Assignment 1

Group 12 - 1817945, 1727716, 1804107, 1802212

7 February 2021

Executive summary

Some missing values were recorded in the dataset as NA, which was easy to filter out. Some other missing data was given a 'dummy value' instead of its true value. All such missing data was omitted before any statistical analysis was carried out.

After analysis of the data set we saw that pctKidsBornNevrMarr (the percentage of kids born to parents who never married) and pctKids2Par (the percentage of kids in family housing with two parents) were the most correlated to the violent and non-violent crime rates in each community.

From the exploratory data analysis we concluded that the required variables to consider when modelling violent and non violent crime in the USA were:

- region
- medIncome
- pctWdiv
- pctNotHSgrad
- pctEmploy
- pctKidsBornNevrMarr
- pctHousOccup
- pctHousOwnerOccup
- pctVacantBoarded
- rentMed
- popDensity
- pctForeignBorn.

1 Findings

1.1 About the data and Missing Values

Looking at the data fields, we see that there are 2 output variables (violentPerPop and nonViolPerPop), 2 factor variables (state and region), and the rest are numeric values. There were only 3 values for the Pacific region (a lot less than the other regions), so we decided to add these values to be part of the "West" region.

When looking at pctUrban, we noticed it resembled more of a boolean value with most values either being 0 or 100. We decided to turn this into a factor with values "Rural" or "Urban" for the rest of the analysis.

Some data is missing in the form of NA values, which was easy to spot. A small number of missing values (in ownHousMed and rentMed) in California were given dummy values instead of NA, which were not significant enough to alter the model, so those values were removed. There were also some repeated values and some 0 values that looked to be unrelated. For our further analysis we removed this data. No other values appeared more than 10 times for any variable so there are no other significant 'dummy' scores.

Most of the explanatory variables are non-linear when compared to the outcomes, violating the assumptions of linearity and mean zero errors for linear models. Moreover, the vertical spread seems to be vary as the variable value changes, showing heteroscedasticity. This again violates the assumption of constant variance for linear models.

1.2 Skew

To aid with the process of linear modelling, it is often helpful to transform the predictor variables to reduce skew and make the distribution close to normal. As a rough benchmark for acceptable skew, we came to the figure of ±0.7, as this would enable modelling which is accurate enough, while at the same time, it would not transform the data too drastically. (This figure can be changed very easily at the modeller's discretion)

6 of the 22 predictor variables were already within this range, and 3 variables were categorical, leaving us with 13 predictor variables to transform. Most were resolved with simple log or square root transforms.

Variable	Transformation
medIncome	log
pctLowEdu	log
pctCollGrad	log
pctUnemploy	log
ownHousMed	log
rentMed	log
popDensity	log
pctForeignBorn	log
pctNotHSgrad	square root
pctHousOccup	Raise to power of 9
pctKidsBornNevrMarr	log
ownHousQrange	Raise to power of -0.5
rentQrange	log

'PctVacantBoarded' has many outliers, all of which are highly correlated to high rates of crime. When these outliers are separately observed, a square root transform fits the rest of the data.

1.3 Outliers

There are many variables with many points either above the maximum percentile if it's positively skewed, or below the minimum percentile if it's negatively skewed. By definition of boxplots, these maximum and minimum percentile are the largest and lowest data point respectively that excludes any outliers, so the points mentioned previously are most likely outliers.

However, transformations that solve skewness can often solve outliers. So to distinguish whether or not they are truly outliers, we could use different methods, such as influenceIndexPlot or outlierTest. We could also talk to the client to clarify if the data is geniune, or alternatively use surveys to gather more data. There is more data for the lower values and less data for the higher values, so gathering more data might get more data on the high values to have a better non-skewed data set.

Most of the outliers are from urban areas. However, the distribution of Rural areas and Urban areas are similar, suggesting that although outliers are from Urban, not all data from Urban are outliers. One of the reasons could be that Urban has more data points, at least double of the data from Rural, and that might suggest the increase in number of potentital outliers from Urban.

1.4 Correlations

When looking at the correlations between variables we noticed that pctKidsBornNevrMarr had the greatest positive correlation to violentPerPop and nonViolPerPop at 0.74 and 0.56 respectively. Also, we observed that pctKids2Par had the greatest negative correlation to these outcome variables at -0.73 and -0.67 respectively. This tells us that these variables are heavily related to the outcome variables. We also saw that some variables are not correlated to the outcome variables, implying the relation between them is very weak, such as pctVacant6up and ownHousQrange.

We found all the pairs are highly correlated (Either positive or negative with a magnitude of greater than 0.7). From these pairs we can tell which variables are highly related with each other and can be used later when deciding what variables to omit. Using the similarly correlated pairs of variables we can see that some measure similar things. This means we don't need all of them in the model, otherwise there would be too many explanatory variables and would be too complex.

From the pairs we can see that pctNotHSgrad is correlated to the education-based variables so decide to keep that one and omit pctLowEdu, pctCollGrad. We also see that pctwdiv was related to a lot of variables and would be a good proxy measure for things like pctUnemploy and the education variables. medincome is highly correlated to to the housing variables rentmed and ownhousmed. When looking at pctKids2Par and pctKidsBornNevrMarr, we decided the latter was a better fit for modeling as it was more closely correlated to the outcome variables and not highly correlated to another variables.

1.5 Model Recomendations

When creating a linear model on the data we suggest that all the errors/missing data be removed. Also, chose variables that have strong relationship with the outcome variables and ones that are linear and have mean 0 error. We suggest that the data should be transformed as described above to achieve minimum skewness in the explanatory variables and homoscedasticity in the outcome variables.

We suggest we use the following for should be in our linear model for violent and non-violent crime: region, medIncome, pctWdiv, pctNotHSgrad, pctEmploy, pctKidsBornNevrMarr, pctHousOccup, pctHousOwnerOccup, pctVacantBoarded, rentMed, popDensity, pctForeignBorn

We removed state because there would be too many factor levels with too few data points, so there isn't enough data to make an accurate model.

2 Statistical methodology

Firstly we showed the libraries and initialized the data.

```
library(ggplot2)
library(VIM)
library(MASS)
library(car)
library(RColorBrewer)
library(e1071)
library(corrplot)
load("USACrime.Rda")
```

2.1 About the data and Missing Values

Take a first look at the dataset:
We looked at the head and tail of the data (Figure 1)

Outcome variables are violent crimes per 100k population (violentPerPop) and non-violent crimes per 100k population (nonViolPerPop). There are 22 potential explanatory variables.

State and region are factors with 48 levels and 5 levels respectively:

```
length(levels(USACrime$State))
```

```
## [1] 48
```

```
length(levels(USACrime$region))
```

```
## [1] 5
```

The remaining variables are numerical values. Variables medIncome, ownHousMed, ownHousQrange, rentMed, and rentQrange only take integer values and are coded as numerical. However, as they have many unique values, and due to the context of the data, they won't be coded as factors.

Another observation by looking at the head and tail of the data is the existence of missing values (NA). Moreover notice the pctUrban variable having mostly values of 0 and 100.

```
summary(USACrime$pctUrban)
```

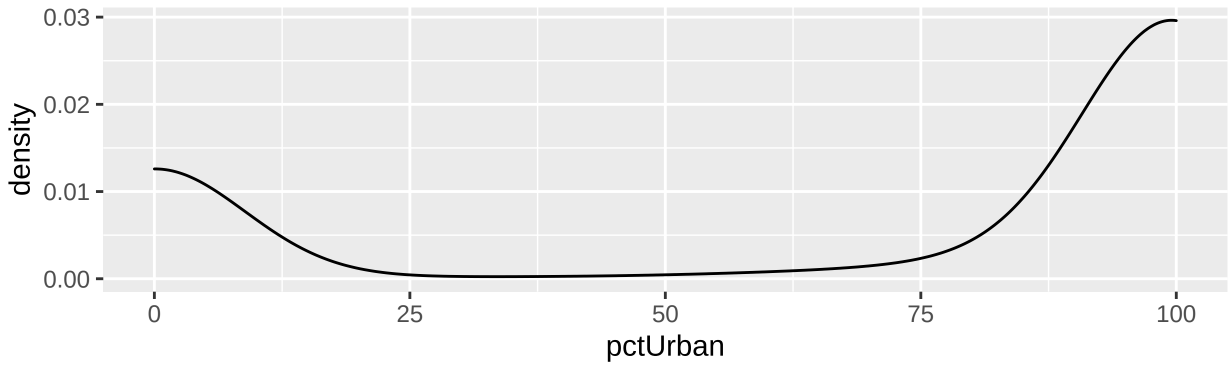
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00  100.00   69.96  100.00  100.00
```

```
tail(sort(table(USACrime$pctUrban)), decreasing = TRUE, n=2)
```

```
##
##    0  100
##  522 1104
```

Note that the median of pctUrban is 100. By summarising pctUrban, you will notice there are 522 occurences of zero and 1104 occurences of 100. By producing a density plot (Figure 4), we observe that the variable is bimodal. Hence, we'll make pctUrban a factor by separating values of less than 50 and more than 50 into "rural" and "urban" respectively:

```
ggplot(USACrime,aes(x = pctUrban, main = "Density of pctUrban"))+geom_density()
```



Furthermore, looking at the summary of the region variable spot that “Pacific” has only 3 observations:

```
summary(USACrime$region)
```

```
##   MidWest NorthEast   Pacific     South      West
##       280       589         3       585       445
```

This amount of data is insufficient for proper statistical analysis and could cause instability in fitted models. So, we’ll group it with “West”

```
USACrime$region[USACrime$region == "Pacific"] <- "West"
```

Check observations for State:

```
summary(USACrime$State)
```

```
##  AK  AL  AR  AZ  CA  CO  CT  DC  DE  FL  GA  IA  ID  IL  IN  KS  KY  LA  MA  MD
##   3  16  25  20 278  25  69   1   1  90  37   3   7   0  48   0  26  19 116   9
##  ME  MI  MN  MO  MS  NC  ND  NH  NJ  NM  NV  NY  OH  OK  OR  PA  RI  SC  SD  TN
##  17   0   6  42  19  46   8  21 211  10   4  29 106  36  31 100  26  28   8  33
##  TX  UT  VA  VT  WA  WI  WV  WY
## 152  24  33   0  39  59  14   7
```

Notice that some levels have 0 observations. Ignore those.
The summary of data shows missing values. There are 52 for medIncome and 21 for pctEmploy. (Figure 5)

These are mutually exclusive and no pattern is observed (Figure 6).

We produced plots to investigate the mechanism for these missing data. See how the boxplots of the variables pctEmploy and medIncome without the missing data (blue) are not much different than the ones involving the NAs (red), which implies data missing completely at random (MCAR).

Check summary statistics of the rest of the variables to look for abnormalities. Looking specifically at the min and max values we see some odd values for the variables ownHouseQrange, rentQrange, ownHousMed, rentMed. Investigating these variables (Figure 7), observe 5 zeros for ownHousQrange and 2 for rentQrange. These correspond to max values for ownHousMed and rentMed respectively. These could me missing data. It seems like they are not mutually exclusive and all of them are from California. (Figure 8)

This suggests a pattern in the missing data. Now investigate max values 500001 and 1001 for ownHousMed and rentMed respectively (Figure 9). There are 14 of each and some of them overlap. Maybe values above 500001 and 1001 for the respective variables were not recorded. Produce plots to suggest mechanism. See how the boxplots of the variables ownHousMed, ownHousQrange, rentMed and rentQrange without the missing data (blue) are quite different from the ones involving the NAs (red), which implies data missing at random (MAR) or data missing not at random (MNAR). We need to look at the data to distinguish between the 2 mechanisms. Below are the boxplots with regards to variables that have missing data. (Figure 5)

We think that all this data should be removed for the rest of the analysis as it will effect other measures and plots and we do not believe it to be accurate.

Change all this data to NA and remove NA's alltogether:

```
USACrime$ownHousMed[USACrime$ownHousMed==500001] <- NA
USACrime$ownHousQrange[USACrime$ownHousQrange==0] <- NA
USACrime$rentMed[USACrime$rentMed==1001] <- NA
USACrime$rentQrange[USACrime$rentQrange==0] <- NA
USACrime.2 <- na.omit(USACrime)
```

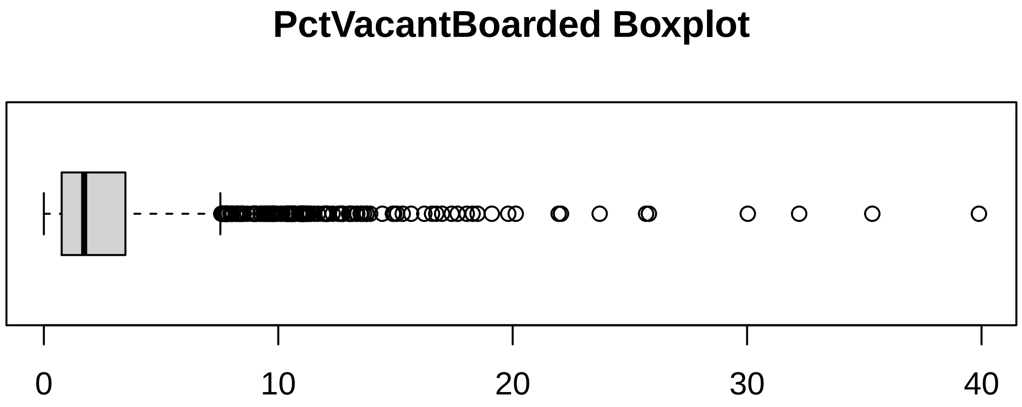
No more values appeared more than 10 times so no other significant "dummy" score. (Figure 10)

2.2 Skewness

To accurately fit a model, linearity is one of the key criteria that variables need to meet. This ensures that there is a linear relationship between each x variable and the y variable. However, if the x variables are skewed in some way, then this condition will not hold. To clean the data and make it easier to handle, removing skewness from the data is a good first step.

Initially, to just get a feel for how skewed each variable was, we plotted density functions, alongside the mean of the variable. This gave us a rough visual idea of which variables were heavily skewed, and which were acceptable. We then used the skewness function to quantify and put a definitive number on the skewness of each distribution.

We chose ± 0.7 as a benchmark, as we felt that transforming the variables to sit within that skewness was narrow enough to fit a linear model to, while not being so narrow that the 'flavour' of the data was not lost, that is – we are not losing the initial properties of the variable to get perfectly unskewed data.



A lot of the data was positively skewed, so we set up a for loop that tested the skewness of the data as it was, after a log transform, and after a square root transform. (Figure 11) This sorted out most of our problems, but there were still a few very skewed variables, for which this did not help. For these, we investigated different transformations, and settled on a power transform of -0.5 for "ownHousQrange", and a power transform of 9 for "pctHousOccup". "pctVacantBoarded" was still very problematic, and a boxplot illustrated why.

96% of the data lied between 0 and 10, while the maximum was 35. The outliers had a massive impact on the skewness and removing even just the 20 highest data points made a square root transform effective in reducing skew. In the top 20 values, comparing basic indicators showed that pctVacantBoarded values > 10 are correlated with very high rates of both violent and non-violent crime.

2.3 Outliers

From the boxplots (Figure 13), we were able to see the large amount of outliers. We can use transformations on the variables to reduce the skewness and the number of outliers, and by using influence index plot or outlier test to justify which observations are an outlier (Figure 14).

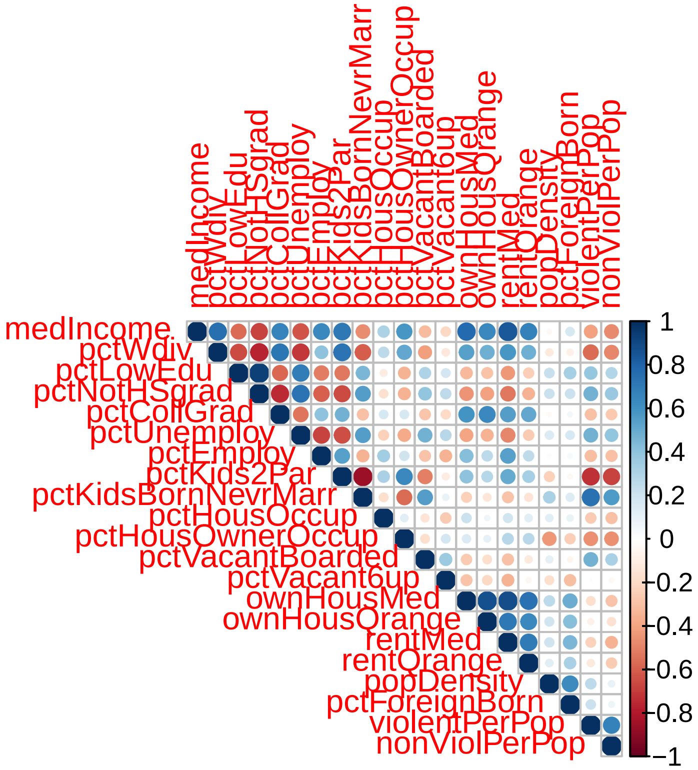
From the scatter plot of violentPerPop against medIncome (Figure 15), by using the identify function on the points that are far away from the best fitted line, we were able to see most of the points are from Urban, hinting that Urban might be an outlier. Note that we use only the variable "medIncome" to show the result, and the other variables are showed in a similar way.

From the bivariate boxplot model (Figure 16), it shows that Rural and Urban has similar distributions for "violentPerPop", so Urban isn't necessarily an outlier. The reason why the outliers are mostly from Urban could be a result from larger data set from Urban areas, and further investigation is required for better findings, such as surveys or checking if the data is genuine. Note that similar method can be used on nonViolPerPop.

2.4 Correlation

Using correlation between the variables and the outcome variables violentPerPop and nonViolPerPop we have decided that variables with low correlation should be not included in a model. For example pctpctVacant6up. (Figure 17)

Variables that are highly correlated can be used in place of each other. From the corrplot (CRAN.) and by looking at the correlations with magnitude over 0.7. (Figure 18) We found that the following pairs are highly correlated (Correlation with a magnitude of greater than 0.7) using the following code.



```
for (i in 4:24) {for(j in i:24){if (i!=j){
      if (abs(cor(USACrime.2[,i],USACrime.2[,j])) >= 0.7){
        print(paste(colnames(USACrime.2)[i],",",colnames(USACrime.2)[j], ",",
              cor(USACrime.2[,i],USACrime.2[,j])))
      }   }  }}
```

```
## [1] "medIncome , pctWdiv , 0.745151895991773"
## [1] "medIncome , pctKids2Par , 0.712009657396248"
## [1] "medIncome , ownHousMed , 0.767621184288536"
## [1] "medIncome , rentMed , 0.843844059106991"
## [1] "pctWdiv , pctNotHSgrad , -0.767154180384872"
## [1] "pctWdiv , pctCollGrad , 0.720216578014387"
## [1] "pctWdiv , pctUnemploy , -0.714310401329646"
## [1] "pctWdiv , pctKids2Par , 0.734528319639785"
## [1] "pctLowEdu , pctNotHSgrad , 0.933248470082486"
## [1] "pctNotHSgrad , pctCollGrad , -0.748345874714662"
## [1] "pctNotHSgrad , pctUnemploy , 0.730622995230578"
## [1] "pctKids2Par , pctKidsBornNevrMarr , -0.86056158603018"
## [1] "pctKids2Par , violentPerPop , -0.727725156658333"
## [1] "pctKidsBornNevrMarr , violentPerPop , 0.73606550097721"
## [1] "ownHousMed , ownHousQrange , 0.869650618532911"
## [1] "ownHousMed , rentMed , 0.882299988251323"
## [1] "ownHousMed , rentQrange , 0.735577791683381"
## [1] "ownHousQrange , rentMed , 0.705037318548239"
```

We then plotted some of the highly correlated variables with the regions highlighted and saw that the correlation was similar in all regions. So the general trend between the variables is the same as the regions. (Figure 19)

3 A paragraph or table on “authors’ contributions”

All members of the team work together at the start as we went over each area of the data and each contributed our own ideas to make sure we all understood the data and had ideas about each stage of the EDA. Then we split up and took the following areas of the EDA which best suited our expertness and interests.

1727716 - Helped with describing the initial look at the data and understanding it. Made some clarifications about the variables and observations. Was responsible for detecting missing values in the dataset by investigating curious scores.

1817945 - Completed work and write up on outliers of variables.

1804107 - Completed work and write up on skewness of variables. Also found that the NAs are unrelated. Helped create functions and pieces of code useful for other team members.

1802212 - Completed work and write up on correlation of variables and brought the ideas together to create the document.

We propose an even split of 100% each for our efforts.

4 References

CRAN. An Introduction to corrplot Package. Available: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html. Last accessed 6th Feb 2021.

Wikipedia: Boxplots. Available: https://en.wikipedia.org/wiki/Box_plot#:~:text=In%20descriptive%20statistics%2C%20a%
Last accessed 7th Feb 2021.

5 Appendix

Figure 1 - Head and Tail of the Data

```
##   State    region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad pctCollGrad
## 1    NJ NorthEast    Urban     75122   70.20      5.81         9.90       48.18
## 2    PA NorthEast    Urban     47917   64.11      5.61        13.72       29.89
## 3    OR      West    Urban     35669   55.73      2.80         9.09       30.13
## 6    MO   MidWest    Urban     21577   41.15      8.76        23.03       20.66
## 7    MA NorthEast    Urban     42805   47.70      4.49        13.89       27.01
## 8    IN   MidWest    Urban     23221   35.74     10.09        28.67       12.00
##   pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr pctHousOccup
## 1        2.70     64.55       90.17                0.36        98.37
## 2        2.43     61.96       85.33                0.24        97.15
## 3        4.01     69.80       78.85                0.88        95.68
## 6        5.72     59.02       69.79                1.58        91.81
## 7        4.85     65.42       79.76                1.18        95.11
## 8        8.19        NA       58.70                4.66        92.22
##   pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed ownHousQrange
## 1             91.01             3.12        37.50     262600        111000
## 2             84.88             0.00        18.33     164200         63600
## 3             57.79             0.92         7.54      90400         37300
## 6             55.50             2.09        26.22      53900         35400
## 7             56.96             1.41        34.45     179000         60400
## 8             63.82             6.39        56.36      37000         26100
##   rentMed rentQrange popDensity pctForeignBorn violentPerPop nonViolPerPop
## 1      NA        316     1845.9          10.66         41.02       1394.59
## 2     560        205     2186.7           8.30        127.56       1955.95
## 3     428        150     2780.9           5.00        218.59       6167.51
## 6     280        134     1995.7           1.49        442.95       6867.42
## 7     669        361     2643.5           9.19        226.63       1890.88
## 8     253        139     1515.3           0.87        439.73       4909.26
```

```
##      State    region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad
## 2210    NJ NorthEast    Urban     37664   48.61     12.84        27.73
## 2211    CA      West    Urban     24727   31.42     17.12        30.87
## 2212    LA     South    Urban     20321   33.25     12.51        27.71
## 2213    CA      West    Urban     27182   44.72      7.82        26.14
## 2214    TX     South    Rural     19899   21.94     24.37        39.63
## 2215    CA      West    Urban     23287   27.54     13.93        33.68
##      pctCollGrad pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr
## 2210        9.28        4.13     65.15       73.53                1.12
## 2211       15.79        9.99     55.53       64.81                4.49
## 2212       19.28        7.90     54.64       63.66                2.98
## 2213       12.42        5.18     50.54       74.20                1.60
## 2214       12.40       12.12     52.53       63.45                2.35
## 2215        8.86        9.27     53.35       60.23                4.85
##      pctHousOccup pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed
## 2210        97.03             70.65             0.00        35.71     144200
## 2211        96.40             44.63             1.46        13.18      91100
## 2212        89.72             54.24             4.59        46.08      52000
## 2213        93.30             76.81             0.84        29.47     123900
## 2214        85.39             58.39             5.61        67.21      37800
## 2215        94.85             55.68             3.67        28.67      87600
##      ownHousQrange rentMed rentQrange popDensity pctForeignBorn violentPerPop
## 2210         41500     553        207     4109.8           6.56        132.87
## 2211         47700     374        157     3365.4          18.90        545.75
## 2212         39100     248        121     1682.8           2.24        124.10
## 2213         72300     451        204     1195.2           7.35        353.83
## 2214         26100     227        182     2142.2           2.28        691.17
## 2215         46500     369        177     1331.0          16.49        918.89
##      nonViolPerPop
## 2210       1992.98
## 2211       7356.84
## 2212       5824.44
## 2213       4654.20
## 2214       5340.87
## 2215       8838.50
```

Figure 2 - Summary of the Data

```
summary(USACrime)
```

```
##      State           region     pctUrban      medIncome        pctWdiv
##  CA     :278   MidWest  :280   Rural: 546   Min.   : 12908   Min.   : 9.02
##  NJ     :211   NorthEast:589   Urban:1356   1st Qu.: 23811   1st Qu.:34.20
##  TX     :152   Pacific  :  0                Median : 31455   Median :42.43
##  MA     :116   South    :585                Mean   : 34074   Mean   :43.46
##  OH     :106   West     :448                3rd Qu.: 41649   3rd Qu.:52.49
##  PA     :100                                Max.   :123625   Max.   :89.04
##  (Other):939                                NA's   :52
##    pctLowEdu       pctNotHSgrad    pctCollGrad     pctUnemploy
##  Min.   : 0.200   Min.   : 2.09   Min.   : 1.63   Min.   : 1.320
##  1st Qu.: 4.720   1st Qu.:14.16   1st Qu.:14.08   1st Qu.: 4.090
##  Median : 7.860   Median :21.55   Median :19.68   Median : 5.470
##  Mean   : 9.432   Mean   :22.66   Mean   :23.03   Mean   : 6.013
##  3rd Qu.:12.148   3rd Qu.:29.63   3rd Qu.:29.00   3rd Qu.: 7.410
##  Max.   :49.890   Max.   :73.66   Max.   :73.63   Max.   :23.830
##
##    pctEmploy      pctKids2Par    pctKidsBornNevrMarr  pctHousOccup
##  Min.   :24.82   Min.   :26.11   Min.   : 0.000      Min.   :37.47
##  1st Qu.:56.42   1st Qu.:63.72   1st Qu.: 1.070      1st Qu.:90.95
##  Median :62.47   Median :72.23   Median : 2.065      Median :94.01
##  Mean   :61.89   Mean   :71.02   Mean   : 3.113      Mean   :92.69
##  3rd Qu.:67.58   3rd Qu.:79.98   3rd Qu.: 3.930      3rd Qu.:95.94
##  Max.   :84.67   Max.   :92.58   Max.   :24.190      Max.   :99.00
##  NA's   :21
##  pctHousOwnerOccup pctVacantBoarded  pctVacant6up     ownHousMed
##  Min.   :16.86     Min.   : 0.000   Min.   : 3.12   Min.   : 26600
##  1st Qu.:54.22     1st Qu.: 0.760   1st Qu.:24.57   1st Qu.: 56900
##  Median :62.11     Median : 1.720   Median :34.33   Median : 85350
##  Mean   :62.74     Mean   : 2.771   Mean   :35.02   Mean   :115151
##  3rd Qu.:71.82     3rd Qu.: 3.480   3rd Qu.:44.15   3rd Qu.:157050
##  Max.   :96.36     Max.   :39.890   Max.   :82.13   Max.   :497900
##                                                     NA's   :14
##  ownHousQrange       rentMed        rentQrange      popDensity
##  Min.   : 14600   Min.   :139.0   Min.   : 55.0   Min.   :   10
##  1st Qu.: 33100   1st Qu.:289.8   1st Qu.:140.0   1st Qu.: 1176
##  Median : 44800   Median :397.5   Median :175.0   Median : 2001
##  Mean   : 58807   Mean   :429.0   Mean   :202.2   Mean   : 2804
##  3rd Qu.: 69200   3rd Qu.:549.0   3rd Qu.:243.0   3rd Qu.: 3277
##  Max.   :331000   Max.   :999.0   Max.   :803.0   Max.   :44230
##  NA's   :5        NA's   :14      NA's   :2
##  pctForeignBorn   violentPerPop     nonViolPerPop
##  Min.   : 0.190   Min.   :   6.64   Min.   :  116.8
##  1st Qu.: 2.180   1st Qu.: 164.24   1st Qu.: 2913.3
##  Median : 4.575   Median : 369.31   Median : 4479.7
##  Mean   : 7.785   Mean   : 583.70   Mean   : 4942.3
##  3rd Qu.: 9.935   3rd Qu.: 792.69   3rd Qu.: 6271.7
##  Max.   :60.400   Max.   :4877.06   Max.   :27119.8
##
```

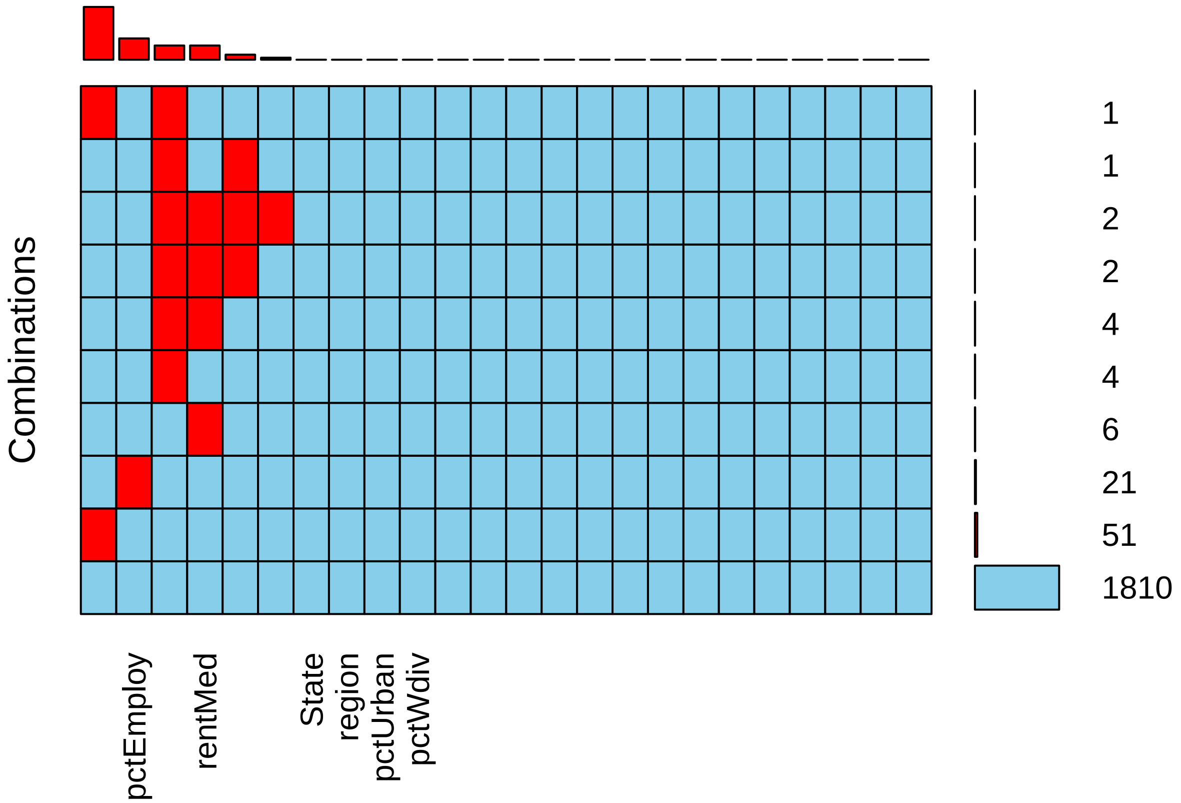
Figure 3 - pctUrban data

```
summary(USACrime$pctUrban)
```

```
## Rural Urban
##   546  1356
```

Figure 5 - Relationship between NA's

```
options(max.print = 3)
aggr(USACrime, prop=FALSE, combined=TRUE, numbers=TRUE, sortVars=TRUE, sortCombs=TRUE, axes=TRUE)
```



```
##
##  Variables sorted by number of missings:
##   Variable Count
##  medIncome    52
##  [ reached 'max' / getOption("max.print") -- omitted 23 rows ]
```

```
pbox(USACrime, pos=8)
```

```
## Warning in createPlot(main, sub, xlab, ylab, labels, ca$at): not enough space to
## display frequencies
```

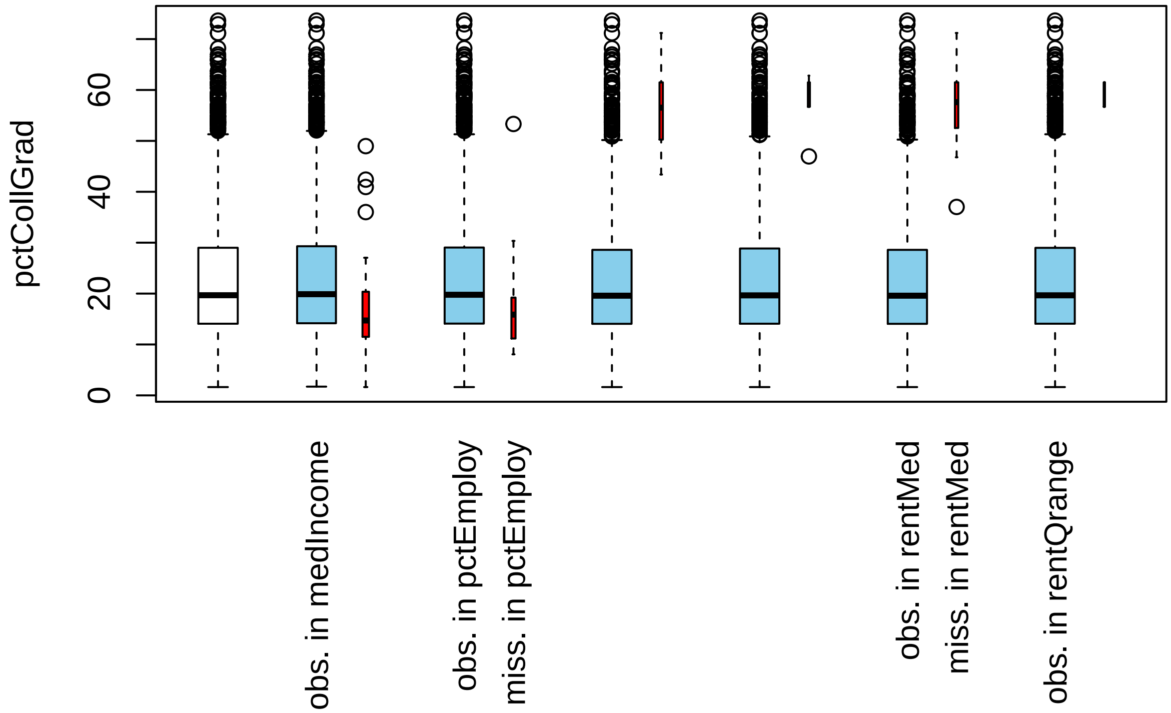


Figure 6 - The relationships between NA's in pctEmploy and medIncome

```
options(max.print = 20)
USACrime[is.na(USACrime$pctEmploy),]
```

```
##      State region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad pctCollGrad
##      pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr pctHousOccup
##      pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed ownHousQrange
##      rentMed rentQrange popDensity pctForeignBorn violentPerPop nonViolPerPop
##  [ reached 'max' / getOption("max.print") -- omitted 21 rows ]
```

```
USACrime[is.na(USACrime$medIncome),]
```

```
##      State region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad pctCollGrad
##      pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr pctHousOccup
##      pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed ownHousQrange
##      rentMed rentQrange popDensity pctForeignBorn violentPerPop nonViolPerPop
##  [ reached 'max' / getOption("max.print") -- omitted 52 rows ]
```

```
filter(is.na(USACrime$pctEmploy),is.na(USACrime$medIncome))
```

```
## Time Series:
```

```
## Start = 1
## End = 1902
## Frequency = 1
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [ reached getOption("max.print") -- omitted 1882 entries ]
```

Figure 7 - The 0s and 500001 in ownHousQrange, ownHousMed, rentQrange, rentMed)

```
sort(USACrime$ownHousQrange, n=10)
```

```
##  [1] 14600 15100 15300 15500 16600 17400 17500 18000 18400 18500 18800 19000
## [13] 19200 19200 19300 19600 19600 19700 19700 19700
##  [ reached getOption("max.print") -- omitted 1882 entries ]
```

```
sort(USACrime$ownHousMed,decreasing = TRUE)
```

```
##  [1] 497900 471300 463600 462500 461500 457800 452500 445300 440800 439900
## [11] 435700 432200 431000 422400 414300 411700 410200 408800 408200 407500
##  [ reached getOption("max.print") -- omitted 1868 entries ]
```

```
sort(USACrime$rentQrange)
```

```
##  [1] 55 71 76 79 80 84 85 87 88 88 91 93 93 96 96 96 97 97 98 98
##  [ reached getOption("max.print") -- omitted 1880 entries ]
```

```
sort(USACrime$rentMed,decreasing = TRUE)
```

```
##  [1] 999 995 994 989 961 958 947 946 943 935 934 930 927 916 913 902 902 902 900
## [20] 899
##  [ reached getOption("max.print") -- omitted 1868 entries ]
```

Figure 8 - Sates where ownHousQrange = 0 are all CA

```
USACrime[USACrime$ownHousQrange==0,]
```

```
##      State region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad pctCollGrad
##      pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr pctHousOccup
##      pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed ownHousQrange
##      rentMed rentQrange popDensity pctForeignBorn violentPerPop nonViolPerPop
##  [ reached 'max' / getOption("max.print") -- omitted 5 rows ]
```

Figure 9 - Data where ownHousMed = 500001 and rentMed = 1001

```
USACrime[USACrime$ownHousMed==500001,]
```

```
##      State region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad pctCollGrad
##      pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr pctHousOccup
##      pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed ownHousQrange
##      rentMed rentQrange popDensity pctForeignBorn violentPerPop nonViolPerPop
##  [ reached ’max’ / getOption("max.print") -- omitted 14 rows ]
```

```
USACrime[USACrime$rentMed==1001,]
```

```
##      State region pctUrban medIncome pctWdiv pctLowEdu pctNotHSgrad pctCollGrad
##      pctUnemploy pctEmploy pctKids2Par pctKidsBornNevrMarr pctHousOccup
##      pctHousOwnerOccup pctVacantBoarded pctVacant6up ownHousMed ownHousQrange
##      rentMed rentQrange popDensity pctForeignBorn violentPerPop nonViolPerPop
##  [ reached ’max’ / getOption("max.print") -- omitted 14 rows ]
```

Figure 10 - Checking number of repeated values

```
for (i in 4:24){
print(colnames(USACrime)[i])
print(head(sort(table(USACrime[,i]),decreasing = TRUE)))
}
```

```
## [1] "medIncome"
##
## 23819 27095 33273 14579 17103 17826
##     3     3     3     2     2     2
## [1] "pctWdiv"
##
## 41.65 35.73  36.1 38.24 40.13 45.85
##     6     4     4     4     4     4
## [1] "pctLowEdu"
##
## 5.78 5.02 7.74  2.4 2.53 3.42
##    9    7    7    5    5    5
## [1] "pctNotHSgrad"
##
## 11.27 28.02  10.4 11.77 12.19 14.97
##     5     5     4     4     4     4
## [1] "pctCollGrad"
##
##  14.2 11.26  8.91 11.45 11.63 12.18
##     6     5     4     4     4     4
## [1] "pctUnemploy"
##
##  4.6 4.36 5.21 5.41 7.84 3.91
##    9    8    8    8    8    7
## [1] "pctEmploy"
##
##  62.6 61.46 65.69 71.79 51.51  57.2
##     6     5     5     5     4     4
## [1] "pctKids2Par"
##
```

```
##    85 63.25 64.84 65.81 69.46 81.76
##     5     4     4     4     4     4
## [1] "pctKidsBornNevrMarr"
##
## 1.15  1.2 1.26 0.97 1.12 1.43
##   12   11   11   10   10   10
## [1] "pctHousOccup"
##
## 95.38 92.29 95.42 95.53 97.03 89.33
##     8     7     7     7     7     6
## [1] "pctHousOwnerOccup"
##
## 56.17 56.64 57.33  58.1 63.34 83.73
##     4     4     4     4     4     4
## [1] "pctVacantBoarded"
##
##    0 0.61 0.35 0.47 0.58 0.28
##  109   12   10   10   10    9
## [1] "pctVacant6up"
##
##  37.5 38.05 38.32 44.25 16.67 20.97
##     5     4     4     4     3     3
## [1] "ownHousMed"
##
## 42300 71500 49800 44000 50800 66000
##     8     8     7     6     6     6
## [1] "ownHousQrange"
##
## 28100 32000 43800 32600 33200 35400
##    10    10    10     9     9     9
## [1] "rentMed"
##
## 316 283 248 255 261 282
##  14  11  10   9   9   9
## [1] "rentQrange"
##
## 139 153 148 133 144 130
##  26  25  24  21  21  20
## [1] "popDensity"
##
## 347.3 456.3 503.3 503.9 614.6 642.6
##     2     2     2     2     2     2
## [1] "pctForeignBorn"
##
## 1.43 2.97 1.59 1.78 4.19 0.63
##    8    8    7    7    7    6
## [1] "violentPerPop"
##
## 223.06  28.45  76.24  84.88 103.33 105.23
##      3      2      2      2      2      2
## [1] "nonViolPerPop"
##
## 2246.14 4295.96 5613.23  116.79  421.09  453.77
##       2       2       2       1       1       1
```

Figure 11 - Best Transformations for Positive Skewness

```r
c <- c(4,5,6,7,8,9,10,11,13,14,16,17,19,21,22,23,24)
for (i in c){
   a <- abs(skewness(USACrime.2[,i]))
   b <- abs(skewness(log(USACrime.2[,i])))
   c <- abs(skewness(sqrt(USACrime.2[,i])))
   z <- NA
   y <- min(a,b,c,na.rm = TRUE)
   if (y == a){
      z <- "normal"
   }
   else if(y == b){
      z <- "log"
   }
   else{
      z <- "sqrt"
   }
   print(paste(colnames(USACrime.2[i]),z,min(a,b,c)))
}
```

```
## [1] "medIncome log 0.17645171928677"
## [1] "pctWdiv sqrt 0.188957628409267"
## [1] "pctLowEdu log 0.373926835746847"
## [1] "pctNotHSgrad sqrt 0.0890612251910951"
## [1] "pctCollGrad log 0.104459710706259"
## [1] "pctUnemploy log 0.0261841977354174"
## [1] "pctEmploy normal 0.461405903212213"
## [1] "pctKids2Par normal 0.646959774079886"
## [1] "pctHousOccup normal 3.4706455458993"
## [1] "pctHousOwnerOccup normal 0.0666874356227232"
## [1] "pctVacant6up sqrt 0.0901690777915914"
## [1] "ownHousMed log 0.353655516481908"
## [1] "rentMed log 0.048240940769942"
## [1] "popDensity log 0.253882851324813"
## [1] "pctForeignBorn log 0.0889746832148378"
## [1] "violentPerPop log 0.33391974391436"
## [1] "nonViolPerPop sqrt 0.477175748987812"
```

Figure 13 - Boxplot of medIncome

```r
boxplot(USACrime.2$medIncome, main="Boxplot of medIncome", horizontal=TRUE)
```

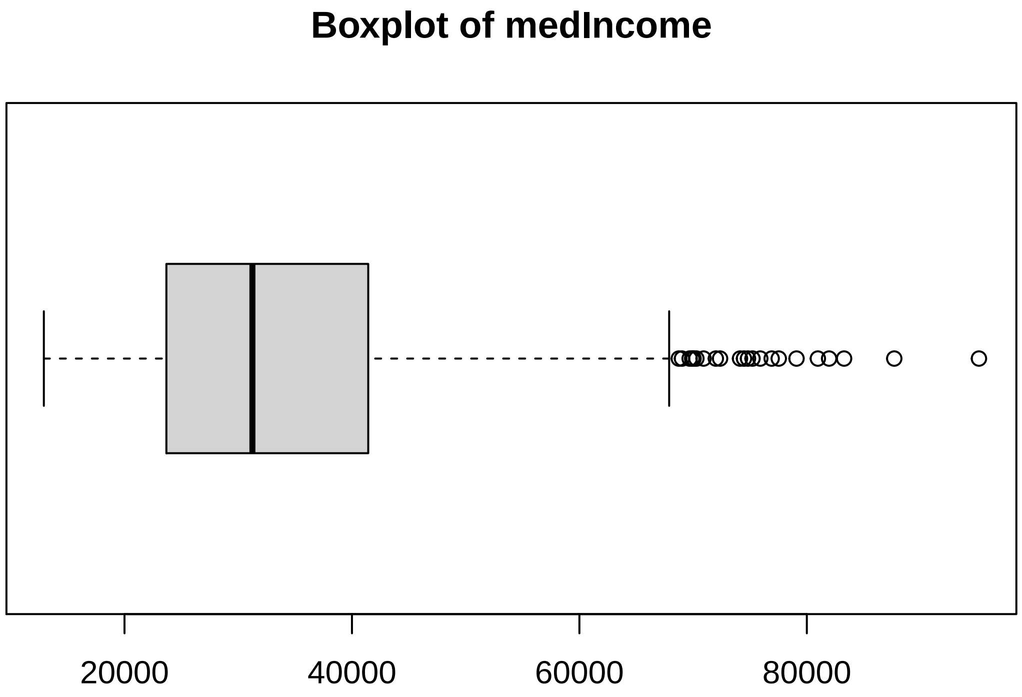
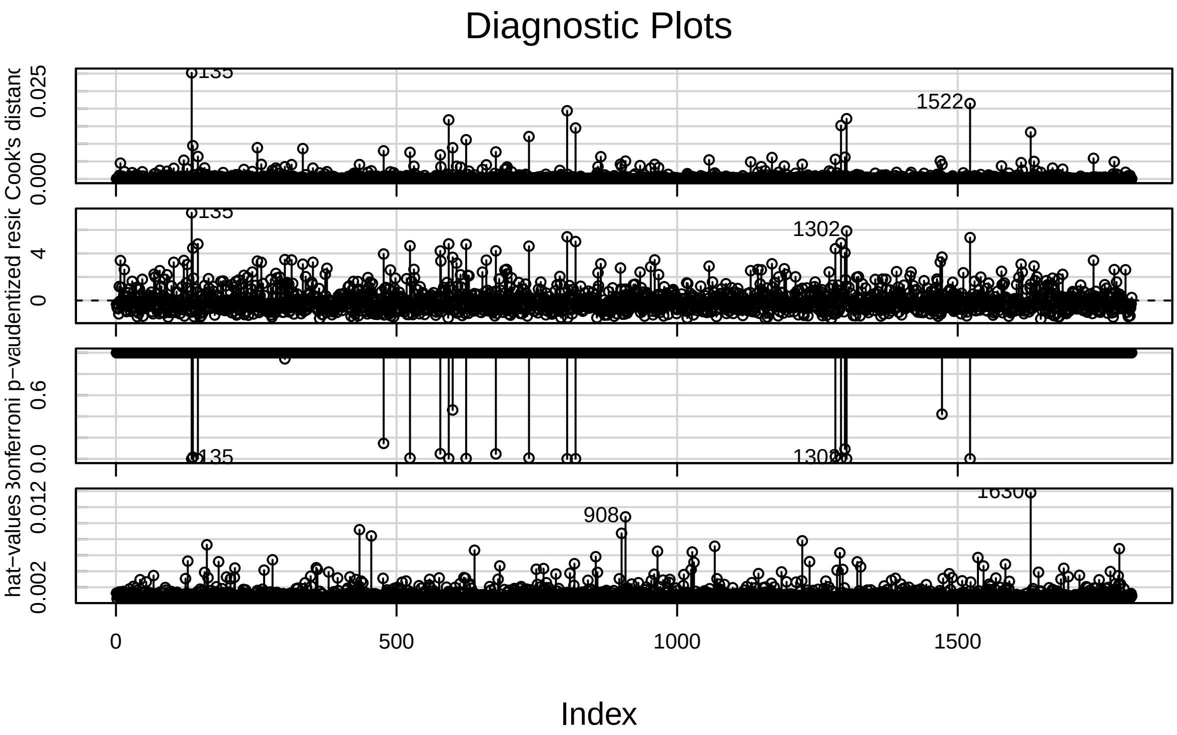


Figure 14 - InfluenceIndexPlot and OutlierTest method

```r
influenceIndexPlot(lm(USACrime.2$violentPerPop ~ USACrime.2$medIncome))
```

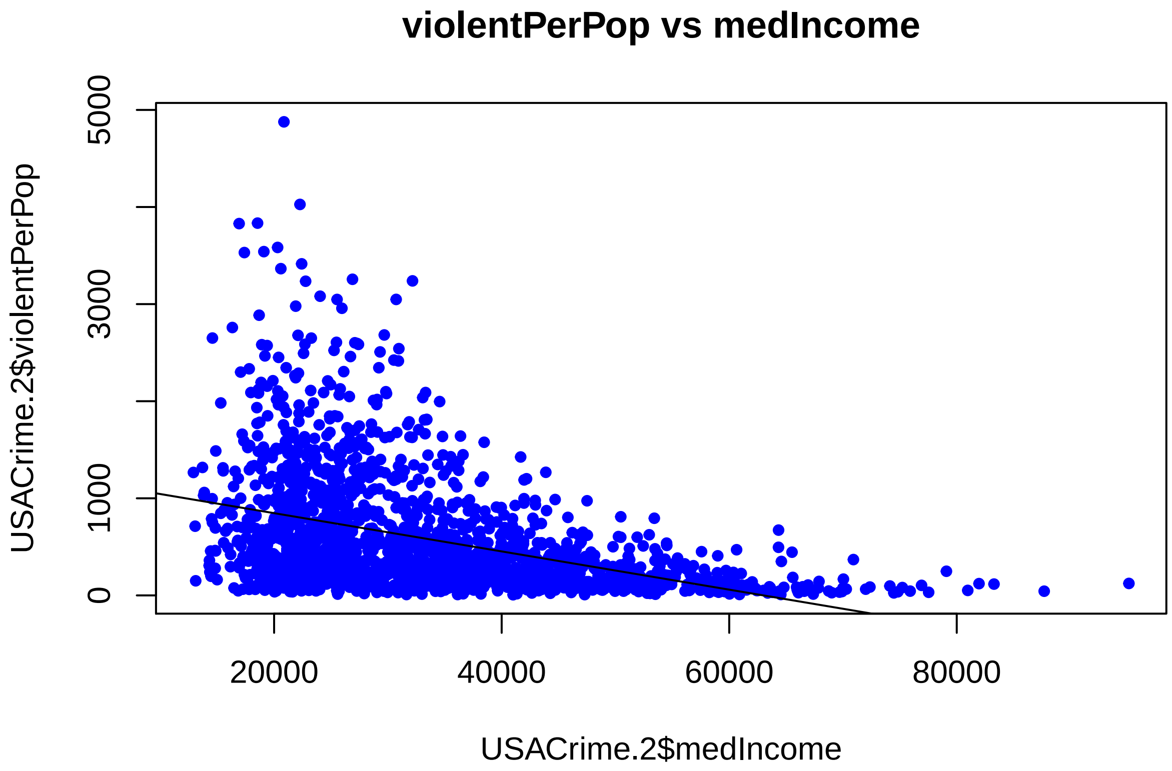


```
outlierTest(lm(USACrime.2$violentPerPop ~ USACrime.2$medIncome))
```

```
##      rstudent unadjusted p-value Bonferroni p
## 135  7.453466         1.4025e-13   2.5385e-10
## 1302 5.904982         4.2035e-09   7.6084e-06
## 804  5.410707         7.1152e-08   1.2878e-04
## 1522 5.342805         1.0311e-07   1.8664e-04
## 819  5.009721         5.9830e-07   1.0829e-03
## 1292 4.886482         1.1172e-06   2.0222e-03
##  [ reached 'max' / getOption("max.print") -- omitted 4 rows ]
```

Figure 15 - Bivariate model of violentPerPop against medIncome

```
plot(USACrime.2$violentPerPop ~ USACrime.2$medIncome, main="violentPerPop vs medIncome", pch=20, col="bl
abline(lm(USACrime.2$violentPerPop ~ USACrime.2$medIncome),col="black",lwd=1)
identify(USACrime.2$medIncome, USACrime.2$violentPerPop, labels=USACrime.2$pctUrban, plot=TRUE, cex=0.8
```



```
## integer(0)
```

Figure 16 - Boxplot of violentPerPop against pctUrban

```
boxplot(USACrime.2$violentPerPop ~ USACrime.2$pctUrban, main="Boxplots of violentPerPop vs pctUrban",lw
```

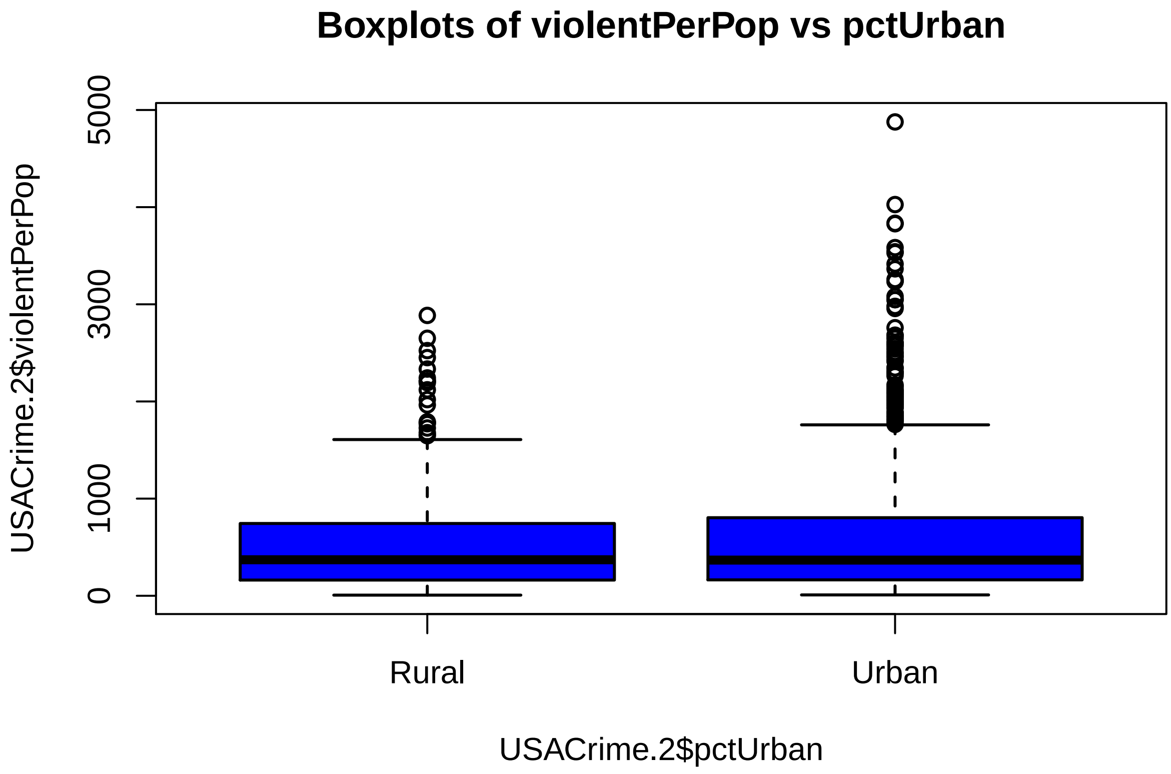


Figure 17 - Correlation between outcome variables and decision variables

```
options(max.print = 50)
corr <- round(cor(USACrime.2[,4:24]),2)
corr.2 <- corr[1:19,20:21]
corr.2
```

```
##                      violentPerPop nonViolPerPop
## medIncome                    -0.41         -0.48
## pctWdiv                      -0.56         -0.49
## pctLowEdu                     0.38          0.29
## pctNotHSgrad                  0.47          0.37
## pctCollGrad                  -0.30         -0.27
## pctUnemploy                   0.47          0.39
## pctEmploy                    -0.31         -0.30
## pctKids2Par                  -0.73         -0.67
## pctKidsBornNevrMarr           0.74          0.56
## pctHousOccup                 -0.26         -0.30
## pctHousOwnerOccup            -0.46         -0.46
## pctVacantBoarded              0.47          0.32
## pctVacant6up                  0.01         -0.04
## ownHousMed                   -0.18         -0.29
## ownHousQrange                -0.08         -0.16
## rentMed                      -0.23         -0.34
## rentQrange                     -0.12         -0.26
## popDensity                      0.26          0.09
## pctForeignBorn                  0.21          0.07
```

Figure 18 - Corrplot of all numeric variables

```
corr <- round(cor(USACrime.2[,4:24]),2)
corrplot(corr, type="upper")
```

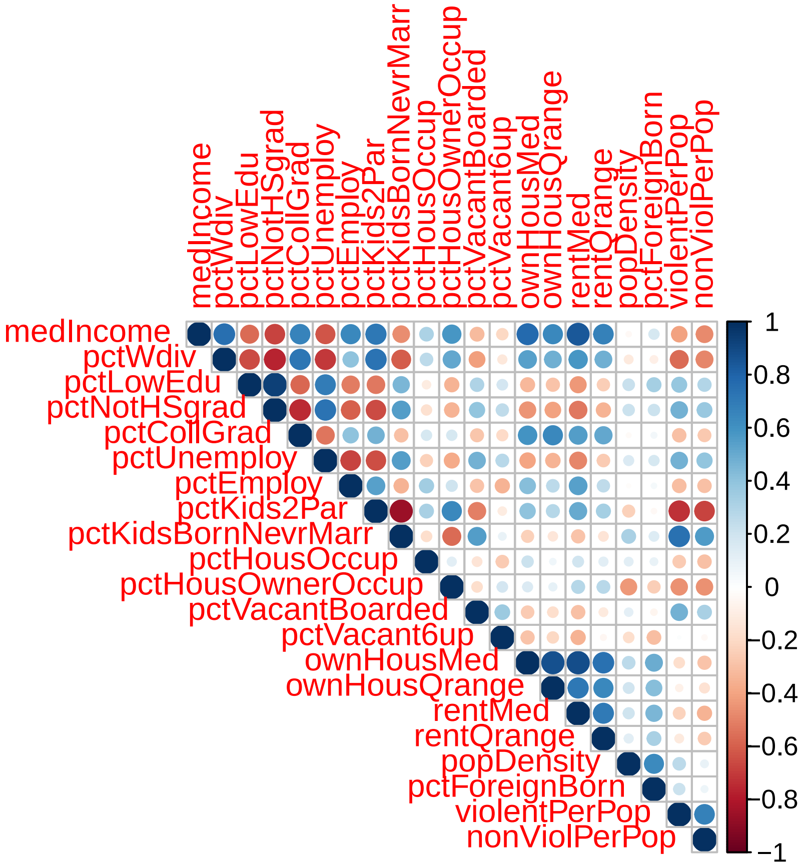
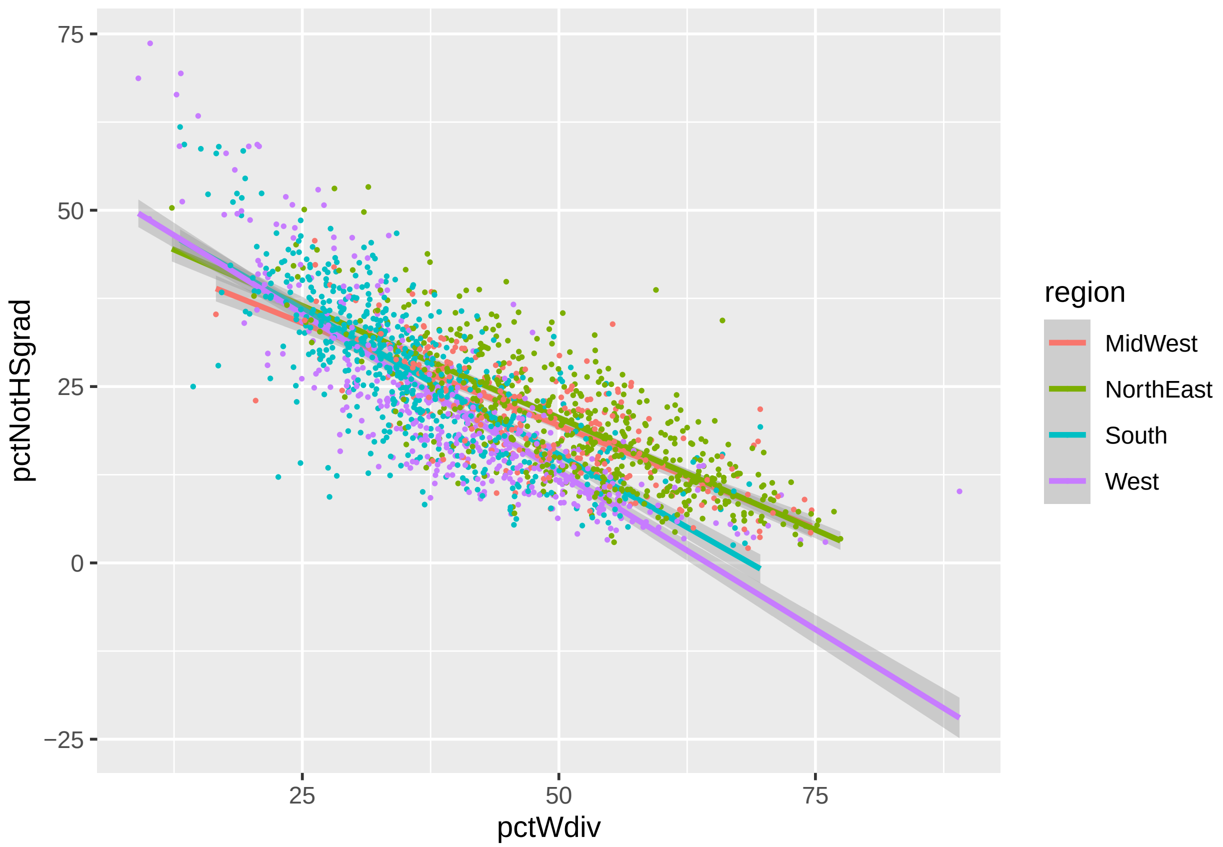


Figure 19 - Plots of highly correlated variables

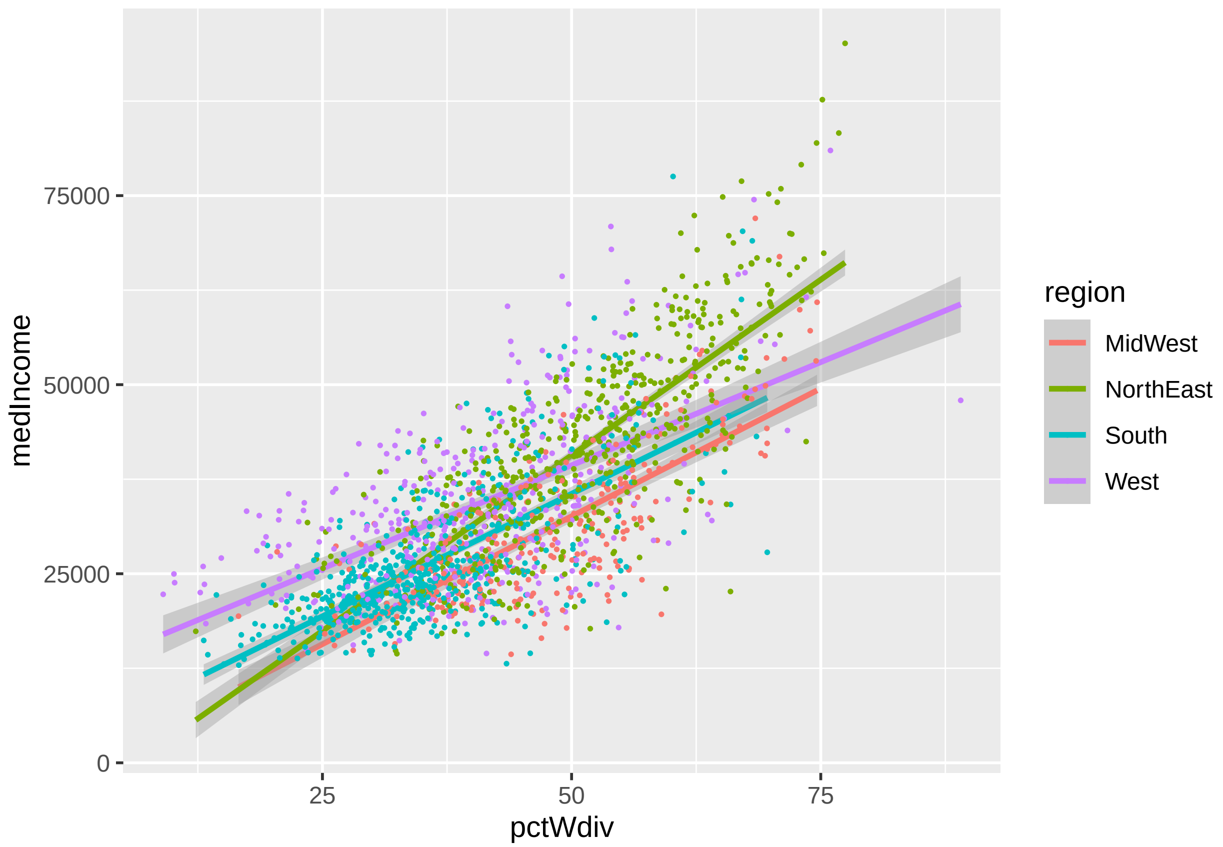
```
USACrime.wr <- USACrime.2[USACrime.2$region!="Pacific",]
ggplot(USACrime.wr, aes(pctWdiv, pctNotHSgrad, col=region)) + geom_smooth(method = "lm")+geom_point(siz
```

```
## `geom_smooth()` using formula 'y ~ x'
```



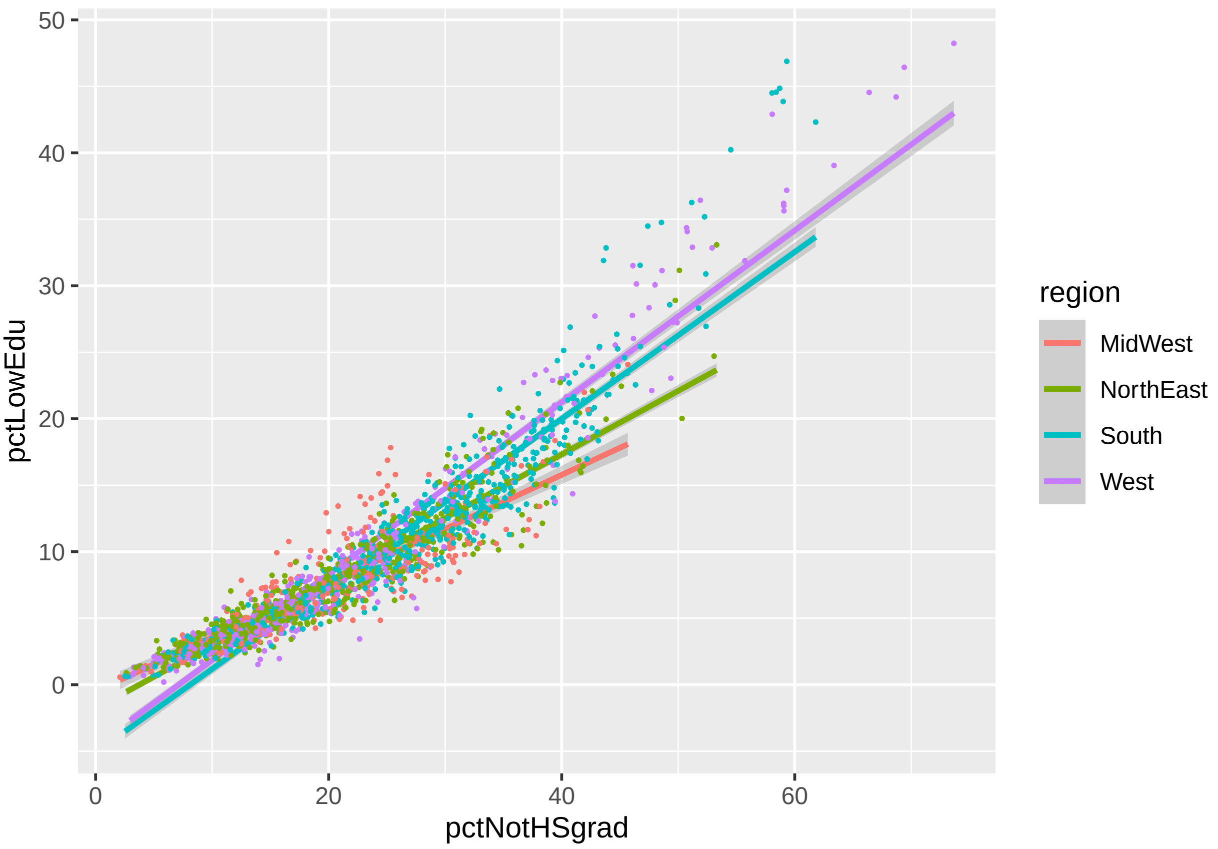
```
ggplot(USACrime.wr, aes(pctWdiv, medIncome, col=region)) + geom_smooth(method = "lm") + geom_point(size
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



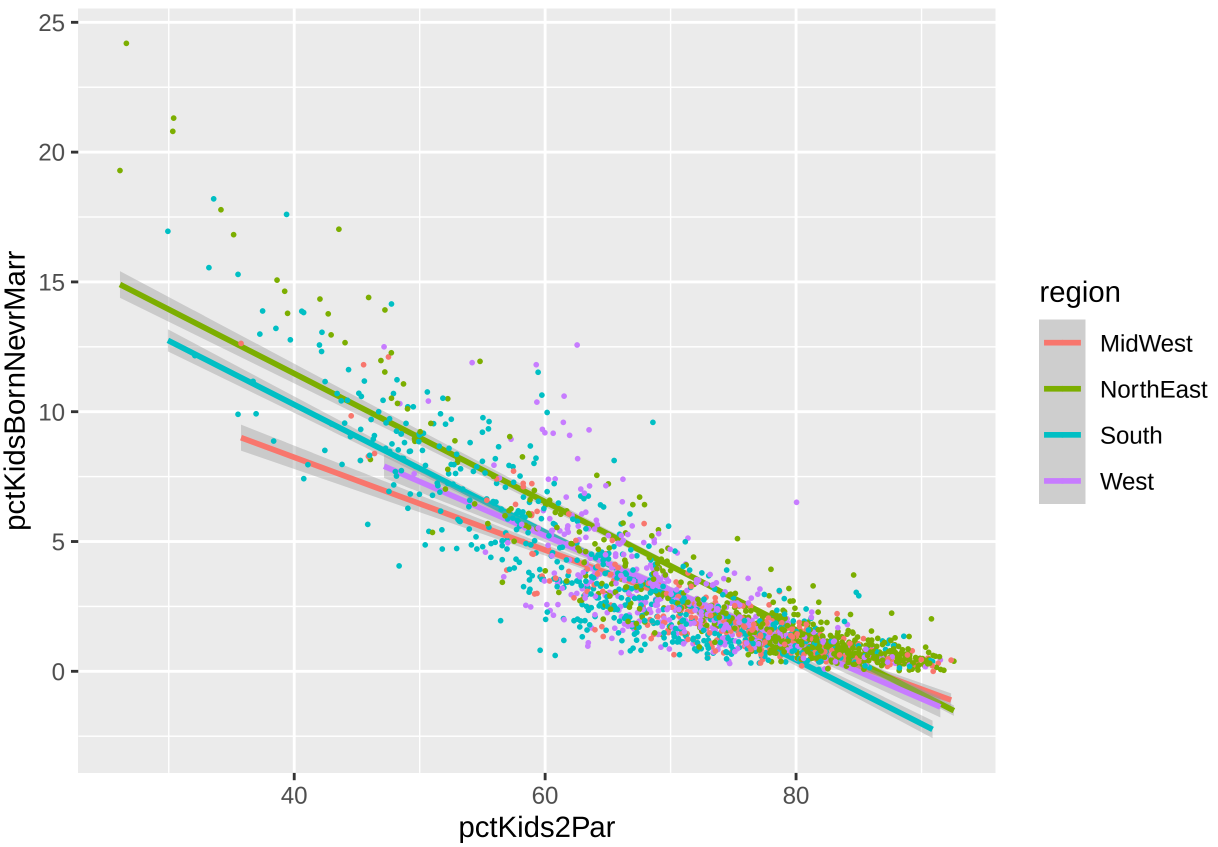
```
ggplot(USACrime.wr, aes(pctNotHSgrad, pctLowEdu, col=region)) + geom_smooth(method = "lm") + geom_point
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
ggplot(USACrime.wr, aes(pctKids2Par, pctKidsBornNevrMarr, col=region)) + geom_smooth(method = "lm") + ge
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```


25
20
15
10
5
0
pctKidsBornNevrMarr
40
60
80
pctKids2Par
region
MidWest
NorthEast
South
West