Importação pseudo-Facebook dataset

```
filename <- "pseudo_facebook.tsv"

pf <- read.csv(filename, sep = "\t")
```

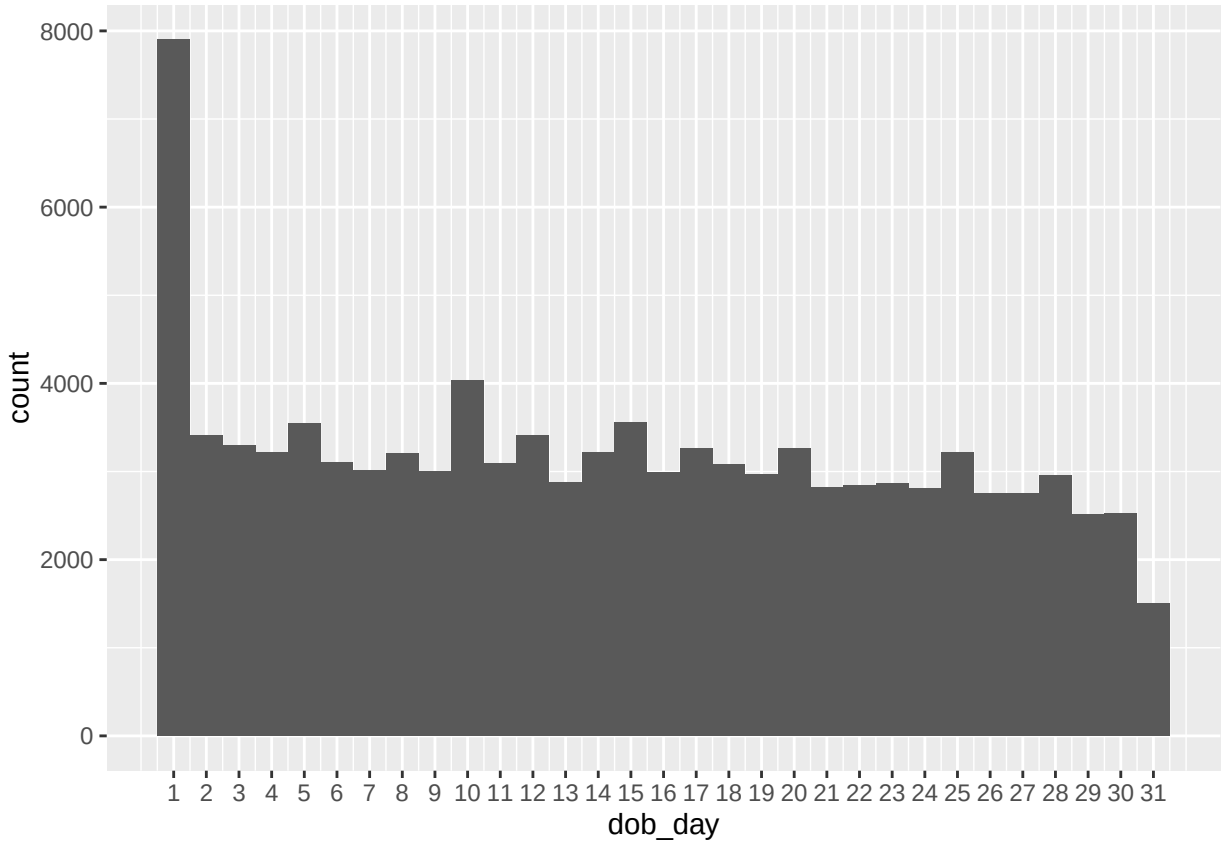
Histograma dos dias de aniversário dos usuários

```
install.packages('ggplot2')
```

```
## Installing package into 'C:/Users/leo/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\leo\AppData\Local\Temp\RtmpQlNEru\downloaded_packages
```

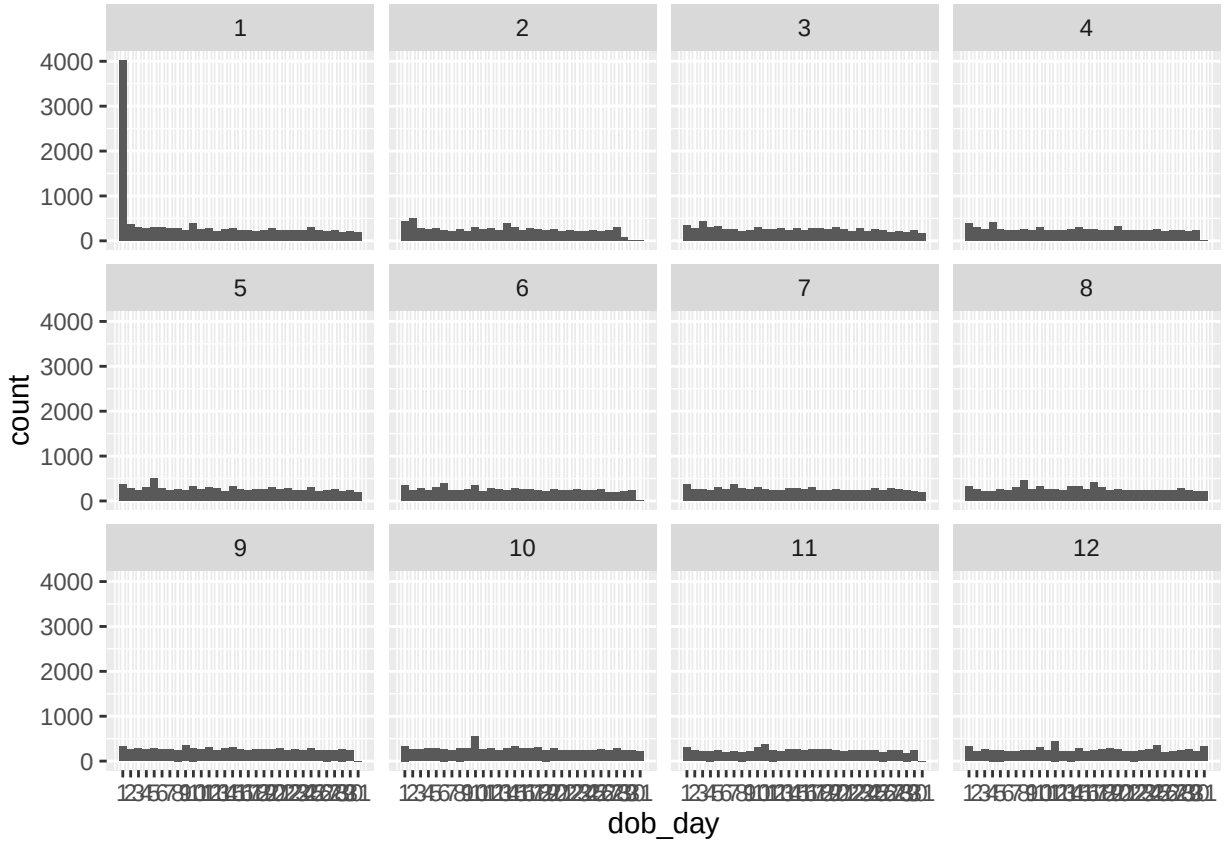
```
library(ggplot2)

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



Faceting

Plot para visualizar o dia do aniversário dos usuários do pseudo-facebook. Observando que o dia 1º de janeiro está muito fora do padrão, ou seja, denominado "Outlier".

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

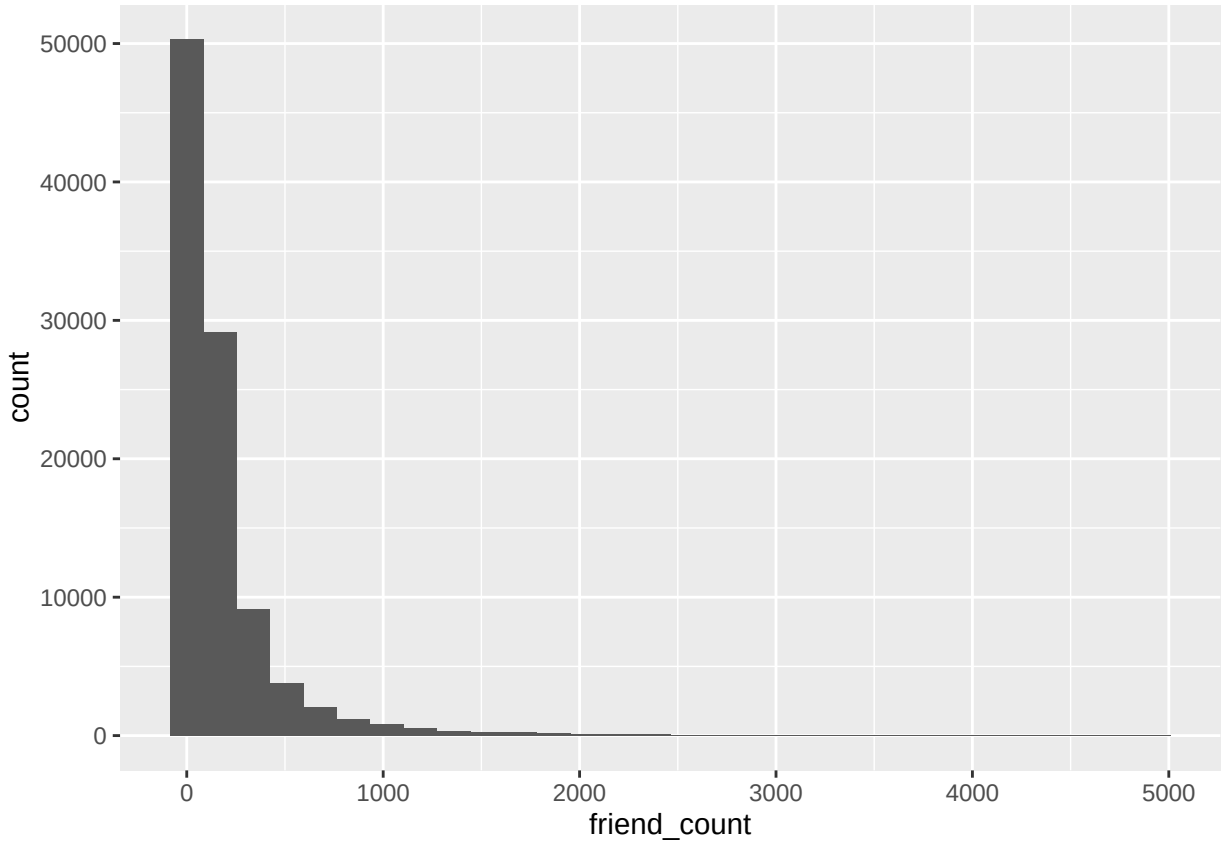


Contador de amigos

Histograma da quantidade de amigos. O gráfico ficou com cauda longa, abaixo vamos melhorá-lo.

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

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



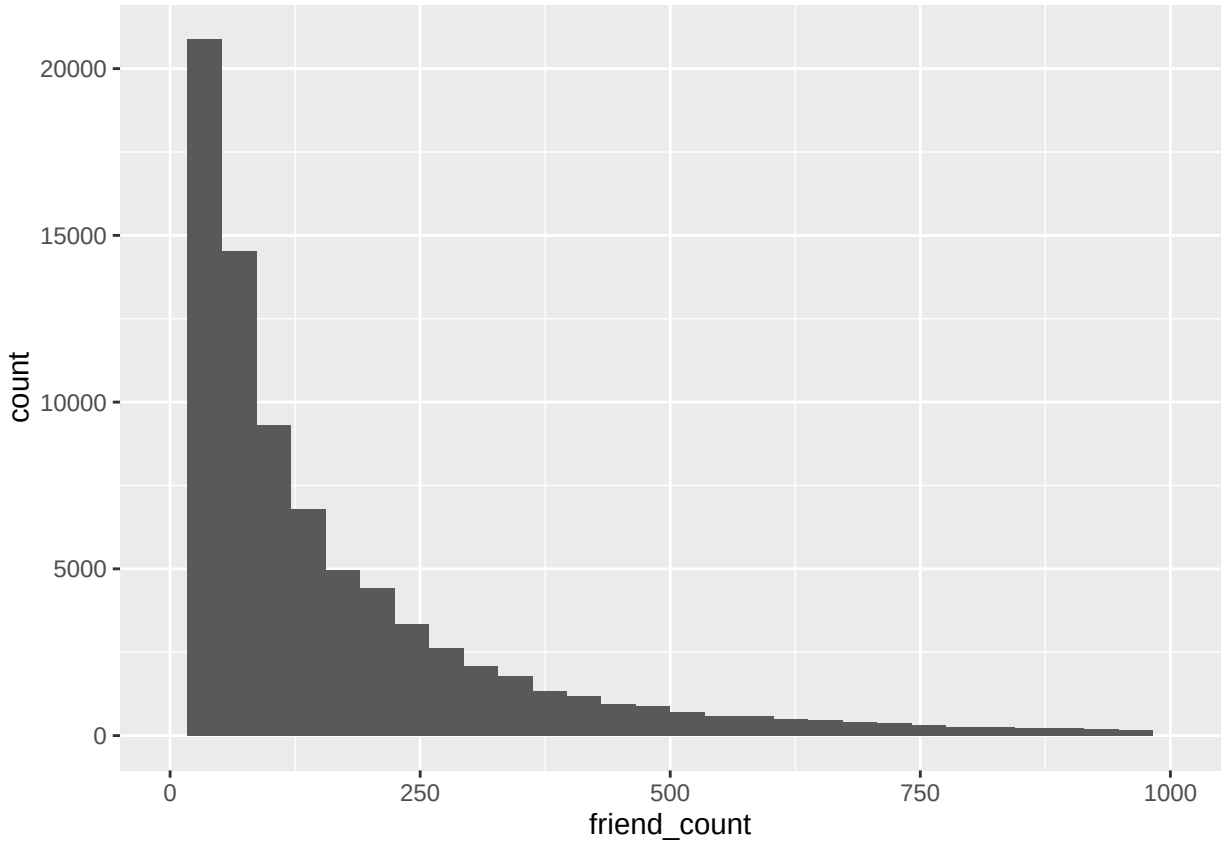
Limitando os eixos

Limitando o eixo X para diminuir a abrangência dos dados.

```
qplot(x = friend_count, data = pf, xlim = c(0, 1000))
```

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

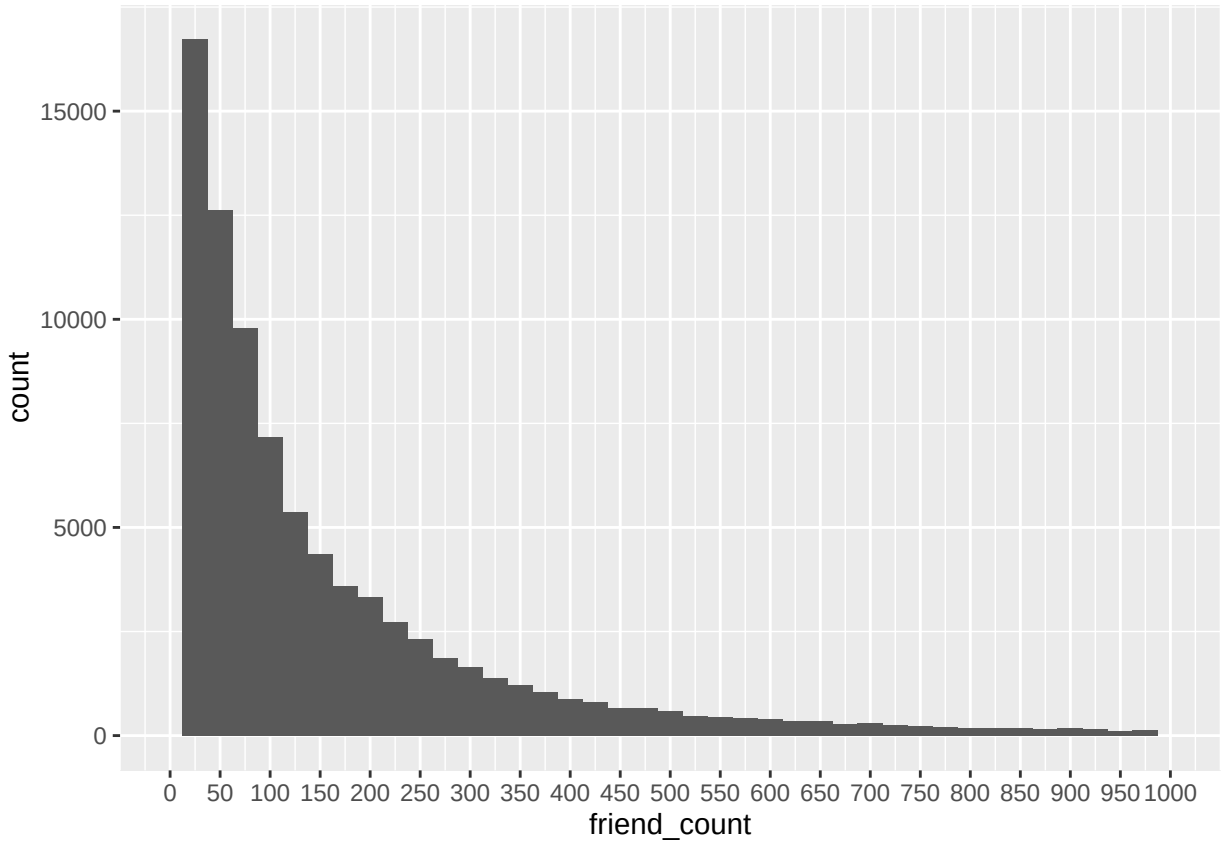
```
## Warning: Removed 2951 rows containing non-finite values (stat_bin).
```



Explorando a largura dos eixos

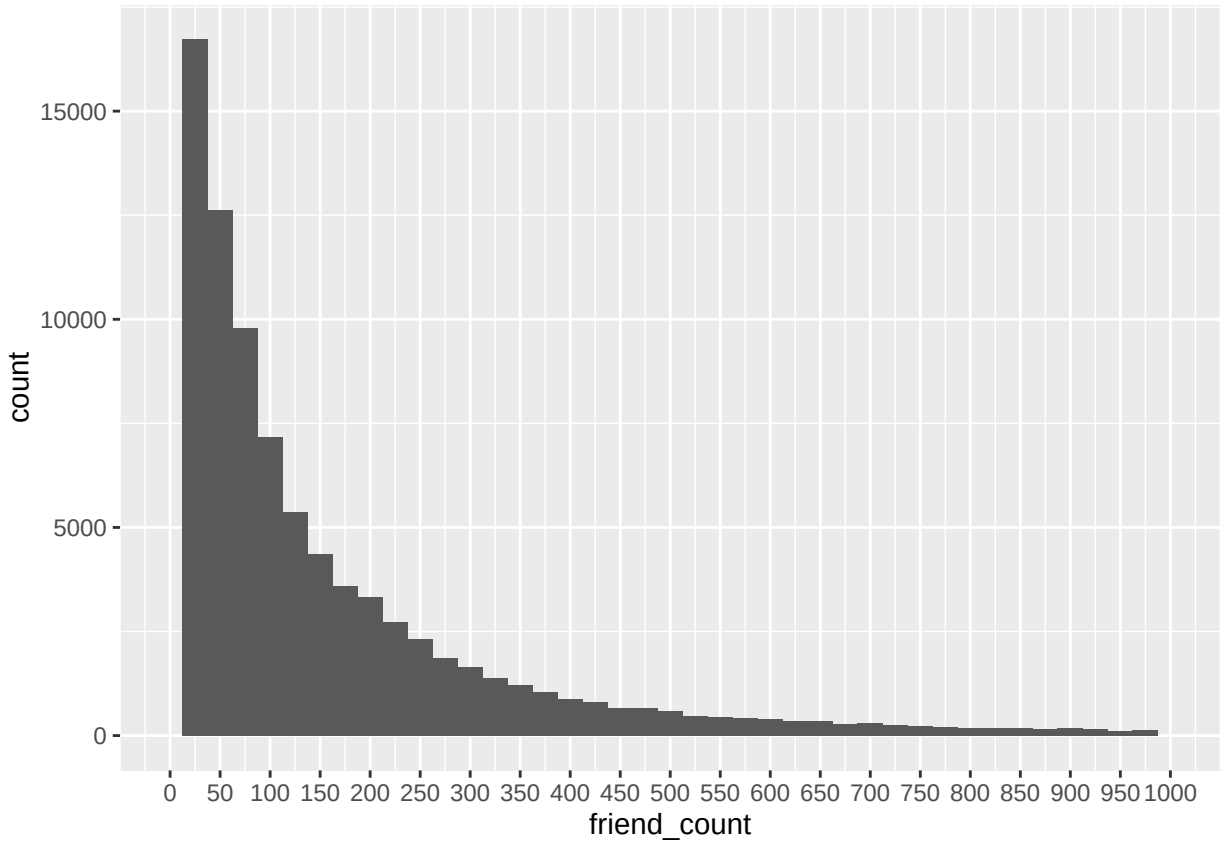
```
ggplot(aes(x = friend_count), data = pf) +
  geom_histogram(binwidth = 25) +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50))
```

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




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

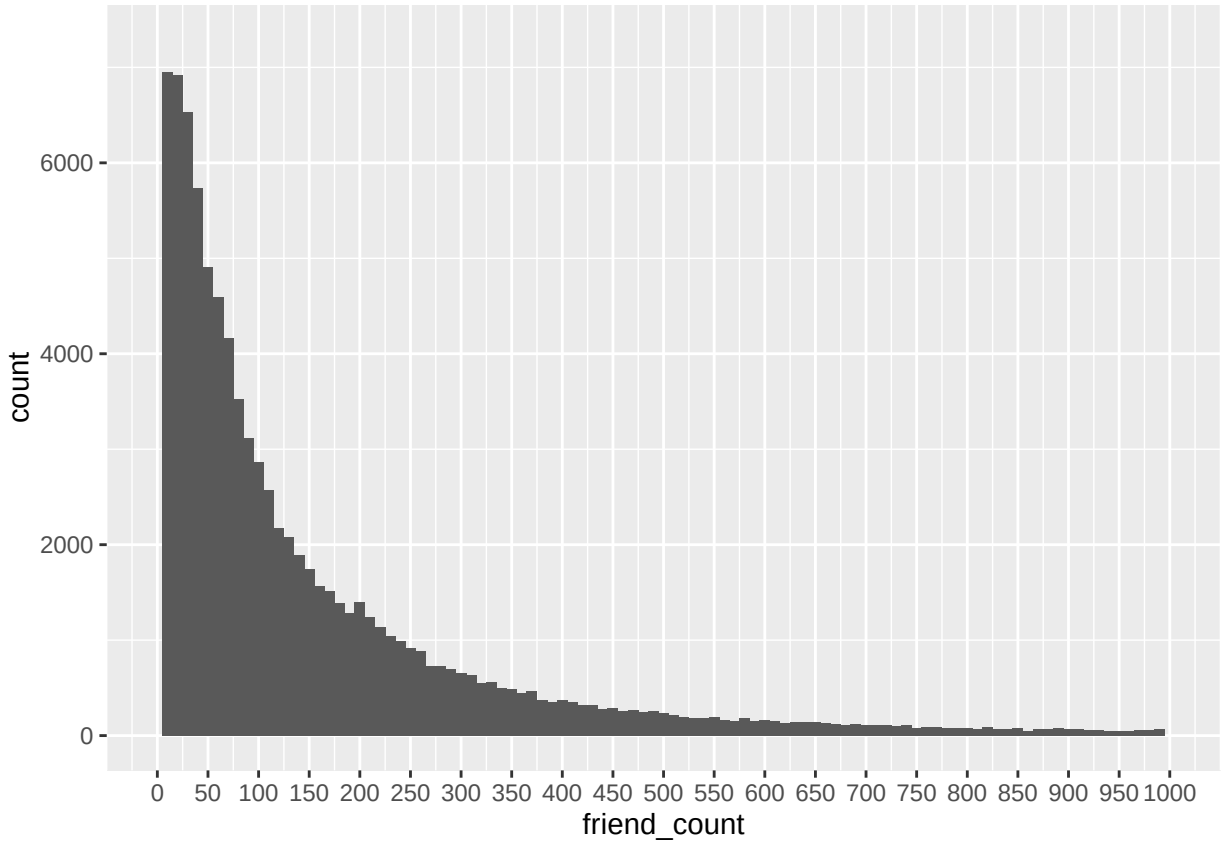
```
## Warning: Removed 2951 rows containing non-finite values (stat_bin).
```



Faceting Friend Count

```
qplot(x = friend_count, data = pf, binwidth = 10) +
  scale_x_continuous(limits = c(0, 1000),
                     breaks = seq(0, 1000, 50))
```

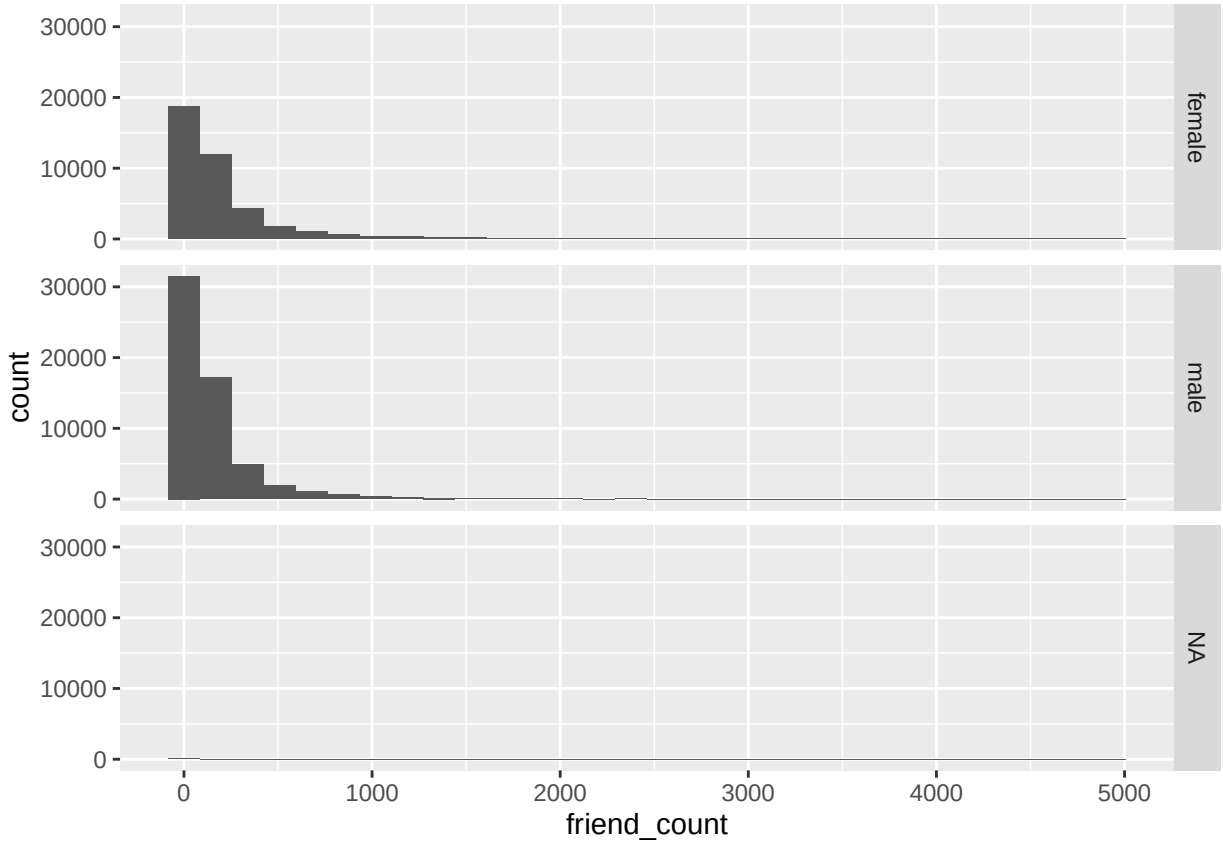
```
## Warning: Removed 2951 rows containing non-finite values (stat_bin).
```



Estatísticas por gênero

```
qplot(x = friend_count, data = pf) +
  facet_grid(gender ~ .)
```

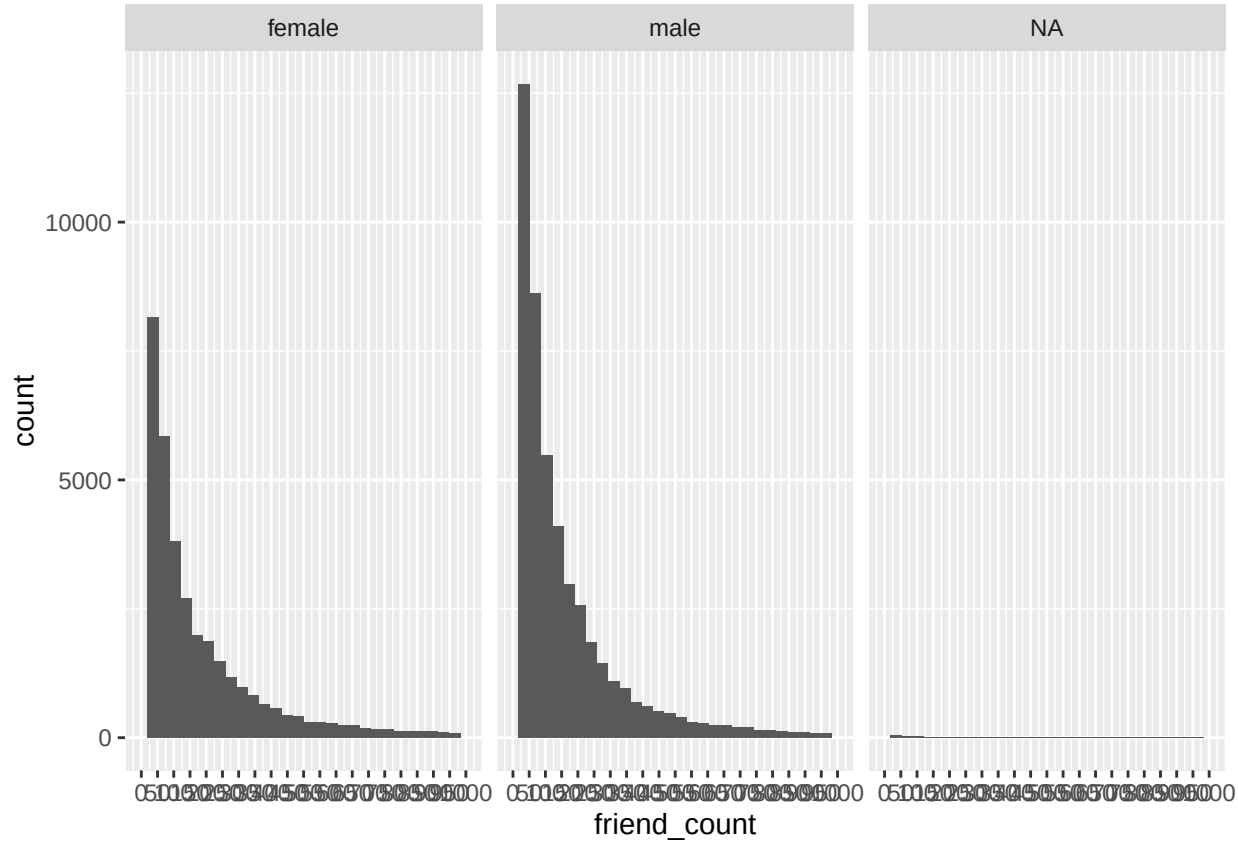
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
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).
```

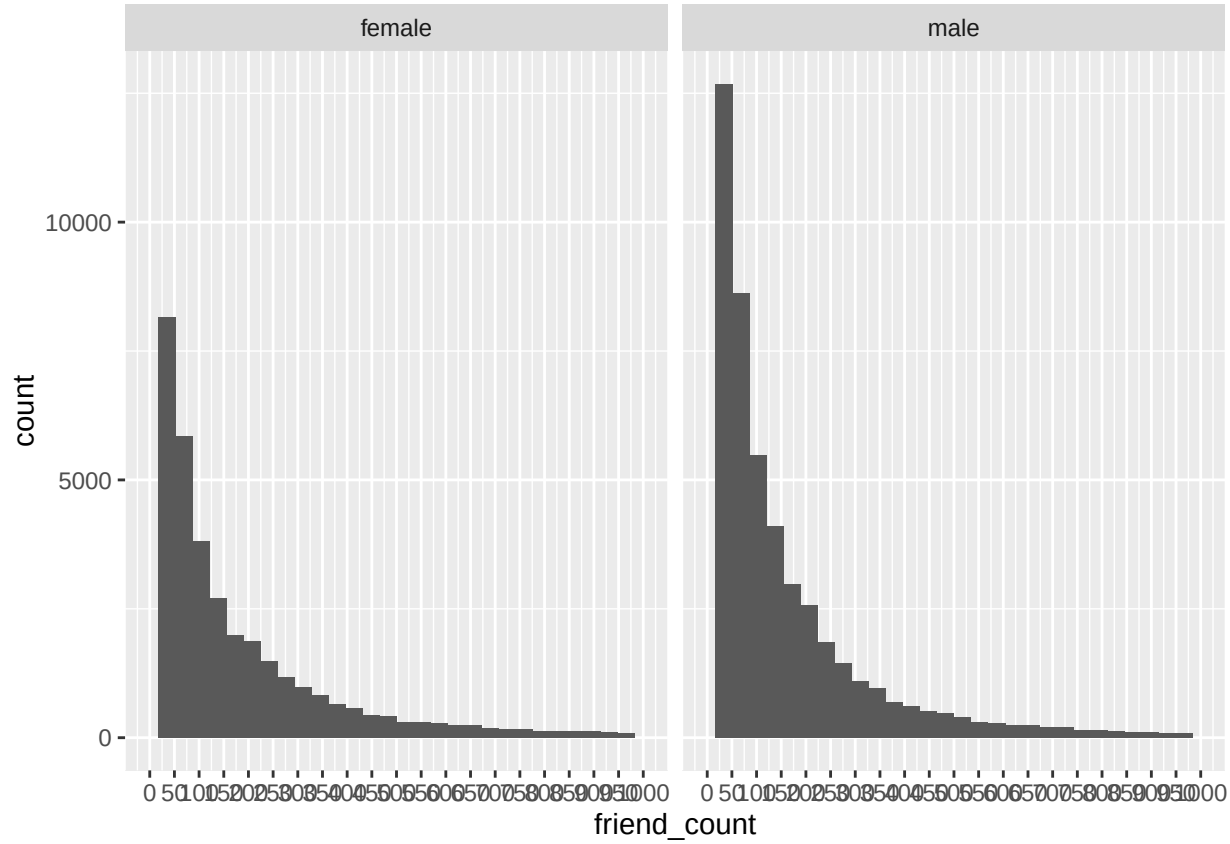


Omitindo vários NA

```
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).
```



Estatísticas por gênero

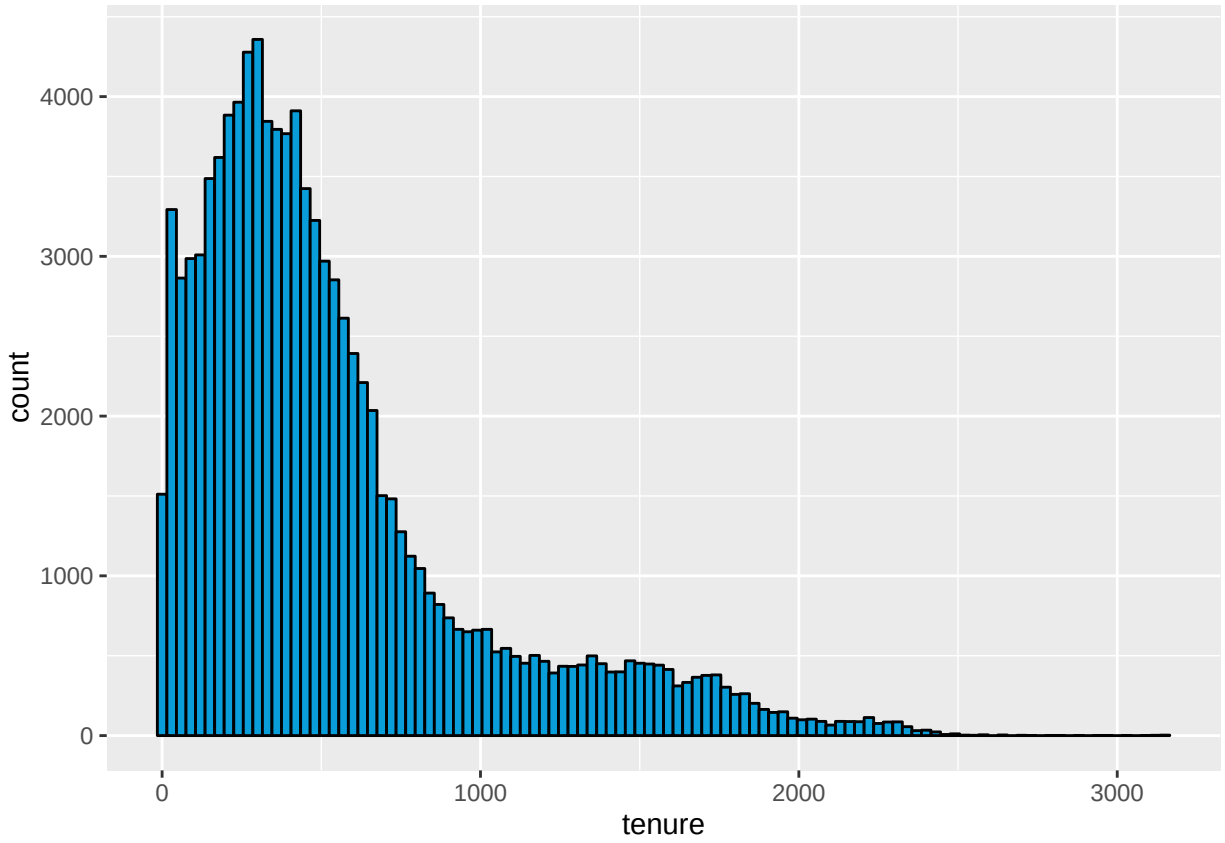
```
by(pf$friend_count, pf$gender, summary)
```

```
## 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
```

Estatística de duração e quantos dias de uso dos usuários

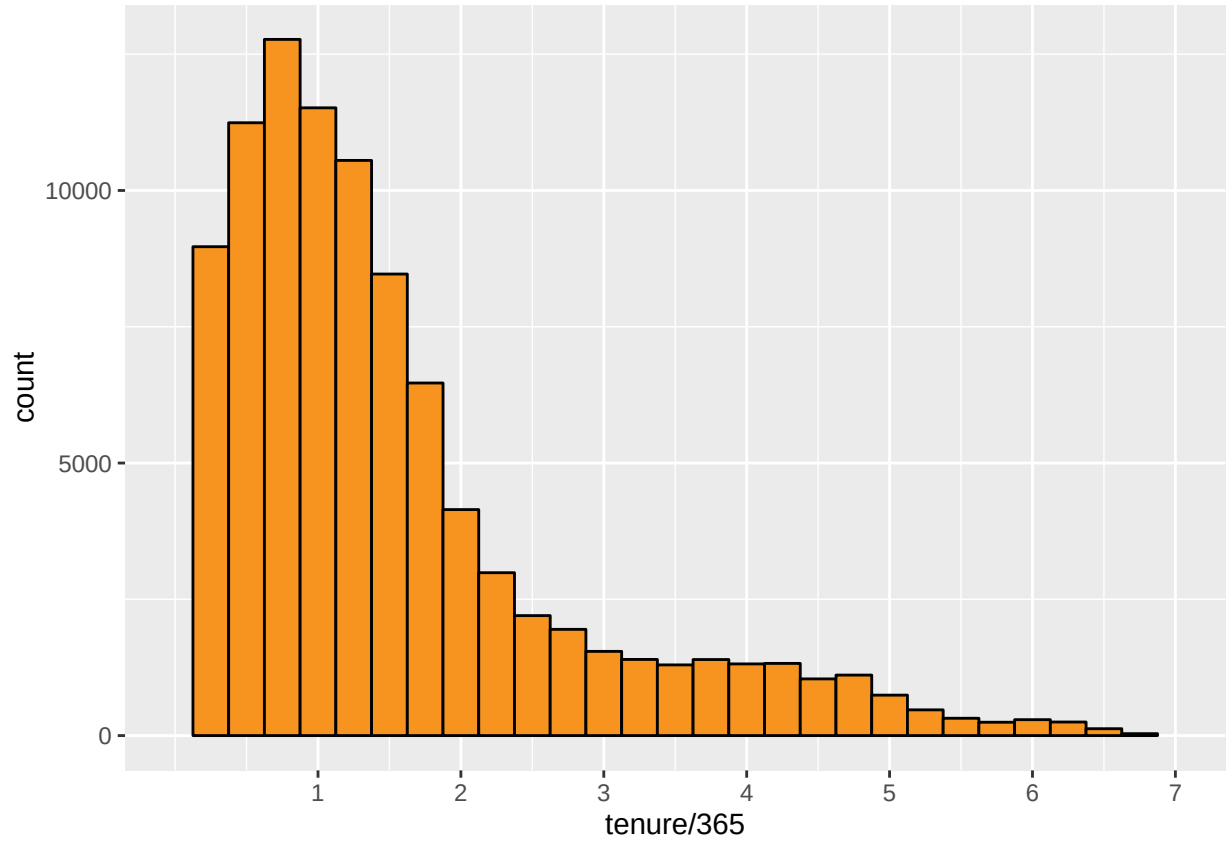
```
ggplot(aes(x = tenure), data = pf) +
   geom_histogram(binwidth = 30, color = 'black', fill = '#099DD9')
```

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




```
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).
```

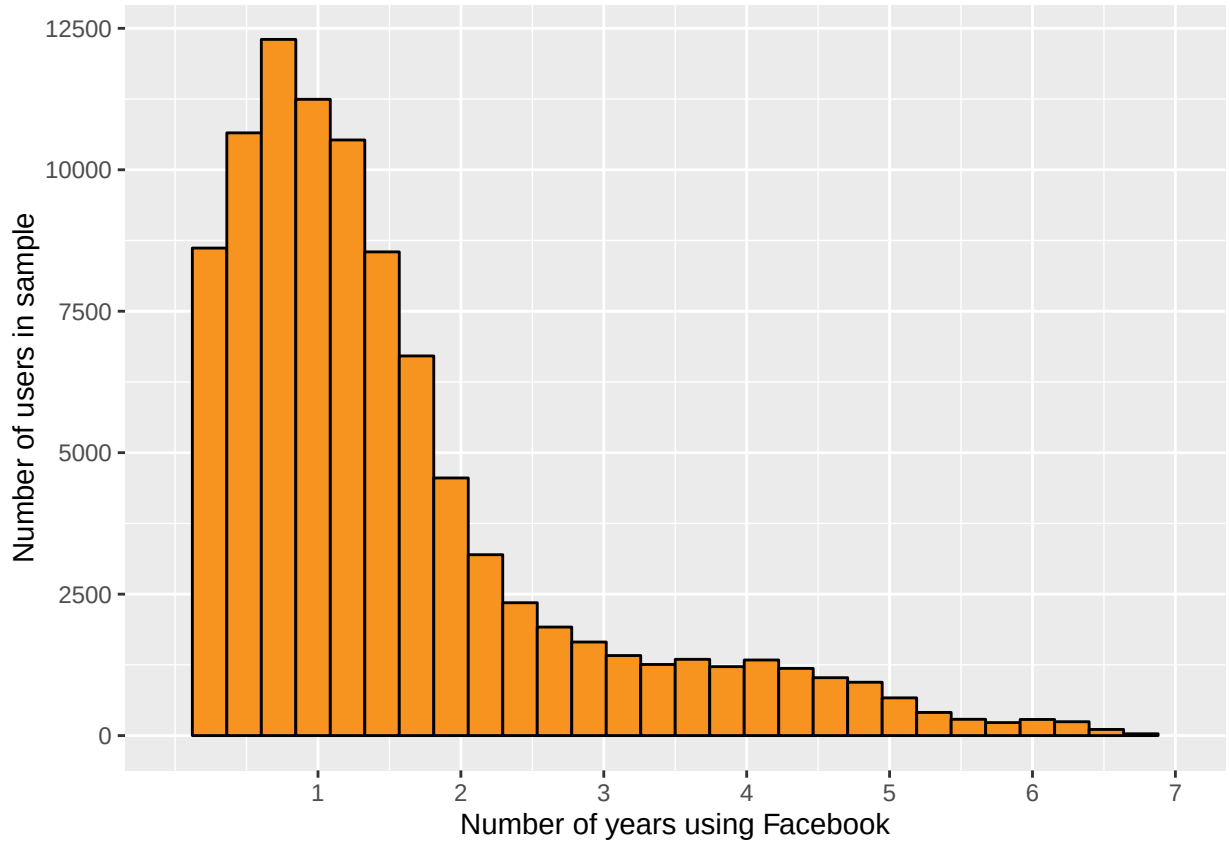


Alterando labels dos plots

```
ggplot(aes(x = tenure / 365), data = pf) +
  geom_histogram(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')
```

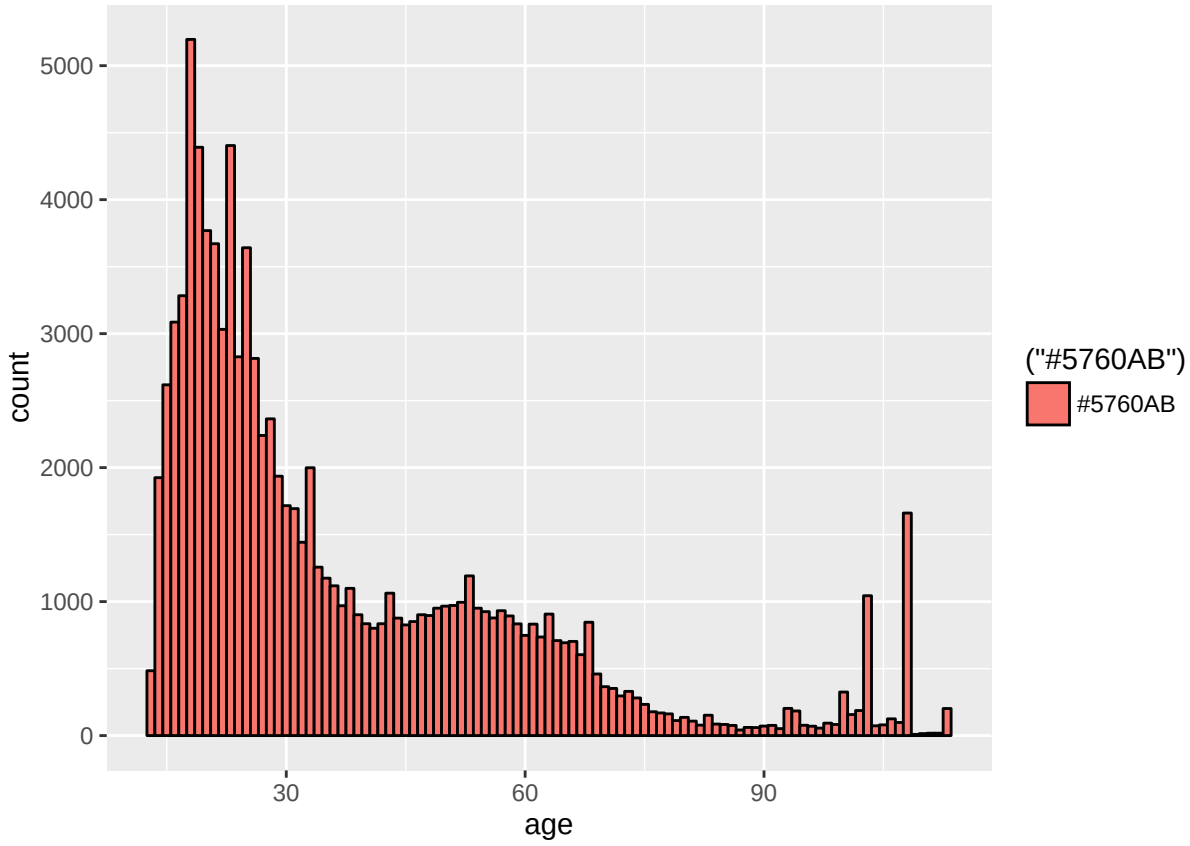
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



Idade dos usuários

```
qplot(x = age, data = pf, binwidth = 1,
      color = I("black"), fill = ("#5760AB"))
```



Visualização de 3 plots em uma imagem

```
install.packages('gridExtra')
```

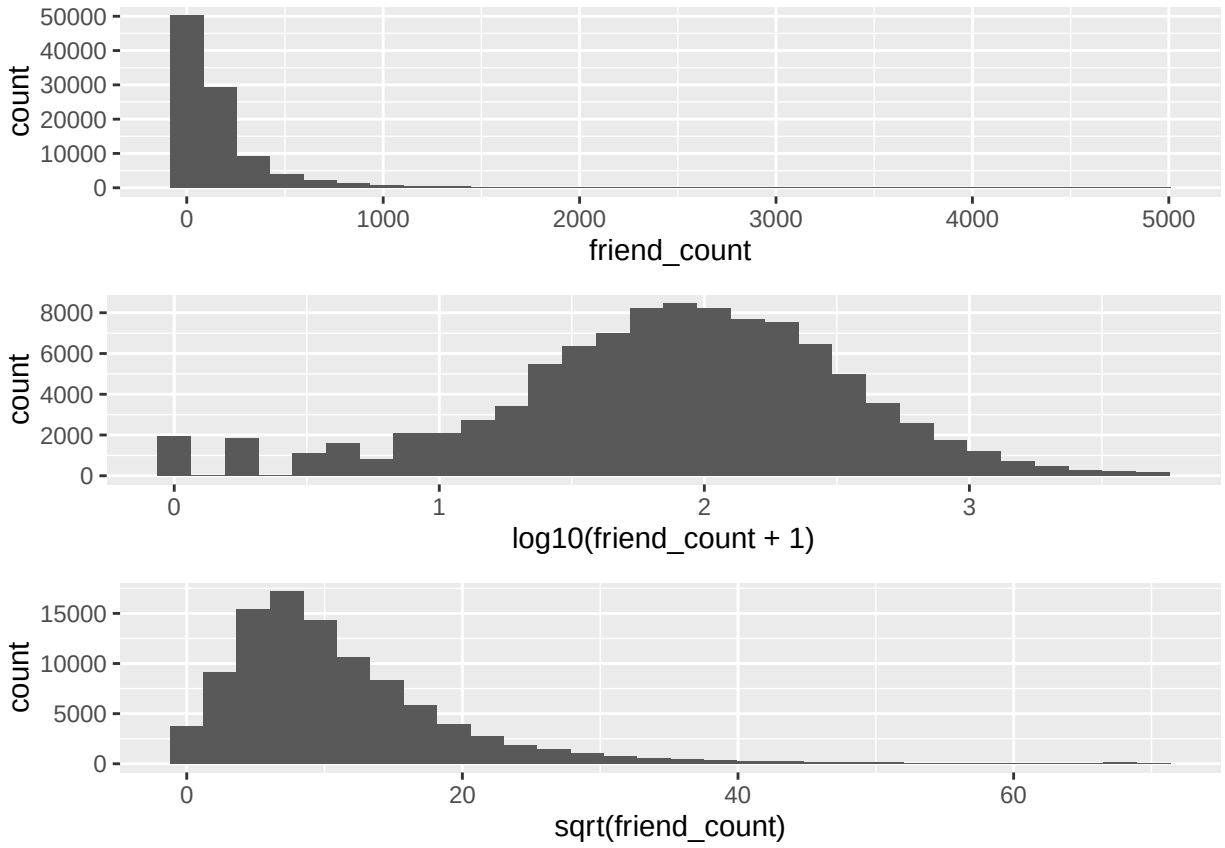
```
## Installing package into 'C:/Users/leo/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)
```

```
## package 'gridExtra' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\leo\AppData\Local\Temp\RtmpQlNEru\downloaded_packages
```

```
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`.
```

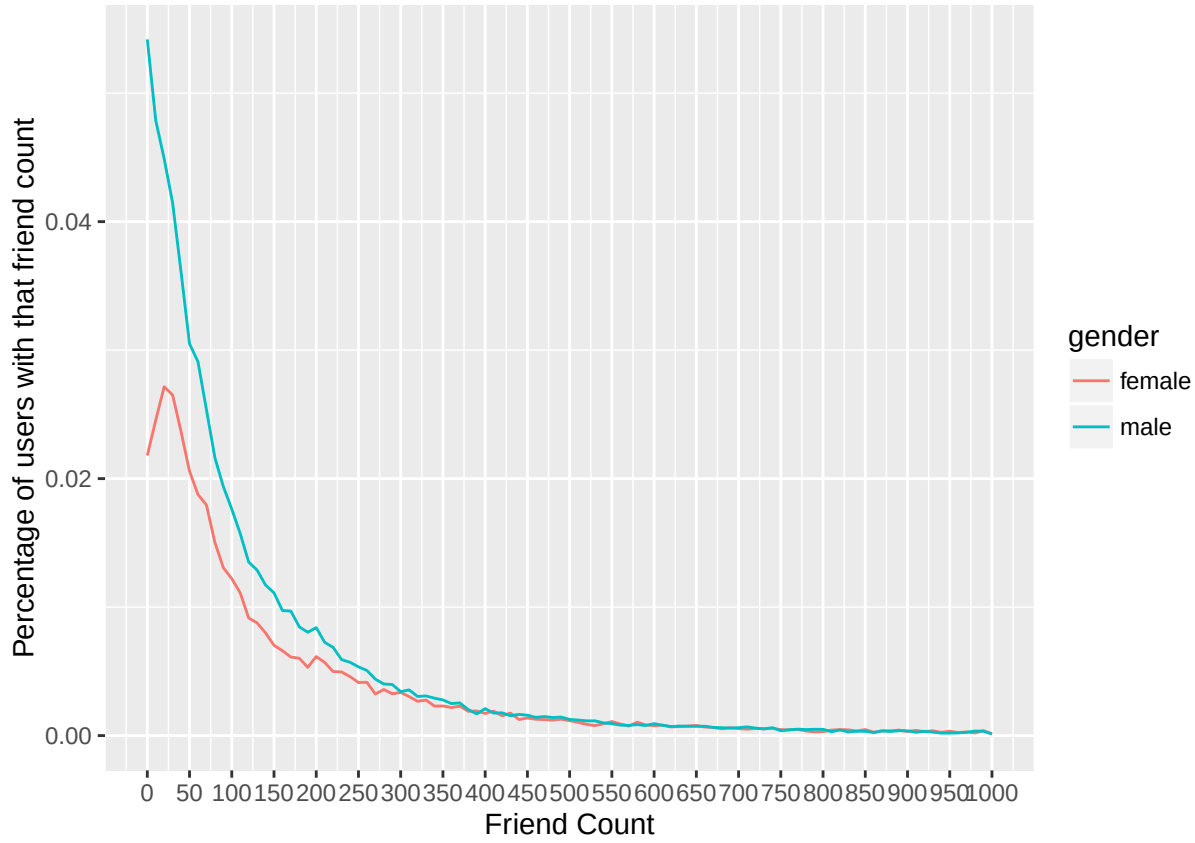


Frequencia de poligonos

```
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('Percentage 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).
```



Likes por gênero

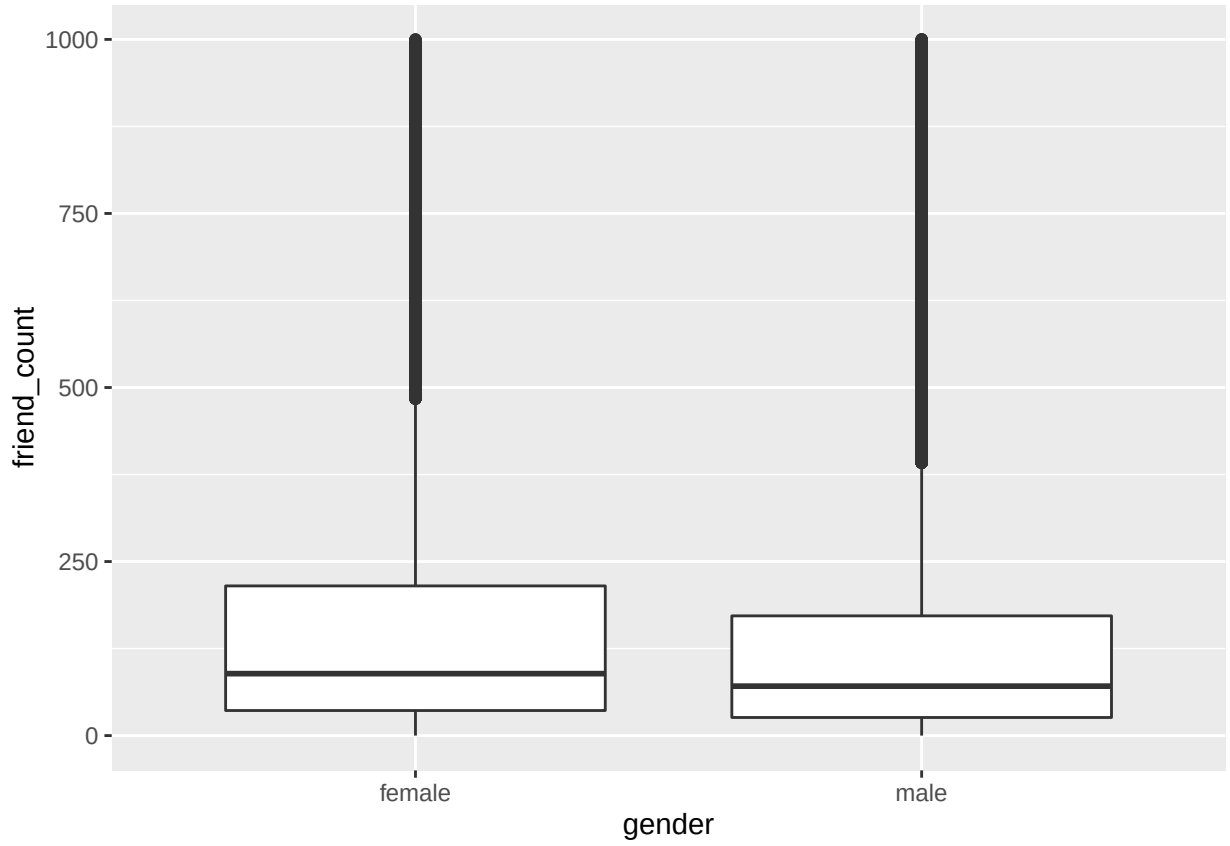
```
by(pf$www_likes, pf$gender, sum)
```

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

Box Plots

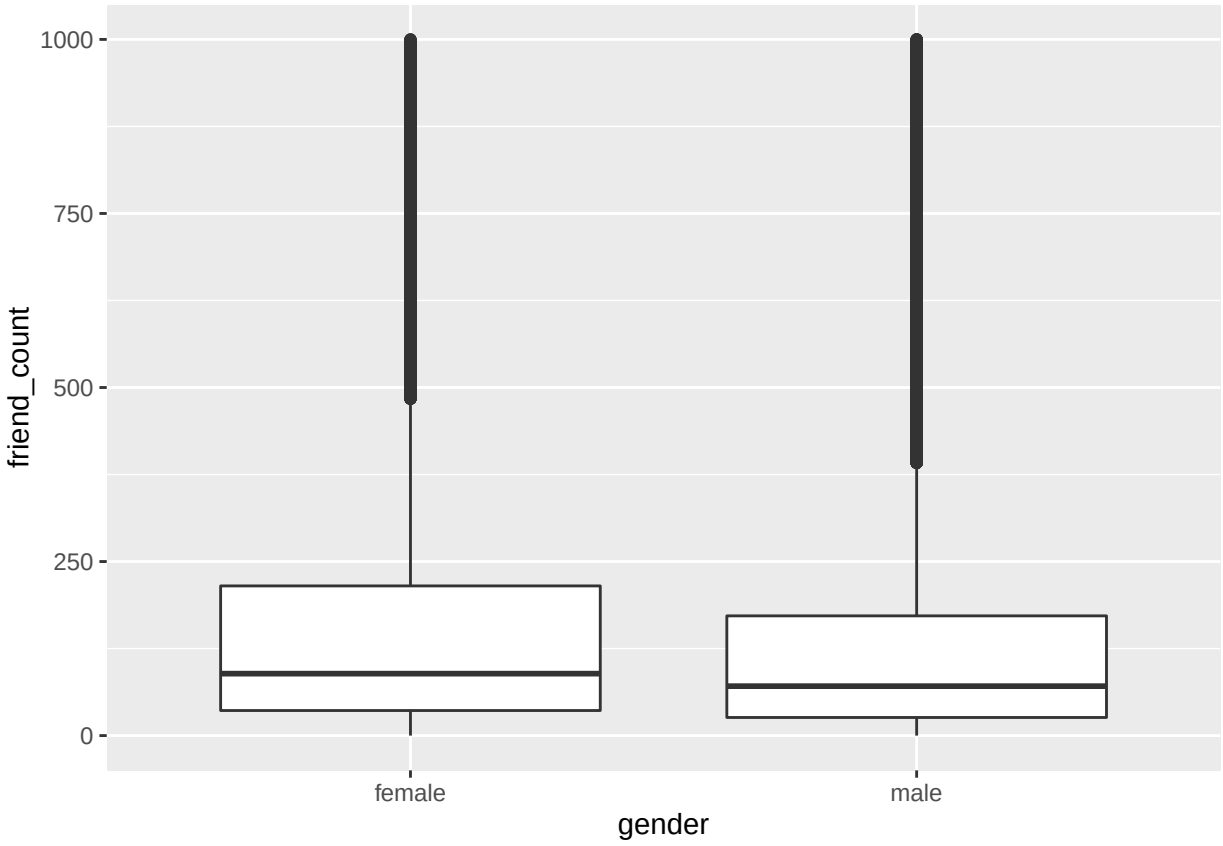
```
qplot(x = gender, y = friend_count,
      data = subset(pf, !is.na(gender)),
      geom="boxplot",
      ylim = c(0, 1000))
```

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




```
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).
```


1000
750
500
250
0
friend_count
female
male
gender