Pstat 131 Project

Mubai Liu & Hongshan Lin
8690091 & 9913807

15 June 2018

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

Many reasons may lead to the difficulty of predicting the voter behavior and election forecasting. The first one may be the number of voters for the 2016's polls. While we were expecting that the equal number of Democrats and Republicians, it turned out that the Republican voters were much higher than the Democratic voters.The second reason may be the decision changing for the voters. It depended on whether or not the voter is a minority, and the income earned, and the gender. We should analysis all the factor that may influence an individual's vote. Also,it turned out that many voters changed their vote in the week that leads up to voting. And the last reason may be the unpredictable future events that will happen. The society may change their attitude due to the news they've found on the TV or website, which is not predcitable.

2 Although Nate Silver predicted that Clinton would win 2016, he gave Trump higher odds than most. What is unique about Nate Silver's methodology?

Compared to the usual approach which will take the maximum probabilty as the outcome, Nate Silver's approach takes a full range of possibilities instead of just taking one maximum. For example, he calculated the possibilites of different dates of support and after calculation, he utilized the whole set of possibilities to model the shift in the polling numbers and thus get the desire result. He also looked at both the nation-level and tste-level votes. The whole idea of his approach is based on the Bayes' Theorem.

3 Discuss why analysts believe predictions were less accurate in 2016. Can anything be done to make future predictions better? What are some challenges for predicting future elections? How do you think journalists communicate results of election forecasting models to a general audience?

In the 2016, as we mentioned in the first question, the media plays a huge role for deciding which side of voters will be, the media overstated Clinton's lead, especially in the Costal state. The news will lead to many voters choose Clintion, and feel uncomfortable with Trump. It is the same situation for the prediction of voting. So if we want to make the future prediction more precise, we might want to find out the potential news in the policitician. People should able to balance with the media's instigate and their own thought. The challenges are clear because as people growing, their experience and knowledge is also growing, so next time maybe they will stick with their choice all the time instead of changing their decision last second. We think that journalists' action may also lead to some violations to the model that we are trying to predict. It could cause some people to change their mind once again.

Data wrangling

4 Remove summary rows from election.raw data: i.e., Federal-level summary into a election_federal. State-level summary into a election_state. Only county-level data is to be in election.

Here are the first few rows in the 'election.raw' data.

county	fips	candidate	state	votes
NA	US	Donald Trump	US	62984825
NA	US	Hillary Clinton	US	65853516
NA	US	Gary Johnson	US	4489221
NA	US	Jill Stein	US	1429596
NA	US	Evan McMullin	US	510002
NA	US	Darrell Castle	US	186545

Here are the first few rows of federal-level summary

county	fips	candidate	state	votes
NA	US	Donald Trump	US	62984825
NA	US	Hillary Clinton	US	65853516
NA	US	Gary Johnson	US	4489221
NA	US	Jill Stein	US	1429596
NA	US	Evan McMullin	US	510002
NA	US	Darrell Castle	US	186545

Here are the first few rows of state-level summary

county	fips	candidate	state	votes
NA	CA	Hillary Clinton	CA	8753788
NA	CA	Donald Trump	CA	4483810
NA	CA	Gary Johnson	CA	478500
NA	CA	Jill Stein	CA	278657
NA	CA	Gloria La Riva	CA	66101
NA	FL	Donald Trump	FL	4617886

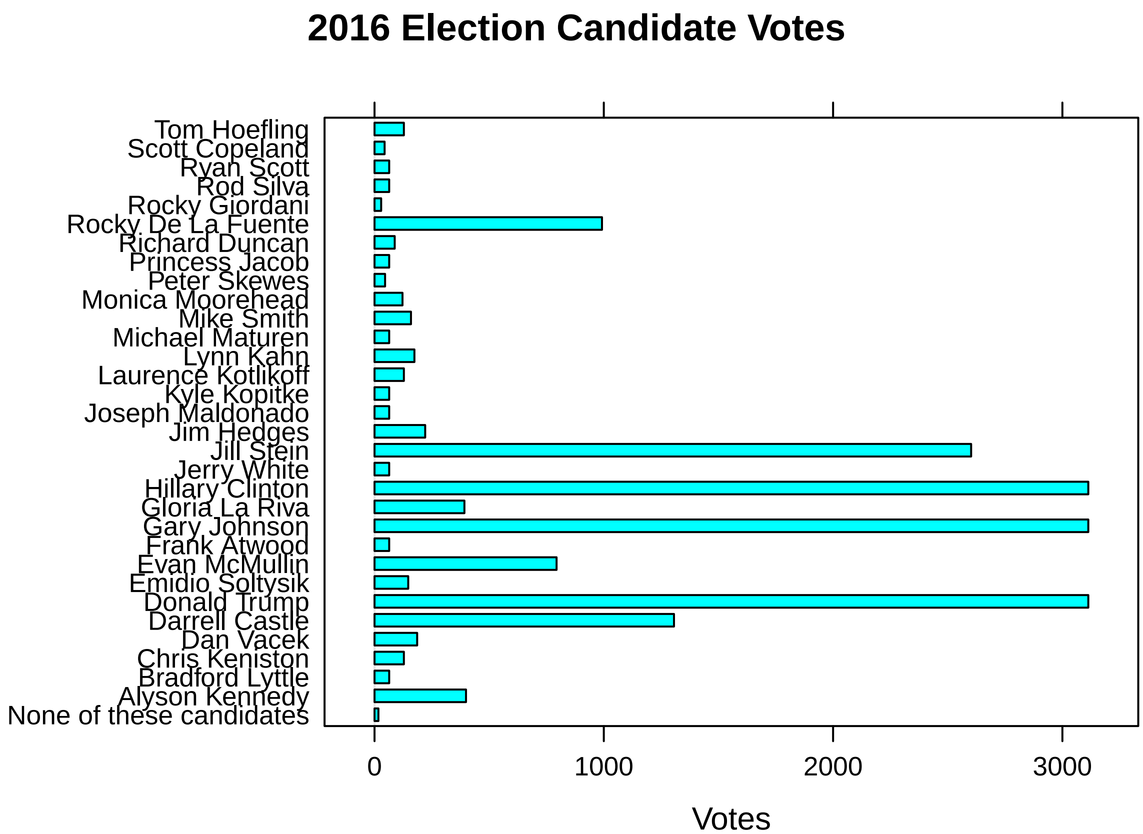
Here are the first few rows of county-level data in election

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
Cook County	17031	Hillary Clinton	IL	1611946

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

```
## [1] 32
```

Thus there are 32 presidential candidates were there in the 2016 election, but only $32 - 1 = 31$ candidates were named. Here is the bar chart:



6 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).

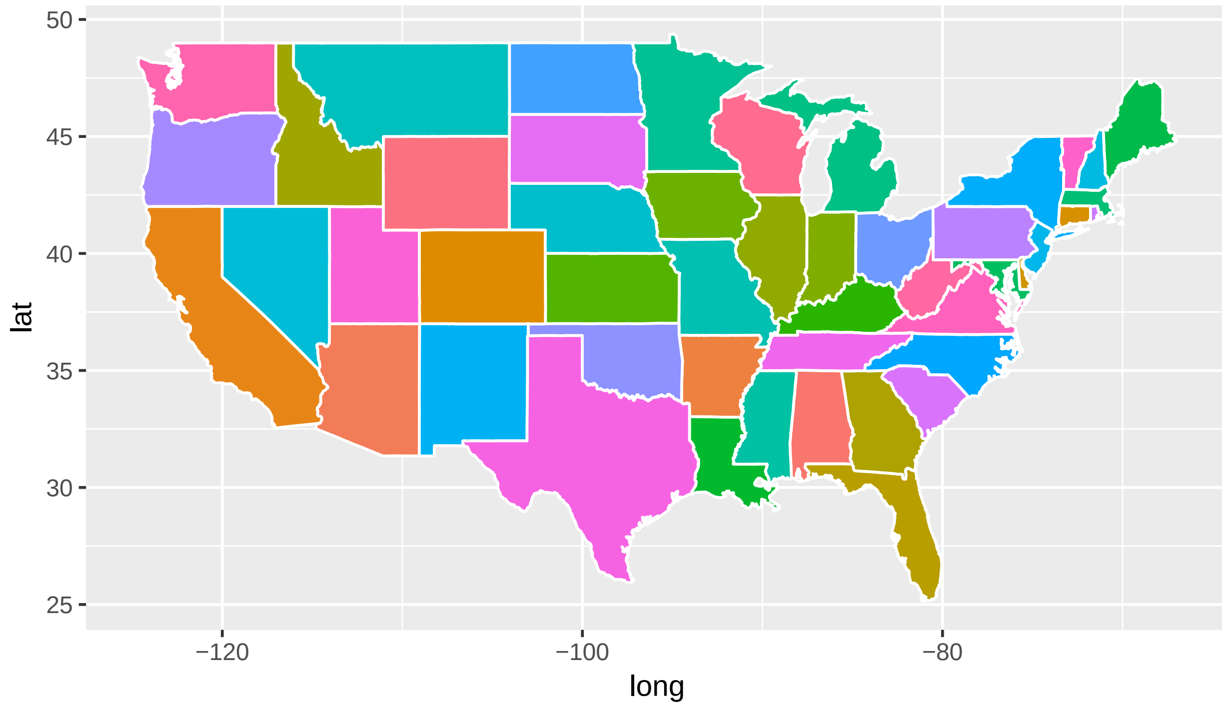
County winner:

county	fips	candidate	state	votes	total	pct
Los Angeles County	6037	Hillary Clinton	CA	2464364	3421533	0.7202514
Cook County	17031	Hillary Clinton	IL	1611946	2156395	0.7475189
Maricopa County	4013	Donald Trump	AZ	747361	1536743	0.4863279
Harris County	48201	Hillary Clinton	TX	707914	1305434	0.5422825
San Diego County	6073	Hillary Clinton	CA	735476	1291078	0.5696604
Orange County	6059	Hillary Clinton	CA	609961	1186203	0.5142130

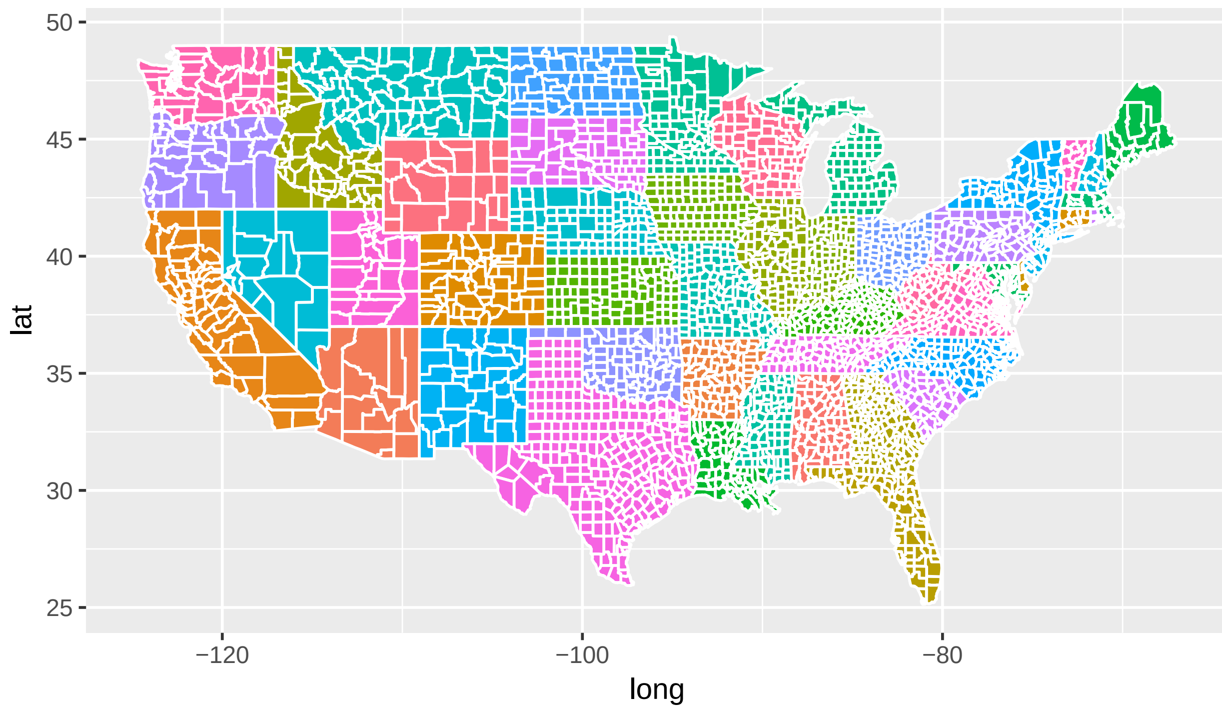
State winner:

candidate	state	votes	VotesInState	pct
Hillary Clinton	CA	8753788	14060856	0.6225644
Hillary Clinton	IL	3090729	5523142	0.5595962
Donald Trump	AZ	1252401	2554240	0.4903224
Donald Trump	TX	4685047	8917965	0.5253493
Hillary Clinton	WA	1742718	3209214	0.5430358
Donald Trump	FL	4617886	9419886	0.4902274

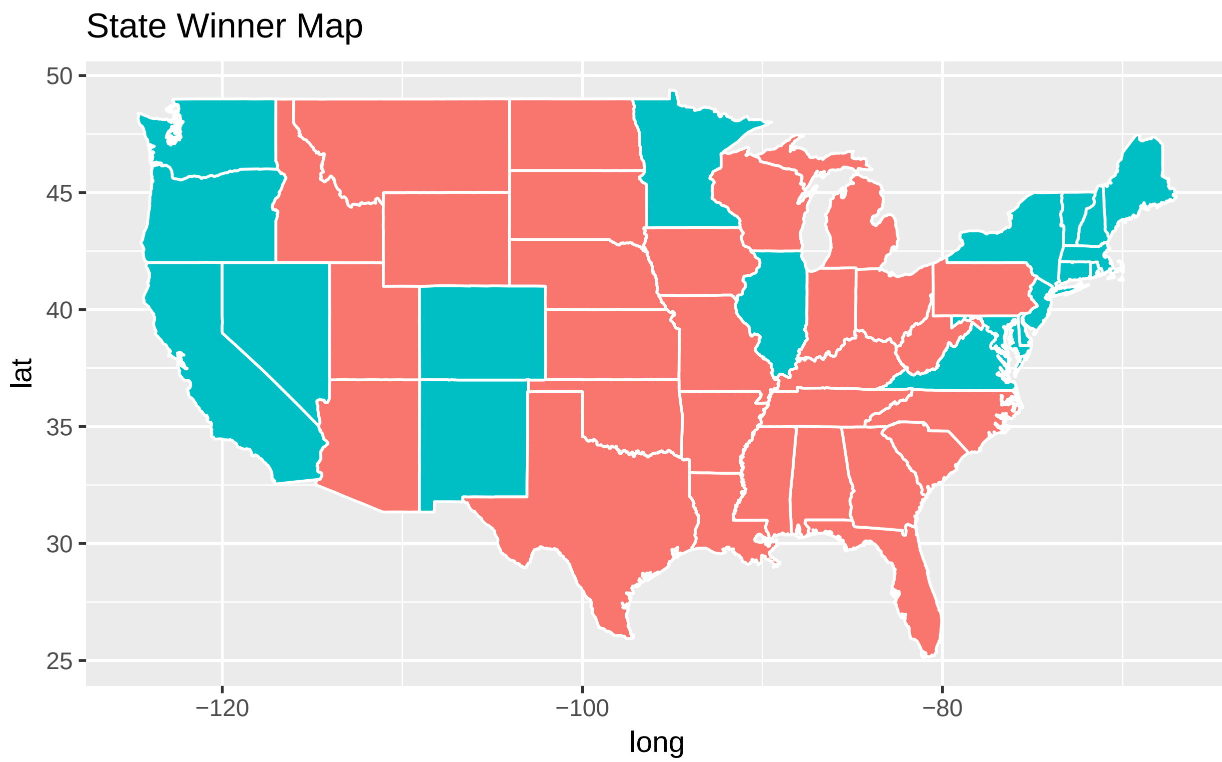
Visualization



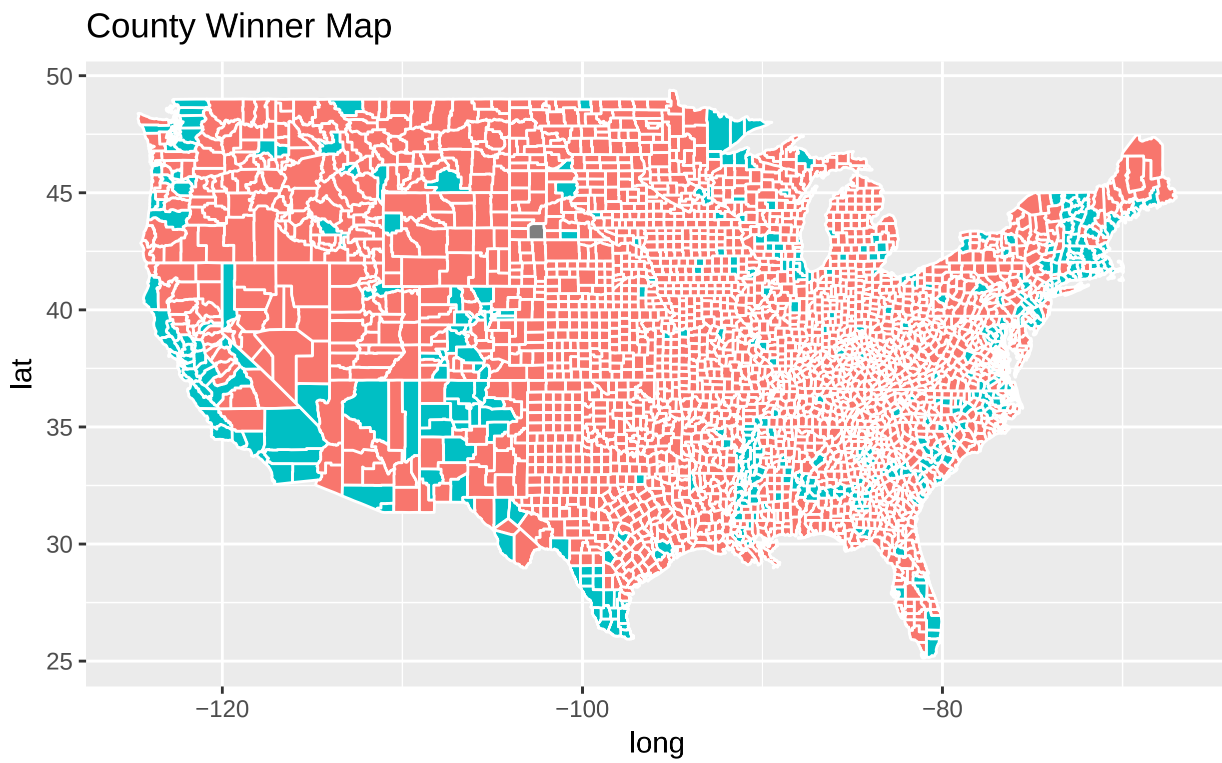
7 Draw county-level map. Color by county.



8 Color the map by the winning candidate for each state

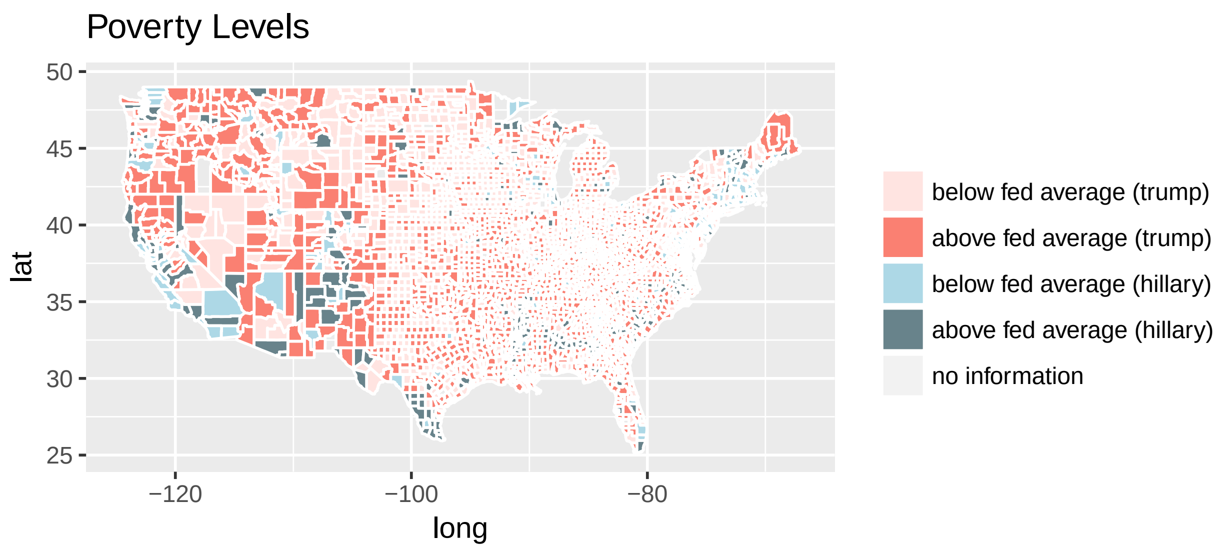


9 Color the map by the winning candidate for each state



10 Create a visualization of your choice using census data.

The following will show the map that visualizes the poverty level of each county, where the darker color of each group shows more federal poverty (above average) while the lighter color of each group represents less federal pvoerty(below average) in that region. For Trump we use the orange and blue for Hillary.As the result, we can see that Hillary has fewer ligiher color county with average lower rate of poverty compared to Trump. Therefore, the demographs play a big role in the elecion. It shows different control variable driving different voting preferences.



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

State	County	Men	White	Minority	Citizen	Income	IncomeErr	IncomePerCap	IncomePerCa
Alabama	Autauga	48.43266	75.78823	22.53687	73.74912	51696.29	7771.009	24974.50	3433
Alabama	Baldwin	48.84866	83.10262	15.21426	75.69406	51074.36	8745.050	27316.84	3803
Alabama	Barbour	53.82816	46.23159	51.94382	76.91222	32959.30	6031.065	16824.22	2430
Alabama	Bibb	53.41090	74.49989	24.16597	77.39781	38886.63	5662.358	18430.99	3073
Alabama	Blount	49.40565	87.85385	10.59474	73.37550	46237.97	8695.786	20532.27	2052
Alabama	Bullock	53.00618	22.19918	76.53587	75.45420	33292.69	9000.345	17579.57	3110

Dimensionality reduction

12 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 features with the largest absolute values in the loadings matrix?

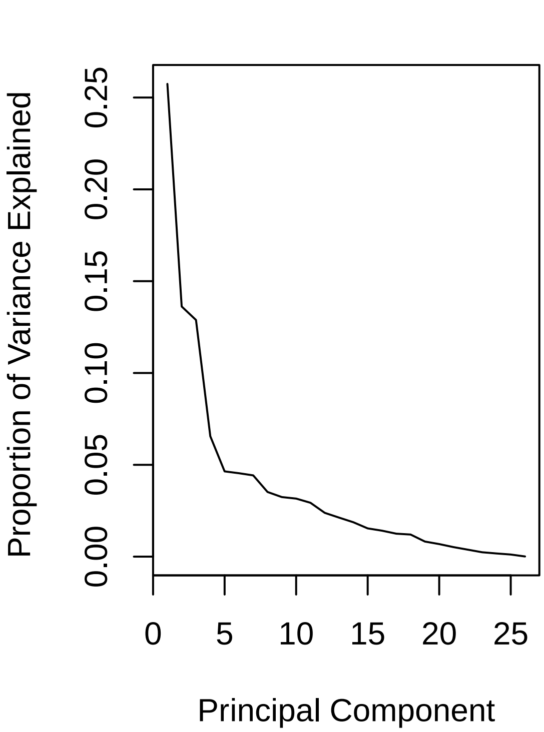
We choose center=TRUE and scale=TRUE because it puts all variables on the same scale and we don't have to worry about the units of the variables. And especially for the mixed types. The largest absolute values in the loading martrix is the first entry in the following r output. The the largest absolute values in ct.pc is the IncomPerCap = 0.350767, and subct.pc is Income Err 0.314502186. We will see the features with largest absolute values in the loading matrix as the first entry in the following r output.

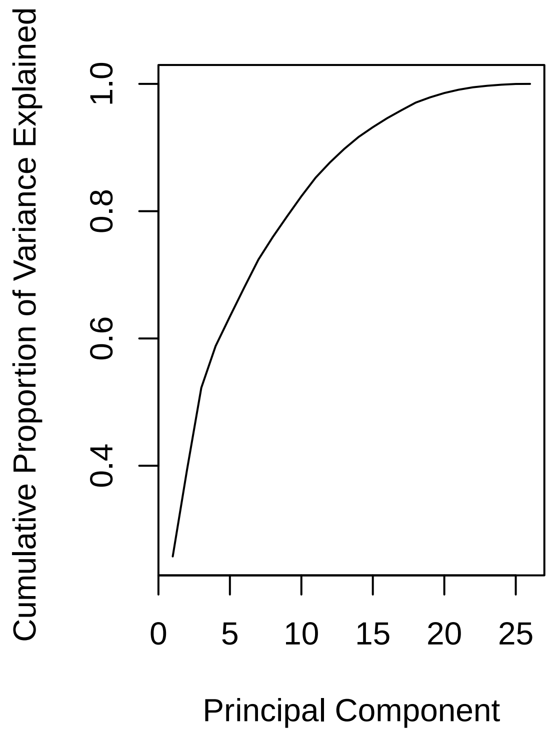
```
##  [1] 0.3530767161 0.3421530456 0.3405832434 0.3274293648 0.3225865807
##  [6] 0.3145021861 0.3087990253 0.2931957490 0.2926767584 0.2889832088
## [11] 0.2876313774 0.2777578140 0.2520238157 0.2452918361 0.2396236777
## [16] 0.2212851691 0.2176990922 0.2157214952 0.2088076127 0.2074056079
## [21] 0.2071183115 0.1969492637 0.1929373697 0.1819627714 0.1801805293
## [26] 0.1738246072 0.1724756889 0.1588719555 0.1434392860 0.1389016724
## [31] 0.1354888230 0.1211691321 0.1094149984 0.0949814857 0.0938983015
## [36] 0.0771785385 0.0765359491 0.0624743558 0.0590834460 0.0589372390
## [41] 0.0589278307 0.0560237641 0.0555820911 0.0462881560 0.0405821706
## [46] 0.0368556038 0.0303197605 0.0115397934 0.0086377486 0.0048240359
## [51] 0.0029776839 0.0003126037
```

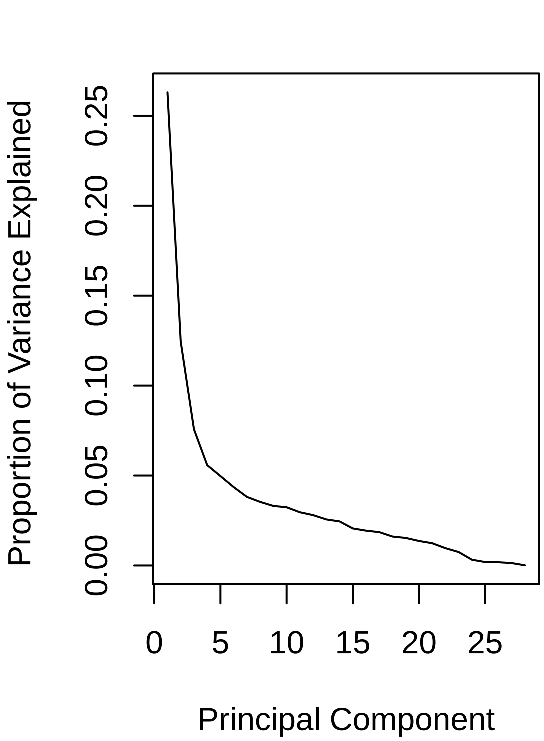
```
##    IncomePerCap    ChildPoverty         Poverty        Employed
##    0.3530767161    0.3421530456    0.3405832434    0.3274293648
##          Income     Unemployment    Professional        Minority
##    0.3225865807    0.2876313774    0.2520238157    0.2212851691
##           White IncomePerCapErr         Service       IncomeErr
##    0.2176990922    0.1969492637    0.1801805293    0.1738246072
##      WorkAtHome      Production           Drive     SelfEmployed
##    0.1724756889    0.1211691321    0.0949814857    0.0938983015
##         Carpool         Transit     CountyTotal     PrivateWork
##    0.0771785385    0.0765359491    0.0624743558    0.0589372390
##     MeanCommute      FamilyWork          Office     OtherTransp
##    0.0555820911    0.0462881560    0.0115397934    0.0086377486
##             Men         Citizen
##    0.0048240359    0.0003126037
```

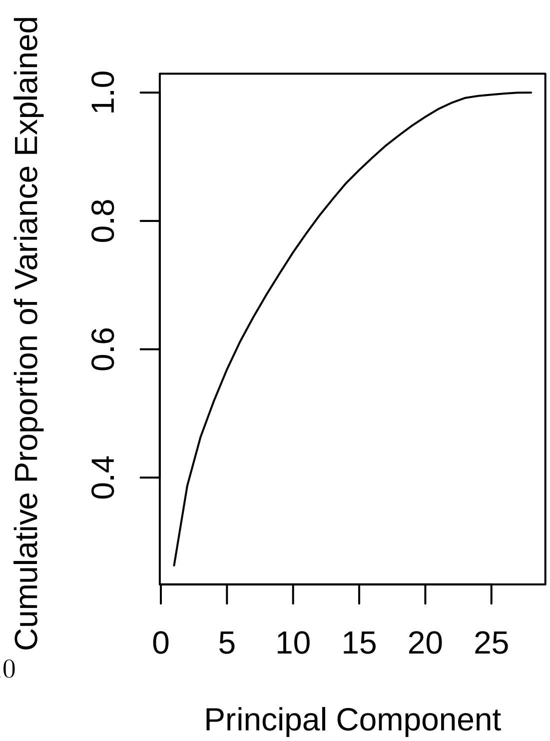
```
##       IncomeErr     SelfEmployed     CountyTotal           White
##     0.314502186     0.308799025     0.293195749     0.292676758
##        Minority         Transit          Office         Citizen
##     0.288983209     0.277757814     0.245291836     0.239623678
##      WorkAtHome      FamilyWork          Income IncomePerCapErr
##     0.215721495     0.208807613     0.207405608     0.207118312
##     MeanCommute     PrivateWork     Unemployment      Production
##     0.192937370     0.181962771     0.158871955     0.143439286
##    IncomePerCap             Men     Professional     OtherTransp
##     0.138901672     0.135488823     0.109414998     0.059083446
##         Service         Poverty     ChildPoverty         Carpool
##     0.058927831     0.056023764     0.040582171     0.036855604
##           Drive         Employed
##     0.030319761     0.002977684
```

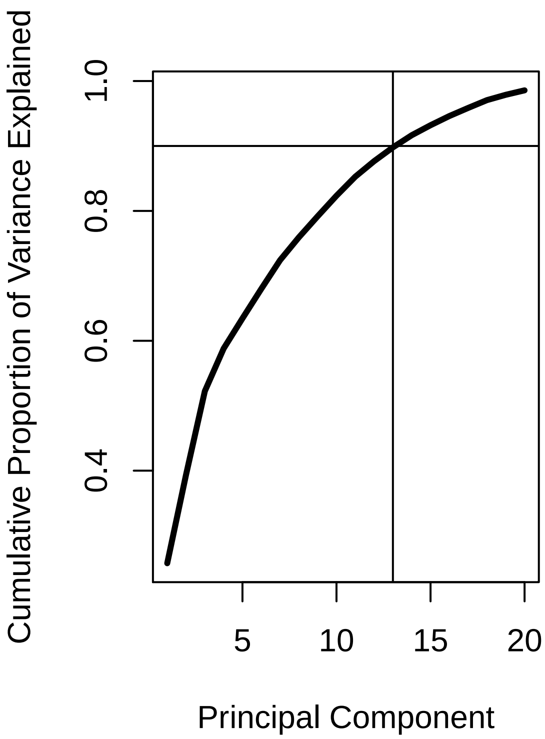
13 Determine the number of minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses. Plot proportion of variance explained (PVE) and cumulative PVE for both county and sub-county analyses.

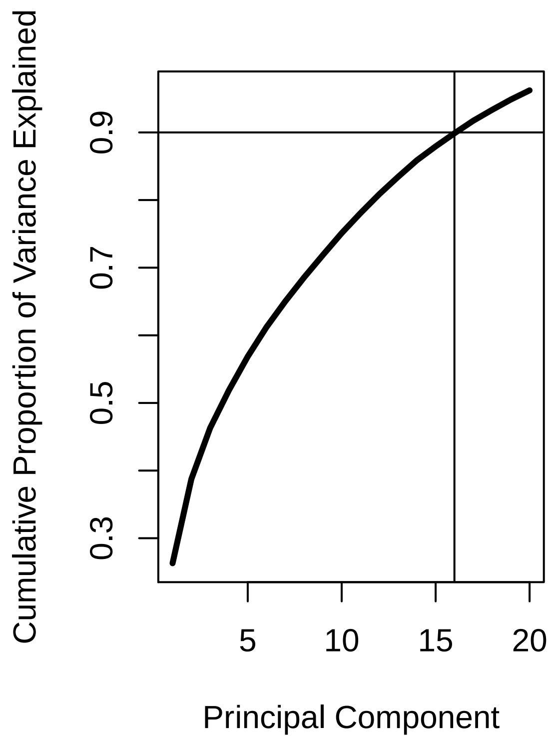












14 With census.ct, perform hierarchical clustering with complete linkage. Cut the tree to partition the observations into 10 clusters. Re-run the hierarchical clustering algorithm using the first 5 principal components of ct.pc as inputs instead of the original features. Compare and contrast the results. 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.

```
## clusters.whole
##    1    2    3    4    5    6    7    8    9   10
## 2632  501    6    7    5    1   11   13   38    4
```

```
## clusters.five
##    1    2    3    4    5    6    7    8    9   10
## 2441  525   97    6    8   31    5   18    7   80
```

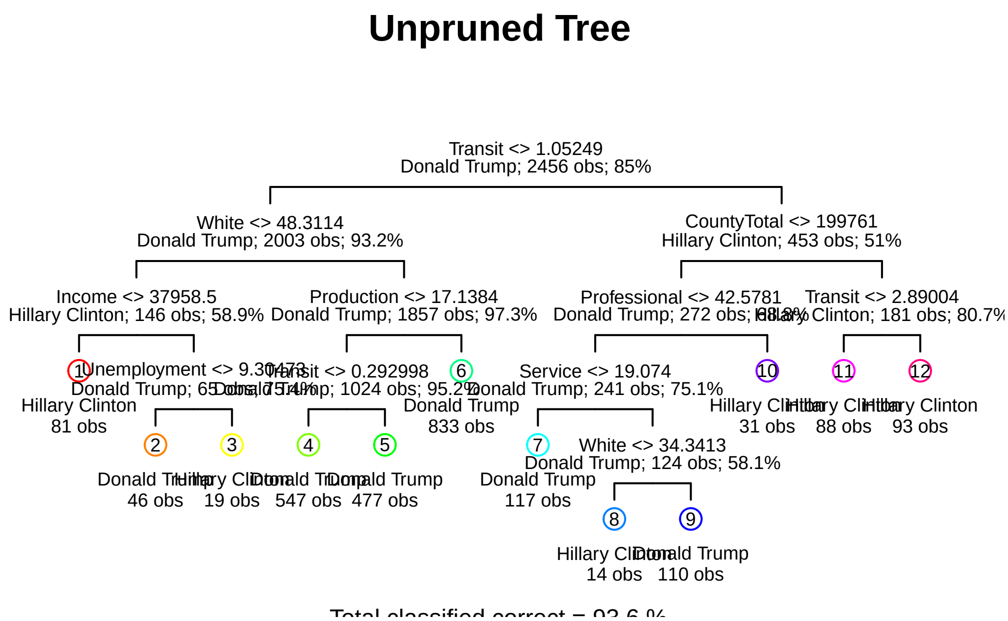
```
## [1] 2
```

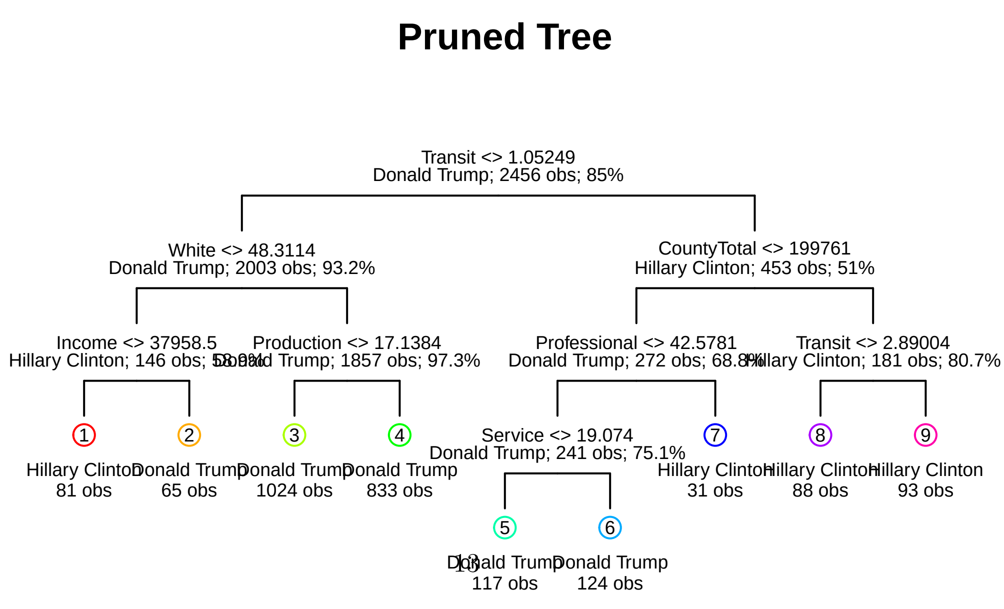
```
## [1] 1
```

It turns out that when we use different number of principal components as inpust we will position San Mateo in different clusters. For example, at first San Mateo is placed into the cluster 2 but when we changing the PCs to PC1-PC5, it changes the clusters to 1. It appears to be more in line with cluster guidelines when we conside the original data. We can observe that there are less Alabama counties inside the cluster 2 with San Mateo, but consider the cluster 1 we can see that many differing counties are in this cluster. This is most likely due to the fact that PC1-PC5 won't describe variance in the data census.ct, thus we have this disagreement in the clustering.

Classification

15 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. Interpret 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?)





```
##                       train.error test.error
## tree                   0.07288274 0.09120521
## Logistic Regression            NA         NA
## LASSO                          NA         NA
```

We prune the tree to mininize the misclassification error, and to prevent overfitting. We redeuce the node from 12 to 9.The transit as a primary split and shows many times after, plays an important role in thte election. As the result, in the nominating contests so far, Senator Clition has won the vast majority of countries with less white and low income. Sencator Trump as a commanding lead in the majority of countries with poeple who rearely use public transportation, and less employed people in production. In the county total, people employed in professional and service job and white are like to vote Trump.

16 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.

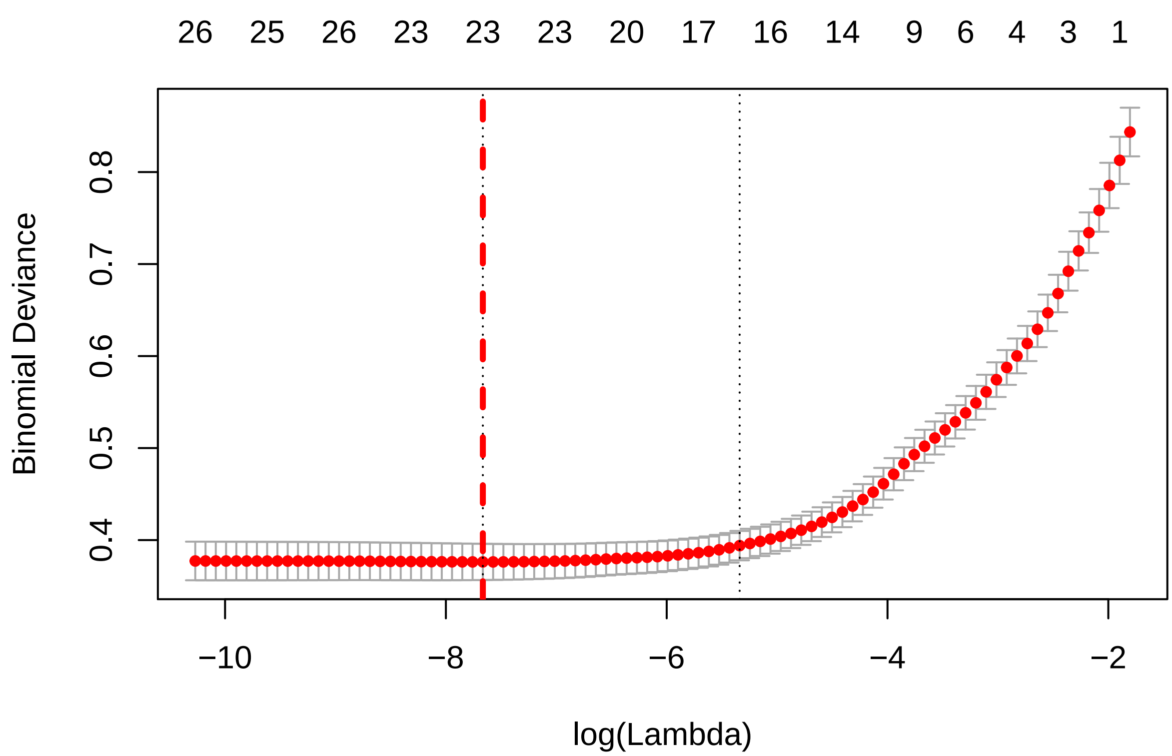
```
##
## Call:
## glm(formula = candidate ~ ., family = binomial, data = trn.cl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7362  -0.2705  -0.1133  -0.0407   3.5782
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.158e+01  9.701e+00  -1.193 0.232686
## Men             8.346e-02  5.386e-02   1.550 0.121229
## White          -2.147e-01  6.550e-02  -3.278 0.001047 **
## Minority       -8.369e-02  6.274e-02  -1.334 0.182229
## Citizen         1.069e-01  3.049e-02   3.508 0.000452 ***
## Income         -7.558e-05  2.752e-05  -2.747 0.006016 **
## IncomeErr      -3.703e-05  6.269e-05  -0.591 0.554786
## IncomePerCap    2.669e-04  6.717e-05   3.974 7.07e-05 ***
## IncomePerCapErr -2.759e-04  1.308e-04  -2.109 0.034904 *
## Poverty         2.083e-02  4.110e-02   0.507 0.612267
## ChildPoverty   -7.147e-03  2.551e-02  -0.280 0.779357
## Professional    2.739e-01  3.972e-02   6.897 5.32e-12 ***
## Service         3.590e-01  4.953e-02   7.248 4.23e-13 ***
## Office          9.549e-02  4.801e-02   1.989 0.046688 *
## Production      1.811e-01  4.317e-02   4.196 2.72e-05 ***
## Drive          -2.542e-01  5.393e-02  -4.714 2.43e-06 ***
## Carpool        -2.441e-01  6.681e-02  -3.653 0.000259 ***
## Transit        -1.745e-02  1.022e-01  -0.171 0.864480
## OtherTransp    -9.864e-02  1.010e-01  -0.976 0.328820
## WorkAtHome     -2.093e-01  7.920e-02  -2.642 0.008238 **
## MeanCommute     6.120e-02  2.494e-02   2.454 0.014133 *
## Employed        1.650e-01  3.287e-02   5.021 5.14e-07 ***
## PrivateWork     8.717e-02  2.216e-02   3.934 8.36e-05 ***
## SelfEmployed    7.917e-03  4.674e-02   0.169 0.865516
## FamilyWork     -1.189e+00  4.080e-01  -2.914 0.003564 **
## Unemployment    1.813e-01  3.811e-02   4.758 1.96e-06 ***
```

```
## CountyTotal      3.666e-07  4.036e-07   0.908 0.363716
## ---
## 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:  853.13  on 2429  degrees of freedom
## AIC: 907.13
##
## Number of Fisher Scoring iterations: 7
```

We can see the significant variables are with the stars following the numbers. It is a little consistent with the tree model, but still somewhat different. For the white category, we can see that it follows from our expectation because whether you are white or black may heavily affect who you going to vote. Also, the citizen is important because policies from future president may affect specific area of people. Thus those two variables are significant.

```
##                       train.error test.error
## tree                   0.07288274 0.09120521
## Logistic Regression  0.06392508 0.07654723
## LASSO                          NA         NA
```

17 You may notice that you get a warning glm.fit: fitted probabilities numerically 0 or 1 occurred.

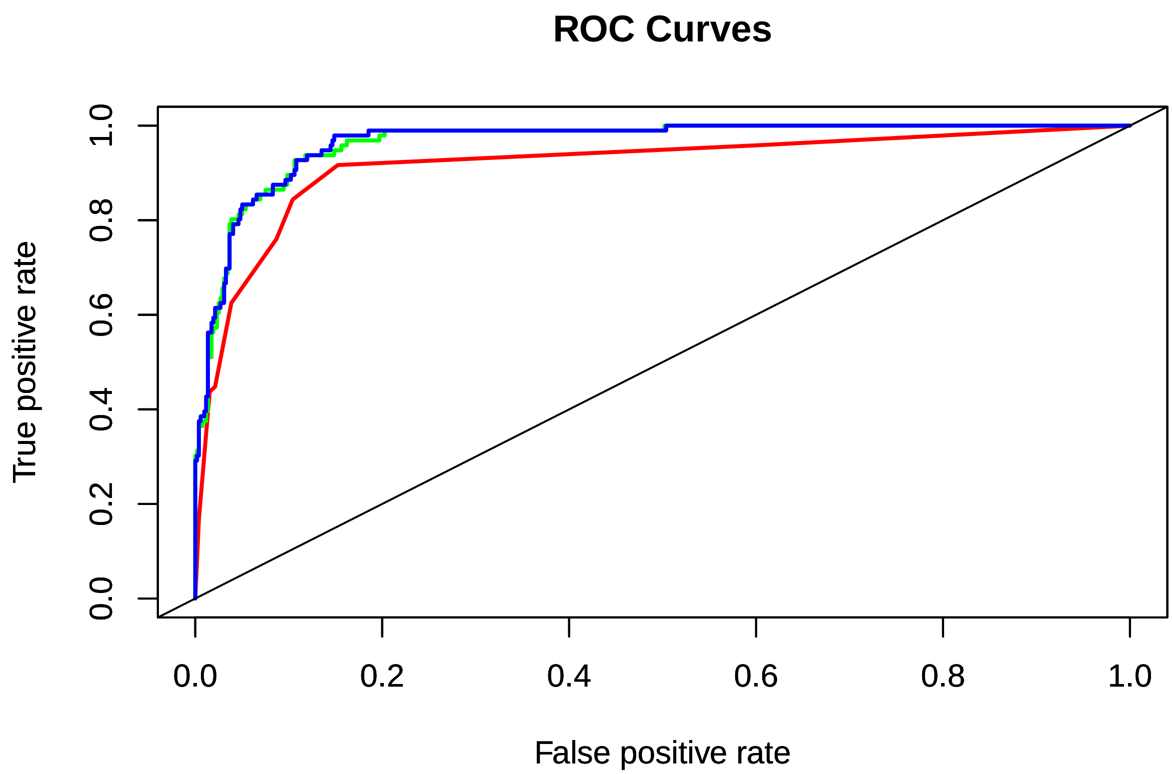


```
## 27 x 1 sparse Matrix of class "dgCMatrix"
##                            1
## (Intercept)     -1.985982e+01
## Men              4.847066e-02
## White           -1.226611e-01
## Minority         .
## Citizen          1.176061e-01
## Income          -4.715141e-05
## IncomeErr       -4.323307e-05
## IncomePerCap     1.998874e-04
## IncomePerCapErr -2.017798e-04
## Poverty          1.648560e-02
## ChildPoverty     .
## Professional     2.479018e-01
## Service          3.294766e-01
## Office           6.815465e-02
## Production       1.482169e-01
## Drive           -2.092266e-01
## Carpool         -1.947458e-01
## Transit          2.688360e-02
## OtherTransp     -4.569328e-02
## WorkAtHome      -1.572083e-01
## MeanCommute      4.214164e-02
## Employed         1.564896e-01
## PrivateWork      7.934463e-02
## SelfEmployed     .
## FamilyWork      -1.040494e+00
## Unemployment     1.718771e-01
## CountyTotal      4.634252e-07
```

We can see that coefficients of Minority, ChildPoverty, and SelfEmployed are zero, and the rest of the variables coefficients are non-zero. Compared to the logistic regression, we find the absolute value of those coefficients turned to be smaller for the LASSO method.

```
##                     train.error test.error
## tree                 0.07288274 0.09120521
## Logistic Regression  0.06392508 0.07654723
## LASSO                0.06514658 0.07817590
```

18 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 different classifiers more appropriate for answering different kinds of problems or questions?



For logistic regression, it has convenient probability scores for observations and efficient implementations available across tools. The cons are also obvious: it doesn't perform well when feature space is too large and relies on transformations for non-linear features and the entire data. For decision trees, the pros are being able to handle non-linear features and taking into account variable interactions. The downsides are it highly biased to traning set and no ranking score. For the LASSO, we have that LASSO does a better job than the usual methods of automatic variable selection such as forward, backward and stepwise, it has a much better result. The cons are it may ignore the variables play a huge role and it does the most job of yours. Indeed, we need different method for different kind of problems. Specficially, from the roc curve we can see that the red line(tree) compared to the other two are less favorable.

Taking it further

19 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 project shows that with so much difficuties to predict the election outcomes, we need to determine the most influential factors in order to form the most accurate predictions. We can see the raw data has some discrepancies like counties were split into 2 subcounties, some cities were classified as counties, and a few counties had missing data for the name variable. These kind of discrepancies make our job much more difficult to identify the voting outcomes for them.

For the previous questions, we've discussed the poverty levels between Hillary and Trump. The conclusion we have drawn is that Hillary had fewer counties vote for and and with less poverty on the counties than the Trump's had. This result is consisitent with our analysis afterwards because it shows that Trump's voters on average have fewer income and the PCA results shows that the income per capita was the most influential factor in the voting.

In fact, the PCA analysis also shows us other important variables such as income per capita and income error on the county level, and income per capita and method of transportation on the subcounty level. To discover the subcounty level, we found that the percentage of the population that commuted via public transportation was highly influential and we were encouraged to know why would this happen. Thus we've figued out that one reason is due to the public transpotation is a bracket for the lower income people. This is what we have found in the PCA analysis.

For the cluster analysis, we also found some discrepancies. For example, we looked at San Mateo county and figured out it will be placed into different categories for the tree model. Specifically, it is different when we consider the whole PCs and PC1 to PC5. And we think this could be an issue about the misclassfication of San Mateo because Democrat-voting county was placed into the cluster 1. And when we consider why this would happen, we figure that income per capita is more influential with Trump voters than Hillary voters. Thus this kind of classification occured.

We are declaring that we want to collecting addional data from past votings like the 2012 election to make the prediction more precise and analyze the data more clearly. We can figure out how many counties swithched their opinions from Democrates to Republicans for example. Also, when we get the data we can contrast the results from different times and locations to get more informations and try to simulate what will happen next.

For the interesting question, we choose to use a different kind of classification method. Using KNN model for classfication. How do these compare to logistic regression and the tree method?

```
## [1] 15
```

This is the best number of k.

```
## [1] 0.1131922
```

```
## [1] 0.1237785
```

This is the error for the knn classification.

We can see that the knn misclassification error is not low compared to the other methods we've used previously. Compared to the logistic regression and tree, we can see that logistic regression has the lowest error rate, which indicates that decision boundary for the candidates is probably on the linear side. Since KNN is non-parametric approach and with a linear boundry, we expect this kind of result from KNN classification. For the classfication trees, we can see the relationship between each variables is well approximated by a linear model then we will figure out that is not good compared to the logistic regression method. Our records error tells exactly what we are expecting to see. However, if we consider the difference between the two methods, it is not that significant, so we may still want to use decision tree because of its interpretability and visualization ability.