Is there any relationship between race and types of crime committed in 1978 America?

Santhosh Saravanan

2021-05-09

Name: Santhosh Saravanan

EID: sks3648

Introduction

The first dataset I chose was from the Vincentaelbuldock Github website concentrated on a dataset related to Guns and Crime in the United States in 1978. The columns of interest are the violent crime rate (per 100,000), murder rate (per 100,000), robbery rate (per 100,000), and incarceration rate in the state in the previous year. There are 664 observations and the data was collected by the US government with complements to Stock and Watson. This data is tidy and I am interested in furthering my data analysis in more crime statistics from the intriguing results of Project 1 and want to explore more gun statistics. I expect to find some correlation between violent crime rates, murder rates, and robbery rates and there may be some correlation between a person's ethnciity and the proportion of general crimes committed.

Tidy

```
#Import necessary packages
library(readxl)
library(tidyverse)
library(ggpubr)
Guns <- as.data.frame(read_csv("~/git/SDS348/Projects/Datasets/Guns.csv")) #read the Excel file and save as a dataframe
glimpse(Guns)
```

```
## Rows: 1,173
## Columns: 14
## $ X1         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, …
## $ year       <dbl> 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986,…
## $ violent    <dbl> 414.4, 419.1, 413.3, 448.5, 470.5, 447.7, 416.0, 431.2, 457…
## $ murder     <dbl> 14.2, 13.3, 13.2, 13.2, 11.9, 10.6, 9.2, 9.4, 9.8, 10.1, 9.…
## $ robbery    <dbl> 96.8, 99.1, 109.5, 132.1, 126.5, 112.0, 98.4, 96.1, 105.4, …
## $ prisoners  <dbl> 83, 94, 144, 141, 149, 183, 215, 243, 256, 267, 283, 307, 3…
## $ afam       <dbl> 8.384873, 8.352101, 8.329575, 8.408386, 8.483435, 8.514000,…
## $ cauc       <dbl> 55.12291, 55.14367, 55.13586, 54.91259, 54.92513, 54.89621,…
## $ male       <dbl> 18.17441, 17.99408, 17.83934, 17.73420, 17.67372, 17.51052,…
## $ population <dbl> 3.780403, 3.831838, 3.866248, 3.900368, 3.918531, 3.925229,…
## $ income     <dbl> 9563.148, 9932.000, 9877.028, 9541.428, 9548.351, 9478.919,…
## $ density    <dbl> 0.0745524, 0.0755667, 0.0762453, 0.0768288, 0.0771866, 0.07…
## $ state      <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Ala…
## $ law        <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no",…
```

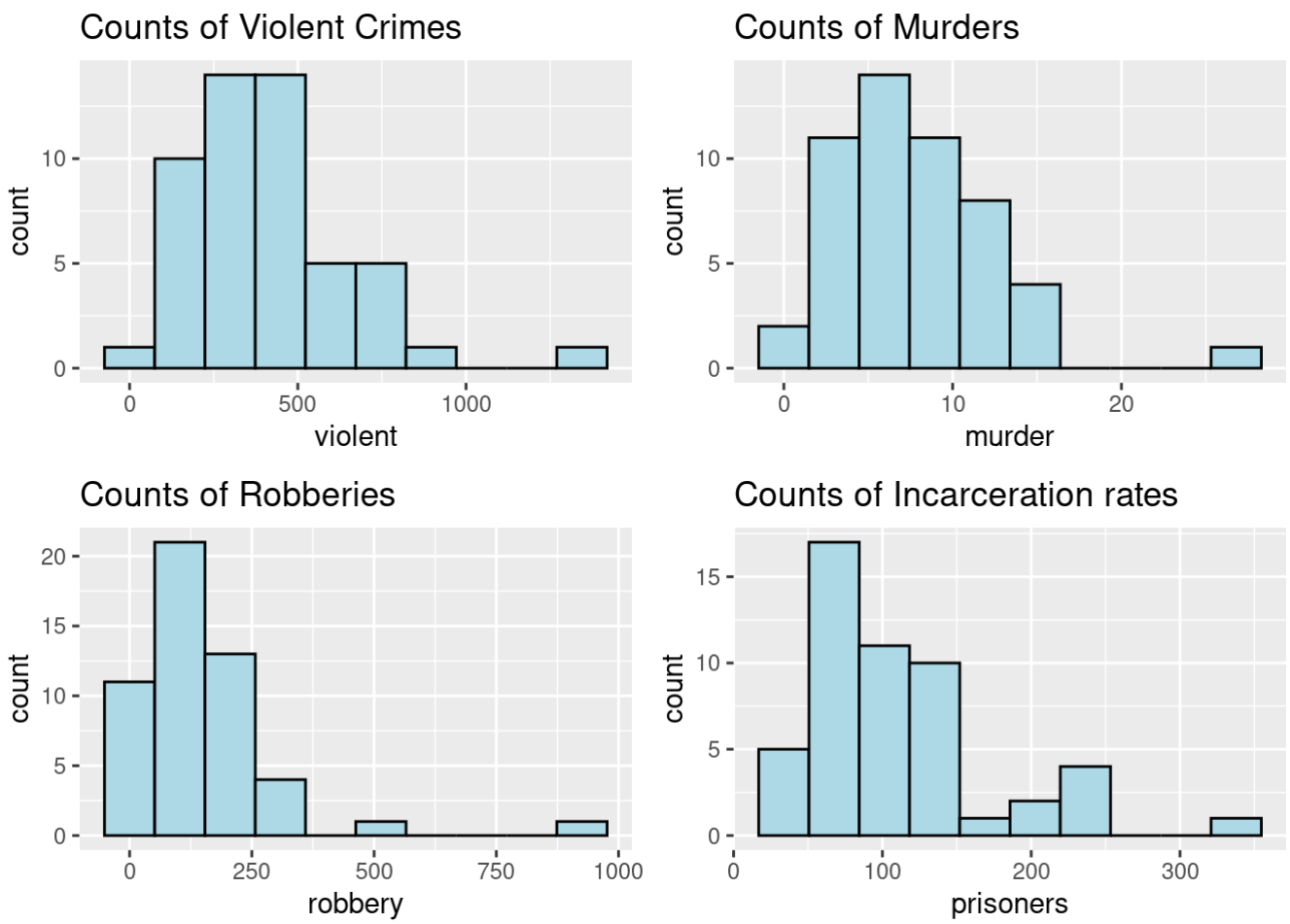
```
print(colnames(Guns)) #print the colNames to get an understanding of the data we're dealing with
```

```
##  [1] "X1"         "year"       "violent"    "murder"     "robbery"
##  [6] "prisoners"  "afam"       "cauc"       "male"       "population"
## [11] "income"     "density"    "state"      "law"
```

```
Guns1978 <- filter(Guns, year == 1978)
Guns1978$X1 = NULL

NE<- c("Connecticut","Maine","Massachusetts","New Hampshire",
             "Rhode Island","Vermont","New Jersey","New York",
             "Pennsylvania")
MW<- c("Indiana","Illinois","Michigan","Ohio","Wisconsin",
             "Iowa","Kansas","Minnesota","Missouri","Nebraska",
             "North Dakota","South Dakota")
S<- c("Delaware","District of Columbia","Florida","Georgia",
            "Maryland","North Carolina","South Carolina","Virginia",
            "West Virginia","Alabama","Kentucky","Mississippi",
            "Tennessee","Arkansas","Louisiana","Oklahoma","Texas")
W<- c("Arizona","Colorado","Idaho","New Mexico","Montana",
            "Utah","Nevada","Wyoming","Alaska","California",
            "Hawaii","Oregon","Washington")

Guns1978 <- Guns1978 %>%
  mutate(Region = case_when(state %in% MW ~ "MidWest",
                        state %in% W  ~ "West",
                        state %in% NE  ~ "NorthEast",
                        state %in% S ~ "South")) %>% arrange(desc(Region))
```

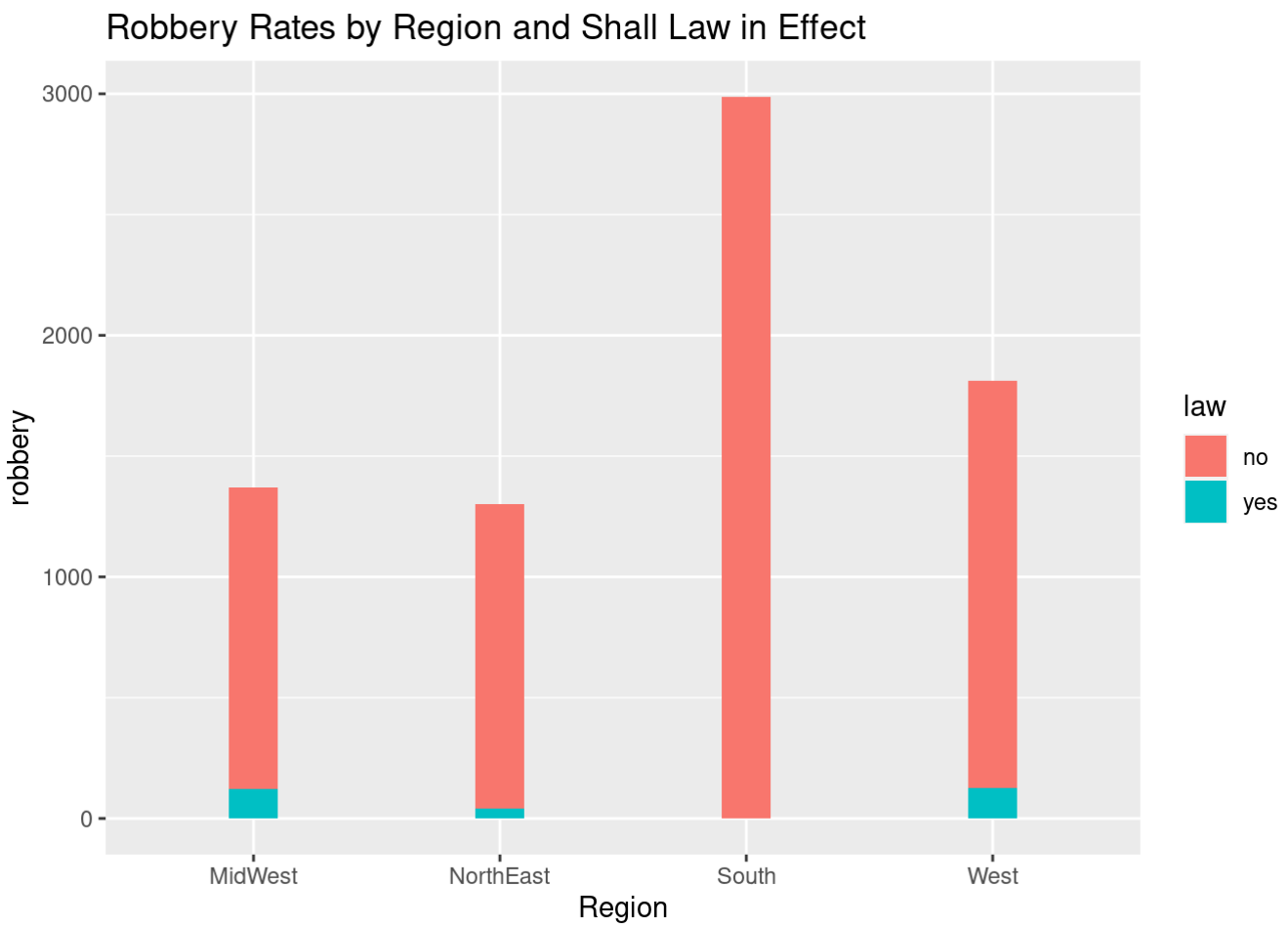
EDA

```
violent <- ggplot(data=Guns1978, aes(x= `violent`)) + ggtitle("Counts of Violent Crimes")+
  geom_histogram(bins=10, color="black", fill="light blue")
murder <- ggplot(data=Guns1978, aes(x= `murder`)) + ggtitle("Counts of Murders ")+ geom_histogr
am(bins=10, color="black", fill="light blue")

robbery <- ggplot(data=Guns1978, aes(x= `robbery`)) + ggtitle("Counts of Robberies") +
  geom_histogram(bins=10, color="black", fill="light blue")
prisoners <- ggplot(data=Guns1978, aes(x= `prisoners`)) + ggtitle("Counts of Incarceration rate
s ")+
  geom_histogram(bins=10, color="black", fill="light blue")
ggarrange(violent, murder, robbery, prisoners,
          ncol = 2, nrow = 2)
```

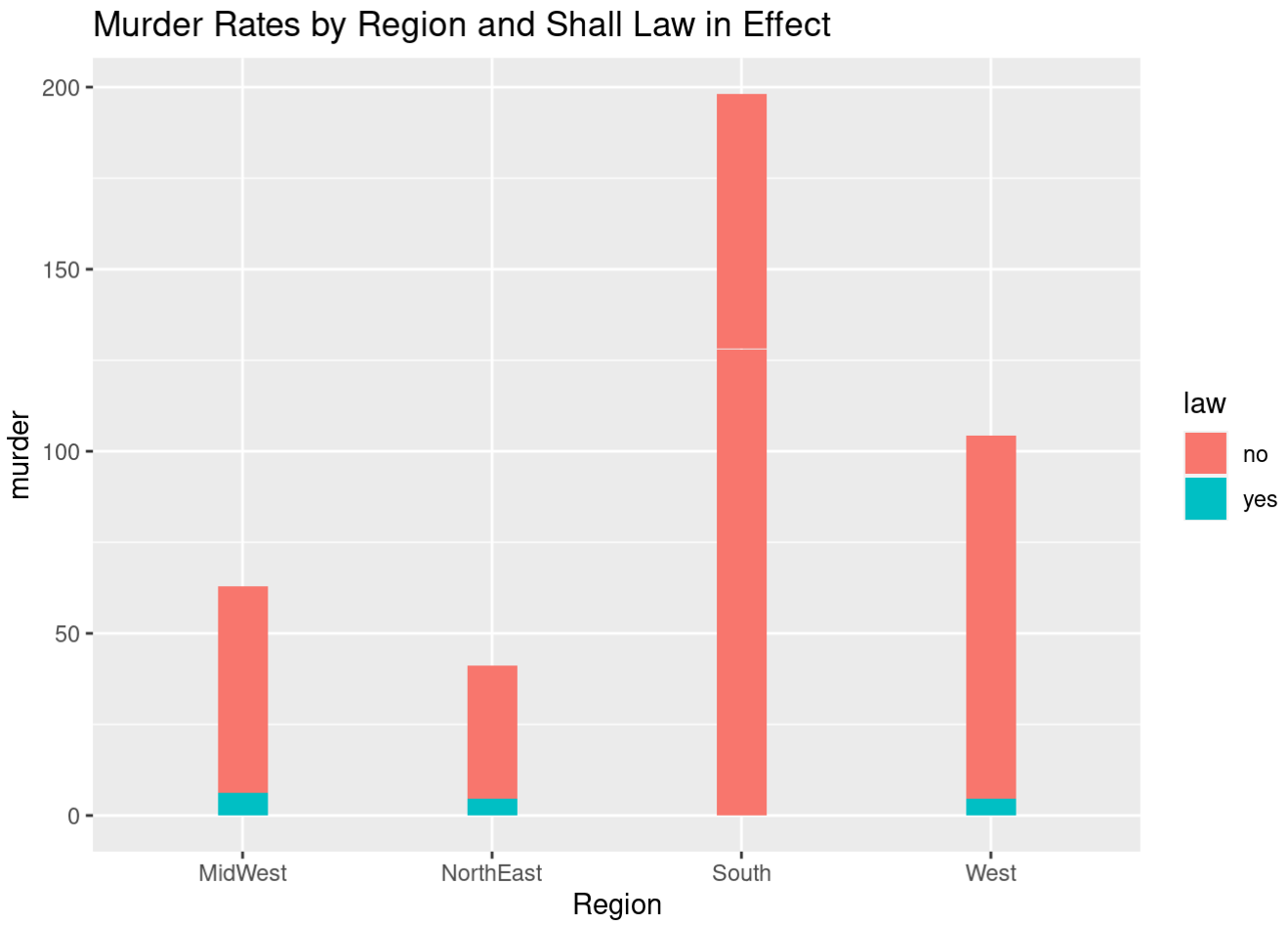


From the plots above describing crime rates in 1978 USA, there seems to be more instances of 0-250 counts of robbery and it's quite uncommon for the number of robbery cases to go above 500. For Incarceration rates counts, there seems to be a relatively even distribution from 0-200, and it's uncommon for one to expect above 300 counts of incarceration rates. For violent rate counts, there is an even distribution of counts from 0-1000 counts and it's quite uncommon for counts to be significantly above 1000 cases. Lastly, for murder cases, it's rare for there to be more than 20 cases of murder in all of the states. These rates are per 100,000 residents in each state.

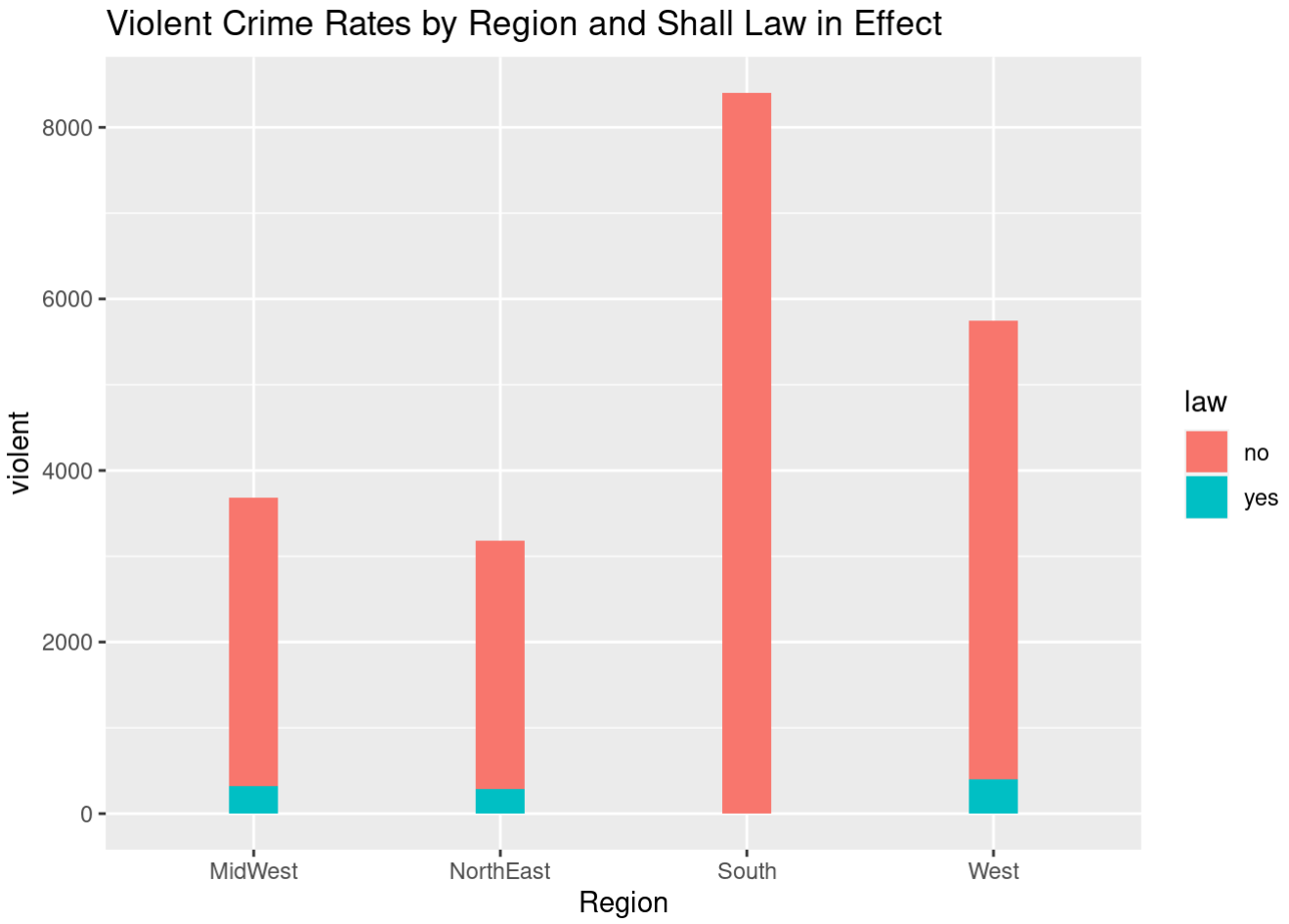
```
ggplot(Guns1978, aes(x=Region, y=robbery, fill = law)) +
  geom_bar(stat = "identity", width=0.2) + ggtitle("Robbery Rates by Region and Shall Law in Ef
fect")
```



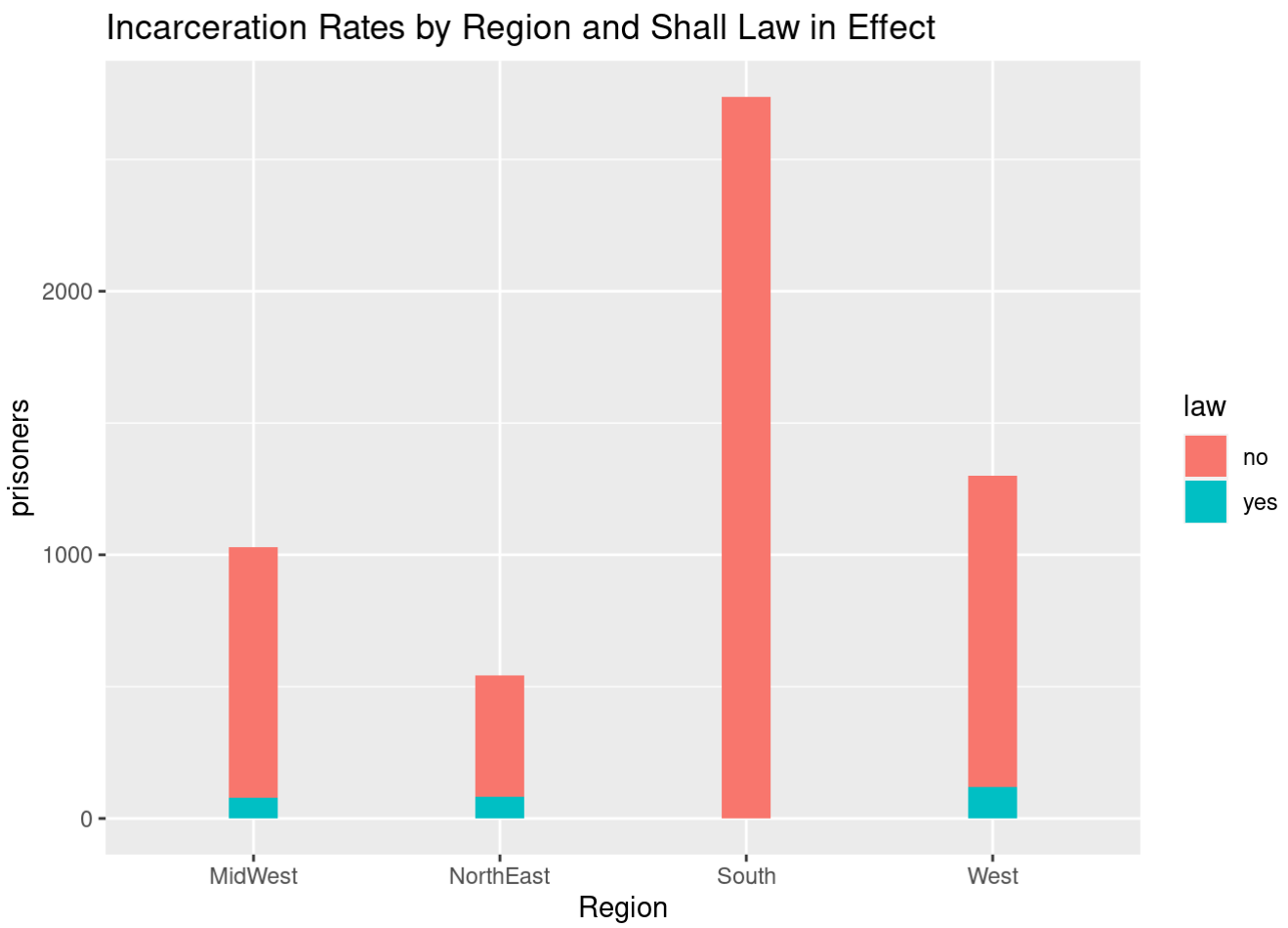

```
ggplot(Guns1978, aes(x=Region, y=murder, fill = law)) +
  geom_bar(stat = "identity", width=0.2) + ggtitle("Murder Rates by Region and Shall Law in Eff
ect")
```



```
ggplot(Guns1978, aes(x=Region, y=violent, fill = law)) +
  geom_bar(stat = "identity", width=0.2) + ggtitle("Violent Crime Rates by Region and Shall Law in Effect")
```

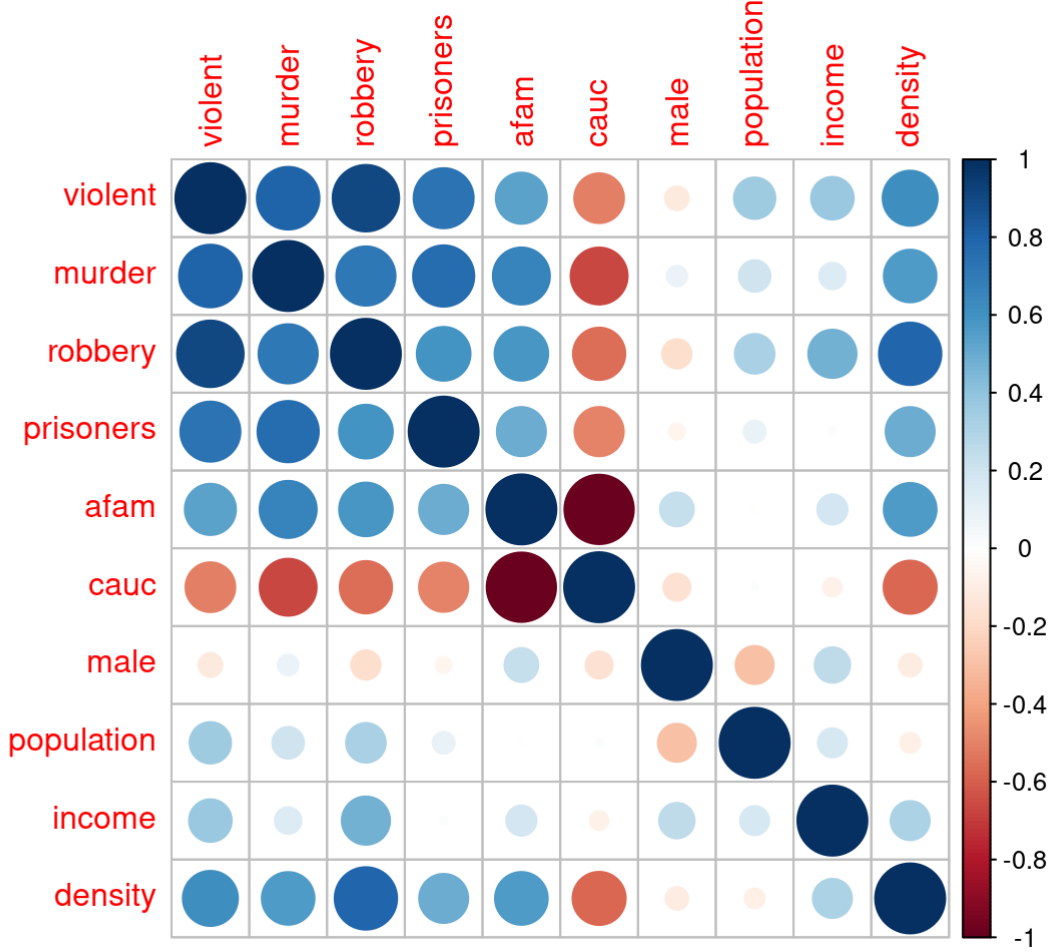



```
ggplot(Guns1978, aes(x=Region, y=prisoners, fill = law)) +
  geom_bar(stat = "identity", width=0.2) + ggtitle("Incarceration Rates by Region and Shall Law in Effect")
```

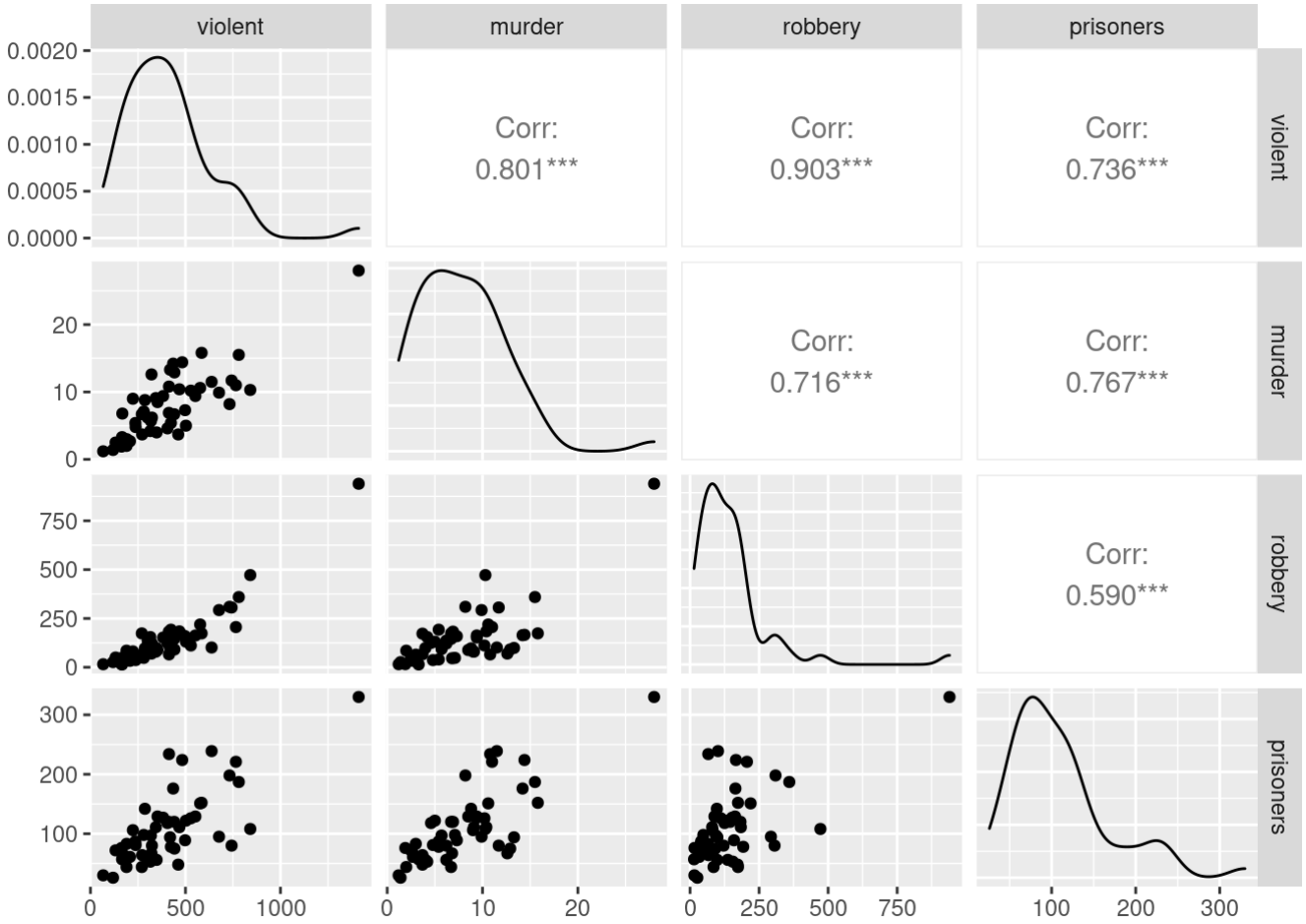


From the respective bar plots shown, the South undoubtedly had the highest crime rates out of any other regions when it came to violent, murder, robbery, and incarceration rate. The distribution of law (Whether a state has a shall carry law) is heavily biased towards not having this law, so there may be a correlation between the crime rates and this categorical variable.

```
library(corrplot)
# calculate correlations
correlations <- cor(Guns1978[,2:11])
# create correlation plot
corrplot(correlations, method="circle")
```




```
# pair-wise scatterplots of all 4 attributes
library(GGally)
ggpairs(Guns1978[,2:5])
```



These are statistically significant correlations between the four variables in question, regarding robbery rates, murder rates, violent rates, and incarceration rates in 1978 USA.

```
manova_guns <- manova(cbind(violent, murder, robbery, prisoners) ~ Region, data = Guns1978)
summary(manova_guns)
```

```
##           Df  Pillai approx F num Df den Df    Pr(>F)
## Region     3 0.82663    4.374     12    138 6.836e-06 ***
## Residuals 47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant differences were found among the 4 regions for at least one of the different type of crime rates per 100,000 residents(Pillai's trace = 0.82663, pseudo F(12,138), p < 0.001).

```
summary.aov(manova_guns)
```

```
##  Response violent :
##             Df  Sum Sq Mean Sq F value Pr(>F)
## Region       3  290288   96763  1.8165 0.1571
## Residuals   47 2503617   53268
##
##  Response murder :
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Region       3 423.52 141.172  8.8382 9.394e-05 ***
## Residuals   47 750.73  15.973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response robbery :
##             Df  Sum Sq Mean Sq F value Pr(>F)
## Region       3   27603  9201.1  0.4144 0.7434
## Residuals   47 1043451 22201.1
##
##  Response prisoners :
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Region       3  74775 24924.9   10.38 2.341e-05 ***
## Residuals   47 112854  2401.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

at least one of the group means for crime rates differ.Post-hoc tests must be used to determine how many groups differ.

```
    pairwise.t.test(Guns1978$murder,Guns1978$Region, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Guns1978$murder and Guns1978$Region
##
##           MidWest NorthEast South
## NorthEast 0.708   -         -
## South     9.9e-05 8.7e-05   -
## West      0.089   0.053     0.017
##
## P value adjustment method: none
```

```
    pairwise.t.test(Guns1978$prisoners,Guns1978$Region, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Guns1978$prisoners and Guns1978$Region
##
##           MidWest NorthEast South
## NorthEast 0.24541 -         -
## South     0.00018 8.9e-06   -
## West      0.47117 0.06817   0.00147
##
## P value adjustment method: none
```

At least one of the group means for crime rates differ by region.

```
#numTests is one test from MANOVA, 4 tests from ANOVA, and 12 tests from cross post-hoc tests
numTests <- 1 + 4 +12
probTypeIError <- 1 - (0.95)^17
Bonferroni <- 0.05/numTests
```

I performed 17 tests, the probability of a Type I error is 58.188% and the Bonferroni Level is 0.00294. None of my post hoc tests were no longer significant when they were significant before the adjustment.

```
#Check sample size assumptions
Guns1978 %>%
  group_by(Region) %>%
  summarise(N = n())
```

```
## # A tibble: 4 x 2
##   Region        N
##   <chr>     <int>
## 1 MidWest      12
## 2 NorthEast     9
## 3 South        17
## 4 West         13
```

The number of states in each region is greater than 4, so the sample size assumptions have been met. # Univariate Outliers

```
library(rstatix)
Guns1978 %>%
  group_by(Region) %>%
  identify_outliers(violent)
```

```
## # A tibble: 2 x 16
##   Region     year violent murder robbery prisoners  afam  cauc  male population
##   <chr>     <dbl>   <dbl>  <dbl>   <dbl>     <dbl> <dbl> <dbl> <dbl>      <dbl>
## 1 NorthEast  1978    841    10.3    472.       108  5.39  62.8  17.0     17.7
## 2 South      1978   1412.   28      940.       330 25.8   22.0  18.0      0.665
## # … with 6 more variables: income <dbl>, density <dbl>, state <chr>, law <chr>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

```
Guns1978 %>%
  group_by(Region) %>%
  identify_outliers(murder)
```

```
## # A tibble: 2 x 16
##   Region     year violent murder robbery prisoners  afam  cauc  male population
##   <chr>     <dbl>   <dbl>  <dbl>   <dbl>     <dbl> <dbl> <dbl> <dbl>      <dbl>
## 1 NorthEast  1978    841    10.3    472.       108  5.39  62.8  17.0     17.7
## 2 South      1978   1412.   28      940.       330 25.8   22.0  18.0      0.665
## # … with 6 more variables: income <dbl>, density <dbl>, state <chr>, law <chr>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

```
Guns1978 %>%
  group_by(Region) %>%
  identify_outliers(robbery)
```

```
## # A tibble: 4 x 16
##   Region     year violent murder robbery prisoners  afam  cauc  male population
##   <chr>     <dbl>   <dbl>  <dbl>   <dbl>     <dbl> <dbl> <dbl> <dbl>      <dbl>
## 1 NorthEast  1978    841    10.3    472.       108  5.39  62.8  17.0     17.7
## 2 South      1978   1412.   28      940.       330 25.8   22.0  18.0      0.665
## 3 South      1978    732     8.2    310.       198  8.57  59.5  18.4      4.18
## 4 West       1978    781.   15.5    360.       187  3.75  70.3  18.3      0.719
## # … with 6 more variables: income <dbl>, density <dbl>, state <chr>, law <chr>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

```
Guns1978 %>%
  group_by(Region) %>%
  identify_outliers(prisoners)
```

```
## # A tibble: 5 x 16
##   Region     year violent murder robbery prisoners  afam  cauc  male population
##   <chr>     <dbl>   <dbl>  <dbl>   <dbl>     <dbl> <dbl> <dbl> <dbl>      <dbl>
## 1 MidWest    1978    577.   10.6   219.        151 4.67   64.8  18.6      9.22
## 2 NorthEast  1978    119.    1.4    26.3        26 0.380  73.9  18.2      0.892
## 3 NorthEast  1978    424.    5.4   193.         78 4.77   64.9  16.9      7.35
## 4 NorthEast  1978    841    10.3   472.        108 5.39   62.8  17.0     17.7
## 5 West       1978    781.   15.5   360.        187 3.75   70.3  18.3      0.719
## # … with 6 more variables: income <dbl>, density <dbl>, state <chr>, law <chr>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

There is at least one extreme outlier present in all of the key numerical variables.

Check Multivariate Normality Assumption

```
library(rstatix)
Guns1978 %>%
  select(murder,robbery,prisoners,violent) %>%
  mshapiro_test()
```

```
## # A tibble: 1 x 2
##   statistic  p.value
##       <dbl>    <dbl>
## 1     0.586 9.33e-11
```

The multivariate normality Shapiro test is significant (p-value < 0.05) so we can't assume multivariate normality.

Identifying Multicollineraity

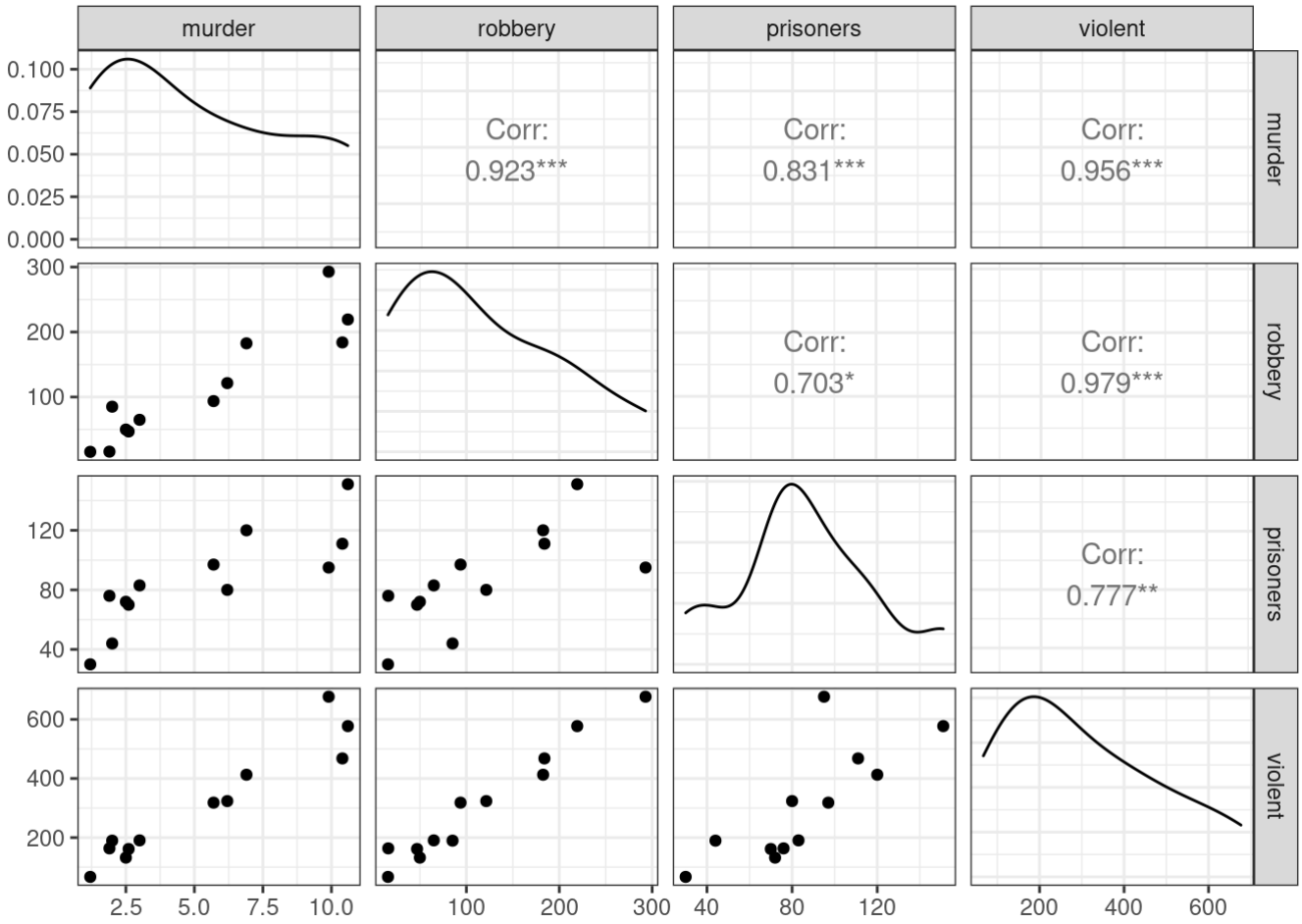
```
Guns1978 %>%
  cor_test(murder,robbery,prisoners,violent)
```

```
## # A tibble: 16 x 8
##    var1      var2        cor    statistic        p conf.low conf.high method
##    <chr>     <chr>     <dbl>        <dbl>    <dbl>    <dbl>     <dbl> <chr>
##  1 murder    murder     1          Inf    0.          1         1     Pearson
##  2 murder    robbery    0.72         7.18 3.52e- 9    0.549     0.828 Pearson
##  3 murder    prisoners  0.77         8.36 5.36e-11    0.623     0.861 Pearson
##  4 murder    violent    0.8          9.36 1.71e-12    0.674     0.882 Pearson
##  5 robbery   murder     0.72         7.18 3.52e- 9    0.549     0.828 Pearson
##  6 robbery   robbery    1          Inf    0.          1         1     Pearson
##  7 robbery   prisoners  0.59         5.12 5.16e- 6    0.376     0.745 Pearson
##  8 robbery   violent    0.9         14.7  1.19e-19    0.836     0.944 Pearson
##  9 prisoners murder     0.77         8.36 5.36e-11    0.623     0.861 Pearson
## 10 prisoners robbery    0.59         5.12 5.16e- 6    0.376     0.745 Pearson
## 11 prisoners prisoners  1    469762048.   0.          1.00      1.00  Pearson
## 12 prisoners violent    0.74         7.62 7.32e-10    0.578     0.841 Pearson
## 13 violent   murder     0.8          9.36 1.71e-12    0.674     0.882 Pearson
## 14 violent   robbery    0.9         14.7  1.19e-19    0.836     0.944 Pearson
## 15 violent   prisoners  0.74         7.62 7.32e-10    0.578     0.841 Pearson
## 16 violent   violent    1          Inf    0.          1         1     Pearson
```

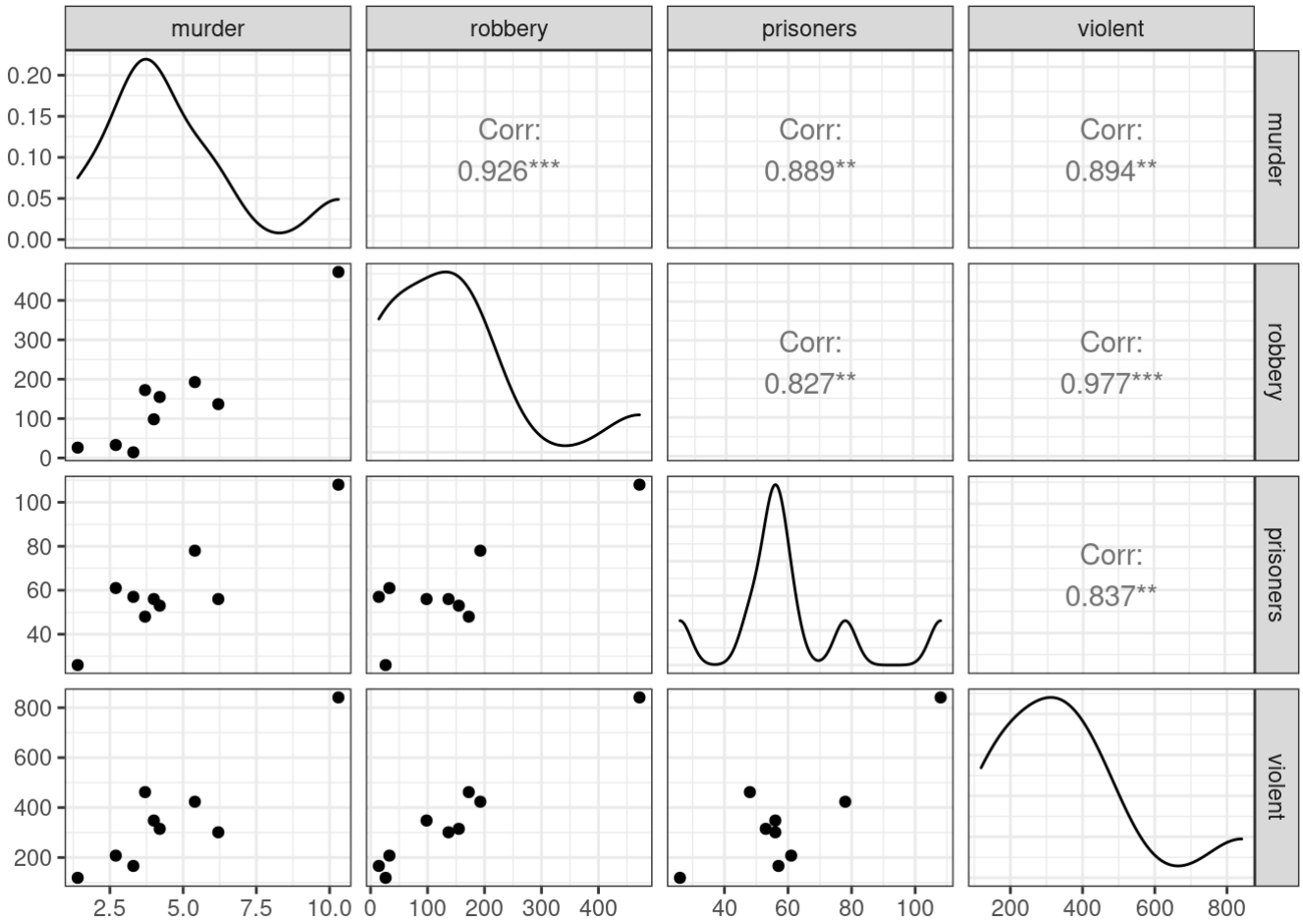
There is no multicollinearity, as assessed by the respective Pearson correlations (p-values less than 0.00001).

```
# Create a scatterplot matrix by group
library(GGally)
results <- Guns1978 %>%
  select(murder,robbery,prisoners,violent,Region) %>%
  group_by(Region) %>%
  doo(~ggpairs(.) + theme_bw(), result = "plots")
results$plots
```

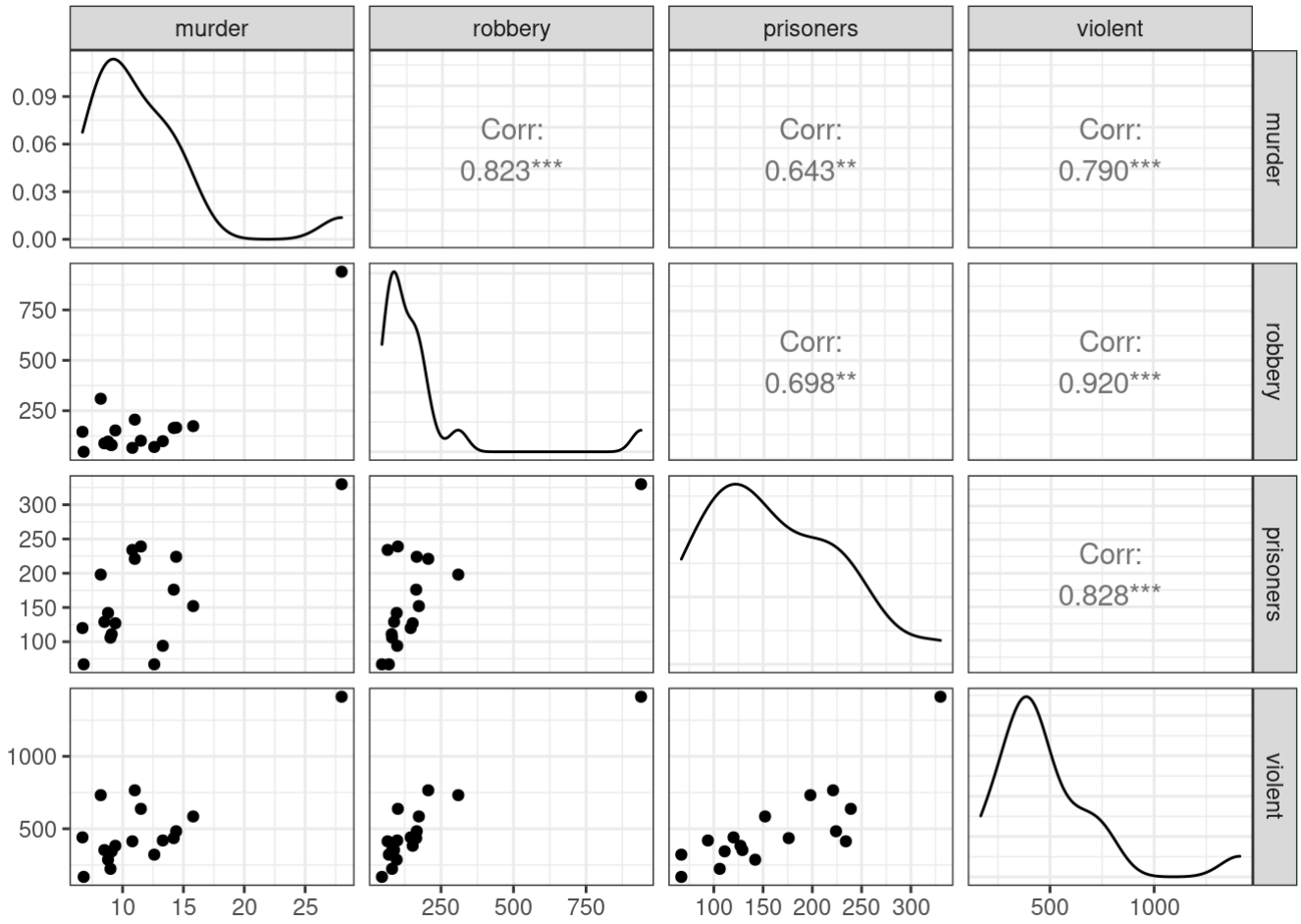
```
## [[1]]
```


murder
robbery
prisoners
violent
Corr:
0.923***
Corr:
0.831***
Corr:
0.956***
Corr:
0.703*
Corr:
0.979***
Corr:
0.777**

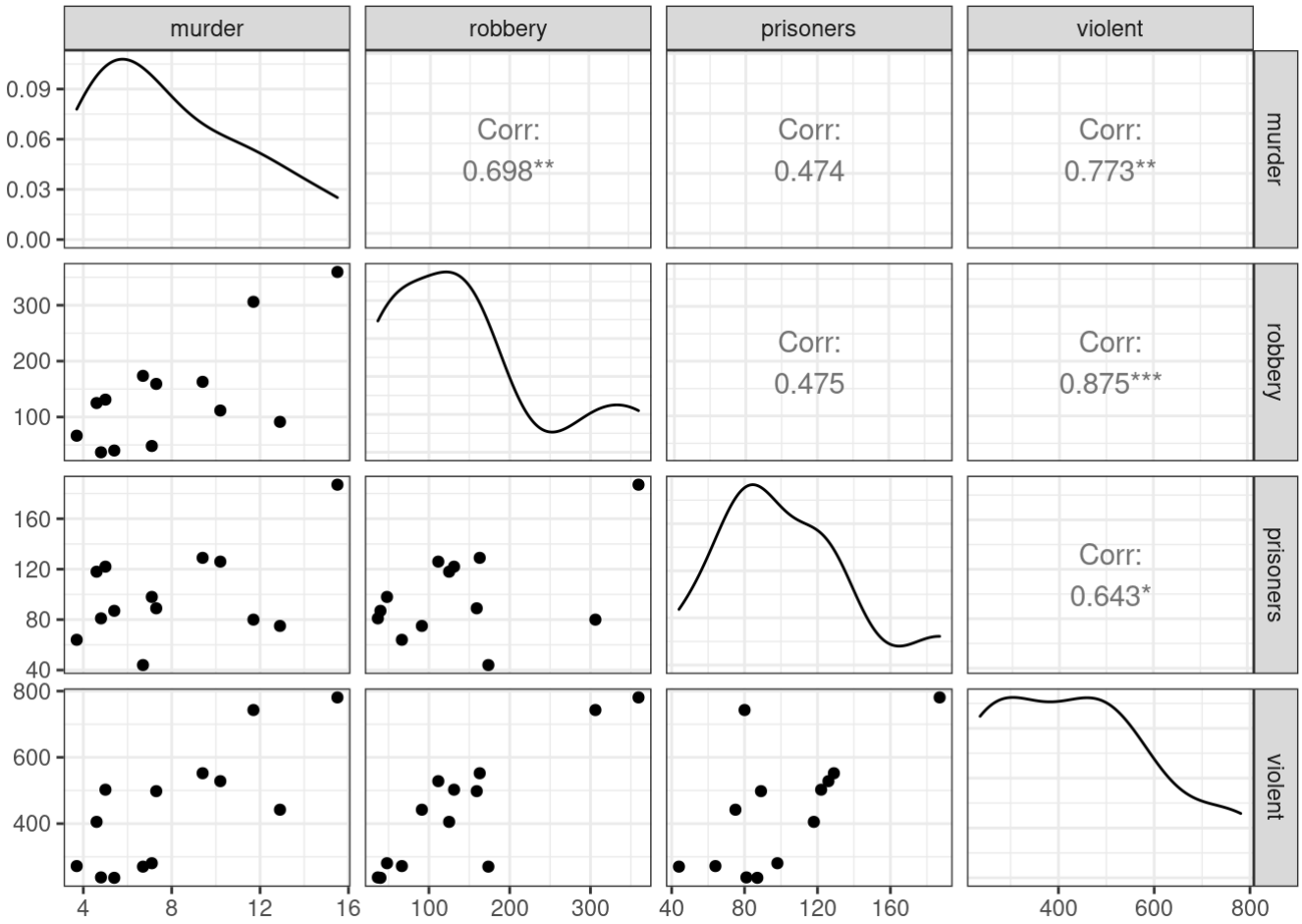
```
##
## [[2]]
```


murder
robbery
prisoners
violent
Corr:
0.926***
Corr:
0.889**
Corr:
0.894**
Corr:
0.827**
Corr:
0.977***
Corr:
0.837**

```
##
## [[3]]
```




```
##
## [[4]]
```



There is a linear relationship between the murder, violent, robbery, and prisoner rates in each Region group, as assessed by the scatter plots.

```
library(rstatix)
box_m(Guns1978[,c("murder","robbery","prisoners","violent")],Guns1978$Region)
```

```
## # A tibble: 1 x 4
##   statistic   p.value parameter method
##       <dbl>     <dbl>     <dbl> <chr>
## 1      73.6 0.0000159        30 Box's M-test for Homogeneity of Covariance Matr…
```

Box's M-test for Homogeneity of Covariance Matrices is statistically significant (p < 0.0001) so the data has violated the assumption of homogeneity of variance-covariance matrices.

```
Guns1978 %>%
  gather(key = "variable", value = "value", murder,robbery,prisoners,violent) %>%
  group_by(variable) %>%
  levene_test(value ~ Region)
```

```
## # A tibble: 4 x 5
##   variable    df1   df2 statistic       p
##   <chr>     <int> <int>     <dbl>   <dbl>
## 1 murder        3    47     0.626 0.602
## 2 prisoners     3    47     4.40  0.00824
## 3 robbery       3    47     0.167 0.918
## 4 violent       3    47     0.112 0.952
```

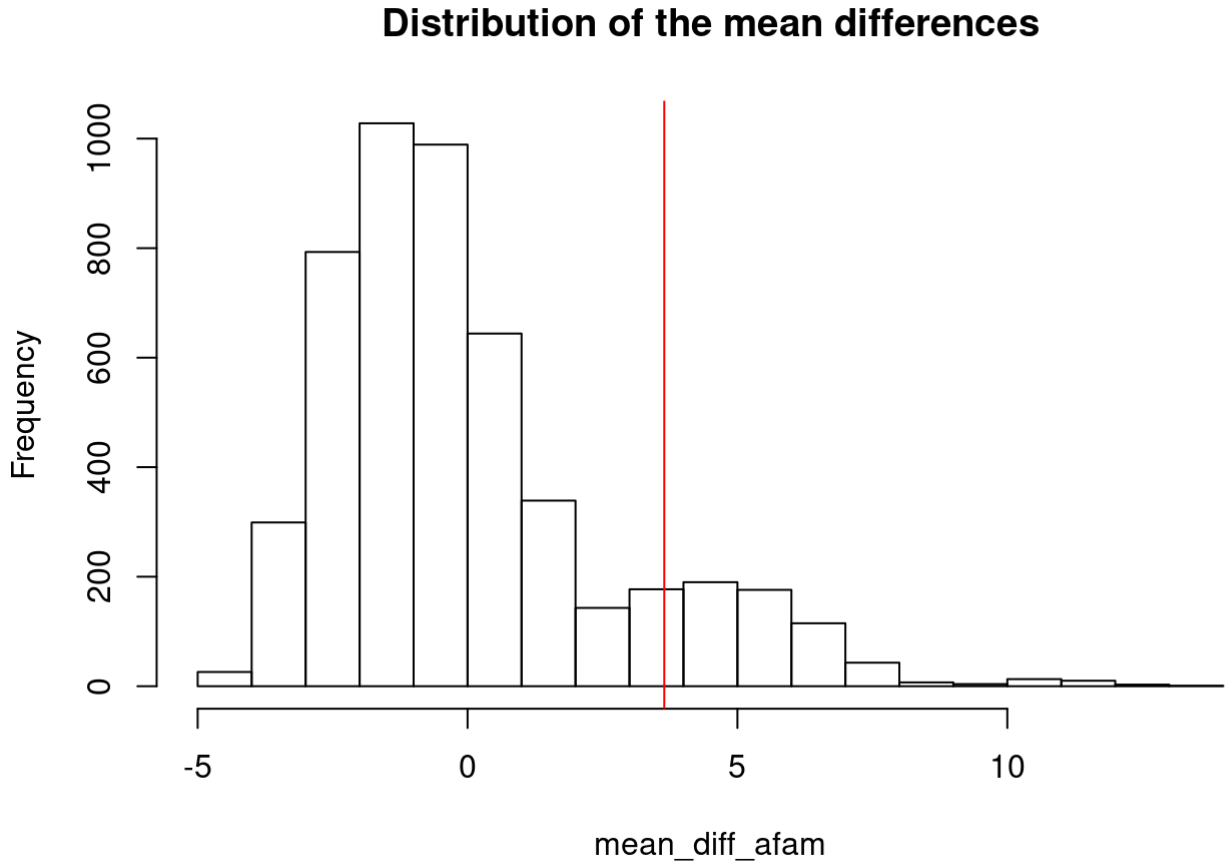
The Levene's test is not significant for any of the variables (p > 0.05) so there is homogeneity of variances. In summary, this dataset violates two of MANOVA's key assumptions. I can only assume that the observations were independent and random.

Randomization Test

```
obs_diff <- mean(Guns1978$afam[Guns1978$law == 'no']) - mean(Guns1978$afam[Guns1978$law == 'ye
s'])
```

```
set.seed(348)
# 5000 Randomizations finding mean with original weight data
# Find the new mean difference
mean_diff_afam <- vector()
# Create many randomizations with a for loop
for(i in 1:5000){
  temp <- data.frame(carryLaw = Guns1978$law, black = sample(Guns1978$afam))
  mean_diff_afam[i] <- temp %>%
    group_by(carryLaw) %>%
    summarize(means = mean(black)) %>%
    summarize(mean_diff = diff(means)) %>%
    pull
}
```

```
{hist(mean_diff_afam, main="Distribution of the mean differences"); abline(v = obs_diff, col="r
ed")}
```




```
mean(mean_diff_afam > obs_diff)
```

```
## [1] 0.128
```

Null Hypothesis: There is not a difference between the mean proportions of African Americans aged between 10 and 64 in states with a shall carry law not in effect and a a shall carry law in effect in 1978 America. Alternative Hypothesis: There is a difference between the mean proportions of African Americans aged between 10 and 64 in states with a shall carry law not in effect and a a shall carry law in effect in 1978 America. We don't have statistically strong evidence to reject the null hypothesis that there is a difference between the mean proportions of African Americans aged between 10 and 64 in states with a shall carry law not in effect and a a shall carry law in effect in 1978 America (p-value > 0.05).

Linear Regression Model

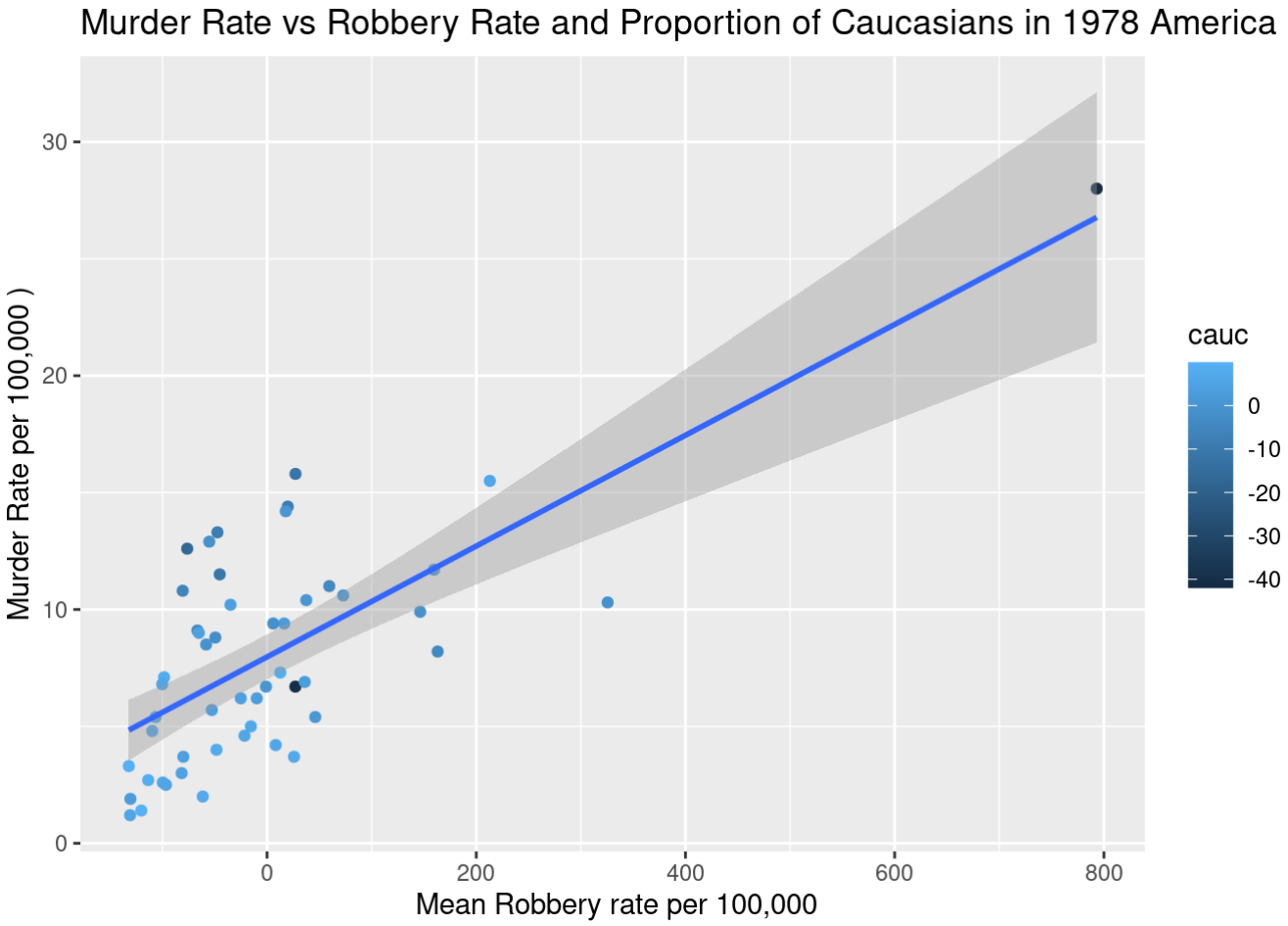
```
Guns <- Guns1978
Guns$cauc <- Guns$cauc - mean(Guns$cauc, na.rm = TRUE)
Guns$robbery <- Guns$robbery - mean(Guns$robbery, na.rm = TRUE)
fit <- lm(murder ~ robbery + cauc + cauc*robbery, data = Guns)
summary(fit)
```

```
##
## Call:
## lm(formula = murder ~ robbery + cauc + cauc * robbery, data = Guns)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5627 -2.0913  0.1817  1.7855  5.8401
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0906829  0.4474949  18.080  < 2e-16 ***
## robbery       0.0195621  0.0046193   4.235 0.000106 ***
## cauc         -0.1952035  0.0534013  -3.655 0.000646 ***
## robbery:cauc  0.0001464  0.0001459   1.004 0.320720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.079 on 47 degrees of freedom
## Multiple R-squared:  0.6205, Adjusted R-squared:  0.5962
## F-statistic: 25.61 on 3 and 47 DF,  p-value: 5.743e-10
```

```
summary(lm(murder ~ robbery + cauc + cauc*robbery, data = Guns))$r.squared
```

```
## [1] 0.6204557
```

```
ggplot(Guns,aes(y=murder,x=robbery,color=cauc)) +
geom_point() + geom_smooth(method = "lm") + xlab(" Mean Robbery rate per 100,000") + ylab("Murd
er Rate per 100,000 )") + ggtitle("Murder Rate vs Robbery Rate and Proportion of Caucasians in
 1978 America")
```



The mean murder rate per 100,000 is 8.0907 if the mean percent of the states' populations that are Caucasian (ages 10 to 64) is 0 and the mean robbery rate per 100,000 in all of the states is 0.

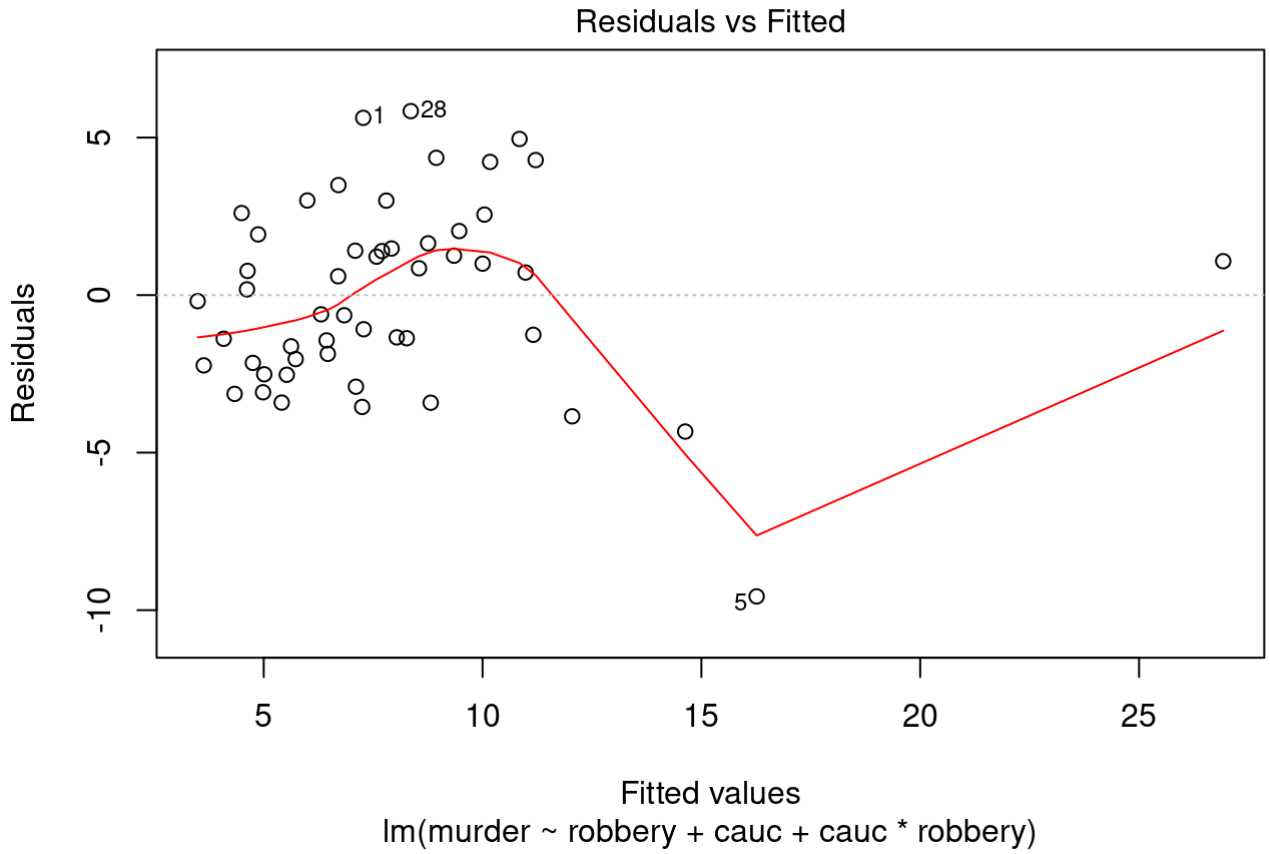
The mean murder rate per 100,000 increases by 0.0195621 (incidents per 100,000) for every increase in mean robbery rate per 100,000 and keeping the mean percent of the states' Caucasian populations constant.

The mean murder rate per 100,000 decreases by 0.1952035 (incidents per 100,000) for every increase in mean percent of the states' Caucasian populations and keeping the mean robbery rate per 100,000 constant.

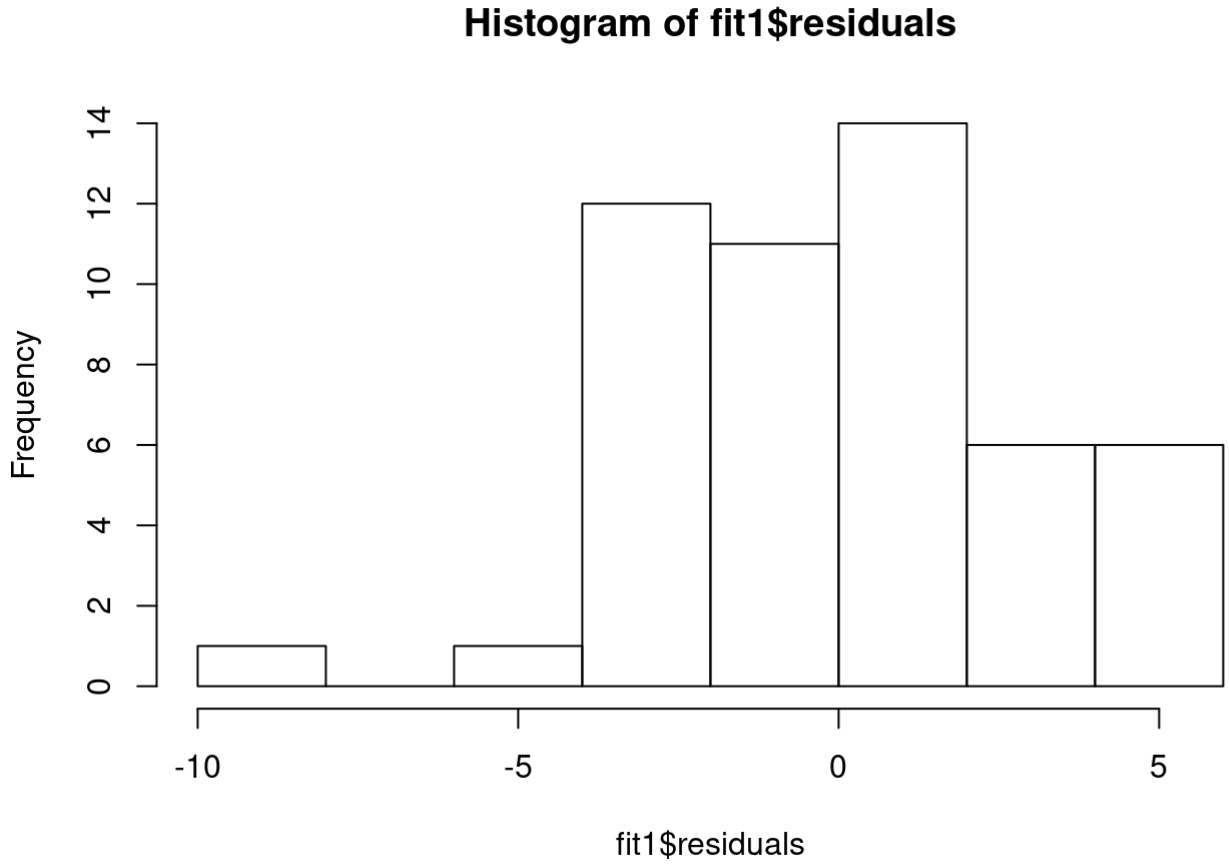
The difference in mean murder rate per 100,000 decreases by 0.0001464 if the percent of a states' populations is Caucasian and commits high robbery rates compared to the states' population not having a high Caucasian proportion and having lower incidents of robbery rates.

62.04557% of the total variation in murder rates per 100,000 in 1978 Americ ca can be explained by robbery rates per 100,000 in 1978 America and the proportion of Caucasians in 1978 America.

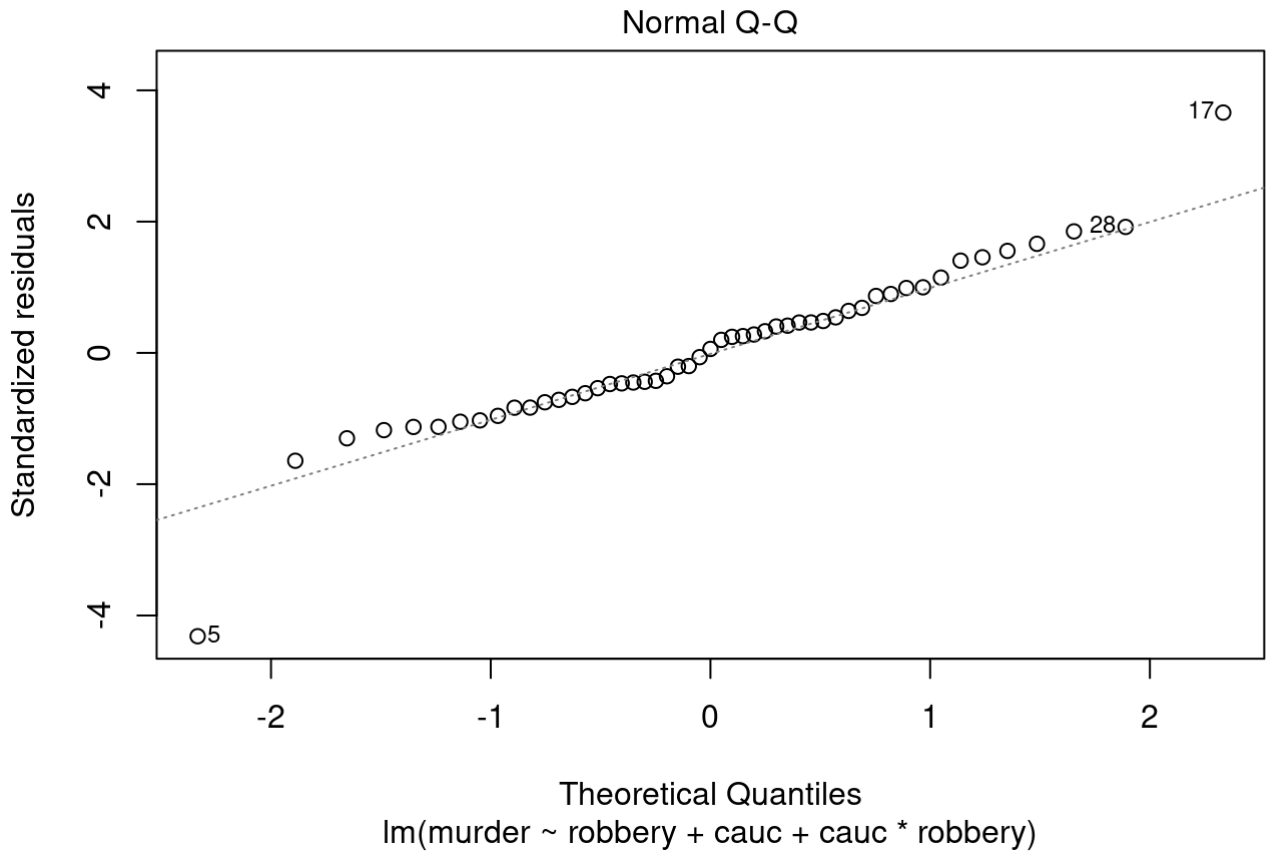
```
Guns$robbery <- Guns$robbery - mean(Guns$robbery,na.rm = TRUE)
Guns$cauc <- Guns$cauc - mean(Guns$cauc,na.rm = TRUE)
fit1 <- lm(murder ~ robbery + cauc + cauc*robbery, data = Guns)

plot(fit1, which = 1)
```




```
hist(fit1$residuals)
```




```
plot(fit1, which = 2)
```




```
library(sandwich)
library(lmtest)
shapiro.test(fit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.97015, p-value = 0.2243
```

```
ks.test(fit1$residuals,"pnorm",mean = 0,sd(fit$residuals))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fit1$residuals
## D = 0.075142, p-value = 0.9148
## alternative hypothesis: two-sided
```

```
bptest(fit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 23.666, df = 3, p-value = 2.934e-05
```

The normality assumption has been met as there isn't a clear pattern in the residuals and a good majority of the points lie on the straight line in the QQ plot (linearity assumption has been met). The Shapiro-Wilk Test does not fail for the residuals so the residuals originated from a normal distribution. The Kolmogorov-Smirnov test fails to reject the null hypothesis that the distribution of residuals follow the normal distribution. However, the equal variance assumption has not been met with the results of the Breusch-Pagan test (homoscedasticity).

```
# Compare with robust SEs
print(coeftest(fit,vcov. = vcovHAC))
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)   8.09068287  0.50568546 15.9994 < 2.2e-16 ***
## robbery       0.01956205  0.00552884  3.5382 0.0009192 ***
## cauc         -0.19520348  0.12399645 -1.5743 0.1221340
## robbery:cauc  0.00014639  0.00015505  0.9441 0.3499248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the robust SE's, the difference in the interaction between robbery and cauc variables for standard error is negligible, in that the interaction term is still not statistically significant. For the cauc variable, this variable is no longer statistically significant compared to the original model because its p-value is now greater than 0.05 and the delta between the standard errors are far larger. For the robbery variable, there isn't too much of a standard error difference so this variable is still statistically significant.

```
##  Bootstrap from residuals
# Repeat bootstrapping 5000 times, saving the coefficients each time
resids_SEs <- replicate(5000, {
  # Bootstrap your residuals (resample with replacement)
  new_resids <- sample(fit$residuals, replace = TRUE)
  # Consider a new response as fitted values plus residuals
  boot_data <- Guns1978
  boot_data$new_y = fit$fitted.values + new_resids
  # Fit regression model
  fitboot <- lm(new_y ~ robbery + cauc + cauc*robbery, data = boot_data)
  # Save the coefficients
  coef(fitboot)
})

# Estimated SEs
resids_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)
```

```
##   (Intercept)     robbery       cauc robbery:cauc
## 1    3.985098 0.006959218 0.06017861 0.0001384663
```

The interaction terms are still not statistically significant when compared between the original model and the Bootstrap SE model. For the cauc variable, there is a difference in the standard error but it's still statistically significant as the p-values would still be way less than 0.05. For the robbery variable, there is a difference in the standard error but it's still statistically significant as the p-values would still be way less than 0.05.

Logistic Regression

```
# 1 means law in effect, 0 means law not in effect
Guns1978 <- Guns1978 %>%
  mutate(y = ifelse(law == "yes", 1, 0))
fit1 <- glm(y ~ robbery + murder, data= Guns1978, family = "binomial")
summary(fit1)
```

```
##
## Call:
## glm(formula = y ~ robbery + murder, family = "binomial", data = Guns1978)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8996  -0.4452  -0.2420  -0.1304   2.3580
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.243315   1.072572  -0.227    0.821
## robbery     -0.003339   0.011473  -0.291    0.771
## murder      -0.333459   0.269343  -1.238    0.216
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.042  on 50  degrees of freedom
## Residual deviance: 23.049  on 48  degrees of freedom
## AIC: 29.049
##
## Number of Fisher Scoring iterations: 7
```

```
exp(coefficients(fit1))
```

```
## (Intercept)     robbery      murder
##   0.7840247   0.9966661   0.7164409
```

While holding other predictors constant, the odds of a shall carry law in effect for a robbery charge increase.While holding other predictors constant ,the odds of a shall carry law for a murder charge increase.

```
# Based on predicted probabilities...
Guns1978$prob1 <- predict(fit1, type = "response")

# ... we can classify a clump as malignant or not (for example apply a cutoff of 0.5)
Guns1978$predicted <- ifelse(Guns1978$prob1 > 8.744202e-03, "yes", "no")
```

```
# Confusion matrix: compare true to predicted condition
table(true_condition = Guns1978$law, predicted_condition = Guns1978$predicted) %>%
  addmargins
```

```
##               predicted_condition
## true_condition no yes Sum
##            no   9  38  47
##            yes  0   4   4
##            Sum  9  42  51
```

```
# Accuracy (correctly classified cases)
(4+9)/51
```

```
## [1] 0.254902
```

```
# Sensitivity (True Positive Rate, TPR)
4/(4+38)
```

```
## [1] 0.0952381
```

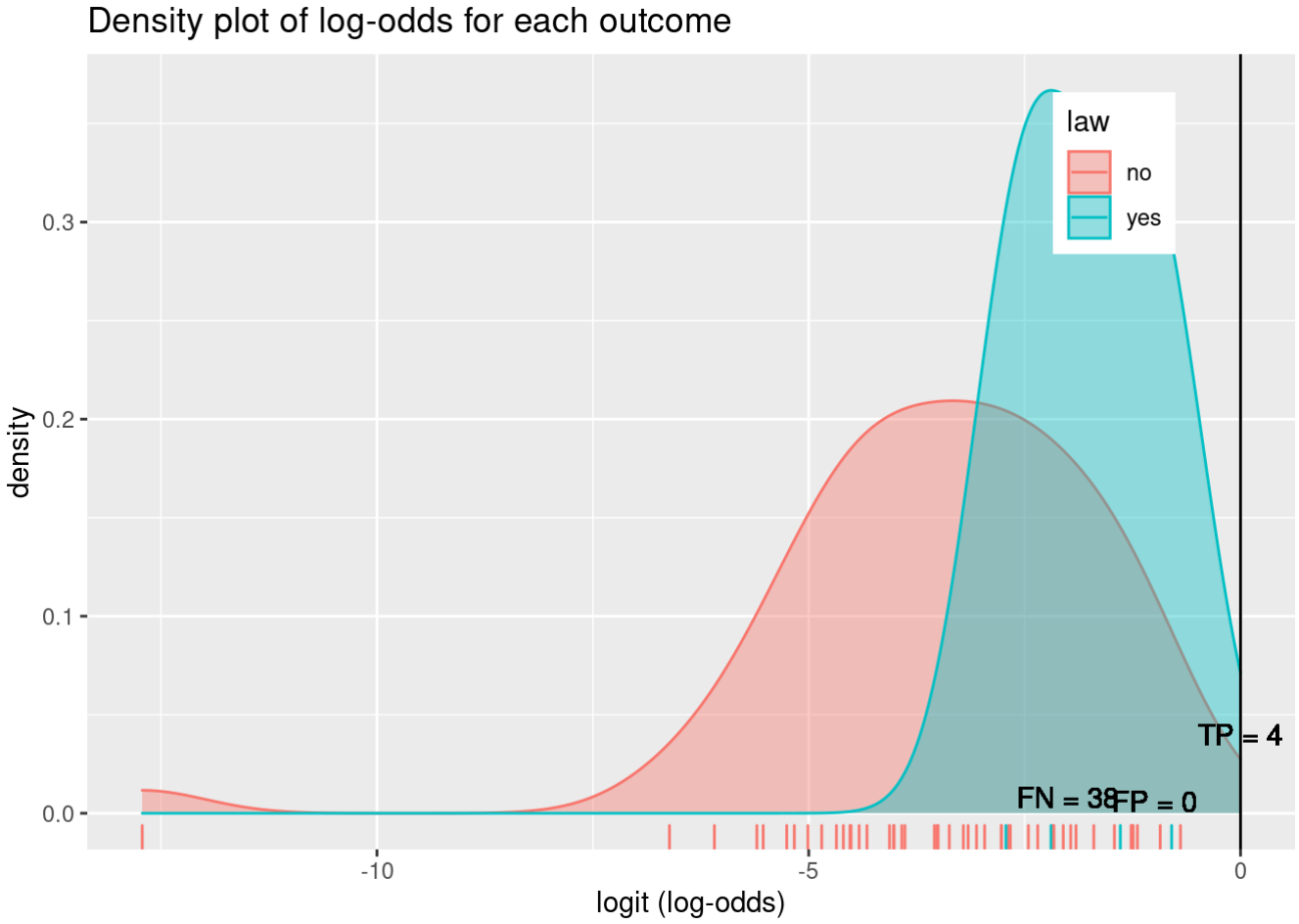
```
# Specificity (True Negative Rate, TNR)
9/(9+38)
```

```
## [1] 0.1914894
```

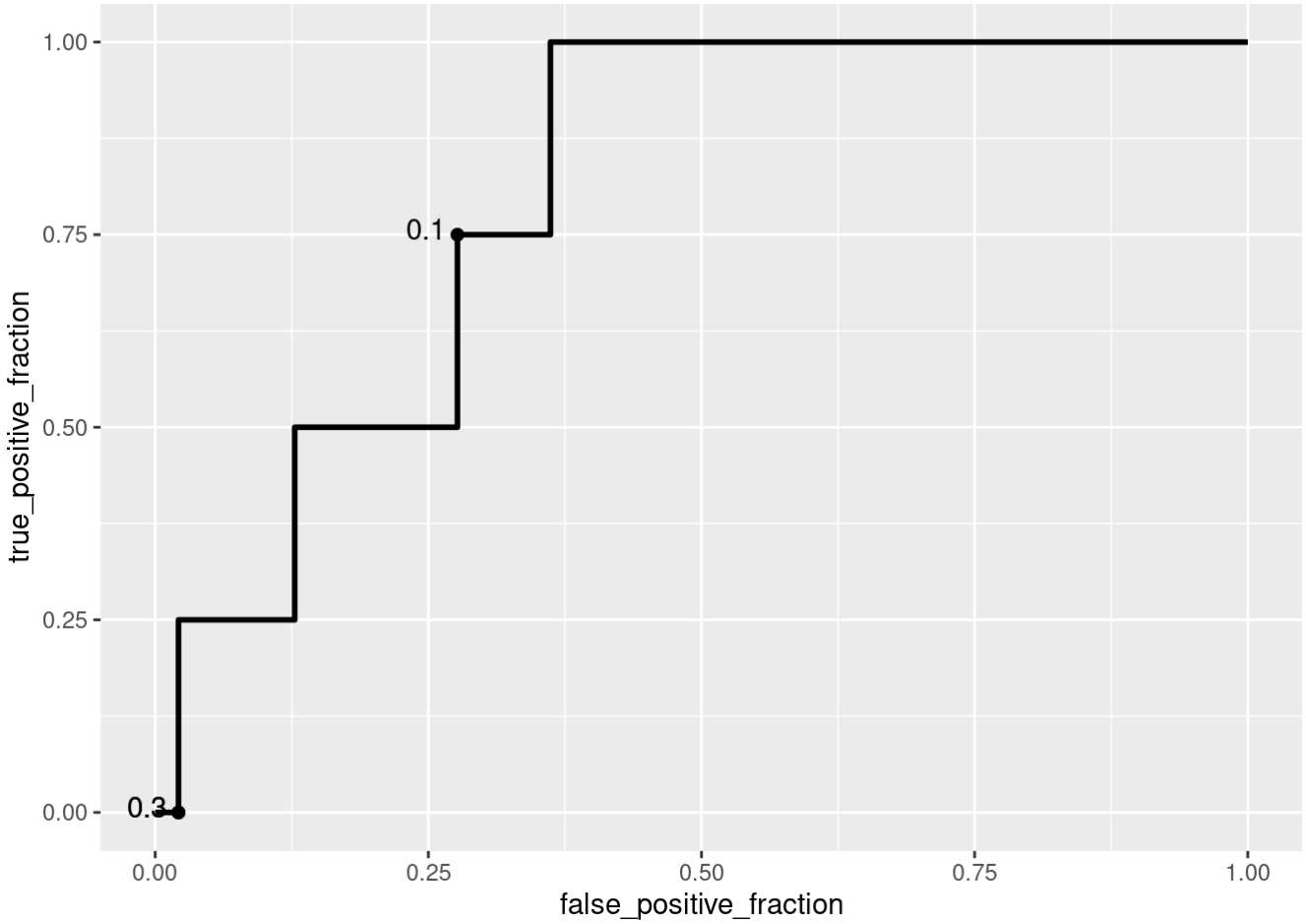
```
# Precision (Positive Predictive Value, PPV)
4/4+0
```

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

```
# Predicted log odds
Guns1978$logit <- predict(fit1, type = "link")
# Density plot of log-odds for each outcome
Guns1978 %>%
  ggplot() +
  geom_density(aes(logit, color = law, fill = law), alpha = .4) +
    geom_rug(aes(logit, color = law)) +
  geom_text(x = -15, y = .07, label = "TN = 9") +
  geom_text(x = -2, y = .008, label = "FN = 38") +
  geom_text(x = -1, y = .006, label = "FP = 0") +
  geom_text(x = 0, y = .04, label = "TP = 4") +
  theme(legend.position = c(.85,.85)) +
  geom_vline(xintercept = 0) +
  xlab("logit (log-odds)") + ggtitle("Density plot of log-odds for each outcome")
```



```
# Call the library plotROC
library(plotROC)

# Plot ROC depending on values of y and its probabilities displaying some cutoff values
ROCplot1 <- ggplot(Guns1978) +
  geom_roc(aes(d = y, m = prob1), cutoffs.at = list(0.1, 0.5, 0.9))
ROCplot1
```




```
# Calculate the area under the curve still using the library plotROC with function calc_auc
calc_auc(ROCplot1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.8031915
```

According to the rule of thumb, this model is excellent in terms of prediction power as its AUC score is greater than 0.9. However, I believe there has been significant overfitting done with my logistic regression model.