1.Executive summary

After completing our data analysis our main findings were

- High correlation exists between certain predictor variables such as, medIncome with rentMed, pctLowEdu with pctNotHSgrad, pctNotHSgrad with pctCollGrad, pctKids2Par with pctKidsBornNevrMarr, ownHousMed with rentMed.
- Predictor variables with the strongest relationship with crime rates[1] are pctHousOwnerOccup, pctNotHSGrad, medIncome, pctKids2Par and pctKidsBornNevrMarr.
- We recommend the following transformations: inversing medIncome, 1/4 power transformation to pctLowEdu, 1/2 power transformation to pctCollGrad, $\log_{10}$ transformation to pctKidsBornNevrMarr, popDensity and pctForeignBorn.

2.Findings

After having an initial look at the data, we found there were missing values for medIncome and pctEmploy. We noticed that there is no continuous predictor variable that has a glaring difference in the boxplots between missing and non-missing values. As for the categorical variables, mosaic plots showed missing and non-missing values have similar distributions. We concluded that that these values are likely to be missing completely at random (MCAR). Therefore, a potential method to resolve the problem of the missing data points is by fitting a linear model for medIncome and pctEmploy against all, or a selected subset of predictor variables, then fill in the missing values with the fitted values.

Secondly, there appeared to be a maximum value set for ownHousMed and rentMed of $500,001 and $1,001, respectively. This explained why some of the datapoints had an interquartile range of $0 for occupied housing value and rent. All 5 of these problematic data values were in the state of California, however since the data was large enough, we removed these datapoints by filtering.

Next for Housing Occupancy[2], we found that the pctHousOwnerOccup has a negative correlation with crime. The negative effect is stronger with non-violent crimes compared to violent crimes. A log transformation on the outcome variables, violentPerPop and nonViolPerPop appears to make the relationship more homoscedastic.

The variables ownHousMed and ownHousQrange do not appear to have much of an effect on crime.

[1] This group contains the variables violentPerPop and nonViolPerPop
[2] This groups contains the variables pctHousOccup, pctHousOwnerOccup, pctVacantBoarded and pctVacant6up.

For Population variables[3], we can see that the South has the highest instances of crime amongst the regions, followed by the West. When we looked at the outcome variables against the states, there weren't any particular states that caused the region factors to be the way that they are. For popDensity and pctForeignBorn, a log transformation could be used to reduce the strong positive skew, as well as the slight heteroscedasticity. Despite this, we find that neither has any particular relationship with crime.

Among Income & Employment variables[4], medIncome and pctWdiv have the strongest relationship with crime. MedIncome shows an inverse relationship, and pctWdiv shows a faint linear relationship.

For Education variables[5], pctNotHSgrad and pctLowEdu has a positive linear relationship with crime. PctCollGrad has a negative linear relationship with crime. A 1/4 power transformation reduces skewness for pctLowEdu, and a 1/2 power transformation reduces skewness for pctCollGrad. PctNotHSgrad is the most correlated with crimes, and since it has a high correlation with pctLowEdu and pctCollGrad, we recommend to only include pctNotHSgrad among these three predictor variables.

For the Family Background variables[6], pctKidsBornNevrMarr was positively skewed. This can be resolved by offsetting 0% values by a small positive number such as 0.00001% and then taking a $\log_{10}$ transformation. pctKidsBornNevrMarr showed a positive linear relationship with crime. PctKids2Par was slightly negatively skewed and showed a negative linear relationship with crime. We found they were highly correlated so we recommend to include at most one predictor variable among the two when fitting a linear model.

Finally, the two outcome variables, violentPerPop and nonViolPerPop, showed similar relationships against the majority of the predictor variables.

3.Statistical methodology

3.1 Data Cleaning

After an initial look at the dataset, we spotted a few red flags. The first red flag was in the State variable. Upon close inspection, there were 7 states without any data points (Hawaii, Illinois, Kansas, Michigan, Montana, Nebraska, Vermont) [Fig 3.1.1]. It is noted in the dataset information that there are

[3] Contains the variables state, region, %urban population density and %foreign born.
[4] Contains variables median income, % with investment or rent income, % unemployed, % employed.
[5] Contains the variables % low education, % not high school graduate, % college graduate.
[6] Contains variables % kids with two parents and %born and never married.

controversies in some states about the counting of rapes, especially in Midwestern states, and so there are missing values for violent crimes. 4 out of the 7 states that are missing are in the Midwest region and so this could be part of the reason for their omission.

Another issue was significant number of states with low number of data points [Fig 3.1.1]. Furthermore, when grouping we noticed that Pacific only had 3 datapoints all in the state Alaska. This would cause instability in our model due to the lower number of data points [Fig 3.1.1]. To prevent this, we merged Pacific with West region due to its close geographical location.

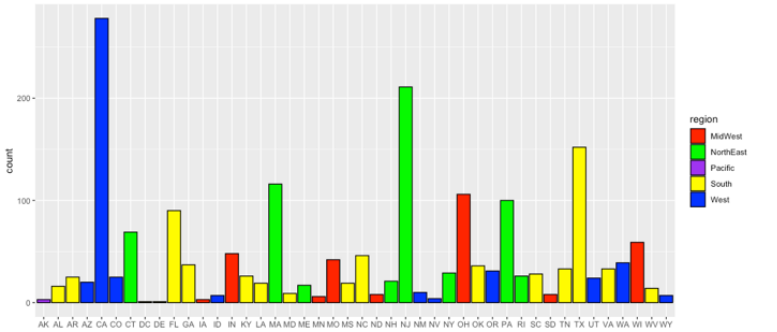


Fig 3.1.1

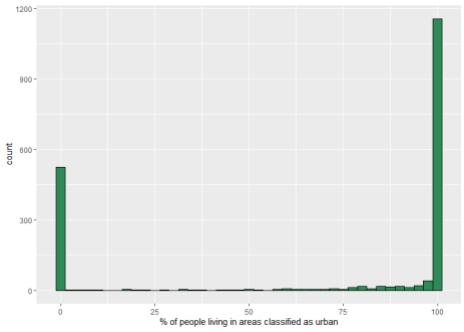


Fig 3.1.2

The histogram of the pctUrban variable shows that the majority of data points are either 0% or 100% [Fig 3.1.2]. Because of this, we recommend that pctUrban needs to be changed to a factor with levels 0%, 1-25%, 25-50%, 50-75%, 75%-99% and 100%.

We also found that there were 5 datapoints with value 0 for ownHousQrange [Fig 3.1.3]. Moreover, 2 of these datapoints also has value 0 for rentQrange. The maximum value of ownHousMed is 500,001, with 11 datapoints having this value [Fig 3.1.3]. Importantly, all 5 datapoints with value 0 for ownHousQrange has a ownHousMed value of 500,001 [Fig 3.1.3]. This suggests these houses could be worth more than $500,001. A similar phenomenon occurred for rentQrange and rentMed with a max value of 1,001 [Fig 3.1.4]. Hence this is a plausible explanation for the resulting zero Interquartile range. These outlier values only occur in the state of California. Since the data was large enough, we decided to remove these datapoints anyway even though the data points aren't missing completely at random.

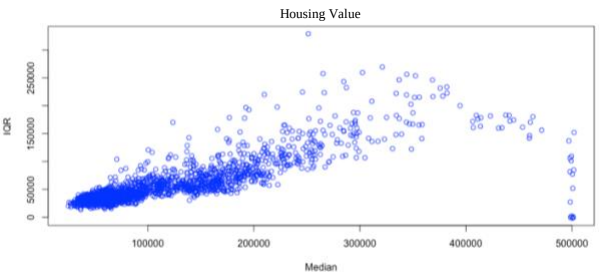


Fig 3.1.3

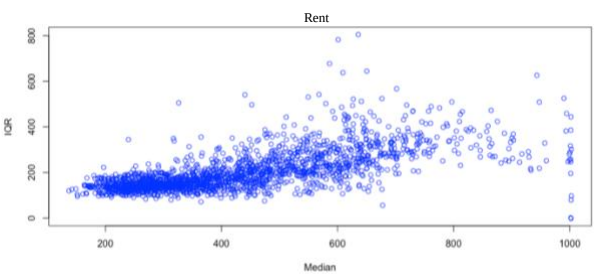


Fig 3.1.4

Lastly, we found there were 73 values in our dataset denoted N/A. The following boxplots and mosaic plots [Fig 3.1.5 – 3.1.7] are a few examples of plots we produced to compare the distribution of the missing values with the rest of the data. There appeared to be little variation. Hence, we conclude that the data is missing completely at random.

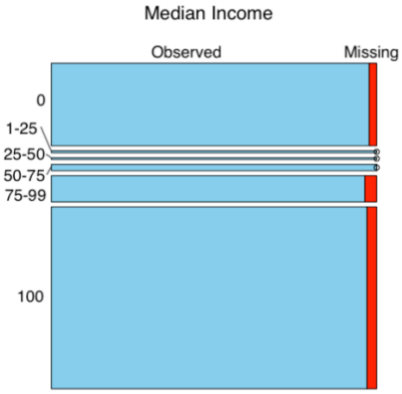


Fig 3.1.5

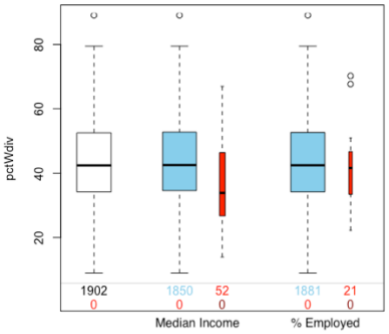


Fig 3.1.6

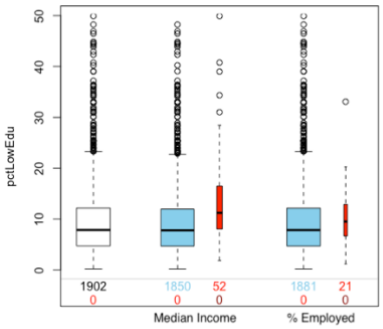


Fig 3.1.7

3.2 Correlation

Careful examination of [Fig 3.2.1] leads to the conclusion that strongly correlated explanatory variables include medIncome with rentMed, pctLowEdu with pctNotHSgrad, pctNotHSgrad and pctCollGrad, pctKids2Par with pctKidsBornNevrMarr, ownHousMed with rentMed.

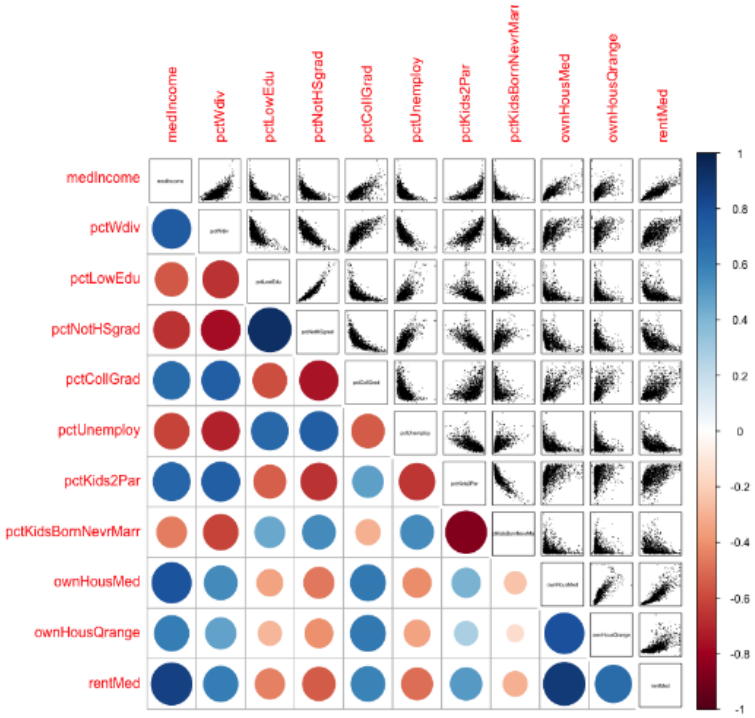
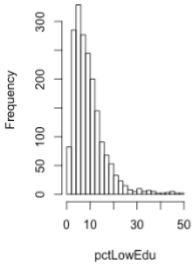


Fig 3.2.1

3.3 Relationship with crime

3.3.1 Education & Parents' Marriage

Firstly, we used histograms to check the skewness of variables and used a normal distributed curve to check the transformations. We can use a 0.25 power transformation to reduce skewness of percentage of Low Grades [Fig 3.3.1.1].



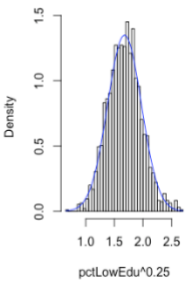


Fig 3.3.1.1

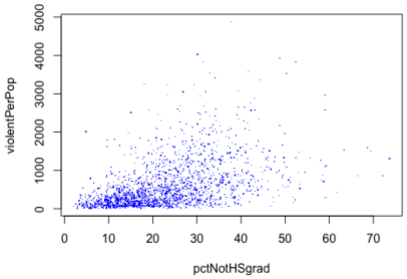


Fig 3.3.1.2

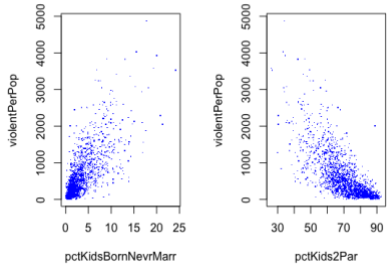


Fig 3.3.1.3

PctNotHSgrad has a correlation value of 0.47 with violentPerPop, with a positive linear relationship [Fig 3.3.1.2], which is the strongest among education variables. Thus, we consider it the major predictor in education.

For pctKids2Par and pctKidsBornNevrMarr, their correlation coefficients with violentPerPop are -0.73 and 0.74 respectively [Fig 3.3.1.3]. Since these two variables are highly correlated with each other, we recommend only including pctKidsBornNevrMarr when fitting a linear model.

The plots and findings are similar when considering nonViolPerPop instead.

3.3.2 Housing Occupancy

We started by looking at the relationship between housing occupancy variables. PctVacantBoarded appears to be positively skewed [Fig 3.3.2.1]. Offsetting 0 values by a small positive number and taking a log transformation improves this.

We then looked at how these variables affected crime rates. there appear to be little to no correlation between pctHousOccup and crime, however there is some weak correlation between pctHousOwnerOccup and crime [Fig 3.3.2.2].

This negative linear relationship appeared to be stronger for nonViolPerPop in comparison to violentPerPop. By adding a log transformation, a more homoscedastic relationship is obtained [Fig 3.3.2.3]. The correlation observed for pctVacantBoarded and pctVacant6up with crime appeared very small. Note in our analysis the trends above were quite similar across all regions.

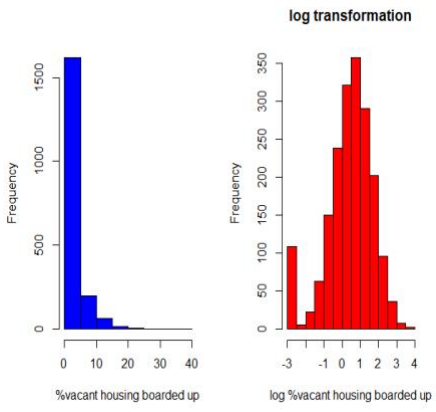


Fig 3.3.2.1

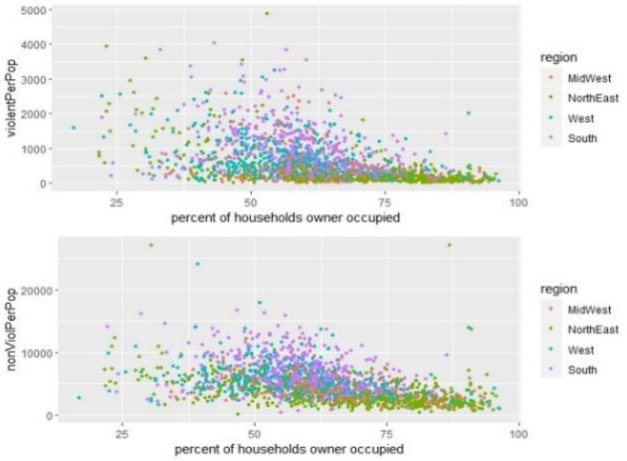


Fig 3.3.2.2

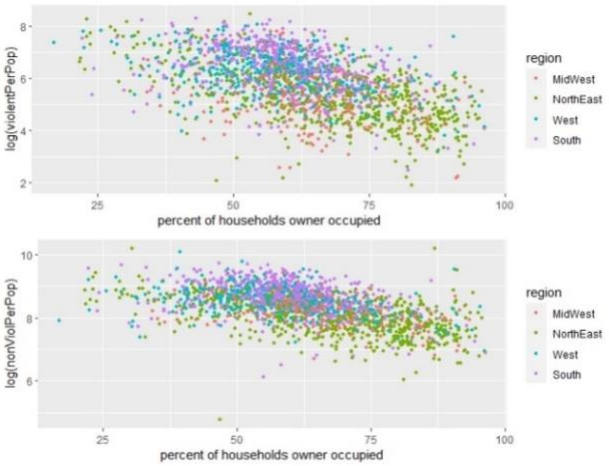


Fig 3.3.2.3

3.3.3 Housing Value

The distribution for the west region appears to be positively skewed. Note this is partly due to California being in the west region with a very high ownHousMed. Finally, there appeared little to no relationship between ownHousMed and crime [Fig 3.3.3.1].

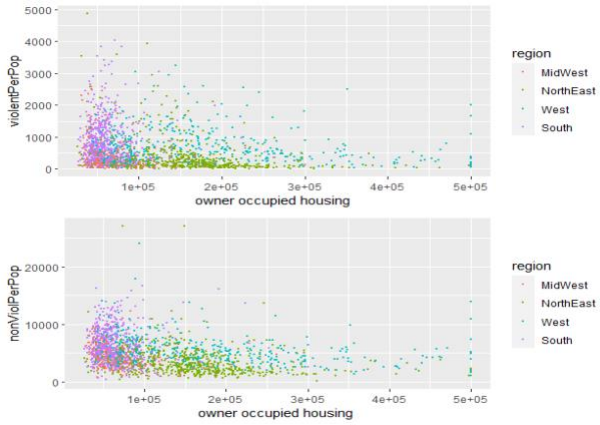


Fig 3.3.3.1

3.3.4 Population

We can see from the two plots in [Fig 3.3.4.2] that a lot of the data is concentrated on lower values, and perhaps slight heteroscedasticity all the plots against violent and non-violent crimes. Also, the histograms of both popDensity and pctForeignBorn both show a strong positive skew, which would explain the concentration on the lower values. A log transformation improves this skew to make it more normal and solves the issue of heteroscedasticity. We found after this transformation that there is no relationship to either of the outcome variables. There is no discernible pattern of either crime variable when looking by state, so instead we look at crime rates by region [Fig 3.3.4.1].

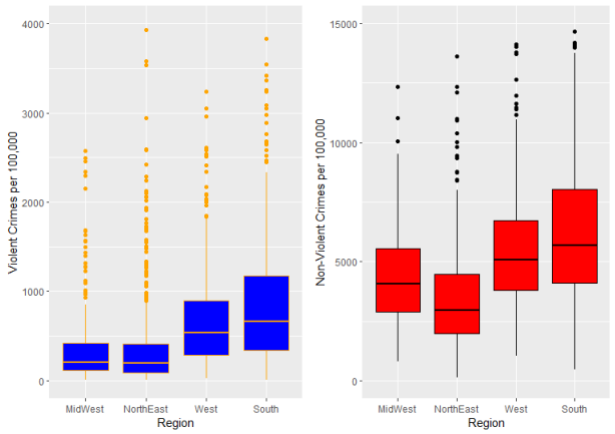


Fig 3.3.4.1

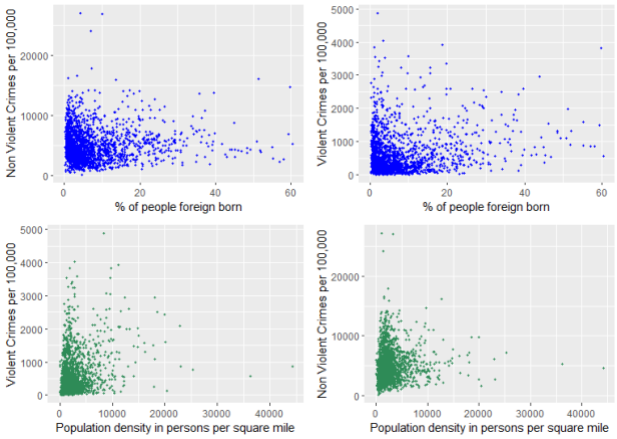


Fig 3.3.4.2

3.3.5 Income & Employment

MedIncome has a weak relationship to the outcome variable [Fig 3.3.5.1]. A more linear relationship can be seen when plotting against the inverse of MedIncome [Fig 3.3.5.2]. The predictor variable pctWdiv shows a faint linear relationship [Fig 3.3.5.3].

The plots for nonViolPerPop have been omitted as they showed a similar trend as violentPerPop.

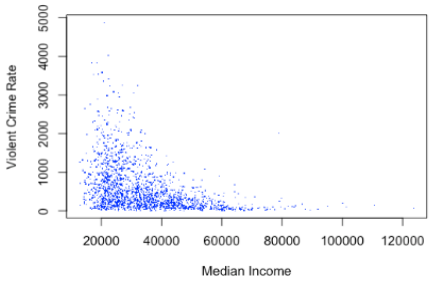


Fig 3.3.5.1

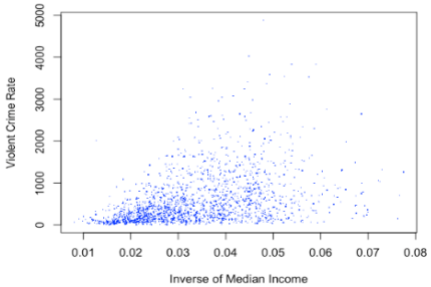


Fig 3.3.5.2

Fig 3.3.5.3

Authors' contributions

Name and student ID	Contribution	Mark weightage
Tom 1816394	Population predictor variables, missing data, report compilation	100%
Colin 1803013	Education and family predictor variables, report compilation	100%
Wei Xiang 1800673	Income & Employment variables, making presentation slides	100%
Scott 1818498	Housing variables, making presentation slides	100%

Appendix

Installing relevant packages

```
> library(ggplot2)
> library(GGally)
> library(corrplot)
> library(VIM)
```

Code for Fig 3.1.1

```
> ggplot(data = USACrime, aes(State, fill = region)) + geom_bar(color
= "black") + scale_fill_manual(values = c("red", "green", "purple",
"yellow", "blue"))
```

Code for Fig 3.1.2

```
> ggplot(data = USACrime, aes(x=pctUrban)) + geom_histogram(binwidth
= 2.5,fill="seagreen",color="black") + xlab("% of people living in
areas classified as urban")
```

Code to change pctUrban into a factor

```
> USACrime$pctUrbanFCTR = cut(USACrime$pctUrban, c(-Inf, 0, 25, 50,
75, 99, Inf), c("0", "1-25", "25-50", "50-75", "75-99", "100"))
```

Code for Fig 3.1.3

```
> plot(jitter(USACrime$ownHousMed, 100),
jitter(USACrime$ownHousQrange, 100), ylab = "IQR", xlab = "Median",
col = “blue”)
```

Code for Fig 3.1.4

```
> plot(jitter(USACrime$rentMed, 10), jitter(USACrime$rentQrange, 10),
ylab = "IQR", xlab = "Median", col = “blue”)
```

Code for Fig 3.1.5

```
> mosaicMiss(UASCrime[,c("medIncome", "pctUrbanFCTR")], highlight =
1, plotvars = 2)
```

Code for Fig 3.1.6

```
> pbox(USACrime, pos = 5)
```

Code for Fig 3.1.7

```
> pbox(USACrime, pos = 6)
```

Code for Fig 3.2.1

```
> Subset = USACrime[complete.cases(USACrime), c(4:9,11,12,17,18,19)]
> pairs(Subset, pch= ".", lower.panel = NULL, xaxt = 'n', yaxt = 'n')
> corrplot(cor(Subset))
```

Code for Fig 3.3.1.1

```
> par(mfrow=c(1,2))
> hist(USACrime$pctLowEdu,nclass = 30, xlab = "pctLowEdu")
> hist(USACrime$pctLowEdu^0.25,nclass = 30, freq = FALSE, xlab =
"pctLowEdu^0.25")
> curve(dnorm(x, mean = mean(USACrime$pctLowEdu^0.25), sd =
sqrt(var(USACrime$pctLowEdu^0.25))), col = "blue", add = TRUE)
```

Code for Fig 3.3.1.2

```
> plot(USACrime$violentPerPop~USACrime$pctNotHSgrad, xlab =
"pctNotHSgrad", ylab = "violentPerPop",main="",pch=".",col="blue")
```

Code for Fig 3.3.1.3

```
> par(mfrow = c(1, 2))
```

```
> plot(USACrime$violentPerPop~USACrime$pctKidsBornNevrMarr, xlab =
"pctKidsBornNevrMarr", ylab = "violentPerPop", pch = ".", col =
"blue")
> plot(USACrime$violentPerPop~USACrime$pctKids2Par, xlab =
"pctKids2Par", ylab = "violentPerPop", pch = ".", col = "blue")
```

Code for Fig 3.3.2.1

```
> par(mfrow = c(1, 2))
> hist(USACrime$pctVacantBoarded, xlab = "%vacant housing boarded
up", col = "blue")
> hist(log(USACrime$pctVacantBoarded + 0.05), xlab = "log %vacant
housing boarded up", col = "red", main = "log transformation")
```

Code for Fig 3.3.2.2

```
> ggplot(data = USACrime, aes(y = violentPerPop, x =
pctHousOwnerOccup, col = region)) + geom_point(size = 1) +
xlab("percent of households owner occupied")
> ggplot(data = USACrime, aes(y = nonViolPerPop, x =
pctHousOwnerOccup, col = region)) + geom_point(size = 1) +
xlab("percent of households owner occupied")
```

Code for Fig 3.3.2.3

```
> ggplot(data = USACrime, aes(y = log(violentPerPop), x =
pctHousOwnerOccup, col = region)) + geom_point(size = 1) +
xlab("percent of households owner occupied")
> ggplot(data = USACrime, aes(y = log(nonViolPerPop), x =
pctHousOwnerOccup, col = region)) + geom_point(size = 1) +
xlab("percent of households owner occupied")
```

Code for Fig 3.3.3.1

```
> ggplot(data = USACrime, aes(y = violentPerPop, x = ownHousMed, col
= region)) + geom_point(size = .5) + xlab("owner occupied housing")
```

```
> ggplot(data = USACrime, aes(y = nonViolPerPop, x = ownHousMed, col
= region)) + geom_point(size = .5) + xlab("owner occupied housing ")
```

Code for Fig 3.3.4.1

```
> ggplot(data = USACrime, aes(x = Region, y = violentPerPop)) +
geom_boxplot(col = "orange", fill = "blue") + ylab("Violent Crimes
per 100,000")
> ggplot(data = USACrime, aes(x = Region, y = nonViolPerPop)) +
geom_boxplot(col = "black",fill="red") + ylab("Non-Violent Crimes per
100,000")
```

Code for Fig 3.3.4.2

```
> ggplot(data = USACrime, aes(x = pctForeignBorn, y = nonViolPerPop))
+ geom_point(shape = 20,color = "blue")+xlab("% of people foreign
born") + ylab("Non Violent Crimes per 100,000")
> ggplot(data = USACrime, aes(x = pctForeignBorn, y = violentPerPop))
+ geom_point(shape = 20,color = "blue") + xlab("% of people foreign
born") + ylab("Violent Crimes per 100,000")
> ggplot(data = USACrime, aes(x = popDensity, y = violentPerPop)) +
geom_point(shape = 20, color = "seagreen") + xlab("Population density
in persons per square mile") + ylab("Violent Crimes per 100,000")
> ggplot(data = USACrime, aes(x = popDensity, y = nonViolPerPop)) +
geom_point(shape = 20, color = "seagreen") + xlab("Population density
in persons per square mile") + ylab("Non Violent Crimes per 100,000")
```

Code for Fig 3.3.5.1

```
> plot(USACrime$violentPerPop ~ USACrime$medIncome, pch = ".", col =
"blue", ylab = "Violent Crime Rate", xlab = "Median Income")
```

Code for Fig 3.3.5.2

```
> USACrime = mutate(USACrime, "invMedIncome" = 1000*medIncome^-1)
> plot(USACrime$violentPerPop~USACrime$invMedIncome, pch = ".", col =
"blue", ylab = "Violent Crime Rate", xlab = "Inverse of Median
Income")
```

Code for Fig 3.3.5.3

```
> plot(USACrime$violentPerPop~USACrime$pctWdiv, pch = ".", col =
"blue", ylab = "Violent Crime Rate", xlab = "Households with
Investment Income")
```