**R Code for Examples in the book**



***“Statistics: The Art and Science of Learning from Data”***

**by Agresti, Franklin and Klingenberg, 5th edition**

**Chapter 3**

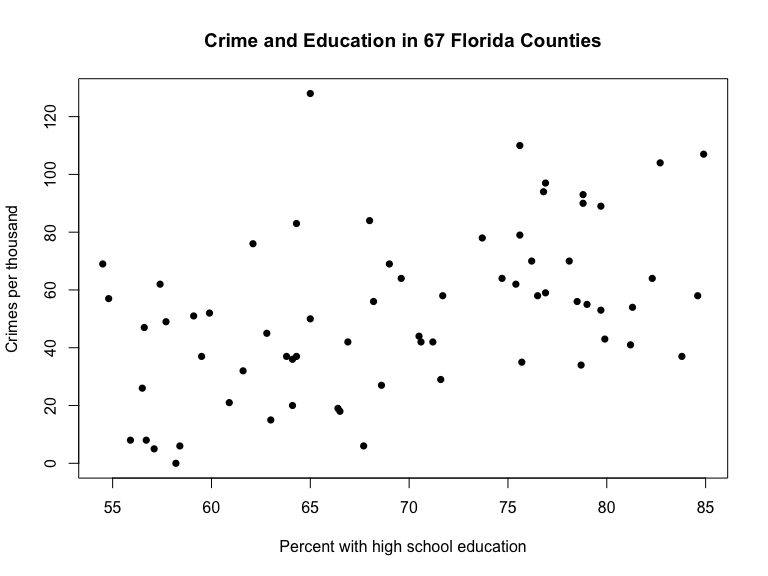
**Example 15: Education and Murder – Correlation and Causation**

## Reading in the data

crime <- read.csv(file='https://raw.githubusercontent.com/artofstat/data/master/Chapter3/fl\_crime.csv')  
attach(crime) # so we can refer to variable names

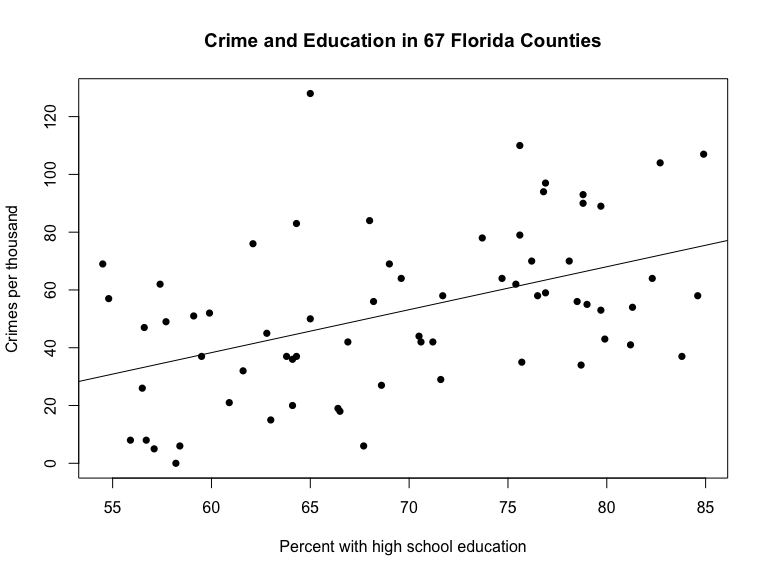
## Basic scatterplot of crime rate and percentage with at least a high school education

plot(x = education...., y = crime.rate..per.1000., pch = 16,  
 main = 'Crime and Education in 67 Florida Counties',  
 xlab = 'Percent with high school education',  
 ylab = 'Crimes per thousand')



## Fitting in regression model and adding to plot

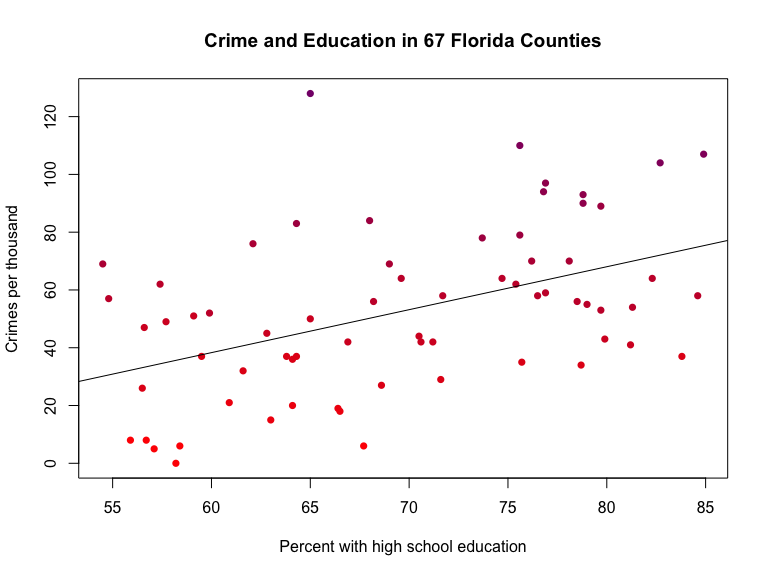
lm.reg <- lm(crime.rate..per.1000. ~ education....)  
plot(x = education...., y = crime.rate..per.1000., pch = 16,  
 main = 'Crime and Education in 67 Florida Counties',  
 xlab = 'Percent with high school education',  
 ylab = 'Crimes per thousand')  
abline(lm.reg)



## 

## This adds a column of color values based on the y values

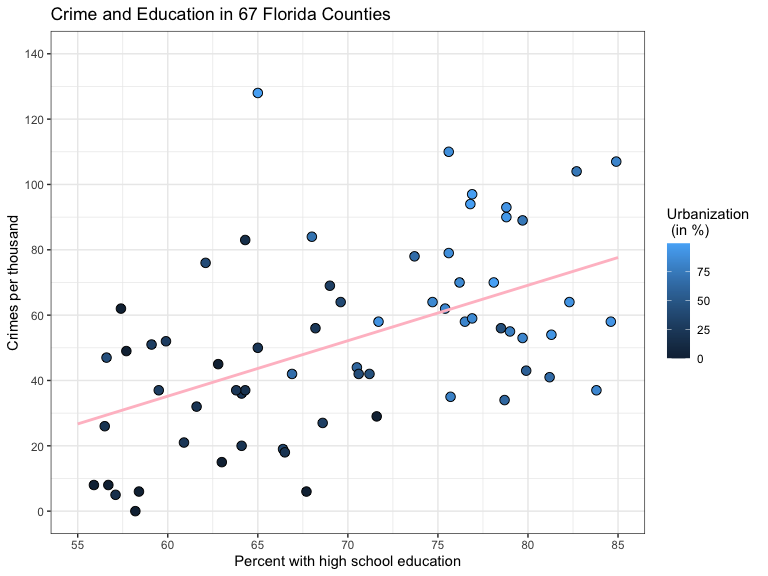
rbPal <- colorRampPalette(c('red','blue'))  
crime$Col <- rbPal(20)[as.numeric(cut(crime$crime.rate..per.1000.,   
 breaks = 10))]  
plot(x = education...., y = crime.rate..per.1000., pch = 16, col = crime$Col,  
 main = 'Crime and Education in 67 Florida Counties',  
 xlab = 'Percent with high school education',  
 ylab = 'Crimes per thousand')  
abline(lm.reg)



## 

## Scatterplot of crime rate and percentage with at least a high school education with dots colored according to the percentage of urbanization of a county

library(ggplot2)  
ggplot(crime, aes(x = education...., y = crime.rate..per.1000.)) +  
 geom\_point(aes(color = urbanization...., fill = urbanization....),   
 pch = 21, colour = 'black', size = 3) +   
 geom\_smooth(method = lm, se=FALSE, fullrange= TRUE, col = 'pink') +  
 labs(x = 'Percent with high school education',  
 y = 'Crimes per thousand',  
 title = 'Crime and Education in 67 Florida Counties',   
 fill = 'Urbanization \n (in %)') +  
 theme\_bw() +  
 scale\_x\_continuous(lim = c(55, 85), breaks = seq(55, 85, 5)) +  
 scale\_y\_continuous(lim = c(0, 140), breaks = seq(0, 140, 20))



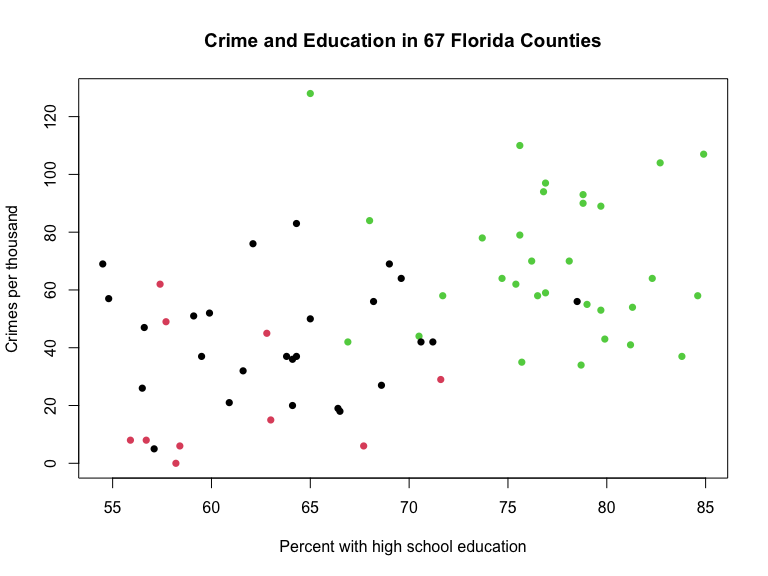
## 

## Adding Urbanization variable depending on urbanization percent using mutate() function from the dplyr package

library(dplyr)  
crime\_new <- crime %>%   
 mutate(Urbanization = case\_when(urbanization.... <= 15 ~ 'rural',   
 urbanization.... <= 50 ~ 'mixed',   
 urbanization.... > 50 ~ 'urban'))

## Basic scatterplot crime rate and percentage with at least a high school education with dots colored according to whether the county is rural, mixed, or urban

attach(crime\_new)  
plot(x = education...., y = crime.rate..per.1000., pch = 16,  
 col = factor(Urbanization),  
 main = 'Crime and Education in 67 Florida Counties',  
 xlab = 'Percent with high school education',  
 ylab = 'Crimes per thousand')



## Separating observations for rural, mixed, and urban counties

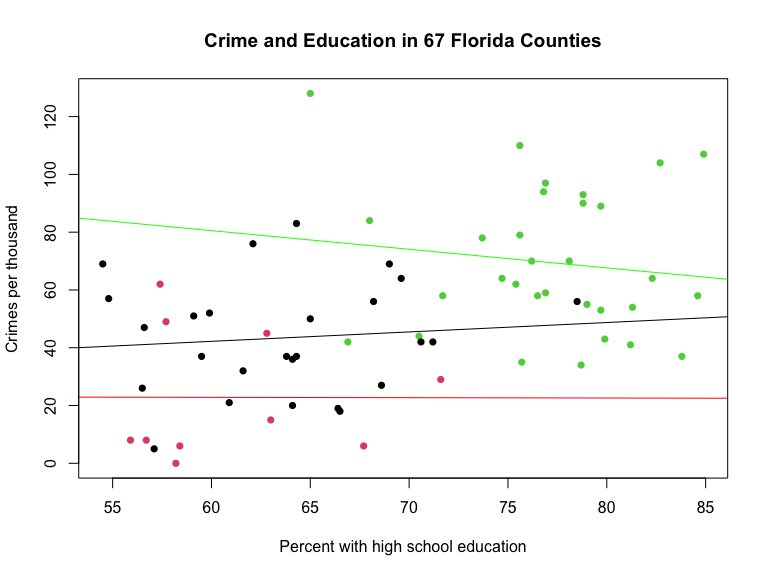
rural\_observations <- subset(crime\_new, Urbanization == 'rural')  
mixed\_observations <- subset(crime\_new, Urbanization == 'mixed')  
urban\_observations <- subset(crime\_new, Urbanization == 'urban')

## Fitting in corresponding regression models for rural, mixed, and urban counties

lm\_rural <- lm(crime.rate..per.1000. ~ education...., data = rural\_observations)  
lm\_mixed <- lm(crime.rate..per.1000. ~ education...., data = mixed\_observations)  
lm\_urban <- lm(crime.rate..per.1000. ~ education...., data = urban\_observations)

## Adding the regression equations to the plot

plot(x = education...., y = crime.rate..per.1000., pch = 16,  
 col = factor(Urbanization),  
 main = 'Crime and Education in 67 Florida Counties',  
 xlab = 'Percent with high school education',  
 ylab = 'Crimes per thousand')  
abline(lm\_rural, col = 'red')  
abline(lm\_mixed, col = 'black')  
abline(lm\_urban, col = 'green')



## Using the ggplot2 package to make the same scatterplot

ggplot(crime, aes(x = education...., y = crime.rate..per.1000.)) +  
 geom\_point(aes(shape = Urbanization, color = Urbanization,  
 fill = Urbanization), size = 3) +  
 geom\_smooth(method = lm, se = FALSE, fullrange = TRUE,   
 aes(color=Urbanization)) +  
 theme\_bw() +   
 labs(title = 'Crime and Education in 67 Florida Counties',  
 x = 'Percent with high school education', y = 'Crimes per thousand')

