HW5 Kelly “Scott” Sims

Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

> I work in the oil industry in which we utilize advanced downhole logging and drilling tools to drill an oil well and hit the payzone. All of these tools have to go through various levels of servicing and parts replacement depending on the amount of abuse they’ve taken downhole throughout its lifetime since the last servicing. There’s various instrumentation on these tools to record the “abuse”. There’s internal thermometers, accelerometers, magnetometers, pressure sensors, etc. A simple regression model could help predict how many hours a tool can be utilized without servicing before it fails. Predictors could be:

1. The amount of time spent at various g-shock severity levels (binned like a histogram)
2. The amount of time spent at various torque levels (binned like a histogram)
3. The amount of time spent at various temperature levels (binned like a histogram)
4. The amount of time spent at various compression levels - the amount of weight pushed down on the tools (binned like a histogram)
5. The total amount of time elapsed since the last servicing
6. The amount of time spenat at various pressure levels (binned like a histogram)

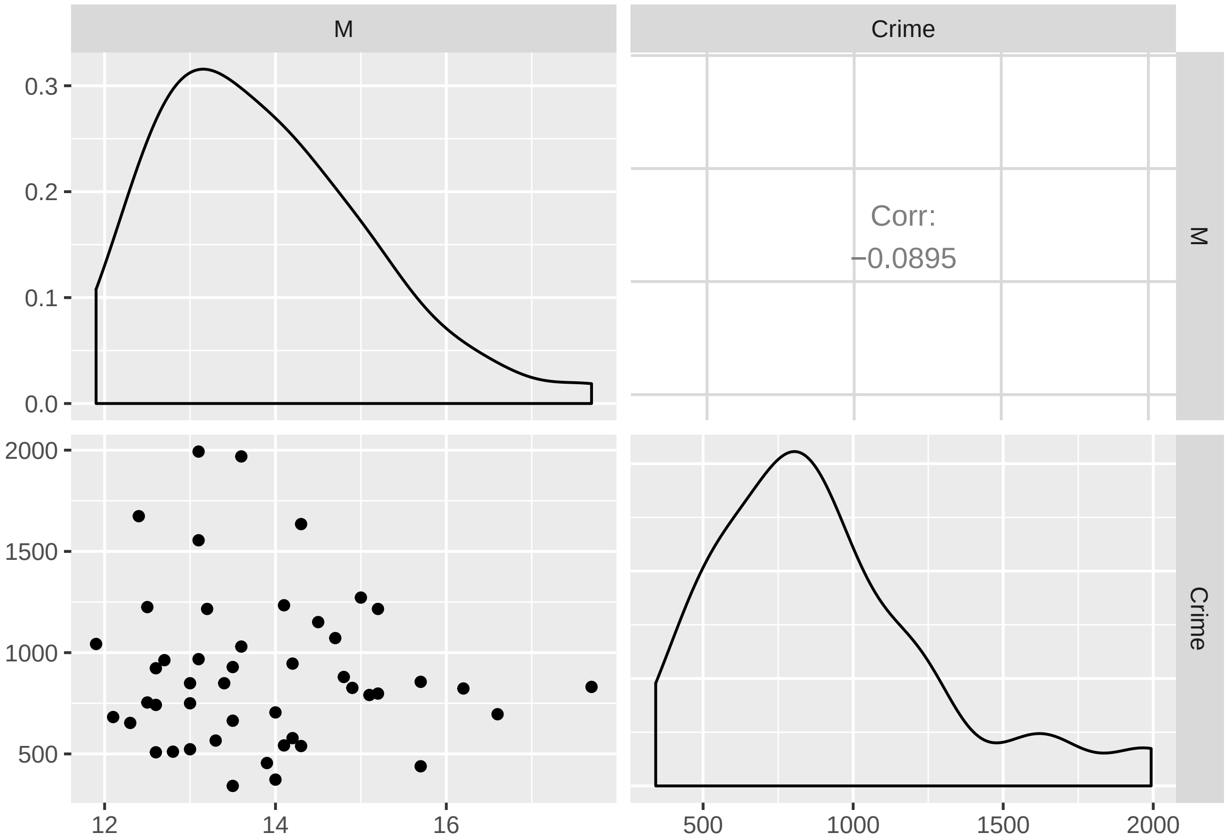
Question 8.2

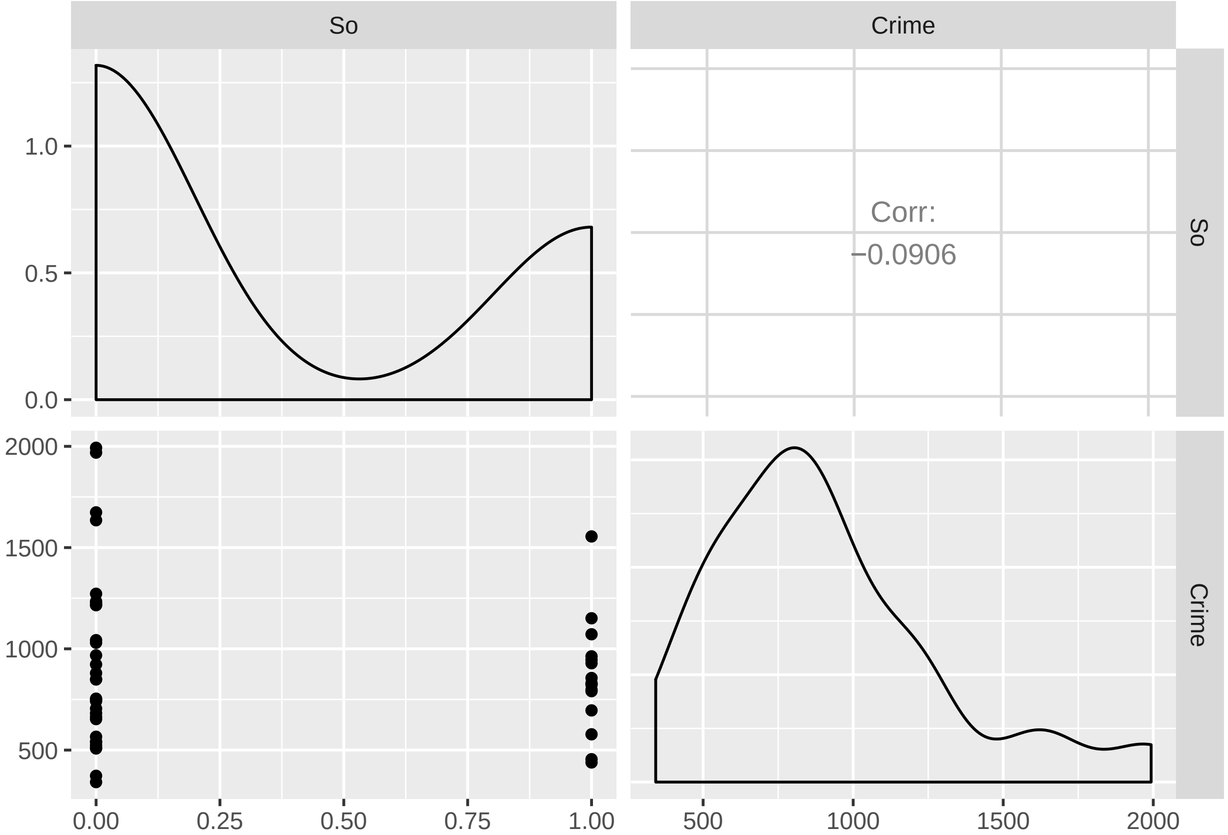
Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data: M = 14.0 So = 0 Ed = 10.0 Po1 = 12.0 Po2 = 15.5 LF = 0.640 M.F = 94.0 Pop = 150 NW = 1.1 U1 = 0.120 U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0

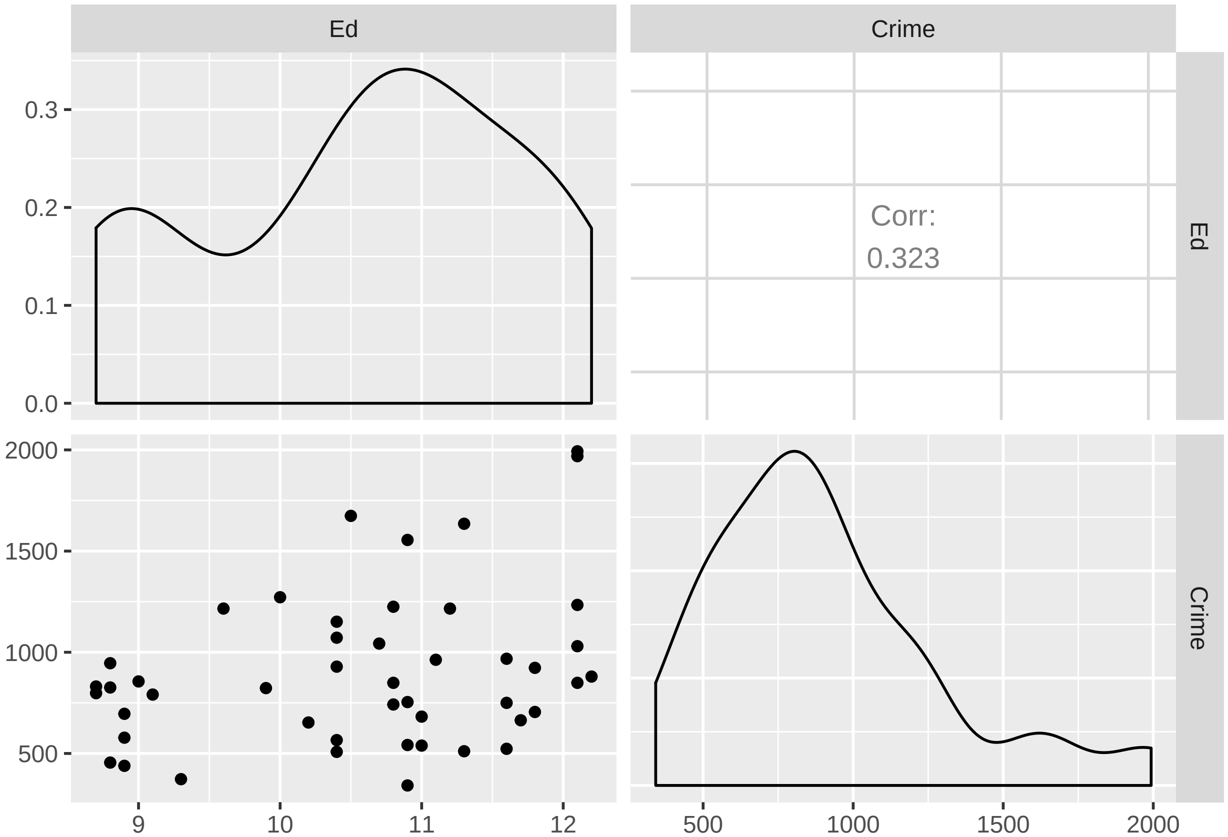
```r
library(ggplot2)
library(GGally)

#load the data
data = read.table('uscrime.txt', header = TRUE, stringsAsFactors = FALSE)
```

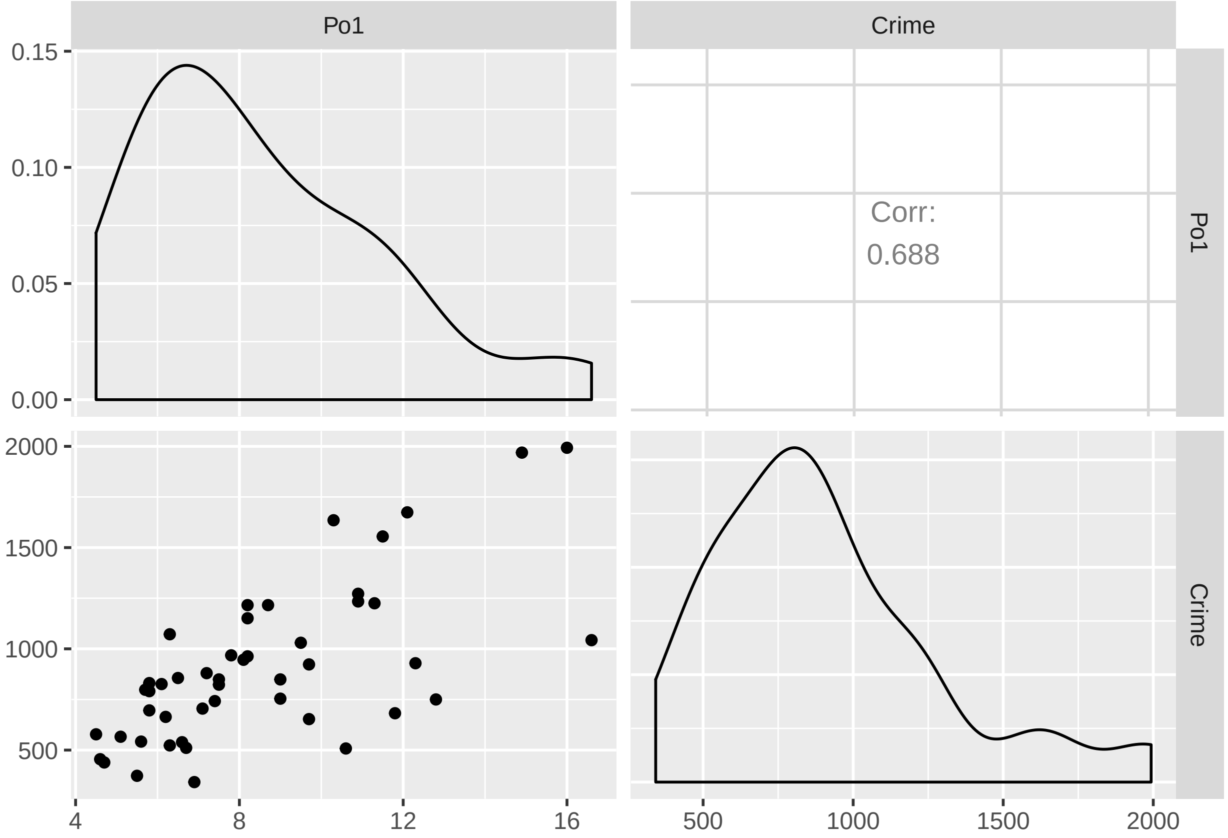
First we will compare each independent variable with the target variable, crime, in search of any correlation.

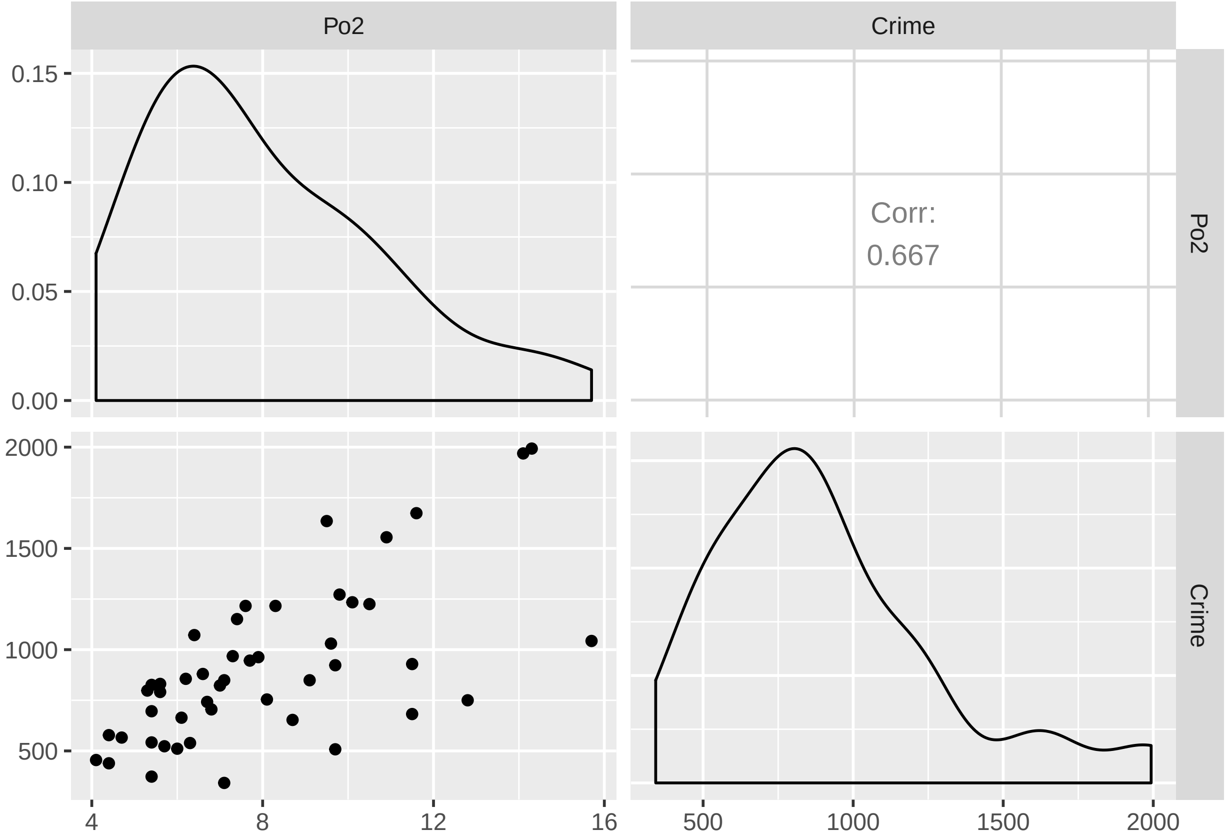
```r
for(i in seq(1,15)){
pm = ggpairs(data=data,
             columns= c(i,16),

             )
print(pm)
}
```

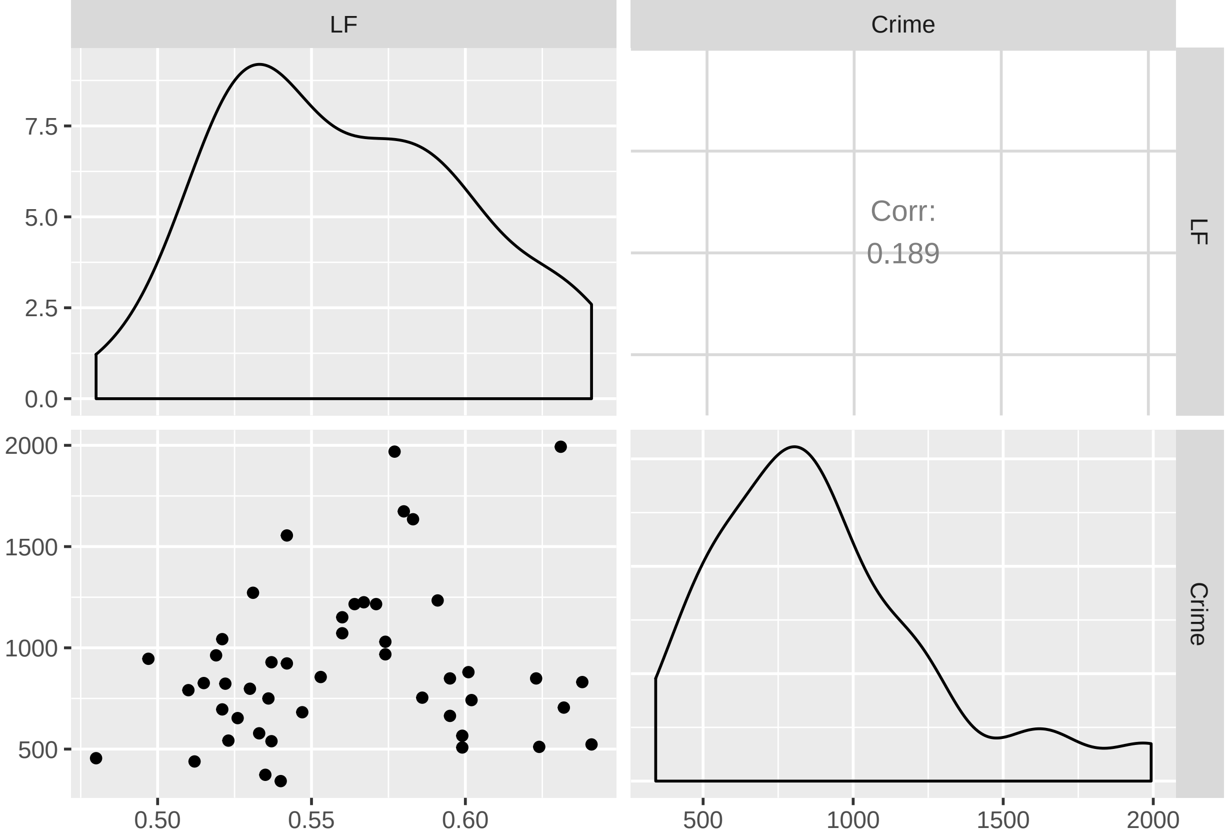

M
Crime
0.3
0.2
0.1
0.0
Corr:
−0.0895
M
2000
1500
1000
500
Crime
12
14
16
500
1000
1500
2000

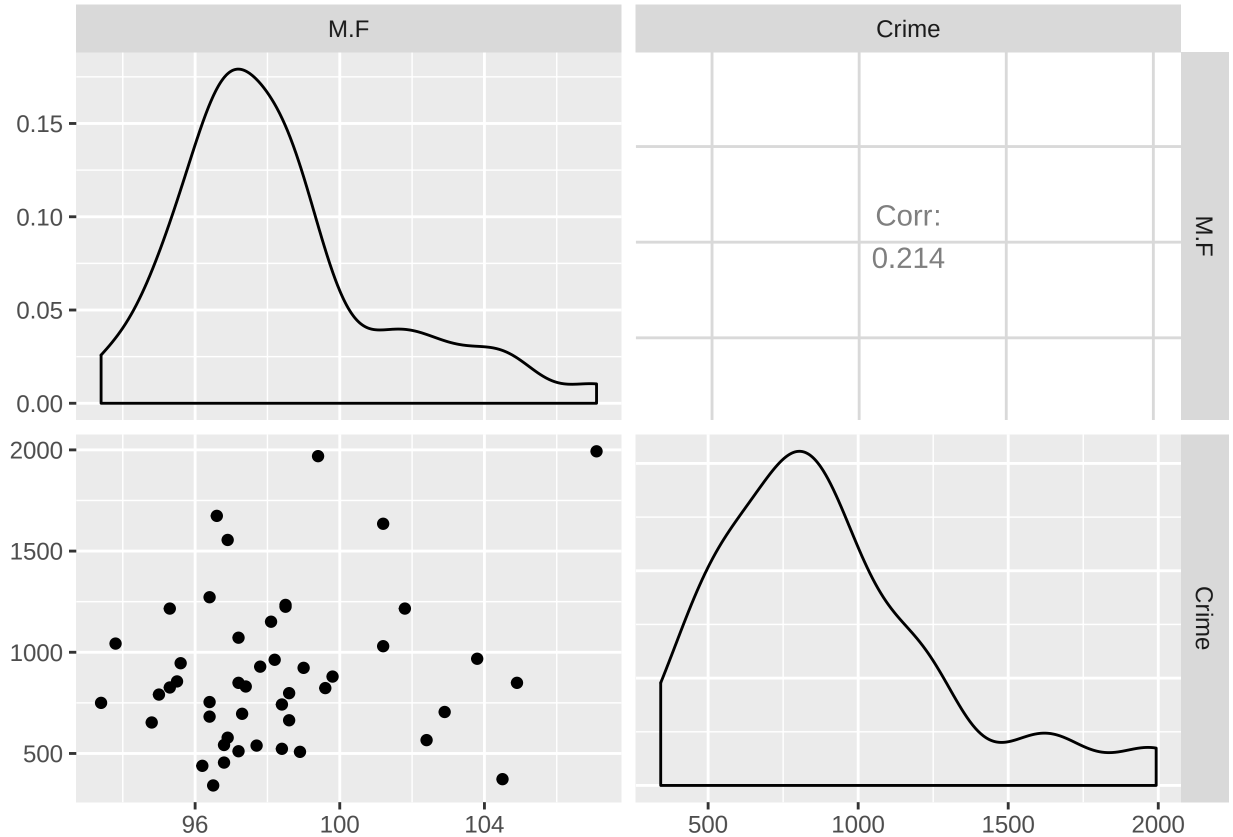

So
Crime
1.0
0.5
0.0
Corr:
−0.0906
So
2000
1500
1000
500
Crime
0.00
0.25
0.50
0.75
1.00
500
1000
1500
2000

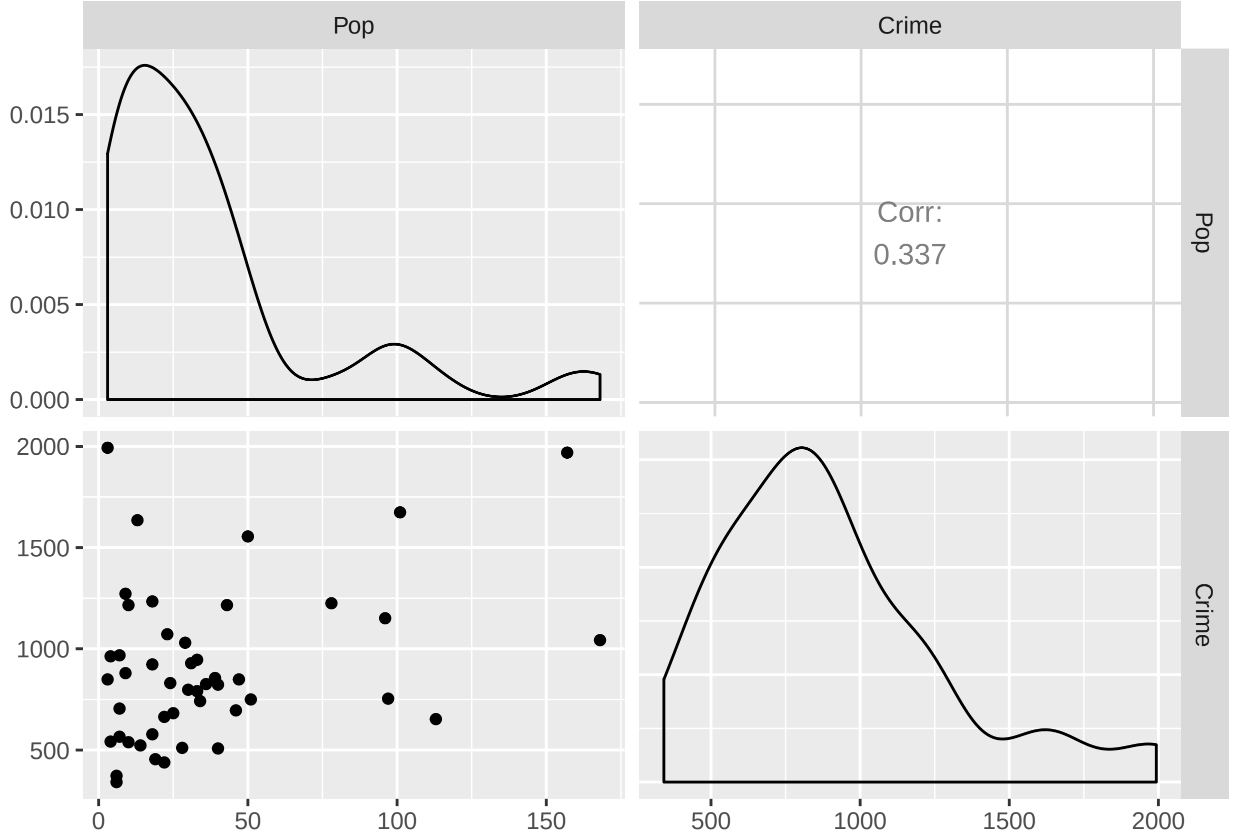

Ed
Crime
0.3
0.2
0.1
0.0
Corr:
0.323
Ed
2000
1500
1000
500
Crime
9
10
11
12
500
1000
1500
2000

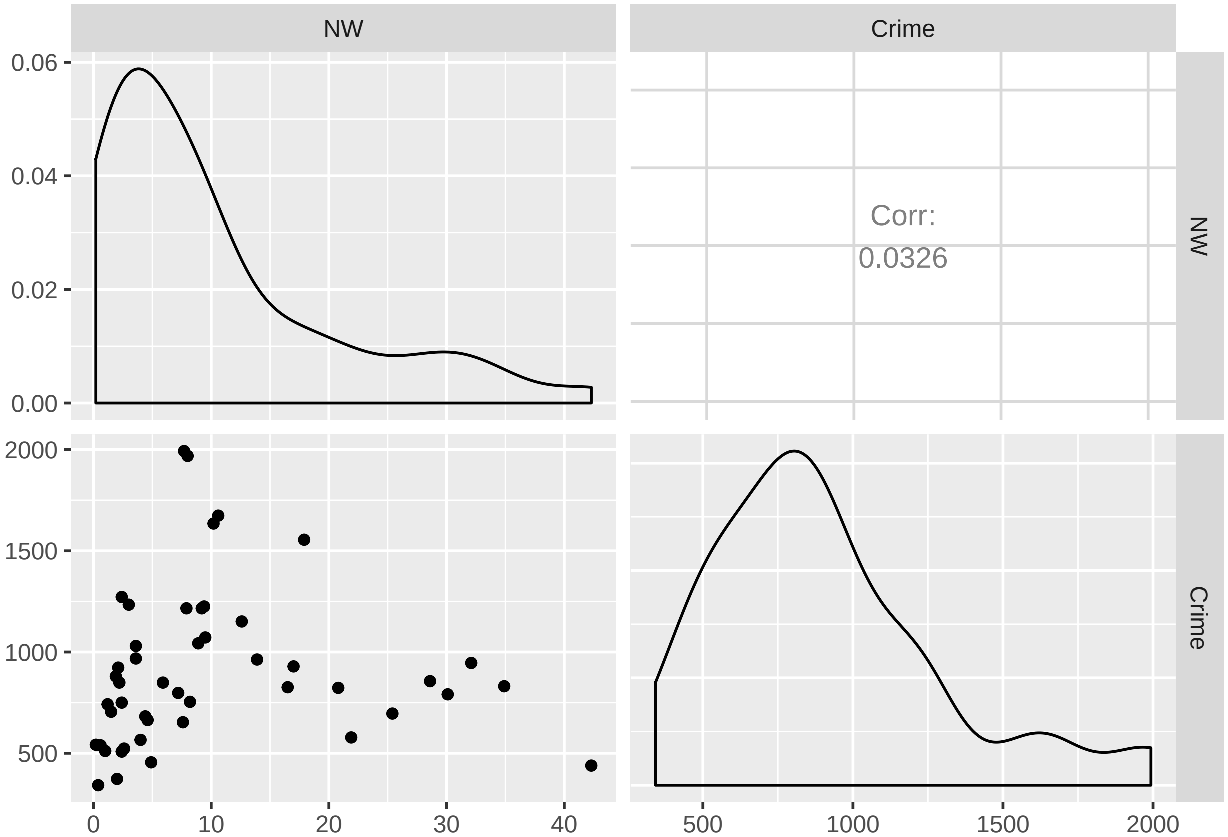

Po1
Crime
0.15
0.10
0.05
0.00
Corr:
0.688
Po1
2000
1500
1000
500
Crime
4
8
12
16
500
1000
1500
2000

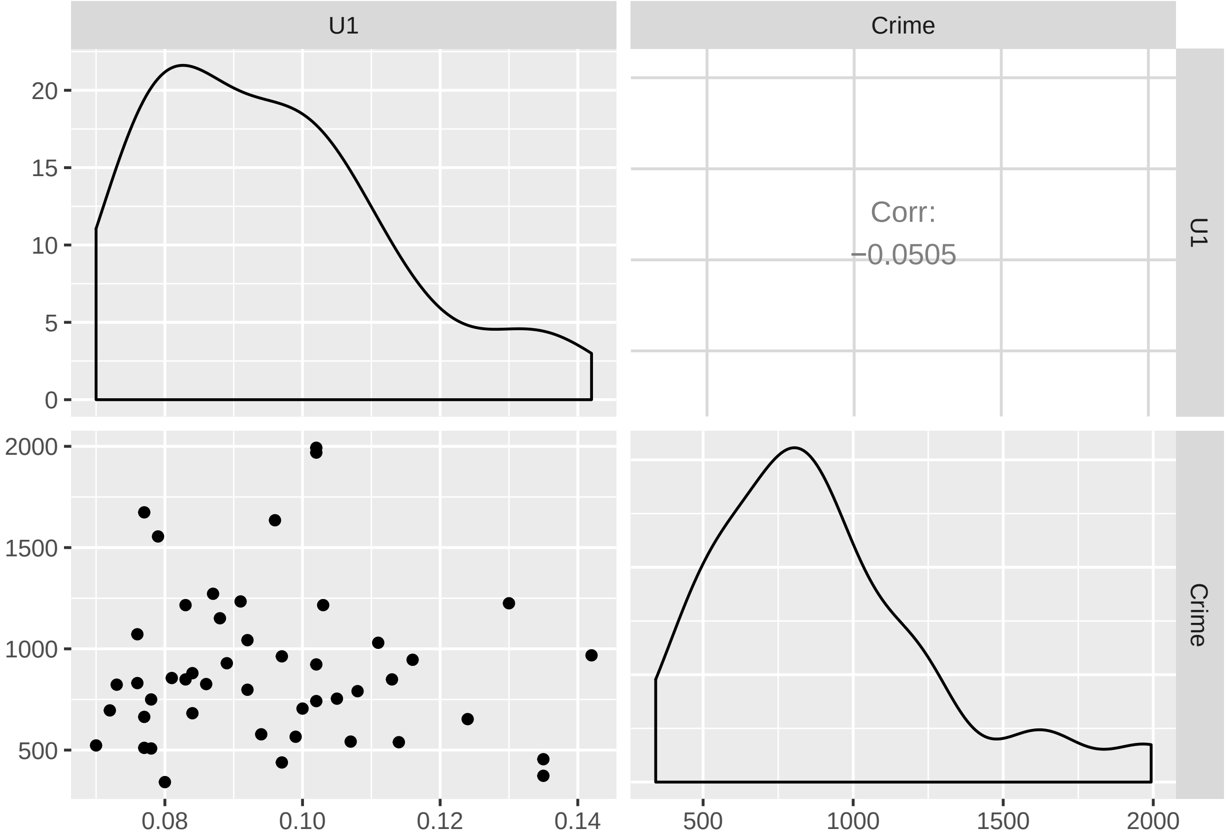

Po2
Crime
0.15
0.10
0.05
0.00
Corr:
0.667
Po2
2000
1500
1000
500
Crime
4
8
12
16
500
1000
1500
2000

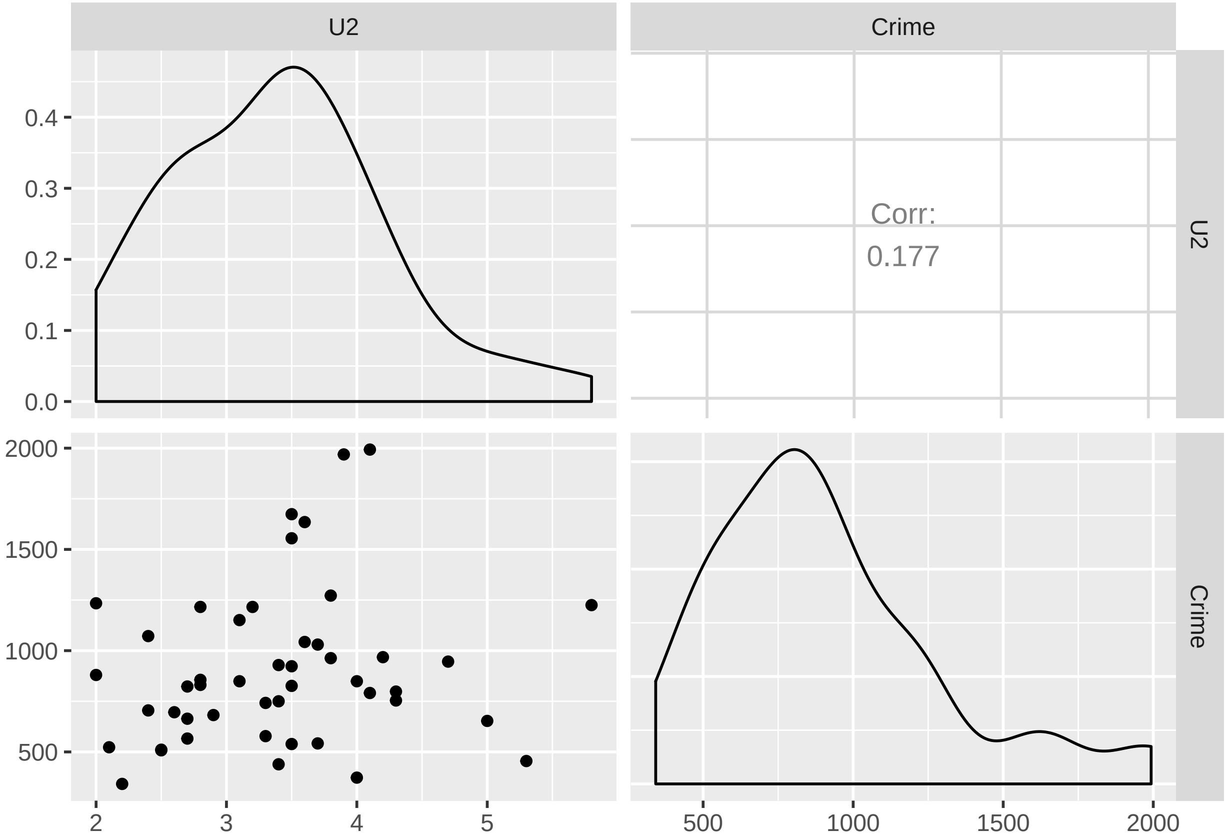

LF
Crime
7.5
5.0
2.5
0.0
Corr:
0.189
LF
2000
1500
1000
500
Crime
0.50
0.55
0.60
500
1000
1500
2000


M.F
Crime
0.15
0.10
0.05
0.00
Corr:
0.214
M.F
2000
1500
1000
500
Crime
96
100
104
500
1000
1500
2000

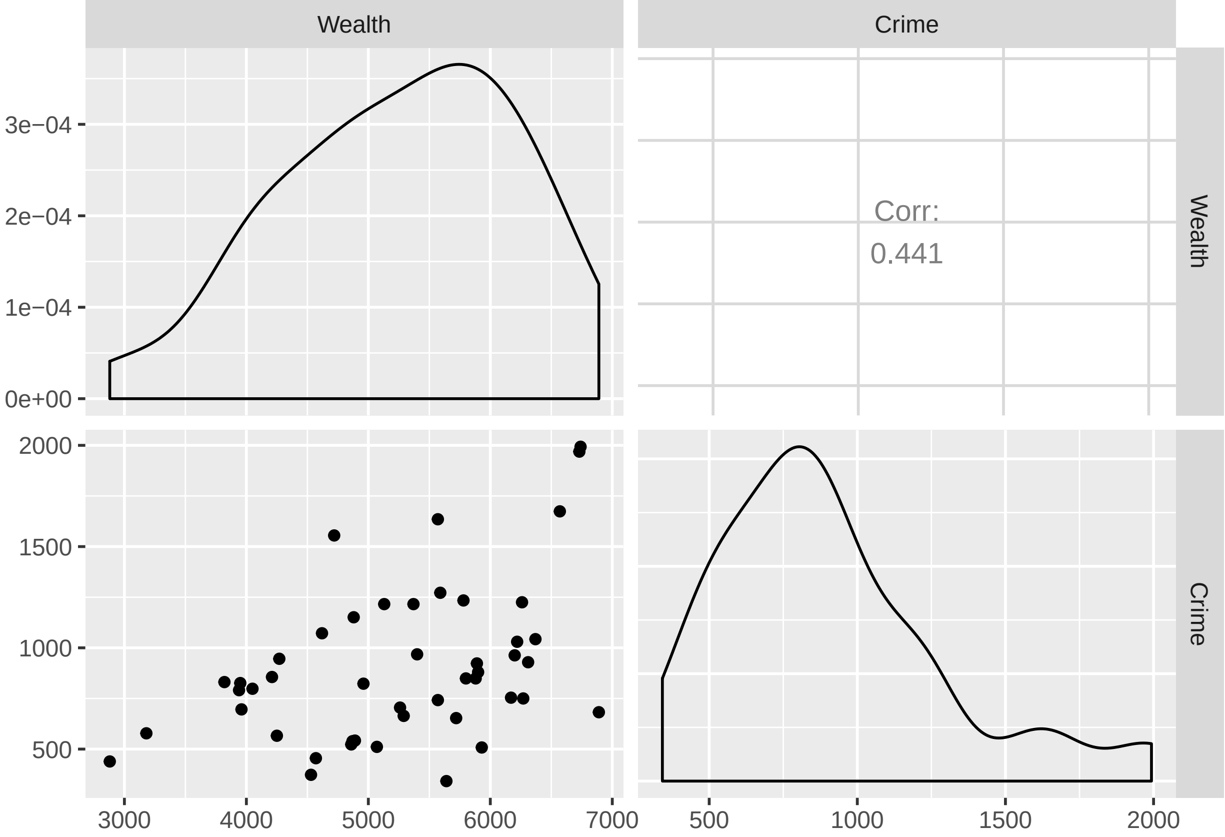

Pop
Crime
0.015
0.010
0.005
0.000
Corr:
0.337
Pop
2000
1500
1000
500
Crime
0
50
100
150
500
1000
1500
2000


NW
Crime
0.06
0.04
0.02
0.00
Corr:
0.0326
NW
2000
1500
1000
500
Crime
0
10
20
30
40
500
1000
1500
2000

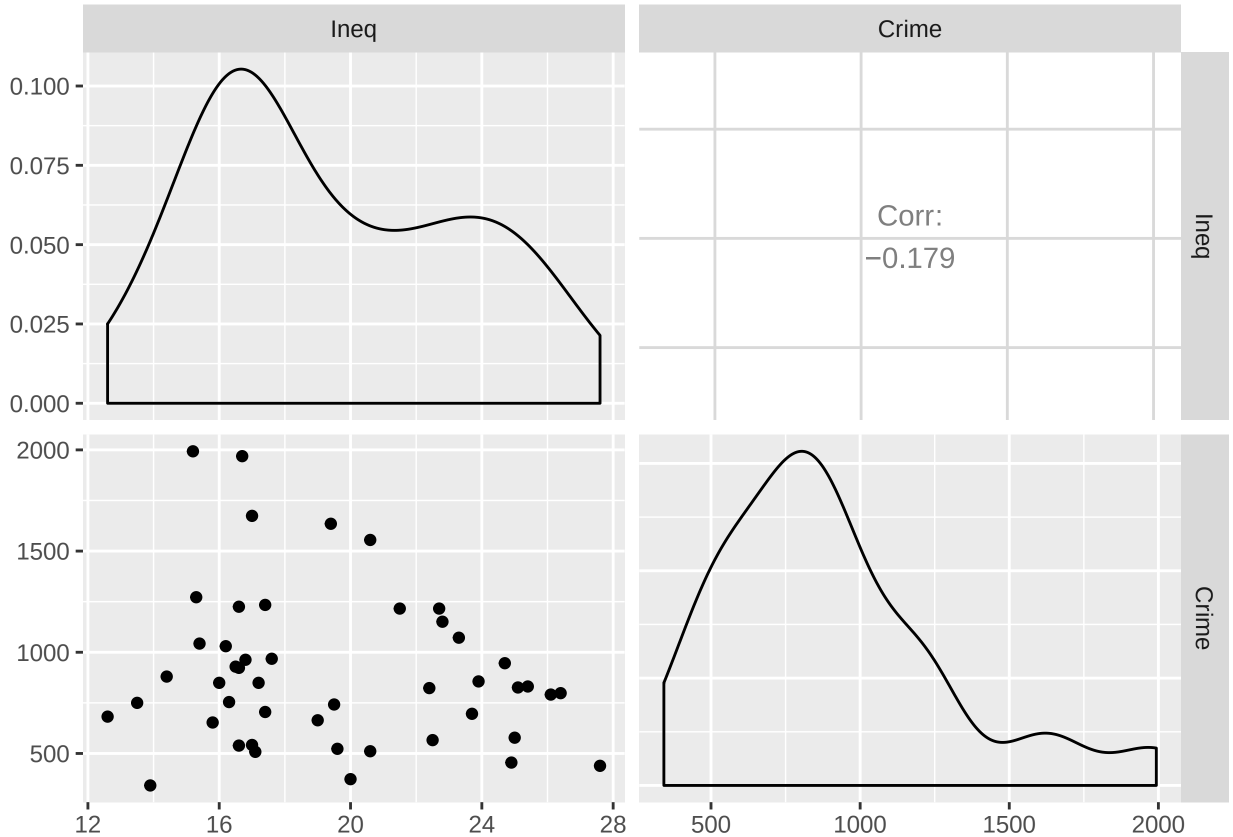

U1
Crime
20
15
10
5
0
Corr:
−0.0505
U1
2000
1500
1000
500
Crime
0.08
0.10
0.12
0.14
500
1000
1500
2000

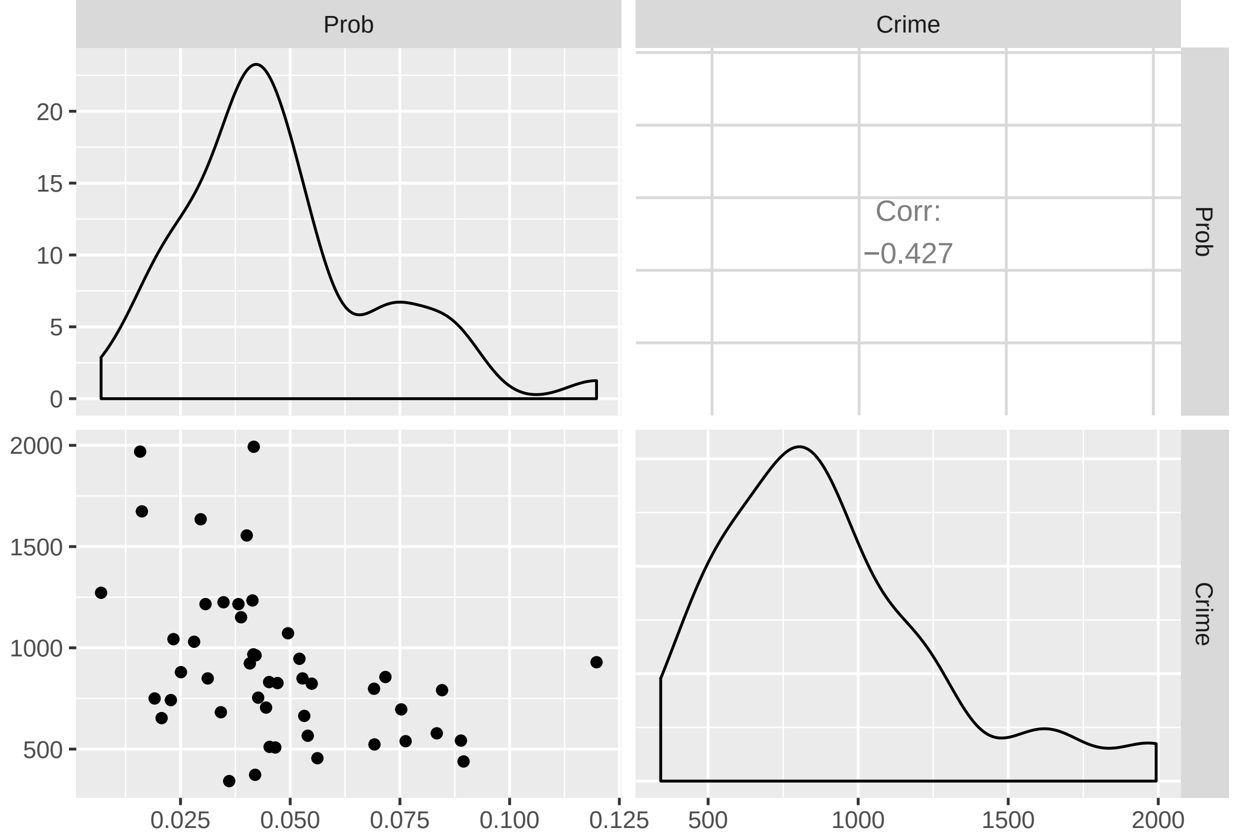


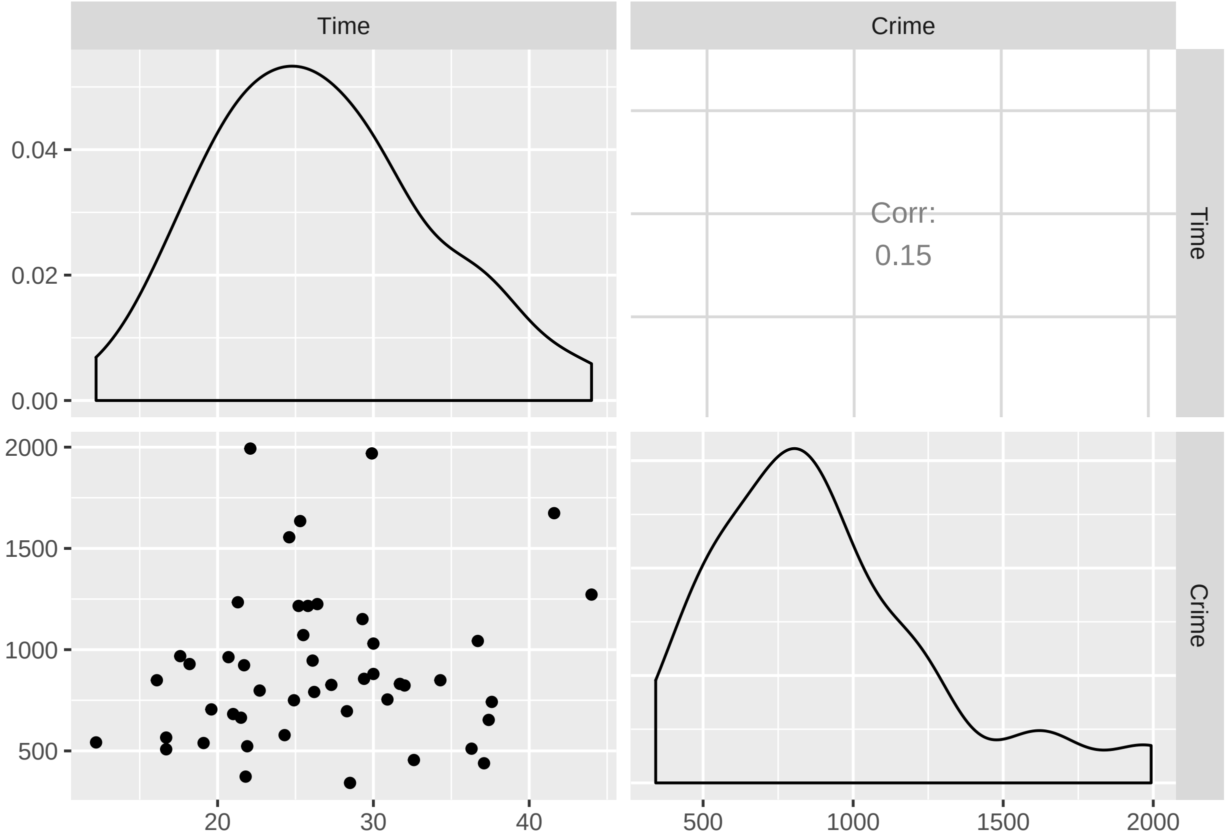
U2
Crime
0.4
0.3
0.2
0.1
0.0
Corr:
0.177
U2
2000
1500
1000
500
Crime
2
3
4
5
500
1000
1500
2000



Wealth
Crime
3e−04
2e−04
1e−04
0e+00
Corr:
0.441
Wealth
2000
1500
1000
500
Crime
3000
4000
5000
6000
7000
500
1000
1500
2000

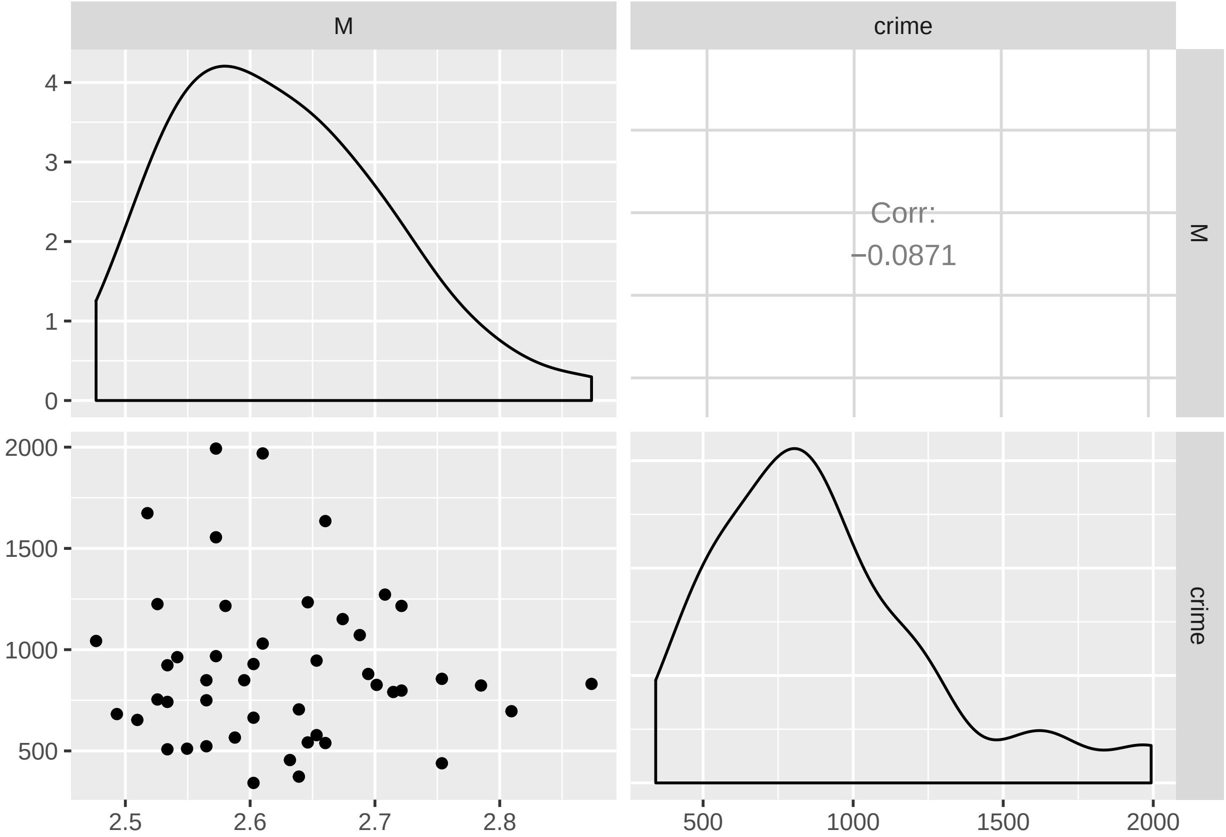

Ineq
Crime
Corr:
−0.179
Ineq
Crime
0.100
0.075
0.050
0.025
0.000
2000
1500
1000
500
12
16
20
24
28
500
1000
1500
2000


Prob
Crime
Corr:
−0.427
Prob
Crime
20
15
10
5
0
2000
1500
1000
500
0.025
0.050
0.075
0.100
0.125
500
1000
1500
2000

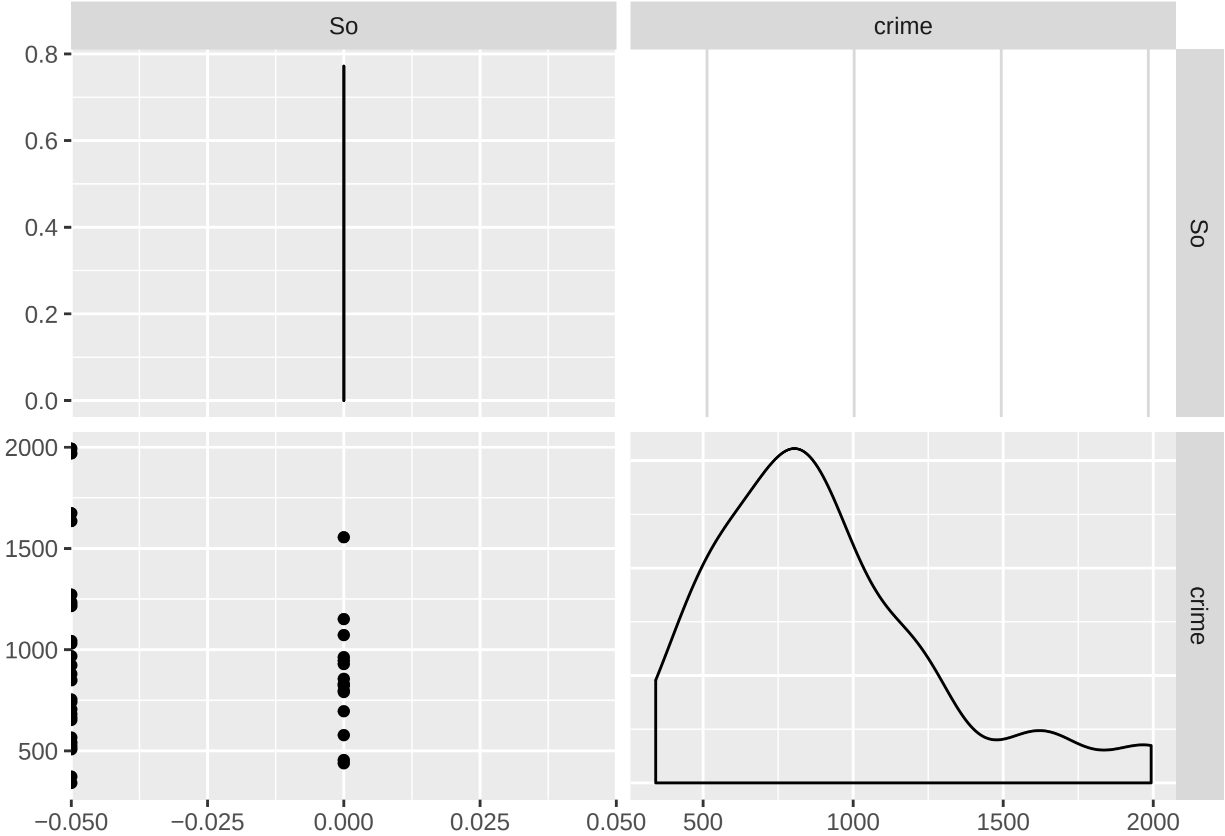


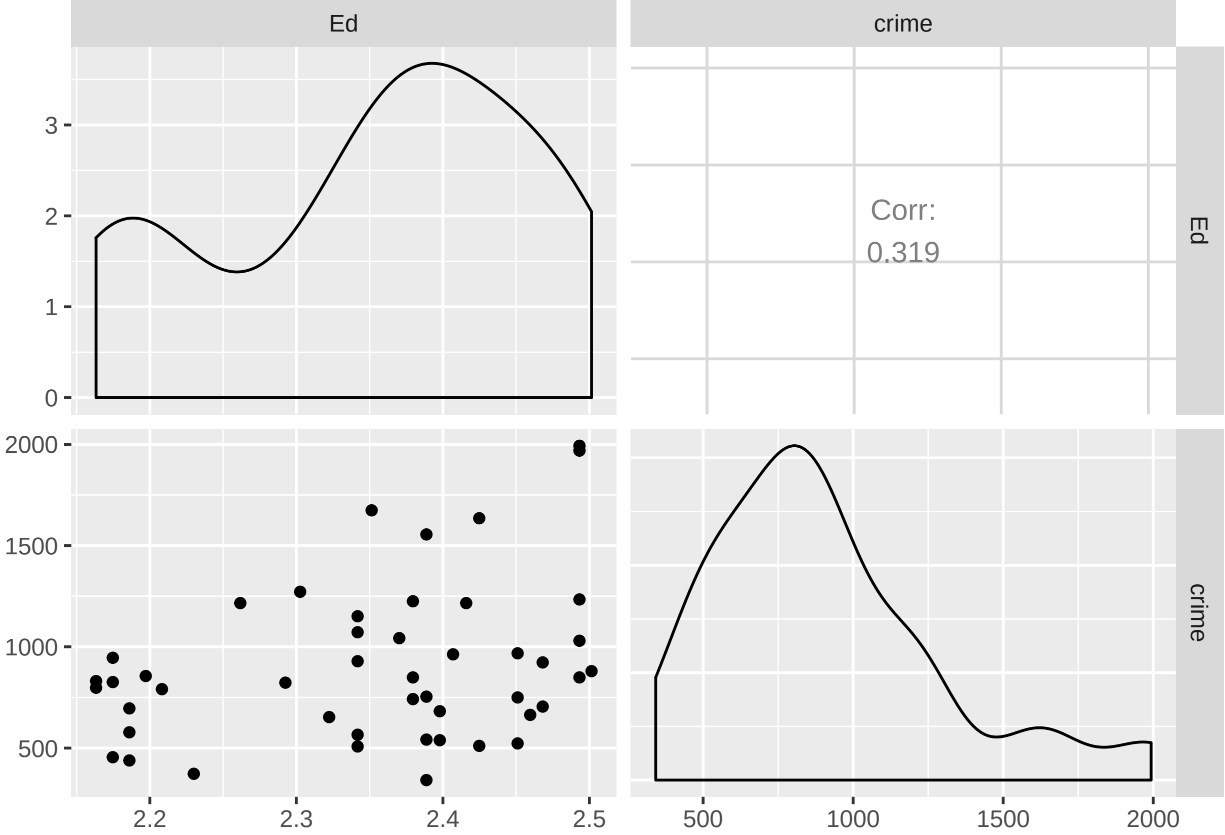
Surprisingly there's a high positive correlation between Po2 and Po1 with Crime. Po1 is per capital expenditure on police protection in 1959 and Po2 is representative of 1960. This is interpreted that there's a correlation that when police expenditure increases, so does crime. Another surpirsing note is that there is a relatively high negative correlation between prob and crime. This means that if there's a higher probability of going to jail, crime rate drops. Finally, there's also a decent correlation between wealth and crime. With increasing wealth comes increasing crime. Next, let's perform the same analysis, but on the log of the data to find any other hidden correlations.

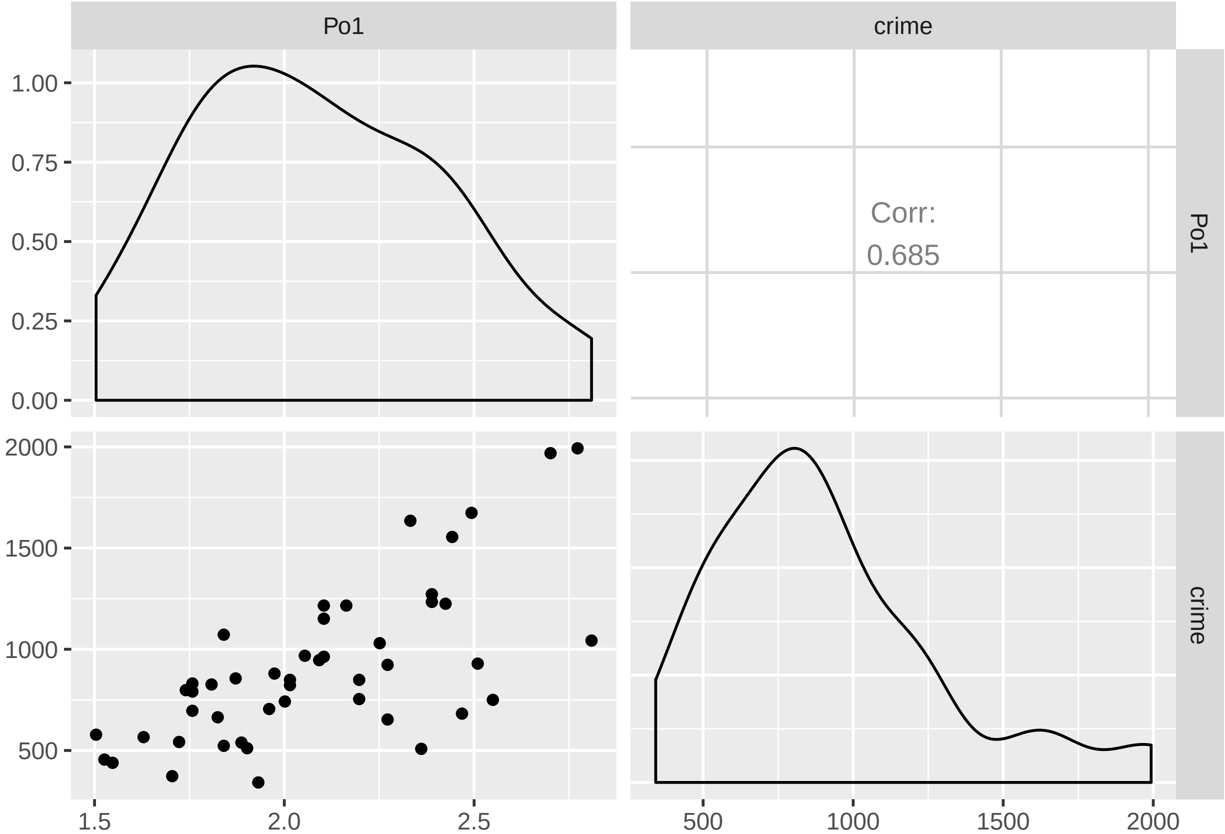
```
log_data = log(data[-16])
log_data['crime'] = data$Crime
for(i in seq(1,15)){
pm = ggpairs(data=log_data,
             columns= c(i,16),

             )
print(pm)
}
```

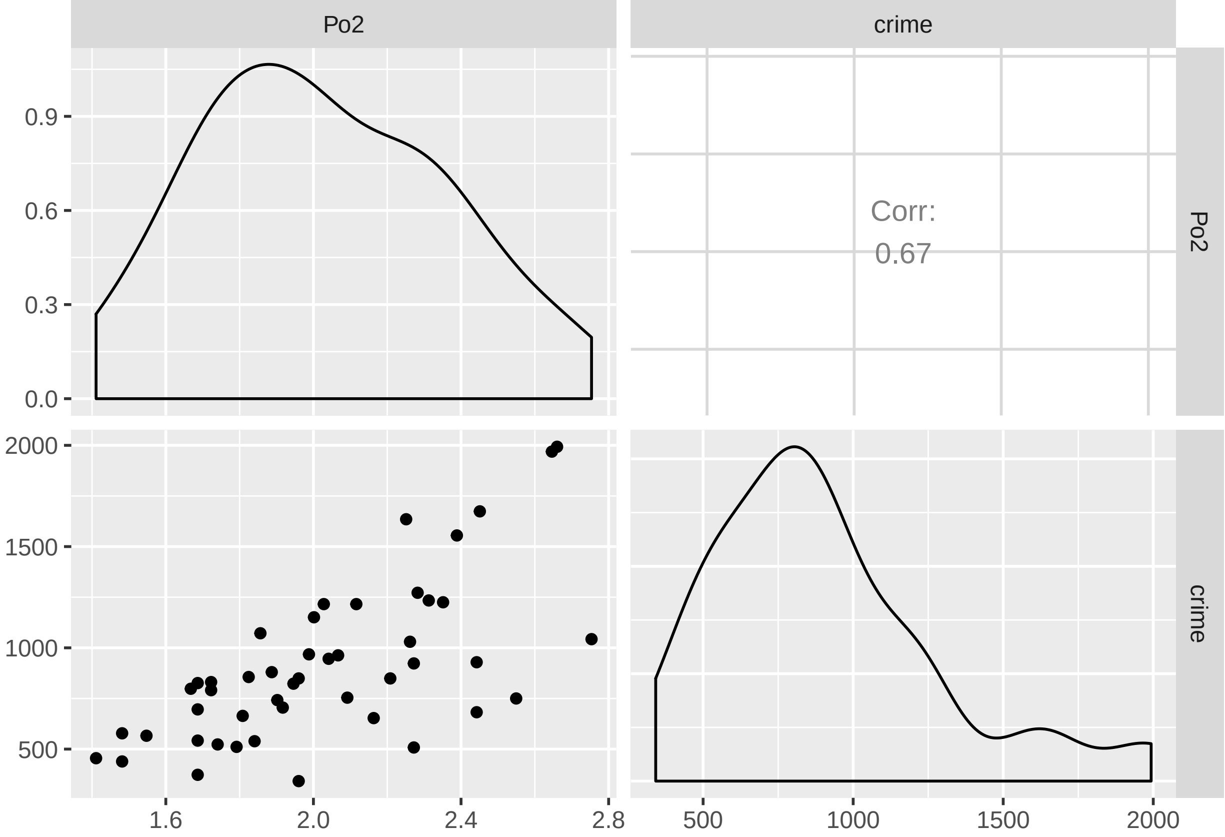


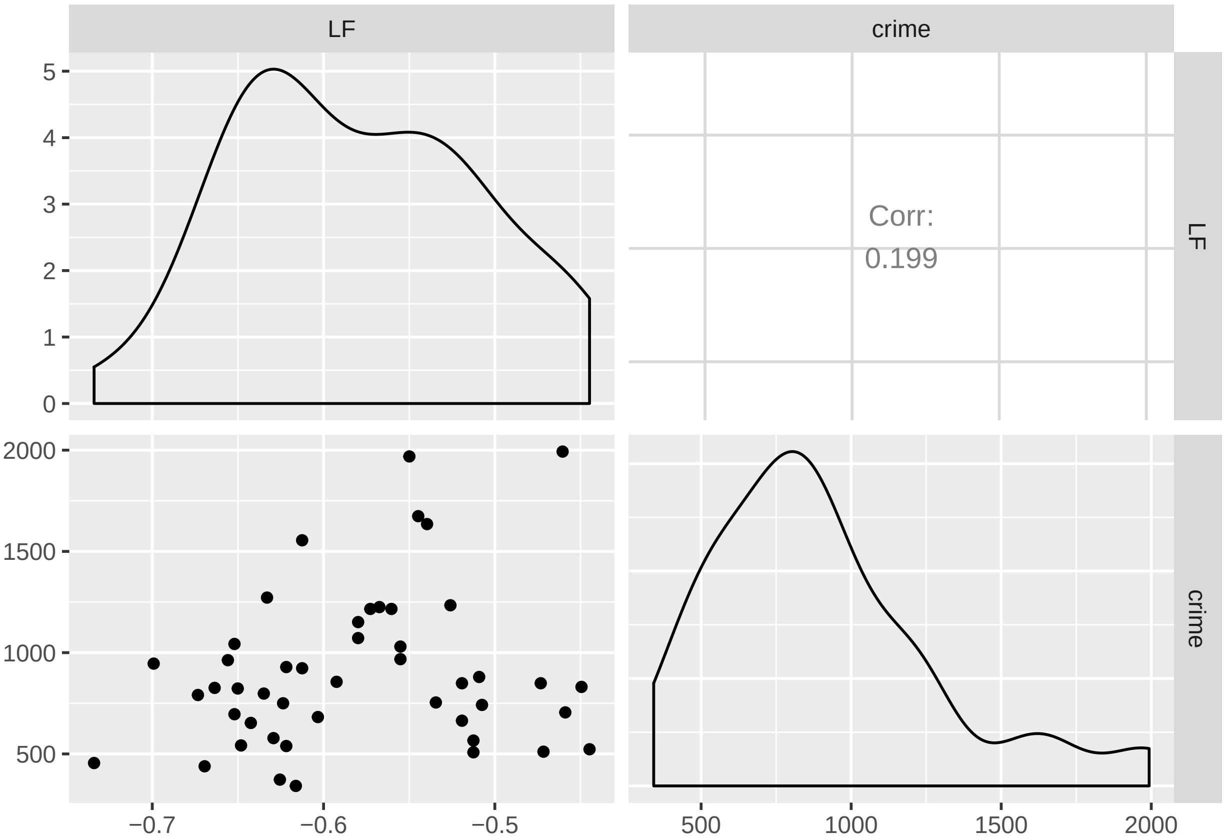
```
## Warning: Removed 31 rows containing non-finite values (stat_density).

## Warning: Removed 1 rows containing missing values (geom_text).
```

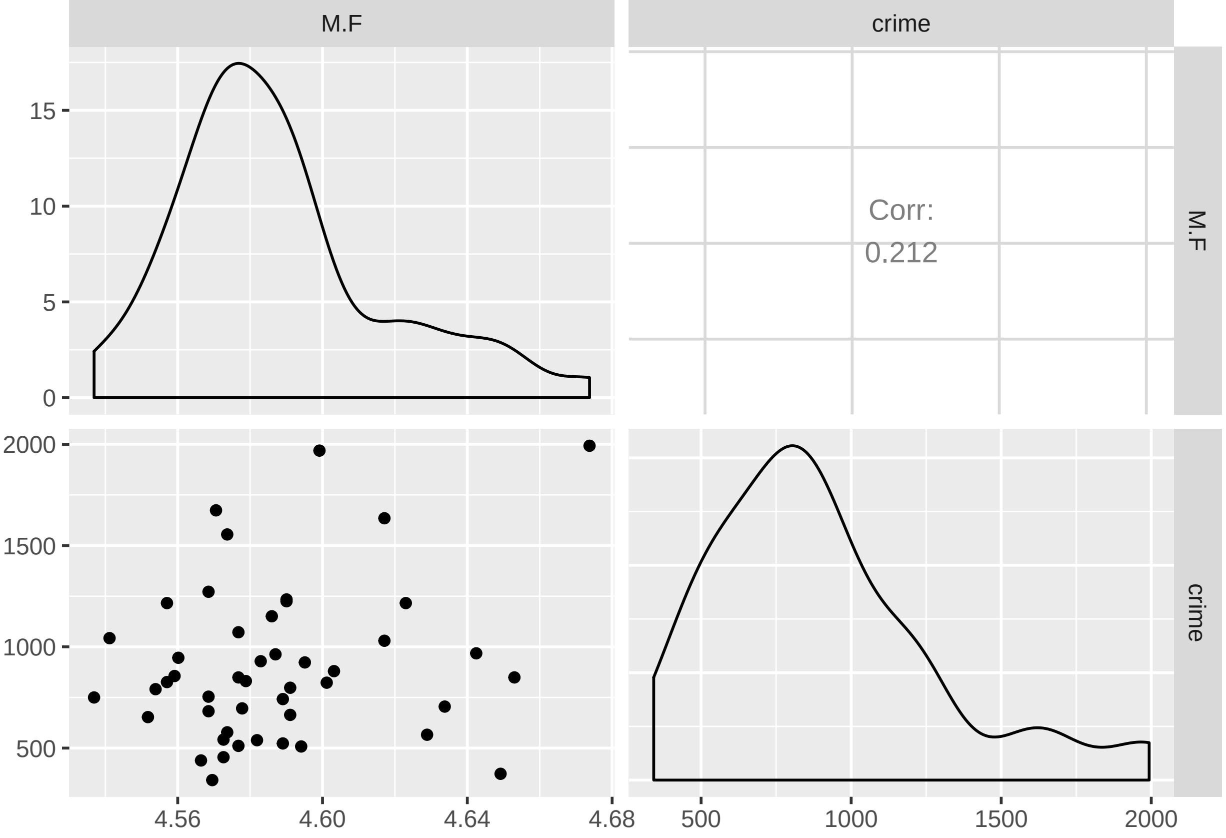

So
crime
0.8
0.6
0.4
0.2
0.0
So
2000
1500
1000
500
crime
−0.050
−0.025
0.000
0.025
0.050
500
1000
1500
2000

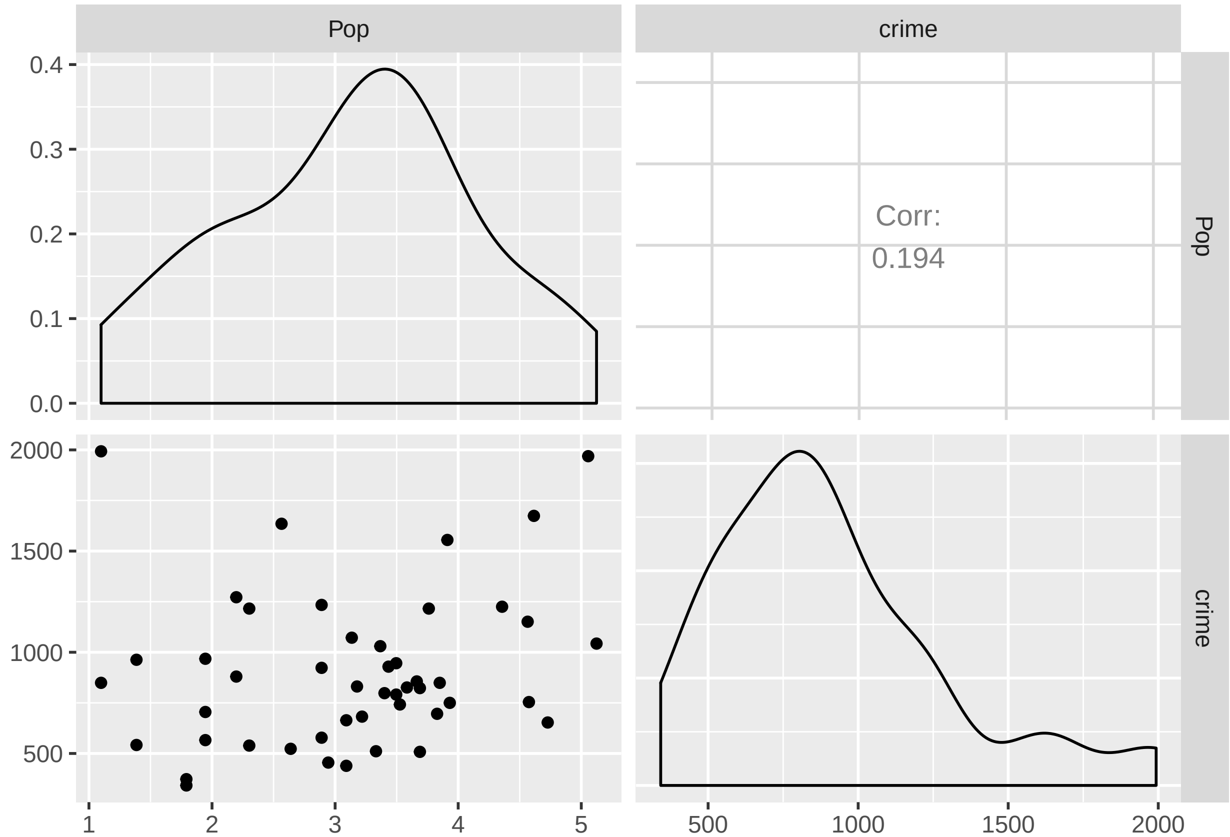

Ed
crime
3
2
1
0
Corr:
0.319
Ed
2000
1500
1000
500
crime
2.2
2.3
2.4
2.5
500
1000
1500
2000

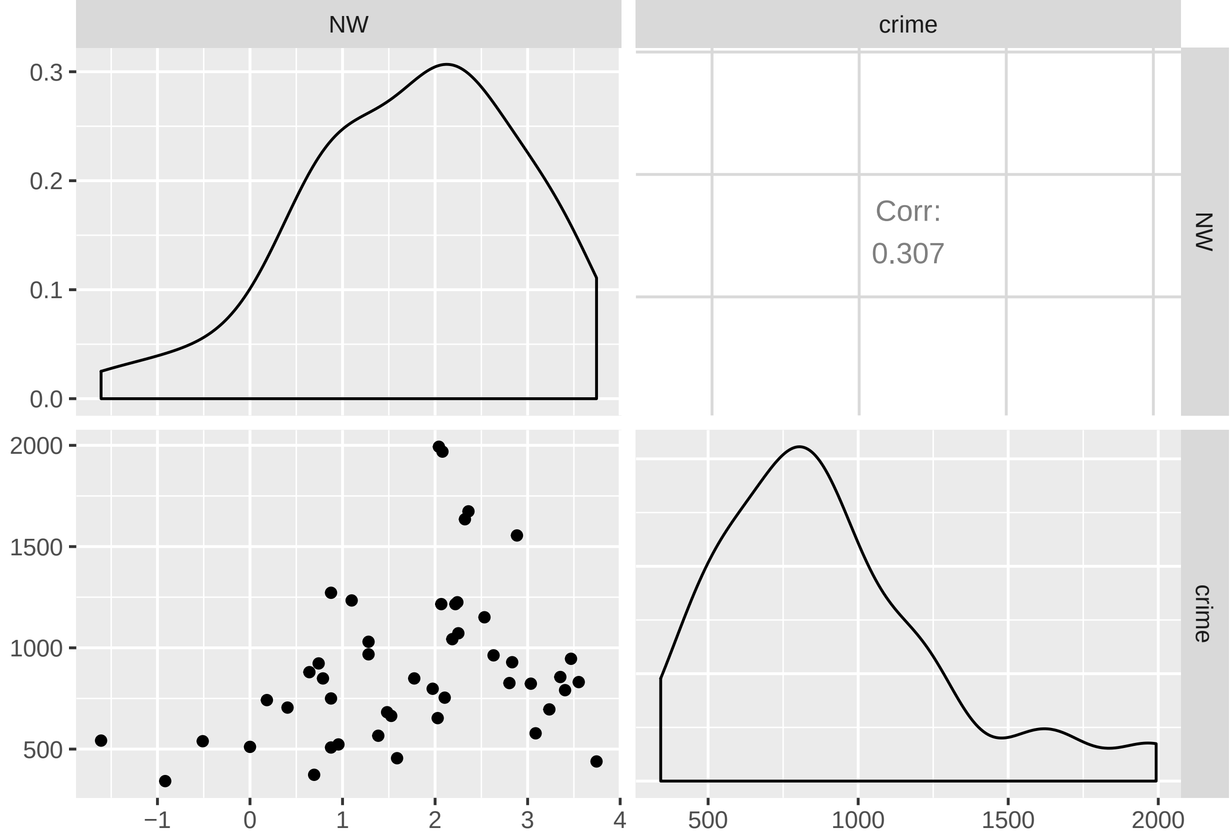

Po1
crime
1.00
0.75
0.50
0.25
0.00
Corr:
0.685
Po1
2000
1500
1000
500
crime
1.5
2.0
2.5
500
1000
1500
2000

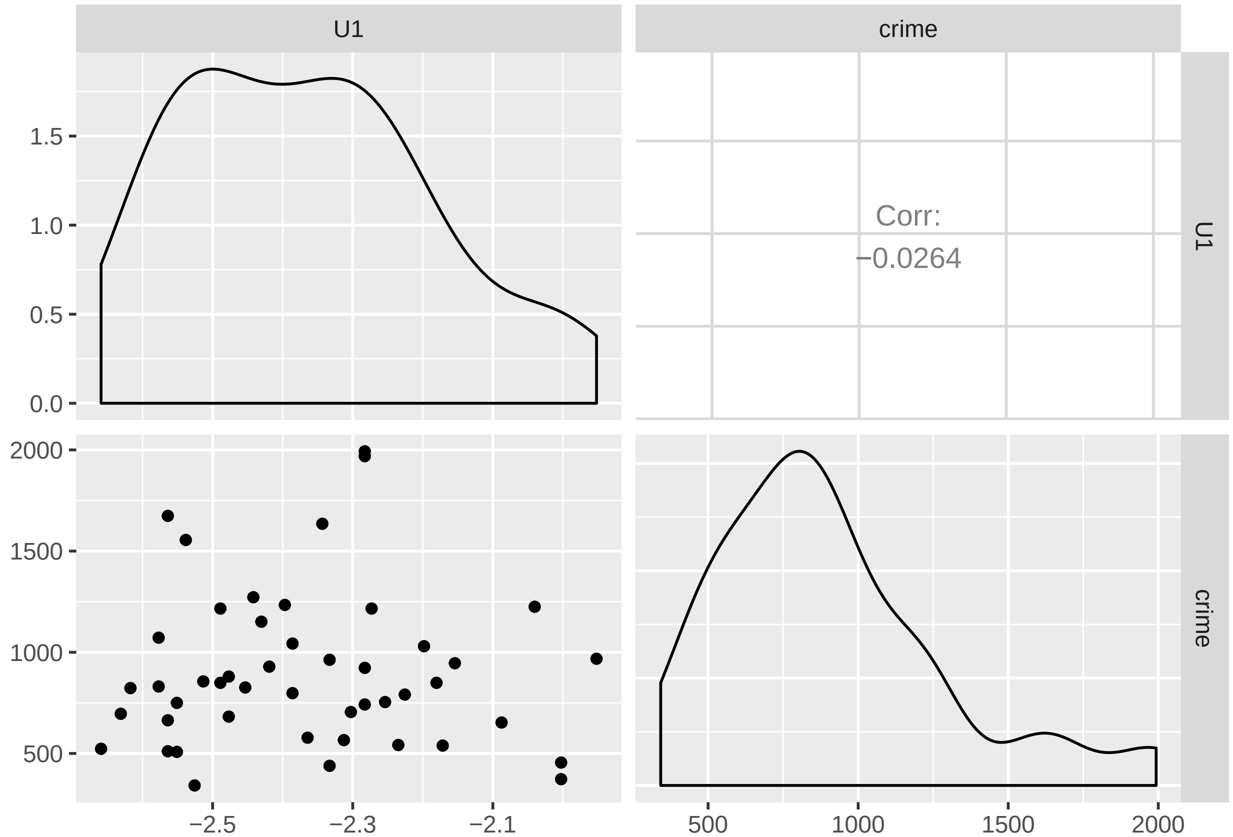

Po2
crime
0.9
0.6
0.3
0.0
Corr:
0.67
Po2
2000
1500
1000
500
crime
1.6
2.0
2.4
2.8
500
1000
1500
2000

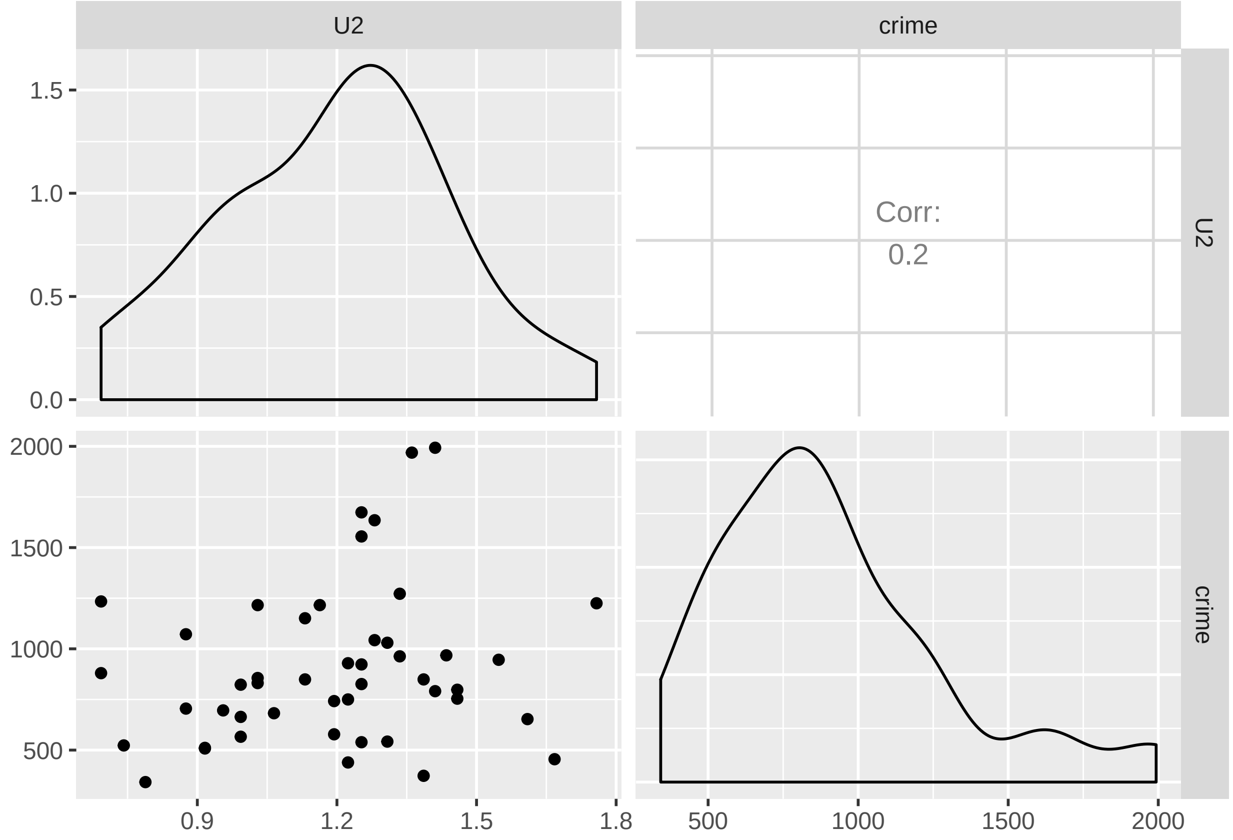

LF
crime
LF
crime
Corr:
0.199
5
4
3
2
1
0
2000
1500
1000
500
−0.7
−0.6
−0.5
500
1000
1500
2000

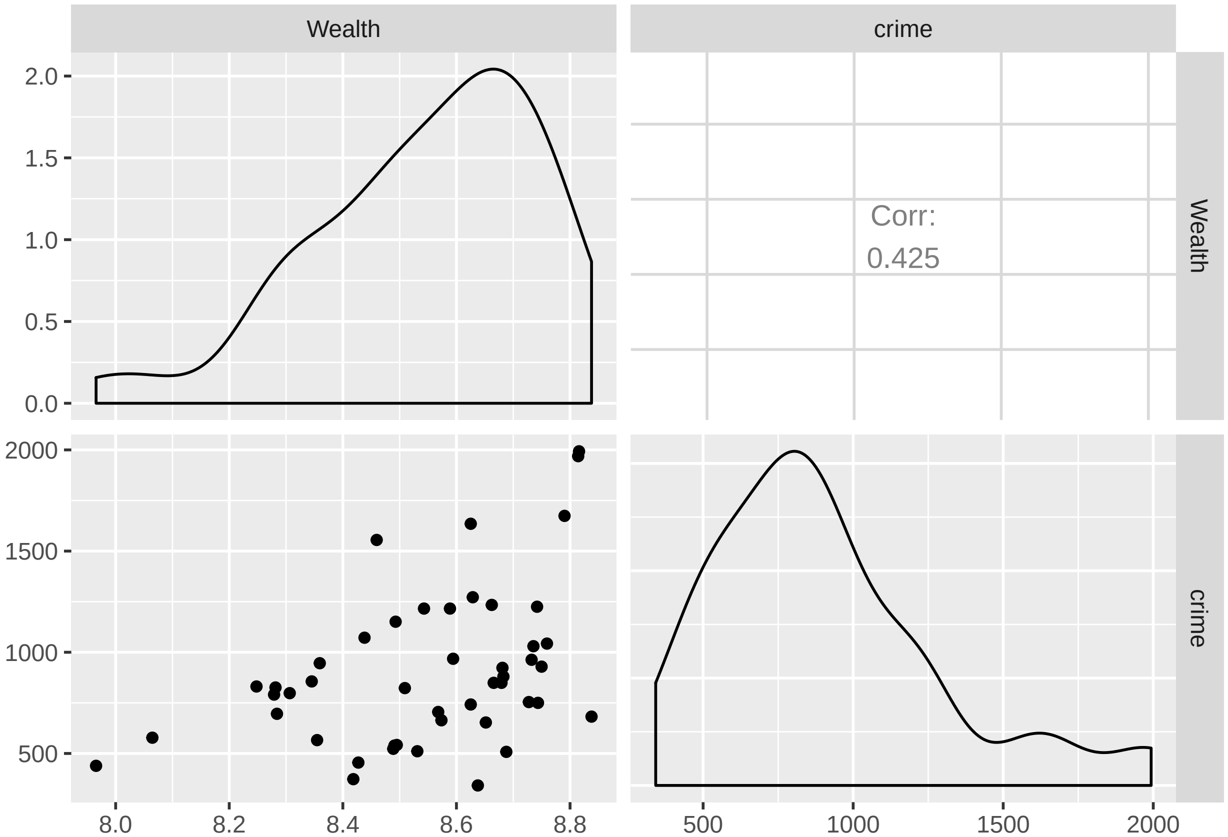

M.F
crime
M.F
crime
Corr:
0.212
15
10
5
0
2000
1500
1000
500
4.56
4.60
4.64
4.68
500
1000
1500
2000

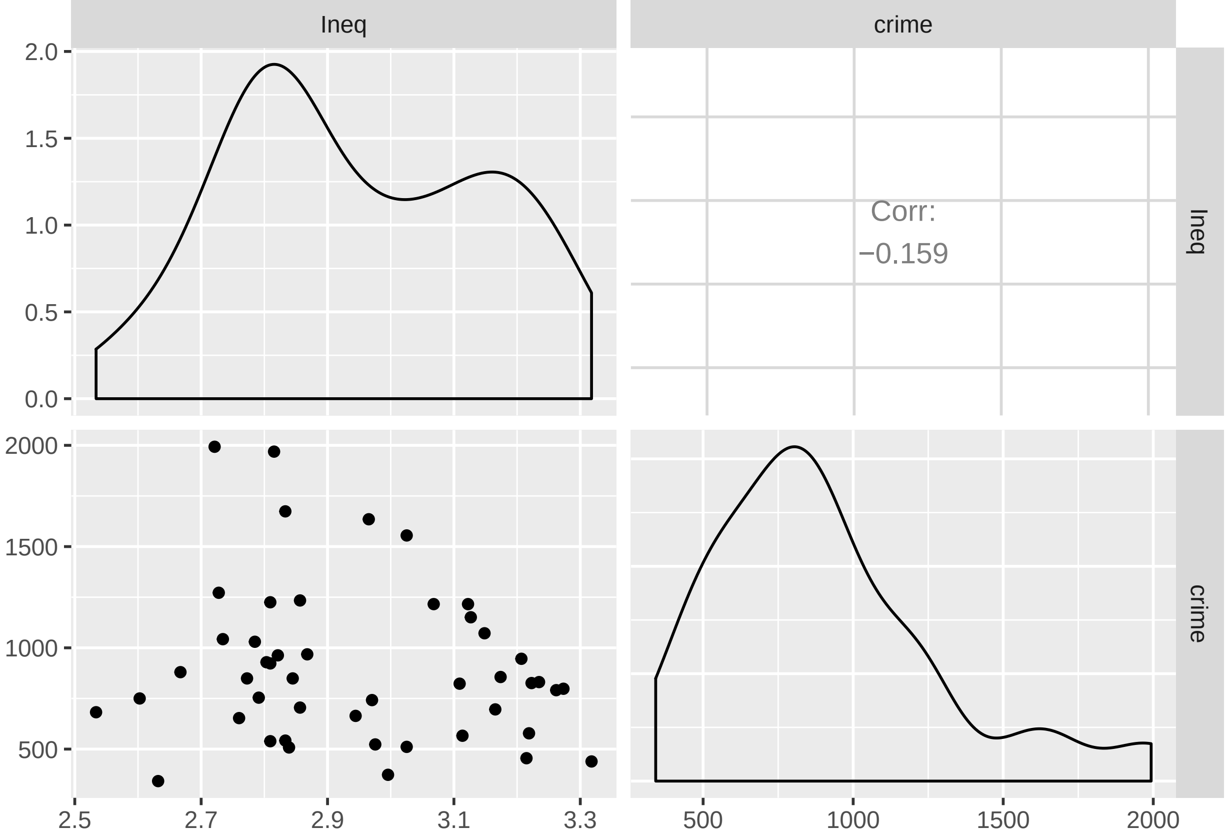

Pop
crime
Corr:
0.194
Pop
crime
0.4
0.3
0.2
0.1
0.0
2000
1500
1000
500
1
2
3
4
5
500
1000
1500
2000

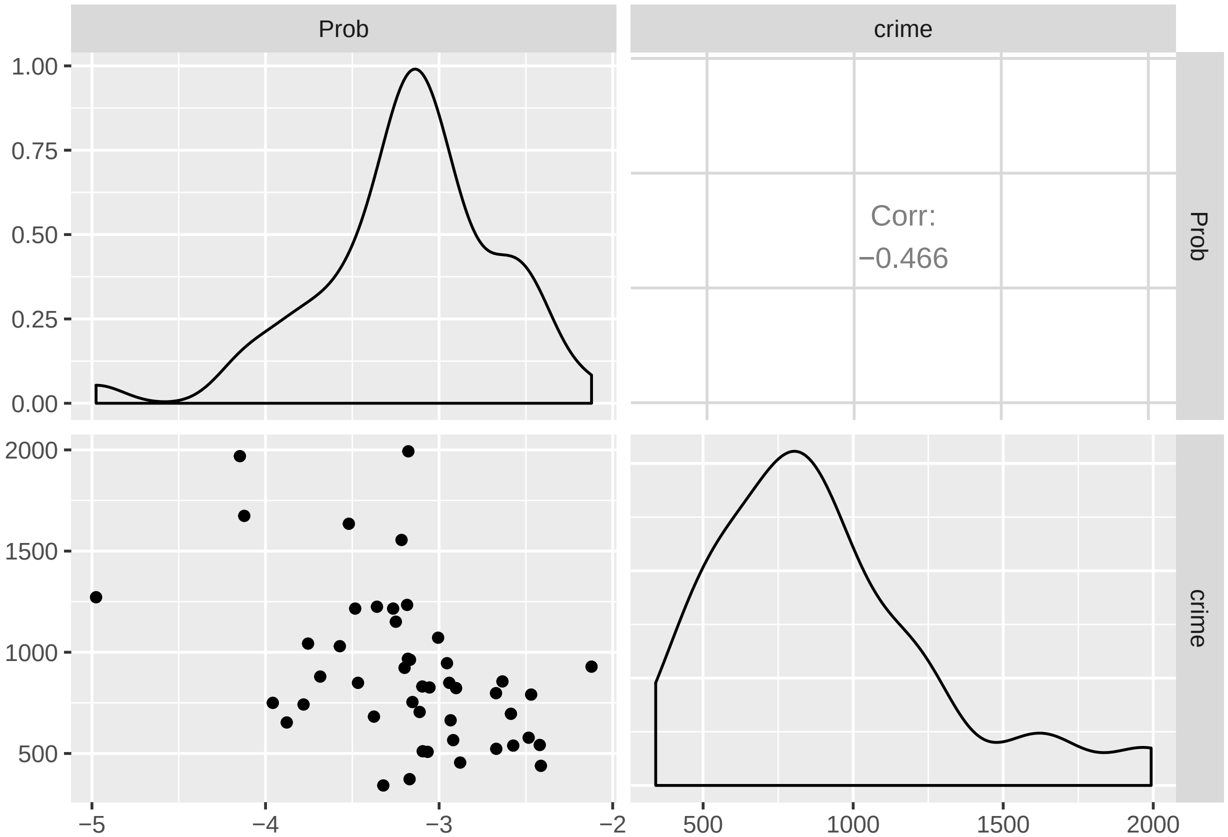

NW
crime
Corr:
0.307
NW
crime
0.3
0.2
0.1
0.0
2000
1500
1000
500
−1
0
1
2
3
4
500
1000
1500
2000

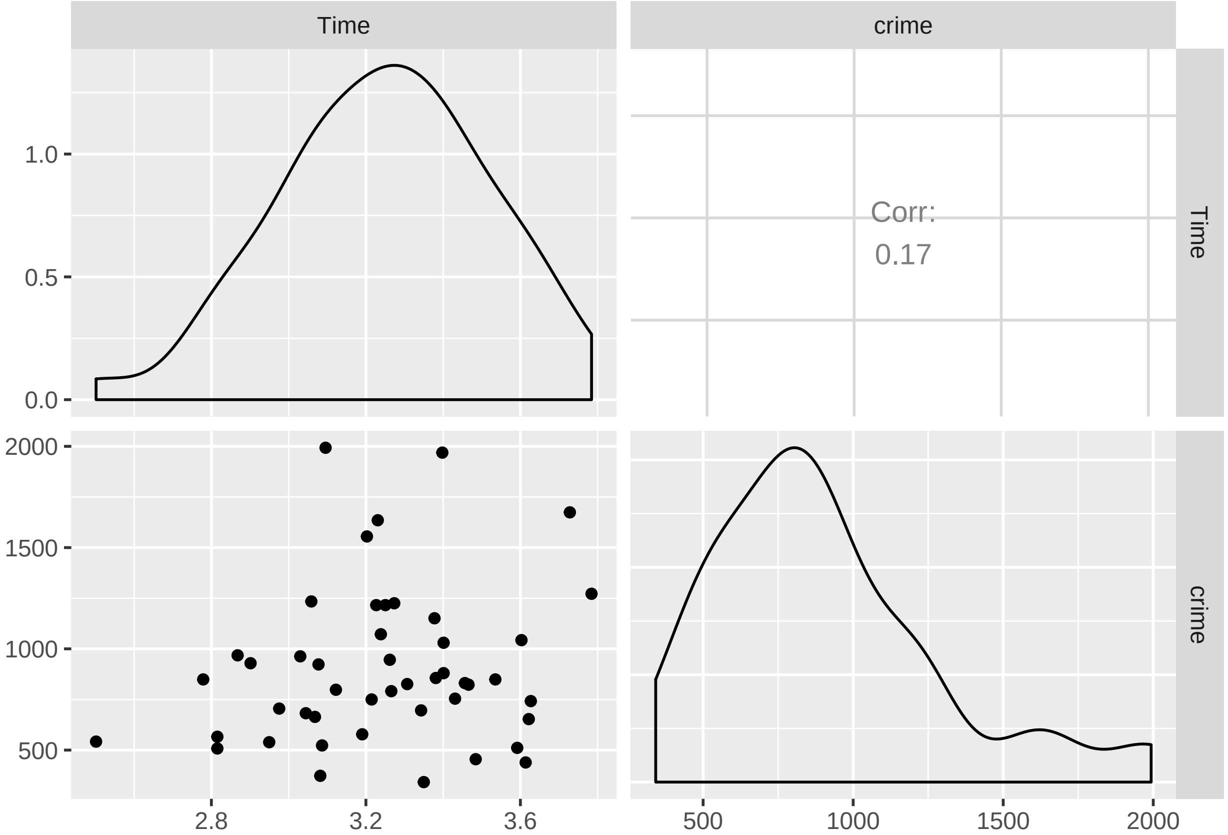

U1
crime
Corr:
−0.0264
U1
crime
1.5
1.0
0.5
0.0
2000
1500
1000
500
−2.5
−2.3
−2.1
500
1000
1500
2000


U2
crime
Corr:
0.2
U2
crime
1.5
1.0
0.5
0.0
2000
1500
1000
500
0.9
1.2
1.5
1.8
500
1000
1500
2000


Wealth
crime
Corr:
0.425
Wealth
crime
2.0
1.5
1.0
0.5
0.0
2000
1500
1000
500
8.0
8.2
8.4
8.6
8.8
500
1000
1500
2000


Ineq
crime
Corr:
−0.159
Ineq
crime
2.0
1.5
1.0
0.5
0.0
2000
1500
1000
500
2.5
2.7
2.9
3.1
3.3
500
1000
1500
2000





There doesn't appear to be any further highly correlated variables in the log transformed data. Let's build our model and then quantify how good it is. We will actually be building 2 models. One model utilizing all the independent variables at their normal scale, and another model with the independent values scaled and normalized

```
#model 1 - all independent variables
m1 <- lm(Crime ~ ., data = data)

scaled_data = as.data.frame(scale(data[1:15], center = TRUE, scale = TRUE))
scaled_data['Crime'] = data$Crime
#model 2 - scaled independent variables
m2 <- lm(Crime ~ ., data = scaled_data)
```

Let's quantify model 1 first

```
summary(m1)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

*Model 1 Analysis

The first thing to notice is that the most significant variables are Ed, Ineq, M, Prob, Po1, and U2. This is quantified by each of the variables Pr(>|t|) value and the significance code next to them. The next thing to note is the R-squared and the Adjusted R-squared of 0.8031 and 0.7078 respectively. The R squared value says our model explains 80.3% of the variance in the data, whereas the adjusted R-squared says our model explains 70.8% of the variance in the data. These are extremely good values and signals that we have a significant model or it could mean that we have an overfitted model. This is validated by the overall p-value -of 3.59e-07. This value is greatly below below a p-value of 0.05. Let's next analyze model2 with the scaled and normalzied independent variables.

```
summary(m2)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = scaled_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  905.085     30.495  29.680  < 2e-16 ***
## M            110.382     52.424   2.106  0.04344 *
## So            -1.822     71.250  -0.026  0.97977
## Ed           210.678     69.458   3.033  0.00486 **
## Po1          572.995    315.347   1.817  0.07889 .
## Po2         -305.958    328.483  -0.931  0.35883
## LF           -26.826     59.394  -0.452  0.65465
## M.F           51.293     59.977   0.855  0.39900
## Pop          -27.906     49.095  -0.568  0.57385
## NW            43.234     66.642   0.649  0.52128
## U1          -105.056     75.906  -1.384  0.17624
## U2           141.714     69.536   2.038  0.05016 .
## Wealth        92.792    100.028   0.928  0.36075
## Ineq         281.954     90.630   3.111  0.00398 **
## Prob        -110.394     51.667  -2.137  0.04063 *
## Time         -24.655     50.780  -0.486  0.63071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

*Model 2 Analysis

The statistics behind model 2 are almost as identical as model one. So we will just use model 1 to predict on the unseen data so as to save a step of scaling the data.

```
new_data = as.data.frame(t(c(14.0,0,10.0,12.0,15.5,0.640,94.0,150,1.1,0.120,3.6,3200,20.1,0.04,39.0)))
colnames(new_data) = colnames(data)[1:15]

prediction = predict(m1, new_data)
print(prediction)
```

```
##        1
## 155.4349
```

The Crime prediction on the new unseen data is 155