# **Data Visualization Assignment**

# Introduction

Visualizations of data provide a more complete picture than descriptive statistics alone. The goal of the assignment is for you to become familiar with the specific qualities of the variables that you use in the visualizations. In this assignment you will produce univariate and bivariate data visualizations.

For this assignment, you will use the Vital Signs data from 2010. For the Decision Tree Assignment, you will use the 2010 data to build a model for rates of violent crime. As you work with the variables think in terms of the variables that would have an impact on rates of violent crime and about what you would like to know about the variables that you might add to the model.

## Directions

The information for the assignment is primarily from the lecture material. Use the assigned readings from Lesson 4, *Practical Data Science with R,* Part 1, Chapter 3, Section 3.2, *Beginning Data Science with R*, Chapter 4, Section 4.3 and Chapter 5, Section 5.2.2, and *R in Action*, Part 2, Chapter 6, Sections 6.1, and 6.3 along with the lecture material and from Lesson 4 as a guide for answering the following questions. (For question 3, *R in Action* Chapter 8, Section 8.2.4 has a short section on scatterplot Matrix if you would like additional information. This is optional and is not assigned reading.)

Use the Vital Signs 2010 data linked in Lesson 4 to answer the questions below. **Note:** the file will import in tbl\_df format. To convert it to a dataframe to use in the assignment, use vs10<-as.data.frame(VS10)

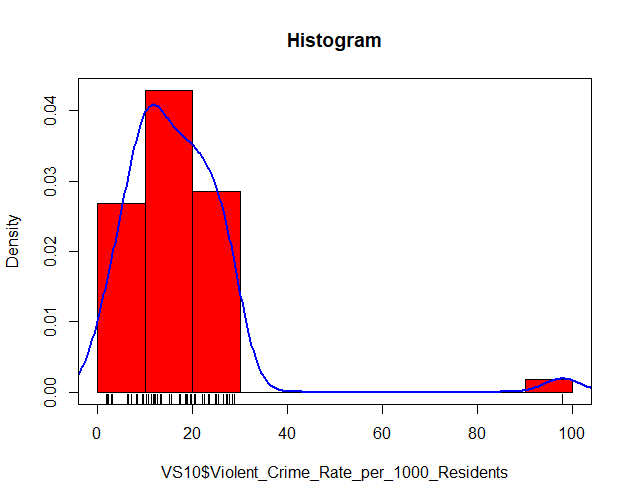
For your answers, make sure to reference the description of the variables on the Vital Signs web page so that you have a clear understanding of the unit of measure of the variable and exactly what the variable is. Make sure to convey this understanding in your answers.

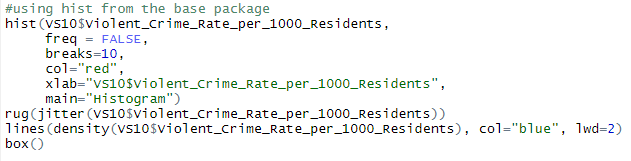
The completed assignment should adhere to the following guidelines:

* 1. Include your answers on the assignment document. Leave the assignment document as it is and include your answers after each question.
  2. Write your answers using complete sentences with correct punctuation, grammar, and spelling.
  3. Submit your completed assignment through the Blackboard portal in Lesson 4.

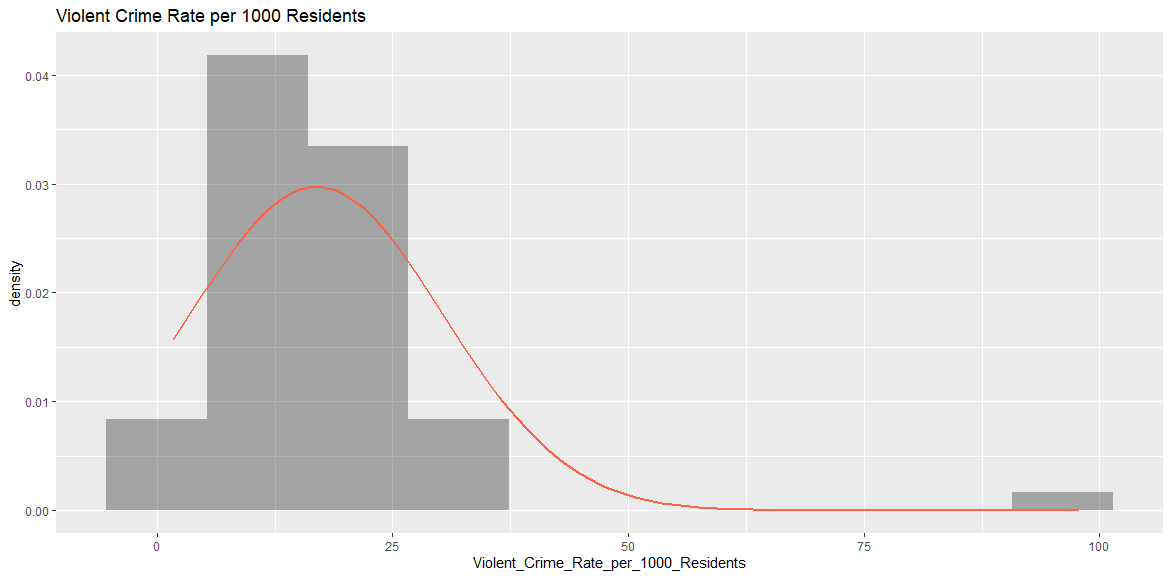
1. Use "Violent Crime Rate per 1,000 Residents" to produce histograms using both the base package and ggplot2. Overlay the histograms with a density curve and a normal distribution. Copy the images to the clipboard in bitmap format and paste the charts and your code below. Discuss what you observe about the distribution. *(3 points)*

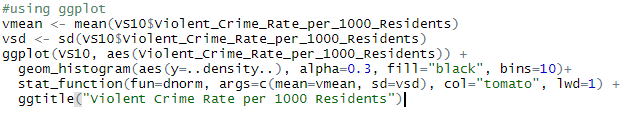
hist()





Ggplot()

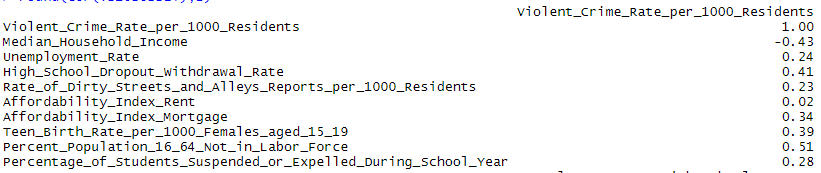


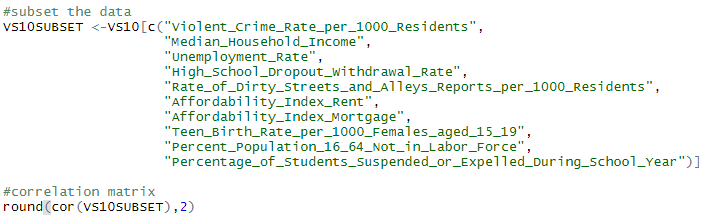


OBSERVATIONS:

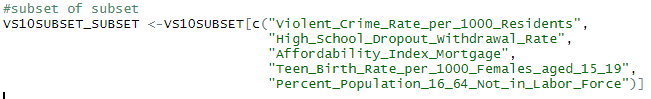
Crime appears to be correlated with areas of high population density.

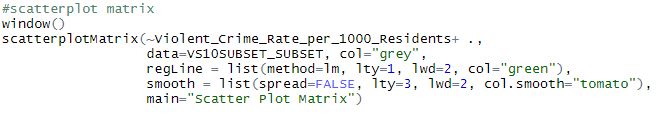
1. Create a dataframe that is a subset of nine variables from the vs10 dataset that you believe might have a relationship with violent crime plus the Violent Crime variable (so ten total variables for the subset). Run a correlation matrix using the **car** package. Paste the code and the output below. Discuss the sign and the magnitude of the four variables that are the most strongly correlated with violent crime. *(2 points)*

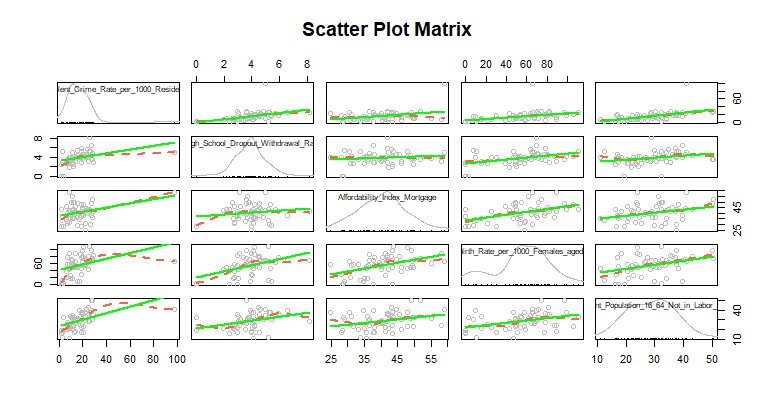
**



1. Subset the data again with these four variables and the violent crime variable. Using the **car** package again run a scatterplot matrix for this group of variables. Copy the image to the clipboard in bitmap format and paste the chart and your code below. Discuss what you observe about the scatter plots for violent crime and the four other variables. *(3 points)*

**





OBSERVATIONS:

There is a positive linear relationship between Crime Rate and School Dropout.

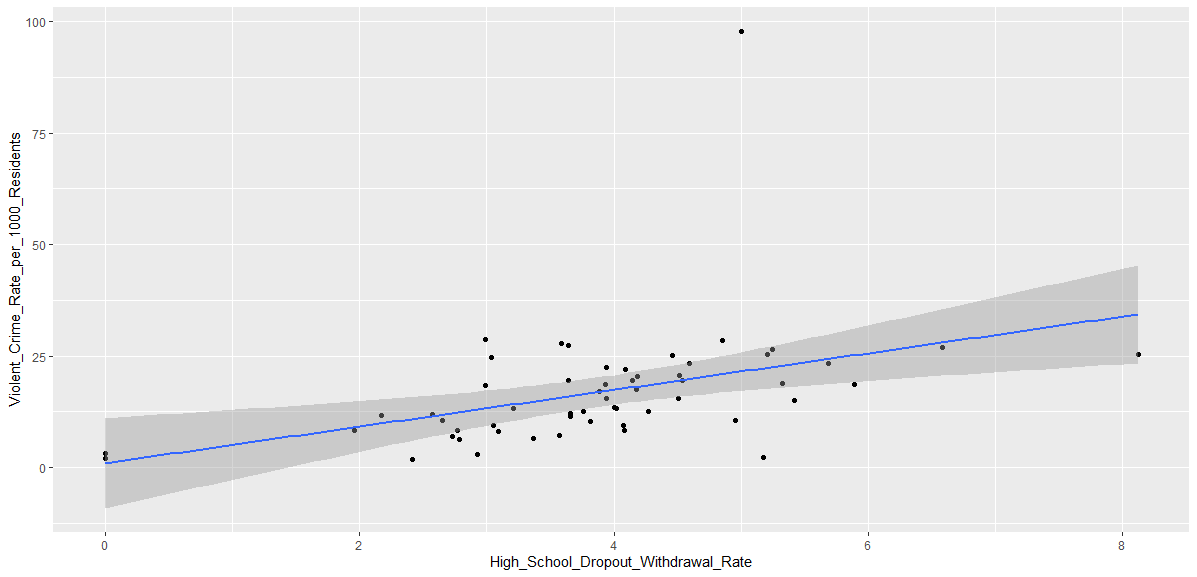
There is a slight, positive, linear relationship between Crime Rate and Mortgage index (note the higher values are not linear). This would be interesting to explore further.

There is a positive linear relationship between Crime Rate and Underage Birth Rate.

There is a positive linear relationship between Crime Rate and Population Not in Labor Force.

1. Select two of the variables from the scatter plot matrix and use ggplot to produce scatterplots with either a regression line or a smoothed line on the plot. Copy the images to the clipboard in bitmap format and paste the charts and your code below. Discuss what you observe about the two bivariate relationships. *(2 points)*

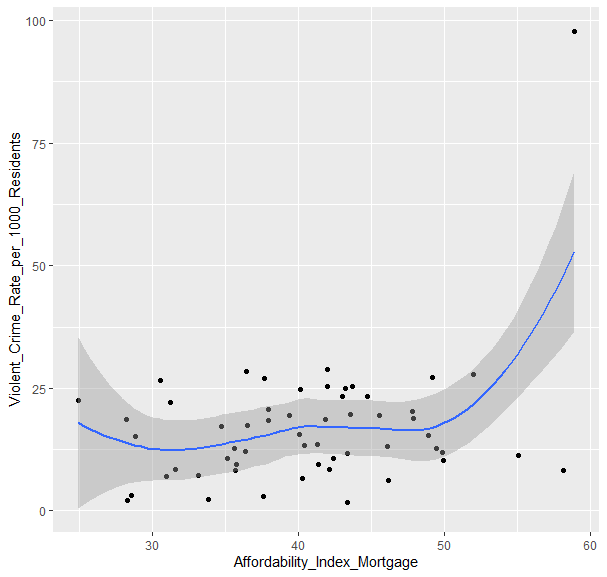




OBSERVATIONS:

This relationship is strongest around the rate of 4. Interestingly this relationship begins to weaken above 4. Also, it should be noted that there are two data points with a Zero drop-out rate. Further investigation is required in these two points.





OBSERVATIONS:

This non-linear relationship is interpretable between the range of 30 to 50. Outside that frame the model breaks down and is not useable.

## Scoring

To receive full credit, the visualizations should be constructed with the elements covered in the lecture or reading materials, complete with titles and any modifications to the content that enhance the visualization. For the discussion of the visualizations, use the specific results in the visualizations and in the Vital Signs documentation to report your observations in a complete and clear way.