EDA USACrime

Group 16

07/02/2021

Executive Summary

- Missing data is observed and upon inspection, removed. Incorrect data entries have been found and are recommended to be excluded.

- It is shown that multiple variables measure similar concepts and thus are repetitive for the purpose of analysis. Additionally, some variables are determined to have no relationship with violent and non-violent crime. Appropriate suggestions to consider removing them are made.

- Irregular data patterns are spotted. For modelling purposes, transformation in both predictor and outcome variables are recommended.

Findings

Data Cleaning Several variables exhibit problems with missing or incomplete values. For the variables `medIncome` and `pctEmploy`, we see 52 and 21 NA values respectively. The 21 NA's for `pctEmploy` form just 1% of the total observations and so represent too small a sample to conduct meaningful analysis on. These can be removed. We can investigate the 52 NA's for `medIncome` using boxplots and conclude that they are Missing Completely at Random. Therefore, it is appropriate to remove these points.

We can also see several incorrect entries for `ownHouseMed`(x14), `rentMed`(x14), `ownHouseQrange`(x5), `rentQrange`(x2). These have been removed. A total of 92 entries are removed through this process.

We can also see that the min values for `pctKidsBornNevrMarr` and `pctVacantBoarded` are 0. Upon investigation, we believe the observations are reasonable and there is no problem with the data.

The variable `State` had some empty levels and those were dropped at the beginning to clean up the variable. The variable `Region` had a level "Pacific" which only contained three observations which turned out to be from Alaska. To fix this the "Pacific" level was merged with "West" as this made the most sense geographically. In the end `region` has four levels: Midwest, Northeast, West and South.

Skewness Investigating the histograms and density plots of the variables, we can see that `pctHousOwnerOccup`, `pctVacant6up`, `pctWdiv`, `rentMed`, `pctEmploy are relatively symmetric, while`pctKids2Par,pctHousOccup‘ are negatively skewed, with the rest of the variables positively skewed. As this might affect the robustness of our model, we might want to look into transforming the variables before fitting the model.

Heteroscedasticity Investigating the residual and Scale-Location plots, we can see that the problem of heteroscedasticity is present in most of the variables. This indicates the need to transform the outcome variables before starting the model building process. Perhaps a log transformation may be appropriate.

Correlation After examining the correlation between the variables, we can see that the five variables `medIncome`, `rentMed`, `rentQrange`, `ownHousMed` and `ownHouseQrange` are all highly correlated, with correlation coefficients in excess of 0.64 between every pair. This makes sense, as all the aforementioned variables relate in some way to a person's wealth or income. Therefore, when fitting a linear model, it is reasonable to conclude that we only need to include one of the five variables above, as they encode similar information about the data point.

A high correlation between `pctKids2Par` and `pctKidsBornNevrMarr` is also observed. It is not recommended to include both as they represent different ways of measuring the same notion, and hence we would exclude `pctKidsBornNevrMarr`.

The last group of highly correlated variables is composed of pctLowEdu, pctNotHSgrad and pctCollGrad. They are all measures of education and as such including all three might be redundant. By the principle of parsimony, it is suggested that when moving on to the modelling stage, we include only one of them and consider adding the other ones if it would improve the model.

Linearity & Relationship with outcome variables Looking at the relationship with the outcome variables, there are variables with strong relationship that can be assumed to be linear such as pctKids2Par and pctKidsBornNevrMarr. On the other hand, there are variables with very weak or even no relationship such as popDensity, pctVacantBoarded, pctHousOccup. We might want to exclude these three variables from our model.

Outliers We found three consistent outliers in violentPerPop and one in nonViolentPerPop. These problematic observations remain outliers regardless of the variable they are plotted against or the transformation used. The suggestion would be to keep this in mind when moving on to the model building stage and try to fit models with and without these outlier observations. Only then would one choose which model to use.

pctUrban For pctUrban which gives the proportion of people living in what is considered an urban area in each county, most values are concentrated at the edges: either at 0% or 100% urban population. In this case it would seem to make sense to encode pctUrban as a factor, with levels describing whether a county is urban or not. Further information as to what criteria were used to assess whether an area was considered urban or not may help to inform the decision on how the factor is categorised.

When we encode Urban as a factor with 85% as the cut-off, it appears that Urban areas account for the areas with the worst rates of violent crime.

State When looking at the variable 'State' we noticed that some states had an unusually high number of observations and in some cases exceeded the number of counties in that state. For example, California has 278 observations but only has 58 counties (CASA, 2014). This brings into doubt whether the data are really observed at the county level.

Statistical Methodology

Data Cleaning There are some data points which are clearly incorrect for the variables `ownHouseMed`(x14), `rentMed`(x14), `ownHouseQrange`(x5), `rentQrange`(x2). For the variables `ownHouseMed` and `rentMed`, missing values seem to have been entered with the value '500001' and '1001' respectively. These are clearly incorrect and should be removed. Note that there is some overlap between these two categories.

For the variables `ownHouseQrange` and `rentQrange`, we have some '0' entries, which are clearly incorrect in the context of an inter-quartile range, as the data is not constant. Therefore, these values have also been removed from the data frame. This results in a total of 92 entries being removed.

To assess the cause of the missing data values for `medIncome`, we can create an indicator variable which shows whether the `medIncome` value is NA for each data point. We can then produce two boxplots for each variable, one for each level of the indicator variable, as shown below. Comparing the boxplots, we can see that none of them exhibit any major difference between the missing and non-missing values, which can be seen in the below example. Note that this example is representative of all the other variables. Therefore, we can conclude the data is MCAR and hence can be removed.

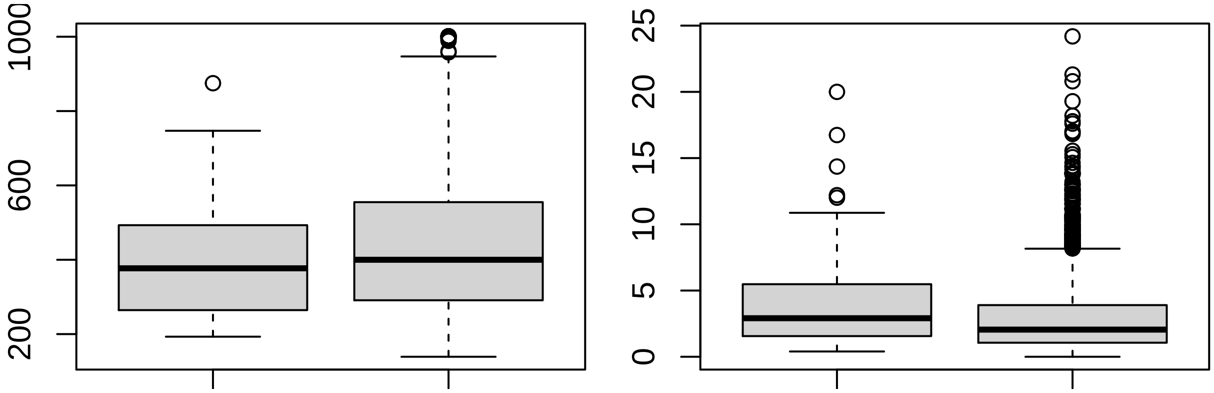

Figure 1

Code used to remove missing/incorrect values (Note this also includes the code to remove outliers):

```
USACrime$ownHousMed[USACrime$ownHousMed==500001]<-NA
USACrime<-na.omit(USACrime)
```

We go on to investigate the 0 values in `pctKidsBornNevrMarr` and `pctVacantBoarded`. For both variables, after sorting the observations, we see a number of similarly small values in the range of 0.03 to 0.05 following the 0 values for both variables, which indicates the 0 values are reasonable, and there is no problem with the data.

```
## [1] 0.00 0.03 0.04 0.05 0.05 0.07
```

For `State` we drop all empty levels and for the region variable we merge the Pacific level into West.

Skewness Skewness of the variables can be identified through looking at the median and mean values in the summary, and through histograms and density plots. Having investigated the univariate plots of all numeric variables, we can identify 3 main groups of skewness in the variables, namely positively skewed, relatively symmetric and negatively skewed, with an example for each included below.

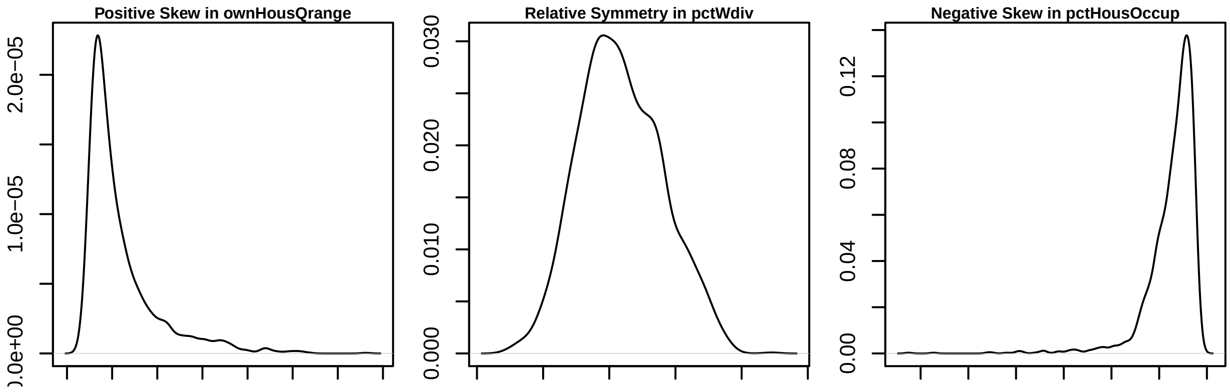

Figure 2

In the following, we will be using the rightly skewed predictor variable `ownHousQrange` to demonstrate how to rectify a skewed distribution using transformation. The positive skew can be seen in the density plot alongside with several outliers on the right end of the distribution. Thus, one might suggest a transformation on this variable to deal with these issues, perhaps an appropriate power transformation from Tukey's Ladder of Powers. To address this, we can try multiple transformations such as power or log transformations. In the case of the power transformation, we can use the `powerTransform` function to provide the optimal transformation to reduce skewness and round appropriately. Applying to this example, the function gives us a value of -0.55811, which we can round to -0.5. As we can see by the density plot below, this improves the symmetry of the plot significantly and stabilises the spread of the observations. We would however recommend choosing a specific transformation only in the model building stage by comparing different ones and only then deciding which works best for the given variable.

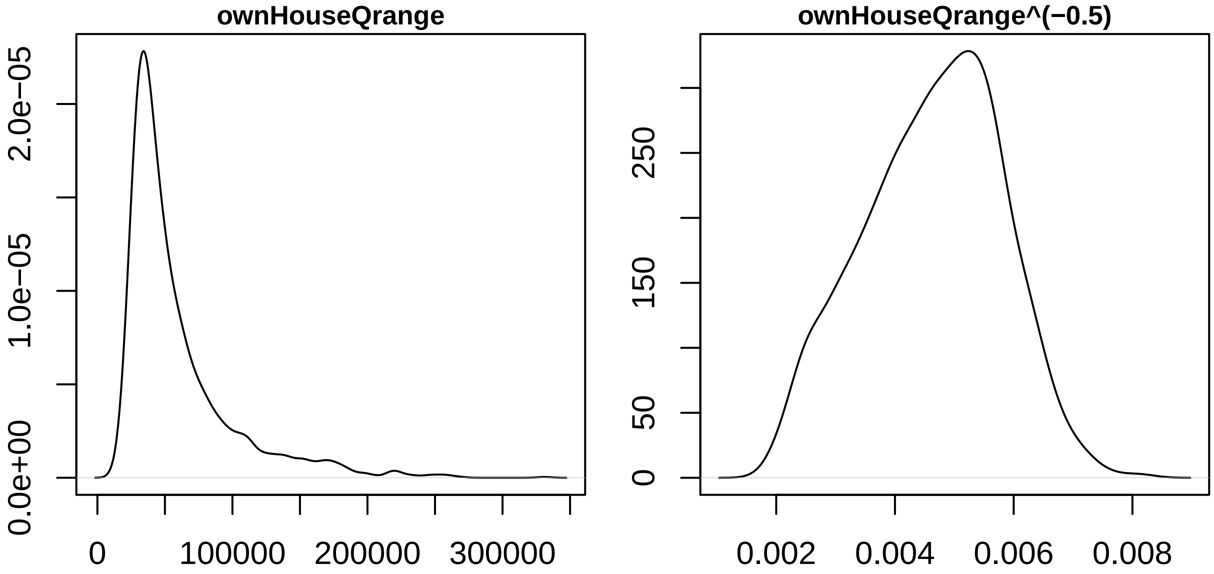

Figure 3

Similar approach can be used for negatively skewed variables to stabilize the spread of the observations. As for the cases where the variables are relatively symmetrical, there is no need to apply a transformation.

Heteroscedasticity To identify heteroscedasticity, we fit individual predictors to the outcome variables and make use of the residual plot and the Scale-Location plot. A horizontal line with equally spread lines in the Scale-Location plot indicates homoscedasticity, and heteroscedasticity otherwise. We see a problem of heteroscedasticity throughout most of the variables. This indicates we may need to transform the outcome variables before fitting the model. To illustrate, we will use the predictor variable `pctKids2Par` against the outcome variable `violentPerPop`. We observe a 'right-opening megaphone' residual plot (Figure 10), meaning the variance of the residuals increases as the predictor increases. Together with an upward sloping line in Scale-Location plot, we can identify a problem of heteroscedasticity. We would therefore recommend transforming the outcome variable by taking log. As expected, after the transformation there is a huge improvement in homoscedasticity, indicated by the horizontal line with equally spread points in the Scale-Location plot.

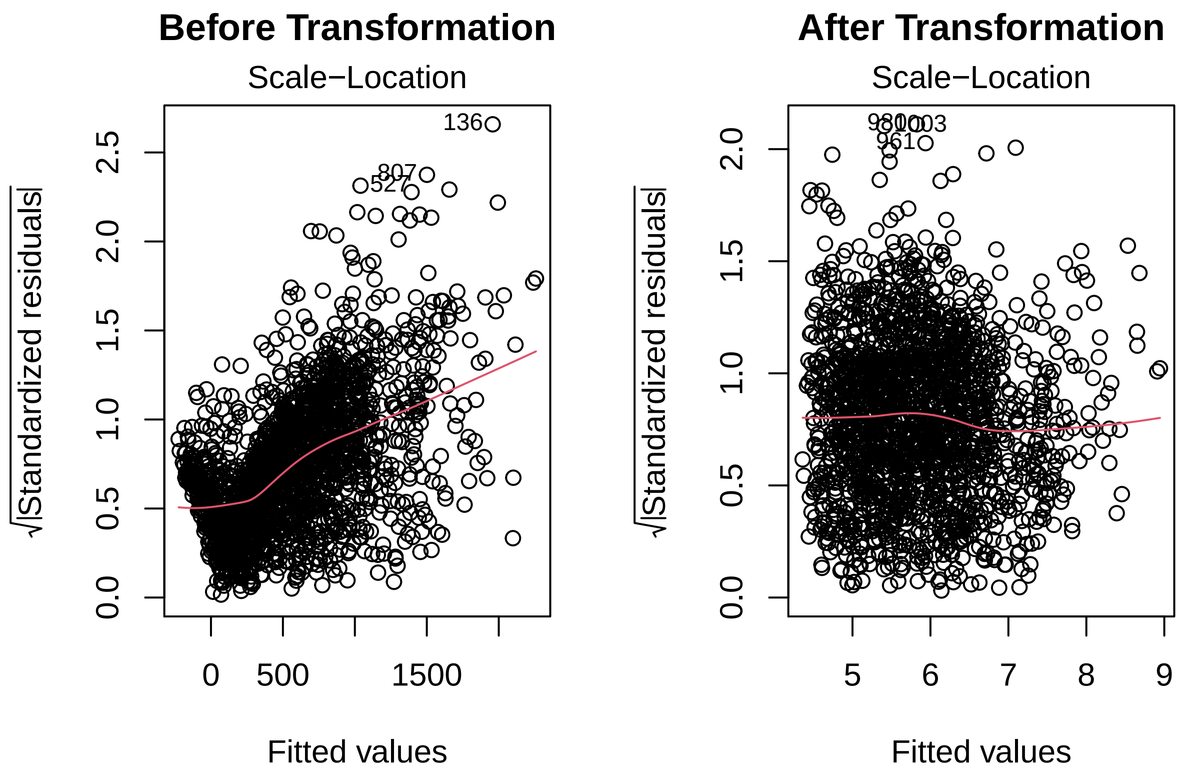

Figure 4

Correlation We can calculate the correlation between variables using the `cor` function. We can also look for non-linear correlation between the variables, such as a quadratic or exponential relationship, by plotting the variables against each other in a pairwise plot. In the below example, we can see a strong positive linear correlation between `medIncome` and both `ownHousMed` and `rentMed`. This positive linear correlation is repeated between all of the variables `medIncome`, `ownHouseMed`, `ownHousQrange`, `rentMed` and `rentQrange`, meaning we would not recommend using all these variables in the model.

```
## [1] 0.773456
```

```
## [1] 0.8479741
```

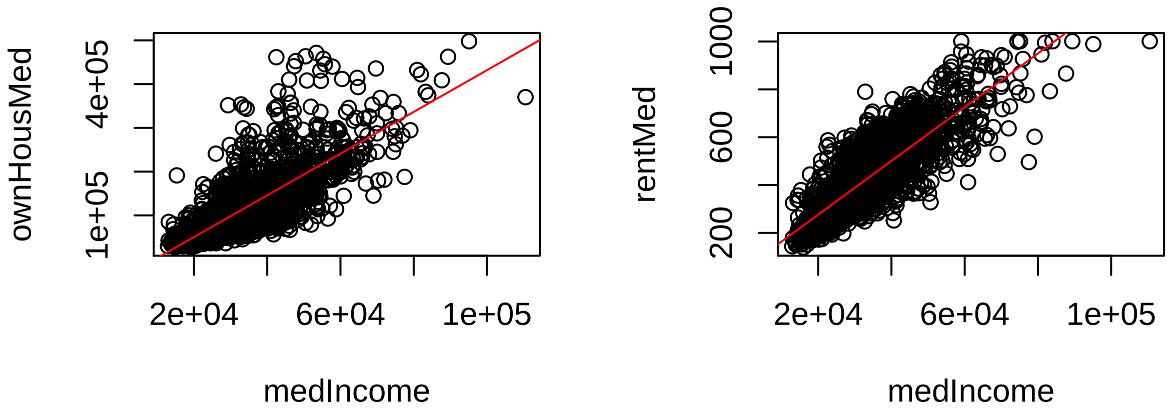

Figure 5

Additionally, `pctWdiv` is quite highly correlated with `medIncome` but not so much with the other predictors relating to income but it does correlate somewhat with the outcomes (-0.64 for violent crime and –0.54 for non-violent crime) so we wouldn't suggest getting rid of it in favour of another outcome variable, and it may be useful to include in a statistical model.

We see a very strong positive relationship between `pctNotHSgrad` and `pctLowEdu`, and so it might be worth excluding one of the two variables from the analysis, especially since they have a similar correlation with other variables. Additionally, `pctCollGrad` and `pctNotHSgrad` have a high negative relationship with similar correlations, although negative and positive, with other variables. Thus, we could decide to exclude `pctNotHSgrad` from our analysis as it is overlapping with two other variables. In the model selection stage, we would recommend considering whether all three variables are needed, as the correlation between the three variables is high and they are all measures of education. Hence only one of the three may be needed to capture the effect of education in the model.

We can also see a strong negative correlation between `pctKids2Par` and `pctKidsBornNevrMarr` of -0.86. This makes sense as they are measuring the quality of the family environment of kids growing up in a community using different benchmarks, i.e. whether there are 2 parents in the family housing or are the kids born to never married. We should only choose one of the two when fitting the model, and we would recommend using `pctKids2Par`, as this is more representative of the current family environment.

Linearity & Relationship with Outcome Variable To assess the relationship between each of the (transformed) predictor variables and the outcome variables, we can produce pairwise plots between them. We can also add the fitted regression line from a hypothetical linear model between the two variables to assess how they relate.

Out of the predictor variables, the ones with the strongest relationship with the outcomes were `pctKids2Par`, `pctKidsBornNeverMarr` and `pctWdiv`.
`popDensity`, `pctVacantBoarded` and `pctHousOccup` were found to have little to no relationship with the outcomes.

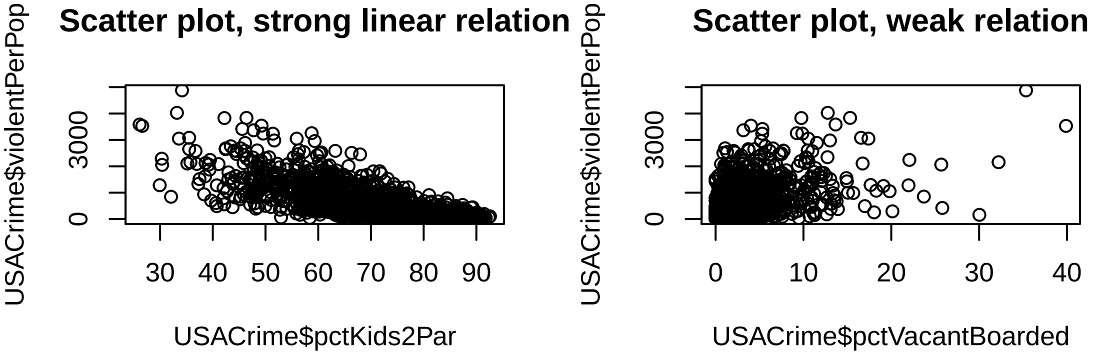

Figure 6

Additionally, while `State` and `region` may be useful for identifying patterns (e.g. for a region with a high crime rate – investigating what combinations of predictors can explain this) we wouldn't include `State` or `region` in the models themselves since there isn't an identifiable association between them and the predictors. Other variables after transformations may all be viable choices for predictors in models and as in the above section some variables can be substituted for others – when they are highly correlated and measure similar information.

Outliers We monitor outliers to see if there are observations that cause problems for multiple variables. As one can observe, there consistently are three outliers with very high non-violent crime, floating above the rest of the data points. As for `violentPerPop`, there is only one outlier that consistently shows up in bivariate plots away from the rest of the observations. When building the model, it will become important to keep this in mind, and perhaps try models with and without these outlier observations to see whether to keep them or to exclude them from the dataset.
Here are the plots showing the three data points in red and the other one in blue.

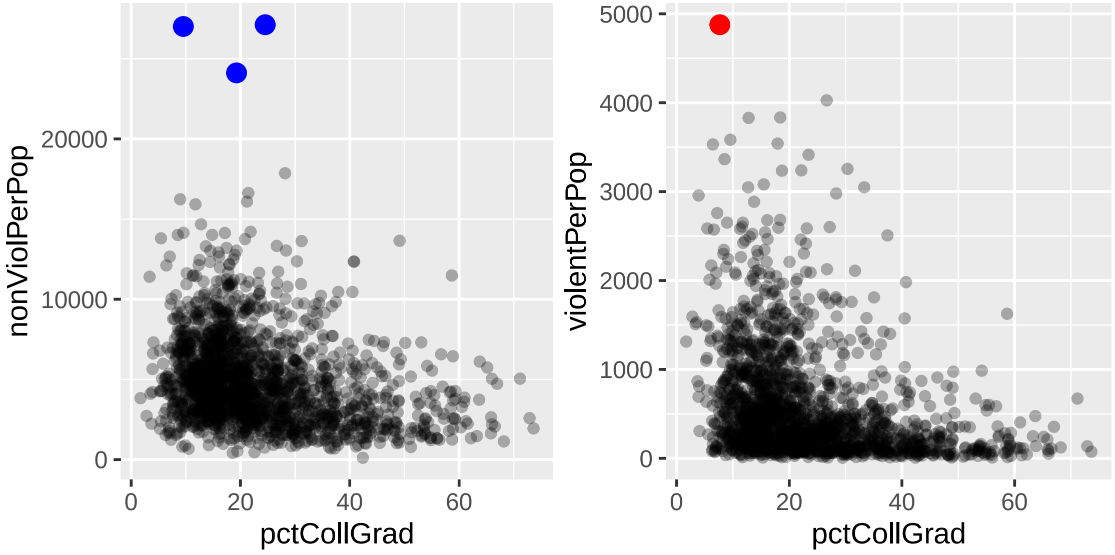

Figure 7

pctUrban A couple of scatter plots of the outcome variables against `pctUrban`:

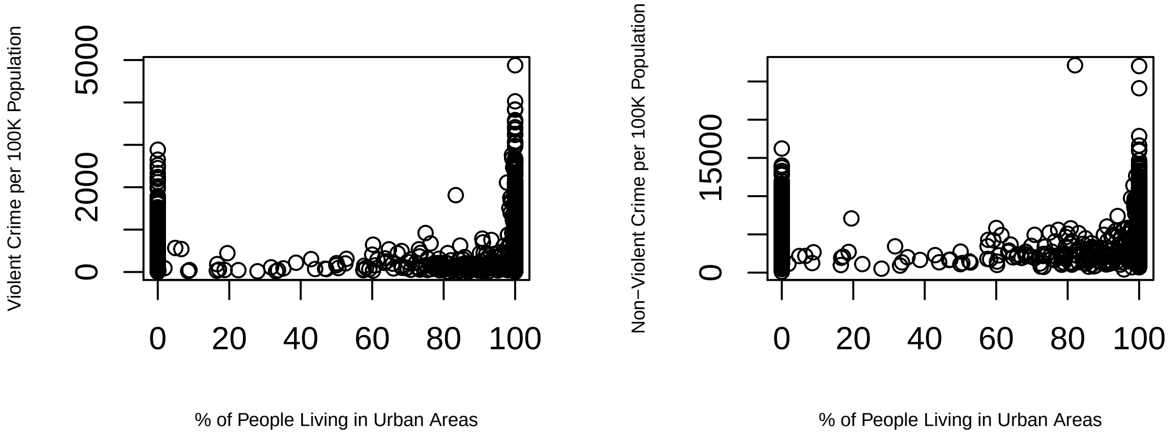

Figure 8

As mentioned above, most values are concentrated at 0 or 100% so we make a categorical variable describing whether an area is urban or not. Chose 85% as cut-off.

```
USACrime <- mutate(USACrime, isUrban = if_else(pctUrban < 85, "Not Urban", "Urban"))
USACrime$isUrban <- as.factor(USACrime$isUrban)
```

Then, we can see how this use of Urban as a factor changes how we see the relationship with crime rates:

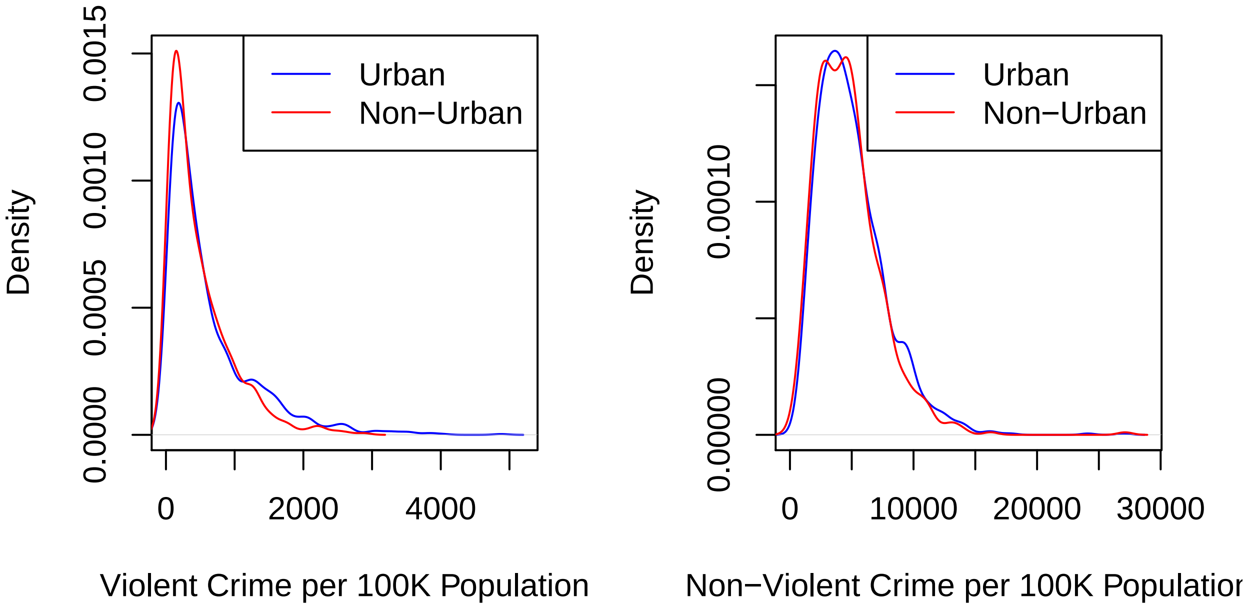

Figure 9

Here we can see that there is little difference for non-violent crime however for violent crime the urban areas have a distribution with a larger tail, so we can say that urban areas account for counties with very high crime rates, as such using the urban factor in a model may be useful for dealing with observations with very high crime rates. This doesn't appear to be particularly sensitive to where the cut-off is chosen (due to the majority of observations having either 0 or 100% urban population) but caution should be taken to make sure that a particular choice of cut-off doesn't affect model results.

Conclusion and Recommendations

(again, feel free to amend or add anything)

Summarizing, before the model building process, we would recommend omitting the missing and incorrect data, excluding some variables with high correlation with other predictors or weak relationship with the outcome variables, and applying transformations to the outcome variables, highly skewed predictors, and variables ‘pctUrban’ and ‘region’ to make the model more comprehensible.

Author contributions

The following table outlines which variables and sections each team member worked on:

David Huk	Matt Wong	Oliver Robinson	Tim Leeman
pctLowEdu	pctKids2Par	medIncome	pctUrban
pctNotHSGrad	pctKidsBornNevrMarr	ownHousemed	State
pctCollGrad	pctHousOccup	ownHouseQrange	Region
pctUnemploy	pctHousOwnerOccup	rentMed	pctWdiv
pctEmploy	pctVacant6up	rentQrange	popDensity
pctForeignBorn	pctVacantBoarded	Missing values	Categorical variables
Outliers	Heteroscedasticity	Data cleaning	Data cleaning
Correlation: education vars	Linearity	Correlation: Income/Housing Vars	
Skewness	Skewness		

The variables `violentPerPop`, `nonViolPerPop`, the writing and formatting of this report and the accompanying presentation and any other work was collaborated on by all four team members.

Marks to be distributed evenly i.e. 100% weighting for each contributor ## References

California State Association of Counties (2014) The Creation of our 58 Counties, [online] Available from: https://www.counties.org/general-information/creation-our-58-counties.

Appendix

Code used: (—ideally, put all code in order of the report here, annotated with #comments—)

For producing boxplots of missing/not missing values

```
library(VIM)

USACrime$medIncomeNA<-ifelse(is.na(USACrime$medIncome),"miss","not miss")
USACrime$pctUnemployNA<-ifelse(is.na(USACrime$pctUnemploy), "miss", "not miss")

attach(USACrime)
par(mfrow=c(1,2))
boxplot(rentMed~medIncomeNA, xlab="Median Income Missing Indicator", ylab="RentMed")
boxplot(pctKidsBornNevrMarr~medIncomeNA, xlab="Median Income Missing Indicator",
ylab="% KidsBorn NevrMarr")
```

Code used to remove missing/incorrect values (Note this also includes the code to remove outliers):

```
USACrime$ownHousMed[USACrime$ownHousMed==500001]<-NA
USACrime$rentMed[USACrime$rentMed==1001]<-NA
USACrime$ownHousQrange[USACrime$ownHousQrange==0]<-NA
USACrime$rentQrange[USACrime$rentQrange==0]<-NA
USACrime$rentQrange[USACrime$rentQrange==55]<-NA

USACrime<-na.omit(USACrime)
```

Analysing `pctUrban`:

```
#Produces plots of pctUrban against violent and non-violent crime
{par(mfrow=c(1,2))
plot(violentPerPop ~ pctUrban, data = USACrime,
     xlab = "% of People Living in Urban Areas",
     ylab = "Violent Crime per 100K Population")
plot(nonViolPerPop ~ pctUrban, data = USACrime,
     xlab = "% of People Living in Urban Areas",
     ylab = "Non-Violent Crime per 100K Population")
par(mfrow=c(1,1))}
```

```
#Summary of Urban variable when 100% and 0% values taken out
summary(filter(filter(USACrime, pctUrban < 85), pctUrban > 0)$pctUrban)
```

```
#Creates a new variable isUrban which takes Urban if an area is more than 85% urban and Not Urban otherwise
USACrime <- mutate(USACrime, isUrban = if_else(pctUrban < 85, "Not Urban", "Urban"))
USACrime$isUrban <- as.factor(USACrime$isUrban)
```

```
#Creates vectors describing densities of violent/non-violent crime in urban/non-urban areas
denseUrbanVio <- density(na.omit(filter(USACrime, isUrban == "Urban")$violentPerPop))
denseNonUrbanVio <- density(na.omit(filter(USACrime, isUrban == "Not Urban")$violentPerPop))
denseUrbanNonVio <- density(na.omit(filter(USACrime, isUrban == "Urban")$nonViolPerPop))
denseNonUrbanNonVio <- density(na.omit(filter(USACrime, isUrban == "Not Urban")$nonViolPerPop))

#creates plots of densities
{par(mfrow=c(1,2))
  plot(denseUrbanVio, col = "blue", ylim = c(0, max(denseUrbanVio$y, denseNonUrbanVio$y)), xlim = c(0, max(de
  lines(denseNonUrbanVio, col = "red")
  legend("topright", legend = c("Urban", "Non-Urban"), col = c("blue", "red"), lty=c(1, 1))
  plot(denseUrbanNonVio, col = "blue", ylim = c(0, max(denseUrbanNonVio$y, denseNonUrbanNonVio$y)), xlim = c(
  lines(denseNonUrbanNonVio, col = "red")
  legend("topright", legend = c("Urban", "Non-Urban"), col = c("blue", "red"), lty=c(1, 1))
}
```

region:

```
#Shows levels in region variable
levels(USACrime$region)
#This merges the pacific level into west
levels(USACrime$region) <- c("Midwest", "NorthEast", "West", "South", "West")
```

State:

```
#Drops empty levels in the State variable
USACrime$State <- droplevels(USACrime$State)
```

```
levels(USACrime$region) <- c("Midwest", "NorthEast", "West", "South", "West")
```

Plots to show skewness:

```
par(mfrow=c(1,3))
attach(USACrime)
plot(density(ownHousQrange), main="Positive Skew in ownHousQrange")
plot(density(pctWdiv), main="Relative Symmetry in pctWdiv")
plot(density(pctHousOccup), main="Negative Skew in pctHousOccup")
```

Plot to demonstrate transformation improving skewness

```
par(mfrow=c(1,2))
plot(density(ownHousQrange), main="ownHouseQrange")
plot(density(ownHousQrange^(-0.5)), main="ownHouseQrange^(-0.5)")
```

Transformations suggested to reduce skewness:

```
USACrime$medIncome.transformed<-medIncome^(-0.5)
USACrime$ownHousMed.transformed<-ownHousMed^(-0.5)
USACrime$ownHousQrange.transformed<-ownHousQrange^(-0.5)
USACrime$rentQrange.transformed<-rentQrange^(-0.5)
```

Examples of correlation between variables:

```
cor(medIncome, ownHousMed)
cor(medIncome, rentMed)
```

```
par(mfrow=c(1,2))
plot(medIncome, ownHousMed)
abline(lm(ownHousMed~medIncome),col="red")

plot(medIncome, rentMed)
abline(lm(rentMed~medIncome), col="red")
```

Code for extracting 0 values in variable `pctKidsBornNevrMarr`

```
#Prints the top few values for pctKidsBornNevrMarr after sorting to see 0 values
head(sort(USACrime$pctKidsBornNevrMarr))
```

Demonstrating plots of variables with strong/weak association with outcome variables:

```
par(mfrow=c(2,2))
#Plots pctKids2Par against violentPerPop which have strong association
plot(USACrime$violentPerPop~USACrime$pctKids2Par, main='Scatter plot, strong linear relation')
#Plots pctVacantBoarded against violentPerPop which have weak association
plot(USACrime$violentPerPop~USACrime$pctVacantBoarded, main='Scatter plot, weak relation')
```

Plots to demonstrate fixing heteroscedasticity with transformations

```
par(mfrow=c(2,2))
#Scale-Location and Residual vs Fitted plots for regression before transformation
plot(lm(USACrime$violentPerPop~USACrime$pctKids2Par),3, main="Before Transformation")
#Scale-Location and Residual vs Fitted plots for regression after transformation
plot(lm(log(USACrime$violentPerPop)~USACrime$pctKids2Par),3, main="After Transformation")
```

Residual plot

```
plot(lm(USACrime$violentPerPop~USACrime$pctKids2Par),1, main="Before Transformation", sub="Figure 10")
```

Before Transformation

Residuals vs Fitted

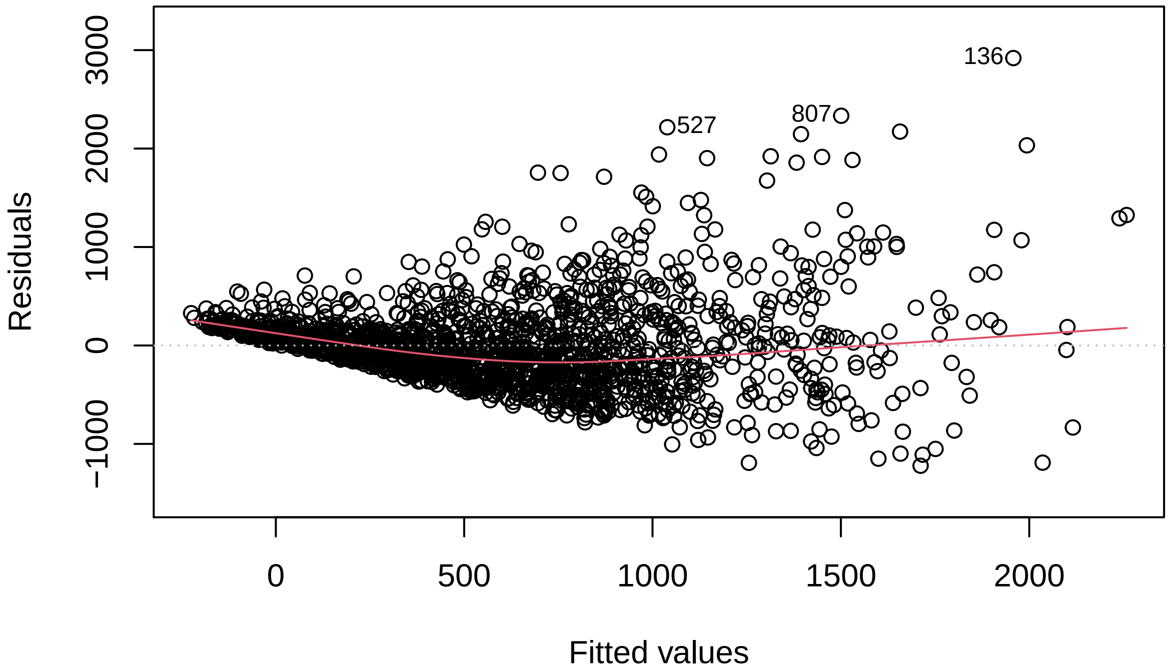

Figure 10