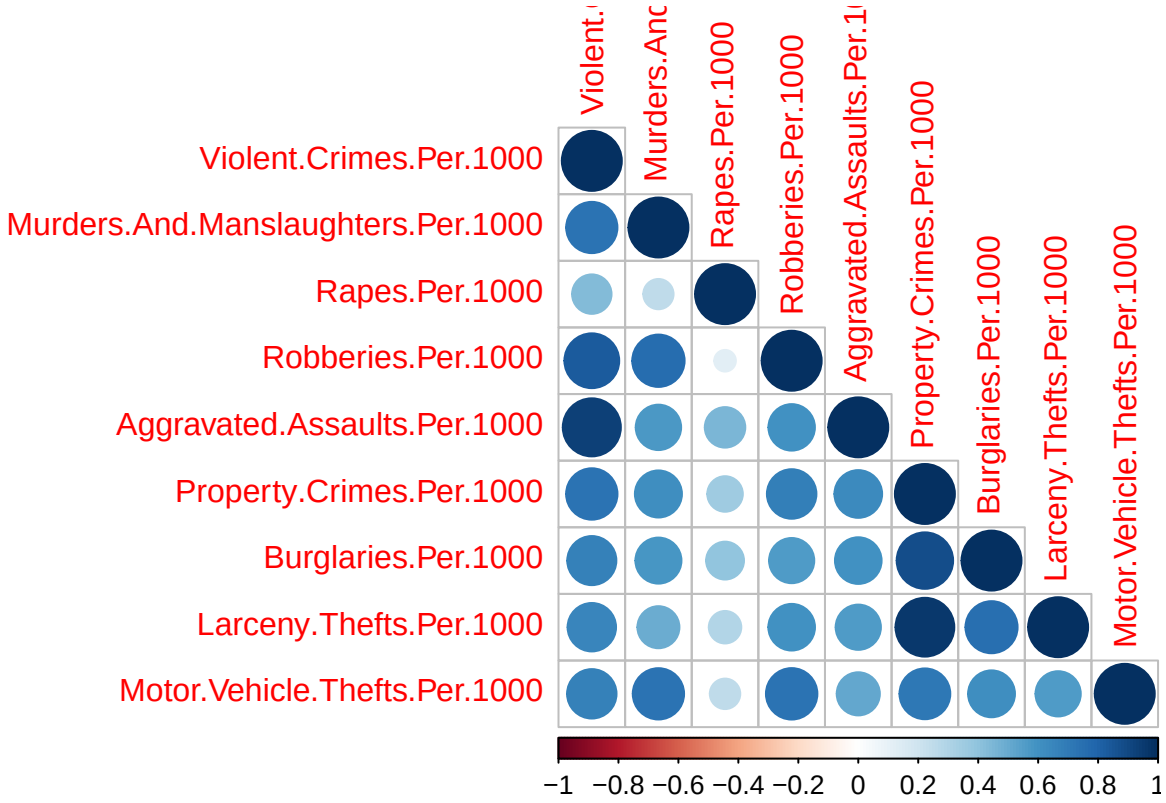
Appendix A: Data Cleaning and Preparation

Jimmy DeLano and Casey DeLano

12/4/2019

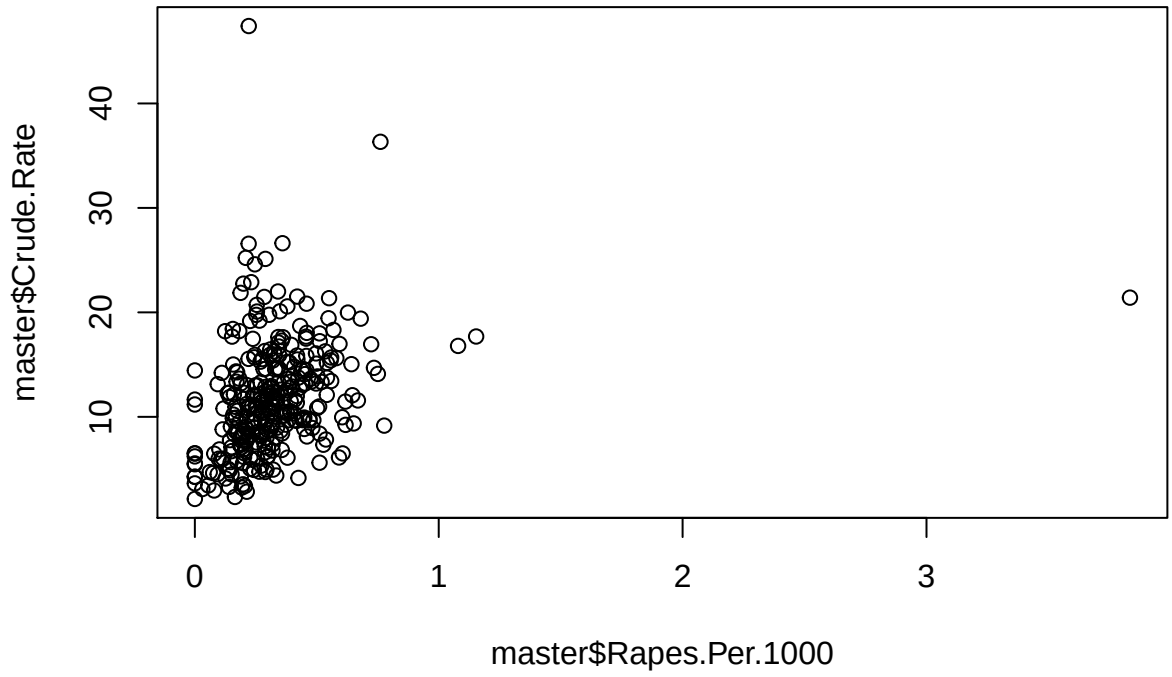
```
cdc_data <- read.csv("~/stat346/final-project/cdc_data.csv")
census <- read.csv("~/stat346/final-project/gundata.csv")
colnames(census)[colnames(census) == 'STCOU'] <- "County.Code"
laws <- read.csv("~/stat346/final-project/laws.csv")

merge1 <- merge(x = census, y = cdc_data, by = "County.Code", all.y = TRUE)

merge2 <- merge(x = merge1, y = laws, by = "State", all.x = TRUE)
merge2$Federal.Govt.Expenditure.Per.Person = merge2$Federal.Govt.Expenditure / merge2$Population

master <- subset(merge2, select = -c(State, County.Code, Areaname,
                                     Deaths, Total.Persons, Federal.Govt.Expenditure))

# divide crime numbers by population
master[,4:12] = (master[,4:12] / master$Population) * 1000

names(master)[4:12] <- c("Violent.Crimes.Per.1000", "Murders.And.Manslaughters.Per.1000", "Rapes.Per.10
                         "Robberies.Per.1000", "Aggravated.Assaults.Per.1000", "Property.Crimes.Per.100
                         "Burglaries.Per.1000", "Larceny.Thefts.Per.1000", "Motor.Vehicle.Thefts.Per.10
```

```
df_pearson_cor_values <-cor(master[4:12],method = "pearson")
corrplot(df_pearson_cor_values, method="circle", type = "lower")
```

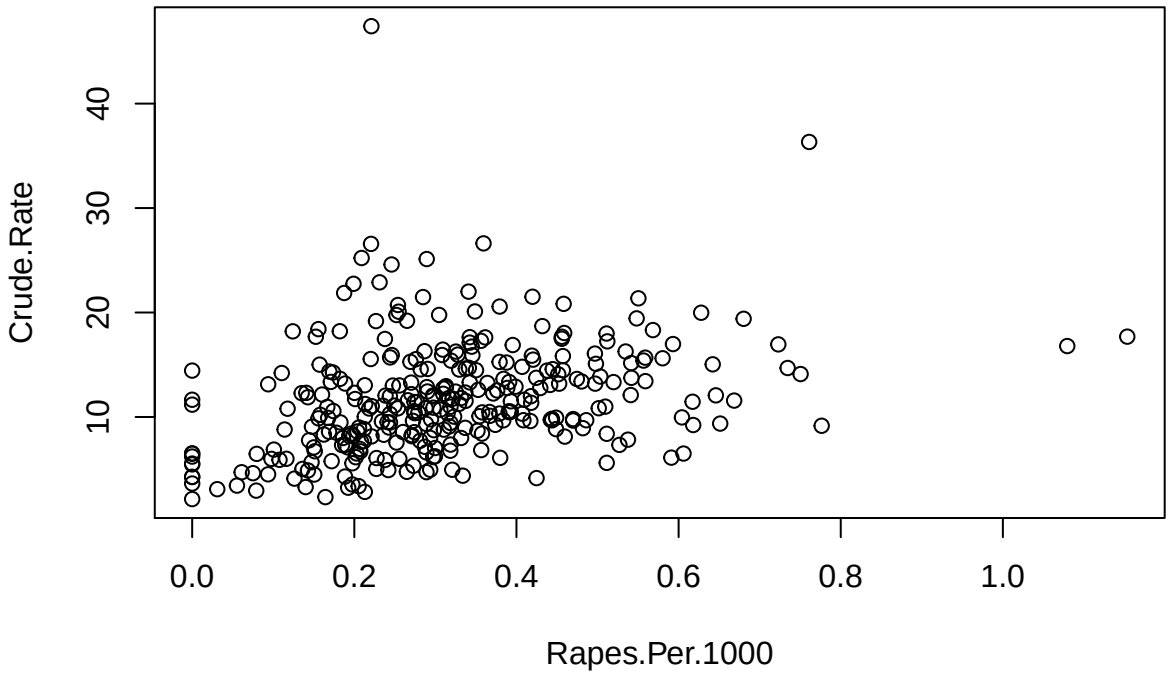


Rapes doesn't have as much collinearity with the other crime predictors. Let's take a look at it:

```
# feel like something could be wrong here --> huge influence point
plot(master$Rapes.Per.1000, master$Crude.Rate)
```



```
plot(Crude.Rate~Rapes.Per.1000, data = master[master$Rapes.Per.1000 < 2, ])
```



We found that Navajo County, AZ had way too many rapes per 1000. More research found that this data was actually be incorrect so the row will be dropped from the dataset. See the final paper for more details (attribution link).

```
master <- master[!(master$County == 'Navajo County, AZ'),]
summary(master)
```

```
##    Median.Age    Avg.Land.Value   Unemployment.Rate
##  Min.   :24.60   Min.   :     0   Min.   : 4.900
##  1st Qu.:34.20   1st Qu.:  3222   1st Qu.: 8.100
##  Median :36.65   Median :  4850   Median : 9.500
##  Mean   :36.97   Mean   :  9044   Mean   : 9.734
##  3rd Qu.:39.30   3rd Qu.:  7939   3rd Qu.:11.100
##  Max.   :55.90   Max.   :457143   Max.   :18.900
##
##  Violent.Crimes.Per.1000 Murders.And.Manslaughters.Per.1000
##  Min.   : 0.000          Min.   :0.00000
##  1st Qu.: 2.985          1st Qu.:0.02480
##  Median : 4.457          Median :0.04421
##  Mean   : 4.977          Mean   :0.06017
##  3rd Qu.: 6.551          3rd Qu.:0.07631
##  Max.   :23.408          Max.   :0.54182
##
##  Rapes.Per.1000   Robberies.Per.1000 Aggravated.Assaults.Per.1000
##  Min.   :0.0000   Min.   :0.0000     Min.   : 0.000
##  1st Qu.:0.2042   1st Qu.:0.6597     1st Qu.: 1.700
##  Median :0.2941   Median :1.2402     Median : 2.720
##  Mean   :0.3139   Mean   :1.5602     Mean   : 3.043
##  3rd Qu.:0.4050   3rd Qu.:1.9881     3rd Qu.: 3.912
##  Max.   :1.1535   Max.   :8.8857     Max.   :13.837
##
```

```
##  Property.Crimes.Per.1000 Burglaries.Per.1000 Larceny.Thefts.Per.1000
##  Min.   : 0.00            Min.   : 0.000      Min.   : 0.00
##  1st Qu.:25.70            1st Qu.: 5.202      1st Qu.:17.37
##  Median :32.98            Median : 7.765      Median :22.26
##  Mean   :35.12            Mean   : 8.180      Mean   :23.61
##  3rd Qu.:43.40            3rd Qu.:10.234      3rd Qu.:28.68
##  Max.   :97.40            Max.   :24.618      Max.   :56.08
##
##  Motor.Vehicle.Thefts.Per.1000 High.School.Or.Higher.Pct
##  Min.   : 0.000                Min.   :59.50
##  1st Qu.: 1.654                1st Qu.:83.45
##  Median : 2.706                Median :86.60
##  Mean   : 3.327                Mean   :85.88
##  3rd Qu.: 4.364                3rd Qu.:89.28
##  Max.   :18.425                Max.   :97.30
##
##  Bachelors.Or.Higher.Pct Republican.Vote.Pct Without.Health.Insurance.Pct
##  Min.   :12.10           Min.   : 0.00       Min.   : 6.90
##  1st Qu.:21.93           1st Qu.:38.15       1st Qu.:13.20
##  Median :27.70           Median :45.30       Median :16.70
##  Mean   :28.74           Mean   :45.81       Mean   :17.11
##  3rd Qu.:33.58           3rd Qu.:55.88       3rd Qu.:20.50
##  Max.   :58.40           Max.   :85.00       Max.   :33.50
##
##  Income.Per.Capita Median.Household.Income  Poverty.Rate
##  Min.   :13130     Min.   : 27421          Min.   : 3.30
##  1st Qu.:23257     1st Qu.: 42827          1st Qu.:10.72
##  Median :26193     Median : 48940          Median :14.25
##  Mean   :27454     Mean   : 52414          Mean   :14.14
##  3rd Qu.:30402     3rd Qu.: 58896          3rd Qu.:17.00
##  Max.   :60047     Max.   :102325          Max.   :35.20
##
##    Land.Area        Population.Density   Female.Pct      White.Pct
##  Min.   :   22.83   Min.   :    7.2    Min.   :46.40   Min.   :18.90
##  1st Qu.:  450.15   1st Qu.:  294.8    1st Qu.:50.60   1st Qu.:61.83
##  Median :  687.66   Median :  575.0    Median :51.15   Median :73.90
##  Mean   : 1238.68   Mean   : 1612.3    Mean   :51.08   Mean   :71.04
##  3rd Qu.: 1019.07   3rd Qu.: 1315.0    3rd Qu.:51.67   3rd Qu.:82.42
##  Max.   :20056.94   Max.   :69467.5    Max.   :53.10   Max.   :96.50
##
##    Black.Pct        Asian.Pct      Mixed.Race.Pct    Hispanic.Pct
##  Min.   : 0.200   Min.   : 0.400   Min.   : 0.900   Min.   : 0.900
##  1st Qu.: 4.625   1st Qu.: 1.600   1st Qu.: 2.100   1st Qu.: 5.425
##  Median :10.800   Median : 2.700   Median : 2.600   Median : 9.450
##  Mean   :15.128   Mean   : 4.139   Mean   : 2.982   Mean   :14.469
##  3rd Qu.:21.125   3rd Qu.: 4.600   3rd Qu.: 3.500   3rd Qu.:18.175
##  Max.   :69.100   Max.   :43.900   Max.   :22.300   Max.   :90.600
##
##  Foreign.Born.Pct                  County      Population
##  Min.   : 1.200   Ada County, ID      :  1   Min.   :  75129
##  1st Qu.: 5.025   Adams County, CO    :  1   1st Qu.: 238444
##  Median : 7.800   Aiken County, SC    :  1   Median : 395458
##  Mean   :10.718   Alachua County, FL  :  1   Mean   : 616680
##  3rd Qu.:14.050   Alameda County, CA  :  1   3rd Qu.: 718588
```

```
##  Max.   :49.400   Allegheny County, PA:  1   Max.   :9818605
##                   (Other)              :312
##    Crude.Rate    Gun.Law.Rank.2010 Federal.Govt.Expenditure.Per.Person
##  Min.   : 2.14   Min.   : 1        Min.   :  0.000
##  1st Qu.: 8.17   1st Qu.:10        1st Qu.:  6.589
##  Median :11.06   Median :19        Median :  8.640
##  Mean   :11.68   Mean   :19        Mean   : 10.745
##  3rd Qu.:14.57   3rd Qu.:27        3rd Qu.: 11.371
##  Max.   :47.41   Max.   :50        Max.   :102.904
##                  NA's   :1
```

From the summary, we notice lots of minimums with 0. Turns out there are 4 rows with missing values. Anchorage Borough, AK had 0% republican vote, so the row was dropped. Avg.Land.Value had 13 missing values, so we'll drop this column completely. The Bronx County, NY, New York County, NY, & Queens County, NY were missing all columns of crime data and federal govt expenditure so they were dropped. We'll drop District of Columbia, DC because it has no Gun.Law.Rank.2010.

```
master <- master[!(master$County %in% c('Anchorage Borough, AK', 'Bronx County, NY',
                                        'New York County, NY', 'Queens County, NY', 'District of Columbi
master <- subset(master, select = -c(Avg.Land.Value))
summary(master)
```

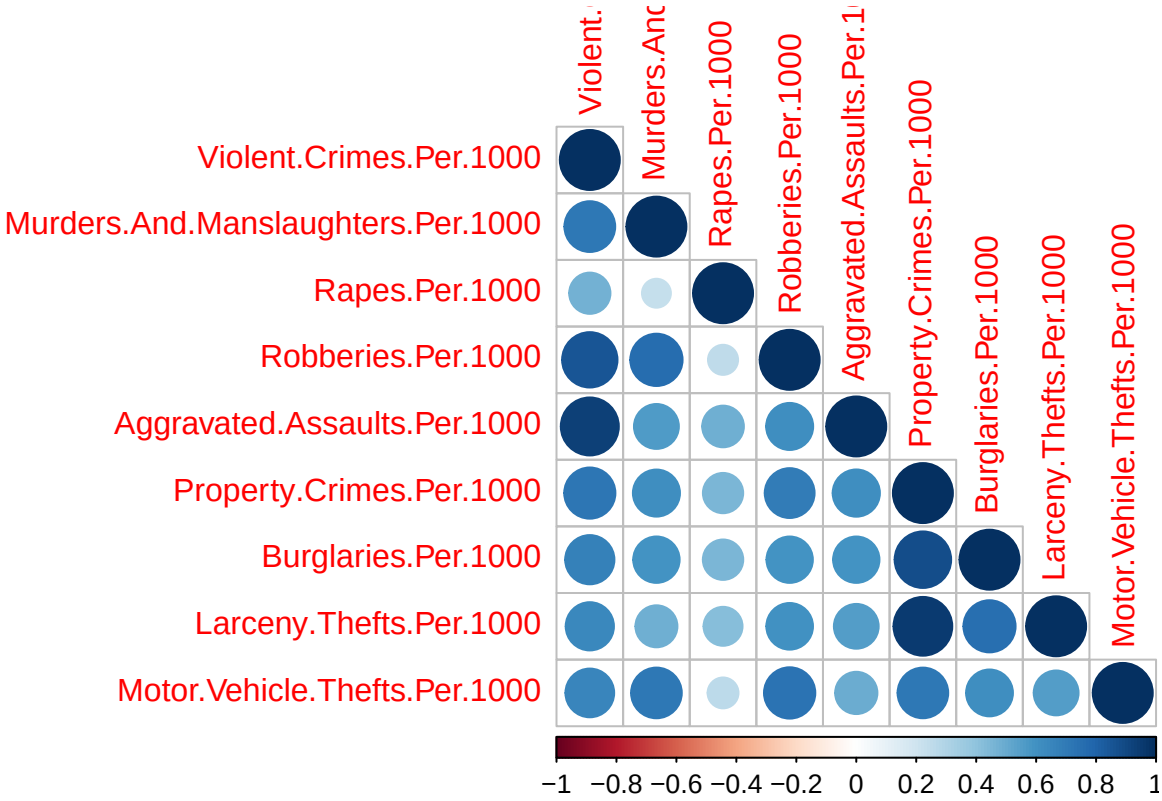
```
##    Median.Age   Unemployment.Rate Violent.Crimes.Per.1000
##  Min.   :24.6   Min.   : 4.900    Min.   : 0.4275
##  1st Qu.:34.2   1st Qu.: 8.100    1st Qu.: 3.0332
##  Median :36.7   Median : 9.500    Median : 4.4917
##  Mean   :37.0   Mean   : 9.742    Mean   : 4.9738
##  3rd Qu.:39.4   3rd Qu.:11.100    3rd Qu.: 6.5260
##  Max.   :55.9   Max.   :18.900    Max.   :23.4079
##
##  Murders.And.Manslaughters.Per.1000 Rapes.Per.1000   Robberies.Per.1000
##  Min.   :0.00000                    Min.   :0.0000   Min.   :0.0583
##  1st Qu.:0.02488                    1st Qu.:0.2058   1st Qu.:0.6650
##  Median :0.04423                    Median :0.2946   Median :1.2516
##  Mean   :0.05993                    Mean   :0.3145   Mean   :1.5553
##  3rd Qu.:0.07649                    3rd Qu.:0.4070   3rd Qu.:1.9917
##  Max.   :0.54182                    Max.   :1.1535   Max.   :8.8857
##
##  Aggravated.Assaults.Per.1000 Property.Crimes.Per.1000 Burglaries.Per.1000
##  Min.   : 0.3036              Min.   : 5.742           Min.   : 0.732
##  1st Qu.: 1.7050              1st Qu.:26.300           1st Qu.: 5.307
##  Median : 2.7246              Median :32.993           Median : 7.796
##  Mean   : 3.0440              Mean   :35.368           Mean   : 8.264
##  3rd Qu.: 3.9089              3rd Qu.:43.345           3rd Qu.:10.248
##  Max.   :13.8368              Max.   :97.396           Max.   :24.618
##
##  Larceny.Thefts.Per.1000 Motor.Vehicle.Thefts.Per.1000
##  Min.   : 4.933          Min.   : 0.07773
##  1st Qu.:17.496          1st Qu.: 1.66560
##  Median :22.396          Median : 2.71386
##  Mean   :23.771          Mean   : 3.33320
##  3rd Qu.:28.653          3rd Qu.: 4.36856
##  Max.   :56.080          Max.   :18.42503
##
##  High.School.Or.Higher.Pct Bachelors.Or.Higher.Pct Republican.Vote.Pct
```

```
##  Min.   :59.50             Min.   :12.10           Min.   :10.40
##  1st Qu.:83.60             1st Qu.:21.90           1st Qu.:38.70
##  Median :86.60             Median :27.60           Median :45.80
##  Mean   :85.94             Mean   :28.61           Mean   :46.37
##  3rd Qu.:89.30             3rd Qu.:33.50           3rd Qu.:56.30
##  Max.   :97.30             Max.   :58.40           Max.   :85.00
##
##  Without.Health.Insurance.Pct Income.Per.Capita Median.Household.Income
##  Min.   : 6.90                Min.   :13130     Min.   : 27421
##  1st Qu.:13.20                1st Qu.:23256     1st Qu.: 42792
##  Median :16.80                Median :26189     Median : 48772
##  Mean   :17.13                Mean   :27327     Mean   : 52335
##  3rd Qu.:20.50                3rd Qu.:30020     3rd Qu.: 58732
##  Max.   :33.50                Max.   :53284     Max.   :102325
##
##   Poverty.Rate    Land.Area        Population.Density   Female.Pct
##  Min.   : 3.3   Min.   :   46.19   Min.   :    7.2    Min.   :46.40
##  1st Qu.:10.7   1st Qu.:  457.85   1st Qu.:  294.7    1st Qu.:50.60
##  Median :14.2   Median :  698.91   Median :  570.8    Median :51.10
##  Mean   :14.1   Mean   : 1252.27   Mean   : 1213.2    Mean   :51.07
##  3rd Qu.:17.0   3rd Qu.: 1020.21   3rd Qu.: 1296.2    3rd Qu.:51.60
##  Max.   :35.2   Max.   :20056.94   Max.   :35369.1    Max.   :53.00
##
##    White.Pct       Black.Pct       Asian.Pct      Mixed.Race.Pct
##  Min.   :18.90   Min.   : 0.20   Min.   : 0.400   Min.   : 0.90
##  1st Qu.:62.60   1st Qu.: 4.60   1st Qu.: 1.600   1st Qu.: 2.10
##  Median :74.00   Median :10.70   Median : 2.700   Median : 2.60
##  Mean   :71.44   Mean   :14.96   Mean   : 4.047   Mean   : 2.95
##  3rd Qu.:82.60   3rd Qu.:20.90   3rd Qu.: 4.600   3rd Qu.: 3.50
##  Max.   :96.50   Max.   :69.10   Max.   :43.900   Max.   :22.30
##
##   Hispanic.Pct   Foreign.Born.Pct                  County
##  Min.   : 0.90   Min.   : 1.20    Ada County, ID      :  1
##  1st Qu.: 5.40   1st Qu.: 5.00    Adams County, CO    :  1
##  Median : 9.40   Median : 7.80    Aiken County, SC    :  1
##  Mean   :14.31   Mean   :10.48    Alachua County, FL  :  1
##  3rd Qu.:17.70   3rd Qu.:13.50    Alameda County, CA  :  1
##  Max.   :90.60   Max.   :49.40    Allegheny County, PA:  1
##                                   (Other)             :307
##    Population        Crude.Rate    Gun.Law.Rank.2010
##  Min.   :  75129   Min.   : 2.33   Min.   : 1.00
##  1st Qu.: 234906   1st Qu.: 8.20   1st Qu.:10.00
##  Median : 389891   Median :11.08   Median :19.00
##  Mean   : 607057   Mean   :11.72   Mean   :19.04
##  3rd Qu.: 713335   3rd Qu.:14.56   3rd Qu.:27.00
##  Max.   :9818605   Max.   :47.41   Max.   :50.00
##
##  Federal.Govt.Expenditure.Per.Person
##  Min.   : 2.016
##  1st Qu.: 6.606
##  Median : 8.647
##  Mean   :10.535
##  3rd Qu.:11.312
##  Max.   :61.377
```

##

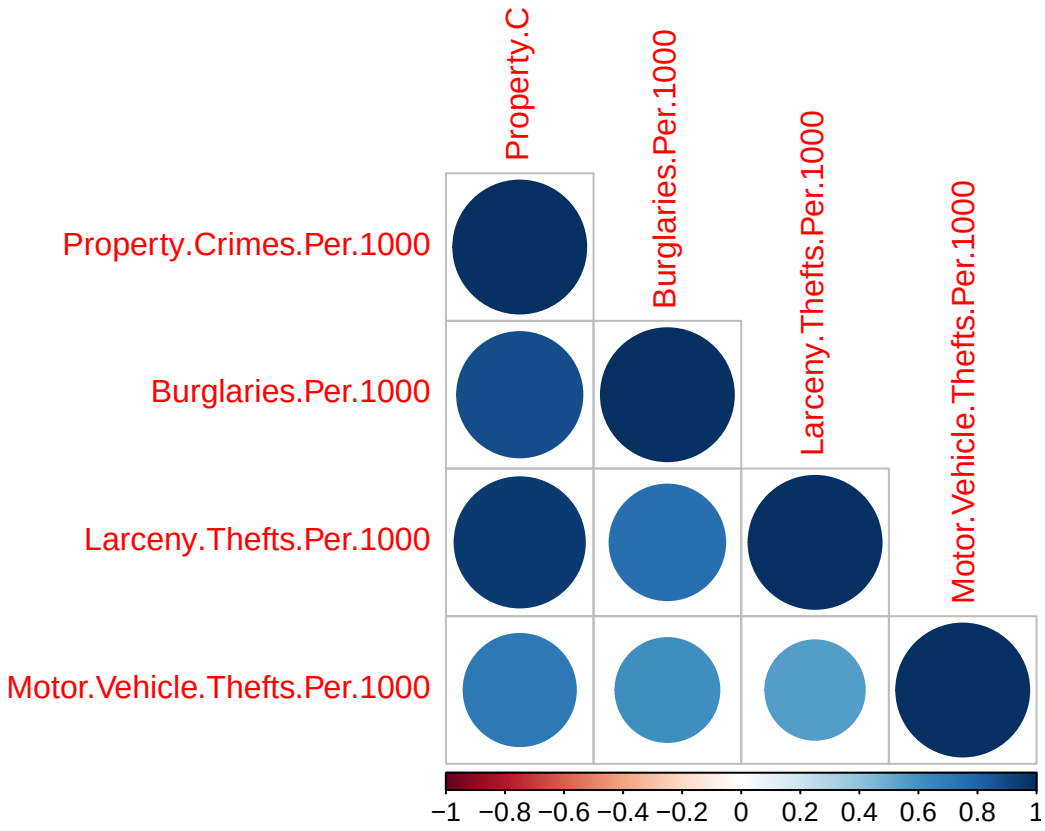
Nice. It looks like we've removed all missing values. Let's dive into the crime statistics a little more.

```
df_pearson_cor_values2 <-cor(master[3:11],method = "pearson")
corrplot(df_pearson_cor_values2, method="circle", type = "lower")
```



We notice that violent crimes = murders + rapes + robberies + aggravated assaults, so we'll drop the individual columns after showing that violent crimes = sum of other crimes.

```
suppressMessages(suppressWarnings(attach(master)))
assert(Violent.Crimes.Per.1000-(Murders.And.Manslaughters.Per.1000+Rapes.Per.1000+
                                Robberies.Per.1000+Aggravated.Assaults.Per.1000)<.00001)

# now lets look at the other crime statistics
df_pearson_cor_values3 <-cor(master[8:11],method = "pearson")
corrplot(df_pearson_cor_values3, method="circle", type = "lower")
```



Because there's a high level of collinearity between the rest of the crime statistics, we'll combine them into a non.violent.cimes.per.1000 variable.

```
master$Non.Violent.Crimes.Per.1000 = Property.Crimes.Per.1000 + Burglaries.Per.1000 +
  Larceny.Thefts.Per.1000 + Motor.Vehicle.Thefts.Per.1000

# remove columns
final.df <- subset(master, select = -c(Property.Crimes.Per.1000, Burglaries.Per.1000, Larceny.Thefts.Pe
                                       Motor.Vehicle.Thefts.Per.1000, Murders.And.Manslaughters.Per.10
                                       Rapes.Per.1000, Robberies.Per.1000, Aggravated.Assaults.Per.100

## create categories for republican vote
# 0%-40% = Democrat, 40%-60% = Swing, 60%-100% = Republican
final.df$Voter.Group.2008 <- cut(final.df$Republican.Vote.Pct,
                   breaks=c(-Inf, 40, 60, Inf),
                   labels=c("Democrat","Swing","Republican"))

final.df <- subset(final.df, select = -c(Republican.Vote.Pct))

write.csv(final.df, file = "~/stat346/final-project/final.csv", row.names = FALSE)
```