FinalProject-131

Shravan Shenoy(707570-8) + John Weisner(7012719) (Both students are 131 only)

6/6/2021

1. What makes voter behavior prediction (and thus election forecasting) a hard problem?

Predicting voter behavior is a hard problem for a variety of reasons. Any poll will have sources of error in it, and these errors often have a tendency to add up instead of canceling each other out. One source of error is sampling bias. The people who are polled fall within the demographic of the means of which the poll was conducted. For example, polls which call landlines are more likely to sample a demographic old enough to still own a landline phone. Other error comes from people who were sampled either lying or changing their mind. In 2016, there was a case of the 'silent majority', which refers to people who voted for Trump, but do not make it known publicly that they support him. This is called the 'Shy Tory Effect'. These people may have told pollsters that they were not voting or that they were voting for Clinton when they actually ended up voting for Trump. There are many difficulties when it comes to predicting elections, such as obtaining a representative sample and accounting for people who change their minds at polls. Other issues include a lack of ample datasets of previous elections, data collection biases, and poll errors. Because of the 4 year time span between elections, gathering sufficient data for a single election takes years, and attempting to analyze further than just the previous election requires an additional 4 years of data collection, so on and so forth for older elections. The large amount of time between elections also makes it difficult to analyze a large set of them, resulting in a very small dataset to work with. Additionally, the variables being used for analysis will vary since methods of data collection and the kind of data collected changes with time and a comprehensive dataset for that entire timespan is not guaranteed, leading to skewed responses and biased, incomplete data. Having to account for these unavoidable characteristics affects the accuracy of the data, leading to a mere approximation being the most accurate result that can be obtained. The lack of accuracy in data collection and high variability of each election makes voter behavior and election forecasting difficult to predict.

2. What was unique to Nate Silver's approach in 2012 that allowed him to achieve good predictions?

Nate Silver had an interesting model because he acknowledged that opinions change over time and that someone who would vote Democrat if the election were today may change their minds over the course of a few months. Nate Silver also used Bayes Theorem and graph theory to help predict what actual probabilities were compared to what the polling numbers suggested. He decided that his model would take into account the actual percentage of people who vote for the candidate + the house effect + sampling variation would be the polled amount of people voting for a candidate. His model was successful because it changed over time, acknowledged uncertainties, and utilized advanced theories to mold probabilities.By allowing a wide range of possibilities to be accounted for, sophisticated utilization of Bayes Theorem, and graph theory, overall error is reduced and the likelihood of error is also minimized. Another favorable aspect that Silver used was consistently using incoming polling data that from election campaigning results. Comparing theorized "polling bias" to actual votes allowed Silver to better understand the margin of bias that needed to be accounted for as well, which also helped accuracy.

3. What went wrong in 2016? What do you think should be done to make future predictions better?

One of the issues in 2016 was the Shy-Tory Effect. Pollsters found that Trump polled better among people who were talking to recorded voices instead of live pollsters. Furthermore, Democrats did not turn up to the polls as much as was anticipated, so their numbers dropped. Additionally, Trump received late support from previously undecided Republican voters as well as voters who favored Gary Johnson, but decided to vote for one of the two main parties. The data given to pollsters is theorized to be incredibly skewed, as many

generalized demographics such as young voters, women, and people of specific swing states, were considered quite reluctant to actually give significant data, thus leading to pollster data being inaccurate while the vote still favored Trump. Outperforming several polls as a result, Trump was able to win the election despite many statisticians and theorists to have believed the opposite would have occured. Some ways to make future predictions better would be to make voter polling more accessible. Landline polling is considerably outdated in terms of accessibility, and it is likely that email, web, and text polls would be significantly more accessible. In addition, creating polls that incentivize people to give an answer (such as some form of raffle or possible cash prize) would encourage more people to answer polls through some medium and thus reach a wider audience and sample size.

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.

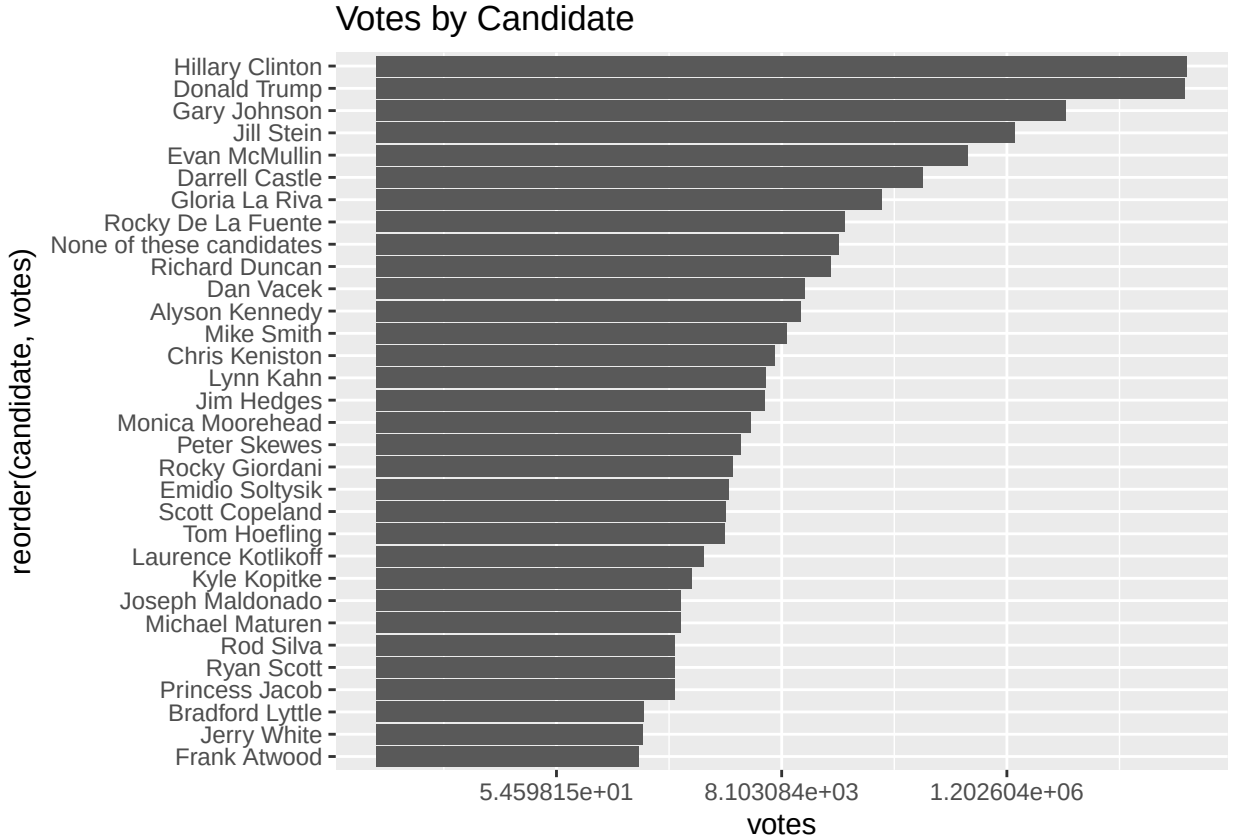
The dimensions of the updated election.raw is 18345x5.

The entries where fips = 2000 all have the state as Alaska and the county as NA. All the other entries with county as NA have the fips equal to the state abbreviation (CA, VT, NY, etc.) or equal to US except for the NA entries where fips = 2000. So, these entries do not match the description provided of being summary rows and are not counties in the United States, so they are excluded.

5. Remove summary rows from election.raw data:

6. 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!

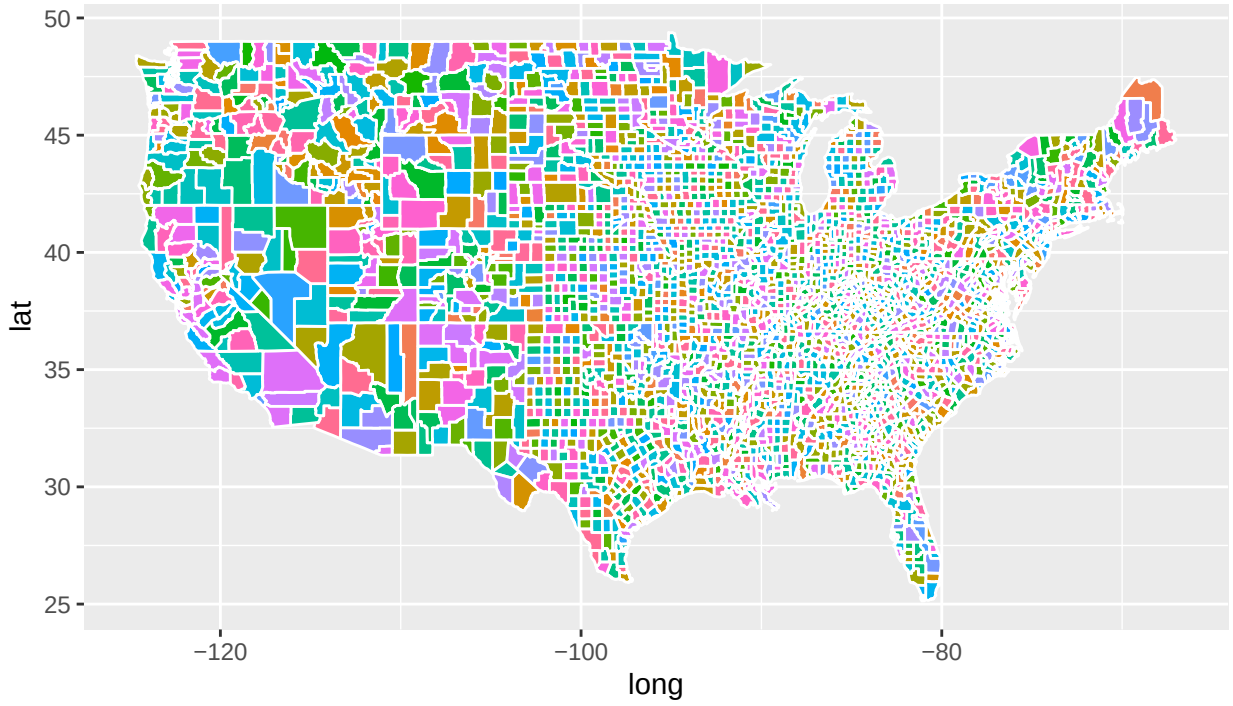
There were 31 different presidential candidates in 2016 in addition to a 32nd group which is 'None of these candidates'.



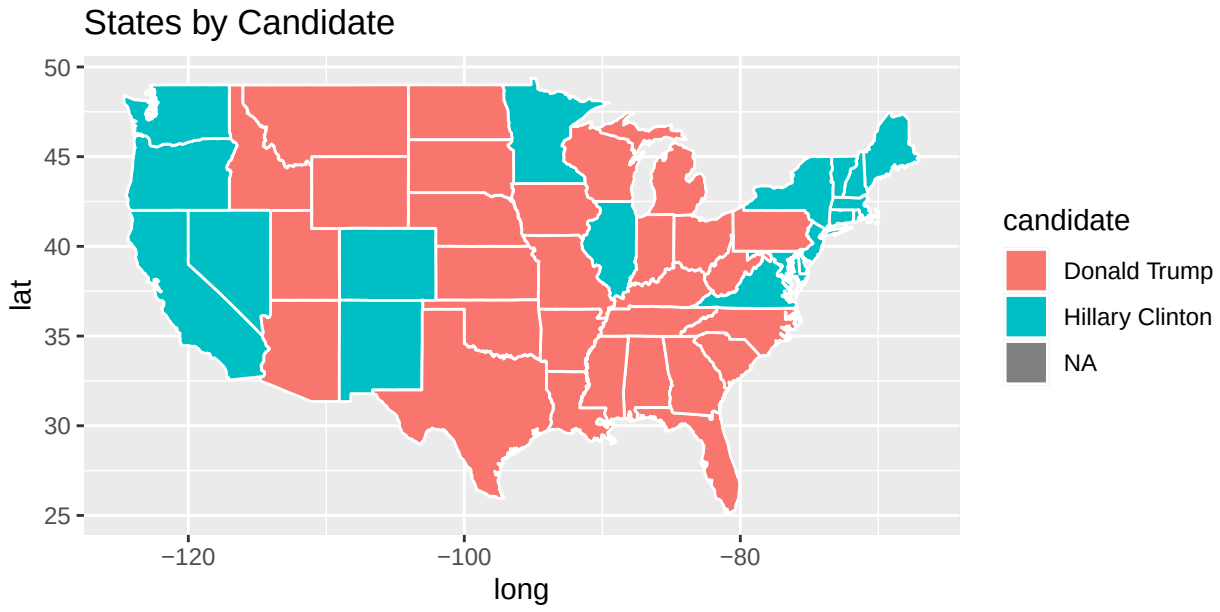
7. Create variables county_winner and state_winner by taking the candidate with the highest proportion of votes. Hint: to create county_winner, start with election, group by fips, compute total votes, and

pct = votes/total. Then choose the highest row using top_n (variable state_winner is similar).

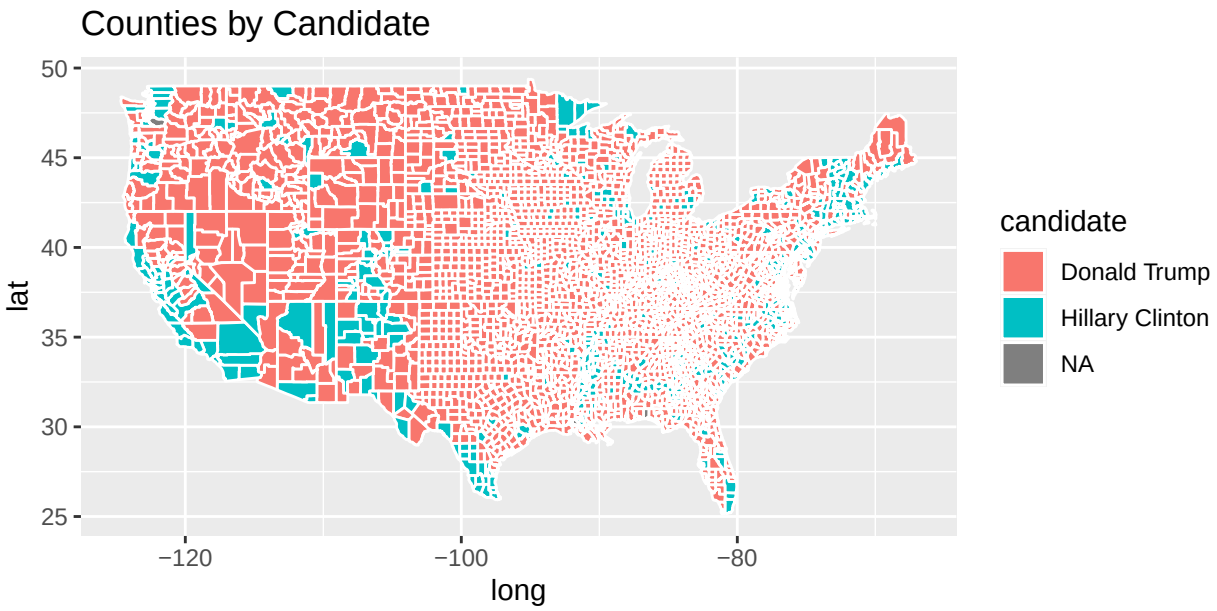
8. Draw county-level map by creating counties = map_data("county"). Color by county



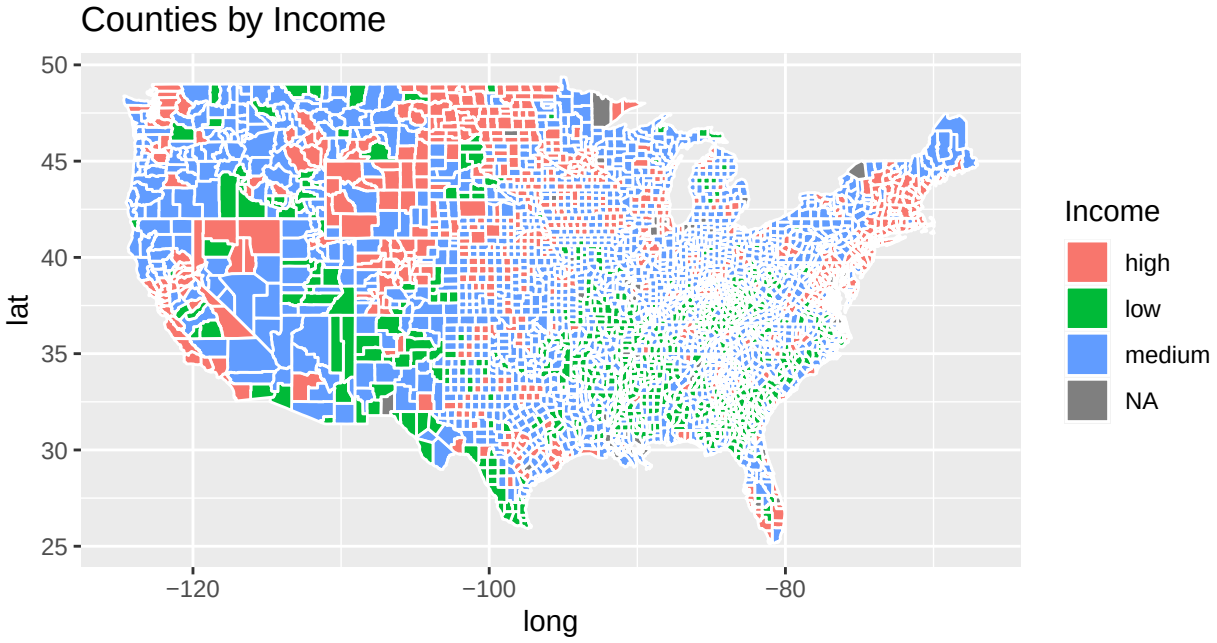
9. Now color the map by the winning candidate for each state.



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



11. Create a visualization of your choice using census data.



Here is a visualization showing the average income of each county in the United States. The income is separated into factors. High corresponds to roughly the 75th percentile at an income of over 27,000. Low corresponds to roughly the 25th percentile at an income of under 20,000. Medium is the values in between those. Some counties had missing data, so those are indicated as well. Many of the high income counties are located along the California coast or in the New Region. When comparing this to the county election map, these high income counties often voted for Hillary Clinton.

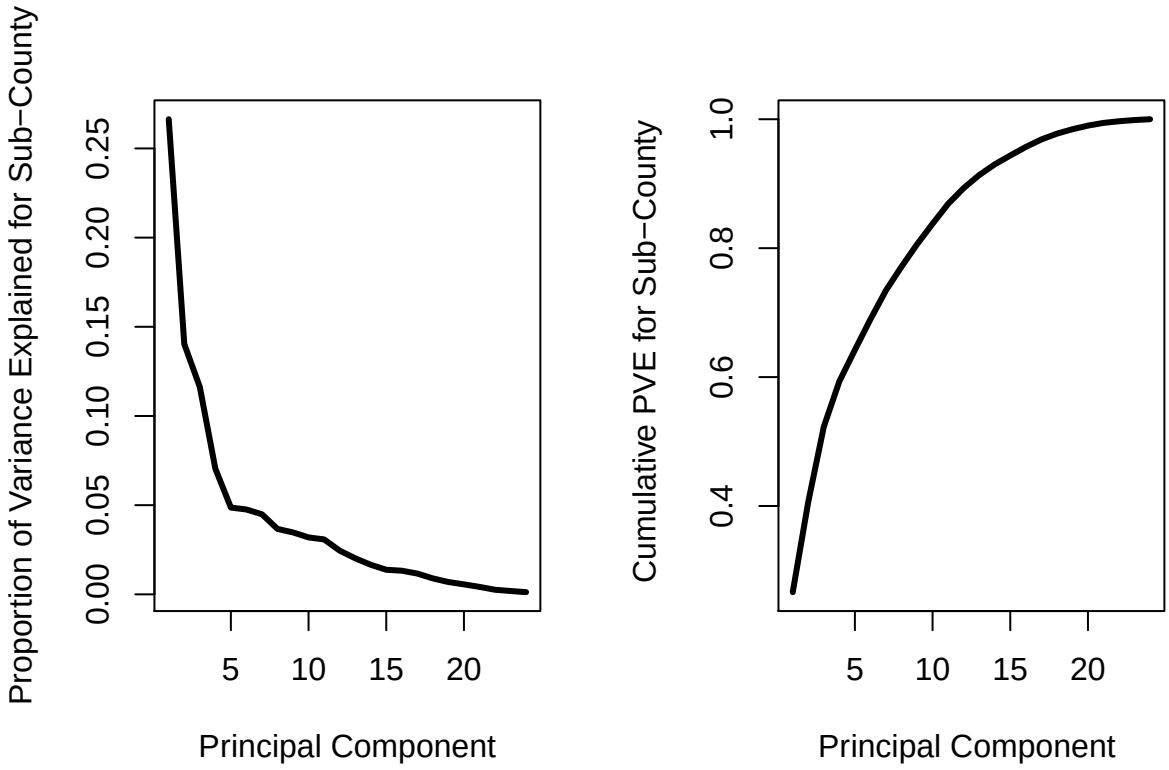
12.Clean census data census.del: start with census, filter out any rows with missing values, convert {Men, Employed, Citizen} attributes to percentages (meta data seems to be inaccurate), compute Minority attribute by combining {Hispanic, Black, Native, Asian, Pacific}, remove these variables after creating Minority, remove {Walk, PublicWork, Construction}.

13. Run PCA for both county & sub-county level data. Save the first two principle components PC1 and

PC2 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 correaltion between these features?

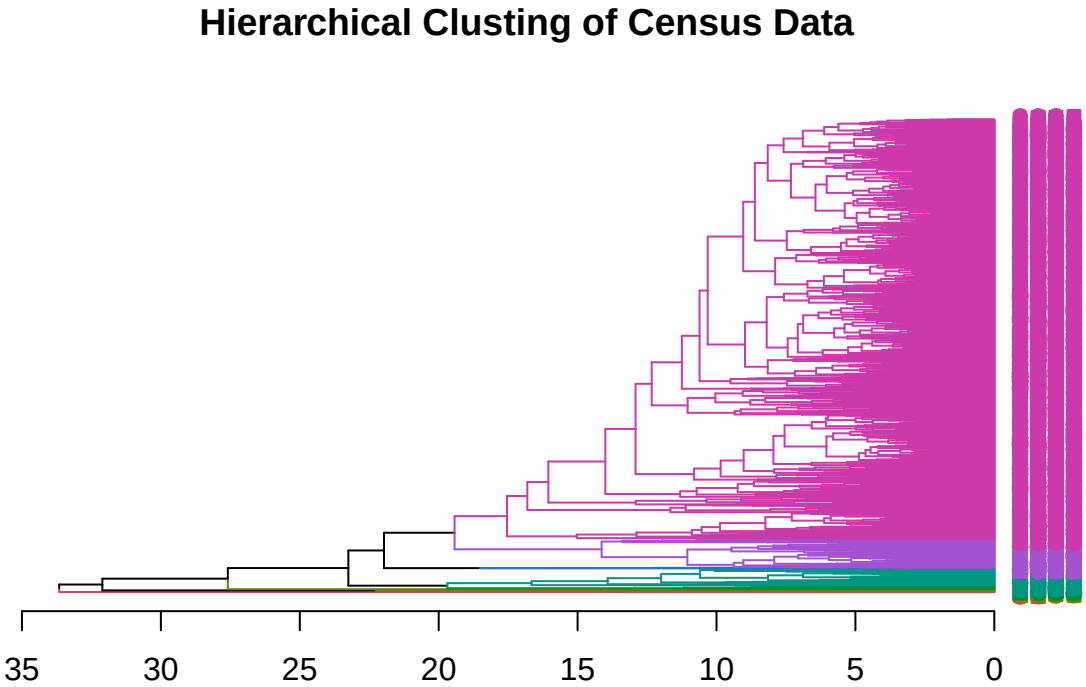
IncomePerCap, ChildPoverty, and Poverty have the highest absolute values for PC1 for both county and sub county level data. IncomePerCap has an opposite sign than ChildPoverty and Poverty, so it is negatively correlated with those variables(high income tends to mean low child poverty and poverty). Poverty and child poverty were both positively correleated, so high values of one indicate high values of the other and low values of one indicate low values of the other.

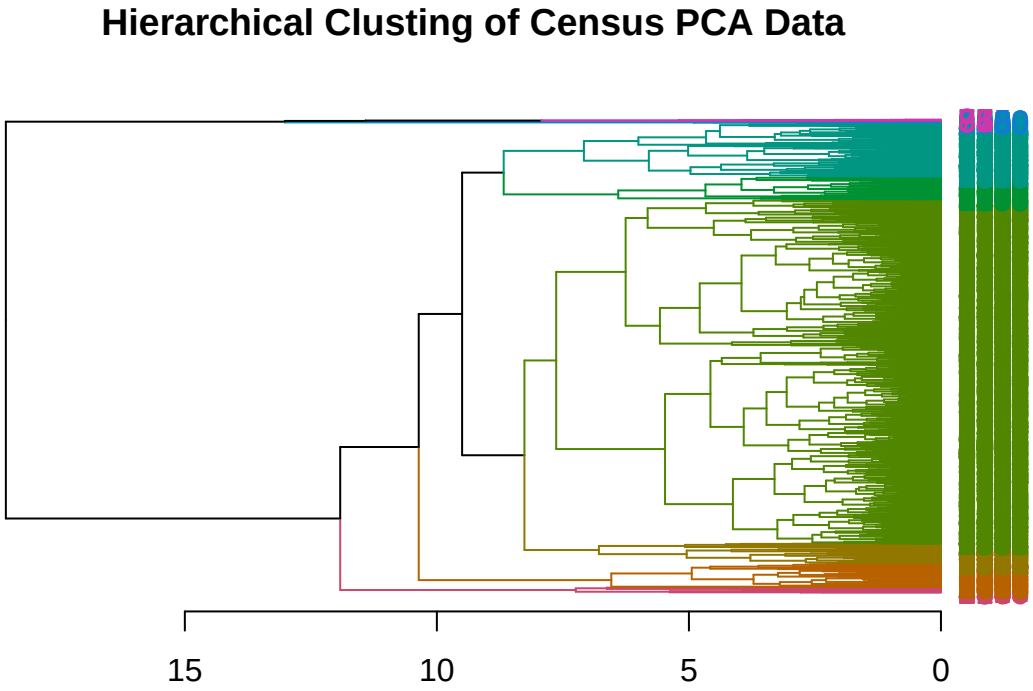
14. Determine the number of minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses. Plot proportion of variance explained (PVE) and cumulative PVE for both county and sub-county analyses.



3 principal components explain over half the data for both the county and sub-county levels.

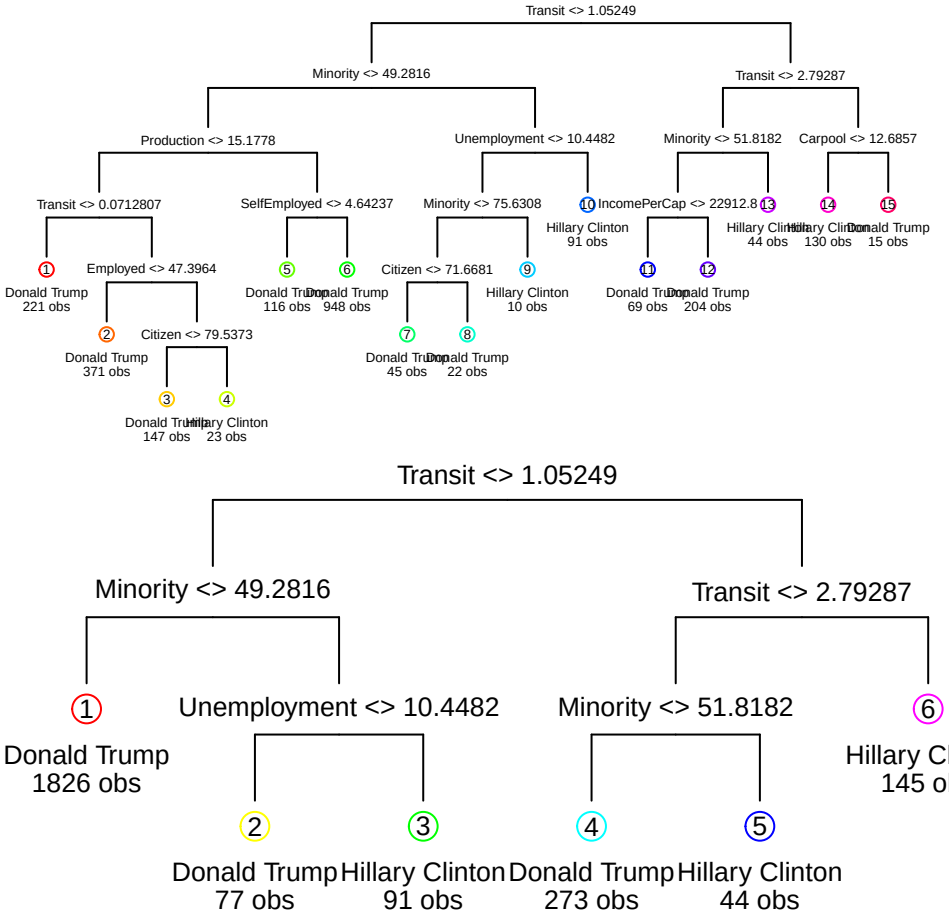
15. With census.ct, perform hierarchical clustering with complete linkage. Cut the tree to partition the observations into 10 clusters. Re-run the hierarchical clustering algorithm using the first 5 principal components of ct.pc as inputs instead of the originald features. Compare and contrast the results. For both approaches investigate the cluster that contains San Mateo County. Which approach seemed to put San Mateo County in a more appropriate clusters? Comment on what you observe and discuss possible explanations for these observations.





San Mateo was part of the 3rd cluster for the original hierarchical clustering method and part of the 4th cluster for the clustering using the first 5 principal components. Both of clusters contain many of the same counties, however there are a few differences. The clustering based PC contains less counties which voted for Trump and, since San Mateo voted for Clinton, the PC clustering method is likely the better one.

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. Visualize the trees before and after pruning. Save training and test errors to records variable. Intepret and discuss the results of the decision tree analysis. Use this plot to tell a story about voting behavior in the US (remember the NYT infographic?)



Based on the pruned decision tree, counties with high Transit values, high Minority rates, and high unemployment were more likely to vote for Clinton.

More specifically, the following can be concluded about each county: higher than a 2.80 value for transit indicates a Clinton county, between 1.05 and 2.80 value for transit and lower than a 51.8% minority rate indicates a Trump county, between 1.05 and 2.80 value for transit and higher than a 51.8% minority rate indicates a Clinton county, less than 1.05 values for transit and lower than a 49.28% minority rate indicates a Trump county, less than 1.05 values for transit, higher than a 49.28% minority rate, and higher than 10.45 average unemployment indicates a Trump county, and less than 1.05 values for transit, higher than a 49.28% minority rate, and higher than 10.45 average unemployment indicates a Clinton county.

```
##          train.error test.error
## tree      0.07939739 0.07317073
## logistic          NA         NA
## lasso             NA         NA
```

17. Run a logistic regression to predict the winning candidate in each county. Save training and test errors to records variable. What are the significant variables? Are the consistent with what you saw in decision tree analysis? Interpret the meaning of a couple of the significant coefficients in terms of a unit change in the variables.

Citizen, Income, IncomePerCap, IncomePerCapErr, Professional, Service, Office, Production, Drive, Carpool, WorkAtHome, MeanCommute, Employed, PrivateWork, FamilyWork, Unemployment, and Minority are all significant variables.

This is not completely consistent with the decision tree model. In the tree, transit was significant, however it is not in the regression model. Minority and unemployment were significant in both the models. All the

other variables not mentioned that were significant in the regression model were not significant in the pruned decision tree.

Unemployed has a coefficient of 0.216. This indicates that each one unit change in percentage of unemployed people in the county corresponds to an increase of 0.216 in the log odds of Clinton winning that county .This is a change of the odds which is eˆ(0.216) which is equal to 1.24, so Clinton's odds increase of winning the county by .24.

Drive has a coefficient of -0.206. This indicates that each one unit change in percentage of people that drive to work in the county corresponds to an decrease of 0.206 in the log odds of Clinton winning that county. This is a change of the odds which is eˆ(-0.206) which is equal to .8138, so Clinton's odds decrease of winning the county by .1862.

```
##          train.error test.error
## tree      0.07939739 0.07317073
## logistic  0.07003257 0.06666667
## lasso             NA         NA
```

18. You may notice that you get a warning glm.fit: fitted probabilities numerically 0 or 1 occurred. As we discussed in class, this is an indication that we have perfect separation (some linear combination of variables perfectly predicts the winner). This is usually a sign that we are overfitting. One way to control overfitting in logistic regression is through regularization. Use the cv.glmnet function from the glmnet library to run K-fold cross validation and select the best regularization parameter for the logistic regression with LASSO penalty. Reminder: set alpha=1 to run LASSO regression, set lambda = c(1, 5, 10, 50) * 1e-4 in cv.glmnet() function to set pre-defined candidate values for the tuning parameter lambda. This is because the default candidate values of lambda in cv.glmnet() is relatively too large for our dataset thus we use pre-defined candidate values. What is the optimal value of lambda in cross validation? What are the non-zero coefficients in the LASSO regression for the optimal value of lambda? How do they compare to the unpenalized logistic regression? Save training and test errors to the records variable.

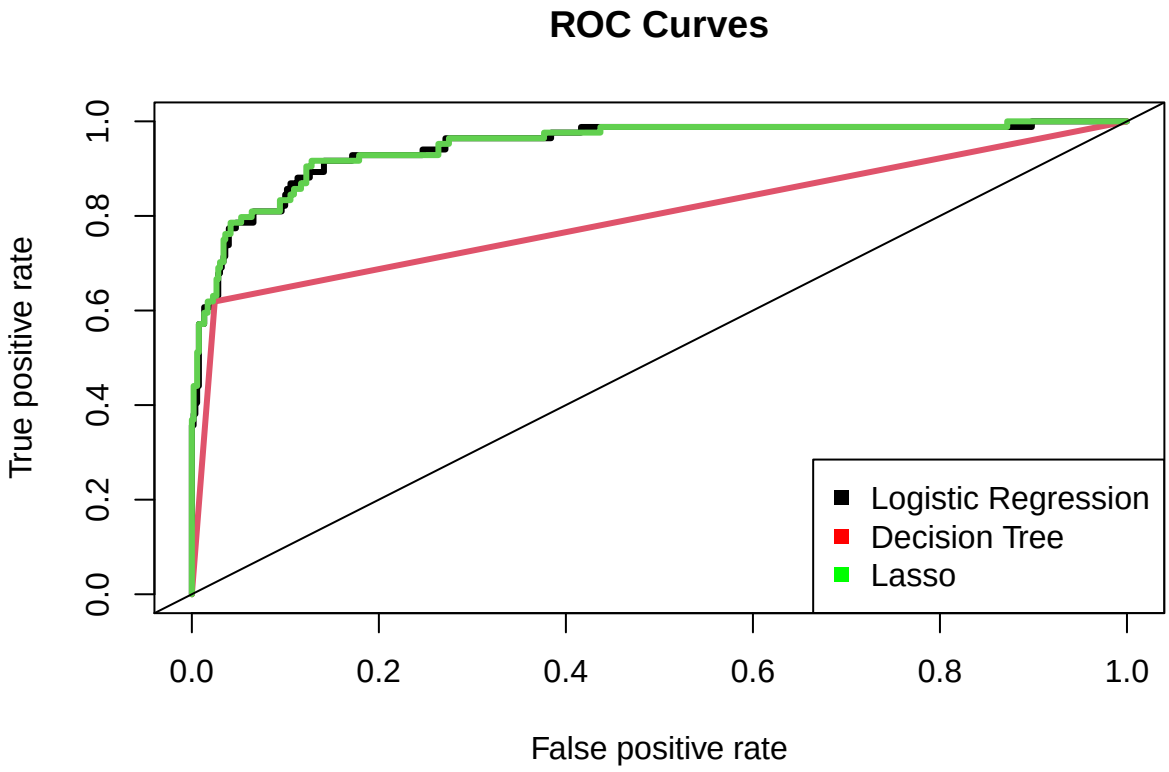
The optimal value of lambda is 5x10ˆ-4.

```
##          train.error test.error
## tree      0.07939739 0.07317073
## logistic  0.07003257 0.06666667
## lasso     0.06881107 0.06341463
```

The non-zero coefficients for the optimal value of lambda are the following: Men, Citizen, Income, IncomeErr, IncomePerCap, IncomePerCapErr, Poverty, Professional, Service, Office, Production, Drive, Carpool, Transit, OtherTransp, WorkAtHome, MeanCommute, Employed, PrivateWork, SelfEmployed, FamilyWork, Unemployment, and Minority.

Compared to the unpenalized logistic regression model, the lasso regression has a slightly higher training error of 0.068 compared to the logistic model training error of 0.066. There seems to be some variance on whether the coefficients are higher or lower for the logistic model compared to the lasso model, but for the most part it seems that the coeficients are closer to zero in the lasso model and (slightly) more spread out in the unpenalized model.

19. Compute ROC curves for the decision tree, logistic regression and LASSO logistic regression using predictions on the test data. Display them on the same plot. Based on your classification results, discuss the pros and cons of the various methods. Are the different classifiers more appropriate for answering different kinds of questions about the election?



The Decision Tree has the worst ROC curve, not capturing much of the data. This model is used because it is simple and easy to understand, but it is not as good as the other methods at classifying the data successfully. The logistic and lasso ROC curves are both very similar with practically no difference. It seems that the logistic regression curve is slightly better, but not by much. Lasso regression is good to use when there are redundant variables. Logistic regression is good to use when you want every variable accounted for, even just in a small amount.

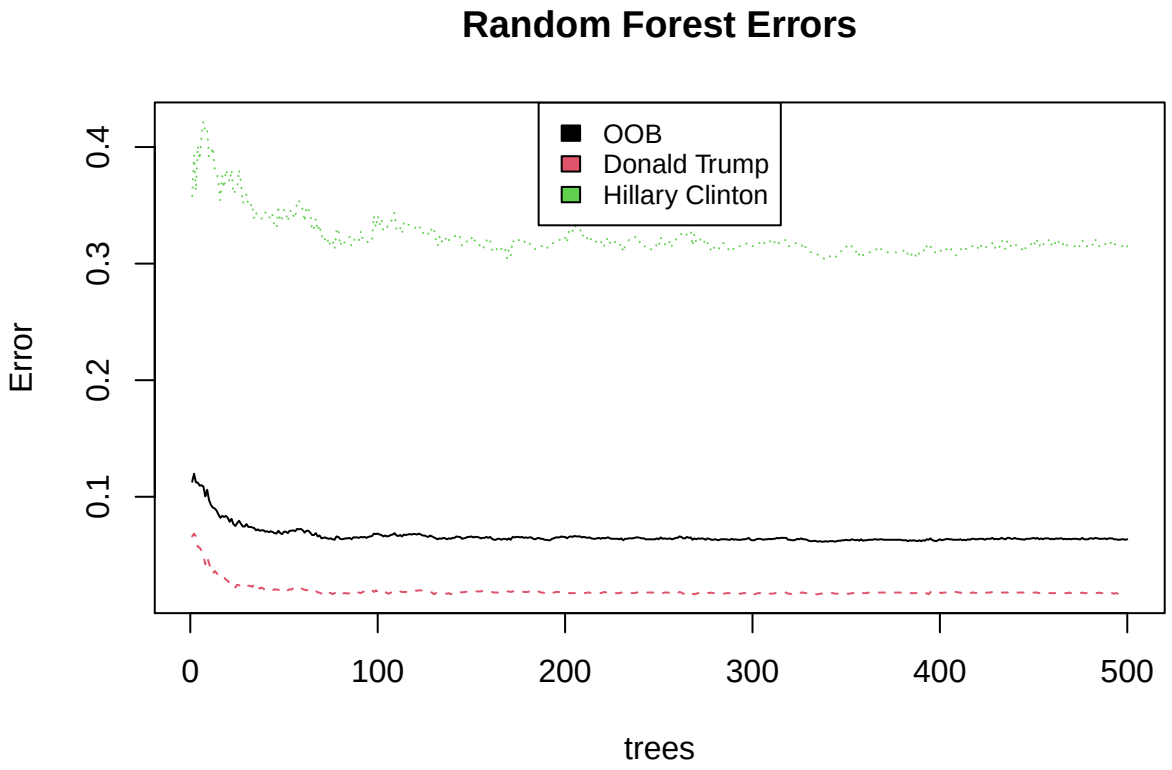
20. Interpret and discuss any overall insights gained in this analysis and possible explanations.

Our first exploration was to fit a boosting model to the data. Minority and tranisit are the two most important predictors in the boosting model. This is very similar to the pruned decison tree, where minority and transit were two of the significant variables.

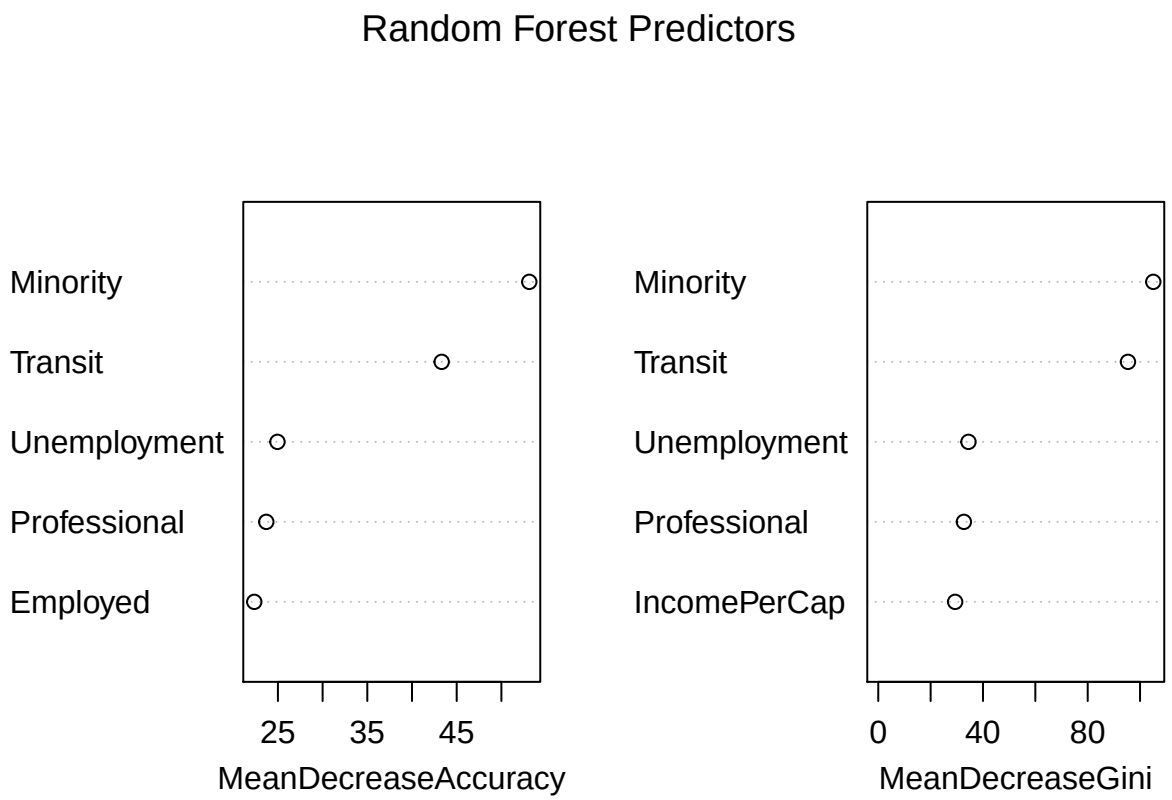
The boosting model correctly predicted 521/51 counties for Trump and 61/84 counties for Clinton. The test error rate is 0.055 and the training error rate is 0.036. Both of these values are lower than the error rates for logistic regression, lasso regression, and decision tree. So the boosting model is a very good predictor for the election data.

Next, we explored fitting a random forest model. 500 trees were created with 4 variables tried at each split. There was an OOB error rate of 6.35%. The classification error rate for Clinton was 0.315 versus an error rate of only 0.017 for Trump, so the amount of states predicted to go Hillary's way but actually went Trump's way was much higher than the opposite.

The random forest model predicted 524/531 correctly for Trump and 65/84 counties correctly for Hillary. The error rate is about 0.05. This is quite similar to the error rate for boosting, so the two models are comparable in their accuracy. The error is also much lower than the errors calculated for logistic regression, lasso regression, and the decision tree model.



As shown above, just like in the other methods, the error rate for Hillary was much higher than the error rate for Trump.



Minority and Transit were by far the two most important predictor varibales in the random forest model. This is very similar to both the pruned decison tree and the boosting model, where minority and transit were

two of the significant variables.

The optimal amount of neighbors found using a k-nearest neighbor model with leave-one-out-cross-validation (LOOCV) was 34. The test error rate for the knn model was 0.130 and the training error rate was T .134. These error rates are much higher than any of the previous models, so knn is likely not the optimal fit for our data.

Conclusion (#20 Continued)

Through the completion of this project, we utilized decision trees, logistic regression, lasso regression, random forest, boosting, and knn LOOCV. Though these methods, we found that improtant predictors in the 2016 election were transit, minority rates, unemployment, income, drive, and production, a long with some other smaller important predictors. Most important among these were transit and minority. One of the additional questions that we tackled was looking at the results of a random forest model and observing how good of a fit that model was for predicting the electon. One thing of note was the error rate difference between the candidates. Clinton had an error rate hovering around 0.4 as more trees were added. Comparatively, Trump had an error rate of less than 0.1. The contrast in these two errors rates is extremely high and can help explain why 2016 went differently than many of the experts predicted.

Between the lasso, logistic regression, and decision tree models, we believe that the lasso regression model was the best. The decision tree ended up overfitting the data and had a high error rate. The logistic regression also overfit the data with perfect separation. Lasso had very similar results to the logistic regression, but controlled overfitting with a penalty parameter lambda. Since lasso kept both error and overfitting under control, it is a good model to use.

Boosting, random forest, and KNN LOOCV were three additional models that we looked at to fit the data. We found that boosting and random forest both did a fairly good job at modelling the data, whereas knn was less effective due to being more linear in nature and not as effectively taking into account other covariates.