facebook_dataset

Suchitra

2/19/2017

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Loading library

```
#install.packages("ggplot2")
library(ggplot2)

#install.packages("gridExtra")
library(gridExtra)

#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
getwd()
```

```
## [1] "/Users/suchitra/Desktop/Suchitra/R/Data Analysis using R/facebook_dataset"
```

```
list.files()
```

```
## [1] "facebook_dataset.Rproj" "fb.html"
## [3] "fb.Rmd"                 "pseudo_facebook.tsv"
```

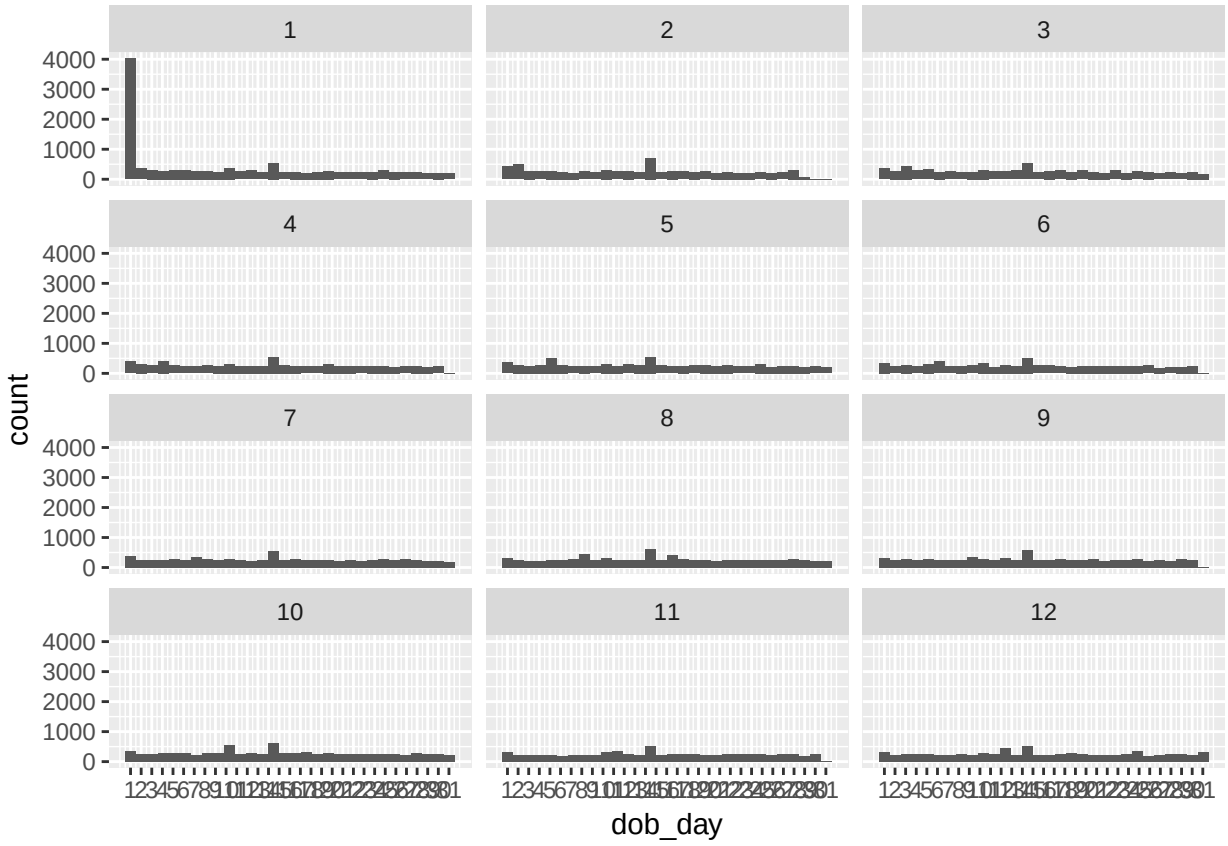
```
fb <- read.csv("pseudo_facebook.tsv", sep="\t")
names(fb)
```

```
##  [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"
```

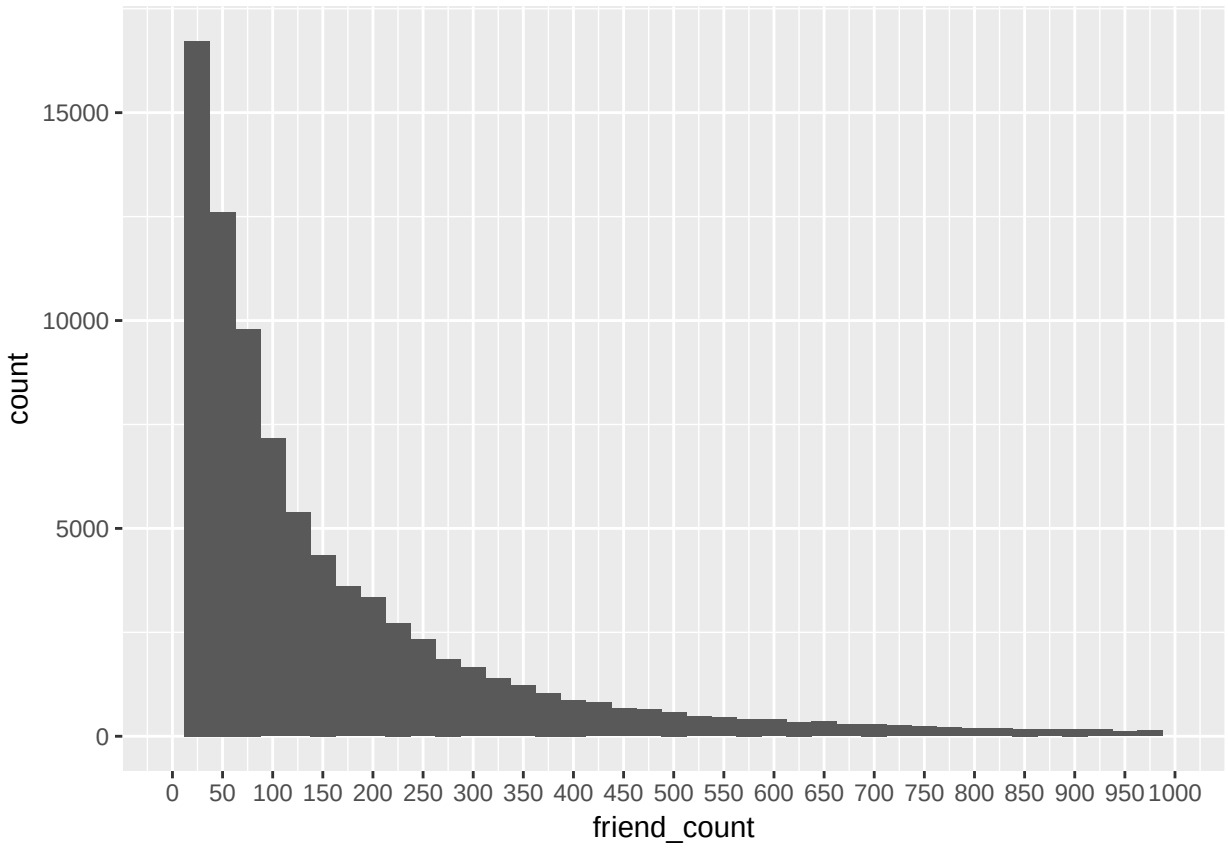
```
qplot(data= fb, x=dob_day) +
   scale_x_continuous(breaks=1:31)+
  facet_wrap(~dob_month, ncol = 3)
```

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




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

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




```
  facet_wrap(~gender)
```

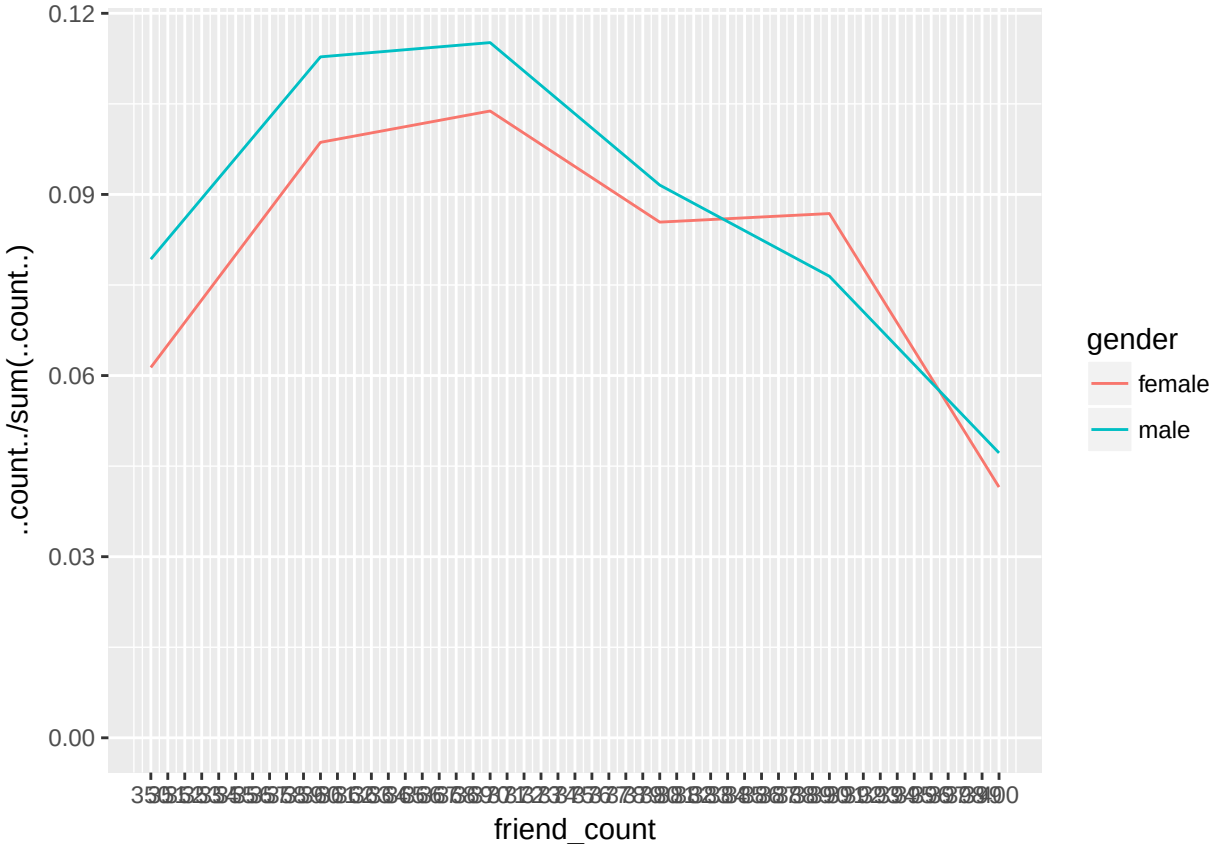
```
## <ggproto object: Class FacetWrap, Facet>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map: function
##     map_data: function
##     params: list
##     render_back: function
##     render_front: function
##     render_panels: function
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train: function
##     train_positions: function
##     train_scales: function
##     super:  <ggproto object: Class FacetWrap, Facet>
```

```
  qplot( data= subset(fb, !is.na(gender)), x= friend_count,
         y= ..count../sum(..count..),
         binwidth=10,
         geom = "freqpoly", color= gender)+
```

```
  scale_x_continuous(limits = c(350,400), breaks= seq(350,400,1))
```

```
## Warning: Removed 96709 rows containing non-finite values (stat_bin).
```

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



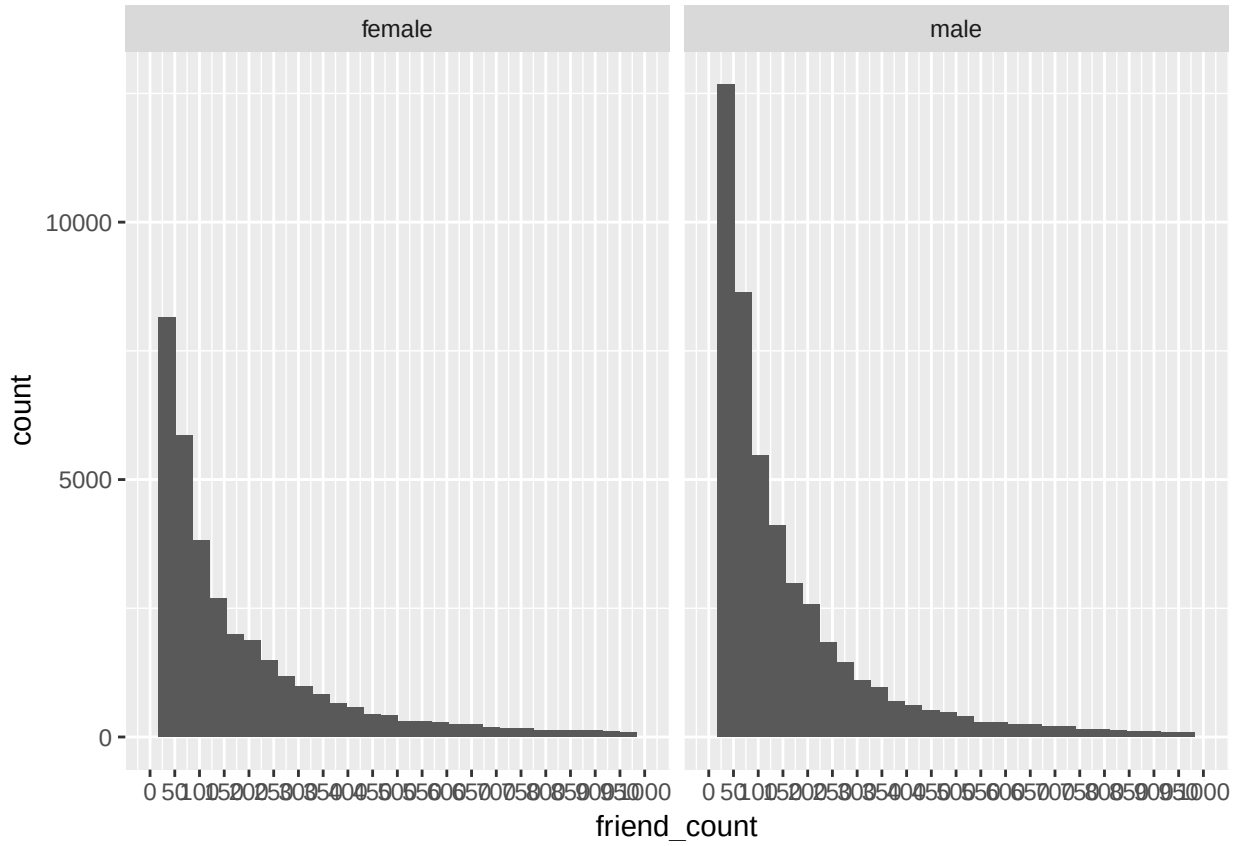

```
  facet_wrap(~gender)
```

```
## <ggproto object: Class FacetWrap, Facet>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map: function
##     map_data: function
##     params: list
##     render_back: function
##     render_front: function
##     render_panels: function
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train: function
##     train_positions: function
##     train_scales: function
```

```
##     super:  <ggproto object: Class FacetWrap, Facet>
```

```
  ggplot(aes(x = friend_count), data = subset(fb, !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).
```




```
  table(fb$gender)
```

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

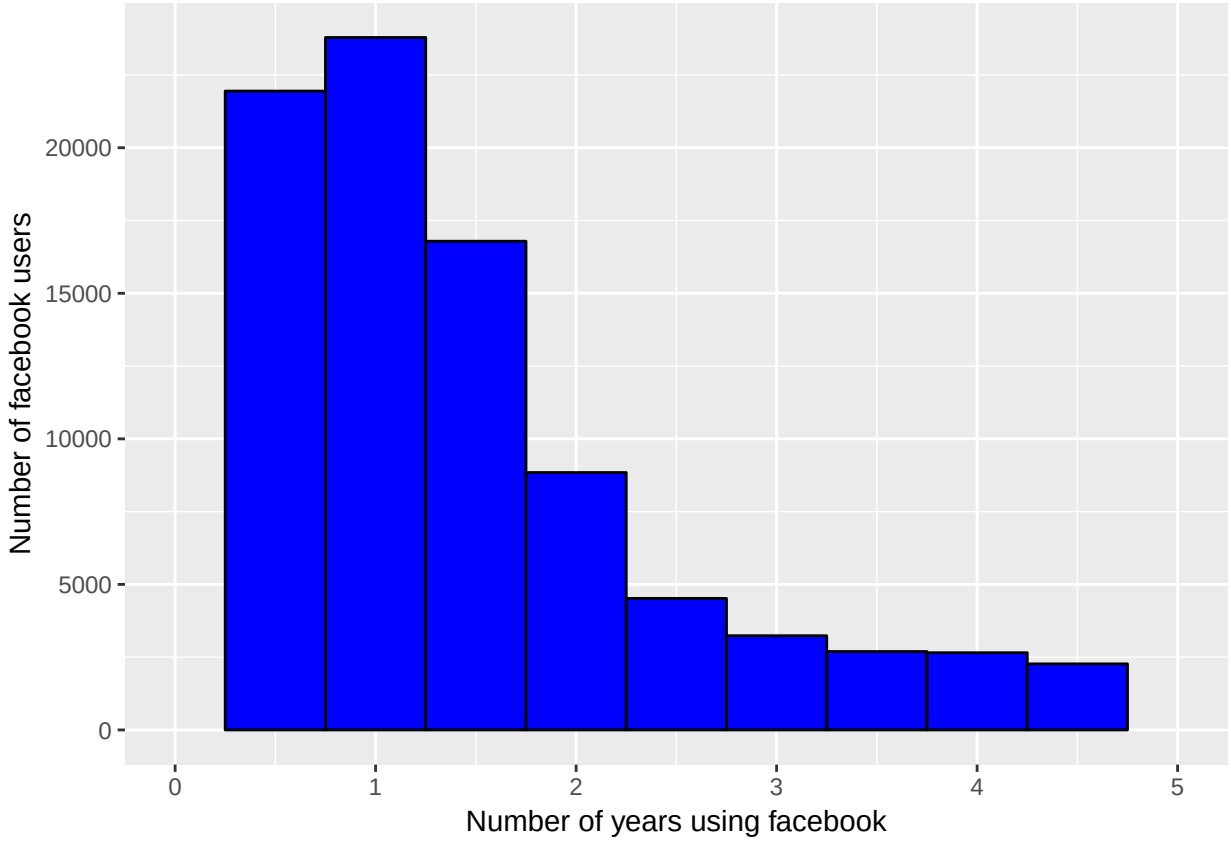
```
  by(fb$friend_count, fb$gender, summary)
```

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

Tenure

```
qplot(data=fb, x= tenure/365, binwidth=0.5,
      xlab="Number of years using facebook",
      ylab="Number of facebook users",
      color=I('black'), fill=I('blue')) +
  scale_x_continuous(limits = c(0,5.0), breaks= seq(0,5.0,1.0))
```

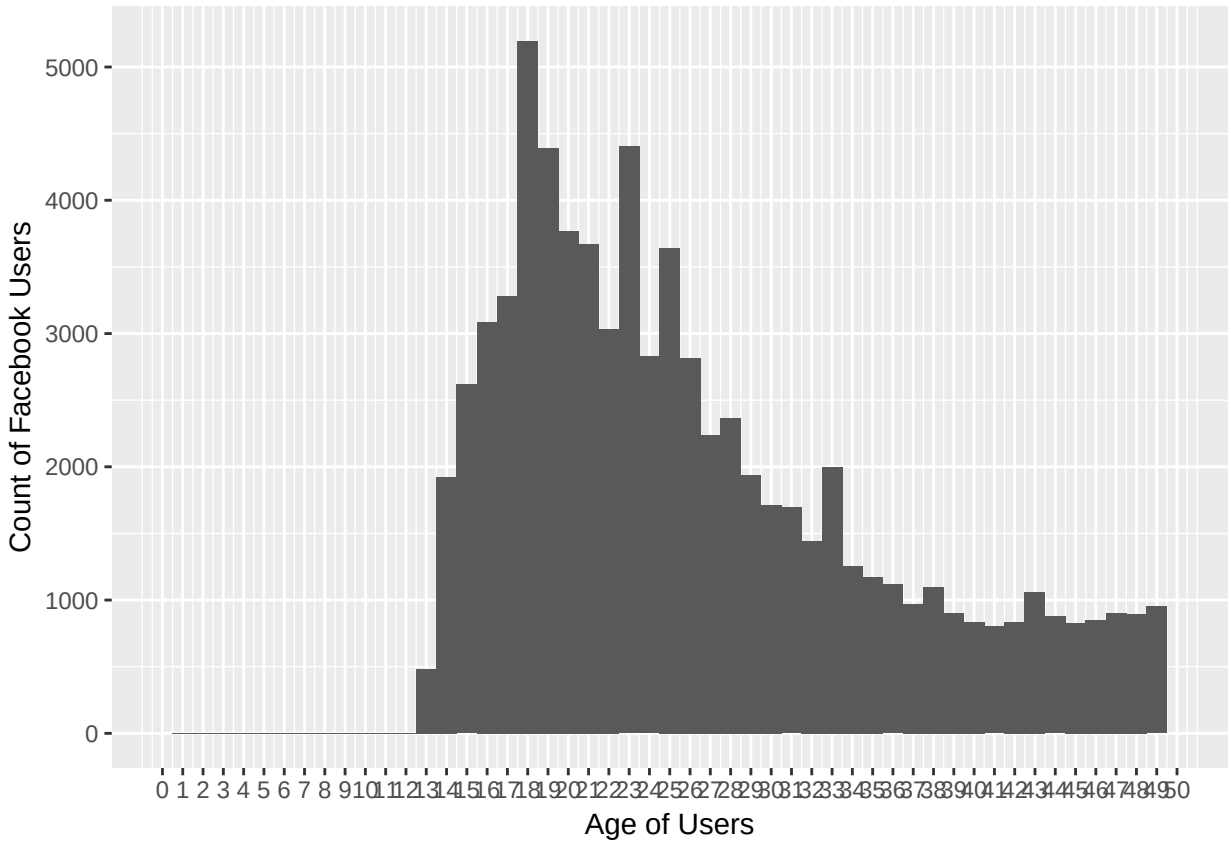
```
## Warning: Removed 2114 rows containing non-finite values (stat_bin).
```



Age

```
qplot(data= fb, x=age, binwidth= 1,
  xlab="Age of Users",
  ylab="Count of Facebook Users") +
  scale_x_continuous( limit= c(0,50), breaks= seq(0,50,1))
```

```
## Warning: Removed 24146 rows containing non-finite values (stat_bin).
```



##Transforming Data

```
summary(fb$friend_count)
```

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

```
summary(log10(fb$friend_count + 1))
```

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

```
summary(sqrt(fb$friend_count))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   5.568   9.055  11.090  14.350  70.160
```

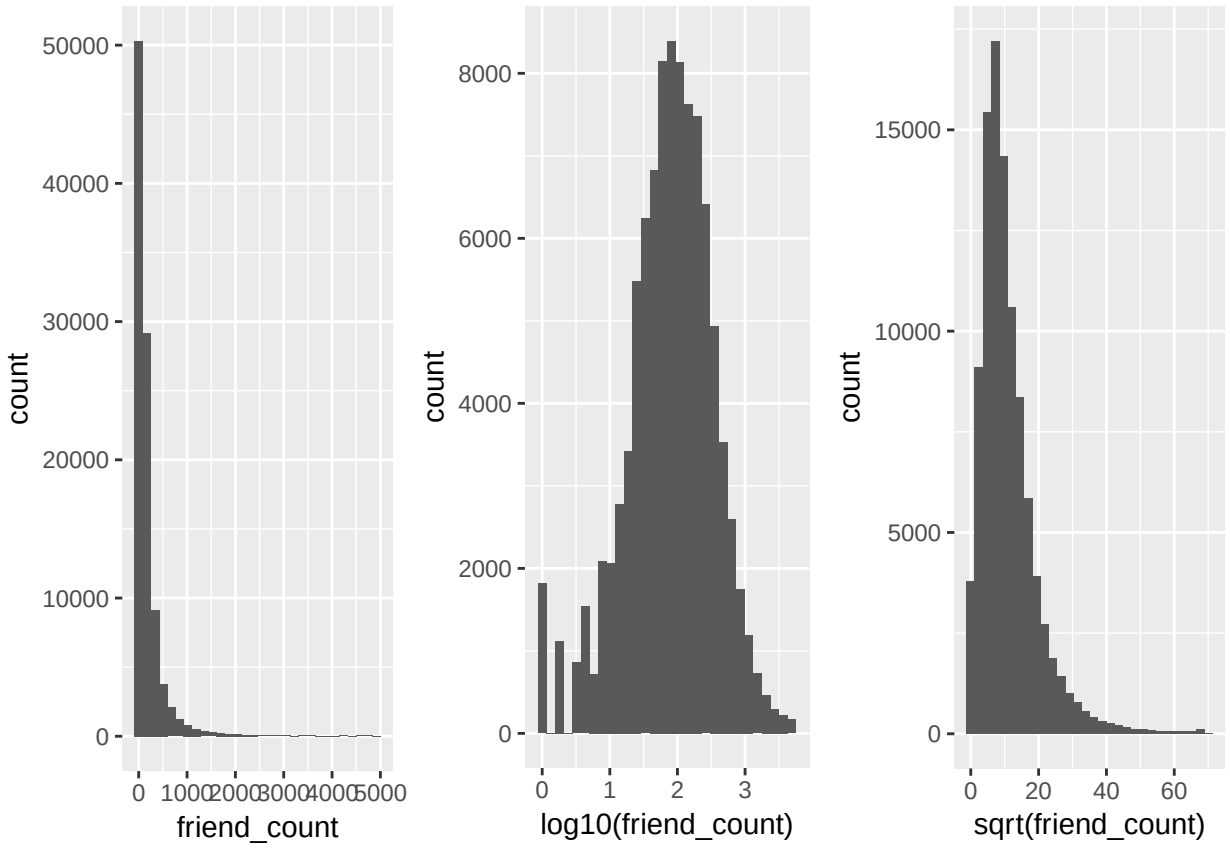
```
p1=qplot(data= fb, x= friend_count)

p2=qplot(data=fb, x=log10(friend_count))
p3=qplot(data=fb, x=sqrt(friend_count))
grid.arrange(p1,p2,p3,ncol=3)
```

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



www likes

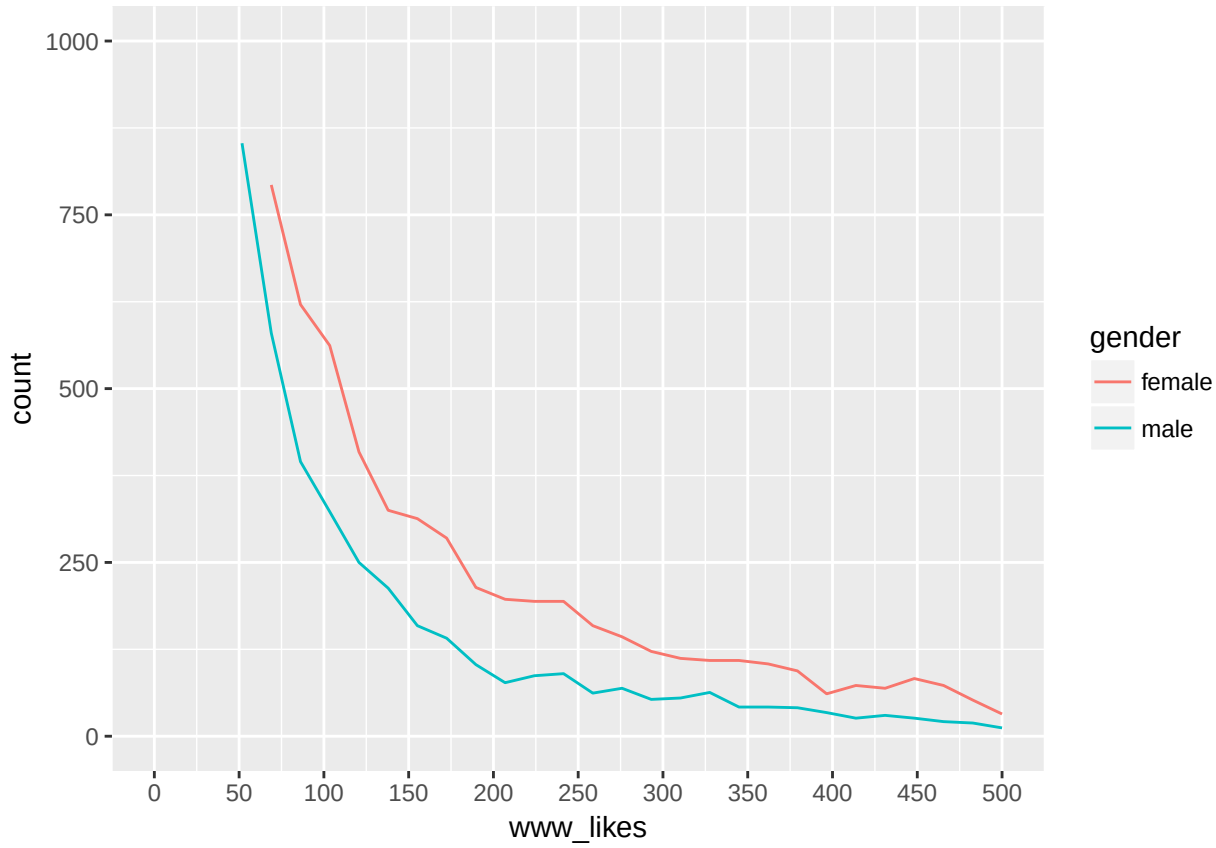
```
qplot(data= subset(fb,!is.na(gender)), x= www_likes, geom="freqpoly", color=gender)+
  scale_x_continuous()+
  scale_x_log10()+
  scale_x_continuous(limit= c(0,500), breaks= seq(0,500, 50))+
  scale_y_continuous(limit= c(0,1000))
```

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

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

```
## Warning: Removed 2279 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 11 rows containing missing values (geom_path).
```




```
facet_wrap(~gender)
```

```
## <ggproto object: Class FacetWrap, Facet>
##     compute_layout: function
##     draw_back: function
##     draw_front: function
##     draw_labels: function
##     draw_panels: function
##     finish_data: function
##     init_scales: function
##     map: function
##     map_data: function
##     params: list
##     render_back: function
##     render_front: function
##     render_panels: function
##     setup_data: function
##     setup_params: function
##     shrink: TRUE
##     train: function
##     train_positions: function
##     train_scales: function
##     super:  <ggproto object: Class FacetWrap, Facet>
```

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

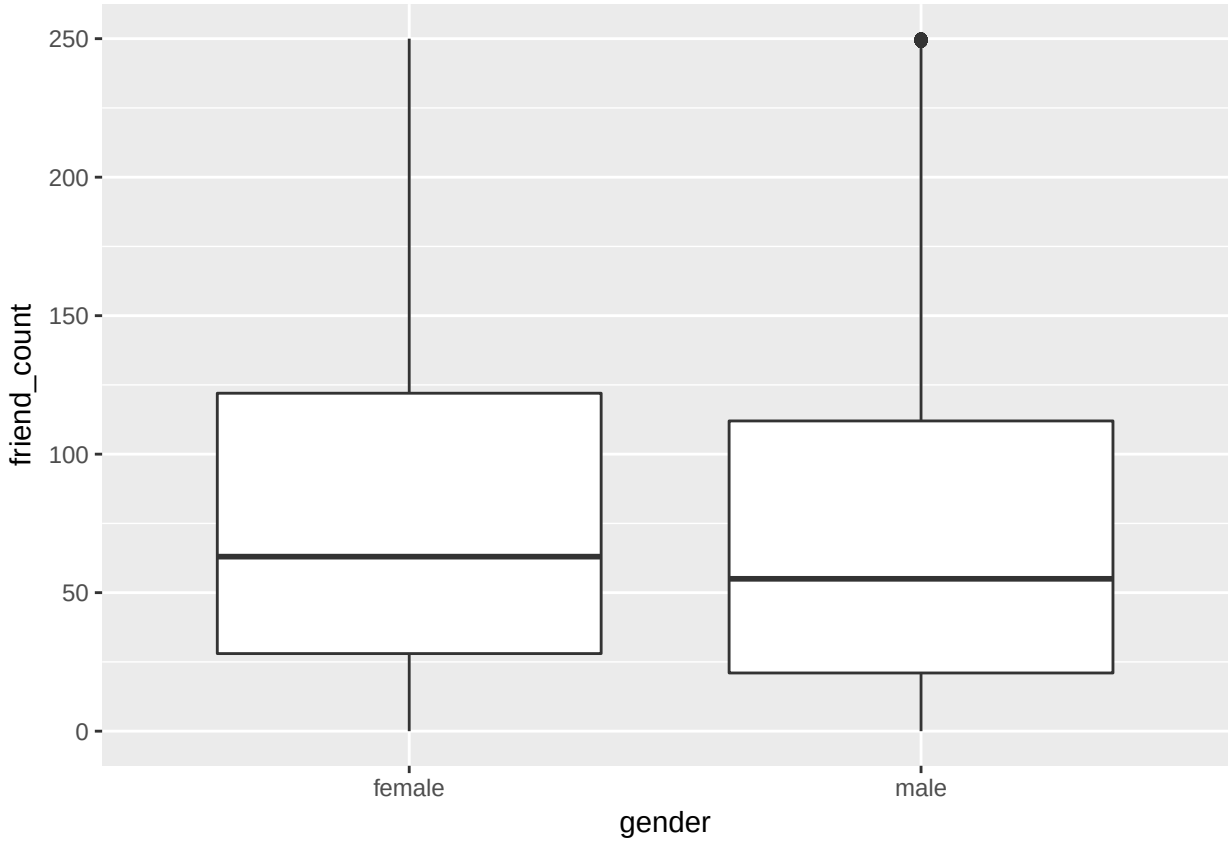
```
## fb$gender: female
## [1] 3507665
```

```
## ------------------------------------------------------------
## fb$gender: male
## [1] 1430175
```

Boxplot

```
library(ggplot2)
qplot(x= gender, y= friend_count, data= subset(fb, !is.na(gender)) ,
                                                  geom='boxplot') +
  scale_y_continuous(limit=c(0,250))
```

```
## Warning: Removed 19870 rows containing non-finite values (stat_boxplot).
```




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

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

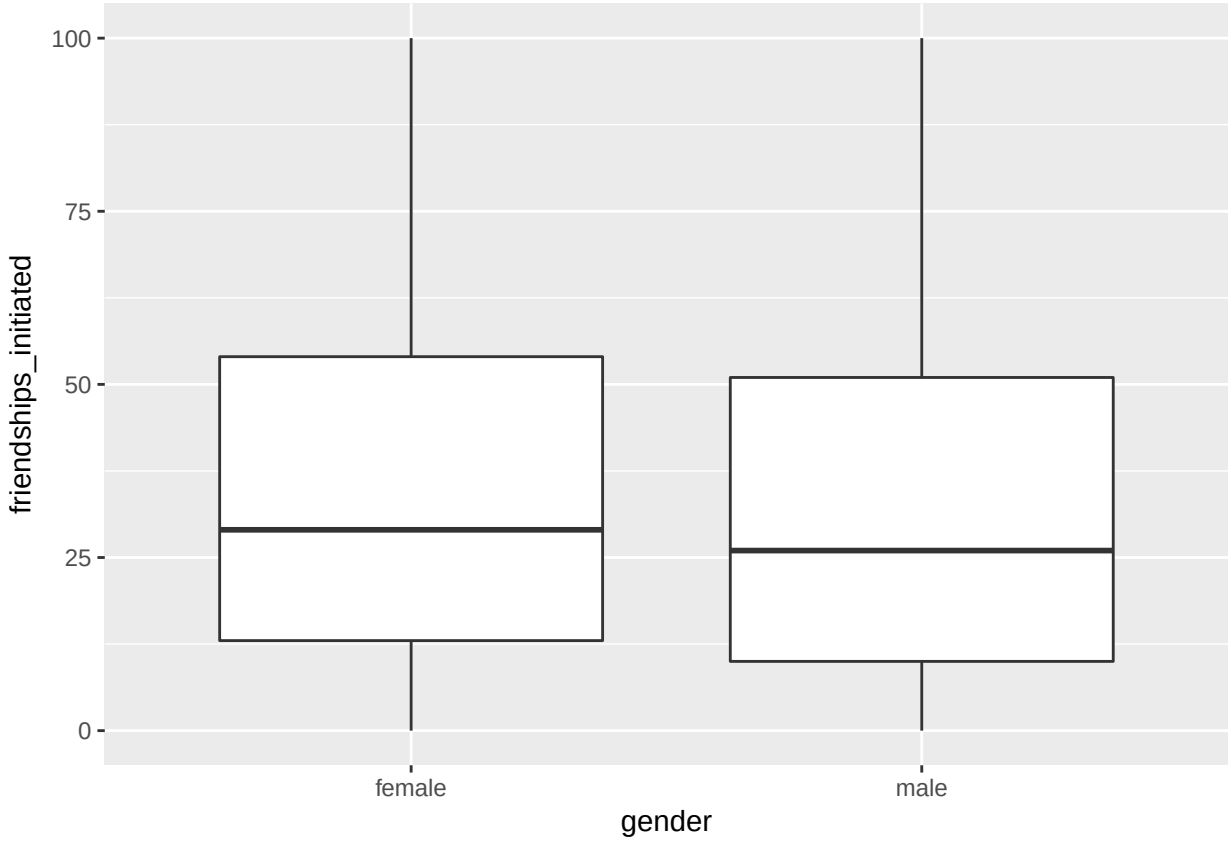
```
names(fb)
```

```
##  [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"
```

```
qplot(y= friendships_initiated, x= gender, data = subset(fb, !is.na(gender)), geom="boxplot") +
  scale_y_continuous(limit=c(0,100))
```

```
## Warning: Removed 28229 rows containing non-finite values (stat_boxplot).
```




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

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

Transformation into binaries

```
fb <- read.csv("pseudo_facebook.tsv", sep="\t")
summary(fb$mobile_likes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     4.0   106.1    46.0 25110.0
```

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

```
##    Mode   FALSE    TRUE    NA's
## logical   35056   63947       0
```

```
mobile_chekin <- NA
fb$mobile_chekin <- ifelse(fb$mobile_likes > 0,1,0)
fb$mobile_chekin <- factor(fb$mobile_chekin)
summary(fb$mobile_chekin)
```

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

```
# % of people who check in
a <- sum(as.numeric(fb$mobile_chekin))
b <- sum(as.numeric(!fb$mobile_chekin))
```

```
## Warning in Ops.factor(fb$mobile_chekin): '!' not meaningful for factors
```

```
perc <- a/(a+b)
perc
```

```
## [1] NA
```

```
ls()
```

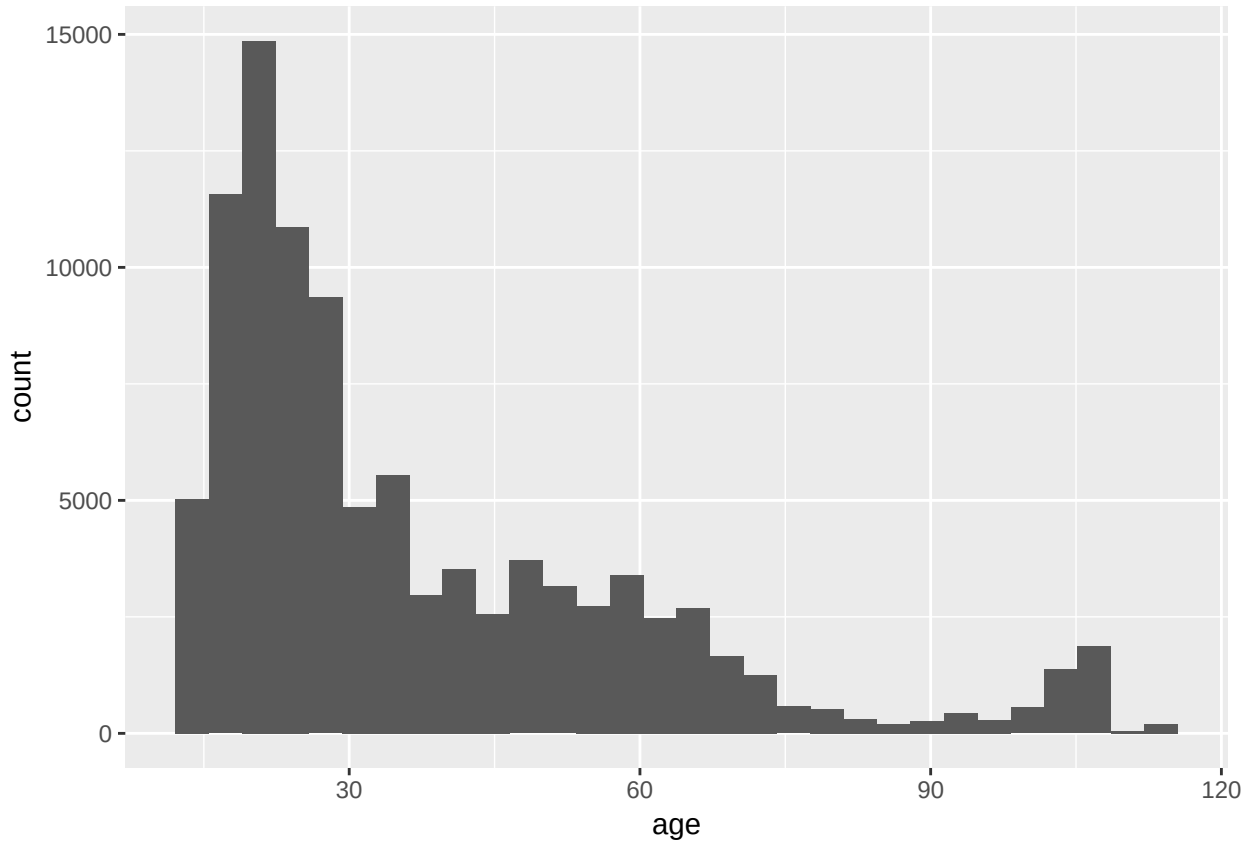
```
## [1] "a"             "b"             "fb"            "mobile_chekin"
## [5] "p1"            "p2"            "p3"            "perc"
```

Multivariate Data

Third Qualitative Variable

```
pf <- read.csv("pseudo_facebook.tsv", sep = "\t")
ggplot(aes(x =age),
       data = subset(pf, !is.na(gender))) + geom_histogram()
```

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




```
#Create a dataframe

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)
```

Plotting Conditional Summaries

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

Reshaping Data

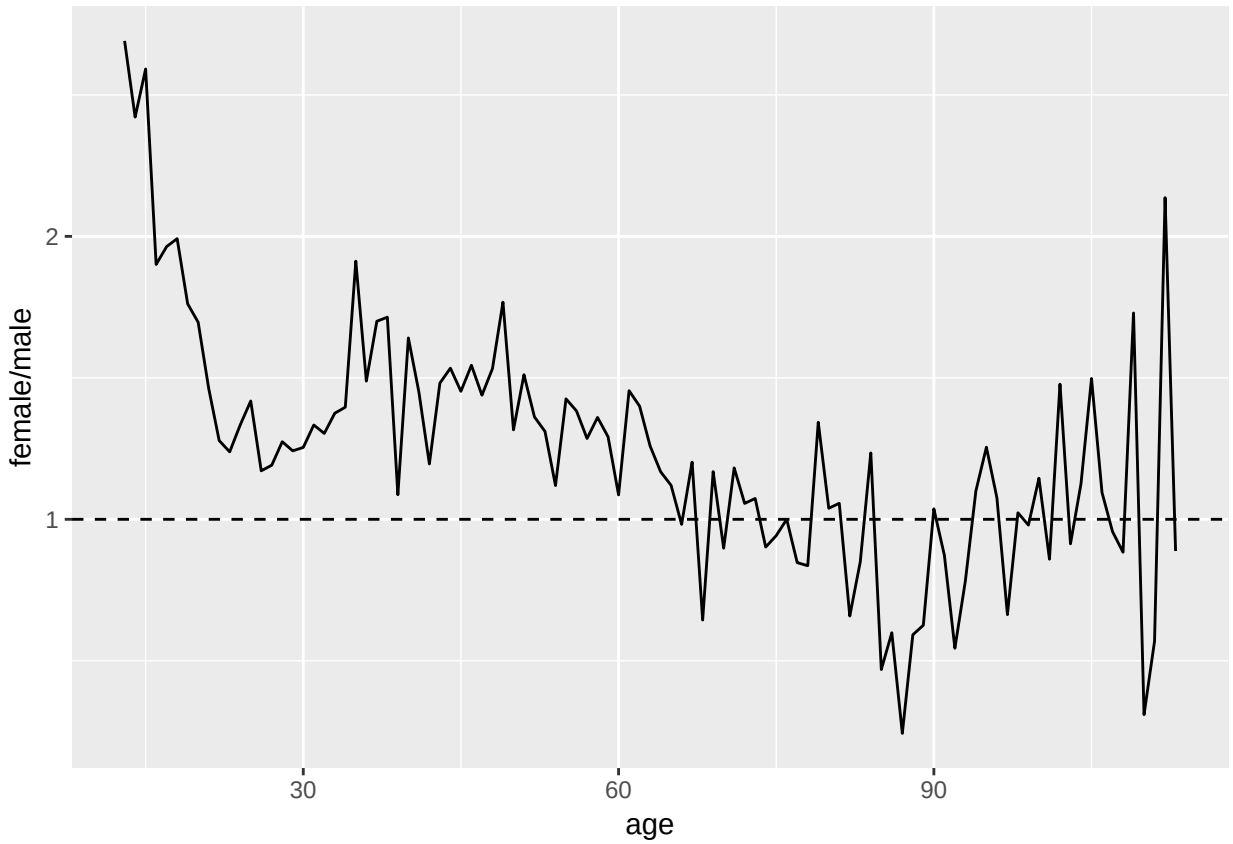
```
#install.packages('reshape2')
library(reshape2)

#dcast function : long to wide
pf.fc_by_age_gender <- dcast(pf.fc_by_age_gender,
                             age ~ gender,
                             value.var = "median_friend_count")

# Nill function : wide to long
```

Ratio Plot

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




```
?geom_hline
```

Third Quantitative Variable

```
year_joined <- floor(2014 - pf$tenure/365)
pf <- pf %>% mutate(year_joined)
summary(pf$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
```

Cut a Variable

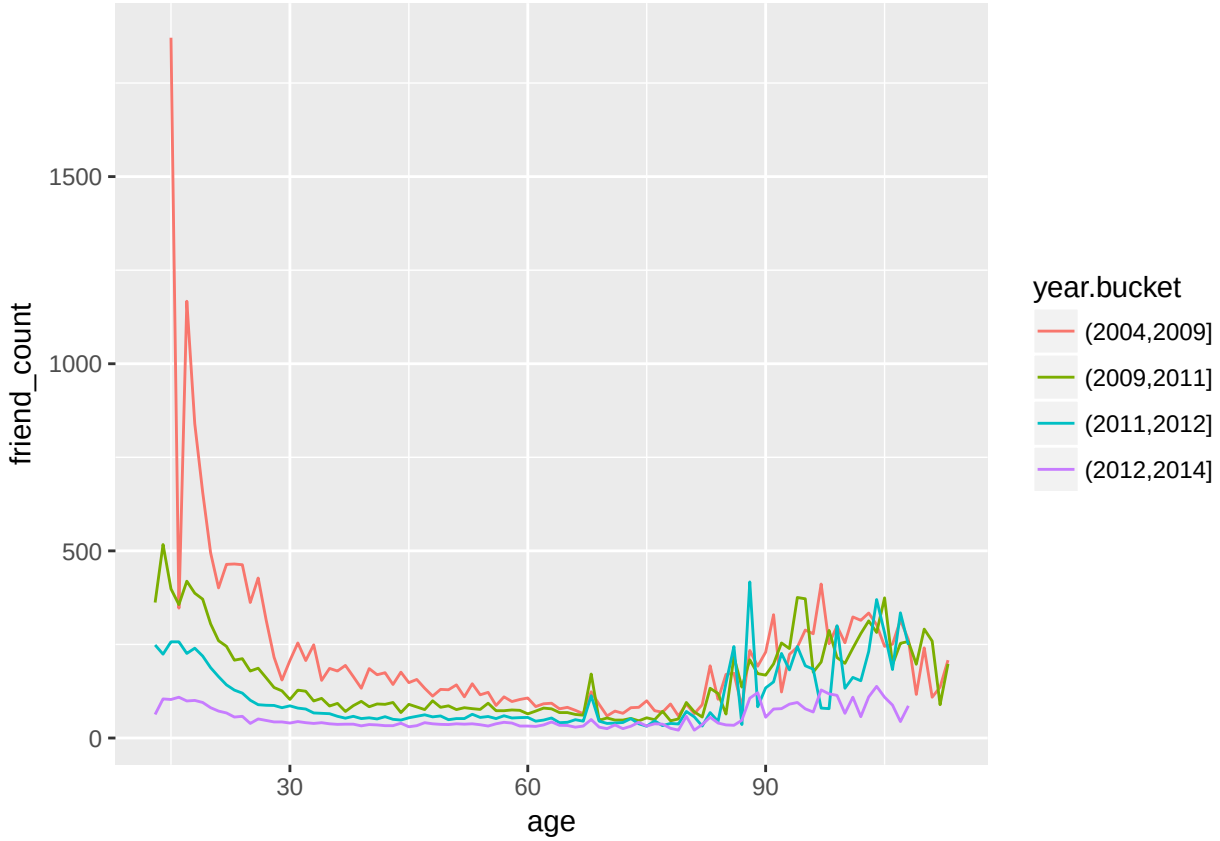
```
pf$year.bucket <- cut(pf$ year_joined, breaks = c(2004, 2009,2011,2012,2014), right= TRUE)
table(pf$year.bucket, useNA = 'ifany')
```

```
##
```

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

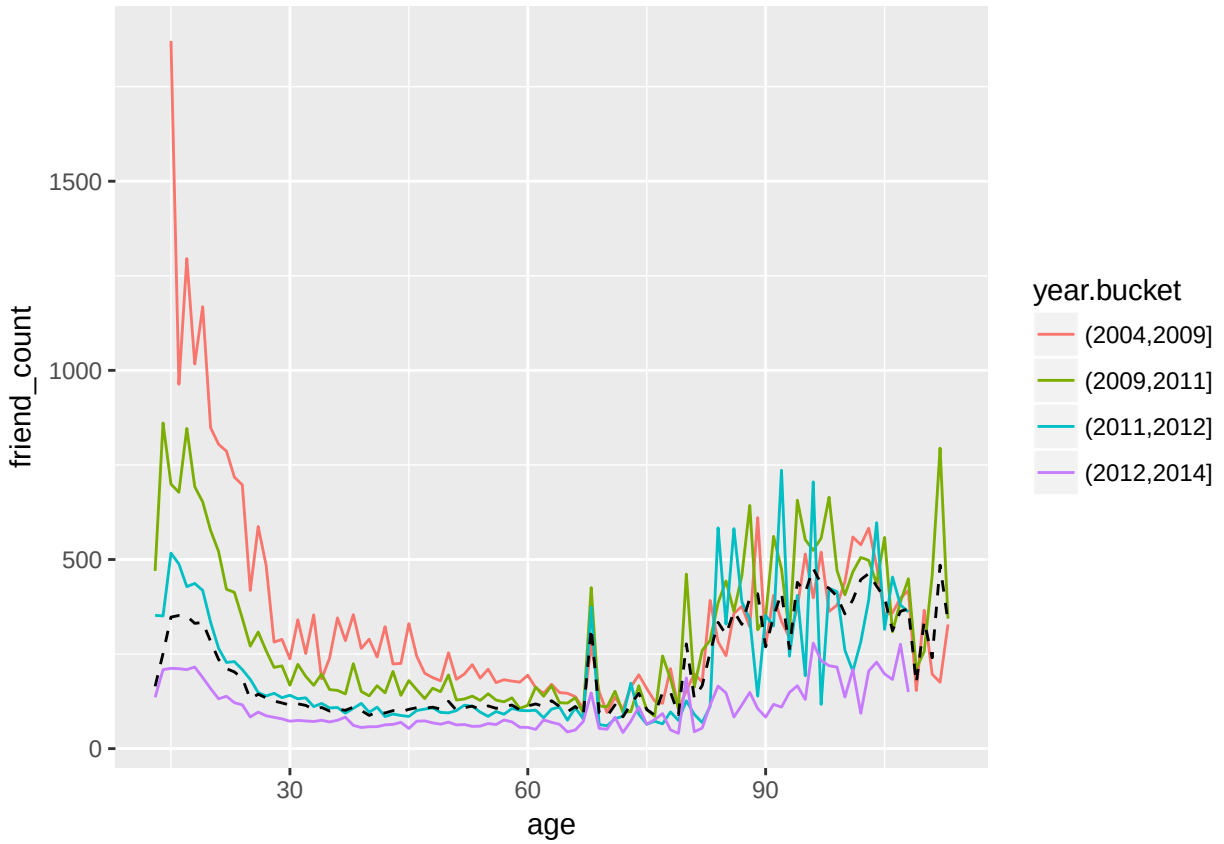
Plotting it All Together

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



Plot the Grand Mean

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



Friending Rate

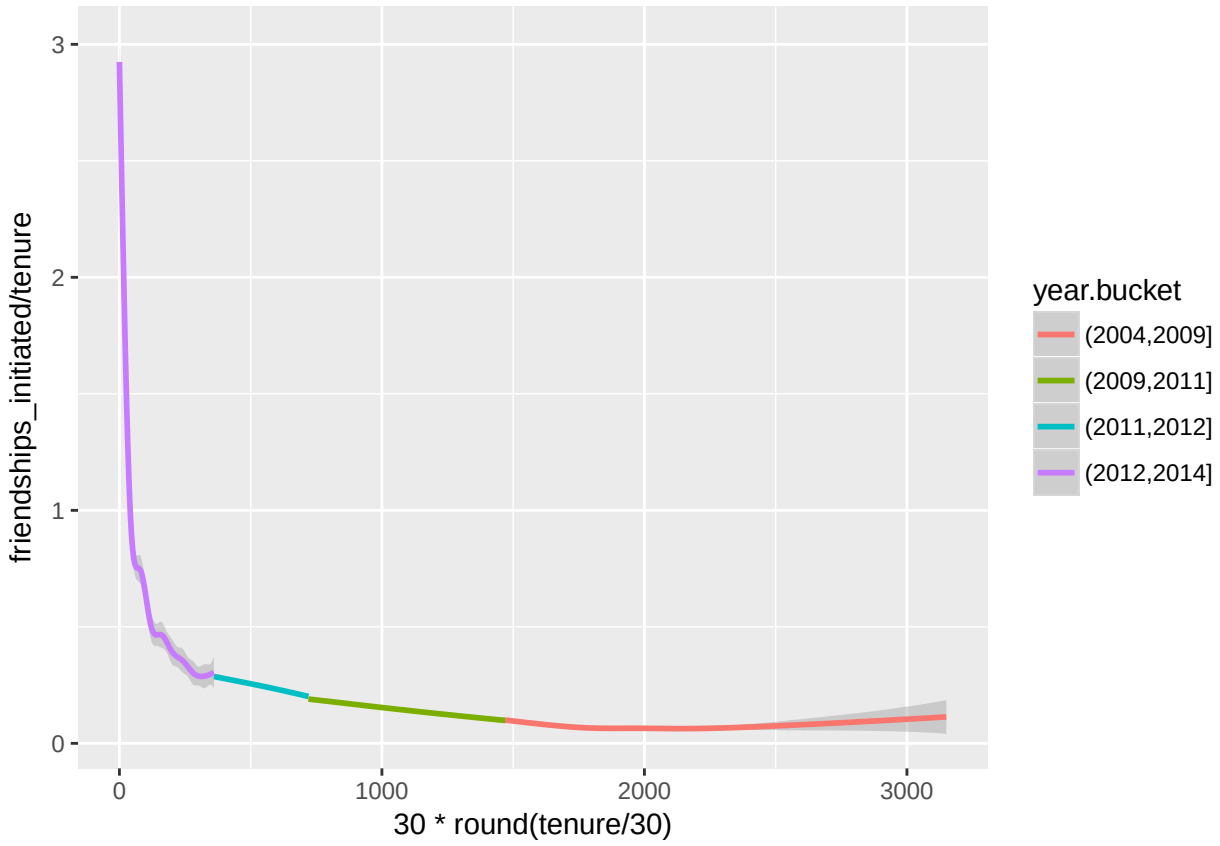
```
with(subset(pf, tenure >=1), summary(friend_count/tenure))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0775   0.2205   0.6096   0.5658 417.0000
```

Friendships Initiated

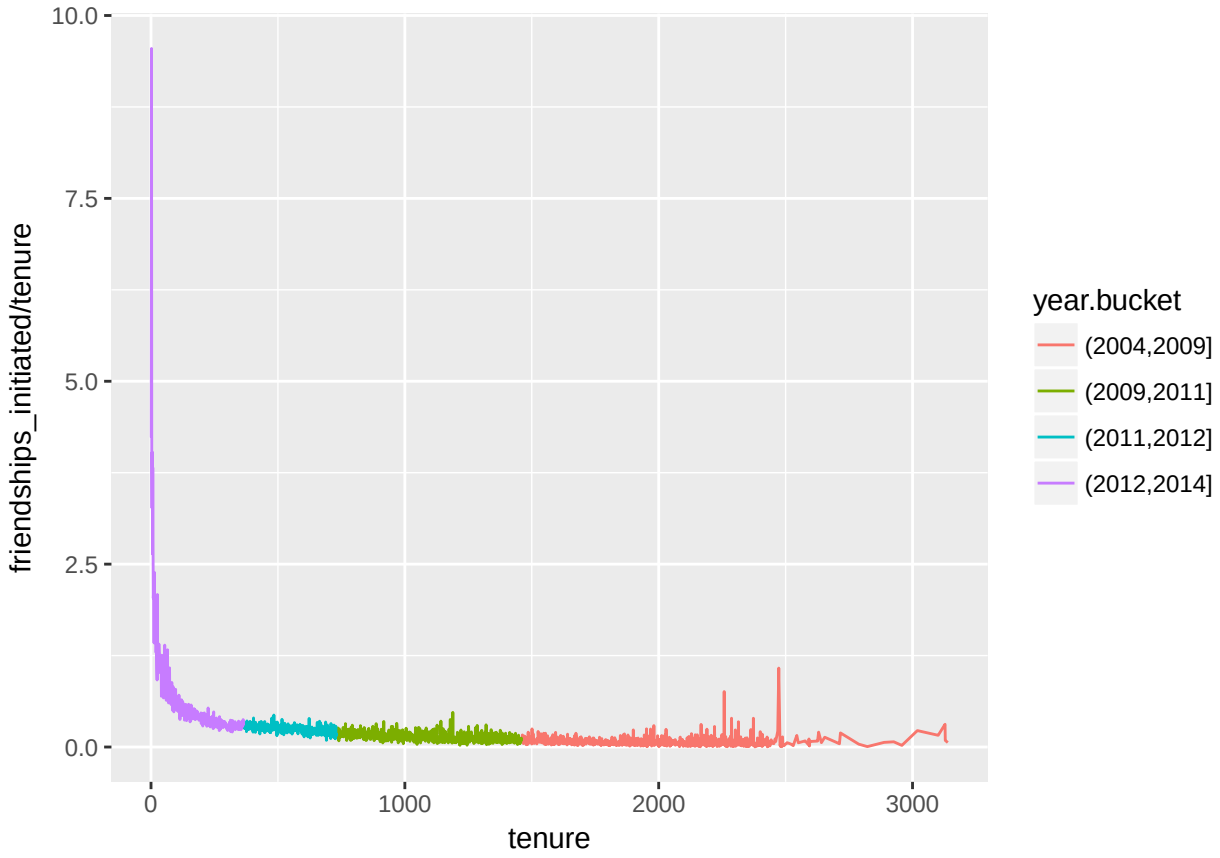
```
ggplot(aes(x= 30*round(tenure/30) , y= friendships_initiated/tenure),data=subset(pf, tenure>=1)) +geom_s
```

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

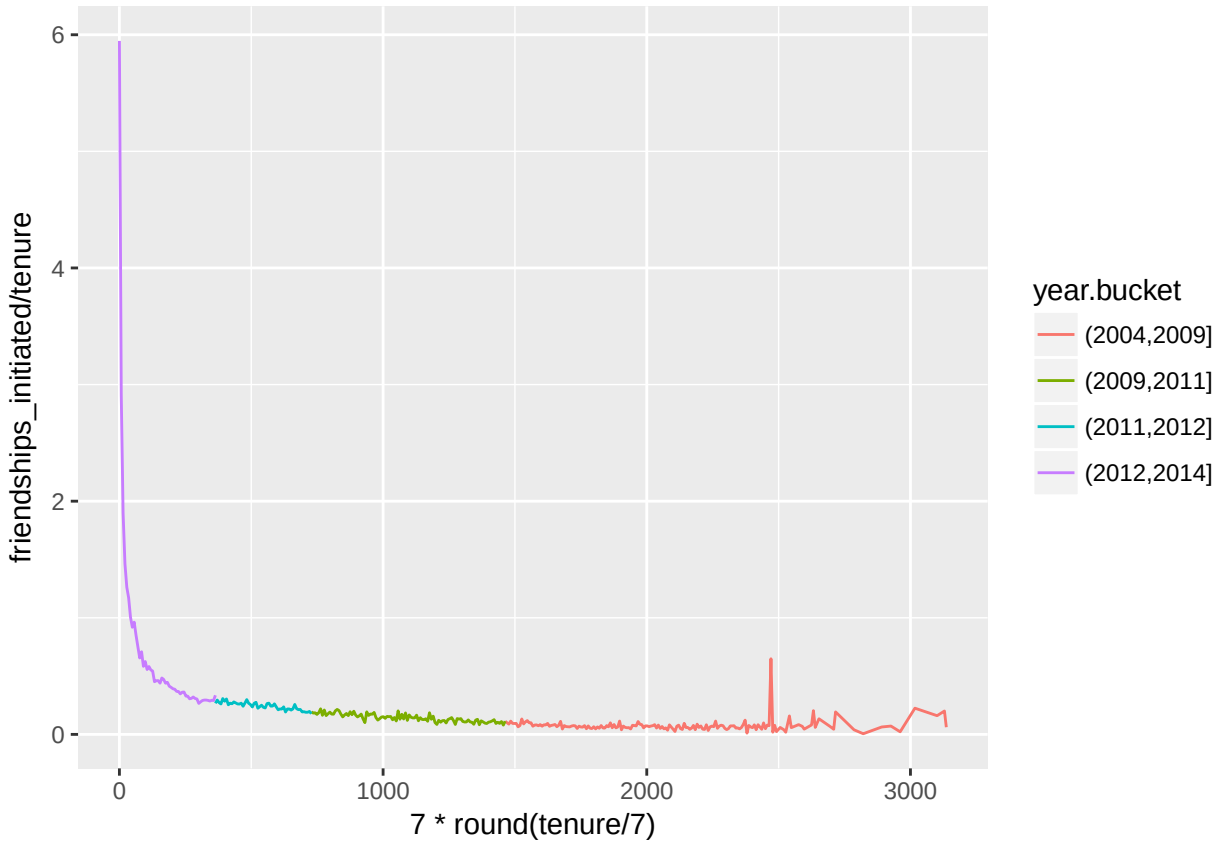


Bias-Variance Tradeoff Revisited

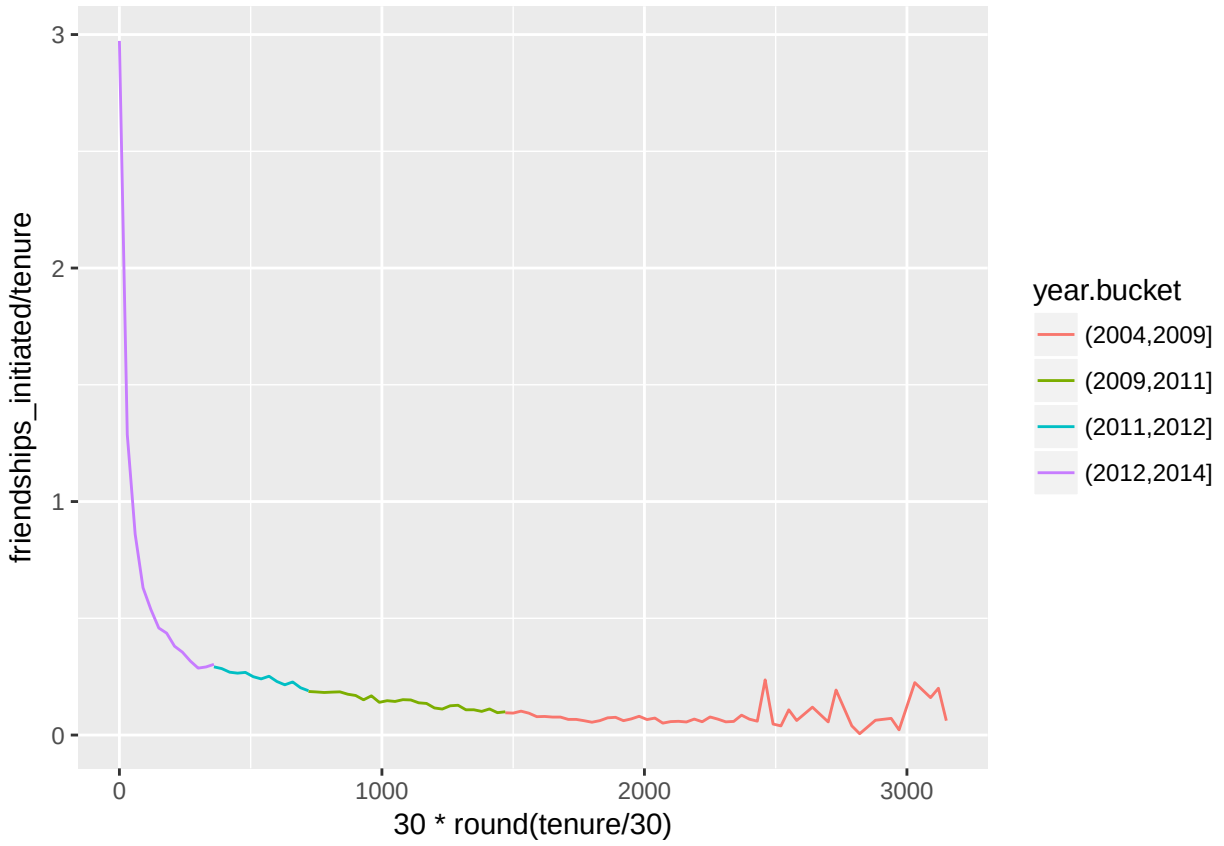
```
ggplot(aes(x = tenure, y = friendships_initiated / tenure),
       data = subset(pf, tenure >= 1)) +
  geom_line(aes(color = year.bucket),
            stat = 'summary',
            fun.y = mean)
```



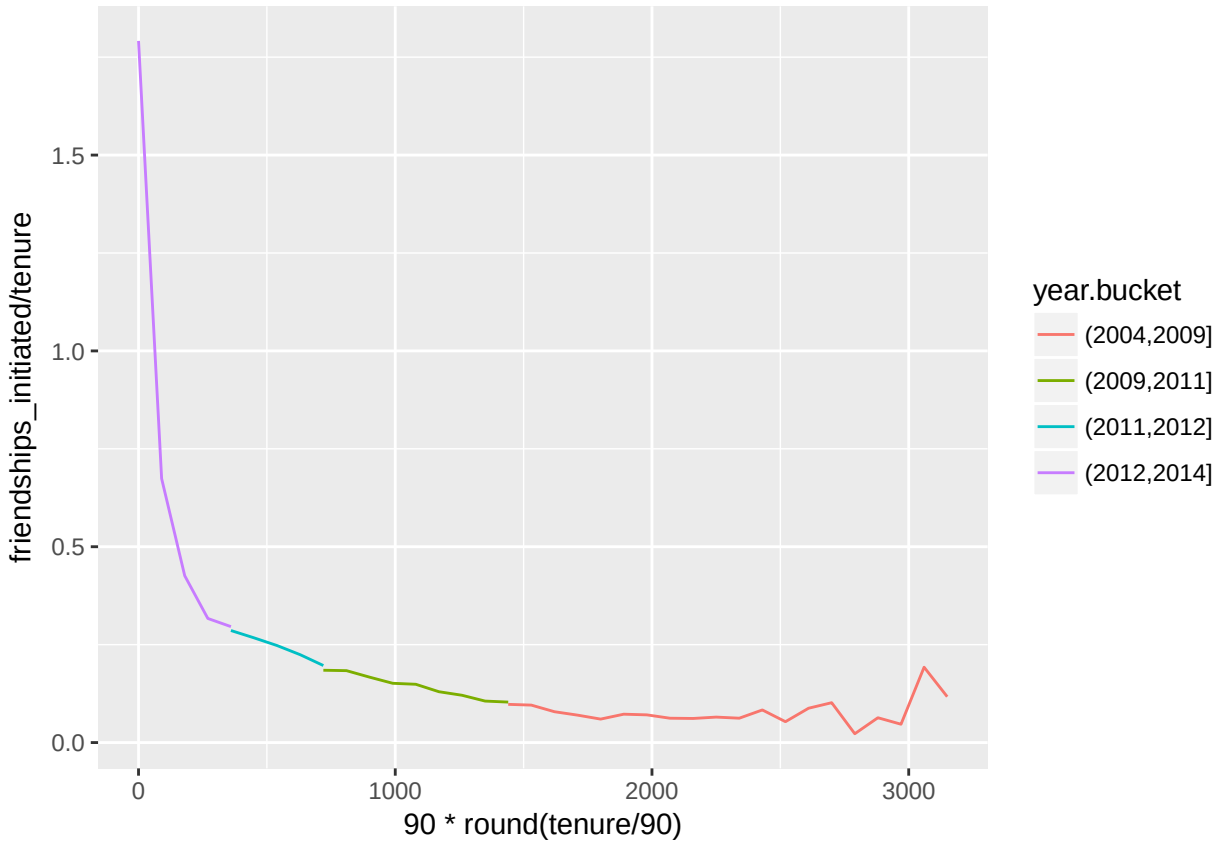

```
ggplot(aes(x = 7 * round(tenure / 7), y = friendships_initiated / tenure),
       data = subset(pf, tenure > 0)) +
  geom_line(aes(color = year.bucket),
            stat = "summary",
            fun.y = mean)
```




```
ggplot(aes(x = 30 * round(tenure / 30), y = friendships_initiated / tenure),
       data = subset(pf, tenure > 0)) +
  geom_line(aes(color = year.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.bucket),
            stat = "summary",
            fun.y = mean)
```


friendships_initiated/tenure
1.5
1.0
0.5
0.0
0
1000
2000
3000
90 * round(tenure/90)
year.bucket
(2004,2009]
(2009,2011]
(2011,2012]
(2012,2014]