R Note:

Install.packates(‘ggplot2’)

Install.packates(‘RColorBrewer’)

> library(ggplot2)

Warning message:

package ‘ggplot2’ was built under R version 3.4.1

> library(RColorBrewer)

Warning message:

package ‘RColorBrewer’ was built under R version 3.4.1

[www.stackoverflow.com](http://www.stackoverflow.com)

[www.statmethods.net](http://www.statmethods.net)

exercise 1:

getwd()



> summary(mtcars)

mpg cyl disp hp

Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0

1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5

Median :19.20 Median :6.000 Median :196.3 Median :123.0

Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7

3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0

Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0

drat wt qsec vs

Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000

1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000

Median :3.695 Median :3.325 Median :17.71 Median :0.0000

Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375

3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000

Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000

am gear carb

Min. :0.0000 Min. :3.000 Min. :1.000

1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000

Median :0.0000 Median :4.000 Median :2.000

Mean :0.4062 Mean :3.688 Mean :2.812

3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000

Max. :1.0000 Max. :5.000 Max. :8.000

> subset(mtcars, mpg > 30 & hp > 100)

> subset(mtcars, mpg < 14 | disp > 390)

你需要安装并加载 knitr 包，以便使用 **KNIT HTML** 按钮。在 RStudio 控制台中运行以下命令，以安装并加载 knitr。

install.packages('knitr', dependencies = T)

library(knitr)

> mtcars$wt

[1] 2.620 2.875 2.320 3.215 3.440 3.460 3.570 3.190 3.150 3.440 3.440

[12] 4.070 3.730 3.780 5.250 5.424 5.345 2.200 1.615 1.835 2.465 3.520

[23] 3.435 3.840 3.845 1.935 2.140 1.513 3.170 2.770 3.570 2.780

> cond <- mtcars$wt < 3

> cond

[1] TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

[12] FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE

[23] FALSE FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE

> mtcars$weight\_class <- ifelse(cond, 'light', 'average')

> mtcars$weight\_class

[1] "light" "light" "light" "average" "average" "average" "average"

[8] "average" "average" "average" "average" "average" "average" "average"

[15] "average" "average" "average" "light" "light" "light" "light"

[22] "average" "average" "average" "average" "light" "light" "light"

[29] "average" "light" "average" "light"

> cond <- mtcars$wt > 3.5

> mtcars$weight\_class <- ifelse(cond, 'heavy', mtcars$weight\_class)

> mtcars$weight\_class

[1] "light" "light" "light" "average" "average" "average" "heavy"

[8] "average" "average" "average" "average" "heavy" "heavy" "heavy"

[15] "heavy" "heavy" "heavy" "light" "light" "light" "light"

[22] "heavy" "average" "heavy" "heavy" "light" "light" "light"

[29] "average" "light" "heavy" "light"



> setwd('C:/edwin/Document/Udacity R study note')

> getwd()

[1] "C:/edwin/Document/Udacity R study note"

> reddata <- read.csv('reddit.csv')

> str(reddata)

'data.frame': 32754 obs. of 14 variables:

$ id : int 1 2 3 4 5 6 7 8 9 10 ...

$ gender : int 0 0 1 0 1 0 0 0 0 0 ...

$ age.range : Factor w/ 7 levels "18-24","25-34",..: 2 2 1 2 2 2 2 1 3 2 ...

$ marital.status : Factor w/ 6 levels "Engaged","Forever Alone",..: NA NA NA NA NA 4 3 4 4 3 ...

$ employment.status: Factor w/ 6 levels "Employed full time",..: 1 1 2 2 1 1 1 4 1 2 ...

$ military.service : Factor w/ 2 levels "No","Yes": NA NA NA NA NA 1 1 1 1 1 ...

$ children : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...

$ education : Factor w/ 7 levels "Associate degree",..: 2 2 5 2 2 2 5 2 2 5 ...

$ country : Factor w/ 439 levels " Canada"," Canada eh",..: 394 394 394 394 394 394 125 394 394 125 ...

$ state : Factor w/ 53 levels "","Alabama","Alaska",..: 33 33 48 33 6 33 1 6 33 1 ...

$ income.range : Factor w/ 8 levels "$100,000 - $149,999",..: 2 2 8 2 7 2 NA 7 2 7 ...

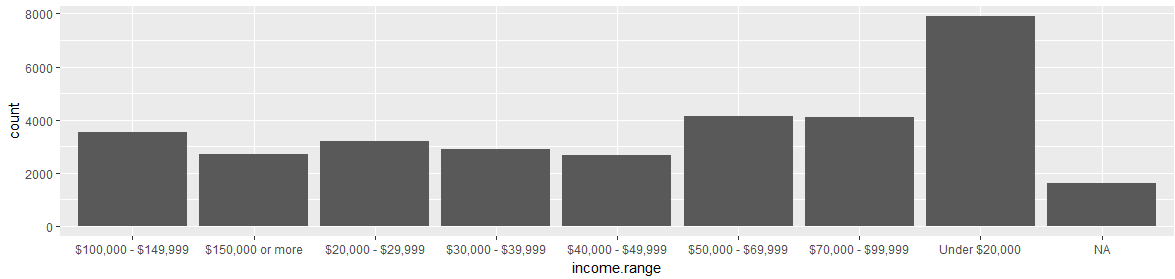
$ fav.reddit : Factor w/ 1834 levels "","'home' page (or front page if you prefer)",..: 720 691 1511 1528 188 691 1318 571 1629 1 ...

$ dog.cat : Factor w/ 3 levels "I like cats.",..: NA NA NA NA NA 2 2 2 1 1 ...

$ cheese : Factor w/ 11 levels "American","Brie",..: NA NA NA NA NA 3 3 1 10 7 ...

> qplot(data = reddata,x=age.range)

> qplot(data = reddata,x=income.range)



reddata$age.range <- ordered(reddata$age.range,levels=c("Under 18","18-24","25-34","35-44","45-54","65 or Above"))

> qplot(data = reddata,x=age.range)

Analyze Facebook data

> list.files()

[1] "~$R Note.docx" "EDA\_Course\_Materials" "EDA\_Course\_Materials.zip" "pseudo\_facebook.tsv" "R Note.docx"

[6] "reddit.csv"

>

>

> pf <- read.csv('pseudo\_facebook.tsv',sep='\t')

> str(pf)

'data.frame': 99003 obs. of 15 variables:

$ userid : int 2094382 1192601 2083884 1203168 1733186 1524765 1136133 1680361 1365174 1712567 ...

$ age : int 14 14 14 14 14 14 13 13 13 13 ...

$ dob\_day : int 19 2 16 25 4 1 14 4 1 2 ...

$ dob\_year : int 1999 1999 1999 1999 1999 1999 2000 2000 2000 2000 ...

$ dob\_month : int 11 11 11 12 12 12 1 1 1 2 ...

$ gender : Factor w/ 2 levels "female","male": 2 1 2 1 2 2 2 1 2 2 ...

$ tenure : int 266 6 13 93 82 15 12 0 81 171 ...

$ friend\_count : int 0 0 0 0 0 0 0 0 0 0 ...

$ friendships\_initiated: int 0 0 0 0 0 0 0 0 0 0 ...

$ likes : int 0 0 0 0 0 0 0 0 0 0 ...

$ likes\_received : int 0 0 0 0 0 0 0 0 0 0 ...

$ mobile\_likes : int 0 0 0 0 0 0 0 0 0 0 ...

$ mobile\_likes\_received: int 0 0 0 0 0 0 0 0 0 0 ...

$ www\_likes : int 0 0 0 0 0 0 0 0 0 0 ...

$ www\_likes\_received : int 0 0 0 0 0 0 0 0 0 0 ...

> names(pf)

[1] "userid" "age" "dob\_day" "dob\_year" "dob\_month"

[6] "gender" "tenure" "friend\_count" "friendships\_initiated" "likes"

[11] "likes\_received" "mobile\_likes" "mobile\_likes\_received" "www\_likes" "www\_likes\_received"

install.packages('ggthemes', dependencies = TRUE)   
library(ggthemes)

scale\_x\_discrete() 图层从 ggplot2 版本 2.0 开始被弃用。你可以改用 scale\_x\_continuous() 获得断点

分面的写法：

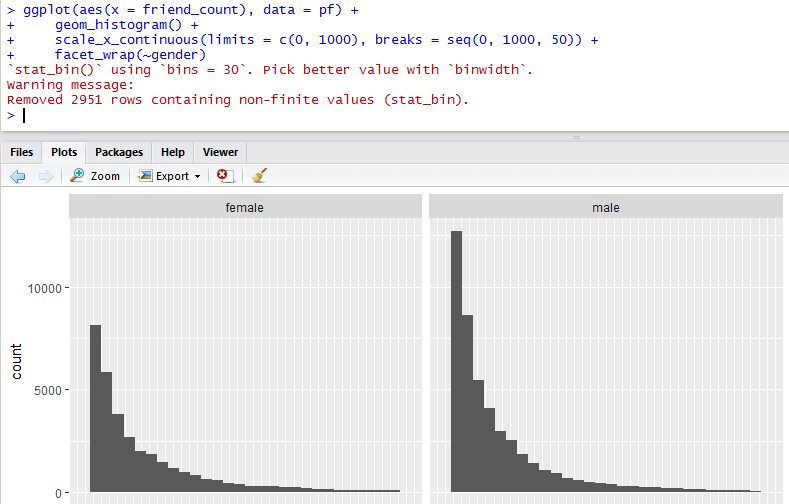
ggplot(data = pf, aes(x = dob\_day)) +   
  geom\_histogram(binwidth = 1) +   
  scale\_x\_continuous(breaks = 1:31) +   
  facet\_wrap(~dob\_month)

qplot(x=friend\_count,data=pf,xlim=c(0,3000))

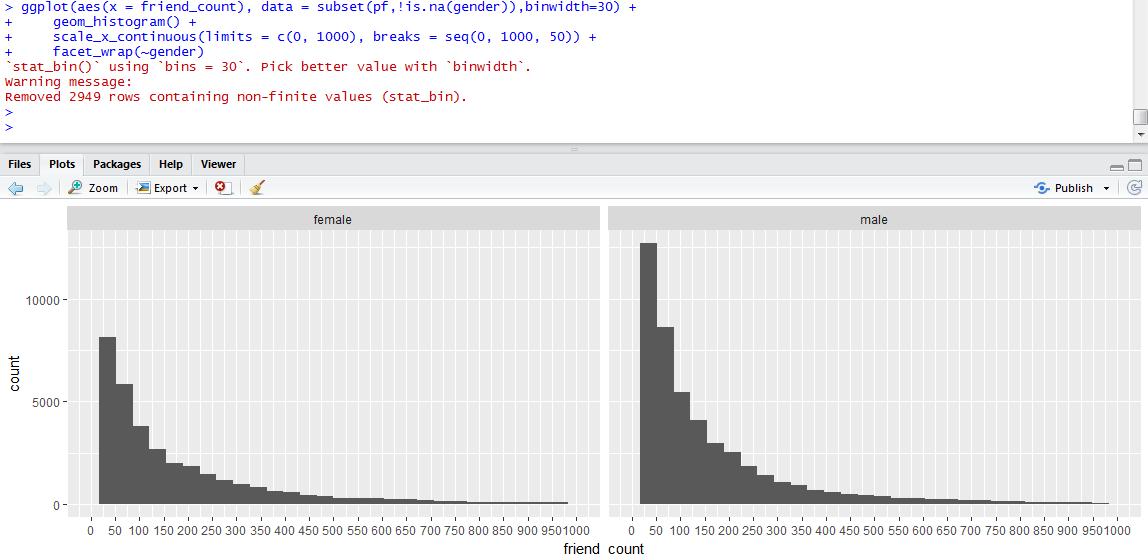
qplot(x=friend\_count,data=pf,binwidth=30)+

scale\_x\_continuous(limits = c(0, 1000),breaks= seq(0,1000,50))

ggplot(aes(x = friend\_count), data = subset(pf,!is.na(gender)),binwidth=30) +   
  geom\_histogram() +   
  scale\_x\_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 50)) +   
  facet\_wrap(~gender)



去掉NA的数据



> table(pf$gender)

female male

40254 58574