PSTAT 131 Final Project

Cass Bivens & Vance Sine

12/3/2021

First, we will read in and clean our 'census' data set and add the District of Columbia to the state.name and state.abb global variables. The 'census' data set contains many demographic variables for each county in the U.S.

```
state.name <- c(state.name, "District of Columbia")
state.abb <- c(state.abb, "DC")
## read in census data
census <- read_csv("/Users/vancesine/Downloads/acs2017_county_data.csv") %>% select(-CountyId, -ChildPo
  mutate(State = state.abb[match(`State`, state.name)]) %>%
  filter(State != "PR")
```

```
## Rows: 3220 Columns: 37
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): State, County
## dbl (35): CountyId, TotalPop, Men, Women, Hispanic, White, Black, Native, As...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(census)
```

```
## # A tibble: 6 x 31
##   State County   TotalPop   Men  Women Hispanic White Black Native Asian Pacific
##   <chr> <chr>       <dbl> <dbl>  <dbl>    <dbl> <dbl> <dbl>  <dbl> <dbl>   <dbl>
## 1 AL    Autauga~    55036 26899  28137      2.7  75.4  18.9    0.3   0.9       0
## 2 AL    Baldwin~   203360 99527 103833      4.4  83.1   9.5    0.8   0.7       0
## 3 AL    Barbour~    26201 13976  12225      4.2  45.7  47.8    0.2   0.6       0
## 4 AL    Bibb Co~    22580 12251  10329      2.4  74.6  22      0.4   0         0
## 5 AL    Blount ~    57667 28490  29177      9    87.4   1.5    0.3   0.1       0
## 6 AL    Bullock~    10478  5616   4862      0.3  21.6  75.6    1     0.7       0
## # ... with 20 more variables: VotingAgeCitizen <dbl>, Poverty <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>,
## #   Unemployment <dbl>
```

We now read in and clean our 'education' data set. This data set contains county-level educational attainment for adults age 25 and older in 1970-2019. We will be specifically using data from 2015-2019.

```
education <- read_csv("/Users/vancesine/Downloads/Education.csv") %>%
  filter(!is.na(`2003 Rural-urban Continuum Code`)) %>%
  filter(State != "PR") %>%
```

```
  select(-`FIPS Code`,
         -`2003 Rural-urban Continuum Code`,
         -`2003 Urban Influence Code`,
         -`2013 Rural-urban Continuum Code`,
         -`2013 Urban Influence Code`) %>%
  rename(County = `Area name`)
```

```
## Rows: 3283 Columns: 47
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): FIPS Code, State, Area name
## dbl (24): 2003 Rural-urban Continuum Code, 2003 Urban Influence Code, 2013 R...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(education)
```

```
## # A tibble: 6 x 42
##   State County         `Less than a high ~ `High school dipl~ `Some college (1-~
##   <chr> <chr>                        <dbl>              <dbl>              <dbl>
## 1 AL    Autauga County                6611               3757                933
## 2 AL    Baldwin County               18726               8426               2334
## 3 AL    Barbour County                8120               2242                581
## 4 AL    Bibb County                   5272               1402                238
## 5 AL    Blount County                10677               3440                626
## 6 AL    Bullock County                4245                958                305
## # ... with 37 more variables: Four years of college or higher, 1970 <dbl>,
## #   Percent of adults with less than a high school diploma, 1970 <dbl>,
## #   Percent of adults with a high school diploma only, 1970 <dbl>,
## #   Percent of adults completing some college (1-3 years), 1970 <dbl>,
## #   Percent of adults completing four years of college or higher, 1970 <dbl>,
## #   Less than a high school diploma, 1980 <dbl>,
## #   High school diploma only, 1980 <dbl>, ...
```

Preliminary data analysis:

1)

```
## [1] 3142   31
```

Thus our data frame is 3142 X 31 meaning the data frame has 31 variables which were measured for each county, and 3142 counties had data collected from them.

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

The census data frame is not missing any values.

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

We have the 50 states and the federal District of Columbia.

2)

```
## [1] 3143   42
```

```
## [1] 3125   42
```

As we can see we only have 3125 counties which contain all of their data, so 18 counties contained missing values in the data set.

```
## [1] 1877
```

```
## [1] 1877
```

Both sets of data have 1877 distinct counties.

DATA WRANGLING 3)

```
## [1] 3125   42
```

4)

```
## # A tibble: 6 x 6
##   State County         `Less than a high ~ `High school dipl~ `Some college or ~
##   <chr> <chr>                        <dbl>              <dbl>              <dbl>
## 1 AL    Autauga County                4291              12551              10596
## 2 AL    Baldwin County               13893              41797              47274
## 3 AL    Barbour County                4812               6396               4676
## 4 AL    Bibb County                   3386               7256               3848
## 5 AL    Blount County                 7763              13299              13519
## 6 AL    Bullock County                1798               2860               1587
## # ... with 1 more variable: Bachelor's degree or higher, 2015-19 <dbl>
```

```
## # A tibble: 6 x 7
##   State County         `Less than a high ~ `High school dipl~ `Some college or ~
##   <chr> <chr>                        <dbl>              <dbl>              <dbl>
## 1 AL    Autauga County                4291              12551              10596
## 2 AL    Baldwin County               13893              41797              47274
## 3 AL    Barbour County                4812               6396               4676
## 4 AL    Bibb County                   3386               7256               3848
## 5 AL    Blount County                 7763              13299              13519
## 6 AL    Bullock County                1798               2860               1587
## # ... with 2 more variables: Bachelor's degree or higher, 2015-19 <dbl>,
## #   Total Population <dbl>
```

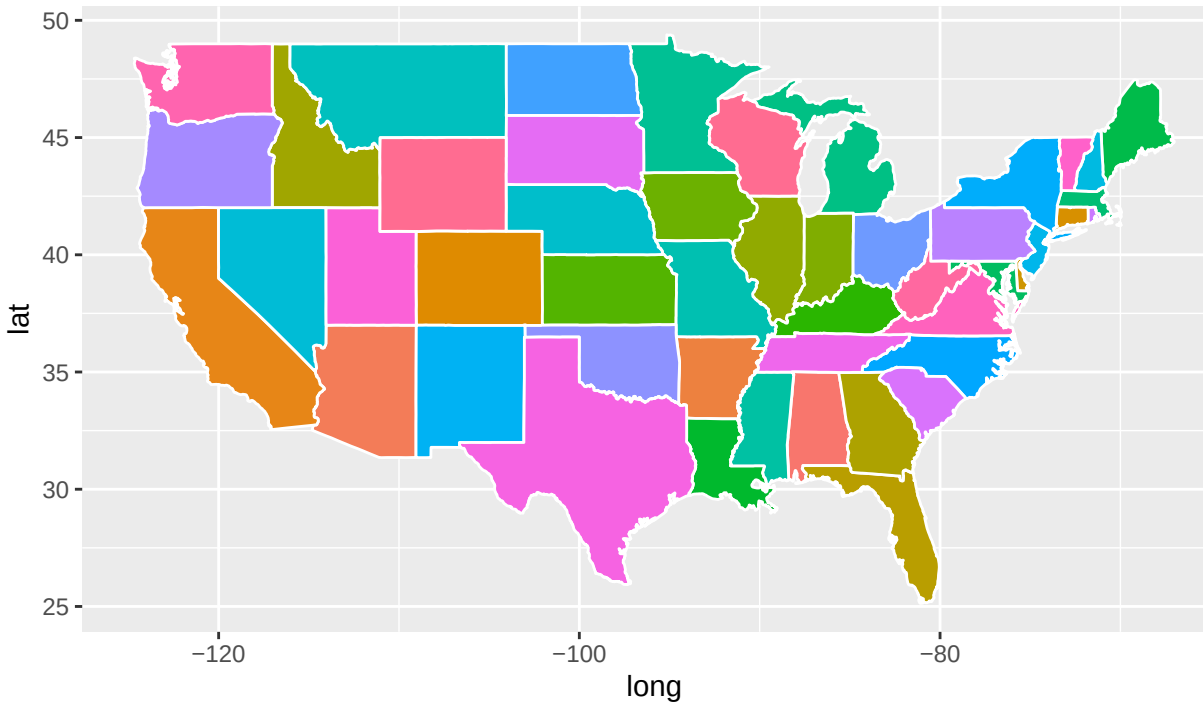
5)

```
##   State Less than a high school diploma, 2015-19
## 1    AK                                   32338
## 2    AL                                  458922
## 3    AR                                  270168
## 4    AZ                                  604935
## 5    CA                                 4418675
## 6    CO                                  314312
##   High school diploma only, 2015-19 Some college or associate's degree, 2015-19
## 1                            126881                                        162816
## 2                           1022839                                        993344
## 3                            684659                                        593576
## 4                           1124129                                       1594817
## 5                           5423462                                       7648680
## 6                            810659                                       1114680
##   Bachelor's degree or higher, 2015-19 Total Population
## 1                               137666           459701
## 2                               845772          3320877
## 3                               463236          2011639
## 4                              1392598          4716479
```

```
## 5                              8980726         26471543
## 6                              1538936          3778587
```

6)

```
##   State Less than a high school diploma, 2015-19
## 1    AK                                   32338
## 2    AL                                  458922
## 3    AR                                  270168
## 4    AZ                                  604935
## 5    CA                                 4418675
## 6    CO                                  314312
##   High school diploma only, 2015-19 Some college or associate's degree, 2015-19
## 1                            126881                                        162816
## 2                           1022839                                        993344
## 3                            684659                                        593576
## 4                           1124129                                       1594817
## 5                           5423462                                       7648680
## 6                            810659                                       1114680
##   Bachelor's degree or higher, 2015-19 Total Population
## 1                               137666           459701
## 2                               845772          3320877
## 3                               463236          2011639
## 4                              1392598          4716479
## 5                              8980726         26471543
## 6                              1538936          3778587
##                       Largest education level
## 1 Some college or associate's degree, 2015-19
## 2           High school diploma only, 2015-19
## 3           High school diploma only, 2015-19
## 4 Some college or associate's degree, 2015-19
## 5        Bachelor's degree or higher, 2015-19
## 6        Bachelor's degree or higher, 2015-19
```



VISUALIZATION 7)

```
##        long      lat group order region subregion
## 1 -87.46201 30.38968     1     1     AL      <NA>
## 2 -87.48493 30.37249     1     2     AL      <NA>
## 3 -87.52503 30.37249     1     3     AL      <NA>
## 4 -87.53076 30.33239     1     4     AL      <NA>
## 5 -87.57087 30.32665     1     5     AL      <NA>
## 6 -87.58806 30.32665     1     6     AL      <NA>
```

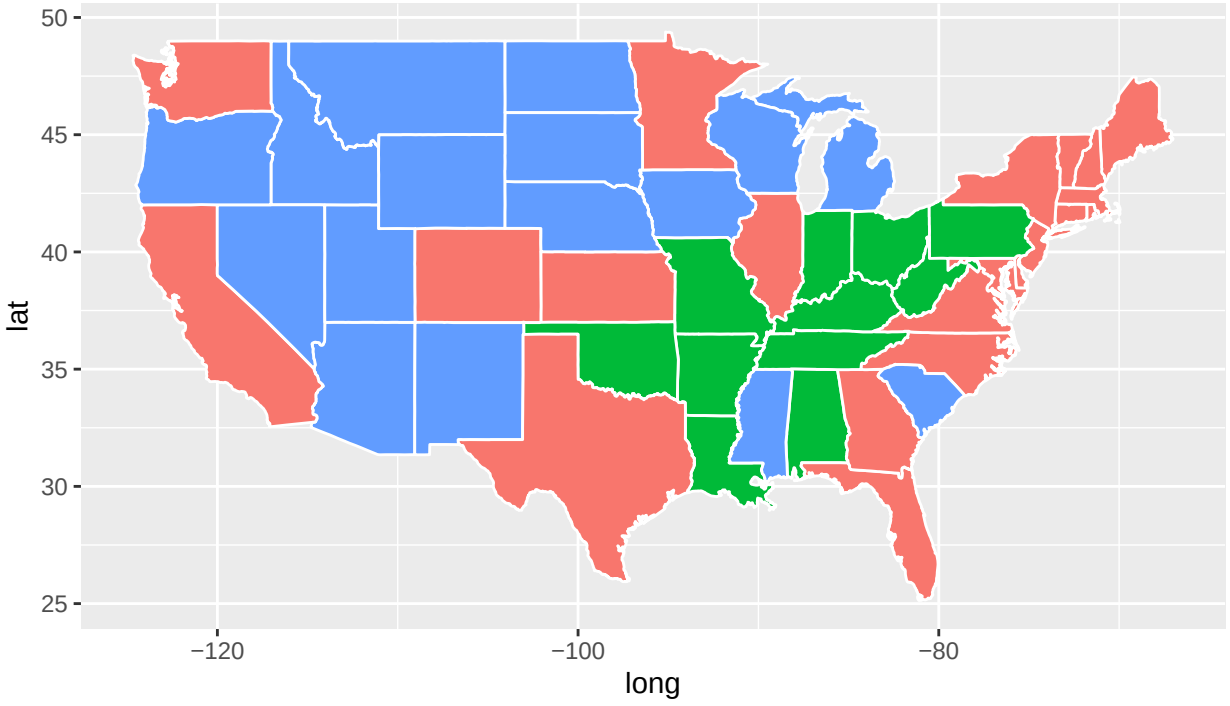
```
## Joining, by = "State"
```

```
##   State Less than a high school diploma, 2015-19
## 1    AK                                   32338
## 2    AL                                  458922
## 3    AL                                  458922
## 4    AL                                  458922
## 5    AL                                  458922
## 6    AL                                  458922
##   High school diploma only, 2015-19 Some college or associate's degree, 2015-19
## 1                            126881                                       162816
## 2                           1022839                                       993344
## 3                           1022839                                       993344
## 4                           1022839                                       993344
## 5                           1022839                                       993344
## 6                           1022839                                       993344
##   Bachelor's degree or higher, 2015-19 Total Population
## 1                               137666           459701
## 2                               845772          3320877
## 3                               845772          3320877
## 4                               845772          3320877
## 5                               845772          3320877
## 6                               845772          3320877
##                      Largest education level      long      lat group order
```

```
## 1 Some college or associate's degree, 2015-19        NA       NA    NA    NA
## 2           High school diploma only, 2015-19 -87.46201 30.38968     1     1
## 3           High school diploma only, 2015-19 -87.48493 30.37249     1     2
## 4           High school diploma only, 2015-19 -87.52503 30.37249     1     3
## 5           High school diploma only, 2015-19 -87.53076 30.33239     1     4
## 6           High school diploma only, 2015-19 -87.57087 30.32665     1     5
##   region subregion
## 1   <NA>      <NA>
## 2     AL      <NA>
## 3     AL      <NA>
## 4     AL      <NA>
## 5     AL      <NA>
## 6     AL      <NA>
```

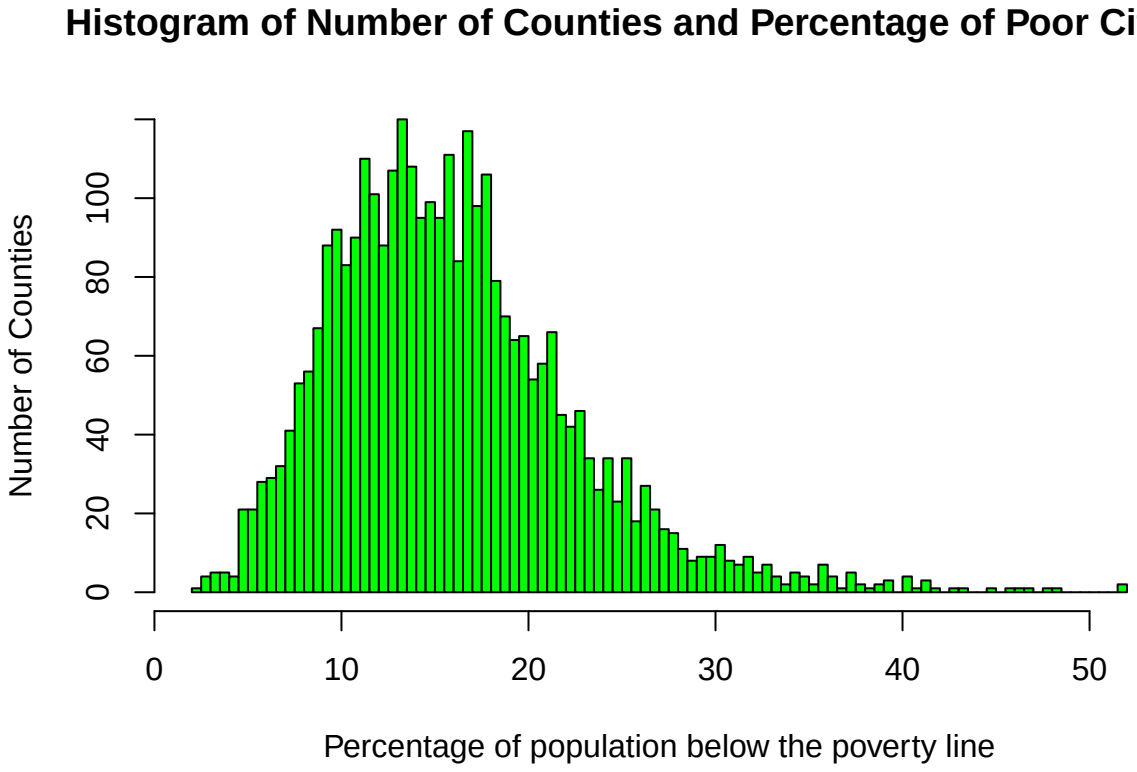
```
## function (x, as.factor = FALSE)
## {
##     if (as.factor) {
##         labs <- rownames(x, do.NULL = FALSE, prefix = "")
##         res <- factor(.Internal(row(dim(x))), labels = labs)
##         dim(res) <- dim(x)
##         res
##     }
##     else .Internal(row(dim(x)))
## }
## <bytecode: 0x7f7de7f496d8>
## <environment: namespace:base>
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

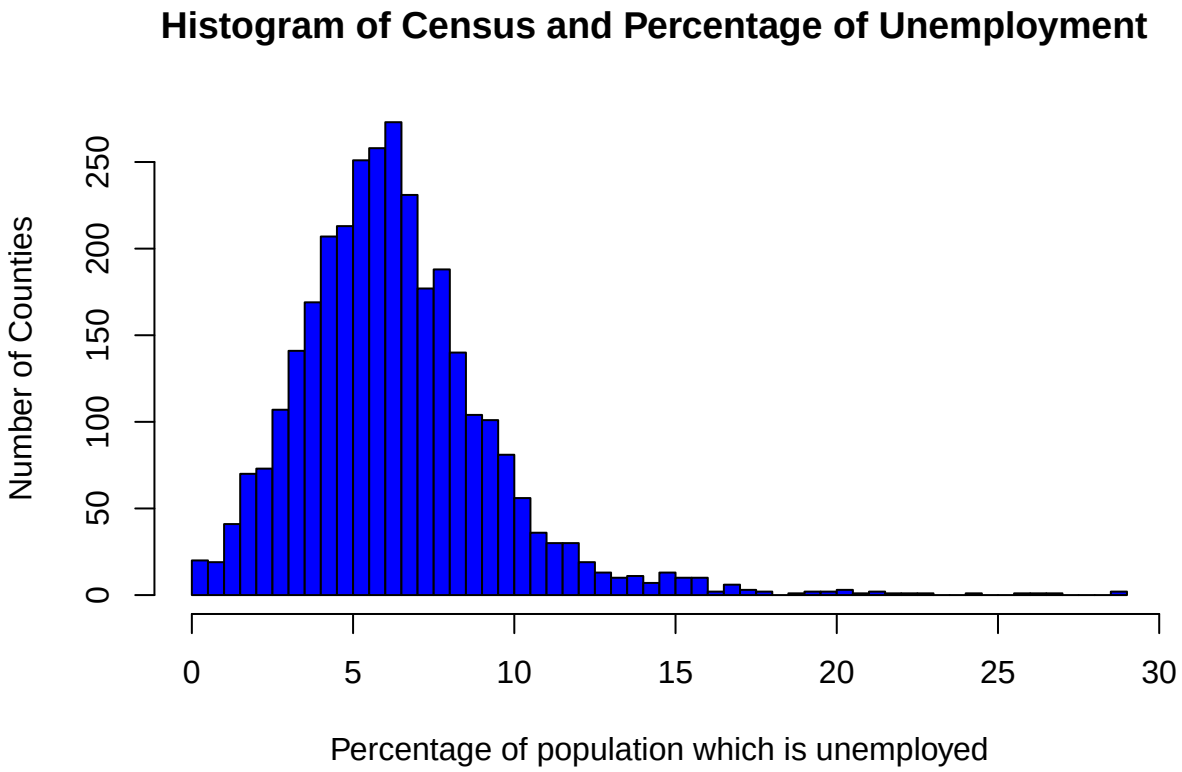


8)

```
hist(census$Poverty, breaks = 100, xlab = "Percentage of population below the poverty line", ylab = "Nu
```




```
hist(census$Unemployment, breaks = 100, xlab = "Percentage of population which is unemployed", ylab = "
```



9)

10)

```
## # A tibble: 5 x 21
##   State County      TotalPop   Men VotingAgeCitizen Poverty Professional Service
##   <chr> <chr>          <dbl> <dbl>            <dbl>   <dbl>        <dbl>   <dbl>
## 1 AL    Autauga Co~    55036 0.489            0.745    13.7         35.3    18
## 2 AL    Baldwin Co~   203360 0.489            0.764    11.8         35.7    18.2
## 3 AL    Barbour Co~    26201 0.533            0.774    27.2         25      16.8
## 4 AL    Bibb County    22580 0.543            0.782    15.2         24.4    17.6
## 5 AL    Blount Cou~    57667 0.494            0.737    15.6         28.5    12.9
## # ... with 13 more variables: Office <dbl>, Production <dbl>, Drive <dbl>,
## #   Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Minority <dbl>
```

Dimensionality Reduction 11)

```
##     State               County             TotalPop             Men        
##  Length:3142        Length:3142        Min.   :      74   Min.   :0.4190  
##  Class :character   Class :character   1st Qu.:   10945   1st Qu.:0.4890  
##  Mode  :character   Mode  :character   Median :   25692   Median :0.4960  
##                                        Mean   :  102166   Mean   :0.5008  
##                                        3rd Qu.:   67445   3rd Qu.:0.5058  
##                                        Max.   :10105722   Max.   :0.8083  
##  VotingAgeCitizen    Poverty       Professional      Service      
##  Min.   :0.4569   Min.   : 2.40   Min.   :11.40   Min.   : 0.00  
##  1st Qu.:0.7313   1st Qu.:11.30   1st Qu.:27.30   1st Qu.:15.70  
##  Median :0.7596   Median :15.20   Median :30.50   Median :17.80  
##  Mean   :0.7495   Mean   :15.99   Mean   :31.54   Mean   :18.12  
##  3rd Qu.:0.7816   3rd Qu.:19.40   3rd Qu.:34.90   3rd Qu.:20.07  
##  Max.   :0.9767   Max.   :52.00   Max.   :69.00   Max.   :46.40  
##      Office        Production        Drive          Carpool      
##  Min.   : 4.80   Min.   : 0.00   Min.   : 4.60   Min.   : 0.000  
##  1st Qu.:19.90   1st Qu.:11.60   1st Qu.:77.20   1st Qu.: 8.100  
##  Median :22.00   Median :15.50   Median :81.00   Median : 9.500  
##  Mean   :21.78   Mean   :15.93   Mean   :79.52   Mean   : 9.899  
##  3rd Qu.:23.80   3rd Qu.:19.60   3rd Qu.:84.00   3rd Qu.:11.300  
##  Max.   :37.20   Max.   :48.70   Max.   :97.20   Max.   :29.300  
##     Transit         OtherTransp       WorkAtHome      MeanCommute   
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 5.10  
##  1st Qu.: 0.1000   1st Qu.: 0.900   1st Qu.: 2.900   1st Qu.:19.60  
##  Median : 0.3000   Median : 1.300   Median : 4.100   Median :23.10  
##  Mean   : 0.9368   Mean   : 1.603   Mean   : 4.803   Mean   :23.35  
##  3rd Qu.: 0.8000   3rd Qu.: 1.900   3rd Qu.: 5.800   3rd Qu.:26.90  
##  Max.   :61.8000   Max.   :43.200   Max.   :33.000   Max.   :45.10  
##     Employed       PrivateWork     SelfEmployed      FamilyWork    
##  Min.   :0.1017   Min.   :31.10   Min.   : 0.000   Min.   :0.0000  
##  1st Qu.:0.3960   1st Qu.:71.70   1st Qu.: 5.200   1st Qu.:0.1000  
##  Median :0.4429   Median :76.30   Median : 6.800   Median :0.2000  
##  Mean   :0.4383   Mean   :75.07   Mean   : 7.758   Mean   :0.2824  
##  3rd Qu.:0.4861   3rd Qu.:80.30   3rd Qu.: 9.175   3rd Qu.:0.3000  
##  Max.   :0.7326   Max.   :88.80   Max.   :38.000   Max.   :8.0000  
##     Minority     
##  Min.   : 0.00  
##  1st Qu.: 5.50  
##  Median :13.55  
##  Mean   :21.24
```

```
##  3rd Qu.:32.50
##  Max.   :99.30
```

```
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
```

```
## Warning in FUN(newX[, i], ...): NAs introduced by coercion
```

```
##            State           County        TotalPop              Men
##               NA               NA     1.077758e+11     5.862892e-04
## VotingAgeCitizen          Poverty     Professional          Service
##     2.817957e-03     4.300811e+01     4.277364e+01     1.358740e+01
##           Office       Production            Drive          Carpool
##     9.297472e+00     3.380418e+01     5.884476e+01     8.485439e+00
##          Transit      OtherTransp       WorkAtHome      MeanCommute
##     9.616366e+00     2.828719e+00     9.462904e+00     3.179096e+01
##         Employed      PrivateWork     SelfEmployed       FamilyWork
##     4.269568e-03     5.706079e+01     1.495226e+01     2.040316e-01
##         Minority
##     3.977816e+02
```

I am going to center and scale our df before preforming the PCA, as the variables have a wide range of mean values and differing variances. Specifically, the TotalPop variable has a much higher mean and variance than any of the other varibles, so if we failed to scale the pricipal components would most likely be artifically driven by TotalPop.

 # A tibble: 6 x 19

TotalPop Men VotingAgeCitizen Poverty Professional Service Office Production

1 55036 0.489 0.745 13.7 35.3 18 23.2 15.4

2 203360 0.489 0.764 11.8 35.7 18.2 25.6 10.8

3 26201 0.533 0.774 27.2 25 16.8 22.6 24.1

4 22580 0.543 0.782 15.2 24.4 17.6 19.7 22.4

5 57667 0.494 0.737 15.6 28.5 12.9 23.3 19.5

6 10478 0.536 0.784 28.5 19.7 17.1 18.6 30.6

 # ... with 11 more variables: Drive , Carpool , Transit ,

 # OtherTransp , WorkAtHome , MeanCommute , Employed ,

 # PrivateWork , SelfEmployed , FamilyWork , Minority

```
##                         PC1         PC2
## TotalPop          0.02647537 -0.17639788
## Men               0.06734237  0.23100987
## VotingAgeCitizen  0.02508638 -0.02324311
## Poverty          -0.24039363  0.37889770
## Professional      0.34469432 -0.25298598
## Service          -0.09122182  0.29681236
```

```
## Office           -0.14792201 -0.23080624
## Production       -0.29166926  0.04164515
## Drive            -0.35781105 -0.19769392
## Carpool          -0.06792515  0.23776802
## Transit           0.10831749 -0.10572358
## OtherTransp       0.11448636  0.17412743
## WorkAtHome        0.42673365  0.05461074
## MeanCommute      -0.17805008 -0.13249119
## Employed          0.26003242 -0.38804129
## PrivateWork      -0.27012383 -0.39056527
## SelfEmployed      0.36051238  0.15252643
## FamilyWork        0.21732612  0.14485148
## Minority         -0.11484242  0.24350578
```

```
##   WorkAtHome SelfEmployed        Drive Professional   Production  PrivateWork
##    0.4267336    0.3605124    0.3578110    0.3446943    0.2916693    0.2701238
```

#The features, Poverty, Service, Office, Production, Drive, Carpool, MeanCommute, Private Work, and Minority appear to be negative in PC1. This is a scaling technique with negative correlation; meaning that if a negative PC1 increases, the component will become more centered.

#The three features with the largest absolute values are Work at Home, Self Employed, and Drive.

12)

#It appears that we need at least 12 PC to capture 90% of the variance. CLustering 13)

MODELING 14)

```
## # A tibble: 5 x 61
##   State County      TotalPop   Men VotingAgeCitizen Poverty Professional Service
##   <chr> <chr>          <dbl> <dbl>            <dbl>   <dbl>        <dbl>   <dbl>
## 1 AL    Autauga Co~    55036 0.489            0.745    13.7         35.3    18
## 2 AL    Baldwin Co~   203360 0.489            0.764    11.8         35.7    18.2
## 3 AL    Barbour Co~    26201 0.533            0.774    27.2         25      16.8
## 4 AL    Bibb County    22580 0.543            0.782    15.2         24.4    17.6
## 5 AL    Blount Cou~    57667 0.494            0.737    15.6         28.5    12.9
## # ... with 53 more variables: Office <dbl>, Production <dbl>, Drive <dbl>,
## #   Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Minority <dbl>,
## #   Less than a high school diploma, 1970 <dbl>,
## #   High school diploma only, 1970 <dbl>, Some college (1-3 years), 1970 <dbl>,
## #   Four years of college or higher, 1970 <dbl>, ...
```

```
## # A tibble: 6 x 7
##   State County          NOHS    HS  NOBA    BA `Total Population`
##   <chr> <chr>          <dbl> <dbl> <dbl> <dbl>              <dbl>
## 1 AL    Autauga County  4291 12551 10596  9929              37367
## 2 AL    Baldwin County 13893 41797 47274 48148             151112
## 3 AL    Barbour County  4812  6396  4676  2080              17964
## 4 AL    Bibb County     3386  7256  3848  1678              16168
## 5 AL    Blount County   7763 13299 13519  5210              39791
## 6 AL    Bullock County  1798  2860  1587   856               7101
```

```
## # A tibble: 5 x 26
##   State County      TotalPop   Men VotingAgeCitizen Poverty Professional Service
##   <chr> <chr>          <dbl> <dbl>            <dbl> <fct>          <dbl>   <dbl>
```

```
## 1 AL    Autauga Co~    55036 0.489           0.745 1              35.3    18
## 2 AL    Baldwin Co~   203360 0.489           0.764 1              35.7    18.2
## 3 AL    Barbour Co~    26201 0.533           0.774 0              25      16.8
## 4 AL    Bibb County    22580 0.543           0.782 1              24.4    17.6
## 5 AL    Blount Cou~    57667 0.494           0.737 1              28.5    12.9
## # ... with 18 more variables: Office <dbl>, Production <dbl>, Drive <dbl>,
## #   Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Minority <dbl>, NOHS <dbl>, HS <dbl>, NOBA <dbl>,
## #   BA <dbl>, Total Population <dbl>
```

We are going to remove some of the Data which we don't think is relevant to the poverty classification task. We have decided that 'State', 'County', 'VotingAgeCitizen', and 'OtherTransp'. This is because state and county are not something which can be measured or compared therefore they really carry no information relevant to the task. The percentage of voting aged citizens is just a metric telling you some information about what percentage of the population is over the age of 18. This does not contribute anything to our analysis of percentage of people living in poverty in my opinion. We also have removed 'OtherTransp'column because we do not know exactly which means of transportation it is referring to. It could be a means of transportation which correlates with being poor or not being poor.

```
## # A tibble: 6 x 21
##   TotalPop   Men Poverty Professional Service Office Production Drive Carpool
##      <dbl> <dbl> <fct>          <dbl>   <dbl>  <dbl>      <dbl> <dbl>   <dbl>
## 1    55036 0.489 1               35.3    18     23.2       15.4  86       9.6
## 2   203360 0.489 1               35.7    18.2   25.6       10.8  84.7     7.6
## 3    26201 0.533 0               25      16.8   22.6       24.1  83.4    11.1
## 4    22580 0.543 1               24.4    17.6   19.7       22.4  86.4     9.5
## 5    57667 0.494 1               28.5    12.9   23.3       19.5  86.8    10.2
## 6    10478 0.536 0               19.7    17.1   18.6       30.6  73.1    15.7
## # ... with 12 more variables: Transit <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Minority <dbl>, NOHS <dbl>, HS <dbl>, NOBA <dbl>,
## #   BA <dbl>
```

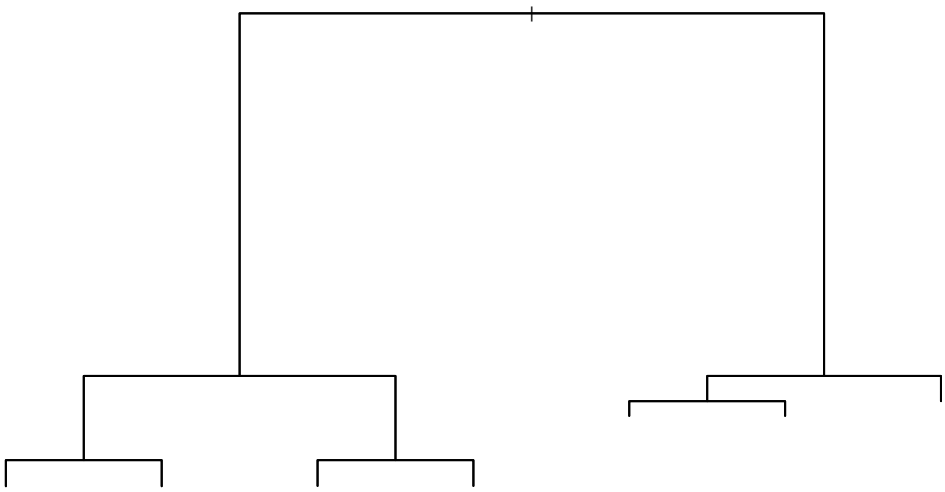
```
calc_error_rate = function(predicted.value, true.value){
  return(mean(true.value!=predicted.value))
}
records = matrix(NA, nrow=3, ncol=2)
colnames(records) = c("train.error","test.error")
rownames(records) = c("tree","logistic","lasso")
records
```

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

The matrix 'records' provides us with a place to store the error rates for the following problems using the 'calc_error_rate' function. 15)

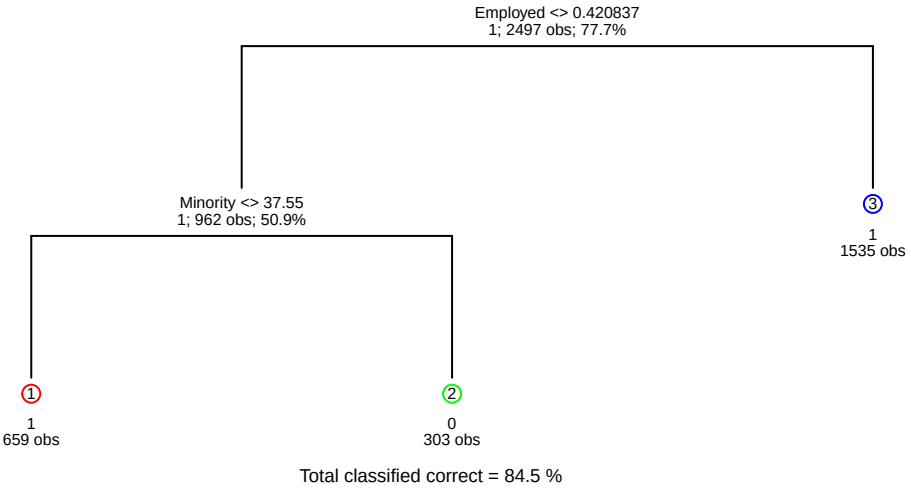
```
## Warning in text.default(x[1L], y[1L], "|", ...): "nodeinfo" is not a graphical
## parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "nodeinfo" is not a
## graphical parameter
```



```
## [1] 402 402 557
```

```
## [1] 7 3 1
```




```
##
## Classification tree:
## snip.tree(tree = poor.tree, nodes = 3:5)
## Variables actually used in tree construction:
## [1] "Employed" "Minority"
## Number of terminal nodes:  3
## Residual mean deviance:  0.7366 = 1837 / 2494
## Misclassification error rate: 0.1554 = 388 / 2497
```

#When employment is less than 0.420837 our tree goes to the minority. When minority is greater than 37.55 our tree goes to poverty. Our tree classified overall 84.5% correctly.

16)

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

```
##
## Call:
## glm(formula = Poverty ~ ., family = "binomial", data = all.tr)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.4509   0.0038   0.1576   0.4240   4.3288
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.698e+01  3.868e+00  -6.976 3.03e-12 ***
## TotalPop     -1.516e-04  1.938e-05  -7.823 5.16e-15 ***
## Men           3.431e+01  2.970e+00  11.551  < 2e-16 ***
## Professional -6.894e-02  2.414e-02  -2.856  0.00428 **
## Service      -1.181e-01  2.634e-02  -4.484 7.32e-06 ***
## Office       -2.668e-02  2.894e-02  -0.922  0.35653
## Production   -9.094e-02  2.227e-02  -4.083 4.44e-05 ***
## Drive         2.284e-02  2.453e-02   0.931  0.35183
## Carpool      -1.565e-02  3.269e-02  -0.479  0.63213
## Transit      -1.078e-01  6.153e-02  -1.753  0.07967 .
## WorkAtHome    1.009e-01  4.454e-02   2.264  0.02355 *
## MeanCommute   2.103e-02  1.536e-02   1.369  0.17106
## Employed      3.033e+01  1.946e+00  15.583  < 2e-16 ***
## PrivateWork   3.079e-02  1.660e-02   1.855  0.06366 .
## SelfEmployed  3.972e-02  3.204e-02   1.240  0.21497
## FamilyWork    1.006e-01  1.814e-01   0.554  0.57927
## Minority     -3.116e-02  4.196e-03  -7.428 1.10e-13 ***
## NOHS          1.937e-04  3.817e-05   5.075 3.87e-07 ***
## HS            1.922e-04  2.934e-05   6.551 5.70e-11 ***
## NOBA          3.455e-04  4.478e-05   7.715 1.21e-14 ***
## BA            1.998e-04  3.025e-05   6.605 3.99e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2650.6  on 2496  degrees of freedom
## Residual deviance: 1375.2  on 2476  degrees of freedom
## AIC: 1417.2
##
## Number of Fisher Scoring iterations: 9

## [1] 0.1201442

## [1] 0.1248

## [1] 0.1553865

## [1] 0.168
```

#Our logistic regression training error is 0.1201442 #Our logistic regression test error is 0.1248 #Our pruned tree training error is 0.1553865 #Our pruned tree test error is 0.168 #The variables, TotalPop, Men, Service, Production, Employed, Minority, NOHS, HS, NOBA, and BA are signigicant variables.The variable men has a coefficient 3.431e+01. For every one unit change in men the log odds of poverty increases by 3.431e+01, holding other variables fixed. The variable HS has a coefficient 1.922e-04 meaning that for every one unit change in HS, the log odds of poverty increases by 1.922e-04. 17)

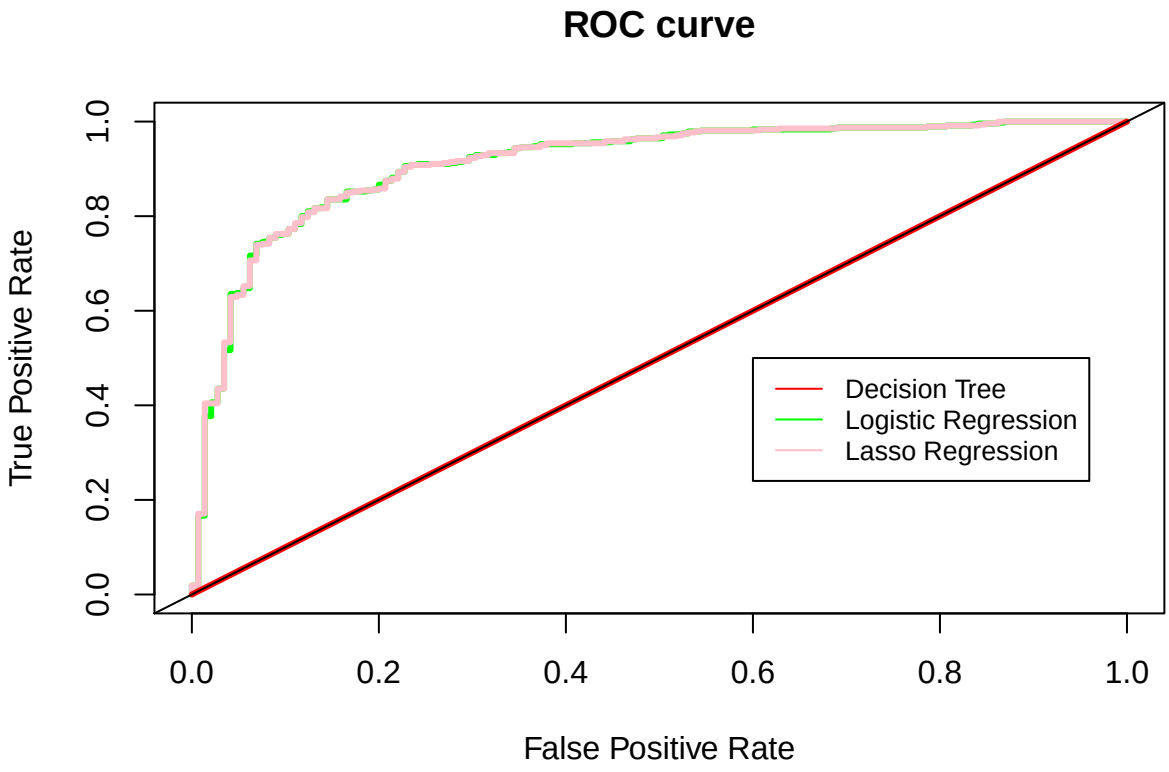
```
## 21 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)  -2.716265e+01
## TotalPop     -1.411091e-04
## Men           3.412703e+01
```

```
## Professional -6.895685e-02
## Service      -1.165866e-01
## Office       -2.567812e-02
## Production   -9.090092e-02
## Drive         2.517933e-02
## Carpool      -1.374781e-02
## Transit      -9.816376e-02
## WorkAtHome    1.032389e-01
## MeanCommute   2.154082e-02
## Employed      3.014389e+01
## PrivateWork   3.137679e-02
## SelfEmployed  4.264213e-02
## FamilyWork    1.020182e-01
## Minority     -3.125180e-02
## NOHS          1.769734e-04
## HS            1.810406e-04
## NOBA          3.239438e-04
## BA            1.841270e-04
```

```
## [1] 1e-05
```

#Our best lambda value is 1e-05 #Every coefficient displayed appears to be nonzero

18)



The green line represents the ROC for logistic regression The pink line represents the ROC for lasso The red line represents the ROC for our trees Logistic and Lasso appear to follow the same curve, with very minor differences between the two

Logistic Regression and Lasso appear to have roughly the same ROC. Where larger AUC(Area under the curve) is preferred. Optimal values are closer to 1 as this curve is made up of rates of classification error. Decision trees are highly interpretable nd do not require you to scale data. However, they are more complex

to implement and can be computationally expensive. Lasso is useful when there is a large amount of training data as this regularization method avoids overfitting. However, this results in a high bias toward a few selected features and results in lower prediction performance. Logistic regression is easy to interpret when you have a high amount of available data to train. From the ROC curve constructed, Lasso Regression appears to have the highest area under curve by a small margin.

19) For exploring the additional classification methods we will employ K-nn and boosting as our strategies. We will continue to utilize the 'all' data set which was used in problems 16-18 and will continue to attempt to classify the poverty rate in a binomial fashion where Poverty = 1 if the poverty rate >20% and Poverty = 0 if the poverty rate < 20%. We start by separating the data into 50% test and 50% training sets, and logging the corresponding variable respionses as well.

```
## # A tibble: 6 x 21
##   TotalPop   Men Poverty Professional Service Office Production Drive Carpool
##      <dbl> <dbl> <fct>          <dbl>   <dbl>  <dbl>      <dbl> <dbl>   <dbl>
## 1    55036 0.489 1               35.3    18     23.2       15.4  86       9.6
## 2   203360 0.489 1               35.7    18.2   25.6       10.8  84.7     7.6
## 3    26201 0.533 0               25      16.8   22.6       24.1  83.4    11.1
## 4    22580 0.543 1               24.4    17.6   19.7       22.4  86.4     9.5
## 5    57667 0.494 1               28.5    12.9   23.3       19.5  86.8    10.2
## 6    10478 0.536 0               19.7    17.1   18.6       30.6  73.1    15.7
## # ... with 12 more variables: Transit <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Minority <dbl>, NOHS <dbl>, HS <dbl>, NOBA <dbl>,
## #   BA <dbl>
```

Now, we are going to use the Knn function to train a Knn classifier, as well as output the confusion matrix for the training data.

```
##           true
## predicted    1    2
##         1  231   54
##         2  120 1156
```

```
## [1] 0.111467
```

We can see for Knn with a K of three we have a training error rate of 11.14%. Now let us look at the test error rate.

```
##           true
## predicted    1    2
##         1  182  155
##         2  169 1055
```

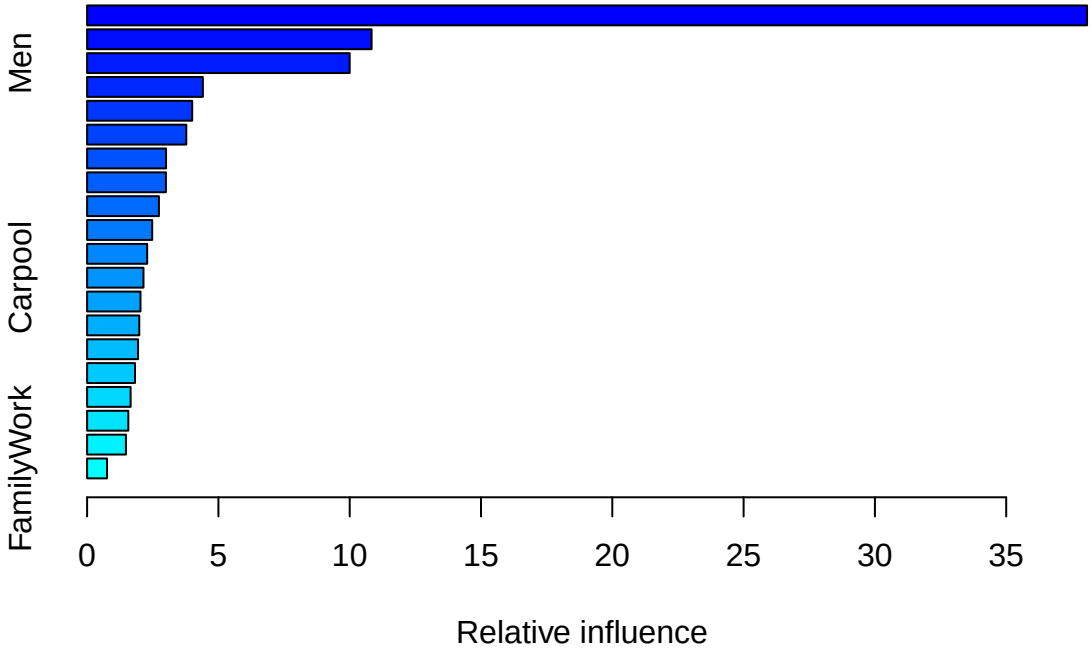
```
## [1] 0.2075593
```

We can see for a Knn with k=3, the test error rate is 20.75%. NEED TO MAKE COMPARISON TO THE THREE METHODS USED IN 16-18.

For the second classification method we will be using the gbm() function to fit boosted classification trees to the our 'all' data set. First, we need to make sure Poverty is coded as a numeric value as either 0 or 1. Next, we split our data into 75% testing data and 25% training data.

```
## # A tibble: 6 x 21
##   TotalPop   Men Poverty Professional Service Office Production Drive Carpool
##      <dbl> <dbl>   <dbl>        <dbl>   <dbl>  <dbl>      <dbl> <dbl>   <dbl>
## 1    55036 0.489       2         35.3    18     23.2       15.4  86       9.6
## 2   203360 0.489       2         35.7    18.2   25.6       10.8  84.7     7.6
## 3    26201 0.533       1         25      16.8   22.6       24.1  83.4    11.1
```

```
## 4    22580 0.543       2         24.4    17.6   19.7       22.4  86.4     9.5
## 5    57667 0.494       2         28.5    12.9   23.3       19.5  86.8    10.2
## 6    10478 0.536       1         19.7    17.1   18.6       30.6  73.1    15.7
## # ... with 12 more variables: Transit <dbl>, WorkAtHome <dbl>,
## #   MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>, SelfEmployed <dbl>,
## #   FamilyWork <dbl>, Minority <dbl>, NOHS <dbl>, HS <dbl>, NOBA <dbl>,
## #   BA <dbl>
```




```
##                        var    rel.inf
## Employed         Employed 38.0786714
## Minority         Minority 10.8337427
## Men                   Men  9.9988474
## MeanCommute   MeanCommute  4.4084922
## Professional Professional  4.0017469
## WorkAtHome     WorkAtHome  3.7774218
## Service           Service  3.0062708
## Production     Production  3.0007396
## PrivateWork   PrivateWork  2.7346632
## NOHS                 NOHS  2.4784927
## NOBA                 NOBA  2.2870573
## Carpool           Carpool  2.1471392
## SelfEmployed SelfEmployed  2.0333596
## Drive               Drive  1.9861110
## Transit           Transit  1.9396718
## Office             Office  1.8238038
## BA                     BA  1.6575148
## TotalPop         TotalPop  1.5689116
## HS                     HS  1.4788972
## FamilyWork     FamilyWork  0.7584451
```

We can see that the employment rate, minority population percentage, and percentage of men in a county were the most important variables in determining if a county was deemed 'poor' or not on the training data. Now let us run our boosted Forrest on the test set and see how it performs.

```
## [1] 0.1267606
```

If we define the boosted tree classifier to select poverty's class with a 'majority rules' style approach then the test error rate is 12.03%. COMPARE TO THE TEST ERROR RATE OF OTHER METHODS.

20) For the next interesting question we are going to use ridge regression to train our computer to estimate poverty rate of each county. first we need to split the data into training and test sets.

```
all <- census.clean %>% left_join(education2, by = c('State' = 'State' , 'County' = 'County')) %>% na.o
x = model.matrix(Poverty~., all)
y = all$Poverty
set.seed(123)
train=sample(1:nrow(x), nrow(x)/2)
test=(-train)
x.tr=x[train,]
y.tr=y[train]
x.te=x[test,]
y.te=y[test]
```

```
'cv.ridge <- cv.glmnet(x.tr, y.tr, alpha = 0, folds = 20)
bestlam <- cv.ridge$lambda.min
bestlam'
```

```
## [1] "cv.ridge <- cv.glmnet(x.tr, y.tr, alpha = 0, folds = 20)\nbestlam <- cv.ridge$lambda.min\nbestl
```

```
'ridge.predict <- predict(cv.ridge, s = bestlam, newx = x.te)
differences = ifelse(abs(y.te-ridge.predict)>sd(y), 1,0)
wrong.rate = sum(differences)/length(y.te)
wrong.rate'
```

```
## [1] "ridge.predict <- predict(cv.ridge, s = bestlam, newx = x.te)\ndifferences = ifelse(abs(y.te-rid
```

We can see here that our ridge regression predicts a poverty rate which is more than a standard deviation of all poverty rates away from the actual poverty rate in the test data 20.11% of the time even when using the lamba selected from 20-fold cross validation. This metric allows us to replicate a classification error rate by saying that an error was made if our prediction was off by more than one standard deviation.

21) Our logistic regression test error is 0.1248. This is close to our our pruned tree test error of 0.168, and is even closer to our boost test error 0.1267606. These values appear to make sense as we would expect a higher test error for a decision tree as decision tree error increases with complexity. For number 19, our goal was to reduce variance at the small tradeoff of bias. One factor that affects our these values is the number of factors that each method considers important. For our pruned tree, there are only 2 factors: minority and employed. Wheras our logistic regression has 11 important factors. Where there were only two factors, this method returned our highest error rate. Boosting is a generative process that improves on the previous model. This returned a very similar error rate to logistic regression. Overall, it appears that logistic regression provides us with the most accurate results while maintaining computational simplicity.