Exploratory Data Analysis Report - Group 1

Authors: 1725298, 1803954, 1824442, 1664771 February 7, 2021

1 Summary

- The primary indicators of the level of crime was much to do with the environment children are brought up in, the overall wealth of the region and rate of completion of high school.
- Many of the variables required some form of manipulation in order to come to conclusions about their relationship with the level of crime.
- Factors that did not affect crime level to a significant degree include cost of housing and population demographics.

2 Findings

We found that the primary indicators of the level of crime was much to do with the environment children are brought up in, the overall wealth of the region, and rate of completion of high school. Additionally, we found that the majority of the variables exhibited some degree of heteroskedasticity and skewness that could be solved with a log transformation or by splitting the variable into equally sized bins. The most highly correlated variables, their transformations, and their corresponding correlations with both violent and non-violent crime are summarised in Table 1.

Variable	Violent Correlation	Non-violent Correlation	Transformation
pctKids2Par	0.741	0.680	log
pctKidsBornNevrMarr	0.649	0.525	untransformed
pctWdiv	0.639	0.535	log
pctHousOwnerOccup	0.524	0.515	log
medIncome	0.489	0.529	log
pctNotHSgrad	0.501	0.396	log

Table 1: Variables with highest correlation with dependent variables

We found that of these variables, many were strongly colinear. As such to avoid the effects of multicollinearity, if making a model one would choose a small subset of variables with low linearity, even if ones chosen have slightly weaker correlations with the dependent variables. An example of such a selection would be `pctKidsBornNevrMarr`, `pctHousOwnerOccup`, `pctVacantBoarded`.

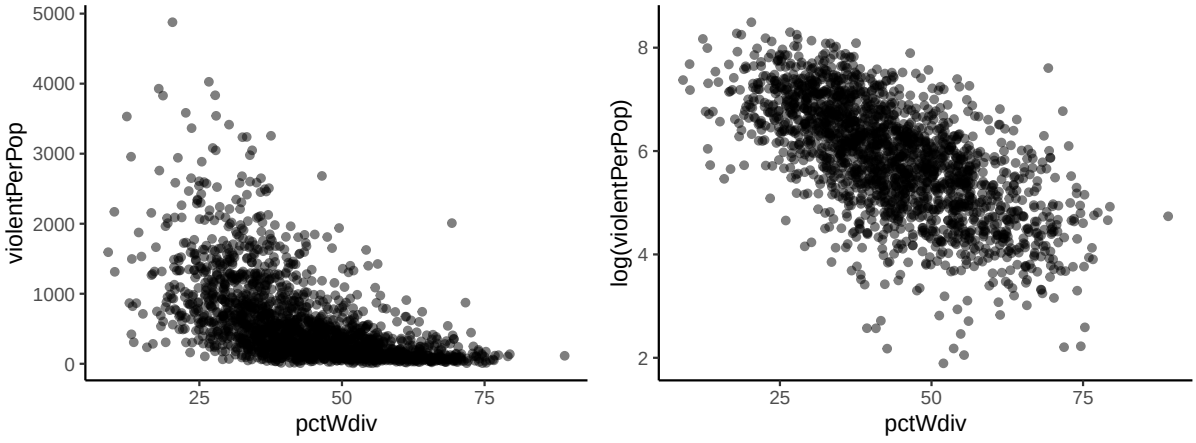

Figure 1: The relationship of the variables is heteroskedastic, this is solved with a log transformation.

In Figure 1 we show the distribution of one of the variables that we decided to transform and the outcome of applying a log transformation. A similar transformation was applied to many variables that exhibited a similar relationship.

Some factors had very little correlation with the dependent variables. In particular, variables concerning population demographics had no correlation with non-violent crime, with mild correlation for violent crime (~0.2). The median rent price, itself only weakly negatively correlated with crime (0.31), was a mildly better indicator than the median house price (corr $\approx$ 0.25) and measures of inequality in the cost of houses such as the interquartile range of house prices or rent in a region showed almost no correlation.

While not useful as an area to target for crime prevention, there is a clear link between region and crime demonstrated in Figure 2.

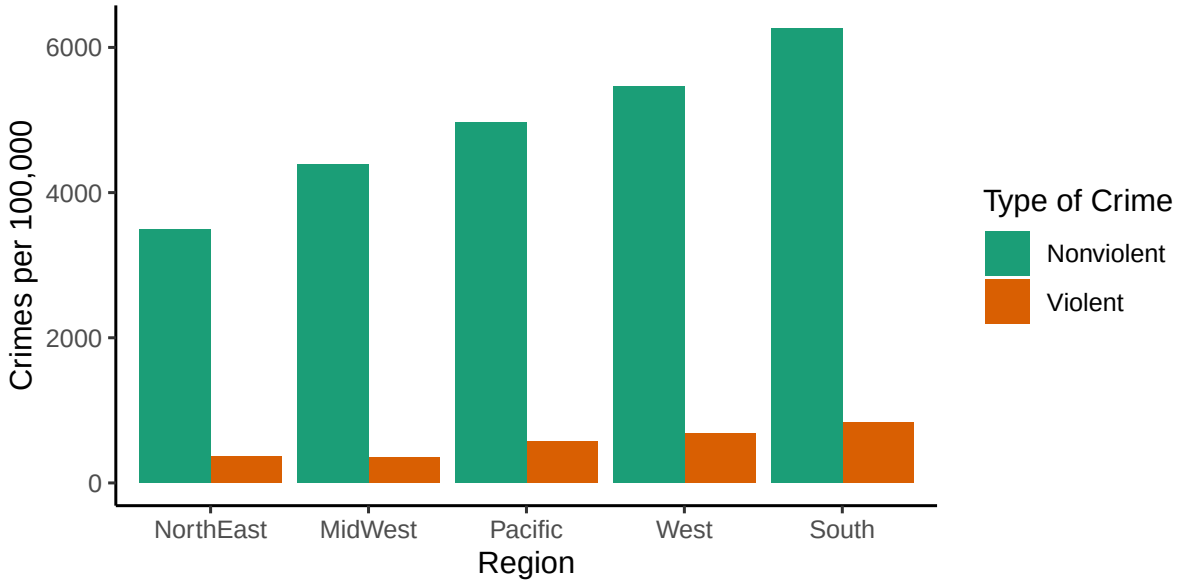

Figure 2: The levels of crime by geographic region

3 Statistical Methodology

Missing Values The variables `medIncome` and `pctEmploy` contained `NA` values explicitly, but `rentMed` and `ownHousMed` had implicit missing data due to truncation (which would make them MNAR). We treat the former case first. Below are tables showing the number of missing values in each variable.

	pctEmploy	
medIncome	Not Missing	Missing
Not Missing	1829	21
Missing	52	0

Table 2: Number of missing values by variable

	pctEmploy	
medIncome	Not Missing	Missing
Not Missing	96.16%	1.10%
Missing	2.73%	0%

Table 3: Missing values in proportion

We can see that there are a nontrivial number of missing values in both cases, so it was important to determine how these entries were missing (i.e. MCAR, MAR, or MNAR) so we could decide whether it was acceptable to ignore or delete the entries, or whether a different technique such as imputation would be used.

Median Income	Violent	Non-violent
Value present	578	4923
Value missing	801	5627
%-increase	38.9%	14.3%

%Over-16 employed	Violent	Non-violent
Value present	584	4941
Value missing	593	5035
%Increase	1.5%	1.9%

Table 4: Comparing crime statistics from regions with missing values to those that have values present

Here we see that on a regional level, those without values for median income also have much larger values of violent crime, with a much smaller effect for the `pctEmploy` variable. There was no pattern in the dataset as to which variables were missing and we were unable to determine why the data were missing. In the next section we discuss why this is not so much an issue.

When deciding how to deal with the truncated variables, we felt it best to simply ignore the missing values as they were not numerous in the dataset, as shown in the figure. Ultimately we chose against imputing missing values in either case as our goal was to describe and investigate the data rather than predict.

	ownHousMed	
rentMed	Not Missing	Missing
Not Missing	1882	6
Missing	6	8

Figure 3: Missing values of truncated variables

Outliers When doing initial univariate analysis many variables had high variance. This meant that detecting the presence of outliers using tests such as

the extreme studentized deviate test, as in Walfish (2006), did not produce satisfactory results. In any case, we did not wish to remove but the most extreme outliers from our data given that our sources are credible and official and the data itself is from a large region. We know that for example, there is extreme poverty in some parts of the US as well as incredibly affluent areas which may skew the results of some areas surveyed.

We also discussed the removal of Washington DC from our dataset as it is a special administrative region in the US which operates differently to most areas. Specifically it has a low permanent population which would seriously affect the per-capita figures of crime. Indeed, of all states and DC, DC had the highest violent and non-violent crime levels, its violent crime being twice as high as the next (3048 compared to 1414). For this reason we decided to exclude it from all of our calculations, but we also note that due to the size of our dataset this did not change our results when rounded in reporting.

An example of the lack of impact that the outliers have is the pctNotHSgrad variable. After plotting the graph of the relationship between this variable and the log transformation of the dependent variable, it is clear to see there are multiple outliers. However, after identifying 10 clear data points that have large residuals, removing them had a minimal effect on the correlation (edges upwards by 0.01). Furthermore, upon further inspection, these data points have no abnormal factors that may have caused them to be outliers and seem to be perfectly valid. Hence, removing outliers from this dataset is often illogical and makes little difference to any conclusions drawn.

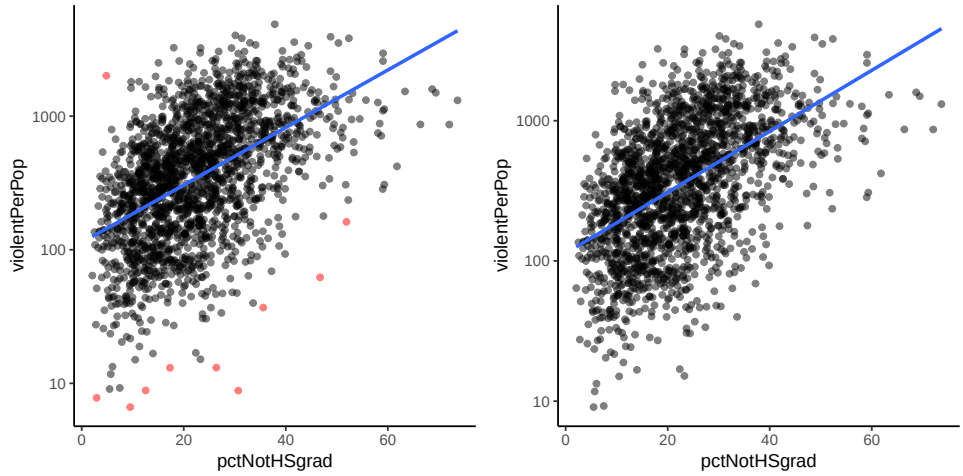

Figure 4: Demonstration of outliers and their effect on correlation. Correlation with: 0.5, correlation without: 0.51

Variable Groups When building models it is important to avoid multicollinearity. In our data we identified that some variables could be grouped together as they were measuring the same kind of concept e.g. number of parents, or education level. If variables within categories were correlated then we could select the variable which is most correlated (positively or negatively) with the dependent variables as the "representative" of this category and avoid collinearity.

The categories of variables we identified were income (medIncome, pctWdiv), education (pctLowEdu, pctNotHSgrad, pctCollGrad), employment (pctUnemploy, pctEmploy), parents (pctKids2Par, pctKidsBornNevrMarr), house occupation (pctHousOccup, pctHousOwnerOccup), vacancy (pctVacantBoarded, pctVacant6Up), house value (ownHousMed, ownHousQrange), rent level (rentMed, rentQrange), and demographics (popDensity, pctForeignBorn) .

For the income category, we identified a linear relation between the two variables which had a correlation of 0.75, and of the two the pctWdiv variable had higher correlation with the dependent variables. This along with the fact that the medIncome variable has missing values meant that we could ignore this variable entirely and use pctWdiv as a proxy if necessary. The relationship between pctWdiv and the dependent variables exhibits heteroskedasticity which can be solved with a log transformation as demonstrated in Figure 1. We used the log transformation because violentPerPop has a clear lower limit of 0 and no defined upper limit. Additionally, we felt that the log of the amount of crime is more interpretable than another transformation reducing heteroskedasticity such as the square root. Intuitively, the log of crime represents the magnitude of crime.

Correlation	violent	non-violent
medIncome	-0.394	-0.462
pctWdiv	-0.557	-0.486

Figure 5: Correlation of income variables and dependent variables

Similarly, the variables of pctNotHSgrad and pctLowEdu in the education category were very strongly correlated (>0.9), and less so were pctNotHSgrad and pctCollGrad with a coefficient of -0.75. In addition to this, as the variables displayed a similar heteroskedasticity as previously, we looked at the correlation of each variable with the log of our dependent variables. Doing so showed us that pctNotHSgrad had the highest correlation with our dependent variables (logged).

For the employment category, the two independent variables we have do not represent the entire population. Particularly it does not include students over the age of 16. As we have seen in the previous category, there is at least some link with people going to college and crime being lower hence we looked for collinearity between these two categories. Indeed there was a correlation of -0.55 between pctUnemploy and pctCollGrad as such one should take care if making a model with both of these variables, or any from both categories. For both (logged) dependent variables pctUnemploy had stronger correlation at 0.48 and 0.39 for violent and non-violent crime than pctEmploy, who themselves had collinearity of -0.67.

In the parents category, the pctKidsBornNevrMarr variable exhibits a linear relationship with the dependent variables, wheras pctKids2Par's relationship resembles exponential decay as shown in Figure 6. Hence, when we compared correlations we decided to compare the former to the untransformed dependent variables, and the latter to their logged counterparts. This analysis gave pctKids2Par a higher correlation and due to the high collinearity (-0.86) we choose only this variable from this category. Note that when only one depedent variable is plotted, it can be inferred that both demonstrated the same rela-

tionship.

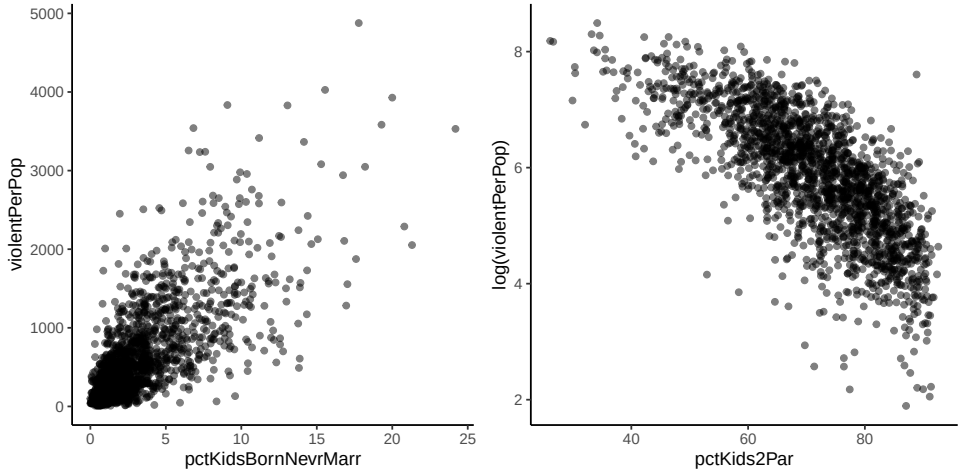

Figure 6: Relationship between parent category variables and violent crime

From hereon, when we speak of correlation with the dependent variables we mean the log of these variables and the transformation is implicit. The two variables in the house owner occupancy category are not strongly correlated (0.133) so it could be feasible to include both in a linear model however `pctHousOccup` has only a correlation of -0.3 with the dependent variables. The other variable, `pctHouseOwnerOccup`, has a higher correlation at ~0.5. Another reason to prefer `pctHouseOwnerOccup` is because the data is far less skewed as seen in Figure 7.

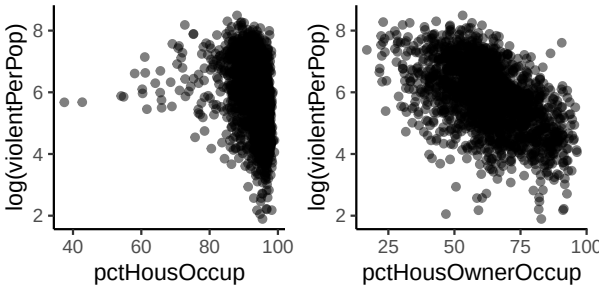

Figure 7: Demonstration of skewness

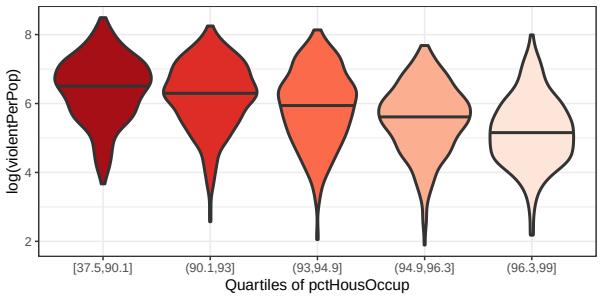

Figure 8: Quartiles of rate of house occupation and relation to crime

A way to solve the skewness if one were to want to include this variable in a linear model is to split the data up into factors of roughly equal distribution. Splitting this variable into its quartiles we can see that the median of each group does demonstrate that higher occupancy rate means lower crime, however as is shown in Figure 8 the differences in these group medians are so small and the range so large that we felt it not worth including.

In the vacancy section, we found a moderate correlation between `pctVacantBoarded` and the *untransformed* dependent variables (~0.5). Additionally, there was no correlation between `pctVacancy6Up` and the dependent variables. The data for `pctVacantBoarded` is skewed towards the lower end, how-

ever the correlation suggests if a county has a significant percentage of houses boarded then crime is higher. The lack of much data on the higher end `pctVacantBoarded` could mean that this correlation was partially due to outliers or random chance.

For the house value category, our analysis ignored the truncated entries. The two variables exhibit strong collinearity (-0.798) hence a model would only include one. In all cases the variables are both weakly negatively correlated with the dependent variables but `ownHousMed` has slightly stronger correlation.

Correlation	Violent	Non-violent
`ownHousMed`	-0.229	-0.332
`ownHousQrange`	-0.143	-0.209
`rentMed`	-0.272	-0.377
`rentQrange`	-0.198	-0.337

Figure 9: Correlations of house value variables

Similarly for the rent level category, we find that the variables exhibit strong collinearity (0.7) and weak correlation with the dependent variables. Notable is that for both of these categories the correlation is 0.1 points stronger for non-violent crime indicating that there is indeed a weak link between nonviolent crime and lower house prices/rent.

Finally in the demographics category, we find moderate collinearity (0.627), and (looking at untransformed dependent variables) no correlation with non-violent crime (<0.1 in absolute value in both cases), and weak correlation with violent crime (0.262 and 0.202 for `popDensity` and `pctForeignBorn` respectively).

4 Authors' Contributions

This assignment was completed by Martin and Angelo. We reached out to Fangtian via Teams and email but did not hear anything back despite them being online on Teams. In our first meeting Andrei informed us that his contributions probably wouldn't be very substantial but that he was happy with taking 90% of the mark. Martin and Angelo decided to split the analysis into half, each taking half of the independent variables. After analysis the report was written by Martin and edited by both, then the powerpoint was produced by Angelo and edited by both.

Proposed grade distribution:

- Martin and Angelo: 110%
- Andrei and Fangtian: 90%

References

[Wal06] Steven Walfish. "A review of statistical outlier methods". English. In: *Pharmaceutical technology (2003)* (2006). ISSN: 1543-2521.

A Code Appendix

```
# Load data and preprocess first half
load("~/R/st404/data/USACrime.Rda")
data <- tibble(bind_cols(USACrime[1:12], USACrime[23:24])) %>%
    mutate(medIncome = as.double(medIncome)) %>%
    clean_names() %>%
    rename(
        nonviolent_per_pop = non_viol_per_pop,
        pct_invest_rent_income = pct_wdiv,
        pct_not_hs_grad = pct_not_h_sgrad,
        pct_college_grad = pct_coll_grad,
        pct_kids_both_parents = pct_kids2par,
        pct_kids_unmarried_parents = pct_kids_born_nevr_marr
    )

dependent_vars = c("violent_per_pop", "nonviolent_per_pop")

# Which columns have missing values?
data[ , apply(data, 2, function(x) any(is.na(x)))] %>% names()

# Which rows have these?
data <- data %>%
    mutate(
        missing_med_income = is.na(med_income),
        missing_pct_employ = is.na(pct_employ)
    )

# Create tables for missing values, first raw values then proportions
# Table 2
data %>%
    select(starts_with("missing")) %>%
    mutate_all(list(~ifelse(.,"Missing","Not Missing"))) %>%
    table()

# Table 3
data %>%
    select(starts_with("missing")) %>%
    mutate_all(list(~ifelse(.,"Missing","Not Missing"))) %>%
    table() %>%
```

```
    prop.table() %>%
    `*`(100) %>%
    round(2)

# Examine Washington DC's violent crime levels

USACrime %>%
    group_by(State) %>%
    summarise(mean_violent_crime = mean(violentPerPop)) %>%
    arrange(desc(mean_violent_crime))

## Missing Values

# Determine whether regions with missing median income have more crime per capita?
data %>%
    group_by(missing_med_income) %>%
    summarise_at(dependent_vars, mean) %>%
    rename(`Violent crime` = violent_per_pop,`Non-violent crime` = nonviolent_per_pop)
        %>%
    pivot_longer(cols = -missing_med_income, names_to = "crime") %>%
    ggplot(aes(x = crime, y = value, fill = missing_med_income)) +
        geom_col(position = "dodge") +
        geom_col(colour="black",alpha=0, position = "dodge") +
        scale_fill_manual(values=c("#488f31","#de425b")) +
        theme_bw() +
        labs(
            title="Do regions with missing median income have more crime per capita?",
            fill="Missing?",
            x="Type of crime",
            y="Crimes per 100,000 residents"
        )

# Create tables to see whether missing values indicate higher levels of crime
data %>%
    group_by(missing_med_income) %>%
    summarise_at(dependent_vars, mean)

data %>%
    group_by(missing_pct_employ) %>%
    summarise_at(dependent_vars, mean)
```

```r
# Tally to see whether truncated values have an affect on levels of crime
truncated <- USACrime %>%
    mutate(
        truncated_rent = ifelse(rentMed>=1000,TRUE,FALSE),
        truncated_ownhouse = ifelse(ownHousMed >=500000,TRUE,FALSE)
    ) %>%
    select(truncated_rent, truncated_ownhouse, rentMed, ownHousMed)

truncated %>% group_by(truncated_ownhouse, truncated_rent) %>% tally()

# Outliers

anomalies = c(805,331, 625, 1747,1866, 1030, 1056, 846, 444, 156)

USACrime2 <- USACrime %>% tibble() %>%
    mutate(anom = ifelse(row_number() %in% anomalies, T, F))



plot1 <- ggplot(USACrime2, aes(x=pctNotHSgrad,y=violentPerPop)) +
    geom_point(aes(x=pctNotHSgrad,y=violentPerPop, colour=anom), alpha=0.5) +
    geom_smooth(method="lm", se=F) +
    scale_colour_manual(values=c("black","red")) +
    scale_y_log10() +
    theme_classic() +
    theme(legend.position = "none")

plot2 <- USACrime %>% slice(-805,-331, -625, -1747,-1866, -1030, -1056, -846, -444,
    -156) %>%
    ggplot(aes(x=pctNotHSgrad,y=violentPerPop)) +
    geom_point(alpha=0.5) +
    geom_smooth(method="lm", se=F) +
    scale_y_log10() +
    theme_classic()

plot_grid(plot1, plot2)

# Correlation table
USACrime %>%
    mutate(
```

```
        logviolentPerPop = log(violentPerPop),
        lognonViolPerPop = log(nonViolPerPop)
    ) %>%
    select(-State, -region) %>%
    cor(use = "complete.obs") %>%
    as.data.frame() %>% # Tibbles don't have row names
    select(violentPerPop,
           logviolentPerPop,
           nonViolPerPop,
           lognonViolPerPop) %>%
    # abs() %>%
    rownames_to_column(var = "variable") %>%
    filter(
        !(variable %in% c(
            "violentPerPop",
            "logviolentPerPop",
            "nonViolPerPop",
            "lognonViolPerPop"
        )
    )) %>%
    tibble() %>%
    mutate(sumcorlog = lognonViolPerPop + logviolentPerPop) %>%
    mutate(sumcor = nonViolPerPop + violentPerPop) %>%
    arrange(desc(abs(sumcorlog))) %>% # Reorder by absolute value of sum of
        correlations with logged dependent variables
    mutate_if(is.numeric, ~ round(.x, 3))

# Figure 2
USACrime %>%
    group_by(region) %>%
    summarise(`Violent` = mean(violentPerPop), `Nonviolent` = mean(nonViolPerPop)) %>%
    ungroup() %>%
    pivot_longer(!region, names_to = "type_of_crime", values_to = "level") %>%
    ggplot(aes(reorder(region, level), level, fill=type_of_crime)) +
    geom_col(position="dodge") +
    labs(
        x="Region",
        y="Crimes per 100,000",
        fill="Type of Crime"
    ) +
```

```
    theme_classic() +
    scale_fill_brewer(type="qual", palette = "Dark2")

## Analysing Categories
# Income

data %>%
    select(med_income, pct_invest_rent_income, dependent_vars) %>%
    ggpairs(progress = FALSE)

# From our plot we see that dividends are a better predictor than median income so we
    see if it needs a log transformation
# Figure 1
p_invest_viol <- data %>%
    ggplot(aes(x=pct_invest_rent_income,y=violent_per_pop)) +
        geom_point(alpha=0.5) +
        labs(
            x="pctWdiv",
            y="violentPerPop"
        ) +
        theme_classic()

p_invest_viol_log <- data %>%
    ggplot(aes(x=pct_invest_rent_income,y=log(violent_per_pop))) +
        geom_point(alpha=0.5) +
        labs(
            x="pctWdiv",
            y="log(violentPerPop)"
        ) +
        theme_classic()
plot_grid(p_invest_viol, p_invest_viol_log)

# Parents, Figure 6
p_bothparents_viol_log <- data %>%
    ggplot(aes(x=pct_kids_both_parents, y=log(violent_per_pop))) +
        geom_point(alpha=0.5) +
        labs(
            x="pctKids2Par",
            y="log(violentPerPop)"
        ) +
```

```
        theme_classic()

p_unmarried_viol <- data %>%
    ggplot(aes(x=pct_kids_unmarried_parents, y=violent_per_pop)) +
        geom_point(alpha=0.5) +
        labs(
            x="pctKidsBornNevrMarr",
            y="violentPerPop"
        ) +
        theme_classic()

plot_grid(p_unmarried_viol, p_bothparents_viol_log)


# Figure 7, Figure 8
# Skewness plot
p_houseoccup_log <- data2 %>%
    ggplot(aes(x=pct_hous_occup, y=log(violent_per_pop))) +
        geom_point(alpha=0.5) +
        labs(
            x="pctHousOccup",
            y="log(violentPerPop)"
        ) +
        theme_classic()

p_houseowneroccup_log <- data2 %>%
    ggplot(aes(x=pct_hous_owner_occup, y=log(violent_per_pop))) +
        geom_point(alpha=0.5) +
        labs(
            x="pctHousOwnerOccup",
            y="log(violentPerPop)"
        ) +
        theme_classic()

plot_grid(p_houseoccup_log, p_houseowneroccup_log)

# Violin Plot
USACrime %>%
    select(pctHousOccup, violentPerPop) %>%
    mutate(pctHousOccupQuartile = cut_number(pctHousOccup,n=5)) %>%
```

```
ggplot(aes(x=as.factor(pctHousOccupQuartile),y=log(violentPerPop))) +
    geom_violin(aes(fill=pctHousOccupQuartile), size=1, draw_quantiles = 0.5) +
    labs(
        x="Quartiles of pctHousOccup"
    ) +theme_bw() +
    theme(legend.position = "none") +
    scale_fill_manual(values =
        rev(c('#fee5d9','#fcae91','#fb6a4a','#de2d26','#a50f15')))
```