01_17_aggregate_delinquency

Siyang Ni

2023-01-17

Note

For every variable, *Column %(NA-)* reports the frequency without any NAs. I also reported the cumulative frequency in column cum.%(NA-).

I also plotted the frequency for each variable, giving a more intuitive way of digesting the frequency information. All NAs are indicated by the "last" bar of the histogram, unless otherwise indicated.

Loading datasets

```
setwd("D:/research/Monitoring-The-Future/mtf_19_12")

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0001/37841-0001-Data.rda")
core <- da37841.0001
rm(da37841.0001)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0002/37841-0002-Data.rda")
ds2 <- da37841.0002
rm(da37841.0002)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0003/37841-0003-Data.rda")
ds3 <- da37841.0003
rm(da37841.0003)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0004/37841-0004-Data.rda")
ds4 <- da37841.0004
rm(da37841.0004)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0005/37841-0005-Data.rda")
ds5 <- da37841.0005
rm(da37841.0005)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0006/37841-0006-Data.rda")
ds6 <- da37841.0006
rm(da37841.0006)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0007/37841-0007-Data.rda")
ds7 <- da37841.0007
rm(da37841.0007)
```

```
load("D:/research/Monitoring-The-Future/mtf_19_12/DS0008/37841-0008-Data.rda")
ds8 <- da37841.0008
rm(da37841.0008)

load("D:/research/Monitoring-The-Future/mtf_19_12/DS0009/37841-0009-Data.rda")
ds9 <- da37841.0009
rm(da37841.0009)
```

Aggregate All

Change factor to numeric first

```
which(colnames(ds3) == "V2279")
```

```
## [1] 65
```

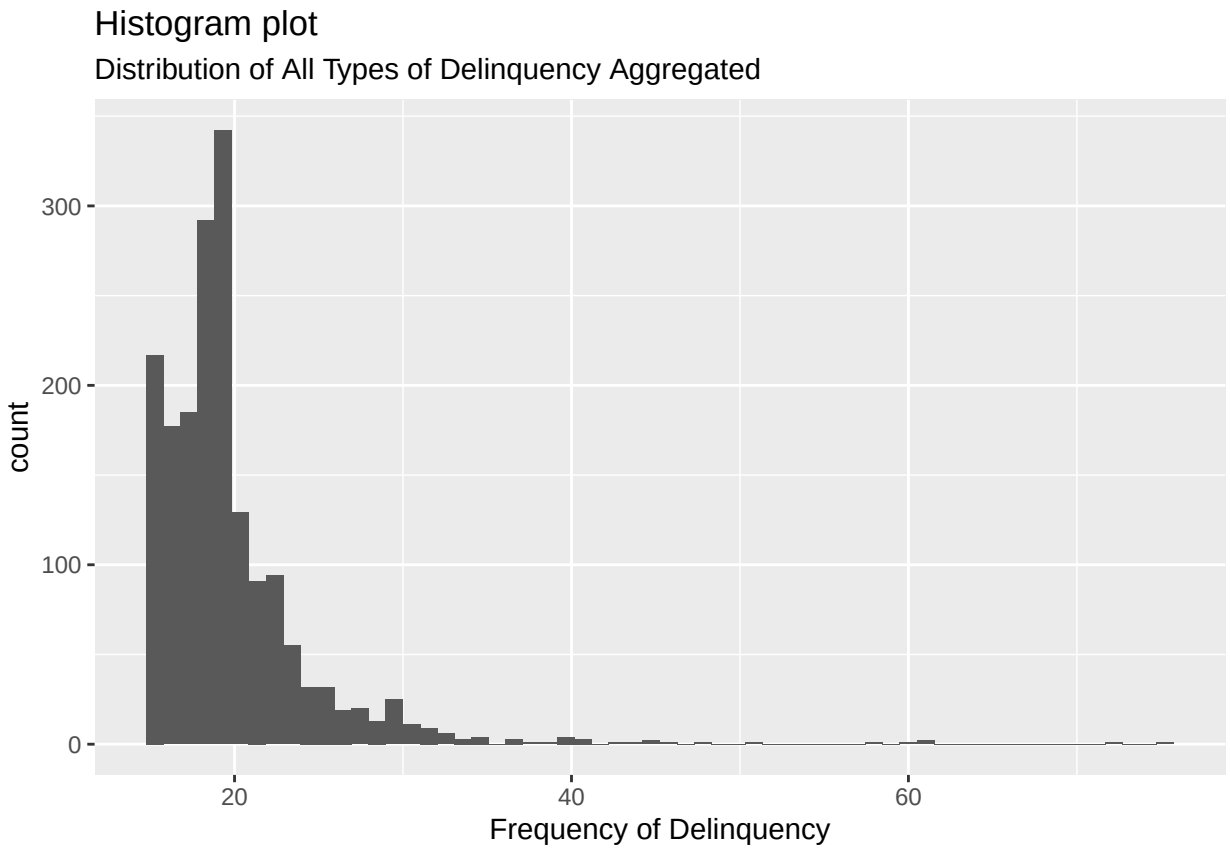
```
which(colnames(ds3) == "V2293")
```

```
## [1] 79
```

```
x <- ds3[c(65:79)]

ndelinquency <- as.data.frame(mapply(as.numeric, x))
```

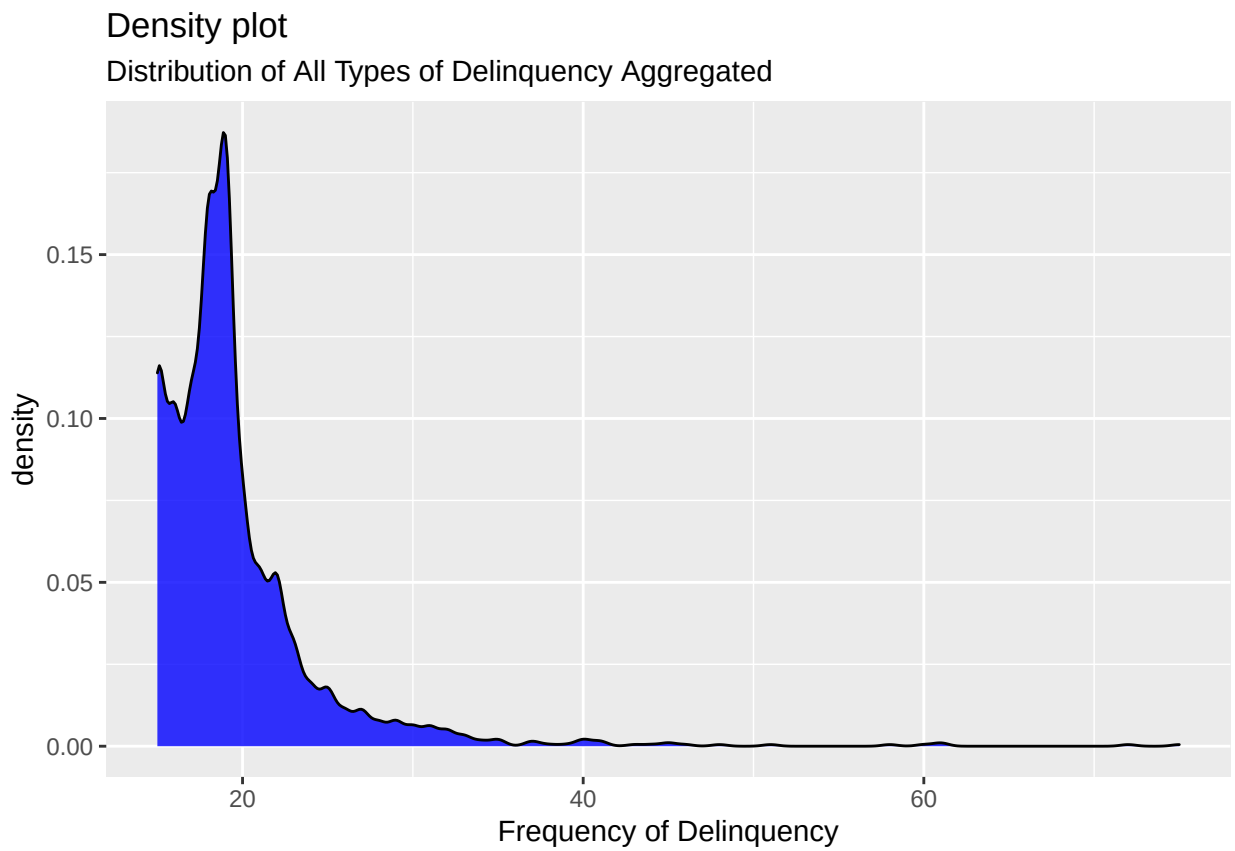
```
attach(ndelinquency)

ndelinquency$delin_all <- V2279 + V2280 + V2281 + V2282 + V2283 + V2284 + V2285 + V2286 + V2287 + V2288


plot1 <- ggplot(ndelinquency, aes(x=delin_all)) +
  geom_histogram(bins=60) +
  labs(title="Histogram plot",
         subtitle="Distribution of All Types of Delinquency Aggregated",
         x="Frequency of Delinquency")

plot1
```

```
## Warning: Removed 518 rows containing non-finite values (stat_bin).
```




```
plot2 <- ggplot(ndelinquency, aes(x=delin_all)) +
  geom_density(fill = 'blue', alpha=0.8) +
    labs(title="Density plot",
         subtitle="Distribution of All Types of Delinquency Aggregated",
         x="Frequency of Delinquency")

plot2
```

```
## Warning: Removed 518 rows containing non-finite values (stat_density).
```



Density plot and histogram shows on a 0 - 75 scale, most data falls within the 0 - 30 range, with two peaks at around 0 and 20. Given data skewness at this level, I recommend dichotomize all delinquency variables.

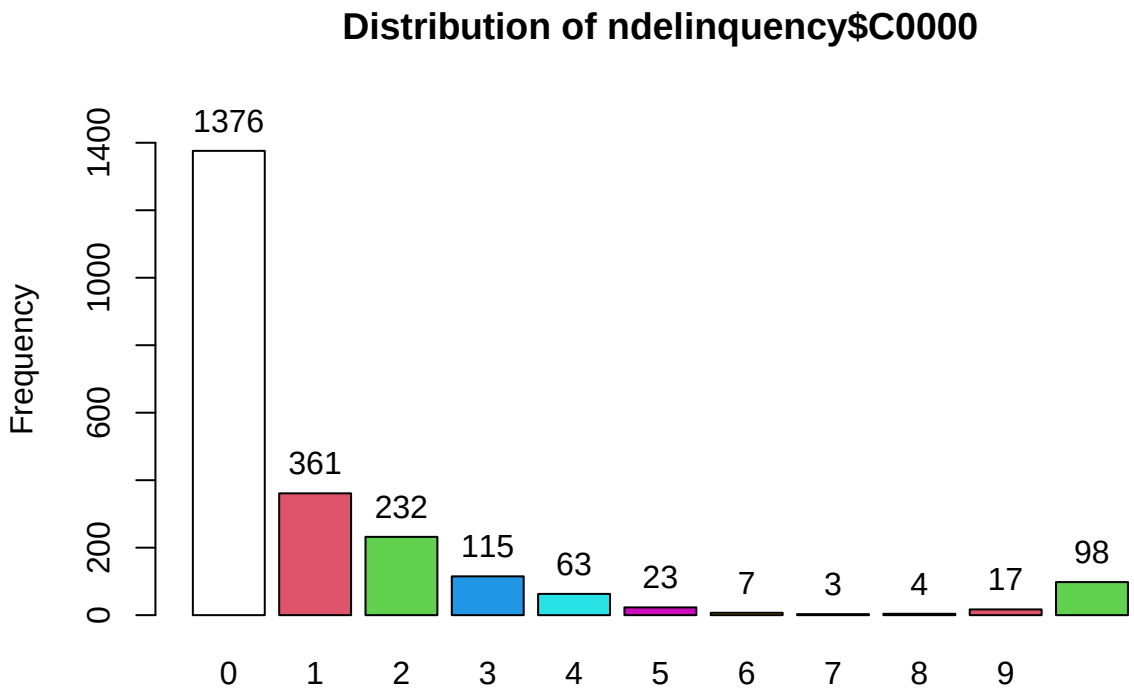
Dichotomize and Aggregate

C0000: Property-related Delinquency

```
library(dplyr)

# Create Numerical Dichotomous Variables


ndelinquency$V2285D<-ifelse(ndelinquency$V2285==1,0,1)
ndelinquency$V2286D<-ifelse(ndelinquency$V2286==1,0,1)
ndelinquency$V2287D<-ifelse(ndelinquency$V2287==1,0,1)
ndelinquency$V2288D<-ifelse(ndelinquency$V2288==1,0,1)
ndelinquency$V2289D<-ifelse(ndelinquency$V2289==1,0,1)
ndelinquency$V2290D<-ifelse(ndelinquency$V2290==1,0,1)
ndelinquency$V2291D<-ifelse(ndelinquency$V2291==1,0,1)
ndelinquency$V2292D<-ifelse(ndelinquency$V2292==1,0,1)
ndelinquency$V2293D<-ifelse(ndelinquency$V2293==1,0,1)

# Sum these dichotomous

attach(ndelinquency)
```

```
## The following objects are masked from ndelinquency (pos = 3):
##
##     V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2286, V2287,
##     V2288, V2289, V2290, V2291, V2292, V2293
```

```
ndelinquency$C0000 <- V2285D + V2286D + V2287D + V2288D + V2289D + V2290D + V2291D + V2292D + V2293D

# Showing the distribution of the new composite variable
tab1(ndelinquency$C0000, cum.percent = TRUE)
```




```
## ndelinquency$C0000 :
##         Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0            1376     59.9       59.9     62.5       62.5
## 1             361     15.7       75.6     16.4       78.9
## 2             232     10.1       85.6     10.5       89.5
## 3             115      5.0       90.6      5.2       94.7
## 4              63      2.7       93.4      2.9       97.5
## 5              23      1.0       94.4      1.0       98.6
## 6               7      0.3       94.7      0.3       98.9
## 7               3      0.1       94.8      0.1       99.0
## 8               4      0.2       95.0      0.2       99.2
## 9              17      0.7       95.7      0.8      100.0
## <NA>           98      4.3      100.0      0.0      100.0
##   Total      2299    100.0      100.0    100.0      100.0
```

C0001: Violent Delinquency

The making of the composite violdent delinquency variable follows the same logic described above in the property-related delinquency.

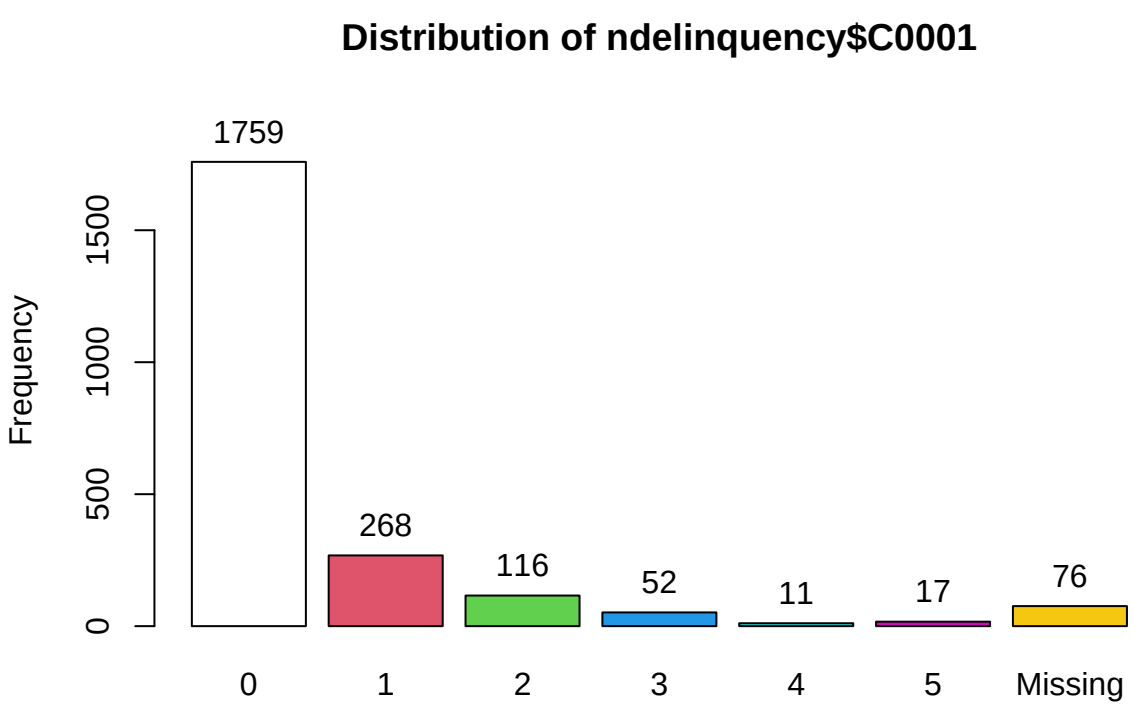
```
# Creating numerical dichotomous variables

attach(ndelinquency)
```

```
## The following objects are masked from ndelinquency (pos = 3):
##
##     delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2285D,
##     V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D, V2290,
##     V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 4):
##
##     V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2286, V2287,
##     V2288, V2289, V2290, V2291, V2292, V2293
```

```
ndelinquency$V2280D<-ifelse(ndelinquency$V2280==1,0,1)
ndelinquency$V2281D<-ifelse(ndelinquency$V2281==1,0,1)
ndelinquency$V2282D<-ifelse(ndelinquency$V2282==1,0,1)
ndelinquency$V2283D<-ifelse(ndelinquency$V2283==1,0,1)
ndelinquency$V2284D<-ifelse(ndelinquency$V2284==1,0,1)

attach(ndelinquency)
```

```
## The following objects are masked from ndelinquency (pos = 3):
##
##     C0000, delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285,
##     V2285D, V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D,
##     V2290, V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 4):
##
##     delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2285D,
##     V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D, V2290,
##     V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 5):
##
##     V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2286, V2287,
##     V2288, V2289, V2290, V2291, V2292, V2293
```

```
# Sum across these variables:
ndelinquency$C0001 <- V2280D + V2281D + V2282D + V2283D + V2284D

# Showing the distribution:
tab1(ndelinquency$C0001, cum.percent = TRUE)
```




```
## ndelinquency$C0001 :
##         Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0            1759     76.5       76.5     79.1       79.1
## 1             268     11.7       88.2     12.1       91.2
## 2             116      5.0       93.2      5.2       96.4
## 3              52      2.3       95.5      2.3       98.7
## 4              11      0.5       96.0      0.5       99.2
## 5              17      0.7       96.7      0.8      100.0
## <NA>           76      3.3      100.0      0.0      100.0
##   Total      2299    100.0      100.0    100.0      100.0
```

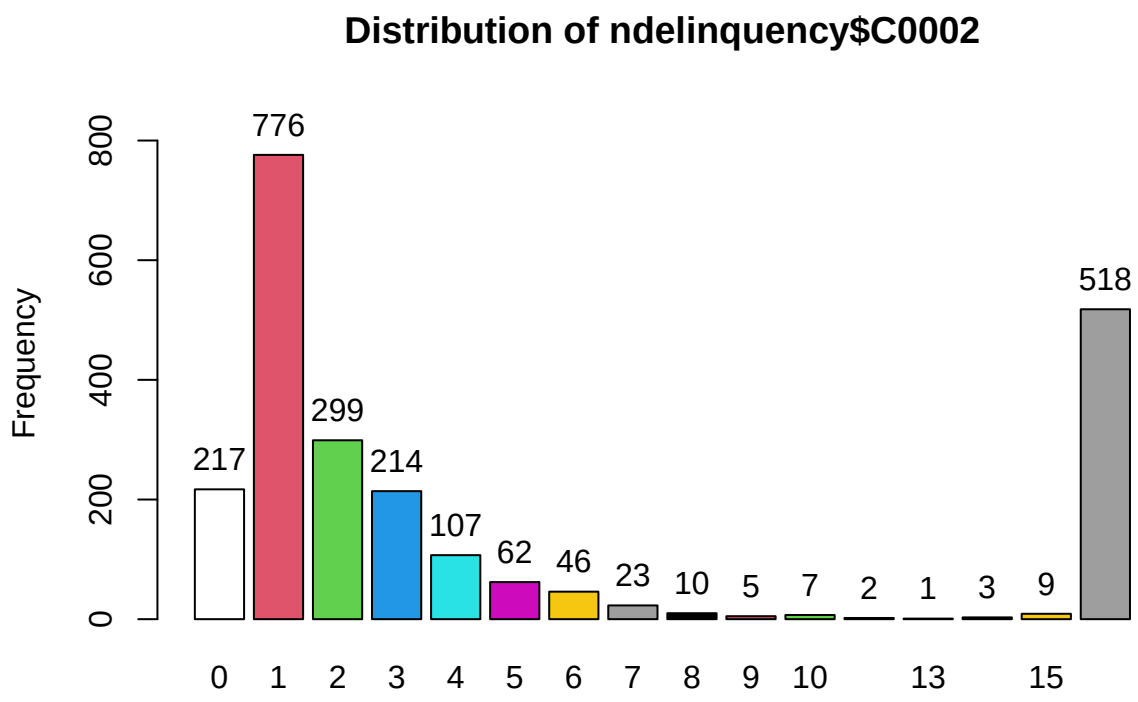
C0002: All Delinquency

```
# Dichotomize V2279 Fight with Parents
ndelinquency$V2279D<-ifelse(ndelinquency$V2279==1,0,1)

attach(ndelinquency)
```

```
## The following objects are masked from ndelinquency (pos = 3):
##
##     C0000, delin_all, V2279, V2280, V2280D, V2281, V2281D, V2282,
##     V2282D, V2283, V2283D, V2284, V2284D, V2285, V2285D, V2286, V2286D,
##     V2287, V2287D, V2288, V2288D, V2289, V2289D, V2290, V2290D, V2291,
##     V2291D, V2292, V2292D, V2293, V2293D
```

```
## The following objects are masked from ndelinquency (pos = 4):
##
##     C0000, delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285,
##     V2285D, V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D,
##     V2290, V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 5):
##
##     delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2285D,
##     V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D, V2290,
##     V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 6):
##
##     V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2286, V2287,
##     V2288, V2289, V2290, V2291, V2292, V2293
```

```r
# Sum across All Delinquency variables:

ndelinquency$C0002 <- V2279D + V2280D + V2281D + V2282D + V2283D + V2284D + V2285D + V2286D + V2287D + V

# Showing the Distribution
tab1(ndelinquency$C0002, cum.percent = TRUE)
```




```
## ndelinquency$C0002 :
```

```
##         Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0             217      9.4        9.4     12.2       12.2
## 1             776     33.8       43.2     43.6       55.8
## 2             299     13.0       56.2     16.8       72.5
## 3             214      9.3       65.5     12.0       84.6
## 4             107      4.7       70.2      6.0       90.6
## 5              62      2.7       72.9      3.5       94.0
## 6              46      2.0       74.9      2.6       96.6
## 7              23      1.0       75.9      1.3       97.9
## 8              10      0.4       76.3      0.6       98.5
## 9               5      0.2       76.5      0.3       98.8
## 10              7      0.3       76.8      0.4       99.2
## 12              2      0.1       76.9      0.1       99.3
## 13              1      0.0       76.9      0.1       99.3
## 14              3      0.1       77.1      0.2       99.5
## 15              9      0.4       77.5      0.5      100.0
## <NA>          518     22.5      100.0      0.0      100.0
##   Total      2299    100.0      100.0    100.0      100.0
```

```
detach (ndelinquency)
```

Finding out the "Useful Aggregate"

I experimented with picking out every individual delinquency variable from the aggregate variable, and checked the distribution. The process is almost automatic with R, but showing it here will make this presentation unnecessarily long, so the following section only shows points I found worthy of attention.

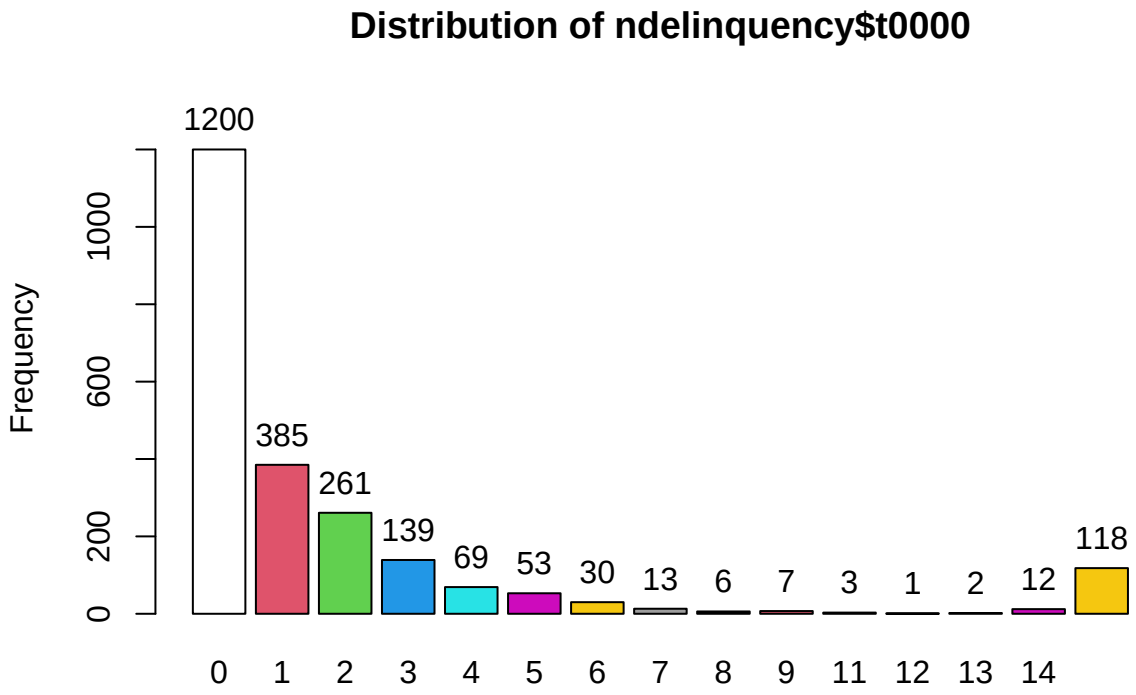
First, fighting with parents (V2279) has a significantly higher prevalence level than other variables, it also has lots of missing data. Picking it out would change the overall prevalence level. The peak would change from 1 to 0.

```
attach(ndelinquency)
```

```
## The following objects are masked from ndelinquency (pos = 3):
##
##     C0000, delin_all, V2279, V2280, V2280D, V2281, V2281D, V2282,
##     V2282D, V2283, V2283D, V2284, V2284D, V2285, V2285D, V2286, V2286D,
##     V2287, V2287D, V2288, V2288D, V2289, V2289D, V2290, V2290D, V2291,
##     V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 4):
##
##     C0000, delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285,
##     V2285D, V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D,
##     V2290, V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D

## The following objects are masked from ndelinquency (pos = 5):
##
##     delin_all, V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2285D,
##     V2286, V2286D, V2287, V2287D, V2288, V2288D, V2289, V2289D, V2290,
##     V2290D, V2291, V2291D, V2292, V2292D, V2293, V2293D
```

```
## The following objects are masked from ndelinquency (pos = 6):
##
##     V2279, V2280, V2281, V2282, V2283, V2284, V2285, V2286, V2287,
##     V2288, V2289, V2290, V2291, V2292, V2293
```

```
ndelinquency$t0000 <- V2280D + V2281D + V2282D + V2283D + V2284D + V2285D + V2286D + V2287D + V2288D + V

tab1(ndelinquency$t0000, cum.percent = TRUE)
```




```
## ndelinquency$t0000 :
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0             1200     52.2       52.2     55.0       55.0
## 1              385     16.7       68.9     17.7       72.7
## 2              261     11.4       80.3     12.0       84.6
## 3              139      6.0       86.3      6.4       91.0
## 4               69      3.0       89.3      3.2       94.2
## 5               53      2.3       91.6      2.4       96.6
## 6               30      1.3       93.0      1.4       98.0
## 7               13      0.6       93.5      0.6       98.6
## 8                6      0.3       93.8      0.3       98.9
## 9                7      0.3       94.1      0.3       99.2
## 11               3      0.1       94.2      0.1       99.3
## 12               1      0.0       94.3      0.0       99.4
## 13               2      0.1       94.3      0.1       99.4
## 14              12      0.5       94.9      0.6      100.0
```

```
## <NA>          118      5.1      100.0       0.0      100.0
##   Total      2299    100.0      100.0     100.0      100.0
```

The other five variables that are relatively more prevalent are:

V2281: During the LAST 12 MONTHS, how often have you . . . gotten into a serious fight in school or at work?

V2282: During the LAST 12 MONTHS, how often have you . . . taken part in a fight where a group of your friends were against another group?

V2283: During the LAST 12 MONTHS, how often have you . . . hurt someone badly enough to need bandages or a doctor?

V2285: During the LAST 12 MONTHS, how often have you . . . taken something not belonging to you worth under $50?

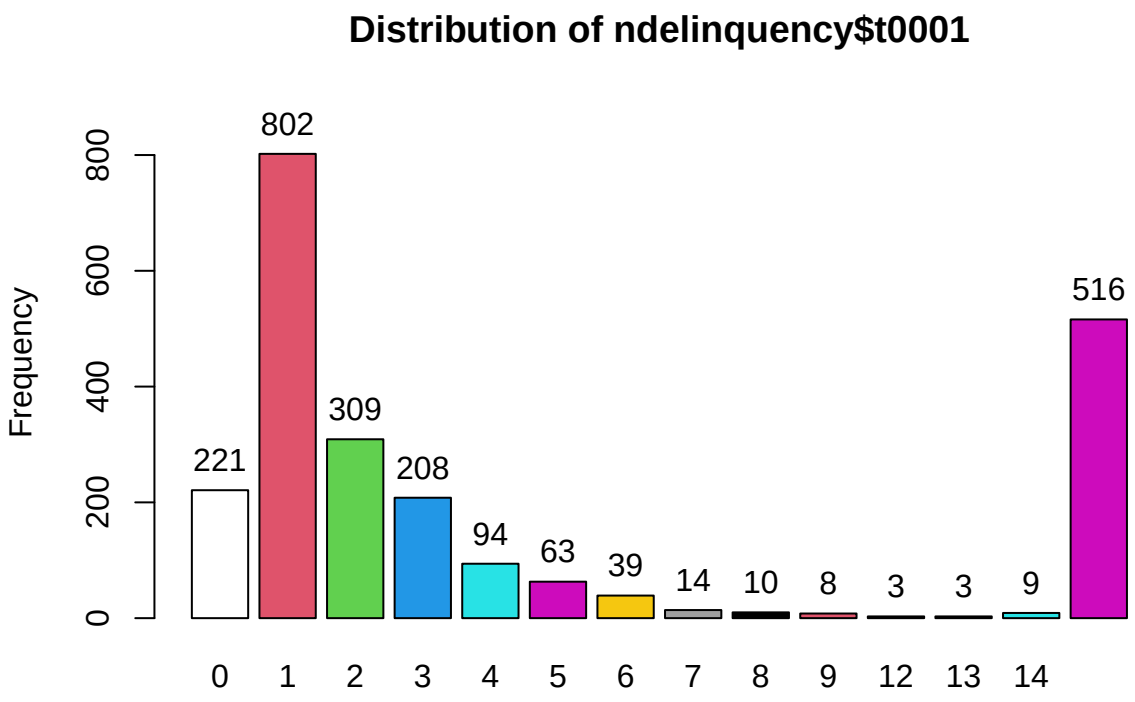
V2287: During the LAST 12 MONTHS, how often have you . . . taken something from a store without paying for it?

However, picking each one of them out from the aggregate variable won't change the key features of the distribution, as it's shown below:

Taking each them out of the aggregate variable does not change the features of the overall distribution.

Without V2281, fighting in school:

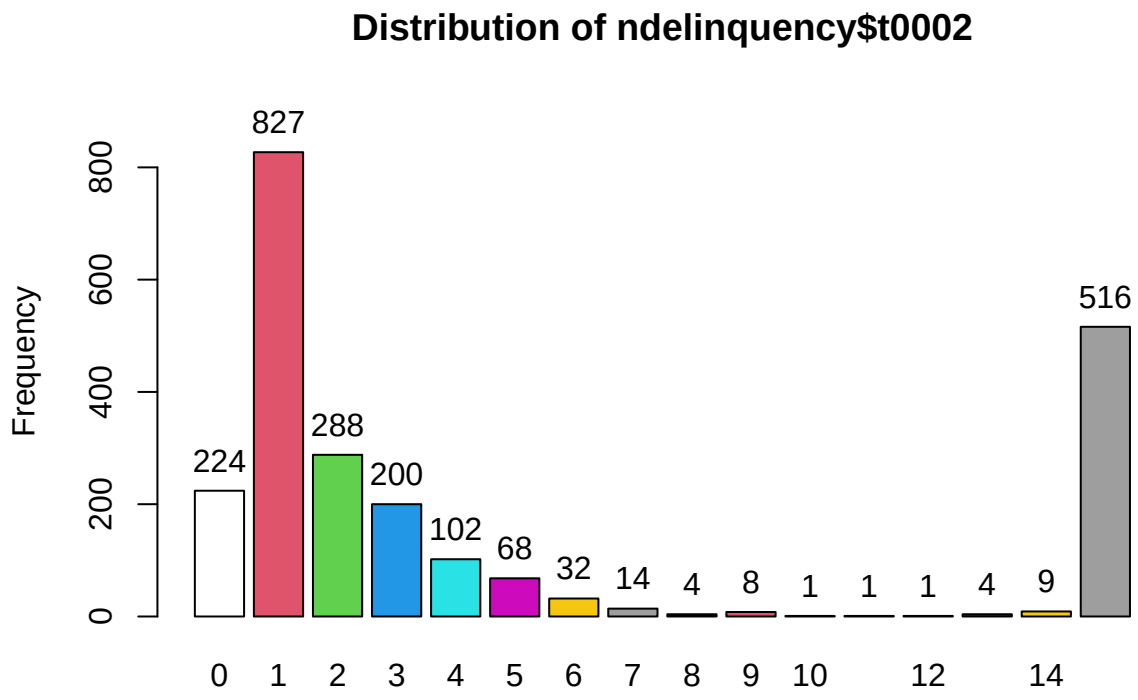
```
ndelinquency$t0001 <- V2279D + V2280D + V2282D + V2283D + V2284D + V2285D + V2286D + V2287D + V2288D + V

tab1(ndelinquency$t0001, cum.percent = TRUE)
```



```
## ndelinquency$t0001 :
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0              221      9.6        9.6     12.4       12.4
## 1              802     34.9       44.5     45.0       57.4
## 2              309     13.4       57.9     17.3       74.7
## 3              208      9.0       67.0     11.7       86.4
## 4               94      4.1       71.1      5.3       91.6
## 5               63      2.7       73.8      3.5       95.2
## 6               39      1.7       75.5      2.2       97.4
## 7               14      0.6       76.1      0.8       98.1
## 8               10      0.4       76.6      0.6       98.7
## 9                8      0.3       76.9      0.4       99.2
## 12               3      0.1       77.0      0.2       99.3
## 13               3      0.1       77.2      0.2       99.5
## 14               9      0.4       77.6      0.5      100.0
## <NA>           516     22.4      100.0      0.0      100.0
##   Total       2299    100.0      100.0    100.0      100.0
```

Without V2282, group fight:

```
ndelinquency$t0002 <- V2279D + V2280D + V2281D + V2283D + V2284D + V2285D + V2286D + V2287D + V2288D + V

tab1(ndelinquency$t0002, cum.percent = TRUE)
```

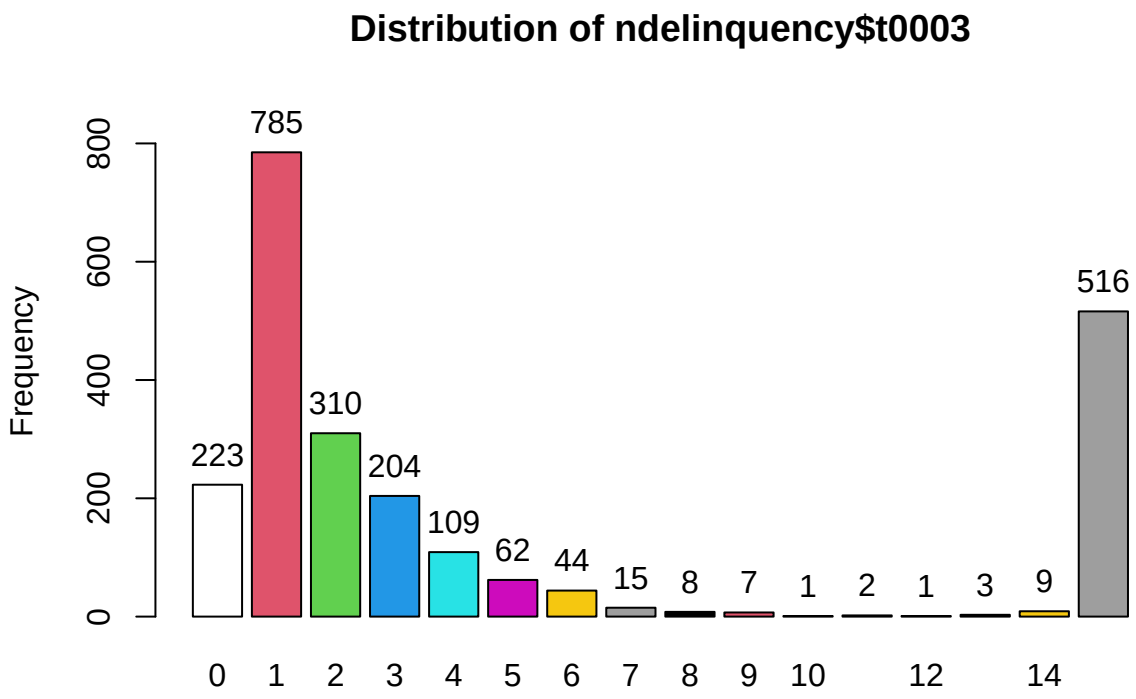



```
## ndelinquency$t0002 :
```

```
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0              224      9.7        9.7     12.6       12.6
## 1              827     36.0       45.7     46.4       58.9
## 2              288     12.5       58.2     16.2       75.1
## 3              200      8.7       66.9     11.2       86.3
## 4              102      4.4       71.4      5.7       92.0
## 5               68      3.0       74.3      3.8       95.8
## 6               32      1.4       75.7      1.8       97.6
## 7               14      0.6       76.3      0.8       98.4
## 8                4      0.2       76.5      0.2       98.7
## 9                8      0.3       76.9      0.4       99.1
## 10               1      0.0       76.9      0.1       99.2
## 11               1      0.0       76.9      0.1       99.2
## 12               1      0.0       77.0      0.1       99.3
## 13               4      0.2       77.2      0.2       99.5
## 14               9      0.4       77.6      0.5      100.0
## <NA>           516     22.4      100.0      0.0      100.0
##   Total       2299    100.0      100.0    100.0      100.0
```

Without V2283, hut someone:

```
ndelinquency$t0003 <- V2279D + V2280D + V2281D + V2282D + V2284D + V2285D + V2286D + V2287D + V2288D + V

tab1(ndelinquency$t0003, cum.percent = TRUE)
```



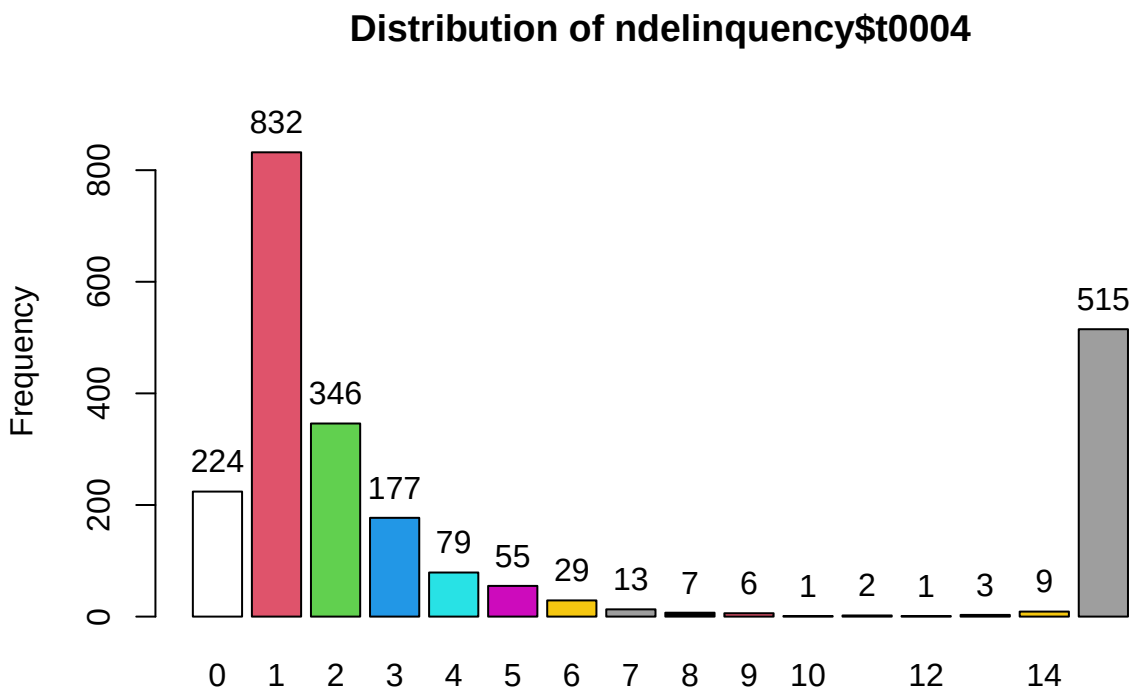

```
## ndelinquency$t0003 :
```

```
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0              223      9.7        9.7     12.5       12.5
## 1              785     34.1       43.8     44.0       56.5
## 2              310     13.5       57.3     17.4       73.9
## 3              204      8.9       66.2     11.4       85.4
## 4              109      4.7       70.9      6.1       91.5
## 5               62      2.7       73.6      3.5       95.0
## 6               44      1.9       75.6      2.5       97.4
## 7               15      0.7       76.2      0.8       98.3
## 8                8      0.3       76.6      0.4       98.7
## 9                7      0.3       76.9      0.4       99.1
## 10               1      0.0       76.9      0.1       99.2
## 11               2      0.1       77.0      0.1       99.3
## 12               1      0.0       77.0      0.1       99.3
## 13               3      0.1       77.2      0.2       99.5
## 14               9      0.4       77.6      0.5      100.0
## <NA>           516     22.4      100.0      0.0      100.0
##   Total       2299    100.0      100.0    100.0      100.0
```

Without V2285, take something under 50

```
ndelinquency$t0004 <- V2279D + V2280D + V2281D + V2282D + V2283D + V2284D + V2286D + V2287D + V2288D + V
```

```
tab1(ndelinquency$t0004, cum.percent = TRUE)
```



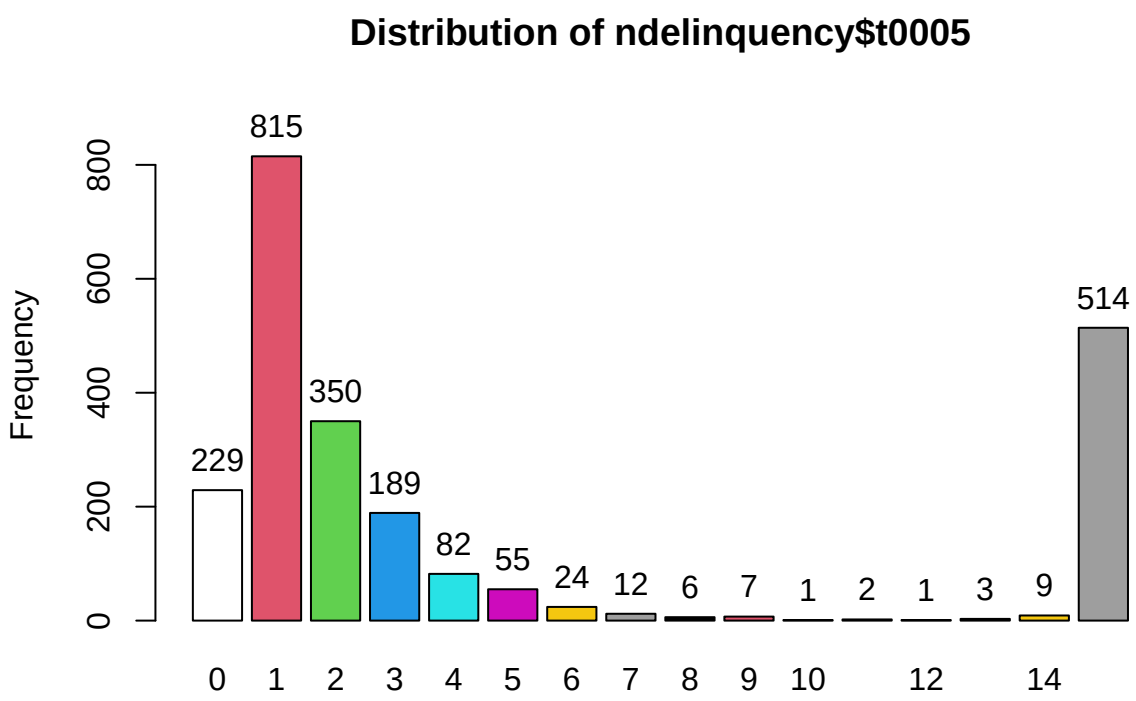

```
## ndelinquency$t0004 :
```

```
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0              224      9.7        9.7     12.6       12.6
## 1              832     36.2       45.9     46.6       59.2
## 2              346     15.1       61.0     19.4       78.6
## 3              177      7.7       68.7      9.9       88.5
## 4               79      3.4       72.1      4.4       92.9
## 5               55      2.4       74.5      3.1       96.0
## 6               29      1.3       75.8      1.6       97.6
## 7               13      0.6       76.3      0.7       98.4
## 8                7      0.3       76.6      0.4       98.8
## 9                6      0.3       76.9      0.3       99.1
## 10               1      0.0       76.9      0.1       99.2
## 11               2      0.1       77.0      0.1       99.3
## 12               1      0.0       77.1      0.1       99.3
## 13               3      0.1       77.2      0.2       99.5
## 14               9      0.4       77.6      0.5      100.0
## <NA>           515     22.4      100.0      0.0      100.0
##   Total       2299    100.0      100.0    100.0      100.0
```

Without V2287, take something from store without paying

```
ndelinquency$t0005 <- V2279D + V2280D + V2281D + V2282D + V2283D + V2284D + V2285D + V2286D + V2288D + V
```

```
tab1(ndelinquency$t0005, cum.percent = TRUE)
```

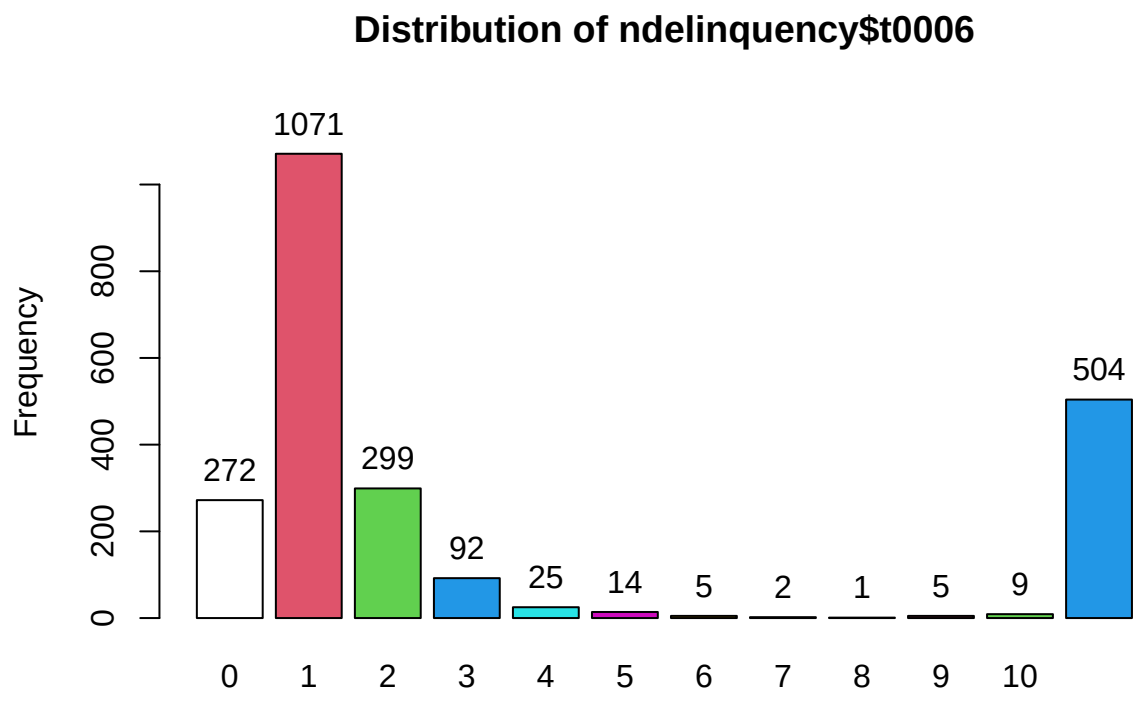



```
## ndelinquency$t0005 :
```

```
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0              229     10.0       10.0     12.8       12.8
## 1              815     35.5       45.4     45.7       58.5
## 2              350     15.2       60.6     19.6       78.1
## 3              189      8.2       68.9     10.6       88.7
## 4               82      3.6       72.4      4.6       93.3
## 5               55      2.4       74.8      3.1       96.4
## 6               24      1.0       75.9      1.3       97.7
## 7               12      0.5       76.4      0.7       98.4
## 8                6      0.3       76.6      0.3       98.7
## 9                7      0.3       76.9      0.4       99.1
## 10               1      0.0       77.0      0.1       99.2
## 11               2      0.1       77.1      0.1       99.3
## 12               1      0.0       77.1      0.1       99.3
## 13               3      0.1       77.3      0.2       99.5
## 14               9      0.4       77.6      0.5      100.0
## <NA>           514     22.4      100.0      0.0      100.0
##   Total       2299    100.0      100.0    100.0      100.0
```

Next step is to try taking all five variables (V2281, V2282, V2283, V2285, V2287) and see if this would change the distribution:

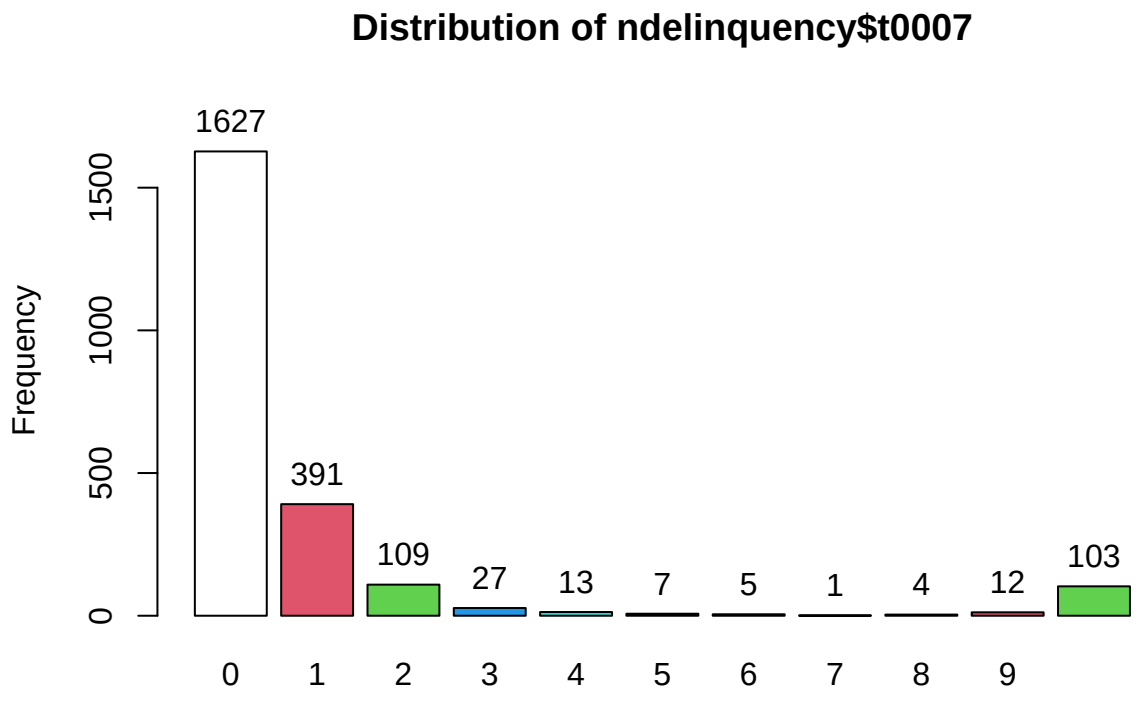
```
ndelinquency$t0006 <- V2279D + V2280D + V2284D + V2286D + V2288D + V2289D + V2290D + V2291D + V2292D + V

tab1(ndelinquency$t0006, cum.percent = TRUE)
```



```
## ndelinquency$t0006 :
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0              272     11.8       11.8     15.2       15.2
## 1             1071     46.6       58.4     59.7       74.8
## 2              299     13.0       71.4     16.7       91.5
## 3               92      4.0       75.4      5.1       96.6
## 4               25      1.1       76.5      1.4       98.0
## 5               14      0.6       77.1      0.8       98.8
## 6                5      0.2       77.3      0.3       99.1
## 7                2      0.1       77.4      0.1       99.2
## 8                1      0.0       77.5      0.1       99.2
## 9                5      0.2       77.7      0.3       99.5
## 10               9      0.4       78.1      0.5      100.0
## <NA>           504     21.9      100.0      0.0      100.0
##   Total       2299    100.0      100.0    100.0      100.0
```

And if we take all five variables (V2281, V2282, V2283, V2285, V2287) plus V2279 (fighting with parents):

```
ndelinquency$t0007 <- V2280D + V2284D + V2286D + V2288D + V2289D + V2290D + V2291D + V2292D + V2293D

tab1(ndelinquency$t0007, cum.percent = TRUE)
```




```
## ndelinquency$t0007 :
##          Frequency   %(NA+) cum.%(NA+)   %(NA-) cum.%(NA-)
## 0             1627     70.8       70.8     74.1       74.1
```

```
## 1             391     17.0       87.8     17.8       91.9
## 2             109      4.7       92.5      5.0       96.9
## 3              27      1.2       93.7      1.2       98.1
## 4              13      0.6       94.3      0.6       98.7
## 5               7      0.3       94.6      0.3       99.0
## 6               5      0.2       94.8      0.2       99.2
## 7               1      0.0       94.8      0.0       99.3
## 8               4      0.2       95.0      0.2       99.5
## 9              12      0.5       95.5      0.5      100.0
## <NA>          103      4.5      100.0      0.0      100.0
##   Total      2299    100.0      100.0    100.0      100.0
```

Only V2279 has the "power" to alter the distribution from a peak at 1 to a peak at 0, and it would also eliminate most missing data.

```
detach(ndelinquency)
```