Stat 557 - Midterm

Carson Sievert

October 8, 2012

Problem 4:

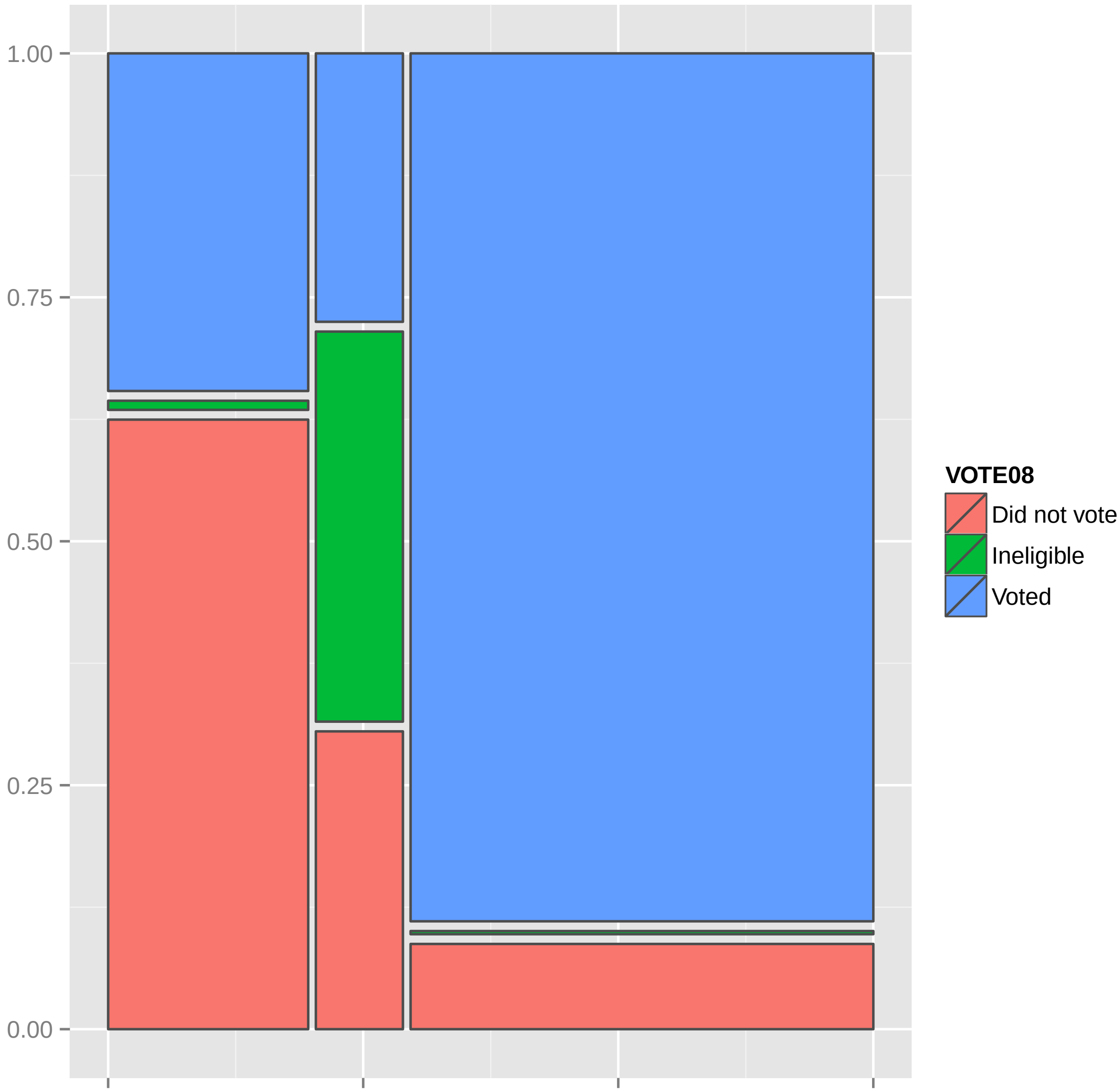
```
vote <- read.delim("http://www.hofroe.net/stat557/GSS%20data%20csv/GSS%202010.sav.csv",
    sep = ",", header = T)
idx <- c("VOTE08", "vote04", "sex", "age", "partyid", "educ")
vote2 <- na.omit(vote[, idx])
# Remove factors in VOTE08 that have a low count and those who 'don't
# know' if they voted in '04 Educ of 98 means 'Don't Know' and 99 means
# 'No answer'. To help with interpretation, remove these cases as well.
vote3 <- subset(vote2, vote04 != "DONT KNOW/REMEMBER" & VOTE08 != "No answer" &
    VOTE08 != "DON'T KNOW" & educ != 98 & educ != 99)
vote3$vote04 <- factor(vote3$vote04)
vote3$VOTE08 <- factor(vote3$VOTE08)
```

```
require(ggplot2)
require(productplots)
require(plyr)
require(reshape2)
```

part (a)

```
f <- subset(as.data.frame(xtabs(~VOTE08 + vote04 + sex + partyid, data = vote3)),
    Freq > 0)
prodplot(f, Freq ~ VOTE08 + vote04, c("vspine", "hspine")) + aes(fill = VOTE08)

## Scale for 'x' is already present.  Adding another scale for 'x', which will replace the
existing scale.
```



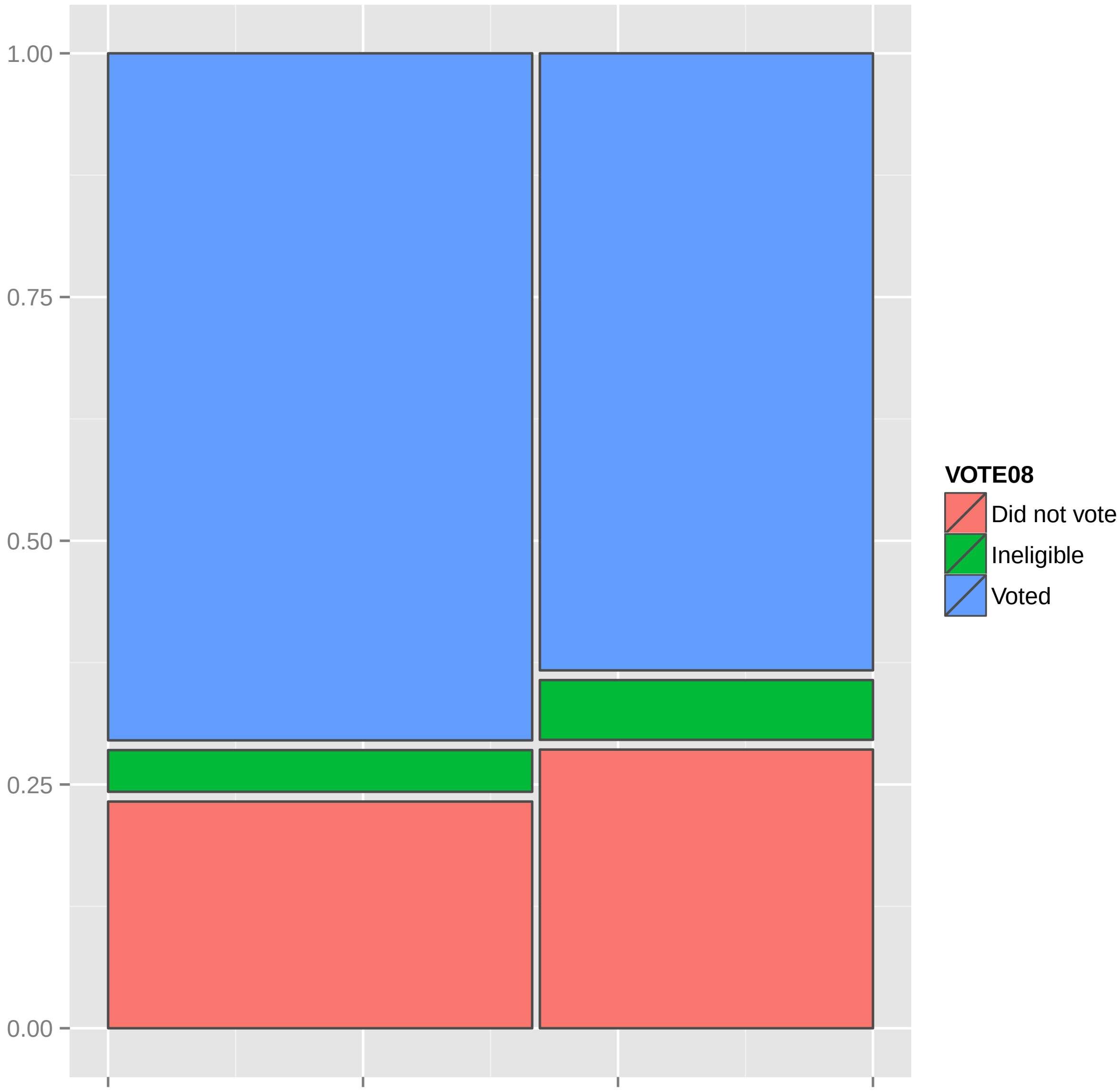

```
levels(f$vote04)

## [1] "DID NOT VOTE" "INELIGIBLE"   "VOTED"
```

One's actions in 2004 are higly indicative of their behavior in 2008. That is, given someone voted in 2004, they are much more likely to vote in 2008. Given ineligiblity in 2004, they are much more likely to be ineligibile in 2008. Also, given that someone did not vote in 2004, they are most likely not going to vote 2008.

```
prodplot(f, Freq ~ VOTE08 + sex, c("vspine", "hspine")) + aes(fill = VOTE08)

## Scale for 'x' is already present.  Adding another scale for 'x', which will replace the
existing scale.
```



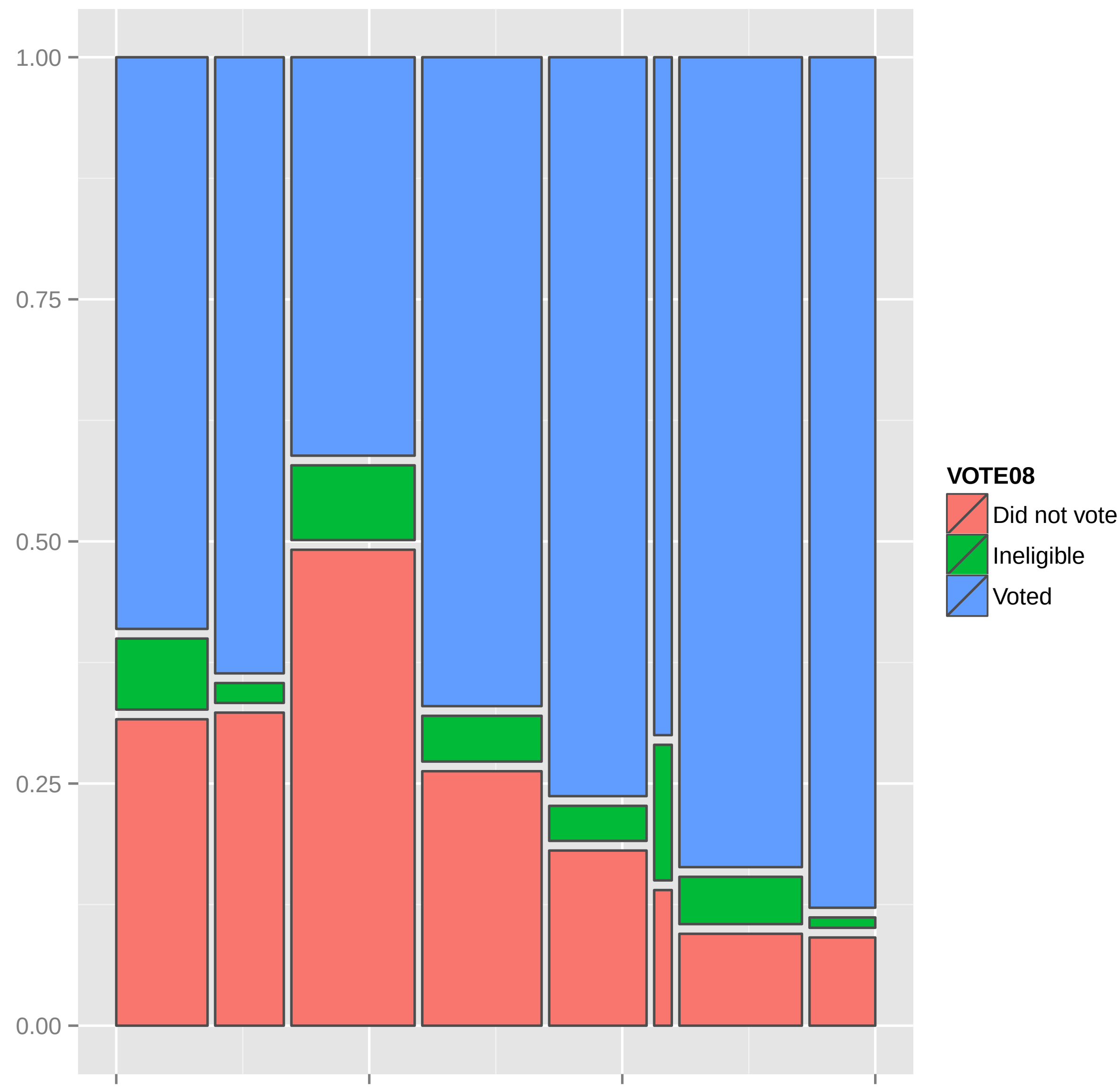

```
levels(f$sex)

## [1] "FEMALE" "MALE"
```

It appears females are more likely to vote compared to males. Males seem to be more likely to be ineligible.

```
prodplot(f, Freq ~ VOTE08 + partyid, c("vspine", "hspine")) + aes(fill = VOTE08)

## Scale for 'x' is already present. Adding another scale for 'x', which will replace the
existing scale.
```

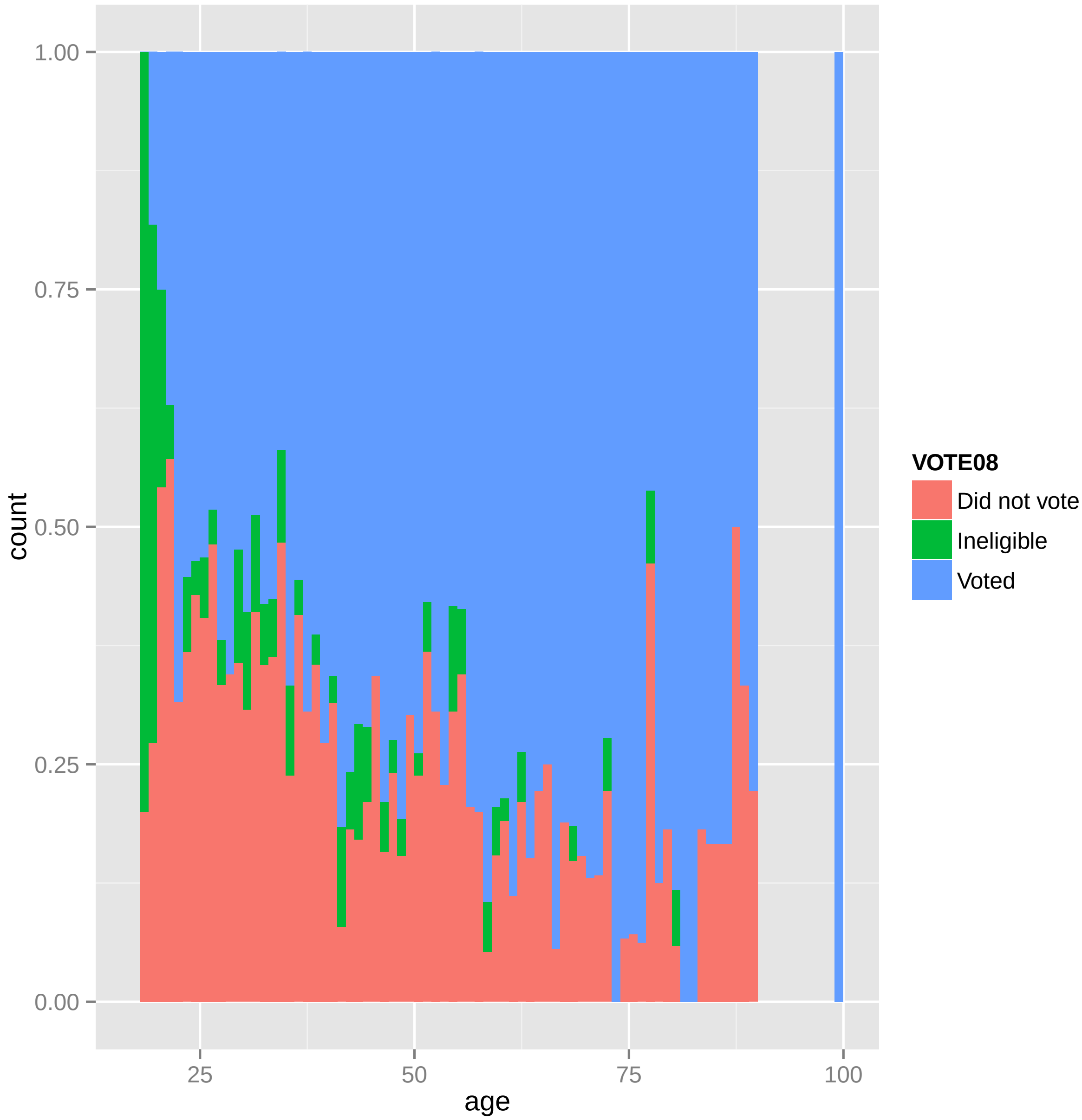



```
levels(f$partyid)

## [1] "IND,NEAR DEM"       "IND,NEAR REP"       "INDEPENDENT"
## [4] "NOT STR DEMOCRAT"   "NOT STR REPUBLICAN" "OTHER PARTY"
## [7] "STRONG DEMOCRAT"    "STRONG REPUBLICAN"
```

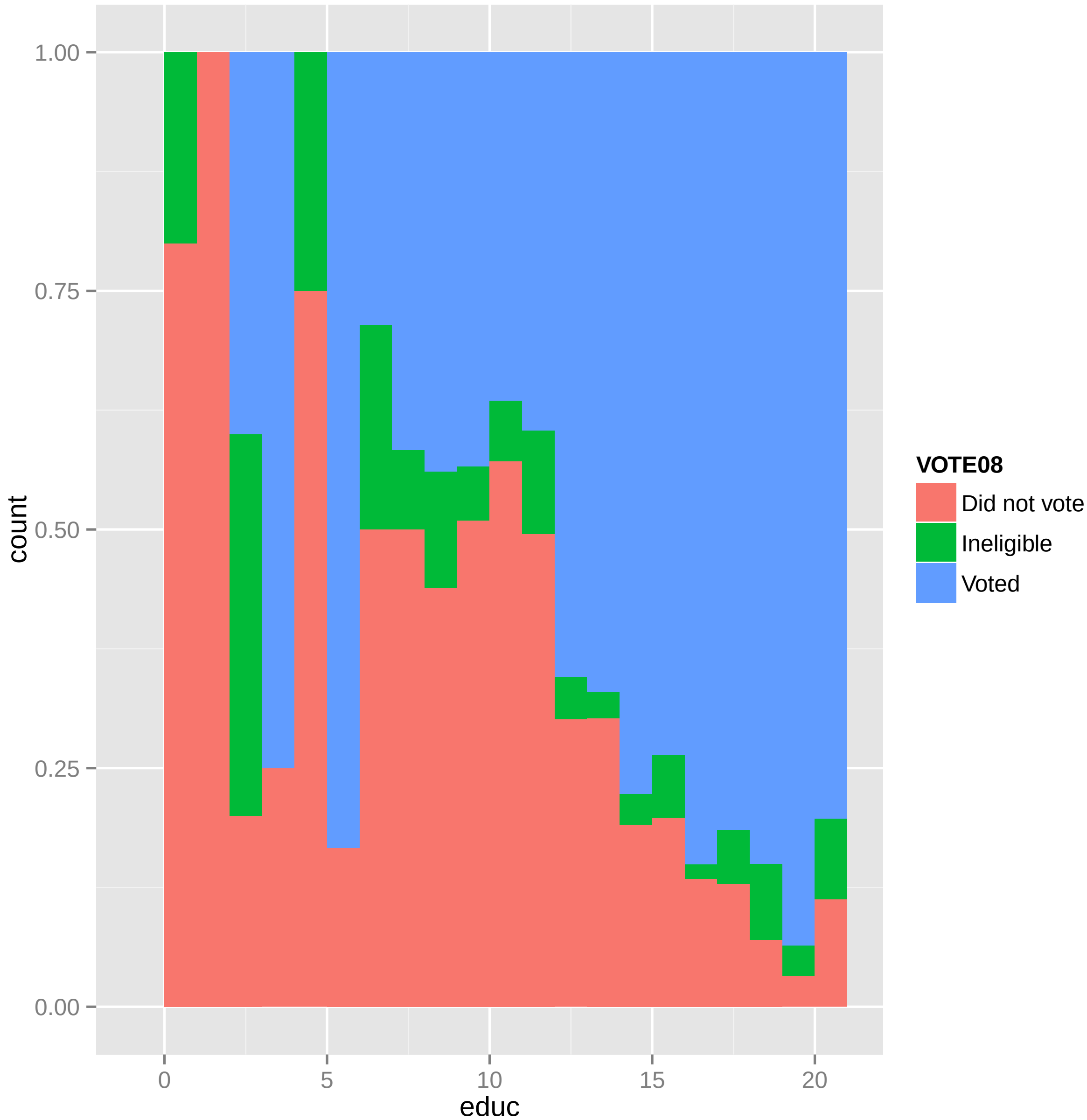
Independents are the least likely to vote out of any party affliation.

```
qplot(age, geom = "histogram", position = "fill", fill = VOTE08, binwidth = 1,
    data = vote3)
```



The proportion of people who vote increases with age.

```
qplot(educ, geom = "histogram", position = "fill", fill = VOTE08, binwidth = 1,
    data = vote3)
```



The proportion of people who vote increases with "highest year of school completed".

part (b)

```
library(nnet)
null <- multinom(VOTE08 ~ 1, data = vote3)
vote_04 <- multinom(VOTE08 ~ vote04, data = vote3)
sex <- multinom(VOTE08 ~ sex, data = vote3)
partyid <- multinom(VOTE08 ~ partyid, data = vote3)
educ <- multinom(VOTE08 ~ educ, data = vote3)
age <- multinom(VOTE08 ~ age, data = vote3)
```

```
anova(vote_04, null)

## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##    Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1      1      3926       2994
## 2 vote04      3922       2009 1 vs 2     4    985.2       0

anova(sex, null)

## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##   Model Resid. df Resid. Dev   Test    Df LR stat.  Pr(Chi)
## 1     1      3926       2994
## 2   sex      3924       2982 1 vs 2     2    12.53 0.001901

anova(partyid, null)

## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##     Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1       1      3926       2994
## 2 partyid      3912       2747 1 vs 2    14    246.8       0

anova(educ, null)

## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##   Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1     1      3926       2994
## 2  educ      3924       2812 1 vs 2     2    182.6       0

anova(age, null)

## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##   Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1     1      3926       2994
## 2   age      3924       2840 1 vs 2     2    154.5       0
```

All of the test for main effects are significant. Next, we compare the fit of the full model to the model with main effects.

```
full <- multinom(VOTE08 ~ vote04 * sex * partyid * educ * age, data = vote3)
main <- multinom(VOTE08 ~ vote04 + sex + partyid + educ + age, data = vote3)
```

```
anova(full, main)

## Likelihood ratio tests of Multinomial Models
##
```

```
## Response: VOTE08
##                                Model Resid. df Resid. Dev   Test    Df
## 1 vote04 + sex + partyid + educ + age      3902       1809
## 2 vote04 * sex * partyid * educ * age      3554       1708 1 vs 2   348
##   LR stat. Pr(Chi)
## 1
## 2    100.6       1
```

```r
anova(main, null)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##                                Model Resid. df Resid. Dev   Test    Df
## 1                                  1      3926       2994
## 2 vote04 + sex + partyid + educ + age      3902       1809 1 vs 2    24
##   LR stat. Pr(Chi)
## 1
## 2     1185       0
```

There is no siginifcant improvement going from the model with main effects to the full model. As a result, it's reasonable to ignore interactions. As anticipated, the main effects model is a great improvement from the null model. The main effects model seems like the appropriate model here.

part (c)

```r
coef(main)
```

```
##            (Intercept) vote04INELIGIBLE vote04VOTED sexMALE
## Ineligible      -8.136           5.9341      0.3948  0.1009
## Voted           -3.233           0.6816      2.4613 -0.3766
##            partyidIND,NEAR REP partyidINDEPENDENT partyidNOT STR DEMOCRAT
## Ineligible             -0.9127            0.07705                -0.01038
## Voted                  -0.2656           -0.57873                 0.32828
##            partyidNOT STR REPUBLICAN partyidOTHER PARTY
## Ineligible                   0.09701             2.3204
## Voted                        0.41771             0.4666
##            partyidSTRONG DEMOCRAT partyidSTRONG REPUBLICAN   educ     age
## Ineligible                  1.171                  -0.1463 0.0544 0.05884
## Voted                       1.288                   0.9482 0.1635 0.01429
```

```r
coef(main)[1, 4]
```

```
## [1] 0.1009
```

Switching from female to male, the logs odds for ineligibility in 2008 relative to *not* voting in 2008 (ceteris paribus) is increased by a factor of about 0.1.

```r
coef(main)[2, 12]
```

```
## [1] 0.1635
```

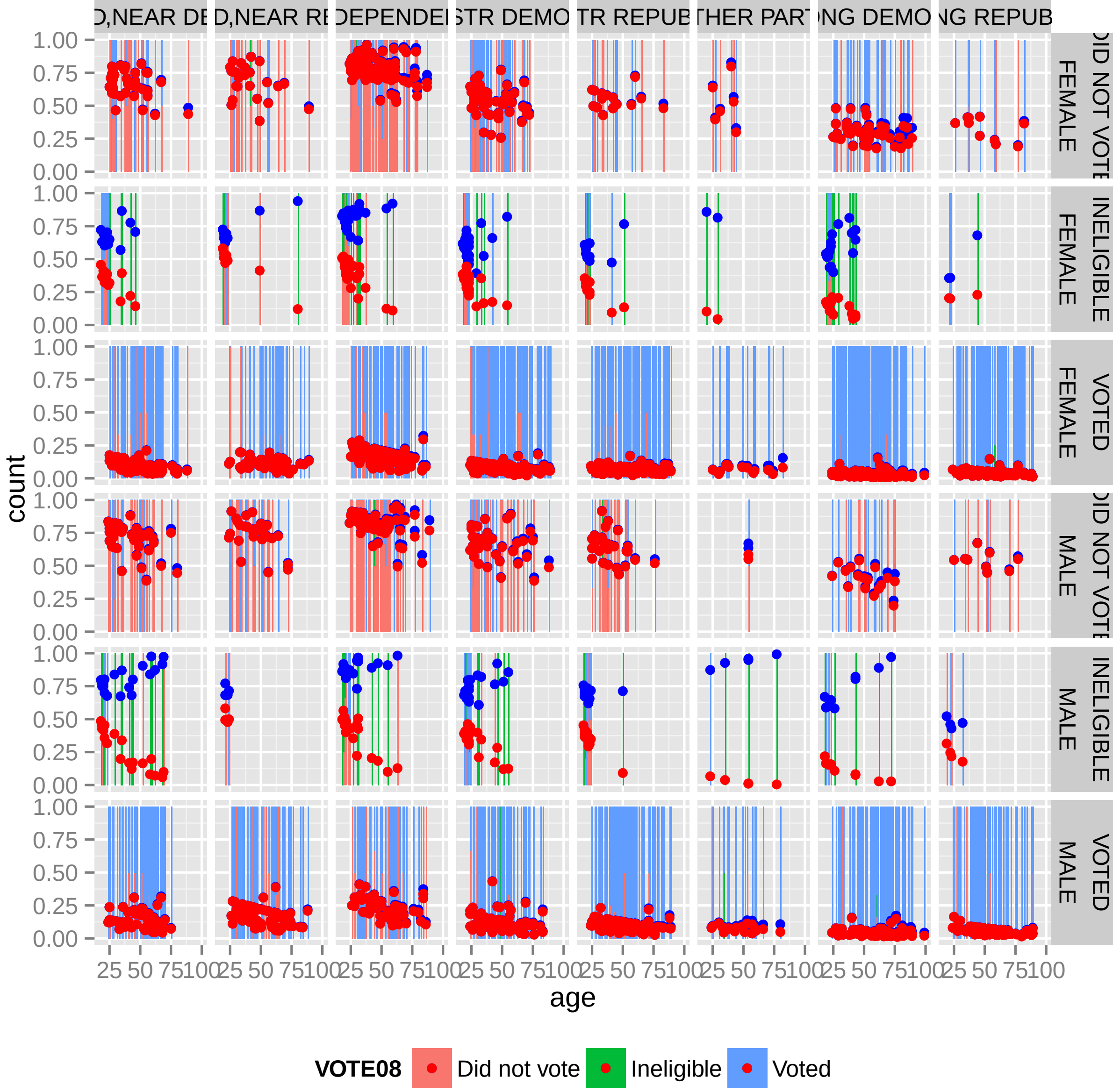
Given one more year of school completed, the logs odds of voting in 2008 relative to *not* voting in 2008 (ceteris paribus) increases by about 0.16.

```r
anova(main, null)

## Likelihood ratio tests of Multinomial Models
##
## Response: VOTE08
##                                  Model Resid. df Resid. Dev   Test    Df
## 1                                    1      3926       2994
## 2 vote04 + sex + partyid + educ + age      3902       1809 1 vs 2    24
##   LR stat. Pr(Chi)
## 1
## 2     1185       0
```

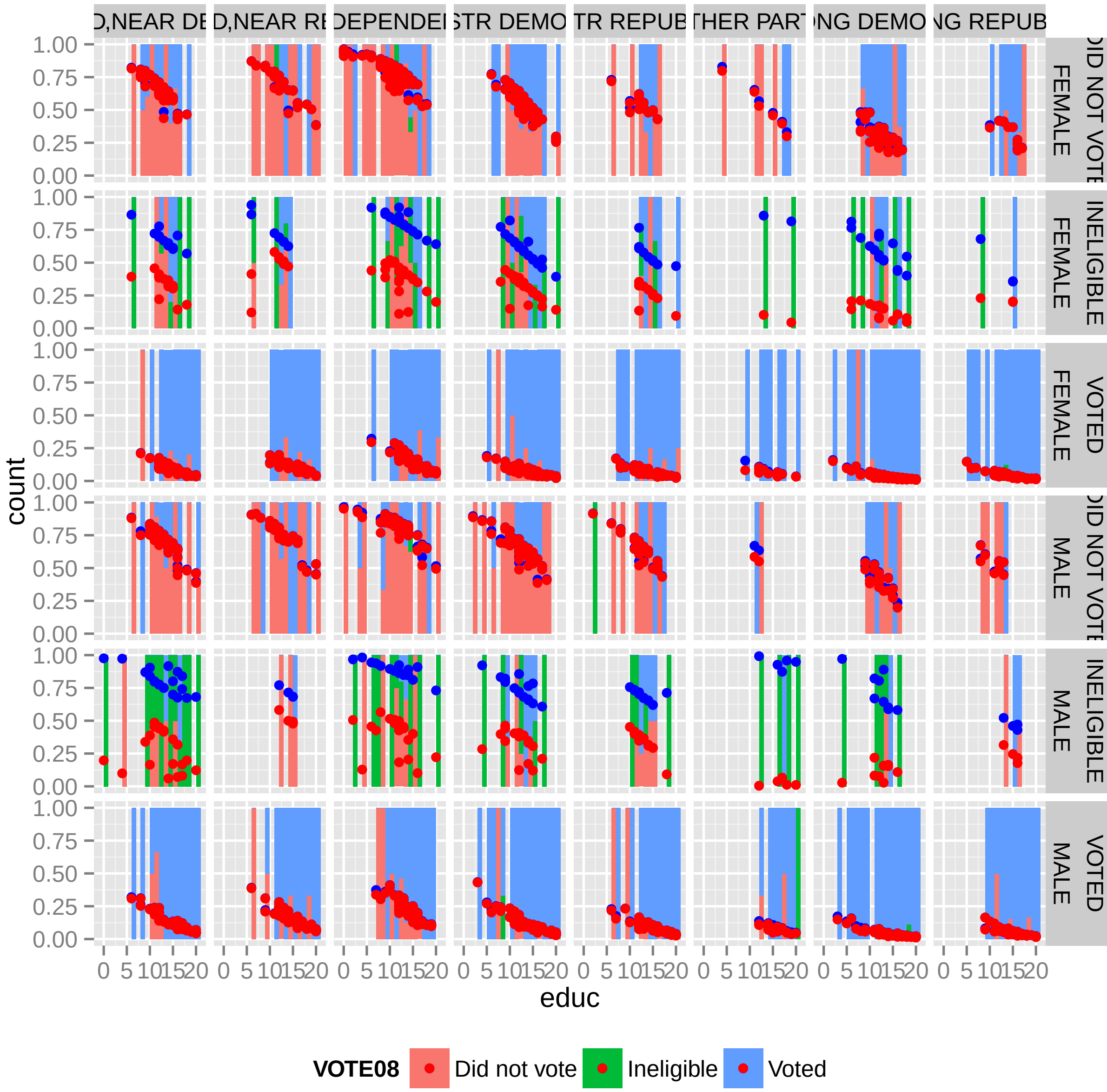
We see a very high reduction in residual deviance (1185.227) going from the null model to the main effects model. Since we are only sacrificing 24 degrees of freedom in this model, the reduction in deviance is highly significant and we have a good overall fit. To further investigate just how well our model fits the data, we'll have a look at the residuals.

```r
pred <- predict(main, newdata = vote3, type = "probs")
play <- cbind(vote3, pred)
names(play)[7] <- "Did_not_vote"
play$Voted <- 1 - play$Voted
qplot(age, geom = "histogram", position = "fill", fill = VOTE08, binwidth = 1,
    data = play) + facet_grid(sex ~ vote04 ~ partyid) + geom_point(aes(x = age,
    y = Voted), color = "blue") + geom_point(aes(x = age, y = Did_not_vote),
    color = "red") + theme(legend.position = "bottom", legend.direction = "horizontal")
```



The "bar" geometry presented in this plot represent the "actual" data. The red and blue dots represent the predicted proportions related to "Did not vote" and "Voted", respectively. The blue dots actually represent the predicted proportion compliment to "Voted" proportion. Thus, a big vertical distance in blue and red represents a high predicted proportion of "Ineligible". As you can see in this plot, the model does a good job of detecting where the greatest proportion of "Ineligible" people are classified. Also, for the cohorts that have a small proportion of ineligible voters, the model does a good job of predicting the proportion that will actually vote (or, equivalently, not vote).

```
qplot(educ, geom = "histogram", position = "fill", fill = VOTE08, binwidth = 1,
    data = play) + facet_grid(sex ~ vote04 ~ partyid) + geom_point(aes(x = educ,
    y = Voted), color = "blue") + geom_point(aes(x = educ, y = Did_not_vote),
    color = "red") + theme(legend.position = "bottom", legend.direction = "horizontal")
```



This is a plot analagous to the previous plot except here we treat "highest year of school completed" as the numerical variable of interest. Again, we can see that the model captures the overall trends and fits the observed data quite well.