Final Project Code

Kunwu Lyu and Evan Hart

2025-03-16

Data Wrangling

```
## Data from ICPSR
survey <- read_tsv("data/ICPSR_37143/DS0001/37143-0001-Data.tsv") %>%
  janitor::clean_names() # To all lower case
receipt <- read_tsv("data/ICPSR_37143/DS0002/37143-0002-Data.tsv") %>%
  janitor::clean_names()
fast_food <- read_tsv("data/ICPSR_37143/DS0003/37143-0003-Data.tsv") %>%
  janitor::clean_names()
grocery <- read_tsv("data/ICPSR_37143/DS0004/37143-0004-Data.tsv") %>%
  janitor::clean_names()
recall <- read_tsv("data/ICPSR_37143/DS0005/37143-0005-Data.tsv") %>%
  janitor::clean_names()

## Combine multiple surveys
full_data <- survey %>%
  full_join(receipt, relationship = "many-to-many") %>%
  full_join(fast_food, relationship = "many-to-many") %>%
  full_join(grocery, relationship = "many-to-many") %>%
  full_join(recall, relationship = "many-to-many")

## Mutating
full_data <- full_data %>%
  mutate(limit = ordered(q75, levels = c("Never", "Seldom", "Sometimes",
                                         "Often", "Always"))) %>% # for ordinal
  mutate(age = as.numeric(q76),
         gender = if_else(q77 == 0, "M", "F"),
         race = case_when(
           !is.na(q79_1) ~ "Native",
           !is.na(q79_2) ~ "Black",
           !is.na(q79_3) ~ "Asian",
           !is.na(q79_4) ~ "White",
           !is.na(q79_a) ~ "Other"
         ),
         edu = as.numeric(q80),
         location = nemslocationindicator,
         city = q1,
         num_kids = q44,
         surveydate = dmy(surveydate)) %>%
  mutate(days_since_ban =
           as.numeric(interval(as.Date("2013-03-12"), surveydate) / days(1))) %>%
  filter(age > 0)
```

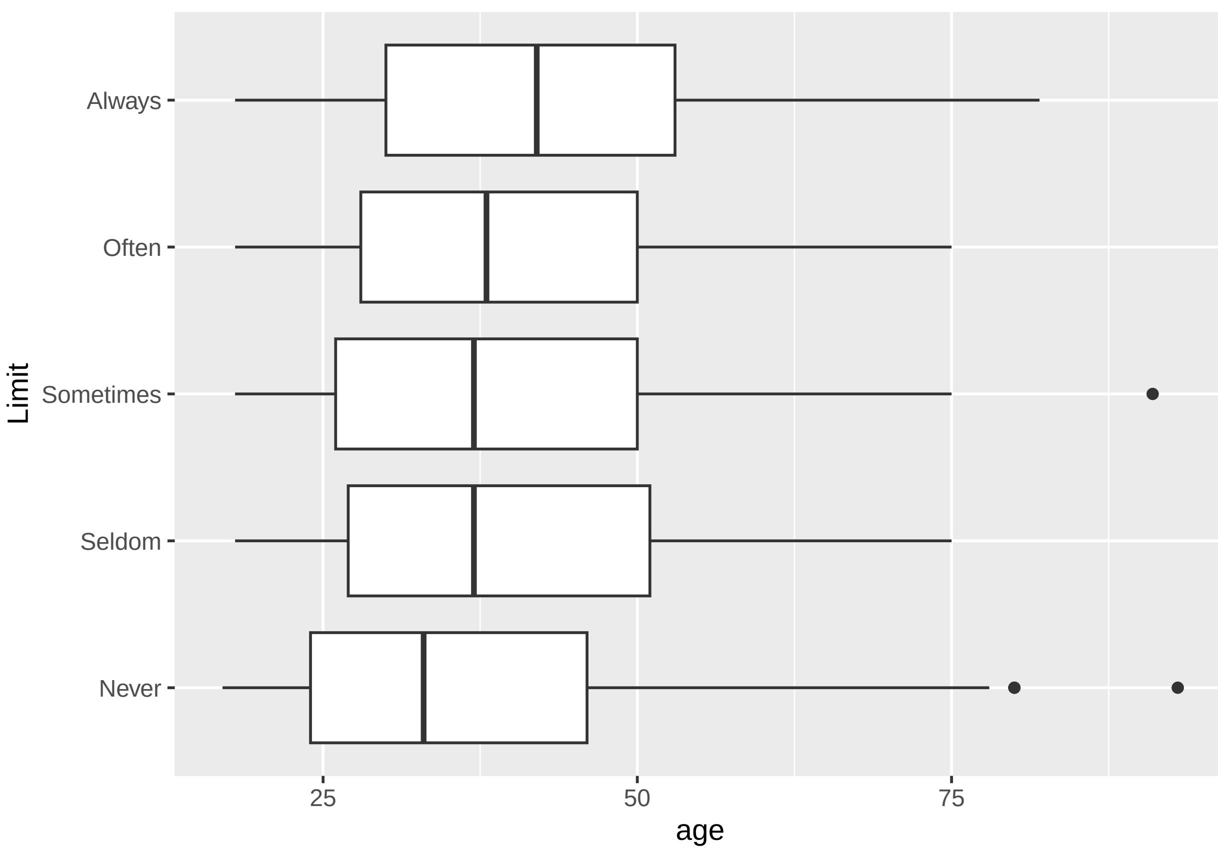
```r
# Standardize numerical for prediction
standardize <- function(x, na.rm = TRUE) {
  (x - mean(x, na.rm = na.rm)) /
    sd(x, na.rm = na.rm)
}

# subset of complete dataset
reduced_data <- full_data %>%
  mutate(age_std = standardize(as.numeric(q76))) %>%
  select(c("receiptid", "person_id", "limit", "age", "age_std", "gender",
           "race", "edu", "city", "caff", "location", "round", "nsigns_ssb",
           "num_kids", "surveydate", "days_since_ban", "kcal", "f_total",
           "v_total", "fatg", "sugarg")) %>%
  group_by(receiptid) %>%
  mutate(caff = mean(caff, na.rm = T), # across each receipt
         f_total = mean(f_total, na.rm = T),
         v_total = mean(v_total, na.rm = T),
         kcal = mean(kcal, na.rm = T),
         fatg = mean(fatg, na.rm = T),
         sugarg = mean(sugarg, na.rm = T)) %>%
  drop_na() %>%
  distinct() %>% # Remove duplicate rows because multiple items are on a receipt
  mutate(receiptid = as.factor(receiptid),
         person_id = as.factor(person_id),
         location = as.factor(location),
         round = as.factor(round),
         edu = case_when(
           edu == 1 ~ "Less than High School",
           edu == 2 ~ "Some High School",
           edu == 3 ~ "High School",
           edu == 4 ~ "Some College",
           edu == 5 ~ "Associates Degree",
           edu == 6 ~ "College Degree",
           edu == 7 ~ "Graduate Degree"
         )) %>%
  ungroup() %>%
  mutate(f_std = standardize(f_total),
         v_std = standardize(v_total),
         caff_std = standardize(caff),
         nsigns_ssb_std = standardize(nsigns_ssb),
         days_since_ban_std = standardize(days_since_ban),
         kcal_std = standardize(kcal),
         fatg_std = standardize(fatg),
         sugarg_std = standardize(sugarg)
         ) %>%
  mutate(fv = f_std + v_std) %>%
  mutate(fv_std = standardize(fv),
         log_age = log(age),
         exp_age = exp(age))

# Cleaned data
write_csv(reduced_data, "dietControl.csv")
```

```r
# One receipt can't appear in multiple locations
multi_receipt_locations <- reduced_data %>%
  group_by(receiptid) %>%
  summarize(n_rounds = n_distinct(location)) %>%
  filter(n_rounds > 1) %>%
  pull(receiptid)

reduced_data %>%
  filter(receiptid %in% multi_receipt_locations) %>%
  count(receiptid, location)
```

```
## # A tibble: 0 x 3
## # i 3 variables: receiptid <fct>, location <fct>, n <int>
```

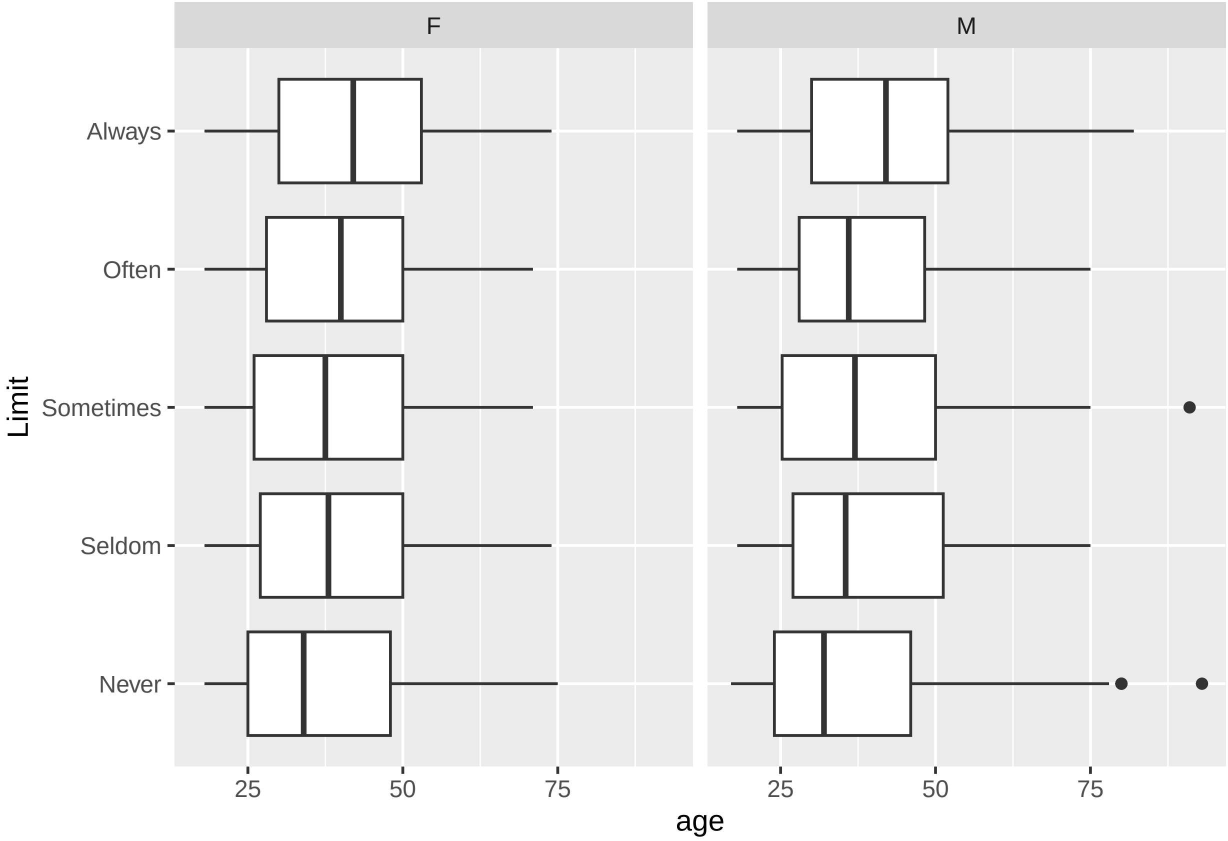
EDA

```r
# Single variables, interactions plotted against limit

# Age
ggplot(data = reduced_data, aes(x = age , y = limit)) +
  geom_boxplot() +
  labs(x = "age", y = "Limit")
```

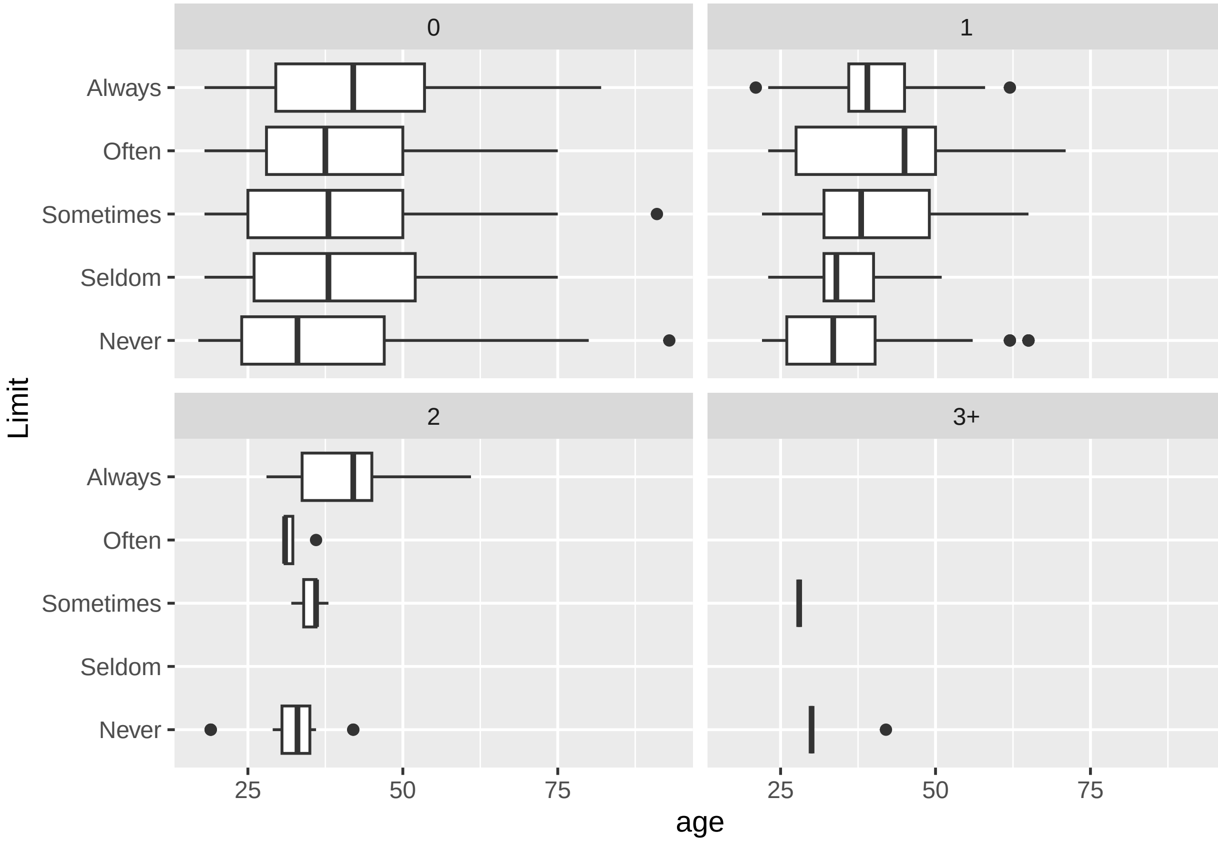



```r
# Age faceted by gender
ggplot(data = reduced_data, aes(x = age , y = limit)) +
  geom_boxplot() +
  facet_wrap(~gender) +
```

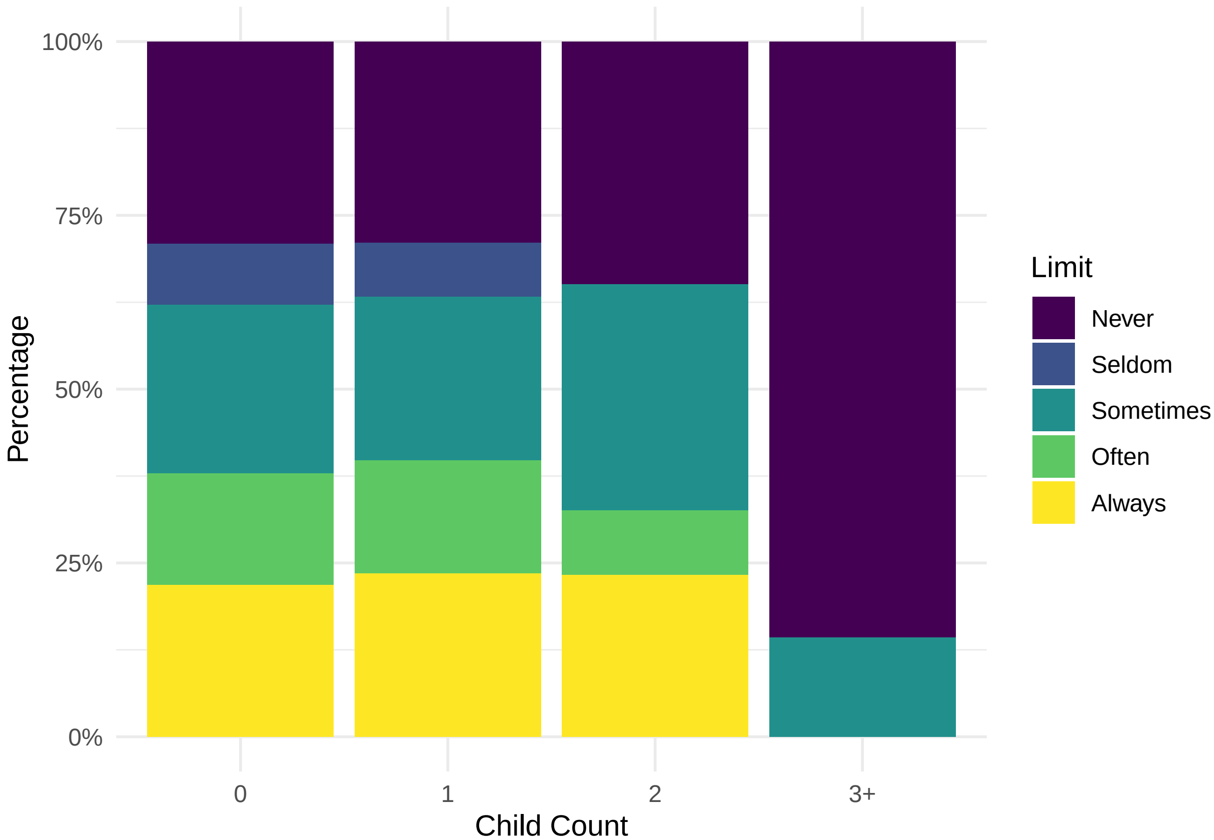
```
  labs(x = "age", y = "Limit")
```



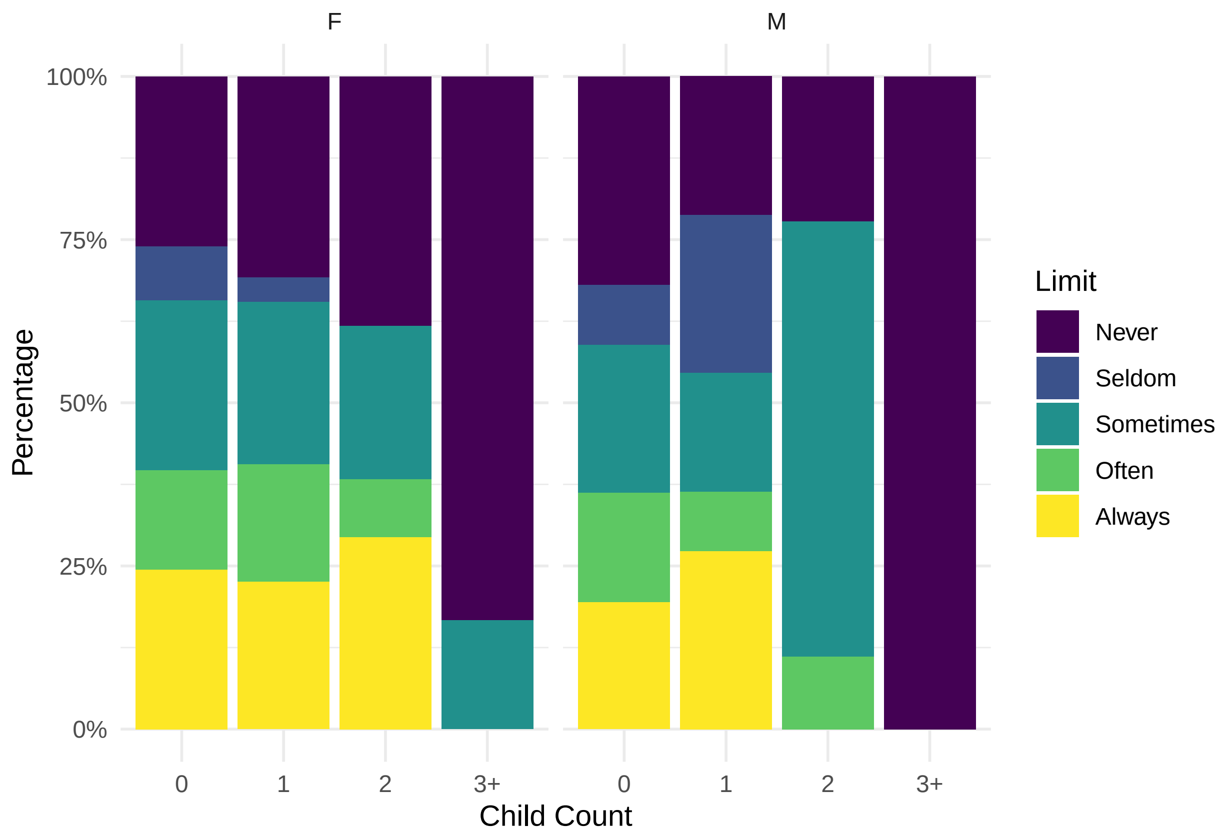

```
# Age faceted by child count
ggplot(data = reduced_data, aes(x = age , y = limit)) +
  geom_boxplot() +
  facet_wrap(~num_kids) +
  labs(x = "age", y = "Limit")
```



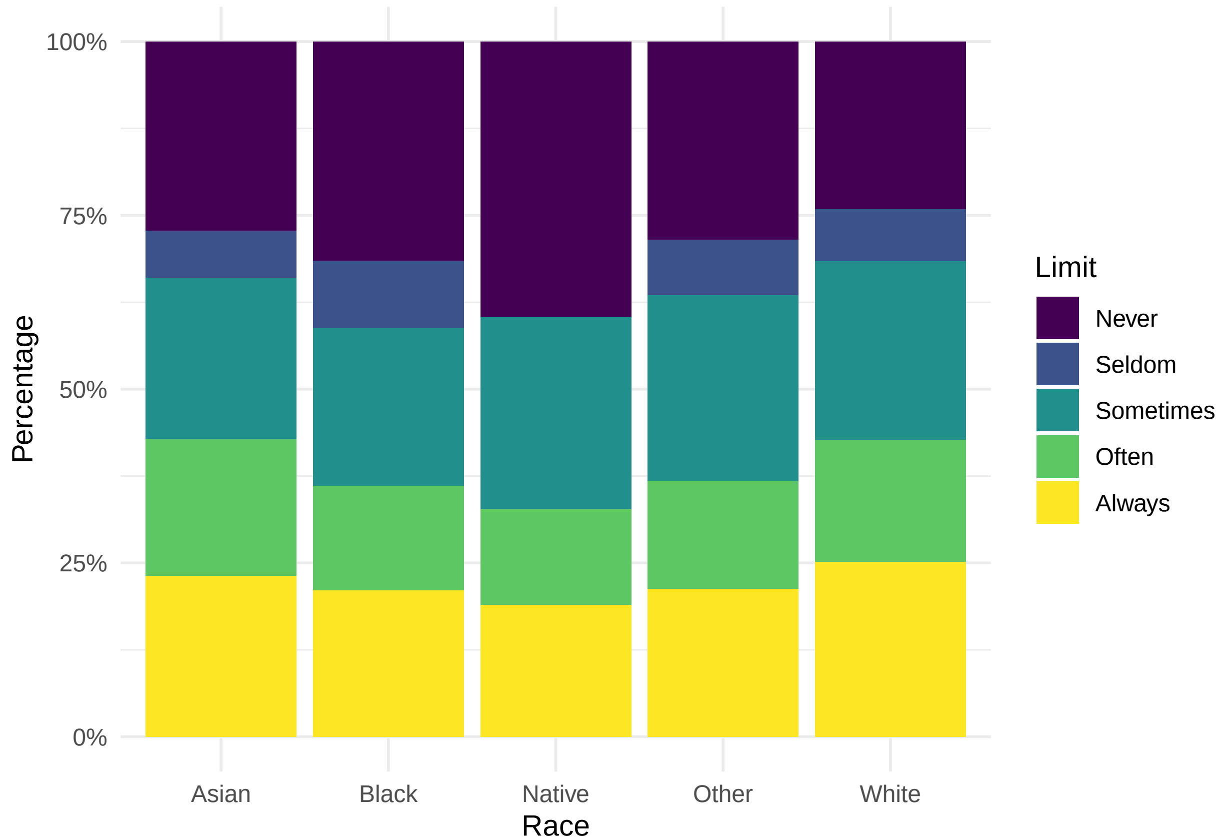

```r
# Child count
ggplot(data = reduced_data, aes(x = num_kids, fill = limit)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Child Count", y = "Percentage", fill = "Limit") +
  theme_minimal()
```



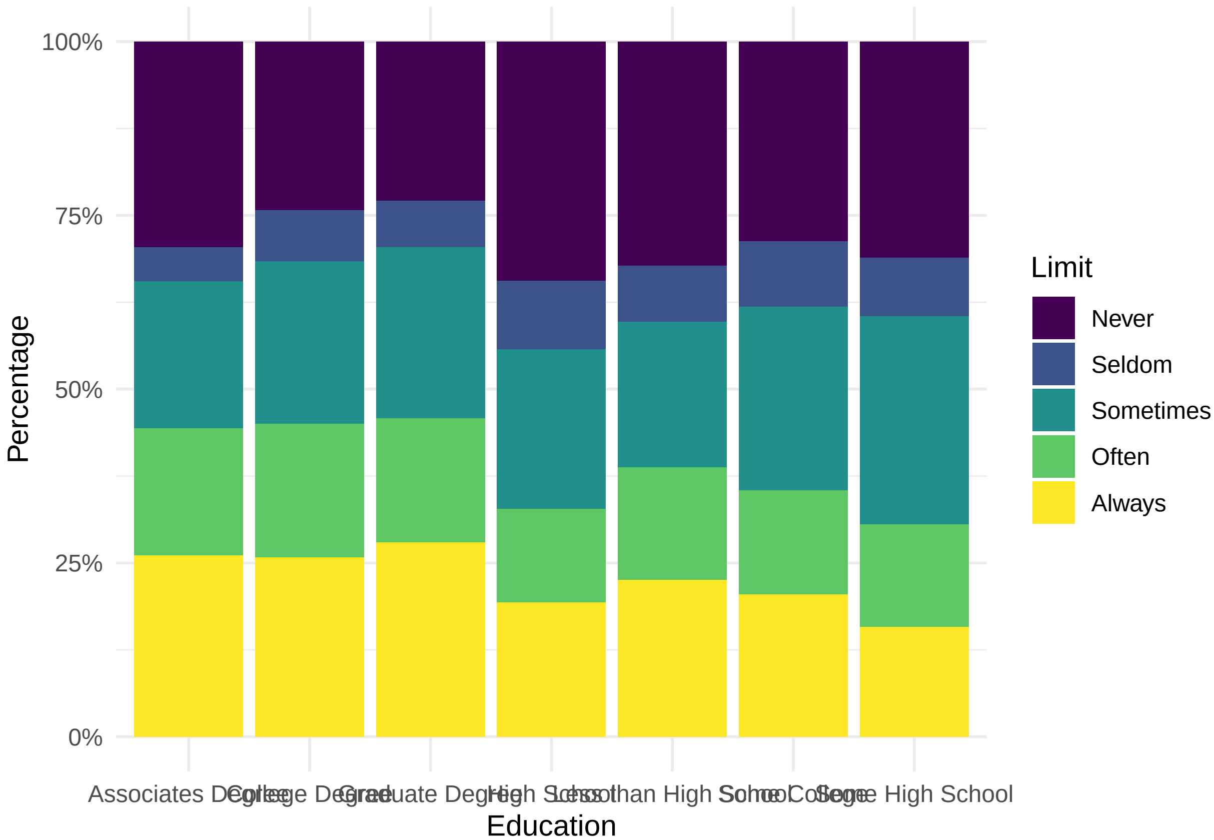

```r
# Child count / gender interaction
ggplot(data = reduced_data, aes(x = num_kids, fill = limit)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Child Count", y = "Percentage", fill = "Limit") +
  facet_wrap(~gender) +
  theme_minimal()
```




```r
# Race
ggplot(data = reduced_data, aes(x = race, fill = limit)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Race", y = "Percentage", fill = "Limit") +
  theme_minimal()
```

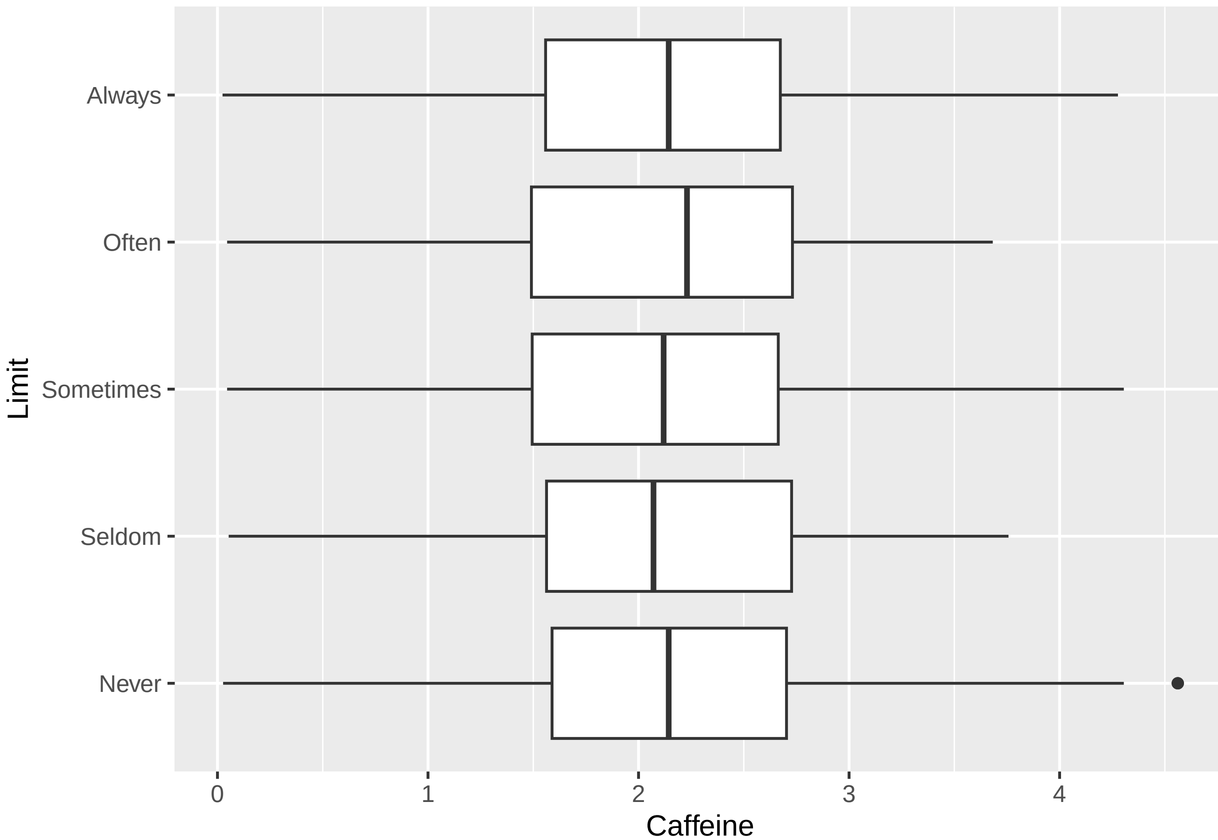



```r
# Education
ggplot(data = reduced_data, aes(x = edu, fill = limit)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Education", y = "Percentage", fill = "Limit") +
  theme_minimal()
```

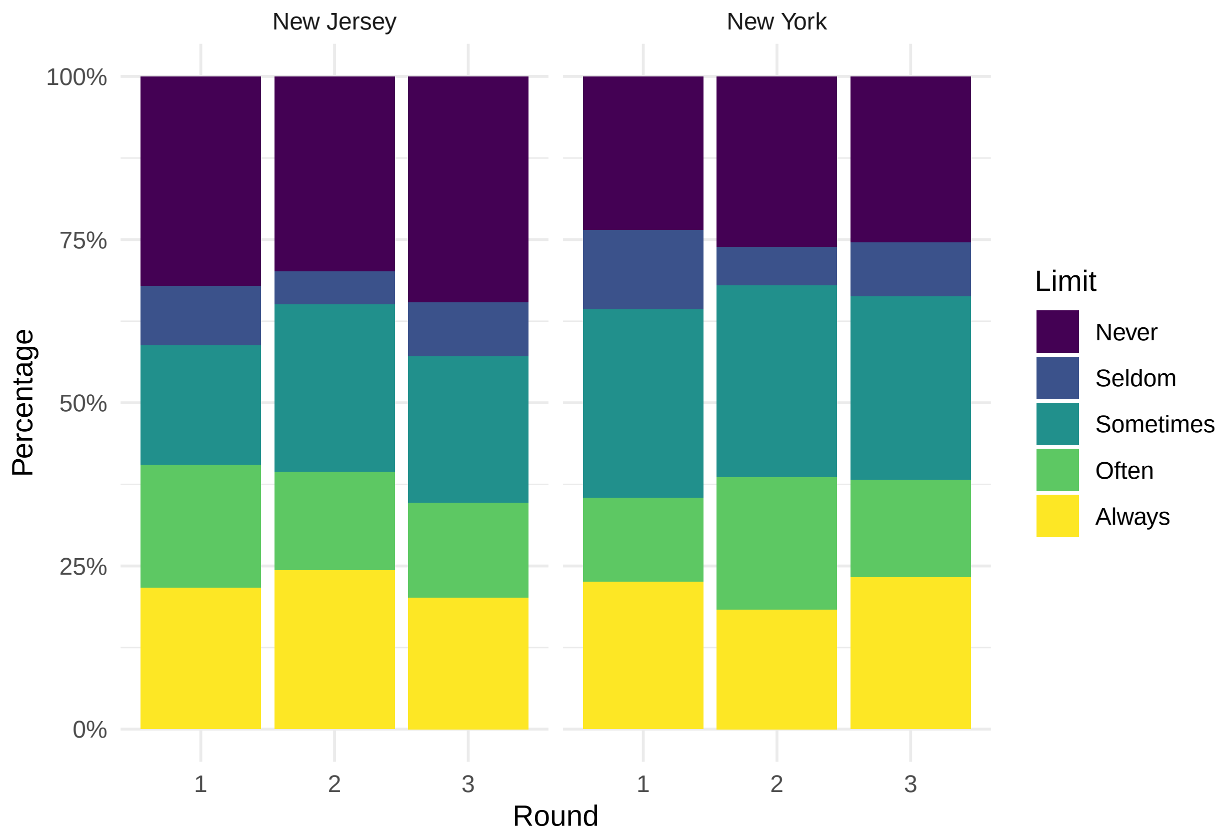



```r
# City
ggplot(data = reduced_data, aes(x = city, fill = limit)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "City", y = "Percentage", fill = "Limit") +
  theme_minimal()
```

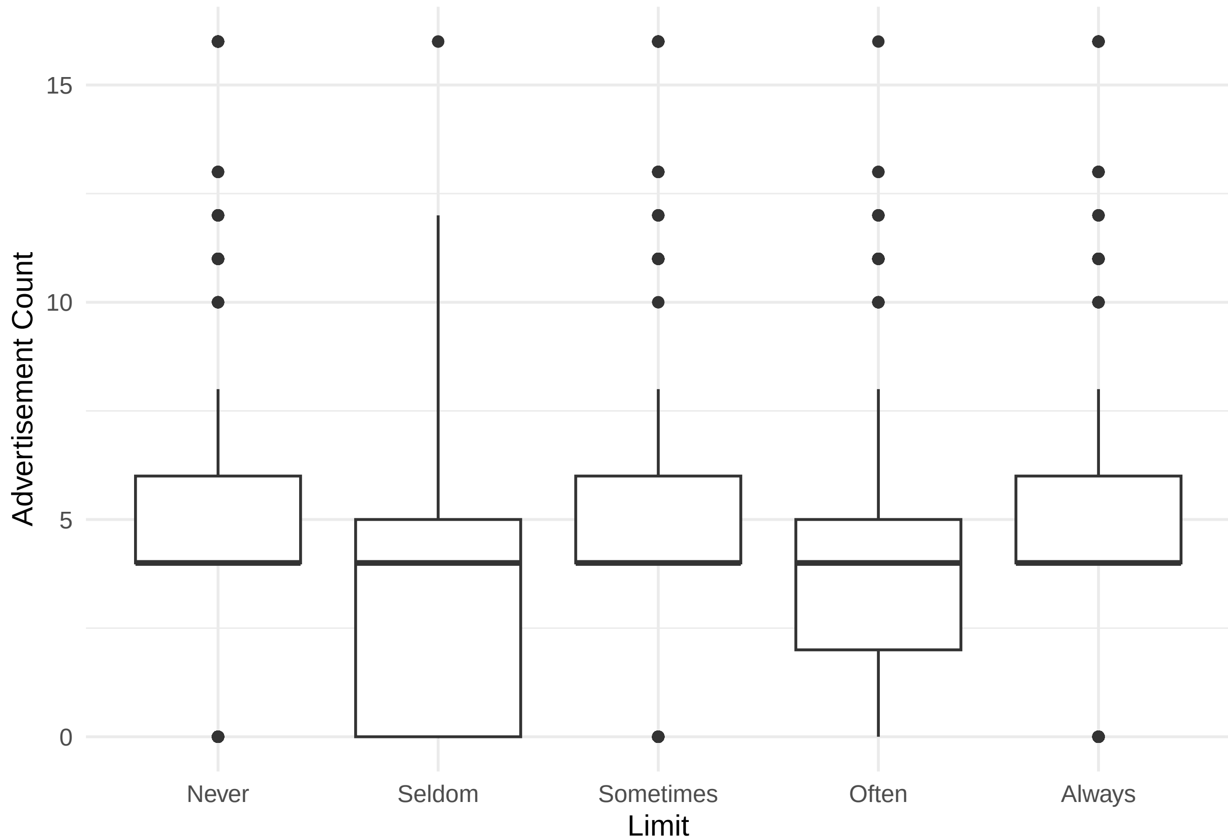

```
# Caffeine
ggplot(data = reduced_data %>% filter(caff > 0), aes(x = log(caff + 1) , y = limit)) +
  geom_boxplot() +
  labs(x = "Caffeine", y = "Limit")
```




```
# Survey round
ggplot(data = reduced_data, aes(x = round, fill = limit)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Round", y = "Percentage", fill = "Limit") +
  facet_wrap(~city) +
  theme_minimal()
```



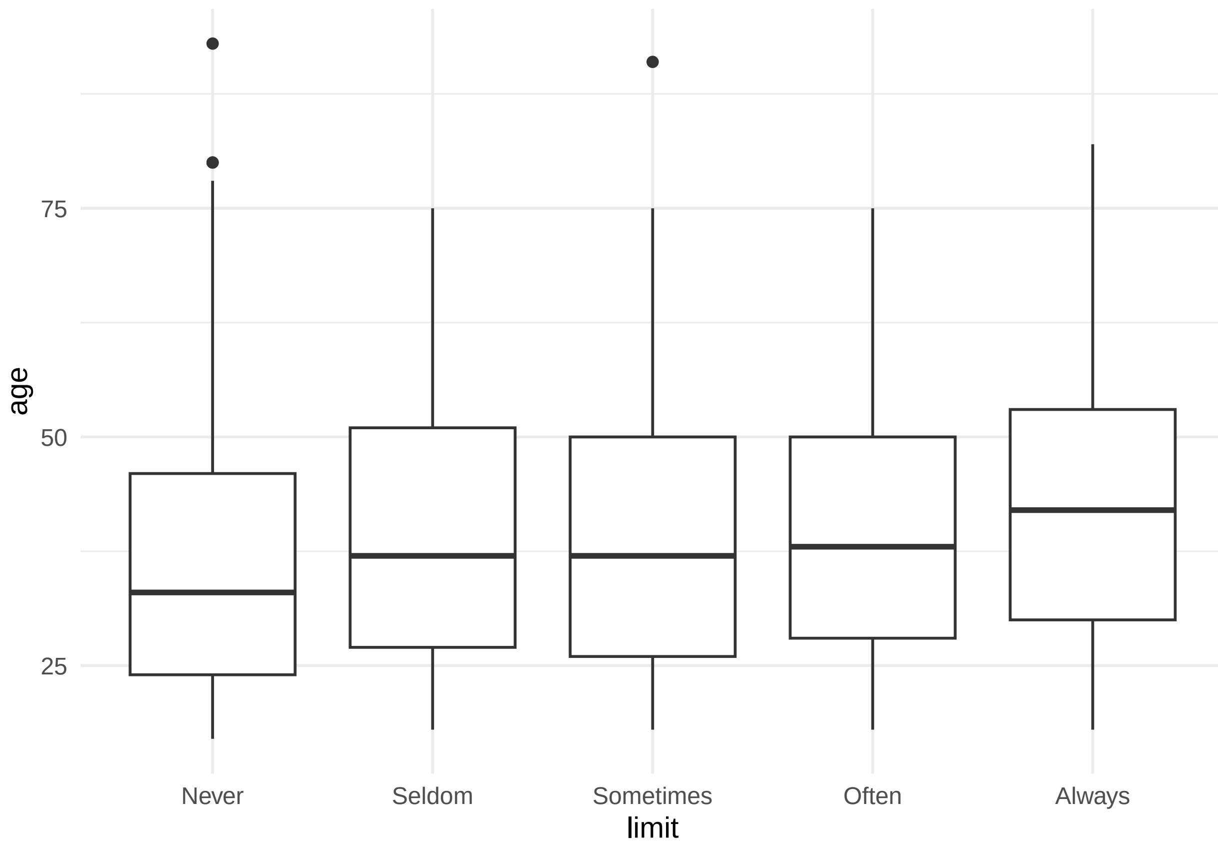

```r
# Advertisement count
ggplot(data = reduced_data, aes(y = nsigns_ssb, x = limit)) +
  geom_boxplot() +
  labs(y = "Advertisement Count", x = "Limit") +
  theme_minimal()
```

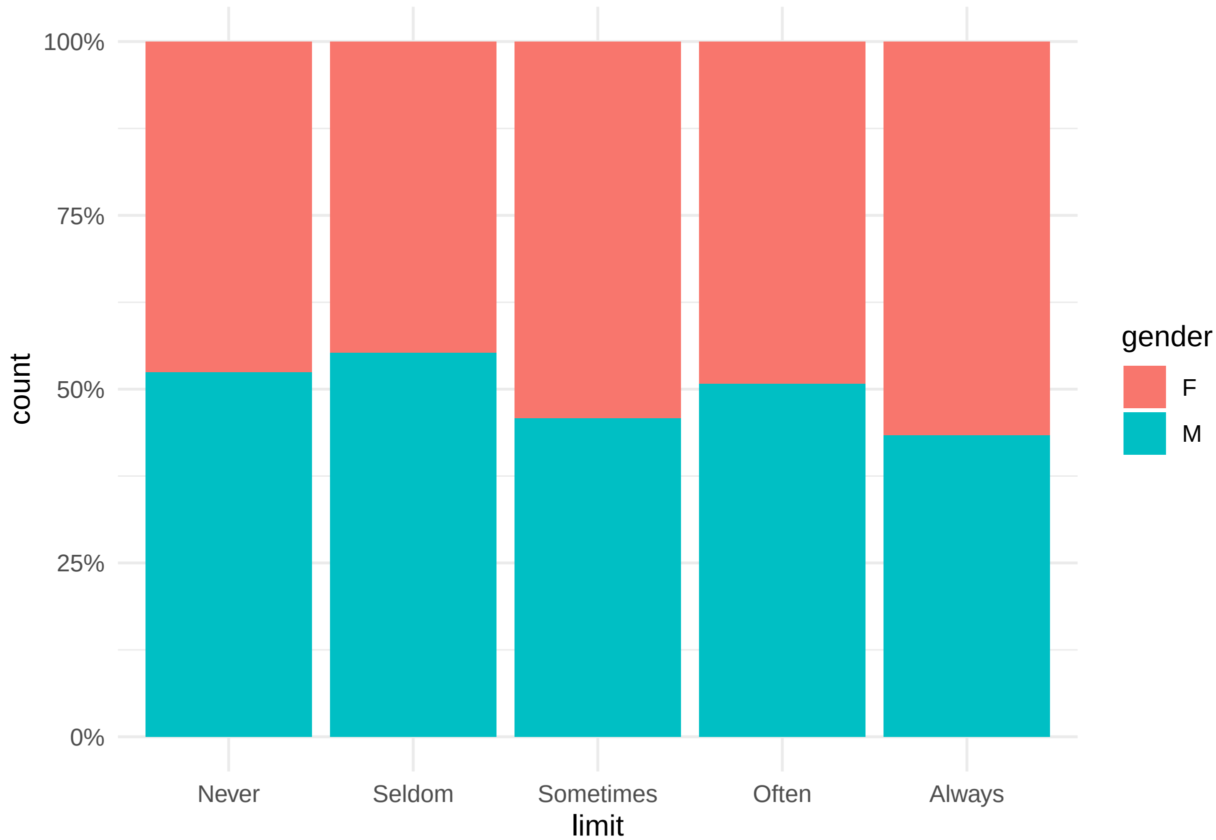


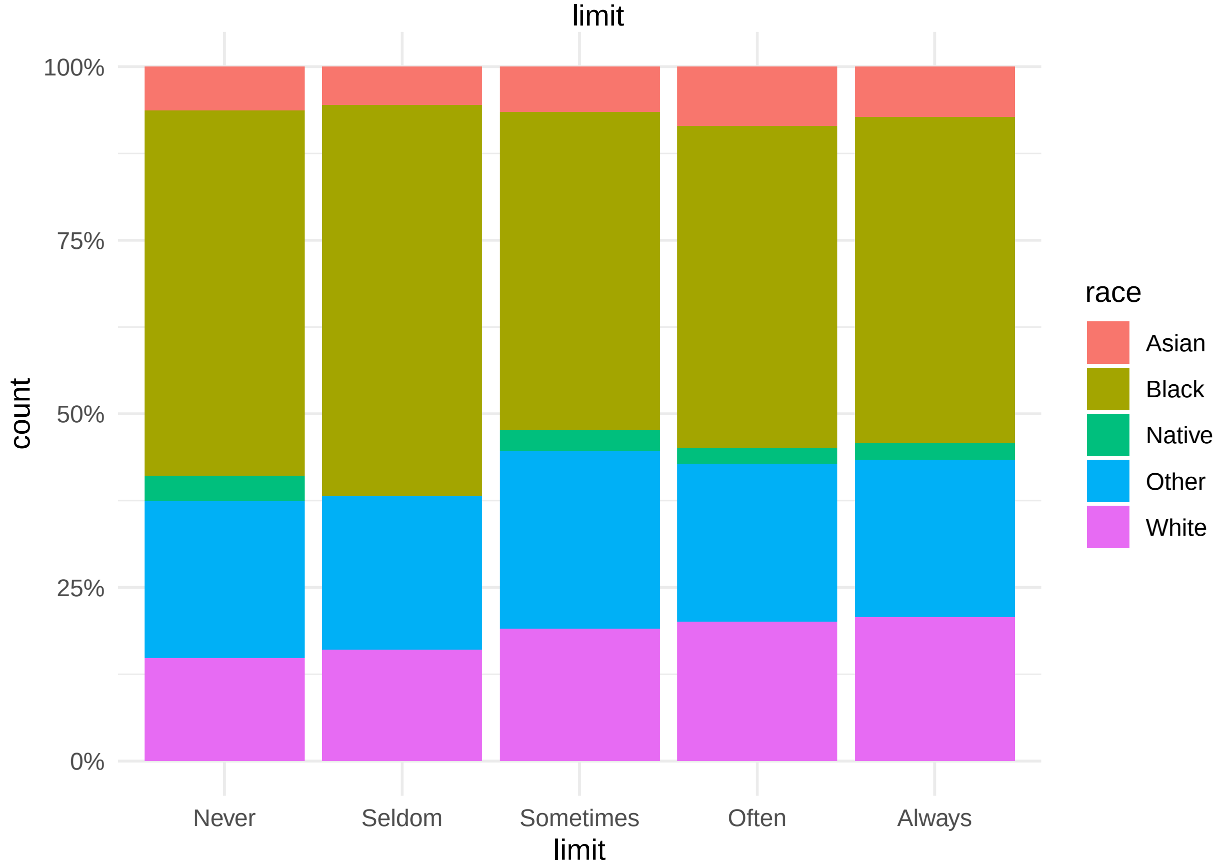

```r
# Plot function for interactions

plot_cats <- c("limit", "gender", "race", "city", "round", "num_kids", "edu", "diet")
plot_nums <- c("age", "caff", "nsigns_ssb", "days_since_ban")

library(rlang)

make_plot <- function(var1, var2){
  if(var1 %in% plot_cats & var2 %in% plot_cats){
    print(ret_plot <- ggplot(data =reduced_data, aes(x = !!sym(var1), fill = !!sym(var2))) +
      geom_bar(position = "fill") +
      scale_y_continuous(labels = scales::percent) +
      theme_minimal())
  }

  if(var1 %in% plot_cats & var2 %in% plot_nums){
    print(ret_plot <- ggplot(data = reduced_data, aes(x = !!sym(var1), y = !!sym(var2))) +
      geom_boxplot() +
      theme_minimal())
  }

  if(var1 %in% plot_nums & var2 %in% plot_cats){
    print(ret_plot <- ggplot(data = reduced_data, aes(x = !!sym(var2), y = !!sym(var1))) +
       geom_boxplot() +
       theme_minimal())
  }
```

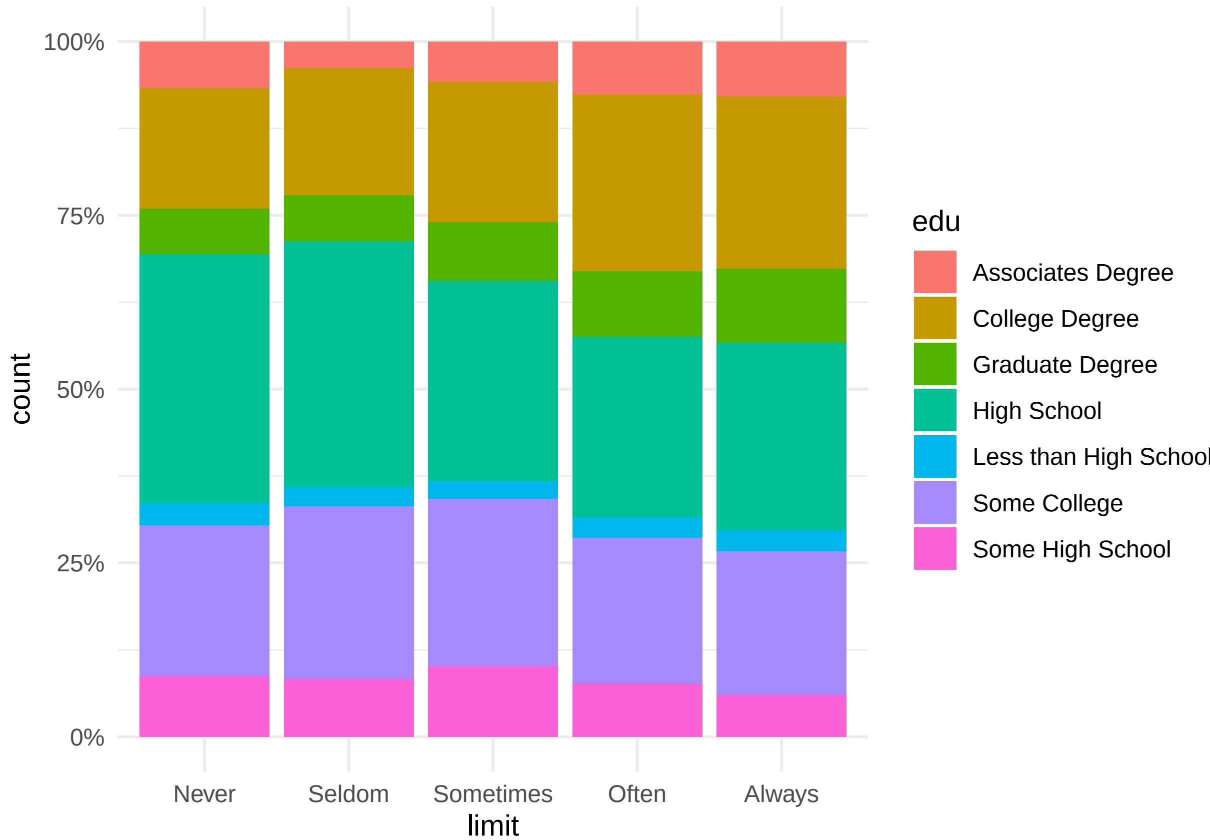
```r
  if(var1 %in% plot_nums & var2 %in% plot_nums){
    print(ggplot(data = reduced_data, aes(x = !!sym(var2), y = !!sym(var1))) +
      geom_point() +
      theme_minimal())
  }
}
```

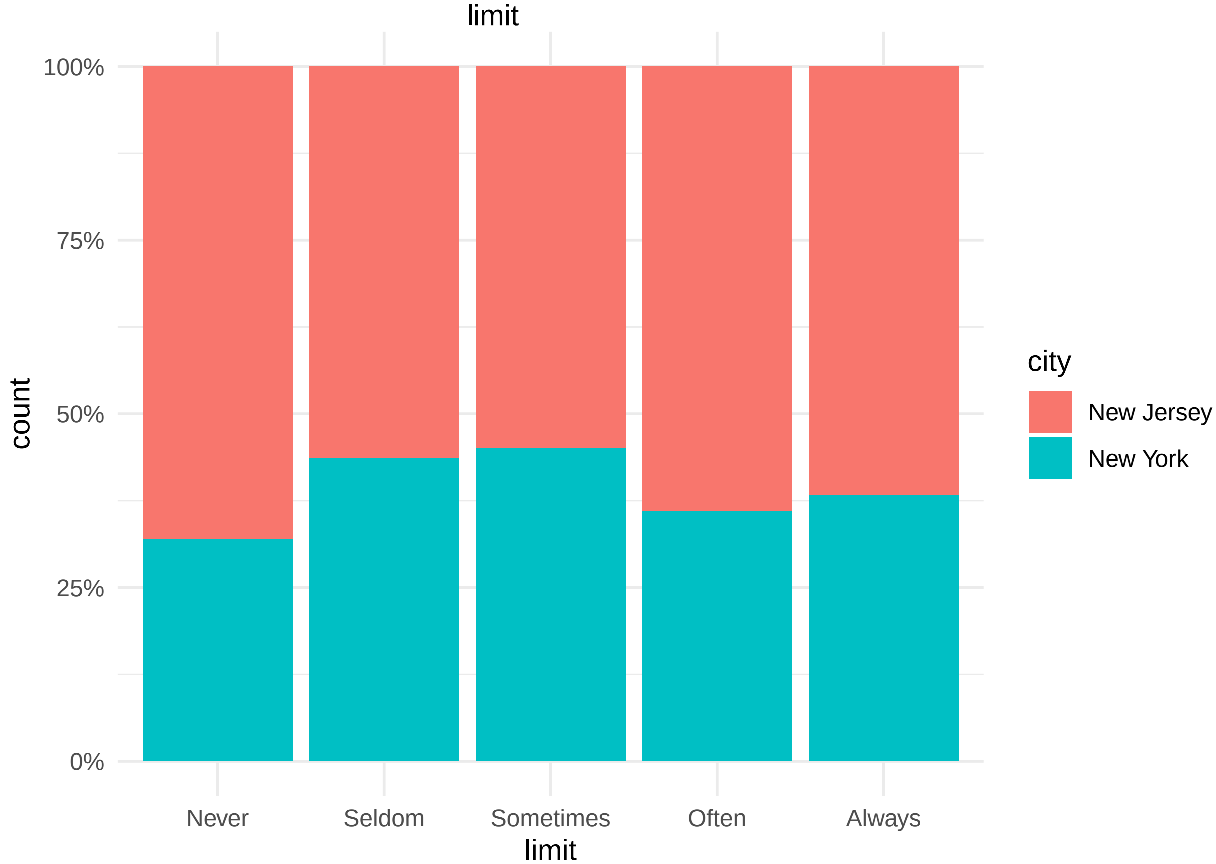
```r
for(i in 1:length(names(reduced_data))){
  if(i != length(reduced_data)){
    for(j in (i+1):length(reduced_data)){
      make_plot(names(reduced_data)[i], names(reduced_data)[j])
    }
  }
}
```

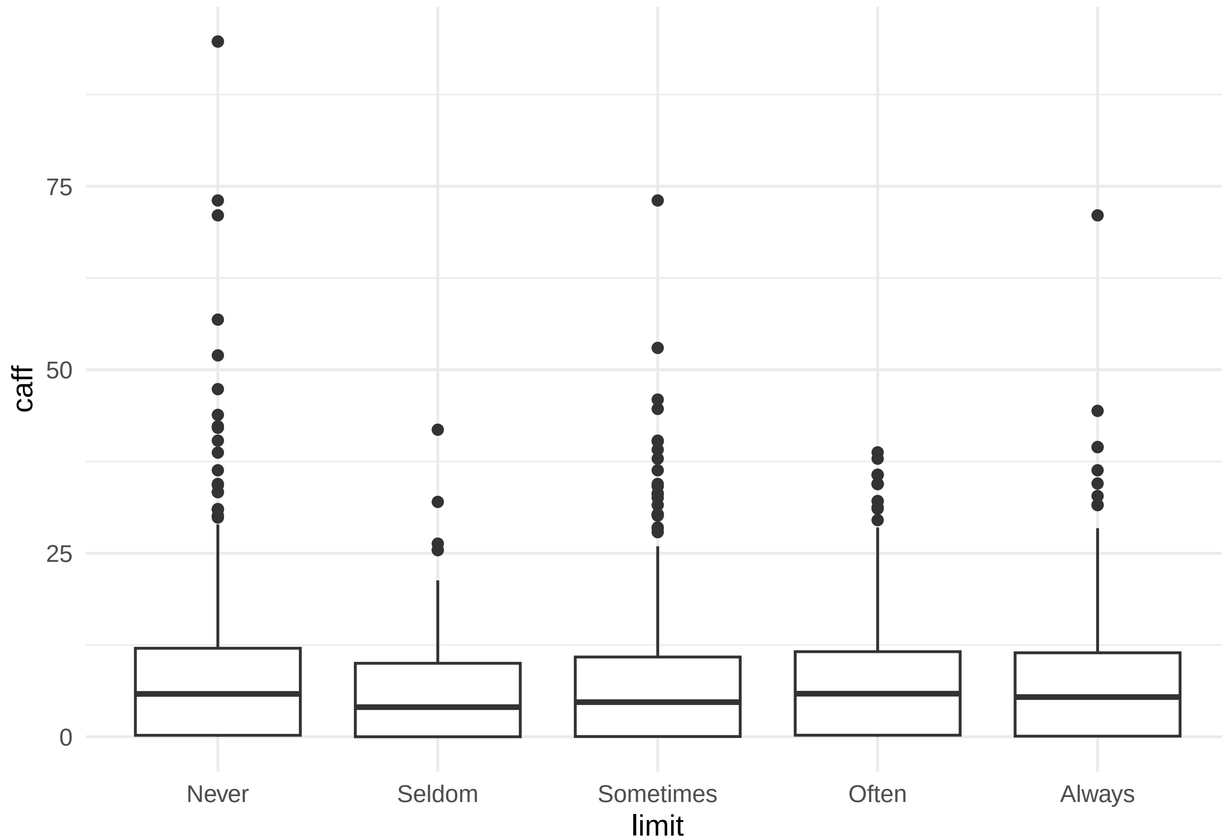


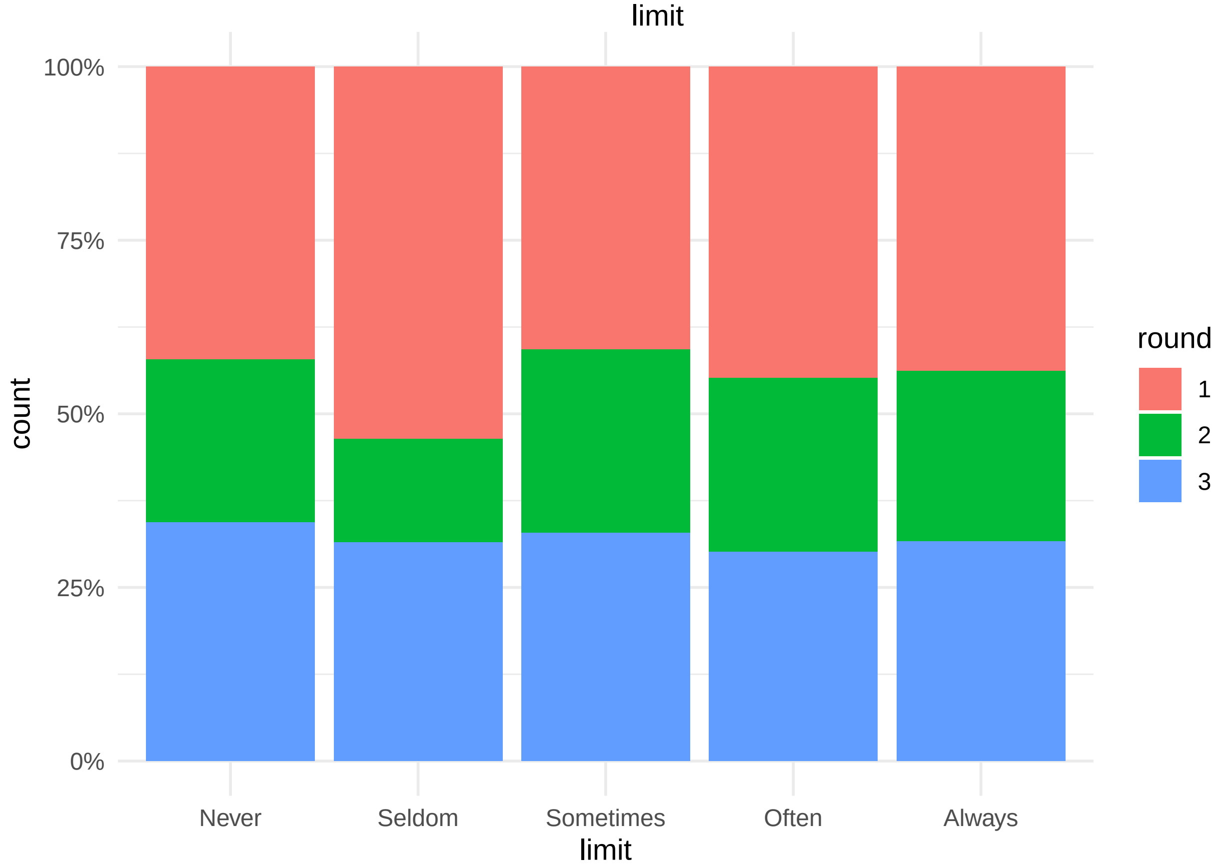

100%
75%
50%
25%
0%
count
Never
Seldom
Sometimes
Often
Always
limit
gender
F
M

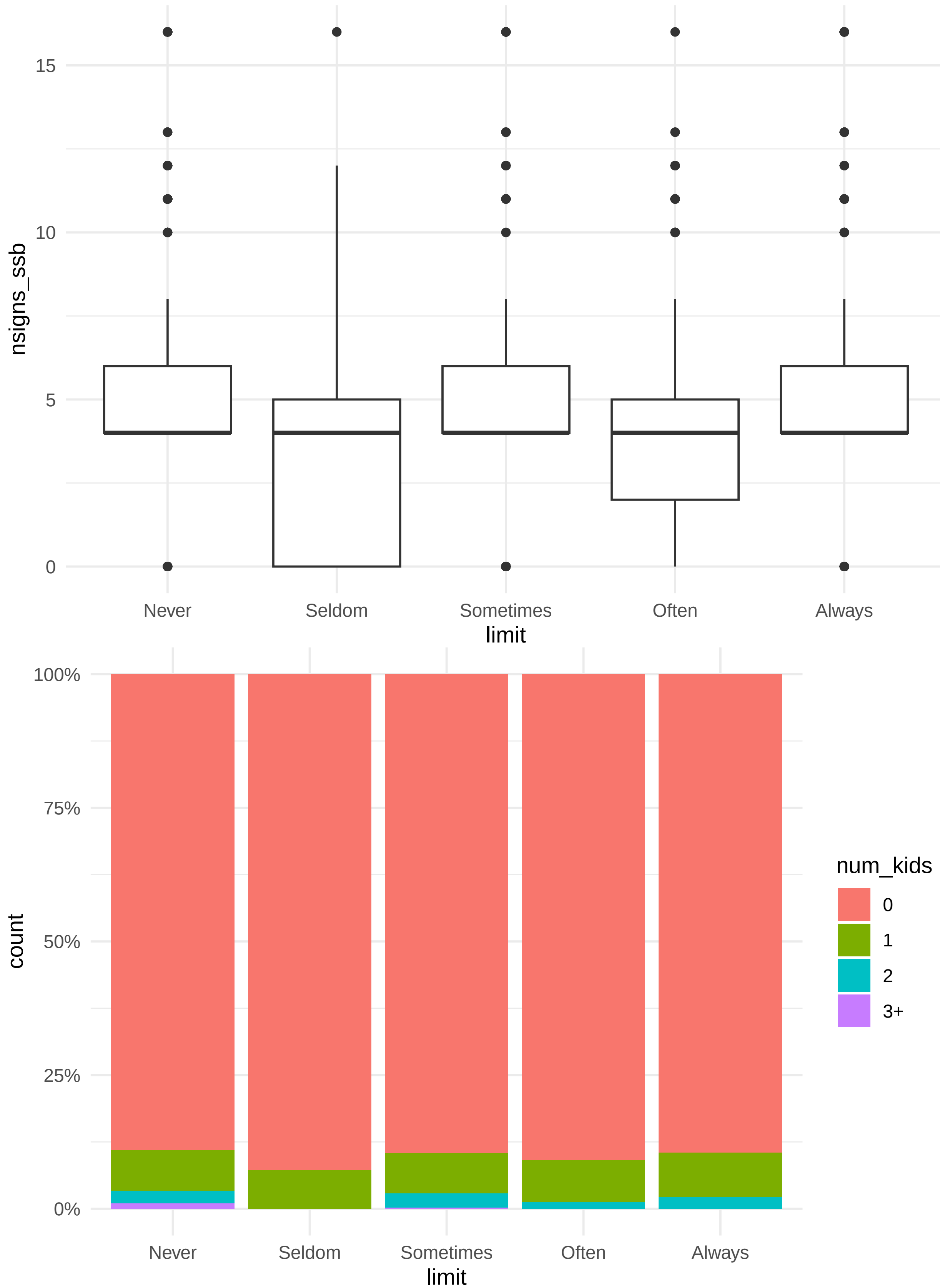

100%
75%
50%
25%
0%
count
Never
Seldom
Sometimes
Often
Always
limit
race
Asian
Black
Native
Other
White

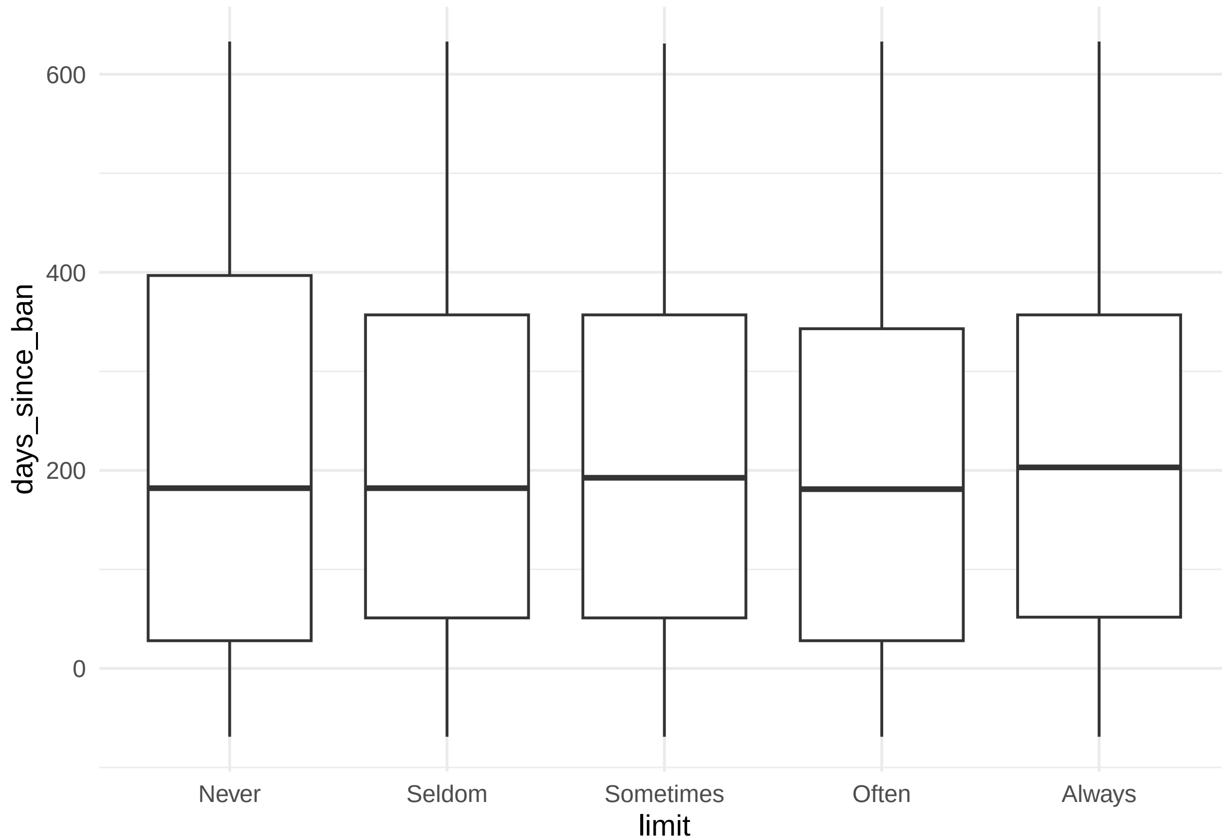

100%
75%
50%
25%
0%
count
Never
Seldom
Sometimes
Often
Always
limit
edu
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School

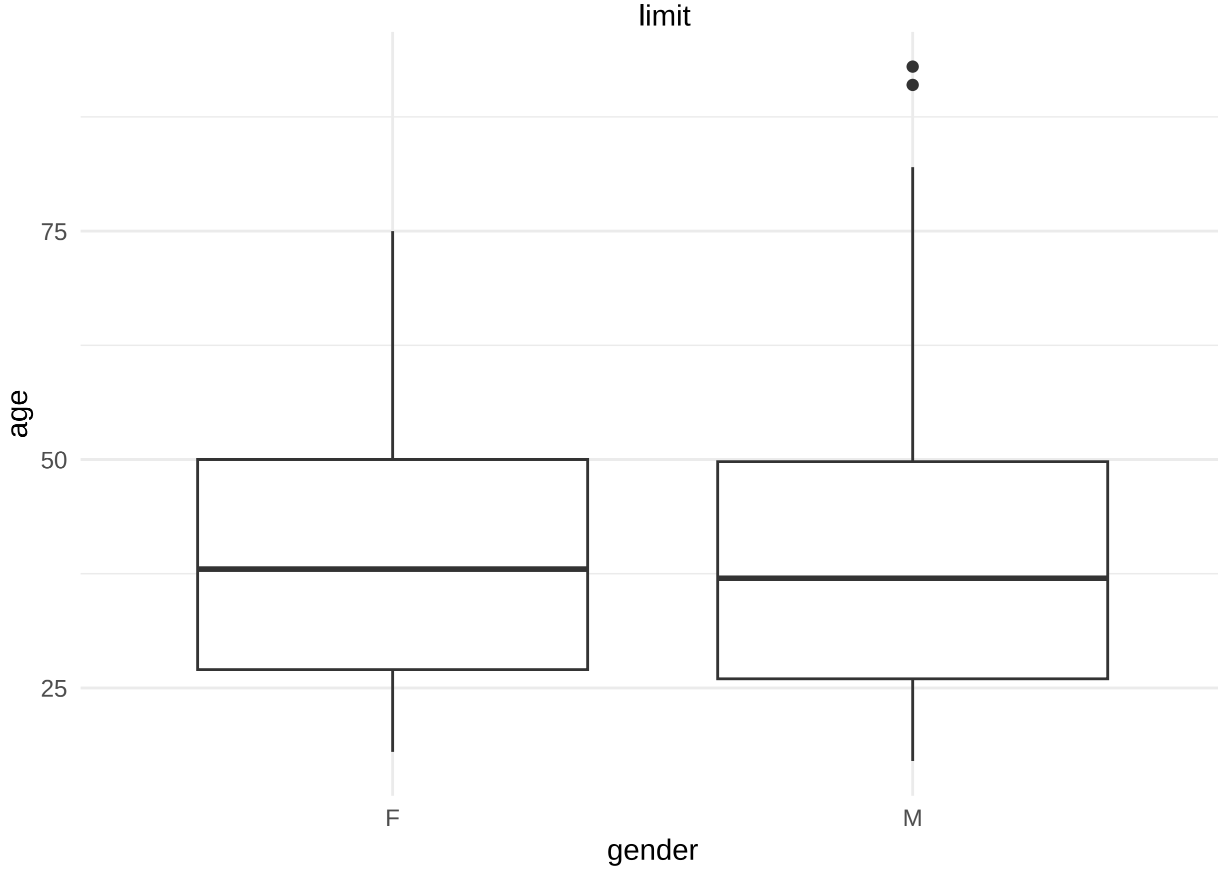

100%
75%
50%
25%
0%
count
Never
Seldom
Sometimes
Often
Always
limit
city
New Jersey
New York

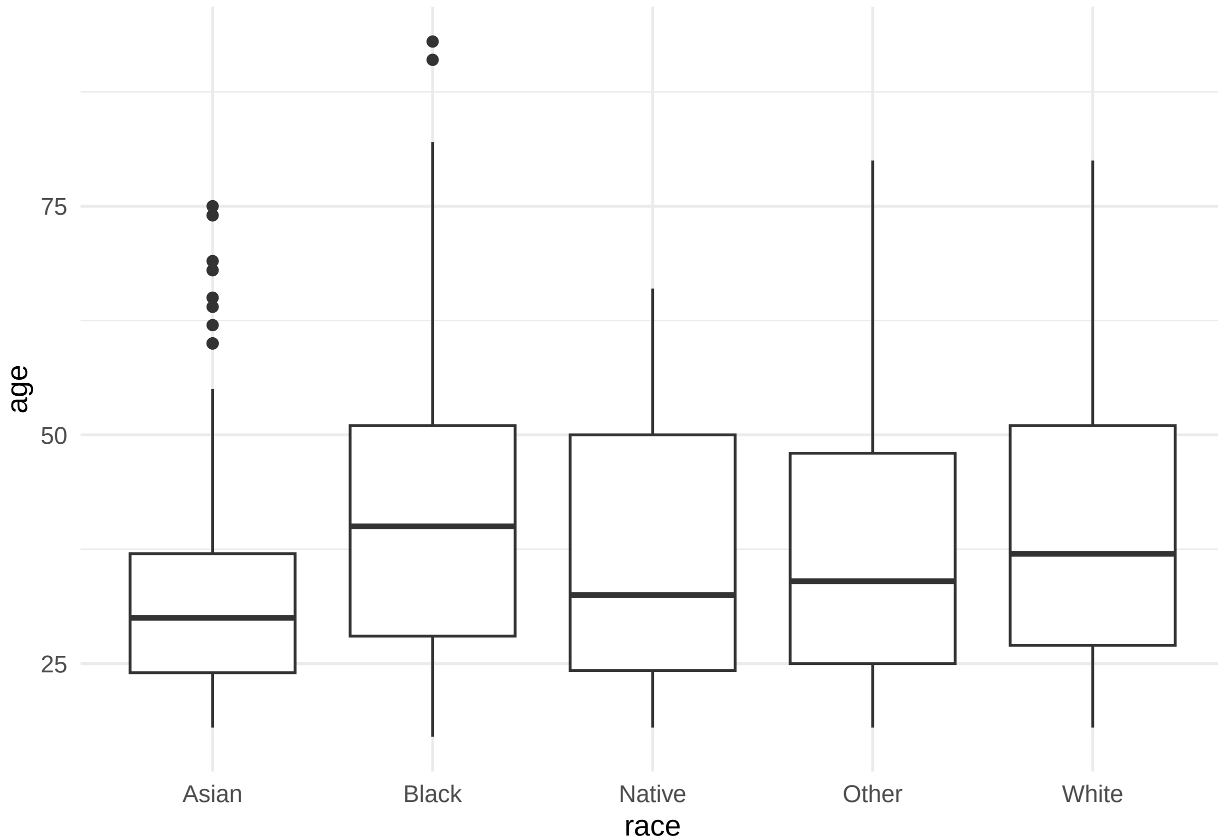

caff
75
50
25
0
Never
Seldom
Sometimes
Often
Always
limit

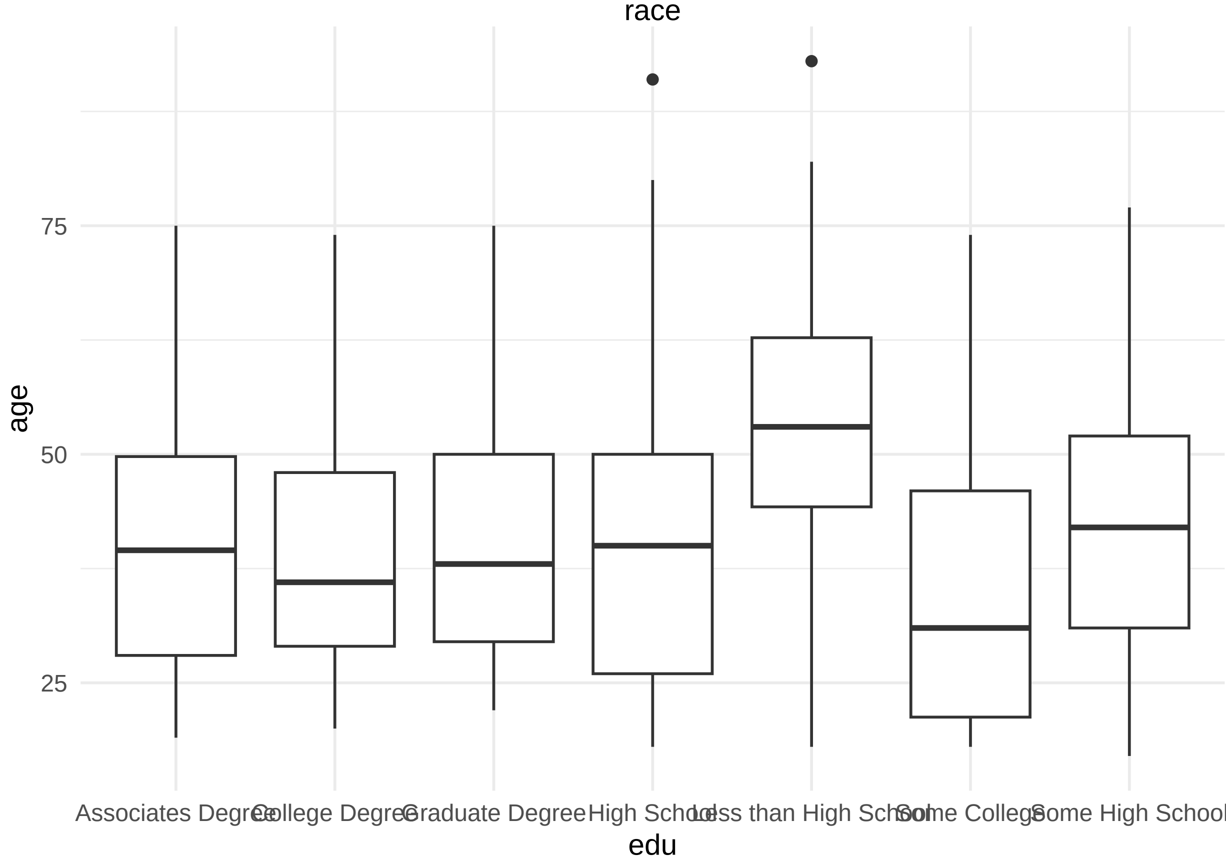

count
100%
75%
50%
25%
0%
Never
Seldom
Sometimes
Often
Always
limit
round
1
2
3

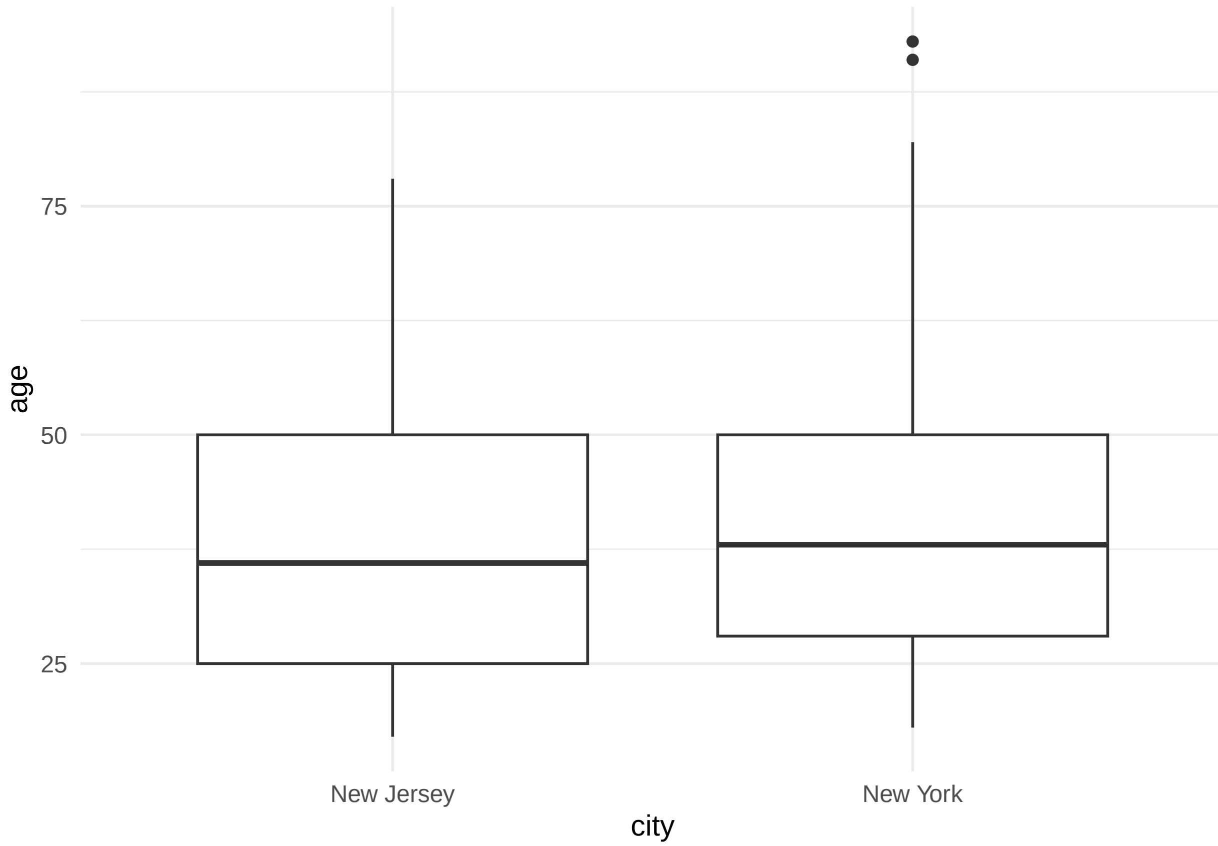

15
10
5
0
nsigns_ssb
Never
Seldom
Sometimes
Often
Always
limit
100%
75%
50%
25%
0%
count
Never
Seldom
Sometimes
Often
Always
limit
num_kids
0
1
2
3+


days_since_ban
600
400
200
0
Never
Seldom
Sometimes
Often
Always
limit

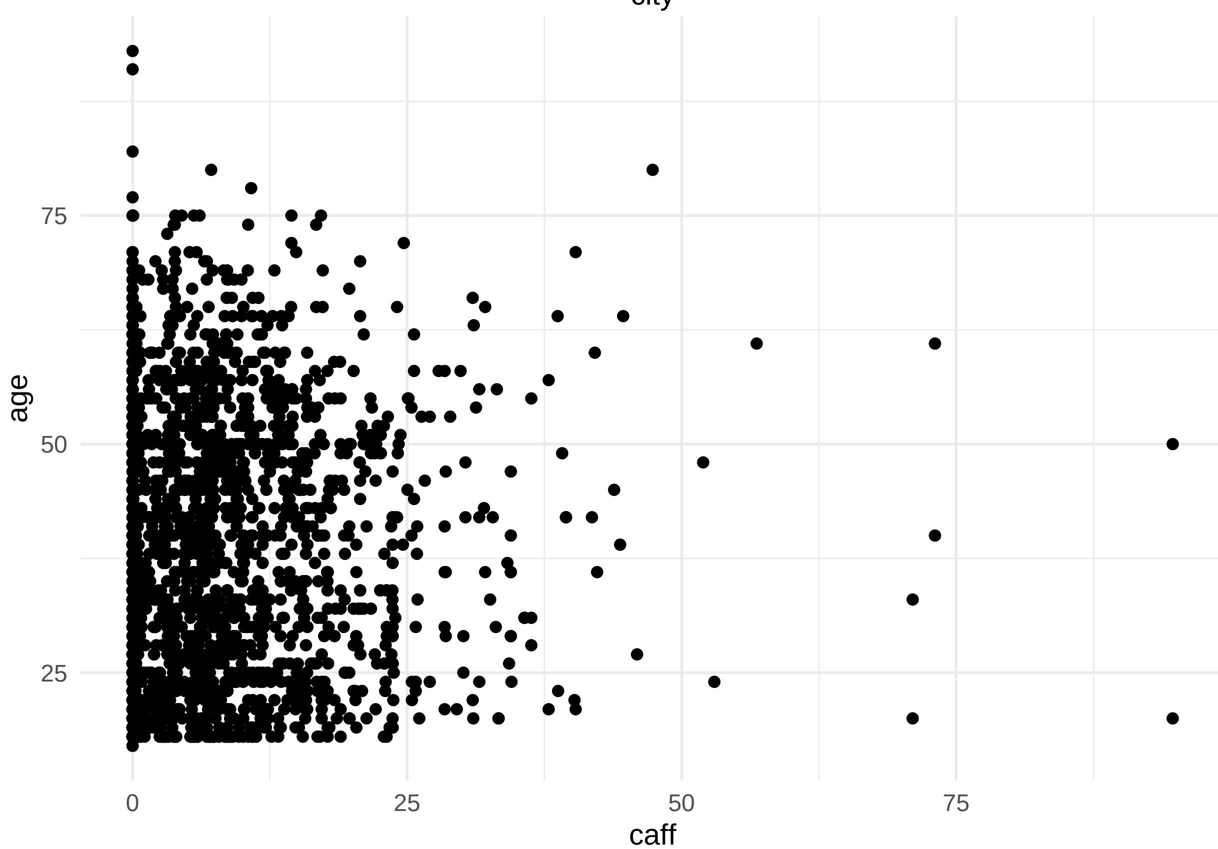

age
75
50
25
F
M
gender


75
50
25
age
Asian
Black
Native
Other
White
race

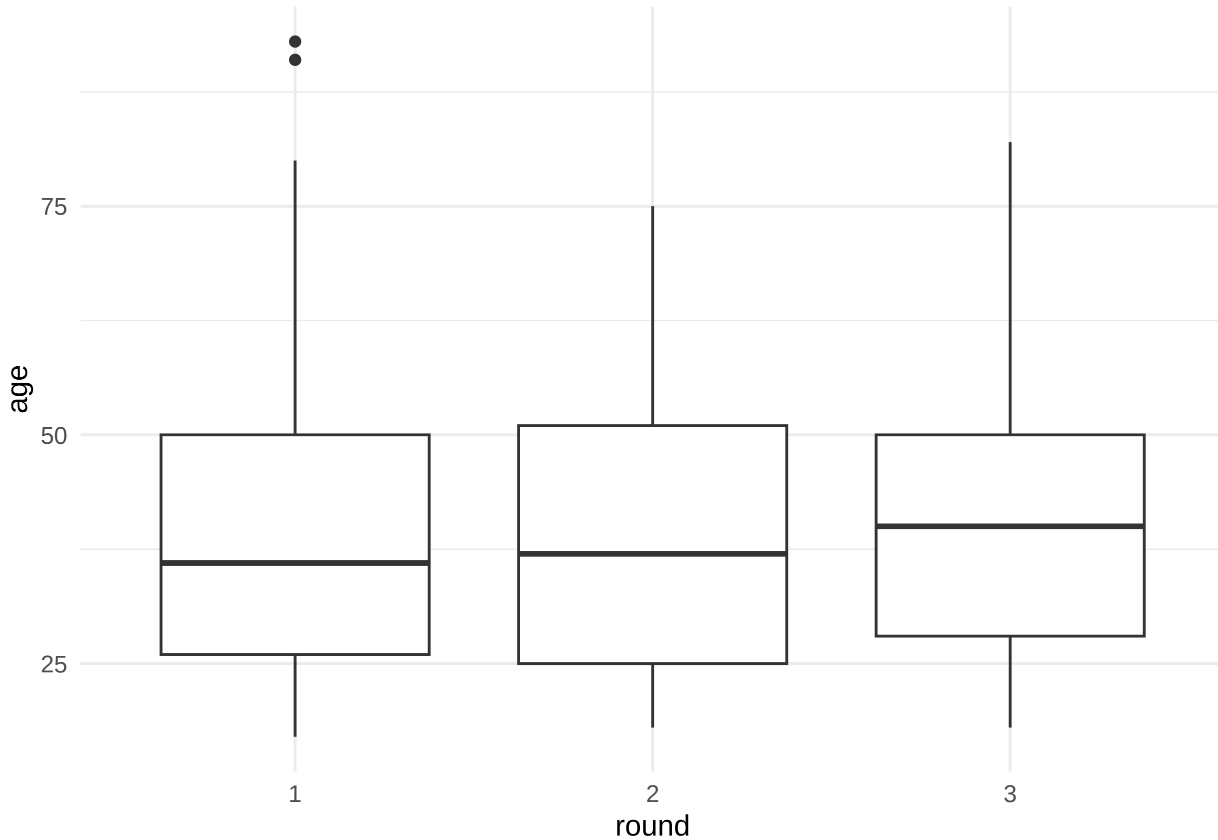

75
50
25
age
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu

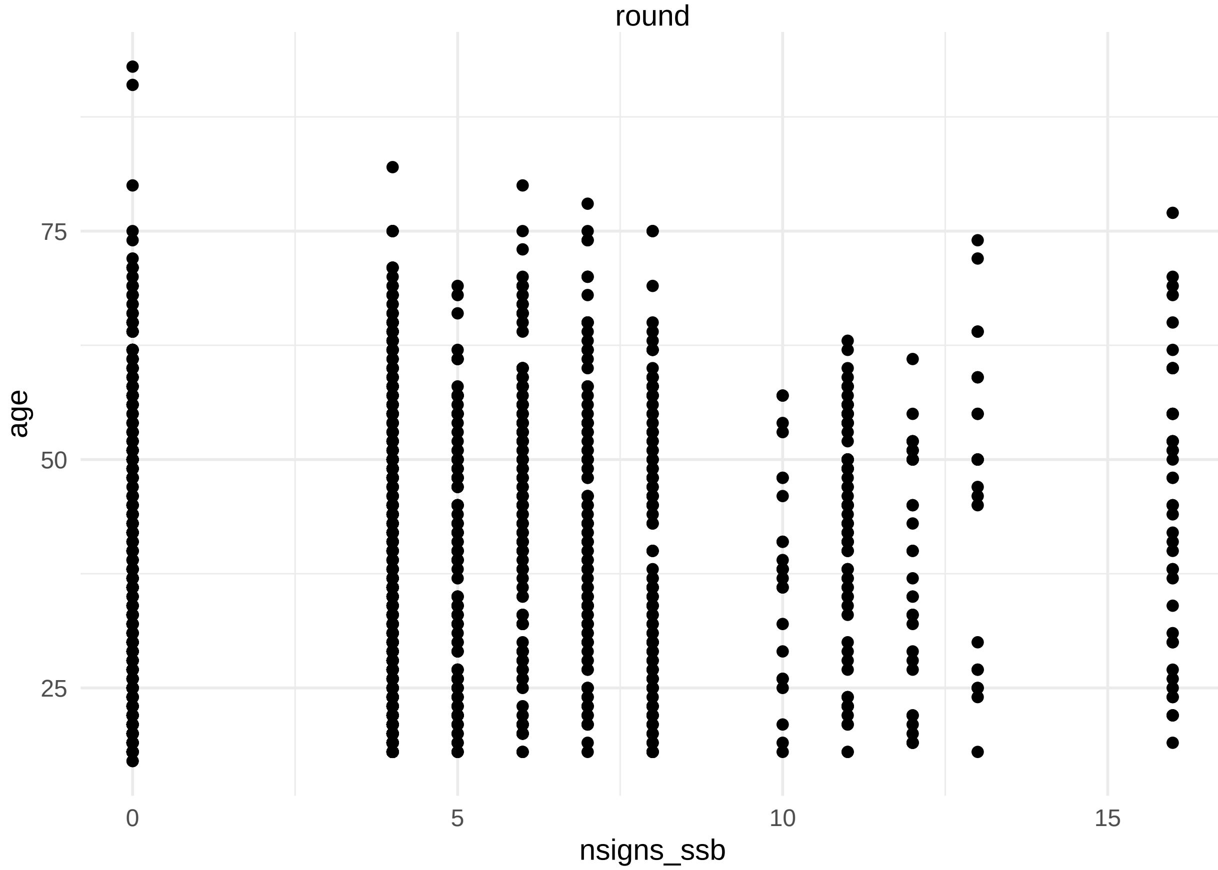


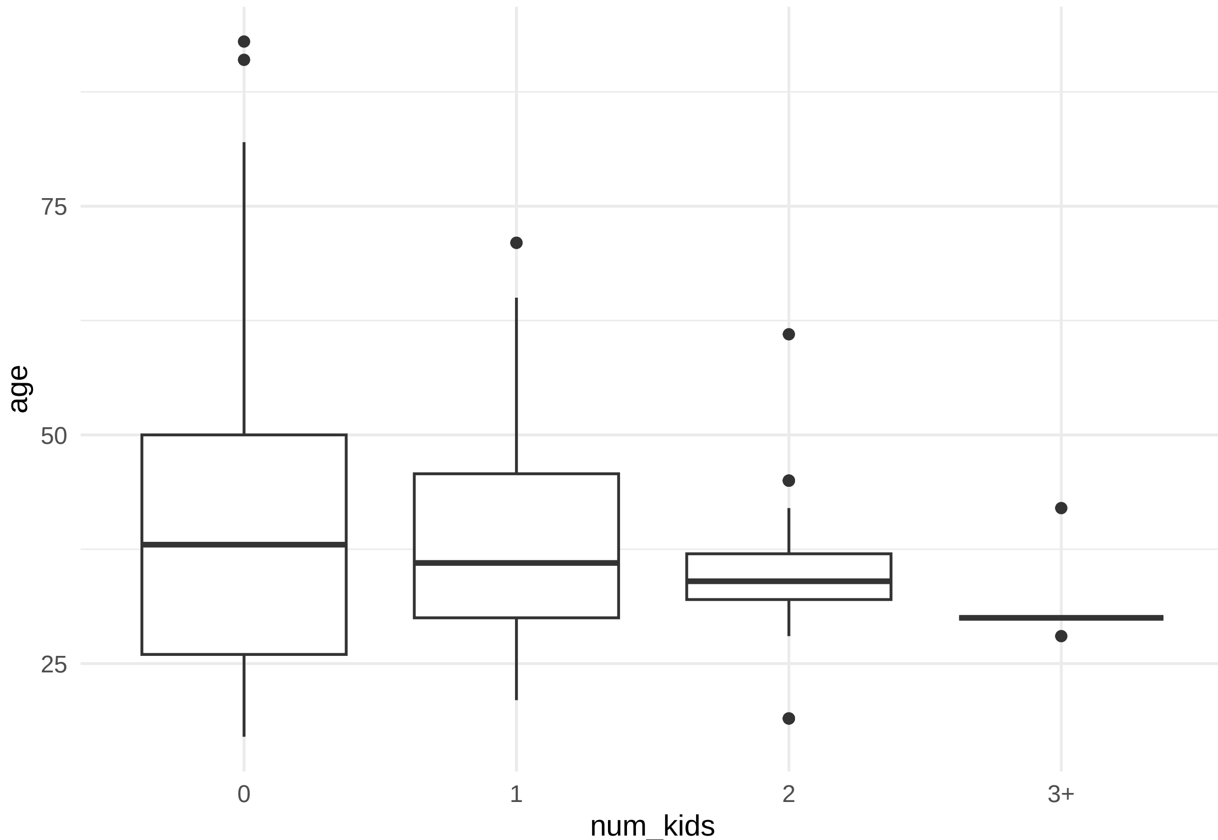
75
50
25
age
New Jersey
New York
city



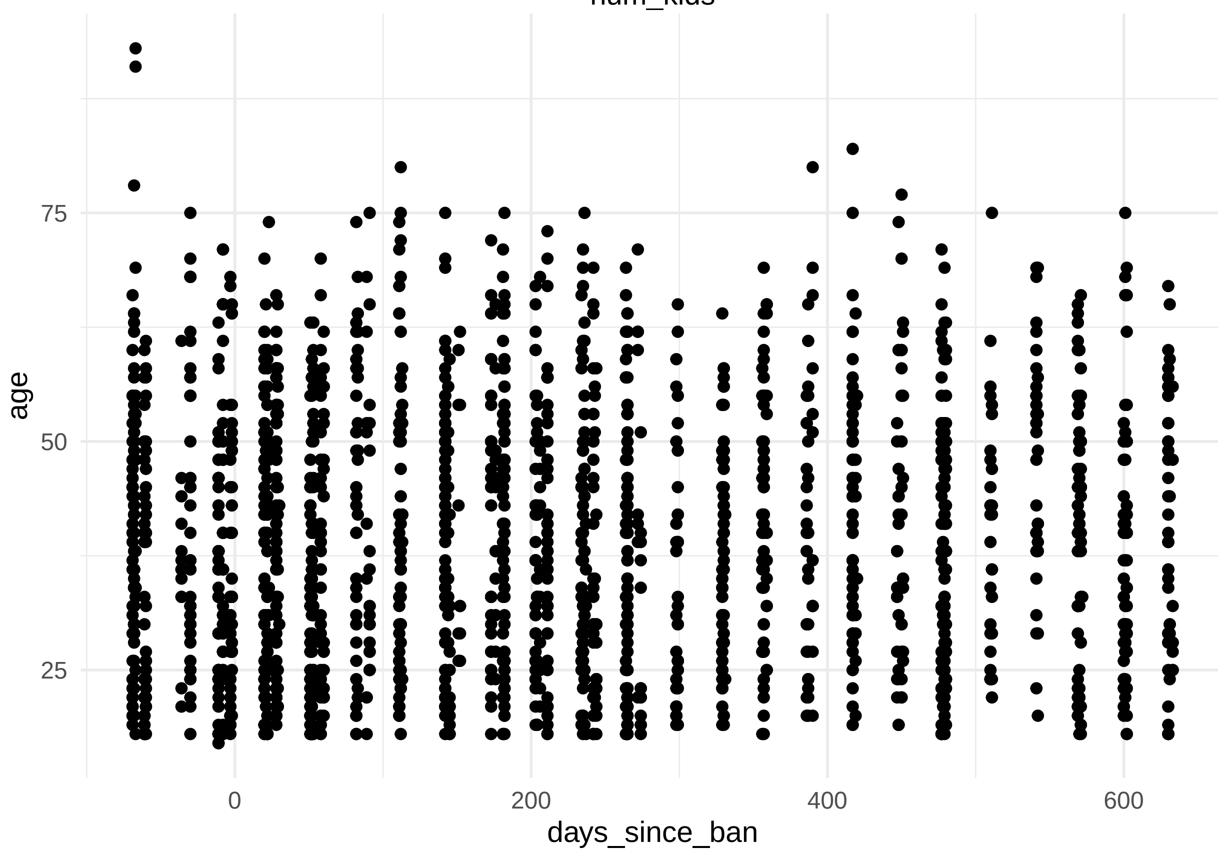
75
50
25
age
0
25
50
75
caff


age
75
50
25
1
2
3
round

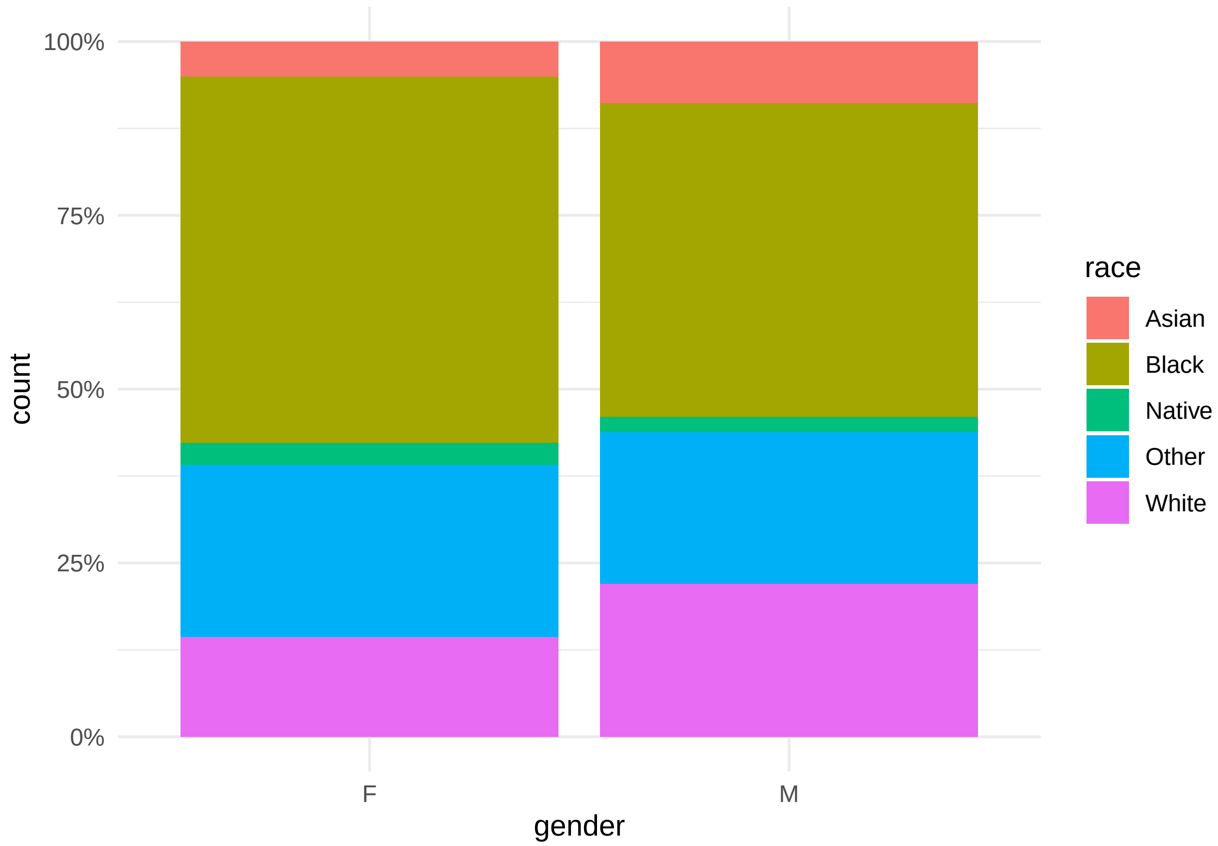

age
75
50
25
0
5
10
15
nsigns_ssb

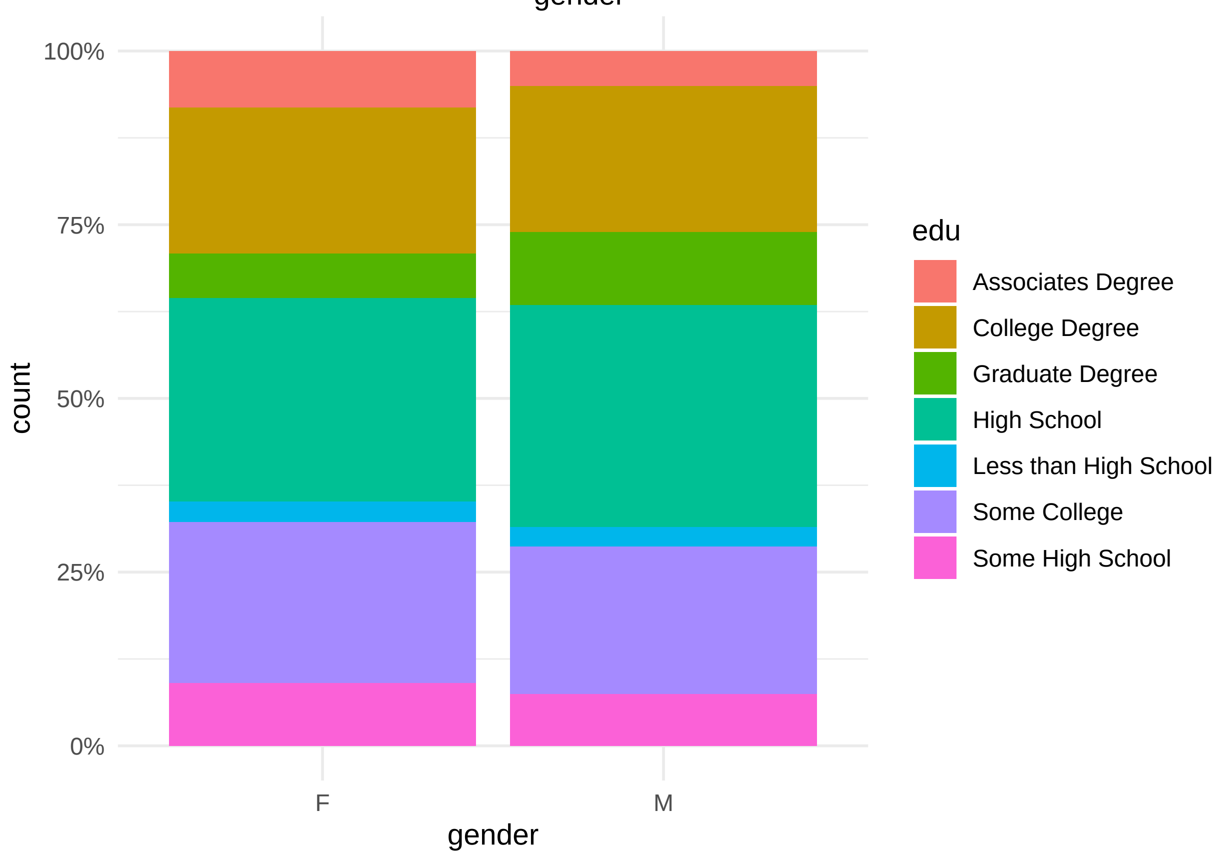


age
75
50
25
0
1
2
3+
num_kids

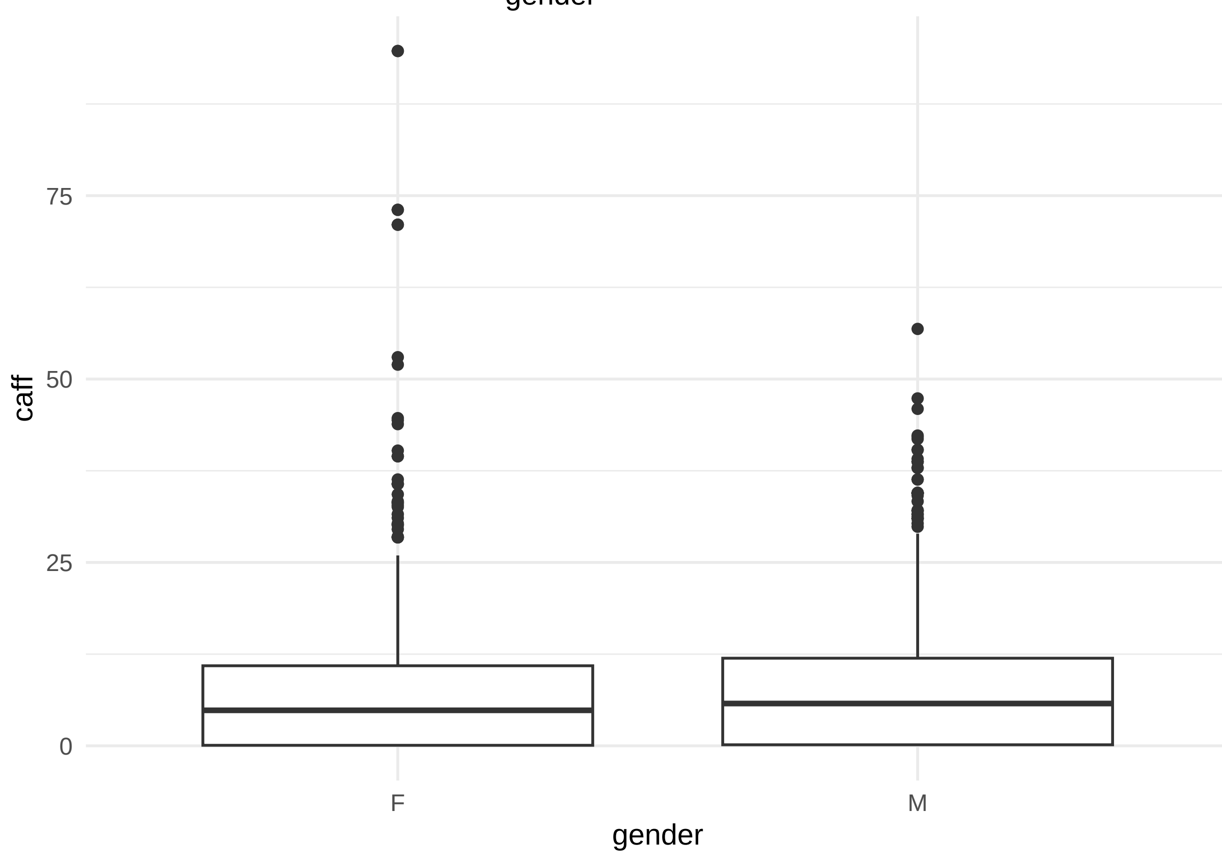


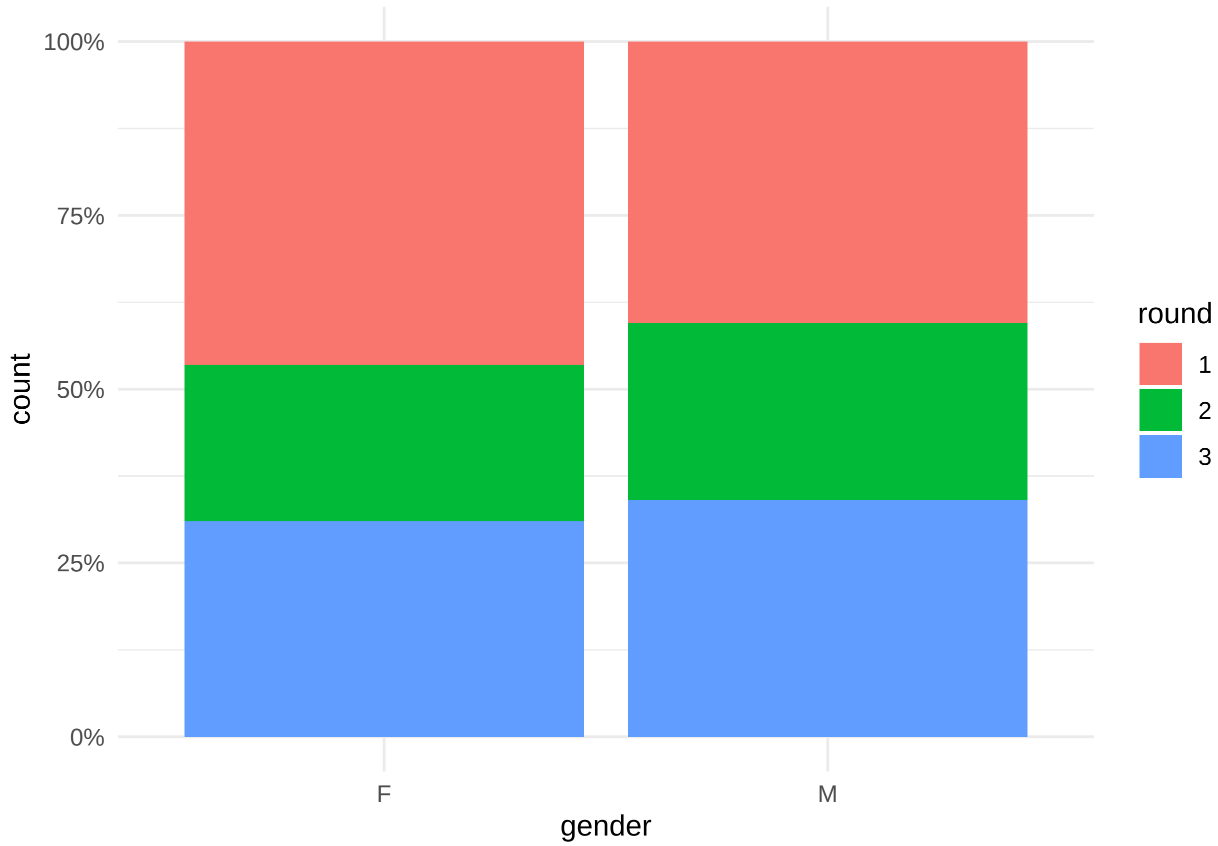
age
75
50
25
0
200
400
600
days_since_ban

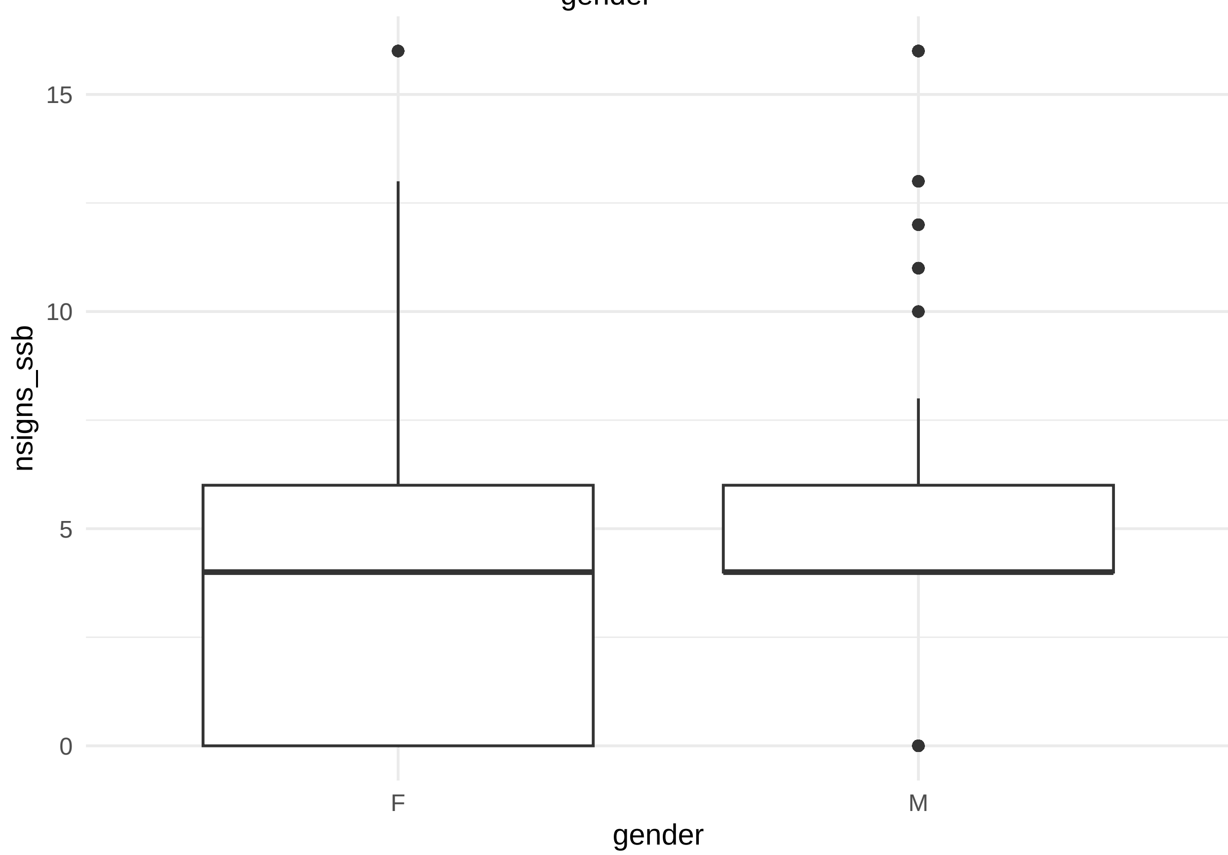

100%
75%
50%
25%
0%
count
F
M
gender
race
Asian
Black
Native
Other
White

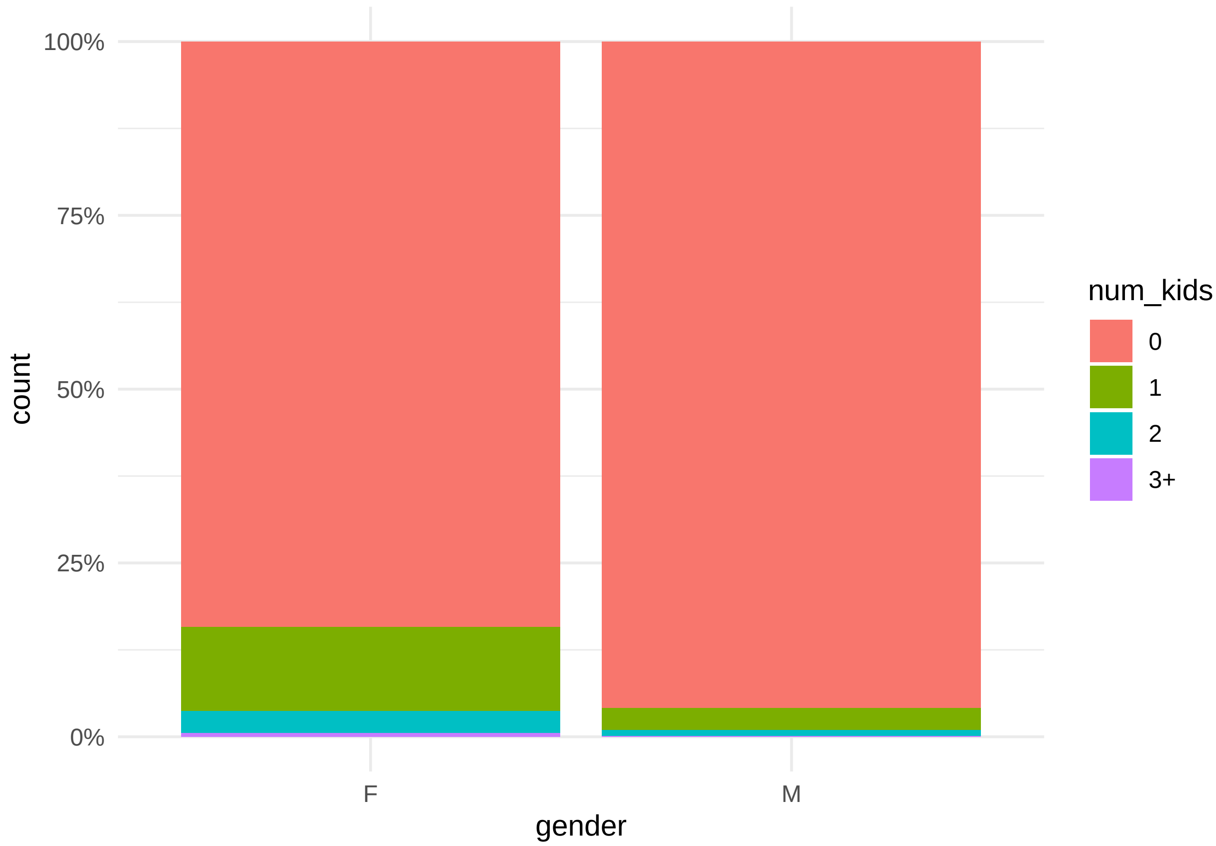

100%
75%
50%
25%
0%
count
F
M
gender
edu
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School

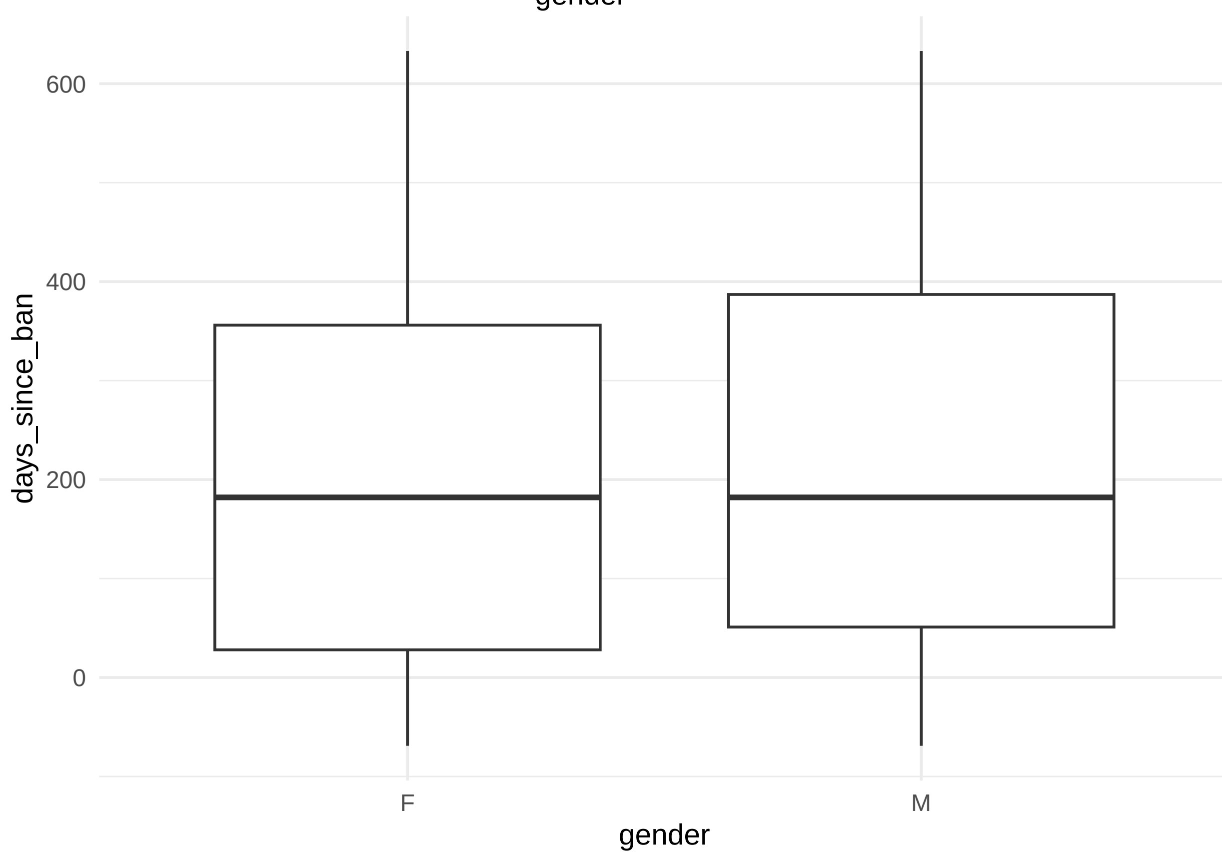

100%
75%
50%
25%
0%
count
F
M
gender
city
New Jersey
New York

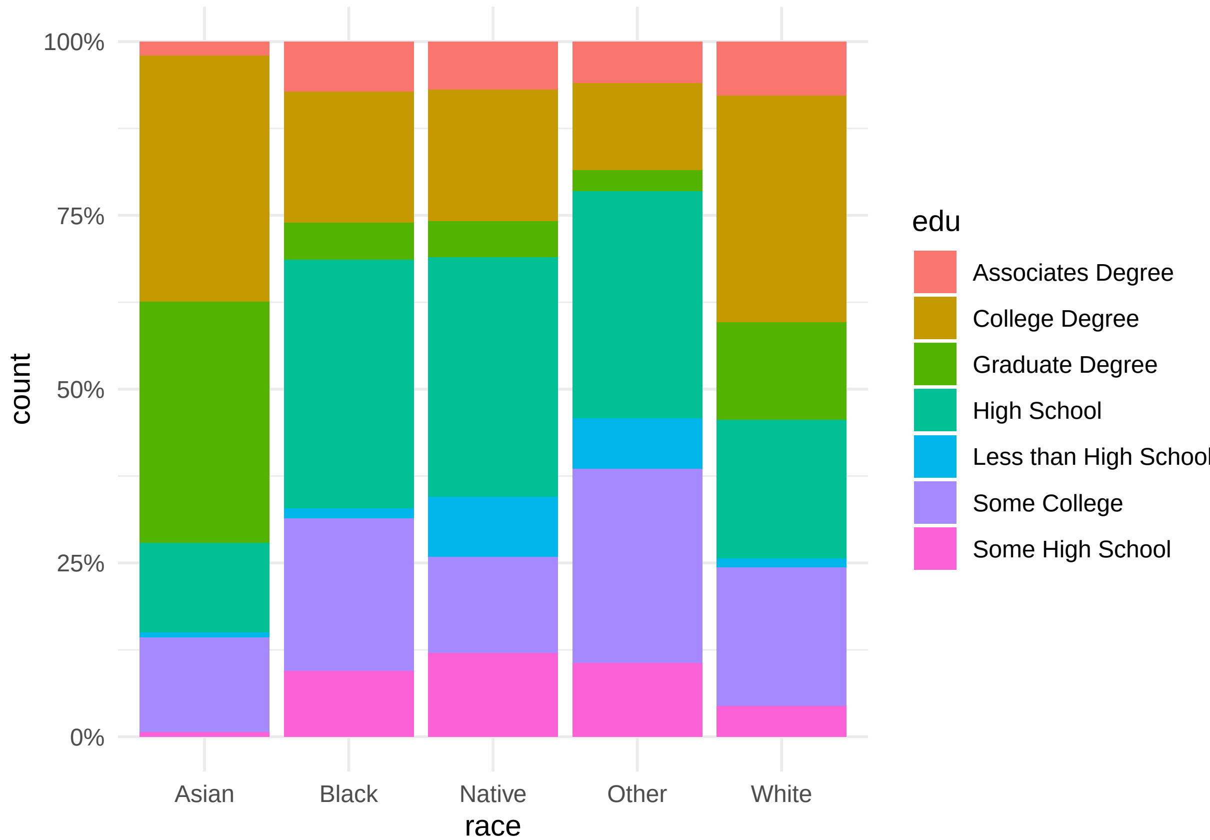

75
50
25
0
caff
F
M
gender

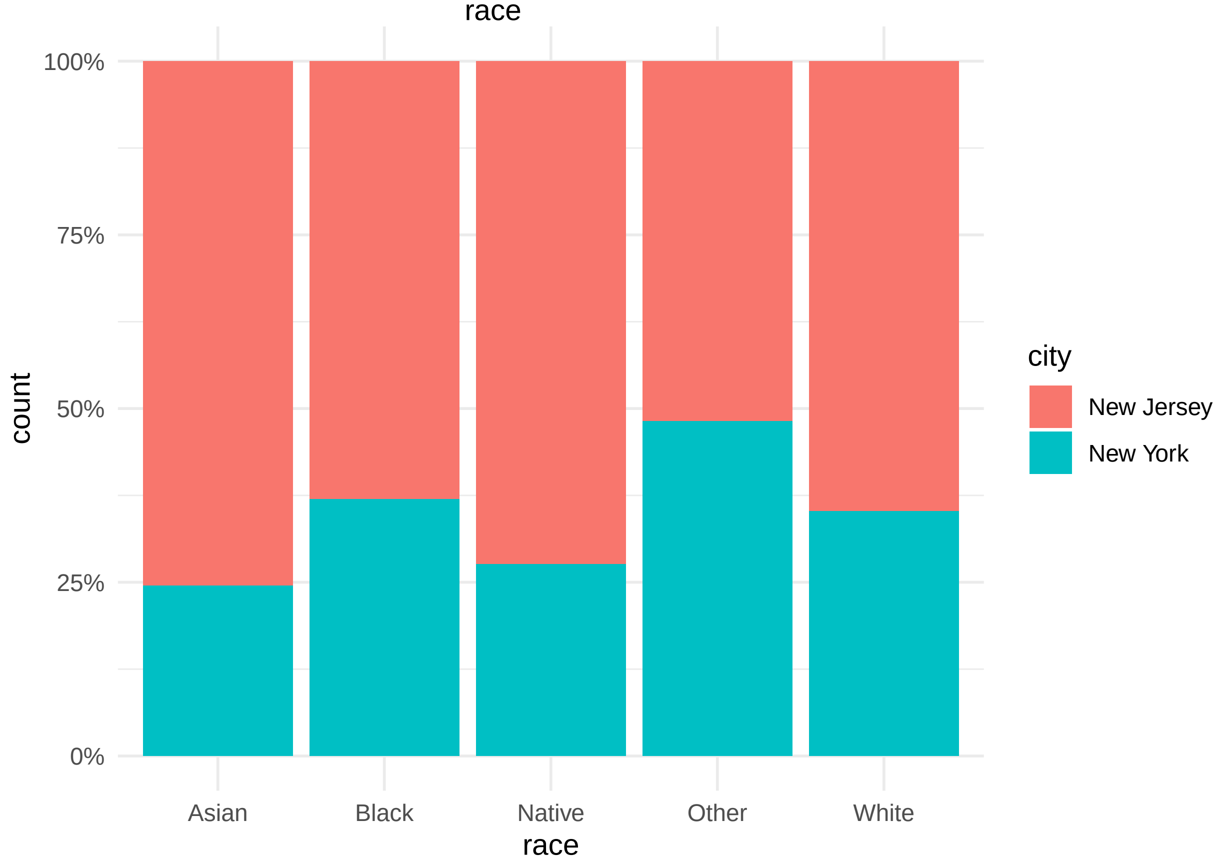

100%
75%
50%
25%
0%
count
F
M
gender
round
1
2
3

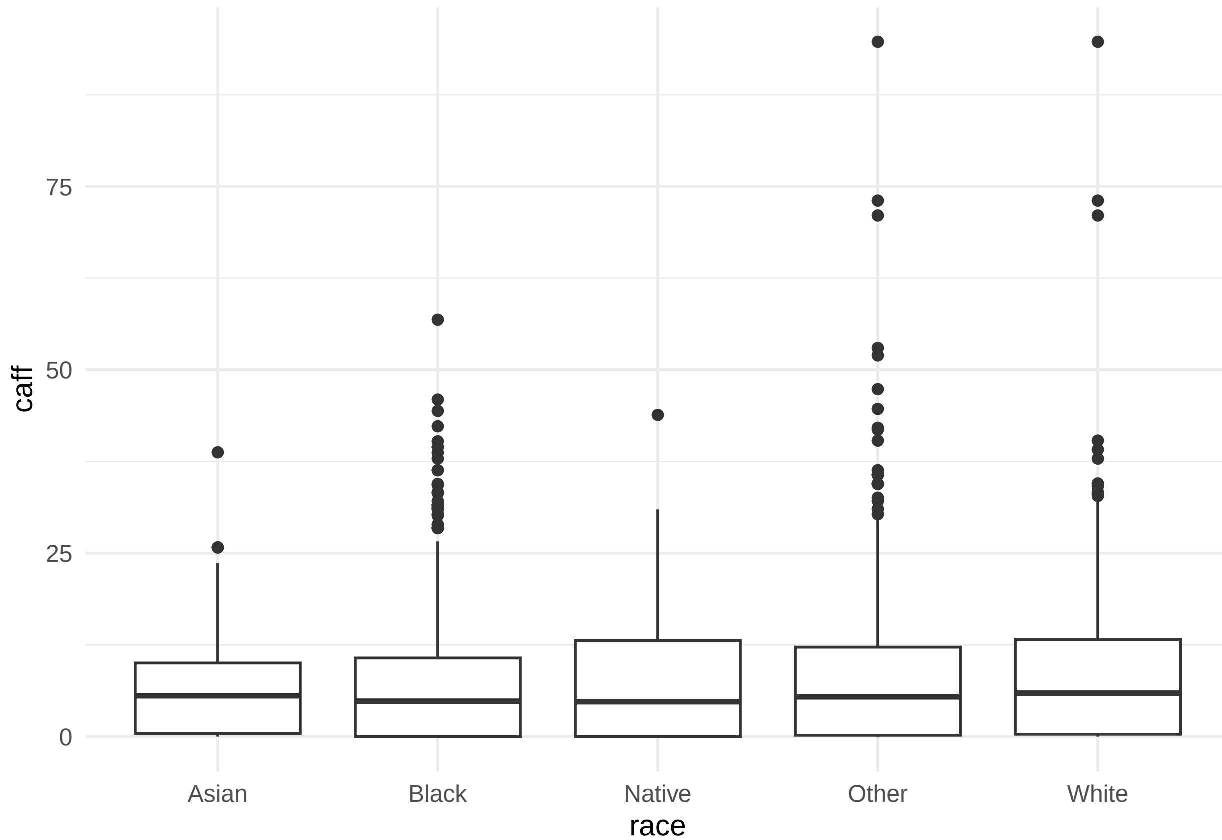

15
10
5
0
nsigns_ssb
F
M
gender


100%
75%
50%
25%
0%
count
F
M
gender
num_kids
0
1
2
3+

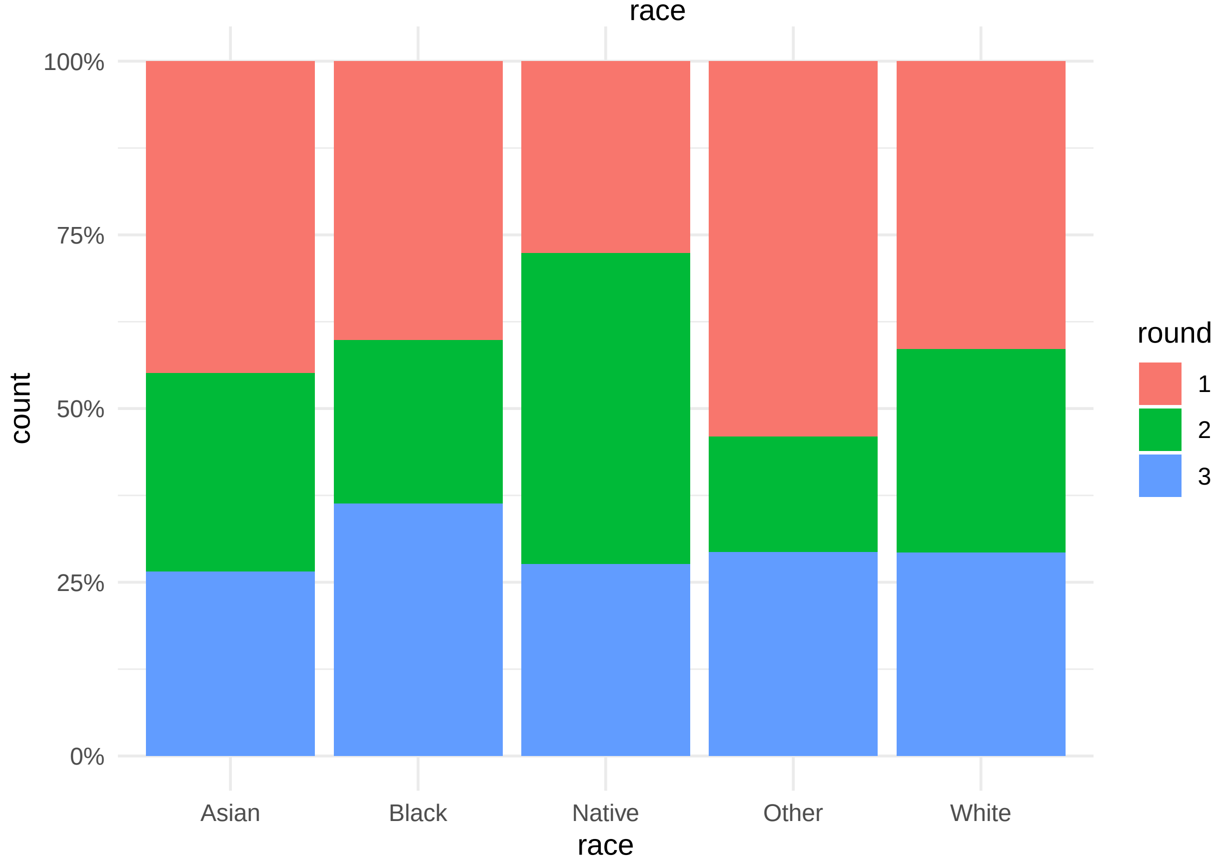

600
400
200
0
days_since_ban
F
M
gender


100%
75%
50%
25%
0%
count
Asian
Black
Native
Other
White
race
edu
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School

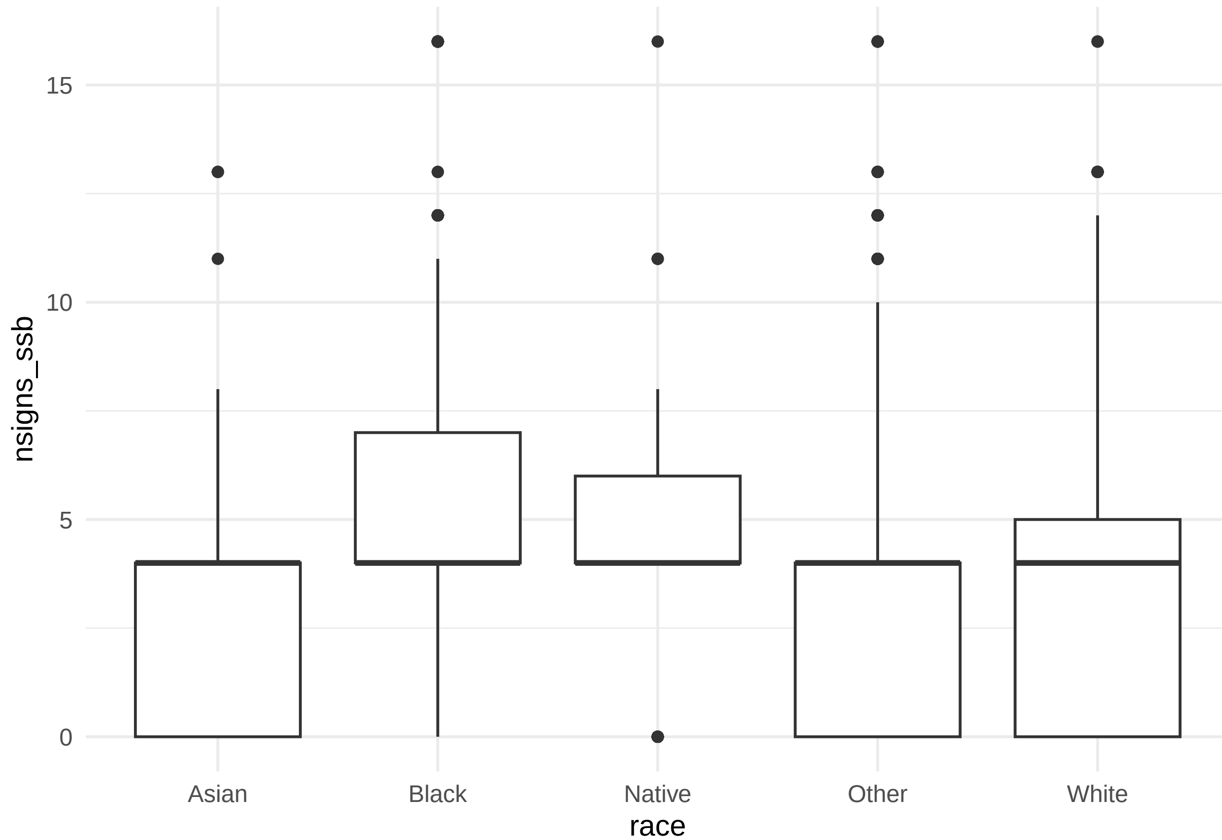

100%
75%
50%
25%
0%
count
Asian
Black
Native
Other
White
race
city
New Jersey
New York

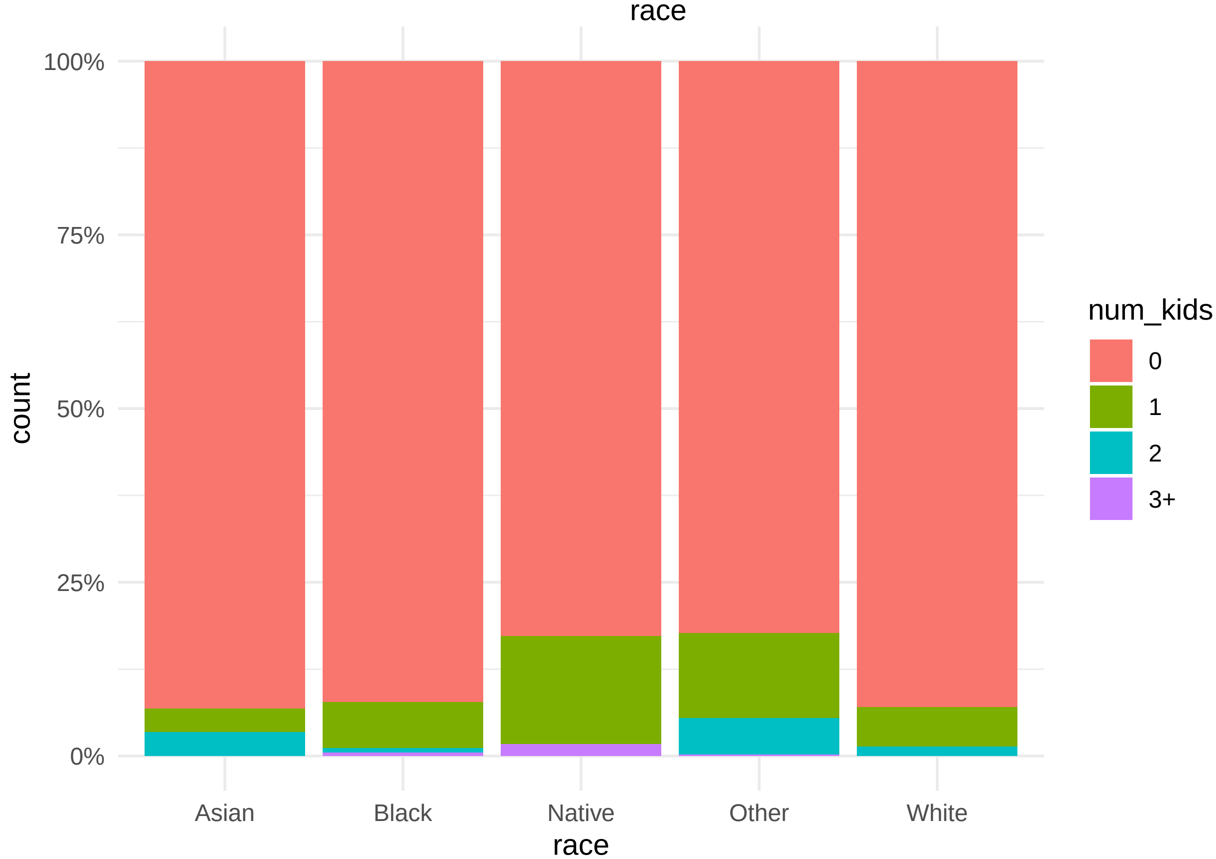


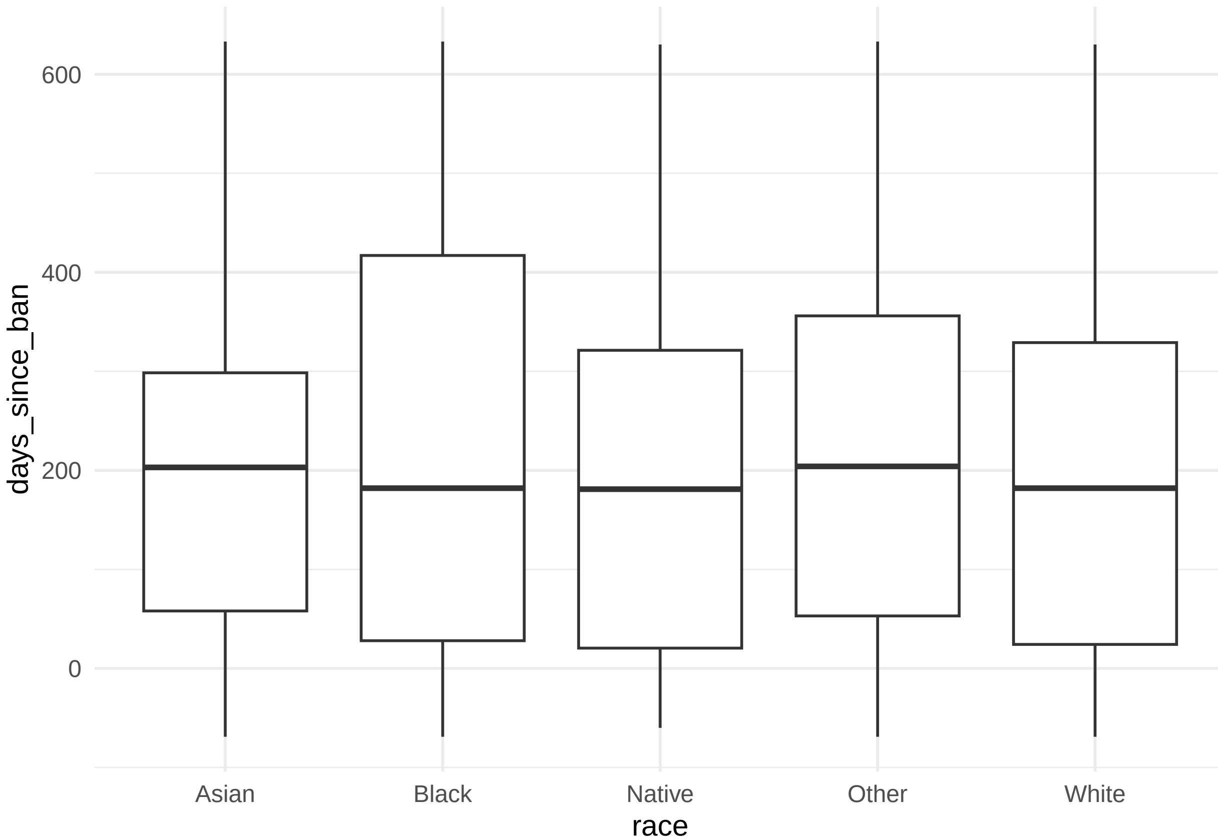
75
50
25
0
caff
Asian
Black
Native
Other
White
race

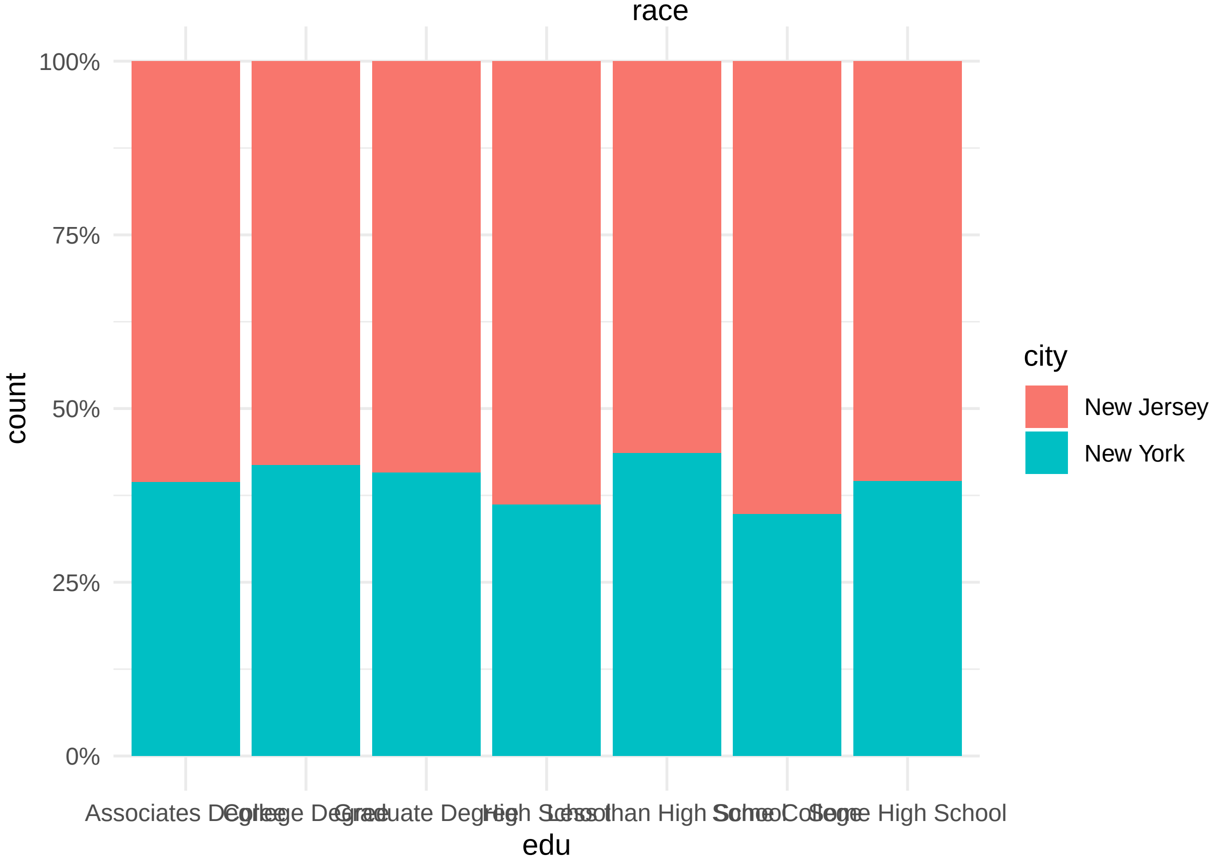


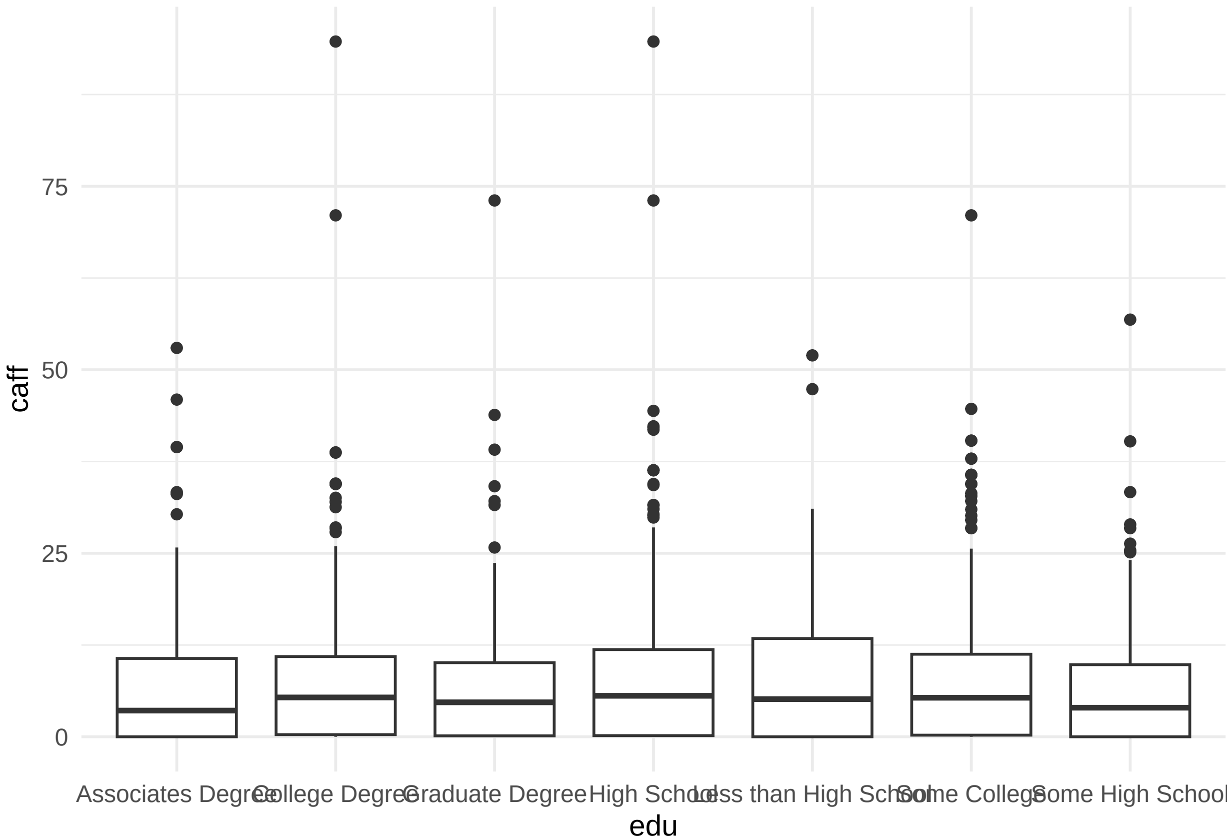
100%
75%
50%
25%
0%
count
Asian
Black
Native
Other
White
race
round
1
2
3

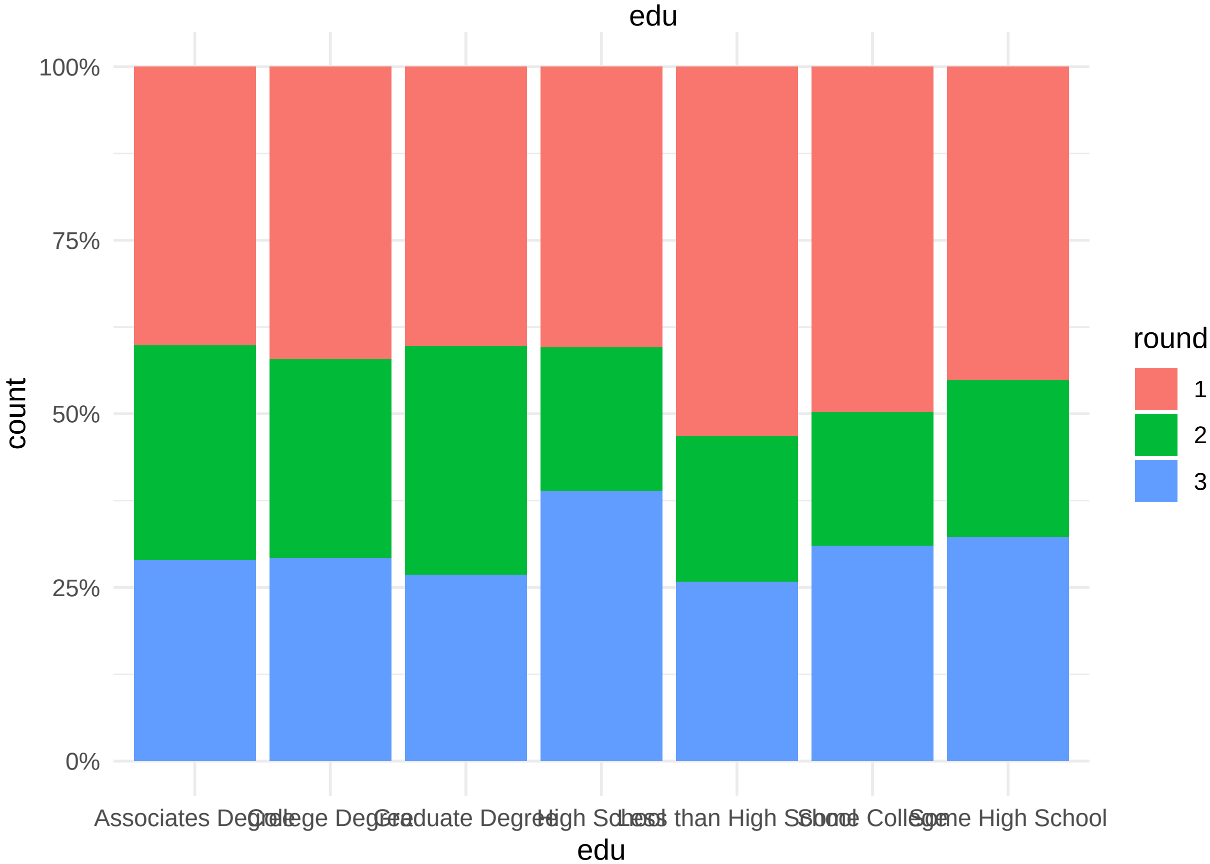

nsigns_ssb
15
10
5
0
Asian
Black
Native
Other
White
race

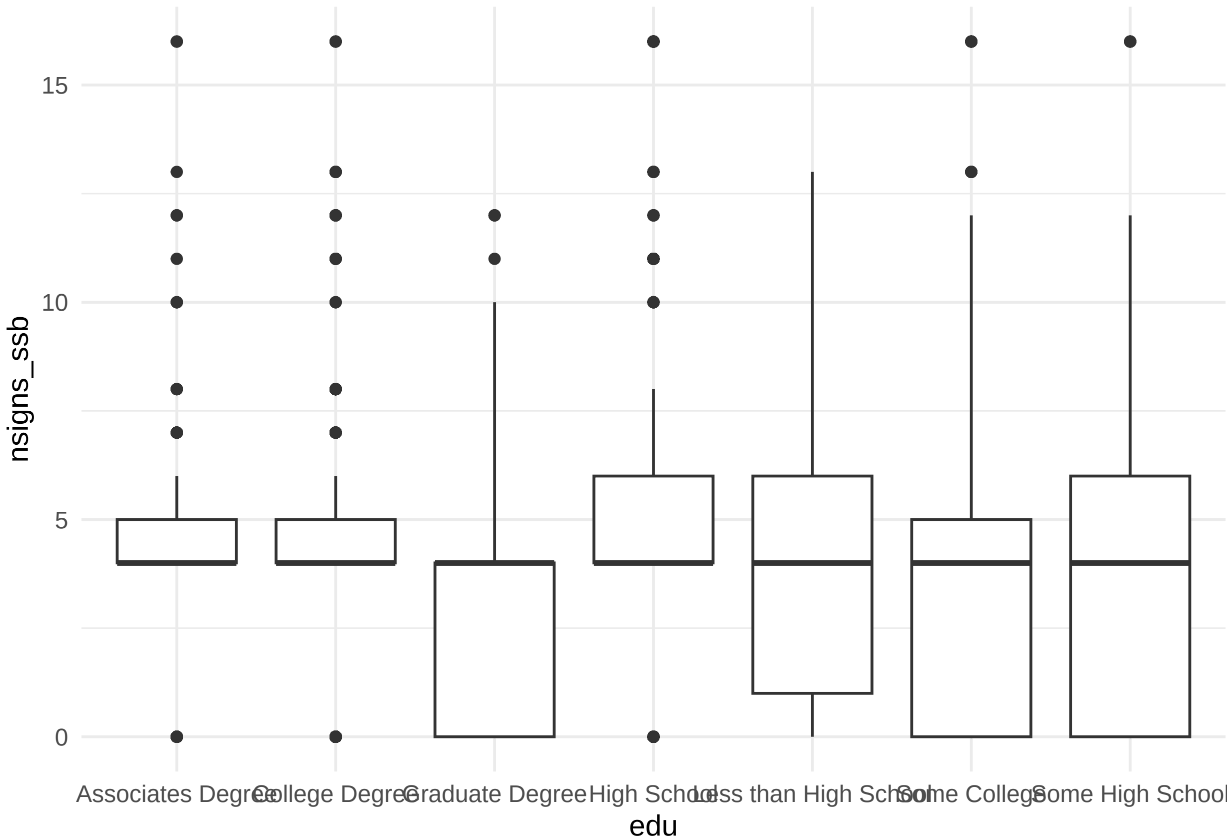

count
100%
75%
50%
25%
0%
Asian
Black
Native
Other
White
race
num_kids
0
1
2
3+

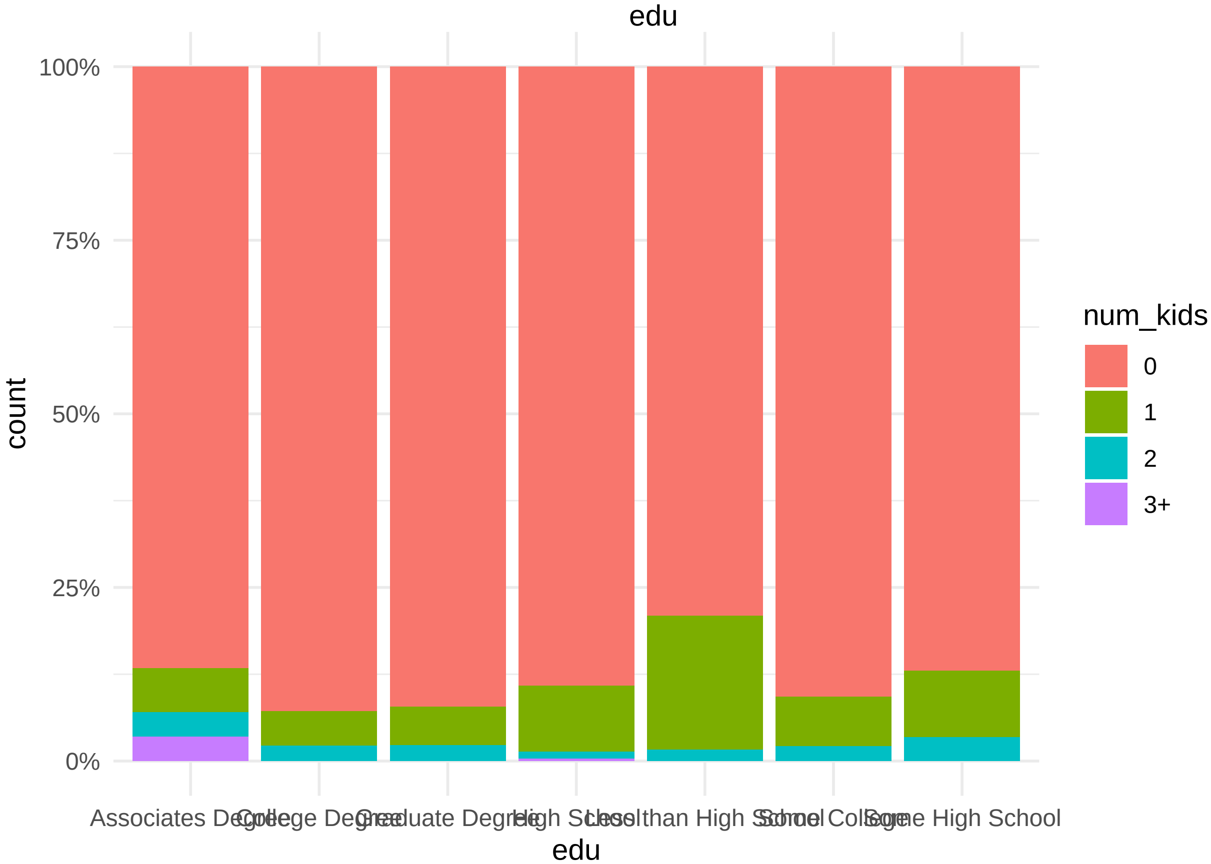

600
400
200
0
days_since_ban
Asian
Black
Native
Other
White
race

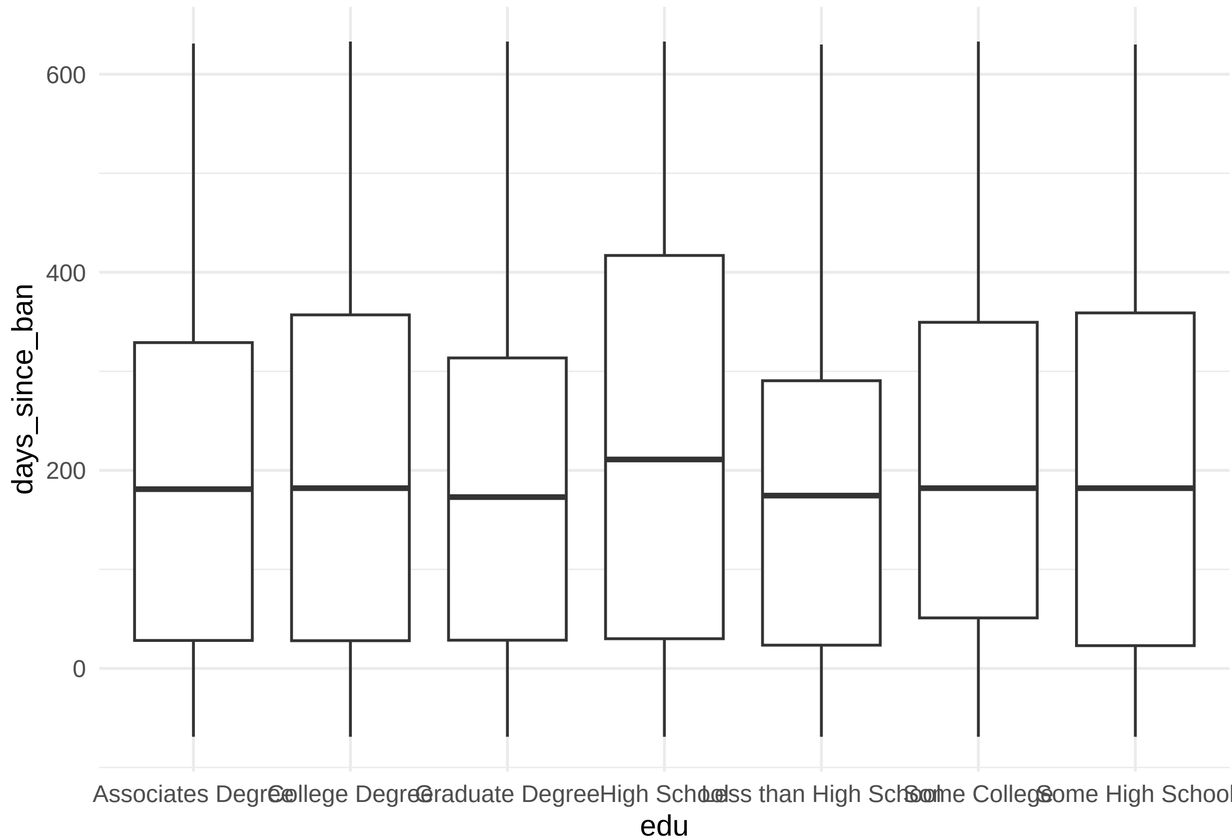

100%
75%
50%
25%
0%
count
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu
city
New Jersey
New York


caff
0
25
50
75
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu

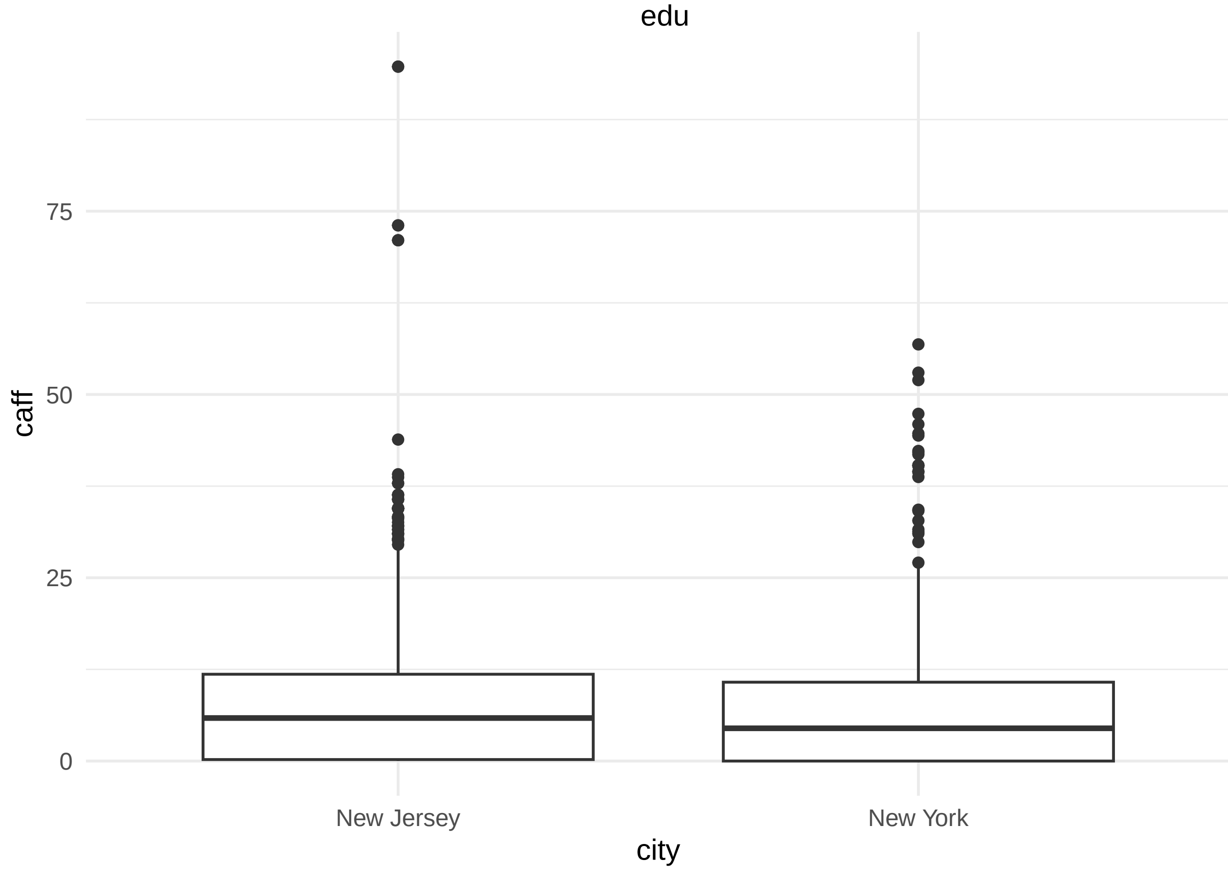

count
0%
25%
50%
75%
100%
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu
round
1
2
3


nsigns_ssb
15
10
5
0
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu

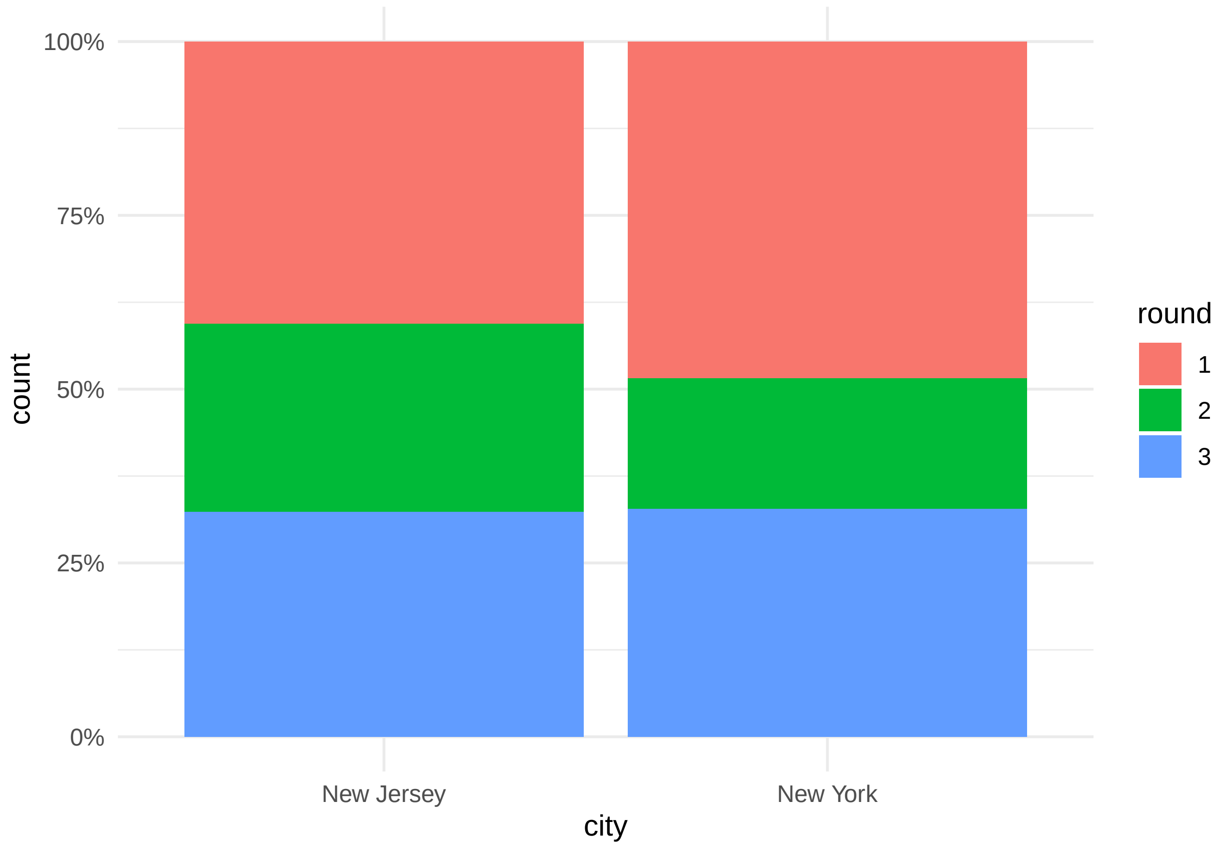

count
100%
75%
50%
25%
0%
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu
num_kids
0
1
2
3+

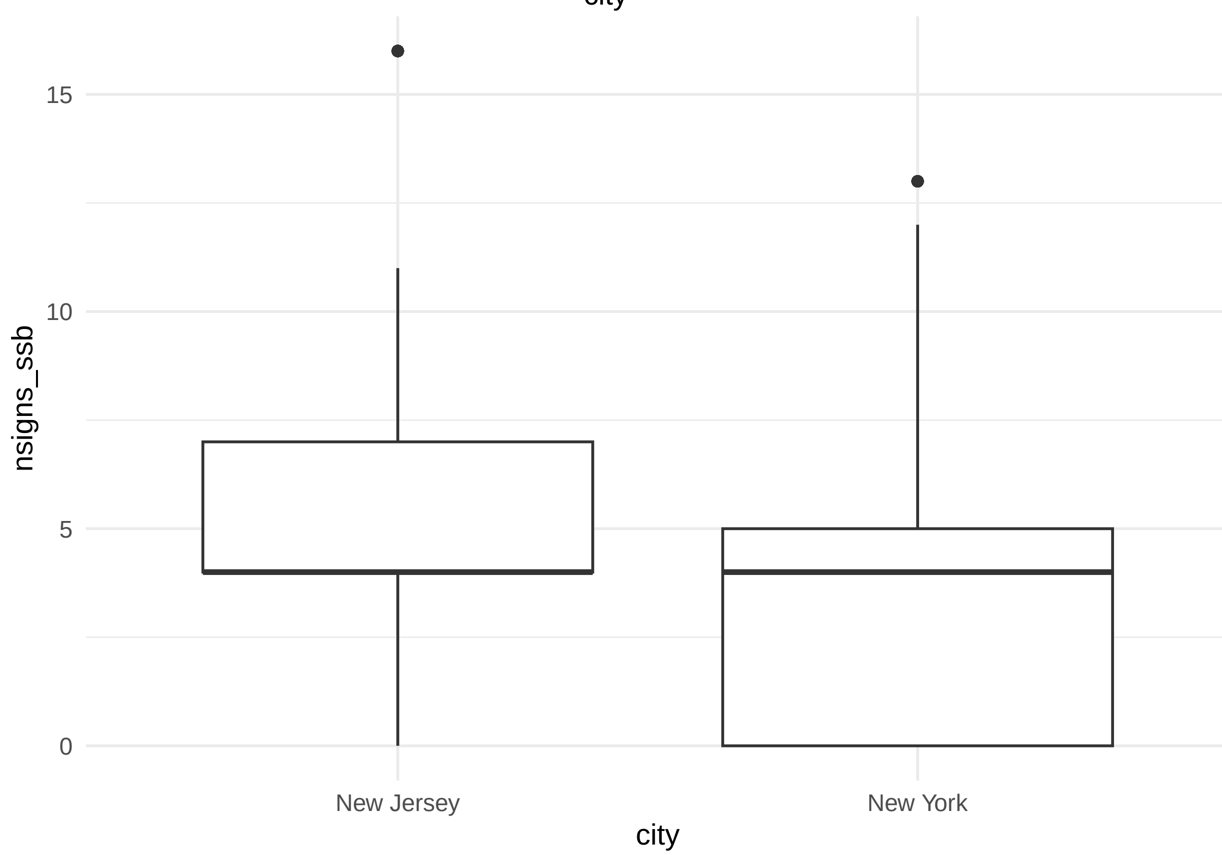


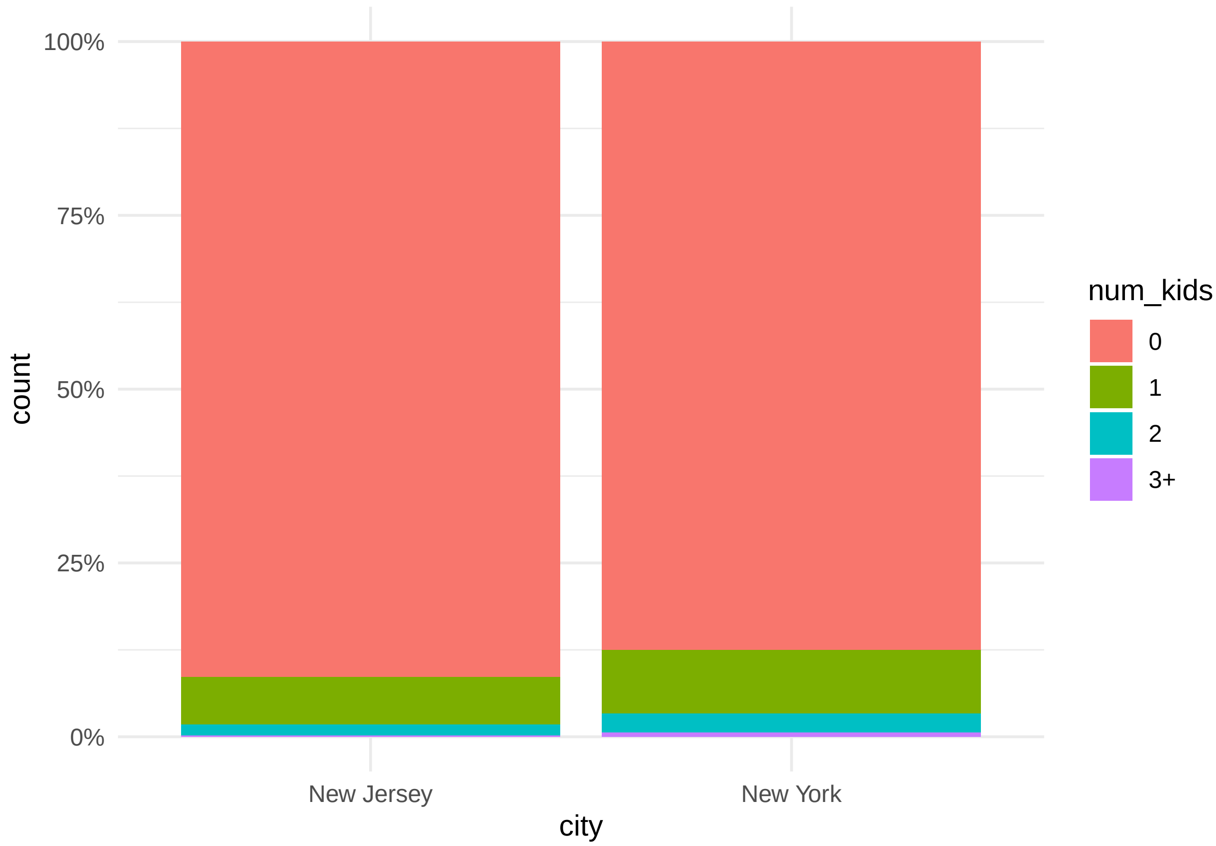
600
400
200
0
days_since_ban
Associates Degree
College Degree
Graduate Degree
High School
Less than High School
Some College
Some High School
edu

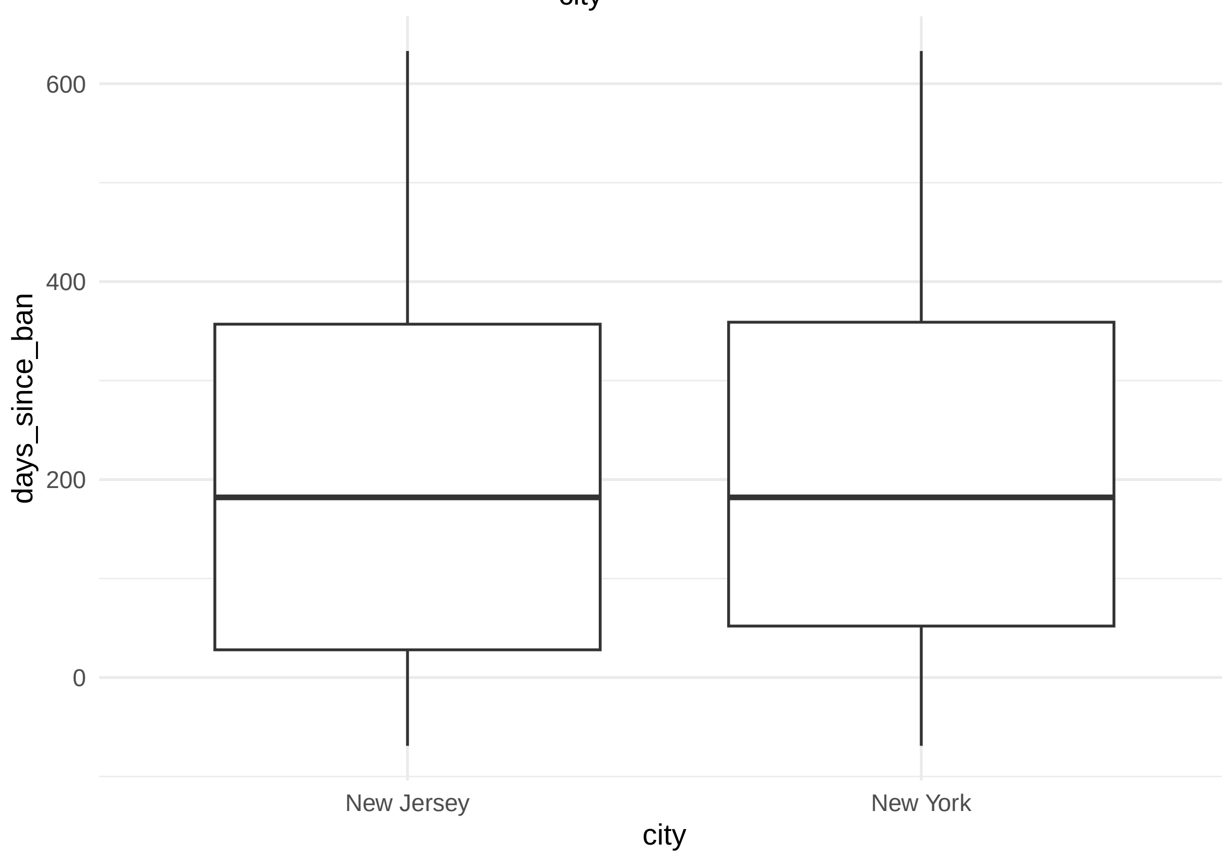


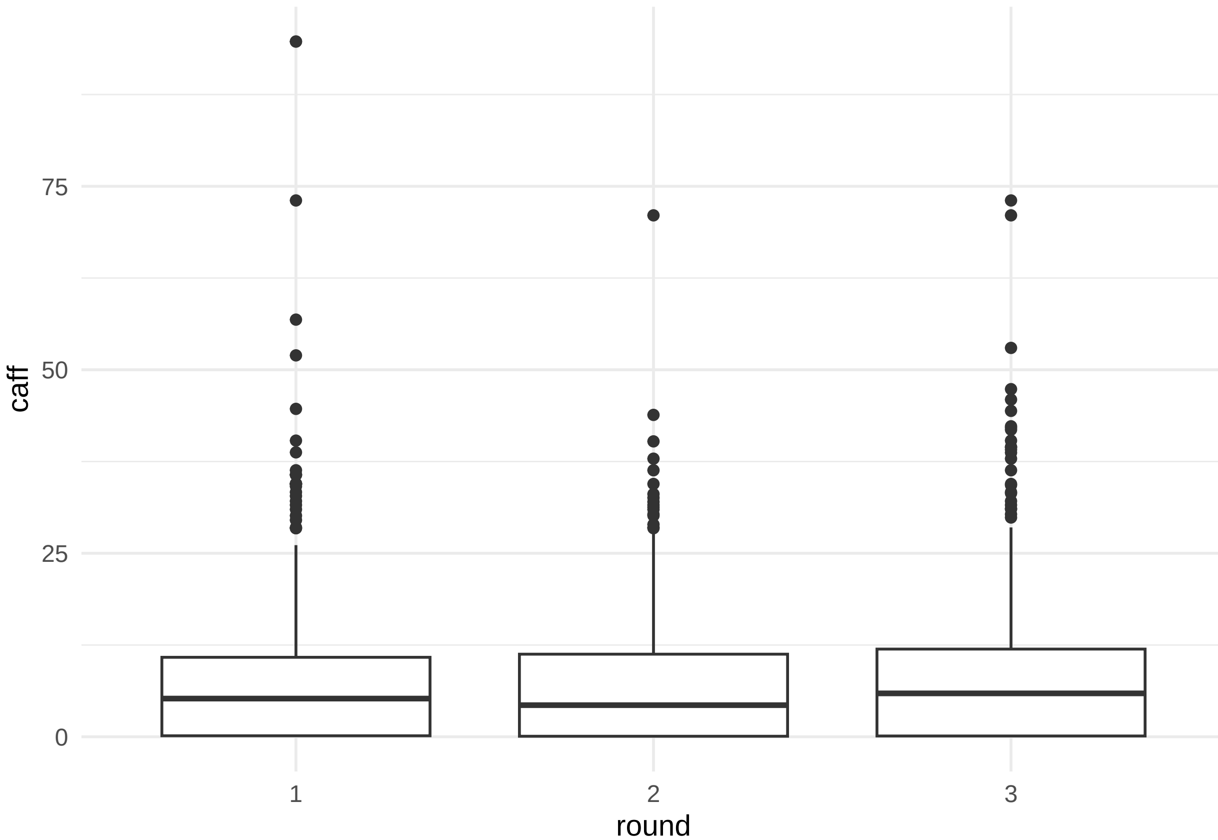
75
50
25
0
caff
New Jersey
New York
city

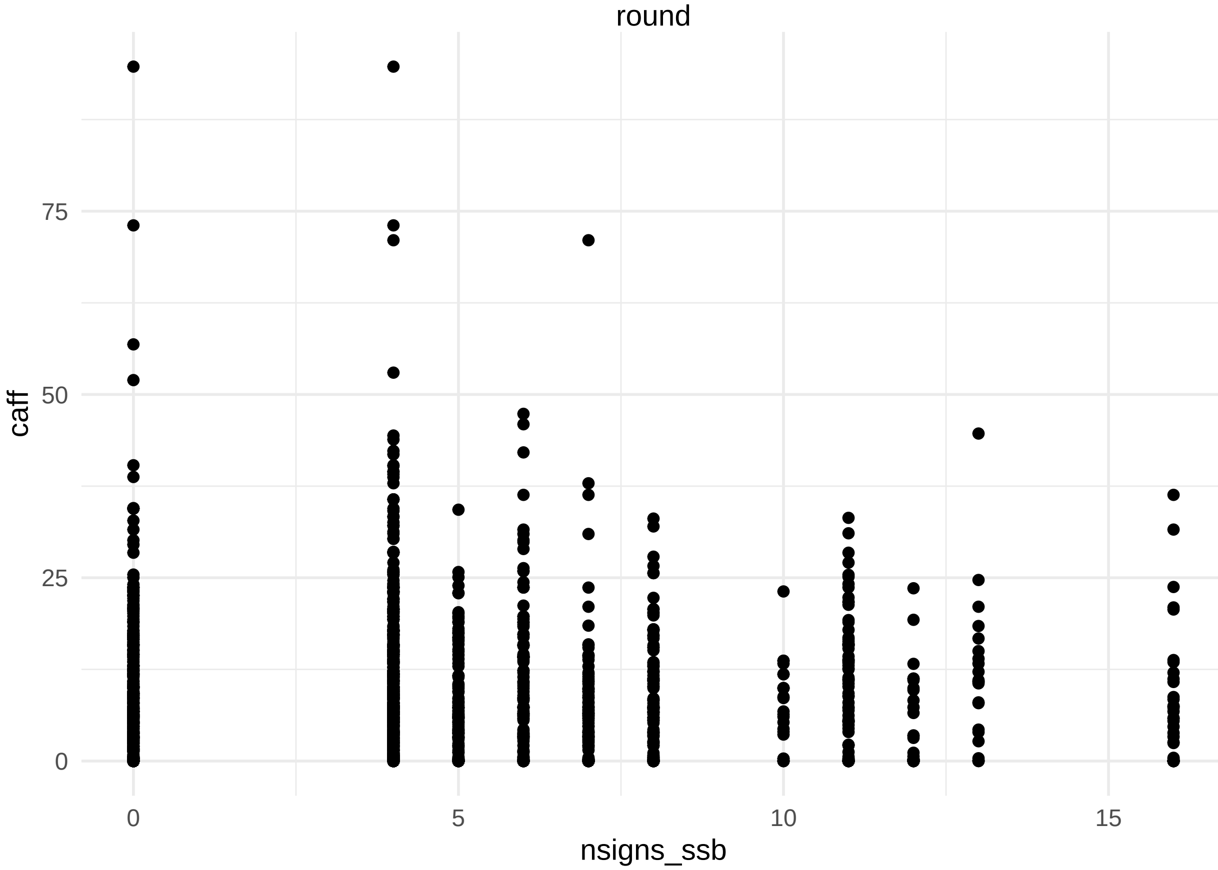

100%
75%
50%
25%
0%
count
New Jersey
New York
city
round
1
2
3

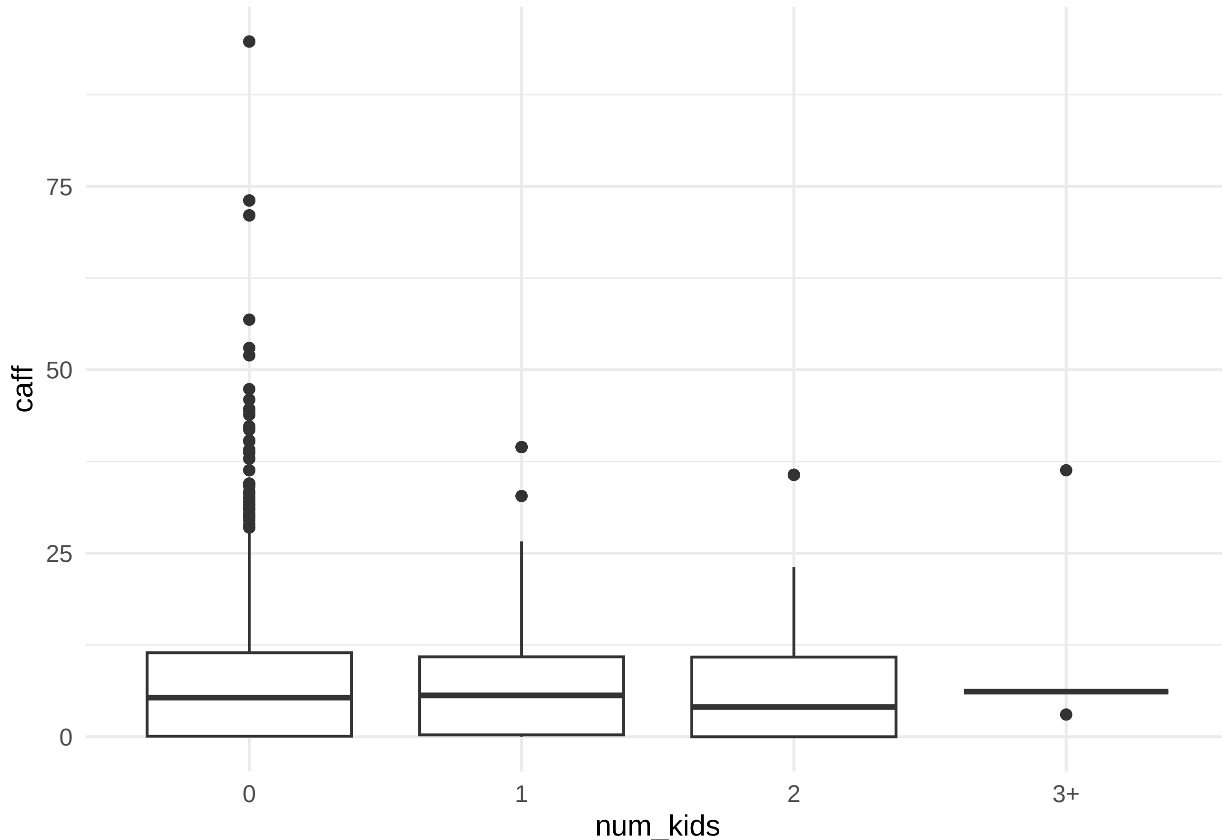

15
10
5
0
nsigns_ssb
New Jersey
New York
city

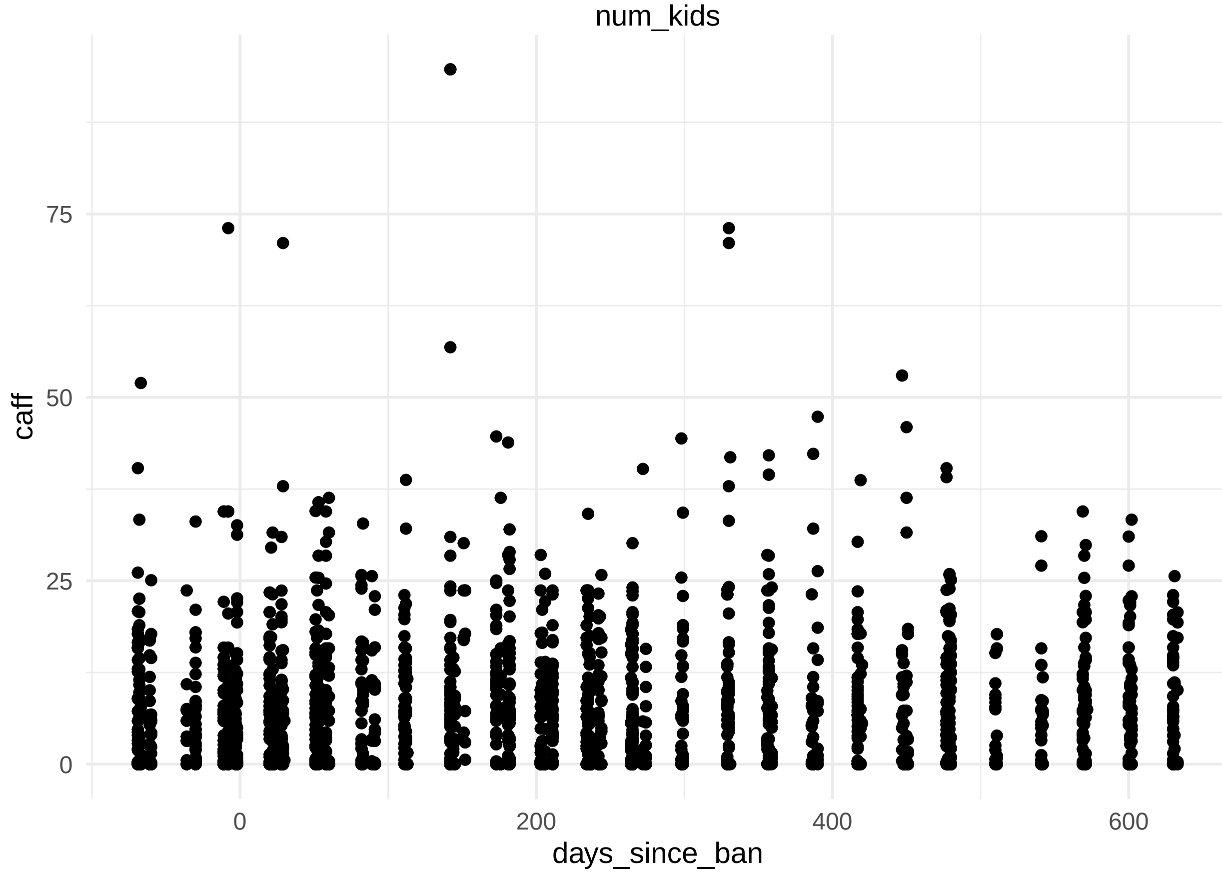

100%
75%
50%
25%
0%
count
num_kids
0
1
2
3+
New Jersey
New York
city

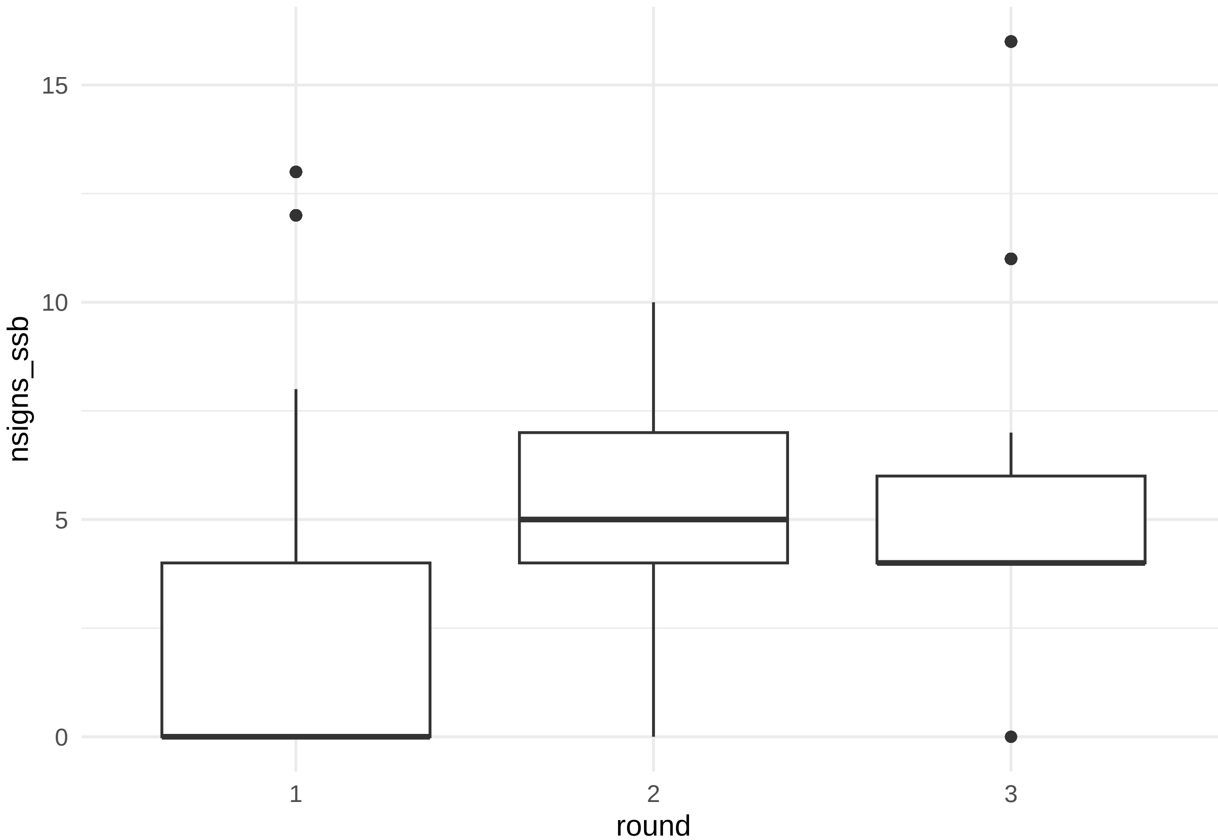

600
400
200
0
days_since_ban
New Jersey
New York
city


caff
75
50
25
0
1
2
3
round

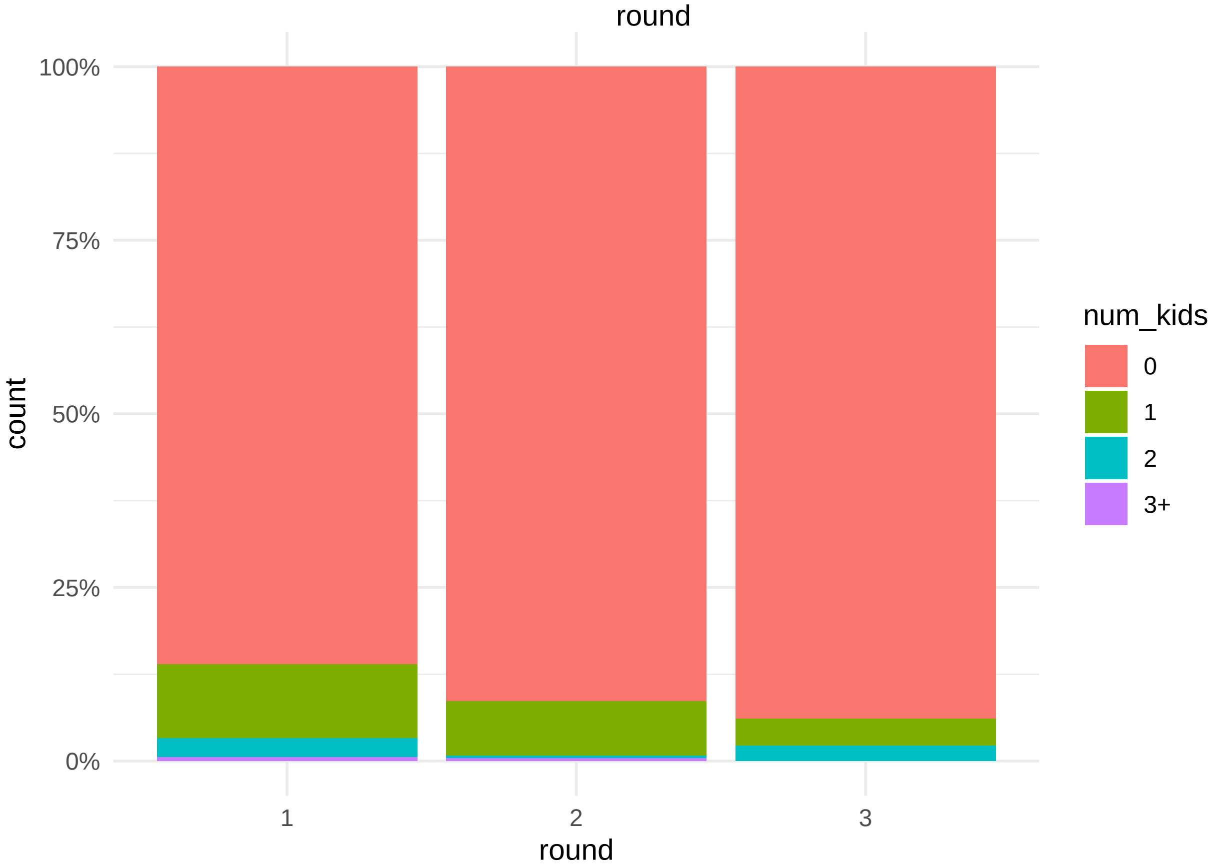

caff
75
50
25
0
0
5
10
15
nsigns_ssb


caff
0
25
50
75
0
1
2
3+
num_kids

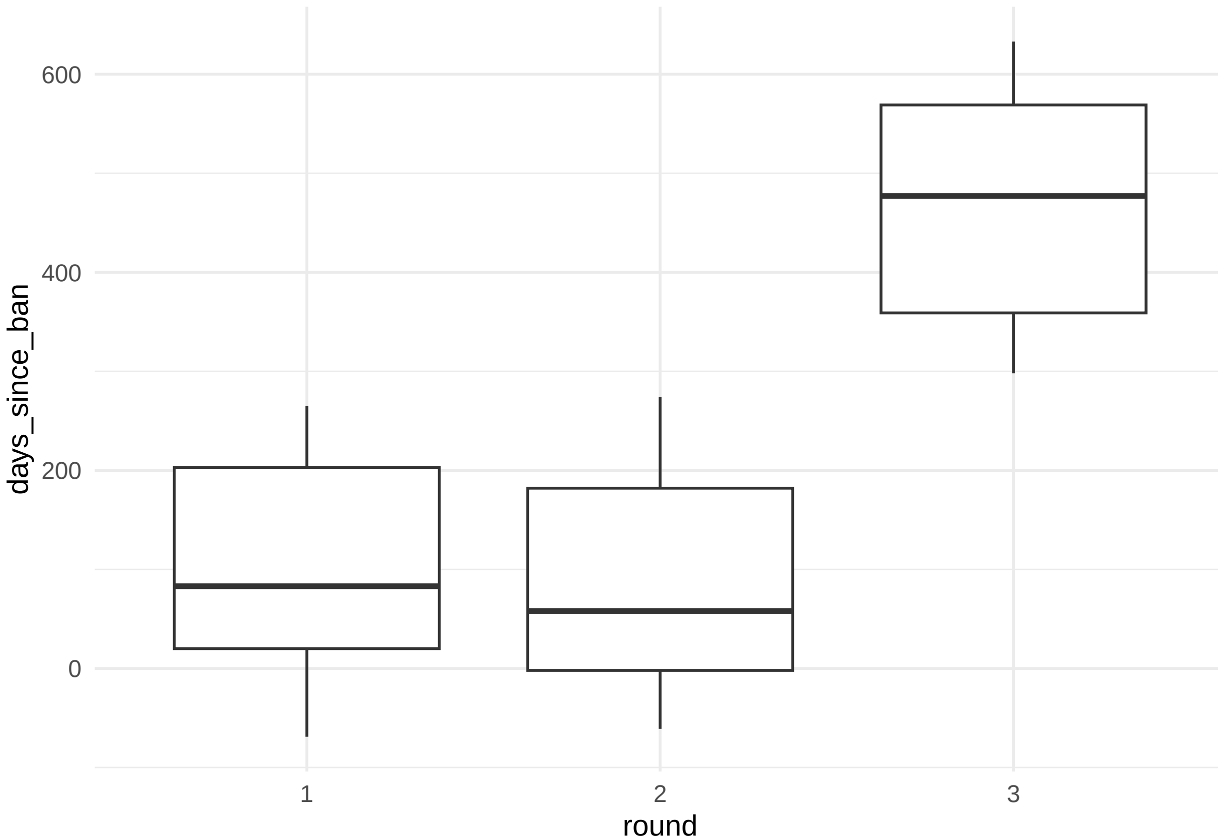

caff
0
25
50
75
0
200
400
600
days_since_ban

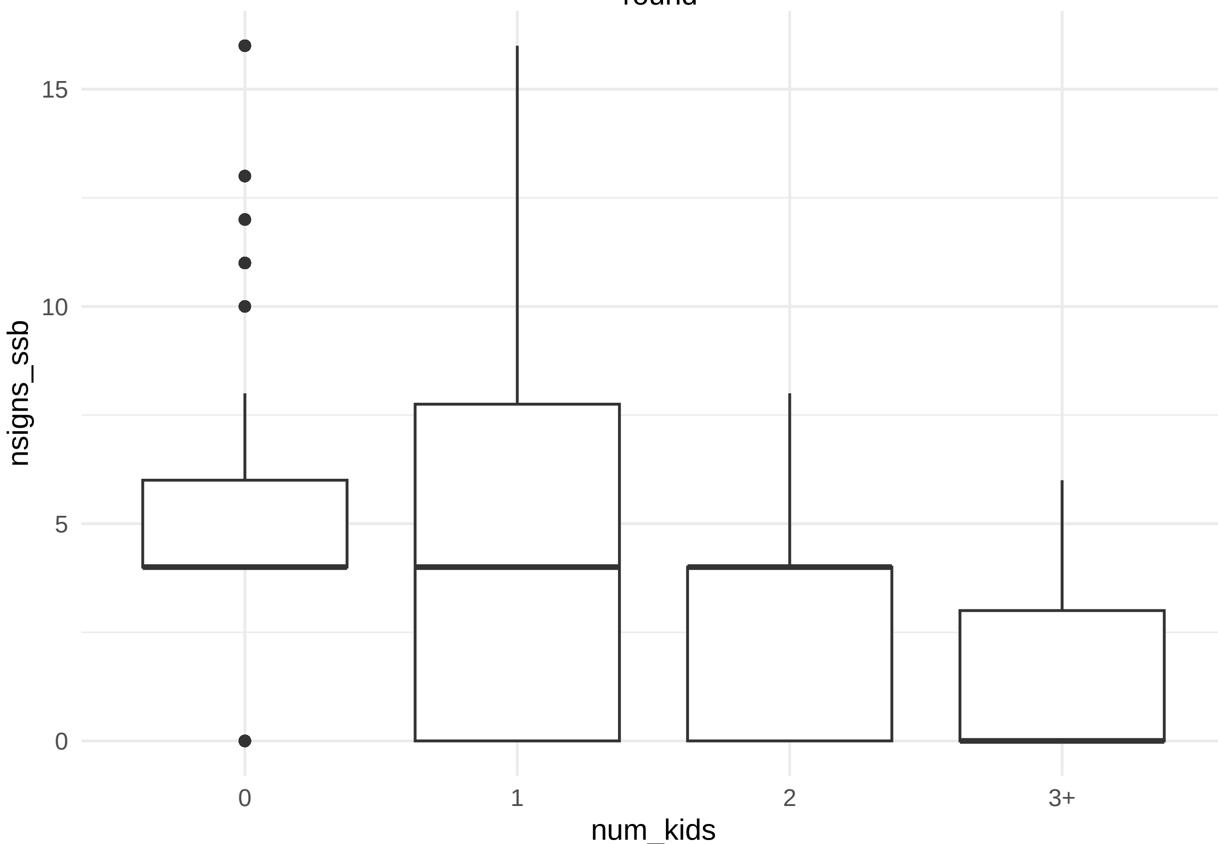


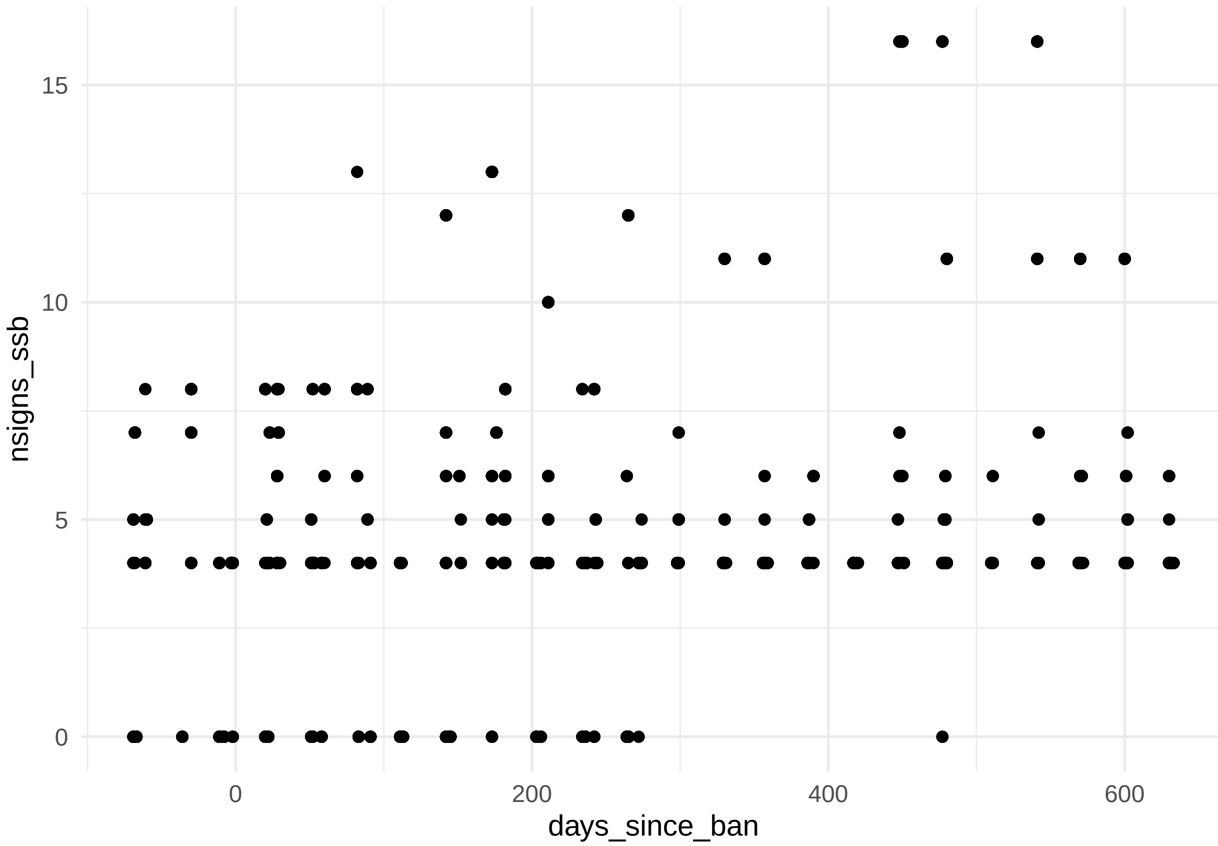
nsigns_ssb
15
10
5
0
1
2
3
round

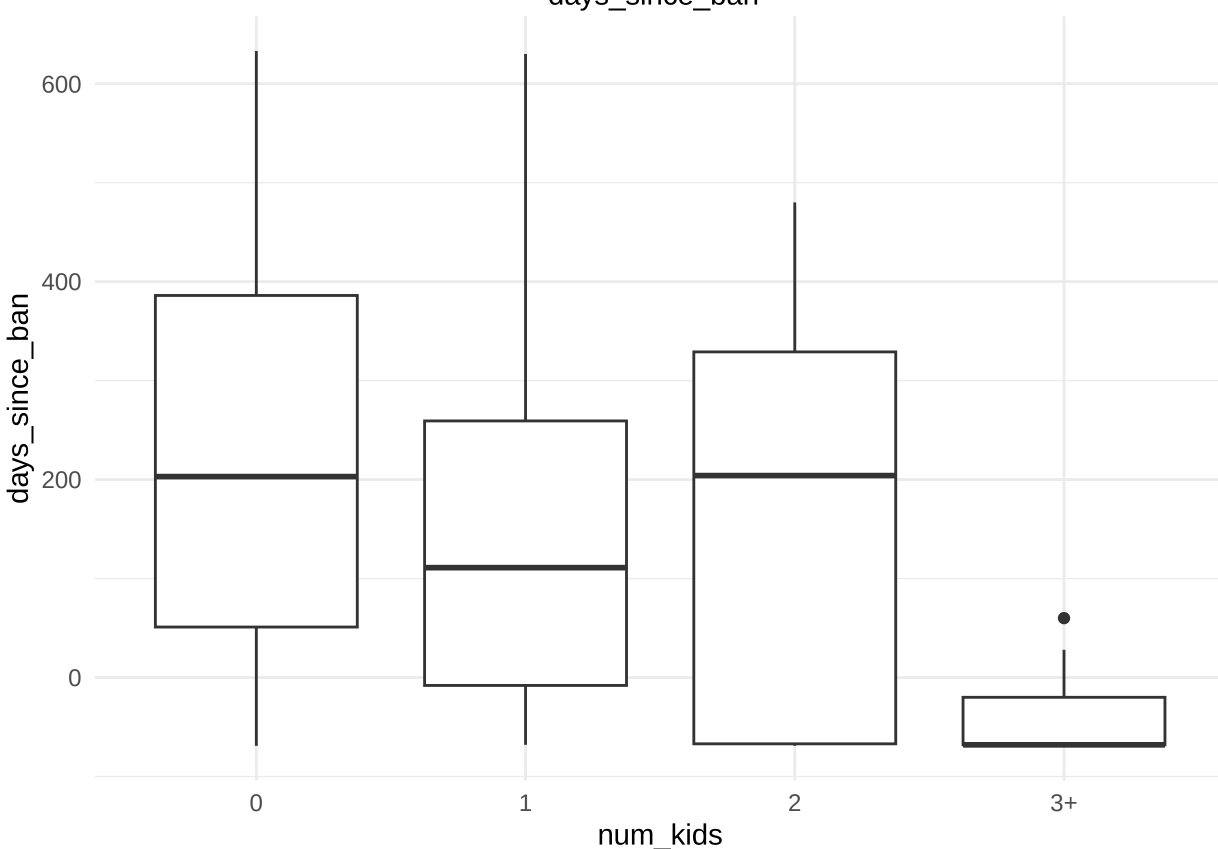


count
100%
75%
50%
25%
0%
1
2
3
round
num_kids
0
1
2
3+


600
400
200
0
days_since_ban
1
2
3
round


15
10
5
0
nsigns_ssb
0
1
2
3+
num_kids


nsigns_ssb
15
10
5
0
0
200
400
600
days_since_ban


days_since_ban
600
400
200
0
0
1
2
3+
num_kids

Modeling Process

Testing Different Optimization Methods

For models with no random effects, best to use Newton's approximation. For models with random effects, best to use `nlminb`, which is the default.

```
# No random effects
control_clm_full <- clm(limit ~ 1 + age + gender + race + edu + caff +
                          nsigns_ssb + num_kids + days_since_ban,
                   data = reduced_data, control = list(
  maxIter = 10000,
  maxLineIter = 2000,
  maxModIter = 2000,
  method = "Newton",
  trace = 1))
control_clm <- clm(limit ~ 1 + age + gender + race + edu + caff +
                     nsigns_ssb + num_kids + days_since_ban, data = reduced_data, control = list(
  method = "ucminf",
  stepmax = 1,
  grad = "central",
  maxeval = 500000,
  gradstep = c(1e-10, 1e-12),
  trace = 1))
control_clm <- clm(limit ~ 1 + age + gender + race + edu + caff +
                     nsigns_ssb + num_kids + days_since_ban, data = reduced_data, control = list(
  method = "nlminb",
  eval.max = 2000,
  iter.max = 1500,
  abs.tol = 1e-20,
  trace = 1))
control_clm <- clm(limit ~ 1 + age + gender + race + edu + caff +
                     nsigns_ssb + num_kids + days_since_ban, data = reduced_data, control = list(
  method = "optim",
  tmax = 100,
  maxit = 100000,
  type = 1,
  ndeps = 1e-10,
  REPORT = 1,
  trace = 1))

## Check with alternative packages. Produced the same intercepts
control_vglm <- vglm(limit ~ 1 + age + gender + race + edu + caff +
                       nsigns_ssb + num_kids + days_since_ban,
                     data = reduced_data, family = cumulative(parallel = TRUE))

## Random effects. Omit the rest for brevity
control_clmm_full <- clmm(limit ~ 1 + age + gender + race + edu + city + caff +
                            nsigns_ssb + num_kids + days_since_ban +
                            (1 | location) + (1 | round),
                          control = list(method = "nlminb",
                                         useMatrix = T,
                                         maxIter = 200,
                                         gradTol = 1e-4,
                                         maxLineIter = 200,
```

```
                                    trace = 1),
                    data = reduced_data, link = "logit")

# Same intercepts
summary(control_clm)
summary(control_vglm)
coef(control_vglm, matrix = T)

summary(control_clmm_full)
coef(control_clmm_full, matrix = T)
```

Full Model

Note that we also tested the non-standardized model. They both produced the similar conclusions. However, the non-standardized model couldn't fit properly because of the `kcal` variable. We proceeded with the standardized model for predictions.

```
control_clmm_full_std <- clmm(limit ~ 1 + age_std + gender + race + edu + city + caff_std +
                               nsigns_ssb_std + num_kids + days_since_ban_std +
                               kcal_std + fv_std + fatg_std + sugarg_std +
                               (1 | location) + (1 | round),
                         control = list(method = "nlminb",
                                        useMatrix = T,
                                        maxIter = 200,
                                        gradTol = 1e-4,
                                        maxLineIter = 200
                                        # , trace = 1
                                        ),
                    data = reduced_data, link = "logit")

summary(control_clmm_full_std)
```

```
## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: limit ~ 1 + age_std + gender + race + edu + city + caff_std +
##     nsigns_ssb_std + num_kids + days_since_ban_std + kcal_std +
##     fv_std + fatg_std + sugarg_std + (1 | location) + (1 | round)
## data:    reduced_data
##
##  link  threshold nobs logLik   AIC     niter       max.grad cond.H
##  logit flexible  2136 -3229.37 6516.74 7705(15567) 8.81e-03 3.3e+03
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  location (Intercept) 0.0098   0.09899
##  round    (Intercept) 0.0000   0.00000
## Number of groups:  location 53,  round 3
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## age_std                  0.2698144  0.0423399   6.373 1.86e-10 ***
## genderM                 -0.2501815  0.0809657  -3.090   0.0020 **
## raceBlack               -0.2727377  0.1714172  -1.591   0.1116
## raceNative              -0.3438367  0.2920322  -1.177   0.2390
```

```
## raceOther               -0.0701338  0.1822037  -0.385   0.7003
## raceWhite                0.0132864  0.1782384   0.075   0.9406
## eduCollege Degree       -0.0002048  0.1786373  -0.001   0.9991
## eduGraduate Degree       0.0274560  0.2111176   0.130   0.8965
## eduHigh School          -0.4272045  0.1713349  -2.493   0.0127 *
## eduLess than High School -0.5169313  0.2856546  -1.810   0.0704 .
## eduSome College         -0.1835094  0.1761134  -1.042   0.2974
## eduSome High School     -0.4519243  0.2044999  -2.210   0.0271 *
## cityNew York             0.1029802  0.0894119   1.152   0.2494
## caff_std                -0.0496544  0.0410145  -1.211   0.2260
## nsigns_ssb_std           0.0453734  0.0501650   0.904   0.3657
## num_kids1                0.0425334  0.1505704   0.282   0.7776
## num_kids2               -0.1692523  0.2886452  -0.586   0.5576
## num_kids3+              -2.5815563  1.0973158  -2.353   0.0186 *
## days_since_ban_std      -0.0324282  0.0417544  -0.777   0.4374
## kcal_std                -0.0700867  0.0465851  -1.504   0.1325
## fv_std                   0.0187035  0.0437145   0.428   0.6688
## fatg_std                 0.0106854  0.0391156   0.273   0.7847
## sugarg_std               0.0341262  0.0385197   0.886   0.3756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                  Estimate Std. Error z value
## Never|Seldom     -1.39621    0.23097  -6.045
## Seldom|Sometimes -0.99878    0.23003  -4.342
## Sometimes|Often   0.03407    0.22871   0.149
## Often|Always      0.83072    0.22952   3.619
```

```
## Non-standardized model
# control_clmm_full_non <- clmm(limit ~ 1 + age + gender + race + edu + city + caff +
# nsigns_ssb + num_kids + days_since_ban + kcal + fv +
# (1 | location) + (1 | round),
#                            control = list(method = "nlminb",
#                                           useMatrix = T,
#                                           maxIter = 200,
#                                           gradTol = 1e-4,
#                                           maxLineIter = 200,
#                                           trace = 1),
#                       data = reduced_data, link = "logit")
```

Fixed Effects

```
control_clmm_red <- clmm(limit ~ 1 + age_std + gender + edu + num_kids +
                           (1 | location) + (1 | round),
                         data = reduced_data, link = "logit")
anova(control_clmm_red, control_clmm_full_std)
```

```
## Likelihood ratio tests of cumulative link models:
##
##                       formula:
## control_clmm_red      limit ~ 1 + age_std + gender + edu + num_kids + (1 | location) + (1 | round)
## control_clmm_full_std limit ~ 1 + age_std + gender + race + edu + city + caff_std + nsigns_ssb_std +
##                       link: threshold:
```

```
## control_clmm_red      logit flexible
## control_clmm_full_std logit flexible
##
##                       no.par    AIC  logLik LR.stat df Pr(>Chisq)
## control_clmm_red          17 6509.0 -3237.5
## control_clmm_full_std     29 6516.7 -3229.4  16.304 12     0.1777
```

```r
summary(control_clmm_red)
```

```
## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: limit ~ 1 + age_std + gender + edu + num_kids + (1 | location) +
##     (1 | round)
## data:    reduced_data
##
##  link  threshold nobs logLik   AIC     niter      max.grad cond.H
##  logit flexible  2136 -3237.52 6509.05 2224(4451) 2.56e-03 3.3e+03
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  location (Intercept) 0.01333  0.1155
##  round    (Intercept) 0.00000  0.0000
## Number of groups:  location 53,  round 3
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## age_std                     0.26565    0.04144   6.411 1.45e-10 ***
## genderM                    -0.22886    0.08005  -2.859  0.00425 **
## eduCollege Degree           0.03375    0.17735   0.190  0.84905
## eduGraduate Degree          0.10883    0.20601   0.528  0.59732
## eduHigh School             -0.44635    0.17087  -2.612  0.00900 **
## eduLess than High School -0.49035    0.28065  -1.747  0.08061 .
## eduSome College            -0.18151    0.17566  -1.033  0.30147
## eduSome High School        -0.46765    0.20403  -2.292  0.02190 *
## num_kids1                   0.05868    0.14887   0.394  0.69348
## num_kids2                  -0.07456    0.28526  -0.261  0.79379
## num_kids3+                 -2.65632    1.09271  -2.431  0.01506 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                    Estimate Std. Error z value
## Never|Seldom        -1.2601     0.1637  -7.699
## Seldom|Sometimes   -0.8656     0.1624  -5.330
## Sometimes|Often     0.1632     0.1609   1.014
## Often|Always        0.9581     0.1623   5.902
```

Random Effects

Note that we couldn't perform bootstrap because the `simulate` command is not implemented in `ordinal`, but the effects are fairly marginal and not significant.

```r
control_clmm_loc <- clmm(limit ~ 1 + age_std + gender + edu + num_kids +
                           (1 | location),
```

```
                          data = reduced_data, link = "logit")

lrt_obs_round <- as.numeric(2*(logLik(control_clmm_red) -
                                logLik(control_clmm_loc)))
.5*(1 - pchisq(lrt_obs_round, 0)) + .5*(1 - pchisq(lrt_obs_round, 1))
```

Level 3 Random Intercept

```
## [1] 0.4986577
```

```
control_clm <- clm(limit ~ 1 + age_std + gender + edu + num_kids,
                   data = reduced_data, link = "logit")
lrt_obs_loc <- as.numeric(2*(logLik(control_clmm_loc) - logLik(control_clm)))
.5*(1 - pchisq(lrt_obs_loc, 0)) + .5*(1 - pchisq(lrt_obs_loc, 1))
```

Level 2 Random Intercept

```
## [1] 0.1154997
```

Separate slopes for each level

Ordinal provides two built-in commands for testing whether we need separate slopes for predictors of each level and whether we need to scale our response by each predictors. None of them showed significance.

```
nominal_test(control_clm)
```

```
## Tests of nominal effects
##
## formula: limit ~ 1 + age_std + gender + edu + num_kids
##          Df  logLik    AIC     LRT Pr(>Chi)
## <none>      -3238.2 6506.5
## age_std   3 -3235.8 6507.5  4.9725   0.1738
## gender    3 -3235.2 6506.4  6.1212   0.1059
## edu      18 -3232.9 6531.9 10.5938   0.9108
## num_kids
```

```
scale_test(control_clm)
```

```
## Tests of scale effects
##
## formula: limit ~ 1 + age_std + gender + edu + num_kids
##          Df  logLik    AIC    LRT Pr(>Chi)
## <none>      -3238.2 6506.5
## age_std   1 -3237.9 6507.8 0.6590   0.4169
## gender    1 -3237.9 6507.8 0.6793   0.4098
## edu       6 -3235.5 6513.0 5.5310   0.4777
## num_kids  3 -3238.2 6512.3 0.1399   0.9867
```

Overall fit Compared to the only intercept model.

```
control_null <- clm(limit ~ 1, data = reduced_data, link = "logit")
# Overall fit
anova(control_null, control_clm)
```

```
## Likelihood ratio tests of cumulative link models:
##
```

```
##              formula:                                      link: threshold:
## control_null limit ~ 1                                     logit flexible
## control_clm  limit ~ 1 + age_std + gender + edu + num_kids logit flexible
##
##              no.par    AIC  logLik LR.stat df Pr(>Chisq)
## control_null      4 6577.4 -3284.7
## control_clm      15 6506.5 -3238.2  92.955 11  4.386e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(control_clm)
```

```
## formula: limit ~ 1 + age_std + gender + edu + num_kids
## data:    reduced_data
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  2136 -3238.24 6506.48 5(1)  8.27e-08 3.3e+03
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## age_std                  0.26134    0.04087   6.394 1.62e-10 ***
## genderM                 -0.22153    0.07940  -2.790  0.00527 **
## eduCollege Degree        0.05781    0.17565   0.329  0.74205
## eduGraduate Degree       0.12703    0.20482   0.620  0.53514
## eduHigh School          -0.44982    0.17025  -2.642  0.00824 **
## eduLess than High School -0.49357    0.28012  -1.762  0.07807 .
## eduSome College         -0.18401    0.17518  -1.050  0.29352
## eduSome High School     -0.47584    0.20319  -2.342  0.01919 *
## num_kids1                0.05404    0.14777   0.366  0.71457
## num_kids2               -0.11069    0.28133  -0.393  0.69399
## num_kids3+              -2.73563    1.08815  -2.514  0.01194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                 Estimate Std. Error z value
## Never|Seldom      -1.2424     0.1615  -7.693
## Seldom|Sometimes  -0.8491     0.1603  -5.296
## Sometimes|Often    0.1757     0.1591   1.105
## Often|Always       0.9683     0.1606   6.027
```

Model Diagnostics

Accuracy Metrics

Because residual analysis are not well understood in ordinal models, we opted for accuracy metrics. Note that our model doesn't predict well.

```
library(tidymodels)
library(workflows)

model_accuracy <- function(model = control_clm, adj = F) {
    comp_metrics <- function(model = model, predict) {
      control_results <- reduced_data %>%
      bind_cols(fit = predict)
```

```r
    # Confusion matrix
    # table(control_results$limit, control_results$fit)
    conf_mat(control_results, truth = limit, estimate = fit) -> conf

    # accuracy metrics
    accuracy(control_results, truth = limit, estimate = fit) -> acc
    sensitivity(control_results, truth = limit, estimate = fit) -> sen
    specificity(control_results, truth = limit, estimate = fit) -> spe
    # ppv(control_results, truth = limit, estimate = fit)

    # Goodness of fit
    chisq.test(control_results$limit, control_results$fit) -> gof

    return(list(conf = conf, acc = acc, sen = sen, spe = spe, gof = gof))
    }
  if (adj) {
    # Predict response
    control_vglm_pred <- predict(model, type = "response")
    level_counts <- table(reduced_data$limit)
    total_counts <- sum(level_counts)
    proportions <- as.numeric(level_counts / total_counts)
    names(proportions) <- names(level_counts)

    adjusted_probs <- control_vglm_pred / proportions[colnames(control_vglm_pred)]
    adjusted_probs <- adjusted_probs / rowSums(adjusted_probs)
    fit <- ordered(colnames(adjusted_probs)[max.col(adjusted_probs)],
                   levels = c("Never", "Seldom", "Sometimes",
                              "Often", "Always"))
    comp_metrics(model = model, predict = fit) -> result
  } else {
    # Predict response
    control_pred <- predict(model, type = "class")
    comp_metrics(model = model, control_pred) -> result
  }
  return(result)
}

model_accuracy(control_clm)
```

```
## $conf
##             Truth
## Prediction   Never Seldom Sometimes Often Always
##   Never        493    131       368   229    288
##   Seldom         0      0         0     0      0
##   Sometimes      0      0         0     0      0
##   Often          0      0         0     0      0
##   Always       135     50       152   110    180
##
## $acc
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.315
```

```
##
## $sen
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 sensitivity macro          0.234
##
## $spe
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 specificity macro          0.812
##
## $gof
##
##  Pearson's Chi-squared test
##
## data:  control_results$limit and control_results$fit
## X-squared = 39.245, df = 4, p-value = 6.199e-08
```

```r
## Similar results under different model specifications
control_clm_probit <- clm(limit ~ 1 + age_std + gender + edu + num_kids,
                   data = reduced_data, link = "probit")
model_accuracy(control_clm_probit)
```

```
## $conf
##            Truth
## Prediction  Never Seldom Sometimes Often Always
##   Never       489    131       369   229    287
##   Seldom        0      0         0     0      0
##   Sometimes     0      0         0     0      0
##   Often         0      0         0     0      0
##   Always      139     50       151   110    181
##
## $acc
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.314
##
## $sen
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 sensitivity macro          0.233
##
## $spe
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 specificity macro          0.811
##
## $gof
##
##  Pearson's Chi-squared test
```

```
##
## data:  control_results$limit and control_results$fit
## X-squared = 37.073, df = 4, p-value = 1.74e-07
```

```r
control_clm_sym <- clm(limit ~ 1 + age_std + gender + edu + num_kids,
                   data = reduced_data,
                   link = "probit", threshold = "equidistant")
model_accuracy(control_clm_sym)
```

```
## $conf
##            Truth
## Prediction  Never Seldom Sometimes Often Always
##   Never       465    115       321   214    239
##   Seldom        0      0         0     0      0
##   Sometimes     0      0         0     0      0
##   Often         0      0         0     0      0
##   Always      163     66       199   125    229
##
## $acc
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.325
##
## $sen
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 sensitivity macro          0.246
##
## $spe
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 specificity macro          0.816
##
## $gof
##
##  Pearson's Chi-squared test
##
## data:  control_results$limit and control_results$fit
## X-squared = 61.965, df = 4, p-value = 1.121e-12
```

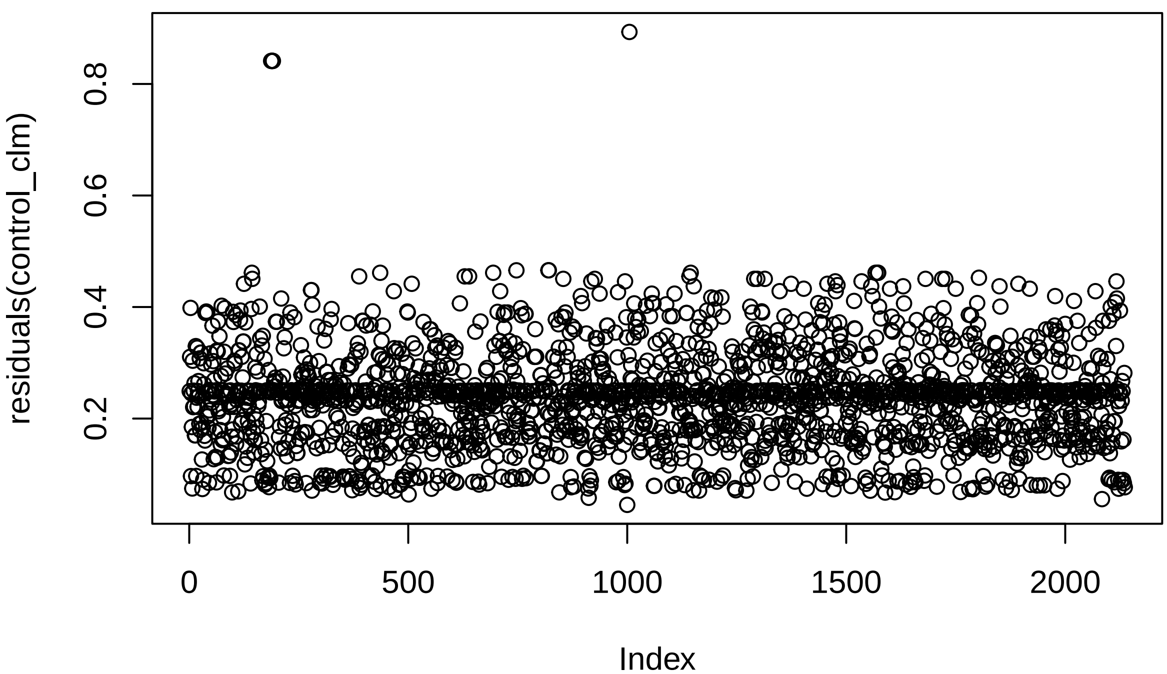
```r
## Use VGAM to get prob for each level of resp, not implemented in Ordinal
## Similarly inaccurate model
control_vglm_sig <- vglm(limit ~ 1 + age_std + gender + edu + num_kids,
                     data = reduced_data,
                     family = cumulative(parallel = TRUE))
model_accuracy(control_vglm_sig, adj = T)
```

```
## Warning in chisq.test(control_results$limit, control_results$fit): Chi-squared
## approximation may be incorrect
```

```
## $conf
##            Truth
## Prediction  Never Seldom Sometimes Often Always
```

```
##   Never        266      82        196    127     143
##   Seldom        15       4         10      8      10
##   Sometimes    137      35        118     69      88
##   Often         59      13         70     46      87
##   Always       151      47        126     89     140
##
## $acc
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.269
##
## $sen
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 sensitivity macro          0.221
##
## $spe
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 specificity macro          0.807
##
## $gof
##
##  Pearson's Chi-squared test
##
## data:  control_results$limit and control_results$fit
## X-squared = 42.047, df = 16, p-value = 0.0003882
```

```r
plot(fitted(control_clm), residuals(control_clm))
```



```r
control_resid <- control_results %>%
  mutate(
    case_id = row_number(),
    fit = ordered(fit, levels = c("Never", "Seldom", "Sometimes", "Often", "Always"))
  ) %>%
  select(case_id, limit, fit) %>%
  mutate(
    limit_num = case_when(
      limit == "Never" ~ 0,
      limit == "Seldom" ~ 1,
      limit == "Sometimes" ~ 2,
      limit == "Often" ~ 3,
      limit == "Always" ~ 4),
    fit_num = case_when(
      fit == "Never" ~ 0,
      fit == "Seldom" ~ 1,
      fit == "Sometimes" ~ 2,
      fit == "Often" ~ 3,
      fit == "Always" ~ 4),
    ) %>%
  mutate(resid = limit_num - fit_num)
```

```
## Error: object 'control_results' not found
```

```r
ggplot(control_resid, aes(x = case_id)) +
  geom_jitter(aes(y = limit, color = "Actual"), alpha = 0.4) +
  geom_jitter(aes(y = fit, color = "Fitted"), alpha = 0.4) +
  scale_color_manual(values = c("Actual" = "blue", "Fitted" = "red")) +
  labs(
    x = "Case ID",
    y = "Response Category",
    title = "Actual vs. Fitted Values",
    color = "Legend"
  ) +
  theme_minimal()
```

```
## Error: object 'control_resid' not found
```

```r
ggplot(control_resid, aes(x = case_id, y = resid)) +
  geom_jitter(alpha = 0.4) +
  labs(
    x = "Case ID",
    y = "Residual"
  ) +
  theme_minimal()
```

```
## Error: object 'control_resid' not found
```

Effects Interpretation

Confidence Intervals

```r
confint(control_clm) %>% kable(digits = 3)
```

	2.5 %	97.5 %
age_std	0.181	0.342
genderM	-0.377	-0.066
eduCollege Degree	-0.287	0.403
eduGraduate Degree	-0.274	0.529
eduHigh School	-0.784	-0.116
eduLess than High School	-1.044	0.055
eduSome College	-0.527	0.160
eduSome High School	-0.875	-0.078
num_kids1	-0.236	0.344
num_kids2	-0.667	0.440
num_kids3+	-5.683	-0.955

```
exp(confint(control_clm)) %>% kable(digits = 3)
```

	2.5 %	97.5 %
age_std	1.199	1.407
genderM	0.686	0.936
eduCollege Degree	0.751	1.496
eduGraduate Degree	0.760	1.697
eduHigh School	0.457	0.891
eduLess than High School	0.352	1.057
eduSome College	0.590	1.173
eduSome High School	0.417	0.925
num_kids1	0.790	1.410
num_kids2	0.513	1.553
num_kids3+	0.003	0.385

```
(100*(exp(confint(control_clm))-1)) %>% kable(digits = 3)
```

	2.5 %	97.5 %
age_std	19.889	40.727
genderM	-31.426	-6.386
eduCollege Degree	-24.914	49.556
eduGraduate Degree	-23.996	69.714
eduHigh School	-54.331	-10.941
eduLess than High School	-64.810	5.661
eduSome College	-40.992	17.312
eduSome High School	-58.298	-7.474
num_kids1	-21.016	41.028
num_kids2	-48.673	55.254
num_kids3+	-99.660	-61.500