EDA Lesson 4

Tom Gocken

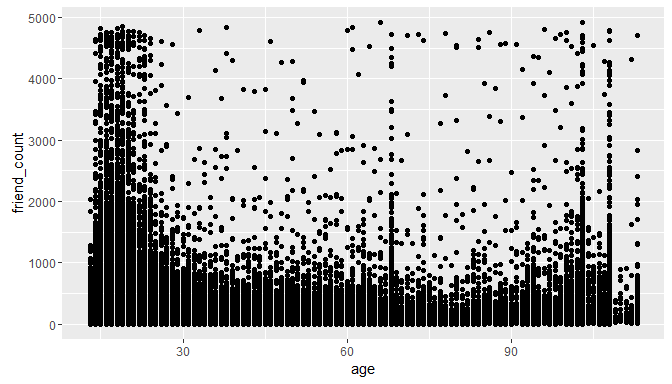
Sunday, January 24, 2016

### Scatterplots

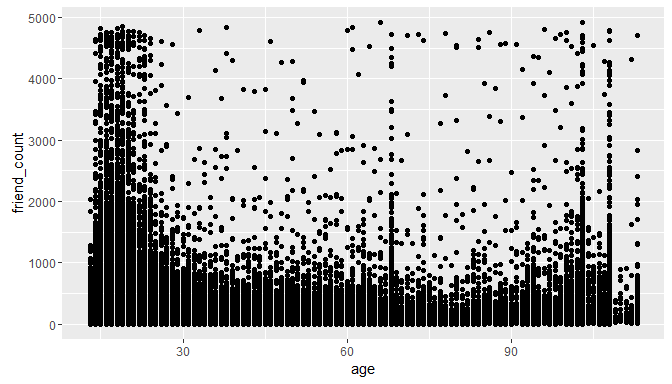
# https://s3.amazonaws.com/udacity-hosted-downloads/ud651/EDA\_Course\_Materials.zip  
# Set directory, load data and ggplot2 library  
  
setwd("C:/Projects/UD651")  
list.files()

## [1] "EDA\_lesson3.docx" "EDA\_lesson3.html" "EDA\_lesson3.Rmd"   
## [4] "EDA\_lesson4.html" "EDA\_lesson4.Rmd" "pseudo\_facebook.tsv"  
## [7] "UD651.Rproj"

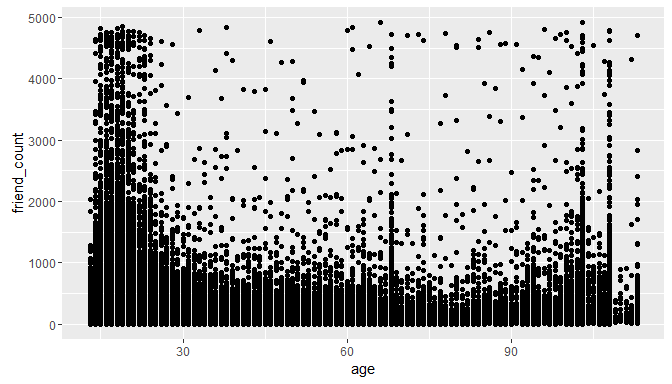
pf <- read.csv('pseudo\_facebook.tsv', sep = '\t')  
library(ggplot2)  
  
qplot(x = age, y = friend\_count, data = pf)



# Alternative syntax  
qplot(age, friend\_count, data = pf)



# ggplot syntax  
ggplot(aes(x = age, y = friend\_count), data = pf) + geom\_point()

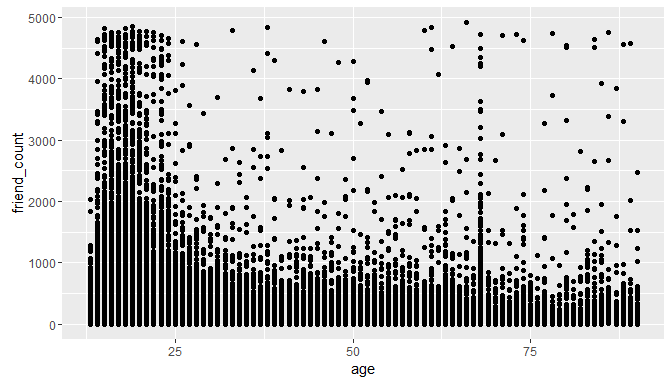


summary(pf$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13.00 20.00 28.00 37.28 50.00 113.00

# Limit x axis range  
ggplot(aes(x = age, y = friend\_count), data = pf) + geom\_point() + xlim(13,90)

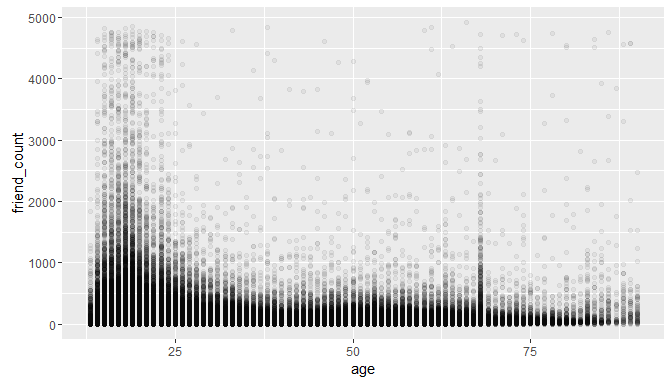
## Warning: Removed 4906 rows containing missing values (geom\_point).



### Overplotting

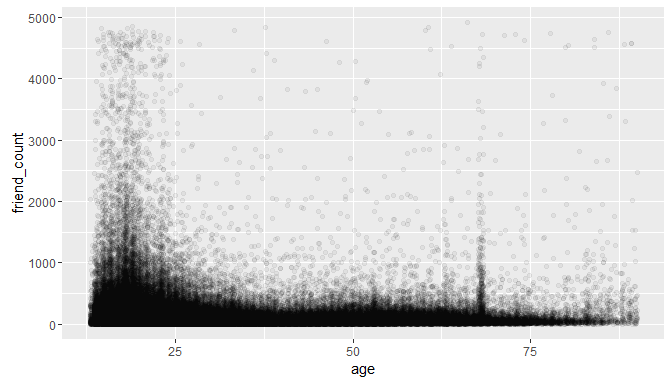
# Reduce overplotting with alpha where 1 graph point equal to 20 data points  
ggplot(aes(x = age, y = friend\_count), data = pf) +   
 geom\_point(alpha = 1/20) +   
 xlim(13,90)

## Warning: Removed 4906 rows containing missing values (geom\_point).



# Add jitter  
ggplot(aes(x = age, y = friend\_count), data = pf) +   
 geom\_jitter(alpha = 1/20) +   
 xlim(13,90)

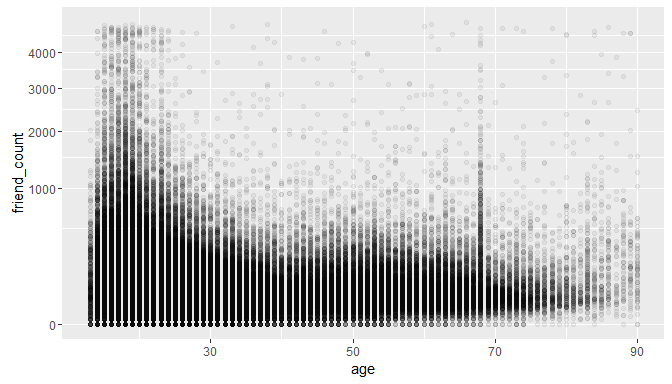
## Warning: Removed 5193 rows containing missing values (geom\_point).



### coord\_trans

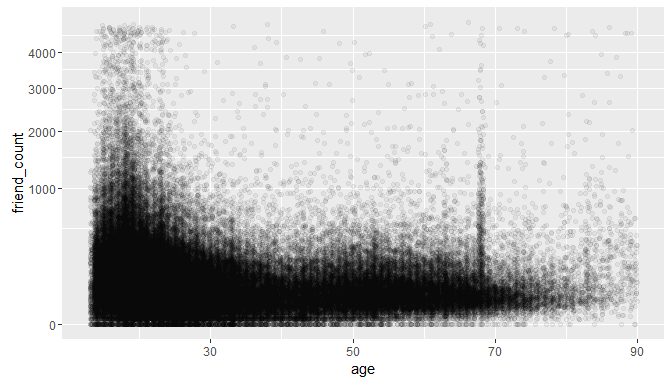
# Square root transformation  
ggplot(aes(x = age, y = friend\_count), data = pf) +   
 geom\_point(alpha = 1/20) +   
 xlim(13,90) +  
 coord\_trans(y = 'sqrt')

## Warning: Removed 4906 rows containing missing values (geom\_point).



# Jitter age, > 0 only  
ggplot(aes(x = age, y = friend\_count), data = pf) +   
 geom\_point(alpha = 1/20, position = position\_jitter(h = 0)) +   
 xlim(13,90) +  
 coord\_trans(y = 'sqrt')

## Warning: Removed 5190 rows containing missing values (geom\_point).

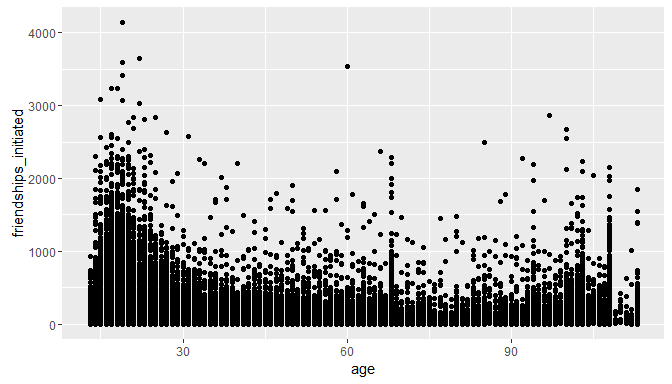


### Alpha and Jitter

names(pf)

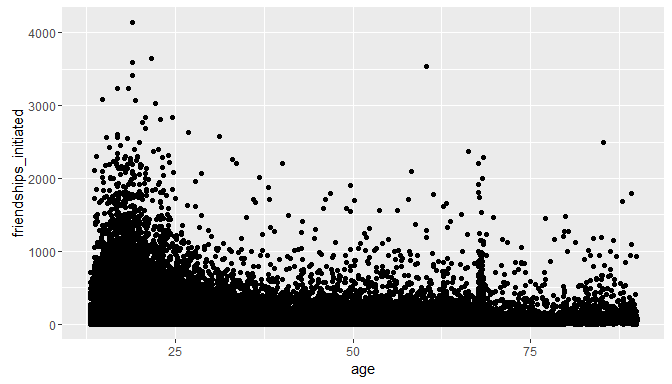
## [1] "userid" "age"   
## [3] "dob\_day" "dob\_year"   
## [5] "dob\_month" "gender"   
## [7] "tenure" "friend\_count"   
## [9] "friendships\_initiated" "likes"   
## [11] "likes\_received" "mobile\_likes"   
## [13] "mobile\_likes\_received" "www\_likes"   
## [15] "www\_likes\_received"

ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 geom\_point()



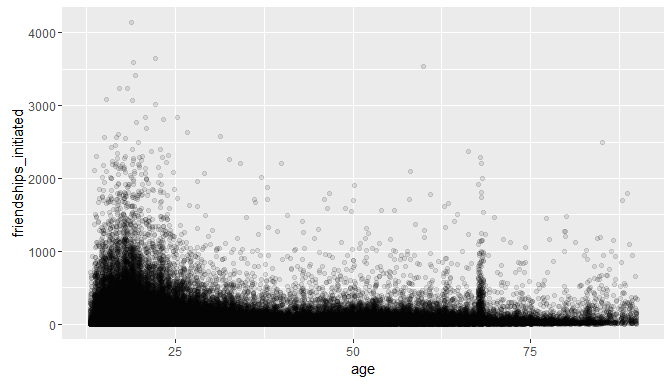
# Add jitter  
ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 geom\_jitter() +  
 xlim(13,90)

## Warning: Removed 5179 rows containing missing values (geom\_point).



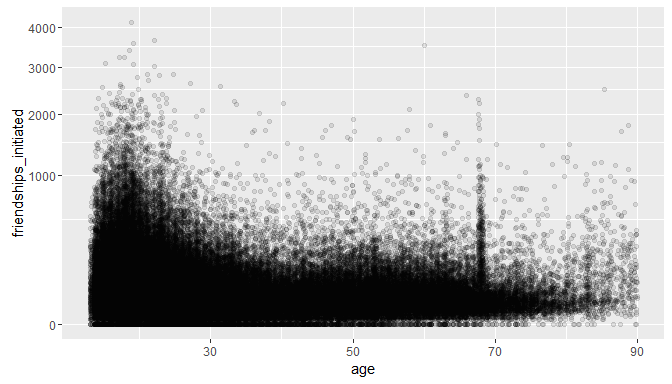
# Add alpha  
ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 geom\_jitter(alpha = 1/10) +  
 xlim(13,90)

## Warning: Removed 5207 rows containing missing values (geom\_point).



# Alt jitter syntax, sqrt transformation  
ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 geom\_point(alpha = 1/10, position = position\_jitter(h = 0)) +  
 xlim(13,90) +  
 coord\_trans(y = 'sqrt')

## Warning: Removed 5185 rows containing missing values (geom\_point).



### Conditional Means

# install.packages('dplyr')  
library(dplyr)

##   
## Attaching package: 'dplyr'  
##   
## The following objects are masked from 'package:stats':  
##   
## filter, lag  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

age\_groups <- group\_by(pf, age)  
  
pf.fc\_by\_age <- summarise(age\_groups,  
 friend\_count\_mean = mean(friend\_count),  
 friend\_count\_median = median(friend\_count),  
 n = n())  
  
pf.fc\_by\_age <- arrange(pf.fc\_by\_age, age)  
  
head(pf.fc\_by\_age)

## Source: local data frame [6 x 4]  
##   
## age friend\_count\_mean friend\_count\_median n  
## (int) (dbl) (dbl) (int)  
## 1 13 164.7500 74.0 484  
## 2 14 251.3901 132.0 1925  
## 3 15 347.6921 161.0 2618  
## 4 16 351.9371 171.5 3086  
## 5 17 350.3006 156.0 3283  
## 6 18 331.1663 162.0 5196

# Alternative syntax, note: %.% equivalent to %>%  
pf.fc\_by\_age <- pf %.%   
 group\_by(age) %.%  
 summarise(friend\_count\_mean = mean(friend\_count),  
 friend\_count\_median = median(friend\_count),  
 n = n()) %.%  
 arrange(age)

## Warning: '%.%' is deprecated.  
## Use '%>%' instead.  
## See help("Deprecated")

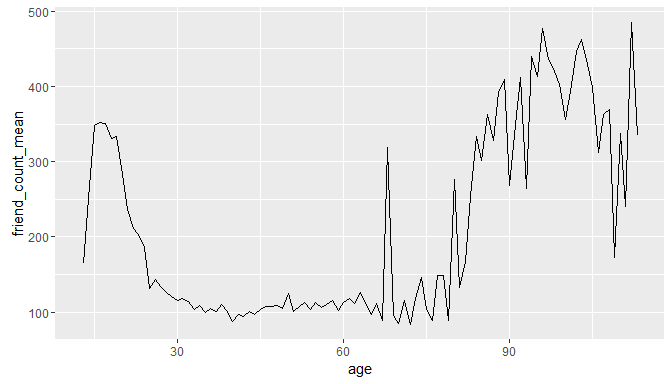
## Warning: '%.%' is deprecated.  
## Use '%>%' instead.  
## See help("Deprecated")

## Warning: '%.%' is deprecated.  
## Use '%>%' instead.  
## See help("Deprecated")

head(pf.fc\_by\_age)

## Source: local data frame [6 x 4]  
##   
## age friend\_count\_mean friend\_count\_median n  
## (int) (dbl) (dbl) (int)  
## 1 13 164.7500 74.0 484  
## 2 14 251.3901 132.0 1925  
## 3 15 347.6921 161.0 2618  
## 4 16 351.9371 171.5 3086  
## 5 17 350.3006 156.0 3283  
## 6 18 331.1663 162.0 5196

# geom\_line on summarized values  
ggplot(aes(x= age, y = friend\_count\_mean), data = pf.fc\_by\_age) +  
 geom\_line()

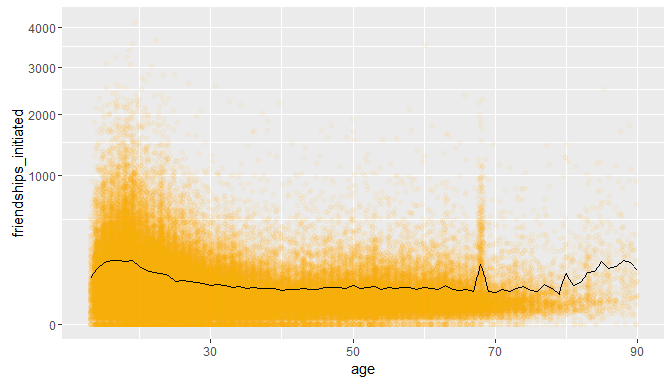


### Overlaying Summaries with Raw Data

# Display raw data and mean  
ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 geom\_point(alpha = 0.05,   
 position = position\_jitter(h = 0),  
 color = 'orange') +  
 xlim(13,90) +  
 coord\_trans(y = 'sqrt') +  
 geom\_line(stat = 'summary', fun.y = mean)

## Warning: Removed 4906 rows containing non-finite values (stat\_summary).

## Warning: Removed 5176 rows containing missing values (geom\_point).



# Add quantiles  
ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 geom\_point(alpha = 0.05,   
 position = position\_jitter(h = 0),  
 color = 'orange') +  
 xlim(13,90) +  
 coord\_trans(y = 'sqrt') +  
 geom\_line(stat = 'summary', fun.y = mean) +  
 geom\_line(stat = 'summary', fun.y = quantile, fun.args=list(probs=0.1),  
 linetype = 2, color = 'blue') +  
 geom\_line(stat = 'summary', fun.y = quantile, fun.args=list(probs=0.5), # = median  
 color = 'blue') +  
 geom\_line(stat = 'summary', fun.y = quantile, fun.args=list(probs=0.9),  
 linetype = 2, color = 'blue')

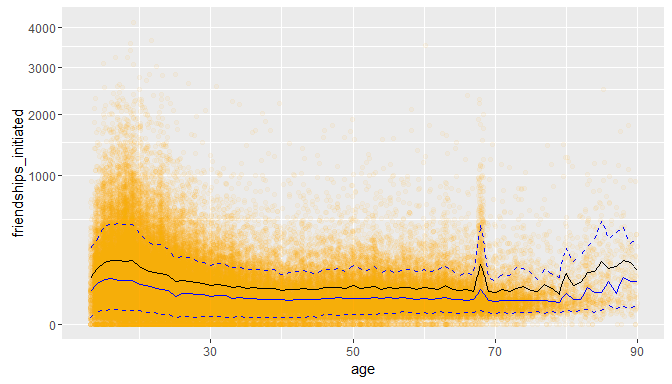
## Warning: Removed 4906 rows containing non-finite values (stat\_summary).

## Warning: Removed 4906 rows containing non-finite values (stat\_summary).

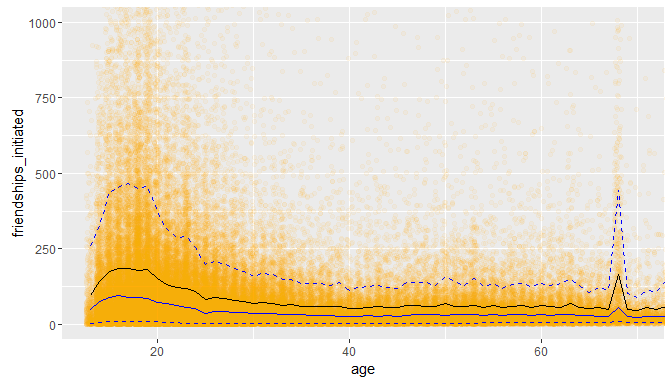
## Warning: Removed 4906 rows containing non-finite values (stat\_summary).

## Warning: Removed 4906 rows containing non-finite values (stat\_summary).

## Warning: Removed 5188 rows containing missing values (geom\_point).



# Use coord\_cartesian to zoom in  
ggplot(aes(x = age, y = friendships\_initiated), data = pf) +  
 coord\_cartesian(xlim = c(13, 70), ylim = c(0, 1000)) +  
 geom\_point(alpha = 0.05,   
 position = position\_jitter(h = 0),  
 color = 'orange') +  
 geom\_line(stat = 'summary', fun.y = mean) +  
 geom\_line(stat = 'summary', fun.y = quantile, fun.args=list(probs=0.1),  
 linetype = 2, color = 'blue') +  
 geom\_line(stat = 'summary', fun.y = quantile, fun.args=list(probs=0.5), # = median  
 color = 'blue') +  
 geom\_line(stat = 'summary', fun.y = quantile, fun.args=list(probs=0.9),  
 linetype = 2, color = 'blue')



### Correlation

cor.test(pf$age, pf$friend\_count, method = 'pearson')

##   
## Pearson's product-moment correlation  
##   
## data: pf$age and pf$friend\_count  
## t = -8.6268, df = 99001, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.03363072 -0.02118189  
## sample estimates:  
## cor   
## -0.02740737

# Alt syntax  
with(pf, cor.test(age, friend\_count, method = 'pearson'))

##   
## Pearson's product-moment correlation  
##   
## data: age and friend\_count  
## t = -8.6268, df = 99001, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.03363072 -0.02118189  
## sample estimates:  
## cor   
## -0.02740737

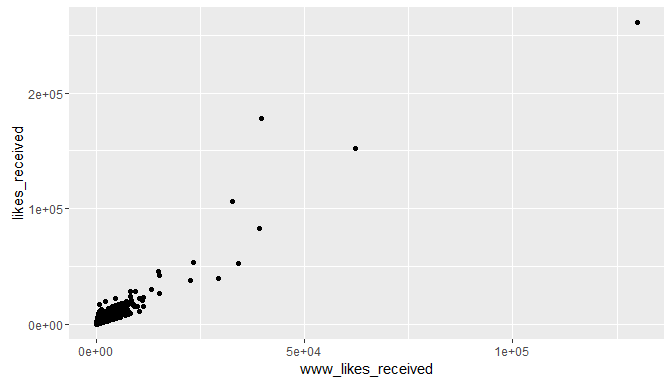
### Correlation on Subsets

with(subset(pf, age <= 70), cor.test(age, friend\_count)) # method = 'peason' is default

##   
## Pearson's product-moment correlation  
##   
## data: age and friend\_count  
## t = -52.5923, df = 91029, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.1780220 -0.1654129  
## sample estimates:  
## cor   
## -0.1717245

### Create Scatterplots

ggplot(aes(x = www\_likes\_received, likes\_received), data = pf) +  
 geom\_point()

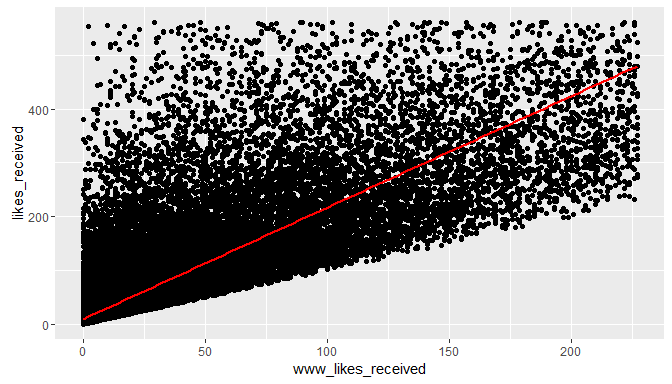


### Strong Correlations

# Limit range using 95% quantiles, add line for correlation coefficient  
ggplot(aes(x = www\_likes\_received, likes\_received), data = pf) +  
 geom\_point() +  
 xlim(0, quantile(pf$www\_likes\_received, 0.95)) +  
 ylim(0, quantile(pf$likes\_received, 0.95)) +  
 geom\_smooth(method = 'lm', color = 'red')

## Warning: Removed 6075 rows containing non-finite values (stat\_smooth).

## Warning: Removed 6075 rows containing missing values (geom\_point).



cor.test(pf$www\_likes\_received, pf$likes\_received)

##   
## Pearson's product-moment correlation  
##   
## data: pf$www\_likes\_received and pf$likes\_received  
## t = 937.1035, df = 99001, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9473553 0.9486176  
## sample estimates:  
## cor   
## 0.9479902

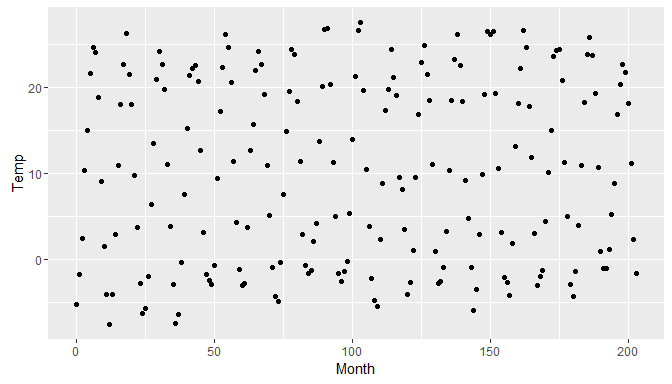
# High correlation not meaningful since www\_likes\_received is a subset of likes\_received

### More Caution with Correlation

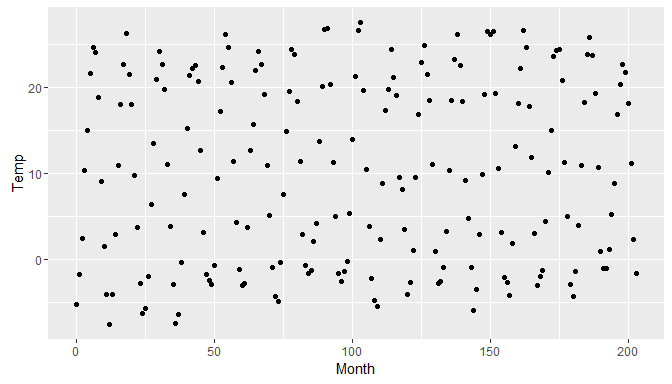
# install.packages('alr3')  
library(alr3)

## Loading required package: car

data(Mitchell)  
# ?Mitchell  
# Mitchell soil temperature  
# Description: Data collected by Kenneth G. Hubbard on soil temperature at 20 cm depth   
# in Mitchell, Nebraska for 17 years (1976-1992) The variable month is the month number.  
  
ggplot(data = Mitchell, aes(x = Month, y = Temp)) +  
 geom\_point()



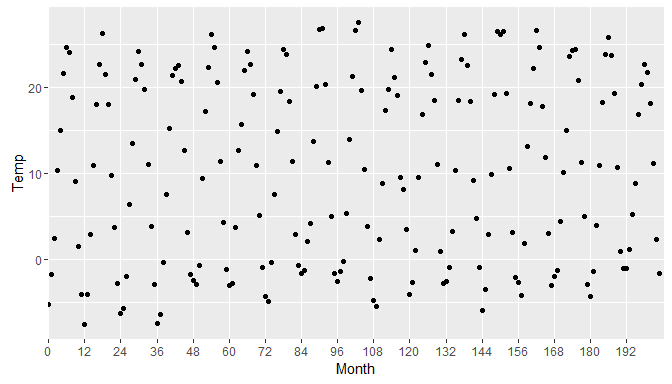
# Alt syntax  
qplot(data = Mitchell, Month, Temp)



### Noisy Scatterplots

### Making Sense of Data

ggplot(data = Mitchell, aes(x = Month, y = Temp)) +  
 geom\_point() +  
 scale\_x\_discrete(breaks = seq(0, 203, 12))

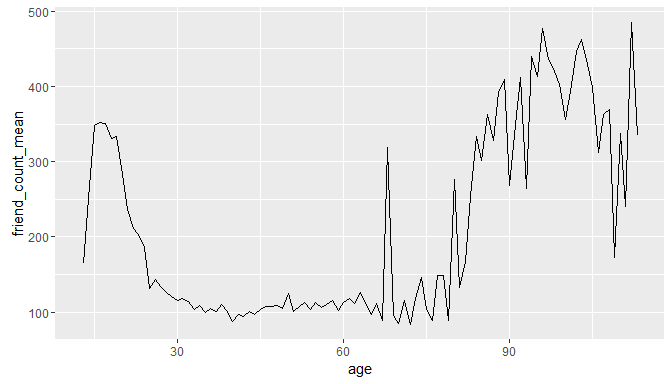


### A New Perspective

# Stretch horizontal axis to reveal cyclic pattern

### Understanding Noise Age to Age Months

ggplot(aes(x = age, y = friend\_count\_mean), data = pf.fc\_by\_age) +  
 geom\_line()



head(pf.fc\_by\_age, 10)

## Source: local data frame [10 x 4]  
##   
## age friend\_count\_mean friend\_count\_median n  
## (int) (dbl) (dbl) (int)  
## 1 13 164.7500 74.0 484  
## 2 14 251.3901 132.0 1925  
## 3 15 347.6921 161.0 2618  
## 4 16 351.9371 171.5 3086  
## 5 17 350.3006 156.0 3283  
## 6 18 331.1663 162.0 5196  
## 7 19 333.6921 157.0 4391  
## 8 20 283.4991 135.0 3769  
## 9 21 235.9412 121.0 3671  
## 10 22 211.3948 106.0 3032

pf.fc\_by\_age[17:19, ]

## Source: local data frame [3 x 4]  
##   
## age friend\_count\_mean friend\_count\_median n  
## (int) (dbl) (dbl) (int)  
## 1 29 120.8182 66.0 1936  
## 2 30 115.2080 67.5 1716  
## 3 31 118.4599 63.0 1694

pf$age\_with\_months <- pf$age + (12 - pf$dob\_month) / 12

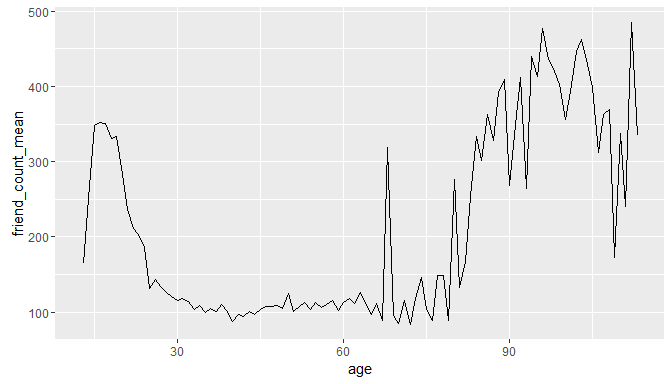
### Age with Months Means

library(dplyr)  
pf.fc\_by\_age\_months <- pf %>%  
 group\_by(age\_with\_months) %>%  
 summarise(friend\_count\_mean = mean(friend\_count),  
 friend\_count\_median = median(friend\_count),  
 n = n()) %>%  
 arrange(age\_with\_months)  
  
head(pf.fc\_by\_age\_months)

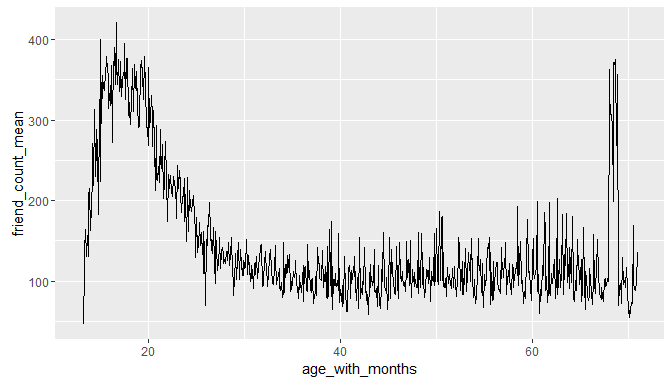
## Source: local data frame [6 x 4]  
##   
## age\_with\_months friend\_count\_mean friend\_count\_median n  
## (dbl) (dbl) (dbl) (int)  
## 1 13.16667 46.33333 30.5 6  
## 2 13.25000 115.07143 23.5 14  
## 3 13.33333 136.20000 44.0 25  
## 4 13.41667 164.24242 72.0 33  
## 5 13.50000 131.17778 66.0 45  
## 6 13.58333 156.81481 64.0 54

### Noise in Conditional Means

ggplot(aes(x = age, y = friend\_count\_mean), data = pf.fc\_by\_age) +  
 geom\_line()



# Revise using age\_with\_months, subset by age\_with\_months < 71  
ggplot(aes(x = age\_with\_months, y = friend\_count\_mean),   
 data = subset(pf.fc\_by\_age\_months, age\_with\_months < 71)) +  
 geom\_line()



### Smoothing Conditional Means

# Save each plot to a variable, add smoothing lines, plot in one column  
# Reducing number of bins reduces noise but may also leave out important details  
# bias / variance trade-off  
  
p1 <- ggplot(aes(x = age, y = friend\_count\_mean),   
 data = subset(pf.fc\_by\_age, age < 71)) +  
 geom\_line() +  
 geom\_smooth()  
  
p2 <- ggplot(aes(x = age\_with\_months, y = friend\_count\_mean),   
 data = subset(pf.fc\_by\_age\_months, age\_with\_months < 71)) +  
 geom\_line() +  
 geom\_smooth()  
  
p3 <- ggplot(aes(x = round(age / 5) \* 5, y = friend\_count),   
 data = subset(pf, age < 71)) +  
 geom\_line(stat = 'summary', fun.y = mean)  
  
library(gridExtra)  
grid.arrange(p2, p1, p3, ncol = 1)

