Final Project

Kevin Ayala(131) and Nathan Hwangbo (231)

December 12, 2018

1. What makes voter behavior prediction (and thus election forecasting) a hard problem? Voter behavior prediction is a hard problem for a couple reasons: First, we don't have clean or reliable data (or for that matter, a single source of truth). Polling data is the best we can use for predictors, which suffers from sampling error (since we can't poll the entire country). Additionally, the people being polled may or may not be telling the truth about their voting intentions, and even if they are being honest, their intentions may change over time (introducing a time-series element as well). Furthermore, this polling data comes from a variety of sources, and each source has its own agenda and biases. Second, election predictions are difficult because of how segmented the election process is. It isn't a simple majority rule vote, but rather a system split up into states. Then we have to deal with trying to mixing national data (eg a nationwide poll) with data specific to each state (eg state polls).

2. What was unique to Nate Silver's approach in 2012 that allowed him to achieve good predictions? First, Nate Silver's results were probabilistic in nature: meaning that he did not create a single best estimate for each of his parameters, but instead gave a range of probabilities for each. This method better incorporates the incertainty in the model (rather than just calculating the MLE and generating a single result, for example). Most predictions tend to offer a single number prediction, which can be misleading or lack information about the prediction. Second, Silver included a pollster's historical record to help weigh which polls were better than others. Rather than just using the most recent polling data, which is the most straightforwards approach, weighing the polls based on their historical reliability helps create reduce variation in the the sampled data, hence making better predictions.

3. What went wrong in 2016? What do you think should be done to make future predictions better? In 2016, the polling data was further off from the truth. In 2012, the polls overwhemlingly showed the truth (ie that Obama was going to win). The difference between the two candidates in the polls was large enough for the signal to overcome any sort of polling noise. However, in 2016, the election was close enough that the reults went the opposite direction of the polls. This means that the polling error was actually significant in this case! So estimating and trying to account for polling error was much more important in 2016 than in 2012. Furthermore, the candidates were very polarizing this year, causing an increase in media and analyst bias (eg many of the "academic-type" analysts might have thought that a Trump victory was impossible because ALL of the people in their circles were voting Clinton. This might have impacted their choices when creating models). In the future, we can hopefully use this election, where polling error ended up being significant, to update the weights assigned to each of the polls indicating their credibility. Additionally, we should be more careful about biases that might show up when creating models to predict an election.

county	fips	candidate	state	votes
Los Angeles County	6037	Hillary Clinton	CA	2464364
Los Angeles County	6037	Donald Trump	CA	769743
Los Angeles County	6037	Gary Johnson	CA	88968
Los Angeles County	6037	Jill Stein	CA	76465
Los Angeles County	6037	Gloria La Riva	CA	21993

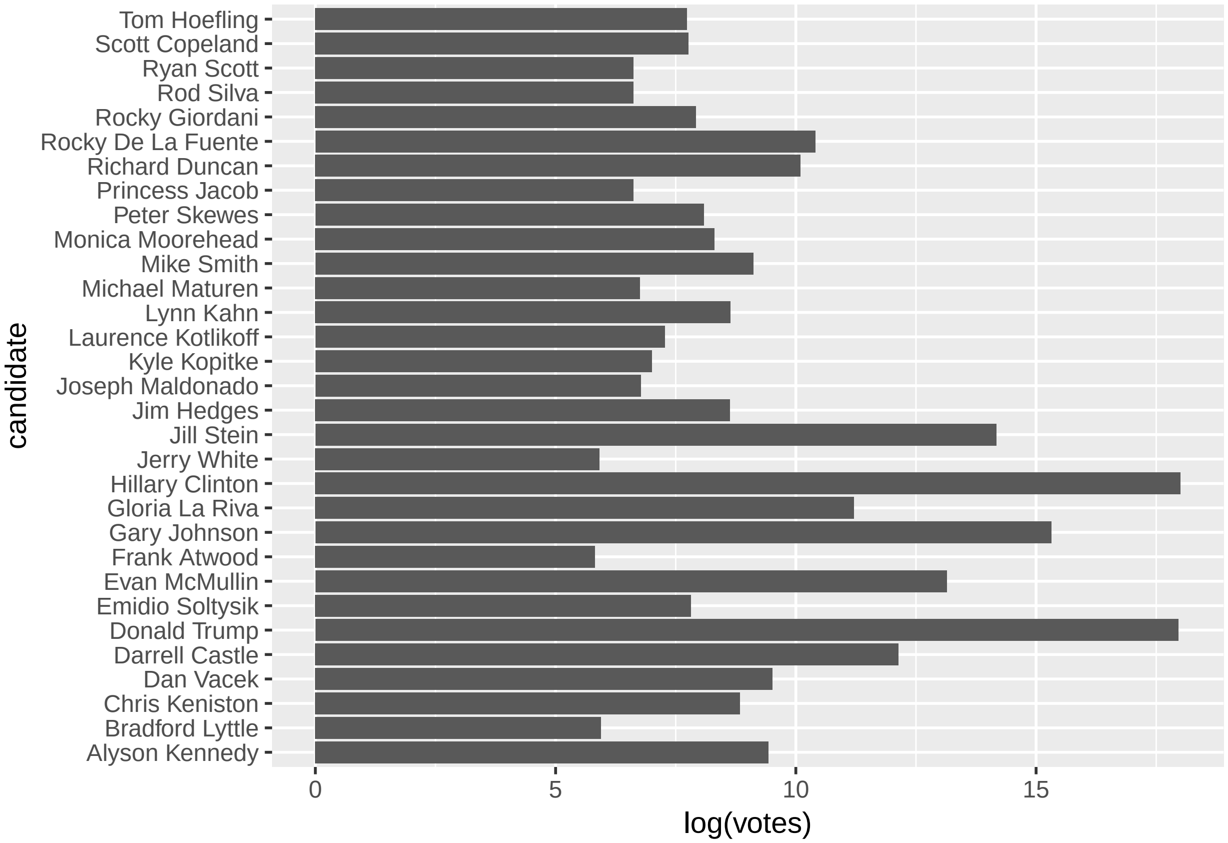
4. Report the dimension of `election.raw` after removing rows with `fips=2000`. Provide a reason for excluding them. Please make sure to use the same name `election.raw` before and after removing those observations

```
## The dimension of election.raw after removing rows with fips=2000 is  18345 5 .
##  the reason we do this is because these are duplicate observations
## (this can be clearly seen by looking at all observations with state='AK'
```

5. Remove summary rows from `election.raw` data:

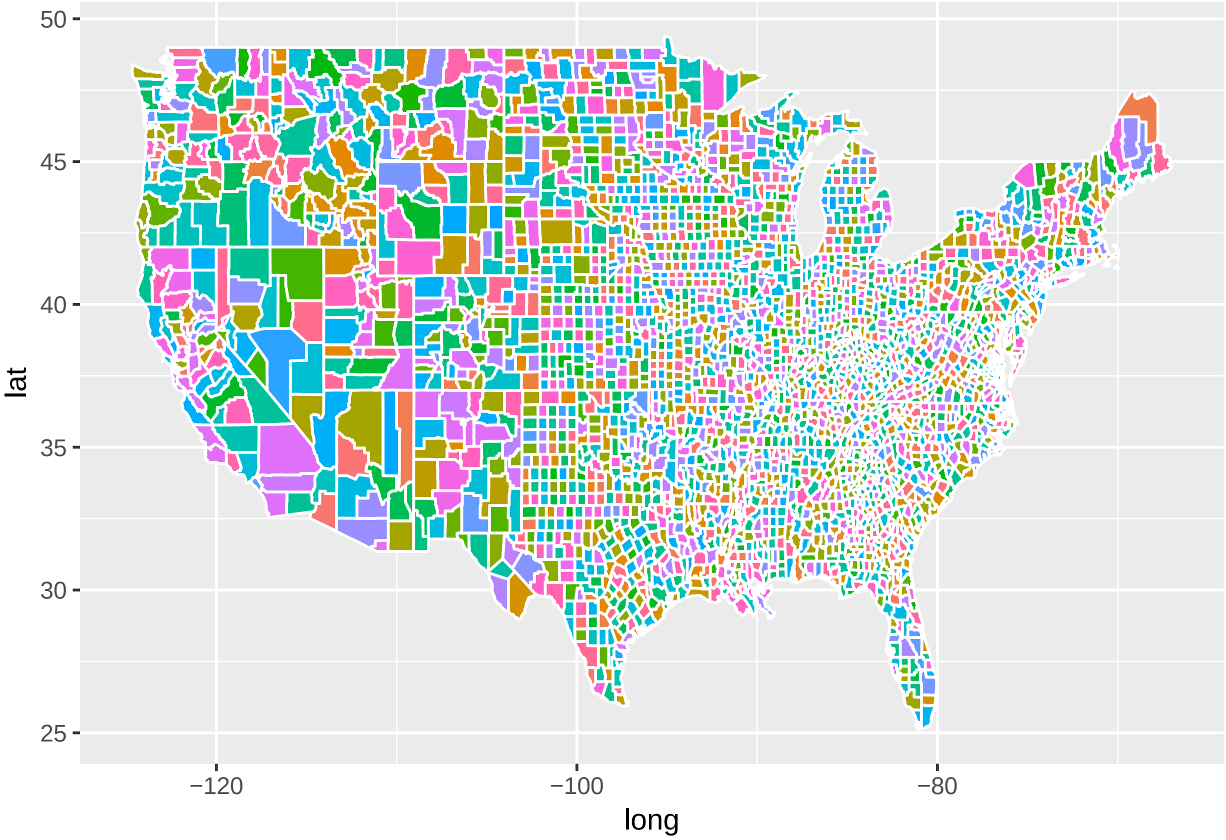
How many named presidential candidates were there in the 2016 election? Draw a bar chart of all votes received by each candidate. You can split this into multiple plots or may prefer to plot the results on a log scale. Either way, the results should be clear and legible!

There were 31 named presidential candidates in the 2016 election



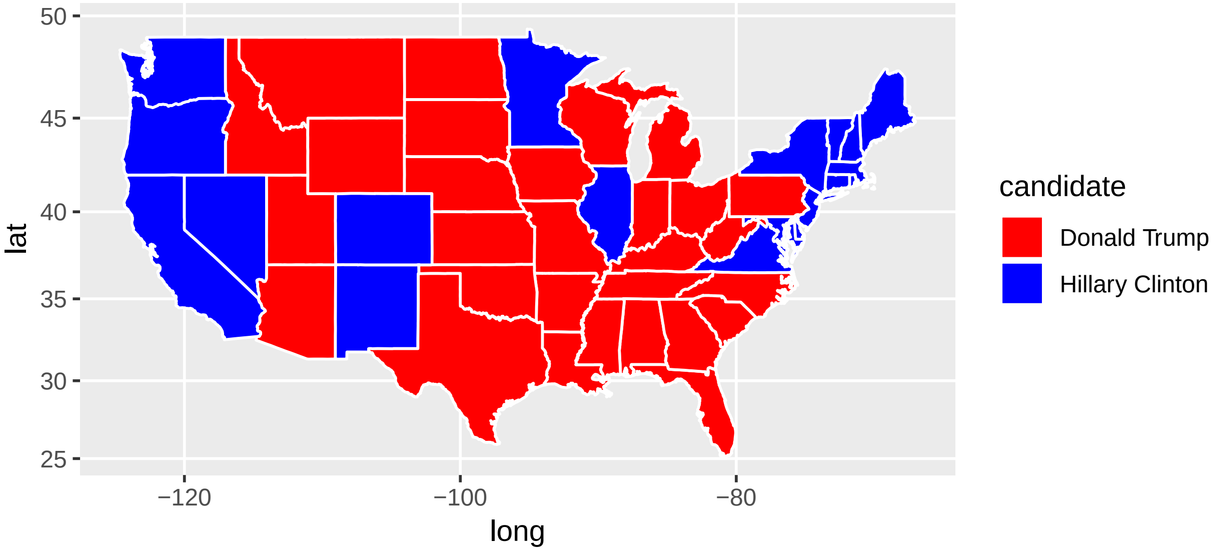
7. Create variables `county_winner` and `state_winner` by taking the candidate with the highest proportion of votes

8. Draw a county-level map by creating 'counties = map_data("county"). Color by county



9. Now color the map by the winning candidate for each state

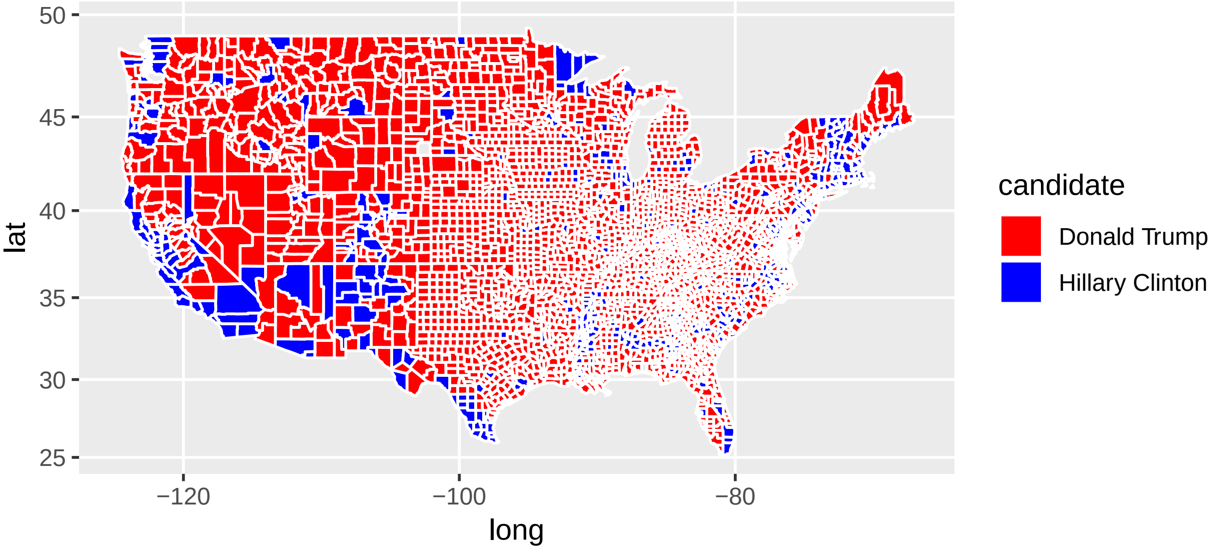
```
## Joining, by = "fips"
```



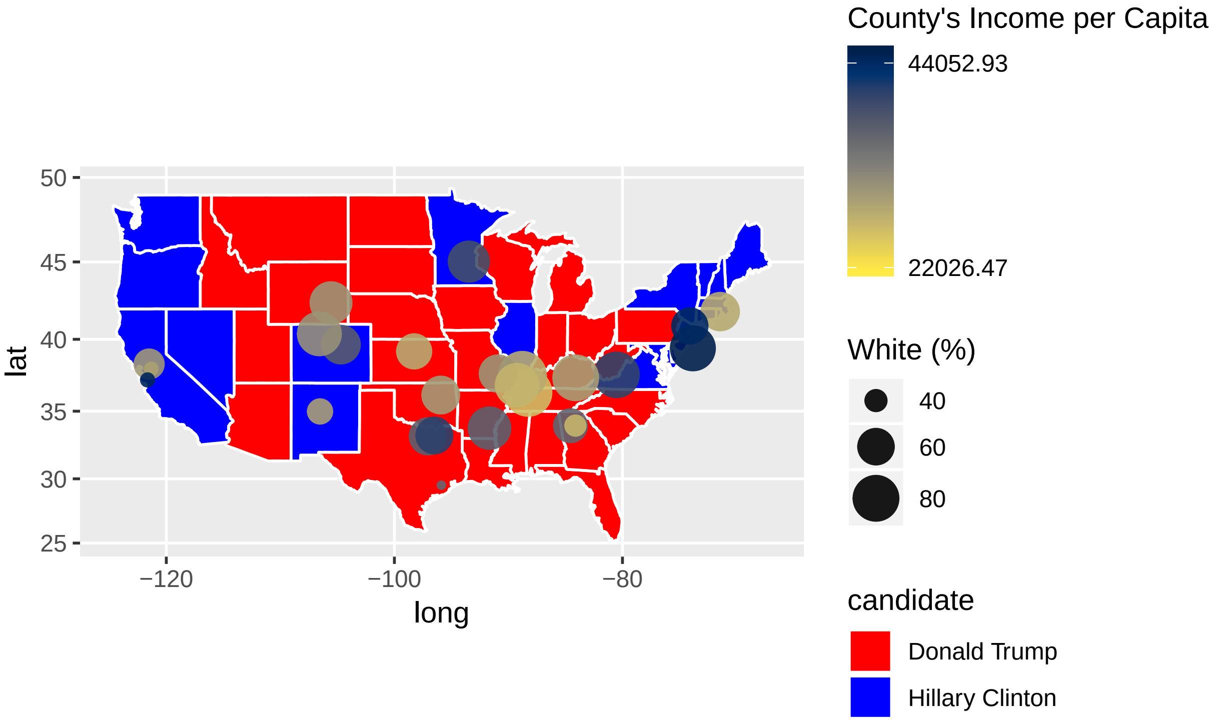
10. The variable county does not have the fips column. So we will create one by pooling information from `maps::county.fips`

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:Matrix':
```

```
##
##     expand
## The following object is masked from 'package:class':
##
##     condense
## The following object is masked from 'package:dplyr':
##
##     rename
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
## Joining, by = c("region", "subregion")
## Joining, by = "fips"
```



11. Create a visualization of your choice using census data. Many exit polls noted that demographics played a huge role in the elction.



This plot has two components. The first is background colors, which show the election results in 2016 (same as the plot in question 9). The second component is the bubbles, which show three different pieces of information: There are 25 bubbles plotted, which are the 25 largest counties, by population. Each bubble has a unique color and size: the color corresponds to that county's average income per capita, and size of the bubble corresponds to the percentage of the population that is White. Our motivation for this vizualization was the media portrayal of the election. The narrative being told was that Trump supporters were mostly wealthy white Americans in rural areas, and Clinton supporters mostly lived in urban areas, were more educated, and tended to be poorer. We wondered what happened in the intersection between these two sets: for example, we wanted to know whether wealthy white people in urban areas tended to vote Trump (because they are wealthy and white) or Clinton (because they live in an urban area). This plot shows that areas like this could go either way. Some of the East Coast plots and Colorado show that a state could go Democrat, having a high percentage of white people and a high income per capita. However, Texas and the Midwest have points that show similar a similar demographic voting Republican.

12. The census data contains high resolution information (more fine-grained than county-level). In this problem , we aggregate the inforamtion into county-level data by computing `TotalPop`-weighted average of each attributes for each county. Create `census.del`, `census.subct`, `census.ct`. Print the first few rows of census.ct

```
##     State  County       Men    White   Citizen   Income IncomeErr
## 1 Alabama Autauga 0.4843266 75.78823 0.7374912 51696.29  7771.009
## 2 Alabama Baldwin 0.4884866 83.10262 0.7569406 51074.36  8745.050
## 3 Alabama Barbour 0.5382816 46.23159 0.7691222 32959.30  6031.065
##   IncomePerCap IncomePerCapErr  Poverty ChildPoverty Professional  Service
## 1     24974.50        3433.674 12.91231     18.70758     32.79097 17.17044
## 2     27316.84        3803.718 13.42423     19.48431     32.72994 17.95092
## 3     16824.22        2430.189 26.50563     43.55962     26.12404 16.46343
##     Office Production    Drive   Carpool    Transit OtherTransp WorkAtHome
## 1 24.28243   17.15713 87.50624  8.781235 0.09525905    1.305969   1.835653
## 2 27.10439   11.32186 84.59861  8.959078 0.12662092    1.443800   3.850477
## 3 23.27878   23.31741 83.33021 11.056609 0.49540324    1.621725   1.501946
##   MeanCommute  Employed PrivateWork SelfEmployed FamilyWork Unemployment
## 1    26.50016 0.4343637    73.73649     5.433254 0.00000000     7.733726
## 2    26.32218 0.4405113    81.28266     5.909353 0.36332686     7.589820
## 3    24.51828 0.3192113    71.59426     7.149837 0.08977425    17.525557
##   Minority
## 1 22.53687
## 2 15.21426
## 3 51.94382
```

13. Run PCA for both county and subcounty level data. Save the first 2 PCs into a two-column data frame, call it `ct.pc` and `subct.pc`, respectively. Discuss whether you chose to center and scale the features before running PCA and the reasons for your choice. What are the three features with the largest absolute values of the first principal component? Which features have opposite signs and what does that mean about the correlation between these features?

Because our data is NOT all in the same units or in the same scale (eg Male is a percentage and Income is an average), we choose to both scale and center our data. (income doesn't have a mean of 0, so we want to center as well)

```
## The top three county features with largest absolute values are:
##   IncomePerCap ChildPoverty Poverty
```

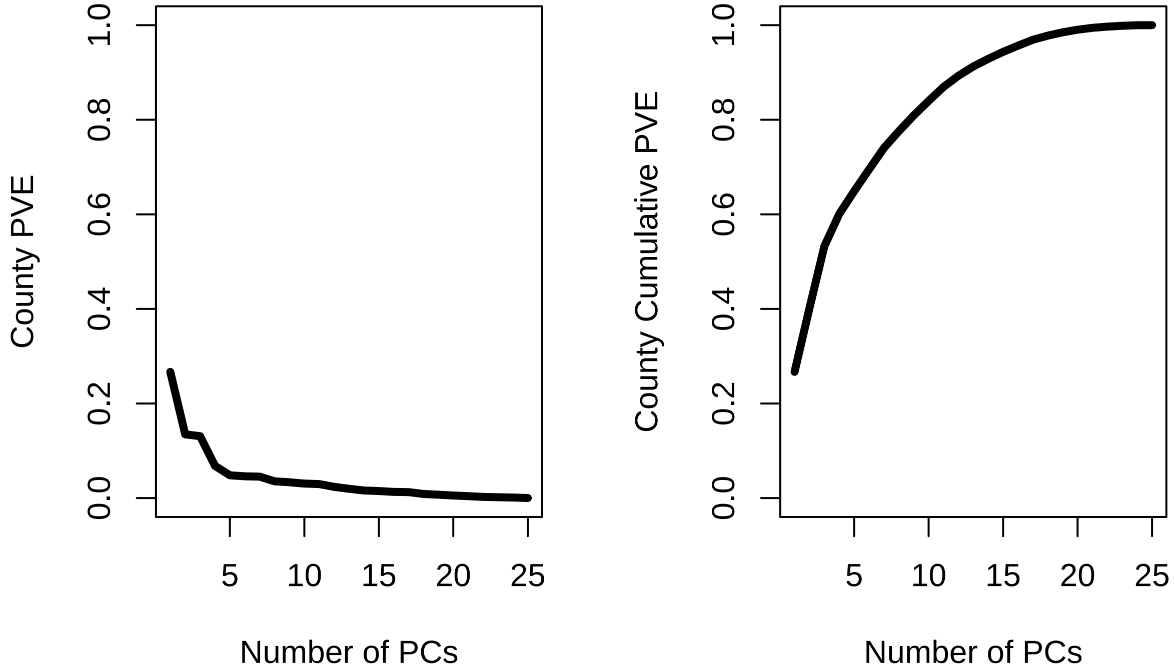
```
## The top three subcounty features with largest absolute values are:
##   IncomePerCap Professional Poverty
```

```
## The county features with negative loadings are:
##  Poverty ChildPoverty Service Office Production Drive Carpool OtherTransp MeanCommute Unemployment Mi
## and the features with postiive loadings are:
##  Men White Citizen Income IncomeErr IncomePerCap IncomePerCapErr Professional Transit WorkAtHome Empl
##  This means that along the PC1 axis,
##  these features are 'opposites' of each other.
## Then we might expect some negative correlation between the two sets.
```

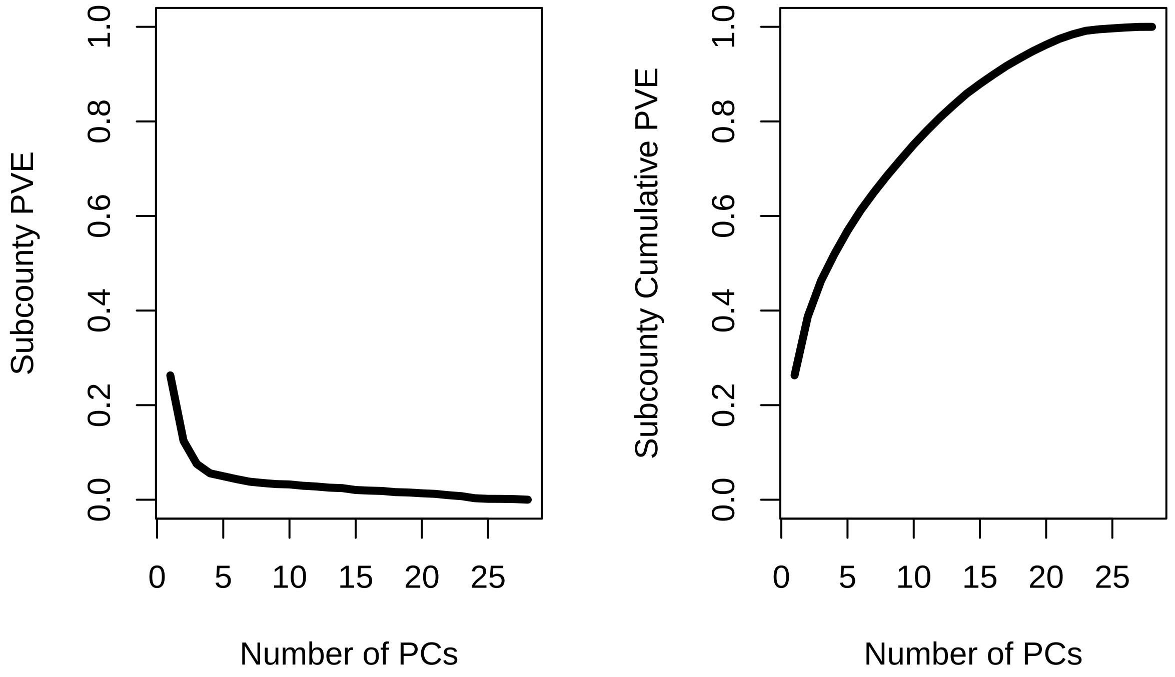
```
##
## The Subcountyfeatures with negative loadings are:
##  TotalPop Men White Citizen Income IncomeErr IncomePerCap IncomePerCapErr Professional Drive WorkAtHo
## and the features with postiive loadings are:
##   Poverty ChildPoverty Service Office Production Carpool Transit OtherTransp PrivateWork Unemploymen
##  and we apply the same reasoning above about the correlation between the two sets
```

14. Determine the number of minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses. Plot PVE and cumulative PVE for both county and subcounty analyses

```
## For the county data, the minimum number of PCs needed to capture 90% of the variance is : 13
```



```
## For the subcounty data, the minimum number of PCs needed to capture 90% of the variance is : 17
```



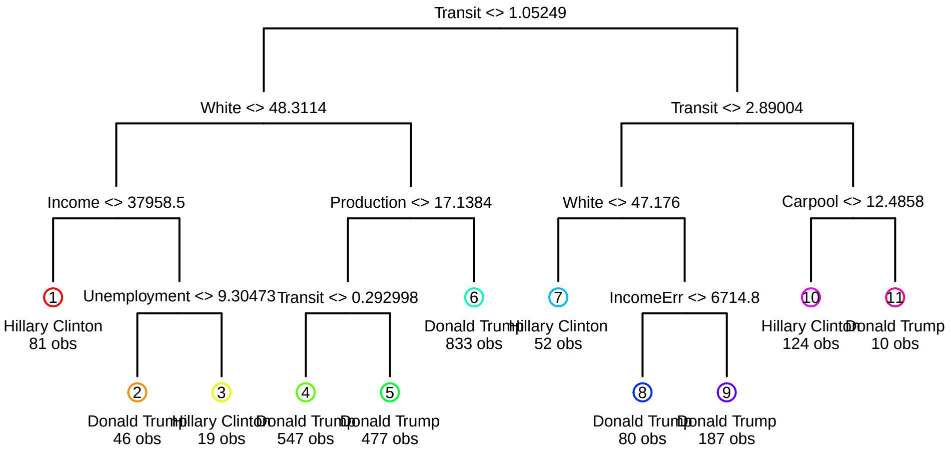
15. With `census.ct`, perform hierarchical clustering with complete linkage.

```
## We can examine the state breakdown of the data in San Mateo's cluster:
```

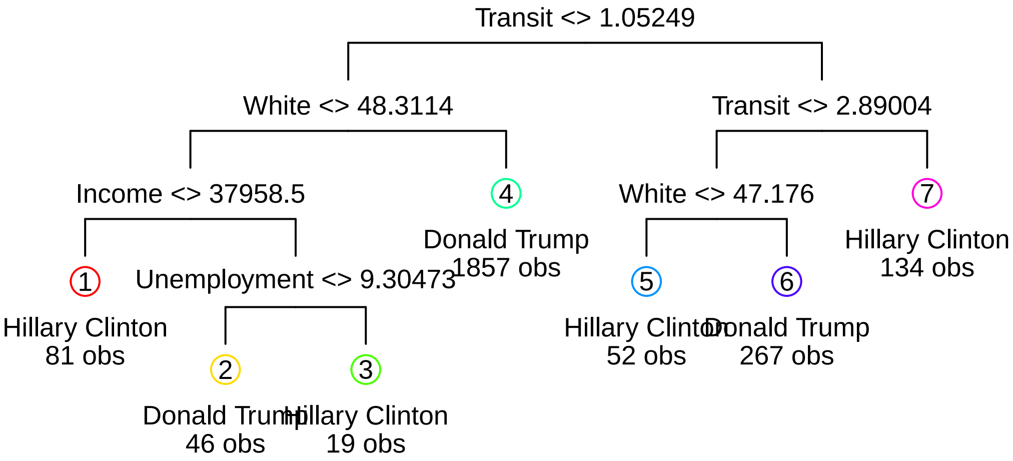
```
##
##    California      Colorado   Connecticut      Maryland Massachusetts
##             2             1             1             3             2
##    New Jersey    New Mexico      New York          Ohio  Pennsylvania
##             5             1             3             1             1
##     Tennessee          Utah      Virginia
##             1             1             4

## We compare the row of San Mateo County in the data with the averaged row
##  across all the data in the cluster.
##  San Mateo is the first row, and the average is the second.

##           Men    White   Citizen   Income IncomeErr IncomePerCap
## 227 0.4919773 40.63851 0.6420050 100369.9  16123.02     47881.29
## 2   0.4921213 67.74808 0.6870218 100254.1  14147.59     45273.18
##     IncomePerCapErr  Poverty ChildPoverty Professional  Service   Office
## 227        6115.552 8.011122     9.705514     45.73565 18.28979 22.30400
## 2          5544.313 6.650767     7.977504     49.24406 14.97435 22.71257
##     Production    Drive   Carpool  Transit OtherTransp WorkAtHome
## 227   7.343290 69.92713 10.681436 9.257082    2.598808   5.077957
## 2     6.441967 74.23085  8.366556 7.780258    1.511996   5.764982
##     MeanCommute  Employed PrivateWork SelfEmployed FamilyWork Unemployment
## 227    26.82681 0.5172497    79.76635     8.367532  0.1716192     6.689483
## 2      31.06520 0.5138841    77.59147     5.816113  0.1346526     5.911511
##     Minority
## 227 55.53405
## 2   29.79726

##
##  Similarly, we examine the State breakdown for the data using the first two PCs to cluster

##
##    California      Colorado   Connecticut       Georgia       Indiana
##             4             5             2             1             2
##          Iowa        Kansas      Kentucky      Maryland Massachusetts
##             1             1             1             6             3
##     Minnesota New Hampshire    New Jersey    New Mexico      New York
##             2             1             5             1             4
##  North Dakota          Ohio  Pennsylvania  Rhode Island  South Dakota
##             3             1             2             1             2
##     Tennessee         Texas          Utah      Virginia     Wisconsin
##             1             3             1             9             1
##       Wyoming
##             1

##
## We also compute the average of each column for this cluster
##  and compare it with San Mateo.
##  Again, San Mateo is the first row.

##           Men    White   Citizen   Income IncomeErr IncomePerCap
## 227 0.4919773 40.63851 0.6420050 100369.9  16123.02     47881.29
## 2   0.4985973 75.76718 0.7087768  91117.5  13499.99     44329.97
##     IncomePerCapErr  Poverty ChildPoverty Professional  Service   Office
## 227        6115.552 8.011122     9.705514     45.73565 18.28979 22.30400
## 2          6114.573 6.890861     7.923726     48.18768 14.52754 21.95869
##     Production    Drive   Carpool  Transit OtherTransp WorkAtHome
## 227   7.343290 69.92713 10.681436 9.257082    2.598808   5.077957
## 2     7.167475 74.02807  7.649492 6.786605    1.682261   6.569067
##     MeanCommute  Employed PrivateWork SelfEmployed FamilyWork Unemployment
## 227    26.82681 0.5172497    79.76635     8.367532  0.1716192     6.689483
## 2      27.19373 0.5281399    77.49696     7.302827  0.1754773     5.006901
##     Minority
## 227 55.53405
## 2   22.06209

## the sum of the absolute percentage differences between san mateo and the full data was:  4.039064
##  This is lower than the number we got for the PC data, which was  4.256098
##  This means that if we took the percent differences between San Mateo's 'gender' column and the clus
## did the same for every column, and added all those numbers up,
## we would get the numbers above. Then the total error for the full dataset performed better,
## placing San Mateo County in the more appropriate cluster.
##  We expect this result, since using the first two principal components to form the tree implies that
## (we can think about each principal component as forming 'topics' from linear combinations of our var
## Then the tree for the PCs misses differences elsewhere in the data, that the full dataset picks up o
```

16. Decision tree: train a decision tree by `cv.tree()`. Prune tree to minimize misclassification error. Be sure to use the `folds` from above for cross validation.

```
## First, we draw the tree before pruning
```



```
## Now, we draw the tree after pruning
```



These trees tell us that this election was largely split on geography, race, and income. The first split is on transit, and tells us that people that don't use public transit tended to vote more for Trump. A possible explanation for this is that Trump's base is largely in rural areas and with wealthy people. Both of these groups are likely to drive private transportation (rural because areas are so far apart, wealthy because they have plenty of money to spend on cars) The left side of the tree agrees with this explanation, as it shows that white (rural communities are mostly white) voters with high income are more likely to vote Trump. \ On the other hand, we can take a look at Clinton's base on the right branch. The first tells us that frequent use of public transportation is an indicator of a Clinton supporter. This is mostly in urban cities (where public transportation is good) or within poorer communities (where private transportation is not affordable). These two demographics are highlighted as we go down the tree, since heavy transit users, minorities (more likely to live in uran areas), and low income voters were more likely to vote Clinton.

17. Run a logistic regression to predict the winning candidate in each county.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## [1] "First, we look at the coefficients of our model"
```

```
##
## Call:
## glm(formula = candidate ~ ., family = binomial(link = "logit"),
##     data = trn.cl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8088  -0.2685  -0.1124  -0.0400   3.5943
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.158e+01  9.734e+00  -1.189 0.234280
## Men              8.585e+00  5.406e+00   1.588 0.112305
## White           -2.217e-01  6.478e-02  -3.422 0.000623 ***
## Citizen          1.019e+01  3.010e+00   3.384 0.000715 ***
## Income          -7.579e-05  2.750e-05  -2.756 0.005846 **
## IncomeErr       -3.877e-05  6.283e-05  -0.617 0.537200
## IncomePerCap     2.676e-04  6.728e-05   3.978 6.95e-05 ***
## IncomePerCapErr -2.785e-04  1.309e-04  -2.127 0.033415 *
## Poverty          1.990e-02  4.119e-02   0.483 0.629007
## ChildPoverty    -7.409e-03  2.560e-02  -0.289 0.772289
## Professional     2.817e-01  3.905e-02   7.214 5.44e-13 ***
## Service          3.659e-01  4.923e-02   7.432 1.07e-13 ***
## Office           1.051e-01  4.714e-02   2.230 0.025754 *
## Production       1.845e-01  4.317e-02   4.273 1.93e-05 ***
## Drive           -2.562e-01  5.424e-02  -4.724 2.32e-06 ***
## Carpool         -2.463e-01  6.714e-02  -3.668 0.000244 ***
## Transit         -8.605e-03  1.015e-01  -0.085 0.932442
## OtherTransp     -9.699e-02  1.014e-01  -0.956 0.339050
## WorkAtHome      -2.101e-01  7.961e-02  -2.639 0.008312 **
## MeanCommute      6.326e-02  2.482e-02   2.549 0.010813 *
## Employed         1.653e+01  3.292e+00   5.021 5.14e-07 ***
## PrivateWork      9.253e-02  2.149e-02   4.306 1.66e-05 ***
## SelfEmployed     1.226e-02  4.668e-02   0.263 0.792768
## FamilyWork      -1.191e+00  4.099e-01  -2.907 0.003655 **
## Unemployment     1.838e-01  3.810e-02   4.824 1.41e-06 ***
## Minority        -8.920e-02  6.214e-02  -1.435 0.151153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2074.96  on 2455  degrees of freedom
## Residual deviance:  854.06  on 2430  degrees of freedom
## AIC: 906.06
```

```
##
## Number of Fisher Scoring iterations: 7
```

```
##
##
## The five most significant variables in our model, in order, are :
##  Service Professional Employed Unemployment Drive
##  This lines up pretty well with the decision tree.
##  We notice that employment, method of transportation, and race were important issues in this race.
##  However, this adds more detail, with variables like the type of job (eg service, professional) matte
```

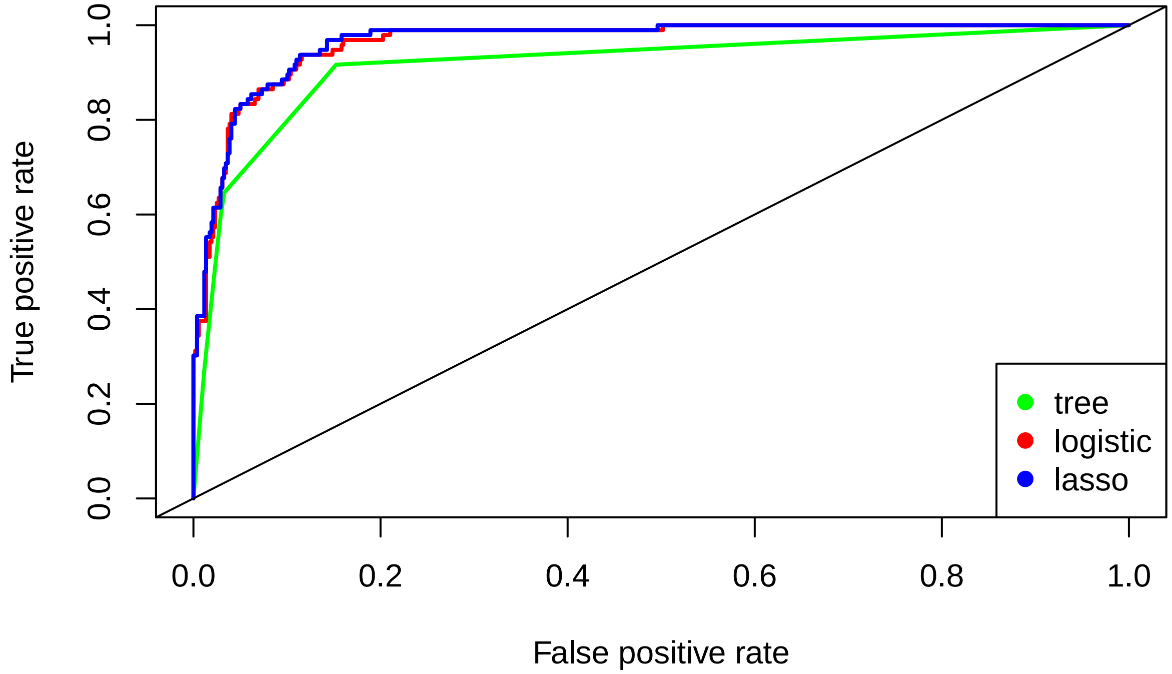
```
##
## For Service (the percentage of people working a service job),
##  the most important variable in our model, the estimated coefficient is  0.365875
##  This means that a unit increase in the Service variable creates a .365875 increase in the log odds o
##  ie her odds improve multiplicatively by e^.365 = 1.4405
##
##
##
## We can do the same for the second most important var, Professional
##  (the percentage of people working in a 'professional' job.
## The estimated coeff is:  0.2816889
##
## This means that a unit increase in the Professional variable leads to a .282 increase in the log odds
## ie her odds improve multiplicatively by e^.282 = 1.326
```

18. Control overfitting using regularization

```
## The optimal value of lambda in our cross validation is:  5e-04
```

```
## The only predictors with a coefficient of 0 are ChildPoverty, Self-Employed, and Minority.
## This is not surprising when compared to the unpenalized logistic regression model,
## since those three were the categories with the highest (ie least significant) p values.
```

19. Compute ROC curves for the decision trees, logistic regression, and LASSO logistic regression using predictions on the test data



The pros of the decision tree model is overwhemingly its interpretability. The model is extremely easy to interpret, and even easier to see visually. The ROC curve shows that its predictions might not be as good, but if we are trying to explain the results in an easy to understand model, the tree is the best way to go. This method might be better for answering questions in a post-mortem analysis, like what we are doing with this project with the 2016 election. It can help us group variables together into "topics", and understand the biggest splits in the data. But for doing predictions on new data, this will not yield the best results.

The pros of the unpenalized logistic regression is in its prediction power. It tended to make very good predictions for our data, so where precision matters, we should use logistic regression. It also has the benefit of being a soft classifer, which makes it easier to see how "close" some ofthe splits are in our data. This gives it an edge over the decision tree in problems where we can easier tweak our model (our p_thresh) to change the results without having huge changes to our entire model like we might have in the decision tree. However, it suffers from interpretability, as increasing "log odds" is hard to think about, and would be even harder to explain. Because there are so many variables, it also makes it hard to see variables interact. This would probably the best model for actually makinng predictions on a new dataset.

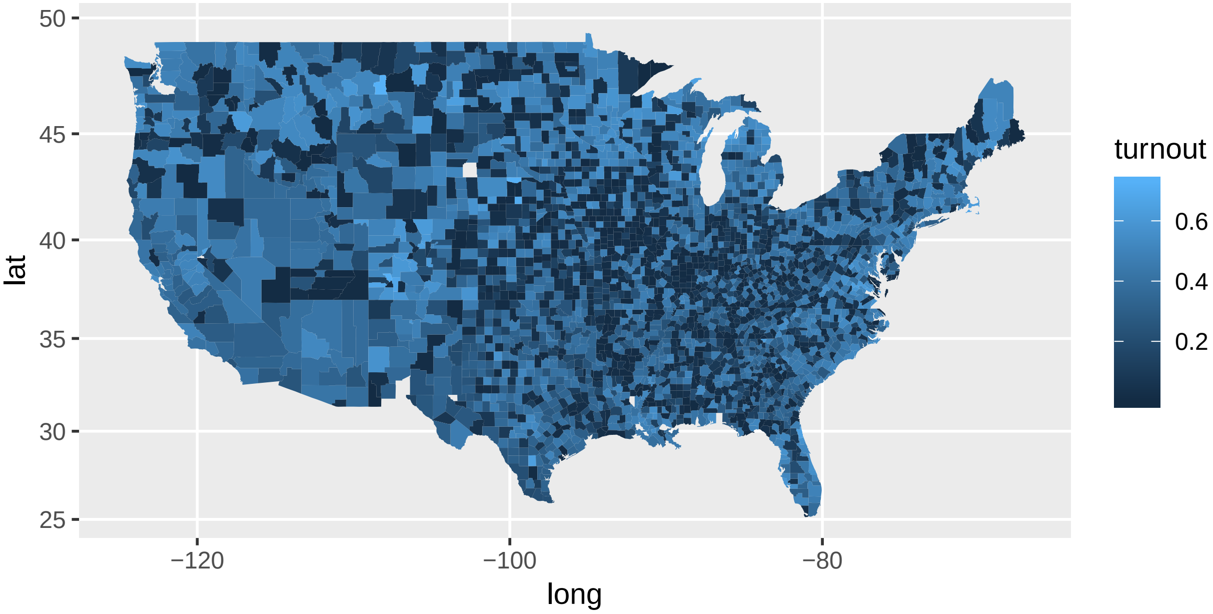
The pros of lasso regression is that it is right in the middle of the two methods. The prediction power (as you can see in the ROC curve) are VERY similar to the unpenalized logistic regression, but we also were able to remove three variables from the model, making it much easier for us to tell which of the variables are significant, or whether some are redundant (ie we found that Minority was insiginficant but White was significant, due to redundancy). So it was slightly easier to make sense of than the unpenalized regression, but it still suffers from most of the same interpretability issues as the logistic regression. So compared to the decision tree, it is very difficult to interpret still. Since this model is not better at prediction than logistic regression, and is not as interpretable as the decision tree, this method is probably best for us to answer questions about variable importance, and should be used when we need to see if we can reduce our data to get rid of redundancies.

20. This is an open question. Interpret and discuss any overall insights gained in this analysis and possible explanations. Use any tools at your disposal to make your case: visualize errors on the map, discuss what does/doesn't seems reasonable based on your understanding of these methods, propose possible directions (collecting additional data, domain knowledge, etc). In addition, propose and tackle *at least* one more interesting question. Creative and thoughtful analyses will be rewarded! *This part will be worth up to a 20% of your final project grade!*

This analysis mostly showed demographics that most clearly split the election. We focused on trying to figure out which counties tended to vote most similar to each other, and what demographics (%male, %white, income, ect) were most likely to predict the voting behavior in a county. What all of this analysis misses out on, though, is the question of WHO voted. We want to see if there are groups of people more likely to vote, and if so, we want to see if we can hypothesize how these groups were able to get a higher turnout. Most importantly, since the US election is through the electoral college (and hence by state), we want to know if certain regions had higher voter turnout than others. For this preliminary exploration, we will be using a simple linear regression model, trying to predict the voting percentage (Total Votes / Total Population) of each county.

We start by plotting the voter turnout as a gradient across all the counties

```
## Joining, by = "County"
```



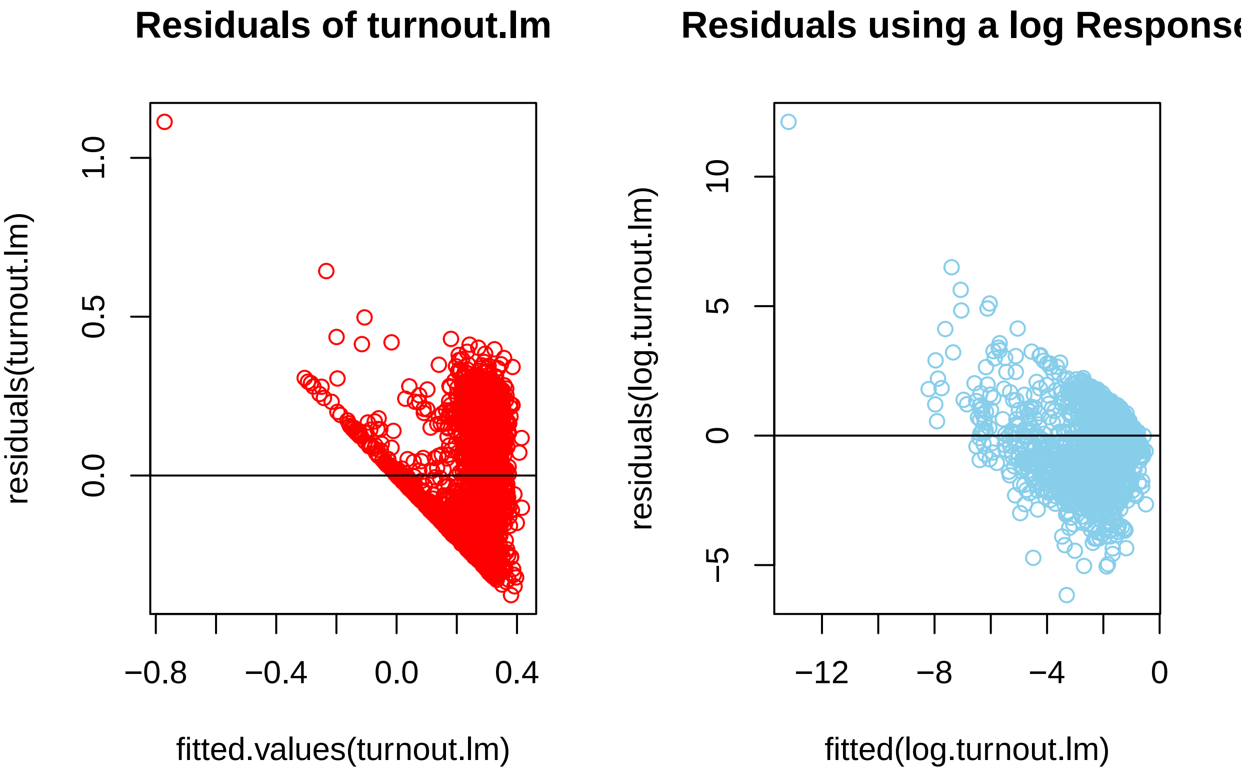
Interestingly, we notice that the midwest tends to have a much lower voter turnout than the coasts. Donald Trump tended to win these areas during the election.
We also note that Montana, Colorado, and Wisconson all had very high average voter turnouts.

Now we turn our attention to trying to create a simple linear model, to see whether gender, employment, race, or total population had a large impact on the voter turnout rate

```
##
## Call:
## lm(formula = turnout ~ Men + Minority + Employed + TotalPop,
##     data = election.20)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3763 -0.1620  0.0085  0.1492  1.1133
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.733e-01  7.998e-02   2.167   0.0303 *
## Men         -3.100e-01  1.418e-01  -2.187   0.0288 *
## Minority     1.301e-03  1.728e-04   7.528 6.72e-14 ***
## Employed     6.005e-01  5.347e-02  11.231  < 2e-16 ***
## TotalPop    -1.162e-07  4.107e-09 -28.299  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.174 on 3065 degrees of freedom
## Multiple R-squared:  0.2272, Adjusted R-squared:  0.2262
## F-statistic: 225.3 on 4 and 3065 DF,  p-value: < 2.2e-16
```

Interestingly, we notice that all four of these factors tend to be pretty good predictors for predicting turnout in this very crude model. This model also suggests that gender was the least important of these predictors. This is interesting, especially in light of the fact that Hillary Clinton was running to be the first female President. According to this model, a unit increase in the percentage of men in a county actually decreases voter turnout, which means that women tended to vote more in this election, as we might expect. Also interesting is that an increase in minority % increases the turnout (perhaps because Trump's immigration proposals were so unorthodox), and that large cities tended to have a smaller voter turnout percentage. Maybe this is just due to there being a large percentage of the population of large counties who are inelligable to vote. More research would have to be done to determine the reasoning.

Of course, with the variables all being percentages, and strictly positive, we do not expect that the assumptions of a linear model are very good. In fact, the residual plots below show that they are quite badly broken. We try to fit a model using the log response as well, but that hardly does anything to help us fulfill our assumptions. But this was a quick, preliminary analysis, and the effects were strong enough to come away with a very weak conclusion. Further research should be done with better fitting models to go in and see how our findings hold.



Prediciting the outcome of an election is a very hard thing to do indeed. Why is this the case? Polling data in 2016 indicated that Hillary Clinton was projected to win the election, and many were trumped the night "The Donald" gave his victory speech. How were the polls so wrong? Frequentist statistics tells us that if we continued to sample, ie. polling, then by law of large numbers it was expected that Hillary Clinton would win since the polls indicated as such. However, this is very hard to observe in practice, especially when the population is just extremely large; a population of about 327 million is big indeed. One thing that was extremley overlooked was biasing. We know that polling data was conducted on large metropolis cities where populations were over one million, where people are easy to communicate with via cellular methods or in person.

We have many forms of bias, such as nonresponse biasing and forms of bias where potential voters would be too embarressed to reveal who they really supported. 2016 was the year where the “establishment” of politics was challenged by those outside the “establishment”. Many people felt that President Obama did not do enough in easing the transition of coming out of the great recession.

Here, we will use bias as our advantage and use a Bayesian mentality to explore and predict counties, based on who we think would vote for Trump. The one thing about earlier analysis and classification in regards to prediction, is that training data is needed, which implictly implies that the election has had to already happened. In prediction a real election, we do not have access to the data and we must rely on polling data, which can be biased. Bias is not a bad thing as it has a negative connotation, and as we have seen, can usually offset varaince known in the Bias-Variance tradeoff.

We start by having a prior beleif, we know that Trump was polling well with voters outside of metropolin areas in 2016. We know Trump was having high voter turnout in his primaries, just seeing the riots started by his supporters on media sources during primaries further makes us beleive this. The average county population is 100,000 and we beleive that people who have not found relief from the great recession still lie within populations where the county population is less than the average county population of 100,000. We feel that these people are the ones who blame the establishment for failure of properly handling the recession which has led to thier current frustrations, and are looking to blame someone or something.

On June 16, 2015, Donald Trump descended the stairs of his tower to announce to the world that he was putting in his bid for the 2016 presidential election. He made his declaration to the world by saying the following words, in which is now recognizable anywhere in todays world and has even become a meme, an idea popularized by the internet.

“Our country is in serious trouble. We don’t have victories anymore. We used to have victories, but we don’t have them. When was the last time anybody saw us beating, let’s say, China in a trade deal? They kill us. I beat China all the time. All the time.

When did we beat Japan at anything? They send their cars over by the millions, and what do we do? When was the last time you saw a Chevrolet in Tokyo? It doesn’t exist, folks. They beat us all the time.

When do we beat Mexico at the border? They’re laughing at us, at our stupidity. And now they are beating us economically. They are not our friend, believe me. But they’re killing us economically."

but most recognizable of all,

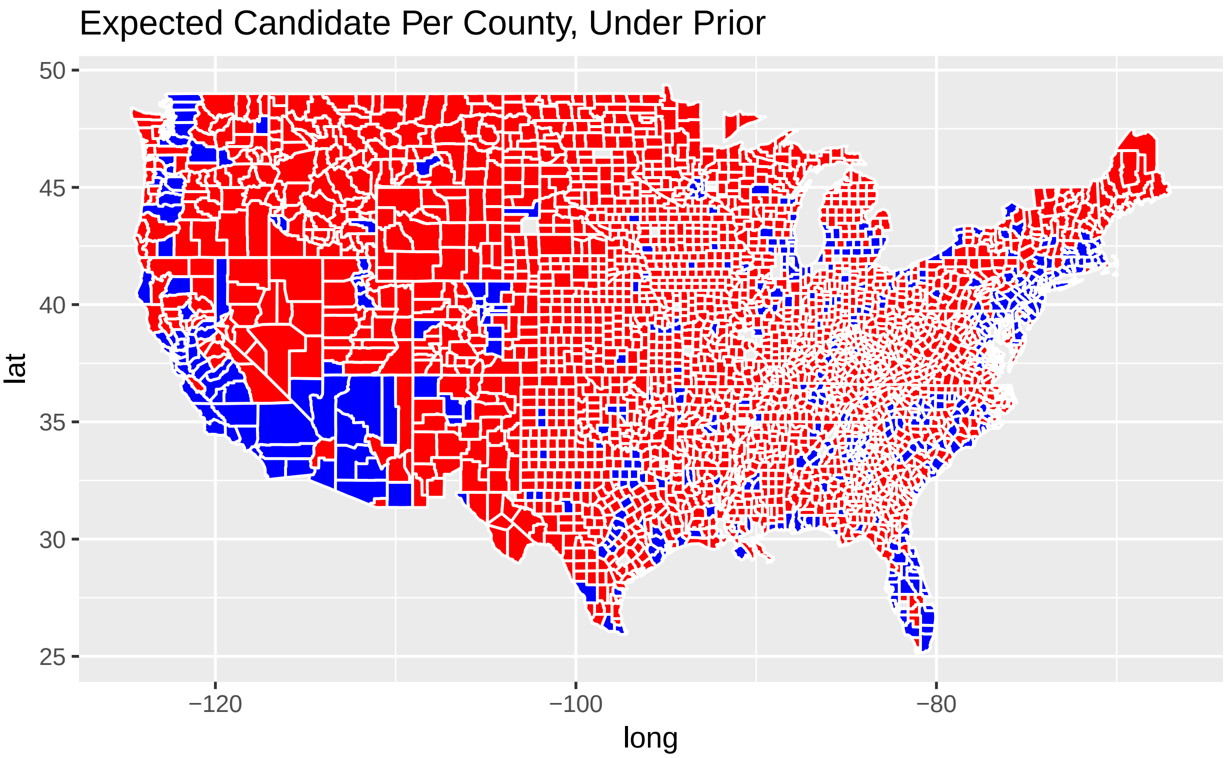
“They’re bringing drugs. They’re bringing crime. They’re rapists. And some, I assume, are good people”

It is impossible to read this last quote without hearing the voice of “The Donald”

Donald Trump effectively targeted minorities and immigrants, blaming then from taking away jobs. We make a prior assumption that counties where the population has less than 10 percent minority population, has a Trump advantage of winning that county since there is not enough minorites to possibly offset angry and frustrated Trump supporters.

We looked at the census data and thus manipulated it by creating the expected candidate to win in each county. We based our expectation on the following; if the population of the county was less than 100,000 or if the minority population was less than ten percent, then it was considered a Trump win. The following US map visualization, shows the expected candidate to win by county with this criteria,

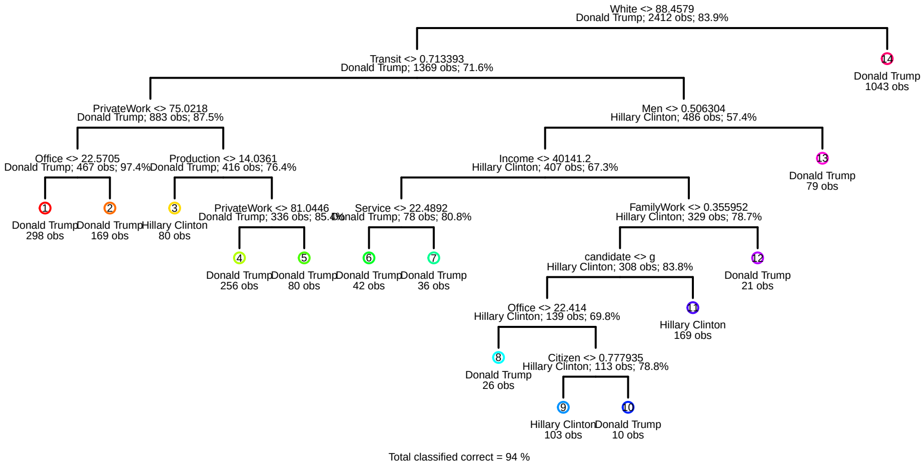
```
## Joining, by = c("State", "County")
## Joining, by = c("State", "County")
```



Considering we did this based on our population size and minority population, this was already a great start, as this was done using census data only with no regard towards the voting outcome data, guided only by our beleifs. We noticed that the places where the population is large for that county such as Los Angeles, Maimi,Las Vegas, and New York ect, went for Clinton. This was in line where polling was most frequent during 2015-2016. People are extremely easy to reach and thus poll. We noticed that places like Kansas, go to Donald Trump due to low relative population and small minority percentage per county. No one is really pollin data to predict the election in say Wyoming.

```
## Joining, by = "fips"
```

```
##
## Classification tree:
## tree(formula = ExpectedCandidate ~ ., data = train.EC, method = "class")
## Variables actually used in tree construction:
##  [1] "White"        "Transit"      "PrivateWork" "Office"       "Production"
##  [6] "Men"          "Income"       "Service"     "FamilyWork"   "candidate"
## [11] "Citizen"
## Number of terminal nodes:  14
## Residual mean deviance:  0.2974 = 713.3 / 2398
## Misclassification error rate: 0.06012 = 145 / 2412
```

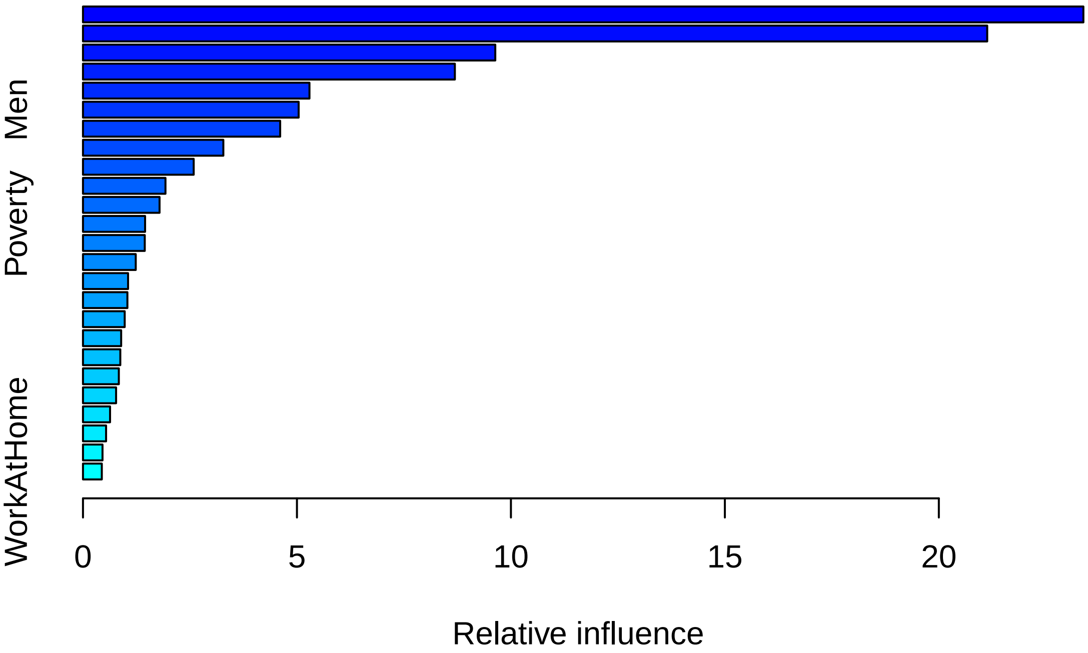


We decided to run a decision tree based on this data where we created an expected winner for each county. Variables such as the population total for each county were dropped as we felt they had served their purpose in being used to create expected candidates under our beleif. We noticed that the first split varaible is on White, which makes sense to us as we beleive the stereotype that Republican party supporters are predominently of caucasian descent. Since Donald Trump mocked hispanics, the largest minority group in the nation during his speeches, it makes sense that large population of caucasion would imply a small hispanic population and thus even less of other minorities.

Regarding the transit variable split, we observed that the split occurs because if there is a small percentage of people using their local public transportation, then it is perhaps possible the area has no public transportation system since it may be a town or small county were the population is small, and thus would be considered a likely Trump victory.

One important variable to stood out to us was the private work variable. We find this indicative of possible small business owners. It may be possible that these small business owners were frustrated from the 2008 recession, and were most likely upset that banks and large corporations were bailed out during the 2009 financial crisis. Many small business owners were perhpas resentful at the 2009 stimulus package since it benefited densley populated areas the most. Most of these places where there are a large number of "PrivateWork" were areas left out of releif during this time period.

```
## Loaded gbm 2.1.4
```



```
##                             var    rel.inf
## Transit                 Transit 23.3777008
## White                     White 21.1298486
## Office                   Office  9.6351779
## PrivateWork         PrivateWork  8.6904030
## Citizen                 Citizen  5.2928152
## Men                         Men  5.0398703
## Income                   Income  4.6081804
## Professional       Professional  3.2802784
## SelfEmployed       SelfEmployed  2.5875736
## IncomeErr             IncomeErr  1.9275647
## IncomePerCap       IncomePerCap  1.7895166
## Poverty                 Poverty  1.4511570
## IncomePerCapErr IncomePerCapErr  1.4439182
## Unemployment       Unemployment  1.2331096
## MeanCommute         MeanCommute  1.0538751
## FamilyWork           FamilyWork  1.0382155
## Carpool                 Carpool  0.9753781
## Service                 Service  0.8924693
## Drive                     Drive  0.8724419
## Production           Production  0.8376132
## OtherTransp         OtherTransp  0.7734043
## Employed               Employed  0.6325413
## ChildPoverty       ChildPoverty  0.5401452
## candidate             candidate  0.4565016
## WorkAtHome           WorkAtHome  0.4403002
```

We notice that under the plot, that the “white” and “transit” are the top variables. This indicates the influence of these variables.

Using our tree, we decided to “boost” it since we were not sure if our prior assumption was correct or not. If we had made this analysis at the time of 2016, we would not have known what the true election outcome would have been. Thus we would have had to assume that our subjective beleif about the voting outcome as a “weak” prediction/assumption. In order to remedy this, we employed the use of boosting the tree in order to create an even stronger classifier. Using a boost on our "weak" tree allowed for it become strong by making each new iteration of the tree stronger; by fitting information from the previous tree. We thought a boost size of one thousand would be good, not too small nor too big.

To be honest, we were disappointed that the previous decision tree split on "White" and beleived that perhaps using a subjective assumption on the prior was the wrong thing to do.

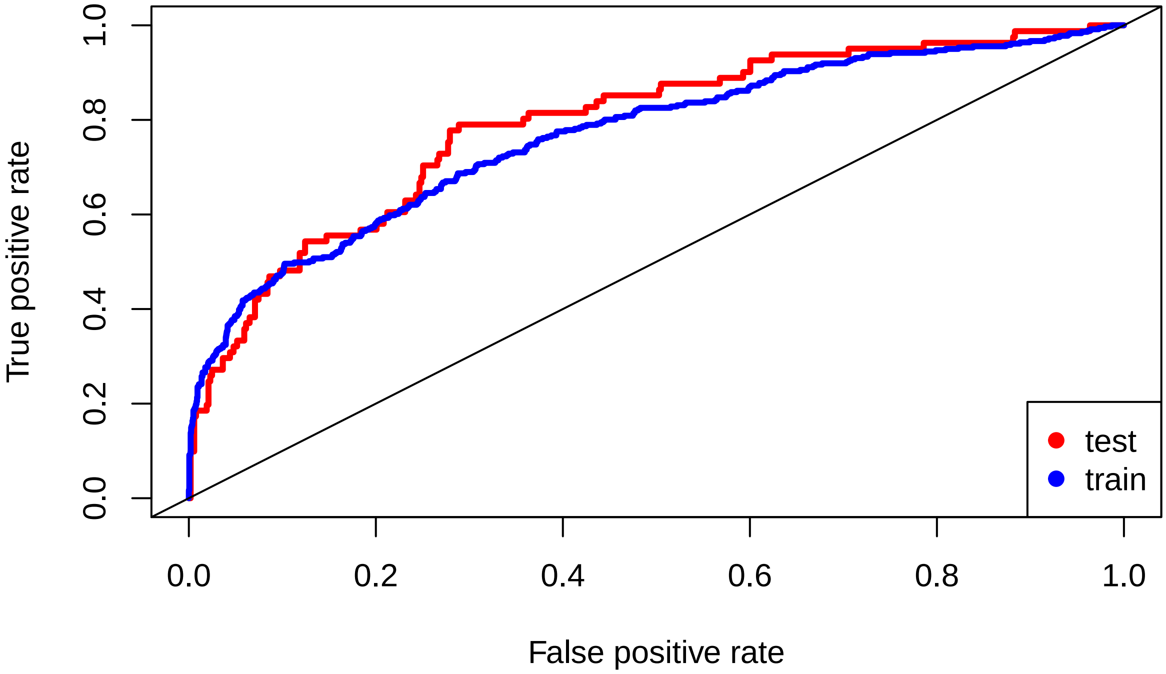
After we boosted, we noticed that the most important variable was now Transit, which is in line with the decision tree trained under the true county outcome from earlier. We found this fascinating as we were not expecting our data with expected candidates to be in line with the true decisive variable, an indicater if the population was rural or not "transit". This variable boosted Trump to victory on Nov. 8 2016. We thus found the true "deciding population" of the election without looking at the real data just by having the correct assumption. None of the previous classifiers could have acheived this on thier own since would would have had no true training data that is unbaised in the frequentist sense and thus suffered by other forms of biasing.

We noticed that under the turnout voter graph of the US map, when we compare it to our map with expected winners along with the true county winner map, that Trump won most places where there was a low voter turnout. This is true especially in the mideast. This is where most of the counties had less than below average county population. The boosted tree helped immensely in identifying important variables, when we had the right subjective beleif.

```
## [1] 0.1490066
```

```
## Warning in predict.gbm(extreme.freedom, newdata = train.EC, n.trees =
## 1000, : NAs introduced by coercion
```

```
## [1] 0.1625207
```



We then decided to use the boosted decision tree model for prediction when it had been trained with the expected candidate win per county. This was then compared to the true winner of each candidate after the election results came in. This was done as a test to see how good our assumption about how voters would vote which as a reminder, was based on county population or a minority population less than ten percent going to Trump. We were ecstatic to observe a test error rate of around 15% and a training error rate of about 16%. We then followed up with an ROC curve regarding true postive rates. Although it is not the best ROC curve we have ever seen, we deem it passable considering it is an election we are trying to predict, which has much variablility, where even experts in the field failed to predict the 2016 election. With these results, we were ecstatic to find out that our method was not complete nonsense and that our “prior” was a good one.

Final Thoughts.

We knew that polling for the presidential election is in itself bias by nature, thus we made an assumption of what the county that would vote for Trump would look like, and similated a situation where the election had not occured and we where left only with census data. Our prior came from what we had observed in our daily lives which was affected by our surroundings. We created a decision tree that was in a way “Bayesian” in spirit although no Bayesian techniques were explicitly used. We then boosted the tree since our decision tree was biased by our assumption and prior. We concluded this made the decision tree a weak learner. From boosting, we made the tree strong and because of this were able to grab the important variables such as transit, which was the true deciding variable in the 2016 election. Following up on this we predicted by training on our Expected Candidate win for each county. Machine learning currently has limitations, but with the right tools and risk taking, can push the bounds to acheive the unacheivable and we look forward to seeing how far it can go in our lifetimes. -Thank You