Module1\_Lesson3\_Assignment

## M1L3 Homework Assignment

Xueyi Fan 05/15/2016

*Exploratory Data Analysis assignment*

* Load the file M01\_quasi\_twitter.csv
* Answer the following questions for the data in each column:
  + How is the data distributed?
  + Test distribution assumptions (e.g. normal distributions or skewed?)
  + What are the summary statistics?
  + Are there anomalies/outliers?
* identify useful raw data & transforms (e.g. log(x))
* identify data quality problems
* identify outliers
* identify subsets of interest
* suggest functional relationships

## Answer

* Load the file M01\_quasi\_twitter.csv

library("ggplot2")  
library("reshape2")  
library("e1071")   
  
file\_L3<- "/Users/fanxueyi/Documents/NEU Bioinformatics/DSCS6030 Intro Data Mining:Machine Learing/Module1\_Getting\_to\_Know\_a\_Data\_Set/Assignment/M01\_quasi\_twitter.csv"  
  
L3Q1\_data <- read.csv(file\_L3)  
str(L3Q1\_data)

## 'data.frame': 21916 obs. of 25 variables:  
## $ screen\_name : Factor w/ 21916 levels "+5400E1.","000D0se7",..: 4341 15303 21127 13570 14085 3607 14942 8653 15547 19146 ...  
## $ created\_at\_month : int 2 11 4 3 4 2 7 5 1 1 ...  
## $ created\_at\_day : int 9 21 1 24 23 9 15 23 23 13 ...  
## $ created\_at\_year : int 2007 2009 2007 2007 2009 2009 2006 2008 2009 2009 ...  
## $ country : Factor w/ 44 levels " Germany","Argentina",..: 44 19 19 44 44 12 44 5 44 44 ...  
## $ location : Factor w/ 378 levels "Akron Ohio","Alabama",..: 188 202 25 233 211 79 365 41 242 83 ...  
## $ friends\_count : int 1087 5210 1015 338 641 917 1574 16300 8316 640 ...  
## $ followers\_count : int 22187643 6692814 6257020 3433218 2929559 2540842 1960373 1934803 1855827 1697620 ...  
## $ statuses\_count : int 60246 93910 118465 78082 93892 59397 41023 62178 56057 82912 ...  
## $ favourites\_count : int 1122 3825 1143 0 226 2122 20160 15 540 3 ...  
## $ favourited\_count : int 105005 40487 87968 25943 32589 19760 13558 25084 8732 24515 ...  
## $ dob\_day : int 29 24 4 22 9 1 2 6 15 26 ...  
## $ dob\_year : int 1999 1991 1997 1998 1963 1995 1999 1986 1991 1986 ...  
## $ dob\_month : int 4 10 3 8 11 1 11 10 2 9 ...  
## $ gender : Factor w/ 2 levels "female","male": 1 1 2 2 1 1 1 2 1 2 ...  
## $ mobile\_favourites\_count: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ mobile\_favourited\_count: int 0 5032191 0 0 0 0 0 1934803 0 0 ...  
## $ education : int 8 15 9 9 13 15 14 10 11 12 ...  
## $ experience : int 0 0 0 44 24 21 31 0 27 20 ...  
## $ age : int 29 0 32 40 45 14 27 31 34 40 ...  
## $ race : Factor w/ 10 levels "arab","asian",..: 10 10 10 10 10 10 10 10 2 1 ...  
## $ wage : num 16.3 17.9 15.7 7 17.9 ...  
## $ retweeted\_count : int 1 1 2 0 1 2 1 2 0 0 ...  
## $ retweet\_count : int 30 6 65 8 7 64 13 14 15 10 ...  
## $ height : int 156 162 168 180 162 158 160 178 156 173 ...

names(L3Q1\_data)

## [1] "screen\_name" "created\_at\_month"   
## [3] "created\_at\_day" "created\_at\_year"   
## [5] "country" "location"   
## [7] "friends\_count" "followers\_count"   
## [9] "statuses\_count" "favourites\_count"   
## [11] "favourited\_count" "dob\_day"   
## [13] "dob\_year" "dob\_month"   
## [15] "gender" "mobile\_favourites\_count"  
## [17] "mobile\_favourited\_count" "education"   
## [19] "experience" "age"   
## [21] "race" "wage"   
## [23] "retweeted\_count" "retweet\_count"   
## [25] "height"

* Answer the following questions for the data in each column:
  + How is the data distributed?
  + Test distribution assumptions (e.g. normal distributions or skewed?)
  + What are the summary statistics?
  + Are there anomalies/outliers?

### Column1 Screen\_name

head(L3Q1\_data$screen\_name)

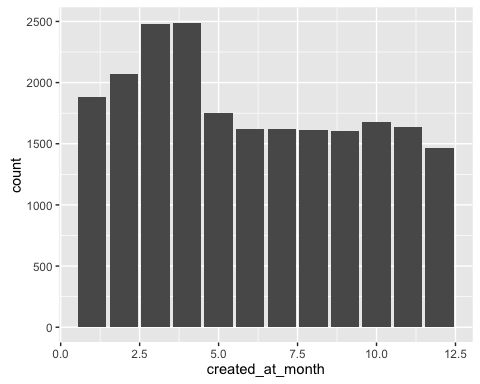
## [1] CNN osbrFe WSJ ninc nssubies BNCC   
## 21916 Levels: +5400E1. 000D0se7 001apdov 001RBTePh 003B0K2 ... zzzrnfoia

This column contains different names, This column doesn't have distribution

### Column2,3,4 created\_at\_month, day, year

* created\_at\_month

#distribution  
qplot(created\_at\_month, data = L3Q1\_data,geom = "bar")



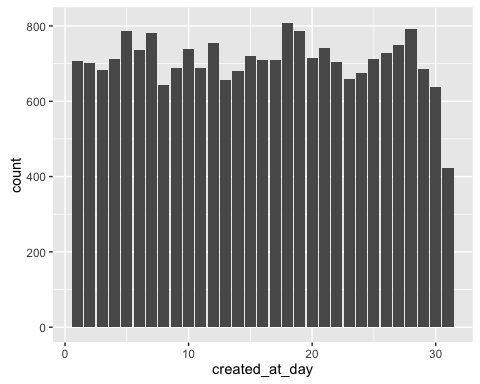
#summary statistics  
summary(L3Q1\_data$created\_at\_month)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 6.000 6.069 9.000 12.000

The data created\_at\_month is from uniform distribution. March and Apirl have more creations than other month.

* created\_at\_day

#distribution  
qplot(created\_at\_day, data = L3Q1\_data,geom = "bar")



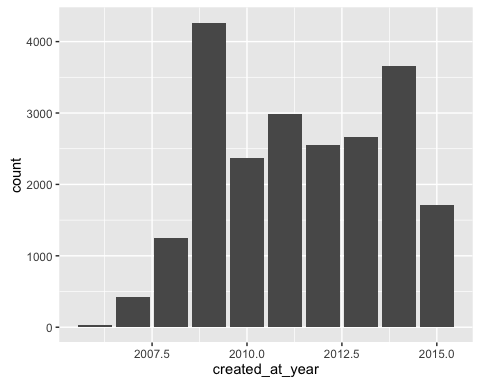
#summary statistics  
summary(L3Q1\_data$created\_at\_day)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 8.00 16.00 15.78 23.00 31.00

The data created\_at\_day is from uniform distribution. It seems users don't have a preference on the data to create a account.

* created\_at\_year

#distribution  
qplot(created\_at\_year, data = L3Q1\_data,geom = "bar")



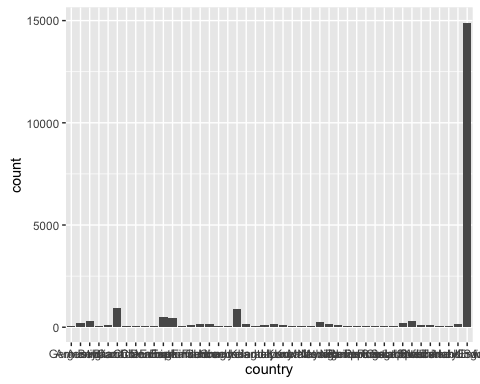
#summary statistics  
summary(L3Q1\_data$created\_at\_year)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2006 2009 2011 2011 2013 2015

It's hard to decide which kind of distribution that the created\_at\_year comes from. But the plot shows that at the beginning years, the number of usrer increases.

### Column5 country

#distribution  
qplot(country, data = L3Q1\_data,geom = "bar")



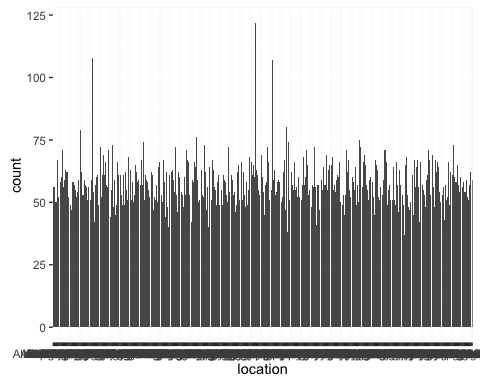
#summary statistics  
summary(L3Q1\_data$country)

## Germany Argentina Australia   
## 54 190 291   
## Belgium Brazil Canada   
## 56 121 943   
## Chile China Colombia   
## 61 57 50   
## Denmark Earth England   
## 59 516 467   
## European Union Finland France   
## 58 110 180   
## Germany Greece Hong Kong   
## 151 59 60   
## India Ireland Israel   
## 890 171 52   
## Italy Japan Kenya   
## 116 162 117   
## Kuwait Luxembourg Malaysia   
## 49 62 55   
## Mexico Netherlands Nigeria   
## 236 170 132   
## Panama Philippines Portugal   
## 59 53 49   
## Russia Scotland Singapore   
## 63 57 53   
## South Africa Spain Sweden   
## 183 283 123   
## Switzerland Turkey United Arab Emirates   
## 115 73 56   
## United Kingdom USA   
## 149 14905

The plot shows that most of the users of twitter come from USA which is far greater than other countries.

### Column6 location

#distribution  
qplot(location, data = L3Q1\_data,geom = "bar")



#summary statistics  
summary(L3Q1\_data$location)

## Mexico Boston Montreal   
## 122 108 107   
## Nevada Bangalore Indianapolis Indiana   
## 80 79 76   
## Pune India Dallas Texas New Hampshire   
## 75 74 74   
## Cambridge MA Istanbul Vancouver BC   
## 73 73 73   
## Brooklyn New York Fremont CA London United Kingdom   
## 72 72 72   
## Minnesota NY Raleigh NC   
## 72 72 72   
## Arizona california Houston Texas   
## 71 71 71   
## Nigeria Philly San Jose CA   
## 71 71 71   
## San Jose California The Netherlands Buenos Aires Argentina   
## 71 71 69   
## Miami Florida Orange County CA Richmond VA   
## 69 69 69   
## Tokyo Japan Cleveland Maryland USA   
## 69 68 68   
## New York USA PA South Africa   
## 68 68 68   
## Alexandria VA CT Espana   
## 67 67 67   
## Houston TX Kansas City Nebraska   
## 67 67 67   
## Phoenix AZ San Diego CA Stamford CT   
## 67 67 67   
## Sweden The World Toronto ON   
## 67 67 67   
## Buffalo NY Chennai East Coast   
## 66 66 66   
## hyderabad Indiana Madison Wisconsin   
## 66 66 66   
## Massachusetts Mississippi new york   
## 66 66 66   
## Pasadena CA Rhode Island Santa Barbara CA   
## 66 66 66   
## SF Sydney Australia Toronto Ontario Canada   
## 66 66 66   
## usa Columbus OH MA   
## 66 65 65   
## Melbourne Australia NJ Orlando Florida   
## 65 65 65   
## Ottawa Canada Rochester NY San Diego California   
## 65 65 65   
## Somewhere Virginia Earth   
## 65 65 64   
## Indianapolis Johannesburg Oklahoma City OK   
## 64 64 64   
## Portland Oregon Seattle Stockton   
## 64 64 64   
## TX ATL Cleveland OH   
## 64 63 63   
## Fort Worth TX Ireland Kansas City MO   
## 63 63 63   
## Mexico City Moscow Orlando FL   
## 63 63 63   
## san francisco Silicon Valley Atlanta   
## 63 63 62   
## Atlanta GA Bangalore India Cincinnati OH   
## 62 62 62   
## Columbus Ohio Denver Colorado Finland   
## 62 62 62   
## (Other)   
## 15131

The location data are from uniform distribution. but there are three locations have a greater value than other locations which seem to be outliers.

### Column7,8,9,10,11 friend\_count, followers\_count, statuses\_count, favourites\_count, favourited\_count

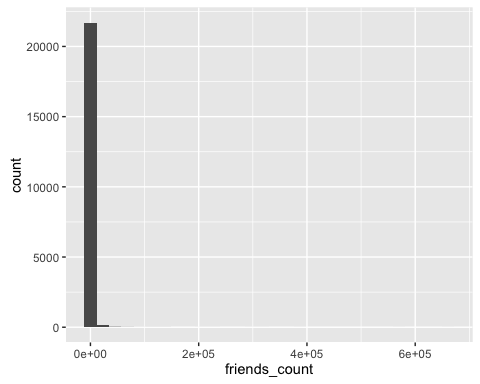
* friend\_count

#summary statistics  
summary(L3Q1\_data$friends\_count)

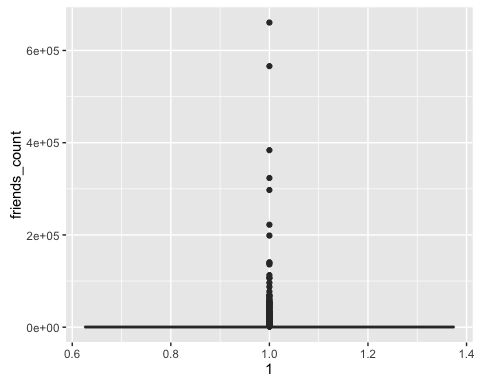
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -84 123 324 1058 849 660500

#distribution  
qplot(friends\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, friends\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from -84 to 660500, the mean is 1058 and median is 324. The minimum number is negative which is outlier.

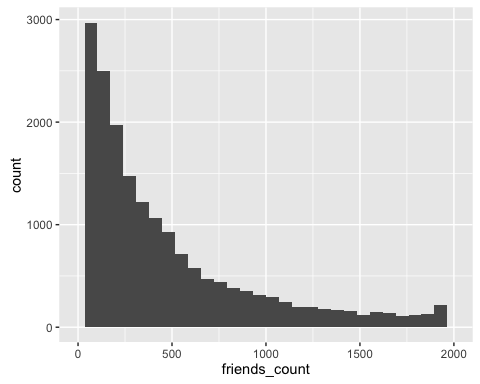
The boxplot also show there are at least 9 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 2e+03

#distribution  
qplot(friends\_count, data = L3Q1\_data, xlim = c(0,2\*10^3) )

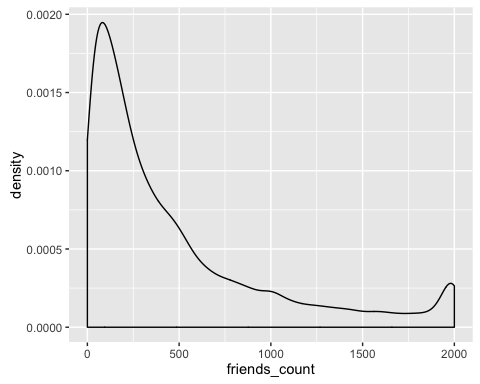
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1732 rows containing non-finite values (stat\_bin).



qplot(friends\_count, data = L3Q1\_data, xlim = c(0,2\*10^3), geom = "density")

## Warning: Removed 1732 rows containing non-finite values (stat\_density).



The major number of friends are around 100.

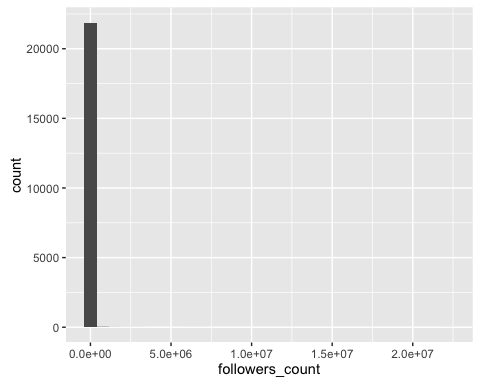
* followers\_count

#summary statistics  
summary(L3Q1\_data$followers\_count)

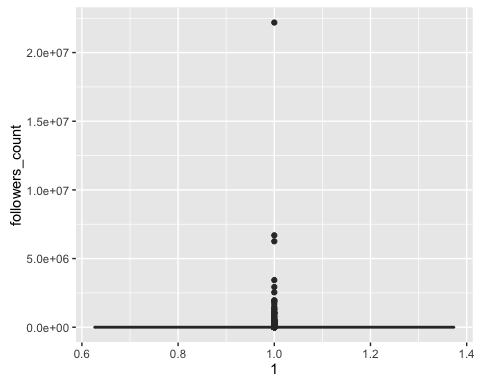
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 105 336 5859 1075 22190000

#distribution  
qplot(followers\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, followers\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 22190000, the mean is 5859 and median is 336. The median is not equal to the mean which implicate that there are some outlieres in the data.

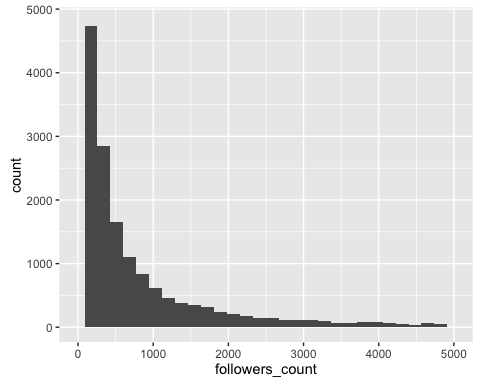
The boxplot also show there are at least 6 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 5.0e+03

#distribution  
qplot(followers\_count, data = L3Q1\_data, xlim = c(0,5\*10^3) )

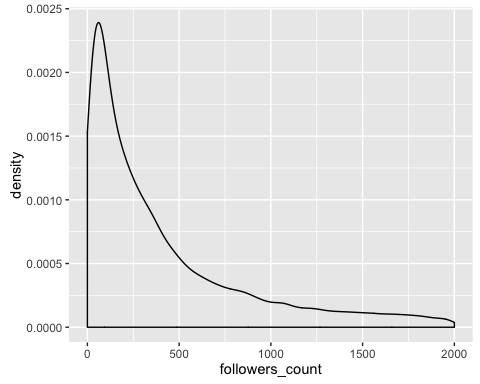
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1870 rows containing non-finite values (stat\_bin).



qplot(followers\_count, data = L3Q1\_data, xlim = c(0,2\*10^3), geom = "density")

## Warning: Removed 3563 rows containing non-finite values (stat\_density).



The major number of followers are around 100.

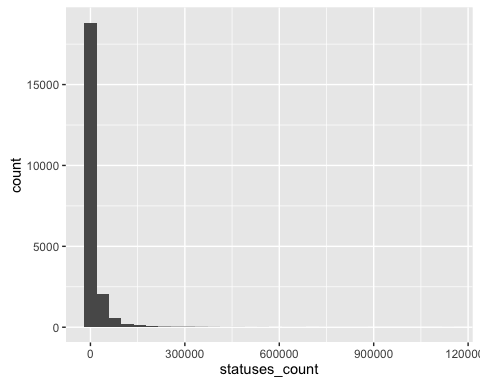
* statuses\_count

#summary statistics  
summary(L3Q1\_data$statuses\_count)

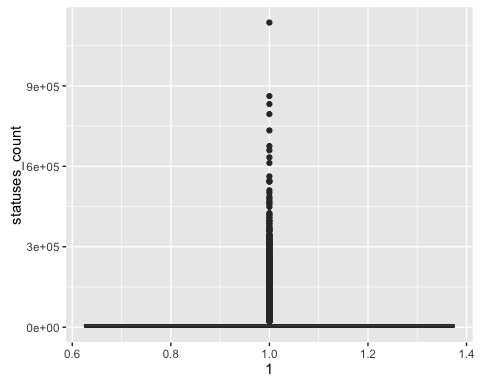
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 558 2341 12490 9348 1136000

#distribution  
qplot(statuses\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, statuses\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 1 to 1136000, the mean is 12490 and median is 2341. The median is not equal to the mean which implicate that there are some outlieres in the data.

The boxplot also show there are at least 15 outliers above the top whisker.

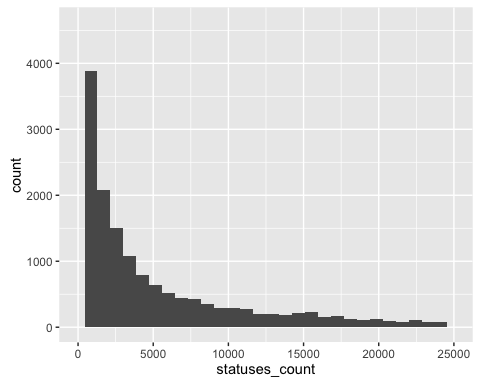
This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 25000

#distribution  
qplot(statuses\_count, data = L3Q1\_data, xlim = c(-1,25000) )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

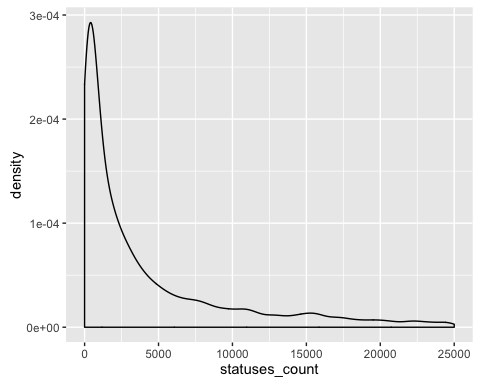
## Warning: Removed 2490 rows containing non-finite values (stat\_bin).

## Warning: Removed 1 rows containing missing values (geom\_bar).



qplot(statuses\_count, data = L3Q1\_data, xlim = c(-1,25000), geom = "density")

## Warning: Removed 2490 rows containing non-finite values (stat\_density).



The major number of followers are around 1000.

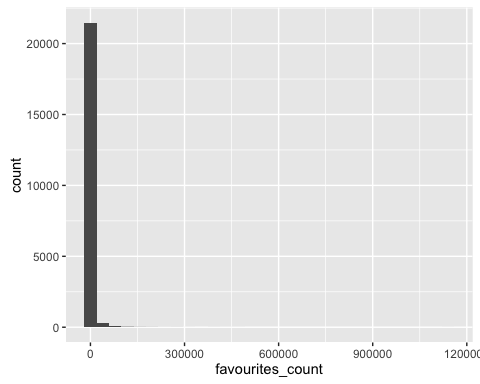
* favourites\_count

#summary statistics  
summary(L3Q1\_data$favourites\_count)

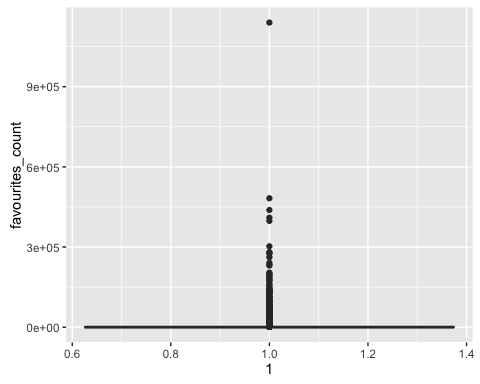
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 16 164 2217 950 1140000

#distribution  
qplot(favourites\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, favourites\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 1140000, the mean is 2217 and median is 164. The median is not equal to the mean which implicate that there are some outlieres in the data.

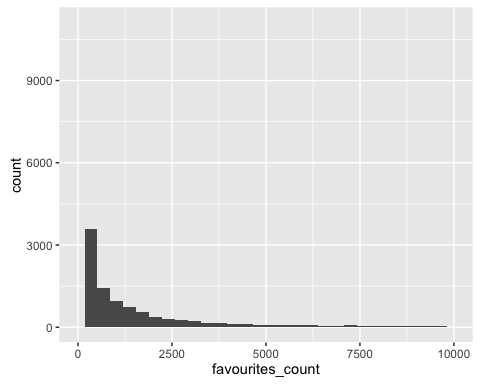
The boxplot also show there are at least 10 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 10000

#distribution  
qplot(favourites\_count, data = L3Q1\_data, xlim = c(0,10^4) )

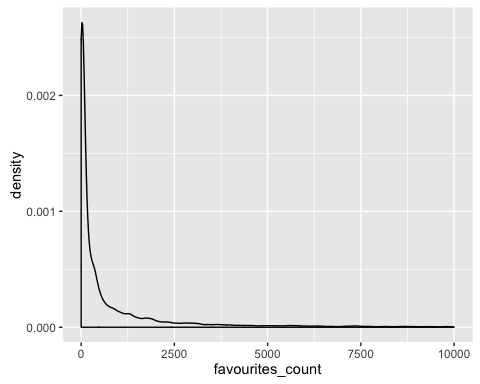
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 934 rows containing non-finite values (stat\_bin).



qplot(favourites\_count, data = L3Q1\_data, xlim = c(0,10^4), geom = "density")

## Warning: Removed 934 rows containing non-finite values (stat\_density).



The major number of followers are around 500.

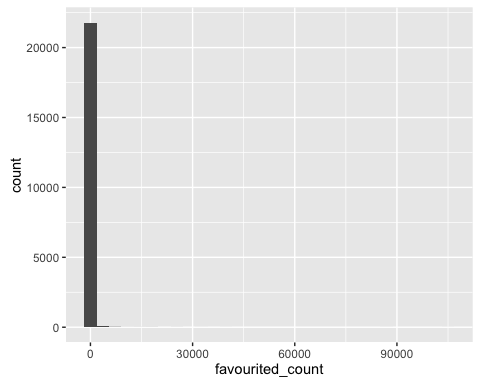
* favourited\_count

#summary statistics  
summary(L3Q1\_data$favourited\_count)

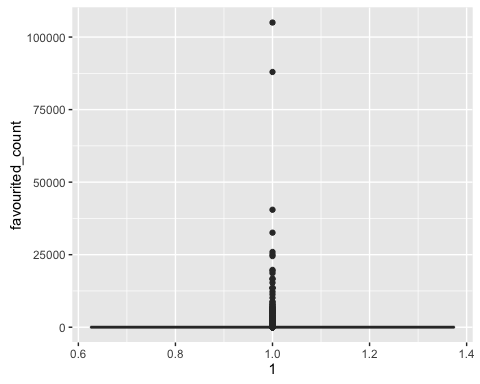
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 2.00 9.00 92.24 36.00 105000.00

#distribution  
qplot(favourited\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, favourited\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 105000, the mean is 92.24 and median is 9. The median is not equal to the mean which implicate that there are some outlieres in the data.

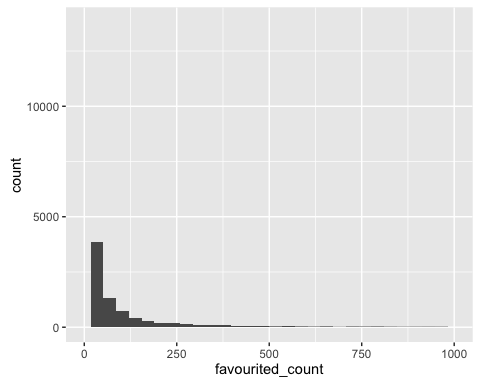
The boxplot also show there are at least 10 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 7000

#distribution  
qplot(favourited\_count, data = L3Q1\_data, xlim = c(0,1000) )

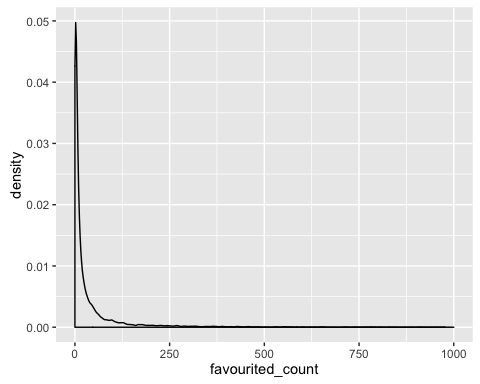
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 258 rows containing non-finite values (stat\_bin).



qplot(favourited\_count, data = L3Q1\_data, xlim = c(0,1000), geom = "density")

## Warning: Removed 258 rows containing non-finite values (stat\_density).



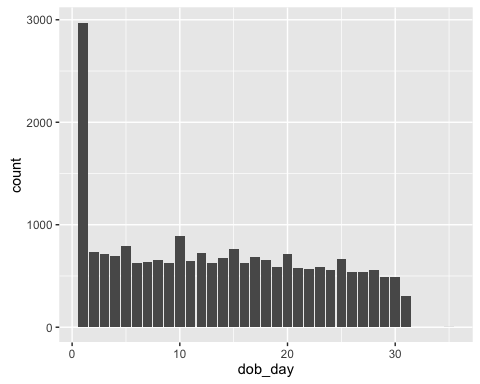
The major number of followers are around 30.

However,the histogram of all count column can't show the distribution of these data. The data need to do some transformation.

### Column12,13,14 dob\_day, dob\_year, dob\_month

* dob\_day

#distribution  
qplot(dob\_day, data = L3Q1\_data,geom = "bar")



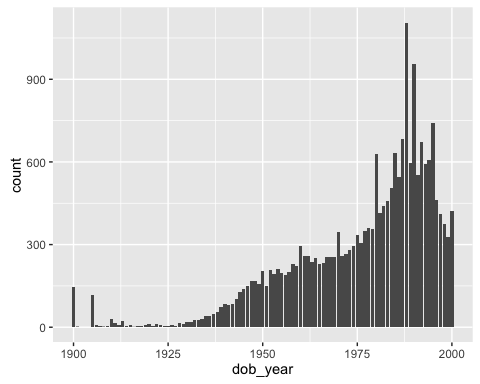
#summary statistics  
summary(L3Q1\_data$dob\_day)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 5.00 13.00 13.49 21.00 35.00

The data dob\_day is from uniform distribution.But there are more people on first day of month than other days.

* dob\_year

#distribution  
qplot(dob\_year, data = L3Q1\_data,geom = "bar")



#summary statistics  
summary(L3Q1\_data$dob\_year)

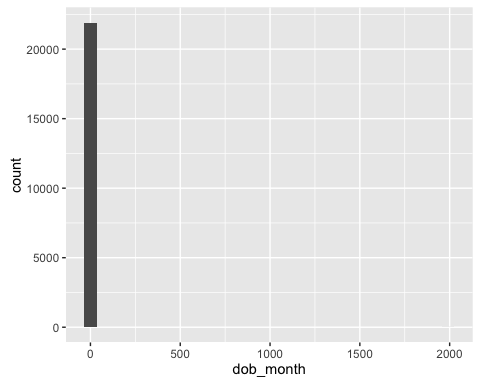
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1900 1965 1982 1976 1990 2000

The data dob\_year are from 1900 to 2000. However, the year 1900 is not what we expect which may be a outlier. Most of people are bone in 1987 and 1989.

* dob\_month

#distribution  
qplot(dob\_month, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#summary statistics  
summary(L3Q1\_data$dob\_month)

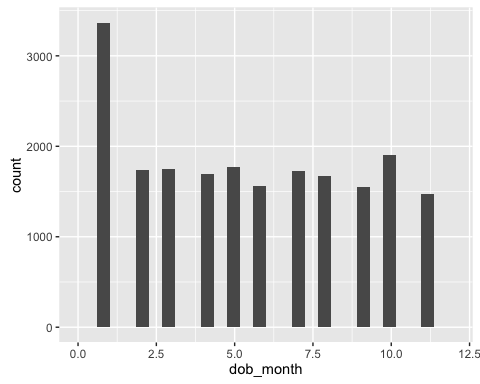
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 6.000 6.398 9.000 1992.000

The data dob\_month are from 1 to 1992. Obviously, the max number is a outliers.

#distribution  
qplot(dob\_month, data = L3Q1\_data, xlim = c(0,12))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

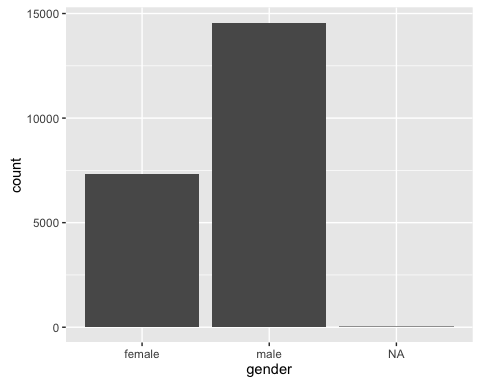
## Warning: Removed 4 rows containing non-finite values (stat\_bin).



The distribution of dob\_month is unifrom distribution. However, more people are born on Janunary than other months.

### Column15 gender

#distribution  
qplot(gender, data = L3Q1\_data)



#summary statistics  
summary(L3Q1\_data$gender)

## female male NA's   
## 7319 14569 28

There are more male than female using the twitter.

### Column16,17 mobile\_favourites\_count, mobile\_favourited\_count

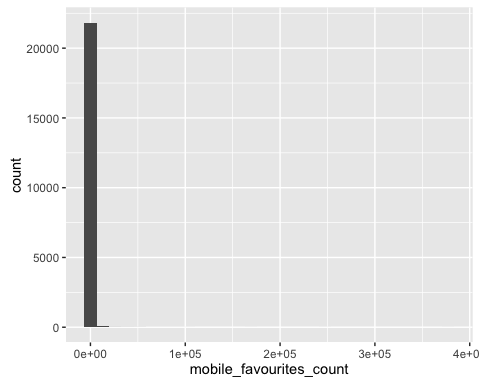
* mobile\_favourites\_count

#summary statistics  
summary(L3Q1\_data$mobile\_favourites\_count)

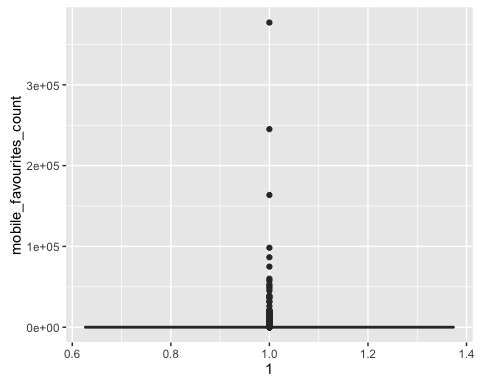
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 152.9 0.0 377100.0

#distribution  
qplot(mobile\_favourites\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, mobile\_favourites\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 377100, the mean is 152.9 and median is 0.0. The median is not equal to the mean which implicate that there are some outlieres in the data.

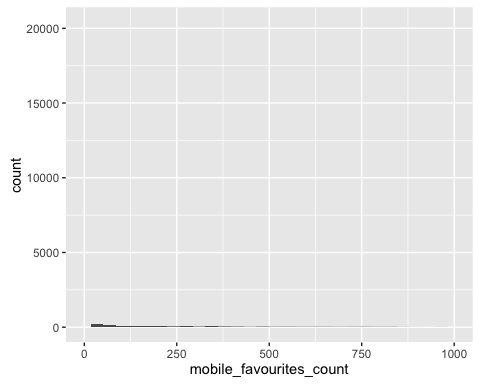
The boxplot also show there are at least 6 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 1000

#distribution  
qplot(mobile\_favourites\_count, data = L3Q1\_data, xlim = c(0,1000) )

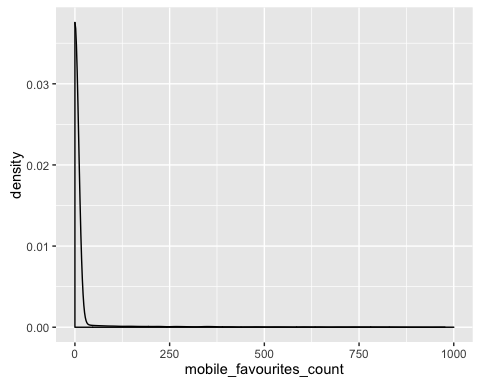
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 359 rows containing non-finite values (stat\_bin).



qplot(mobile\_favourites\_count, data = L3Q1\_data, xlim = c(0,1000), geom = "density")

## Warning: Removed 359 rows containing non-finite values (stat\_density).



The major number of followers are around 0.

However,the histogram of all count column can't show the distribution of these data. The data need to do some transformation.

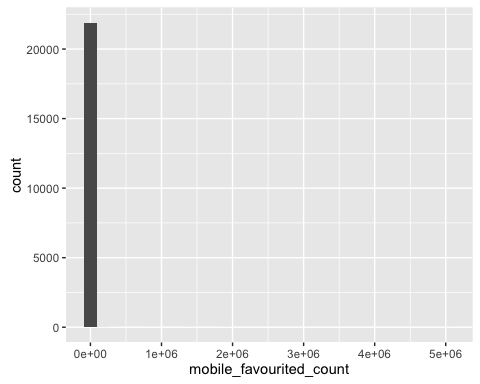
* mobile\_favourited\_count

#summary statistics  
summary(L3Q1\_data$mobile\_favourited\_count)

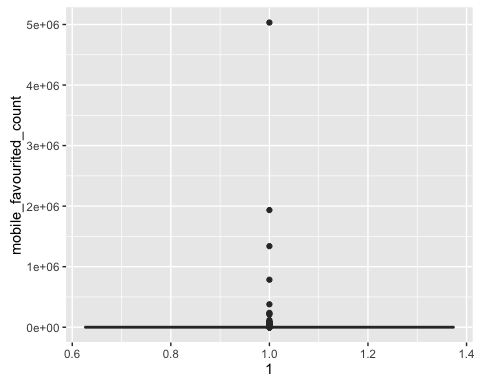
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 649 0 5032000

#distribution  
qplot(mobile\_favourited\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, mobile\_favourited\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 5032000, the mean is 649 and median is 0. The median is not equal to the mean which implicate that there are some outlieres in the data.

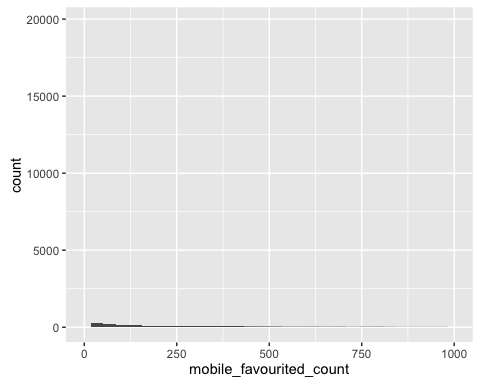
The boxplot also show there are at least 6 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 1000

#distribution  
qplot(mobile\_favourited\_count, data = L3Q1\_data, xlim = c(0,1000) )

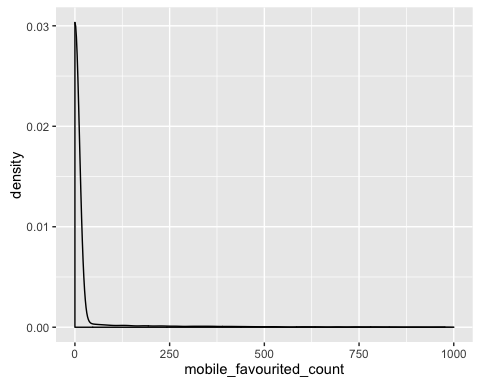
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 489 rows containing non-finite values (stat\_bin).



qplot(mobile\_favourited\_count, data = L3Q1\_data, xlim = c(0,1000), geom = "density")

## Warning: Removed 489 rows containing non-finite values (stat\_density).



The major number of followers are around 0.

However,the histogram of all count column can't show the distribution of these data. The data need to do some transformation.

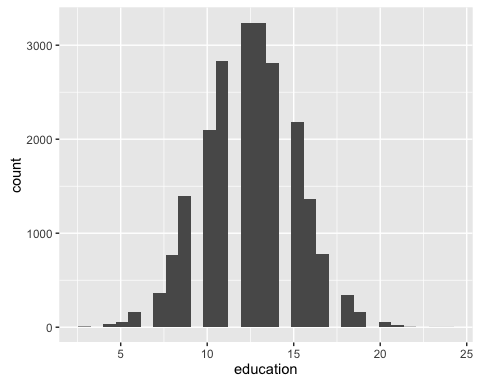
### Column18 education

#summary statistics  
summary(L3Q1\_data$education)

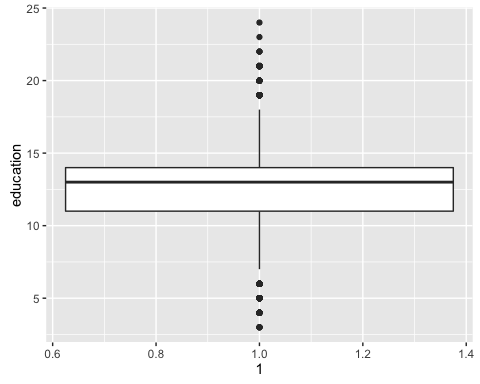
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.0 11.0 13.0 12.5 14.0 24.0

#distribution  
qplot(education, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, education, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 3 to 24, the mean is 12.5 and median is 13. Histogram shows that these data looks like poisson distribution. And the number of people having 12.5 years education are greater than other years.

The boxplot shows that there are some outliers outside the whiskers.

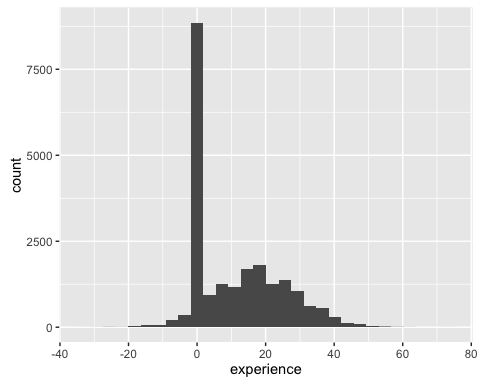
### Column19 experience

#summary statistics  
summary(L3Q1\_data$experience)

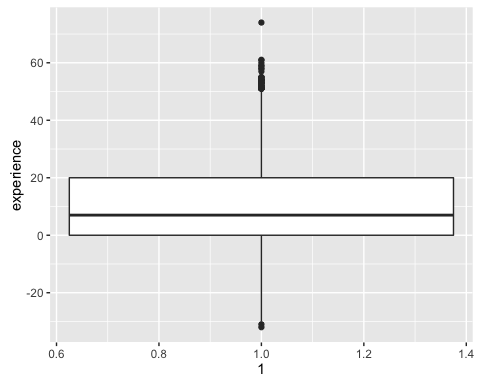
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -32.00 0.00 7.00 10.88 20.00 74.00

#distribution  
qplot(experience, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, experience, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from -32 to 74, the mean is 10.88 and median is 7. Histogram shows that these data looks like normal distribution. And the number of people having no experience are greater than other years. The one possible reason is that people don't choose the experience years and 0 is setted by default.

However there are some outliers whose value is less than 0. The boxplot also shows some outliers above the top whisker.

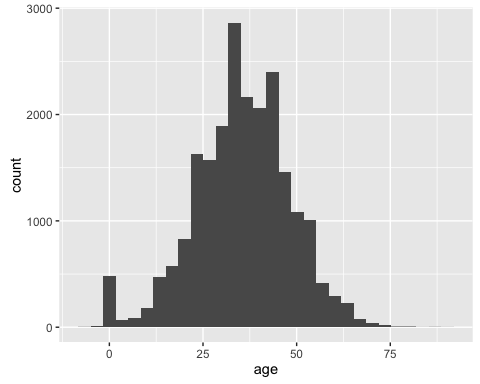
### Column20 age

#summary statistics  
summary(L3Q1\_data$age)

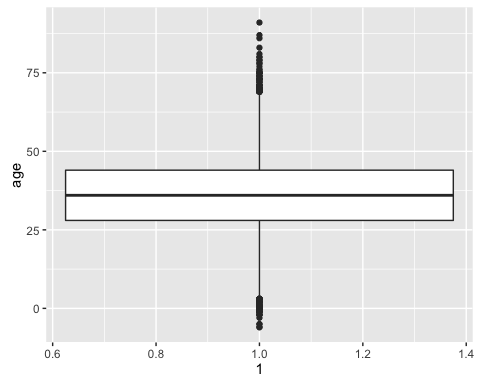
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -6.00 28.00 36.00 35.54 44.00 91.00

#distribution  
qplot(age, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, age, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from -6 to 91, the mean is 35.54 and median is 36. Histogram shows that these data looks like normal distribution. And the number of people about 30 years old are greater than other years. The one possible reason is that people don't choose the age and 0 is setted by default.

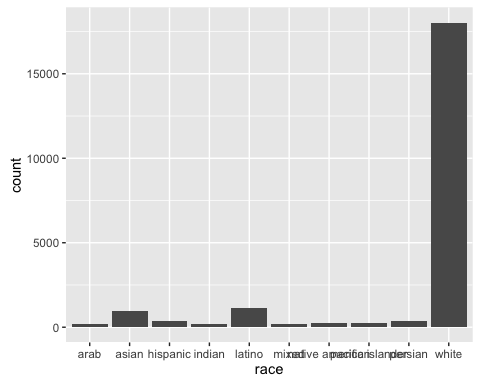
However there are some outliers whose value is less than 0. The boxplot also shows some outliers above the top whisker and below the bottom whisker.

### Column21 race

#summary statistics  
summary(L3Q1\_data$race)

## arab asian hispanic indian   
## 187 960 353 162   
## latino mixed native american pacific islander   
## 1115 199 256 276   
## persian white   
## 376 18032

#distribution  
qplot(race, data = L3Q1\_data)



Histogram shows that white is dominant among other races users.

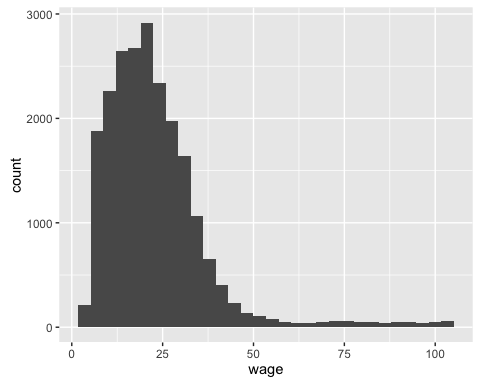
### Column22 wage

#summary statistics  
summary(L3Q1\_data$wage)

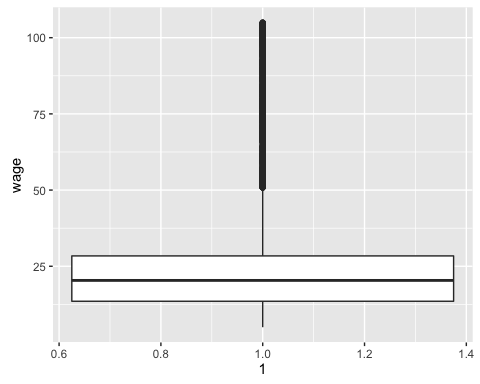
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 13.52 20.36 22.97 28.40 105.00

#distribution  
qplot(wage, data = L3Q1\_data)

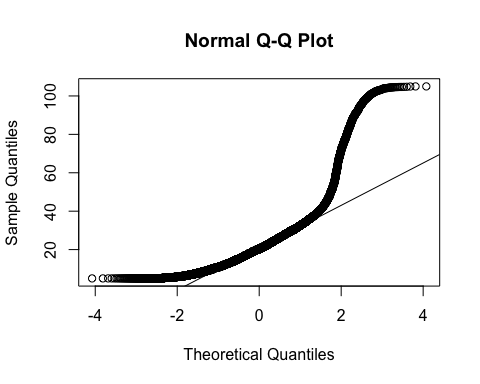
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, wage, data = L3Q1\_data, geom = "boxplot")



#normality  
qqnorm(L3Q1\_data$wage)  
qqline(L3Q1\_data$wage)



The summary statistics show that the range of data is from 5 to 105, the mean is 22.97 and median is 20.36. Histogram shows that these data looks like normal distribution. And the number of people having wage around 25 are greater than other years.

However there are some outliers whose value is less than 0. The boxplot also shows some outliers above the top whisker.

### Column23,24 retweeted\_count, retweet\_count

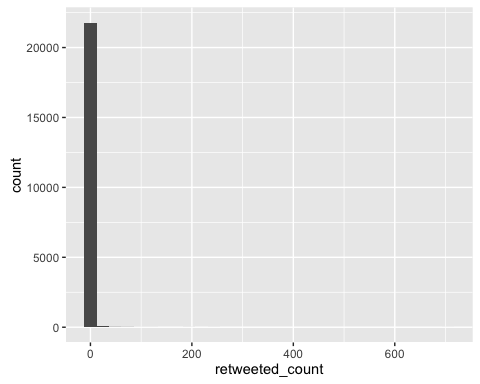
* retweeted\_count

#summary statistics  
summary(L3Q1\_data$retweeted\_count)

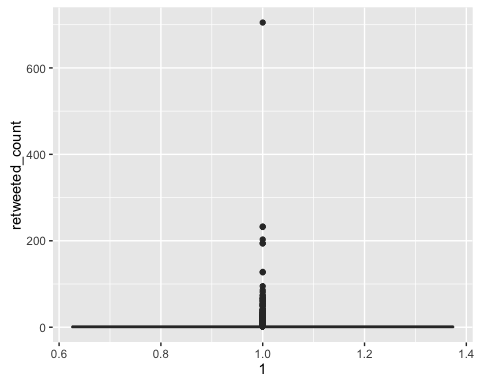
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 1.0000 0.9715 1.0000 705.0000

#distribution  
qplot(retweeted\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, retweeted\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 705, the mean is 0.9715 and median is 1.

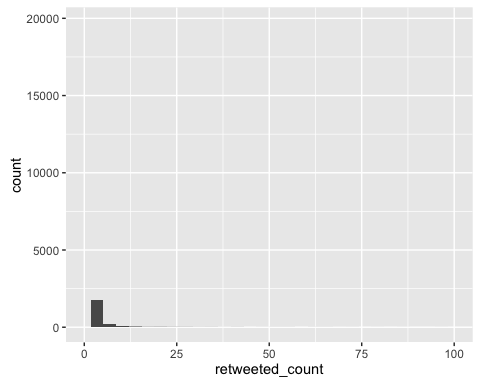
The boxplot also show there are at least 5 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 100

#distribution  
qplot(retweeted\_count, data = L3Q1\_data, xlim = c(0,100) )

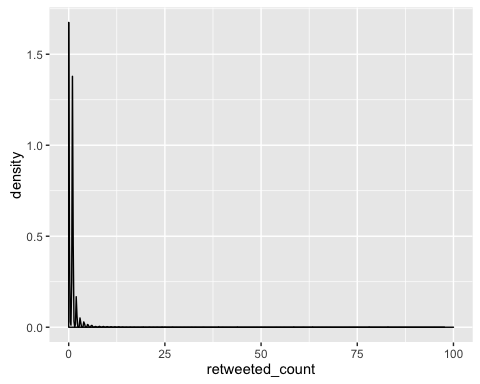
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 8 rows containing non-finite values (stat\_bin).



qplot(retweeted\_count, data = L3Q1\_data, xlim = c(0,100), geom = "density")

## Warning: Removed 8 rows containing non-finite values (stat\_density).



The major number of followers are around 0.

However,the histogram of all count column can't show the distribution of these data. The data need to do some transformation.

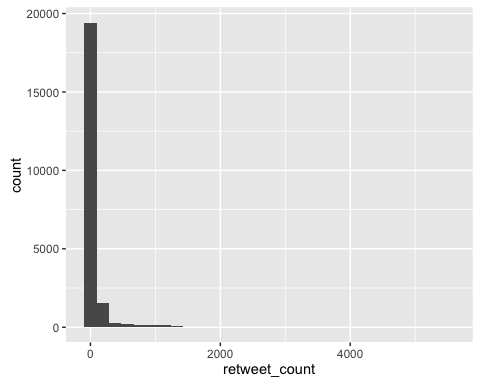
* retweet\_count

#summary statistics  
summary(L3Q1\_data$retweet\_count)

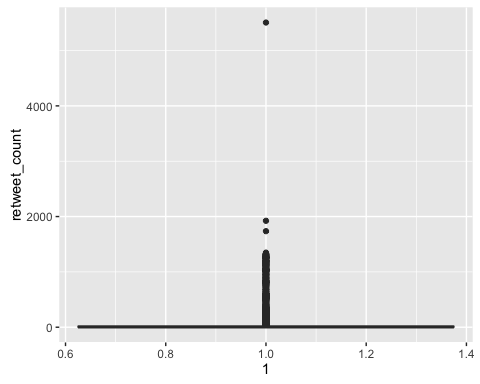
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 3.00 52.73 19.00 5506.00

#distribution  
qplot(retweet\_count, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, retweet\_count, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 0 to 5506, the mean is 52.73 and median is 3. The median is not equal to the mean which implicate that there are some outlieres in the data.

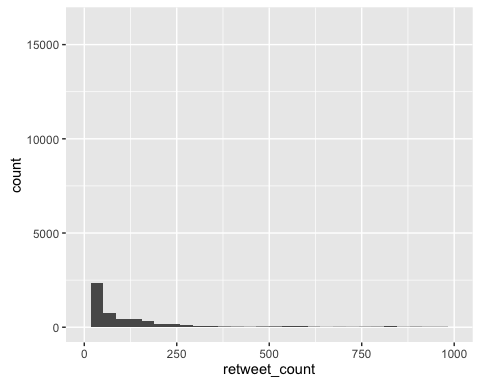
The boxplot also show there are at least 3 outliers above the top whisker.

This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 0 to 1000

#distribution  
qplot(retweet\_count, data = L3Q1\_data, xlim = c(0,1000) )

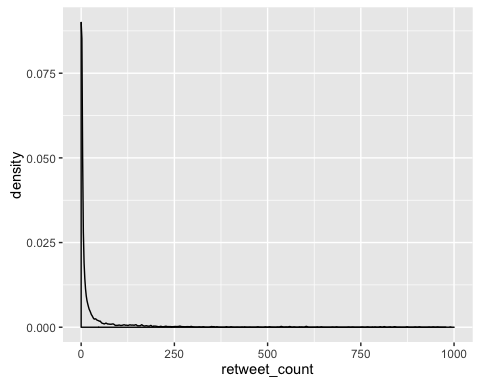
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 290 rows containing non-finite values (stat\_bin).



qplot(retweet\_count, data = L3Q1\_data, xlim = c(0,1000), geom = "density")

## Warning: Removed 290 rows containing non-finite values (stat\_density).



The major number of followers are around 0.

However,the histogram of all count column can't show the distribution of these data. The data need to do some transformation.

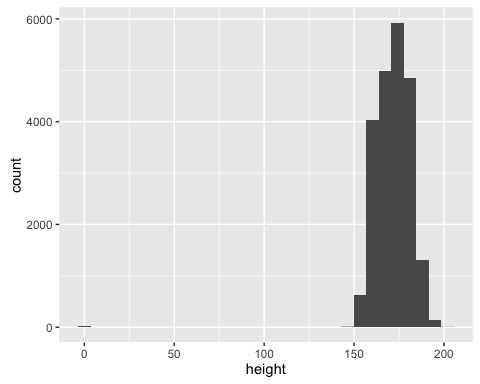
### Column25 height

#summary statistics  
summary(L3Q1\_data$height)

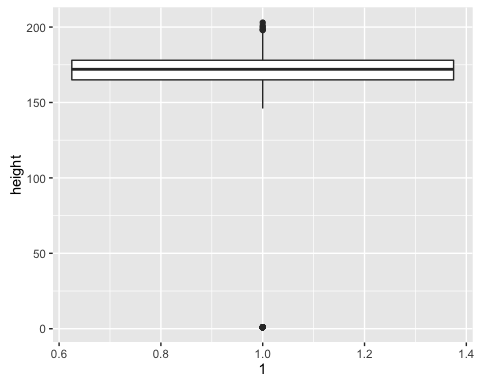
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 165.0 172.0 171.5 178.0 203.0

#distribution  
qplot(height, data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Outliers  
qplot(1, height, data = L3Q1\_data, geom = "boxplot")



The summary statistics show that the range of data is from 1 to 203, the mean is 172 and median is 171.5.

The boxplot also show there are at least 3 outliers above the top whisker and below the bottom whisker.

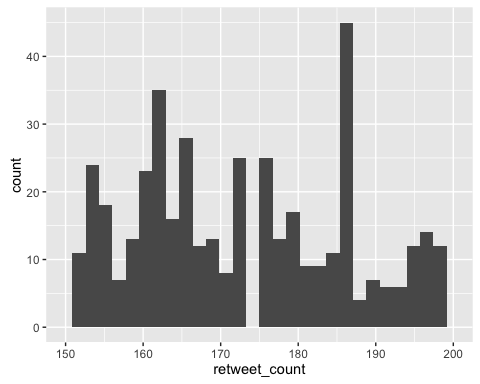
This histogram couldn't tell us the distribution of the data because the existence of outliers. So I zoom in the range from 150 to 2000

#distribution  
qplot(retweet\_count, data = L3Q1\_data, xlim = c(150,200) )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

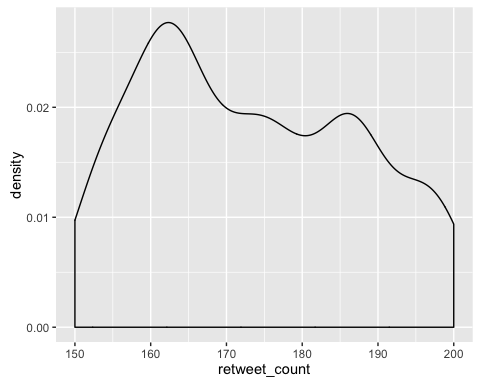
## Warning: Removed 21472 rows containing non-finite values (stat\_bin).

## Warning: Removed 1 rows containing missing values (geom\_bar).



qplot(retweet\_count, data = L3Q1\_data, xlim = c(150,200), geom = "density")

## Warning: Removed 21472 rows containing non-finite values (stat\_density).



However,the histogram of all count columns can't show the distribution of these data. The data need to do some transformation.

### identify useful raw data & transforms (e.g. log(x))

When I explore the data from each column, I found some \*\_count data are fat tailed and can be log transformed.

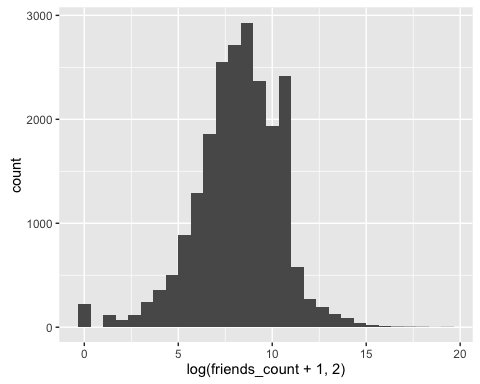
* friend\_count

#new distribution  
qplot(x=log(friends\_count+1,2), data=L3Q1\_data)

## Warning: NaNs produced  
  
## Warning: NaNs produced  
  
## Warning: NaNs produced

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1 rows containing non-finite values (stat\_bin).

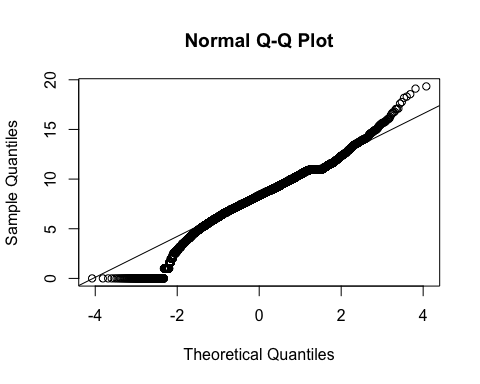


#normality  
  
qqnorm(log(((L3Q1\_data$friends\_count)+1),2))

## Warning in qqnorm(log(((L3Q1\_data$friends\_count) + 1), 2)): NaNs produced

qqline(log(((L3Q1\_data$friends\_count)+1),2))

## Warning in quantile(y, probs, names = FALSE, type = qtype, na.rm = TRUE):  
## NaNs produced

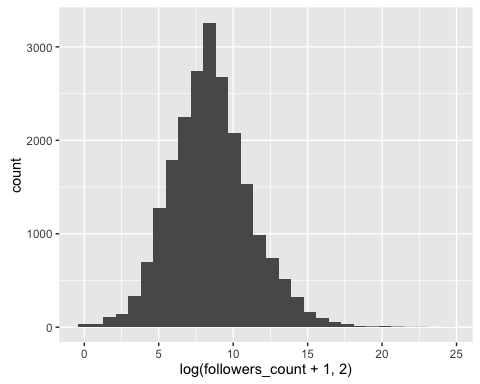


After log transformation, the data seems to be a normal distribution. But qq-plot show the log-data are not fitted very well.

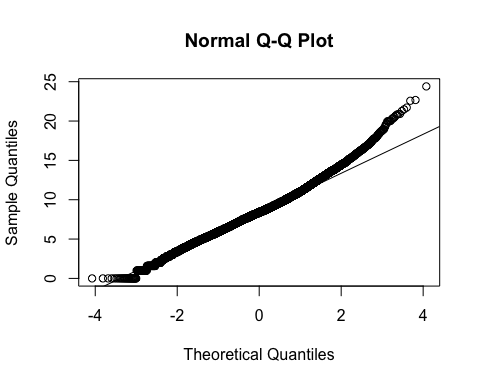
* followers\_count

#new distribution  
qplot(x=log(followers\_count+1,2), data=L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#normality  
  
qqnorm(log(((L3Q1\_data$followers\_count)+1),2))  
qqline(log(((L3Q1\_data$followers\_count)+1),2))

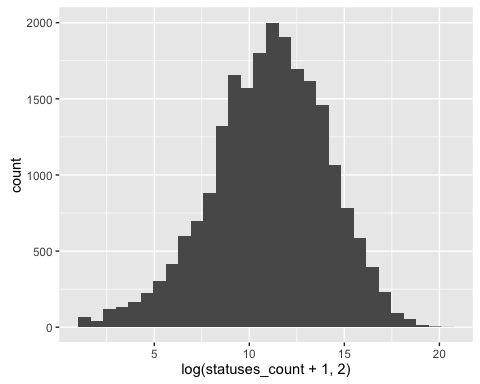


After log transformation, the data seems to be a normal distribution. And qq-plot show the log-data whose range is from -2 to 2 are fitted very well.

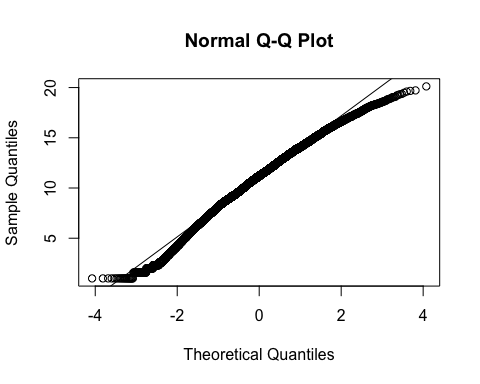
* statuses\_count

#distribution  
qplot(x=log(statuses\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#normality  
  
qqnorm(log(((L3Q1\_data$statuses\_count)+1),2))  
qqline(log(((L3Q1\_data$statuses\_count)+1),2))

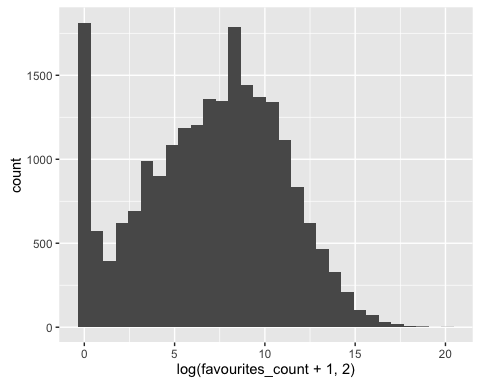


After log transformation, the data seems to be a normal distribution. And qq-plot show the log-data whose range is from -2 to 2 are fitted very well.

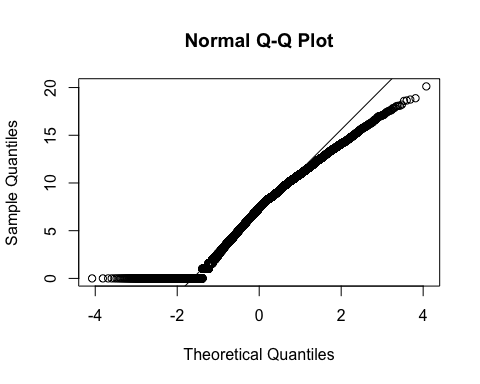
* favourites\_count

#distribution  
qplot(x=log(favourites\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#normality  
  
qqnorm(log(((L3Q1\_data$favourites\_count)+1),2))  
qqline(log(((L3Q1\_data$favourites\_count)+1),2))

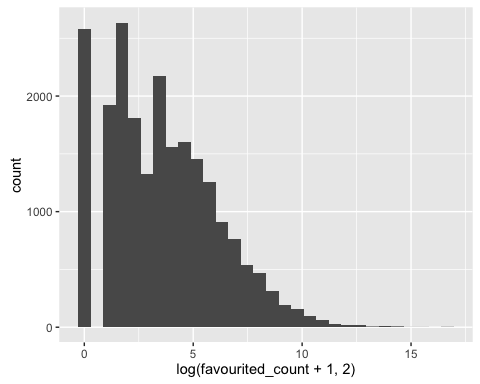


After log transformation, the data seems to be a normal distribution. But qq-plot show the log-data are not fitted very well.

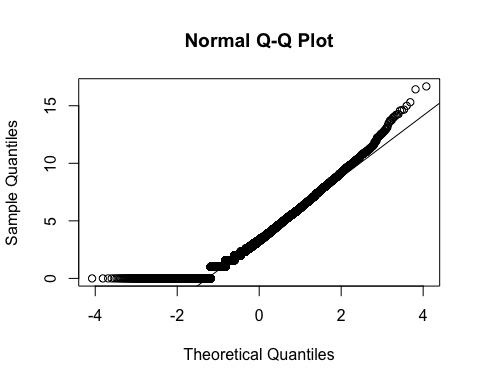
* favourited\_count

#distribution  
qplot(x=log(favourited\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#normality  
  
qqnorm(log(((L3Q1\_data$favourited\_count)+1),2))  
qqline(log(((L3Q1\_data$favourited\_count)+1),2))

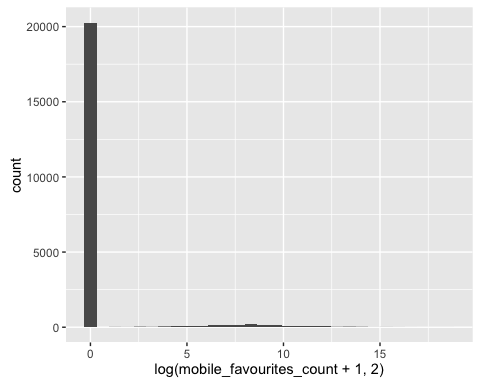


After log transformation, part of the data seems to be a normal distribution. And qq-plot show the log-data whose range is from 0 to 2 are fitted very well.

* mobile\_favourites\_count

#distribution  
qplot(x=log(mobile\_favourites\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

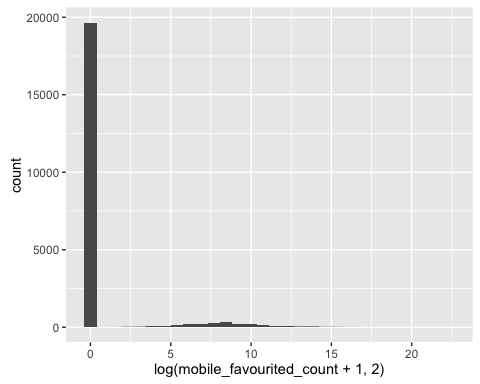


After log transformation, the data are still not good. The data should use another way of transformation.

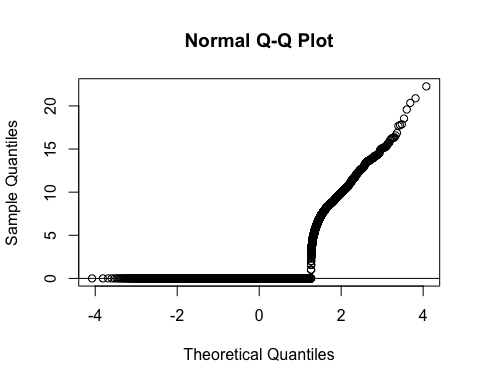
* mobile\_favourited\_count

#distribution  
qplot(x=log(mobile\_favourited\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#normality  
  
qqnorm(log(((L3Q1\_data$mobile\_favourited\_count)+1),2))  
qqline(log(((L3Q1\_data$mobile\_favourited\_count)+1),2))

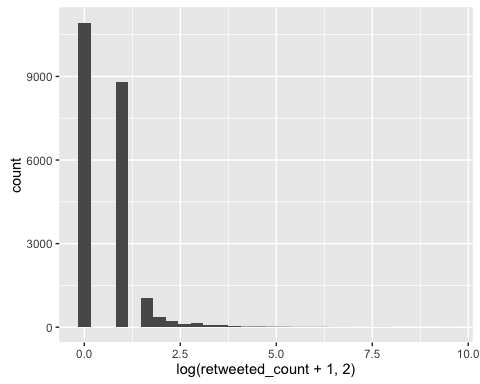


After log transformation, the data are still not good. The data should use another way of transformation.

* retweeted\_count

#distribution  
qplot(x=log(retweeted\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

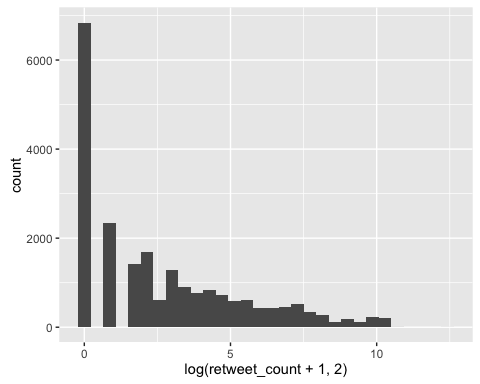


After log transformation, the data are still not good. The data should use another way of transformation.

* retweet\_count

#distribution  
qplot(x=log(retweet\_count+1,2), data = L3Q1\_data)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



After log transformation, the data are still not good. The data should use another way of transformation.

### identify data quality problems

When I explore the data of each column, some of them have obvious mistake.

For example:

1. negative value in friends\_count, experience, age
2. year of birthday is unbelivable which is too early
3. age which is equal 0 is unbelivable

### identify outliers

Just as shown in the above, I use boxplot to find the outliers of each column.

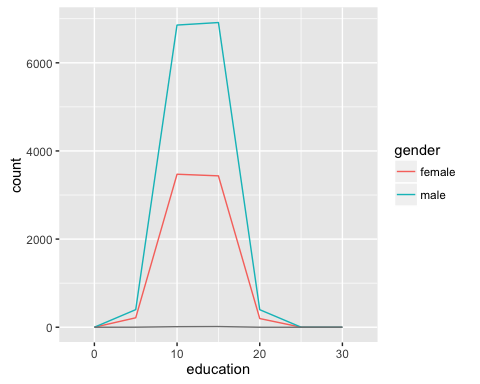
For example

There are outliers in these columns: 1. experience 2. age 3. height

### identify subsets of interest & suggest functional relationships

* Gender | Education

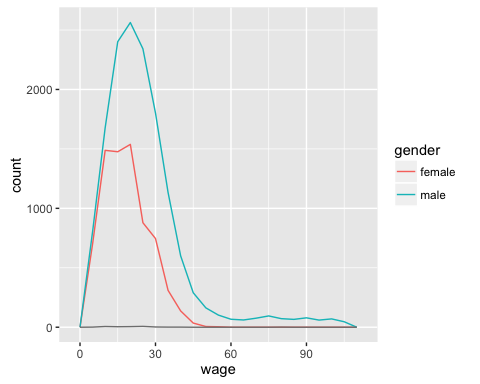
qplot(education , data = L3Q1\_data, color = gender, geom = "freqpoly", binwidth=5)



Compare gender and education, the plot shows that the number female is less than male at the same education level.

* Gender | Wage

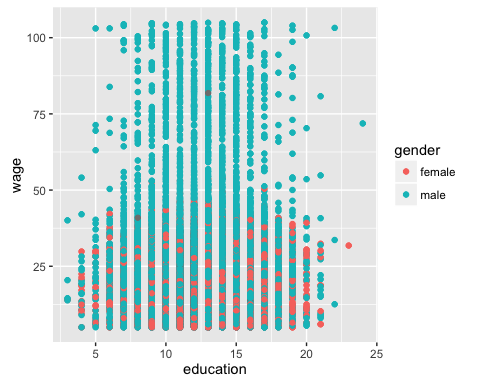
qplot(wage , data = L3Q1\_data, color = gender, geom = "freqpoly", binwidth=5)



Compare gender and wage, the plot shows that the number female is less than male at the same wage level.

* Education | Wage

qplot( education , wage, data = L3Q1\_data, color = gender) + geom\_point()



Compare education and wage, they have no relationship in this dataset, however, when I consider the gender, at the same education level, female earn less than male.