Mapping the 2016 Election

February 3, 2017

We will analyze returns from the past two electoral cycles to try to understand the social and demographic trends that may have contributed to Donald Trump's victory. We will first examine how Republican vote share at the county level has changed from 2012 to 2016. Then, we will look at four variables that were prominent in the discourse around the election – race, education, unemployment, and immigration – to see how well they predict GOP electoral gains at the county level.

We will be working with three datasets. The first, `electoral returns 2012.csv`, has one observation per county and contains the following variables:[1]

Name	Description
`FIPS`	FIPS code (unique county identifier)
`state`	State abbreviation
`county`	County name
`votes_dem`	Number of votes cast for Democratic candidate, 2012 election
`votes_gop`	Number of votes cast for Republican candidate, 2012 election
`votes_total`	Total number of votes cast in 2012 election

The second, `electoral returns 2016.csv`, has the same structure but reports data for the 2016 presidential election.

The third dataset, `county level data.csv`, includes social and demographic characteristics for each county:[2]

Name	Description
`FIPS`	FIPS code (unique county identifier)
`pct_for_born15`	Percent of county's population that is "foreign born" according to the U.S. Census, meaning anyone who is not a U.S. citizen at birth (measured over 2011-2015)
`pct_bach_deg15`	Percent of county population holding a Bachelor's degree or above (2011-2015)
`pct_non_white15`	Percent of county population that is not white (2011-2015)
`pct_unemp16`	Percent of county population that is unemployed, BLS estimates (average, Jan-Oct 2016)
`pct_unemp12`	Percent of county population that is unemployed, BLS estimates (average, Jan-Oct 2012)

Load and merge data

First we'll load the data and merge the three datasets by FIPS code to construct one complete data file for analysis.

```
# read in data
returns12 <- read.csv("data/electoral returns 2012.csv")
returns16 <- read.csv("data/electoral returns 2016.csv")
```

[1] 2012 and 2016 electoral returns come from Tony McGovern (https://github.com/tonmcg/County_Level_Election_Results_12-16).

[2] Assembled from BLS Local Area Unemployment Statistics (https://download.bls.gov/pub/time.series/la/) and PolicyMap (https://www.policymap.com/) with generous assistance from Bernie Langer.

```r
covars <- read.csv("data/county level data.csv")

# merge all datasets by FIPS code first, let's change the names that overlap
# so we keep them straight
names(returns12) <- c("state", "county", "FIPS", "votes_dem_12", "votes_gop_12",
    "votes_total_12")
names(returns16) <- c("state", "county", "FIPS", "votes_dem_16", "votes_gop_16",
    "votes_total_16")

# now, lets merge
returns <- merge(returns12, returns16[, !names(returns16) %in% c("state", "county")],
    by = "FIPS")
## we don't need state and county in both datasets

# did we lose any observations?
dim(returns12)
```

```
## [1] 3141    6
```

```r
dim(returns16)
```

```
## [1] 3141    6
```

```r
dim(returns)
```

```
## [1] 3141    9
```

```r
## nope

# now let's merge on covariates
merged <- merge(returns, covars, by = "FIPS", all.x = TRUE)
## want to keep all observations that are in returns

# remove missing values (listwise deletion)
final <- na.omit(merged)

# how much data did we lose?
dim(merged)
```

```
## [1] 3141   14
```

```r
dim(final)
```

```
## [1] 3112   14
```

```r
## not too bad

# identify lost data
lost <- merged[merged$FIPS %in% final$FIPS == FALSE, ]
head(lost)
```

```
##    FIPS state county votes_dem_12 votes_gop_12 votes_total_12 votes_dem_16
## 68 2013    AK Alaska           NA           NA             NA        93003
## 69 2016    AK Alaska           NA           NA             NA        93003
## 70 2020    AK Alaska           NA           NA             NA        93003
## 71 2050    AK Alaska           NA           NA             NA        93003
## 72 2060    AK Alaska           NA           NA             NA        93003
## 73 2068    AK Alaska           NA           NA             NA        93003
```

```
##    votes_gop_16 votes_total_16 pct_for_born15 pct_bach_deg15
## 68       130413         246588          45.13          14.05
## 69       130413         246588          35.40          14.74
## 70       130413         246588           9.96          33.20
## 71       130413         246588           1.85          11.65
## 72       130413         246588           1.96          20.03
## 73       130413         246588           6.75          33.33
##    pct_non_white15 pct_unemp12 pct_unemp16
## 68           78.45         6.7         2.7
## 69           68.44         9.3         3.7
## 70           35.06         4.8         5.1
## 71           87.98        14.3        13.2
## 72           43.61        13.8         9.5
## 73           12.67        14.4        10.5
```

```r
unique(lost$county)
```

```
## [1] Alaska
## 1848 Levels: Abbeville County Acadia Parish Accomack County ... Ziebach County
```

Looks like we lost Alaska, for which we didn't have 2012 returns. Otherwise, a successful merge.

Difference in Republican vote share from 2012 to 2016: overall and in battleground states

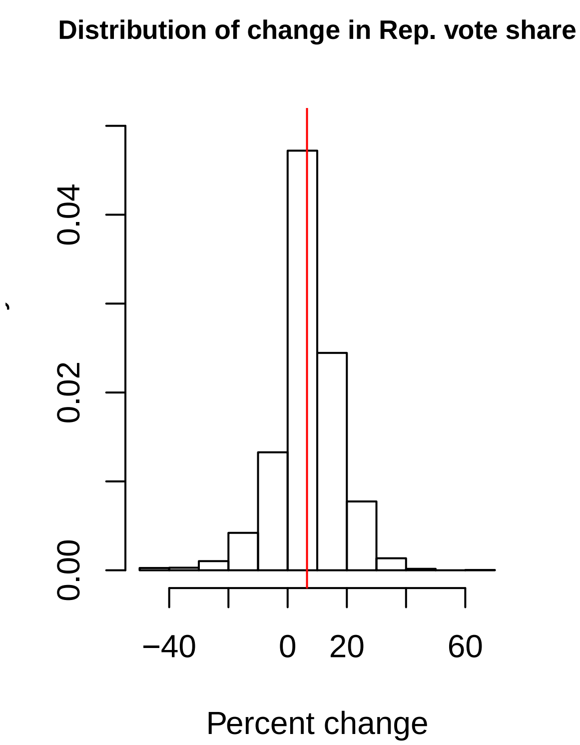
We'll now compute the Republican vote share as a proportion of total votes, in 2012 as well as in 2016. We'll also compute the percent difference in this Republican vote share variable from the 2012 to 2016 election, and plot its distribution with a red line at the median. We'll repeat the same analysis on the subset of battleground states: Florida, North Carolina, Ohio, Pennsylvania, New Hampshire, Michigan, Wisconsin, Iowa, Nevada, Colorado, and Virginia.

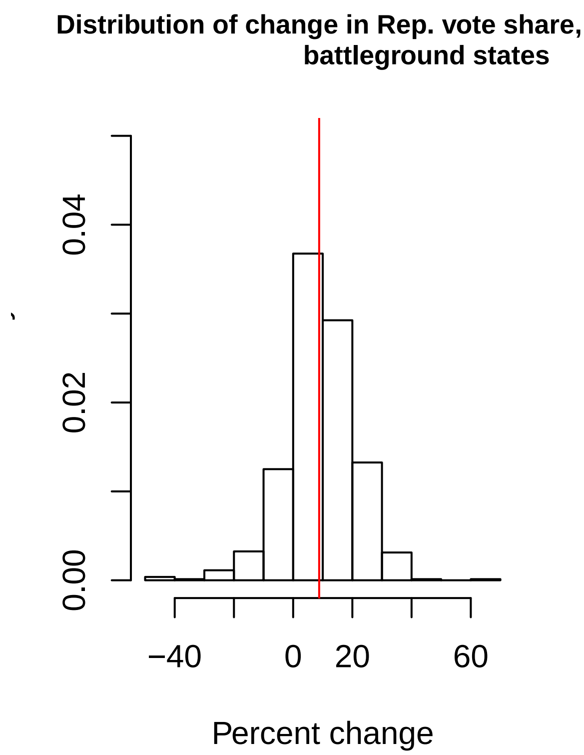
```r
# vote share variables
final$gop_vs_12 <- final$votes_gop_12 / final$votes_total_12
final$gop_vs_16 <- final$votes_gop_16 / final$votes_total_16

# percent change from 2012 to 2016
final$gop_vs_pct_ch <- ((final$gop_vs_16 - final$gop_vs_12) / final$gop_vs_12) * 100

# subset data to battleground states
battlestates <- c("FL", "NC", "OH", "PA", "NH", "MI", "WI", "IA", "NV", "CO", "VA")
battle <- subset(final, state %in% battlestates)
table(battle$state) ## check that you subset properly
```

```
## 
##  AK  AL  AR  AZ  CA  CO  CT  DC  DE  FL  GA  HI  IA  ID  IL  IN  KS  KY 
##   0   0   0   0   0  64   0   0   0  67   0   0  99   0   0   0   0   0 
##  LA  MA  MD  ME  MI  MN  MO  MS  MT  NC  ND  NE  NH  NJ  NM  NV  NY  OH 
##   0   0   0   0  83   0   0   0   0 100   0   0  10   0   0  17   0  88 
##  OK  OR  PA  RI  SC  SD  TN  TX  UT  VA  VT  WA  WI  WV  WY 
##   0   0  67   0   0   0   0   0   0 133   0   0  72   0   0
```

```r
# plot the distributions
par(mfrow = c(1, 2), pin = c(2, 2.5)) ## put plots side by side

hist.full <- hist(final$gop_vs_pct_ch,
```

```r
                   freq = FALSE,
                   main = "Distribution of change in Rep. vote share", cex.main = .8,
                   xlab = "Percent change",
                   ylim = c(0, .05))
abline(v = median(final$gop_vs_pct_ch), col = "red")

hist.battle <- hist(battle$gop_vs_pct_ch,
                     freq = FALSE,
                     main = "Distribution of change in Rep. vote share,
                             battleground states", cex.main = .8,
                     xlab = "Percent change",
                     ylim = c(0, .05))
abline(v = median(battle$gop_vs_pct_ch), col = "red")
```





Republicans made electoral gains in this election over the last in more counties across the nation than Democrats, and this was even (slightly) more true in the sample of swing states.

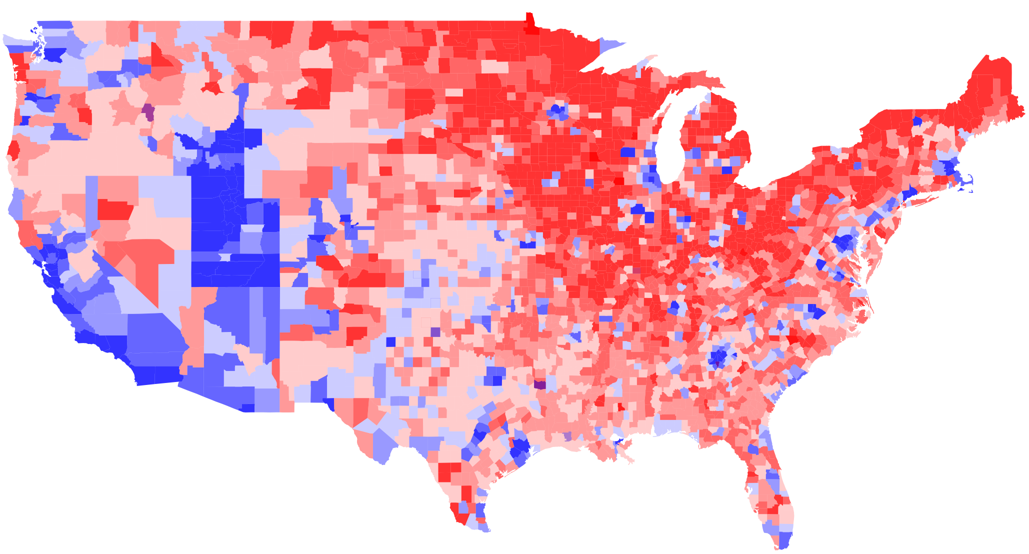
Mapping GOP electoral gains, 2012-2016

We'll now create a county-level map of the United States, with counties where Democrats got a larger vote share in 2016 than 2012 in blue, and counties where the Republican vote share increased in red. We also want the intensity of the color to depend on the magnitude of the Democratic or Republican gains.

```r
# load libraries
library(maps)
library(ggmap)
```

```
## Loading required package: ggplot2
```

```r
# merge data onto county.fips
cf <- county.fips
names(cf) <- c("FIPS", "name")

toplot <- merge(cf, final, by = "FIPS", all.x = TRUE)

# function to compute alpha
alpha.compute <- function(invar) {
  var <- abs(invar) ## absolute value
  quants <- quantile(var, na.rm = TRUE)
  alpha <- ifelse(var < quants[2], .2,
             ifelse(var >= quants[2] & var < quants[3], .4,
               ifelse(var >= quants[3] & var < quants[4], .6, .8)))
  alpha <- ifelse(is.na(alpha), 0, alpha)
  return(alpha)
}

# compute alpha
alpha1 <- alpha.compute(invar = toplot$gop_vs_pct_ch)

# create vector of colors
toplot$cols1 <- ifelse(toplot$gop_vs_pct_ch < 0,
                       rgb(red = 0, blue = 1, green = 0, alpha = alpha1),
                       rgb(red = 1, blue = 0, green = 0, alpha = alpha1))



# create map
map(database = "county", lty = 0) # activate empty map
for(i in 1:nrow(toplot)) {
  map(database = "county", regions = toplot$name[i], col = toplot$cols1[i],
      fill = TRUE, add = TRUE, lty = 0)
}
```



The largest Republican gains occurred in the Midwest. There were actually quite a few counties where the Democratic party made gains since 2012, but these counties were predominantly the larger, less populous counties of the West that did not matter much for the Electoral College. Texas is a particularly interesting state, as it contains counties with significant Republican gains alongside those with significant Democratic gains, suggesting possible geographical sorting. By contrast, the most static regions appear to be the middle of the country and the Northwest.

Modeling the effects of immigration, education, race, and unemployment on the 2016 election

```
lm1 <- lm(gop_vs_pct_ch ~ pct_for_born15 + pct_bach_deg15 + pct_non_white15 + pct_unemp16,
          data = final)
summary(lm1)
```

```
##
## Call:
## lm(formula = gop_vs_pct_ch ~ pct_for_born15 + pct_bach_deg15 +
##     pct_non_white15 + pct_unemp16, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.841  -4.739  -1.164   3.965  58.360
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     22.703192   0.666417  34.068   <2e-16 ***
## pct_for_born15  -0.506000   0.027763 -18.226   <2e-16 ***
## pct_bach_deg15  -0.577759   0.018451 -31.314   <2e-16 ***
## pct_non_white15 -0.089935   0.009726  -9.247   <2e-16 ***
## pct_unemp16     -0.098992   0.096866  -1.022    0.307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.841 on 3107 degrees of freedom
## Multiple R-squared:  0.4457, Adjusted R-squared:  0.4449
## F-statistic: 624.4 on 4 and 3107 DF,  p-value: < 2.2e-16
```

Percent foreign-born, percent with a Bachelor's degree or above, and percent non-white are all statistically significant predictors of Republican losses since 2012. Somewhat surprisingly, unemployment is not a strong predictor.

Predicting the 2016 election

We will now see which counties had the most surprising election results in 2016 given our predictions based on the previous election. To do so, we'll first regress 2012 Republican vote share on percent foreign-born, percent with a Bachelor's degree or above, percent non-white, and percent unemployed in 2012. Then, we'll predict 2016 Republican vote share in each county using these same variables in 2016.[3]

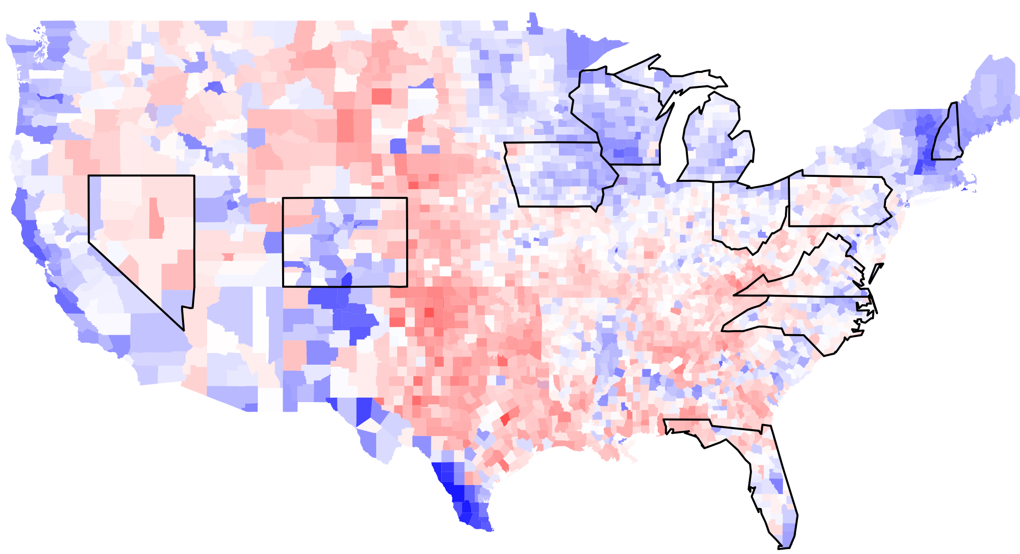
```
# run 2012 model
lm2 <- lm(gop_vs_12 ~ pct_for_born15 + pct_bach_deg15 + pct_non_white15 + pct_unemp12,
          data = final)
```

[3] Only unemployment actually has updated data; the most recent available Census estimates span 2011-2015, so we'll have to reuse them.

```r
# compute 2016 predictions
pred.df <- toplot[, c("pct_for_born15", "pct_bach_deg15", "pct_non_white15", "pct_unemp16")]
names(pred.df) <- c("pct_for_born15", "pct_bach_deg15", "pct_non_white15", "pct_unemp12")
toplot$preds <- predict(lm2, pred.df)
```

We can now create a county-level map of the prediction error (predicted Republican vote share subtracted from the observed value in 2016), with counties colored in red where the observed value was higher than the prediction and blue otherwise. We'll use double the absolute value of the prediction error as the intensity of the color.

```r
toplot$pred.error <- toplot$gop_vs_16 - toplot$preds
## higher positive values where Trump overperformed
toplot <- na.omit(toplot)

# create vector of colors
toplot$pred.cols <- ifelse(toplot$pred.error > 0,
                           rgb(red = 1, blue = 0, green = 0, alpha = abs(2 * toplot$pred.error)),
                           rgb(red = 0, blue = 1, green = 0, alpha = abs(2 * toplot$pred.error)))

# make map
map(database = "county", lty = 0) # activate empty map
for(i in 1:nrow(toplot)) {
  map(database = "county", regions = toplot$name[i], col = toplot$pred.cols[i],
      fill = TRUE, add = TRUE, lty = 0)
}

# let's put boundaries just around the swing states
swings <- c("florida", "north carolina", "ohio", "pennsylvania", "new hampshire",
            "michigan", "wisconsin", "iowa", "nevada", "colorado", "virginia")
map("state", boundary = TRUE, regions = swings, add = TRUE)
```

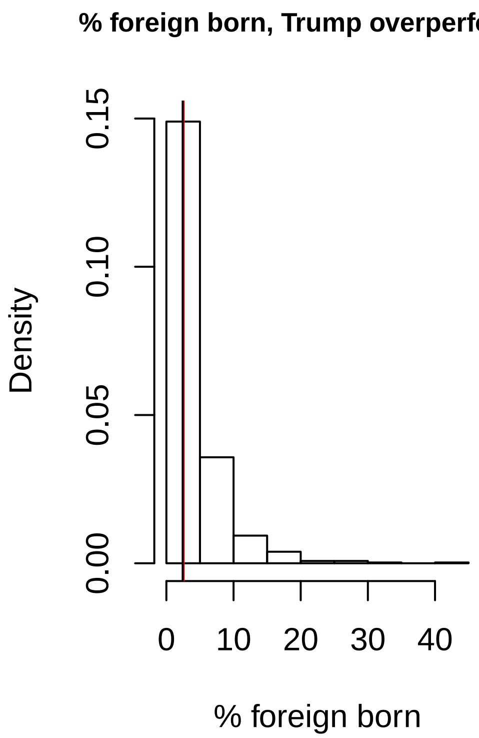


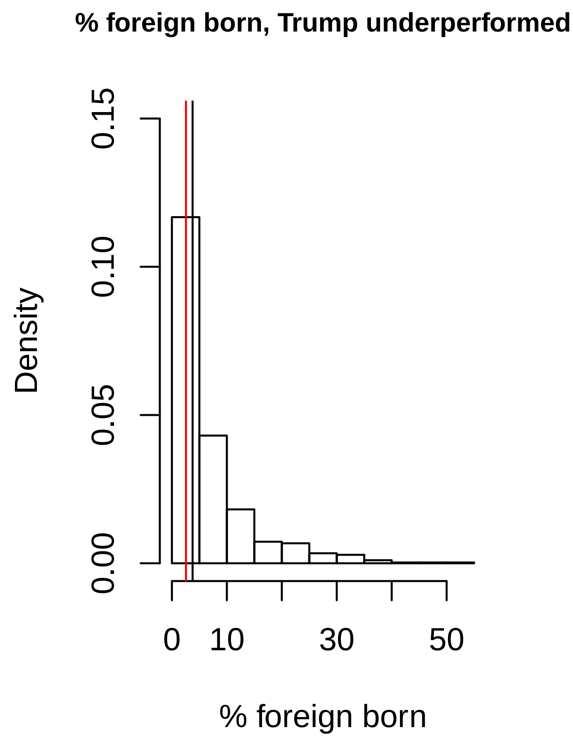
We can interpret heavily colored areas as places where new dynamics were introduced in the 2016 election: for instance, note the intensely blue areas along the Texas border, where immigration issues might have played a fundamentally different role in 2016 than 2012. Trump generally overperformed in Texas, Nevada, and along a band in the middle of the country, but there is also a surprising amount of blue on the map, even in swing states. Part of the story might be population dynamics, not represented here: if Trump overperformed in more population-dense counties and underperformed in rural areas — consistent with what we know about the relatively low urban/minority turnout in this election — then we can reconcile this map with a Trump victory. Of course, underperformance doesn't imply that there wasn't a Republican majority in the county; we can have a fairly blue map that is still consistent with an overall Trump victory, and the Electoral College amplifies this possibility.

What are the characteristics of surprise counties?

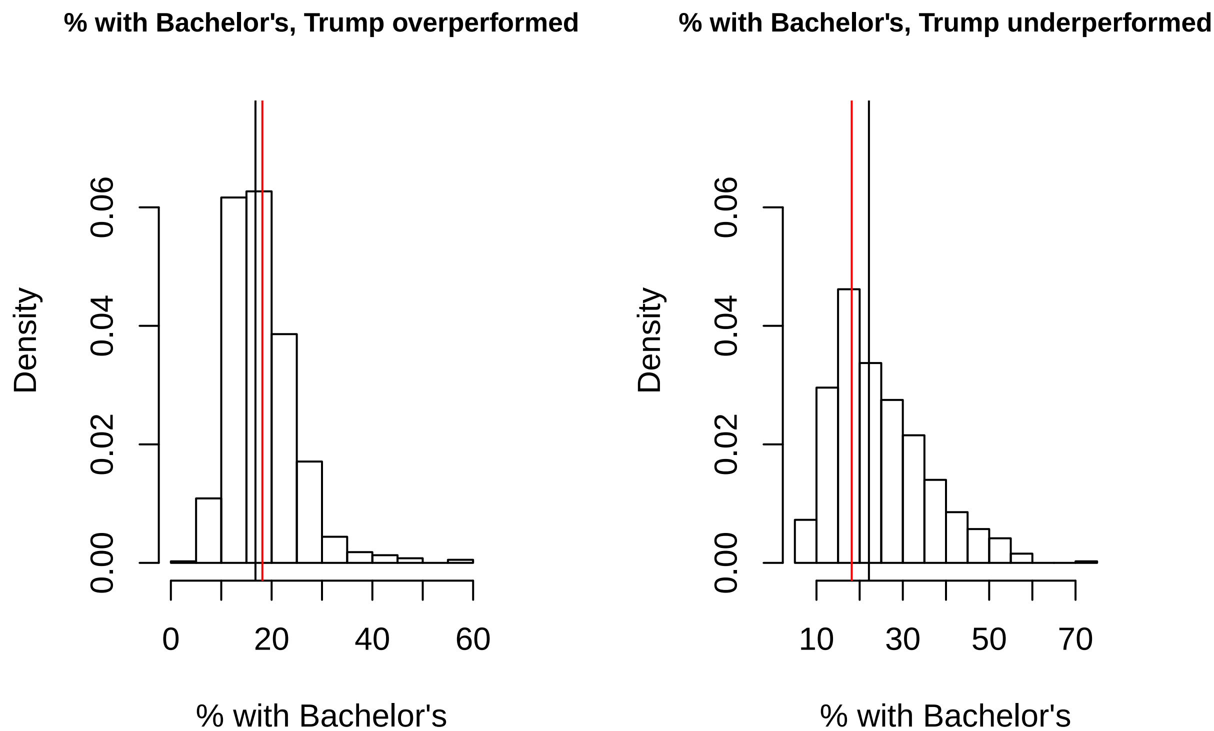
Let's take a closer look at the counties where Trump most over- and underperformed. Are these counties that defied our expectations unusual in any interesting ways?

```
# make quantiles of prediction error
pred.error.quant <- quantile(toplot$pred.error)

# subset data
final.top <- toplot[toplot$pred.error >= pred.error.quant[4],]
final.bottom <- toplot[toplot$pred.error < pred.error.quant[2],]

# plot percent foreign-born
par(mfrow = c(1, 2), pin = c(1.7, 2.5)) ## put plots side by side

hist(final.top$pct_for_born15, freq = FALSE, breaks = 12,
     main = "% foreign born, Trump overperformed", cex.main = .8,
     xlab = "% foreign born",
         ylim = c(0, .15))
abline(v = median(final$pct_for_born15, na.rm = TRUE), col = "red")
abline(v = median(final.top$pct_for_born15, na.rm = TRUE), col = "black")

hist(final.bottom$pct_for_born15, freq = FALSE, breaks = 12,
     main = "% foreign born, Trump underperformed", cex.main = .8,
     xlab = "% foreign born",
         ylim = c(0, .15))
abline(v = median(final$pct_for_born15, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_for_born15, na.rm = TRUE), col = "black")
```

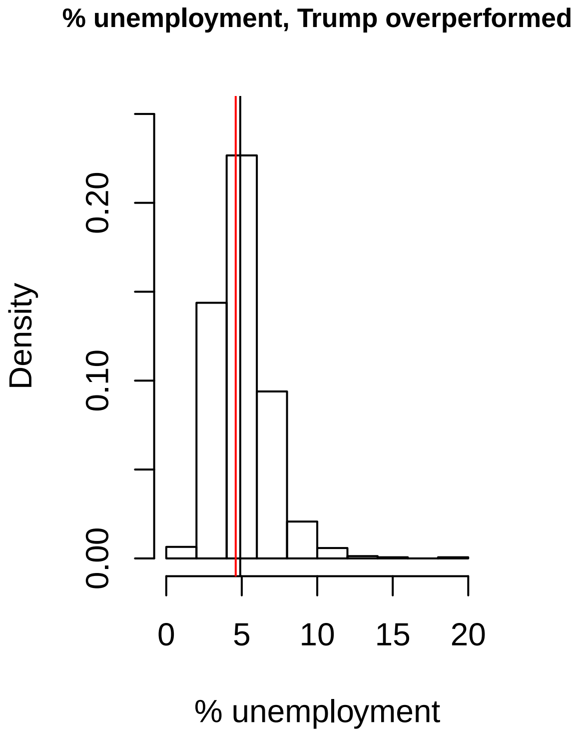


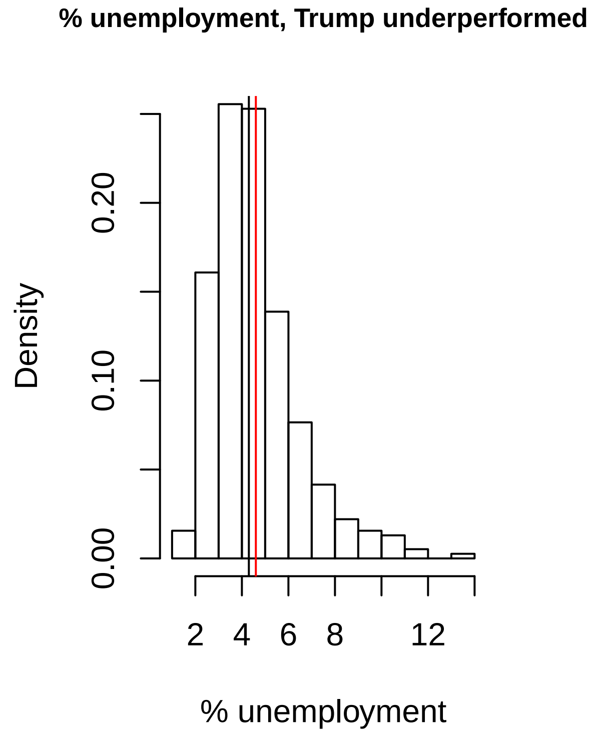



```r
# plot education
par(mfrow = c(1, 2), pin = c(1.7, 2.5)) ## put plots side by side

hist(final.top$pct_bach_deg15, freq = FALSE, breaks = 12,
     main = "% with Bachelor's, Trump overperformed", cex.main = .8,
     xlab = "% with Bachelor's",
     ylim = c(0, .075))
abline(v = median(final$pct_bach_deg15, na.rm = TRUE), col = "red")
abline(v = median(final.top$pct_bach_deg15, na.rm = TRUE), col = "black")

hist(final.bottom$pct_bach_deg15, freq = FALSE, breaks = 12,
     main = "% with Bachelor's, Trump underperformed", cex.main = .8,
     xlab = "% with Bachelor's",
     ylim = c(0, .075))
abline(v = median(final$pct_bach_deg15, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_bach_deg15, na.rm = TRUE), col = "black")
```

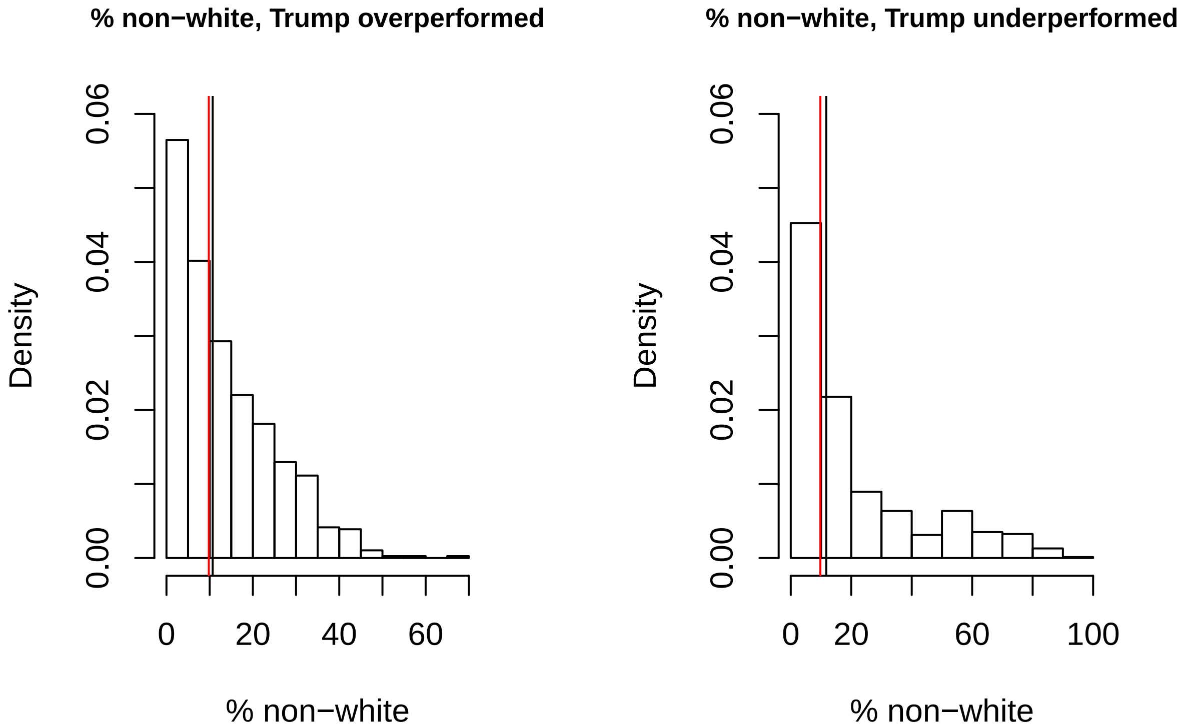



```
# plot unemployment
par(mfrow = c(1, 2), pin = c(1.7, 2.5)) ## put plots side by side

hist(final.top$pct_unemp16, freq = FALSE, breaks = 12,
     main = "% unemployment, Trump overperformed", cex.main = .8,
     xlab = "% unemployment",
         ylim = c(0, .25))
abline(v = median(final$pct_unemp16, na.rm = TRUE), col = "red")
abline(v = median(final.top$pct_unemp16, na.rm = TRUE), col = "black")

hist(final.bottom$pct_unemp16, freq = FALSE, breaks = 12,
     main = "% unemployment, Trump underperformed", cex.main = .8,
     xlab = "% unemployment",
     ylim = c(0, .25))
abline(v = median(final$pct_unemp16, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_unemp16, na.rm = TRUE), col = "black")
```






```
# plot percent non-white
par(mfrow = c(1, 2), pin = c(1.7, 2.5)) ## put plots side by side

hist(final.top$pct_non_white15, freq = FALSE, breaks = 12,
     main = "% non-white, Trump overperformed", cex.main = .8,
     xlab = "% non-white",
         ylim = c(0, .06))
abline(v = median(final$pct_non_white15, na.rm = TRUE), col = "red")
abline(v = median(final.top$pct_non_white15, na.rm = TRUE), col = "black")

hist(final.bottom$pct_non_white15, freq = FALSE, breaks = 12,
     main = "% non-white, Trump underperformed", cex.main = .8,
     xlab = "% non-white",
     ylim = c(0, .06))
abline(v = median(final$pct_non_white15, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_non_white15, na.rm = TRUE), col = "black")
```



In counties where Trump did much worse than predicted, education, percent foreign born, and percent non-white tended to be a little higher than in counties where he did much better than predicted, and unemployment was a little lower. But none of these differences appears particularly large.