GR5291 Final Project Report

Chen Chen (cc4291), Yang Meng (ym2696), and Zhichao Hu (zh2351)

May 4, 2019

Part I: Summary

The goal of our study is to find an appropriate model to predict the factors that reduce the number of serious crimes in a county. In other words, we hope to find the factors that might have impacts on the crime rates. Furthermore, we want to know what kinds of approaches we could take in order to reduce the crime rate in a county.

The results of our study show that the possible factors include population density, population, the average number of hospital beds, percentage of people who completed 12 or more years of school, and percentage of the population with income below poverty level. All these five factors have positive effects on the crime rate, which we need to decrease their effects to reduce the crime rate. Moreover, crime rate in the South is higher than in other regions of the United States. We also find that the population has an even higher positive effect in the South. The reduction of crime rate can be achieved by population control and eliminating poverty.

Part II: Introduction

Serious crime is always a complex issue faced by every government. The results of this analysis are important as people will have a better understanding of the possible factors that might influence the number of serious crimes. Without such researches on crime data, governors would hardly find solutions to crimes.

For the purpose of our research, we use crime rate (the number of serious crimes divided by the population) instead of crime as our response variable. Due to the different scales of variables, we divide the population by 10,000. Besides 21 predictors, we have created 4 new variables by dividing the area, the number of physicians, and the number of hospital beds by population in 10,000. Then we divided the population by area to eliminate the influence of the size of that county. Moreover, we use income per capita instead of total income as total income is strongly correlated with the population.

1. Data description

Crime dataset provides us with demographic information for 440 of the most populous counties in the United States and we sampled 300 rows as our training data. Appendix contains the detailed description of variables.

- **Significant Level**

All the analysis and results in our project are based on the significant level at $alpha = 0.05$.

- **Response**

Log transformation can make the response variable more symmetric, which is good for visualization. However, the data is still not normally distributed after the log transformation. As such, we do not need to apply it to Crime Rate in the future regression part.

- **Predictors**

In order to have a clear view of our predictors, we classified 21 variables into 6 different groups based on their similarity and potential correlations.

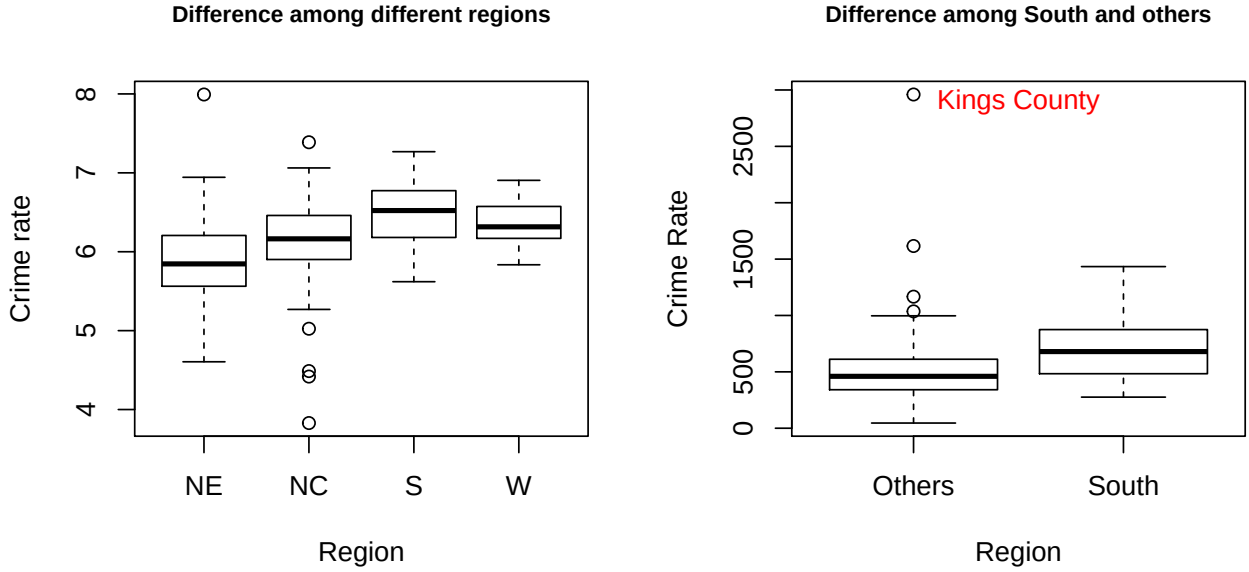
Group	Category	Predictor Names	Abbreviation
1	Region	South(Dummary)	south
2	Population	Total Population,Population Density	pop,popdens
3	Age (Percent)	Young(18-34)	young
4	Medical Level (Per person)	Hospital Beds	avgbeds
5	Education Level	High School Graduates	highgrad
6	Status	Poor,Unemployment	poor,unemp

For each group, we have explored their relationships through EDA (See Details in Appendix: Predictors' selection), and then we chose the most significant ones into our model building. After these careful selections, the table above shows the predictors that we will use later.

2. Exploratory data analysis:

- **Divide Region**

From the `left` boxplot of crime rates in different regions, we could see that crime rates in South are generally higher than in other regions. Thus, we collapse "region" into one dummy variable "South" in order to discover the effect of "South" on crime rate and the interaction with other variables.



- **Outliers**

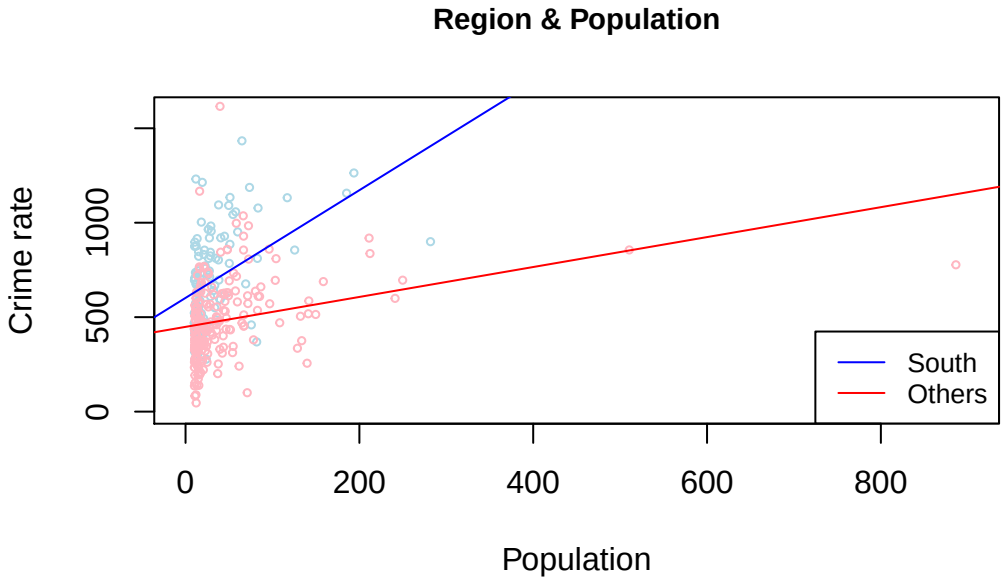
From the `Right` boxplot we found that "Kings County" in New York state is an outlier that deviates from the rest of data by a large degree. It approximately has a crime rate of 2960 per 10,000 people. We believe that this outlier could undermine our model. Hence, we deleted this outlier from the dataset.

- **Test for Multicollinearity**

We utilize the variance inflation factor (VIF) to justify multicollinearity. Smaller VIF and R_i^2 indicates less chance of multicollinearity.

After the selection we did before, we deleted all the variables that might have collinearity, and the new VIF and R_i^2 values indicate that we are less likely to have multicollinearity in our model, which also supports our classification before (See Details in Appendix: Multicollinearity).

- **Interaction**



In the graph above, we plotted crime rate versus population. Additionally we regress crime rate on population in South and in other regions to see if the population effect in different regions is the same. Based on the graph, we conclude that although population has positive effect on crime rate in both South and not South regions, the population effect in South seems to be significantly larger than in other regions. Therefore, we include this interaction term in our regression model.

Part III: Analysis

1. Model Selection

- **Initial Model**

$$\begin{aligned} crimerate = &-551.336 + 0.026\ popdens + 0.528\ pop + 3.054\ young + 2.637\ avgbeds + 8.106\ highgrad \\ &+ 22.428\ poor - 0.116\ unemp + 123.077\ south + 1.722\ pop \times south,\ \ south = \begin{cases} 0, & south \\ 1, & others \end{cases} \end{aligned}$$

The table in the following shows the estimates of the **initial** model, the corresponding p-values, and the 95% confidence intervals.

Variable	Coefficient	P-value	95% Confidence Interval
Intercept	-551.336	0.016	[-1001.161 ,-101.510]
popdens	0.026	0.000	[0.012 , 0.039]
pop	0.528	0.002	[0.202 , 0.853]
young	3.054	0.271	[-2.400 , 8.508]
avgbeds	2.637	0.000	[1.448 , 3.826]
highgrad	8.106	0.001	[3.171 , 13.042]
poor	22.428	0.000	[15.612 , 29.244]
unemp	-0.116	0.986	[-12.853 , 12.622]
south	123.077	0.000	[63.334, 182.820]
pop:South	1.722	0.000	[0.797 , 2.647]

The coefficients of variable *young* and *unemp* are not significant differ from zero at significant level $\alpha = 0.05$, with p-values 0.271 and 0.986, respectively, both larger than 0.05. These p-values give us sufficient reasons to remove *young* and *unemp* in the following step.

- **Final Model**

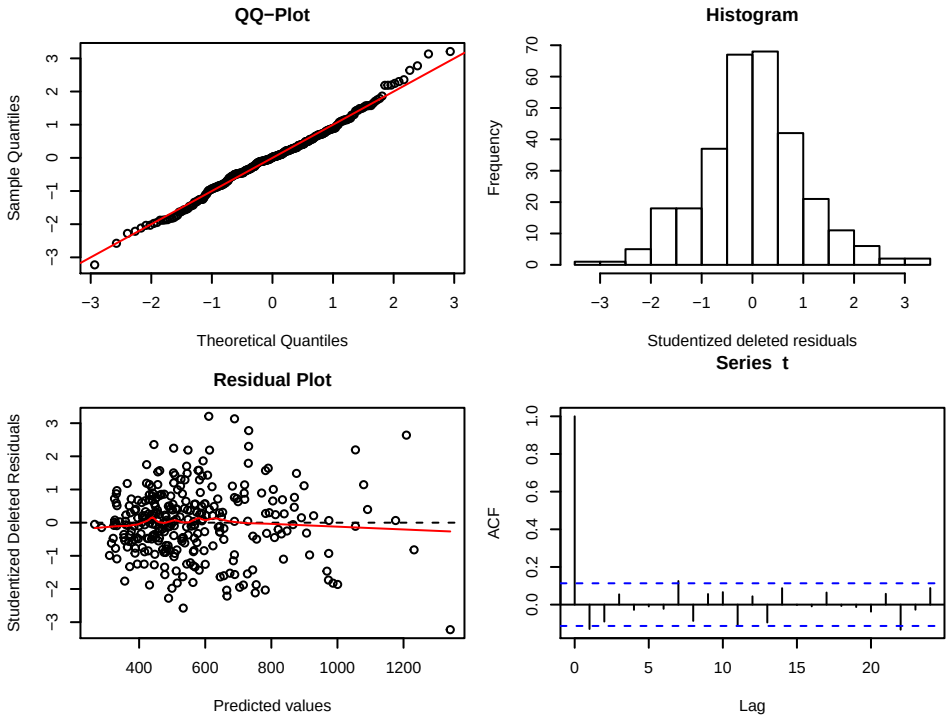
$$\begin{aligned} crimerate = &- 558.511 + 0.027\ popdens + 0.537\ pop + 2.596\ avgbeds + 9.179\ highgrad \\ &+ 23.484\ poor + 128.079\ south + 1.704\ pop \times south,\ \ south = \begin{cases} 0,\ south \\ 1,\ others \end{cases} \end{aligned}$$

The table in the following shows the estimates of the **Final** model, the corresponding p-values, and the 95% confidence intervals.

Variable	Coefficient	P-value	95% Confidence Interval
Intercept	-558.511	0.002	[-908.592, -208.430]
popdens	0.027	0.000	[0.014 , 0.040]
pop	0.537	0.001	[0.213 , 0.861]
avgbeds	2.596	0.000	[1.453 , 3.740]
highgrad	9.179	0.000	[5.151 , 13.206]
poor	23.484	0.000	[17.076 , 29.891]
South	128.079	0.000	[72.398 , 183.759]
pop:South	1.704	0.000	[0.784 , 2.625]

All the p-values are smaller than the significant level α=0.05, so we are 95% confident that these predictors have significant relationship with Crime rate, and we are 95% confident that the regression coefficients are within the corresponding confidence intervals.

2. Residual Analysis



All 3 important assumptions of residuals are met, so our final model is somewhat plausible and reliable.

- Normal Assumption:

QQ plot and histogram of the studentized deleted residuals shows that the studentized deleted residuals are normally distributed, with just slightly heavy tail which is hard to avoid in practice.

- Equal Variance Assumption:

Residual plot shows the studentized deleted residuals share equal variance, as the residuals seem like randomly distributed. The unclear pattern shows randomness and homoscedasticity in residuals.

- Independent Assumption:

Auto-correlation function plot shows the independency in studentized deleted residuals, because the function values with lag larger than 1 are all within or just on the blue bands.

Part IV: Results

1. Model

$$\begin{aligned} crimerate = &- 558.511 + 0.027\ popdens + 0.537\ pop + 2.596\ avgbeds + 9.179\ highgrad \\ &+ 23.484\ poor + 128.079\ south + 1.704\ pop \times south,\ south = \begin{cases} 0,\ south \\ 1,\ others \end{cases} \end{aligned}$$

2. Interpretation

- **Population density (*popdens*):**

Increasing *popdens* (population density) by 1 unit (1 more person per square miles), the crime rate will increase by 0.027 units (0.027 crimes per 10,000 people), while holding all the other variables fixed.

For the interval estimation, we are 95% confident that increasing *popdens* by 1 unit, the crime rate will increase by a value in the interval [0.014, 0.040], while holding all the other variables fixed.

- **Population (*pop*) and its interaction with region (*pop* × *south*):**

- In the South region, increasing *pop* (population) by 1 unit (10,000 more people in the county), the crime rate will increase by 2.241 units (2.241 crimes per 10,000 people), while holding all the other variables fixed.

For the interval estimation, we are 95% confident that, in the South region, increasing *popd* by 1 unit, the crime rate will increase by a value in the interval [0.997, 3.486], while holding all the other variables fixed.

- In the other region, increasing *pop* (population) by 1 unit (10,000 more people in the county), the crime rate will increase by 0.537 units (0.537 crimes per 10,000 people), while holding all the other variables fixed.

For the interval estimation, we are 95% confident that, in the other region, increasing *popd* by 1 unit, the crime rate will increase by a value in the interval [0.213, 0.861], while holding all the other variables fixed.

- **Hospital beds per 10,000 people (*avgbeds*):**

Increasing *avgbeds* (the number of hospital beds per 10,000 people) by 1 unit, the crime rate will increase by 2.596 units (2.596 crimes per 10,000 people), while holding all the other variables fixed.

For the interval estimation, we are 95% confident that increasing *avgbeds* by 1 unit, the crime rate will increase by a value in the interval [1.453, 3.740], while holding all the other variables fixed.

- **Percentage of high school graduates (*highgrad*):**

Increasing *highgrad* (percentage of high school graduates) by 1 unit, the crime rate will increase by 9.179 units (9.179 crimes per 10,000 people), while holding all the other variables fixed.

For the interval estimation, we are 95% confident that increasing *highgrad* by 1 unit, the crime rate will increase by a value in the interval [5.151, 13.206], while holding all the other variables fixed.

- **Percentage of people below the poverty level (*poor*):**

Increasing *poor* (percentage of people below the poverty level) by 1 unit, the crime rate will increase by 23.484 units (23.484 crimes per 10,000 people), while holding all the other variables fixed.

For the interval estimation, we are 95% confident that increasing *poor* by 1 unit, the crime rate will increase by a value in the interval [17.076, 29.891], while holding all the other variables fixed.

- **Region (*south*) and its interaction with population (*pop* $\times$ *south*):**

- The South region has $(128.079 + 1.704\ pop)$ more crime rate compared to the other region, where *pop* is the total population in the county, while holding all the other variables fixed.

For the interval estimation, we are 95% confident that the South region has a value in the interval [72.398+0.784*pop*, 183.759+2.625*pop*] more crime rate compared to the other region, while holding all the other variables fixed.

- What is more, increasing *pop* (population) by 1 unit (10,000 more people in the county), the difference of crime rate between the South and the other region will increase by 1.704 units, while holding all the other variables fixed.

For the interval estimation, we are 95% confident that increasing *pop* by 1 unit, the difference of crime rate between the South and the other region will increase by a value in the interval [0.784, 2.625], while holding all the other variables fixed.

Part V: Conclusion

As a result of our analysis, we arrive at the final regression model. We conclude that all the variables in our model, population density, population, the average number of hospital beds per 10,000 people, percentage of high school graduates, and percentage of poor, have significantly positive effects on the crime rate.

Moreover, counties in the South tend to have higher crime rates, and the population has a significantly higher positive effect in the South.

Theoretically, reducing any one of the variables can decrease the crime rate. In practice, to reduce the crime rate within a county, we suggest more practical ways:

- Control fertility rate to avoid overpopulation, especially in the South region.

Reducing the population by controlling fertility rate can reduce the crime rate according to the model. In the South, the crime rate decreases about 4 times as fast as in the other regions, so the policy may be more effective in the South region.

- Eliminate poverty by creating jobs and cutting taxes.

Eliminating poverty in a proper way is always plausible. The related policies will increase the index of the well-being of citizens, and decrease the crime rate. The positive relationship between wealth and the low crime rate is also shown in the model.

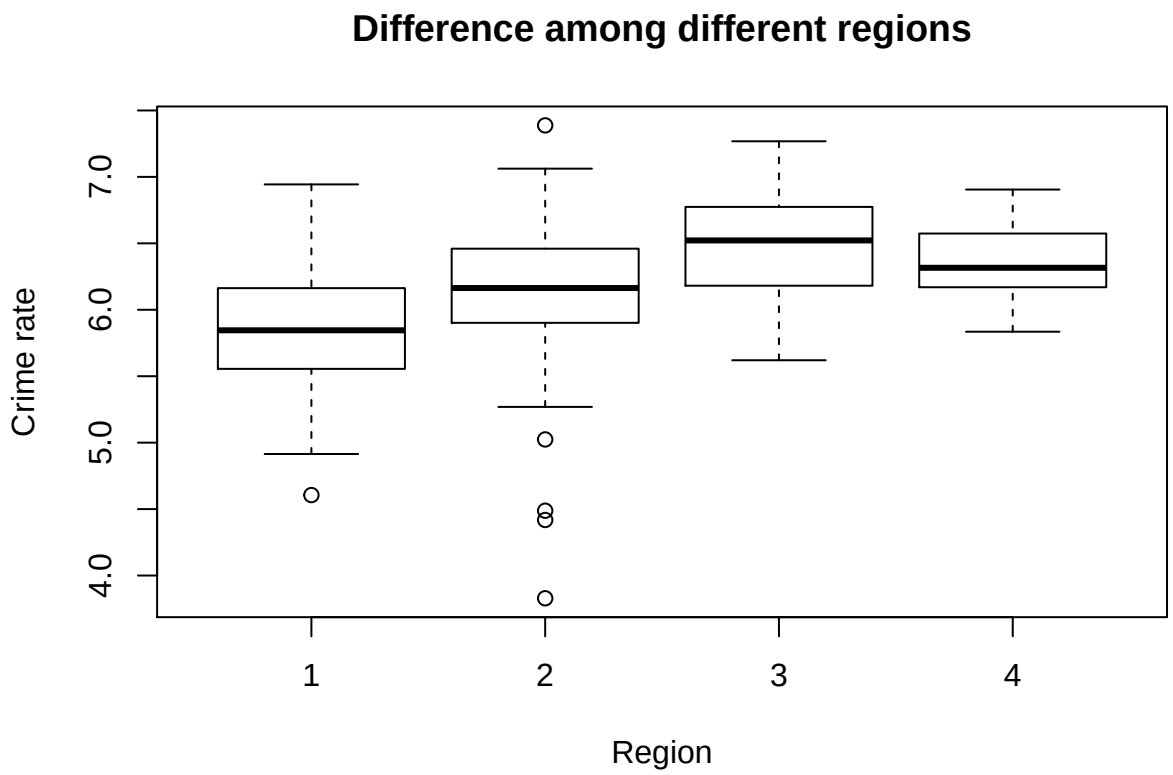
Part VI: Appendix

1. Predictors' Selection

Group 1: Region

Pick one from County, State, and Region.

```
# Boxplot against region
y <- log(crimerate)
boxplot(y~region, xlab = "Region", ylab = "Crime rate", main="Difference among different regions")
```



The boxplot shows a significant difference of crime rate among different regions.

```
# Least Significant Difference Method
pairwise.t.test(crimerate,region,pool.sd = T, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  crimerate and region
##
##   1       2       3
## 2 0.00024 -       -
## 3 < 2e-16 1.6e-07 -
## 4 2.5e-07 0.05146 0.00878
##
## P value adjustment method: none
```

```
# Bonferroni Method
pairwise.t.test(crimerate,region,pool.sd = T, p.adjust.method = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  crimerate and region
##
##   1       2       3
## 2 0.0015  -       -
## 3 < 2e-16 9.8e-07 -
## 4 1.5e-06 0.3087  0.0527
##
## P value adjustment method: bonferroni
```

```
# Tukey Method
TukeyHSD(aov(crimerate~region))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = crimerate ~ region)
##
## $region
##           diff        lwr        upr     p adj
## 2-1  137.10915   41.69917 232.519128 0.0013923
## 3-1  318.87124  230.97643 406.766046 0.0000000
## 4-1  217.40379  111.02235 323.785237 0.0000015
## 3-2  181.76209   94.22514 269.299039 0.0000010
## 4-2   80.29465  -25.79132 186.380610 0.2074102
## 4-3 -101.46745 -200.84886  -2.086035 0.0433404
```

In order to determine the region effect on crime rate, we try three different approaches: Least Significant Difference, Bonferroni, and Tukey method. In summary, difference between NE and S, NE and W, NC and S are significant, so region should be an important factor effecting crime rate.

- Least Significant Difference

Pairwise difference is significant except NC and W, with p-value of 0.1049.

- Bonferroni

According to the result, we are confident to conclude that the differences between NE and NC, between NC and W, and between S and W are not statistically significant, with p-values of 0.1330, 0.6295, and 0.1741.

- Tukey

Tukey method calculates the 95% confidence intervals for each difference. The confidence intervals for the differences between NE and NC, between NC and W, and between S and W contain 0, which implies that these 3 differences are not statistically significant. The result of Tukey and Bonferroni method are similar.

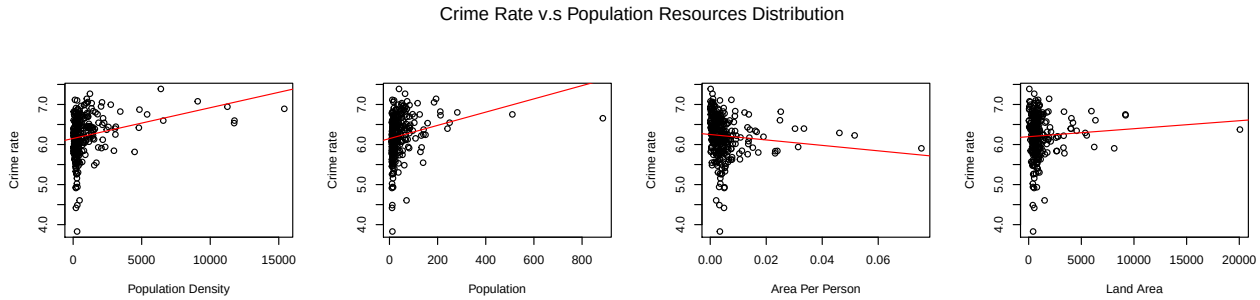
Group 2: Population Resources Distribution

Pick one from Total Population, Land Area, Population Density, Area Per Person.

In order to see which predictor has a strong influence on Crimerate, we did four plots to make a comparison.

```
par(mfrow=c(1,4), oma = c(0, 0, 3, 0))
plot(y ~ popdens, xlab="Population Density", ylab="Crime rate")
abline(lm(y ~ popdens), col=2)
```

```r
plot(y ~ pop,xlab="Population", ylab="Crime rate")
abline(lm(y ~ pop), col=2)

plot(y ~ avgarea, xlab="Area Per Person", ylab="Crime rate")
abline(lm(y ~ avgarea), col=2)

plot(y ~ area, xlab="Land Area", ylab="Crime rate")
abline(lm(y ~ area), col=2)
mtext("Crime Rate v.s Population Resources Distribution", side = 3, line = 0, outer = T,cex = 1)
```



Based on the figure above, we can easily see that Total Population and Population Density have strong positive relationship with Crime rate compared to other two.

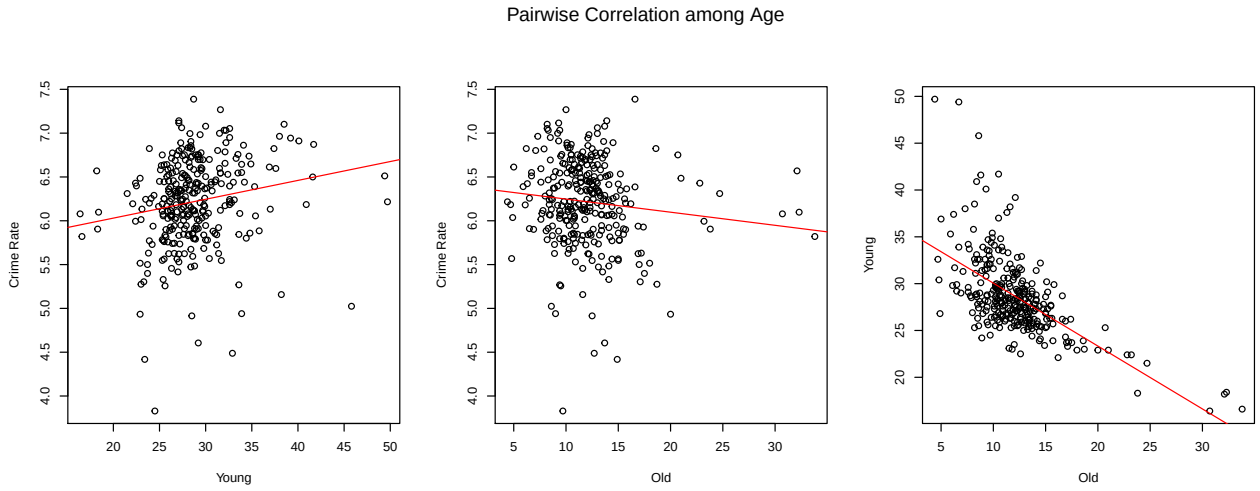
The reason why we don't pick Land Area as the trend shows a plain relationship between Land Area and Crime Rate; The reason why we don't pick Area Per Person is that this variable is equivalent to Population Density in some way, and it shows a weak relationship to Crime Rate.

In this case, we pick Total Population and Population Density.

** Group 3: Age (Percent)**

Now we need to choose from Young(18-34) and Old(65+):

```r
par(mfrow=c(1,3), oma = c(0, 0, 3 ,0))

plot(y ~ young, xlab="Young", ylab="Crime Rate")
abline(lm(y ~ young),col=2)

plot(y ~ old, xlab="Old", ylab="Crime Rate")
abline(lm(y ~ old), col=2)

plot(young ~ old, xlab="Old", ylab="Young")
abline(lm(young ~ old), col=2)

mtext("Pairwise Correlation among Age", side = 3, line = 0, outer = T,cex = 1)
```

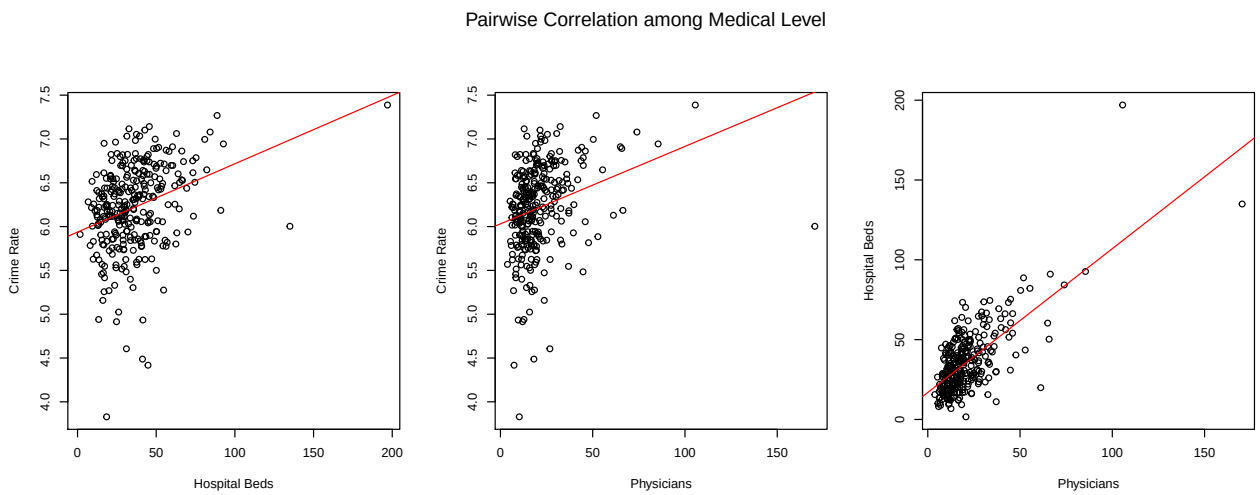


The plots show that they are all highly correlated, so we just randomly pick one. Here we pick Young(18-34).

** Group 4: Medical Level (Per person)**

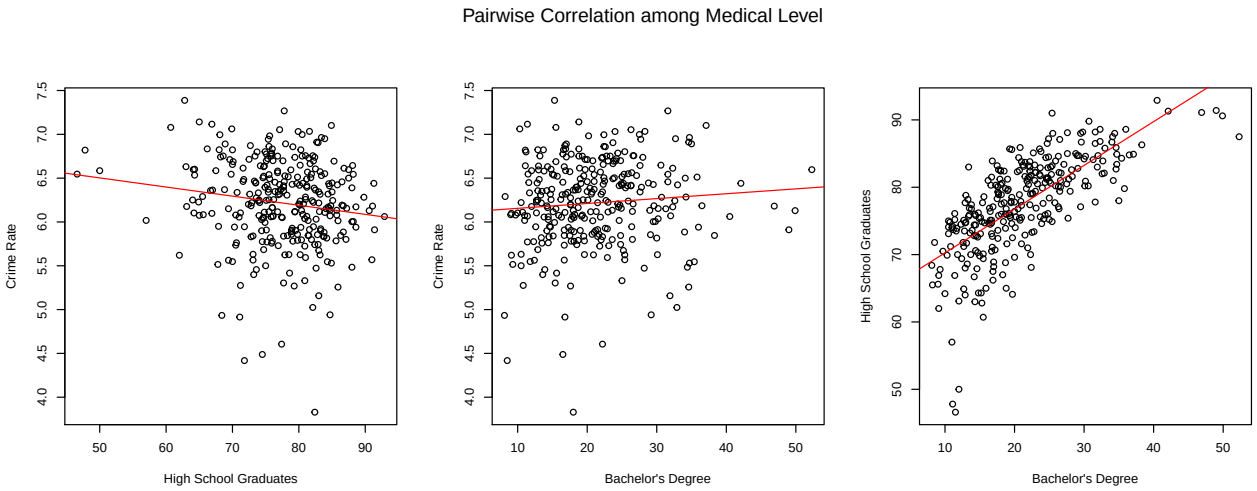
Similarily for Group 4, we need to choose from Physicians and Hospital Beds, and the plots show that they are all highly correlated, so we just randomly pick one. Here we pick Hospital Beds.

```
# highly correlated, pick one (e.g. avgbeds)
par(mfrow=c(1,3), oma = c(0, 0, 3 ,0))

plot(y ~ avgbeds, xlab="Hospital Beds", ylab="Crime Rate")
abline(lm(y ~ avgbeds), col=2)

plot(y ~ avgphys, xlab="Physicians",ylab="Crime Rate" )
abline(lm(y ~ avgphys), col=2)

plot(avgbeds ~ avgphys, xlab="Physicians", ylab="Hospital Beds")
abline(lm(avgbeds ~ avgphys), col=2)

mtext("Pairwise Correlation among Medical Level", side = 3, line = 0, outer = T,cex = 1)
```



**** Group 5: Education Level****

Similarily, for Group 4, we need to choose from High School Graduates and Bachelor's Degree, and the plots show that they are all highly correlated, so we just randomly pick one. Here we pick High School Graduates.



**** Group 6: Wealth Level****

Poor, Unemployment, Per Capita Income, Total Income

```
# par(mfrow=c(1,3), pin=c(1.4,1.4))
# plot(log(data[,c(13:16)]))
cor(data[,c(13:16)])
```

```
##                poor      unemp     income      totinc
## poor    1.00000000  0.4268144 -0.5955298 -0.02450506
## unemp   0.42681435  1.0000000 -0.3483085 -0.02224950
## income -0.59552980 -0.3483085  1.0000000  0.34838439
## totinc -0.02450506 -0.0222495  0.3483844  1.00000000
```

Pick between poor and income, because their relationship is too strong, then we pick between income and total income, because their relationship is too strong.

2. Multicollinearity

```
library(usdm)
vif.after <- vif(data.frame(popdens,pop,young,avgbeds,highgrad,poor,unemp))
vif.after$r2 <- sqrt(vif.after$VIF)
vif.after # It's good now after the selection.
```

```
##   Variables      VIF       r2
## 1   popdens 1.266706 1.125480
## 2       pop 1.107204 1.052237
## 3     young 1.296904 1.138817
## 4   avgbeds 1.357930 1.165303
## 5  highgrad 2.775240 1.665905
## 6      poor 2.491334 1.578396
## 7     unemp 1.737280 1.318059
```