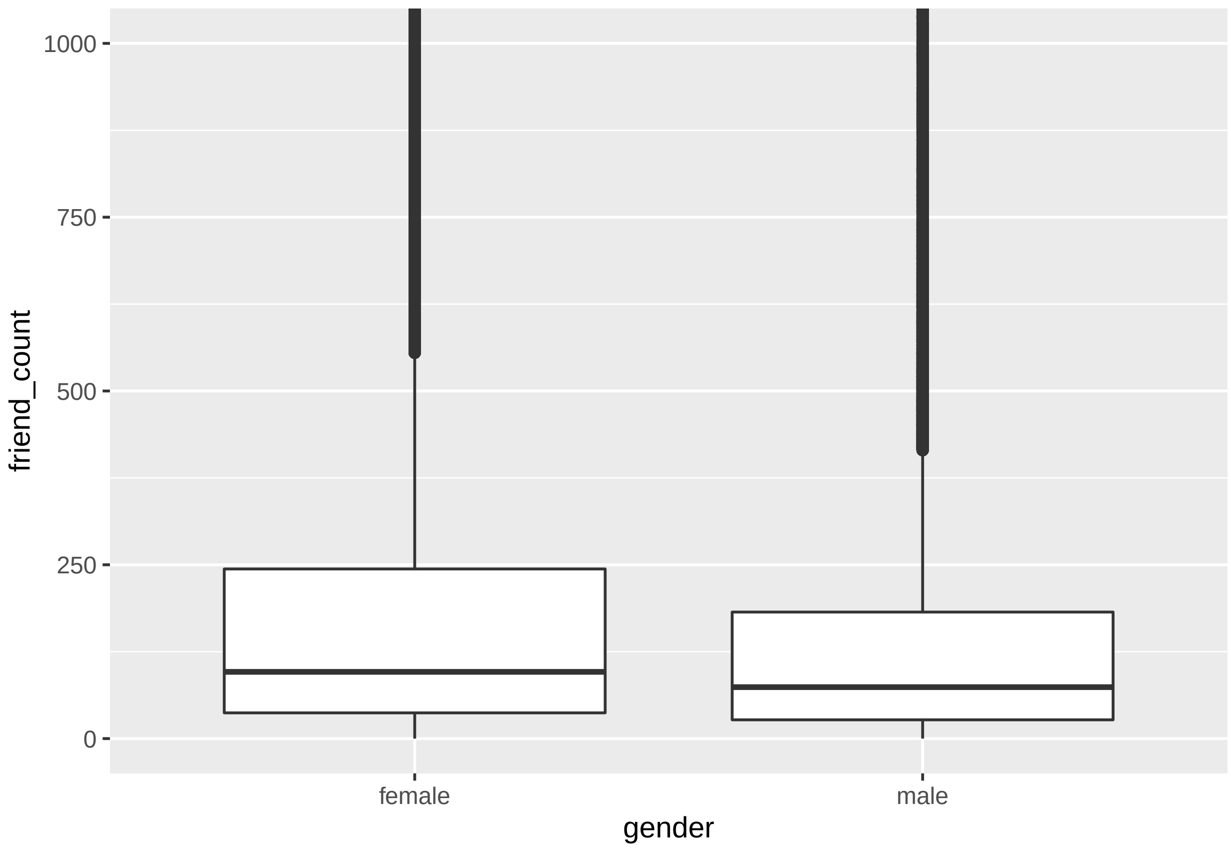
```
#scale_y_continuous()

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

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



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

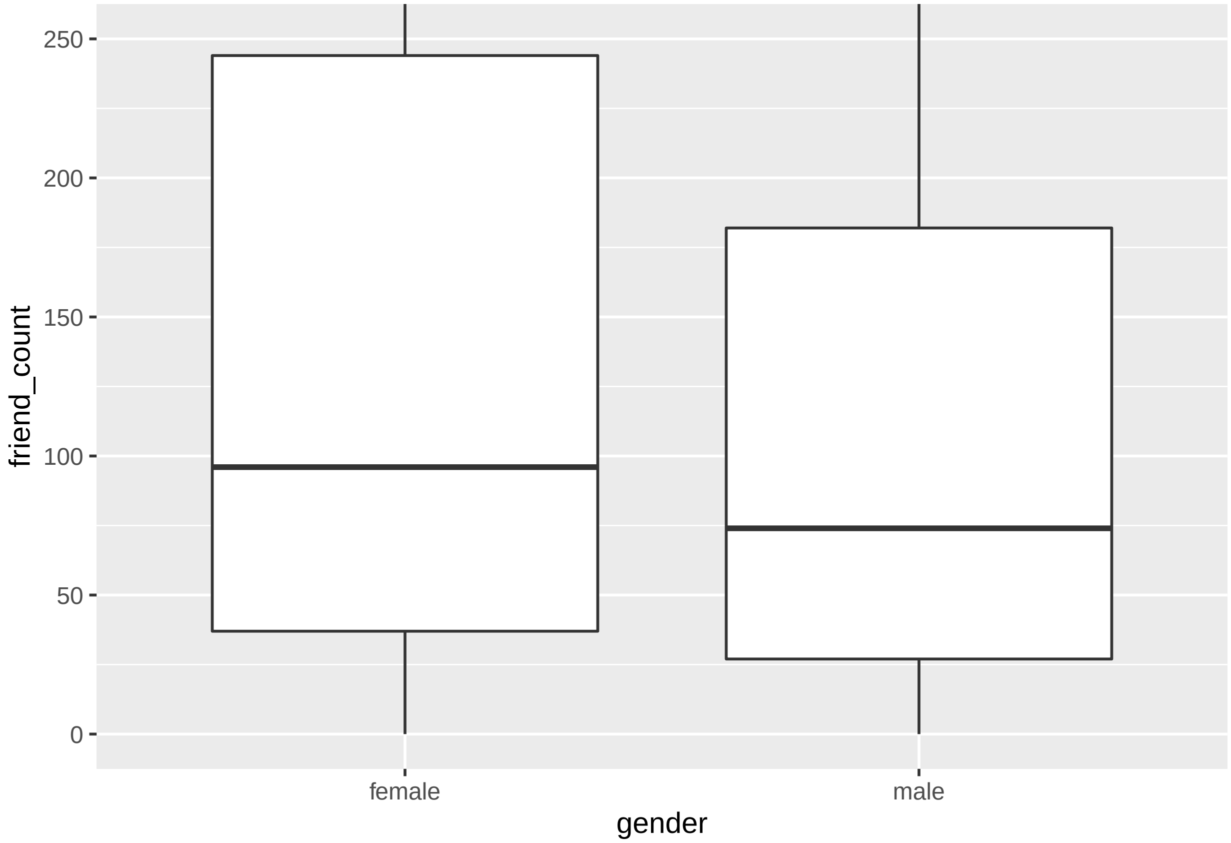
```
qplot(x=gender,y=friend_count,
                    data=subset(pf,!is.na(gender)),
                    geom = 'boxplot')+
      coord_cartesian(ylim=c(0,1000))
```



Box Plots, Quartiles, and Friendships

Notes:

```
qplot(x=gender,y=friend_count,
                    data=subset(pf,!is.na(gender)),
                    geom = 'boxplot')+
      coord_cartesian(ylim=c(0,250))
```



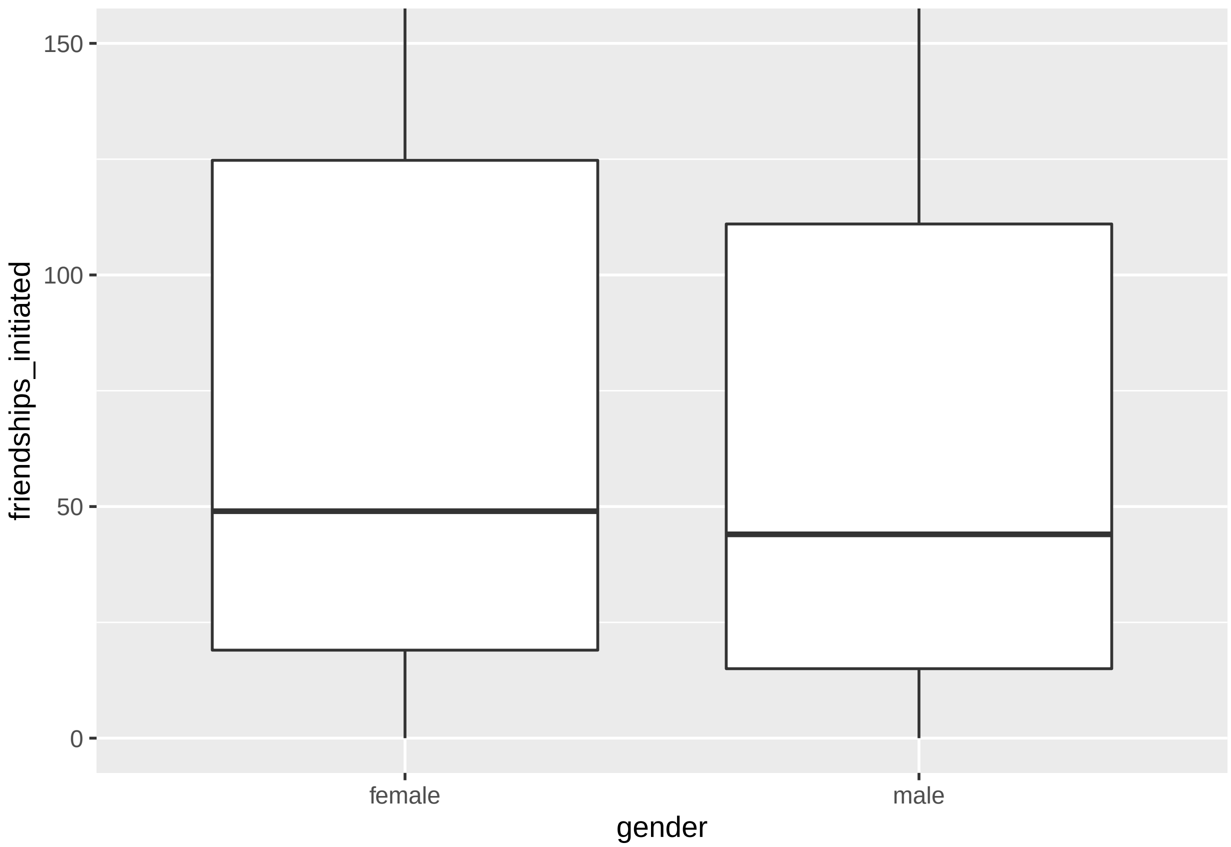
```r
by(pf$friend_count,pf$gender,summary)
```

```
## 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
```

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

Response: #### Write about some ways that you can verify your answer. Response:

```r
qplot(x=gender, y=friendships_initiated,
         data = subset(pf,!is.na(gender)),geom = 'boxplot')+
          coord_cartesian(ylim = c(0,150))
```



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

```
## pf$gender: female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    19.0    49.0   113.9   124.8  3654.0
## -------------------------------------------------------------
## pf$gender: male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    15.0    44.0   103.1   111.0  4144.0
```

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)
b=length((pf$mobile_check_in))
b
```

```
## [1] 99003
```

```
a=sum(pf$mobile_check_in == 1)
a
```

```
## [1] 63947
```

```
summary(pf$mobile_check_in)
```

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

```
35056+63947
```

```
## [1] 99003
```

```
63947/(35056+63947)
```

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

```
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!