Module1L3

Neha Parulekar

January 25, 2016

## Homework Question 2

*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

To run the code in you may need additional packages.

If necessary install ggplot2 package.

`install.packages("ggplot2");

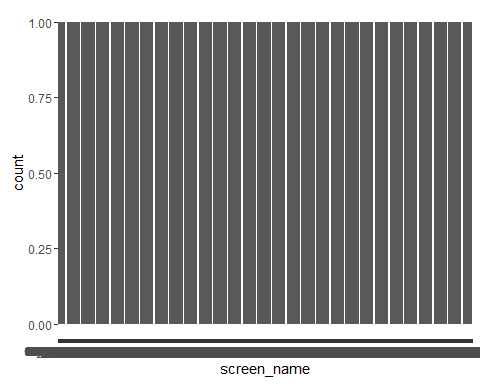
require(ggplot2)

## Loading required package: ggplot2

setwd("C:/Users/Neha/Desktop")  
twitterdataset <-read.csv("M01\_quasi\_twitter.csv")  
head(twitterdataset)

## screen\_name created\_at\_month created\_at\_day created\_at\_year country  
## 1 CNN 2 9 2007 USA  
## 2 osbrFe 11 21 2009 India  
## 3 WSJ 4 1 2007 India  
## 4 ninc 3 24 2007 USA  
## 5 nssubies 4 23 2009 USA  
## 6 BNCC 2 9 2009 England  
## location friends\_count followers\_count statuses\_count  
## 1 Miami Florida 1087 22187643 60246  
## 2 Mumbai 5210 6692814 93910  
## 3 Bangalore 1015 6257020 118465  
## 4 North Carolina 338 3433218 78082  
## 5 Nevada 641 2929559 93892  
## 6 Coventry 917 2540842 59397  
## favourites\_count favourited\_count dob\_day dob\_year dob\_month gender  
## 1 1122 105005 29 1999 4 female  
## 2 3825 40487 24 1991 10 female  
## 3 1143 87968 4 1997 3 male  
## 4 0 25943 22 1998 8 male  
## 5 226 32589 9 1963 11 female  
## 6 2122 19760 1 1995 1 female  
## mobile\_favourites\_count mobile\_favourited\_count education experience age  
## 1 0 0 8 0 29  
## 2 0 5032191 15 0 0  
## 3 0 0 9 0 32  
## 4 0 0 9 44 40  
## 5 0 0 13 24 45  
## 6 0 0 15 21 14  
## race wage retweeted\_count retweet\_count height  
## 1 white 16.31000 1 30 156  
## 2 white 17.91000 1 6 162  
## 3 white 15.71000 2 65 168  
## 4 white 7.00000 0 8 180  
## 5 white 17.87000 1 7 162  
## 6 white 14.10839 2 64 158

library(ggplot2)  
qplot(screen\_name, data=twitterdataset)



**Answers**

1. The data is distributed normally.
2. The data follows uniform distribution, thus no normal or skewed distribution.
3. The summary of the data set is

summary(twitterdataset$screen\_name)

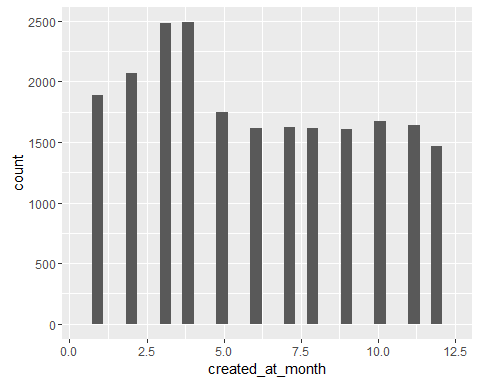
## +5400E1. 000D0se7 001apdov 001RBTePh 003B0K2 007unfasa 00kubisn3   
## 1 1 1 1 1 1 1   
## 00Leo6v 00ODN8 00r3rm1 00situfur 018mekyem 01cho8ga0 01E1+.715   
## 1 1 1 1 1 1 1   
## 01m9ya1 01nlp2s1a 01uvdt 01Uz2Ii1D 020yamv1i 031a2bc 0332321m5   
## 1 1 1 1 1 1 1   
## 03aBoybcM 03esn5a0c 04a9u7nsi 04na4tlar 05339ishC 05DJ2 05L42174G   
## 1 1 1 1 1 1 1   
## 066sv1yua 06ldav0 07ih0jhcr 08AAJoovg 08tq7auSJ 0970inoer 097aryi5z   
## 1 1 1 1 1 1 1   
## 0a36apsr 0ac7iuF 0ag1pmna 0ANYIEMM 0auscA07r 0awyx0 0bd3sa3Me   
## 1 1 1 1 1 1 1   
## 0be2a1yse 0Bo20aniN 0cac8remj 0cikeg7 0d44rye 0dauhri6a 0deagllie   
## 1 1 1 1 1 1 1   
## 0DUGN0I1A 0e94l782h 0ea7nW08y 0eerset2e 0emgvaa50 0en2l4eht 0end3v6wi   
## 1 1 1 1 1 1 1   
## 0epd3edlr 0estseu0n 0eTr9s7 0Flng0mCD 0g1z9gBi2 0gJoKiiai 0gocsilpa   
## 1 1 1 1 1 1 1   
## 0h6y8ipls 0hm12yloi 0i2ne50lj 0iannMtl1 0II3GOGOP 0in5qun1 0ipptl9ou   
## 1 1 1 1 1 1 1   
## 0j206l32 0j2c0to0n 0JD30S14 0jneel0 0k90ts 0kuke0akr 0L0yoo2Tr   
## 1 1 1 1 1 1 1   
## 0l1k03k2 0lbva0 0ld4d158 0leeak1jb 0m0eoLohy 0m0kneo1o 0m63gim   
## 1 1 1 1 1 1 1   
## 0Mf6on770 0min9ef3r 0mryodag7 0n1ow1a2a 0n364a82t 0n6d0eyd8 0neLoXX   
## 1 1 1 1 1 1 1   
## 0nesn2ivg 0nh11i 0njos3k30 0NllCBT 0NNrwKest 0o26arc 0ojJl2cVa   
## 1 1 1 1 1 1 1   
## 0olf5Aw (Other)   
## 1 21817

1. There are no outliers.
2. There are no transforms, quality problems, outliers, subset intrest.

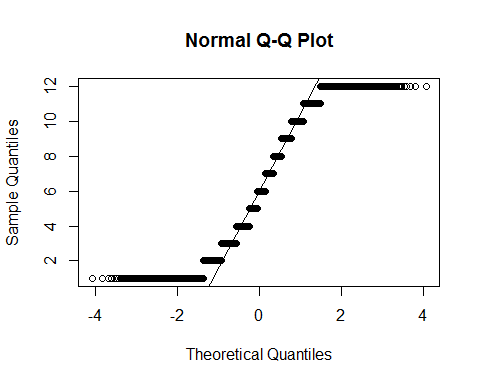
**created\_at\_month**

qplot(created\_at\_month, data=twitterdataset)

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



qqnorm(twitterdataset$created\_at\_month)  
qqline(twitterdataset$created\_at\_month)



1. skewed distribution to the left
2. this is not a normal distribution
3. summary

summary(twitterdataset$created\_at\_month)

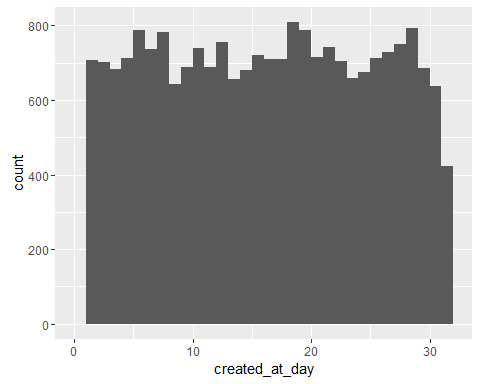
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 6.000 6.069 9.000 12.000

1. No
2. No

**created\_at\_day**

ggplot(twitterdataset)+aes(x=created\_at\_day)+geom\_histogram()

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



1. uniform distribution
2. this is not a normal distribution
3. summary

summary(twitterdataset$created\_at\_day)

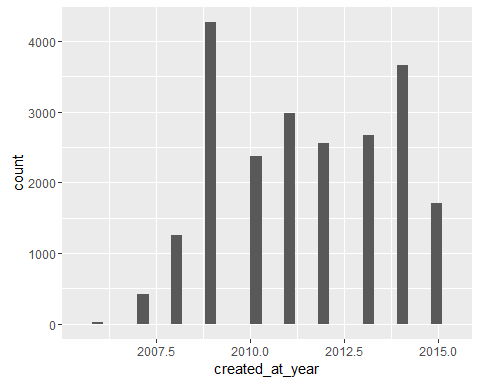
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 8.00 16.00 15.78 23.00 31.00

1. drop at no 34
2. no

**created\_at\_year**

ggplot(twitterdataset)+aes(x=created\_at\_year)+geom\_histogram()

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



1.uniform distribution with peaks in 2009 and 2014. 2. Not a normal distribution. 3. summary

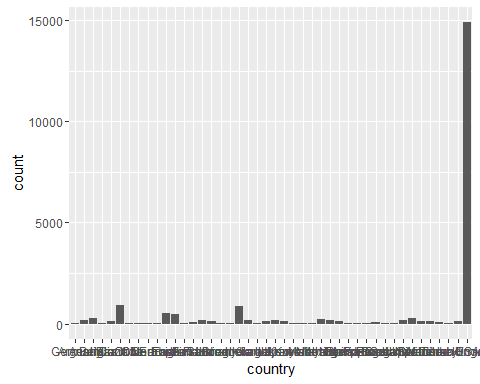
summary(twitterdataset$created\_at\_year)

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

4.outlier at 2006 5. No

**country**

qplot(country, data=twitterdataset)



1.uniform distribution with peak at USA 2. not a normal distribution 3. summary

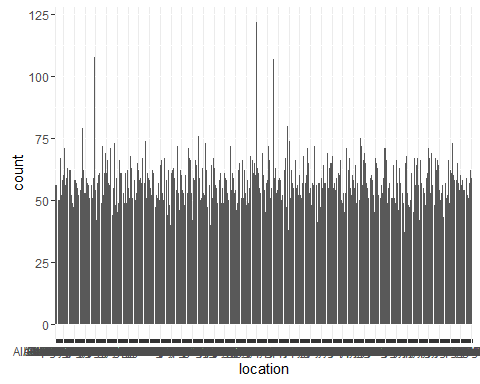
summary(twitterdataset$country)

## 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

1. no
2. no

**location**

qplot(location, data=twitterdataset)



1. its a uniform distribution
2. not a normal distribution
3. sumary

summary(twitterdataset$location)

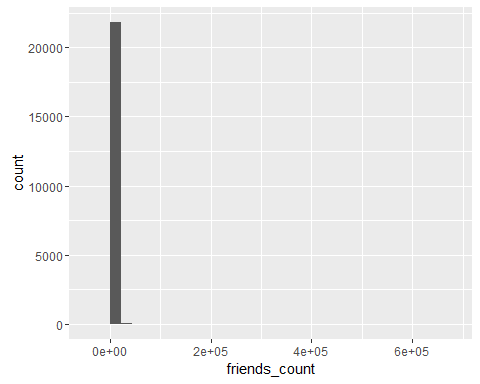
## Mexico Boston Montreal   
## 122 108 107   
## Nevada Bangalore Indianapolis Indiana   
## 80 79 76   
## Pune India Dallas Texas New Hampshire   
## 75 74 74   
## Cambridge MA Istanbul Vancouver BC   
## 73 73 73   
## Brooklyn New York Fremont CA London United Kingdom   
## 72 72 72   
## Minnesota 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

1. No
2. No

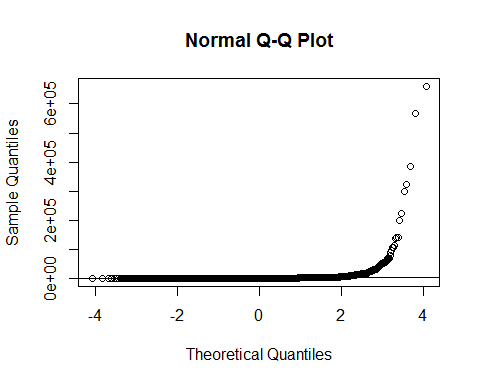
**friends\_count**

qplot(friends\_count, data=twitterdataset)

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



qqnorm(twitterdataset$friends\_count)  
qqline(twitterdataset$friends\_count)



1. skewed distribution to the left
2. not a normal distribution
3. summary

summary(twitterdataset$friends\_count)

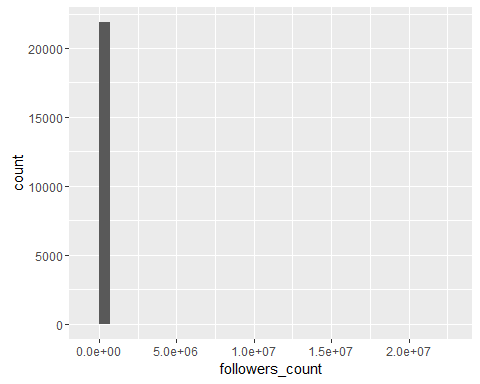
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -84 123 324 1058 849 660500

1. No
2. No

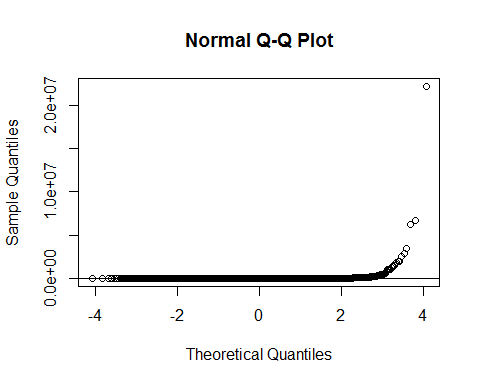
**followers\_count**

qplot(followers\_count, data=twitterdataset)

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



qqnorm(twitterdataset$followers\_count)  
qqline(twitterdataset$followers\_count)

 1. skewed distribution 2. not a normal distribution 3. summary

summary(twitterdataset$followers\_count)

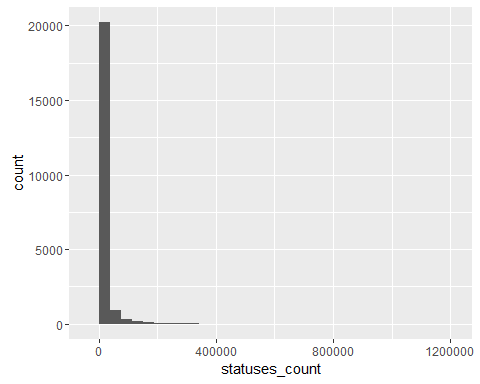
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 105 336 5859 1075 22190000

1. outliers beyond 5million
2. No

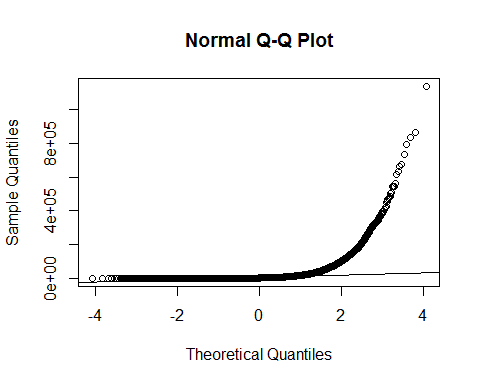
**statuses\_count**

qplot(statuses\_count, data=twitterdataset)

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



qqnorm(twitterdataset$statuses\_count)  
qqline(twitterdataset$statuses\_count)

 1. skewed distribution 2. not a normal distribution 3. summary

summary(twitterdataset$followers\_count)

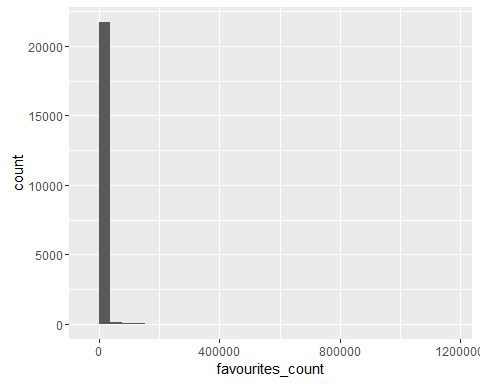
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 105 336 5859 1075 22190000

1. outliers present
2. No

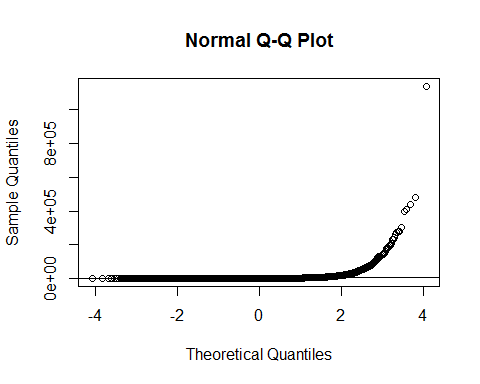
**favourites\_count**

qplot(favourites\_count, data=twitterdataset)

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



qqnorm(twitterdataset$favourites\_count)  
qqline(twitterdataset$favourites\_count)



1. skewed distribution
2. Not a normal distribution
3. summary

summary(twitterdataset$favourites\_count)

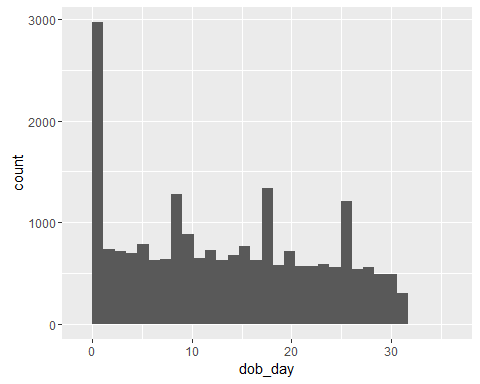
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 16 164 2217 950 1140000

1. outliers present
2. No

**dob\_day**

ggplot(twitterdataset)+aes(x=dob\_day)+geom\_histogram()

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

 1. uniformly distributed 2. not a normal distribution 3. summary

summary(twitterdataset$dob\_day)

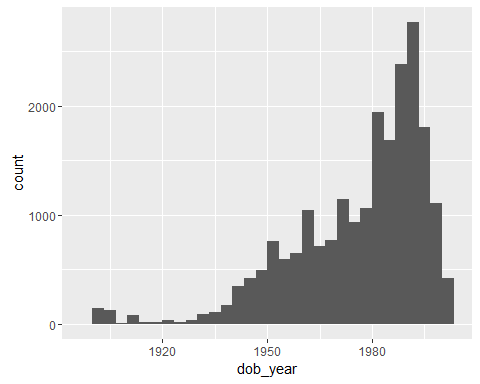
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 5.00 13.00 13.49 21.00 35.00

1. outliers at 31
2. No

**dob\_year**

ggplot(twitterdataset)+aes(x=dob\_year)+geom\_histogram()

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

 1. uniform distribution with a peak at 1990 2. No 3. summary

summary(twitterdataset$dob\_day)

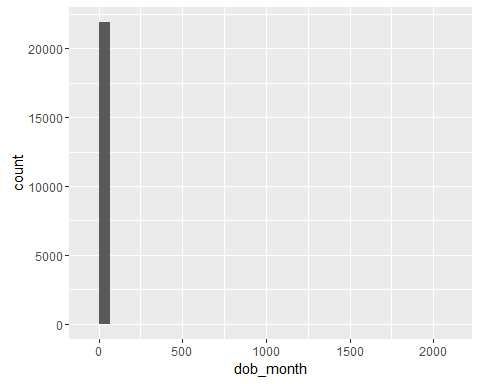
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 5.00 13.00 13.49 21.00 35.00

1. no 5.no

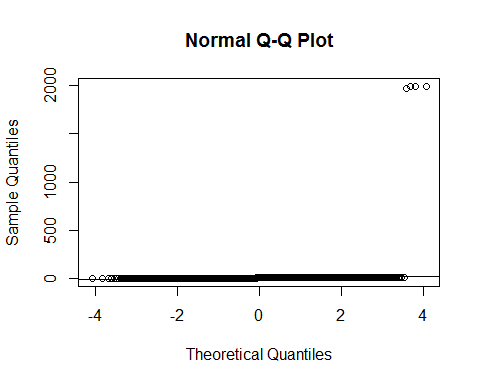
**dob\_month**

qplot(dob\_month, data=twitterdataset)

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



qqnorm(twitterdataset$dob\_month)  
qqline(twitterdataset$created\_at\_month)

 1. uniform distribution 2. no 3. summary

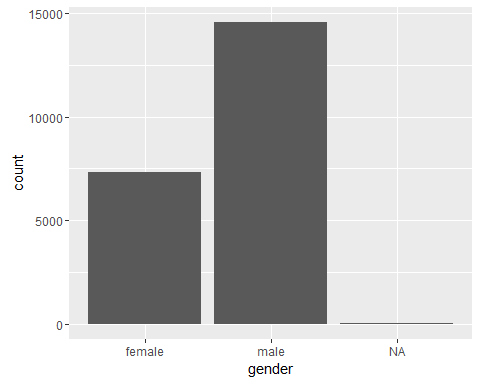
summary(twitterdataset$dob\_month)

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

4 outliers beyond 1500 5. no

**gender**

qplot(gender, data=twitterdataset)

 1. uniform distribution 2. no 3. summary

summary(twitterdataset$gender)

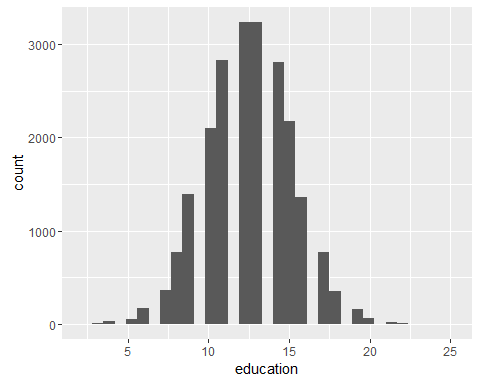
## female male NA's   
## 7319 14569 28

1. outliers at 0
2. No

**education**

qplot(education, data=twitterdataset)

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

 1. No 2. Normal distribution 3. summary

summary(twitterdataset$education)

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

1. no 5.no

**experience**

qplot(experience, data=twitterdataset)

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

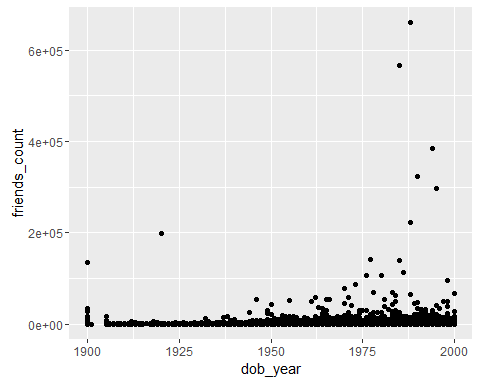
 1. uniform distribution 2. No 3. summary

summary(twitterdataset$experience)

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

1. outliers present
2. no
3. There can be a functional relationship between the age and the friends count or age and followers count

qplot(dob\_year, friends\_count, data=twitterdataset)



qplot(dob\_year, followers\_count, data=twitterdataset)

