Analyzing the 2016 US Presidential Election

We analyze returns from the 2012 and 2016 elections in order to understand the social and demographic trends that may have contributed to Donald Trump's victory in 2016. 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, `election2012.csv`, has one observation per county and contains the following variables:

Name	Description
`FIPS`	FIPS code (unique county identifier)
`state`	State abbreviation
`county`	County name
`votes_dem_12`	Number of votes cast for Democratic candidate, 2012 election
`votes_gop_12`	Number of votes cast for Republican candidate, 2012 election
`votes_total_12`	Total number of votes cast in 2012 election

The second, `election2016.csv`, has the same data structure and similar variable names but reports data for the 2016 presidential election.

The third dataset, `county.csv`, includes social and demographic characteristics for each county:

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)

Question 1

Start by load all three datasets. Merge the three datasets by FIPS code to construct one complete data file for analysis. Check your merge to see how many observations came in from all three sources. Did you lose much data? Finally, perform listwise deletion (hint: check section 3.2 of *QSS*) of missing values on the full dataset. Did you lose much data? Get whatever characteristics you can on the data you lost. What can you say about these observations?

Answer 1

```r
## read in data

returns12 <- read.csv("data/election2012.csv")
returns16 <- read.csv("data/election2016.csv")
covars <- read.csv("data/county.csv")

## which columns do we want in both? No need to duplicate state and county.
columnsToKeep <- c("FIPS", "votes_dem_16", "votes_gop_16", "votes_total_16")

## now, lets merge
returns <- merge(returns12, returns16[, columnsToKeep], by = "FIPS")
#left_join(return12, returns16, by = "")

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

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

```r
dim(returns16)
```

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

```r
dim(returns)
```

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

```r
## no!

## 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, ]
lost$county
```

```
##  [1] "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska"
##  [9] "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska"
## [17] "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska" "Alaska"
## [25] "Alaska" "Alaska" "Alaska" "Alaska" "Alaska"
```

It looks like the lost data are mostly observations from the whole state of Alaska (as well as one observation for Oglala County, SD), which were probably erroneously included in this county-level dataset to begin with. All in all, a successful merge.

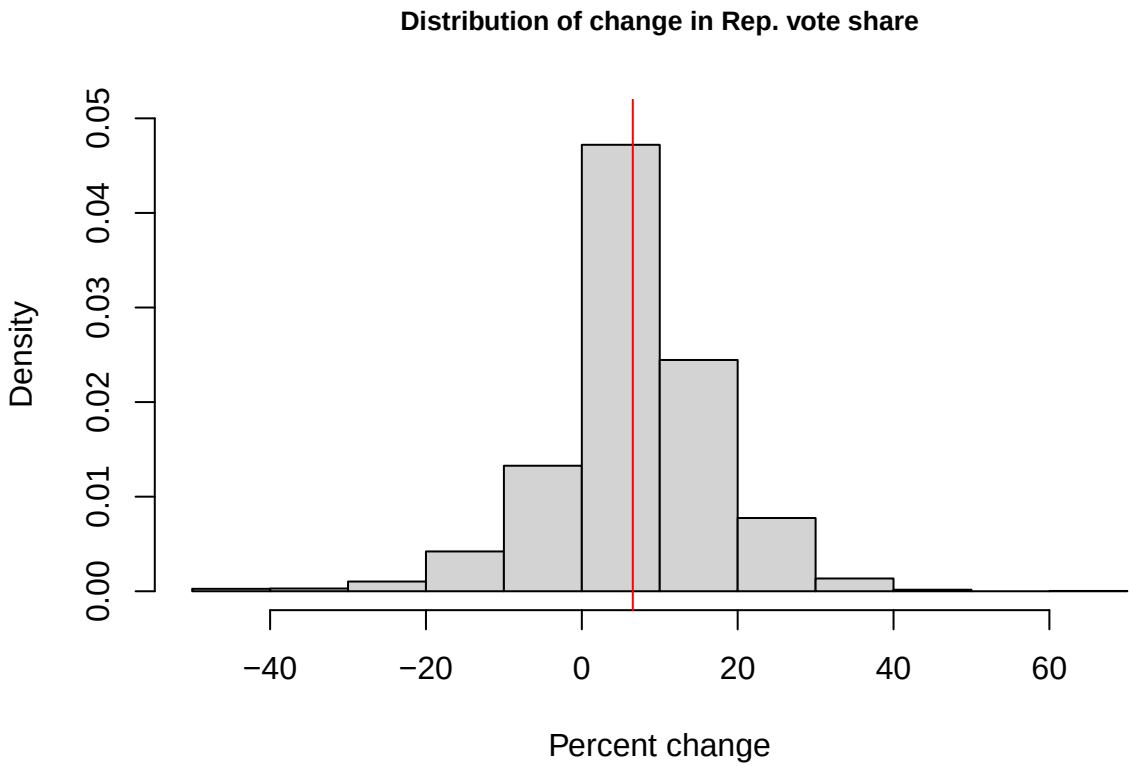
Question 2

Compute the Republican vote share as a proportion of total votes, in 2012 as well as in 2016. Also compute the percent difference in this Republican vote share variable from the 2012 to 2016 election. Plot the distribution of this percent difference, with a red line at the median.

Then, subset your data to just the battleground states: Florida, North Carolina, Ohio, Pennsylvania, New Hampshire, Michigan, Wisconsin, Iowa, Nevada, Colorado, and Virginia. Plot the distribution of the same variable in this sample, with a red line at the sample median.

Answer 2

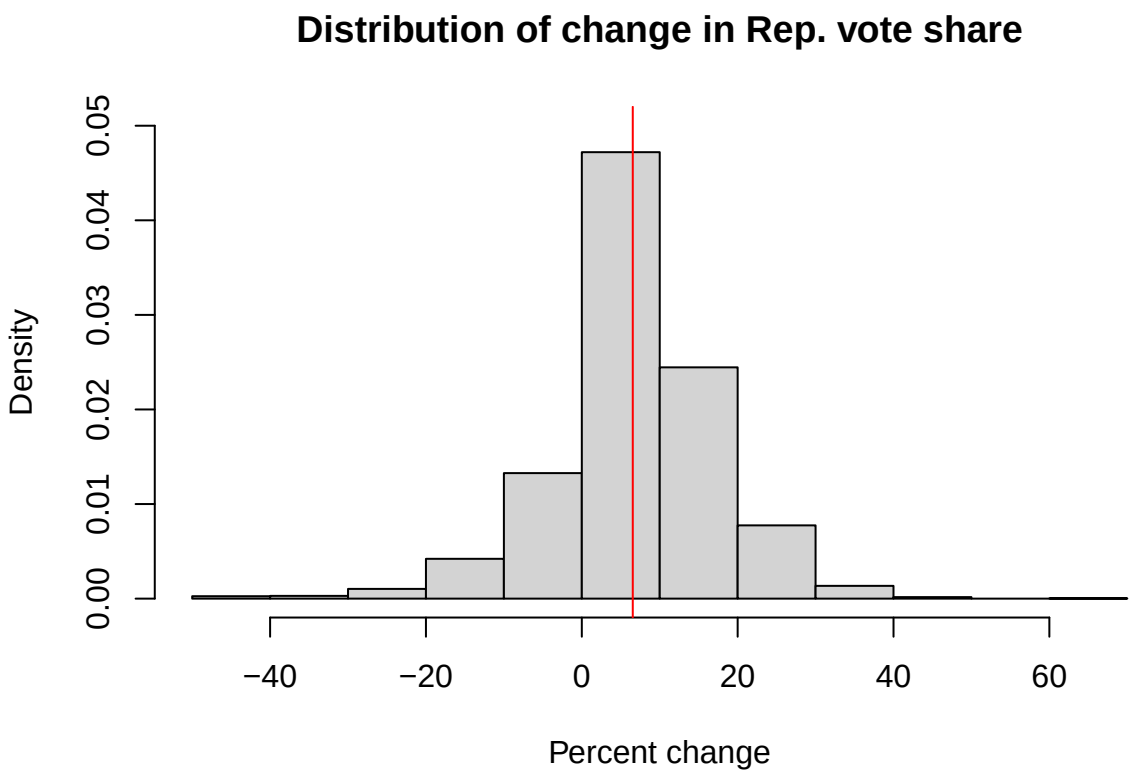
```
## vote share variables: Republican / Total
final$gop_vs_12 <- final$votes_gop_12 / final$votes_total_12
final$gop_vs_16 <- final$votes_gop_16 / final$votes_total_16

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

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



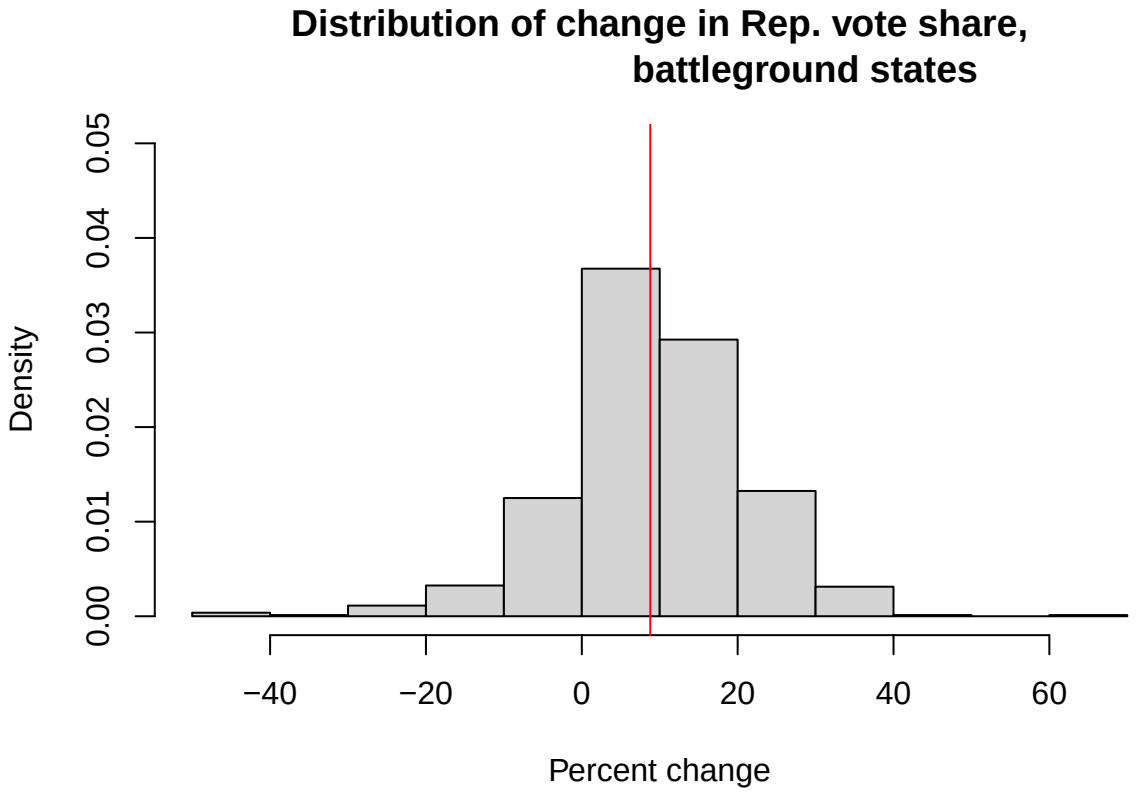

```
## subset data to battleground states
battlestates <- c("FL", "NC", "OH", "PA", "NH", "MI", "WI", "IA", "NV", "CO", "VA")
battle <- subset(final, state %in% battlestates)
table(battle$state) ## check that you subset properly
```

```
##
##  CO  FL  IA  MI  NC  NH  NV  OH  PA  VA  WI
##  64  67  99  83 100  10  17  88  67 133  72
```

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




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

Question 3

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. To create this map, you will need to take the following steps:

1. If you have not yet done so, install the `maps` package and load it.
2. Take the `county.fips` dataset, which comes with the `maps` library, and perform a merge with the dataset you were working with in the last question, by FIPS code – that is, make sure that your merged data contains all the observations from `county.fips`, in their original order (hint: use the `all.x` argument).
3. As you know, `alpha` values are typically used in the interval [0,1]. One way to normalize data to this range is to calculate, for a vector x:

$$\frac{x - \min(x)}{\max(x) - \min(x)}$$

Use this normalization strategy on the Republican vote share variable you calculated above and store the result in an object.

Then, use the `rgb()` function to create a vector of appropriate colors (red for Republican gains, blue for Republican losses), using the vector you created for the `alpha` argument inside that function. Think carefully about what a Republican *gain* and *loss* mean in relation to the vote share variable you calculated. Finally, use this vector of colors within the `map()` function. Include the `lty = 0` option to get rid of black borders around the states.

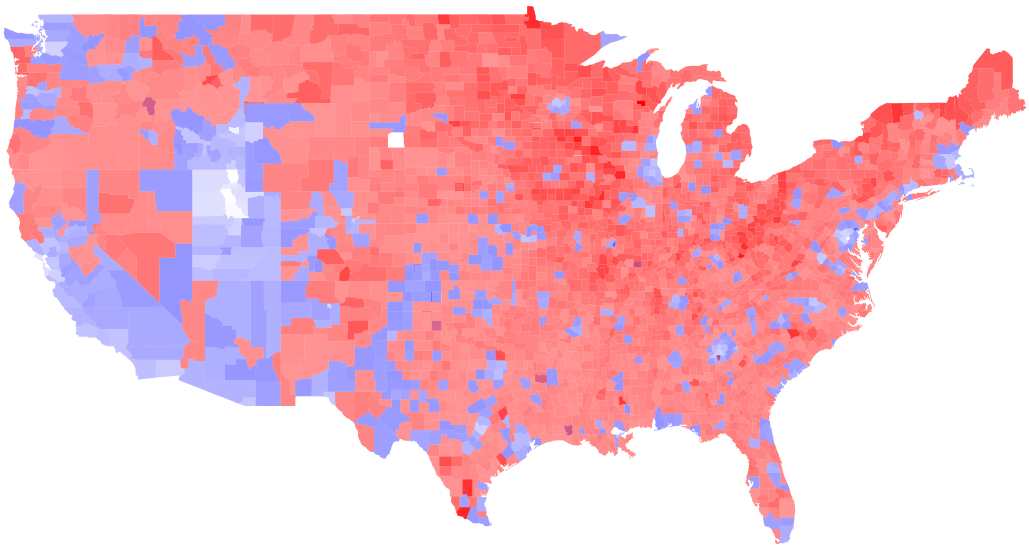
Comment on your results. In what parts of the U.S. did Republicans make the most significant gains? What other interesting patterns do you observe?

Answer 3

```
## load library
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.4.1
```

```
## merge data onto county.fips
cf <- county.fips
names(cf) <- c("FIPS", "name")
toplot <- merge(cf, final, by = "FIPS", all.x = TRUE)

## normalize between 0 and 1
alpha1 <- (toplot$gop_vs_pct_ch - min(toplot$gop_vs_pct_ch, na.rm = T)) /
    (max(toplot$gop_vs_pct_ch, na.rm = TRUE) - min(toplot$gop_vs_pct_ch, na.rm = TRUE))

## 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)) {
  if(i==2389) next
  map(database = "county", regions = toplot$name[i], col = toplot$cols1[i],
      fill = TRUE, add = TRUE, lty = 0)
}
```



This map shows that 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 region

appears to be the middle of the country.

Question 4

Run a regression of percent change in Republican vote share from 2012 to 2016 on percent foreign-born, percent holding a Bachelor's degree or above, percent non-white, and percent unemployed. Intepret your results.

Answer 4

```
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
```

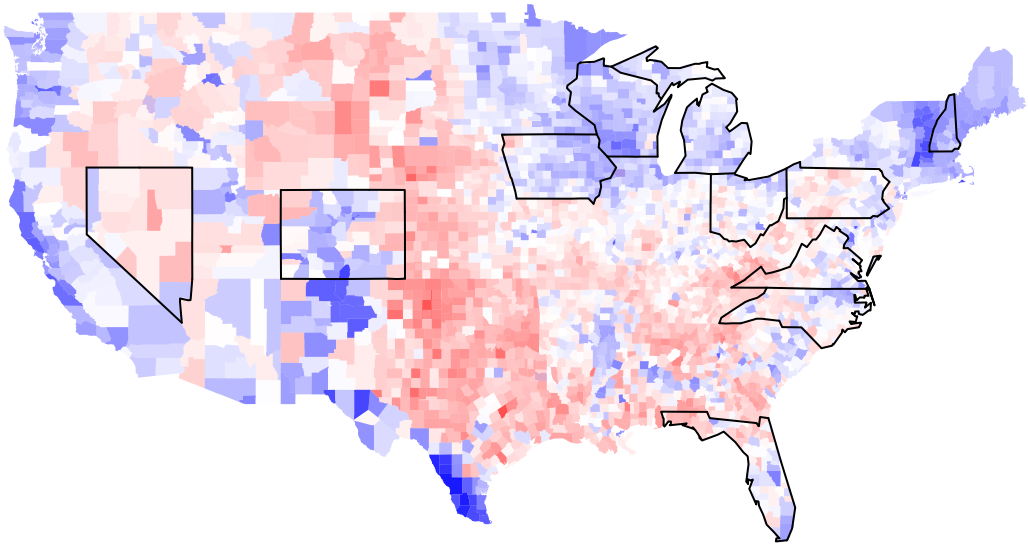
Percent foreign-born, percent with a Bachelor's degree or above, and percent non-white are all statistically significant predictors of Republican losses since 2012. Somewhat surprisingly, unemployment is not a strong predictor.

Question 5

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, 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. For the first three, you can use the variables ending in `15` since these are the most recent available Census estimates, which are averaged over the period 2011-15. Then predict 2016 Republican vote share in each county using the same 2011-15 variables and percent unemployed in 2016. Compute the prediction error, which is the predicted Republican vote share subtracted from the observed value in 2016. Create a county-level map with counties colored in red where the observed value was higher than the prediction and blue otherwise. Use double the absolute value of the prediction error as the intensity of the color (the `rgb()` `alpha` parameter). Comment on the results.

Answer 5

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

## 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$preds <- predict(lm2, pred.df)
toplot$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)) {
  if(i==2389) next
  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 — which is 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.

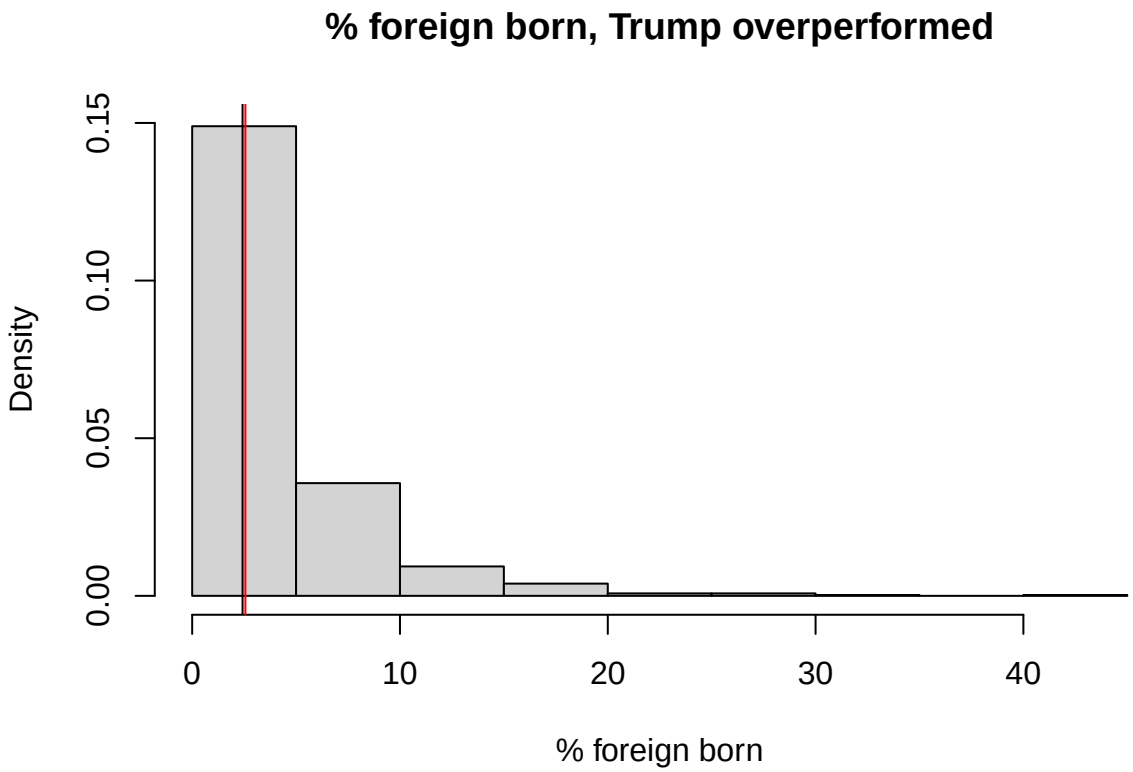
Question 6

Subset the data to the counties with the largest overpredictions and underpredictions of Republican vote share based on the last question (take the top and bottom quantiles of prediction error).

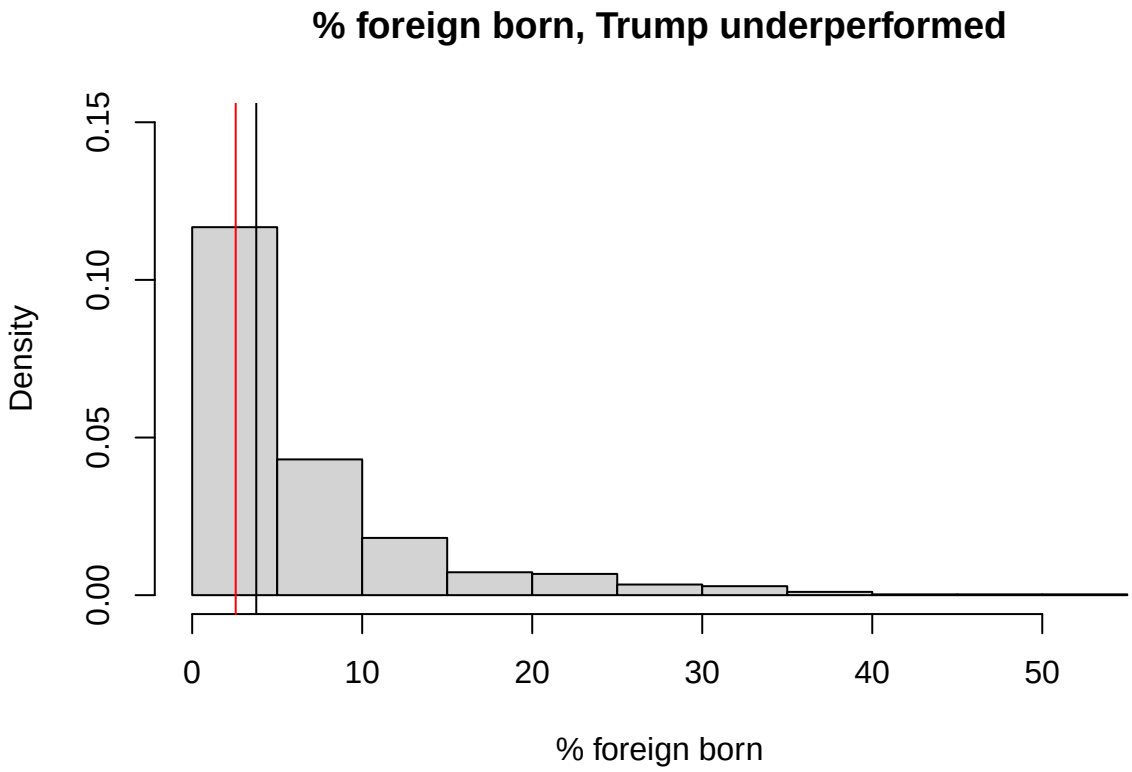
Create some histograms using these subsets, with black lines at the medians in the subset and red lines at the medians from the full data. Are the counties that defied our expectations unusual in any interesting ways?

Answer 6

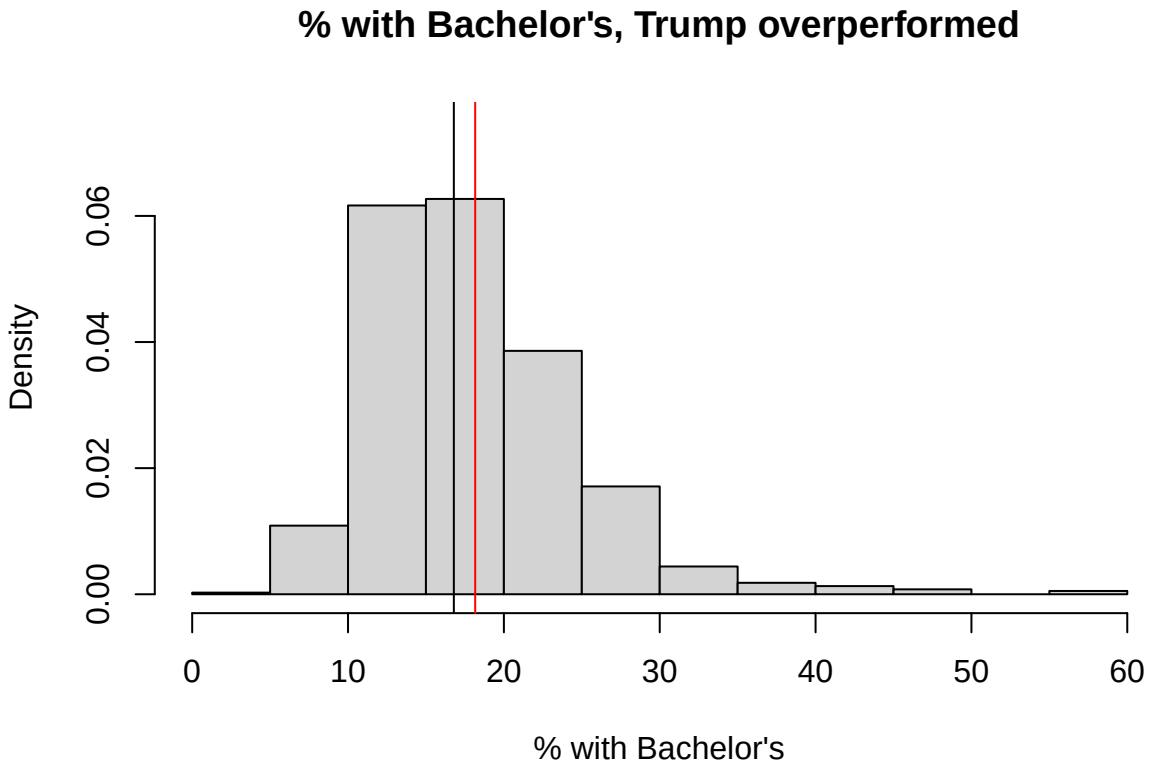
```
## make quantiles of prediction error
pred.error.quant <- quantile(toplot$pred.error)

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



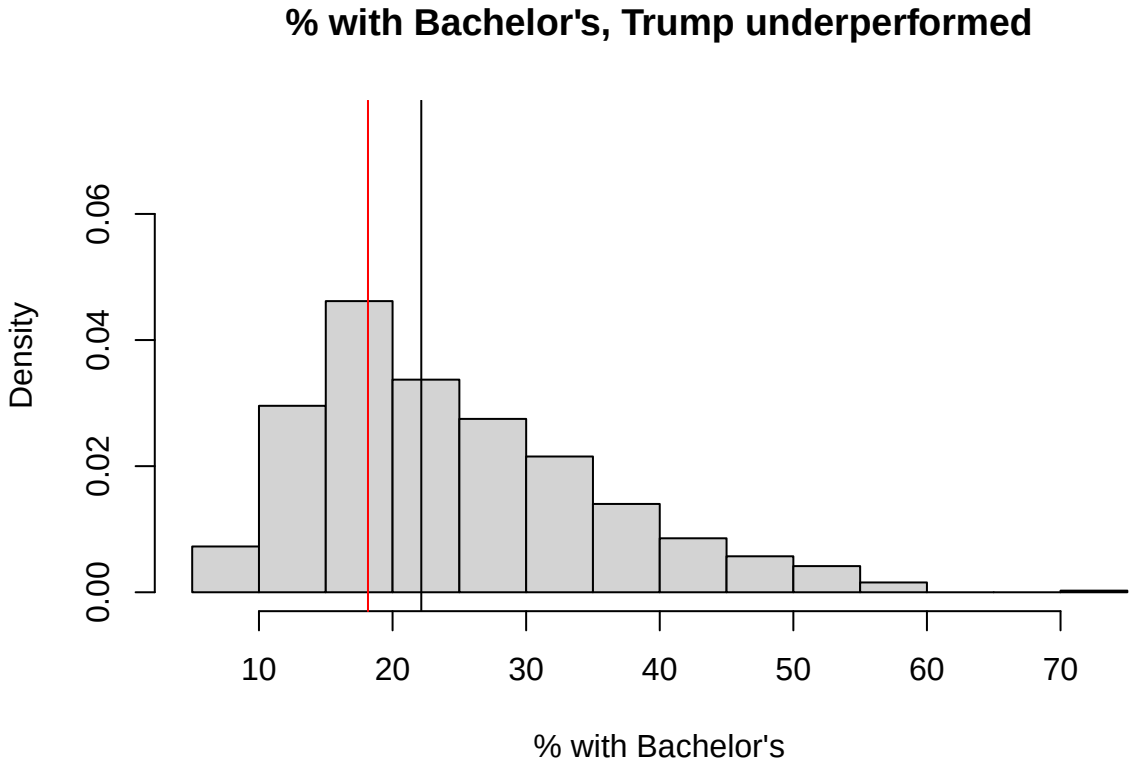
```r
hist(final.bottom$pct_for_born15, freq = FALSE,
     main = "% foreign born, Trump underperformed",
     xlab = "% foreign born",
         ylim = c(0, .15))
abline(v = median(final$pct_for_born15, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_for_born15, na.rm = TRUE), col = "black")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



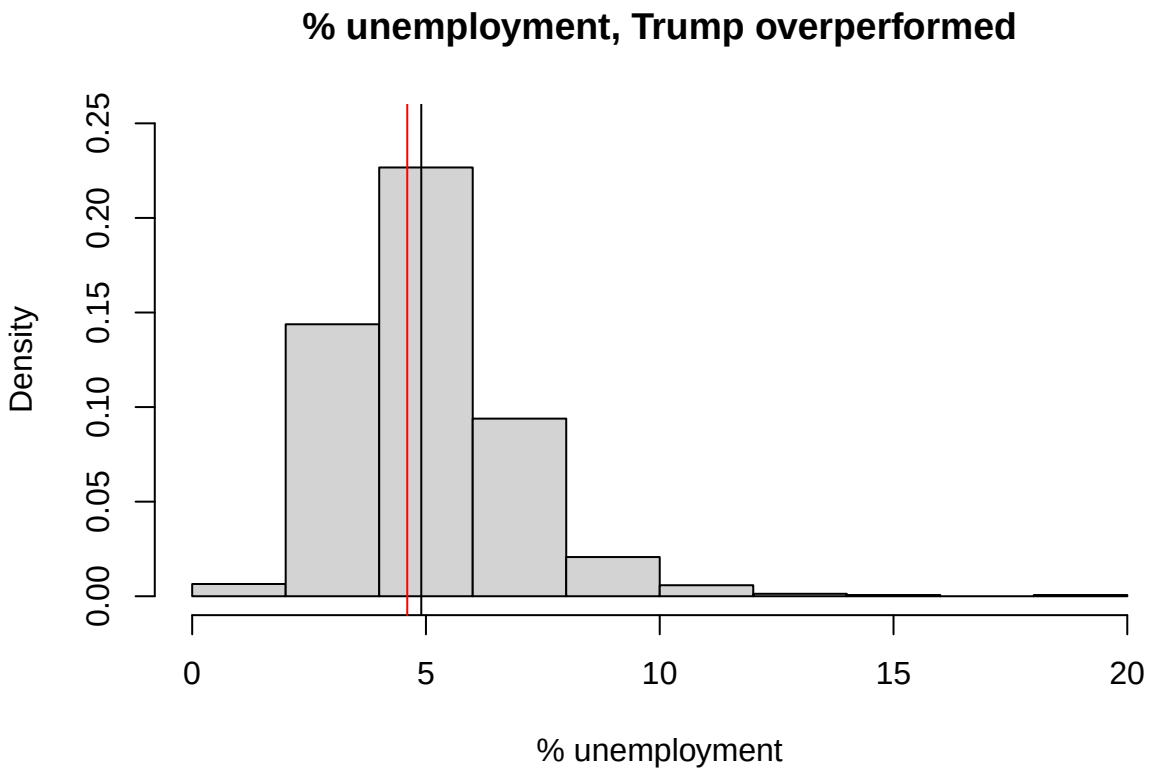

```r
## plot education
hist(final.top$pct_bach_deg15, freq = FALSE,
     main = "% with Bachelor's, Trump overperformed",
     xlab = "% with Bachelor's",
     ylim = c(0, .075))
abline(v = median(final$pct_bach_deg15, na.rm = TRUE), col = "red")
abline(v = median(final.top$pct_bach_deg15, na.rm = TRUE), col = "black")
```



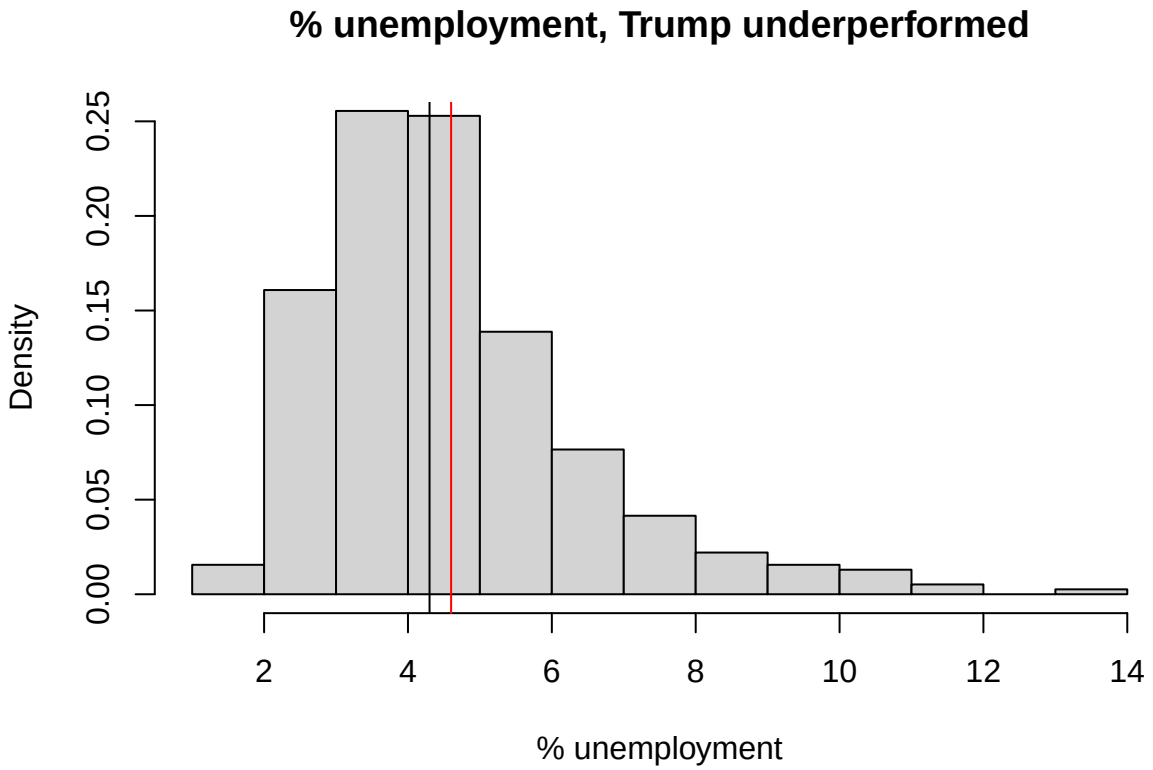

```
hist(final.bottom$pct_bach_deg15, freq = FALSE,
     main = "% with Bachelor's, Trump underperformed",
     xlab = "% with Bachelor's",
     ylim = c(0, .075))
abline(v = median(final$pct_bach_deg15, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_bach_deg15, na.rm = TRUE), col = "black")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



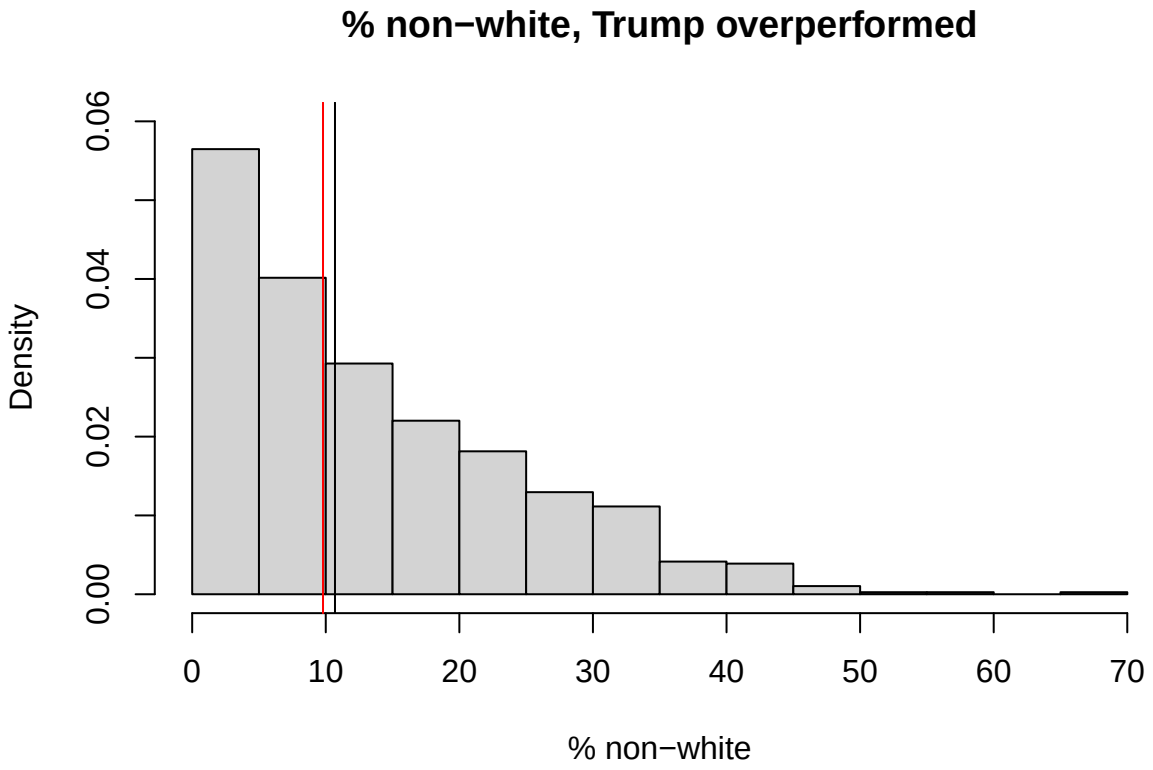

```
## plot unemployment
hist(final.top$pct_unemp16, freq = FALSE,
     main = "% unemployment, Trump overperformed",
     xlab = "% unemployment",
         ylim = c(0, .25))
abline(v = median(final$pct_unemp16, na.rm = TRUE), col = "red")
abline(v = median(final.top$pct_unemp16, na.rm = TRUE), col =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



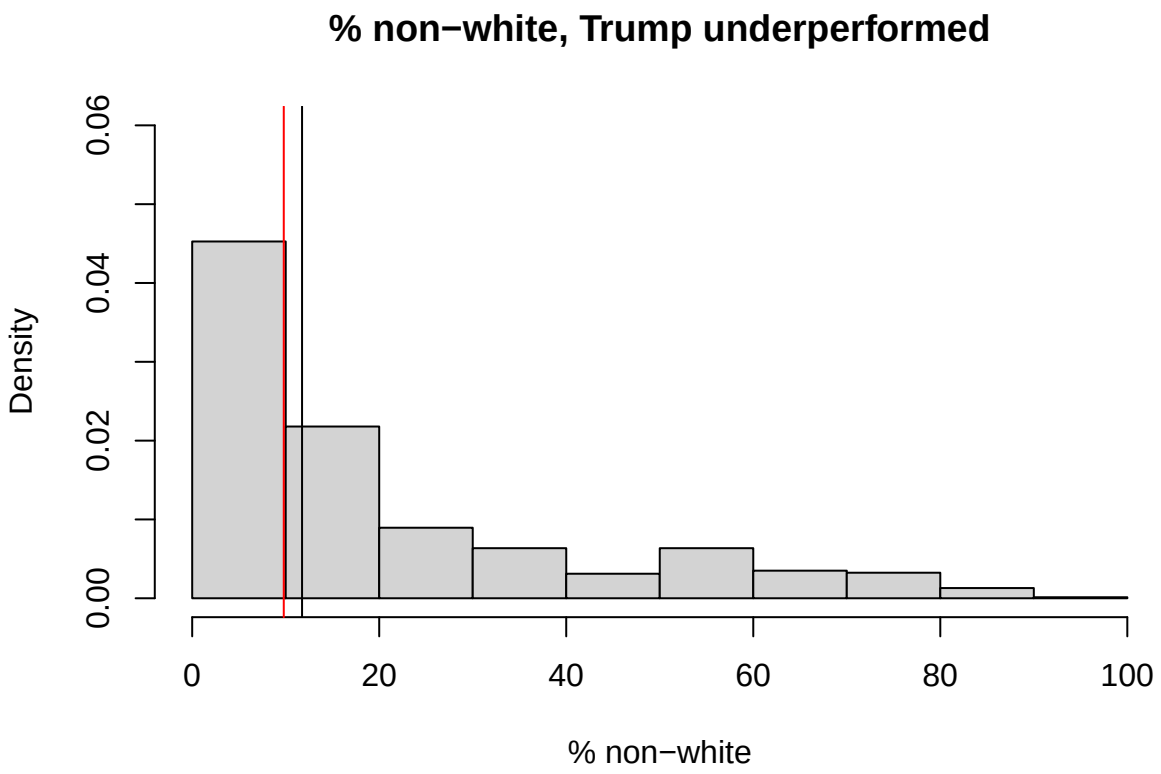

```
hist(final.bottom$pct_unemp16, freq = FALSE,
     main = "% unemployment, Trump underperformed",
     xlab = "% unemployment",
     ylim = c(0, .25))
abline(v = median(final$pct_unemp16, na.rm = TRUE), col = "red")
abline(v = median(final.bottom$pct_unemp16, na.rm = TRUE), col = "black")
```




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




```
hist(final.bottom$pct_non_white15, freq = FALSE,
     main = "% non-white, Trump underperformed",
     xlab = "% non-white",
     ylim = c(0, .06))
abline(v = median(final$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.