2016 Election Analysis

Zian He

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

First, the data we get to predict the outcome of the poll is based on how people think they will vote at the time they are asked, which may be months before the election, and people might change over time. Also, many tangible and latent variables will affect the final outcome. Moreover, sampling error will influence our prediction in that pollsters necessarily don't ask everyone for their intentions, so instead they take a random sample, and the sample they take might has tendency. People might also lie about who they will vote for due to the Shy Tory Effect. In all, predicting vote behavior will be a hard problem.

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?

The single most important reason that Silver's model gave Trump a better chance than others is because of his assumption that polling errors are correlated. No matter how many polls you have in a state, it's often the case that all or most of them miss in the same direction. Furthermore, if the polls miss in one direction in one state, they often also miss in the same direction in other states, especially if those states are similar demographically. Also, his model considers the number of undecided and third-party voters when evaluating the uncertainty in the race. There were far more of these voters than in recent, past elections.

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?

They are lots of reasons that why analysts believe the predictions were less believable. The sampling error would definitely be the most influential one, the sampling error is larger because more people are undecided or lying during the polling before the final date. Personally, I think if pollsters should be more independent about the media and should give less significance about the live survey polling, to reduce the bias and the lies. Also, I think the pollsters should also give survey about the latent variables, such as whether they want to speak out, the level of their favors, in order to predict whether they will change in the future. For the future prediction, the sampling error would still be the priorest challenge, the latent variables would also be the case. From my opinion, I don't think the journalists communicate well to the audience about their model. They should give more words and details about the terminologies they used, since students like us from the statistical department can't understand really well about that. If people understand better about the model, maybe they could express more about why they make the decision or other information that can be used to facilitate the prediction model.

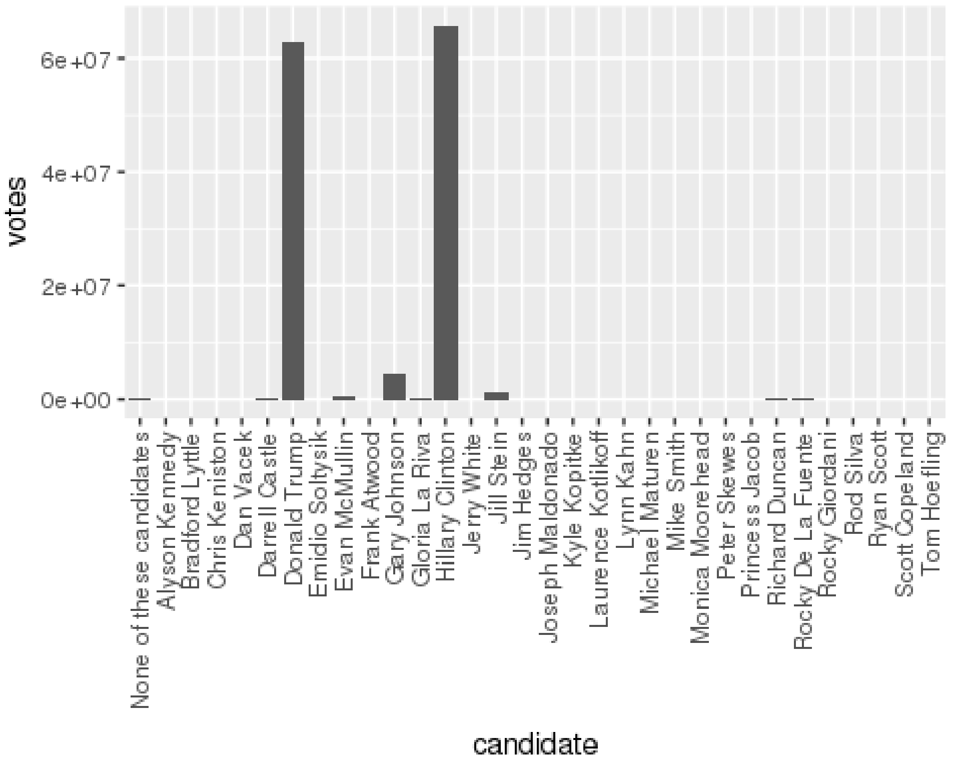
4. We first split the originial election.raw data into 3 split datasets.
   Federal-level summary into election_federal.

   State-level summary into election_state.

   Only county-level data is to be in election.

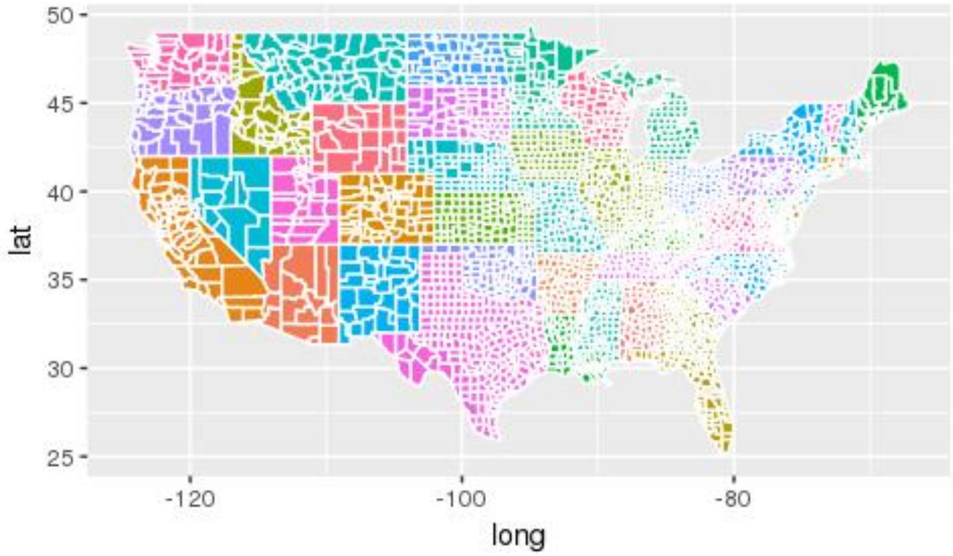
5. We have 32 presidential candidates, and 31 of them are named, one of them belongs to people other than these 31 people.
   Here is the bar chart for the number of votes for each candidate.
   We can see that Donald Trump and Hillary Clinton own most of the votes.

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

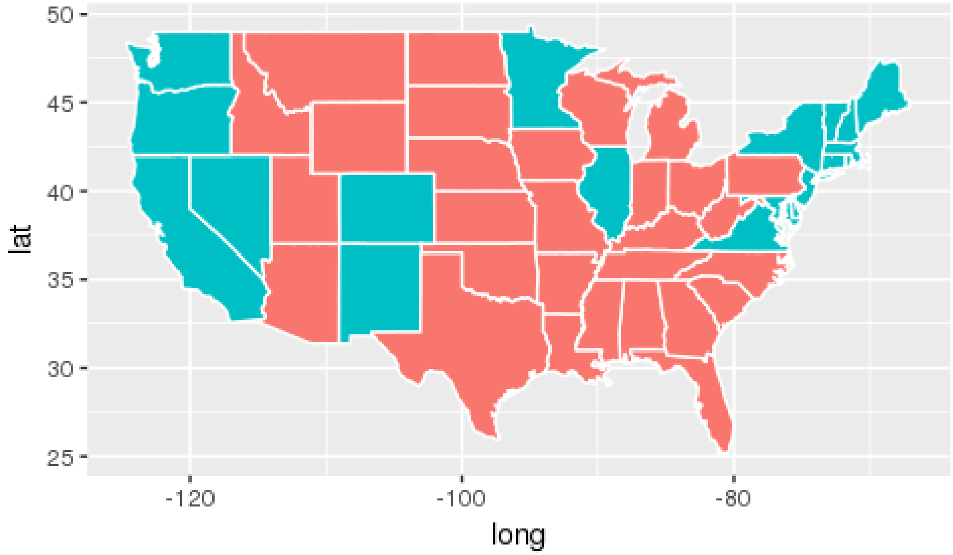


6. Then we create variables county_winner and state_winner by taking the candidate with the highest proportion of votes.

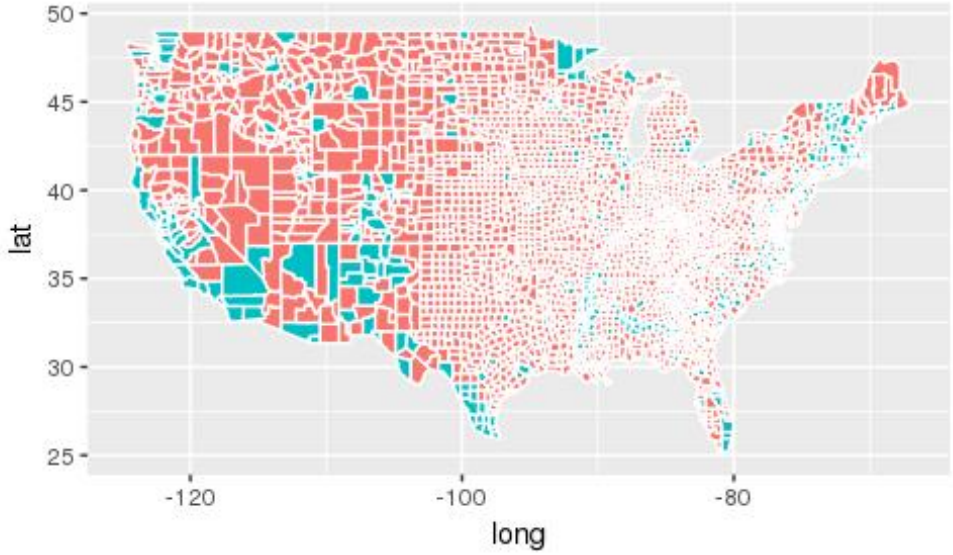
7. Below is the county-level map of United States. We color this map by county.



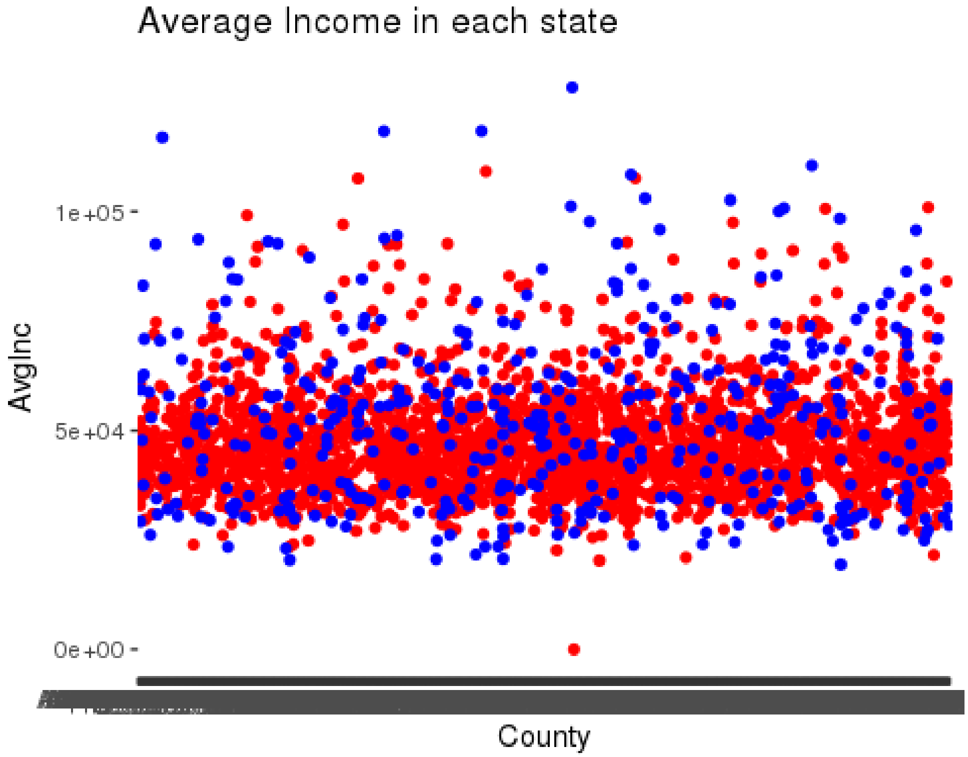
8. We then color the map by the wining candidate for each state. We all know that the red color stand for Trump, whereas the blue one stand for Clinton.



9. We then color the map by the wining candidate for each county.



10. We first calculate the average income in each county, then assign each county to its winner, Clinton or Trump. Now we plot two plots of average income together, and the blue plots denotes the average income of each county that supports Clinton, where the red plots denotes the average income of each county that supports Trump. We can find that the counties that support Trump generally have more aggregate mid-lower average income, where the counties that support Clinton have either low average income or mid-higher income.

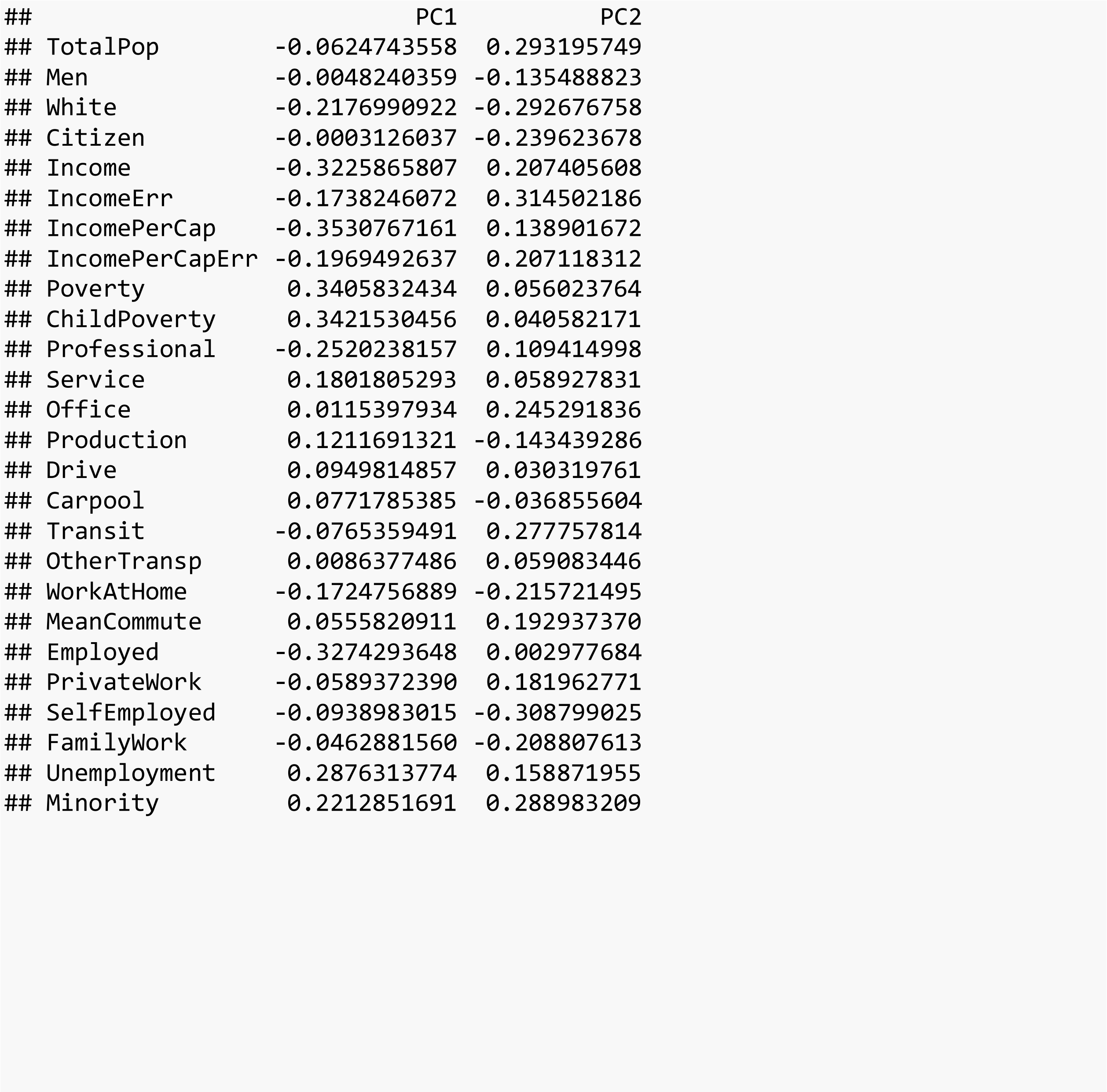


11. Now we aggregate the information into county-level data by computing TotalPop-weighted average of each attributes for each county.

12. Now we run principle component analysis for both county and sub-county level data.

    I think we should set the scale and center both equal to true since the data has percentage data as well as count data. From the rotation matrix of the first two PCs, we can see that Income Per Capita has the larget absolute loading in PC1. For PC2, the median household income error has the largest loading.

    Below is the rotation matrix for PC1 and PC2.

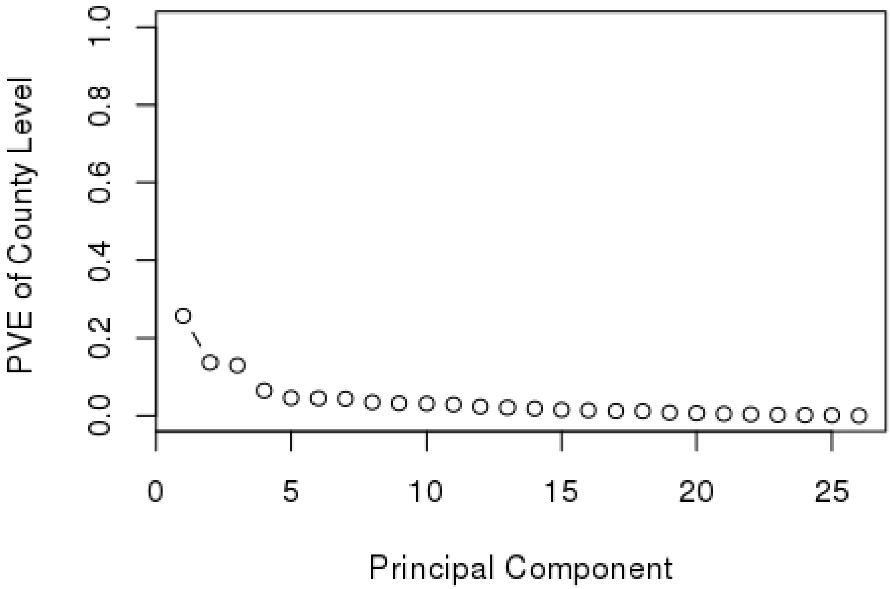


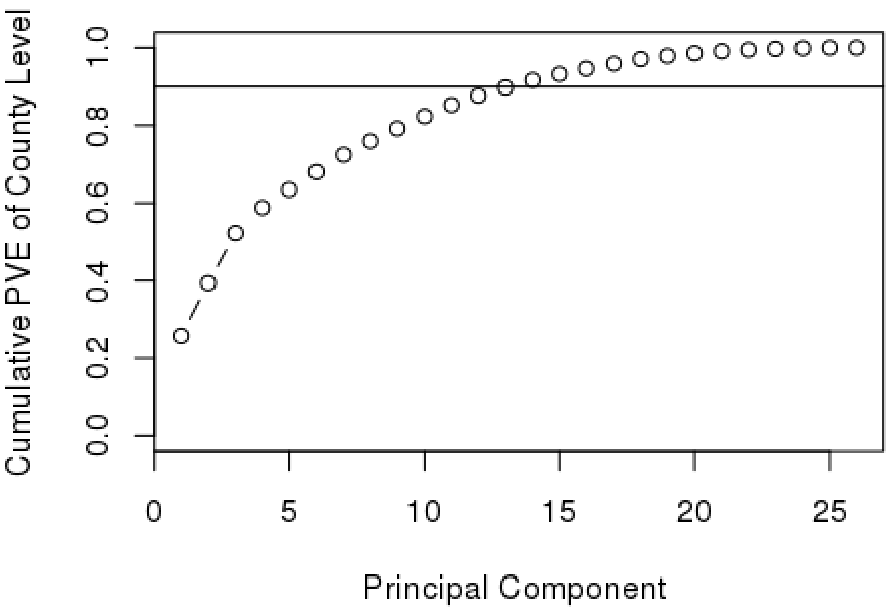
```
##                          PC1          PC2
## TotalPop        -0.0624743558  0.293195749
## Men             -0.0048240359 -0.135488823
## White           -0.2176990922 -0.292676758
## Citizen         -0.0003126037 -0.239623678
## Income          -0.3225865807  0.207405608
## IncomeErr       -0.1738246072  0.314502186
## IncomePerCap    -0.3530767161  0.138901672
## IncomePerCapErr -0.1969492637  0.207118312
## Poverty          0.3405832434  0.056023764
## ChildPoverty     0.3421530456  0.040582171
## Professional    -0.2520238157  0.109414998
## Service          0.1801805293  0.058927831
## Office           0.0115397934  0.245291836
## Production       0.1211691321 -0.143439286
## Drive            0.0949814857  0.030319761
## Carpool          0.0771785385 -0.036855604
## Transit         -0.0765359491  0.277757814
## OtherTransp      0.0086377486  0.059083446
## WorkAtHome      -0.1724756889 -0.215721495
## MeanCommute      0.0555820911  0.192937370
## Employed        -0.3274293648  0.002977684
## PrivateWork     -0.0589372390  0.181962771
## SelfEmployed    -0.0938983015 -0.308799025
## FamilyWork      -0.0462881560 -0.208807613
## Unemployment     0.2876313774  0.158871955
## Minority         0.2212851691  0.288983209
```

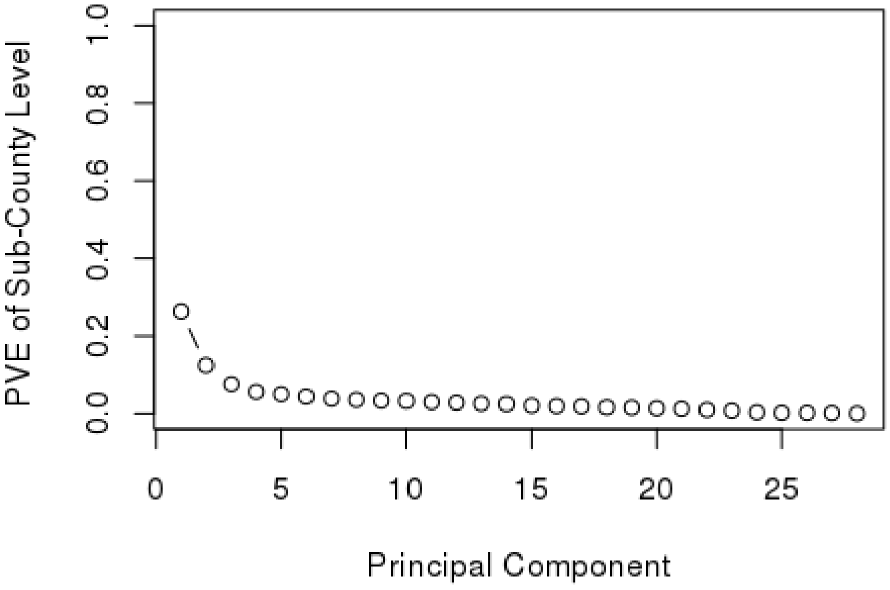
13. We try to determine the number of minimum number of PCs needed to capture 90% of the variance for both the county and sub-county level data.

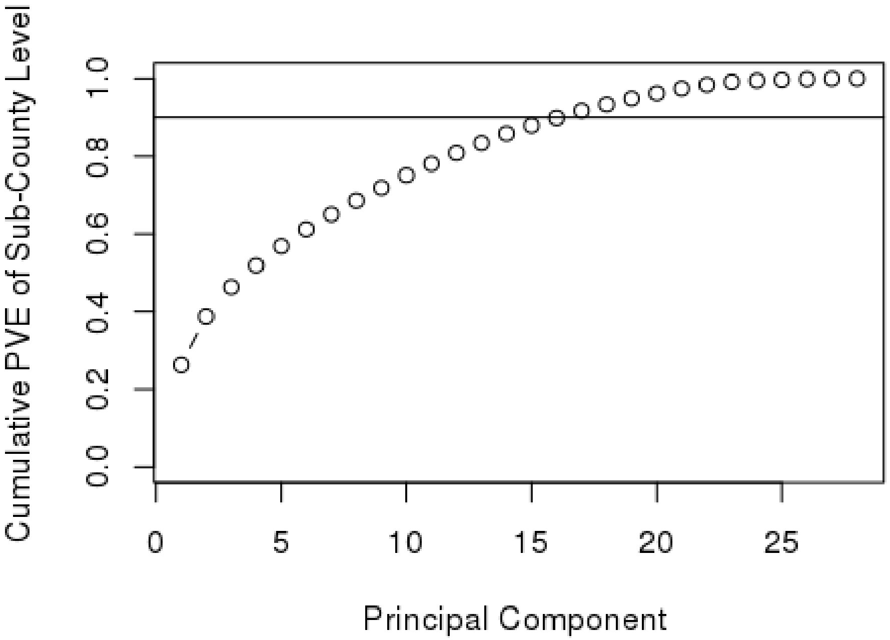
    We Plot proportion of variance explained (PVE) and cumulative PVE for both county and sub-county analyses.

    For sub-county level data, we need at least 17 PCs to capture 90% of the variance. For county level data, we need at least 14 PCs to capture 90% of the variance.

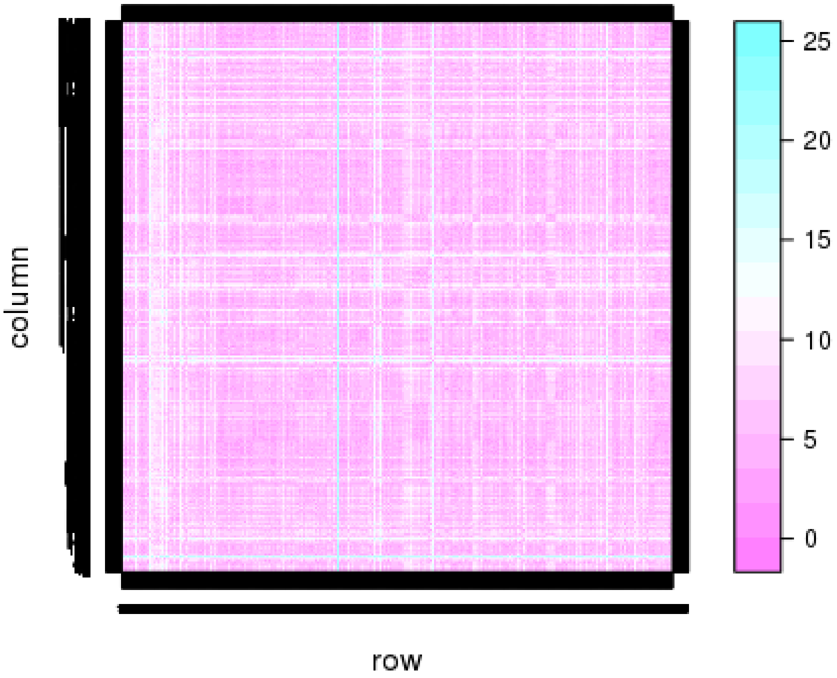


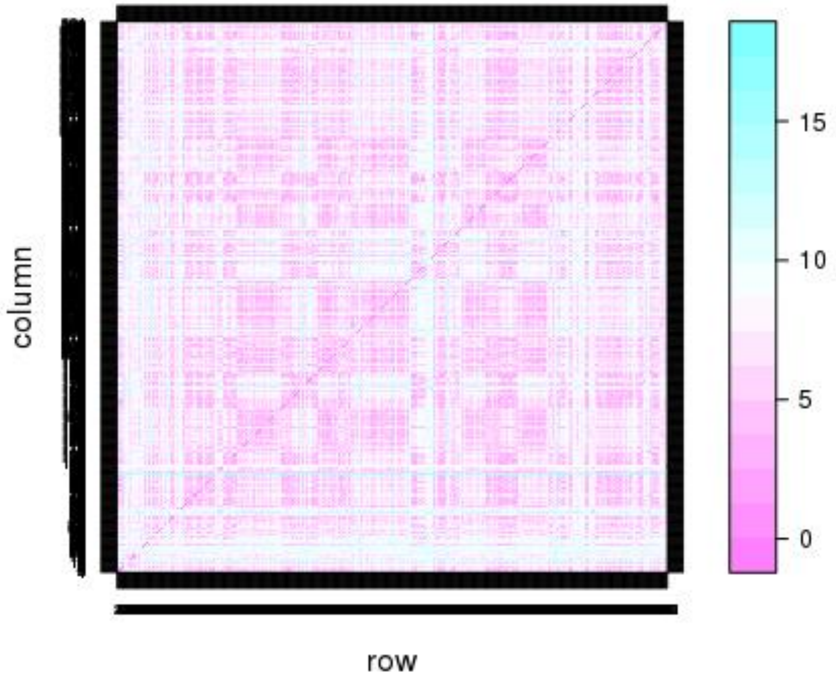

Cumulative PVE of County Level
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Principal Component


PVE of Sub-County Level
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
Principal Component



14. After we find the groups that San Mateo city in for each model, we build two levelplots for these two groups. The first levelplot is for the model built on the first 5 PCs, and the second levelplot is for the model built on the census.ct data. Although the level plots are in different scales, we can still find from the levelplots that the group in the first 5 pcs model centers better, so this model has more appropriate clusters.

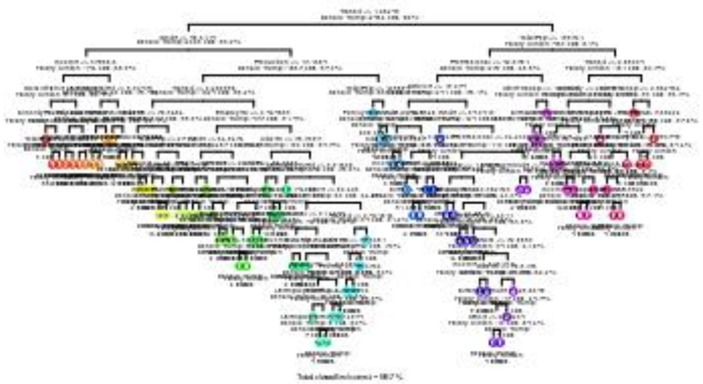


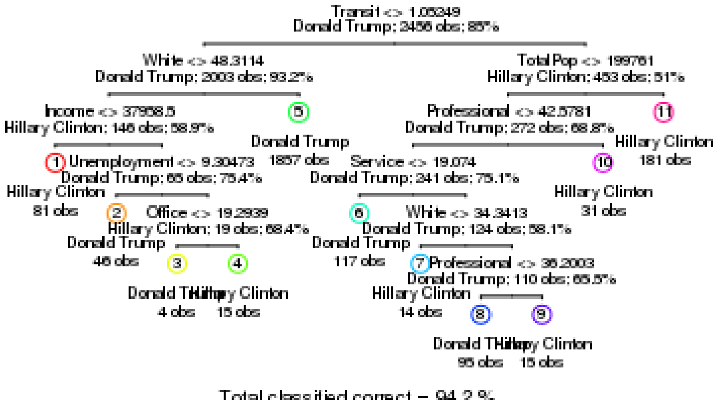


15. Now we try to build the decision tree and prune the tree.

    Before pruning, the tree has over 90 tree nodes so that it's impossible for us to visualize, and it must overfit our result. Then we prune our tree by setting the best size at 11, we got a tree that can lead to a lower test error rate, which is exactly what we want. From this decision tree image, we can see that among the states, if citizens commuting on public transportation everyday less than 1.0525% out of total number of people, they will vote for Donald Trump mostly, and this is a state where majority of people choose to drive to work. During these states, if more than 48.3144% of population are white, then Donald Trump is more likely to win. It is a state with more than half of white people. However, if there are less than 48.3144% of population of white people, votes are inclined to vote Hillary Clinton. Amid these states, if median household income is below 37958.5, then Hillary will have more chance to win. Next, if the state with the median household income greater than 37958.5, and simultaneously with a lower unemployed rate compared to the threshold 9.30473%, then Donald Trump will be more possiible to win. In the contrast, people in states which have higher unemployment rate will tend to vote Hillary. If a state with less than 19.2939% of people employed in sales and office, they will tend to vote Trump more likely.

We can also see the train error and test error after we prune the tree in the table below.





```
##            train.error test.error
## tree        0.05822476 0.07980456
## logistic            NA         NA
## lasso               NA         NA
```

16. Then we want to use the logistic regression to build the model.

The most important variables are Professional, Service, Citizens, Drive and Carpool, and they are slightly inconsistent with the important variables in the tree model. The logistic model predicts mostly based on people’s profession and the number of citizens, whereas the tree model focuses primarily based on the way people commuting with public transportation.

Below is the error table and the summary of the logistic regression model.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##          train.error test.error
## tree      0.05822476 0.07980456
## logistic  0.06392508 0.07654723
## lasso             NA         NA
## 
## 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    
## TotalPop         3.666e-07  4.036e-07   0.908 0.363716    
## Men              8.346e-02  5.386e-02   1.550 0.121229    
## White           -2.147e-01  6.550e-02  -3.278 0.001047 ** 
## 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+00   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 ***
## Minority        -8.369e-02  6.274e-02  -1.334 0.182229
## ---
## 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
```

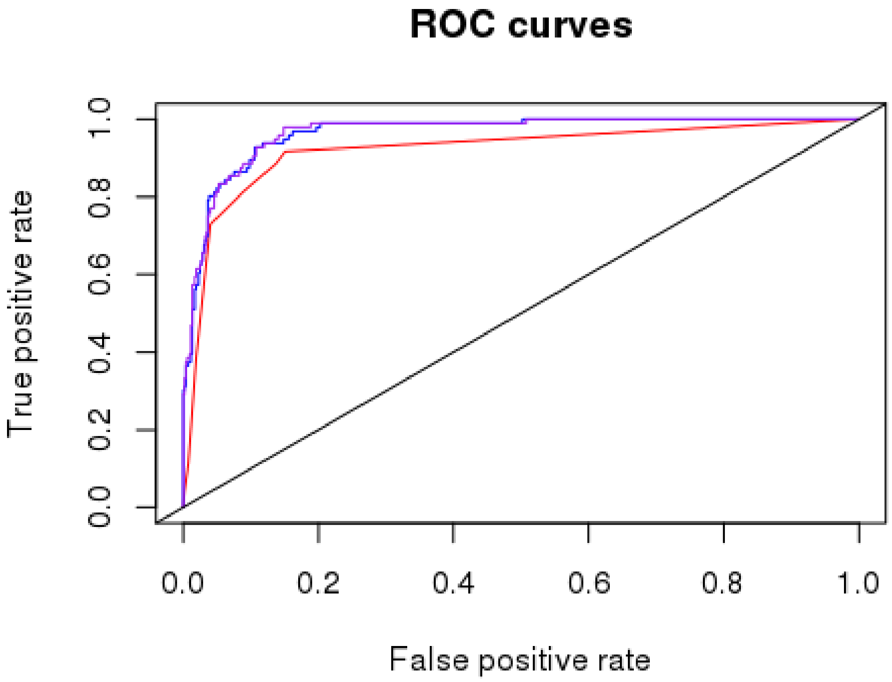
17. Finally, we build the lasso model to try to compromise the overfitting.

    Only Childpoverty, Self Employed, Minority have zero-coefficient. Total Population, Income, Income Error, Income per capita, Income per capita error have really small coefficient, but it’s fine because the values of their data are great also. All others have non-zero coefficients. The train error for lasso regression is greater than logistic regression, but the test error for lasso regression is smaller than logistic regression, which means that the laso regression relieve the overfitting problem that happens in the logistic regression.

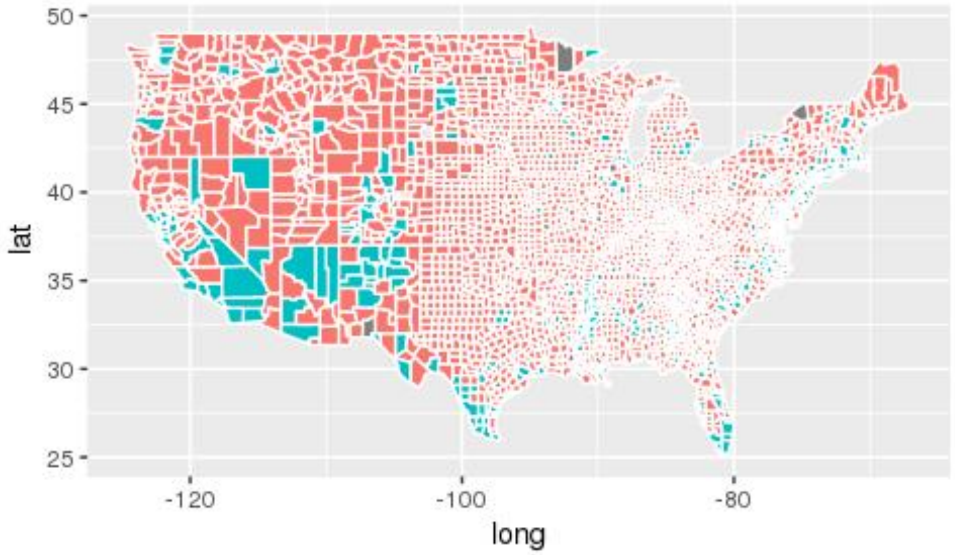
```
##     (Intercept)        TotalPop             Men           White
##   -1.992489e+01    4.768649e-07    3.897490e-02   -1.211283e-01
##         Citizen          Income       IncomeErr    IncomePerCap
##    1.212554e-01   -4.006344e-05   -4.224593e-05    1.813948e-04
## IncomePerCapErr         Poverty    ChildPoverty    Professional
##   -1.792067e-04    1.711590e-02    0.000000e+00    2.399216e-01
##         Service          Office      Production           Drive
##    3.203617e-01    5.857007e-02    1.387448e-01   -1.967658e-01
##         Carpool         Transit     OtherTransp      WorkAtHome
##   -1.823394e-01    3.606572e-02   -3.014910e-02   -1.426446e-01
##     MeanCommute        Employed     PrivateWork    SelfEmployed
##    3.719081e-02    1.547020e+01    7.878242e-02    0.000000e+00
##      FamilyWork    Unemployment        Minority
##   -1.004608e+00    1.693336e-01    0.000000e+00
##          train.error test.error
## tree      0.05822476 0.07980456
## logistic  0.06392508 0.07654723
## lasso     0.06677524 0.07817590
```

18. Now we try to plot the ROC curves on each model and compare these three models.

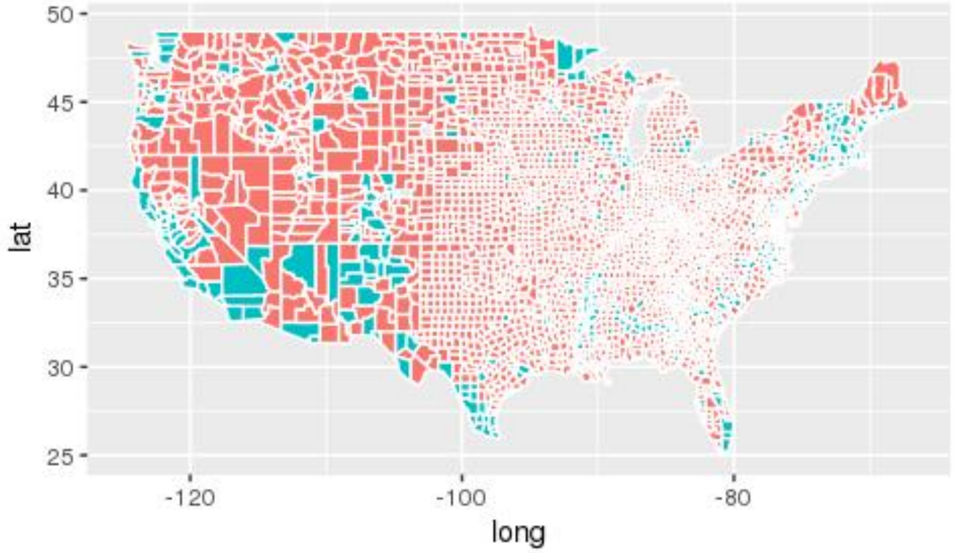
    The decision tree model can be interpreted easily, but it has the corresponding larger test error and easily be overfitting. The logistic regression model has the lowest train error, but it is also easily overfitting. The lasso regression model has the best ROC curve, which means that for this particular problem, the lasso model gives us the best predict, although it slightly has larger train error. There is no general best classifier when we build the model, and we should try different classifiers in different cases.



19. Firstly, I want to apply our model to our census data and try to predict the candidate and map our result. Since the logistic model previouly has the lowest test error, I map the result on the first map, and compare it to the second map. We can see that we predict the most counties right. But there is a apparent drawback in our model. In the real winner map, we can see that the counties that vote for Clinton or Trump are aggregated, which means that they are more likely to be neighborhood. But in our predicted map, we can see that the some small counties that are close each other have different voting result, which is not the common sense. In the real world, people in the small counties nearby should almost have the same idea toward voting, so that we should consider more about the correlations between the small counties that have the similar locations.(The black area denotes there is missing data problem, which is not a big problem compare to the data we already have)



map drawn by the logistic model

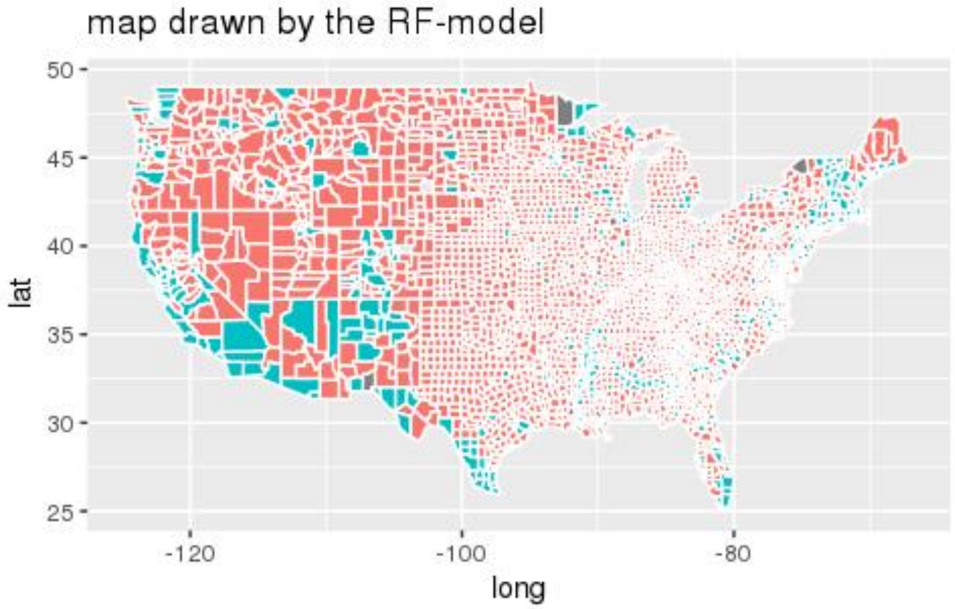


map drawn by the voting result

Now I want to use the random forest method to build the model again. We can find that the train error and test error by random forest method in this particular problem are both smaller than the previous three models.

```
## [1] 0.001221498
## [1] 0.04560261
```

Now let me remap our result for predicting on the census data. We can see that this map is almost same to the map that drawn by the voting data, and it also shows taht the small counties that are close to each other are more likely to vote the same person, relieving the symptom of the previous modeling map.



In all, the random forest model has done the best prediction compared to previous models. From all the models we build above, I think I can experimentally generate the prediction result based on the amount of resources they have. When people in a county tend to have really small amount or they are mid-class or above, this county is more likely to vote for Clinton, and that is true in the real life. Since people who are very poor and have many relatives who are illegal residents, they hope Clinton could bring them more chance to live. People who are in mid-class or above are more likely to care about the factors such as environment, ethinic and stability, so that they will also tend to favor the policy made by Clinton. On the contrary, people who are in the mid-lower class want less competition and less tax, so that they will be inclined to choose the guy who want to dispel all the illegal residents.