UD651-Lesson3

This document includes all the assignments done as part of the UD651 Lesson 3 course on Udacity.

Dataset used is a Pseudo facebook data downloaded from the course site and saved in the current working directory. https://s3.amazonaws.com/udacity-hosted-downloads/ud651/lesson3_student.rmd

Loading required libraries:

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

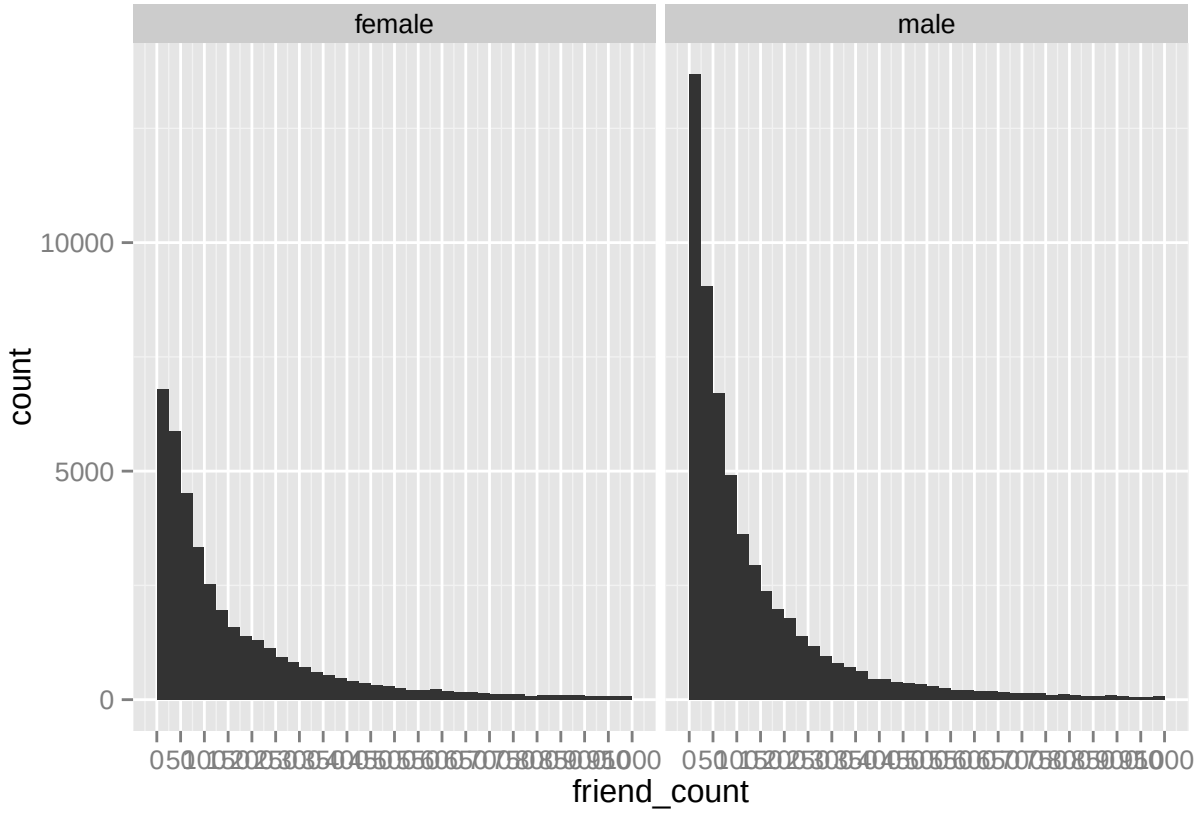
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

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

Creating a histogram for the friend count separated by Gender:

```
qplot(x = friend_count, data = na.omit(fb), binwidth = 25) +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50)) +
  facet_wrap(~gender)
```



Generating the summary data:

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

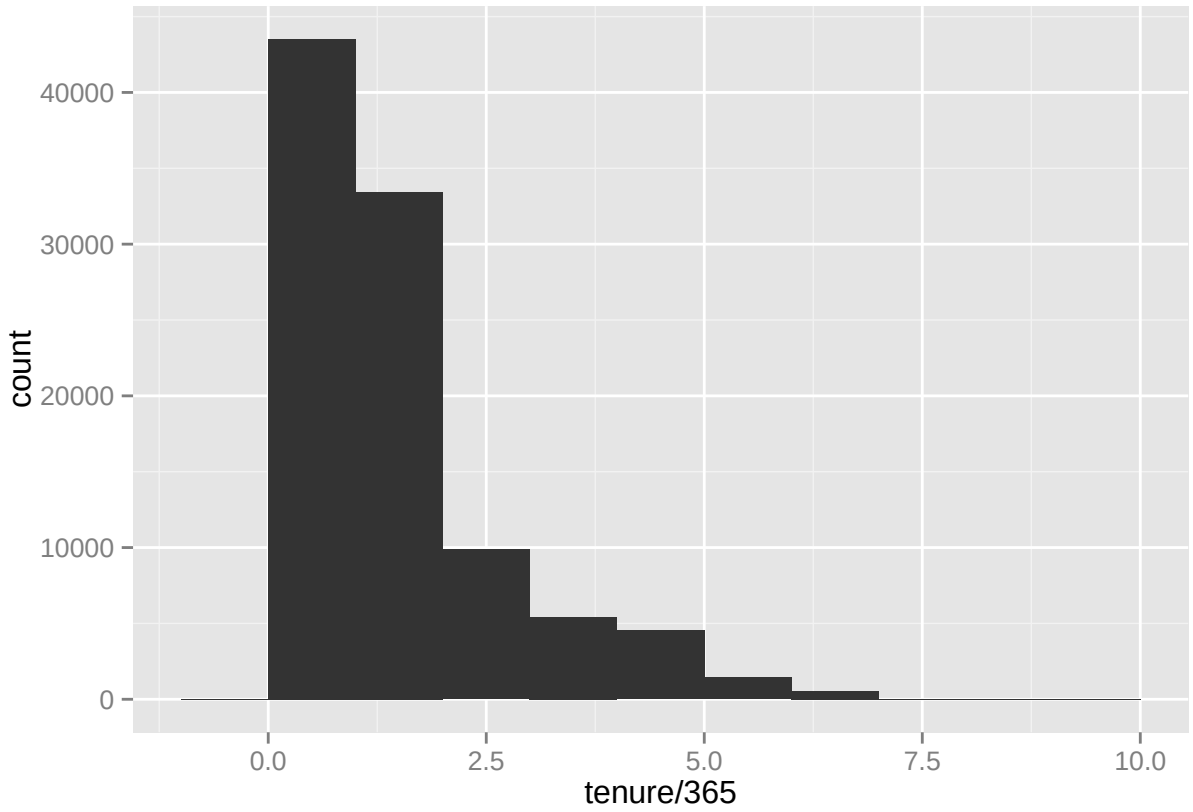
```
##
## female   male
##  40254  58574
```

```
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
```

Analyze the tenure and its spread in the dataset. Converting the data into years since its in days:

```
qplot(x = tenure/365, data = na.omit(fb), binwidth = 1)
```



Explore several scaling functions in ggplot and comparing them to the original plot.

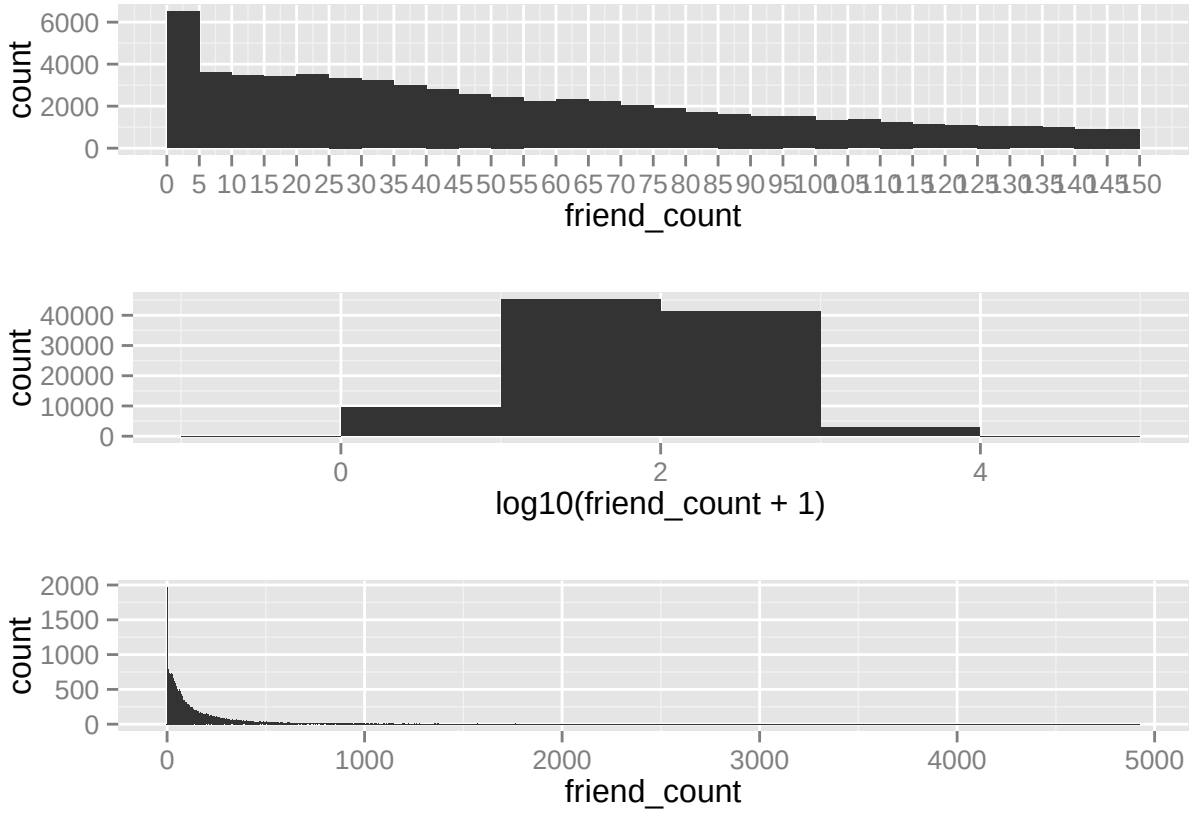
```
p1 <- qplot(x = friend_count, data = na.omit(fb), binwidth = 5) +
  scale_x_continuous(limits = c(0, 150), breaks = seq(0, 150, 5))

p2 <- qplot(x = log10(friend_count+1), data = na.omit(fb), binwidth = 1)
```

```
p3 <- qplot(x = friend_count, data = na.omit(fb), binwidth = 1)
```

Arrange the above graphs in a grid:

```
grid.arrange(p1, p2, p3, ncol=1)
```



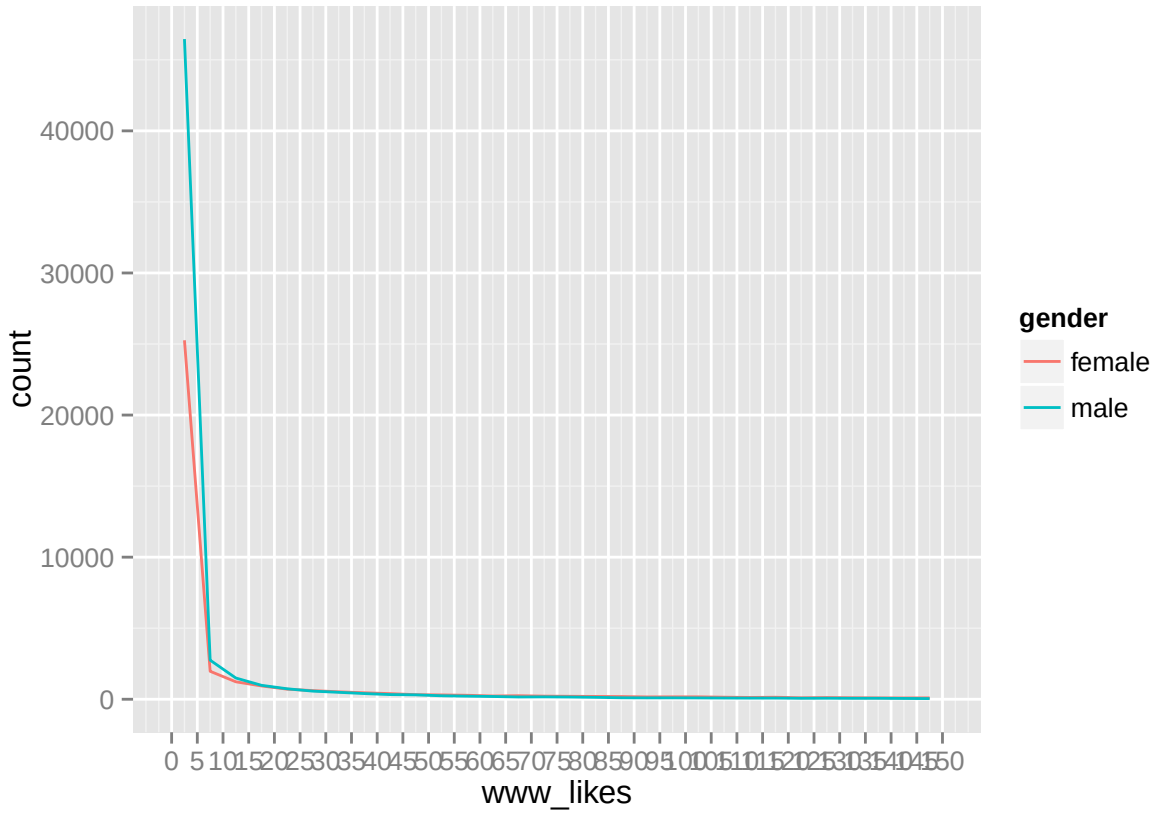
Plotting a frequency polygon using the geom options on different scales

```
qplot(x = www_likes, data = na.omit(fb), geom="freqpoly", color=gender) +
  scale_x_continuous(limits = c(0, 150), breaks = seq(0, 150, 5))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

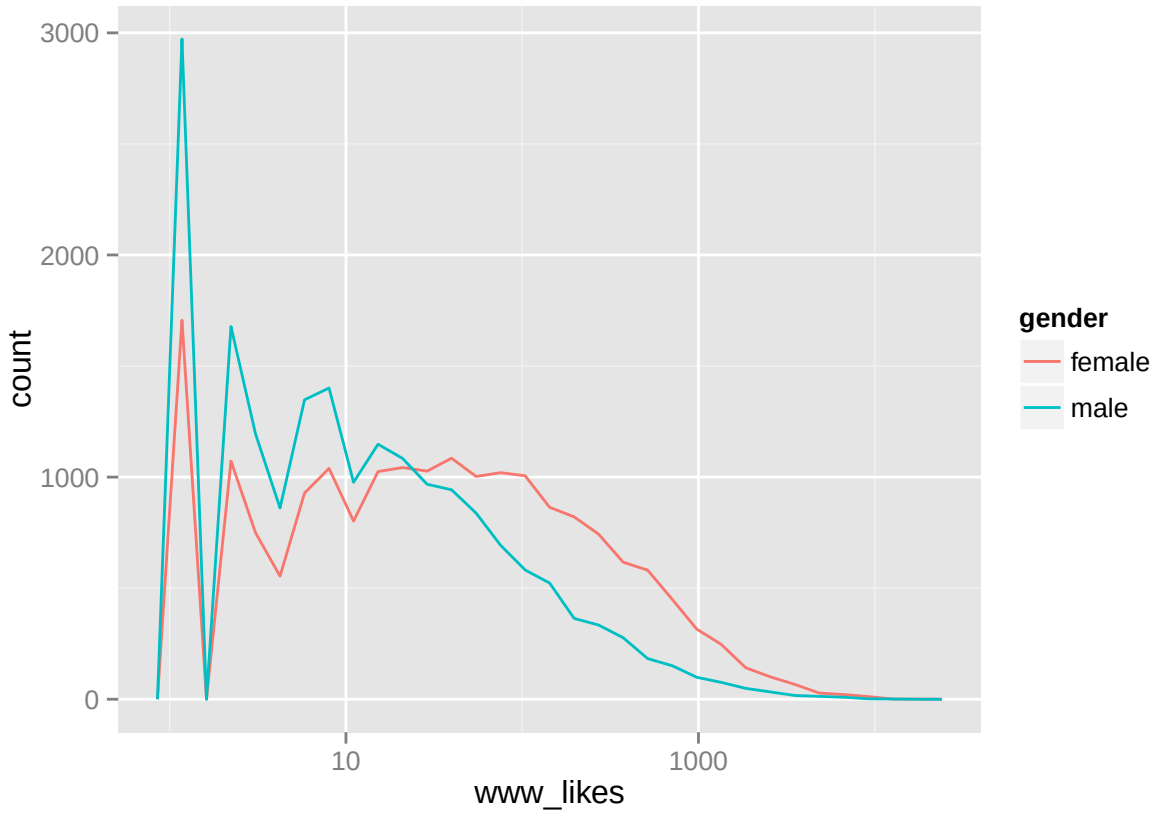
```
## Warning: Removed 2 rows containing missing values (geom_path).
```

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




```
qplot(x = www_likes, data = na.omit(fb), geom="freqpoly", color=gender) +
  scale_x_log10()
```

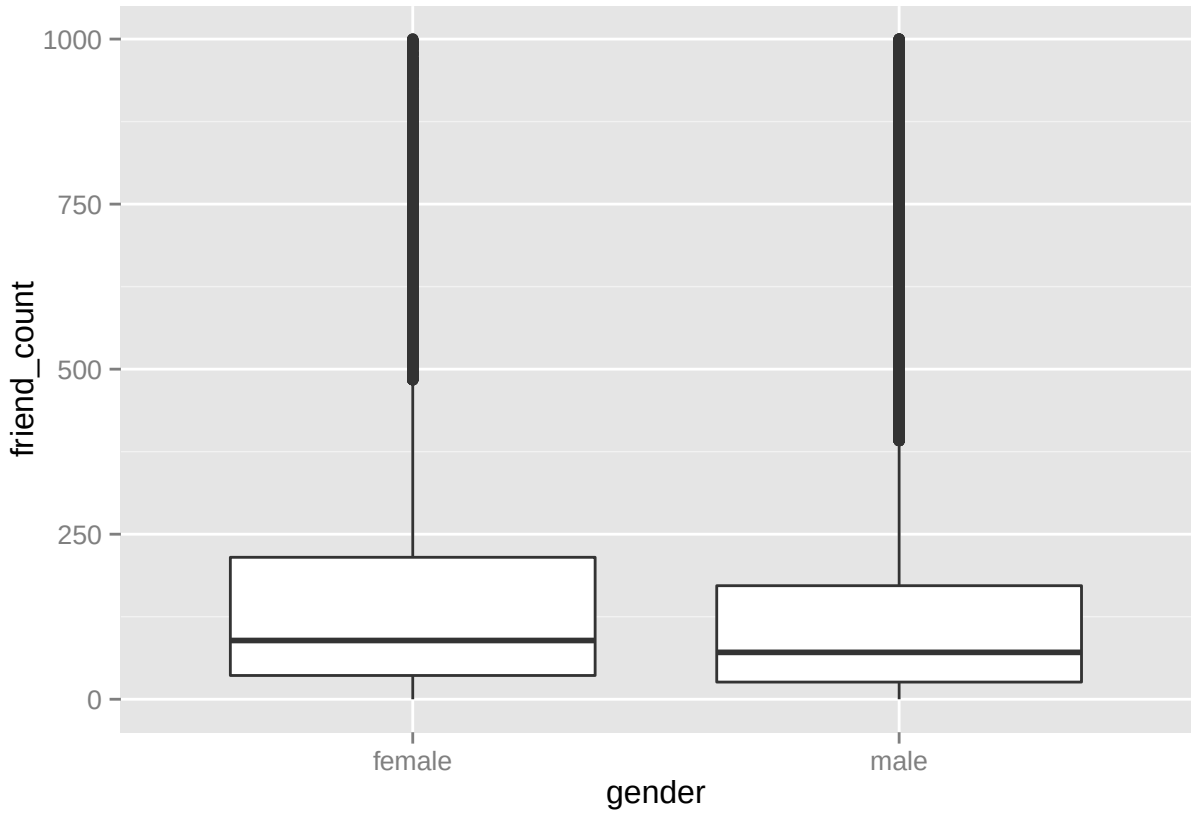
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Using the ylim operator to limit the Y limits

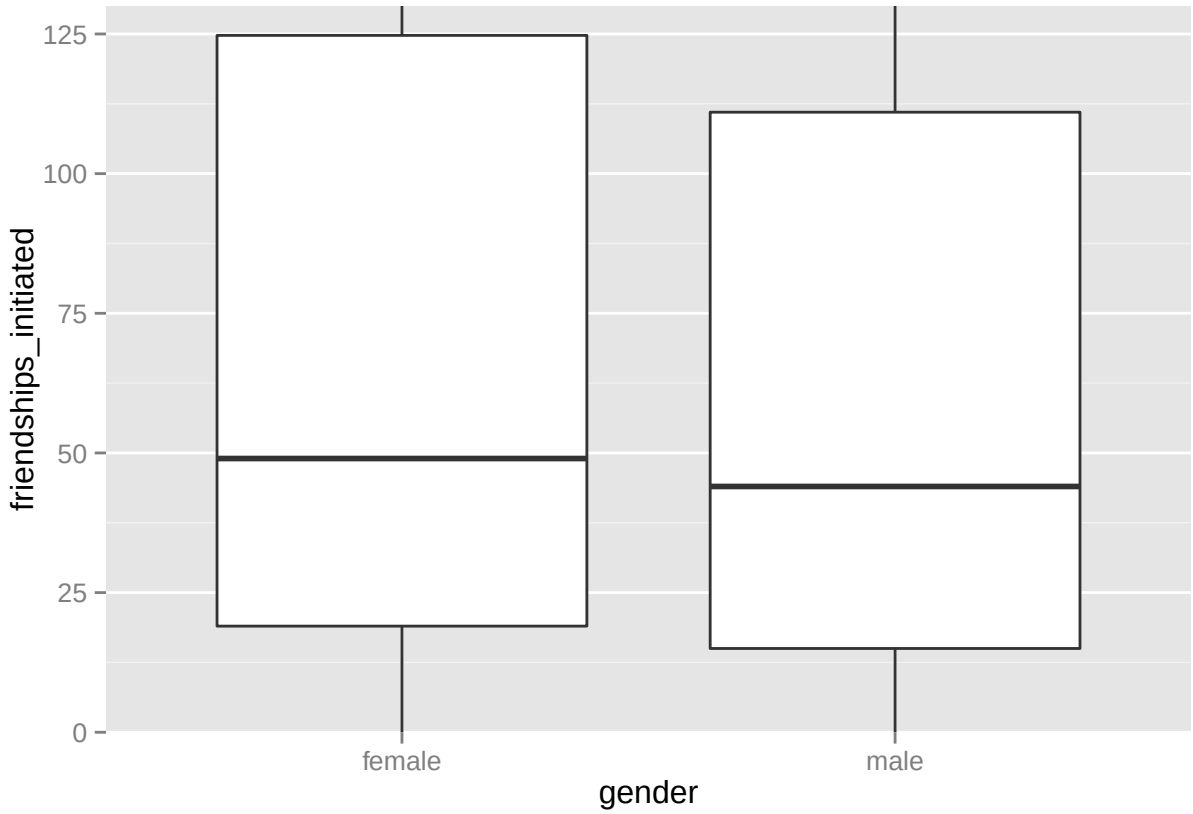
```
qplot(x=gender, y = friend_count, data = subset(fb, !is.na(gender)), geom="boxplot",
      ylim = c(0,1000))
```

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



Using coord cartesian to limit the y values on friendships_initiated:

```
qplot(x=gender, y = friendships_initiated, data = subset(fb, !is.na(gender)), geom="boxplot") +
      coord_cartesian(ylim = c(0,130))
```



Calculate the mobile_check_in s:

```
fb$mobile_check_in <- NA

fb$mobile_check_in <- ifelse(fb$mobile_likes >0,1,0)

summary(fb$mobile_check_in)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.6459  1.0000  1.0000
```

```
sum(fb$mobile_check_in == 1)/
length(fb$userid)
```

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