PSTAT 131 Final Project

Yubo Wei (6990006) & Haoze Zhu (3141892)

06/11/2021

Background

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

First, voter behavior varies over time. A change in unemployment rate in a particular state can affect voter intention in that states. National change such as rise in federal tax could lead to variation in voter intention nationwide.And some changes in voter behavior are difficult to measure such as a successful campaign ad. Second, the poll can be a problematic source of data. There could be biases during the data collection process, people could provide false information, people could change their opinions from time to time, and finally the variables used for analysis could be not representative for the prediction.

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

Silver uses hierarchical modelling with adjustment to voter behavior, the house effect, and sampling variation. Instead of maximising probability of variation of the support, Silver looks at the full range of probabilitiesuses and uses Bayer's Theorem and graph theory to calculate the new probabilities of each level of support.This model can also be simulated forward in time for each estimated level of support. As much polling data become available towards the end of the election campaign, Silver can get better estimates of public support for Obama.

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

All polls has errors. It is possilbe that there are systematic polling errors in state polls because the forecasts based on them missed in the same direction. The polls underestimated Trump's support in groups like whites without college degrees. Trump voters , especially women, might be too shy to tells pollsters whom they were supporting. We need a better statistical model to estimate these errors efficiently in the future and bulid a more conclusve polling system.

Data Wrangling

4. Report the dimension of election.raw after removing rows with fips=2000. Provide a reason for excluding them.

```
## # A tibble: 6 x 5
##   county fips  candidate          state  votes
##   <chr>  <chr> <fct>              <chr>  <dbl>
## 1 <NA>   2000  Donald Trump       AK    163387
## 2 <NA>   2000  Hillary Clinton    AK    116454
## 3 <NA>   2000  Gary Johnson       AK     18725
## 4 <NA>   2000  Jill Stein         AK      5735
```

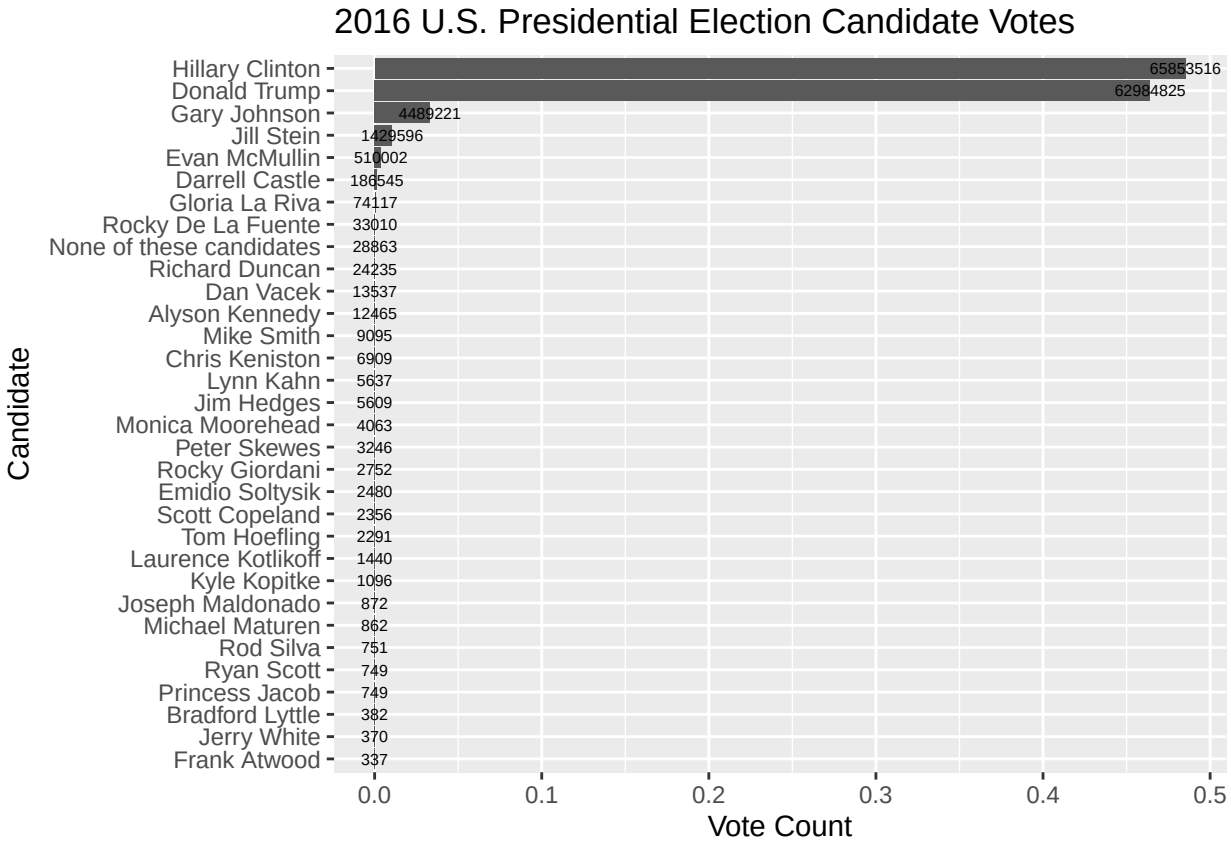
```
## 5 <NA>   2000  Darrell Castle     AK      3866
## 6 <NA>   2000  Rocky De La Fuente AK      1240
```

```
## [1] 18345     5
```

Reason: Those rows had NA values. The dimension is 18345 * 5

5. Remove summary rows from election.raw data: i.e.,

6. How many named presidential candidates were there in the 2016 election? Draw a bar chart of all votes received by each candidate.



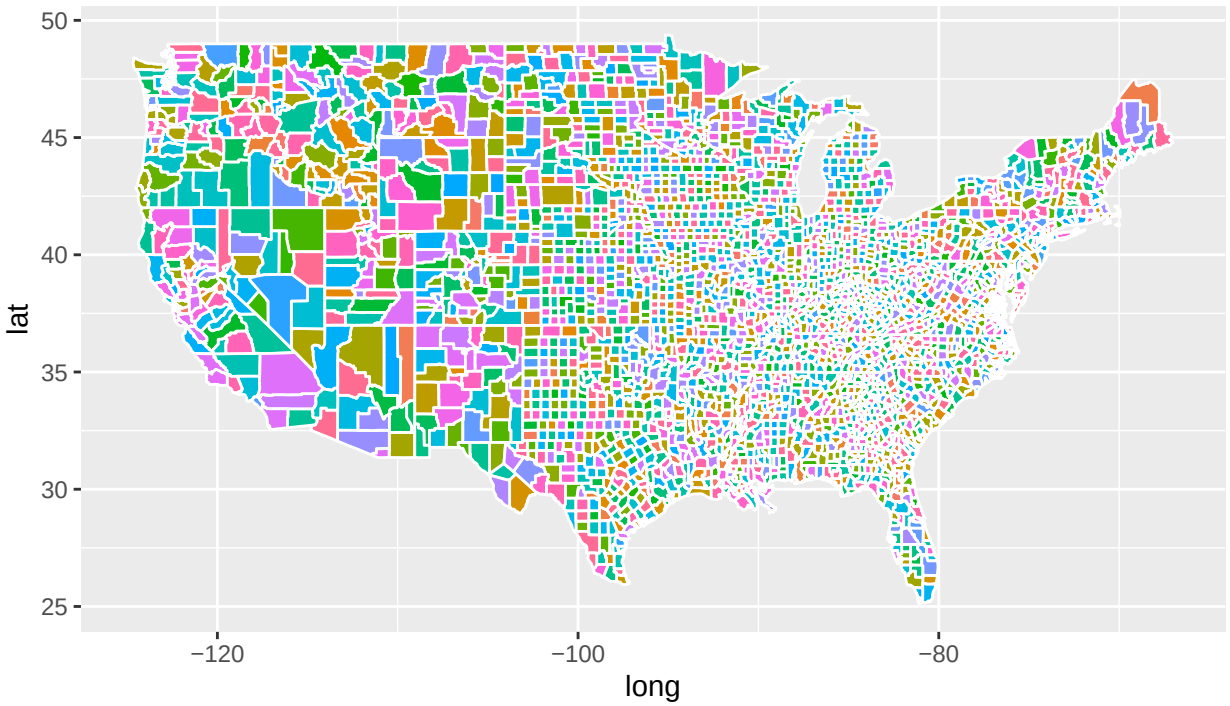
There are 32 named presidential candidates in the 2016 election.

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

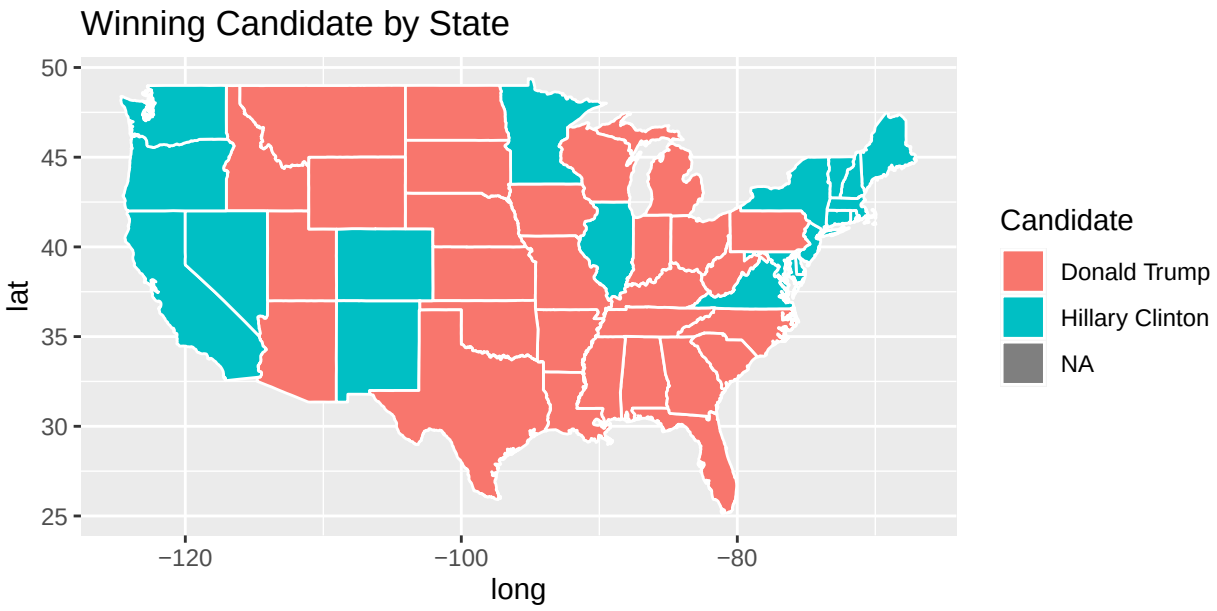
```
## Selecting by pct
## Selecting by pct
```

Visualization

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

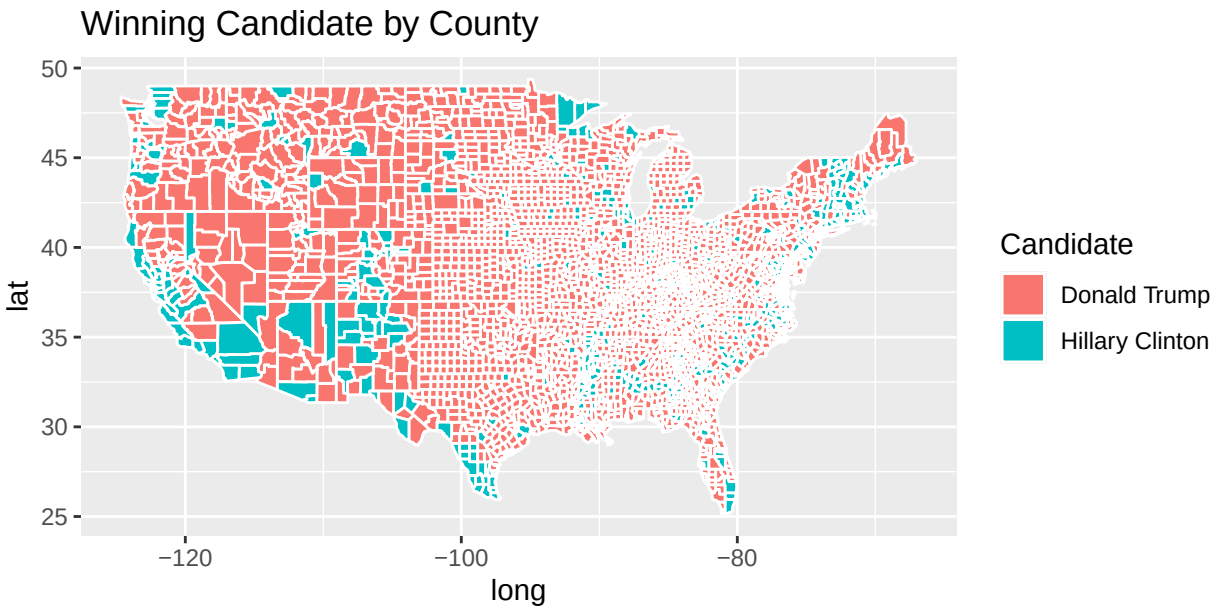


9. Now color the map by the winning candidate for each state.

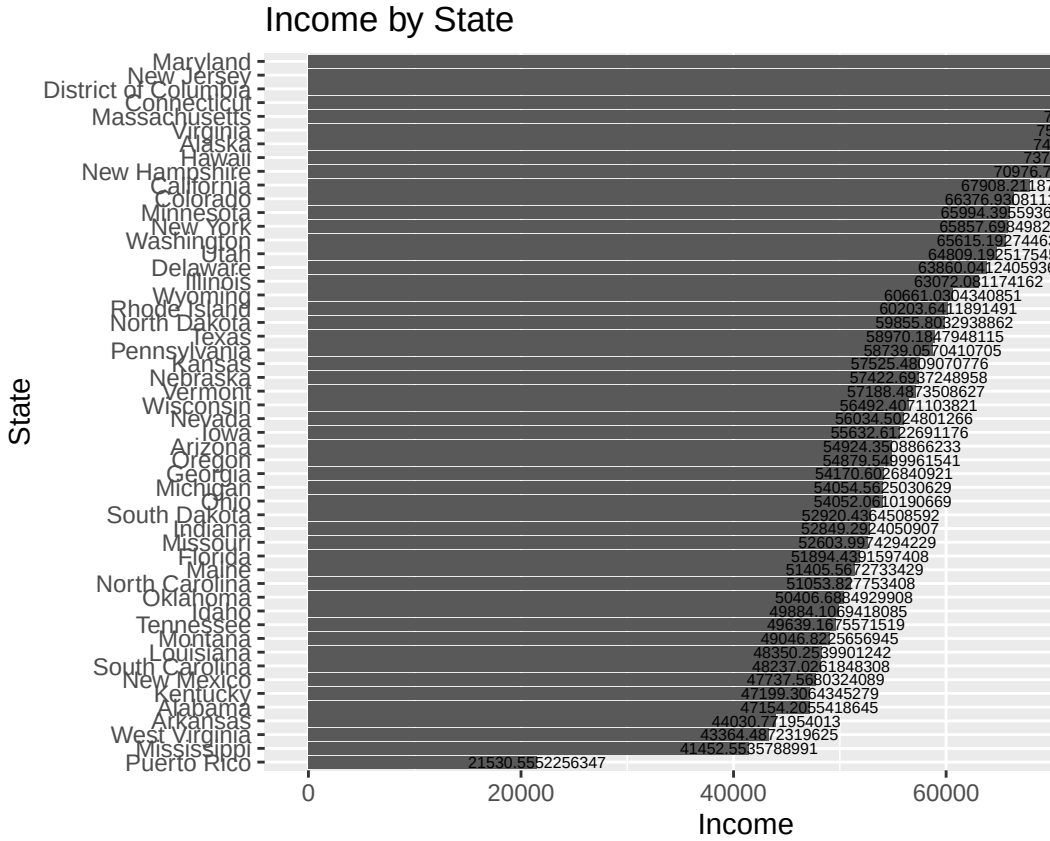


10. Now color the map by the winning candidate for each county The variable county does not have fips column. So we will create one by pooling information from maps::county.fips.

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



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



Here we show the income by state.

12. In this problem, we aggregate the information into county-level data by computing TotalPop-weighted average of each attributes for each county.

```
## # A tibble: 6 x 26
## # Groups:   State [1]
##   State   County    Men Citizen Income IncomeErr IncomePerCap IncomePerCapErr
```

```
##   <chr>   <chr>   <dbl>   <dbl>  <dbl>     <dbl>        <dbl>           <dbl>
## 1 Alabama Autauga  48.4    73.7 51696.     7771.       24974.           3434.
## 2 Alabama Baldwin  48.8    75.7 51074.     8745.       27317.           3804.
## 3 Alabama Barbour  53.8    76.9 32959.     6031.       16824.           2430.
## 4 Alabama Bibb     53.4    77.4 38887.     5662.       18431.           3074.
## 5 Alabama Blount   49.4    73.4 46238.     8696.       20532.           2052.
## 6 Alabama Bullock  53.0    75.5 33293.     9000.       17580.           3111.
## # ... with 18 more variables: Poverty <dbl>, ChildPoverty <dbl>,
## #   Professional <dbl>, Service <dbl>, Office <dbl>, Production <dbl>,
## #   Drive <dbl>, Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>,
## #   WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>,
## #   SelfEmployed <dbl>, FamilyWork <dbl>, Unemployment <dbl>, Minority <dbl>
```

13. Run PCA for both county & sub-county level data. 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?

```
## [1] 3218   24
```

```
## [1] 72727    24
```

```
## IncomePerCap ChildPoverty      Poverty
##    0.3671276    0.3442901    0.3398956
```

```
## IncomePerCap Professional       Income
##    0.3360229    0.3235265    0.3204694
```

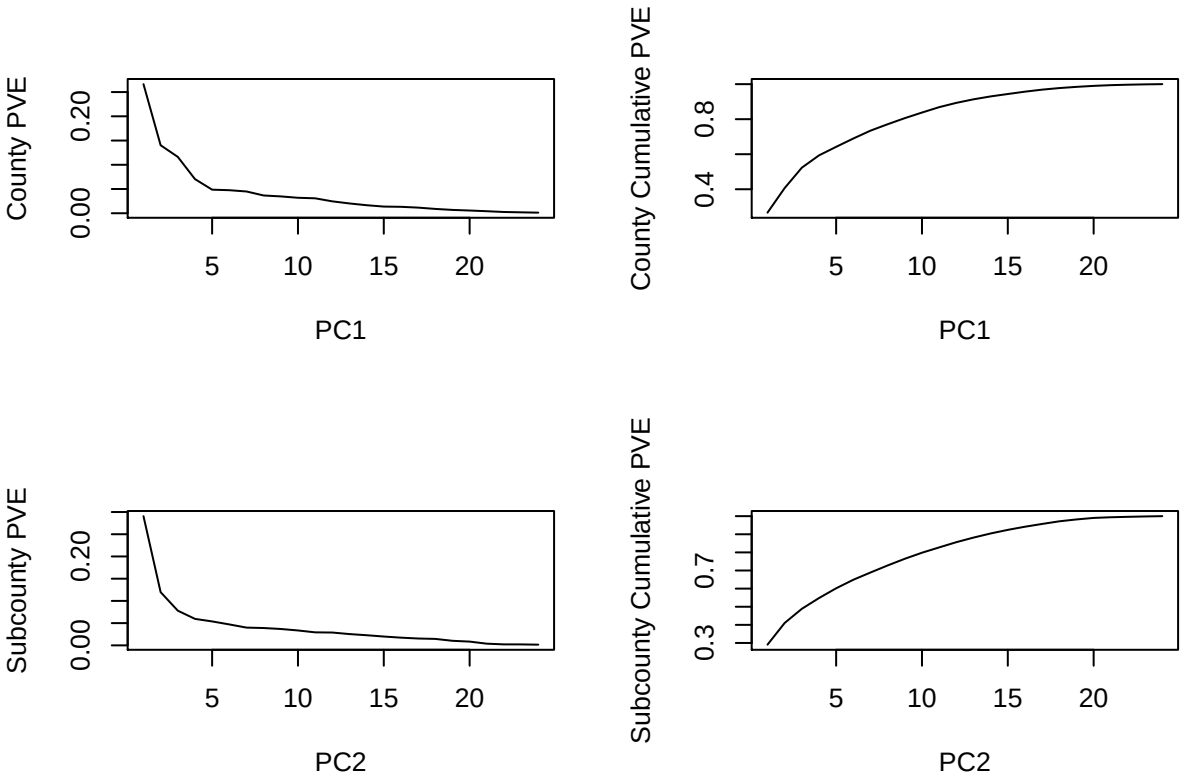
We need to center and scale the features before running PCA since some of varibles in the dataset are not comparable. Standardization is important because it puts an emphasis on variables with higher variances than those low variances to help with identifying the right principal components. For example, we have varible 'Men' as percentage and varible 'Income' with 5-digit large number. If we do not center and scale the data, most of the principal components that we observed would be driven by the Income variable, since it has by far the largest mean and variance.

In the county level PCA, varible 'IncomePerCap','ChildPoverty', and 'Poverty' has the largest absolute values of the first principal component. 'ChildPoverty' and 'Poverty' are negative.

In the sub-county level PCA, varible 'IncomePerCap','Professional', and 'Income' has the largest absolute values of the first principal component. All three are positive.

The positive and negative signs refer to whether or not the features have a positive or negative correlation with one another within the Principal Component.

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




```
## [1] 13
```

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

13 is the minimum number of PCs needed to capture 90% of the variance for the county. 14 is the minimum number of PCs needed to capture 90% of the variance for the subcounty.

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

```
## [1] 3
```

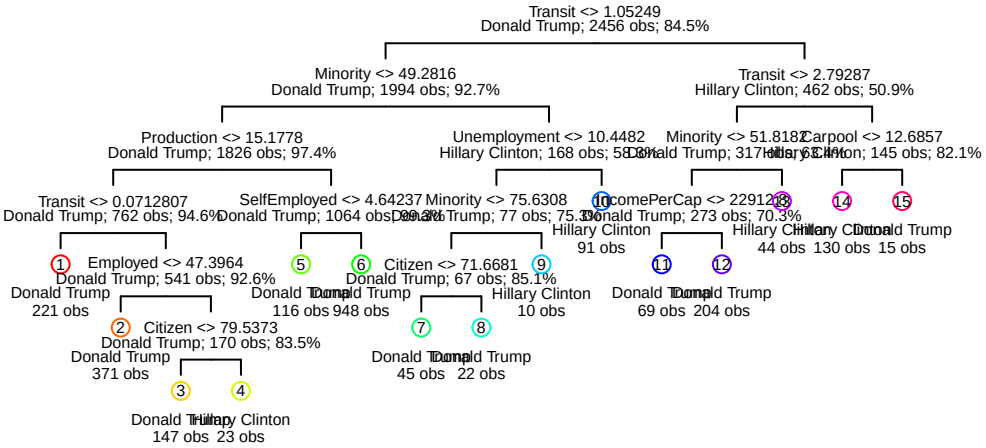
```
## [1] 4
```

15. Here we have to scale the data before clustering. If we do not scale it, the result may be biased towards those variabels with larger scale. The hierachical clustering using the first 5 principal components seemed to put San Mateo County in a more appropriate cluster. The first cluster contains almost all data points in GROUP 1, although San Mateo is in group 3, it actually gives us no information about San Mateo. However, the principal component model puts San Mateo to cluster 2 Group 4, which is a group with 155 out of 3218 observations, and the cluster itself looks variable. It means the group 4 must be away from other clusters to be distinguished so that it must have some special properties. That is why San Mateo is put in a more appropriate cluster in the principal component model because it contains some useful information about this county.

Classification

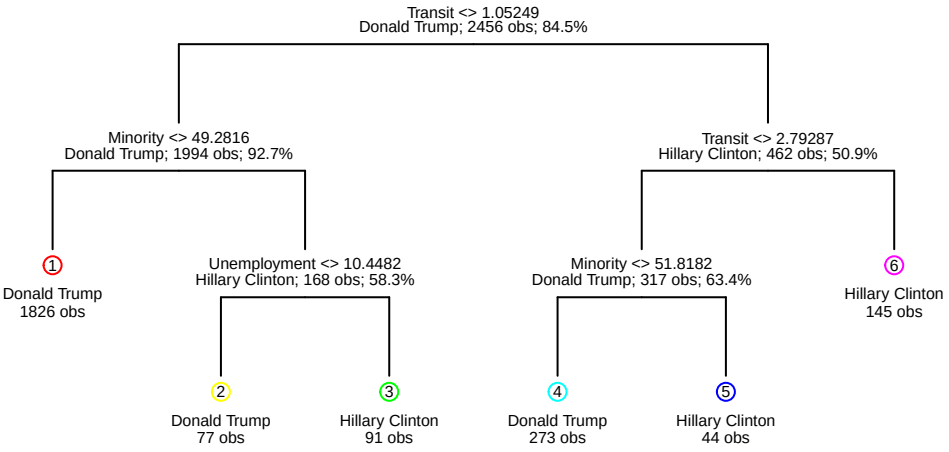
16. Decision tree: train a decision tree by cv.tree(). Visualize the trees before and after pruning. Intepret and discuss the results of the decision tree analysis. Use this plot to tell a story about voting behavior in the US

Unpruned Tree



Total classified correct = 92.9 %

Pruned Tree



Total classified correct = 92.1 %

	train.error	test.error
tree	0.0793974	0.0731707
logistic	NA	NA
lasso	NA	NA

We can see from the plots that the variable 'transit' is the most important variable, then, the minority status is the second most important one. From the second plot, we can notice that Trump is preferred in the regions that have lower minority rate.

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

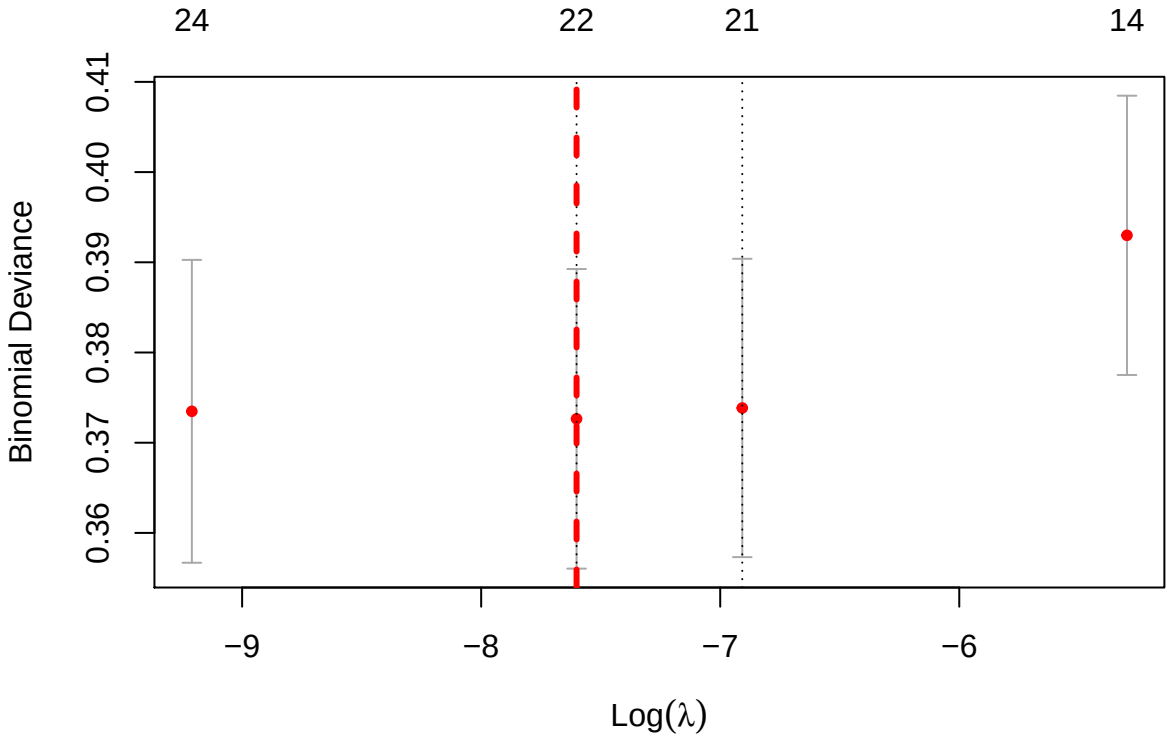
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = factor(candidate) ~ ., family = "binomial", data = trn.cl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2513  -0.2638  -0.1133  -0.0420   3.5118
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.174e+01  7.177e+00  -5.816 6.03e-09 ***
## Men              9.332e-02  4.844e-02   1.926  0.05406 .
## Citizen          1.193e-01  2.762e-02   4.319 1.57e-05 ***
## Income          -8.212e-05  2.715e-05  -3.025  0.00248 **
## IncomeErr       -1.855e-05  6.376e-05  -0.291  0.77109
## IncomePerCap     2.678e-04  6.707e-05   3.993 6.52e-05 ***
## IncomePerCapErr -3.952e-04  1.715e-04  -2.305  0.02118 *
## Poverty          4.305e-02  4.084e-02   1.054  0.29182
## ChildPoverty    -1.096e-02  2.466e-02  -0.445  0.65662
## Professional     2.934e-01  3.802e-02   7.716 1.20e-14 ***
## Service          3.396e-01  4.718e-02   7.197 6.15e-13 ***
## Office           9.849e-02  4.356e-02   2.261  0.02375 *
## Production       1.705e-01  4.088e-02   4.171 3.03e-05 ***
## Drive           -2.061e-01  4.641e-02  -4.440 9.00e-06 ***
## Carpool         -1.622e-01  5.921e-02  -2.739  0.00616 **
## Transit          1.157e-01  9.265e-02   1.249  0.21182
## OtherTransp     -5.614e-02  9.482e-02  -0.592  0.55382
## WorkAtHome      -1.681e-01  7.243e-02  -2.321  0.02027 *
## MeanCommute      5.513e-02  2.408e-02   2.290  0.02205 *
## Employed         2.013e-01  3.352e-02   6.005 1.92e-09 ***
## PrivateWork      9.820e-02  2.086e-02   4.706 2.52e-06 ***
## SelfEmployed     1.866e-02  4.665e-02   0.400  0.68918
## FamilyWork      -8.567e-01  3.760e-01  -2.278  0.02271 *
## Unemployment     2.163e-01  3.999e-02   5.411 6.28e-08 ***
## Minority         1.333e-01  9.561e-03  13.945  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2119.56  on 2455  degrees of freedom
## Residual deviance:  861.16  on 2431  degrees of freedom
## AIC: 911.16
##
## Number of Fisher Scoring iterations: 7
```

	train.error	test.error
tree	0.0793974	0.0731707
logistic	0.0700326	0.0666667
lasso	NA	NA

Significant variables are Citizen, Income, IncomePerCap, IncomePerCapErr, Professional, Service, Production, Office, Carpool, Drive, WorkAtHome, MeanCommute, Employed, PrivateWork, FamilyWork, Unemployment, and Minority at 95% confidence level. This is not consistent with what I got in the last problem.Because Transit is not significant in Logistic model. For example, the coefficient of citizen is 1.193e-01, which means if the percentage of citizens increases by 1 percent, the logit will increase by 1.193e-01, which corresponds to a multiplicative change in the odds of $e^{1.193e-01}$. Also, the coefficient of Production is 1.705e-01, which means if the percentage of Production increase by 1 percent, the logit will decrease by 1.705e-01, which corresponds to a mutiplicative change in the odds of $e^{-1.705e-01}$.

18.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. What is the optimal value of in cross validation? What are the non-zero coefficients in the LASSO regression for the optimal value of ? How do they compare to the unpenalized logistic regression?

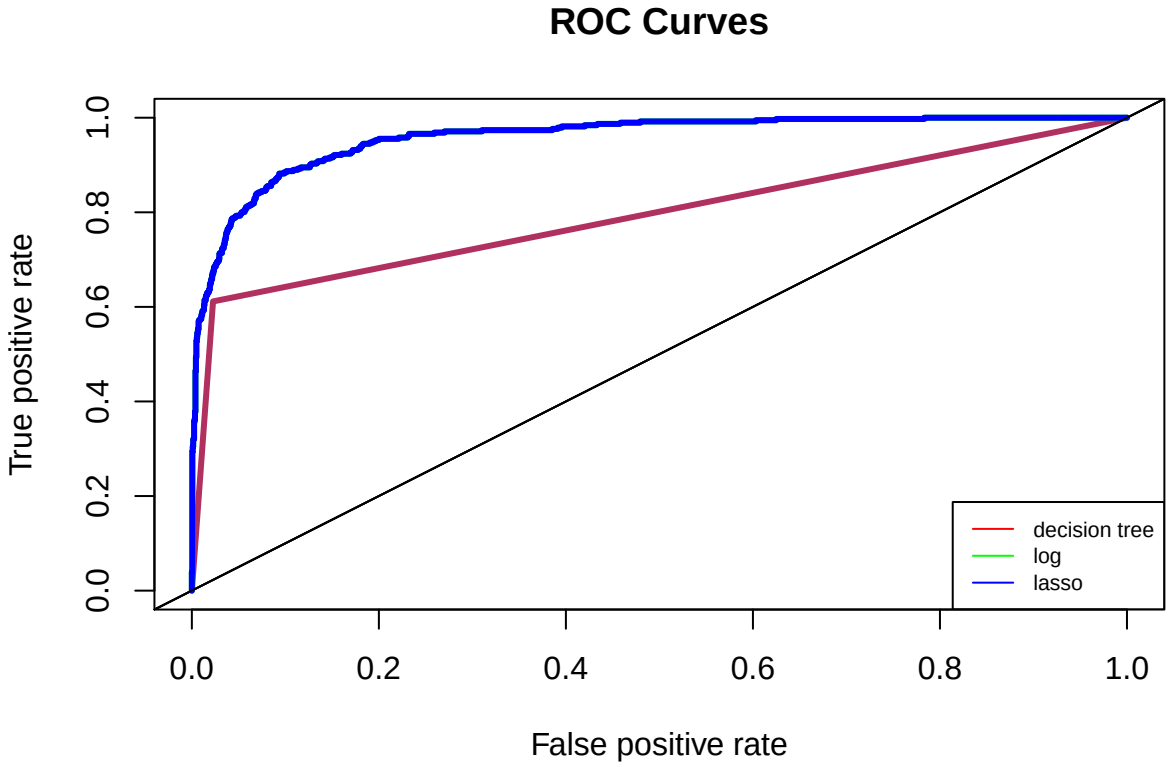



```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
```

	train.error	test.error
tree	0.0793974	0.0731707
logistic	0.0700326	0.0666667
lasso	0.0712541	0.0699187

The optimal value for lambda is 5e-04. All the coefficients in the LASSO regression for the optimal value of lambda are nonzero, except 'SelfEmployed' and 'ChildPoverty'. They are generally smaller in magnitude than the unpenalized logistic regression.

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




```
## [[1]]
## [1] 0.794449

## [[1]]
## [1] 0.9576941

## [[1]]
## [1] 0.9576941
```

Decision trees are very simple to use but they do not have the best accuracy. Since they also have high variance and tend to overfit, any small changes can lead to a completely different tree. This form of classification will only work well if the data can easily be split into rectangular regions. Logistic regression is good for classifying between two different values. In this class, we are classifying the election result for each county. However, if the data is linear or has complete separation, it will be hard to classify. Lasso Regression is most useful when some predictors are redundant and can be removed. Much like all regularization methods as well as logistic regression, Lasso Regression tends to have a lower variance and does not overfit as much. But it ignores non significant variables, that may be problematic because we'll never know how interesting or uninteresting they are. Decision trees perform poorly with only .794449 while logistic and lasso regression perform pretty much the same with values 0.9576941 and 0.9576941 Anyway, we won't use decision trees for classifying election results.

Taking it further

For our final open question, we decided to explore four more classification methods: KNN, Random Forest, Boosting, and SVM. Our goal is to find a model with the smallest error, using the same training and testing

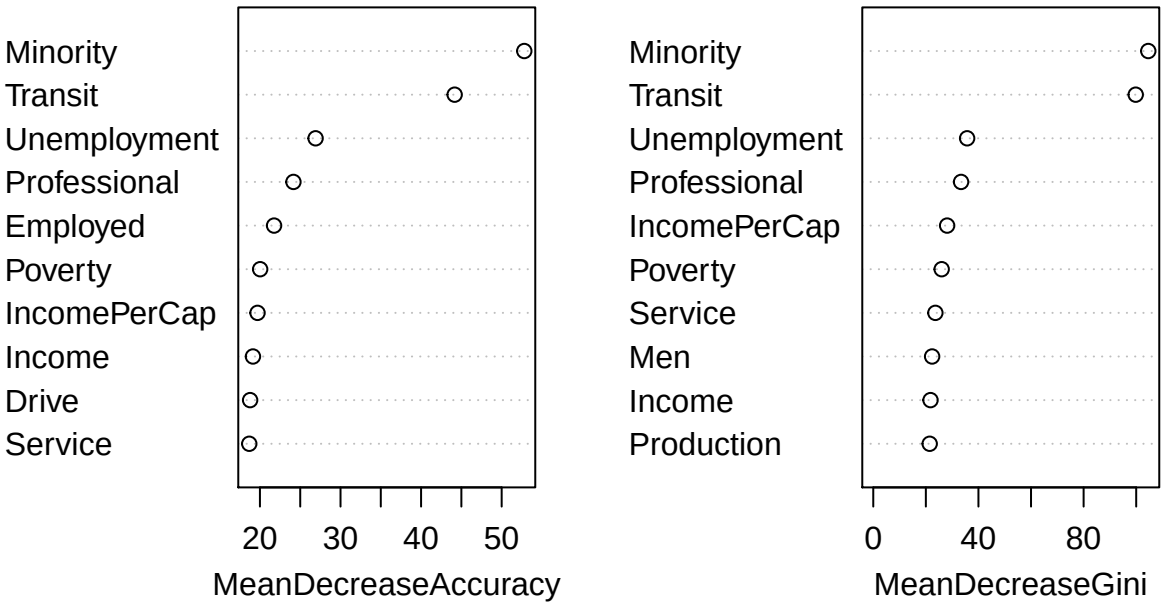
dataset in question 16-18. First, we fit knn model using 10-fold Cross-Validation. During Cross Validation, we found 32 is the best number of neighbors with the smallest error rate in the training dataset.Then, we train a 32-NN classifier, and calculate the test error rate.

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------


##
## Attaching package: 'plyr'

## The following object is masked from 'package:maps':
##
##     ozone

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact

## [1] 24
```

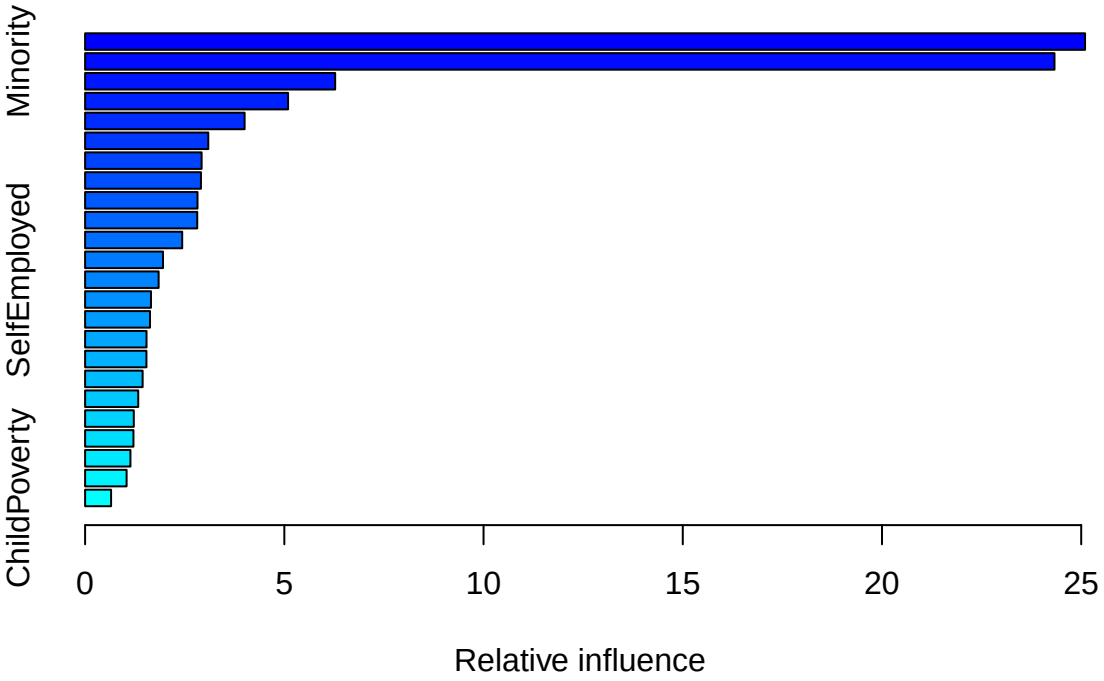
	train.error	test.error
knn	0.1364007	0.1349593
random forest	NA	NA
boosting	NA	NA
SVM	NA	NA

Both of training error rate and test error rate of KNN is significantly higher compare to previous models like logistic model. This may be due to the fact that KNN method do not perform well in high dimensions. As varible increases, closest neighbor usually become "far away". Next, we examine random forest model. The random forest model has the lowest training error rate of 0.081% and test error rate of 4.23%. In the above graph we also notice that 'Transit','White',and 'Minority' are the most significant variable in decreasing the Gini impurity. In context of the election, this makes sense due to the fact that a large proportion of white voters tend to vote for Donald Trump while Minority tend to voted for Hillary Clinton. Also, white voters make up a higher proportion in the Republican than in the Democrat.

	train.error	test.error
knn	0.1364007	0.1349593
random forest	0.0008143	0.0422764
boosting	NA	NA
SVM	NA	NA



In the boosting model, we also gain a low test error of 5.52% Notice that White, Transit, and Minority are the more influential variables in this model, which argee with the random forest model. This indicates that race is a important facter in this election.




```
##                              var    rel.inf
## Transit                  Transit 25.0962809
## Minority                Minority 24.3294318
```

```
## Unemployment       Unemployment  6.2759168
## Professional       Professional  5.0940903
## Employed               Employed  4.0029305
## Service                 Service  3.0911173
## Men                         Men  2.9244068
## IncomePerCapErr IncomePerCapErr  2.9070937
## Citizen                 Citizen  2.8210461
## MeanCommute         MeanCommute  2.8140863
## Poverty                 Poverty  2.4366742
## Office                   Office  1.9553081
## SelfEmployed       SelfEmployed  1.8455583
## IncomePerCap       IncomePerCap  1.6534610
## OtherTransp         OtherTransp  1.6288525
## WorkAtHome           WorkAtHome  1.5411518
## Drive                     Drive  1.5394963
## Production           Production  1.4434411
## PrivateWork         PrivateWork  1.3332091
## Carpool                 Carpool  1.2221266
## Income                   Income  1.2133641
## IncomeErr             IncomeErr  1.1376342
## FamilyWork           FamilyWork  1.0412363
## ChildPoverty       ChildPoverty  0.6520859
```

	train.error	test.error
knn	0.1364007	0.1349593
random forest	0.0008143	0.0422764
boosting	0.0077362	0.0552846
SVM	NA	NA

However, like KNN model, the training error rate and test error rate of SVM is significantly higher compare to other models. SVM also do not perform well in higher dimension, especially with distance involved.

	train.error	test.error
knn	0.1364007	0.1349593
random forest	0.0008143	0.0422764
boosting	0.0077362	0.0552846
SVM	0.0008143	0.1252033

Overall, random forest is the best model so far to efficently predict election winner, and we notice that race is a inportant factor in this race.