Homework 1 - Crime

Due:

* 3 Variable importance plot crimeRf
* 3.1 Variable importance plot Pairs plot
* 3.4 Variable importance plot
* 3.5 Hexbin scatterplot matrix with smoothes
* 3.6 Find different 6 variable model
* 4.5 One 2d proximity-base MDS scatterplot your choice
  + ONE 3d proximity-based MDS plot
* 5 The 3d local variable importance based scatterplot

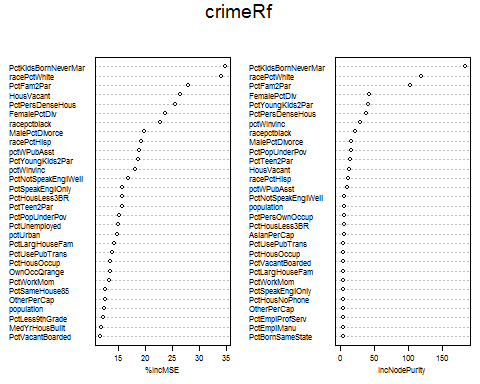
# 3- Predictor variable importance and selection

As described in class, random forests assess variable importance. We can look at a list or a plot of variable importance and conjecture about simple models may do almost as well.

imp <- importance(crimeRf)  
imp

## %IncMSE IncNodePurity  
## population 12.427174 6.6167075  
## householdsize 8.490989 2.8248554  
## racepctblack 22.735479 22.7361603  
## racePctWhite 33.976884 119.4819708  
## racePctAsian 9.572296 3.3987277  
## racePctHisp 19.318857 12.5734111  
## agePct12t21 9.581649 3.6556312  
## agePct12t29 9.590405 3.2991657  
## agePct16t24 11.102081 3.1446706  
## agePct65up 8.775902 2.8420992  
## pctUrban 14.785445 2.1180532  
## medIncome 7.736138 3.8979315  
## pctWWage 9.198961 3.0410587  
## pctWFarmSelf 3.606367 3.7726003  
## pctWInvInc 18.132652 29.9454584  
## pctWSocSec 8.821200 2.8048193  
## pctWPubAsst 18.813439 10.1969047  
## pctWRetire 6.973192 3.3099509  
## perCapInc 9.881499 2.7923750  
## blackPerCap 8.647845 4.3831407  
## indianPerCap 6.868526 3.3210121  
## AsianPerCap 10.906265 5.6262944  
## OtherPerCap 12.580386 4.6241000  
## HispPerCap 9.913765 4.3257954  
## PctPopUnderPov 15.097168 16.3420015  
## PctLess9thGrade 12.162596 3.7729685  
## PctNotHSGrad 10.879421 4.0017246  
## PctUnemployed 14.973120 4.1354495  
## PctEmploy 8.313353 2.9494911  
## PctEmplManu 7.473787 4.5416280  
## PctEmplProfServ 7.625722 4.5708515  
## PctOccupManu 9.876418 3.5791459  
## PctOccupMgmtProf 8.295285 3.4863534  
## MalePctDivorce 19.815245 16.7277197  
## MalePctNevMarr 10.827888 3.7779088  
## FemalePctDiv 23.613579 43.2927365  
## PersPerFam 8.833164 2.2778694  
## PctFam2Par 27.908554 103.6589084  
## PctYoungKids2Par 18.595451 41.4635047  
## PctTeen2Par 15.674344 14.8138985  
## PctWorkMomYoungKids 9.134194 3.9526421  
## PctWorkMom 13.384917 4.9368934  
## PctKidsBornNeverMar 34.742520 184.0611978  
## PctImmigRecent 9.635950 3.4906593  
## PctImmigRec5 6.136891 3.1241701  
## PctImmigRec8 8.101901 3.1363845  
## PctImmigRec10 7.174797 3.3135429  
## PctRecentImmig 11.282982 2.7625591  
## PctSpeakEnglOnly 15.766231 4.8920492  
## PctNotSpeakEnglWell 16.829337 6.9487293  
## PctLargHouseFam 14.165230 5.0105070  
## PersPerOccupHous 9.649052 2.3149140  
## PersPerOwnOccHous 10.307780 2.8682609  
## PersPerRentOccHous 11.024849 3.1954373  
## PctPersOwnOccup 11.333573 6.6054264  
## PctPersDenseHous 25.499634 39.1096878  
## PctHousLess3BR 15.706014 6.4539572  
## MedNumBR 4.881657 0.2742761  
## HousVacant 26.382295 14.1453729  
## PctHousOccup 13.471532 5.2909297  
## PctVacantBoarded 11.595829 5.0550446  
## PctVacMore6Mos 6.950554 3.0476729  
## MedYrHousBuilt 11.894134 3.1299423  
## PctHousNoPhone 10.826541 4.6422152  
## PctWOFullPlumb 9.039217 3.6126685  
## OwnOccMedVal 10.779284 3.0127214  
## OwnOccQrange 13.440671 3.9887981  
## RentMedian 10.008817 3.0546998  
## RentQrange 9.631775 3.9021191  
## MedRentPctHousInc 11.073155 4.2718372  
## MedOwnCostPctInc 6.413286 3.5698125  
## MedOwnCostPctIncNoMtg 9.735719 3.1826024  
## PctForeignBorn 11.327171 4.1732313  
## PctBornSameState 11.364504 4.4795909  
## PctSameHouse85 12.644429 3.0614964  
## PctSameCity85 10.806338 3.7574659  
## PctSameState85 7.299747 3.2325098  
## LandArea 9.856201 4.1636722  
## PopDens 7.032408 4.1864400  
## PctUsePubTrans 13.956981 5.4735941

varImpPlot(crimeRf,cex=.5) #updated for sizing



## 3.1 Reduced model 1

n <- 20 #Corrected from 15  
  
ord1 <- order(imp[, 1],decreasing=TRUE)  
nam1 <- row.names(imp[ord1, ])[1:n]  
nam1

## [1] "PctKidsBornNeverMar" "racePctWhite" "PctFam2Par"   
## [4] "HousVacant" "PctPersDenseHous" "FemalePctDiv"   
## [7] "racepctblack" "MalePctDivorce" "racePctHisp"   
## [10] "pctWPubAsst" "PctYoungKids2Par" "pctWInvInc"   
## [13] "PctNotSpeakEnglWell" "PctSpeakEnglOnly" "PctHousLess3BR"   
## [16] "PctTeen2Par" "PctPopUnderPov" "PctUnemployed"   
## [19] "pctUrban" "PctLargHouseFam"

ord2 <- order(imp[, 2],decreasing=TRUE)  
nam2 <- row.names(imp[ord2,])[1:n]  
nam2

## [1] "PctKidsBornNeverMar" "racePctWhite" "PctFam2Par"   
## [4] "FemalePctDiv" "PctYoungKids2Par" "PctPersDenseHous"   
## [7] "pctWInvInc" "racepctblack" "MalePctDivorce"   
## [10] "PctPopUnderPov" "PctTeen2Par" "HousVacant"   
## [13] "racePctHisp" "pctWPubAsst" "PctNotSpeakEnglWell"  
## [16] "population" "PctPersOwnOccup" "PctHousLess3BR"   
## [19] "AsianPerCap" "PctUsePubTrans"

varNam <- union(nam1,nam2)   
varNam

## [1] "PctKidsBornNeverMar" "racePctWhite" "PctFam2Par"   
## [4] "HousVacant" "PctPersDenseHous" "FemalePctDiv"   
## [7] "racepctblack" "MalePctDivorce" "racePctHisp"   
## [10] "pctWPubAsst" "PctYoungKids2Par" "pctWInvInc"   
## [13] "PctNotSpeakEnglWell" "PctSpeakEnglOnly" "PctHousLess3BR"   
## [16] "PctTeen2Par" "PctPopUnderPov" "PctUnemployed"   
## [19] "pctUrban" "PctLargHouseFam" "population"   
## [22] "PctPersOwnOccup" "AsianPerCap" "PctUsePubTrans"

# same results below  
# commonSubs <- sort(unique(c(ord1,ord2)))  
  
  
# We can check spearman rho rank correlation.  
checkCor <- round( cor(crimeReg[,varNam],  
 method="spearman"),2)  
checkCor

## PctKidsBornNeverMar racePctWhite PctFam2Par HousVacant  
## PctKidsBornNeverMar 1.00 -0.74 -0.85 0.47  
## racePctWhite -0.74 1.00 0.58 -0.41  
## PctFam2Par -0.85 0.58 1.00 -0.50  
## HousVacant 0.47 -0.41 -0.50 1.00  
## PctPersDenseHous 0.64 -0.75 -0.54 0.40  
## FemalePctDiv 0.63 -0.50 -0.78 0.51  
## racepctblack 0.68 -0.78 -0.57 0.42  
## MalePctDivorce 0.59 -0.39 -0.76 0.50  
## racePctHisp 0.23 -0.48 -0.09 0.27  
## pctWPubAsst 0.75 -0.50 -0.79 0.35  
## PctYoungKids2Par -0.83 0.54 0.94 -0.47  
## pctWInvInc -0.72 0.59 0.72 -0.36  
## PctNotSpeakEnglWell 0.24 -0.45 -0.10 0.24  
## PctSpeakEnglOnly -0.11 0.36 -0.07 -0.16  
## PctHousLess3BR 0.59 -0.41 -0.71 0.50  
## PctTeen2Par -0.77 0.57 0.90 -0.48  
## PctPopUnderPov 0.74 -0.54 -0.83 0.42  
## PctUnemployed 0.64 -0.43 -0.68 0.36  
## pctUrban 0.02 -0.13 0.03 0.18  
## PctLargHouseFam 0.46 -0.57 -0.26 0.17  
## population 0.33 -0.37 -0.23 0.74  
## PctPersOwnOccup -0.72 0.56 0.78 -0.45  
## AsianPerCap -0.31 0.12 0.37 -0.11  
## PctUsePubTrans 0.05 -0.11 0.05 0.09  
## PctPersDenseHous FemalePctDiv racepctblack  
## PctKidsBornNeverMar 0.64 0.63 0.68  
## racePctWhite -0.75 -0.50 -0.78  
## PctFam2Par -0.54 -0.78 -0.57  
## HousVacant 0.40 0.51 0.42  
## PctPersDenseHous 1.00 0.54 0.39  
## FemalePctDiv 0.54 1.00 0.46  
## racepctblack 0.39 0.46 1.00  
## MalePctDivorce 0.44 0.92 0.43  
## racePctHisp 0.69 0.21 0.09  
## pctWPubAsst 0.58 0.57 0.36  
## PctYoungKids2Par -0.54 -0.72 -0.52  
## pctWInvInc -0.69 -0.64 -0.47  
## PctNotSpeakEnglWell 0.59 0.11 0.08  
## PctSpeakEnglOnly -0.47 0.06 0.01  
## PctHousLess3BR 0.52 0.56 0.29  
## PctTeen2Par -0.50 -0.72 -0.57  
## PctPopUnderPov 0.61 0.57 0.43  
## PctUnemployed 0.52 0.44 0.32  
## pctUrban 0.07 0.06 0.08  
## PctLargHouseFam 0.71 0.20 0.25  
## population 0.33 0.29 0.27  
## PctPersOwnOccup -0.63 -0.64 -0.43  
## AsianPerCap -0.27 -0.24 -0.10  
## PctUsePubTrans -0.03 -0.08 0.07  
## MalePctDivorce racePctHisp pctWPubAsst  
## PctKidsBornNeverMar 0.59 0.23 0.75  
## racePctWhite -0.39 -0.48 -0.50  
## PctFam2Par -0.76 -0.09 -0.79  
## HousVacant 0.50 0.27 0.35  
## PctPersDenseHous 0.44 0.69 0.58  
## FemalePctDiv 0.92 0.21 0.57  
## racepctblack 0.43 0.09 0.36  
## MalePctDivorce 1.00 0.11 0.54  
## racePctHisp 0.11 1.00 0.14  
## pctWPubAsst 0.54 0.14 1.00  
## PctYoungKids2Par -0.72 -0.08 -0.83  
## pctWInvInc -0.58 -0.21 -0.80  
## PctNotSpeakEnglWell 0.02 0.82 0.17  
## PctSpeakEnglOnly 0.14 -0.82 -0.03  
## PctHousLess3BR 0.56 0.29 0.52  
## PctTeen2Par -0.69 -0.10 -0.67  
## PctPopUnderPov 0.54 0.12 0.83  
## PctUnemployed 0.42 0.18 0.83  
## pctUrban 0.00 0.22 -0.16  
## PctLargHouseFam 0.12 0.55 0.46  
## population 0.23 0.32 0.19  
## PctPersOwnOccup -0.57 -0.30 -0.62  
## AsianPerCap -0.25 0.00 -0.39  
## PctUsePubTrans -0.12 0.21 -0.13  
## PctYoungKids2Par pctWInvInc PctNotSpeakEnglWell  
## PctKidsBornNeverMar -0.83 -0.72 0.24  
## racePctWhite 0.54 0.59 -0.45  
## PctFam2Par 0.94 0.72 -0.10  
## HousVacant -0.47 -0.36 0.24  
## PctPersDenseHous -0.54 -0.69 0.59  
## FemalePctDiv -0.72 -0.64 0.11  
## racepctblack -0.52 -0.47 0.08  
## MalePctDivorce -0.72 -0.58 0.02  
## racePctHisp -0.08 -0.21 0.82  
## pctWPubAsst -0.83 -0.80 0.17  
## PctYoungKids2Par 1.00 0.75 -0.07  
## pctWInvInc 0.75 1.00 -0.14  
## PctNotSpeakEnglWell -0.07 -0.14 1.00  
## PctSpeakEnglOnly -0.09 -0.01 -0.90  
## PctHousLess3BR -0.63 -0.46 0.33  
## PctTeen2Par 0.81 0.63 -0.10  
## PctPopUnderPov -0.85 -0.79 0.10  
## PctUnemployed -0.72 -0.75 0.18  
## pctUrban 0.11 0.16 0.27  
## PctLargHouseFam -0.29 -0.51 0.51  
## population -0.22 -0.14 0.34  
## PctPersOwnOccup 0.71 0.60 -0.30  
## AsianPerCap 0.40 0.42 0.05  
## PctUsePubTrans 0.14 0.27 0.32  
## PctSpeakEnglOnly PctHousLess3BR PctTeen2Par  
## PctKidsBornNeverMar -0.11 0.59 -0.77  
## racePctWhite 0.36 -0.41 0.57  
## PctFam2Par -0.07 -0.71 0.90  
## HousVacant -0.16 0.50 -0.48  
## PctPersDenseHous -0.47 0.52 -0.50  
## FemalePctDiv 0.06 0.56 -0.72  
## racepctblack 0.01 0.29 -0.57  
## MalePctDivorce 0.14 0.56 -0.69  
## racePctHisp -0.82 0.29 -0.10  
## pctWPubAsst -0.03 0.52 -0.67  
## PctYoungKids2Par -0.09 -0.63 0.81  
## pctWInvInc -0.01 -0.46 0.63  
## PctNotSpeakEnglWell -0.90 0.33 -0.10  
## PctSpeakEnglOnly 1.00 -0.23 -0.04  
## PctHousLess3BR -0.23 1.00 -0.64  
## PctTeen2Par -0.04 -0.64 1.00  
## PctPopUnderPov 0.05 0.62 -0.73  
## PctUnemployed -0.09 0.48 -0.57  
## pctUrban -0.27 0.06 -0.03  
## PctLargHouseFam -0.45 0.14 -0.21  
## population -0.30 0.27 -0.25  
## PctPersOwnOccup 0.17 -0.79 0.72  
## AsianPerCap -0.14 -0.27 0.31  
## PctUsePubTrans -0.38 0.09 0.00  
## PctPopUnderPov PctUnemployed pctUrban PctLargHouseFam  
## PctKidsBornNeverMar 0.74 0.64 0.02 0.46  
## racePctWhite -0.54 -0.43 -0.13 -0.57  
## PctFam2Par -0.83 -0.68 0.03 -0.26  
## HousVacant 0.42 0.36 0.18 0.17  
## PctPersDenseHous 0.61 0.52 0.07 0.71  
## FemalePctDiv 0.57 0.44 0.06 0.20  
## racepctblack 0.43 0.32 0.08 0.25  
## MalePctDivorce 0.54 0.42 0.00 0.12  
## racePctHisp 0.12 0.18 0.22 0.55  
## pctWPubAsst 0.83 0.83 -0.16 0.46  
## PctYoungKids2Par -0.85 -0.72 0.11 -0.29  
## pctWInvInc -0.79 -0.75 0.16 -0.51  
## PctNotSpeakEnglWell 0.10 0.18 0.27 0.51  
## PctSpeakEnglOnly 0.05 -0.09 -0.27 -0.45  
## PctHousLess3BR 0.62 0.48 0.06 0.14  
## PctTeen2Par -0.73 -0.57 -0.03 -0.21  
## PctPopUnderPov 1.00 0.79 -0.22 0.33  
## PctUnemployed 0.79 1.00 -0.19 0.46  
## pctUrban -0.22 -0.19 1.00 0.07  
## PctLargHouseFam 0.33 0.46 0.07 1.00  
## population 0.16 0.15 0.40 0.25  
## PctPersOwnOccup -0.73 -0.53 -0.01 -0.26  
## AsianPerCap -0.48 -0.38 0.13 -0.11  
## PctUsePubTrans -0.24 -0.13 0.49 0.17  
## population PctPersOwnOccup AsianPerCap PctUsePubTrans  
## PctKidsBornNeverMar 0.33 -0.72 -0.31 0.05  
## racePctWhite -0.37 0.56 0.12 -0.11  
## PctFam2Par -0.23 0.78 0.37 0.05  
## HousVacant 0.74 -0.45 -0.11 0.09  
## PctPersDenseHous 0.33 -0.63 -0.27 -0.03  
## FemalePctDiv 0.29 -0.64 -0.24 -0.08  
## racepctblack 0.27 -0.43 -0.10 0.07  
## MalePctDivorce 0.23 -0.57 -0.25 -0.12  
## racePctHisp 0.32 -0.30 0.00 0.21  
## pctWPubAsst 0.19 -0.62 -0.39 -0.13  
## PctYoungKids2Par -0.22 0.71 0.40 0.14  
## pctWInvInc -0.14 0.60 0.42 0.27  
## PctNotSpeakEnglWell 0.34 -0.30 0.05 0.32  
## PctSpeakEnglOnly -0.30 0.17 -0.14 -0.38  
## PctHousLess3BR 0.27 -0.79 -0.27 0.09  
## PctTeen2Par -0.25 0.72 0.31 0.00  
## PctPopUnderPov 0.16 -0.73 -0.48 -0.24  
## PctUnemployed 0.15 -0.53 -0.38 -0.13  
## pctUrban 0.40 -0.01 0.13 0.49  
## PctLargHouseFam 0.25 -0.26 -0.11 0.17  
## population 1.00 -0.30 0.02 0.32  
## PctPersOwnOccup -0.30 1.00 0.34 -0.04  
## AsianPerCap 0.02 0.34 1.00 0.22  
## PctUsePubTrans 0.32 -0.04 0.22 1.00

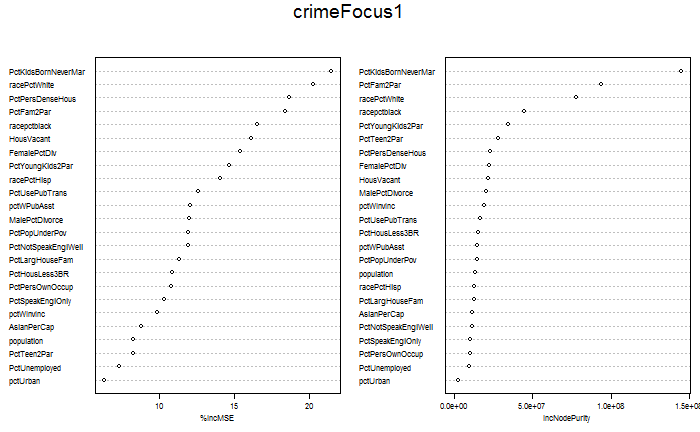
set.seed(4543)  
crimeFocus1 <- randomForest(x = crimeReg[,varNam],   
 y=crimeReg[,81],  
 importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
crimeFocus1

##   
## Call:  
## randomForest(x = crimeReg[, varNam], y = crimeReg[, 81], ntree = 500, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 8  
##   
## Mean of squared residuals: 125863.5  
## % Var explained: 65.98

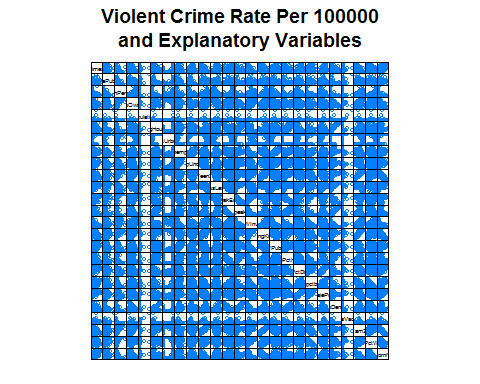
impF1 <- importance(crimeFocus1)  
impF1

## %IncMSE IncNodePurity  
## PctKidsBornNeverMar 21.435871 144646766  
## racePctWhite 20.253396 77901870  
## PctFam2Par 18.366111 93927922  
## HousVacant 16.087994 21519867  
## PctPersDenseHous 18.625013 23198036  
## FemalePctDiv 15.409261 22321858  
## racepctblack 16.526392 45012551  
## MalePctDivorce 11.982029 20511482  
## racePctHisp 14.062629 12953297  
## pctWPubAsst 12.022994 14921321  
## PctYoungKids2Par 14.609801 34529726  
## pctWInvInc 9.833733 19543041  
## PctNotSpeakEnglWell 11.882496 11664636  
## PctSpeakEnglOnly 10.318585 10615423  
## PctHousLess3BR 10.830969 15693032  
## PctTeen2Par 8.209393 28074767  
## PctPopUnderPov 11.934819 14625345  
## PctUnemployed 7.299631 9730584  
## pctUrban 6.288872 2662213  
## PctLargHouseFam 11.295610 12681918  
## population 8.232411 13766986  
## PctPersOwnOccup 10.781728 10059979  
## AsianPerCap 8.800569 11679373  
## PctUsePubTrans 12.567497 16447342

varImpPlot(crimeFocus1,cex=.5) #updated for sizing



varDep <- colnames(crimeReg)[81]  
varNamDep <- c(varNam,varDep )  
  
windows(width=9,height=9)  
splom(crimeReg[,rev(varNamDep)],  
 xlab='',cex=.5,as.matrix=TRUE,  
 main=paste("Violent Crime Rate Per 100000",  
 "and Explanatory Variables",sep="\n"),  
 pscale=0, varname.cex=0.38  
)



## 3.4 A six variable model

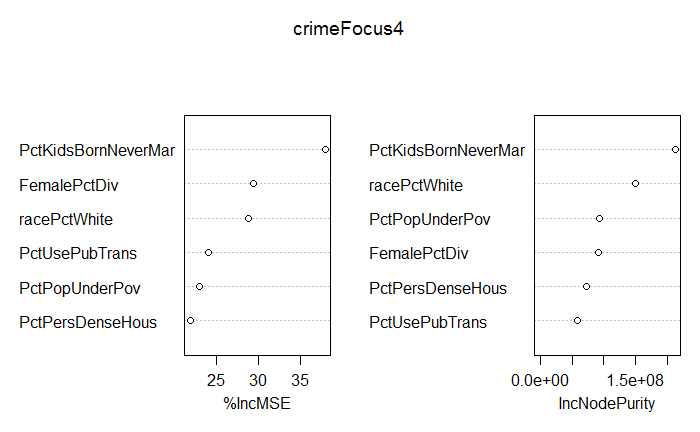
varNam4 <- c(  
 "racePctWhite",   
 "FemalePctDiv", "PctKidsBornNeverMar",   
 "PctPersDenseHous",  
 "PctUsePubTrans", "PctPopUnderPov")   
  
#Set seed 1  
set.seed(4543)  
#set.seed(37)  
crimeFocus4 <- randomForest(x = crimeReg[, varNam4], y=crimeReg[,81],  
 mtry=2, importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
  
crimeFocus4

##   
## Call:  
## randomForest(x = crimeReg[, varNam4], y = crimeReg[, 81], ntree = 500, mtry = 2, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 131575.9  
## % Var explained: 64.44

imp4 <- importance(crimeFocus4)  
imp4

## %IncMSE IncNodePurity  
## racePctWhite 28.82642 149160824  
## FemalePctDiv 29.39239 91672616  
## PctKidsBornNeverMar 37.97813 212293102  
## PctPersDenseHous 21.89859 72948749  
## PctUsePubTrans 24.10107 58279541  
## PctPopUnderPov 22.99763 93592064

varImpPlot(crimeFocus4)



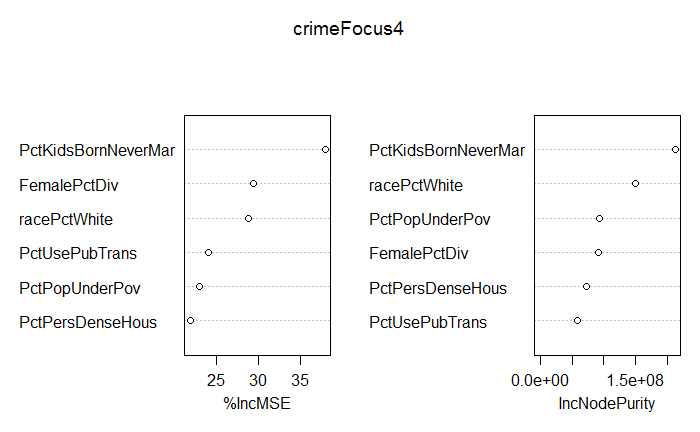
#Set seed 2  
set.seed(37)  
crimeFocus4 <- randomForest(x = crimeReg[, varNam4], y=crimeReg[,81],  
 mtry=2, importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
  
crimeFocus4

##   
## Call:  
## randomForest(x = crimeReg[, varNam4], y = crimeReg[, 81], ntree = 500, mtry = 2, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 131448.5  
## % Var explained: 64.47

imp4 <- importance(crimeFocus4)  
imp4

## %IncMSE IncNodePurity  
## racePctWhite 30.96345 151196098  
## FemalePctDiv 30.48101 93574327  
## PctKidsBornNeverMar 35.50930 212632880  
## PctPersDenseHous 23.46690 72023375  
## PctUsePubTrans 21.92147 56776917  
## PctPopUnderPov 23.08421 92305207

varImpPlot(crimeFocus4)



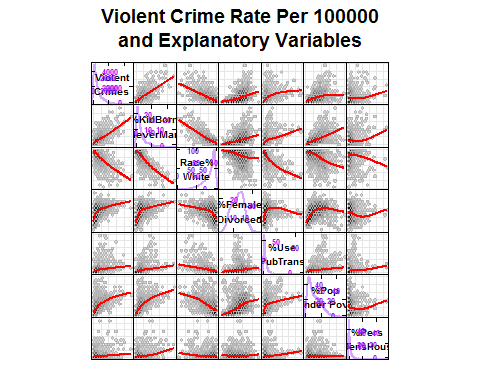
The result is % Var explained: 64.54 Changing the random number seed to 37 yields % Var explained: 64.64

## 3.5 Hexbin scatterplot matrix with smoothes

ord <- order(imp4[,1],decreasing=TRUE)  
varNam4Ord <- varNam4[ord]  
depNam <- colnames(crimeReg)[81]  
varNamPlus <- c(depNam, varNam4Ord)  
varNamPlus

## [1] "ViolentCrimesPerPop" "PctKidsBornNeverMar" "racePctWhite"   
## [4] "FemalePctDiv" "PctPersDenseHous" "PctPopUnderPov"   
## [7] "PctUsePubTrans"

labs <- c(  
paste("Violent","Crimes",sep="\n"),  
paste("%KidBorn","NeverMar",sep="\n"),  
paste("Race%","White",sep="\n"),  
paste("%Female","Divorced",sep="\n"),  
paste("%Use","PubTrans",sep="\n"),  
paste("%Pop","Under Pov.",sep="\n"),  
paste("%Pers","DensHous",sep="\n")  
)  
  
offDiag <- function(x,y,...){  
 panel.grid(h=-1,v=-1,...)  
 panel.hexbinplot(x,y,xbins=15,...,border=gray(.7),  
 trans=function(x)x^.5)  
 panel.loess(x , y, ..., degree=1,span=.65,  
 lwd=2,col='red')  
 }  
  
onDiag <-   
function(x, ...){  
 yrng <- current.panel.limits()$ylim  
 d <- density(x, na.rm=TRUE)  
 d$y <- with(d, yrng[1] + 0.95 \* diff(yrng) \* y / max(y) )  
 panel.lines(d,col=rgb(.83,.66,1),lwd=2)  
 diag.panel.splom(x, ...)  
 }  
   
windows(width=9,height=9)   
splom(crimeReg[, varNamPlus],as.matrix=TRUE,  
 varnames=labs,   
 xlab='',  
 main=paste("Violent Crime Rate Per 100000",  
 "and Explanatory Variables",sep="\n"),  
 pscale=2, varname.cex=0.62,varname.font=2,  
 axis.text.cex=0.5, axis.text.col="purple",axis.text.font=2,  
 axis.line.tck=.5,  
 panel=offDiag,  
 diag.panel = onDiag  
)



## 3.6 Assignment: Find a different six variable model

At least one variable should differ the variables shown in 3.5. Try at least two models and pick the one with a large percent of variance explained. For this model

1. Report the explanatory variables, the random number seed, and the % of variance explained. Using the same variables as in a) report the % of variance explained after making the following change and rerunning the script
2. Change the seed
3. Change the mtry parameter to 3, the kinds of items in a)
4. Raise the power of the dependent variable to 0.2 and report kinds if item in a)

#Test two different 6 variable models  
##Test A  
varNam36A <- c(  
 "racePctWhite", "PctFam2Par",  
 "HousVacant", "PctKidsBornNeverMar",   
 "PctPersDenseHous","PctPopUnderPov")   
  
set.seed(4543)  
crimeFocus36A <- randomForest(x = crimeReg[, varNam36A], y=crimeReg[,81],  
 mtry=2, importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
  
crimeFocus36A

##   
## Call:  
## randomForest(x = crimeReg[, varNam36A], y = crimeReg[, 81], ntree = 500, mtry = 2, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 137361.1  
## % Var explained: 62.87

##Test B  
varNam36B <- c(  
 "FemalePctDiv", "PctFam2Par",  
 "HousVacant", "PctKidsBornNeverMar",   
 "PctYoungKids2Par","PctPopUnderPov")   
  
set.seed(4543)  
crimeFocus36B <- randomForest(x = crimeReg[, varNam36B],   
 y=crimeReg[,81],  
 mtry=2,  
 importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
  
crimeFocus36B

##   
## Call:  
## randomForest(x = crimeReg[, varNam36B], y = crimeReg[, 81], ntree = 500, mtry = 2, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 143673.1  
## % Var explained: 61.17

###Model A explained more variance – 62.87 vs 61.17

#Change the seed  
set.seed(5)  
crimeFocus36A <- randomForest(x = crimeReg[, varNam36A],  
 y=crimeReg[,81],  
 mtry=2,  
 importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
  
crimeFocus36A

##   
## Call:  
## randomForest(x = crimeReg[, varNam36A], y = crimeReg[, 81], ntree = 500, mtry = 2, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 136668  
## % Var explained: 63.06

#Change mtry parameter to 3  
set.seed(4543)  
crimeFocus36A <- randomForest(x = crimeReg[, varNam36A],  
 y=crimeReg[,81],  
 mtry=3,  
 importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
crimeFocus36A

##   
## Call:  
## randomForest(x = crimeReg[, varNam36A], y = crimeReg[, 81], ntree = 500, mtry = 3, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 3  
##   
## Mean of squared residuals: 138310.3  
## % Var explained: 62.62

# Raise the power of the dependent variable to 0.2   
set.seed(4543)  
crimeFocus36A <- randomForest(x = crimeReg[, varNam36A],  
 y=crimeReg[,81]\*\*.2,  
 mtry=2,  
 importance=TRUE, proximity=FALSE, ntree=500, keepForest=FALSE)  
crimeFocus36A

##   
## Call:  
## randomForest(x = crimeReg[, varNam36A], y = crimeReg[, 81]^0.2, ntree = 500, mtry = 2, importance = TRUE, proximity = FALSE, keepForest = FALSE)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## Mean of squared residuals: 0.15918  
## % Var explained: 67.49

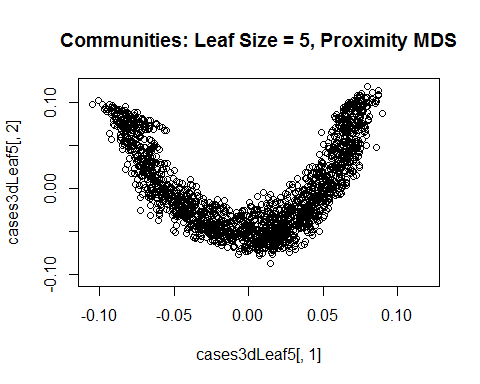
# Case proximities, dissimilarity matrices multidimensional scaling data clustering case clustering and graphs.

## 4.5 Proximity-based case distance plots

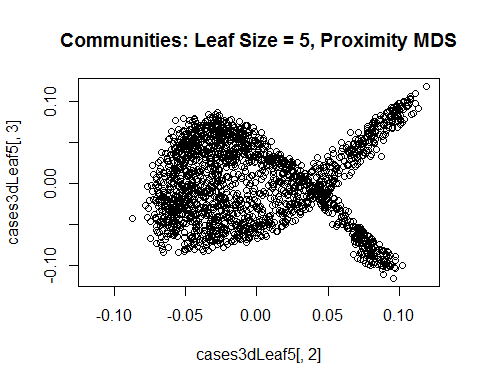
#Graphics minimum node size = 5  
proximityDist <- 1-crimeFocus4a$proximity  
cases3dLeaf5 <- cmdscale(proximityDist,k=3)  
head(cases3dLeaf5)

## [,1] [,2] [,3]  
## 1 0.06372490 0.09075553 0.08730912  
## 2 0.07224562 0.06386898 0.02402676  
## 3 0.02584782 -0.03714209 -0.03021634  
## 6 0.01184428 -0.05942331 -0.03223935  
## 7 0.05416656 0.01604695 -0.03320099  
## 8 -0.06155815 -0.02501095 0.05529197

windows()  
rval <- range(cases3dLeaf5[,1:2])  
plot(cases3dLeaf5[,1],cases3dLeaf5[,2],pty="s",  
 xlim=rval,ylim=rval,  
 main="Communities: Leaf Size = 5, Proximity MDS")



windows()  
rval <- range(cases3dLeaf5[,2:3])  
plot(cases3dLeaf5[,2],cases3dLeaf5[,3],pty="s",  
 xlim=rval,ylim=rval,  
 main="Communities: Leaf Size = 5, Proximity MDS")



library(rgl)  
open3d(FOV=0)

## wgl   
## 1

aspect3d(x=c(1,1,1))  
bg3d(color=c("white","black"))  
plot3d(cases3dLeaf5,radius=.004,col="red",type='s',  
 main="Minimum Leaf Size = 5")

rgl.snapshot("CrimeProx5.png")

proximityDist <- 1-crimeFocus4b$proximity  
cases3dLeaf20 <- cmdscale(proximityDist,k=3)  
head(cases3dLeaf20)

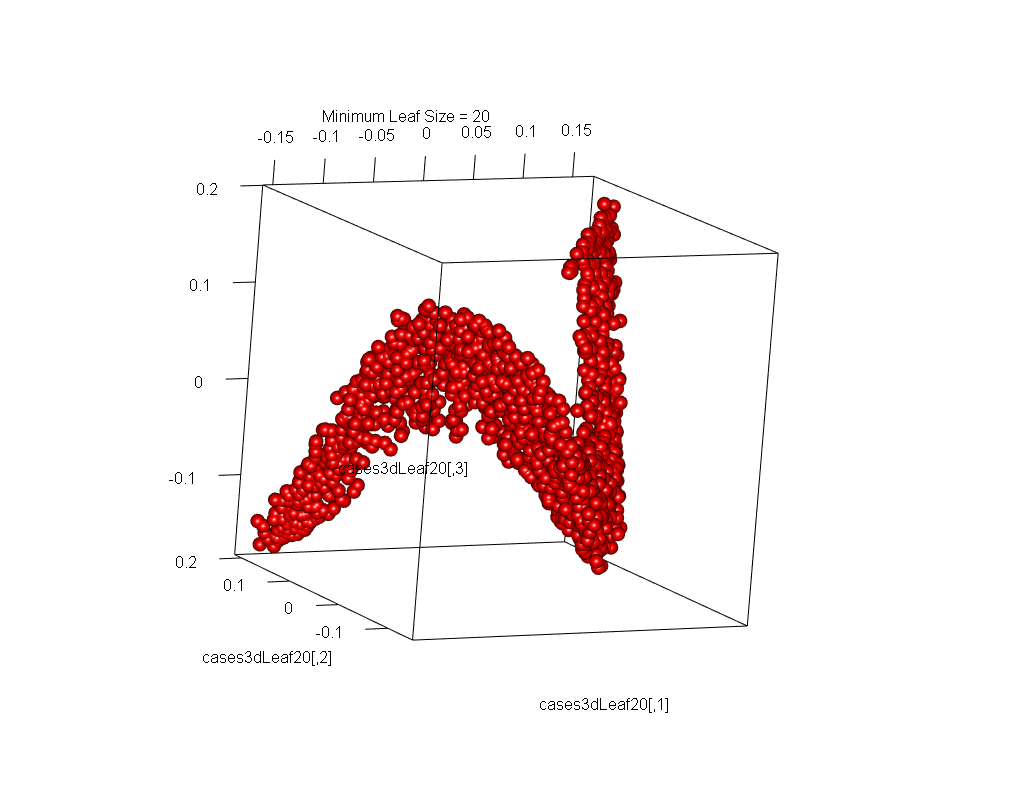
## [,1] [,2] [,3]  
## 1 -0.10802797 0.14833922 -0.12359604  
## 2 -0.10542329 0.09192364 -0.02084566  
## 3 -0.05216331 -0.08034773 0.05864373  
## 6 -0.02611698 -0.12889158 0.05066464  
## 7 -0.10075158 0.02024339 0.08680665  
## 8 0.08964895 -0.01329818 -0.07176246

#Graphics minimum node size = 20  
open3d(FOV=3)

## wgl   
## 2

aspect3d(x=c(1,1,1))  
bg3d(color=c("white","black"))  
plot3d(cases3dLeaf20,radius=.007,col="red",type='s',  
 main="Minimum Leaf Size = 20")

rgl.snapshot("CrimeProx20.png")



# 5- Local variable importance

More exploratory (and questionable) ideas

The script producing the CrimeFocus4a and 4b models, specified calculating both global variable importance and local (individual community) variable importance. Below we access local variable importance matrix.

impDat4a <- t(crimeFocus4a$localImportance)  
head(impDat4a)

## racePctWhite FemalePctDiv PctKidsBornNeverMar PctPersDenseHous  
## 1 6474.942 33337.588 88069.62 13209.49  
## 2 11308.608 24893.515 63535.45 36902.49  
## 3 63888.305 5626.137 138335.54 57086.75  
## 6 110308.709 24916.328 67587.40 23112.39  
## 7 126525.477 24066.473 16166.07 18704.51  
## 8 44347.150 -348687.360 -513306.04 -74284.85  
## PctUsePubTrans PctPopUnderPov  
## 1 3232.721 53996.60449  
## 2 2442.247 20907.27385  
## 3 4711.278 54056.04634  
## 6 6518.193 -18.91132  
## 7 -4168.371 37074.28844  
## 8 40960.393 -362704.45485

clus4a <- kmeans(impDat4a,3,nstart=25)  
clus4a$centers

## racePctWhite FemalePctDiv PctKidsBornNeverMar PctPersDenseHous  
## 1 671042.61 397612.08 1208403.44 130057.29  
## 2 2445565.53 1545041.65 4383635.97 547076.73  
## 3 19361.89 13838.57 48506.27 20813.87  
## PctUsePubTrans PctPopUnderPov  
## 1 219197.30 535573.45  
## 2 1557372.10 1659111.29  
## 3 12304.95 12018.84

table(clus4a$cluster) # $clusters

##   
## 1 2 3   
## 67 8 1826

# impDat4a obtained above  
head(impDat4a)

## racePctWhite FemalePctDiv PctKidsBornNeverMar PctPersDenseHous  
## 1 6474.942 33337.588 88069.62 13209.49  
## 2 11308.608 24893.515 63535.45 36902.49  
## 3 63888.305 5626.137 138335.54 57086.75  
## 6 110308.709 24916.328 67587.40 23112.39  
## 7 126525.477 24066.473 16166.07 18704.51  
## 8 44347.150 -348687.360 -513306.04 -74284.85  
## PctUsePubTrans PctPopUnderPov  
## 1 3232.721 53996.60449  
## 2 2442.247 20907.27385  
## 3 4711.278 54056.04634  
## 6 6518.193 -18.91132  
## 7 -4168.371 37074.28844  
## 8 40960.393 -362704.45485

xyzCrime4a <- svd(impDat4a)$u[,1:3]  
head(xyzCrime4a)

## [,1] [,2] [,3]  
## [1,] -0.004180385 1.785661e-08 0.0050060426  
## [2,] -0.003014838 -8.079127e-04 0.0025393706  
## [3,] -0.006774523 6.307726e-03 -0.0010296130  
## [4,] -0.004711820 3.023842e-03 -0.0009251313  
## [5,] -0.003711616 1.442412e-03 0.0012475656  
## [6,] 0.024676479 1.545228e-02 -0.0617910874

open3d()

## wgl   
## 4

aspect3d(x=c(1,1,1))  
bg3d(color=c("white","black"))  
par3d(FOV=3)  
plot3d(xyzCrime4a,radius=.006,  
 col=c("green","red","blue")[clus4a$cluster],  
 type='s',  
 main="Minimum Leaf Size = 5",  
 xlab="x",ylab="y",zlab="z")  
  
rgl.snapshot("CrimeProx3Color.png")

