PSTAT 131 Final Project

Kevin Zhang and Riccardo Omenti

3/5/2019

Background

The presidential election in 2012 did not come as a surprise. Some correctly predicted the outcome of the election correctly including Nate Silver, and many speculated his approach.

Despite the success in 2012, the 2016 presidential election came as a big surprise to many, and it was a clear example that even the current state-of-the-art technology can surprise us. Predicting voter behavior is complicated for many reasons despite the tremendous effort in collecting, analyzing, and understanding many available datasets. For our final project, we will analyze the 2016 presidential election dataset.

Problem 1

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

From the general knowledge, a voter behaviour is hard to predict as it may be influenced by many factors. Indeed, while making forecasts about the election, many diffferent variables come into play. First of all, we must be aware of the fact that people's opinion may change over time; thus, time is a very important component which is not very easy to account for. Secondly, people do not always tell the truth about who they are going to vote. In addition, we must consider the fact that some events may alterate a given voter's opinion such as an unemployment decrease in a given state. Since election forecasting must take care of such a high number of variables, this may be easily lead to a high error in the model we try to build, which eventually will produce inaccurate and biased results.

Problem 2

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

Nate Silver's approach in predicting 2012 election's outcome consisted in computing the full range of probabilities for each state. In order to achieve this gaol, he calculated a probability for each percentage support for Obama in each state, so he can use this data to ask how much of this probability is above 50%. This model was then simulated forward in time to the election day for each level of support, including state and national. After this, he weighted each forward simulation by the probability that the starting point (the initial probabilities) is the true one. This model may be used to predict the probability that Obama will win the election.

Problem 3

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

Many different factors contributed to the wrong predictions concerning 2016 elections. First of all, the presence of a high nonresponse bias and other errors, which may occur in polls, are a consisent cause in the failure of forecasting 2016 election. Secondly, pollers' opinions may change over time. For example, the scandal involving Clinton's email may decreased her chance of winning Thridly, some of the pollers, especially female and minorities, may have lied about their actual opinion, as they were afraid of expressing their support for Trump. Fourthly, people joining a poll may tend to be more educated and this may have brought about an overestimation of the number of voter in favor of Clinton. Finally, as it is pointed out in the article 'The Polls Missed Trump. We Asked Pollsters Why', the outperformance of Trump over Clinton in states with higher percentage of white males without a college degree was significantly higher than Clinton's in the states, where she was predicted to win.

Data

We create the variables for our data: election.raw, census_meta, and census.

Election Data

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

The dimensions of election.raw is 18345 by 5.

Problem 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.

There are 18,345 observations and 5 variables in election.raw after removing fips = 2000. We removed the observations with fips = 2000 because Alaska has a fips value of 2000, so the rows where fips = 2000 are indeed state-level summary of election results. However, the state-level summary rows of Alaska are already available when we read the data, so it makes no sense to have duplicate records.

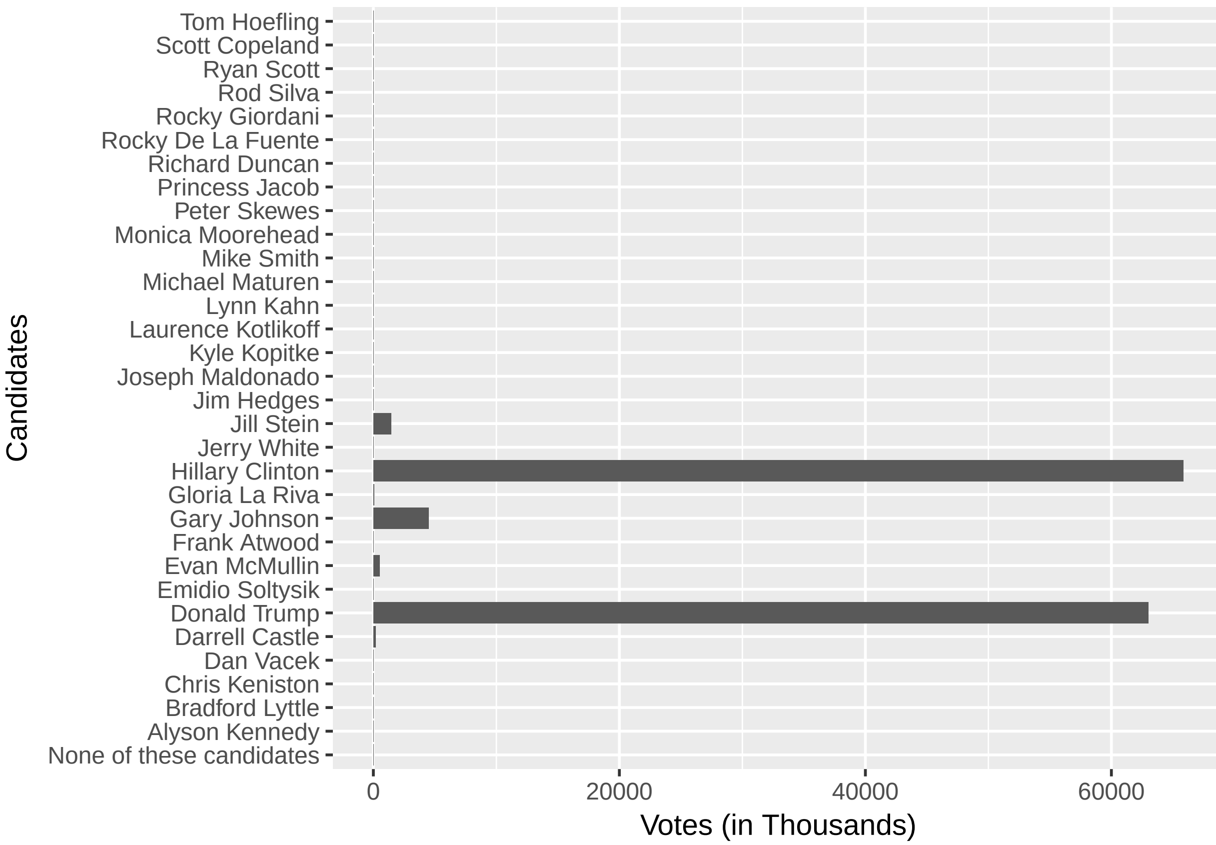
Data Wrangling

Number of Presidential Candidates

```
## [1] 31
```

The number of presidential candidates is 31.

Boxplot of All Votes Received by Each Candidate



County_winner and State_winner

The following is the output of county_winner and state_winner.

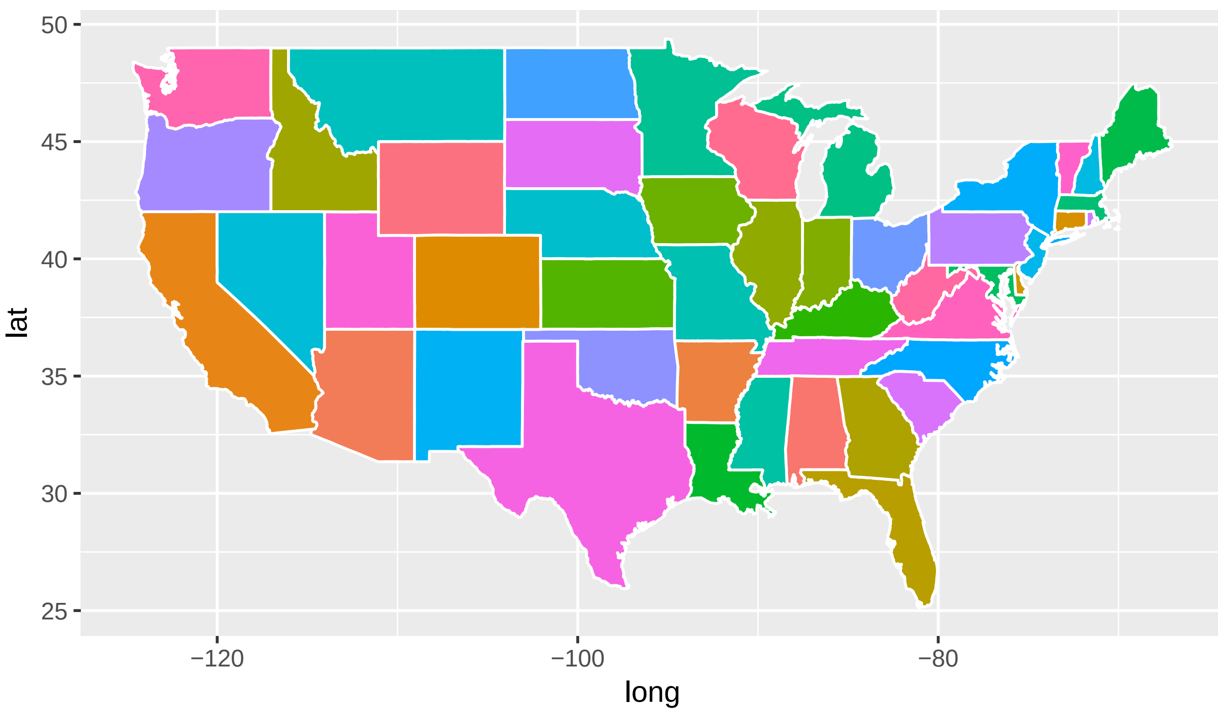
```
## # A tibble: 3,112 x 7
## # Groups:   fips [3,112]
##    county             fips  candidate       state   votes     total     pct
##    <fct>              <fct> <fct>           <fct>   <int>     <int>   <dbl>
##  1 Los Angeles County 6037  Hillary Clinton CA    2464364 135382571 0.0182
##  2 Cook County        17031 Hillary Clinton IL    1611946 135382571 0.0119
##  3 Maricopa County    4013  Donald Trump    AZ     747361 135382571 0.00552
##  4 Harris County      48201 Hillary Clinton TX     707914 135382571 0.00523
##  5 San Diego County   6073  Hillary Clinton CA     735476 135382571 0.00543
##  6 Orange County      6059  Hillary Clinton CA     609961 135382571 0.00451
##  7 King County        53033 Hillary Clinton WA     718322 135382571 0.00531
##  8 Miami-Dade County  12086 Hillary Clinton FL     624146 135382571 0.00461
##  9 Broward County     12011 Hillary Clinton FL     553320 135382571 0.00409
## 10 Kings County       36047 Hillary Clinton NY     640553 135382571 0.00473
## # ... with 3,102 more rows
```

```
## # A tibble: 51 x 7
## # Groups:   fips [51]
##    county fips  candidate       state   votes     total    pct
##    <fct>  <fct> <fct>           <fct>   <int>     <int>  <dbl>
##  1 <NA>   CA    Hillary Clinton CA    8753788 135691978 0.0645
```

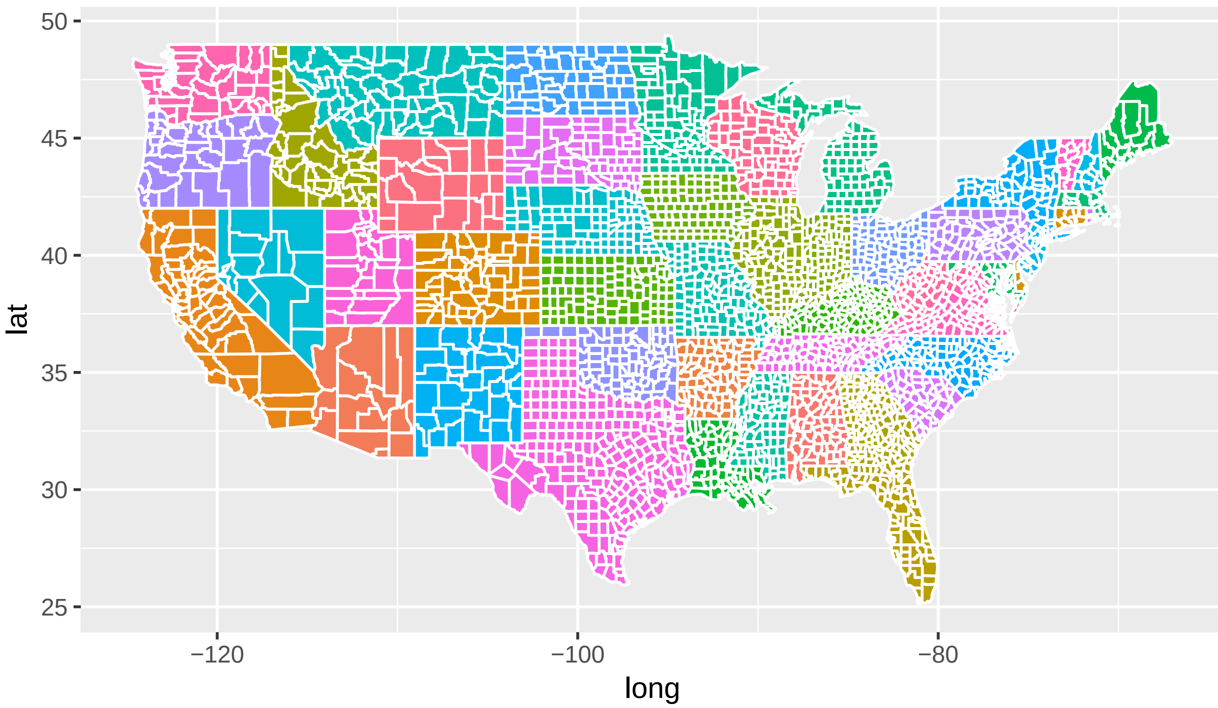
```
##   2 <NA>    FL    Donald Trump    FL    4617886 135691978 0.0340
##   3 <NA>    TX    Donald Trump    TX    4685047 135691978 0.0345
##   4 <NA>    NY    Hillary Clinton NY    4556124 135691978 0.0336
##   5 <NA>    PA    Donald Trump    PA    2970733 135691978 0.0219
##   6 <NA>    IL    Hillary Clinton IL    3090729 135691978 0.0228
##   7 <NA>    OH    Donald Trump    OH    2841005 135691978 0.0209
##   8 <NA>    MI    Donald Trump    MI    2279543 135691978 0.0168
##   9 <NA>    NC    Donald Trump    NC    2362631 135691978 0.0174
##  10 <NA>    GA    Donald Trump    GA    2089104 135691978 0.0154
## # ... with 41 more rows
```

Visualization

Visualization is crucial for gaining insight and intuition during the data mining process. To that end, we will generate cartographic representations (maps) of the states and counties, and map our data onto these representations. Below is the visualization of the State-level map.

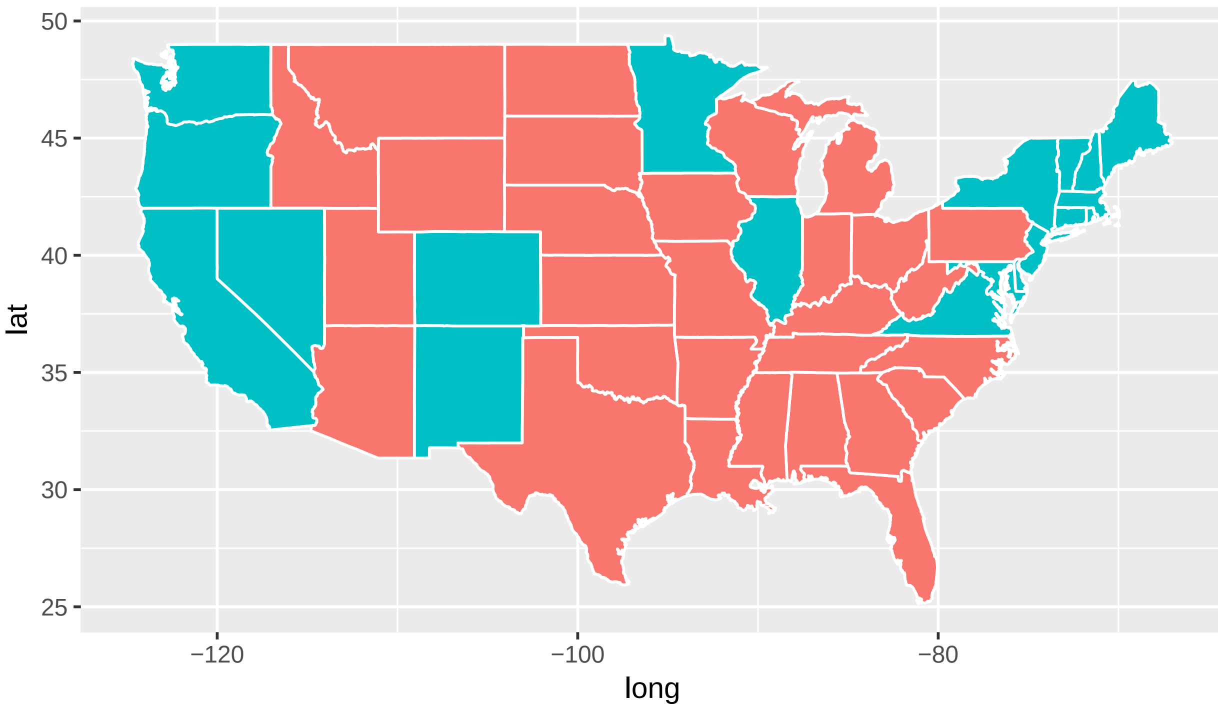


Below is the visualization of the County-level map.



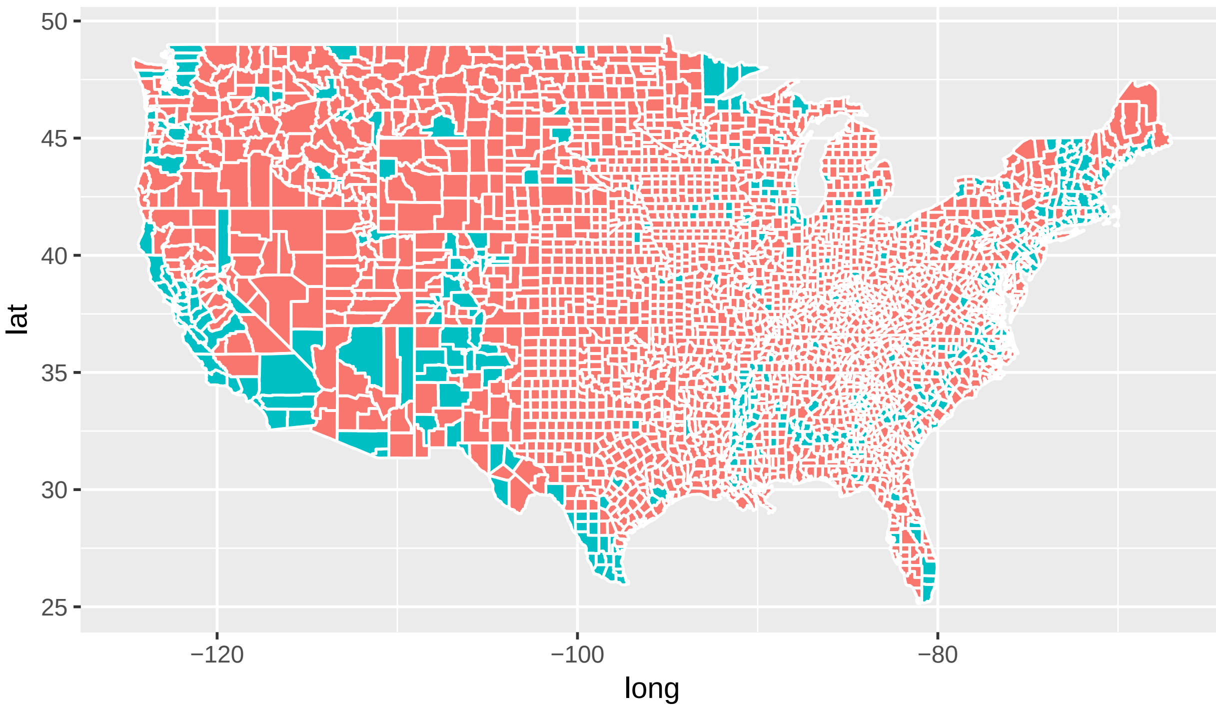
Color the Map by Winning Candidates for State

Here we colored the map according to the winning candidates in each state, such that blue is Hillary Clinton and red is Donald Trump.



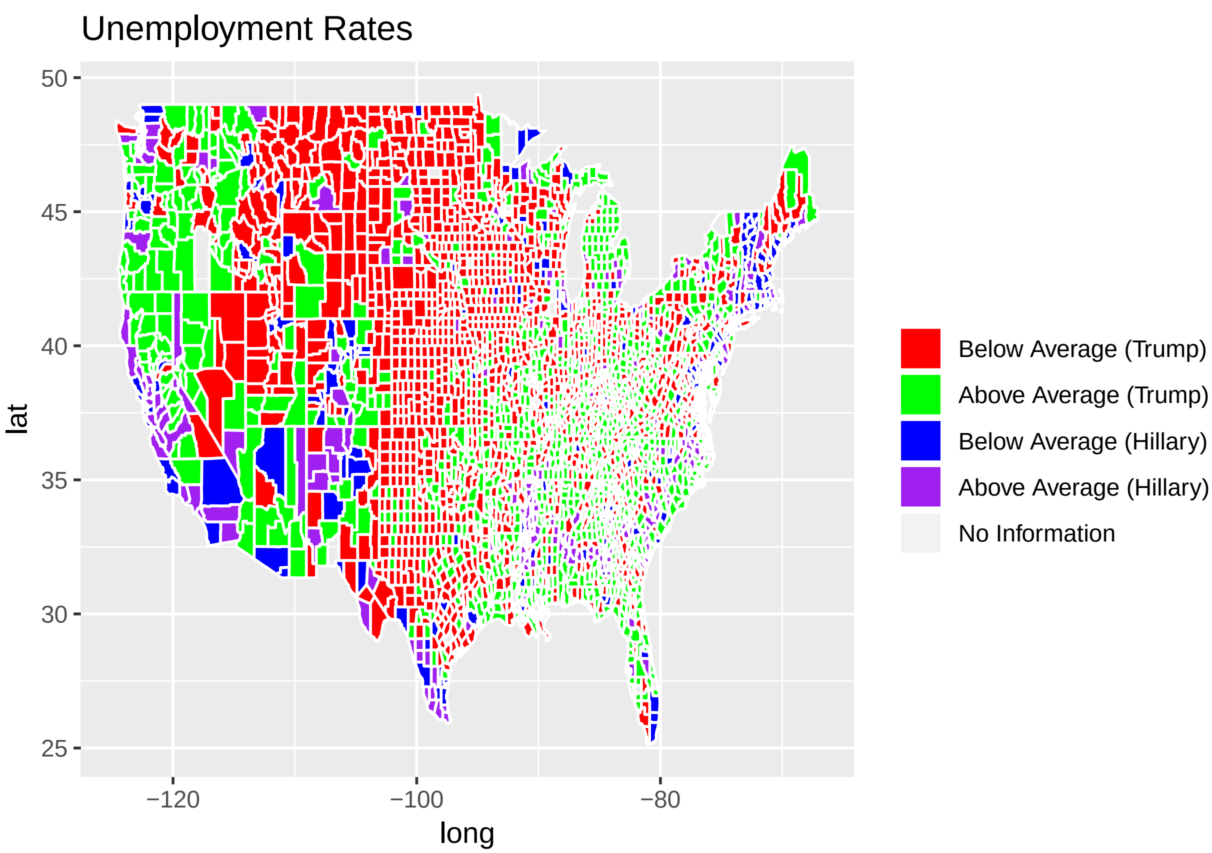
Color the Map by Winning Candidate for County

Here we colored the map according to the winning candidates in each county, such that blue is Hillary Clinton and red is Donald Trump.



Problem 11

Create a visualization of your choice using census data. Many exit polls noted that demographics played a big role in the election. Use this Washington Post article and this R graph gallery for ideas and inspiration.
We chose to create a ggplot visual of the Unemployment rate in the counties. We partitioned the results based on two conditions: whether the county voted Donald Trump or Hillary Clinton, and whether the county was above the average unemployment rate or below the average unemployment rate.



Creating Variables for Census

```
##     State  County      Men    White    Native Minority  Citizen   Income
## 1 Alabama Autauga 48.43266 75.78823 0.4218812 22.53687 73.74912 51696.29
## 2 Alabama Baldwin 48.84866 83.10262 0.5594682 15.21426 75.69406 51074.36
## 3 Alabama Barbour 53.82816 46.23159 0.1881405 51.94382 76.91222 32959.30
## 4 Alabama    Bibb 53.41090 74.49989 0.4310697 24.16597 77.39781 38886.63
## 5 Alabama  Blount 49.40565 87.85385 0.2911748 10.59474 73.37550 46237.97
## 6 Alabama Bullock 53.00618 22.19918 1.1635700 76.53587 75.45420 33292.69
##   IncomeErr IncomePerCap IncomePerCapErr  Poverty ChildPoverty
## 1  7771.009     24974.50        3433.674 12.91231     18.70758
## 2  8745.050     27316.84        3803.718 13.42423     19.48431
## 3  6031.065     16824.22        2430.189 26.50563     43.55962
## 4  5662.358     18430.99        3073.599 16.60375     27.19708
## 5  8695.786     20532.27        2052.055 16.72152     26.85738
## 6  9000.345     17579.57        3110.645 24.50260     37.29116
##   Professional  Service   Office Production    Drive   Carpool    Transit
## 1     32.79097 17.17044 24.28243   17.15713 87.50624  8.781235 0.09525905
## 2     32.72994 17.95092 27.10439   11.32186 84.59861  8.959078 0.12662092
## 3     26.12404 16.46343 23.27878   23.31741 83.33021 11.056609 0.49540324
## 4     21.59010 17.95545 17.46731   23.74415 83.43488 13.153641 0.50313661
## 5     28.52930 13.94252 23.83692   20.10413 84.85031 11.279222 0.36263213
## 6     19.55253 14.92420 20.17051   25.73547 74.77277 14.839127 0.77321596
##   OtherTransp WorkAtHome MeanCommute Employed PrivateWork SelfEmployed
## 1   1.3059687  1.8356531    26.50016 43.43637    73.73649     5.433254
## 2   1.4438000  3.8504774    26.32218 44.05113    81.28266     5.909353
## 3   1.6217251  1.5019456    24.51828 31.92113    71.59426     7.149837
## 4   1.5620952  0.7314679    28.71439 36.69262    76.74385     6.637936
```

```
## 5   0.4199411  2.2654133    34.84489 38.44914    81.82671     4.228716
## 6   1.8238247  3.0998783    28.63106 36.19592    79.09065     5.273684
##   FamilyWork Unemployment CountyTotal
## 1 0.00000000     7.733726       55221
## 2 0.36332686     7.589820      195121
## 3 0.08977425    17.525557       26932
## 4 0.39415148     8.163104       22604
## 5 0.35649281     7.699640       57710
## 6 0.00000000    17.890026       10678
```

Problem 13

If you were physically located in the United States on election day for the 2016 presidential election, what state and county were you in? Compare and contrast these county results, demographic information, etc., against the state it is located in. If you were not in the United States on election day, select a county that appears to stand apart from the ones surrounding it. Do you find anything unusual or surprising? If not, what do you hypothesise might be the reason for the county and state results?

```
##                White    Native Minority
## California 55.65070 1.5756507 41.26007
## Alameda    32.97244 0.3044554 62.58824
```

Looking at the comparison between Alameda County and California, the proportion of minorities in Alameda County is greater than the proportion of minorities in California as a whole (including Alameda County). However, the prporition of Natives and Whites is greater in California than in Alameda County. This is not particularly surprising because Alameda County is a popular residence for minority groups; whereas, the majority of California is White.

Dimensionality Reduction

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
##    0.3508652
```

```
## PrivateWork
##   0.4276034
```

```
## IncomePerCap
##    0.3183354
```

```
##     Drive
## 0.3771882
```

```
## [1] "IncomePerCap"
```

```
## [1] "PrivateWork"
```

```
## [1] "IncomePerCap"
```

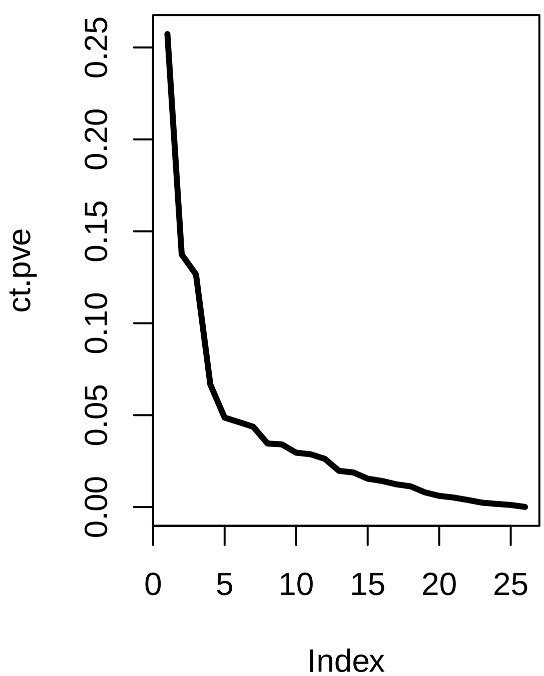
```
## [1] "Transit"
```

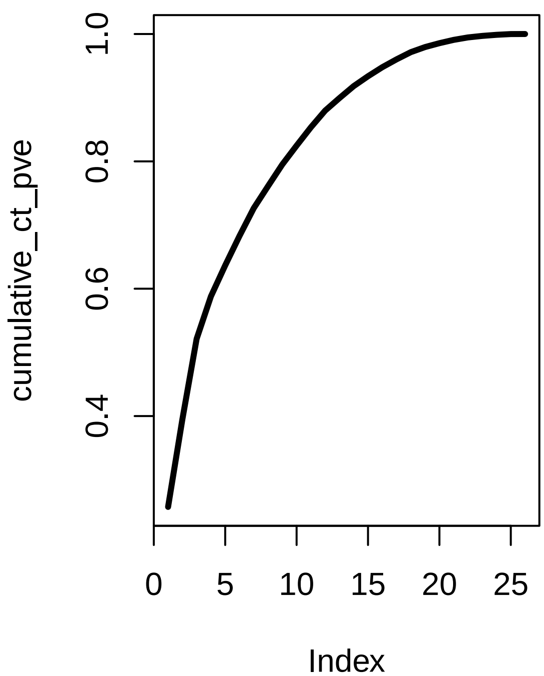
We chose to center and scale the before running PCA because the variables in census.ct contained some categorical values. Also the three highest features of the first principal component were IncomePerCap, PrivateWork, and IncomePerCap. The features that have opposite signs are Men, IncomePerCapErr, Professional, Office, Production, Drive, Transit, WorkAtHome, MeanCommute, SelfEmployed, and FamilyWork. This means that those, which are negative in the first principal component, will make it smaller, while they will increase the second principal component.

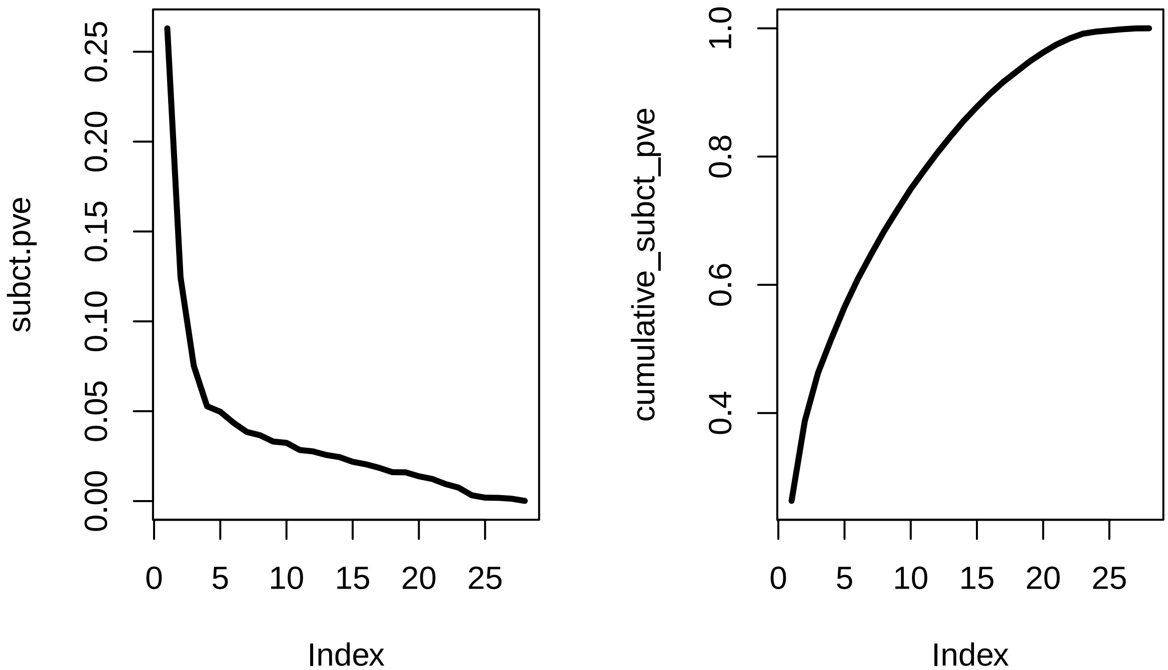
The Minimum Number of PCs and Plotting PVE and Cumulative PVE

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

```
## [1] 17
```







The minimum number of PCs needed to capture 90% of the variance for both the county and sub-county is 14 and 17 respectively.

Clustering

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 cluster? Comment on what you observe and discuss possible explanations for these observations.

```
## clusters1
##    1    2    3    4    5    6    7    8    9   10
## 2751  406    2    7    4   16   20    5    3    4

## clusters2
##    1    2    3    4    5    6    7    8    9   10
## 2564  357  125    4    7   40   20    1   19   81

## [1] 2

## [1] 1
```

Complete linkage is a more appropriate method for clustering San Mateo County because it results in a smaller cluster result. By comparing the two clusters we have obtained, we can notice that the clusters1 results in two rather big groups of observations and the remaining smaller 8 groups. On the other hand, clusters2 seem to have yielded a similar results regarding the split of groups, as among the 10 a couple have a very high size whereas the others seem to have a very small size. In general, such a result may tell us that in both clusters there is a high number of similar observations, which are either in the first group or the second group respectively, while the other few remaining observations may be very different.

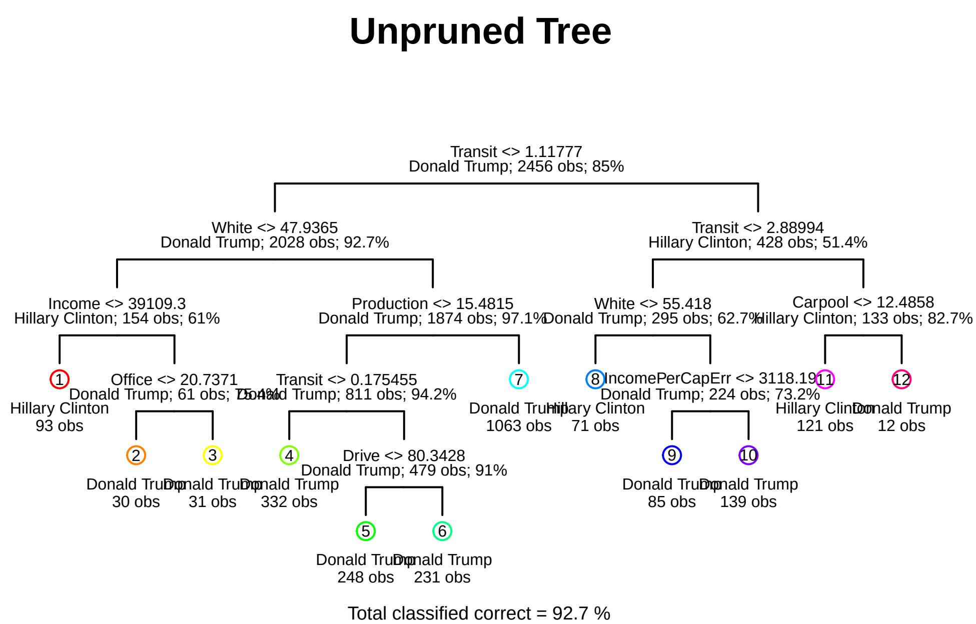
Classification

In order to train classification models, we need to combine county_winner and census.ct data. Then we partition data into 80% training and 20% testing to do 10 cross-validation folds.

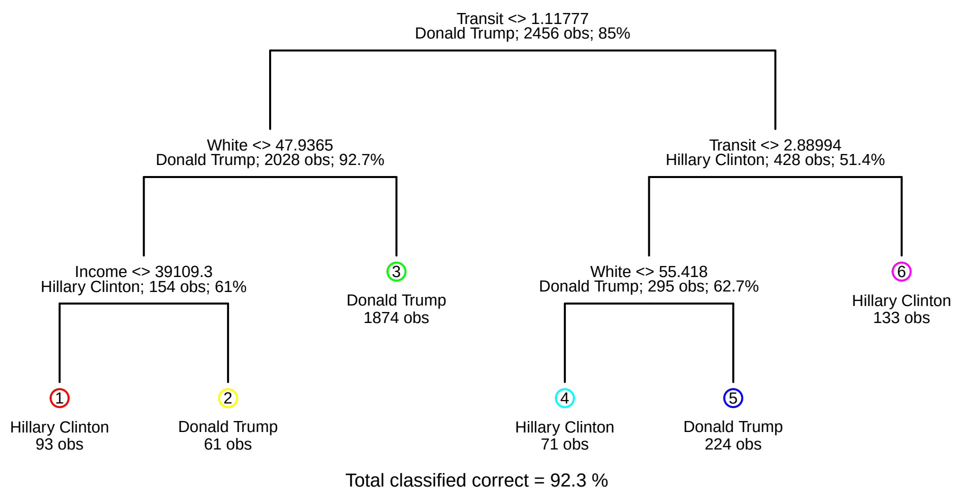
Decision Tree

```
##
## Classification tree:
## tree(formula = candidate ~ ., data = trn.cl)
## Variables actually used in tree construction:
## [1] "Transit"         "White"          "Income"         "Office"
## [5] "Production"      "Drive"          "IncomePerCapErr" "Carpool"
## Number of terminal nodes:  12
## Residual mean deviance:  0.3816 = 932.5 / 2444
## Misclassification error rate: 0.07329 = 180 / 2456
```

```
## [1] 6
```



Pruned Tree



When comparing our pruned tree to the unprund tree, we can see that production is one of the variables that did not show up in the pruned tree. This is due to it not being significant enough in determining whether a voter chose Hillary or Trump. We also grouped carpool into transit because we would be overfitting the data if carpool was it's own node. We can intrepret such a pruned tree as follows.

If the transit in a city is less than 1.1177 and the percentage of White people is 47.93% and the average income is less than 39109.3, then 93 observations are likely to vote for Hillary and 61 observations are likely to vote for Trump with a 92.3% confidence. On the other hand, if the transit in a city is less than 2.89 and the percentage of white people is less than 55.418, then 71 observations are likely to vote for Hillary and 224 observations are likely to vote for Donald Trump with a 92.3% confidence.

```
##          train.error test.error
## tree      0.07654723 0.08780488
## logistic          NA         NA
## lasso             NA         NA
```

Logistic Regression

What are the significant variables? Are these consistent with what you observed in the decision tree analysis? Interpret the meaning of a couple of the significant coefficients in terms of a unit change in the variables. Did your particular county (from question 13) results match the predicted results?

```
##
## Call:
## glm(formula = candidate ~ ., family = binomial, data = trn.cl)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.7906  -0.2654  -0.1145  -0.0436   3.5625
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.016e+00  9.364e+00  -0.856 0.391968
## Men              2.953e-02  5.134e-02   0.575 0.565146
## White           -2.619e-01  6.414e-02  -4.083 4.45e-05 ***
## Native          -4.632e-02  1.331e-02  -3.479 0.000504 ***
## Minority        -1.271e-01  6.165e-02  -2.061 0.039314 *
## Citizen          1.143e-01  2.853e-02   4.006 6.19e-05 ***
## Income          -3.249e-05  2.683e-05  -1.211 0.225826
## IncomeErr       -3.793e-05  6.242e-05  -0.608 0.543400
## IncomePerCap     1.819e-04  6.257e-05   2.907 0.003648 **
## IncomePerCapErr -2.150e-04  1.341e-04  -1.603 0.108940
## Poverty          5.444e-02  4.035e-02   1.349 0.177334
## ChildPoverty    -1.160e-02  2.462e-02  -0.471 0.637461
## Professional     2.486e-01  3.669e-02   6.774 1.25e-11 ***
## Service          3.088e-01  4.545e-02   6.795 1.08e-11 ***
## Office           9.081e-02  4.528e-02   2.006 0.044892 *
## Production       1.610e-01  4.118e-02   3.909 9.29e-05 ***
## Drive           -1.892e-01  4.451e-02  -4.252 2.12e-05 ***
## Carpool         -1.757e-01  5.887e-02  -2.985 0.002838 **
## Transit          7.453e-02  8.922e-02   0.835 0.403480
## OtherTransp     -1.004e-01  9.283e-02  -1.082 0.279330
## WorkAtHome      -8.175e-02  7.003e-02  -1.167 0.243096
## MeanCommute      2.313e-02  2.409e-02   0.960 0.336994
## Employed         1.666e-01  3.220e-02   5.172 2.32e-07 ***
## PrivateWork      7.199e-02  2.178e-02   3.306 0.000945 ***
## SelfEmployed     5.992e-03  4.529e-02   0.132 0.894746
## FamilyWork      -9.665e-01  3.884e-01  -2.489 0.012826 *
## Unemployment     1.845e-01  3.845e-02   4.799 1.60e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2078.43  on 2455  degrees of freedom
## Residual deviance:  855.79  on 2429  degrees of freedom
## AIC: 909.79
##
## Number of Fisher Scoring iterations: 7
```

```
##          train.error test.error
## tree      0.07654723 0.08780488
## logistic  0.06351792 0.08292683
## lasso             NA         NA
```

The signigificant variables are: White, Native, Minority, Citizen, IncomePerCap, Professional, Service, Office, Production, Drive, Carpool, Employed, PrivateWork, FamilyWork, and Unemployment.
These variables are consistent with the variables that were significant in the Decision Tree method.
The meaning of a couple of the significant coefficients in terms of a unit change in the variables can be explained for example, a unit increase in the Native variable is the logit (log odds) of the response is decreased by the coefficient of -1.271e-01.
Additionally, a unit increase in the Professional variable is the logit (log odds) of the response is increased by the coefficient of 2.486e-01.
Also, a unit increase in the Unemployment variable is the logit (log odds) of the response is increased by the coefficient of 1.845e-01.

If the being White or being Employed affected individuals who voted for the Hillary Clinton or Donald Trump, then our results (of Alameda County) were simialr to the predictions.

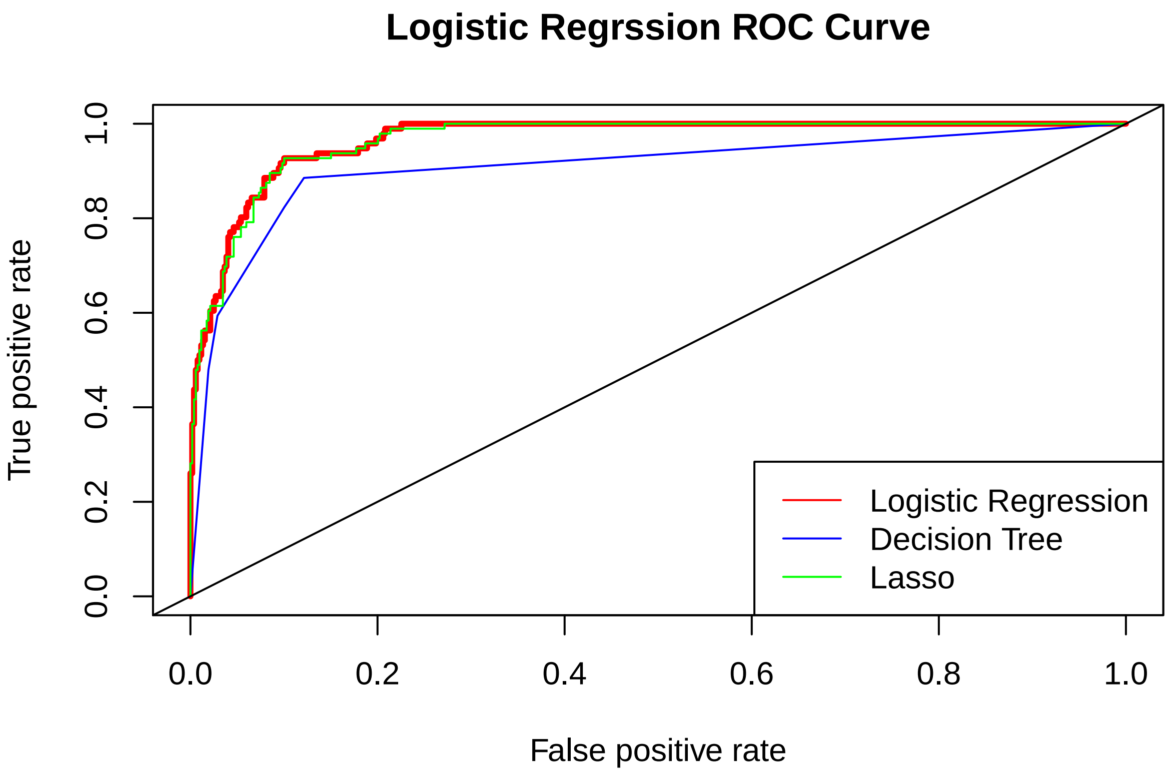
Lasso

```
## 27 x 1 sparse Matrix of class "dgCMatrix"
##                           s0
## (Intercept)     -2.176534e+01
## Men              .
## White           -1.190485e-01
## Native          -3.876376e-02
## Minority         .
## Citizen          1.222497e-01
## Income           .
## IncomeErr       -3.603392e-05
## IncomePerCap     9.301125e-05
## IncomePerCapErr -9.974042e-05
## Poverty          4.374621e-02
## ChildPoverty     .
## Professional     2.070226e-01
## Service          2.590183e-01
## Office           5.256119e-02
## Production       1.025792e-01
## Drive           -1.162394e-01
## Carpool         -9.776333e-02
## Transit          1.396624e-01
## OtherTransp      .
## WorkAtHome      -3.746574e-03
## MeanCommute      .
## Employed         1.516132e-01
## PrivateWork      6.476362e-02
## SelfEmployed    -6.722474e-03
## FamilyWork      -7.468008e-01
## Unemployment     1.668978e-01
```

```
##          train.error test.error
## tree      0.07654723 0.08780488
## logistic  0.06351792 0.08292683
## lasso     0.07084691 0.08292683
```

The non-zero coefficients in the LASSO regression for the optimal value of λ are Men, Minority, Income, ChildPoverty, OtherWTransp, and MeanCommute. In the unpenalized logistic regression, the standard error of the coefficients are higher; therefore, there are more significiant coefficients in the unpenalized logstic regression. This is because the lambda shrinks the coefficients towards zero by decreasing the variance at the cost of some bias.

ROC Curves



Looking at the ROC curves, we see that the Logistic Regression and the Lasso curves overlap because we can think of the Lasso as a method of a simpler model by pushing the coefficients to be zero. In the end, they are very similar because Lasso is a method, which provides us with a more parcimonious model, in which we fit logistic regression, by the introduction of penalalty parameter. On the other hand, the ROC curve for the Decision Tree method seems to be a bit further from the other two curves, resulting in a slightly higher false positve rate and a slightly lower true positive rate. This may be due to the fact that such a method is non-paramteric and may produce less powerful results due to bias-variance trade-off.

Problem 21

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 or doesn't seem 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.

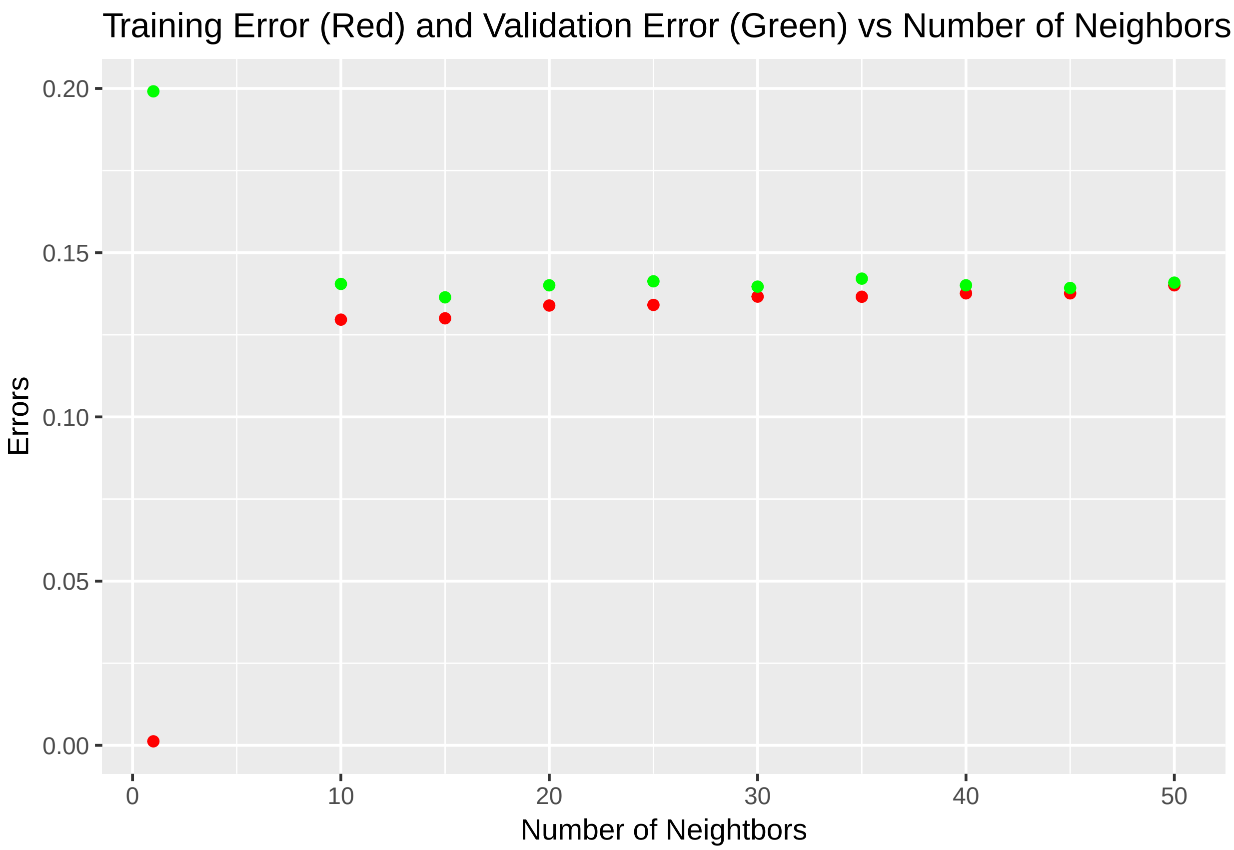
Exploring additional classification methods: KNN. How does this compare to logistic regression and the tree method?

For the last question, we have decided to use a KNN algorithm. Such a method may allow us improving the interpretability of our classification analysis. Indeed, such an analysis regarding the American political election has generally a high impact. Therefore, it may be preferable to pick methods, whose results are simpler to interpret. In this way, the results may be better understood even by those who are not statisticians. Furthermore, KNN is a non-parametric method with no distributional assumptions regarding the data. This may help not to overfit our data. Indeed, in the previous questions, we have reduced the number of candidates in the data sets as we have also taken into account the two main candidates Hilary Clinton and Donald Trump. By applying the KNN we may have some more reliable results.

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

```
## [1] 0.1311075
```

```
## [1] 0.1447154
```



The training error is 0.1311075 and test error is 0.1447154 when we chose neighbors = 15 for the KNN method. These errors are higher than the Decision Tree and Logistic Regression method. This is likely due to fact that when we perform 10-fold cross validiation, the misclassification rate changes based on how we split our training and test data. The reasoning for choosing this method is because KNN is non-parametric, which has no assumptions about the data distribution. Thus, the flexibility of KNN is an advantage in classifying our data.