Admissions Consulting

Ishika Gupta

02/11/22

```
data1 <- read_csv("/Users/ishika/Desktop/Applied Data Science/Project 1/Admissions_data_main_1.csv")
```

```
## New names:
## * `` -> ...1
```

```
data2 <- read_csv("/Users/ishika/Desktop/Applied Data Science/Project 1/Admissions_data_main_2.csv")
```

```
## New names:
## * `` -> ...1
```

```
data3 <- read_csv("/Users/ishika/Desktop/Applied Data Science/Project 1/Admissions_data_main_3.csv")
```

```
## New names:
## * `` -> ...1
```

```
data4 <- read_csv("/Users/ishika/Desktop/Applied Data Science/Project 1/Admissions_data_main_4.csv")
```

```
## New names:
## * `` -> ...1
```

```
data <- rbind(data1, data2, data3, data4)
```

```
nrow(data)
```

```
## [1] 263136
```

```
colnames(data)
```

```
##  [1] "...1"      "MyID"      "C.O..22"   "Male"      "URM"       "Zip"
##  [7] "State"     "X28.36"    "X23.27"    "Soph"      "MajorCode" "Density"
## [13] "X..Black"  "X..Latino" "X.PrivHS"  "X..Bach."  "X..Adv."   "HH.."
## [19] "Fam.."     "FamK.."    "Lower"     "LowMid"    "Mid"       "UpMid"
## [25] "Upper"     "CBSA"      "Metro"     "CSA"       "WooDist"   "Lat"
## [31] "Long"      "Inquiry"   "Applicant"
```

```
str(data)
```

```
## spec_tbl_df [263,136 x 33] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1     : num [1:263136] 1 2 3 4 5 6 7 8 9 10 ...
##  $ MyID     : num [1:263136] 1 2 3 4 5 6 7 8 9 10 ...
##  $ C.O..22  : num [1:263136] 1 1 1 0 0 0 0 1 0 1 ...
##  $ Male     : num [1:263136] 1 1 0 0 0 1 1 0 0 0 ...
##  $ URM      : num [1:263136] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Zip      : num [1:263136] 55376 44654 46217 98030 60565 ...
##  $ State    : chr [1:263136] "MN" "OH" "IN" "WA" ...
##  $ X28.36   : num [1:263136] 0 0 0 0 0 0 0 0 0 0 ...
##  $ X23.27   : num [1:263136] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Soph     : num [1:263136] 0 0 0 0 0 0 0 0 0 0 ...
##  $ MajorCode: num [1:263136] 27 13 26 0 9 13 28 13 4 13 ...
##  $ Density  : num [1:263136] 538 125 1367 4741 3251 ...
##  $ X..Black : num [1:263136] 2.2 0.9 6.2 16.5 5.3 4.8 5.2 5.2 1.6 1.5 ...
##  $ X..Latino: num [1:263136] 1.7 0.9 3.5 15.3 4.1 3.5 3.5 3.5 10.1 8.7 ...
##  $ X.PrivHS : num [1:263136] 0 11.6 19.6 7.2 5.1 16.8 9.7 9.7 5.6 36 ...
##  $ X..Bach. : num [1:263136] 33 10 33 20 68 65 65 65 26 76 ...
##  $ X..Adv.  : num [1:263136] 9 3 12 5 30 30 24 24 9 40 ...
##  $ HH..     : num [1:263136] 96188 53373 61055 55797 120891 ...
##  $ Fam..    : num [1:263136] 100668 61208 72920 61043 135794 ...
##  $ FamK..   : num [1:263136] 88666 63007 82663 57838 134667 ...
##  $ Lower    : num [1:263136] 4 10 9 20 5 7 5 5 8 7 ...
##  $ LowMid   : num [1:263136] 13 29 24 22 7 19 10 10 15 10 ...
##  $ Mid      : num [1:263136] 32 39 40 32 20 27 24 24 35 18 ...
##  $ UpMid    : num [1:263136] 32 15 21 19 24 17 24 24 26 18 ...
##  $ Upper    : num [1:263136] 19 8 7 7 45 30 37 37 17 48 ...
##  $ CBSA     : num [1:263136] 33460 18740 26900 42660 16980 ...
##  $ Metro    : num [1:263136] 1 0 1 1 1 1 1 1 1 1 ...
##  $ CSA      : num [1:263136] 378 NA 294 500 176 198 378 378 176 NA ...
##  $ WooDist  : num [1:263136] 665 19 236 2025 326 ...
##  $ Lat      : num [1:263136] 45.2 40.5 39.7 47.4 41.7 ...
##  $ Long     : num [1:263136] -93.7 -81.9 -86.2 -122.2 -88.1 ...
##  $ Inquiry  : num [1:263136] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Applicant: num [1:263136] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   MyID = col_double(),
##   ..   C.O..22 = col_double(),
##   ..   Male = col_double(),
##   ..   URM = col_double(),
##   ..   Zip = col_double(),
##   ..   State = col_character(),
##   ..   X28.36 = col_double(),
##   ..   X23.27 = col_double(),
##   ..   Soph = col_double(),
##   ..   MajorCode = col_double(),
##   ..   Density = col_double(),
##   ..   X..Black = col_double(),
##   ..   X..Latino = col_double(),
##   ..   X.PrivHS = col_double(),
```

```
##   ..   X..Bach. = col_double(),
##   ..   X..Adv. = col_double(),
##   ..   HH.. = col_double(),
##   ..   Fam.. = col_double(),
##   ..   FamK.. = col_double(),
##   ..   Lower = col_double(),
##   ..   LowMid = col_double(),
##   ..   Mid = col_double(),
##   ..   UpMid = col_double(),
##   ..   Upper = col_double(),
##   ..   CBSA = col_double(),
##   ..   Metro = col_double(),
##   ..   CSA = col_double(),
##   ..   WooDist = col_double(),
##   ..   Lat = col_double(),
##   ..   Long = col_double(),
##   ..   Inquiry = col_double(),
##   ..   Applicant = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
res <- cor(data[, c('Lat','Long', 'C.O..22', 'MyID', 'Male', 'URM', 'Zip', 'X28.36', 'X23.27', 'Soph',

round(res, 4)
```

```
##              [,1]
## Lat        0.0237
## Long       0.0179
## C.O..22    0.0024
## MyID      -0.0027
## Male      -0.0147
## URM        0.0017
## Zip       -0.0053
## X28.36     0.0117
## X23.27     0.0255
## Soph       0.0558
## MajorCode -0.0013
## Density   -0.0089
## X..Black   0.0146
## X..Latino -0.0312
## X.PrivHS   0.0049
## X..Bach.  -0.0180
## X..Adv.   -0.0105
## HH..      -0.0388
## Fam..     -0.0319
## FamK..    -0.0287
## Lower      0.0244
## LowMid     0.0284
## Mid        0.0295
## UpMid     -0.0138
## Upper     -0.0324
## CBSA      -0.0226
## Metro     -0.0198
## CSA       -0.0241
```

```
## WooDist   -0.0530
## Inquiry    0.4764
```

```
### Variables to drop:
```

```
#This shows us the number of missing variables in every columns.
sapply(data, function(x) sum(is.na(x)))
```

```
##      ...1      MyID   C.O..22      Male       URM       Zip     State    X28.36
##         0         0         0         0         0         0         0         0
##    X23.27      Soph MajorCode   Density  X..Black X..Latino  X.PrivHS  X..Bach.
##         0         0         0         0         0         0       128        19
##   X..Adv.      HH..      Fam..    FamK..     Lower    LowMid       Mid     UpMid
##        19        70       101       540         0         0         0         0
##     Upper      CBSA     Metro       CSA   WooDist       Lat      Long   Inquiry
##         0         0         0     15843         0         0         0         0
## Applicant
##         0
```

Data Cleaning

```
# Dropping the unnecessary column
data = subset(data, select = -c(...1, MyID))
```

```
colPerc(xtabs(~Applicant+Male,data=data))
```

```
##           Male
## Applicant      0      1
##     0      98.94  99.23
##     1       1.06   0.77
##     Total 100.00 100.00
```

```
xtabs(~Applicant+Male,data=data)
```

```
##           Male
## Applicant      0      1
##         0 136003 124701
##         1   1460    972
```

```
colPerc(xtabs(~Applicant+Inquiry,data=data))
```

```
##           Inquiry
## Applicant   0      1
##     0     100  76.62
##     1       0  23.38
##     Total 100 100.00
```

```
xtabs(~Applicant+Inquiry,data=data)
```

```
##           Inquiry
## Applicant      0      1
##         0 252734   7970
##         1      0   2432
```

```
colPerc(xtabs(~Applicant+URM,data=data))
```

```
##           URM
## Applicant      0      1
##     0      99.08  99.04
##     1       0.92   0.96
##     Total 100.00 100.00
```

```
xtabs(~Applicant+URM,data=data)
```

```
##           URM
## Applicant      0      1
##         0 223981  36723
##         1   2076    356
```

```
colPerc(xtabs(~Applicant+X28.36,data=data))
```

```
##           X28.36
## Applicant     0      1
##     0      99.1  98.33
##     1       0.9   1.67
##     Total 100.0 100.00
```

```
xtabs(~Applicant+X28.36,data=data)
```

```
##           X28.36
## Applicant      0      1
##         0 254058   6646
##         1   2319    113
```

```
colPerc(xtabs(~Applicant+X23.27,data=data))
```

```
##           X23.27
## Applicant      0      1
##     0      99.12  97.58
##     1       0.88   2.42
##     Total 100.00 100.00
```

```
xtabs(~Applicant+X23.27,data=data)
```

```
##          X23.27
## Applicant      0      1
##         0 253721   6983
##         1   2259    173
```

```
colPerc(xtabs(~Applicant+Soph,data=data))
```

```
##          Soph
## Applicant      0      1
##     0      99.32  97.84
##     1       0.68   2.16
##     Total 100.00 100.00
```

```
xtabs(~Applicant+Soph,data=data)
```

```
##          Soph
## Applicant      0      1
##         0 218487  42217
##         1   1500    932
```

```
colPerc(xtabs(~Applicant+MajorCode,data=data))
```

```
##          MajorCode
## Applicant      0      1      2   3     4      5      6     7      8      9
##     0      99.05  99.33  98.97 100  99.1  98.94  98.63  99.2  98.99  99.07
##     1       0.95   0.67   1.03   0   0.9   1.06   1.37   0.8   1.01   0.93
##     Total 100.00 100.00 100.00 100 100.0 100.00 100.00 100.0 100.00 100.00
##          MajorCode
## Applicant     10     11     12     13    14  15     16  17     18     19     20
##     0      99.19  99.27  98.99  99.06  98.6  99  99.24 100  99.46  98.61  99.16
##     1       0.81   0.73   1.01   0.94   1.4   1   0.76   0   0.54   1.39   0.84
##     Total 100.00 100.00 100.00 100.00 100.0 100 100.00 100 100.00 100.00 100.00
##          MajorCode
## Applicant     21     22     23     24     25     26     27     28     29
##     0      98.83  98.66  99.03  99.02  99.49  99.11  98.53  99.01  99.13
##     1       1.17   1.34   0.97   0.98   0.51   0.89   1.47   0.99   0.87
##     Total 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00 100.00
```

```
xtabs(~Applicant+MajorCode,data=data)
```

```
##          MajorCode
## Applicant     0     1     2     3     4     5     6     7     8     9    10
##         0  4068  1182  3267    12 24125 13392  2877 12214  7337 32614  2341
##         1    39     8    34     0   219   144    40    99    75   305    19
##          MajorCode
## Applicant    11    12    13    14    15    16    17    18    19    20    21
##         0  3256   686 39772  1549  6236   523    11  2594   637  1421   842
##         1    24     7   376    22    63     4     0    14     9    12    10
##          MajorCode
## Applicant    22    23    24    25    26    27    28    29
##         0   368  5490  8545   589  6147   335 14322 63952
##         1     5    54    85     3    55     5   143   559
```

```
colPerc(xtabs(~Applicant+C.O..22,data=data))
```

```
##          C.O..22
## Applicant     0      1
##     0      99.1  99.06
##     1       0.9   0.94
##     Total 100.0 100.00
```

```
xtabs(~Applicant+C.O..22,data=data)
```

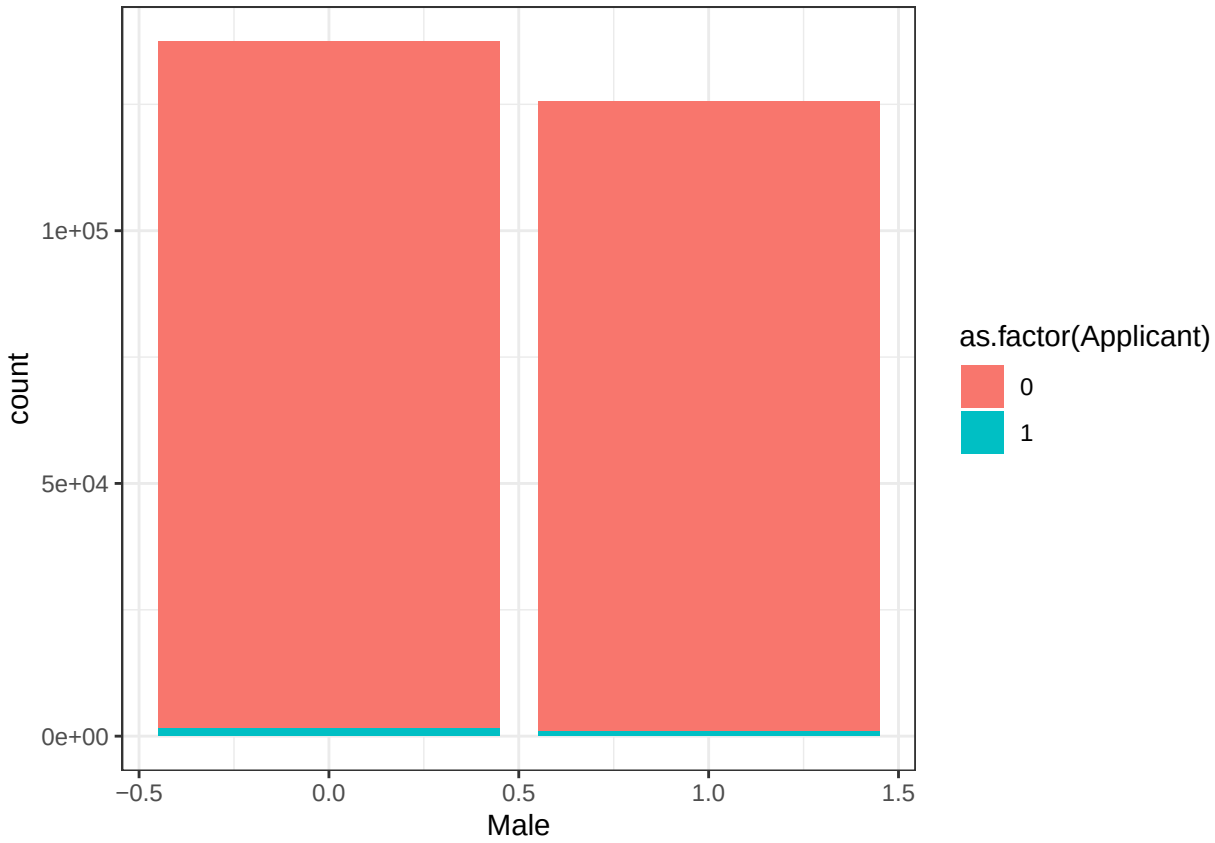
```
##          C.O..22
## Applicant      0      1
##         0 122752 137952
##         1   1116   1316
```

```
colPerc(xtabs(~Applicant+C.O..22,data=data))
```

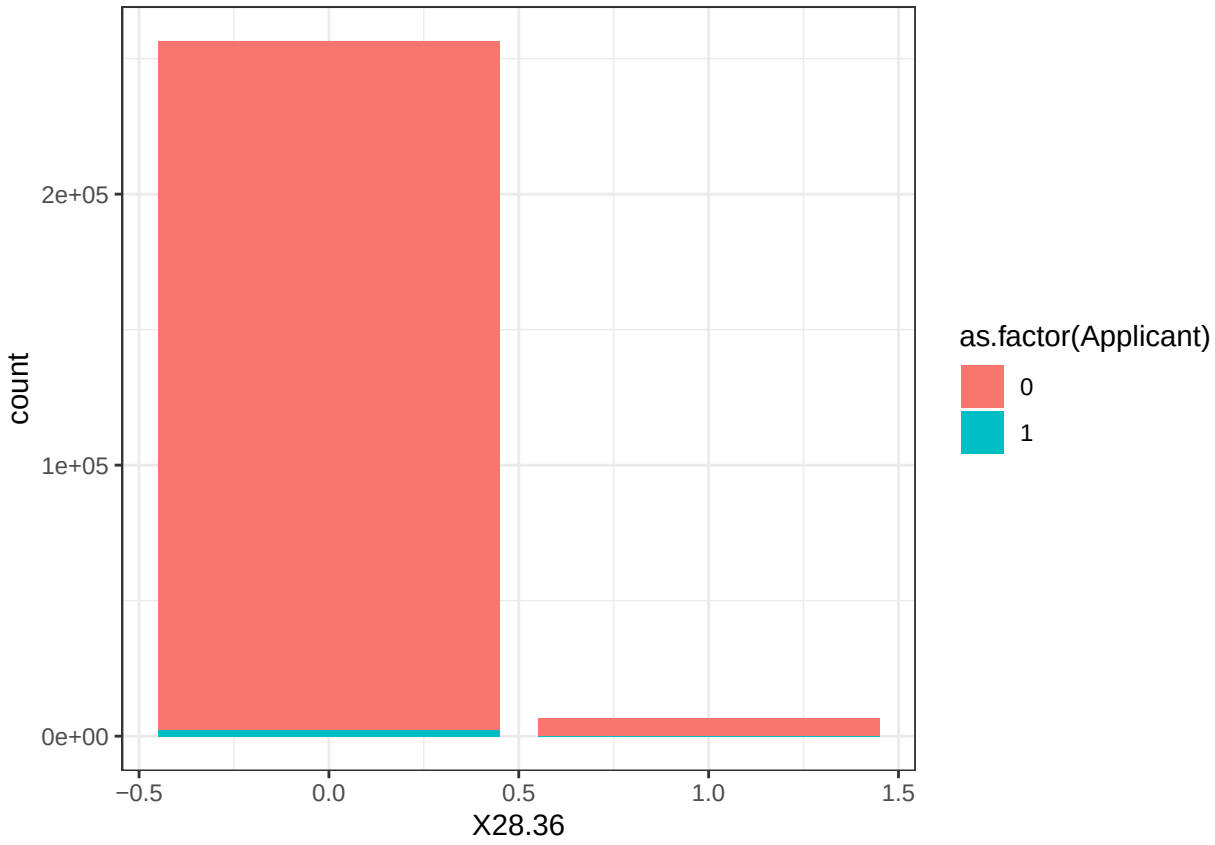
```
##          C.O..22
## Applicant     0      1
##     0      99.1  99.06
##     1       0.9   0.94
##     Total 100.0 100.00
```

```
#For comprehensive pdf data visualizations
#DataExplorer::create_report(data)
```

```
ggplot(data, aes(x = Male, fill = as.factor(Applicant) )) + theme_bw()+ geom_bar()
```

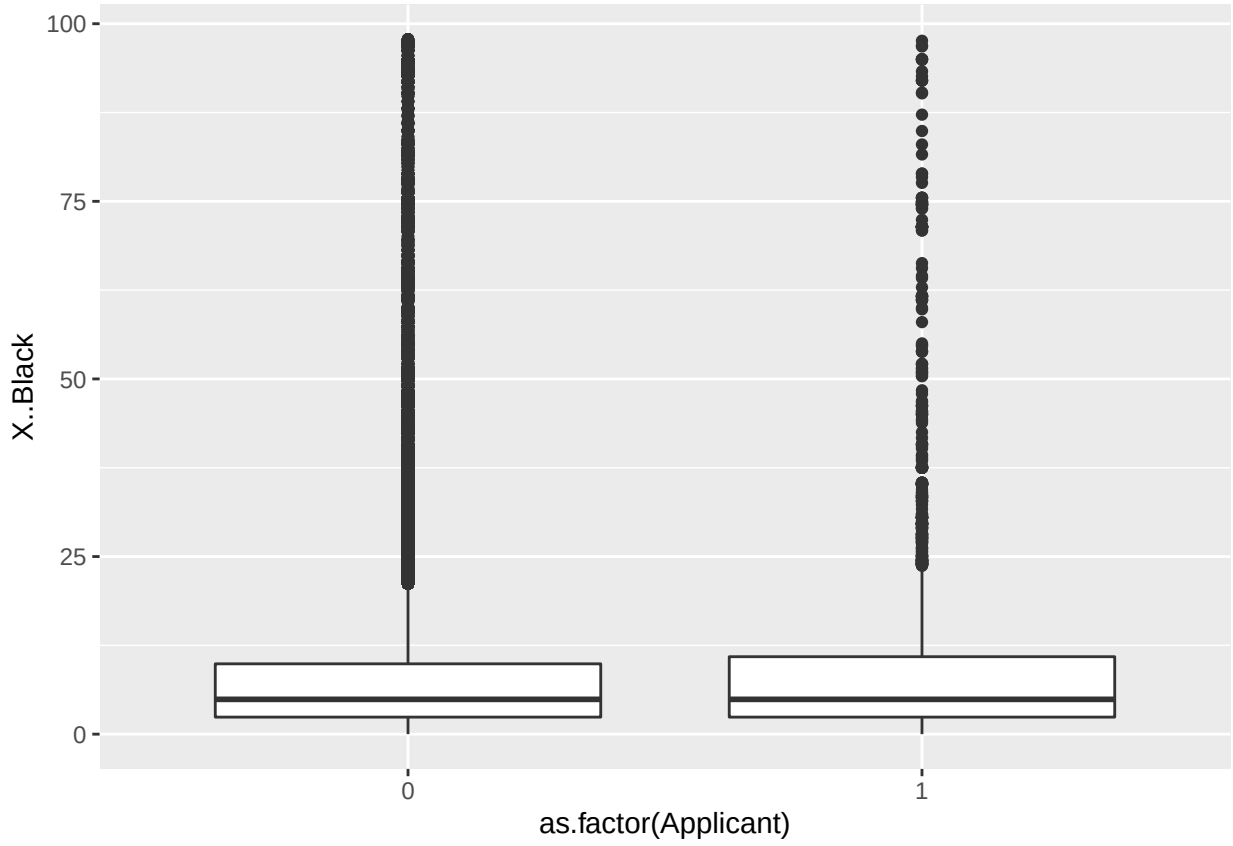



```
ggplot(data, aes(x = X28.36, fill = as.factor(Applicant) )) + theme_bw()+ geom_bar()
```

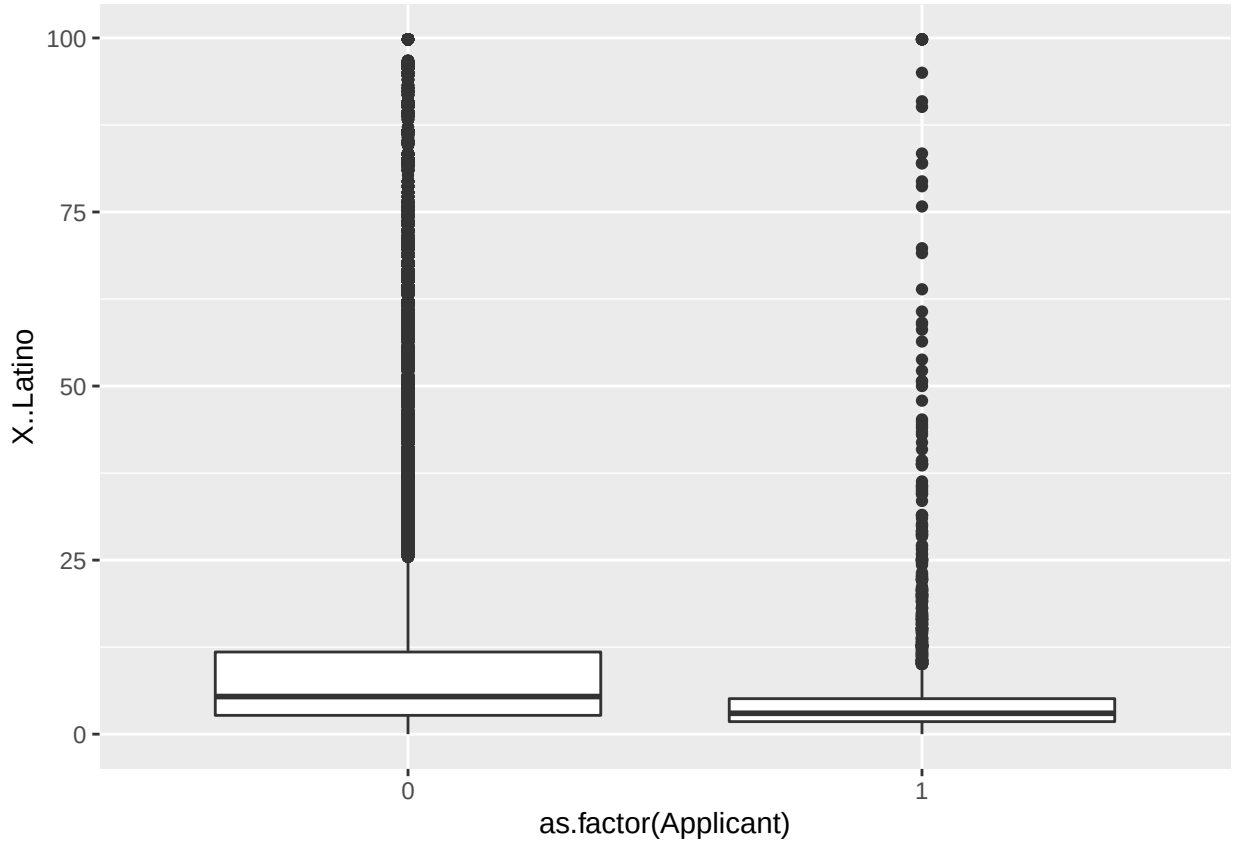



```
cor_data <- subset(data, select = -c(State) )
```

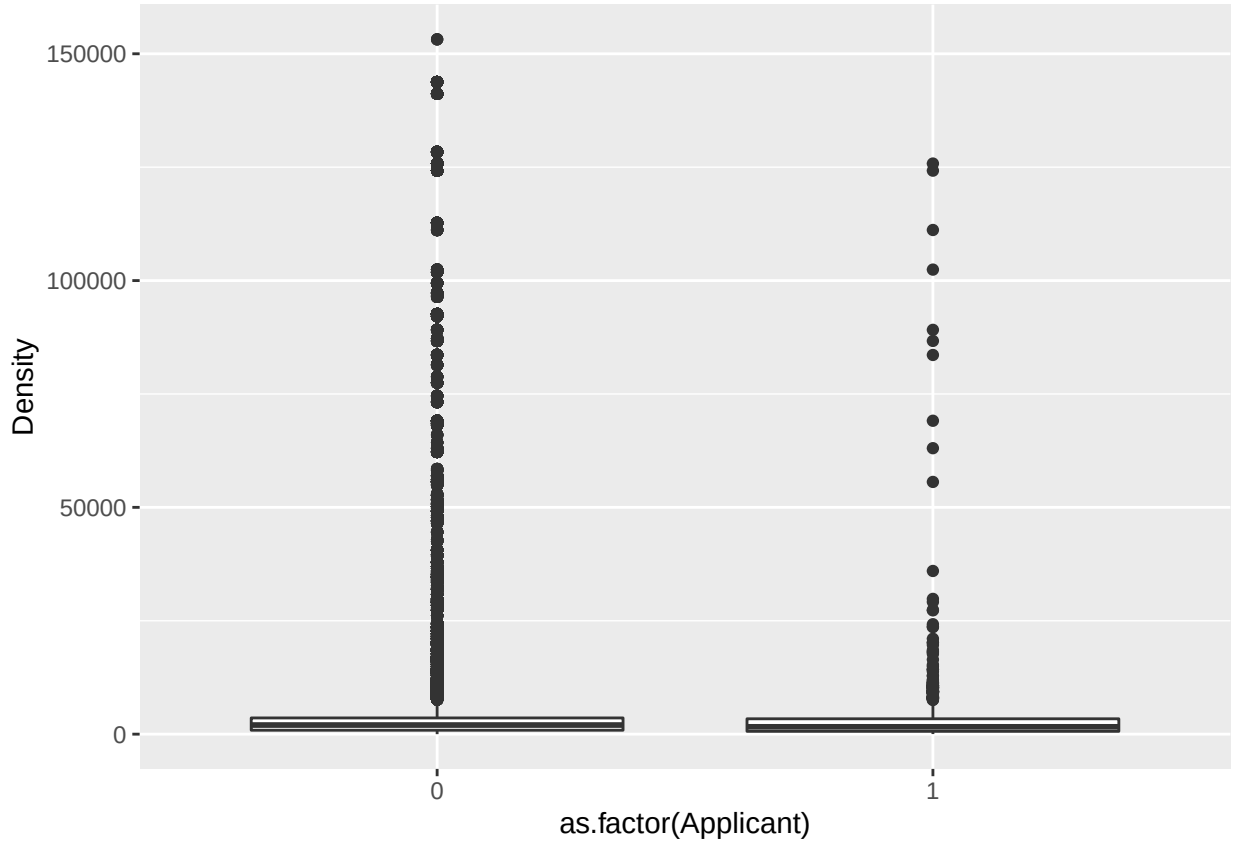
```
# Boxplot (X..Black and Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=X..Black)) +
  geom_boxplot()
```




```
# Boxplot (X..Latinoand Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=X..Latino)) +
  geom_boxplot()
```

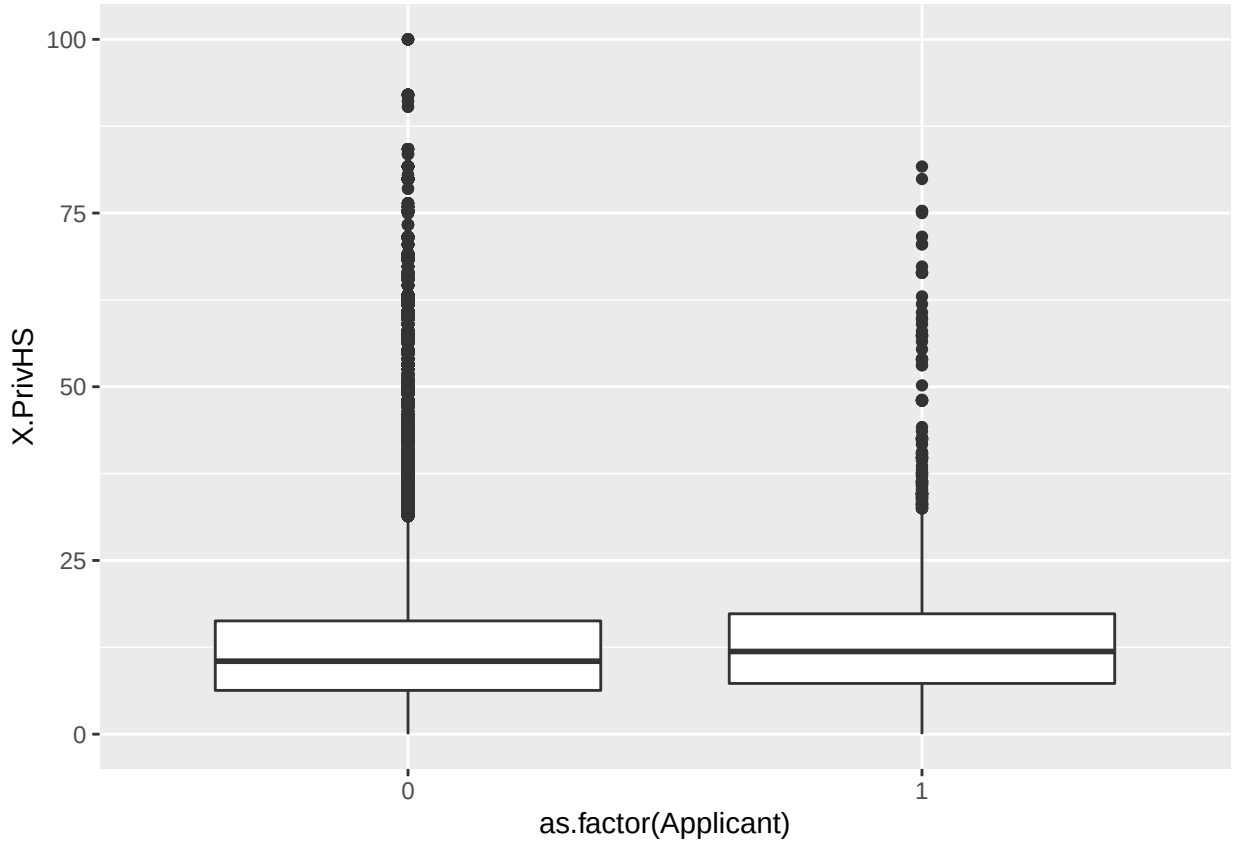



```
# Boxplot (X..Latinoand Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=Density)) +
  geom_boxplot()
```



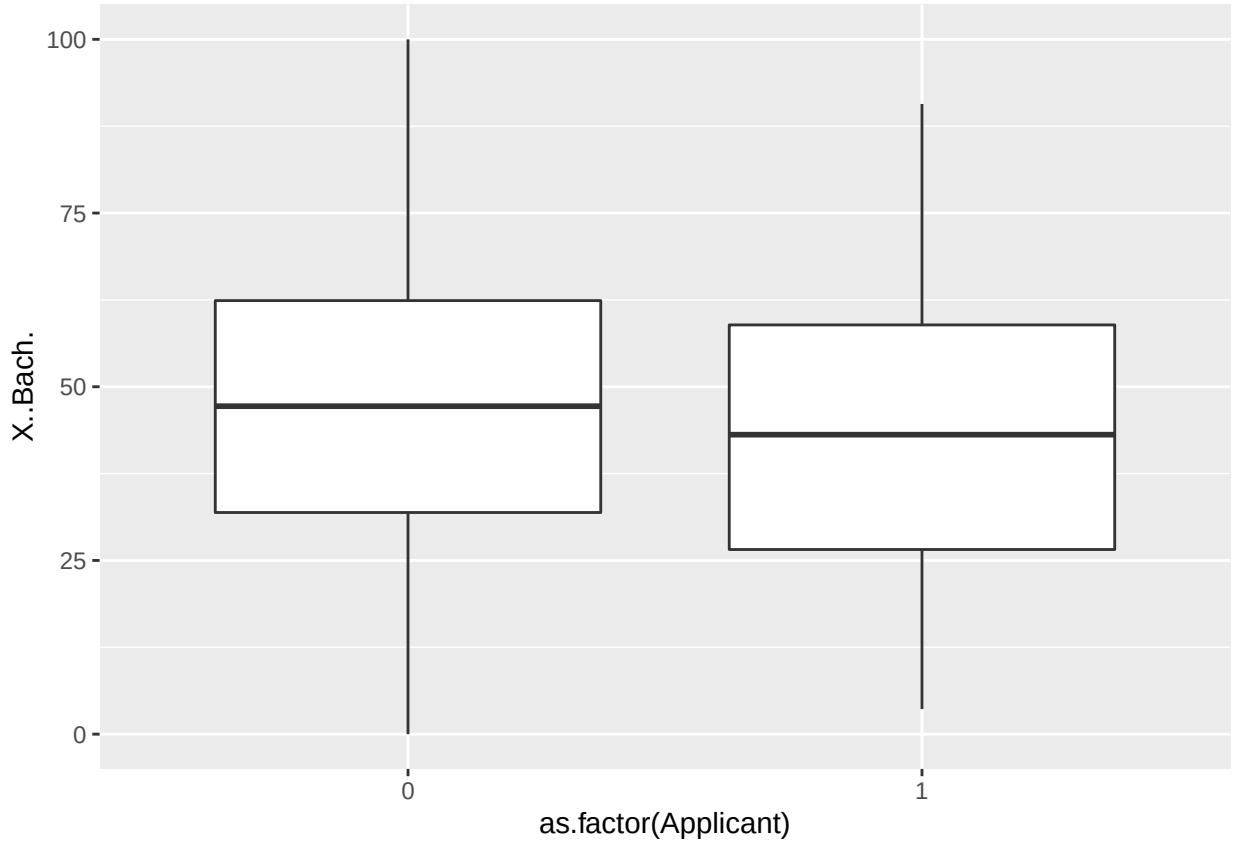

```
# Boxplot (X..Latinoand Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=X.PrivHS)) +
  geom_boxplot()
```

```
## Warning: Removed 128 rows containing non-finite values (stat_boxplot).
```



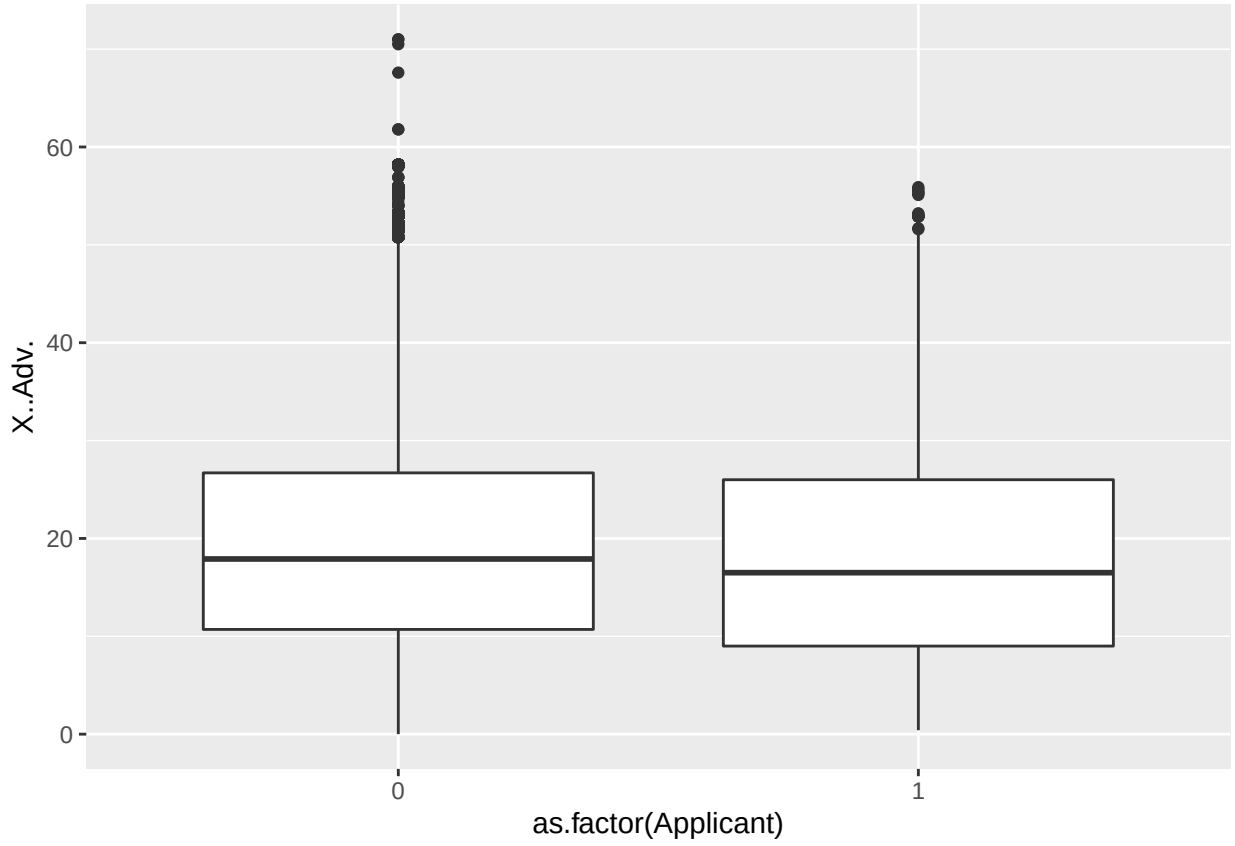

```
# Boxplot (X..Latinoand Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=X..Bach.)) +
  geom_boxplot()
```

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```



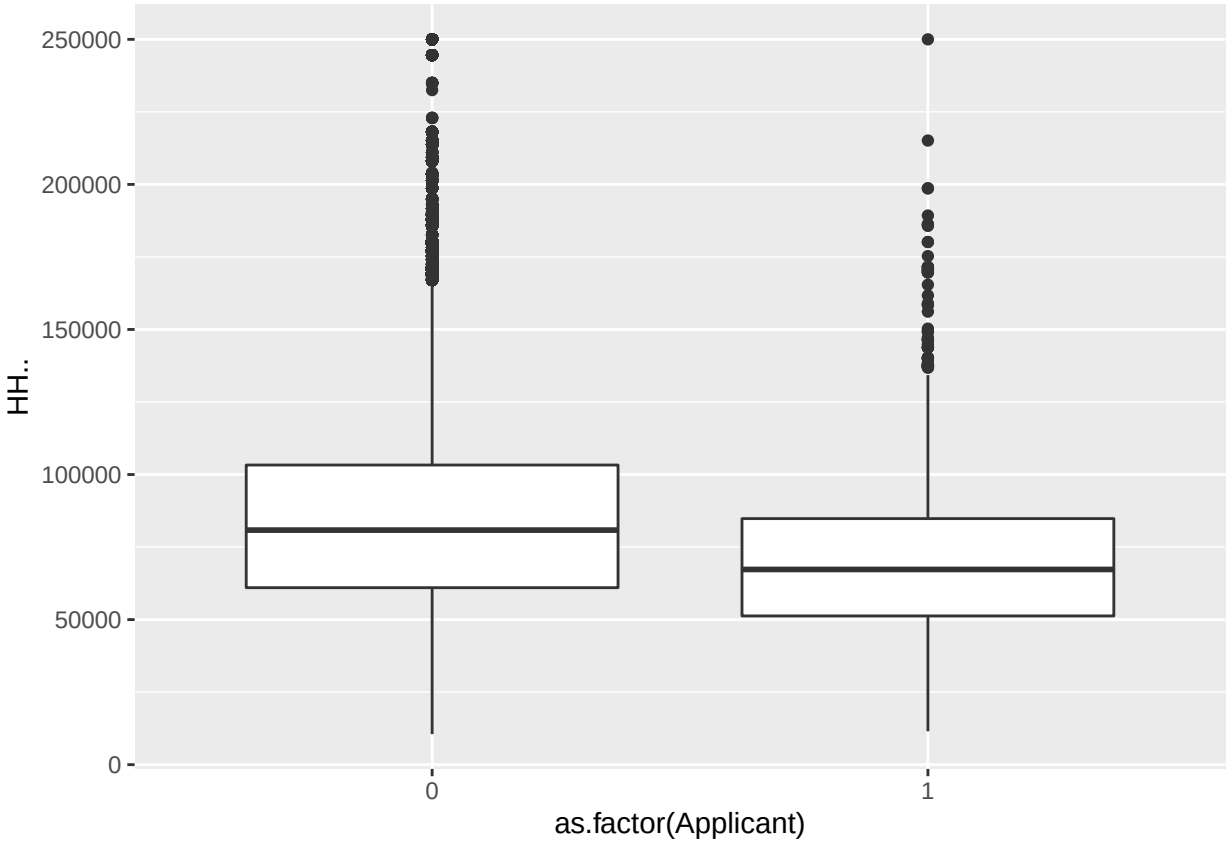

```
ggplot(data, aes(x=as.factor(Applicant), y=X..Adv.)) +
  geom_boxplot()
```

```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```



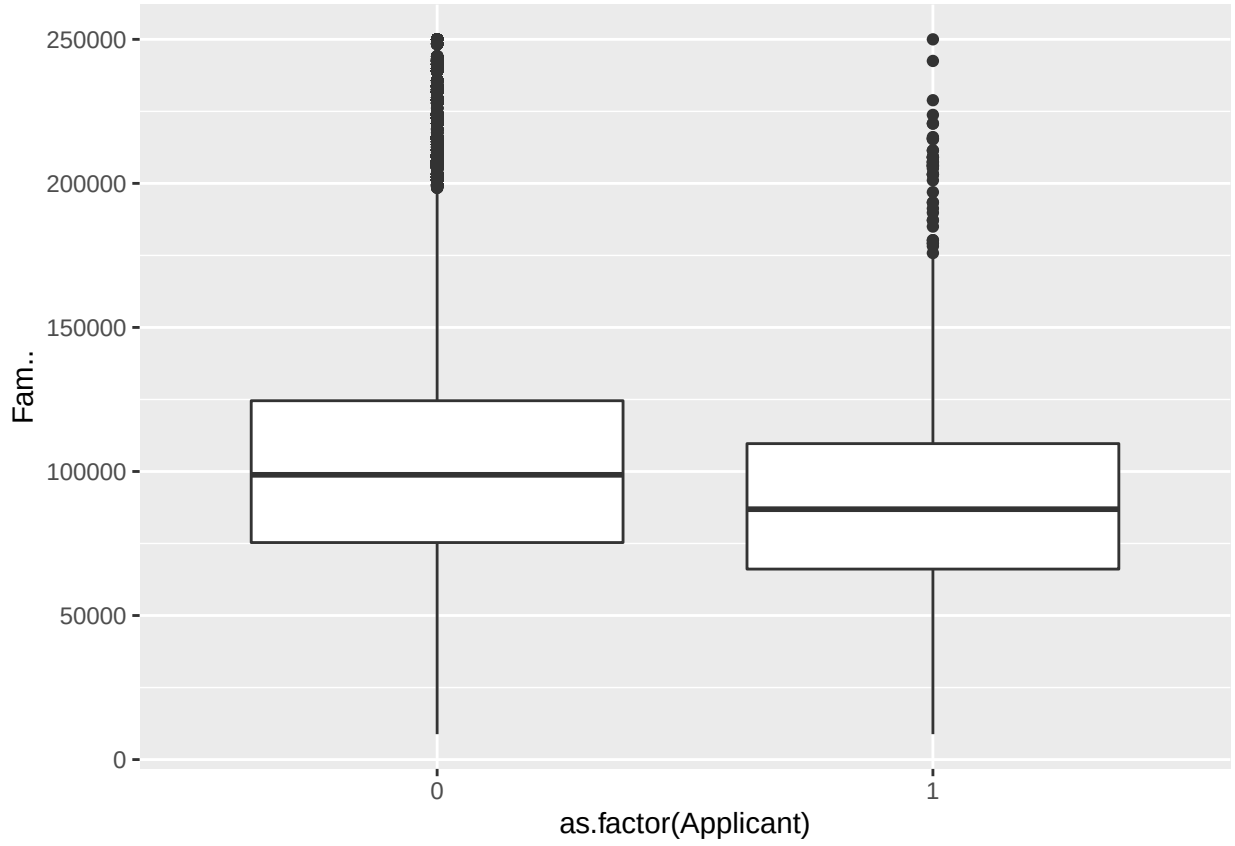

```
# Boxplot (X..Latinoand Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=HH..)) +
  geom_boxplot()
```

```
## Warning: Removed 70 rows containing non-finite values (stat_boxplot).
```

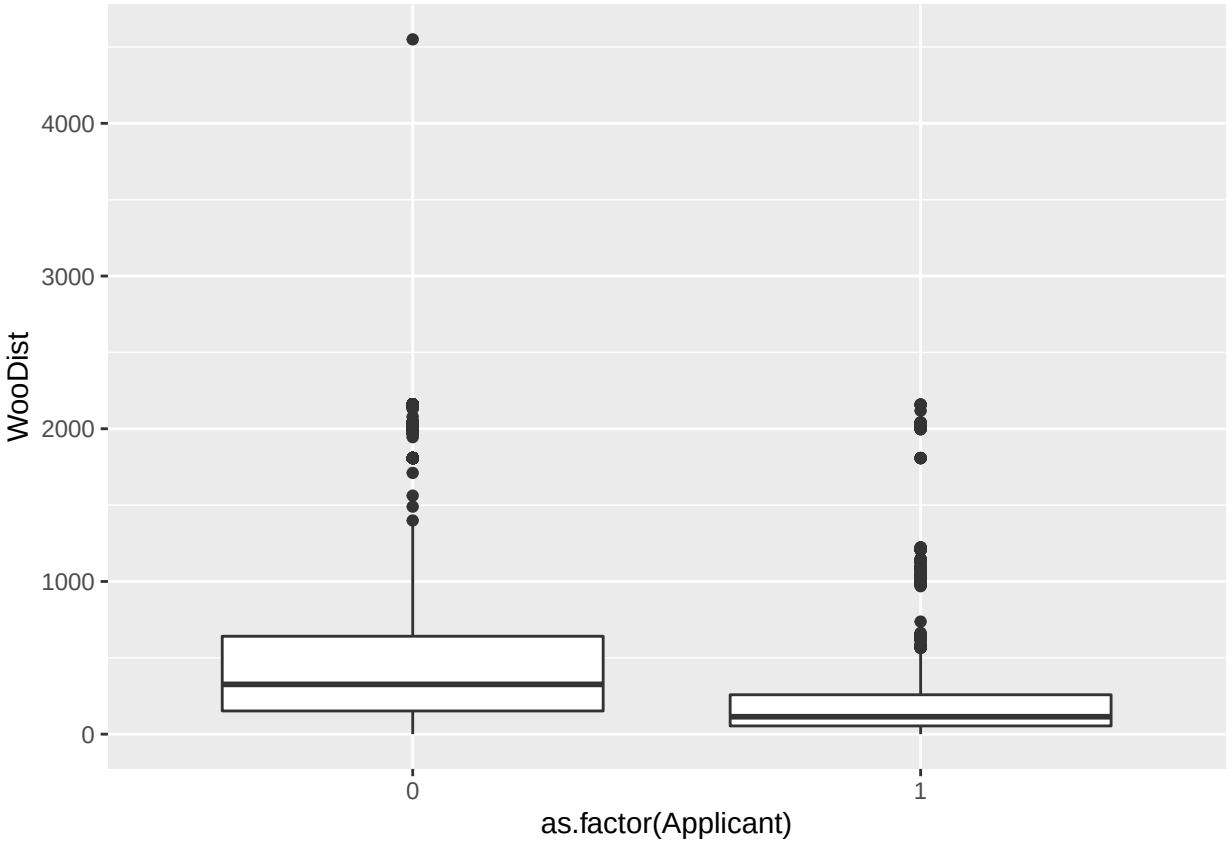



```
# Boxplot (X..Latinoand Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=Fam..)) +
  geom_boxplot()
```

```
## Warning: Removed 101 rows containing non-finite values (stat_boxplot).
```




```
# Boxplot (WooDist and Applicant)
ggplot(data, aes(x=as.factor(Applicant), y=WooDist)) +
  geom_boxplot()
```




```
unique(data$CSA)
```

```
##  [1] 378  NA 294 500 176 198 206 548 220 408 288 160 184 430 178 330 148 348 476
## [20] 266 370 122 278 216 240 212 400 350 310 464 320 534 532 566 488 336 170 360
## [39] 515 338 390 394 359 425 248 268 434 428 316 296 306 545 462 106 236 376 458
## [58] 496 162 356 422 260 452 357 204 412 142 118 314 450 304 312 209 297 365 221
## [77] 150 420 520 154 554 276 144 340 438 482 300 218 406 238
```

```
model1 <- glm(Applicant ~ WooDist + (Male) + (URM) + (X28.36) + (X23.27)  + Density + Soph + X..Black +
summary(model1)
```

```
##
## Call:
## glm(formula = Applicant ~ WooDist + (Male) + (URM) + (X28.36) +
##     (X23.27) + Density + Soph + X..Black + X..Latino + HH.. +
##     Fam.. + Metro + Lower + Upper + UpMid + C.O..22, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5029  -0.1590  -0.1160  -0.0770   4.2205
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -3.054e+00  2.464e-01 -12.393  < 2e-16 ***
## WooDist     -1.548e-03  1.051e-04 -14.734  < 2e-16 ***
## Male        -2.805e-01  4.182e-02  -6.706 2.00e-11 ***
## URM          2.828e-01  6.302e-02   4.488 7.18e-06 ***
## X28.36       1.915e-01  9.861e-02   1.942 0.052197 .
## X23.27       4.281e-01  8.145e-02   5.255 1.48e-07 ***
## Density     -9.542e-06  3.420e-06  -2.790 0.005268 **
## Soph         5.773e-01  4.525e-02  12.760  < 2e-16 ***
## X..Black     1.337e-03  1.896e-03   0.705 0.480534
## X..Latino   -1.244e-02  3.193e-03  -3.897 9.74e-05 ***
## HH..        -2.768e-05  2.396e-06 -11.552  < 2e-16 ***
## Fam..        7.595e-06  3.524e-06   2.155 0.031136 *
## Metro       -3.008e-01  9.057e-02  -3.321 0.000898 ***
## Lower        3.770e-03  5.796e-03   0.650 0.515379
## Upper        2.115e-02  5.828e-03   3.629 0.000285 ***
## UpMid        6.671e-03  6.013e-03   1.109 0.267275
## C.O..22     -6.940e-02  4.131e-02  -1.680 0.092955 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27622  on 263023  degrees of freedom
## Residual deviance: 25838  on 263007  degrees of freedom
##   (112 observations deleted due to missingness)
## AIC: 25872
##
## Number of Fisher Scoring iterations: 9
```

```
model2 <- glm(Applicant ~ WooDist + (Male) + (URM) + (X28.36) + (X23.27)  + Density + Soph + X..Black +
summary(model2)
```

```
##
## Call:
## glm(formula = Applicant ~ WooDist + (Male) + (URM) + (X28.36) +
##     (X23.27) + Density + Soph + X..Black + X..Latino + HH.. +
##     Fam.. + Metro + Lower + LowMid + Mid + Upper + UpMid + C.O..22,
##     family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5586  -0.1591  -0.1161  -0.0767   4.2260
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.727e+00  3.130e+00  -0.871 0.383666
## WooDist     -1.547e-03  1.051e-04 -14.727  < 2e-16 ***
## Male        -2.808e-01  4.182e-02  -6.715 1.88e-11 ***
## URM          2.838e-01  6.299e-02   4.505 6.65e-06 ***
## X28.36       1.901e-01  9.862e-02   1.928 0.053906 .
## X23.27       4.279e-01  8.145e-02   5.253 1.50e-07 ***
## Density     -9.572e-06  3.429e-06  -2.792 0.005245 **
## Soph         5.767e-01  4.524e-02  12.747  < 2e-16 ***
## X..Black     1.433e-03  1.898e-03   0.755 0.450225
```

```
## X..Latino   -1.175e-02  3.218e-03  -3.650 0.000262 ***
## HH..        -2.802e-05  2.411e-06 -11.621  < 2e-16 ***
## Fam..        7.182e-06  3.554e-06   2.021 0.043288 *
## Metro       -3.107e-01  9.088e-02  -3.419 0.000629 ***
## Lower        4.193e-03  3.149e-02   0.133 0.894061
## LowMid      -1.348e-02  3.154e-02  -0.428 0.668963
## Mid          3.724e-03  3.136e-02   0.119 0.905457
## Upper        1.972e-02  3.108e-02   0.635 0.525714
## UpMid        3.688e-04  3.112e-02   0.012 0.990544
## C.O..22     -7.055e-02  4.131e-02  -1.708 0.087689 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27622  on 263023  degrees of freedom
## Residual deviance: 25835  on 263005  degrees of freedom
##   (112 observations deleted due to missingness)
## AIC: 25873
##
## Number of Fisher Scoring iterations: 9
```

```
model3 <- glm(Applicant ~ WooDist + (Male) + (URM) + (X28.36) + (X23.27)  + Density + Soph  + X..Latino
summary(model3)
```

```
##
## Call:
## glm(formula = Applicant ~ WooDist + (Male) + (URM) + (X28.36) +
##     (X23.27) + Density + Soph + X..Latino + HH.. + Metro + LowMid +
##     Upper + C.O..22, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5763  -0.1593  -0.1165  -0.0760   4.1896
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.200e+00  2.305e-01  -9.545  < 2e-16 ***
## WooDist     -1.544e-03  1.047e-04 -14.741  < 2e-16 ***
## Male        -2.816e-01  4.182e-02  -6.734 1.65e-11 ***
## URM          2.947e-01  5.995e-02   4.916 8.82e-07 ***
## X28.36       1.845e-01  9.856e-02   1.872 0.061254 .
## X23.27       4.227e-01  8.138e-02   5.195 2.05e-07 ***
## Density     -8.055e-06  3.319e-06  -2.427 0.015221 *
## Soph         5.739e-01  4.518e-02  12.703  < 2e-16 ***
## X..Latino   -1.311e-02  3.142e-03  -4.174 2.99e-05 ***
## HH..        -2.534e-05  1.858e-06 -13.640  < 2e-16 ***
## Metro       -3.098e-01  8.638e-02  -3.586 0.000335 ***
## LowMid      -1.537e-02  6.541e-03  -2.350 0.018759 *
## Upper        2.641e-02  3.613e-03   7.309 2.69e-13 ***
## C.O..22     -7.022e-02  4.129e-02  -1.701 0.089030 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27623  on 263065  degrees of freedom
## Residual deviance: 25841  on 263052  degrees of freedom
##   (70 observations deleted due to missingness)
## AIC: 25869
##
## Number of Fisher Scoring iterations: 9
```

```
model_i1 <- glm(Inquiry ~ WooDist + C.O..22 + (Male) + (URM) + (X28.36) + (X23.27)  + Density + Soph + I
summary(model_i1)
```

```
##
## Call:
## glm(formula = Inquiry ~ WooDist + C.O..22 + (Male) + (URM) +
##     (X28.36) + (X23.27) + Density + Soph + Density + X.PrivHS +
##     X..Black + X..Latino + HH.. + Fam.. + Metro + Lower + LowMid +
##     Mid + Upper + UpMid, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3666  -0.2898  -0.2266  -0.1783   3.4502
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.455e+00  1.555e+00  -3.508 0.000451 ***
## WooDist     -8.283e-04  3.996e-05 -20.730  < 2e-16 ***
## C.O..22     -9.998e-02  2.059e-02  -4.855 1.20e-06 ***
## Male        -3.021e-01  2.081e-02 -14.516  < 2e-16 ***
## URM          2.825e-01  3.139e-02   8.999  < 2e-16 ***
## X28.36       1.590e-01  5.002e-02   3.180 0.001474 **
## X23.27       3.886e-01  4.311e-02   9.014  < 2e-16 ***
## Density     -1.369e-05  1.900e-06  -7.205 5.82e-13 ***
## Soph         1.110e+00  2.262e-02  49.069  < 2e-16 ***
## X.PrivHS     5.922e-03  1.112e-03   5.327 1.00e-07 ***
## X..Black    -2.130e-03  9.738e-04  -2.187 0.028733 *
## X..Latino   -3.087e-03  1.242e-03  -2.485 0.012940 *
## HH..        -1.276e-05  1.318e-06  -9.678  < 2e-16 ***
## Fam..        7.235e-06  1.766e-06   4.096 4.20e-05 ***
## Metro       -4.196e-01  4.537e-02  -9.249  < 2e-16 ***
## Lower        5.472e-02  1.561e-02   3.506 0.000455 ***
## LowMid       1.807e-02  1.567e-02   1.153 0.248789
## Mid          3.292e-02  1.560e-02   2.110 0.034863 *
## Upper        2.744e-02  1.550e-02   1.771 0.076605 .
## UpMid        3.137e-02  1.549e-02   2.025 0.042847 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 87552  on 262963  degrees of freedom
## Residual deviance: 80213  on 262944  degrees of freedom
##   (172 observations deleted due to missingness)
## AIC: 80253
```

```
##
## Number of Fisher Scoring iterations: 7

# Model using all variables: predicting Applicant

model_all1 <- glm(Applicant ~ C.O..22 + (Male) + (URM) + (X28.36) + (X23.27) + Soph + Density + X..Blac
summary(model_all1)


##
## Call:
## glm(formula = Applicant ~ C.O..22 + (Male) + (URM) + (X28.36) +
##     (X23.27) + Soph + Density + X..Black + X..Latino + X.PrivHS +
##     X..Bach. + X..Adv. + HH.. + Fam.. + FamK.. + Lower + LowMid +
##     Mid + UpMid + Upper + Metro + WooDist, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6230  -0.1592  -0.1144  -0.0754   4.1481
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.714e+00  3.145e+00  -1.181 0.237611
## C.O..22     -7.078e-02  4.135e-02  -1.712 0.086974 .
## Male        -2.843e-01  4.184e-02  -6.795 1.09e-11 ***
## URM          2.847e-01  6.292e-02   4.524 6.07e-06 ***
## X28.36       1.737e-01  9.868e-02   1.760 0.078332 .
## X23.27       4.452e-01  8.152e-02   5.462 4.72e-08 ***
## Soph         5.690e-01  4.542e-02  12.528  < 2e-16 ***
## Density     -1.250e-05  3.821e-06  -3.271 0.001072 **
## X..Black    -1.413e-04  1.951e-03  -0.072 0.942264
## X..Latino   -1.038e-02  3.234e-03  -3.210 0.001327 **
## X.PrivHS     2.243e-03  2.220e-03   1.010 0.312461
## X..Bach.     1.650e-02  4.887e-03   3.376 0.000735 ***
## X..Adv.      2.216e-02  6.903e-03   3.211 0.001324 **
## HH..        -1.503e-05  2.879e-06  -5.221 1.77e-07 ***
## Fam..       -4.038e-06  4.701e-06  -0.859 0.390418
## FamK..       7.690e-06  2.529e-06   3.041 0.002357 **
## Lower        1.296e-02  3.162e-02   0.410 0.681990
## LowMid      -3.130e-03  3.169e-02  -0.099 0.921332
## Mid          5.891e-03  3.147e-02   0.187 0.851511
## UpMid       -7.894e-03  3.129e-02  -0.252 0.800847
## Upper       -2.275e-02  3.156e-02  -0.721 0.470927
## Metro       -3.548e-01  9.224e-02  -3.846 0.000120 ***
## WooDist     -1.439e-03  1.051e-04 -13.686  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27604  on 262569  degrees of freedom
## Residual deviance: 25720  on 262547  degrees of freedom
##   (566 observations deleted due to missingness)
```

```
## AIC: 25766
##
## Number of Fisher Scoring iterations: 9
```

As the p-values are too high, variables 'X..Black, LowMid, Mid, UpMid','XPrivHS' are eliminated from the model. Also, variables 'Fam..' and 'FamK..' might cause multicollinearity issues so the variable 'Fam..'

```
model_12 <- glm(Applicant ~ C.O..22 + (Male) + (URM) + (X28.36) + (X23.27) + Soph + Density + X..Latino
summary(model_12)
```

```
##
## Call:
## glm(formula = Applicant ~ C.O..22 + (Male) + (URM) + (X28.36) +
##     (X23.27) + Soph + Density + X..Latino + X..Bach. + X..Adv. +
##     HH.. + FamK.. + Lower + Upper + Metro + WooDist, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7979  -0.1589  -0.1148  -0.0755   4.1629
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.680e+00  2.064e-01 -17.831  < 2e-16 ***
## C.O..22     -7.176e-02  4.131e-02  -1.737 0.082346 .
## Male        -2.839e-01  4.184e-02  -6.785 1.16e-11 ***
## URM          2.853e-01  6.085e-02   4.688 2.76e-06 ***
## X28.36       1.745e-01  9.862e-02   1.769 0.076888 .
## X23.27       4.469e-01  8.147e-02   5.485 4.13e-08 ***
## Soph         5.723e-01  4.538e-02  12.613  < 2e-16 ***
## Density     -1.192e-05  3.745e-06  -3.184 0.001454 **
## X..Latino   -1.051e-02  3.167e-03  -3.319 0.000905 ***
## X..Bach.     1.486e-02  4.406e-03   3.372 0.000745 ***
## X..Adv.      2.221e-02  6.477e-03   3.429 0.000605 ***
## HH..        -1.733e-05  2.278e-06  -7.610 2.73e-14 ***
## FamK..       6.507e-06  2.016e-06   3.228 0.001248 **
## Lower        1.271e-02  4.683e-03   2.714 0.006654 **
## Upper       -2.428e-02  6.856e-03  -3.542 0.000398 ***
## Metro       -3.663e-01  8.756e-02  -4.183 2.87e-05 ***
## WooDist     -1.448e-03  1.049e-04 -13.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27605  on 262591  degrees of freedom
## Residual deviance: 25726  on 262575  degrees of freedom
##   (544 observations deleted due to missingness)
## AIC: 25760
##
## Number of Fisher Scoring iterations: 9
```

```
#res <- cor(data[, c('Lat','Long', 'C.O..22', 'MyID', 'Male', 'URM', 'Zip', 'X28.36', 'X23.27', 'Soph',

#round(res, 4)
```

```
data %>% count(State, sort = TRUE)
```

```
## # A tibble: 43 x 2
##    State     n
##    <chr> <int>
##  1 OH    41232
##  2 TX    34720
##  3 MI    27409
##  4 IL    24271
##  5 PA    15083
##  6 NY    14900
##  7 IN    12179
##  8 NJ    10806
##  9 CA     9707
## 10 MA     9345
## # ... with 33 more rows
```

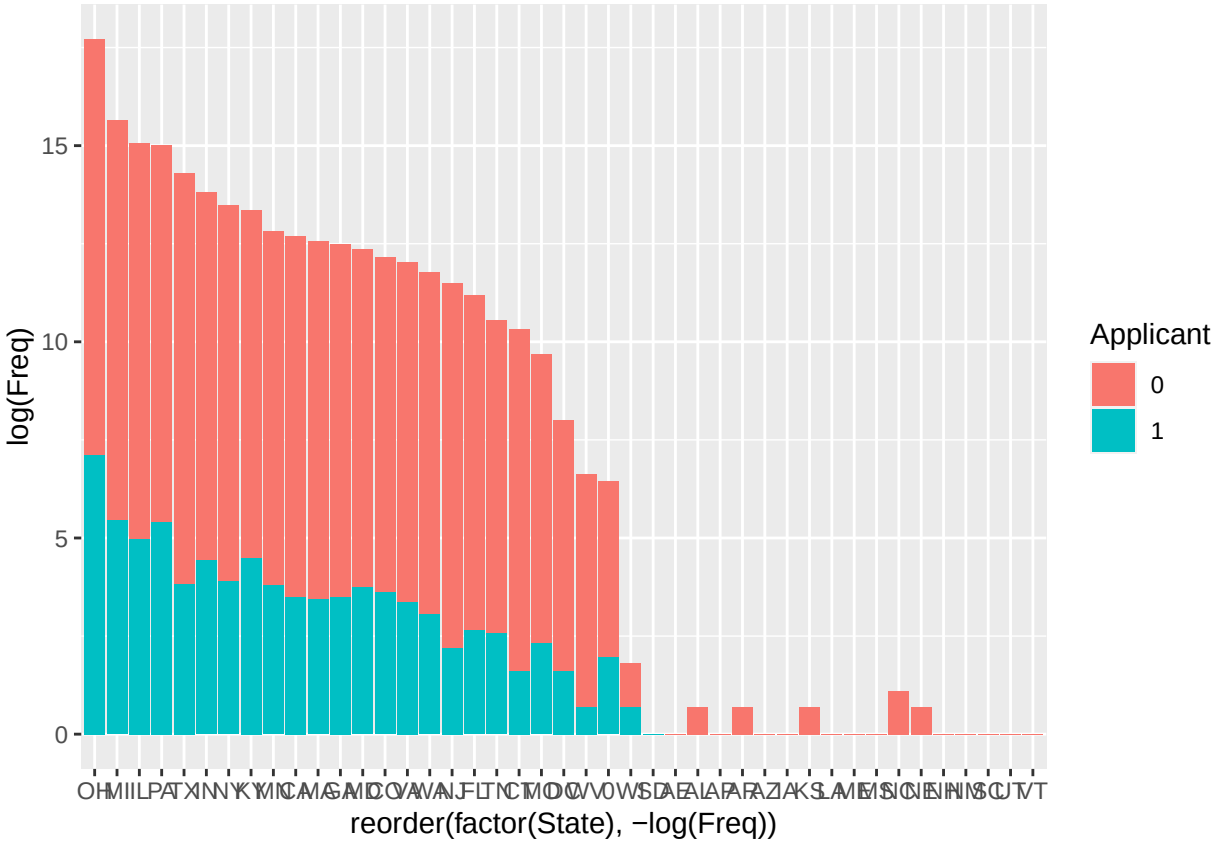
```
#ggplot(as.data.frame(tbl), aes(factor(Depth), Freq, fill = Species)) +
  #geom_col(position = 'dodge')
```

```
#ggplot(data, aes(fill=condition, y=value, x=specie)) +
    #geom_bar(position="dodge", stat="identity")
```

```
#ggplot(data, aes(x = reorder(State, MyID)), y = Applicant) + theme_bw()+ geom_bar()
#ggplot(tips2, aes(x = reorder(day, -perc), y = perc)) + geom_bar(stat = "identity")
```

```
tbl <- with(data, table(State, Applicant))
ggplot(as.data.frame(tbl), aes(x = reorder(factor(State), -log(Freq)), log(Freq), fill = Applicant)) +
  geom_col(position = 'stack')
```

```
## Warning: Removed 17 rows containing missing values (geom_col).
```




```
#ggplot(theTable, aes(x=reorder(Position, -table(Position)[Position]))) + geom_bar()
```