Crime and Communities

Group Member 1 Name: Khang V. Tran **Group Member 1 SID:** 25181590

Group Member 2 Name: Christian Philip Hoeck **Group Member 2 SID:** ____________________

The crime and communities dataset contains crime data from communities in the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. More details can be found at https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized.

The dataset contains 125 columns total; $p = 124$ predictive and 1 target (ViolentCrimesPerPop). There are $n = 1994$ observations. These can be arranged into an $n \times p = 1994 \times 127$ feature matrix $\mathbf{X}$, and an $n \times 1 = 1994 \times 1$ response vector $\mathbf{y}$ (containing the observations of ViolentCrimesPerPop).

Once downloaded (from bCourses), the data can be loaded as follows.

```
library(readr)
CC <- read_csv("../data_files/crime_and_communities_data.csv")
print(dim(CC))
```

```
## [1] 1994  125
```

```
y <- CC$ViolentCrimesPerPop
X <- subset(CC, select = -c(ViolentCrimesPerPop))
```

Dataset exploration

In this section, you should provide a thorough exploration of the features of the dataset. Things to keep in mind in this section include:

- Which variables are categorical versus numerical?
- What are the general summary statistics of the data? How can these be visualized?
- Is the data normalized? Should it be normalized?
- Are there missing values in the data? How should these missing values be handled?
- Can the data be well-represented in fewer dimensions?

YOUR CODE GOES HERE

Examine Categorical vs. Quantitative data

Let's look at the structure of the data

```r
str(X)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1994 obs. of  124 variables:
##  $ population          : num  11980 23123 29344 16656 140494 ...
##  $ householdsize       : num  3.1 2.82 2.43 2.4 2.45 2.6 2.45 2.46 2.62 2.54 ...
##  $ racepctblack        : num  1.37 0.8 0.74 1.7 2.51 ...
##  $ racePctWhite        : num  91.8 95.6 94.3 97.3 95.7 ...
##  $ racePctAsian        : num  6.5 3.44 3.43 0.5 0.9 1.47 0.4 1.25 0.92 0.77 ...
##  $ racePctHisp         : num  1.88 0.85 2.35 0.7 0.95 ...
##  $ agePct12t21         : num  12.5 11 11.4 12.6 18.1 ...
##  $ agePct12t29         : num  21.4 21.3 25.9 25.2 32.9 ...
##  $ agePct16t24         : num  10.9 10.5 11 12.2 20 ...
##  $ agePct65up          : num  11.3 17.2 10.3 17.6 13.3 ...
##  $ numbUrban           : num  11980 23123 29344 0 140494 ...
##  $ pctUrban            : num  100 100 100 0 100 100 100 100 100 100 ...
##  $ medIncome           : num  75122 47917 35669 20580 21577 ...
##  $ pctWWage            : num  89.2 79 82 68.2 75.8 ...
##  $ pctWFarmSelf        : num  1.55 1.11 1.15 0.24 1 0.39 0.67 2.93 0.86 1.54 ...
##  $ pctWInvInc          : num  70.2 64.1 55.7 39 41.1 ...
##  $ pctWSocSec          : num  23.6 35.5 22.2 39.5 29.3 ...
##  $ pctWPubAsst         : num  1.03 2.75 2.94 11.71 7.12 ...
##  $ pctWRetire          : num  18.4 22.9 14.6 18.3 14.1 ...
##  $ medFamInc           : num  79584 55323 42112 26501 27705 ...
##  $ perCapInc           : num  29711 20148 16946 10810 11878 ...
##  $ whitePerCap         : num  30233 20191 17103 10909 12029 ...
##  $ blackPerCap         : num  13600 18137 16644 9984 7382 ...
##  $ indianPerCap        : num  5725 0 21606 4941 10264 ...
##  $ AsianPerCap         : num  27101 20074 15528 3541 10753 ...
##  $ OtherPerCap         : num  5115 5250 5954 2451 7192 ...
##  $ HispPerCap          : num  22838 12222 8405 4391 8104 ...
##  $ NumUnderPov         : num  227 885 1389 2831 23223 ...
##  $ PctPopUnderPov      : num  1.96 3.98 4.75 17.23 17.78 ...
##  $ PctLess9thGrade     : num  5.81 5.61 2.8 11.05 8.76 ...
##  $ PctNotHSGrad        : num  9.9 13.72 9.09 33.68 23.03 ...
##  $ PctBSorMore         : num  48.2 29.9 30.1 10.8 20.7 ...
##  $ PctUnemployed       : num  2.7 2.43 4.01 9.86 5.72 4.85 8.19 4.18 8.39 7.19 ...
##  $ PctEmploy           : num  64.5 62 69.8 54.7 59 ...
##  $ PctEmplManu         : num  14.7 12.3 15.9 31.2 14.3 ...
##  $ PctEmplProfServ     : num  28.8 29.3 21.5 27.4 26.8 ...
##  $ PctOccupManu        : num  5.49 6.39 8.79 26.76 14.72 ...
##  $ PctOccupMgmtProf    : num  50.7 37.6 32.5 22.7 23.4 ...
##  $ MalePctDivorce      : num  3.67 4.23 10.1 10.98 11.4 ...
##  $ MalePctNevMarr      : num  26.4 28 25.8 28.1 33.3 ...
##  $ FemalePctDiv        : num  5.22 6.45 14.76 14.47 14.46 ...
##  $ TotalPctDiv         : num  4.47 5.42 12.55 12.91 13.04 ...
##  $ PersPerFam          : num  3.22 3.11 2.95 2.98 2.89 3.14 2.95 3 3.11 2.99 ...
##  $ PctFam2Par          : num  91.4 86.9 78.5 64 71.9 ...
##  $ PctKids2Par         : num  90.2 85.3 78.8 62.4 69.8 ...
##  $ PctYoungKids2Par    : num  95.8 96.8 92.4 65.4 79.8 ...
##  $ PctTeen2Par         : num  95.8 86.5 75.7 67.4 75.3 ...
##  $ PctWorkMomYoungKids : num  44.6 51.1 66.1 59.6 63 ...
```

```
##  $ PctWorkMom           : num  58.9 62.4 74.2 70.3 70.5 ...
##  $ NumKidsBornNeverMar  : num  31 43 164 561 1511 ...
##  $ PctKidsBornNeverMar  : num  0.36 0.24 0.88 3.84 1.58 1.18 4.66 1.64 4.71 2.47 ...
##  $ NumImmig             : num  1277 1920 1468 339 2091 ...
##  $ PctImmigRecent       : num  8.69 5.21 16.42 13.86 21.33 ...
##  $ PctImmigRec5         : num  13 8.65 23.98 13.86 30.56 ...
##  $ PctImmigRec8         : num  21 13.3 32.1 15.3 38 ...
##  $ PctImmigRec10        : num  30.9 22.5 35.6 15.3 45.5 ...
##  $ PctRecentImmig       : num  0.93 0.43 0.82 0.28 0.32 1.05 0.11 0.47 0.72 0.53 ...
##  $ PctRecImmig5         : num  1.39 0.72 1.2 0.28 0.45 1.49 0.2 0.67 1.07 1.05 ...
##  $ PctRecImmig8         : num  2.24 1.11 1.61 0.31 0.57 2.2 0.25 0.93 1.63 1.66 ...
##  $ PctRecImmig10        : num  3.3 1.87 1.78 0.31 0.68 2.55 0.29 1.07 2.31 1.94 ...
##  $ PctSpeakEnglOnly     : num  85.7 87.8 93.1 95 96.9 ...
##  $ PctNotSpeakEnglWell  : num  1.37 1.81 1.14 0.56 0.6 0.6 0.28 0.43 2.51 0.81 ...
##  $ PctLargHouseFam      : num  4.81 4.25 2.97 3.93 3.08 5.08 3.85 2.59 6.7 3.66 ...
##  $ PctLargHouseOccup    : num  4.17 3.34 2.05 2.56 1.92 3.46 2.55 1.54 4.1 2.51 ...
##  $ PersPerOccupHous     : num  2.99 2.7 2.42 2.37 2.28 2.55 2.36 2.32 2.45 2.42 ...
##  $ PersPerOwnOccHous    : num  3 2.83 2.69 2.51 2.37 2.89 2.42 2.77 2.47 2.5 ...
##  $ PersPerRentOccHous   : num  2.84 1.96 2.06 2.2 2.16 2.09 2.27 1.91 2.44 2.31 ...
##  $ PctPersOwnOccup      : num  91.5 89 64.2 58.2 57.8 ...
##  $ PctPersDenseHous     : num  0.39 1.01 2.03 1.21 2.11 1.47 1.9 1.67 6.14 3.41 ...
##  $ PctHousLess3BR       : num  11.1 23.6 47.5 45.7 53.2 ...
##  $ MedNumBR             : num  3 3 3 3 2 3 2 2 2 2 ...
##  $ HousVacant           : num  64 240 544 669 5119 ...
##  $ PctHousOccup         : num  98.4 97.2 95.7 91.2 91.8 ...
##  $ PctHousOwnOcc        : num  91 84.9 57.8 54.9 55.5 ...
##  $ PctVacantBoarded     : num  3.12 0 0.92 2.54 2.09 1.41 6.39 0.45 5.64 2.77 ...
##  $ PctVacMore6Mos       : num  37.5 18.33 7.54 57.85 26.22 ...
##  $ MedYrHousBuilt       : num  1959 1958 1976 1939 1966 ...
##  $ PctHousNoPhone       : num  0 0.31 1.55 7 6.13 ...
##  $ PctWOFullPlumb       : num  0.28 0.14 0.12 0.87 0.31 0.28 0.49 0.19 0.33 0.3 ...
##  $ OwnOccLowQuart       : num  215900 136300 74700 36400 37700 ...
##  $ OwnOccMedVal         : num  262600 164200 90400 49600 53900 ...
##  $ OwnOccHiQuart        : num  326900 199900 112000 66500 73100 ...
##  $ OwnOccQrange         : num  111000 63600 37300 30100 35400 60400 26100 39200 38800 41400 ...
##  $ RentLowQ             : num  685 467 370 195 215 463 186 241 192 234 ...
##  $ RentMedian           : num  1001 560 428 250 280 ...
##  $ RentHighQ            : num  1001 672 520 309 349 ...
##  $ RentQrange           : num  316 205 150 114 134 361 139 146 177 142 ...
##  $ MedRent              : num  1001 627 484 333 340 ...
##  $ MedRentPctHousInc    : num  23.8 27.6 24.1 28.7 26.4 24.4 26.3 25.2 29.6 23.8 ...
##  $ MedOwnCostPctInc     : num  21.1 20.7 21.7 20.6 17.3 20.8 15.1 20.7 19.4 17.1 ...
##  $ MedOwnCostPctIncNoMtg: num  14 12.5 11.6 14.5 11.7 12.5 12.2 12.8 13 12.9 ...
##  $ NumInShelters        : num  11 0 16 0 327 0 21 125 43 1 ...
##  $ NumStreet            : num  0 0 0 0 4 0 0 15 4 0 ...
##  $ PctForeignBorn       : num  10.66 8.3 5 2.04 1.49 ...
##  $ PctBornSameState     : num  53.7 77.2 44.8 88.7 64.3 ...
##  $ PctSameHouse85       : num  65.3 71.3 36.6 56.7 42.3 ...
##  $ PctSameCity85        : num  78.1 90.2 61.3 90.2 70.6 ...
##  $ PctSameState85       : num  89.1 96.1 82.8 96.2 85.7 ...
##  $ LemasSwornFT         : num  NA NA NA NA NA NA NA NA 198 NA ...
##   [list output truncated]
```

The structure of the data is partialy ommited due to the high number of features. Let's try getting the class of each feature

```r
apply(X = X, MARGIN = 2, FUN = class)
```

```
##            population         householdsize          racepctblack
##             "numeric"             "numeric"             "numeric"
##          racePctWhite          racePctAsian           racePctHisp
##             "numeric"             "numeric"             "numeric"
##           agePct12t21           agePct12t29           agePct16t24
##             "numeric"             "numeric"             "numeric"
##            agePct65up             numbUrban              pctUrban
##             "numeric"             "numeric"             "numeric"
##             medIncome              pctWWage          pctWFarmSelf
##             "numeric"             "numeric"             "numeric"
##            pctWInvInc            pctWSocSec           pctWPubAsst
##             "numeric"             "numeric"             "numeric"
##            pctWRetire             medFamInc             perCapInc
##             "numeric"             "numeric"             "numeric"
##           whitePerCap           blackPerCap          indianPerCap
##             "numeric"             "numeric"             "numeric"
##           AsianPerCap           OtherPerCap            HispPerCap
##             "numeric"             "numeric"             "numeric"
##           NumUnderPov        PctPopUnderPov       PctLess9thGrade
##             "numeric"             "numeric"             "numeric"
##          PctNotHSGrad           PctBSorMore         PctUnemployed
##             "numeric"             "numeric"             "numeric"
##             PctEmploy           PctEmplManu       PctEmplProfServ
##             "numeric"             "numeric"             "numeric"
##          PctOccupManu      PctOccupMgmtProf        MalePctDivorce
##             "numeric"             "numeric"             "numeric"
##        MalePctNevMarr         FemalePctDiv           TotalPctDiv
##             "numeric"             "numeric"             "numeric"
##            PersPerFam            PctFam2Par           PctKids2Par
##             "numeric"             "numeric"             "numeric"
##      PctYoungKids2Par           PctTeen2Par   PctWorkMomYoungKids
##             "numeric"             "numeric"             "numeric"
##            PctWorkMom   NumKidsBornNeverMar   PctKidsBornNeverMar
##             "numeric"             "numeric"             "numeric"
##              NumImmig        PctImmigRecent          PctImmigRec5
##             "numeric"             "numeric"             "numeric"
##          PctImmigRec8         PctImmigRec10        PctRecentImmig
##             "numeric"             "numeric"             "numeric"
##          PctRecImmig5          PctRecImmig8         PctRecImmig10
##             "numeric"             "numeric"             "numeric"
##      PctSpeakEnglOnly   PctNotSpeakEnglWell       PctLargHouseFam
##             "numeric"             "numeric"             "numeric"
##     PctLargHouseOccup      PersPerOccupHous     PersPerOwnOccHous
##             "numeric"             "numeric"             "numeric"
##    PersPerRentOccHous       PctPersOwnOccup       PctPersDenseHous
##             "numeric"             "numeric"             "numeric"
##        PctHousLess3BR              MedNumBR            HousVacant
##             "numeric"             "numeric"             "numeric"
##          PctHousOccup          PctHousOwnOcc      PctVacantBoarded
##             "numeric"             "numeric"             "numeric"
```

```
##         PctVacMore6Mos          MedYrHousBuilt          PctHousNoPhone
##              "numeric"               "numeric"               "numeric"
##         PctWOFullPlumb          OwnOccLowQuart             OwnOccMedVal
##              "numeric"               "numeric"               "numeric"
##          OwnOccHiQuart            OwnOccQrange                RentLowQ
##              "numeric"               "numeric"               "numeric"
##             RentMedian               RentHighQ              RentQrange
##              "numeric"               "numeric"               "numeric"
##                MedRent      MedRentPctHousInc       MedOwnCostPctInc
##              "numeric"               "numeric"               "numeric"
## MedOwnCostPctIncNoMtg           NumInShelters               NumStreet
##              "numeric"               "numeric"               "numeric"
##         PctForeignBorn        PctBornSameState          PctSameHouse85
##              "numeric"               "numeric"               "numeric"
##          PctSameCity85          PctSameState85            LemasSwornFT
##              "numeric"               "numeric"               "numeric"
##        LemasSwFTPerPop      LemasSwFTFieldOps    LemasSwFTFieldPerPop
##              "numeric"               "numeric"               "numeric"
##          LemasTotalReq       LemasTotReqPerPop        PolicReqPerOffic
##              "numeric"               "numeric"               "numeric"
##            PolicPerPop      RacialMatchCommPol          PctPolicWhite
##              "numeric"               "numeric"               "numeric"
##          PctPolicBlack            PctPolicHisp          PctPolicAsian
##              "numeric"               "numeric"               "numeric"
##          PctPolicMinor    OfficAssgnDrugUnits       NumKindsDrugsSeiz
##              "numeric"               "numeric"               "numeric"
##       PolicAveOTWorked                LandArea                 PopDens
##              "numeric"               "numeric"               "numeric"
##         PctUsePubTrans               PolicCars          PolicOperBudg
##              "numeric"               "numeric"               "numeric"
##    LemasPctPolicOnPatr    LemasGangUnitDeploy    LemasPctOfficDrugUn
##              "numeric"               "numeric"               "numeric"
##        PolicBudgPerPop
##              "numeric"
```

Neither str() nor apply(class) shows any factor. Just to be certain, I examine the documentation from the source (UC Irvine): https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized

There exist in the original data the feature of states, county code, and community code, which are catergorical. However, they are not included in the given data. On the other hads, all other quantitative features in the original data are. We can say that the data set is entirely quantitative.

Summary Statistics

How many do I show?

```r
summary(X)
```

```
##    population      householdsize    racepctblack    racePctWhite
##  Min.   :  10005   Min.   :1.600   Min.   : 0.00   Min.   : 2.68
##  1st Qu.:  14359   1st Qu.:2.490   1st Qu.: 0.94   1st Qu.:75.88
##  Median :  22681   Median :2.650   Median : 3.15   Median :89.61
##  Mean   :  52251   Mean   :2.707   Mean   : 9.51   Mean   :83.49
##  3rd Qu.:  43154   3rd Qu.:2.850   3rd Qu.:11.96   3rd Qu.:95.99
##  Max.   :7322564   Max.   :5.280   Max.   :96.67   Max.   :99.63
##
##   racePctAsian      racePctHisp      agePct12t21     agePct12t29
##  Min.   : 0.0300   Min.   : 0.120   Min.   : 4.58   Min.   : 9.38
##  1st Qu.: 0.6125   1st Qu.: 0.920   1st Qu.:12.23   1st Qu.:24.38
##  Median : 1.2400   Median : 2.340   Median :13.62   Median :26.77
##  Mean   : 2.7508   Mean   : 8.482   Mean   :14.43   Mean   :27.62
##  3rd Qu.: 2.7375   3rd Qu.: 8.610   3rd Qu.:15.39   3rd Qu.:29.18
##  Max.   :57.4600   Max.   :95.290   Max.   :54.40   Max.   :70.51
##
##   agePct16t24     agePct65up      numbUrban          pctUrban
##  Min.   : 4.64   Min.   : 1.660   Min.   :      0   Min.   :  0.00
##  1st Qu.:11.34   1st Qu.: 8.922   1st Qu.:      0   1st Qu.:  0.00
##  Median :12.54   Median :11.855   Median :  17348   Median :100.00
##  Mean   :13.99   Mean   :12.005   Mean   :  46672   Mean   : 69.62
##  3rd Qu.:14.36   3rd Qu.:14.547   3rd Qu.:  41932   3rd Qu.:100.00
##  Max.   :63.62   Max.   :52.770   Max.   :7322564   Max.   :100.00
##
##    medIncome         pctWWage      pctWFarmSelf      pctWInvInc
##  Min.   : 11576   Min.   :31.68   Min.   :0.0000   Min.   : 7.91
##  1st Qu.: 23597   1st Qu.:73.22   1st Qu.:0.4700   1st Qu.:34.19
##  Median : 30896   Median :78.38   Median :0.7000   Median :42.38
##  Mean   : 33699   Mean   :78.08   Mean   :0.8933   Mean   :43.36
##  3rd Qu.: 41215   3rd Qu.:83.70   3rd Qu.:1.1100   3rd Qu.:52.07
##  Max.   :123625   Max.   :96.62   Max.   :6.5300   Max.   :89.04
##
##    pctWSocSec     pctWPubAsst       pctWRetire      medFamInc
##  Min.   : 4.81   Min.   : 0.500   Min.   : 3.46   Min.   : 13785
##  1st Qu.:20.98   1st Qu.: 3.362   1st Qu.:12.99   1st Qu.: 29307
##  Median :26.79   Median : 5.720   Median :15.66   Median : 36010
##  Mean   :26.66   Mean   : 6.806   Mean   :16.06   Mean   : 39553
##  3rd Qu.:31.84   3rd Qu.: 9.150   3rd Qu.:18.78   3rd Qu.: 46683
##  Max.   :76.39   Max.   :26.920   Max.   :45.51   Max.   :131315
##
##    perCapInc      whitePerCap     blackPerCap      indianPerCap
##  Min.   : 5237   Min.   : 5472   Min.   :     0   Min.   :     0
##  1st Qu.:11548   1st Qu.:12596   1st Qu.:  6706   1st Qu.:  6336
##  Median :13977   Median :15028   Median :  9664   Median :  9834
##  Mean   :15522   Mean   :16535   Mean   : 11472   Mean   : 12257
##  3rd Qu.:17774   3rd Qu.:18610   3rd Qu.: 14464   3rd Qu.: 14690
##  Max.   :63302   Max.   :68850   Max.   :212120   Max.   :480000
##
##   AsianPerCap      OtherPerCap       HispPerCap      NumUnderPov
```

```
##  Min.   :     0   Min.   :     0   Min.   :    0   Min.   :     78.0
##  1st Qu.:  8441   1st Qu.:  5500   1st Qu.: 7253   1st Qu.:    936.2
##  Median : 12331   Median :  8144   Median : 9676   Median :   2217.5
##  Mean   : 14284   Mean   :  9375   Mean   :10989   Mean   :   7398.4
##  3rd Qu.: 17346   3rd Qu.: 11378   3rd Qu.:13360   3rd Qu.:   5097.5
##  Max.   :106165   Max.   :137000   Max.   :54648   Max.   :1384994.0
##                   NA's   :1
##  PctPopUnderPov   PctLess9thGrade   PctNotHSGrad    PctBSorMore
##  Min.   : 0.640   Min.   : 0.200   Min.   : 2.09   Min.   : 1.63
##  1st Qu.: 4.692   1st Qu.: 4.770   1st Qu.:14.20   1st Qu.:14.09
##  Median : 9.650   Median : 7.920   Median :21.66   Median :19.62
##  Mean   :11.796   Mean   : 9.444   Mean   :22.70   Mean   :22.99
##  3rd Qu.:17.078   3rd Qu.:12.245   3rd Qu.:29.66   3rd Qu.:28.93
##  Max.   :48.820   Max.   :49.890   Max.   :73.66   Max.   :73.63
##
##  PctUnemployed      PctEmploy      PctEmplManu    PctEmplProfServ
##  Min.   : 1.320   Min.   :24.82   Min.   : 2.05   Min.   : 8.69
##  1st Qu.: 4.090   1st Qu.:56.35   1st Qu.:11.94   1st Qu.:20.11
##  Median : 5.485   Median :62.27   Median :16.66   Median :23.41
##  Mean   : 6.024   Mean   :61.78   Mean   :17.79   Mean   :24.58
##  3rd Qu.: 7.430   3rd Qu.:67.50   3rd Qu.:22.75   3rd Qu.:27.63
##  Max.   :23.830   Max.   :84.67   Max.   :50.03   Max.   :62.67
##
##   PctOccupManu    PctOccupMgmtProf MalePctDivorce   MalePctNevMarr
##  Min.   : 1.370   Min.   : 6.48    Min.   : 2.130   Min.   :12.06
##  1st Qu.: 9.072   1st Qu.:21.92    1st Qu.: 7.162   1st Qu.:25.41
##  Median :13.040   Median :26.30    Median : 9.240   Median :29.00
##  Mean   :13.747   Mean   :28.25    Mean   : 9.180   Mean   :30.67
##  3rd Qu.:17.465   3rd Qu.:32.89    3rd Qu.:11.110   3rd Qu.:33.47
##  Max.   :44.270   Max.   :64.97    Max.   :19.090   Max.   :76.32
##
##   FemalePctDiv    TotalPctDiv      PersPerFam      PctFam2Par
##  Min.   : 3.35   Min.   : 2.83   Min.   :2.290   Min.   :32.24
##  1st Qu.: 9.94   1st Qu.: 8.64   1st Qu.:2.990   1st Qu.:67.67
##  Median :12.63   Median :11.04   Median :3.095   Median :74.77
##  Mean   :12.40   Mean   :10.88   Mean   :3.129   Mean   :73.90
##  3rd Qu.:14.80   3rd Qu.:13.06   3rd Qu.:3.220   3rd Qu.:81.64
##  Max.   :23.46   Max.   :19.11   Max.   :4.640   Max.   :93.60
##
##   PctKids2Par    PctYoungKids2Par  PctTeen2Par    PctWorkMomYoungKids
##  Min.   :26.11   Min.   : 27.43   Min.   :30.64   Min.   :24.42
##  1st Qu.:63.62   1st Qu.: 74.42   1st Qu.:69.92   1st Qu.:55.45
##  Median :72.06   Median : 83.77   Median :76.67   Median :60.70
##  Mean   :70.91   Mean   : 81.75   Mean   :75.34   Mean   :60.43
##  3rd Qu.:79.82   3rd Qu.: 91.44   3rd Qu.:82.52   3rd Qu.:65.80
##  Max.   :92.58   Max.   :100.00   Max.   :97.34   Max.   :87.97
##
##    PctWorkMom    NumKidsBornNeverMar PctKidsBornNeverMar    NumImmig
##  Min.   :41.95   Min.   :      0.0    Min.   : 0.000      Min.   :      20
##  1st Qu.:64.96   1st Qu.:    146.2    1st Qu.: 1.083      1st Qu.:     407
##  Median :69.25   Median :    361.0    Median : 2.080      Median :    1040
##  Mean   :68.80   Mean   :   2041.5    Mean   : 3.140      Mean   :    6314
##  3rd Qu.:73.34   3rd Qu.:   1070.2    3rd Qu.: 3.980      3rd Qu.:    3389
##  Max.   :89.37   Max.   :527557.0    Max.   :24.190      Max.   :2082931
```

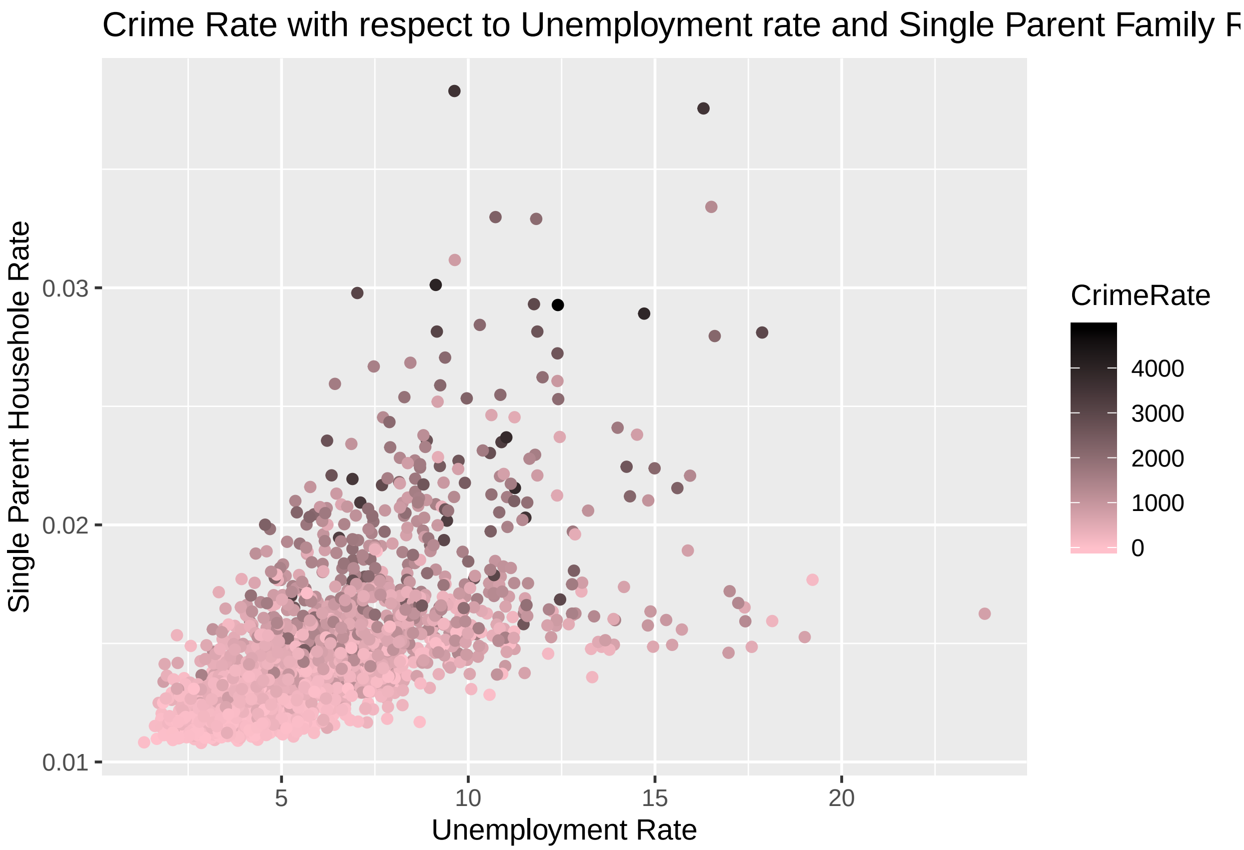
```
##
##  PctImmigRecent    PctImmigRec5    PctImmigRec8    PctImmigRec10
##  Min.   : 0.000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 6.942   1st Qu.:11.70   1st Qu.:17.91   1st Qu.:23.54
##  Median :12.440   Median :19.64   Median :27.46   Median :35.58
##  Mean   :13.734   Mean   :20.83   Mean   :28.12   Mean   :35.48
##  3rd Qu.:18.090   3rd Qu.:27.69   3rd Qu.:37.07   3rd Qu.:46.81
##  Max.   :64.290   Max.   :76.16   Max.   :80.81   Max.   :88.00
##
##  PctRecentImmig    PctRecImmig5     PctRecImmig8    PctRecImmig10
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.180   1st Qu.: 0.290   1st Qu.: 0.410   1st Qu.: 0.540
##  Median : 0.530   Median : 0.780   Median : 1.080   Median : 1.380
##  Mean   : 1.149   Mean   : 1.781   Mean   : 2.424   Mean   : 3.094
##  3rd Qu.: 1.370   3rd Qu.: 2.180   3rd Qu.: 2.870   3rd Qu.: 3.680
##  Max.   :13.710   Max.   :19.930   Max.   :25.340   Max.   :32.630
##
##  PctSpeakEnglOnly PctNotSpeakEnglWell PctLargHouseFam  PctLargHouseOccup
##  Min.   : 6.15    Min.   : 0.000      Min.   : 0.960   Min.   : 0.440
##  1st Qu.:83.70    1st Qu.: 0.510      1st Qu.: 3.390   1st Qu.: 2.360
##  Median :91.78    Median : 0.955      Median : 4.290   Median : 3.050
##  Mean   :86.55    Mean   : 2.538      Mean   : 5.465   Mean   : 3.975
##  3rd Qu.:95.41    3rd Qu.: 2.467      3rd Qu.: 5.957   3rd Qu.: 4.280
##  Max.   :98.98    Max.   :38.330      Max.   :34.870   Max.   :30.870
##
##  PersPerOccupHous PersPerOwnOccHous PersPerRentOccHous PctPersOwnOccup
##  Min.   :1.580    Min.   :1.610     Min.   :1.580      Min.   :13.93
##  1st Qu.:2.400    1st Qu.:2.540     1st Qu.:2.120      1st Qu.:56.56
##  Median :2.560    Median :2.700     Median :2.290      Median :64.99
##  Mean   :2.614    Mean   :2.734     Mean   :2.382      Mean   :65.50
##  3rd Qu.:2.770    3rd Qu.:2.890     3rd Qu.:2.540      3rd Qu.:75.30
##  Max.   :4.520    Max.   :4.480     Max.   :4.730      Max.   :96.59
##
##  PctPersDenseHous PctHousLess3BR     MedNumBR       HousVacant
##  Min.   : 0.050   Min.   : 3.06   Min.   :1.000   Min.   :    36.0
##  1st Qu.: 1.300   1st Qu.:37.93   1st Qu.:2.000   1st Qu.:   310.0
##  Median : 2.470   Median :46.78   Median :3.000   Median :   582.5
##  Mean   : 4.325   Mean   :45.84   Mean   :2.626   Mean   :  1733.0
##  3rd Qu.: 4.920   3rd Qu.:54.09   3rd Qu.:3.000   3rd Qu.:  1280.5
##  Max.   :59.490   Max.   :95.34   Max.   :4.000   Max.   :172768.0
##
##   PctHousOccup    PctHousOwnOcc   PctVacantBoarded PctVacMore6Mos
##  Min.   :37.47   Min.   :16.86   Min.   : 0.000   Min.   : 3.12
##  1st Qu.:90.98   1st Qu.:54.09   1st Qu.: 0.780   1st Qu.:24.74
##  Median :93.98   Median :62.08   Median : 1.740   Median :34.52
##  Mean   :92.71   Mean   :62.63   Mean   : 2.791   Mean   :35.15
##  3rd Qu.:95.91   3rd Qu.:71.59   3rd Qu.: 3.520   3rd Qu.:44.26
##  Max.   :99.00   Max.   :96.36   Max.   :39.890   Max.   :82.13
##
##  MedYrHousBuilt PctHousNoPhone   PctWOFullPlumb   OwnOccLowQuart
##  Min.   :1939   Min.   : 0.000   Min.   :0.0000   Min.   : 15700
##  1st Qu.:1956   1st Qu.: 0.980   1st Qu.:0.1800   1st Qu.: 41800
##  Median :1964   Median : 3.090   Median :0.3300   Median : 65900
##  Mean   :1963   Mean   : 4.446   Mean   :0.4377   Mean   : 91116
```

```
##  3rd Qu.:1971   3rd Qu.: 7.080   3rd Qu.:0.5700   3rd Qu.:126800
##  Max.   :1987   Max.   :23.630   Max.   :5.3300   Max.   :500001
##
##   OwnOccMedVal    OwnOccHiQuart     OwnOccQrange       RentLowQ
##  Min.   : 26600   Min.   : 36700   Min.   :     0   Min.   :  99.0
##  1st Qu.: 56700   1st Qu.: 74800   1st Qu.: 32925   1st Qu.: 210.0
##  Median : 84600   Median :109500   Median : 44250   Median : 305.0
##  Mean   :116102   Mean   :149007   Mean   : 57891   Mean   : 328.1
##  3rd Qu.:156250   3rd Qu.:192850   3rd Qu.: 67475   3rd Qu.: 420.0
##  Max.   :500001   Max.   :500001   Max.   :331000   Max.   :1001.0
##
##    RentMedian       RentHighQ        RentQrange       MedRent
##  Min.   : 120.0   Min.   : 182.0   Min.   :  0.0   Min.   : 192.0
##  1st Qu.: 286.0   1st Qu.: 361.2   1st Qu.:139.0   1st Qu.: 363.0
##  Median : 394.0   Median : 484.0   Median :173.0   Median : 467.0
##  Mean   : 428.4   Mean   : 528.4   Mean   :200.3   Mean   : 502.7
##  3rd Qu.: 547.8   3rd Qu.: 667.8   3rd Qu.:241.0   3rd Qu.: 621.0
##  Max.   :1001.0   Max.   :1001.0   Max.   :803.0   Max.   :1001.0
##
##  MedRentPctHousInc MedOwnCostPctInc MedOwnCostPctIncNoMtg
##  Min.   :14.90     Min.   :14.10    Min.   :10.10
##  1st Qu.:24.30     1st Qu.:19.10    1st Qu.:11.90
##  Median :26.20     Median :21.20    Median :12.80
##  Mean   :26.33     Mean   :21.21    Mean   :13.03
##  3rd Qu.:28.10     3rd Qu.:23.30    3rd Qu.:13.80
##  Max.   :35.10     Max.   :32.70    Max.   :23.40
##
##  NumInShelters        NumStreet        PctForeignBorn   PctBornSameState
##  Min.   :    0.00   Min.   :    0.00   Min.   : 0.180   Min.   : 6.75
##  1st Qu.:    0.00   1st Qu.:    0.00   1st Qu.: 2.080   1st Qu.:48.87
##  Median :    0.00   Median :    0.00   Median : 4.490   Median :62.52
##  Mean   :   67.72   Mean   :   18.71   Mean   : 7.606   Mean   :60.50
##  3rd Qu.:   24.00   3rd Qu.:    1.00   3rd Qu.: 9.585   3rd Qu.:74.38
##  Max.   :23383.00   Max.   :10447.00   Max.   :60.400   Max.   :93.14
##
##  PctSameHouse85  PctSameCity85   PctSameState85   LemasSwornFT
##  Min.   :11.83   Min.   :27.95   Min.   :32.83   Min.   :   65.0
##  1st Qu.:44.68   1st Qu.:71.92   1st Qu.:84.73   1st Qu.:  131.0
##  Median :51.87   Median :79.31   Median :89.64   Median :  173.0
##  Mean   :51.32   Mean   :77.11   Mean   :87.73   Mean   :  458.7
##  3rd Qu.:58.51   3rd Qu.:84.70   3rd Qu.:92.73   3rd Qu.:  314.0
##  Max.   :78.56   Max.   :96.59   Max.   :99.90   Max.   :25655.0
##                                                  NA's   :1675
##  LemasSwFTPerPop  LemasSwFTFieldOps LemasSwFTFieldPerPop LemasTotalReq
##  Min.   :  29.4   Min.   :   14.0   Min.   :  19.21      Min.   :   8100
##  1st Qu.: 149.1   1st Qu.:  113.5   1st Qu.: 130.43      1st Qu.:  49864
##  Median : 196.0   Median :  152.0   Median : 170.16      Median :  89205
##  Mean   : 248.1   Mean   :  395.9   Mean   : 211.32      Mean   : 240510
##  3rd Qu.: 260.8   3rd Qu.:  283.0   3rd Qu.: 226.81      3rd Qu.: 174171
##  Max.   :3437.2   Max.   :22496.0   Max.   :3290.62      Max.   :8328470
##  NA's   :1675     NA's   :1675      NA's   :1675         NA's   :1675
##  LemasTotReqPerPop PolicReqPerOffic  PolicPerPop     RacialMatchCommPol
##  Min.   :   2705   Min.   :  41.4   Min.   :  29.4   Min.   : 42.15
##  1st Qu.:  65486   1st Qu.: 342.9   1st Qu.: 149.2   1st Qu.: 79.44
```

```
##  Median :  91035   Median : 444.8   Median : 196.0   Median : 87.95
##  Mean   : 122280   Mean   : 526.8   Mean   : 248.1   Mean   : 85.49
##  3rd Qu.: 131894   3rd Qu.: 646.0   3rd Qu.: 260.8   3rd Qu.: 93.62
##  Max.   :1926282   Max.   :2162.5   Max.   :3437.2   Max.   :100.00
##  NA's   :1675      NA's   :1675     NA's   :1675     NA's   :1675
##  PctPolicWhite    PctPolicBlack     PctPolicHisp    PctPolicAsian
##  Min.   :  1.60   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.: 76.36   1st Qu.: 2.055   1st Qu.: 0.450   1st Qu.: 0.0000
##  Median : 86.18   Median : 4.840   Median : 2.110   Median : 0.0000
##  Mean   : 82.53   Mean   : 8.983   Mean   : 5.683   Mean   : 0.7088
##  3rd Qu.: 93.09   3rd Qu.:13.355   3rd Qu.: 6.490   3rd Qu.: 0.6650
##  Max.   :100.00   Max.   :67.310   Max.   :98.400   Max.   :18.5700
##  NA's   :1675     NA's   :1675     NA's   :1675     NA's   :1675
##  PctPolicMinor   OfficAssgnDrugUnits NumKindsDrugsSeiz PolicAveOTWorked
##  Min.   : 0.00   Min.   :   0.00     Min.   : 1.000    Min.   :  0.0
##  1st Qu.: 5.05   1st Qu.:   6.00     1st Qu.: 7.000    1st Qu.: 55.1
##  Median :11.39   Median :  12.00     Median : 9.000    Median : 99.0
##  Mean   :15.20   Mean   :  25.87     Mean   : 8.784    Mean   :119.8
##  3rd Qu.:19.68   3rd Qu.:  23.00     3rd Qu.:10.500    3rd Qu.:153.6
##  Max.   :98.40   Max.   :1773.00     Max.   :15.000    Max.   :634.7
##  NA's   :1675    NA's   :1675        NA's   :1675      NA's   :1675
##     LandArea          PopDens      PctUsePubTrans     PolicCars
##  Min.   :   0.90   Min.   :   10   Min.   : 0.000   Min.   :  20.0
##  1st Qu.:   7.40   1st Qu.: 1171   1st Qu.: 0.350   1st Qu.:  54.0
##  Median :  13.70   Median : 1996   Median : 1.220   Median :  86.0
##  Mean   :  27.96   Mean   : 2790   Mean   : 3.063   Mean   : 177.3
##  3rd Qu.:  25.77   3rd Qu.: 3270   3rd Qu.: 3.377   3rd Qu.: 191.0
##  Max.   :3569.80   Max.   :44230   Max.   :54.330   Max.   :3187.0
##                                                     NA's   :1675
##  PolicOperBudg       LemasPctPolicOnPatr LemasGangUnitDeploy
##  Min.   :2.380e+06   Min.   :10.85       Min.   : 0.000
##  1st Qu.:7.247e+06   1st Qu.:83.87       1st Qu.: 0.000
##  Median :1.075e+07   Median :89.44       Median : 5.000
##  Mean   :2.896e+07   Mean   :86.77       Mean   : 4.404
##  3rd Qu.:2.047e+07   3rd Qu.:93.06       3rd Qu.:10.000
##  Max.   :1.617e+09   Max.   :99.94       Max.   :10.000
##  NA's   :1675        NA's   :1675        NA's   :1675
##  LemasPctOfficDrugUn PolicBudgPerPop
##  Min.   : 0.00       Min.   :  15260
##  1st Qu.: 0.00       1st Qu.:  86869
##  Median : 0.00       Median : 114582
##  Mean   : 1.01       Mean   : 154590
##  3rd Qu.: 0.00       3rd Qu.: 156961
##  Max.   :48.44       Max.   :2422367
##                      NA's   :1675
```

Visualization

Due to such as massive number of feature, there is no way to visualize data from every feature without dimentionality reduction. In the next coming graphs, we only examine some groups of feature that will hopefully tell us something about the data.


Crime Rate with respect to Unemployment rate and Single Parent Family F
Single Parent Househole Rate
0.03
0.02
0.01
5
10
15
20
Unemployment Rate
CrimeRate
4000
3000
2000
1000
0

Missing Data Processing

check if target y contains missing data

```r
any(is.na(y))
```

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

check if any of the features contains missing data

```r
any(is.na(X))
```

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

Now that we have detected there is NA in some the features, we decide to replace it by the median of other existing data in that corresponding feature

```r
X <- X %>% mutate_all(function(x) ifelse(is.na(x), median(x, na.rm = TRUE), x))
any(is.na(X))
```

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

Data Normalization - Scaling

After the previous step of examination, it is obvious that many features are different in nature. For example, some feautures are Percentage (PctForeignBorn, PctBornSameState). Some are counts (NumInShelters, population). Some are in US Dollars (MedRent, . . .). Each of the features have different range, scale, and unit. Such condition will affect how much each of the feature influence the predition later on .Therefore, it is highly crucial that we normalize the features.

```r
# X <- scale(X)
```

Dimensionality reduction - Principal Component Analysis

Apply PCA

```r
res.pca <- PCA(X = X,graph = F,ncp = 10)
```

Plot Screeplot for Eigenvalues. Due to the very high number of components (125), we only pick out the first 20

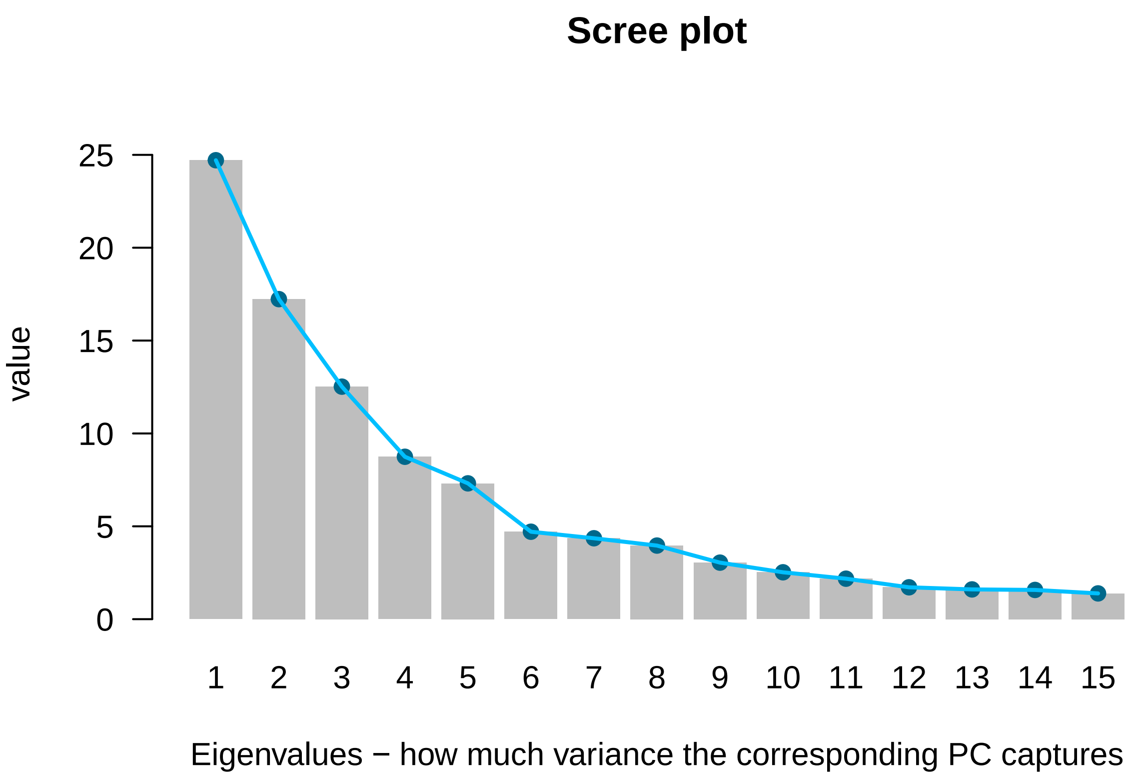
```r
eig <- res.pca$eig
```

Visualize Eigen value

```r
eigvalue <- eig[1:15, "eigenvalue"]

barchart <- barplot(eigvalue, las = 1, border = NA,
                    names.arg = 1:length(eigvalue),
                    ylim = c(0, 1.1 * ceiling(max(eigvalue))),
                    ylab = "value",
                    xlab = "Eigenvalues - how much variance the corresponding PC captures",
                    main = "Scree plot")

points(barchart, eigvalue, pch = 19, col = "deepskyblue4")
lines(barchart, eigvalue, lwd = 2, col = "deepskyblue")
```



As you can see, each of the eigencalue represents the amount of variance in the dataset that was captured by the corresponding PC. Also, let's examine the eigen value result overall

```r
head(eig, n = 35)
```

```
##         eigenvalue percentage of variance
## comp 1  24.7120404             19.9290649
## comp 2  17.2316062             13.8964566
## comp 3  12.5198102             10.0966212
## comp 4   8.7447951              7.0522541
## comp 5   7.3128094              5.8974269
## comp 6   4.7086788              3.7973216
## comp 7   4.3557958              3.5127386
## comp 8   3.9634359              3.1963193
## comp 9   3.0441163              2.4549325
## comp 10  2.5238261              2.0353436
## comp 11  2.1792252              1.7574397
## comp 12  1.7159664              1.3838439
## comp 13  1.6026493              1.2924591
## comp 14  1.5734883              1.2689422
## comp 15  1.3877036              1.1191158
## comp 16  1.3699387              1.1047893
## comp 17  1.1287800              0.9103064
## comp 18  1.1118188              0.8966281
## comp 19  1.0902493              0.8792333
## comp 20  1.0450405              0.8427746
## comp 21  0.9847874              0.7941834
## comp 22  0.9813942              0.7914469
## comp 23  0.9485897              0.7649917
## comp 24  0.9139398              0.7370482
## comp 25  0.8637256              0.6965529
```

```
## comp 26  0.8142463              0.6566503
## comp 27  0.7998861              0.6450695
## comp 28  0.7623629              0.6148088
## comp 29  0.7264818              0.5858724
## comp 30  0.6896531              0.5561718
## comp 31  0.6219795              0.5015964
## comp 32  0.5925124              0.4778326
## comp 33  0.5579743              0.4499792
## comp 34  0.5359479              0.4322160
## comp 35  0.4999603              0.4031938
##          cumulative percentage of variance
## comp 1                           19.92906
## comp 2                           33.82552
## comp 3                           43.92214
## comp 4                           50.97440
## comp 5                           56.87182
## comp 6                           60.66915
## comp 7                           64.18188
## comp 8                           67.37820
## comp 9                           69.83314
## comp 10                          71.86848
## comp 11                          73.62592
## comp 12                          75.00976
## comp 13                          76.30222
## comp 14                          77.57116
## comp 15                          78.69028
## comp 16                          79.79507
## comp 17                          80.70538
## comp 18                          81.60200
## comp 19                          82.48124
## comp 20                          83.32401
## comp 21                          84.11819
## comp 22                          84.90964
## comp 23                          85.67463
## comp 24                          86.41168
## comp 25                          87.10823
## comp 26                          87.76488
## comp 27                          88.40995
## comp 28                          89.02476
## comp 29                          89.61064
## comp 30                          90.16681
## comp 31                          90.66840
## comp 32                          91.14624
## comp 33                          91.59622
## comp 34                          92.02843
## comp 35                          92.43163
```

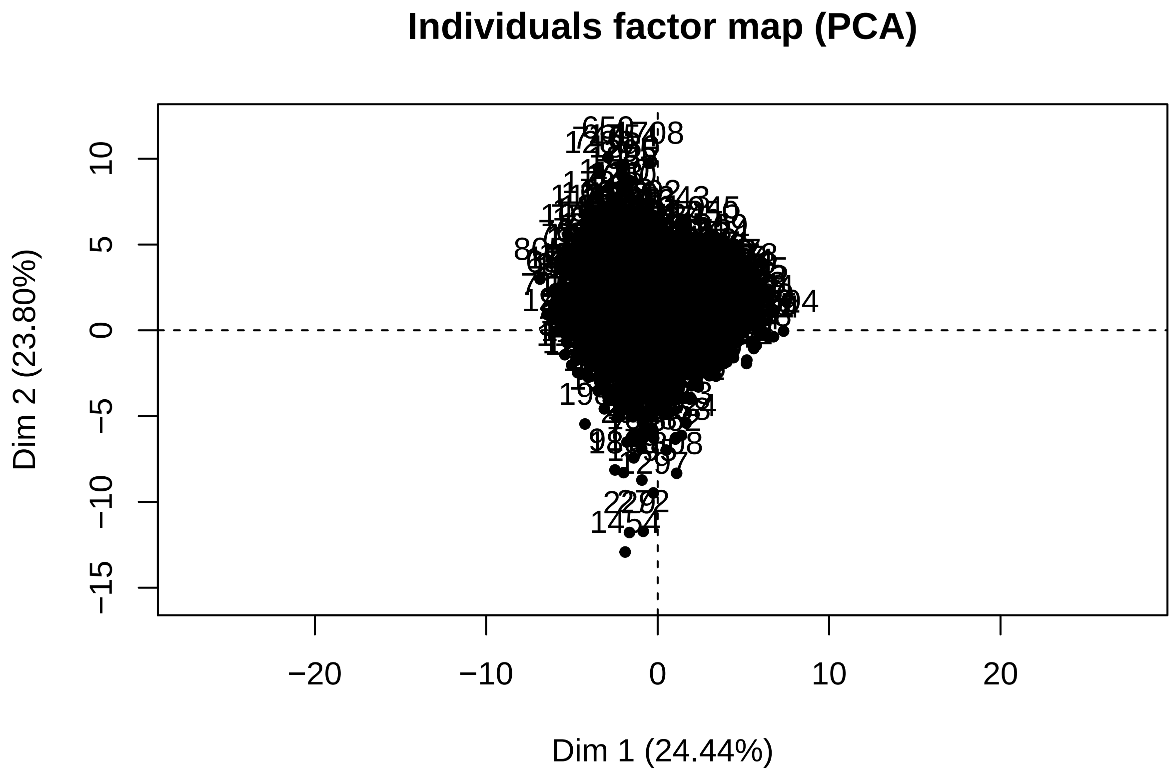
With the given information above we can choose the number of component based on:

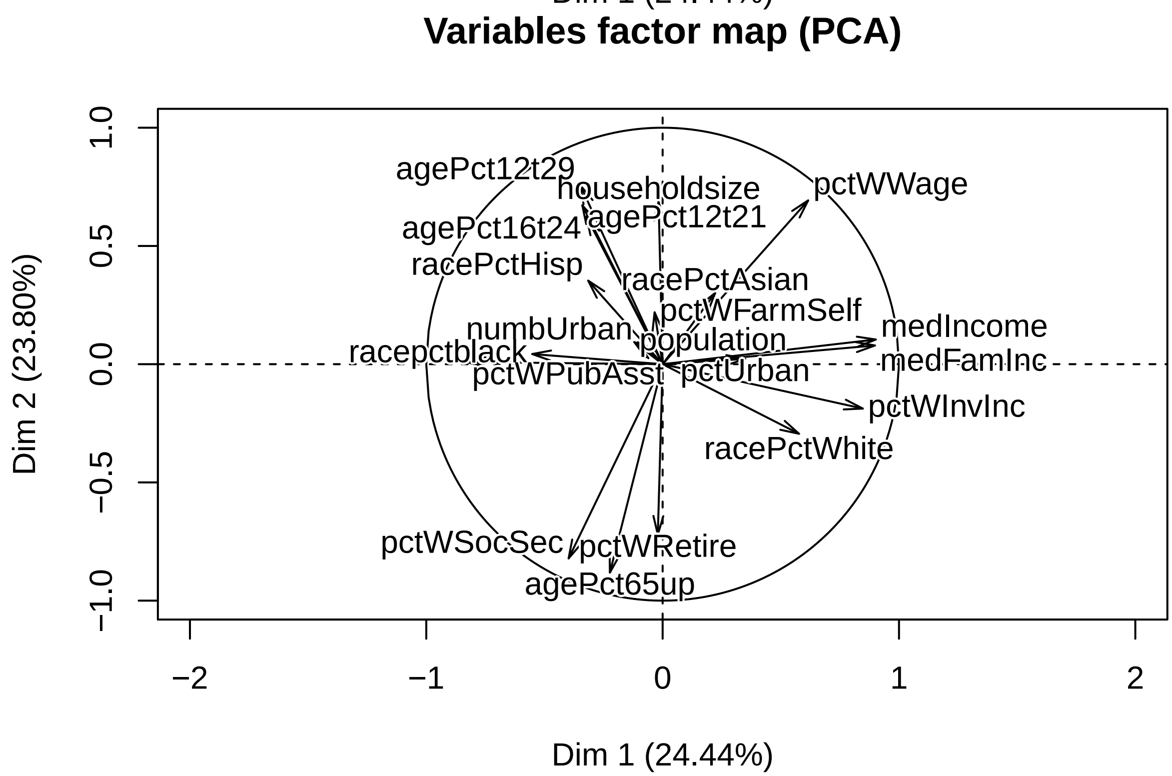
- Elbow method: the first 6 PCs
- Kaiser's rule $\lambda_k > 1$: the first 20 PCs
- Jollie's rule $\lambda_k > 0.7$: the first 30 PCs
- A, if we wish to keep the number of PCs that accumulatively capture 70% of the variance in the data,

we can keep the first 10 PCs

Interpretation of this plot

```r
p <- PCA(X[, 1:20], graph = T)
```





Question for Ryan: Now my plan is that: Keep 10 PCs only and perform regression on PCs instead of doing so on the original data matrix. Which one of the] "coord", "cor", "cos2" ,"contrib" should I use?

Also, do I scale principal component again

```
# head(res.pca$var)
PCs <- res.pca$var
names(PCs) # "coord"   "cor"     "cos2"    "contrib"
```

```
## [1] "coord"   "cor"     "cos2"    "contrib"
```

```
PCs <- res.pca$var[["coord"]]
```

Regression task

In this section, you should use the techniques learned in class to develop a model to predict ViolentCrimesPerPop using the 124 features (or some subset of them) stored in **X**. Remember that you should try several different methods, and use model selection methods to determine which model is best. You should also be sure to keep a held-out test set to evaluate the performance of your model.

YOUR CODE GOES HERE