# Rough Draft

*Joe Comer, Mike McCormack, David Duffrin*

## Contents



```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(leaps)

data <- read.csv('./data/pu_ssocs10.csv', sep='\t')
```

# 1 Data Exploration and examination

Our data comes from the National Center for Education Statistics' 2009-2010 School Survey on Crime and Safety. The data is accompanied by a thorough writes-up on initial treatment of the data.

First we pulled in the full CSV of all 400 variables. Many of the columns are simply imputation flags for previous columns. A great deal of imputation has been done on the data, and entries of -1 were entered for questions left unanswered by a school. We examined the data to see which, if any columns were too incomplete to be useful.

```
cols <- c()
nas <- c()
```

```r
for (col in colnames(data)) {
  perc_na <- sum(data[,col] %in% c(-2, -1, NA)) / nrow(data)
  cols <- c(cols, col)
  nas <- c(nas, perc_na)
}

data_na <- data.frame(cols, nas)
colnames(data_na) <- c('col', 'perc_na')
data_na[data_na$perc_na == 0, 'col'][2:153]
```

```
##   [1] SCHID     C0110     C0112     C0114     C0116     C0120     C0122
##   [8] C0124     C0126     C0128     C0130     C0132     C0134     C0136
##  [15] C0138     C0140     C0141     C0142     C0143     C0144     C0146
##  [22] C0148     C0150     C0151     C0153     C0154     C0158     C0162
##  [29] C0166     C0169     C0170     C0171     C0173     C0174     C0176
##  [36] C0178     C0180     C0181     C0182     C0184     C0186     C0190
##  [43] C0192     C0194     C0196     C0198     C0200     C0202     C0204
##  [50] C0206     C0208     C0210     C0212     C0214     C0216     C0218
##  [57] C0220     C0266     C0268     C0269     C0270     C0272     C0274
##  [64] C0276     C0277     C0280     C0282     C0284     C0286     C0288
##  [71] C0290     C0292     C0294     C0296     C0298     C0300     C0302
##  [78] C0304     C0306     C0308     C0374     C0376     C0378     C0379
##  [85] C0380     C0382     C0384     C0386     C0388     C0389     C0390
##  [92] C0391     C0393     C0394     C0398     C0402     C0406     C0410
##  [99] C0414     C0418     C0422     C0426     C0430     C0434     C0438
## [106] C0442     C0446     C0450     C0454     C0518     C0520     C0526
## [113] C0528     C0532     C0534     C0536     C0538     C0560     C0562
## [120] C0568     C0570     C0572     C0578_YY  CRISIS10  DISTOT10  INCID10
## [127] INCPOL10  OTHACT10  OUTSUS10  PROBWK10  REMOVL10  STRATA    STUOFF10
## [134] SVINC10   SVPOL10   TRANSF10  VIOINC10  VIOPOL10  DISFIRE10 DISDRUG10
## [141] DISWEAP10 GANGHATE  DISRUPT   DISATT10  DISALC10  SEC_FT10  SEC_PT10
## [148] FR_LVEL   FR_SIZE   FR_URBAN  PERCWHT   FINALWGT
## 401 Levels: C0014_R C0016_R C0110 C0112 C0114 C0116 C0120 C0122 ... X
```

```r
# names of columns without NAs, removing columns relating to resampling, imputation, and X (index)
```

From the remaining columns, we found a few interesting features to examine more closely. In particular, we are interested in:

1) Can the total number of violent incidences on a campus be predicted by anything? Does training for teachers make a difference in how many cases occur/how many get reported to the police? Is there any association between attendance and crime on campus?
2) Whether or not attendance is a predictor of crime, does crime affect attendance? Does anything else–taking away bus privileges, for example? High crime in the area?
3) Do violence drills (bomb threat drills, shooter drills, etc) have any effect on the occurence of violence on campuses?
4) Does having random drug sniffs affect crime incidents on campus, violent or otherwise?

Each response variable has a number of possible predictor variables. We attempt backward elimination to select the best model. Similarly, the majority of schools had very few incidents reported to the police, but one school reported as many as 1240 in one year.

We grab the relevant columns and rename them for clarity.

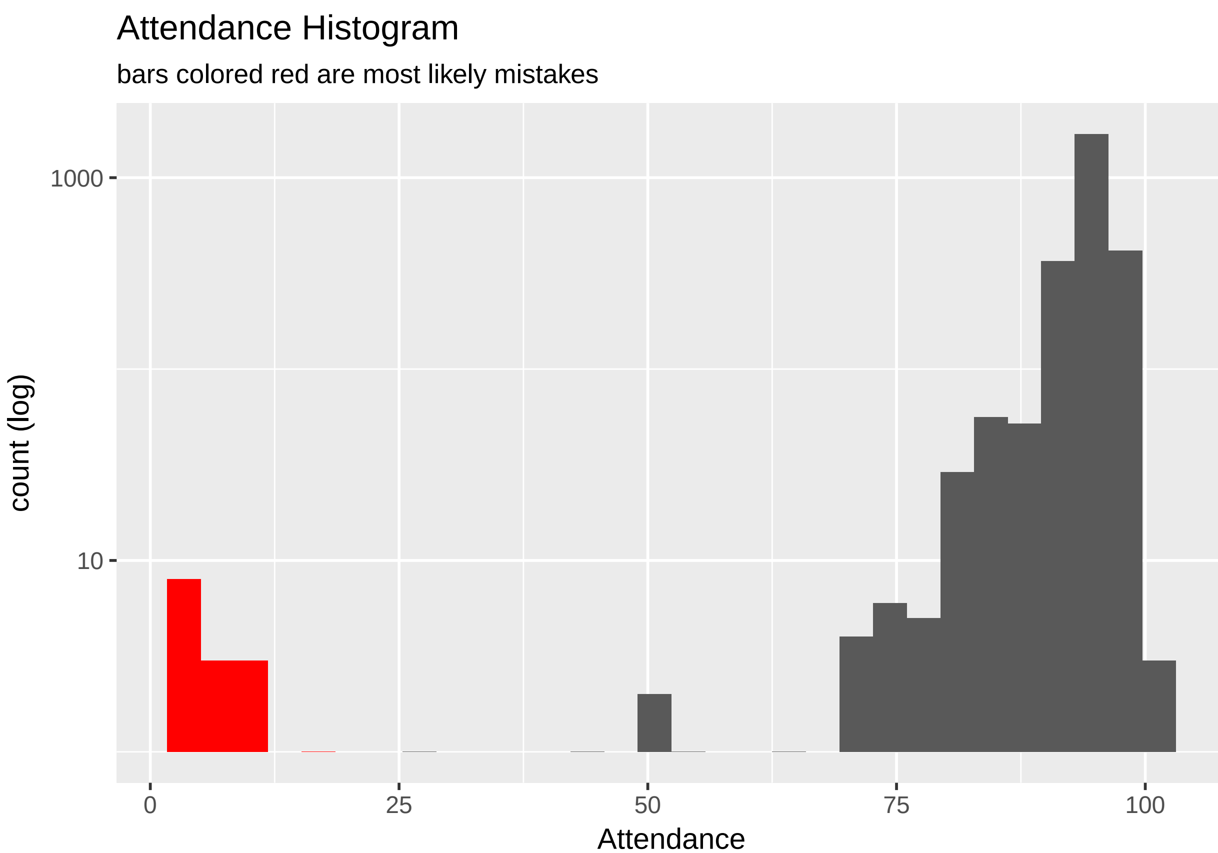
```r
newdata <- select(data,
                  #Predictors
```

```
                    C0568, C0452, C0436, DISRUPT,SVPOL10, C0560,
                    C0538,INCPOL10, GANGHATE, C0294, C0156, C0172,
                    C0124, C0276, C0277, C0272, DISTOT10, DISDRUG10)

newdata <- rename(newdata, c("C0568" = "Attendance","C0452" = "Losspriv",
                   "C0436" =  "Lossbus", "DISRUPT" = "Numdisrpt",
                   "SVPOL10" = "srs_to_police", "C0560" = "Crime_nbhd",
                   "C0538" = "Classchange", "INCPOL10" = "Incpol",
                   "GANGHATE" = "GANGHATE", "C0294" = "Lack_of_funds", "C0156" ="shootingdrills", "C0172
```

In the data description it is mentioned that although a lot of imputation has been done on the data to ensure logical coherence, etc, one column that was left alone was the percent daily attendance. It is further suggested that there is reason to believe some responders may have misinterpretted the question to be about daily percent *absences*, resulting in outlandishly low estimates in some cases.

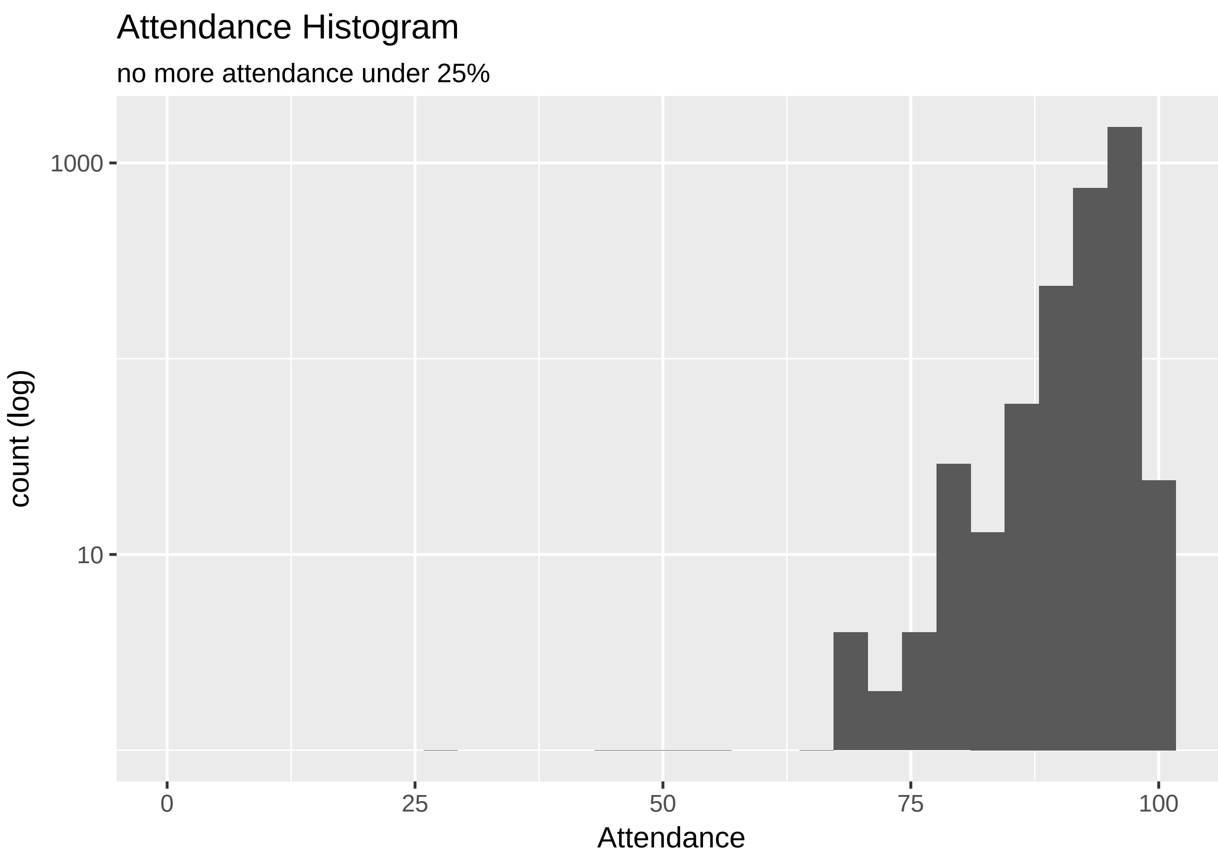
```
ggplot(newdata, aes(Attendance)) +
   geom_histogram(bins=30) +
   geom_histogram(data=subset(newdata,Attendance<25),
   fill="red", bins=30) + scale_y_log10() +
   ggtitle('Attendance Histogram', 'bars colored red are most likely mistakes') + ylab('count (log)')
```




The number of such cases appears to be small, but possibly not insignificant. We will perform our own imputation according to this assumption by imputing x->100-x for reports below 25%. This is a very conservative adjustment, and will probably leave many erroneous reports uncorrected.

```
newdata$Attendance <- sapply(newdata$Attendance, function(x) ifelse(x > 25, x, 100-x))
```

```r
ggplot(newdata, aes(Attendance)) +
   geom_histogram(bins=30) +
   scale_y_log10() + ggtitle('Attendance Histogram', 'no more attendance under 25%') +
   ylab('count (log)') + xlim(0, NA)
```



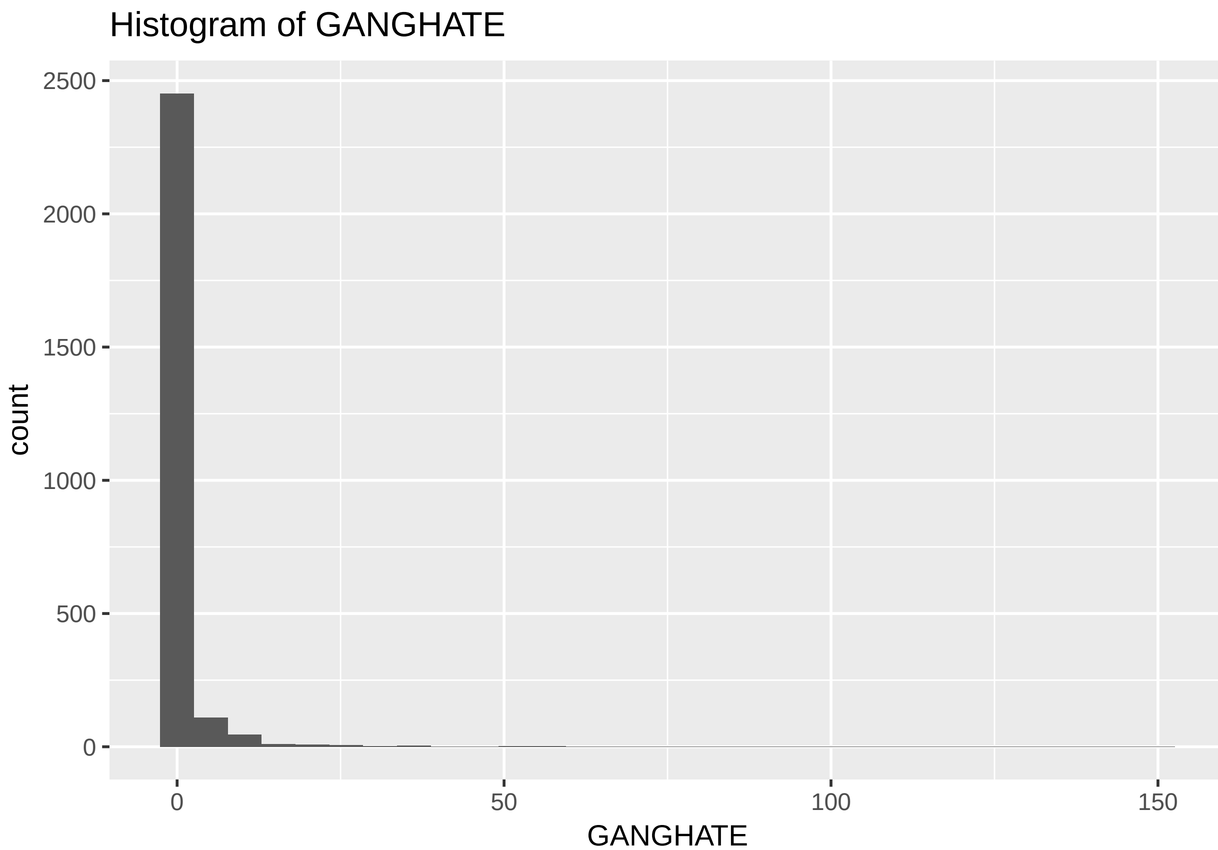

Many of the types of analysis we would like to do are limited by the privacy suppression of the data. Many of the predictor and response variables that might have been continuous/numerical have been binned and made categorical. Many others are discrete counts (like total number of incidents), making linear regression inappropriate, since, among other things, descrete errors cannot be normally distributed. We may consider Poisson regression for some analyses with these variables.

The method of sampling also requires attention: schools were thoroughly stratefied. The pdf gives details of adjustments that need to be made on estimates of various statisticts.

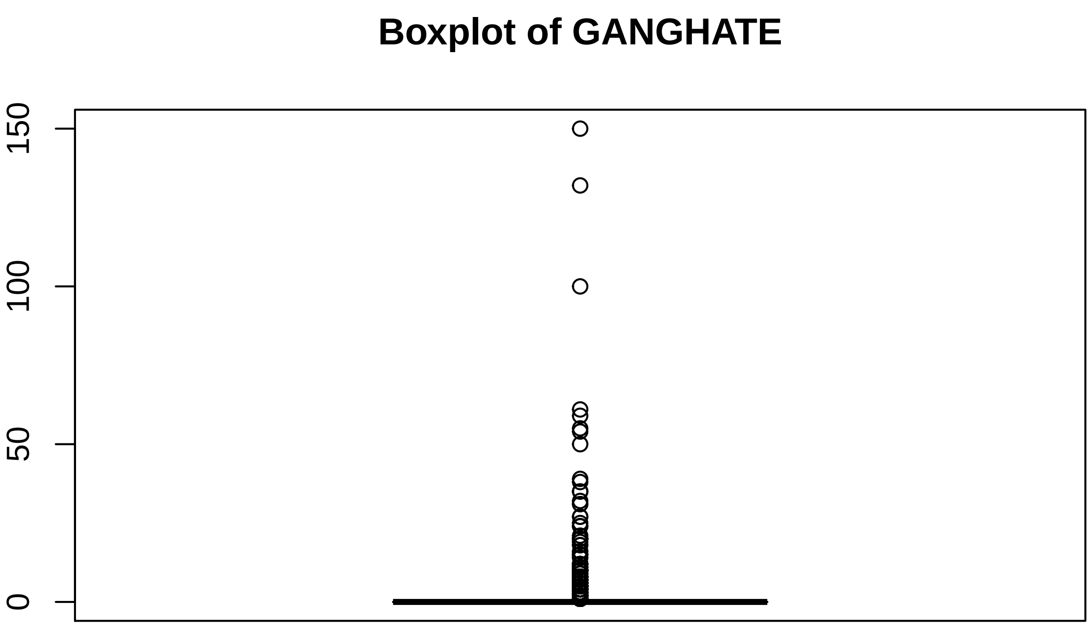
In any case, we can still look at the effect size of various predictors: for example, is the average number of crimes significantly higher in schools that drill for crimes? Is it lower in schools that have violence prevention training for teachers?

Another peculiarity of the data is that much of it is extremely skewed. The vast majority of schools surveyed reported 0 gang related incidents on campus, but many others reported high numbers, making most of the "interesting" cases technically outliers (by the 2.5*IQR standard). Similarly, the majority of schools had very few incidents reported to the police, but one school reported as many as 1240 in one year.

```r
ggplot(newdata, aes(GANGHATE))+geom_histogram(bins=30)+ggtitle('Histogram of GANGHATE')
```

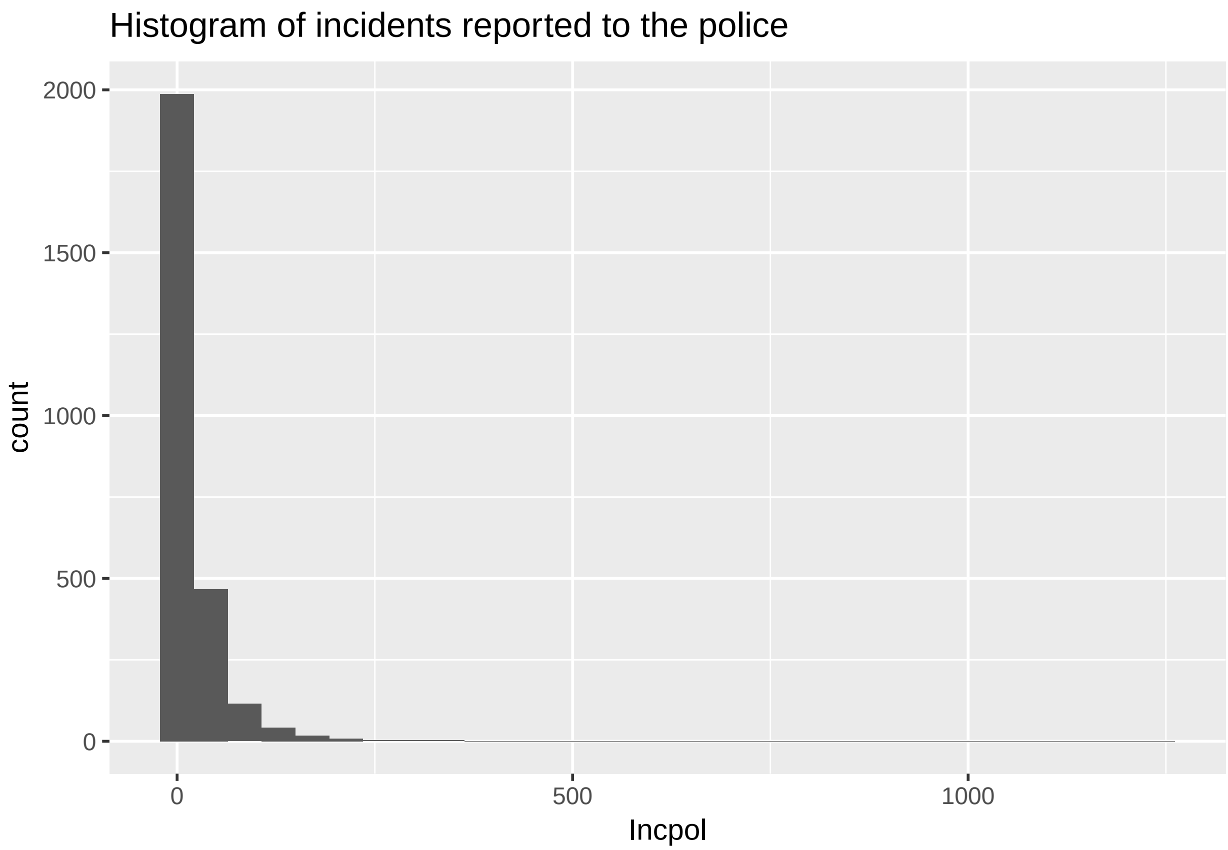



```r
boxplot(newdata$GANGHATE, main='Boxplot of GANGHATE')
```

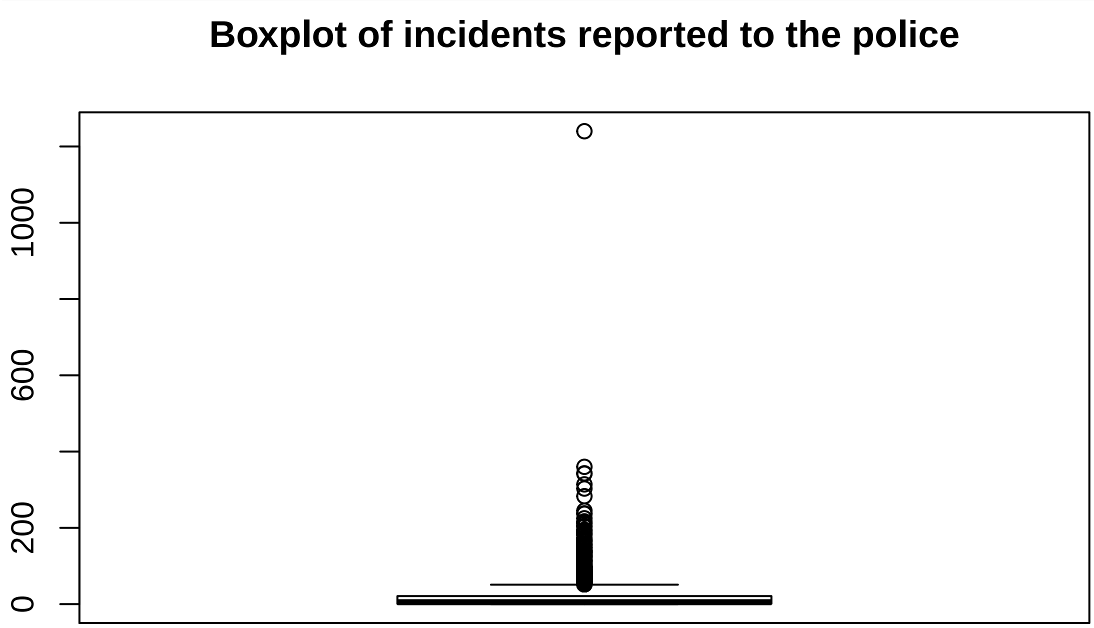



Similarly, the majority of schools had very few incidents reported to the police, but one school reported as many as 1240 in one year.

```r
ggplot(newdata, aes(Incpol))+geom_histogram(bins=30)+ggtitle('Histogram of incidents reported to the pol
```

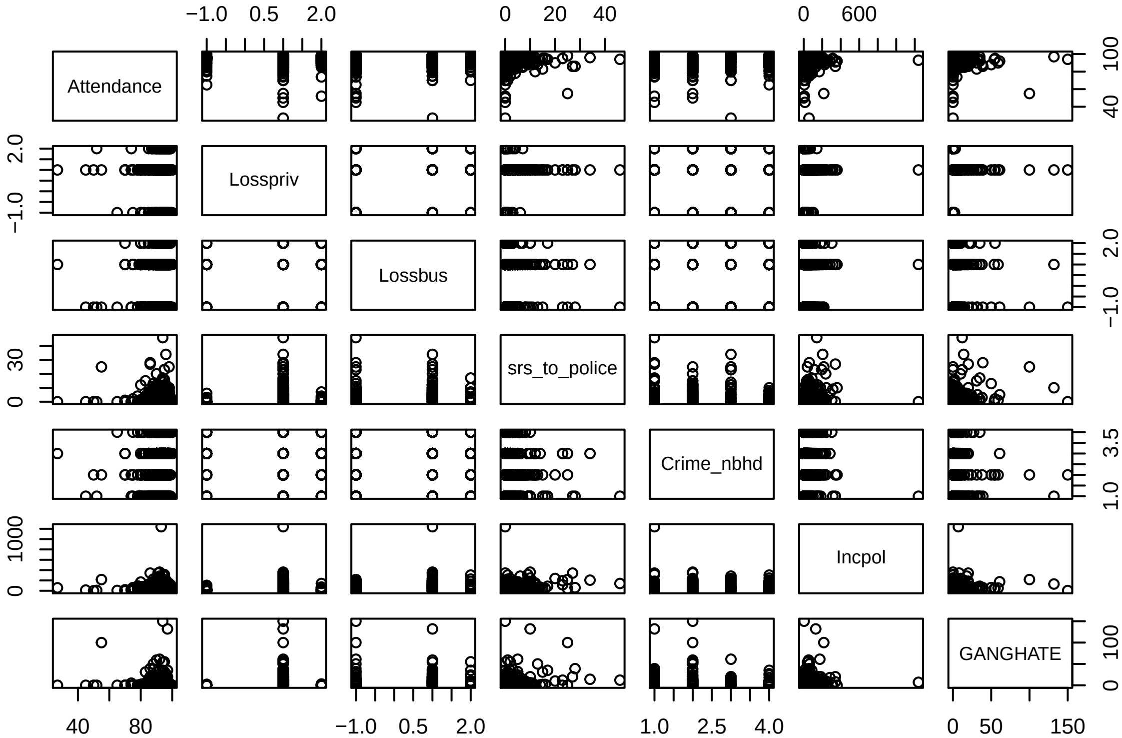



```r
boxplot(newdata$Incpol, main="Boxplot of incidents reported to the police")
```



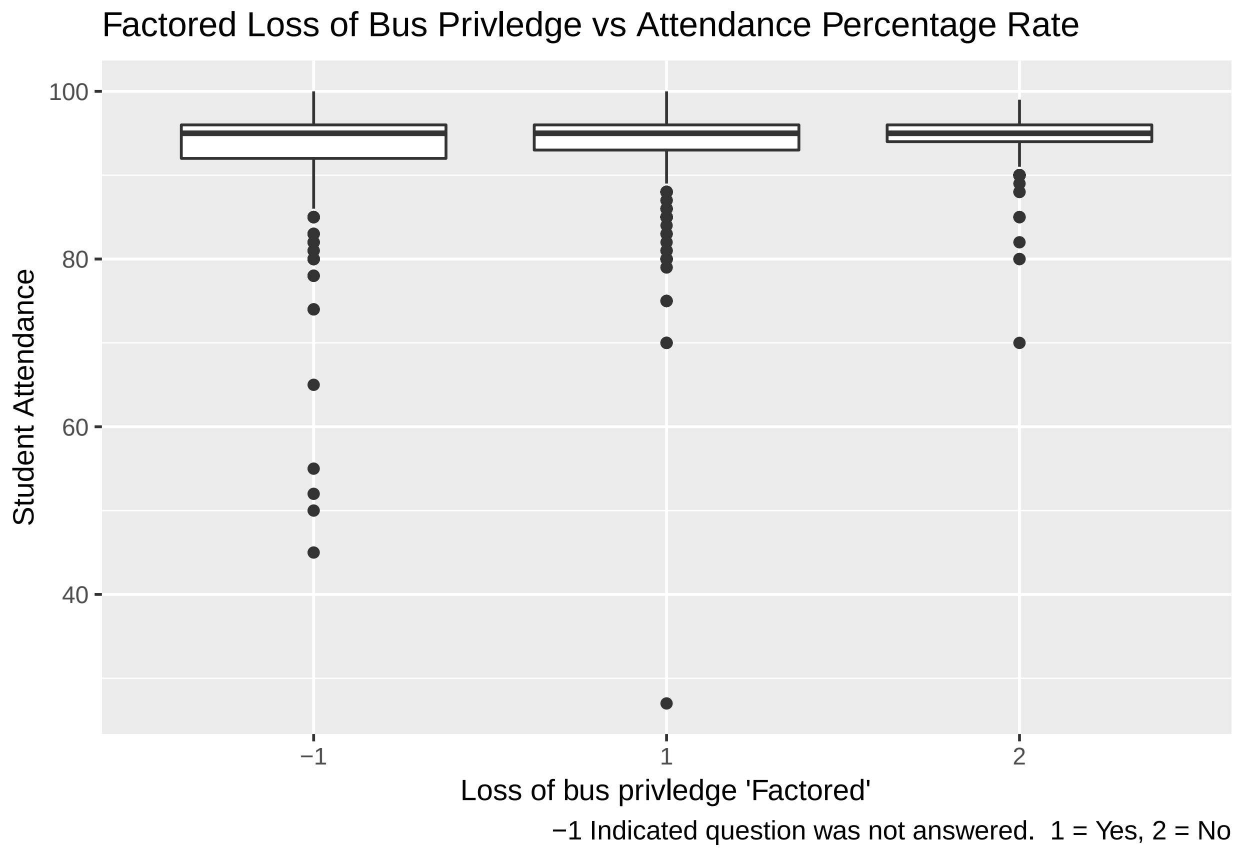

In exploring the question about attendance as a response variable, we examine the scatter plots to check for any patterns.

```
att <- data.frame(newdata$Attendance,newdata$Losspriv,newdata$Lossbus,newdata$srs_to_police,newdata$Cri
colnames(att) <- c("Attendance", "Losspriv", "Lossbus", "srs_to_police", "Crime_nbhd", "Incpol", "GANGH
plot(att)
```



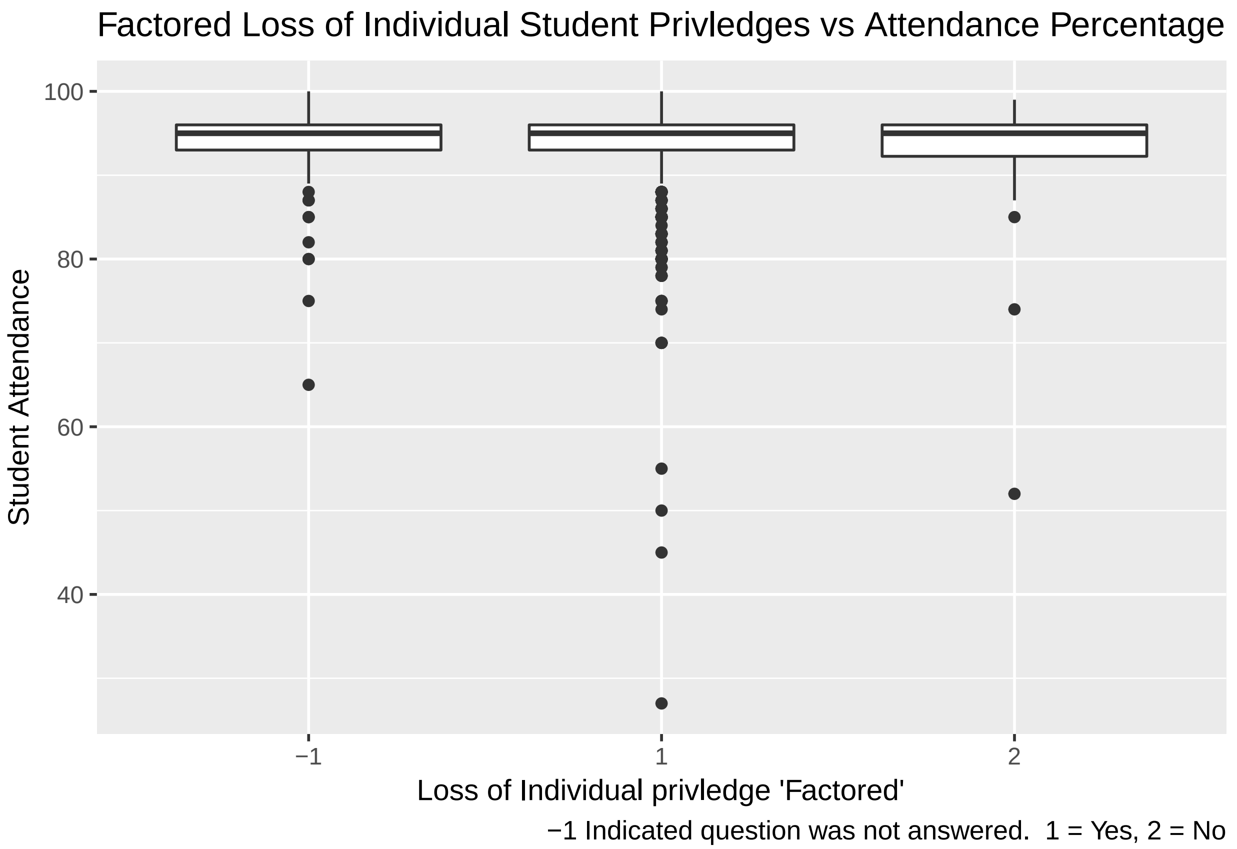

```
newdata %>%
  group_by(Lossbus) %>%
  ggplot(., aes(factor(Lossbus), Attendance)) +
  geom_boxplot() + ggtitle("Factored Loss of Bus Privledge vs Attendance Percentage Rate") +
  xlab("Loss of bus privledge 'Factored'") + ylab("Student Attendance") +
  labs(caption = "-1 Indicated question was not answered.  1 = Yes, 2 = No") +
  ggsave("Factored Loss of Bus Privledge vs Attendance Percentage Rate.png")
```

```
## Saving 6.5 x 4.5 in image
```



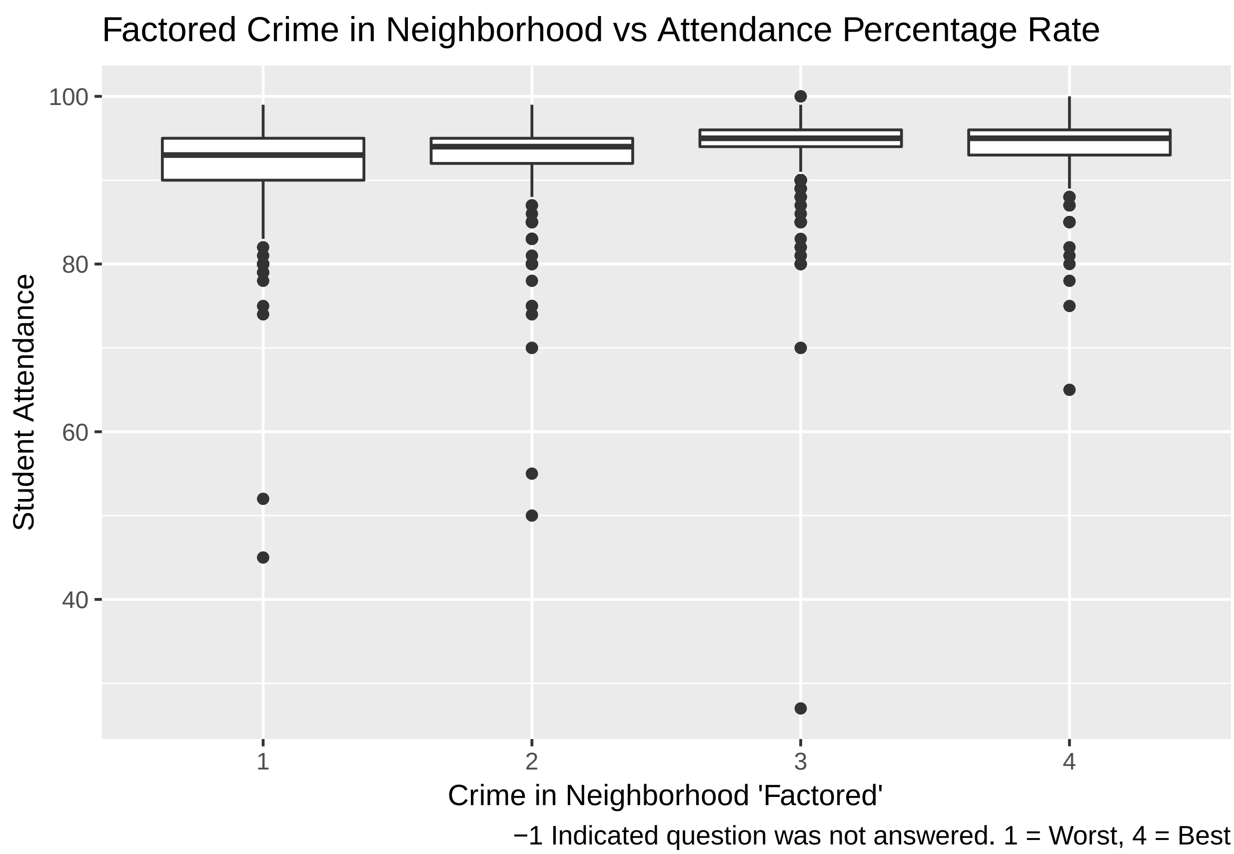

```
#Loss of individual privledges vs Attendance
newdata %>%
  group_by(Losspriv) %>%
  ggplot(., aes(factor(Losspriv), Attendance)) +
  geom_boxplot() + ggtitle("Factored Loss of Individual Student Privledges vs Attendance Percentage Rat
  xlab("Loss of Individual privledge 'Factored'") + ylab("Student Attendance") +
  labs(caption = "-1 Indicated question was not answered.  1 = Yes, 2 = No") +
  ggsave("Factored Loss of Individual Student Privledges vs Attendance Percentage Rate.png")
```

```
## Saving 6.5 x 4.5 in image
```




```r
#Crime in Neighborhood vs Attendance
newdata %>%
  group_by(Crime_nbhd) %>%
  ggplot(., aes(factor(Crime_nbhd), Attendance)) +
  geom_boxplot() + ggtitle("Factored Crime in Neighborhood vs Attendance Percentage Rate") +
  xlab("Crime in Neighborhood 'Factored'") + ylab("Student Attendance") +
  labs(caption = "-1 Indicated question was not answered. 1 = Worst, 4 = Best") +
  ggsave("Factored Crime in Neighborhood vs Attendance Percentage Rate.png")
```

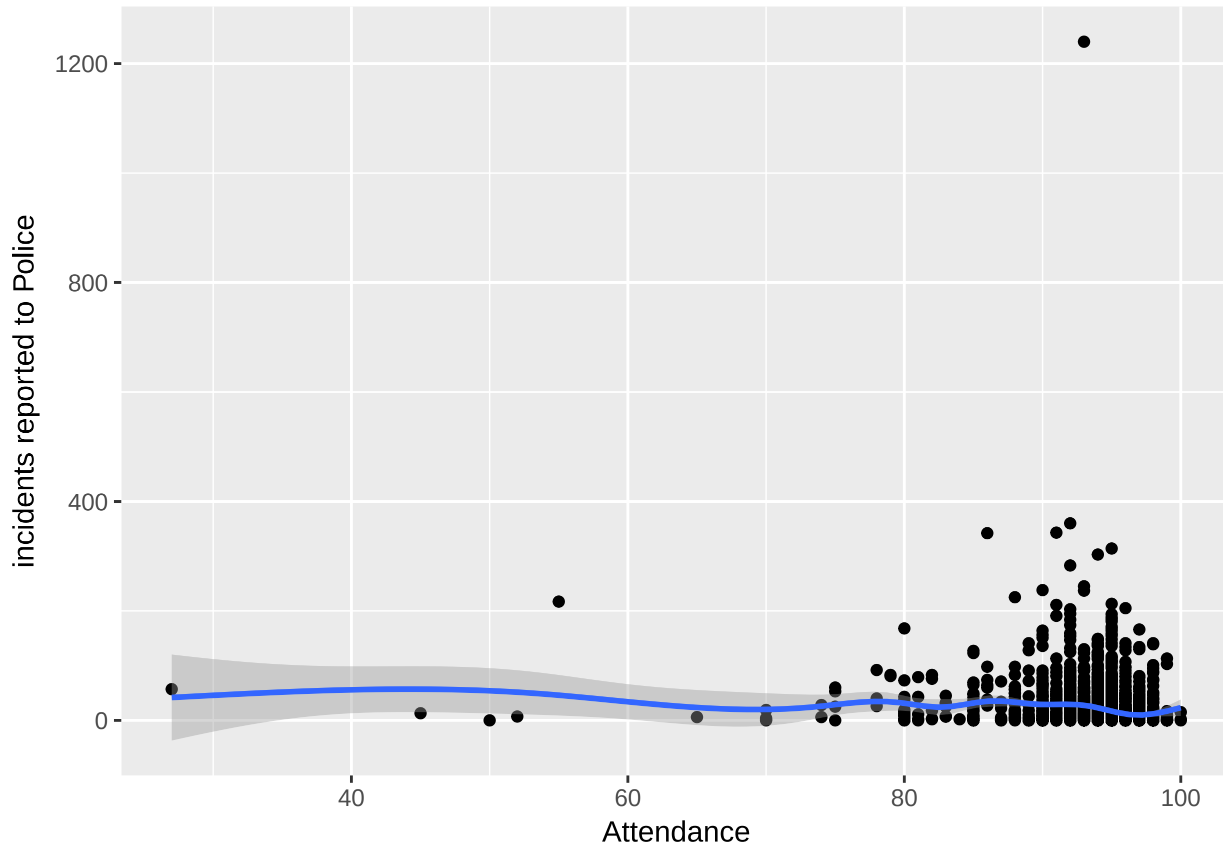
```
## Saving 6.5 x 4.5 in image
```




```r
#incidents reported to Police vs Attendance
qplot(Attendance,Incpol, data = newdata,
      geom = c("point", "smooth")) + ylab("incidents reported to Police") +
  ggsave("Attendance vs Police.png")
```

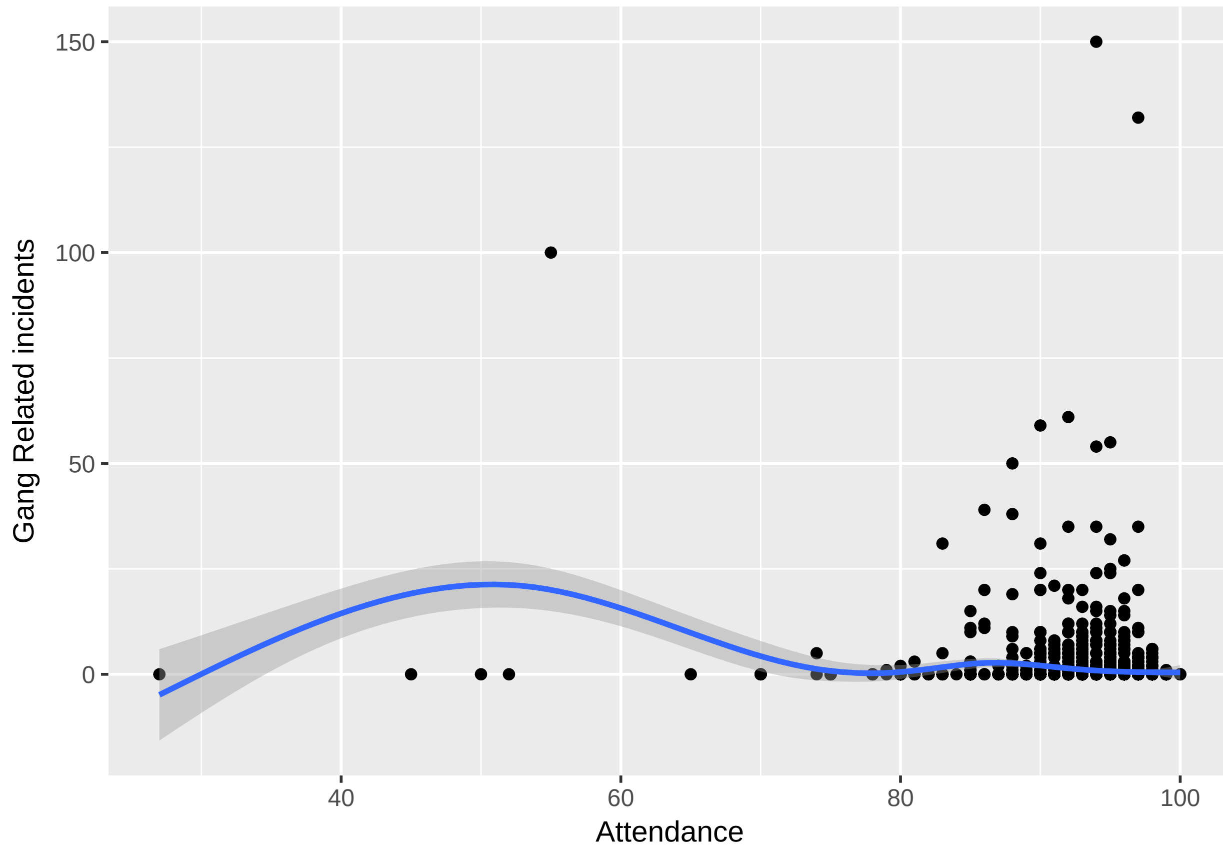
```
## Saving 6.5 x 4.5 in image
```

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

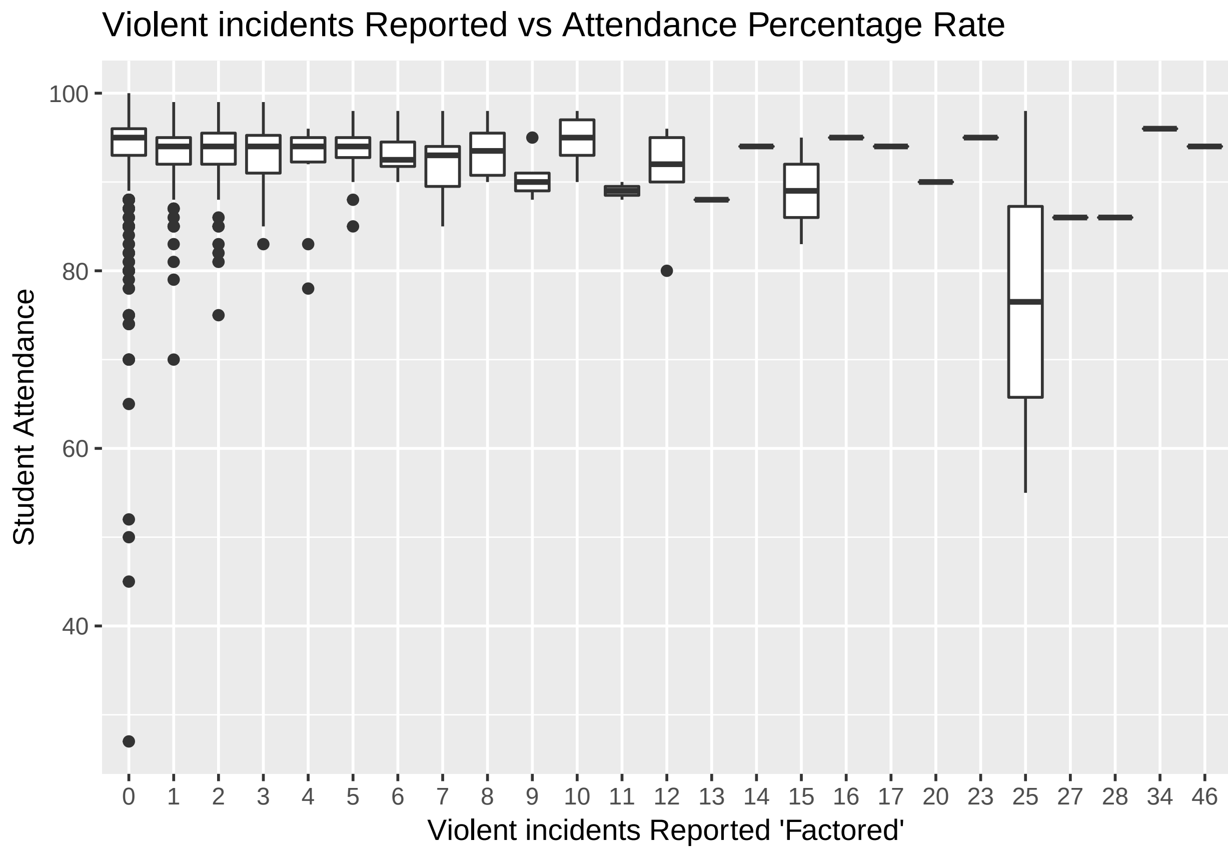



```r
#Gang Related incidents vs Attendance
qplot(Attendance, GANGHATE, data = newdata,
      geom= c("point", "smooth")) + ylab("Gang Related incidents") +
  ggsave("Attendance vs Gang activity.png")
```

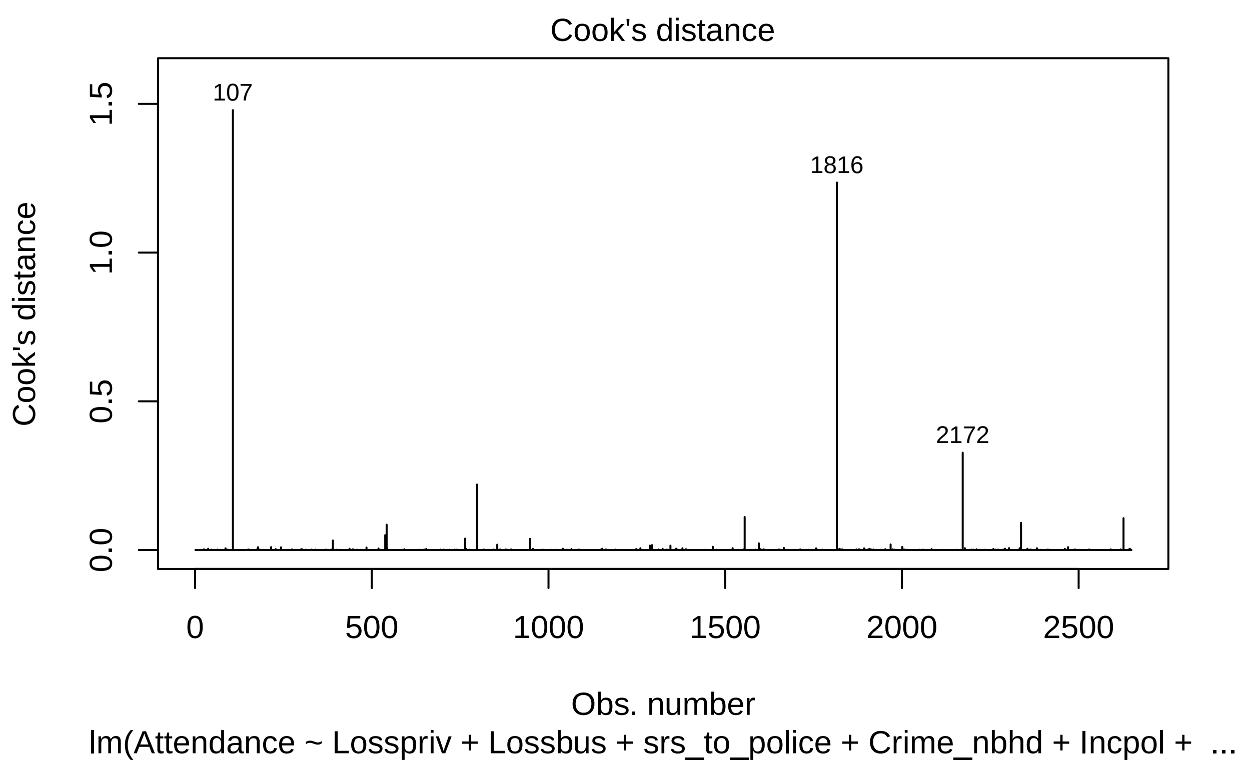
```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using method = 'gam'
## `geom_smooth()` using method = 'gam'
```




```r
#Violent incidents Reported to Police vs Attendance
newdata %>%
  group_by(srs_to_police) %>%
  ggplot(., aes(factor(srs_to_police), Attendance)) +
  geom_boxplot() + ggtitle("Violent incidents Reported vs Attendance Percentage Rate") +
  xlab("Violent incidents Reported 'Factored'") + ylab("Student Attendance")
```

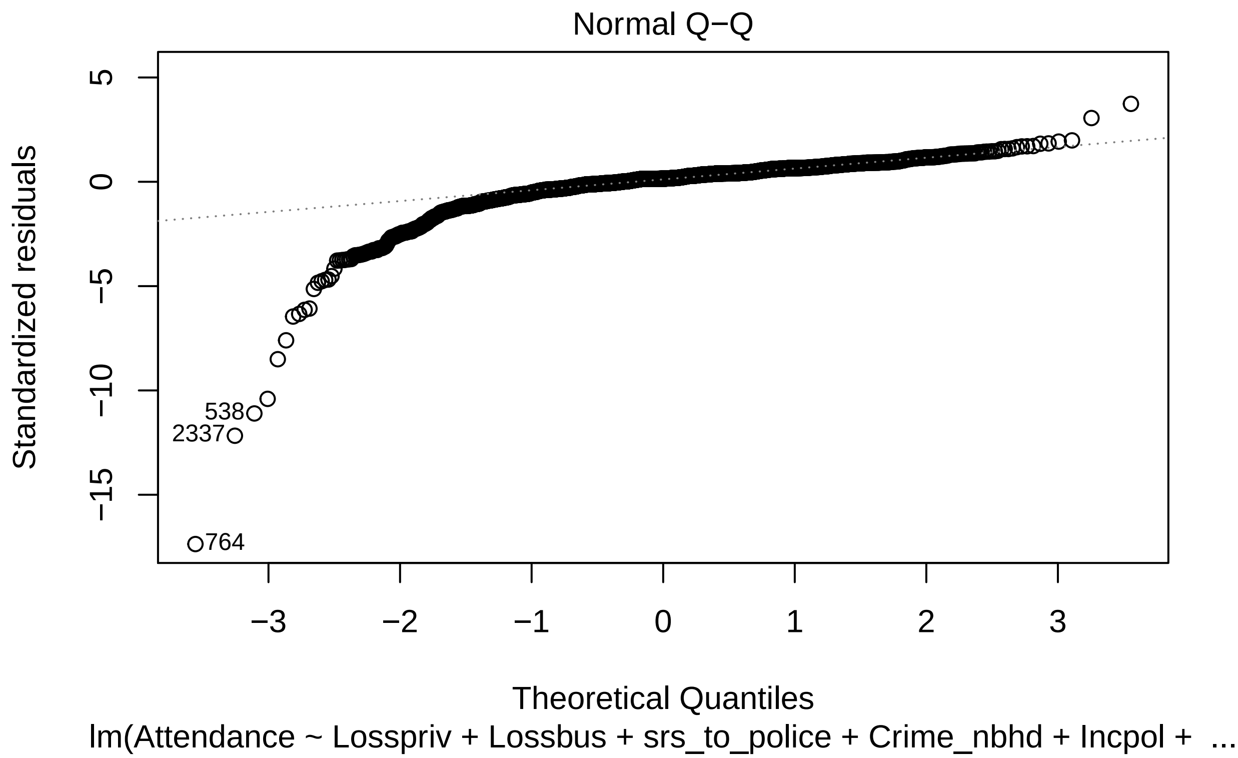



```
att.viol.lm <- lm(Attendance ~ Losspriv + Lossbus + srs_to_police + Crime_nbhd + Incpol + GANGHATE, data
att.viol.summ <- summary(att.viol.lm)
plot(att.viol.lm, which = 4)
```



A look at the Cook's distance for the data in our naive linear model for predicting attendance shows that a number of data points may have outsize influence on the model.

```
plot(att.viol.lm, which = 2)
```




We also have some pretty heavy tails.

```
att.viol.summ
```

```
##
## Call:
## lm(formula = Attendance ~ Losspriv + Lossbus + srs_to_police +
##     Crime_nbhd + Incpol + GANGHATE, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.967  -0.908   0.583   1.790  11.326
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   91.542711   0.335982 272.463  < 2e-16 ***
## Losspriv       0.081532   0.176049   0.463   0.6433
## Lossbus        0.442743   0.095034   4.659 3.34e-06 ***
## srs_to_police -0.067424   0.036924  -1.826   0.0680 .
## Crime_nbhd     0.801016   0.098284   8.150 5.56e-16 ***
## Incpol        -0.008830   0.001955  -4.516 6.58e-06 ***
## GANGHATE      -0.034937   0.013715  -2.547   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.859 on 2641 degrees of freedom
## Multiple R-squared:  0.06117,    Adjusted R-squared:  0.05904
```

```
## F-statistic: 28.68 on 6 and 2641 DF,  p-value: < 2.2e-16
```

In any event, although some of the predictors have low p-values, the R-squared is terrible, and all of the coefficients are so tiny that any effect is unlikely to have practical significance, should they turn out to in fact have statistical significance. Also, loss of bus privileges appears to correlate *positively* with increased attendance, according to this model, which is unexpected.

Before trying a reduced model, we take a side-track to test the hypothesis that taking away bus privileges as a form of punishment affects mean attendance.

```
all.bus <- data.frame(matrix(c(newdata$Attendance,newdata$Lossbus),ncol=2))
colnames(all.bus) <- c("Attendance", "Lossbus")
all.bus <- all.bus[all.bus$Lossbus != -1,]

t.test(Attendance ~ Lossbus, paired = FALSE, var.equal = FALSE, data = all.bus)
```

```
##
##  Welch Two Sample t-test
##
## data:  Attendance by Lossbus
## t = -2.6606, df = 401.37, p-value = 0.008112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9099333 -0.1366427
## sample estimates:
## mean in group 1 mean in group 2
##        94.08917        94.61246
```

The ninety-five percent confidence interval for the difference in group means includes zero, so we conclude that there is not enough information to reject the null hypothesis that taking away bus privileges as a form of punishment has no effect on attendance.

Before moving on to eliminating variables, we try the same model again with severe outliers removed to see if we get any improvement.

```
##Finding and dealing with outliers
#scale(newdata)
outdet <- function(x) abs(scale(x)) >= 3
newdata1 <- newdata[!apply(sapply(newdata, outdet), 1, any), ]

att.viol.lm.clean <- lm(Attendance ~ Losspriv + Lossbus + srs_to_police + Crime_nbhd + Incpol + GANGHATE
summary(att.viol.lm.clean)
```

```
##
## Call:
## lm(formula = Attendance ~ Losspriv + Lossbus + srs_to_police +
##     Crime_nbhd + Incpol + GANGHATE, data = newdata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7342  -0.8511   0.2825   1.4175   6.9983
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   93.493024   0.383919 243.523  < 2e-16 ***
## Losspriv      -0.149147   0.308682  -0.483  0.62902
## Lossbus        0.061462   0.067333   0.913  0.36144
```

```
## srs_to_police -0.160407   0.060066  -2.670  0.00763 **
## Crime_nbhd     0.481912   0.070587   6.827  1.1e-11 ***
## Incpol        -0.016692   0.002564  -6.511  9.1e-11 ***
## GANGHATE      -0.091854   0.032962  -2.787  0.00537 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.481 on 2322 degrees of freedom
## Multiple R-squared:  0.06399,    Adjusted R-squared:  0.06157
## F-statistic: 26.46 on 6 and 2322 DF,  p-value: < 2.2e-16
```

Not much improvement.

We seek a reduced model. Let's try backwards elimination.

```
att.viol.lm.nopriv <- lm(Attendance ~ Lossbus + srs_to_police + Crime_nbhd + Incpol + GANGHATE, data = n
att.viol.lm.nobus <- lm(Attendance ~ Losspriv + srs_to_police + Crime_nbhd + Incpol + GANGHATE, data = n
att.viol.lm.nopolice <- lm(Attendance ~ Losspriv + Lossbus + Crime_nbhd + Incpol + GANGHATE, data = newd
att.viol.lm.nocrim <- lm(Attendance ~ Losspriv + Lossbus + srs_to_police + Incpol + GANGHATE, data = new
att.viol.lm.nopol <- lm(Attendance ~ Losspriv + Lossbus + srs_to_police + Crime_nbhd + GANGHATE, data =
att.viol.lm.nogang <- lm(Attendance ~ Losspriv + Lossbus + srs_to_police + Crime_nbhd + Incpol, data = n
cat(' Adjusted R-Squared without Losspriv:', summary(att.viol.lm.nopriv)$adj.r.squared, '\n',
'Adjusted R-Squared without Lossbus:', summary(att.viol.lm.nobus)$adj.r.squared, '\n',
'Adjusted R-Squared without srs_to_police:', summary(att.viol.lm.nopolice)$adj.r.squared, '\n',
'Adjusted R-Squared without Crime_nbhd:', summary(att.viol.lm.nocrim)$adj.r.squared, '\n',
'Adjusted R-Squared without Incpol:', summary(att.viol.lm.nopol)$adj.r.squared, '\n',
'Adjusted R-Squared without GANGHATE:', summary(att.viol.lm.nogang)$adj.r.squared)
```

```
##  Adjusted R-Squared without Losspriv: 0.05931855
##  Adjusted R-Squared without Lossbus: 0.0516648
##  Adjusted R-Squared without srs_to_police: 0.0582074
##  Adjusted R-Squared without Crime_nbhd: 0.0357383
##  Adjusted R-Squared without Incpol: 0.05213159
##  Adjusted R-Squared without GANGHATE: 0.05708384
```

Let's further reduce after removing Losspriv (our best submodel)

```
att.viol.lm.nopriv.nobus <- lm(Attendance ~ srs_to_police + Crime_nbhd + Incpol + GANGHATE, data = newda
att.viol.lm.nopriv.nopolice <- lm(Attendance ~ Lossbus + Crime_nbhd + Incpol + GANGHATE, data = newdata)
att.viol.lm.nopriv.nocrim <- lm(Attendance ~ Lossbus + srs_to_police + Incpol + GANGHATE, data = newdata
att.viol.lm.nopriv.nopol <- lm(Attendance ~ Lossbus + srs_to_police + Crime_nbhd + GANGHATE, data = newd
att.viol.lm.nopriv.nogang <- lm(Attendance ~ Lossbus + srs_to_police + Crime_nbhd + Incpol, data = newda
cat(' Adjusted R-Squared without Losspriv and Lossbus:', summary(att.viol.lm.nopriv.nobus)$adj.r.square
'Adjusted R-Squared without Losspriv and srs_to_police:', summary(att.viol.lm.nopriv.nopolice)$adj.r.sq
'Adjusted R-Squared without Losspriv and Crime_nbhd:', summary(att.viol.lm.nopriv.nocrim)$adj.r.squared
'Adjusted R-Squared without Losspriv and Incpol:', summary(att.viol.lm.nopriv.nopol)$adj.r.squared, '\n
'Adjusted R-Squared without Losspriv and GANGHATE:', summary(att.viol.lm.nopriv.nogang)$adj.r.squared)
```

```
##  Adjusted R-Squared without Losspriv and Lossbus: 0.05158284
##  Adjusted R-Squared without Losspriv and srs_to_police: 0.05849761
##  Adjusted R-Squared without Losspriv and Crime_nbhd: 0.03607232
##  Adjusted R-Squared without Losspriv and Incpol: 0.05243599
##  Adjusted R-Squared without Losspriv and GANGHATE: 0.05737685
```

We are now excluding both Losspriv and srs_to_police, let's see if we can reduce further.

```
att.viol.lm.nopriv.nopolice.nobus <- lm(Attendance ~ Crime_nbhd + Incpol + GANGHATE, data = newdata)
att.viol.lm.nopriv.nopolice.nocrim <- lm(Attendance ~ Lossbus + Incpol + GANGHATE, data = newdata)
```

```r
att.viol.lm.nopriv.nopolice.nopol <- lm(Attendance ~ Lossbus + Crime_nbhd + GANGHATE, data = newdata)
att.viol.lm.nopriv.nopolice.nogang <- lm(Attendance ~ Lossbus + Crime_nbhd + Incpol, data = newdata)
cat(' Adjusted R-Squared without Losspriv, srs_to_police, and Lossbus:', summary(att.viol.lm.nopriv.nop
'Adjusted R-Squared without Losspriv, srs_to_police, and Crime_nbhd:', summary(att.viol.lm.nopriv.nopol
'Adjusted R-Squared without Losspriv, srs_to_police, and Incpol:', summary(att.viol.lm.nopriv.nopolice.
'Adjusted R-Squared without Losspriv, srs_to_police, and GANGHATE:', summary(att.viol.lm.nopriv.nopolic
```

```
##  Adjusted R-Squared without Losspriv, srs_to_police, and Lossbus: 0.05047294
##  Adjusted R-Squared without Losspriv, srs_to_police, and Crime_nbhd: 0.03438672
##  Adjusted R-Squared without Losspriv, srs_to_police, and Incpol: 0.04866997
##  Adjusted R-Squared without Losspriv, srs_to_police, and GANGHATE: 0.05552353
```

If we really need to reduce the model further, we could also remove GANGHATE. However seeing the Adjusted R-Squared fall by 0.002 makes me want to keep the variable in the model. This means that our best model using backward elimination uses the variables Lossbus, Crime_nbhd, Incpol, and GANGHATE to predict Attendance.

Let's use ANOVA to check if the full model is significantly better than the reduced model.

$H_0$ : The coefficients for all variables in the full model that are not in the reduced model are zero.

$H_a$ : The coefficients are not zero.

$\alpha = 0.05$

```r
anova(att.viol.lm.nopriv.nopolice, att.viol.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Attendance ~ Lossbus + Crime_nbhd + Incpol + GANGHATE
## Model 2: Attendance ~ Losspriv + Lossbus + srs_to_police + Crime_nbhd +
##     Incpol + GANGHATE
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1   2643 39372
## 2   2641 39319  2    52.407 1.76 0.1722
```

We fail to reject the null hypothesis and will use the reduced model.

Let's check to see if this model has the lowest BIC score.

```r
allpossreg <- regsubsets(Attendance ~ .,
                         nbest=6, data=att)

aprout <- summary(allpossreg)

with(aprout,round(cbind(which,rsq,adjr2,cp,bic),3))
```

```
##   (Intercept) Losspriv Lossbus srs_to_police Crime_nbhd Incpol GANGHATE
## 1           1        0       0             0          1      0        0
## 1           1        0       0             0          0      1        0
## 1           1        0       1             0          0      0        0
## 1           1        0       0             1          0      0        0
## 1           1        0       0             0          0      0        1
## 1           1        1       0             0          0      0        0
## 2           1        0       0             0          1      1        0
## 2           1        0       1             0          1      0        0
## 2           1        0       0             1          1      0        0
## 2           1        0       0             0          1      0        1
## 2           1        1       0             0          1      0        0
```

```
## 2          1        0       1              0          0      1        0
## 3          1        0       1              0          1      1        0
## 3          1        0       0              0          1      1        1
## 3          1        0       1              1          1      0        0
## 3          1        0       0              1          1      1        0
## 3          1        0       1              0          1      0        1
## 3          1        1       0              0          1      1        0
## 4          1        0       1              0          1      1        1
## 4          1        0       1              1          1      1        0
## 4          1        1       1              0          1      1        0
## 4          1        0       1              1          1      0        1
## 4          1        0       0              1          1      1        1
## 4          1        1       0              0          1      1        1
## 5          1        0       1              1          1      1        1
## 5          1        1       1              0          1      1        1
## 5          1        1       1              1          1      1        0
## 5          1        1       1              1          1      0        1
## 5          1        1       0              1          1      1        1
## 5          1        1       1              1          0      1        1
## 6          1        1       1              1          1      1        1
##      rsq adjr2      cp      bic
## 1 0.034 0.034  72.411  -76.834
## 1 0.017 0.017 120.934  -29.951
## 1 0.013 0.012 133.205  -18.224
## 1 0.013 0.012 133.606  -17.842
## 1 0.012 0.012 135.054  -16.462
## 1 0.000 0.000 168.521   15.236
## 2 0.048 0.047  37.199 -105.479
## 2 0.043 0.043  49.005  -93.836
## 2 0.042 0.042  51.928  -90.961
## 2 0.042 0.041  54.172  -88.755
## 2 0.035 0.034  73.634  -69.710
## 2 0.030 0.029  87.887  -55.849
## 3 0.057 0.056  13.877 -122.742
## 3 0.052 0.050  28.069 -108.620
## 3 0.051 0.049  30.939 -105.774
## 3 0.050 0.049  31.517 -105.200
## 3 0.050 0.049  33.135 -103.597
## 3 0.048 0.047  38.102  -98.681
## 4 0.060 0.058   6.520 -124.214
## 4 0.059 0.057   9.668 -121.064
## 4 0.057 0.055  15.742 -114.996
## 4 0.054 0.052  23.546 -107.220
## 4 0.053 0.052  25.942 -104.837
## 4 0.052 0.051  28.882 -101.916
## 5 0.061 0.059   5.214 -119.644
## 5 0.060 0.058   8.334 -116.518
## 5 0.059 0.057  11.489 -113.361
## 5 0.054 0.052  25.394  -99.490
## 5 0.053 0.052  26.704  -98.186
## 5 0.038 0.036  71.422  -54.085
## 6 0.061 0.059   7.000 -111.978
```

The model with by far the lowest BIC score includes the same variables that were chosen with backward selection. The R-squareds are all terrible no matter what. The model with the second lowest BIC also removes GANGHATE from the model, which we also would have if we continued with the backward elimination. We conclude that the best model is the one that includes if the school has a punishment of lossing bus privileges, the number of gang-related and hate crimes, the number of incidents reported to the police, and the self-reported rating for incidence of crime in the neighorhood of the school, and this it is not a very good model anyway.

We now move on to the question of whether disaster drills relating to violence affect the mean number of violent incidents. We first check to see if any schools have ommited information about drills and remove them.

```
unique(newdata$shootingdrills)
```

```
## [1]  1 -1  2
```

```
unique(newdata$threatdrills)
```

```
## [1]  2 -1  1
```

```
drills <- data.frame(matrix(c(newdata$srs_to_police, newdata$Incpol, newdata$shootingdrills, newdata$thr
colnames(drills) <- c("srs_to_police", "Incpol", "shootingdrills", "threatdrills")
drills.shoot <- drills[drills$shootingdrills != -1,]

t.test(srs_to_police~ shootingdrills, paired = FALSE, var.equal = FALSE, data = drills.shoot)
```

```
##
##  Welch Two Sample t-test
##
## data:  srs_to_police by shootingdrills
## t = 0.41079, df = 1591.4, p-value = 0.6813
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1612006  0.2466094
## sample estimates:
## mean in group 1 mean in group 2
##       0.5987903       0.5560859
```

Once again, we conclude that there is insufficient evidence to reject the null hypothesis, and that shooter drills do not appear to affect the number of serious violent incidents on campus.

```
drills.threat <- drills[drills$threatdrills != -1,]
t.test(srs_to_police~ threatdrills, paired = FALSE, var.equal = FALSE, data = drills.threat)
```

```
##
##  Welch Two Sample t-test
##
## data:  srs_to_police by threatdrills
## t = 1.0143, df = 1618.1, p-value = 0.3106
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1074112  0.3374705
## sample estimates:
## mean in group 1 mean in group 2
##       0.6557576       0.5407279
```

And again we fail to reject the null hypothesis. It seems that other violence-related drills also do not affect serious crime reports.