Understanding Youth Voter Turnout

Theresa Henle

April 23, 2017

I. Abstract

Previous works have cited the significance of race, gender and education in understanding youth voter turnout. In this paper, we continue to understand the impact of these demographic attributes by considering the relative breakdown of the youth population, and by examining predicted probabilities of voting through these variables. Predicted probabilities for voting are computed by first imputing missing values through the EMB algorithm, and then utilizing a logistic regression. The findings suggest that higher education is the most significant factor in youth who vote, and women on average have higher levels of education. Because youth identifying as white substantially outnumber youth identifying as black, we conclude that white women are the most influential youth voting block.

II. Introduction

American youth historically turnout in lower proportions than any other age group for US presidential elections. In the 2012 presidential election, young people age 18-29 accounted for 22% of the eligible voting population in the United States but only 15% of the votes cast. This discrepancy between voting power and turnout is higher for the youth voting block than is observed in any other age group. The ability to mobilize this large and untapped pool of potential voters can often determine whether a candidate wins or loses an election.

In the following paper, I investigate the importance of race, gender, education and state in prediciting youth voter turnout. I begin with a mosaic plot that visualizes the descriptive statistics of this dataset. I then provide a map of the United States which illustrates the probability of a young person voting by state. I then include a predicted probaility plot to identify the predicted probability of a youth voting based on their race, gender and education level. Lastly, I include the same predicted probabilities in a mosaic plot in order to also visually account for the relative breakdown of the youth population.

The data utilized was collected through a survey of 18 to 24 year olds after the 2012 American presidential election. Participants were asked about their political participation, voting behavior, and political and campaign knowledge. Data was also collected on the respondent's background and demographics, including their experience with civic education in schools, families and community settings. In all, 4,483 participants were surveyed over a period of six weeks. Females, and people identifying as "black" were oversampled, and therefore weights were applied to these observations in each of the analyses below. The data was collected by The Center for Information and Research on Civic Learning and Engagement (CIRCLE) and can be found here.

III. Descriptive Statistics

1. Mosaic Plot

The below figure describes the breakdown of race, education and gender in the United States. Because people identifying as black or females were overrepresented in the study, weights were applied to the data to make this plot an accurate representation of the total youth population. The width of the columns represent the ratio of people in the United States who identify as black versus white, at each level of education. The height

of the rows represent the ratio of people at each education level, within the respective race categories (black or white). Interestingly, the relative breakdown of people who identify as black (or white) is the same across all levels of education. Similarly, the ratio of people at a certain education level is more or less the same regardless of race. The colors of the tiles repesent the percent of females in the study for that race/education combination. The colors are on a yellow-blue scale, with yellow being few females, to blue be predominantly females. It seems that there are fewer females in the lower education categories as compared to males. And that the most highly educated subjects are mostly female. The largest education group, 14 years of education, or equivalently "some college", also appears to be predominantly female.

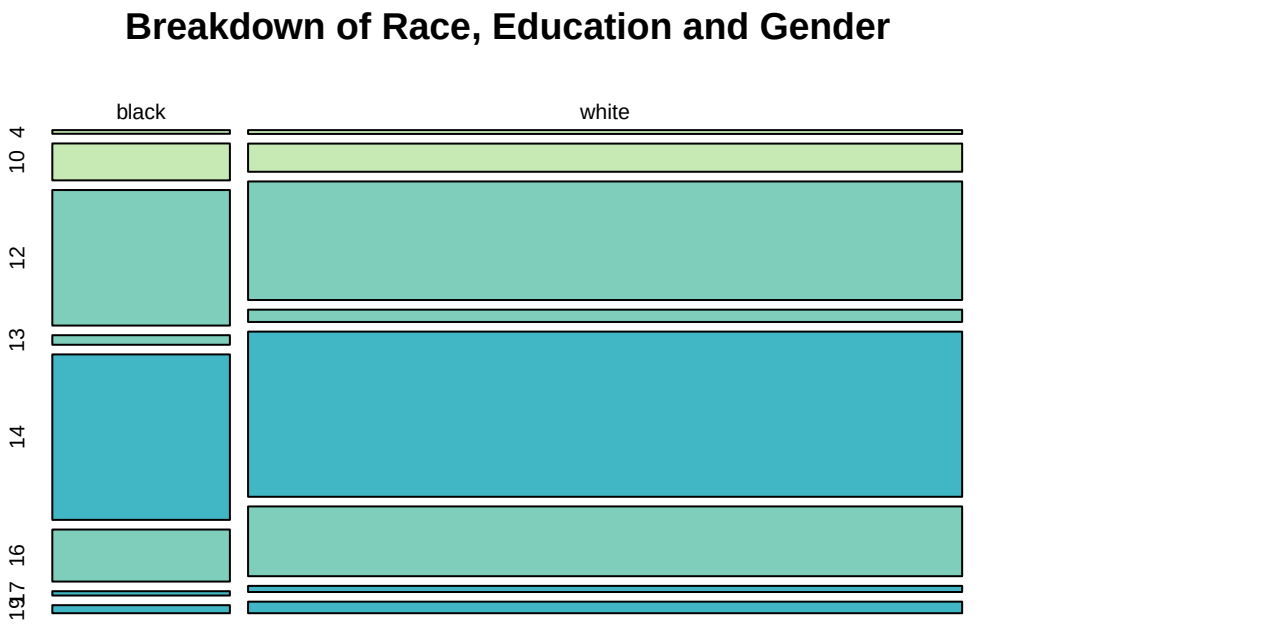
The main take aways from this graph are 1) that youth identifying as white vastly outnumber youth identifying as black, 2) regardless of race, more than half of america's youth have at least some college, and almost all youth have at least graduated from high school 3) men are more likely than women to not graduate high school, and women are more likely than men to pursue college and graduate degrees.

Limitations of this mosaic plot are that the colors cannot be assigned to individual tiles, but only to whole rows. Therefore the colors we are not able to see the difference in percent of females between people who identify as black and people who identify as white, for each level of education. The colors represent a combined percent for both black and white youth.

```
ynew_omit <- na.omit(ynew)
count <- aggregate(weight_final~race +education,
                   data = ynew_omit[c("race", "education", "weight_final")],sum)
castcount <- cast(data = count, formula = race ~ education, fun.aggregate = NULL )
mytable <- as.matrix(castcount[,2:ncol(castcount)])
colnames(mytable) <- names(castcount[,2:ncol(castcount)])
rownames(mytable) <- castcount[[1]]

race <- as.factor(ynew$race)
group_cut <-group_by(ynew_omit, education, race)
Mosaic <- summarise(group_cut, gendermean = mean(gender))

green<-brewer.pal(9,"YlGnBu")
mosaicplot(mytable,col=green[as.numeric(cut(Mosaic$gendermean,breaks=seq(0,1,length=10)))],
           main = "Breakdown of Race, Education and Gender")
```



2. Map of USA

Below is a map of the United States describing youth voter turnout for the 2012 presidential election. The colors utilized are on a red-blue scale, where red represents the high voter turnout and blue represents the low voter turnout. Voter turnout was calculated by taking the weighted number of youth who voted in a given state and dividing it by the weighted total number of youth included in the study for that state. By applying proper weights, we can say this map accurately reflects youth voter turnout by state.

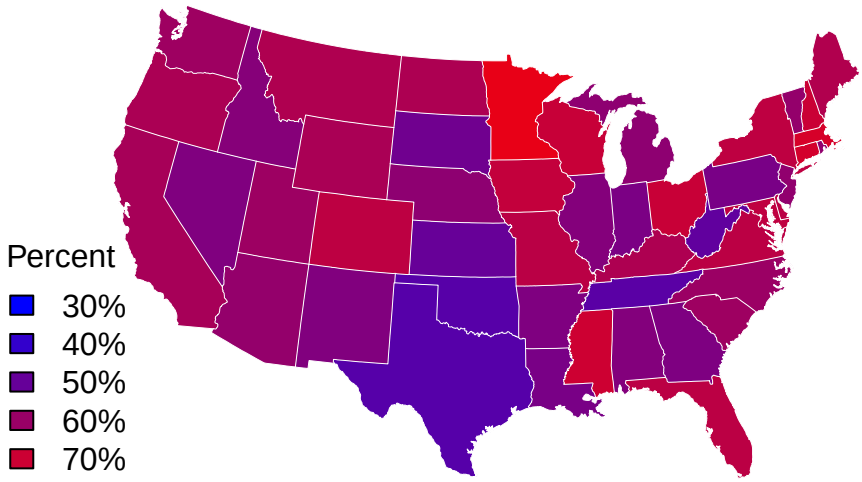
States with particularly low voter turnout 40% and below, include Texas, Oklahoma, Tennessee, and West Virginia. On the other hand, states with high voter turnout include Minnesota, Mississippi, and Massachusetts. Further research is needed to understand why these states have exceptionaly low or high turnout. I would propose using either a regression or classification model to better understand the differences between youth voter turnout by state. One theory I would be interested in exploring is if states with higher voter turnout have significantly more universities per capita than lower voter turnout states, because as we saw in the mosaic plot, most youth have gone to college.

```
ynew$weight2 <- ynew$DidVote*ynew$weight_final
group <- group_by(ynew,State)
one<-summarise(group,num=sum(weight2))
two<-summarise(group,denom=sum(weight_final))
ratio<-one$num/two$denom
rat<-as.data.frame(cbind(state=one[,1],ratio))
rat$State<-tolower(rat$State)

col<-rep(NA, length(map("state", plot=FALSE)$names))
for (i in c(1:7,9:length(map("state", plot=FALSE)$names))){
st<-strsplit(map("state", plot=FALSE)$names[i],":")[[1]][1]
#st<-state.abb[tolower(state.name)==st]
pct<-rat[rat$State==st,"ratio"]
pct<-(pct-.3)/.5
col[i]<-rgb(pct,0,1-pct)
}
```

```
# draw map
map("state", col = col, fill = TRUE, resolution = 0,
lty = 0, projection = "polyconic")
# fill in white lines for states
map("state", col = "white", fill = FALSE, add = TRUE, resolution = 0, lty = 1, lwd = 0.2,
projection="polyconic")
title("Youth Voter Turnout for 2012 Presidential Election")

vec<-seq(0.3,0.7,length=5)
colll<-rgb((vec-.3)/.5,0,1-(vec-.3)/.5)
# Legend
legend("bottomleft",
legend = paste0(c(3:7)*10,"%"),
title = "Percent",
fill = colll,
cex = 1,
bty = "n")
```



IV. Predicting Probabilities of Voting

```
# mulitple imputation
vote.amelia <- amelia(ynew[,-c(1,7)], m=5, noms=c('race','education', 'gender'),
                          emburn=c(500,500))
#### predicted probabilty model
allimplogreg <- lapply(vote.amelia$imputations,
                       function(x){glm(DidVote ~ race + gender + education,
                                       weights = weight_final, family=binomial, data = x)})

mice.betas.glm <- MIextract(allimplogreg, fun=function(x){coef(x)})
mice.se.glm <- MIextract(allimplogreg, fun=function(x){sqrt(diag(vcov(x)))})
mi_inf <- mi.inference(mice.betas.glm, mice.se.glm)
as.data.frame(mi_inf)

# Logistic Regression Model
combmod <- allimplogreg$imp1
is(combmod$coefficients)
# (Intercept)   racewhite     gender   education
combmod$coefficients <- c(-2.9653162, -0.1576565,  0.1548375, 0.2448947)
combmod
```

1. Predicted Probability Plot

This predicted probability plot contains four categories of race and gender combinations: black females, black males, white females and white males. Each category is represented by a line on the graph, however the lines for black males and white females are perfectly situated atop one another. The bands of color around the lines represent 95% confidence bands for each of the lines, meaning we are 95% confident that the true probability of voting falls in that interval at each level of education. The x-axis of the graph represents education in years and the y-axis represents the predicted probability of voting for a person based on a person's combined race, gender and education level. There is a clear positive relationship between education and predicted

probability of voting. The more education a person receives, regardless of their gender or race, the more likely a person is to vote. Given equal education levels between groups, black females are the most likley group to vote for all levels of education. White males are the least likely to vote, and black males and white females turnout equally. The predicted probability of voting is closer at the extremes of education; there is not much difference between groups when education is less than 8 years, or greater than 20 years.

In order to find the predicted probabilities of voting, I used logistic regression to model the probability of voting by race, gender and education. The logistic regression resulted in the following model:

Log(Odds Ratio) = -2.9653162 - 0.1576565*(Race) + 0.1548375*(Gender) + 0.2448947*(education)

Where the Odds Ratio = the probability of voting / the probability of not voting

According to this model, we can say that holding all other variables constant, being white makes a youth 16% less likely to vote than a youth who is black. More so, females are 16% more likely to vote than males. The strongest increase in voting liklihood comes from increasing years of education. For each additional year of schooling, we see a 27% increase in the probability of someone voting.

Missing values were present in both the race and gender variables. To account for this, I used multiple imputation and bootstrapping to impute values where the missing values exist, and then rubin's combining rules to get a logistic regression model. A disussion of these techniques will continue in the following section.

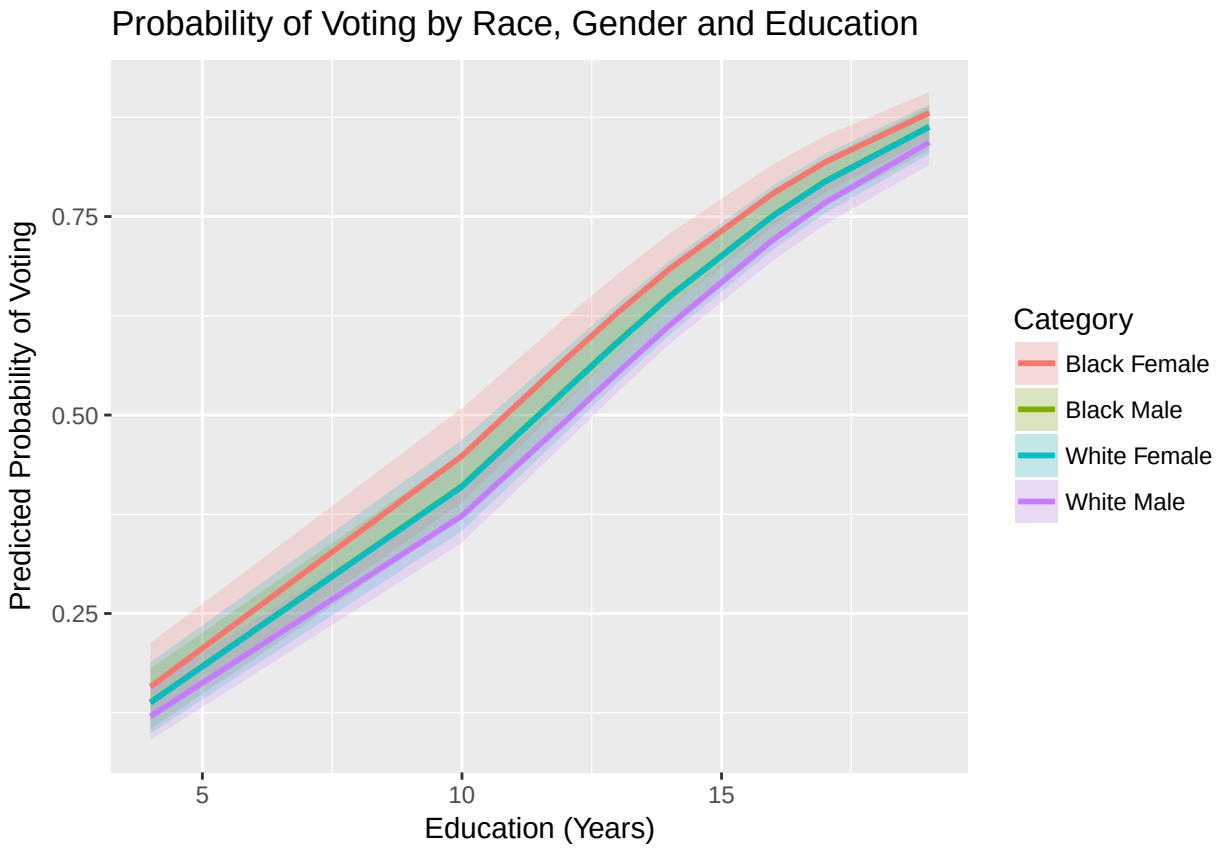
```
# Predicted Probability Plot #
whitemale <- data.frame(gender = rep(1,8), race = rep("white", 8),
                        education = c(4,10,12,13,14,16,17,19))
whitefemale <- data.frame(gender = rep(2,8), race = rep("white", 8),
                          education = c(4,10,12,13,14,16,17,19))
blackmale <- data.frame(gender = rep(1,8), race = rep("black", 8),
                        education = c(4,10,12,13,14,16,17,19))
blackfemale <- data.frame(gender = rep(2,8), race = rep("black", 8),
                          education = c(4,10,12,13,14,16,17,19))
Combinations <- rbind(whitemale, whitefemale, blackmale, blackfemale)

# predictions
Combinations$Predicted <- predict(combmod, newdata = Combinations, type = "response")
Combinations$Category <- "Other"
Combinations$Category[Combinations$gender == 1 & Combinations$race == "white"] <- "White Male"
Combinations$Category[Combinations$gender == 2 & Combinations$race == "white"] <- "White Female"
Combinations$Category[Combinations$gender == 1 & Combinations$race == "black"] <- "Black Male"
Combinations$Category[Combinations$gender == 2 & Combinations$race == "black"] <- "Black Female"
Combinations$Category <- as.factor(Combinations$Category)

newdata3 <- cbind(Combinations, predict(combmod, newdata = Combinations, type="link", se=TRUE))

newdata3 <- within(newdata3, {
  PredictedProb <- plogis(fit)
  LL <- plogis(fit - (1.96 * se.fit))
  UL <- plogis(fit + (1.96 * se.fit))
})

# graphing the prediced values
ggplot(newdata3, aes(x = education, y = PredictedProb)) +
  geom_ribbon(aes(ymin = LL, ymax = UL, fill = Category), alpha = .2) +
  geom_line(aes(colour = Category), size=1) +
  ggtitle("Probability of Voting by Race, Gender and Education") +
  labs(x = "Education (Years)", y= "Predicted Probability of Voting", linetype = "blah")
```



2. Justification for Imputation Procedure

Before selecting a method for imputation, it was important to understand the missingness mechanism at play. Based on the aggr plot below, it is apparent that race is the variable with the highest degree of missingness (25%), where as education is only missing around 5% of the time. In 25% of records, race is missing and no other variables are. Education is missing alone around 4% of the time, and education and race of missingness in combination together about 2% of the time. In about 70% of records we see no missingess of either variable.

The series of matrix plots examine the missing values in one variables in relation to the observed values in other variables. The first matrix plot is sorted by race, and then education; the two variables where missingess exists. The proportion of missingess in education appears to be about the same for each of the race categories, which is not suprising considering what we saw in the aggr plot.

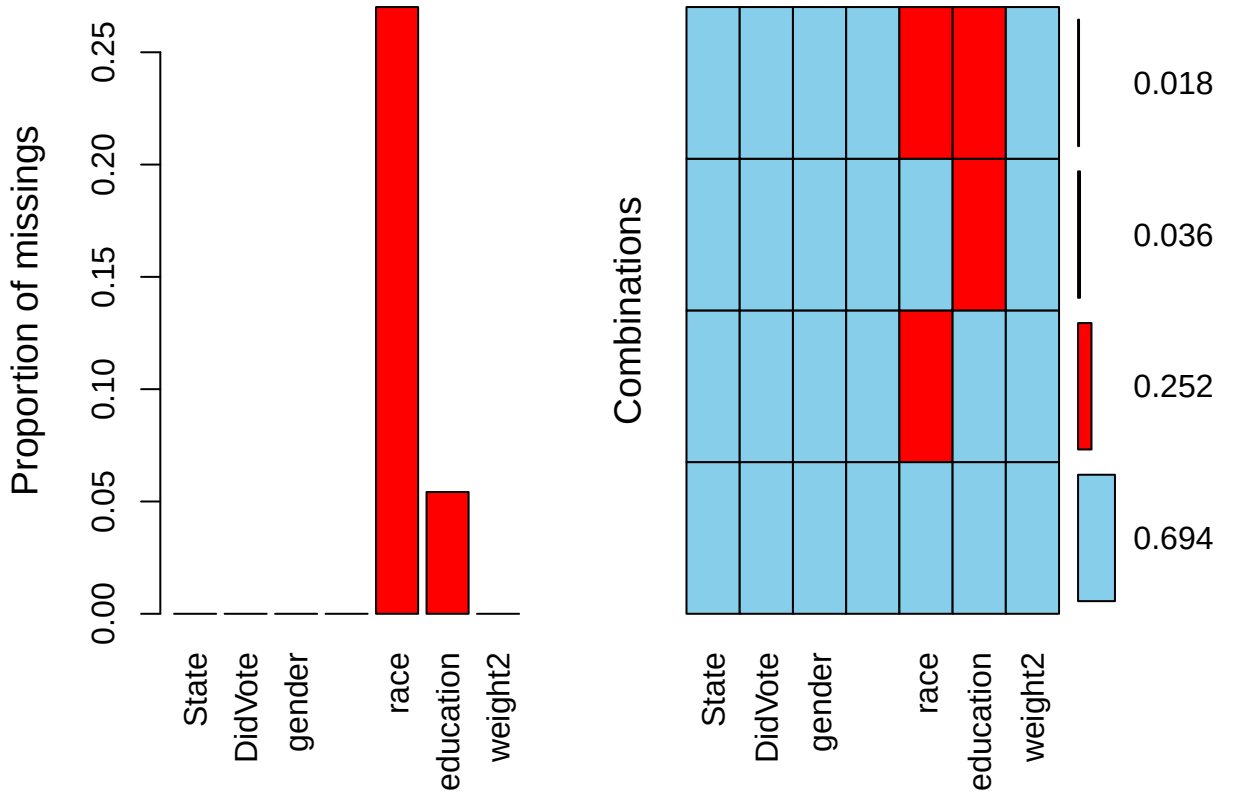
In the next two matrix plots, I look at the missingess in race when the data is sorted by race, and either voting behavior or gender. By comparing the breakdown of voting behavior where race is missing to where race is not missing, we see that these proportions are not the same. Where race is missing, the liklihood of a person voting is about 50%, compared to where race is observed, the liklihood of voting is higher. Because of this result, it is reasonable to argue that missingess in race could be related to a person's voting behavior. However, the gender breakdown when race is missing seems about the same as the gender breakdown when race is not missing, so we cannot argue a relationship here.

The last two matrix plots investigate the missingness in education, when the data is sorted by education along with either voting behavior or gender. The proportion of missingness in education is quite small, so comparisons to observd variables are somewhat difficult. The proportion of people voting where education is omitted is about 50/50, which appears to be about the same proportion for voting behavior where education is observed. When sorting by education and gender, the gender breakdowns also look to be about equal between where missingess does and does not exist, however it is difficult to tell from the plot.

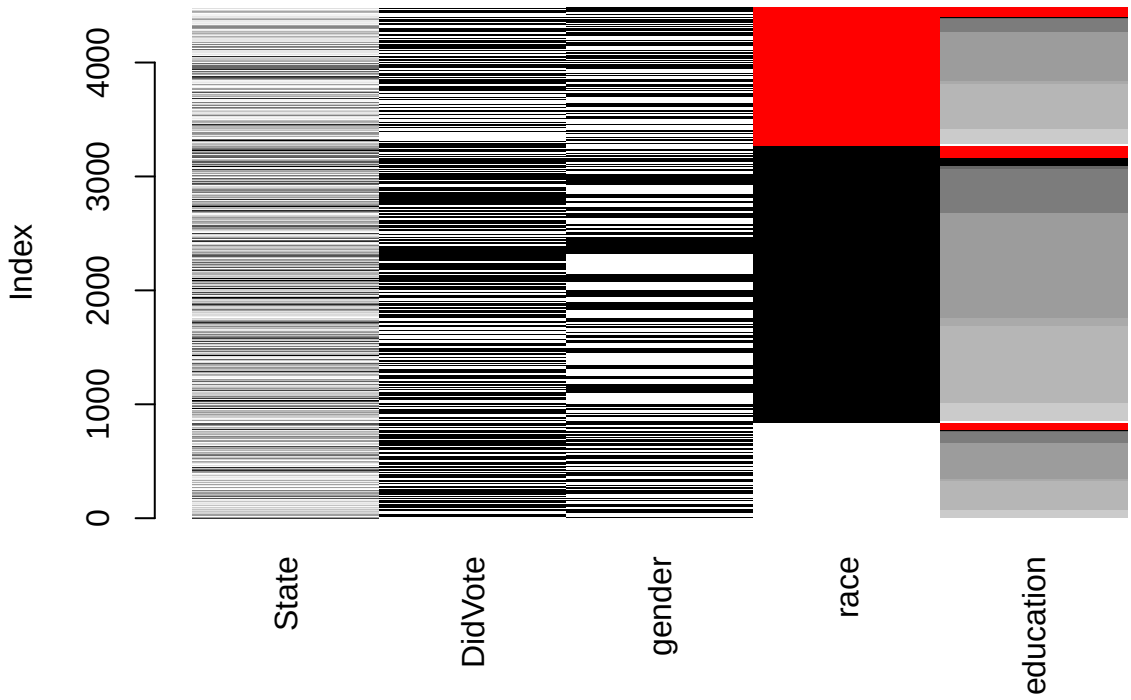
These results lead us to believe that the data is not MCAR(missing completely at random). MCAR occurs when all values are equally likley to be missing. As we saw, there appears to be a relationship beteen missingess in race and voting behavior, so the MCAR assumption does not hold. Though it is impossible to say for sure, our best guess is that the data is MAR(missing at random), meaning the missing values have to do with other observed values in the dataset. MNAR occurs when the missing values are missing because of the true value they carry, which is impossible to identify without a deeper understanding of the way the data was collected. We will assume MAR and utilize imputation techniques accordingly.

The missing values were imputed through the Amelia package. The Amelia package runs the EMB(expectation-maximization with bootstrapping) algorithm on incomplete data in order to create multiple imputed datasets. The procedure works by first generating multiple bootstrapped samples of the original incomplete data and then utilizes the EM algorithm to "draw values from the complete-data parameters". The EM algorithm is an iterative procedure for finding the maximum likelihood estimation of parameters, given the observed data, in the presence of missing values. Then, values are imputed for missing values by drawing from each of the bootstrapped parameters. The Amelia package is a flexible imputation method that works well for categorical data imputation. Therefore it was a good choice as an imputation procedure for this data.

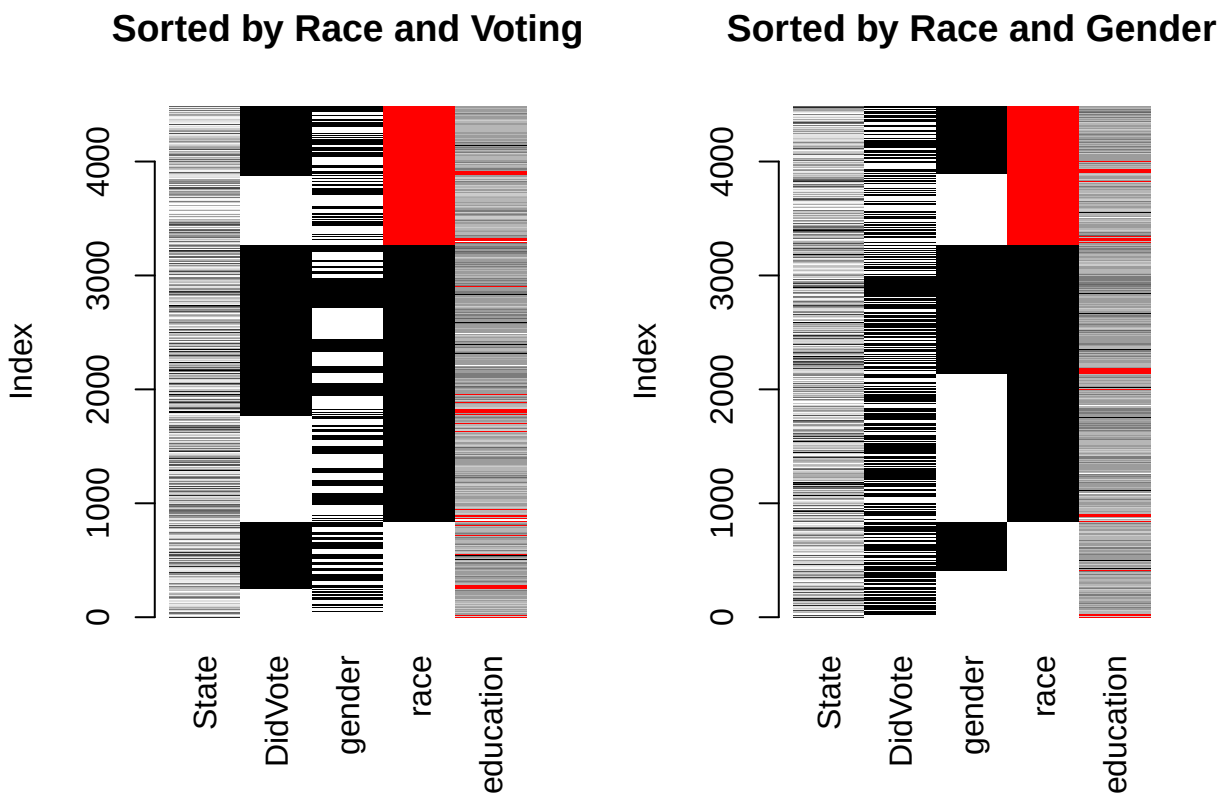
```
#Investigation of Missingness Mechanism#
ynew$race <- as.factor(ynew$race)
ynew$State <- as.factor(ynew$State)
aggr(ynew, numbers = TRUE)
```



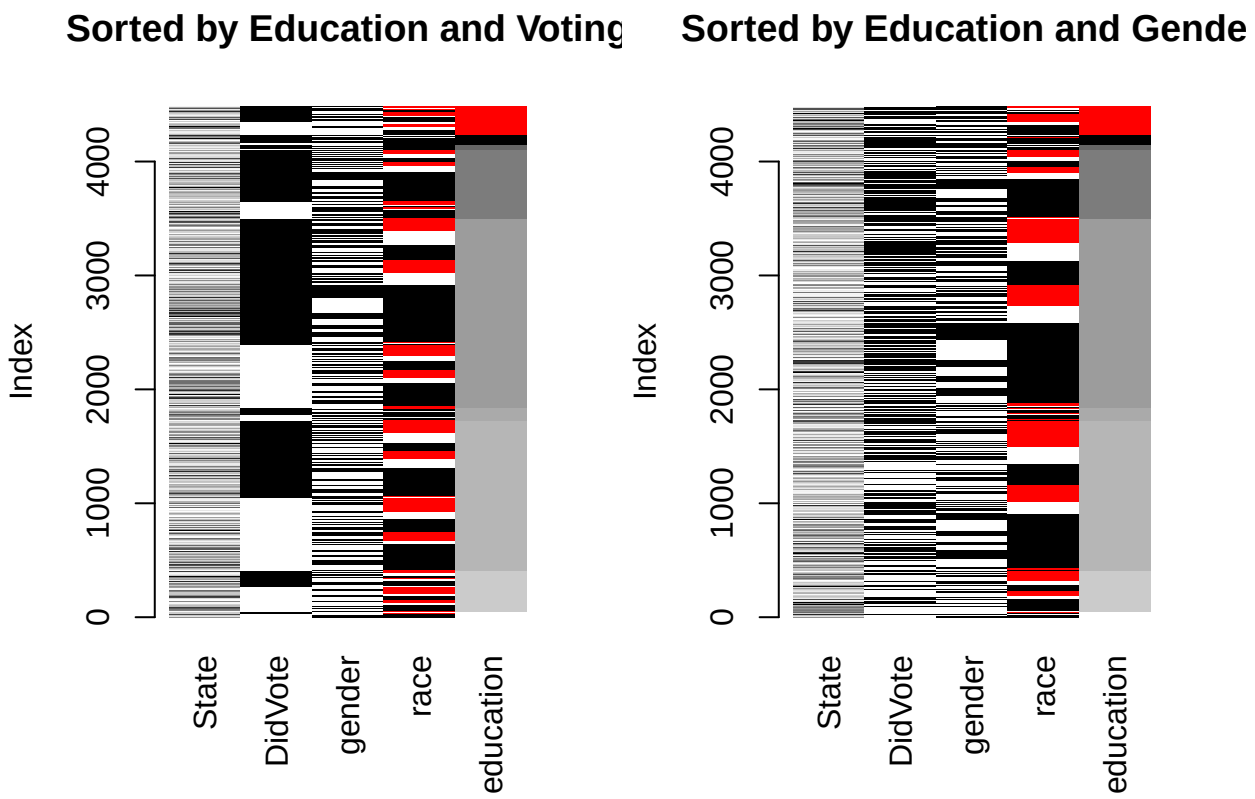

```
# 1: sort by race and education
newdata <- ynew[order(ynew$race, ynew$education),]
matrixplot(newdata[,-c(4,7)], sortby = "race", main = "Sorted by Missing: Race and Education")
```




```
# 1: sort by race and voting behavior
par(mfrow = c(1,2))
newdata <- ynew[order(ynew$race, ynew$DidVote),]
matrixplot(newdata[,-c(4,7)], sortby = "race", main = "Sorted by Race and Voting")
# 2: sort by race and gender
newdata <- ynew[order(ynew$race, ynew$gender),]
matrixplot(newdata[,-c(4,7)], sortby = "race", main = "Sorted by Race and Gender")
```




```
# 3: sort by education and voting behavior
par(mfrow = c(1,2))
newdata <- ynew[order(ynew$education, ynew$DidVote),]
matrixplot(newdata[,-c(4,7)], sortby = "education", main = "Sorted by Education and Voting")
# 4: sort by education and gender
newdata <- ynew[order(ynew$education, ynew$gender),]
matrixplot(ynew[,-c(4,7)], sortby = "education", main = "Sorted by Education and Gender")
```



3. Mosaic Plot

Lastly, by combining predicted probabilities and population breakdown in one visualization, we can paint an interesting picture for the importance of voting behavior between different youth demographic groups. This mosaic plot contains a person's race and gender on the x-axis, and a person's years of education on the y-axis. The color, on a yellow-blue scale, displays the predicted probability of a person voting in that education category voting. Because in mosaic plots colors cannot be assigned to individual tiles, the colors represent a combined percent for both black and white youth. Predicted probabilities are based on on the same logistic regression model used above. This mosaic plot displays similar information to the predicted probability plot shown above, but adds an element of population breakdown. Though we found that black females are the most likely race/gender combination to vote, there are many more white females in the population, so one could argue that their relative impact is likely higher. For all race/gender combinatoins, higher levels of education appear to be the most influential factor in a person voing.

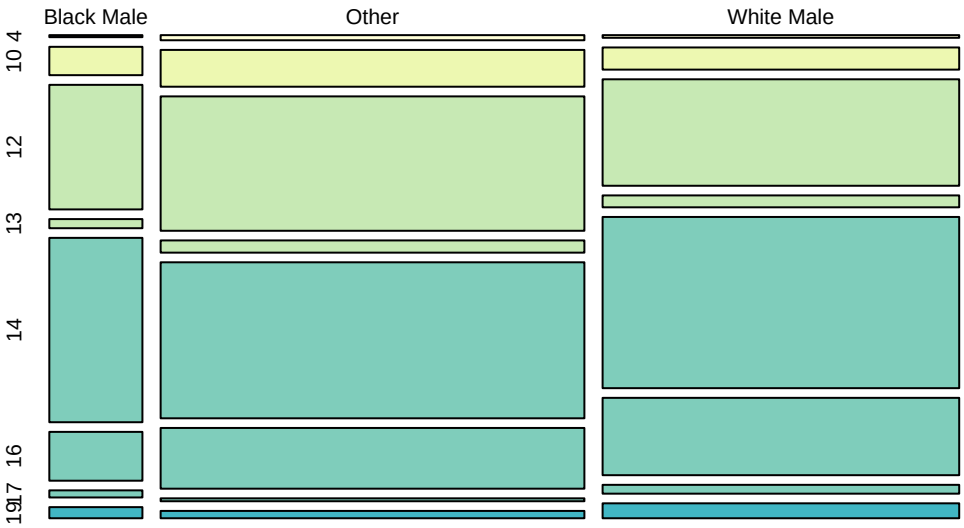
```
ynew_omit <- na.omit(ynew)
ynew_omit$Category <- "Other"
ynew_omit$Category[ynew_omit$gender == 1 & ynew_omit$race == "white"] <- "White Male"
ynew_omit$Category[ynew_omit$gender == 2 & ynew_omit$race == "white"] <- "White Female"
ynew_omit$Category[ynew_omit$gender == 1 & ynew_omit$race == "black"] <- "Black Male"
ynew_omit$Category[ynew_omit$gender == 2 & ynew_omit$race == "black"] <- "Black Female"

count <- aggregate(weight_final~ Category +education,
                   data = ynew_omit[c("Category", "education", "weight_final")], sum)
castcount <- cast(data = count, formula = Category ~ education, fun.aggregate = NULL )
```

```
## Using weight_final as value column.  Use the value argument to cast to override this choice
```

```
mytable <- as.matrix(castcount[,2:ncol(castcount)])
colnames(mytable) <- names(castcount[,2:ncol(castcount)])
rownames(mytable) <- castcount[[1]]
```

```
# Mosaic Plot
green<-brewer.pal(9,"YlGnBu")
mosaicplot(mytable,main = "Predicted Probabilities of Voting by Race and Education",
           col=green[as.numeric(cut(Combinations$Predicted,breaks=seq(0,1,length=6)))])
```



V. Conclusion

This paper investigates the relationship between race, gender and education in predicting and understanding youth voting behavior. From our initial mosiac, we were able to see the population breakdown of race, education and gender. It was identified that most youth in this country have at least some college or technical schooling, with the average woman pursing higher levels of education than men. It is also important to note that youth identifying as white greatly outnumber youth identifying as black. Through the USA map, we were then able to identify states with higher or lower youth voter turnout. Using multiple imputation and logistic regression techniques, we were then able to predict youth voter turnout using our three demographic attributes. These predictions were displayed in the predicted probability plot and in the mosaic plot. From these plots, we found that while black females have the highest predicted probability of voting, white youth greatly outnumber black youth, and therefore white females probably carry greater relative voting power. However, the most significant trend found in these plots is the more education a youth receives, the more likely they are to vote.

Overall, the findings suggest that higher education is the most significant factor in youth who vote, and women on average have higher levels of education. Because youth identifying as white substantially outnumber youth identifying as black, we conclude that white women are the most influential youth voting block.

In further research I am interested in building a classification or regression model with more variables to get a broader understanding of what factors impact a youth's voting behavior. Specifically I am interested in integrating aspects of voting preferences, such as which candidates a person supported or issues they prioritized. Understanding and increasing youth voter participation has the potential to dramatically shift the

way politics are played in the United States. Not only can the youth vote have a significant effect on election results, youth who practice civic engagment are more likely to continue to be active citizens throughout their lives.

VI. Citations

2013. CIRCLE .What do Young Adults Know about Politics? Evidence from a National Survey Conducted After the 2012 Election.CIRCLE Fact Sheet. Medford, MA: Tufts University, Center for Information & Research on Civic Learning & Engagement.

2014. Levine, Peter, CIRCLE .Young People and Immigration Reform: Evidence from a Post Election Survey.CIRCLE Fact Sheet. Medford, MA: Center for Information & Research on Civic Learning & Engagement.

2015. Levine, Peter, Kawashima-Ginsberg, Kei .Teaching civics in a time of partisan polarization.Social Education. 77, (4), 215-217.