Crime and Communities

Group Member 1 Name: Ian Driscoll **Group Member 1 SID:** 3031896752

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.

```r
library(readr)
CC <- read_csv("crime_and_communities_data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )

## See spec(...) for full column specifications.
```

```r
print(dim(CC))
```

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

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

Dataset exploration

All of these variables are numerical. The summary statistics are displayed below, and these variables are a mix of percentages, counts, and counts per capita. Due to the mixed nature of these variable types, normalizing the data and mean-centering it will be best in order to factor in all variables equitably. Most of the entries (1675/1994=84%) are missing data for LEMAS predictors, so LEMAS predictors will have to be omitted. Additionally, several of the variables are goal variables, so they will be omitted as well for interest's sake.

```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
```

Data Processing

Due to the LEMAS survey not being applicable to most of the datapoints, we will remove the columns where the 1675 entries have missing data. We are then left with the first 98 columns of the dataset.

```
X = X[,1:98]
X = X[-is.na(X),]
y = y[-is.na(X)]
```

Regression task

We will first split the data 75%/25% and leave the 25% as a test set for model performance comparison later.

```
smp_size <- floor(0.75 * nrow(X))

set.seed(123)
```

```r
train_ind = sample(seq_len(nrow(X)), size = smp_size)

trainX = X[train_ind, ]
testX = X[-train_ind, ]
trainY = y[train_ind]
testY = y[-train_ind]

train = data.frame(trainX, "y"=trainY)
train = na.omit(train)
test = data.frame(testX, "y"=testY)
test = na.omit(test)
```

Our first step will be to build a simple multiple regression model and examine the significance of the various features. We will examine feature importances and select the most relevant/significant ones. Many of these features are hughly correlated, so by including the highest significance features we are eliminating redundancy.

```r
model = lm(y ~ ., data=train)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1442.59  -181.06   -40.15   125.29  2132.30
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.463e+03  3.929e+03   1.390 0.164612
## population            -1.006e-03  2.354e-03  -0.427 0.669266
## householdsize         -1.152e+02  1.375e+02  -0.838 0.402185
## racepctblack           7.608e+00  3.808e+00   1.998 0.045931 *
## racePctWhite           4.220e-01  3.635e+00   0.116 0.907605
## racePctAsian           1.479e+00  6.098e+00   0.243 0.808357
## racePctHisp           -9.677e-01  3.238e+00  -0.299 0.765104
## agePct12t21            1.151e+01  1.512e+01   0.761 0.446763
## agePct12t29           -2.967e+01  1.525e+01  -1.946 0.051879 .
## agePct16t24            1.459e+01  2.186e+01   0.667 0.504782
## agePct65up            -1.720e+01  1.267e+01  -1.358 0.174819
## numbUrban              9.986e-04  2.322e-03   0.430 0.667232
## pctUrban               8.957e-01  5.515e-01   1.624 0.104612
## medIncome             -1.549e-02  8.142e-03  -1.903 0.057286 .
## pctWWage              -1.062e+01  6.988e+00  -1.520 0.128819
## pctWFarmSelf           2.561e+01  1.854e+01   1.381 0.167497
## pctWInvInc            -3.356e+00  3.055e+00  -1.098 0.272184
## pctWSocSec             1.062e+01  6.812e+00   1.558 0.119365
## pctWPubAsst            1.685e+01  6.945e+00   2.427 0.015364 *
## pctWRetire            -1.636e+01  4.378e+00  -3.736 0.000194 ***
## medFamInc              1.341e-02  8.077e-03   1.660 0.097132 .
## perCapInc             -5.905e-03  1.845e-02  -0.320 0.748995
## whitePerCap           -3.046e-03  1.470e-02  -0.207 0.835861
```

```
## blackPerCap          -7.613e-04  1.193e-03  -0.638 0.523579    
## indianPerCap         -1.965e-04  5.878e-04  -0.334 0.738230    
## AsianPerCap           2.256e-03  1.145e-03   1.971 0.048966 *  
## OtherPerCap           2.166e-03  1.316e-03   1.646 0.099998 .  
## HispPerCap            1.355e-03  2.311e-03   0.586 0.557759    
## NumUnderPov          -9.384e-04  2.711e-03  -0.346 0.729290    
## PctPopUnderPov       -9.111e+00  5.596e+00  -1.628 0.103731    
## PctLess9thGrade      -1.845e+01  7.399e+00  -2.494 0.012745 *  
## PctNotHSGrad          5.538e+00  5.772e+00   0.959 0.337503    
## PctBSorMore          -3.440e-01  4.148e+00  -0.083 0.933907    
## PctUnemployed        -4.548e+00  9.467e+00  -0.480 0.631008    
## PctEmploy             7.903e+00  6.068e+00   1.302 0.193023    
## PctEmplManu          -5.043e+00  2.524e+00  -1.998 0.045874 *  
## PctEmplProfServ      -9.225e-01  3.426e+00  -0.269 0.787769    
## PctOccupManu          4.498e+00  5.436e+00   0.827 0.408174    
## PctOccupMgmtProf      7.046e+00  5.729e+00   1.230 0.218994    
## MalePctDivorce        1.479e+02  8.140e+01   1.817 0.069408 .  
## MalePctNevMarr        7.221e-01  6.031e+00   0.120 0.904703    
## FemalePctDiv          9.204e+01  8.516e+01   1.081 0.279942    
## TotalPctDiv          -2.318e+02  1.645e+02  -1.409 0.159051    
## PersPerFam           -3.007e+02  4.215e+02  -0.714 0.475616    
## PctFam2Par            1.108e+01  1.019e+01   1.087 0.277125    
## PctKids2Par          -2.416e+01  8.492e+00  -2.845 0.004506 ** 
## PctYoungKids2Par      2.955e+00  2.782e+00   1.062 0.288449    
## PctTeen2Par          -5.993e-01  2.596e+00  -0.231 0.817474    
## PctWorkMomYoungKids   6.094e+00  3.259e+00   1.870 0.061710 .  
## PctWorkMom           -1.190e+01  4.734e+00  -2.514 0.012035 *  
## NumKidsBornNeverMar  -6.961e-03  4.494e-03  -1.549 0.121639    
## PctKidsBornNeverMar   4.418e+01  1.122e+01   3.939 8.60e-05 ***
## NumImmig             -1.992e-04  1.189e-03  -0.168 0.866982    
## PctImmigRecent        3.208e+00  2.876e+00   1.116 0.264806    
## PctImmigRec5         -1.876e+00  3.725e+00  -0.504 0.614664    
## PctImmigRec8          3.809e-01  3.491e+00   0.109 0.913134    
## PctImmigRec10         6.210e-01  2.206e+00   0.282 0.778368    
## PctRecentImmig       -1.745e+01  5.793e+01  -0.301 0.763250    
## PctRecImmig5          3.165e+00  7.322e+01   0.043 0.965530    
## PctRecImmig8          2.376e+01  6.637e+01   0.358 0.720340    
## PctRecImmig10        -4.242e+01  3.877e+01  -1.094 0.274177    
## PctSpeakEnglOnly     -1.985e+00  3.861e+00  -0.514 0.607204    
## PctNotSpeakEnglWell  -7.040e+00  1.206e+01  -0.584 0.559432    
## PctLargHouseFam       3.408e+01  3.670e+01   0.929 0.353200    
## PctLargHouseOccup    -4.666e+01  3.953e+01  -1.180 0.238068    
## PersPerOccupHous      5.771e+02  4.951e+02   1.166 0.243997    
## PersPerOwnOccHous     1.038e+02  3.329e+02   0.312 0.755337    
## PersPerRentOccHous   -3.177e+02  1.354e+02  -2.347 0.019087 *  
## PctPersOwnOccup      -3.419e+01  2.018e+01  -1.694 0.090433 .  
## PctPersDenseHous      2.429e+01  8.682e+00   2.798 0.005212 ** 
## PctHousLess3BR        1.283e+00  2.442e+00   0.525 0.599433    
## MedNumBR              7.126e+00  3.095e+01   0.230 0.817938    
## HousVacant            1.557e-02  7.195e-03   2.164 0.030652 *  
## PctHousOccup         -5.614e+00  3.061e+00  -1.834 0.066865 .  
## PctHousOwnOcc         2.940e+01  2.014e+01   1.460 0.144478    
## PctVacantBoarded      1.737e+01  4.258e+00   4.080 4.75e-05 ***
## PctVacMore6Mos       -2.497e+00  1.077e+00  -2.317 0.020631 *  
```

```
## MedYrHousBuilt        -7.909e-01  1.860e+00  -0.425 0.670817
## PctHousNoPhone         1.487e-01  6.572e+00   0.023 0.981955
## PctWOFullPlumb        -1.651e+01  3.036e+01  -0.544 0.586726
## OwnOccLowQuart         1.140e-03  1.484e-03   0.769 0.442286
## OwnOccMedVal          -1.793e-04  1.789e-03  -0.100 0.920180
## OwnOccHiQuart         -9.669e-04  7.705e-04  -1.255 0.209691
## OwnOccQrange                  NA          NA      NA       NA
## RentLowQ              -8.162e-01  3.154e-01  -2.588 0.009756 **
## RentMedian            -3.694e-02  5.717e-01  -0.065 0.948497
## RentHighQ             -2.372e-01  3.296e-01  -0.720 0.471839
## RentQrange                    NA          NA      NA       NA
## MedRent                1.169e+00  5.011e-01   2.332 0.019846 *
## MedRentPctHousInc      1.220e+00  6.048e+00   0.202 0.840214
## MedOwnCostPctInc      -1.785e+00  6.942e+00  -0.257 0.797101
## MedOwnCostPctIncNoMtg -3.480e+01  1.021e+01  -3.409 0.000670 ***
## NumInShelters          1.508e-01  7.862e-02   1.917 0.055385 .
## NumStreet             -1.336e-02  1.666e-01  -0.080 0.936101
## PctForeignBorn         1.471e+01  7.766e+00   1.894 0.058399 .
## PctBornSameState      -5.089e-01  1.597e+00  -0.319 0.750056
## PctSameHouse85        -1.308e+00  3.069e+00  -0.426 0.669940
## PctSameCity85          1.274e+00  2.338e+00   0.545 0.585865
## PctSameState85         1.006e+00  3.717e+00   0.271 0.786653
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361.7 on 1397 degrees of freedom
## Multiple R-squared:  0.679,  Adjusted R-squared:  0.657
## F-statistic: 30.78 on 96 and 1397 DF,  p-value: < 2.2e-16
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```r
importances = data.frame(varImp(model, scale=FALSE))
importances$variable = row.names(importances)

imp = importances[order(importances$Overall, decreasing=TRUE),]

head(imp, 40)
```

```
##                        Overall              variable
## PctVacantBoarded      4.080452      PctVacantBoarded
## PctKidsBornNeverMar   3.938600   PctKidsBornNeverMar
## pctWRetire            3.736076            pctWRetire
## MedOwnCostPctIncNoMtg 3.409151 MedOwnCostPctIncNoMtg
## PctKids2Par           2.845005           PctKids2Par
## PctPersDenseHous      2.798061      PctPersDenseHous
## RentLowQ              2.587910              RentLowQ
## PctWorkMom            2.514401            PctWorkMom
## PctLess9thGrade       2.494044       PctLess9thGrade
```

```
## pctWPubAsst            2.426676           pctWPubAsst
## PersPerRentOccHous     2.346547    PersPerRentOccHous
## MedRent                2.331936               MedRent
## PctVacMore6Mos         2.317314        PctVacMore6Mos
## HousVacant             2.163754            HousVacant
## PctEmplManu            1.998331           PctEmplManu
## racepctblack           1.997806          racepctblack
## AsianPerCap            1.970603           AsianPerCap
## agePct12t29            1.945803           agePct12t29
## NumInShelters          1.917448         NumInShelters
## medIncome              1.902692             medIncome
## PctForeignBorn         1.894248        PctForeignBorn
## PctWorkMomYoungKids    1.869874   PctWorkMomYoungKids
## PctHousOccup           1.834011          PctHousOccup
## MalePctDivorce         1.817149        MalePctDivorce
## PctPersOwnOccup        1.694291       PctPersOwnOccup
## medFamInc              1.660033             medFamInc
## OtherPerCap            1.645954           OtherPerCap
## PctPopUnderPov         1.628092        PctPopUnderPov
## pctUrban               1.623950              pctUrban
## pctWSocSec             1.558400            pctWSocSec
## NumKidsBornNeverMar    1.548871   NumKidsBornNeverMar
## pctWWage               1.519674              pctWWage
## PctHousOwnOcc          1.460136         PctHousOwnOcc
## TotalPctDiv            1.409022           TotalPctDiv
## pctWFarmSelf           1.381009          pctWFarmSelf
## agePct65up             1.357572            agePct65up
## PctEmploy              1.302317             PctEmploy
## OwnOccHiQuart          1.254992         OwnOccHiQuart
## PctOccupMgmtProf       1.229760      PctOccupMgmtProf
## PctLargHouseOccup      1.180334     PctLargHouseOccup
```

We will choose the most important features and build our next regression models, tweaking our feature set when necessary. We will also normalize the features.

```
features = imp$variable[1:35]
train = train[, c(features, "y")]
test = test[, c(features, "y")]
library(BBmisc)
```

```
##
## Attaching package: 'BBmisc'

## The following object is masked from 'package:base':
##
##     isFALSE
```

```
train = data.frame(normalize(train[,features]), "y"=train$y)
test = data.frame(normalize(test[,features]), "y"=test$y)
```

The three different types of regressions we will be using are simple linear regression, random forest regression, and ridge regression. First we will build a Random Forest regression model.

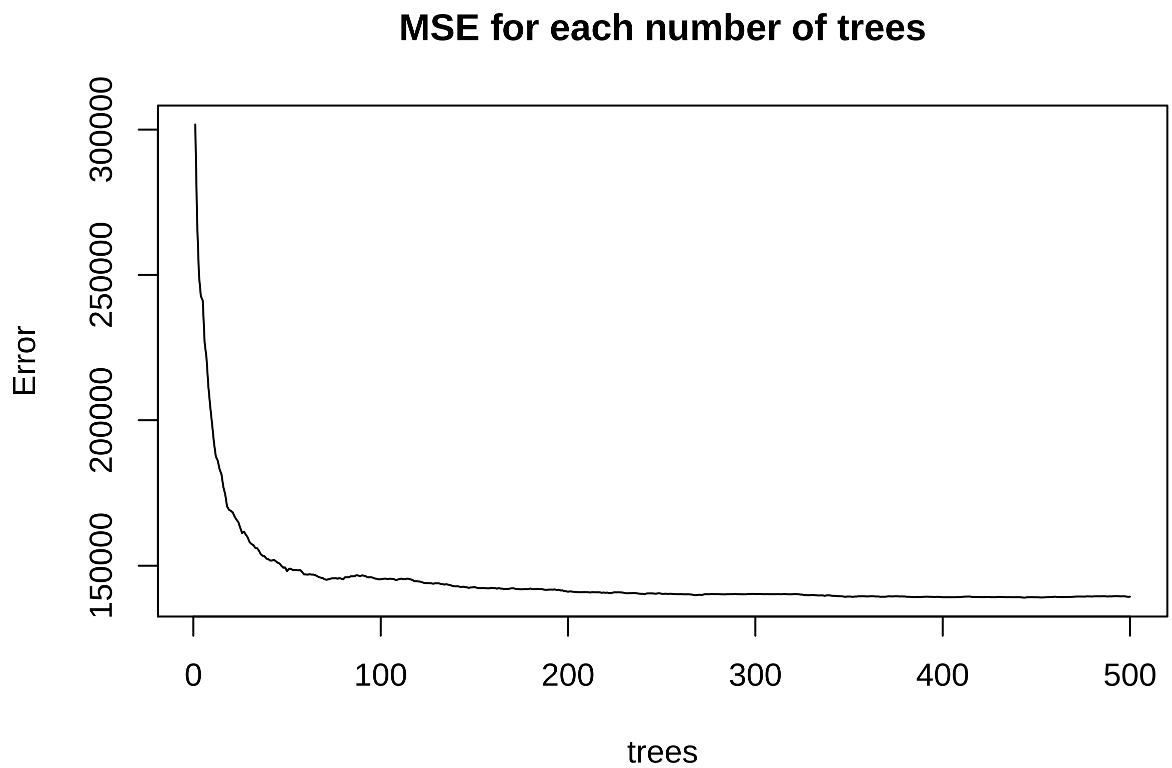
```r
set.seed(123)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
rf = randomForest(
  formula = y~.,
  data    = train
)
rf
```

```
##
## Call:
##  randomForest(formula = y ~ ., data = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 11
##
##           Mean of squared residuals: 139332.1
##                     % Var explained: 63.45
```

```r
plot(rf, main="MSE for each number of trees")
```




```r
which.min(rf$mse)
```

```
## [1] 444
```

We can see that the MSE decreases to a certain point and then is minimized with 339 trees. We will now rebuild our model with the optimal number of trees.

```r
set.seed(123)
rf = randomForest(
  formula = y~.,
  data    = train,
  ntree = 339
)
rf
```

```
##
## Call:
##  randomForest(formula = y ~ ., data = train, ntree = 339)
##                Type of random forest: regression
##                      Number of trees: 339
## No. of variables tried at each split: 11
##
##           Mean of squared residuals: 139802.8
##                     % Var explained: 63.32
```

Our next model will be a ridge regression model. We will use the glmnet package and set alpha to 0 to perform ridge regression. We will test a set of lambda values using cross validation in order to find the optimal lambda for our model.

```
library(glmnet)
```

```
## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-18
```

```
lambda_seq = 10^seq(2, -2, by = -.1)
ridge_cv = cv.glmnet(data.matrix(train[,features]), train$y, alpha = 0, lambda = lambda_seq)
best_lambda = ridge_cv$lambda.min
best_lambda
```

```
## [1] 0.1995262
```

```
ridge = glmnet(data.matrix(train[,features]), train$y, alpha = 0, lambda  = best_lambda)
summary(ridge)
```

```
##           Length Class     Mode
## a0         1     -none-    numeric
## beta      35     dgCMatrix S4
## df         1     -none-    numeric
## dim        2     -none-    numeric
## lambda     1     -none-    numeric
## dev.ratio  1     -none-    numeric
## nulldev    1     -none-    numeric
## npasses    1     -none-    numeric
## jerr       1     -none-    numeric
## offset     1     -none-    logical
## call       5     -none-    call
## nobs       1     -none-    numeric
```

Our final regression model will be a simple multiple regression using our predictor variables.

```
linear = lm(y~., data=train)
summary(linear)
```

```
##
## Call:
## lm(formula = y ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1422.16  -188.72   -41.62   124.75  2108.84
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               599.146      9.339  64.154  < 2e-16 ***
## PctVacantBoarded           64.705     13.621   4.750 2.23e-06 ***
## PctKidsBornNeverMar       130.645     28.914   4.518 6.73e-06 ***
## pctWRetire                -83.530     16.376  -5.101 3.83e-07 ***
## MedOwnCostPctIncNoMtg     -49.327     12.678  -3.891 0.000104 ***
## PctKids2Par              -193.566     45.994  -4.209 2.73e-05 ***
## PctPersDenseHous          115.455     28.416   4.063 5.10e-05 ***
## RentLowQ                 -123.247     36.682  -3.360 0.000800 ***
## PctWorkMom                -61.089     26.848  -2.275 0.023026 *
## PctLess9thGrade           -72.048     22.427  -3.213 0.001344 **
## pctWPubAsst                75.723     24.146   3.136 0.001747 **
## PersPerRentOccHous        -25.266     29.236  -0.864 0.387618
## MedRent                   134.922     40.468   3.334 0.000877 ***
## PctVacMore6Mos            -30.105     13.287  -2.266 0.023607 *
## HousVacant                 80.441     24.126   3.334 0.000877 ***
## PctEmplManu               -28.496     12.185  -2.339 0.019488 *
## racepctblack              101.685     20.249   5.022 5.75e-07 ***
## AsianPerCap                23.001     10.606   2.169 0.030272 *
## agePct12t29               -26.163     23.403  -1.118 0.263784
## NumInShelters              75.638     34.299   2.205 0.027594 *
## medIncome                 -78.555     87.198  -0.901 0.367802
## PctForeignBorn             11.639     21.666   0.537 0.591204
## PctWorkMomYoungKids        30.731     23.967   1.282 0.199970
## PctHousOccup              -18.114     13.046  -1.388 0.165201
## MalePctDivorce            161.401     49.718   3.246 0.001195 **
## PctPersOwnOccup          -186.891    109.864  -1.701 0.089134 .
## medFamInc                  14.860     76.114   0.195 0.845235
## OtherPerCap                27.204     10.045   2.708 0.006845 **
## PctPopUnderPov            -83.488     32.320  -2.583 0.009887 **
## pctUrban                   45.910     12.311   3.729 0.000200 ***
## pctWSocSec                 42.840     39.824   1.076 0.282232
## NumKidsBornNeverMar      -121.045     41.659  -2.906 0.003721 **
## pctWWage                   -6.839     43.370  -0.158 0.874721
## PctHousOwnOcc             194.025    107.417   1.806 0.071081 .
## TotalPctDiv              -148.157     57.152  -2.592 0.009628 **
## pctWFarmSelf                8.712     11.872   0.734 0.463172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361 on 1458 degrees of freedom
## Multiple R-squared:  0.6664, Adjusted R-squared:  0.6584
## F-statistic: 83.21 on 35 and 1458 DF,  p-value: < 2.2e-16
```

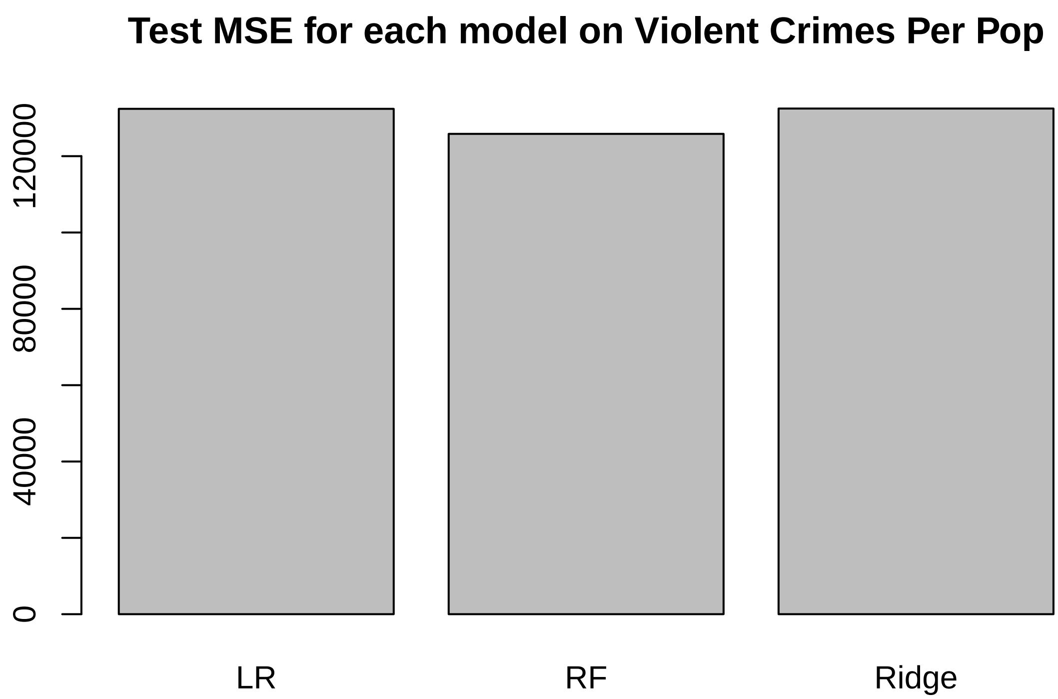
We will now use these three built models and compare their performance on the test set. We will predict using the models, find the residuals, and calculate the Mean Squared Error (MSE) as our performance metric.

```
linear_pred = predict(linear, test[,features])
rf_pred = predict(rf, test[,features])
ridge_pred = predict(ridge, data.matrix(test[,features]))

mse = function(pred, actual) {
  resids = pred - actual
  sqer = resids^2
```

```
  return (mean(sqer))
}

linear_mse = mse(linear_pred, test$y)
rf_mse = mse(rf_pred, test$y)
ridge_mse = mse(ridge_pred, test$y)
mses = c("LR"=linear_mse, "RF"=rf_mse, "Ridge"=ridge_mse)
mses
```

```
##       LR       RF    Ridge
## 132399.6 125837.6 132479.7
```

```
barplot(mses, main="Test MSE for each model on Violent Crimes Per Pop")
```



We can see that our Random Forest regression model performs best out of all 3 models on the test set. The simple linear and ridge regression models performed similarly, but in order to predict violent crimes per population we will choose to use the Random Forest regression model that we built.