PSTAT 131: 2016 Election Analysis

Peter Zhang(5880497), Yuchen Tian(5126735)

2020/12/17

In thie project, we will explore various ways of displaying and analyzing election data in 2016. And we will first improve the way R outputs results using options().

We will also install necessary packages and load them before going into the project.

Background

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

Predicting voter behavior is complicated since there are various type of personalities of voters: angry, resentful, impulsive, passive, frightened. And so, the opinion polls simply cannot predict the results correctly. According to the New York Times "Many voters have stated that they will not vote or vote for a third party candidate, only 9 percent of Americans chose Either Trump or Clinton as their nominee.". This shows the voters' disappointing preference for the candidates. So, this hardship in predicting voter behavior directly result in hardship in forecasting election bahavior. There are 81 percent of Americans said they were afraid of having to choose, the variability of sudden drop of surge of candidates support base on their own public action is unpredictable, which also contribute to the complexity of election forcasting.

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

Nate Silver doesn't rely on "intuition" to make predictions, he relies purely on the simple information he have on hand like people's intentions on voting preference. This special way of modeling change of people's preference over time to predict the actual votes is brilliant in a way that even though it's hard to account for so many variablities, it considers just enough to get a good result in the end. He uses a mathematical model of actual percentage + the house effect + sampling variation to process data and he did a predictive model that reported the results, the exact numbers. In the case of the US election, the model relies on a variety of basic data such as general election polls. Nate Silver, uses basis of the bayesian theory, it is a theory of probability, and the result of the poll data will follow adjust, getting a probability sequence, and the probability sequence needs to be constantly updated. With the election date getting closer, eventually his valid data information is becoming stronger and the algorithms

for forecasting is becoming more accurate. At the end, final prediction and the true result is very close.

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

The failure of prediction was first due to a sense of elitism, which relies on outdated experience, unchanging polls and media dominated by the same group of elite for analysis and prediction. Second, it is misled by the media and polls. In this information society, the media has a great influence on people. The media's tendency of reporting strengthens the standing ground of the elites, while the elites who frequently appear in the media solidify the orientation of the report which didn't accurately predict the results..

Data

We are now setting the working directory as the file location and manipulating data and convert candidate from string to factor

Election data

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

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.

```
[1] 18345     5
```

Alaska is geographically far removed from the rest of the United States. It's not connected to the rest of the United States since It was bought from Russia. Therefore, AK's various attributes are quite different from those of other American states and are not suitable for this analysis altogether.

Census data

Data wrangling

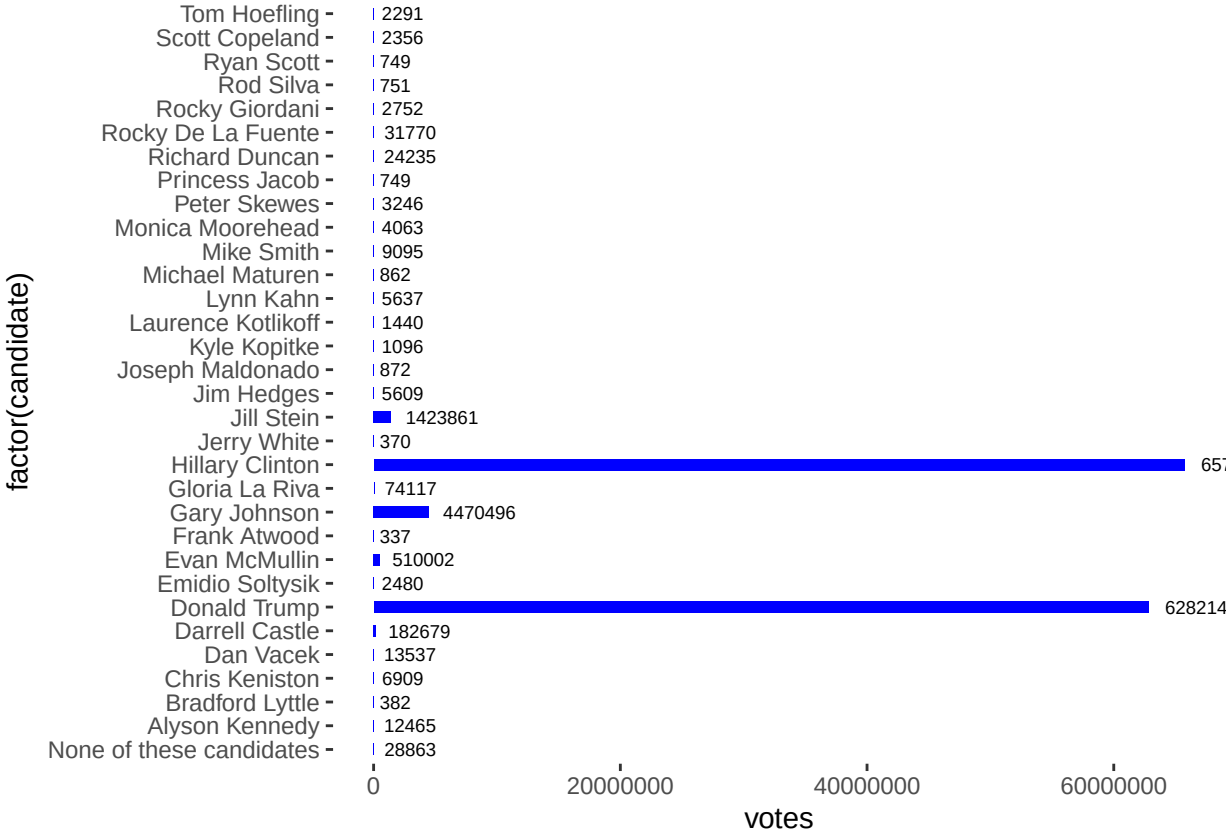
5. Remove summary rows from election.raw data which we have done, and please check the results in the rmd file submitted along side with output pdf.

Please check rmd file for raw code.

6. How many named presidential candidates were there in the 2016 election? Draw a bar chart of all votes received by each candidate. You can split this into multiple plots or may prefer to plot the results on a log scale. Either way, the results should be clear and legible!

```
[1] 32
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

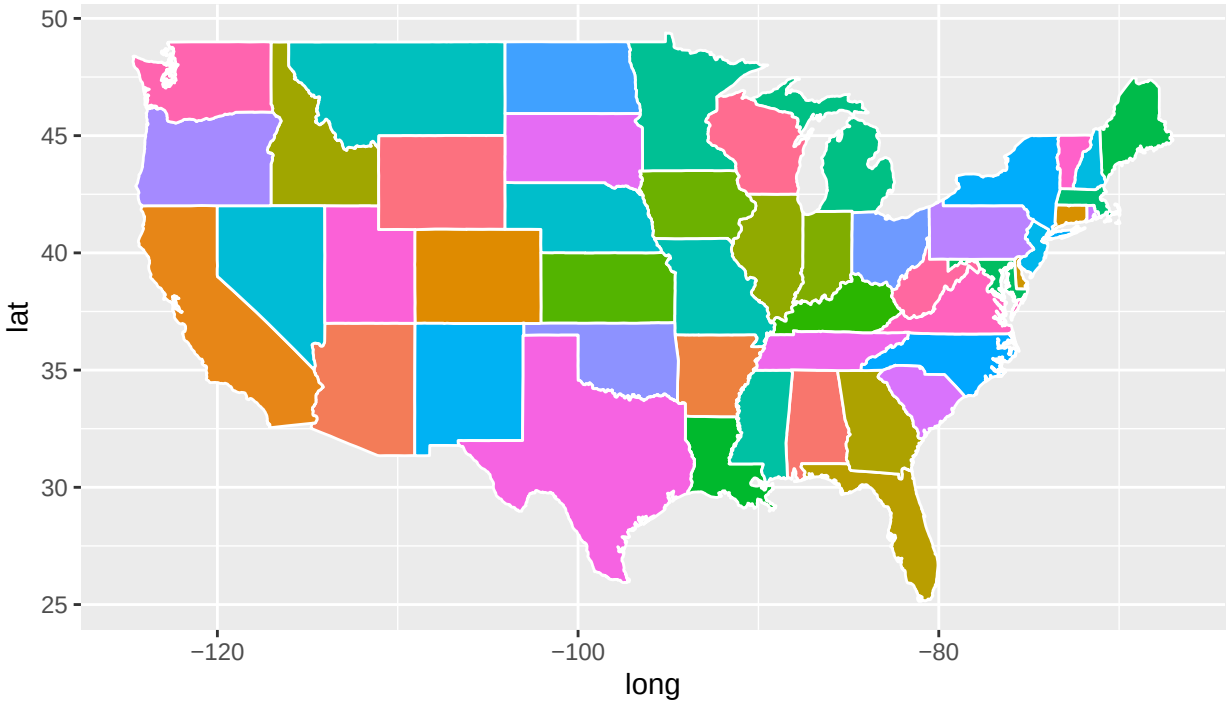


There are 32 named presidential candidates were there in the 2016 election.

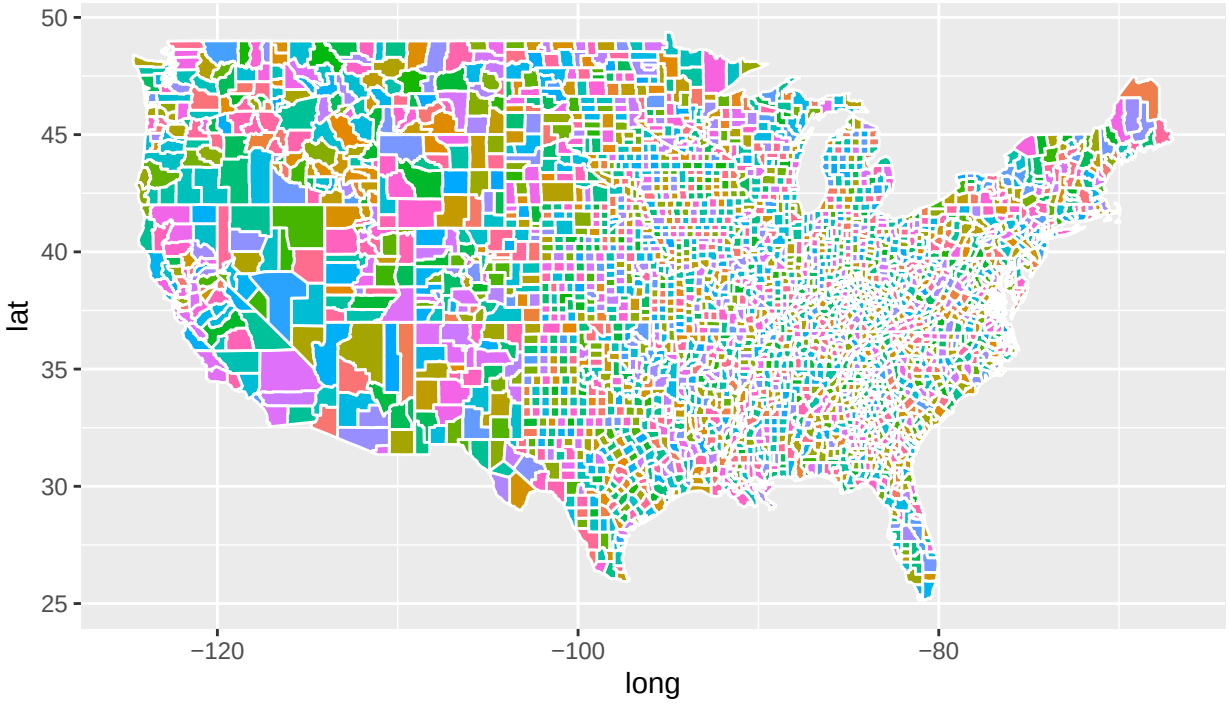
7. Create variables county_winner and state_winner by taking the candidate with the highest proportion of votes. Hint: to create county_winner, start with election, group by fips, compute total votes, and pct = votes/total. Then choose the highest row using top_n (variable state_winner is similar).

Please check rmd file for raw code.

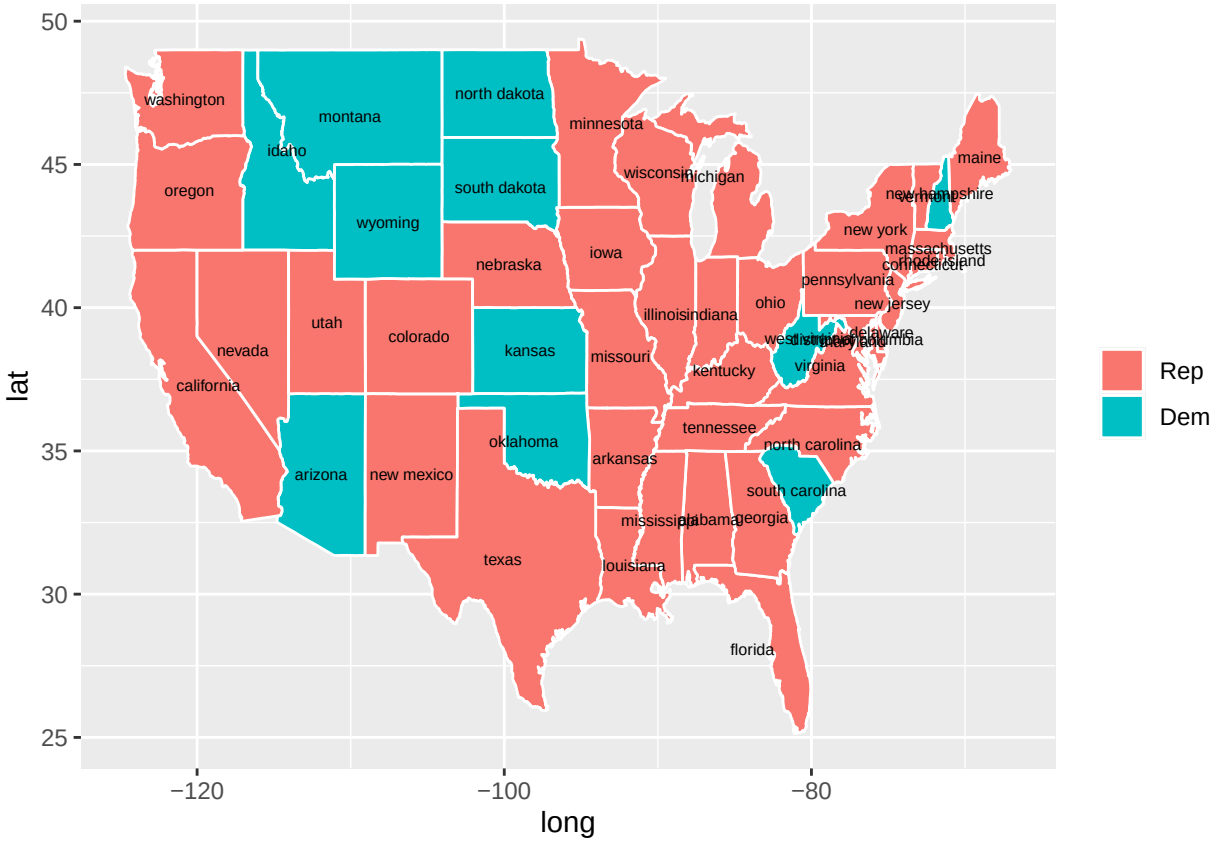
Visualization



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

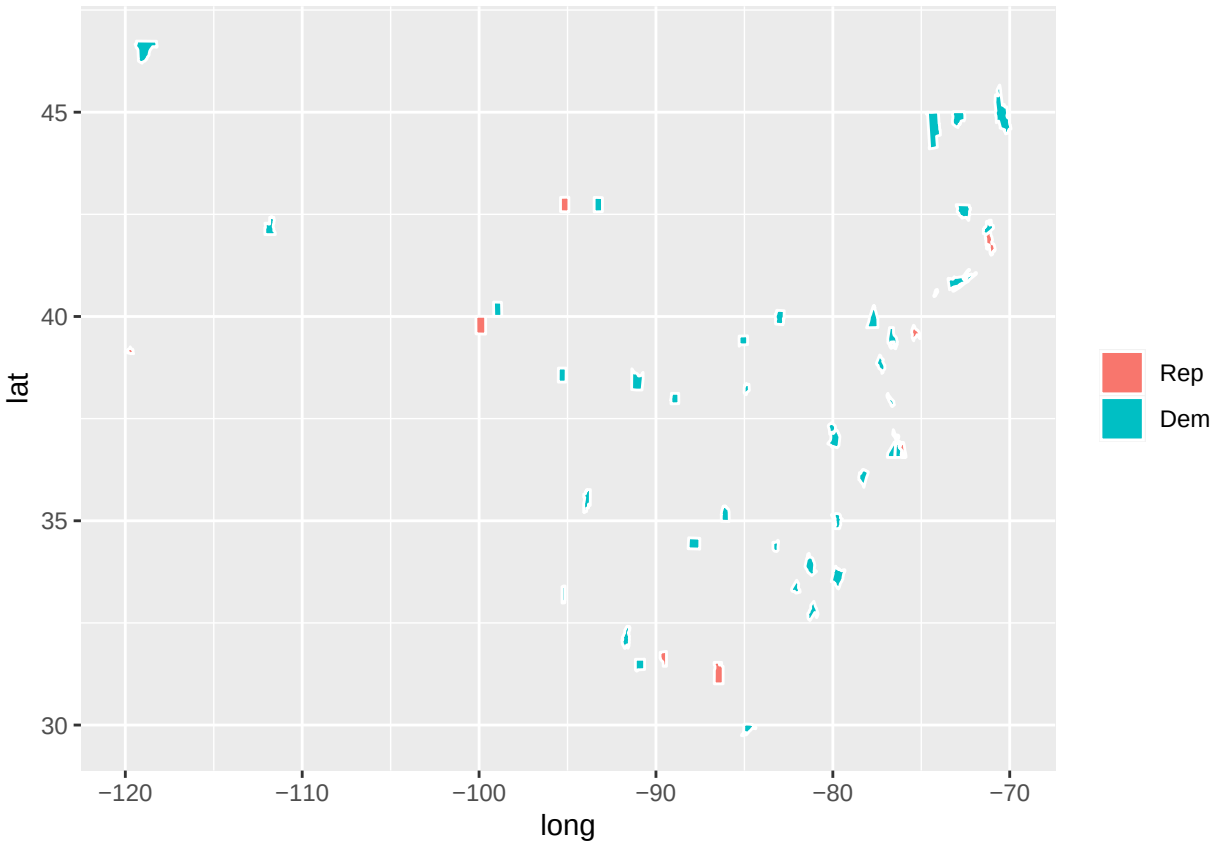


9. Now color the map by the winning candidate for each state. First, combine states variable and state_winner we created earlier using left_join(). Note that left_join() needs to match up values of states to join the tables. A call to left_join() takes all the values from the first table and looks for matches in the second table. If it finds a match, it adds the data from the second table; if not, it adds missing values:



10. The variable county does not have fips column. So we will create one by pooling information from maps::county.fips. Split the polyname column to region and subregion. Use left_join() combine county.fips into county. Also, left_join() previously created variable county_winner. Your figure will look similar to county-level New York Times map.

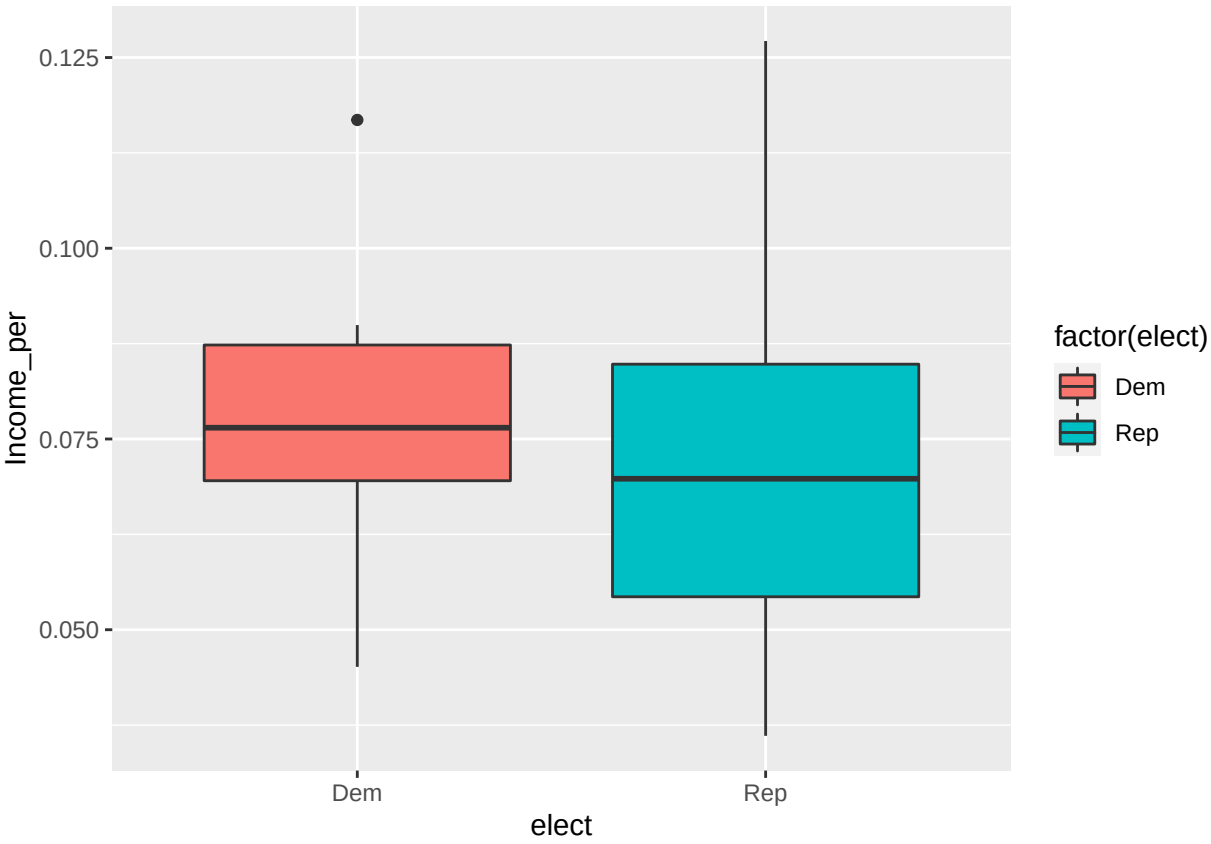
The resulting plot is not showing correctly on the outout pdf, but it ran well on rmd file. So, I will attach the plot in another pdf file for reference.



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.

```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
Warning: Removed 1828 rows containing non-finite values (stat_boxplot).
```



There was no significant difference in average Income_per between counties that chose Trump and those that did not.

12. The census data contains high resolution information (more fine-grained than county-level). In this problem, we aggregate the information into county-level data by computing TotalPop-weighted average of each attributes for each county.

Please check rmd file for raw code.

Dimensionality reduction

13. Run PCA for both county & sub-county level data. Save the first two principle components PC1 and PC2 into a two-column data frame, call it ct.pc and subct.pc, respectively. Discuss whether you chose to center and scale the features before running PCA and the reasons for your choice. What are the three features with the largest absolute values of the first principal component? Which features have opposite signs and what does that mean about the correaltion between these features?

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.5790667 2.1653077 1.8217090 1.26201682 1.18587674
Proportion of Variance 0.2660634 0.1875423 0.1327449 0.06370746 0.05625215
Cumulative Proportion  0.2660634 0.4536057 0.5863506 0.65005809 0.70631023
```

```
                          Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.01967570 1.01306785 0.93206393 0.91227077 0.88100349
Proportion of Variance 0.04158954 0.04105226 0.03474973 0.03328952 0.03104669
Cumulative Proportion  0.74789977 0.78895203 0.82370176 0.85699128 0.88803796
                          Comp.11    Comp.12    Comp.13    Comp.14     Comp.15
Standard deviation     0.83364990 0.74446559 0.67759034 0.61125137 0.484908786
Proportion of Variance 0.02779889 0.02216916 0.01836515 0.01494513 0.009405461
Cumulative Proportion  0.91583685 0.93800601 0.95637116 0.97131629 0.980721750
                          Comp.16     Comp.17     Comp.18     Comp.19
Standard deviation     0.430538962 0.328443078 0.257297365 0.220947122
Proportion of Variance 0.007414552 0.004314994 0.002648077 0.001952705
Cumulative Proportion  0.988136302 0.992451296 0.995099373 0.997052079
                          Comp.20      Comp.21      Comp.22      Comp.23
Standard deviation     0.176426545 0.1428272294 0.1135694980 0.0794347331
Proportion of Variance 0.001245053 0.0008159847 0.0005159212 0.0002523951
Cumulative Proportion  0.998297132 0.9991131162 0.9996290375 0.9998814326
                           Comp.24       Comp.25
Standard deviation     0.0500992167 0.02131325023
Proportion of Variance 0.0001003973 0.00001817019
Cumulative Proportion  0.9999818298 1.00000000000

Loadings:
               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
TotalPop        0.375                0.110
Citizen         0.377
Income          0.379                0.104
IncomeErr       0.380                0.119
IncomePerCap    0.382
IncomePerCapErr 0.381                0.107
Men                        -0.131  0.296 -0.144 -0.389  0.495 -0.303  0.162
White                -0.320  0.255  0.141                              -0.389
Poverty               0.384 -0.188         0.162  0.124                -0.108
ChildPoverty          0.388 -0.151         0.175  0.126                -0.120
Professional    0.124 -0.267 -0.150 -0.391                       -0.155  0.183
Service               0.227 -0.169                -0.252  0.399  0.191 -0.459
Office                0.106  0.171 -0.432                        0.488
Production     -0.102  0.110  0.281  0.464         0.333 -0.152 -0.234
Drive                 0.162  0.376 -0.131  0.374         0.241         0.160
Carpool                            0.407 -0.290 -0.267 -0.255  0.578  0.310
Transit         0.225              -0.102 -0.227 -0.170 -0.360 -0.370 -0.278
OtherTransp                -0.213        -0.534  0.330                -0.274
WorkAtHome           -0.296 -0.282         0.184        -0.110
MeanCommute           0.135  0.164         0.106 -0.571 -0.462 -0.107 -0.135
Employed             -0.357  0.116 -0.121 -0.192  0.164                0.167
PrivateWork                 0.449        -0.170  0.111
SelfEmployed         -0.236 -0.280  0.129  0.322        -0.112
FamilyWork           -0.140 -0.197  0.145  0.313  0.196 -0.135  0.169 -0.248
Minority              0.323 -0.250 -0.139                              0.391
               Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
TotalPop
Citizen
Income
IncomeErr
IncomePerCap
```

```
IncomePerCapErr
Men              0.368  -0.214  -0.243  -0.165   0.169   0.207
White                    0.193          -0.279
Poverty                  0.119          -0.273   0.257   0.194   0.204
ChildPoverty             0.146          -0.244   0.262   0.266   0.175
Professional                     0.351  -0.290   0.252                   -0.607
Service         -0.429                   0.344                           -0.350
Office           0.399          -0.480  -0.178                   0.157  -0.222
Production                      -0.163                  -0.224   0.240  -0.586
Drive                            0.181          -0.152   0.126          -0.111
Carpool         -0.167           0.130  -0.204                           -0.114
Transit         -0.238  -0.246  -0.303  -0.254  -0.270          -0.124
OtherTransp      0.448   0.227   0.285   0.130           0.173
WorkAtHome               0.169  -0.305   0.313   0.511  -0.362
MeanCommute      0.360           0.176   0.364           0.150   0.196
Employed        -0.225  -0.129  -0.100   0.216           0.411   0.640   0.129
PrivateWork              -0.145  -0.150   0.124   0.488   0.370  -0.524
SelfEmployed             0.258  -0.315   0.121  -0.384   0.518  -0.252  -0.208
FamilyWork       0.194  -0.750   0.253
Minority                 -0.193          0.286
                Comp.18 Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24 Comp.25
TotalPop                         0.457   0.708   0.303   0.119
Citizen                          0.663  -0.509  -0.359
Income                          -0.375   0.217  -0.539   0.281          -0.520
IncomeErr                       -0.283   0.124  -0.156  -0.779           0.314
IncomePerCap                    -0.296  -0.209           0.526           0.646
IncomePerCapErr                 -0.179  -0.357   0.673                   -0.460
Men
White                                                                    0.706
Poverty          0.402   0.591
ChildPoverty    -0.485  -0.499
Professional
Service
Office
Production
Drive           -0.538   0.445
Carpool         -0.197   0.169
Transit         -0.293   0.226
OtherTransp     -0.221   0.182
WorkAtHome      -0.314   0.249
MeanCommute
Employed        -0.103
PrivateWork      0.114
SelfEmployed
FamilyWork
Minority                                                                 0.701

Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.6321432 1.7701931 1.39845232 1.21272795 1.15427055
Proportion of Variance 0.2771271 0.1253433 0.07822676 0.05882836 0.05329362
Cumulative Proportion  0.2771271 0.4024705 0.48069722 0.53952558 0.59281920
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.08032163 1.01754716 0.97503985 0.93684700 0.91102384
```

```
Proportion of Variance 0.04668379 0.04141609 0.03802811 0.03510729 0.03319858
Cumulative Proportion  0.63950299 0.68091908 0.71894719 0.75405448 0.78725306
                          Comp.11    Comp.12    Comp.13    Comp.14    Comp.15
Standard deviation     0.89628034 0.83138008 0.82841691 0.74914025 0.74189855
Proportion of Variance 0.03213274 0.02764771 0.02745098 0.02244844 0.02201654
Cumulative Proportion  0.81938580 0.84703351 0.87448450 0.89693294 0.91894948
                          Comp.16    Comp.17    Comp.18    Comp.19     Comp.20
Standard deviation     0.67907076 0.66174430 0.61850755 0.53865878 0.465433625
Proportion of Variance 0.01844548 0.01751622 0.01530206 0.01160613 0.008665138
Cumulative Proportion  0.93739496 0.95491118 0.97021325 0.98181938 0.990484517
                           Comp.21     Comp.22     Comp.23     Comp.24
Standard deviation     0.301650578 0.231914893 0.228275952 0.194172830
Proportion of Variance 0.003639723 0.002151381 0.002084396 0.001508124
Cumulative Proportion  0.994124240 0.996275621 0.998360017 0.999868140
                           Comp.25
Standard deviation     0.0574150556
Proportion of Variance 0.0001318595
Cumulative Proportion  1.0000000000

Loadings:
               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
TotalPop                      0.293  0.364  0.102         0.488
Men                          -0.130  0.276 -0.405 -0.163  0.144         0.652
White           0.242  0.319 -0.114 -0.141 -0.145 -0.236
Citizen         0.169  0.233 -0.221 -0.380  0.133 -0.147
Income          0.315 -0.140  0.144  0.183
IncomeErr       0.212 -0.224                       0.213 -0.295         0.182
IncomePerCap    0.333 -0.165                       0.156 -0.127
IncomePerCapErr 0.227 -0.207 -0.105 -0.124         0.298 -0.276
Poverty        -0.309        -0.188 -0.153         0.187
ChildPoverty   -0.303        -0.159 -0.118         0.193 -0.115
Professional    0.321 -0.141                       0.128  0.102        -0.146
Service        -0.277
Office                        0.240 -0.219  0.468         0.115  0.343  0.491
Production     -0.219  0.198         0.116 -0.275 -0.170 -0.306
Drive                  0.443  0.150  0.146  0.155  0.250 -0.105 -0.112
Carpool        -0.172                0.334 -0.226  0.121 -0.129  0.104
Transit               -0.443        -0.221        -0.416
OtherTransp           -0.166 -0.209 -0.174 -0.324  0.180  0.493  0.167
WorkAtHome      0.184 -0.108 -0.320  0.112                       0.134  0.215
MeanCommute           -0.285  0.174  0.178  0.203 -0.484 -0.195         0.137
Employed        0.219         0.203        -0.238         0.236        -0.298
PrivateWork                   0.411 -0.233 -0.381        -0.198  0.276  0.148
SelfEmployed                 -0.452  0.305  0.144
FamilyWork                   -0.212  0.178        -0.116 -0.128  0.832 -0.220
Minority       -0.244 -0.316  0.114  0.143  0.148  0.234
               Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17
TotalPop                0.533   0.160   0.202   0.217   0.255   0.113   0.130
Men            -0.389  -0.180   0.158  -0.162           0.140
White           0.154   0.156           0.171   0.165                  -0.229
Citizen        -0.128                   0.255                          0.292
Income                                                 -0.145
IncomeErr      -0.116          -0.124   0.139   0.410  -0.230   0.473  -0.255
IncomePerCap                                            0.167  -0.108   0.102
```

IncomePerCapErr					0.180	0.437	-0.205	0.405
Poverty		0.211				0.151		-0.287
ChildPoverty		0.228				0.210		-0.358
Professional	-0.155		0.108		-0.335			-0.210
Service			0.296	-0.159	0.646	-0.263	-0.184	0.177
Office	0.240	-0.311				0.137	0.116	
Production	0.110	0.183	-0.385		-0.110	0.139	0.328	0.341
Drive	-0.182		-0.204	-0.184	0.106		-0.264	
Carpool	0.348	-0.397	0.307	0.542			-0.105	
Transit						0.176	0.269	
OtherTransp			-0.593	0.223		-0.166	-0.158	
WorkAtHome	0.359	0.252	0.208	-0.242	-0.267	-0.458		0.244
MeanCommute			-0.258	0.188			-0.543	
Employed	0.113	-0.315		-0.332		0.226		-0.101
PrivateWork	0.240	0.242		-0.256			-0.222	-0.126
SelfEmployed	0.410		-0.176	-0.313	0.213	0.336		-0.212
FamilyWork	-0.375							
Minority	-0.152	-0.148		-0.176	-0.166			0.229
	Comp.18	Comp.19	Comp.20	Comp.21	Comp.22	Comp.23	Comp.24	Comp.25
TotalPop	0.157							
Men								
White	-0.164	-0.181						-0.708
Citizen	0.516	0.384	-0.225		0.125			
Income	-0.127		-0.702	-0.156	0.475			
IncomeErr	0.351		0.204					
IncomePerCap			-0.346		-0.701	0.355	-0.120	
IncomePerCapErr	-0.152	-0.196	0.299		0.297	-0.141		
Poverty	0.162	-0.120		-0.720			0.225	
ChildPoverty	0.208	-0.264	-0.255	0.595			-0.148	
Professional		0.189			-0.229	-0.712		
Service			-0.198		-0.178	-0.360		
Office		-0.111	-0.103		-0.122	-0.228		
Production	0.164	-0.124	-0.201	-0.107	-0.158	-0.357		
Drive				0.134	-0.157		0.629	
Carpool							0.229	
Transit	-0.158			0.201			0.618	
OtherTransp							0.126	
WorkAtHome	0.197	-0.215					0.191	
MeanCommute	0.216	-0.137	0.138					
Employed	0.516	-0.348						
PrivateWork		0.474						
SelfEmployed		0.387						
FamilyWork								
Minority	0.163	0.187						-0.706

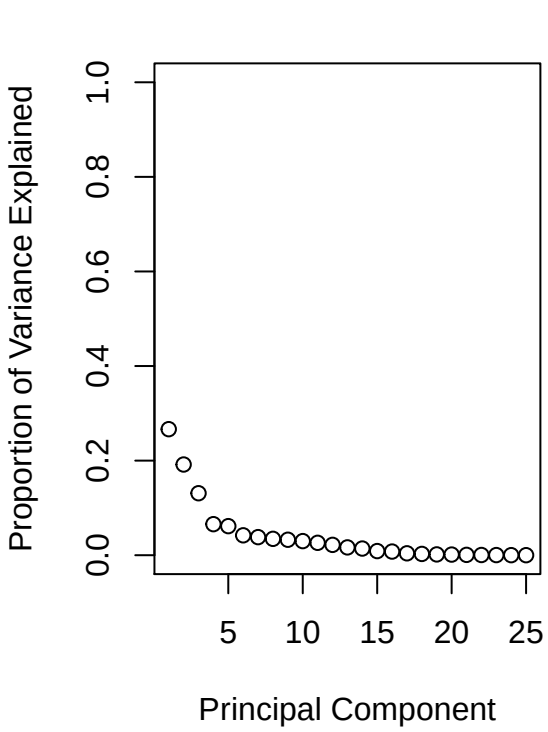
county level data

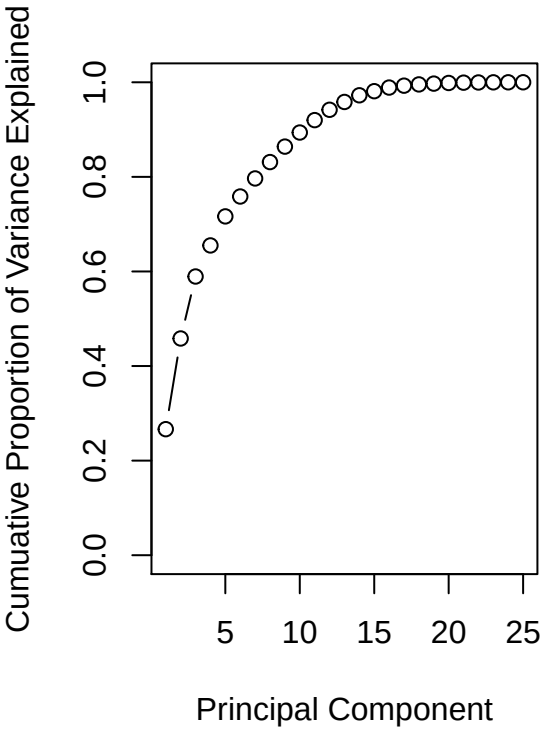
I chose to center and scale the features before running PCA because I want to eliminate the difference in the order of magnitude of each variable. The three features with the largest absolute values of the first principal component are: IncomePerCap, IncomePerCapErr, IncomeErr. For the Comp.1, the features have opposite signs are: Production, SelfEmployed, Men, Drive, FamilyWork, White, Carpool, ChildPoverty, Poverty, WorkAtHome, Service. The correlation between these features is that: features are the opposite of the principal component.

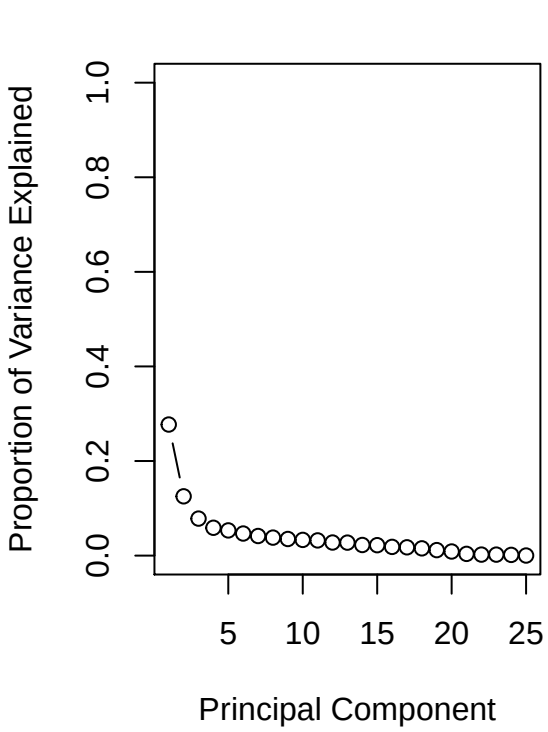
sub-county level data

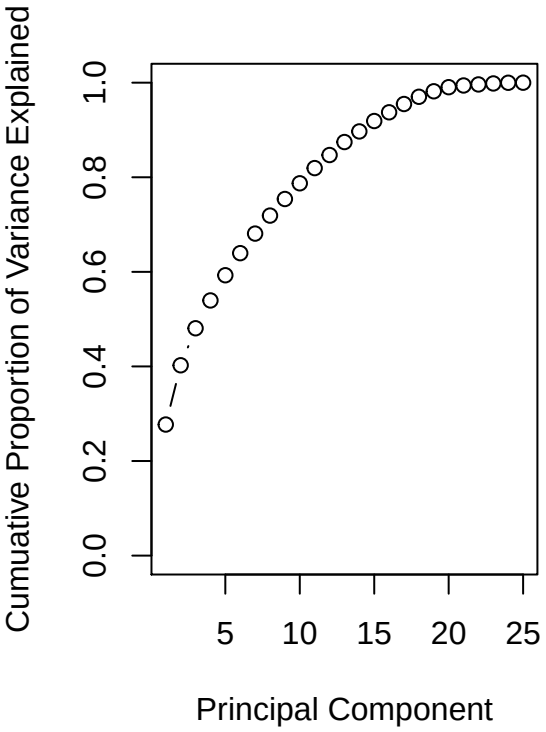
I chose to center and scale the features before running PCA, because I want to eliminate the difference in the order of magnitude of each variable. The three features with the largest absolute values of the first principal component are: IncomePerCap, Professional, Income. For the Comp.1, the features have opposite signs are: Poverty, ChildPoverty, Service, Minority, Production, Carpool, Transit, PrivateWork, OtherTransp, Office. The correaltion between these features is that: that features are the opposite of 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. Plot proportion of variance explained (PVE) and cumulative PVE for both county and sub-county analyses.









county level data (The first 2 plots)

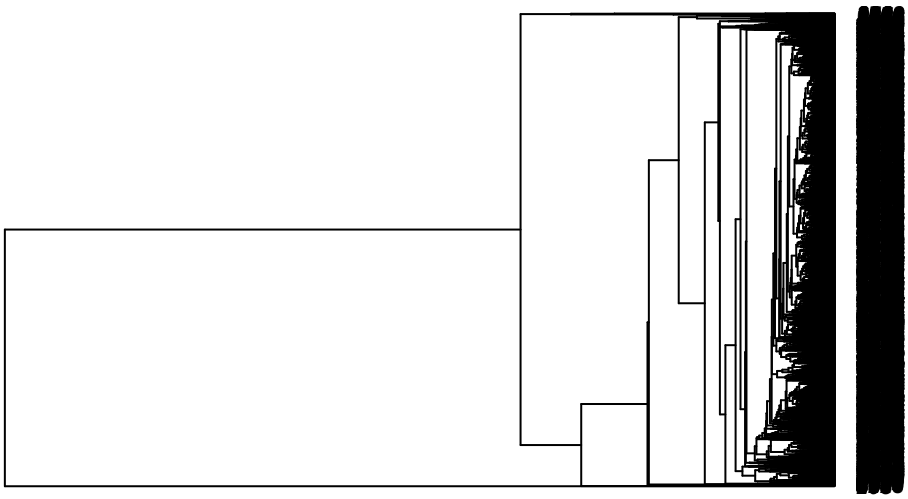
The number of minimum number of PCs needed to capture 90% of the variance for the county analyses is 11.

sub-county level data (The second 2 plots)

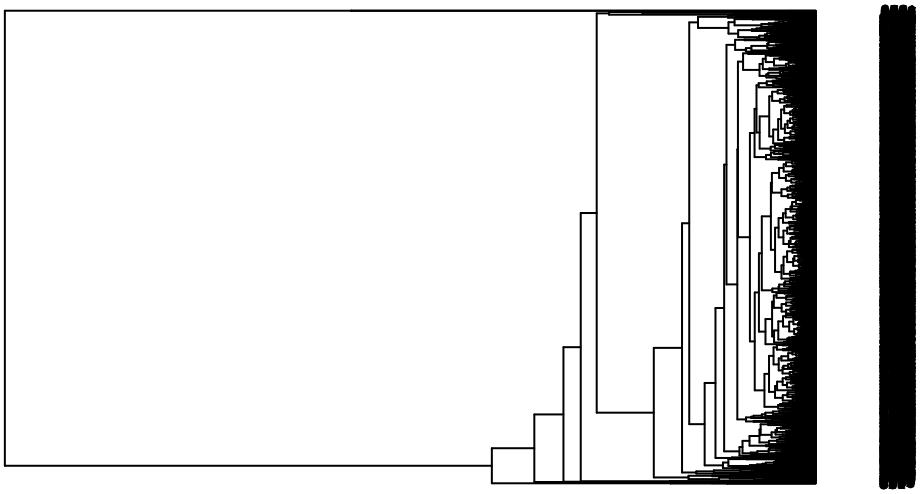
The number of minimum number of PCs needed to capture 90% of the variance for the sub-county analyses is 15.

Clustering

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



```
[1] 1
```



```
[1] 1
```

First one is the Dentrogram with census.ct

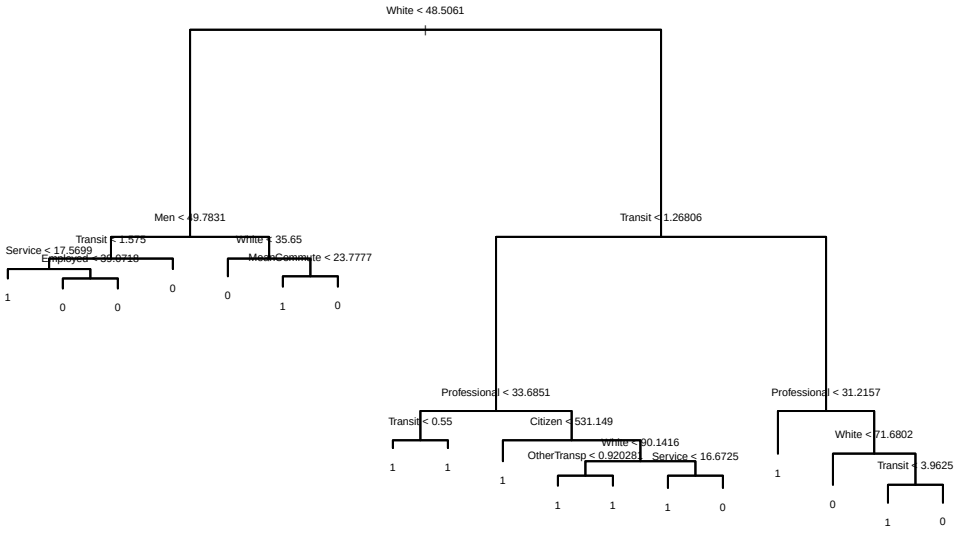
Second one is the Dentrogram with first 5 principle components of ct.pc

The second way seemed to put San Mateo County in a more appropriate clusters because the counties that have similar attributes to San Mateo County are grouped together.

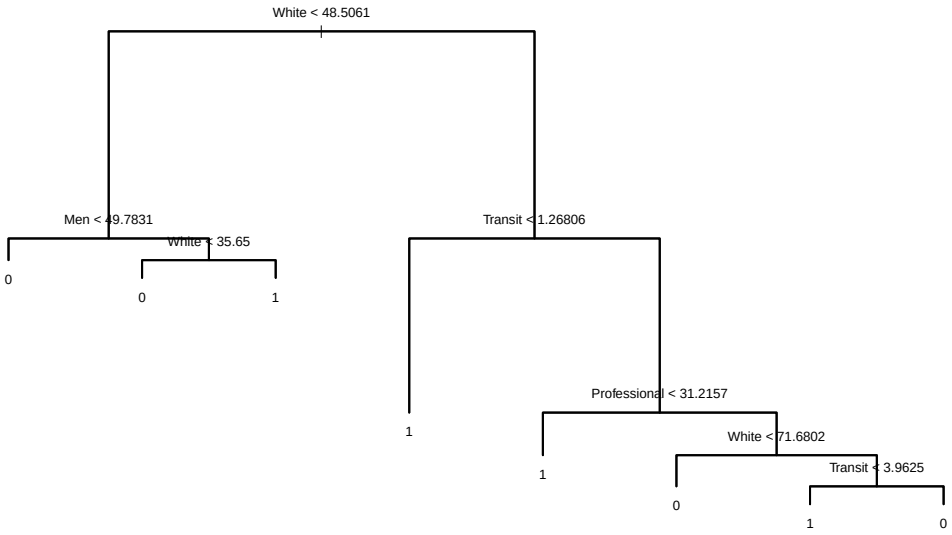
Classification

16. Decision tree: train a decision tree by cv.tree(). Prune tree to minimize misclassification error. Be sure to use the folds from above for cross-validation. Visualize the trees before and after pruning. Save training and test errors to records variable. Intepret and discuss the results of the decision tree analysis. Use this plot to tell a story about voting behavior in the US (remember the NYT infographic?)

```
Classification tree:
tree(formula = factor(class) ~ ., data = trn.cl)
Variables actually used in tree construction:
[1] "White"        "Men"          "Transit"      "Service"      "Employed"
[6] "MeanCommute"  "Professional" "Citizen"      "OtherTransp"
Number of terminal nodes:  18
Residual mean deviance:  0.2784 = 397 / 1426
Misclassification error rate: 0.04848 = 70 / 1444
```




```
$size
[1] 18 15 13  9  8  6  4  2  1

$dev
[1] 126 126 121 123 121 124 129 142 233

$k
[1]  -Inf  0.00  2.00  2.75  3.00 11.00 11.50 19.00 66.00

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```



If the population of white <46.08% and the population of employed in service jobs and Mean commute time (minutes) <23.98, then the person elected is elected; If the population of white <46.08% and the population of employed in service jobs and Mean commute time (minutes) >=23.98, then the chosen man was not elected; If the population of white >=46.08% and the population of commuting on public transportation <2.71, then the person elected is elected; If the population of white >=46.08% and the population of commuting on public transportation >=2.71 and the population of employed in management <31.22, then the person elected is elected; If the population of white >=46.08% and the population of commuting on public transportation >=2.71 and the population of employed in management >=31.22 and the population of white <74.49%, then the chosen man was not elected; If the population of white >=46.08% and the population of commuting on public transportation >=2.71 and the population of employed in management >=31.22 and the population of white >=74.49% and of the population of children under poverty level <21.8%, then the chosen man was not elected; If the population of white >=46.08% and the population of commuting on public transportation >=2.71 and the population of employed in management >=31.22 and the population of white >=74.49% and of the population of children under poverty level >=21.8%, then the person elected is elected.

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

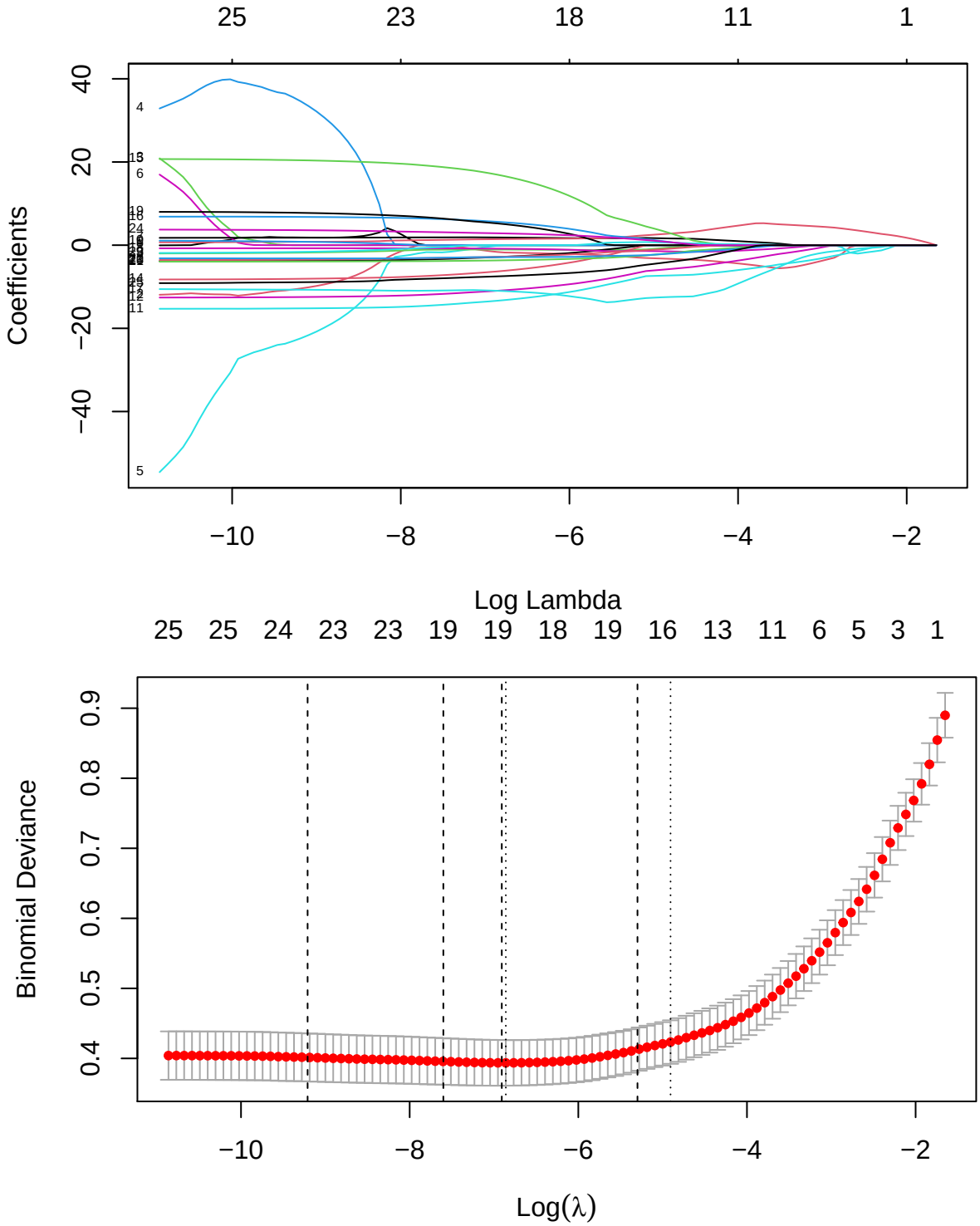
```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred


Call:
glm(formula = factor(class) ~ ., family = binomial(link = "logit"),
    data = trn.cl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1184   0.0592   0.1310   0.2746   2.5192

Coefficients:
                    Estimate    Std. Error z value         Pr(>|z|)
(Intercept)     11.9376900633 10.5739086470   1.129         0.258908
```

```
TotalPop        -0.0000008709  0.0000028046  -0.311         0.756176
Citizen         -0.0000480723  0.0002269621  -0.212         0.832257
Income           0.0000005117  0.0000007153   0.715         0.474397
IncomeErr        0.0000004459  0.0000026123   0.171         0.864457
IncomePerCap    -0.0000017855  0.0000018177  -0.982         0.325964
IncomePerCapErr  0.0000056062  0.0000075500   0.743         0.457758
Men              0.0620580784  0.0601112085   1.032         0.301890
White            0.0077821408  0.0642883593   0.121         0.903651
Poverty         -0.0385844224  0.0464630009  -0.830         0.406294
ChildPoverty     0.0160579840  0.0325438228   0.493         0.621711
Professional    -0.2778079742  0.0432253040  -6.427 0.00000000013017 ***
Service         -0.3848000098  0.0551217163  -6.981 0.00000000000293 ***
Office          -0.1237257063  0.0537008385  -2.304         0.021224 *
Production      -0.2030492260  0.0470097623  -4.319 0.00001565254935 ***
Drive            0.2351116382  0.0518732409   4.532 0.00000583100409 ***
Carpool          0.2588612217  0.0704388624   3.675         0.000238 ***
Transit         -0.1761572456  0.1180980294  -1.492         0.135799
OtherTransp     -0.0650402794  0.1147919706  -0.567         0.570991
WorkAtHome       0.2154874965  0.0855741662   2.518         0.011798 *
MeanCommute     -0.1081951797  0.0268522333  -4.029 0.00005594783130 ***
Employed        -0.0925992570  0.0353106641  -2.622         0.008731 **
PrivateWork     -0.0533671288  0.0261772485  -2.039         0.041482 *
SelfEmployed    -0.0532936682  0.0575909829  -0.925         0.354767
FamilyWork       0.6235119717  0.4989339292   1.250         0.211413
Minority        -0.0923540650  0.0640159593  -1.443         0.149113
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1289.35  on 1443  degrees of freedom
Residual deviance:  514.74  on 1418  degrees of freedom
AIC: 566.74

Number of Fisher Scoring iterations: 8
```

The significant variables are: Professional , Service, Office, Production, Drive, Carpool, WorkAtHome, MeanCommute, Employed, PrivateWork. These are not consistent with what I saw in decision tree analysis. The coefficient of "Service" is -0.38, this means: with other conditions unchanged, if the value of "Service" increases by a unit, then the value of OR reduce by 0.38 unit.

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

variable.




```
 [1]  1  2  3  5  6  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

 [1]   5.2022436   1.8513714 -10.4807105  34.8523071 -22.5766555   1.7937778
 [7]   0.8050371  -1.8004533   0.8314311 -15.2062831 -12.4781248  -3.7736540
[13]  -8.0977423  20.4236665   6.7966251 -10.7103608  -0.7464240   7.7783351
[19]  -3.4783836  -3.9372267  -3.1779642  -1.6234772   3.6078881  -8.9131386
```

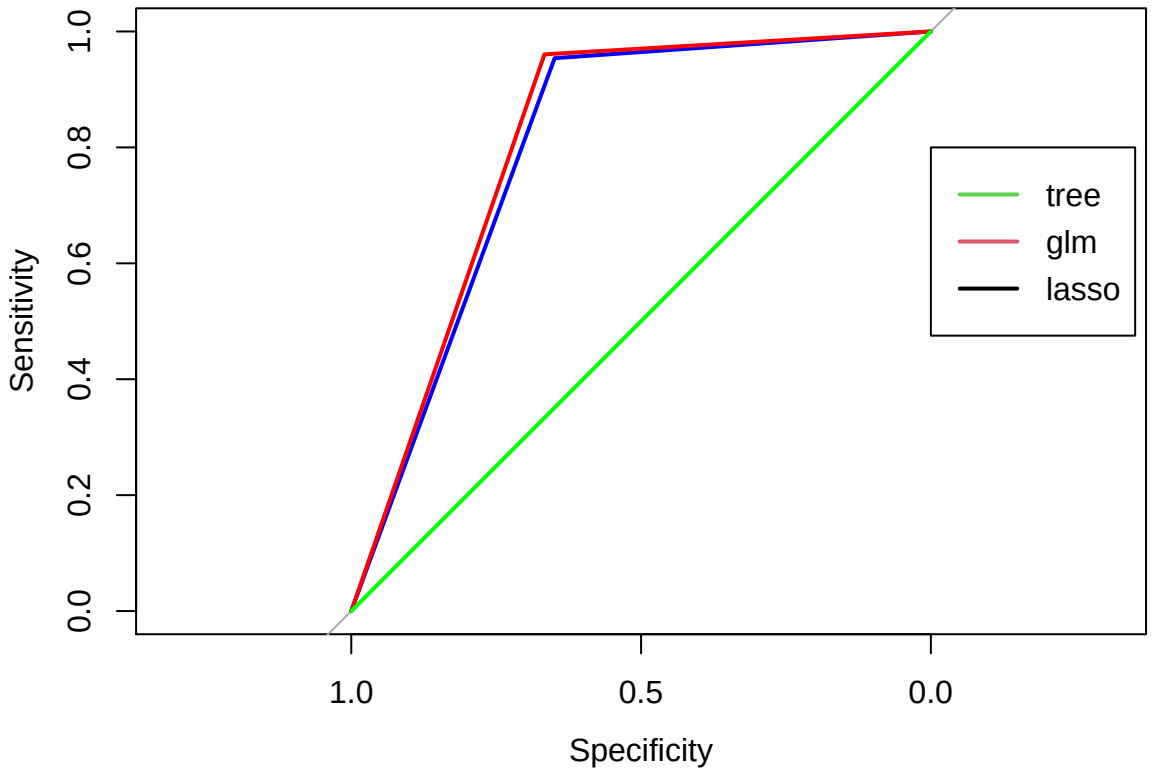
```
 [1] "(Intercept)"  "TotalPop"     "Citizen"      "IncomeErr"    "IncomePerCap"
 [6] "Men"          "White"        "Poverty"      "ChildPoverty" "Professional"
[11] "Service"      "Office"       "Production"   "Drive"        "Carpool"
[16] "Transit"      "OtherTransp"  "WorkAtHome"   "MeanCommute"  "Employed"
[21] "PrivateWork"  "SelfEmployed" "FamilyWork"   "Minority"

 [1]  1  6  8  9 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

 [1]   4.5996348  -1.7143498   1.8711479   1.2503829  -1.0870399 -14.4626406
 [7] -11.8190675  -3.3716153  -7.3163883  18.9634782   6.3206952 -10.9254501
[13]  -0.7731715   6.5521767  -3.0949433  -3.8271127  -3.0400398  -0.5413303
[19]   3.1367453  -8.0247541

 [1] "(Intercept)"  "IncomePerCap" "Men"          "White"        "Poverty"
 [6] "Professional" "Service"      "Office"       "Production"   "Drive"
[11] "Carpool"      "Transit"      "OtherTransp"  "WorkAtHome"   "MeanCommute"
[16] "Employed"     "PrivateWork"  "SelfEmployed" "FamilyWork"   "Minority"

 [1]  1  3  6  8  9 10 12 13 14 15 16 17 18 19 20 21 22 23 25 26

 [1]   4.5286273  -1.5197860  -0.2489407   1.8632798   1.4342949  -1.0510886
 [7] -13.4496958 -11.0318420  -2.9021950  -6.4017377  17.1491161   5.6819770
[13] -10.9344535  -0.8361392   5.3242181  -2.8322816  -3.6775592  -2.8832589
[19]   2.8765262  -7.5409848

 [1] "(Intercept)"  "Citizen"      "IncomePerCap" "Men"          "White"
 [6] "Poverty"      "Professional" "Service"      "Office"       "Production"
[11] "Drive"        "Carpool"      "Transit"      "OtherTransp"  "WorkAtHome"
[16] "MeanCommute"  "Employed"     "PrivateWork"  "FamilyWork"   "Minority"

 [1]  1  3  8  9 10 12 13 14 15 16 17 18 19 21 22 23 24 25 26

 [1]   7.0570652  -2.8180588   1.6953533   2.0584015  -1.1060735  -8.4057910
 [7]  -7.1004615  -0.4569762  -1.1683681   5.7208584   1.8548159 -13.2067188
[13]  -1.4259573  -1.4635598  -2.7425365  -2.5093171   0.6865240   1.5484457
[19]  -5.3339000

 [1] "(Intercept)"  "Citizen"      "Men"          "White"        "Poverty"
 [6] "Professional" "Service"      "Office"       "Production"   "Drive"
[11] "Carpool"      "Transit"      "OtherTransp"  "MeanCommute"  "Employed"
[16] "PrivateWork"  "SelfEmployed" "FamilyWork"   "Minority"

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


Call:
glm(formula = factor(class) ~ IncomePerCap + Men + White + Poverty +
    Professional + Service + Office + Production + Drive + Carpool +
    Transit + OtherTransp + WorkAtHome + MeanCommute + Employed +
    PrivateWork + SelfEmployed + FamilyWork + Minority, family = binomial(link = "logit"),
    data = trn.cl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1168   0.0584   0.1326   0.2782   2.5876

Coefficients:
             Estimate Std. Error z value          Pr(>|z|)
(Intercept)    5.6986     8.8909   0.641          0.521555
```

```
IncomePerCap  -1.4313     5.8638  -0.244          0.807159
Men            1.8247     1.6772   1.088          0.276595
White          0.5112     6.2703   0.082          0.935022
Poverty       -1.1636     1.3916  -0.836          0.403068
Professional -15.8059     2.3547  -6.713 0.000000000019124 ***
Service      -12.8386     1.7950  -7.152 0.000000000000854 ***
Office        -3.9546     1.6525  -2.393          0.016706 *
Production    -8.4034     1.9035  -4.415 0.000010118442853 ***
Drive         20.8911     4.5352   4.606 0.000004095899424 ***
Carpool        7.0301     1.8402   3.820          0.000133 ***
Transit      -11.2351     6.9483  -1.617          0.105887
OtherTransp   -0.7198     1.3147  -0.548          0.584024
WorkAtHome     8.1566     3.1518   2.588          0.009656 **
MeanCommute   -3.4309     0.8802  -3.898 0.000096990712201 ***
Employed      -4.0036     1.4644  -2.734          0.006256 **
PrivateWork   -3.3300     1.5048  -2.213          0.026906 *
SelfEmployed  -1.5860     2.0065  -0.790          0.429268
FamilyWork     3.4477     2.9611   1.164          0.244293
Minority      -9.1451     6.2535  -1.462          0.143634
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1289.35  on 1443  degrees of freedom
Residual deviance:  517.55  on 1424  degrees of freedom
AIC: 557.55

Number of Fisher Scoring iterations: 7
```

5*1e-4 is the optimal value of lambda in cross validation. The non-zero coefficients in the LASSO regression for the optimal value of lambda are: IncomePerCap, Men, White, Poverty, Professional, Service, Office, Production, Drive, Carpool, Transit, OtherTransp, WorkAtHome, MeanCommute, Employed, PrivateWork, SelfEmployed, FamilyWork, Minority. Comparing to the unpenalized logistic regression, they have the same error rate.

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

```
Setting levels: control = 0, case = 1

Setting direction: controls > cases

Setting levels: control = 0, case = 1

Setting direction: controls < cases

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```



According to the ROC curve, logistic regression has the highest AUC value.

The decision tree is a tree-like structure with roots, branches and leaves as its basic components. Its structure simulates the growth process of trees. Each leaf node corresponds to a classification, while non-leaf nodes correspond to a division in a certain characteristic attribute. The decision tree uses the attributes of the sample as nodes. A tree structure that uses the value of an attribute as a branch. It is produced by analyzing and summarizing the attributes of a large number of samples based on information theory. The root of the decision tree is the attribute with the most information in all samples. The middle node of the tree is the attribute with the most information in the sample subset contained in the subtree whose root is this node. The leaves of the decision tree are the class values of the samples. The decision tree is a form of knowledge representation. It is a high level summary of all the sample data. In other words, the decision tree can accurately identify the categories of all samples. It can also effectively identify the categories of new samples. The key to construct a good decision tree is how to choose good logical judgment or attribute. The decision tree model is easy to explain in business, and it also presents intuitive decision rules, which have been understood and used by people. However, due to the overfitting problem of decision tree model, the effect of decision tree model on the training set is better than that on the test set, which reduces its predictive ability to some extent.

The logistic regression is a classical classification algorithm, which is commonly used in dichotomy problems. Logistic regression can directly model the classification possibility without assuming the distribution of data, thus avoiding the problem caused by inaccurate assumption distribution.Moreover, the logistic regression has a simple form, and the model has a very good interpretability. The influence of different features on the final result can be seen from the weight of features. But at the same time, logistic regression is difficult to fit the real distribution of data because of its very simple form, so its accuracy is not very high. The logistic regression is a prediction method with a long history of application, that is, a relatively

mature and robust model. The prediction result of logistic regression is not a discrete value or an exact category, but a probability list of each prediction sample, and then the user can set different criteria to analyze this probability score, obtain a threshold, and then categorize the output in the way that best suits the business problem.Logistic regression is characterized by its wide range of applications and its flexibility and convenience. However, with the development of data science and the rapid development of prediction models, the performance of logistic regression is not so good among many prediction models. Especially when the number of variables is large, the predictive power of logistic regression is usually not good.

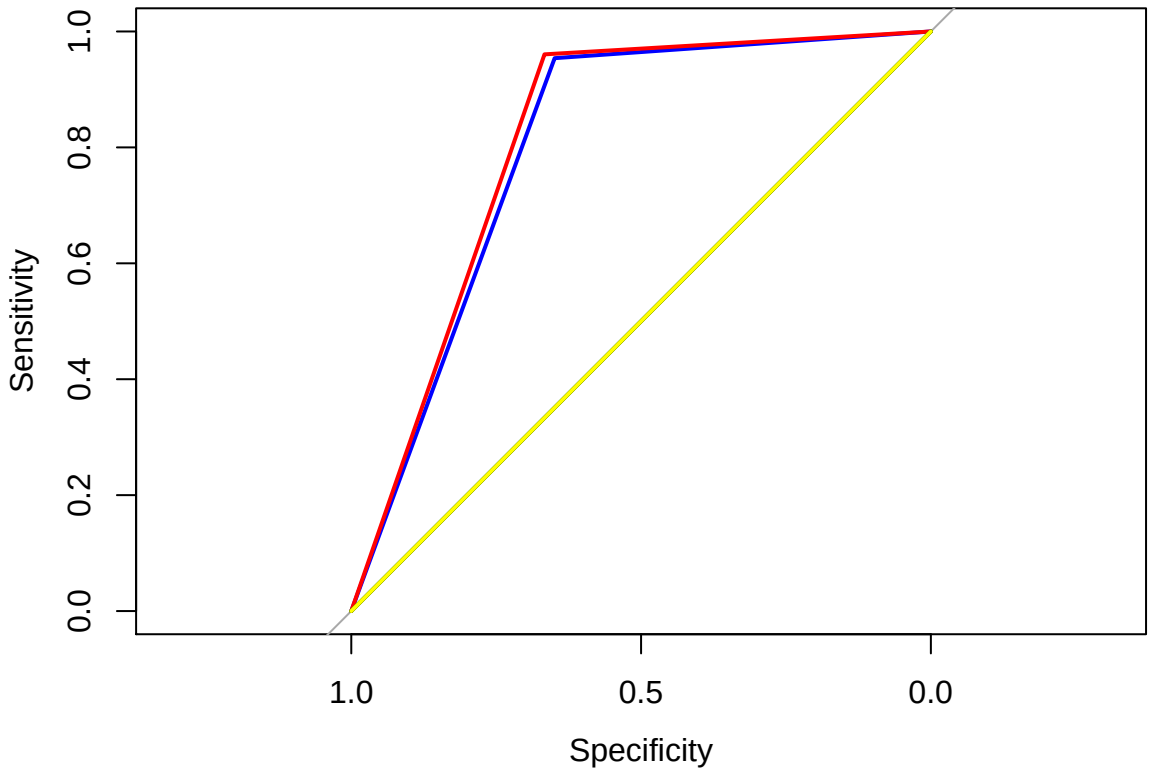
From the results, the different classifiers are not appropriate for the questions about the election.

20. 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!

I explored an additional classification methods——KNN. The KNN model is better than tree method, but worse than logistic regression method. Please check rmd file for raw code.

I explored an additional classification methods——SVM. For the training set, the SVM model is better, but for the testing set, it's worse than logistic regression method.

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

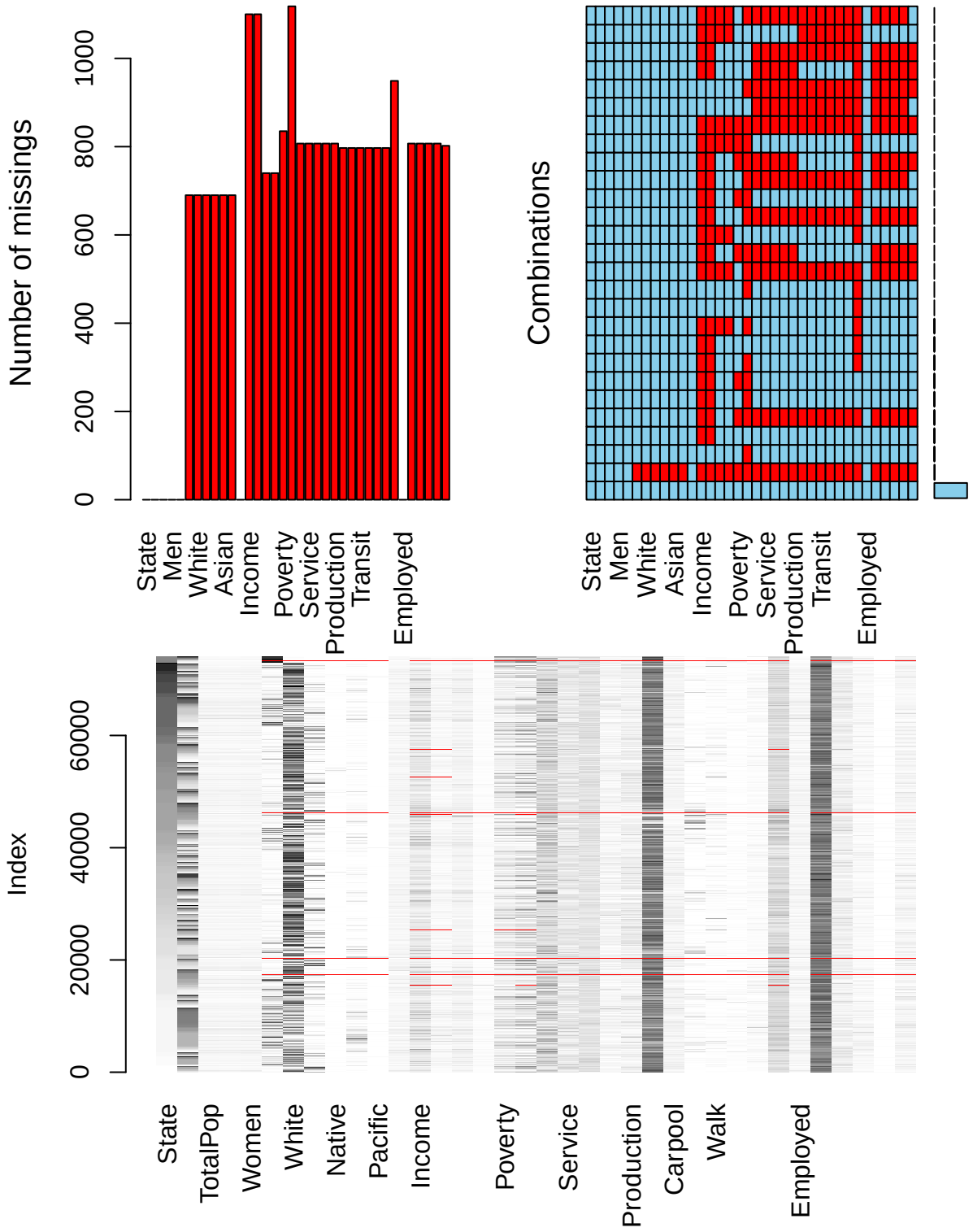



```
NULL

NULL

NULL

95% CI: 0.5-0.5 (DeLong)

95% CI: 0.5-0.5 (DeLong)
```

The AUC CI of decision tree is 0.7032-0.8355, the AUC CI of logistic regression is 0.7472-0.8742, the AUC CI of LASSO regression is 0.5-0.5, the AUC CI of KNN is 0.5-0.5, the AUC CI of SVM is 0.7135-0.8449. Taking the AUC value of ROC curve as the evaluation criterion, the best model is logistic regression, followed by SVM. In the face of various machine learning models, it is important to choose the model that corresponds to the application scenario. I will judge by the dimension size, quality and characteristic attributes of the data. But the effect is not satisfactory, eventually will be only one fittest algorithm to try.

There are a small number of missing values in the election-related data, as is often the case, and even inevitable. The reasons for missing values are various, including systematic reasons and artificial reasons.The system reason is the data loss caused by the failure of data collection or preservation due to the system reason, while the human reason is the data loss caused by the subjective error, historical limitation or deliberate concealment. Missing data can adversely affect model training and subsequent prediction. However, in view of the large amount of data in this dataset and the slight degree of data missing, the missing data was not interpolated, but directly deleted.

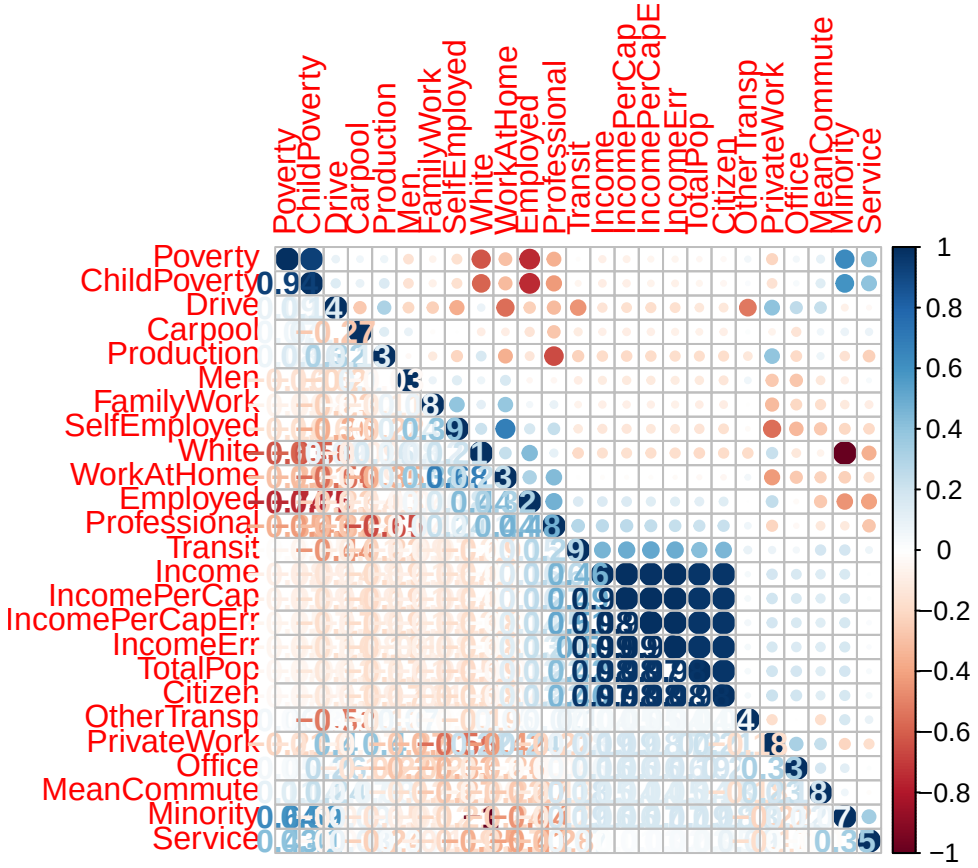
```
Warning in plot.aggr(res, ...): not enough vertical space to display frequencies
(too many combinations)
```



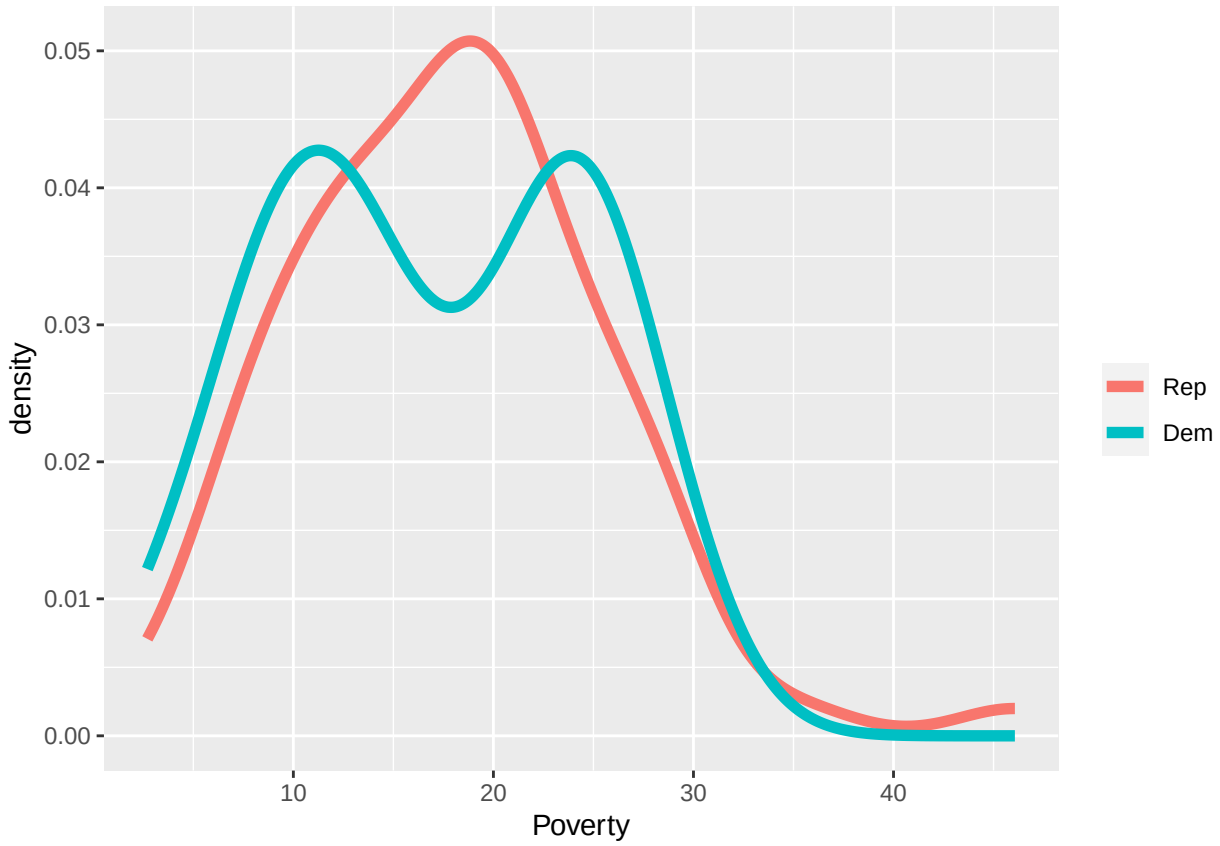
One of the above plots are also not showing on the output pdf, and I will submit the visualization aside from the pdf and rmd files.

Pearson's correlation coefficient is a measure of the degree of linear correlation, and its geometric meaning is the cosine of the Angle between the vectors formed by the concentrated

mean values of two variables. By observing the correlation coefficient between two independent variables to explore the internal correlation of self-varying data. It seems obvious that there is a strong correlation between income and the proportion of people living in poverty, but there is also a strong correlation between income and race and the proportion of people working, which seems reasonable. There is also a greater correlation between income and the proportion of urban population.



Explore the distribution of poverty between counties that chose Trump and those that did not. The main research object is county level attribute——"Poverty(% under poverty level)". According to whether the variable "candidate" in the data set "county_winner" is trump, the data is divided into two groups – Republican and Democrats(win and fail). Compare the distribution of the Poverty variable in the two groups of data. Look at it visually through a probability density plot, the two groups of data show a certain right-skewed pattern. Compared with the counties that chose Trump but did not choose Trump, the distribution of the proportion of poor population shows a more significant "Peak fat-tail" feature.


0.05
0.04
0.03
0.02
0.01
0.00
density
10
20
30
40
Poverty
Rep
Dem