PSTAT 131 - Final Project

Ruoxin Wang (Perm #9408246) & Tina Zhou (Perm #5726039)

December 08, 2022

Data

```
## # A tibble: 6 x 5
##   state    county     candidate     party total_votes
##   <chr>    <chr>      <fct>         <fct>       <dbl>
## 1 Delaware Kent       Joe Biden     DEM         44552
## 2 Delaware Kent       Donald Trump  REP         41009
## 3 Delaware Kent       Jo Jorgensen  LIB          1044
## 4 Delaware Kent       Howie Hawkins GRN           420
## 5 Delaware New Castle Joe Biden     DEM        195034
## 6 Delaware New Castle Donald Trump  REP         88364
```

```
## # A tibble: 6 x 37
##   CountyId State   County  Total~1   Men  Women Hispa~2 White Black Native Asian
##      <dbl> <chr>   <chr>     <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1     1001 Alabama Autaug~   55036 26899  28137     2.7  75.4  18.9    0.3   0.9
## 2     1003 Alabama Baldwi~  203360 99527 103833     4.4  83.1   9.5    0.8   0.7
## 3     1005 Alabama Barbou~   26201 13976  12225     4.2  45.7  47.8    0.2   0.6
## 4     1007 Alabama Bibb C~   22580 12251  10329     2.4  74.6  22      0.4   0
## 5     1009 Alabama Blount~   57667 28490  29177     9    87.4   1.5    0.3   0.1
## 6     1011 Alabama Bulloc~   10478  5616   4862     0.3  21.6  75.6    1     0.7
## # ... with 26 more variables: Pacific <dbl>, VotingAgeCitizen <dbl>,
## #   Income <dbl>, IncomeErr <dbl>, IncomePerCap <dbl>, IncomePerCapErr <dbl>,
## #   Poverty <dbl>, ChildPoverty <dbl>, Professional <dbl>, Service <dbl>,
## #   Office <dbl>, Construction <dbl>, Production <dbl>, Drive <dbl>,
## #   Carpool <dbl>, Transit <dbl>, Walk <dbl>, OtherTransp <dbl>,
## #   WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>,
## #   PublicWork <dbl>, SelfEmployed <dbl>, FamilyWork <dbl>, ...
```

Election data

Question 1:

Dimension of `election.raw`

```
## [1] 32177     5
```

Number of missing value in `election.raw`

```
## [1] 0
```

Number of distinct state in `election.raw`

```
## [1] 51
```

The dimension of `election.raw` is that it contains 32177 rows and 5 columns (variables). And there is no missing values in this data set. After compute the number of distinct values in `state`, there are in total 51 different values, which verifies the data set contains all states and a federal district.

Census data

Question 2:

Dimensions of `census`

```
## [1] 3220   37
```

Number of missing value in `census`

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

Number of distinct county in `census`

```
## [1] 2006
```

Compare with the number of distinct county in `election.raw`

```
##   census_county election_county
## 1          2006            2856
```

The dimension of `census` is that it contains 3220 rows and 37 columns (variables). There is 1 missing value in the data set. Since there are States like Maryland, Michigan, and Texas that all have a county with the same name "Kent", we calculate the number of distinct county by pairing the State and the County together to count the final number. The total number of distinct values in `county` in `census` is 2006. Comparing to the total number of distinct county in `election.raw` of 2856, it is easy to see that the number of county that participate in the election is more than the number of county that participate in the census.

Data wrangling

Question 3: Construct aggregated data sets from `election.raw`

Create a state-level summary `election.state`

```
## # A tibble: 6 x 4
## # Groups:   state, candidate [6]
##   state   candidate      party state_total_votes
##   <chr>   <fct>          <fct>             <dbl>
## 1 Alabama Donald Trump   REP             1441168
## 2 Alabama Jo Jorgensen   LIB               25176
```

```
## 3 Alabama Joe Biden       DEM              849648
## 4 Alabama Write-ins       WRI                7312
## 5 Alaska  Brock Pierce    IND                 825
## 6 Alaska  Don Blankenship CST                1127
```

Create a federal-level summary into a `election.total`

```
## # A tibble: 6 x 3
## # Groups:   candidate [6]
##   candidate      party federal_total_votes
##   <fct>          <fct>               <dbl>
## 1 Alyson Kennedy SWP                  6791
## 2 Bill Hammons   UTY                  6647
## 3 Blake Huber    APV                   409
## 4 Brian Carroll  ASP                 25256
## 5 Brock Pierce   IND                 49552
## 6 Brooke Paige   GOP                  1175
```

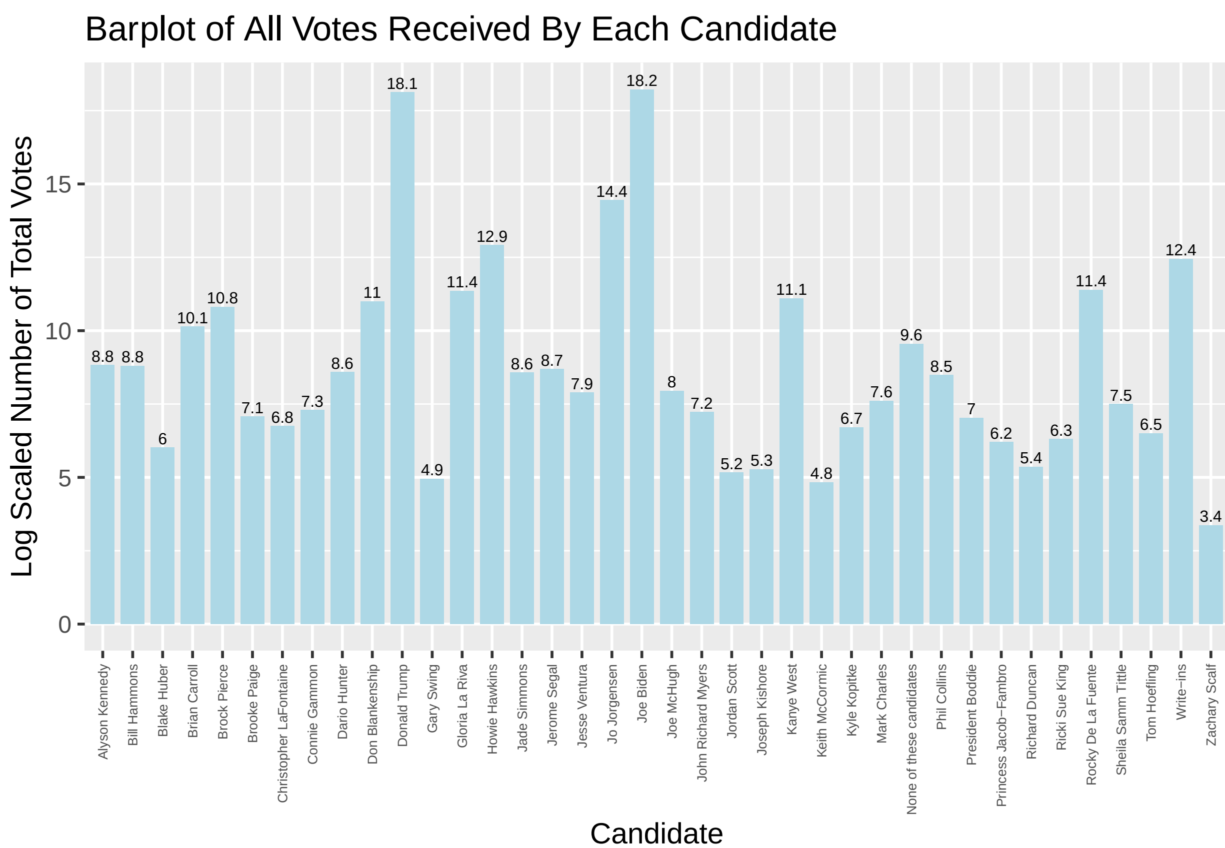
Question 4:

Number of named presidential candidates in the 2020 election

```
## [1] 38
```

```
##  [1] Joe Biden               Donald Trump            Jo Jorgensen
##  [4] Howie Hawkins           Write-ins               Gloria La Riva
##  [7] Brock Pierce            Rocky De La Fuente      Don Blankenship
## [10] Kanye West              Brian Carroll           Ricki Sue King
## [13] Jade Simmons            President Boddie        Bill Hammons
## [16] Tom Hoefling            Alyson Kennedy          Jerome Segal
## [19] Phil Collins            None of these candidates Sheila Samm Tittle
## [22] Dario Hunter            Joe McHugh              Christopher LaFontaine
## [25] Keith McCormic          Brooke Paige            Gary Swing
## [28] Richard Duncan          Blake Huber             Kyle Kopitke
## [31] Zachary Scalf           Jesse Ventura           Connie Gammon
## [34] John Richard Myers      Mark Charles            Princess Jacob-Fambro
## [37] Joseph Kishore          Jordan Scott
## 38 Levels: Alyson Kennedy Bill Hammons Blake Huber ... Zachary Scalf
```

Barplot of all votes received by each candidate



There were 38 distinct value candidate column. However, since there is a value called "None of these candidates", there should be **37** named presidential candidates in total in the 2020 election. And the log scaled bar chart of all votes received by each candidate is shown above.

Question 5:

Create data set `county.winner`

```
## # A tibble: 6 x 7
## # Groups:   county [6]
##   county         state          candidate    party total_votes total   pct
##   <chr>          <chr>          <fct>        <fct>       <dbl> <dbl> <dbl>
## 1 Abbeville      South Carolina Donald Trump REP          8215 12433 0.661
## 2 Abbot          Maine          Donald Trump REP           288   417 0.691
## 3 Abington       Massachusetts  Joe Biden    DEM          5209  9660 0.539
## 4 Acadia Parish  Louisiana      Donald Trump REP         22596 28425 0.795
## 5 Accomack       Virginia       Donald Trump REP          9172 16962 0.541
## 6 Acton          Massachusetts  Joe Biden    DEM         11105 15563 0.714
```

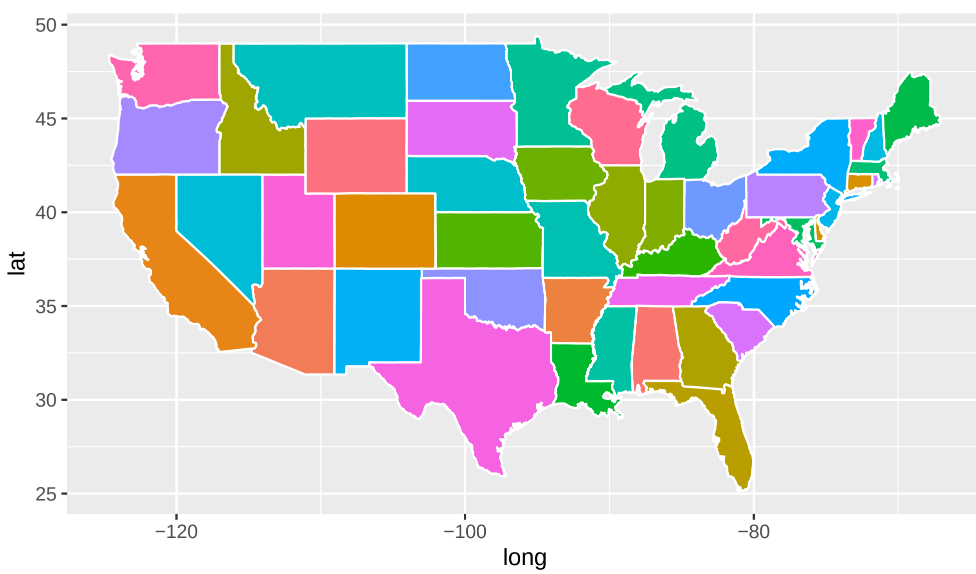
Create data set `state.winner`

```
## # A tibble: 6 x 6
```

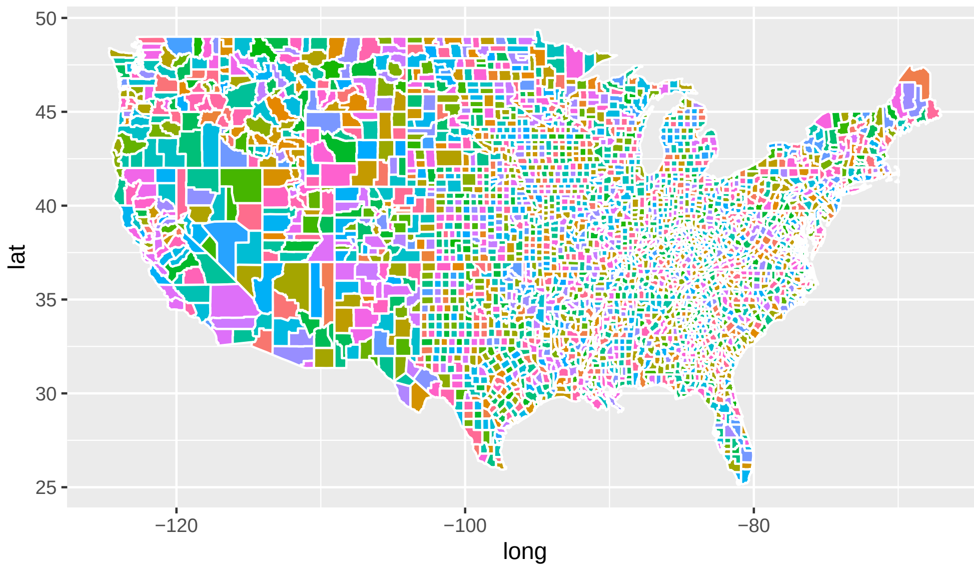
```
## # Groups:   state [6]
##   state      candidate    party state_total_votes    total   pct
##   <chr>      <fct>        <fct>             <dbl>    <dbl> <dbl>
## 1 Alabama    Donald Trump REP             1441168  2323304 0.620
## 2 Alaska     Donald Trump REP              189892   391346 0.485
## 3 Arizona    Joe Biden    DEM             1672143  3387326 0.494
## 4 Arkansas   Donald Trump REP              760647  1219069 0.624
## 5 California Joe Biden    DEM            11109764 17495906 0.635
## 6 Colorado   Joe Biden    DEM             1804352  3256953 0.554
```

Visualization

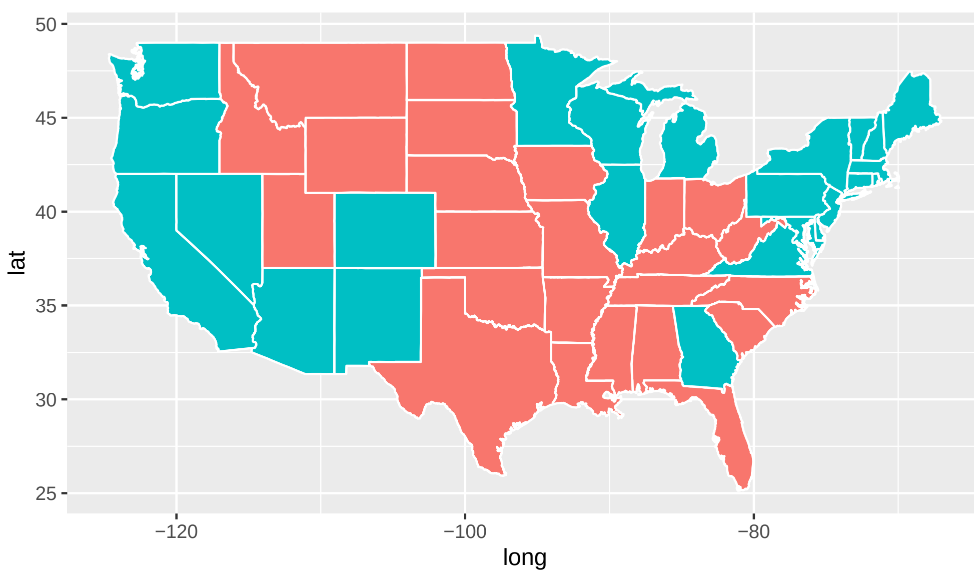
Example



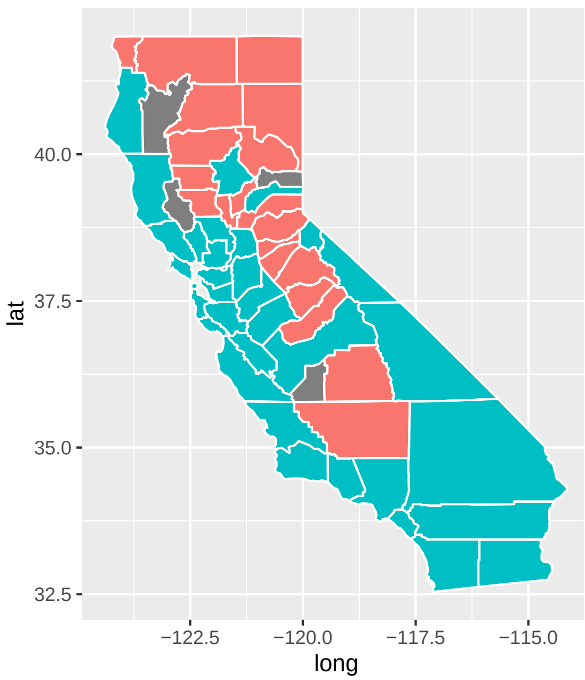
Question 6: Draw county-level map



Question 7: Color the map by the winning candidate for each state.

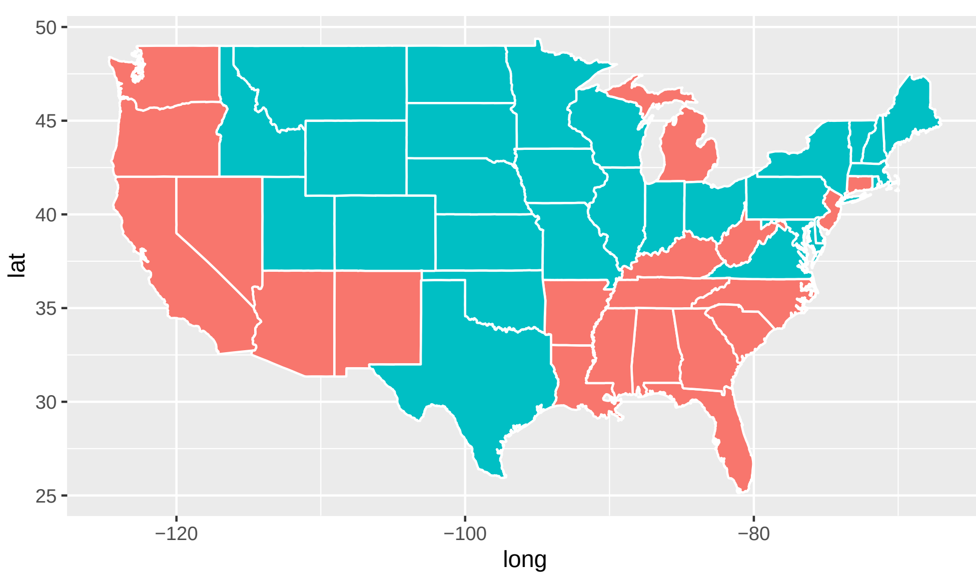


Question 8: Color the map of the state of California by the winning candidate for each county.



Question 9: Create a visualization

Average unemployment level of each state



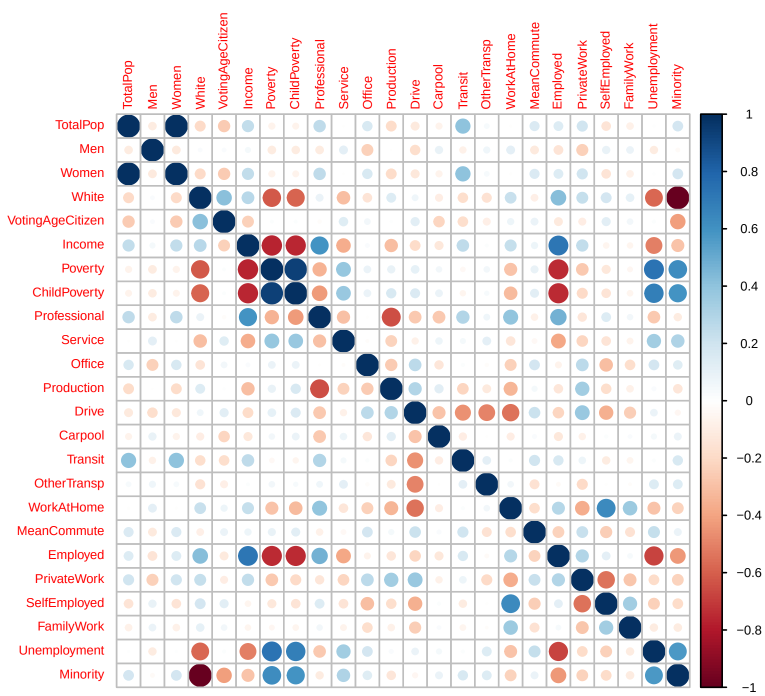
The map above visualizes the average unemployment level of each state, The red/orange color represents the states which have an unemployment rate higher than the mean unemployment rate while the blue/green color represents the states which have an unemployment rate lower than the mean unemployment rate. According to the visualization above, we can see that the states with high unemployment rate concentrate in the western and southeastern parts of the United States.

Question 10:

Clean county-level census data census.clean

```
## # A tibble: 6 x 27
##   Count~1 State County Total~2   Men  Women White Votin~3 Income Poverty Child~4
##     <dbl> <chr> <chr>    <dbl> <dbl>  <dbl> <dbl>   <dbl>  <dbl>   <dbl>   <dbl>
## 1    1001 Alab~ Autau~   55036  48.9  28137  75.4    74.5  55317    13.7    20.1
## 2    1003 Alab~ Baldw~  203360  48.9 103833  83.1    76.4  52562    11.8    16.1
## 3    1005 Alab~ Barbo~   26201  53.3  12225  45.7    77.4  33368    27.2    44.9
## 4    1007 Alab~ Bibb ~   22580  54.3  10329  74.6    78.2  43404    15.2    26.6
## 5    1009 Alab~ Bloun~   57667  49.4  29177  87.4    73.7  47412    15.6    25.4
## 6    1011 Alab~ Bullo~   10478  53.6   4862  21.6    78.4  29655    28.5    50.4
## # ... with 16 more variables: Professional <dbl>, Service <dbl>, Office <dbl>,
## #   Production <dbl>, Drive <dbl>, Carpool <dbl>, Transit <dbl>,
## #   OtherTransp <dbl>, WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>,
## #   PrivateWork <dbl>, SelfEmployed <dbl>, FamilyWork <dbl>,
## #   Unemployment <dbl>, Minority <dbl>, and abbreviated variable names
## #   1: CountyId, 2: TotalPop, 3: VotingAgeCitizen, 4: ChildPoverty
```

Identify perfect collinearity



Print first 5 rows of `census.clean`

```
## # A tibble: 5 x 27
##   Count~1 State County Total~2   Men  Women White Votin~3 Income Poverty Child~4
##     <dbl> <chr> <chr>    <dbl> <dbl>  <dbl> <dbl>   <dbl>  <dbl>   <dbl>   <dbl>
## 1    1001 Alab~ Autau~   55036  48.9  28137  75.4    74.5  55317    13.7    20.1
## 2    1003 Alab~ Baldw~  203360  48.9 103833  83.1    76.4  52562    11.8    16.1
## 3    1005 Alab~ Barbo~   26201  53.3  12225  45.7    77.4  33368    27.2    44.9
## 4    1007 Alab~ Bibb ~   22580  54.3  10329  74.6    78.2  43404    15.2    26.6
## 5    1009 Alab~ Bloun~   57667  49.4  29177  87.4    73.7  47412    15.6    25.4
## # ... with 16 more variables: Professional <dbl>, Service <dbl>, Office <dbl>,
## #   Production <dbl>, Drive <dbl>, Carpool <dbl>, Transit <dbl>,
## #   OtherTransp <dbl>, WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>,
## #   PrivateWork <dbl>, SelfEmployed <dbl>, FamilyWork <dbl>,
## #   Unemployment <dbl>, Minority <dbl>, and abbreviated variable names
## #   1: CountyId, 2: TotalPop, 3: VotingAgeCitizen, 4: ChildPoverty
```

According to the above graph, no other features are prefectly colinear. As a result, we do not need to further drop columns.

Dimensionality reduction

Question 11:

Run PCA for the cleaned county level census data

```
## Importance of components:
##                          PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8
## Standard deviation     2.380 1.843 1.796 1.3256 1.1500 1.1364 1.0449 1.0146
## Proportion of Variance 0.227 0.136 0.129 0.0703 0.0529 0.0517 0.0437 0.0412
## Cumulative Proportion  0.227 0.362 0.491 0.5617 0.6146 0.6663 0.7099 0.7511
##                           PC9   PC10   PC11   PC12   PC13   PC14   PC15   PC16
## Standard deviation     0.9571 0.9283 0.8760 0.8573 0.7572 0.7322 0.6416 0.6009
## Proportion of Variance 0.0366 0.0345 0.0307 0.0294 0.0229 0.0214 0.0165 0.0144
## Cumulative Proportion  0.7877 0.8222 0.8529 0.8823 0.9052 0.9267 0.9432 0.9576
##                          PC17    PC18    PC19    PC20    PC21    PC22    PC23
## Standard deviation     0.5617 0.46710 0.43222 0.35911 0.30731 0.26574 0.20706
## Proportion of Variance 0.0126 0.00873 0.00747 0.00516 0.00378 0.00282 0.00172
## Cumulative Proportion  0.9702 0.97895 0.98643 0.99158 0.99536 0.99819 0.99990
##                          PC24    PC25
## Standard deviation     0.0488 0.00899
## Proportion of Variance 0.0001 0.00000
## Cumulative Proportion  1.0000 1.00000
```

Save the first two principle components PC1 and PC2 into a new data frame

```
##       PC1     PC2
## 1 -0.4041  0.1640
## 2 -1.1228  0.6692
## 3  3.7767 -0.6798
## 4  1.2353 -1.2664
```

```
## 5  0.1880 -0.5048
## 6  4.4953 -0.4130
```

Whether scale and center the features

We chose to both `scale` and `center` the features by using the options `scale = TRUE` and `center = TRUE` before running PCA. The reason to use `scale` function is that it will help to scale all features to have a unit variance of 1 and to be on the same scale. The reason to use `center` function is that it will help to shift all features to be zero centered, which has a mean of 0. These two functions together help to normalize the features and make sure all features are running on the same scale for better running PCA.

Three features with the largest absolute values of the first principal component

```
##                  PC1       PC2      PC3      PC4      PC5      PC6    PC7
## Poverty       0.3815  0.008562  0.09056  0.08825 -0.05767 -0.08112 0.1355
## ChildPoverty  0.3802 -0.006989  0.05307  0.04741 -0.09387 -0.09200 0.1268
## Employed     -0.3513  0.091561 -0.05722 -0.05556  0.15840 -0.02982 0.1377
##                  PC8      PC9     PC10    PC11     PC12    PC13     PC14
## Poverty      0.04697 -0.06531 0.017546 0.06285 -0.01368 0.04689  0.16124
## ChildPoverty 0.04102 -0.04689 0.008497 0.06685 -0.02030 0.08548  0.10772
## Employed     0.01603 -0.02603 0.226957 0.02151 -0.06666 0.01659 -0.08894
##                 PC15    PC16     PC17    PC18   PC19     PC20     PC21    PC22
## Poverty      0.13703  0.1309 -0.21327 0.05794 0.3444 -0.01968 -0.04544 -0.2771
## ChildPoverty 0.13024  0.1264 -0.32338 0.08139 0.3988  0.30817  0.16498  0.3142
## Employed     0.07962 -0.3504 -0.02717 0.18647 0.5870 -0.45296 -0.07270  0.1771
##                  PC23      PC24       PC25
## Poverty      -0.69884 -0.007938  0.0047554
## ChildPoverty  0.51483  0.003440 -0.0022040
## Employed      0.01536 -0.004566 -0.0002702
```

The three features with the largest absolute values of the first principal component (PC1) are:
- `Poverty` with absolute value of **0.3815**
- `ChildPoverty` with absolute value of **0.3802**
- `Employed` with absolute value of **0.3513**

Features have opposite signs in PC1

```
##  [1] "TotalPop"          "Men"               "Women"             "White"
##  [5] "VotingAgeCitizen" "Income"            "Professional"      "Transit"
##  [9] "WorkAtHome"        "Employed"          "PrivateWork"       "SelfEmployed"
## [13] "FamilyWork"
```

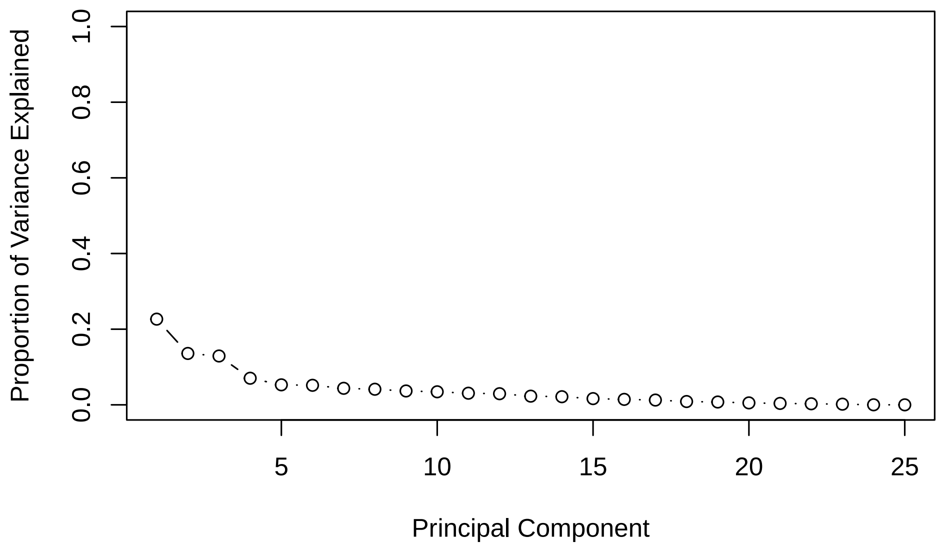
Features have opposite signs in PC2

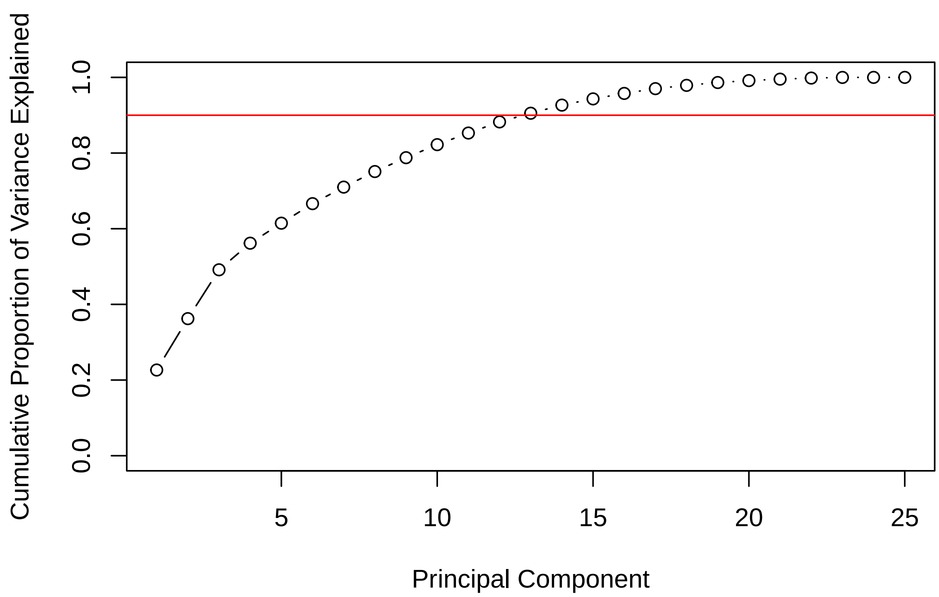
```
##  [1] "Men"               "White"             "VotingAgeCitizen" "ChildPoverty"
##  [5] "Production"        "Drive"             "Carpool"           "WorkAtHome"
##  [9] "SelfEmployed"      "FamilyWork"
```

The negative sign in the loadings of PCA could interpret that these features are having a negative correlation with each other. This means that when one of those features are increasing, the other features would decrease based on this negative correlation.

Question 12:

Plot proportion of variance explained (PVE) and cumulative PVE





Minimum number of PCs need to capture 90% of the variance for the analysis

```
## [1] "First 12 PCs explain 0.882315455428802 of the variance"
```

```
## [1] "First 13 PCs explain 0.905249026224439 of the variance"
```

Since the first 12 PCs explain less than 90% of the total variation and the first 13 PCs explain a little bit more than 90% of the total variation, it means that the minimum number of PCs that needed to explain 90% of the total variation for this analysis is **13**.

Clustering

Question 13:

Perform hierarchical clustering with complete linkage with `census.clean`

```
##
## Call:
## hclust(d = census.dist)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 3219
```

Cut the tree into 10 clusters

```
## newclus
##    1    2    3    4    5    6    7    8    9   10
## 3111   69    2    9   12    1    2    5    7    1
```

Re-run the hierarchical clustering algorithm with PC1 and PC2

```
##
## Call:
## hclust(d = pc.dist)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 3219
```

```
## pc.newclus
##    1    2    3    4    5    6    7    8    9   10
## 1286 1259  374  101  109   52    9   23    1    5
```

By comparing the result of both approach after cutting tree into 10 clusters, it is easy to see that one cluster is much more dense than all the other clusters when using `census.clean` as the input for the hierarchical clustering. However, when using `pc.county` as the input for the hierarchical clustering, the number of observations in each cluster is reasonably reduce in a certain pattern instead of suddenly reduce to a relatively small number.

Investigate the cluster that contains Santa Barbara County

```
## [1] "For census.clean, Santa Barbara County is in cluster: 1"
```

```
## [1] "For pc.county, Santa Barbara County is in cluster: 3"
```

```
## # A tibble: 3,111 x 28
##    CountyId State   County     Total~1   Men  Women White Votin~2 Income Poverty
##       <dbl> <chr>   <chr>        <dbl> <dbl>  <dbl> <dbl>   <dbl>  <dbl>   <dbl>
##  1     1001 Alabama Autauga C~   55036  48.9  28137  75.4    74.5  55317    13.7
##  2     1003 Alabama Baldwin C~  203360  48.9 103833  83.1    76.4  52562    11.8
##  3     1005 Alabama Barbour C~   26201  53.3  12225  45.7    77.4  33368    27.2
##  4     1007 Alabama Bibb Coun~   22580  54.3  10329  74.6    78.2  43404    15.2
##  5     1009 Alabama Blount Co~   57667  49.4  29177  87.4    73.7  47412    15.6
##  6     1011 Alabama Bullock C~   10478  53.6   4862  21.6    78.4  29655    28.5
##  7     1013 Alabama Butler Co~   20126  46.8  10710  52.2    76.8  36326    24.4
##  8     1015 Alabama Calhoun C~  115527  48.1  59934  72.7    76.5  43686    18.6
##  9     1017 Alabama Chambers ~   33895  48.1  17575  56.2    77.5  37342    18.8
## 10     1019 Alabama Cherokee ~   25855  49.7  12993  91.8    79.8  40041    16.1
## # ... with 3,101 more rows, 18 more variables: ChildPoverty <dbl>,
## #   Professional <dbl>, Service <dbl>, Office <dbl>, Production <dbl>,
## #   Drive <dbl>, Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>,
## #   WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>,
## #   SelfEmployed <dbl>, FamilyWork <dbl>, Unemployment <dbl>, Minority <dbl>,
## #   Cluster <int>, and abbreviated variable names 1: TotalPop,
## #   2: VotingAgeCitizen
```

```
## # A tibble: 374 x 28
##    CountyId State   County     Total~1   Men  Women White Votin~2 Income Poverty
##       <dbl> <chr>   <chr>        <dbl> <dbl>  <dbl> <dbl>   <dbl>  <dbl>   <dbl>
##  1     1015 Alabama Calhoun C~  115527  48.1  59934  72.7    76.5  43686    18.6
##  2     1069 Alabama Houston C~  104108  47.9  54203  67.2    75.4  42803    18.5
##  3     1073 Alabama Jefferson~  659460  47.4 347127  50.4    74.5  49321    17.6
##  4     1097 Alabama Mobile Co~  414328  47.8 216360  57.5    74.8  45802    19.3
##  5     1101 Alabama Montgomer~  227120  47.3 119650  35.2    73.7  46545    20.8
##  6     1113 Alabama Russell C~   58480  48.6  30056  47.8    73.4  38988    20.9
##  7     1125 Alabama Tuscaloos~  204424  48.3 105699  62.4    76.5  50513    17.3
##  8     2070 Alaska  Dillingha~    4974  52.0   2389  16.2    68.2  58708    16.6
##  9     2164 Alaska  Lake and ~    1301  50.8    640  21.8    70.0  45208    16.5
## 10     2198 Alaska  Prince of~    6473  54.2   2964  45.3    75.3  52114    16
## # ... with 364 more rows, 18 more variables: ChildPoverty <dbl>,
## #   Professional <dbl>, Service <dbl>, Office <dbl>, Production <dbl>,
## #   Drive <dbl>, Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>,
## #   WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>,
## #   SelfEmployed <dbl>, FamilyWork <dbl>, Unemployment <dbl>, Minority <dbl>,
## #   Cluster <int>, and abbreviated variable names 1: TotalPop,
## #   2: VotingAgeCitizen
```

For `census.clean` Santa Barbara County is in cluster 1. For `pc.county`, Santa Barbara County is in cluster 3. When taking back the cluster to the original dataset, it is easy to see that the cluster 3 we get from `pc.county` would be more appropriate than the cluster 1 we get from `census.clean`. Since the rule of clustering is trying to put observations in the groups where other observations in the same group have relatively similar patterns. However, for the cluster 1 we get from `census.clean`, the observations vary too much and there is no general similarity among them. For example, a lot of Alabama counties are included in this cluster. However, for the cluster 3 we get from `pc.county`, there is more similar patterns among the observations and less counties from Alabama are included, which is identical with what we want to see in the clustering. As a result, hierarchical clustering on `pc.county` may be more appropriate. The possible reason for this is that after running PCA, a low-dimensional representation of the dataset has been found and create a more representative pattern for the data to be better clustered.

Classification

Question 14:

The code above first changes the format of state and county in both `county.winner` and `census.clean` into the same ones. Then, it uses the `left_join()` to combine the two dataset together and drops the rows with missing values at the same time.

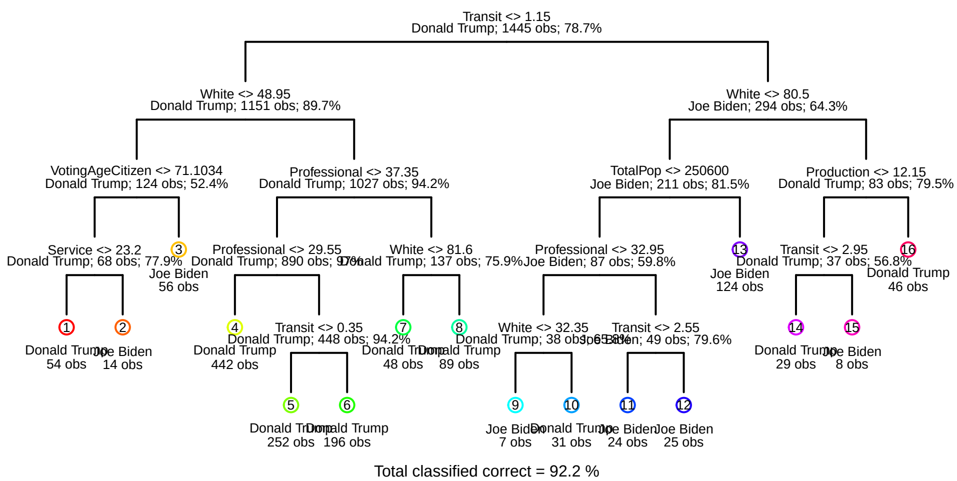
The reason to drop `party` is that after we finished the left join, it is easy to find out that there are only two unique value for `candidate` in `election.cl`: Joe Biden and Donald Trump. Since they are from two different party, this means that there is a colinearity between `candidate` and `party` and `party` is basically the same as what represent by `candidate`, which means that it cannot be a predictor for `candidate`. As a result, we need to exclude the predictor `party` from `election.cl`.

Classification

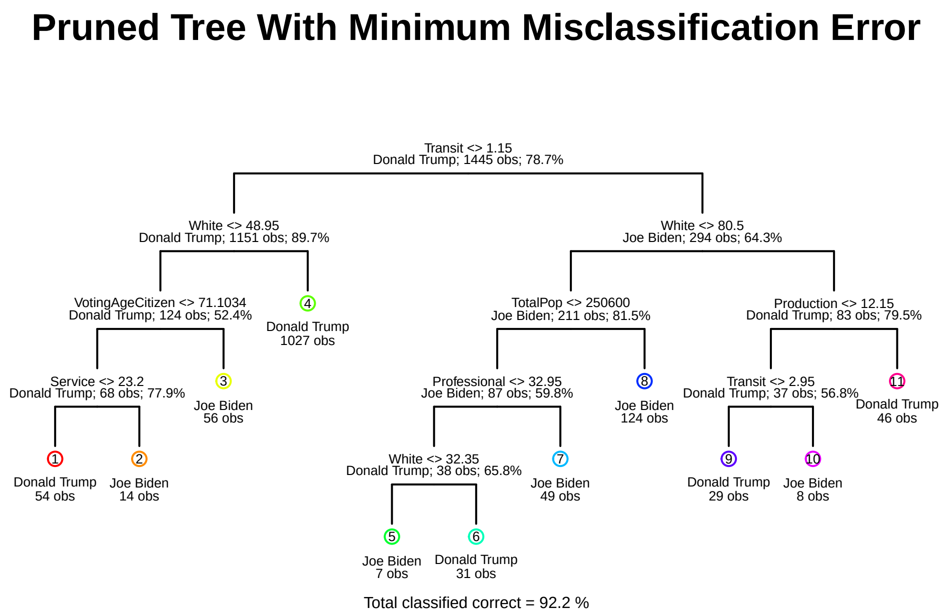
Question 15:

Train a decision tree and visualize

Visualization of Decision Tree Before Pruning



Prune the tree



Save training and test errors to records object

```
##          train.error test.error
## tree         0.07751    0.08287
## logistic          NA         NA
## lasso             NA         NA
```

Interpretation

Since we calculate previously the best size of leaf nodes are 11, this decision tree has 11 leaf nodes. The training error rate is 0.07751 and the test error rate is 0.08287. The two error rate is similar, which means that there is no overfitting for this decision tree model. Also, since the total classified correct equals to 92.2%, this means that the model works well on this election data set. Some significant variable used in this decision tree are: `Transit`, `White`, `VotingAgeCitizen`, `TotalPop`, `Production`, `Service`, and `Professional`. By comparing the total support of two candidate, Donald Trump is having more support than Joe Biden based on this decision tree.

This plots tells a story about the voting behavior of White people in the US. If the majority of the county is White people and they are get a lower Transit percentage, Donald Trump are more possible to be the winner. On the other hand, if the majority of the county is not White people and they are get higher percentage in Service, Transit, and Professional, Joe Biden are more likely to be the winner.

The decision tree start the first split with the `Transit` feature, which decided on whether the percentage is larger or smaller than 1.15. Then, the second split on both sides are performed with `White` feature, which stands for percentage of the White people in the total population. One of the second split is decided on whether the percentage of `White` is larger or smaller than 48.95, the other second split is decided on whether the percentage of `White` is larger or smaller than 80.5. After this split, one leaf node has already been split out from the tree which is leaf 4 which contains 1027 observations who support Donald Trump. Then, one of the third split is performed on `VotingAgeCitizen`, which decided on whether the percentage is larger or smaller than 71.1034. This splitted out leaf 3 which contains 56 observations who support Joe Biden. Another third split is performed on `TotalPop` feature and it is decided on whether the number of total population is larger or smaller than 250600, which split out leaf 8 that contains 124 oberservations who support Joe Biden. The final third split is performed on `Production` feature and it is decided on whether the percentage is larger or smaller than 12.15. After this split, leaf 11 is splitted out which contains 48 observations who support Donald Trump. Then, three fourth split continue performed. One of them is performed with `Service` features and it decided on whether the percentage is smaller or larger than 23.2, which gives out 2 final leaf nodes: leaf 1 which contains 54 observations who support Donald Trump and leaf 2 which contains 14 observations who support Joe Biden. Another fourth split is performed with `Professional` after the previous split with `TotalPop`, which decided on whether the percentage is larger or smaller than 32.95. This split give one final leaf 7 that contains 49 observations who support Joe Biden. The final fourth split is performed with `Transit` feature after the previous split with `Production`, which decided on whether the percentage is greater or smaller than 2.95. This gives out two final leaf: leaf 9 that contains 29 observations who support Donald Trump and leaf 10 that contains 8 observations who support Joe Biden. The fifth (final) split is performed with `White` feature agian, which decided on whether the percentage is greater or smaller than 32.35. It gives out two final leaf nodes: leaf 5 that contains 7 observations who support Joe Biden and leaf 6 that contains 31 observations who support Donald Trump.

Question 16:

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

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

```
##
## Call:
## glm(formula = candidate ~ ., family = "binomial", data = election.tr)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.773  -0.266  -0.099  -0.019   3.229
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.08e+01   1.10e+01   -1.89  0.05919 .
## TotalPop         -4.98e-07   3.48e-05   -0.01  0.98859
## Men              -5.09e-02   6.06e-02   -0.84  0.40060
## Women             4.51e-06   6.86e-05    0.07  0.94755
## White            -1.74e-01   8.49e-02   -2.05  0.04024 *
## VotingAgeCitizen  2.10e-01   3.57e-02    5.87  4.3e-09 ***
## Income           -1.14e-05   2.04e-05   -0.56  0.57447
## Poverty           1.44e-02   5.50e-02    0.26  0.79349
## ChildPoverty      4.06e-03   3.25e-02    0.13  0.90048
## Professional      2.86e-01   5.10e-02    5.62  2.0e-08 ***
## Service           2.90e-01   6.04e-02    4.81  1.5e-06 ***
## Office            1.81e-01   6.40e-02    2.82  0.00479 **
```

```
## Production        1.80e-01   5.16e-02    3.49  0.00048 ***
## Drive            -1.84e-01   4.61e-02   -3.98  6.8e-05 ***
## Carpool          -1.95e-01   6.42e-02   -3.03  0.00246 **
## Transit           3.07e-01   1.23e-01    2.49  0.01280 *
## OtherTransp      -1.66e-02   1.19e-01   -0.14  0.88869
## WorkAtHome       -6.46e-02   7.52e-02   -0.86  0.39065
## MeanCommute       1.34e-02   3.13e-02    0.43  0.66872
## Employed          2.27e-01   4.05e-02    5.61  2.0e-08 ***
## PrivateWork       3.97e-02   2.73e-02    1.45  0.14692
## SelfEmployed      8.21e-05   5.80e-02    0.00  0.99887
## FamilyWork       -1.73e+00   6.93e-01   -2.50  0.01244 *
## Unemployment      1.90e-01   5.37e-02    3.54  0.00040 ***
## Minority         -4.12e-02   8.28e-02   -0.50  0.61863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1497.31  on 1444  degrees of freedom
## Residual deviance:  518.04  on 1420  degrees of freedom
## AIC: 568
##
## Number of Fisher Scoring iterations: 7
```

Save training and test errors to records

```
##          train.error test.error
## tree         0.07751    0.08287
## logistic     0.06782    0.09669
## lasso             NA         NA
```

If we take $\alpha = 0.05$ as the threshold here for the logistic regression, then if the p-value of the predictor is less than 0.05, the predictor is statistically significant. Based on this idea, the significant variables are `White`,`VotingAgeCitizen`, `Professional`, `Service`, `Office`, `Production`, `Drive`, `Carpool`, `Transit`, `Employed`, `FamilyWork`, and `Unemployment`.

All the variables that existed in the Decision Tree are included in the set of significant variables that we get in the logistic function. However, `Office`,`Drive`, `Carpool`, `Employed`, `FamilyWork`, and `Unemployment` do not exist in the decision tree. And the interpretation of significant coefficients are:

- The variable `White` has a coefficient of -1.74e-01. For every one unit change in `White`, the log odds of `candidate` winning decreases by 1.74e-01, holding other variables fixed.

- The variable `VotingAgeCitizen` has a coefficient 2.10e-01. For a one unit increase in `VotingAgeCitizen`, the log odds of `candidate` winning increases by 2.10e-01, holding other variables fixed.

- The variable `Professional` has a coefficient 2.86e-01. For a one unit increase in `Professional`, the log odds of `candidate` winning increases by 2.86e-01, holding other variables fixed.

- The variable `Service` has a coefficient 2.90e-01. For a one unit increase in `Service`, the log odds of `candidate` winning increases by 2.90e-01, holding other variables fixed.

- The variable `Office` has a coefficient 1.81e-01. For a one unit increase in `Office`, the log odds of `candidate` winning increases by 1.81e-01, holding other variables fixed.

- The variable `Production` has a coefficient 1.80e-01. For a one unit increase in `Production`, the log odds of `candidate` winning increases by 1.80e-01, holding other variables fixed.

- The variable `Drive` has a coefficient -1.84e-01. For a one unit increase in `Drive`, the log odds of `candidate` winning decreases by 1.840e-01, holding other variables fixed.

- The variable `Carpool` has a coefficient -1.95e-01. For a one unit increase in `Carpool`, the log odds of `candidate` winning decreases by 1.95e-01, holding other variables fixed.

- The variable `Transit` has a coefficient 3.070e-01. For a one unit increase in `Transit`, the log odds of `candidate` winning increases by 3.07e-01, holding other variables fixed.

- The variable `Employed` has a coefficient 2.27e-01. For a one unit increase in `Employed`, the log odds of `candidate` winning increases by 2.27e-01, holding other variables fixed.

- The variable `FamilyWork` has a coefficient -1.73e-00. For a one unit increase in `FamilyWork`, the log odds of `candidate` winning decreases by 1.73e-00, holding other variables fixed.

- The variable `Unemployment` has a coefficient 1.90e-01. For a one unit increase in `Unemployment`, the log odds of `candidate` winning increases by 1.90e-01, holding other variables fixed.

Question 17:

Run LASSO penalty

```
##
## Call:  cv.glmnet(x = x.tr, y = y.tr, lambda = seq(1, 50) * 1e-04, nfolds = 10,      alpha = 1, famil
##
## Measure: Binomial Deviance
##
##     Lambda Index Measure     SE Nonzero
## min  0.002    31   0.394 0.0208      19
## 1se  0.005     1   0.408 0.0174      16
```

Optimal value of λ in cross validation

```
## [1] 0.002
```

The optimal value of λ in cross validation is 0.002.

Non-zero coefficients in the LASSO regression for optimal λ

```
## 25 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)      -1.899e+01
## TotalPop          .
## Men              -5.709e-02
## Women             3.670e-06
## White            -1.135e-01
## VotingAgeCitizen  1.852e-01
## Income            .
## Poverty           2.480e-02
## ChildPoverty      .
## Professional      2.019e-01
## Service           2.019e-01
## Office            9.974e-02
## Production        9.307e-02
## Drive            -1.286e-01
## Carpool          -1.313e-01
## Transit           2.541e-01
## OtherTransp       2.353e-03
## WorkAtHome       -6.042e-03
## MeanCommute       .
## Employed          1.835e-01
## PrivateWork       2.936e-02
## SelfEmployed     -4.230e-02
## FamilyWork       -1.151e+00
## Unemployment      1.625e-01
## Minority          .
```

```
##      (Intercept)         TotalPop              Men            Women
##       -2.082e+01       -4.977e-07       -5.094e-02        4.514e-06
##            White VotingAgeCitizen           Income          Poverty
##       -1.742e-01        2.098e-01       -1.144e-05        1.439e-02
##     ChildPoverty     Professional          Service           Office
##        4.059e-03        2.862e-01        2.904e-01        1.805e-01
##       Production            Drive          Carpool          Transit
##        1.800e-01       -1.838e-01       -1.945e-01        3.065e-01
##      OtherTransp       WorkAtHome      MeanCommute         Employed
##       -1.659e-02       -6.460e-02        1.340e-02        2.272e-01
##      PrivateWork     SelfEmployed       FamilyWork     Unemployment
##        3.966e-02        8.209e-05       -1.733e+00        1.901e-01
##         Minority
##       -4.121e-02
```

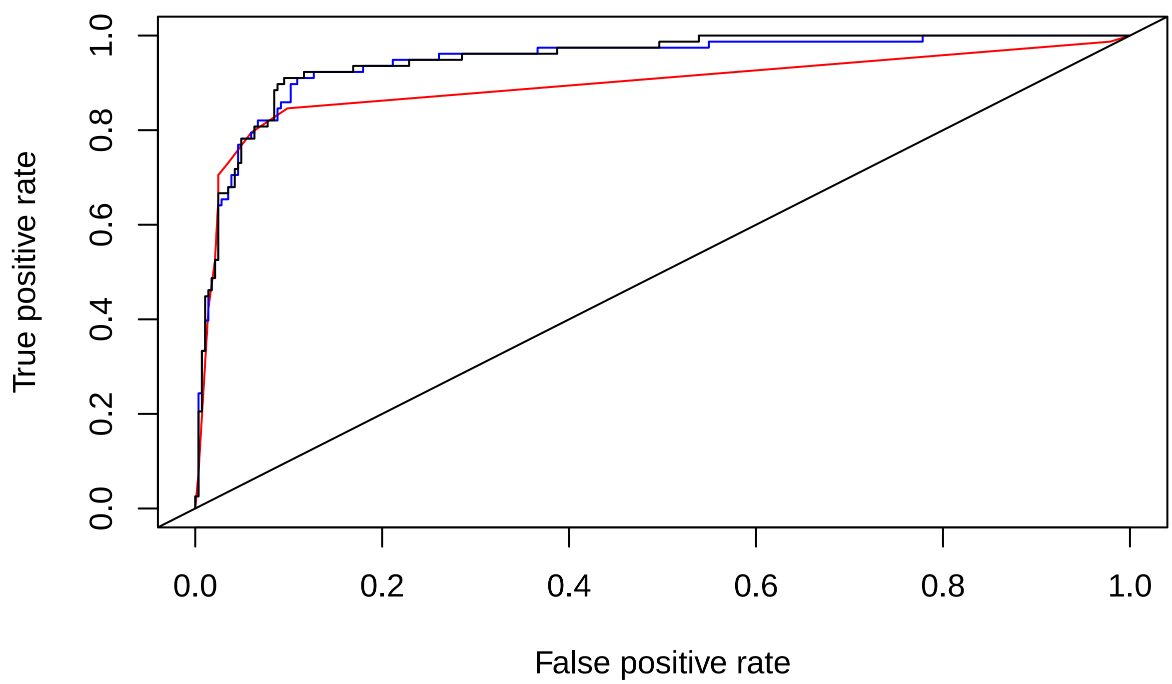
There are 19 out of 24 coefficients in total are non-zero coefficients, which are `Men`, `Women`, `White`, `VotingAgeCitizen`, `Poverty`, `Professional`, `Service`, `Office`, `Production`, `Drive`, `Carpool`, `Transit`, `OtherTransp`, `WorkAtHome`.

Compare the value of coefficients to the ones from unpenalized logistic regression, it is easy to see that the value of coefficients after penalized are smaller than the unpenalized ones. The reason for this is that after using the LASSO penalty, the penalty will limit the influence of variables and also only catch significant variables. For here, after lasso penalty, only 19 out of 24 variables have been caught by the penalized regression. These limitation from LASSO penalty all help to prevent the overfitting that may exist in the case of unpenalized logistic regression.

Save training and test errors to the `records`

```
##          train.error test.error
## tree         0.07751    0.08287
## logistic     0.06782    0.09669
## lasso        0.06990    0.09669
```

Question 18: Compute ROC curves on test data




```
##          AUC value
## tree        0.8943
## logistic    0.9442
## lasso       0.9483
```

Based on the classification results, for decision tree, the pro is that the decision tree is usually very easy to interpret how observations have been splitted and it is also easy to use to handle qualitative predictors. It helps to capture interactions between features in the data. However, the con of decision tree is the predictive accuracy comparied to logistic regression and LASSO logistic regression is relatively low and the differentce of tree results are varies a lot between unpruned and pruned ones. It also fails to deal with the linear relationships like logistic regression and only captures part of the significant variables.

For logistic regression, the pro is that it is easy to interpret how one unit change in variables may cause to the logit of response and it also has a relatively high predictive accuracy. However, the con of logistic regression is that it may sometimes lead to overfitting of data set and can only be used for qualitative response. It also requires to have no perfect colinearity and only make binomial classification.

For LASSO logistic regression, the pro is that it helps to prevent the overfitting. Comparing to logistic regression, its coefficient estimates are sparse, which means that it set some of the features to be zero to remove the features that are not significantly related to the response but with similar predictive accuracy. As a result, it is good for LASSO to find significant variables. The con of LASSO logistic regression is that LASSO coefficient estimates are not scale equivariant. The LASSO should be used after standardized the predictors. Also, since LASSO gets rid of many variables, sometimes it may causes problesm on the prediction results.

Based on the above pros and cons, it is more appropriate to have different classifiers for answering different kinds of questions about the election. For exmaple, decision tree here may be more appropriate to predict the possible outcomes of the candidate winning in the first round since it will involving multiple candidates intead of the binomial classification like logistic regression. Logistic regression may be good to predict the final winning in the final voting round, since during that time only two candidates will be involved in the prediction. And LASSO logistic regression may be good to find out the voters' preference when voting. As a result, different classifier is more appropriate to answer differnet kinds of questions about the eleciton.

Taking it further

Question 19: Explore additional classification methods.

KNN

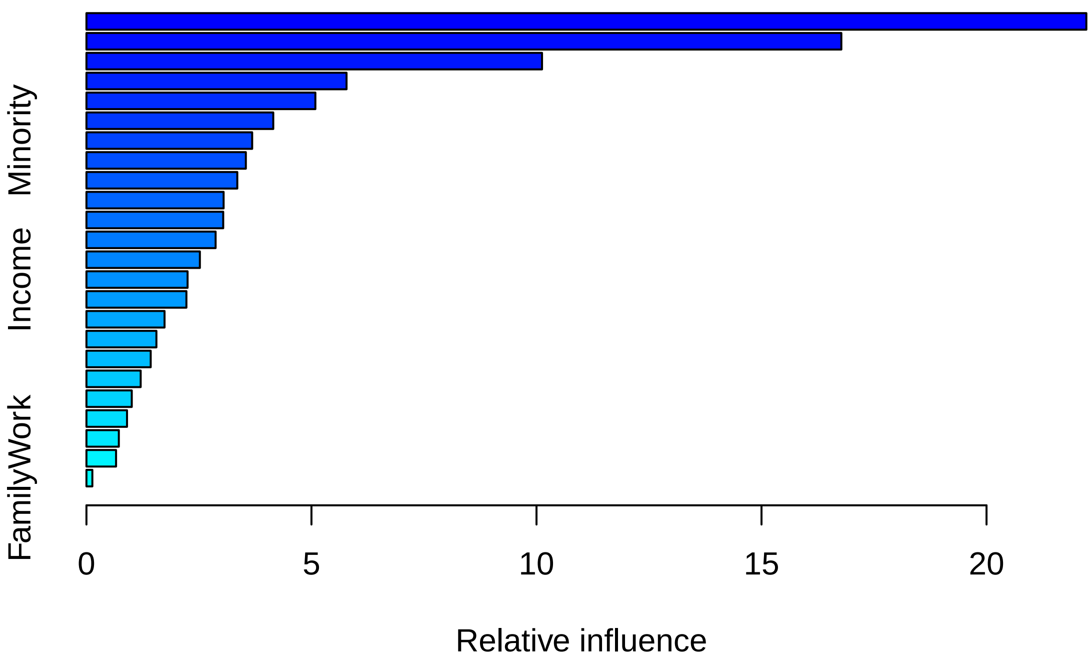
```
##                 true
## predicted       Donald Trump Joe Biden
##   Donald Trump          1113        78
##   Joe Biden               24       230
```

```
##                 true
## predicted       Donald Trump Joe Biden
##   Donald Trump           275        27
##   Joe Biden                9        51
```

```
## [1] "knn train error rate: 0.0705882352941176"
```

```
## [1] "knn test error rate: 0.0994475138121547"
```

Boosting




```
##                                var rel.inf
## Transit                    Transit  22.218
## White                        White  16.774
## Women                        Women  10.123
## Professional          Professional   5.780
## TotalPop                  TotalPop   5.087
## VotingAgeCitizen VotingAgeCitizen   4.151
```

```
## [1] "boosting train error rate: 0.00899653979238754"
```

```
## [1] "boosting test error rate: 0.0828729281767956"
```

Compare Error Rate

```
##             train.error test.error
## tree           0.077509    0.08287
## logistic       0.067820    0.09669
## lasso          0.069896    0.09669
## knn            0.070588    0.09945
## boosting       0.008997    0.08287
```

The first method we use is the KNN method. Compare to the previous three methods, the pro is that it can be used for both classifications and regressions. However, for KNN methods, the cons are pretty obvious.It only works slower than the logistic regression and also high dimensions of dataset will cause KNN to struggle on the predictive accuracy. And since KNN is a non-parametric method, we do not expect it to work as well as logistic regressions. This can be verified from its error rate that is higher than the other methods, which means that the KNN methods may not be very suitable for this dataset.

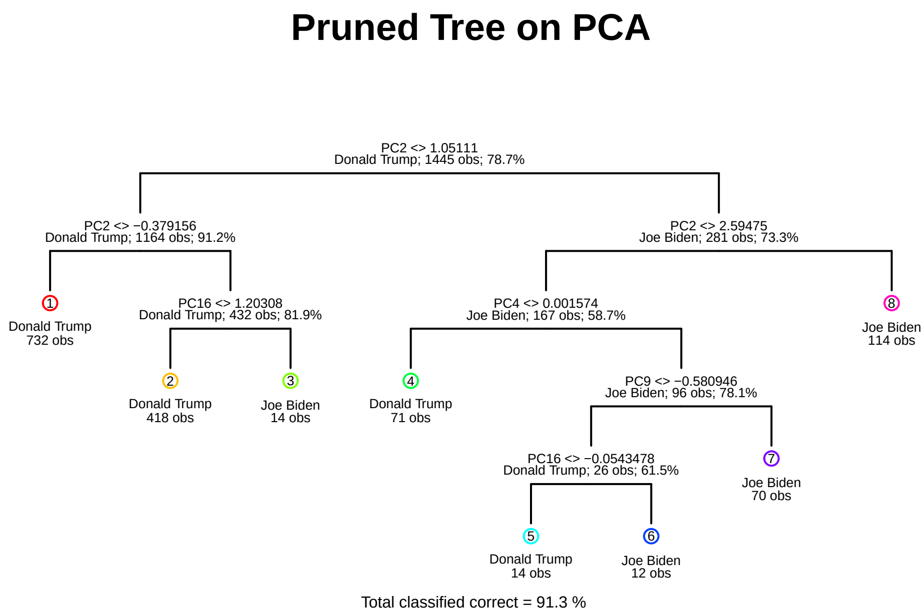
The second method we use is the Boosting method. Boosting method gives out a ranking of how much each feature influence the prediction. For here, it easy to see that the top 6 predictor with high relative influence are: `Transit`, `White`, `Women`, `Professional`, `TotalPop`, and `VotingAgeCitizen`. This result is quite similar to what we get from the decision tree method and the logistic regression method. As a result, we expect it to work as well as the previous methods, which can be verified from its relative similar error rate with the ones of previous methods. Therefore, compared to the previous methods, boosting methods works as well as them and gives out quite similar result.

Question 20:

Instead of using the native attributes (the original features), we can use principal components to create new (and lower dimensional) set of features with which to train a classification model. Use decision tree methods to compare the result trained on the original features with those trained on PCA features.

```
## Importance of components:
##                          PC1   PC2   PC3    PC4    PC5    PC6    PC7   PC8
## Standard deviation     2.294 1.991 1.774 1.2550 1.1533 1.0997 1.0472 1.004
## Proportion of Variance 0.219 0.165 0.131 0.0656 0.0554 0.0504 0.0457 0.042
## Cumulative Proportion  0.219 0.384 0.516 0.5812 0.6366 0.6870 0.7327 0.775
##                           PC9   PC10   PC11   PC12   PC13   PC14   PC15  PC16
## Standard deviation     0.9659 0.8843 0.8545 0.7507 0.7324 0.6639 0.5940 0.579
## Proportion of Variance 0.0389 0.0326 0.0304 0.0235 0.0223 0.0184 0.0147 0.014
## Cumulative Proportion  0.8135 0.8461 0.8765 0.9000 0.9224 0.9407 0.9554 0.969
##                           PC17   PC18    PC19    PC20    PC21    PC22    PC23
## Standard deviation     0.47009 0.4436 0.36708 0.27803 0.24941 0.19741 0.05546
## Proportion of Variance 0.00921 0.0082 0.00561 0.00322 0.00259 0.00162 0.00013
## Cumulative Proportion  0.97862 0.9868 0.99243 0.99565 0.99824 0.99987 1.00000
##                          PC24
## Standard deviation     0.0105
## Proportion of Variance 0.0000
## Cumulative Proportion  1.0000
```

Decision Tree




```
##                train.error test.error
## original tree     0.07751    0.08287
## pca tree          0.08720    0.36188
```

By using principal components to train a classification model, both training error rate and testing error rate increase compare to the original decision tree. Also, another problem that exist here is that this decision tree using the principal components seems to be overfitted, which contains a relatively low train error rate and a relatively high test error rate. This means that maybe for this data set, the decision tree method is not suitable to use principal components as the input and it does not increase the predictive accuracy for the decision tree method in this data set. Also, when looking at the tree plot, it is also hard to interpret what the tree is splitting on to further understand what the voting behavior for the counties. As a result, at least for this classification method, it is better to use the original features to fit the model.

Question 21: Interpret and discuss any overall insights gained in this analysis and possible explanations.

Based on all the above steps and calculations, we can learn that predicting election results is challenging due to all the variables that affect them, but this study teaches us how to target and zero in on the most important ones to make reliable predictions. And the first and most important takeaway from this research was that we can realize there were just two major candidates: Biden and Donald. After drawing the plot in "Question 7: Color the map by the winning candidate for each state." I plotted the unemployment rate on a map of states in question 9 and we can conclude that it looked a lot like the map of candidates. However, demographics are likely the most important impact. As can be seen in the decision tree, the main criterion for deciding between Biden and Donald is whether or not the county is predominantly white. Although, there are some differences because certain counties were broken down into two smaller ones, and some cities were counted as counties, which may have skewed the results by making it harder to determine the vote outcome in certain counties.

In the following quesions, as we used all the three methods—tree, knn, and logistic regression, we all get a pretty low test error rate, suggesting that the traits for each county are strong predictors of whether counties would vote for Biden or Trump. Therefore, tt is possible to conclude that the prediction of the presidential election's victor can be predicted using any of the three approaches since they all accurately predict which counties would vote for which candidate.In question 12, we found 10 clusters, but we are pretty sure that 10 clusters are not be the optimal amount for identifying meaningful county-level subgroups. And we need to experiment with various cluster sizes to find the optimal solution. The dendogram produced by hierarchical clustering did not provide a very helpful visual representation of the resulting groupings, as the various counties provided data. According to the data in the county level, we can find out that the the counties voted for Biden had a lower average poverty rate than the counties that voted for Trump, according to our depiction of of the unemployment rate in question 9 and poverty rate, and the income factor came out to be quite influential. And the factor transit also stands out, as the county's poverty level may also be related with the transit status.

We believe that collecting additional data is a pretty influential step to begin the proposed direction, as the election happened every four years, and we can image a lot will happen in each county in the four years that will definitely influence voters' preferences in a county. As a county may suddenly face poverty because of weather factors or policy changes, and these methodologies will not be useful for predicting future elections since voters will choose the policy fits that are appropriate for the situation at the time. Additionally, other data is necessary to be gathered and used to better forecast and categorize the counties' voting preference, as whether there are flipped allegiances between the two elections, which leads to the importance of gathering the data such as the candidate's physical presence in that county, and the outcome of previous elections in that county. Overall, a lot more factors should be included and thought during the proess of predicting the election results, as it is a very difficult and important prognostic process. For possible process, we need to learn more about each county and each domain, just like what we have mentioned above, we should also think about other factors like the changing factors like election news or unexpected disasters to be included into the statistics factors in order to have a more solid prediction for future elections.