2016 Democratic Presidential Primary

Cole Sanders

March 12, 2018

Background

The 2016 Presidential Democratic Primaries offer an interesting look into the changing political alignments of the Democratic Party. Bernie Sanders, a proponent of universal health care, state funded higher education, and a self-described "Democratic Socialist" was a surprise challenger to the eventual Democratic Nominee Hillary Clinton. Clinton campaigned on expanding the Affordable Care Act, expanding funding to higher education, and was generally considered a more traditional candidate for the party. While the two both offered liberal platforms, Sanders' platform was generally considered more progressive, and he labeled himself as a "Democratic Socialist." That Sanders still saw large scale success in the primaries opens up questions as to whether Democratic voters are starting to prefer more progressive politicians. To unravel this question we need to first look at the areas where Sanders saw his success to gain a better understanding of who his base was. The following is the 2016 Democratic Primary Results by County.

Outcome Map

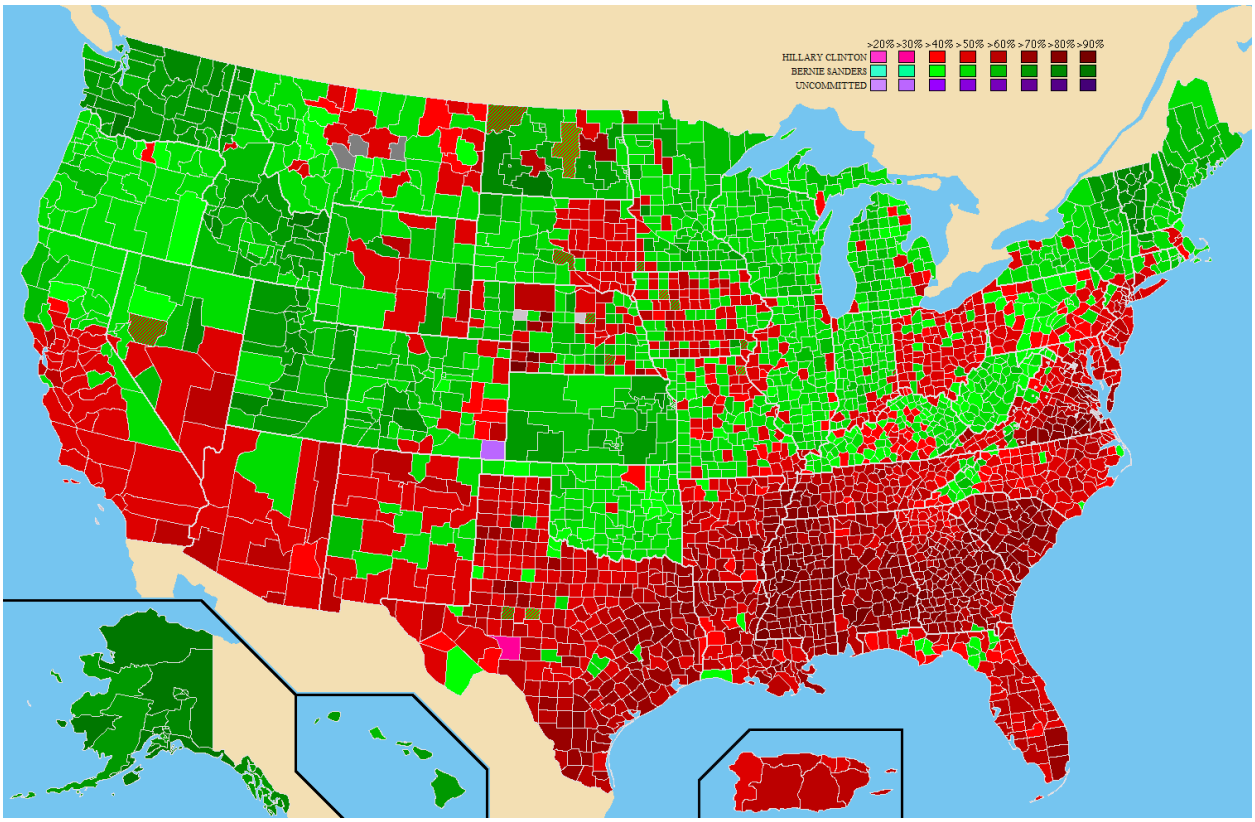


Figure 1: Map Sourced from US Election Atlas.

The Question

Looking at the Democratic Primary result map, it seems that there are trends or groupings in the counties voting for Clinton versus those voting for Sanders. Specifically it looks like Clinton performs very well across the southern half of the United States as well as the middle and upper East Coast. Sanders on the other hand seems to have most of his wins on the Northern Half of the United States. This trend raises the question of what exactly is the difference between the populations voting for a more traditional liberal candidate like Clinton and those voting for a more progressive candidate like Sanders?

Using US Census Data on County Demographics we can explore trends between county characteristics and how strongly the county voted for one candidate versus the other. It is important to note that we are specifically looking at the population voting in the Democratic Primaries and as such this is not representative of the general public or even left leaning Americans as a whole. However it does give a look into the most energetic voting base of the Democratic Party. This population does ultimately decide who will be the Democratic Presidential Candidate, so they are still a worthwhile population to study.

Methodology

For this study we will be using demographic characteristics of a county to fit a mutilinear regression to predict how strongly a county voted for Clinton or Sanders. Our response variable will be the difference in percentage of votes each candidate received. We shall call this variable Outcome and it is derived from the following formula.

Outcome = Clinton % - Sanders %

Thus any positive value for Outcome indicates a Clinton win for that county and any negative value for Outcome indicates a Sanders win for that county. This is a common response variable modeling in political sciences and news casting and will help us reflect the more dynamic relationship between county demographics and preferred candidate. This sort of coding is traditionally referred to as a point difference and we will continue this terminology. So if we say Clinton is favored by 2 points this equates to her getting an extra 1% of the vote over Sanders. Thus

Outcome / 2 = Precentage of Vote above 50

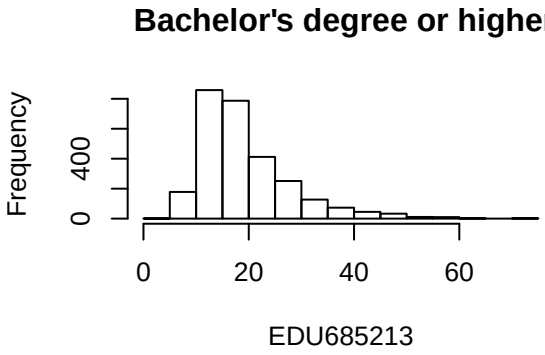
where any positive value is the percentage above 50% for Clinton and any negative value is the percentage above 50% for Sanders.

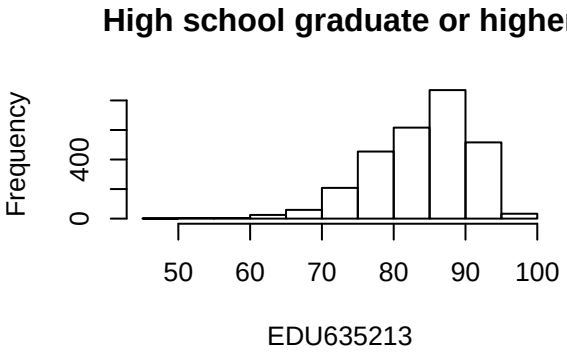
Simple Observations based on Outcome

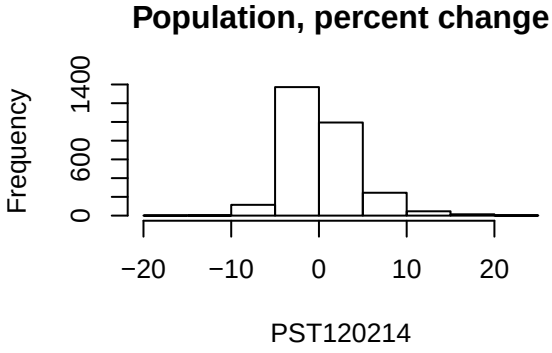
- The mean of our Outcome variable is .10, which means on average Clinton won counties with a 10 point lead over Sanders.
- There are 1640 observations where Outcome is positive, signifying a Clinton win in 1640 counties.
- In counties that Clinton won, Clinton averaged a 30 point lead over Sanders
- There are 1127 observations where Outcome is negative, signifying a Sanders win in 1127 counties.
- In counties that Sanders won, Sanders averaged a 17.6 point lead over Clinton

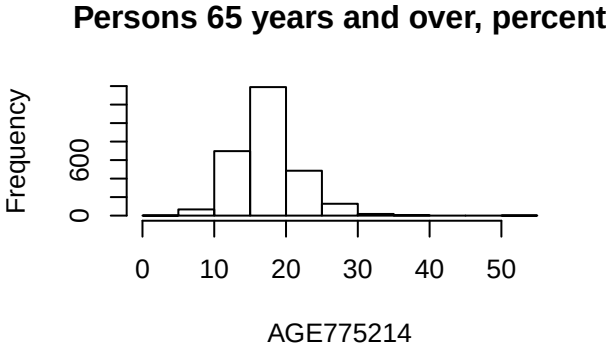
Right away, we can see that Clinton on average beat Sanders by a decent margin on the county level, further she also won over 500 more counties than Sanders did. This however leads us to another important distinction that counties are drawn across geographic lines and not based on populations like electoral districts. This does not hamper our analysis because we are interested in the demographic make up of the county and not necessarily its population. It is simply important to remember that districts are the basis of elections and not counties.

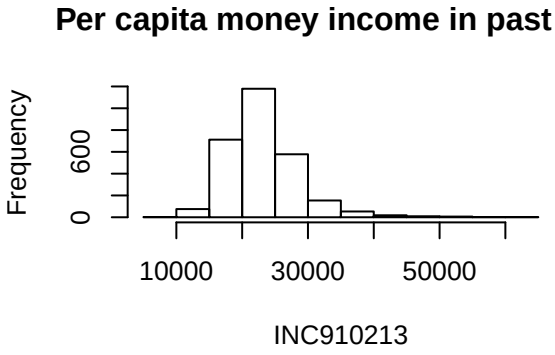
Main Predictor Distributions (After Data Cleaning)

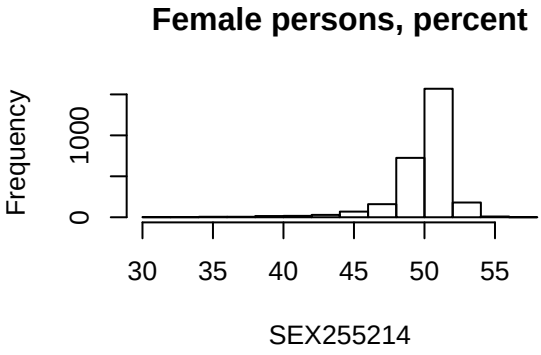


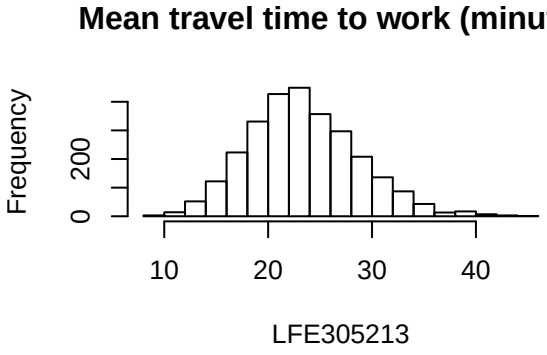






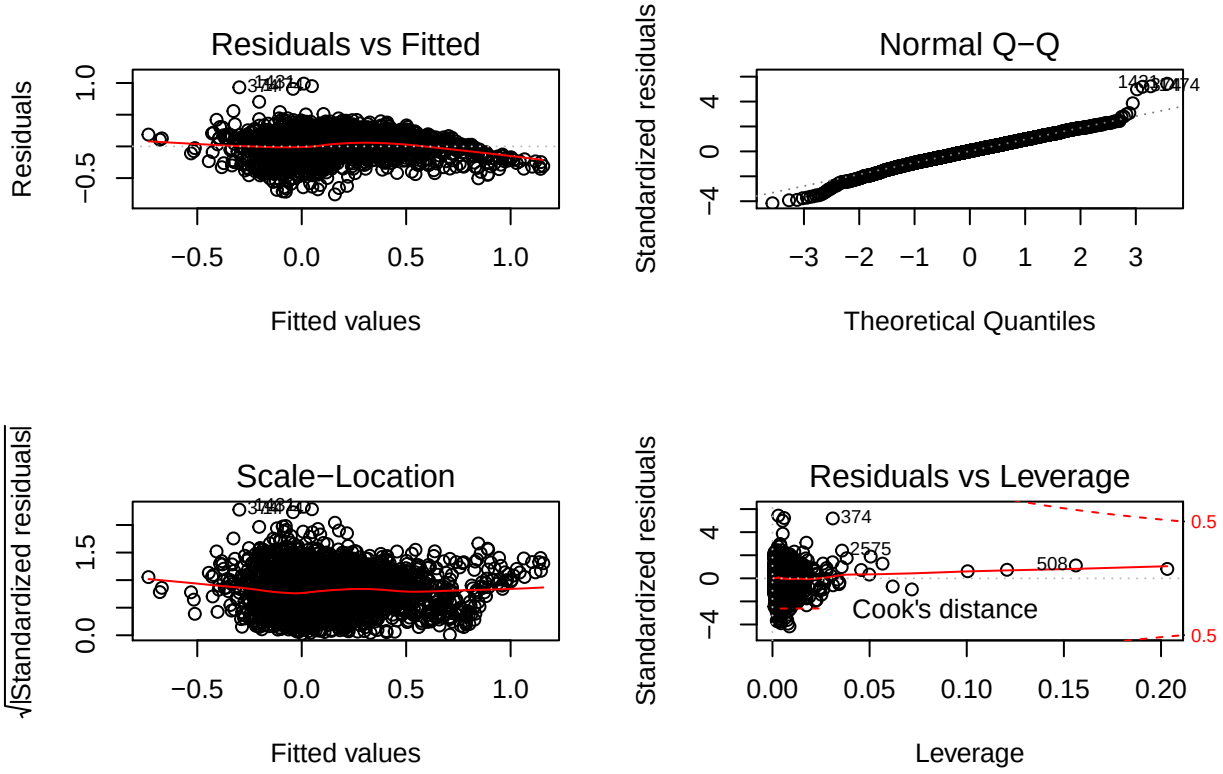






Our Model

By sampling several different demographic factors, we find that we can build a multilinear model with 13 predictor variables (of which the 7 most prominent were previously shown) with an R^2 of just under .7. We will continue with only these 13 predictors because additional variables do not significantly increase R^2 and also tend to correlate with predictors already in our model.



We look at several metrics to judge the overall health of our model and in general our model seems to behave very well. There is no homoscedacity in our residual values and the normal Q-Q of our model is almost perfectly linear. We also notice that there are a few outliers in the data, these specific data points are associated with Clinton wins and all of these counties have a large majority African American populations which ends up being one of the strongest predictors for a Clinton win. Since these follow the general trend of the model and have minimal leverage on the overall model we will leave them in our data.

Significant Predictors

The following is a list of our predictors and how on average they affect the outcome for each county:

- Counties with high median value housing favored Sanders, with every $20,000 increase in median value of owner-occupied housing units, the outcome shifts 1.5 points in Sanders favor.
- For each increase of 10% in county's population (age 25+) with a bachelor's degree or higher, the outcome shifts 5.7 points in Sanders favor.
- For each increase of 1% in county's population (age 25+) with a high school degree, the outcome shifts 1 point in Sanders favor.
- Counties that increased in population from 2010 to 2014 favored Clinton, with each 1% increase in population size corresponding to about a point in her favor.
- Counties with older populations heavily favored Clinton, with 8 points in her favor for every 10% of the population that was over 65
- Counties with relatively large African American populations heavily favored Clinton, with 1.5 points in her favor for each 1% of the population identifying as African American
- Counties with relatively large Hispanic and Latino populations slightly favored Clinton, with .29 points in her favor for each 1%of the population identifying as Hispanic and/or Latino
- For every extra $1000 in per capita income in over 12 months a county had, the outcome shifted 1.9 points in Clinton's favor.
- For each 1% above 50% of the population in a county identifying as female, the outcome shifted 7.4 points in Clinton's favor.
- For each 1% increase in the population of voters identifying as two or more races in a county, the outcome shifts 3 points in Sanders' favor. (This group is relatively small and often white is one of the two races being identified as which could be a factor in this group favoring Sanders).

Realtionships between a County's Demographics and thier Outcome

Ideal County for Sanders

Sanders had the best response in counties that are predominantly white and that have high rates of higher education. It is important to note that higher education and identifying as white are correlated in our model and that previous studies have shown these two variables are linked. Thus Sanders performed, on average, stronger with non ethnically diverse counties according to our model. Sanders also performed strongly on average in counties with low senior populations, seeing very favorable gains when seniors made up less than 10% of the population. He also had a slight advantage over Clinton in counties with significant Asian populations, but struggled severely in all other ethnically diverse counties.

Predictors that had a strong positive relationship with a Sanders win in our model:

- Higher male percentage
- High White Identifying percentage
- High Median Household Value
- Low Senior Populations
- High education levels

Ideal County for Clinton

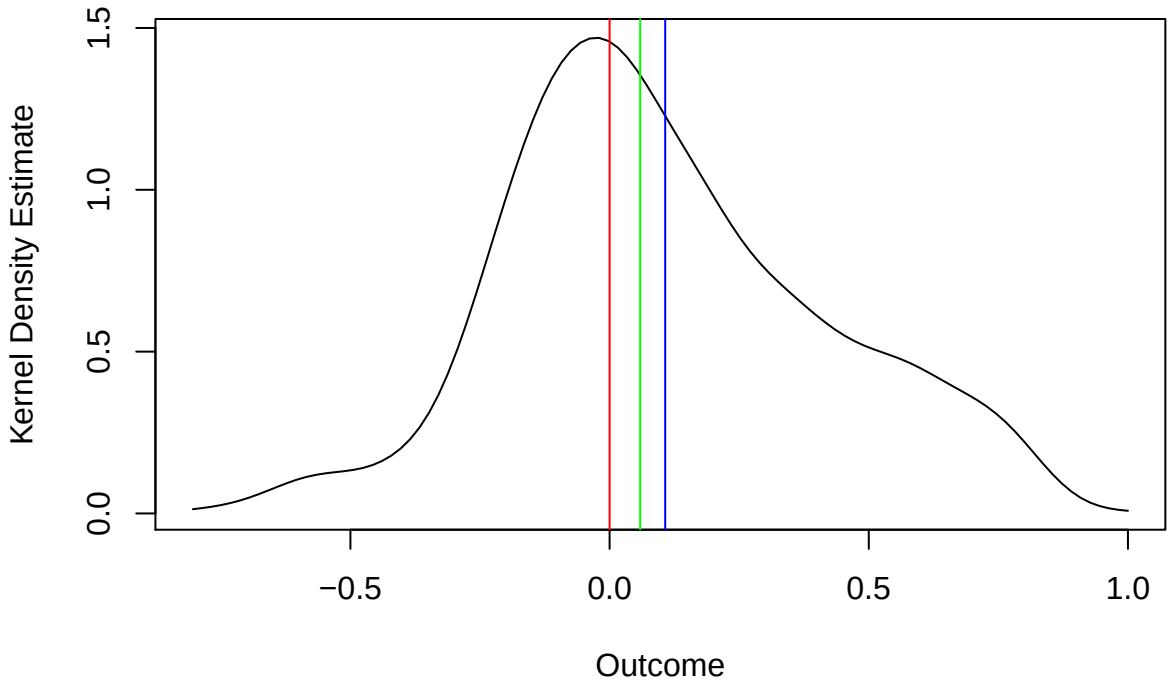
Clinton's strongest performance was in ethnically diverse counties. Any county with a significant African American population went to Clinton by large margins. While not as large, Clinton also performed incredibly well in counties with a large Hispanic population. Clinton also saw favorable turnout in counties that were older and wealthier on average.

Predictors that had a strong positive relationship with a Clinton win in our model:

- Higher Female Percentage
- High African American Populations
- High Senior Population
- High Hispanic Populations
- High Per Capita Income
- Large Population Growth

Kernel Smoothing

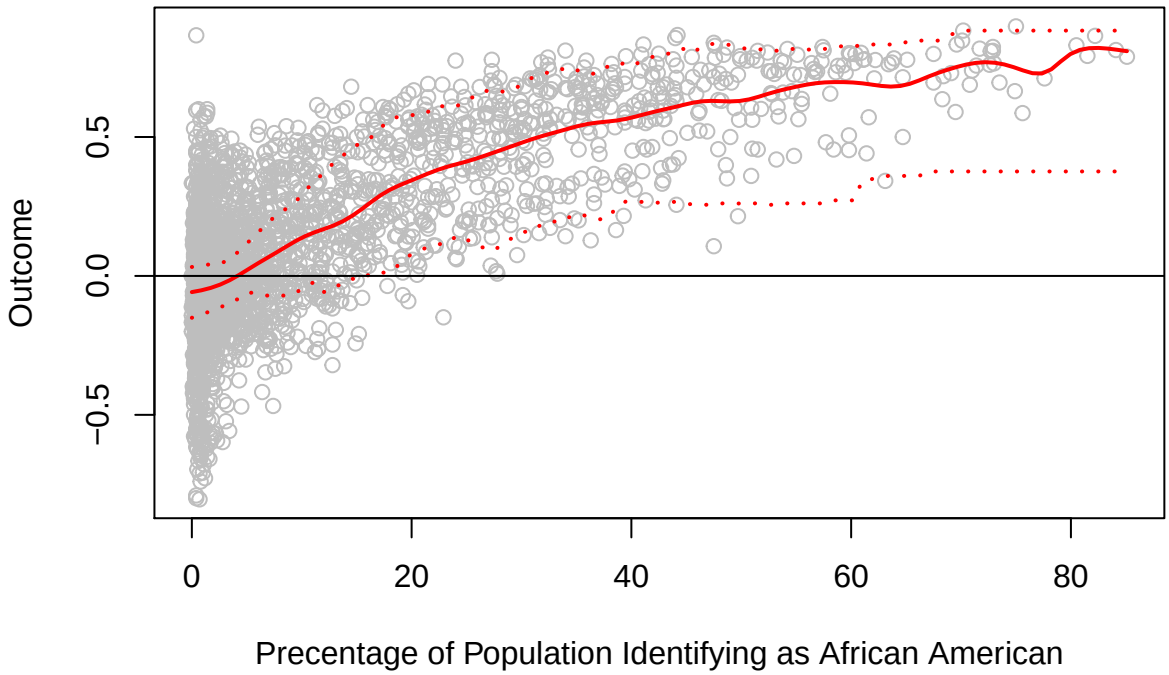
We will now use a non-parametric kernel smoothing to help get an idea of the distribution of observed county wins.



From this we can see an interesting trend in the outcomes of the county elections. First the Outcome variable takes on an approximately normal yet rather skewed distribution. The skewness of the distribution is interesting because it implies that Clinton winning by large margins was somewhat common. For example, from the distribution we can see that Clinton winning a county by about 75 points higher than Sanders happened somewhat frequently, however, Sanders beating Clinton by over 50 points was pretty rare. This distribution shows that Sanders won more close races, while Clinton won by landslides relatively often. Interestingly the curve has its peak at a slightly negative value, indicating that the most common range of would be a very small Sanders win. However when you look at the median (green) and the mean (blue) of the distribution the skewness in the curve becomes apparent. These observations just further support the story that Sanders saw mainly small wins and Clinton won big often.

Kernal Regresssion on Important Predictors

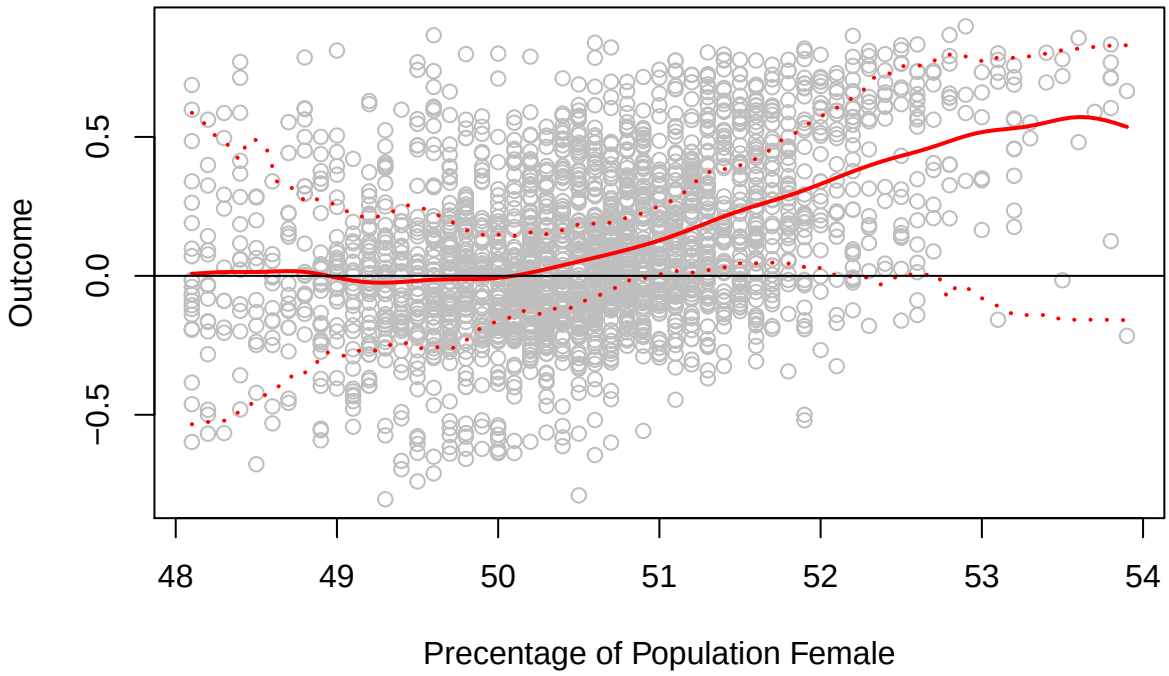
The percent of a county's population identifying as African American ended up being one of our strongest predictors in the model. To make sure we are correctly capturing the relationship between this predictor and the Outcome, we will use kernel regression to remove our parametric assumptions.



From the data and the kernel regression we can see a clear upward trend where the Outcome shifts further in Clinton's favor as the percent of a county's population identifying as African American increases. Further this relationship is nearly linear except for at the highest percentages, however since our outcome has a finite max of 1 this sort of curvature is expected. Also it is interesting to note that from our graph we can see that Sanders only won a single county with an African American population bigger than 20% and that in counties where Sanders had large margins over Clinton, greater than 50 points, all of these counties had consistently small African American populations, always under 10%.

Just based on the data alone we can see that there is a very positive trend for Clinton wins as African American populations within a county increase. Based on our confidence bands we can say with high certainty that after the population within a county identifying as African American has exceed 15%, on average Clinton will win.

Next we look at the relationship of having a slight gender majority in a county in either direction and outcome. Gender data at the county level is severely lacking and due to the nature of the previous election cycle, with issue of gender being a serious topic of debate, I suspect this is an important predictor although it is under represented in this model.



From our kernel regression we can see that even in counties with only a slightly greater female population that the average outcome shifts in Clinton's favor for each percentage point above 50%. The most notable area of interest is in the 50%-52% range where we have enough data points so that our confidence bands are relatively small. In this area marginal increase over the 50% of the population identifying as female are associated with rater large point increase for Clinton. Based on our confidence bands, if the female population of a county is 51-52% of the total population, then on average Clinton will win. I suspect this trend would continue past this range as well but due to the limited data points past this range our confidence bands expand drastically.

This is our only predictor that has anything to do with gender and even this really isn't representative but still has a profound relationship with Outcome. Rather than drawing any further conclusions from this graph I think it shows a good example of where our model is lacking, and that with more detailed data on gender for a county, we could most likely build a more accurate model.

Cautions

This study was based off of a regression analysis of county demographics and their relationship with the county outcome of the Democratic Presidential Primary. Being as such, this model is observational and does not establish any causal links. Nor was this the purpose. We hoped to look at interesting trends in county demographic data that might suggest interesting trends in populations within the Democratic Party that might reveal more about the subgroup within the Democratic Party who supported Sanders and seem to prefer more progressive politicians.

We cannot ignore the huge role played by gender in this election cycle, which is in part why I refrain from drawing conclusions based on percentage of females in a county. From our previous kernel regression we can see that there likely is a significant relationship there that our model did not properly capture. This is why in my ideal counties I refrained from listing anything about gender parity.

Similar warnings should be given about the connections I draw from ethnicity, however the distinction here is so large that I think some speculation is justified. Minority communities at large rejected Sanders and his platform at very large margins. Looking more in depth at counties with Hispanic and Latino populations might yield interesting results about subgroups who might have supported Sanders. It is also relevant to note that all outliers in our model where Clinton win's 80% of the vote or more are all in counties with 60% of the population identifying as African American.

Conclusions and Sugestions for Future Study

Although Bernie Sanders' surprise success in the 2016 Democratic Primaries does seem to show on some level a desire by some Democrats for more progressive candidates, our analysis offers some evidence that those Democrats may make up a very specific subgroup. In our analysis higher minority populations were the strongest predictors for a Clinton win and low minority populations greatly favored Sanders, which may suggest that Sanders and his progressive platform were primarily accepted by white liberal Americans. It is interesting that larger African American populations correlated so highly with a Clinton win when traditional political analysis has shown African American populations to be very liberal. I believe such strong rejection of Sanders and his platform warrants further study about African Americans and more progressive politics. In counties with high Hispanic populations a Clinton win was assured, but at rates much lower than compared to those counties with high African American populations. This is an interesting trend because traditionally Hispanic populations lean conservative with a more liberal swing in recent years. Yet, in counties with majority Hispanic populations Clinton won by much smaller margins than those in counties with large African American populations. This may suggest that Sanders made better inroads with this community and would be an interesting case for further study.

Ultimately our study supports the need for a more comprehensive study into views towards more progressive politics based on ethnicity. As the Democratic Party and the United States as a whole become more ethnically diverse the need to understand shifts in political preference among ethnic lines will become more important for the Democratic Party to properly represent their base. However, this also opens up the questions of whether or not one party has the ability to represent the changing political beliefs of so many different groups when they are not trending towards the same political alignment.

Apendex

Model Output

```
##
## Call:
## lm(formula = Outcome ~ HSG495213 + EDU685213 + EDU635213 + PST120214 +
##     AGE775214 + RHI225214 + RHI725214 + INC910213 + SEX255214 +
##     LFE305213 + SBO315207 + RHI625214 + SBO215207 - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75697 -0.11033  0.00535  0.11912  0.99027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## HSG495213 -8.314e-07  8.456e-08  -9.832  < 2e-16 ***
## EDU685213 -5.854e-03  8.124e-04  -7.206 7.39e-13 ***
## EDU635213 -9.774e-03  6.801e-04 -14.371  < 2e-16 ***
## PST120214  8.975e-03  1.144e-03   7.844 6.19e-15 ***
## AGE775214  8.708e-03  9.723e-04   8.956  < 2e-16 ***
## RHI225214  1.556e-02  3.503e-04  44.402  < 2e-16 ***
## RHI725214  2.844e-03  2.770e-04  10.268  < 2e-16 ***
## INC910213  1.876e-05  1.288e-06  14.566  < 2e-16 ***
## SEX255214  7.351e-03  1.093e-03   6.723 2.15e-11 ***
## LFE305213  2.700e-03  7.478e-04   3.611 0.000311 ***
## SBO315207 -3.391e-03  6.676e-04  -5.079 4.04e-07 ***
## RHI625214 -3.006e-02  3.011e-03  -9.985  < 2e-16 ***
## SBO215207  9.512e-03  1.788e-03   5.319 1.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1826 on 2777 degrees of freedom
## Multiple R-squared:  0.6901, Adjusted R-squared:  0.6887
## F-statistic: 475.7 on 13 and 2777 DF,  p-value: < 2.2e-16
```

Note we removed the intercept and saw improvements in standard error and R^2 values.

Code Index

R Code

```
setwd("C:\\Users\\Cole\\Documents\\MAS\\MAS 404\\Project")
vote <- read.csv("2016_primary_data.csv")

vote$Outcome <- vote$HC_fraction_votes - vote$BS_fraction_votes

dim(vote)
vote <- vote[vote$HC_votes != 0,]
dim(vote)

hist(vote$Outcome)

par(mfrow=c(2,2))
hist(EDU685213, main=" Bachelor's degree or higher")
hist(EDU635213, main="High school graduate or higher")
hist(PST120214, main="Population, percent change")
hist(AGE775214, main="Persons 65 years and over, percent")
hist(INC910213, main="Per capita money income in past year")
hist(SEX255214, main="Female persons, percent")
hist(LFE305213, main="Mean travel time to work (minutes)")

library(car)

attach(vote)

mod1 <- lm(Outcome ~ HSG495213 +
             EDU685213 + EDU635213 + PST120214 + AGE775214
           + RHI225214 + RHI725214 + INC910213
           +SEX255214 + LFE305213 + SBO315207 + RHI625214 + SBO215207
           -1)

summary(mod1)

plot(mod1)

#Load in Kernal Density functions
path = "C:\\Users\\Cole\\Documents\\MAS\\MAS 404\\Project"
setwd(path)
system("R CMD SHLIB project404.c")

dyn.load("project404.dll")

nad_wat = function(n, X, Y, m, g2, res2, bw){
  ans = .C("nad_wat", as.integer(n), as.double(X), as.double(Y), as.integer(m), as.double(g2), as.double
           as.double(bw))
  ans[[6]]
}

kd = function(m, n, bw, g, x, y){
  kern_est <-.C("kd", as.integer(m), as.integer(n), as.double(bw), as.double(g),
```

```
                as.double(x), as.double(y))
  kern_est[[6]]
}

#Kernal Density of Outcome

bw_r = bw.nrd(Outcome)

x = seq(min(Outcome), max(Outcome), (max(Outcome) - min(Outcome))/99)

y = rep(0, 100)

yhat = kd(100,length(Outcome),bw_r,Outcome,x,y)

plot(x, yhat, ylab="Kernel Density Estimate", xlab="Outcome", type = 'l')
abline(v=0, col='red')
abline(v=mean(Outcome), col='blue')
abline(v=median(Outcome), col='green')
```

```
#Kernal Regresion African American Pop
X = RHI225214[Outcome < 1 ]
Y = Outcome[Outcome < 1 ]

bw = bw.nrd(X)
g2 = seq(min(X), max(X), (max(X) - min(X))/99)
res2 = rep(0,100)

gf = data.frame(X,Y)

kern_mat = matrix(nrow=200,ncol=100)

for(i in 1:200){
  samp = gf[sample(1:dim(gf)[1],100,replace = T),]
  kern_mat[i,] = nad_wat(48, samp$X, samp$Y, 100, g2, res2, bw)
}

upper_band = rep(0,100)
lower_band = rep(0,100)
for(i in 1:100){
  upper_band[i] = sort(kern_mat[,i])[5]
  lower_band[i] = sort(kern_mat[,i])[195]
}

plot(X,Y, xlab="Precentage of Population Identifying as African American", ylab="Outcome", col = 'gray')
lines(g2, nad_wat(dim(gf)[1], X, Y, 100, g2, res2, bw),col='red', lwd=2)
lines(g2, upper_band, lty = 'dotted',col='red', lwd=2)
lines(g2, lower_band, lty = 'dotted',col='red', lwd=2)
abline(h=0)


# Kernal Regression Female Pop
X = RHI225214[Outcome < 1 ]
Y = Outcome[Outcome < 1 ]
```

```
bw = bw.nrd(X)
g2 = seq(min(X), max(X), (max(X) - min(X))/99)
res2 = rep(0,100)

gf = data.frame(X,Y)

kern_mat = matrix(nrow=200,ncol=100)

for(i in 1:200){
  samp = gf[sample(1:dim(gf)[1],100,replace = T),]
  kern_mat[i,] = nad_wat(48, samp$X, samp$Y, 100, g2, res2, bw)
}

upper_band = rep(0,100)
lower_band = rep(0,100)
for(i in 1:100){
  upper_band[i] = sort(kern_mat[,i])[5]
  lower_band[i] = sort(kern_mat[,i])[195]
}

plot(X,Y, xlab="Precentage of Population Identifying as African American", ylab="Outcome", col = 'gray'
lines(g2, nad_wat(dim(gf)[1], X, Y, 100, g2, res2, bw),col='red', lwd=2)
lines(g2, upper_band, lty = 'dotted',col='red', lwd=2)
lines(g2, lower_band, lty = 'dotted',col='red', lwd=2)
abline(h=0)
```

C Code

```
#include <R.h>
#include <Rmath.h>

void kd(int *m, int *n, double *bw, double *g, double *x, double *y){
  for(int i=0; i < *m; i++){
    y[i] = 0;
    for(int j=0; j < *n; j++){
      y[i] += dnorm(g[j] - x[i],0, *bw, 0) / *n;
    }
  }
}

void nad_wat (int *n, double *x, double *y, int *m, double *g, double *est, double *b){
  int i,j;
  double a1,a2,c;
  for(i = 0; i < *m; i++){
    a1 = 0.0;
    a2 = 0.0;
    for(j=0; j < *n; j++){
        c = dnorm((x[j] - g[i])/ *b,0,1,0) / *b;
        a1 += y[j] * c;
        a2 += c;
    }
    est[i] = a1/a2;
```

```
  }
}
```