The web address for Calvin’s RStudio is <http://rstudio.calvin.edu>.

Go to Packages and click on the **mosaic** box. This will load the mosaic and mosaicData

packages that we will be using for the course. There are some other packages that we will be

using later.

Data sets in R are in the form of a **data frame**. Each column represents a variable; the column heading is the name of the variable.

> USArrests

Murder Assault UrbanPop Rape

Alabama 13.2 236 58 21.2

Alaska 10.0 263 48 44.5

Arizona 8.1 294 80 31.0

Arkansas 8.8 190 50 19.5

California 9.0 276 91 40.6

Colorado 7.9 204 78 38.7

Connecticut 3.3 110 77 11.1

Delaware 5.9 238 72 15.8

Florida 15.4 335 80 31.9

Georgia 17.4 211 60 25.8

Hawaii 5.3 46 83 20.2

Idaho 2.6 120 54 14.2

Illinois 10.4 249 83 24.0

Indiana 7.2 113 65 21.0

Iowa 2.2 56 57 11.3

Kansas 6.0 115 66 18.0

Kentucky 9.7 109 52 16.3

Louisiana 15.4 249 66 22.2

Maine 2.1 83 51 7.8

Maryland 11.3 300 67 27.8

Massachusetts 4.4 149 85 16.3

Michigan 12.1 255 74 35.1

Minnesota 2.7 72 66 14.9

Mississippi 16.1 259 44 17.1

Missouri 9.0 178 70 28.2

Montana 6.0 109 53 16.4

Nebraska 4.3 102 62 16.5

Nevada 12.2 252 81 46.0

New Hampshire 2.1 57 56 9.5

New Jersey 7.4 159 89 18.8

New Mexico 11.4 285 70 32.1

New York 11.1 254 86 26.1

North Carolina 13.0 337 45 16.1

North Dakota 0.8 45 44 7.3

Ohio 7.3 120 75 21.4

Oklahoma 6.6 151 68 20.0

Oregon 4.9 159 67 29.3

Pennsylvania 6.3 106 72 14.9

Rhode Island 3.4 174 87 8.3

South Carolina 14.4 279 48 22.5

South Dakota 3.8 86 45 12.8

Tennessee 13.2 188 59 26.9

Texas 12.7 201 80 25.5

Utah 3.2 120 80 22.9

Vermont 2.2 48 32 11.2

Virginia 8.5 156 63 20.7

Washington 4.0 145 73 26.2

West Virginia 5.7 81 39 9.3

Wisconsin 2.6 53 66 10.8

Wyoming 6.8 161 60 15.6

To get information about the data in this data frame

> help(USArrests)

A data frame with 50 observations on 4 variables.

|  |  |  |  |
| --- | --- | --- | --- |
| [,1] | Murder | numeric | Murder arrests (per 100,000) |
| [,2] | Assault | numeric | Assault arrests (per 100,000) |
| [,3] | UrbanPop | numeric | Percent urban population |
| [,4] | Rape | numeric | Rape arrests (per 100,000) |

**BASIC PLOTTING IN R**

1. **HISTOGRAMS**

The basic picture for quantitative data is a histogram.

There are three types of histograms:

Frequency Histogram

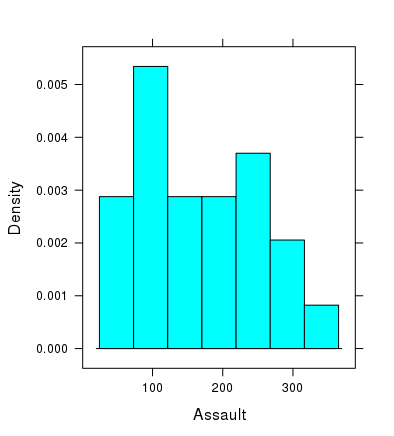
Relative Frequency Histogram

Density Histogram

To plot a histogram of the data in the column with variable name VAR directly from the data frame FRAME

histogram(~VAR, data = FRAME)

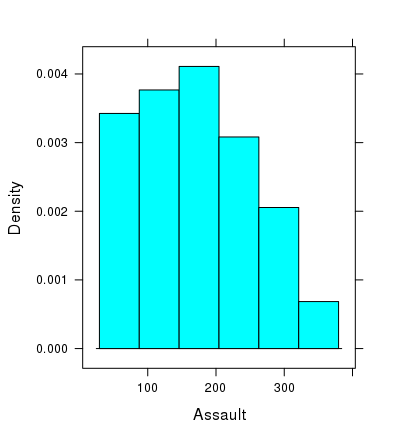
> histogram(~Assault, data = USArrests)



**To copy a picture to a Word document**: Under Export click “copy to clip board”. You can drag the picture into a Word document.

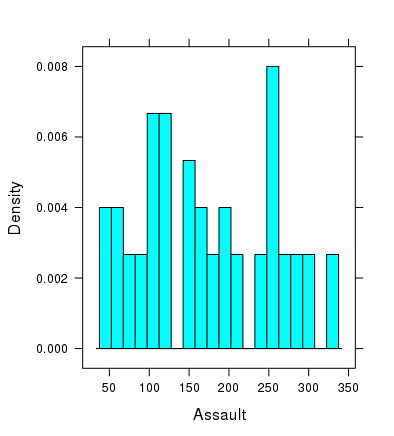
You can change the number of bins using **n**.

> histogram(~Assault, data = USArrests,n = 6)



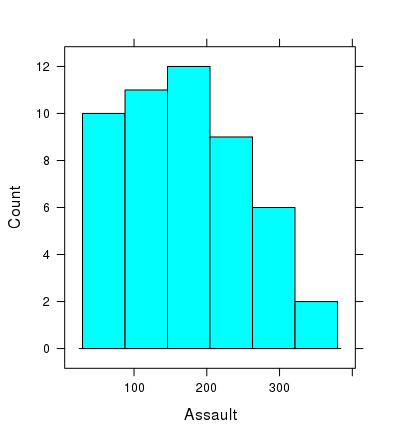
You can change the width of the bins using **width = size.**

> histogram(~Assault, data = USArrests,width=15)

****

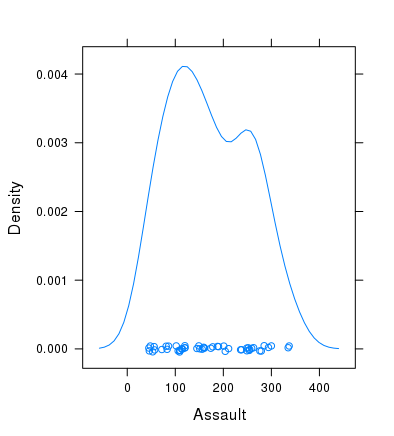
You can change to a frequency histogram using **type = “count”.**

> histogram(~Assault, data = USArrests,n=6,type="count")

****

You can get a picture of a “smoothed” version of a histogram using the command **densityplot.**

> densityplot(~Assault,data=USArrests)

****

**SHAPES OF DISTRIBUTIONS**

**Symmetric Skewed (right/left)**

**Uniform Unimodal Bimodal**

**Outliers**

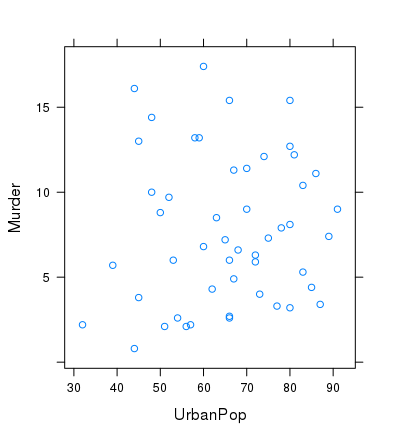
1. **SCATTER PLOTS**

Statisticians often want to explore the connection between two variables. The basic picture is a scatter plot

To create a scatter plot of the variable VARY (on vertical axis) with respect to VARX (on the horizontal axis) from the data frame FRAME

xyplot(VARY~VARX, data = FRAME)

> xyplot(Murder~UrbanPop,data=USArrests)



**EXERCISES 1**

* 1. The data frame **Galton** contains data on the heights (in inches) of parents and their adult children. Each row corresponds to an adult child. The variables give the family for the child, the heights of the father, mother, and child, the child’s gender, and the number of children in the family.

1. You can get the first 6 rows of the data frame with the variable names using the command: head(Galton). This gives the variables involved and a few of the values. Execute this command and include a copy of the output as your solution. (Use copy and paste).
2. Create and print out a histogram of the heights of the children. Does it look like this distribution is unimodal or bimodal?
3. Create and print out a scatter plot of the heights of the children (response or y-variable) versus the height of the father (explanatory or x-variable). Based on the scatter plot does it look like there is a relation between father’s heights and children’s height? If so, what is the nature of the connection? Is the connection strong or weak?
4. Do (b) using mother’s heights.
5. Some people claim that taller women tend to marry taller men and shorter women tend to marry shorter men. Create and print out a scatter plot to test this claim. Does the scatter plot provide strong evidence that this claim is true?
   1. The data frame **faithful** contains data about 272 eruptions of the geyser Old Faithful in Yellowstone National Park. For each eruption there are two values: length of eruption and the waiting time until the next eruption. Look at, but do not print out the data.
6. Create and print out a histogram of the eruption durations that contains 12 bins. Describe the type of distribution the duration has.
7. Repeat (a) for waiting times.
8. Create and print out a scatter plot of waiting times (response or y-variable) versus eruption times (explanatory or x variable). Based on this scatter plot, what can be said about the relation between the length of an eruption and the waiting time until the next eruption? What could be the explanation for this relation?