M1L3 Homework Assignment

Joshua Conte

September 17, 2017

1 M1L3 Homework Assignment

R studio was configured with the following parameters before beginning the project:

```
# clears the console in RStudio
cat("\014")
```

```
# clears environment
rm(list = ls())

# Set working directory
setwd("C:/R/DA5030/module_01")

# Load required packages
require(ggplot2)
require(fitdistrplus)
library(knitr) # Used for making tables in this report
```

1.1 load the file M01_Lesson_02_Q1.csv.

I opened the data using read.csv2 and changed the data from integer to numeric so I could analyze it with R.

Below is my R code:

```
# Some VCF files are really big and take a while to open.  This command checks to
# see if it is already opened, if it is, it does not open it again.
# I also omitted the first column
if (!exists("df1")) {
  df1 <-
    read.csv2(
      'M01_quasi_twitter.csv ',
      sep = ",",
      stringsAsFactors = FALSE,
      row.names = NULL,
      header = TRUE
    )
}

# Check to make sure the data is in numeric form for analysis:
sapply(df1, class)
```

```
##             screen_name        created_at_month          created_at_day
##             "character"               "integer"               "integer"
##         created_at_year                 country                location
##               "integer"             "character"             "character"
##           friends_count         followers_count          statuses_count
##               "integer"               "integer"               "integer"
##        favourites_count        favourited_count                 dob_day
##               "integer"               "integer"               "integer"
##                dob_year               dob_month                  gender
##               "integer"               "integer"             "character"
## mobile_favourites_count mobile_favourited_count               education
##               "integer"               "integer"               "integer"
##              experience                     age                    race
##               "integer"               "integer"             "character"
```

```
##                   wage         retweeted_count           retweet_count
##            "character"               "integer"               "integer"
##                 height
##              "integer"
```

```
# change int to numeric
df1[2:4] <- sapply(df1[2:4], as.numeric)
df1[7:14] <- sapply(df1[7:14], as.numeric)
df1[16:20] <- sapply(df1[16:20], as.numeric)
df1[22:25] <- sapply(df1[22:25], as.numeric)

# Confirm changes
sapply(df1, class)
```

```
##             screen_name        created_at_month          created_at_day
##             "character"                "numeric"               "numeric"
##         created_at_year                  country                location
##                "numeric"              "character"             "character"
##           friends_count         followers_count          statuses_count
##                "numeric"                "numeric"               "numeric"
##        favourites_count        favourited_count                 dob_day
##                "numeric"                "numeric"               "numeric"
##                dob_year               dob_month                  gender
##                "numeric"                "numeric"             "character"
## mobile_favourites_count mobile_favourited_count               education
##                "numeric"                "numeric"               "numeric"
##              experience                     age                    race
##                "numeric"                "numeric"             "character"
##                    wage         retweeted_count           retweet_count
##                "numeric"                "numeric"               "numeric"
##                  height
##                "numeric"
```

1.2 How is the data distributed?

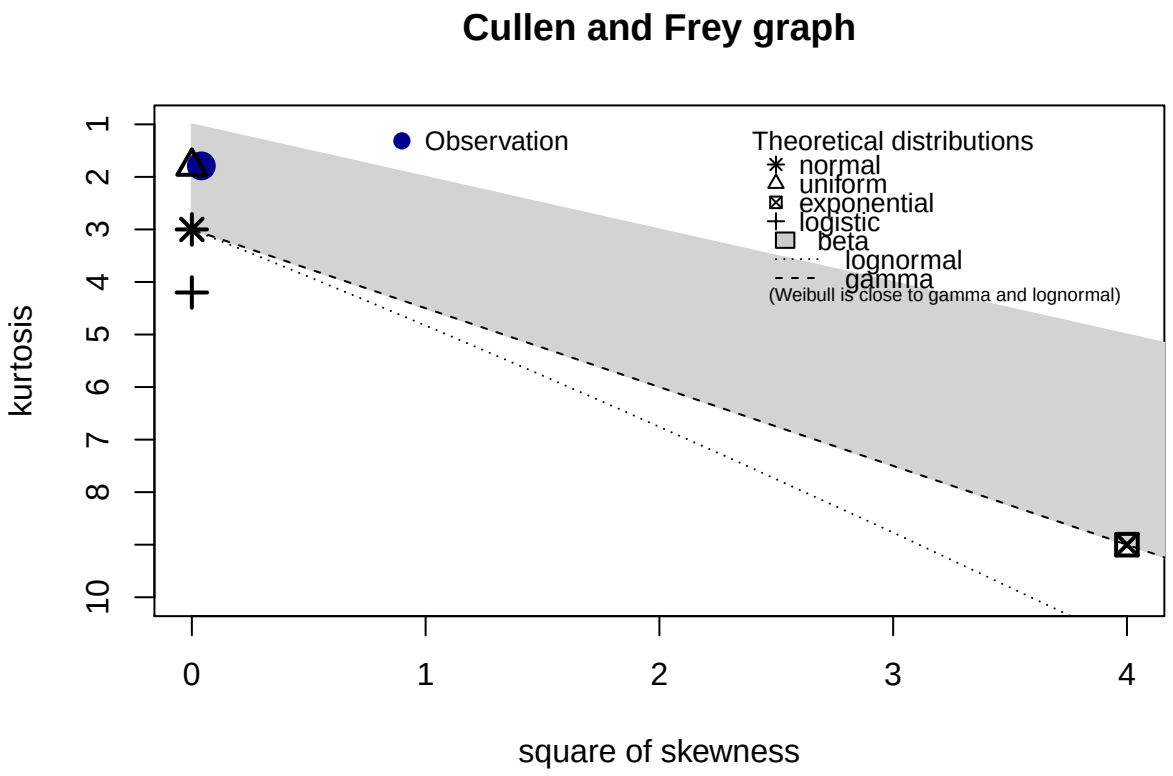
I ran fitdistrplus on all numeric data and printed out Cullen and Frey graphs to get an idea of how the data is distributed.

Most of the data appears to be not normal and un-uniform. created_at_month, created_at_day, created_at_year, dob_day appear to be uniform and education is normal. Everything else appears to be bad and will require additional analysis.

However, if some of the outliers were removed, the data could be normal or uniform. For example, age, there are a lot of people that are listed as 0, this is impossible, if this anomaly was removed, the data would be normal.

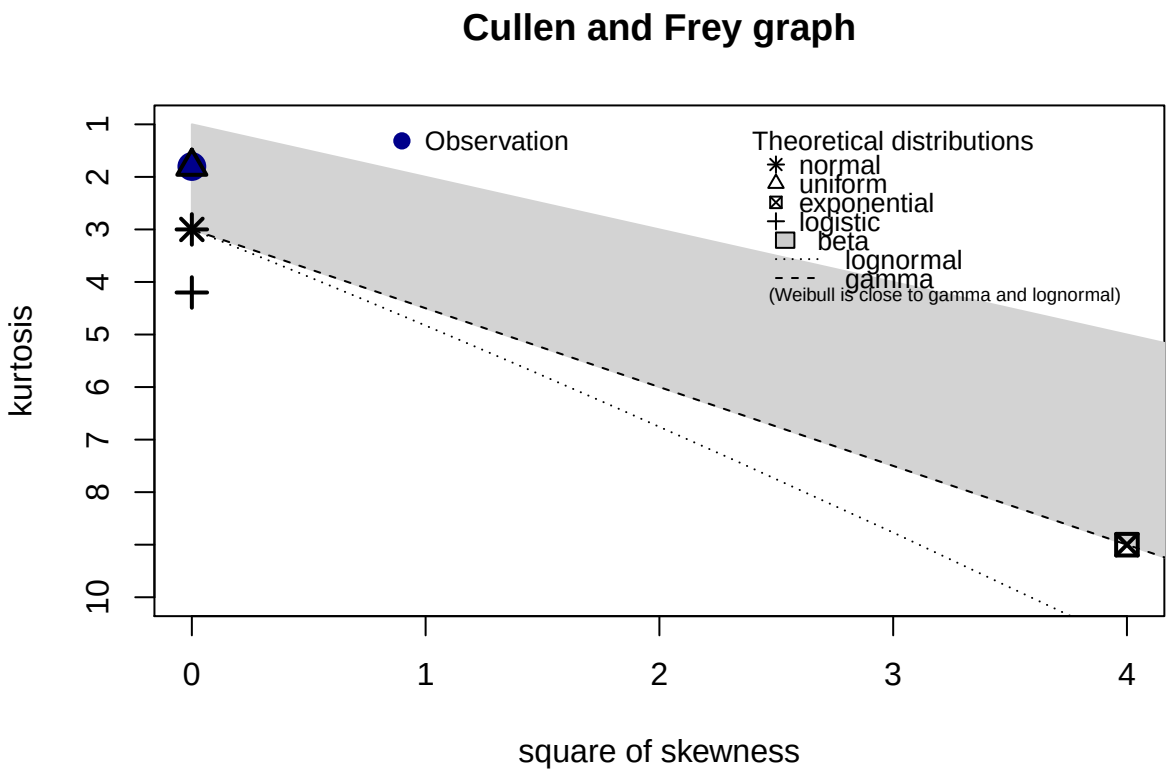
Please refer to the plots below:

```
# created_at_month
descdist(df1$created_at_month, discrete = FALSE)
```



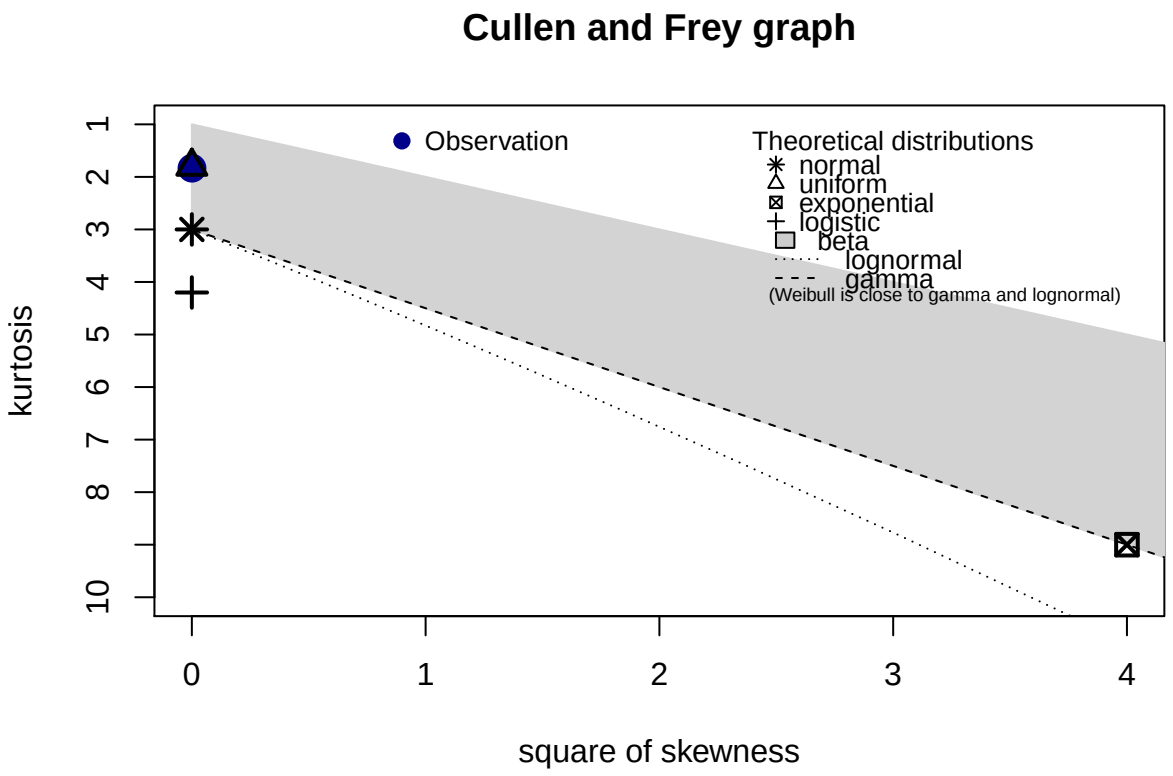

```
## summary statistics
## ------
## min:  1   max:  12
## median:  6
## mean:  6.068626
## estimated sd:  3.426386
## estimated skewness:  0.2015503
## estimated kurtosis:  1.790103
```

```r
# created_at_day
descdist(df1$created_at_day, discrete = FALSE)
```



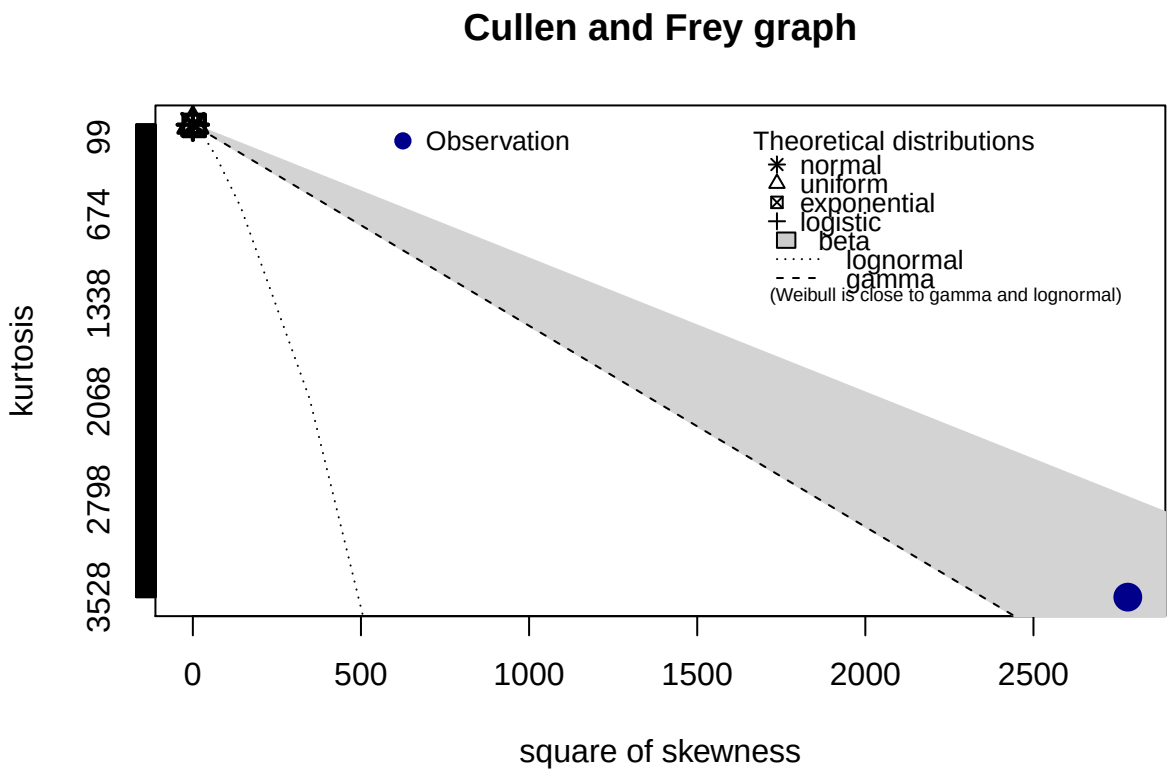

```
## summary statistics
## ------
## min:  1   max:  31
## median:  16
## mean:  15.78386
## estimated sd:  8.803773
## estimated skewness:  0.0003880161
## estimated kurtosis:  1.802239
```

```
# created_at_year
descdist(df1$created_at_year, discrete = FALSE)
```



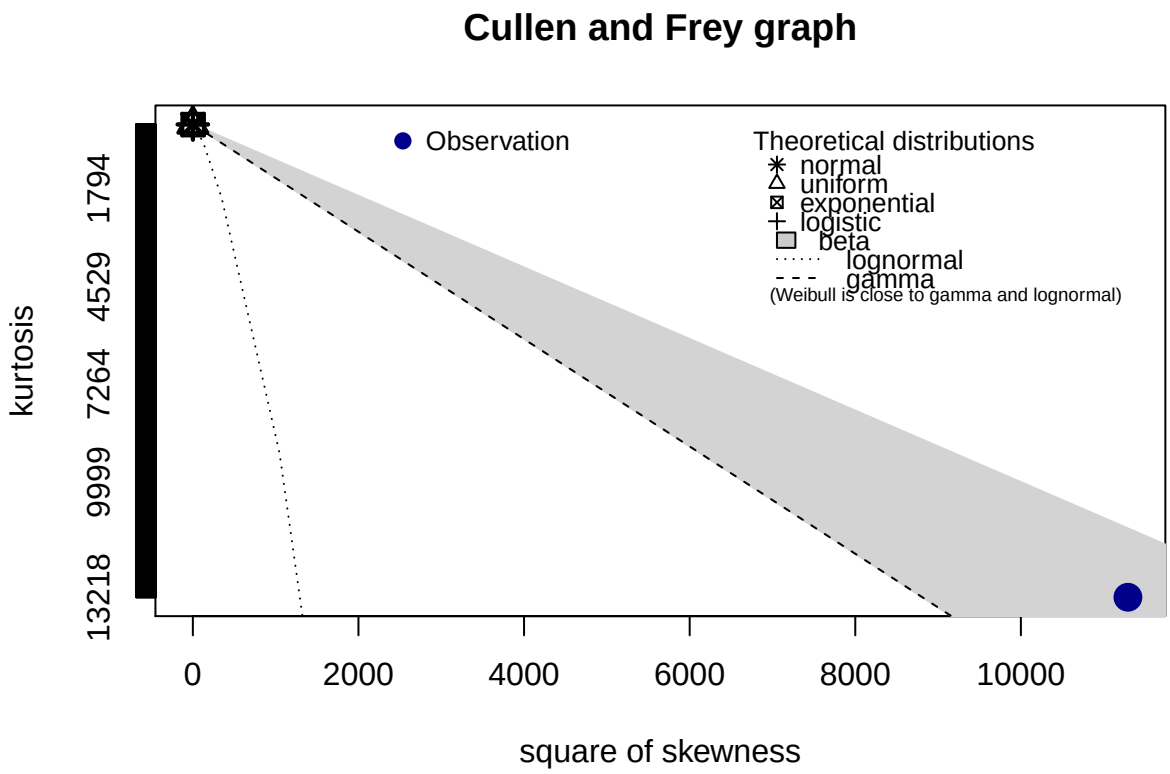

```
## summary statistics
## ------
## min:  2006   max:  2015
## median:  2011
## mean:  2011.42
## estimated sd:  2.217927
## estimated skewness:  -0.02564643
## estimated kurtosis:  1.834411
```

```
# friends_count
descdist(df1$friends_count, discrete = FALSE)
```



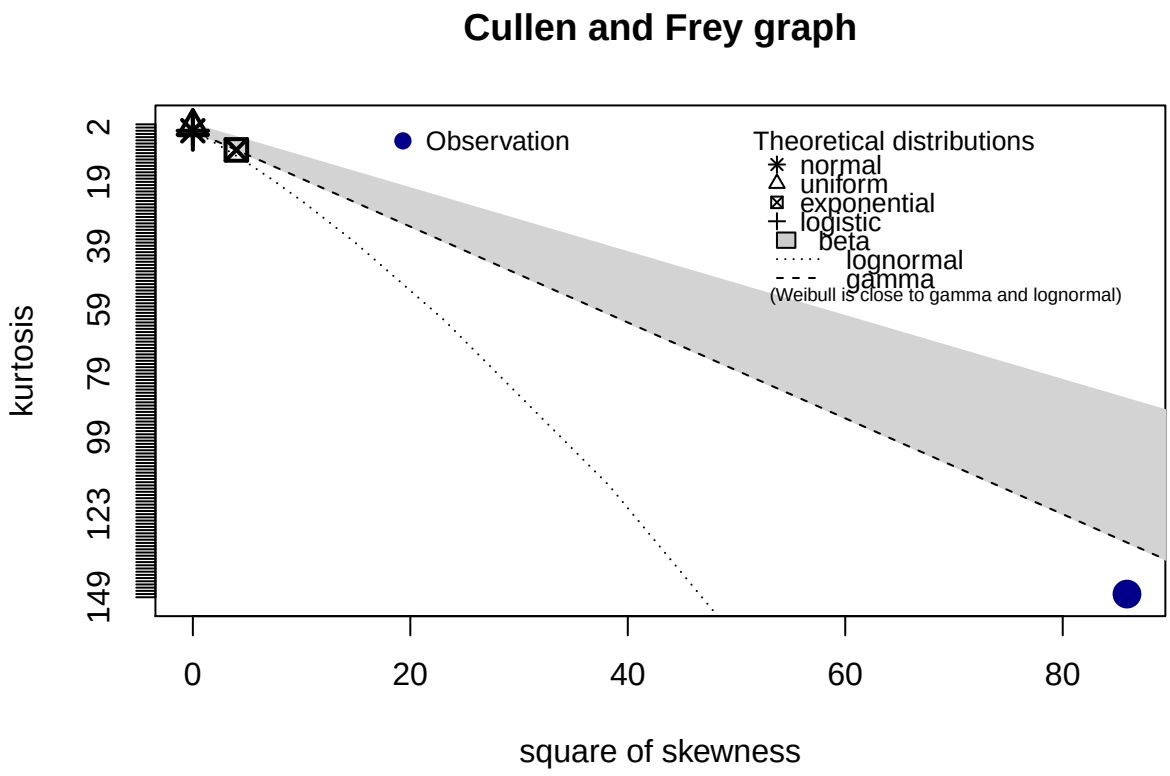

```
## summary statistics
## ------
## min:  -84   max:  660549
## median:  324
## mean:  1057.911
## estimated sd:  8125.054
## estimated skewness:  52.73123
## estimated kurtosis:  3527.241
```

```
# followers_count
descdist(df1$followers_count, discrete = FALSE)
```



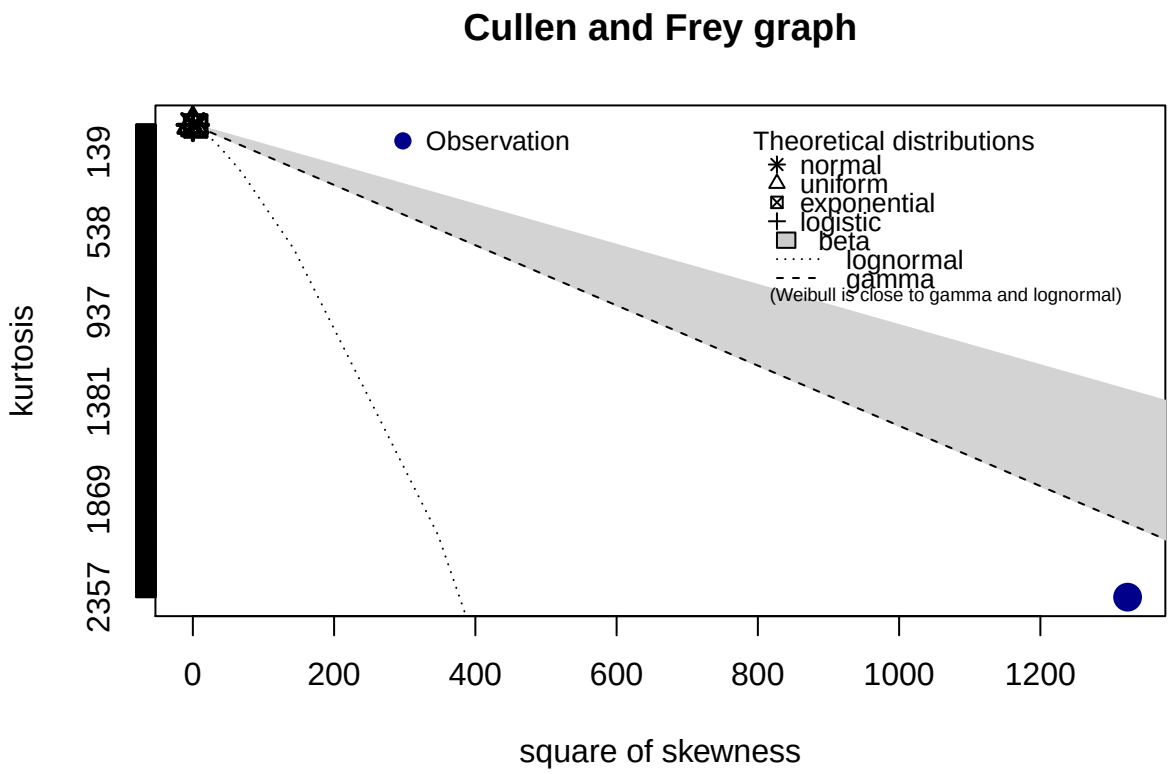

```
## summary statistics
## ------
## min:  0   max:  22187643
## median:  336
## mean:  5859.258
## estimated sd:  170711.3
## estimated skewness:  106.2666
## estimated kurtosis:  13217.98
```

```
# statuses_count
descdist(df1$statuses_count, discrete = FALSE)
```



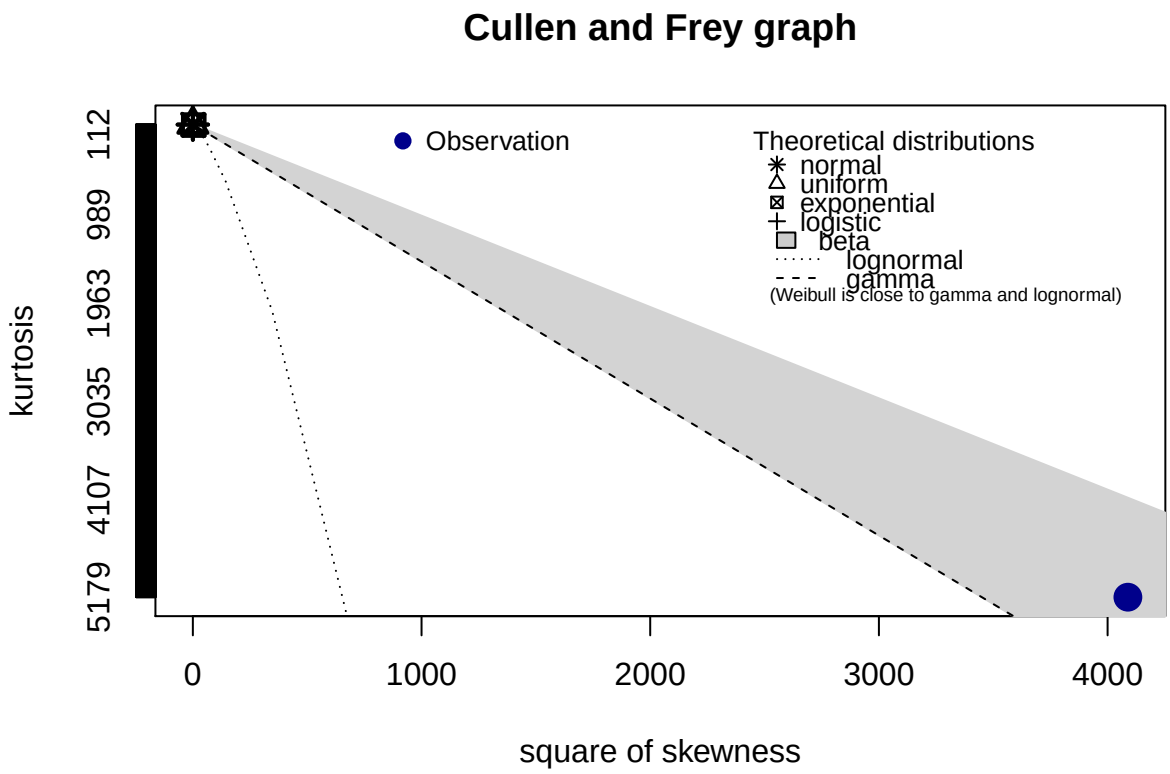

```
## summary statistics
## ------
## min:  1   max:  1136198
## median:  2341
## mean:  12485.88
## estimated sd:  36178.72
## estimated skewness:  9.268858
## estimated kurtosis:  148.0129
```

```
# favourites_count
descdist(df1$favourites_count, discrete = FALSE)
```



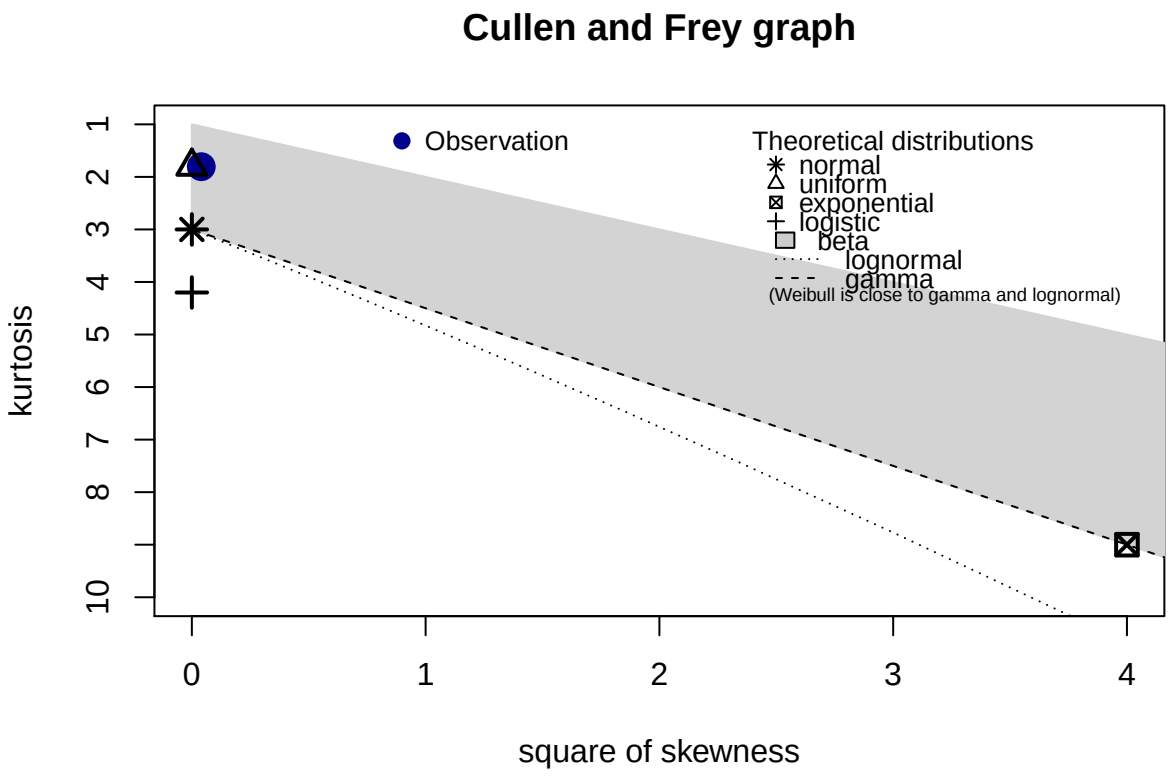

```
## summary statistics
## ------
## min:  0   max:  1140139
## median:  164
## mean:  2216.749
## estimated sd:  13811.44
## estimated skewness:  36.38002
## estimated kurtosis:  2356.703
```

```
# favourited_count
descdist(df1$favourited_count, discrete = FALSE)
```



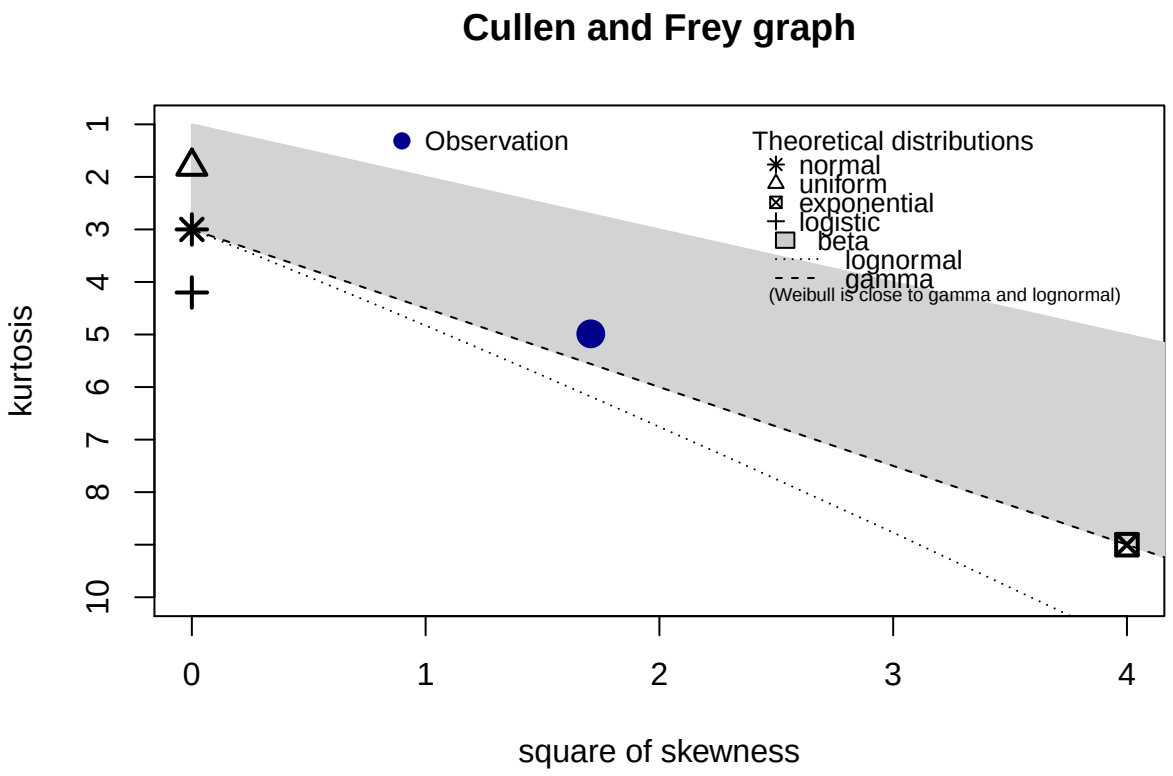

```
## summary statistics
## ------
## min:  0   max:  105005
## median:  9
## mean:  92.23937
## estimated sd:  1132.389
## estimated skewness:  63.94325
## estimated kurtosis:  5178.486
```

```
# dob_day
descdist(df1$dob_day, discrete = FALSE)
```



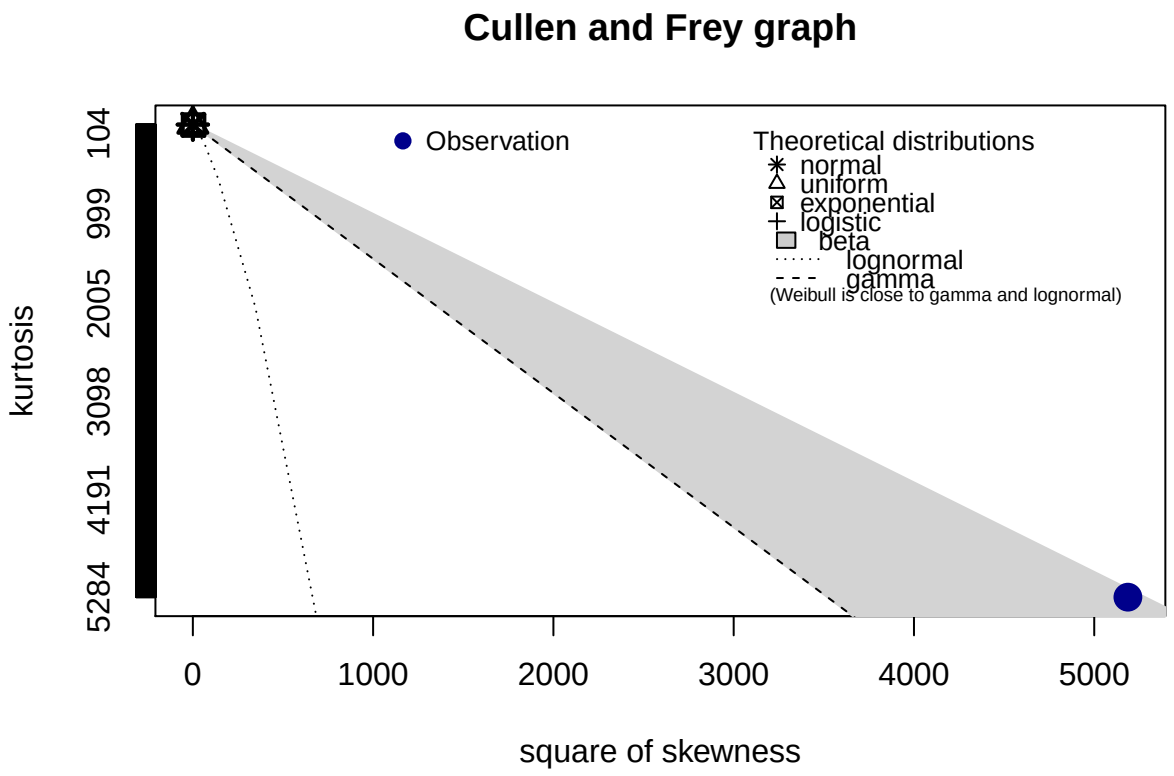

```
## summary statistics
## ------
## min:  1   max:  35
## median:  13
## mean:  13.49024
## estimated sd:  9.231642
## estimated skewness:  0.2007477
## estimated kurtosis:  1.805992
```

```
# dob_year
descdist(df1$dob_year, discrete = FALSE)
```



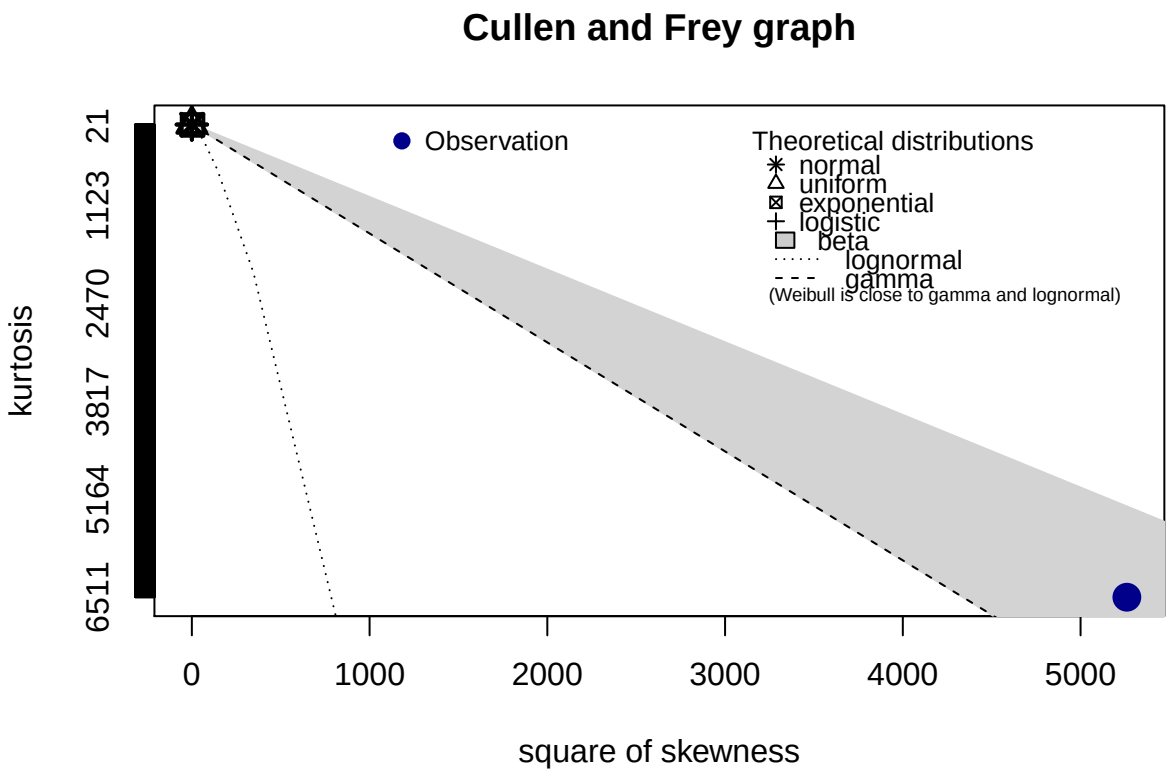

```
## summary statistics
## ------
## min:  1900   max:  2000
## median:  1982
## mean:  1976.145
## estimated sd:  19.05346
## estimated skewness:  -1.306317
## estimated kurtosis:  4.987041
```

```
# dob_month
descdist(df1$dob_month, discrete = FALSE)
```



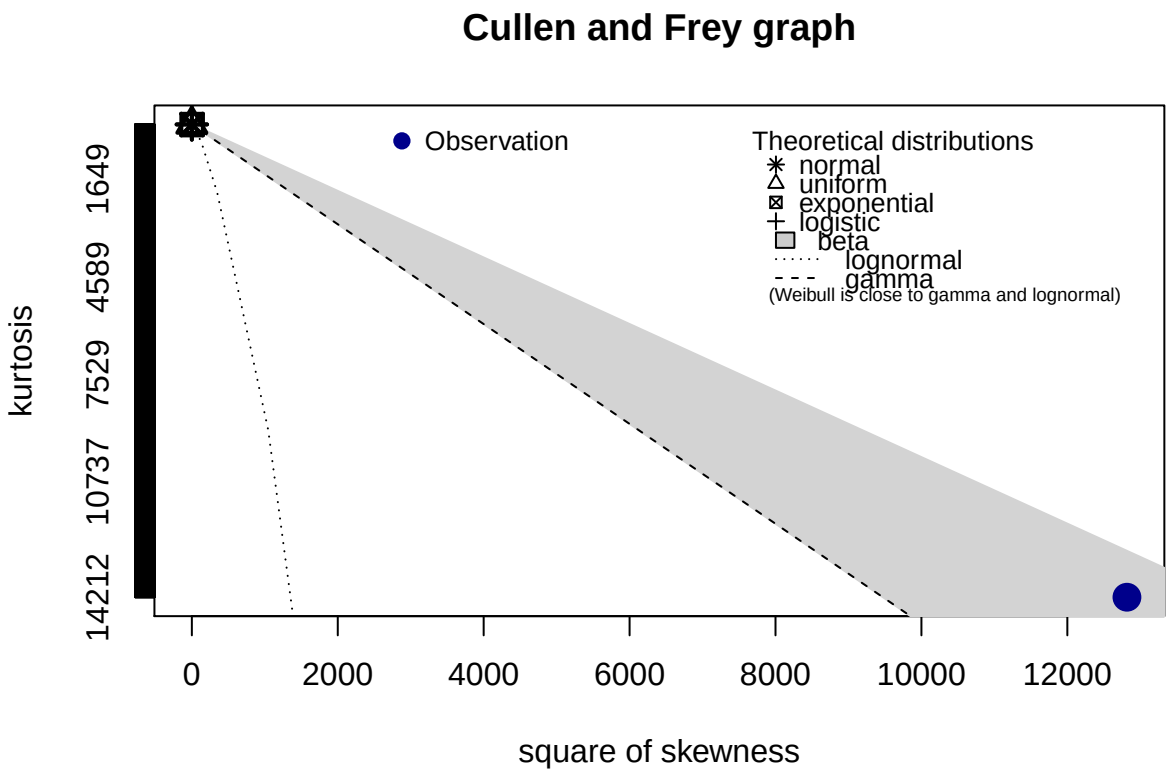

```
## summary statistics
## ------
## min:  1   max:  1992
## median:  6
## mean:  6.397655
## estimated sd:  26.99822
## estimated skewness:  72.01827
## estimated kurtosis:  5283.324
```

```
# mobile_favourites_count
descdist(df1$mobile_favourites_count, discrete = FALSE)
```



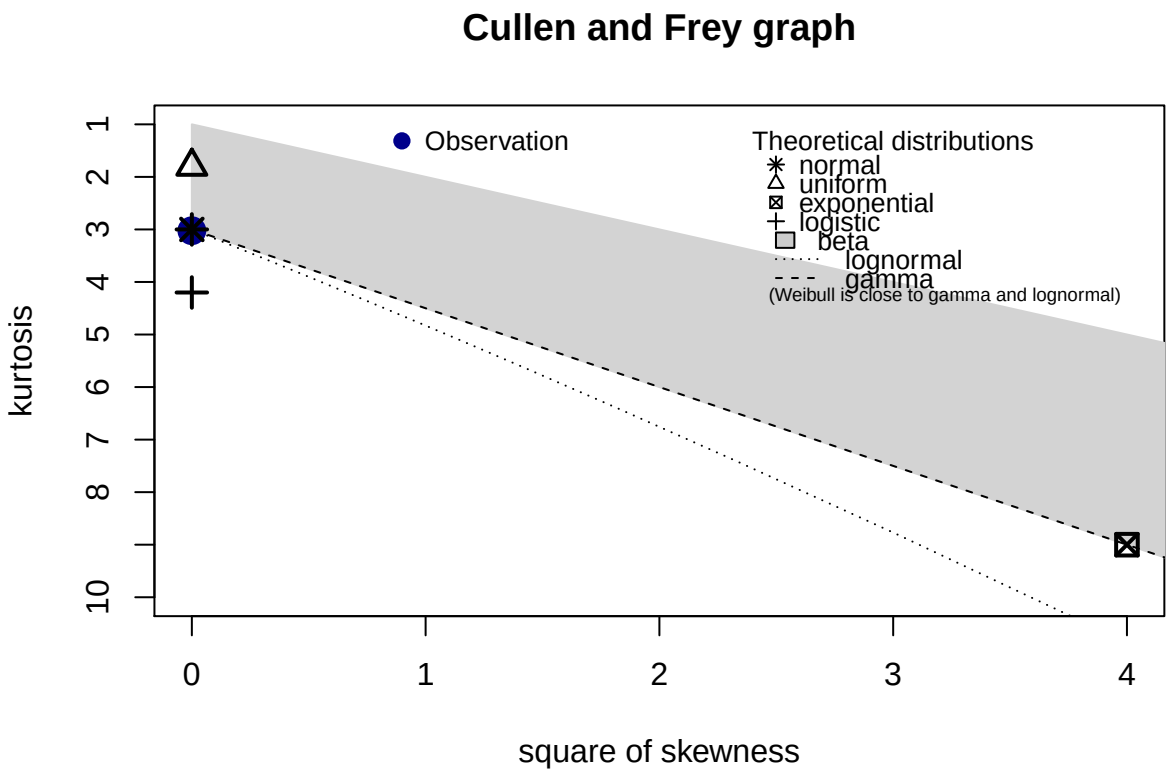

```
## summary statistics
## ------
## min:  0   max:  377123
## median:  0
## mean:  152.8648
## estimated sd:  3629.494
## estimated skewness:  72.52879
## estimated kurtosis:  6510.898
```

```
# mobile_favourited_count
descdist(df1$mobile_favourited_count, discrete = FALSE)
```



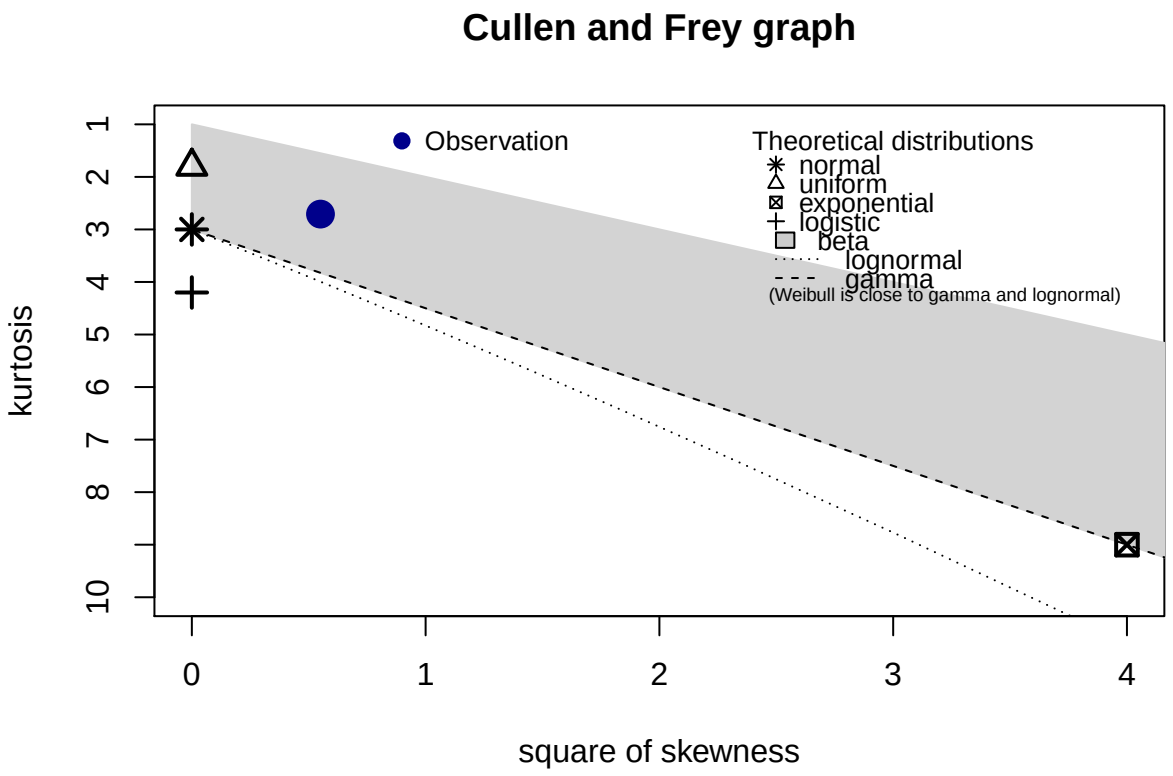

```
## summary statistics
## ------
## min:  0   max:  5032191
## median:  0
## mean:  648.8135
## estimated sd:  38135.54
## estimated skewness:  113.2116
## estimated kurtosis:  14211.25
```

```
# education
descdist(df1$education, discrete = FALSE)
```



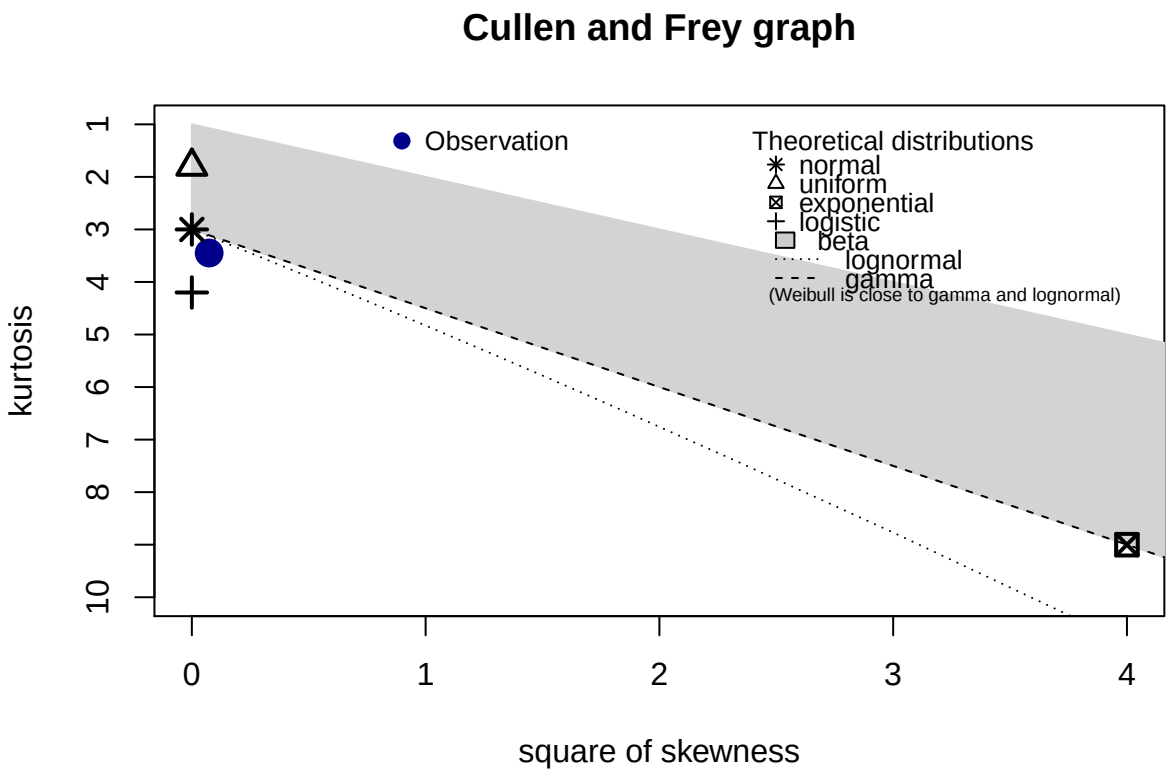

```
## summary statistics
## ------
## min:  3   max:  24
## median:  13
## mean:  12.49786
## estimated sd:  2.638964
## estimated skewness:  -0.008735253
## estimated kurtosis:  3.024296
```

```
# experience
descdist(df1$experience, discrete = FALSE)
```



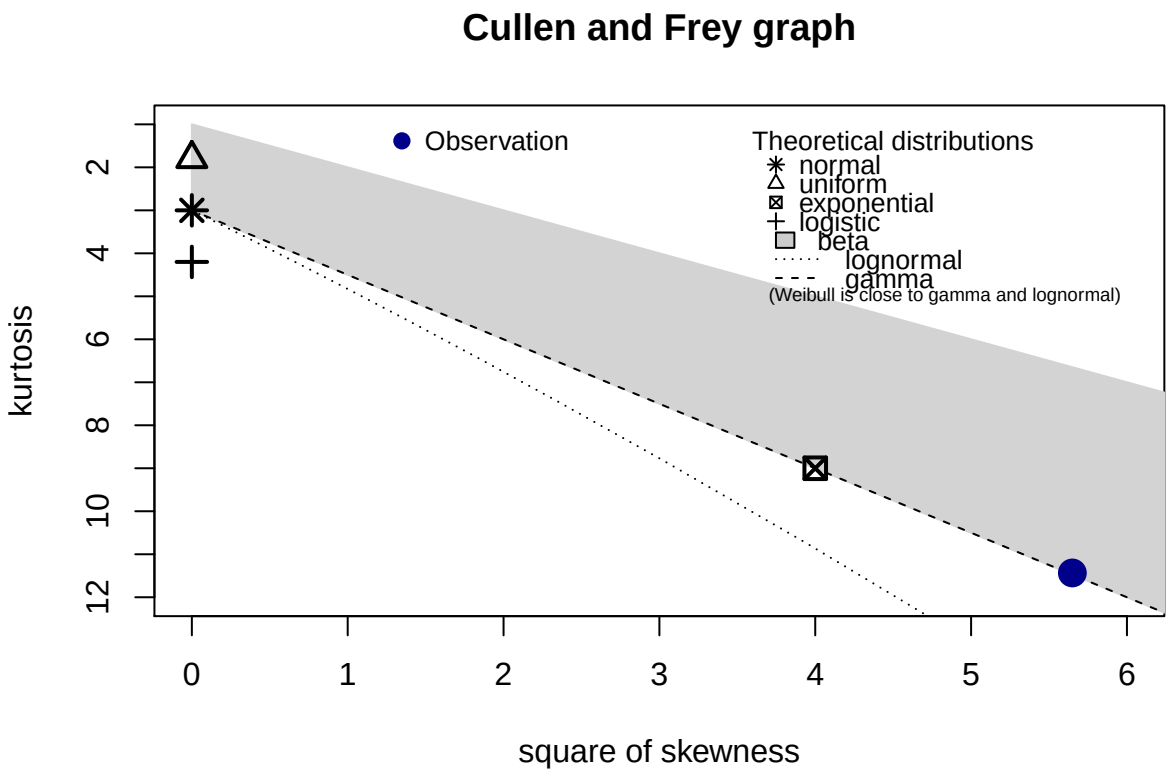

```
## summary statistics
## ------
## min:  -32   max:  74
## median:  7
## mean:  10.8821
## estimated sd:  12.83229
## estimated skewness:  0.7418106
## estimated kurtosis:  2.708021
```

```
# age
descdist(df1$age, discrete = FALSE)
```



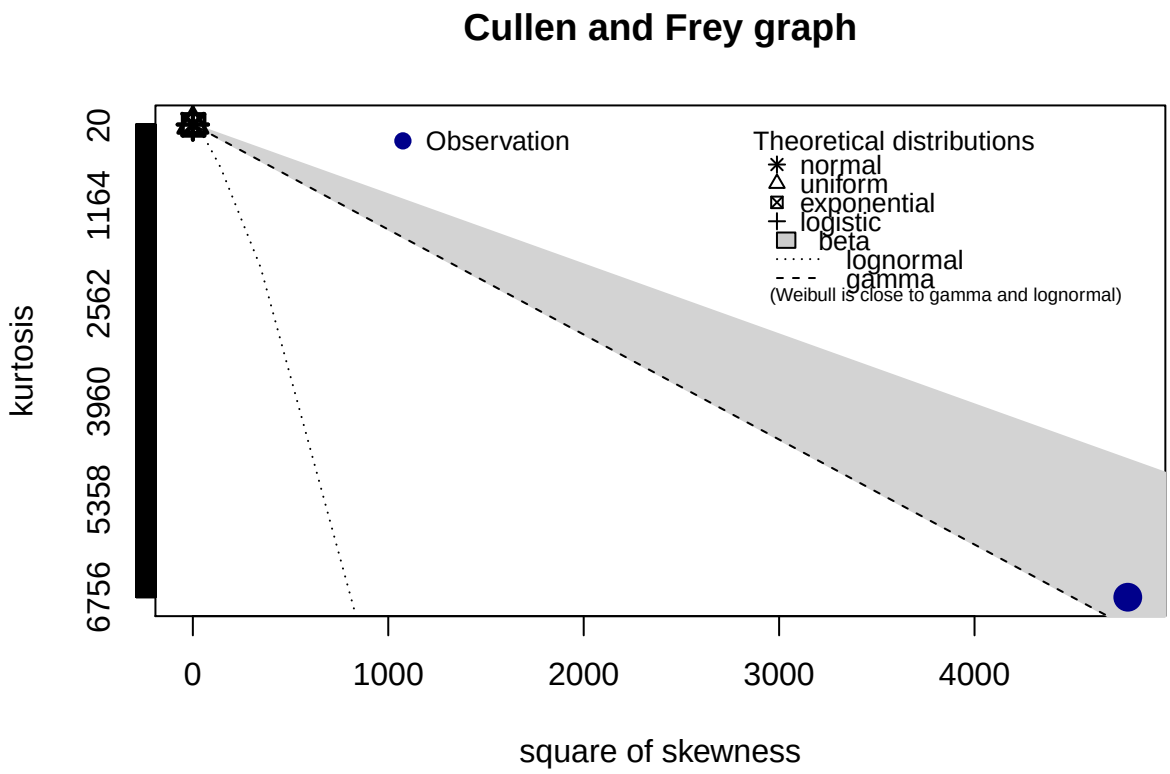

```
## summary statistics
## ------
## min:  -6   max:  91
## median:  36
## mean:  35.53874
## estimated sd:  12.76811
## estimated skewness:  -0.2720907
## estimated kurtosis:  3.448563
```

```
# wage
descdist(df1$wage, discrete = FALSE)
```



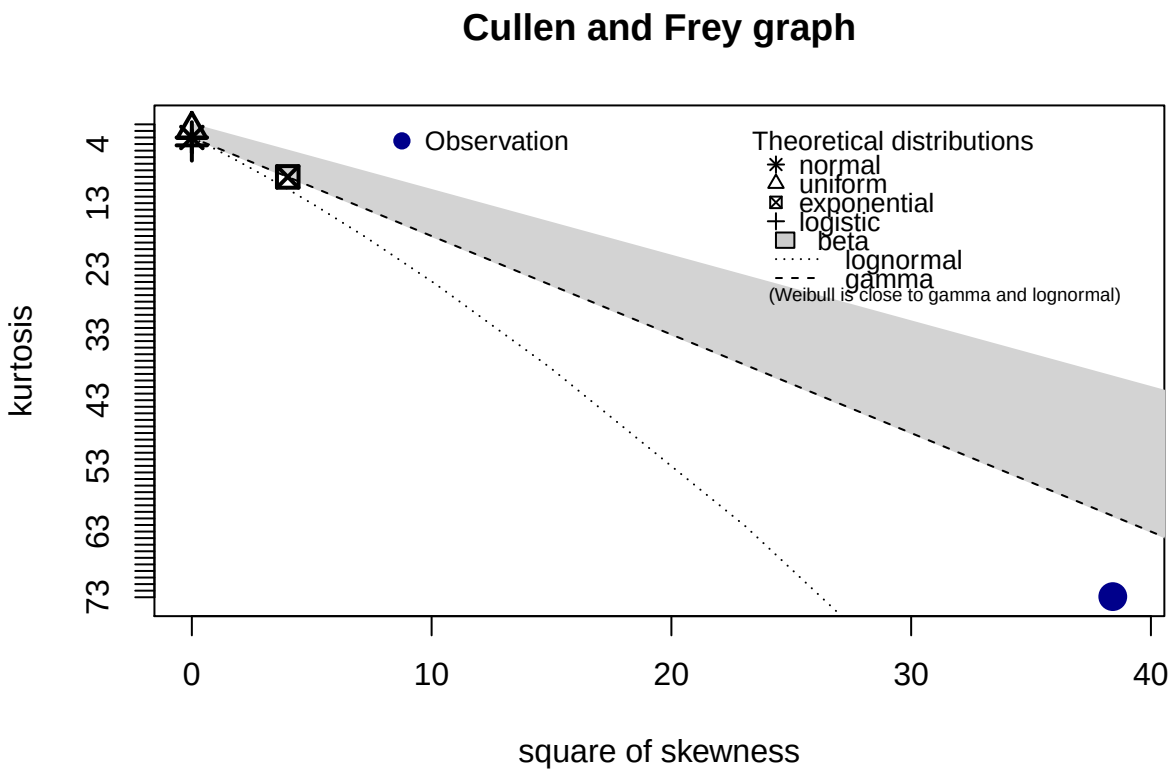

```
## summary statistics
## ------
## min:  5   max:  104.9703
## median:  20.36
## mean:  22.97297
## estimated sd:  14.60475
## estimated skewness:  2.376917
## estimated kurtosis:  11.4351
```

```
# retweeted_count
descdist(df1$retweeted_count, discrete = FALSE)
```



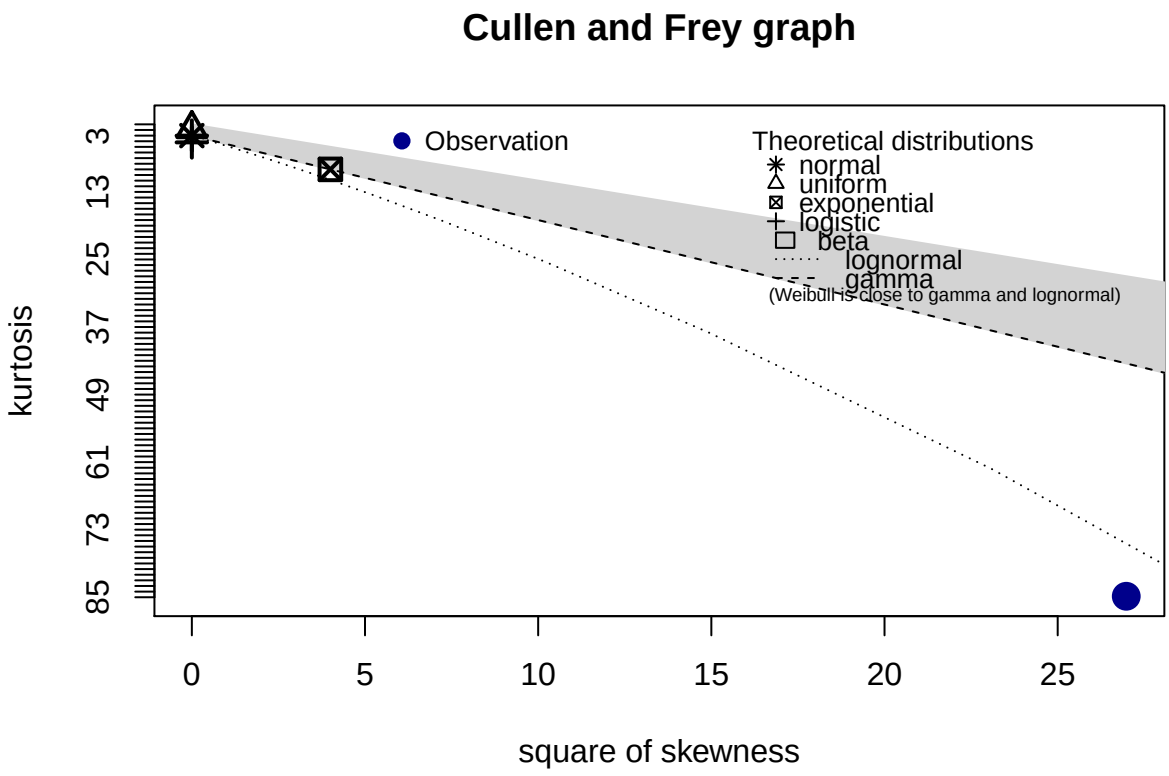

```
## summary statistics
## ------
## min:  0   max:  705
## median:  1
## mean:  0.971482
## estimated sd:  6.453969
## estimated skewness:  69.16765
## estimated kurtosis:  6755.931
```

```
# retweet_count
descdist(df1$retweet_count, discrete = FALSE)
```




```
## summary statistics
## ------
## min:  0   max:  5506
## median:  3
## mean:  52.73316
## estimated sd:  173.9064
## estimated skewness:  6.197483
## estimated kurtosis:  72.90395
```

```
# height
descdist(df1$height, discrete = FALSE)
```




```
## summary statistics
## ------
## min:  1   max:  203
## median:  172
## mean:  171.5292
## estimated sd:  10.65654
## estimated skewness:  -5.194209
## estimated kurtosis:  84.8246
```

1.3 Testing distribution assumptions

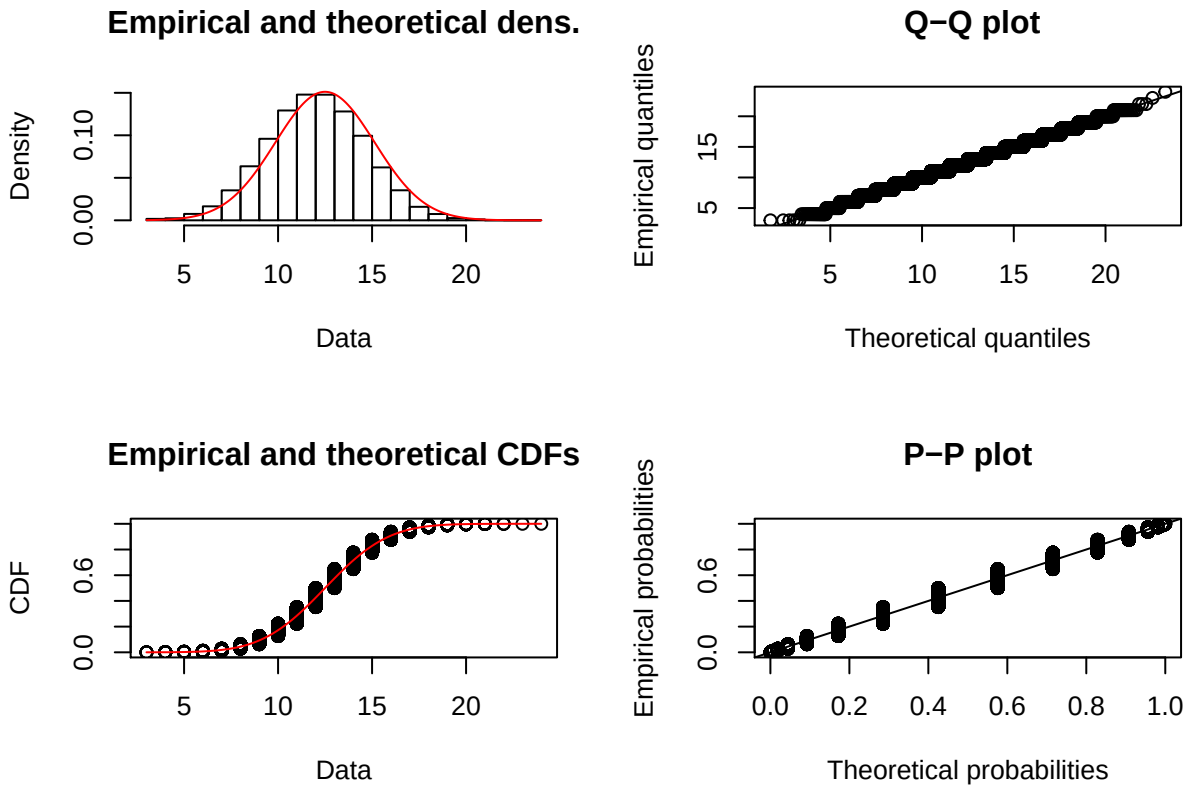
What happens when you test distribution assumptions (e.g. normal distributions or skewed, etc.)?

I ran fitdist plots for the normal data (education) and it looks good. However, when I ran the data that should have been uniform, it does not look very uniform (see created month and year).

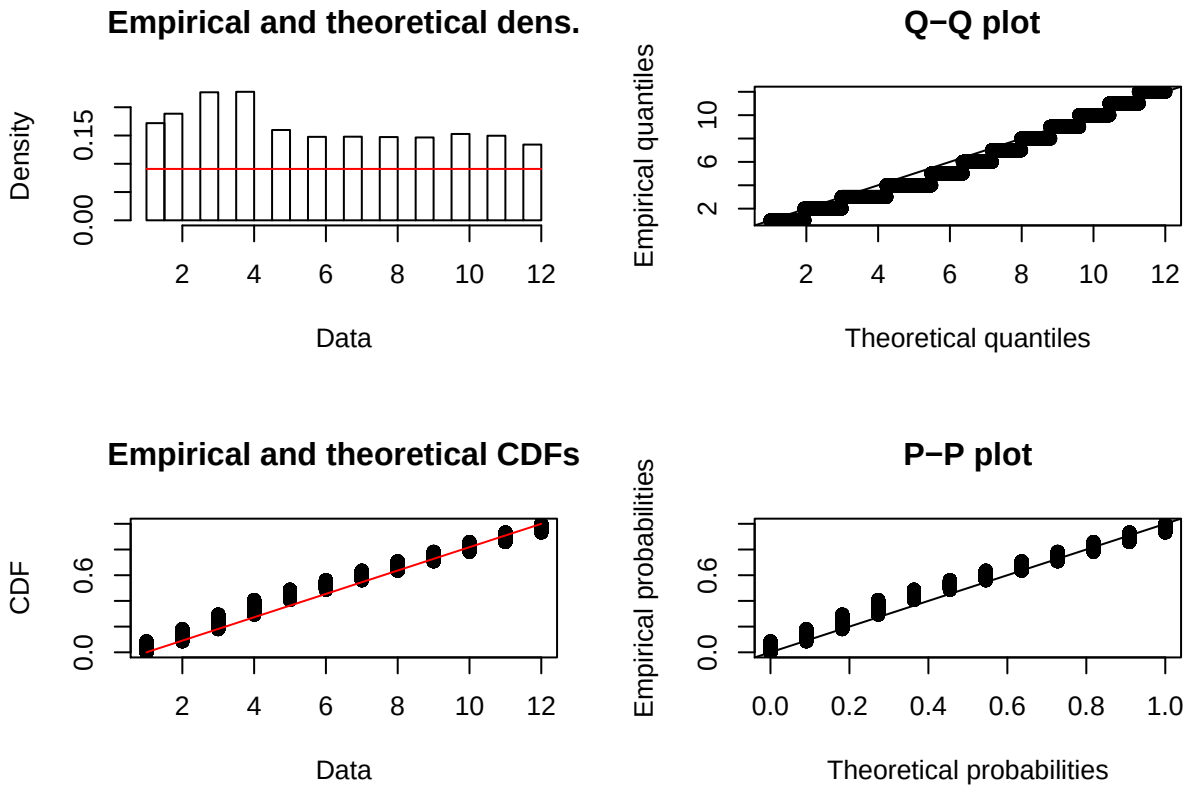
I ran age and it looked like it could be normal if there wasn't a lot of people that were zero.

I also tried to run other plots, but I could not find a good match for the data (Poisson, binomial, beta,...etc.)

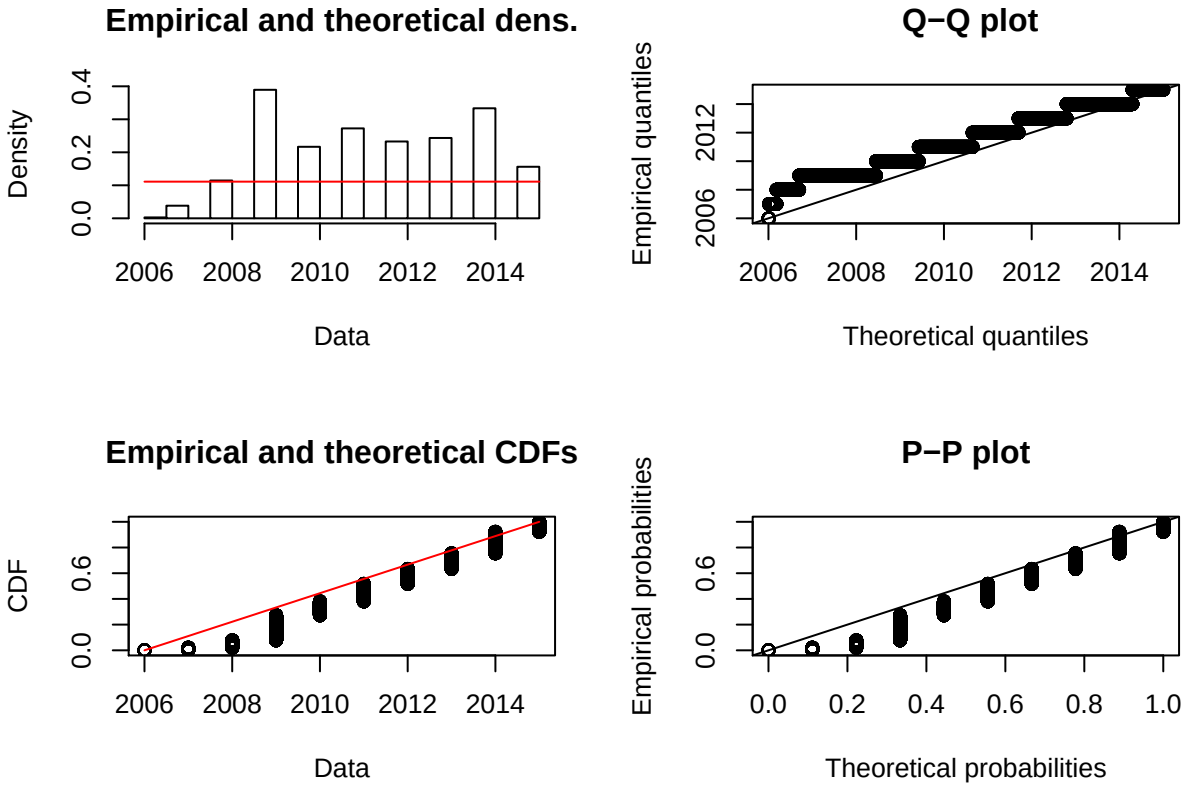
```
# Normal Distribution: education
fit.norm<-fitdist(df1$education, "norm")
plot(fit.norm)
```



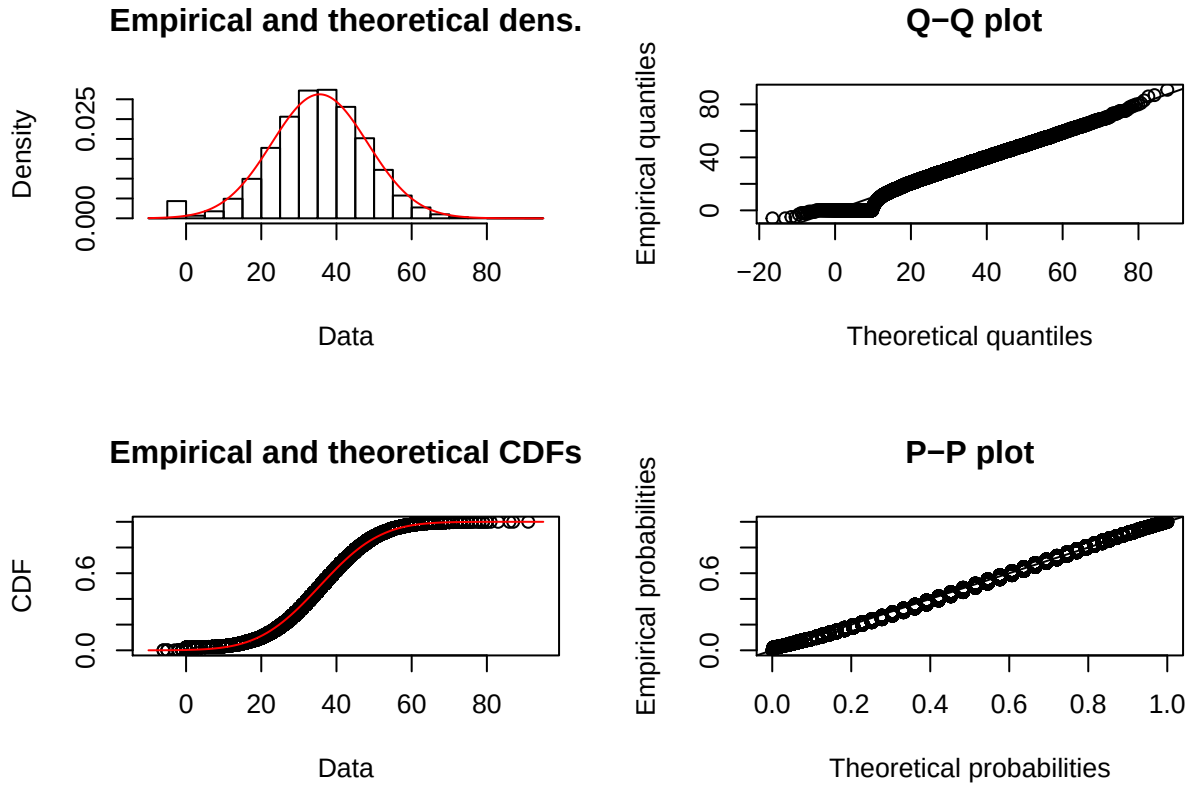

```
# Uniform Distribution: created_at_month
fit.unif<-fitdist(df1$created_at_month, "unif")
plot(fit.unif)
```




```r
# Uniform Distribution: created_at_year
fit.unif2<-fitdist(df1$created_at_year, "unif")
plot(fit.unif2)
```




```
# Uniform Distribution: created_at_year
fit.norm2<-fitdist(df1$age, "norm")
plot(fit.norm2)
```



1.4 What are the summary statistics?

Below are the summary information and first 6 rows of the data:

```
# Check the summary of the data:
summary(df1)
```

```
##  screen_name        created_at_month created_at_day  created_at_year
##  Length:21916       Min.   : 1.000   Min.   : 1.00   Min.   :2006
##  Class :character   1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.:2009
##  Mode  :character   Median : 6.000   Median :16.00   Median :2011
##                     Mean   : 6.069   Mean   :15.78   Mean   :2011
##                     3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:2013
##                     Max.   :12.000   Max.   :31.00   Max.   :2015
##    country            location         friends_count    followers_count
##  Length:21916       Length:21916       Min.   :   -84   Min.   :       0
##  Class :character   Class :character   1st Qu.:   123   1st Qu.:     105
##  Mode  :character   Mode  :character   Median :   324   Median :     336
##                                        Mean   :  1058   Mean   :    5859
##                                        3rd Qu.:   849   3rd Qu.:    1075
##                                        Max.   :660549   Max.   :22187643
##  statuses_count    favourites_count  favourited_count       dob_day
##  Min.   :      1   Min.   :      0   Min.   :     0.00   Min.   : 1.00
##  1st Qu.:    558   1st Qu.:     16   1st Qu.:     2.00   1st Qu.: 5.00
##  Median :   2341   Median :    164   Median :     9.00   Median :13.00
##  Mean   :  12486   Mean   :   2217   Mean   :    92.24   Mean   :13.49
```

```
##  3rd Qu.:   9348   3rd Qu.:    950   3rd Qu.:    36.00   3rd Qu.:21.00
##  Max.   :1136198   Max.   :1140139   Max.   :105005.00   Max.   :35.00
##     dob_year      dob_month           gender
##  Min.   :1900   Min.   :   1.000   Length:21916
##  1st Qu.:1965   1st Qu.:   3.000   Class :character
##  Median :1982   Median :   6.000   Mode  :character
##  Mean   :1976   Mean   :   6.398
##  3rd Qu.:1990   3rd Qu.:   9.000
##  Max.   :2000   Max.   :1992.000
##  mobile_favourites_count mobile_favourited_count   education
##  Min.   :     0.0        Min.   :      0         Min.   : 3.0
##  1st Qu.:     0.0        1st Qu.:      0         1st Qu.:11.0
##  Median :     0.0        Median :      0         Median :13.0
##  Mean   :   152.9        Mean   :    649         Mean   :12.5
##  3rd Qu.:     0.0        3rd Qu.:      0         3rd Qu.:14.0
##  Max.   :377123.0        Max.   :5032191         Max.   :24.0
##    experience          age            race                wage
##  Min.   :-32.00   Min.   :-6.00   Length:21916       Min.   :  5.00
##  1st Qu.:  0.00   1st Qu.:28.00   Class :character   1st Qu.: 13.52
##  Median :  7.00   Median :36.00   Mode  :character   Median : 20.36
##  Mean   : 10.88   Mean   :35.54                      Mean   : 22.97
##  3rd Qu.: 20.00   3rd Qu.:44.00                      3rd Qu.: 28.40
##  Max.   : 74.00   Max.   :91.00                      Max.   :104.97
##  retweeted_count    retweet_count         height
##  Min.   :  0.0000   Min.   :   0.00   Min.   :  1.0
##  1st Qu.:  0.0000   1st Qu.:   0.00   1st Qu.:165.0
##  Median :  1.0000   Median :   3.00   Median :172.0
##  Mean   :  0.9715   Mean   :  52.73   Mean   :171.5
##  3rd Qu.:  1.0000   3rd Qu.:  19.00   3rd Qu.:178.0
##  Max.   :705.0000   Max.   :5506.00   Max.   :203.0
```

```
# Look at the first 6 rows:
head(df1)
```

```
##   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
```

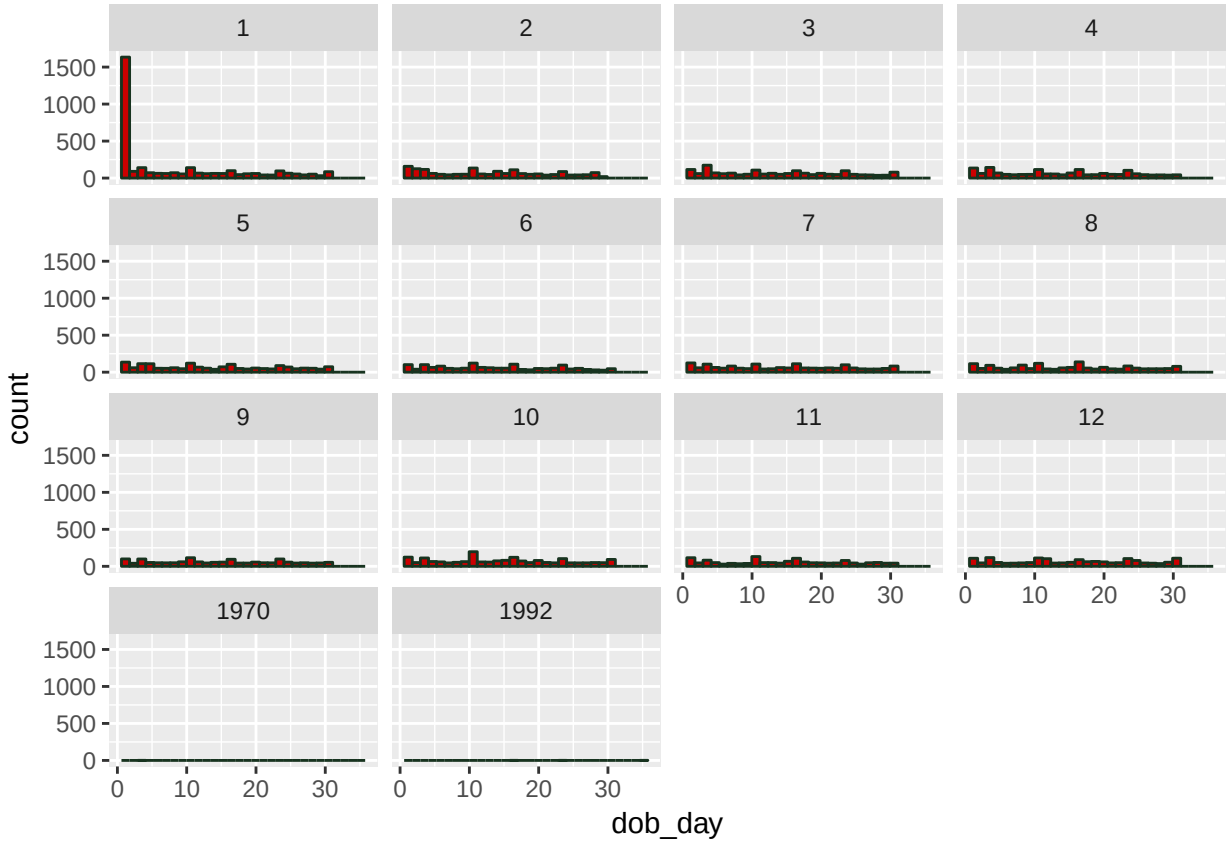
```
##   mobile_favourites_count mobile_favourited_count education experience age
## 1                       0                       0         8          0  29
## 2                       0                 5032191        15          0   0
## 3                       0                       0         9          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
```

1.5 Are there anomalies/outliers?

I do not see any anomalies (that fall into the constraints of insert, delete, or update), just outliers due to data quality problems. One big example is that the DOB year and age to not line up. For example screen_name CNN has a dob_year of 1999 and an age of 29, this is not correct. Age should be calculated from the birthdate and this should be updated in the database. Other examples are listed below:

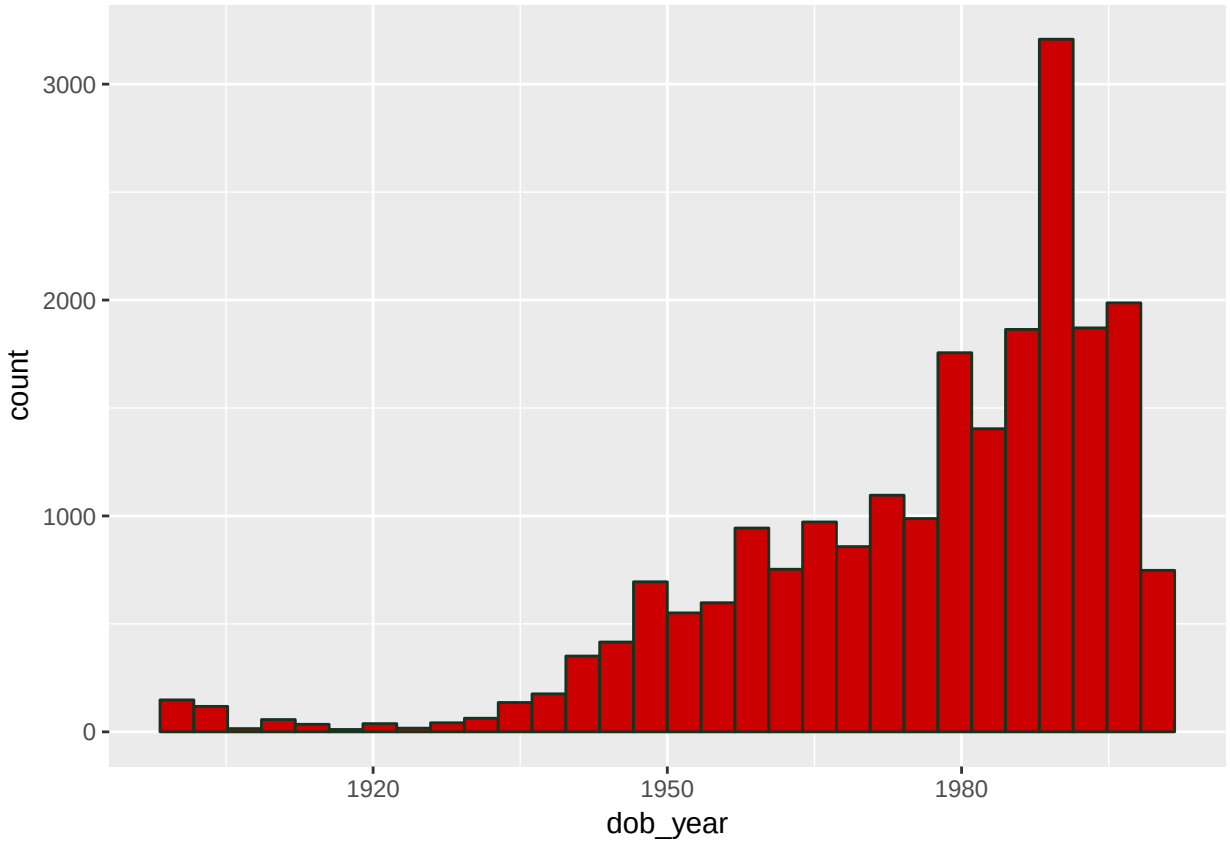
Below in dob_day, it appears there are more people born on the 1st day of the first month. This is probably because people do not put their correct birthday in. There is also an error in the data too, 1970 and 1992 should be NA.

```
qplot(
  x = dob_day,
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
) + facet_wrap( ~ dob_month) # a grid on one extra variable
```



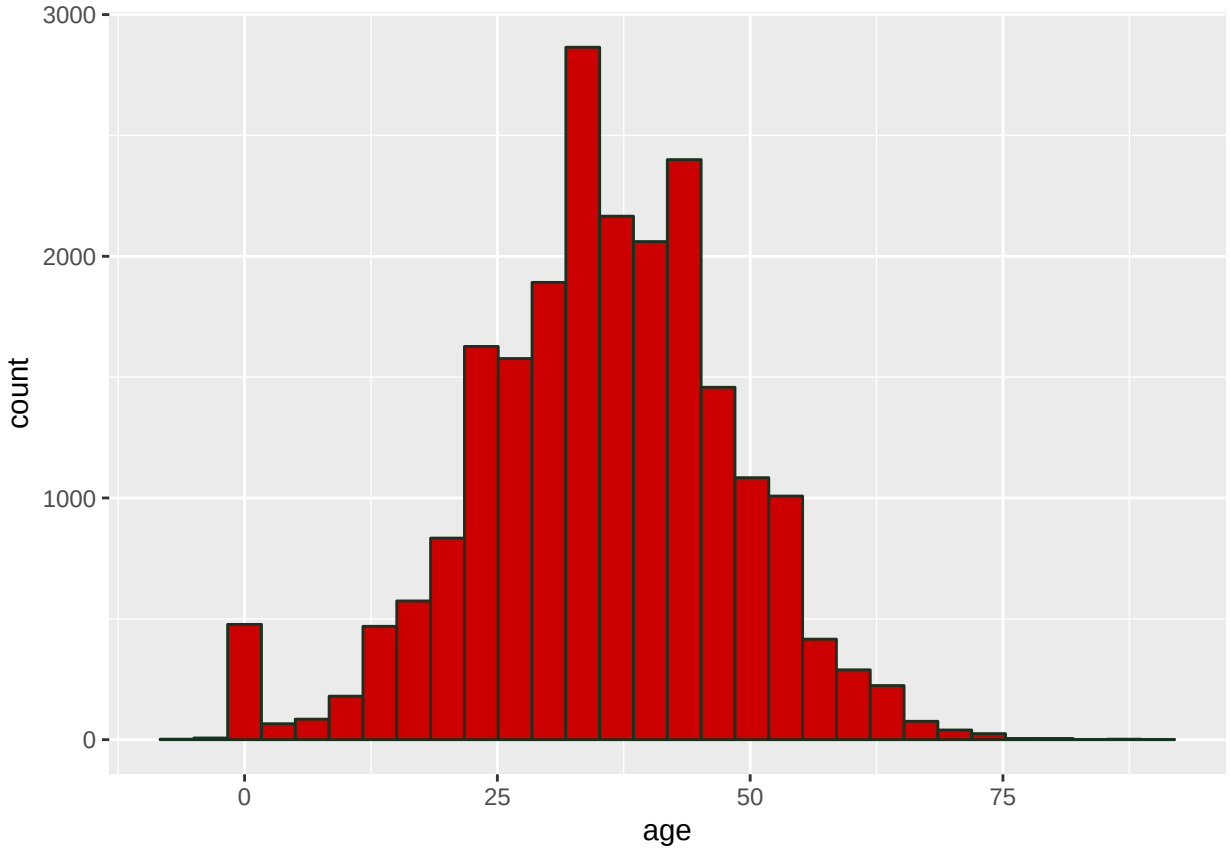
In the dob_year data, there appears to be a lot of people born around 1900, again, this is due to people not entering the correct information.

```
qplot(
  x = dob_year,
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```



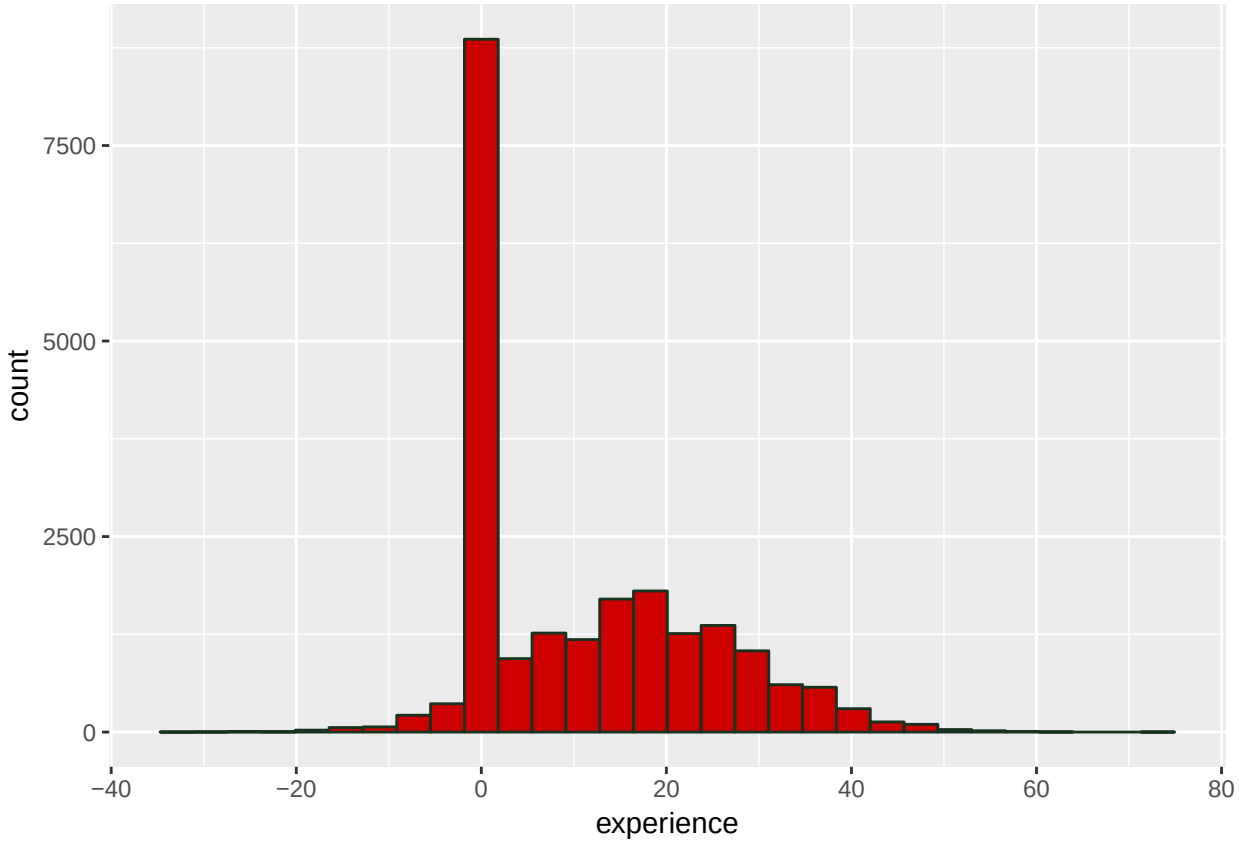
Age data has some negatives and a lot of zeros which is impossible. This information was probably entered incorrectly.

```
qplot(
  x = age,
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```



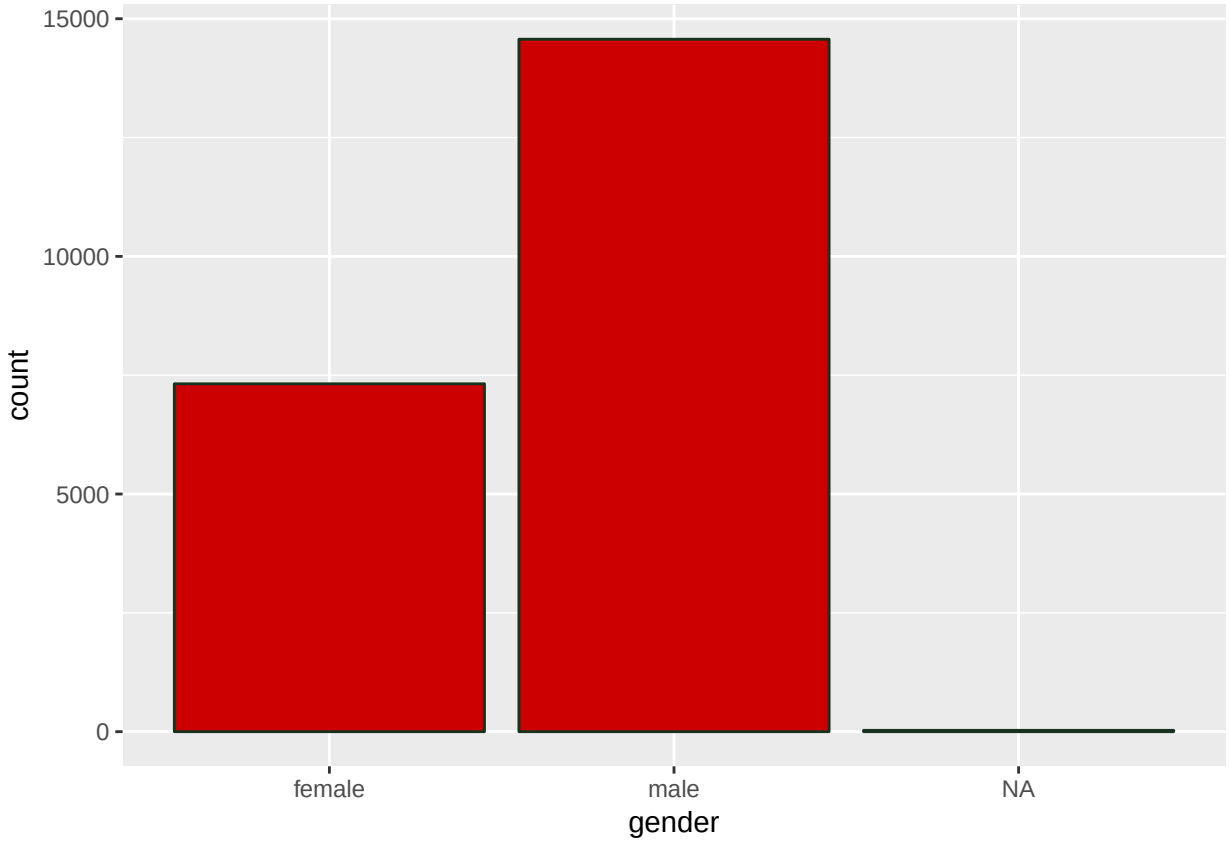
The experience data has negative information which is also impossible.

```
qplot(
  x = experience,
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```



I would expect gender data to be uniform, however, there are about twice as many males in this data.

```
qplot(
  x = gender,
  data = df1,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```

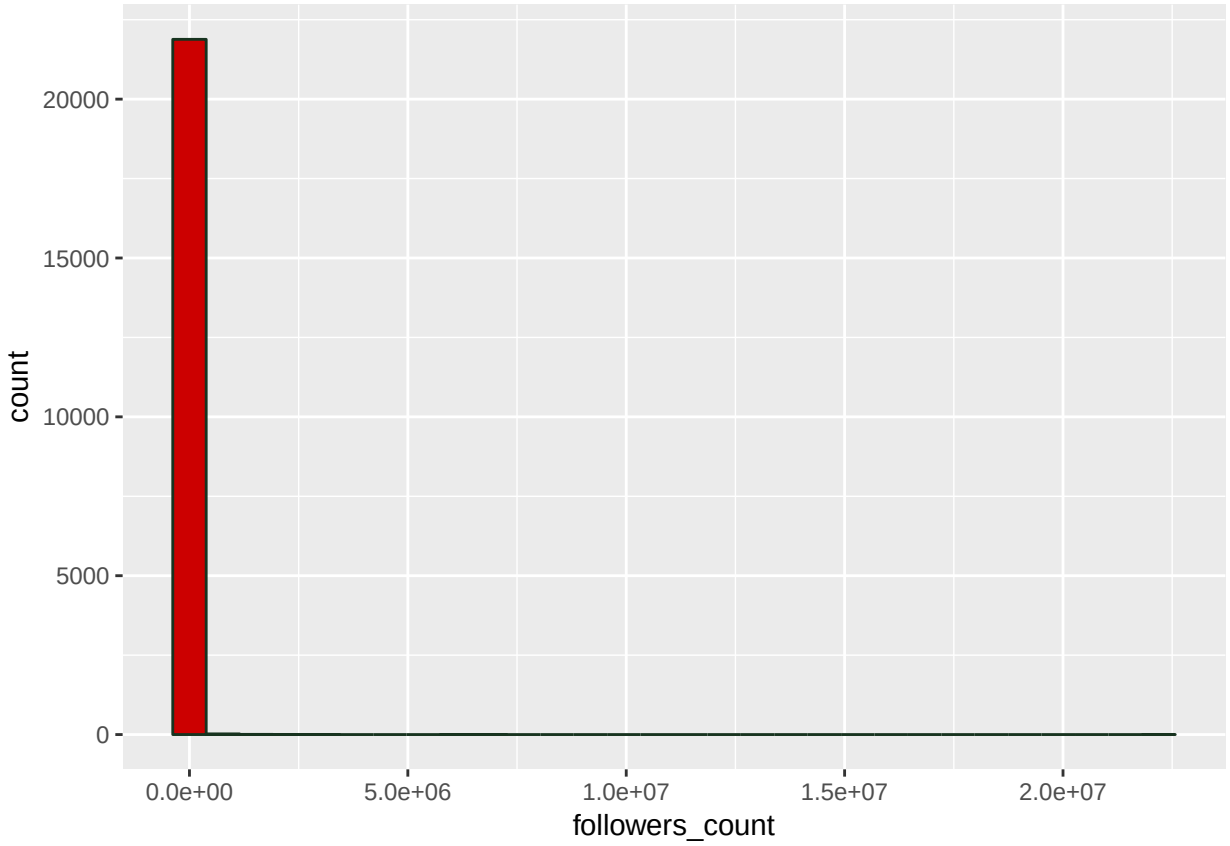


1.6 Can you identify the following:

1.6.1 useful raw data and transforms (e.g. log(x))

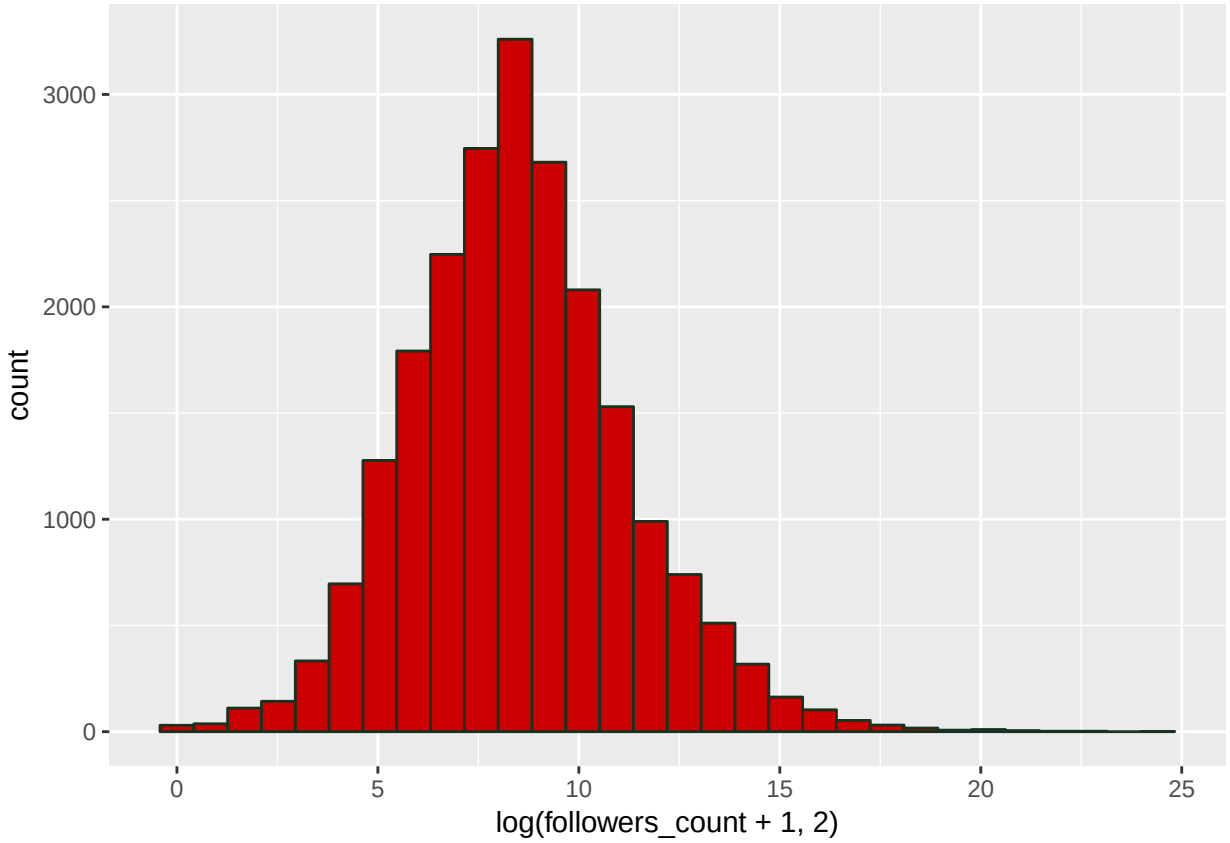
The followers count is a good example of this. The data should be good, but when it is plotted it looks like everyone has zero followers. This is because a lot of people have very little to zero follows and a very small amount of people have millions of follows, this makes the data difficult to graph:

```
qplot(
  x = followers_count,
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```



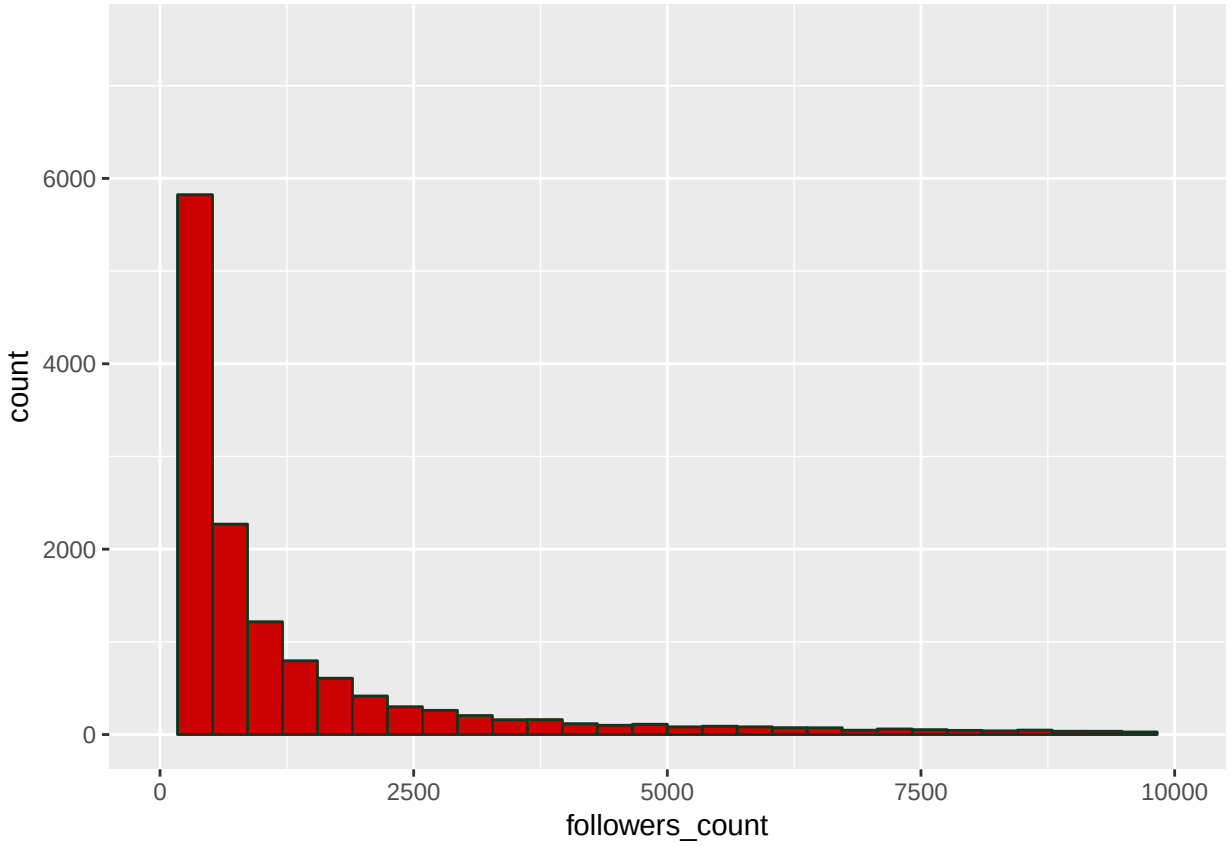
To fix this issue we can apply a Fat-tailed Log-transform:

```
qplot(
  x = log(followers_count + 1, 2),
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```



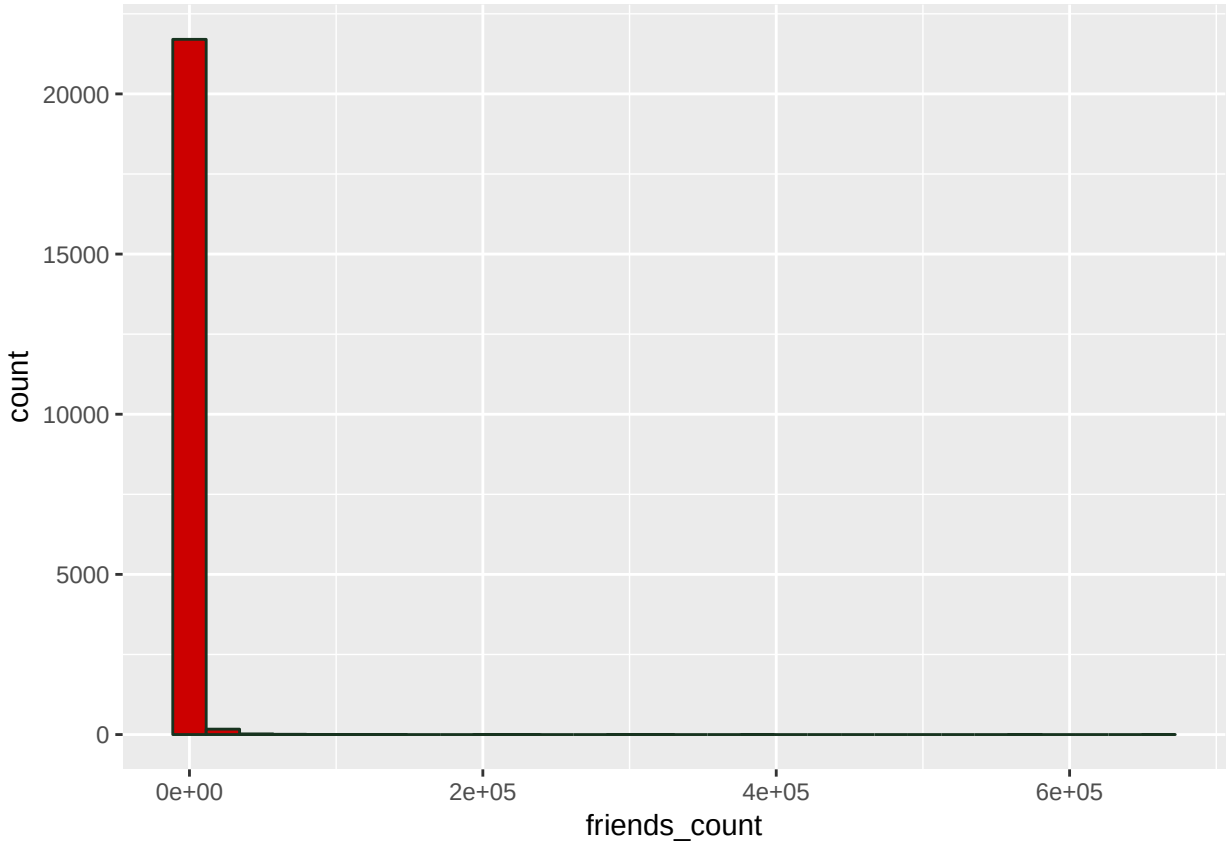
We could also zoom in on the info too:

```
qplot(
  x = followers_count,
  data = df1,
  bins = 30,
  xlim = c(0, 10000),
  color = I('#17331F'),
  fill = I('#CC0000')
)
```

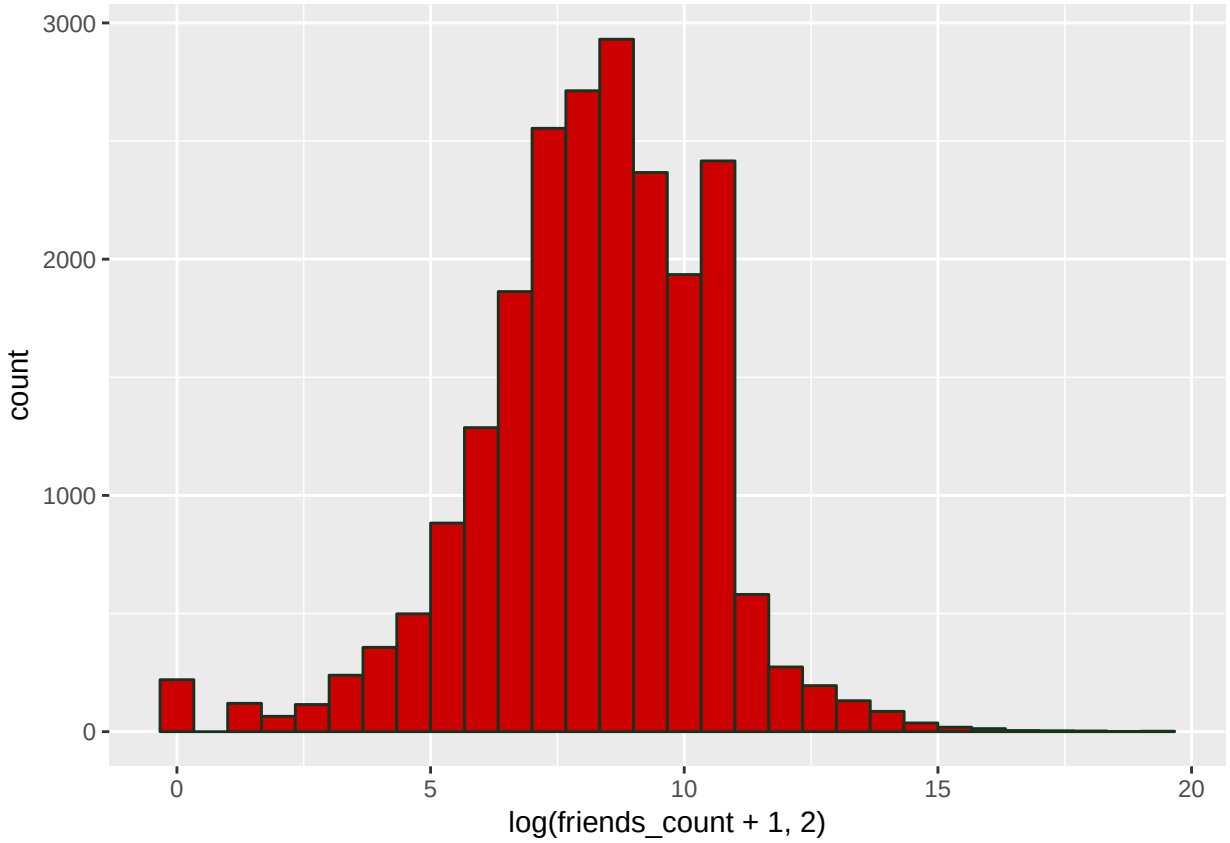


This can also be applied to the friends count too:

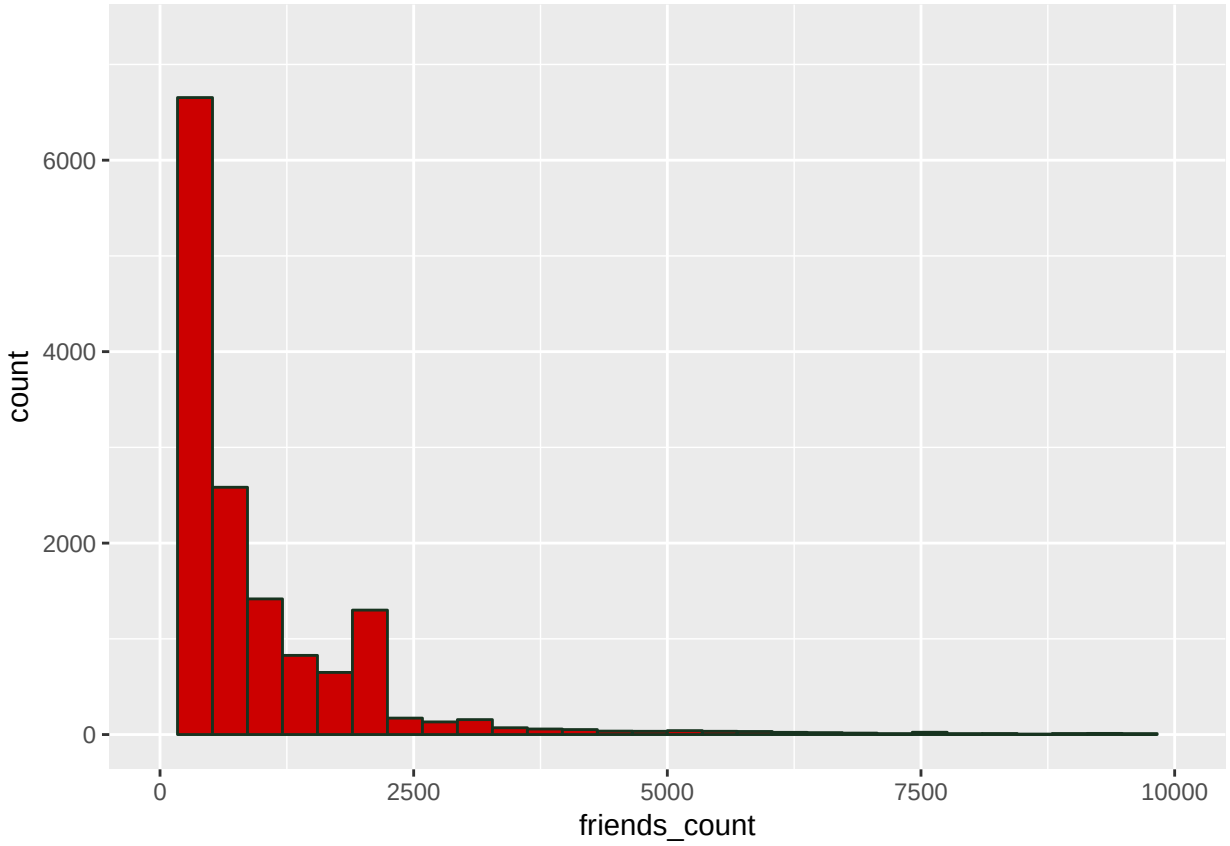
```
qplot(
  x = friends_count,
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```




```
qplot(
  x = log(friends_count + 1, 2),
  data = df1,
  bins = 30,
  color = I('#17331F'),
  fill = I('#CC0000')
)
```




```
qplot(
  x = friends_count,
  data = df1,
  bins = 30,
  xlim = c(0, 10000),
  color = I('#17331F'),
  fill = I('#CC0000')
)
```



1.6.2 data quality problems

I showed many examples of data quality problems in section 1.5, "Are there anomalies/outliers". These were all examples of inaccurate customer data. Birthdays are inaccurate, age is inaccurate, experience is inaccurate, and arguably anything that is dependent on the customer input has quality problems.

1.6.3 outliers

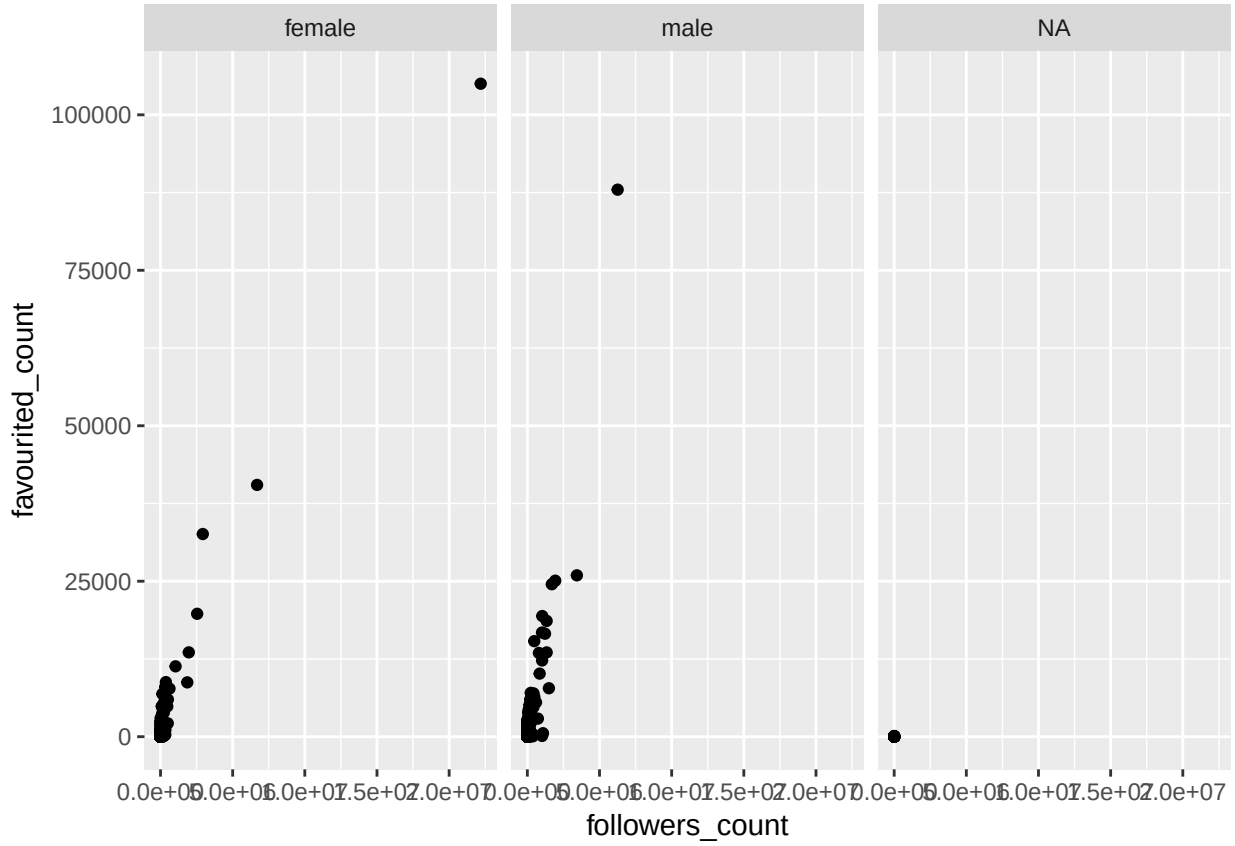
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. The outliers that I found were consistent with the data quality issues as described above.

1.6.4 subsets of interest

I think it is interesting to look at subsets of gender and education.

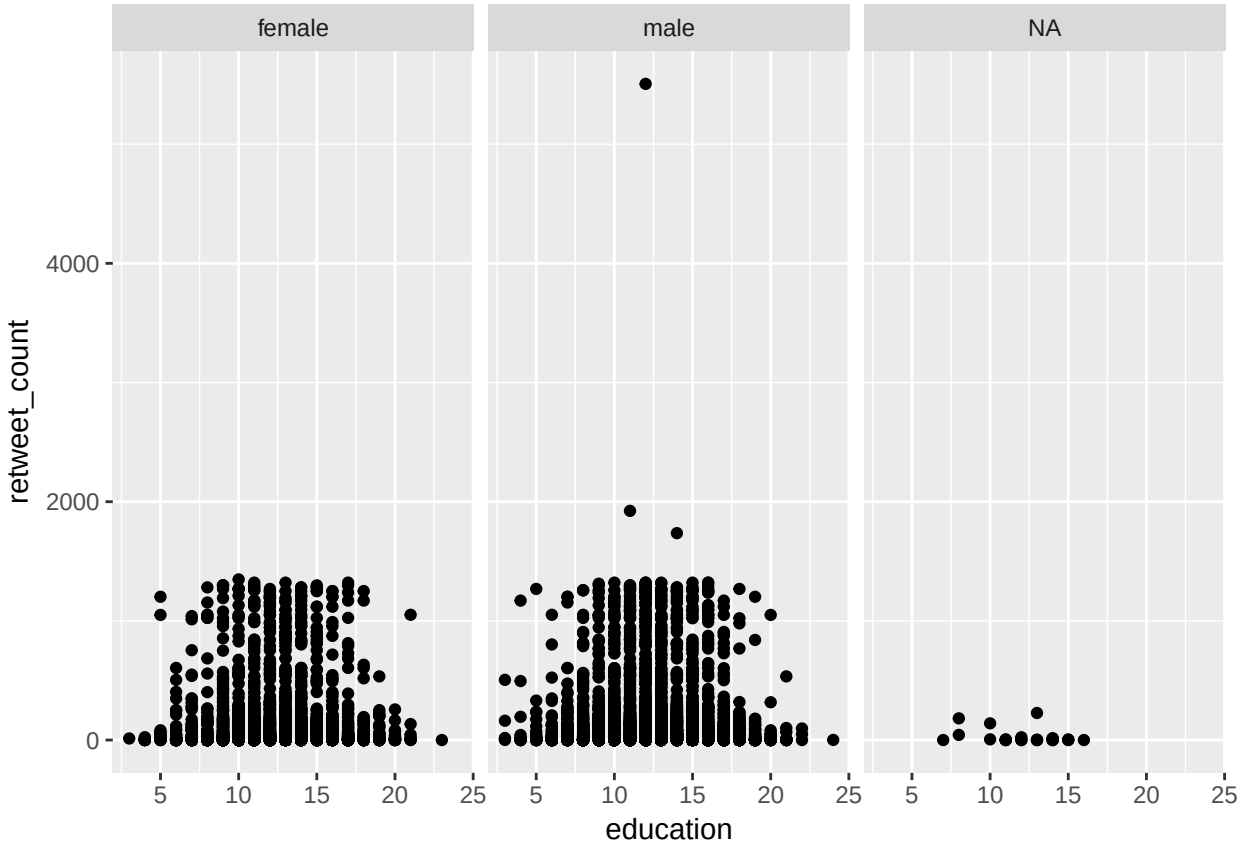
Below is a subset of followers count and favourited counted by gender:

```
qplot(followers_count, favourited_count, data=df1, facets = ~ gender)
```



This is an example of education level and retweer count by gender.

```
qplot(education, retweet_count, data=df1, facets = ~ gender)
```

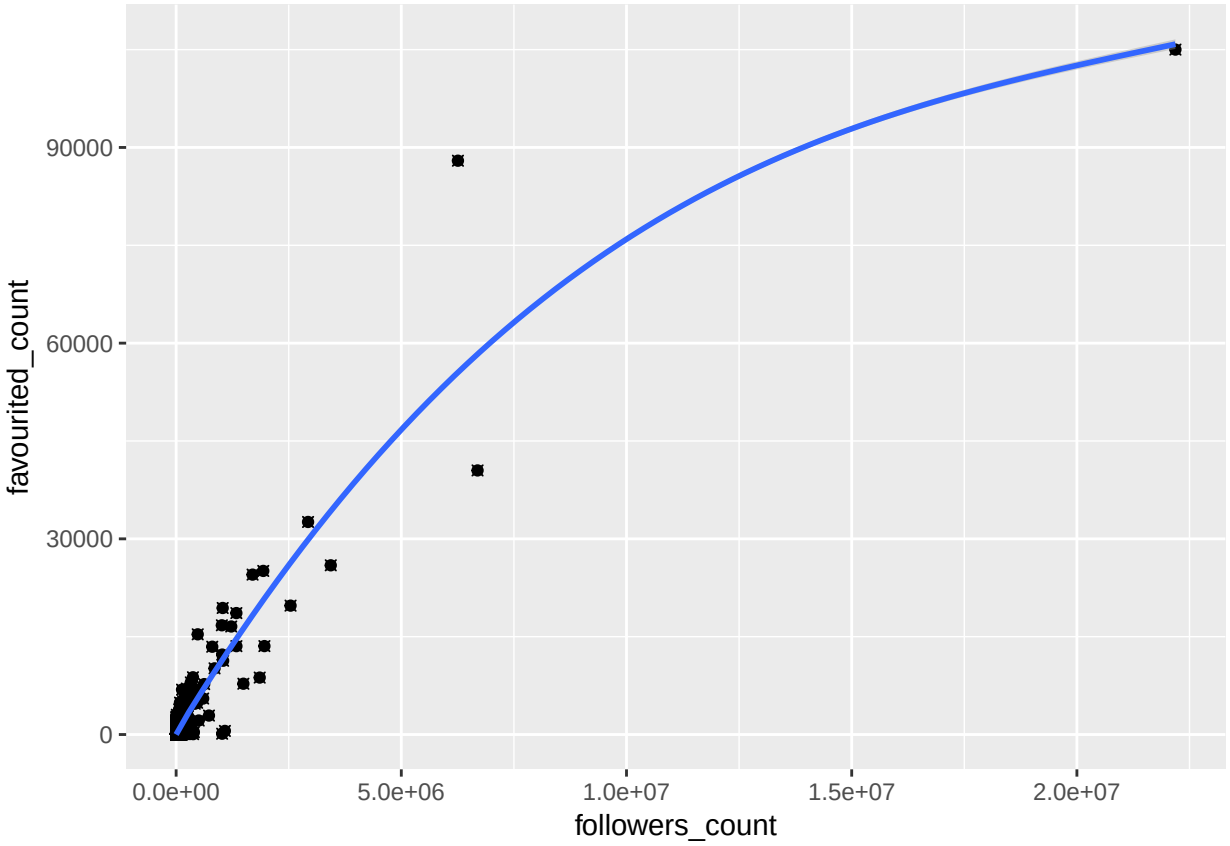


1.7 Finally, suggest any functional relationships.

It is hard to find functional relationships with this data. I did see a trend with followers count and favourited counted as shown below:

```
qplot(followers_count, favourited_count, data=df1) + geom_point(shape=4) + geom_smooth()
```

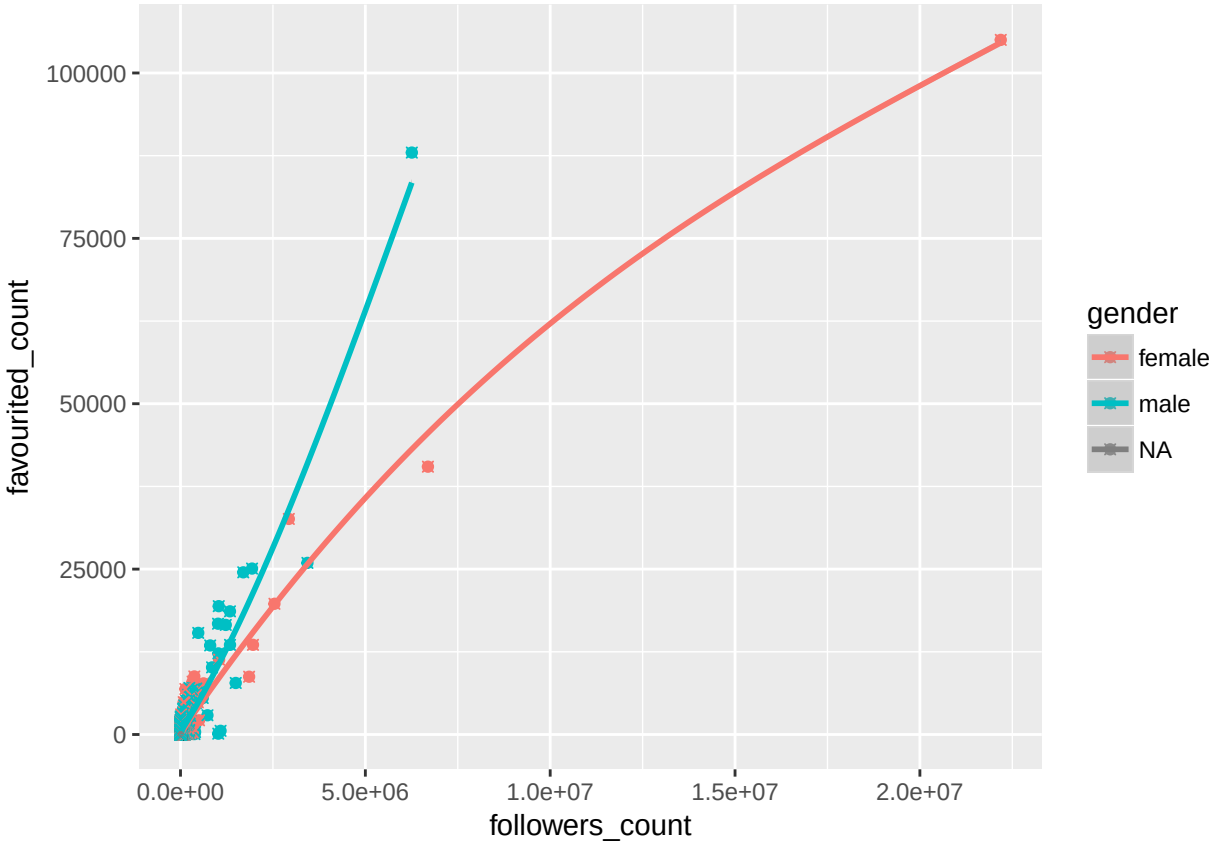
```
## `geom_smooth()` using method = 'gam'
```



I thought it would be interesting to see it broken down by gender and this is the result:

```
qplot(followers_count, favourited_count, data=df1,colour=gender) + geom_point(shape=4) + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```


100000
75000
50000
25000
0
favourited_count
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
followers_count
gender
female
male
NA