152project

Group 5

4/24/2019

```
library(tidyverse)
```

```
## ── Attaching packages ──────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0     ✔ purrr   0.2.5
## ✔ tibble  2.0.1     ✔ dplyr   0.7.8
## ✔ tidyr   0.8.2     ✔ stringr 1.3.1
## ✔ readr   1.3.1     ✔ forcats 0.3.0

## ── Conflicts ─────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

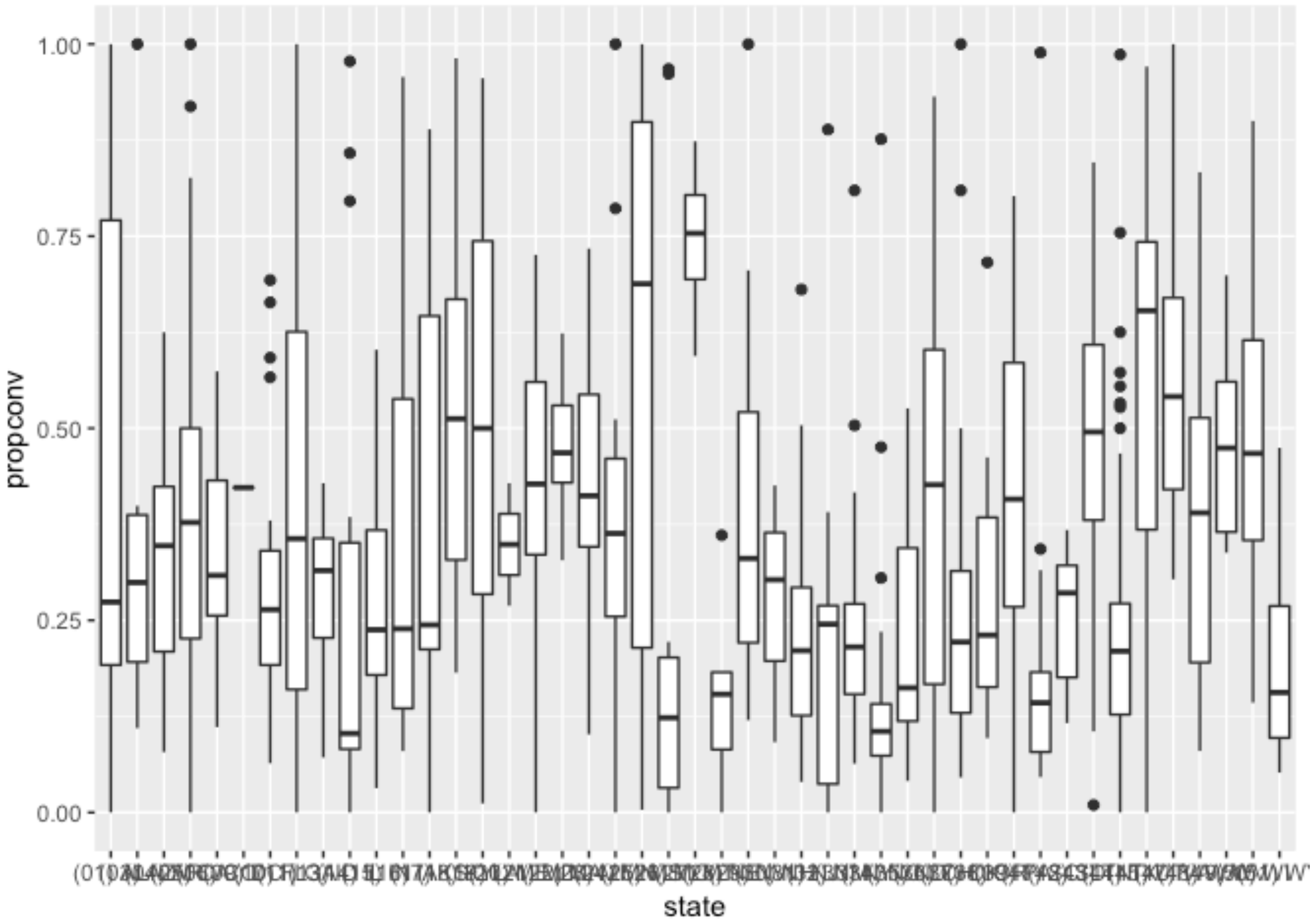
```
library(survey)
```

```
## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
```

```
##
##     dotchart
```

```
library(ggplot2)
load("/Users/simons/Desktop/152project/jail16/DS0001/37135-0001-Data.rda"
jail=da37135.0001

jail1=select(jail,RTIID,GID,JURISID,STATE,STRATUM,
             FINALWT,CONFPOPJUNE,CONFPOP,TOTPOP,
             NONCITZ,ADULTM,ADULTF,
             JUVM,JUVF,TOTGENDER,CONV,UNCONV,FELONY,
             MISD,OTHEROFF,TOTOFF,PEAKPOP,ADPM,ADPF,ADP,
             RELEASEM,RELEASEF,RELEASE)

#nonresponse
nrconf = nrow(jail[jail$CONFPOP_FLAG=="(4) Unit imputed"|jail$CONFPOP_FLA
nrmale = jail[jail$ADULTM_FLAG!="(4) Unit imputed"&jail$CONFPOP_FLAG!="(3


#organize the data frame in 846 rows
group = jail1%>%group_by(JURISID)%>%
  summarise(n=n(),weight=max(FINALWT),strata=STRATUM[1],state = STATE[1],
            release = sum(RELEASE),convicted=sum(CONV),noncitizen=sum(NON


totm = group$male + group$juvm
totf = group$female + group$juvf
group$totm = totm
group$totf = totf
propm = (group$male + group$juvm)/group$confined
group$propm = propm
propf = (group$female + group$juvf)/group$confined
group$propf = propf
#calculate difference between gender proportions
diffprop = group$propm-group$propf
#NaN's are caused by jurisdictions with zero confined population, remove

#reason for picking the cutoff:

##First attempt
group$state = as.character(group$state)
vec = c()
for (i in 1:846){
  if (is.na(group$propm[i])) {vec = c(vec,NA)}
  else {
  if (group$propm[i]<0.4){vec = c(vec,"rare")}
  if (group$propm[i] >= 0.4 & group$propm[i] <= 0.6) {vec = c(vec,"simila
  if (0.6<group$propm[i] & group$propm[i]<0.85) {vec = c(vec,"different")
    if (group$propm[i] >= 0.85) {vec = c(vec,"significantly different")}}
}
group$gendtype = vec

#Does not work very well because "rare" and "similar" numbers are too sma
#we decided to make them NAs so not included in chi-square test

vec = c()
for (i in 1:846){
  if (is.na(group$propm[i])|group$propm[i]<0.4) {vec = c(vec,NA)}
  else {
  if (group$propm[i] >= 0.4 & group$propm[i] <= 0.6) {vec = c(vec,NA)}
  if (0.6<group$propm[i] & group$propm[i]<0.85) {vec = c(vec,"different")
    if (group$propm[i] >= 0.85) {vec = c(vec,"significantly different")}}
}
group$gendtype = vec



#felony, misdemeanor and other
felony_prop=group$felony/group$totoff
misd_prop=group$misd/group$totoff
other_prop=group$otheroff/group$totoff
group$felonyp = felony_prop
group$misdp = misd_prop
group$otherp = other_prop

#proportion convicted
group$propconv = group$convicted/group$confined
ggplot(group,aes(x= state,y=propconv)) + geom_boxplot()
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```

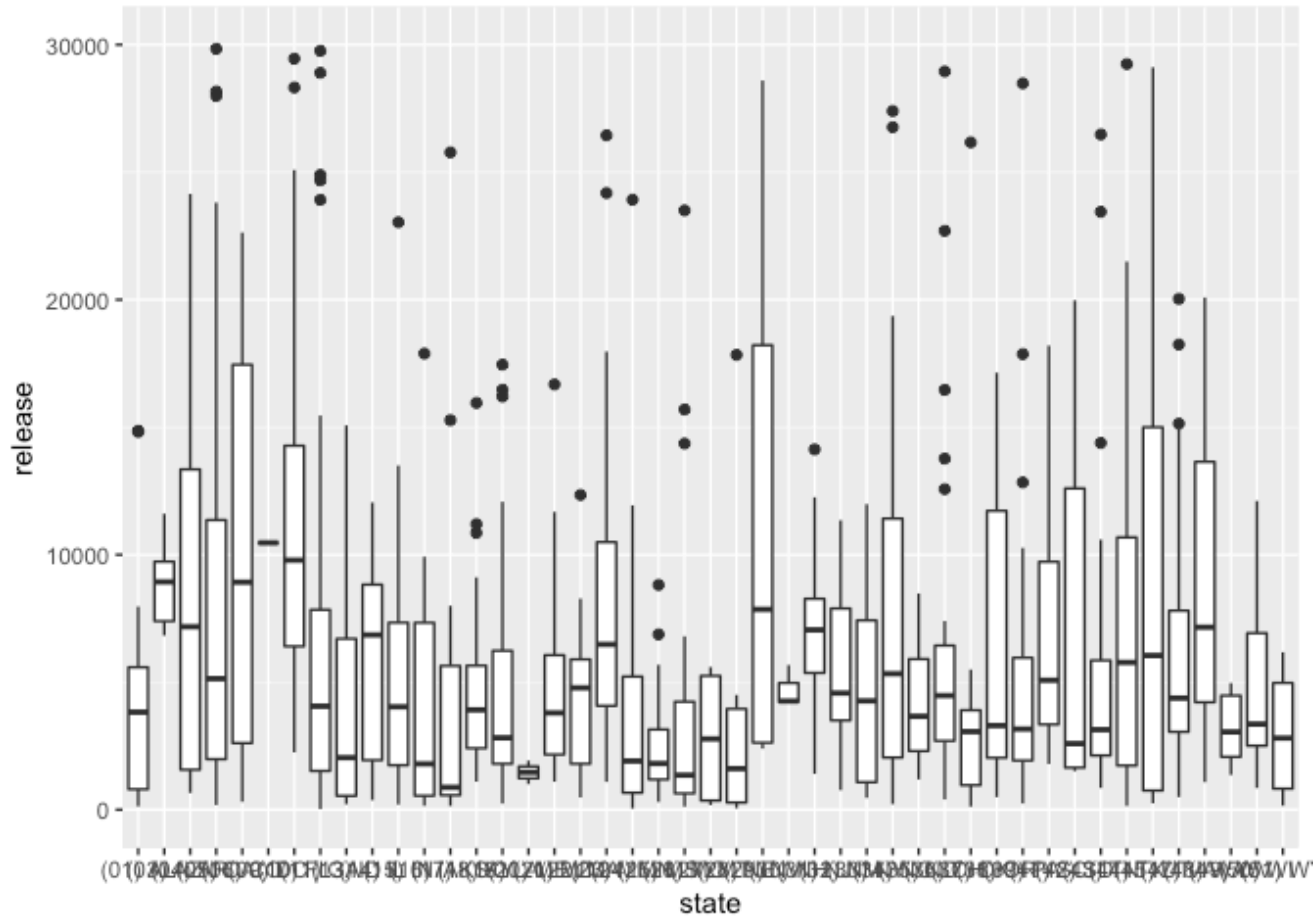


```
head(group,5)
```

```
## # A tibble: 5 x 27
##   JURISID     n weight strata state  male female  juvm  juvf totgen
##   <fct>   <int>  <dbl> <fct>  <chr> <dbl>  <dbl> <dbl> <dbl>  <dbl>
## 1 011002…     1   1.05 (01) … (01)…   382     80     4     0    466
## 2 011004…     1   9.30 (09) … (01)…    61      8     0     0     69
## 3 011015…     1   9.30 (09) … (01)…    41      9     0     0     50
## 4 011022…     1   1.30 (07) … (01)…   246     61     0     0    307
## 5 011028…     1   1.05 (01) … (01)…   598     88     0     0    686
## # … with 17 more variables: confined <dbl>, felony <dbl>, misd <dbl>,
## #   otheroff <dbl>, totoff <dbl>, release <dbl>, convicted <dbl>,
## #   noncitizen <dbl>, totm <dbl>, totf <dbl>, propm <dbl>, propf <dbl>
## #   gendtype <chr>, felonyp <dbl>, misdp <dbl>, otherp <dbl>,
## #   propconv <dbl>
```

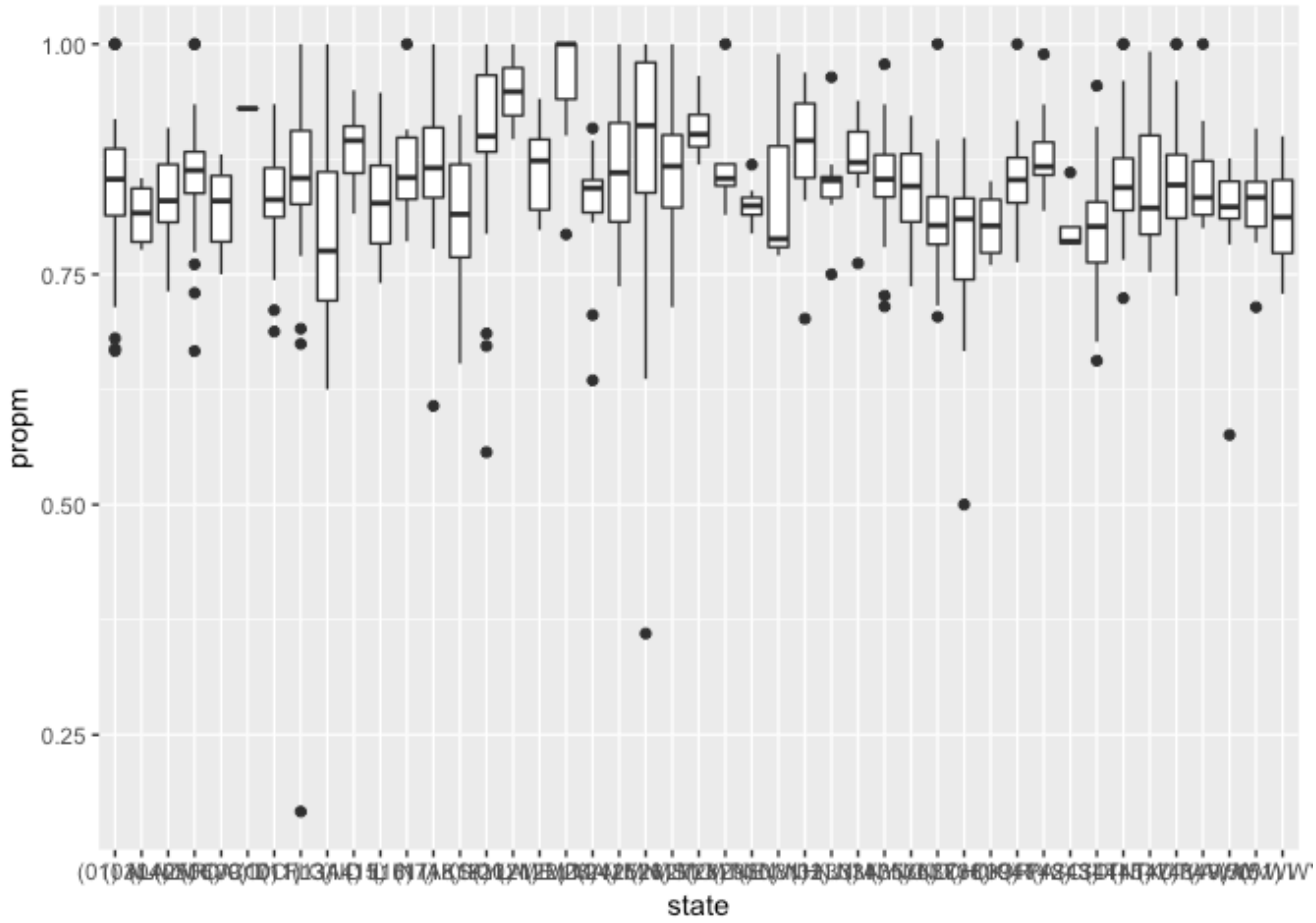
```
ggplot(group,aes(x= state, y=release))+geom_boxplot() +ylim(0,30000)
```

```
## Warning: Removed 36 rows containing non-finite values (stat_boxplot).
```



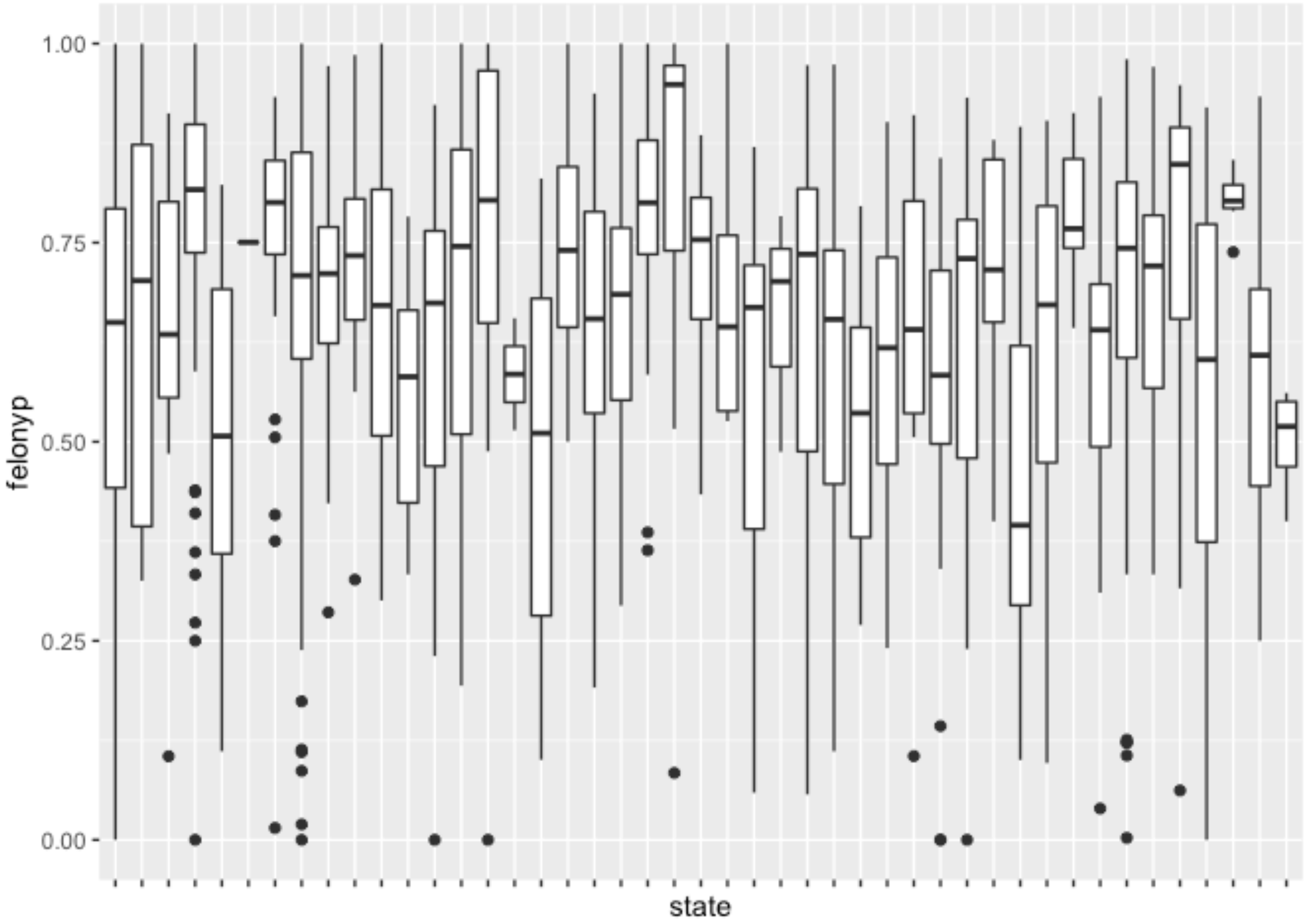
```
ggplot(group,aes(x= state,y=propm)) + geom_boxplot()
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



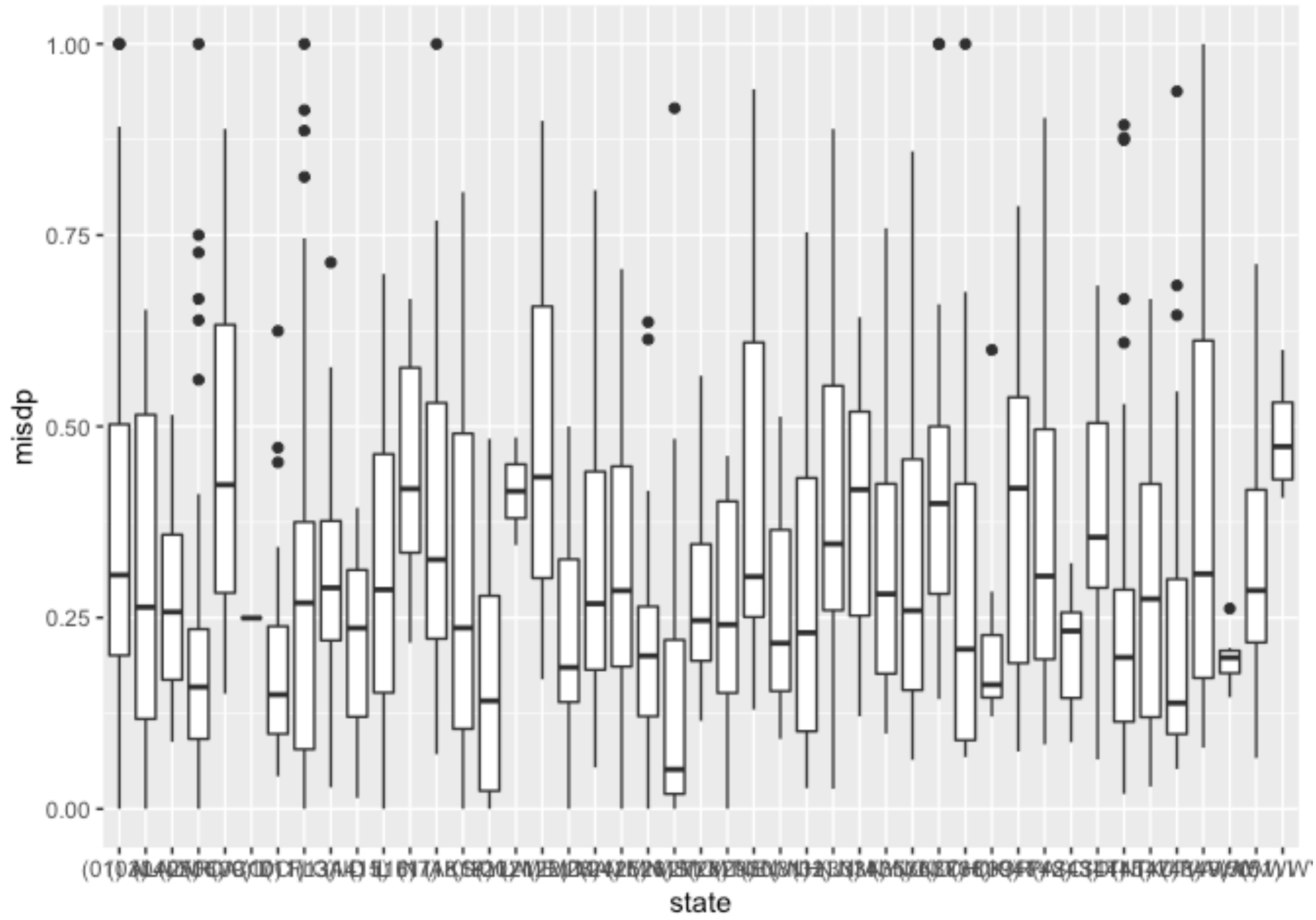
```
ggplot(group,aes(x= state,y=felonyp)) + geom_boxplot() +scale_x_discrete(
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



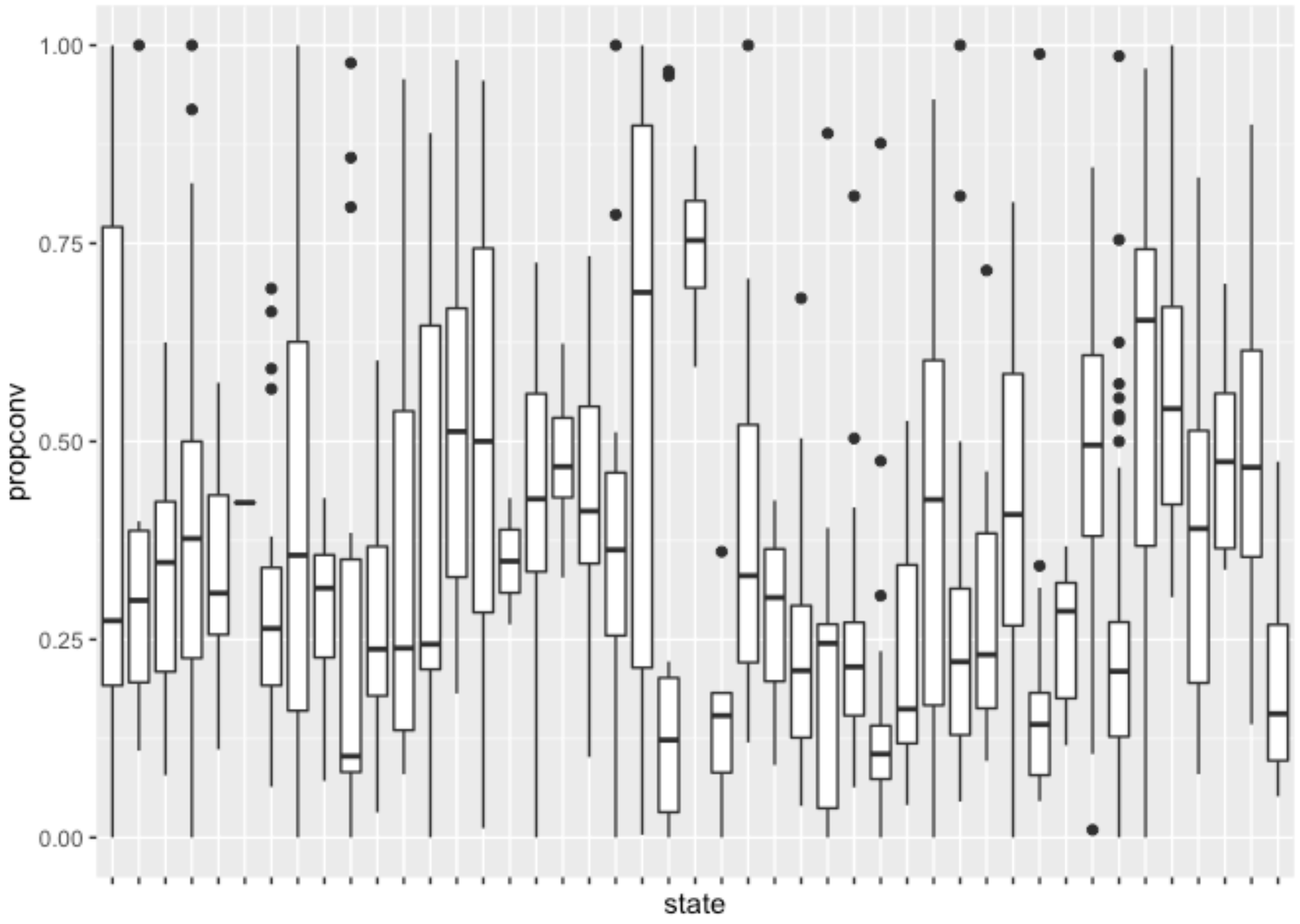
```
ggplot(group,aes(x= state,y=misdp)) + geom_boxplot()
```

```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



```
ggplot(group,aes(x= state,y=propconv)) + geom_boxplot() +scale_x_discrete
```

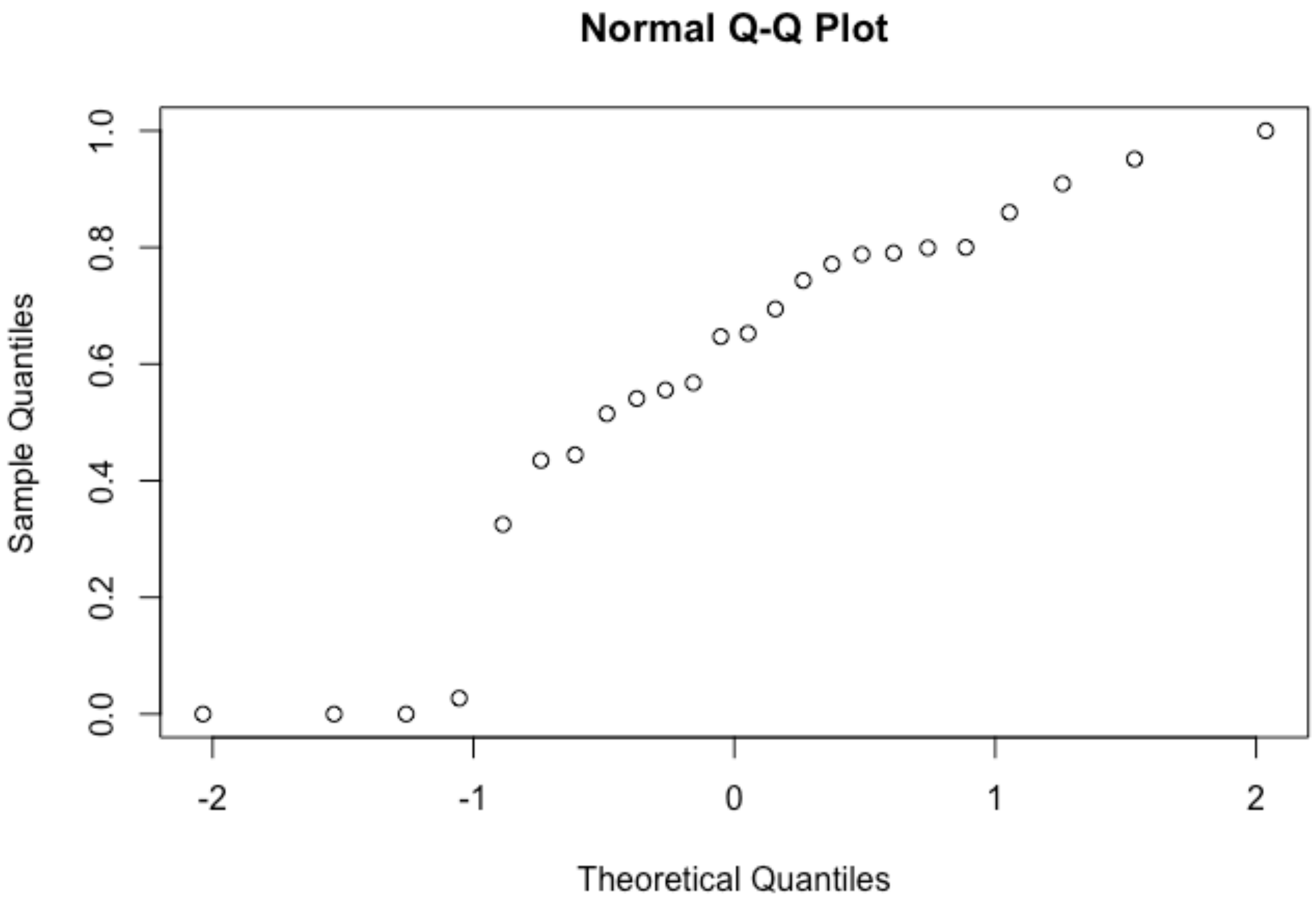
```
## Warning: Removed 5 rows containing non-finite values (stat_boxplot).
```



```
#create the survey object
design = svydesign(id= ~1, strata = ~strata,weights=~weight,data = group)

#Appendix
```

Now we explore the distribution of gender across states by performing a chi-square test within the survey object

```
#explore felony
qqnorm(group$felonyp[group$state=="(01) AL"])
```



```
#choosing one state as an example, the distribution of felony proportion
svyranktest(felonyp~state,design,test="KruskalWallis")
```

```
##
##  Design-based KruskalWallis test
##
## data:  felonyp ~ state
## df = 44, Chisq = 858.57, p-value < 2.2e-16
```

```
#p-value is really small, so we concluded that there are certain states w

#explore gender distributions
svyranktest(propm~state,design,test="KruskalWallis")
```

```
##
##  Design-based KruskalWallis test
##
```

```
## data:  propm ~ state
## df = 44, Chisq = 1795.8, p-value < 2.2e-16
```

```
#make a contingency table using the column of categories
gendertbl=round(svytable(~state+gendtype,design))
summary(gendertbl,statistic="Chisq")
```

```
##          gendtype
## state     different significantly different
##   (01) AL        58                      60
##   (03) AZ         8                       1
##   (04) AR        10                      17
##   (05) CA        24                      40
##   (06) CO        46                      15
##   (09) DC         0                       1
##   (10) FL        29                      11
##   (11) GA        93                      77
##   (13) ID        46                      41
##   (14) IL        26                      30
##   (15) IN        57                      27
##   (16) IA        28                      52
##   (17) KS        48                      63
##   (18) KY        35                      17
##   (19) LA        10                      53
##   (20) ME         0                       2
##   (21) MD        12                      15
##   (22) MA         1                      10
##   (23) MI        36                      28
##   (24) MN        64                      64
##   (25) MS        23                      24
##   (26) MO        52                     102
##   (27) MT         0                      24
##   (28) NE        24                      21
##   (29) NV        23                       1
##   (30) NH         2                       3
##   (31) NJ         5                      17
##   (32) NM        16                      11
##   (33) NY        10                      26
##   (34) NC        37                      43
##   (35) ND         7                       8
##   (36) OH        93                       9
##   (37) OK        68                      12
##   (38) OR        25                       1
##   (39) PA       42                       33
##   (41) SC        6                       35
##   (42) SD       33                        7
##   (43) TN      108                       14
##   (44) TX      146                      123
##   (45) UT       14                       16
##   (47) VA       25                       34
##   (48) WA       45                       34
##   (49) WV        6                        3
##   (50) WI       66                       20
##   (51) WY       12                       15
## 
##  Pearson's X^2: Rao & Scott adjustment
## 
## data:  svychisq(~state + gendtype, design = design, statistic = "Chisq
## X-squared = 131.79, df = 44, p-value = 0.005418
```

```
svychisq(~state+gendtype,design,statistic="Chisq")
```

```
## 
##  Pearson's X^2: Rao & Scott adjustment
## 
## data:  svychisq(~state + gendtype, design, statistic = "Chisq")
## X-squared = 131.79, df = 44, p-value = 0.005418
```

```
#reject the null, states have different distribution of male and female

#offensetbl=round(svytable(~state+offense,design))
#svychisq(~state+offense,design,statistic="Chisq")
svyranktest(propconv~state,design,test="KruskalWallis")
```

```
## 
##  Design-based KruskalWallis test
## 
## data:  propconv ~ state
## df = 44, Chisq = 705.81, p-value < 2.2e-16
```

```
#summary of example variables that might be of interest
```

```
summary(jail1$STRATUM)
```

```
## (01) 1: certainty jails: large jails and california jails 
##                                                       352 
##                         (02) 2: 264 =< adp < 500 & juv >0 
##                                                        34 
##                         (03) 3: 141 =< adp < 264 & juv >0 
##                                                        18 
##                          (04) 4: 79 =< adp < 141 & juv >0 
##                                                         8 
##                            (05) 5: 0 =< adp < 79 & juv >0 
##                                                        12 
##                        (07) 7: 227 =< adp < 750 & juv = 0 
##                                                       214 
##                        (08) 8: 103 =< adp < 227 & juv = 0 
##                                                        78 
##                         (09) 9: 40 =< adp < 103 & juv = 0 
##                                                        61 
##                            (10) 10: =< adp < 40 & juv = 0 
##                                                        58 
##                                   (12) 12: regional jails 
##                                                        67
```

```
summary(jail1$CONFPOP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   129.2   325.0   578.1   671.5 15754.0
```

```
summary(jail1$propm)
```

```
## Length  Class   Mode 
##      0   NULL   NULL
```

```
summary(jail1$propf)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
data(api)
xtabs(~sch.wide+stype, data=apipop)
```

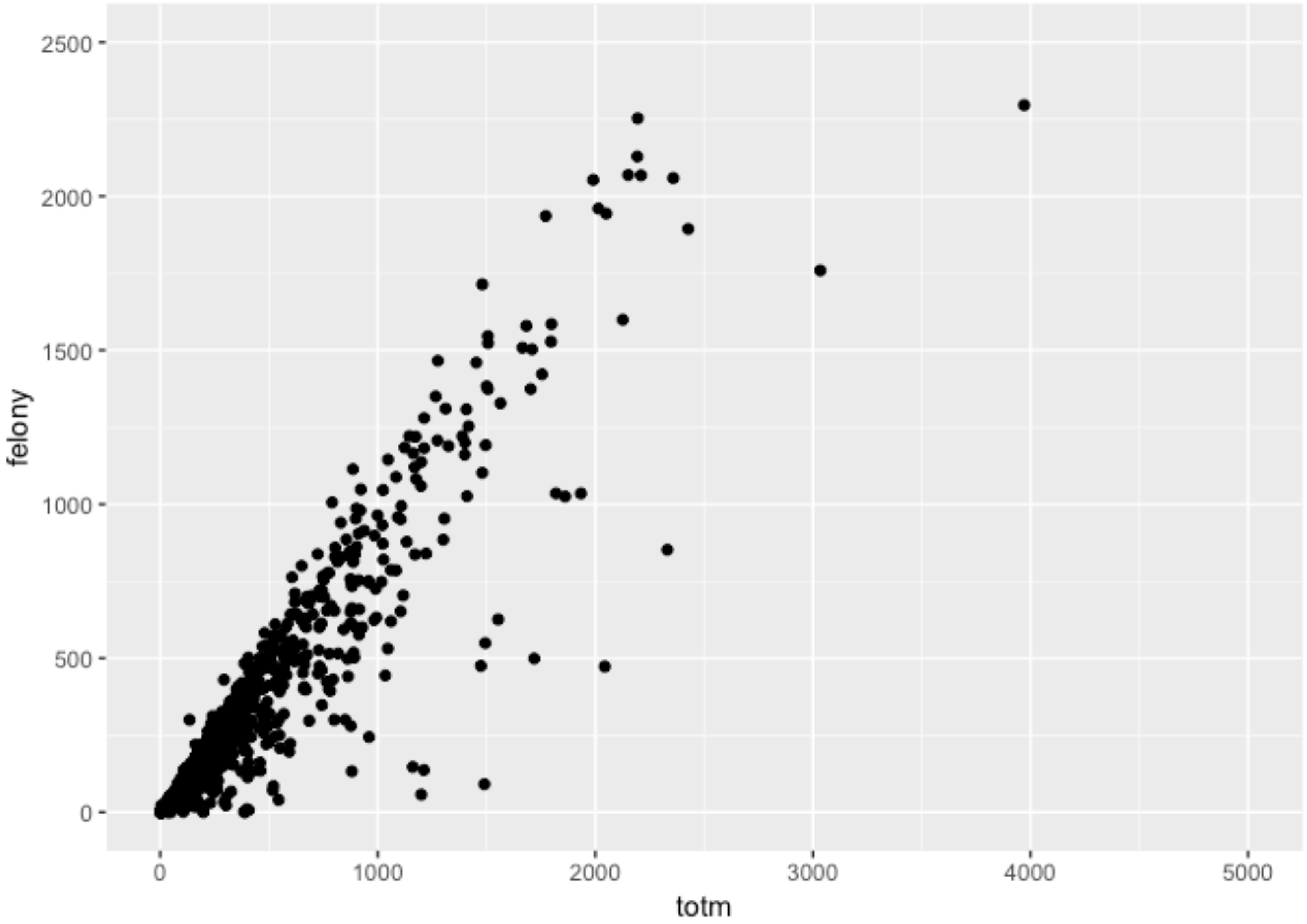
```
##         stype
## sch.wide    E    H    M
##      No   472  334  266
##      Yes 3949  421  752
```

```
diffprop = group$propm-group$propf
```

scatter plots

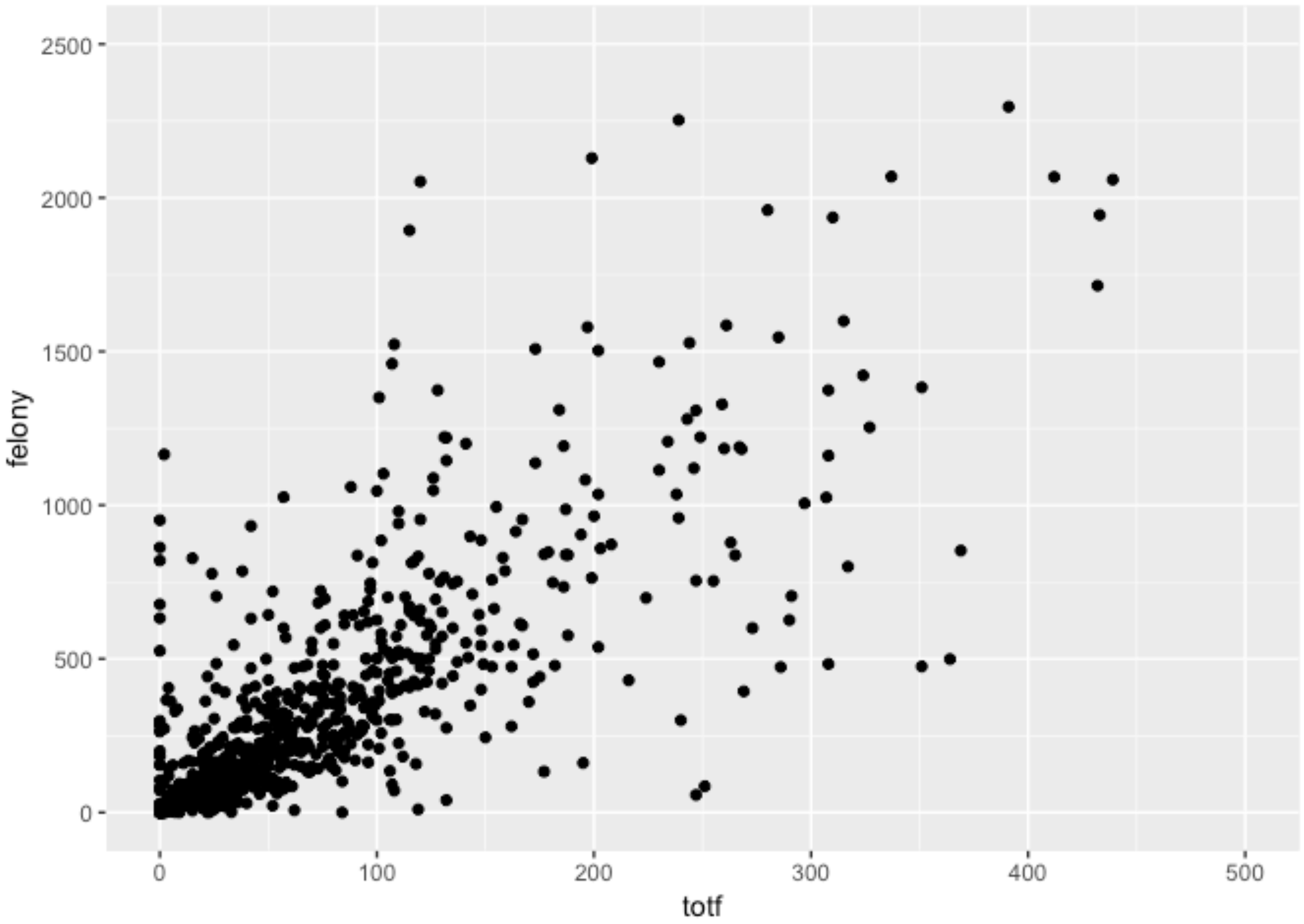
```
#male
ggplot(group,aes(x= totm,y=felony)) + geom_point()+xlim(0,5000)+ylim(0,25
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```



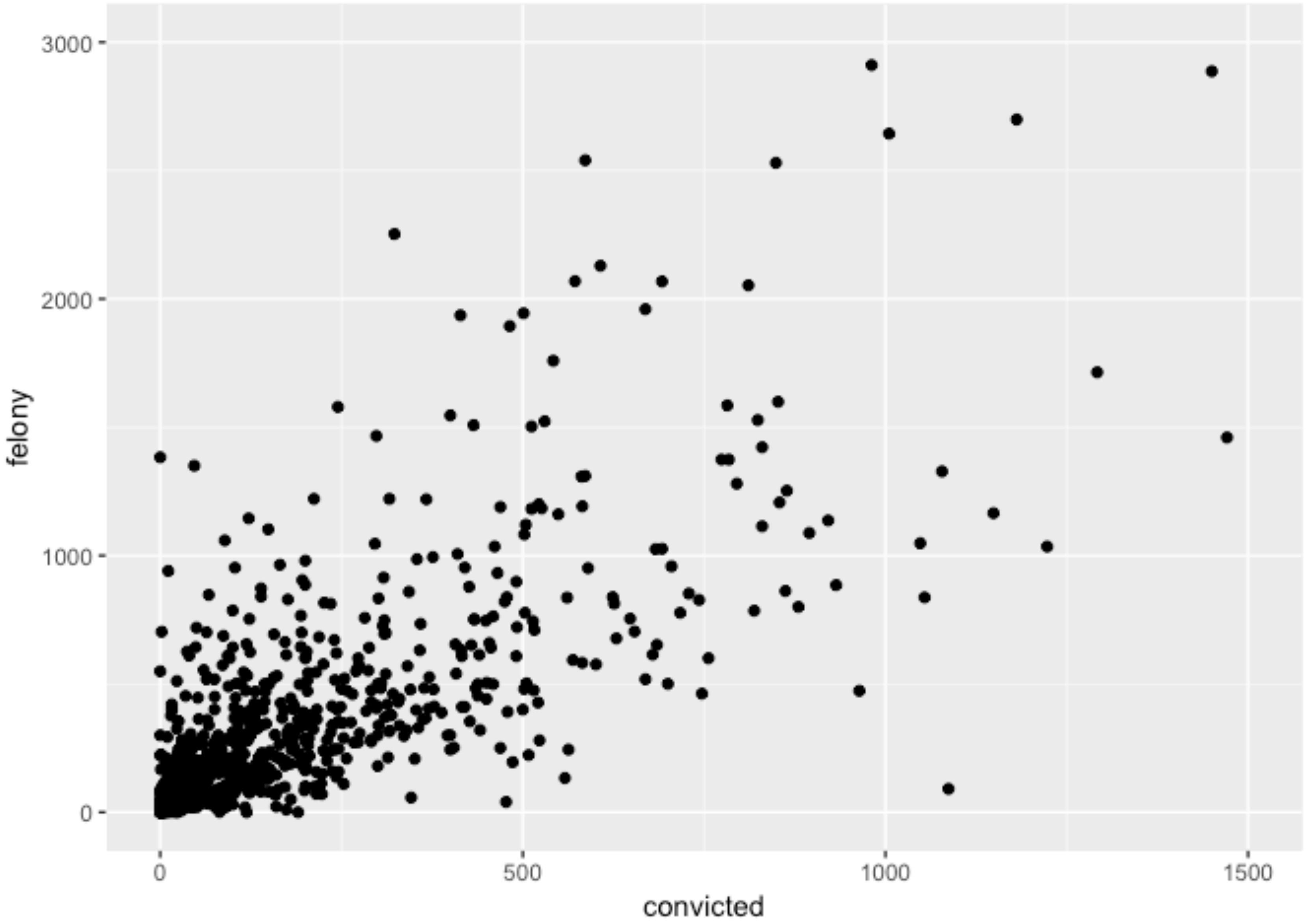
```
#female
ggplot(group,aes(x= totf,y=felony)) + geom_point()+xlim(0,500)+ylim(0,250
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```



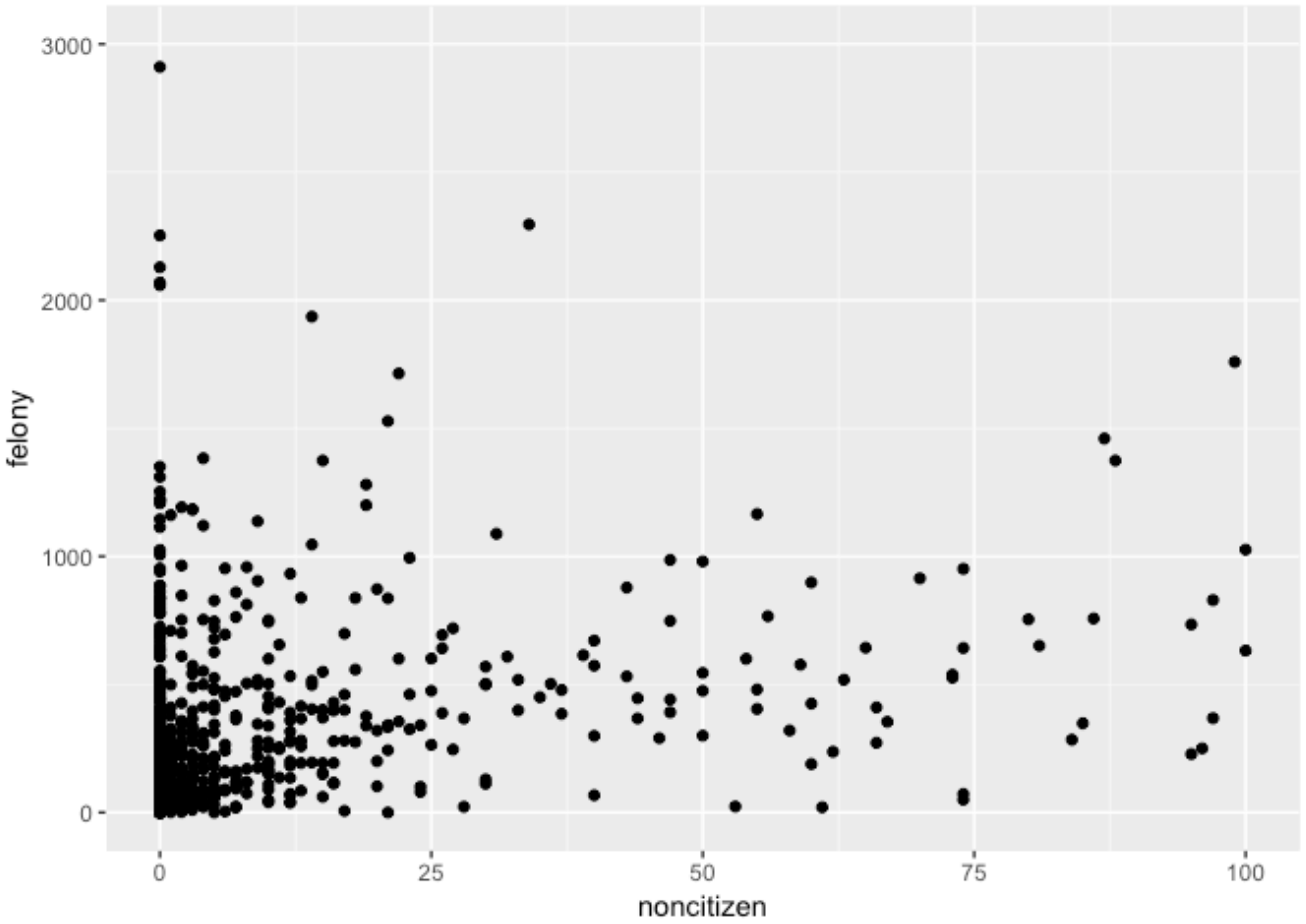
```
#convicted
ggplot(group,aes(x= convicted,y=felony)) + geom_point()+xlim(0,1500)+ylim
```

```
## Warning: Removed 15 rows containing missing values (geom_point).
```



```
#noncitizen
ggplot(group,aes(x= noncitizen,y=felony)) + geom_point()+xlim(0,100)+ylim
```

```
## Warning: Removed 99 rows containing missing values (geom_point).
```



```
#survey estimations
svymean(~propm,design,na.rm=T)
```

```
##         mean     SE
## propm 0.8376 0.0064
```

```
svymean(~felonyp,design,na.rm=T)
```

```
##            mean     SE
## felonyp 0.60914 0.0149
```

```
svymean(~propconv,design,na.rm=T)
```

```
##             mean     SE
## propconv 0.34342 0.0149
```

```
svymean(~propf,design,na.rm=T)
```

```
##          mean     SE
## propf 0.1624 0.0064
```